Master's Thesis

Customer Segmentation With Unsupervised Learning

Onurhan Irkin
04.02.2019

Project Management and Data Science (MPMD)

1st Supervisor: Prof. Dr. Tilo Wendler
2nd Supervisor: Cornelia Niklas

Hochschule für Technik und Wirtschaft Berlin

University of Applied Sciences

Index

1 Introduction 1
1.1 Customer Segmentation 1
1.2 Unsupervised Learning: Clustering Methods 1
1.3 Objective of the Thesis 2
1.4 Outline of the Thesis 2

2 Theoretical Background and Literature Review 3
2.1 Customer Segmentation 3
2.1.1 Purpose and Goals of Segmentation 3
2.1.2 Different Segmentation Bases 4
2.1.3 RFM Analysis 5
2.2 Unsupervised Learning Technics for Segmentation 7
2.3 K-Means Clustering 8
2.3.1 Algorithm 9
2.3.2 Selecting the Best K 11
2.3.3 Selecting Initial Centroids: k-means++ 12
2.4 Agglomerative Hierarchical Clustering 13
2.4.1 Defining the Proximity Between Clusters 14
2.4.2 Agglomerative Hierarchical Clustering Algorithm 15
2.5 Density-based Spatial Clustering of Applications with Noise (DBSCAN) . 16
2.5.1 Algorithm 17
2.5.2 Choosing Parameters 18
2.6 Clustering using Gaussian Mixture Models 19
2.6.1 Gaussian Mixture Models 20
2.6.2 Expectation-Maximization Algorithm 23
2.6.3 Model Selection Criterion 24
2.7 Evaluation of the Clustering Methods 25

3 Implementation of Unsupervised Learning Methods to Identify Customer Segments 27
3.1. Business Overview 27
3.2 Data Understanding 28
3.3 Data Preparation 33
3.4 K-Means Clustering 37
3.5 Agglomerative Hierarchical Clustering 41
3.6 Density-based Spatial Clustering of Applications with Noise (DBSCAN) . 44
3.7 Clustering using Gaussian Mixture Models (GMM) 46
3.8 Comparison of the Methods 48

Summary and Future Directions 51

List of literature 53

Statutory Declaration 55

List of figures

Fig. 1: Focus on market-driven strategy(Wind and Bell, 2008) 4
Fig. 2: Two-feature space with initial centroids assigned 10
Fig. 3: Cluster centroids updated with K-means algorithm 11
Fig. 4: Elbow criterion for selecting K 12
Fig. 5: Hierarchical clustering represented as dendrogram and venn diagrams 13
Fig. 6: Hierarchical clustering output with two different cluster number selection 14
Fig. 7: Graphical definitions of proximity approaches (Han, Kamber and Pei, 2011) 15
Fig. 8: Data set with non-linear high density regions 16
Fig. 9: Object labels with density-based clustering approach (Han, Kamber and Pei, 2011) 17
Fig. 10: Noises and clusters based on k-dist graph (Ester *et al.*, 1996). 18
Fig. 11: K-means clustering assignments on two different data set 19
Fig. 12: An example of Gaussian density for a single variable (figure on the left) and a multivariate Gaussian density over two variables (figure on the right). 21
Fig. 13: Example BIC and AIC values for different number of components 25
Fig. 14: Product offered in different online retail stores 27
Fig. 15: Customer numbers against revenue by country 31
Fig. 16: Revenue and purchase quantity by month 32
Fig. 17: Total sales by product types 32
Fig. 18: Distributions and pairwise correlations of R,F,M features 35
Fig. 19: Data set on three-dimensional space before and after logarithmic transformation 36
Fig. 20: Distributions and pairwise correlations of transformed R,F,M features 37
Fig. 21: Within-cluster sum of squared distances by number of clusters 38
Fig. 22: Results of (a) 3-means, (b) 5-means and (c) 7-means clustering 39
Fig. 23: R,F,M features of clusters compared by different number of clusters generated 40

List of tables

Table 1: Sources influencing customer's decision making process 4
Table 2: A customer base with assigned RFM scores 7
Table 3: Parameter definitions of DBSCAN 17
Table 4: Randomly selected entities from data set 28
Table 5: Records with non-available CustomerID 30
Table 6: Sales by country..... 30
Table 7: Customers with maximum and minimum revenue contribution 31
Table 8: Non-purchase transactions in data set 32
Table 9: Amount of impractical entities 33
Table 10: Manual entities 34
Table 11: Sample from transformed customer data set 34
Table 12: Summary statistics of customer RFM data set 35

1 Introduction

Organizations develop long term marketing strategies with the object of being in a better position in the market especially in highly competitive business areas. Developing an effective marketing strategy, targeting customers accurately with a suitable product mix and pricing necessitates understanding the customer base as a threshold matter. Understanding the customers is very important for all companies from small businesses to largest retailers to be able to develop long term profitable marketing and sales strategies. Each one of the customers in the customer base of a company reflects different characteristics and have possession of different motivations for making purchase or having the service of the organization. This infinite variety of customers cause a highly complex and heterogeneous customer base which is very hard to understand. Marketing experts are using a lot of techniques to solve this complexity with identifying more homogeneously structured customer based. Customer segmentation is the most widely used method for generating a solution to this heterogeneity with identifying more homogeneous subgroups of customers by different features and attributes. Developed machine learning techniques under unsupervised learning category argues the same objective of identifying homogeneities in complex data sets which serves customer segmentation analysis inherently.

1.1 Customer Segmentation

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits. Customer segmentation analysis can be made by using number of different categories of basic customer features based on the objective and available toolset of the study. Most common segmentation categories are geographic, demographic, psychographic and behavioural segmentation. The selection of segmentation base depends on the expectation from the analysis and the available customer information. The subject data set of the thesis uses a financial transaction data of an online retail business customers. Based on the characteristics of the given data set, the segmentation analysis can be categorized as behavioural segmentation. Due to the fact that this type of segmentation's objective is identifying customer segments based on purchase behaviours of existing customers, given data set needs to be transformed to a new data set which provides customer level purchase related attributes. A popular marketing phenomena called RFM analysis inspired the transformation of transaction data into a new customer level data set with useful features about each customer's historical purchase behaviour. Three major attributes of the RFM analysis, which are purchase recency, frequent and monetary are calculated by historical transactions and attached to each customer in customer base of the company. Hereby, the objective of homogeneity in customer segments can be identified with respect to purchase habits of existing customers.

1.2 Unsupervised Learning: Clustering Methods

Unsupervised Learning is a branch of machine learning which consist of methods that are used for number of purposes such as anomaly detection, dimensionality reduction, categorization and clustering. The significant difference between unsupervised learning methods and supervised learning methods makes the validation of unsupervised learning methods a more complex process requiring heuristic approaches combined with statistical

methods. The major dissimilarity is that these methods were designed to be utilized when there is no labelled data set to supervise applied algorithms. The unsupervised learning methods does not necessitate a data set in which the output variable to be predicted is already labelled. Thus, this nature of this category of machine learning serves in case of problems such as detecting hidden, unknown patterns in the data set. In contrast to commonly used supervised learnings, these methods do not promise to conclude a globally optimum solution for the given problem. Instead, they are based on the principle of generating locally optimum solutions to use as a bae for future analysis or decision making processes. Thus, these methods can also be considered as a procedure of explanatory analysis. Thus, it requires heuristics, base knowledge, a good understanding of the core problem and multiple number of experiments. In addition to anomaly detection and dimensionality reduction purposes, clustering methods are widely used for problems like customer segmentation where the objective is understanding the data set better after all. Thesis argues four clustering techniques to explore customer segments which are K-means clustering, hierarchical clustering, density-based spatial clustering and expectation-maximization clustering using Gaussian Mixture Models.

1.3 Objective of the Thesis

The main objective of thesis is identifying homogeneous customer segments using different unsupervised learning methods. By the very nature of unsupervised learning methods applied and segmentation analysis, instead of arguing a globally optimum structure of customer segments, the basic expectations from the final output of the analysis are improving a better understanding of customers with providing alternative solutions to homogeneity of customer base. Some methods necessitate number of assumptions to be made, which can result with high level of complexity of suggested models which leads to unadaptable solutions. Thus, the complexity also considered as an important indicator of learning processes adopted.

1.4 Outline of the Thesis

The thesis consist of two main parts.

Chapter 2 argues the customer segmentation, objectives and traditional approaches together with four unsupervised learning methods that are applied to given data set in Chapter 3.

Chapter 3 includes exploratory analysis of given data set and practice of four unsupervised learning methods.

The last part following Chapter 3 summarizes the findings from applied methods.

2 Theoretical Background and Literature Review

The main objective of the thesis is exploring customer segments through the instrumentality of unsupervised learning methods. Thereby, this chapter discusses the concept of customer segmentation and several unsupervised learning methods which are potentially practicable for exploring customer segments in a heterogenous customer base.

2.1 Customer Segmentation

According to marketing managers and marketing academics, customer segmentation, also referred as market segmentation, is one of the most widely used marketing science tools having largest impact on decision-making processes in marketing field (Roberts, Kayande and Stremersch, 2014). Segmentation has been first argued as a marketing strategy by Smith (1956). By definition, customer segmentation is dividing a customer base, a mass market into smaller groups. Smith (1956) defines the segmentation as viewing a heterogenous market as smaller number of homogenous markets. As it has been argued many times before, all markets are heterogeneous groups (Wind and Bell, 2008). Customer segmentation assists business experts to split the potential customers into smaller groups, or segments, where these segments are homogeneous within and heterogenous between each other (Grigsby, 2016). Therefore, each customer segment is a sub-set of customers who share similar customer characteristics with other customers in the same segment and ideally, different characteristics with customers in other segments. Segmentation has always been viewed as a crucial marketing concept since 1960s, however, technological advances, globalization and high business competition for growing in new markets have increased the demand for more localized marketing efforts. Thus, these changes are leading both a higher level of localization, i.e., ‘one-on-one marketing’ and more advanced applications of segmentation methods (Wind and Bell, 2008). While there are advanced statistical analysis tools that can be used to perform segmentation, it is also possible to accomplish this with more common sense or traditional approaches, i.e., segmentation based on RFM (recency, frequency and monetary) scores. In practice, segmentation analysis and the method are initiated based on the organization’s knowledge base, assumptions, available resources and constraints.

2.1.1 Purpose and Goals of Segmentation

The general purpose of marketing efforts should be underscored to understand the function and the purpose of customer segmentation clearly. The purpose of marketing in general terms is providing appropriate offers to customers which are suited to satisfy their needs and desires (Dolnicar, Grün and Leisch, 2018). To be able to accomplish this, a logical sequence of activities is required with a clear objective and plan. Customer segmentation analysis enables organizations to reflect fundamental strategic business questions and forces organization to seek answers and practice required marketing activities accordingly. Fig. 1 demonstrates a market-driven (or customer-driven) approach where segmentation could be seen as a tool for answering the very core questions to identify potential customers and to develop an understanding about heterogeneity of the customer base and customer needs (Wind and Bell, 2008). This practice forces organizations to understand what the current status is and decide what sort of actions can organization take both in short term and in the future.

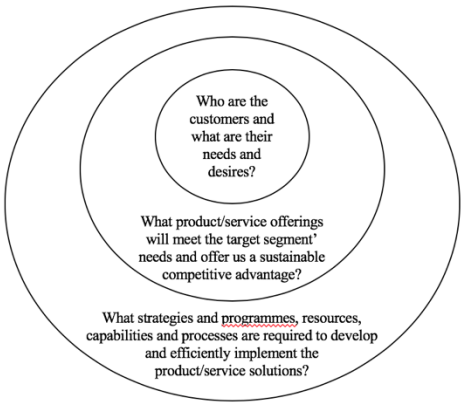


Fig. 1: Focus on market-driven strategy(Wind and Bell, 2008)

The different needs and sensitivities of potential customers lead them to show different behaviors while purchasing a product or a service. The customer behavior is associated with multiple different factors and simultaneous motivations and it is very complex. The three main usage of segmentation analysis, that Grigsby (2016) addresses, provide a general approach to cope with the complexity of customer behavior: (1) Finding homogeneous members in customer base; (2) Making modelling more efficient; (3) Using marketing strategy to target each segment differently. Customer segmentation analysis serves the purpose of identifying homogeneous subgroups in customer base.

2.1.2 Different Segmentation Bases

The heterogeneity in customer base leads varying customer preferences which are usually driven by aggregation of customer's expectations across various product or service attributes such as price, time savings, durability. Diverging purchase decisions are driven by different underlying sources, as summarized in Table 1. The purchase decision is the result of trade-off decisions across all attributes. That is to say, understanding the expectations of each customer and taking marketing actions accordingly, is a complex problem to solve with a huge customer base.

Source	*Description*
Individual differences	A person's stable and consistent way of responding to the environment in a specific domain
Life experiences	An individual's life experiences capture events and experiences unique to their life that have a lasting impact on the value and preferences they place on products and services, which, in turn, affect preferences independent of individual differences
Functional needs	An individual's personal decision weightings across functional attributes based on their personal circumstances
Self-identity/image	Customers actively seek products that they feel will support or promote their desired self-image
Marketing activities	Firms' attempts to build linkages between their brands and prototypical identities or meanings

Table 1: Sources influencing customer's decision making process

There are many factors that leads varying customer characteristics. Variety of customer segmentation approaches offer different ground customer categorization options to enable management of this complexity and heterogeneity. Segmentation approaches diverge from one another based on different group of variables used to identify similarities in markets. These approaches can be categorized as follows:

- *Geographic segmentation* is grouping customers on the basis of geographical units. Variables such as country, region, city, urban, suburban, rural population densities, climate are used to divide markets into specific groups. With the output of segmentation analysis, services, products or marketing efforts can be localized to meet with desires driven by geographical purchasing patterns of customer base.

- *Demographic segmentation* calls for dividing large populations into smaller groups based on age, family size, gender, income, education level, religious preferences, ethnicity, nationality and occupation. Demographic factors are quite frequently used for segmenting customers since they are highly influential on customer needs and it is relatively easy to measure.(Kotler *et al.*, 2017)

- *Psychographic segmentation* is dividing market into segments on the basis of social class, lifestyle and personalities (Kotler *et al.*, 2017). Psychographics are used to identify inner drivers, feelings behind customer behavior. This type of segmentation is particularly helpful for marketers to execute targeted advertising and promotional campaigns.(McDonald, 2013)

- *Behavioral segmentation* serves to understand customer based on customer's historical engagement with the service or product so that people with similar purchasing behavior can be grouped into same segments. Variables such as customer loyalty, usage rate, buying occasions, benefits sought are used for this type of segmentation. In practice, customer transaction and marcomm response data is used to identify what is important for customers. Behavioral segmentation leads further marketing activities such as targeting, optimization of promotions, marketing channel preferences, customer journey analysis and product selection. (Grigsby, 2016)

The decision of which customer segmentation approach to apply, mainly linked to the cost and availability of customer attributes. For this study, behavioral segmentation is used with the given financial transaction data due to the fact the features in the data set enables to extract customer attributes related to purchase behaviours.

2.1.3 RFM Analysis

Another popular traditional marketing concept, which is called RFM (recency, frequency, monetary) analysis is also carried out by marketers using behavioral variables mentioned together with behavioral segmentation in prior chapter. This method can be seen just as a traditional way of identifying customer segments but also can be used to identify new variables which serves to quantify customer's buying characteristics in order to run more precise behavioral segmentation. The analysis based on a popular marketing phenomenon called as Pareto Principle. The concept was proposed more than 75 years ago for direct marketers (Grigsby, 2016). The principle also referred as "80/20 rule" and states that 80% of the revenue comes from 20% of the customers who would be more desirable customer profile for organization in terms of profitability. This classification also can be seen as

generating one customer segment such that the members of the segment have a higher tendency of bringing revenue compared to the rest of the members of customer base.

RFM (recency, frequency, monetary) serves as the quantitative technique that is applied to filter these best existing customers with evaluating three major variables generated based on customer purchase data. Blatterberg, Kim and Neslin disclose these variables as follows (2008);

- *Recency(R)* represents how recently the customer has purchased. As it has been frequently suggested, there is a relationship between the recency and the customer response. Although different relationship types, such as U-shaped, has been also discussed in specific business areas, relationship is assumed to be negative in traditional applications of RFM analysis. The variable can be measured as the total time elapsed between the last purchase of the customer and the time of analysis.
- *Frequency(F)* represents how often the customer purchase in the given time period. There are various but simple ways to measure frequency depending on the business case. Most frequently, it is measured as the total number of purchases of the customer. Alternatively, number of purchase occasions divided by the duration of being customer can be used as a frequency measure. Similar to the recency variable, it is assumed that there is an exact relationship between frequency and customer response. In other words, the customers who purchase more frequently, tent to have a higher probability of responding marketing efforts or purchase again.
- *Monetary(M)* represents how much money does the customer spend on previous purchases. The variable can be measured either as the total amount of money spent by customer in the given time period or the average expenditure per order.

Once the recency, frequency and monetary data is obtained for existing customers, in order to run RFM analysis, customers are assigned a ranking number of one, two, three, four or five where five is the best ranking for each variable. Below is a step by step representation the usual RFM algorithm that has a number of assignment and scoring steps for each variable:

1. Sort the database by the recency of transactions in ascending order.
2. Assign the top 20% with R=5 and on the down to the bottom 20% with R=1.
3. Re-sort the data in descending order based on the purchase frequency variable.
4. Assign the top 20% with F=5 and on the down to the bottom 20% with F=1.
5. Re-sort the data in descending order based on the monetary variable that is amount of money spent by customer.
6. Assign the top 20% with M=5 and on the down to the bottom 20% with M=1.
7. Create a new variable T = R+F+M
8. Re-sort the data in descending order by column T representing total RFM score of each customer ranging from 15 to 3.
9. The top 20% gets a RFM score five and the bottom 20% are assigned RFM = 1 as shown on a sample customer data on Table 2.

CustomerID	*Recency*	*Frequency*	*Monetary*	*R*	*F*	*M*	*RFM score*
15311	1	2379	28607	5	5	5	5
17511	3	963	21595	5	5	5	5
13408	2	478	8442	5	5	5	5
13767	5	368	5563	4	5	4	5
12583	3	232	2872	5	4	4	4
14688	8	327	2705	4	4	4	4
16029	39	241	7350	2	4	5	4

12431	36	233	2743	3	4	4	4
17809	17	61	2120	3	3	3	3
13047	32	172	2010	3	3	3	3
15291	26	103	1856	3	3	3	3
13705	10	28	400	4	2	2	3
16098	88	67	1708	2	3	2	2
17924	12	31	177	4	2	1	2
17420	51	30	606	2	2	2	2
16250	262	24	1991	1	1	3	2
13748	96	28	268	2	2	1	1
18074	374	13	440	1	1	2	1
15100	334	3	56	1	1	1	1
13747	374	1	12	1	1	1	1

Table 2: A customer base with assigned RFM scores

The customers carrying 5 RFM score, are the best customers from the organization's perspective(Grigsby, 2016). Although the analysis is based on customers' buying behavior data, the methodology offers an approach that is entirely focused on company's profit interest rather than customer's desires. In this respect, this traditional approach might not show parallelism with customer centricity that is offered by behavioral segmentation approach. Nevertheless, this well recognized approach can still be used to understand customers. More precisely, R, F and M scores would be helpful as base variables for behavioral segmentation or can be used for evaluation of segmentation results to understand if the customers in the same segments have similar tendencies to response further marketing activities.

2.2 Unsupervised Learning Technics for Segmentation

Machine learning algorithms are commonly divided into two major categories based on the way of processing the given data: Supervised or unsupervised. Supervised learning methods are applied to solve problems when the given training data includes the response measurement or outcome, which is aimed to be predicted, together with the input measurements either in a form of discrete labels or continuous values. Regression, support vector machines, decision trees, random forest are commonly used technics in this category(Sarkar, Bali and Sharma, 2018). However, it is not always possible to have advantage of labelled data or the data might not have an associated response variable at all. In this case, unsupervised learning methods are used to solve such problems. These methods are called unsupervised because the algorithms are not concerned about predicting the outcome of future events with supervision of past outcomes. Instead, they concern with obtaining insightful information to better understand the patterns and key features that are not interpretable from raw data. Most unsupervised learning technics tackle following problems as Igual and Seguí (2017) summarizes the problems that most unsupervised learning technics are used to tackle as follows:

- *Clustering* aims to divide the set of data into smaller groups based on a similarity criterion.
- *Dimensionality reduction* has a goal of reducing the dimensionality of the data with technics such as Principal Component Analysis and nonnegative matrix factorization.
- *Outlier detection* aims to identify abnormal events in data according to a certain criterion.
- *Novelty detection* aims to identify the abnormalities in future data.

From many available unsupervised learning methods, clustering technics are commonly used for segmentation by the nature of the segmentation problem that is finding subgroups with similarity within a heterogeneous data. There are number of different clustering methods which can be classified under following approaches (Sarkar, Bali and Sharma, 2018):

- *Centroid based clustering:* Each cluster in the dataset is formed by one central vector, which is called centroid, and observations, which are assigned to cluster such that each observation belongs to the cluster of nearest centroid. Different proximity approaches are available for measuring the distances within dataset. K-means and K-medoids are commonly used methods classified under this category.

- *Hierarchical clustering:* Creates a structured hierarch of nested clusters from un-structured data with either repetitively dividing or combining clusters such that the smaller cluster size becomes, the more overall within-cluster homogeneity is. The operations are performed with either a top-down or bottom-up approach while building the hierarchy and they are called agglomerative or divisive, respectively.

- *Distribution based clustering:* Identifies clusters based on how the observations in the dataset are likely to belong to the same distribution. The most common method is known as Gaussian mixture models using expectation-maximization algorithm.

- *Density based clustering:* considers clusters as dense regions in multidimensional space which are separated by regions of lower density of data points. In other words, cluster is defined as a set of density connected points. DBSCAN and OP-TICS are most common techniques under this category.

There are number of techniques populated based on common methods and new methods have been proposed under each category. Following chapters 2.3, 2.4, 2.5 and 2.6 explain four clustering techniques from above four major categories. In implementation phase, each technique is tailored with respect to given dataset's requirements.

2.3 K-Means Clustering

K-means clustering has been the most widely used clustering tool by scientific researchers and industry experts as it has been revealed by Pavel Berkhin's survey (Berkhin, 2002). The core algorithm has been proposed by Lloyd as a solution for vector quantization (Lloyd, 1982). The aim of the algorithm is solving an optimization problem: Partitioning the observations into *k* number of groups (clusters) with minimizing within-cluster variation. The clusters are defined by the centroids, which are the points at the center of the clusters, and the cluster of each observation is defined by the most closely located centroid. The distance of observations to these centroids are measured with different measure proximations; hence K-means clustering is referred as distance-based or centroid based clustering.

Assuming that the desired output is partitioning *n* observations into *K* clusters where $S = \{S_1, S_2, \dots S_k\}$ are sets of observations in the dataset. Meaning that, if *i*th observation is in *k*th cluster, then $x_i \in S_k$. The algorithm has to satisfy two constraints (James *et al.*, 2013):

1. Each observation in the dataset should belong to at least one cluster;

$$S_1 \cup S_2 \cup \ldots \cup S_k = \{x_1, x_2, \ldots, x_n\}$$

2. Each observation should belong to only one cluster;

$$S_k \cap S_{k'} = \emptyset$$

In consideration of these qualifications, the optimization problem can be formulated as following:

$$\arg\min_{S_i \in S} \sum_{i=1}^{k} W(S_i)$$

$W(S_i)$ in above formulation denotes the within-cluster variation(James *et al.*, 2013). Therefore, the formula argues that the observations are partitioned into k cluster with minimum possible sum of within-cluster variations across all clusters. In order to solve this problem, these variations need to be measured and minimized. This measurement can be done in many different ways using different proximity functions such as Manhattan, cosine, Bregman divergence and squared Euclidean, which can be seen as the most popular method practiced with K-means clustering.

2.3.1 Algorithm

Considering squared Euclidean distance as the proximity measure for within-cluster variation, the objective function is derived as minimizing a distance function:

$$\sum_{i=1}^{k} \sum_{x \in S_i} dist(x, c_i)^2$$

where k is the number of clusters, c_i is the centroid of the cluster i and "*dist* is the standard Euclidean distance between two objects in Euclidean space" (Tan *et al.*, 2018). All observations are interpreted as d-dimensional vectors where d is the number of variables in the dataset. In this case, the centroid is also a d-dimensional vector where each dimension is respective dimension's mean of the data points in the cluster. In other words, centroid is the mean of all observation belonging to the respective cluster. To measure the within-cluster variation, sum of squared deviations between each observation in the cluster and the center of the cluster is calculated.

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} |x - \mu_i|^2$$

The traditional k-means clustering algorithm follows the steps given below:

1. Start with selecting k points as initial centroids.
2. Assign each observation to the closest centroid using the Euclidean distance function.
3. Calculate the mean of each cluster again to be the new centroid.
4. Repeat the above two steps until clusters no longer alternate.

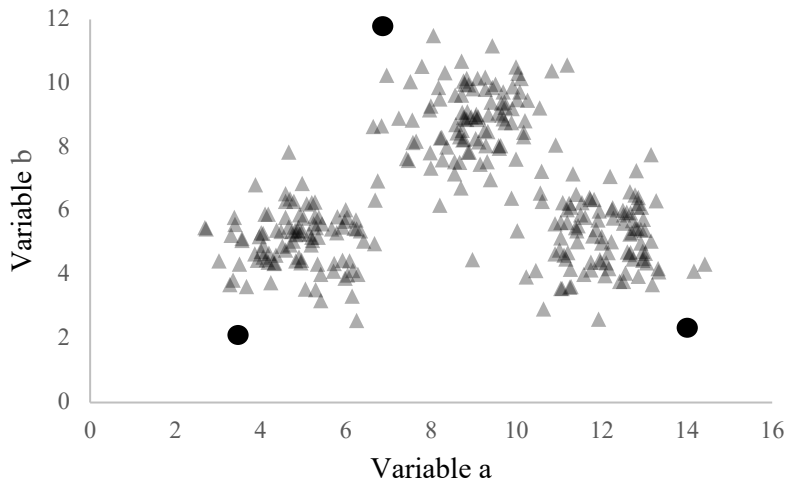


Fig. 2: Two-feature space with initial centroids assigned

K points, which are likely to be in different clusters, should be selected to initiate the algorithm. Fig. 2 demonstrates a given set of observations with two variables on a two-dimensional Euclidean space and arbitrarily selected 3 initial centroids marked with squares. With a heuristic approach, the observations on the example dataset likely to be in 3 different groups. The best *k* value varies depending on case-specific requirements and the properties of given dataset. The main algorithm starts with "arbitrarily" selecting *k* initial centers and there is no theoretical method that can give the optimal number of clusters in advance(Arthur and Vassilvitskii, 2007). However, there are still some methods that would help determining the number of clusters and selecting the initial centroids in a more specific way. They are explained in section 2.3.3 and 2.3.4 respectively. Recall the algorithm process, after arbitrarily selecting the initial centroids, the second step is assigning the observations to closest clusters measuring Euclidean distances between centroids and observations. The distance function for an observation to centroid c can be calculated with

$$dist(x, c_i)^2 = \sum_{j=1}^{d} (x_j - c_{ij})^2$$

where $(x_j - c_{ij})$ is the arithmetic difference between the recorded variable j value of observation x and variable *j* value of centroid of cluster *i*. Each observation is assigned to the closest cluster, where "closest" means the cluster with centroid having the shortest Euclidean distance to observation. After assigning the observations to clusters, the next step is calculating the new centroids. The centroid is the mean of all observations assigned to respective cluster such that;

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

More precisely, the mean for feature *j* in cluster S_i is

$$\mu_j = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i^{(j)}$$

where $x_i^{(j)}$ is the value of *jth* feature of observation x_i (James *et al.*, 2013). Applying the 2nd and 3rd steps of the algorithm to given dataset on Fig. 2, Fig. 3 shows the new centroids that has been assigned. The algorithm does no guarantee the optimal solution to the problem (Hartigan and Wong, 1979). As it is stated on 4th step, step 2 and step 3 can be iterated, until the clusters no longer changes, or another criterion is met. By the nature of the algorithm, each time the observations are reallocated on the Euclidean space, the cluster-wise variation will be decreased so when reallocations does not improve the result anymore, then we can claim that the local optimum has been reached and the algorithm can be stopped (James *et al.*, 2013).

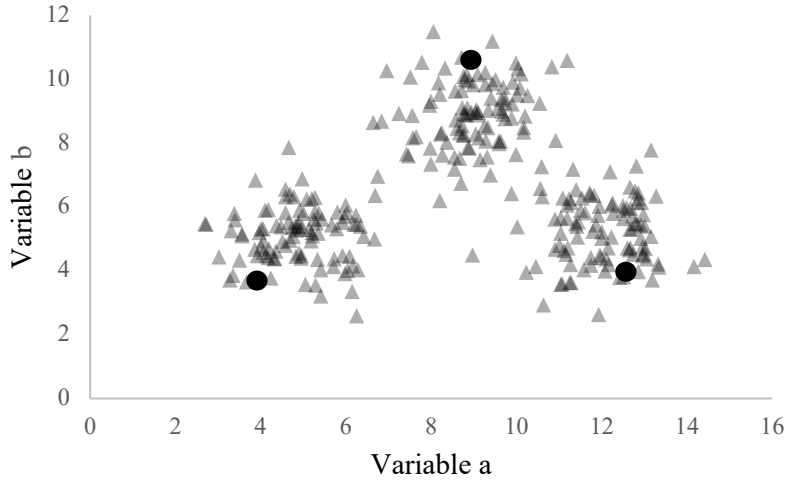


Fig. 3: Cluster centroids updated with K-means algorithm

The main algorithm does not require a selection process for the initial number of clusters however, the result of the algorithm depends on the initial k. For the sample data demonstrated on Fig. 2 & Fig. 3, It is easy to come up with an idea about the number of clusters. Also depending on the use case, the number of clusters might be predefined by the other factors. However, considering a situation such that there is no ground knowledge and the given multidimensional dataset is visually uninterpretable on an n-dimensional Euclidean space, it wouldn't be similarly easy to make a meaningful decision about the number of clusters beforehand. Following chapter explains a more analytical method for selecting the number of clusters.

2.3.2 Selecting the Best K

Since the algorithm aims to find local optimums rather than a global optimum, the initial predeterminations and cluster assignments of observations diversify the final result of the analysis(James *et al.*, 2013). One of the initial decisions is the number of clusters which is traditionally defined arbitrarily or based on the ground knowledge and expectation. K require the number of clusters in a data set as the parameter for initializing the process. The experiments should be started with defining a K or multiple candidate K values to interpret the outputs comparatively. Thereby, it is desirable to estimate a group of reasonable cluster amounts before a clustering algorithm is used to derive detailed information. A useful visual approach for estimating the best possible number of clusters is referred as elbow method, which is a tool for finding an appropriate number of clusters with a visual interpretation of total squared distances between observations. The objective is choosing a *k* such that, increasing the number of clusters would not increase the percent of variance explained significantly. A very optimistic diagram demonstrated on Fig. 4 represents that

the graph of within-cluster variance bends on k=3, i.e., elbow point, and does not significantly decreases after this point and it can be seen as an ideal number of clusters to explain the dataset.

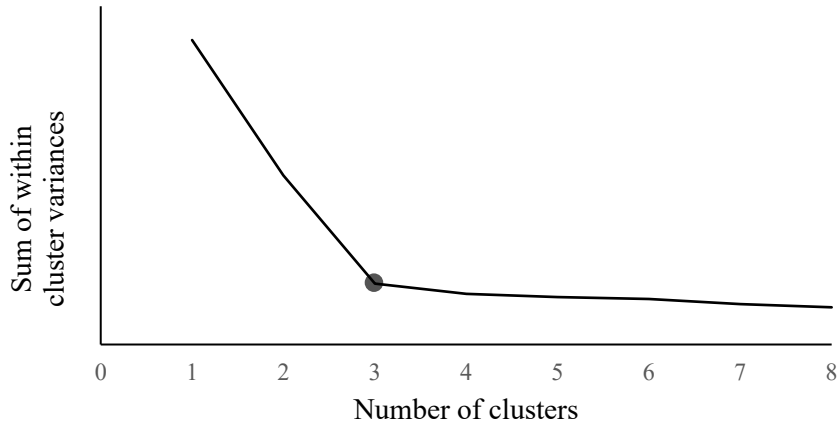


Fig. 4: Elbow criterion for selecting K

The criteria is based on the distance of cluster points to cluster centres without including a measurement for distances to data points in other clusters. The method is useful for having an initial idea about the output of K-means clustering. However, since both criteria and the method do not promise a globally optimum solution, multiple iterations can give more information about the hidden patterns. Another criteria, which is called Silhouette scoring, takes the distance to neighbouring clusters into account in addition to within-cluster distances (Rousseeuw, 1987). Silhouette scores can also be used to find out a reasonable number of clusters. The method is detailed in chapter 2.7 Evaluation of Clustering Methods because the visual interpretation of the technique can be used to compare results of different clustering outputs as well.

2.3.3 Selecting Initial Centroids: k-means++

The traditional K-means method starts with a arbitrarily assigned centroids at the beginning(James *et al.*, 2013). Arthur and Vassilvitskii (2007) proposed a more advanced way of initializing the k-means algorithm, which increases the performance and accuracy of the main k-means. The combination of this initialization logic with the same cluster assignment and centroid calculation steps of original k-means algorithm is called k-means++. The process begins with randomly selecting an observation as the first initial centroid. For the other centroids, the probability of an observation been selected as a centroid is proportional to the squared distance of the observation to the closest centroid. Thus, the algorithm assigns initial centroids that are far away from each other with large squared distances. Assuming that $X = \{x_1, x_2, \ldots, x_n\}$, Arthur and Vassilvitskii (2007) summarizes the combined algorithm as follows:

1. Choose an initial centroid c_1 uniformly at random from the dataset.
2. Choose the next center c_i, selecting $c_i = x_i \in X$
   with probability of $\frac{D(x_i)^2}{\sum_{x \epsilon X} D(x)^2}$.
3. Repeat step 2 until choosing total k centroids.
4. Proceed with the standard assigning and distance calculation steps of k-means algorithm.

K-means++ approach is especially useful in case of technical limitations such as maximum number of iterations of the algorithm with a large data set.

2.4 Agglomerative Hierarchical Clustering

Hierarchical clustering is another widely used clustering technique. In contrast to K-means clustering, this method does not require a predefined parameter such as the number of desired clusters. Broadly, the method identifies a set of nested clusters and a hierarchical representation of these clusters visualized with a dendrogram which is used for interpreting the output of the analysis (James *et al.*, 2013). There are two main approaches for applying hierarchical clustering:

- *Divisive method:* A top-down approach which starts with one cluster including all observations and divides clusters into pair of clusters until each observation becomes one individual cluster.

- *Agglomerative method:* A bottom-up approach starts with treating each observation as a separate cluster. In each step, clusters are merged based on the similarity criterion until one all-inclusive cluster is reached.

Both approaches are applied with defining a notion of cluster proximity to measure the similarity between clusters to merge or divide clusters in each step depending on preferred method (Tan *et al.*, 2018). The similarity measurement in hierarchical clustering is similarly can be found with different distance measurements such as Euclidean distances. Most frequently used approaches for measuring the proximity between two clusters are MIN or Single Link, MAX or Single Link, Group Average and alternatively Ward's method. In depth explanations of these techniques can be found in next chapter.

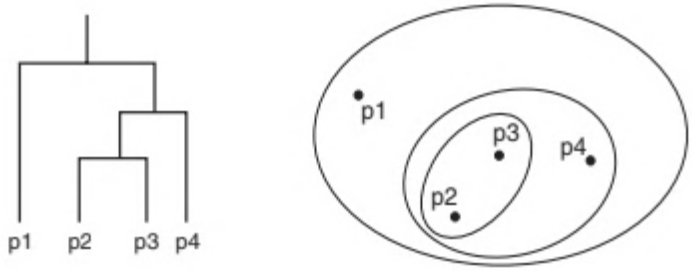


Fig. 5: Hierarchical clustering represented as dendrogram and venn diagrams

The agglomerative version of hierarchical clustering is the most common type of hierarchical clustering (James *et al.*, 2013). A hierarchical clustering is often graphically demonstrated using a dendrogram, which displays the hierarchical relationships between clusters and the order that the clusters are merged or divided (Tan *et al.*, 2018). Fig. 5 displays a simple dendrogram obtained with a hierarchical clustering and a representation of the relationships between clusters as nested venn diagrams. Each single leaf of the dendrogram represents an observation. As it can be seen on the dendrogram, the leaves representing p2 and p3 joins together and fuse into a new branch. This implies that these observations have more similarity than the other observations such that they form a new level of cluster. Moving higher branches of the dendrogram, all branches fuse with other leaves or branches until generating an all-inclusive cluster at the top of the diagram. "Observations that fuse at the bottom of the tree are quite similar to each other, whereas observations that fuse close to the top of the tree will tend to be quite different." (James *et*

al., 2013). Thus, the similarity relationships between observations can be explained with interpreting the branches of the dendrogram.

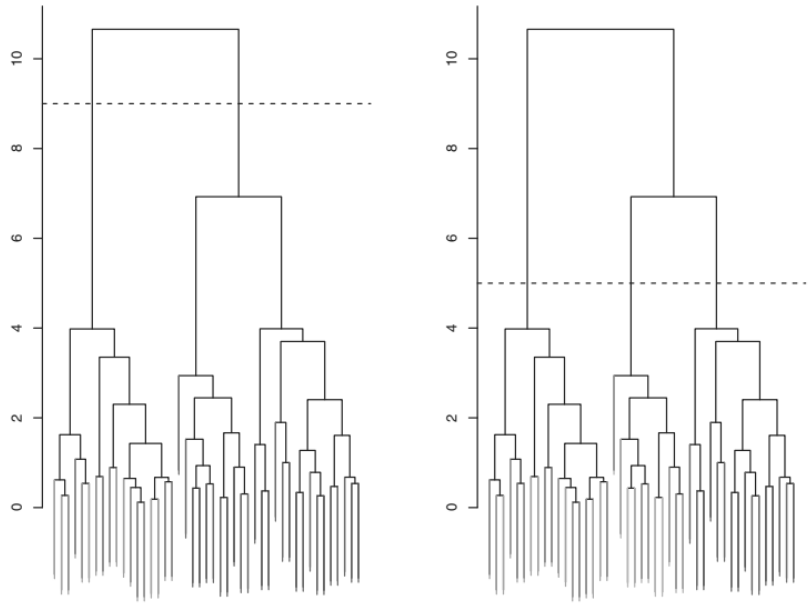


Fig. 6: Hierarchical clustering output with two different cluster number selection

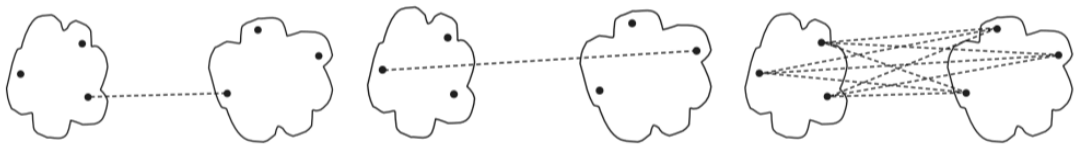
Each vertical branch of the dendrogram can be interpreted as one cluster where all the observations branching out are the members of the respective cluster. Fig. 6 displays same dendrograms with different cut lines obtained by a hierarchical clustering analysis. The y axis represents the level of dissimilarity such that each observation is shown with an individual leaf as separate clusters where the total dissimilarity is 0. The cut lines indicated with dashed lines serves to understand the number of clusters when the level of total dissimilarity is equal to the value on y axis. The number of branches which cut the dashed lines vertically indicates the number of clusters with the given level of sum of within cluster variances. While the hierarchical clustering does not necessitate a predefined number of clusters, the analysis does not also provide a globally optimum number of clusters as an output of the analysis. Besides, the visual representation of the cluster relationships serves interpreting the change in within cluster dissimilarities -or cluster homogeneities- in parallel with increased number of generated clusters. The length of vertical lines represent the decrease in within cluster dissimilarity with generating the respective clusters. A heuristic approach for inferring a good number of clusters can be choosing a cut-off point on y axis where each intersecting vertical line, i.e., clusters, decrease the total dissimilarity within clusters.

2.4.1 Defining the Proximity Between Clusters

As it has been discussed before in k-means clustering, most often, Euclidean distance is used to measure the dissimilarity between each pair of observations similarly in hierarchical clustering(James *et al.*, 2013). Hierarchical clustering requires another measurement that is used to define proximity between each pair of clusters to be merged (or divided) accordingly. Measuring the proximity of the clusters, or also referred as the linkage between clusters, is a key operation of the algorithm where different approaches would result with differentiated outputs. Four most frequently used techniques of measuring the proximity between clusters are MIN or Single Linkage, MAX or Complete Linkage, Group Average and Ward's method or Centroid Linkage. MIN, MAX and Group Average

are graph based approaches that are represented on Fig. 7. MIN or Single linkage version of hierarchical clustering defines proximity as the minimum distance, maximum of the similarity in other words, between any two points in two different clusters. This is simply measured by the distance between closest points from the clusters. On the other hand, MAX version of hierarchical clustering defines the proximity with recording the largest dissimilarity between two clusters. Group average hierarchical clustering identifies the proximity of two clusters as the mean pairwise proximity among all pairs of observations in different clusters(Tan *et al.*, 2018). Mathematically, given that the proximity of cluster c_i and c_j is proximity(c_i, c_j) where the number of observations in clusters are n_i and n_j respectively;

$$\text{proximity}(c_i, c_j) = \frac{\sum_{\substack{x \in c_i \\ y \in c_y}} proximity(x, y)}{n_i * n_j}$$



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Fig. 7: Graphical definitions of proximity approaches (Han, Kamber and Pei, 2011)

Lastly, centroid linkage, also referred as Ward's method identifies the proximity between two clusters as the dissimilarity between the centroid for two clusters. Similar to k-means algorithm, centroid linkage version of hierarchical clustering also assumes that the cluster is represented by its centroid "but it measures the proximity between two clusters in terms of the increase in the SSE that results from merging the two clusters"(Tan *et al.*, 2018). The approach is very similar to K-means clustering's objective function which aims to minimize within cluster dissimilarity.

2.4.2 Agglomerative Hierarchical Clustering Algorithm

Agglomerative hierarchical clustering technique has a simplistic algorithm which does not depend on the pairwise observation distance measurement method or cluster proximity methods. The goal is starting with assuming each observation is a separate cluster, successively merging the closest pair of clusters until only one cluster remains. More formally, the approach can be expressed as following:

1. Start with n observations and a measure pairwise dissimilarities with a distance measurement, such as Euclidean distance, of all the observations. Each observation is assumed as cluster.
2. Identify the pair of clusters that are least dissimilar and merge them to be a new cluster.
3. Compute the new inter-cluster proximities.
4. Repeat step 2 & 3 until only one cluster remains.

When the algorithm reaches a single cluster after final iteration, the output is a hierarchically structured data set where members of each cluster is identified for any number of

clusters between two and the size of data set. Similar to K-means algorithm, the hierarchical clustering also does not provide a globally optimum number of clusters. Practicality of the analysis outputs and the performance of the clustering should be carefully considered before drawing a conclusion, i.e., highest within-cluster similarity, which means the number of clusters is equal to sample size, does not represent the best desired output. Total sum of within-cluster pairwise distances (TSS) would be a useful indicator to understand a good number of clusters.

2.5 Density-based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN was introduced by Ester et al. (Ester *et al.*, 1996) as an alternative clustering approach "relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape". DBSCAN was proposed as an appropriate technique to succeed discovering arbitrarily shaped clusters without a prior knowledge about the desired output and with a good efficiency on larger datasets. As it has been argued in previous methods, the dataset with n input variables is considered appearing in a n-dimensional space. The method aims to locate clusters in the space as regions of high density which are separated from one another by regions with lower density of data objects. The methods gives the freedom of measuring the distance among data objects with any appropriate distance function (Ester *et al.*, 1996)

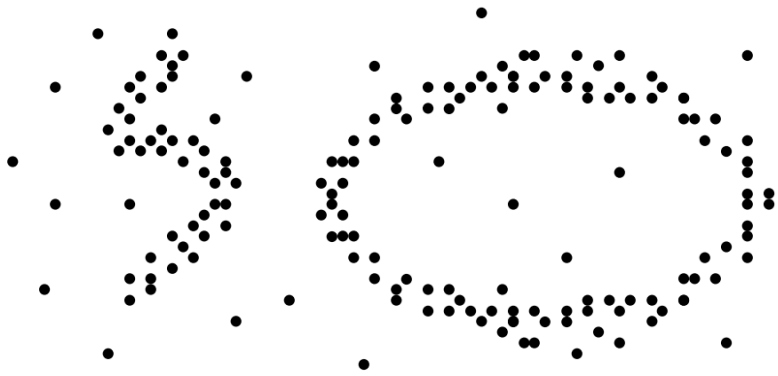


Fig. 8: Data set with non-linear high density regions

Fig. 8 illustrates a dataset placed in a two-dimensional space with high-density regions demonstrating arbitrary shapes (slightly obvious shapes for the sake of simplicity in the example) as distinct from the cluster shapes that centroid based clustering generates as discussed in K-means Clustering chapter. Thus, the density-based argues to provide more meaningful clusters in such case illustrated in the example by associating the observations in a dense data neighborhood and considering the others as noise. DBSCAN algorithm works with two predefined parameters that are briefly described on Table 3.

Parameters	*Definition*
***ε** (Eps)*	ε is used to specify the radius of a neighbourhood that is assumed for each data object. As Fig. 8 illustrates, the ε-neighborhood of the observation is the space within a radius ε centered at the data point. The methods gives the freedom of measuring the distance among data objects with any appropriate distance function (Ester *et al.*, 1996)

MinPts	To identify whether the neighborhood is dense or not, algorithm used the parameter *MinPts,* which declares the threshold for being a dense region.

Table 3: Parameter definitions of DBSCAN

The density-based approach provides three different classes of data points; *Core point* as the interior of the dense region, *border point* located on the edge of the dense region and finally *noise* which are the points sparsely located out of the dense regions. Fig. 8 represents a group of points emphasizing a border point (point B), a noise point (point C) with considering point A as the core point in a bi-dimensional space.

- Core points are in the interior of a density-based cluster. The point is identified as a core point if the number of points within ε-neighborhood, which is determined by the distance function, exceeds the threshold *MinPts.* The point A is a core point where *MinPts* < *7* on Fig. 9.
- Border point is the point which falls within ε-neighborhood of a core point. A point can be identified as a border point of multiple number of core points. B represents a border point on Fig. 9.
- Noise point is a point which is neither a core or border point (Tan *et al.*, 2018) Point C is a noise point on Fig. 9.

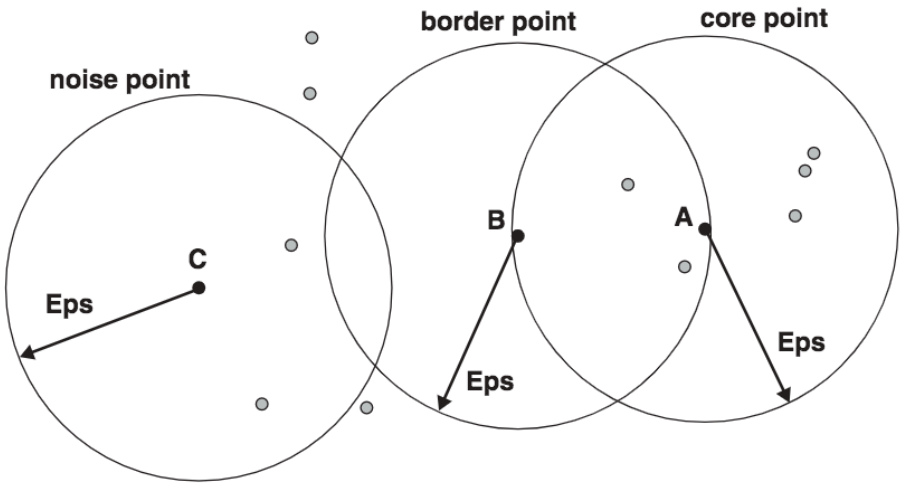


Fig. 9: Object labels with density-based clustering approach (Han, Kamber and Pei, 2011)

2.5.1 Algorithm

Given that $X = \{x_1, x_{2,...} x_n\}$ is the dataset, ε is the neighborhood radius parameter and *MinPts* is the neighborhood density threshold, the algorithm for determining density-based clusters was defined by Han et. al, (2011) as a pseudocode as follows;

```
1. Mark all objects as unvisited;
2.     do
3.         randomly select an unvisited object p;
4.         mark p as visited;
5.         if the ε-neighborhood of p has at least MinPts objects
6.             create a new cluster C, and add p to C;
7.             assume that N is the set of data points in the ε-neighbourhood of p;
```

8. for each point p_1 in N
9. if p_1 is unvisited
10. mark p_1 as visited;
11. if the ε-neighbourhood of p_1 has at least *MinPts* points, add those points to N;
12. if $p_1$is not yet a member of any cluster, add p_1 to C;
13. end for
14. output C;
15. else mark p as noise;
16. until all data points are visited;

Briefly, with the repetitive operations explained, algorithm labels all points as core, border, or noise points; eliminates the noise points; draws an edge between core points that are in ε-neighbourhood of each other; makes each group of connected core points into a separate cluster; assigns border points to one of the clusters that it falls within its ε-neighbourhood (Han, Kamber and Pei, 2011) A significant advantage of this algorithm is being able to identify the outliers with labelling them as *noise* such that, if a point does not fall into neighbourhood of any core point, it stays as a *noise* without influencing within-cluster dense negatively.

2.5.2 Choosing Parameters

Selecting the parameters for DBSCAN analysis is determinant and influences the output of the analysis in the same way as the number of clusters for K-means algorithm due to as it assigns clusters grounded on local distances of observations. Identically, the algorithm should be experimented with different parameters in order to observe alternating outputs. Final decision is made subject to performance of analysis with the given data set and domain knowledge. Though, there is a basic approach suggested by Ester *et al.* (1996) on their paper introducing the method. The heuristic approach basically advocates selecting ε based on the behaviour of *k-dist*, which is the distance between each data point and its k^{th} neighbour. In fact, if the k is not larger than the cluster size, the *k-dist* would be small for the members of the cluster. Certainly, a level of variation is expected depending on the density of the clusters. However, if the cluster densities are not radically incompatible, the range of variance will not be huge for the points in clusters (Han, Kamber and Pei, 2011). On the other hand, *k-dist* value is larger for observations that are not in the cluster, i.e., noise points. When the data points are sorted in descending order by their calculated *k-dist* values, the graphical representation of the behaviour of *k-dist*, which is called *k-dist* graph, is expected to reveal an appropriate ε. If the cure does not provide a good starting point for experiments, algorithm can be iterate with different parameters with updating parameters with interpreting the cluster assignments.

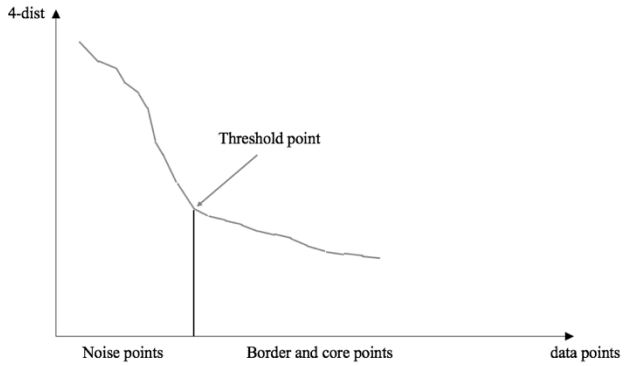


Fig. 10: Noises and clusters based on k-dist graph (Ester *et al.*, 1996).

When a random data point selected from data set and ε is set as the *k-dist* value of selected data point and MinPts to k, all points with smaller or equal *k-dist* will be core points and the rest of the points are either border or noise points. Thereby, if the highest *k-dist* value among the members of smallest cluster could be set as a threshold, the desired parameters would be deduced. As shown in Fig. 9, The threshold point is the first "elbow" of the curve such that all the data points with equal or lower *k-dist* are the member of one of the clusters and the points with higher *k-dist* are considered to be noise points (Ester *et al.*, 1996). Therefore ε is set as the *k-dist* value of threshold point and Min Pts takes the value of *k.* The ideal parameters are set, just as the other unsupervised learning methods, based on some domain knowledge, motivation of clustering, i.e., customer segmentation and visual heuristic approaches suggested.

2.6 Clustering using Gaussian Mixture Models

Another widely used clustering technique is based on probability density estimation using Gaussian Mixture Models (GMM) and the algorithm called Expectation-Maximization (EM) to best fit the model parameters. The basic assumption is that each cluster is generated by a multivariate normal distribution so the overall data set is considered as a mixture of Gaussians with different mean and variances. Deriving each distribution's parameters and identifying which distribution generated each data point result in clustering of the given set of observations.(Deng and Han, 2013)

As discussed in section 2.3, K-means algorithm identifies a cluster using a single point in feature space, i.e., the centroid and assigns each observation to the nearest cluster. The simplicity of K-means brings along some deficiencies: First of all, If two (or more) centroids are too close to each other, their clusters can overlap in the feature space. The algorithm does not provide a further information about which cluster assignment is correct for the data points in the overlapping area. Secondly, while running a K-means clustering analysis, when the Euclidean distance is used to measure distance to cluster center, the clusters have hyper-spherical shape where the radius is equal to Euclidean distance between cluster center and the farthest cluster member. However, some data groups in data set might be defined by a non-spherical shape. As shown in Fig. 10, the figure on the left represents three groups in a sample data set with a balanced spread on variable1 and variable2 axes. K-means can easily detect these clusters if the algorithm is run with k = 3. On the other hand, As shown in the figure on the right, data set might have groups of data points forming a non-spherical shape. K-means fails to successfully assign data points to visually noticeable clusters, because the algorithm generates clusters that are inclusive of the data points within a circular area around the cluster centre as shown.

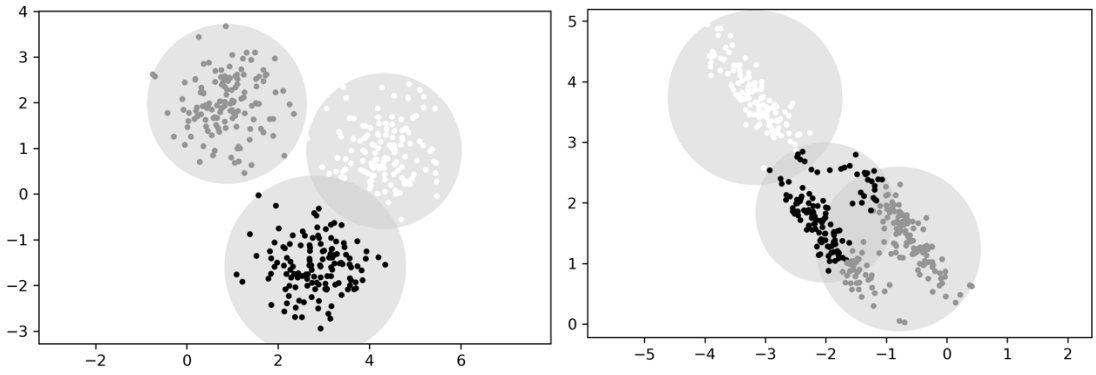


Fig. 11: K-means clustering assignments on two different data set

Gaussian Mixture Models are an extension of the K-means clustering where the clusters are modelled with Gaussian distributions with two parameters: mean and covariance to describe clusters' ellipsoidal shapes so that we can fit the model by maximum likelihood estimation of parameters with an algorithm called Expectation-Maximization (EM). The cluster assignment is very similar with K-means but the major difference is that the algorithm assign each data point to each cluster with a based on a probability function. Hence, it is also called soft cluster assignment.

2.6.1 Gaussian Mixture Models

Suppose that $X = \{x_1, x_2, \ldots, x_n\}$ is the given data set of n observations of a multidimensional random variable x and the random variable x is distributed according to a mixture of K components where each component, i.e., cluster, is represented by a distribution with distinctive set of parameters. Therefore, entire data set is a model identified by a mixture of these component distributions and the probability density function of random variable x can be formulized as

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\theta_k)$$

where π_k is the prior probability or weight coefficient of component k, θ_k is the parameters of component k and $p(x|\theta_k)$ represents the component distribution. Let z be a K-dimensional binary vector with a 1-of-K representation i.e., $z = (0,\ldots,1,0,0)^T$, in which only k^{th} element z_k is equal to 1 and rest of the elements are equal to zero denoting that origin of the observation x is k^{th} component. Thus, z_k must satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$. Depending on the location of nonzero element in the vector, there are K possible different versions of z. Thus, marginal distribution over z can be defined using the mixing coefficient as

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

where π_k must satisfy

$$(0 \leq \pi_k \leq 1), \qquad \sum_{k=1}^{K} \pi_k = 1 .$$

Accordingly, conditional distribution of the x with a given z can be described as

$$p(x|z) = \prod_{k=1}^{K} p(x|\theta_k)^{z_k}.$$

Given that the z is the latent variable, i.e., cluster assignment, and the joint distribution of x and z is defined as $p(\mathrm{z})p(x|\mathrm{z})$, the marginal distribution of x can be written as

$$p(x) = \sum_z p(\mathrm{z})p(x|\mathrm{z}) = \sum_{k=1}^{K} \pi_k p(x|\theta_k).$$

This formulation represents a mixture model in which each observation has a corresponding hidden latent discrete variable z. If we paraphrase in context of clustering, the assumption is that each possible customer x is a member of a cluster k with $p(z_k = 1)$ a

prior probability and the observed customer x is a member of cluster k with the posterior probability $p(z_k = 1|x)$. It is also referred as weight or responsibility that cluster k takes for explaining x (Bishop, 2006). The responsibility of cluster k is denoted as r_k in the following explanations. The value of r_k could be derived as

$$r_k = p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)}$$

$$= \frac{\pi_k p(x|\theta_k)}{\sum_{j=1}^{K} \pi_j p(x|\theta_j)}$$

Using the Bayes' theorem(Bishop, 2006). Suppose that, that entire data set is a mixture of multivariate Gaussian distributions with D dimensions. In other words, the data points are samples from multiple number of normal distributions. The graph on the left in Fig. 10 represents a probability density function of a one dimensional random variable x in which the data points are coming from three Gaussian distributions with different mean and variances. Thus, instead of one normal curve, there are multiple curves with different spreads. On the other hand, the figure on the right shows a mixture of two multivariate Gaussian densities over two variables where the multivariate nature of the data can be observed with the assistance of 3D representation.

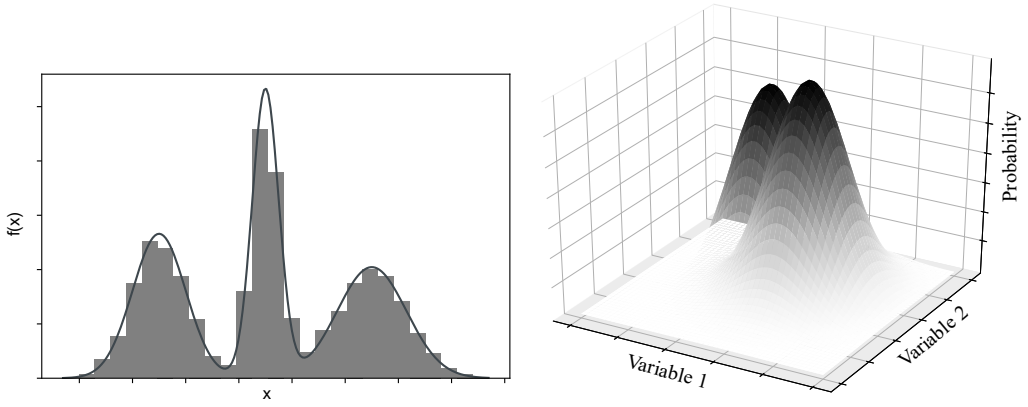


Fig. 12: An example of Gaussian density for a single variable (figure on the left) and a multivariate Gaussian density over two variables (figure on the right).

The multivariate Gaussian distribution of vector x with D dimensions can be written as

$$N(x|\mu, \Sigma) = (2\pi^{D/2})^{-1}(|\Sigma|^{\frac{1}{2}})^{-1}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\},$$

where μ is the vector containing component means, Σ is a D×D covariance matrix and |Σ| represents the determinant of Σ. In this case, the conditional distribution of x given z is such that the marginal distribution can be formulized as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

and the conditional probability of z given x is

$$r_k = p(z_k = 1|x) = = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}$$

In the light of previously introduced definitions. To be able to fit the given sample to a mixture of Gaussian models, the parameters μ_k, r_k, Σ_k must be inferred for each Gaussian distribution behind the data set. A very important statistical concept, that is called maximum likelihood estimation, assists estimating the parameters of model such that the best estimate maximizes the probability of generating all the observations (Deng and Han, 2013). Given that the N number of observations $X=\{x_1, x_2,.., x_n\}$ are coming from a Gaussian mixture model $N(x_n|\mu_k, \Sigma_k)$ having k components with parameters is $\Phi = \{\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k\}$, the log-likelihood function can be written as

$$\log p(X|\Phi) = \sum_{n=1}^{N} \log p(x_n|\Phi) = \sum_{n=1}^{N} log \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)$$

where the objective is

$$\arg\max_{\Phi} \log p(X|\Phi).$$

As it has been described by Bishop (2006) and Deng and Han (2013), to finding the derivatives of log-likelihood estimation with respect to each parameter π_k, μ_k, Σ_k, respectively, discloses the maximum likelihood estimation of parameters for the local maxima of components. For instance, setting the derivative of $\log p(X|\Phi)$ with respect to μ_k to zero and multiplying by Σ_k, the mean is obtained as

$$\mu_k = \frac{\sum_{n=1}^{N} r_k x_n}{\sum_{n=1}^{N} r_k},$$

where

$$r_k = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)},$$

as described in prior inference for the component responsibility coefficient. Similar to mean estimation, the derivative with respect to Σ_k gives the formula of covariance estimation as

$$\Sigma_k = \frac{\sum_{n=1}^{N} r_k (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} r_k}.$$

Finally, the mixing probability of components π_k is derived by derivative of log likelihood function with respect to prior mixing probability o components π_k as well, but with additional inference using Lagrange multiplier (Bishop, 2006). π_k is obtained as

$$\pi_k = \frac{\sum_{n=1}^{N} r_k}{N}.$$

If the equations obtained by log-likelihood function briefly summarized, the mean of the component k is the weighted mean of all points in data set where the weighting factor is r_k. Similarly, obtained Σ_k reveals that each data point is weighted by the conditional probability generated by the corresponding component divided by the number of data points associated with the component. Lastly the π_k is the average responsibility that component k takes for explaining all data points in the given data set (Deng and Han,

2013). Even though the equations of each parameter can be obtained by maximum likelihood approach, finding the actual values of each parameter is still a very complex problem to solve due to the fact that the results of π_k, μ_k, Σ_k rely on the responsibility coefficient r_k while the r_k depends on these parameters therewithal. Nevertheless, a very functional method called Expectation-Maximization (EM) algorithm has been used for finding solutions for locally maximization of parameter estimations explaining models with hidden or unknown variable, i.e., cluster assignment (Dempster, Laird and Rubin, 1977). Next chapter provides an overview of the EM algorithm and its application in context of finding Gaussian mixture model parameters.

2.6.2 Expectation-Maximization Algorithm

EM was first introduced by Dempster, Laird and Rubin (1977) to present an general approach for iteratively computing maximum-likelihood estimates when the observations have incomplete data, i.e., a missing feature such as cluster assignments in context of clustering analysis. The name of the algorithm comes from two iterative steps of the procedure which are expectation step (E) and maximization stem (M). The process starts by initializing with guessed initial parameter estimates. These initial estimations could be obtained by domain knowledge or in case of lack of information, a reasonable starting estimates could be achieved by the output of K-means clustering. For instance, within-cluster covariances would give an estimate for Σ_k and fractions of data points assigned to each cluster would be a reasonable estimation for π_k. Following E and M steps iterates alternately to maximize the likelihood of these parameters. In E-step, r_k is calculated with using available values of parameters. In M-step, the log-likelihood is maximized with the updated component responsibilities and μ_k, Σ_k, π_k are calculated again using new r_k. Then, the log-likelihood is calculated again and the convergence of the algorithm is evaluated. If the convergence criterion is not satisfied we repeat the step 2,3 and 4. Otherwise, final parameters are considered as maximum likelihood estimations. As Deng and Han (2013) denotes, each iteration of the EM algorithm is increases the log-likelihood and in practice, EM algorithm is reaches a point where the log-likelihood does not improve significantly anymore. Taking account of the inferred equations for Gaussian mixture model parameters, iterative steps of the algorithm could be formally described as

1. Initialize the μ_k, Σ_k, π_k and calculate the log-likelihood

$$\log p(X|\Phi) = \sum_{n=1}^{N} \log p(x_n|\Phi) = \sum_{n=1}^{N} log \sum_{k=1}^{K} {\pi_k}^0 N(x_n|{\mu_k}^0, {\Sigma_k}^0).$$

2. E-step: With the current parameters, calculate the component responsibilities

$$r_k = \frac{\pi_k \mathcal{N}(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j,\Sigma_j)}.$$

3. M-step: Estimate the parameters again with calculated component responsibilities r_k

$${\mu_k}^{new} = \frac{\sum_{n=1}^{N} r_k x_n}{\sum_{n=1}^{N} r_k}, {\Sigma_k}^{new} = \frac{\sum_{n=1}^{N} r_k (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} r_k}, {\pi_k}^{new} = \frac{\sum_{n=1}^{N} r_k}{N}.$$

4. Evaluate the log likelihood with

$$\log p(X|\Phi) = \sum_{n=1}^{N} \log p(x_n|\Phi) = \sum_{n=1}^{N} log \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)$$

and monitor the convergence of log-likelihood. If the convergence criterion is not satisfied return to step 2. The result of the Expectation Maximization process associates each cluster with a Gaussian Mixture Model with locally optimum parameters, instead of a hyper spherical shapes as in k-means algorithm. Similar to k-means, the method promises a local maxima with given initial parameters, instead of a global optimum meaning that different initializations would result in variety of outputs. Thus, interpreting the results of technique with several experiments would give a better understanding about the hidden patterns in dataset.

2.6.3 Model Selection Criterion

While maximizing the likelihood of parameters, Gaussian Mixture Model approach does not suggest a best number of clusters as is K-means algorithm using a qualitative indicator. Higher number of clusters does not necessarily implies a better solution for the given clustering problem. As a usual challenge in model selection, while the mixture of too many gaussian models can overfit the data, too few models could imply very heterogeneous clusters which also does not serve the objective of the analysis. Therefore, a statistical procedure needs to be adopted to make inference about the number of clusters which provides the best possible model with a minimum level of information loss. There are number of criteria that have used for comparing the level of information provided by models generated such as Akaike's information criterion (AIC), Schwarz's Bayesian approximation criteria (BIC), Rissanen's minimum description length (MDL), integrated classification likelihood (ICL), etc.(Deng and Han, 2013) Among many available methods, information-theory based methods AIC and BIC are commonly used criterion that can serve for performing model selection. AIC was first introduced by Akaike with the purpose of comparing maximum likelihood estimation of obtained candidate models and identify the model with the minimum AIC value that can be defined as

$$AIC = -2logL(\widehat{\Phi}) + 2k$$

where Φ is the set of parameters, $L(\widehat{\Phi})$ is the maximum likelihood of the model estimated with Φ and k is the number of the parameters estimated in the candidate model (Akaike, 1974). First component of the formula represents the probability of obtaining the data set from the candidate model multiplied by minus two so that higher maximum likelihood minimizes the AIC value. As mentioned, the more number of parameters generally offers a higher maximum likelihood however, the approach eliminates the threat of overfitting with adding k multiplied by two which penalizes the model only if more parameters are added to the model. On the other, Schwarz introduced BIC, which is very closely related to AIC approach, shortly after Akaike's publication. BIC provides a Bayesian argument for adopting a larger penalty term such that increasing the number of parameters has a higher influence on the computation of criterion value (Schwarz, 1978). BIC is calculated very similar to AIC with

$$BIC = -2logL(\widehat{\Phi}) + k\log(N)$$

where N denotes the number of data points. The only difference from the AIC appears on the penalty component which impose a higher penalty for larger number of parameters. Both criterion could be used simultaneously to observe the performance of candidate models from multiple perspectives.

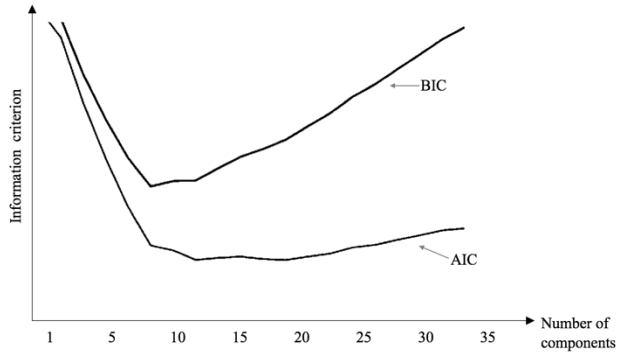


Fig. 13: Example BIC and AIC values for different number of components

In practice, a heuristic approach with using a visual representation of AIC and BIC values for different models could be used for choosing the optimal number of components. As shown in Fig. 12, AIC and BIC results can suggest different number of models based on the minimum AIC and BIC values. The selection can be made aiming to minimize both criterion or minimizing either one of them based on the objective of the comparison of the models. The common approach is making selection which gives a low value for both criterion.

2.7 Evaluation of the Clustering Methods

In supervised learning methods, there are different evaluation criterion and techniques such as means squared errors, area under curve, confusion matrix, etc., are interpreted along the learning process. However, evaluation of the accuracy is not a commonly used part of analysis on non labelled data sets by the very nature of unsupervised learning techniques (Tan *et al.*, 2018). The consideration of the already mentioned evaluation or validation methods in prior chapters gives information regarding the compactness or tightness of the generated clusters with evaluating the within-cluster distances. Rousseeuw (1987) introduced a measure called silhouette which explains how similar a data point to its cluster compared to other clusters generated. Silhouette scores of each data point are measured with the given distance measure, i.e., Euclidean distance using the basic formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance between point i and the other members of its cluster, $b(i)$ is the minimum of average distances between point i and members of clusters not containing i. Silhouette score can take values between -1 and 1. A negative silhouette score implies that the average distance of some neighboring cluster's members is smaller than the average distance to members of object's own cluster. Thereby, silhouette score for data point is expected to be greater than zero. Moreover, $a(i)$ should be as close as possible to zero if highly similar objects are grouped in the same cluster successfully. After recording the silhouette scores of each data point, mean silhouette score of entire data set and each cluster separately calculated. In the light of created parameters, we can summarize the remarkable outputs as follows:

- The mean silhouette score is expected to be as close to one as possible for a better clustering.
- The mean scores of each clusters should be higher than the overall average as much as possible.
- The negative silhouette scores indicates the data points decreasing the homogeneity in cluster.

The Silhouette scoring can be considered to find out tightness and separation across generated clusters in distance centric approaches. Also, the method can be used to decide on number of clusters to generate, if it is not known or there is no targeted number of clusters.

3 Implementation of Unsupervised Learning Methods to Identify Customer Segments

The analysis represents performing unsupervised learning methods which are explained in previous chapters, using an open source real world dataset. Before performing each clustering technique to identify customer segments in given data set, presenting a general overview of business field and collected data would lead to a better understanding of raw data and correspondingly a more efficient data preparation practice. The dataset has been retrieved from UC Irvine Machine Learning Repository where the organization has been described as a non-store online retail company (*Online Retail Data Set*, 2015).

3.1. Business Overview

"Online Retail Data Set" has been retrieved from the UCI Machine Learning Repository where the organization has been described as a non-store online retail company . As per traditional definition, the retail entity is the part of the distribution chain which purchases goods from wholesalers, who have taken products from manufacturers, and sells these goods to final consumers. With the progressing logistic processes and business complexity in developing industries, the entities' roles have been evolving and the distribution chains are becoming more complex with new rings on the chain. Thus, with more general terms, retailing can be described as selling products, which have been purchased by wholesalers in high quantities, to customers in smaller quantities through a store such that the buyers could be the end users of the products but also they could be another entity in distribution chain as a secondary retailer or another small business model before final consumption. According to the given information about data set on UC Irvine Machine Learning Repository website, the customer base of the mentioned company consist of many wholesalers *Online Retail Data Set*, 2015). Consequently, while some of the customers purchase in small quantities, many of them order products in significantly high quantities. It is highly possible to see significant differences among customers in terms of revenue. As it is mentioned on source website, the company's product catalogue mainly consists of unique all-occasion gift products. A quick internet search reveals that customers are able to reach number of different online stores which offer a range of gift products for all occasions.

Fig. 14: Product offered in different online retail stores

A more detailed search with product names that appear in data set shows that the same products are available in many different online stores with different unit prices. For instance, as Fig. 14 shows that one of the products from the data set can be purchased by different online stores which offer the same product. Some of these online stores offer different prices based on the purchase volume as well. The company apparently is running the online retail business in a market where there are number of competitors that offer their products both in UK and in other countries. If the company does not perform appropriate marketing practices and pricing strategies, customers can easily reach another

online store with a better purchase experience or lower prices. As it has been explained in chapter 2.1, company can increase the competitive advantage with a market driven strategy, which starts with understanding the needs and expectations of customers. Customer segmentation has a crucial role in understanding the customer base to be able to answer their needs accurately.

Among the different segmentation methods, i.e., geographic, demographic, psychographic or behavioural, appropriate type of segmentation should be performed according to both company's expectation from the analysis and the context of available data set. Following analysis represents an example of behavioural segmentation. The main reason for proceeding with behavioural segmentation is the characteristics of the given data set. A detailed geographic segmentation, for instance, cannot be performed accurately since the only serviceable information of customer is the country level location and there are very few countries where significant number of customers reside. Similarly, data set does not contain variables for performing demographic and psychographic segmentation. Thus, the possible results of these segmentation analysis would not likely to provide useful new information. On the other hand, behavioural segmentation would be the most reasonable amongst the others since a higher level information regarding customers' buying behaviours can be obtained from given financial transactions.

3.2 Data Understanding

The dataset consist of online transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. Product catalogue mainly consist of all-occasion gift products which are purchased by either large scale wholesalers or normal size retailers. The raw data set contains 541909 records of 8 variables. As it can be seen on the Table 4 representing a small sample from the data set, each records consist of timestamp of transaction, invoice number, stock code of the product, product description, purchased quantity of the product, price of the products, customer identifier and the country that the customer resides.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	Customer	Country
569212	22295	HEART FILIGREE DOVE LARGE	6	02/10/2011 12:05:00	1.65	15164	United Kingdom
C536379	D	DISCOUNT	-1	01/12/2010 09:41:00	27.5	14527	United Kingdom
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010/12/01 08:26:00	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26:00	3.39	17850	United Kingdom
555719	35810A	ENAMEL PINK COFFEE CONTAINER	12	06/06/2011 15:31:00	0.83	12609	Germany

Table 4: Randomly selected entities from data set

Each row represents a product with a certain quantity that has been recorded within the financial transaction by a particular customer. Thus, a purchase, which can be identified by unique InvoiceNo, appear in multiple records if customer purchases multiple type of products in one purchase. The basic characteristics of variables can be described as follows:

- *InvoiceNo* is a nominal variable which holds a 6-digit integral number uniquely assigned to each transaction. 541909 records provide details of 25500 unique transactions. According to the data set source, if the code starts with letter "C", it

indicates a cancellation. Cancelled transactions present 13% of 25500 transactions recorded with unique invoice numbers.

- *StockCode* is the product or item code which appears in records as 5-digit integer number. There are 4070 unique StockCode each representing a different product offered by company. Although the majority of the codes are 5-digit integers, there are also 571 records where StockCode has recorded as "M" which are manually processed transactions according to description. Also, 70 records hold "D" letter which are associated with transactions for discounts.
- *Description* field shows the name of the product with a unique StockCode. There are 4223 descriptions against 4070 unique stock codes which means that majority of the records have one-to-one matching StockCode and associated Description. A closer inspection reveals that there are some records with misspelling or slightly different naming of same products.
- *Quantity* is a numeric variable showing the purchase quantity of the product. The purchase quantity varies between -80995 and 80995. These maximum and minimum values point out that the cancelled transactions have negative entities on quantity field. If a transaction was cancelled, a duplicate entity has been recorded with "C" letter attached to original InvoiceNo and negative value of the first recorded purchase quantity.
- *InvoiceDate* is the date and time when each transaction was generated. Transactions occured between 01/12/2010 and 09/12/2011. Each record with the same InvoiceNo have the same InvoiceDate as well.
- *UnitPrice* is a numeric variable which is the price of the product per unit in pound sterling currency. Unit price varies between -11062 and 38970. These extreme values are associated with either transactions that are not related to a customer or entities that have been manually entered with limited further information. On the other side, unit prices of products are fluctuant such that the same product could be purchased with different unit prices by different customers.
- *CustomerID* is the unique identifier, that is a 5-digit number, of each customer in company's customer base. The entire dataset contains transactions from 4372 unique customers.
- *Country* is the name of the country where each customer resides. Company's customer base consist of customers from 36 different countries. While the vast majority of the customers reside in United Kingdom, company also sell products to 422 customers from overseas.

Real world data sets are likely to contain records that are either incorrect or not available at all. There are 5268 duplicated records which could be symptoms of errors or more likely, they could indicate that the available data set is a subset of fields and entities from a larger and higher-dimensional data set such that the non-available fields of duplicated entities contain different values. In addition to duplicates, a significant number of records reflect non-available values. More precisely, 33% of the data set does not contain a valid value on CustomerID field and minor proportion of the rows have non-available data as Description which correspond to 0.0027%. As the small sample from the records with non-available Description on Table 5 shows that these records do not have a valid value on CustomerID field and the UnitPrice of the products have been recorded as 0.00.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536414	22139	NaN	56	01/12/2010 11:52:00	0.0	NaN	United Kingdom
536545	21134	NaN	1	01/12/2010 14:32:00	0.0	NaN	United Kingdom

536546	22145	NaN	1	01/12/2010 14:33:00	0.0	NaN	United Kingdom
536547	37509	NaN	1	01/12/201 08:26:00	0.0	NaN	United Kingdom
536549	85226A	NaN	1	2011/06/06 15:31:00	0.0	NaN	Germany

Table 5: Records with non-available CustomerID

As it has been stated previously, the vast majority of the customers reside in United Kingdom and there are both wholesalers and small retailers in existing customer base of the online store. Country level sales numbers evince the fact that a minority of customers from a variety of locations bring significantly higher amount of revenue while majority of the customer base reside in UK. As Table 6 shows that 84% of the overall revenue in given period of time has been generated by local customers. While 11.33% of the revenue have been acquired by 5 other countries, the contribution of customers from other – or unspecified – locations contribute to revenue is 4.68%. The total purchase quantity, which appears as "Sales Volume" on table, shows a similar trend with the revenue distribution. However, the number of customers per each location, does not alternate directly proportional to either total sales volume or total revenue generated which indicates that there are remarkable purchase transactions from oversees.

Country	**Transactions**	**Customers**	**Revenue (£)**	**Total quantity**	**Revenue (%)**	**Sales Volume (%)**
U.K.	23494	3950	8,187,806	4263829	84	82.37
Netherlands	101	9	284,661	200128	2.92	3.87
Ireland	360	3	263,276	142637	2.70	2.76
Germany	603	95	221,698	117448	2.27	2.27
France	461	87	197,403	110480	2.03	2.13
Australia	69	9	137,077	83653	1.41	1.62

Table 6: Sales by country

For instance, few number of customers contribute to revenue at a significant level from some markets such as Netherlands, Ireland and Australia. Despite the average revenue per customer is £2,226, Fig. 15 shows that limited number of customers from Ireland contribute £87,759 to revenue on average and the averages are significantly higher than average in Netherlands and Australia as well. Besides that, the average value per transaction is £349 for the customers in UK, while the numbers are significantly higher for the customers from Ireland, Netherlands and Australia. Most particularly, 2.92% of overall revenue have been generated by 9 customers from Netherlands with an average of £2,818 per each transaction. This average transaction value of a customer from Netherlands is 7 times higher than the average for UK. In other words, the average amount of money that 1 customer from Netherlands spend is equal to the amount of money that 8 customers from UK spend per a purchase.

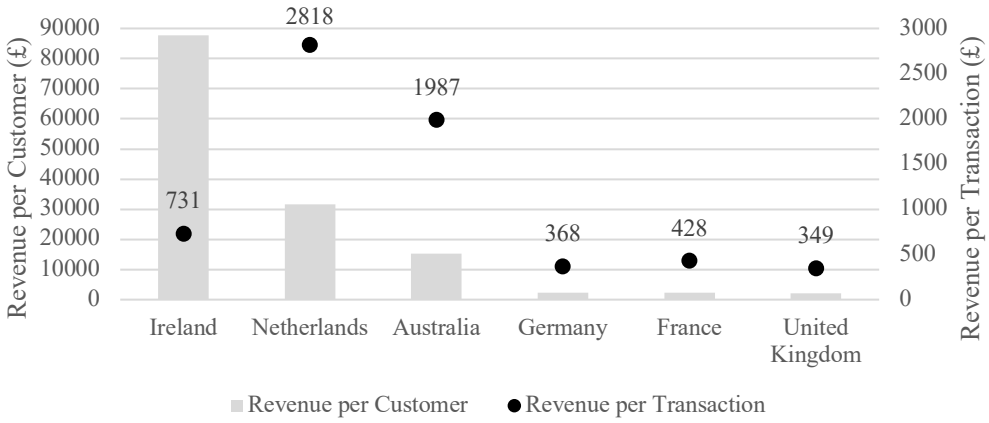


Fig. 15: Customer numbers against revenue by country

When overall revenue per customer is considered, the Fig. 15 shows that the average revenue per 1 customer from Ireland is 42 times higher than the average revenue of customer from UK. These findings indicate two major characteristics of the customer base: Although the large proportion of revenue has been generated from local customers, there are still customers influencing revenue from other countries. Besides, some of these customers are likely to be the wholesalers that has been mentioned previously.

CustomerID	Purchase frequency	Sales amount	Revenue Contribution (%)
12748	4594	£129,075	1.93%
14096	5095	£112,614	1.69%
17841	7847	£105,113	1.57%
13694	568	£85,569	1.13%
14911	5675	£77,127	1.15%
18084	1	£1	~ 0%
16881	1	£1	~ 0%
13307	1	£1	~ 0%
17443	1	£1	~ 0%
17925	1	£1	~ 0%

Table 7: Customers with maximum and minimum revenue contribution

Table 7 shows sales data of five customers with highest and another five customers with the lowest sales contribution. Most notably, customer 12748's contribution to overall revenue is 1.93%. Including following top contributors on the list, there are certainly some customers which can be considered as outliers. The customers on the bottom of the list reveals that company has also individual customers ordering very few items. A segmentation analysis based on sales numbers is expected to differentiate these customers from retails and wholesalers. Given data set includes transactions between 30/11/10 and 09/12/11 meaning that the dataset covers the 51 calendar weeks and as shown in Fig. 16 , revenue and total number of items sold shows a similar trend over the year. Sales show a significant increase in 9th month such that the best sales day is 20th of September with £184,349.28 total revenue and increase until the end of the year. This increase in sales could be explained with the increasing demand by retailers before the festival season. Data set has sales data attached to first three weeks of December 2010 and first nine days of December 2011 such that the sales numbers seem decreasing significantly after November.

Fig. 16: Revenue and purchase quantity by month

Stock code and description level overview of product sales gives a useful hint about variant customer behaviors. For instance, sales of "Paper Craft, Little Birdie" product contributes £168,470 to total revenue, which is the highest sales contribution by product type. Besides, the whole sales is associated with only one order such that one customer made the purchase with paying £168,470 for total amount of 8995 "Paper Craft, Little Birdie" products.

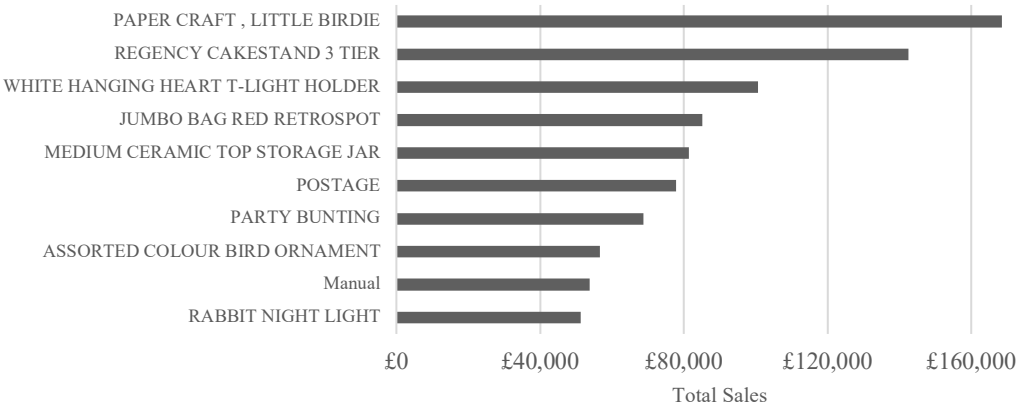


Fig. 17: Total sales by product types

Fig. 17 represents the total sales of the ten highest revenue contributor products. Based on the product descriptions, there are some manually added entities without the product level information. Also, there are some descriptions indicating that the purpose of transaction is not a purchase.

Stock code	Description	Number of orders	Number of items	Sales Amount	Unit Price (mean)
POST	POSTAGE	1099	3120	77803.96	31.57048226
M	Manual	253	7173	53779.93	175.2915845
DOT	DOTCOM POSTAGE	16	16	11906.36	744.1475
C2	CARRIAGE	133	134	6686	50.13533835

Table 8: Non-purchase transactions in data set

Table 9 presents four different stock code entities with associated sales data. "Manual" entities could be seen on both Fig. 17 and Table 9 with a high total transaction value.

Since the further analysis require only invoice level total transaction value, unclarity of the values on StockCode or Description do not have an effect on the final findings. However, the other transactions on the Table 8 could be excluded from the further analysis since these transactions appear to be additional costs to purchases.

3.3 Data Preparation

As a real world business data set, there are number of records containing values that are either unfunctional or not available for the further analysis. First of all, 25% of the entire dataset do not have a valid recorded CustomerID. 1454 of these records also have missing values on Description. These entities are excluded from the analysis, since the CustomerID is vital for measuring customer level purchase transactions. Table 10 shows that there are also some other entities which should be emphasized. Almost 1 percent of the rows contain duplicated records. This proportion could indicate that the available data set has been extracted from a higher-dimensional data set where these records have varying records. These records are not going to be excluded unless there are any other issue. In terms of purchase values, 2.18% of the dataset represents a negative transaction which are mostly associated with cancelled orders. Negative transactions and cancelled orders are not included in further analysis. Finally, as Table 9 shows that there are some transactions related to postage and carriage costs. These records also would not be useful for understanding customer behavior.

	Number of records	**(%) of total**
Duplicate transactions	5268	0.97%
Negative transactions	11805	2.18%
Non-purchase transactions	1966	0.36%
Cancelled transactions	9288	1.71%
Total number of transactions	541909	100%

Table 9: Amount of impractical entities

The Quantity and the UnitPrice are the only numeric variables that are going to be used to measure revenue attached to each customer. UnitPrice across whole product types shows a highly skewed distribution. Based on the z-scores of UnitPrice records, 206 entities are outliers assuming that any z-score greater than 3.5 or less than -3.5 is considered to be an outlier. A closer look at the outliers points out the manual entities that influence the distribution of UnitPrice significantly. As Table 10 shows, manually entered records have a very high value on unit price variable which also equals to the total value (Quantity * UnitPrice) of the transaction such that the Quantity = 1 for all entities with StockCode = M. This derivable satisfies the requirements of further analysis. Additionally, transaction quantities also include some extreme values which are associated with customers who order with high quantities. These entities will also be useful for a better understanding of these customers. In other words, the entities, which can be considered as outliers, are not excluded from the analysis for having more information about all customers appearing on available records.

InvoiceNo	**StockCode**	**Description**	**Quantity**	**InvoiceDate**	**UnitPrice**	**CustomerID**	**Country**
573080	M	Manual	1	27/10/2011 14:20	4161.06	12536	France
573077	M	Manual	1	27/10/2011 14:13	4161.06	12536	France

571751	M	Manual	1	19/10/2011 11:18	3949.32	12744	Singapore
569382	M	Manual	1	03/10/2011 16:44	3155.95	15502	United Kingdom
562946	M	Manual	1	11/08/2011 09:38	2500	15581	United Kingdom
548813	M	Manual	1	04/04/2011 13:03	2382.92	12744	Singapore
571751	M	Manual	1	19/10/2011 11:18	2118.74	12744	Singapore
548820	M	Manual	1	04/04/2011 13:04	2053.07	12744	Singapore
566927	M	Manual	1	15/09/2011 15:20	2033.1	17846	United Kingdom
549468	M	Manual	1	08/04/2011 14:27	1867.86	17940	United Kingdom

Table 10: Manual entities

Cleaned data still requires a feature engineering to be suitable for further unsupervised learning analysis. Since the subject clustering methods demand a customer level data set, the available table should be transformed into a table form which has variables per each customer in each row as shown on Table 11.

Customer	Recency	Frequency	Monetary
17850	373	34	5302
13047	32	10	2010
12583	3	15	2872
13748	96	5	268
15100	334	3	56

Table 11: Sample from transformed customer data set

Customer features on the Table 11 has been generated based on the variables of RFM analysis, which is the marketing phenomena explained on chapter 2.1.4. The definitions of Recency, Frequency and Monetary variables in context of given data set are:

- *Recency* represents how recently the customer has made a transaction. The variable was measured as the total time elapsed between the last transaction of the customer and one day after the last transaction date available in data set.
- *Frequency* indicates how frequently the customer make purchase from the company which is derived by the total number of transactions per each customer.
- *Monetary* finally represents the total revenue distribution of the customers over the given time period which is in other words, total transaction value per customer.

A descriptive analysis of the features that are used in further analysis discloses that customers made purchase for approximately four times on average in given time period. While there are only one or two invoices associated with %50 of customers, there are also customers who purchase very frequently such that the maximum is 210. Besides, the monetary values, which is the total value of transactions in £, disclose that company generate £1538.85 per customer on average. Half of the customers' contribution to revenue is less than or equal to £493.5 with a high standard deviation of £4509.55. Skewness of the distribution of monetary and frequency features indicate that a clustering based on RFM variables would lead to large clusters consisting of customers with small revenue

contribution and small clusters including fewer number of customers who contribute to revenue significantly.

	Recency	Frequency	Monetary	Recency Standardized	Frequency Standardized	Monetary Standardized
Mean	93.09	4.27	1538.85	0.00	0.00	0.00
Std. deviation	100.015	7.7	4509.55	1.00	1.00	1.00
Min	1	1	0	-2.70	-0.41	-0.36
25% quartile	18	1	121	-0.64	-0.33	-0.32
50% quartile	51	2	493.5	0.07	-0.22	-0.23
75% quartile	143	5	1491.25	0.81	0.04	0.00
Max	374	210	129075	1.61	32.46	27.07

Table 12: Summary statistics of customer RFM data set

As shown on the first three columns of Table 12, each variable has a different scale with varying quartiles, means and standard deviations. Traditional RFM approach deals with the difference in variable scales with ranking each variable values from the scale of one to five. In following clustering analysis, the scale difference issue is handled with a more statistical approach such that each variable is standardized with deriving z-scores of R,F and M variables where mean is considered as zero and the standard deviations are reduced to one as seen on the last three columns of Table 12.

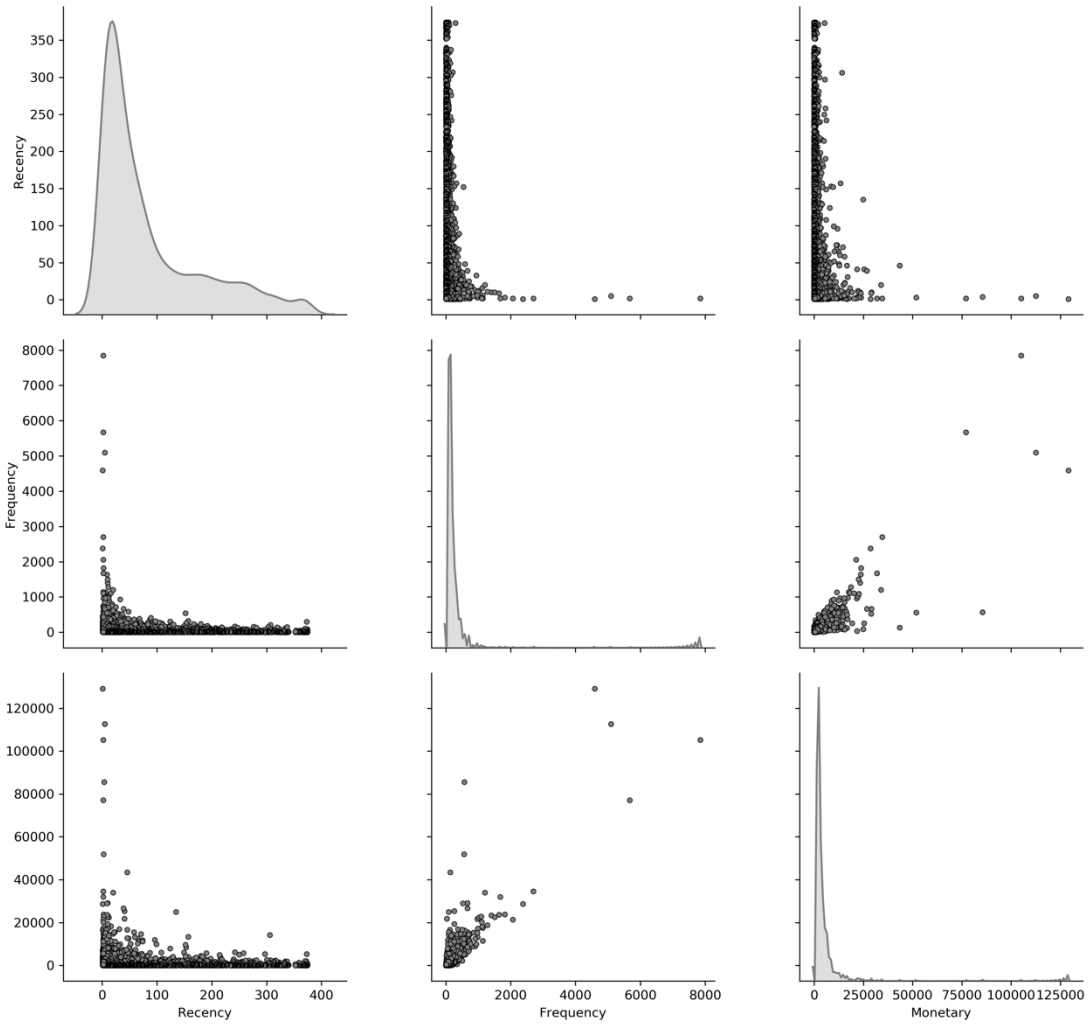


Fig. 18: Distributions and pairwise correlations of R,F,M features

Figure 18 gives an overview of the distribution of recency, frequency and monetary features together with pairwise correlations. The graph demonstrates the skewness of each variable, especially frequency and monetary which is not unforeseen considering the prior inferences. This would be a noteworthy issue, in context of different analysis subjects, i.e., predictive analysis yet, the clustering analysis in question does not necessitate such assumptions. Each variable can still provide additional information for understanding the purchase behaviours of customers although, the frequency and monetary variables show a very similar behaviour. For instance, as shown on Table 7, five customer with the highest sales contribution are not the customers who purchase most frequently.

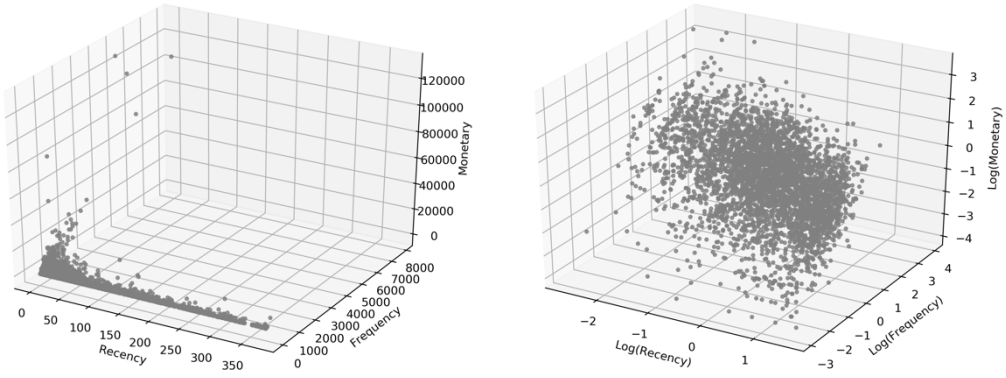


Fig. 19: Data set on three-dimensional space before and after logarithmic transformation

Even though the methods do not require a symmetrical distributions, the shape of overall distribution would influence the performance and the outputs of the methods. The pairwise distances of data points and the density areas of the data set can be manipulated for improving the performance of learning techniques. Logarithmic transformation of variables have generated to provide differently shaped and skewed data sets that the unsupervised learning methods can use as base. Fig. 19 shows the customer data on a three-dimensional Euclidean without any transformation and with logarithmic transformation, respectively. While the initial shape of distribution seem dense around a same area, the logarithmic transformation provides a fairly outspread dataset. Finally, after applying the transformations, more uniformly distributed pairwise relations and more normal distributions of features can be seen on Fig.20.

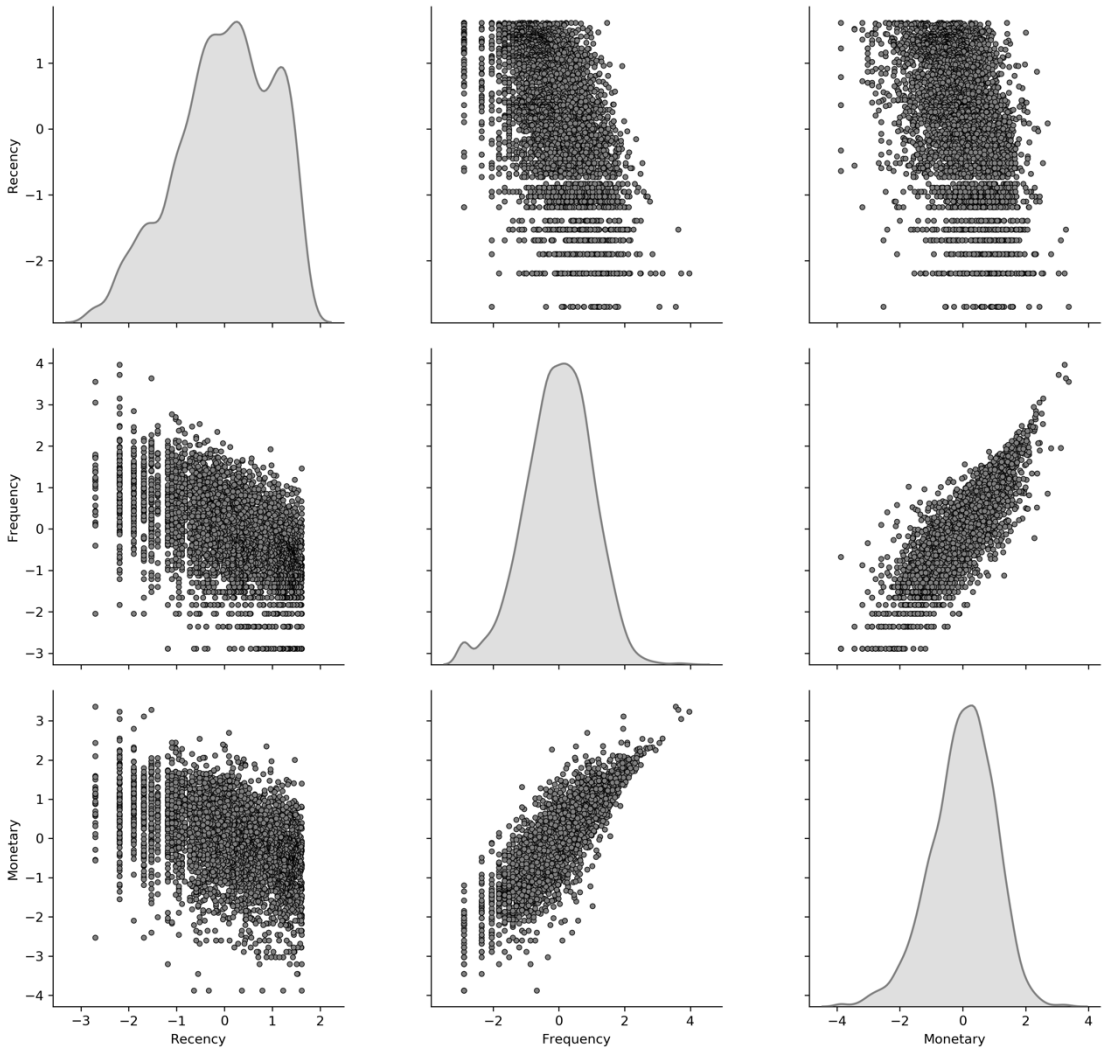


Fig. 20: Distributions and pairwise correlations of transformed R,F,M features

3.4 K-Means Clustering

As it has been mentioned in chapter 2.3.1, K-Means clustering is a distance based clustering method that can be performed to compose customer segments by determining clusters with the objective of having high similarity within clusters and low pairwise similarity among members of different clusters. The applied method represents a modified version of k-means clustering with respect to an alternative initial seeding method k-means++ disclosed in chapter 2.3.4. K-means clustering necessitate a number of predeterminations which have influence on the output of algorithm. First of all, Euclidean distance was used as the base distance measure to determine pairwise and within-cluster similarity (and dissimilarity) of customer profiles. Since there are three major variables, which are Recency, Frequency and Monetary per each customer, running the algorithm based on Euclidean distances would enable a visualization of clusters on a three dimensional Euclidean space where each point represents a customer lying on (R,F,M) coordinates in the space. Another initial requirement of K-means algorithm is the “K”, which is the number clusters. Because of the fact that there is no domain knowledge or expected number of clusters to supervise the analysis, the algorithm was run with different “K”

values and outputs compared to see differences between cluster assignments of customers.

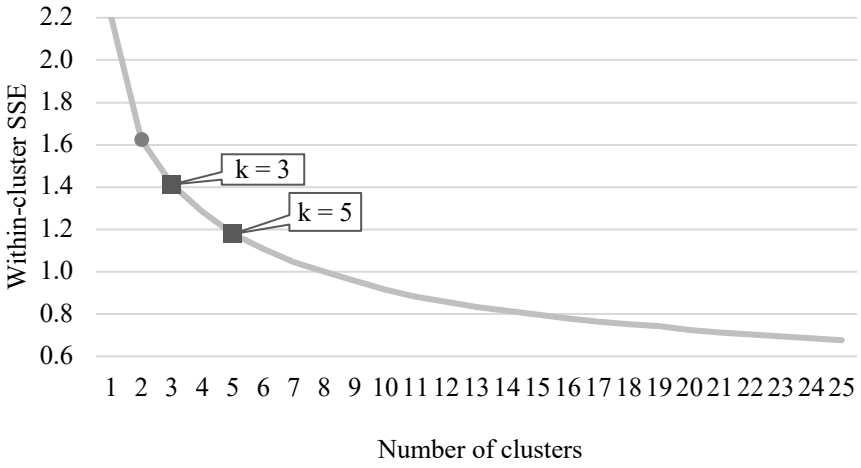
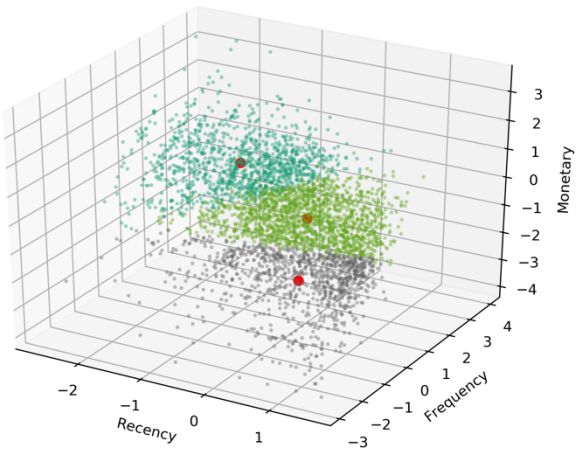
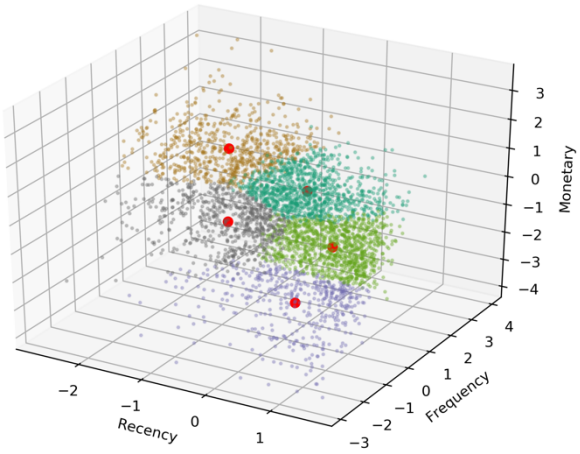


Fig. 21: Within-cluster sum of squared distances by number of clusters

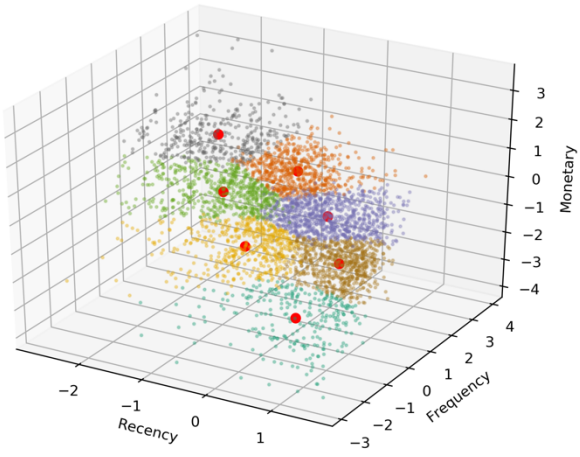
Fig. 21 represents an application of Elbow method that is the visual interpretation of overall within-cluster sum of squares with respect to number of clusters generated. As seen on the graph, the curve bends dramatically when two clusters are created. However, considering the highly skewed distribution of the data set of 3936 customers, only two clusters would not give a very homogeneous subgroups. The posterior elbow points are three and five clusters which also decrease the distances between cluster centres and cluster members significantly. Other initial decision for running K-means clustering is the seeding process of initial centroids such that it can be either performed traditionally with determining completely random k points on Euclidean space of data set or the initial clusters can also be determined with respect to the seeding method of k-means++. K-means++ is the preferred method in this analysis for determination of initial clusters as described in section 2.3.4 instead of random selection of all initial centroids. With seeding the algorithm with respect to k-means++, the algorithm iterated four seven and four times when number of clusters are 2,3 and 5 respectively until total sum of squared distances (TSS) of data points to their closest cluster center cannot be decreased with the same number of clusters anymore. Final total TSS per each clustering algorithm run shows that k-means clustering with five clusters have a better within-cluster similarity compared to two and three clusters as expected. Although five clusters would be statistically a better solution for grouping similar customers, the business relevance and practicality of the output must be considered to understand which output serves better for further applications. Fig. 22 shows customer data lying on three-dimensional Euclidean space with respect to recency, frequency and monetary feature values. The different colours represents different customer segments assigned by K-means++ clustering and the red points are the final central points of each customer segment. Visual representation of the clustering with K-means clearly shows some characteristics of the algorithm. First of all, customers with extremely low or high R,F or M values were grouped under same clusters with the customers who appear closer to the center. It is a natural outcome of K-means algorithm which does not identify or exclude outlier members and tends to create fairly evenly distributed clusters.



(a)



(b)



(c)

Fig. 22: Results of (a) 3-means, (b) 5-means and (c) 7-means clustering

Fig. 23 represents how each feature changes from cluster to cluster by different total number of clusters. The initial clustering with three clusters create three customer segments where each feature shows a similar trends. Customers with highest monetary, recency and frequency values are assigned to Cluster 0 and customers with lowest R,F,M values are grouped in Cluster 2. Increasing the number of clusters results with new customer segments with alternating recency values. As seen on the second and third blocks of box

plots on Fig. 23, while the frequency and monetary features show a similar distribution among clusters, recency feature differentiate the clusters in the center. For instance, Cluster 2 and 3 in 5-Means clustering, are two customer segments in which one group made purchase more recently than the other. 5-Means clustering reveals more alternating customer segments in terms of recency of purchase behaviour. The customers grouped in Cluster 4 contributed to revenue slightly higher than the ones in Cluster 3. However, the customers in Cluster 3 made purchase from company more recently. The grey marks on the top and bottom end of the boxes evidence the inclusive behave of K-means clustering against outliers such that the customer segments with the highest monetary, frequency or recency features show very skewed distributions of these features. The heterogeneity of these customer segments can only be decreased with increasing the number of clusters assigned. Still, the analysis provided number of key inferences about the behaviours of customers in the data set:

- The customers who purchase most frequently and contribute the revenue most are assigned to the same cluster groups.
- The individual customers, who are likely to be small retails are assigned to the same clusters. However, since the algorithm is assigning all customers to exactly one cluster, outliers are also assigned to the same cluster with the small retails. Similarly, the customers with highest revenue contribution are also assigned to the cluster together with customers who have relatively lower revenue contributions.
- Increasing number of clusters improves the homogeneity within clusters. Creating five clusters revealed new information about different purchase trends. Especially the customers on the center of customer base in terms of R,F,M features are differentiated by purchase recency.
- K-means algorithm evenly distributed customers to each cluster so that there is no cluster with significantly lower number of cluster members.

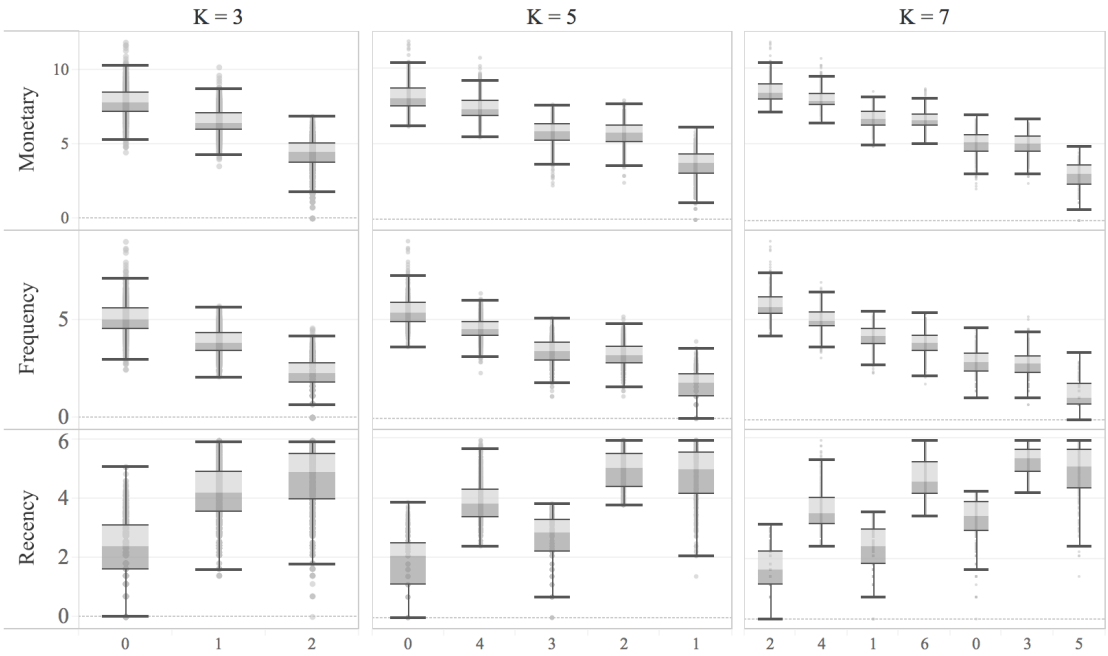


Fig. 23: R,F,M features of clusters compared by different number of clusters generated

3.5 Agglomerative Hierarchical Clustering

Hierarchical clustering offers more alternating clustering clusters compared to K-means algorithm. The dendrogram representing hierarchical relationships with clusters reveals that different proximity approaches populate alternative clusters in terms of size and homogeneity. For a better understanding of how each proximity approach define hierarchical relationships differently, each four approaches applied and visualized with dendrogram. Cut off distance points were selected based on the improvement of within-cluster similarities and placed into dendrograms with dotted horizontal lines. As shown on Fig. 24, when clusters extracted with hierarchical clustering using single linkage, the algorithm creates one very large cluster containing almost entire data set and other clusters with very few cluster members.

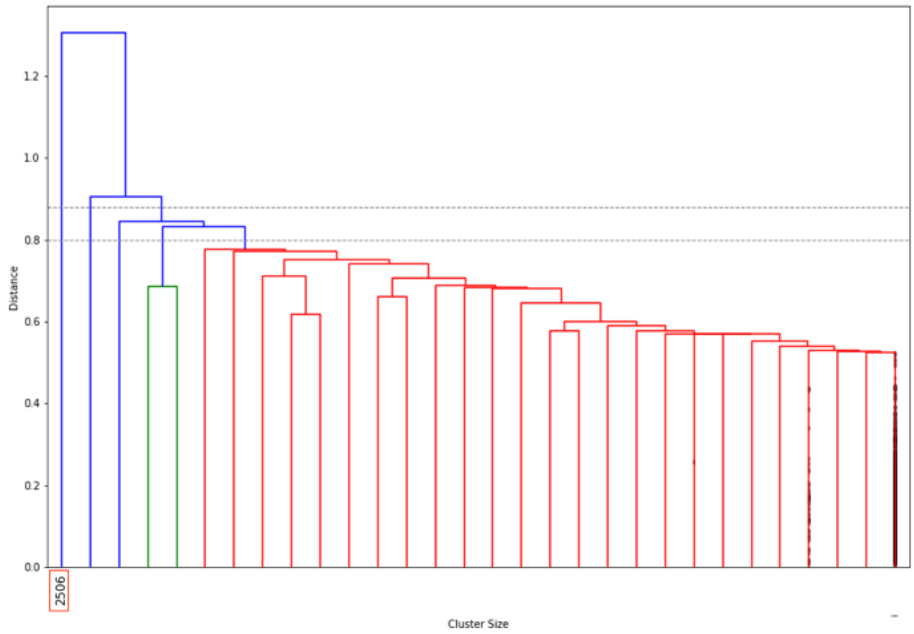


Fig. 24: Truncated dendrogram generated by single linkage hierarchical clustering

Fig. 25 represents the dendrogram created based on complete linkage approach of hierarchical clustering such that the cluster proximities are measured by the farthest members of clusters to join pairs of clusters. Complete linkage approach helps generating more evenly distributed clusters and increasing the number of clusters until 7 improves the within cluster distances significantly in contrast to single linkage. Dendrogram implies that distance between joined clusters increase significantly firstly seven and secondly four clusters are created. In other words, if four or seven clusters are requested by hierarchical clustering with complete linkage proximity approach, algorithm starts to show more dissimilar clusters compared to previously merged and generated clusters.

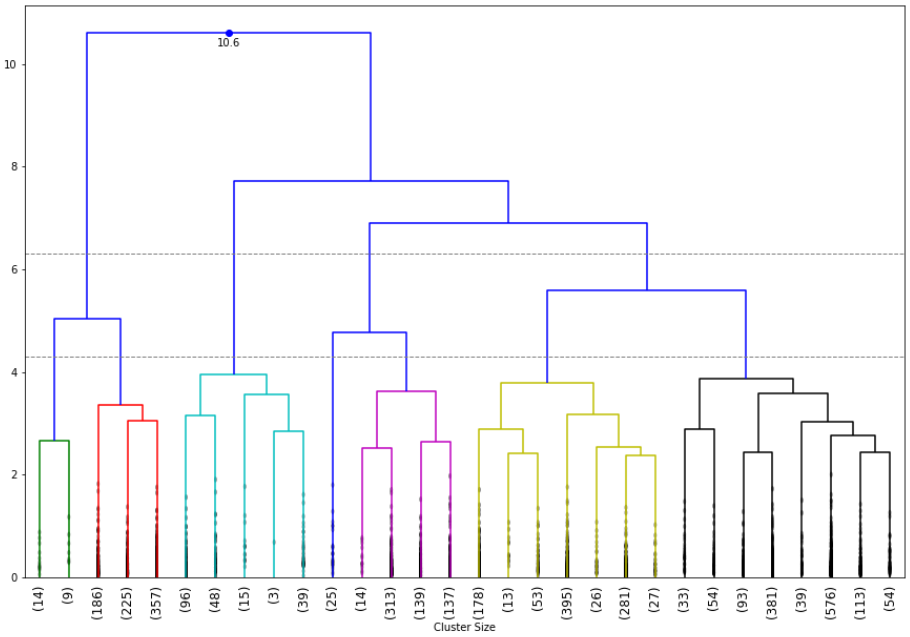


Fig. 25: Truncated dendrogram generated by complete linkage hierarchical clustering

As shown on Fig. 26, average linkage clustering results with a slightly different dendrogram in which one cluster contains only four members. Based on the distance between merged clusters, six clusters could assign merge the data points that are increasing the heterogeneity significantly to new clusters such that the dissimilarity significantly.

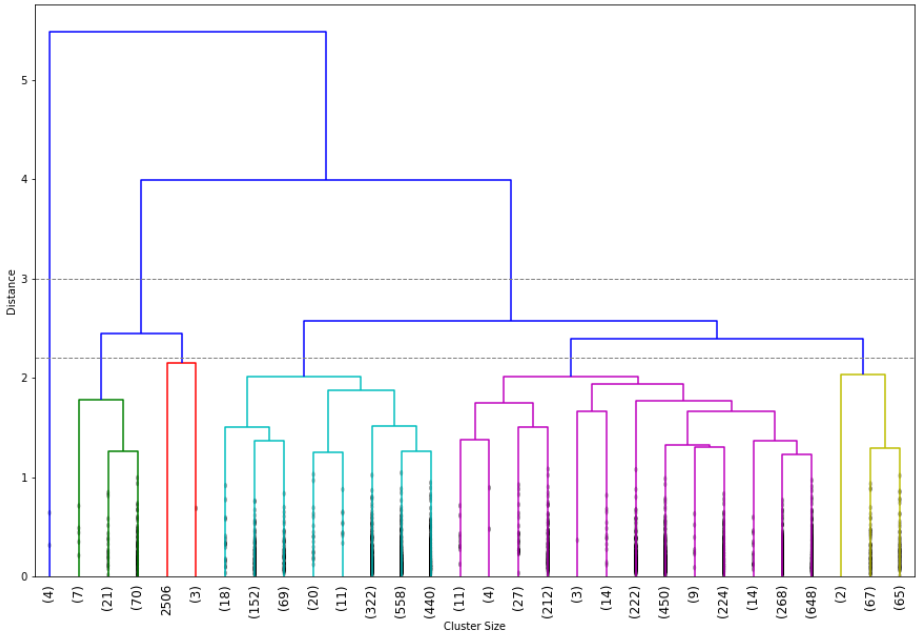


Fig. 26: Dendrogram generated by average linkage hierarchical clustering

Last alternative approach, which is measuring proximities with Ward's method, show the most evenly distributed cluster sizes without differentiating edge data points from least extreme data points. Cluster assignments are made evenly and there is no cluster with very few cluster members. The general picture of the output is quite similar to K-means clustering output. For instance, when cluster assignments are done with using six clusters, all clusters demonstrate a similar within-cluster distance as shown on Fig. 27.

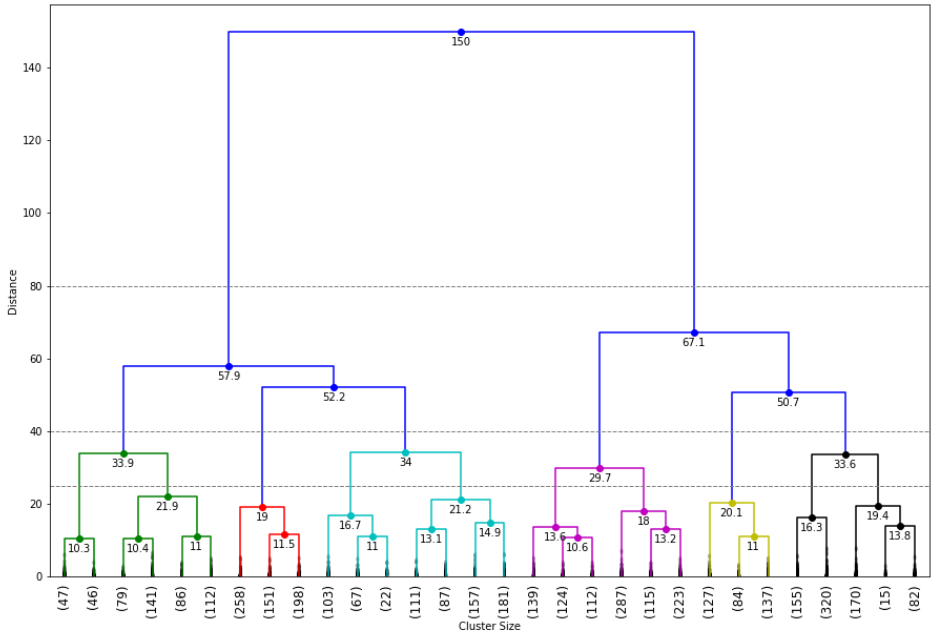
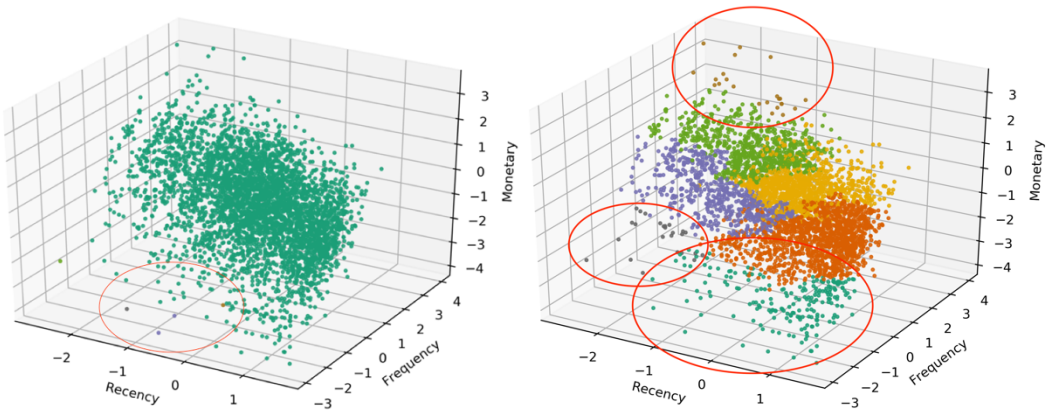


Fig. 27: Dendrogram generated by hierarchical clustering with Ward's proximity approach

Fig. 28 demonstrates the customer data on Euclidean space with assigned clusters using the appropriate cluster quantities inferred by dendrograms generated. Fig. 28 (a) reveals that single linkage method does not serve efficiently with the given customer data set such that most of the customers are assigned to some cluster and the method only separated very few customers who have significantly lower revenue contribution. On the other hand, the next two figures (a) and (b) shows that average linkage and complete linkage approaches detected customer segments with more similar customers. Moreover, these methods created several small customer segments with homogeneously high and low revenue contribution, purchase frequency and recency. Compared to the output of K-means clustering, average and complete linkage hierarchical clustering methods provided better solutions in terms of identifying customers who can be considered as outliers. As seen on Figure 28 (c), when six customer segments are created with average linkage approach, each customer segment show a similar and interpretable characteristics in context of purchase behaviours. Four customers with highest purchase frequency and monetary values are separated from other customers and formed a single cluster. Similarly, there are two other small clusters which consist of only individual customers without any other customer making purchase frequently, i.e., a retailer. Complete linkage method provides a very similar solution with creating slightly larger clusters compared to average linkage as seen on Figure 28 (b).



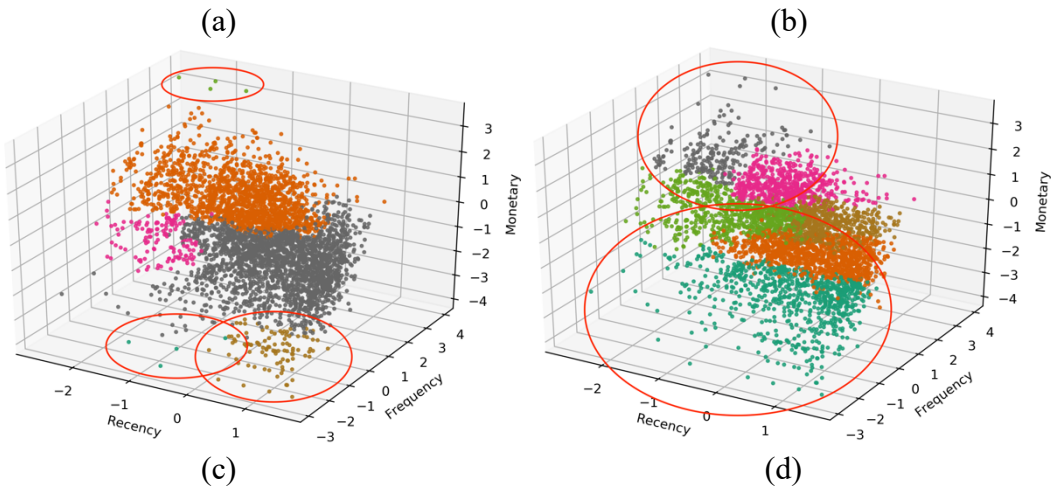


Fig. 28: Customer RFM data set labelled by hierarchical clustering using (a) Single linkage; (b) Complete linkage; (c) Average linkage; (d) Ward's method approaches

Last method using Ward's proximity approach shows a similar cluster densities with K-means algorithm by the nature of measurement since the distances between clusters are calculated by using sum of squared distances similar to latter. It creates wider customer subgroups including customers on head and tail ends into bigger segments such so that the method does not introduces niche customer segments separating outliers from average customers. In this respect, Wards's approach can be considered as less successful at identifying niche customer segments but still as plausible as K-means algorithm. Taking each approach's inputs into account, the key outputs of customer segmentation with hierarchical clustering can be listed as follows:

- Selected proximity approach is crucial for extracting clusters from customer data set. Each approach provides different cluster assignments so both dendrograms and labelled data sets should be examined for generating alternative solutions.
- Even though the approach does not dictate a best number of customer segments, heuristic interpretation of dendrogram can reveal a meaningful number of customer segments.
- Hierarchical clustering suggests customer segments with diverse sizes such that customers showing very different purchase behaviours can be differentiated from average customers using complete or average linkage approaches for cluster proximity measurement.
- Hierarchical clustering can provide a very similar perspective with K-means clustering with Ward's proximity approach. It introduces relatively homogeneous customer segments but does not separate outliers from average customers.

3.6 Density-based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN approach is the most distinctive and challenging approach among the other applied methods in context of segmentation based on real world customer purchase data. The analysis is a challenge for the method due to the lack of different density areas in customer purchase related data set. Another difficulty is estimating parameters for running the algorithm without a domain knowledge or definable number of clusters.

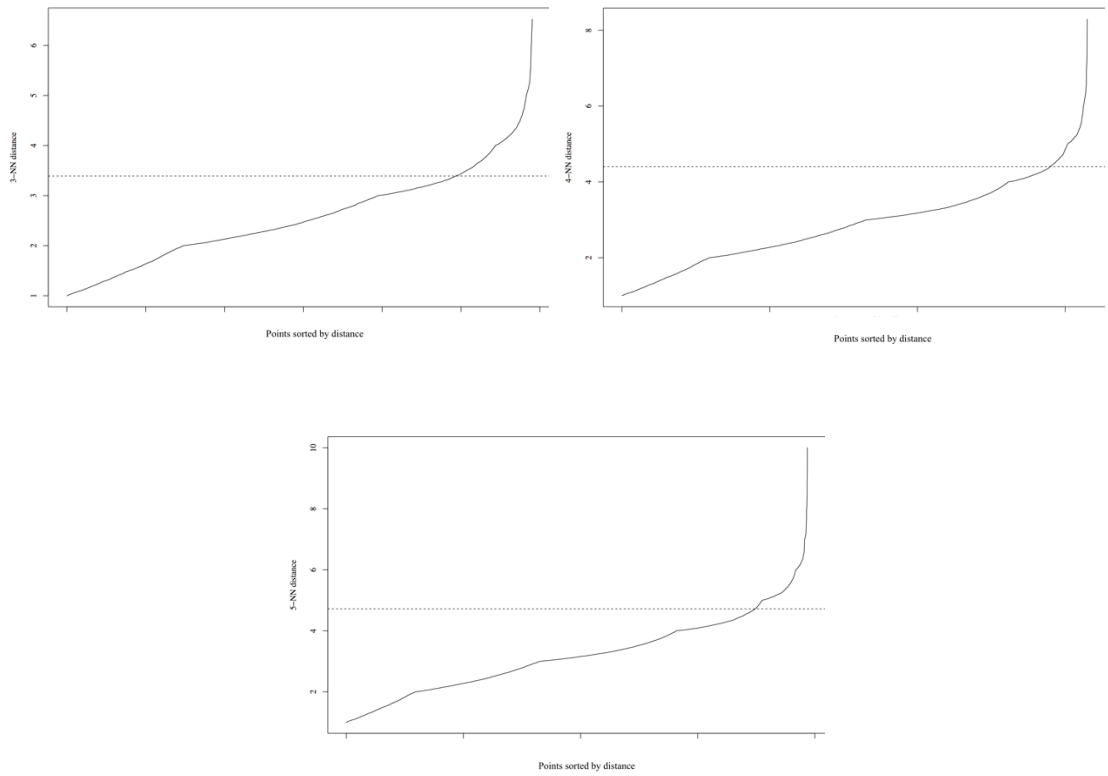


Fig. 29: K-dist curves for 3,4 and 5 nearest neighbours

Fig. 29 represents k-dist curve for three, four and five nearest neighbours respectively. The curves do not have a distinctive elbow points that can help for deciding on a reliable starting parameters such that when ε values are selected based on assumed threshold, that are marked with dashed lines over curves, concludes with one single cluster and few noise points. However, it does not mean that the method is not giving any information regarding homogeneous subgroups in customer base. By the nature of unsupervised learning phenomena, experimentations with heuristic approaches can reveal useful information about the given data set. Thus, the method experimented with variety of different MinPts and ε values. As it has been proposed with the approach, the user of the algorithm can interactively determine parameters with experimenting on the analysis tool.(Ester *et al.*, 1996) Figure N shows some of the cluster assignments with using different set of parameters. The huge majority of the customers, that are remarked with green markers, was not assigned to any cluster by the algorithm. Besides, different parameters help to identify few small subgroups of data points which represent a dense area on feature space. There are for instance, some customer groups with high monetary and low recency or low monetary and high recency values creating a dense subgroups. None of the outputs consider outlier customers as a part of a subgroup with DBSCAN approach. With parameters ε, MinPts selected as 0.175 and 4, respectively, algorithm started to identify dense areas as seen on Fig. (b). Increasing the ε to 0.2 included more points into defined subgroups and represented more dense subgroups. However, increasing both ε and MinPts cancelled these cluster assignments. Another direction with a small ε (0.1) and higher MinPts (5) created clusters in the core of the distribution, which in context of customer segmentation would not be helpful to understand customers better.

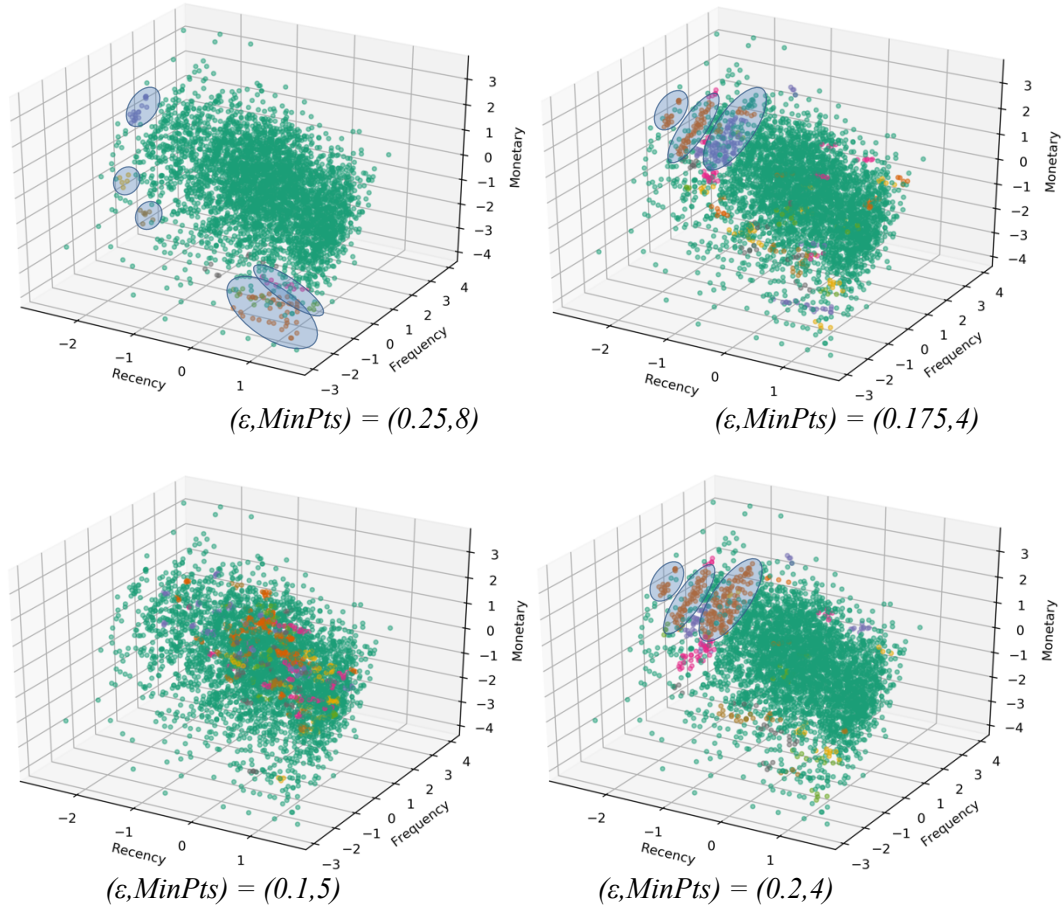


Fig. 30: Cluster assignments with different DBSCAN parameters

Experiments with DBSCAN method provided several informative takeaways in context of subject customer base:

- When the high density of customers show a similar purchase behaviour, density based approach do not introduces evenly distributed customer segments.
- Unless a subgroup of customers show a highly similar behaviour, density based approach is not inclusive against outliers.
- High number of experiments reveal unique homogeneous subgroups that are not grouped by any other unsupervised learning technique applied.

3.7 Clustering using Gaussian Mixture Models (GMM)

As described in relevant chapter, GMM approach can be considered as a developed version of K-means clustering with a probabilistic foundation. Correspondingly, it is highly possible to see a similar picture by assigning clusters to data set based on introduced output of the analysis. Initially, to determine the number of multivariate Gaussian distributions behind the data set, BIC and AIC criterion was used. As shown on Fig., five components would be suitable to explain given data set since the mixture of five models lead to the minimum BIC value with a relatively low AIC. Higher number of components can also be used with only considering AIC criteria however, this approach would be only possible with ignoring the higher penalty given by the nature of BIC criteria.

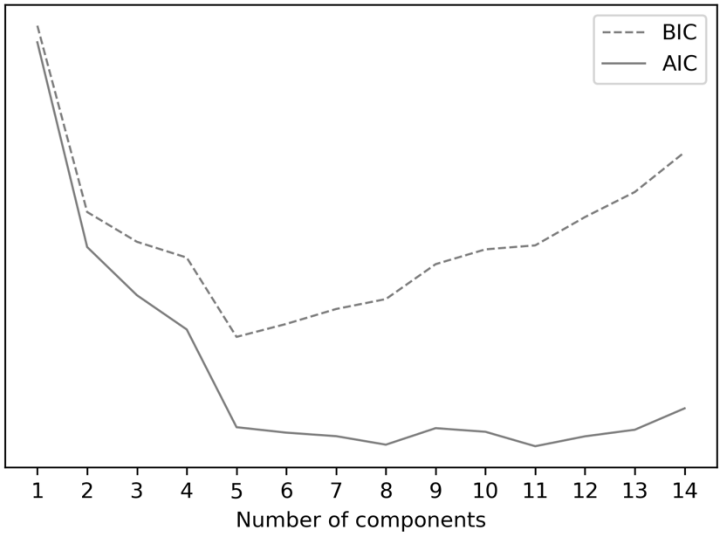


Fig. 31: BIC and AIC scores by number of components of assumed mixture of Gaussian Models

Explaining the data set with five components mean that there are five Gaussian distributions with different probability functions which take different responsibilities to explain each customers' purchase behaviour. Though the method makes soft cluster assignments to each data point, the model that takes the highest responsibility for explaining a customer's purchase behaviour is considered as the best segment for associated customer. Hereby, the resulting cluster assignments are slightly similar with the output of K-means algorithm with five clusters. However, the difference between hard cluster assignment from K-means and new GMM approach is apparent with different shapes of clusters. As shown on Fig, the general picture is similar with K-mean clusters but the edges of cluster areas are not as sharp as K-means clusters. There are overlapping data points which are assigned to different clusters. This flexibility across clusters slightly affected the characteristics of each customer segment. As compared on Fig., the revenue contribution (monetary) and purchase frequency features are more various across clusters just because all of the customers who purchase more, are not assigned to the same segments but some of them was distributed to others due to differentiating supposedly hidden Gaussian distributions behind the data set.

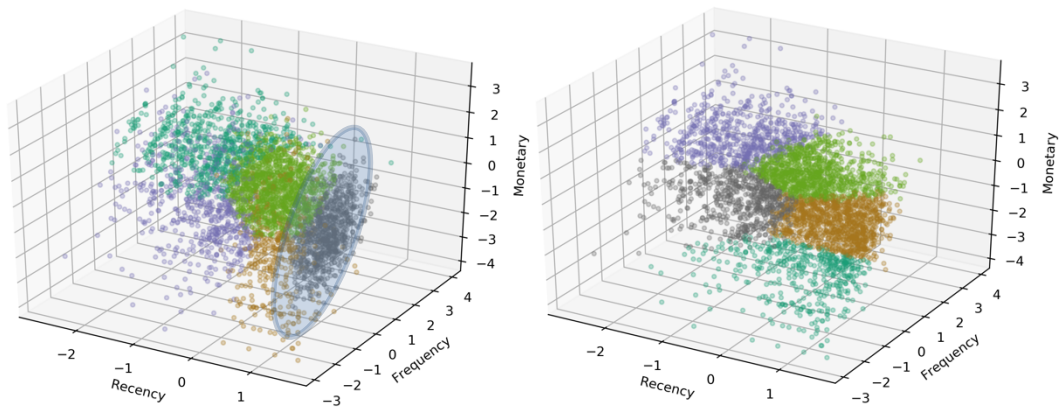


Fig 32: Five clusters assigned by (a) GMM approach and (b) K-means clustering

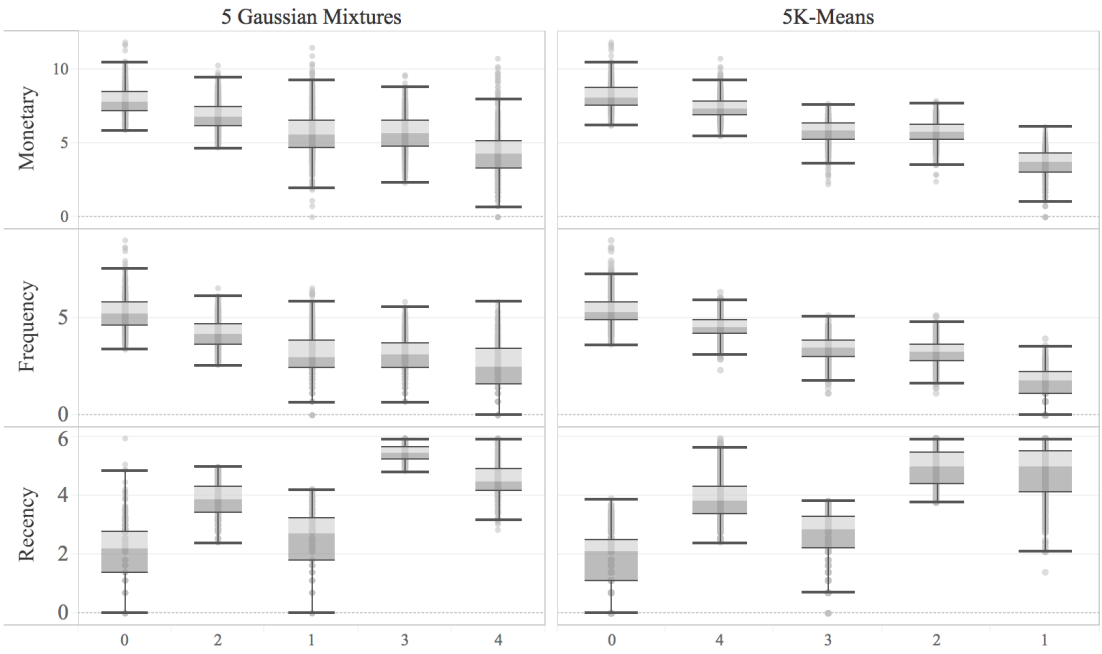


Fig. 33: Monetary, Frequency and Recency values of GMM components and K-means clusters compared

Gaussian mixture model provided another alternative set of customer segments with several different learnings:

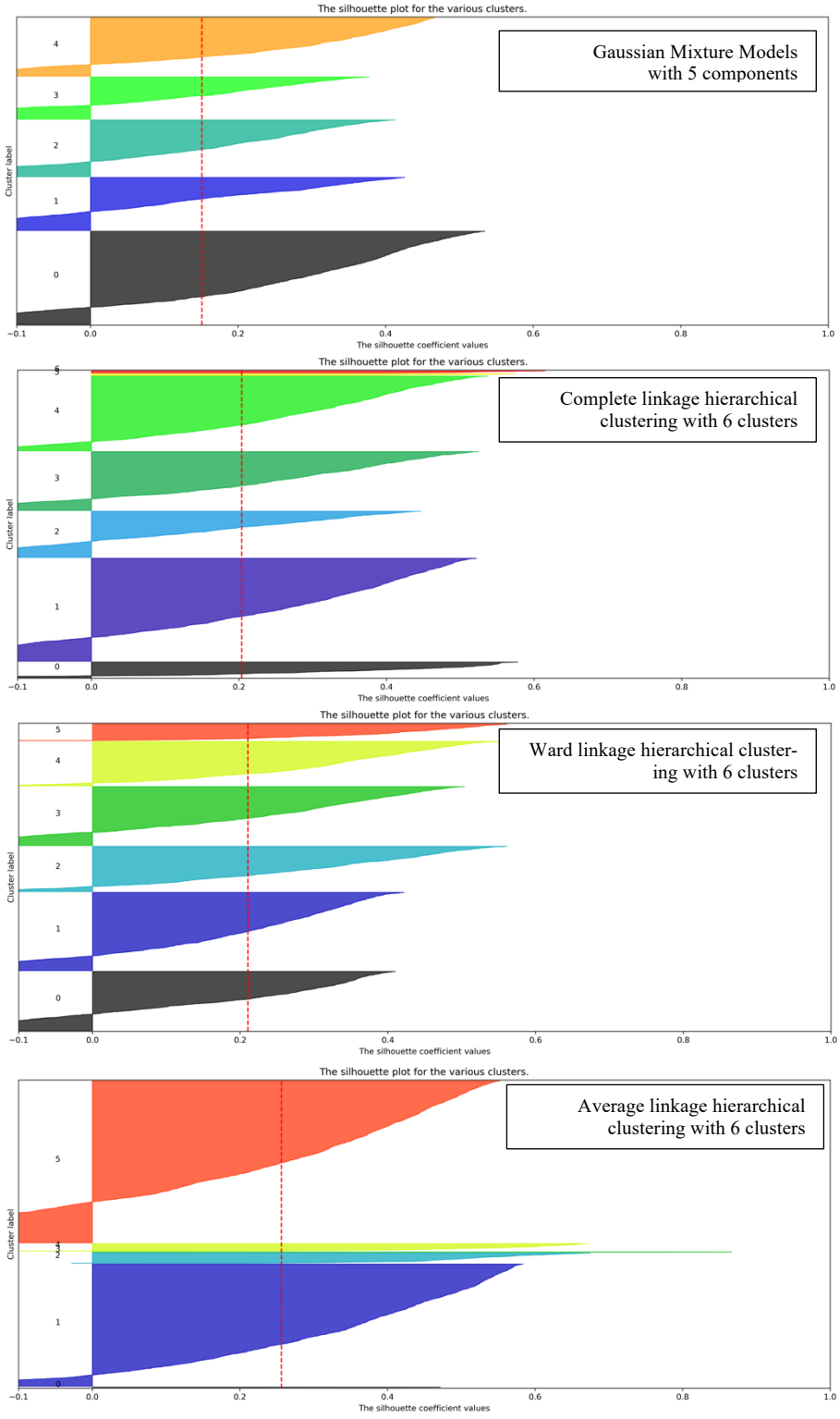
- Even though the approach does not suggest a hard cluster assignments, if a Gaussian Mixture Models, which explains a customer's purchase behaviour best, is considered as the respective customer's segment, the segmentation analysis provide a similar picture that can be interpreted by K-means clustering.
- Some customers are assigned to completely different clusters such that all customers who can be considered as outliers by a feature are not in the same clusters.
- Gaussian Mixture Model could suggested a non-spherical shaped cluster, which was not suggested by any other unsupervised learning methods used.

3.8 Evaluation of the Methods

With the obtained segmentation suggestions from variety of methods, the performance of each method in context of customer purchase behaviour similarities can be evaluated with the assistance of Silhouette scores. Even though the validation against a supervised data is not available in applied unsupervised learning methods, Silhouette score can provide knowledge on the homogeneity within generated customer segments and heterogeneity between neighbour, or candidate segments to give a better understanding about the differences caused by different perspectives. For evaluating the suggested customer segmentation solutions to given data set by different methods, silhouette plots can be analyzed to interpret how similar are the customers with other customers in the same segments. Silhouette scores are analysed with interpreting some basic indicators on silhouette plots:

- Vertical (red) dotted line represents the average silhouette score across entire customer base. The higher average silhouette score of generated clusters imply a higher overall customer segment homogeneity.

- Each differently coloured area represents the silhouette scores of respective cluster. Thus, it is expected to have more coloured area over the average silhouette score for more homogeneous customer segments.
- Negative values show that these cluster members are increasing the heterogeneity of their cluster such that they are not similar to the center of customer segment. Therefore, these areas should be as small as possible for the purpose of homogeneity.



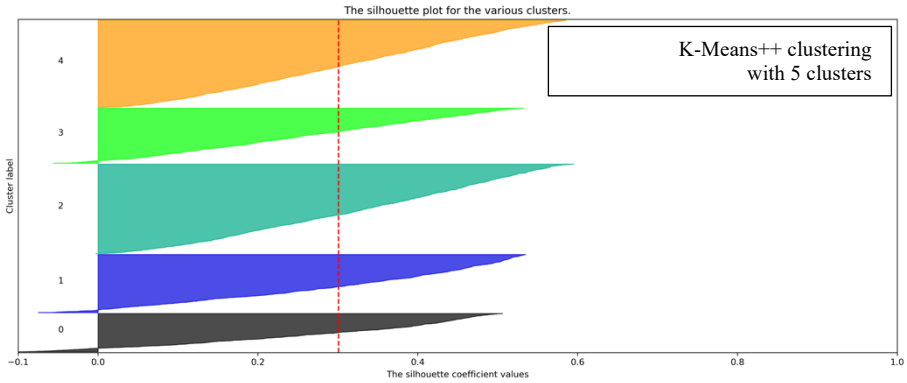


Fig. 34: Silhouette plots of five different clustering approaches

Fig. 34 represents the overall and cluster level silhouette scores of five different methods with different parameters. K-means++ clustering algorithm with five clusters resulted with the highest average silhouette score. Besides, almost all of the customers are assigned to a cluster where the customer's purchase behaviour features are most similar to its own cluster's center. However, there are still few customers with negative silhouette scores. It is an outcome of inclusive algorithm of K-means clustering which process the cluster assignments without excluding any outliers such that the size of clusters are evenly distributed and there is no nich cluster consisting of highly extreme samples. Second highest average silhouette score is seen on clustering assignments with average linkage hierarchical clustering. The main reason is that the cluster two, three and four which cover very small areas on plot, are the very small clusters including outliers that are not detected and separated by K-means or GMM approaches. As shown on Fig. 34, with a high area over average silhouette score, these clusters show very high homogeneity within cluster and separated significantly from the neighbouring clusters. Nevertheless, these methods are as good as K-means clusters with the data points that are closer to center without generating large number of clusters. It can be interpreted by large areas under zero silhouette score. Lastly, GMM suggested a similar number of components with K-means algorithm however, it has the lowest silhouette score compared to other methods by the very nature of the algorithm. While the other methods seek a similarity with using distance calculations, GMM identifies five groups of customers which are likely to be samples from five multinormal distributions. Thus, they do not necessarily show a distance-wise similarity like the other methods and it is not suitable to conclude that the suggested model is not performing accurately.

Summary

Variety of methods with different parameters suggested different number of customer segments with more or less homogenous characteristics. Due to the different natures of the methods and uncertain global optimum, comparison and validation of the applied methods without a supervision makes these unsupervised learning methods very complex statistical problems. However, with a good understanding of available data set and experimenting multiple number of methods make the entire unsupervised learning process more of an explanatory analysis rather than a predictive one. In context of customer segmentation, alternative learning approaches provide a general understanding together. Following key takeaway summarizes the findings from applied unsupervised learning methods to identify customer segments based on purchase behaviour:

- RFM analysis is a traditional marketing phenomena that can be used to generate customer purchase behaviour features from historical purchase data.
- Euclidean distance is a plausible distance measurement for the data set with several numeric variables. The results can be easily interpreted by the visual representation on Euclidean space with assigned customer segment labels attached.
- K-means algorithm successfully divides the customer base into homogeneous subgroups without too many iterations and computational effort. Algorithm tend to create all customers into similarly sized customer segments with generating moderate number of clusters. Thus, it does not successfully handle outliers. Increasing the number of clusters helps identifying more interesting patterns associated with subgroups without extreme feature values.
- Hierarchical clustering can show variety of different results based on the applied proximity approach. Except the Ward's method, which shows a similarity with K-means in terms of measuring cluster proximities, three common approaches detect the outliers more successfully than K-means method. However, it shows more heterogeneity across other customer segments.
- Density based approach is useful for detecting different density areas in data set, which could be more useful in different context, i.e., image recognition however, when the data set does not have clearly separated density areas, it suffers from detecting subgroups that can be used as customer segments for further business objectives. Still, density based approach detects most different patterns in feature space of purchase behaviours which can be more informative for detecting interesting patterns in data set. Also, the method challenges user with selecting the correct parameters for initiating the algorithm which necessitates number of experiments to obtain information.
- GMM using EM identifies customer segments even if the data set shows an equal density across whole area when projected on Euclidean space. Similar to K-means, it does not exclude any outliers or noise points. In fact, it detects hidden Gaussian distributions that the customers are coming from which results with a similar segment characteristics with K-means but differently soft cluster assignments with respect to probabilistic foundation of the algorithm.
- Evaluation and validation of the results requires a domain knowledge and heuristic approach for making the outputs of analysis useful in accordance with the objective of the organization. Thus, multiple number of methods can be applied and the mixture of outputs can be evaluated with the input from experience and other stakeholders.

Attachment

List of literature

Akaike, H. (1974) 'A New Look At The Statistical Model Identification', *IEEE Transactions on Automatic Control*, 19(6), pp. 716–723. doi: 10.1109/TAC.1974.1100705.
Arthur, D. and Vassilvitskii, S. (2007) 'k-means++: the advantages of careful seeding', in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. doi: 10.1145/1283383.1283494.
Berkhin, P. (2002) *Survey of Clustering Data Mining Techniques*. San Jose, CA. Available at: https://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf (Accessed: 31 October 2018).
Bishop, C. M. (2006) 'Mixture Models and EM', in *Pattern Recognition and Machine Learning*. 1st edn. Springer, pp. 423–444.
Dempster, A. P., Laird, ; N M and Rubin, ; D B (1977) 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), pp. 1–38. Available at: http://web.mit.edu/6.435/www/Dempster77.pdf (Accessed: 30 November 2018).
Deng, H. and Han, J. (2013) 'Probabilistic Models for Clustering', in Aggarwal, C. C. and Reddy, C. K. (eds) *Data Clustering: Algorithms and Applications*. 1st edn, pp. 61–82.
Dolnicar, S., Grün, B. and Leisch, F. (2018) 'Market Segmentation', in *Market Segmentation Analysis*, pp. 3–9. doi: 10.1007/978-981-10-8818-6_1.
Ester, M. *et al.* (1996) *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. München. Available at: www.aaai.org (Accessed: 20 October 2018).
Grigsby, M. (2016) 'Introduction to segmentation', in *Advanced Customer Analytics : Targeting, Valuing, Segmenting and Loyalty Techniques*. 1st edn. London: Kogan Page Ltd, pp. 216–230.
Han, J., Kamber, M. and Pei, J. (2011) 'Cluster Analysis: Basic Concepts and Methods', in *Data mining : Concepts and Techniques*. Elsevier Science.
Hartigan, J. A. and Wong, M. A. (1979) 'A K-Means Clustering Algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pp. 100–108.
Igual, L. and Seguí, S. (2017) 'Unsupervised Learning', in *Introduction to Data Science*. Springer, Cham, pp. 115–139. doi: 10.1007/978-3-319-50017-1_7.
James, G. *et al.* (2013) 'Unsupervised Learning', in *An Introduction to Statistical Learning*. 1st edn. Springer Science+Business Media New York, pp. 373–418. doi: 10.1007/978-1-4614-7138-7_10.
Kotler, P. *et al.* (2017) 'Customer-driven marketing strategy: creating value for target customers', in *Principles of Marketing European Edition*. 7th edn. Pearson Education Limited, pp. 190–224.
Lloyd, S. (1982) 'Least squares quantization in PCM', *IEEE Transactions on Information Theory*, 28(2), pp. 129–137. doi: 10.1109/TIT.1982.1056489.
McDonald, M. (2013) *Market Segmentation : How to Do It, How to Profit from It*. 4th edn. John Wiley & Sons.
Online Retail Data Set (2015) *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/ml/datasets/Online Retail (Accessed: 3 September 2018).
Roberts, J. H., Kayande, U. and Stremersch, S. (2014) 'From academic research to marketing practice: Exploring the marketing science value chain', *International Journal of Research in Marketing*, 31, pp. 127–140.
Rousseeuw, P. J. (1987) 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20, pp. 53–65. Available at: https://pdfs.semanticscholar.org/f168/41e022038e94a59f7e0a82002102b78d79a4.pdf

(Accessed: 1 October 2018).
Sarkar, D., Bali, R. and Sharma, T. (2018) 'Customer Segmentation and Effective Cross Selling', in *Practical Machine Learning with Python*. Berkeley, CA: Apress, pp. 373–405. doi: 10.1007/978-1-4842-3207-1_8.
Schwarz, G. (1978) 'Estimating the Dimension of a Model', *The Annals of Statistics*. Institute of Mathematical Statistics, 6(2), pp. 461–464. doi: 10.1214/aos/1176344136.
Smith, W. R. (1956) 'Product Differentiation and Market Segmentation as Alternative Marketing Strategies', *Journal of Marketing*. American Marketing Association, 21(1), pp. 3–8. doi: 10.2307/1247695.
Tan, P.-N. *et al.* (2018) 'Cluster Analysis: Basic Concepts and Algorithms', in *Introduction to data mining*. Pearson.
Wind, Y. (Jerry) and Bell, D. R. (2008) 'Market Segmentation', in *The Marketing Book*. Butterworth-Heinemann, pp. 222–244. doi: 10.1016/B978-0-7506-8566-5.50015-7.

Statutory Declaration

I herewith formally declare that I have written the submitted thesis independently. I did not use any outside support except for the quoted literature and other sources mentioned in the paper.
I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content.
I am aware that the violation of this regulation will lead to failure of the thesis.

Student's name	Student's signature
Matriculation number	Berlin, date