MASTER THESIS

Discovering customer clusters using unsupervised machine learning to aid the marketing strategy: a case study with an online retail webshop SME.

Silke Reuvers

FACULTY OF BEHAVIOURAL, MANAGEMENT AND SOCIAL SCIENCES
BUSINESS ADMINISTRATION
UNIVERSITY OF TWENTE

MSc. Business Administration
Strategic Marketing Management – Marketing for Digital Business

Supervisors
Dr. E. Constantinides
Dr. A. Leszkiewicz

27-06-2021

UNIVERSITY OF TWENTE.

Acknowledgments

First of all, I would like to thank my supervisors dr. E. Constantinides and dr. A. Leszkiewicz for their guidance and valuable feedback during this research.

I would also like to thank my family and friends for their positive support. I am proud of this research!

Thank you,

Silke Reuvers

27th July 2021

Abstract

The rise of digital technologies has given businesses access to large amounts of data. However, the ability of (small) businesses to gain valuable marketing insights from this data is limited. The first aim of this study is to identify customer clusters using a clustering approach and algorithm (unsupervised machine learning) in data collected during online customer-business interactions. To do this, a literature review and a case study are conducted, using the Knowledge Discovery in Databases process as a research methodology. The data used in this study is collected from an online retail webshop (SME), that is retrieved from Google Analytics (1 Jan 2018 till 1 Jan 2021). The results of the literature review show that partitioning-based clustering, using k-prototype, is a suitable approach in analyzing a mixed and large dataset. Two different cluster analyzes have been conducted. The results of the first cluster analysis, based on individual customer data, show that two customer clusters have been found. The results of the second cluster analysis, based on grouped visitors data, show that two visitor clusters have been found. The second aim of this study is to develop a marketing strategy for the discovered clusters. This research provides marketing strategies for the discovered clusters, focusing on digital technologies. The marketing strategies for the discovered clusters include providing personalized products, offerings, and content using e-mail marketing campaigns, loyalty programs, and recommender engines. This research also argues that optimization strategies (SEO, SEA) can be developed based on the discovered clusters. The results of this study contribute to the growing field of using digital technologies in marketing. Furthermore, this study shows that SME online businesses with limited resources, can aid their marketing strategy by using a clustering approach in analyzing customer interaction data. This study is limited by the variety and volume of the dataset. The availability of the variety of the variables determines the completeness of the results. Furthermore, the validity of the results could be questioned because of the skewed data and the (low) silhouette score. Future research could assess how SME online businesses could gather and collect the right type of data in a profitable way for marketing purposes. Moreover, future research also could focus on supervised machine learning with labeled data for marketing purposes (e.g. classification, recommender engine).

Keywords: Clustering, Unsupervised machine learning, Customer segmentation, K-prototype, Marketing strategy, Digital technologies, Personalization, Digital marketing

Table of contents

1. INTRODUCTION

Businesses have nowadays access to large amounts of data, because of the development of new technologies. Artificial Intelligence (AI) plays an important part in these new technologies and is transforming the marketing. According to Davenport, Guha, Grewal, and Bressgott (2019), AI is changing marketing and customer behavior. Because of the combination of big data and technology, firms can gain more insights into customer behavior. More complete and actual data about consumers and their devices are available (Kumar, Ramachandran, & Kumar, 2021). If businesses have a full understanding of their customers, they can develop personalized offerings and content (Kumar et al., 2021). AI has the ability to deliver personalized content, which is the main factor behind its popularity (Kumar, Rajan, Venkatesan, & Lecinski, 2019). Customer segmentation is a cornerstone for these personalized offerings. In order to create customer segments, customers need to be segmented according to their similar behavior.

Segmentation plays a vital role in personalized marketing as a technique that divides customers or other units into groups based on shared attributes for example geographic or behavioral data (Umuhoza, Ntirushwamaboko, Awuah, & Birir, 2020). Each segment includes customers who have similar attributes (Pascal, Ozuomba, & Kalu, 2015). According to Pascal et al. (2015), businesses can provide personalization marketing programs that will be most suitable for each customer segment. This will in the end create value for the customer and reduce marketing costs (Umuhoza et al., 2020). Clustering is a method in segmentation. Clustering is an unsupervised machine learning approach and has proven to be efficient in finding patterns and relationships within unlabeled datasets (Pascal et al., 2015). Each cluster must contain data points that are similar but differ significantly from other data points in other clusters (Pascal et al., 2015). Without an output variable in advance, the algorithm is able to find clusters (McKinsey & Company, 2020). Machine learning is a subset of AI and can identify behavioral patterns in clustering that humans cannot compute (Pascal et al., 2015). Because of the rise of big data, machine learning is used by businesses to analyze data. Ma and Sun (2020) argue that machine learning methods are flexible and excel in prediction with regard to traditional models used in marketing. However, the success of the use of machine learning is limited by the volume and characteristics of the data, and the capability of businesses to create meaningful insights from the customer data (Kumar et al., 2019)

How to deal with this big data and machine learning algorithms remains challenging. Businesses are confronted with a large amount of data, but how to extract meaningful behavioral patterns from this data continues to be a challenge (Trusov, Ma, & Jamal, 2016). A proper understanding of various measurement approaches on how to extract customer data can create value in the decision-making of businesses and their marketing strategy (Marketing Science Institute, 2020). It is vital for businesses to know how to incorporate new technologies in order to guide the marketing strategy, because of the shift of marketing strategies towards personalization and customer engagement (Gupta, Leszkiewicz, Kumar, Bijmolt, & Potapov, 2020). Businesses acknowledge the importance and possibility of new-age technologies, despite the uncertainty in implementation and related expenses (Gupta et al., 2020). Marketing departments face the difficulty to select a suitable machine learning algorithm (Kuiper, 2018). Fahad et al. (2014) state that when choosing a particular algorithm the Volume, Variety, and Velocity of the dataset needs to be considered. In order to overcome this difficulty, Kuiper (2018) developed a framework of algorithms (unsupervised machine learning) regarding various datasets, sizes, and dimensionalities. Kuiper (2018) used complete linkage and K-modes algorithms to cluster university students. Based on these results personalized marketing campaigns can be created for more effective university recruiting (Kuiper, 2018).

In recent news, Google has announced that it will stop using third-party cookies (Google, 2021a). The main purpose of cookies is to collect data from customers and then process this data to improve the online experience (Stoltenkamp, 2021). Most of the time the third-party cookies are used, because the first-party cookies do not provide enough information about the customers to achieve the maximum relevant results for a marketing campaign (Stoltenkamp, 2021). First-party cookies collect data from one website, whereas third-party cookies collect data from external websites to get a better picture of the whole customer journey (Stoltenkamp, 2021). Retargeting a customer from another website will not be available anymore. The capability of analyzing first-party cookies data will play a significant role in marketing in the future, because of the ban of third-party cookies by Google.

This research will follow a case-study design using secondary data from an online retail webshop (SME) that sells umbrellas online to Dutch and Belgian customers. The data is collected during online customer-business interactions using first-party cookies. Two different cluster analyzes will be conducted using different datasets. The first dataset consists of individual customer data. The second dataset consists of grouped visitor data. The data is retrieved from Google Analytics. The dataset is unlabeled (data without labels identifying for example the characteristics of the data), therefore unsupervised machine learning will be used to cluster the data. This dataset contains the following challenges regarding the variety, volume, and velocity: (1) Mix type dataset. The data consists of numerical and categorical data. (2) Large dataset. The dataset consists of 9000 observations and over 20 variables. (3) Variable selection. Selecting the right variables for the marketing strategy. This research tries to prove that small online businesses with limited resources, can aid their marketing strategy by using unsupervised machine learning approaches.

The first objective of this study is to identify customer clusters using a clustering approach and algorithm (unsupervised machine learning) in data collected during online customer-business interactions. The second objective is to develop a marketing strategy for the discovered clusters. The research questions are as follows:

1. ''To what extent can online customer clusters of an online webshop SME be identified with the use of clustering (unsupervised machine learning)?''

Sub-questions:

- *How can online customer clusters be identified?*
- *What clustering approach and algorithm are suitable for the data set?*
- *What are the characteristics of the discovered clusters?*

2. ''What is an appropriate marketing strategy for the discovered clusters?''

2. THEORETICAL FRAMEWORK

2.1 Customer and marketing segmentation

Customer segmentation is in marketing a topic that is studied extensively. Customer segmentation (i.e. marketing segmentation) was first introduced by Smith (1956) and broadly refers to a group of customers based on similar features. The basis of segmentation lies in the heterogeneity of customers' needs, and for a more precise satisfaction of the various needs (Smith, 1956). The traditional segmentation studies used characteristics, for example, geographic, demographic, and psychological to segment customers (Mo, Kiang, Zou, & Li, 2010; Chen et al., 2018; Hung et al., 2019). Because of the recent development of big data, customers can be segmented according to their behavioral data. Behavioral data includes, for example, which items have been purchased, page visits, etc. (Chen et al., 2018).

Customers can also be segmented according to models that are used in marketing. Customer segmentation models are made that classify the customer based on standard and selected variables (Wu & Lin, 2005). The recency, frequency, and monetary value (RFM) is a model that is applied in segmentation studies. The RFM model is frequently used in marketing and customer relationship management (CRM) to cluster customers. According to Pater, Vari-Kakas, Poszet, and Pintea (2019), the RFM model is useful for customer segmentation and a vital model in marketing analysis. The RFM model identifies different customer profiles that can be used in personalized marketing campaigns and services. The future behavior and needs of the customer can be predicted to make personalized offers (Pater et al., 2019). '' The 'R' indicates the latest purchase amount, the 'F' indicates the total number of purchases during a specific period, and the 'M' indicates the monetary value spent during one specific period. '' (Walters & Bekker, 2017, p. 116). Another model that is used in marketing to segment customers is the customer lifetime value (CLV) model. Customers are reviewed as an asset of a firm (Kumar, 2018). When measuring the customer value to a business, the value in the future periods plays a vital role (Kumar, 2018). The future value is conceptualized in the CLV metric. Hwang, Jung, and Suh (2004) define the CLV model as '' as the sum of the revenues gained from company's customers over the lifetime of transactions after the deduction of the total cost of attracting, selling, and servicing customers, taking into account the time value of money'' (p. 182). Gupta et al. (2006) define the CLV as the present value of each future transaction during the customer's lifetime relationship with the business. The customer's lifetime relationship with the business is the period in which the consumers staying as consumers (Hwang et al., 2004). To conclude, the RFM and CLV are vital models in marketing and are based on predefined variables for segmentation.

2.2 Customer segmentation and the marketing strategy

Segmenting customers plays a significant part in CRM and marketing. Businesses can create different strategies for each cluster which in the end maximize the value of the customers (Hung et al., 2019). Similarly, segmentation can be used by businesses to identify groups of customers that might react the same way to a certain marketing campaign (Umuhoza et al., 2020). In other words, segmentation helps by providing personalized marketing strategies for each discovered cluster. If businesses understand their customers they will be able to provide personalized marketing services (Pascal et al., 2015). Because customers are divided into different groups, businesses can tailor marketing and services (Wang, Wang, & Zhong 2020). Wang et al. (2020) state that customer value, loyalty, and satisfaction could be improved by customer segmentation. Marketing strategies are more personalized and suit different customer needs which result in more revenue (Wang et al., 2020). In conclusion, personalization is the main marketing strategy of customer segmentation.

2.2.1 Online behavioral targeting

A topic that gained attention in personalized marketing is Online Behavioral Targeting (OBT), also known as online behavioral advertising and online profiling. OBT is ''the practice of monitoring people’s online behavior and using the collected information to show people individually targeted advertisements’’ (Boerman, Kruikemeier, & Borgesius, 2017, p. 364). A user profile (i.e. customer segment) is a cornerstone of OBT. According to Trusov et al. (2016), a user profile is a collection of user’s activities and preferences based on online activity data. Online activity data include browsing data, searches, purchases, Click-Through Rate (CTR) to ads, etc. (Boerman et al., 2017). A cookie is able to collect these types of data. Lately, more advanced techniques are used; advertisements can be personally shown with advanced profiling and real-time bidding techniques (Varnali, 2021). In OBT research, Yan et al. (2009) show that the CTR can be improved by 670% when users are correctly segmented. The needs of the customers for personalized advertising and offerings have risen (Kumar et al., 2021). The customer is changing and is technology-oriented. The customer prefers quick, uncomplicated, and personalized offerings to their demands (Kumar et al., 2021). This will in the end create an engaging connection with the consumers (Kumar et al., 2019). However, OBT and their use of personal information and data are constrained by privacy concerns, which could negatively influence personalization (Boerman et al., 2017).

2.2.2 Precision marketing

Similarly, a topic that is related to OBT but did not gain much attention in the literature is the precision marketing strategy. Marketing is changing, because of the advancement in AI, big data, machine learning, and data mining. The focus of the precision marketing strategy is to study subgroups of customers that are suitable for different marketing activities, using data mining techniques (Wang et al., 2020). ‘’Precision marketing is based on a database, which processes customer data in accordance with certain standards and selects the people in need for marketing’’ (Wang et al., 2020, p. 208). Precision marketing consists of providing the right products, through the right content and channels to the right consumers (Yin & Pan, 2020). Whereas, in traditional marketing, the focus is on targeting the total market and not precisely targeting subgroups. Wang et al. (2020) state that the traditional model had changed to a more personalized and is more small-scale. Precision marketing is based on the following features (Yin & Pan, 2020) (1) Creating a complete database. (2) Identifying different customer needs through segmenting customers using data mining. (3) Creating different strategies according to the segments (4) Offering different products and services according to the segments (5) Further explore and understand the needs of the customer using data mining (Yin & Pan, 2020). So in other words, precision marketing uses data mining to create different customer segments that are suitable for various personalized marketing strategies.

Wang et al. (2020) provide three marketing strategies based on precision marketing. The first marketing strategy is providing personalized product offerings. This strategy refers to that personalized products meet the preferences of the customers. The second marketing strategy is personalized price setting. This strategy refers to that the price strategy meets the preferences of the customers and product price elasticity. The third marketing strategy is providing accurate marketing information. This strategy refers to that the target customer and the information/content should meet the preferences of the customer (Wang et al., 2020). Because these marketing strategies can be implemented, the marketing strategy can be improved. According to Wang et al. (2020), precision marketing has some features which improve the marketing strategy (1) Communicate effectively with customers which results in customer trust and long-term cooperation. (2) Segmenting customers in order to suggest targeted marketing campaigns (3) Achieve personalized and effective marketing based on these segments (4) Reduce marketing costs because of the marketing planning and personalization.

Segmenting customers helps in developing personalized marketing strategies. Each segment includes customers who have similar attributes. Because of the advancements in big data and machine learning businesses are now capable of identifying behavioral patterns that humans cannot compute. The traditional market analyzes are regularly inefficient because of the amount of data that is now generated by customers (Pascal et al., 2015). Because of precision marketing, businesses are able to segment customers based on big data. Data mining plays a major part in the precision marketing strategy (Wang et al., 2020). You et al., (2015) propose a decision-making framework, using data mining techniques for precision marketing. Different types of data mining techniques are discussed that influence the precision marketing strategy: prediction, cluster, and classifying models.

2.3 Knowledge Discovery in Databases and datamining

In order to segment a customer base, a technique has to be implemented to discover patterns in raw databases created by customer interactions or customer data. The Knowledge Discovery in Databases (KDD) is such a technique. KDD ''refers to the overall process of discovering useful knowledge from data'' (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p. 39). It is important the know the differences between KDD and data mining. Data mining refers to '' the application of specific algorithms for extracting patterns from data'' (Fayyad et al., 1996, p. 39). Data mining techniques have the ability to extract hidden predictive patterns from big data (Rygielski, Wang, & Yen, 2002). As a result, businesses can identify important customers, can make informed decisions, and can predict behaviors. Data mining is a stage in the KDD method and is focused on the application of a specific algorithm. One big difference between data mining and the KDD is that data mining cannot find patterns without validation (Rygielski et al., 2002). In other words, data mining helps to develop hypotheses, however without validating the hypothesis. To conclude, the KDD method can discover patterns in raw databases, in which data mining is a step in this process.

2.3.1. Fundamentals of Knowledge Discovery

The steps in the KDD process are as follows (Fayyad et al., 1996). (1) Understanding the application area and set goals. (2) Selecting a (sub)set of data. (3) Pre-processing and cleaning data including removing meaningless data, missing data, and taking into account unforeseen changes. (4) Transform the data (5) Matching the goals to a specific data-mining method such as classification or clustering (6) Choosing algorithms (e.g. for categorical vs numerical data) and methods. (7) Data mining: finding patterns in the data including for example classification trees or clustering. (8) Interpreting the extracted patterns, involves also visualization (9) Acting on the discovered results (Fayyad et al., 1996). Furthermore, Fayyad et al. (1996) state that the KDD process is iteration and contains loops between different steps. More recently, various methods are introduced to discover patterns in raw databases, for example, SEMMA[1] and CRISP-DM[2]. Wang et al. (2020) provide six steps in the data mining process based on the CRISP-DM: business understanding, data understanding, data preparation, modeling, evaluation, and application. The SEMMA and CRISP-DM are based on the KDD process. The KDD methodology is more precise and complete and is used frequently in studies compared to other methods (Shafique & Qaiser, 2014). Therefore, the KDD process will be used as a methodology in this study.

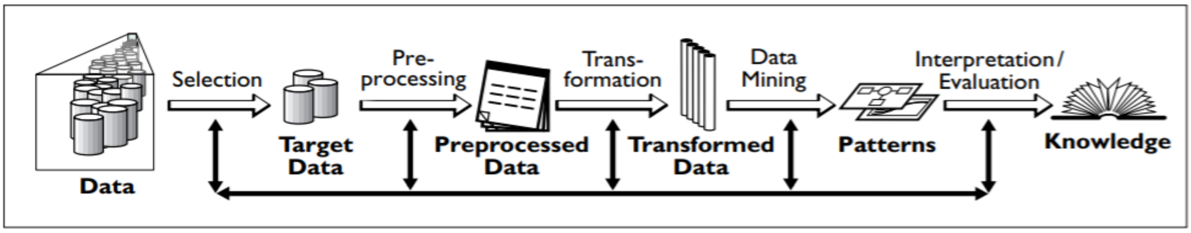


Figure 1. KDD process (Fayyad et al., 1996b)

[1] Sample, Explore, Modify, Model, Assess

[2] Cross-Industry Standard Process for Data Mining

2.3.1 Fundamentals of data mining

Data mining is a step in de KDD process and is concerned with the application of a specific algorithm. The steps in de KDD ensure that useful patterns and knowledge are obtained from the raw data (Fan & Li, 1998). Fan and Li (1998) state that the application of data mining methods with no knowledge beforehand can lead to meaningless patterns. The two goals of data mining are primarily prediction and description (Fayyad et al., 1996). Prediction includes predicting the unknown or future with the use of particular variables. Description includes finding understandable patterns that describe the data (Fayyad et al., 1996). A good choice of data-mining methods can achieve the goals of description and prediction. Examples are classification and clustering (Fayyad et al., 1996). In brief, classification tries to classify a data item into predefined classes. Clustering tries to find clusters that represent each other. To conclude, the two goals of data mining are prediction and description.

2.4 **Data-analytics: Artificial Intelligence and machine learning**

Data-mining methods are based on the technology of AI and more precisely on machine learning. ''AI operates in the domain of continuous learning and automation, acting as the intelligence that drives data-based analytics and enables automatic decision-making'' (Kumar et al., 2021, p. 868). AI can perform aspects of human intelligence (Huang & Rust, 2018). Kumar et al. (2019) argue that AI can discover and train machines to identify patterns in data, using machine learning algorithms. Machine learning is a subcategory in AI (Kumar et al., 2019). Machine learning can generally be separated into supervised and unsupervised machine learning. In supervised machine learning an algorithm uses training data and output from humans to learn the relationship of the given input variables and output variables, whereas in unsupervised machine learning an algorithm uses input data and explores relationships without an output variable (McKinsey & Company, 2020).

It is largely agreed among marketers that using data analytics is important to drive marketing campaigns (Gupta et al., 2020). The new approaches in these data analytics are not constrained to selected industries. Data analytics refer to '' the technology-enabled analyses of data and processes using new-age technologies (such as AI, machine learning (ML), internet of things (IoT), blockchain, drones, etc.) and other online and offline data sources to design and deliver continuous, one-on-one personalized engagement in real-time'' (Gupta et al., 2020, p. 27). Gupta et al. (2020) propose a framework for how businesses can unitize data analytics to drive consumer insights (appendix 1).

2.4.1. Supervised machine learning

Supervised machine learning concentrates on prediction. ''Supervised learning seeks to learn from this training dataset a function, Y = f(X), to predict the output when given an input'' (Ma & Sun, 2020, p.484). According to Ma and Sun (2020), the focus is more on discovering a function that maximizes the accuracy of the prediction, than finding patterns between variables. One way to achieve the maximum accuracy of the prediction is to use various testing datasets. Classification is a method in supervised machine learning. An example of a classification method is a target marketing application (Aggarwal, 2015). The goal of the application is to predict if a customer is interested in a particular product based on relating features to the class label. A training model will be developed to test and predict the class labels (Aggarwal, 2015).

2.4.2. Unsupervised machine learning

Unsupervised machine learning concentrates on discovery. ''In unsupervised learning tasks, the training dataset contains only the input variables, while the output variables are either undefined or unknown. The typical goal is to find hidden patterns in or extract information from the data'' (Ma & Sun, 2020, p.484). Clustering is a method in unsupervised machine learning. Clustering divides multiple clusters into similar patterns or relationships, maximize the similarity within clusters and minimalize the similarity between other clusters (Ma & Sun, 2020). Clustering in marketing includes discovering similar groups of customers in marketing databases (Fayyad et al., 1996). Personalized marketing strategies can be given to the different clustering groups.

This study is concerned with discovering patterns within an unlabeled dataset (output variable is unknown), therefore unsupervised machine learning clustering will be applied in this study.

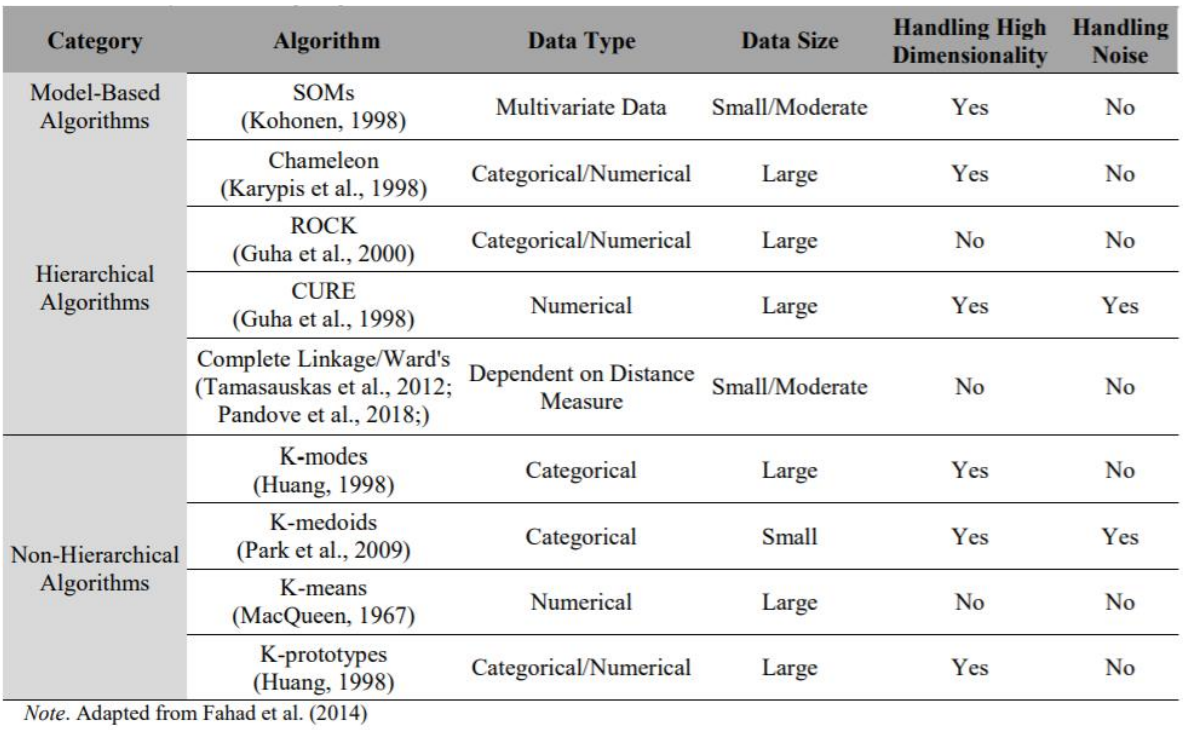
2.5 Clustering

Clustering is an effective tool in unsupervised machine learning to segment customers. Clustering is a method that can discover patterns or relationships within an unlabeled dataset (Kansal et al., 2018; Pascal et al., 2015). According to Wang et al. (2020) ‘’clustering is the process of grouping a collection of concrete or abstract object into multiple classes or clusters composed of similar object” (p. 209). With the use of mathematical methods, the objects are classified into clusters. The objects must exclusively belong to a cluster (Wang et al., 2020). Each cluster includes similar data points, but varies significantly from data points of other clusters (Pascal et al., 2015). When the clusters are made, the algorithm can find hidden patterns. K-means clustering is a commonly used algorithm in segmentation studies (Pascal et al., 2015; Wang et al., 2020; Umuhoza et al., 2020). The K-Means algorithm is faster, simpler, and has to ability to deal with big data sets with respect to other clustering methods (Umuhoza et al., 2020). Other algorithm approaches are used to cluster the customers. For example, Kansal et al. (2018) used a number of cluster algorithms in their research: k-means clustering, agglomerative clustering, and mean-shift clustering. To conclude, a number of algorithms can be used in clustering.

Clustering algorithms and techniques can be divided into different classes. A critical challenge for clustering algorithms is the absence of consensus in the properties and categorization (Fahad et al., 2014). Therefore, Fahad et al. (2014) provide a framework of clustering techniques with respect to large datasets. The essential clustering algorithms can be divided into (1) Partitioning-based, (2) Hierarchical-based, (3) Density-based, (4) Grid-based, (5) Model-based. In brief, the partitioning-based algorithm divides data objects into a number of groups, where each group has to have at least one object, and each object has to belong to one group. The K-means algorithm is a well-known algorithm in this category. The hierarchical-based algorithm builds a hierarchy of clusters. It can be divided into divisive (top-down) and agglomerative (bottom-up). The divisive algorithm begins with the data in one cluster and splits into more clusters down the hierarchy. The agglomerative algorithm begins with one cluster and merges two or more clusters up the hierarchy. The Chamelon, CURE, and ROCK are some well-known algorithms in this category. The density-based algorithm separates the data based ‘’on their regions of density, connectivity and boundary’’(Fahad et al., 2014, p. 268). This algorithm is capable of discovering clusters of arbitrary shapes. One advantage of the density-based algorithm is that it is robust to outliers. DBSCAN is a well-known algorithm in the category. The Grid-based algorithm divides the data object into grids. The advantage is that it can rapidly process the data. Wave-clustering and STING are well-known algorithms in this category. The model-based algorithm, also known as mixture models, assumes that data is created through a mixture of probability distributions. Each element represents diverse clusters (Fahad et al., 2014). SOM and Gaussian mix are some well-known algorithms in this category. Partitioning-based, Hierarchical-based, and Model-based are mainly used in segmentation studies (Fahad et al., 2014).

Selecting the suitable algorithm depends on the dataset. Kuiper (2018) noted that the type, size, and dimensionality of the data is reliant on choosing a particular algorithm for unsupervised machine learning. Similarly, Fahad et al. (2014) state that when choosing a particular algorithm the Volume, Variety, and Velocity of the dataset needs to be taken into account. Volume refers to the capability of the algorithm to handle a big dataset. Variety refers to the capability of the algorithm to handle various types of data (e.g. numerical, categorical). Velocity refers to the speed capacity. Kuiper (2018) developed a framework of algorithms (unsupervised machine learning) regarding various datasets, sizes, and dimensionalities (table 1). This framework is limited to Partitioning-based and Hierarchical-based clustering algorithms. For brevity, this study will follow the framework of Kuiper (2018).

Table 1. A framework of unsupervised machine learning algorithms (Kuiper, 2018)



Category	Algorithm	Data Type	Data Size	Handling High Dimensionality	Handling Noise
Model-Based Algorithms	SOMs (Kohonen, 1998)	Multivariate Data	Small/Moderate	Yes	No
Hierarchical Algorithms	Chameleon (Karypis et al., 1998)	Categorical/Numerical	Large	Yes	No
	ROCK (Guha et al., 2000)	Categorical/Numerical	Large	No	No
	CURE (Guha et al., 1998)	Numerical	Large	Yes	Yes
	Complete Linkage/Ward's (Tamasauskas et al., 2012; Pandove et al., 2018;)	Dependent on Distance Measure	Small/Moderate	No	No
Non-Hierarchical Algorithms	K-modes (Huang, 1998)	Categorical	Large	Yes	No
	K-medoids (Park et al., 2009)	Categorical	Small	Yes	Yes
	K-means (MacQueen, 1967)	Numerical	Large	No	No
	K-prototypes (Huang, 1998)	Categorical/Numerical	Large	Yes	No

Note. Adapted from Fahad et al. (2014)

Kuiper (2018) developed a two-stage clustering framework, where the first stage involves a hierarchical-based algorithm to decide the total number of clusters. One critical challenge in clustering is to determine the number of clusters (Fahad et al., 2014; Kuiper, 2018). For a mixed dataset, the hierarchical clustering algorithms SOM, Chameleon, and Rock are sufficient. The second step in the framework of Kuiper (2018) is to use a non-hierarchal clustering algorithm. As stated above, K-means clustering is used in a variety of segmentation studies. K-means algorithm is efficient and is able to deal with large datasets (Huang, 1998). However, the K-means algorithm only works on numerical values. Therefore, Huang (1998) introduced K-modes, an algorithm that can cluster categorical values. Most of the datasets in real-live consist of both numerical and categorical values (Fahad et al., 2014; Huang, 1998). Therefore, a hybrid approach is introduced, the K-prototype, a mix of the K-means and K-modes algorithm (Huang, 1998). This algorithm uses a Euclidean distance for a numeric variable, and a hamming distance for categorical variables (Huang, 1998). Since this study has a mixed dataset, the K-prototype will be used.

To conclude this study will follow the framework of Kuiper (2018) and will perform a two-stage clustering method. The first step is using hierarchical clustering. The algorithm SOM will be used. The second step is using a non-hierarchical clustering approach. The algorithm K-prototypes will be used.

Table 2. Clustering methods and algorithms for this study, based on the framework of Kuiper (2018).

Two-stage clustering approach based on a mixed dataset	Algorithms
First stage: hierarchical clustering	SOM
Second stage: non-hierarchical clustering (partitioning-based)	K-prototype

2.6 Related work

Pater et al. (2019) segmented the customers by using a k-means algorithm that is based on the RFM model. The goal of their paper is to prove that businesses with a small number of customers can use data-mining techniques to create customer profiles for personalized marketing. Pater et al. (2019) select the k-means algorithm because they state that it is a simple and usable method for segmenting customers. In order to get the optimal clusters, the distances between objects are calculated. This results in five clusters. A description of the discovered clusters is given however, a marketing strategy is missing.

Walters and Bekkkers (2017) performed two cluster analyzes, based on the RFM-model and k-means. The first small dataset consists of 100 customers from a grocery store. The second dataset consists of 100000 participation behavior of campers. The silhouette plot is applied to determine the total clusters. In total four and three clusters were found respectively. The discovered clusters are described however, a profound marketing strategy for these clusters is missing.

Umuhoza et al. (2020) segmented credit card customers on the basis of transactional data. The goal of their paper is to find patterns and extract customer purchasing behaviors that lead to different customer segments. These customer segments are used to suggest personalized marketing strategies. Umuhoza et al. (2020) selected the K-means algorithm, they argue that this algorithm is faster and can handle large datasets. Before applying K-means, the data had to be pre-processed and transformed. Umuhoza et al. (2020) used Principal Component Analysis (PCA) to reduce the dimensionality of the variables. To get the optimal clusters they use the methods: Silhouette, Elbow, and Calinksi-Harbaz. In total four clusters were identified that can aid the marketing strategy. According to Umuhoza et al. (2020), businesses can provide personalized marketing campaigns to these homogeneous clusters. A limited marketing strategy is given for the founded clusters.

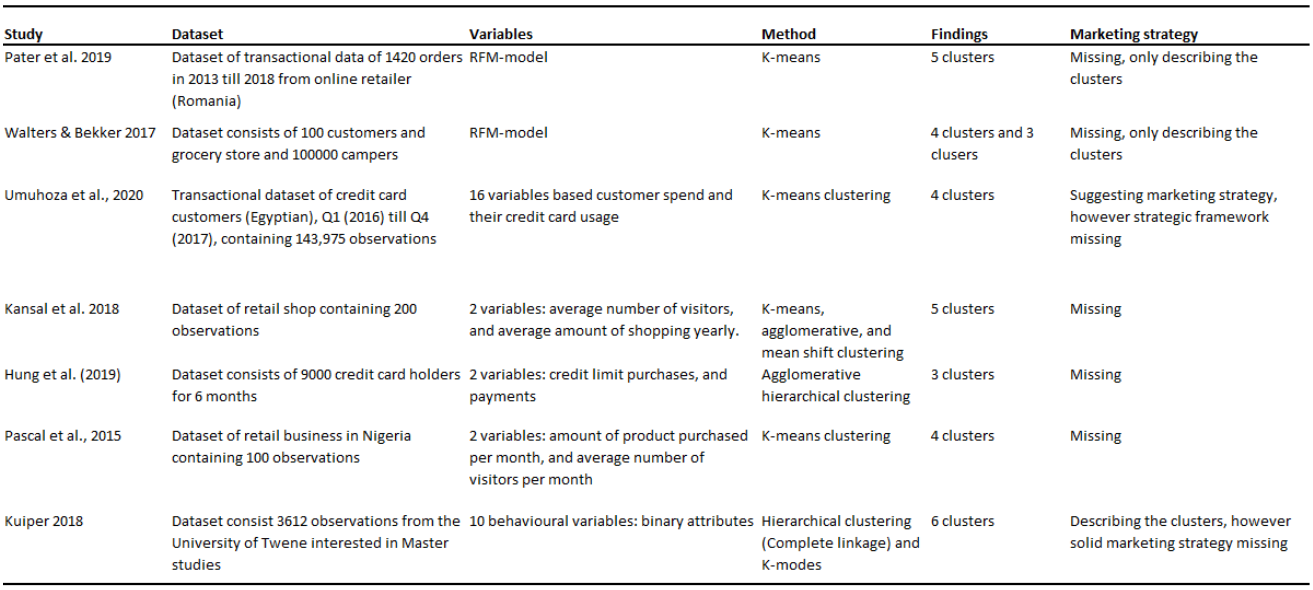
Kansal et al. (2018) segmented customers of a local retail shop based on two variables: average number of store visits and average number of purchases on an annual basis. In their study, they used three clustering algorithms: K-means clustering, agglomerative clustering, and mean-shift clustering. In order to get the optimal clustering, Kansal et al. (2018) used the methods: Elbow (for K-means clustering), Dendrogram (for agglomerative clustering), and Brandwidth (for mean-shift clustering). The results of the Silhouette score indicate that K-means and Agglomerative clustering performed well on the given dataset. In total five clusters were found however, a marketing strategy for these clusters is missing.

Hung et al. (2019) segmented credit card customers based on two variables: limited purchases and payments. Hung et al. (2019) argue that they have a small dataset, which consists of 9000 credit card customers with 18 variables over a period of six months. In their study, they use Hierarchical agglomerative clustering because they argue that this algorithm performs better on a small dataset than for example k-means clustering. A challenge in clustering methods is to evaluate if the dataset is clusterable. To overcome this challenge, Hung et al. (2019) performed the visual assessment of cluster tendency (VAT). In order to get the optimal clusters, Hung et al. (2019) used the methods: Elbow, Sillhouette, and Gap Statistic. In total three clusters were identified however, a proper marketing strategy for these clusters is missing.

Pascal et al. (2015) segmented customers of a retail business according to two variables: the number of product purchases for each month and the average number of visitors for each month. Before performing the k-means algorithm, they normalized the variables using a z-score normalization technique. In total, four clusters were found, however, a solid marketing strategy is missing

Kuiper (2018) segmented (potential) University master students based on various behavioral attributes that were performed on the website. Kuiper (2018) segmented the students based on a two-stage clustering approach. First, a hierarchical clustering method (complete linkage) is applied to determine the total clusters. Next, a non-hierarchical clustering method (k-modes) is used to segmented the students. In total, six clusters were found, in which two clusters were significant in terms of conversion.

Table 3. A Literature review of related work in clustering/segmentation studies.



Study	Dataset	Variables	Method	Findings	Marketing strategy
Pater et al. 2019	Dataset of transactional data of 1420 orders in 2013 till 2018 from online retailer (Romania)	RFM-model	K-means	5 clusters	Missing, only describing the clusters
Walters & Bekker 2017	Dataset consists of 100 customers and grocery store and 100000 campers	RFM-model	K-means	4 clusters and 3 clusers	Missing, only describing the clusters
Umuhoza et al., 2020	Transactional dataset of credit card customers (Egyptian), Q1 (2016) till Q4 (2017), containing 143,975 observations	16 variables based customer spend and their credit card usage	K-means clustering	4 clusters	Suggesting marketing strategy, however strategic framework missing
Kansal et al. 2018	Dataset of retail shop containing 200 observations	2 variables: average number of visitors, and average amount of shopping yearly.	K-means, agglomerative, and mean shift clustering	5 clusters	Missing
Hung et al. (2019)	Dataset consists of 9000 credit card holders for 6 months	2 variables: credit limit purchases, and payments	Agglomerative hierarchical clustering	3 clusters	Missing
Pascal et al., 2015	Dataset of retail business in Nigeria containing 100 observations	2 variables: amount of product purchased per month, and average number of visitors per month	K-means clustering	4 clusters	Missing
Kuiper 2018	Dataset consist 3612 observations from the University of Twene interested in Master studies	10 behavioural variables: binary attributes	Hierarchical clustering (Complete linkage) and K-modes	6 clusters	Describing the clusters, however solid marketing strategy missing

The study of Umuhoza et al. (2020) incorporated a marketing strategy according to the segmented clusters however, a profound marketing strategy is missing. The success of the use of machine learning is limited by businesses to create meaningful insights from the data (Kumar et al., 2019). Moreover, a proper understanding of generating meaningful insights creates value in the decision-making of business and the marketing strategy (Marketing Science Institute, 2020). First, this study will perform a two-stage clustering approach. Second, this study will develop a marketing strategy for the discovered clusters.

3. **METHODOLOGY**

The KDD process will be used as a research methodology. It consists of the following steps: Understanding the application process and set goals, Selecting (sub)set of data, Cleaning and pre-process data, Reducing data, Matching goals to data-mining method and algorithm, Transforming and scaling data. Furthermore, the Data Protection rule, the Google updates, and the Methodology choices will be discussed in this chapter.

3.1 Understanding the application process and set goals

This research will use secondary data from an online retail webshop (SME) that sells umbrellas online to Dutch and Belgian customers. The data is retrieved from Google Analytics. The data is collected from 1 Jan 2018 till 1 Jan 2021 and contains 9000 observations and over 20 variables. Two different cluster analyzes will be conducted using different datasets. The first dataset consists of individual customer data. The individual purchasing customers are tracked based on this data. The second dataset consists of grouped visitor data. One drawback of this dataset is that it is not possible to distinguish between a visitor who made a purchase and who did not (only the conversion rate is tracked). The data is collected from first-party cookies.

The online business has not a CRM system however, the back-end system of the webshop can be considered as a CRM system: data of the customers and their orders are stored there. The webshop runs on the software of CVV shop. When a customer makes a purchase, an (anonymous) transaction-ID is created. This transaction-ID is created through an API by the software of CVV shop that is connected with Google Analytics. All the transaction-IDs are stored in Google Analytics. Also, Google Analytics provides various data that is connected with the transaction-ID, for example, the revenue of the order, which device is used, country etc. (table 4). The first dataset consists of individual customer data. Furthermore, the website visitors are tracked by Google Analytics using first-party cookies. Google Analytics provides data that is connected with the visitors (table 5). The second dataset consists of grouped visitor data. The programs R and Excel will be used to analyze the data.

The goal of this KDD process is to create meaningful insights for the marketing strategy. Customer clusters will be created that can be used to guide the marketing strategy. One vital goal for the company overall is having a high-profit rate. For example, the company would rather sell fewer products with a relatively higher profit rate, than sell many products with a low-profit rate. Furthermore, the company wants to implement personalization strategies that are feasible for a small business (2-4 different clusters).

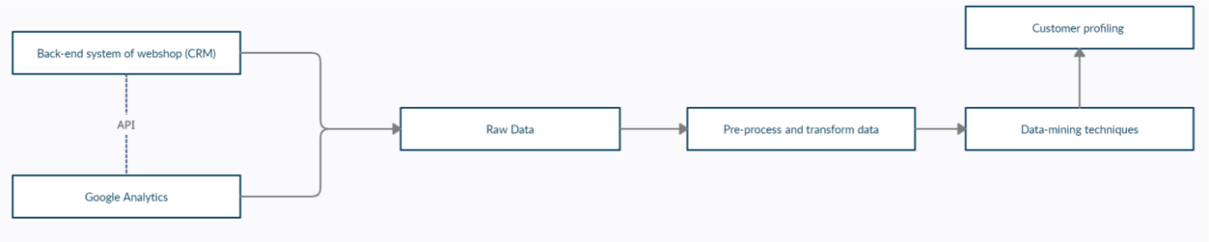


Figure 2. Illustration of the data analyzing process

3.2 Selecting (sub)set of data

To achieve the goals a set of data should be selected. As stated before, the data is extracted from Google Analytics and contains the data from the purchasing customers (individual data) and visitors (grouped data). Many variables are available that are connected with the transaction-ID and the visitors. First, the variables that contain a lot of missing values will not be used for the validity of this research. For example, the variable ‘keywords’ contains a lot of missing values. Second, variables that are not meaningful will not be used. For example, the variable ''screen resolution'' will not be used. Third, the capability of Google Analytics will be critically evaluated. For example, the variable ''gender'' cannot also be tracked by Google Analytics accurately. Fourth, when these variables are excluded, the variables will be selected that have the potential to aid the marketing strategy. This research focuses on a small webshop with limited resources. Marketing strategies could include: personalization strategies, paid search strategies (SEA – Google Adwords, Shopping), e-mail marketing strategies, organic traffic strategies (SEO).

Table 4. Description of variables in dataset 1 – Cluster analysis 1 (4622 observations, 7 variables retrieved from Google Analytics)

Variable	Comment	Data type	Data type in R (after cleaning)
Transaction-ID	A unique number to track the purchasing consumer	Numerical	Numerical
Revenue	The total amount of order	Numerical	Numerical
Number of products	Number of products	Numerical	Numerical
Traffic Channel	Direct, paid search, organic search, referral	Categorical	Factor
Devices	Desktop, mobile, tablet	Categorical	Factor
Country	Dutch, Belgium	Categorical	Factor
Type of visitor	Returning or new visitor	Categorical	Factor

Table 5. Description of variables in dataset – Cluster analysis 2 (360 observations, 11 variables retrieved from Google Analytics)

Variable	Comment	Data type	Data type in R (after cleaning)
Type of visitor	Returning or new visitor	Categorical	Factor
Traffic channel	Direct, paid search, organic search, referral	Categorical	Factor
Search function website	With or Without using the website's search function	Categorical	Factor
Country	The Netherlands, Belgium	Categorical	Factor
Amount of sessions	The total number of sessions. A session is a period in which the visitor is active on the website	Numerical	Numerical
Average session duration	Duration of one session in seconds	Numerical	Numerical
Number of visitors	Amount of visitors	Numerical	Numerical
Pages per session	Amount of pages that are visited during one session	Numerical	Numerical
Conversion rate	The percentage of sessions that has led to a transaction	Numerical	Numerical
Revenue per visitor	Average revenue per unique visitor	Numerical	Numerical
Bounce percentage	Percentage of sessions that only interact/viewed one page	Numerical	Numerical

3.3 Cleaning and pre-process data

The preparation of data is an important step to make valid analyzes and draw conclusions. Data preparation is a process of cleaning and transforming the data. The following procedures will be done during the data preparation stage: (1) data will be checked on missing values, and if necessary deleted. (2) data will be checked on outliers, and if necessary deleted. (3) data will be properly assigned to its data type (e.g. character, number, string) (4) The variable transaction-ID is removed, because this variable is meaningless for the clusters (5) observations that do not include the Netherlands or Belgium are deleted. Dataset 1 is reduced to 4545 observations of 6 variables. Dataset 2 is reduced to 207 observations of 11 variables.

Because cleaning and processing data is an important step in the KDD process, in Appendix 2 is a clear description of this process.

3.4 Matching goals to data-mining method and algorithm

This study will perform a two-stage clustering approach. The first step is a hierarchical-based approach to decide the number of clusters. In addition to that, the Elbow method and Silhouette score will be used to strengthen the choice for the number of clusters. For brevity, this study will use the Ward distance in the hierarchical clustering.

The second step is a partitioning-based clustering approach. The dataset consists of both numerical and categorical data, therefore the algorithm K-prototype is the most suitable algorithm (Huang, 1998). This algorithm uses a Euclidean distance for a numeric variable, and a hamming distance for categorical variables (Huang, 1998).

3.5 Transforming/scaling data

The first step is transforming the dataset for hierarchical clustering. Since the dataset contains both numerical and categorical variables, the variables have to be scaled. The function "Daisy " is used in R to scale the data. This function computes all the pairwise distances (dissimilarities) between the observations in data, resulting in a matrix.[3] The dissimilarity coefficient of Gower (1971) is used to handle the mixed dataset.

The second step is transforming the numerical variables for K-prototype. The categorical variables do not need any common scaling, because the algorithm can handle the categorical variables[4]. However, the numerical variables need to be commonly scaled (Saha, Tariq, Hadi, & Xiao, 2019). Without normalization, the variable "revenue" would dominate the "amount of products" variable. The Z-score is used to scale the data, shown in Equation (1)

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Where:

μ is the mean of the data
σ is the standard deviation of the data

[3] https://www.rdocumentation.org/packages/cluster/versions/2.1.0/topics/daisy

[4] https://www.rdocumentation.org/packages/clustMixType/versions/0.2-11/topics/kproto

3.6 Data protection rule and Google

3.6.1 Data protection rule in general

The General Data Protection Regulation (GDPR) is a rule that has entered in 2018 for all businesses in the European Union. This regulation gives protection to individuals concerning the use of personal data by businesses and the movement of personal data (European Commission, 2017). The GDPR strengthens individual rights online and gives companies more clarification/rules on how to deal with the privacy and data of customers. According to the European Commission (2017), the individual data must be in a transparent way, in which the company must explain why it uses the data, how it intends to use the data, and how long the data will be stored. With respect to the online business in this research, they inform their customer about their rights on the page “privacy”. Furthermore, first-party cookies are installed on the website. As the researcher can tell, the data that is used in this research meets the conditions of the GDPD.

3.6.2 Google cookies update

Google has announced that it will stop using third-party cookies (Google, 2021a). Cookies are small files that are used to identify a computer. Google (2021b) states that cookies make the online experience simpler for consumers. Because of cookies, businesses can keep track of the consumers during their website visits, can remember the website preferences, and can provide personalized and relevant content (Google, 2021b). Google (2012b) uses two types of cookies 1) First-party cookies. These cookies are generated by the domain/website that the customer is visiting. 2) Third-party cookies. These cookies are generated created by other websites, data is collecting from external websites. The capability of analyzing first-party cookies data will play a significant role in marketing in the future, because of the ban of third-party cookies by Google. The data that is used in this study is generated from first-party cookies.

3.7 Methodology choices

As stated in the literature, clustering is the most suitable method within machine learning for this study. Clustering is an effective tool in unsupervised machine learning to cluster customers, based on an unlabeled dataset. K-prototype will be used, because this algorithm can handle a large and mixed dataset with respect to other algorithms (Huang, 1998; Kuiper, 2018). Furthermore, this study will develop a suitable marketing strategy for the discovered clusters. For example, Kuiper (2018) used the AIDA model to describe each cluster and to recognize in what stage the visitor is during the customer funnel in order to provide more information about the customer behavior. This study does not use the AIDA model to identify the conversion funnel and customer behavior. This is because Google Analytics did not track vital behavioral attributes (e.g. viewed product pages, add to cart, create an account etc.) for every individual visitor. Therefore, it could not be recognized at which stage the visitor is during the customer funnel (based on the dataset in this study). In the first dataset (individual customer data) every customer completed the conversion funnel. In the second dataset (grouped visitor data) behavioral attributes are missing that could describe in which stage the visitor is during the customer funnel. This study will describe the clusters based on vital differences that could potentially guide the marketing strategy. The clusters will be characterized based on the mean and max value from the Z-score (numerical variables), and the percentages (categorical variables). In the last chapter, a marketing strategy will be given for the discovered clusters, focusing on how digital technologies can aid the marketing strategy.

4. **RESULTS**

The results chapter consists of the last three steps of the KDD process: Data mining, Results of K-prototype, and the Interpretation of patterns.

4.1 Datamining - Cluster analysis 1

4.1.1 Number of clusters

One of the challenges in clustering is to determine the number of clusters (Saha et al., 2019). Many clustering algorithms only calculate the grouping of data according to a preset number of clusters (Kodinariya & Makwana, 2013). Therefore, determining the number of clusters is an essential step in clustering. In this study, the Elbow method, Silhouette score, and Hierarchical clustering will be used to determine the number of clusters.

4.1.2 Elbow method

The elbow method is an empirical approach that requires minimum prior knowledge about the data (Saha et al., 2019). "Elbow method is a method which looks at the percentage of variance explained as a function of the number of clusters"(Bholowalia & Kumar, 2014, p. 18). The elbow method is derived from the total within-cluster sums of square (WSS). A line is drawn between the WSS and K (number of clusters) (Saha et al., 2019). At a certain value for the number of clusters the graph drops considerately (Kodinariya & Makwana, 2013). The bend in the graph, ''elbow criteria'', is considered as an indicator for the number of clusters. The idea behind this is that adding a new cluster does not give better modeling of the data (Bholowalia & Kumar, 2014).

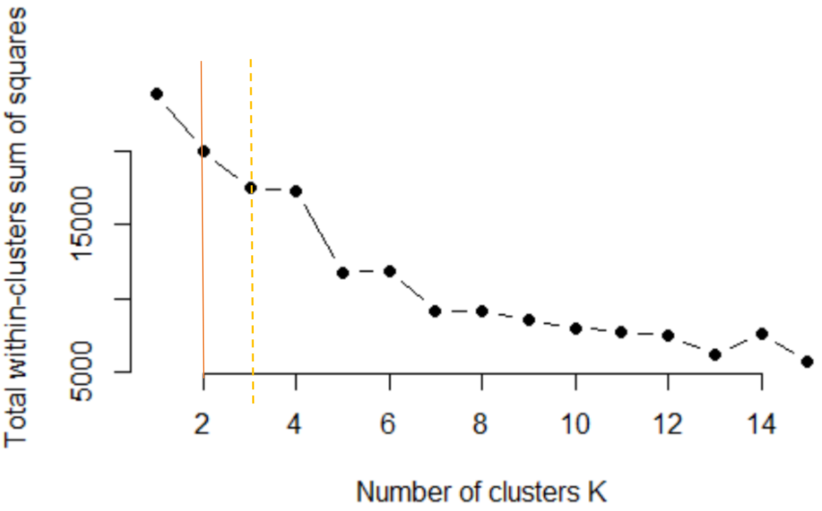


Figure 3. Elbow graph for cluster analysis 1.

The bend in the graph is not clearly shown. The number of clusters will be at 2 or 3. Because the number of clusters is not clearly shown, the silhouette score and hierarchical clustering will be used as well to determine the number of clusters

4.1.3 Silhouette score

The silhouette score can be used next to the Elbow method to determine the number of clusters. The silhouette score is based on the separation distance between the clusters (Ogbuabor & Ugwoke, 2018). The silhouette score is built on the mean score for each point in the data (Baarsch & Celebi, 2012). It measures the distance between each point in one cluster and other points in other clusters. The differences between points are then divided by the normalizing term (Baarsch & Celebi, 2012). The silhouette score values between -1 and 1 (Ogbuabor & Ugwoke, 2018). The value -1 indicates that the cluster is not assigned properly, whilst the value 1 indicates that the cluster is properly assigned. (Ogbuabor & Ugwoke, 2018). In other words, the higher the value of the Silhouette score, the better the cluster validity.

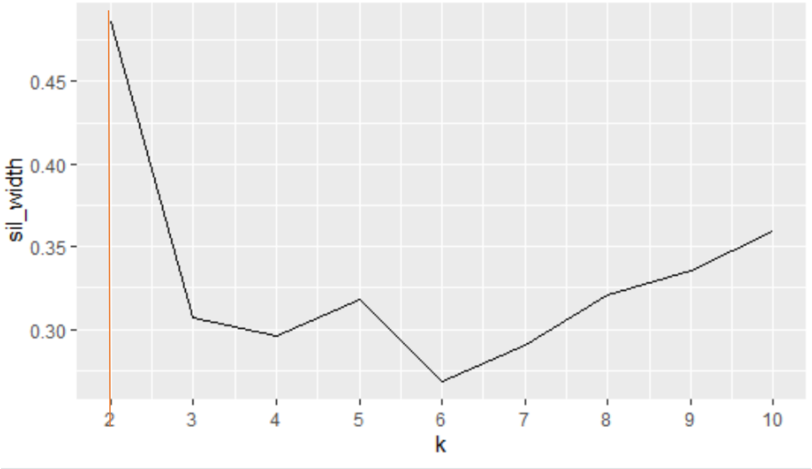


Figure 4. Silhouette score (sil_width) for the number of clusters (K) for cluster analysis 1.

The silhouette score of 0,49 is the highest at two clusters (figure 4).

4.1.4 Hierarchical clustering dendrogram

The dendrogram from hierarchical clustering is also used to determine the number of clusters. The dendrogram can be used for agglomerative clustering (Kansal et al., 2018). ''Dendrogram is the hierarchical representation of object, it is used to determine the output of the hierarchical clustering'' (Kansal et al., 2018, p. 136). To find the optimal number of clusters, the longest vertical line must not be cut by the horizontal lines (Kansal et al., 2018). When cutting the longest vertical line, the horizontal line provides the number of clusters.

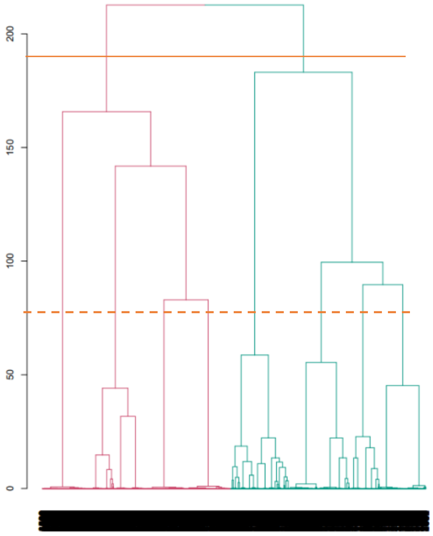


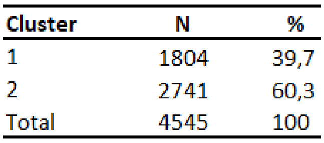
Figure 5. Hierarchical clustering dendrogram for cluster analysis 1.

When cutting the longest vertical line (dashed line), the number of clusters will be 8. However, for a small company providing 8 different marketing strategies for each cluster will not be profitable. Therefore for this analysis, the outcome of the silhouette score will dominate, which is at 2 clusters. The validity of this cluster analysis could be questioned, because the methods (Elbow, Silhouette score, and Hierarchical clustering dendrogram) do not show an equal number of clusters

4.2 Results of K-prototype - Cluster analysis 1

This section will discuss cluster analysis 1 using K-prototype (dataset 1). The number of clusters has been determined at 2.

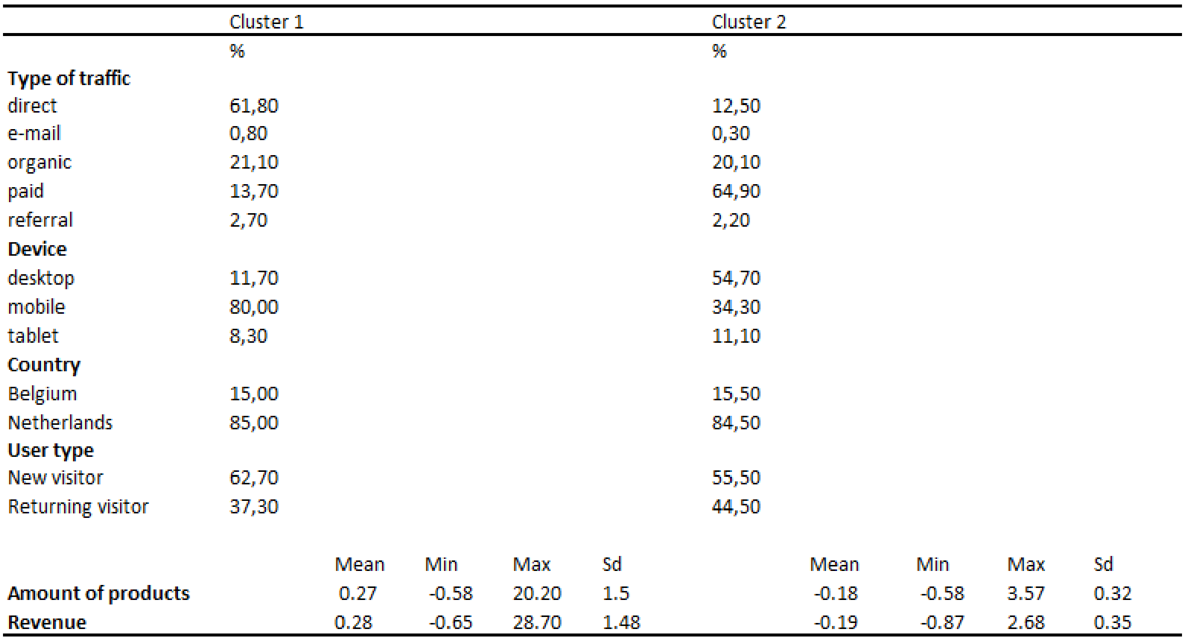
Table 6. Distribution of clusters



Cluster	N	%
1	1804	39,7
2	2741	60,3
Total	4545	100

Cluster 1 has 1804 (39,70%) customers and cluster 2 has 2741 (60.30%) customers, as seen in table 6.

Table 7. Cluster results of analysis 1



	Cluster 1					Cluster 2				
	%					%				
Type of traffic										
direct	61,80					12,50				
e-mail	0,80					0,30				
organic	21,10					20,10				
paid	13,70					64,90				
referral	2,70					2,20				
Device										
desktop	11,70					54,70				
mobile	80,00					34,30				
tablet	8,30					11,10				
Country										
Belgium	15,00					15,50				
Netherlands	85,00					84,50				
User type										
New visitor	62,70					55,50				
Returning visitor	37,30					44,50				
		Mean	Min	Max	Sd		Mean	Min	Max	Sd
Amount of products		0.27	-0.58	20.20	1.5		-0.18	-0.58	3.57	0.32
Revenue		0.28	-0.65	28.70	1.48		-0.19	-0.87	2.68	0.35

The customers in cluster 1 mainly come from direct traffic (61,80%), and organic traffic (21,10%). Use the mobile phone (80%) and mainly come from the Netherlands (85%). The visitor type is 62,70% new and 37,30% returning. The customers in cluster 1 purchase more products than customers in cluster 2 (mean: 0,27 and max: 20,20). This results that the revenue is higher for customers in cluster 1 with regard to customers in cluster 2 (mean: 0,28 and max: 28,70).

The customers in cluster 2 mainly come from paid traffic (64,90%) and organic traffic (20,10%). Use the desktop (54,70%) and mobile (34,30%) and mainly come from the Netherlands (84,50%). The visitor type is 55,5% new and 44,50% returning. The customers in cluster 2 purchase fewer products than the customers in cluster 1 (mean: -0.18 and max: 3.57). The revenue is less for customers in cluster 2 with regard to customers in cluster 1 (mean: -0.19 and max: 2.68)

4.3 Interpretation of patterns – Cluster analysis 1

Two clusters have been found in which some differences have been noticed that could potentially aid the marketing strategy. One critical difference between the clusters is the revenue and amount of products purchased. The revenue and amount of products for cluster 1 is higher than from cluster 2. So, cluster 1 can be indicated as high revenue customers and cluster 2 as low revenue customers (the revenue variable is including VAT and excluding related expenses). Another vital difference between the clusters is the traffic type: paid search. For cluster 1 the paid traffic is 13,70% and respectively 64,90% for cluster 2. So, customers in cluster 1 can be indicated as low-cost customers, and customers in cluster 2 can be indicated as high-cost customers. Additionally, an important difference between the clusters is the use of device type. Cluster 1 can be indicated as a mobile usage cluster (80%), cluster 2 as a mix of desktop (54,70%) and mobile (34,30%). Furthermore, for cluster 1 is 62,70% new visitor and 37,30% returning visitor. For cluster 2 is 55,50% new visitor and 44,50% returning visitor. The customers from both of the clusters come mainly from the Netherlands (85%). To conclude, the customer in cluster 1 can be summarized as high-value customers (high revenue, low cost) using mainly mobile. The customer in cluster 2 can be summarized as low-value customers (low revenue, high cost), using desktop and mobile (table 8).

Table 8. Overview of cluster characteristics (cluster analysis 1)

Cluster	**1 - High-value customer (high revenue, low cost) – using mainly mobile**	**2 - Low-value customer (low revenue, high cost) – using desktop and mobile**
Type of traffic	Direct (61,80%)	Paid (64,90%)
Type of device	Mobile (80%)	Desktop (54,70%)
Country	Netherlands (85%)	Netherlands (84,50%)
Type of visitor	New Visitor (62,70%)	New Visitor (55,50%)
Amount of products (scaled)	0,27	-0,18
Amount of revenue (scaled)	0,28	-0,19

Note: The numerical variables present the means

4.4 Data mining – Cluster analysis 2

This section will analyze the results of the second cluster analysis.

First of all, the Elbow method, Silhouette score, and Hierarchical clustering dendrogram are used to determine the number of clusters.

4.4.1 Elbow method

The graph below shows a clear “elbow-cut”, which is at two clusters.

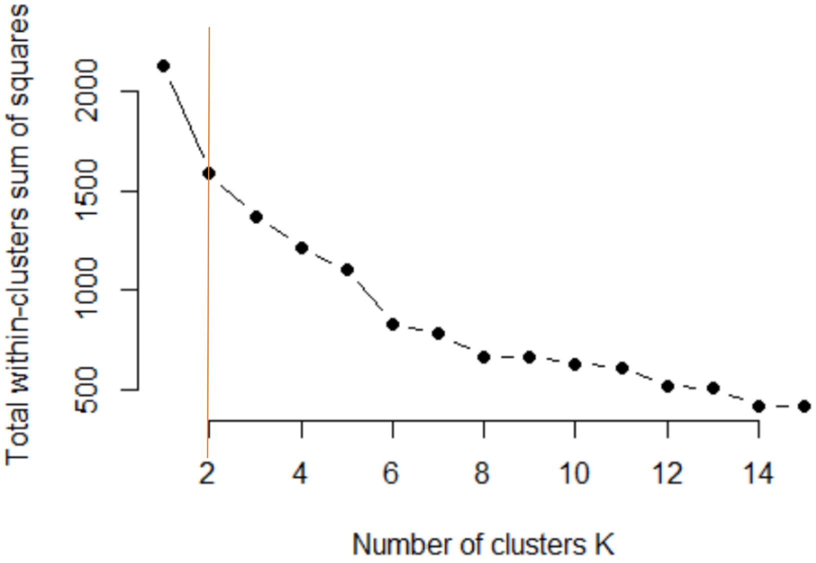


Figure 6. Elbow graph for cluster analysis 2.

4.4.2 Silhouette score

The graph of the silhouette score below shows that two clusters have the highest silhouette score (0,32).

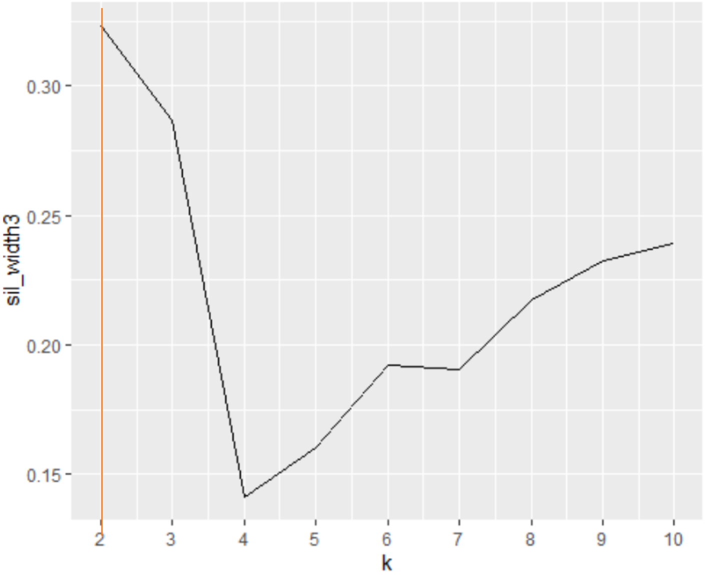


Figure 7. Silhouette score (sil_width) for the number of clusters (K) for clusters analysis 2.

4.4.3 Hierarchical clustering dendrogram

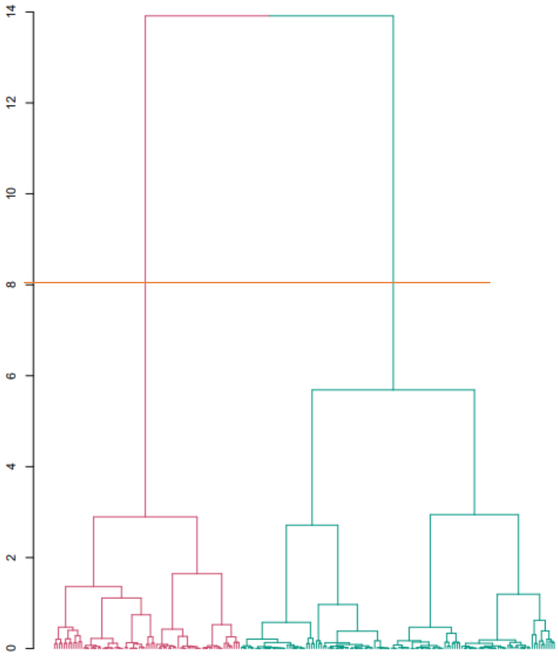


Figure 8. Hierarchical clustering dendrogram for cluster analysis 1.

When cutting the longest vertical line, the number of clusters will be two.

The Elbow method, silhouette score and hierarchical clustering dendrogram show an equal number of 2 clusters. The silhouette score is 0,32, which is not a (very) high score. The cluster validity can be questioned.

4.5 Results of K-prototype – Cluster analysis 2

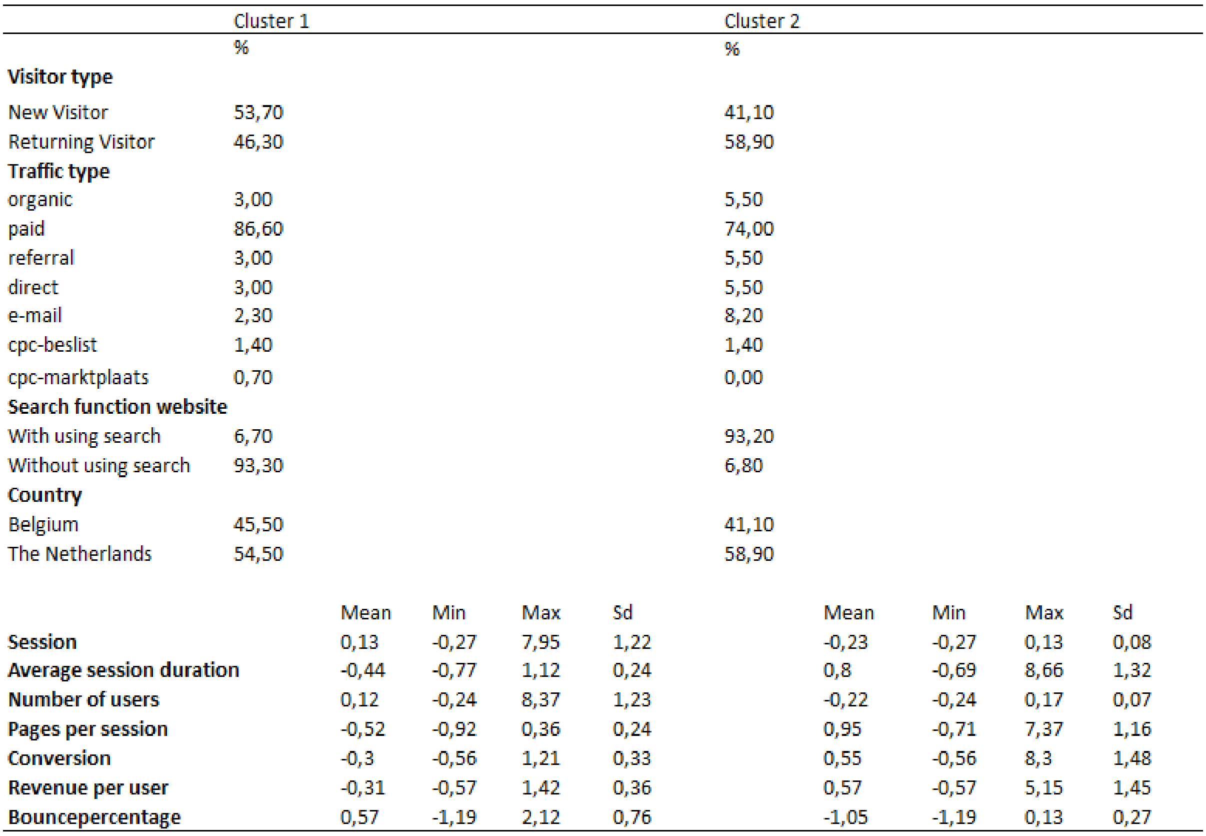
This section will discuss cluster analysis 2 using K-prototype (dataset 2)

Table 9. Distribution of clusters,

Cluster	N	%
1	134	64,73
2	73	35,26
Total	207	100

Cluster 1 has 134 (64,73%) visitors and cluster 2 has 73 (35,26%) visitors, as seen in table 9.

Table 10. Cluster results of clusters analysis 2.



	Cluster 1					Cluster 2				
	%					%				
Visitor type										
New Visitor	53,70					41,10				
Returning Visitor	46,30					58,90				
Traffic type										
organic	3,00					5,50				
paid	86,60					74,00				
referral	3,00					5,50				
direct	3,00					5,50				
e-mail	2,30					8,20				
cpc-beslist	1,40					1,40				
cpc-marktplaats	0,70					0,00				
Search function website										
With using search	6,70					93,20				
Without using search	93,30					6,80				
Country										
Belgium	45,50					41,10				
The Netherlands	54,50					58,90				
		Mean	Min	Max	Sd		Mean	Min	Max	Sd
Session		0,13	-0,27	7,95	1,22		-0,23	-0,27	0,13	0,08
Average session duration		-0,44	-0,77	1,12	0,24		0,8	-0,69	8,66	1,32
Number of users		0,12	-0,24	8,37	1,23		-0,22	-0,24	0,17	0,07
Pages per session		-0,52	-0,92	0,36	0,24		0,95	-0,71	7,37	1,16
Conversion		-0,3	-0,56	1,21	0,33		0,55	-0,56	8,3	1,48
Revenue per user		-0,31	-0,57	1,42	0,36		0,57	-0,57	5,15	1,45
Bouncepercentage		0,57	-1,19	2,12	0,76		-1,05	-1,19	0,13	0,27

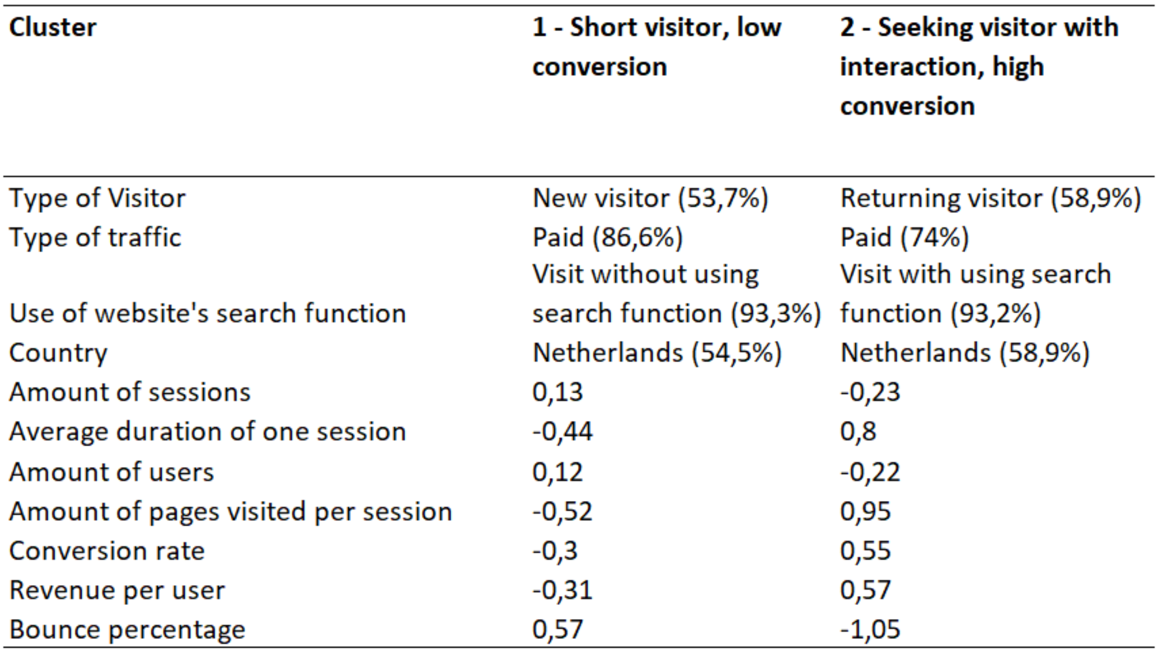
In cluster 1, the type of visitors is 53,70% new and 46,30% returning. The traffic type comes mainly from paid search (86,60%). The visitors do not use the search function (93,3%) during their journey. 45,50% comes from Belgium and 54,50% comes from the Netherlands. The visitors in cluster 1 have more sessions in their journey with regard to the visitors in cluster 2 (mean: 0,13 and max: 7,95). The average session duration is shorter with regard to the visitors in cluster 2 (mean: -0,44 and max: 1,12). The number of users during this journey is higher than in cluster 2 (mean: 0,12 and max: 8,37). Furthermore, the number of visited pages during one session is less than in cluster 2 (mean: -0,52 and max: 0,36) The conversion rate is also less than in cluster 2 (mean: -0,30 and max: 1,21). This results that the revenue per user is also less than in cluster 2 (mean: -0,31 and max: 1,42). The bounce percentage is more in cluster 1 than in cluster 2 (mean: 0,57 and max: 2,12).

In cluster 2, the type of visitors is 41,10% new and 58,90% returning. The traffic type mainly comes from paid search (74%). The visitors do use the search function (93,20%) during their journey. 41,10% of the visitors come from Belgium and 58,90% come from the Netherlands. The number of sessions is less than in cluster 1 (mean: -0,23 and max: 0,13). However, the average session duration is longer than in cluster 1 (mean: 0,80 and max: 8,66). The number of users is less than in cluster 1 (mean: -0,22 and max: -0,17). The pages that are visited are higher than in cluster 1 (mean: 0,95 and max: 7,37). Furthermore, the conversion rate is higher than in cluster 1 (mean: 0,55 and max: 8,3). This results, that the revenue per user is higher than in cluster 1 (mean: 0,57 and max: 5,15). Lastly, the bounce percentage is less in cluster 2 with regard to cluster 1 (mean: -1,05 and max: 0,13).

4.6 Interpretation of patterns – Cluster analysis 2

Two clusters have been found in which two visitor clusters can be identified. First of all, there is a significant difference between the use of the website's search function. In cluster 1, visitors do not use the search function (93,3%), whereas in cluster 2 the visitors do use the search function (93,2%). Furthermore, the amount of sessions in cluster 1 is higher than in cluster 2. However, the average duration of one session and the amount of page visits during one session is higher in cluster 2 than in cluster 1. Furthermore, the bounce percentage is for cluster 1 higher than for cluster 2. Cluster 1 can be identified as short visitor, whereas cluster 2 can be identified as seeking visitor with a lot of interaction. The number of visitors is in cluster 1 higher as in cluster 2, so cluster 1 can be identified as the largest group of visitors. The conversion rate and revenue per user are both for cluster 2 higher than for cluster 1. Therefore, cluster 1 can be indicated as low conversion visitors and cluster 2 can be indicated as high conversion visitors.

Table 11. Overview of cluster characteristics.



Cluster	**1 - Short visitor, low conversion**	**2 - Seeking visitor with interaction, high conversion**
Type of Visitor	New visitor (53,7%)	Returning visitor (58,9%)
Type of traffic	Paid (86,6%)	Paid (74%)
Use of website's search function	Visit without using search function (93,3%)	Visit with using search function (93,2%)
Country	Netherlands (54,5%)	Netherlands (58,9%)
Amount of sessions	0,13	-0,23
Average duration of one session	-0,44	0,8
Amount of users	0,12	-0,22
Amount of pages visited per session	-0,52	0,95
Conversion rate	-0,3	0,55
Revenue per user	-0,31	0,57
Bounce percentage	0,57	-1,05

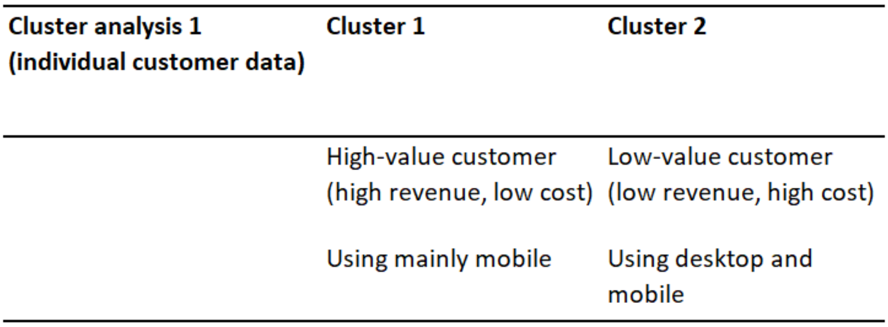
Note: The numerical variables present the means.

4.7 Summary of the cluster analyzes

To conclude, two different cluster analyzes were conducted. In the first cluster analysis, based on individual customer data, two customer clusters were found. Cluster 1 can be identified as high-value customers (high revenue, low cost) and use mainly mobile. Cluster 2 can be identified as low-value customers (low revenue, high cost) and use desktop and mobile (table 12).

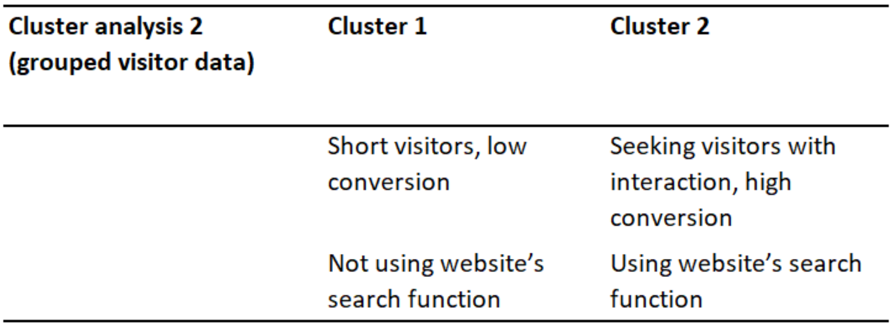
In the second cluster analysis, based on grouped visitor data, two visitor clusters were found. Cluster 1 can be identified as short visitors with low conversion. The visitors in this cluster do not use the website's search function. Cluster 2 can be identified as seeking visitors with interaction and high conversion. The visitors in this cluster do use the website's search function (table 13).

Table 12. Summary of cluster analysis 1.



Cluster analysis 1 (individual customer data)	**Cluster 1**	**Cluster 2**
	High-value customer (high revenue, low cost)	Low-value customer (low revenue, high cost)
	Using mainly mobile	Using desktop and mobile

Table 13. Summary of cluster analysis 2.



Cluster analysis 2 (grouped visitor data)	**Cluster 1**	**Cluster 2**
	Short visitors, low conversion	Seeking visitors with interaction, high conversion
	Not using website's search function	Using website's search function

5. MARKETING STRATEGY

As stated in the literature review, providing personalized marketing for different customer clusters is the main marketing strategy of customer segmentation. This section will focus on what a suitable marketing strategy is for the discovered clusters. The section will also focus on how digital technologies (e.g. Artificial Intelligence) can aid the marketing strategy.

5.1 Marketing strategy for cluster analysis 1

Cluster 1 can mainly be identified as high-value customers, whereas cluster 2 can be identified as low-value customers. Both clusters are interesting from a marketing perspective however the purposes of the strategies are different.

Cluster 1 can be identified as high-value customers. The amount of products and revenue is high in this cluster. Customer retention and providing personalized offerings is a vital marketing strategy for this cluster. As stated above, Google (2021) will ban third-party cookies which makes it more important to analyze and retain existing customers. Because this cluster can be indicated as high-value customers, it is even more important to retain these customers. In order to maintain these high-value customers, personalized offerings can be given in which customers receive personalized discounts and product recommendations. This can be done through, for example, personalized e-mail marketing campaigns and personalized loyalty programs. These personalized marketing offerings can improve the customer experience. Kumar et al. (2021) argue that the need for personalized marketing and offerings has increased. The millennial customer wants simple, rapid, and personalized offerings that are in line with their needs (Kumar et al., 2021). These personalized offerings can create an engaging relationship with the customer, which will improve the customer experience (Kumar et al., 2019). It would be interesting to research what these customers buy in order to give relevant product recommendations. Also, it would be interesting to research what campaigns in the paid search lead to a conversion in order to tailor the marketing communication.

Cluster 2 can be identified as low-value customers. The amount of products and revenue is low in this cluster. Cross- and upselling is a vital marketing strategy for this cluster. In order to increase sales, personalized recommendations could be provided through a recommender engine. A recommender engine can recommend personalized products that visitors may find attractive and are expected to buy (Behera, Gunasekaran, Gupta, Kamboj, & Bala, 2020). Because a recommender engine can show products that visitors are interest in as well as the correlations between products, the sales can increase (Deng, Shi, Chen, Kwak, & Tang, 2019). In order words, the recommender engine supports customers in their decision-making process. Behera et al. (2020) state that a recommender engine should show relevant and personalized products at various touchpoints in the customer journey. Furthermore, these product recommendations could also be given in an e-mail marketing campaign.

For both clusters, an optimization marketing strategy can be given. 64,90% of the customers of cluster 2 comes from paid traffic (i.e. SEA traffic). It would be interesting to research if these customers are still profitable because the marketing costs of the paid search are high for Google Adwords and Shopping. Furthermore, the traffic from both clusters comes approximately 20% from organic traffic (i.e. SEO traffic). It would for example be interesting to conduct an SEO cluster analysis with relevant keywords to optimize the SEO position. Additionally, the customers in cluster 1 mainly use mobile (80%). It would be interesting to research if the website pages are mobile optimized.

5.2 Marketing strategy for cluster analysis 2

In the second cluster analysis, cluster 1 can be identified as short visitors with low conversion, whereas cluster 2 can be identified as seeking visitors with interaction and high conversion. Both clusters are interesting from a marketing perspective however, the purposes of the strategies are different.

Cluster 1 can be identified as short visitors with low conversion, therefore it is important to provide relevant and personalized products, offerings, and content at multiple touchpoints in the customer journey. As stated above, offering personalized products, content, and offerings has the ability to improve the customer experience. Similarly, this cluster would benefit from personalized product recommendations using a recommender engine to drive sales.

Moreover, an optimization strategy could be given for this cluster. In order to improve the customer experience during the customer journey, the bounce percentage needs improvement. The bounce percentage is the percentage of a session that only interact/viewed one page. The goal is to have a low bounce percentage. It would be interesting to research what pages have a high bounce percentage and optimize these pages with relevant content.

Cluster 2 can be identified as seeking visitors with interaction and high conversion, therefore it is important to retain and provide personalized marketing. In order to maintain these customers, personalized offerings can be given through personalized price settings and product recommendations. This cluster would also benefit from personalized product recommendations using a recommender engine because it supports customers in their decision-making process and will reduce the search space.

Additionally, an optimization strategy could be given for this cluster. The visitors in this cluster use the website’s search function (93,2%). It would be interesting to research what they search for. First of all, it is interesting researching if relevant products are shown with the corresponding keyword and if necessary, optimize the search function. Furthermore, those keywords give insight into the needs of the customers which gives a better picture of the customer profile. Also, the keywords can be used to optimize the SEO strategy.

6. DISCUSSION

The first objective of this research was to identify customer clusters using a clustering approach and algorithm (unsupervised machine learning) in data collected during online customer-business interactions. As mentioned in the literature review, partitioning-based is a suitable approach in analyzing a mixed and large dataset, and therefore suitable for this research. Partitioning-based is used in various clustering studies (Pater et al., 2019; Walters & Bekker, 2017; Umuhoza et al., 2020; Kansal et al., 2018; Pascal et al., 2015; Kuiper, 2018). The k-prototype algorithm was used because it can handle both numerical and categorical data. Two cluster analyzes were conducted with different variables. Individual customer data is used in the first cluster analysis. In the first cluster analysis, two customer clusters have been found in which cluster 1 can be indicated as high-value customers and cluster 2 as low-value customers. The results reflect those of for example Pater et al. (2015) who found four types of clusters that are distinguished between high or low buyers. One interesting finding from the first cluster analysis in this study was that 61,80% comes from direct traffic in cluster 1 and 64,90% comes from paid traffic in cluster 2. In the second cluster analysis, grouped visitor data is used. Similarly, two visitor clusters have been found in which the visitors in cluster 1 can be identified as short visitors with low conversion, whereas the visitors in cluster 2 can be identified as seeking visitors with interaction and high conversion. One interesting finding is that visitors in cluster 1 do not use the search function during their journey (93,2%), whereas in cluster 2 visitors do use the search function (93,2%). These results are in line with the findings of Kuiper (2018) who found six behavioral clusters in which two clusters are most significant with regard to conversion. Some behavioral patterns (e.g. Pdf download) are found that could explain the high conversion.

The second objective of this study was to develop an appropriate marketing strategy for the discovered clusters. As stated in the literature review, providing personalized strategies for different customer segments is the main marketing strategy of clustering. Wang et al. (2020) provide three marketing strategies based on precision marketing. 1) personalized product offerings 2) personalized price setting 3) accurate marketing information/content. In the current study, marketing strategies are given for the discovered clusters, focusing on digital technologies. For the high-value customer cluster (cluster analysis 1 cluster 1), the marketing strategy is focused on customer retention marketing and providing personalized offerings. In order to retain these customers, personalized offerings and recommendations could be given through for example an e-mail marketing campaign and personalized loyalty programs. For the low-value customer cluster (cluster analysis 1 cluster 2), the marketing strategy is focused on cross- and upselling. Personalized product recommendations could be provided through a recommender engine. Furthermore, optimization strategies (SEO and SEA) are given for both clusters. In the second cluster analysis, short visitors with low conversion (cluster analysis 2 cluster 1), the marketing strategy is focused on providing relevant and personalized products, offerings, and content at multiple touchpoints in the customer journey. For the seeking visitor with interaction and high conversion (cluster analysis 2 cluster 2) the marketing strategy is focused on customer retention marketing and providing personalized marketing strategies using personalized offerings and products. For both clusters, some optimization strategies are given concerning optimizing the bounce percentage and researching the website's search function. To conclude, this research provides marketing strategies for the discovered clusters, focusing on digital technologies. The marketing strategies include personalized products, offerings, and content through e-mail marketing campaigns, loyalty programs, and recommender engines. This research also argues that optimization strategies can be given based on the discovered clusters.

The results of this research are not generalizable however, the methods and strategies can be used as a basis by researchers and businesses to guide the marketing strategy.

6.1 Theoretical implications

This research contributes to the understanding of how a clustering approach (unsupervised machine learning) can aid the marketing strategy. The results of this study contribute to the growing field of using digital technologies in marketing. According to the Marketing Science Institute (2020), there is a need to research how various measurement approaches can create value in the decision-making of businesses and their marketing strategy. This study provides a basis on how unsupervised machine learning can be used to identify clusters that can aid the marketing strategy. The framework of Kuiper (2018) is used for determining the right type of clustering method and algorithm for the mixed and large dataset. One theoretical implication is the use of K-prototype in this research, which provides researchers with a basis for handling a mixed and large dataset. This paper also laid a foundation for providing marketing strategies for the discovered clusters, which is focused on personalized marketing strategies and optimization strategies

6.2 Practical implications

This research tries to prove that SME online businesses with limited resources, can aid their marketing strategy by using a clustering approach (unsupervised machine learning) in analyzing customer interaction data. With the rapid development of digital technologies and data, there is a need for businesses to use this technology and data in a profitable way to keep up with the competition. This research tries to prove that not only big companies with large amounts of marketing budget can adjust to this trend. The data of this research is retrieved from Google Analytics (a tool that is used by most online businesses) and analyzed in the statistical program R. It is not only essential to implement a clustering approach, however it is even more important to know how to analyze this data in order to aid the marketing strategy successfully. This research informs marketing managers and businesses how to implement a cluster approach and how the results of the clusters can aid the marketing strategy, providing personalized and optimization marketing strategies.

Furthermore, this study informs how marketing managers and businesses can use first-party cookies data to aid the marketing strategy. The capability of analyzing first-party cookies data will play a significant role in marketing in the future, because of the ban of third-party cookies by Google. Moreover, this research is limited to data from Google Analytics. A key priority for SME online businesses should therefore be investing in gathering and collecting the right type of data to have more complete and valid results.

7. CONCLUSION

The first aim of this study was to identify customer clusters using a clustering approach and algorithm (unsupervised machine learning) in data collected during online customer-business interactions. The data used in this study is collected from an online retail webshop (SME), that is retrieved from Google Analytics (1 Jan 2018 till 1 Jan 2021). The first dataset consists of individual customer data, whereas the second dataset consists of grouped visitor data. The first research question was formulated as: "To what extent can online customer clusters of an online webshop SME be identified with the use of clustering (unsupervised machine learning)?". After critical reviewing literature regarding customer segmentation/clustering and machine learning, it can be concluded that partitioning-based clustering using k-prototype is a suitable approach for a mixed and large dataset. Using the KDD-process as research methodology, the k-prototype algorithm found for both datasets two clusters: High-value and low-value customers for cluster analysis 1 and short visitors with low conversion and seeking visitors with interaction and high conversion for cluster analysis 2. The second aim of this study was to develop a marketing strategy for the discovered clusters. The second research question was formulated as "What is an appropriate marketing strategy for the discovered clusters?". This research provides marketing strategies for the discovered clusters, focusing on digital technologies. The marketing strategies include personalized products, offerings, and content through e-mail marketing campaigns, loyalty programs, and recommender engines. This research also argues that optimization strategies (SEO and SEA) can be given based on the discovered clusters.

7.1 Limitations

A limitation of this study is the variety and volume of the dataset. The availability of the variety of the variables determines the completeness of the results. Many variables in the original dataset in Google Analytics have missing values (not set), which makes it harder to select these variables. Furthermore, this research is limited to "Transaction-ID", which only tracked the individual purchasing customers. A "Customer-ID or Visitor-ID" that records the customer journey of every individual visitor would lead to a more complete dataset and results. Additionally, tracking vital behavioral attributes (e.g. viewed product pages, add to cart, create an account etc.) for every individual visitor would lead to a more complete view of the conversion funnel and customer behavior.

Furthermore, is it important for the validity of the results to critically evaluate how the clustering method is performing. First of all, some outliers were removed from the dataset however, the data was still a bit skewed. This is because there is a difference between business and customer transactions. Since businesses often purchase a large number of products, this leads to skewed data with regard to customers who purchase one product. Since K-prototype is sensitive for outliers and skewed data, this could lead to invalid results. Second, the methods for determining the number of clusters (Elbow method, Silhouette Score, and Hierarchical dendrogram) did not show an equal number of clusters for the first cluster analysis. The silhouette score was 0,49 and 0.32 respectively for cluster analysis 1 and 2 which is not a (very) high score and could therefore lead to invalid results.

The COVID-19 pandemic has changed businesses and customers. The competition online is increasing and customer behavior is changing because of the pandemic. The dataset is from 1 Jan 2018 till 1 Jan 2021, however the COVID-19 pandemic and other external forces are not taken into consideration in the analysis. Bibby, Gordon, Schuler, and Stein (2021) argue that customer behavior has changed significantly during the COVID-19 pandemic and therefore existing models and data are not valid anymore.

7.2 Future research

Further research could focus on supervised machine learning with labeled data for marketing purposes. This research provides a basis for supervised machine learning and provides labeled data (discovered clusters). Future research could focus on for example classification. Aggarwal (2015) states that a target marketing application is an example of a classification method. The goal of the application is to predict if a customer is interested in a particular product based on relating features to the class label. A training model will be developed to test and predict the class labels (Aggarwal, 2015). Clusters are found that have similar patterns consequently, a prediction model can be made for the new customers. These new customers can be classified into one predefined cluster, to give for example real-time personalized marketing strategies. Further research could also focus on how to implement and optimize recommender engines for SME online businesses. Moreover, future research could compare how the results differ if data is used before the COVID-19 pandemic and during/after the COVID-19 pandemic. Businesses and marketers need to know if their methods and models are still valid, because of the changing customer behavior. Furthermore, because the results of this research are limited by the availability of data, future research could assess how SME online businesses could gather and collect the right type of data in a profitable way for marketing purposes. Also, future research could identify clusters with individual visitor data, which would lead to a more complete view of the conversion funnel and customer behavior. The individual visitor data include a distinction between a visitor who makes a purchase and a visitor who did not and behavioral attributes (e.g. viewed product pages, add to cart, create an account etc.).

8. REFERENCES

Aggarwal, C. C. (2015). Data Classification. *Data Mining*, 285–344. https://doi.org/10.1007/978-3-319-14142-8_10

Baarsch, J., & Celebi, M. E. (2012). Investigation of Internal Validity Measures for K-Means Clustering. *Proceedings of the International MultiConference of Engineers and Computer Scientists, 1*, 471–476.

Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S., & Bala, P. K. (2020). Personalized digital marketing recommender engine. *Journal of Retailing and Consumer Services*, *53*, 101799. https://doi.org/10.1016/j.jretconser.2019.03.026

Bibby, C., Gordon, J., Schuler, G., & Stein, E. (2021). *The big reset: Data-driven marketing in the next normal.* McKinsey & Company. https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-big-reset-data-driven-marketing-in-the-next-normal

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2017). Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising, 46*(3), 363–376. https://doi.org/10.1080/00913367.2017.1339368

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications, 105*(9), 17–24.

Chen, X., Fang, Y., Yang, M., Nie, F., Zhao, Z., & Huang, J. Z. (2018). PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data. *IEEE Transactions on Knowledge and Data Engineering, 30*(3), 559–572. https://doi.org/10.1109/tkde.2017.2763620

Deng, W., Shi, Y., Chen, Z., Kwak, W., & Tang, H. (2019). Recommender system for marketing optimization. *World Wide Web*. https://doi.org/10.1007/s11280-019-00738-1

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2019). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

European Commission. (2017). Data protection. European Commission. https://ec.europa.eu/info/law/law-topic/data-protection

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., & Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing, 2*(3), 267–279. https://doi.org/10.1109/tetc.2014.2330519

Fan, J., & Li, D. (1998). An overview of data mining and knowledge discovery. *Journal of Computer Science and Technology, 13*(4), 348–368. https://doi.org/10.1007/bf02946624

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence, 17*(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27–34. https://doi.org/10.1145/240455.240464

Google (2021a). *Charting a course towards a more privacy-first web.* https://blog.google/products/ads-commerce/a-more-privacy-first-web/

Google. (2021b*). Clear, enable, and manage cookies in Chrome - Computer - Google Chrome Help.* https://support.google.com/chrome/answer/95647?co=GENIE.Platform%3DDesktop&hl=en

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics, 27*(4), 857. https://doi.org/10.2307/2528823

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., &mo Sriram, S. (2006). Modeling Customer Lifetime Value*. Journal of Service Research, 9(*2), 139–155. https://doi.org/10.1177/1094670506293810

Gupta, S., Leszkiewicz, A., Kumar, V., Bijmolt, T., & Potapov, D. (2020). Digital Analytics: Modeling for Insights and New Methods. *Journal of Interactive Marketing, 51,* 26–43. https://doi.org/10.1016/j.intmar.2020.04.003

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery, 2*(3), 283–304. https://doi.org/10.1023/a:1009769707641

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Hung, P. D., Lien, N. T. T., & Ngoc, N. D. (2019). Customer Segmentation Using Hierarchical Agglomerative Clustering. *Proceedings of the 2019 2nd International Conference on Information Science and Systems*, 33–37. https://doi.org/10.1145/3322645.3322677

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications, 26*(2), 181–188. https://doi.org/10.1016/s0957-4174(03)00133-7

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 135–139. https://doi.org/10.1109/ctems.2018.8769171

Kodinariya, T., & Makwana, P. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1*(6), 90–95.

Kuiper, F. J. (2018). Behavioural profiles of potential students as basis for more effective university recruiting. (Master's Thesis). Retrieved from https://essay.utwente.nl/77027/1/Kuiper_MA_BMS.pdf

Kumar, V. (2018). A Theory of Customer Valuation: Concepts, Metrics, Strategy, and Implementation. *Journal of Marketing, 82*(1), 1–19. https://doi.org/10.1509/jm.17.0208

Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the Role of Artificial Intelligence in Personalized Engagement Marketing. *California Management Review*, *61*(4), 135–155. https://doi.org/10.1177/0008125619859317

Kumar, V., Ramachandran, D., & Kumar, B. (2021). Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research*, *125*, 864–877. https://doi.org/10.1016/j.jbusres.2020.01.007

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

Marketing Science Institute (2020), "Research Priorities 2020–2022." Cambridge, MA: Marketing Science Institute, [available at https://www.msi.org/wp-content/uploads/2020/06/MSI_RP20-22.pdf

McKinsey & Company. (2020). *An executive's guide to AI*. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai

Mo, J., Kiang, M. Y., Zou, P., & Li, Y. (2010). A two-stage clustering approach for multi-region segmentation. *Expert Systems with Applications, 37*(10), 7120–7131. https://doi.org/10.1016/j.eswa.2010.03.003

Ogbuabor, G., & Ugwoke, F. N. (2018). Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science and Information Technology, 10*(2), 27–37. https://doi.org/10.5121/ijcsit.2018.10203

Pascal, C., Ozuomba, S., & kalu, C. (2015). Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. *International Journal of Advanced Research in Artificial Intelligence, 4*(10), 40–44. https://doi.org/10.14569/ijarai.2015.041007

Pater, A.-M., Vari-Kakas, S., Poszet, O., & Pintea, I. G. (2019). Segmenting Users of an Online Store Using Data Mining Techniques. *2019 15th International Conference on Engineering of Modern Electric Systems (EMES)*, 205–208. https://doi.org/10.1109/emes.2019.8795200

Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society, 24*(4), 483–502. https://doi.org/10.1016/s0160-791x(02)00038-6

Saha, R., Tariq, M. T., Hadi, M., & Xiao, Y. (2019). Pattern Recognition Using Clustering Analysis to Support Transportation System Management, Operations, and Modeling. Journal of Advanced *Transportation, 2019*, 1–12. https://doi.org/10.1155/2019/1628417

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217–222.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing, 21*(1), 3. https://doi.org/10.2307/1247695

Stoltenkamp, J. (2021). *Online privacy tracking & online adverteren: wat gaat er veranderen?* Marketingfacts. https://www.marketingfacts.nl/berichten/online-privacy-tracking-en-online-adverteren-wat-gaat-er-veranderen

Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting. *Marketing Science*, *35*(3), 405–426. https://doi.org/10.1287/mksc.2015.0956

Umuhoza, E., Ntirushwamaboko, D., Awuah, J., & Birir, B. (2020). Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa. *SAIEE Africa Research Journal, 111*(3), 95–101. https://doi.org/10.23919/saiee.2020.9142602

Varnali, K. (2021). Online behavioral advertising: An integrative review. *Journal of Marketing Communications, 27*(1), 93–114. https://doi.org/10.1080/13527266.2019.1630664

Walters, M., & Bekker, J. (2017). Customer super-profiling demonstrator to enable efficient targeting in marketing campaigns. *South African Journal of Industrial Engineering*, *28*(3), 113–127. https://doi.org/10.7166/28-3-1846

Wang, H., Wang, J., & Zhong, Z. (2020). Research on Precision Marketing Strategy Based on Cluster Analysis Algorithm. *2020 International Conference on E-Commerce and Internet Technology (ECIT),* 208–211. https://doi.org/10.1109/ecit50008.2020.00054

Wu, J., & Lin, Z. (2005). Research on customer segmentation model by clustering. *Proceedings of the 7th International Conference on Electronic Commerce - ICEC '05,* 316–318. https://doi.org/10.1145/1089551.1089610

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How much can behavioral targeting help online advertising*? Proceedings of the 18th International Conference on World Wide Web - WWW '09*, 261–270. https://doi.org/10.1145/1526709.1526745

Yin, S., & Pan, H. (2020). Application of Big Data to Precision Marketing in B2C E-commerce. *Advances in Intelligent Systems and Computing, 1117*, 731–738. https://doi.org/10.1007/978-981-15-2568-1_100

You, Z., Si, Y.-W., Zhang, D., Zeng, X. X., Leung, S. C. H., & Li, T. (2015). A decision-making framework for precision marketing. *Expert Systems with Applications, 42*(7), 3357–3367. https://doi.org/10.1016/j.eswa.2014.12.022

9. APPENDIX

9.1 Appendix 1 Framework of data analytics

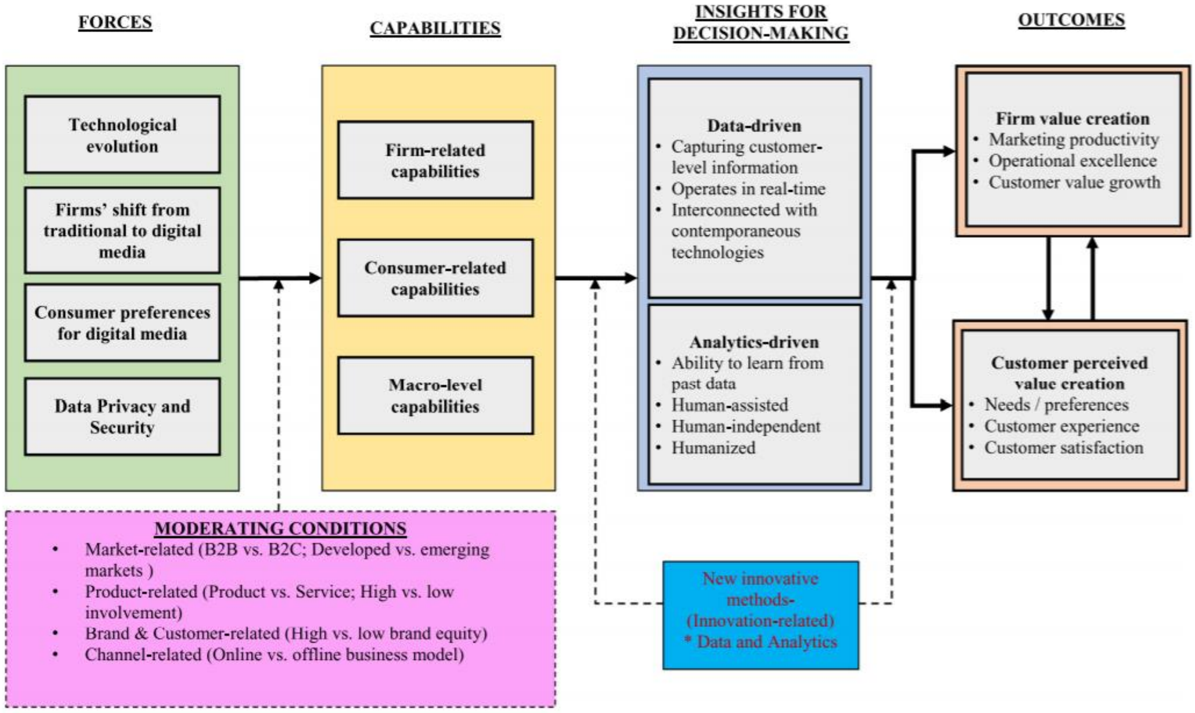

FORCES
Technological evolution
Firms' shift from traditional to digital media
Consumer preferences for digital media
Data Privacy and Security
CAPABILITIES
Firm-related capabilities
Consumer-related capabilities
Macro-level capabilities
INSIGHTS FOR DECISION-MAKING
Data-driven
• Capturing customer-level information
• Operates in real-time
• Interconnected with contemporaneous technologies
Analytics-driven
• Ability to learn from past data
• Human-assisted
• Human-independent
• Humanized
OUTCOMES
Firm value creation
• Marketing productivity
• Operational excellence
• Customer value growth
Customer perceived value creation
• Needs / preferences
• Customer experience
• Customer satisfaction
MODERATING CONDITIONS
• Market-related (B2B vs. B2C; Developed vs. emerging markets)
• Product-related (Product vs. Service; High vs. low involvement)
• Brand & Customer-related (High vs. low brand equity)
• Channel-related (Online vs. offline business model)
New innovative methods- (Innovation-related) * Data and Analytics

Figure 9. Framework of data analytics (Gupta et al., 2020).

9.2 Appendix 2 Cleaning and processing of data in R

Cleaning and processing of data

1) Missing values

# Treating missing values

```
> sum(is.na(mydata))
   0 missing values
```

2) Outliers

# Making boxplots

```
boxplot(mydata$Opbrengst)
```

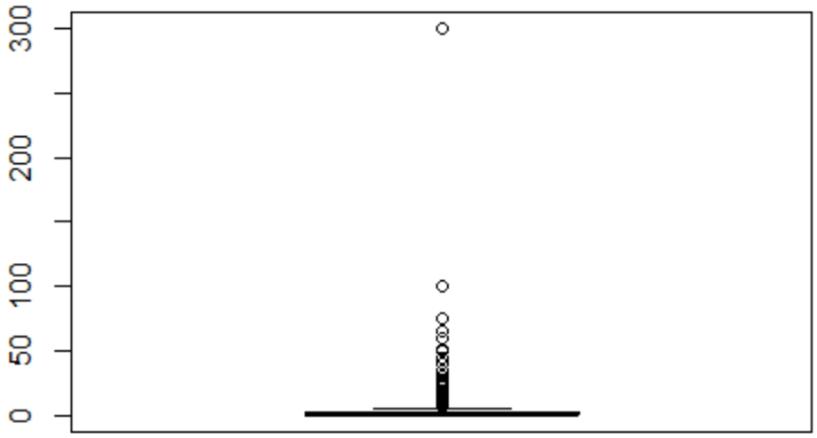


Figure 10. Boxplot for revenue (Outlier detection).

```
boxplot(mydata$Aantal)
```

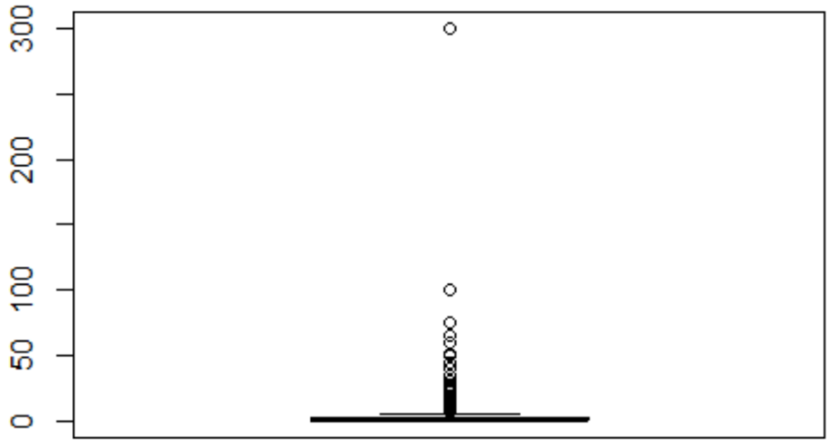
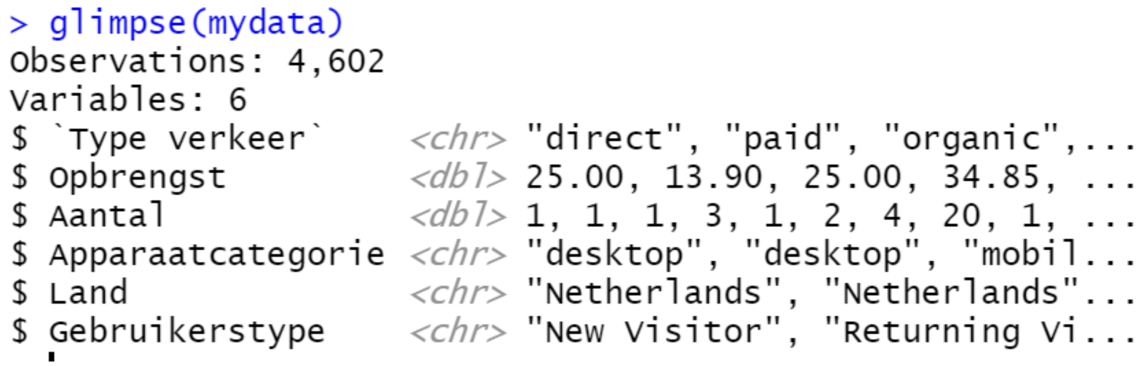


Figure 11. Boxplot for amount of product (outlier detection).

3) Assigning data to its data type (e.g. factor, number, string)

# checking out the dataset



```
> glimpse(mydata)
Observations: 4,602
Variables: 6
$ `Type verkeer`      <chr> "direct", "paid", "organic",...
$ Opbrengst           <dbl> 25.00, 13.90, 25.00, 34.85, ...
$ Aantal              <dbl> 1, 1, 1, 3, 1, 2, 4, 20, 1, ...
$ Apparaatcategorie <chr> "desktop", "desktop", "mobil...
$ Land                <chr> "Netherlands", "Netherlands"...
$ Gebruikerstype      <chr> "New Visitor", "Returning Vi...
```

The character <chr> variables will be transformed into factor variables.

The double class <dbl> variable will be transformed into numerical variables

# Converting the data to its data type

```
new_mydata$`Type verkeer` <- as.factor(new_mydata$`Type verkeer`)
new_mydata$Apparaatcategorie <- as.factor(new_mydata$Apparaatcategorie)
new_mydata$Land <- as.factor(new_mydata$Land)
new_mydata$Gebruikerstype <- as.factor(new_mydata$Gebruikerstype)
new_mydata$scaled_opbrengst <- as.numeric(new_mydata$scaled_opbrengst)
new_mydata$scaled_aantal <- as.numeric(new_mydata$scaled_aantal)
```

9.3 Appendix 3 R script for cluster analyzes

R script clustering analysis 1

Hierarchical clustering:

#attachting the library

```
library(stats)
library(dplyr)
library(ggplot2)
library(clustMixType)
library(cluster)
library(tidyverse)
library(dendextend)
library(FactoMineR)
```

# import the dataset

```
library(readxl)
excel_bestand_shop_2_zonder_transactie <- read_excel("~/Thesis/R/excel best
and shop 2 zonder transactie.xlsx",
+     sheet = "Gegevensset1")
View(excel_bestand_shop_2_zonder_transactie)
```

# Change the name of the dataset

```
mydata <- excel_bestand_shop_2_zonder_transactie
```

# Treating missing values

```
sum(is.na(mydata))
```

```
0 missing values
```

# Outlier detection

```
 boxplot(mydata$Aantal)
```

```
boxplot(mydata$Opbrengst)
```

# convert charactercs into factors

```
mydata$`Type verkeer` <- as.factor(mydata$`Type verkeer`)
mydata$Apparaatcategorie <- as.factor(mydata$Apparaatcategorie)
mydata$Land <- as.factor(mydata$Land)
mydata$Gebruikerstype <- as.factor(mydata$Gebruikerstype)
```

#scale the data using daisy function

```
gower_dist <- daisy(mydata, metric = "gower")
```

# Plot the dendrogram using complete method

```
hc_shop <- hclust(gower_dist, method = "complete")
dend_shop <- as.dendrogram(hc_shop)
plot(dend_shop)
```

# Plot the dendrogram using Ward.D function

```
hc_shop2 <- hclust(gower_dist, method = "ward.D")
dend_shop2 <- as.dendrogram(hc_shop2)
plot(hc_shop2)
```

# Color the dendrogram

```
dend_colored <- color_branches(dend_shop2, k=2)
plot(dend_colored)
```

K-prototype

#attachting the library

```
library(stats)
library(dplyr)
library(ggplot2)
library(clustMixType)
library(cluster)
library(tidyverse)
library(dendextend)
library(wesanderson)
library(readxl)
```

# import the dataset

```
library(readxl)

excel_bestand_shop_2_zonder_transactie <- read_excel("~/Thesis/R/excel best
and shop 2 zonder transactie.xlsx",
+     sheet = "Gegevensset1")
View(excel_bestand_shop_2_zonder_transactie)
```

# Change the name of the dataset

```
mydata <- excel_bestand_shop_2_zonder_transactie
```

Data preparation

# Data type that is assigned to the dataset

```
> glimpse(mydata)
```

# Treating missing values

```
> sum(is.na(mydata))

0 missing values
```

# Outlier detection

```
 boxplot(mydata$Aantal)

boxplot(mydata$Opbrengst)
```

# Normalize the data using z-score

# drop the columns that are numerical

```
> notscaledata <- mydata[c(1,4,5,6)]
```

# scale the numerical data

```
scaled_opbrengst <- scale(mydata$Opbrengst, center=TRUE, scale=TRUE)
scaled_aantal <- scale(mydata$Aantal, center =TRUE, scale=TRUE)
```

```
# mutate the matrix into one

new_mydata <- mutate(notscaledata, scaled_aantal, scaled_opbrengst)
```

```
# convert data into characters into factors and numeric
new_mydata$`Type verkeer` <- as.factor(new_mydata$`Type verkeer`)
new_mydata$Apparaatcategorie <- as.factor(new_mydata$Apparaatcategorie)
new_mydata$Land <- as.factor(new_mydata$Land)
new_mydata$Gebruikerstype <- as.factor(new_mydata$Gebruikerstype)


new_mydata$scaled_opbrengst <- as.numeric(new_mydata$scaled_opbrengst)
new_mydata$scaled_aantal <- as.numeric(new_mydata$scaled_aantal)
```

```
# Wss plot to choose the optimum number of clusters[5]

data <- new_mydata
# Elbow Method for finding the optimal number of clusters
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max <- 15
data <- na.omit(data) # to remove the rows with NA's
wss <- sapply(1:k.max,
          function(k){kproto(data, k)$tot.withinss})
wss
plot(1:k.max, wss,
    type="b", pch = 19, frame = FALSE,
    xlab="Number of clusters K",
    ylab="Total within-clusters sum of squares")
```

```
# Silhouette score

pam_k2 <- pam(new_mydata, k = 2)
plot(silhouette(pam_k2))


pam_k3 <- pam(new_mydata, k=3)
plot(silhouette(pam_k3))
```

```
# Use map_dbl to run many models with varying values of k

sil_width <- map_dbl(2:10,  function(k){
+     model <- pam(x = new_mydata, k = k)
+     model$silinfo$avg.width
+ })


sil_df <- data.frame(
+     k = 2:10,
+     sil_width = sil_width
)


> ggplot(sil_df, aes(x = k, y = sil_width)) +
```

[5] Retrieved from: https://stats.stackexchange.com/questions/293877/optimal-number-of-clusters-using-k-prototypes-method-in-r

```
+     geom_line() +
+     scale_x_continuous(breaks = 2:10)
>
```

# The optimum number of clusters will be then 2

# K-prototype algorithm

```
model_data <- kproto(new_mydata, k=2)


clust_data <- model_data$cluster
segment_data <- mutate(new_mydata, cluster = clust_data)
count(segment_data, cluster)


summary(model_data)
table(model_data$cluster)
```

# Connecting to Tableau:

```
library(Rserve)
```

# exporting results into excel file

```
library(openxlsx)
library(rio)
export(segment_data3, "segment_data3_file.xlsx")
```

# Cluster analysis 2

Hierarchical clustering:

#attachting the library

```
library(stats)
library(dplyr)
library(ggplot2)
library(clustMixType)
library(cluster)
library(tidyverse)
library(dendextend)
library(FactoMineR)
```

# import the dataset

```
> library(readxl)
> excel_bestand_shop_2_zonder_transactie <- read_excel("~/Thesis/R/excel be
stand shop 2 zonder transactie.xlsx",
+     sheet = "Gegevensset1")
> View(excel_bestand_shop_2_zonder_transactie)
```

# Change the name of the dataset

```
mydata <- excel_bestand_shop_2_zonder_transactie
```

Data preparation

# Data type that is assigned to the dataset

```
> glimpse(mydata)
```

# Treating missing values

```
> sum(is.na(mydata))

0 missing values
```

# Outlier detection

```
 > boxplot(mydata$Aantal)


> boxplot(mydata$Opbrengst)
```

# convert characterics into factors

```
> mydata5$Gebruikerstype <- as.factor(mydata5$Gebruikerstype)
> mydata5$`Type verkeer` <- as.factor(mydata5$`Type verkeer`)
> mydata5$`Status van sitezoekfunctie` <- as.factor(mydata5$`Status van sit
ezoekfunctie`)
> mydata5$Land <- as.factor(mydata5$Land)
> mydata5$Sessies <- as.numeric(mydata5$Sessies)
> mydata5$`Gem. sessieduur` <- as.numeric(mydata5$`Gem. sessieduur`)
> mydata5$Gebruikers <- as.numeric(mydata5$Gebruikers)
> mydata5$`Pagina's/sessie` <- as.numeric(mydata5$`Pagina's/sessie`)
> mydata5$`Conversiepercentage van e-commerce` <- as.numeric(mydata5$`Conve
rsiepercentage van e-commerce`)
> mydata5$`Opbrengst per gebruiker` <- as.numeric(mydata5$`Opbrengst per ge
bruiker`)
> mydata5$Bouncepercentage <- as.numeric(mydata5$Bouncepercentage)
```

#scale the data using daisy function

```
gower_dist <- daisy(mydata5, metric = "gower")
```

# Plot the dendrogram using complete method

```
hc_shop <- hclust(gower_dist, method = "complete")
dend_shop <- as.dendrogram(hc_shop)
plot(dend_shop)
```

# Plot the dendrogram using Ward.D function

```
hc_shop2 <- hclust(gower_dist, method = "ward.D")
dend_shop2 <- as.dendrogram(hc_shop2)
plot(hc_shop2)
```

# Color the dendrogram

```
dend_colored <- color_branches(dend_shop2, k=2)
plot(dend_colored)
```

K-prototype

#attachting the library

```
library(stats)
library(dplyr)
library(ggplot2)
library(clustMixType)
library(cluster)
library(tidyverse)
library(dendextend)
library(wesanderson)
library(readxl)
```

# Normalize the data using z-score

# drop the columns that are numerical

```
> notscaledata <- mydata[c(1,2,3,4,5)]
```

# scale the numerical data

```
scaled_session <- scale(mydata$Sessies, center = TRUE, scale = TRUE)
> scaled_av_sessionduraction <- scale(mydata$`Gem. sessieduur`, center = TR
UE, scale = TRUE)
> scaled_users <- scale(mydata$Gebruikers, center = TRUE, scale = TRUE)
> scaled_pages_sessions <- scale(mydata$`Pagina's/sessie`, center = TRUE, s
cale = TRUE)
> scaled_conversion <- scale(mydata$`Conversiepercentage van e-commerce`, c
enter = TRUE, scale = TRUE)
> scaled_revenueperuser <- scale(mydata$`Opbrengst per gebruiker`, center =
TRUE, scale = TRUE)
> scaled_bounce <- scale(mydata$Bouncepercentage, center = TRUE, scale = TR
UE)
```

# mutate into one matrix

```
new_mydata <- mutate(notscaledata, scaled_session, scaled_av_sessionduraction, scaled_users, scaled_pages_sessions, scaled_conversion, scaled_revenueperuser, scaled_bounce)
```

# convert data into characters into factors and numeric

```
new_mydata$Gebruikerstype <- as.factor(new_mydata$Gebruikerstype)
new_mydata$`Type verkeer` <- as.factor(new_mydata$`Type verkeer`)
new_mydata$`Status van sitezoekfunctie` <- as.factor(new_mydata$`Status van sitezoekfunctie`)
new_mydata$Campagne <- as.factor(new_mydata$Campagne)
new_mydata$Land <- as.factor(new_mydata$Land)
new_mydata$scaled_session <- as.numeric(new_mydata$scaled_session)
new_mydata$scaled_av_sessionduraction <- as.numeric(new_mydata$scaled_av_sessionduraction)
new_mydata$scaled_users <- as.numeric(new_mydata$scaled_users)
new_mydata$scaled_pages_sessions <- as.numeric(new_mydata$scaled_pages_sessions)
new_mydata$scaled_conversion <- as.numeric(new_mydata$scaled_conversion)
new_mydata$scaled_revenueperuser<- as.numeric(new_mydata$scaled_revenueperuser)
new_mydata$scaled_bounce <- as.numeric(new_mydata$scaled_bounce)
```

# Wss plot to choose the optimum number of clusters[6]

```
data <- new_mydata
# Elbow Method for finding the optimal number of clusters
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max <- 15
data <- na.omit(data) # to remove the rows with NA's
wss <- sapply(1:k.max,
          function(k){kproto(data, k)$tot.withinss})
wss
plot(1:k.max, wss,
    type="b", pch = 19, frame = FALSE,
    xlab="Number of clusters K",
    ylab="Total within-clusters sum of squares")
```

# **Silhouette score**

```
pam_k2 <- pam(new_mydata, k = 2)
plot(silhouette(pam_k2))


pam_k3 <- pam(new_mydata, k=3)
plot(silhouette(pam_k3))
```

# **Use map_dbl to run many models with varying values of k**

```
sil_width <- map_dbl(2:10,  function(k){
+     model <- pam(x = new_mydata, k = k)
+     model$silinfo$avg.width
+ })
```

[6] Retrieved from: https://stats.stackexchange.com/questions/293877/optimal-number-of-clusters-using-k-prototypes-method-in-r

```
sil_df <- data.frame(
+     k = 2:10,
+     sil_width = sil_width
)
```

```
> ggplot(sil_df, aes(x = k, y = sil_width)) +
+     geom_line() +
+     scale_x_continuous(breaks = 2:10)
>
```

# The optimum number of clusters will be then 2

# K-prototype algorithm

```
model_data <- kproto(new_mydata, k=2)
```

```
clust_data <- model_data$cluster
segment_data <- mutate(new_mydata, cluster = clust_data)
count(segment_data, cluster)
```

```
summary(model_data)
table(model_data$cluster)
```

# Connecting to Tableau:

```
library(Rserve)
```

# exporting results into excel file

```
library(openxlsx)
library(rio)
export(segment_data3, "segment_data3_file.xlsx")
```