CUSTOMER SEGMENETATION WITH THE HELP OF DATA ANALYSIS FOR COMPANY X

Master Thesis

Lappeenranta–Lahti University of Technology LUT

MBAN, Master Thesis

2024

Elviira Kairamo

Examiner(s): Associate Professor Jan Stoklasa

Assistant Professor Mahinda Mailagaha Kumbure

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Business and Management

Business Administration

Elviira Kairamo

Customer segmentation with the help of data analysis for Company X

Master's of Business Administration

2024

84 pages, 12 figures, 5 tables and 1 appendices

Examiner(s): Associate Professor Jan Stoklasa and Assistant Professor Mahinda Mailagaha Kumbure

Keywords: Customer segmentation, Clustering, K-means, Data Analytics, Marketing

This master's thesis examines how the customer segmentation of private forest owners can be renewed in Company X, in order to better understand the changes in the customer base and their attitude towards forest ownership. In recent decades, the demographic changes of forest owners have had a significant impact on the wood trade, such as the decrease in the number of full-time forestry entrepreneurs and the increase in the size of forest farms. This can also be seen in Company X's customer base, because Company X has begun to suspect that certain of its current customer groups do not represent customers as well as would be necessary, especially for marketing purposes. This thesis focuses especially on Company X's Customer Group 2, which is the largest and most heterogeneous group. A customer questionnaire was developed for the thesis, which was used to collect data about customers. The data used in the study were the data that Company X already has, as well as the answers to the questionnaire. The chosen segmentation method was K-means algorithm because of its of simplicity, efficiency, clarity, flexibility, and proven effectiveness. The results of the analysis and clustering of the data showed that the current customer segments are not optimal, especially Customer Group 2 needs an update. With

the help of data analysis and a survey, the customer segments could be divided a little more precisely, and even in the sample group there were a few customers who do not belong to this group at all. K-means clustering did not prove to be the best method for segmenting Customer Group 2, when evaluating it from the viewpoint of Company X's desires, although it identified some differences between the clusters. Clustering did not produce the desired results that Company X was after. Desired results were clear, evenly sized clusters that could be deemed to include customers that differ from other clusters majorly. So other segmentation methods are recommended in the future. A data-based approach can enable deeper customer understanding and help improve marketing, if descriptive information is obtained from customer groups.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Elviira Kairamo

Asiakkaiden segmentointi data-analyysin avulla Yritykselle X

Kauppatieteiden pro gradu -tutkielma

2024

84 sivua, 12 kuvaa, 5 taulukkoa ja 1 liitettä

Tarkastaja(t): Associate Professor Jan Stoklasa ja Assistant Professor Mahinda Mailagaha Kumbure

Avainsanat: Asiakas segmentointi, Klusterointi, K-means, Data Analytiikka, Markkinointi,

Tässä maisterin tutkielmassa tarkastellaan, kuinka yksityisten metsänomistajien asiakassegmentointi voidaan uudistaa Yrityksessä X, jotta voidaan paremmin ymmärtää asiakaskunnan muutoksia ja heidän suhtautumistaan metsänomistukseen. Viime vuosikymmeninä metsänomistajissa tapahtuneet muutokset ovat vaikuttaneet merkittävästi

puukauppaan. Tämä näkyy myös Yrityksen X asiakaskunnassa, koska Yritys X on alkanut epäilemään, että sen tietyt nykyiset asiakasryhmät eivät kuvasta asiakkaita niin hyvin kuin olisi tarpeellista erityisesti markkinoinnin avuksi. Tämä tutkimus keskittyy erityisesti Yrityksen X Asiakasryhmään 2, joka on sen suurin ja heterogeenisin asiakasryhmä. Tutkimukseen kehitettiin asiakaskysely, jonka avulla kerättiin dataa asiakkaista. Datana tutkimuksessa toimi Yrityksen X aikaisemmin keräämä data asiakkaista, sekä kyselyn vastaukset. Valituksi segmentointimenetelmäksi valikoitui K-means algoritmi sen yksinkertaisuuden, selkeyden, joustavuuden ja todistetun tehokkuuden vuoksi. Tulokset kyselyn analysoinnista ja klusteroinnista osoittivat, että nykyiset asiakassegmentit eivät ole optimaalisia, erityisesti Asiakasryhmä 2 tarvitsee päivitystä. Data-analyysin ja kyselytutkimuksen avulla asiakassegmentit voidaan jakaa hieman tarkemmin ja otosryhmässäkin oli muutamia asiakkaita, jotka eivät kuulu nykyään ollenkaan tähän ryhmään. K-means klusterointi ei osoittautunut parhaaksi tavaksi segmentoida Asiakasryhmä 2:sta, kun sitä arvioitiin Yrityksen X toiveiden näkökulmasta, vaikka se tunnisti joitakin yhtäläisyyksiä saatujen ryhmien sisällä. Toivotut tulokset Yritykselle X olisivat olleet selkeitä, tasakokoisia klustereita, jotka sisältäisivät asiakkaita, jotka eroavat suuresti muista klustereista. Klusterointi ei tuottanut toivottuja tuloksia, joten muita segmentointimenetelmiä suositellaan jatkossa. Dataan perustuva lähestymistapa voi syventää asiakasymmärrystä ja auttaa parantamaan markkinointia, jos asiakasryhmiltä saadaan kuvaavaa tietoa.

ACKNOWLEDGEMENTS

Kiitos kumppanille, ystäville ja perheelle tuesta.

Kiitos ihana Enklaavi ja Tetra, kun teitte opiskeluajoistani aivan mahtavia.

Kiitos Katri ja Jan kaikesta avusta tähän työhön.

Table of contents

Abstract

Acknowledgements

1. Introduction 10
   1.1. Purpose of this Master's Thesis 12
   1.2. Previous studies of forest owners in Finland 14
   1.3. Limitations 14
   1.4. Current situation in Company X 15
   1.5. Research questions 17
   1.6. Ethics of the thesis 18
   1.7. Structure of the thesis 18
2. Literature review 20
   2.1. Marketing 20
   2.2. Customer-focused marketing 21
   2.3. Customer segmentation 22
   2.4. Customer Segmentation in marketing 24
   2.5. Customer segmentation methods 25
   2.6. Clustering 26
   2.7. Proximity measures 30
      2.7.1. Euclidean Distance 31
      2.7.2. Manhattan Distance 31
      2.7.3. Cosine Similarity 31
      2.7.4. Minkowski Distance 32
   2.8. K-means clustering 32
   2.9. Agglomerative Clustering 34
   2.10 DBSCAN Clustering 35
   2.11 Fuzzy Clustering 35
   2.12 Validity of clustering 36
3 The Questionnaire 39
   3.1 Questionnaire making process 39

3.2 The questions 41
3.3 The sample 43
3.4 Execution of the questionnaire 44
4 Data 45
5 Data Analysis 47
5.1 Data preprocessing 48
5.2 Exploratory data analysis 50
5.3 Analysis of the questionnaire data 52
5.4 Chosen clustering method 56
5.5 Data Clustering 57
6 Results of clustering 59
6.1.1 Two clusters made with Manhattan distance 64
6.1.2 Two clusters made with Euclidean distance 65
6.1.3 Three clusters made with Manhattan distance 67
6.1.4 Three clusters made with Cosine similarity 69
7 Conclusions 72
7.1 Answering the research questions 73
7.2 Future research 77
References 80
Appendices 85
Figures 87
Tables 89

Appendices

Customer Questionnaire

Figures

Figure 1 Current situation on Company X

Figure 2 Cluster analysis process

Figure 3 Clustering methods and algorithms

Figure 4 Process of making the questionnaire (Chang et al. 2023)

Figure 5 Data Analysis process (Andrienko, 2020)

Figure 6 Distribution of Customer Groups for sample data

Figure 7 Distributions of data points with different proximity measures using 2 clusters

Figure 8 Distributions of data points with different proximity measures using 3 clusters

Figure 9 Distributions of data points with different proximity measures using 4 clusters

Figure 10 Distributions of data points with different proximity measures using 5 clusters

Figure 11 Distributions of data points with different proximity measures using 6 clusters

Figure 12 Distributions of data points with different proximity measures using 7 clusters

Tables

Table 1 Customer Groups of Company X

Table 2 Different categories of data

Table 3 Different Clustering Algorithms

Table 4 Silhouette Scores

Table 5: Davies–Bouldin index

Table 5: Calinski-Harabasz index

Equations

1. Euclidean Distance

2. Manhattan Distance

3. Cosine Similarity

4. Minkowski Distance

5. Silhouette score 1

6. Silhouette score 2

7. Davies-Bouldin index 1

8. Davies-Bouldin index 2

9. Calinski-Harabasz index 1

10. Calinski-Harabasz index 2

11. Calinski-Harabasz index 3

1. Introduction

Private forest owners in Finland are important for Company X. Private forest owners own more than 60 percent of forest land and 52 percent of forestry land. About 80 percent of domestic raw wood used by industry comes from private forests. (Karppinen, Hänninen, Horne, 2020) Getting to know the people who own the forests helps Company X succeed in the market. Knowing their customers helps the companies to get advantage over competitors. (Company X)

Significant changes can be observed in the forest owner population, which affect the timber trade in Finland. The change in the economic structure can be seen especially in the decrease in the number of full-time forestry entrepreneurs and decrease of number of people living in their forest estate. Only one in ten of the forest owners is a full-time farmer, and a third live permanently on their estate. (Karppinen, Hänninen, Horne, 2020). This can also be seen in the customer base structure of Company X. Most valuable customers for Company X are full-time forestry entrepreneurs, customers living further away from their estate and in a need of help with maintenance of the forest and ones whose goal for the forest is financial gain. These customers can be offered certain services, help and more catered customer service. Company X wants to find out if their customer base has more of these types of customers or if other valuable groups can be found. (Company X)

Forest owners' goals have changed in the 21st century towards features that emphasize security and income. The shares of multi-purpose and recreational users have decreased, while the share of those who emphasize security and income has increased. The age structure of forest owners has become outdated, but the proportion of baby boomers has shrunk. With EU membership, the size of forest estates increased, which affected the use of forests. Full-time agricultural entrepreneurs became active wood sellers, and this is reflected in the wood trade. (Karppinen, Hänninen, Horne, 2020)

As the number of forest owners has decreased since the beginning of 2000, the size of the wood trade has increased, indicating an increase in the size of the forest and an increase in the efficiency of wood harvesting. The effect of gender and level of education is significant in forest ownership, as men are overrepresented among forest owners, and as the level of education increases, the income level of the household increases, which reduces the need to sell wood. (Karppinen, Hänninen, Horne, 2020)

On average, Company X's forest owners customers own 30 hectares of forest in Finland. (Company X) When Company X started dividing its customers into groups, the limit for forest area that was needed to be sorted into Customer Groups 4, 5 and 6 was higher than the average customers forest. This means that even if the customer has some characteristics represented by these groups, it will not end up in to those groups, because the customer's area is too small. Customer Groups 4 to 6 include the most valuable customers for Company X. The group is once determined to the customer and Customer Groups have not been updated once since the creation of them. The problem that has arisen from this is that no updates have been made since the customer first gets the assigned group, some customers may now be in groups that they no longer belong in. The reason may be the growth of the estate, owners death or the person's desire to get rid of their forest by selling or leaving it as inheritance. Since the size of the forest estate area has been a significant factor in the grouping in Company X, more than 60% of Company X's customers have ended up in Customer Group 2. This group has various different customers inside it and the description of the group that was given to it might not apply to all of the customers inside it. The group currently does not provide important information to Company X. The need for this thesis has risen from this problem. The problem is that because customer groups have not been updated once since they were given to each person, Company X is not sure if they are accurate anymore. Customer Group 2 is one group in particular that is most unknown to Company X these days. Company X wants to get better insight to this group so that they can utilize the info mainly in marketing, service offers and in customer service. This is the group that this thesis will focus on.

Private forest owners play a significant role in Finland, owning the majority of forest land, timber growth, and carbon sinks, and supplying most of the industry's domestic wood needs. Around 350,000 forest estates are owned by ordinary families, engaging in about

100,000 timber transactions annually, with over 600,000 family forest owners in total. Changes in regulations, organizations, and increased competition over the past decade have led to new forest ownership options and greater access to digital services. The forest industry is making substantial investments, indicating a need for updated knowledge. (Karppinen, Hänninen, Horne, 2020)

Research on forest owners has analyzed demographic changes and their impact on timber sales and management practices. Studies have identified trends such as the emphasis on non-material goals in forest ownership and the decreasing timber sales due to aging forest owners. Gender and occupation also influence activity levels, with women selling less wood on average, and full-time farmers being more active in timber sales than other professional groups. (Karppinen, Hänninen, Horne, 2020)

1.1. Purpose of this Master's Thesis

Purpose of this master's thesis is to analyse the customer base of Company X and mainly the customers inside Customer Group 2. Purpose is to find out if it is possible to find more of the potential and valuable customers than the Company X currently has out of its customers from Customer Group 2 and if new segments can be found. This is done to better serve customers and find ways that marketing can be made more specific to the customer groups. Purpose of this thesis is outlined more specifically in the chapter about research questions.

The subject of this master's thesis is to analyse the customers and the already existing customer groups that the Company X has. Company X has 6 different customer groups. The Customer Group is determined to the customer when they first starts doing business with Company X. Customer segments have not been updated after the segment is first given to the customer. Before the year 2014 segmentation had not been done, so all customers that had done business with Company X before that were then assigned group by the professionals at Company X. This has led to the suspicion that the segments are not accurate anymore and need to be updated to then be used in marketing to better target Company X's segments. Customer group 2 also has the vaguest description of all the

groups, and for this reason the group may also have included customers about whom not so much is known in advance. Also giving the segments by hand to some of the Company X's customers might have led to some human errors when evaluating the customer. The customer segments are used in Company X's marketing to make the marketing material and deals more accurate to each customer. Segments help to have better knowledge on the customers and find personal aspects of each group for marketing, service offers and in customer service. (Company X)

This thesis tries to find out if the customers in Customer Group 2 belong into the segment that they were given, should they be moved into another segment or could the current segments be refined for example with subgroups by using customer data. Company X's aim is to find evenly sized customer groups from Customer Group 2 that would be different from each other and would include customers that are similar to each other and have some key characteristics that describes the group. Analysing this group is done by making a questionnaire to a sample of the group and using the data gained from that with addition of the data that Company X already has to analyse the problem. The questionnaire goes deeper into what kind of customer the people in the group are. After the questionnaire is done the results will be analysed and the data added to the existing data of customers. This data will be used in segmenting the customers with using chosen method.

The topic is relevant to the Company X because this research belongs in a Company X new project on the bettering of its marketing efforts. Company X has concluded that the segments that it has now are not good and relevant and they need to be changed to succeed. This thesis aims to do that. The topic is relevant to my field of thesis because it includes analytics, business development and marketing. The topic is also current to this field of studies because the usage of data in different areas of business is a global growing trend. Both quantitative research and qualitative research will be used in this thesis to answer the research questions by collecting the data and analysing it.

1.2. Previous studies of forest owners in Finland

Nationwide questionnaire studies about the forest owners in Finland have been conducted in 1999 and 2009. In 2020, nationwide questionnaire of forest owners was published as a postal and telephone questionnaire. The Suomalainen Metsänomistaja 2020 study was carried out by the University of Helsinki, Natural Resources Center, Pellervo Taloustutkimus and Työtehoseura. (Karppinen, Hänninen, Horne, 2020) This study will be significant in this thesis because it closely examines similar topics as thesis. This study is also referenced widely in Company X when dealing with forest owners. The study displays an up-to-date close-up view of Finland's forest owners. The study describes the background characteristics of forest owners and farms, the goals of forest ownership and forest sales. This study also uses sample and questionnaire to find information of forest owners.

1.3. Limitations

The subject that is analysed during this master thesis sets some limitations. Since the process is specific to the case company's core business process, findings can't be provided in detailed manner in this master thesis so that no advantage to Company X's competitors are provided. The data used in this thesis is based on the Company X's customers the data and its variables can't be detailed. Customer data is personal data, so the laws and regulations protecting it must be taken into account. The real names given to the segments will not be used and description of data will be done discreetly. The aim is to make this thesis so that it is possible to research the topic and provide all the needed information, to make conclusions without revealing any sensitive information about the business process.

In this thesis, priority is given to considering the requirements and objectives of Company X. The research approach involves initially focusing on understanding Company X's specific challenges, needs, and objectives within the context of the core business process. Subsequently, broader theoretical frameworks, methodologies, and insights from relevant literature are explored to address Company X's specific requirements effectively. This

sequential approach ensures that the research outcomes are tailored to meet Company X's needs while also contributing to the broader academic understanding of the subject matter.

1.4. Current situation in Company X

Company X is a prominent Finnish corporation operating within a specific industry. It is responsible for procuring wood from forest owners to support various aspects of its business operations. Throughout this thesis, the term "Company X" is utilized when talking about the company.

The customer base of Company X comprises both private individuals and businesses in Finland who possess the forest land that Company X procures. These customers engage in product sales, service purchases, and also can hold ownership stakes in Company X. Within the company's supply chain, customers represent the initial stage. Company X also wants to keep in better contact with customers and thus wants to maintain a good customer relationship. This essentially includes knowing what the customers are like. In Company X, products and services are marketed to customers, and because of this, Company X wants to get better picture of what different customers are like and how they differ from each other, so that targeted marketing, deals and more catered customer service can be offered as relevant as possible to each customer.

For this reason, Company X's customers have been divided into different customer groups. There are currently 6 customer groups in Company X. Each customer group has its own characteristics and distinguishing factors, with the help of which it is possible to market to these groups the services that suit them and manage business with them in a certain way. The characteristics of each customer group can’t be discussed in this thesis but the focus group, Customer Group 2 has the broadest and most vague description. Only the area of the forest is important in this group and otherwise the group does not have the same descriptive features as other groups.

These customer groups are assigned when the customer starts doing business with Company X. At this point the customer answers 5 question questionnaire and the customer

group is decided by that. This is decided with a kind of tree model, where each answer affects which group the customer ends up in. Tree model can't be shown in this thesis. There are problems that have been noticed in the Company X with this procedure. This questionnaire was implemented in 2014. Any customer that had done business with Company X before that was given a group by customer specialist. Problems with that raised when Company X realized that the customer specialist might not know the customer that they evaluated for the group. Human error is also prevalent in this practice. The questionnaire implemented in 2014 mainly categorized people with the area of the forest that the customer owned and anything below the certain limit ended up in Customer Group 2. Company X has reason to believe that this group should be updated. The current situation in Company X is that over 60% of the customer base belongs to Customer Group 2. This customer group is taken into deeper analysis to find the more profitable customers that have ended up in this group.

The current breakdown into customer groups can be seen in figure 1. Undone and Not done in this figure are customers who did not complete the survey or who for some other reason do not have a group, for example human error. Company X has no direct explanation for these. Customers in Customer group 4, Customer group 5 and Customer group 6 are the most profitable customers for Company X. Sizes of these groups are small. 13.4% belonging in Customer Group 4, 9.3% belonging in Customer group 5 and only 2.7% belonging in Customer Group 6 out of all the customer of Company X. The Customer Group 2 segment is overwhelmingly huge and heterogenous. Customer Group 2 group is not that valuable to the Company X because Company X does not know what the customers' goals for their forest are and what kind of services can be offered to them. The smaller size of the forest limits certain actions that are profitable for Company X to do and provide, so it would be benefitial to find customers in the large mass to whom Company X can offer products and services that are easier to offer to other groups based on their characteristics. Can there be found these types of customers inside Customer Group 2 and could new segments be made from those. One thing that should be considered is whether the current groups could be divided further inside the group and not create completely new additional groups to the existing 6.

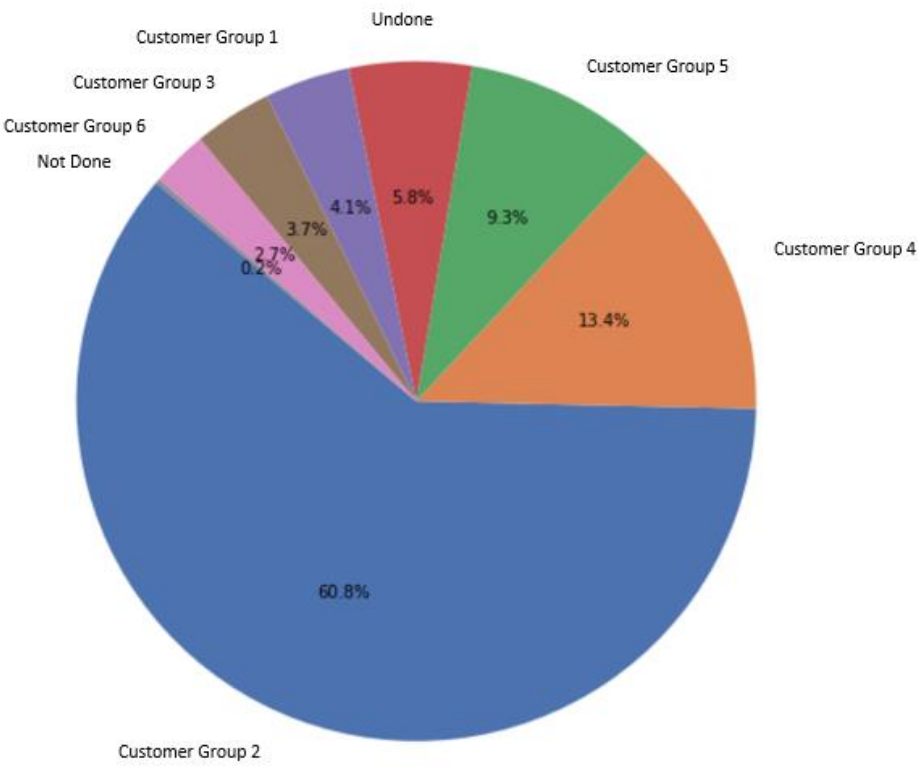

Figure 1 Current situation on Company X

1.5. Research questions

Purpose of this thesis and discussions with Company X have resulted in research questions that need to be answered to improve the current situation in Company X. These questions help to focus this thesis and help find answers to Company X's needs. There are three main research questions in this master's thesis:

1. *Are the customer groups in a need of an update and can data be used to divide customer into segments?*

2. *Can Customer Group 2 be divided further into groups that would provide valuable insights into the customers?*

3. *Can the chosen method of clustering and collecting more data be used for the rest of the customer segments?*

To find answers to these questions, customer data is analysed and more data is collected with questionnaire developed for this thesis. Customer questionnaire is made to dive deeper into the Customer Group 2 to get more insight if some customer can be separated from the mass.

1.6. Ethics of the thesis

Handling of people's personal data rises some ethical questions regarding privacy, consent, and security. Privacy is handled by using only the Company's computer, and software's when handling the data and writing the thesis. These are all protected. Company X has strict rules regarding handling and protecting personal data, and those are obeyed when writing the thesis. The data will be stored in the Company x's secured data storages. The General Data Protection Regulation is also followed when doing this thesis. GDPR is a comprehensive data protection and privacy law that came into effect in the European Union (EU) on May 25, 2018. Its primary aim is to give individuals greater control over their personal data and to establish clear rules for organizations that handle personal data (European Commission, 2023). Consent to use the data collected by the questionnaire is asked and also the use of the collected data is informed to the customers.

1.7. Structure of the thesis

This thesis is divided into seven main parts. Introduction, where the purpose for this thesis was discussed, current situation in Company X was explained and the research questions were presented. The second part is literature review where relevant literature is analysed and presented so that the theoretical framework for the thesis is clear. This will help to make the decisions on clustering and analysis methods. Third part explores the making of the customer questionnaire, framework and the aim of the questionnaire is discussed. In the fourth part data is presented. Fifth part delves into the analysis of the Company X's customer data and the analysis of the collected data. In this part clustering analysis on

Customer Group 2 is done. Clustering results are discussed in part six. In this part all the answers to the questions are found and next steps for Company X are considered. Seventh part discusses conclusions of the thesis. Research questions are answered, and future research is suggested.

2. Literature review

In this chapter theoretical background of the thesis is introduced. Marketing and customer segmentation and, why customer segmentation is a useful tool in marketing are the main topics in this part. Different segmentation methods are presented. Clustering methods being the selected method for this thesis. Different clustering methods are presented to then choose the most suitable one. This chapter helps Company X and the reader understand the theoretical background if this thesis so that the rest of the thesis is easy to follow and to understand.

2.1. Marketing

The fundamental question of why companies engage in marketing is a pivotal one. A market-driven strategy forms the cornerstone of a company's understanding of its position in the market and its customer base. A well-crafted marketing strategy draws its strength from the wealth of information it gathers about the market and its customers, and it wields this information to boost innovation, strengthen branding, drive sales, and fortify the company's standing within the market. In essence, marketing assumes a central role in determining a company's success (Fojt, 2005).

At the heart of marketing lies a crucial imperative: knowing your customers. It is at this juncture that the concept of customer segmentation comes into play. Marketing capabilities assume a prominent role in driving a company's performance. But the question persists: how can a company enhance its marketing process? (Vorhies, et al, 2011).

Looking ahead, Company X is gearing up for a marketing revolution in the coming years. A strategic pivot is underway, with a heightened focus on investing in marketing with precision, aligned with this fresh strategy. Placing customers at the centre of this paradigm shift, the company aims to hone its ability to engage in targeted marketing with even greater precision. The significance of zeroing in on customers and the specific segments they inhabit cannot be overstated (Company X).

2.2. Customer-focused marketing

A customer-focused strategy, as elucidated by Bose (2012), is a meticulously devised plan that places the needs and desires of customer segments at the forefront. It stands as a resounding declaration that the customer takes precedence above all else, and the company is committed to addressing each and every customer need comprehensively. This strategy, in essence, acts as a powerful catalyst for enhancing communication between the company and its customers, resulting in a clearer understanding of customer needs and preferences.

The versatility of this customer-centric can be used in many different parts of a company. The adoption of customer-focused strategies within the marketing domain empowers the department to make campaigns that are accurate to the customers. By honing in on the customer and their wants and needs, the marketing department can make strategies, ensuring that every marketing endeavour is aligned with the customer's expectations and needs. (Kansal, Bahuguna, Singh, Choudhury, 2018).

Marketing department of a company can utilize customer focused strategies because focusing on the customer and their needs in marketing helps the department make accurate and relevant marketing campaign. Segmentation can directly or indirectly influence the marketing strategy as it reveals many facts like for which segment the product will be good, customising the marketing plans according to each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company (Kansal, Bahuguna, Singh, Choudhury, 2018).

In essence, the customer-focused strategy represents an all-encompassing approach that also Company X is focusing on.

2.3. Customer segmentation

In general, segmentation aims to maximize the within-segment homogeneity and the between segment heterogeneity, so the segments are as similar inside the segment and as different as possible with each other (Jonker, Piersma et al. 2004). The concept of segmentation was introduced to marketing in 1965 by Wendell Smith. Segmentation was implemented as product differentiation at that time and Smith defined product differentiation as "the bending of demand to the will of supply". (Nairn, A., & Berthon 2003). Differentiating from product differentiation, which is based on extensive use of advertising and promotion aimed at the market, segmentation made it possible for a company to find new markets by targeting different marketing activities and advertisement to different groups of consumers called segments. Now this concept is used in the field of marketing to find the target audiences and delivering and communication value to them (Kotler, Keller 2009). Now in Company X this concept of segmentations is more important than before and is needed for communicating to different customers.

Customer segmentation is a process of dividing a broad customer base into manageable sub-groups of customers to find which groups are profitable, have high growth potential or other good qualities. Segmentation is a process of clustering similar types of customers to segments where they share common characteristics such as buying patterns, values, and demography (Smith, 1956). Customer segmentation is especially a marketing strategy. The goal of customer segmentation is to better understand and cater to the distinct needs, preferences, and behaviours of different customer segments to then find the most profitable ones. Well done customer segmentation is more profitable to a company and also the knowledge gained from that can be a factor in success or failure. Company X needs to be aware of behavioural changes in the customer base (Alves & Meisen, 2023). This allows businesses to create targeted marketing campaigns, tailor product offerings, and improve overall customer experiences. Segmenting the customer base also helps to perform different kinds of analysis on the customers of the company. Analysis on the customers aims to get better understanding on the customers of the company to then further implement more personalized service, marketing and products (Yossi, Keren, 2023).

Customer segmentation is closely tied to customer lifetime. Customer lifetime value refers to each customer sum of value that they bring for the company now and in the future. The higher the value is, the higher the profit for the company is by that customer. (He, Li, 2016) By segmenting customers, the company can better understand the relative importance of each customer and each segment. Separating valuable customers from the less valuable ones to then treat these group differently based on their value to the company (Yossi, Keren, 2023). He and Li suggested a three-dimensional approach to improving customer lifetime, customer satisfaction and customer behaviour. It was concluded that consumers are different from each other and so are their needs so that need to be taken into consideration. Segmentation assists in discovering more segment focused demands and expectations and service and products catered to them.

Customer satisfaction is one of the important standards to measure customers' opinions and experiences (He, Li, 2016). High satisfaction can increase the possibility that customers continue to use the services and product of the company and in the case of Company X, they continue to sell the forest land to Company X and not to its competitors. In the case of Company X customer satisfaction should be guaranteed for the more valuable segments such as Customer group 4, Customer Group 5 and Customer Group 6 customers because they are contacted more often than the other groups and services are more specified for them. This is why it is important to evaluate the current customer groups of Company X's customer base to see if there are customers that should belong in these more valuable groups.

Customer activity is an index of how close the relationship between the customer and the company is. Highly active customer activity means close relationship with the customer and the company, and the loyalty may be higher. (He, Li, 2016) Recognizing the segments with high activity is important for Company X so that it can invests in them but take into account the not so active segments in a different way and trye to activate them.

2.4. Customer Segmentation in marketing

Bergström and Leppänen, in their comprehensive book on "Yrityksen asiakasmarkkinointi," shed light on the evolving landscape of customer marketing strategies. Historically, businesses sought to cater to a broad customer base with products that had universal appeal. However, this approach has since proven to be ineffective. In response, the paradigm of customer segmentation emerged as a necessary shift in strategy. Initially, customers were categorized based on easily measurable variables like age, gender, profession, or income level. Yet, in today's fiercely competitive environment, distinguishing one's brand from the competition has become increasingly challenging. (Bergström, Leppänen, 2021)

In this contemporary landscape, it has become evident that these traditional demographic factors are no longer sufficient to explain the intricate nuances of consumer purchasing behaviour. To gain a deeper understanding, businesses must delve into the realms of customer values, purchase motives, and individual lifestyles to craft even more precise customer segments. (Bergström, Leppänen, 2021)

In the case of Company X, an abundance of readily measurable data, such as age, occupation, and past purchase history, forms the foundation of the initial customer segmentation process. However, the company is also committed to unravelling the less tangible variables that influence its customers. To this end, the questionnaire presented in this thesis aims to extract valuable insights that will further enhance the effectiveness of customer segmentation strategies.

But why is customer segmentation so important in marketing? It can aid decision-makers in making well-timed, informed choices to enhance marketing. (Wang, 2022). For instance, if a business wishes to send targeted direct emails or formulate one-to-one marketing plans for its most valuable customers, it is imperative to have a deep understanding of the various customer types within its base and the ability to efficiently target these groups. Customer segmentation makes marketing efforts more efficient by allowing for organized and cost-effective strategies, targeting larger groups of similar customers simultaneously.

Ultimately, the primary goal of this customer segmentation is to differentiate between the profitable and non-profitable customer segments. This differentiation empowers businesses to direct their marketing activities with precision, focusing their efforts on the more lucrative segments (Jonker et al., 2004). Finding non-profitable segments can help companies to also find solutions to make use of those. In essence, it's about optimizing resource allocation and enhancing the overall efficiency of marketing to drive growth and profitability.

2.5. Customer segmentation methods

Customer segmentation can be based in many different things. Company X could segment customers by demographic, geographic or psychometric measurements. Also looking at customer behaviours like purchase history or brand loyalty to make the groups is a way to segment customers.

Customer segmentation is a case where data mining and machine learning could be utilized. Data mining is the process of extracting or discovering similar patterns in large data sets. Data mining helps to separate the customers with similar patterns into similar groups also known as the segments and using data mining techniques helps the company to handle the large customer base. With data mining it is possible to find the variables that affect the segment the most. There are different methods that can be used to segment customers of a company. (Kansal et al, 2018)

One step in segmentation is customer representation. Representation tells us what the customer is like and can be done with feature selection. There are different methods of feature selection, these include for example, RFM-analysis and Purchase tree.

In an RFM analysis, three essential features are extracted from customer data. These are recency, frequency, and monetary factors. Recency pertains to the timing of the most recent customer activity, such as a purchase. Frequency measures how often a customer engages within a specified time frame, while monetary quantifies the amount of money a customer spends during that period. (Alves & Meisen, 2023) The underlying concept of

Purchase Trees involves representing purchased products as a tree structure, where products serve as the leaves and product categories function as the nodes within the tree. (Alves & Meisen, 2023)

Clustering has been proven to be an effective way to segment customers. There are many different clustering algorithms that can be used in making the segments. (Kansal et al, 2018) K-means is popular clustering method and this thesis will focus on that more than the others (Alves & Meisen, 2023). In this thesis first the focus is on clustering methods to segment the whole customer base to see if the existing groups need to be changed and then focus on different methods that can be used when analysing the questionnaire.

Even though clustering is one of the most used segmentation methods, there are also different methods to used. Self-organizing map is one method that can be used in customer segmentation. These are based on neural networks. Even though neural networks are supervised learning method it is also possible to use them in an unsupervised manner for clustering by pushing fully connected neurons towards the data points that are closest to them. Deep Embedding Clustering combines a deep neural network and a self-supervised probabilistic clustering. This approach produces explainable customer segments. The expectation-maximization algorithm performs a maximum likelihood estimation on given data which must consists of latent variables. Spectral Clustering is a graph-based clustering where distances between data points are represented by the edges. Evolutionary Algorithm are a class of optimization methods that find an approximate solution to a problem which also includes clustering. (Alves & Meisen, 2023).

2.6. Clustering

In today’s world data is more present and used than ever. Data can provide a basis for analysis, reasoning, decisions, and ultimately, for the understanding of all kinds of objects and phenomena. Company X is also utilizing data more and more in all operations of the company. Now data processing and analysis are in the forefront of marketing and the data about customers becomes more important in the marketing decisions. Segmenting data is one of the most important data analysis activities when it comes to obtaining information

of the data and utilizing it (Xu & Wunsch, 2009). Clustering is useful technique in many data mining and machine learning tasks such as data summarization, and customer segmentation and it has many different applications across many different fields like engineering, computer science and in this case in economics, more defined in marketing, where customer characteristics and purchasing patterns recognition and grouping of features is useful. Xu and Wunsch in their book *Clustering* say that the goal of clustering is to separate a finite, unlabelled data set into a finite and discrete set of natural hidden data structures, rather than to provide an accurate characterization of unobserved samples generated from the same probability. Clustering is a helpful tool to understanding the data that is new, because it can be used to can find out descriptive features that define the data and groups that can be found in it. Determinate this from the data provided by Company X answers the research questions.

Clustering is an unsupervised classification system. This means that it uses data that does not have labels. Training is not used in unsupervised systems, so it does not learn from any previous data. Clustering as an unsupervised algorithm aims to separate the unlabelled dataset to finite and discrete set of hidden data structures. The reason to use unsupervised clustering as a method is because it is a requirement to explore the unknown natures of the data. (Xu & Wunsch, 2009). Even though Company X has the data labels for the current groups available and so some supervised learning method could be used, it is not advisable because Company X does not want method that learns from the current data labels that it is not happy with. Supervised learning uses patterns from the data to make predictions for the data. Supervised learning would limit out finding new groups from the data, because it would try to find data that fits the labels given to it in the training set. Supervised learning is also often used to predict the data, and this is not the aim of this thesis.(Verdhan, 2020) Data labels for the current customer groups are not included into the data set. Because one of the problems with this data is to find out whether the current segments can be changed, that unsupervised learning methods like clustering are acceptable in this case.

Xu and Wunsch in their book *Clustering* have presented a four-step process for clustering analysis and this will be the basis for this thesis analysis as well. The proses is presented in figure 2. First step is to do feature selection and extraction, where the variables that are wanted to the clustering are selected. Second step is to design and decide the best suited

clustering algorithm. This includes determining appropriate proximity measure and making the criterion function. It is important to carefully investigate the characteristics of a problem to select or design an appropriate clustering strategy. Third step is the cluster validation. Objective assessments are needed to be used to critically assess the clustering results. The fourth and last step is the result interpretation where meaningful insights are provided from the original data and solve the problems that were set in the beginning. (Xu & Wunsch, 2009).

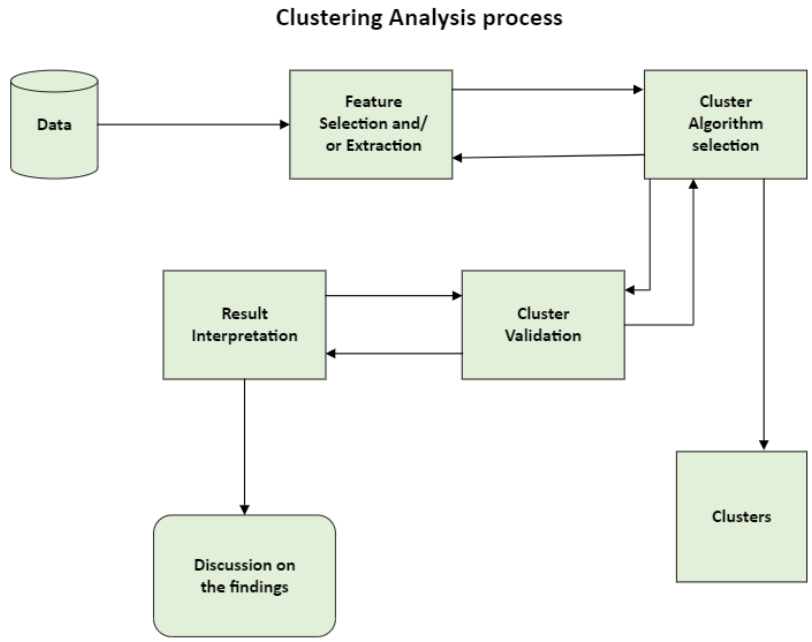

Figure 2 Cluster analysis process

There are many different types of clustering methods that can be utilized based on what type of problem the clustering problem is and what kind of data is used. Mainly the clustering algorithms are divided into partitional clustering, hierarchical clustering and density-based clustering shown in figure 3.

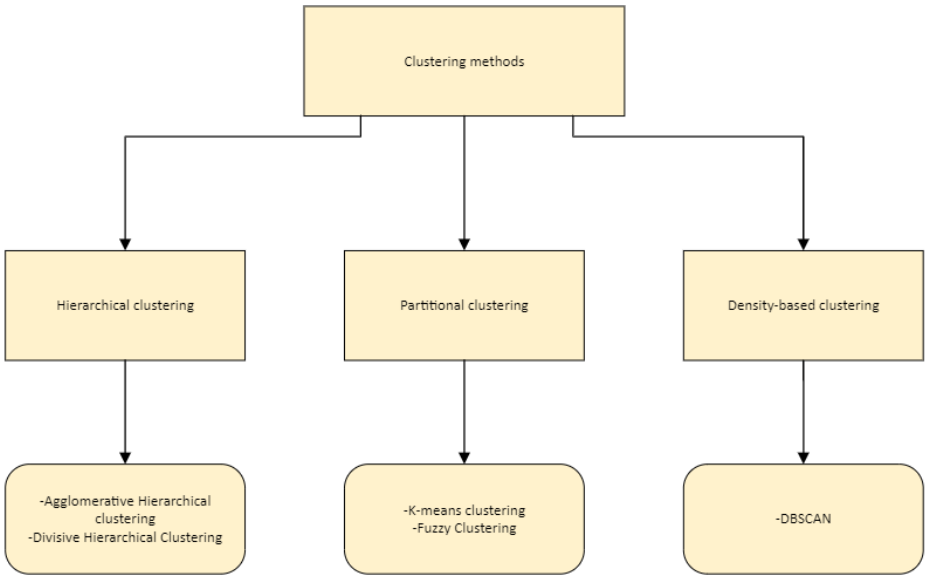

Figure 3 Clustering methods and algorithms

Partitional clustering is a clustering method that involves directly segregating data points into a predetermined number of clusters. In this approach, there is no inherent hierarchical structure. In this type of clustering, in the beginning it is decided how many clusters are wanted to divide the data into. This is suitable method for this Company X's problem because the goal is to have certain number of customer segments from the data and not a continuous number of them. Choosing also the number of clusters that they want the data to be divided into can help Company X to handle the large customer base. K-means is well-known partitional clustering algorithms. K - means seeks an optimal partition of the data by minimizing the sum - of - squared - error criterion. (Xu & Wunsch, 2009).

On the other hand, hierarchical clustering organizes data by partitioning the data into clusters with increasingly finer granularity. Hierarchical clustering is usually presented as a binary tree or a dendrogram that shows the division to the clusters (Cohen-Addad et all. 2019). Hierarchical clustering has both agglomerative and divisive clustering methods. Agglomerative clustering initiates the process with N individual clusters, each containing a single data point. It then employs a series of merge operations that progressively combine these clusters until all data points belong to a single group. In contrast, divisive clustering

follows a reverse course. Initially, the entire dataset is treated as a single cluster, and a procedure is employed to iteratively divide it until each cluster consists of only one data point. Hierarchical clustering methods lack robustness, and they are quite sensitive to noise and outliers. These are also requiring quite much computing power so might not be suitable to large data sets (Xu & Wunsch, 2009).

There can also be neural network-based clustering, kernel-based clustering, sequential data clustering, large-scale data clustering and high-dimensional data clustering. The types of clustering and decision which types are suitable with the thesis problem and the data are discussed later.

2.7. Proximity measures

It is important to measure the clustering result and see how well the data is divided into the clusters. The objects inside the clusters need to be as similar as possible to each other and as dissimilar to other objects that are in different clusters as possible. These differences are measured with closeness or distance between object and the cluster. These measures are called proximity measures. Proximity is the generalization of both dissimilarity and similarity. (Xu & Wunsch, 2009) When deciding the proximity measure, first it is needed to determinate what category does our data belong in. These categories are nominal, binary, ordinal and numeric. (Mehta et al. 2020)

Nominal	Binary	Ordinal	Numeric
Contain names that represent categories with no order	Contain only two values, 0 or 1	Same as nominal, but there is a specific order	Consist of numerical values

Table 2 Different categories of data

This thesis will focus on numeric and binary proximity measurement types, because our data is mainly numerical and binary. Numerical proximity measures include Euclidean distance, Manhattan distance, Cosine similarity, Minkowski distance, Jaccard similarity and some others. The choice of proximity measure is also dependent on the algorithm is

going to be used. The choice of proximity measure can significantly impact the results and the quality of the clusters obtained in a clustering analysis. (Mehta et al. 2020)

2.7.1. Euclidean Distance

Euclidean distance calculates the straight-line distance between two data points $a = [a_1, a_2, \dots, a_{n]}$ and $b = [b_1, b_2, \dots, b_{n]}$ in a data set. It is well-suited for continuous data and works best when the features are on a similar scale. (Mehta et al. 2020). It uses the length of the shortest path between two points in a straight line.

$$d^2(a, b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (1)$$

2.7.2. Manhattan Distance

Manhattan distance, or the city block distance that it is also known as, measures the distance by summing the absolute differences between two points a and b that cannot be raised diagonally, instead they need to be measure along either of the two dimensions. It's particularly useful when working with data that is distributed in a grid or network structure. (Mehta et al. 2020)

$$d(a, b) = \sum_{i=1}^{n} |a_i - b_i| \quad (2)$$

2.7.3. Cosine Similarity

Cosine similarity measures the cosine of the angle between two datapoints with respect to the origin (Jaskowiak, 2013). It measures the similarity between the two attributes a and b. It is often used for text data and other high-dimensional data, as it focuses on the direction rather than the magnitude of the vectors. It is particularly useful when dealing with sparse data that includes significant number of zeros. (Mehta et al. 2020)

$$S_c(a,b) = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2}\sqrt{\sum_{i=1}^{n}(b_i)^2}} \quad (3)$$

(Jaskowiak, 2013).

2.7.4. Minkowski Distance

Minkowski distance is a generalized distance metric that includes both Euclidean and Manhattan distances as special cases (Mehta et al. 2020) It measures the dissimilarity between two datapoints a and b (Jaskowiak, 2013). It allows you to adjust the distance measure by changing a parameter p to control the sensitivity to different dimensions, for p=1 it becomes Manhattan distance and for p=2 it becomes Euclidean distance. (Mehta et al. 2020)

$$d_p(a,b) = \sum_{i=1}^{n}((|a_i - b_i|)^p)^{1/p} \quad (4)$$

Based on the data that is used and the types of proximity measures, Euclidean distance, Manhattan distance, or Cosine similarity would be appropriate choices for proximity measures in this thesis.

2.8. K-means clustering

K-means is one of the simplest and maybe the best-known clustering algorithms and it is based on the partitioning principle. Partitional clustering is based on the idea that the algorithm tries to organize the data point to K-number of clusters while maximizing or minimizing the criterion chosen to the algorithm. In the case on K-means the criterion is minimizing sum - of - squared – error. The partition that minimizes the sum - of - squared - error is regarded as the optimal number of clusters and is called the minimum variance partition. (Xu & Wunsch, 2009)

K-means is sensitive to the initialization of the centroid. Even though K-means is a unsupervised learning method, the initialization of the centroids makes it less unsupervised than the other methods, because the algorithm is interfered with. K-means algorithm is always affected by initializations. (Sinaga, Yang, 2020).

The number of centroids can be calculated for example with elbow method, silhouette value or the GAP statistic or in the case on Company X determinate the number of customer segments can be done by testing different numbers and finding out what serves Company X's wishes the best. (Kansal et al, 2018).

Number of clusters for initialization can be done also by setting a limit for the centroids because the Company X does not benefit from too many cluster differences if they do not have a separate strategy planned for these. For this reason, the number of clusters would remain within certain limit values. The cluster labels are assigned by the algorithm by calculating the Euclidian distance from the closest centroid point and the proses is repeated until there is no change in the centroid positions of the clusters. (Kansal et all, 2018)

K-means clustering procedure has 4 steps (Xu & Wunsch, 2009):

Initialize a K - partition randomly or based on some prior knowledge. Calculate the cluster prototype matrix

Assign each object in the data set to the nearest cluster

Recalculate the cluster prototype matrix based on the current partition,

Repeat steps 2 and 3 until there is no change for each cluster

Even though K-means is a simple algorithm to use, it has some good features and some disadvantages that need to be considered before choosing this one. The algorithm works well for may differently types of problems and the results are compact. These methods works very well for large-scale data sets because K-means is approximately linear. One of

the main problems with K-mean is that it is sensitive to the initialization and if that is chosen poorly, the optimum cannot be guaranteed, and the algorithm can become stuck in local minimum. Other problem for K-means is that it requires lot of computational power if the initialization is done poorly. K-means is also very sensitive to outliers and noise in the data and those are needed to be dealt with in the data before using the algorithm. (Xu & Wunsch, 2009)

2.9. Agglomerative Clustering

The basic idea of agglomerative hierarchical clustering is that the clustering starts with N-number of clusters with one data point in them. The merge operations of the agglomerative clustering forces all the datapoints into one group. This is done by merging the two closest clusters together. The method gives us a hierarchy of the clusters (Ackerman et al. 2014). This is shown as a dendrogram that displays groupings of patterns in data. Dendrogram is formed by finding then the similar patterns in bigger and bigger groups (Bouguettaya et al, 2015). It is a tree-like structure that shows the hierarchical relationships between the objects in the data. The algorithm then cuts in the appropriate level of clusters where the groups are most distinct. Agglomerative hierarchical clustering methods also doesn't need any predefined parameters like the number of clusters that K-means needs (Xu & Wunsch, 2009).

But to define the agglomerative strategy it is needed to specify the distance measure between clusters. There are three most used distance measures for strategy, single linkage strategy, complete linkage strategy and average linkage strategy. The single linkage strategy, at its core, defines the distance between two clusters by examining the distance between their closest pair of data points. It focuses on the proximity of clusters by assessing the distance between the most neighbouring data elements.

In stark contrast, the complete linkage strategy measures the distance between two clusters based on the farthest pair of data points within each cluster. This approach considers the maximum distance between data points, which offers a perspective on the separation or dissimilarity between clusters. The average linkage strategy takes a different route by

defining cluster distance as the average of all possible pairs of data points, with one data point from each cluster. It considers the collective or average proximity between data objects in the two clusters. (Ackerman et al. 2014)

Agglomerative clustering can also be used to get the final number of clusters if it is given pre-clustered dataset. In this clusters with only few number of datapoints are considered outliers. (Xu & Wunsch, 2009) There are many different algorithms that are agglomerative clustering algorithms like CURE, BRICH, CF tree, CHAMELEON and ROCK.

2.10 DBSCAN Clustering

Density Based Spatial Clustering of Applications with Noise, also known as DBSCAN clustering is a density-based algorithm that relies on the density of the datapoints for clustering and creates clusters that are easily scalable and with arbitrary shapes. This algorithm is also able to identify noise or outliers in the data (Xu & Wunsch, 2009). Same as agglomerative clustering, it does not need specified number of clusters for this algorithm beforehand, so this works well for data where number of clusters are not known.

2.11 Fuzzy Clustering

The previous algorithms discussed can be called hard or crisp clustering methods, because in those the datapoints can only belong in one cluster at a time. In a case Where the datapoint could belong to multiple clusters and not strictly in just one, Fuzzy clustering could be used. In these the restrictions are relaxed and the object can belong in different clusters with a certain degree of membership. Basically, easily said datapoint A can belong in cluster 1 with the membership degree of 0,5 and to clusters 2 and 3 with degree of 0,25. The sum membership degrees needs to be one. These algorithms are useful in cases where the clusters or groups are not that different from each other and different relationships in the data are wanted to be found. (Xu & Wunsch, 2009) Membership degrees however are not similar to probabilities, so one can’t see the probability to belonging in each of the clusters (Klawonn et al. 2015). One of these methods is called Fuzzy C-means. Fuzzy C-

means (FCM) is similar to the K-means algorithm, but the difference is that the membership function is soft rather than crisp (Omran et al., 2007).

2.12 Validity of clustering

After clustering in done to the data, the results are needed to be evaluated. It is important to also see if the clustering structure derived is meaningful so that the results of the clustering don't become meaningless. (Xu & Wunsch, 2009) Cluster validity index measures how well the clustering partition done to the dataset reflects the underlying structure of the modelled domain(Gagolewski et al. 2021).

Depending on the chosen clustering algorithm and its class there are different testing criteria to use. These include external criteria, internal criteria and relative criteria. The process of validation starts with the identifying of the clustering structure and choosing the validation type. Then validation type is determined, and null hypothesis is made. Null hypothesis usually is a statement of randomness regarding the structure of the data. This can be example H0 states that there is no structure in the data. After this the baseline distribution under the null hypothesis is established and index calculated. This is done with some type of statistical technique like Monte Carlo or bootstrapping. After these steps the hypothesis is tested for no structure. (Xu & Wunsch, 2009)

Cluster validation can be made with different evaluating techniques. Silhouette score serves as a metric for assessing the similarity of an object to the cluster it belongs to in comparison to other clusters. The value falls within the range of -1 to +1. The higher the silhouette value is, the better match the object is to its own cluster. Negative silhouette score, on the other hand, implies that the datapoint should belong to some other cluster than the one that is assigned now. Moreover, silhouette score equal to 0 indicates the border of two separate clusters and could belong to either one. (Petrovic, Slobodan, 2006) Silhouette score is calculated by taking the difference between average distance inside the cluster and the minimum distance between clusters(Wang, Xu. 2019). Silhouette score is calculated as:

$$SS = \frac{1}{n}\sum_{i=1}^{n} s(x_i) \quad (5)$$

Where:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \quad (6)$$

Where SS falls between -1 and 1. In this formula $a(x_i)$ represents the average distance inside the cluster and $b(x_i)$ shows how distinct or well-separated a cluster is from the other clusters. (Ros, F. et al. 2023)

On the other hand, the Gap statistic is a measure of cluster validity that relies on a statistical hypothesis test. The Gap statistic functions by examining how much the dispersion within clusters changes in comparison to what would be expected under a relevant reference null distribution at each given cluster count. (Sinaga, Yang, 2020).

The Davies–Bouldin index was introduced in 1979 by David L. Davies and Donald W. Bouldin. It tells the ratio between the within-cluster distances and the between-cluster distances and computing the average over all clusters. It is similar to Silhouette score but the goal value of the Davies-Bouldin index is either 0 or very close to 0. (Petrovic et al. 2006) It first minimizes the variance inside the cluster and then maximizes the distance between cluster with formula:

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k} \max\left\{\frac{S_i + S_j}{d(x_i, x_j)}\right\} \quad (7)$$

Where S_i is the average distance inside the cluster:

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_j} d(x_j, \bar{x}_i) \quad (8)$$

In which d is the dimension of the data, k is the number of clusters and n is the size of the data. In these functions x_i and x_j represent the number of items and the mean value of the i th or j th cluster. (Ros, F. et al. 2023)

The Calinski-Harabasz index, introduced in 1974, compares the between-clusters variance to the within-cluster variance to assess clustering quality. The larger the index is, the higher the degree of dispersion between clusters is. It is calculated with the formula:

$$CH(K) = \frac{B(K)(N-K)}{W(K)(K-1)} \quad (9)$$

Where

$$B(K) = (\sum_{k=1}^{K} a_k ||\overline{x_k} - \bar{x}\,||^2) \quad (10)$$

And

$$W(K) = \left(\sum_{k=1}^{K} \sum C(j) = k||x_j - \overline{x_k}\,||^2\right) \quad (11)$$

Where K is the number of the cluster, B(K) and W(K) is the inter-cluster covariance and N is the number of samples. The higher B(K) is bigger the dispersion between the cluster is and the smaller the W(K) is the closer the relationship inside the cluster is. This creates a ratio of these and if the ratio is high the larger the Calinski-Harabasz index is. High Calinski-Harabasz index indicates good clustering. (Wang, Xu, 2019)

Each of these validation techniques offers unique insights into the quality of clustering results. While the Silhouette score provides an intuitive measure of cluster cohesion and separation, the Gap statistic and Davies–Bouldin index offer complementary perspectives on cluster validity. Silhouette score shows which objects are well within their cluster, and which ones are merely somewhere in between clusters. The average silhouette score provides an evaluation of clustering validity overall and can be used to select an appropriate number of clusters for the data (Rousseeuw, 1987).

3 The Questionnaire

In this thesis data from customer of Company X is analyzed. Company X has already some data that is collected from the customers but is also interested in utilizing a questionnaire to gain more information of the customers in Customer Group 2. Company is interested in looking into the values and ideas of the customers and finding what they are looking for in the maintenance, use and goals of their forest estate. The data that Company X wants will be focused more on variables that are not so easily counted, such as values of the customers, lifestyles, ideas, future goals and feelings. Gathering of the data will be done in a separate questionnaire that will be analyzed to find out more about the customers and used in segmentation. This data will be then added to the existing data that Company X has. The focus group for this questionnaire will be the drawn from Customer Group 2 and questionnaire will be send to a sample group. This is done because goal of Company X is to focus on trying to understand Customer Group 2. Analysis whether this kind of questionnaire will be useful for Company X in the future is done to hopefully the expand it to the whole customer base. Results from the segmentation and answer analysis will determinate that. If Customer Group 2 can be segmented into more manageable sized or useful groups, then this questionnaire is a good tool to collect more data to support segmentation to the whole customer base.

3.1 Questionnaire making process

Making of a questionnaire should start with general introduction on what Company X wants the questionnaire to achieve. Several parameters like questionnaire objectives, questionnaire sections and the time needed to complete the questionnaire and the answers are outlined in the beginning of making this type of questionnaire. The questionnaire should also be pretested to assess the content of the questionnaire and assessing the questionnaire process. This proses usually involves multiple rounds of testing. (Chang et al. 2023) In figure 4 can be seen the proses that this questionnaire will follow.

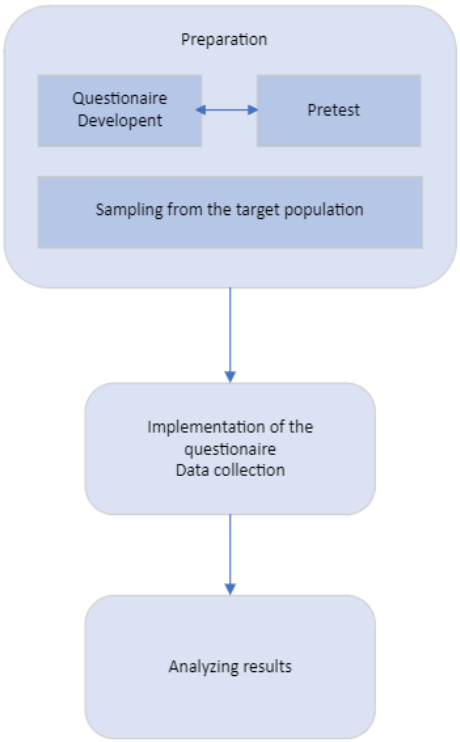

Figure 4 Process of making the questionnaire (Chang et al. 2023)

There is a dilemma concerning the incorporation of open-ended questions to enhance information quality to this questionnaire. Despite the potential for open-ended questions to elicit rich and detailed responses, they are often criticized for yielding poor-quality answers, posing challenges in coding, and requiring costly administration. Company X needs to evaluate if the answers for the types of questions are good enough. Company X wants this questionnaire to be short and easy to answer so that it gets as much answers as possible. The use of open-ended questions holds the promise of providing valuable insights into respondents' perspectives and thought processes (Davino et al. 2013).

In the case of this thesis the sample group will get the first round of questions that are agreed with the Company X and it will be assessed during this thesis if these questions are good or do they need changes to gain more insights. This is done because the thesis process time is limited and extended testing is not possible. This is the piloting phase of the

questionnaire. After this thesis is done Company X might want to edit the questionnaire based on the answers.

3.2 The questions

Company X wants these questions to richen our current quite sparce data that it has from the customers. The current data that Company X has contains many categorial variables made into dummy variables and multiple empty rows that describe inactive customers. However, Company X would like to somehow have these inactive customers separated from each other, because even though they are inactive in some aspects of the company's business areas, they might be different from each other in some way. Most of the inactive customers end up in the Customer Group 2. The segment currently has the majority of all Company X's customers and very little is known about them. The questions should be structured in such a way that Company X gets as much information as possible from the customers so that it can find differences and similarities in these customers.

The questionnaire was constructed to incorporate various questions from different sub-areas, and a joint decision was made with Company X to determine the most suitable ones for inclusion in the final questionnaire. The sub-areas covered included economy, environment, life values, goals, sustainability, and life situation. The construction of the questionnaire was significantly influenced by study "Suomalainen Metsänomistaja 2020" by Karppinen, Hänninen, and Horne. This study shared similar objectives to this thesis and followed a comparable methodology, utilizing a questionnaire aimed at forest owners.

In this thesis the target statements are related to six thematic areas. Goals related to recreation and leisure involved the significance of forests as residential or recreational environments and the opportunities the land offers for activities. Customers of Company X were also asked about the importance of being able to perform maintenance themselves. Regarding forest sales and income from sales, respondents were asked to assess the importance of regular sales income and income from land work, the importance of the land in financing major purchases, and the importance of obtaining firewood for personal use. Concerning nature and landscape, customers were asked about the importance of

preserving biodiversity in their own land and the importance of it as conservation areas. An assessment was also requested regarding the significance of landscapes beauty. Economic security included the significance of lands in facilitating access to credit and the security it provides for old age or unexpected situations. Assessments were also sought regarding the importance of lands as inflation-protected assets and as inheritance for family members. Emotional values were explored by inquiring about the inherent value of owning the land and the importance of it as a connection to one's homeland. In addition, respondents were asked about the significance they attributed to the economic value of residential and cottage plots.

Goal in making the questionnaire for this thesis is to get as much information as possible with short and easy to answer questionnaire with keeping in mind the similar themes that were studied in the study presented above.

In this thesis the final questionnaire has 14 questions ranging in different topics that would get the usable information on the customers that Company X wants. Five of these questions are same as the ones used to make the current customer segments. This helps Company X to compare the current answers with the previous ones and see if there has been any change in the customers. This will show if there are changes in the customer that have been sorted into Customer Group 2 since the first segmentation has been done. This will show if the current customer groups need to be updated. The rest of the questions in the questionnaire are aimed at getting as broad a picture of each customer as possible, so that grouping would be as easy as possible.

11 out of the 14 questions where multiple choice questions. One was to ask the customer id and one to ask age and the last question was open feedback question for overall feedback for Company X The questionnaire language was Finnish.

The questions for the thesis can be found in the appendix.

3.3 The sample

It is not possible to do the questionnaire for the whole customer base of Company X. There are number of reasons for this. First is the time limit of this thesis and second that the questionnaire can't be made for the whole customer base. Company X also can't reach each individual easily so it is best to do the questionnaire for a sample out of this customer base. Second Company X wants this questionnaire to especially be done for customer in the existing in Customer Group 2 because the knowledge that Company X has from that group is vague at the moment and it can't be separated other than with the amount of forest land owned. This is why Company X wants the questionnaire to be done to a sample and not done to all customers. Thirdly as said before there is not enough time to do testing for the questionnaire so the first round of questions is send to a sample for also to evaluate the questions before Company X decides based on this thesis if this is a good way to get more data from the whole customer base. If this questionnaire is success and conclusions are that these questions are sufficient way to get more and better data to separate the customers in to groups, it will be implemented to all customer and will stay as a permanent way to segment customers.

Sampling can be approached through two main methods: nonprobability sampling and probability sampling. Nonprobability sampling doesn't ensure that everyone has a specific chance of being chosen and encompasses techniques such as convenience sampling. Purposeful sampling is selecting participants with specific attributes such as the segment Customer Group 2. On the other hand, probability sampling, commonly employed in quantitative questionnaire studies, randomly selects participants. This method includes techniques like random sampling, stratified sampling, cluster sampling, systematic sampling, and multistage sampling. (Chang et al. 2023) In the case of this thesis mix of purposeful sampling and random sampling will be used, because sample is limited to be out of Customer Group 2 but Company X wants them to be randomly picked. Sample also need to be made out of customers that have shared their e-mail address with Company X so that the questionnaire can only be send by e-mail. Sending the questionnaire by post has been ruled out because of the time limit of this thesis. The location, amount of forest land or any other variable will not be taken into consideration in the choosing. The chosen customers had Finnish as their preferred language. The sample should be as random as the

customer group is so that the sample portrays accurate knowledge on how different the people on the existing Customer Group 2 are. Selection was done by Analytics Specialist at Company X that determined preferred language, existing customer group to be 2 and need of a email address and pulled the random customers from CRM database. The sample size was 3000 customers. Hoped answering rate would be 10% or 300 individual answers to the questionnaire. The questionnaire is not anonymous because first question is asking the customer-id of the customer so the answering rate might be lower because of this.

3.4 Execution of the questionnaire

Questionnaire has been made in collaboration with different professionals currently employed at Company X. Questions have been reviewed many times and the questions decided are so that Company X will get the information that it wants.

Questionnaire is done with Microsoft Forms questionnaire program. Questionnaire was sent on pre-decided day to 3000 selected customers and answering time was two weeks. Reminder email was sent week after the first email. Questionnaire was not anonymous so that the answers can be added to already existing data of the selected customers.

In the first round of sending the questionnaire got 148 answers and it was decided that a reminder would be necessary to spike the answering rate up. After the reminder questionnaire got 220 answers. The hope 300 answer was not reached so second round of emails were send to new customers. Company X took 1000 additional customers with the same requirements as before and the questionnaire was sent to them on also on pre-decided day and answering time was one week. After the one-week questionnaire had 477 individual answers and Company X was satisfied.

The total number of the final sample had 4000 customers, and the answering rate was 11.9% with 477 answers.

4 Data

Data used in this thesis include the current data that Company X has for Customer Group 2 and the additional data from questionnaire. Answers from questionnaire will be analysed first separately from the current data but will be added to the current data for segmentation. The current data that Company X has that will be used in this thesis includes variables that describe the level of relationship with Company X, amount sales of forests and bought services. Historical sales, types of cutting of the forest, potentials for selling forestland and buying services that Company X calculates and other variables that Company X has deemed to give information about the customer. This data has 39 columns the number of customers in that group as rows.

The questionnaire had 477 individual answers. First it is needed to make some changes to the data, because some of the answers are not usable in this thesis. If the customer had not filled their customer id, answers from that customer can't be used in this thesis. Company X want to combine the data that it already has from Customer Group 2 to the data collected from the customers with the questionnaire, so it is crucial that the "*Question 1: Customer Id?/ Asikasnumerosi?"* was answered with correct id number. Customer id is used and the connecting variable between tables.

There were 105 unusable answers where either the Question 1 was blank or answered wrongly with name, other number or something else that is not customer id. The dataset containing the answers of the questionnaire has 372 rows representing the individual customers. Average age of the customer who answered was 63,77 years and the youngest was 26 and oldest 91 years old.

These answers of the questionnaire are first analysed by themselves, and conclusions are made from that. After this the questionnaire data variables are added to the original data of Customer Group 2. The data obtained from the answer to the questionnaire gives information on the customers' goals for their forest, the distance of the person from their own forest is also considered important, because the further away the person lives, the

more often they needs help. Does the person use online services and do they do wood sales in the web. Do they want to part with the forest by selling or as giving it as heritage in the close future and do they think that financial gain is important and what do they do to achieve this. These are some variables that Company X has deemed useful in describing their customers. This combined dataset is used to make assumptions of Customer Group 2 and to make the final clusters of the customer segments.

5 Data Analysis

The analysis part of the thesis is done in Databricks cloud platform and analysis is done in Python-coding language. This is chosen because the policy in Company X is to use these tools.

Data analysis can be viewed as a multi-step process that is presented in figure 5. This thesis will follow this proses to do the analysis. This process starts by exploring the characteristics of our data, aiming to ascertain their suitability and determining the necessary preparations for analysis. In this step we determinate what the data is going to be used for. This focus was presented in the research questions.

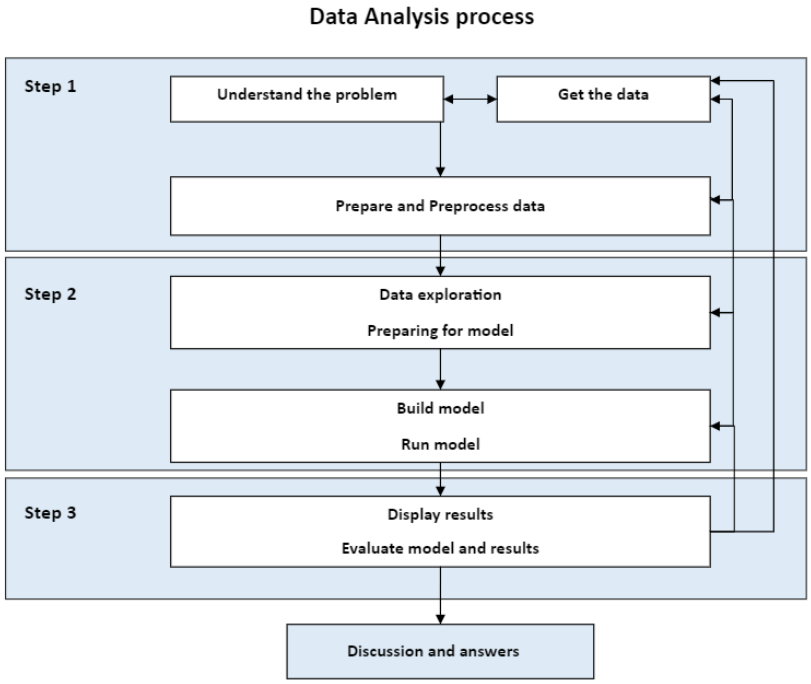

Figure 5 Data Analysis process (Andrienko, 2020)

Following this, the data is prepared and preprocessed. Subsequently, the prepared data will be subjected to analysis, with the primary objective being the construction of a model capable of facilitating decision-making to address our inquiries and test hypotheses.

Following the analysis, the models are evaluated, and the results are presented. A discussion will then ensue to address the main questions and assess whether this model effectively segments the customers of Company. (Andrienko, 2020)

The raw data gathered from various sources for this thesis needs to be preprocessed before any analysis can be done. The main point of preprocessing the data is to deal with missing data, scale and or normalize the data so that it is good to use in models.

After preprocessing the data testing the different customer segmentation methods is started. This mainly includes different clustering algorithms. The result evaluated and analyzed and the model that gives the best and the most helpful results to Company X is chosen.

5.1 Data preprocessing

Because the data is collected from different sources and describes different aspects, the data is quite sparse in the beginning. Data from different sources is displayed in different data tables. The data tables are combined according to the member ID, so that all the necessary information from each data table can be obtained separately for each different customer. Tables are joined using the left-join method. This is chosen because left-join returns all rows from the left table, and the matching rows from the right table. The main table is the right-side table. The result is NULL from the right side, if there is no match. The tables are connected to the main table, which has the member IDs of every customer of Company X. One row in the main table represents one customer of Company X. This data is quite sparse, and some customers might have empty cells. The main problem for the data is that if the customer does not have activity in some column, for example the customer has not bought any services, they will not have anything in those columns. Customers with empty cells in some columns can’t be eliminated from the data because they describe some customers less active behavior.

Before joining the tables unnecessary and duplicate variables are dropped from the data tables. This is done with the help of professionals in Company X to decide what variables give crucial information about the customers. Company X only needs variables that give

useful information on what type of customer the individual is. In this step dummy variables are also made from categorical variables in the data tables, so that they can be analyzed in the clustering models. This separates the categories from the original variable into their own variables and give the row value of 1 in that variable which corresponds to the original category of that customer.

Variables are grouped together so that each customer only has one row of data before joining these tables together. This mainly affects the data of purchased services and sold forest land because one customer might have done this many times and each individual purchase or selling of forest becomes with one row with a sum. After these changes have been made to the data tables these are joined as one data tables that now is the main data to be preprocessed and analyzed.

There are different types of customers in Company X. Some individuals exhibit recurrent and substantial transactions of their forest lands, while others remain devoid of any sales activity. Similarly, there exists a spectrum of engagement with services, ranging from frequent consumption to complete abstention. So the data table has many empty cells across various variables. Managing these is important to ensure that the data portrays each customer's profile. It is imperative to refrain from arbitrary imputation of values or extrapolation from neighboring cells, as such practices risk distorting the integrity of individual customer data. For instance, should a customer exhibit no history of service purchases, it would not be accurate to the customers to assume averages from other customers for them. Likewise, deletion of rows containing empty values is untenable, as it would introduce biases by disproportionately excluding less active customers. The most suitable approach within this analytical framework involves populating empty cells with zeros. This strategy transparently communicates that certain customers have not engaged in activities pertinent to the variables in question.

Some other steps included in the preparing phase of the analysis is to change all the columns names into same format and delete letters and alphabetic characters that are complicated.

Data is normalized before clustering can be done. This is done with standardization. Standardization involves transforming the data to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of each feature from the data and then dividing by the standard deviation of the feature. Data needs to be standardized because all the features need to be equally weighted in the clustering process. After analyzing the results some features might be more important than others and it would be important for Company X to evaluate if they want to emphasize some features more than others.

5.2 Exploratory data analysis

Now after the initial changes were made to the data and the final dataset containing all the relevant data is made, let's look at the characteristics of the data before considering the model that wanted to use for segmenting the customers. This is done with exploratory data analysis.

The final data has 39 columns and over 90 000 rows and each row display the information about different customer. The summary statistics of the data that display basic summary statistics, such as mean, median, mode, range, and standard deviation, are looked at to understand the central tendency and dispersion of the data.

In columns that describe bought services and sold forest land min and max value have a big difference. Difference is thousands in most of these. This tells us that there is wide range of values that the customer can get in these variables. A wide range can affect the mean and median differently. The mean is sensitive to extreme values, so if there are outliers, it can be skewed. The median is less affected, as it represents the middle value when the data is sorted. This can be also interpreted so that Company X many different types of customers that buy and sell differently. This is why it is so important to find similar customers from this mass. This can be seen in the data.

In many instances, it is observed that summary statistics reveal a substantial disparity between the mean and median values across numerous columns, the latter of which

consistently registers at zero. When the mean substantially exceeds the median, an indication is generally provided that the data distribution is positively skewed. Positive skewness signifies the presence of a few exceptionally large values exerting influence on the mean in the direction of the skewness, while the median, as the central value, remains comparatively less impacted by extreme values.

From this, it is inferred that outliers are present within the dataset. This is attributed to the presence of certain customers of Company X who significantly surpass others in magnitude, thereby exerting a distorting effect on the data.

Numerous variables consist of categorical data, predominantly populated by zero values. Within the dataset, numerous rows are void, represented by zeros, prompting consideration for the necessity of augmenting the data quality to facilitate informed decision-making.

Two variables exhibit leftward skewness, while others skew rightward. The variable demonstrating the greatest variance is variable x which delineates the geographical location of the customers. Additionally, it is noteworthy that an examination of another variable reveals Company X's greater propensity for potential customers compared to non-potential ones. These potential customers have engaged in forest land transactions with Company X and its competitors, signaling their potential significance to Company X, contingent upon appropriate marketing and sales strategies.

Most of the variables demonstrate no significant correlations with each other, albeit a notable subset exhibits correlation. The most prominent correlation is observed between two variables 1 and 2 which elucidate the presence and extent of procurement transactions by the customer, thereby explicating the rationale behind their correlation.

Variable customer Id exhibits the lowest correlation with other columns, predominantly displaying negative correlations. Given that it solely represents the customer ID number, this lack of correlation is expected. Consequently, this column is slated for removal from the dataset prior to segmentation. Careful consideration is warranted regarding variables displaying high correlation.

Drawing from the insights gleaned during exploratory data analysis, decisions must be made regarding the selection of variables for inclusion in the segmentation model aimed at delineating customer segments.

5.3 Analysis of the questionnaire data

Before questionnaire data can be added into the existing data of Customer Group 2, some data preprocessing and initial analysis has to be done. First the questionnaires data is cleaned so that it has no empty or wrong customer id's and the data is made uniform. In this step some changes were needed to be made to the answers in Question 2: Age? And drop anything other than the number. The open feedback question was dropped in this step, and it will be analyzed separately from the rest of the data, because long feedback is not optimal to analyze using Python. Column names were also made clearer and shorter for easier data handling.

The data is analyzed first the similar way that the customer segments have been made in Company X when the customer first joins. Some of the questions in questionnaire are same as used in the original customer segmentation in Company X. All of the customers that answered the questionnaire belong in Customer Group 2 and it is crucial to compare whether their answers have changed to these questions since they first answered them as they started doing business with Company X. This can help you determinate if the customer groups that Company X has are accurate anymore or should they be updated. This will help us answer the research question 1. Are the customer segments in a need of a update and can data be used to more accurately divide customer into segments? The only difference from before is that answer options have been added to two questions. This was felt to be a necessary change even before the questionnaire was sent, because the earlier questionnaire with which the customers are divided might be misunderstood.

There are 4 questions to first analyze. These are questions 3, 8, 9 and 11 found in appendix. These are similar to the questions that Company X initially used to determinate group for customer when they first started doing business with them. These questions determinate the customer groups by assigning the group based on a tree model. Using these

questions and analyzing them similarly as Company X has done, can give more precise groups than originally because some of the answer options are much more specific to a certain group than in the original way of dividing the customers. Also the limit for forest area is not limiting the groups in this analysis. Dividing the customers the same way as original groups are given but not limiting groups 4, 5 and 6 could now give subgroups for Customer Group 2 because customers could be divided into all groups. Of course, this cannot directly replace the previous division that Company X has done, but it can give indications that some update to the groups should be made. Also, at this stage it can be seen if some of the customers now currently in Customer Group 2 should be moved to some other group based on the more specific answers that describe other groups. Analyzing the questionnaire answers first this way can help Company X decide if this type of questionnaire is useful in obtaining more useful information from Customer Group 2 than before.

Looking at the grouping results based on this analysis, seen in figure 6 it can already be seen how much more evenly the customers of this sample are divided into different groups. Company X would want more evenly sized groups rather than one group dominating the others. Only 24.2 percent of those who responded to the questionnaire would end up in Customer Group 2. Four of the segments are similar sizes and 3 of these are the most important for Company X. There are still big number of customers falling into Customer Group 2. 0.3% of the customers who answered this questionnaire did not end up in any group. This can be caused by empty answers in some of the 4 questions taken into this analysis. This cannot be directly compared or mirrored to the rest of the customer base or to the previous division, because in the previous assignment of group forest area was of great importance, but if it is not allowed to affect the final groups so much, then with a more detailed questionnaire, it could be possible to get groups that better match the characteristics of each customer.

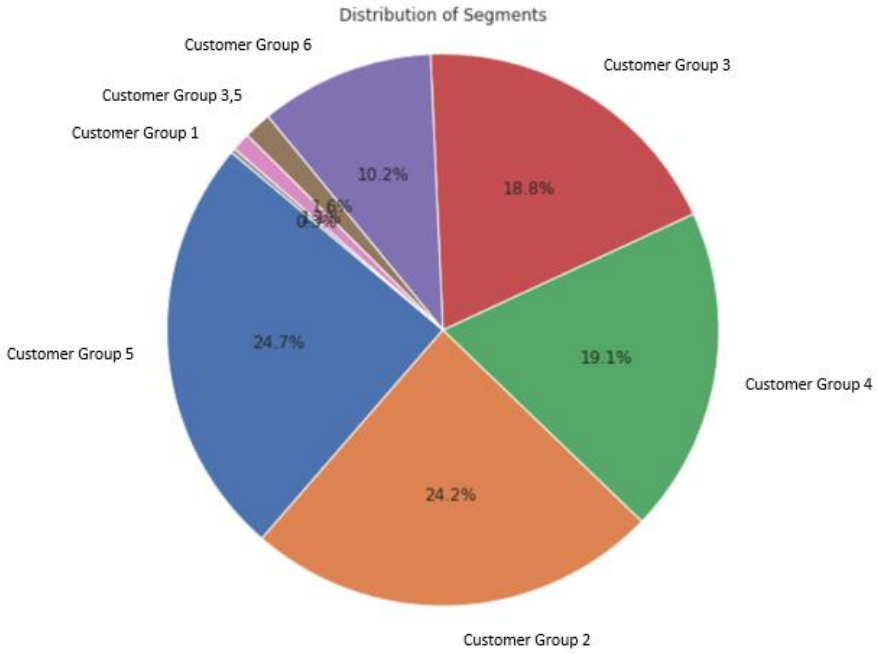

Figure 6 Distribution of Customer Groups for sample data

Results of this analysis can also help answer research question *2. Can Customer Group 2 be divided further into groups that would provide valuable insights into the customers??* Because this questionnaire has helped to make the customers in Customer Group 2 divide in more specific groups. Based on the questions and answers one customer group was added based on question 8. This question defines two ways to completely give up the forest, either as an inheritance or by selling it. In the past, Company X has only had one customer group for these types of customers, but this question allows us to separate these two different procedures from each other. There are different services that can be offered based on this difference for these customers in Company X. The question 8 answers help to differentiate customers that would in the original sorting end up in Customer Group 3 into two smaller segment. This differentiation will help Company X to offer better and more accurate services to those who end up in these groups. One group can be offered inheritance transfer services or sales services, and another group can be helped in transferring the inheritance. 18.8% of the sample group belong into Customer Group 3a and 1.6% into Customer Group 3b.

Based on analyzing grouping the customers with similar tree model as before it could be said that the group can be made more accurate than before because some of the customer would end up in different group than in Customer Group 2. Update for customer groups is needed in Company X. Analyzing the goals of the customers asked in question 13 with statements that the customers could choose it can be seen that the customers have chosen some similar statements that describe the group that was now assigned for them. This can determinate if the groups that are made are accurate to what the customers think themselves of their goals.

Question 13 includes different statements that correspond with the customer groups. Statements are meant to describe the characteristics of groups. When looking at how the statements divide between the new grouping there can be seen some similarities with what group the customer was now sorted into and what type of statements they has chosen. To analyze the answers to question 13, statements that correspond with customer groups were given a numerical value corresponding with that group. We have 6 of these statements. The other statements can’t fully be corresponded to one group. This way we can see how the customers themselves have evaluated their own group and how well this correlates with Company X's customer group. By analyzing this I found that 41,12% of customer have chosen similar statements as what the group would be for them given by the tree model. This is a good indicator that the groups made with this type of tree model are not as accurate to the ideas of the customers that we would want them to be but that there is some similarity. After analyzing the statement it can be said that this trial questionnaire should be modified if we want to use the tree model in segmenting the customers if the clustering is not deemed usable. The statements that are asked in question 13 should be even more focused on just describing the customer groups. This analysis also raises the question of whether customer grouping should be completely rethought. Instead of using old groups for Customer Group 2, should completely new groups be established? Fortunately, clustering is able to create new groups for us based only on data, without us setting any prior requirements about the content of the group. For example, deepen existing groups in some way, for example by dividing existing groups within them according to certain characteristics. To test different groupings for Customer Group 2, 2-7 clusters are tested for the data. The upper limit was decided on 7 clusters because currently Company X has 6 groups and based on the analysis of the questionnaire one additional was found. Company

X does not want too many groups for its customers. With this, it is hoped to find the most reasonable way to divide Customer Group 2 and hopefully in the future also other customer groups, depending on what kind of results are obtained from the research. These will be explored in more detail in the clustering chapter of this thesis.

5.4 Chosen clustering method

In Suomalainen Metsänomistaja 2020 K-means clustering was used to make groups out of their sample of forest owners in Finland. K-means was used to find groups with similar focus and goals for their forest. This has influenced my decision on what algorithm to choose. (Karppinen, Hänninen, Horne, 2020) This study is one off the most important studies regarding customer segmentation in in the field where Company X operates and is widely referenced in Company X when discussing customer segmentation. The study was also influence for this thesis.

After going through the methods presented in table 1 and reading academic literature, I have come to the conclusion that the partitioning model is best suited for customer segmentation in this case, where the number of groups can be decided in advance. This is partly since Company X only wants a certain number of customer groups. However, since the data used is very sparse and empty for many customers, this should also be considered. The segments will also be of different sizes, so the model must be able to handle them. The segments should also be as clear as possible and the best thing for Company x is that one customer belongs to only one group and also the algorithm should be simple to execute. Considering these, the best clustering method for this case is K-means, but it should be noted that it is not a perfect method to use either and for future research it is advisable to test different methods.

Chosen proximity measures are Euclidean distance, Manhattan distance, or Cosine similarity. K-means clustering is done with using each of these and the results of clustering are analyzed to determinate which of these is the best one to use in this case. K-means clustering is also robust to different distance metrics, other algorithms may behave differently.

5.5 Data Clustering

Company X wants to use clustering on the data to find as accurate customer groups from Customer Group 2 as possible. This is done by testing different cluster numbers to determine what number of new groups from Customer Group 2 gives us the best results. The Customer Group 2 group is big, has customers different from each other and has no distinct qualities other than the area of forest being small. Company X needs a way to make handling the group easier. This could be 2 groups where one is more active than the other. This will give us more info on how to start connecting with this heterogenous customer group.

Clustering is done using K-means algorithm with Euclidean distance, Manhattan distance, or Cosine similarity as proximity measures. This was chosen because based on the literature review it stands out for several compelling reasons. Firstly, its simplicity and versatility make K-means highly accessible and applicable across various domains, including customer segmentation. Its straightforward approach aligns well with the diverse nature of customer data that Company X deals with and will be easy it implement in the future if Company X want to use it with the other customer groups. K-means also offers a practical solution for determining the optimal number of clusters. Company X can choose what number of clusters it wants to test with this method. Techniques like the elbow method assist in identifying the ideal cluster count, ensuring that Company X can define specific segmentation targets, but we can choose this based on the needs of the Company.

Efficiency is another key advantage of K-means clustering. It is computationally efficient and scales linearly, making it well-suited for processing large dataset. The test sample data is small, but if the method is deemed good, it could be used with the bigger dataset that includes all of Company X's customers. The results produced by K-means are also clear and easy to analyze. This is important because if this way of making the groups is deemed good in Company X, there are many different people on different areas of business that might use this. Each customer is assigned to a single cluster based on similarity to centroids, simplifying the interpretation of segmentation outcomes. K-means exhibits flexibility in handling sparse data. It can accommodate varying cluster sizes and

incomplete customer information, ensuring that segments can be formed even with uneven distribution or missing data points that this data has. Some customers might have empty variables in the data. The advantages of simplicity, efficiency, clarity, flexibility, and proven effectiveness make K-means clustering the ideal choice for Company X's customer segmentation needs.

6 Results of clustering

The data was divided into clusters with K-means algorithm. Clustering was done 2-7 clusters, and we will go through to in order from 2 to 7. Silhouette score, Davies–Bouldin index and Calinski-Harabasz index were used to validate clusters. The silhouette score is a popular metric used to evaluate the quality of clusters obtained from clustering algorithms. It shows how well-separated the clusters are and how compact the clusters are relative to their neighbouring clusters. The silhouette score ranges from -1 to 1. A score close to 1 indicates that the data points are well-clustered, with data points in the same cluster being closer to each other than to data points in other clusters. (Sinaga, Yang, 2020). Silhouette score being close to 0 indicates that the datapoints is close to the boundary of the clusters. Reason for using silhouette validation in this case is because of its broad usability since calculation of silhouette coefficient is possible with any distance metrics like in this case Euclidean distance which we are using (De Amorim, Henning, 2015).

Looking at table 4 displaying the Silhouette scores it can be seen that many of the scores are not that good. The silhouette scores indicate that the clustering quality may not be very high regardless of the number of clusters chosen or the distance metric used. When the Silhouette score is close to 0, like in most cases in this thesis, it means that none of the datapoints in this data are distinctly in the clusters that belongs in. Data points are close to the cluster boarders. The ideal silhouette score should be close to 1 after clustering so that the data points would be well-clustered.

Number of clusters	**Euclidean**	**Manhattan**	**Cosine**
2	0.542	0.771	0.152
3	0.085	0.634	0.137
4	0.076	0.221	0.120
5	0.055	0.166	0.112
6	0.083	0.173	0.106
7	0.062	0.140	0.099

Table 4: Silhouette Scores

Average Silhouette score for all of the different cluster numbers is 0.151 for Euclidean distance, 0.351 for Manhattan distance and 0.121 for Cosine similarity. This can be considered quite bad in the scale of -1 to 1. Even though none of the scores are negative, being close to 0 is not good. The best scores can be seen when looking at K-means done with Manhattan Distance. Highest silhouette score is 0.771 with clustering done with 2 clusters and the second highest silhouette score is with clustering done with 3 clusters. These can be considered good. Using Manhattan distance, there is a notable decrease in silhouette score from 3 to 4 clusters. However, the silhouette scores remain inconsistent as the number of clusters increases, indicating that the clusters may not be clearly delineated regardless of the number of clusters chosen. Two clusters made with Euclidean distance have the silhouette score of 0.542. The drop when adding additional cluster to this is very significant.

All other silhouette scores are under 0.5 and the worst one being five cluster made with Euclidean distance. Euclidean distance in general shows the worst silhouette scores and should not be included in further analysis. Cosine distance also shows to be not good proximity measure for this type of clustering using this data. Analysis will focus on clusters made with Manhattan distance.

Overall, the silhouette scores suggest that further exploration and refinement of the clustering approach may be necessary to improve the separation between clusters and enhance the overall clustering quality. Trying different clustering method would be useful in this case.

Davies–Bouldin index in table 5 shows that the best number of clusters is in the minimum value of that index. Davies–Bouldin index is the ratio between the within-cluster distances and the between-cluster distances and computing the average over all clusters. Manhattan distance shows to be the best one with this index as well because the minimum value 0.582 indicating strong cluster cohesion and separation can be found there with 2 clusters. Second smallest is with 3 clusters made with Manhattan distance. The index raises quite lot after that. Cosine distance seem to be the worst one because in all cluster numbers it yields the index to be more than 2.

Number of clusters	**Euclidean**	**Manhattan**	**Cosine**
2	1.765	0.582	2.174
3	2.82	0.638	2.053
4	2.41	1.284	2.053
5	2.59	1.446	2.254
6	1.66	1.316	2.136
7	2.23	1.493	2.231

Table 5: Davies–Bouldin index

Calinski-Harabasz index in table 6 shows the best number of clusters to be the values where the index is the maximum indicating better-defined and more separable clusters. It measures the quality of clustering solutions based on the ratio of between-cluster dispersion to within-cluster dispersion. 2 clusters consistently yields the highest Calinski-Harabasz index values across all distance metrics considered. The highest values can also in this metric be found in Manhattan distance with 2 or 3 cluster and dropping after that one. Euclidean distance shows the lowest values of all in each clustering case and Cosine distance fails to match the performance of Manhattan distance.

Number of clusters	**Euclidean**	**Manhattan**	**Cosine**
2	41.602	368.983	67.182
3	31.517	337.293	57.197
4	27.481	283.852	48.245
5	25.626	240.935	41.860
6	24.337	219.120	37.690
7	24.421	198.508	34.674

Table 5: Calinski-Harabasz index

In conclusion to cluster evaluating, the Silhouette, Davies–Bouldin, and Calinski-Harabasz scores indicate that some of the clusters are good, but there might be better clustering methods for this data because most of the clustering don’t yield that good results. Specifically, Manhattan distance-based clustering configurations show promise and

warrant further investigation, while alternatives to Euclidean and Cosine distance metrics should not be investigated more in this thesis.

When we look at the pie charts in figures 7-12 displaying how the data is divided between the different cluster numbers focus will be on K-means with Manhattan distance because it displays the best scores and indexes in evaluation of the clusters. Focus is also on two clusters in figure 7 and three clusters in figure 8 because they show to be the best number of clusters for this data.

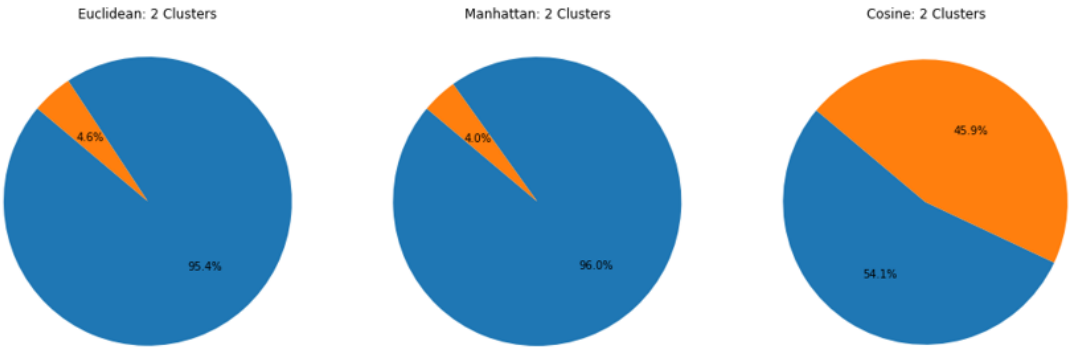

Figure 7 Distributions of data points with different proximity measures using 2 clusters

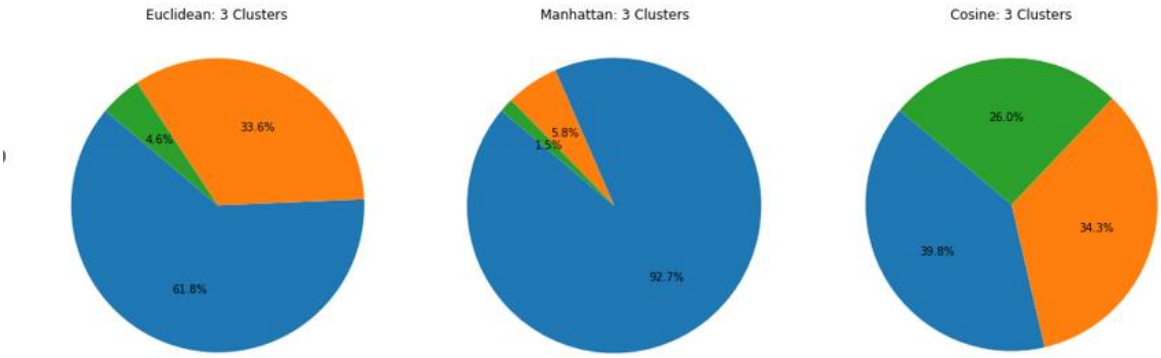

Figure 8 Distributions of data points with different proximity measures using 3 clusters

With two clusters in figure 7 made with Manhattan distance 4% of data point belong in cluster 2.2 and 96% belong in cluster 2.1. Also in three cluster made with Manhattan

distance in figure 8, 5.8% belong in cluster 3.3, 1.5% belong in cluster 2.2 and 92.7% belong in cluster 2.1. Majority of datapoints in both cases belong in one cluster. When looking at two clusters made with the other 2 proximity metrics Cosine similarity shows the best division of data points with 45.9% belonging in cluster 2.2 and 54.1% belonging in cluster 2.1. Euclidean distance is similar to Manhattan in cluster sizes. With three cluster both Euclidean and Cosine similarity display better distributed datapoints. In Euclidean distance 33.6% belong in cluster 3.3, 4.6% belong in cluster 2.2 and 61.8% belong in cluster 2.1. With Cosine distance this is even better because data is almost divided in thirds. 34.5% belong in cluster 3.3, 26.0% in cluster 2.2 and 39.8% belong in cluster 2.1. It is good to also examine these because the data divide much more evenly to these clusters from the viewpoint of Company X. Looking at Silhouette score, Davies-Bouldin index and Calinski_Harabasz index for these cases it can be concluded that in these metrics these are worse than the cluster done with Manhattan distance, but when choosing one to look into more, Silhouette score and Davis-Bouldin index suggests Euclidean with 2 clusters and Calinski-Harabasz suggests two clusters made with Cosine distance.

Next two clusters made with Manhattan distance and Euclidean distance and three clusters made with Manhattan distance and Cosine distance are analysed closer. Looking into the clusters and the characteristics of the data inside them will help determinate the best way to cluster this data. This is done by examining the summary statistics of each variable in the data and seeing if there is differences between the clusters.

It is worth to note that when looking ate figures 9-12 in appendix displaying the clustering done with 4-7 clusters in initialization, the distributions of the datapoints looks better than with smaller number of clusters, but the clustering evaluation results are not good. These bigger cluster numbers give worse Silhouette score, Davies-Bouldin index and Calinski_Harabasz index, but might indicate that whit some other clustering method or different segmentation approach it would be possible to find different types of groups in this data.

6.1.1 Two clusters made with Manhattan distance

Analysing how the data divides in these different clustering methods by comparing the summary statistics of the clusters. Summary statistics display the mean, standard deviation, median, count of values, minimum and maximum of the values. Comparisons in the clusters can be made by looking into these statistics and seeing what they show about the customers inside the clusters. Focus is on mean and median values because these give the best average person of the groups made with clustering. First analysis is done to two clusters made with Manhattan distance and after that to two clusters with Euclidean distance.

314 datapoints belong in cluster 2.1 and only 13 datapoints belong in cluster 2.2 with Manhattan distance. When looking at the summary statistics of different variables that affect the clustering cluster 2.2 has done much more sales of forest land than cluster 2.1. When looking both mean and median is clear that cluster 2.2 has customers that have sold their forest land more, differences are thousands. Also, cluster 2.2 is more active in different types of sales. Looking also at different services bought it is clear in this also that cluster 2.2 is the more active one, with means being higher in all the different bought services listed. Cluster 2.2 has also more area than cluster 2.1 when looking at means and medians, mean area of cluster 2.2 is 87.6 and cluster 2.1 is 47.8 and medians are 35 and 33. It is clear that in cluster 2.2 there are some customers that might not belong in Customer Group 2 all together and should be moved into another segment because the mean for cluster 2.2 is larger than the limit for Customer Group 2. This tells us that one of the conclusion of this thesis is, the customer segmentation in Company X has to be updated in some way. People in cluster 2.1 are more likely to live on their estate than people in cluster 2.2 and people in cluster 1 have historically in 1980-1990 done more sales than people in cluster 2.2. When taking into account the forest land this might affect the current sales and why they are so low. Cluster 2.2 has more potential in possible sales in the future with mean calculating this being 47.6 and median being 49. In cluster 2.1 these are 29.2 and 28. In services procurement cluster 2.2 is also more potential with mean of 64.1 and median of 63 compared to cluster 2.1's 36.6 and 37. People in cluster 2.1 might be more potential customers for Company X because they have sold their forest land more to competitors than cluster 2.2, but both can be seen as containing customers that Company X should try

to activate to do sales with them. Customers in cluster 2.2 are younger with mean age being 54 and in cluster 2.1 this being 65. People in cluster 2.2 are more likely to live in cities and further away than people in cluster 2.1. Both clusters have owned the forest land almost the same time and over 20 years, but cluster 2 little bit less. Cluster 2.1 uses less internet services than cluster 2.2 and does less sales through web. People in cluster 2.2 have done much more web sales. Cluster 2.2 is more likely to manage their estate themselves and cluster 2.1 is more likely to need help with management. Both clusters have sold forest land in the past 5 years but all 13 customers in cluster 2.2 have sold. For cluster 2.2 the sale of the forest land is more significant compared to other income than for cluster 2.1. For people in cluster 2.2 the investment product of Company X are more relevant than for people in cluster 2.1. Both clusters are interested in keeping their estate in the next 5 years. Based on the analysis done for the questions statements chapter 5.12 people in cluster 2.2 would be more likely to belong in Customer Group 4 or 5 and people in cluster 2.1 would belong in Customer group 3.5 or 4. When analysing the claims in questionnaire, cluster 2.2 displays interest in making money, taking care of their estate by themselves and taking as good care of the estate as possible and cluster 2.1 displays interest in making money, needing services, taking care of the estate by themselves and giving up their estate either by selling or as inheritance.

In conclusion the analysis shows a significant disparity between cluster 2.1 (314 data points) and cluster 2.2 (13 data points). Cluster 2.2 demonstrates higher forest land sales, greater activity across services, and larger sales. Historical sales favours cluster 2.1. Looking ahead, cluster 2.2 holds more potential for future sales and service procurement. Demographically, cluster 2.2 customers are younger and more internet-savvy. Preferences for estate management differ, with cluster 2.2 leaning towards self-care. Analysis suggests cluster 2.2 fits better into Customer Groups 4 or 5, while cluster 2.1 aligns with 3.5 or 4.

6.1.2 Two clusters made with Euclidean distance

When using Euclidean distance 312 datapoints belong in cluster 2.1 and 15 datapoints belong in cluster 2.2. Silhouette score, Davies-Bouldin index and Calinski_Harabasz index were much worse than when using Manhattan distance. When looking at the summary

statistics of different variables that affect the clustering it is clear that also in this case cluster 2.2 has done much more sales of forest land than cluster 2.1. Both mean and median show that cluster 2.2 has customers that have sold their forest land in larger, differences are thousands in this one also. There are also clear difference in services. Customers in cluster 2.2 have bought more services than customers in cluster 2.1. The summary statistics of the variables follow same pattern as in clusters made with Manhattan distance. Sizes of the clusters are similar in both cases and there are no significant differences in the statistics. Differences can be seen in the areas of forest land owned. Cluster 2.1 has the mean area of 48.1 and median of 33 and cluster 2.2 has the mean of 78.8 and median of 27 which is smaller than in cluster 2.1. Some datapoints with larger areas are now in cluster 2.1. Other variables seem to be similar with cluster 2.2 being more potential in forest land sales and services procurement, but now cluster 2.2 is more potential to entice the sell the forest land to Company X. Ages are now for cluster 2.1 65 and 56 for cluster 2.2. Where the people live are more similar in this clustering case. Neither cluster is more likely to live in cities and live almost as far. Clusters have owned similar lengths of time the forest land with cluster 2.1 owning little bit longer. Cluster 2.2 is more likely to use internet services and do sales in the web. Importance of sales income and investment products are also more important for cluster 2.2 similarly than in Manhattan distance. Base on analysis of the statements people in cluster 2.1 are more likely to belong in Customer Group 3.5 or 4 and people in cluster 2.2 are more likely to belong in Customer Group 4. Analysing the statements shows that people in cluster 2.1 display interest in making money, buying services or needing help, doing maintenance themselves and also giving up their estate either by selling or as inheritance. In cluster 2.2 the customers are interests in making money, especially taking care of the estate by them selves and seeking services.

In conclusion, both clustering methods reveal a notable contrast between cluster 2.1 and cluster 2.2. Cluster 2.2 consistently shows higher forest land sales, increased service activity, and larger sales, mirroring the findings from the other method. Both analyses highlight cluster 2.2's potential for future sales and service procurement. Overall, these findings show that there are some customers that are in a need of updating the customer group within Company X to accommodate the diverse preferences and behaviours observed across clusters. Also both of these methods don’t distribute the clusters as evenly

as Company X would want. The issue of trying to manage and divide the huge Customer Group 2 with the large number of customers would not be well resolved, because if the rest of the customers follow the same formula as the sample, then more than 90% of the customers from this group would still only end up in a large cluster, which is not so useful for Company X. It is good to know that there are some definitely hugely different customers in this group that then become the small cluster. But after taking those customers out, it would be useful to use some other segmentation method.

6.1.3 Three clusters made with Manhattan distance

Clusters made with Manhattan distance show that there are 303 data points in cluster 2.1, only 5 in cluster 2.2 and 19 in cluster 2.3. When looking at services procured, cluster 2.1 clearly is the lowest and cluster 2.2 much higher than higher other two combined. With that cluster having only 5 customers in it shows that Customer Group 2 has some customers that should not belong in it. These 5 customers are very different from the others in this sample. Cluster 2.3 is also much higher is services bought than cluster 2.1 but not close to cluster 2.2. Differences in this one are also in thousands. Also the same pattern can be seen in forest land sales. Cluster 2 is the highest, then 2.3 and the least services procured are in cluster 2.1. Differences in cluster 2.2 and 2.3 are not as big as in services bought. Areas with cluster 2.1 and 2.3 are similar with means being 47.4 and 51.2 and medians being 31 and 29. Cluster 2.2 has interestingly mean area of 174 and median of 27. This cluster shows that there are clearly under 5 customers in this sample that definitely should be moved from Customer Group 2 to some other cluster. There might be some customer that have been shorted to group 2 in the first place because the area has been under 50, but now they have bought more forest land and the customer group does not change because it has not been undated. Cluster 2.3 has sold most historically in 1980-1990 and cluster 2.2 the least. Cluster 2.2 shows to be most potential in future forest land sales with the value being 51.2, then cluster 2.3 with 42.4 and the least potential is cluster 2.1 with value of 28.8. Cluster are similar in the potential of future service procurement with cluster 2.2 being most potential wit value of 62.8, cluster 2.3 with 54.3 and then cluster 2.1 with value of 36.2. Potentiality when trying to get the customers do more sales with Company X, cluster 2.3 is the most potential to entice, then cluster 2.2 and then cluster 2.31. Ages of cluster 2.1 and 2.3 are same 64 and cluster 2.2 is the youngest cluster

with age of 51. Cluster 2.2 lives most likely in cities and furthest away and cluster 2.3 in countryside and closet to the estate or in the estate. Cluster 2.1 has owned the longest their forest land and cluster 2.2 the least. Cluster 2 uses internet services the most and more frequently than in the previous clustering case. People in cluster 2.2 also do most online sales. There is no significant difference in use of internet services and sales in web with clusters 2.1 and 2.3. All clusters are interested in keeping their estate in the future, but cluster 2.3 is more interested than others to sell it in the future. Cluster 2.2 shows that customers in it are more interested to do maintenance themselves and people in cluster 2.1 are more interested to seek professional help with maintenance. All customers in cluster 2.2 have done forest land sales in the past 5 years. Cluster 2.3 is more active in this than cluster 2.1, but all clusters show that they have sold forest land. For cluster 2.2 the income from sales is more significant, then for cluster 2.3 and then cluster 2.1. Cluster 2.3 and 2.2 are quite similar in this variable. The investment products of Company X is most important for cluster 2.2, but all the cluster are quite similar in this and on the scales of the question are in the middle, so significance overall is not that important for the clusters. Questionnaire statement analysis shows that cluster 2.1 aligns most with Customer group 3.5 or 4, cluster 2.2 with Customer Group 4 and cluster 2.3 with Customer group 3.5 or 4. Analysing the statements shows that people in cluster 2.1 display interest in making money, services, maintenance done by themselves and giving the estate as inheritance. Cluster 2.2 displays interest in mainly doing maintenance themselves and making money and cluster 2.3 displays interest in making money, services, doing maintenance themselves and hobby related statements.

In conclusion clusters generated using Manhattan distance show 303 data points in cluster 2.1, 5 in cluster 2.2, and 19 in cluster 2.3. Cluster 2.2 stands out with higher services procured and forest land sales. Notably, its mean area is significantly larger, suggesting misclassifications. Historical sales favour cluster 2.3, while cluster 2.2 shows the most potential for future sales. Cluster 2.2 also leads in internet usage and online sales. Preferences for estate management vary, with cluster 2.3 more interested in selling estates. All clusters have engaged in forest land sales, with cluster 2.2 showing the most significant income. Analysis suggests cluster 2.1 aligns with Customer Groups 3.5 or 4, cluster 2.2 with Group 4, and cluster 2.3 with 3.5 or 4. Cluster 2.1 emphasizes money-making, services, and self-maintenance, while cluster 2.2 focuses mainly on self-maintenance and

profit. Cluster 2.3 shows interest in money-making, services, self-maintenance, and hobbies.

6.1.4 Three clusters made with Cosine similarity

Three clusters done with Cosine similarity show little bit more even clustering. There are 130 data points in cluster 2.1, 85 in cluster 2.2 and 112 in cluster 2.3. This has shown the most even division of the analyzed clusters. It is worth to note that for Company X more similarly sized clusters are easier to manage. A more even distribution is more useful, but it should be remembered that creating three clusters with Cosine distance gave a worse Silhouette score, Davies-Bouldin index and Calinski_Harabasz index than using Manhattan distance.

Looking at summary statistics cluster 2.3 of these is the most active buyer of services. However, the difference is no longer thousands, but hundreds compared to other clusters. The second most active is cluster 2.2, followed by cluster 2.1. The difference between these two is not as big as their difference to cluster 2.3. When looking at the different services bought there is not that much difference in cluster 2.2 and 2.1 and they have bought even number of different services. Cluster 2.3 is also the most active in forest land sales, this follows the same pattern as in previous clustering cases, where one cluster is active in both. Cluster 2.2 is the next active and cluster 2.1 the least. Cluster 2.3 has also the most area with mean being 71.2 and median being 44.5. Cluster 2.2 has the mean of 42.6 and median of 35- Cluster 2.3 has the least area with mean of 35.5 and median 24.5. Cluster 2.1 is the most likely to live in cities and furthest away. Cluster 2.2 lives closest to the estate or in the estate and in countryside. Cluster 2.3 lives also in cities and further away from estate. Cluster 2.2 has done the most sales historically in 1980-1990, then cluster 2.3 and the least cluster 2.1. Cluster 2.3 shows to have most potential for future forest land sales with the value of showing that being 44.5. It also shows the most potential for future service seeking with value of 54.1. With cluster 2.1 and 2.2, cluster 1 shows more potential on forest land sales, but lest in service seeking and cluster 2.2 show least potential in forest land sales and more in service seeking than cluster 2.1. Values for cluster 2.1 are 23.6 for forest land sales and 28.9 for service seeking and the same values

for cluster 2.2 are 20,1 and 29.4. Values for cluster 2.1 and 2.2 are quite similar. They also show to be more potential customers to entice for Company X because they might have done forest land sales for competitors. From the three clusters, cluster 2.3 is the youngest with average age being 59. Cluster 2.1 is in the middle with average age of 65 and cluster 2.2 is the oldest with average age of 70. People on cluster 2.2 have owned the forest land the longest, then cluster 2.1 and last cluster 2.3, but all of them have owned similar lengths of time. Similarly, as in previous clustering cases the most active cluster, here cluster 2.3 has used the most of internet services and done the most online sales compared to the other two. Cluster 2.1 and 2.2 are similar in this one also and not as active as cluster 2.3. All of the cluster are aiming to keep their estate in the next 5 years and all of the cluster want to do maintenance themselves but might also require professional help. Almost all datapoints in cluster 2.3 have done forest land sales in the last 5 years. Cluster 2.2 is also similar in this one with almost all and the least active in forest land sales is cluster 2.1. The income from forest land sales is most important financially for cluster 2.2, then cluster 2.3 and last cluster 2.1. Cluster 2.3 values certain services of Company X the most. Cluster 2.1 clearly values the least and cluster 2.2 is in the middle. If looking at the questionnaire analysis, cluster 2.1 would belong in Customer Group 3.5, cluster 2.2 in Customer Group 3.5 or 4 and cluster 2.3 in Customer Group 4. Analysing the statements shows that customers in cluster 2.1 display interest in making money, buying more services, maintenance done by themselves. Cluster 2.2 displays interest in mainly in making money and doing maintenance themselves, with some service seeking. Cluster 2.3 majorly in making money, more than the other two, but also similarly in doing maintenance and buying services.

In conclusion on clusters formed with Cosine similarity exhibit a more balanced distribution with 130 in cluster 2.1, 85 in cluster 2.2, and 112 in cluster 2.3. While even distribution is what Company X is it yields poorer clustering scores than Manhattan distance. Summary statistics show cluster 2.3 as the most active in-service procurement, with smaller differences compared to other clusters. Cluster 2.3 also leads in forest land sales and area owned. Cluster 2.2 has the highest historical sales, but cluster 2.3 shows the most potential for future sales and service seeking. All clusters aim to maintain their estates, with varying degrees of professional help. Most datapoints in clusters 2.3 and 2.2 have engaged in recent forest land sales. Income from forest land sales is most significant for cluster 2.2, while cluster 2.3 values certain services the most. Analysis suggests

differing customer group alignments, with cluster 2.1 focusing on money-making and service procurement, cluster 2.2 on profit and self-maintenance, and cluster 2.3 on profit with similar interests in maintenance and service purchasing.

7 Conclusions

This thesis's purpose was to analyse Company X's customer base and existing segmentations, questioning their accuracy and relevance. By leveraging customer data and analysis techniques, the aim is to determine if customers belong in their assigned segments or should be reclassified and if we can find more accurate grouping inside customers groups that Company X is not content with. Further analysis seeks to explore whether segments can be subdivided based on distinct variables. This involved deploying questionnaire to delve deeper into customer characteristics of one particular group. Ultimately, the goal is to enhance marketing efforts, maximize profits, and adapt to changing market dynamics by getting know to the customers better and then make better marketing decisions.

For Company X, improving customer segmentation is integral to ongoing marketing initiatives. Recognizing the relevance of analytics, business development, and marketing, this thesis aimed to get Company X more involved with current global trends in leveraging data for business optimization. The research employed both quantitative and qualitative methods to gather and analyse data, providing comprehensive insights into customer segmentation and customers behaviour in relevant areas for Company X.

First in this thesis relevant literature was examined to make informative decision if this type of segmentation is possible. What way to do it? Best practice was concluded to be K-means clustering for the data that was used in this thesis. The decision to use K-means clustering was informed by its alignment with the thesis's objectives, previous successful applications in relevant research, and its ability to meet the specific requirements and challenges posed by Company X's customer data.

7.1 Answering the research questions

This master thesis had three research questions that Company X needed answers to:

1. *Are the customer groups in a need of an update and can data be used to divide customer into segments?*

When looking at the current situation of the customer segments in Company X it can be concluded that they are not as sufficient as they could be. At least Customer Group 2 needs an update, and data can indeed be used to more accurately divide customers into segments because both questionnaire and clustering found some customers that have similarities with other groups. It would be recommended for Company X to also check if the other groups need updating. There is a high chance that this is the case. Clustering analysis shows that there are differentiating factors in the data when analyzing the sample of customer group 2. It's essential to acknowledge that while the existing customer segments may have provided valuable insights in the past, market dynamics and customer behaviors evolve over time. Therefore, updating the segmentation approach is necessary to ensure that Company X's marketing strategies remain relevant and effective. Also, the data itself is in a need of a update because when analyzing the answer from the questionnaire that were the same as the Company X has used before, it was seen that the answers of some of the customers have changed in these years since they have joined.

By leveraging data analytics techniques and collecting additional data, such as through the questionnaire mentioned, Company X can gain deeper insights into customer preferences, behaviors, and needs. Question 13, where the customers were asked to choose different statements that describe them, was found to be good in Company X. This allows Company X to get information from the customers that they did not have before. Data-driven approach enables a more nuanced understanding of the customer base, allowing for the identification of new segmentation criteria or the refinement of existing ones. This approach also follows the trends in the business world that aims for companies to utilize their data in business decisions.

However, it's important to note that updating the customer segments does not necessarily mean discarding the old groups entirely. Instead, it involves revisiting and revising them based on new insights and data. By retaining some elements of the old segments while incorporating fresh data-driven insights and clustering, Company X can ensure continuity while adapting to changing market conditions. Company X could keep the current groups as they are, because the clustering for the sample group did not give as good results as it was hoped, but it could use the findings from the questionnaire and clustering in marketing, service offerings and customer service. Although the clustering did not give as good results as Company X would have liked, it still found a few customers who clearly do not belong to Customer Group 2.

Clustering done to the sample showed that dividing the data into similar number of groups that Company X already has did not give good groups and remained similar as the current situation that is aimed to better. It was seen that smaller number of clusters divided the data better into groups when examining Silhouette score, Davies-Bouldin index and Calinski_Harabasz index but when looking at the distribution of the data into new clusters it could be seen that with bigger number of groups the distribution looks better at first, but the scores and indexes were worse.

In summary, based on the questionnaire answers and results gained from clustering, updating the customer segments is crucial for staying competitive and relevant in the marketplace. Utilizing data to more accurately divide customers into segments enables Company X to tailor its marketing efforts more effectively, ultimately driving better customer engagement and satisfaction. Using K-means clustering can be concluded not to be the most efficient way to segment these customers and other methods, shall they be clustering methods or different segmentation methods should be considered.

2. *Can Customer Group 2 be divided further into groups that would provide valuable insights into the customers?*

Based on the literature review it was hypothesized that K-means clustering would be the optimal method for segmenting Company X's customers. Using K-means using proximity

measure Euclidean distance, Manhattan distance and Cosine similarity was not the most suitable for this data. Although results were obtained and the Silhouette score, Davies-Bouldin index and Calinski_Harabasz index gave somewhat good results for 2 clusters and 3 clusters with the Manhattan distance, the results were not as desired by Company X because the clusters made with these were very unevenly sized. When analyzing the clustering results, similarities were found within the clusters so the algorithm was able to find customers that were similar, but the cluster sizes were not what was wanted. These results can be utilized by moving the few customers that clearly do not belong in Customer Group 2 anymore to other groups. Company X would benefit more from groups where the data would be better distributed, than from groups where more than 90% of customers end up in the same group. However, on a positive note, the clustering found a small number of customers who probably no longer belong to Customer Group 2 at all because their forest land area has grown larger than the threshold that determined belonging in Customer Group 2. Also if looking only the results of the questionnaire answers there were more customers that could be moved from Customer Group 2. These people have more surface area now than before, and therefore should be placed in a new group, for example by re-doing the original customer survey for them or using the survey of this thesis.

It is also good to note that when examining each of the 4 different clustering cases, one more active cluster was found in the data. Although for the 2 clusters made with Euclidean and Manhattan distance, there were only a maximum of 15 people, so they were clearly different from the rest of the data. This was also visible in the case of three clusters. These people were united by a larger area, more active forest land sales, more active purchasing of services and a lower age. The data gained from questionnaire showed that they also used online services and made purchases online more than others. Forest land had been sold in the last 5 years and they are interested in making money and maintaining the forest lands themselves. This shows that there is potential clearly to find some differentiating factors in the data.

The data could not be effectively subdivided into smaller, more evenly distributed groups, despite identifying similarities within the clusters. Nonetheless, the analysis did identify a small number of customers who likely no longer belong to Customer Group 2 due to changes in their characteristics, such as increased surface area. Even with the clustering

method's limitations, both the questionnaire created for this thesis and clustering itself provided valuable insights into customer segmentation. Data obtained from the questionnaire, such as goals for the forest, interest in selling, keeping or giving away the forest, online activity, interest in doing maintenance work yourself or the need for professional help are useful information for Company X when handling its customers. While K-means clustering did not perform as well as hoped, it demonstrated little potential to refine customer segmentation strategies. Therefore, it is recommended to explore alternative clustering methods in future research endeavors.

3. *Can this method of clustering and collecting more data be used for the rest of the customer segments?*

Continuing with the consideration of whether this method of clustering and data collection can be applied to the rest of the customer segments, it's important to acknowledge the preliminary nature of the questionnaire used in this thesis. As mentioned, the questionnaire was piloted, indicating that certain questions may require better formatting or refinement, and some may need to be altered or omitted entirely in future surveys. Some of the statements in question 13 could be defined even more closely to represent original customer groups. It could be useful to find some way to merge the questionnaire data with the customer's previous data without having to ask for the customer ID. This clarity affected the answers of some customers, because either it was not filled in, it was wrong, or the question was answered incorrectly, for example with the customer’s name.

Despite these considerations, the questionnaire serves as a valuable starting point from which Company X can further develop its data collection efforts. This refinement process can enable Company X to gather more comprehensive and actionable insights from a broader range of customers within Customer Group 2, not just a sample. With some changes the questionnaire could also be broadened for the whole customer base of Company X. Some of the questions are so that they update current knowledge and some give more understanding even for groups where the customer type is more distinct.

Regarding the application of K-means clustering to other, more distinct customer groups were not gained. The groups formed with K-means clustering were not as useful as hoped,

even though with K-means with 2 clusters using Manhattan distance gave satisfactory results in Silhouette score, Davies-Bouldin index and Calinski_Harabasz index. The distribution of datapoint in cluster is not what was wanted. This method of clustering cannot be effectively used for the rest of the customer segments. The results obtained from this approach were unsatisfactory, yielding poor segmentation outcomes. Despite efforts to refine the process, including adjusting clustering algorithms and refining survey questions, the results remained subpar. It would be useful to try this method for bigger group than the sample or even the whole customer base to determinate if the method is suitable or should be discarded. Also, some other clustering method or even different segmentation approach is recommended for Company X to try to segment Customer Group 2.

7.2 Future research

After completing this thesis some potential future research topic for Company X on the topic of customer segmentation were found. It would be useful for Company X to delve deeper into the insights gained from the questionnaire utilized in this study by employing qualitative research methods. While this thesis focused primarily on quantitative data analysis, incorporating qualitative analysis might yield invaluable insights that may have been overlooked. Specifically, examining the responses to the open-ended question in the questionnaire could provide nuanced and contextual information that could inform decision-making processes for Company X. Open-ended question 14 was not examined in this thesis because it asked about joys and worries related to forest ownership and was not so relevant to the thesis topic itself. However, Company X wanted to add this question because it rarely surveys its customers and wanted to see if there are some worries it could focus on. This question mainly focused on receiving overall feedback for Company X, but there were some answers where one might find characteristics of the customer. Exploring the diverse topics covered in the open-ended responses could offer a deeper understanding of customer preferences, concerns, and motivations. By systematically analyzing these responses, Company X can gain actionable insights to tailor its strategies and offerings to better meet customer needs.

Additionally, if Company X intends to implement the customer questionnaire in the future, conducting further research on question formulation and refinement is recommended. Fine-tuning the questionnaire to ensure clarity, relevance, and effectiveness can enhance the quality of data collected and improve the validity of subsequent analyses. The questionnaire used in this thesis was the pilot version.

Furthermore, it would be interesting for Company X to explore alternative methods for customer segmentation beyond clustering, which was the primary focus of this thesis. Using and comparing different segmentation methods could provide insights into their respective strengths, weaknesses, and suitability for Company X's specific context. With different approach better segmentation result could be obtained.

Fuzzy C-means clustering could be good candidate to explore based on the cluster evaluations in this thesis. Fuzzy C-means is similar to K-means but assigns data points to clusters with varying degrees of membership rather than strictly assigning them to a single cluster. Hierarchical clustering could be a good alternative to try also. Hierarchical clustering does not require specifying the number of clusters beforehand and can reveal the hierarchical structure of the data, which may offer insights into natural groupings.

Refining the decision tree that is used in Company X with the questionnaire could be also a better alternative. The tree model used in analyzing the questionnaire gave more divided results than the original used in Company X and that could be tested for whole customer base to see what type of result might be gained from that. Also trying the original way of dividing customers but not giving as much weight on the area could give better results because the limit for area is strict, although the average customer does not meet the limit and then is forced in Customer Group 2

After finding a better segmentation approach shifting attention from Customer Group 2 to the entire customer base and extending the clustering process to include additional data points could also be useful for Company X. By broadening the scope of analysis, Company X can gain a more comprehensive understanding of its customer base and identify untapped opportunities for growth and optimization. It would be also interesting to collect

additional data with the questionnaire from all different customer groups of Company X. The conclusion presented that this way of working could be used for the whole customer base in sections ad that should be done next.

In summary, future research endeavors within Company X should encompass a multidimensional approach, incorporating qualitative analysis, questionnaire refinement, alternative segmentation methods, and a broader focus on the entire customer base. By embracing these avenues for exploration, Company X can position itself for continued success and relevance in an ever-evolving market landscape.

References

Ackermann, Marcel R. ; Blömer, Johannes ; Kuntze, Daniel ; Sohler, Christian. (2014) Analysis of Agglomerative Clustering. Algorithmica. [Online] 69 (1), 184–215.

Alves Gomes, M. & Meisen, T. (2023) A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information systems and e-business management. [Online] 21 (3), 527–570. https://link.springer.com/content/pdf/10.1007/s10257-023-00640-4.pdf

Andrienko, N. (2020) Visual Analytics for Data Scientists. 1st edition. Netherlands: Springer Nature.

Bakeman, R. & Robinson, B. F. (2005) Understanding Statistics in the Behavioral Sciences. 1st edition. [Online]. Mahwah: Psychology Press. [cited 20.4.2024] Available: https://ebookcentral.proquest.com/lib/lut/reader.action?docID=227516

Bergström, S. & Leppänen, A. (2021) Yrityksen asiakasmarkkinointi. 19., uudistettu painos. Helsinki: Edita.

Bose, T. K. (2012). Market segmentation and customer focus strategies and their contribution towards effective value chain management. International Journal of Marketing Studies, 4(3), 113.

Bouguettaya, A. et al. (2015) Efficient agglomerative hierarchical clustering. Expert systems with applications. [Online] 42 (5), 2785–2797. Referenced 02.06.2024 https://www-sciencedirect-com.ezproxy.cc.lut.fi/science/article/pii/S0957417414006150?via%3Dihub

Chang, Yaping ; Wang, Li ; Phillips, Mark R ; Wykoff, Charles C ; Garg, Sunir J ; Thabane, Lehana ; Bhandari, Mohit ; Chaudhary, Varun. (2023) What makes a good questionnaire? Eye (London). [Online] 37 (16), 3299–3301. https://www.nature.com/articles/s41433-023-02518-8.pdf cited 13.11.2023

Cohen-addad, Vincent ; Kanade, Varun ; Mallmann-trenn, Frederik ; Mathieu, Claire , (2019) Hierarchical Clustering: Objective Functions and Algorithms. Journal of the ACM. [Online] 66 (4), 1–42.

Company X, 2023, websites and internal information

Davino, Cristina., and Luigi. Fabbris. Questionnaire Data Collection and Integration. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. Web. https://link-springer-com.ezproxy.cc.lut.fi/book/10.1007/978-3-642-21308-3

De Amorim, R.C. and Hennig, C. (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors." Information Sciences, 324, pp. 126-145.

Dunteman, G. H. (George H. (1989) *Principal components analysis*. Newbury Park, [Calif.] ; SAGE. Referenced 02.06.24 https://methods-sagepub-com.ezproxy.cc.lut.fi/book/principal-components-analysis/n1.xml

European Comisson, 2023, Referenced [01.10.2023]

https://commission.europa.eu/law/law-topic/data-protection_en

Fojt, Martin. Marketing, Emerald Publishing Limited, 2005. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/lut/detail.action?docID=233894.

Gagolewski, Marek ; Bartoszuk, Maciej ; Cena, Anna (2021) Are cluster validity measures (in) valid? Information sciences. [Online] 581620–636. Cited 22.11.2023

https://ieeexplore-ieee-org.ezproxy.cc.lut.fi/xpl/ebooks/bookPdfWithBanner.jsp?fileName=5236612.pdf&bkn=5236612&pdfType=book

He, X., & Li, C. (2016). The research and application of customer segmentation on E-commerce websites. In 2016 6th International Conference on Digital Home (ICDH) (pp. 203–208). IEEE.

Hadad, Y. & Keren, B. (2023) A decision-making support system module for customer segmentation and ranking. Expert systems. [Online] 40 (2),

https://onlinelibrary-wiley-com.ezproxy.cc.lut.fi/doi/full/10.1111/exsy.13169

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). "Data clustering: a review." ACM Comput. Surv. 31(3), 264–323. https://doi.org/10.1145/331499.331504

Jaskowiak, P. A. et al. (2013) Proximity Measures for Clustering Gene Expression Microarray Data: A Validation Methodology and a Comparative Analysis. IEEE/ACM transactions on computational biology and bioinformatics. [Online] 10 (4), 845–857. Referenced 02.06.2024 https://ieeexplore-ieee-org.ezproxy.cc.lut.fi/stamp/stamp.jsp?tp=&arnumber=6461019

Jonker J.J., Piersma, N. and Van Den Poel, D., 2004. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert Systems with Applications, 27(2), pp. 159-168.

Jonker, Jedid-Jah ; Piersma, Nanda ; Van den Poel, Dirk (2004) Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert systems with applications. [Online] 27 (2), 159–168.

Karppinen, Heimo; Hänninen, Harri; Horne, Paula (2020). Suomalainen metsänomistaja 2020. Luonnonvara- ja biotalouden tutkimus 30/2020. http://urn.fi/URN:ISBN:978-952-326-961-3

Kotler, P. & Keller, K. L. (2009) *Marketing management*. 13th ed. Upper Saddle River (N.J.): Pearson/Prentice Hall.

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, December). Customer segmentation using K-means clustering. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 135-139). IEEE. https://ieeexplore.ieee.org/abstract/document/8769171

Klawonn, Frank ; Kruse, Rudolf ; Winkler, Roland. (2015) Fuzzy clustering: More than just fuzzification. Fuzzy sets and systems. [Online] 281272–279.

Mehta, Vivek ; Bawa, Seema ; Singh, Jasmeet (2020) Analytical review of clustering techniques and proximity measures. The Artificial intelligence review. [Online] 53 (8), 5995–6023.

Mori, Yūichi, kirjoittaja.; Kuroda, Masahiro, kirjoittaja.; Makino, Naomichi, (2016) Nonlinear Principal Component Analysis and Its Applications. [Online]. Singapore: Springer Singapore.

https://link-springer-com.ezproxy.cc.lut.fi/book/10.1007/978-981-10-0159-8

Nairn, A., & Berthon, P. (2003). Journal of Business Ethics, 42(1), 83–100. https://link.springer.com/article/10.1023/A:1021620825950

Omran, Engelbrecht, and Salman. (2007). An overview of clustering methods. Intell. Data Anal.. 11. 583-605. 10.3233/IDA-2007-11602.

Petrovic, Slobodan. "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters." Proceedings of the 11th Nordic workshop of secure IT systems. Vol. 2006. Citeseer, 2006. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b

Ros, F. et al. (2023) PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing (Amsterdam)*. [Online] 528178–199.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65. ISSN 0377-0427. https://doi.org/10.1016/0377-0427(87)90125-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

Sinaga, K. P. & Yang, M.-S. (2020) Unsupervised K-Means Clustering Algorithm. IEEE access. [Online] 880716–80727.

Smith WR (1956) Product Differentiation and Market Segmentation as Alternative Marketing Strategies. J Mark 21:3. https://journals.sagepub.com/doi/10.1177/002224295602100102

Takane, Yoshio. (2014) Constrained principal component analysis and related techniques. 1st edition. [Online]. Boca Raton: Chapman and Hall/CRC.

https://ebookcentral.proquest.com/lib/lut/reader.action?docID=1402688

Verdhan, V. (2020) Supervised learning with Python : concepts and practical implementation using Python. 1st ed. 2020. [Online]. Place of publication not identified: Apress. References 02.06.24 https://link-springer-com.ezproxy.cc.lut.fi/book/10.1007/978-1-4842-6156-9

Vorhies, D. W. et al. (2011) Improving customer-focused marketing capabilities and firm financial performance via marketing exploration and exploitation. Journal of the Academy of Marketing Science. [Online] 39 (5), 736–756.

Wang, C. (2022) Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. Information processing & management. [Online] 59 (6), 103085–.

Wang, X. & Xu, Y. (2019) An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP conference series. Materials Science and Engineering*. [Online] 569 (5), 52024-. Referenced 5.5.24 https://www.proquest.com/docview/2561126122?accountid=27292&parentSessionId=lF32Y4gmZipJp991k9J5wc7uNJrpFDALKc%2BBTSLwRp4%3D&pq-origsite=primo&sourcetype=Scholarly%20Journals

Xu, R. & Wunsch, D. C. (2009) Clustering. Oxford: Wiley.

Appendices

Customer Questionaire:

1. **Jäsennumero?**
2. **Ikä?**
3. **Asutko?**
   A) Maaseudulla
   B) Kaupungissa

4. **Kuinka kaukana asut metsästäsi?**

   A) Asun metsätilani läheisyydessä
   B) Asun alle 50 km päässä metsästäni
   C) Asun 50-200 km päässä metsästäni
   D) Asun yli 200 km päässä metsästäni
   E) Asun ulkomailla

5. **Kuinka pitkään olet omistanut metsää?**

A. Alle 5-vuotta
B. 5-20-vuotta
C. Yli 20 vuotta

6. **Kuinka usein käytät sähköisiä palveluita?**
   A) Useita kertoja kuukaudessa
   B) Kerran kuukaudessa
   C) Kerran puolessa vuodessa
   D) Kerran vuodessa
   E) Harvemmin
   F) En koskaan

7. **Oletko tehnyt sähköisen puukaupan?**

   A) En ole tehnyt sähköistä puukauppaa
   B) Olen tehnyt 1-3 kertaa
   C) Olen tehnyt yli 3 kertaa

8. Onko tarkoituksesi?

A) Omistaa metsää jatkossakin
B) Luopua metsänomistuksesta seuraavan 5 vuoden aikana siirtämällä metsätila saman suvun sisällä eteenpäin
C) Luopua metsänomistuksesta seuraavan 5 vuoden aikana siirtämällä metsätila myymäläl metsätila suvun ulkopuolelle
D) Hankkia lisää metsätiloja

9. Hoidatko aktiivisesti metsääsi, jotta se tuottaisi mahdollisimman hyvin?

A. Annan metsän ammattilaiselle hoidettavaksi
B. Kyllä, mutta tarvitsen apua
C. Kyllä, teen mahdollisimman paljon itse
D. En juurikaan hoida, pidän luonnontilassa

10. Oletko myynyt puuta viimeisen 5 vuoden aikana?

A. Kyllä
B. Ei

11. Tuoko metsäsi sinulle merkittävästi tuloja?

A. Kyllä
B. Ei

12. Kuinka tärkeitä Company X Palvelu Y on sinulle?

Ei tärkeitä 1- 5 Todella tärkeitä

13. Mitkä seuraavat tekijät kuvaavat minua metsänomistajana? Valitse 5

A. Omistan metsääni erityisesti taloudellisista syistä ja/tai osana muita sijoituksiani
B. Omistan metsäni erityisesti luontoarvojen takia
C. Omistan metsää, jotta voin hoitaa itse metsää ja/tai tehdä tarvittavat polttopuut
D. Omistan metsää retkeilyn ja/tai metsästysharrastuksen takia
E. Metsänomistaminen on minulla merkittävä elinkeino
F. Teen metsänhoitotöitä itse
G. Tarvitsen apua metsänhoitoon
H. Hyvä metsänhoito on minulle tärkeää
I. Metsäni monimuotoisuus on minulle tärkeää
J. Metsäni rooli hiilinieluna on minulle tärkeää
K. Olen valmis luovuttamaan metsätilani eteenpäin joko sukuni sisällä tai ulkopuoliselle
L. En koe suurta kiinnostusta metsäomaisuuteni hoitamiseen
M. Joku muu, mikä? (avautuu avonaiseksi kysymykseksi, missä voi selittää itse

14. Millaisia metsän omistukseen liittyviä iloja tai huolenaiheita sinulla on?

Figures

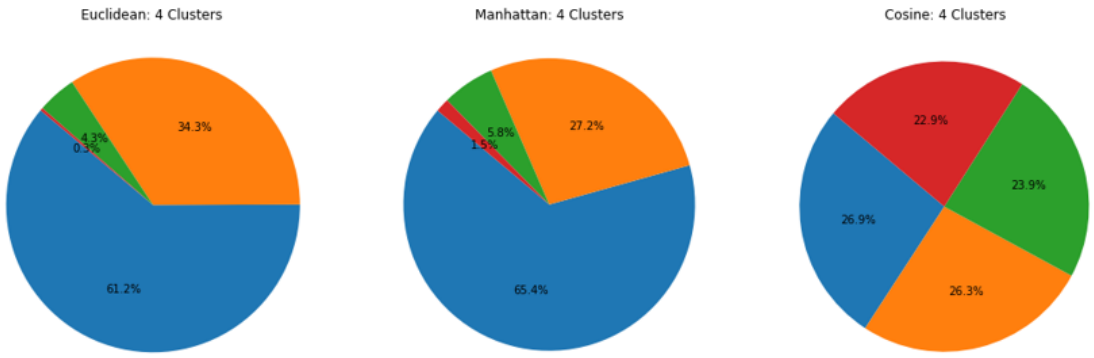

Figure 9 Distributions of data points with different proximity measures using 4 clusters

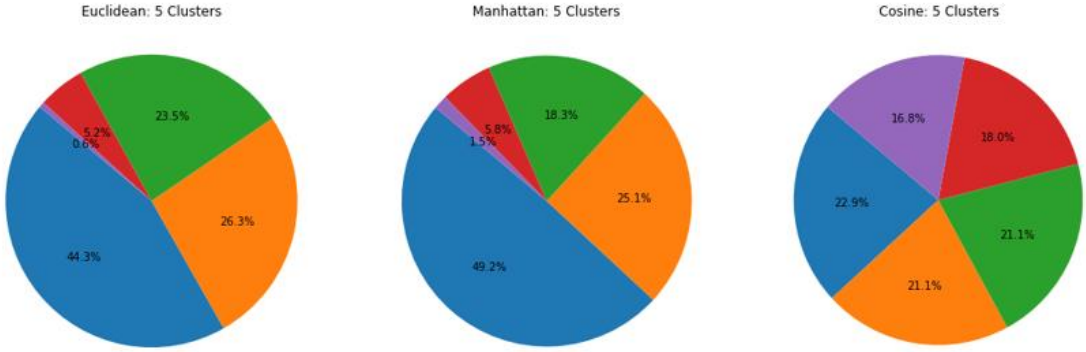

Figure 10 Distributions of data points with different proximity measures using 5 clusters

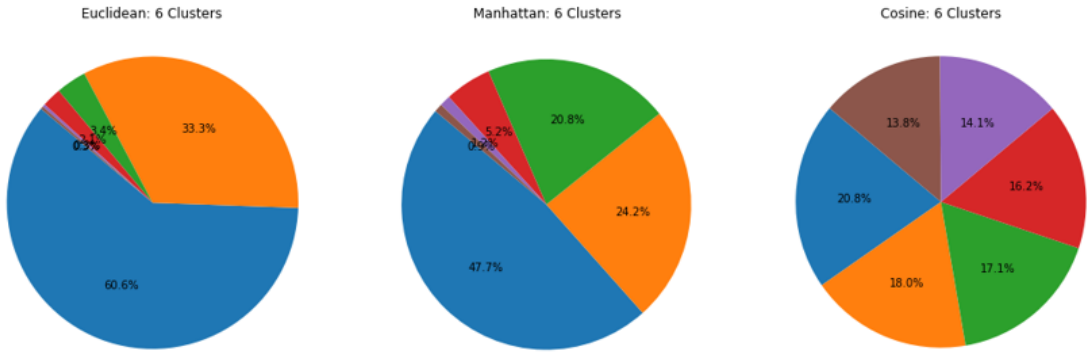

Figure 11 Distributions of data points with different proximity measures using 6 clusters

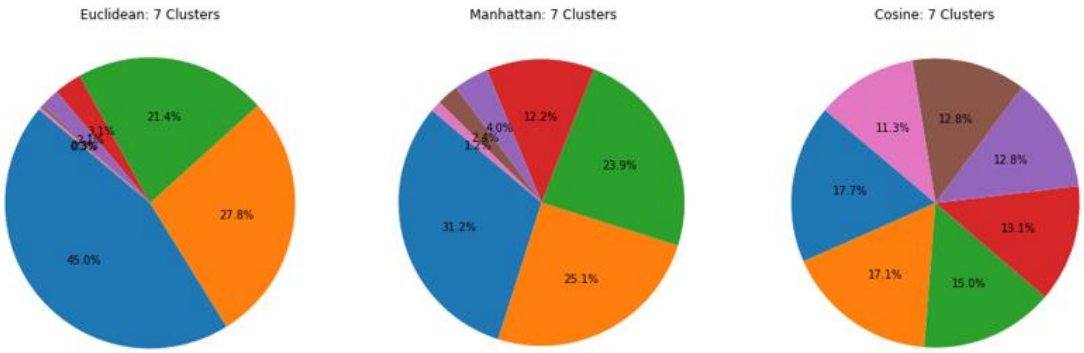

Figure 12 Distributions of data points with different proximity measures using 7 clusters

Tables

Table 3: Different Clustering Algorithms

What	K-means	DBSCAN	Fuzzy Clustering	Agglomerative Clustering
Approach	Partitioning	Density-based	Fuzzy	Hierarchical
Cluster Shape	Assumes spherical clusters	Can handle arbitrary shapes	Can handle arbitrary shapes	Can handle arbitrary shapes
Number of clusters	Must be specified in advance	Automatically determines based on density	Must be specified in advance	Can be determined based on dendrogram structure
Outliers	Sensitive to outliers	Robust to outliers	Sensitive to outliers	Robust to outliers
Cluster Size validity	Assumes clusters have similar size	Can handle clusters of varying sizes	Assumes clusters have similar size	Can handle clusters of varying sizes
Scalability	Sensitive to the number of dimensions and data size	Efficient for large datasets	Sensitive to the number of dimensions	Sensitive to the number of dimensions and data size
Interpretability	Clusters are well-defined and easily interpretable	May be harder to interpret due to varying cluster shapes	Fuzzy membership degrees allow for uncertainty	Hierarchical structure may provide insights

User Cases	Well-suited for compact and well-separated clusters	Suitable for clusters of arbitrary shapes and sizes	Suitable for scenarios with uncertainty	Suitable for scenarios with hierarchical structure
Data distribution	Assumes clusters are isotropic and equally sized	No assumptions about cluster shapes or sizes	Assumes data points may belong to multiple clusters	No assumptions about cluster shapes or sizes