LUT
Lappeenranta
University of Technology

LUT School of Business and Management
Master's Program in International Marketing Management

Master's Thesis

Utilising knowledge discovery for effective customer acquisition and in data-driven business development

Sebastian Paavilainen

1st Examiner: Anssi Tarkiainen
2nd Examiner: Pasi Luukka

ABSTRACT

Author:	Sebastian Paavilainen
Title:	Utilising knowledge discovery for effective customer acquisition and data-driven business development
Faculty:	School of Business and Management
Master's Programme:	International Marketing Management
Year:	2019
Master's Thesis University:	LUT University
	75 pages, 14 figures, 7 tables, 3 appendices
Examiners:	Associate Professor Anssi Tarkiainen
	Professor Pasi Luukka
Keywords:	geodemographic, business analytics, customer acquisition, network effect

Analytics and solid business performance have had strong correlation over time, as data and analytics enable to streamline the processes, allocating the resources more efficiently, and moreover, gaining competitive edge over others. The benefit from the customer perspective is to gain more personalised services and products, and organisations are able to focus to most prominent customer segments fully. More precise analytics require multi-dimensional data, for instance, geodemographic dataset to deliver more value and with more accuracy to the customers. This research studied how the geodemographic analytics can be used to acquire customers more effectively and how the business can be developed into data-driven environment. Furthermore, aside from other studies in the same field, this study had network effect in the business environment, which also had to be taken into consideration in the area and customer selection.

The aim of the research is to find how the analytics can be used in the customer development and developing an organisation to be data-driven. The results suggest that there is positive correlation with certain attributes to the positive output of the business operations, yet, there is not specific type of cluster of ideal customers. There are variables, which are affecting the results indirectly, and therefore, more profound analysis of the results is necessary. Embracing data-driven decision making require quite significant steps to transform data to knowledge, which might not result by incremental business development.

Tiivistelmä

Tekijä:	Sebastian Paavilainen
Otsikko:	Tiedon louhinnan hyödyntäminen tehokkaammassa asiakashankinnassa ja dataan pohjautuvassa liiketoiminnan kehityksessä
Tiedekunta:	School of Business and Management
Maisteriohjelma:	International Marketing Management
Vuosi:	2019
Pro Gradu tutkielma:	LUT University
	75 sivua, 14 kuviota, 7 taulukkoa, 3 liitettä
Tarkastajat:	Associate Professor Anssi Tarkiainen
	Professor Pasi Luukka
Hakusanat:	geodemografinen, business analytiikka, uusasiakashankinta, verkostovaikutus

Analytiikalla ja liiketoiminnan tuottavuudella on ollut pitkään suoranainen yhteys, kun dataa ja analytiikkaa on voitu hyödyntää prosessien tehostamisessa, resurssien tehokkaammassa allokoinnissa ja tärkeimpänä, saavuttaa kilpailuetu muihin toimijoihin nähden. Asiakkaat ovat hyötyneet analytiikan käytöstä tehostettuna ja kohdennettuna hyödykkeinä, ja yritykset ovat voineet fokusoida tuottavimpiin asiakassegmentteihin enemmän. Tarkemmat analyysimenetelmät vaativat monimuotoista dataa, kuten esimerkiksi geo-demografista dataa, jotta asiakkaille pystytään tuottamaan enemmän ja tarkemmin lisäarvoa. Tässä työssä tutkittiin, miten geo-demografista dataa pystyttiin hyödyntämään tehokkaammassa uusasiakashankinnassa ja dataan pohjautuvassa liiketoiminnan kehittämisessä. Samankaltaisia tutkimuksia on tehty mutta tässä työssä verkostovaikutuksella oli oma merkityksensä, joka tuli ottaa huomioon analyyseissä.

Työn tarkoituksena oli selvittää, miten analytiikkaa voidaan hyödyntää asiakkaiden kehityksessä ja miten saavutetaan datalla johtamisen ympäristö. Tuloksien pohjalta voidaan sanoa, että tietyillä muuttujilla on suora sekä epäsuora vaikutus liiketoiminnan ulostulolle, mutta tiettyä ideaalia asiakasklusteria ei voida määrittää. Kuitenkin, useamman analyysimenetelmän avulla tulokset tuottivat selvempää kuvaa. Tulosten hyödyntäminen yrityksen datalla johtamisessa vaatii merkittäviä askelia datan muuttamisessa tiedoksi, mikä vaatinee enemmän, kuin asteittain tehtävää liiketoiminnan kehitystä.

ACKNOWLEDGEMENTS

To begin with, even though my journey at LUT University and Lappeenranta was short lived, I managed to enjoy my time in the city, appreciate the school, befriend with numerous of people and expand my networks. I've learned and developed a lot during my short journey at LUT, and therefore, it is time to move forward with my career.

First, I want to thank my examiners, Anssi and Pasi. Without your assistance and support, I would not have got this research done in this timeframe. You managed to steer me into the right path even in last steps. Secondly, Sanna, I want to thank you for this opportunity and your sparring throughout this process. Heikki, thank you for enabling this process and supporting me throughout it.

Lastly, and probably the most importantly, M.K, thank you for being there during this process. I've spent numerous evenings and weekends on this project and might have complained about the research, but you've still have been supportive. I have not expressed my gratitude enough.

Sebastian Paavilainen

In Helsinki, 4.2.2019

Table of contents

1. Introduction 1
   1.1 Background 2
   1.2 Research objectives and questions 4
   1.3 Literature review 5
   1.4 Definitions and delimitations 7
   1.5 Structure 8
2 Theoretical framework 11
   2.1 Business development 11
      2.1.1 Business development in customer acquisition 14
      2.1.2 Data-driven decision making 15
   2.2 Data to knowledge-flow 17
      2.2.1 Data processing 21
      2.2.2 Data integration 22
      2.2.3 Data cleaning 22
      2.2.4 Data in marketing 23
   2.3 Business analytics 25
      2.3.1 Data mining 26
      2.3.2 Descriptive analytic methods 28
      2.3.3 Predictive analytic methods 29
      2.3.4 Business analytics in marketing 32
3 Research methodology 37
   3.1 Research design 37
   3.2 Data collection 39
   3.3 Data manipulation & analysis 41
4 Results 45
   4.1 Statistical results 45
   4.2 Decision tree 50
   4.3 Random forest 58
   4.4 Combination of the analytical methods 61
5 Discussion 64
   5.1 Reliability and validity 69
   5.2 Theoretical implications 71
   5.3 Managerial implications 71
6 Conclusions 74

6.1 Limitations and suggestions for future research 75
List of references 76
Appendices 87

List of figures

Figure 1 An example of network externality

Figure 2 Theoretical framework and causalities of the elements

Figure 3 Structure of the study

Figure 4 Business Development Project process flow

Figure 5 Modified conceptual framework for data-driven decision making

Figure 6 The process of knowledge discovery in databases

Figure 7 Revised DIKW pyramid

Figure 8 Different types of business analytics

Figure 9 Confusion matrix

Figure 10 Area comparison of primary and secondary data sources

Figure 11 Correlation heatmap

Figure 12 Decision tree version 1

Figure 13 Decision tree version 2

Figure 14 The most relevant attributes and indirect effect of correlation

List of tables

Table 1 Performance measures

Table 2 Correlation classification

Table 3 Complexity table of decision tree

Table 4 Decision tree variable importance

Table 5 Decision tree classification performance

Table 6 Random forest variable importance

Table 7 Random forest classification performance

1. Introduction

The emergence of new trends such as *big data, artificial intelligence and* 5G technology has led to a situation where the data is an abundant source. This is due to the fact that each year, more data is created, and the sheer amount of data is growing exponentially. It is said that 90% of data ever created happened in 2018. (Marr 2018).

Although, there is tremendous possibility to utilise the generated data, not all companies are not utilising business analytics fully, even though, the benefits are imminent. The return of the initial investment, in terms of the benefits to an organisation, is significant. (Ramanathan et al. 2017). Using data in one's operations is not a novel idea as Amazon has been utilising artificial intelligence and machine learning to advertise items to targeted groups, enabling the company to use their resources more efficiently. Furthermore, data has opened new opportunities as companies emerge with data and its utilisation as a core business (Linden, Smith & York 2003; Columbus 2018). It can be concluded that nowadays transition towards the data centric business environment is more of a prerequisite than an option, in order to compete with the other organisations. Utilisation of data does not only limit to be used for business purposes per se, but for individualising and streamlining the services offered to consumers. Moreover, it can be concluded that potential of data expands to societies as well. Computing power, IoT devices and cloud services enable real-time analysis in different fields, such as hospital, commerce and entertainment services. (Reinsel, et al. 2017).

As it was pointed out that incorporating data and analysis to business function should be more of a prerequisite than optionality. The competition is fierce on many markets and industries, therefore using the resources to the maximum efficiency is essential. Moreover, using data for business analytics enables businesses to allocate resources to the right business functions. A link between analytics and business performance, however, it should be noted that there are several different factors affecting the performance as well. (Krishnamoorthi & Mathew 2018). Nonetheless, having competitive advantage is possible by introducing business analytics into the strategy (Côrte-Real et al. 2017).

1.1 Background

As it was mentioned, the possibilities of data and utilisation of business analytics is not fully exploited, even though many organisations obtain the necessary capabilities and on the other hand, there are forefront organisations, which show explicit performance increase after encompassing the pro-data environment. (LaValle et al. 2011) The initial starting point for this research is as the situation described above. Moreover, this research was initiated due to lack of more holistic analytic methods and perspectives. Some level of analytics in different elements, such as in marketing has been in use, but a more data-driven business direction is desired. Area selection and marketing decision are seen to benefit the most from the analytical perspective of decision making. Therefore, data-driven marketing is sought where the data is used to steer the decision, but also measure the results (Kumar et al. 2013). The business environment is business-to-customer (B2C) where large infrastructural investments are involved in the business actions. The main target group is a household, in and beyond the capital region in Finland. Therefore, marketing and sales are the active elements in delivering value to the household.

In order to understand the business environment and the purpose of this study more comprehensively, the product and the business idea behind the product should be elaborated. As mentioned, the product in question is a fixed infrastructural item used in everyday actions on the background. The idea is to find a cluster of potential households in a close proximity to each other rather than one or two wealthy households. In this environment, the number of the potential customers and leads determine the success and the entrance to the market area. However, the more there are customers are willing to purchase, it does not correlate directly to the price. It rather decides whether there is enough business potential to enter the area. A theory called Network Effect can be applies to this situation. The term network externality is used interchangeably with the network effect as well. The idea of the theory is that the more there are different entities involved in an infrastructure, or the infrastructure is on a certain level, more value can be delivered to the customers. (Na et al. 2013) In this case, the network effect theory cannot be applied in terms of pricing or increased value delivery (Lee et al. 2011).

In this research, although, the network effect works slightly differently as the network effect will be affecting the market entrance. The network externality in this business setting is related more to the acceptance and adoption of the product in the community. As a contrary, the more there are households, the less it may actually correlate with the adoption of the technology itself As Ramadani et al. (2018) and Brink & Rensburg (2017) elaborate, it may not be sufficient to analyse the customer in terms of classifying the different entities, but also to locate the most potential customers. Therefore, focusing the marketing and sales efforts to a specific area, with relevant number of potential customers, is to be the most effective.

The figure 1 depicts the situation of the business model, where the market selection method is based on the number of potential leads. As the figure 1 depicts, the more there are smaller communities willing to adopt the technology, the higher it will rank in market selection.

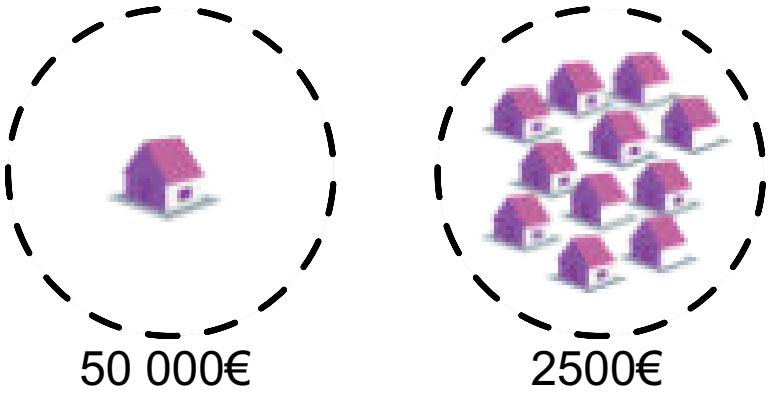

Figure 1 An example of network externality

There are two different data sources used in this research, previous purchase data and location based demographic data, which create the base for the study. The demographic data contains various different attributes linked to Finnish households and purchase data explains the previous sales performance in a specific area. The geodemographic data can be considered as a primary data and the sales data as secondary. These are not related to the data collecting methods as the purpose of each data is different and the geographical data is the more prominent. The data analysis is done with different methods, in order to gain more comprehensive insights and understand the situation better.

The aim of the study can be divided into two categories. *How business analytics can be utilised in the business setting* and *are what can be considered the main elements leading to the sales?* The result would be supporting the decision-making in order to acquire customers more effectively. Although, finding the most potential customers is not sufficient in this business environment. Moreover, it is important to identify clusters containing sufficient number of potential customers to focus the scope of the business. Moreover, the objective is to find whether there is a correlation between data and the sales. On the other hand, there is a chance that the customers are acting entirely irrationally, which would result in no correlation whatsoever. Nonetheless, it can be concluded that logical area selection is the culmination of the purpose why the research is done. The research questions will direct the study in the correct way in order to achieve the aim and the objectives.

1.2 Research objectives and questions

The goal of this study is to find whether business analytics in the given business environment can be used, in order to improve and take the customer acquisition further. The current methods to prospect new customers, is related to gaining knowledge of the service, online and offline channels of marketing and entering new areas where competition is not present yet. Also, taking into consideration different variables in the sourcing of the new potential areas, is acknowledged at its current state, but the aim is to delve into the analytics more. Business analytics would enable transforming different data, internal and external, rows of numbers and characters into actionable information. The key is to find common demographic nominals, which affect the result of the business actions. Therefore, the research questions would be the following, as proposed:

> Q1: How business analytics can be used effectively in customer acquisition in a network externality influenced environment?

Furthermore, the research question can be further be divided, in order to gain knowledge on the scale of the effectiveness of the business analytics. The proposed further research question would be the following:

> Q2: What kind of data can be used for customer acquisition and lead scoring?

Q3: What kind of analytical models and analytics can be used in customer acquisition and business development?
Q4: How can the analytical models and analytics be used for customer acquisition and business development?

The main research seeks to give answer whether the results of the analytics are valid and useful or entirely irrelevant. The latter three are more precise, and in return be more informative and drive the business towards right decisions.

As it has been pointed out, analytics, whether it is about the purchase data or geodemographic data is not a novelty, nor does it contribute to the research community as such. However, the proposed research methods and the business environment with network externality, which aim to satisfy the business problem, would fill the research gap. The business analytics methods would be using different types of analytics methods that serve different purpose and result in alternative insights.

1.3 Literature review

The base for this study is created by theoretical framework. In order to have a complete view on the topic, the theory must cover the whole process from data to actual decisions. Therefore, the proposed theoretical framework will include the following theoretical fields: data processing and transformation, business analytics and lastly data driven decision making. The theoretical framework is formulated to simulate the flow of information towards the results. The most crucial theoretical frameworks are depicted below. The figure below can be elaborated as following: processing of data and transformation of data is the base of the business analytics, and business analytics measures have to be changed according to the outcome. Therefore, it can be said that data processing and business analytics are a continuous flow until the specific results can be achieved. Further, the results of the business analytics can be used as a base for the data-driven decision making. Yet, the theoretical framework is introduced in different order in order to depict the requirements for the data-driven environment.

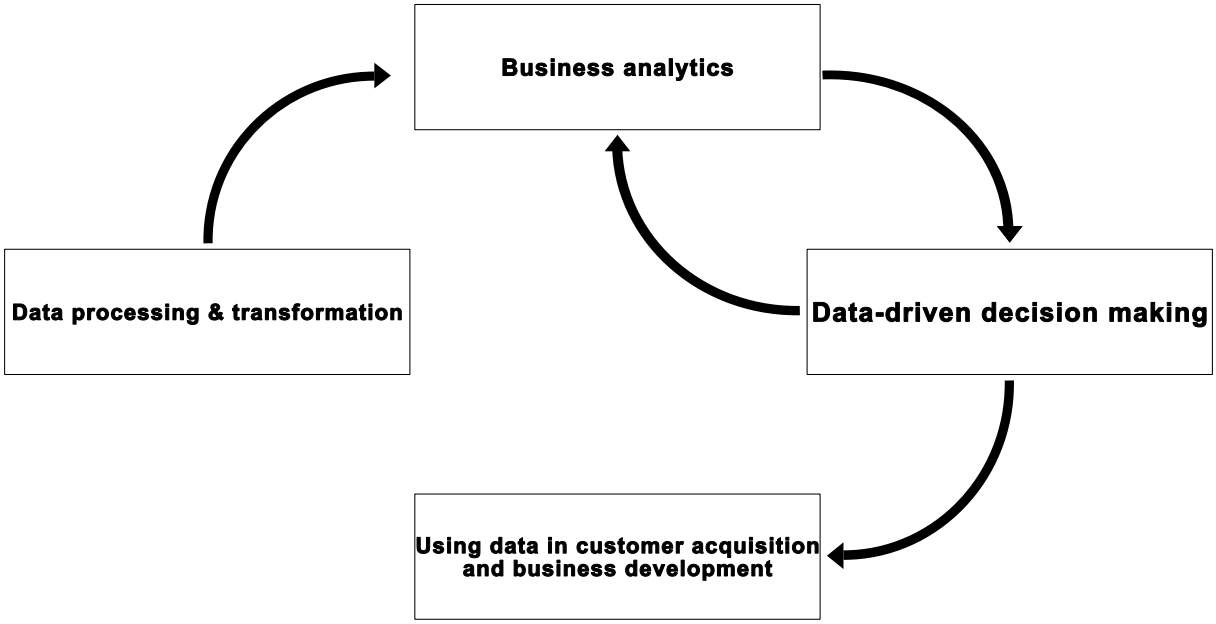

Figure 2 Theoretical framework and causalities of the elements

The material for the theoretical framework is collected from LUT's library databases. This includes secondary research data on data pre-processing, data to knowledge transformation, how data can be used in decision making and lastly business analytics and data mining. In order to extract the most relevant and recent data, the following key words were used, *data preparation, data pre-processing, data to information, data transformation, knowledge discovery, data-driven decisions, consumer analytics, target marketing, business analytics prospecting, business analytics and business development.*

Each of the element of the theoretical framework are not novel as each one has been around for decades. The only changing element is the complexity. Data processing and transformation has not changed, as similar tools and processes at hand are still in use. Only the quantity of the data has evolved over the years and tools have become more effective with the hardware. This is reflected in the data-driven decisions, as there are plenty of data to analyse, but the preciseness of the analysis is the key in the area. Lastly, business analytics has not changed much either, but the tools used in the analysis has developed significantly. The equipment is nowadays able to process larger data-sets, therefore, business analytics can be utilised further.

Bijmolt et al. (2010) researched different methods of analytics used in different stages of customer management. The research included different channels of marketing in order to capture the market value of the customer, moreover, the customer acquisition revolved around the customer segmentation. The segmentation was a base for the customer scoring methods in the study. Fan & Zhang (2009) and Brink & Rensburg (2017) introduced spatial data and geographical analytics, in order to segment the customers more accurately the researches included the GIS based geographical analytics, which was used as a classification, but also a predictive method for the customer behaviour and to locate the most potential clusters of customers. Both researches concluded that statistical analytics incorporating geodemographic data enabled more effective managerial decisions. The other angle to this research is the slight network externality factor. Kim & Lee (2007) have researched how network externalities affect the customer management, especially targeted customer acquisition. The argumentation was that even the less profitable customers may become more potential under influence of the network effect. However, there has been a similar study done in the Finnish environment, related to the adaptation of the hybrid cars. The study concluded that the attributes related to the adaptation were concerning education, income and household structure (Saarenpää et al. 2013).

1.4 Definitions and delimitations

The theme in this research is the market selection and other factors related to it, in order to access the highest number of potential clusters of households. Therefore, the most fundamental concept, which is addressed in this research is the geodemographic analytics. Moreover, it is not sufficient to know the background of the potential, but also where the customers are. (Brink & Rensburg 2017) It is also essential to understand the market and the decision behind the market selection. There are different factors related to the market selection, such as number of clusters of households, and therefore, the network externalities. This is also as essential concept to understand as it affects market entrance as there should be relevant number of potential leads, which affect the pricing, but also the probability of entering the market. (Na et al. 2013) In addition, the capacity and resources to enter multiple areas is limited, and therefore, the phenomenon of customer or lead ranking is introduced (Rhee & McIntyre 2008).

In order to maintain the scale of this research, the scope of the study has to be set, and limitations should be defined. The scope of this study is to cover only geographical and demographic data and therefore, it does not take into consideration more detailed substituting technological or infrastructural data in the prominent areas. However, the current technological infrastructure is managed in the current area selection, which concludes that there would not be overlapping aspects. For example, without going too much into detail, the supplementary technological infrastructures are the limiting factors, which affect the area decision. With the technical side of area analysis, this research would be too complex to be handled within the set timeframe and resources. Another crucial aspect is left out of the equation, which is any activity related to proceedings of a sale. In other terms, quantity and quality of marketing and the actual sales procedures are excluded out of the study as complexity of the research would be hard to manage. Sales person has a significant relevance to the purchase decision and marketing is driving consciousness to the product at hand. Therefore, having the results based merely on the data and the analysis, and not on the personal factors, the scope of the study is maintained.

1.5 Structure

This research is divided into six different chapter. Every chapter function as an essential role in the flow of the study. The figure 2 below, represents the process of the study. Furthermore, as it can be seen in the figure, the chapters are divided into steps, which have corresponding output.

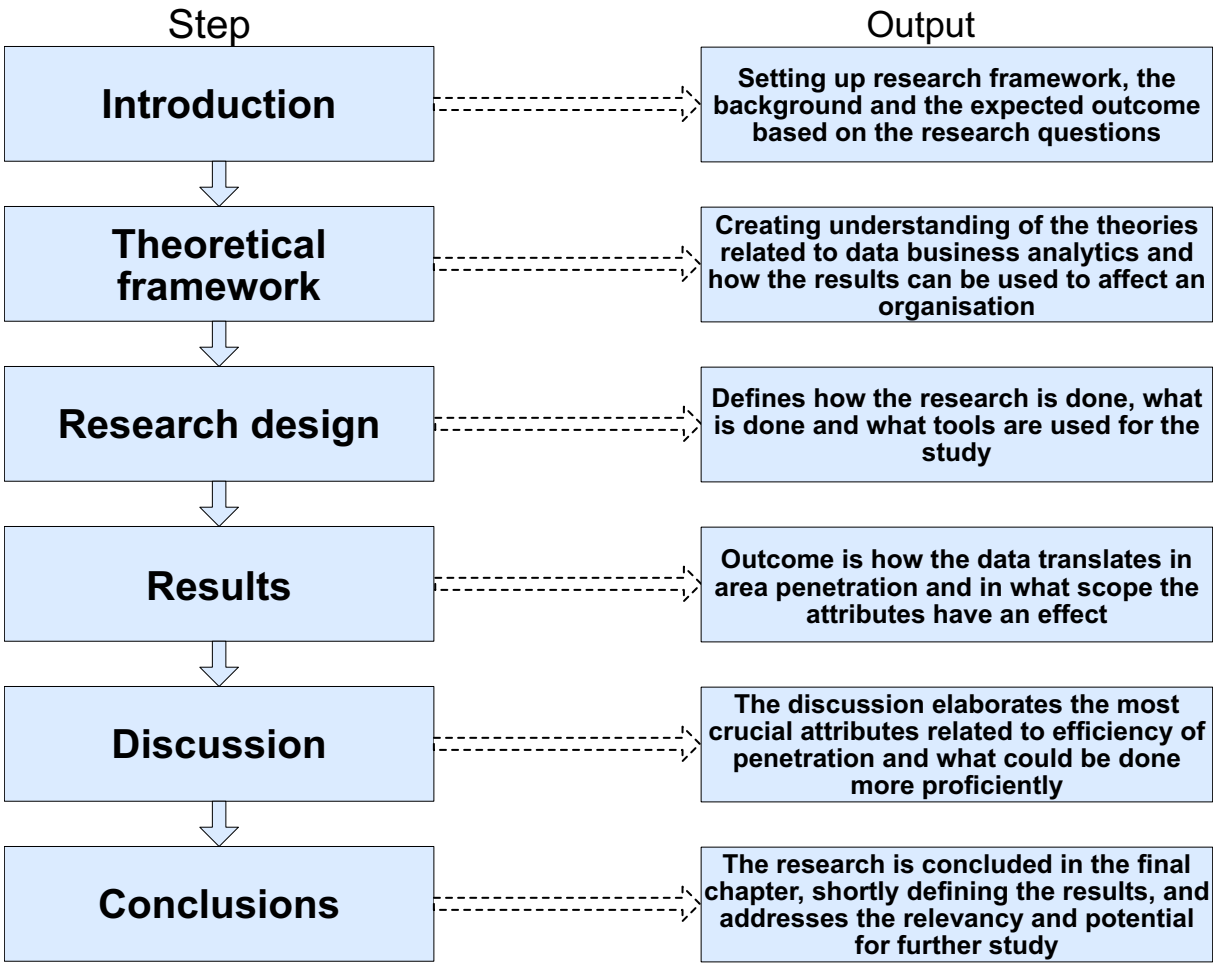

Figure 3 Structure of the study

As it can be seen, the study starts with introduction, which lays the foundation to the research. The current trends in the data and analytics in business environment are discussed, which is the initial reasoning behind this research. Aim and goals of the study define the research questions, which in return, define the theoretical framework. The second chapter binds the theories affecting the study as a whole. For example, data, business analytics and data-driven business development are defined, and the theory includes only the relevant measures and tools used in each theoretical category. Next chapter elaborates how the whole research has been formulated, starting from the characteristics of the study, what data is used and what tools are used to manipulate and process the data. The results from the data mining are presented in the result's chapter, without analysing the numbers any further. Discussion address more comprehensively the link between the results and how the results can be utilised elsewhere. Moreover, finding the causalities and reasoning behind the results is elaborated in the discussion part. In addition, the discussion defines the most important

factors to consider in future business decisions based on the research results. In the conclusion chapter, the whole research is summarised, reliability and validity of the study is analysed and lastly, reasoning for the future research is presented.

2 Theoretical framework

In this chapter, all the theories, related to the process of this research, are introduced. Starting from utilisation of the data and analytics, focused on how the methods can be used for developing an organisation to embrace the data-driven environment. Next, the study continues to what are the features of data and what are the necessary actions to process the data for business analytics. In the business analytics chapter, definition of the phenomenon is elaborated and then continued with business analytics methods as well as how these can be implemented empirically.

2.1 Business development

Business development (BD) as such does not account much in academic literature as the concept can be considered to be quite vague and does not hold value in eyes of scholars. BD is more generic term and an umbrella term for a various development activity. BD in this context should not be mistaken for new business development, which considers creating a new business area or service business development, for example (Paiola et al. 2012; Li et al. 2013). There is no common definition for the concept, as different entities have varying viewpoints. However, there are some studies done on the business development front and there is a proposed definition for the phenomenon. BD is defined as organisational value creation in long-term perspective from customers, markets and relationships. It can be concluded that BD can be used in almost any business application, based on the definition. (Ahtenhagen et al. 2017) As a contrary, Forsman (2008) argue that BD has two different lengths to it. Organisational development to deliver value can be done with two different methods, in a continuous gradual and progressive manner or in a radical way, which includes more than an incremental change in processes. It could be said that BD is used in every organisation as no organisation cannot exist without delivering value to different stakeholders. Although, business development and its characteristics and purpose vary from organisation to other. For example, organisations' size has a dramatic effect on the application as larger companies perceive BD in an entrepreneurial manner. On the other hand, smaller organisations have different view of the matter, as BD is linked to the growth phase of a company. (Ahtenhagen et al. 2017)

There are different tools, which can be used in the BD environment. As mentioned, BD can be incremental, or larger change, but also intentionality of the BD must be considered as well. A more incremental, long-term development process may come as a by-product of normal business actions, but more radical change is an outcome of a purposeful development project. One applicable tool, which can be used as a mixture of the two approaches, is a business development project (BDP). The figure 4. describes the DBP tool, which suggest BDP being a continuous flow.

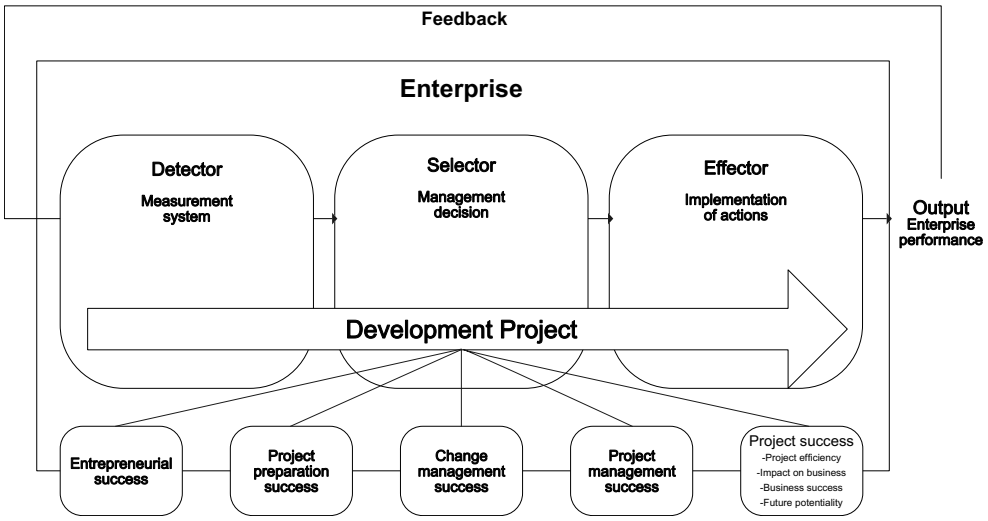

Figure 4 Business Development Project process flow (Forsman 2008)

The development project has to have goals and objectives, and moreover, the goals have to be aligned with business objectives. When goals and objectives are set, it is also crucial to measure the efficiency and effectiveness of the project. Furthermore, efficiency should take budgeting into account, but it must be pointed out that BDP has to correlate with company resources and plans. The success of BDP can be measured by effectiveness after a period of time. Benchmarking and assessing the current situation compared to the previous methods. Moreover, effectiveness can be divided down to impact of the BD, business success and potential in the future. The BDP has five different elements: entrepreneurial success, project preparation success, change management success, project management success and finally, project success. The four first elements include for instance managerial know-how, and also the usual methods to track and assess the project at hand. Further, four first elements are contributing to the project success in general. (Forsman 2008) Furthermore, setting realistic objectives of the BD outcome is something to consider when assessing the

risk and the benefit of the process. However, any BD activity carries out a potential risk and the risk and the benefit should always be weighted as a whole. Corporate culture plays a significant role as risk taking or risk averting culture might hinder BD process or push it forward in seeking the benefits. Breaking the corporate cultural barriers might be resulted in success or failure of BD activities. (Valentine 2003)

The business activities related to BD are different from company to company, but also from industries. What should be noted though, is that value creation for the company and the customers require the following activities: *seeking new opportunities* refer to utilising new customer trends or technological innovations. This way, value can be delivered in terms of new products or services. *Understanding the industry and competition* is crucial in case of more incremental BD process, which enable organisation to base their decision-making and place them in the market correctly. *Marketing and sales* are fundamental when considering new, innovative products, which are potentially more radical BD process, as the processes might change as an aftermath. Furthermore, BD activity in marketing and sales perspective is to find the potential customer but to take into consideration the organisational changes due to the development. In addition to the main three activities, there are three sub-activities, which can be linked to each activity as well. The approach to BD, incremental or radical could have relevance to the three sub-activities, as project-like BD is costlier and require constant monitoring in a reasonably short time. On the other hand, incremental BD does require monitoring and managing, and consumes resources in the long-run. *Capital management* is crucial in both approaches as mentioned, both methods require resources, but in different intensities. In order to successfully implement BD, resources, moreover, the monetary resources are important to secure throughout the process. *Talent management* is also related to resource management and gaining the widest knowledge and expertise. This would mean recruiting from external sources or finding the suitable candidate from internal reservoirs. Training the labour is applicable also to talent management, but considering the organisation's size, training and recruiting is affected by the size and therefore, the resources. Last sub-activity is *organisational structure and process development* that addresses the issue with new changed operations or products. When introducing changes in delivering value, what must be considered is how the process is changing and how organisation can support the change. Without managing the organisational structure well enough, the highest

potential of the BD may be missed. Additionally, processes are equally important to be handled as a whole, whereas incomplete processes might hinder organisations' future endeavours. (Achtenhagen et al. 2017)

2.1.1 Business development in customer acquisition

There are ways to incorporate business development into customer acquisition, however, more subtle and indirect approach could be considered as customer business development (CBD). The concept considers delivering value to customers by developing operations into more customer-oriented manner and integrating marketing with different stakeholders. Contrary to direct marketing activities, CBD is usually related to customer relationship management (CRM), by understanding the orientation of the markets and buyers and including the seller into to the equation. Sales personnel hold actionable knowledge, who work in-between of the different organisations. The stakeholder must share the market knowledge and be able to assess the market situation and the strategic customer orientations, in order to maintain the customer relationships (Hunter 2014). In addition, acquiring and maintaining the customers is more complex than it is thought to be. Not all customers are equal as potentiality of the customer and profitability vary. Segmenting the customers is crucial, but also incorporating the business analytics into segmenting and classification could be considered. Therefore, the marketing activities can be specified and made more effective, which in result, support the CBD activities. (Atul & Jagdish 2001).

On the other hand, marketing activities and marketing strategies are more concrete manifestations of BD. Changing the marketing strategies might be more radical than narrowing down the marketing activities. Change marketing strategies include for example segmenting the customer base again or re-doing the marketing mix altogether in order to change the approach in the market. Further, changing the focus on more customer focused rather than product or service, causes changes in organisations and processes. Value proposal and how the value is delivered faces also a dramatical change, when focusing more onto customer-oriented approach in marketing. (Souba et al. 2001)

When analysing customers, location, industry, socio-demographic and business factor are to be considered. When analysing the above-mentioned factors and including the result of analysis, a concept of geomarketing is used. There are couple more profound definitions for it, such as marketing method utilising the assumption that customers have similar demographic attributes in the same geographical area. Alternatively, techniques used in geo-based data mining assist in analysis more than in strategy formulation and decision making. As mentioned, greater changes in segmenting or targeting, using geomarketing is called geosegmentation, which might have significant effect on value delivery through BD. Geosegmentation can be used to find the most suitable customers in certain geographical area, but enables also to acquire the most potential and profitable customers as well. In BD, marketing activities in order to acquire customers could be limited to acquiring the most profitable customers in a new area or narrowing down the scope to a more specific location. Additionally, geomarketing can be used in planning of distribution or retailing points, but focus should be on personalising the geo-specific activities, such as marketing mix or advertising. When considering the risk/profit set-up, integrating GIS-based (Geographic Information System) analytics requires resources, but the benefit could overtake the risks and costs. (Ramadani et al. 2018)

2.1.2 Data-driven decision making

There are different terms for data-driven decision making and depending on the literature, the term varies from DDD to DDDM. The term and concept can be determined as disciplined data collection, its analysis, examination of the data and interpretation of the results (Mandinach 2012). Furthermore, it can also be described to be discovering potential knowledge from various sources of data by utilising business analytics methods. Data-driven decision making is entirely dependent on comprehensiveness and objectiveness of the data, but collecting such data is resource heavy. The progression in the technology however, eases the process as simulations and predictions are more prominent nowadays. (Long 2018) The benefit of use of DDDM is imminent as it has been studied that business leaders' decisional performance has increased when data has been the base for the decision opposed to intuition. The analytics in decision making is more or less becoming a norm these days

and it can be applied in various different applications, such as production estimates. (Dutta & Bose 2015)

As mentioned, DDDM utilises business analytics principles and data mining methods. The DDDM approach is using the same knowledge discovery flow as represented in figure 2, consequently, it takes the flow further. The figure represents the proposed model for DDDM. Based on the figure below, it can be seen that DDDM is a constant process as the knowledge opens possibilities for improvements. Furthermore, implementation of the decision and then analysing the impact on the information and knowledge, is going to improve over long term. (Mandinach 2012)

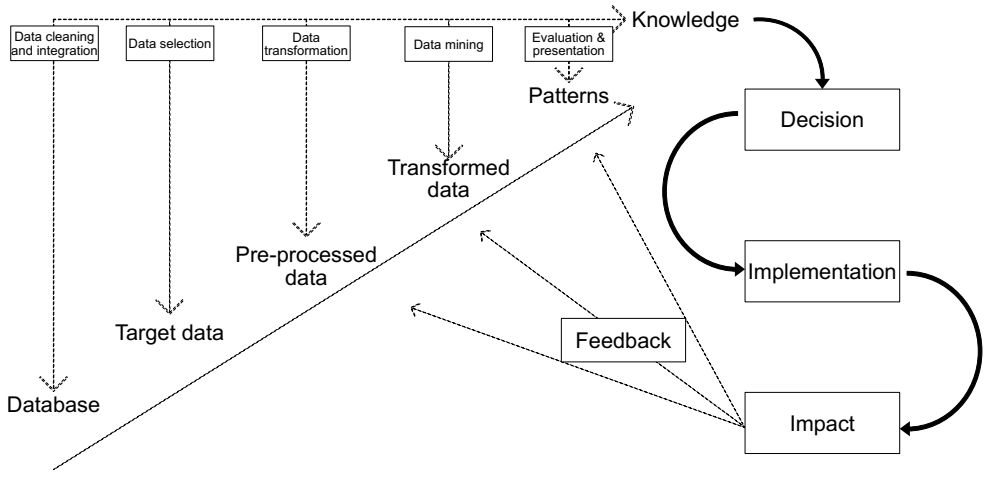

Figure 5 Modified conceptual framework for data-driven decision making (Mandinach 2012)

The implementation of DDDM methods can be considered resource heavy and depending on the industry and application area, there are additional factors to be considered. However, there are common nominals, which affect positively the adoption of DDD methods. Company culture, and proneness to technology are driving forces, however, the investments of the system will have limitation to it. (Brynjolfsson & McElheran 2016) Mandinach (2012) compliment the argument of the limitative factors with the issues related to DDDM. First and foremost, the human factor and the knowledge is the most relevant hindrance on the matter. Without sufficient number of qualified professionals being incorporated in applying the DDDM process, the outcome may not be the desired one.

When considering the link between BD and DDDM, there could be two different approaches. Ramadani et al. (2018) introduce in their perspective that decision-

making and BD are separate functions, which resulted after geomarketing determinants. However, the BD element in their study includes sales, products and clients, which are determined by geo-based analytics. Decision making acts as a solo element in their model. On the other hand, when assessing the Forsman's (2008) BDP model, the DDDM method could be considered as a one project. Therefore, it could be considered that the flow would first be DDDM and then transforming the results into measurable outcome.

2.2 Data to knowledge-flow

In order to understand the framework introduced, the fundamentals of data have to be considered first. The data itself can be described as information in an organised format, which has been collected in order to fulfil a need, relevant to a specific entity. The data can be in different forms, which serve separate purposes. (Davidson 1996, 3-5)

The data as a phenomenon has not changed from the definition above, but for example, the complexity and volume has changed over the years. Term, *big data,* has been around quite a while, but recent events in technology and other trendy phenomena related to it have made it popular. The term has emerged in 90s, but large datasets have been in use since 1960s, and the phenomenon has been rebranded using more complex technologies. Digitalisation and technological transformation have steered into environment where the big data is present. (De Mauro et al. 2016, 39; Smirnova et al. 2018) Despite the fact that the term is not new, and it is being used more and more in the 21st century, there is no one definite meaning. There are four different areas, which have an effect on the definition. The first one is *information*, degree in which data is gathered, spread around and used for different purposes. Secondly, t*echnology* and its changes have enabled large datasets to be handled and stored properly. Thirdly, m*ethods* of processing large datasets demanded firstly technological innovations, but in order to be effective, competence in analytics was necessary. This element is more related to the organisation and personal investments in the field of analytics. Lastly, *impact* of the big data affected the creation of the term. The data and its analysis can be applied to various different fields, which enable, in the best case, a competitive advantage over others. (De Mauro et al. 2016)

However, there are common properties of data, which emerge from variety of studies. These properties define the quality of the data, and therefore, classify the data to the term big data. Batini & Scannapieco (2015, 439) refer to the properties as 3Vs:

- Variety refers to diversity of the data, collection of data and how the data is visualised
- Volume is sheer size of the data, and the amount of data is increasing exponentially in years to come
- Velocity refers to the pace in which data is created.

However, depending on the viewpoint, there are other properties introduced. Lukoianova & Rubin (2014) argue that there fourth V to the quality of the data is *veracity*, referring to the accuracy of the data. They have broken down the property to objectivity, truthfulness and credibility. The fourth V is relevant due to the fact that without accurate data, the results may be just indicative. Even though, veracity can be considered quite relevant property, some literature does not take it into account when defining the term. Furthermore, Torrecilla & Romo (2018) add the element of complexity of data to the properties of data. The addition is quite similar to variety, as the collection is said to be done from sensors and different sources, but they argue that the complexity comes down to nature of the data. The data may be entirely unstructured, which creates complications when decoding the data into understandable and processable format.

The phenomenon of data to information to knowledge is described in the literature as *knowledge discovery.* The phenomenon has been considered as a flow or a process as it requires different steps, to the actionable intelligence. There have been different perspectives on the process/flow state of the data. Based on the causality or consequential aspect of the elements the information to knowledge could be considered as a process (figure 6), but on the other hand, the information flow is characterised as there are no specific, tangible elements to the process. (Bosancic 2016)

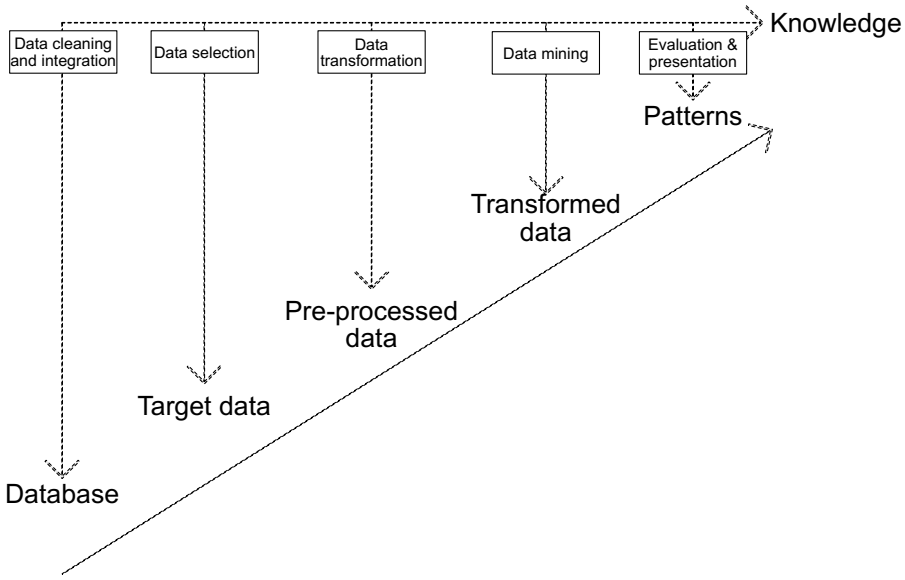

Figure 6 The process of knowledge discovery in databases (Maimon 2010, 3)

The process/flow of data to information is represented visually in a form of a pyramid. The term for the pyramid is a *DIKW* pyramid or hierarchy – Data, Information, Knowledge and Wisdom (figure 7). Therefore, the pyramid introduces the wisdom element, to more widely used data – information – knowledge flow. The DIKW consist of:

- Data: who, what, when and where, but also symbolic representation of different element and objects.
- Information: data in relation to other data in a specific environment or processed data.
- Knowledge is information, which is internalised, based on the insights it provides.
- Wisdom locates the knowledge to a schema, which results in actionable intelligence.

The data is the base layer of the pyramid while the wisdom being the top. The drawback of the DIWK hierarchy is that the data flows only one direction, from top down as data is non-existent without having base knowledge of the specific area. (Aven 2013; Jennex 2017) Bosancic (2016) add the element of understanding to the DIKW hierarchy, the purpose of which, is to answer the ‘why’ questions related to the phenomenon.

As mentioned, the opposed flaw of the current data to wisdom theory was that the information flows only one direction. Nonetheless, no other current trends, big data and so forth, or other factors are considered in the model. Jennex (2017) introduces a

revised model of the hierarchical pyramid, which deals with the drawbacks and the new technological advancements.

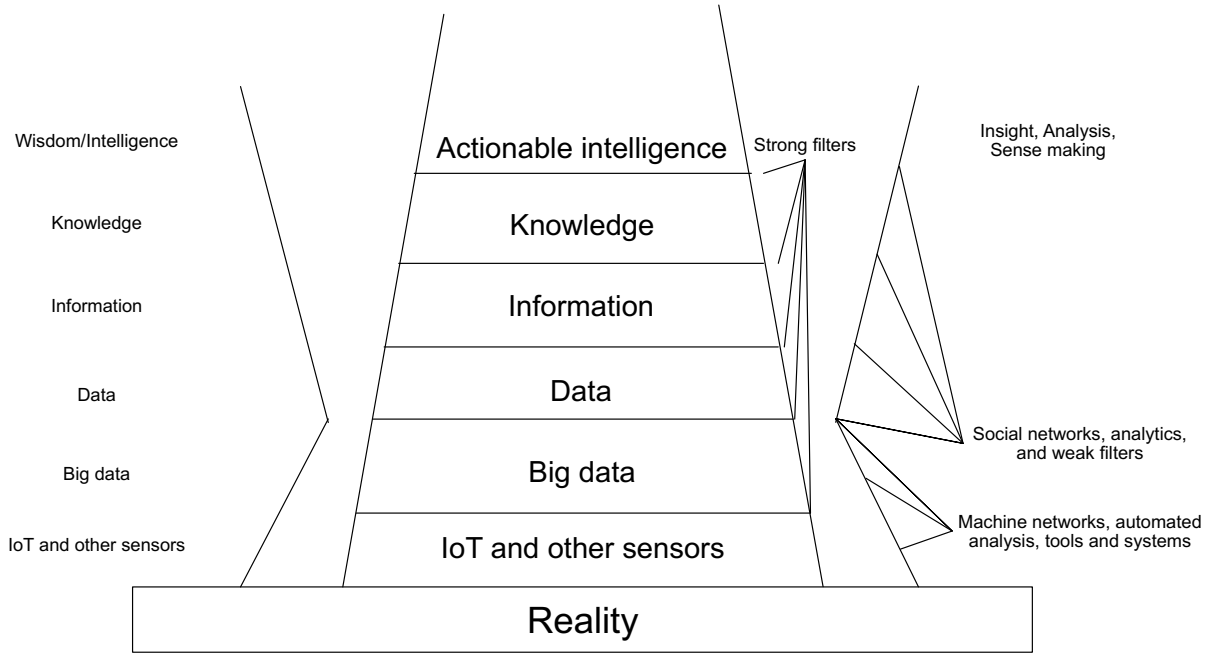

Figure 7 Revised DIKW pyramid (Jennex 2017)

The value created by the new improved model is that the complexity enables to address the technical effects on organisational and societal decision making. Different entities in the society and organisations are opposed with the need for more precise data and therefore, more knowledge-based decisions.

In some cases of information and knowledge processing, the dataset may be massive, creating a dilemma of how to present the results. Thus, delivering the valuable information in understandable format, i.e. visual format, plays significant role in the KDD process. However, the nature of the data usually dictates how the data can be visualised, and the visualisation can be divided to two different techniques – scientific and informational. Scientific visualisation is more complex and does not necessarily deliver understandable visuals for the general public as the technique illustrates, for example, what is the different attributes' relationship to the studied phenomenon. On the other hand, information visualisation is the driver for more explorative visualisation method. The technique enables interactive means, more visual items, such as color-coding, but as well optionality to manipulate the raw data. This creates more

comprehendible method, which enables hypothesis analysis from the visual representation. (Zhang et al. 2012)

2.2.1 Data processing

The base and the most crucial part of a research is the data. Processing of data and its transformation to a valuable state requires multiple steps and without necessary tools, it is impossible to do a proper analysis. The data is not always in the final form for the purpose and irregularities and values might be missing, which can cause problems. Moreover, processing itself is a more complex process, but an absolute measure. In addition, data processing or preparation can be considered a process, where the end result is an actionable information. The whole process from raw data to knowledge according to Squire (2015, 3) is the following:

1. Gathering data and storing the data accordingly
2. Identifying the problem, which needs to be solved
3. Cleaning the data and preparing the data for analysis
4. Data analysis
5. Visualisation of the results
6. Dissemination of the results

The list above can be a quite straightforward process methods, which guarantees basic results. However, Yu et al. (2006) introduce additional steps to data preparation, which include the following steps:

- Requirement analysis
- Data collection
- Selecting variables
- Data integration

When examining the two lists above, it can be concluded that there are certain factors, which are quite complementary, but the second approach adds complexity. García et al. (2016, 11) add measure of time to the data processing by stating that majority of the time spend with data analysis is allocated with data pre-processing, which should be considered in the timeframe assessment.

2.2.2 Data integration

Data fusion or interchangeably used as data integration is used, in order extract the most value from a single dataset. When sourcing data from a single outlet, the cost structure related to it can be considerably high when correlating the quality and complexity of the dataset. Not only, the costs related to a single source can be high and one-dimensional as the contents rely on the respondents. However, type of data has significant effect on the above-mentioned traits as sourcing from a database, or utilising pre-sourced dataset diminish the flaws. Although, the data could still not be exploiting the full capacity of data. Thus, the data integration does not extract novel information, but increases the value of the single dataset. (Davino & Fabbris 2013, 122) Moreover, Sorber et al. (2015), emphasise that integrating dataset is creating a possibility to extract more complex insights and increase the accuracy of the results. There are constraints for the process as the datasets have to be matching on the attributes, schema or dimensions. Therefore, integrating datasets, which have different size matrices, have different schema or the attributes represent entirely separate purpose, is possibly extremely hard. Further, recent technological changes are introducing problems with data fusion as the process is usually done manually and in offline environment. *Properties of data* play significant role in the data integration as the complexity, volume, velocity and veracity of the data is opposing new challenges, which require additional care to the data infusion. (Sagi & Gal 2018)

2.2.3 Data cleaning

Data revolves around every buzzword and creates a base for big data, data science and artificial intelligence. What should be noted though, is that in order to have valid and accurate results, the base work has to be done correctly. Furthermore, data has to be cleaned before any analytical measures can take place. (García, et al. 2016) Moreover, data cleaning should not be considered a one-time action as reiteration of data cleaning and storing will have to take place during the analytical process (Squire 2015, 4). Yu et al. (2006) argue that the process for initiating the data processing requires prerequisite analysis, which should be noted again in the data cleaning phase. The two theories indicate clearly that leaning ought to be done to serve the purpose of the analysis. The need for data cleaning comes from the datasets as integrating and

sourcing of the data may not be on the required level. The quality of data and the properties of the data are in play of the problems occurring the data. Furthermore, in case of integrating different dataset, it is crucial to understand the overlapping features and variables that might be occurring. (Rahm & Do 2000)

The process of data cleaning is occasionally called data standardisation and the techniques included in the process are essentially correcting, removing or amending the data inside. Furthermore, the type of the data dictates what techniques could and should be used. (García et al. 2016) The actual process of data cleaning has several different methods, which can be used in order to achieve usable and good quality data. *Reformatting* or *normalisation* can be used to standardise the dataset into a uniform manner. The actual information of the data does not change, but the data is represented differently and uniformly. Reformatting date is exemplary case, where there are several different methods to format the date. *Removing values* is crucial in order to have a reliable result, which is directly linked to designing the data processing. Finding the necessary nominals or requirements is the key. Furthermore, having blank values or values with no uninformative values are to be removed. (Randall et al. 2013) García et al. (2015, 13) add data noise identification to the techniques, which deals with unsystematic errors or variance. Noise identification does not correct the irregularities but detect them.

The presented techniques do not modify the dataset itself, but make it more uniform and standardised, which results in less errors and challenges in the analysis phase. García et al. (2016, 13-16) point out that data reduction is part of data cleaning but should be kept separate from the techniques presented above. Data reduction techniques affect the dataset itself, where for instance features are readjusted, and occurrences reselected. Additionally, numeral data can be transformed into qualitative variables.

2.2.4 Data in marketing

Data and analytics have been present in the marketing environment for quite some time, but many of the aforementioned factors enable more precise and real time analytics of the customers. The use of analytics has potential impact on the companies'

performance, but on the downside, implementing analytics can be costly. Utilising data analytics can be divided to three categories in certain business area, *customer acquisition, customer development and customer retention.* Each category bases its analytics to different sources of data. (Bijmolt et al. 2010) Emergence of purchase data has been one of the key instigators in the customer analytics sector as it dates back to 1940s. The historical data is the most widely used data type, which support the customer acquisition analysis in the future. The RFM (Recency, Frequency and Monetary value) is used to classify customers based on the profit and future potential. (Asllani & Diane 2011). Furthermore, due to the simplicity of the RFM data, the data should be complemented by other data sources. Demographic and geographic data is used in different applications but offer truly insightful knowledge. Demographic data has been a base for the customer loyalty program. (Brink & Rensburg 2017).

It can be concluded that integrating the first two sources of data enables various of analysis methods, which enable companies to provide customers customised items. (Wedel & Kannan 2016) Furthermore, demographic data can be further divided to individual, but also to household level. In addition, utilising geospatial data alongside demographic data can be used to find certain characteristics in a definite area. Usually the areas are divided into grids so that the areas can be more easily determined. This way, geodemographic data can be used to find common nominals in certain area, which are the most potential households/individuals. The analysis could be a base for customer segmentation and incorporate the geographic aspect as well. (Brink & Rensburg 2017). Consequently, besides the RFM and geodemographic data, in order to achieve more complex and accurate results, exographics can be used. Exographics in marketing perspective means contextual factors related to a single entity. The single entity can be an individual, a household or a community and exographic is aspects related beyond them. This data dimension consists of different scales, such as region and city, but as well as of domains, for instance, nature and culture. Exographics also add different stages in the relationship, which focus differently to the customer management. The stages are prospecting, acquiring, retaining, loyalty and partner. The exographic and demographic data are argued to be the most cost efficient. (Greene & Milne 2005)

2.3 Business analytics

Data analytics, business analytics (BA) and business intelligence (BI) are used synonymously, but the specific definition vary in each phenomenon. Data analytics is a pure form of science in which raw, data is examined and conclusions from the data are extracted (Praseeda & Shivakumar 2014). Business intelligence is more of a background function of BA as BI focuses more on tools related to information technology, data warehousing, visualisation and data mining, which enable BA. Moreover, BI is the systems in the background, which gather and disseminate data and BI systems are designated to the organisational level. (Arnott et al. 2017; Seddon et al. 2017) Nonetheless, BA and BI have some similarities, and some say that BA is just a subset of BI, on the terms, such as analysis of the data. However, there are multiple interpretations for BA and not one established definition, but the main message is that BA focuses on processes related to supporting business actions by examining, calculating or inference of data. Furthermore, BA utilises mathematical or statistical analysis, which enable BA to deliver business value in different applications. (Holsapple et al. 2014) Seddon et al. (2016) argue that capability in the organisation is the most crucial asset to deliver business value. Moreover, they add that specific processes related to business analytics exist, which include recognising and defining the problem, creating a solution to address the problem, incorporating different resources to the problem solving and then monitor and assess the results.

Term artificial intelligence (AI) and machine learning are usually associated with BA, so in order to understand the data extraction methods, it is crucial to understand the basic terms. AI emerged already in 1950s, but focus on big data for example, has accelerated AI's importance, moreover, the processing capacity has exponentially improved in the later years. AI is defined as study of making intelligent machines, but also system, which is able to analyse the environment and take humanly actions autonomously. In general, AI is just an algorithm. (Tiwari et al. 2018) General misconception of machine learning (ML) is that it is a separate phenomenon from AI, but it should be understood that ML is just a subset of AI. The purpose of ML is to perform better than AI by learning from the experience. Teaching of the ML can be done under supervision or let the algorithm learn by itself. (Bini 2018) ML is capable to process large datasets, and due to computing power, the algorithm is possible to be

more accurate than a human. Encounters with ML and AI occurs on a daily basis, as virtual assistants or autonomous vehicles for instance, are based on an algorithm. (Tiwari et al. 2018)

2.3.1 Data mining

Data mining (DM) can easily be confused with data collection, but data mining is the most crucial element of knowledge discovery. However, some categorise DM to be knowledge discovery and others define it as a part of process of knowledge discovery. By the definition, DM means analysing data, discovering data patterns and extracting hidden insights in order to solve issues at hand. (García et al. 2016, 1-2) One could say that DM and business analytics are synonyms to each other, but on the contrast, Holsapple et al. (2014), define that business analytics aim to solve problems in a specific business issue. Therefore, it can be concluded that DM is a broader term for achieving understanding on the issue and is more of a discipline rather than method to extract crucial information.

The approach to data mining is determined on the business application, which require BA. BA/DM include three different types of analytical methods, which vary in complexity and the value adding prospect. In addition, usability of the results varies as well, as three first analytical methods are more information and insight based, whereas the most complex introduce decision capabilities to the results. The figure 8. depicts the complexity and value adding matrix of different business analytics methods.

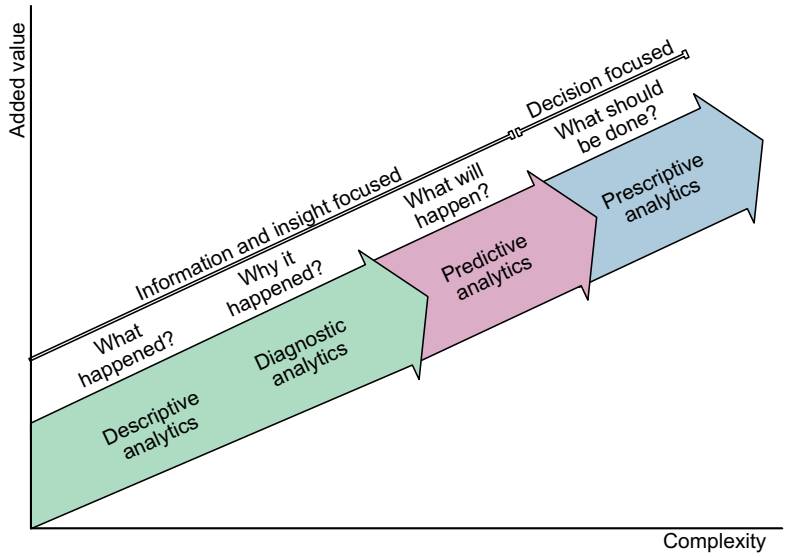

Figure 8 Different types of business analytics (Delen & Zolbanin 2018)

The most straightforward and less contributing analysis methods are descriptive and diagnostic analytics (Delen & Zolbanin 2018). Descriptive analysis aims to define what has happened and diagnostic tell why the specific occurrence happened. The benefit of descriptive analysis is that it provides the first step in order to gain insights on the company processes and performance as the data is based on historical data. Additionally, results of descriptive method are categorisation and classification, but it also consolidates the data at hand. Predictive analytics on the other hand, uses the historical data to assess the future outcome and possible scenarios and indicate what might be happening next. Pattern discovery is the essential in predictive analytics, which enables managerial input as the possible result is a forecast or simulation of the business issue. However, it should be pointed out that even though the complexity of the analytics, and the forecasting nature of the results, is not a direct guideline what to do next. Additionally, the results do not suggest any actions, but the analysis of the results have to be done on a managerial level and implement the actions accordingly. (Praseeda & Shivakumar 2014)

In each of these different elements in complexity, the concept of machine learning can be used. The ML is using the prior, historical data and which is then trained by new set of data. This enables more accurate predictions. Even though ML can be used for any type of analytics, ML is usually used for the predictive methods. *Neural networks or random forest* is exemplary use of supervised learning. Neural network, artificial neural network (ANN) or deep learning mean the same method, but the complexity changes

drastically from neural network to deep learning, however, deep learning is still neural network. The idea behind neural network is that the data-set is run through set of weights for each attribute and the result is based on the different weights.

2.3.2 Descriptive analytic methods

Clustering or cluster analysis is another widely used analysis method, which seeks to find a group of objects based on specified criteria. Clustering and ML are usually associated together as there are two different methods how the clustering algorithm can operate: supervised or unsupervised. One of the ML teaching methods is interlinked to how clustering analysis is operating. Unsupervised learning is clustering in its pure form and the data is not given any output parameters nor data labels, so the algorithm has to autonomously decide, which are the separate nominals. Unsupervised ML and clustering analysis enable to find hidden patterns. (Delen & Zolbanin 2018; Tiwari et al. 2018) In general, when there are set labels to data, it cannot be defined as clustering as the biggest revelations has already been determined (Tiwari et al. 2018). However, there have been studies where possibility of using supervised ML with clustering is being introduced (Peralta et al. 2016; Grbovic et al. 2013). Grbovic et al. (2013) argue that regular clustering does not consider ranking of the variables, and therefore it cannot be applied to every business application. Peralta et al. (2016) point out that performance of supervised clustering with multimethod is higher than with single method.

Even though the descriptive nature of the analytics, link analytics introduce predictive attributes to the analysis. Link analysis is a method used to investigate the relationship of nodes of a network. Furthermore, the analysis also predicts on a certain level the connections between the nodes. A node can be a human or an organisation, and therefore, link analysis can be used for social network analysis and ranking of the most potential entity. Used interchangeably, link analysis is also known as network analysis. (Delen & Zolbanin 2018) Link analysis can be used together with clustering in order to find common nominals and analyse relationship with the clusters. Huang et al. (2015) introduced a method to find linkages between different clusters. The clustering was done with different parameters to create base clustering and resulting in consensus

clustering. The link analytics was introduced to identify relationships with separate clusters.

2.3.3 Predictive analytic methods

Predictive analytic methods can be divided further down to statistical and symbolic methods. Statistical models are usually represented by mathematical models whereas, symbolic method uses less mathematical measures, and therefore, holds more direct interpretable results. (García et al. 2015, 4) Decision tree is one of the basics of the symbolic predictive analytics. The decision tree is an algorithm, which can be used in regressive or classification methods. The tree is a test on a specific case, and branches of the tree represents outcome of the test case. Furthermore, each leaf of the tree, i.e. node may include predictive measure of the possible outcome. The algorithm runs the test through and separates the results into multiple different branches and accuracy of the decision tree is determined by the number of nodes and therefore, determine the quality of the end results. (Delen & Zolbanin 2018)

Wu et al. (2016) argue with remark that normal decision tree has flaws, which may affect the accuracy of the results, by over-fitting the model to the data-set for example. This would mean that significant information may be over-seen due to the complexity of the tree. Wu et al. (2016) introduce constraint related to the algorithm, which would increase the accuracy and include the desired information. The combination of the two different methods, is introduced in the literature as Classification and Regression Tree (CART), which utilises the both methods in order to produce more accurate and reliable result. Algorithm of CART seeks both nonlinear and linear common relationships of the data-set, moreover, CART is usually used to find nonlinear structures and relationships. Ruiz-Samblás et al. (2014) add to the definition that the method uses categorical and continuous features in the analysis. Moreover, they add that in each partition of the branch of the tree, the algorithm tests all the attributes and the chosen attribute is decided by the fit and performance. CART can also be used for provisional analysis, which include predictive variables in larger content than the sample data.

Where combination of the two models differs from the basic classification tree, CART is able to reveal dependant relationships and most effective predictors. (Russell &

Macgill 2015) In order to have more comprehensive and sound results, the data and the methods have to be validated. As the decision tree is using training and test data, the values for each data set are selected by random, and multiple iterations are done to subset the data differently each time. This enables a different result that have to be aggregated to produce more reliable results. (Bramer 2017, 80-82, 213)

To overcome the drawback of the decision tree, random forest (RF) method is used in order to overcome those flaws. Random forest is similar to decision trees, as the model is using decision trees as a base and it is generating numerous different decision trees from a subset of the data. For example, random forest can iterate 100 different times capturing different attributes and irregularities from the average values, which the algorithm seems to be the most interested about (Anantha et al. 2006) As the decision trees are seeing the issues of overfitting the model and hence lowering the accuracy, random forest is able to produce more accurate results due to the iterations. Due to the complexity and number of the iterations, random forest is more demanding on the hardware. (Habeeb et al. 2018) Due to the *n* number of iterations the RF is computing, there is no need for randomised sub-setting of the datasets. Therefore, RF is a more complete analytical method from the beginning as there is no imminent threat of over-fitting of the model. However, the model can be further tuned by determining the number of iterations and number of selected predictors. It is suggested to produce large trees and forests i.e. numerous iterations but keeping in mind that the model is capable of overfitting the data if the predictor attributes for instance are not managed. (Zhang 2012, 163-168) RF is using supervised classification and therefore, it is extremely useful in finding significant variables affecting the desired output. The use of RF is quite unlimited and hence, it can be used in various fields. As it has been pointed out, the RF is seeing very high accuracy rate compared to other analytical methods. (Lovatti et al. 2019)

When comparing the output of the both analytical models, the decision tree is benefiting from the visual output, whereas the RF is more numerical based. However, the decision tree can produce numerical values as well, yet the variable importance is usually overlooked in the literature. In decision tree, variable importance is sensitivity analysis where the relative and independent attribute has importance over other predictors. The output determines how often the variable has occurred in the analysis

(Delen et al. 2013). RF on the other hand, is relying on the variable importance. As the name states, the value is how important the attribute is to the splitting of the nodes. (Zhang 2012, 163-168) The values the RF model produces are different from the decision tree (see chapters 4.2 and 4.3) as RF uses more mathematical formulas. RF is using mean squared error (MSE), which produces accuracy of prediction after opposing an attribute to alteration. The higher MSE value, the more important it is. (Ananta et al. 2006)

Cross-validating the data itself is not sufficient to determine whether the analytical model is sound, and whether the results are valid. The usual method of determining the performance of the analytical model is using the confusion matrix, figure 9. The matrix can be used to determine whether variable X has been classified correctly and moreover, how often the classification is done correctly. (Bramen 2017, 89-90)

	Classified Positive	Classified Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Figure 9 Confusion matrix

The confusion matrix is used for various different performance measures, which are elaborated below. There are multiple other performance classifiers, but the ones in the table 6 are the most common. The widely used accuracy measure is useful, but unfortunately it is not sufficient alone and other supplement measures are to be used (Carbanero-Ruz et al. 2017). However, it is argued that recall and precision are the ones, which holds the most information (El-Yaniv et al. 2017).

Table 1 Performance measures (Bramen 2017, 179; Carbonero-Ruz et al. 2017)

Measure	Formula	Meaning
Recall	TP/TP+FN	Correctly classified as positive of the all predictions

Specicity	TN/TN+FP	Correctly classified as negative outcome of all the predictions
Error Rate	(FP+FN)/(P+N)	Proportion of instances classified correctly negative
Precision	TP+/TP+FP)	Proportion of instances classified correctly positive
Accuracy	(TN+TP)/(TN+TP+FN+FP)	Proportion of correctly classified instances from the whole mass

2.3.4 Business analytics in marketing

Marketing and analytics are not a novel combination as analytics has been introduced in many different forms over the years. Moreover, the applications in marketing are increasing due to the sheer quantity of the data, and enhanced processing capabilities of it. For example, technical applications can enable eye movement tracking, which help in understanding the individuals' preferences and the data can be analysed on real-time basis. Merging analytics and marketing can allow companies to optimise and allocate resources more effectively. Analytics can give insights of how well the company is performing and how well the portfolio is perceived. This would translate to sales and further down to how well the initial investment in marketing is paying out. Similarly, as data analytics being more of a compulsory part of businesses, integrating marketing and analytics is towards a similar scenario. It has been discovered that by having analytic driven marketing, competitive advantage can be achieved. This has led to a situation where the analytics from a marketing perspective are constantly developed in order to gain the maximum value. (Wedel & Kannan 2016)

In general, marketing analytics can be divided into two basic categories, customer management, which include customer acquisition, development and retention and to social/web analytics. However, the line between, for example customer acquisition

analytics and web analytics is quite thin as both applications may be utilising the same methods. (Holsapple et al. 2018; Bijmolt et al. 2010) On the other hand, Erevelles et al. (2016) introduce a more abstract approach to combining marketing and analytics. They argue that marketing analytics provide more of a value creation approach, based on resource-based theory (RBT).

As mentioned, customer management can be divided to customer acquisition or selection, development and retention. Customer selection means selecting the most suitable customer based on the business and customer relations. The result of efficient customer selection would be a high *customer lifetime value* (CLV), which means the probability of net profit during the lifetime of being a customer. Furthermore, customer engagement is directly linked to customer value management, therefore, incorporating CLV for instance, to customer acquisition analytics is fundamental. (Bijmolt et al. 2010) Demographic data in addition to historical data can be argued to be one of the main data sources in customer acquisition, as it is crucial to know the customers. However, the demographic data itself is not sufficient, but requires something to complement the analytics. Transaction data from purchases, credit cards, IP or from log-in in online market place enable fundamental data of the customers to be collected. This opens possibility for example, for loyalty program, which in turn result in more data of the customers. (Bradlow et al. 2017) Fan & Zhang (2009) compliment the background by defining different attributes to a single customer. According to them, on top of sales and personal demographic values, there are other non-spatial attributes that should be included. The non-spatial attributes are divided to more geographical level as the demographic attributes include residential, work environment and point-of-sales area. There are multiple different methods to find the most suitable customer, starting from classification methods, decision trees and clustering as well as expanding to more profound predictive methods such as regression models and ANN. In addition, incorporating multiple different methods, more complex and accurate analysis of the customer selection and acquisition can be achieved. (Linder et al. 2004)

There are multiple methods for utilising geographic data in order to acquire customers more accurately. For instance, location analytics, map segmentation or geospatial data mining are used when discussing geographic based customer analytics. Geospatial analytics itself requires geographic analysis, which is further focused to customer level.

Geospatial data mining can be used to discover potential areas by analysing the demographics. The accuracy of the analytics can be increased by incorporating movement and traffic data in a specific area. (Lee et al. 2012; Ferguson 2013) Consequently, there are multiple ways to assess and predict the probability of the most potential leads. These analytical methods incorporate statistical methods in order to find the trend between different attributes affecting the positive outcome. Both classification and predictive models can be used in customer scoring. Lead scoring or customer ranking has been studied extensively and the outcome of the analysis is a go or a no-go decision of different inputs. For example, financial institutions have used scoring methods in order to determine whether a customer is a good customer or a bad one. These types of analytics take the customer history, but also the relevant demographic attributes into consideration. Additionally, the analysis can be extended to geographical applications as well as the potential customers can emerge as a cluster in a certain region. (Rhee & McIntyre 2008; Ahn et al. 2011)

When talking about customer relationship management (CRM) usually the initial phase of acquiring the customers is overlooked. The concept usually considers only customer retention and development due to the cost-profit setting. Nonetheless, it could be argued that in the CRM environment, customer acquisition and all the necessary traits should be kept in as high esteem as other customer management measures. (D'Haen & Van del Poel 2013) CLV and CRM are usually interlinked as the acquisition, the profitability and threat of customer leaving are crucial elements in CRM (Asllani & Diane 2011). Furthermore, the recency, frequency and monetary value is used as a fundamental data for the CLV calculations. The classical model of sales funnel can be used in the CRM and CLV based customer acquisition, where larger pool of potential customers is analysed and lesser and lesser end up being actual customers. D'Haen & Van del Poel (2013) introduce the sales funnel and focus more on the measures, which analyse the potentiality of the customer. A lead scoring is used in this case as well, but the difference from the lead scoring in the banking industry, the lead scoring does not have historical data to support the decisions. The model uses similarity or nearest neighbour model, which seeks the similar instances and ranks them. Depending on the model, Tillmans et al. (2017) add that social, demographic and geographic data can be used to score the leads.

Both customer development and retention are quite similar to each other as they use similar data-sets and analytics methods. Therefore, both are linked as one in this chapter. One of the most crucial data both elements use is the historical data of the consumption habits. Furthermore, one of the most significant analytics methods that both utilise, are ANN and regression models. As a contrary to ANN and regression model in customer acquisition phase, customer development and retention focus on the possible outcomes of the consumption and combining data to other similar clusters of consumers. (Bijmolt et al. 2014) One of the driving forces behind customer development & retention is personalisation and market intelligence. In order to serve the customer better, offer more personal items in terms of recommendations and integrate a customer loyalty program, knowing the customer is a must. (Bijmolt et al. 2014; Chen et al. 2012). On the contrary to just managing and developing the customers, predicting and analysing the customer churn and defection is also necessary. Bijmolt et al. (2014) argue that main element of maintaining the customer relationship and ensuring high CLV are commitment and effective loyalty program.

Value creation of marketing analytics is two-fold. The marketing analytics and analytics in general can create value in terms of market and customer intelligence, but also to the customer. Value creation could be considered as a constant flow, as gaining better market and customer intelligence means better customer value and increased customer value will translate better market intelligence. (Chen et al. 2012; Zeng & Glaister 2018) There are numerous methods of extracting insights of customer orientation, but the information is more or less obvious. What may be the most insightful, is the non-structured information. Non-structured data can be divided into verbal and non-verbal categories. Verbal is everything about written and spoken text, such as emails and tweets, but also actual spoken words. Non-verbal can be linked to verbal as facial cues and gestures may tell a different story than actual words. Also, geographical information is classified as non-verbal information. (Balducci & Marinova 2018)

As mentioned, increasing customer insight and customer value is a constant flow. Guenzi & Troilo (2007) elaborate on the customer value creation through marketing and sales. The study also mentioned product development being one of the features, which affect the customer value, but the most crucial aspect is inter-department

functionality. In addition, delivering customer value originates from the company culture and orientation as well. Firstly, value creation through marketing and sales can be linked back to customer engagement. Classification of the customer is one of the most functional methods to segment the customers correctly, and therefore, more precise marketing actions can be performed (Corrigan et al. 2014). Additionally, as Bijmolt et al. 2014) argue that commitment and loyalty programs, i.e. purchase data, is one of the effective methods to develop the customer relationship and retain customers. With analysing purchasing behaviour, and incorporating classifications, more personalised items can be marketed, but also suggested in e-commerce platform. This way, individuals can feel more value delivered.

3 Research methodology

This chapter includes the necessary features regarding this research. The characteristics of this research are introduced. Furthermore, the research data and how the data was processed and what measures were taken into consideration are presented. Lastly, the chapter elaborates the different tools and methods used in the research.

3.1 Research design

The characteristics of a study are defined by many different factors, such as type of data and how the research questions are formatted. Quantitative analysis usually answers to questions like why, what and how. However, the research questions cannot always tell the difference as the line between quantitative and qualitative is vague. Nevertheless, the limiting factor is the data. In quantitative research, the data is always in numeric form. The queries and surveys might can be in a text format, but the answer type of 1-4 is defining a study to quantitative study. Therefore, the data must be collected in a numeric format. Additionally, on top of collecting numeric data, the definition of quantitative research method state that the data is analysed in statistical methods in order to explain a certain phenomenon (Mujis 2011, 1-2). Saunders et al. (2015, 166-168) add to the definition that the phenomenon is in a form of an input-output analysis of the attributes. Furthermore, the results of the quantitative analysis can be represented in a graphical format.

Quantitative research method was usually related to deductive approach, but the recent events are leading also to an inductive approach. In general, deductive approach means using the data to test set hypotheses whereas inductive approach creates theory from the data. (Saunders et al. 2015, 166) Hoy (2010, 5-6) adds that deductive approach has assumptions in the background and the result is more specific prediction and conclusion from the vague starting point. Furthermore, inductive approach uses generalisation based on observations and trends, which can be tested in the future. Quantitative research method incorporates different measures in order to maintain the validity of the data and the results. This means that ensuring the questions for a survey for example, are entirely understood and the data is collected in a

standardised way. (Saunders et al. 2015, 166) Mujis (2011, 66-67) add that there are different types of validity, such as content validity, criterion validity and construct validity. Criterion validity is related to the theory, meaning that observable matters are related to theory. In addition, criterion validity can be divided into predictive and concurrent validity. Predictive validity addresses the theoretically expected outcome and how selected research instrument is affecting the prediction. Concurrent validity on the contrary means that the assumptions are less rigid and there are expectations that may have an effect on the results.

According to Saunders et al. (2015, 167) quantitative research is usually linked to experiential or survey types of researches. However, the limitation in use of experiential research is that the method is associated with the hypotheses testing although it is good to keep in mind that experiential research is considering the causalities of different variables to the phenomenon in question. (Thomas 2003, 51-52) Experiential research is occurring in more academic researches, but as this research is incorporating a corporate environment in a study, the case study is more suitable.

Case study researches the phenomenon in a research, and it seeks to find answers to what is causing the phenomenon to emerge and in what kind of situations the phenomenon deals with. Furthermore, a case study usually incorporates empirical study and an organisation to a real-life phenomenon. A case study usually answers to research questions how and why. (Farquhar 2012, 5-6) On the other hand, Simons (2009, 19-20) continue that data used for a case study is unstructured, for example interviews, and the research purpose more qualitative, while the aim is not to generalise the results to wider population. Therefore, it could be argued that only a business case and real-life phenomenon study can be applied from a case study, and the research method is more experiential.

Experiment research design means that each object in the dataset is treated in a specific way and then assessing the results and how and why the chosen method affected the objects. An object can be basically anything as experiential does not limit only to quantitative research. The aim in the experiential approach is to find what and how the changes in the treatment methods affect the causality in the objectives and

how much are the objectives affected. Furthermore, experiential method can be considered as a process where the first element is the objective itself, next the starting condition of the object, followed by the treatment and lastly incurring object. Saunders et al. (2015, 178-179) continue that the initial object can be considered as an independent variable whereas changes in other variables due to independent variable is called dependent variable. These two variables create a null hypothesis in which the assumption is that there is no significant difference to one another and the relationship between the variables in minimal. As a contrary, the alternative hypothesis is an option is the opposite i.e. a hypothesis of the phenomenon. The cases are tested statistically, and the threshold value is set to 0.05. Over the value rejects the hypotheses and values equal or below confirms the alternative hypothesis.

There are multiple different methods to approach this research, but due to the characteristics of the research method and the starting point of this study, the purpose is considered to be an explanatory and constructive research. Consequently, as there are no hypotheses to test and the aim is to find whether there is correlation between different inputs, the explanatory approach is on point. Purpose of explanatory research design is to find the causalities between different attributes and answering to questions why and how. In addition, centre of an explanatory study is to understand the phenomenon using different types of analytical methods, such as correlation and prediction. (Saunders et al. 2015, 176) As contrary, constructive approach is also taking different research tools in consideration, but the approach is focusing on interpreting the empirical part. Constructive approach is the connective part between research questions, theory and empirical part. The approach can be used to find novel methods to both practical and theoretical issues by representing the solutions in a graphical form or diagrams. (Oyegoke 2011)

3.2 Data collection

The data used in this research is comprised from two different sources. What should be noted is that there was no specific data collection done for the study, nor were there any interviews done regarding the area selection, as the complexity of the main data is significant. The data sets used in this research are geodemographic data sourced from Statistics Finland and the secondary data frame is from the business operations'

purchase data. Sourcing in the primary data means that the data required access, which was obtained for the research purposes. The statistical data is collected by Statistics Finland by various different methods, such as surveys and interviews, but also from governmental databases (Statistics Finland 2018a).

The primary data functions with two purposes, location based and demographic based statistics. The data is divided into either 250*250m, 1000*1000m or 5000*5000m grids depending on the dispersion of inhabitants in the area. Naturally, in the capital region the grid size is the lowest in order to maintain the scale of the data in the grid. The relevant sales data includes only the capital region of Finland, from which the demographic data was fitted to the sales data. In order to locate the grids on a map, the coordination system in the dataset is in ETRS89-TM35FIN format, which means that in order to gain full access to the data, specific geodemographic software has to be in use. The demographic attributes are divided into the following groups:

- Population structure
- Educational structure
- Inhabitants' Disposable Monetary Income Size and stage in life of households
- Households' Disposable Monetary Income Buildings and housing
- Workplace structure
- Main type of work activity

Furthermore, the above mentioned categorial attributes are divided into specific attributes. In total, there are 103 different attributes spread to categorial attributes. In this research, only household related, and the most relevant attributes are taken into the final dataset (Appendix I) In general, the dataset contained both individual and household level attributes. Moreover, the different types of attributes have been collected in different years as the primary data is valid from 2017 forward. The data is the most recent one available. However, the data within the different attribute categories might not be the most one up to date, but the most recent is used. Moreover, the household attributes may belong to multiple different attribute categories. For instance, household with small children may be included in the household with under school age children if there are different aged children in a household.

Based on reasoning, there are certain demographic factors that can be considered of having effect on the sales performance. As the product in question is a fairly expensive infrastructural investment, it could be argued that only certain level of income could

have an effect on the sales. However, due to the type of the product, as mentioned, households are classified to be the target customer, only the demographic attributes related to households are taken into consideration, and inhabitant levels are left out. The appendix I contains list of attributes, which are selected into the analysis.

The secondary data source is the sales data, which is collected directly from the given IT systems. Moreover, the data is collected by area, meaning that each area data is specifically selected from the source data. It is crucial to combine and manage the two datasets as mentioned, the primary data is divided into grids. The grids do not compare to sales areas perfectly and therefore, the primary data has to be scaled down to the secondary data. The function of the sales data is to collect information what is the number of potential households in the specific area, but also to determine the sales performance, i.e. penetration in the area.

3.3 Data manipulation & analysis

Data integration is the start for this research. Sorber et al. (2015) argue that combining different datasets into one, opens more possibilities and insights. Same principle applies here with geodemographic and the sales data, so that the area specific classification could be done. Combining the sales data to geodemographic data enables prediction of the possible scenario. The sales data provides the number of possible households in the area, but also the penetration. The penetration is calculated by the actual sales/potential sales. However, due to current circumstances, such as GDPR, the more precise identification down to household specific level data is not possible.

The research data could be considered to be naïve as each grid is divided into certain number of households, and those divided into sales, “1” or no sale, “0”. Due to aforementioned reasons, the household proximity and accuracy of the results could be argued. With more precise data, it could be determined who belonged to each household. Furthermore, data requires cleaning as the primary and secondary data do not match in terms of location. The secondary data is more area specific, whereas the primary data lays a grid over a map. This results in undesired households, which is depicted in figure 10, such as apartment complexes and the grids overlapping two

different sales areas. In order to maintain the scale of the two sets of data, the attributes in each data set are scaled to fit each other. Due to aforementioned reasons, the primary data can distort the data and therefore, the secondary data can be considered as the correct data. However, as the attributes are related to a household, not all attributes can be scaled down or up. For example, average square meter or number of persons in a household does not vary even though the data is fit to the secondary data. Additionally, if there are occasions when distorting attributes are in the data set, necessary rows are erased to balance out the data. These distorting attributes can be considered as outliers and data trimming is a practice to remove the outliers so that exaggerated outliers can be avoided (Korhari et al. 2005). Furthermore, the data is cleaned afterwards by changing the relevant attributes to relevant data format and all the NAs (not available) are removed from the analysis.

Figure 10 Area comparison of primary and secondary data sources

The data cleaning, scaling and integration are done by the area level as the data and the required software are the limiting factors. Nonetheless, the purpose of this research is achieved by utilising statistical software and in which, the required algorithm is coded in order to gain insights of the affecting attributes. The data mining and the base for the empirical part of this study is done using coding language R, and in a software RStudio. R is an environment for statistical analysis and computing, which also enables graphical output of the data mining. R is open source, which makes it extremely flexible platform for different uses. The language is usually used for statistics, clustering and graphical analysis (RStudio 2018) Another open source tool, which is used for this

study is a geographical information system (GIS), in order to extract the necessary information from the data. The system can be used to map different types of data on a grid based on a coordinates system in a dataset. (QGIS 2018) The process includes spacing the sales area in the GIS software and collecting only the necessary grids to be processed further.

There are multiple different methods, which can be utilised in fulfilling the purpose of this study. However, it is crucial to understand what is happening and what could be happening. Hence, this research uses a decision tree to understand the relation of different elements and predict the possible outcome. The decision tree in this study uses a CART-method, which is using classification and regression tree modelling for different purposes. Classification tree is using attributes, which are in non-numeric format, for example character type, whereas, the regression model benefits from the numeric attributes. However, due to the data format used in this research, regression model is used as the attributes are in numeric or factor format. An alternative analysis method is introduced to this research in order to gain more complex insights and understand the situation better. There are multiple different methods, which can be used in order to supplement the results from the CART analysis. An alternative tree method, random forest analysis method is used in order to overcome the issue with decision trees. Random forest is less graphical, and the output is in numeric format. Nonetheless, the results of the random forest, the importance of the other attributes to the output variable (sales) is compared to the results of the decision tree.

This study includes a comparison of the analytical methods, taking into consideration the necessary actions to transform the dataset to an integer and numeric format for classification tree. Additionally, the basic analysis of the dataset is done in order to understand the complexity, but also whether the data is valid to work with. In order to increase the accuracy of the results, the data is processed within the R software and introducing ML capabilities. A ML algorithm is used in order to train the algorithm first to recognise the data provided and further tested with separate data. The dataset is the same, however, the whole dataset is separated to training set and test set. Training test is naturally larger part in order to ensure the algorithm has enough data to ‘learn’ from. The separation is done on a 70%-30% division. Training data is 70% of the mass and 30% is reserved for the test data. The data is randomly split to 70-30 ratio several

times in order to cross-validate the results. The data splitting iteration is done, and the average of the results is calculated.

The confusion matrix (figure 9) is used to measure the performance of the analysis and classification of the results in the both analytical methods. The most fundamental measure is the accuracy, which is evaluated based on the true positives and negatives compared to all the results. However, not always is the accuracy a sufficient performance indicator and other measures have to be taken into use as well. The rest of the performance measures are elaborated in chapter 2.2.3 and in this research, all of the performance measures are used. The most important measures in this search, in terms of the predictability validity, are recall and specificity, as they determine whether the model is capable to identify the true and false negatives and true positives. Moreover, the model uses an overall accuracy measure, which is overlooked in the majority of researches. When assessing the figure 9, confusion matrix, the overall accuracy takes into account all of the correctly classified (True Negative and True Positive), which is divided by all of the instances. In the decision tree model, the data is split at least 30 times and the performance classifier is used for all of the instances. From the results, the average values for *accuracy, error rate, recall, specificity and precision* are calculated. Additionally, RF model is set to 500 iterations, so data splitting is not as relevant as in the decision tree. In order to compare the models, the same performance measures are used to analyse the results and the accuracy of the model.

4 Results

In this chapter, the results from the statistical analysis are introduced. Furthermore, the process of elaborating the results starts from the data itself and how the attributes affect each other.

4.1 Statistical results

Starting from the analysis of the data itself and the attributes themselves. When analysing the data with built-in function summary in R, the results are represented in Appendix II. The demographics can be divided into educational attributes, family structure, income levels, work-related attributes and attributes related to the household itself, such as average square meter of the property and size of the household.

The results of the summary of the data can be seen in the Appendix II. Starting from the educational background, the analysis elaborates that higher number of highly educated are located in the very centre of the capital region and in close proximity to universities. Therefore, average number of highly educated people in one grid is 26,21 and highest number available is 81. On the other hand, vocational diploma is highly represented on the list as mean of vocational diplomas in one grid is 26,34, whereas the highest number is 72. Number of only basic education or holding only matriculation examination diploma is quite significant as well. Mean for each is 13,16 and 9,31 respectively. Surprisingly, lower level of education (bachelor's degree) is not represented highly in the data as maximum number of lower level education holders per grid is 34 and average is just below 13.

The family structure shows that the majority of the households are either only adult or pensioner households. The average number of adult households in one grid is 15,08 and highest number is 55. On the other hand, the lowest number is 2, which might land in an area where the child households are the dominant family structure. Pensioners as a contrary have high numbers as well, average is 10,72 per grid and maximum is 37 in a grid. The household with children can be divided into three, small children (te_plap), children under school age (te_aklap) and school age (te_klap). The majority of the households have children in school-age, average of 8,5 and highest number being 25. On the other hand, the small children household represented with the lowest

numbers in the family structure with any children. The average was only 3,48 and highest number is 12. This could be affected by the desire to live nearby the services and near the city centre and moving to the suburbs when the children grow older.

The income levels indicate that majority of the households are belonging to the middle- or high-income level. The average of both are 16,84 (middle) and 22,93 (high). As a contrast, the low-income level category represents only with average of 2,3 and high of 25. More interestingly, the median income level of the household is placing to quite high a level as mean is 66 549 and high 105 565. The lowest value is 37 571. This can, again, be result of the area in which this research is done. Households are located in suburbs and moreover, in detached houses. This can be seen that average of the floor area is 112,8 and the highest one being 174.

Attributes related to work can be seen having the normal distribution in this setting and environment. As the basic services are close by and no basic industries are in the close proximity, it can be stated that services, wholesale, processing, transportation and health services had the highest numbers of individuals. Service sector being the highest occurrence is no surprise, but the processing and transportation can be considered as an out of the norm. Consequently, taking into consideration the nature of the product in question, the information and communication sector is not as represented in the data as expected.

The correlation of each attribute is more relevant than analysing each attribute as such. In general, the scale of the correlation is done on scale of -1 to 1. -1 representing negative correlation and 1 being the most correlation. The correlation can be used to determine, which values occur with relevant significance with certain attributes. The scale is broken down to 0.3, 0.5 and 0.7. The figure 11 represents the correlation heatmap of the attributes. The values represented below in the text are all in correlation values and the figure 11 represents the correlation values contrast to others. The values are represented with colours where the darker blue is high correlation and as a contrast, the reddish colour has no correlation. What can be seen is that some of the values are marked with “?”, which in this case means that the values throughout the data is constant and therefore, does not affect any other attribute or, as seen in the

appendix II, the certain attributes do not have any values and therefore, creates the question mark.

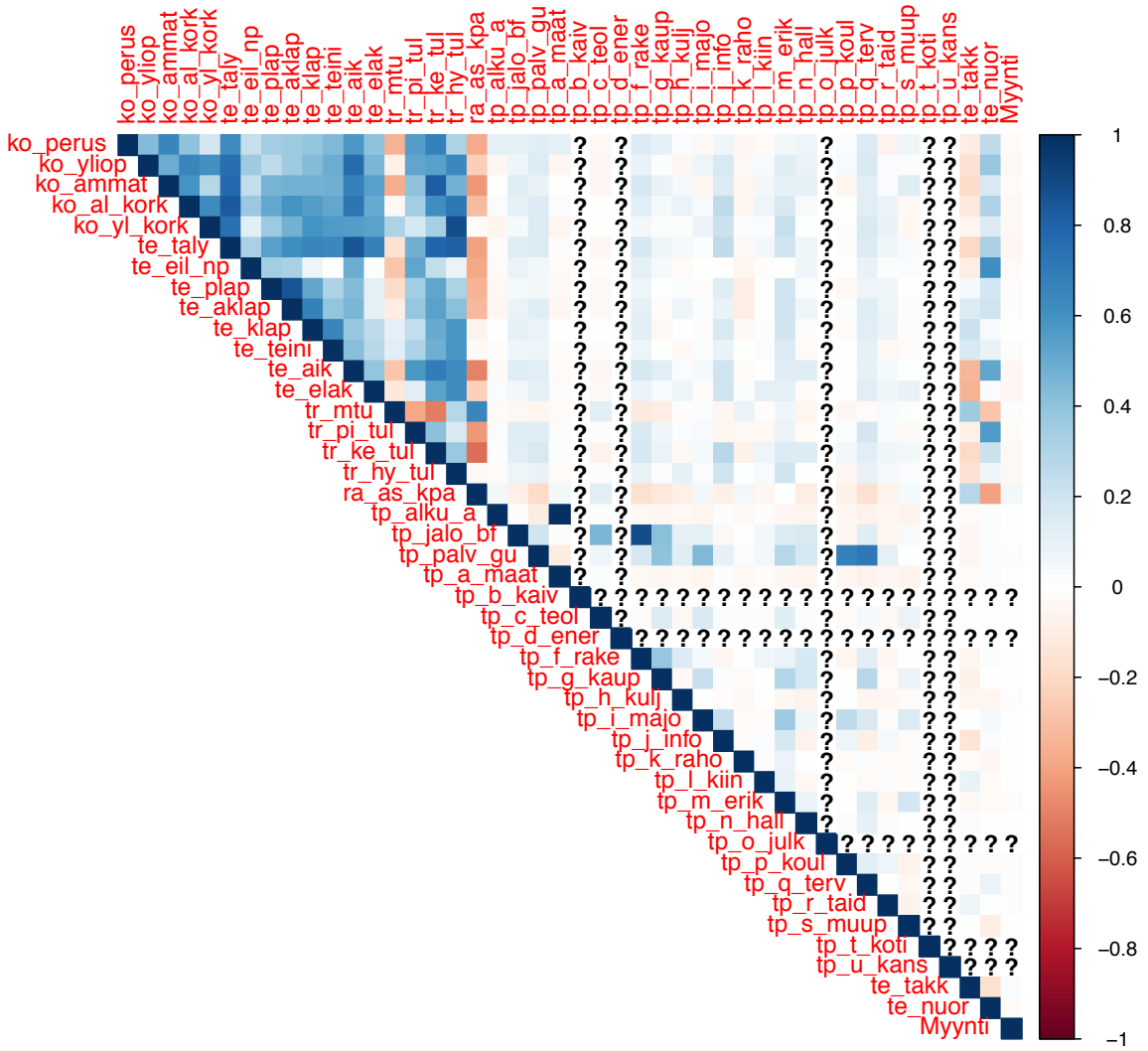

Figure 11 Correlation heatmap

The most interesting values in the figure are the attributes correlating to the sales. In the data, the sales are marked as ‘Myynti’ and based on the figure above, there is no single attribute, which has significant correlation to the end results. When analysing the values in the appendix III, the whole list of correlation values, it can be seen that there are only three different attributes, which have even a positive correlation to sales. Medium income (tr_mtu) has 0.018 value, average floor area (ra_as_kpa) value of 0.054 and primary production (tp_alku_a) has value of 0.007. As a contrast, the least correlation to sales can be found in the following attributes: pensioner households

(te_elak, -0.0483148566), adult household (te_aik, -0.0418157600) and number of households in the area (te_taly, -0.0442479523). Due to the nature of the product in question, surprisingly, no single industry, belonging to high income level, household size nor level of education seem to have any correlation.

On the other hand, the correlation between other attributes may indicate a naïve classification or clustering of the household. This means that in the data, there are certain attributes, which have numerous high correlation values over others. The conclusion can be made that there is a continuum or linkage between attributes. As seen in the table 7, the conclusion can be made that person with matriculation diploma (ko_yliop) has either lower or/and higher university degree (ko_al_kork, ko_yl_kork) and the person belongs to high-income category (tr_hy_tul). As a contrary, higher level education can be seen having more positive correlation to higher income. Other classification can be drawn from the vocational education (ko_ammat), and the person most likely has a basic education on the base, fall into medium income category and is in an adult household. The naïve approach can be due to the fact that this cannot be analytically re-validated, and the induction is done merely on the data.

The highest correlation between other attributes can be found from high income level (tr_hy_tul) and high education level (ko_yl_kork) with a value of 0.882196111. Consequently, the high-income level can be seen to have significant correlation to the number of households (te_taly) in the area and some correlation between matriculation diploma (ko_yliop), lower university level education (ko_al_kork), number of adult households (te_aik) and number of pensioners in the area (te_elak). All of these have correlation value between 0.62 and 0.71.

The second highest correlation can be found between the number of households in selected grid (te_taly) and number of adult households in the area (te_aik) with value of 0.85655889. Furthermore, number of households in an area can be seen to have the highest number of significant correlations, values over 0.6. The attributes, which correlate with number of are the following: basic education (ko_perus, 0.85655889), matriculation diploma (ko_yliop, 0.75832154), vocational diploma (ko_ammat, 0.77946580), lower university degree (ko_al_kork, 0.83402986), higher university degree (ko_yl_kork, 0.74972981), households with small children (te_plap,

0.61851003), households with small children under school age (te_aklap, 0.65544288), households with teenagers (te_teini, 0.64975460), adult households (te_aik, 0.85655889), pensioner households (te_elak, 0.67288144), household belonging to middle income category (tr_ke_tul, 0.79574886), and household belonging to high income category (0.81708523). This does not give too much information as the context is not displayed. Meaning that number of households does not tell whether low or high number of households correlate with abovementioned attributes. However, it can be concluded that finding the hidden insight in this context would enable apprehension. Moreover, finding a real demographic attribute, which holds significant number of high correlation values can be considered to be a naïve cluster of households. Besides the number of households, matriculation diploma (ko_yliop), appears to have the second highest number of values over 0.6. It can be determined that lower and higher university degrees (te_al_kork, te_yl_kork) number of households in the area (te_taly), adult households (te_aik) and high-income category (tr_hy_tul) are correlating with the attribute.

The third highest correlation can be found from households with small children and with children under school age (0.852423106). What should be noted is that the data allows a household to belong into different categories, such as abovementioned. The true correlation of these attributes is not imminent, but it may indicate that households have children with different age group, or the child household are located in same area. There are more such combinations, in the same attribute category, which are disregarded in the further analysis and implication of the results.

When taking into consideration the highest correlations (>0.6), a primitive classification can be, which gives an indicative result of the different individuals. The table 2 the most relevant attributes and values respectably. The table 2 is using colour scales, which indicate the correlation value. The darker the green, the higher the correlation value. The thresholds for the hues are the following: the darkest green <0.8, green 0.7-0.8 and the light green >0.7.

Table 2 Correlation classification

	ko_perus		ko_yliop		ko_ammat		ko_al_kork		ko_yl_kork		te_aik
ko_ammat	0.65739448	ko_al_kork	0.65638985	ko_perus	0.65739448	ko_yliop	0.656389849	ko_yliop	0.603543895	ko_yliop	0.767515984
te_taly	0.62505659	ko_yl_kork	0.60354389	te_taly	0.77946580	te_taly	0.834029856	ko_al_kork	0.620260773	ko_al_kork	0.724769224
tr_ke_tul	0.67276633	te_taly	0.75832154	te_aik	0.67981397	te_aik	0.724769224	te_taly	0.749729807	te_taly	0.856558888
		te_aik	0.76751598	tr_ke_tul	0.82211312	tr_ke_tul	0.619540697	tr_hy_tul	0.882196111	tr_ke_tul	0.697616993
		tr_hy_tul	0.64525308			tr_hy_tul	0.718540717			tr_hy_tul	0.627998916

What should be noted in this case, but in others as well, is that number of households may not classify an individual or a household, but it states that the number of specific attributes is well represented in the geographical area throughout. For instance, lower university degree (ko_al_kork), and number of adult households (te_aik) are well represented in the data and hence, correlation value >0.8 can be seen.

4.2 Decision tree

When assessing the importance of this research, the results from the data mining activities can be considered to be the core and hold more significant information. For the sake of the results regarding the data itself, it is important to understand what the analytics are based on. Starting from the classification and predictive model of decision tree, and with complexity table.There are 8109 different instances used, i.e. data itself contains 8109 different rows of valid data.

Table 3 Complexity table of decision tree

	CP	nsplit	rel error	xerror	xstd
1	0.0032946883	0	1.0000000	1.0006443	0.01315466
2	0.0018087903	1	0.9967053	0.9995424	0.01315444
3	0.0013206068	2	0.9948965	1.0012654	0.01319515
4	0.0012831883	4	0.9922553	0.9995653	0.01317518
5	0.0010436069	5	0.9909721	1.0008837	0.01321338
6	0.0010007246	6	0.9899285	1.0018465	0.01325014
7	0.0009486278	7	0.9889278	1.0028983	0.01326741
8	0.0008761243	8	0.9879792	1.0036775	0.01327651
9	0.0007721813	9	0.9871030	1.0035782	0.01327576
10	0.0007500000	10	0.9863309	1.0054051	0.01330405
1	0,0031382	0	1	1,0002701	0,01337725
2	0,0020030	1	0,9968618	1,0006295	0,01340506
3	0,0012604	2	0,9948588	0,9998558	0,01339068
4	0,0008108	3	0,9935984	1,0016756	0,01342992
5	0,0008040	4	0,9927875	1,0036326	0,01350376
6	0,0008006	6	0,9911795	1,0036326	0,01350376

7	0,0007679	7	0,9903789	1,0054028	0,01353207
8	0,0007172	8	0,989611	1,0077336	0,01356448
9	0,0007000	9	0,9888938	1,0092182	0,01359304

For the comparison's sake, the complexity table was split two times as the values in this section are not as crucial. It can be seen that the values do not vary too much from the other and therefore, does not hold that much value. As seen in the table 3, there are 10 or 9 different splits in the tree. There are issues with overfitting the decision tree, however, without sufficient number of splits, the decision tree may remain small, and therefore, does hold enough relevant information. The results presented in table 3 are derivated from the rpart-package in R and therefore, the definitions are unique. Starting from CP, which indicates the complexity parameter of the model and the split. In the first split, the complexity parameter is the highest and more there are splits, i.e. number 9 or 10, the less complex the splits are. Nsplit indicates only the number of the split done. The rel error gives out the resubstituting error in which proportion of the initial observations had misclassification by different factors in the tree. This means that the model evaluates the error in the observations in the tree. The value is decreasing from the first split and therefore, meaning that the training data is fitting better to the model the further down the splits go. However, this situation opposes the overfitting risk and there is no need to find the minimal error. Xerror is used to indicate the cross-validated error rate, which can be used to find the optimal tree. Lastly, the xstd (standard error) is used to tell the standard deviation of the error over the cross-validation sets. However, the values presented in the table 3 do not tell the whole truth of the data and its performance. The given values can be considered as indicative in the future. Furthermore, as explained in chapter 3, the new dataset is constructed, and due to the nature of the data, it can be considered to be naïve and thus providing the values above.

More importantly, in the table 4, represents the most relevant attributes regarding the variable importance. In this case, the variable importance value is a measure for how well the specific attribute is performing in the given the size of tree, as a primary or secondary splitter.

Table 4 Decision tree variable importance

tr_hy_tul	tp_g_kaup	te_taly	ra_as_kpa	ko_yl_kork	ko_yliop	ko_al_kork	tp_palv_gu	te_klap	te_aik	ko_ammat	te_elak
9	9	9	9	8	6	6	5	4	4	4	3
ko_perus	te_plap	tp_k_raho	te_teini	te_aklap	tr_mtu	tr_ke_tul	tp_m_erik	te_takk	tp_f_rake	tp_n_hall	tp_c_teol
3	3	3	3	2	2	2	2	1	1	1	1
ra_as_kpa	te_taly	tr_ke_tul	te_aik	ko_yl_kork	tr_mtu	tr_hy_tul	ko_ammat	tr_pi_tul	te_elak	te_klap	te_teini
12	10	9	8	8	8	7	6	5	4	4	3
ko_perus	ko_al_kork	ko_yliop	tp_m_erik	tp_r_taid	te_plap	tp_h_kulj	te_aklap	tp_g_kaup	tp_jalo_bf	tp_s_muup	
3	3	3	2	2	1	1	1	1	1	1	
te_taly	ra_as_kpa	tr_hy_tul	ko_perus	te_elak	te_aik	tr_mtu	tr_ke_tul	ko_ammat	ko_al_kork	ko_yl_kork	tr_pi_tul
13	12	9	8	7	7	6	6	5	5	3	3
tp_g_kaup	ko_yliop	tp_palv_gu	te_plap	te_aklap	tp_q_terv	te_nuor	tp_h_kulj				
3	3	3	2	1	1	1	1				
te_taly	ra_as_kpa	tr_hy_tul	ko_perus	te_elak	te_aik	tr_mtu	tr_ke_tul	ko_ammat	ko_al_kork	ko_yl_kork	tr_pi_tul
13	12	9	8	7	7	6	6	5	5	3	3
tp_g_kaup	ko_yliop	tp_palv_gu	te_plap	te_aklap	tp_q_terv	te_nuor	tp_h_kulj				
3	3	3	2	1	1	1	1				

It can be said that there is no single definite set of attributes, which comprise the importance table. As introduced in the chapter 2.3.3, the decision tree variable importance only defines how many an attribute occurs in the analysis, and therefore, no straight conclusion can be drawn. It can be argued that the variable importance holds less value than the actual decision tree, figures 12 and 13, as the variable importance have secondary splitters among them. The table 4 above takes a sample of the multiple iterations in order to show the variance in the different emerging factors. However, it could be argued that one of the most evident one is the average floor area (ra_as_kpa), followed by number of households in the area (te_taly) Furthermore, the household structure such as adult household (te_aik) or pensioner household (te_elak) can be seen ranking relatively high in the variance importance. Number of households (te_taly) is ranking significantly high but cannot be seen in the figures 12 & 13 as the attribute is acting as a secondary splitter attribute. However, the number of households (te_taly) could be seen as the factor affecting as the network effect and affect entrance to the market. Moreover, it is crucial to keep in mind the correlation values as the number of households was seen to have high correlation over many other variables (Appendix III).

Interpreting the decision tree is quite straightforward, the attributes and values of the left of the splitting point or a node are the positive ones, and the right side represents, the negative outcome. As a comparison, there were two models selected for the decision tree model, which were seen to have the highest occurrence. Starting from figure 12, it can be seen as a contrast from figure 13 as the main splitter appear to be higher university degree (ko_yl_kork) and too high a concentration is seen to have a

positive effect. From the decision tree, the second most important splitter is the average floor area of the house (ra_as_kpa) with value of 113m^2. Consequently, the number of basic education (ko_perus, over 4,5 household) is seen to have positive outcome. The low-income level of basic education holders is seen the affecting factor here. As a contrast, the vocational diploma holders (ko_ammat, under 29 households) is seen to be main splitter and furthermore, the work-related group of other service activity (tp_s_muup) and minimal occurrence of them is to have positive outcome to the sales. Furthermore, the high-income category (tr_hy_tul, under 15) households is seen in to have an effect to the decision tree and too high a concentration of them is not beneficial. It can be seen that smaller than 113m^2 houses result in more positively than larger ones, on the pre-requisite that basic education holders (ko_perus) the concentration is low, relatively high number of vocation diplomas (ko_ammat) are in the area and there is only fraction of certain work-related group. Furthermore, if there is relevant number of households belonging in the high-income category (ty_hy_tul, >15) the sales are more predictable.

As a contrast, the negative path of the figure 12, splitting done to the right-hand side. The most negative variables to the sales can be considered to be the median income of the household. Interestingly, the median income (tr_mtu) can be seen in three levels of nodes with different scope of values. The main and first node with more than 102 000€ median income can be seen to have very drastic effect on the sales, however, the instances seems to be low (21 instances). Therefore, it can be concluded that large house and significantly high income is negative to the output, yet it can be argued that the occurrence is not too high hence, the validity is questionable. However, if the median income is below above 95000€ and less than 84000€, the model predicts that the output is positive.

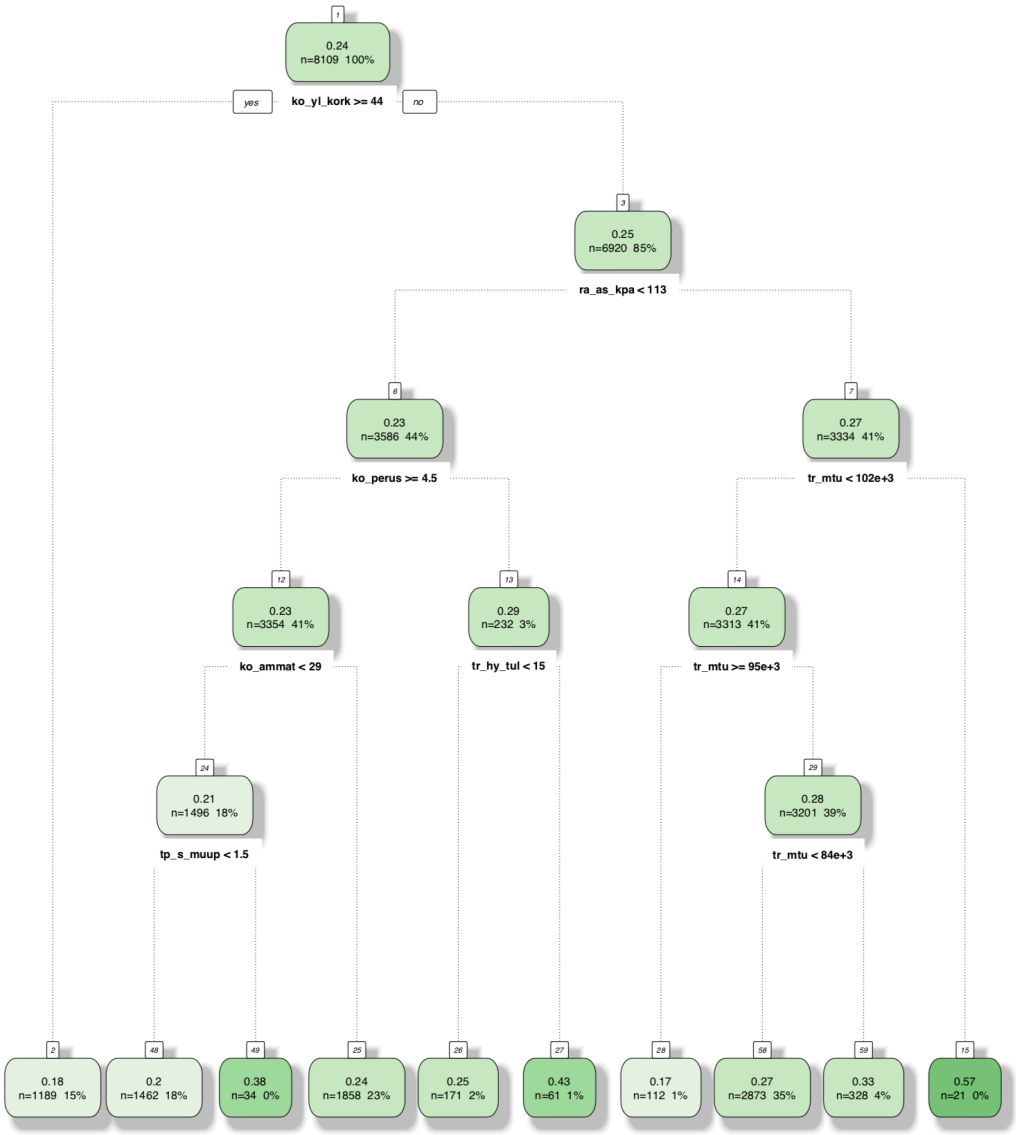

Figure 12 Decision tree version 1

In the second most high occurring decision tree, figure 13, the main splitter is the average floor area (ra_as_kpa). The main splitter is average floor area being less than 131.5 square meters, which indicate that larger houses (over the threshold of 131 square meters) are more likely to result negatively in the output. However, the Finnish society and culture should be noted when analysing the household splitter. When comparing to the national average floor size regarding detached and semi-detached, the value can be considered quite high. Based on the Statistics Finland (2018c), the average floor size nationally for detached houses was 111.8 and for semi-detached houses 71.4. What should be noted is that large house square meter has positive

correlation to median income (0.650372316) and has the highest importance value on the decision tree, yet there are no visible remarks in the decision tree.

The second most important splitter is the number of pensioners in the area, (te_elak), which acts as a main splitter in two different levels. Tt can be analysed that if there are more than 3,5 pensioner households, and later on, more than 12, it can be seen as positive. However, if there are more than 12 households, less than 25 households with vocational diploma (ko_ammat), and less than 19 households with small children (te_klap) the outcome is still positive. On the other hand, if there are more than 3,5 pensioner households, and less than 19 high-income households, the output is positive. Based on the results and figure 13, the less there is variance in the household structure in one area, the more negative effect it will have, whereas the diversity in the area is positive.

Continuing with the other end of the decision tree in the figure 13. The second most important splitter after the average dwelling size is the industry, or the service sector to be more precise. If there are less than 1,5 households where someone works in service sector is seen positive, whereas, the larger concentration is negative. On the other hand, if there are more than 34 households in the grid, the outcome is positive on the notation that there should be more than 14 households with lower university degree (ko_al_kork). It can be said that less there is service sector workers and the smaller concentration of household is located in the grid, the outcome can be predicted to be positive.

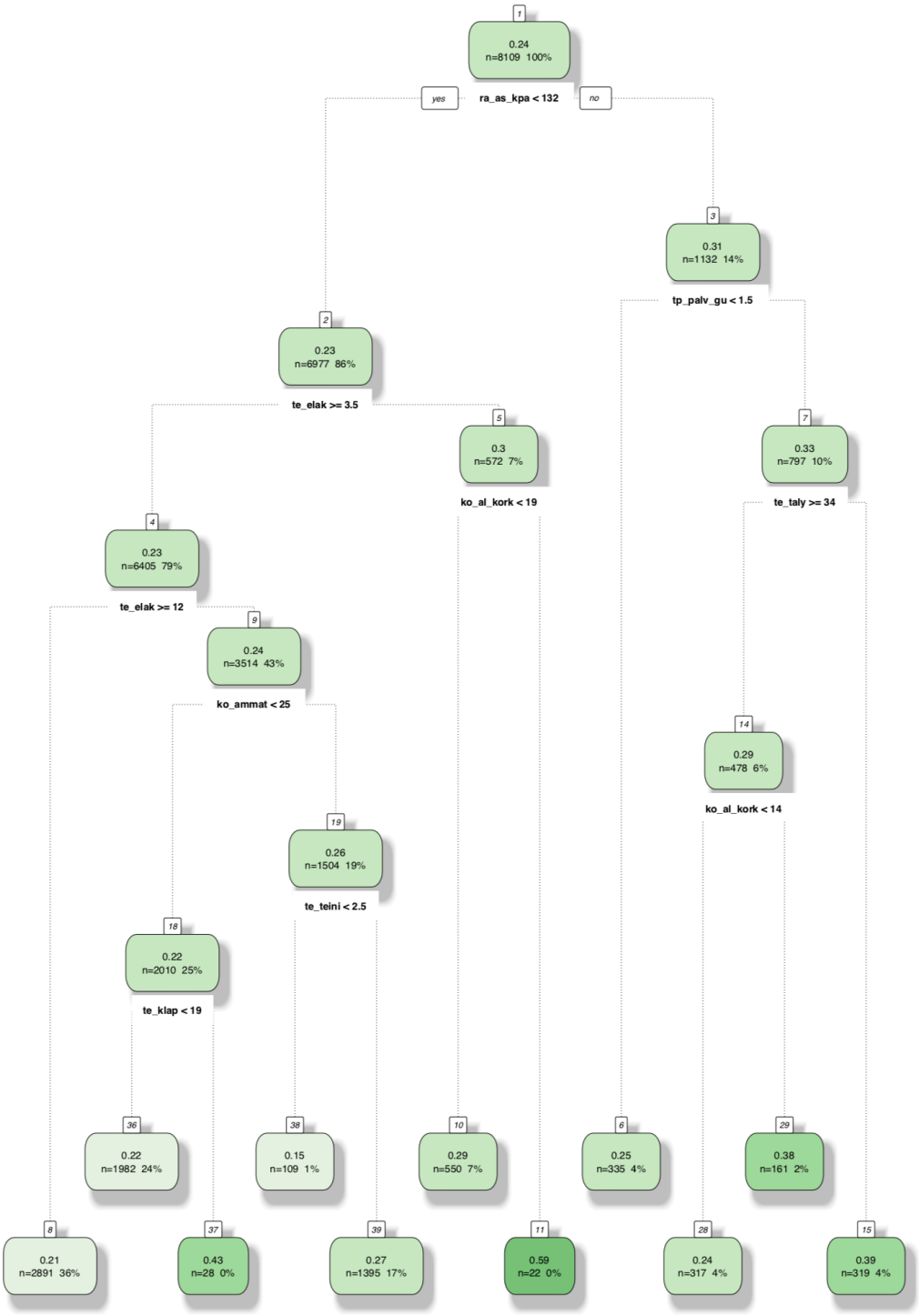

Figure 13 Decision tree version 2

The results are very useful in the managerial perspective, as the business focus could be adjusted based on the values and figures. However, all the results are useless if the analysis method is not accurate or it is not accurate enough. Therefore, the data and the results are validated. As mentioned in the chapter 3.3, the data was divided

into training and test data (70% & 30%) in order to gain more accurate data and therefore, have more reliable results to base the decisions on. Different classification measures were applied to this as results are the following:

Table 5 Decision tree classification performance

	Accuracy	Error Rate	Recall	Specificity	Precision
Average	0,765062051	0,234937949	0,419291437	0,769041445	0,756969641
Std. deviation	0,033664646	0,033664646	0,287484402	0,197859949	0,055645258
Confidence interval (L)	0,759318019	0,229193918	0,3702394	0,735281584	0,747475167
Confidence interval (H)	0,770806082	0,240681981	0,468343473	0,802801306	0,766464115

In order to check the performance of the model, the confusion matrix, figure 9 and the basic performance classifiers, table 1, are used. Moreover, in order to cross-validate the results, the results of the performance were iterated multiple times with different sub-sets of the data. This enables the data to be cross-validated by taking the average of the values into consideration. The most crucial ones in the confusion matrix are the true negatives and true positives, other ones are the demeaning factors in terms of performance.

The table above (table 5) shows that the overall accuracy (defined in table 1) holds a score of 0,76, meaning that the model is 76% of the time correct. Other times, 23,5% times the results fall into different category than true positive or negative. However, as mentioned, the accuracy does not always tell enough and other performance measures are complimenting the accuracy. Starting from recall (sensitivity) the value of 0.419 is relatively good, yet it still has a room for improvement. The results may inflict due to the nature of the data. Recall value is used for how effective the analysis is in identifying the number of positive outputs and contrary, specificity is used to measure the ability to tell the negatives from the analysis. It can be said that the model is more effective to identify the negative outcome. The specificity measure scores value of 0,769. The difference in the prediction performance in terms of the true negative and positives can be explained by the number of positive instances in the sales data was quite low and the number of negative outcomes relatively high (1189 positive observations vs. 6920 negative observations). Precision on the other hand is a value for the positives from the dataset. The precision value is 0,756, which also is

quite good, on the other hand, the error rate, which measures the proportion of correct classification of negative. The model results in error rate of 0,23, from which can be concluded that the error rate is quite low.

As the average values can be the most informative, it is crucial to understand what comprises the average value. The variation and dispersion could be significant, and therefore, the accuracy could be superficial. To understand the values, confidence interval is an estimate of the scope of the observed values, which might contain true values and the standard deviation is used for enumerating the variation in the values. When assessing the values given in the table 5, the standard deviation is relatively low in most of the performance indicators, except in recall measure. The recall has quite high dispersion in the values and confidence interval is quite large as well. Specificity on the other hand shows that the confidence interval is narrower and standard deviation is smaller. It can be concluded on the two measures that the variation can be seen as a demeaning factor in terms of interpreting the results and overall accuracy. On the most parts, the standard deviation of the values and confidence interval is relatively narrow, so on that front the values can be considered valid.

4.3 Random forest

In order to test and compare the results of the two different classification and predictive analytical methods, random forest was used to create a juxtaposition environment or alternatively, support the results of each other. The model was set to 500 iterations and the analysis produced the following results.

Table 6 Random forest results

Attribute	%IncMSE	IncNodePurity	Attribute	%IncMSE	IncNodePurity
ra_as_kpa	4.5564070	1.2325741	tp_j_info	1.0257134	0.3277153
ko_yl_kork	4.3044403	0.8749446	ko_yliop	0.8209485	0.7563512
te_taly	4.2951912	0.8502232	te_klap	0.6907726	0.7714518
te_plap	3.8472082	0.6002444	tp_h_kulj	0.4691358	0.3906435
ko_perus	3.7558818	0.7309686	tp_c_teol	0.4597067	0.2154126
te_elak	3.1696734	0.8175068	tp_b_kaiv	0.0000000	0.0000000
te_aklap	3.0574963	0.6556219	tp_d_ener	0.0000000	0.0000000
te_aik	3.0513641	0.7543536	tp_o_julk	0.0000000	0.0000000
tr_hy_tul	3.0085571	0.8305508	tp_t_koti	0.0000000	0.0000000
tp_q_terv	3.0028904	0.4506323	tp_u_kans	0.0000000	0.0000000

tr_mtu	2.9499293	1.0068642	tp_alku_a	-3.8598389	0.1313652
tp_palv_gu	2.8616744	0.7754525	tp_a_maat	-3.5006123	0.1166900
tr_ke_tul	2.8489332	0.7507375	tp_n_hall	-3.4482247	0.2638762
te_takk	2.7387969	0.2623825	tp_k_raho	-2.8333876	0.1776358
ko_ammat	2.7155854	0.7439699	tp_s_muup	-2.5110284	0.2148125
tp_p_koul	2.4863074	0.3283566	tp_l_kiin	-1.5760902	0.2092670
ko_al_kork	2.0074856	0.6873452	tp_i_majo	-0.8559685	0.3070611
tp_r_taid	1.8861454	0.4147951	tp_jalo_bf	-0.7589712	0.4618198
tr_pi_tul	1.8173079	0.6170505	tp_g_kaup	-0.5808932	0.4200759
te_teini	1.7722782	0.6672963	tp_f_rake	-0.4179286	0.3963629
te_nuor	1.3555908	0.3734527	tp_m_erik	-0.1217568	0.4569536
te_eil_np	1.3111609	0.4118415			

The values indicate the importance of each attribute against the desired counterpart factor, in this case the sales. The higher the value, the higher the importance of that particular attribute. In the table 6 can be seen two different values, MSE value (%IncMSE) and importance value (IncNodePurit). From the two, the MSE value can be considered more robust and therefore, hold more value. The attributes and values are sorted in descending order to see the most fundamental variables on top, i.e. the higher the MSE value, the better.

Based on the table 6, the highest MSE value can be found from the average floor size (ra_as_kpa, 4,556), which supports the decision tree (figure 13) as well, random tree does not however, disclose in what scope does the average floor area affect the sales. Moreover, when compared to the importance value (IncNodePurit), the same attributes score the highest on the scale. The second highest value can be seen in the higher university degree (ko_yl_kork), which supports the first decision tree (figure 12). The value of 4,304 tells that the variable is important to the output variable, whereas, the other importance value scores only 0.874, which is controversially the third highest.

Furthermore, when comparing to table 4, where the median income of household (tr_mtu) scored lower than expected, the random forest seems to identify the income to have quite high importance. On the other hand, the MSE value does not rank the median income that high, possibly due to the fact that the number of the incidents in figure 12 has not that much significance. It could be argued that this is aligned with the correlation analysis and therefore, income affects indirectly to the sales. According to the Statistic Finland article based on their study, households living in detached houses

are less likely to belong to the lowest income level in Finland. Moreover, as contrary, the areas in the study show that the wealthiest areas where the structure is focused to medium and high-income levels, are located in the geographical areas in this research. (Statistics Finland 2018b). Continuing with the pensioners, which is one of the main attributes in the decision tree in figure 12, and surprisingly it did not result in higher on the MSE scale. This is supported by the importance table of decision trees, table 4, where the pensioners ranked in the middle of the importance table. The pensioner household structure holds values of 3.169 and 0.817. It should be noted is that correlation analysis elaborated that pensioner households are usually placed in the high-income category, which is supported by Statistics Finland's study (2015). The study discloses that a couple of pensioners or individual pensioner have an increasing real income, however, the analysis consisted of all inhabitants in Finland and thus, distorts the results in the capital region. Usually, the capital region is seeing highest and lowest income level in pensioner categories. One of the most important variables in the importance table can be considered the number of households (te_taly). The attribute has value of 4.295 and 0.850, yet, compared to table 4, the variable was constantly on the top of the scale. It can be concluded that the number of households seem to have the high importance over the sales, which could indicate that the network effect is quite strong. The other relevant attributes, which score highly in the variable importance table are related to the household structure. Households with small children (te_plap, 3.847), households with under school age children (te_aklap, 3.057) and the adult households (te_aik, 3.051) are significantly high in the table 6, therefore, it may be said that the household structure and moreover, the possible children are seen to have positive effect on the sales.

Table 7 Random forest classification performance

Accuracy	Error Rate	Recall	Specificity	Precision
0,7740993	0,2259007	0,254717	0,9193173	0,7956778

When assessing the performance of the RF model, the same measures as in table 7 are used. Starting from the accuracy, the RF model holds value of 0.774, which is seen to have slightly better performance compared to decision tree (0,765). Consequently, the most important measure of recall and specificity resulted in the value of 0.254 and 0.919. Compared to the other model, (0.419, 0,769) it can be said that RF model is not

as capable to identify the positives as well as the decision tree, whereas, the RF supersedes in the negative occurrences. The RF model is able to define the positive outputs quite well with value of 0.795 and falsely classify the negatives with value of 0.225. Compared to the decision tree, the overall performance is better, but the recall value is severely lacking behind the decision tree. Moreover, the decision tree is able to identify the negatives (error rate and specificity) with more stable results throughout.

4.4 Combination of the analytical methods

In order to interpret the results more profoundly, results of both decision tree and random forest have to be combined, with notation to the correlation analysis as well. There are similarities in the results in both analysis methods, which means that the results support each other. For instance, the average floor size, pensioner households and high-income category/median income are well represented, but the conclusion cannot be generalised beyond the population in the study area. Moreover, there are differences that have to be taken into consideration, when analysing the implications of the results.

In general, based on both analysis methods, floor area and indirectly the median income level has the highest relevance to the sales. However, the decision tree introduces into the household structure that the area should have both pensioners, and households with children in order to have the highest performance. What should be noted though, is that based on the decision tree, too large floor area might be a hindrance. As a contrary, a large floor area combined with children and pensioner household, has a positive outcome is possible. Additionally, vocational diplomas increase the potential in the given area. Vocational diploma was seen to have correlation with both medium and high-income categories, which can be considered to be the explaining factor to be living in a slightly larger house than the average. As the environment where the business operates, can be considered challenging as number of households and interest in the area are functioning as the network effect, the number of households in the specific area can be seen to have high importance as well. Number of households (te_taly) could be seen to have high correlation with many different attributes but have high importance value in both decision tree and random forest. Therefore, it can be concluded that the more there are households in one area,

the more prominent the area is. As a contrary, area with high income (median income over 102 000€), and therefore, large house (over 113 square meter), may be equally prominent in the eyes of the business, but the network effect might be counter effective in that environment.

To understand the comprehensively the results decision tree and random forest, all of the analytical methods should be considered. However, as the random forest did not disclose scope of the different attributes, it is hard to determine how the attributes actually affect each other. Yet, the random forest results are supporting the decision tree and correlation analyses. The figure 14 depicts the most relevant attributes including the scope and correlation values. On the side note, the random forest results are indirectly shown in the figure as for instance, higher university degree has significant correlation with high income category.

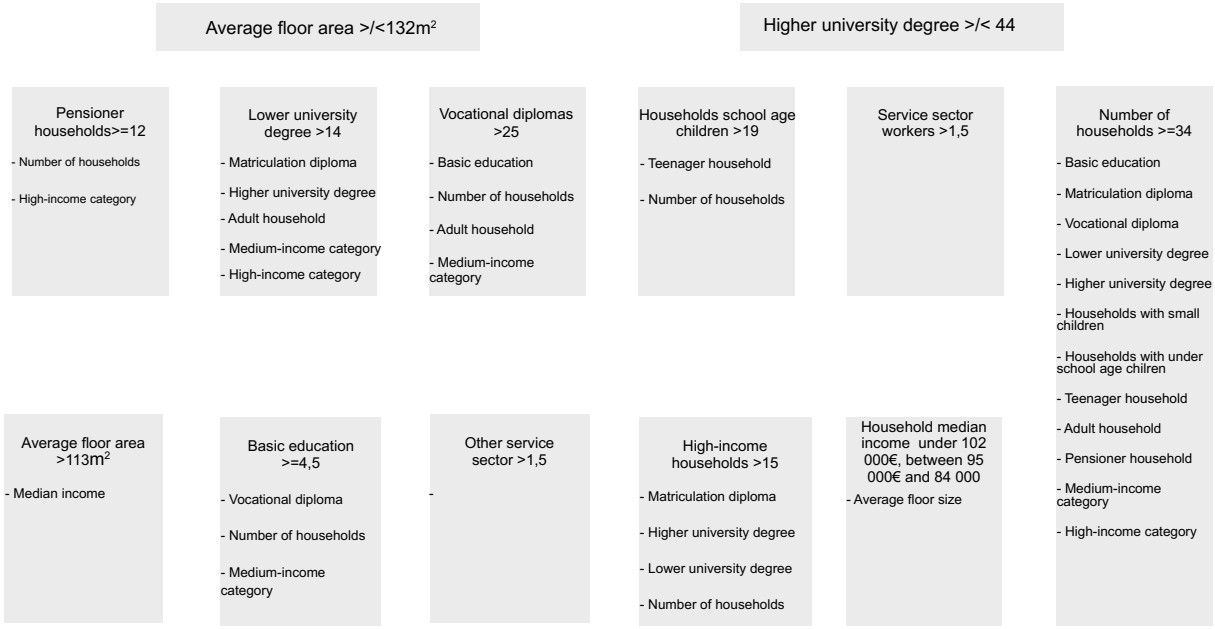

Figure 14 The most relevant attributes and indirect effect of correlation

Based on the analysis and figure 14, it can be said that there is no single ideal cluster of households and the areas consists of numeral different attributes with different scope value and indirect affecting attributes. The figure above is a summary of the relevant attributes seen in figures 12 and 13 of which result in sales. Furthermore, the figure elaborates how the other variables are indirectly affecting the result. Both correlation and random forest results are taken into consideration in the figure 14. The

relevant correlation variables have the threshold value of <0.6, which can be considered a significant correlation between two variables.

When analysing the figure above, it can be said that the heterogeneity of the markets is highlighted. As the main splitter is the large dwelling area in the second decision tree, the household structure may be the defining factor for the smaller home. Pensioners do not require that much space and small and young households (households with under school age children) may not need that much space either. On the other hand, the too high concentration of higher university degree holder (ko_yl_kork, over 44 in a grid) is not seen positive. Consequently, the pensioners were associated with the high-income category, hence, pensioners still have the necessary means to afford living in quite large houses, and therefore, but still there is no cluster of pensioners in the most prominent area. Assessing the results, the negatives are highlighted in the analysis. Too high median income and therefore large house may result in by the fact that in the area there is too few households and controversially, the number of households. The could be concentration of small apartment block, which distorts the results, or the area have smaller than average houses and based on the results, the outcome is usually negative. Consequently, too high intensity of household structure or educational level might not be the ideal one, yet, the vocational diploma seems to have positive output.

5 Discussion

When assessing the research and its impact on different entities, the results and implications can be divided into managerial and theoretical. Furthermore, it is crucial to address the set research questions, which are then divided into different sections later on this chapter. There is no direct definition, which research question is dedicated to each subchapter, but more adaptive approach is taken in terms of usability of the research questions and implications. The managerial implications by its name can be utilised directly in a business environment, whereas, the theoretical implications have suggestive information to the research community.

To begin with, usually, the network effect or externalities are linked to value creation in different settings by increasing the number of users of defined service or product. The more there are users, the more value can be delivered, in terms of intangible value or concrete monetary value. (Na et al. 2013) However, in this research the setting is not as described above, rather the network effect is more radical as the number of potential and interested customers define whether the area is selected for the business actions or not. The relation between the pricing and network size, i.e. network externality was studied, and the resulting factor was that network size itself has positive effect on customer interest towards a product or a company. (Maicas & Sese 2015) Therefore, addressing the first research question (Q1) can be analysed from a few different perspectives.

> Q1: How business analytics can be used effectively in customer acquisition in a network externality influenced environment?

To answer this research question, a broader approach has to be taken where both theoretical and empirical part is utilised. It has been pointed out that business analytics in customer development, moreover, in customer acquisition is not a novel thing as the it has been practiced for decades (Linden, Smith & York 2003). On the other hand, the change in more precise customer acquisition actions have been utilised by introducing geographical and demographical aspects to the analysis (Rhee & McIntyre 2008). Furthermore, the knowledge discovery where the whole process (figure 6) from the database and the data itself is transformed to actionable information should be

considered. This knowledge discovery can be then utilised further in the decision making of an organisation. Mandinach (2012) elaborate that data-driven decision making (DDDM) is knowledge discovery developed further by refining the data and analytics over time to match the changing environment. It should be pointed out that achieving the effective customer acquisition requires initial resources as the knowledge discovery, and precisely the DDDM is resource heavy and require constant improvements.

If not taking the demanded resources into consideration, the analytics and the decision based on analytical results have been seen to have positive outcome compared to intuition or assumption. This is due to the fact that with data and analytics, the business actions can be more thought over and be more precise. For instance, the analytics can be used in the customer relationship management (CRM) and not only in the customer retention and development. The study shows that integrating different types of data enables more focused customer acquisition measures. Usually, the historical data should be sufficient, but more profound analytics demand more complex data types, such as geodemographic and exographic data types. (Asllani & Diane 2011; Bijmolt et al. 2010) Therefore, the marketing activities can be directed to a specific customer segments based on different classification categories or demographic factors and hence, achieve the data-driven marketing environment.

From the empirical approach, it can be stated that there are different analytical methods, which support the customer acquisition greatly. This research used the basic classification and predictive methods, decision tree and random forest. The results suggest that there are certain demographic attributes, which affect the output of the business actions significantly, which can be used in market selection and customer scoring directly. However, focusing on the most obvious attributes might not be sufficient as there are different factors affecting the specific attributes on the background. When assessing the network effect in the given business environment, the market size does not directly correlate with the phenomenon as adaptation and activity in the communities can be considered to be the driving forces. On the other hand, in the theoretical environment, there are numerous different analytical methods, which support the customer acquisition. Depending on complexity elaborated by Delen & Zolbanin (2018) in the figure 8, clustering, decision trees and different neural

networks can be used to classify the customer groups or score the pool of potential customers. Yet, the format of the data can be considered to be a limiting factor to choosing the analytical model. Nevertheless, this research supports the claim that RFM data can be used for more efficient customer acquisition. However, the data might not be sufficient alone, and introducing supportive data types, such as geodemographic dataset truly helps building the data-driven business environment.

The second research question focuses in more detailed manner to the utilisation of the data in customer acquisition.

Q2: What kind of data can be used for customer acquisition and lead scoring?

The data used in customer management can be divided by the stages elaborated by Greene & Milne (2005) and the different application studied by Bijmolt et al. (2010). The most basic data, which can be used in customer acquisition is called RFM data (Recency, Frequency and Monetary). The data is collected from the historical behaviour of a customer, and in the developing a data-driven decision making, the RFM data is usually sufficient. The historical data can be used for the initial assessment of the customer. The research suggests that customer lifetime value (CLV) and CRM cannot exist without the other as the initial calculations of the potential of a customer is done based on the RFM and CLV. However, the RFM might be a limiting factor in the analytics and more focused customer acquisition, which can be solved in integrating different data types. As Sorber et al. (2015) argue that having more than one data source increases the information and knowledge gained from the analytics. Brink & Rensburg (2017) introduce a concept of geodemographic data, which encompasses both geographic and demographic aspect to a dataset. This enables decision made by customer location or specific demographic variable. Moreover, the complexity and understanding of the customer can be elaborated with yet comprehensive data called exographics. Exographics take into consideration the surrounding of an entity in question, such as the culture or the region.

From the business development perspective, the data can focus to customer business development (CBD). The concept examines how value can be delivered to customer as in customer-oriented manner. (Hunter 2014) Therefore, RFM and geodemographic data holds the most informative value as the customers can be clustered and classified

by the demographic variables, location and the customer behaviour. The method is called geosegmentation, which enable more data-driven marketing methods and hence, improve the efficiency of customer acquisition. On the other hand, alongside the more focused customer segmentation, there should be scoring on the potential customer pool. RFM and demographic data can be used as a base to determine the ideal customer and focus on the similar kinds of customers to maximise the business efforts.

When reflecting the third research question, it can be concluded that the theoretical part can be used to deliver a comprehensive picture for the Q3.

> Q3: What kind of analytical models and analytics can be used in customer acquisition and business development?

When assessing the analytical models for customer acquisition, the Delen & Zolbanin (2017) classification of different types of analytics can be used. The types can be divided by the complexity, descriptive or predictive models or the purpose of the analytics. The most basic analytical model, descriptive methods, classify and cluster the studied matter in predefined or autonomic way. This enables organisations to understand their current customers even more profoundly, but also develop their customer acquisition model even further. What should be noted, is that classification or the cluster analysis requires slightly complex data, such as demographic data in order to customers to be categories effectively. Bradlow et al. (2017) argue that the demographic analysis can be used as a base for customer loyalty programs, which in turn personalise the content and provide market intelligence to the organisation. Consequently, more advanced methods will take the classification further, but factor limiting the classification is the data, and precisely the accuracy of the data points. When increasing the complexity of the data, the preciseness of the data should be addressed (Bramen 2017, 89-90).

Predictive models on the other hand, enable to analyse the customer in more detailed way. As it has been pointed out, the predictive analytics are more demanding on the resources, but provide indicative results. In general, predictive methods are usually linked to minising the customer churn or identifying the best fit of a customer in a banking industry. The latter one is exemplary case of lead scoring on which D'Haen &

Van del Poel (2013) elaborated in their study. The study included a classical model of sales funnel, which included different measures to rank the customers based on their potential and profitability and the results suggested that neural network can be used to rank the potential customers. Furthermore, in this research the results indicated that in area selection, selecting clusters of customers a classification and regression (CART) model can be used.

The customer development side of the business analytics can be continued from the different analytical models. As it has been pointed out by Forsman (2008), the BD can be done incrementally, i.e. developing the business as a by-product of business actions, or more comprehensive business development project. Despite the nature of the BD, as the definition of the BD by Ahtenhagen et al. (2017) is delivering value to different entities related to business actions, the organisation or the customers on a long time-period. There is no single approach how the business analytics can be used in BD, but the purpose of the data and analytics should be fit to the organisations' needs. The analytic methods could be considered a stand-alone improvement to business processes but embracing the data-driven environment require constant actions on the data and analytics front.

Taking into consideration all the other research questions, Q4 is focused more on the implications and applications of the analytical methods.

> Q4: How can the analytical models and analytics be used for customer acquisition and business development?

In order to answer the research question above, this research can be considered an example how the analytics can be implemented in order to develop an organisation to more data-driven environment and therefore, acquire customers more effectively. Assessing the data to knowledge flow or the knowledge discovery (figure 6), starting from the data and its management is crucial. As it has been pointed out, having more than one source of data increases the value and complexity of the information the data holds. Depending on the business situation, the RFM data should be sufficient, but this study shows that having the RFM and geodemographic data enables more precise and complex results. Furthermore, the use of the analytics dictates the analytical

models and methods, but as a contrary, some analytical models can be used for customer management, for instance the neural networks.

This research suggests that the analytical models suitable for this business environment and organisation novel to analytics can be the basic ones. In this study two different classification and regression models were used to either contradict each other or support the results. Based on the results, the historical data was in a minor role, as the demographic attributes were more informative. The data mining resulted in certain attributes affecting the sales. In this setting, the average floor area, higher education, pensioners, vocational diploma and median income had the highest effect on the desired output. This gives great deal of information to act on. From the customer acquisition perspective, results suggest a model for area selection and finding the ideal cluster of potential customers. In the area selection, all the relevant attributes should be considered, but also take into account the indirect attributes affecting the most relevant attributes.

The business development on the other hand, requires different steps when it comes to the analytical models. One of the most crucial concepts is the DDDM (data-driven decision making), which is the desired state in most of the organisations. In this research it was revealed that the BD actions necessary may not be that great, but it still would require resources to acquire the necessary analytical capabilities. In this study, a statistical and geographical software was used in order to analyse the data correctly. Further, as the DDDM (figure 5) dictates, the analytical methods should be used constantly in an organisation. As the historical and sales data is constantly generated, the results of the analytics would be elaborated more. The constant stream of results can improve the analytical model to be more effective and hence, develop the business further and make the customer acquisition efficient.

5.1 Reliability and validity

Reliability and validity are used to determine the overall quality of the research. Moreover, whether the research in question is reliable or in other terms replicable and validity is used to define if the study, its scope and measurements have been specified to the intention of the study. The validity means the methods and concepts are to be

aligned with the research area. There are different approached related each quality measure, of which few are selected to this chapter. (Andres 2012, 30-33; 122-123)

Reliability can be divided into internal and external, both of which measure slightly different aspects. The internal reliability measures whether the values in the results are consistent when validating the research and external is related to replicability of the whole research (Seale 2013, 140-141). In this research, the internal validity is subject entirely to the analytical methods and the data management. The internal reliability might differ from individual to another individual, for instance in integration of the two datasets into one. However, the attributes selected to the analysis could be considered to remain constant as there are certain attributes, which can be linked to the households. From the analytical model perspective, the depending factor, which might differ is the controls for the classification and predictive models. Moreover, if the dataset is sub-set, the algorithm randomly separates the instances, which would result in differing output. External replicability may be affected by the availability of the data, however, for the arguments sake, if both sets of data would be publicly available, the results could be retested over and over again. In conclusion, the reliability of the study is quite straightforward as in quantitative studies in general. Only the data management and controlling the analytical models could be considered as the differing factors if this research would be replicated.

The validity measure addresses few different aspects: whether the research has been producing valid information to answer the research question, the data sample has been elaborated and the research can be applied to other setting as well. On the other hand, it can also be evaluated from different perspectives, such as criteria, content and construct. (Andres 2012, 115-116) Consequently, Seale (2013, 38-42) adds internal and external validities, which both address slightly different matters. Internal validity is regard whether there is another factor intervening the research and therefore, cannot be applied elsewhere. External validity means whether the result proposition remain the same when applied to different settings. When assessing the validity of this research, it can be said that the data and the analytical methods are answering the opposed research questions, the data is elaborated, and the construct of this research can be applied to other field if the sales data would be replaced. Regarding the internal validity, there were no external factors affecting the results, however, if applied

elsewhere, the software and the methods could be varying nominator. The external validity can be considered as sound as is, this research can be fully applied to other environments compared to this.

5.2 Theoretical implications

The theoretical contribution relies on the network externality, as geodemographic marketing, market selection and customer classification are quite researched phenomena. There have been similar studies, such as Saarenpää et al. 2013 and Lassila et al. 2011, where the study focused on the adaptation of less static element compared to the product in this research. The findings on the 2013 article was supported by this research that certain attributes emerge more than others. However, the attributes were somewhat different, and the attributes were related more to the monetary side, such as house size and median income. Consequently, neither of the studies touched the network externalities and how does the concept affect the adaptation. This research combined both geodemographic analytics and network externality in terms of market entry rather than partial value delivery. The theoretical contribution of this research lies in the combination of the both and how geodemographic factors in the network effect environment affect the adaptation of the costly infrastructural element.

The alternative approach of theoretical contribution of this study, is the complementing the theoretical framework in terms of the use of analytical methods in customer acquisition. The study suggests that less complex models can be used for customer acquisition, in this research, the classification and regression trees. Furthermore, this study suggests that in certain environments, the RFM data may not be sufficient and must be complemented by other data types, for example, by geodemographic data. In general, the results can be applied to competing companies and supplementing industries.

5.3 Managerial implications

This research resulted in more managerial input than theoretical contribution. Still this research holds some theoretical implications as well. Starting from the base setting

and the purpose of this research – finding whether there are any demographic factors affecting the purchase decision. The data and analytics were introduced to find the connection between the geodemographic attributes, and based on the different analytical methods, there is link between different attributes to the sales. However, the results did not support the basic induction considering the product in question, more money the households have the better the sales performance. On the contrary, the results turned out that there are certain factors, which affect the output of the sales activities, which should be taken into account in different business operations. Moreover, as there is network effect, of sort, affecting the business environment, the results and the applicability have to be thoroughly assessed. Based on the business model, the network effect does effect more on the market selection, hence, the network size does not affect that much or directly to the sales. Nonetheless, it is crucial to understand that the network does affect the other attributes and therefore also affect the sales indirectly.

There were several different attributes, which have quite significant correlation value, and therefore, the actions should be aligned with the results. This meaning that market selection and lead scoring should be done by taking the most relevant attributes and the indirect forces in play into consideration when deciding which markets to enter. Additionally, before applying the analytical model and continuous knowledge discovery process, the actual sales performance scope should be determined and scale for the lead scoring should be set. That way, the actual performance of the analytics and markets' potential could be validated. Based on this, the market selection could be changed from network size and income level to high concentration and number of most prominent clusters of households (see figure 13) and testing the interest in the area. Consequently, the business development could be steered more towards the customer management and change in the marketing strategy. As the analytics enable more accurate customer analytics, the marketing strategy can be more targeting, even within the same market.

The activities and business orientation are different things when it comes to adaptation to new means and methods. The above-mentioned implications can be done on a short notice without too much planning behind. This could be considered changing the business on the go and based on the up-to-date performance. However, the business

orientation and culture change would be more comprehensive business development, which would be needing more resources and possibly a more controlled process. Depending on the business, the data-orientation could be well established, but others might not have analytics embedded into the actions at all. Therefore, in the first approach, there would not be much need to change, merely embracing the data-driven decision making would be sufficient and focus on the quality of the data, whereas the latter case would be needing more thorough organisational change. What should be noted though, is that like Mandinach (2012) argued that in the initial phase, the results may not be what is expected as the performance improves over time due to the accuracy and the volume of the data. Still, the change would require continuous improvements over a long period of time and necessary actions to gather the sufficient data. Furthermore, as this research was done based on the sales and geodemographic data from the capital region, the expanding markets might cause some issues. This is due to the population structure and other demographic attributes compared to other areas, such as income and education, are not identical. This creates a need to gather more data as the results may not apply entirely, hence the continuous incremental development over time.

When assessing the complexity of the pre-requirements demanded for the DDDM, such as data collection, software for analysing the data and lastly, necessary competence to interpret the results is required. The Forsman's (2008) business development process flow (figure 9) could be applied in this case as there are certain forces, managerial and theoretical, which steer towards the more analytic business environment. Based on the Forsman's BDP model, there should be actual need and the capacity to initiate a business development project. However, there are certain pre-requisite that are demanded before the actual analysis can be performed. The data collection and management should be in order, and the necessary software should be in use to use the data. In addition, the actual data mining is part of the knowledge discovery project as interpreting the results could be seen as more demanding task. If the DDDM would be implemented as a project-like manner, there should be performance indicators. In managerial perspective, this means that use of the analytical model and therefore, the success should be benchmarked, which would require sales performance indicators as well.

6 Conclusions

This research included different analytical methods to study how the analytical methods can improve customer acquisition and how the business can be developed into more data-driven environment. The business environment was the Finnish society where the data was collected from the capital region and later integrated to geo-demographic data from Statistics Finland. The theoretical framework provided insights to the concept of data, how the data is managed and how the data can actually be used in marketing. Furthermore, data is the main component of business analytics and in the theoretical framework, different methods of business analytics and data mining, were introduced, which were further used in this research. From the business perspective, the data and analytics are used in decision-making in order to make more sound results. However, the businesses may be in different situations and there are times when the data-driven environment require large projects to be implemented.

As the quantity of the data is constantly growing and utilisation of the data is expanding, the businesses are finding more applications to data and improving the current ones. Nonetheless, embracing the data and analytics require plenty of resources, moreover, it requires the open and agile business orientation. The concept and field of business analytics have been vastly studied, and the methods have significant impact on the business. The results revealed that there is significant importance to understand the business environment where the business operates, moreover, take into consideration the different variables seen in the analyses. The analytical results are not that straightforward as there are other contributing factors affecting the business operations, such as network externality. Yet, the network externality does not have direct correlation nor impact on the sales, the concept is still present when examining the results. What can be concluded, is that the analytical methods, more precisely, the methods in this research are applicable to environments where the customer acquirement is larger business than the rest of the customer management. However, it can be argued that the full knowledge discovery is more demanding as there are intangible and tangible resources required to acquire for the data-driven decision-making. Nonetheless, implementing the DDDM requires much consideration, therefore, the quick improvement in the business activities and step towards the data embracing culture is possible.

6.1 Limitations and suggestions for future research

As in all of the research, this study has also limitations. Limiting factors can be divided to two – the scope of the study in terms of the data and the actual business environment. The research consisted of data only from the capital region of Finland due to availability of the sales data, therefore, the results from the analytics cannot be directly applied to other regions. What should be noted is that some demographical attributes remain more or less the same, such as size of the household and the average floor area, but regional differences, such as income, education and work-related factors should be considered. Consequently, the quality of the sales activities can be considered as quite significant limitation. This is due to the fact that the sales persons might not perform as expected and online marketing activities may not reach sufficient number of potential customers.

The natural continuity to this research and suggestion for the future is to implement the data-driven decision making by introducing analytics to the operations. This way the results could be validated in the natural environment and performance tested. The suggestive research would also consider how different areas affect the output of the sales activities, for instance, areas where the urbanisation is not such a strong force. In addition, the success of the comparing areas could be one subject to study alongside the quality of the sales activities. Moreover, the actual pre-sales activities could be studied in order to find what kind of sales activities work the best. As it was pointed in out that the crucial data of non-structured and exographics form holds even more intricate value compared to RFM and geodemographic, it could be extracted in order to support the study regarding the pre-sales activities.

List of references

Achtenhagen, L., Ekberg, S. & Melander, A. (2017). Fostering Growth through Business Development: Core Activities and Challenges for Micro-firm Entrepreneurs. *Journal of Management and Organization,* 23(2), pp. 167-185.

Ahn, H., Ahn, J.J., Byun, H.W. & Oh, K.J. (2011). A novel customer scoring model to encourage the use of mobile value-added services. *Expert Systems with Applications,* 38(9), pp. 11693-11700.

Anantha M. Prasad, Iverson, R. & Liaw A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems,* 9(2), pp. 181-199.

Andres, L. (2012). *Designing & doing survey research.* 1. publ. edn. Los Angeles [u.a.]: SAGE.

Arnott, D., Lizama, F., & Song, Y. (2017). Patterns of business intelligence systems use in organizations. *Decision Support Systems,* 97, pp. 58.

Asllani, A. & Halstead, D. (2011). Using RFM data to optimize direct marketing campaigns: a linear programming approach. *Academy of Marketing Studies Journal,* 15(S2), pp. 59.

Aven, T. (2013). A conceptual framework for linking risk and the elements of the data–information–knowledge–wisdom (DIKW) hierarchy. *Reliability Engineering and System Safety,* 111, pp. 30-36.

Balducci, B. and Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science,* 46(4), pp. 557-590.

Batini, C. & Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques.* Cham: Springer.

Bijmolt, T.H.A., Leeflang, P.S.H., Block, F., Eisenbeiss, M., Hardie, B.G.S., Lemmens, A. & Saffert, P. (2010). Analytics for Customer Engagement. *Journal of Service Research,* 13(3), pp. 341-356.

Binder, S., Macfarlane, G.S., Garrow, L.A. & Bierlaire, M. (2014). Associations among Household Characteristics, Vehicle Characteristics and Emissions Failures: An Application of Targeted Marketing Data. *Transportation Research: Part A: Policy and Practice,* 59, pp. 122-133.

Bini, S.A. (2018). Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *The Journal of Arthroplasty,* 33(8), pp. 2358-2361.

Bosancic, B. (2016). Information in the knowledge acquisition process. *Journal of Documentation,* 72(5), pp. 930-960.

Bradlow, E.T., Gangwar, M., Kopalle, P. & Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. *Journal of Retailing,* 93(1), pp. 79-95.

Bramer, M. (2016). *Principles of Data Mining.* 3rd ed. 2016 edn. London: Springer London, Limited.

Brink, M.P & Van Rensburg, A. (2017). An Approach to Improving Marketing Campaign Effectiveness and Customer Experience Using Geospatial Analytics. *South African Journal of Industrial Engineering,* 28(2), pp. 95.

Brynjolfsson, E. & McElheran, K. (2016). The rapid adoption of data-driven decision-making. *The American economic review,* 106(5), pp. 133-139.

Carbonero-Ruz, M., Martínez-Estudillo, F.J., Martínez-Estudillo, A.C., Fernández-Navarro, F. & Becerra-Alonso, D. (2017). A two dimensional accuracy-based measure for classification performance. *Information Sciences,* 382-383, pp. 60-80.

Chen, H., Chiang, R.H.L. T Storey, V.C., (2012). Business intelligence and analytics. *Management information systems,* 36(4), pp. 1165-1188.

Columbus, L. (2018). 10 Charts that will change your perspective of big data's growth. Forbes. [www-document]. [Accessed 10 Sep 2018]. Available https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#30f4b2062926

Corrigan, H.B., Craciun, G. & Powell, A.M. (2014). How Does Target Know So Much About Its Customers? Utilizing Customer Analytics to Make Marketing Decisions. *Marketing Education Review,* 24(2), pp. 159-166.

Côrte-Real, N., Oliveira, T. & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. *Journal of Business Research,* 70, pp. 379-390.

Davidson, F. (1996). Principles of Statistical Data Handling. *Principles of Statistical Data Handling.* Thousand Oaks: SAGE Publications, Inc.

Davino, C. & Fabbris, L. (2013). *Survey Data Collection and Integration.* 1. Aufl. edn. DE: Springer-Verlag.

De Mauro, A., Greco, M. & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review,* 65(3), pp. 122-135.

Delen, D., Kuzey, C. & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems With Applications,* 40(10), pp. 3970-3983.

Delen, D. & Zolbanin, H.M. (2018). The analytics paradigm in business research. *Journal of Business Research,* 90, pp. 186-195.

Dutta, D. & Bose, I. (2015). Managing a Big Data project: The case of Ramco Cements Limited. *International Journal of Production Economics,* 165, pp. 293-306.

El-Yaniv, R., Geifman, Y. & Wiener, Y. (2017). The Prediction Advantage: A Universally Meaningful Performance Measure for Classification and Regression.

Erevelles, S., Fukawa, N. & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research,* 69(2), pp. 897-904.

Fan, B. & Zhang, P. (2009). Spatially enabled customer segmentation using a data classification method with uncertain predicates. *Decision Support Systems,* 47(4), pp. 343.

Farquhar, J.D. (2012). *Case study research for business.* 1. publ. edn. Los Angeles, Calif. [u.a.]: Sage.

Ferguson-Boucher, R. (2013). Location Analytics: Bringing Geography Back. *MIT Sloan Management Review,* 54(2), pp. 1.

Forsman, H. (2008). Business development success in SMEs: a case study approach. *Journal of Small Business and Enterprise Development,* 15(3), pp. 606-622.

García, S., Luengo, J. & Herrera, F. (2015). Data preprocessing in data mining

García, S., Luengo, J. & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems,* 98, pp. 1-29.

Grbovic, M., Djuric, N., Guo, S. & Vucetic, S. (2013). Supervised clustering of label ranking data using label preference information. *Machine Learning,* 93(2), pp. 191-225.

Greene, H., & Milne, G. R. (2005). Alternative data sources in targeted marketing: The value of exographics. Journal of Targeting, Measurement and Analysis for Marketing, 14(1), 33-46.

Guenzi, P. & Troilo, G. (2007). The joint contribution of marketing and sales to the creation of superior customer value. *Journal of Business Research,* 60(2), pp. 98-107.

Habeeb, R.A., Nasaruddin, F., Gani, A., Targio Hashem, I.A., Ahmed, E. & Imran, M., (2018). Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management.*

Holsapple, C., Lee-Post, A. & Pakath, R. (2014). A unified foundation for business analytics. *Decision Support Systems,* 64, pp. 130.

Holsapple, C.W., Hsiao, S. & Pakath, R. (2018). Business social media analytics: Characterization and conceptual framework. *Decision Support Systems,* 110, pp. 32-45.

Hoy, W.K. (2010). *Quantitative research in education.* Los Angeles [u.a.]: SAGE.

Huang, D., Lai, J. & Wang, C. (2015.) Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis. *Neurocomputing,* 170, pp. 240-250.

Hunter, G.K. (2014). Customer business development: identifying and responding to buyer-implied information preferences. *Industrial Marketing Management,* 43(7), pp. 1204-1215.

Jennex, M. (2017). Big Data, the Internet of Things, and the Revised Knowledge Pyramid. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems,* 48(4), pp. 69-79.

Kim, E. & Lee, B. (2007). An economic analysis of customer selection and leveraging strategies in a market where network externalities exist. *Decision Support Systems,* 44(1), pp. 124-134.

Krishnamoorthi, S. & Mathew, S.K. (2018). Business analytics and business value: A comparative case study. *Information & Management,* 55(5), pp. 643-666.

Kumar, V., Chattaraman, V., Neghina, C., Skiera, B., Aksoy, L., Buoye, A. & Henseler, J. (2013). Data-driven services marketing in a connected world. *Journal of Service Management,* **24**(3), pp. 330-352.

Lassila, J., Haakana, J., Partanen, J., Koivuranta, K. & Peltonen, S. (2011). Network Effects Of Electric Vehicles - Case From Nordic Country, *International Conference on Electricity Distribution* 2011.

Lavalle, S., Lesser, E., Shockley, R., Kruschwitz, N. & Hopkins, M.S. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review,* 52(2), pp. 21.

Lee, K., Lee, I. & Torpelund-Bruin, C. (2012). Map segmentation for geospatial data mining through generalized higher-order Voronoi diagrams with sequential scan algorithms. *Expert Systems with Applications,* 39(12), pp. 11135-11148.

Li, Y., Wang, X., Huang, L. & Bai, X. (2013). How does entrepreneurs' social capital hinder new business development? A relational embeddedness perspective. *Journal of Business Research,* 66(12), pp. 2418-2424.

Linden, G., Smith, B. & York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing,* 7(1), pp. 76-80.

Linder, R., Geier, J. & Kölliker, M. (2004). Artificial neural networks, classification trees and regression: Which method for which customer base? *Journal of Database Marketing & Customer Strategy Management,* 11(4), pp. 344.

Long, Q. (2018). Data-driven decision making for supply chain networks with agent-based computational experiment. *Knowledge-Based Systems,* 141, pp. 55-66.

Lovatti, B.P.O., Nascimento, M.H.C., Neto, ÁC., Castro, E.V.R. & Filgueiras, P.R. (2019). Use of Random forest in the identification of important variables. *Microchemical Journal,* 145, pp. 1129-1134.

Lukoianova, T. & Rubin, V.L. (2014). Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances in Classification Research Online,* 24(1), pp. 4.

Maicas, J.P. & Sese, F.J., 2015. Customer-Base Management in Network Industries: The Moderating Role of Network Size and Market Growth. European Management Review, 12(4), pp. 209-220.

Marr, B. (2018). How much data do we create every day, the mind-blowing stats everyone should read. Forbes. [www-document]. [Accessed 20 Sep 2018]. Available https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#14a3e43560ba

Maimon, O.Z. (2010). *Data mining and knowledge discovery handbook.* 2. ed. New York: Springer.

Mandinach, E.B. (2012). A Perfect Time for Data Use: Using Data-Driven Decision Making to Inform Practice. *Educational Psychologist,* 47(2), pp. 71-85.

Muijs, D. (2011). *Doing quantitative research in education with SPSS.* 2. ed. edn. Los Angeles [u.a.]: SAGE.

Na, K., Han, C. & Yoon, C. (2013). Network effect of transportation infrastructure: a dynamic panel evidence. *The Annals of Regional Science,* 50(1), pp. 265-274.

Novaković, J., Veljović, A., Ilić, S., Papić, Z. & Tomović, M. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science,* 7(1), pp. 39.

Oyegoke, A. (2011). The constructive research approach in project management research. *International Journal of Managing Projects in Business,* 4(4), pp. 573-595.

Paiola, M., Gebauer, H. & Edvardsson, B. (2012). Service Business Development in Small- to Medium-Sized Equipment Manufacturers. *Journal of Business-to-Business Marketing,* 19(1), pp. 33-66.

Peralta, B., Caro, A. & Soto, A. (2016). A proposal for supervised clustering with Dirichlet Process using labels. *Pattern Recognition Letters,* 80, pp. 52-57.

Praseeda, C.K. & Shivakumar, B.L. (2014). A Review of Trends and Technologies in Business Analytics. *International Journal of Advanced Research in Computer Science,* 5(8),

QGIS. (2019). A Free and Open Source Geographic Information System. [www-document]. [Accessed 12 Dec 2018]. Available https://qgis.org/en/site/

Rahm, E. & Do, H.H. (2000). Data Cleaning: Problems and Current Approaches .*IEEE Computer Society,* 23(4), pp. 1-42.

Ramadani, V., Zendeli, D., Gerguri-Rashiti, S. & Dana, L. (2018). Impact of geomarketing and location determinants on business development and decision making. *Competitiveness Review: An International Business Journal,* 28(1), pp. 98-120.

Ramanathan, R., Philpott, E., Duan, Y. & Cao, G. (2017). Adoption of business analytics and impact on performance: a qualitative study in retail. *Production Planning & Control,* 28(11-12), pp. 985-998.

Randall, S.M., Ferrante, A.M., Boyd, J.H. & Semmens, J.B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making,* 13(1), pp. 64.

Reinsel, D., Gantz, J. & Rydning, J. (2018) Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big. [www-document]. [Accessed 12 Nov 2018]. Available https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

R. The R Project for Statistical Computing. [www-document]. [Accessed 12 Dec 2018]. Available https://www.r-project.org

Rhee, S. & Mcintyre, S. (2008). Including the effects of prior and recent contact effort in a customer scoring model for database marketing. *Journal of the Academy of Marketing Science,* 36(4), pp. 538-551.

Ruiz-Samblás, C., Cadenas, J., Pelta, D. & Cuadros-Rodríguez, L. (2014). Application of data mining methods for classification and prediction of olive oil blends with other vegetable oils. *Analytical and Bioanalytical Chemistry,* 406(11), pp. 2591-2601.

Russell, J. & Macgill, S. (2015). Demographics, policy, and foster care rates; A Predictive Analytics Approach. *Children and Youth Services Review,* 58, pp. 118-126.

Saarenpää, J., Kolehmainen, M. & Niska, H. (2013). Geodemographic analysis and estimation of early plug-in hybrid electric vehicle adoption. *Applied Energy,* **107**, pp. 456-464.

Sagi, T. & Gal, A. (2018). Non-binary evaluation measures for big data integration. *The VLDB Journal,* 27(1), pp. 105-126.

Saunders, M.N.K., Lewis, P. & Thornhill, A. (2015). *Research Methods for Business Students.* Seventh edition. Harlow, United Kingdom: Pearson Education M.U.A.

Seale, C. (2013). The quality of qualitative research. Sage Publications.

Seddon, P.B., Constantinidis, D., Tamm, T. & Dod, H. (2017). How does business analytics contribute to business value? *Information Systems Journal,* 27(3), pp. 237-269.

Simons, H., 2009. *Case study research in practice.* 1. publ. edn. Los Angeles [u.a.]: SAGE.

Smirnova, E., Ivanescu, A., Bai, J. & Crainiceanu, C.M. (2018). A practical guide to big data. *Statistics and Probability Letters,* 136, pp. 25-29.

Sorber, L., Van Barel, M. & De Lathauwer, L. (2015). Structured Data Fusion. *IEEE Journal of Selected Topics in Signal Processing,* 9(4), pp. 586-600.

Squire, M. (2015). *Clean Data.* 1 edn. GB: Packt Publishing.

Souba, W.W., Haluck, C.A. & Menezes, M.A.J. (2001). Marketing strategy: An essential component of business development for academic health centers. *The American Journal of Surgery,* 181(2), pp. 105-114.

Statistics Finland. (2015). Tulonjakotilasto. [www-document]. [Accessed 30 Dec 2018]. Available http://www.stat.fi/til/tjt/2015/02/tjt_2015_02_2017-03-24_kat_005_fi.html

Statistics Finland. (2018a) Grid Database – Data content. [www-document]. [Accessed 04 Dec 2018]. Available. https://www.stat.fi/tup/ruututietokanta/tietosisalto_en.html

Statistics Finland. (2018b). OMAKOTITALOISSA ASUVIEN PIENITULOISUUS HARVINAISTA PÄÄKAUPUNKISEUDULLA. [www-document]. [Accessed 30 Dec 2018]. Available http://www.stat.fi/tietotrendit/artikkelit/2018/omakotitaloissa-asuvien-pienituloisuus-harvinaista-paakaupunkiseudulla/

Statistics Finland (2018c). Asuntokanta 2017. [www-document]. [Accessed 22 Dec 2018]. Available https://www.stat.fi/til/asas/2017/01/asas_2017_01_2018-10-10_kat_001_fi.html

Thomas, R.M. (2003). *Blending qualitative & quantitative research methods in theses and dissertations.* Thousand Oaks, Calif: Corwin Press.

Tillmanns, S., Ter Hofstede, F., Krafft, M., & Goetz, O. (2017). How to separate the wheat from the chaff: Improved variable selection for new customer acquisition. Journal of Marketing, 81(2), 99-113.

Tiwari, T., Tiwari, T. & Tiwari, S. (2018). How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different? *International Journal of Advanced Research in Computer Science and Software Engineering,* 8(2), pp. 1.

Torrecilla, J.L. & Romo, J. (2018). Data learning from big data. *Statistics and Probability Letters,* 136, pp. 15-19.

Valentine, E.L. (2003). Business development: A barometer of future success. *Journal of Commercial Biotechnology,* 10(2), pp. 123-130.

Wedel, M. & Kannan, P.K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing,* 80(6), pp. 97-121.

Wu, C., Chen, Y., Liu, Y. & Yang, X. (2016). Decision tree induction with a constrained number of leaf nodes. *Applied Intelligence,* 45(3), pp. 673-685.

Yu, L., Wang, S & Lai, K.K. (2006). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering,* 18(2), pp. 217-230.

Zeng, J. & Glaister, K.W. (2018). Value creation from big data: Looking inside the black box. *Strategic Organization,* 16(2), pp. 105-140.

Zhang, C. (2012). *Ensemble machine learning.* 2012 edn. New York [u.a.]: Springer.

Zhang, X., Simpson, T., Frecker, M. & Lesieutre, G. (2012). Supporting knowledge exploration and discovery in multi-dimensional data with interactive multiscale visualisation. *Journal of Engineering Design,* 23(1), pp. 23-47.

Appendices

Appendix I Selected attributes

Educational structure 2016	Size and stage in life of households 2016	Households' Disposable Monetary Income 2015	Buildings and housing 2016	Workplace structure 2015
Basic level studies ko_perus	Households, total	Median income of households tr_mtu	Average floor area ra_as_kpa (m2)	Primary production tp_alku_a
		Households belonging to the lowest income category tr_pi_tul (>=16466€ PA)		
Matriculation examination ko_ylio	Average size of households			Processing tp_jalo_bf
		Households belonging to the middle income category tr_ke_tul (16467-34087€ PA)		
Vocational diploma ko_ammat	Young single persons te_nuor (>35yrs)			Services tp_palv_gu
	Young couples without children te_eil_np	Households belonging to the highest income category tr_hy_tul (<34087€ PA)		
Lower academic degree ko_al_kd	(>35y)			A Agriculture, forestry and fishing tp_a_maat
Higher academic degree ko_yl_kd	(>3yrs)			B Mining and quarrying tp_b_min
	Households with children under school age te_aklap (>7yrs)			C Manufacturing tp_c_teol
	(7-12yrs			D Electricity, gas, steam and air conditioning supply tp_d_ener
	Households with teenagers te_teini (13-17yrs)			E Water supply; sewerage, waste management and remediation activities tp_e_vesi
	Adult households te_aik (18-64yrs)			F Construction tp_f_rake
	Pensioner households te_elak (>64yrs)			G Wholesale and retail trade; repair of motor vehicles and motorcycles tp_g_kaup
				H Transportation and storage tp_h_kuli
				I Accommodation and food service activities tp_i_majo
				J Information and communication tp_j_info
				K Financial and insurance activities tp_k_raho
				L Real estate activities tp_l_kiin
				M Professional, scientific and technical activities tp_m_erik
				N Administrative and support service activities tp_n_hall
				O Public administration and defence; compulsory social security tp_o_juik
				P Education tp_p_koul
				Q Human health and social work activities tp_q_terv
				R Arts, entertainment and recreation tp_r_taid
				S Other service activities tp_s_muup
				T Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use tp_t_koti

Appendix II Summary of the data

```
    ko_perus        ko_yliop       ko_ammat       ko_al_kork     ko_yl_kork       te_taly         te_eil_np        te_plap          te_aklap         te_klap          te_teini         te_aik      
 Min.   : 1.00   Min.   : 0.000   Min.   : 2.00   Min.   : 0.00   Min.   : 1.00   Min.   :  6.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 2.00  
 1st Qu.: 8.00   1st Qu.: 5.000   1st Qu.:18.00   1st Qu.: 8.00   1st Qu.:15.00   1st Qu.: 30.00   1st Qu.: 0.000   1st Qu.: 2.000   1st Qu.: 4.000   1st Qu.: 6.000   1st Qu.: 4.000   1st Qu.:10.00  
 Median :12.00   Median : 8.000   Median :25.00   Median :12.00   Median :25.00   Median : 42.00   Median : 0.000   Median : 3.000   Median : 7.000   Median : 8.000   Median : 7.000   Median :13.00  
 Mean   :13.16   Mean   : 9.319   Mean   :26.34   Mean   :12.99   Mean   :26.21   Mean   : 42.69   Mean   : 0.886   Mean   : 3.481   Mean   : 7.546   Mean   : 8.518   Mean   : 7.389   Mean   :15.08  
 3rd Qu.:17.00   3rd Qu.:12.000   3rd Qu.:35.00   3rd Qu.:17.00   3rd Qu.:35.00   3rd Qu.: 53.00   3rd Qu.: 1.000   3rd Qu.: 5.000   3rd Qu.:10.000   3rd Qu.:11.000   3rd Qu.:10.000   3rd Qu.:20.00  
 Max.   :47.00   Max.   :35.000   Max.   :72.00   Max.   :34.00   Max.   :81.00   Max.   : 91.00   Max.   :11.000   Max.   :12.000   Max.   :21.000   Max.   :25.000   Max.   :23.000   Max.   :55.00  
    te_elak          tr_mtu         tr_pi_tul       tr_ke_tul       tr_hy_tul        ra_as_kpa        tp_alku_a        tp_jalo_bf       tp_palv_gu       tp_a_maat        tp_b_koiv        tp_c_teol     
 Min.   : 1.00   Min.   : 37571   Min.   : 0.000   Min.   : 1.00   Min.   : 2.00   Min.   :  68.0   Min.   :0.00000   Min.   : 0.000   Min.   :  0.00   Min.   :0.00000   Min.   : 0   Min.   : 0.0000  
 1st Qu.: 6.00   1st Qu.: 57417   1st Qu.: 1.000   1st Qu.:10.00   1st Qu.:15.00   1st Qu.: 101.0   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:  2.00   1st Qu.:0.00000   1st Qu.:0   1st Qu.: 0.0000  
 Median : 9.00   Median : 65765   Median : 2.000   Median :15.00   Median :22.00   Median : 110.0   Median :0.00000   Median : 0.000   Median :  4.00   Median :0.00000   Median :0   Median : 0.0000  
 Mean   :10.72   Mean   : 66550   Mean   : 2.341   Mean   :16.84   Mean   :22.93   Mean   : 112.8   Mean   :0.07306   Mean   : 1.775   Mean   : 10.12   Mean   :0.07306   Mean   :0   Mean   : 0.3091  
 3rd Qu.:15.00   3rd Qu.: 74909   3rd Qu.: 3.000   3rd Qu.:22.00   3rd Qu.:30.00   3rd Qu.: 124.0   3rd Qu.:0.00000   3rd Qu.: 2.000   3rd Qu.:  9.00   3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.: 0.0000  
 Max.   :37.00   Max.   :105565   Max.   :25.000   Max.   :53.00   Max.   :53.00   Max.   : 174.0   Max.   :3.00000   Max.   :34.000   Max.   :118.00   Max.   :3.00000   Max.   :0   Max.   :23.0000  
   tp_d_ener        tp_f_rake        tp_g_kaup        tp_h_kulj        tp_i_majo        tp_j_info        tp_k_raho        tp_l_kiin        tp_m_erik        tp_n_hall        tp_o_julk        tp_p_koul     
 Min.   :0   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.0000   Min.   :0   Min.   : 0.000  
 1st Qu.:0   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0   1st Qu.: 0.000  
 Median :0   Median : 0.000   Median : 0.000   Median : 0.0000   Median : 0.0000   Median :0.0000   Median :0.0000   Median : 0.0000   Median : 1.000   Median : 0.0000   Median :0   Median : 0.000  
 Mean   :0   Mean   : 1.457   Mean   : 1.032   Mean   : 0.9389   Mean   : 0.4436   Mean   :0.3972   Mean   :0.1056   Mean   : 0.1403   Mean   : 1.085   Mean   : 0.3936   Mean   :0   Mean   : 2.082  
 3rd Qu.:0   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 1.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.000   3rd Qu.: 0.0000   3rd Qu.:0   3rd Qu.: 0.000  
 Max.   :0   Max.   :33.000   Max.   :39.000   Max.   :33.0000   Max.   :19.0000   Max.   :7.0000   Max.   :8.0000   Max.   :20.0000   Max.   :14.000   Max.   :11.0000   Max.   :0   Max.   :94.000  
   tp_q_terv        tp_r_taid        tp_s_muup        tp_t_koti       tp_u_kans        te_takk          te_nuor          Myynti      
 Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0   Min.   :0   Min.   :1.000   Min.   : 0.0000   Min.   :0.0000  
 1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0   1st Qu.:0   1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000  
 Median : 0.000   Median : 0.0000   Median :0.0000   Median :0   Median :0   Median :2.000   Median : 0.0000   Median :0.0000  
 Mean   : 3.076   Mean   : 0.3653   Mean   :0.1172   Mean   :0   Mean   :0   Mean   :2.336   Mean   : 0.6414   Mean   :0.2441  
 3rd Qu.: 1.000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.:0   3rd Qu.:0   3rd Qu.:3.000   3rd Qu.: 1.0000   3rd Qu.:0.0000  
 Max.   :60.000   Max.   :16.0000   Max.   :3.0000   Max.   :0   Max.   :0   Max.   :4.000   Max.   :13.0000   Max.   :1.0000  
```

Appendix III Correlation values

	ko_perus	ko_yliop	ko_ammat	ko_al_kork	ko_yl_kork	te_taly	te_eil_np	te_plap	te_aklap	te_klap	te_teini	te_aik	te_elak	tr_mtu	tr_pi_tul	tr_ke_tul	tr_hy_tul	ra_as_kpa	tp_alku_a	tp_jalo_bf	tp_palv_gu	tp_a_maat
ko_perus	1.00000000	0.43848267	0.65739448	0.400279022	0.186012501	0.62505659	0.231568995	0.341676469	0.354561185	0.394043828	0.449371960	0.570078879	0.405593584	-0.34237301	0.55586343	0.67276633	0.312544823	-0.374456694	0.114261681	0.108542048	0.13061924	0.114261681
ko_yliop	0.43848267	1.00000000	0.49837181	0.656389849	0.603543895	0.75832154	0.355631698	0.259384357	0.359606223	0.435592924	0.494912153	0.767515984	0.454967208	-0.07749831	0.52125437	0.54242518	0.645253082	-0.243237987	-0.042119654	0.070388881	0.14445099	-0.042119654
ko_ammat	0.65739448	0.49837181	1.00000000	0.596004695	0.282580245	0.77946580	0.240892062	0.461968139	0.475806072	0.473667934	0.489176580	0.679813971	0.547407691	-0.36428158	0.40670263	0.82211312	0.471809777	-0.433430702	-0.007653920	0.131872611	0.09788572	-0.007653920
ko_al_kork	0.40027902	0.65638985	0.59600470	1.000000000	0.620260773	0.83402986	0.377744700	0.515529864	0.591787401	0.552377238	0.515718159	0.724769224	0.502246370	-0.03552871	0.34857582	0.61954070	0.718540717	-0.313791433	0.024077048	0.110143586	0.14365740	0.024077048
ko_yl_kork	0.18601250	0.60354389	0.28258024	0.620260773	1.000000000	0.74972981	0.124604300	0.360126714	0.472990091	0.594453849	0.563194111	0.550096368	0.532657543	0.31572059	0.17002990	0.33723668	0.882196111	-0.028059292	-0.052346853	0.045695686	0.15669387	-0.052346853
te_taly	0.62505659	0.75832154	0.77946580	0.834029856	0.749729807	1.00000000	0.329502204	0.546677355	0.618510031	0.655442883	0.649754601	0.856558888	0.672881443	-0.15343098	0.48234526	0.79574886	0.817085231	-0.404205737	-0.010822048	0.136242659	0.15403056	-0.010822048
te_eil_np	0.23156899	0.35563170	0.24089206	0.377744700	0.124604300	0.32950220	1.000000000	0.340642309	0.329000906	0.072853367	0.009684586	0.493918884	0.012674428	-0.27951302	0.52123024	0.24969447	0.085335773	-0.387599666	-0.022598313	0.086559600	0.05594065	-0.022598313
te_plap	0.34167647	0.25938436	0.46196814	0.515529864	0.360126714	0.54667735	0.340642309	1.000000000	0.852423106	0.523138589	0.300466612	0.437295965	0.102769980	-0.18078282	0.34764753	0.52812519	0.318210102	-0.343926107	-0.036735369	0.048684305	0.05917471	-0.036735369
te_aklap	0.35456118	0.35960622	0.47580607	0.591787401	0.472990091	0.61851003	0.329000906	0.852423106	1.000000000	0.679117450	0.391905117	0.456862214	0.117288813	-0.09299775	0.38195254	0.53622273	0.416250901	-0.323394161	-0.048962813	0.080247029	0.13671898	-0.048962813
te_klap	0.39404383	0.43559292	0.47366793	0.552377238	0.594453849	0.65544288	0.072853367	0.523138589	0.679117450	1.000000000	0.654770366	0.380210500	0.260462062	0.11002258	0.24760903	0.50080809	0.582954342	-0.068933423	0.007876627	0.114905748	0.08368845	0.007876627
te_teini	0.44937196	0.49491215	0.48917658	0.515718159	0.563194111	0.64975460	0.009684586	0.300466612	0.391905117	0.654770366	1.000000000	0.428550456	0.274639635	0.12345960	0.30980320	0.48516278	0.584876082	-0.028400917	-0.012981577	0.063996858	0.07164744	-0.012981577
te_aik	0.57007888	0.76751598	0.67981397	0.724769224	0.550096368	0.85655889	0.493918884	0.437295965	0.456862214	0.380210500	0.428550456	1.000000000	0.396778907	-0.27376351	0.57808024	0.69761699	0.627998916	-0.493189356	-0.023645700	0.133588086	0.12941718	-0.023645700
te_elak	0.40559358	0.45496721	0.54740769	0.502246370	0.532657543	0.67288144	0.012674428	0.102769980	0.117288813	0.260462062	0.274639635	0.396778907	1.000000000	-0.12099836	0.12345681	0.53971011	0.625282828	-0.233718575	0.024611579	0.055288155	0.11850411	0.024611579
tr_mtu	-0.34237301	-0.07749831	-0.36428158	-0.035528712	0.315720588	-0.15343098	-0.279513023	-0.180782821	-0.092997755	0.110022579	0.123459601	-0.273763512	-0.120998364	1.00000000	-0.38782148	-0.51651608	0.281722144	0.650372316	-0.029502528	-0.035986780	-0.05376664	-0.029502528
tr_pi_tul	0.55586343	0.52125437	0.40670263	0.348575821	0.170029899	0.48234526	0.521230239	0.347647534	0.381952539	0.247609027	0.309803197	0.578080235	0.123456805	-0.38782148	1.00000000	0.41962776	0.162903272	-0.433376593	0.022739532	0.138250346	0.13704971	0.022739532
tr_ke_tul	0.67276633	0.54242518	0.82211312	0.619540697	0.337236684	0.79574886	0.249694469	0.528125194	0.536222733	0.500808089	0.485162778	0.697616993	0.539710110	-0.51651608	0.41962776	1.00000000	0.370451293	-0.553172569	0.018004905	0.123860968	0.16784310	0.018004905
tr_hy_tul	0.31254482	0.64525308	0.47180978	0.718540717	0.882196111	0.81708523	0.085335773	0.318210102	0.416250901	0.582954342	0.584876082	0.627998916	0.625282828	0.28172214	0.16290327	0.37045129	1.000000000	-0.039103151	-0.021044260	0.077928021	0.09085793	-0.021044260
ra_as_kpa	-0.37445669	-0.24323799	-0.43343070	-0.313791433	-0.028059292	-0.40420574	-0.387599666	-0.343926107	-0.323394161	-0.068933423	-0.028400917	-0.493189356	-0.233718575	0.65037232	-0.43337659	-0.55317257	-0.039103151	1.000000000	0.053205103	-0.082474200	-0.18500581	0.053205103
tp_alku_a	0.11426168	-0.04211965	-0.00765392	0.024077048	-0.052346853	-0.01082205	-0.022598313	-0.036735369	-0.048962813	0.007876627	-0.012981577	-0.023645700	0.024611579	-0.02950253	0.02273953	0.01800491	-0.021044260	0.053205103	1.000000000	-0.019853977	-0.10101349	1.000000000
tp_jalo_bf	0.10854205	0.07038888	0.13187261	0.110143586	0.045695686	0.13624266	0.086559600	0.048684305	0.080247029	0.114905748	0.063996858	0.133588086	0.055288155	-0.03598678	0.13825035	0.12386097	0.077928021	-0.082474200	-0.019853977	1.000000000	0.19234734	-0.019853977
tp_palv_gu	0.13061924	0.14445099	0.09788572	0.143657397	0.156693871	0.15403056	0.055940649	0.059174712	0.136718978	0.083688447	0.071647437	0.129417178	0.118504115	-0.05376664	0.13704971	0.16784310	0.090857929	-0.185005812	-0.101013487	0.192347337	1.00000000	-0.101013487
tp_a_maat	0.11426168	-0.04211965	-0.00765392	0.024077048	-0.052346853	-0.01082205	-0.022598313	-0.036735369	-0.048962813	0.007876627	-0.012981577	-0.023645700	0.024611579	-0.02950253	0.02273953	0.01800491	-0.021044260	0.053205103	1.000000000	-0.019853977	-0.10101349	1.000000000

	tp_b_kaiv	tp_c_teol	tp_d_ener	tp_f_rake	tp_g_kaup	tp_h_kulj	tp_i_majo	tp_j_info	tp_k_raho	tp_l_kiin	tp_m_erik	tp_n_hall	tp_o_julk	tp_p_koul	tp_q_terv	tp_r_taid	tp_s_muup	tp_t_koti	tp_u_kans	te_takk	te_nuor	Myynti
ko_perus	NA	-0.0359685718	NA	0.142807179	0.105489300	0.0296103216	-0.0302929782	0.068662695	-0.0457492179	-0.0203303141	0.054313181	0.075162284	NA	0.0223762470	0.15306474	-0.065396755	0.075348134	NA	NA	-0.108759975	0.255587320	-0.0258016479
ko_yliop	NA	-0.0406564059	NA	0.098883136	0.012259825	-0.0449942394	0.0706335536	0.169253625	-0.0317338142	0.0320378296	0.139431800	0.085406680	NA	0.0379885886	0.16030319	0.025725891	-0.018928688	NA	NA	-0.152774575	0.387505619	-0.0340577395
ko_ammat	NA	-0.0494024387	NA	0.177873043	0.099658001	0.0453945095	-0.0450936385	0.153875949	-0.0500170601	0.0508088042	0.100508023	0.108203953	NA	-0.054727141	0.12708458	0.022192351	0.136322967	NA	NA	-0.182894689	0.168861070	-0.0258376611
ko_al_kork	NA	-0.0054318957	NA	0.127858630	0.044128855	0.0856240681	0.0344361832	0.252916810	-0.0548309380	0.0781059934	0.164991493	0.038230402	NA	-0.002642992	0.12873019	0.088242057	-0.031869634	NA	NA	-0.120851990	0.304325368	-0.0378697888
ko_yl_kork	NA	-0.0059974531	NA	0.052136165	0.015629211	0.0141264279	0.0892874821	0.211137934	0.0215191295	0.0204677385	0.239100138	0.008285331	NA	0.0108190339	0.16155863	0.094654551	-0.031133775	NA	NA	-0.059949308	0.140912781	-0.0322096711
te_taly	NA	-0.0419119608	NA	0.176653982	0.069142261	0.0443586967	0.0269251261	0.253389128	-0.0338138579	0.0570765735	0.200322198	0.068295900	NA	-0.010849413	0.15601563	0.061420998	0.030571288	NA	NA	-0.209846506	0.325846287	-0.0442479523
te_eil_np	NA	-0.0165996314	NA	0.107186258	0.040174240	-0.0215638614	-0.0065580118	-0.001692732	-0.0525185975	-0.0111342255	-0.003545647	0.065885119	NA	-0.005612869	0.09316550	0.014841791	0.023642985	NA	NA	-0.094382234	0.610829620	-0.0203843851
te_plap	NA	-0.0373160207	NA	0.076748134	0.003829594	0.0116736955	-0.0102533562	0.132665922	-0.0931455676	-0.0058525778	0.194738197	0.029152088	NA	-0.033554065	0.07736917	0.077941843	-0.001208048	NA	NA	0.047313777	0.181274343	-0.0154057113
te_aklap	NA	-0.0348966745	NA	0.111016880	-0.012508233	0.0768946699	0.0277571061	0.097912453	-0.0992312383	0.0078644907	0.202551476	0.078651526	NA	0.0181783204	0.13571756	0.073603344	-0.012090946	NA	NA	0.149588619	0.200992539	-0.0150141979
te_klap	NA	0.0045644780	NA	0.127783745	-0.039878078	0.0046477994	-0.0117270655	0.162281786	-0.0306498645	0.0119413158	0.194672236	0.026574742	NA	-0.023398817	0.12176890	0.085882404	-0.011754450	NA	NA	0.212663131	0.036678107	-0.0097257124
te_teini	NA	-0.0134240580	NA	0.079239495	-0.001504676	-0.0367090451	0.0213405477	0.096726386	-0.0039493411	0.0266875771	0.154332715	-0.021674243	NA	-0.007641031	0.11728480	-0.033746334	-0.023373956	NA	NA	0.166016181	0.094707820	-0.0109968462
te_aik	NA	-0.0402399856	NA	0.172483091	0.121862170	0.0221694750	0.0477298626	0.209697189	-0.0288981882	-0.0012101710	0.133628933	0.133112905	NA	-0.002910639	0.10680220	0.043715728	0.003496610	NA	NA	-0.331288314	0.525121916	-0.0418157600
te_elak	NA	-0.0350406552	NA	0.081560244	0.033626623	0.0706152203	-0.0148748039	0.207947900	0.0118801115	0.1134253339	0.110158760	-0.021151254	NA	-0.008148421	0.12013516	0.067701417	0.095738949	NA	NA	-0.346277757	0.021617652	-0.0483148566
tr_mtu	NA	0.1250969010	NA	-0.112509977	-0.094005619	-0.0128803974	0.0207606102	-0.044623988	0.0899542703	-0.0189023697	0.065310920	-0.057040978	NA	-0.040075113	-0.02845010	0.037169432	0.015569929	NA	NA	0.353198152	-0.275908255	0.0187504035
tr_pi_tul	NA	-0.0175043398	NA	0.165278142	0.066496802	0.0246533946	0.0323075234	-0.052158380	-0.0370835891	-0.0376924251	0.080310071	-0.020519543	NA	0.0160243497	0.18790691	0.032849344	-0.036720338	NA	NA	-0.088049894	0.567209275	-0.0262093745
tr_ke_tul	NA	-0.0613859206	NA	0.174964162	0.087773410	-0.0307109819	0.0443552128	0.275586987	-0.0731236820	0.0738136242	0.144914731	0.141320842	NA	0.0535606475	0.12481849	0.063304829	0.058334781	NA	NA	-0.199285689	0.210084399	-0.0379996033
tr_hy_tul	NA	-0.0106209307	NA	0.092118801	0.027710872	0.1040625309	0.0205414130	0.202375253	0.0088624236	0.0348850682	0.192671439	-0.005610225	NA	-0.044465302	0.09379041	0.048182059	0.022164888	NA	NA	-0.158469453	0.075661265	-0.0311543509
ra_as_kpa	NA	0.1242684218	NA	-0.160501500	-0.122149328	-0.0838957812	0.0527569345	-0.084874278	0.0540023126	-0.0453863739	0.002020371	-0.091786039	NA	-0.074890943	-0.15971570	-0.066407488	0.048307410	NA	NA	0.288208961	-0.400213715	0.0542437357
tp_alku_a	NA	0.0177192041	NA	-0.031114201	-0.040992941	-0.0306282416	-0.0523439942	0.013286374	-0.0378723217	-0.0197485834	-0.040674388	-0.058016020	NA	-0.046320654	-0.06971863	-0.043001533	-0.064381053	NA	NA	-0.020659709	0.017425909	0.0075453716
tp_jalo_bf	NA	0.4587936336	NA	0.881691647	0.410602969	0.1133140102	0.1133996628	-0.031242400	-0.0024519428	0.0536457689	0.144172856	0.161332380	NA	-0.045847922	0.13772992	-0.054080823	0.036373118	NA	NA	-0.047184719	0.010574414	-0.0010681585
tp_palv_gu	NA	0.0670510946	NA	0.175073791	0.409107369	0.1322749557	0.4461407699	0.101712564	0.0240088090	0.0566575134	0.284517447	0.170740108	NA	0.6750318892	0.71138566	0.114584105	0.006521090	NA	NA	-0.041474535	0.029484449	-0.0169472772
tp_a_maat	NA	0.0177192041	NA	-0.031114201	-0.040992941	-0.0306282416	-0.0523439942	0.013286374	-0.0378723217	-0.0197485834	-0.040674388	-0.058016020	NA	-0.046320654	-0.06971863	-0.043001533	-0.064381053	NA	NA	-0.020659709	0.017425909	0.0075453716