UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

BUSINESS ANALYTICS IN SMALL COMPANIES: THE CASE OF A SELECTED COMPANY

Ljubljana, September 2021 IULIANA MASLENNICOV

AUTHORSHIP STATEMENT

The undersigned Iuliana Maslennicov, 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 "Business Analytics in small companies: the case of a selected company", prepared under supervision of Jurij Jaklic.

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 the 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, September 21, 2021 Author's signature: 

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

CEO - Chief Executive Officer

COGS - Cost Of Goods Sold

DSS - Decision Support System

IT - Information Technology

NPS - Net Promoter Score

OECD - Organisation for Economic Cooperation and Development

ROI - Return On Investment

SBE - Small Business Enterprise

SME - Small and Medium-sized Enterprise

VADER - Valence Aware Dictionary for Sentiment Reasoning

INTRODUCTION

The history of the term Business Analytics has its roots in the distant 19th century when Frederick Winslow Taylor published his theory of Scientific Management that analyzes and synthesizes different workflows. In the 1960s, computers started providing support in the decision-making processes, therefore, the term Business Analytics started receiving more attention and evolving more significantly (Foote, 2018).

Nowadays, there are a variety of definitions proposed for Business Analytics, however, all of them have a common essence. It is the consolidation of all the supporting mechanisms that help to transform the data into a valuable piece of information that will improve and accelerate the decision-making and problem-solving processes. It involves mathematics, statistics, machine learning, and professional knowledge to discover various data insights (Delen & Ram, 2018).

Business Analytics is based on the four main types of data analytics, according to Gartner's Analytics Ascendancy Model (Tamm, Seddon & Shanks, 2013). The first type is called Descriptive analytics, which is used to describe the historical and existing data (Whitelock, 2018). According to the model, descriptive analytics has the lowest level of difficulty and provides the smallest value among all four types (Eriksson, Bigi & Bonera, 2020). Another type is Diagnostic analytics, it focuses on the reasons for the problem occurrences. It helps managers to adjust the company operations and improve the situation. The third type of analytics described by Gartner is Predictive analytics, where the main focus is forecasting and predicting trends and probabilities that could happen in the future, which is useful for preparing what-if analysis. Finally, the most valuable and difficult to implement form of data analytics is Prescriptive analytics, which provides the possible decisions to maximize good outcomes and minimize the number of bad outcomes (Whitelock, 2018).

There are plenty of reasons why different companies around the world decide to implement analytical tools in their businesses. First of all, organizations are struggling with managing huge amounts of data while various business intelligence and analytical tools provide access to automated data collection systems. These systems make data management easier and benefit the company processes (Delen & Ram, 2018). Secondly, the data should be analyzed and transformed into information, while this information needs to be converted into knowledge, knowledge into insight, and then into an action that leads to better decision-making and improved performance. This is a complex and time-consuming process that also requires plenty of specialists, while business analytical tools make it way faster and easier. (Whitelock, 2018). The third reason is the cultural tendency to change towards evidence-based management (Delen & Ram, 2018). The new generation of managers tends to make their business decisions based on critical thinking and the best available evidence, which includes data, information, assumptions, and hypotheses (Barends & Rousseau, 2018).

Literature review indicates that despite the number of opportunities that Business Analytics brings for the companies, there is room for challenges and barriers. One of the challenges is the lack of talented professionals, who can deal with the data, make decisions, perform the analysis and present it (Delen & Ram, 2018). Secondly, the resistance to change the organizational culture is playing a big role as a challenge since business analytics implementation does not only require hiring a good business analyst, it also requires other staff members from different departments to understand and accept the need for change. Another challenge behind business analytics adoption is the inability to justify its return on investment since analytics projects are complex, costly, and require some time to see the results, which can be directly or indirectly related to the actions that were taken before (Whitelock, 2018). The fourth challenge is the lack of strategy for handling huge amounts of structured and unstructured data. Fifthly, the less technical businesses can struggle with slow technology adoption due to the lack of personnel skills, technological instruments, and costs. Finally, the last challenge is related to security and privacy issues, which is one of the most common criticisms towards business analytics (Delen & Ram, 2018).

The question is if those opportunities and barriers are similar for the corporations that manage on average seven times more data than an average small company (Yanovitch, 2016). Small Business Enterprises have several unique characteristics that could influence the adoption of business analytical tools, therefore SBEs could potentially have other issues and benefits related to Business Analytics, which is the problem that I will address. Those characteristics are limited income and profit, limited project scope, limited variety of talented individuals, special needs. According to several studies, some of the main problems that small companies are facing, are identifying the need for technological changes, defining the technology adoption criteria (Ajimoko, 2018), and financing those projects (Ayoubi & Aljawarneh, 2018). Moreover, it is important to find out which type of analytics is the most appropriate for small enterprises. Generally speaking, Business Analytics is a promising concept, however, small companies could struggle with that or gain more benefits than other types of companies. For this reason, my purpose would be to investigate and identify the key opportunities and barriers of using business analytics in small companies to help small and medium-sized enterprises to go through this process more easily and smoothly.

One of the most important steps towards creating a competitive advantage is to understand the data that a company is generating in its own business. Having the unique information at the right time and being able to make the right decision make the basis for achieving the competitive advantage (Palmer & Hartley, 1999). Nowadays, the Global Market of Big Data Analytics is growing at a compound annual growth rate of 12.3% starting from 2019 to 2027. In the year 2018, the market was valued at US 37,34 billion, and, taking into consideration the compound annual growth rate, by 2027 the market is expected to reach US 105,08 billion. The main suppliers of business analytical tools remain the large businesses that hold more than 60% of the global market due to their financial capabilities and greater puissance to host large projects (Wood, 2020).

For a long time, large businesses had been realizing the value and importance of the existing data, they started to invest huge amounts of money into the analytical systems. Certain small companies around the world have only recently begun investing in business analytical tools to improve their business processes and gain unique competitive advantages in the currently fast-changing global economy (Guarda, Santos, Pinto, Augusto & Silva, 2013). Since this topic is gaining popularity on the global level, in this master's thesis, I am going to identify and analyze the benefits and challenges of implementing such tools in SMEs and particularly in one small company. Therefore, the master's thesis will address the research question "What are the opportunities and barriers of using Business Analytics in small companies?". The research question is directly aligned with the purpose of this thesis which is to investigate and identify the key opportunities and barriers of using business analytics in small companies to help SMEs to go through this process more easily and smoothly. The thesis includes an example of one small company that is going to be analyzed by using several analytical tools. The analysis itself, the results, and unique recommendations should help other small companies to make several decisions regarding the use of business analytics in their case.

In order to identify the opportunities and barriers of using business analytics in small companies, it was necessary to analyze and define the potential of business analytics and business intelligence, scopes, strengths, weaknesses, and their adaptability in the context of small businesses, which are some of the goals of this thesis. Every stage of the research has its list of goals that should be achieved in the end.

First of all, the goal is to review the literature, define the key terminology for the thesis, and identify the list of opportunities and barriers of using business analytics in small companies. Secondly, it is important to describe a particular business case, the main business issue, and the data that is going to be used for further analysis. The next set of goals is related to data analytics, where the goals are the following: select the data for the analysis, perform data cleansing steps, transform the data, split the data into training and testing sets, analyze the data that the company is using and build several models, evaluate and deploy the best models. Fourthly, the results of data analytics are interpreted with dashboards, figures, tables, and other tools. The additional goal for the analysis and results development stages is to find some additional opportunities and barriers of using business analytics that could complement the list derived from the literature review. Next, the goal is to evaluate the contribution of the thesis by performing an interview with the CEO of the company. The last set of goals is related to bringing all the findings together in order to answer the main research question, writing a list of suggestions for possible improvements in the company, and preparing the valuable conclusion of the thesis.

The methodological approach that is going to be used is case analysis. The first chapter will discuss the definition and the history of Business Analytics, its types, concepts, challenges, benefits, and importance. It is very important to justify the topic and show its importance.

The second chapter describes the main characteristics of small companies, defines their role in the global economy, discusses the implementation of business analytics in small companies, and presents the list of opportunities and barriers of using business analytics in small companies. The method that is used in the first two chapters is secondary data analysis.

The third chapter "Methodology" consists of two subchapters. The actual business case and the company description are presented in the first subchapter, while the second part includes the description of all the data that is used for the analysis. The fourth chapter includes the analysis of five datasets that were provided by Company X, which is a retail company that provides a service of personal stylists. The datasets include huge amounts of structured and unstructured data, therefore, the analysis involves several data analytical steps like exploratory data analysis, data cleansing, data preprocessing, data splitting, and application of machine learning algorithms. All the results and findings are presented in the fifth chapter, where the first subchapter includes the actual final model and its description, while the second subchapter represents another method that is used in the thesis which is an interview with the CEO of the company where he provides his feedback regarding the analysis, results, and suggestions. The last chapter is a discussion of the topic, the analysis, and the results. The relationship between the actual and prior findings from the literature is defined and discussed in the first subchapter, while in the second subchapter, we can find a list of suggestions for improvements with some potential costs and benefits of those improvements. The main data analytical tools that are used for the analysis are Google Sheets and Rapid Miner.

1 BUSINESS ANALYTICS

Business Analytics can be defined as a generalization and abstraction of many activities that occur in companies. The term consists of two independent terms that together can be explained as the application of analytics to various business problems (Power, Heavin, McDermott & Daly, 2018). The term "business" refers to commercial activities, while the term "analytics" is a scientific discipline of fact-based problem-solving (Nelson, 2017). There are various definitions of the term Business Analytics since each of them has a different main focus. It can be the nature of data, scope, the main problem, coverage, enabling methods, and others. However, all of them have one single denominator which helps to define Business Analytics as the consolidation of all the supporting mechanisms that help to transform the data into a valuable piece of information that will improve and accelerate the decision-making and problem-solving processes (Delen & Ram, 2018). Business Analytics involves statistics, mathematics, machine learning, and professional knowledge to measure daily insights in plenty of business areas like sales, banking, marketing, finance, eCommerce, human resources, and others (Amber, 2017).

The term “Analytics” has its own long history and can be used in any area where the process of problem-solving plays a big role, however, the history of the term “Business Analytics” can be described as relatively recent. In the late 19th and the beginning of the 20th century, Frederick Winslow Taylor initiated the history of the term Business Analytics when he introduced the first formalized system of business analytics in the USA (Foote, 2018). In 1909, Frederick Winslow Taylor published "The Principles of Scientific Management" which helped to analyze and synthesize different workflows (Giannantonio & Hurley-Hanson, 2011). This theory of Scientific Management led Henry Ford to continue making the history of business analytics and to introduce Ford’s car assembly line time measurements (Foote, 2018). Henry Ford was measuring the time needed for each component of the Ford Model T car to be completed on the assembly line. Even though nowadays it could be treated as a simple task, at the beginning of the 20th century, Henry Ford revolutionized the automotive and manufacturing industries around the world (Davenport, 2006). Therefore, in the early days of the evolution of business analytics organizations were mostly focused on higher productivity and improved efficiency. Beginning in the middle of the 20th century, the next stage in the history of business analytics can be explained as a stage of operational reporting. At this point, organizations were gathering and saving certain information to be able to create daily reporting. However, the information was not easily shared company-wide which led to little integration and small amounts of historical data saved (Amber, 2017).

In the 1960s, computers started providing support in the decision-making processes, therefore, the term Business Analytics started receiving more attention and evolving more significantly (Foote, 2018). One of the most crucial steps towards the development of advanced business analytics was the invention of Decision Support Systems since they were able to filter and sort huge amounts of data, which could make the process of decision-making easier for executives (Keen, 1980). During the 1980s and 1990s, computers together with business analytics continued to evolve, which allowed organizations to save the historical data and prepare it for analysis. One of the first analytical tools that was based on the DSS platform was Microsoft Excel which was introduced in 1985 (Amber, 2017). At the end of the 20th and the beginning of the 21st centuries, there were introduced a variety of other analytical tools that are commonly used worldwide. Some of the tools that can be used for data science, text mining, data cleansing, data preprocessing, machine learning, and others, are Python (McKinney, 2017), R (Ihaka, 2012), RapidMiner (Kotu & Deshpande, 2014), Power BI (Becker & Gould, 2019), SQL (Linoff, 2015), and others.

There are plenty of terms related to Business Analytics that play a big role in understanding the term completely. To make the right decisions, identify issues, be efficient and profitable, organizations around the world start implementing such things as Business Analytics, Business Intelligence, Data Analytics, Big Data Analytics, and others. All these relatively new terms are related to data storage, data gathering, data management, and decision-making processes (Sun, Zou & Stang, 2015). However, despite those similarities, each term has its

uniqueness. As was mentioned above, business analytics is the consolidation of all the supporting mechanisms that help to transform the data into a valuable piece of information that will improve and accelerate the decision-making and problem-solving processes (Delen & Ram, 2018). Business analytics helps companies to anticipate business outcomes and trends through statistical analysis, data mining, and predictive modeling, whereas business intelligence helps to evaluate complex data and create an input for the decision-making process (Negash & Gray, 2008). Therefore, the main difference between these two terms is that business analytics prioritizes predictive analytics and involves data modeling with machine learning to predict future outcomes, while business intelligence focuses mostly on descriptive analytics which summarises the actual and historical data (Tableau, n.d.).

Another important term that is a component of business analytics is Data Analytics. It can be defined as a technique that examines, summarizes, and draws conclusions from databases to describe and anticipate something. The main difference between data analytics and business analytics is that the first term is broad, it focuses on finding insights in any kind of data, while business analytics prioritizes the identification of operational insights (Sun, Zou & Stang, 2015). One more equally important term is Big Data Analytics which brings new opportunities for companies and adds value to business analytics. Big Data can be described as huge datasets with complicated structures, which causes problems in analyzing and visualizing this data (Sagiroglu & Sinanc, 2013). Therefore, Big Data Analytics is an advanced and integrated form of Data Analytics, which helps to collect, organize and analyze big data to discover new patterns and other pieces of information within the big data. Big data analytics can be represented as a summary of data analytics, big data, data storage, data mining, statistical modeling, machine learning, visualization, and optimization (Sun, Zou & Stang, 2015). All those terms and techniques are also the components of business analytics. The main difference, in this case, is that business analytics focuses on operational insights and their practical application.

As it was mentioned before, Business Analytics includes various components including data analytics. According to Gartner's Analytics Ascendancy Model, business analytics, as well as data analytics, can be divided into four main types of analytics (Tamm, Seddon & Shanks, 2013, p. 4). The first type is called Descriptive analytics, which is used to describe the historical and existing data (Whitelock, 2018). According to the model, descriptive analytics has the lowest level of difficulty and provides the smallest value among all four types (Eriksson, Bigi & Bonera, 2020). Descriptive analytics is one of the most commonly used types of analytics to describe different financial metrics like price and cost changes, sales growth, number of customers per day, or a customer average bill. Descriptive analytics and different visualization tools help managers to see the whole view of what has occurred in a company during a certain period (Kaur & Phutela, 2018). Therefore, as we can see from Table 1 below, the main questions that it can answer are "What is happening in the business?" and "What was happening" (Eriksson, Bigi & Bonera, 2020). This type of analytics helps in identifying strengths and weaknesses in an organization, moreover, it is an

essential component of performance analysis. Descriptive analytics help in decision-making but it does not solve problems and offer solutions by itself (Kaur & Phutela, 2018).

Another type is Diagnostic analytics, which focuses on the reasons for the problem occurrences. It helps managers to adjust the company operations and improve the situation (Whitelock, 2018). This type of analytics requires exploratory data analysis to find the root causes of an issue, therefore the main question that it answers is "Why is something happening?" (Banerjee, Bandyopadhyay & Acharya, 2013).

There are four main techniques used for performing diagnostic analytics. Firstly, the data discovery technique should be applied, which means that the data sources should be identified. The second technique is called drill-down, which involves prioritizing a particular facet of the data or certain widget. The third technique is data mining, which can be defined as getting information from a huge amount of raw data in an automated way. The last technique is related to finding correlations in different datasets, which helps in defining the parameters of the analysis (Sisense, 2021). According to Gartner's Analytics Ascendancy Model, diagnostic analytics has medium-low value and medium-low level of complexity, since it does not involve decision-making and forecasting (Eriksson, Bigi & Bonera, 2020).

Table 1: Types of Data Analytics

	Descriptive Aanalytics	**Diagnostic Analytics**	**Predictive Analytics**	**Prescriptive Analytics**
Goal	Description of historical and existing data	To find the reasons behind certain outcomes	To predict trends and probabilities that could happen in the future	To provide the possible decisions to maximize good outcomes
Question	What is happening in the business?	Why is it happening?	What can happen?	What to do next?
Value	Low	Medium-low	Medium-high	High
Complexity	Low	Medium-low	Medium-high	High
Outcome	Visualization of the past or current situation	Defined causes of certain issues	Predicting future trends	Testing all the potential outcomes and choosing the right decision

Source: Own work.

The third type of analytics described by Gartner is Predictive analytics, where the main focus is forecasting and predicting trends and probabilities that could happen in the future, which is useful for preparing what-if analysis (Whitelock, 2018). The method that is used in

predictive analytics is to anticipate future outcomes by using past events. This type of analytics involves statistics, mathematics, artificial intelligence, robotics, and other techniques to explore valuable relationships and patterns in the data. However, it still requires a certain degree of human involvement because only professionals have certain knowledge about the business and know how to prepare the data, which tools to apply, and how to interpret the results. Predictive models are time-consuming and require hard work, and in the end, there is no guarantee that the results will provide any business value (Eckerson, 2007). Therefore, predictive analytics has a medium-high value and a medium-high level of complexity, which we can see from Table 1 (Eriksson, Bigi & Bonera, 2020).

Finally, the most valuable and difficult to implement form of data analytics is Prescriptive analytics, which provides the possible decisions to maximize good outcomes and minimize the number of bad outcomes (Whitelock, 2018). Prescriptive analytics is considered the most helpful type of analytics because it can assist companies in maximizing their values and mitigating risks by recommending various actions that can be taken (Soltanpoor & Sellis, 2016). The key question that it answers is "What to do next?". Therefore, it is characterized as the most valuable and complex type of analytics (Eriksson, Bigi & Bonera, 2020).

One of the most challenging processes related to business analytics is the implementation process. There are several roles in business analytics that are crucial for successful implementation. The business people that also include top managers of a company see themselves as some sort of "consumers of analytics" since in most cases they initiate the business analysis process by requesting a solution to a certain problem (Saxena & Srinivasan, 2012). Another role in business analytics plays the IT team, which is responsible for IT support, analytics capabilities development, qualified software solutions, data warehousing, and business intelligence. In this case, IT teams can be treated as "suppliers of analytics". However, the most important role play business analysts, who translate business strategies into solutions. A good business analyst is a person who is capable of filling the information gap between goals and execution by identifying the projects with the biggest rewards and minimal risks, by finding surprising data patterns and insights, by visualizing the situations and problems, by defining problems and offering decisions (Lindbergh, VanderHorst, Hass & Ziemski, 2018).

The implementation process starts when managers or financial professionals identify the need for change or the need of using the company data in a new way. It does not mean that a company should start adopting complex analysis since every company needs its kind of analysis that fits its own special needs. This is the most suitable approach to get a competitive advantage with the help of business analytics (Brands & Holtzblatt, 2015). Business analytics implementation is a complex and time-consuming process that consists of seven steps. Firstly, a business analytics team together with managers should define the company's objectives related to business analytics. At this point, the participants should discuss the organization's mission and vision to define the focus and requirements for business analytics. This step is essential for small companies due to limited resources.

Secondly, the organizational structure should be clearly defined. According to Gartner, the most appropriate structure for business analytics deployment is the combination of centralized and decentralized teams. The centralized model provides essential consistency, governance, and in-house control of data. Nonetheless, decentralized structure allows business analysts to manage the data without getting stuck into a centralized structure's rules. The third critical step is the creation of cross-functional teams that include management accountants who are aware of the company's financial activities, IT professionals who can set up data flows and take care of data warehousing, and business analysis who can fill the gap between the data and company goals (Brands & Holtzblatt, 2015).

Fourthly, the team should prepare a detailed business analytics plan that is represented as a document that specifies and translates the requirements of the objectives into analytical models that can benefit the company. It should be a formal detailed plan that includes such aspects as data privacy, internal control, time frames, key objectives, data transmission, and others. A misleading or incorrect analysis can lead to making the wrong decision. Fifthly, it is important to select appropriate business analytics software. This step requires the participation of all three types of roles because the analytical software should meet all the company requirements. The budget of a small company may not allow running complex data mining and business intelligence models but a small company can find another solution that fits its budget and needs. For example, such software can be Microsoft Power BI or Tableau, which are relatively affordable and easy to use. Before purchasing and setting up software a company should make sure that it also meets other important aspects like existing hardware, cloud configurations, maintenance opportunities, staff competencies, and others. The sixth step includes the actual system implementation and testing. A business analyst should review and test the system to be sure that analytical models use the right databases and in the right way, generate highly accurate results, meet business requirements, and add value to the business. The last step which is called "Evaluate and revise" reminds business analysts and other team members to constantly review models, track changes in the requirements, and develop new models that are relevant for different periods (Brands & Holtzblatt, 2015).

The Business Analytics process consists of several steps that are essential for performing the analysis and developing the right solution. The first step is the problem identification or the planning phase from Figure 1 below, where the managers should realize the need for business analytics. This step is followed by the identification of issues and opportunities for collecting data, and the actual data collection step where the company should decide on which data to collect, how, and where to store it (Raj, Wong & Beaumont, 2016). In the case of unstructured or semistructured data, such steps as data cleansing and data preparation steps should be applied in order to prepare it for the analysis (Tan, Han & Elmasri, 2000). The next steps are the actual analysis and the model deployment, followed by data interpretation in order to solve the problem. The last step is the evaluation of the results. Each such process should be finished with the efficiency and model accuracy evaluation that helps to make the right decision (Raj, Wong & Beaumont, 2016).

Figure 1: The Business Analytics Process

Adapted from Raj, Wong & Beaumont, (2016).

Considering all the benefits of business analytics that were mentioned above, such as discovering insights from data, using fact-based decision making, making real-time decisions, and others, there are other benefits related not only to information but the company's overall performance. Business analytics in big companies benefits customer experience, accelerates innovation, decreases costs of operations, and improves revenue growth (Someh & Shanks, 2015). The practical examples of using business analytics in different sized companies bring other advantages and benefit fraud detection, talent detection, consumer basket analysis, response modeling, retention modeling, customer segmentation, market risk modeling, operational risk modeling, tax avoidance detection, terrorism detection, demand forecasting, web analytics, and others (Baesens, 2014).

There are three main reasons why different companies around the world decide to implement analytical tools in their businesses. The first reason behind the popularity of the term business analytics is that different-sized companies around the world are struggling with managing huge amounts of data while various business intelligence and analytical tools provide access to automated data collection systems. These systems make data management faster and easier, which benefits the company's internal processes (Delen & Ram, 2018). Another reason is that the data should be analyzed and transformed into valuable information, while this information needs to be converted into knowledge, knowledge into insight, and then into an action that leads to better decision-making and improved performance. This is a complex and time-consuming process that also requires plenty of specialists, while business analytical tools make it way faster and easier. Moreover, the tools can provide more accurate results and find surprising patterns in datasets (Whitelock, 2018). The third reason is the cultural tendency to change towards evidence-based management. A variety of successful organizations around the world are making an effort to shift into data-driven business practices (Delen & Ram, 2018). The new generation of managers tends to make their business decisions based on critical thinking and the best available evidence, which includes data, information, assumptions, and hypotheses (Barends & Rousseau, 2018).

Another important aspect of business analytics that should be mentioned is the challenges. As was mentioned above, there are several opportunities and benefits that companies can get from business analytics, however, there are also some challenges that prevent the

implementation process. The first challenge is the lack of talented professionals, which means that it is hard to find a good data analyst or a team of qualified analysts who have all the skills and knowledge required to be able to deal with the right data, make decisions, and convert data into valuable insight. Business analytics has its value when the analysis has high quality (Delen & Ram, 2018). Secondly, the resistance to change the organizational culture is playing a big role as a challenge. The implementation of business analytics requires several company departments to collaborate. Those departments are the IT department, accounting department, human resource department, and others. Employees may not understand the value of business analytics, here where the resistance to change is born (Brands & Holtzblatt, 2015). The implementation of business analytical tools permanently requires the companies to make a step from the traditional management style towards contemporary or evidence-based management. Thirdly, it is hard to calculate and clearly define the return on investment from business analytics since those kinds of projects are costly, complex, and require some time to see the results. While it is hard to define the correlation between the business analytics implementation and positive or negative results (Whitelock, 2018). Fourthly, in most cases, companies are lacking the right strategy for handling huge amounts of structured and unstructured data. Here is also the point where companies start struggling with the lack of data interoperability. According to the international World Wide Web Consortium, the Internet expansion provokes an increase in the amount of unstructured data available, which widens the gap between structured and unstructured data (Brands & Holtzblatt, 2015). The fifth reason is the issues related to technology adoption since a company that decided to implement business analytics should be aware of the costs, have proper technological instruments, and have special personnel skills. Finally, the last challenge is related to security and privacy issues, which is one of the most common criticisms towards business analytics (Delen & Ram, 2018). Companies should understand that their data should be protected and treated as an asset not only on the IT department level but also on the management level since the loss of privacy can lead to different types of risks such as damaged reputation and others (Brands & Holtzblatt, 2015).

2 THE USE OF BUSINESS ANALYTICS IN SMALL COMPANIES

According to the European Commission Recommendation to the Member States, the European Investment Bank, and the European Investment Fund, an enterprise is defined as an entity that is engaged in an economic activity, irrespectively of its legal form, while an economic activity is defined as the sale of products or services at a certain price, on a certain market (CSES, 2012).

The definition of a small company is included in the broader term SME, which stands for a Small and Medium-sized Enterprise. All the enterprises must be categorized according to three main conditions which are the staff headcount, financial turnover, and total balance sheet. According to the European Commission's definition, an SME is an enterprise that

officially has less than 250 employees, which annual turnover does not exceed EUR 50 million, and an annual balance sheet does not exceed EUR 43 million. Whereas the ceilings for the definition of a small enterprise or a small company are the following: the number of employees does not exceed 50, the annual turnover is less than EUR 10 million, and the annual balance sheet does not exceed EUR 10 million (CSES, 2012). OECD defines an SME as a company that has up to 249 employees, a micro-enterprise as a company that has up to 9 employees, and up to 49 employees in a small enterprise. The United States defines an SME as a company that has less than 500 people (Natarajan & Wyrick, 2011).

The main objective of stating clearly the definition of a small enterprise and an SME is the fact that most countries apply different regulations, policies, programs, and supporting measures for every company category. For instance, the European Union offers different funding programs that support small companies. Such programs are divided into direct funding which includes grants and contracts, and indirect funding which includes such funds as European Regional Development Fund, European Social Fund, Cohesion Fund, and others (European Commission, 2021). However, according to the statistics, around 20% of small and medium-sized enterprises in developed countries last less than one year, another 20% last around two years, 50% of them last less than five years, and only 10% of SMEs survive in the long run and have a chance to change the economy (Bayraktar & Algan, 2019).

The President of the European Commission, Jean-Claude Juncker stated that SMEs are the backbone of the economy since they are creating more than 85% of new working places in Europe. Moreover, in the European Union, nine out of every ten enterprises are categorized as SMEs (European Commission, 2017). While, according to the World Trade Organization, SMEs represent around 90% of companies worldwide, and provide on average 60-70% of working places in developed economies (Azevedo, 2016). Furthermore, small and medium-sized enterprises provide entrepreneur skills, stimulate competition prices, improve efficiency, diversify product designs, encourage freedom of choice, and speed up innovation. A study made in the United States shows that SMEs make four times greater profit for a dollar invested than huge companies (Neagu, 2016). Therefore, SMEs play a big role in stimulating economic growth worldwide, moreover, they play an important role in achieving the Sustainable Development Goals (Bayraktar & Algan, 2019).

There are several specific challenges and characteristics of small and medium-sized enterprises that make them unique players in the global market. The first challenge is the inability to access finance or to make the necessary investments in research and innovation. Moreover, they can also lack the resources to fulfill environmental regulation requirements (European Commission, 2017). Small and medium-sized enterprises are less likely to get bank loans than large corporations. The International Finance Corporation estimates that 40% of SMEs in developing countries have an unfulfilled financing need of EUR 4,3 trillion every year (World Bank, 2019). Moreover, SMEs are struggling with building and maintaining customer loyalty which is directly related to the lack of financing (Goebel, Norman & Karanasios, 2015). Another challenge is related to the structural barriers, such as

a lack of management, lack of technical and expert skills, labor market rigidities, lack of information, and limited amount of knowledge (European Commission, 2017). Most SMEs do not have specialized IT departments which means that in many cases they are run by owners who might not have expert technological knowledge. Despite those challenges, small enterprises are able to quickly adapt to changes due to their size in comparison to some large companies (Goebel, Norman & Karanasios, 2015).

Small and medium-sized enterprises are considered as a backbone for economic growth and success as was mentioned above. However, this growth and success take place in the 21st century, when most enterprises around the world are using informational technology in order to accelerate economic growth. Business analytics refers to information technologies that allow analyzing big amounts of data and discovering new and valuable insights (Goebel, Norman & Karanasios, 2015). Despite business analytics applications are primarily accessible to large enterprises since they meet their specific needs and resources, according to the surveys performed by Gartner, business intelligence and analytics systems are ranked as the highest technological priority of different-sized enterprises in the last few years worldwide (Papachristodoulou, Koutsaki & Kirkos, 2017). Whereas, according to the "Oxford Economic Survey-2013", information technology and innovation are the factors of SME growth. Moreover, the study defined big data as an important factor in company success (Iqbal et al., 2018).

As it was mentioned before, there are several challenges and special characteristics of small companies that differ them from large enterprises. As the literature review shows, small and medium-sized enterprises have plenty of shared opportunities and barriers of using business analytics, however, they mostly differ from the large companies. Therefore, it is important to define the reasons why small companies decide to implement business analytical tools despite the lack of finance, management, expert skills, information, and technical knowledge. Moreover, small companies, as well as large corporations, struggle with excessively huge volumes of structured and unstructured data (Papachristodoulou, Koutsaki & Kirkos, 2017). According to several pieces of research and studies, there are several benefits of using business analytics in small companies.

First of all, small and medium-sized companies tend to implement business analytics due to improved data support, which solves one of the SMEs' challenges. This factor also includes the reduced amount of time and effort needed for the analysis and reporting, high-quality reports are available much faster and to the right people. The second benefit is the improved decision-making process. Business analytics supports this process by providing precise and valuable analysis, calculating possible risks, and offering possible solutions. Another benefit is that such innovative solutions provide flexibility and freedom of choice. Generally, information technology departments in small companies do not get along with other departments. For instance, when the marketing department needs a new automated system, it typically seeks the system independently, applies it, and starts using it. Such situations in small companies increase the amount of unstructured data that exists in different data types

throughout the company. Business analytics and big data solutions allow small companies to focus only on the most important capabilities and apply one main solution and data type for the whole company in order to be able to extract insights from data (Ogbuokiri, Udanor & Agu, 2015). Finally, business intelligence and analytics help small companies in gaining a competitive advantage in the market and saving costs on employees in different departments such as IT (Papachristodoulou, Koutsaki & Kirkos, 2017). Moreover, the right analytical solution that meets the organizational needs, can even lower other types of costs and benefit the company overall (Ogbuokiri, Udanor & Agu, 2015).

However, despite the fact that business intelligence and analytics solve several small company's initial issues, such solutions can also bring other types of challenges for the company, which we can see from Table 2 below. The first set of issues is related to the SMEs' awareness of business analytics and big data. Managers are not sure if their company data meets the definition of Big Data and is able to benefit the decision-making process. As a result, they are not able to calculate the ROI from such solutions and see them as uncertain. There is a lack of business cases about successful business analytics implementation in SMEs, which is an additional reason for a small interest in implementing BA in small companies and medium-sized companies (Iqbal et al., 2018). Another barrier is the inadequate use of financial resources. Business analytics can help a small enterprise to increase its revenue in the long run, however, it can be done only if a company is using the right type of business analytics that meets company needs (Russegger et al., 2015).

The next issue is related to the specialists that are typically working in small companies. Generally, small and medium-sized enterprises operate in one specialized field, which plays a role in their strength. Therefore, most employees working in a small company are field specialists with no expert knowledge in business analytics and IT, while hiring an expert would be costly (Russegger et al., 2015). Moreover, varying from country to country, analytics on customers' data can be protected by the law. This means that a small company requires additional expert knowledge in this field in order to perform data analytics (Iqbal et al., 2018). Most small and medium-sized companies are not capable of doing business analytics by themself. Whereas, the practice shows that SMEs also struggle in getting access to data analytics consulting services. The reason behind this is that, in general, consulting firms are large, they are focused on working with large enterprises and solving complex problems. The business practices of these consulting firms are not aligned with the needs and financial resources of small and medium-sized companies (Papachristodoulou, Koutsaki & Kirkos, 2017).

Another challenging factor is the data itself and the data usage. Due to the lack of experts, data in small companies are usually unstructured, the reports are complex and hard to read, and the handling of the solution is too complicated. Moreover, various business analytics and intelligence tools and functions do not meet the business needs of small enterprises, they struggle with software errors, inadequate security functions, and lack of expertise (Papachristodoulou, Koutsaki & Kirkos, 2017).

The last challenge is related to data security. In comparison to large enterprises, small and medium-sized companies see data security issues more seriously. As it was mentioned before, small and medium-sized enterprises are frequently struggling with the lack of financial resources that are essential for qualified data security. Generally, small enterprises have outdated database management systems, which brings threats of data breaches and cyber-attacks (Lacey & James, 2010).

Table 2: Opportunities and Barriers of using Business Analytics in small companies

OPPORTUNITIES	BARRIERS
+ Improved data support + Improved decision-making process + Flexibility and freedom of choice + Focus only on the most important capabilities + Extracting insights from data + Gaining a competitive advantage + Saving costs	- Unawareness of business analytics and inability to calculate the ROI from such solutions - Inadequate use and the lack of financial resources - Absence of expert knowledge - Law protection - Poor data quality and data usage - Poor data security

Source: Own work.

Another challenging factor is the data itself and the data usage. Due to the lack of experts, data in small companies are usually unstructured, the reports are complex and hard to read, and the handling of the solution is too complicated. Moreover, various business analytics and intelligence tools and functions do not meet the business needs of small enterprises, they struggle with software errors, inadequate security functions, and lack of expertise (Papachristodoulou, Koutsaki & Kirkos, 2017). The last challenge is related to data security. In comparison to large enterprises, small and medium-sized companies see data security issues more seriously. As it was mentioned before, small and medium-sized enterprises are frequently struggling with the lack of financial resources that are essential for qualified data security. Generally, small enterprises have outdated database management systems, which brings threats of data breaches and cyber-attacks (Lacey & James, 2010).

Before choosing a business analytical solution for a small enterprise, the key performer indicators and the company strategy should be clearly defined. There are several common characteristics that bring together all the small and medium-sized companies, however, each company has its own unique strategy and needs which makes the process of implementation of business analytical tools, even more, complex (Goebel, Norman & Karanasios, 2015). Each small company may be using business analytics for different purposes, which means that not every company should implement a particular type of analytics (Chernyshova, 2013).

In order to be able to compete with other companies in the market, small enterprises should monitor and use their resources in the most effective way, especially information resources that assist them in decision-making (Raj, Wong & Beaumont, 2016). There are several types of analytical solutions that should be chosen based on the company's resources, needs, and goals. Practically, considering the costs and complexity of business analytical solutions, small enterprises tend to implement simple spreadsheets such as Microsoft Excel which are integrated with the company's data. However, typically such spreadsheets do not include a rich set of data analytics and visualization tools to facilitate insight discovery. According to several different studies, the most suitable solution for small enterprises would be using analytics-as-a-service solutions, particularly software as a service (Goebel, Norman & Karanasios, 2015). Such solutions are able to help the companies to understand the customers in a better way, enter the new markets, and cut down the costs (Iqbal et al., 2018). Such solutions are less costly and easier to implement in small enterprises (Delen & Ram, 2018). The main vendors that provide such solutions that can be used in a small company are Microsoft, IBM, SAP, SAS, Oracle, MicroStrategy, and others (Tutunea & Rus, 2012). For instance, the cheapest monthly offer from Microsoft Power BI costs EUR 8,40 per one user and includes the ability to publish and share reports, the ability to connect to data sources, controls, AI visuals, data security, metrics for content creation, and other (PowerBI, 2021). The main concerns of using such cloud-based services are data privacy, data security, and data ownership (Raj, Wong & Beaumont, 2016).

Small companies with no or small IT departments may be using analytics-as-a-service solutions for descriptive analytics and visualizations to describe different financial metrics like price and cost changes, sales growth, number of customers per day, or a customer average bill. Descriptive analytics and different visualization tools help managers to see the whole view of what has occurred in a company during a certain period (Kaur & Phutela, 2018). While analytics-as-a-service solutions can also be used by small and medium-sized companies with bigger IT abilities. In this case, data analytics is used for finding correlations, evaluating risks, forecasting the demand, optimizing processes, performing predictive inventory planning, and others (Sen, Ozturk & Vayvay, 2016).

3 METHODOLOGY

3.1 Business Case

The empirical part involves the implementation of business analytics and the evaluation of its opportunities and challenges in a small company. Company X provides a service of personal stylists that work on providing the best combinations of clothes for the clients. The company name and origin can not be revealed due to this information is hidden from third parties. Company X is a startup that is has around 35 employees, which classifies the

company as a small enterprise, according to the literature review in the previous chapters. Considering the company's modus operandi, it is operating in the retail trade industry.

The business process consists of several steps where the first one is that a client fills out a questionnaire with a detailed description of the parameters and preferences. The form consists of a mobile number, which is also the client's ID, preferred style, clothes size, age, expectations from the service, the amount of money willing to spend, personal address, occasion, and other data. Filling out the questionnaire takes 15-20 minutes, and all the personal information remains completely confidential. Then the platform sends it to one of the company's stylists, who creates a personalized selection of clothes based on the wishes, tastes, and budget of the client. The stylist creates it by using the company's internal platform, where stylists can see what is available in the warehouse. Company X is present only online, however, it has a warehouse where different clothes, shoes, and accessories are stored. The company has several suppliers from whom it purchases clothes, shoes, and accessories wholesale, which allows the company to reduce costs and accelerate the work of stylists. If a client is satisfied with the quality of the stylist's work, then the same stylist will be assigned to the next order of the same client. This practice helps the stylist to find out more personal information and build long-term customer relationships. Then the staff puts all those pieces of clothes in a box and sends them to the client. Each selection includes clothes, shoes, and accessories that the client can try on at home and return them back completely free of charge the next day if something doesn't suit her or him. The client has 24 hours to try those clothes and decide what to leave and what to send back. Together with each box, the client receives a letter with the stylist's advice on the selection of the image, the combination of things with the client's wardrobe, as well as answers to questions that the client can ask the professional stylist in the questionnaire (Company's website).

The main customer segment consists of women for the moment, the company is also introducing the same service for men. Access to the service is available through the company's Instagram page or through the company's website, where customers can read about the service and make an order. The price of each order depends on the client's preferences which are taken into consideration by each stylist. Each box typically consists of 6 pieces of clothes, shoes, and accessories, and the client can leave only those that suit her the best (Company's website). Each client is not aware of the contents of the box until it's delivered and opened. After making an order, a client should pay the service fee of EUR 10,95. Then after receiving the box and trying on all the pieces, the client sends her feedback about the clothes that are going to be purchased, whereupon the company's staff sends the invoice. The average delivery time is around 5 days. If a customer is willing to purchase at least one piece of clothes, shoes, or accessories, then the service fee is deducted from the final bill and the service becomes free of charge for the customer (Company's website).

The main idea of the company is that instead of going to shopping malls and stores or wasting time in different online stores, a client can simply visit Company X's website, fill out a short questionnaire, and get the selection of clothes, shoes, and accessories that fit both in size and

style. The CEO of Company X points out: “I believe that on the horizon of 10 years, clothes will become a kind of utility. Large percentage of people will stop looking at them as a means of "showing themselves", but will be buying clothes in order not to freeze. Therefore, people will just need to replenish stocks of clothes as they do with toilet paper, toothpaste, socks, and others. Here will be the point where our company will be even more relevant” (Interview with the CEO). The company’s strategy of working with clients is focused on building long-term personalized relationships. Each client has a personal manager in WhatsApp, who guides clients and makes their customer experience with the brand. Moreover, each stylist writes a unique note to each client on how to combine things inside the box, which also benefits the customer experience.

The company showed the need for additional business analytics when the CEO realized that the company has a lot of data that is not used for adding value to the business, however, at the moment the company uses some descriptive analytics to support several basic business decisions. Therefore, there are no defined business problems that should be solved with the analysis, the main focus is to find useful data insights that could create additional opportunities for the business.

The thesis includes several ways for answering the research question which is “What are the opportunities and barriers of using Business Analytics in small companies?”. Firstly, the set of opportunities and barriers was defined in the literature review in the previous chapter, then the analysis helps to identify the additional technical opportunities and barriers of using business analytics. Thirdly, the first result subchapter includes several data insights that could become potential opportunities or barriers for the business, then the interview with the CEO of Company X provides additional information about the opportunities and barriers of using business analytics in a small company that were not defined in the thesis before. Moreover, the interview will show the relevance of the opportunities and barriers defined during the data analysis and results development. All the findings will be summarized and discussed in the sixth chapter in order to provide a complete answer to the research question. In order to justify the topic’s importance, the findings from the first two chapters were presented from the review of literature. The method that was used is secondary data analysis and the methodological approach that is going to be used in the whole thesis is case analysis.

3.2 The Data

The company’s database that is going to be used for further analysis and result development, includes 5 datasets. Each dataset will be used for the analysis, however, all of them should go through several data preparation and data cleansing steps in order to be ready for further analysis since the data is partly structured and unstructured.

The first dataset is called “Cloth Feedback”, which consists of 13 columns and 22.185 rows. This dataset represents the customer feedback regarding each item received, which includes the following columns: a two-digit feedback ID, Date column, Cloth ID, Cloth Status, Item

Fit, Quality Rating up to 5, Price Rating up to 5, Style Rating up to 5, Purchase ID, Client's Comment, Cloth Number in the box from 1 to 6, COGS, and Price. Each item from one box has a different Cloth ID and the same Purchase ID because it stands for one order. The Cloth Status column shows if an item is bought in the end or returned. In the Item Fit column, we can see the customer feedback about the size of an item, it can be too small, too big, or a perfect fit. Moreover, each client is able to evaluate the item quality, price, and style from 1 to 5. In total, around 85% of all the clients provide feedback regarding each cloth, which can help to detect the mistakes in service. The first feedback in the dataset is dated July 9, 2020, and the last one is February 8, 2021.

The second dataset "General Feedback" includes the data about each purchase of a box. It is composed of 15 columns and 3.480 rows, in the time period between July 9, 2020, and February 2, 2021. This dataset provides the customer feedback about the whole purchase and the service that Company X is providing. The columns are the following: Date and column, Purchase ID, Service Rating from 1 to 5, Style Rating from 1 to 5, Overall Rating from 1 to 5, Recommendation to the stylist, General Comment, New or Old stylist, Net Promoter Score from 1 to 10, NPS Reason, Improvements, Delivery Rating from 1 to 10, Comments about the delivery, and a two-digit feedback ID. The NPS Reason column includes the comments and reasons of customers for assessment of low or high net promoter scores, while the Improvements column includes the recommendations from customers for further service improvements. The next dataset which is called "Order Status" shows the Client's Name, Shipping Date, Purchase ID, Order Value, Advance Payment, Purchase Price, Number of Clothes Sent, Number of Clothes Purchased, COGS, and Margin, which creates 10 columns. "Order Status" dataset includes 4.428 rows starting from January 10, 2020, until January 29, 2021.

The fourth and the fifth datasets are the results of two types of questionnaires that the company was using in order to get more information about the customers while they were making orders. From February 4, 2019 to November 30, 2020 the company was conducting a survey using Google Form, dataset "Quiz 2" represents the results of the survey. The dataset consists of 2.617 rows and 235 columns, where some of them are missing the data. On September 5, 2020, the company introduced another questionnaire from Typeform. Therefore, another dataset "Quiz 1" was created. It consists of 78 columns and 1.950 rows until April 28, 2021. Both datasets include several similar columns such as Date, Customer ID which is also a telephone number, Birthday, Age, Occasion, Expected Style, Height, Favorite Brands, Source from where a customer found out about the brand, Email, Instagram, Additional social network, Occupation, Dress Size, Top Size, Pullover Size, Skirt Size, Pants Size, Hair Color, Comments, Parameters, Body Specialties, Favorite fit, Willingness to wear accessories, Undesirable color, Undesirable material, Ear Piercing, Home Address, and others. However, there are several differences between the datasets. For instance, "Quiz 2" includes information about the customer's shoe size, breast size, and amount of money willing to pay for each item category, which is not available in another dataset.

The main tools that are going to be used for the following analysis are Google Sheets and RapidMiner. Google Sheets is a useful tool in this case because the original datasets are in another language, while in Google Sheets they can be translated into the English language. Moreover, the company is initially using this tool, therefore some data cleansing and data preparation steps are going to be performed in Google Sheets. The actual analysis of data, model development, text mining, and visualizations are going to be done in RapidMiner.

Table 3: Initial Database

Dataset	Period	№ of rows	№ of columns	Columns
Cloth Feedback	09.07.2020-02.02.2021	22.185	13	Feedback ID, Date, Cloth ID, Status, Fit, Quality Rating, Price Rating, Style Rating, Purchase ID, Comments, Cloth Number, COGS, and Price.
General Feedback	09.07.2020-02.02.2021	3.480	15	Date, Purchase ID, Service Rating, Style Rating, Overall Rating, Recommendation, Comment, New or Old stylist, NPS, NPS Reason, Improvements, Delivery Rating, Comments about the delivery, Feedback ID.
Order Status	10.01.2020-29.01.2021	4.428	10	Name, Shipping Date, Purchase ID, Order Value, Advance Payment, Price, Clothes Sent, Clothes Purchased, COGS, Margin.
Quiz 1	05.09.2020-28.04.2021	1.950	78	Date, Customer ID, Birthday, Age group, Expected Style, Height, Favorite Brands, Email, Facebook, Instagram, Occupation, Dress Size, Comments, Parameters, Body Specialties, Undesirable material, Ear Piercing, Address, and others.
Quiz 2	04.02.2020-30.11.2020	2.617	235	Date, Customer ID, Birthday, Age group, Occasion, Willingness to spend, Style, Height, Favorite Brands, Email, Facebook, Instagram, Occupation, and others.

Source: Own work.

4 ANALYSIS

The analysis of the database consists of several stages and involves several tools. One of the tools that the company is using at the moment is Google Sheets, where the company stores its data. After getting access to the datasets, the first step of the analysis was to perform some data cleansing adjustments directly in Google Sheets. According to previous chapters, the business analytics process involves several steps such as data collection, data cleansing, data

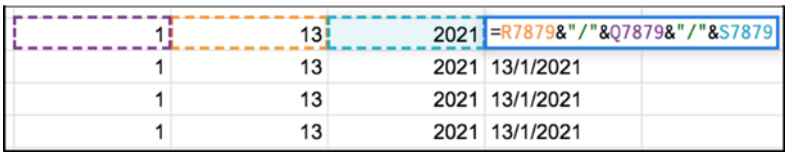
preparation, data analysis, model development, and deployment. The data for the analysis was collected by Company X in several years, therefore, this step will not be discussed in the following analysis. However, the data structure and quality require going through data cleansing and processing steps. Further, the analysis will include the actual data analysis and the application of various algorithms.

The first step of the analysis was sorting all 5 datasets by the "Date" column in order to define the common time period that can be applied to all five datasets. According to Table 3 above, the datasets have a common time period between 09.07.2020 and 29.01.2021. However, some datasets are missing data for the middle of January 2021, therefore, the common time period that would be covered by the following analysis is from 09.07.2020 to 09.01.2021. The whole analysis will be based on this period of time in order to have the consistency of the analysis. Here is defined the first barrier for qualified and accurate business analytics which is the inconsistency of dates in different datasets. Each dataset went through the data cleansing adjustments starting from the first dataset which is "Cloth feedback".

The starting point was deleting the missing row header, deleting "COGS", "Price", "ID", and "Cloth Number" columns. The values of costs and prices were the same or the costs were exceeding the price, moreover, those values were negative in some cases. In order to have a more accurate final model, those columns were excluded from the analysis since the quality of data in such columns was low, which shows another technical barrier. The "ID" column includes unique numbers for each feedback, which means that those IDs are not important for the analysis, therefore, the column should be removed. The "Cloth Number" column was excluded as well since it consisted of the numbers of each cloth in each box that could be absolutely random for every box.

While sorting the dataset by date, it was visible that the dates were expressed in different formats such as American MM/DD/YYYY and DD/MM/YYYY. The decision, in this case, was to change the format of those values that were expressed as MM/DD/YYYY to DD/MM/YYYY by using the "Date and Time" format menu. However, the function did not help in changing the format which could mean that there was a mistake not in the format but in actual value placing. In this case, the adjustments had to be done manually by separating all the values with wrong formatting by "/" and merging them again with the right order using the following function "=R&"/"&Q&"/"&S" what could be seen on Figure 2 below. After applying sorting once again, all the dates were written in the correct order. Therefore, the third challenge that was defined is the format incoherence which could damage the future model accuracy. After making all those steps, the dataset was ready to be adjusted to the common time period which is from 09.07.2020 to 07.02.2021 by removing all the unnecessary rows.

Figure 2: Manual adjustments of Dates



1	13	2021	=R7879&"/"&Q7879&"/"&S7879
1	13	2021	13/1/2021
1	13	2021	13/1/2021
1	13	2021	13/1/2021

Source: Own work.

The last adjustment of the first dataset in Google Sheets was translating all the column headers and the cells in the "Status", "Fit", and "Comments" columns. The "Status" and "Fit" columns were translated by using the "Find and Replace" function because those columns include multiple choice answers, while the comments were translated with the use of a translation function by Google Sheets. After that, the columns were filtered to check if all the values were translated.

The next "General Feedback" dataset required several data cleansing and data processing modifications as well. Firstly, the "Date" column included the time and date of every feedback, in this case, the column was split, and the "Time" column was deleted from the analysis. Moreover, the "Date" column required some format modifications as well as the same column in the previous dataset. While changing the date format, the same problem occurred which means that the format had to be changed the same way as it is shown in Figure 2. Secondly, all the columns were sorted by the date and all the rows, that are not in the interest of the following analysis, were removed. Other columns that were excluded from the further analysis were "Comments about the delivery" due to the fact that the dataset includes the "Delivery" column which consists of different ratings about the delivery that were assigned by customers, and "ID" because it includes unique values that do not add value to the analysis.

Finally, "Recommendation", "General Comment", "NPS Reason", and "Improvements" columns were translated into English, which is an essential step since some business analytics programs like RapidMiner do not support all the languages, especially for text mining. After translating the columns, the cells with missing values were replaced by nothing. Due to the fact that the recommendations and improvements include similar content, those two columns were merged into one which is called "Suggestions". Huge amounts of missing values in the datasets could also decrease the model accuracy, which could be treated as an additional challenge in this case. Some customers preferred answering only one of those questions and leaving either "Recommendations" or "Improvements" blank free, while the merged column has a reduced amount of missing values.

The third dataset "Order Status" required the smallest amount of modifications for data preparation. The column with names of customers was excluded from the analysis since such data does not make any sense in the analysis. Another step was sorting the dataset by dates and excluding those rows which included the dates that are not important for further analysis.

Thirdly, columns such as “Order Value”, “Advance PMT”, “Purchase Price”, “COGS”, and “Margin” include the values that were expressed in another currency This means that those values are supposed to be converted into euros, which was done by using the “googlefinance” function. After applying this function to those five columns, the final numbers should also be adjusted by removing extra digits after the comma and leaving only two of them.

After finishing the data cleansing and data processing steps in Google Sheets in the “General Feedback” and “Order Status” datasets, both of them were merged together in order to create additional opportunities in finding different insights in financial data and customer feedback. The merge was based on the values from the “Purchase ID” column. According to the results of the merge, 2.940 out of 2.978 rows were matching, while non-matching rows were excluded from the analysis. The new dataset is called “Feedback & Order Status”.

After reviewing and sorting the last two datasets which are “Quiz 1” and “Quiz 2”, there was made a decision to merge these two datasets into one, since they both were having several columns with common names, and only the merged dataset was able to meet the time period condition mentioned above. The fifth data issue that was discovered during the analysis is that the company was using two different types of questionnaires in order to get the same information about the customers which resulted in two different datasets instead of one. The reason behind such a decision could be the lack of expert knowledge since the company did not know how to collect the customer data in the correct way from the beginning. The first modification that was done was deleting the missing headers in both datasets, which was followed by sorting both datasets by dates and removing the rows that are not important for further analysis. Thirdly, it was important to define the common columns and the columns that are not going to be included in the merged dataset, which was followed by renaming the columns due to the fact that most of the columns in both datasets did not have short names, they were named after the whole questions from the questionnaire.

The first column was created by merging the values in the “Date” columns from both datasets. The date format was correct in both datasets, therefore, it was possible to make such step without any additional difficulties. The merged “Date” column included the exact date and time of taking a questionnaire, which made it possible to split the column into two, and create an additional “Time” column with the exact hour, minute, and second when the questionnaire was submitted. However, this data would be difficult to analyze, thus, the new column was split into three more columns such as “Hour”, “Minute”, and “Second”, where the last two were excluded from the analysis. By creating a time range ”Parts of the day” that includes such values as “Night”, “Morning”, “Afternoon”, and “Evening”, the “Hour” column was transformed into parts of the day and renamed as a new “Time” column from Figure 3 below.

Figure 3: Parts of the day

HOUR	TIME	PARTS OF THE DAY	
?	=VLOOKUP(U3,TIME,2)	0	Night
14	Afternoon	6	Morning
22	Evening	12	Afternoon
23	Evening	18	Evening
23	Evening	23	Evening
0	Night		
10	Morning		
11	Morning		

Source: Own work.

Another set of modifications had to be applied to the mobile numbers of customers which are the customer IDs at the same time. Both datasets have absolutely different formats of mobile numbers, some of them include spaces in between, some include brackets or dashes, some are written with "+" in the beginning, and others include the country special code in the beginning. Here the data format and data quality issues arise once again. The first step was to analyze what is the optimal telephone number format for most of the numbers, which is a 10 digit number that starts from 9. However, the datasets also include some foreign numbers that require some special modifications. Firstly, such symbols as pluses, dashes, spaces, and brackets were removed from the merged "Phone Number" column. The next step was splitting the column into several other columns which create 11 columns, or even more in some cases because some of the numbers were foreign with a different amount of digits. The goal was to delete the first column, in order, the second column will start from 9, and the overall number of columns will be 10. However, it was not possible because of the foreign numbers. In this case, the function "Filter" was used to find those rows with foreign or nonstandard telephone numbers that were not having 9 as a second digit. As soon as they were detected, those "special" values were transferred to the next column, which allowed the deletion of the whole first column. Such modification could also be done by using the "IF" function, however, there is a probability that a phone number will have 9 as the second digit and be a foreign one at the same time. This situation took place in this dataset, while one of the telephone numbers had 9 as the second digit and consisted of 13 digits.

After normalizing the digits, all those columns were merged together. All the customer IDs were written in the same format, which enabled the removal of all the duplicates in the dataset by "Phone Number". As a result, 151 rows with duplicates were removed from the dataset. After removing the duplicates, the column was also excluded from the analysis. The presence of duplicates in the dataset also leads to the reduced model accuracy which means that it is another important barrier in this case.

The next columns that require several modifications are the "Age" and "Birthday" columns. The first step is the removal of the word "years" in "Age" columns in both datasets "Quiz 1" and "Quiz 2", which enables defining the following age groups: 18-25, 26-35, 36-45, 46-55, >55. In the dataset "Quiz 2", the values in the "Birthday" column are absolutely random,

unstructured, and unstandardized which is an important challenge while implementing business analytics. Some of them are missing, others are missing the year of birth, some of them have months written with words, others have only the last 2 digits of the year of birth, such as 98. However, in this dataset, all the age groups are defined and there are no missing values in the "Age" column. While the situation with those columns in the dataset "Quiz 1" is completely different. The "Age" column has a huge amount of missing values, while the "Birthday" column does not have any missing values, moreover, the dates are written in the correct format. The only issue, in this case, is the human error when some years of birth were written in unrealistic forms, such as 1800 or 2028. This means that such values will be treated as missing values in the analysis. In order to fill the blanks in the "Age" column from the "Quiz 1" dataset, the age of each customer was calculated based on the day of birth, then those values were categorized into the following age groups: 18-25, 26-35, 36-45, 46-55, >55. The last step was merging the two columns and removing the "Birthday" columns from the datasets.

The next step was merging the "Occupation", "Source", "Expected Style", Weekend Style", "Job Style", "Season", "Height", "Dress Size", "Top Size", "Down Size", "Cest", "Waist", "Hips", "Hair Color", "Favorite Brands", "Unfavorable Colors", "Unfavorable Materials", "Ear Piercing", "Metal Color", and "Jewelry" columns. In the "Weekend Style" column some values are too long which is not useful for the analysis. Therefore, the first step in this column was to translate it, and replace the values "casual (jeans, tops, street style)" by "casual". The same procedure was done with the values "sporty style (comfort is over everything)" which was replaced by "sporty", and "business casual (jackets, formal dresses)" by "business-casual". Those cells that were including words like "dresses" and "skirts" were replaced by "romantic" style. Another column called "Job Style" also required some similar adjustments like translation and values replacement: "not strict dress code" was replaced by "non-strict", "strict dress code" by "strict", and "I do not have a job" by "unemployed". The issue, in this case, was that the values are too long for the analysis. After translating the "Season" column, a translation issue has occurred. The values with the autumn season were translated as "fall" or as "autumn". In this case, all the "fall" values were replaced, and the whole column was transformed to lowercase values in order to have more consistency.

While merging the columns that include body parameters, it has been discovered that dataset "Quiz 1" includes duplicates of the following columns: "Cest", "Waist", "Hips", and "Hair Color". If the value about a customer is missing in one column, then it is available in another with the same name, therefore, all those columns were merged into single ones and then added to the "Merged Quiz". Moreover, the "Cest", "Waist", and "Hips" columns were also required to remove "cm" or "centimeters" because some responders were specifying these measures. The main issue, in this case, is the fact that the format of answers in the questionnaires was polynomial, which causes additional issues in the analysis. The "Hair Color" column also required some additional adjustments. In "Quiz 2" some of the cells in the "Hair Color" column included two values, for example, "dirty blond, blond" or "ginger,

red". While the same column in "Quiz 1" included only one-word values. Therefore, the same logic had to be applied for the merged dataset, by separating the "Hair Color" column by comma and leaving only the first words in the column, which made the whole column more consistent. Moreover, another issue was related to the translation of the entire column to the English language by using the "google translate" function. The "dirty blond" color was translated as "blond", and "light blonde" as "blonde", which could be misleading during further analysis. Therefore, such values had been renamed.

While translating the merged "Unfavorable Color" column the problem with translation occurred once again. The color of the sea was translated by the program in several different ways: marine, sea, nautical, and offshore. In this case, after translating the whole column, all those translation versions were replaced by the one "sea". The same problem has occurred with the maroon color, which was also translated as vinous, marsala, and burgundy in some cases. Acidic colors were translated as acid or acidic. Some customers responded, "all" or "any", which could mean that they did not understand the question, meaning that they accept all the colors. Therefore, these kinds of values were replaced with "no", meaning that a respondent does not have any unfavorable colors. In some cases black color was written with "the" in front, in this case, all "the" were removed from the column. The last step is to transmit all the values to lowercase because some values were translated with uppercase and some not.

The next column that also requires the translation is "Unfavorable Material". Some respondents were replying "leather" and others "natural leather", as well as with fur. In this case, all the "natural leather" values were replaced by "leather", and "natural fur" by "fur", the values for non-natural materials there are the following: "faux-leather" and "faux-fur". The values "no" are expressed in many different ways: "no such", "all are okay", "any", "not sure", "no such material", "all are ok", "nothing", "no one", "all are allowed", "love all" and others. In this case, all of them had to be replaced by a simple "no" to make the analysis easier. The next step is to merge the "Ear Piercing" column. The trick, in this case, is that in the "Quiz 2" dataset, the values are expressed as "yes" and "no" while in the "Quiz 1" dataset, the values are expressed as "true" and "false". When merging the columns, all the "yes" and "no" answers were replaced with boolean "true" and "false" by using the function "Find and Replace". A similar situation is with the last "Jewelry" column, however, in this case, the "yes" and "no" answers were replenished with "maybe", while the same column in "Quiz 1" consists of binary "true" and "false" values. In the "Merged Quiz" dataset, "maybe" and "yes" answers will be treated as "true", and "no" as "false".

Table 4 below represents the information about all three datasets that came as a result of data cleansing and data processing in Google Sheets. The table can be compared to Table 3 above in order to see the substantial changes. For instance, the number of datasets decreased from 5 to 3, considering the fact that the data inside stayed the same. Another important aspect is that the columns that did not bring enough value were removed from the analysis.

Table 4: Database after the data cleansing and data processing in Google Sheets

Dataset	Period	№ of rows	№ of columns	Columns
Cloth Feedback	09.07.2020-09.01.2021	15.064	9	Date, Cloth ID, Purchase ID, Status, Fit, Quality, Price, Style, Comments.
Feedback & Order Status	09.07.2020-09.01.2021	2.940	18	Date, Service, Style, Overall Rating, Delivery, NPS, New Stylist, General Comment, NPS Reason, Suggestions, Shipping Date, Purchase ID, Order Value, Advance Payment, Price, Clothes Sent, Clothes Purchased, COGS, Margin.
Merged Quiz	09.07.2020-09.01.2021	2.489	23	Date, Customer ID, Birthday, Age group, Expected Style, Height, Favorite Brands, Email, Facebook, Instagram, Occupation, Dress Size, Comments, Parameters, Body Specialties, Undesirable material, Ear Piercing, Address, and others.

Source: Own work.

Another essential step is the analysis of these three datasets in another tool for Business and Data Analytics which is called RapidMiner. Before performing some text mining, descriptive or predictive analytics, the database should be completely ready for further analysis. Therefore, the first operator that is added to the process is "Read Excel" in order to read the first dataset "Cloth Feedback". The dataset requires some adjustments, for instance, changing the datatype for "Status" from polynomial to binomial.

Another operator that should be added to the process is the "Quality Measures" operator that helps to decide which column to remove and which to leave for further analysis. The operator provides information about ID-ness, Stability, Missing, and Text-ness for all attributes. The higher the ID-ness the less important an attribute is, the same rule applies to Stability if the number is too close to 1. For instance, the highest stability has "Status" that equals 0,7, and "Fit" attribute that equals 0,68, however, both of the attributes are important for further analysis. Moreover, their ID-ness equals 0. The highest ID-ness is for the "Comments" which equals 0,38 but this number is not too close to 1 which means that the attribute can be used for further analysis. The higher the Text-ness, the less probability that an attribute is going to be useful for the analysis, which means that an attribute will require some text mining later on. In the case of the "Cloth Feedback" dataset, the highest Text-ness indicators have the "Comments" attribute which is 0,98. The "Missing" quality measure shows the percentage of missing values for each variable. The higher the number of missing values, the better to remove such variables. In the case of the first dataset, there are no such attributes.

After removing the “Quality Measures” operator, the first type of data analytics that can be applied to the dataset is descriptive analytics with the use of visualizations. One of the main business opportunities, in this case, could be the possibility to see visually what is happening in the company at the moment in order to speed up the decision-making process. Firstly, it is applied on the “Cloth ID” column that shows the most popular cloth for stylists. However, the attribute data type is polynomial since some IDs include letters. In this case, after the dataset is read, the “Replace” attribute is added in order to remove all the letters from the whole column, then the dataset is saved and retrieved with a new data type for the “Cloth ID” column, which is integer now. The issue, in this case, is the potential lack of expert knowledge about the most useful data types and their roles for the analysis. Secondly, the “Status” attribute was grouped by “Status” with the use of the “count” aggregation function in order to see the number of sold and returned items. Thirdly, such attributes as “Quality", “Price”, and “Style” are grouped by “Status” with the “average” aggregation function.

The next step of the analysis is to open the visualizations for the “Date” attribute with 6 bins in order to find out in which month the company has received the highest amount of cloth feedback from customers. Fifthly, the “Fit” attribute was aggregated by the “count” function, which provides the visualizations about the most frequent fits among customers, which could help the company to determine if it is using the right size measures. Thereafter, the “Quality”, “Price” and “Style” ratings were analyzed to the average measures.

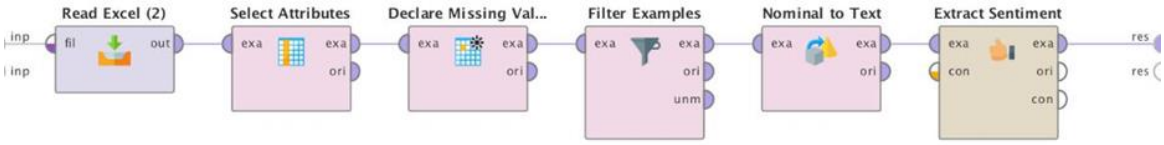
Descriptive analytics can also be applied to the text, namely to the “Comments” to see the most frequent comments and the fact if they are mostly positive or negative, which helps to determine customer attitude towards the brand. In order to detect the most frequent comments, the model is built the way it is represented below in Figure 4. Firstly, the “Comments” attribute type is changed to text, otherwise, the next operator will not be able to proceed with the data. Secondly, the “Process Documents from Data” operator is added to the process in order to generate the word vectors from string attributes. The subprocess includes the operators that create tokens in one single word, transform the text to lowercase, remove English stop words, filter tokens based on their length, remove word suffixes in order to reduce the word length, and others. The subprocess shown in Figure 4 below, includes the following operators: “Tokenize”, “Transform Cases”, “Filter Stopwords (English)”, “Filter Tokens by Length”, “Stem (Porter)”, and “Generate n-Grams”.

Figure 4: Defining the most common words in “Comments”

Source: Own work.

While the process discussed above is focused on defining the most common words and phrases, the process presented in Figure 5 below focuses on calculating the negativity and positivity scores of the comments. In the "Select Attributes" operator only the comments were selected due to the fact that other attributes were not important for the following part of the analysis. With the use of the "Declare Missing Values" operator, it was possible to show the program how to detect a missing value, while the "Filter Examples" operator helped to remove all the missing values from the analysis since the "Extract Sentiment" operator treats missing values as a normal value and assigns sentiment to each of them what reduces the accuracy of the analysis. The "Nominal to Text" attribute was used in order to transform comments to the text. Finally, the "Extract Sentiment" attribute was used in order to detect the positivity and negativity score of each comment.

Figure 5: Extracting Sentiment from the comments

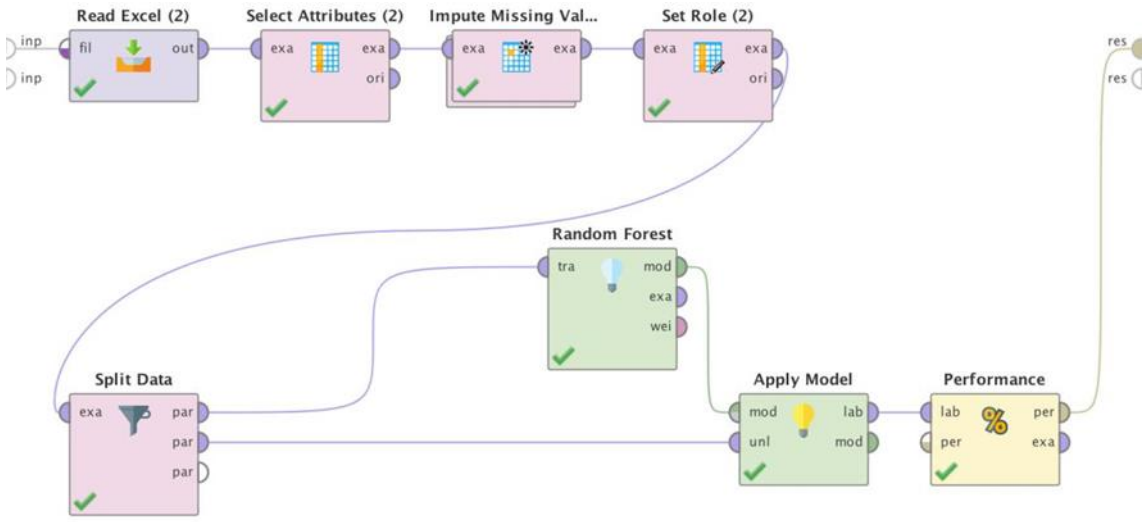


Source: Own work.

The next step is related to predictive analytics because the final model will be focused on predicting if a cloth is sold or returned, which creates several opportunities for the business. Knowing the most important factors that influence the purchase or return, the company can easily determine its internal strengths and weaknesses to improve the service. Firstly, it is important to select the most valuable attributes for the analysis, which are the following: "Fit", "Price", "Quality", "Status", and "Style". Secondly, the model required some missing values adjustments of missing values for the better accuracy of the model. "Impute Missing Values" is a subprocess that could include several operators. The imputation, in this case, is based on the fact that the program learns from the cases with no missing values, and replaces the missing values with the results. The subprocess includes the "k-NN" operator that chooses 5 nearest neighbors in order to fill the missing value. The biggest problem related to this operator is the amount of time it takes to finish the process. Thirdly, it is important to set the role of the attributes by using the "Set Role" operator and assigning "Status" as a label. Fourthly, the dataset should be divided into training and testing sets. There are several approaches to split the data such as 80:20 ratio, 70:30 ratio, or a k-fold splitting which is 9:1. For the analysis, the data was split the following way: 0.7 for the training set, and 0.3 for the testing set. Fifthly, it is important to choose the best algorithm that provides the highest accuracy. As the model is built under supervised learning, where the dependent and independent variables are predefined, the Random Forest algorithm is used for the analysis. The parameters of the algorithm are assigned the following way: 10 decision trees, gain ratio criterion, maximal depth of 5, pruning, and the confidence of 0.1. The result consists of 10 individual decision trees that operate as an ensemble, where the predictions made by the individual trees have low correlations with each other. As the training set is connected to

the“Random Forest” operator first, the testing set is connected to the “Apply Model” operator directly in order to get the prediction on the unseen data. The last operator in the model is “Performance Binomial Classification” which is used to statistically evaluate the strengths and weaknesses of binary classification after the training model has been applied to the data.

Figure 6: Predictive Modeling for “Cloth Feedback” dataset



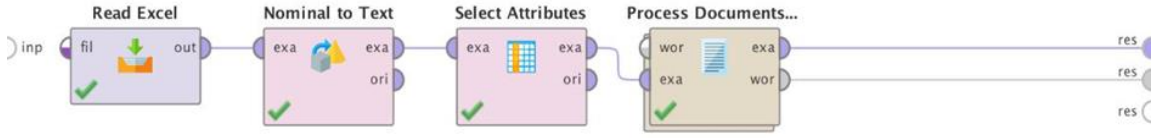
Source: Own work.

The second dataset that is going to be analyzed in RapidMiner is “Feedback & Order Status”. The procedure, in this case, is partially similar to the previous dataset. Firstly, the dataset is read by adding a “Read Excel” operator. Some data types require several modifications, for instance, “New Stylist” should be changed from polynomial to binary, since it includes only two types of values: “yes” and “no”. After reading the dataset, the operator is connected to the output side of the process in order to check the statistics and visualizations. These two components help to determine the average number of clothes sent, clothes purchased, average service rating, style rating, overall rating, delivery, and average NPS. All those measures help the managers to understand the current state of the company. Moreover, by using visualizations, it is possible to see the relationships between the “Style”, “Overall”, “Service” and “New Stylist". Another visualization shows the relationships between those three ratings and the number of items purchased from one box, the aggregation function used in this case is “average”.

Another visualization shows the number of purchased items grouped by the “New Stylist”. The data is aggregated, and the aggregation function used for the bar chart visualization is the average function. The last set of visualizations show the company’s profit by date using the “sum” aggregation function, profit by “New Stylist” using average aggregation function, and profit by “Service”, “Style”, and “Overall” ratings using “sum” function.

The next step of the analysis is to apply text mining to the dataset, particularly to the "NPS Reason", and "Suggestions" attributes. In order to define the most common reasons why the customers assign low or high NPS, and what they suggest to improve the service quality, the following process from Figure 7 was created. The business opportunity that could be discovered in this case is the fact that the company finds out about its weaknesses from the customers. Firstly, the "NPS Reason" column was processed by the following operators: "Nominal to Text", which transforms the values to text, "Select Attributes" which filters the attributes for further analysis, and Process Documents from Data, which is a subprocess. The Process Documents from Data subprocess includes the operators that create tokens in one single word, transform the text to lowercase, remove English stop words, filter tokens based on their length, and others. As a result of the analysis, we get the most common words that clients were using in the "NPS Reason" column. After receiving the results, the "Suggestion" column went through the same process which is shown below in Figure 7.

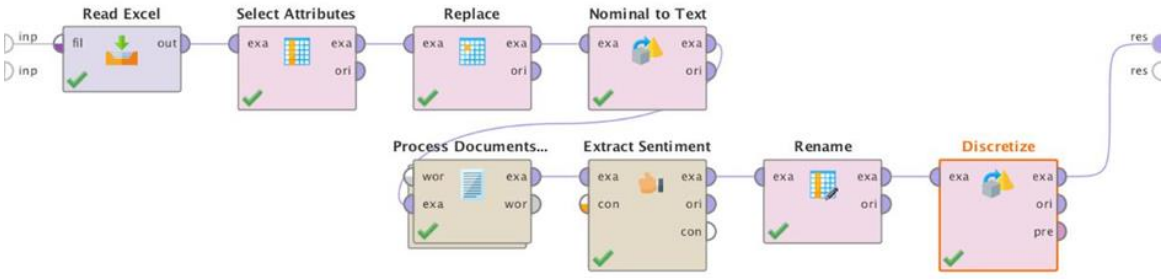
Figure 7: Text Mining of "NPS Reason" and "Suggestions"



Source: Own work.

Different text mining techniques are applied to the "General Comment" attribute. In this case, the "NPS Reason", and "Suggestions" attributes are unselected, all the punctuation characters are replaced by nothing, the values from the "General Comment" column are transformed into text, the text is processed in the Process Documents from Data subprocess the same was as it was processed for the "NPS Reason", and "Suggestions" attributes, the sentiment is extracted by using the VADER model, the result score column is renamed and discretized by user specification.

Figure 8: Text Mining of "General Comment"



Source: Own work.

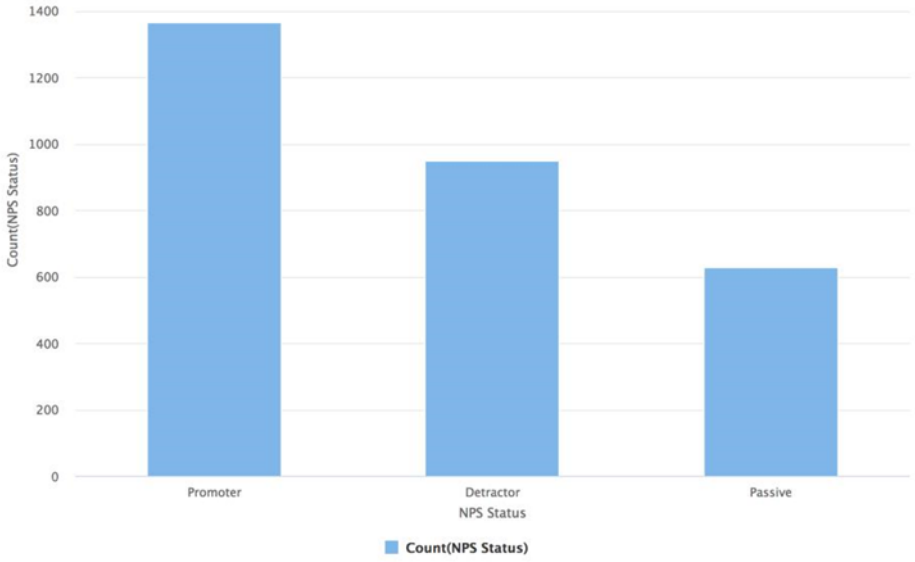
The outcome of extracting sentiment with the VADER model is the score which is measured on a scale from -4 to +4, where -4 stands for the negative sentiments, 0 for neutral, and +4 as the most positive. The outcome score column is renamed to "GC Score" and discretized by three parameters: negative, neutral, and positive. The negative category includes all the values up to 0, the neutral category includes all the values that are equal to 0, and the positive category includes all the values which are greater than 0.

After performing some descriptive analytics and text mining, the next step is to evaluate the attributes that can be done by using an operator that was used before which is "Quality Measures". The operator provides information about ID-ness, Stability, Missing, and Text-ness for all 17 attributes. The highest ID-ness indicators have "Purchase ID" which equals 1 and "Order Value" which equals 0.855. The highest results in the stability measure have two attributes: "Clothes" that equals 0.809, and "New Stylist" attributes that equals 0.718. These indicators mean that those variables should be removed from the analysis. In the case of the "Feedback & Order Status" dataset, the highest Text-ness indicators have "General Comment", "NPS Reason", and "Suggestions" attributes, which are 0.993, 0.990, and 0.994. The "Missing" quality measure shows the percentage of missing values for each variable. The higher the number of missing values, the better to remove such variables. In the case of this dataset, "NPS Reason" has the highest amount of missing values which is expressed with 0.568.

After checking the measures, the "Quality Measures" operator is replaced with the "Select Attributes" operator, where "Clothes", "COGS", "Date", "NPS Reason", "Order Value", "Purchase ID", "Suggestions", and "General Comment" attributes are not selected for the future model development, while "GC Score" is added to the analysis. The next operator in the model that is presented in Figure 10 below, is a "Set Role" operator that changes the roles of "text" and "GC Score" attributes from target to regular, and by using the next operator, the "text" attribute is removed from the analysis.

The eleventh operator in the process generates a new attribute called "Profit", by using the following expression: "if(Purchased>0, "true", "false"), according to which the column shows if an order is profitable or not. By using a "Generate Copy" attribute, there are two identical columns created in the dataset which are "NPS" and "NPS Status". Hereinafter, the new "NPS Status" attribute is discretized into user-specified classes such as "Detractor" with the upper limit of 6, "Passive" with the upper limit of 8, and "Promoter" with an upper limit of 10. Such discretization helps to make different customer segments and evaluate their potential behavior towards the brand. Figure 9 below represents the share of each type of customer in accordance with the "NPS Status" column.

Figure 9: The number of Promoters, Detractors, and Passive customers

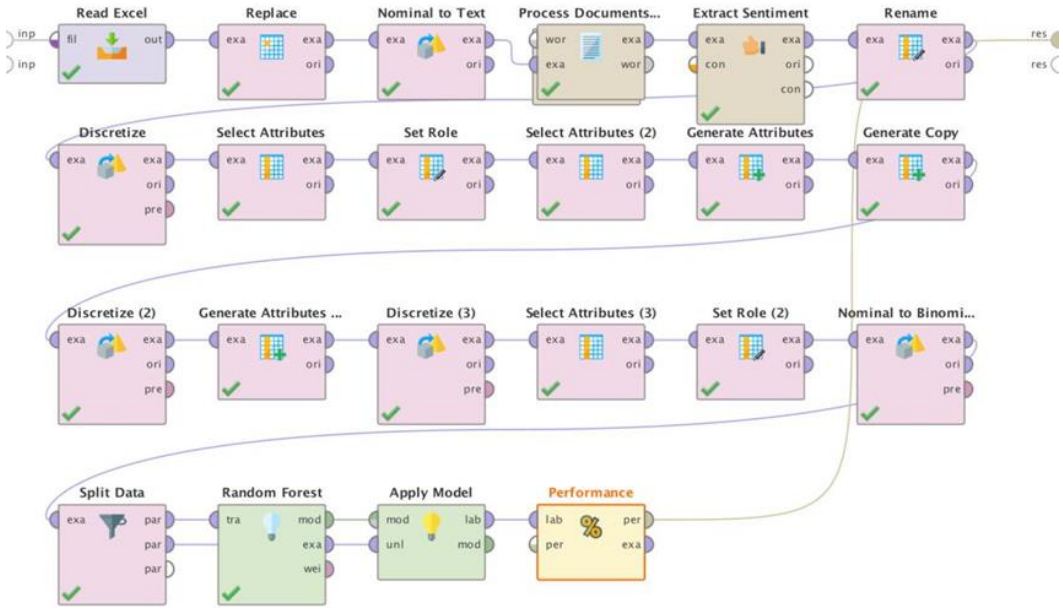


Source: Own work.

The "Generate Attributes (2)" operator constructs a new attribute called "Promote" by using the "if" function, which creates a nominal attribute that is going to be changed to a binomial. This attribute shows if a customer is willing to promote the brand or not, what are the essential parts of the brand reputation and further marketing. The third discretization, which could be seen from Figure 13 below, is added to the process of model development in order to discretize the "Delivery" attribute into other numerical classes: "1" with upper limit 2, "2" with upper limit 4, "3" with upper limit 6, "4" with upper limit 8, and "5" with upper limit 10. This allows the variables to look more consistent since the highest class in "Overall", "Service", and "Style" attributes is "5". The inconsistency of data is one of the challenges in this case. While selecting the attributes one more time, the "NPS" and "Purchased" columns are excluded from the analysis, since their values are the parts of other new attributes.

The next step is setting the target role for the "Profit" attribute because the model will predict if a condition is beneficial for the profit generation or not, while "Nominal to Binomial" changes the type of the attribute. The nineteenth attribute splits the data into subsets in accordance with the 70:30 data split, where 70% of data is used for the training set and 30% for testing. While choosing the right algorithm, the model was trained with the Decision Tree algorithm first, however, one tree was not able to provide enough information. Therefore, the final model is based on the "Random Forest" algorithm that generates 15 trees with a maximal depth of 6, based on the gain ratio criterion. The last two operators are used in order to apply the model and evaluate its performance, which is essential since it should be accurate to create some value.

Figure 10: Predictive Modeling for "Feedback & Order Status" dataset



Source: Own work.

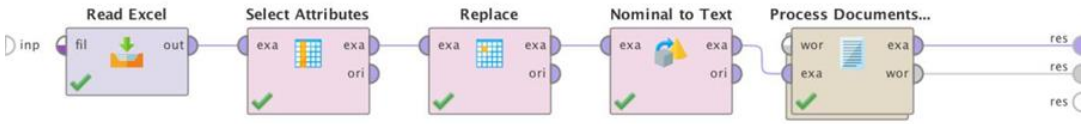
Another model that is created by using the same process, predicts if a customer will be a promoter of the service or not, which helps to understand customer needs better, analyze different customer segments, and find a way to target each of them correctly. In this case, in the "Select Attributes (3)" operator, the "NPS Score" attribute is excluded from the analysis, the "Promote" attribute is assigned to have a label role in the process, and the attribute type is changed to binomial. The total number of trees is changed to 10, with a maximum depth of 5. Pruning is as well applied in the algorithm since it reduces the model complexity by replacing sub-trees with leaves.

The third dataset "Merged Quiz" represents the data about the customers. The first step of analyzing the dataset in RapidMiner is reading the dataset and assigning the right data types such as "Telephone Number" from real to integer, "Ear Piercing" and "Jewelry" from polynomial to binomial. The next step is adding the "Quality Measure" operator in order to see which attributes should be excluded from the analysis. According to the measures, the biggest amount of missing values that equals 0.65 have the "Season", "Top", "Dress", and "Down" columns. The highest stability of 0.92 has the "Ear Piercing" attribute, which means that 92% of customers have ear piercings. The highest textness of 0.7 belongs to the "Brands" column, which means that it requires some text mining.

After checking all the measures for each attribute and disconnecting the operator, the "Select Attributes" operator is added to the analysis to exclude the following attributes: "Date", "Phone Number", "Season", "Top", "Dress", and "Down". The "Top", "Dress", "Down", "Season" and "Ear Piercing" because of the results of the "Quality measures", while the "Date" attribute is excluded because of the fact that the analysis covers only half of the year.

The “Replace” operator is used for a subset of attributes that require text mining; all the punctuation characters except one were replaced by nothing. The only character that was not replaced is “&” because some brands include this sign in their names. In order to analyze such attributes as “Brands”, “Job Style”, “Unfavorable Colors”, “Unfavorable Materials”, and “Weekend Style”, their type has to be changed to text, which the “Nominal to Text” operator does. The next step, according to Figure 11 below, is generating word vectors from string attributes, by adding a subprocess “Process Documents from Data”. The subprocess includes similar operators that were used before: “Tokenize”, “Transform Cases”, “Filter Stopwords (English)”, “Filter Tokens”, “Stem (Porter)”, and “Generate n-Grams”. The roles of each operator were explained before during the analysis of another dataset.

Figure 11: Text Mining for “Merged Quiz”



Source: Own work.

The first attribute that goes through the text mining process is “Brands”; the result of such a process is a list of the most popular cloth brands among the customers. By specifying other attributes and running the process several times, we also get a list of the most common job styles, weekend styles, unfavorable colors, and unfavorable materials. The maximum amount of words for “Job Style”, “Unfavorable Materials”, and “Brands” attributes was 2, while “Unfavorable color” and “Weekend Style” attributes were focused only on one-word values. In order to be able to describe other measures, the “Size” column should be added to the analysis. “Top”, “Dress”, and “Down” attributes were including the cloth sizes of customers, however, all three columns were including a huge share of missing data, which is one of the challenges in data analytics. Therefore, the new size column is based on three parameters: chest, waist, and hips. The definition for each size is based on the size chart values of the most favorable cloth brand among customers that was defined before, during the text mining. According to the size chart from Table 5 below, the new attribute “Size” is based on the following expression: "if((Chest<83)&&(Waist<63)&&(Hips<91), "XS", ...".

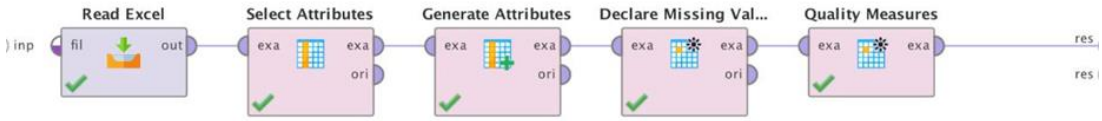
Table 5: Size Chart in centimeters

	XS	S	M	L	XL
Chest	82	86	90	96	108
Waist	62	66	70	76	88
Hips	90	94	98	104	116

Source: ZARA (n.d.).

By adding the "Declare Missing Values" attribute, the question mark is treated as a missing value. The last step in the process is checking the quality measures one more time in order to check the number of missing values in the new "Size" column. The missing measure for the "Size" attribute is 0.4, which is lower than it was for the "Top", "Dress", and "Down". Therefore, the final process is created and presented in Figure 12 below; the dataset is ready for performing some descriptive analytics. After removing the "Quality Measures" and creating a connection with the output, it is possible to evaluate the statistical indicators and visualizations in order to see the current state of the company: "Time", "Occupation", and "Size" by "count" function; "Source", "Expected Style", "Age", "Hair Color", "Metal" and "Jewelry" by distribution; "Height", "Chest", "Waist", and "Hips" by "mode" function.

Figure 12: The final process of "Merged Quiz"



Source: Own work.

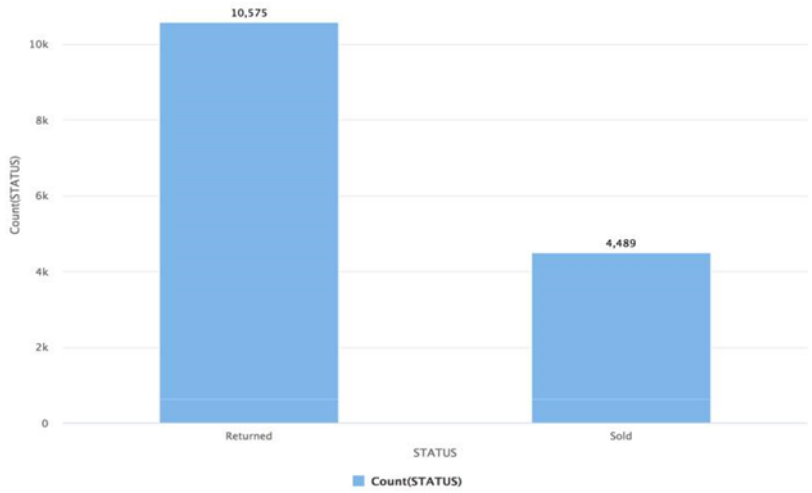
5 RESULTS

There are several data insights that have been discovered during the analysis of the company's data, particularly three main datasets. The following chapter complements the previous findings and provides additional opportunities and barriers of using business analytics in small enterprises.

5.1 Models and Findings

The first and the only initial dataset that has been analyzed is "Cloth Feedback". After the data cleansing and data preparation steps, the first type of data analytics that has been applied to the dataset is descriptive analytics. Firstly, it has been applied on the "Cloth ID" column that shows that the most popular cloth for stylists is the cloth with ID 4631152135329, which was sent 55 times. Another most popular cloth used by the stylists is ID 4631152135350 that was sent to customers 51 times. However, while grouping the cloth IDs by the cloth status, using the mode, the results showed that the most "returned" cloth has ID 4631152135329, while the most "sold" cloth has ID 4631152135350. This information helps to see which clothes are not preferable by the customers in order to stop sending the items that are not successfully sold to customers. Secondly, the "Status" attribute was grouped by "Status" with the use of the "count" aggregation function in order to see the number of sold and returned items, which is presented in Figure 13 below. The bar chart shows that around 30% of clothes were sold, while 70% of all the clothes sent in the period between 09.07.2020 and 09.01.2021 were returned back, which helps to evaluate the company's success.

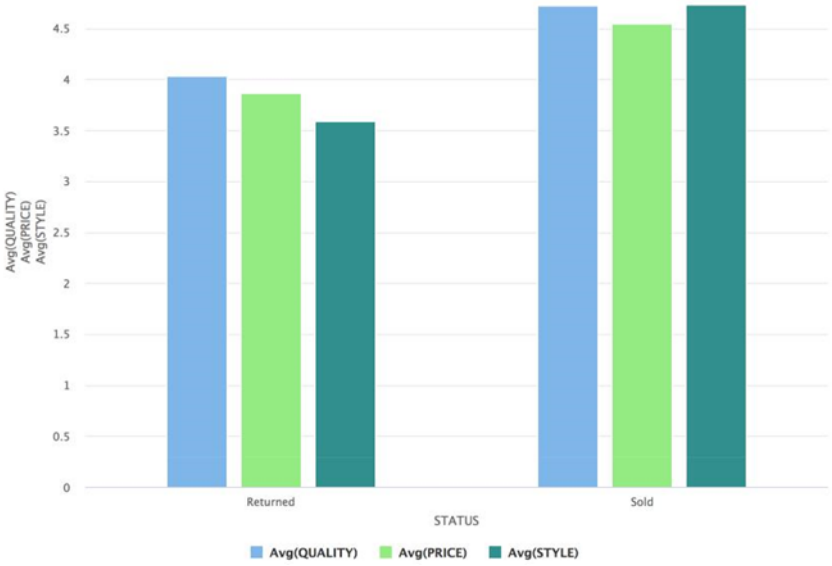
Figure 13: Number of Returned and Sold items



Source: Own work.

Thirdly, such attributes as “Quality", “Price”, and “Style” were grouped by “Status” by using the “average” aggregation function, which is presented on the bar chart below. The bar chart shows that on average people who were buying items, were evaluating “Quality” and “Style” higher than “Price”, which means that these two indicators were the most important while buying an item. While the most important indicator that influenced the item return the most was “Style”, which is shown in Figure 14 below. This information helps the company to assign its priorities that influence the profit generation.

Figure 14: Quality, Price, and Style are grouped by Sold and Returned items



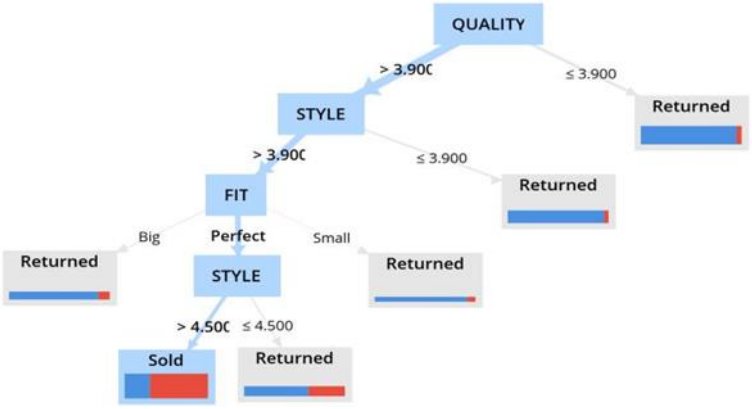
Source: Own work.

According to the dataset statistics, the biggest amount of cloth feedback was received in November 2020 which equals 3 376, while the lowest amount of customer feedback was received in July 2020 which equals 1 491. Such results could be influenced by various factors that could be analyzed by the managers in order to create additional business opportunities. The most frequent "Fit" value is "Perfect" which represents 68% of values that equals 9 373, the next one is "Big" with 3 015 values and "Small" with 1 398 values. The opportunity, in this case, could be the development of a more accurate size chart that will help to reduce the amounts of too big or too small fits. The average "Quality" rating is 4.25, the "Price" rating is 4.07, and the average "Style" rating is 3.94 which is the lowest among these three, which means that this parameter can be improved in the future.

The last column that was analyzed by descriptive analytics and text mining was "Comments". The results of the analysis show that the most frequently used comments are related to the cloth's "color" which was mentioned 1 833 times, "size", and "style", which means that these are the important indicators for customers. Therefore, we can notice that in this case business analytics assists in detecting key customer preferences while using the service. Additionally, the word "similar" and such combinations of words as "similar item" or "have similar" appeared more than 315 times, which could mean that customers already have some clothes that look alike with those that are sent in a box, which shows the lack of communication between customers and stylists. Another text mining technique that was used for this attribute was extracting the sentiments from comments. According to the model shown in Figure 5 above, the average comment score according to the VADER model is 0.443, average negativity is 0.142, and average positivity is 0.585. The score is measured on a scale from -4 to +4, where -4 stands for the most negative comments, 0 for the neutral ones, and +4 for the most positive comments. Such results show that the biggest share of comments provides positive feedback from customers, while on average all the scores are close to being treated as neutral. The technique could be useful for the frequent service performance evaluation, which provides an additional opportunity for the business.

The goal of predictive modeling for the "Cloth Feedback" dataset was to build a model that predicts if a cloth is sold or returned with an accuracy higher than 80%. Predictive modeling in this particular case brings several additional opportunities and barriers. The main barrier is the fact that predictive modeling requires additional expert knowledge in order to choose the best algorithm and build a model that is able to add some value. The opportunity of this model is that by knowing the condition in which an item can be sold or returned, the company can improve its performance and decrease the number of returned items. The actual process is discussed in Figure 6 above. According to the performance results of the model, the model accuracy equals 82.15%, and the recall of returned items equals 86.04% which represents the high quality of predictions related to the returned items. The classification error of the model equals 17.85% which represents the amount of false positive and false negative responses. The actual process produces 10 different decision trees that provide several data insights.

Figure 15: The model for "Cloth Feedback"



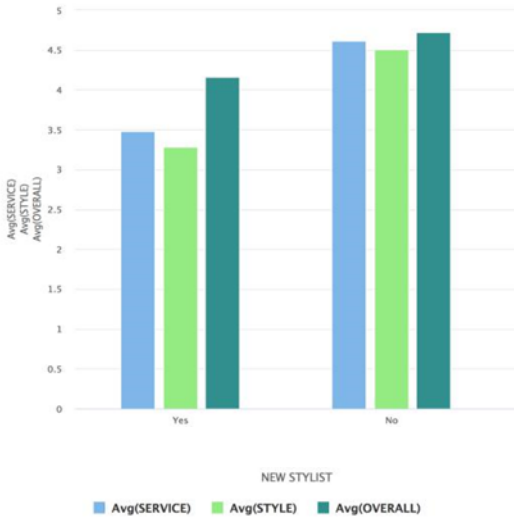
Source: Own work.

One of the trees presented in Figure 15 above shows the conditions in which an item is more likely to be returned. For instance, 2 413 items were returned and only 126 were sold in the case when the customers were evaluating the quality for less or equal to 3.9. The same condition applies to the style of items when customers are more likely to return an item if they assign less than 3.9, even if they rated the quality higher than 3.9. Even if the ratings for style and quality are higher than 3.9 but an item is too big or too small, there is a high probability that it will be returned back. An item has a 70% probability to be sold if it meets the following condition: quality rating higher than 3.9, perfect fit, and style rating higher than 4.5. While if all the conditions are met but the style is rated between 3.9 and 4.5, then the item is more likely to be returned.

According to other trees, the biggest influence on the item return is the "Style" attribute. The results of another tree show that an item has a 96.77% probability to be returned if the style is rated less than 3.9 out of 5. The descriptive and text analytics that were described above also showed that one of the most important indicators for customers is style. This brings an opportunity for the company to adjust its strategy and pay more attention to this indicator. Another important observation is that an item is more likely to be sold if it has a perfect fit. The "Price" has the lowest influence on the model. The most common result for this attribute is that the price should be rated for more than 1.5 in order for an item to be sold.

The next set of results belongs to the second merged dataset which is "Feedback & Order Status". The first step of the analysis was related to descriptive analytics with the use of statistics and visualizations. The statistics show that the average number of clothes sent is 6 and the average number of clothes purchased is around 2. The average service rating is 3.5/5, the style rating is 3.4/5, and the overall rating of 4/5. The average delivery rating is 7.7/10 and the average NPS of 7/10. While checking the relationships between the "Style", "Overall", "Service" and "New Stylist", the bar chart shows that people who work with the stylist that they already know, give higher ratings for all three criteria, in comparison to those who work with the new stylist, what is shown on Figure 16 below.

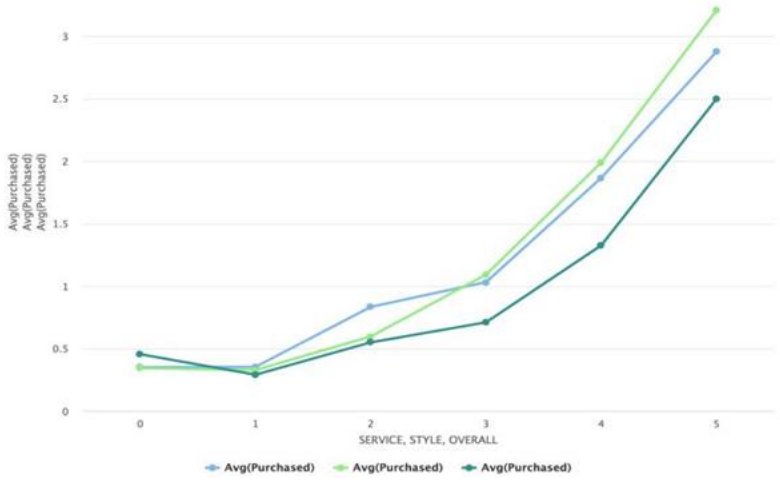
Figure 16: Service, Style, and Overall ratings grouped by "New Stylist"



Source: Own work.

The second visualization shows the relationships between those three ratings and the number of items purchased from one box. According to Figure 17 below, the biggest amount of items purchased have those people who assign 3 or more for the style of clothes, while for the ratings less than 3, the biggest influence on the purchased items have the service rating. This creates several opportunities for the further decision-making process, such as working on the style improvement, since it is one of the most crucial indicators for the customers.

Figure 17: "Purchased" grouped by Service, Style, and Overall ratings

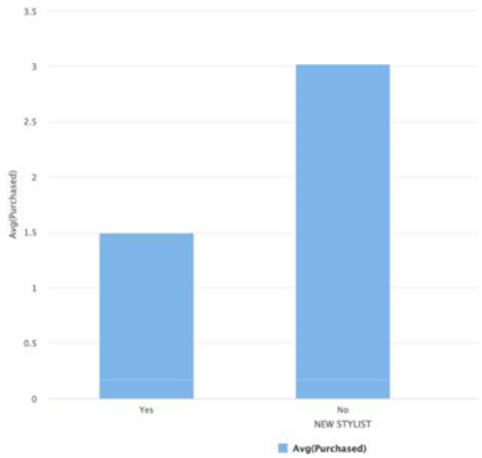


Source: Own work.

Another visualization shows the number of purchased items grouped by the "New Stylist" attribute. The data is aggregated, and the aggregation function used for the bar chart visualization on Figure 18 below is the "average" function. The visualization shows that people are more likely to buy an item if they already know the stylist, which shows that

customers would like to know the stylist better before they start working with him or her, which creates additional opportunities for decision making processes regarding the prior communication between stylists and customers.

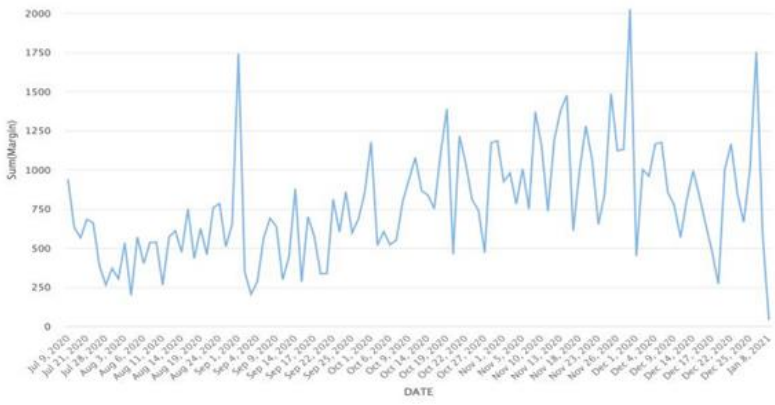
Figure 18: Purchased items by "New Stylist"



Source: Own work.

The last set of visualizations represent the company's profit by other attributes. For instance, while grouping the average profit by the "New Stylist" attribute, the result of the visualizations was similar to the bar chart from Figure 18 above. The average margin of using the new stylist is around EUR 26.2, while the average margin of using the "old" stylist is around EUR 43.13. Another profit-related visualization is shown in Figure 19 below, it represents the profit function by date. According to the function below, a slight increase in profit starts at the beginning of October 2020 and ends in the middle of December. The peak was achieved on November 30, 2020, bringing around EUR 2 027. Such visual representation helps to evaluate not only customer data but also the financial data that can support the decision-making regarding profit maximization and other financial aspects.

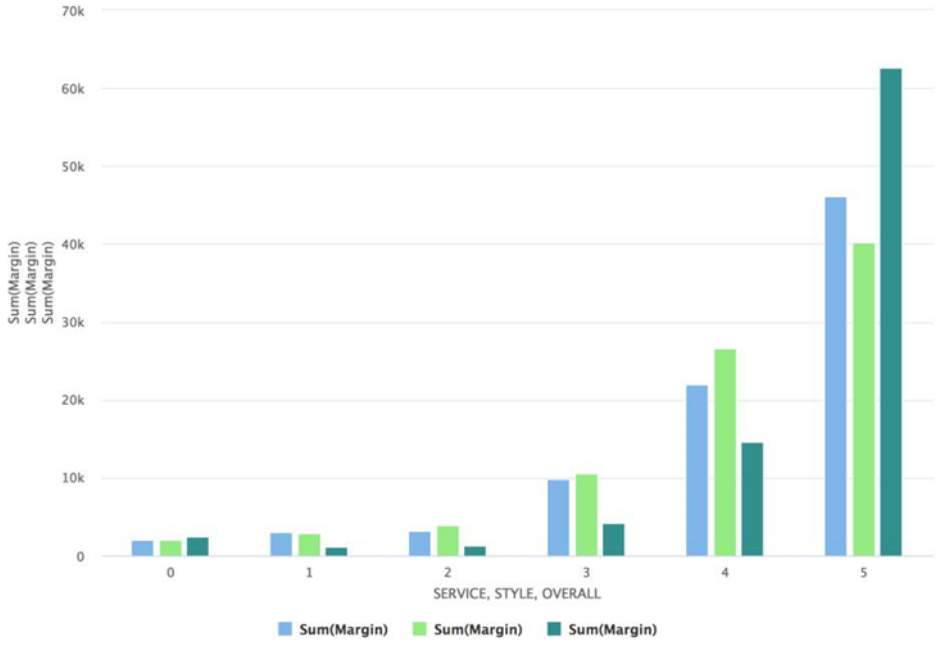
Figure 19: Profit curve by Date



Source: Own work.

The last visualization related to the "Margin" attribute is presented in Figure 20 below. The attribute was grouped by "Service", "Style" and "Overall" ratings using the "sum" aggregation function. The visualization shows that the higher the ratings, the higher the margin. In the "rating 5" group, the biggest influence on the margin has the "Overall" rating, while in the "rating 4" group the biggest influence has the "Style" rating.

Figure 20: Margin by "Service", "Style" and "Overall" ratings



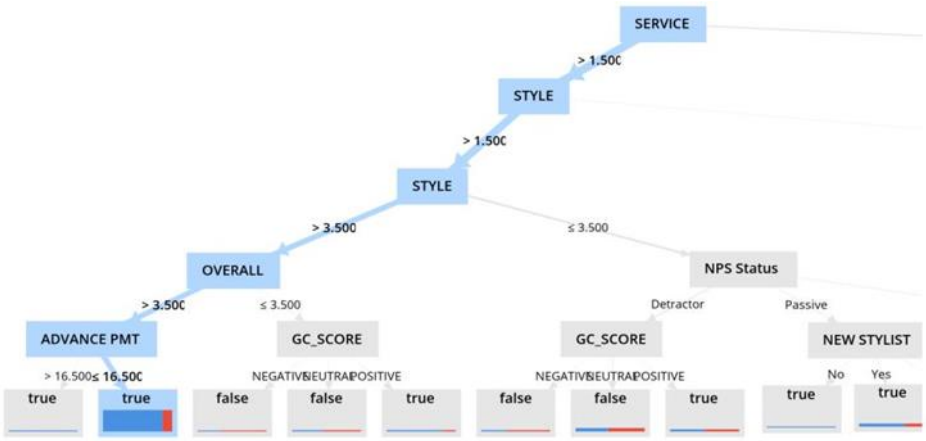
Source: Own work.

The next technique that has been applied to the "Feedback & Order Status" dataset was text mining. Firstly, the "NPS Reason" attribute was analyzed to find the most common reasons for assigning low or high scores. The most commonly used words are the following: service, box, time, stylist, and size. These words show the most important aspects of the service that influence customer's further willingness to promote or not to promote the brand.

While analyzing the "Suggestions" attribute, the most common values or phrases are missing instructions, missing recommendations, stylist instructions, image, image examples, notes, and communication. Such results of text mining show that most of the customer suggestions are aimed at improving the stylists' recommendations regarding the way how to wear the box items and with what to combine, the customers are missing more communication with the stylists regarding the further instructions. Therefore, the most important opportunity that text analytics brings is the detection of business issues that are hard to define. However, there are several barriers regarding text analytics implementation: the need for expert knowledge, the access to the tools for text mining such as RapidMiner, and others.

The predictive model of the dataset is focused on predicting two parameters. The first model predicts the conditions in which the company is able to make any profit, even the minimal. Knowing the condition for profit generation, the company can adjust its business processes and strategy in order to maximize profit. The model accuracy equals 80%, and the classification error equals 20%. The class recall for prediction of having profit equals 92.7%. The result of the process discussed before is 15 different decision trees. One part of the first tree is presented in Figure 21 below, where the “Style > 1.5" subtree represents 81.1% of values. The first tree shows that a profitable transaction will take place if a customer evaluates the service for more than 1.5, the style for more than 3.5, as well as the overall rating, and pays in advance less than EUR 16.5. The second tree shows that the condition is true if the style is rated more than 3.5, the overall rating is more than 4,5, and a customer belongs to the “promoter” category, which means that he or she should rate NPS for more than 8 points. The third tree says that the condition is true if the delivery rating equals 5, the service rating is more than 3.5 and the style rating is more than 1.5. All other trees show strong connections between making a profit and evaluating "Style" and "Overall" for more than 3.5, "Service" for more than 2.5, being a brand promoter, paying less than EUR 16.5 for an advance payment. The algorithm also shows that even if a comment is treated as neutral but a customer gives only one point for the delivery, the profit will not be generated with a probability of 85.35%, which shows that the delivery is also an important indicator. Another predicted situation of not generating any profit is if a customer provides an overall score of less than 1.5 and belongs to the “detractor” group of customers. The barrier of implementing this kind of analytics is the necessity of expert knowledge to evaluate several algorithms, build an accurate business-relevant model, and "translate" the findings into insights.

Figure 21: Part of the first tree for predicting the “Profit”



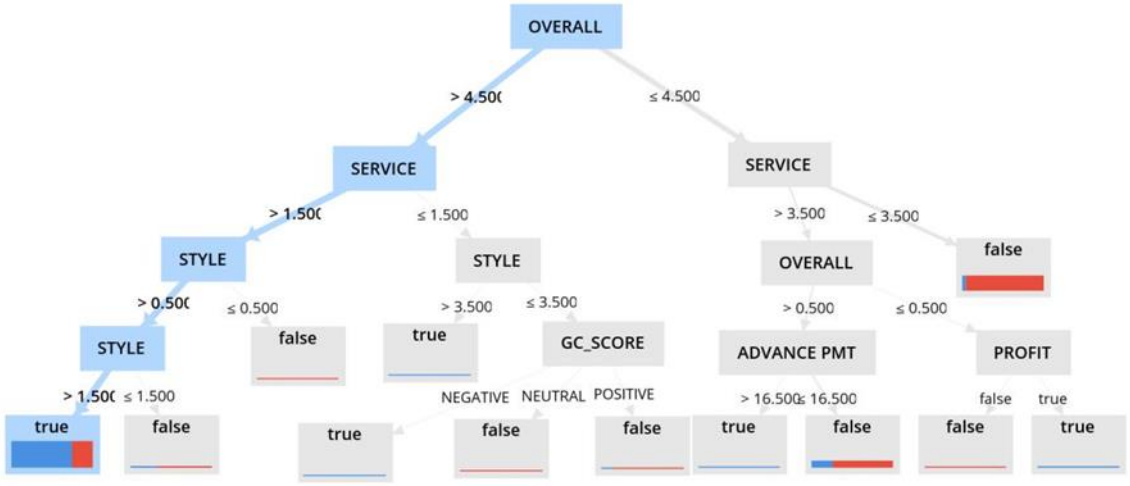
Source: Own work.

The second model that was built based on the “Feedback & Order Status” dataset predicts if a customer will be a brand promoter or not in the future, after the purchase. The business opportunity, in this case, would be addressed to the improvement of customer relationships.

The model is based on the same process but with several adjustments. The current model involves 10 trees with a maximum depth of 5. The model works with an accuracy of 82.99%. The main focus of the model is the true values since the model is supposed to define the "promoter" type of customers. The class recall for true values equals 89.51% and the class precision of 89.46%. The classification error, in this case, equals 17.01%.

The first tree shows that a customer will be a promoter if he or she purchases at least one item from the box and assigns the "overall" score of more than 4.5. The second tree in Figure 22 below demonstrates the relationships between the true condition and three ratings that are available in the dataset. A customer is more likely to be treated as a promoter if he or she evaluates the style and service for more than 1.5 and the overall for more than 4.5 points. The tree shows that even if the service is evaluated for 3.5 but the overall rating is less than 4.5, then a customer is less likely to be a promoter. The next tree shows another interesting data insight. A customer is more likely to be treated as a brand promoter if he or she assigns more than 4.5 points for an "Overall" rating, more than 3.5 points for style, and buys at least one item. While if a customer assigns less than 4.5 points for "Overall" and less than 2.5 points for "Style", then there is a probability of 96.34% that that customer is not a promoter.

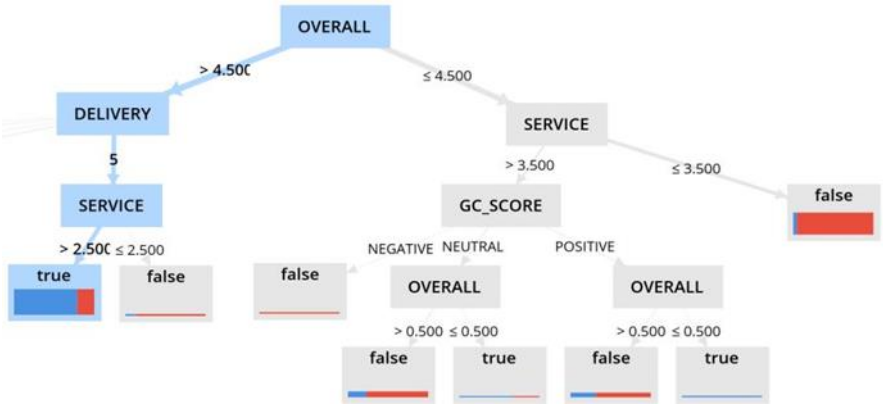
Figure 22: The second tree for detecting a "Promoter"



Source: Own work.

The fifth tree below shows the importance of the "Delivery" rating. The condition is true and a customer can be treated as a promoter if he or she assigns 5 points for delivery, more than 4.5 points for the "Overall" rating, and more than 2.5 points for the service. The same tree shows the same condition for a customer not to be a promoter as it was discussed in Figure 22 above. Another tree proves that a customer is less likely to be a promoter if he or she assigns less than 4.5 points for the "Overall" criteria and less than 3.5 points for the service. Other trees show mostly similar results, however, one of them also created a connection with the "New Stylist" attribute. A customer is more likely to be a promoter if he or she buys at least one item from the box and works with an "old" stylist.

Figure 23: The fifth tree for detecting a "Promoter"

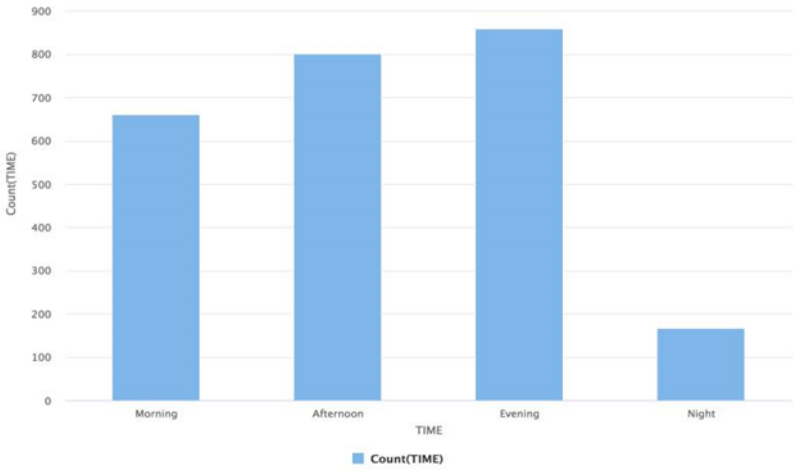


Source: Own work.

The last dataset that has been analyzed is called "Merged Quiz". The result of the analysis includes only descriptive and text analytics. After finishing the data cleansing, preparation, and application of text mining techniques, the result of the analysis is a list of the most popular brands among customers, the most common job styles, weekend styles, unfavorable colors, and unfavorable materials. The text analysis of the "Brand" column shows that the most preferable brands by the customers are Zara, Mango, Pull&Bear, Dutti Massimo, Reserved, and others. These results create an opportunity to reconsider the company's vendors, or to analyze the customer style preferences. After applying the same process on the "Job Style" column, the most common values are casual and non-strict style. Another column that includes the information about the most common style is the "Weekend Style". The results show that most of the clients, which equals 2 103, mentioned "casual" as the most preferred weekend style, 521 people mentioned "sporty" style, and 484 "business-casual" style. The next set of results refers to the "Unfavorable Colors" attribute, where orange is considered as the most unfavorable color since it was mentioned 876 times, pink color was mentioned 710 times, purple 706 times, yellow 680 times, and gold 670 times. The last column that went through the same process is "Unfavorable Materials". The results are the following: faux fur, faux leather, polyester, and wool. The least common unfavorable materials that were mentioned in the text were silk, linen, and synthetics. Such information could influence the decisions regarding the purchase of the inventory and inventory management since in this case text analytics helps to detect the customer preferences.

The first attribute that is described is "Time"; the bar chart in Figure 24 below shows that the most preferable time for making an order is evening. The statistics show the distribution of customers in different age groups. The results are the following: around 66% of customers belong to the "26-35" age group, 18% belong to the "18-25" age group, around 13% of customers belong to the "36-45" age group, around 2% of customers are between 46 and 55, and less than 1% of customers are over 55 years old. Most of the customers are managers, unemployed, in the decree, economists, or marketing specialists. This information allows the company to define the main customer segments which should be targeted in different ways.

Figure 24: Most preferable time for making an order



Source: Own work.

Another attribute that is described in the analysis is the "Source". The statistics also show that the most common source, from which the customers became aware of the company, is Instagram which represents 76% of all the values. Around 10% of customers specified "Blog" as the main source, and another 10% of customers mentioned "Advice" from people as the main source. In the "Expected Style" column customers are supposed to specify what kind of style they are expecting to get from the service: "Average" which is composed of the new and old styles of a customer, "Absolutely New" style, and "Old" style that a customer is used to. The biggest share of customers which is around 58% prefer a mix of both styles, around 32% of customers would like to try something absolutely new for them, and 10% of customers prefer the usual style.

The next attribute that has been analyzed is "Height" which is one of the physical parameters. The results show that the most common height among customers is 165 centimeters. Other physical parameters that have been analyzed are "Chest", "Waist", and "Hips". The most common values for all three body parameters among the customers are the following: 90 centimeters for the chest, 70 centimeters for the waist, and 100 centimeters for the hips. While the most common chest and waist parameters refer to "M" cloth size, the most common hips parameters meet the "L" size condition, according to Table 5 above. Therefore, the most common size among customers is "L", then "M", and then "S", which can benefit inventory management. The minority of customers wear "XL>" and "XS" sizes. The distribution of the "Hair Color" attribute is the following: around 33% of clients have dirty blond hair color, 22% have brown hair color, around 19% of customers are brunet, 18% are blond, 5% are red, and the remaining small number of people gave other hair colors. The distribution of the answers regarding the favorite metal is more or less equal between the tree values: 37% prefer both metals, 35% prefer silver, and 26% prefer gold. Less than 2% of customers do not prefer any metal. The last attribute that has been analyzed is "Jewelry", which shows that around 78% of customers are willing to wear jewelry, and the rest are not.

Table 6 below represents the results discussed above for all three datasets. The results are grouped by three criteria: descriptive analytics, which includes the statistics and visualizations of different attributes from all three datasets, text analytics or text mining, which includes the analysis of all the attributes with customer feedback, and predictive analytics, which includes three predictive models.

Table 6: The results of data analytics

	Descriptive Analytics	**Text Analytics**	**Predictive Analytics**
Cloth Feedback	• Most frequently sold and returned items • Number of sold and returned items • Relationships between the ratings and "Cloth Status" • Feedbacks by date • Distribution of "Fit" values • Average ratings	• Keywords from "Comments": color, size,and style • "Comments": Somecustomers already have similar clothes in their closets • Most of the comments are positive	• Conditions for an item to be Sold or Returned
Feedback & Order Status	• The average number of clothes sent and purchased • Average ratings • Correlation between the "New Stylist" and ratings • Correlation between the purchased items and ratings • The averahe number of purchased items by "New Stylist" • Profit by "New Stylist", by date, by ratings	• Aspects defining NPS Reason: service, box, time, stylist, and size • Suggestions: more communication, instructions, and recommendations from stylists	• Conditions for Profit • Conditions for a custumer to become a Promoter
Merged Quiz	• Most preferable time for making an order • Distribution of "Age" values • Most popular occupation, source, and expected style • Most common height, size, and body parametrs • Distribution of "Hair Color" and "Metal" • Willingness to wear jewerly	• Brands: Zara, Mango, Pull and Bear, Dutti Massimo, Reserved • Job Style: casual and non-strict • Weekend Style: casual, sporty, and business casual • Unfavorable colors: pink, purple, yellow, gold • Unfavorable materials: faux-fur, wool, polyester, faux-leather	-

Source: Own work.

5.2 Interview

The interview with the CEO of Company X complements the results acquired in the previous subchapter. It provides the additional opportunities and barriers of using business analysis that could be discovered in a small enterprise, moreover, it shows if the analysis and findings are relevant for the company, it shows if there is a real need for business analytics in small companies.

The interview is structured in accordance with the business analytics process discussed in the first chapter, which includes the following steps: business problem identification, identification of issues and opportunities for collecting data, data collection, data processing, data analysis, data interpretation, and results development. The first set of questions and subquestions was an introductory set, it was related to the current situation with business analytics in the company, its awareness in the company management, and the tools that are currently used in the company. The CEO explained that currently data analytics is used in the company since some part of decision-making in the company is based on the results of data analysis, however, the company is currently using only one type of business analytics which is descriptive. For instance, the inventory and supply management is based on the results of data analytics, since "it is the most convenient way to find out the customer preferences for a certain season". Moreover, it is important to know if an item or material is selling well or not, where business analytics helps to track the "success" of each item and material. The company is trying to do everything in order to teach and help each employee to deal with data and interpret the results. Currently, the company data is processed and visualized in the following tools Google Sheets, and Yandex Metrics. Both tools are free of charge since currently, the company is lacking the investment to implement "good BA", as the CEO clarified.

The second question follows the business analytics process, therefore, it is the following: "Are there any current business problems that could be solved with data analytics?". The CEO mentioned that "almost any business problem can be solved with business analytics". However, the most important one that can be solved with BA is the problem of correct sizes and the most appropriate size chart. The company wants to reduce the amount of "too big" and "too small" fits by adjusting its size chart. In the third question, the CEO was asked to choose the most appropriate for the company opportunities and barriers for data analytics from Table 2 above. The answer was "the lack of finance is the biggest issue", while data security "is not considered as an issue" due to the fact that the managers decided to ignore this aspect, which actually could be risky for the future of the company's data. The CEO mentioned that all the opportunities mentioned in the table such as improved data support, improved decision-making process, flexibility and freedom of choice, focus only on the most important capabilities, data insights, competitive advantage, and reduced costs are relevant for the company.

While the first three questions with the sub-questions were asked before the presentation of the analysis and results to the CEO, the next two questions were asked during the presentation and the last four after. Therefore, the next question with sub-questions is focused on the data collection, which is the third step in the business analytics process. The CEO clarified that he prepared the databases, the set questions for the first questionnaire, and together with the team, he prepared the questions for the second questionnaire. The data for the datasets is stored and collected automatically in Google Sheets. At the moment the company is implementing the third type of questionnaire, which is going to replace two previous ones. At this stage, the product manager is going to take care of the questionnaires and the data collection process. The fifth question was structured in the way to find out if there is a team member with expert knowledge in business analytics who takes care of data cleansing, processing, analysis, interpretation, and results development. The current situation is that only the CEO by himself takes care of business analytics in the company, he is the only team member with expert knowledge in this area.

Based on the results presented in Table 6 above, the next question in the interview was the following: "Which type of data analytics do you find the most interesting and useful for your company? Why?". Taking into consideration the fact that most of the decisions in the company are supported by descriptive analytics and visualizations prepared in Google Sheets and Yandex Metrics, the most interesting types of data analytics that were used in the analysis were "definitely predictive and text analytics". The CEO was already aware of descriptive analytics and its capabilities and results, however, the company does not use special tools that support text and predictive analytics. Therefore, the CEO finds both of them the most useful. Based on the fact that the top manager is aware of the opportunities of data analytics, and was interested in the results of the analysis, the next question was the following: "Is the company ready to invest in the new software or human resources that could benefit the situation with business analytics in the company?". The positive answer was followed by additional questions about the amount of money that the company is willing to invest in data analytics on a monthly basis. The answer was the following: "It really depends. If an employee tells me that he or she can increase a certain ratio by x% and then do it, then based on that I can increase his or her salary. Regarding the software, it also depends on its benefits. For a simple graphical data processing, I am ready to pay EUR 200 per month".

The eighth question measures if the results of this analysis are going to influence the business decisions in the future. The CEO stated: "The results will definitely influence the business decisions in the company". The first thing that he found useful is the relatively low correlation between the "Price" rating and the fact of an item to be purchased, while the "Style" rating has the highest influence on the purchased items. Another useful result that the company will start implementing is the new size chart proposed in Table 5 above. The CEO liked the approach that was taken in the analysis: first applying text analytics in order to determine the most popular brand, and applying the size chart of that brand on the

customer's body parameters in order to perform descriptive analytics. As he mentioned earlier in the interview, this was one of the most important problems that the company was willing to solve with business analytics. It was also discovered in the results that the lack of communication between stylists and customers is a problem. The CEO mentioned that he is aware of this problem, and the analysis proved it. The stylists are not willing to communicate more with customers for the amount of money they get, therefore, the problem can not be solved because of the lack of finance, however, the company now is trying to solve it another way. He also mentioned that the problem of the absence of good detailed instruction in the box was really taking place in the business. The company managers noticed that several months ago, and at the moment it is already solved.

The last question is similar to the third one, the CEO was asked to list any additional business opportunities and barriers for data and business analytics in the company. The answer complements the lists of opportunities and barriers derived from the literature, analysis, and results. While talking about the additional opportunities, the CEO did not find any of them to add to the list, since all the most important ones had been already mentioned. However, he added some additional barriers that are relevant for the company. Firstly, the "high costs of data science experts and the fact that the company needs the whole team of them" is a barrier to implementing better business analytics in the company. Secondly, it is hard to treat the data correctly because every customer has his or her own subjective view while answering the questions. The CEO stated: "What is the style for him or her? Nobody knows". Another issue is that "if a client is angry, then there is a higher probability for him to assign low ratings, and we do not know how to deal with it". Moreover, some customers can write their "aspirational parameters and height", therefore, it is hard to treat the data in a correct way. Thirdly, "missing data is also a problem for data analytics because a big share of customers simply do not answer all the questions from surveys". The last barrier that he mentioned is that the company should try many different approaches to collect the right data before finding the right one. This procedure harms the historical data and creates difficulties for data analysis in the future.

6 DISCUSSION

6.1 Relationship between the actual and prior findings

Each chapter of the master's thesis identifies new opportunities and barriers of using Business Analytics in small companies in order to answer the main research question in the most precise way. In this case, the results of the literature review are considered as the prior findings and the results of the analysis as the actual findings.

According to the literature review above, the opportunities of business analytics in small companies are the following: improved data support, improved decision-making process, flexibility and freedom of choice, focus on the most important capabilities, insights extraction from data, a gain of competitive advantage, and cost savings. In order to see the relationship between the prior and actual findings, the opportunities from the analysis and the results chapters should also be listed. Starting from the analysis and results development of the first dataset, there were discovered several opportunities for the company. The first opportunity is the visualization of the current situation in the company, or the visualization of historical data for several years, which improves decision-making in the company. This opportunity is relevant to all three datasets, moreover, during the interview, the CEO mentioned that most of the decisions in the company are based on descriptive analytics and visualizations.

The next two opportunities that were discovered during the analysis of the first dataset are that business analytics can help to determine the attitude of customers towards the brand and that it can help to determine the customer preferences, which together could improve the customer relationships. Since this opportunity was not discovered before, it will be added as an additional opportunity of using business analytics, which is presented in Table 7 below. Another opportunity for the company is predicting the conditions for profit and return of items, which helps to see the future prospects for the company. The analysis also helped to see which clothes are not preferred by customers in order to stop sending to items that have not been successfully sold. Such information can influence the inventory and supply management in the company, as was also mentioned in the interview. As the results of the analysis show, business analytics assists in measuring the success and performance of the company. According to predictive analytics applied on the first dataset, the trees show the ratings that have the highest and lowest influence on the returned items, such as “Style” and “Price”. During the interview, the CEO of Company X paid high attention to these results, which helped him to focus on the most important capabilities and set new priorities.

Most of the opportunities that have been defined during the analysis of the first dataset are also relevant for the second dataset “Feedback & Order Status”. However, in this case, text mining techniques help to find out the company weaknesses that are defined by customers. This opportunity can be treated as one from the literature review, which is “extracting insights from data”. Text analysis discovered one of the current business problems which is the lack of communication between stylists and customers, which was also proven during the interview. This fact creates several opportunities for using business analytics, such as identifying business problems that are hard to define and improving customer experience. Predictive models for making a profit or for detecting promoters among customers also bring several opportunities that are already listed in Table 7 below. These are the opportunities for determining future prospects, focusing on the most important capabilities and others.

The analysis of the last dataset which is called “Merged Quiz” brings several similar opportunities for business, such as improved decision-making by using visualizations, data

insights extraction, and improved inventory and supply management. However, there is one more opportunity that could be added to the final list, which is improved customer segmentation. The application of descriptive analytics on the last dataset helped to see the visual representation of different customer segments based on their age, occupation, expected style, and others. Therefore, the literature review, analysis, results development, and the interview helped to create a list of the key opportunities of using business analytics in small companies, where the last opportunity is a combination of all of them because together they can adjust a company's business strategy.

Another objective of this work aims at identifying the barriers to using business analytics in small enterprises. According to the findings from the literature that are presented in Table 2, the most common barriers and challenges are the following: low awareness of business analytics, inability to calculate the ROI from such solutions, inadequate use and the lack of financial resources, lack of expert knowledge, legal concerns, poor data quality and data usage, and problems related to data security.

The analysis and the development of the results were performed to confirm those opportunities or complement the list of them. The first barrier that was discovered was the dates mismatch in different datasets, which led to finding the common period of time for all the company data, which made the analysis shorter and shallower. In this case, the first barrier that was defined is the lack of data in several datasets. Starting from the first dataset "Cloth Feedback", one of the first barriers that was discovered was poor data quality, which stands for the presence of negative values, unrealistic values, missing values, outliers, and others. This means that one of the barriers derived from the literature is confirmed during the first steps of data cleansing. Thirdly, the inconsistency of formats for dates created another barrier for using business analytics. During the interview with the CEO, it was detected that he was not aware that this problem exists in the datasets, which can be treated as several barriers: absence of expert knowledge, poor data quality, or unawareness of data-related issues.

The most common barriers that were discovered in the second dataset "Feedback & Order Status" are similar to those that were already mentioned: huge amount of missing values, format inconsistency, and inconsistency of measures for the ratings. Additional barriers that were detected and confirmed during the interview are the lack of expert knowledge for text and predictive analytics, as well as the lack of access to text mining tools.

The analysis of the last dataset "Merged Quiz" brings additional arguments and confirms the barriers that were already listed. Firstly, the inconvenient format of values which is in some cases polynomial instead of binomial or numerical brings us back to the issue of poor data quality. Additionally, the presence of human error also influenced the data quality. The questionnaires were developed in such a way that customers could enter any kind of data while entering their birthday, for instance. The values were unrealistic since it was not taken into consideration while developing the questionnaires. Moreover, the high share of missing

values and duplicates likewise reduces the data quality. Secondly, as it was mentioned during the analysis, the data was unstructured and unstandardized, which could be explained as the lack of expert knowledge. Thirdly, the presence of two different types of questionnaires with similar data that had to be merged into one dataset creates an additional barrier, which was confirmed by the company's CEO during the interview. The issue is that small enterprises should try many different approaches for data collection before finding the right one, which could be treated as a barrier of "complex data collection". Fourthly, it was confirmed during the interview that the lack of financial resources needed to pay the qualified experts in data analytics is a huge barrier to implementing business and data analytics in small enterprises. Therefore, in this case, the prior findings were confirmed with the actual findings. Fifthly, the CEO of Company X mentioned that at the moment the managers do not pay enough attention to the company's data security. This fact confirms the issue of poor data security that was identified from the literature review. The last barrier, that was discovered during the interview and can be added to the final list of all the opportunities and barriers, is data bias or data subjectivity. Since every customer as well as every data expert has his or her own way of understanding the questions, terminology, and the data itself.

Table 7: All the Opportunities and Barriers of using Business Analytics

OPPORTUNITIES	BARRIERS
+ Improved data support + Improved decision-making process + Flexibility and freedom of choice + Focus only on the most important capabilities + Extracting insights from data + Gaining a competitive advantage + Saving costs + Improved customer relationships + Determining future prospects + Improved inventory and supply management + Measuring the success and performance + Setting new priorities + Identifying business problems that are hard to define + Improving customer experience + Improved customer segmentation + Adjusting business strategy	- Unawareness of business analytics and inability to calculate the ROI from such solutions - Inadequate use and the lack of financial resources - Absence of expert knowledge - Law protection - Poor data quality and data usage - Poor data security - The lack of data - Unawareness of data-related issues - The lack of access to complex analytical tools - Complex data collection - Data bias

Source: Own work.

Table 7 above includes the list of all the opportunities and barriers of using business analytics in small companies that arose from this research. It sums up all the main findings from the literature review, analysis, results development, and the interview in order to answer the main research question. These opportunities and barriers can vary from company to company

or industry, since it is mainly based on the analysis of one selected company. However, the combination of several approaches can bring up a chance for any small company to find several opportunities and barriers that are relevant only for this business.

6.2 Suggestions for improvements

Taking into consideration only those barriers and issues that belong to Company X, those barriers that were discovered during the data analysis and the interview, the list is the following: lack of financial resources, lack of expert knowledge, poor data quality, lack of data, unawareness of data-related issues, lack of access to complex analytical tools, complex data collection, and data bias.

The lack of financial resources is one of the most common issues for all SMEs (Russegger et al., 2015). According to several studies, the issue can be solved in two main ways such as external and internal financing. External financing stands for the attraction of investors, while internal financing stands for the improvement of internal business processes. Several studies highly recommend focusing only on the internal business capabilities, since the perception of risk of investing in small businesses pushes investors to increase the costs of lending the money, which can cause additional financial problems for the company in the future (Bakhtiari, Breunig, Magnani & Zhang, 2020). According to the literature review that was presented above, business analytics can help small companies in saving costs on employees and increasing revenue in the long run (Papachristodoulou, Koutsaki & Kirkos, 2017). This could mean that the problem of the lack of financial resources can be solved with business analytics, which is an internal capability. The problem of lack of access to complex analytical tools is directly related to the problem of limited financial resources in SMEs since most of the complex analytical solutions are not free of charge. For instance, according to the results of the interview, the most interesting and useful results of the analysis were produced by implementing text and predictive analytics in the RapidMiner Studio analytical tool. The pricing structure of using this tool is divided into three categories: “small” which covers 100.000 rows for approximately EUR 2.100 per one user per year, “medium” which covers 1.000.000 data rows for EUR 4.200 per user per year, and the last “large” category which costs EUR 8.400 per year and allows to analyze the unlimited number of rows for one user (RapidMiner., 2021). Considering the fact, that none of the five initial datasets were exceeding 100.000 data rows, and the company’s CEO mentioned that the company is willing to spend on business analytics around EUR 200 per month which equals EUR 2.400 per year, the company can afford the “small” package from RapidMiner Studio.

In the first stages of implementing more advanced analytical tools, the problem of the lack of expert knowledge can be solved by organizing additional workshops and training for the current employees (Erol, Jager, Hold, Ott & Sihn, 2016). Moreover, the company can invest in the expert knowledge gain and the requalification of one of the company’s employees, for

instance, from the IT department. For example, RapidMiner offers a huge variety of video tutorials that are free of charge in the RapidMiner academy (RapidMiner., 2021). After solving the problem of the lack of expert knowledge, the company can overcome the barrier of unawareness of data-related issues. Additional literature research shows that the professional knowledge of working with analytical tools, company's systems, and software, helps to understand the root cause of several data-related issues (Botsis, Hartvigsen, Chen & Weng, 2010). There is another couple of barriers that can be overcome after adding some expert knowledge to the data analysis. Firstly, by improving the current questionnaire, and cleaning the historical data, which can be the responsibilities of a data analyst in the company, the issue of lack of data can be solved because all the subsequent data that the company acquires, will be collected and structured in the proper way. In this case, the barrier of complex data collection has an opportunity to be overcome automatically, since the expert knowledge in data collection can help to reduce the number of approaches the company takes before finding the right one.

The poor data quality problem can be solved if the company starts applying data cleansing and data processing strategies, which include the utilization of missing, duplicate, and unstructured data issues (Botsis, Hartvigsen, Chen & Weng, 2010).

The last barrier that is called "data bias" was discovered during the interview. The company's CEO mentioned that it is one important issue that can not be solved at the moment. The additional literature review offers a couple of solutions that can be applied in this case. First of all, since the customers have their individual subjective way of understanding the data, the company can introduce its own vocabulary that defines each of the most important terms for the analysis. The second solution could be the application of text mining techniques in order to extract the subjective definitions from the customer comments (Botsis, Hartvigsen, Chen & Weng, 2010).

CONCLUSION

Business Analytics is the consolidation of all the supporting mechanisms that help to transform the data into a valuable piece of information that will improve and accelerate the decision-making and problem-solving processes (Delen & Ram, 2018). Business analytics helps companies to anticipate business outcomes and trends through statistical analysis, data mining, and predictive modeling (Negash & Gray, 2008). Each business is unique, which means that every company requires the use of different types of analytics that suits its business strategy. The results of using different types of analytics, tools, and expert knowledge, bring us to the list of opportunities and barriers of using business analytics, which is specific for every enterprise.

The European Commission defined several types of enterprises by size: micro, small, medium, and large, while small and medium are merged into one group, which stands for a Small and Medium-sized Enterprise (CSES, 2012). According to the literature review above, the opportunities and barriers of using business analytics in large companies differ from the opportunities and barriers of using the same analytics in SMEs, while they almost do not differ in small and medium-sized companies. This implies the existence of two categories of using business analytics: in small and medium-sized companies, and in large enterprises. While large businesses had been realizing the value and importance of the existing data for a long time, they started to invest huge amounts of money into the analytical systems several years ago, most small companies around the world have only recently discovered the potential of using business analytics tools to improve business processes and gain unique competitive advantages (Guarda, Santos, Pinto, Augusto & Silva, 2013). Since the topic of using business analytics in small companies is relatively new, and it is gaining popularity, the main purpose of the research was to investigate and identify the key opportunities and barriers of using business analytics in small companies to help other small and medium-sized enterprises to go through this process more easily and smoothly, what was aligned with the main research question "What are the opportunities and barriers of using Business Analytics in small companies?".

Every stage of the research has its list of goals that were achieved stepwise. First of all, the literature was reviewed, the key terminology for the thesis was defined, and the list of opportunities and barriers of using business analytics in small companies was identified, according to the literature. Secondly, the goal was to describe a particular business case and the data, which was achieved in the third chapter. Then the data for the analysis was selected, processed, transformed, split into training and testing sets, analyzed, and evaluated. Moreover, the goal was to deploy the best model with the highest accuracy, which was also achieved. Fourthly, the results of data analytics were interpreted with figures and tables. The additional goal for the analysis and results development stages was to find some additional opportunities and barriers of using business analytics that could complement the list derived from the literature review. Next, the goal was to evaluate the contribution of the thesis by performing an interview with the CEO of the company. The last set of achieved goals was related to bringing all the findings together in order to answer the main research question, writing a list of suggestions for possible improvements in the company, and preparing the valuable conclusion of the thesis.

The answer to the main research question required to be as precise as possible. Therefore, firstly, the opportunities and barriers were derived from an in-depth literature review, secondly, the list of them was prolonged after performing the analysis of data from one small company, then the results and the actual model development complemented the list of opportunities and barriers. The last stage of answering the research question involves the interview with the CEO of the small enterprise that provided its datasets for the analysis. Basically, the whole thesis is a stepwise answer to the research question.

There are eleven barriers that have been discovered in the thesis, which are the unawareness of business analytics and inability to calculate the ROI from such solutions, inadequate use and the lack of financial resources, absence of expert knowledge, law protection, poor data quality and data usage, poor data security, lack of data, unawareness of data-related issues, lack of access to complex analytical tools, complex data collection, and data bias. The thesis also provides sixteen most relevant opportunities of using business analytics in small companies, which are improved data support, improved decision-making process, flexibility and freedom of choice, focus only on the most important capabilities, extracting insights from data, gaining a competitive advantage, saving costs, improved customer relationships, determining future prospects, improved inventory and supply management, measuring the success and performance, setting new priorities, identifying business problems that are hard to define, improving customer experience, improving customer segmentation, and adjusting business strategy. The list of opportunities shows that business analytics can and should be implemented not only in large enterprises but also in smaller ones. It gives them a chance to overcome the current challenges and accelerate business growth in the long run, which is worth getting over all the barriers that were discovered above.

To better understand the implications of these results, future studies could address the question of using business analytics in small companies in particular industries, since the results of this work are broad for all small enterprises. Some opportunities and barriers could be highly relevant for one industry, while not that relevant for another, which creates additional possibilities for future studies. Another possibility could be that future studies could address the question of comparing opportunities and barriers of using business analytics in small and large companies. Such studies could complement the results of this research.

The main contribution of this work is the set of results that can help small companies to identify the most relevant set of benefits and issues for them. As it was mentioned before, each company is unique as a combination of all the business components, therefore, it requires a unique set of opportunities and barriers from the defined list. This brings the main limitation of this research since the whole set of results can not be applied to every small enterprise. Nevertheless, the research could also increase the awareness of using business analytics in small companies, as well as boost the increasing popularity of this topic. The research shows the potential of implementing different types of data analytics, which could guide the managers of small enterprises in making the decision regarding the implementation of business analytics.

REFERENCE LIST

1. Ajimoko, J. (2018). Considerations for the Adoption of Cloud-based Big Data Analytics in Small Business Enterprises. *The Electronic Journal Information Systems Evaluation, 21*(2), 63-79.
1. Amber. (2017, December 21). *The history of the evolution of business analytics.* Cyfe. Retrieved April 7, 2021 from https://www.cyfe.com/blog/history-evolution-business-analytics/
2. Ayoubi, E. & Aljawarneh, S. (2018). Challenges and opportunities of adopting business intelligence in SMEs: collaborative model. *Proceedings of the First International Conference on Data Science, E-learning and Information Systems, 42*(5). https://doi.org/10.1145/3279996.3280038
3. Azevedo, R. (2016). Levelling the trading field for SMEs. *WTO,* 1-12.
4. Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. John Wiley & Sons.
5. Bakhtiari, S., Breunig, R., Magnani, L., & Zhang, J. (2020). Financial constraints and small and medium enterprises: A review. *The Economic Record*, *96*(315), 506–523. https://doi.org/10.1111/1475-4932.12560
6. Banerjee, A., Bandyopadhyay, T. & Acharya, P. (2013). Data Analytics: Hyped Up Aspirations or True Potential? *Indian Institute of Management, 38*(4), 1-12. https://doi.org/10.1177/0256090920130401
7. Barends, E. & Rousseau, D. (2018). *Evidence-Based Management: How to Use Evidence to Make Better Organizational Decisions* (1st ed.). Kogan Page Publishers. https://doi.org/10.1007/s10551-020-04488-3
8. Bayraktar, M. & Algan, N. (2019). The Importance Of SMEs On World Economies. *International Conference on Eurasian Economies*, 56-59. https://dx.doi.org/10.36880/c11.02265
9. Becker, L. & Gould, E. (2019). Microsoft Power BI: Extending Excel to Manipulate, Analyze, and Visualize Diverse Data. *Serials Review, 45*(3), 184-188. https://doi.org/10.1080/00987913.2019.1644891
10. Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. (2010). Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on translational bioinformatics,* 1–5.
11. Brands, K. & Holtzblatt, M. (2015). Business Analytics: Transforming the Role of Management Accountants. *IMA Management Accounting Quarterly, 16*(3), 1-12. shorturl.at/opzVX
12. CSES. (2012). Evaluation of the SME definition. *Centre for Strategy and Evaluation Services,* 2-11.
13. Chernyshova, G. (2013). Application of business intelligence tools for small and average businesses. *Scientific and Practical Journal, 2*(13), 23-26.

14. Davenport, T. (2006). Competing on analytics. *Harvard Business Review*, *84*(1), 98–107.
15. Delen, D. & Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of Business Analytics, 1*(1), 2-12. https://doi/10.1080/2573234X.2018.1507324
16. Eckerson, W. (2006). Predictive analytics: Extending the Value of Your Data Warehousing Investment. *TDWI,* 15-21.
17. Eriksson, T., Bigi, A. & Bonera, M. (2020). Think with me, or think for me? On the future role of artificial intelligence in marketing strategy formulation. *The TQM Journal, 32*(4), 795–814. https://doi/10.1108/TQM-12-2019-0303
18. Erol, S., Jäger, A., Hold, P., Ott, K. & Sihn, W. (2016). Tangible Industry 4.0: a scenario-based approach to learning for the future of production. *Procedia CIRP, 54*, 13-18. https://doi.org/10.1016/j.procir.2016.03.162
19. European Commission. (2017). User guide to the SME Definition. *Publications Office of the European Union*, 3-31. https://doi.org/10.2873/620234
20. European Commission. (2021, March 26). *EU funding programmes*. Retrieved April 16, 2021 from https://europa.eu/youreurope/business/finance-funding/getting-funding/eu-funding-programmes/index_en.htm\
21. Foote, K. D. (2018, September 25). A Brief History of Analytics. *DataVersity.* Retrieved April 16, 2021 from https://www.dataversity.net/brief-history-analytics/
22. Giannantonio, C. & Hurley-Hanson, A. (2011). Frederick Winslow Taylor: Reflections on the Relevance of The Principles of Scientific Management 100 Years Later. *Journal of Business and Management, 17*(1), 7-11.
23. Goebel, R., Norman, A. & Karanasios, S. (2015). Exploring the Value of Business Analytics Solutions for SMEs. *UK Academy for Information Systems*, 1-26.
24. Guarda, T., Santos, M., Pinto, F., Augusto, M. & Silva, C. (2013). Business Intelligence as a Competitive Advantage for SMEs. *International Journal of Trade, Economics, and Finance, 4*(4). http://doi.org/10.7763/IJTEF.2013.V4.283
25. Ihaka, R. & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314. https://doi.org/10.2307/1390807
26. Iqbal, M., Kazmi, S., Manzoor, A., Soomrani, A., Butt, S. & Shaikh, K. (2018). A study of big data for business growth in SMEs: Opportunities & Challenges. *International Conference on Computing, Mathematics and Engineering Technologies*, 1-7. https://doi.org/10.1109/ICOMET.2018.8346368
27. Kaur, H. & Phutela, A. (2018). Commentary upon descriptive data analytics. *2nd International Conference on Inventive Systems and Control (ICISC)*, 678-683. https://doi.org/10.1109/ICISC.2018.8398884
28. Keen, P. (1980). *Decision support systems: a research perspective*. Center for Information Systems Research, Alfred P. Sloan School of Management. https://doi.org/10.1016/b978-0-08-027321-1.50007-9

29. Kotu, V. & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner.* Elsevier Inc. https://doi.org/10.1016/C2014-0-00329-2
30. Lacey, D. & James, B. (2010). Review of Availability of Advice on Security for Small/Medium Sized Organisations. *Information Commissioner's Office*, 2-26.
31. Lindbergh, L., VanderHorst, R., Hass, K. & Ziemski, K. (2018). *From Analyst to Leader: Elevating the Role of the Business Analyst*. Berrett-Koehler Publishers.
32. Linoff, G. (2015). *Data Analysis Using SQL and Excel* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119183419
33. McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media, Inc.
34. Natarajan, G. & Wyrick, D. (2011). Framework for Implementing Sustainable Practices in SMEs in the United States. *Proceedings of the World Congress on Engineering*, 750-754.
35. Neagu, C. (2016). The importance and role of small and medium-sized businesses. *Theoretical and Applied Economics, 23*(3), 331-338.
36. Negash, S. & Gray, P. (2008). Business Intelligence. *Communications of the Association for Information Systems, 13*, 175-193. https://doi.org/10.17705/1CAIS.01315
37. Nelson, G. (2017, July 7). *Difference between analytics and big data, data science and informatics* [published on blog]. Retrieved April 4, 2021 from https://www.thotwave.com/blog/2017/07/07/difference-between-analytics-and-bigdata-datascience-informatics/
38. Ogbuokiri, B., Udanor, C. & Agu, M. (2015). Implementing big data analytics for small and medium enterprise (SME) regional growth. *IOSR Journal of Computer Engineering, 17*(6), 35-43. http://doi.org/10.9790/0661-17643543
39. Palmer, A. & Hartley, B. (1999). *The Business and Marketing Environment* (3rd ed.). McGraw-Hill.
40. Papachristodoulou, E., Koutsaki, M. & Kirkos, E. (2017). Business intelligence and SMEs: Bridging the gap. *Journal of Intelligence Studies in Business, 7*, 70-78. http://doi.org/10.37380/jisib.v7i1.216
41. Power BI. (2021). *Power BI pricing*. Retrieved April 27, 2021, from https://powerbi.microsoft.com/en-us/pricing/
42. Power, D., Heavin, C., McDermott, J. & Daly, M. (2018) Defining business analytics: an empirical approach. *Journal of Business Analytics*, *1*(1), 40-53. https://doi.org/10.1080/2573234X.2018.1507605
43. Raj, R., Wong, S. & Beaumont, A. (2016). Business Intelligence Solution for an SME: A Case Study. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 41-50. https://doi.org/10.5220/0006049500410050
44. RapidMiner. (2021). *RapidMiner pricing.* Retrieved July 17, 2021 from https://rapidminer.com/pricing/

45. Russegger, S., Freudenthaler, B., Güntner, G., Kieseberg, P., Stem, H. & Strohmeier, F. (2015). Big Data und Data-driven Business fiir KMU. *Digital networked Data-Verein fiir Innovation und Erforschung vernetzter digitaler Daten*, 3-13.
46. Sagiroglu, S. & Sinanc, D. (2013). Big data: A review. *International Conference on Collaboration Technologies and Systems*, 42-47. https://doi.org/10.1109/CTS.2013.6567202
47. Saxena, R. & Srinivasan, A. (2012). *Business Analytics: A Practitioner's Guide*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-6080-0
48. Sen, D., Ozturk, M. & Vayvay, Ö. (2016). An Overview of Big Data for Growth in SMEs. *Procedia, 235*, 159-167. https://doi.org/10.1016/j.sbspro.2016.11.011
49. Sisense. (2021). *Diagnostic Analytics.* Retrieved April 17, 2021 from https://www.sisense.com/glossary/diagnostic-analytics/
50. Soltanpoor, R. & Sellis, T. (2016). Prescriptive Analytics for Big Data. *Australasian Database Conference, 9877*, 245-256. https://doi.org/10.1007/978-3-319-46922-5_19
51. Someh, I. & Shanks, G. (2015). How Business Analytics Systems Provide Benefits and Contribute to Firm Performance? *ECIS*, 1-12. https://doi.org/10.18151/7217270
52. Sun, Z., Sun, L. & Strang, K. (2018). Big Data Analytics Services for Enhancing Business Intelligence. *Journal of Computer Information Systems, 58*(2), 162-169. http://doi.org/10.1080/08874417.2016.1220239
53. Sun, Z., Zou, H. & Strang, K. (2015). Big Data Analytics as a Service for Business Intelligence. *Conference on e-Business, e-Services, and e-Society*, 200-211. http://doi.org/10.1007/978-3-319-25013-7_16
54. Tableau. (n.d.). *Business Intelligence or Business Analytics: What's The Difference?* Retrieved April 7, 2021 from https://www.tableau.com/learn/articles/business-intelligence/bi-business-analytics
55. Tamm, T., Seddon, P. & Shanks, G. (2013). Pathways to value from Business Analytics. *Thirty-Fourth International Conference on Information Systems*, 1-16.
56. Tan, K., Han, H. & Elmasri, R. (2000). Web data cleansing and preparation for ontology extraction using WordNet. *IEEE, 2*, 11-18. https://doi.org/10.1109/WISE.2000.882844
57. Tutunea, M. & Rus, R. (2012). Business Intelligence Solutions for SMEs. *Procedia Economics and Finance, 3*, 865-870. https://doi.org/10.1016/S2212-5671%2812%2900242-0
58. Whitelock, V. (2018, September 19). Business analytics and firm performance: role of structured financial statement data. *Journal of Business Analytics, 1*(2), 81-92. https://doi.org/10.1080/2573234X.2018.1557020
59. Wood, L. (2020, March 2). The Global Big Data Analytics Market, 2027: A $105+ Billion Opportunity Assessment. *Research and Markets.* Retrieved February, 25 2021 from https://www.prnewswire.com/news-releases/the-global-big-data-analytics-market-2027-a-105-billion-opportunity-assessment
60. World Bank. (2019). *Small and Medium Enterprises Finance*. Retrieved April 18, 2021, from https://www.worldbank.org/en/topic/smefinance
61. Yanovitch, S. (2016). Data & Analytics Survey. *IDG,* 1-7.

62. ZARA. (n.d.). *Size chart*. Retrieved May 19, 2021, from https://www.zara.com/ size-guide

APPENDICES

Appendix 1: Povzetek (Summary in Slovene language)

Dolgo časa so velika podjetja spoznavala vrednost in pomen obstoječih podatkov, zato so začela vlagati ogromne količine denarja v analitične sisteme. Nekatera mala podjetja po vsem svetu so šele pred kratkim začela vlagati v poslovna analitična orodja, da bi izboljšala svoje poslovne procese in pridobila edinstvene konkurenčne prednosti. Ker ta tema pridobiva na priljubljenosti na svetovni ravni, sem v tej magistrski tezi opredelila in analizirala prednosti in izzive izvajanja takšnih orodij v malih podjetjih in zlasti v enem malem podjetju. Zato magistrsko delo obravnava raziskovalno vprašanje "Kakšne so možnosti in ovire za uporabo poslovne analitike v malih podjetjih?". Raziskovalno vprašanje je neposredno usklajeno z namenom te teze, ki je preučiti in opredeliti ključne priložnosti in ovire za uporabo poslovne analitike v malih podjetjih, da bi MSP lažje in nemoteno prestala ta proces. V tej tezi je primer majhnega podjetja, ki je bilo analizirano z uporabo več analitičnih orodij. Odgovor na glavno raziskovalno vprašanje je moral biti čim bolj natančen. Zato je celotna teza postopen odgovor na raziskovalno vprašanje, kjer vsaka faza teze odkriva nove koristi in ovire, ki dopolnjujejo končni odgovor.

V tezi je bilo odkritih enajst ovir, ki so nevednost poslovne analitike, neustrezna uporaba in pomanjkanje finančnih sredstev, odsotnost strokovnega znanja, varstvo prava, slabo kakovost podatkov in uporaba podatkov, pomanjkljiva varnost podatkov, pomanjkanje podatkov, nevednost o vprašanjih, povezanih s podatki, pomanjkanje dostopa do zapletenih analitičnih orodij, zapleteno zbiranje podatkov, in podatkovne pridržke. V tezi je predstavljenih šestnajst najpomembnejših možnosti uporabe poslovne analitike v malih podjetjih, ki so izboljšana podpora za podatke, izboljšan postopek odločanja, prilagodljivost in svoboda izbire, osredotočenost le na najpomembnejše zmogljivosti, pridobivanje vpogledov iz podatkov, pridobivanje konkurenčne prednosti, varčevanje stroškov, izboljšane odnose s strankami, določanje prihodnjih možnosti, izboljšano upravljanje zalog in dobave, meritev uspešnosti in uspešnosti, določanje novih prednostnih nalog, prepoznavanje poslovnih težav, ki jih je težko opredeliti, izboljšanje izkušenj strank, izboljšanje segmentacije strank in prilagajanje poslovne strategije. Seznam priložnosti kaže, da se poslovna analitika lahko in mora izvajati ne samo v velikih podjetjih, ampak tudi v manjših. Omogoča jim, da na dolgi rok premagajo sedanje izzive in pospešijo rast poslovanja, kar je vredno preboleti vse ovire, ki so bile zgoraj odkrite.

Glavni prispevek tega dela je nabor rezultatov, ki malim podjetjem lahko pomagajo pri prepoznavanju najustrezenjšega sklopa koristi in vprašanj zanje. Vsako podjetje je edinstveno kot kombinacija vseh poslovnih komponent, zato zahteva edinstven nabor priložnosti in ovir z opredeljenega seznama.