UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER THESIS

APPLICATION OF DATA MINING TECHNIQUES DURING MARKETING PLANNING STAGES

Ljubljana, April 2020 TAISIYA STEPANKINA

AUTHORSHIP STATEMENT

The undersigned Taisiya Stepankina, a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title "Application of Data Mining Techniques during Marketing Planning Stages", prepared under supervision of Jurij Jaklič and co-supervision of Tanja Dmitrović

D E C L A R E

1. this written final work of studies to be based on the results of my own research;

2. the printed form of this written final work of studies to be identical to its electronic form;

3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;

4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;

5. to be aware of the consequences a proven plagiarism charge based on this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;

6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;

7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;

8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;

9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;

10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, April 21st, 2020 Author's signature: 

(Month in words / Day / Year,
e. g. June 1st, 2012

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

MIT – Massachusetts Institute of Technology

MkIS – Marketing Information System

MDSS – Marketing Decision Support System

CHAID – Chi-square Automatic Interaction Detector

GBM –Gradient Boosting Machine

SVM – Support Vector Machine

FM – Factorization Machines

NN – Neural Networks

GMM – Gaussian Mixture Models

ANOVA – ANalysis Of Variance

PCA – Principal Component Analysis

QQ-plot – Quantile-Quantile Plot

WSS – Within-Cluster Sum of Square

RBTH – Russia Beyond The Headlines

ROI – Return On Investment

CEO – Chief Executive Officer

COO – Chief Operating Officer

CTO – Chief Technology Officer

BPMS – Business Process Management System

API – Application Program Interface

UX – User Experience

AI – Artificial Intelligence

CTR – Click-Through Rate

SQL – Simple Query Language

NoSQL – Not Only SQL

CRISP-DM – Cross-Industry Standard Practice for Data Mining

SWOT – Strengths, Weaknesses, Opportunities, and Threats

SEMMA – Sample, Explore, Modify, Model, and Assess

ETL – Extract, Transform, Load

KNN – k-Nearest Neighbors

RFM – Recency, Frequency, Monetary

CART – Classification and Regression Trees

MBR – Memory-Based Reasoning

CRM – Customer Relationship Management

OLAP – Online Analytical Processing

LTV – Lifetime Value

KPI – Key Performance Indicators

PESTEL – Political, Economic, Social, Technological, Environmental and Legal

WOM – Word of Mouth

BCG – Boston Consulting Group

BI – Business Intelligence

AIDA – Attention, Interest, Desire, Action

GDPR – General Data Protection Regulation

5C – Company, Customers, Competitors, Collaborators, Climate

INTRODUCTION

Marketing is the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large (AMA, 2013). In order to create value for all stakeholders, continuous monitoring, forecasting, and adapting to fast changing business environment is needed. Technological changes and development of Internet led to appearance of Web 2.0, which provoked introduction of e-marketing, defined by Chaffey, Ellis-Chadwick, Mayer, & Johnston (2009, p.9) as a "process of achieving marketing objectives through use of electronic communications technology". Another similar, broadly-used term is digital marketing which is very similar to e-marketing but is more focused on online channels to market – web, e-mail, databases, mobile, wireless/digital TV, according to Chaffey, Ellis-Chadwick, Mayer, & Johnston (2009).

In Web 1.0, businesses busied themselves with getting the basic internet technologies in place, so that they could establish a web presence, build electronic commerce capability, and improve the efficiency of their operations. In the era of Web 2.0, new systems and companies began taking advantage of the interactive nature of the Web (Provost & Fawcett, 2013). Comparing to Web 1.0, Web 2.0 includes a wide range of interactive tools and social communication techniques like blogs, podcasts, etc. (Chaffey et al., 2009)

Linoff and Berry (2011) stated long time ago that one of the reasons for learning to understand the behavior of individual customers is to be able to generalize so as to make predictions about the behavior of other, similar people. Twenty years later businesses became even more customer-oriented. Comparing to Web 1.0, the most noticeable changes are development of social networks and the rise of the "voice" of the individual consumer (Provost & Fawcett, 2013a). In 2004, building the business around customer.

Marketing management is defined by Kotler & Keller (2012, p.5) as "the art and science of choosing target markets and getting, keeping, and growing customers through creating, delivering, and communicating superior customer value". All of the above-mentioned factors change marketing management in terms of sources though which data is gathered and the volume of this data; research focus; set of tools and techniques available, and the depth and level of captured marketing insights.

Existence of Web 2.0 make people generating enormous amount of data at increasing pace. 18 years ago, 25% of the world's stored data was digital. Today 98% of all stored data is digital (Xu et al., 2016). Thus, companies generating terabytes of data should take advantage of using it (Linoff & Berry, 2011).

Generally, every decision made in a company is based on information gathered, internally and externally. The data is gathered though a mix of sources, among which are company

internal databases, and external sources which can be accessed via Internet – articles, publications, books, opinions buzz in social media.

Internal information forms a marketing information system (MkIS) which consists of people, equipment, and procedures to gather, sort, analyze, evaluate, and distribute needed, timely, and accurate information to marketing decision makers. It relies on internal company records, marketing intelligence activities, and marketing research (Kotler & Keller, 2012).

Internal company records of orders, sales, prices, costs, inventory levels, receivables, and payables are organized in databases which are combined and analyzed. Marketing research is the function that links the consumer, customer, and public to the marketer through information - information used to identify and define marketing opportunities and problems; generate, refine, and evaluate marketing actions; monitor marketing performance; and improve understanding of marketing as a process (AMA. October, 2004). It is done through qualitative and quantitative methods of data gathering, processing and analysis. Then, everyday information about developments in the marketing environment is gathered through marketing intelligence activities.

Marketing research can provide general direction of where to dig, and data mining can help to refine high-level strategies (Chiu & Tavella, 2008). While marketing research is mostly used to analyze data at macro level, data mining, defined as the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules, enables researchers to find such patterns and rules which are not obvious at the first sight (Linoff & Berry, 2011; Kotler & Keller, 2012). Now, in the consumer-focused era, a possibility to gather deeper marketing insights on individual levels is especially important.

Thus, data mining and marketing research are two complementary disciplines (Chiu & Tavella, 2008). Marketing intelligence then refers to information gained through marketing research and data mining, and manages it in a way it can be used for decision making.

Data mining is used for general customer relationship management to analyze customer behavior in order to manage attrition and maximize expected customer value (Provost & Fawcett, 2013). It can be used at every stage of marketing planning process.

Many authors divide marketing planning process into four stages and ten steps, including McDonald (2007): goal setting (mission, corporate objectives), situation review (marketing audit, market overview, SWOT analysis), strategy formulation (assumptions, marketing objectives and strategies, estimation of expected results and identification of alternative plans and tactical decisions), resource allocation and monitoring (budget, implementation programme). Sometimes measurement and review activities are considered to form a separate stage.

There is also a deeper-refined version of McDonald's marketing planning process which appears in different marketing online sources (Smartsheet, n.d.) and which consists of five

steps: mission, situation analysis, marketing strategy, marketing mix, implementation and control.

This master thesis will focus on data mining application at situation analysis, marketing strategy, marketing mix, and implementation and control stages, skipping the step one of defining the mission as this decision is not directly data-driven.

There are many books and journal articles written about data mining, big data, data mining for marketing and CRM. There are also different case studies examined in those books and research articles. However, there are few works which connect theory, practice and actual, newly written case studies of implementation of data mining techniques for every marketing planning process stage.

Thus, the purpose of this master thesis is to conduct a generic research and to bring together technical and business knowledge, covering data mining techniques implementation for different marketing planning process stages both from managerial and technical perspectives. To this end, the master thesis will cover each stage marketing planning process and will include theoretical framework with overview of existing techniques, practical part with real-life examples or experiments and their solutions.

This thesis is meant for business people who don't have enough theoretical background but do understand the necessity of applying data mining techniques while solving current business problems, and, oppositely, for data analysis who would like to get an overview of how data mining techniques can be implemented for answering business questions at situation analysis, designing marketing strategy and marketing mix, implementation and control stages of marketing planning process. The final chapter will also highlight current and arising marketing problems and their possible solutions for further investigation.

The goals of the master thesis are:

- To identify possible approaches and data mining techniques for each marketing planning process stage.
- To evaluate different approaches and data mining techniques through analyzing existing case studies, literature and own experiments with the data.
- To show practical example of data mining techniques used for marketing purposes.
- To summarize current findings about data mining techniques for digital marketing purposes.
- To design a matrix of proposed data mining methods to be used at each marketing planning process stage.

Exploratory research is the prevailing method of getting the first impression about the data and for conceptualization. Case studies, journal articles, user stories, books will be used as the main sources.

However, own experiments will be applied as well. These experiments will aim to answer particular marketing questions such as “which products are frequently bought together”, “what is the probability of an e-mail to be opened in the next marketing campaign”, “what is our target audience”, “what ads are the most successful in the selected channels”, and others. The data used for the experiments will be sourced from the open databases or gathered in other ways. Then, this data will be analyzed and presented via common data mining tools, such as: R Studio, RapidMiner, Power BI.

Data mining techniques which are going to be used include:

- Classification and segmentation, decision trees, rule induction
- Mining frequent patterns, associations and correlations
- Predictive analytics – regressions, neural networks, K-nearest neighbor
- Web scrapping, text mining, sentiment analysis
- Social media mining
- OLAP, digital analytics, web analytics from a data mining prospective

Thesis consists of nine chapters. Chapters 1-3 offer a brief overview of selected works about marketing, data mining, business intelligence, marketing intelligence, etc., and also an overview of commonly used data mining tasks and techniques.

Chapter 4 serves as a bridge between theoretical chapters and more practical ones, also introducing the Table 2 which is meant to complete the first goal of this master thesis.

Chapters 5-8 are mostly focused on practice. Each section represents one of marketing planning process stages and consists of a business problem statement, theoretical framework, practical applications (case studies) and/or experiments conducted. Situation analysis (chapter 5) includes opportunities identification, 5C (company, customers, competitors, collaborators, clients), SWOT, PESTEL analysis. Such techniques as competitors’ social media scraping, web scrapping, opinion mining, OLAP, common patterns identification, clustering and descriptive analytics based on existing databases are suitable for these purposes. Internal company’s databases, but also – news, blog entries, journal articles, open market researches, even tweets can serve as sources of data.

(Developing) marketing strategy (chapter 6) includes defining target audience, setting of measurable goals, budget development. Association rules, churn rate analysis, clustering and segmentation, market basket analysis will be applied at this stage. (Developing) marketing mix (chapter 7) includes product development, pricing, place, promotion, people and processes. OLAP, neural networks, correlation matrix, genetic algorithms can be applied at this step. Chapter 8 refers to execution of the plan, results monitoring and adjusting. This section of master thesis is focused on visualization techniques and prescriptive analytics.

Chapter 9 (Findings) represents the main findings drawn while writing this master thesis.

Conclusion is a review of goals of this thesis and explanations whether they were accomplished or not. Ideas for future examination and investigation discussed.

1. MARKETING: TRADITIONAL AND MODERN

1.1. Modern marketing origins

Kotler (2000) mentions managerial definition of marketing proposed by American Marketing Association (2013): marketing (management) is the process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods, and services to create exchanges that satisfy individual and organizational goals.

However, throughout years marketing concept has been developing, becoming broader and more holistic. There are several discussions of what marketing means nowadays. One of the definition proposed by marketing practicians is: "Modern Marketing is a holistic, adaptive methodology that connects brands with real customers and drives business results by blending strategy, creative, technology, and analysis" (Green, 2015). By combining broadly-used 'traditional' marketing definitions and those proposed by modern marketing companies, it is possible to conclude that marketing is a holistic and adaptive, customer-centric process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods, and services to create exchanges that satisfy individual and organizational goals.

The break point of the beginning of modern marketing can be considered to be in the beginning of 20th century, when many statistical techniques and theorems from diverse spheres, like econometrics, economics, social sciences, started to be applied by marketers in their decisions of how to sell. According to Wedel and Kannan (2016), for every marketing request, sooner or later, another technique was introduced or theorem implemented. Bayesian theorem, time-series, decision calculus, first automation systems and other statistical methods started to form a basic pool of techniques for data-driven marketing already in the early 1900th.

The introduction of the personal computer by IBM in 1981 enabled marketers to store the data on current and potential customers. In the late 1990, CRM software boomed. From that point up to the introduction of World Wide Web the late 20th century, data-driven analytics in marketing has progressed through approximately three stages: the description of observable market conditions through simple statistical approaches, the development of models to provide insights and diagnostics using theories from economics and psychology, and the evaluation of marketing policies, in which their effects are predicted and marketing decision making is supported using statistical, econometric, and Operational Research approaches (Wedel & Kannan, 2016).

Altogether, the process of using digital technologies to create new — or modify existing — business processes, culture, and customer experiences to meet changing business and market requirements is called digital transformation. While digitization is a process of moving from analog to digital, digital transformation starts and ends with a customer, and adds value to every customer interaction (Salesforce, n.d.).

Overall trend to business digital transformation provoked completely new marketing channels to emerge: online, Internet, digital marketing – broad yet very similar species. Chaffey & Smith (2017, p.13) emphasize that the term "digital marketing is now used worldwide by marketers to reference the range of digital media, technology and digital platforms used to reach and interact with consumers and businesses. It is at the heart of digital business – getting closer to customers and understanding them better, adding value to products, widening distribution channels and boosting sales through running digital marketing campaigns using digital media channels such as search marketing, online advertising and affiliate marketing… Digital marketing is a way of thinking, a way of putting the customer at the heart of all online activities".

Together with new forms of marketing, data-driven approach became essential for companies to survive. Different research centers (e.g. MIT), famous-in-the-field professionals emphasize that intuitive approach is not compatible with digital era and analytical approach lead to better overall company performance.

The routine capture of digital information through online and mobile applications produces vast data streams on how consumers feel, behave, and interact around products and services as well as how they respond to marketing efforts (Wedel & Kannan, 2016). Through collecting, storing and analyzing this data, marketers aim to fulfill their business function - build and maintain customer relationships; personalize products, services, and the marketing mix; and automate marketing processes in real time.

In 21st century, the era of Industry 4.0, marketers dive much deeper in investigation of consumer behavior. Every person is producing some data, and, along with traditional marketing research, marketers have got a unique opportunity to investigate their potential customers as close as it was never possible. Data has been called "the oil" of the digital economy. "Data is at the heart of many core business processes. It is generated by transactions in operational systems regardless of industry — retail, telecommunications, manufacturing, health care, utilities, transportation, insurance, credit cards, and financial services, for example" (Linoff & Berry, 2011, p.10).

With new marketing channels, new performance indicators were introduced. Among those are (Davis, 2017):

- Gross page impressions
- Word of mouth

- Total clicks
- Click through rate
- Cost per click
- Cost per action
- Pay per lead
- Activity ratio for social media
- Deductive social media return on investment
- Resolution time
- Social media profitability
- Bounce rate
- Return on advertising spend,

And many others. The full list of what can be measured online can be find in digital advertising and analytics, such as Facebook Ads, Google Analytics, etc.

KPIs are needed to make sure that marketing strategy is in align with firm's corporate goals and to understand how marketing activities contribute to company's overall success. There is no doubt that together with qualitative research, quantitative methods are highly useful to help in achieving established KPIs. Likely, nowadays qualitative research became cheaper and easier to hold because of introduction of online technologies. It also became easier to deal with the results of qualitative research as modern techniques allow to go beyond structured data.

Competition and other factors which influence the industry and a particular company stimulate businesses to think carefully about their competitive advantage. With high as never customer's expectations, companies try to find the most effective ways of reaching their potential customers.

After deciding which segments present the greatest opportunity, firms elaborate on designing customer-oriented marketing strategy and building mutually beneficial relationships with the customers. There are many theories and buzz around this topic in the era of social media and, from one side, consumer generated content and openness to the dialogue with the brand, and, from another side, the need to privacy protection, the need in some intimacy in the era of GDPR. That is why product positioning is extremely important in order to survive. To summarize, the main differences between 'traditional' and 'modern' marketing are:

- Modern marketing is more customer-centric. It seeks for the optimal point of the highest customer satisfaction and business goals to be achieved.
- Deriving from the first point, customer segmentation became more accurate, narrow and sophisticated, as company's products or services should not be annoying for customer. Selling to the market segments which are the most interested in the product or service is essential for modern marketing companies.

- New technologies brought upgraded marketing mix. It now also includes digital distribution channels, and that leads to completely new types of advertising available.
- The whole new concept of digital marketing was formed. It includes not only digital distribution channels but also a variety of tools for market segmentation, customer behavior tracking, and analytics.
- Since social media became a part of everyday life, marketing is not a monologue anymore but rather a dialogue with consumer. Analyzing customers' digital identities, marketers find better segments to sell the product or service to.
- Customer relationships management has changed with adoption of digital technologies and rising amount of consumer-generated content.
- As a consequence, marketing environment became data-rich, which led to invention of new tools, methods and techniques for holding and analyzing marketing data. Generally, marketing became more holistic. It interferes all stages of business development, starting from strategy and product elaboration and finishing with retaining existing and attracting new ones.

A simple, five-steps marketing process is described by Kotler and Armstrong (2012) as following:

- Understand the marketplace and customer needs and wants
- Design a customer-driven marketing strategy
- Construct an integrated marketing program the delivers superior value
- Build profitable relationships and create customer delight
- Capture value from customers to create profits and customer equity

1.2. Marketing planning process

Marketing planning process is essential for a company to map its mission with the means of how it will be completed. Marketing planning consists of several stages which different authors define in different manner, however, agreeing in what tasks each step should solve.

For example, Proctor (2000) divides the whole planning process on corporate planning and marketing planning.

Corporate planning consists of the following steps:

- Goal setting. Goals must be realistic and measurable
- Auditing, which includes all functional areas of management: marketing, production, finance, personnel. Marketing audit is divided on
    - Internal marketing audit comprises detailed analysis by product service of its market share, profitability. This stage, according to Proctor (2000), also includes marketing mix elements analysis through marketing research, and budgeting.
    - External marketing audit refers to external company's environment

- Gap analysis. The author emphasizes that it is important for company to forecast its demand, taking in account both external and internal factor that might affect company's performance.

Marketing planning is also divided by Proctor (2000) by several sub-plans, which are:

- A product mix plan which indicates product deletions, product modifications and product additions, when they are to occur, and the volume, turnover and profit objectives, broken down by product groups and even product items. Products may be grouped together and each grouping should have its own set of objectives (Proctor, 2000).
- A sales plan which aims to specify desired servicing levels for existing accounts and targets for obtaining the new ones.
- An advertising plan where the nature of advertising, target channels, communication objectives, and advertising amount are specified.
- A sales promotion plan, which by its nature is similar to marketing plan

In addition, Proctor (2000) states out that marketing plan includes several steps:

- Setting corporate objectives
- Performing marketing audit
- Performing SWOT analysis
- Setting assumptions
- Setting marketing objectives
- Listing strategic options
- Appraisal of strategic options
- Recommendation of strategy
- Contingency plans
- Implementation plan
- Feedback/control

Another textbook describes similar marketing planning steps. McDonald (2007) determines the following phases in marketing planning process:

- Goal setting
  - Mission formulation
  - Corporate objectives definition
- Situation review
  - Marketing audit
    - External audit (investigation of business and economic environment, market, consumer, and competition)

        - Internal audit (deep-dive into marketing operational variables, such as sales, market shares, profit margins, etc.; investigation of marketing mix variables: product management, price, distribution, promotion; customer analytics)
    - SWOT analyzes
    - Assumptions
- Strategy formulation
    - Marketing objectives and strategies
    - Estimation of expected results
    - Identification of alternative plans and mixes
- Resource allocation and monitoring
    - Budgeting
    - First year detailed implementation program
- Measurement and review of achieved results

The two observed descriptions of marketing planning process stages are very similar. Several other authors define marketing planning process in a comparable manner. For this master thesis purposes, another conceptualization of marketing planning process is used, which includes five steps (Smartsheet, n.d.):

1. Mission: includes mission statement and corporate objectives definition.
2. Situation analysis: consists of opportunities identification, 5C Analysis (Company, Customers, Competitors, Collaborators, Climate), SWOT and PESTEL analysis.
3. Marketing strategy: includes target audience definition, measurable goals setting, and budget developing.
4. Marketing mix: consists of product development, pricing, promotion, place (distribution) defined.
5. Implementation and Control: consists of putting plan into action and monitoring the results.

As it can be concluded from the literature, in its essence, modern marketing planning process is not very different by its essence from the traditional marketing planning. What has changed is the set of tools used; marketing environmental variables; marketing mix complexity.

1.3. Mission

Mission is a statement through which a company transmits its goals, policies, and values to the customers. A good mission also aims to provide company's employees with a sense of purpose, direction and opportunities (Kotler, 2000). In addition, as Kotler (2000) claims, a well-thought corporate mission defines company's major competitive scopes, which are:

- Industry scope
- Products and applications scope

- Competence scope
- Market-segment scope
- Vertical scope
- Geographical scope

Writing a descent corporate mission involves a good understanding of environment in which company operates or will be operating. Getting to know competitive scopes could include any type of marketing research – qualitative and quantitative, and data obtained as a result of those activities could be analyzed using various data mining techniques, which are going to be discussed in latter chapters of this master thesis.

1.4. Situation analysis

The environment in which company operates is turbulent and hostile, and is a subject to a deep investigation (McDonald, 2007, p.30).

The process by which a company can understand how it relates to the environment is called a marketing audit. Through marketing audit, a company can identify its own strengths and weaknesses as they relate to external opportunities and threats. The marketing audit includes external and internal audits (McDonald, 2007).

The external audit's aim is to examine the forces operating outside the company (PESTEL), the market itself (total market, size, growth, trends; characteristics, developments; products, prices, physical distribution; channels, customers, communication, industry practices), and the competition (major competitors, sixe, market shares/coverage, market standing/reputation, production capabilities, distribution policies, marketing methods, key strengths and weaknesses, etc.

The internal audit includes marketing operational variables (own company, sales by different metrics, market shares, profit margins/costs, marketing procedures, marketing organization, marketing information/research, and marketing mix.

The objective of the audit is to indicate what a company's marketing objectives and strategies should be. One of useful tools to process the audit's findings is SWOT analysis (McDonald, 2007) which describes how the company operates internally, its strengths, weaknesses, opportunities and threats, referring to the global environment and the company's position among competitors.

After the situation analysis was performed, the decisions of what potentially could bring company more profit, which actions should be undertaken to increase company's values, are made.

Since the end of 20th century, marketing and marketplace where all stakeholders – suppliers, consumers, companies – are operating, has changed dramatically. In the era of industry 4.0 and digital transformation, the marketplace and the marketspace are two different subjects. The marketplace, according to Kotler (2000, page 5), is physical, while the marketspace is digital. Additionally, Kotler (2000, p.5) proposed a concept of metamarket which is a "cluster of complementary products and services that are closely related in the mind of consumers but are spread across diverse set of industries".

McDonald (2007, p.5) also states that "the matching process between a company's capabilities and customer want is fundamental to commercial success". This matching process is called 'a marketing environment', a milieu in which the firm is operating (McDonald, 2007, page 7).

From decade to decade the marketing environment focus was changing. Besides McDonald (2007) also puts a customer into the center of marketing environment hierarchy, Kotler, Kartajaya, and Setiawan (2017) distinguish different historical stages of marketing based on its focus: product-driven marketing (1.0), customer-centric marketing (2.0), human-centric marketing (3.0), and the newest marketing 4.0. In the Marketing 3.0, Kotler, Kartajaya, and Setiawan (2017) argued that the future of marketing lies in creating products, services, and company cultures that embrace and reflect human values. In Marketing 4.0, the authors' major premise is that marketing should adapt to the changing nature of customer paths in the digital economy. The role of marketers is to guide customers throughout their journey from awareness and ultimately to advocacy.

Thus, although marketing (starting from Marketing 2.0, using Kotler's classification) was always "the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large (AMA, 2013), nowadays, due to introduction of new technologies, the way of how the value is created and delivered to customers, and the way of how the value is captured from customers, has significantly changed. Building strong and mutually beneficial relationships with customers is now more important than ever.

1.5. Marketing strategy

1.5.1. Market segmentation

As it is impossible to maximize the value and satisfaction for all customers, marketers start with market segmentation to identify and profile distinct groups of buyers who might prefer or require varying products and marketing mixes (Kotler, 2000). After selecting the target markets, marketing offering (some combination of products, services, information, or experiences offered to a market to satisfy a need or a want) is designed and positioned to the target audience with the aim to fulfill its demands (human wants that are backed by buying

power with marketing offering (Kotler & Armstrong, 2012) and, therefore, bring the maximum satisfaction.

The general rule for 'market' definition is that it should be described in terms of a customer need in a way which covers the aggregation of all the alternative products or services which customers regard as being capable of satisfying that same need (Kotler, 2000). To design a successful marketing plan, it is crucial to define which markets, parts of markets, and consumers to serve. It is economically unprofitable for firms both to design a product for an individual or to try to satisfy all potential buyers. The solution here is to segment the market by identifying and profiling different groups of buyers who might prefer or require varying products and marketing mixes (Kotler, 2000).

Many authors, including those mentioned here, state that customer segmentation is one of the most important marketing planning procedures, and refer to the famous '5 Ws' model. The main goal of customer segmentation is to answer the questions: who buys, what do they buy, (when and where do they buy – these questions are answered at the stage of discussing marketing mix), and why do they buy. In order to find this out, marketers conduct marketing research and/or use data mining techniques, which are going to be further discussed later. Traditionally, marketing research was used to segment customers.

Kotler (2000) describes market-segmentation procedure which consists of three steps: survey, analysis and profiling. Survey stage includes exploratory interviews and focus groups to gain insight into customer motivations, attitudes, and behavior; a questionnaire through which data on attributes, brand awareness and other metrics is collected. Analysis stage refers mostly to applying different statistical techniques to create a specified number of maximally different segments. At profiling step, customer portrait is defined in terms of attitudes, behavior, demographics, psychographics, media patterns and other patterns, and each segment is given a name based on its dominant characteristic.

Every segment shares a pool of common characteristics: demographic, psychographic, behavioral, occasion, benefit, etc.; and common needs. Kotler and Armstrong (2012) state out that, besides above-mentioned, there is a difference between business and international markets segmentation.

Other than being segmented geographically, demographically (or in another way), business marketers might also use some additional variables, such as customer operating characteristics, purchasing approaches, situational factors, etc., while geographical (regional) segmentation might be more important for international markets, as well as economic, political, and legal factors. For international markets, Kotler and Armstrong (2012) introduce a term of intermarket, or cross-market segmentation, where consumers have similar needs and common buying behavior while being located in different countries.

As McDonald (2007, pp.162-163) suggests, a viable marketing segment has certain universally accepted criteria: 1) size of segments should be reasonable to be able to fulfill

company goals and to provide firm with the desired return for its effort; 2) members of each group should have a big share of common characteristics but be easily distinguished from the rest of the market; 3) segments criteria should be realistic in compliance with purchase situation, 4) segments should be reachable.

There are different ways and levels to segment the market (Kotler, 2000). Niche marketing addresses groups of people seeking for a distinctive mix of benefits or needs. Local marketing is focused on trading areas or neighborhoods; individual marketing in 21st century refers to a term of mass customization (the ability to prepare individually designed products and communications of a mass basis to meet each customer's requirements). Kotler (2000) also highlights three types of patterns of market segmentation. Homogeneous preferences, diffused preferences, or clustered preferences, also called natural market segments.

However, as Peppers and Rogers (2011) state, it is hard to identify customers only through surveys and focus groups due to particular business and consumers' specifics. There are additional sources to discover customers' information, well described by Peppers and Rogers (2011):

- Internal records
- Transactional data
- Records based on relationship marketing actions (offline events – mostly b2b sector)
- Open-source data which requires no customer active involvement in digital era, like social media
- Payable and free customer data databases

From another side, consumers differ not only by above-described characteristics but also by their value to the company.

Value is the ratio between benefits over costs, or sum of functional and emotional benefits divided by sum of monetary, time, energy, and physic costs. Thus, to increase the value for customer, marketers can 1) increase benefits 2) decrease costs 3) apply the mix of 1 and 2, 4) decrease the benefits less than reduce the costs. Then, choosing between two value offerings (V1/V2), a customer will prefer V1 if the ratio is larger than 1; V2 if it is less than 1, and he will be indifferent in the ration will be equal to one (Kotler, 2000, page 9).

Customer lifetime value (LTV) is one of the key metrics to measure while segmenting the customers. Customer's LTV is the "net present value of the expected future stream of financial contributions from the customer" (Peppers & Rogers, 2011, p.116).

As LTV is hard to model, one of widely-spread techniques to define current customer's value is a Recency (the last time customer purchased from a company), Frequency (how often does customer purchase), Monetary value (how much does customer spend) model.

Calculating these values is a part of database marketing. Modelling those help to decide upon which customers to pay more attention during marketing campaigns; which ones it makes more sense to enliven, and from whom there is a chance to capture more potential value.

As customers changed and became pro-active instead of just reacting to companies' actions, they started to communicate to enterprises what they are willing to pay for and what they want in return by themselves, pushing businesses to hear each individual voice, record it and implement different variations of mass customization strategies. Understanding each customer as an individual and treating him with respect makes company able to generalize and make predictions about the behavior of other, similar people (Linoff & Berry, 2011). The knowledge which company gets from individual customer relationships helps it to build better overall experiences.

However, the ability to generalize doesn't mean that customers won't be treated as individuals. The idea is to find common patterns in a group of individuals' behavior and make customers feeling that they are something more than a number. The strategy of what kind of relationships to build with a customer depends on what level of Maslow's pyramid a customer will satisfy and, thus, what value will be brought to a customer and through which value positioning.

1.5.2. Targeting

One of modern marketing trends is the way customers are treated which has significantly changed towards personalization. Wedel and Kannan (2016) define three main methods of personalization.

- Pull personalization provides a personalized service when a customer explicitly requests it. An example is Dell, which enables customers to customize the computer they buy in terms of pre-specified product features.
- Passive personalization displays personalized information about products or services in response to related customer activities, but the consumer has to act on that information. Recommendation systems represent another example of this approach.
- Push personalization takes passive personalization one step further by sending a personalized product or service directly to customers without their explicit request.

Thus, personalization can happen on three different levels: mass personalization, also known as mass customization; segment-level personalization which relates to sophisticated and accurate procedure of market segmentation, and individual personalization where each element of marketing mix is customized to personal customer's taste.

A powerful way of treating customer as an individual is a recommendation system. There are two basic types of recommendation engines: 1) content filtering based on the similarity

between customer's past preferences, and 2) collaborative filtering which makes predictions about customer's preferences using known preferences of similar customers.

Modern marketers are put in the situation when, from one side, customers indirectly declare their wishes and needs via different ways, and then companies need to make sure that the right service or product will find the proper customer. But at the same time, customers don't want to make effort on choosing a particular product within a range of products, and thus the responsibility to decide of what is the best fit for a particular customer's need or want is transferred to marketing teams. After all, modern marketers pursue one goal: to catch and to interpret correctly the message that customer indirectly sends about his preferences, and to propose the most suitable offers to him.

Targeting concept is connected to customer relationship management: once potential customers are found, how to treat them?

1.5.3. Customer relationship management

Customer relationship management is defined by Kotler and Armstrong (2012) as the overall process of building and maintaining profitable customer relationships by delivering superior customer value and satisfaction. It involves managing detailed information about individual customers and carefully managing customer 'touchpoints' to maximize customer loyalty. In the era of Marketing 4.0, marketers should cover every aspect of the customer's journey.

Generally speaking, key aspects of customer management are (Chaffey & Smith, 2017):

- Customer acquisition which means converting customers into qualified leads
- Customer retention: returning visitors are a very important KPI. There are many techniques to retarget the customer and remind him about the company, such as personalized and relevant e-mails
- Customer extension which related to selling other relevant products and services to the same customer.

As Kotler & Armstrong (2012) suggest, customer relationship management includes two important blocks: customer value (evaluation of benefits over costs) and customer satisfaction (how well a product's performance fits into customer's expectations). According to Maechler, Neher, & Park (2016), higher customer satisfaction leads to higher revenues for the company.

At the same time, Kotler and Armstrong (2012) mention that company's aim should not be to deliver the highest possible value to a customer (because it could be done by lowering the price, for example) but rather to maximize customer satisfaction.

The aim of relationships with customers is an exchange: customers get value driven by a service or product they purchase, while a company gets benefits in terms of revenues, brand

reputation, awareness, etc. Kotler (2000) suggests that marketing management is the process of planning and executing the conception, pricing, promotion, and distribution of ideas, goods, and services to create exchanges and satisfy individual and organizational goals. Customer relationship management serves as a glue between all described components.

Since companies became customer-centric and got much wider set of tools to study their customers, to dig into their minds, they got to think about how to make customer-company relationships not only profitable for both sides but also more 'ecological' in psychological way.

Kotler and Armstrong (2012) describe different levels of relations with customers, distinguishing them by the depth of these relationships. A basic level almost doesn't include personal interactions with customers through, for example, phone calls. Another extreme is a full partnership with key customers.

Kotler and Armstrong (2012) state that companies move towards deeper and more direct relationships with selected customers. Nowadays, this trend is developing in the direction of much more detailed customization as customers want to be treated not as numbers but rather like individuals with individual preferences and requirements (Peppers & Rogers, 2011).

Peppers and Rogers (2011) suggest, there are four different levels of customer relationships: intimate, like between a doctor and a patient; face-to-face, which don't require a disclosure of personal information; distant, which happen through an intermediary like phone of Internet; or no-contact, when customers interact with a distributor, like buying a brand of soda in a supermarket. It is important to understand here that this classification is based on the amount of personal information revealed by a customer to a provider of services/product manufacture and/or on personal contact.

The strategy of what kind of relationships to build with a customer depends on what level of Maslow's pyramid should be satisfied for a customer and, thus, what value will be brought to a customer and through which value positioning.

However, from marketing prospective, sometimes a customer may feel that he is treated as an individual by several brands much more than by a doctor. Peppers and Rogers (2011) highlight the importance of what a company understands under 'a relationship' and the level of emotional engagement to a product/service from a customer prospective.

Thus, when a right product/service finds a customer at a right time and place, and this product or service was designed and placed in accordance with customers' demands, but a customer doesn't have any emotional connection to a provider of this good or service, this is still considered to be some kind of relationship.

So, it is very important for companies to find a balance between being visible for a customer, but at the same time trying not to be intrusive.

1.5.4. Positioning

Kotler and Armstrong (2012) define product position as the way the product is defined by consumers on important attributes – the place the product occupies in consumers' minds relative to competing products. The differentiation and positioning task consists of three steps: identifying a set of differentiating competitive advantages on which to build a position, choosing the right competitive advantages, and selecting an overall positioning strategy. The company must then effectively communicate and deliver the chosen position to the market.

Targeting and positioning strategies are interrelated. "The choice of one or more target markets is based, at least in part, on the feasibility of the organization designing and implementing an effective positioning strategy to meet the target's needs" (Proctor, 2000, p.198).

1.5.5. Setting measurable goals and budgeting

It is important to establish reliable and achievable Key Performance Indicators (KPIs) to know how to measure the success of marketing activities. Properly defined KPIs will help to make sure that marketing strategy is aligned with firm's corporate goals and to understand how marketing activities contribute to company's overall success.

Common marketing metrics can be divided by categories (Davis, 2017):

- Corporate financial metrics (revenue, return on sales, net profit…),
- Marketing planning measures (market share, market demand, causal forecast…),
- Brand metrics (brand equity, brand scorecard…),
- Customer metrics (segment profitability, customer profitability, churn rate…),
- Product metrics (usage, marketing cost per unit),
- Price metrics, advertising/promotion metrics,
- Digital marketing metrics (cost per click, click-through rate, total clicks, bounce rate…),
- Place metrics (cost per sales dollar, transactions per customer, return to net sales….),
- Sales metrics (percent of sales, turnover rate…),

and many others. Chiu and Tavella (2008) propose different classification of metrics, dividing them into return metrics, investment cost metrics, operational metrics, and business impact metrics. The authors also emphasize the importance of understanding how these metrics contribute to business processes. Return metrics are often referred to as key performance indicators (KPI) or success metrics.

The costs of marketing programs, goods sold, and capital are investment cost metrics that must be optimally related to metrics measuring investment returns. Operational metrics influence the performance of return metrics (most of the metrics are considered to fall under

this category), and a thorough understanding of their impact on return metrics is essential in order to track those with the highest potential. One common mistake is to invest significant resources to track hundreds of operational metrics without precisely quantifying whether they significantly influence success (Chiu & Tavella, 2008).

1.6. Marketing mix

Marketing mix is set as an operationalization of strategy at tactical level and traditionally includes 4Ps: Product, Price, Place, and Promotion. Through years, theses four P's were extended to 7 Ps (Chaffey et al., 2009), which include elements that better reflect service delivery: People, Process and Physical evidence.

As Kotler (2000, p.9) suggests, "marketing mix is the set of marketing tools that the firm uses to pursue its marketing objectives in the target market". Kotler (2000) also parallelizes this concept with the Four Cs: customer solution, customer cost, convenience and communication.

P standing for product includes deep and constant product analytics which is also a part of marketing. New systems allow to track each user's action in the application or on the website which are products (or services) by themselves or represent those. Since 2000, the automated capture of online clickstream, messaging, word-of-mouth (WOM), transaction, and location data has greatly reduced the variable cost of data collection and has resulted in unprecedented volumes of data that provide insights on consumer behavior at exceptional levels of depth and granularity (Wedel & Kannan, 2016).

In addition, Chaffey and Smith (2017, page 3) claim that the extended Product in upgraded marketing mix contribute to the perceptions of quality and highlight the importance of 'building credibility before visibility'. The product audit is conducted from different perspectives. At the situation analysis point, macro- and microeconomic factors and their influence on the product are studies. At the macro level, it is important to understand the product lifecycle, its competitive advantage (using BCG matrix or other technique), its place in the consumers' system of values. It is related to theory of brand and of firm as a brand. In the micro-level, product analytics converts into constant tracking and monitoring consumer-product relationships.

Product analytics can be defined as "a specialized application of business intelligence (BI) and analytical software that consumes service reports, product returns, warranties, customer feedback and data from embedded sensors to help manufacturers evaluate product defects, identify opportunities for product improvements, detect patterns in usage or capacity of products, and link all these factors to customers. Product analytics can also incorporate feeds from social platforms to track complaints about products" (Gartner IT Glossary, n.d.). It is important to adjust products to meet customers' expectations in order to increase consumers' satisfaction and capture bigger value.

Pricing is another core thing to be defined as it is a monetary equivalent of product value communicated to the customer, from one side, and from another – a monetary equivalent of customer value. Finding an optimal pricing is one of big deals in marketing field. There are several pricing theories. Customer value-based pricing is set based on buyers' perceptions of value rather than on the sellers' cost (Kotler & Armstrong, 2012); good-value pricing is used when offering the right combination of quality and good service at a fair price, while value-added pricing encourages consumers to pay for feature which bring them additional value.

On another side of value-based pricing there is cost-based pricing which is set based on the costs for production, distribution, and selling the product, plus profit margin. Competition-based pricing is set based on competitors' strategies, prices, costs, and market offerings.

There are several pricing strategies such as market-skimming pricing, market-penetration pricing; product line, optional, captive, by-product pricing, etc. As pricing is closely related to demand and its elasticity (the sensitivity of demand to changes in price), it depends on many macro- and microeconomic factors, and thus it is not completely 'stationary' and is adjusted as situation changes. Some techniques of price adjustment include discounts, allowances, segmented pricing, promotional, international… Pricing analytics helps to discover and predict consumers' reaction to changes in price.

New buying models evoked new pricing approaches. A model offering a range of purchase options at different price points is widely used by many digital products. Pricing became more complex, and so did its management and monitoring.

Introduction of time series enabled marketers to monitor and notice every change in sales at every point of time. Well-known techniques taken from economics that were used in 20th century to identify optimal prices (working with elasticities and demand functions; moving average method, profit optimization, etc.) are still in use, but programming languages and software allow to do it much faster and with higher level of precision as different models can be created, compared and validated easier. Machine learning algorithms put pricing prediction to another level, and together with recommendation systems pushed the appearance of new services which propose to optimize advertising budget in social media or pricing automatically.

In order to reach target markets, marketers use marketing channels. Kotler (2000) distinguishes communication channels which aim is to deliver and receive messages to and from customers (meaning any type of media); dialogue channels (e-mail, messengers); monologue channels (ads); distribution, trade and selling channels. The last also include transactional intermediaries such as insurance companies and banks. One of the challenges which marketers face is a design problem in choosing the best mix of communication, distribution, and channels.

Marketing channels distribution should help company to achieve such goals as (Proctor, 2000): to create awareness of product or service; to build brand recognition; to evoke AIDA (attention, interest, desire, action) towards target product or service; to generate leads and convert leads into purchases; to maintain good retention rate (also part of building profitable relationships with customer through customer relationships management); to remind people about products' benefits; to inform about new product lines or features, etc.

Geo-fencing and retargeting are modern techniques to find new and retain 'old' customers in their natural places of inhabitance like Facebook or Instagram. Potential customer can come to a company himself through search marketing, which changes the definition of what is 'awareness' for companies. To be found, special keywords and meta-tags will be inserted into landing page. Recommendation systems based on his previous choices will help to fill the basket; chat-bot will automatically lead them through website navigation and will sort product out, understanding natural customer's language. Social media will work on all aspects of consumer lifecycle: brand awareness, brand recognition, engagement into product or service, lead conversion, purchase and post-purchase experience, and retention (Proctor, 2000).

P for Promotion includes sales promotion, advertising, sales force, public relations, direct marketing (Kotler, 2000), and gives marketers a lot of food for reflection as they need to decide upon the best mix of online and offline techniques. For some industries, winning customers (acquiring them) is still done in a traditional, offline way (for example, hospitality industry). Although networking is done offline, the awareness might be built through online channels.

Digital technologies facilitate large-scale field experiments that produce big data and have become powerful tools for eliciting answers to questions on the causal effects of marketing actions. For example, large-scale A/B testing enables firms to "test and learn" for optimizing website designs, (search, social, and mobile) advertising, behavioral targeting, and other aspects of the marketing mix. Hui et al. (2013) use field experiments to evaluate mobile promotions in retail stores (Wedel & Kannan, 2016).

'People' refer to all product stakeholders. Chaffey and Smith (2017, p.88) propose a simple yet effective scheme of "Happy Staff = Happy Customers = Happy Shareholders". A company is a big mechanism where all elements are interconnected, and nowadays people matter more than ever as 'people do business with people'. The authors argue that in the digital era automation of services through new technologies is always the best option. New tools like live chats, personalized e-mail campaigns, and others are of a great value for customers as they feel a personal attitude.

It is also stated that every employee is company's ambassador and, in a way, a salesperson. Employees' satisfaction and motivation are important drivers of high-quality customer

service. “People” needs to be budgeted as a part of marketing mix to ensure that company will provide the excellent customer experience.

P for Physical evidence, in offline, physical evidence refers to buildings, logos, and other elements of company’s branding which help to make a proper first impression and build initial customer’s trust. In online, physical evidence refers to company’s digital representation: adaptive, consumer-friendly website and landing page interface; relevant, informative yet not overloaded content structure; well-managed social networks with available and liable customers’ reviews. Online or offline, physical evidence needs to be managed continuously.

Chaffey and Smith (2017) reserve another ‘P’ for Process, referring to internal and external processes management as a part of marketing mix. Processes are continuous and include post-sales services and user experience, such as generating customer feedback, reviews and ratings; up- and cross-selling; product development and improvement. ‘P’ for processes also include all technical background behind product or service – e.g. interaction between frontend and backend; supply chain and inventory management, generating an extra competitive advantage resulting in high quality customer service.

In summary, modern marketing mix is an extended version of classical 4P model which goes beyond traditional marketing tasks and perceives a product or a service as a holistic ecosystem, where customers and their satisfaction are placed in the center. The process of building customer satisfaction starts with building a healthy environment within an organization and continues through amplified range of digital tools for marketing distribution channels, advertising, product analytics, pricing management, building trust through physical evidence, processes optimization and management.

2. GATHERING AND MANAGING MARKETING INTELLIGENCE

2.1. Marketing intelligence and marketing information systems

Business intelligence solution is a combination of strategy and technology for gathering, analyzing, and interpreting data from internal and external sources, with the end result of providing information about the past, present, and future state of the subject being examined (Oracle, n.d.). As a part of business intelligence, marketing intelligence term is introduced and defined as a set of procedures and sources utilized by managers in order to obtain day-to-day information about development in the marketing field (Uhl & Schoner, 1969). It refers to insights generated from marketing research or data mining and provides maximum value when its components and parts are weaved together to depict an overall picture of market opportunities and challenges (Chiu & Tavella, 2008).

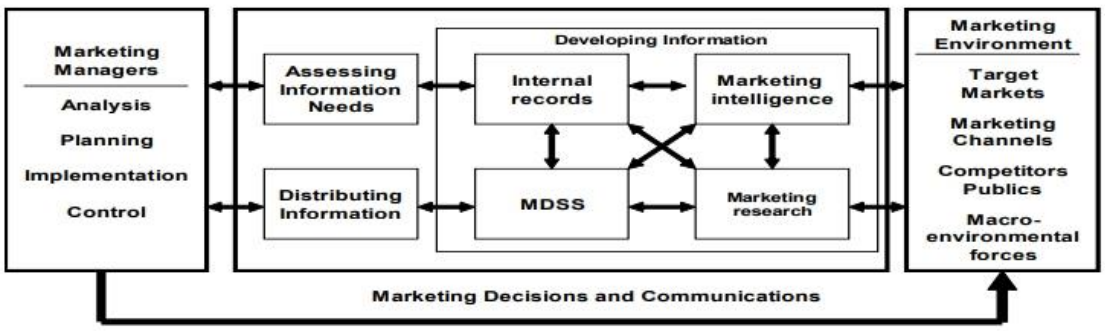
Marketing intelligence includes any data relevant for the company, gathered from Web, during networking events, by updating internal records and keeping them in order. Generating and collecting big amounts of data, companies need to store it somewhere for later usage. Internal information forms a marketing information system (MkIS) which consists of people, equipment, and procedures to gather, sort, analyze, evaluate, and distribute needed, timely, and accurate information to marketing decision makers. It relies on internal company records, marketing intelligence activities, and marketing research, as shown on Figure 1 (Kotler & Keller, 2012).

Uhl & Schoner (1969) highlight a difference between marketing intelligence and marketing information system (as many modern authors consider those to be synonyms). Marketing Information System is structured and oriented on internal environment of the organization, while Marketing Intelligence System is semi-structured and mostly operates in the external environment of the organization. While MkIS facilitates organizational decision-making by generating various reports and analyzing marketing trends, Marketing Intelligence system facilitates strategic planning and top management decision-making. It is oriented to future trends, while MkIS allows to analyze the situation in past and present.

During the last 50 years MkIS changed significantly, of course, due to introduction of new technologies and increased amount of data. Together with MkIS, another term was introduced – MDSS (marketing decision support system), which is defined as a set of core applications in the MkIS that provides computer-based tools, models, and techniques to support the marketing decision making process (Harmon, 2003).

MkIS supports every element of marketing strategy. Representing a holistic system of information, it contains internal company records (transactional and sales data, customer behavioral tracking data, salesforce feedback), data acquired through marketing intelligence activities, and primary data collected by company-sponsored marketing research. It allows company to monitor markets, competitor activities, changes in the environment and consumer behavior.

Figure 1: The marketing information system



Source: Kotler (1997).

MkIS also supports strategy development for new products, product positioning, marketing communications (advertising, public relations, and sales promotion), pricing, personal selling, distribution, customer service and partnerships and alliances. The MkIS provides the foundation for the development information system-dependent e-commerce strategies (Harmon, 2003). At the same time, MDSS provides models for forecasting, simulation, and optimization.

However, information on revenue growth, competitors, etc. in isolation does not provide significant value. Chiu & Tavella (2008, p.9) emphasize that to "facilitate building market and customer intelligence, it is necessary to have integrated database systems that link together data from sales, marketing, customer, research, operations, and finance". Figure 2 represents business intelligence architecture.

Currently there are multiple ways to organize company's databases. Generally, databases can be divided on relational and non-relational. Relational databases are the most used in the field. However, lately non-relational databases arose. Relational databases can be re-though as a collection of structured data in tables format. This type of databases takes its origins in 1970-s, and is operated by Structured Query Language (SQL). Non-relational, or NoSql databases, emerged as web applications became too complex to be hold in a table format.

The choice of type of database, database architecture, and tools to deal with the database depends on data volume and complexity (Hammink, 2018).

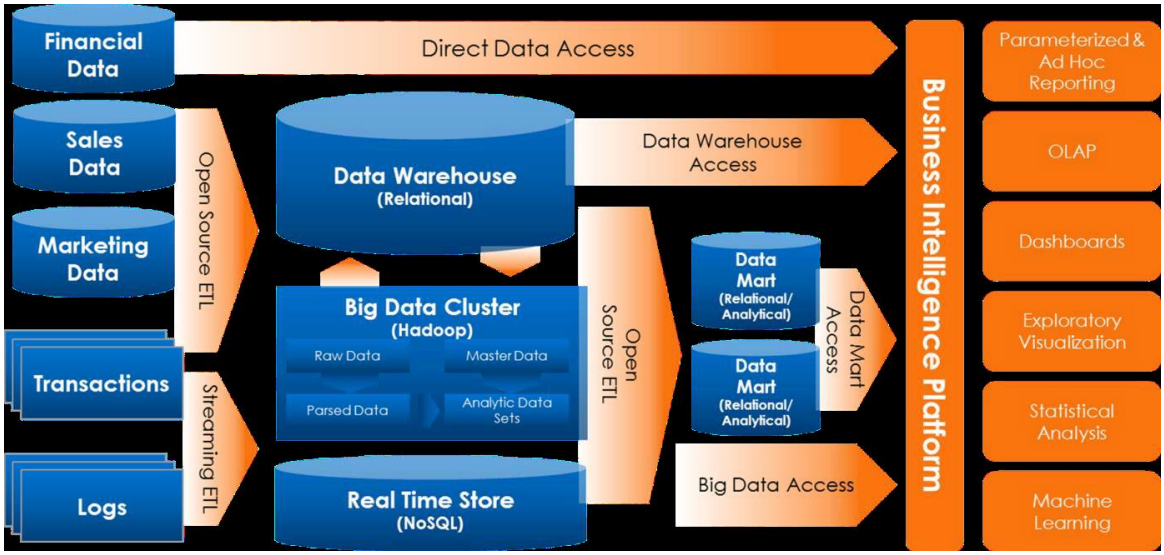
Obviously, it is possible to deal with small data using classical pool of tools for data analytics. However, for multi-source databases containing big amount of data classical tools might not be enough. Thus, data volume and its complexity leads to two related terms: big data and data mining.

Data mining is the process which helps to extract value from the data. It is defined as "a business process for exploring large amounts of data to discover meaningful patterns and rules" by Linoff and Berry, (2011, p. 2). Chaffey and Smith (2017, p.633) define data mining as "searching organizational databases in order to uncover hidden patterns or relationships in groups of data. Data-mining software attempts to represent information in new ways so that previously unseen patterns or trends can be identified".

Linoff and Berry (2011) emphasize that data mining process is ongoing. It starts with data, then through analysis informs or inspires action, which, in turn, creates data that begets more data mining.

A vicious circle of data mining is that it is not enough to capture data and get insights. "You must respond to the patterns and act on them, ultimately turning data into information, information into action, and action into value", - state Linoff and Berry (2011, p.22). The main goal of using data mining techniques in marketing is to gain the value: for customers, and from customers.

Figure 2: Business Intelligence architecture



Source: Eckerson (2011).

2.2. Data mining for marketing planning process

"Data mining tasks do not exist in a vacuum; they exist in a business context" (Linoff & Berry, 2011, p. 34). After the overall business context is examined during situation analysis stage, company should focus on building customer-centric marketing strategy, as described in the part one of this master thesis.

Business relationships with customers are developing through time. As Linoff and Berry (2011) suggest, customers pass through five major phases:

- Prospects are in the target market but not yet customers
- Responders are prospects who have exhibited some interest
- New customers are responders who have made a commitment (have made a first purchase, signed a contract…)
- Established customers are returning new customers with whom relationship is broadening or deepening
- Former customers are those who have left

The first step is to decide how a customer is going to be acquired. Identifying good prospects, choosing a communication channel for reaching prospects, picking appropriate messages for different groups of prospects are some business problems which could be solved by corresponding data mining tasks, such as building a response model which will allow marketers to estimate the probability of a prospect to respond to a marketing campaign. To achieve the goal, each prospect is being ranked using such techniques as regression models, decision trees, and neural networks. Every response can then be translated into profit terms, contributing to campaign profitability analysis.

Business processes help customers to transfer from one stage of the lifecycle to another, making them more valuable over time (Linoff & Berry, 2011). When customers are acquired, their data is used to learn about potential customers, as well. However, it is important to keep in mind that such approach might be misleading when forgetting that current customers reflect past marketing decisions, so it is important to choose to learn on that might be the most common for current and future customers. For these purposes, customer segmentation is conducted.

Segmentation techniques will also be useful in customizing the bundle of marketing messages addressed to a particular group of existing customers for future marketing campaigns. Then, churn rate prediction can be used to finding customer segments where customers are least likely to churn, and then to focus retention strategy on them.

Customer retention form a large pool of business tasks, such as cross-selling, up-selling, customer LTV estimation, and making recommendations. These business problems can be translated in such data mining activities as classification and clustering.

2.3. Marketing research and data mining

In order to gain marketing insights and estimate the efforts, marketing research is conducted. While marketing intelligence is a continuous process, marketing research is situational and problem-oriented (Harmon, 2003).

AMA (2013) defines marketing research as "the function that links the consumer, customer, and public to the marketer through information–information used to identify and define marketing opportunities and problems; generate, refine, and evaluate marketing actions; monitor marketing performance; and improve understanding of marketing as a process. Marketing research specifies the information required to address these issues, designs the method for collecting information, manages and implements the data collection process, analyzes the results, and communicates the findings and their implications". Chiu and Tavella (2008) highlight that customers are key components of a market, and, thus, customer research should also be considered as part of marketing research.

Traditionally, marketing research is divided on exploratory, descriptive and causal research by its nature.

- Exploratory research is a very flexible one in terms of data collection. It helps on the first step of problem identification and hypothesis generation. It is some kind of "detective work" (Stevens, 2006, p.27) which should help to answer the question of what is the current situation, what is happening or what happened. Traditionally it is done through exploring secondary information, conducting personal interviews, leading focus groups, etc.

- Descriptive research is used to describe the unit of observation: customers, sales, etc. Different quantitative techniques are used at this stage to test the hypothesis.
  Causal, or prescriptive research seeks for causes of the situation. Researches might seek for correlations and hidden relationships between variables. The design of such research involves a lot of experiments

Marketing research involves the following steps (Bradley, 2013; Chiu & Tavella, 2008; Malhotra, 2013; Mishra, 2008; Uhl & Schoner, 1969; Wedel & Kannan, 2016):

- Identification of problems. This means to identify a business problem and to convert it into hypotheses to be researched.
- Research design is the specification of methods and procedures for acquiring the information needed for solving the problem. It decides sources of information and methods for gathering data. Good research design insures that the information obtained is relevant to the research questions and that it was collected by objective.
- Determining sources of data. Data sources are traditionally divided by primary and secondary. In brief, primary data is the one collected especially for this particular research question. Primary data sources can be internal, like internal company records, and external, which is collected outside of the company – for example, through competitors' websites, publicly available industry reports, social media, etc.
- Sample design and collection of data. Traditionally, marketers sample the entire population. Sampling process is an important step in the research as a sample should be representative in terms of characteristics of the entire population. The findings which will occur on the later stage must be applicable from sample to the entire population.
- Analysis and Interpretation of data. After the data was collected, it should be analyzed and interpreted. Depending on the research question, the nature of research and the data itself, there are different methods of data analysis and interpretation which are going to be described in further chapters of this thesis.
- Research report summarizes findings and conclusions drawn during the research process. They are then translated into major recommendations which will affect decision-making process in the company.
- Recommendation follow up validates the research report throughout time.

World Wide Web and online services give researches a lot of opportunities to collect data about customer behavior. Firms continuously assess customer satisfaction; new digital interfaces require this to be done with short surveys to reduce fatigue and attrition (Wedel & Kannan, 2016). Global digital transformation affects all aspects of marketing research, starting from problem statement and finishing with the analytics report. Many current statistical and econometric models and the estimation methods used in the marketing literature are not designed to handle large volumes of data efficiently (Wedel & Kannan, 2016). Not every program can handle terabytes, petabytes of data. The issue of where and how to collect and store the data is solved nowadays by big data analytical tools and data mining techniques.

Big data analytical tools together with data mining broaden marketing research borders significantly by introducing data reduction, faster algorithms, model simplification, computational solutions. For example, MapReduce algorithm allows to process large data in an efficient way, by dividing it on stacks, running parallel processing, and then joining it altogether again. MapReduce-based clustering, naive Bayes classification, artificial neural networks help solving common marketing tasks in the field of big data more efficiently.

Thus, statistical and econometric techniques, with the help of big data analytical tools and data mining, are now scalable; new techniques, methods and algorithms are introduced. Linoff and Berry (2011) state that traditional marketing research can help on macro-level, while data mining allows to examine the objects of interest on micro-level. All of this brings marketing research to a completely new level of investigation granularity, making marketing research much deeper, detailed, insightful, allowing to extract much more value.

In a logical, smooth way, forced by many factors – existence of data-rich environment, change of marketplace and marketspace, digitalization, change in consumer behavior, appearance of new forms of marketing, technology and methodology development – data mining incorporated into marketing field and is now used together with traditional marketing research techniques.

These are some of the common marketing tasks solved through data mining:

- Defining customer lifecycle
- Improving direct marketing campaigns
- Choosing a communication channel
- Identifying good prospects
- Optimizing campaign profitability
- Churn rate prediction
- Micro-market identification and study
- Marketing spending model optimization
- Clustering and segmentation
- Forecasting sales
- Market basket analysis
- Recommendation system
- Customer retention
- Buyer behavior
- Identifying overall customer mood towards product
- Collecting opinions
- Conducting web-research
- Response modelling,

And many others.

Thus, marketing research techniques and data mining are complementary subjects, a set of tools which help to maintain healthy relationships with existing customers and to attract new ones.

3. DATA MINING ESSENTIALS

Similar to traditional marketing research, there are analytical technologies which best suit for each of four types of analytics:

- Descriptive: reports, OLAP. This is the most popular type of analytics.
- Explanatory and predictive: data mining, statistics. The aim of explanatory analytics is to find the cause why the situation happened, while predictive analytics tries to forecast possible outcomes of the situation.
- Prescriptive: simulations, goal seeking

Quinn (2009) and other authors proposed three levels at which analytics happen: strategic, analytical, and operational. Some researchers also add a tactical level. These analytical levels coincide with the levels of organizational planning. Each level completes its function: the operational one aims to automate and accelerate processes; empower employee decisions; enhance customer partner relations; monitor performance of initiative.

At analytical level researchers try to identify historical trends and forecast future potential, and direct the focus of operational initiatives. Strategic level is the highest one in the system. At this level, researchers monitor performance and communicate strategy; promote continued improvement; direct analytical business intelligence towards potential problems.

There are three known approaches to data mining process:

- CRISP-DM standing for Cross-Industry Standard Practice for Data Mining. It is the most-widely used approach which includes the following steps (De Ville, 2001):
  - Business understanding
  - Data understanding
  - Data preparation
  - Modeling
  - Evaluation
  - Deployment
  - Performance measurement
- Six-Sigma-Based which is heritage of traditional marketing research. Its steps are:
  - Define
  - Measure
  - Analyze
  - Improve
  - Control

- SEMMA approach developed by SAS institute. It is an acronym that stands for Sample, Explore, Modify, Model, and Assess.

According to Linoff and Berry (2011), data mining process consists of the following steps:

- Preparing data for mining. Includes data collection and ETL (extract, transform, load) processes. Takes the most time because rarely data is perfect and ready-to-use.
- Exploratory data analysis. Includes summarization techniques, basic statistics overview, exploratory dashboards creation.
- Binary response modeling (also called binary classification): basic classification referring to descriptive research.
- Classification of discrete values and predictions: refers to descriptive, predictive and prescriptive researches.
- Estimation of numeric values: used for descriptive research and creation of analytical dashboards.
- Finding clusters and associations: through being descriptive by nature, refers to experimental research design
- Applying a model to new data: data validation.

Table 1: Summary of data mining techniques grouped by data mining tasks

Task	Best Fit	Also Consider
Classification and prediction	Decision trees, logistic regression, neural networks	Similarity models, table look-up models, nearest neighbor models, naïve Bayesian models
Estimation	Linear regression, neural networks	Regression trees, nearest neighbor models
Binary response	Logistic regression, decision trees	Similarity models, table look-up models, nearest neighbor models, naïve Bayesian models
Finding clusters and patterns	Clustering algorithms	Association rules

Source: Linoff & Berry (2011).

There are three main styles of data mining: hypothesis testing, directed data mining, undirected data mining. Hypothesis testing comes from the statistics field and refers to exploratory data analysis.

Directed data mining is more objective-oriented: the aim is to build a model that explains and/or predicts target variables. Undirected data mining discovers overall patterns and hidden relationships between variables.

Depending on the goal, task, data itself, different techniques and data mining styles are applied. Linoff and Berry (2011) summarize data mining techniques grouped by data mining tasks in a table represented by Table 1.

3.1. Data mining and statistics

Later in this master thesis there will be a short overview of data mining techniques in context of marketing problems, however, before that, it is important to emphasize the difference between data mining and statistical techniques.

All data mining techniques are based on statistics and the science of probability (Linoff & Berry, 2011). However, there are several differences in the approaches which statisticians and data miners use.

Data which data miners collect explains them what is happening generally rather than explaining everything. Next, data miners assume dependency on the time, and scientific data should bring similar results regardless the time. Finally, data miners use business data which might be incomplete at the time (final customer tenure is greater than current; some customers will churn tomorrow, and some – in 10 years) while scientific data is usually assumed to be complete.

Generally, data miners should be sceptic about the data they use and always to keep in mind that this data comes from business processes. There are two major data mining methodologies: directed and undirected. Directed techniques are good in answering a specific question which is transformed to a target variable, allowing to find the best model – meaning, such model which the best estimates the value of target variable. Undirected techniques don't answer a specific question but rather help to get a general idea about what is happening inside the data. Both types of techniques will be discussed in details later in this master thesis.

3.2. Data mining tasks

All data mining tasks can be generally categorized to classification and prediction, estimation and prediction, and clustering tasks. This section explains each of them.

3.2.1. Binary response modeling, classification of discrete values and predictions

Classification is a general data mining task which includes a range of business problems referring to putting variables into classes, e.g. anytime when an answer to a business question

are several categorical options. Some of such problems are: will a customer churn or not, will customer open an e-mail or not, is this ad good or not; who are generally company's customers; is this customer highly valuable for a company, is he in the risk-of-loss group, or he is not profitable, at all; is e-mail a spam or not; what products are selling best together, etc.

Classification often addresses prediction problems, e.g. predicting in which class a new variable will fall (not only "is this customer profitable or not", but also "will a new customer be profitable or not"), and, thus, these two data mining tasks frequently come together.

Classification techniques can be used as a standalone solution of a classification task itself, or during data exploration and preparation for further analysis. Some sub-tasks when classification is used are: feature selection, dimensionality reduction, anomaly and outlier detection, dealing with missing data.

Only a few algorithms can naturally deal with 'raw' input categorical variables, so pre-processing is needed. One of the most popular methods to 'translate' categorical variables into numeric are creating dummy columns. For example, when a categorical variable contains three levels – 'red', 'blue', and 'yellow', three new columns – 'red', 'blue', and 'yellow' respectively – are created. Then, if a variable has a value 'yellow', then in the respective column it will get a value of 1, and in 'blue' and 'red' columns – of 0.

Another method which might be more appropriate sometimes (it depends on the business problem which should be solved) is to do something similar to reverse classification, e.g. to rethink a categorical variable in numeric terms. For example, instead of categories 'short hair' or 'long hair', a hair length could be introduced.

Depending of the data input type, different techniques are supposed to be "natural fit" for the problem-solving.

Classification is closely related to a concept of similarity. Similarity, in mathematical terms, refers to a distance between two variables and shows how close to each other are variables of interest. The k-based methods are the most widely used, particularly K-Nearest Neighbor (KNN) algorithm. There are different distances functions used for different data types. For continuous variables, Euclidean (the "default" distance, the most widely spread) distance, Manhattan, or Minkowski distances are used. For categorical variables, Hamming distance is used.

The concept of similarity also forms basis for memory-based reasoning (MBR) and collaborative filtering (which is a particular case of MBR) techniques. Collaborative filtering takes into account not only neighbors similarities but also their preferences, and therefore are widely used for providing personalized recommendations, fraud detection, customer

response prediction, free-text classification of responses (for example, complaints from customers). One of the biggest advantages of these algorithms is that records format doesn't matter.

For categorical data classification, logistic regression is the most commonly used technique which classifies a data entry itself or estimates the probability of a new entry to fall into a particular class.

Naïve Bayes is another method which is applicable to categorical data. In its essence, this algorithm is based on the assumption of variables independency from one another, which is often not the case, but treating each variable individually allows algorithm to make quite accurate predictions about "the mostly likely class" rather than estimating the probabilities of many different classes.

Table lookup models which assign the same score to all observations which fall into the same cell of a table also deal well with categorical input variables. One of a classical examples of table lookup models is Recency, Frequency, Monetary (RFM) analysis.

Decision trees are appropriate both for numeric and categorical input variables, making this algorithm used for a big variety of business problems solution. Besides classification, it is successfully used for estimation and prediction tasks. It represents a hierarchical collection of rules (Linoff & Berry, 2011), and by its essence recursively splits data into smaller and smaller sets which continuously become more and more similar to one another.

Decision trees can be classified based on the splitting criteria (Linoff & Berry, 2011) - Gini, Chi-square, pruning, and others – on CART (Classification and Regression Trees), C5.0 (based on "pessimistic pruning"), CHAID (based on Chi-square), and others. Categorical variables are split by forming groups of classes; continuous variables – by dividing their range of values on as homogeneous as possible sub-nodes.

Besides classification and prediction, decision trees are often used for supportive tasks like specific data selection or as part of selection criteria.

Decision trees are easy to understand and implement, however, with increased data complexity, they can misperform. For such cases, there are many algorithms grown from a single decision tree, such as Random Forest. In its essence, it consists of a large number of decision trees which operate as an ensemble (Yiu, 2019). When predictions transparency, e.g. the drivers for the predictions, are less important than ease of implementations, random forest is often used.

Gradient boosting algorithm is also tree-based. It continuously grows big number of decision trees, taking into account even the weakest of them, and is considered to be high-performing.

Such algorithms as XGBoost or GBM are based on gradient boosting and in their essence represent an ensemble of algorithms. Those are very good with complex data, and Catboost is considered to be a good choice for complex and large categorical data.

Another algorithm for classification are Support Vector Machines (SVM) which are a geometric method of separating two classes (Linoff & Berry, 2011). It is a natural choice for classifying variables into two classes, however, if number of classes is bigger, it might be better to use Naïve Bayes algorithm.

Factorization Machines algorithm proposed in 2010 (Rendle, 2010) combines advantages of factorization models and SVMs (Amazon SageMaker, n.d.). It is widely used on sparse data, e.g. data with a lot of missing variables, and it is famous for be good in finding meaningful interactions between variables, making FM one of good choices for sub-tasks when classification is needed, such as feature selection. They are also used for regression-connected types of tasks, and very often – for CRM, clicks and other metrics predictions, and for complex tasks, such as recommender systems building (Bhatt, n.d.).

For numeric variables, technically any of above-mentioned algorithms can be used, however, some of them come as natural choice, like neural networks. Based on human brain neurons, (artificial) neural networks are considered to be a set of generic algorithms for classification (given a labeled dataset), clustering (not labeled dataset), and predictions (Moro, Cortez, & Rita, 2014;)Nicholson, n.d.). The biggest tradeoff of NN is that prediction/classification drivers is very hard to understand, so if drivers of classification are important, it might be better to choose a more transparent algorithm.

Besides neural networks, any other algorithm which is used for numeric variables estimation and prediction, can be used also for their classification.

3.2.2. Estimation and prediction of numeric values

When working with numeric variables, besides basic data preprocessing (dealing with outliers, missing data, etc.) it is important to pay special attention to data distribution. There are two types of techniques: stochastic (mostly - classical statistical models) and algorithmic (Breiman, 2001). Stochastic techniques include logistic and linear regressions, and algorithms based on them, such as regression trees. Models based on stochastic techniques make assumptions about data distribution, and, therefore, when normal distribution is required but data is not pre-processed to fit normal distribution, such models won't perform well. Some of data transformations were discussed earlier in this master thesis.

From another side, algorithmic techniques fit a function to observed data and do not make any assumptions about how data is distributed. Such techniques include clustering and

associations which will be covered later in this chapter; decision trees, random forests, gradient boosted trees, SVMs.

The choice of what technique to use depends on business task to be solved, data complexity, desired output, and other factors (Azzalini et al., 2012).

Estimation and prediction of numeric variables data mining tasks solve such marketing problems as customer LTV, risk of customer churn, CTR prediction, etc. Natural fit for such tasks from data mining prospective are different kinds of regression, neural networks, regression trees (CART), and other algorithms which were already discussed above.

3.2.3. Finding clusters and associations

By its essence, clustering involves grouping of data points (Seif, n.d.; Kaushik, 2016). When dataset doesn't contain labeled data for running classification process, but marketers still want to know how the dataset population can generally be segmented, or to run supportive task like binning variables for dimensionality reduction, clustering task can be completed. An example of typical business problem which can be solved through clustering is clustering direct marketing campaigns by customer response.

Generally, there are many algorithms for clustering. The first alternative to the most popular one – k-means clustering - are Gaussian mixture models (GMM) which perform better than k-means on different (mixed) types of distributions because k-means algorithm doesn't account for variance (Maklin, n.d.). Another alternative is mean-shift clustering, which biggest advantage over k-means is that there is no necessity in manual selection of optimal number of classes.

As clustering is one of the most popular data mining tasks, it got automated in different forms. Automatic cluster detection is data mining technique which is used to learn about structure of complex data, leading to meaningful insights without answering any specific questions but helping to formulate them.

Hierarchical agglomerative clustering, in its turn, can come in two forms: bottom-up or top-down. The title reflects how algorithm treats the data. Bottom-up clustering takes each data point as an individual cluster and then merges pairs of clusters until all data points are merged into a single cluster, while top-down does exactly the opposite. Cluster structure is then represented as a dendrogram. Such clustering method is useful for the data which by its nature has a hierarchical structure. Unlike other clustering algorithms, hierarchical clustering is not sensitive to the choice of distance metric (distance of each data point to the center of the cluster) (Ojeda et al., 2014).

Another popular clustering algorithm is finding associations and affinity groups which are used for market basket analysis, building recommendation systems, market segmentations,

rules generation (if two actions are made in sequence, then the third one will be defined). Sequential patterns detection also helps solving such business problems as seasonality discovering – for example, when customers buy several products together are particular time of the year. Such information can then be used for promotion campaigns set-up (Wang et al., n.d.).

In addition to association rules, link analysis can be conducted. Link analysis is based on graph theory which represents relationships between different objects as edges in a graph (Linoff & Berry, 2011). It is widely used in social media mining to find opinion leaders and influencers, examine who is the most likely to make a 'match' on dating websites, investigate social buzz, etc.

A special section of clustering is text mining (the application of data mining to text data coming from many different sources) (Linoff & Berry, 2011). The goal of text mining techniques is to understand and extract meaningful information from text documents, websites, comments, blogs, etc. Text documents summarization (in some way) and their clustering into similarity groups are common processes during sentiment analysis and opinion mining (describing group of people's attitude towards some subject) which helps to understand crowd's mood. These techniques are useful during marketing planning process stages and marketing strategy elaboration to scrap competitors' comments and prepare SWOT analysis based on weak points people have mentioned in social media, product analysis, new feature elaboration, and other tasks (Ravindran & Garg, 2015).

3.3. Choosing the proper data mining task and technique

The selection of proper algorithm depends on many factors (Ghani & Soares, 2010). Data mining tasks and techniques are limited by data which organization initially possesses. Data might be changed through time, but at a stage of high-level marketing strategy the pool of tools is restricted by the data which is already in MkIS. All directed data mining techniques (regression and regression-based algorithms, decision trees, neural networks) require a dataset containing labeled values for training of target variables, otherwise undirected data mining techniques will be used. Such techniques as clustering or exploratory data analysis help to understand the data but they cannot explain a particular problem on the go (Linoff & Berry, 2011).

Data input type is important as some algorithms require more data preparation then the others, as discussed earlier. The desired output variable is another criterion of choosing the proper algorithm (Zaki & Meira, 2014). For numeric target variables, when the goal is to estimate the value of a continuous variable, linear regression, neural networks, or any other algorithm that produces continuous value, will be appropriate. Regression trees and table lookup models might not be a natural choice as their output is a relatively small number of discrete variables. Memory based reasoning produces bigger range of values, but those will

never be outside of the range of original data. For categorical target variable or binary response, decision trees, logistic regression, neural networks, and other techniques which output is a probability distribution of variables to fall into each class, are handy. Depending on other aspects of the problem, and on the nature of the inputs, other techniques such as similarity models, memory-based reasoning, and naïve Bayesian models may be good choices (Linoff & Berry, 2011).

The next criterion of data mining technique selection depends on business task to be solved as some algorithms require more data preparation than the others. It is a tradeoff between model performance and the effort of data preprocessing requires. For example, neural networks can only process numeric variables of a small range without missing values and are sensitive to outliers, while decision trees require less data preparation but may not be as good as neural networks for particular data mining tasks, and genetic algorithms are generally used rarely is there is any alternative due to extremely high effort on data preparation step (Linoff & Berry, 2011)

The next tradeoff is associated with the ease of explanation of the results. Decision trees do a relatively good job here, however, complex decision trees-based models are already hard to interpret. Regression is quite easy to interpret in terms of results drivers as every small change in the value affects the score, but regression models still don't explain many things about what contributes to a score. Neural networks are essentially inexplicable but they deal great with complex functions. The tradeoff between best scores and best explanations is solved by always referring to which business problem should be solved (Linoff & Berry, 2011).

Generally, technique best choice depends on business needs and data miner's experience (Spendler, 2010). Sometimes it is hard to predict which model will show the best results, so it makes sense to try several of them on solving the same tasks (Spendler, 2010). The next chapters show some practical examples of data mining techniques applied to real life marketing problems.

4. ANALYTICAL FRAMEWORK IN DATA-RICH ENVIRONMENT

The first chapter of this master thesis has provided an overview of modern marketing, marketing intelligence, and data mining essentials. This chapter is meant to serve as a bridge between the first 3 chapters which were theoretical, and the following, more practical chapters.

Marketing starts with the planning process (its steps have been discussed in the first chapter 1 of this master thesis). It includes careful investigation of external and internal factors

affecting company's performance, as well as tools and techniques selection for completing marketing planning process.

Traditionally, marketing planning has been accompanied by marketing research. In the latest years, as discussed in previous chapters, data mining and marketing research have started to be considered complimentary subjects. Not only knowing what data to collect, how to collect potentially useful data, but also knowing how to prepare it and to clean in order to extract meaningful insights from the data, could also be considered a part of marketing planning.

Customer analytics, product analytics, pricing optimization, promotion campaigns prediction, consumer web buzz and customer reviews' monitoring, competitors' blogs and macroeconomic situation analysis – everything is overlapping, different processes happen parallel to each other. That is why it is hard to separated business analytics from marketing analytics and from product (and all other types of) analytics; why marketing research and data mining processes are getting so interrelated. It is difficult to separate one concept from another since everything is interconnected.

Table 2 provides a framework for conducting and managing marketing planning process in data-rich environment. For each stage in marketing planning process it identifies business tasks to be solved with marketing research and data mining, it proposes the type of research which should be, type of data which might be used, and, finally, marketing research and data mining techniques which could be used.

The table summarizes theoretical insights presented in the previous chapters. Its goal is to refine technical knowledge through managerial and marketing perspectives, by adopting structured approach. The table aims to complete the first goal of this master thesis: to identify possible approaches and data mining techniques for each marketing planning process stage. The table also points out the direction in which it would be worthy going at the early stages of the analysis. It works like a short cheat sheet telling with what to start at each of marketing planning process stage.

Since the table is based on many scientific works, as well as on practical cases, it comes handy when deciding upon which techniques to try first at different marketing planning stages. Some examples are going to be presented in the next chapters. As mission statement is a completely managerial process, this stage is skipped in this part of the master thesis. However, every stage of marketing planning process within an organization can contribute to future formulation of mission statement. The majority of other stages are illustrated with practical examples; others are discussed in the context of the real world.

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Marketing planning process stage				
	Mission	**Situation Analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
Tasks to be solved	Mission statement and corporate objectives definition	• Opportunities identification • 5C Analysis (Company, Customers, Competitors, Collaborators, Climate) • SWOT and PEST-EL analysis	• Target audience definition • Measurable goals setting • Budget development	• Product development and analytics • Pricing defined • Distribution channels defined • Promotion development and analytics	Putting plan into action and monitoring the results
Marketing Research stage	Problem identification	• Research design • Determining sources of data • Sample design and data collection • Analysis and interpretation of data	• Problem identification • Research design • Identifying data sources • Sample design and Data Collection • Analysis and interpretation of data • Research report preparation Recommendation follow-up		Recommendation follow-up

(table continues)

(continued)

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Mission	**Situation Analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
Data mining process stage	Business understanding	• Data understanding • Data preparation • Modeling	• Business understanding • Data understanding • Data preparation • Modeling • Evaluation • Deployment • Performance measurement		Performance measurement
Type of research	Exploratory	• Descriptive • Exploratory • Explanatory • Operational and analytical levels of analysis	• Descriptive • Prescriptive • Analytical level of analysis		• Descriptive • Prescriptive • All levels of analytics • Experimental design of research

(table continues)

(continued)

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Mission	**Situation Analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
Data used	Internal	• Internal company records of orders, sales, prices, costs, inventory levels, receivables, and payables • Publicly available industry and companies' reports • User generated content in social media • Competitors' websites, blogs, etc.	Internal company records of orders, sales, prices, costs, inventory levels, receivables, and payables		• Internal company records • Publicly available industry and companies' reports • User generated content in social media • Competitors' websites, blogs, etc.

(table continues)

(continued)

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Mission	**Situation analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
Marketing research and data mining techniques	<u>Data understanding and preparation:</u> • data cleansing • dealing with missing variables • dimensionality reduction • building initial relationship model in case of structured relational data, in case of unstructured data - mostly dealing with text transformations (tokenization, elimination of unnecessary characters, etc.)				• Data monitoring through interactive dashboards and streaming analytics • Models re-consideration
Marketing research and data mining techniques (cont.)	Data gathering: Qualitative marketing research techniques: surveys (online and offline), in-depth interviews, panels, secondary data research	Data gathering: Qualitative marketing research techniques plus data mining techniques: mining internal databases, web and social media scrapping			

(table continues)

(continued)

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Mission	**Situation analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
Marketing research and data mining techniques (cont.)		Descriptive analytics: data summarization and initial visualizations for frequency distribution (boxplots, histograms, scatterplots); basic statistics representation (describing categorical and continuous variables), cumulative distribution normality checks	Target audience definition: classification and clustering (descriptive and predictive), distance-based clustering methods (k-means), Gaussian mixture models (GMM), hierarchical clustering, naïve Bayesian algorithms, decision trees, artificial neural networks, random forest, link analysis, SVM	Product analytics Descriptive analytics techniques 7 (8) Ps: naïve Bayesian algorithms, decision trees, artificial neural networks, random forest, time series analysis, regressions, memory-based reasoning, choice modelling, rule induction, SVM	All of the described in the previous steps methods might be applicable due to dynamic changes in internal and external environment, new releases of product, changed marketing mix, and other factors.

(table continues)

(continued)

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Mission	**Situation analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
Marketing research and data mining techniques (cont.)		Explanatory analytics: exploratory factor analysis, correlation matrixes, PCA, statistical tests (t-tests, Chi, F...), linear and non-linear regressions, ANOVA, association rules investigation	Measurable goals setting (descriptive and predictive): time series analysis, decision trees, descriptive analytics techniques from previous steps Budget developing (descriptive and predictive): CART, memory-based reasoning		At these stage analysts might experiment through undirected data mining, coming back to one of the previous marketing planning process stages at any point. Some additional tools that might be used: streaming analytics, recommendation systems, dashboards, machine learning techniques

(table continues)

(continued)

Table 2: Marketing planning tasks solved through marketing research and data mining techniques

	Mission	**Situation analysis**	**Marketing strategy**	**Marketing mix**	**Implementation and control**
	Data visualization: • Different kinds of graphs and charts • OLAP operational and analytical dashboards				

Source: Own work

5. SITUATION ANALYSIS

Situation analysis is meant to discover situation on the market and to examine marketing opportunities. Mostly descriptive and explanatory analytics are used to complete this step in marketing planning, however, other data mining techniques may also be used, for example, clustering or web scrapping, as it is going to be discussed further.

5.1. Data preparation, transformations, and descriptive analytics

Descriptive analytics, as the name suggests, help to understand current situation in a company and drive initial conclusions about the data. Before the analysis can be conducted, data extraction and transformation steps are carried out. However, for the purpose of this master thesis, these parts won't be tackled in detail. It is assumed that for this part of analysis the data is gathered from internal company's databases. Every company choose the pool of tools used according to its needs. OLAP together with R or Python for additional analysis is a widely-used combination.

The section 5.1 is based on exercises and datasets from Chapman & McDonnell Feit (2015). The tool used is R Studio. A sample dataset represents observations of total sales by week for two products at a chain of 20 stores over 2 years, with price and promotion status. It provides a brief overview of how descriptive analytics through summarization and visualization, and automatic segmentation can help to compose the first impression about the data itself, customers, and other subjects of interest.

Summary functions return basic statistics of the datasets (Figure 3): mean, median, quantiles, maximum and minimum data points, standard deviation, standard error, skew kurtosis. This is a good point to start the first meeting with the data.

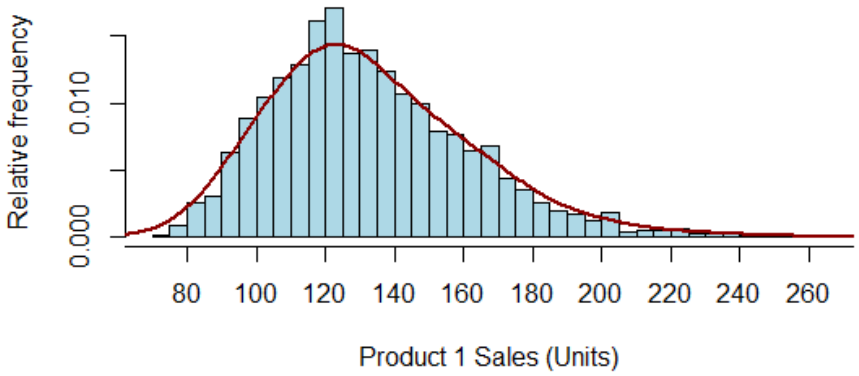
Figure 3: Descriptive statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
storeNum	1	2080	110.5000000	5.7676679	110.50	110.50000000	7.41300	101.00	120.00	19.0	0.0000000	-1.2077396	0.126464485
Year	2	2080	1.5000000	0.5001202	1.50	1.50000000	0.74130	1.00	2.00	1.0	0.0000000	-2.0009613	0.010965862
Week	3	2080	26.5000000	15.0119401	26.50	26.50000000	19.27380	1.00	52.00	51.0	0.0000000	-1.2026174	0.329158561
p1sales	4	2080	133.0485577	28.3725990	129.00	131.08052885	26.68680	73.00	263.00	190.0	0.7393500	0.6565010	0.622110387
p2sales	5	2080	100.1567308	24.4241905	96.00	98.05168269	22.23900	51.00	225.00	174.0	0.9902065	1.5133064	0.535535803
p1price	6	2080	2.5443750	0.2948819	2.49	2.53296875	0.44478	2.19	2.99	0.8	0.2773973	-1.4426746	0.006465714
p2price	7	2080	2.6995192	0.3292181	2.59	2.68939904	0.44478	2.29	3.19	0.9	0.3168365	-1.3978820	0.007218585
p1prom	8	2080	0.1000000	0.3000721	0.00	0.00000000	0.00000	0.00	1.00	1.0	2.6647438	5.1033138	0.006579517
p2prom	9	2080	0.1384615	0.3454668	0.00	0.04807692	0.00000	0.00	1.00	1.0	2.0920368	2.3777619	0.007574861
country*	10	2080	4.5500000	1.7172413	4.50	4.62500000	2.22390	1.00	7.00	6.0	-0.2922095	-0.8077303	0.037653007

Source:. Chapman & McDonnell Feit (2015).

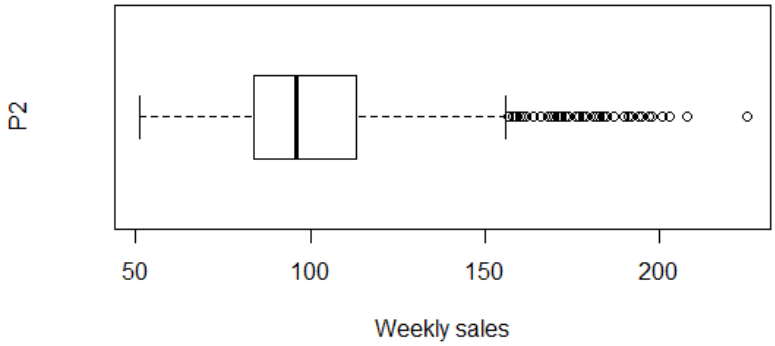
Two basic plots to describe a single continuous variable are histogram and boxplot. They are useful to judge about the normality of distribution. Figures 4 and 5 show a histogram for product 1 sales, and a boxplot for product 2 sales.

Figure 4: A histogram of product one weekly sales frequencies



Source: Chapman & McDonnell Feit (2015).

Figure 5: A boxplot of weekly sales of product two



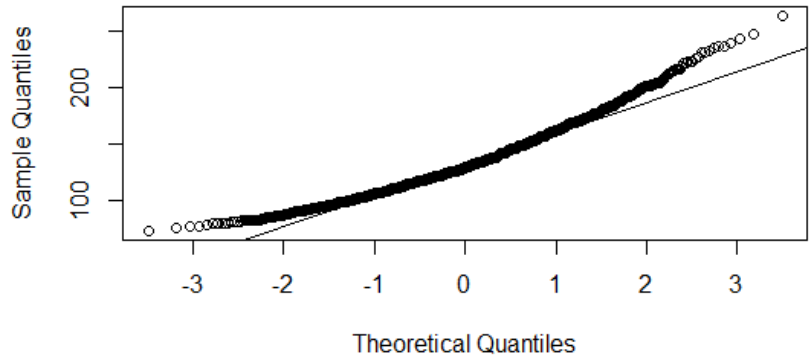
Source: Chapman & McDonnell Feit (2015).

The distribution for both products is right-tailed. Frequency distribution helps to understand the data structure; to prepare the data for further analysis; and from the marketing perspective – to judge about current sales (or other unit of observation) situation. Here, it is visible that the majority of sales for both products occur in the first-second quantile, which for marketers mean that mostly customers order 80-120 product units.

To check if the distribution is still normal, QQ-plot (quantile-quantile plot) might be used. Figure 6 shows that the distribution is far from normal, and it is skewed.

As mentioned, it is important to understand if the distribution is normal because this is going to affect further analysis. Some transformations can help to convert the function in a way that it will fit the normal distribution quantiles. Many of those involve putting the function into a power, but in order not to try all possible values manually, the Box–Cox transformation which generalizes the use of power functions might be applied.

Figure 6: Q-Q plot for Product 1 normality distribution check



Source: Chapman & McDonnell Feit (2015).

A frequently occurring marketing task is to examine correlation between two or more variables. One of the ways to do it involves finding the r value, the Pearson product-moment correlation coefficient. The interpretation of r value is based on the assumption that the variables are normally distributed. The dataset offered in the book simulates a typical customer dataset which is similar to what a simple CRM system output might look like. It represents the data for 1000 customers of a retailer operating both online and offline. The variables describe some characteristics of customers, like credit score, age, distance to store; online visits, spending in store, satisfaction, and others.

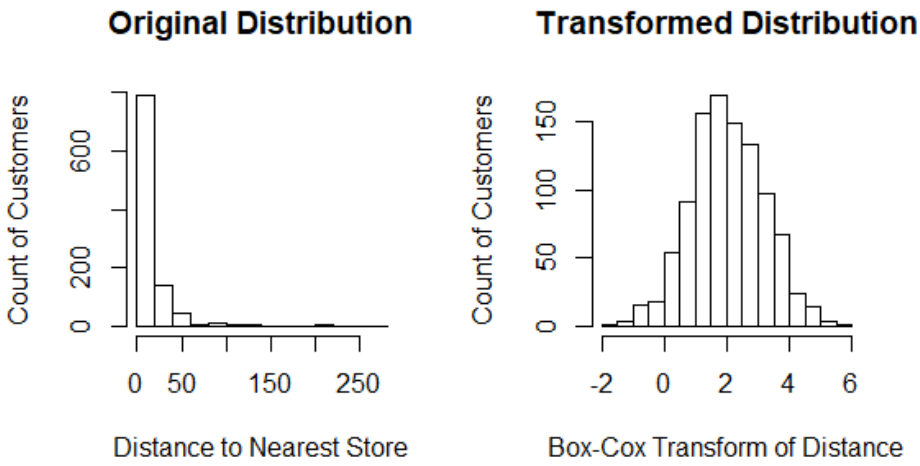
For example, a marketer wants to explore associations between distance to store and in-store transactions. Distance to store is not distributed normally. So, using a function, lambda coefficient equal to -0.003696395 is computed, and the variable is transformed. Figure 7 shows the difference between transformed and untransformed variable.

After the variables are transformed, it is possible to compute the correlation coefficient between distance to store and spending in-store. The correlation is equal to -0.4683126 and is treated as strong negative.

It is important to test whether the correlation coefficient is statistically significant. R output for correlation coefficient significance test is represented below and proves that, with 95% confidence interval, r equal to -0.468 is statistically significant (Table 3).

However, the goodness of fit coefficient which, in its essence, show how much variance of one variable is explained by another one, might be a better way to measure association between variable. R^2 value is equal to 0.22 and shows that only 22% of one variable is explained by another one.

Figure 7: Original distribution vs transformed distribution of distance variable



Source: Chapman & McDonnell Feit (2015).

Table 3: Pearson's product-moment correlation

t = -16.732, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: (-0.5150896; -0.4181827)
sample estimates: correlation = -0.4680421

Source: Chapman & McDonnell Feit (2015).

Another useful information that marketers usually need is at what data point do the 90% of sales happen. Cumulative distribution is what is often used in this case. Figure 8 shows that 90% of data is equal or lower than 171 units.

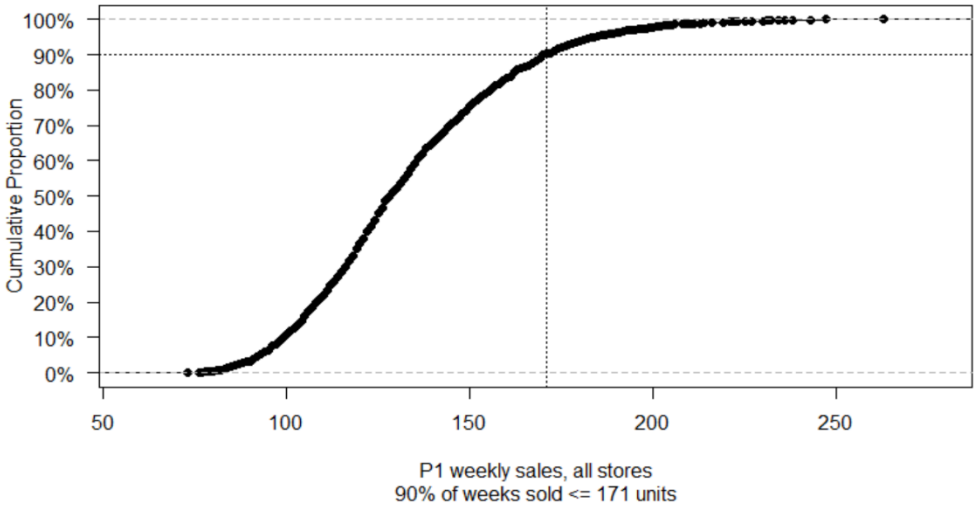
For international companies, visualizing data per-country might be a useful step to compare sales and other KPIs. One of the commonly-used technique for this is to plot the data on a map. Figure 9 represents sales of product one per country.

After company exploration, initial customer exploration and segmentation is conducted. This part of analysis may include basic segmentation based on demographic characteristics: age, geolocation, job title, and on other parameters which dataset might include. Sometimes segments are pre-defined. In this case descriptive analytics would help to draw initial conclusions about the audience. For example, figure 10 shows average income per group of customers refined by ownership of home and subscription.

Another dataset (Kaggle, n.d.-b) contains data about customers and their attributes, such as gender, if customer has a partner or not, if customer has dependents or not, tenure (number of months customer has stayed with the company), information about monthly charges and total charges, whether customer has churned or not, and several others.

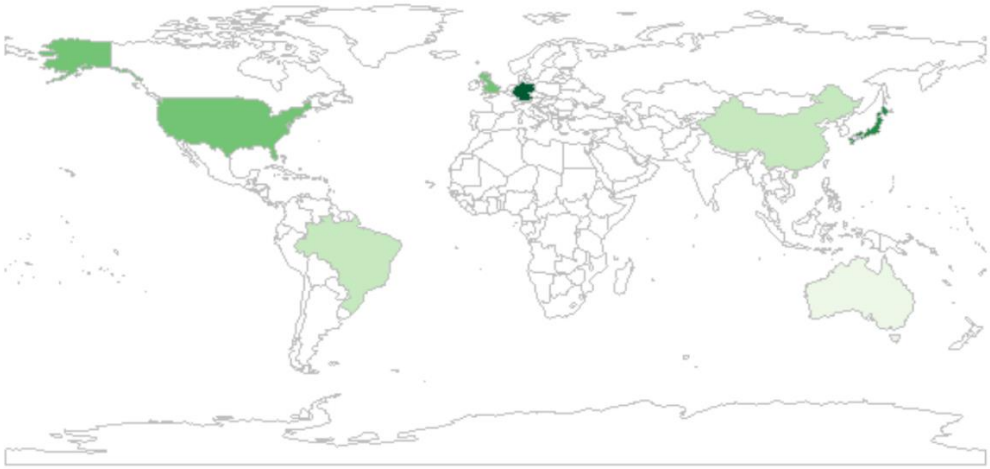
There is approximately equal number of female and male customers, and the bar plot represented on Figure 11 shows the distribution of customers by their payment method.

Figure 8: Product 1 weekly sales cumulative distribution



Source: Chapman & McDonnell Feit (2015).

Figure 9: Product 1 total sales by country



Source: Chapman & McDonnell Feit (2015).

It might be interesting to explore the average churn rate in correlation with customer gender. For this a rank of 1 is assigned if a customer didn't churn, and 0 in case he did. Table 4 represents the results which can be interpreted as approximately 73% of customers didn't churn regardless of their gender.

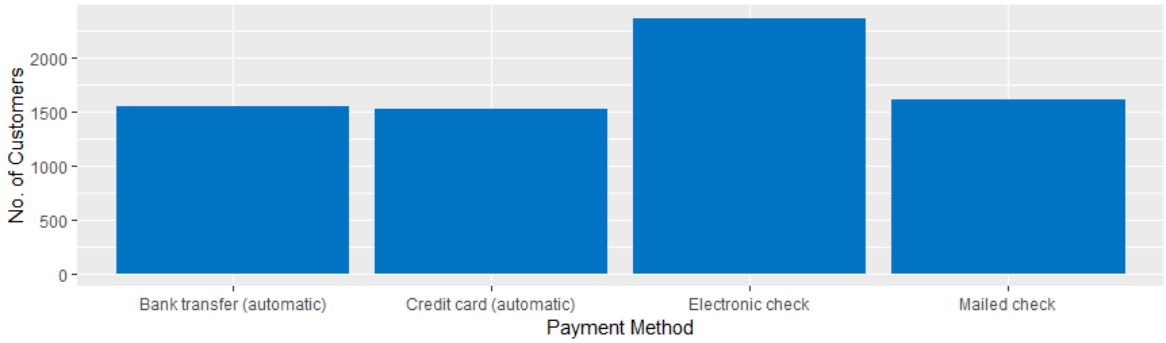
To explore whether churn rate depends on gender, the variable is transformed into a dummy one (male = 1, female = 0), and a simple linear regression analysis is conducted. Its results are presented in Figure 12 from which it is visible that alpha is very high and equal to 0.47 which indicates that this model doesn't have statistical significance and there is no correlation between gender and churn rate.

Figure 10: Average income per customer group

	Segment	ownHome	subscribe	income
1	Moving up	ownNo	subNo	55402.89
2	Suburb mix	ownNo	subNo	54579.99
3	Travelers	ownNo	subNo	65852.54
4	Urban hip	ownNo	subNo	21604.16
5	Moving up	ownYes	subNo	49898.85
6	Suburb mix	ownYes	subNo	55354.86
7	Travelers	ownYes	subNo	61749.71
8	Urban hip	ownYes	subNo	23993.93
9	Moving up	ownNo	subYes	50675.70
10	Suburb mix	ownNo	subYes	63753.97
11	Travelers	ownNo	subYes	48091.75
12	Urban hip	ownNo	subYes	20271.33
13	Moving up	ownYes	subYes	51359.44
14	Suburb mix	ownYes	subYes	52815.13
15	Travelers	ownYes	subYes	62944.64
16	Urban hip	ownYes	subYes	19320.64

Source: Chapman & McDonnell Feit (2015).

Figure 11: Bar chart of a count of customers segmented by payment method



Source: Own work

It would be interesting to explore whether there is correlation between monthly charges and churn rank gives interesting results.

There is a question whether linear regression can be used for this variable, at all. For this it is necessary to check its distribution for normality. As it is hard to conclude if frequency distribution is normal or not, a QQ-plot was drawn. Basing on it, it is possible to conclude that, most probably, the distribution is not normal. To dive deeper into the issue, Shapiro-Wilk normality test is run, and p-value close to zero indicates that it is not possible accept the null hypothesis which states that values are distributed normally, and the alternative hypothesis stating that the distribution is far from normal should be accepted.

Then, a function which applies all possible transformations to transform the distribution into closely normal one, is run, and, as it suggested, Ordered Quantile normalizing transformation is performed, and transformed variables are written as a new column in the data frame.

Table 4: Churn rate by gender

Gender	Churn rate
Female	0.7307913
Male	0.7383966

Source: Own work

Running Shapiro-Wilk normality test again gives an output of p-value equal to 0.9983 which, from one hand, is obviously bigger than 5%, meaning that it is not possible to reject the null hypothesis stating that the values are distributed normally and that it should be accepted, and, from another hand, indicating a possible overfitting problem (which is a common problem for training datasets). After transforming Monthly Charges variable, its distribution seems to be normal.

Now, it is possible to run linear regression with Transformed Monthly Charges as independent variable. Its p-value is very close to zero showing that the association between two variables is statistically significant. However, the R-squared value is very low: 0.0268, indicating that less than 3% of one variable variance can be explained by another variable variance.

In order to explore other correlations, correlation plot is drawn, from which it is possible to conclude that none of the correlations can be interpreted as medium or strong. It is important to note that only relationships between numeric variables were taken into account (monthly charges, tenure, churn rate variables).

One of common marketing tasks is to analyze survey responses. For this analysis, sales survey responses were mocked-up (meaning that the data was simulating a real-life situation while it wasn't really collected but rather it was created using different techniques).

The data contains the following information: if the purchase was made during the weekend; number of days waiting for delivery; overall satisfaction with the experience with the brand; total accessories purchased; average product rating per customer; craft package points; overall satisfaction through brand lines. Figures 12 and 13 show correlation between variables.

Several satisfaction items are moderately to strongly associated with one another ($r>0.4$), but at the same time none of the items can be considered to be identical or almost identical (none of r is bigger than 0.8). Such results show that the idea to proceed with exploring relationships between these variables is reasonable. The results are showed on Figure 13. The variables used are: number of days of delivery, units purchased, total accessories purchased, product rating, craft package, overall satisfaction, overall experience with the brand.

Figure 12: Correlations of sales survey data variables

	no_of_days_delivery	overall_experience	total_accessories_purchased	product_rating_100	craft_package	overall_satisfaction
no_of_days_delivery	1.000000000	-0.04026024	0.004658171	-0.02097292	-0.01345167	0.3194804
overall_experience	-0.040260243	1.00000000	0.45518511	0.31419951	0.78956505	0.5859863
total_accessories_purchased	0.004658171	0.45518511	1.00000000	0.29910498	0.51697987	0.4374679
product_rating_100	-0.020972921	0.31419951	0.29910498	1.00000000	0.36788467	0.5726217
craft_package	-0.013451671	0.78956505	0.51697987	0.36788467	1.00000000	0.6393982
overall_satisfaction	0.319480357	0.58598628	0.43746787	0.57262166	0.63939818	1.0000000

Source: Own work

Figure 13: Correlation plot for sales survey data

Source: Own work

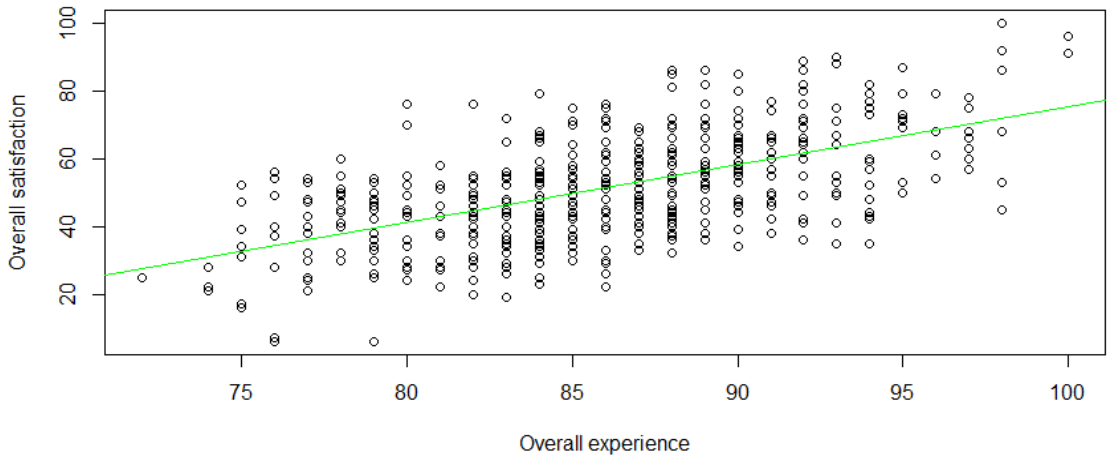
Exploring further relationships between overall experience and overall satisfaction, scatterplot of linear model output (dependency of overall satisfaction on overall experience) is created (Figure 14).

Next step is to examine how strong is the relationship between overall customers' satisfaction and their overall experience (Figure 15). Coefficient of 1.70 shows that with every additional rating point for overall experience, the overall satisfaction increases by 1.7. Mathematically it means that linear model line crosses y-axes at this point. P-value (Pr>|t|) refers to Wald test which checks if the coefficient is significantly different than zero. By default, the confidence interval is equal to 95%, so with 95% of confidence it is possible to make it sure that the coefficient lies between ±1.96 * Standard Error. For this model, it means

that 1.7033±1.96×0.1055 = (1.495,1.910): the coefficient (intercept with y-axis) for overall experience is 1.495-1.910.

Residuals in the output show how close are the values to the best fit (linear model) line. Generally, residuals are the difference between the predicted and the actual value. R-squared is also called a goodness of fit coefficient. In case of only one predictor it is equal to squared r coefficient. For this model, it is possible to say that 34% of variation in overall satisfaction is explained by variation in overall experience. That indicates that the relationship between 2 variables is not that strong. F-statistic with p-value <0.05 allows to reject the null hypothesis stating that a model without predictors performs as well as the composed model.

Figure 14: Scatterplot of linear model visualization showing correlation between overall satisfaction and overall experience



Source: Own work

Figure 15: Linear model summary for relationships between overall satisfaction and overall experience investigation

```
Call:
lm(formula = sales_df$overall_satisfaction ~ sales_df$overall_experience)

Residuals:
    Min      1Q  Median      3Q     Max
-33.597 -10.048   0.425   8.694  34.699

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -94.9622     9.0790  -10.46   <2e-16 ***
sales_df$overall_experience   1.7033     0.1055   16.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.88 on 498 degrees of freedom
Multiple R-squared:  0.3434,    Adjusted R-squared:  0.3421
F-statistic: 260.4 on 1 and 498 DF,  p-value: < 2.2e-16
```

Source: Own work

5.2. Exploratory analysis

Business intelligence tools like Power BI or Tableau allow to conduct descriptive and explanatory analytics in an easy and fast way.

5.2.1. Automatic segmentation

To finalize descriptive and explanatory analytics section, it is appropriate to highlight the importance of these early analytical stages as they can indicate what to pay attention on, where relationships between variables might become an object of interest, which variables are likely to be correlated, and to specify the direction of further analysis.

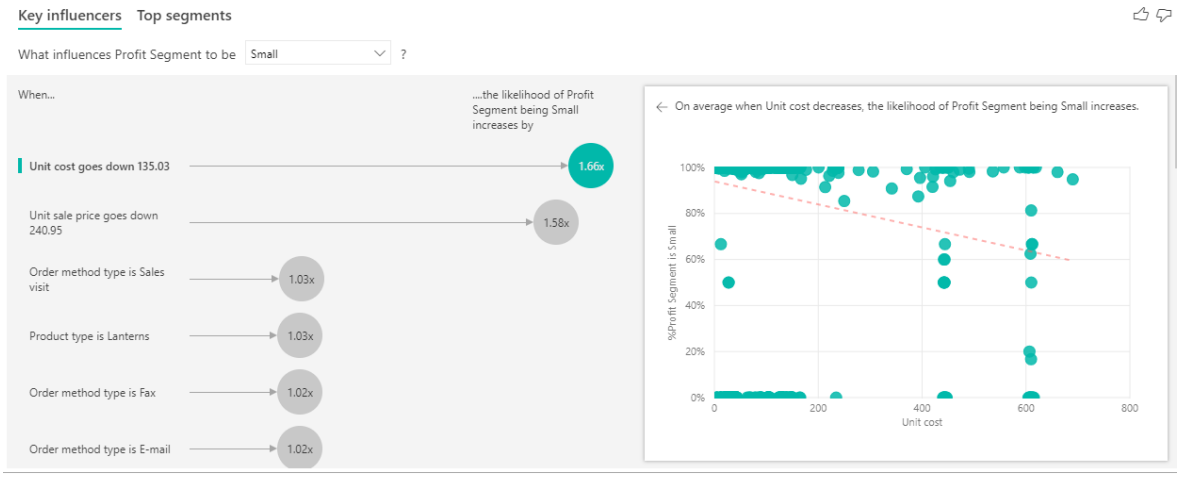
For example, sample sales data provided by IBM (2017) includes records about product type, product line, product, and their parameters: gross profit, order method type, planned revenue, quantity, retail country, and others. A company knows that when gross profit is smaller than 500000$, it can be considered as small; between 500000 and 1500000 – as medium, and everything else is a high profit. The aim here is to examine the drivers, e.g. what influences a product type to fall into one of these categories. Using linear and logistic regression analyzes, Figures 16-22 were created.

As it is visible from the Figure 16, surprisingly, lower the unit cost is, higher is the probability that the profit will be low, too. However, frequency distribution shown on the scatter plot doesn't seem to be normal, so without additional transformations described earlier it is hard to say if this proposed key influencer is really an influencer. Despite, it might be useful on initial stages of the analysis to dive deeper into automatically generated (using decision tree) segments.

Figures 17-22 represent key characteristics of each segment of product types which bring low revenue. The graphs show outliers which affect the distribution the most – those might become units of analyzes in the future. In this case, United States (retailer country) and Tents (product type) mostly affect the distribution.

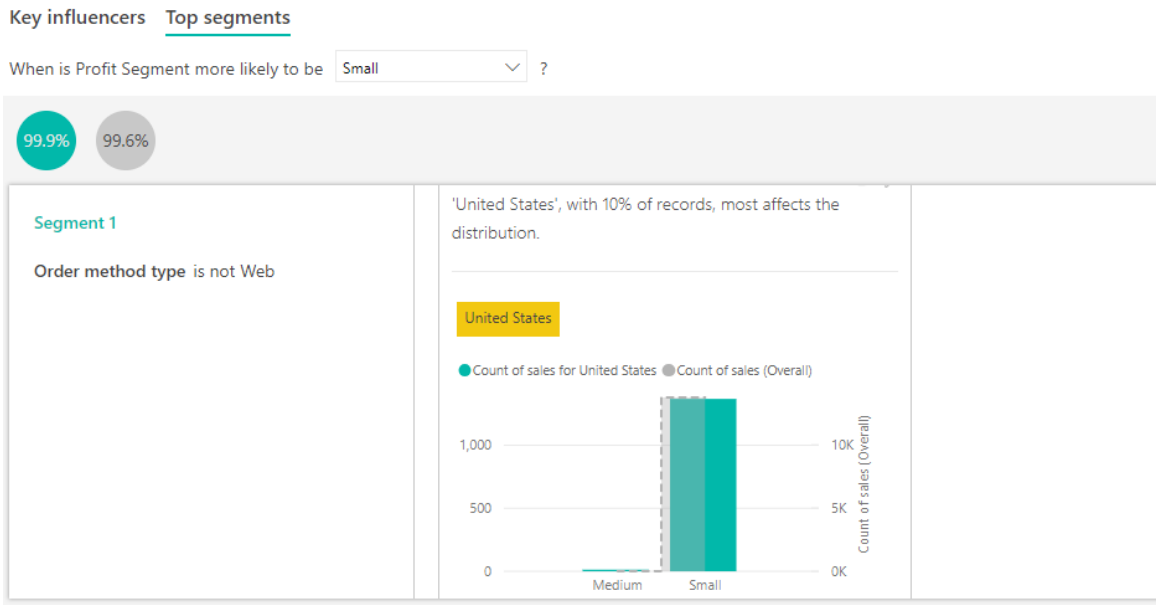
In order to examine how segments are different from one another, in further analysis it would be useful to use ANOVA. In the initial stages, to get the general impression about the data, it might make sense to check other profit segments at first.

Figure 16: Key influencers for product type to fall into 'Small' profit segment



Source: Own work

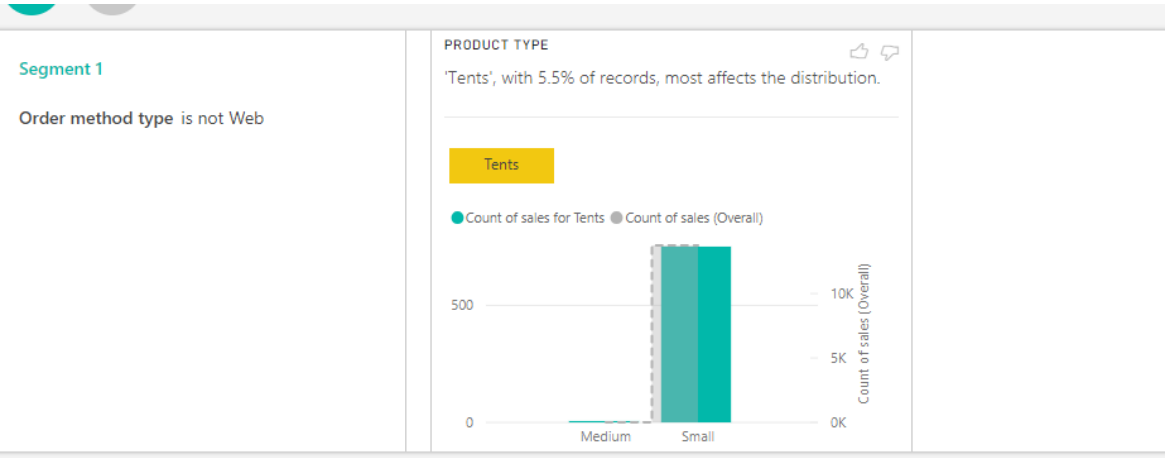
Figure 17: Segment 1 for low-profit product type – Country variable



Source: Own work

Automatic segmentation brings something interesting this time. In difference of Small profit category, there is more clear segmentation by product type, order method type, and unit sales price. When profit segment is likely to be medium, it has higher unit cost, retailer country is likely to be China or United States, and Product Type will most likely be Woods, Eyewear, or Watches.

Figure 18: Segment 1 for low-profit product type – Product type variable



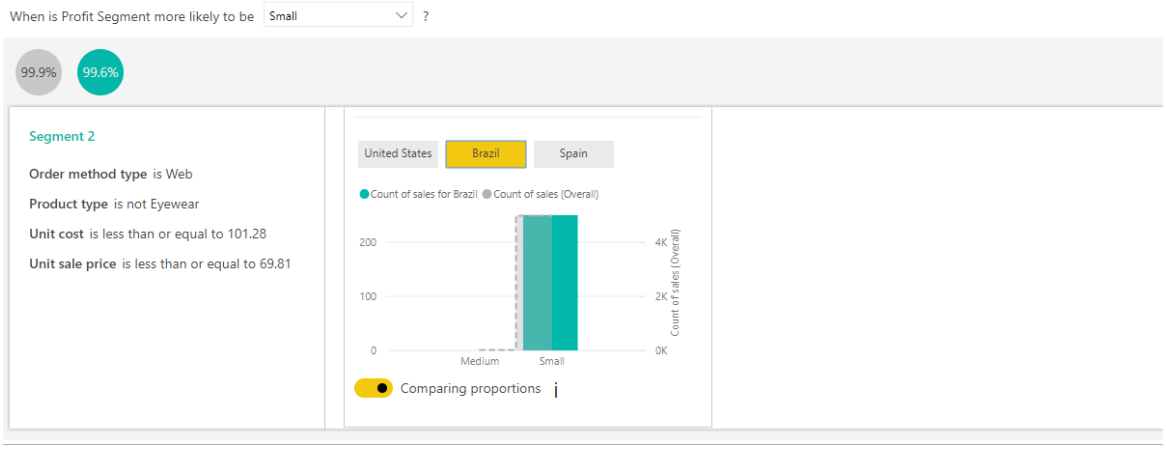
Source: Own work

5.2.2. Automatic cluster detection

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters (Kaushik, 2016).

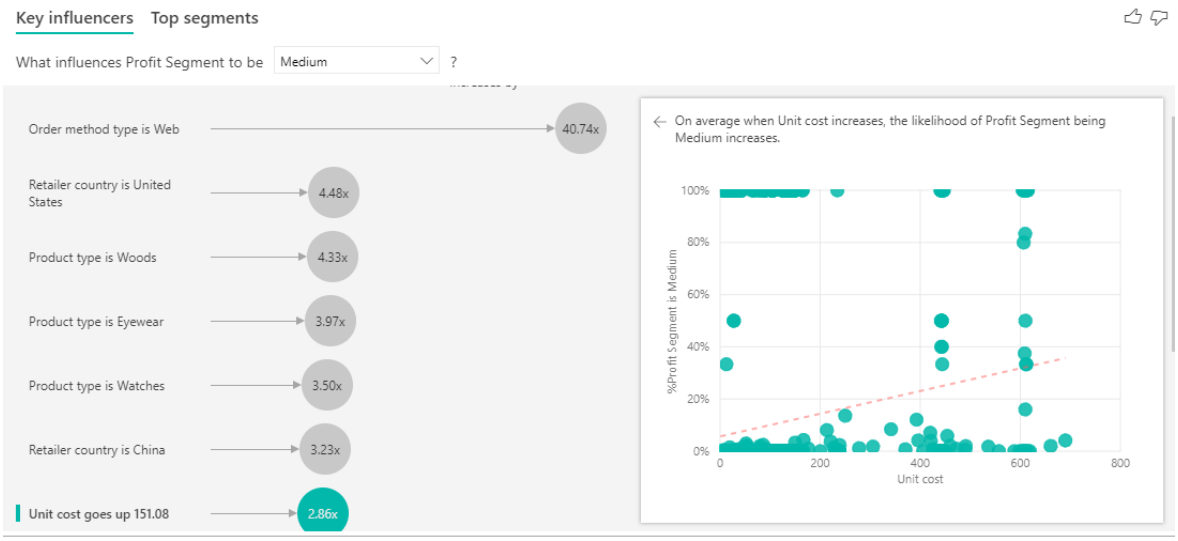
The clusters are composed of records similar to each other. Those similarities are discovered by algorithms. In marketing, clusters formed for a business purpose are usually called "segments," and customer segmentation is a popular application of clustering (Linoff & Berry, 2011).

Figure 19: Segment 2 for low-profit product type



Source: Own work

Figure 20: Key drivers for profit segment to fall into Medium category

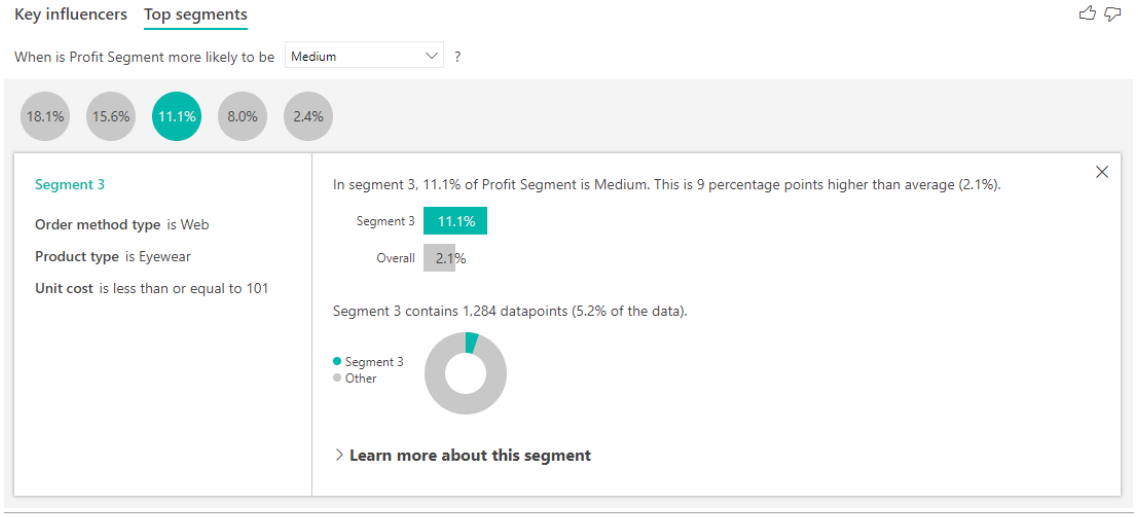


Source: Own work

When talking about data mining, automatic clusters detection is one of the most popular tasks. It is rarely used as a standalone technique because once clusters have been detected, other methods must be applied in order to figure out what do clusters actually mean.

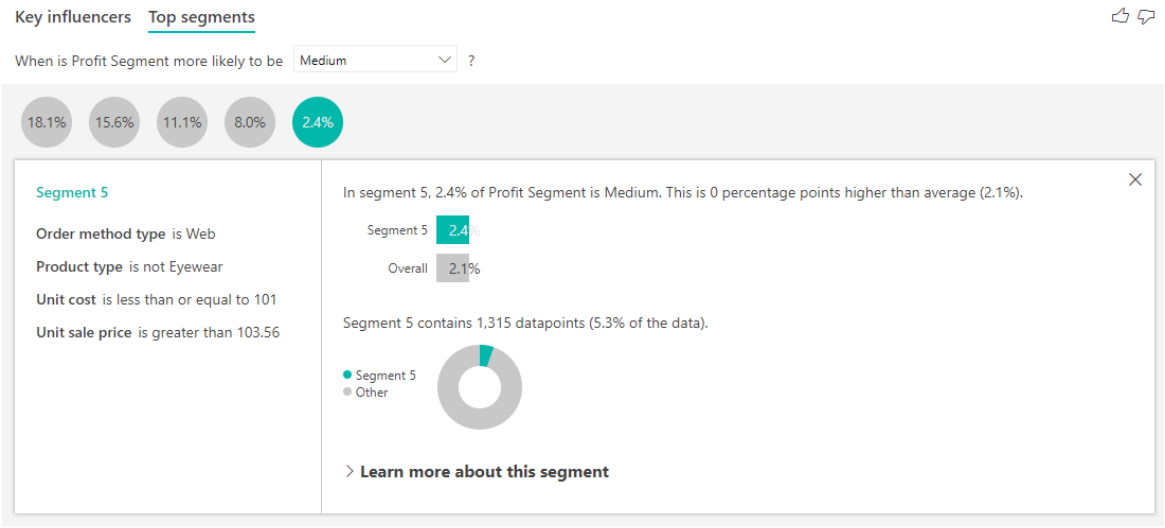
One of the most popular techniques of clustering is K-means algorithm which refers to centroid models, in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. The "K" in algorithm title refers to the fact that the algorithm looks for a fixed number of clusters which are defined in terms of proximity of data points to each other. There are several techniques to choose the 'best' clustering algorithm, including high-level functions packages for this (in R: Sekula, Datta, & Datta, 2017). When using k-means algorithm, there are also number of packages to help determining the optimal number of clusters (Kassambara, n.d.).

Figure 21: Segment 3 for medium-profit product types



Source: Own work

Figure 22: Segment 5 for medium-profit product types



Source: Own work

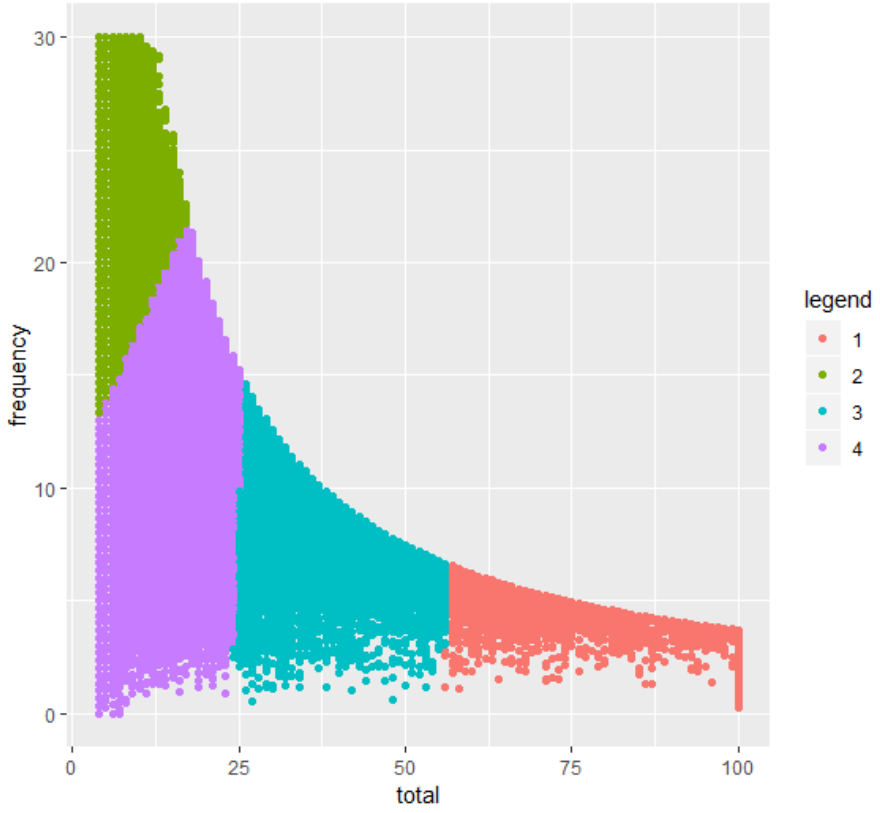
The data for the next case associated with clustering was taken from Kaggle competition (Kaggle, 2017), and clustering algorithms' code used is partially taken from Kaggle's competitor's repository (Dourado, 2017). In this competition, Instacart is challenging the Kaggle community to use this anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order. Several datasets are presented, each containing data about different aspects of an order: product description, product category, previous orders items, current order items, number of times an item was reordered, days since prior sale, and many others. Before the analysis, the data is prepared to be used: cleaned, joined, and re-formatted.

Using the concept of RFM – recency, frequency, monetary value of a customer (Hosseini et al., 2010) – the recency vs frequency became metrics of interest in clustering, literally answering question 'how often do people rebuy'. K-means clustering is used, number of clusters is 4. Clusters are formed (Figure 23): Buy twice a week (1), Buy almost monthly (2), Buy almost weekly (3), Buy almost every two weeks (4).

There are several techniques to choose optimal number of clusters in k-mean clustering. The most frequently used are Elbow method, Average silhouette method, and Gap statistic method (Kassambara, n.d.).

The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and it is wanted to be as small as possible. The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

Figure 23: Automatic clusters formed in R – How often do people rebuy



Source: Dourado (2017).

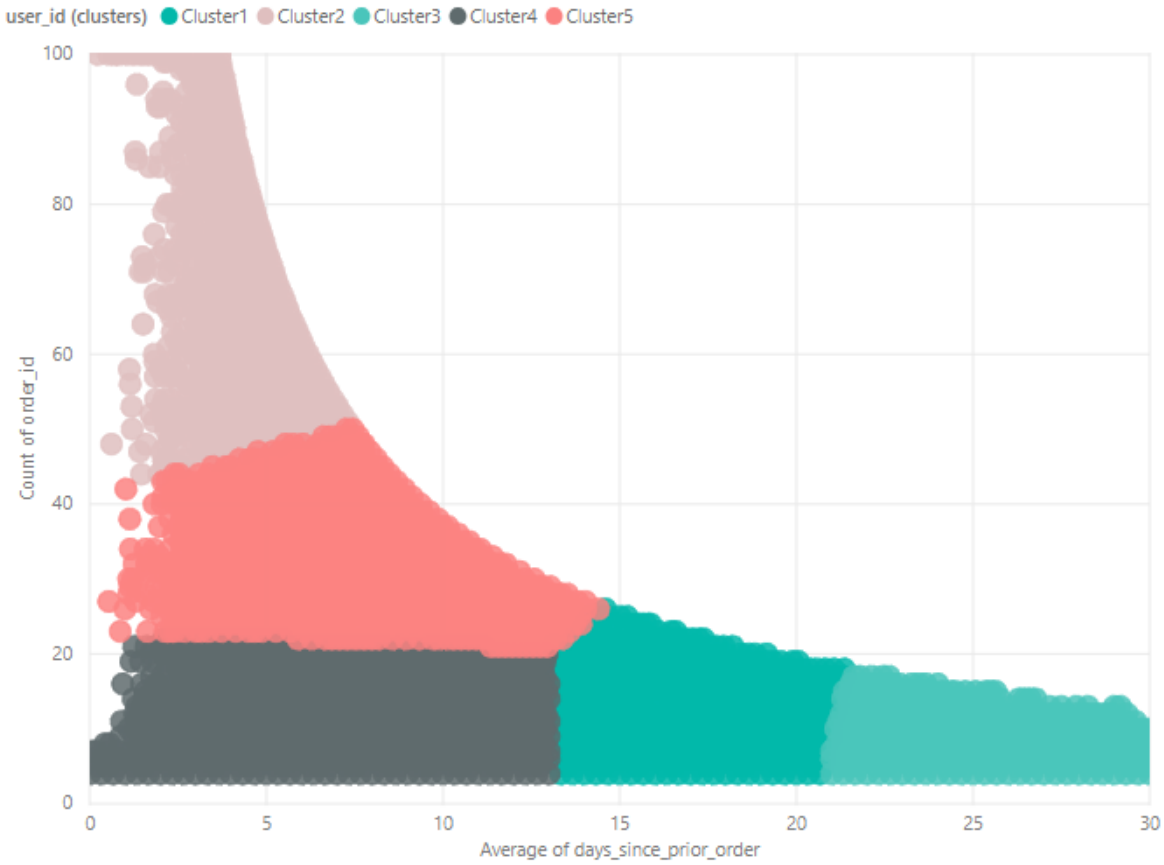
The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points (Kassambara, n.d.).

The gap statistic and average silhouette methods gave the optimal number of clusters equal to 2, while the result of Elbow method was somewhere between 3 and 4.

It was interesting to try automatic cluster detection without specifying number of clusters or parameters, at all. It was done (Figure 24) in Microsoft Power BI (Microsoft Power BI, 2016). As it is visible from the picture, Power BI determined 5 clusters similar to those determined earlier in R Studio. However, they are not completely the same: the majority of people tend to buy within 13 days since prior order, and Power BI has detected 3 sub-clusters based on count of orders within one cluster, which was split by two in R (‘people who buy almost weekly’ and ‘people who buy twice a week’).

Figure 24: Automatic clusters formed in PowerBI – How often do people rebuy



Source: Own work

Such automatic clustering is useful to find overall trends in data, however, they need further investigation from business perspective – using, perhaps, segmentation tools like in one of examples described earlier in this master thesis.

5.3. Web and social media data analysis

Social media and web scrapping and mining, in simple terms, mean systematic collection and analysis of information generated from social media and / or web (Danneman, 2014; Russell, 2011; Technopedia, n.d.). Data mining techniques associated with this complex topic come handy for marketers to complete the following tasks:

- Customer sentiment measurement. Opinion mining (sentiment analysis) is a fast way to get an impression of what do customers feel about a product/service to measure customer loyalty and satisfaction

- Online branding monitoring: brand mentions, customer services, competitors' buzz, customers' touch points where customers engage with the brand (Batrinca & Treleaven, 2015)

- Identifying market trends through tracking industry influencers and comparing them over the time

- Portraying target audience through analyzing customers' profiles in social media

- Market segmentation. Social media mining may help to define market segments in terms of the most common geographical locations, web behavior, job titles, and other social-demographic characteristics.

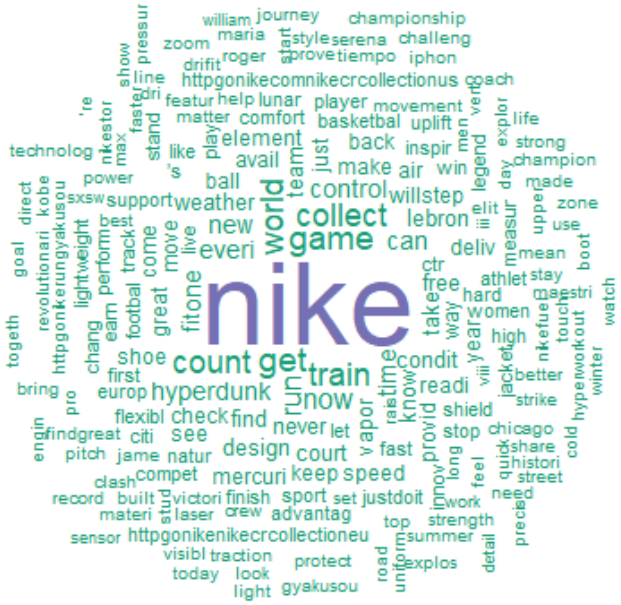
- Defining competitors' average customer satisfaction rate through scrapping those from specialized websites.

Conducting such analysis is a necessary part of preparing a marketing plan. Data mining techniques make this part faster and easier. There is number of limitations associated with social media scrapping which will be discussed later in this master thesis. However, using a combination of tools and data mining techniques, such kind of research can still be conducted.

5.3.1. Scrapping Facebook pages

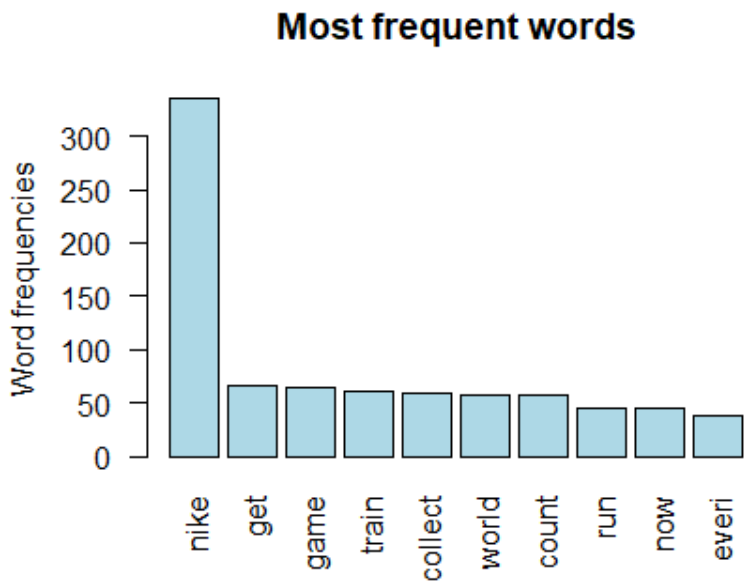
To get an idea of what competitors or/and collaborators write about on their Facebook pages, two pages were analyzed: Russia Beyond the Headlines, and Nike. For RBTH, 10 posts for the dates of 1st and 2nd of November, 2017, were collected; for Nike, 634 posts for several months in 2019, got into a .csv file.

Figure 25: Word cloud for Nike page



Source: Own work

Figure 26: Nike's most frequent words

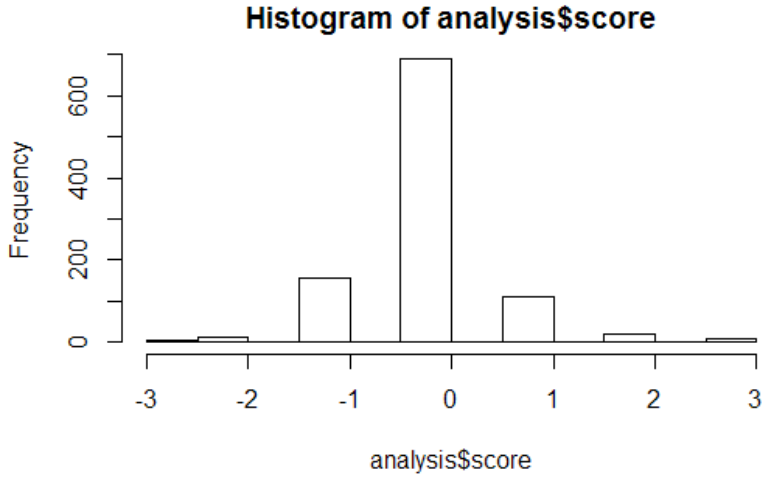


Source: Own work

Then, word clouds were composed out of the posts scrapped from two pages. The Nike word cloud is presented in the figure 25. Using a similar algorithm, posts from other social media like LinkedIn could be collected.

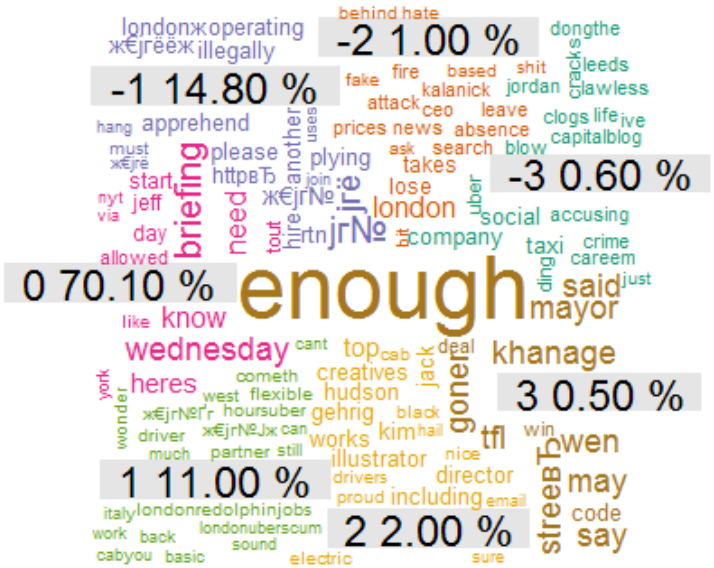
Next example is related to sentiment analysis and defining customers' emotional attitude towards a brand. A thousand tweets about Uber (#uber) in London were collected in June, 2017. Then semantic dictionaries for negative and positive categories of words.

Figure 27: Histogram of sentiment analysis scores distribution



Source: Own work

Figure 28: Word cloud of Uber London sentiment analysis



Source: Own work

Then the data was cleaned and tokenized, after that it was compared to semantic dictionaries, and semantics were categorized. After this procedure was done, scores were assigned to each line of text (negative, neural, or positive) and expressed in percentage. The results were drawn in a histogram and pictured as a word cloud. The results are presented in Figures 27 and 28.

As the figures suggest, the majority of people feel neutral or slightly negative towards Uber in London. Many tweets contained words ‘long’, ‘waiting’, ‘expensive’, as well as ‘illegal’. This might be a good insight for a company to reconsider its politics, both business and advertorial.

Sentiment analysis might be helpful to collect people's opinions about brand or service without organizing costly surveys or questionnaires, to eliminate social desirability bias associated with surveys; to build a frame of what people really like or do not like about something, excluding external influence factors, and to improve service quality and managing negative (reputation control).

5.3.2. Scrapping and analyzing Amazon reviews

Web scrapping allows to get an initial idea about people’s attitude towards the product or service. For example, small drones for children are a trend toy for the last 3 years. There are more than 600 reviews on Amazon for Force1 Mini Drones for Kids – UFO 3000 (Amazon.com, n.d.). When creating SWOT analysis, it might be useful to know what customers buzz around. Scraping around 400 customer reviews resulted in another word cloud (Figure 29). The scraping was done using materials from Saito (2019b).

Figure 29: Word cloud of Amazon reviews

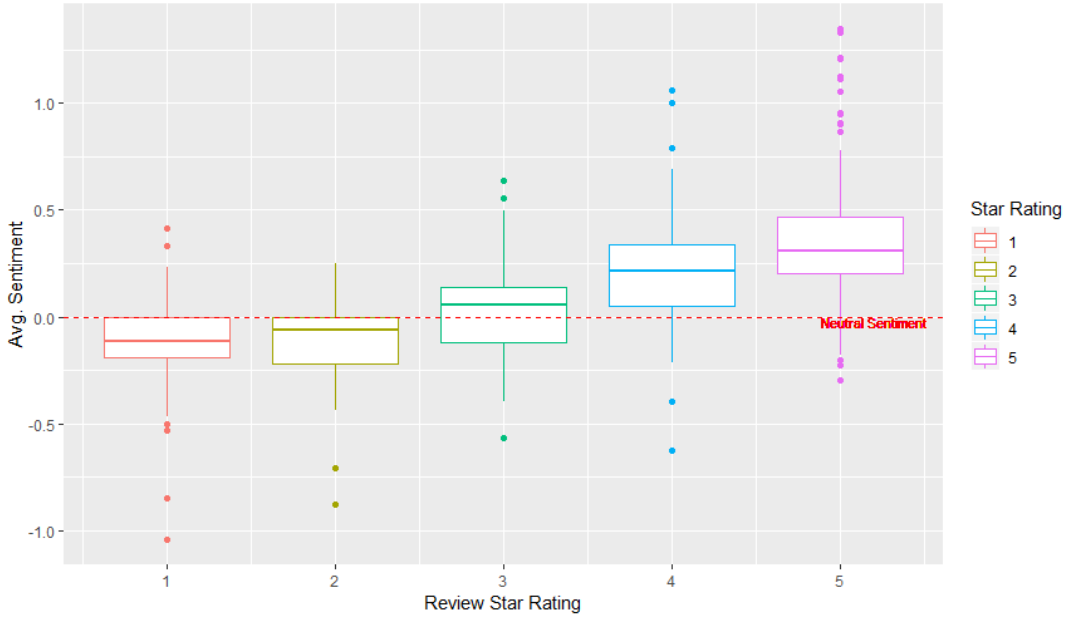
Source: Own work

Word cloud gives an initial idea of what customers talk about but still doesn't represent their attitude well enough. Figure 30 represents customers' sentiment over star rating. Star rating was scrapped from Amazon reviews. Using statistical functions, average star rating was extracted.

Average sentiment was found by comparing every word to sentiment vocabulary, where each word has a rating assigned, based on Saito (2019a). To exclude bias, Drones, Helicopter, Drone and several other synonymous were re-assigned a rating equal to 0. Then every word acquired its rating, and average rating per every review star rating was computed. As a result, a graph with comparison of average sentiment over the star rating was composed.

It might be useful for marketing planning to connect PESTEL and SWOT analyses to actual customers' attitude. For example, the boom of AirBnb put hotels in a situation of higher competition, which could lead to higher amount of neutral and negative reviews written by hotels guests.

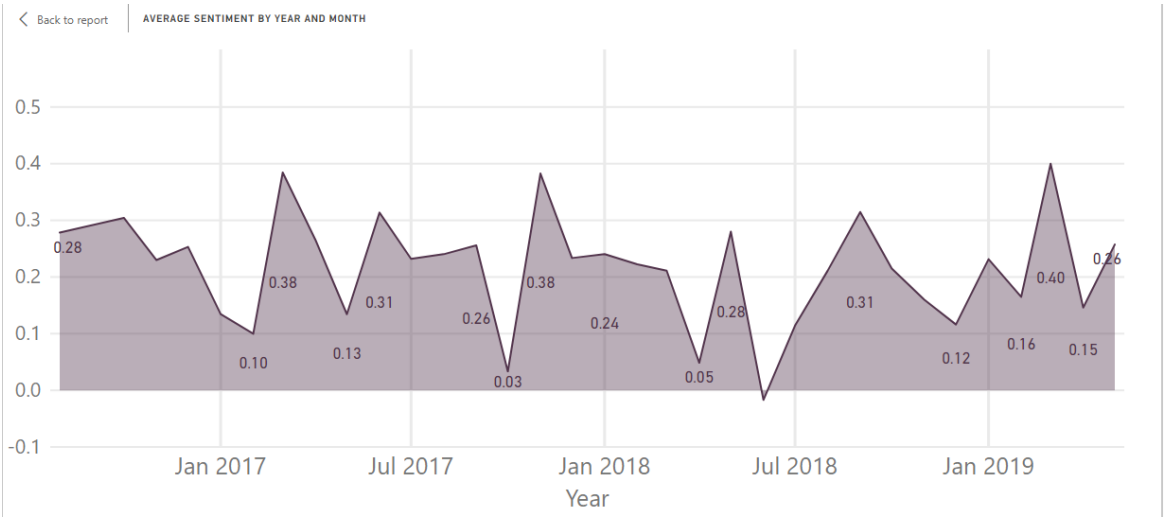
Figure 30: Average sentiment of Force 1 Mini Drones Amazon Reviews over star rating



Source: Own work

Another reason for poor reviews could come from the inside of a hotel and could be related to a high amount of maintenance issues reported by guests. Conducting time series analysis, e.g. to compare average sentiment over time, could help discovering such relationships. In Amazon reviews case, it is visible from Figure 31 that there might be a trend for seasonality, probably caused by national holidays, and this could be a good starting point for further analysis.

Figure 31: Average sentiment by year and month



Source: Own work

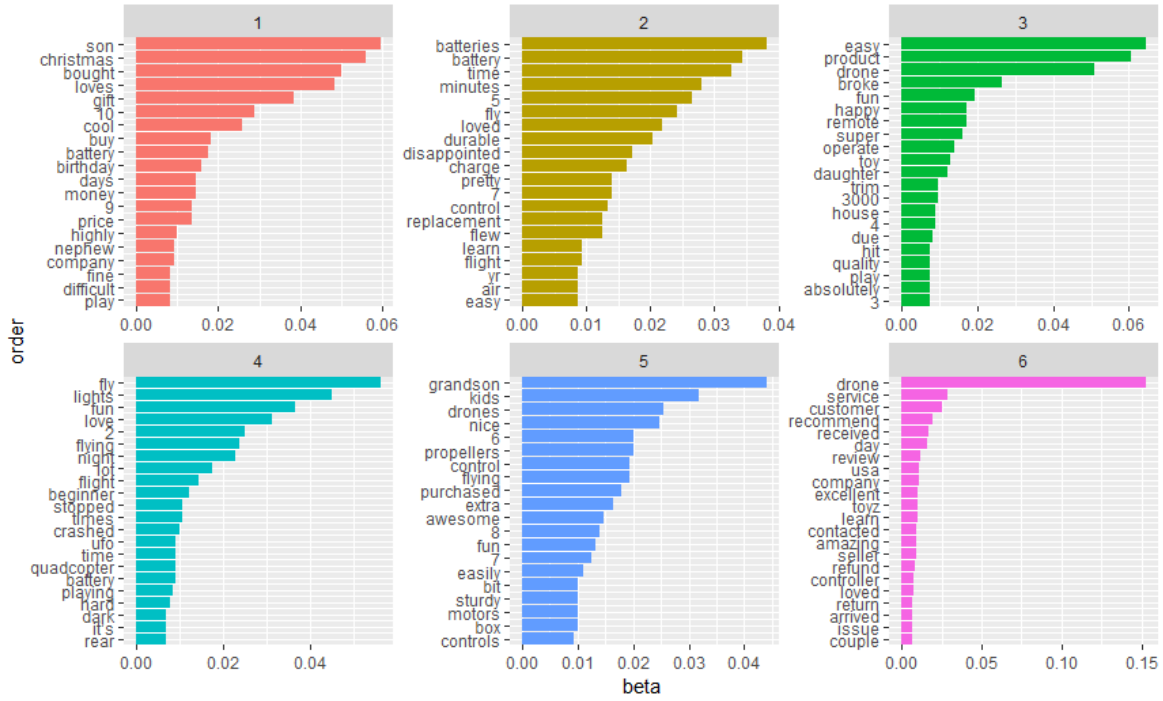
The next thing to cover about web scrapping is topic modelling. Topics are modelled across the term document matrix formed out of all words that appear in more than 1% of reviews, and then selected top 20 words from the rest of words based on their 'beta' value which

roughly represents how 'informative' the word is to the topic (Saito, 2019b). Words in a topic share common contextual meaning.

Each review is treated as a separate document, and words are clustered together, but a word can appear in more than one document, and each document can be categorized by more than one topic. A model determines the likelihood of terms co-occurrence, and based on this detection a topic is formed. For this case, Latent Dirchlet Allocation (Blei et al., 2003), one of the most popular methods for collections of discrete data such as text corpora, is used. LDA is a three-level hierarchical general probabilistic Bayesian model.

For example, the topic under number one (Figure 32) is clearly about giving drones as a Christmas or Birthday gift, the second one – about their technical characteristics, the third and the sixth – about technical characteristics and quality.

Figure 32: Topics modelled for Amazon reviews on Force 1 Mini Drones



Source: Own work

Generally, topic modelling might be useful when composing texts for advertisement in order to hit people's 'pain', catch their attention and effectively reply to their requests. Of course, it is not the only use of web and social media mining, but these two methods might become very handy when elaborating on marketing planning process stages.

6. DEVELOPING MARKETING STRATEGY

Marketing strategy is a broad plan of managerial initiatives and actions relating an organization to its customers and markets. Marketing strategy focuses on strategic decisions necessary to allocate resources. It concerns managerial actions that have long-term effects,

and decisions relating to marketing strategy are made by marketing executives and implemented by many others through the organization and beyond (Shankar & Carpenter, 2012).

The high-level and directional insights into market opportunities collected at the stage of situation analysis serve as the foundation for building a high-level marketing strategy, which, however, requires analytical work (Chiu & Tavella, 2008).

When elaborating on marketing strategy, knowledge management process is applied. Knowledge discovery and learning is an iterative process that extends the collection of data mining techniques into a knowledge management framework. After deciding to use a complete database or a representative sample, the next step is to explore data in order to get a first feel of data and select appropriate variables for proper data mining tasks.

The result of data mining efforts is "evaluated to identify the usefulness of the resulting patterns to the solution of the marketing problem and the accuracy of prediction of future customer behavior from a known set of data. This assessment gives further insights into the data set and helps the marketer to refine the data mining model. The iterative learning process continues until the model is acceptable" (Shaw, Subramaniam, Tan, & Welge, 2001, p.131).

It is crucial to compare results derived from completion of data mining tasks against selected key metrics against the initial business goal (Shaw et al., 2001). A systematic way to retain, refine and use the data model is crucial for effective decision making in the future (Chiu & Tavella, 2008).

6.1. Market segmentation, positioning; budgeting and setting measurable goals

As discussed in the first chapter, marketing strategy includes market segmentation, product positioning, and, based on the results on those tasks and conducted earlier situation analysis – setting measurable goals and measuring the results, with the techniques discussed earlier in this section. As also discussed in the first chapter, marketing orientation has shifted towards customer centrism, which leads to creating customer-centric marketing strategy, as well.

Such marketing strategy puts customer in the center of marketing activities, paying special attention to customer relationship management – and, thus, to customer analytics. Data generated from interactions with customers can bring meaningful insights about customers, as it will be presented in the examples below.

Marketing tools such as RFM and Market Basket analyses are often used to conduct initial market segmentation. Translating this business task into data mining terms, it refers to classification or clustering data mining tasks, as evident from the cases below.

6.1.1. RFM: how often do customers buy and how much they are willing to spend

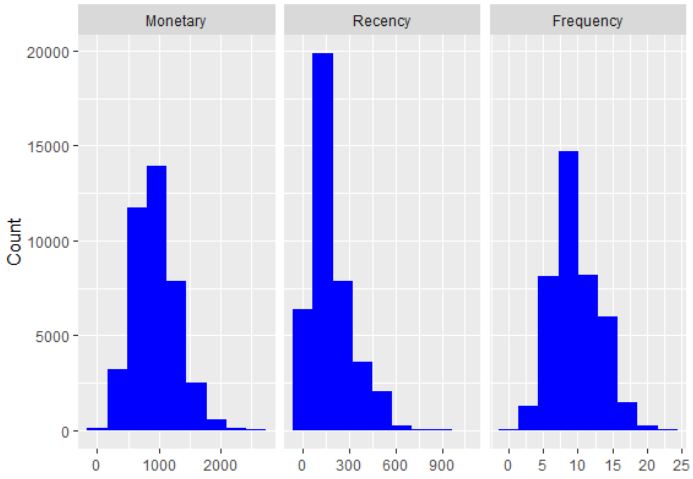
RFM (recency, frequency, monetary value) are ones of the most popular database marketing techniques used to quantify customer transaction history. RFM analysis might be the first 'predictive model' used in database marketing (Blattberg et al., 2008).

RFM analysis can be classified as one of value-based segmentation techniques. It tells which customers are valuable and allows to focus only on those who are of value for the business. Such customer analytics can enable the profitability of customer base by reducing acquisition costs, selling more per time unit to the customers, retaining customers for longer time. The ultimate purpose of customer analytics is to give agility to the business – the ability to react fast to market changes (Laursen, 2011).

A sample data offered by R Studio 'rfm' package (Hebbali, 2019) was used for this case. It contained information about a unique customer id, number of transactions/orders, total revenue from the customer, number of days since the last visit (Rsquared Academy Blog, 2019). Firstly, RFM score is computed for each customer:

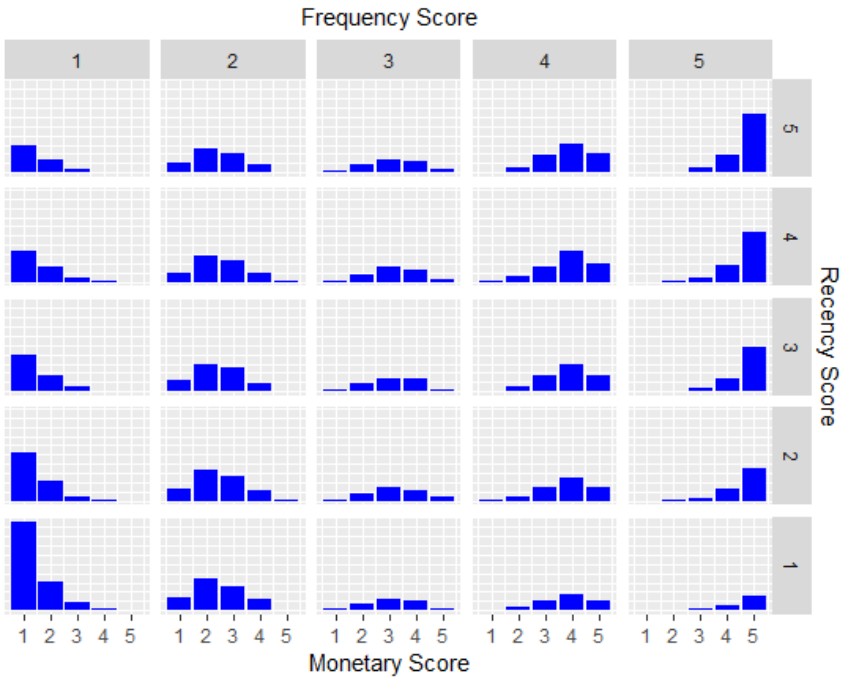
- Frequency score is computed by binning data to 5 categories, each category gets a score assigned. Higher is the purchase frequency, higher is the score assigned.
- Recency is calculated by binning most recent visit date into 5 categories, where 1 is assigned to purchase occurred in distant past.
- Monetary value is computed based on the revenue generated by each customer and then, again, binned into 5 categories, where 5 is assigned to the customers with higher spending.
- Lastly, based on three above-mentioned scores, a single RFM score is calculated by binning concatenated values of Frequency, Recency, and Monetary variables.

Figure 33: RFM histogram



Source: Hebbali (2019).

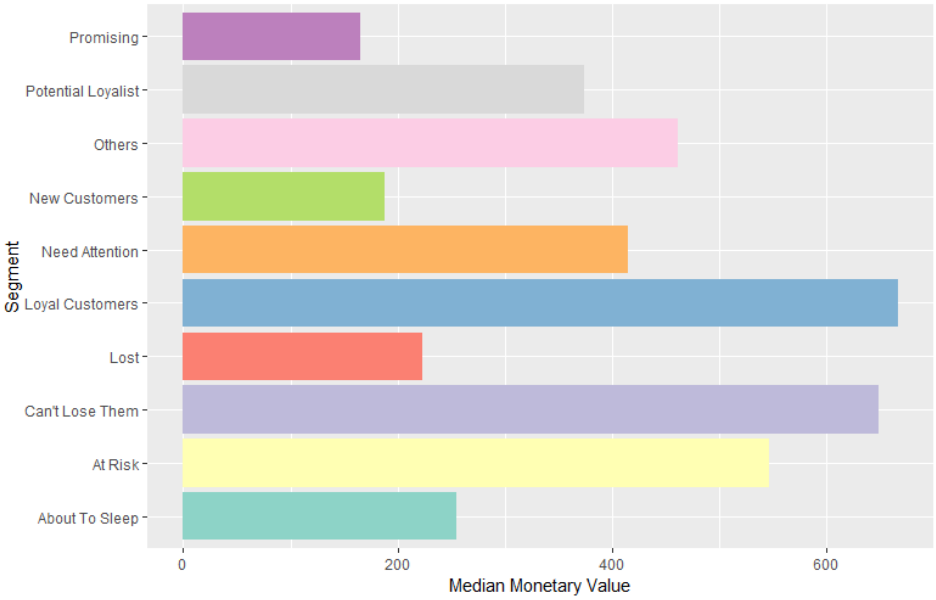
Figure 34: RFM histograms matrix



Source: Hebbali (2019).

Customers with the highest RFM are considered to be the most responsive to new offers (Correia, 2016); those with high frequency and monetary but low recency values require some kind of reminder (Hebbali, n.d.). After all scores are computed, histograms and bar charts illustrating distributions of each can be plotted (Figures 33 and 34).

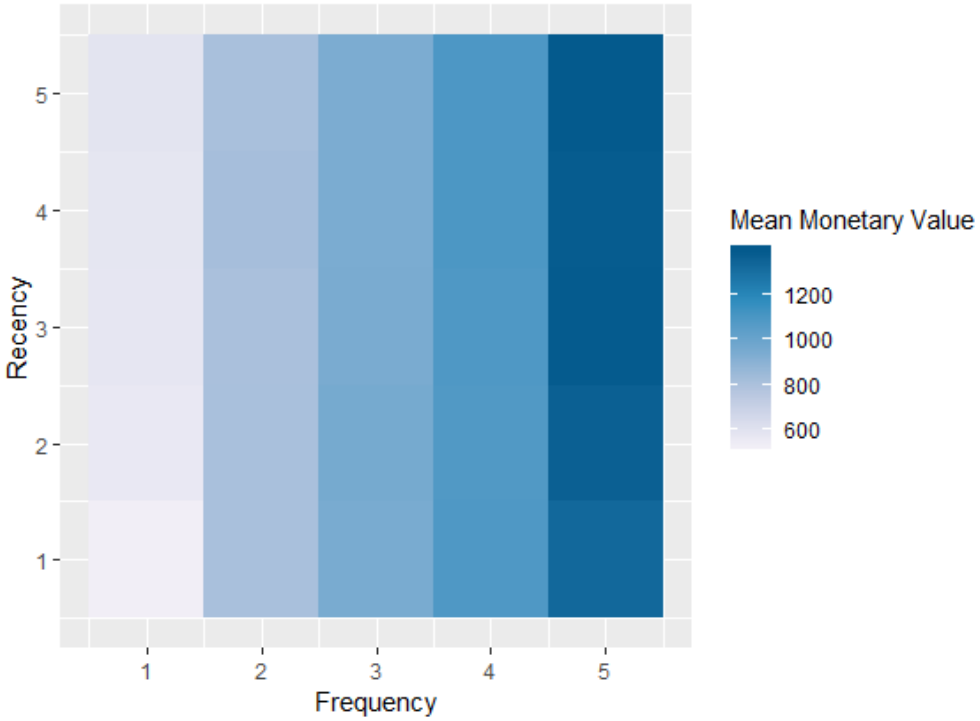
Figure 35: Median monetary value by customer segment



Source: Hebbali (2019).

Another way to graphically represent monetary expenditure in relation to recency and frequency values is heat map which shows average spending for different combinations of recency and frequency values. Logically, more customers buy, more they spend (Figure 36).

Figure 36: RFM heat map



Source: Hebbali (2019).

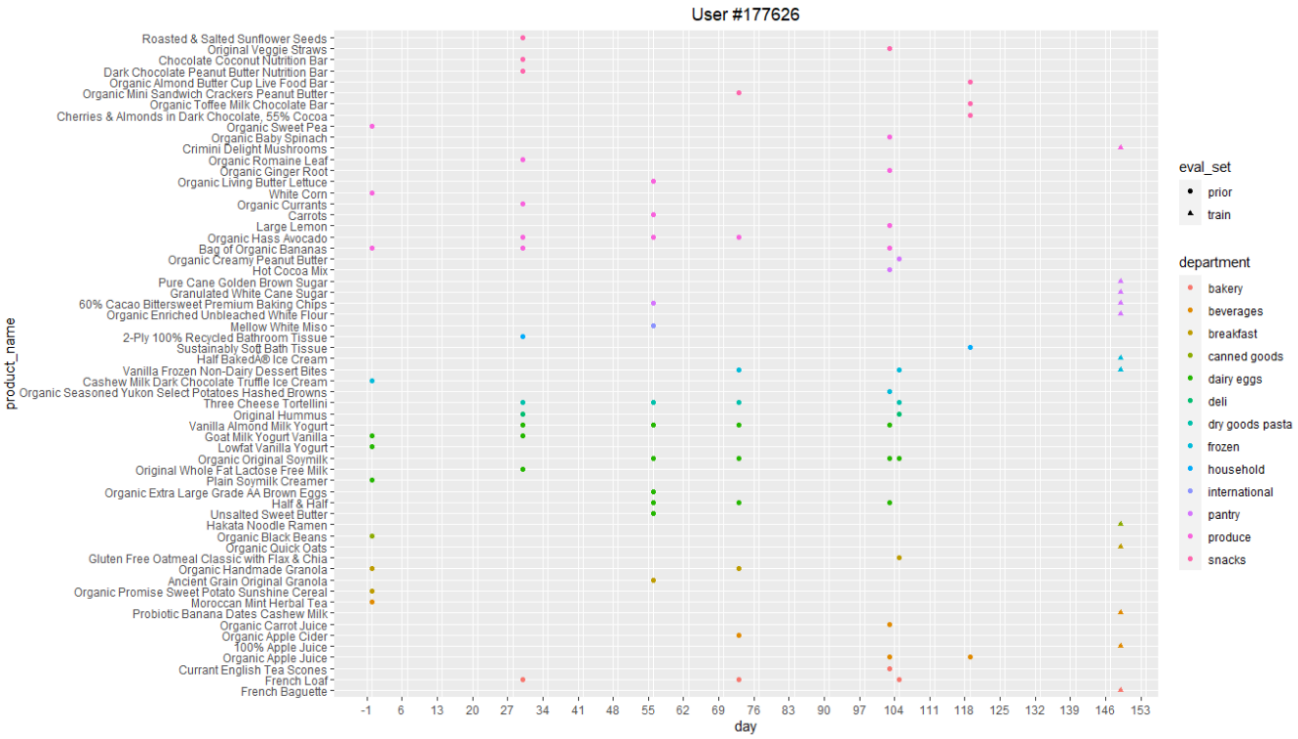
RFM is a powerful marketing analysis which stands in good stead for many more complex data mining techniques.

6.1.2. User story: what does each customer buy

Sometimes it is useful to examine individual customers rather than generalized categories. Using the dataset from clustering case (Kaggle, 2017) and another Kaggle competitor's notebook (Weisberg, 2017), user stories were created. Figure 37 represents a user story of user 165010. It contains info about user's purchases throughout the time segmented by departments and product names.

For example, this user quite often buys products which could be categorize as organic/bio: organic juices, eggs, milk, avocado, etc. Possessing this information, markers could advertise healthy lifestyle-related products and goods; new flavors of almond milk, etc. Such analysis helps brainstorming during marketing strategy development; investigation of outliers; customer segmentation; persona profiling.

Figure 37: User story



Source: Own work

6.2. Positioning

To get an idea of how product is perceived by customers, besides already discussed techniques like opinion mining, such basic marketing tool as market basket analysis is often implemented.

Product analytics is closely related to the circumstances under which people buy this product. Market basket analysis is a powerful way to understand point-of-sale transaction data, most often using association rules, which examines customers' buying behavior and explores what is frequently bought together. This helps to increase sales through cross-selling.

Generally, every rule can be read as 'if this, then that' – 'if customer bought beer, then he purchases also diapers'. The strength of association rules is calculated based on three metrics: support, confidence, and lift. Support is the percentage of transactions that contain all of the items in the item set (products that have been bought together).

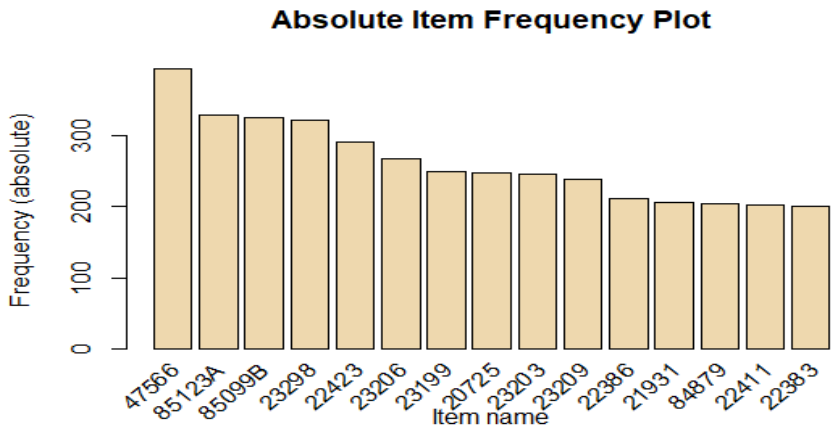
Confidence is defined as a "probability that a transaction that contains the items on the left side also contains the item on the fight side" (McColl, n.d.). In more formal words, confidence is the ratio of the number of the transactions supporting the rule to the number of transactions where the conditional part of the rule holds. Another way of saying this is that "confidence is the ratio of the number of transactions with all the items to the number of transactions with just the "if" items" (Linoff & Berry, 2011, p.300).

Lift is "the probability of all of the items in a rule occurring together" (McColl, n.d.). Another way of saying this is that "lift is the ratio of the records that support the entire rule to the number that would be expected, assuming that there is no relationship between the products" (Linoff & Berry, 2011, p.300).

Market basket analysis performs the best when the items appear in approximately equal number of transactions which prevents rules to be skewed by the most frequently purchased items. Generalization could help here: drilling up rarely bought items into their parent categories (specific type of bread – to 'bread' category) will make those items to appear more frequently. (Linoff & Berry, 2011, p.306).

To illustrate market basket analysis, a dataset is taken from (UCI Machine Learning Repository, n.d.) and consists of following variables: invoice number, stock code (item code), description (item name), quantity (purchased per transaction), invoice date and time, unit price, customer id, and country of customer residence. The case itself is based on Kumar (2018). First, data preparation steps take place: cancelled transactions are removed, data is converted into transactions format. Item frequency plot is drawn (Figure 38), then, using a priori algorithm, association rules are found under support of 0.023 and confidence of 0.82 (Figure 39). Item frequency plot shows how often, in absolute terms, each item (stock code for item name) was bought.

Figure 38: Absolute item frequency plot



Source: Own work

Figure 39: A priori algorithm

	lhs		rhs	support	confidence	lift	count
[1]	{23173,23174}	=>	{23175}	0.02732423	0.9213483	26.084588	82
[2]	{23170,23171}	=>	{23172}	0.02865711	0.8431373	25.558130	86
[3]	{23173,23175}	=>	{23174}	0.02732423	0.9213483	25.366663	82
[4]	{23170,23172}	=>	{23171}	0.02865711	0.9662921	25.216023	86
[5]	{23172}	=>	{23171}	0.03032323	0.9191919	23.986913	91
[6]	{23175}	=>	{23174}	0.03032323	0.8584906	23.636057	91
[7]	{23174}	=>	{23175}	0.03032323	0.8348624	23.636057	91
[8]	{23174,23175}	=>	{23173}	0.02732423	0.9010989	21.985348	82
[9]	{23175}	=>	{23173}	0.02965678	0.8396226	20.485427	89
[10]	{23171,23172}	=>	{23170}	0.02865711	0.9450549	20.403668	86

Source: Own work

Examining top two rules (Figures 40 and 41) with stock code and their description gives an actual information about which products are frequently bought together. In this case, those are tea plates of different colors, and milk jug together with sugar bowl and teapot, which does logically make sense.

Figure 40: Decifration of the first best rule

	StockCode	Description
182453	23172	REGENCY TEA PLATE PINK
182454	23171	REGENCY TEA PLATE GREEN
182455	23170	REGENCY TEA PLATE ROSES

Source: Own work

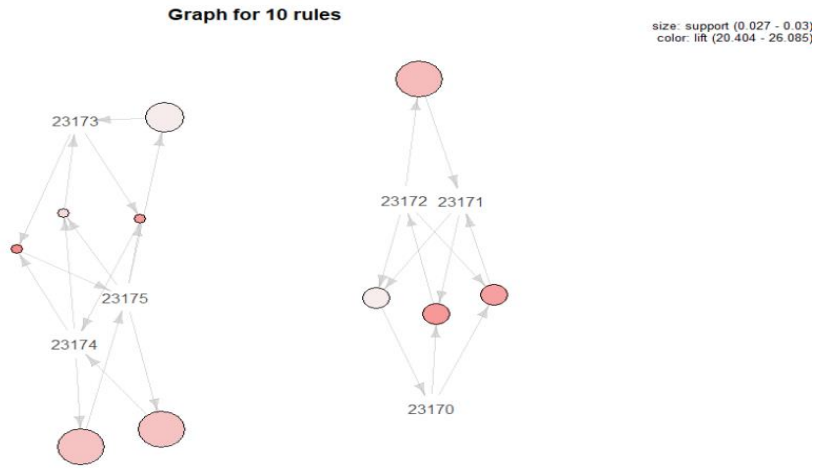
Figure 41: Decifration of the second best rule

	StockCode	Description
182450	23175	REGENCY MILK JUG PINK
182451	23174	REGENCY SUGAR BOWL GREEN
182452	23173	REGENCY TEAPOT ROSES
290582	23175	<NA>

Source: Own work

Visual representation of association rules is shown on a Figure 42. It allows to highlight the most important rules easily. In the Figure 42, arrows represent the association between lhs (left-hand side) and rhs (right-hand side) rules; the color of bubble indicates how high is the lift (brighter it is, higher is the lift), while size of the bubble is a support indicator (bigger it is, higher is the support).

Figure 42: Graph for 10 rules



Source: Own work

Sometimes these rules lead to unexpected results like in the most well-known example of beer and diapers. Association rules used for market basket analysis are powerful thing which is often further developed to recommendation system. There are plenty of types of recommenders, each used for different purposes by such giants as Amazon, Netflix, IMDb, Facebook, Google, YouTube, and others. Recommenders help to improve conversions through cross- and upselling, helping users to find the product and services which suit them the most.

6.3. Budgeting and setting measurable goals

Budgeting and measurable goals, as discussed in the chapters 1 and 2, refer to measuring the contribution of particular marketing activities to overall company's success. These are managerial decisions which, however, more and more often are based not only on managerial assumptions and past experience but are more data-driven. It is arguable what to consider to be a 'success' of a marketing campaign – this is defined by goals set by the company.

Generally, referring to digital marketing metrics, any target action completed by customer can be considered to be a 'success': number of clicks on ad in comparison with their costs, number of the filled-in forms on the website versus their cost, etc. They are mostly solved to estimation and prediction problems discussed earlier in this chapter. Then, marketing planning measures such as market share, market demand, or causal forecast; corporate financial metrics (revenue, net profit, return on sales), place metrics (transactions per customer), sales metrics (turnover rate) also refer either to regression-related problems or can be solved through descriptive analytics and OLAP technologies.

Other metrics results could be optimized using one of the techniques discussed earlier in this section. Several examples of how to find an optimal price, or how to choose the best advertising set, will be shown in the next chapter.

7. DEVELOPING MARKETING MIX

This chapter focuses in data mining techniques that can be used for developing the elements of marketing. The cases presented here, are 'student' cases, and data and models quality have a lot of room for improvement. However, they might be helpful for decisions 'what to start with'.

7.1. Product: Next Screen Button example

Product analytics can be done on many levels. In digital age, there are many analytical tools to investigate product from customer's perspective. Besides classical qualitative research, companies collect data about user behavior in the applications, on websites, landing pages, etc. Tracking user's clicks helps to understand his path and find bottlenecks, elaborate on existing and invent new features. Even if a product is, by its essence, a tangible object, and exists mostly offline, it might be anyway presented online, giving marketers plenty of possibilities for side, indirect analytics. The case presented in this section was composed based on a test task given by one of game developing companies. The main idea was to investigate users' behavior based on their clicks in the app.

The data given contained the following information (Table 5). The first row contains user's id and a timestamp with information about the app being used for the first time, in the first and second column respectively, second row contains user id and a timestamp when the user

went to another screen, in the first and second column respectively. In the case the user did not decide to proceed to another screen, there is only one entry (the first one) for such user in the dataset. The tasks to proceed with the case were:

- Add entries incremental ID column (1, 2, 3, 4…) – index
- Add a calculated column which returns the datm value from previous row for the same user, if it exists, otherwise returns the same datm
- Add a calculated column which return the time difference between datm and previous_action_datm
- Add a column which return only those userIDs for which 2 entries were registered (Table 6).

Table 5: Initial data given in Next Screen example

userID	datm_opened
userID	datm_button_clicked

Source: Own work

Table 6: The second step in Next Screen example

index	userID	datm	previous_action_datm	time_difference	Idcheck
1	101	11/1/2018 22:59			
2	101	11/3/2018 20:57	11/1/2018 22:59	46	101
3	102	11/2/2018 21:05	11/3/2018 20:57	-23	

Source: Own work

Once there is enough data, this table will show only those users who clicked on the button. It is possible to re-organize time difference and to re-calculate it in seconds. Then those users whose time of watching the tutorial is less than 5 seconds (closed immediately) could be removed.

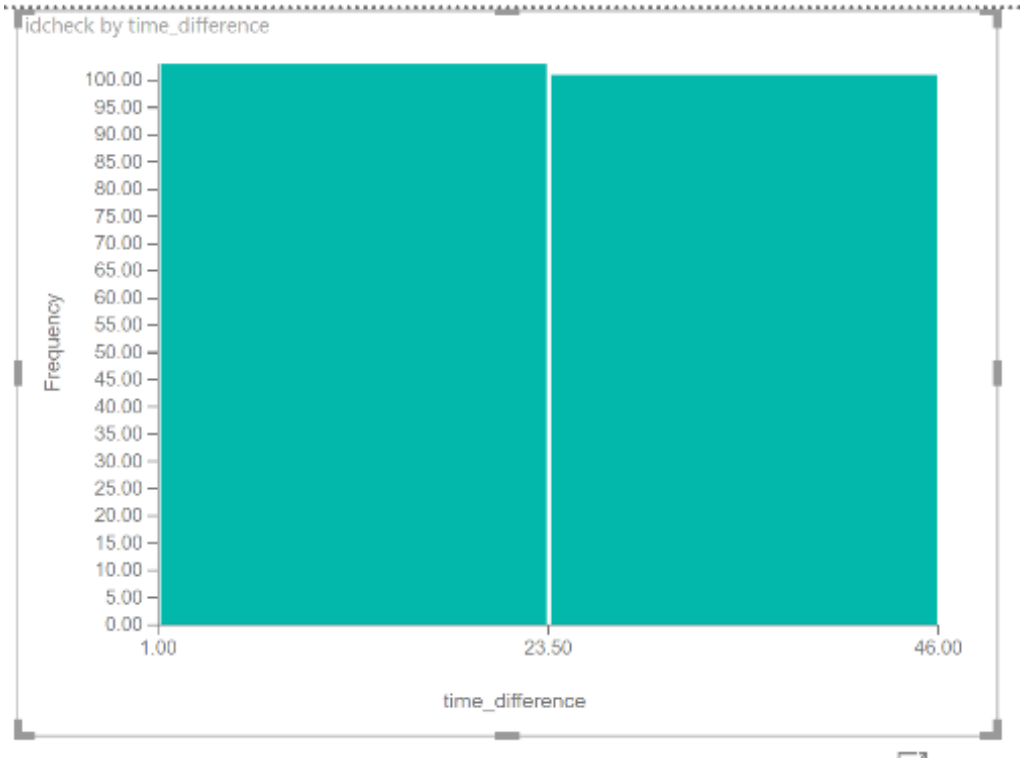
Table 7: Filter idcheck column – unselect Blank rows

index	userID	datm	previous_action_datm	time_difference	Idcheck
2	101	11/3/2018 20:57	11/1/2018 22:59	46	101

Source: Own work

From the rest of the entries, the average time difference can be derived. Mean interaction time with the button is, however, not the best option for the decision-making process. Figure 43 shows the data distribution.

Figure 43: Data distribution in Next Screen example



Source: Own work

Proceeding with examining frequency distribution through interaction time intervals in depth in the same manner as in Figure 43, would make it possible to decide upon the optimal time the button should be on the screen.

7.2. Price

Defining optimal product price is crucial for any commerce-related activity. It depends on many factors, such as price elasticity of demand which affect sales profit. Elasticity is a measure of how much buyers and sellers respond to changes in market conditions. Price elasticity of demand is a measure of how much the quantity demanded of a good responds to a change in the price of that good, computed as the percentage change in quantity demanded divided by the percentage change in price (Mankiw, 2009).

There are several methods to assess pricing analytics, with business or analytical approach (Srivastava, 2016). Generally, in study cases price-quantity relationship is assumed to be linear, and, thus, expressed through a linear function, as shown in equation (1) (Srivastava, 2016).

$$Q(p) = \alpha p + \beta \quad (1)$$

Then the revenue function will be given by equation (2).

$$R(p) = pQ(p) = \alpha p^2 + p\beta \tag{2}$$

Total profit function will be expressed as shown in equation (3):

$$L(p) = (p - c)Q(p) = \alpha p^2 - \alpha pc + \beta(p - c) \tag{3}$$

Thus, the optimal price to maximize revenue is shown in equation (4).

$$P_{\text{max rev}} = \frac{-\beta}{2\alpha} \tag{4}$$

Profit maximization is presented in equation (5).

$$P_{\text{max prof}} = \frac{-\beta + \alpha c}{2\alpha} \tag{5}$$

Table 8: Sample Sales dataset structure

sales	price	ad_type	price_apple	price_cookies
222	9.83	0	7.36	8.80
201	9.72	1	7.43	9.62
247	10.15	1	7.66	8.90
169	10.04	0	7.57	10.26

Source: Own work

Example sales data was taken in order to complete sample optimal pricing case (Zhang, n.d.) and consists of sales price, ad type, and three product prices: grape (indicated as 'price'), apple juice (indicated as 'price_apple') and cookies (price_cookies).

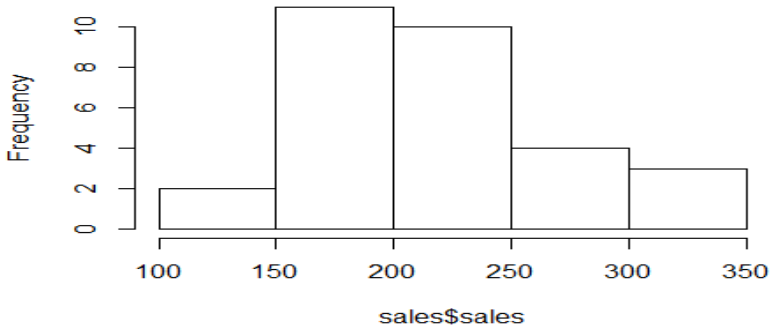
In this example, only grape price is taken into account. Data structure is shown in Table 8, and its descriptive statistics - in Table 9. Figure 44 represents sales distribution, and price-quantity relationship seems to be linear (Figure 45).

Table 9: Sample Sales dataset descriptive statistics

Descriptive	Sales	Descriptive	Price
Min.	131.0	Min.	8.200
1st Qu.	185.2	1st Qu.	9.585
Median	204.5	Median	9.855
Mean	216.7	Mean	9.738
3rd Qu.	244.2	3rd Qu.	10.268
Max	335.0	Max	10.490

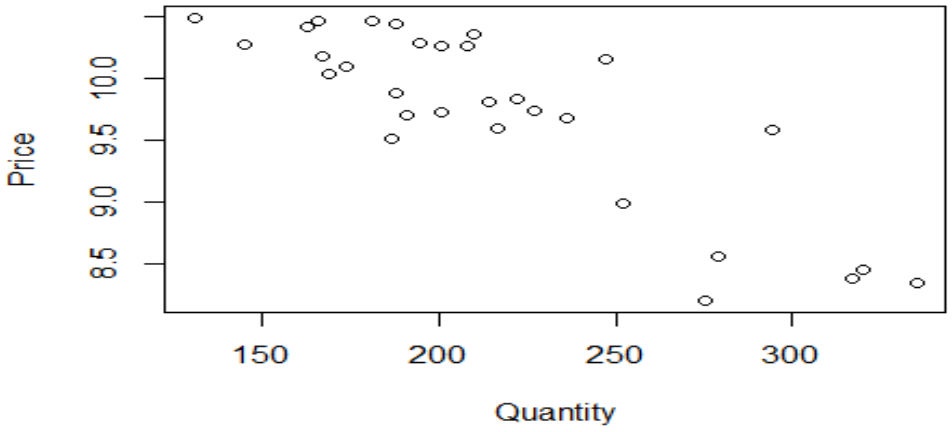
Source: Zhang, (n.d.).

Figure 44: Sales distribution



Source: Own work

Figure 45: Price-Quantity relationship



Source: Own work

Hence, linear regression might be the proper model. Results are shown in Figure 46. P-values are tending to zero which indicates reliability of selected model, and R-squared values show that the model explains more than 70% of the data.

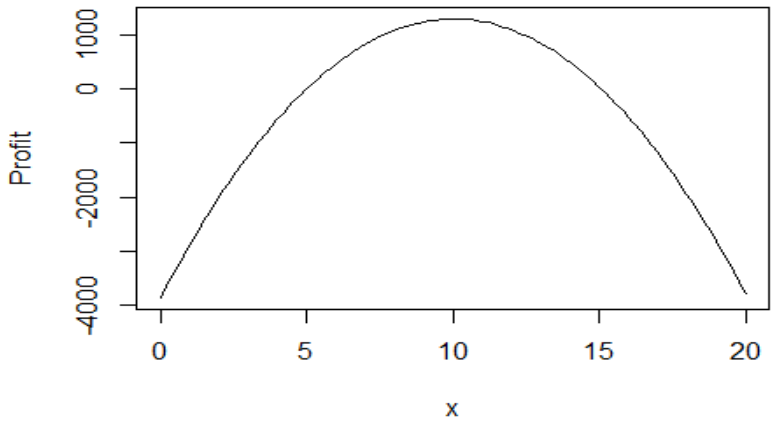
Figure 46: Linear regression output for Sample Sales dataset

```
## Coefficients:
##                Estimate          Std. Error      t value        Pr(>|t|)
## (Intercept)   833.362        72.091                 11.560              3.58e-12 ***
## salesSprice  -63.327         7.384                  -8.576         2.54e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.19 on 28 degrees of freedom
## Multiple R-squared:  0.7243, Adjusted R-squared:  0.7144
## F-statistic: 73.55 on 1 and 28 DF,  p-value: 2.545e-09
```

Source: Own work

Profit-maximization function is derived after solving linear regression (Figure 47). Optimal price for profit maximization results is 10.03942, according to the output. This case could be developed further by composing multiple linear regression and adding apple price and cookies price into the equation and deriving cross-price elasticity of demand. It would be also possible to examine effect of ads on sales.

Figure 47: Profit-maximization function



Source: Zhang, (n.d.).

From data-preparation side, outliers detection and data transformation should be conducted. From modelling side, proper model validation is missing here (Fonseca, 2017; Keramati et al., 2014). Also, more complex techniques for price auto-adjustment might be implemented. However, the purpose of this case was just to illustrate a direction, to give a basic idea of how price analysis might be conducted.

7.3. Place

The marketing mix element Place includes channels, coverage, assortments, locations, inventory, and transportation. Most frequently, when talking about Place-related marketing problems, those can be translated as regression-based or decision trees-based (estimation and prediction) problems.

For example, possessing data about performance of different marketing channels, the best marketing channel for particular marketing campaign will be the one which meets the most marketing goals (discussed earlier in the Marketing Strategy chapter). The optimal choice of marketing spending with the most target actions completed can be found through regression, decision trees, neural networks, and other techniques which were also discussed earlier.

Linoff & Berry (2011) also propose as a possible approaches the following:

- To segment existing customers and then to profile the profitable ones using their demographic and geographical characteristics to build a profile of a typical profitable customer
- To profile each marketing channel using the same variable which were used to profile customers
- Using similarity models discussed earlier, to estimate the distance from each advertising channel to the typical profitable customer

Smaller is the distance – better the channel corresponds profitable customer profile. Thus, advertising through the channels with the lowest scores will bring the most profit.

7.4. Promotion

In digital marketing, Facebook is one of the main advertising platforms. In the second quarter of 2017, Facebook reported advertising revenue of $9.16 billion which is a 47 percent increase over the same quarter last year. The number of daily active users also increased and resulted in 1.32 billion in June (17% increase), and the number of monthly users hit the mark of 2.01 billion on June 30 (Facebook Investor Relations, 2017).

The mechanics of Facebook advertising can be described in a following way:

- Choose one of campaign goals – the most popular are ‘Traffic’ (leading users from Facebook to a target platform), ‘Engagement’ (reaching as many people as possible), ‘Conversions’ (increasing conversions on a target platform).
- Choose a proper audience. There are many parameters available, among which the demographics and category of interests are the most important. The category of interests is a ‘bubble’ of user interests formed based on their behavior. For example, by choosing a ‘Science’ category it’s possible to reach people who liked the Pages of this

category. It is also possible to choose users' education, working position, online behavior patterns, etc.

- Choose a proper budget. Facebook as a company is interested in the maximum ads spending possible, as it is visible from the upper-mentioned statistics. At the same time, Facebook has its own advertising politics: it will more likely show more expensive and more creative, interesting one over the cheaper and less pleasant, relevant ad for the users.
- Choose a proper combination of text and visuals. Facebook ads look like a video + text, a pic + text, a gallery + text, or a combination of those. This choice is the crucial, as Facebook will show more effective ads over the less effective ads much more often and for much less price. The 'more effective' means 'more attractive' for the users. How to choose such a combination? Marketers usually do it manually by testing many combinations. For example, in one campaign they test one picture, combining it with different texts. Then they choose the best text and test it with many different pictures/videos in another campaign.

The main aim of any advertising campaign is to get as many target actions (likes, clicks, conversion…) for as low price as possible. The next case is related to Facebook advertising and aims to:

- Find correlations between conversions and other variables in order to understand what to put more attention on while creating an advertising campaign on Facebook
- Find and optimal combination of price and approved conversions in order to choose the optimal text + picture combination
- Predict approved conversions number in order to be able to save on ineffective combinations of text + picture in the future, to know what to expect to make it easier to conduct strategic analysis of business and refining the strategy, seeking for the most effective way of business development

The analysis is conducted with the help of free, user-friendly, easy to use, and relatively powerful software – RapidMiner Studio. The data is taken from the open source Kaggle database (Kaggle, n.d.-a). The dataset required further changes, as explained below

In the real life world, such analysis is recommended to conduct each time when launching advertising on Facebook. There are many project planning techniques, including Scrum, PMBOK, Agile and other methods.

This process might become a routine – in such a way it will help to plan and save budget. In the end of each circle of the process (collecting the data, conducting analysis, writing recommendations) the business people should get a list of recommendations with optimal creative combinations proposed and predictions for the next advertisement campaigns.

The data for this project is presented in a form of Excel file with the following variables:

- ad_id: a unique ID for each ad
- xyz_campaign_id: an ID associated with each ad campaign of XYZ company (assigned by advertiser)
- fb_campaign_id: an ID associated with how Facebook tracks each campaign (assigned automatically by Facebook)
- age: age of the person to whom the ad is shown
- gender: gender of the person to whom the add is shown
- interest: a code specifying the category to which the person's interest belongs (interests are as mentioned in the person's Facebook public profile, the code is assigned manually by advertiser)
- Impressions (number of times when users reacted on ad)
- Clicks: number of clicks on for that ad
- Spent: amount paid by company xyz to Facebook, to show that ad
- Total_Conversion: total number of people who enquired about the product after seeing the ad
- Approved_Conversion (number of conversions after the last step – payment)

All metadata was provided by Kaggle contributors.

The essential statistic data related to number of conversions, clicks, impressions, demographics, ad and campaign ids can be retrieved from Facebook ads manager. It is enough to conduct simple analysis. However, for this project several adjustments were required – they are mentioned below. Moreover, some additional encoding (assigning codes to particular category of interest) was pre-made before uploading the dataset to Kaggle by their authors. In the real-life world, everything depends on what has to be analyzed. Thus, variables such as encoded texts, images/videos, etc. could be added manually.

7.4.1. Correlation matrix to find the key influencing advertisement campaign features

In order so solve the first problem – to find how much each element of Facebook campaign matters – the correlation matrix was applied. Before conducting the analysis, two new columns with simulated data were added: Text and Pic. The Text column contains invented IDs of different ads texts which marketers might be using. The Pic column contains invented IDs of different creative pictures/photos for ads which marketers might be using. These is needed to understand the overall impact of text and picture on the conversion results. As it is visible from the dataset (Figure 48), for each marketing hypothesis (xyz_campaign_id column) a few pictures and several texts are used – for simplification, the opposite is not tested in this example. The results (sorted) are presented in Figure 49.

Figure 48: Facebook ads set

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ad_id	aign_id	fb_campa	age	gender	interest	Impressio	Clicks	Spent	Total_Con	Approved	Pic	Text
2	708746	916	103916	30-34	M	15	7350	1	1.43	2	1	1	1
3	708749	916	103917	30-34	M	16	17861	2	1.82	2	0	1	2
4	708771	916	103920	30-34	M	20	693	0	0	1	0	1	3
5	708815	916	103928	30-34	M	28	4259	1	1.25	1	0	1	4
6	708818	916	103928	30-34	M	28	4133	1	1.29	1	1	1	5
7	708820	916	103929	30-34	M	29	1915	0	0	1	1	1	6
8	708889	916	103940	30-34	M	15	15615	3	4.77	1	0	1	7
9	708895	916	103941	30-34	M	16	10951	1	1.27	1	1	1	8
10	708953	916	103951	30-34	M	27	2355	1	1.5	1	0	1	9
11	708958	916	103952	30-34	M	28	9502	3	3.16	1	0	1	10
12	708979	916	103955	30-34	M	31	1224	0	0	1	0	1	11
13	709023	916	103962	30-34	M	7	735	0	0	1	0	1	12
14	709038	916	103965	30-34	M	16	5117	0	0	1	0	1	13
15	709040	916	103965	30-34	M	16	5120	0	0	1	0	1	14
16	709059	916	103968	30-34	M	20	14669	7	10.28	1	1	1	15
17	709105	916	103976	30-34	M	28	1241	0	0	1	1	1	16
18	709115	916	103978	30-34	M	30	2305	1	0.57	1	0	1	17
19	709124	916	103979	30-34	M	31	1024	0	0	1	1	1	18
20	709179	916	103988	35-39	M	15	4627	1	1.69	1	0	1	19
21	709183	916	103989	35-39	M	16	21026	4	4.63	2	1	1	20
22	709320	916	104012	35-39	M	15	1422	0	0	1	1	1	21
23	709323	916	104012	35-39	M	15	7132	2	2.61	1	0	1	22
24	709326	916	104013	35-39	M	16	12190	2	3.05	1	0	1	23
25	709327	916	104013	35-39	M	16	12193	2	3.06	1	1	1	24

Source: Own work

As the results show, spending, number of clicks and impressions have almost no influence on the number of approved conversions, while gender is the most important factor here.

Despite the choice of target audience is undoubtedly important, the importance of text and pictures is underestimated in this matrix. This might happen because it is really hard to estimate in numerical terms their importance in accordance with Facebook algorithms (the exact criteria by which Facebook chooses the best or the worst texts are not known, so it is not possible to add these variables into the model).

Figure 49: Attribute by its weight in advertising campaign for conversion variable

attribute	weight ↓
gender	1
age	0.944
interest	0.888
Total_Conversion	0.329
Pic	0.270
Text	0.167
Impressions	0.036
Spent	0.013
Clicks	0

Source: Own work

The number or impressions could be considered as a possible criterion for texts and visual parts importance evaluation. However, if the same Correlation Matrix analysis for the Impressions variable (labeled) is conducted, the correlation between those is still not obvious (Figure 50).

Figure 50: Attribute by its weight for the impression variable

attribute	weight ↓
gender	1
age	0.969
interest	0.857
Approved_Conver...	0.396
Pic	0.229
Total_Conversion	0.214
Text	0.134
Spent	0.010
Clicks	0

Source: Own work

Thus, the conducted analysis definitely shows the importance of each elements, but limitations (lack of information about Facebook internal algorithms) underestimate several of them. This fact should be also taken in account.

7.4.2. Finding the cost of conversion

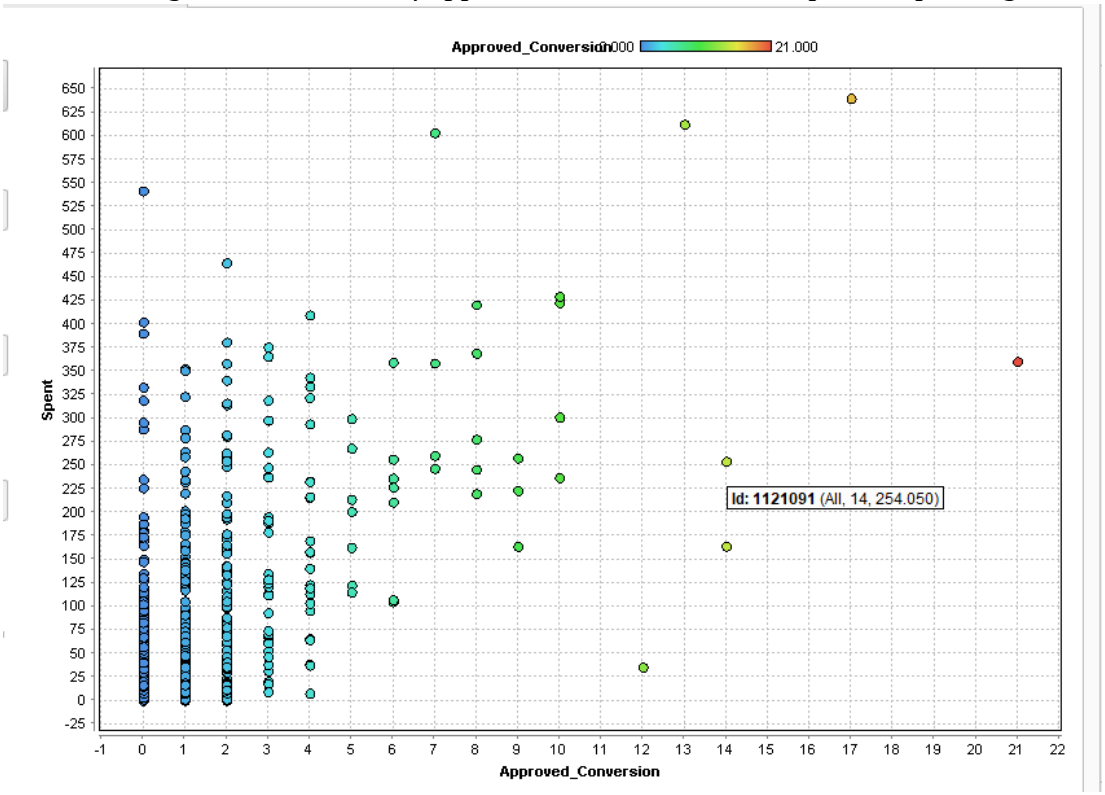
There are two ways of finding the optimal combination of text-and-visual creative pair: the visual one (through the graph), and the mathematical one. More conversions for less spending is the main rule to find an optimal combination. A simple newly-added parameter – cost of conversion (spending divided by number of approved conversions) – might help to identify a suitable pair of text and picture. Smaller it is – better it is.

This simple scatter plot (Figure 51) shows the number of approved conversions with respect to the spending. The goal is to maximize the number of approved conversion with the minimum possible spending.

The outliers are not considered (as they are either too expensive or gained too low number of conversions), so it is needed to look somewhere in the middle. A good option is the ad

with id 1121091 under which picture number 5, text number 495 were applied. This ad was targeting males of 30-34 years old, of the group of interests labeled with a number ‘10’.

Figure 51: Number of approved conversion with respect to spending



Source: Own work

Figure 52: The minimum cost of conversion

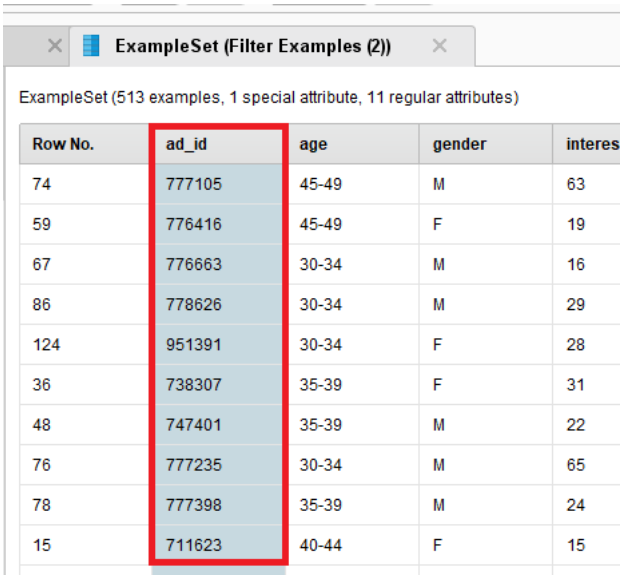
Filter (513 / 513 examples): all

Spent	Total_Conve...	Approved_C...	Pic	Text	Cost of c...
0.180	1	1	3	244	0.180
0.490	1	1	3	192	0.490
0.570	1	1	3	221	0.570
0.720	1	1	3	295	0.720
1.590	2	2	4	461	0.795
0.860	1	1	3	113	0.860
0.980	1	1	3	139	0.980
0.990	1	1	3	252	0.990
1.050	1	1	3	256	1.050
1.130	2	1	2	26	1.130
1.150	1	1	3	136	1.150
1.180	1	1	3	226	1.180
1.230	1	1	3	330	1.230

Source: Own work

The mathematical way applies a simple formula to find the conversion price. By using a Generate Attribute operator, a new column with conversion cost variable was created. The results are presented in Figure 52. The first 10 ads ids with the lowest cost of conversion are the possible solution (Figure 53). By looking on these ids, another possible ads campaign attributes sets can be found.

Figure 53: The ads with the lowest cost of conversion



ExampleSet (Filter Examples (2))

ExampleSet (513 examples, 1 special attribute, 11 regular attributes)

Row No.	ad_id	age	gender	interes
74	777105	45-49	M	63
59	776416	45-49	F	19
67	776663	30-34	M	16
86	778626	30-34	M	29
124	951391	30-34	F	28
36	738307	35-39	F	31
48	747401	35-39	M	22
76	777235	30-34	M	65
78	777398	35-39	M	24
15	711623	40-44	F	15

Source: Own work

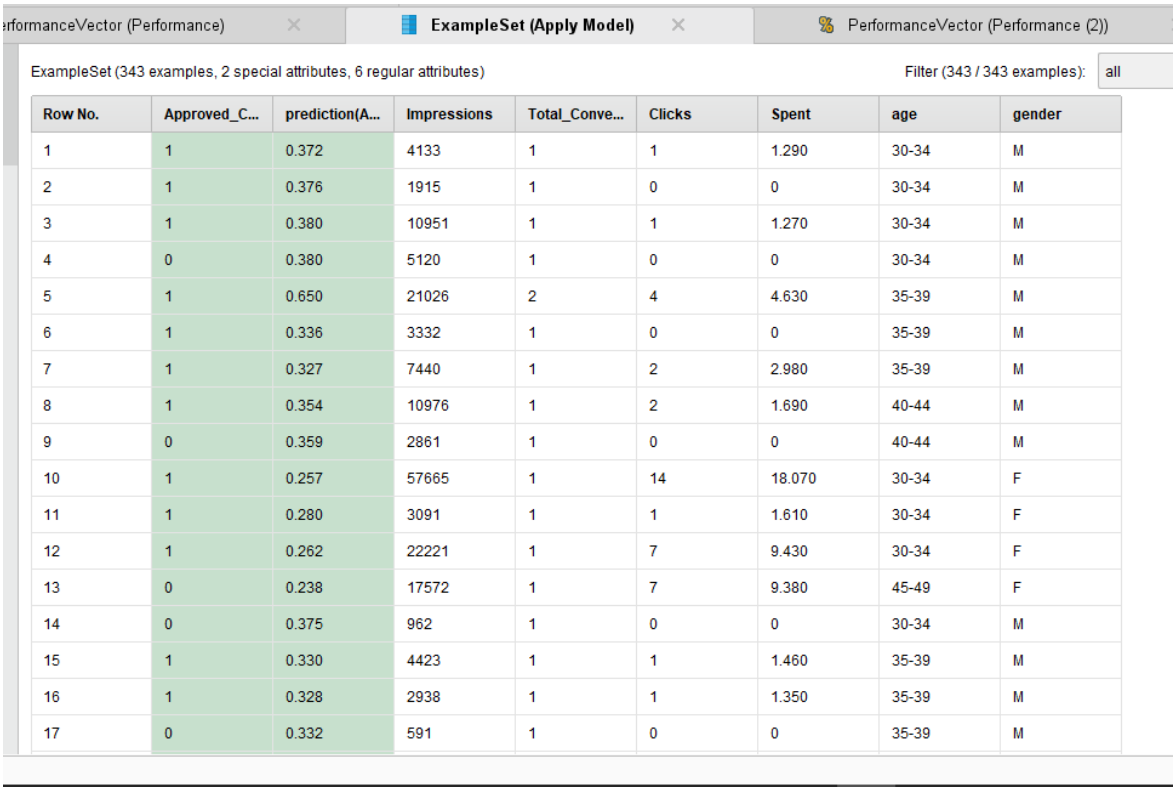
7.4.3. Predicting number of conversions

Simple linear regression analysis with respect to the approved conversions variable (labeled) is a common model for forecasting numeric variables. The results are presented in the Figure 54.

The performance evaluation gives root mean squared error equal to 0.795, which is very high, so this model is not enough accurate, and, thus, not suitable for this prediction analysis. The next algorithm to try is artificial neural networks.

Before conducting the analysis, some changes were applied in the dataset. Several columns with historical data about approved conversions were added. This was needed so the model takes in account not only the parameters related to the particular variable but also historical data about conversions. The results and performance estimation are presented in Figures 55 and 56. As it is visible, this model performs much better than the Regression analysis. Thus, Neural Networks can be considered to be a suitable model for prediction analytics in this case, and the model could be trained further.

Figure 54: Regression analysis results



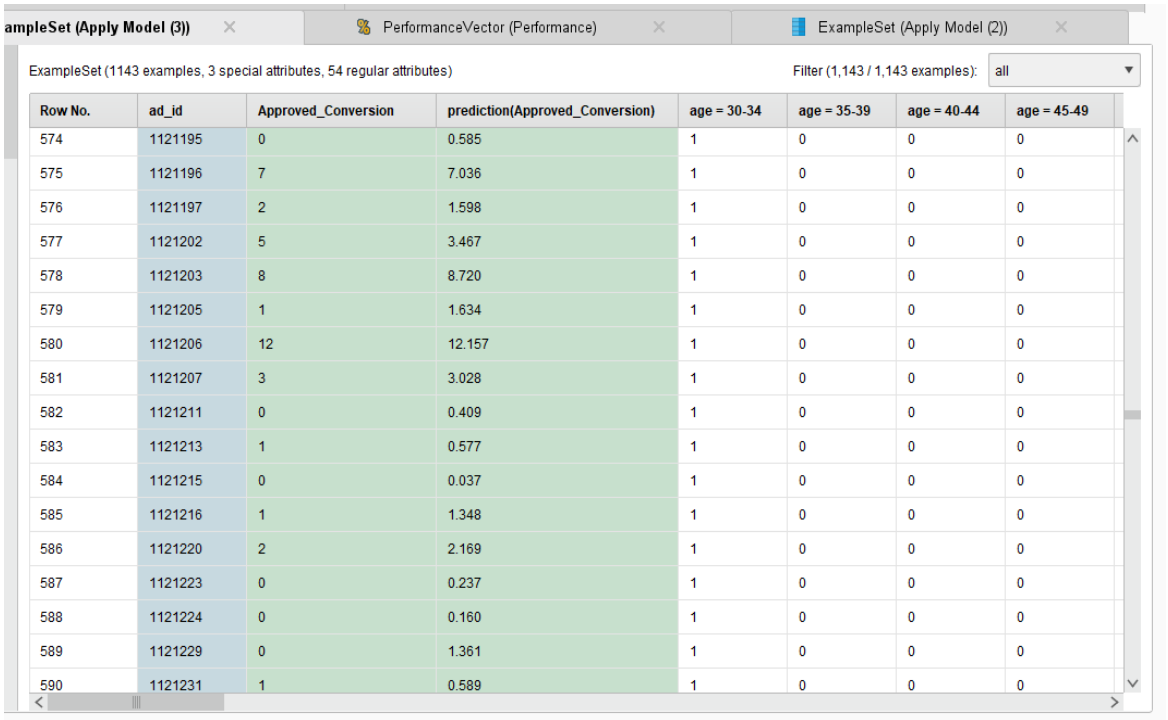
erformanceVector (Performance) | ExampleSet (Apply Model) | PerformanceVector (Performance (2))

ExampleSet (343 examples, 2 special attributes, 6 regular attributes) Filter (343 / 343 examples): all

Row No.	Approved_C...	prediction(A...	Impressions	Total_Conve...	Clicks	Spent	age	gender
1	1	0.372	4133	1	1	1.290	30-34	M
2	1	0.376	1915	1	0	0	30-34	M
3	1	0.380	10951	1	1	1.270	30-34	M
4	0	0.380	5120	1	0	0	30-34	M
5	1	0.650	21026	2	4	4.630	35-39	M
6	1	0.336	3332	1	0	0	35-39	M
7	1	0.327	7440	1	2	2.980	35-39	M
8	1	0.354	10976	1	2	1.690	40-44	M
9	0	0.359	2861	1	0	0	40-44	M
10	1	0.257	57665	1	14	18.070	30-34	F
11	1	0.280	3091	1	1	1.610	30-34	F
12	1	0.262	22221	1	7	9.430	30-34	F
13	0	0.238	17572	1	7	9.380	45-49	F
14	0	0.375	962	1	0	0	30-34	M
15	1	0.330	4423	1	1	1.460	35-39	M
16	1	0.328	2938	1	1	1.350	35-39	M
17	0	0.332	591	1	0	0	35-39	M

Source: Own work

Figure 55: Neural Networks algorithm output



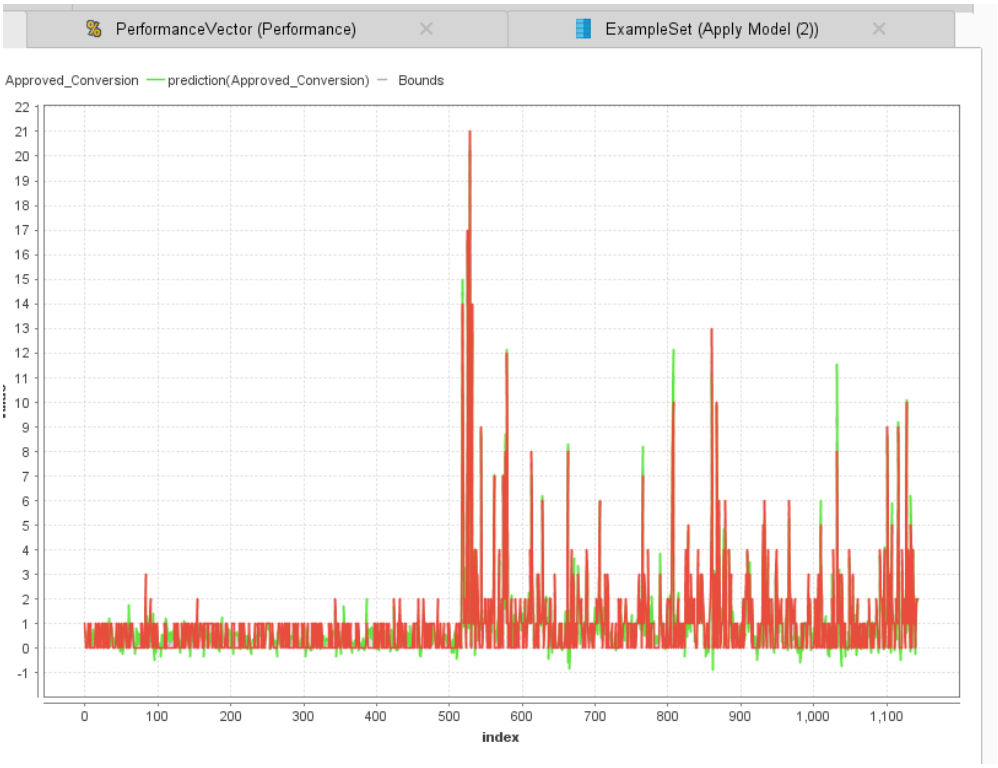
ampleSet (Apply Model (3)) | PerformanceVector (Performance) | ExampleSet (Apply Model (2))

ExampleSet (1143 examples, 3 special attributes, 54 regular attributes) Filter (1,143 / 1,143 examples): all

Row No.	ad_id	Approved_Conversion	prediction(Approved_Conversion)	age = 30-34	age = 35-39	age = 40-44	age = 45-49
574	1121195	0	0.585	1	0	0	0
575	1121196	7	7.036	1	0	0	0
576	1121197	2	1.598	1	0	0	0
577	1121202	5	3.467	1	0	0	0
578	1121203	8	8.720	1	0	0	0
579	1121205	1	1.634	1	0	0	0
580	1121206	12	12.157	1	0	0	0
581	1121207	3	3.028	1	0	0	0
582	1121211	0	0.409	1	0	0	0
583	1121213	1	0.577	1	0	0	0
584	1121215	0	0.037	1	0	0	0
585	1121216	1	1.348	1	0	0	0
586	1121220	2	2.169	1	0	0	0
587	1121223	0	0.237	1	0	0	0
588	1121224	0	0.160	1	0	0	0
589	1121229	0	1.361	1	0	0	0
590	1121231	1	0.589	1	0	0	0

Source: Own work

Figure 56: Neural Networks algorithm graphic performance evaluation



Source: Own work

This example shows how it is possible to find more valuable variables for targeting by using simple available tools. Focusing on these ads requires spending less money and leads to improved efficiency. It helps to find an optimal combination and to make a decision regarding the best ad to spend the budget on. Neural networks turned out to be sufficiently powerful to predict the campaign results – that is, helping to judge about future campaign effectiveness and to develop an optimal marketing strategy. However, there is a plenty of room for improvement and improving the model accuracy.

There are, of course, many other ways to analyze Facebook ads– for example, as described by author in another case (Bow, 2018). Using several R packages, the author presents techniques for fast and effective analysis of many Facebook ads campaigns, including establishing KPIs, features correlation matrix (audience analytics), analysis of interests, and, finally, improving ROI (return on investment). The author highlights that the analysis could be extended by combining Facebook data with additional sources such as Google Analytics.

7.5. People

People and process are relatively newly added element in marketing mix model. How employees are connected within organization, how productive are they, what processes happen within a company. As an example is a case describing kudos hackathon organized by Celtra Inc.

The company positions itself as Creative Management Platform which combines modern technologies to make advertisers' experience better. Particularly, Celtra uses data mining techniques to elaborate on algorithms for data-driven marketing (e.g. data-driven decision about which picture to choose for Facebook ads) and Business Intelligence to deliver superior value to its customers (e.g. all digital ads and their measurement delivered through personalized insights and custom reports). In the year 2018, Celtra organized a hackathon, where a particular business problem was not stated but rather an idea to elaborate on was given.

In tech companies, there is a way to express gratitude to a colleague who contributed to a part of code, helped to find a solution or simply shared a file - 'kudos'. This is what Oxford dictionaries say about kudos: "kudos comes from Greek and means 'praise'" (Dictionaries, n.d.). Kudos help to increase the recipient's rating in some communities, earn another badge and (like in PowerBI community), at some point, to get a DataNaut title.

Celtra wanted to create an environment for giving kudos within a company, then to integrate it to Slack; kudos, entered through this environment, should be recorded in a database (AWS, MySQL, Azure, MongoDB…) according to algorithms written in the backend, and then data insights should be extracted from the data acquired on frontend and sent to a database from backend. Such environment could help to increase employees' internal motivation by encouraging them to give and receive more kudos; improve cross-departmental experience exchange; make it possible to elaborate on individual employee's rating system.

This case is also a good example of data cleansing and preparation. It is, probably, one of those cases when real-life data might be less demanding (ETL-wise) than the mocked one. The tools used for both cases are R Studio and Microsoft Power BI.

7.5.1. Kudos hackathon: data preparation

Some dummy data to play around and to get initial insights from it while developers were working on the backend side was provided. The dataset consisted of 10000 dummy entries in one .csv file, containing information about sender team (working department: Ad Serving, DevOps, Quality Insurance, etc.), sender position (Senior Software Developer, Director, Project Lead, etc.), kudos text (message body, in this case taken from a fortune cookies texts generator but in the live case - entered by an employee in the frontend), kudos sent date and time; recipient id; recipient name; recipient last name; recipient team; recipient position.

Some data was omitted on purpose - around 200 entries for recipient/sender ids, positions and teams. Initially, there were not column names, as well. The second .csv file represented information about first name, last name, email, team, and manager id. In this file there was no information omitted.

As there were two clearly related tables, Power BI was chosen as a well-known tool for relational databases. For the live-case scenario, when backend started to work, the data would be stored in MongoDB and then analyzed in Python or R. There is a connector for MongoDB-Power BI, as well, but it would be strange to connect a tool for analyzing relational databases with a NoSQL database. Yet, it is possible.

After the data was imported into PowerBI Desktop, data preparation and cleansing steps were hold. There were over 200 missing rows for each of the columns: sender_id, recipient_id, sender_team, recipient_team, sender_position, recipient_position, and several others. Altogether around 10% of data was missing, and it should be restored.

Of course, in the 'real life' such situation would be unlikely to happen as ids are usually assigned automatically. But anyway, several DAX formulas were applied, after which only around 6-13 rows with missing values were left per one issue. For the main columns of interest - sender's full name, sender's and recipient's positions - there were less than 50 rows with blank values left, which was a satisfactory result.

Before going further, it was important to add several columns to employees table. Initially, every employee's id corresponded to a manager's id. It would be useful to obtain manager's names, as well. But as id is given in a not-numeric format, first indexes were assigned to each employee by using Query Editor and adding a simple index column with increment numbers from 1 to 117.

Indexes to managers were assigned, and then, to make visualization of results easier, managers' names and positions were looked up. At this point, the initial data cleansing was finished.

7.5.2. Employee's scoring development

The next step was to elaborate on employee's scoring. For this purpose, a table of themes was introduced. It consisted of 3 columns: theme_id, theme, and scores. Each theme had several scores. The idea behind was that, besides it was not given, it might be in the real life that on front-end side a sender should choose one of the themes of his kudo: "favor", referring to all kudos given for any type of help, like helping to find a file, sending a proper source, giving an advice, etc., and this theme was worth 10 scores; "knowledge sharing" was referring to kudos given within one team for helping with some task where extra-knowledge was required, worth 20 points; "knowledge input" of 30 points for helping somebody from another team, and "managerial task" of 40 scores for successfully completing a task given by a manager of an employee.

This table was completely invented for simulation purposes. Then, in initial kudos.xls table an additional column of theme_id was added for simulation purposes.

Then, final kudos scoring was added applying new conditions. The idea behind this scoring system was to take into account if kudo was sent by somebody who was on the higher position than the chosen employee. The chart which showed possible managerial roles in the company was created: Director, CEO, COO, CTO. There were also other roles applicable to particular (lower by rank) positions: Lead Software Developer with correspondence to Software Developer, Senior Software Developer, and Software Engineer Intern positions, and Lead DevOps Engineer to DevOps Engineer. Another column which showed if a kudo was sent by somebody of a higher position was created. It was also important to exclude those who sent kudos to themselves.

The final column of main kudos scoring uses the following logic behind:

- If kudo was sent by a leader and was not sent by an employee to himself, then kudo scoring is equal to 40;
- If kudo was not sent within one team, its score is 30
- If kudo was sent within one team, its score is 20
- If none of above-described conditions are met, the score will be taken from Scores column of Themes table

Next step was relatively close to the final scoring system variant. It was assumed that employee's score might depend on several conditions:

- Number of kudos received
- Number of kudos sent
- "Quality" of kudos received
    - Kudos received from a leading position worth more than from a co-worker
    - Kudos received from a co-worker from another team worth more than from your own team
    - Kudos received and sent by the same person are not counted

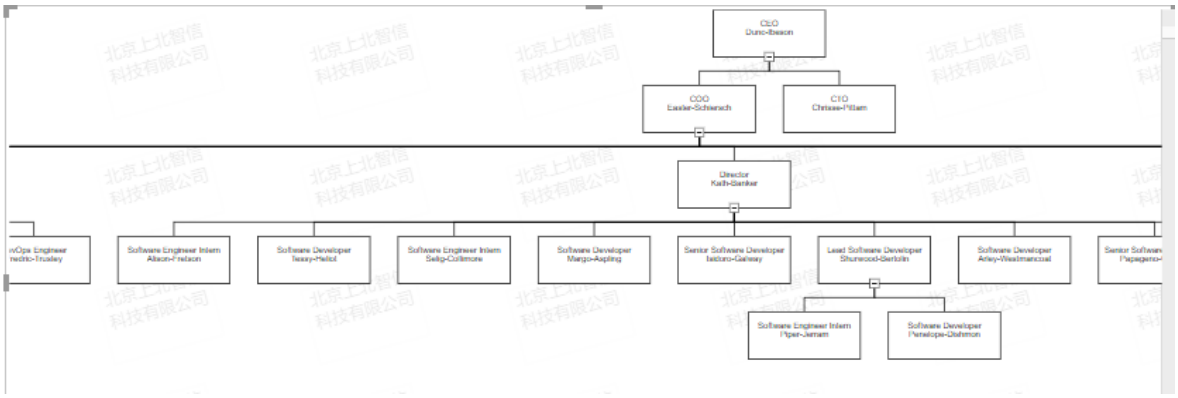
For these purposes, each kudos entry was evaluated in accordance with above-described conditions. Than the initial scoring for each employee was calculated using the following logic: sum of (count of kudos of each out of 4 categories multiplied by category (theme) scoring) filtered by employee's full for a specific time period

The next thing was to normalize those scores to make it easier to visualize. Normalization generic formula is presented in the equation (6).

$$\frac{x-min(x)}{max(x)-min(x)} \quad (6)$$

As the final hackathon result was supposed to be an interactive dashboard, instead of 'static' columns more dynamic 'measures' were used. All columns and measures are filtered by recipients' ids as mainly kudos are sent by managers, and filtering by senders' ids would lead to have TOP management in the top of employees' rating. The proposed evaluation system is oriented on employee's, not managers' rating.

Figure 57: Company structure graph

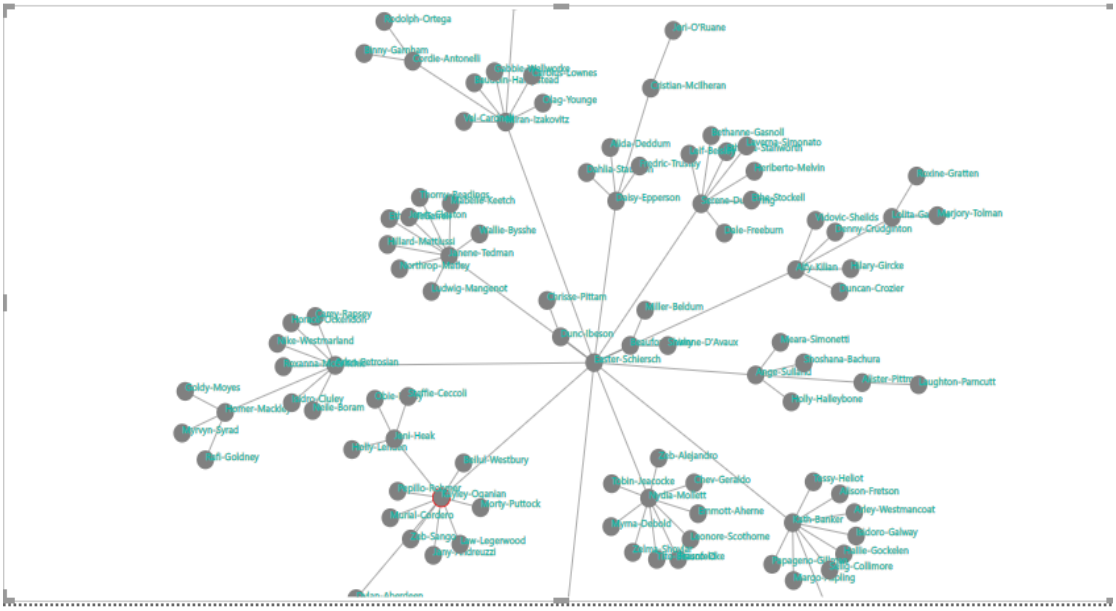


Source: Own work

As a guest network was used, there was no chance to connect Power BI to established backend, and, thus, cleaned 'static' data was used. The result was an interactive dashboard. The first page represented company hierarchical structure with names and positions, based on earlier described path function (Figure 57).

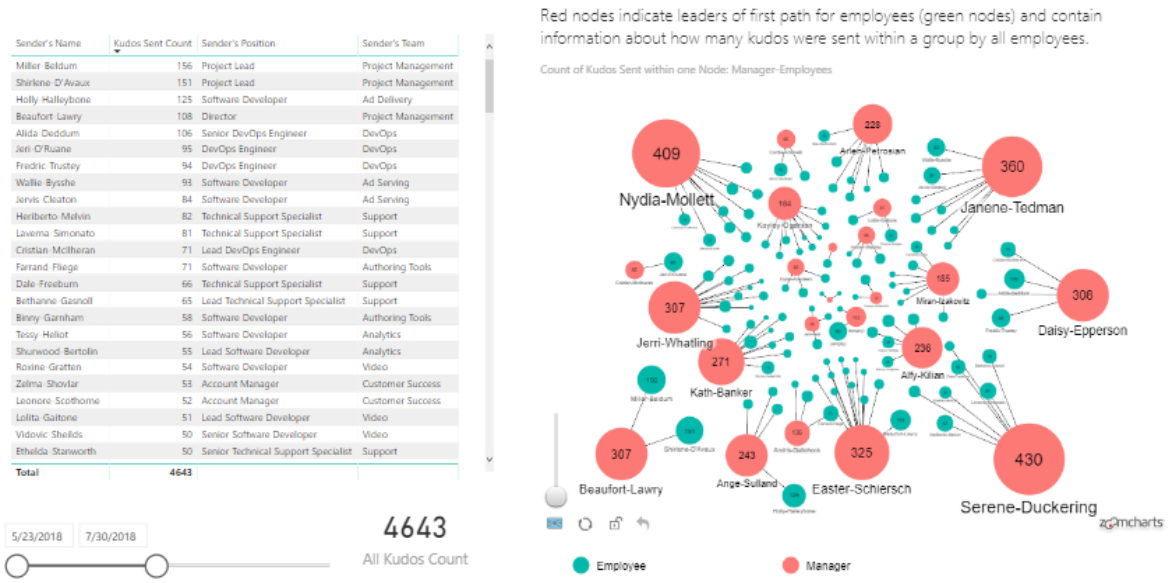
The second page showed employees-manager relationships. It is useful to form teams for working on projects, for team building purposes, etc. In marketing, networks visualization helps to segment target audience (Figure 58).

Figure 58: Global hierarchical network



Source: Own work

Figure 59: Employee-manager relationships from managerial perspective



Source: Own work

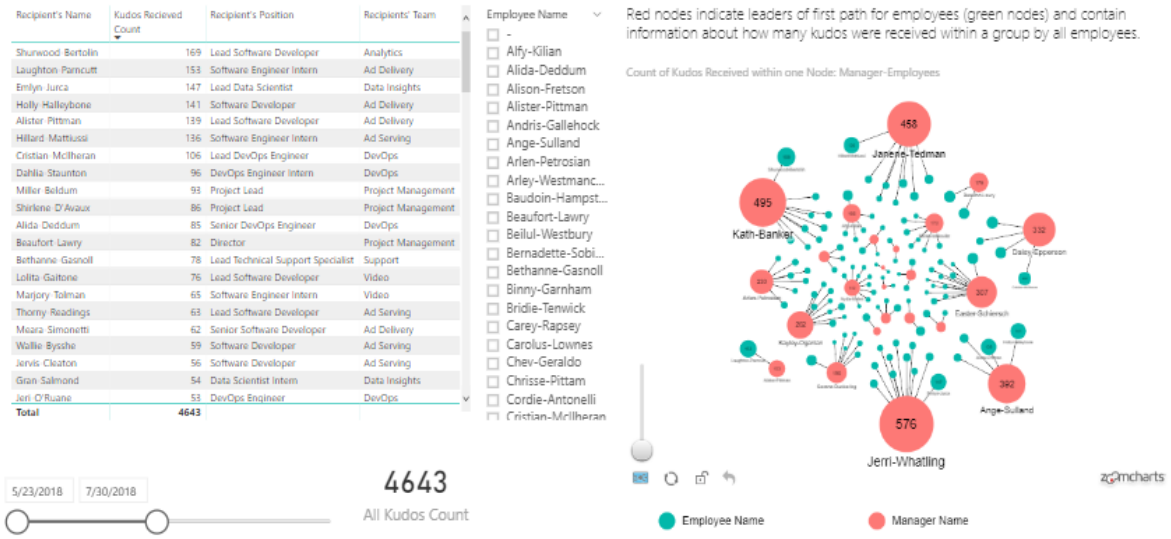
The third page represented employees-manager relationships from managers' perspective. It shows sender full name, count of kudos sent by this employee, his/her position, and a team (Figure 59). The graph on the right side, red circles indicated leaders of the first (initial, root) path for employees (green circles) and contain information about how many kudos were sent within these local nodes by the employees. This graph could help to identify the most kudos-active teams.

It might be also helpful to know whose employees got the most kudos. The fourth page was a duplicate of the third one, with the only difference that all visuals were filtered now by recipients' names (Figure 60).

Such visualizations might be useful to show customers' networks, as well as to represent sales and advertisement data in a comprehensive manner, where nodes could be represented by groups of customers or ads in different sources.

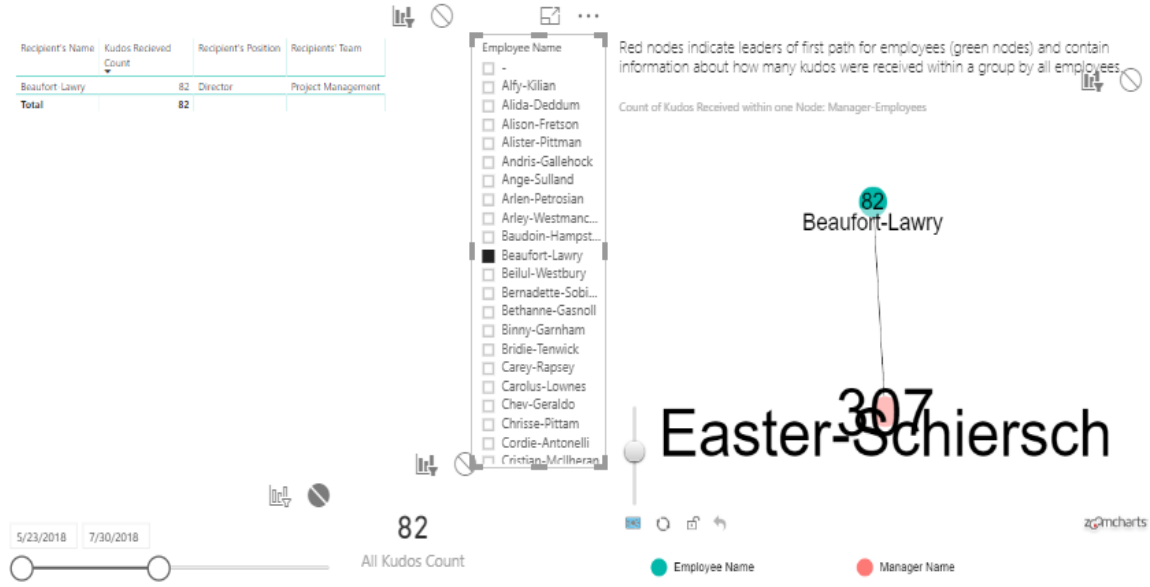
The slicer allowed to check particular employee's number of kudos received (Figure 61). Chart's visual representation allowed to gather important information immediately. For example, from May, 23, to July, 30, Jerri Whatling's team got the most kudos, and from the table it is visible that Jerri Whatling is the Director of Data Science team. The next after Jerri was Kath Banker. When clicking on her node, table was filtered by her team, so all team members, total kudos received, and her team were visible. In this case, her Analytics team got altogether 495 kudos in the period from May, 23, to July, 30 (Figure 62).

Figure 60: Employee-manager relationships from employees' perspective



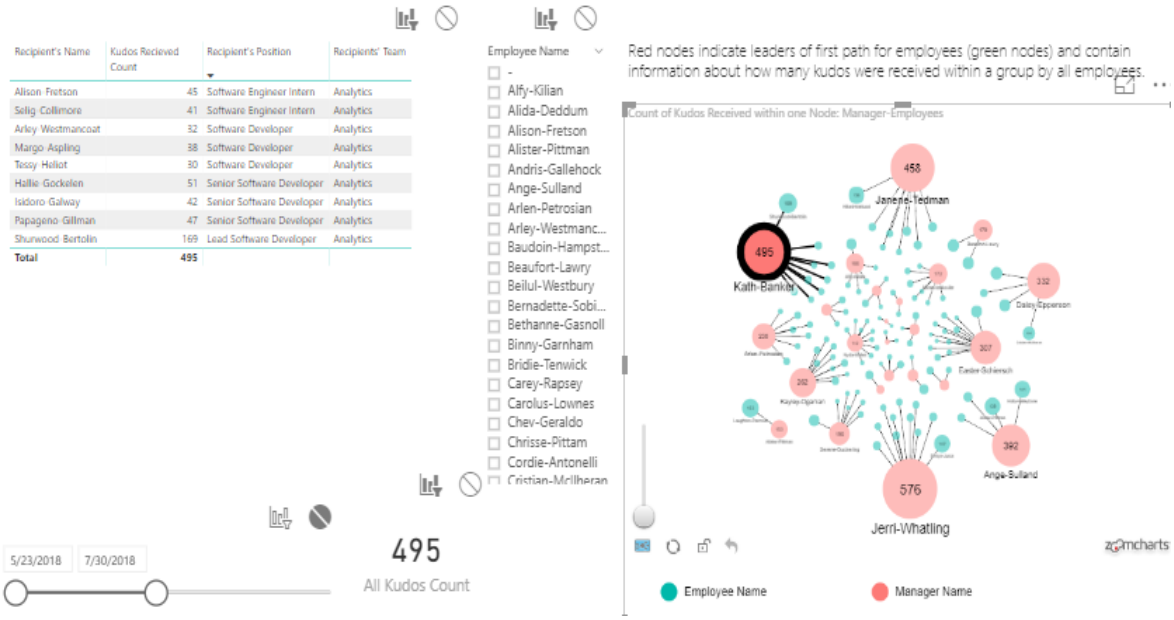
Source: Own work

Figure 61: Kudos received by an employee



Source: Own work

Figure 62: Employee-manager relationships from managerial perspective – team view

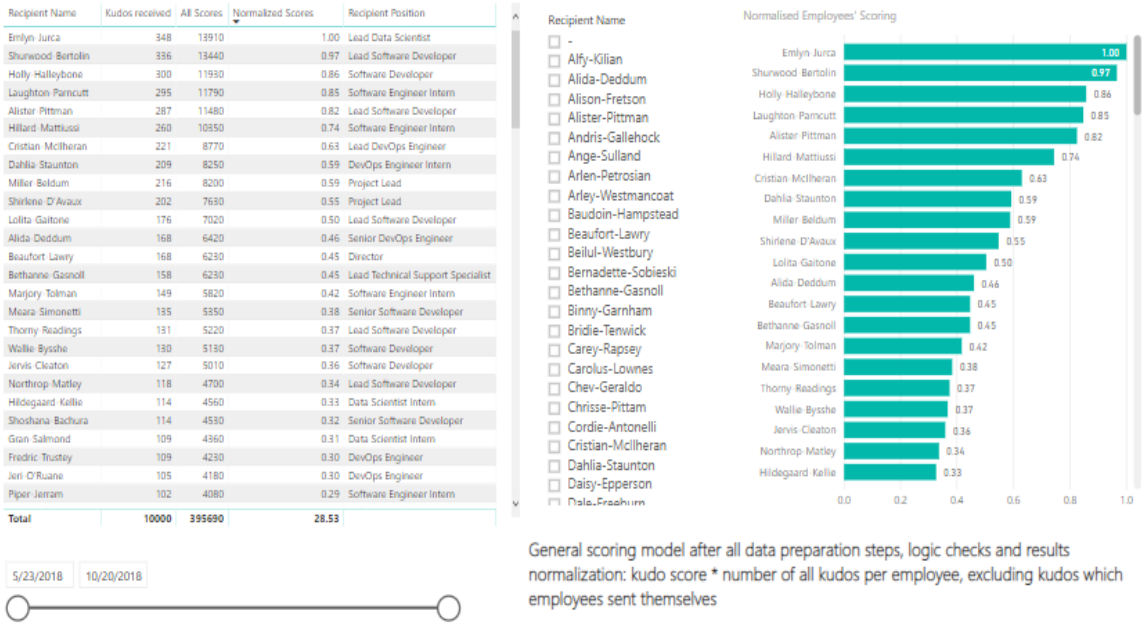


Source: Own work

The fifth slide shows scores and normalized scores over selected period per employee and his/her position with graphical representation in clustered bar chart (Figure 63). As we can see, Emlyn Jurca was the TOP employee as her normalized scoring tended towards 1 (which meant that she was receiving the most kudos, and the majority of kudos received were coming from team leader).

Figure 64 showed communication between the teams. Wider the line was - more kudos was sent to or received by a team. This kind of visualization might be useful when visualising sales data from different sources.

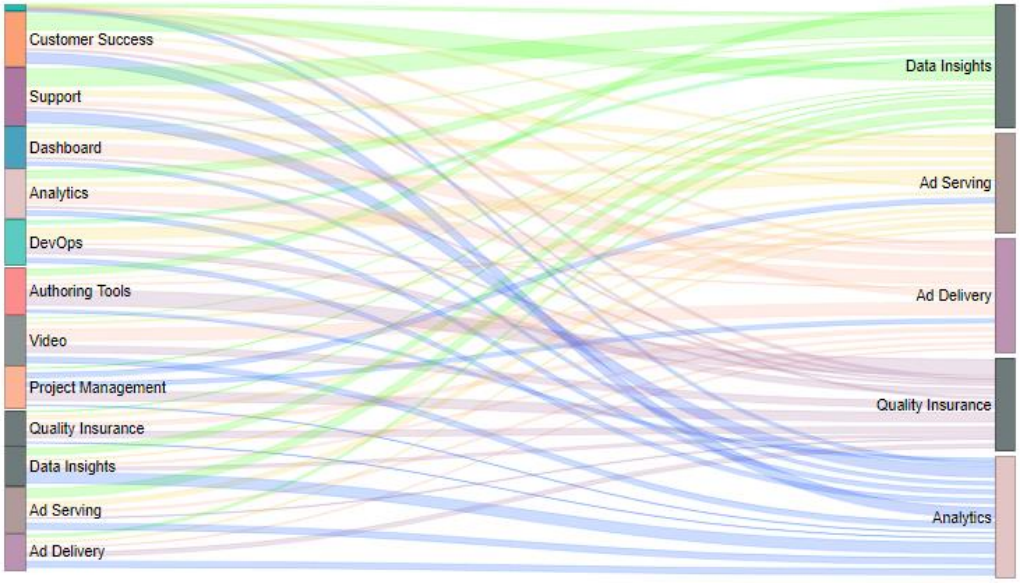
Figure 63: Employees' normalized scoring



Recipient Name	Kudos received	All Scores	Normalized Scores	Recipient Position
Emlyn Jurca	348	13910	1.00	Lead Data Scientist
Shurwood Bertolin	336	13440	0.97	Lead Software Developer
Holly Halleybone	300	11930	0.86	Software Developer
Laughton Parncutt	295	11790	0.85	Software Engineer Intern
Alister Pittman	287	11480	0.82	Lead Software Developer
Hillard Mattiussi	260	10350	0.74	Software Engineer Intern
Cristian McIlheran	221	8770	0.63	Lead DevOps Engineer
Dahlia Staunton	209	8250	0.59	DevOps Engineer Intern
Miller Beldum	216	8200	0.59	Project Lead
Shirlene D'Avaux	202	7630	0.55	Project Lead
Lolita Gaitone	176	7020	0.50	Lead Software Developer
Alida Deddum	168	6420	0.46	Senior DevOps Engineer
Beaufort Lawry	168	6230	0.45	Director
Bethanne Gasnoll	158	6230	0.45	Lead Technical Support Specialist
Marjory Tolman	149	5820	0.42	Software Engineer Intern
Meara Simonetti	135	5350	0.38	Senior Software Developer
Thorny Readings	131	5220	0.37	Lead Software Developer
Wallie Bysshe	130	5130	0.37	Software Developer
Jervis Cleaton	127	5010	0.36	Software Developer
Northrop Matley	118	4700	0.34	Lead Software Developer
Hildegaard Kellie	114	4560	0.33	Data Scientist Intern
Shoshana Bachura	114	4530	0.32	Senior Software Developer
Gran Salmond	109	4360	0.31	Data Scientist Intern
Fredric Trustey	109	4230	0.30	DevOps Engineer
Jeri O'Ruane	105	4180	0.30	DevOps Engineer
Piper Jerram	102	4080	0.29	Software Engineer Intern
Total	10000	395690	28.53	

Source: Own work

Figure 64: Communication between teams (team-kudos-sender – team-receiver)



Source: Own work

7.5.3. Looking for employee's scoring rules

In order to find patterns that lead an employee to a higher score, ANOVA and decision trees were used.

Hypotheses which were tested:

H0: There is no relationship between employee's rating and the role of a person in company by whom the kudo was sent to the employee

H1: There is relationship between employee's rating and the role of a person in company by whom the kudo was sent to the employee

The assumptions were tested using ANOVA which might be a convenient technique as the analysis to be run bases on continuous numeri dependent and categorical independent variable (Bergen, n.d.). ANOVA results are shown in the Figure 65.

Figure 65: ANOVA for employee's scoring results

Response: employee_rating

	Df	Sum Sq	Mean Sq	Sq F value	Pr(>F)
FromLeader	1	4.3356	4.3356	61102	< 2.2e-16 ***
Residuals	117	0.0083	0.0001		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Source: Own work

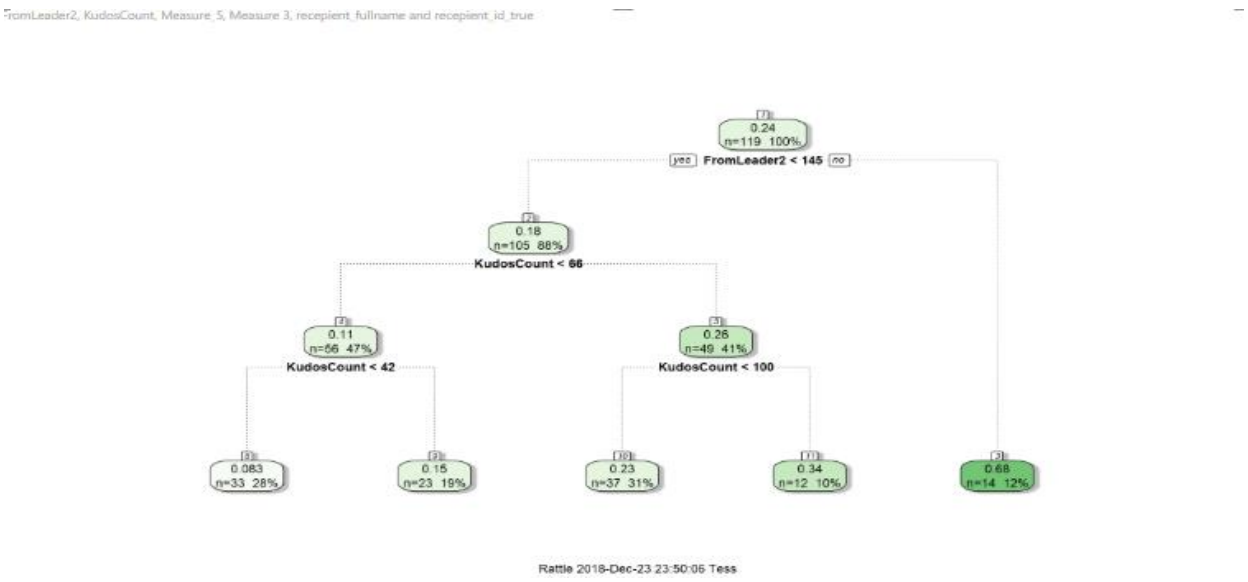
Decision trees are one of the most common and powerful descriptive and predictive analytics techniques. For marketing purposes, decision trees are used to track products and services offered by the competitors as it identifies the best combination of products and marketing channels that target specific sets of consumers, and then used for direct marketing. They are also used to improve customer retention rate by providing good quality products, discounts, and gift vouchers. These can also analyze buying behaviors of the customers and know their satisfaction levels.

At significance level of 0, H0 can be rejected in favor of H1, meaning that there is relationship between employee's rating and the role of a person in company by whom the kudo was sent to the employee.

The decision tree (Figure 66) shows dependence of employee's rating on number of kudos received and on number of kudos received from a leader. Root nodes represent average of normalized scoring. The upper number is the predicted employee rating; the lower 2 numbers show population in particular branch and it percentage of grand total.

Taking into account regression data calculated previously, it becomes clear that without leaders' contribution employees' scoring will be very low. It also indicates the imperfection of elaborated scoring system - which might work better if data were not mocked. Otherwise, frequency distribution, regression and decision trees could help to establish fair business KPIs and evaluate employees' performance indicators.

Figure 66: Employee's rating dependency on number of kudos received from a manager



Source: Own work

This case shows how 'People' element of marketing mux can be analyzed using various tools. There are many opportunities for interactive internal employees' leaderboard usage, but sample models shown in this case could be used for other marketing mix activities as well.

7.6. Physical Evidence

Physical evidence helps to build initial consumer trust. It may refer to buildings, logos, but also to company's digital presence.

There are many services which allow to track users' behavior in the apps and on the websites, such as Google Analytics or Mixpanel. The basic metrics like bounce rate, session time, target actions completion, users' demographic and geographic metrics, average time spent in each app block are present in such services. In case more complex analysis is needed, there are different ways to connect them to PowerBI, Python, R Studio, and other data mining tools. Then, any of techniques discussed earlier could be handy.

One typical example is to extract customers' data, information about the exit page (the page where customer finishes his session), information about sources from which a user has come to this page, and users' path to this page. As Brys (2017) shows in his book, after extracting data from Google Analytics, basic statistics (mean, average…) for page visits or other metrics, could be calculated, or a heat map of average session duration by day and hour. Then, finding the key influencers, the cause can be found, as well, or visitors' segmentation performed (Brys, 2017), or even seasonality detected (Edmondson & Wilson, n.d.).

Data mining techniques could help to identify the key influencer of bad user experience which will help marketer to find the origin of the problem. Usability design and constant testing of every website feature is a must-conduct activity for healthy physical evidence: all buttons should work, the way of how to use website should be as simple as possible, the audience brought to the website should be verified target audience, etc.

7.7. Process: analysis of BPMS data based on Flexkeeping example

Marketing processes benefit from business process management at strategic, tactical and operational planning levels (Meek & Chartered Institute of Marketing., 2006). Process analytics is needed to find bottlenecks and optimize the way company operates, helping to re-think resource allocation, streamline all processes, reduce spending and become more competitive.

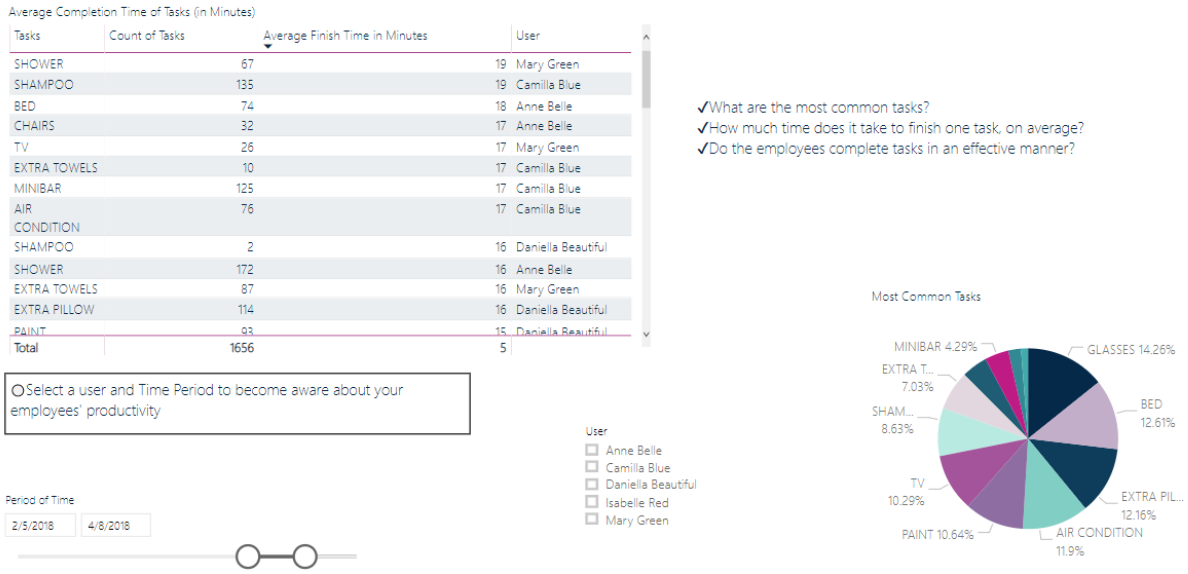
Business process analysis involves diagnosing the strengths and weaknesses of existing processes against requirements of the new process, while customer requirements and delivery of customer value are prioritized in this analysis (Meek & Chartered Institute of Marketing., 2006). Constant business process monitoring and analysis helps during business process redesign, and then – implementation and review. The following case is composed with the mock data and shows a small part of hotel process analytics done by Flexkeeping company for its clients.

Flexkeeping is a hotel management software which, among other services, offers in-depth analytics for its clients. As this data is really big, there is a necessity to find a way of creating customizable reports for each of the client in such a way that each client can adjust the report to the particular hotel needs without seeing other hotel's data. It is also important that such reports are very easy to interpret and use. Traditionally, Excel was used for each particular report, however, this solution was very time-consuming and had many limitations (such as, Excel cannot hold really big data; it wasn't interactive or scalable). As the new solution, Power BI was chosen.

The first step was to connect MySQL database with Power BI. Next, the relationships between all tables were verified in such a way that every table is also translated to other languages which is crucial for multi-languages platform. Currently the company uses several data schemas with multiple fact tables. Hotel process management reports are created with DAX language and Power BI visualizations. Each report is structured in such a way that clients can choose variables of interest by themselves.

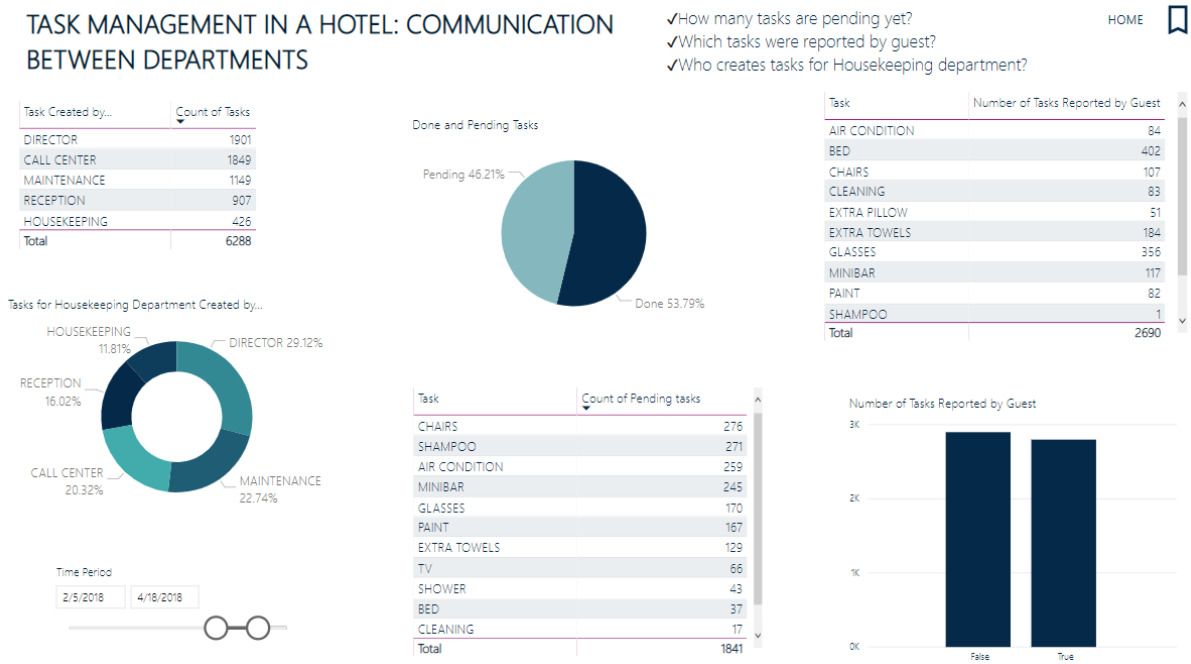
Hotel management is aware of average kitchen and room service reaction time, average cleaning time per different room status, current minibar consumption, the most popular maintenance issues reported by guests, and many other metrics important for hoteliers. Through reports, hoteliers receive meaningful data insights and use them to improve hotel services, optimize operations and increase guest satisfaction (Figures 67-72).

Figure 67: Task management in a hotel



Source: Own work

Figure 68: Communication between departments



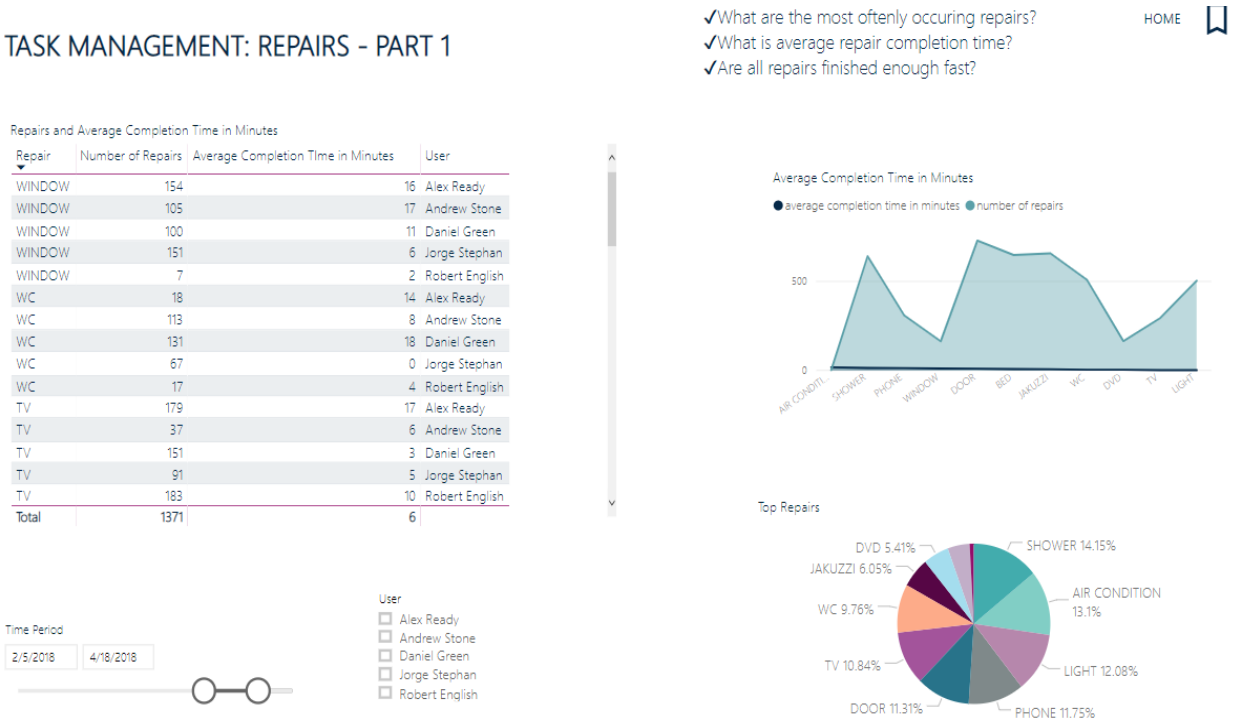
Source: Own work

Currently such reports are embedded into Flexkeeping web application on managerial and operational levels, offering the following:

- Single or multi-property levels
- Multi-property comparative KPI analytics
- Departmental analytics

- Analytics of staff performance, productivity, and efficiency (per department and employee)
- Task management analytics
- Repair management analytics
- Quality control and quality audit reporting system
- Cost of Quality reporting system
- Guest satisfaction reporting system
- Lost & Found stats and overview

Figure 69: Maintenance segment in a hotel – part 1

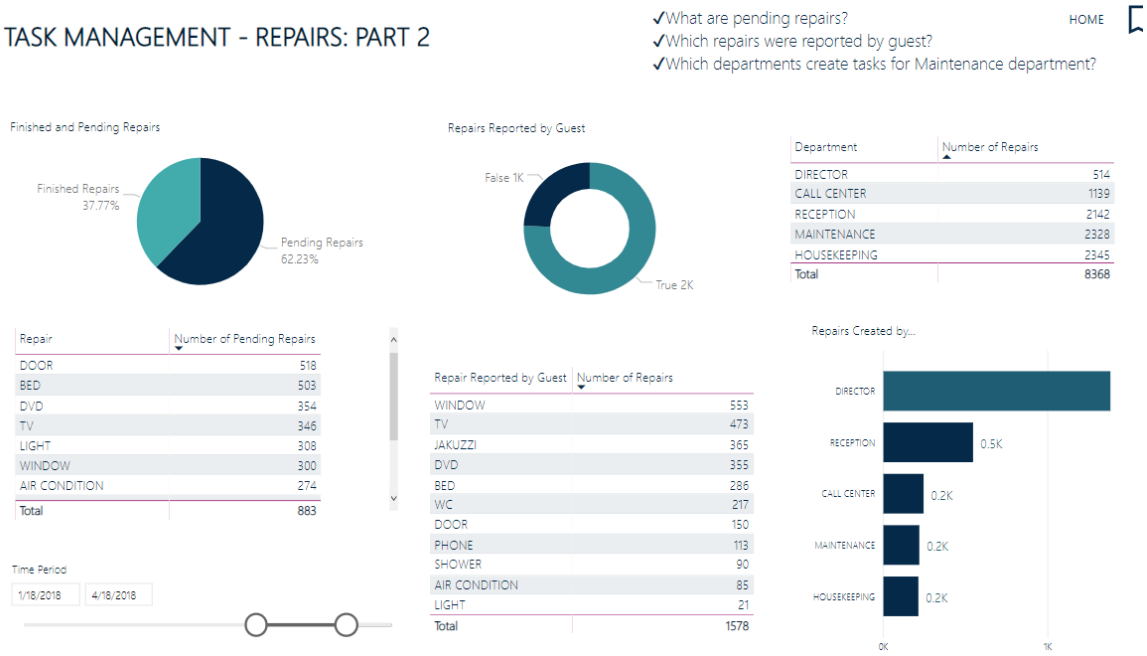


Source: Own work

From Flexkeeping side as a software supplier, it is important to understand how customers are using the app. For example, if maid's cleaning time is 5 seconds, or if there is negative difference between timestamp when task became assigned till it became completed, it might indicate inconsistent application usage.

In this case, company's UX and support teams take care about the matter to collect customers' opinions about app usability, and then new improved design is implemented. Such attitude helps Flexkeeping to have high customers' retention and satisfaction rate. Working closely with customers' opinions also allows Flexkeeping to position themselves as one of the best hotel operations software which covers hoteliers' real needs.

Figure 70: Maintenance segment in a hotel – part 2

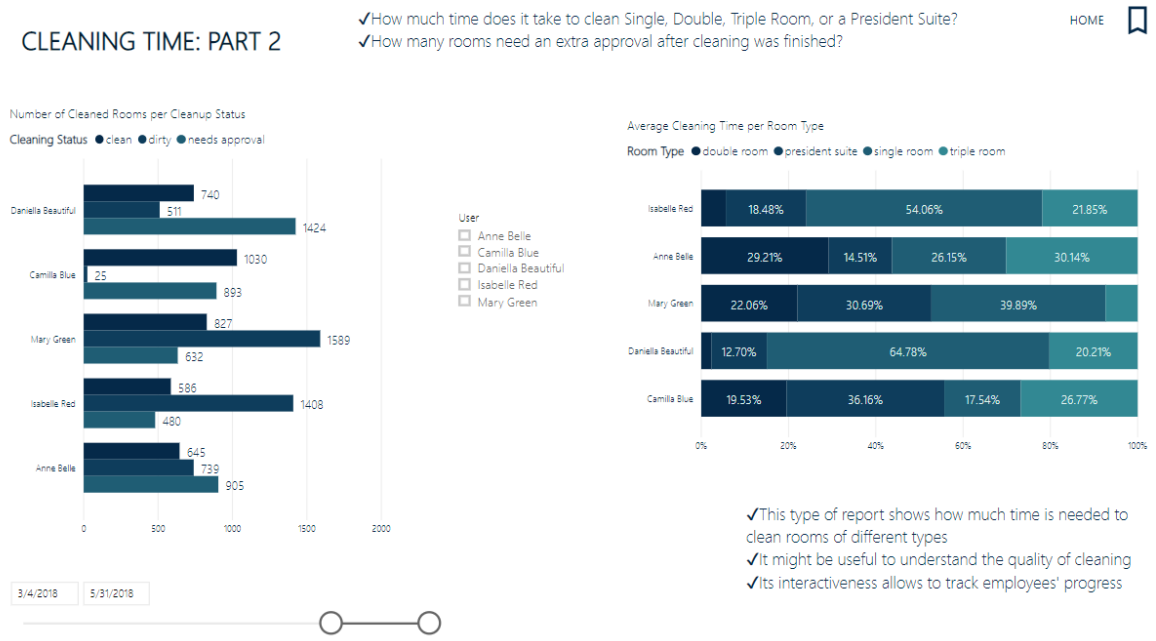


Source: Own work

Figure 71: Cleaning time – part 1

Source: Own work

Figure 72: Cleaning time – part 2



Source: Own work

Sample interactive dashboard is available by the following link: https://app.powerbi.com/view?r=eyJrIjoiNDA3M2Q3YTgtMjUyOC00NzI3LWI5MDItNWY0Y2VkNmJlMmI1IiwidCI6ImE2Y2M5MGRmLWY1ODAtNDlkYy05MDNmLTg3YWY1YTc1MzM4ZSIsImMiOjh9

8. Implementation and control

Controlling and adjusting the results is a big part of marketing planning, ignoring which leads to ineffective decisions and money losses. It is important to control all parts of marketing mix to change them dynamically. Different CRM systems such as Hubspot allow to store useful data for customer analytics; Google Analytics and internal social media analytics help to monitor and adjust online marketing campaigns; Mixpanel provides means for product analytics, and different business intelligence tools and computing environments (such as Power BI, R Studio, Python, Redash, etc.) connected to databases are great for complex price, people, processes, sales, and other parts of marketing mix, analytics.

Implementation and control stage is closely related to the marketing intelligence concept discussed in the chapters 1-3 of the master thesis. Every experience brings insights, and to make those insights improving future marketing performance, it is important to convert them into knowledge, which will enter MkIS and re-used for further data mining.

In some cases, monitoring, controlling and adjusting results becomes a core part of the business. For example, Zemanta, which started as Slovenian start-up and has recently become a part of Outbrain Inc., offers unified campaign tracking, impression-level decision-making abilities and integrations with first and third party data platforms to help companies

meet their native advertising objectives (Zemanta, n.d.). Using a complex mix of machine learning algorithms, this platform analyzes more than 200+ attributes of ads every nanosecond and automatically adjusts ads bids based on every ads impression ability to contribute to business goal; allocates budgets between networks based on ads performance to reduce marketing spending; helps to identify the best performing content at a glance.

Outbrain Inc. focuses on native advertising, e.g. the ads which imitate a part of page user currently is viewing; paid ads that match the look, feel and function of the media format in which they appear (Outbrain.com, n.d.). Every month Outbrain Inc. generates more than 275-billion content recommendations which appear as native placements on premium publishers like CNN, Ha'aretz, Le Parisien, and many others. Recommended articles appear below the article which a visitor has just read provided by Outbrain's AI-powered content discovery platform.

Together Outbrain Inc. and Zemanta operate in the field of programmatic advertising, defined as the automated buying and selling of online advertising space (Outbrain, n.d.). Constant monitoring, controlling and adjusting the variable of interest (content unit, bids) form the key value which companies offer to their customers.

For SMEs, same as for large companies, monitoring, adjusting and controlling lead not only to optimal marketing spending but also to higher customer satisfaction through proposed content of higher quality and value.

In the chapter 7, Flexkeeping company case was discussed. That case is based on OLAP methodology which is widely used in companies of all sizes both for internal and external (like Flexkeeping) purposes. OLAP (Online Analytic Processing) is a complementary subject to data mining, related to data warehousing concept (Linoff & Berry, 2011).

Data warehousing is the process of bringing together disparate data from throughout an organization for decision support purposes (Linoff & Berry, 2011). The data which enters data warehouse is classified by the authors, from bottom-up, as:

- Operational/Transaction data
- Operational summary data
- Decision-support summary data
- Schema
- Metadata
- Business rules which describe why relationships exist and how they are applied

This data is then used for further analysis with the means of different tools, one of which is OLAP. Earlier-mentioned and below-mentioned dashboards are examples of OLAP, which is a powerful way to distribute both summarized and detailed information among users in easy-understandable format. Such Business Intelligence tools as Microsoft Power BI offer in-build data mining techniques as automatic cluster detection, descriptive analytics,

correlations analysis, and also offer integration with such analytical tools as R Studio and Python where complex analysis can be conducted through the techniques discussed earlier, and then presented in a nice visual way, targeting broad, not only technical fields related, audiences.

It works the other way around, as well. Such techniques as decision trees could help to improve the dimensions of OLAP, themselves. While conducting analysis, it will become clear what data is missing in the data warehouse and how OLAP should be organized for the best performance.

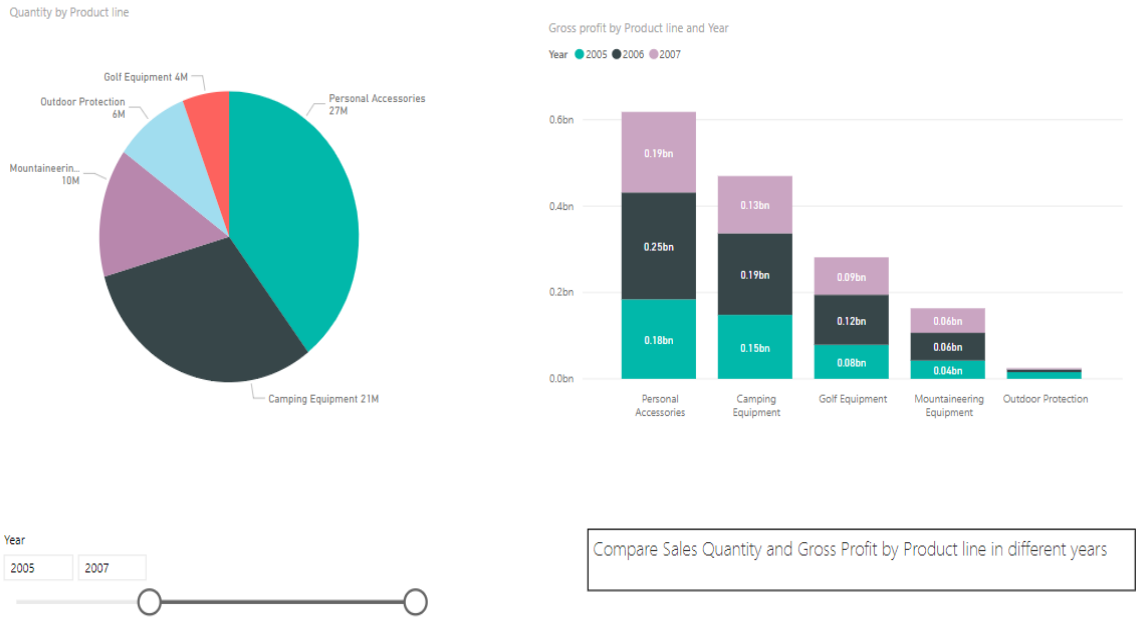
This case is based on open-source data taken from IBM resources (IBM, 2017). Its purpose is to show how dashboard with key business performance indicators, such as gross profit, sales amount, and costs, might look like for a large international company.

Figure 73: Gross profit and costs by product line and country

Source: Own work

Data consists of a table containing such attributes as gross profit, unit price, unit cost, product line, quantity, revenue, year, and others. In the Figure 73, average product line cost against product line profit is shown across different countries. For example, Camping Equipment in Canada made average gross profit of 7444 $ at an average product cost of 132,745 during 2 years from 2005 to 2007.

Figure 74: Sales quantity and gross profit by product line



Source: Own work

Figure 74 summarizes gross profit by product line and year, and quantity by product line across the time. Such simple way of visual information representation helps stakeholders to be aware of all processes in the company at any moment of time and take actions on time, avoiding big losses, adapting and optimizing current company's strategies.

Interactive dashboard is available by the following link: https://app.powerbi.com/view?r=eyJrIjoiMjBiMTI0MTMtYzI0My00ZDJlLTgxZDktNTAxNzFhZTM3MTNkIiwidCI6ImE2Y2M5MGRmLWY1ODAtNDlkYy05MDNmLTg3YWY1YTc1MzM4ZSIsImMiOjh9

9. FINDINGS

Initially, the planned output produced by this master thesis would be presented as Table 10, a shorten yet corrected and deeper form of the Table 2, which finalizes the first chapters of this master thesis and refines data mining techniques by associated marketing research stages, data mining process stages, data collection, and marketing tasks. After the research work and practical experiments were completed, it became understandable that such classification is not enough to have the big picture of when to use particular technique because data mining and marketing became much more holistic.

Data mining process, as marketing research and marketing planning, are continuous tasks which are not done only once but rather form a part of company's culture. For each of marketing planning process stages, depending of the reasons discussed later in this part,

multiple types of research and, thus, different data mining techniques could be applied. Data mining techniques can be integrated into all marketing-related activities. Thus, a more generic logic of summarizing knowledge about data mining techniques applications for marketing was desirable to find. As a result, Table 11 was produced and discussed in details below.

The decision of what technique to use depends not only of the nature of marketing planning process stage but also on:

- The nature of data input:
    - its source (every company collects data using its own business logic which might have its own limitations);
    - its type: structured or unstructured data
    - pre-processing effort associated (it was discussed in detail in related section of chapter 3)
    - data amount
    - data date range (is it a single data point, or time series data)
- The business environment in which company is operating:
    - The nature of product for which marketing is needed
    - Product's users whose actions affect the data collected
        - For example, in Flexkeeping's cases, the product and analytics users are hoteliers who are a very special type of people. They not only need simple but deep metrics at the same time – the metrics should be represented in such a format that a hotel manager can easily consume those reports while drinking his morning coffee and gain meaningful insights about his business
- Marketing task formulation differs from company to company: for example, by 'market segmentation' different companies can mean different things, and product analytics will depend on the product itself
- The desired output:
    - Is analytics part of the company's value proposition to the customers, or it will be used for internal purposes
    - Should output be reproducible, and if yes, how often and how automated it should be (will it be streaming analytics dashboard or historical dashboard)
    - Who will be the users of the data: how the data should be presented
- Company's resources
    - Computational power
    - Human resources
    - Time constraints
- And other factors.

Almost any of existing data mining technique can be applied at each of marketing planning process stages. The choice of the proper technique also depends on the business environment, e.g. how the problem is formulated and how it is applicable taking into account the current situation on macro and micro levels. The same marketing task, e.g. audience segmentation, will be tackled differently depending on the marketing environment, the influencers discussed above, but also the way how the problem is stated, the main goal (on which the type of research depends – is this needed for initial audience understanding, or there is a new marketing campaign being planned, and segments should have clear characteristics?) which business wants to achieve, its purposes, etc.

However, some techniques still appear to be more useful than others for particular marketing planning process stages, thus being summarized into a shorten version of the table from the chapter 4. Table 10 can be a good point to start from when designing the process of solving particular marketing problem.

At the same time, understanding what kind of data mining tasks are associated with a particular marketing task from a marketing planning process stage for this particular company operating in a particular marketing environment leads to a more generic approach than just refining data mining techniques only by marketing planning process stages. As a product of this approach, Table 11 was derived based on literature mentioned in the reference list section, case studies, and own experiments shown in the practical part of this master thesis.

This table is more generic (and, thus, useful for a broad audience) and refines data mining tasks, taking into account their specifics, by possible data mining techniques, in addition providing examples of how a typical business task for which a particular technique could be used, could be formulated. This table goes a bit more into details regarding each technique specialties, providing the reader with a generic hint of which data mining techniques (e.g. KNN) from a particular data mining techniques group (classification/clustering) could be more suitable for his business problem.

Table 11 represents at a glance the most popular data mining techniques, based on communities and blog pages discussions, discussions with experts from different companies, own experiments, summer schools, hackathons, online competitions, and academic literature.

Tables 10 and 11 finalize this master thesis summarizing the cumulative knowledge about data mining for marketing into at-a-glance, visual format, making them useful both for business and technical people. They are also the main contribution which this master thesis does to the data science community.

Table 10: Summary of data mining techniques grouped by marketing planning tasks – refined

Situation Analysis	**Marketing strategy**	**Marketing mix**	**Implementation and control**
• Web and social media scrapping • data summarization • initial visualizations for frequency distribution examination (boxplots, histograms, scatterplots) • basic statistics representation (describing categorical and continuous variables) • cumulative distribution normality checks • exploratory factor analysis • correlation matrixes • PCA • linear and non-linear regressions • ANOVA • association rules investigation • automatic cluster detection	<u>Target audience definition</u> • Classification and clustering (descriptive and predictive) • distance-based clustering methods (k-means) • Gaussian mixture models (GMM) • hierarchical clustering • naïve Bayesian algorithms • decision trees • artificial neural networks • web and social media scrapping + sentiment analysis • random forest • link analysis • SVM <u>Measurable goals setting</u> • time series analysis • decision trees • descriptive analytics <u>Budget developing</u> • Regressions • decision trees • memory-based reasoning • association rules	<u>Product analytics</u> • Descriptive analytics techniques • Link analysis • Association rules • Decision trees <u>Price, Place, Promotion</u> • naïve Bayesian algorithms • decision trees • artificial neural networks • random forest • time series analysis • regressions • memory-based reasoning • choice modelling • rule induction • SVM	• streaming analytics • OLAP • Recommendation systems

Source: Own work

Table 11: Data mining algorithms per data mining task

Categories (yes) Are classes pre-defined? (no)		Predicting numeric value	
Classification	Clustering	Typical tasks	Techniques
How similar is/will be a value to one of the classes • Response modelling (yes/no question answer): SVM, logistic regression • Naïve Bayes • Distance-based (similarity models): KNN (typical task: find the best channels to advertise; predict to which category customer will fall) • Logistic regression (the probability of value falling into each class) • Table lookup (typical task: RFM analysis) Personal recommendations, fraud detection, customer response: memory-based reasoning Conditions-based (clear rules-based classification): decision trees, random forest	Text data • Memory-based reasoning (typical task: classifying customer complaints) • Text mining: topic modelling, sentiment analysis • Swarm intelligence, genetic algorithms Non-text data • K-means clustering • GMM Hierarchical clustering (good for hierarchical data; output = tree-like looking model) Rules generation - associations (typical task: market basket analysis, market segmentation, seasonality detection, recommendation systems building)	• Customer LTV • Churn rate prediction • Number of target actions (no. of clicks, forms fulfilled, etc.) • Target metric prediction (CPC, CTR…)	• Different types of regression • Neural networks • Factorization machines • Decision trees and trees-based algorithms (XGBoost, Random Forest…)

(table continues)

(continued)

Table 11: Data mining algorithms per data mining task

Categories (yes) Are classes pre-defined? (no)		Predicting numeric value	
Classification	Clustering	Typical tasks	Techniques
Features selection (finding meaningful interactions between variables): Factorization Machines (typical task: CTR prediction) Working with complex categorical data: Factorization Machines, CatBoost Working with complex numeric data: Gradient Boosting algorithms (XGBoost), Neural networks Associated supportive tasks • Binning • Feature selection • Correlation between attributes • Outliers detection	Link analysis (typical task: finding opinion leaders, 'matched' on dating websites) Associated supportive tasks: optimization (typical tasks: budgeting, resources allocation): swarm intelligence, genetic algorithms		

Source: Own work

CONCLUSION

The purpose of this master thesis, as it was stated in the introduction, was to (1) conduct a generic research, and (2) to bring together technical and business knowledge, covering data mining techniques implementation for different marketing planning process stages both from managerial and technical perspectives. The first part of the purpose was completed with the first four parts of this master thesis. To complete the second part of the purpose, chapters five to eight were introduced, where each chapter has covered one of marketing planning process stage. Chapters 5-8 are more practice-oriented, and have also completed the following goals of this master thesis:

- To evaluate different approaches and data mining techniques through analyzing existing case studies, literature and own experiments with the data;
- To show practical examples of data mining techniques used for marketing purposes;
- To summarize current findings about data mining techniques for digital marketing purposes.

Different approaches were evaluated by conducting cases for each step of marketing planning process.

As a result, a shorten, more useful version of the table from chapter 4, was created (Table 10). The table summarizes the literature overview by refining data mining techniques by marketing planning process stages and by marketing tasks associated with each of marketing planning process stages. The aim of the table is to provide the basis with which it is possible to start digging into each technique, knowing what kind of marketing task is planned to be solved. Its difference from the table from chapter 4 is that it only shows marketing planning process stages and possible data mining techniques to be used on each of them, without any additional information, and its aim is to serve as a cheat-sheet for the interested parties.

Additionally, a table which refines data mining algorithms per data mining task with typical marketing tasks, was created (Table 11), therefore completing the last goal of this master thesis, which is formulated as ‘to summarize current findings about data mining techniques for digital marketing purposes’. In difference of the shorten version of the conclusive table from chapter 4, this table drills down data mining techniques by data mining tasks, while possible marketing tasks to be solved come as the refining criteria of the second level. The reason for this is that, after the research work was completed, it is possible to conclude that the majority of data mining techniques can come handy at any of marketing planning process stage.

From the data perspective, the main limitation was that data used was either created for academic purposes or taken from data mining competitions, meaning that it sometimes was

mocked, sometimes – not realistic, but still suitable for showcase purposes. Also, there is a lot of room for improvement regarding data preparation step.

From the techniques perspective, due to showcase goal (to show as many techniques as possible), not all possible approaches were tested. Also, the techniques used are not the ultimate ones. Sometimes, the techniques shown were not properly validated in terms of their performance: for some cases, other techniques might have worked better, but for showcase purposes the goal was to show as many techniques as possible.

Some techniques were impossible to showcase because of technical difficulties. The main tool used was R Studio which is mostly community-driven (the packages are written by the community members and not by the software providers). Thus, some packages became obsolete with no substitution for this language. Another languages and environments (Python, Hadoop in the pair with Git) could be used for algorithms optimization and big data processing, however, such tools are out of scope of this particular master thesis. In some cases, the hardware used was not able to process some algorithms and/or the amounts of data needed to train the algorithm. Some techniques became obsolete and were substituted by better-performing models, which, however, also require much more computational capacities. For web-related methods of data collection (social media and web scarping), GDPR now should be taken into account (which was introduced when the work on this master thesis has already started). Particularly, there are API restrictions applied by Facebook in 2018. New rules do not allow free parsing of users' data which becomes a big problem for a student research such as this one. However, there are ways for companies to comply with the new rules.

From the cases perspective, some of examples were guided by other data mining community participants (their cases were taken as the base for new cases development). For some marketing planning process stages, there are no cases, only explanations and possible algorithms as the main part of them was already shown in other cases.

This master thesis might be a good start both for business and technical people and will help them to integrate data mining into their companies' culture. Lately, data mining and machine learning have become a common part of organizational culture. Many processes are moving towards automation; the only question is "by how much it is possible to automate people". As authors state lately in Harvard Business Review Journal (Fountaine et al., 2019), to scale-up AI within a company, three shifts must be made: from soloed work to interdisciplinary collaboration; from experience-based, leader-driven decision making to data-driven decision making at the front line; from rigid and risk-averse to agile, experimental, and adaptable.

The authors of the article also state that "the ways AI can be used to augment decision making keep expanding. New applications will create fundamental and sometimes difficult changes in workflows, roles, and culture, which leaders will need to shepherd their organizations through carefully. Companies that excel at implementing AI throughout the

organization will find themselves at a great advantage in a world where humans and machines working together outperform either humans or machines working on their own". Thus, the future work might be related to digging into machine learning and data mining techniques which will serve as the basis for AI-based decision making.

For now, transforming data insights into knowledge, and knowledge – into value, is still a humans' job. Besides it might change in the future, marketing ecosystem changes together with customer, and any technology used on the way should always put customer in the center of the processes.

REFERENCE LIST

1. AMA. (2013, July). *Definitions of Marketing*. Obtained March 4, 2019 from https://www.ama.org/the-definition-of-marketing/
2. Amazon.com. (no date). *Force1 Mini Drones for Kids*. Obtained May 29, 2019 from https://www.amazon.com/Force1-4000-Mini-Drones-Kids/product-reviews/B07GZTG3HL/ref=cm_cr_dp_d_show_all_btm?ie=UTF8&reviewerType=all_reviews
3. Amazon SageMaker. (no date). *Factorization Machines Algorithm*. Obtained August 14, 2019 from https://docs.aws.amazon.com/sagemaker/latest/dg/fact-machines.html
4. Azzalini, A., Walton, G., & Scarpa, B. (2012). *Data Analysis and Data Mining : An Introduction*. Oxford University Press.
5. Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4
6. Bergen, B. (no date). *Statistical Techniques in Language and Cognition Research*. Obtained August 3, 2019 from http://www.cogsci.ucsd.edu/~bkbergen/lcl/statshowto.html
7. Bhatt, P. (no date). *Robust Factorization Machines - WalmartLabs*. Medium. Obtained August 14, 2019 from https://medium.com/walmartlabs/robust-factorization-machines-1a9ef9f75abf
8. Blattberg, R. C., Kim, B.-D., Neslin, S. A., Kim, P., & Neslin, S. A. (2008). *Database marketing : analyzing and managing customers*. Springer. https://doi.org/10.1007/978-0-387-72579-6_12
9. Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3). Obtained May 30, 2019 from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
10. Bow, C. (2018). *An introduction to Facebook ad analysis using R*. Kaggle. Obtained July 11, 2019 from https://www.kaggle.com/chrisbow/an-introduction-to-facebook-ad-analysis-using-r
11. Bradley, N. (2013). *Marketing research: tools and techniques* (3rd ed.). Oxford.

12. Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Source: Statistical Science*, *16*(3), 199–231. Obtained August 14, 2019 from http://www2.math.uu.se/~thulin/mm/breiman.pdf
13. Brys, M. (2017). *Using Google Analytics with R*. https://michalbrys.gitbooks.io/r-google-analytics/content/chapter4/exploratory_data_analysis.html
14. Chaffey, D., Ellis-Chadwick, F., Mayer, R., & Johnston, K. (2009). *Internet Marketing: Strategy, Implementation and Practice*. Pearson Education.
15. Chaffey, D., & Smith, P. R. (Paul R. (2017). *Digital marketing excellence : planning, optimizing and integrating online marketing* (5th ed.). Routledge.
16. Chapman, C., & McDonnell Feit, E. (2015). *R for Marketing Research and Analytics*. Springer International Publishing. https://doi.org/10.18637/jss.v067.b02
17. Chiu, S., & Tavella, D. (2008). *Data mining and market intelligence for optimal marketing returns*. Butterworth-Heinemann/Elsevier.
18. Correia, J. (2016, July 6). *How RFM Analysis Boosts Sales*. BlastAm. Obtained August 6, 2019 from https://www.blastam.com/blog/rfm-analysis-boosts-sales
19. Danneman, N. (2014). *Social Media Mining with R*. Packt Publishing.
20. Davis, J. (2017). *Measuring Marketing: The 100+ Essential Metrics Every Marketer Needs* (3rd ed.). Walter de Gruyter GmbH & Co KG.
21. De Ville, B. (2001). *Microsoft data mining : integrated business intelligence for e-Commerce and knowledge management*. Digital Press.
22. Dictionaries, O. (no date). *kudos*. Obtained November 18, 2018 from https://en.oxforddictionaries.com/definition/kudos
23. Dourado, R. M. (2017). *How often people rebuy ?* Kaggle. Obtained July 9, 2019 from https://www.kaggle.com/rafaelmdourado/how-often-people-rebuy/notebook
24. Eckerson, W. W. (2011). *Performance dashboards : measuring, monitoring, and managing your business*. Wiley.
25. Edmondson, M., & Wilson, T. (no date). *Digital Analytics: R and Statistics*. http://www.dartistics.com/
26. Facebook Investor Relations. (2017). *Facebook Reports Second Quarter 2017 Results*. Investor.Fb.Com. https://investor.fb.com/investor-news/press-release-details/2017/Facebook-Reports-Second-Quarter-2017-Results/default.aspx
27. Fonseca, Y. (2017, August 27). *Pricing Optimization: How to find the price that maximizes your profit*. R-Bloggers. Obtained August 6, 2019 from https://www.r-bloggers.com/pricing-optimization-how-to-find-the-price-that-maximizes-your-profit/
28. Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-Powered Organization. *Harvard Business Review*.
29. Gartner IT Glossary. (no date). *Product Analytics*. Obtained February 5, 2019 from https://www.gartner.com/it-glossary/product-analytics
30. Ghani, R., & Soares, C. (2010). *Data Mining for Business Applications* (Issue v. 218). IOS Press.
31. Green, E. (2015, September 15). *What is Modern Marketing?* Obtained March 4,

2019 from https://www.oliveandcompany.com/blog/what-is-modern-marketing
32. Hammink, J. (2018, March 7). *The Types of Modern Databases*. Obtained March 12, 2019 from https://www.alooma.com/blog/types-of-modern-databases
33. Harmon, R. (2003). Marketing Information Systems. *Encyclopedia of Information Systems*, *3*(1), 137–151. https://doi.org/10.1016/B0-12-227240-4/00110-6
34. Hebbali, A. (no date). *rfm • Recency, Frequency and Monetary Value Analysis*. Obtained August 6, 2019 from https://rfm.rsquaredacademy.com/
35. Hebbali, A. (2019). *RFM - Customer Level Data*. Obtained August 6, 2019 from https://cran.r-project.org/web/packages/rfm/vignettes/rfm-customer-level-data.html
36. Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070
37. IBM. (2017). *Cognos Analytics - Business Analytics*. Obtained August 1, 2019 from https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2017/06/19/guide-to-ibm-cognos-analytics-sample-data-sets
38. Kaggle. (no date-a). *Sales Conversion Optimization*. Obtained July 11, 2019 from https://www.kaggle.com/loveall/clicks-conversion-tracking
39. Kaggle. (no date-b). *Telco Customer Churn*. Obtained August 17, 2019 from https://www.kaggle.com/blastchar/telco-customer-churn
40. Kaggle. (2017). *Instacart Market Basket Analysis*. Obtained July 9, 2019 from https://www.kaggle.com/c/instacart-market-basket-analysis
41. Kassambara, A. (no date). *Determining The Optimal Number Of Clusters: 3 Must Know Methods*. Datanovia. Obtained July 9, 2019 from https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
42. Kaushik, S. (2016, November 3). *Clustering Introduction and different methods of clustering*. Obtained July 8, 2019 from https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
43. Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, *24*, 994–1012. https://doi.org/10.1016/J.ASOC.2014.08.041
44. Kotler, P. (1997). *Marketing management: Analysis, planning, implementation, and control* (9th ed.). Prentice Hall.
45. Kotler, P. (2000). *Marketing Management, Millenium Edition*. Prentice Hall PTR. https://doi.org/10.1016/0024-6301(90)90145-T
46. Kotler, P., Kartajaya, H., & Setiawan, I. (2017). *Marketing 4.0: Moving from Traditional to Digital*. https://doi.org/10.1515/9783110258394.189
47. Kotler, P., & Keller, K. L. (2012). *Marketing Management, 14th Edition*. Prentice Hall PTR. https://doi.org/10.1080/08911760903022556

48. Kotler, P., & Armstrong, G. (2012). *Principles of Marketing* (14th ed.). Pearson Prentice Hall.
49. Kumar, A. (2018, March 18). *Unsupervised Learning - Market-Basket analysis on e-Commerce dataset*. RStudio Pubs. Obtained April 13, 2020 from https://rstudio-pubs-static.s3.amazonaws.com/370943_cbb1b4f7ad284442843c4120bf7f2c40.html
50. Laursen, G. H. N. (2011). *Business analytics for Sales and Marketing Managers : How to Compete in the Information Age*. John Wiley & Sons.
51. Linoff, G. S., & Berry, M. J. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.). Wiley.
52. Maechler, N., Neher, K., & Park, R. (2016). From touchpoints to journeys: Seeing the world as customers do. *McKinsey Digital*. https://doi.org/http://dx.doi.org/10.1103/PhysRevLett.28.1516
53. Maklin, C. (no date). *Gaussian Mixture Models Clustering Algorithm Explained*. Obtained August 14, 2019 from https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e
54. Malhotra, N. K. (2013). *Review of Marketing Research* (Vol. 10). Emerald Group Publishing Limited.
55. Mankiw, N. G. (2009). *Principles of economics*. South-Western Cengage Learning.
56. McColl, L. (no date). *Market Basket Analysis: Understanding Customer Behaviour*. Obtained July 5, 2019 from https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/
57. McDonald, M. (2007). *Marketing Plans : How to Prepare Them, How to Use Them* (6th ed.). Butterworth-Heinemann.
58. Meek, H., & Chartered Institute of Marketing. (2006). *Managing marketing performance 2006-2007*. Butterworth-Heinemann.
59. Microsoft Power BI. (2016, November 28). *Power BI Desktop November Feature Summary*. Obtained July 11, 2019 from https://powerbi.microsoft.com/en-us/blog/power-bi-desktop-november-feature-summary/
60. Mishra, M. N. (2008). *Modern Marketing Research*. Himalaya Publishing House.
61. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001
62. Nicholson, C. (no date). *Skymind | A Beginner's Guide to Neural Networks and Deep Learning*. Obtained August 14, 2019 from https://skymind.com/wiki/neural-network
63. Ojeda, T., Dasgupta, A., Bengfort, B., & Murphy, S. P. (2014). *Practical Data Science Cookbook*. Packt Publishing.
64. Oracle. (no date). *What Is Business Intelligence?* Obtained March 24, 2020 from https://www.oracle.com/business-analytics/business-intelligence/what-is-business-intelligence.html
65. Outbrain.com. (no date). *Performance-Based Native Advertising Platform*. Obtained August 1, 2019 from https://www.outbrain.com/
66. Outbrain. (no date). *What is Programmatic Advertising and how to Start? | Outbrain*

Blog. Obtained August 1, 2019 from https://www.outbrain.com/blog/programmatic-advertising/

67. Peppers, D., & Rogers, M. (2011). *Managing Customer Relationships: A Strategic Framework* (2nd ed.). Wiley. http://web.b.ebscohost.com.nukweb.nuk.uni-lj.si/ehost/detail/detail?vid=5&sid=da64a0c9-ee2b-4818-9599-8fcba6ce337c%40sessionmgr120&bdata=Jmxhbmc9c2wmc2l0ZT1laG9zdC1saXZl#AN=354167&db=nlebk
68. Proctor, T. (2000). *Strategic Marketing: An Introduction*. Routledge. http://web.b.ebscohost.com.nukweb.nuk.uni-lj.si/ehost/detail/detail?vid=3&sid=da64a0c9-ee2b-4818-9599-8fcba6ce337c%40sessionmgr120&bdata=Jmxhbmc9c2wmc2l0ZT1laG9zdC1saXZl#AN=74775&db=nlebk
69. Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. In *O'Reilly*. O'Reilly. https://doi.org/10.1017/CBO9781107415324.004
70. Quinn, K. (2009). *How the Fastest Growing Companies Use Business Intelligence*. http://www.umsl.edu/~sauterv/DSS4BI/links/pdf/BI/Worstpractices_R4.pdf
71. Ravindran, S. K., & Garg, V. (2015). *Mastering Social Media Mining with R Extract valuable data from social media sites and make better business decisions using R*. Packt Publishing. www.packtpub.com
72. Rendle, S. (2010). Factorization Machines. *2010 IEEE International Conference on Data Mining*, 995–1000. https://doi.org/10.1109/ICDM.2010.127
73. Rsquared Academy Blog. (2019, February 11). *RFM Analysis in R*. Obtained August 6, 2019 from https://www.r-bloggers.com/rfm-analysis-in-r/
74. Russell, M. A. (2011). *Mining the Social Web, Second Edition* (2nd ed.). O'Reilly Media. https://doi.org/10.1017/CBO9781107415324.004
75. Saito, R. (2019a, March 3). *Web Scraping Amazon Reviews in R*. Just R Things. Obtained May 30, 2019 from https://justrthings.com/2019/03/03/web-scraping-amazon-reviews-march-2019/
76. Saito, R. (2019b, March 4). *Sentiment Analysis, Word Embedding, and Topic Modeling on Venom Reviews*. Just R Things. Obtained May 29, 2019 from https://www.r-bloggers.com/sentiment-analysis-word-embedding-and-topic-modeling-on-venom-reviews/
77. Salesforce. (no date). *What is Digital Transformation?* Obtained April 1, 2019 from https://www.salesforce.com/products/platform/what-is-digital-transformation/
78. Seif, G. (no date). *The 5 Clustering Algorithms Data Scientists Need to Know*. Obtained August 14, 2019 from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
79. Sekula, M., Datta, S., & Datta, S. (2017). optCluster: An R Package for Determining the Optimal Clustering Algorithm. *Bioinformation*, *13*(3), 101–103. https://doi.org/10.6026/97320630013101
80. Shankar, V., & Carpenter, G. S. (2012). *Handbook of marketing strategy*. Edward

Elgar Pub.
81. Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, *31*(1), 127–137. https://doi.org/10.1016/S0167-9236(00)00123-8
82. Smartsheet. (no date). *Here's How the Marketing Process Works*. Obtained June 10, 2018 from https://www.smartsheet.com/strategic-marketing-processes-and-planning
83. Spendler, L. I. (2010). *Data Mining and Management*. Nova Science Publishers, Inc.
84. Srivastava, T. (2016, July 4). *Solving Case study : Optimize the Products Price for an Online Vendor (Level : Hard)*. Obtained August 6, 2019 from https://www.analyticsvidhya.com/blog/2016/07/solving-case-study-optimize-products-price-online-vendor-level-hard/
85. Stevens, R. E. (2006). *The marketing research guide*. Best Business Books.
86. Technopedia. (no date). *What is Web Scraping?* Obtained May 26, 2019 from https://www.techopedia.com/definition/5212/web-scraping
87. UCI Machine Learning Repository. (no date). *UCI Machine Learning Repository: Online Retail Data Set*. Obtained April 13, 2020 from http://archive.ics.uci.edu/ml/datasets/Online+Retail
88. Uhl, K. P., & Schoner, B. (1969). *Marketing Research*. Nirali Prakashan.
89. Wang, S.-C., Wang, S.-S., Chang, C.-M., Yan, K.-Q., & Lin, Y.-P. (no date). *Systematic Approach for Digital Marketing Strategy through Data Mining Technology*. Obtained May 31, 2019 from http://www.csroc.org.tw/journal/JOC25-3/JOC25-3-4.pdf
90. Wedel, M., & Kannan, P. K. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413
91. Weisberg, J. (2017). *Visualizing User Histories*. Kaggle. Obtained July 11, 2019 from https://www.kaggle.com/jweisber/visualizing-user-histories/notebook
92. Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, *69*(5), 1562–1566. Obtained April 23, 2018 from https://www-sciencedirect-com.nukweb.nuk.uni-lj.si/science/article/pii/S0148296315004403
93. Yiu, T. (2019). *Understanding Random Forest*. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
94. Zaki, M. J., & Meira, W. J. (2014). *Data Mining and Analysis: fundamental concepts and algorithms.* https://doi.org/10.1371/journal.pone.0058904
95. Zemanta. (no date). *Platform*. Obtained August 1, 2019 from http://www.zemanta.com/platform/
96. Zhang, Y. (no date). *Pricing Analysis*. RPubs. Obtained August 6, 2019 from http://www.rpubs.com/angelayy/185880

APPENDICES

Appendix 1: Povzetek (Summary in Slovene language)

Trženje je koncept, ki se je z leti spreminjal. Trg danes in trg pred dvajsetimi leti se med seboj bistveno razlikujeta. Eden od vzrokov za to je pojav spletnih tehnologij v vsakdanjem življenju in močna usmerjenost v digitalizacijo. Podatki, ki jih s tem v zvezi obvladuje podjetje, so ključni element trženjskega informacijskega sistema . Zbrani in organizirani so v podatkovnih bazah, ki ji lahko povezujemo in analiziramo. Tehnike raziskovanja trga so bile že prej tradicionalno uporabljene za te namene, vendar so količina podatkov, njihova globina, vrste in načini shranjevanja omogočili nove načinih ekstrakcije pomembnih informacij. Danes, v dobi usmerjenosti k potrošnikom, je možnost vedno bolj podrobnega razumevanja podatkov na individualnem nivoju izjemno pomembna.

Obstaja veliko knjig in člankov o rudarjenju podatkov, analitiki masovnih podatkov, rudarjenju podatkov za namene trženja in managementa odnosov z odjemalci. Po drugi strani, le njihov manjši del povezuje teorijo s prakso in z novejšimi tehnikami podatkovnega rudarjenja na različnih stopnjah procesa načrtovanja trženja. Tako je namen tega magistrskega dela izvedba generične raziskave in povezava tehničnega in poslovnega znanja na področju podatkovnega rudarjenja na različnih stopnjah procesa načrtovanja trženja z obeh vidikov, tehničnega in poslovnega.

Preiskovalna analiza je ena najbolj priljubljenih metod pridobivanja prvega vtisa o podatkih in za konceptualizacijo idej. Prav tako so bili uporabljeni lastni eksperimenti so bili prav tako uporabljeni, saj je njihov cilj odgovoriti na najbolj pogosta poslovna vprašanja. Podatki za eksperimente so pridobljeni iz odprtih virov oziroma so bili zbrani ali pripravljeni za namen tega dela. Obdelani in vizualizirani so bili z uporabo tipičnih orodij za podatkovno rudarjenje in sicer R studio, RapidMiner in Power BI.

Prva inačica načrtovanega rezultata tega magistrskega dela, je tabela, ki je nastala na podlagi študija literature in povezuje tehnike podatkovnega rudarjenja, ustrezne stopnje trženjskega načrtovanja, faze procesa podatkovnega rudarjenja, vire podatkov in naloge trženja. Raziskovalno delo je pokazalo, da taka klasifikacija ne zagotavlja celotne slike oz. ustreznih usmeritev, kdaj se naj uporabi posamezna tehnika, saj sta trženje in podatkovno rudarjenje postali veliko bolj holistični.

Za vsako stopnjo v procesu trženjskega načrtovanja je možno uporabiti več različnih načinov raziskav in tehnik podatkovnega rudarjenja. Te tehnike je mogoče vključiti v vse postopke, ki so povezani s trženjem. Zato je kot končni rezultat tega magistrskega dela nastal bolj splošen pregled možnosti uporabe podatkovnega rudarjenja oz. tehnik podatkovnega rudarjenja v okviru trženjskega načrtovanja.

Ena od ključnih omejitev tega dela je, da so uporabljeni podatki v nekaterih primerih namenoma umetno generirani ali nerealistični in je zato še veliko prostora za izboljšanje v razumevanju koraka priprave podatkov za analizo. Poleg tega niso preizkušene vse metode, ki so na voljo. V nekaterih primerih ni bila ovrednotena učinkovitost metod.