# Customer Behaviors Analysis


Ece Turgut

EC Utbildning

Examensarbete

2024 January

# Abstract

This study aims to understand and analyze customer behaviors and actions based on E-commerce big data. Results of such an analysis can be applied to improve business strategies, marketing, Customer relations management and even recommender systems. How does customer segmentation work in theory and in applications? In that regard this paper studied K means algorithms together with the applications of Recency Frequency Monetary Model. Customers are segmented into 4 groups based on their previous purchasing behaviors.

Keywords: e-commerce, customer segmentation, RFM, clustering

# Acknowledgement

I am grateful and feeling lucky to be surrounded by bright and creative minds at this phase of my life.

Thank you, my classmates, we probably spent a generous amount of time in the past two years online, all together in lectures and meetings. Of course, it is not the same as a good old classroom environment but still. Coding and handling data is a lonely work, so it was important for me to feel supported and not so alone.

Thank you, Anders Petterson for your endless chatter, good amount of sarcasm and your support, and thank you Riri as being my emotional support animal.

Special thanks to my thesis supervisor, sensei Antonio Prgomet for his extreme efforts for his students and such a good guidance! The person made me love statistics and the world of data. Feeling privileged to be your student.

# Abbreviations

B2B: Business to Business

B2C: Business to Customer

CLV: Customer Loyalty Value

CRM: Customer Relationship Management

DBSCAN: Density Based Spatial Clustering of Applications with Noise

ML:  Machine Learning

RFM: Recency Frequency Monetary Model

# Table of Contents


Abstract ..... II

Acknowledgement ..... III

Abbreviations ..... IV

1. Introduction ..... 1

1.1 Research Questions ..... 2

2. Theoretical Framework ..... 3

2.1 Customer Behavior and Segmentation ..... 3

2.2 Clustering Algorithms ..... 3

2.2.1 Hierarchical Clustering ..... 4

2.2.2 Density-Based Spatial Clustering (DBSCAN) ..... 5

2.2.3 Partitional Clustering ..... 5

2.2.4 K Means Clustering ..... 6

2.3 Recency Frequency Monetary Model ..... 7

3. Methodology ..... 9

3.1 Data Collection ..... 9

3.2 Exploratory Data Analysis ..... 10

3.3 Data Pre-Processing ..... 11

3.3.1 Standardisation with Standard Scaler ..... 12

3.4 Model Implementation ..... 12

3.5 Model Visualization: Inertia and Cohesion Score ..... 13

4. Results and Discussion ..... 15

4.1 K-Means Performance Metrics and Comparison ..... 15

4.2 Silhouette Analysis ..... 16

4.3 Model Evaluation ..... 17

4.4 Challenges and Limitations ..... 19

4.5 Model Improvements ..... 20

5. Conclusions ..... 21

5.1 Key Findings ..... 21

5.2 Future Improvements ..... 22

6. List of Figures and Tables ..... 23

7. Bibliography ..... 24

# 1. Introduction

Customer segmentation is one of the most innovative tools to help businesses to understand their customer groups and tailor-cut their marketing campaigns (H. Thanh, Nguyen S., Nguyen H., et al, 2023). It would help to define businesses' different customers segments, target specific ones and channel their focus and marketing on a science-based approach through this method. Effectively set segmentation would enable healthy and long-term Customer relations. (P. Anitha, Malini M. Patil, 2022) In many ways it could benefit both consumers and business to avoid unnecessary reclamation when not needed and detect when to make use of it more.

*Recommender systems* on the internet are becoming more and more popular these days. When we are on YouTube for example, algorithms suggest the user a list of contents depending on our previous watch-history. One of the ways of doing that is Clustering algorithms. (Mirenda, Viterbo, Bernadini, 2020)

Therefore, it is the aim of this paper to scrutinize data of e commerce customers and by using different models such as K-means, RFM (Recency, Frequency and Monetary) Model and get the bigger picture. While doing that, a prior exploratory analysis of dataset is crucial to understand the data. Therefore, I studied and eliminated redundant dimensions of data to focus on more important ones.

## 1.1 Research Questions

Through this essay I will try to answer these questions as some of which motivated me to do this research. Because I know from my own experience that I tend to get skeptical about almost every sales call and promotion message which probably does not make me an easy or ideal consumer. But I also could be a loyal customer once I find a good product or sustainable business of ethical practices, avoiding animal testing or keeping track of their carbon emissions for example. There are so many factors and even psyches that are affecting our purchases of course and it is not possible to check every box in this paper, which is not the aim either, in fact this one will mainly focus on historical transactions of customer data with more than 500K observations.

- How can we identify different customer behaviors and make data-driven decisions for businesses?

- What are the most common algorithms and methods to customer behaviors, customer detention and segmentation?

# 2. Theoretical Framework

## 2.1 Customer Behavior and Segmentation

Customer segmentation is evidently an important tool to understand customer purchase behaviors and it may lead to better CRM (customer relations management) when applied correctly. Preprocessed and one hot encoded customer data is divided into segments with similar features (subscribed or unsubscribed customer, gender clusters i.e.) might help to reveal more distinct patterns in terms of their purchasing habits.

Customer retention is another concept that I will investigate in this paper. The customer retention rate is a key terminology for businesses since it means keeping the customers and turning them in into regular buyers and preventing them switching to another competitor. In other words, it means customer loyalty.

According to Anitha and Patil (2022), in an era of tremendous competition and ever-increasing technologies, it is crucial for businesses to study and understand customer segmentation and strategies to increase customer loyalty at an enterprise-level.

## 2.2 Clustering Algorithms

Let's imagine we are doing a one thousand piece of jigsaw puzzle. A landscape puzzle with blue skies and endless meadows and a lake with ducks in it. Independent from how big or complicated the puzzle is we should first categorize those colors and patterns of puzzle pieces in different categories. It could be done in a different way based on similar colors and tones or edges, corners and inner pieces. There we did a clustering task.

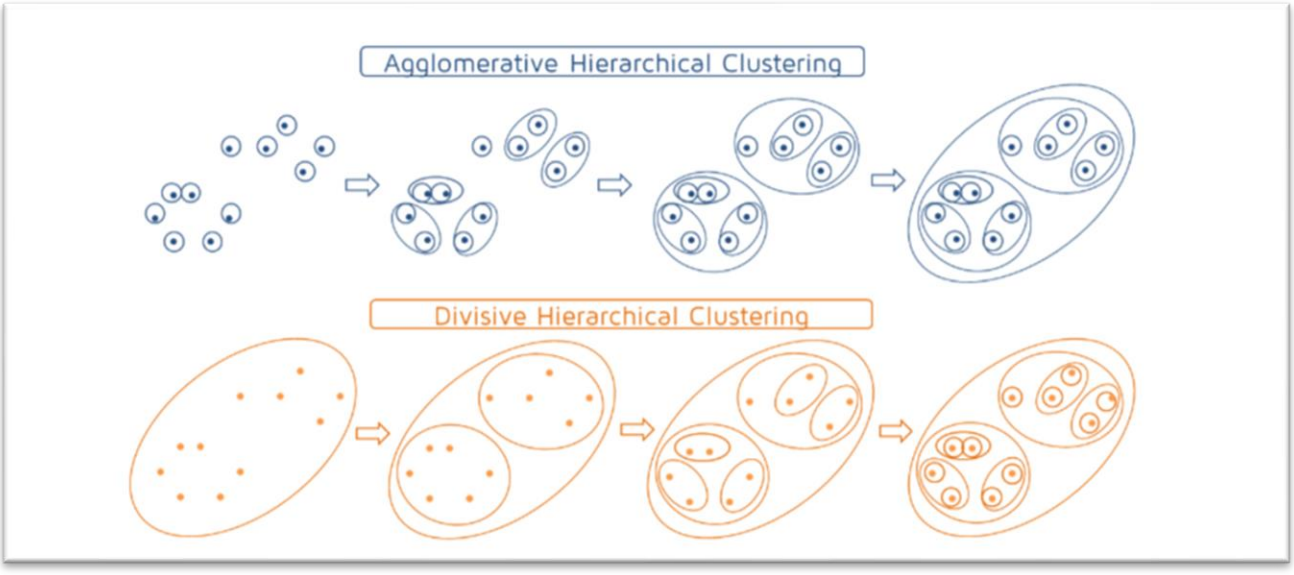
Clustering simply put is to “identifying similar instances and assigning them into *clusters* i.e. groups of similar instances. Just like in classification, each instance gets assigned to a group” (Aurélien, 2019).

It is part of unsupervised learning, and it represents most real-world data without being labeled and yet unstudied. Therefore, it is highly significant to research and understand its application areas. Before focusing on the actual algorithm used for this analysis it could be helpful to give a general look at Clustering algorithms in the field.

### 2.2.1 Hierarchical Clustering

Hierarchical types of clustering can be summarized by agglomerating (bottom-up) or divisive (top-down) approaches. The agglomerating approach collects data points based on their similarity and finally all is collected in single cluster. The Divisive approach, on the other hand, is completely opposite of agglomerating behavior. It starts out as one big single cell or cluster and splits down to the least similar cluster until the last data point as shown in Figure 1. (Lopez Yse, 2023)

Figure 1 Agglomerating and Divisive Hierarchical Clustering stages (Source: QuantDare)



### 2.2.2 Density-Based Spatial Clustering (DBSCAN)

DBSCAN is a density-based clustering algorithm which can identify noise unlike mean-shift clustering. It decides on clusters by checking the density of data points. While higher density areas are grouped as clusters, sparser areas are labeled as noise (Zangana Abdulazeez, 2023). One advantage is that that DBSCAN does not require a pre-set number of clusters (k hyperparameter) and it can also identify discretionary (arbitrary) clusters compared to other clustering methods. In that regard it offers some level of uniqueness.

(Practicus, 2018)

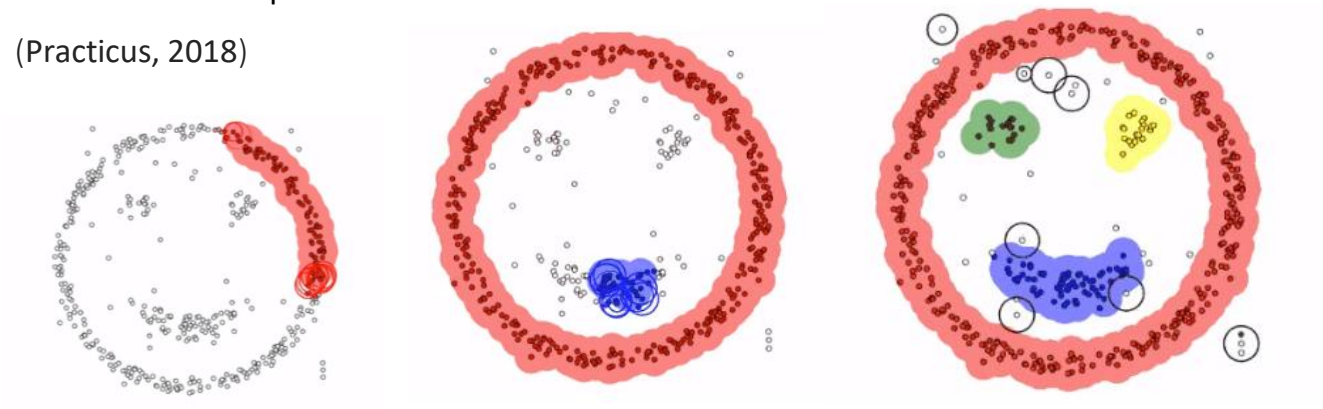


Figure 2 DBSCAN Clustering images, clustered density basis in a closed sphere. (Source: Primo.ai)

### 2.2.3 Partitional Clustering

In partitional clustering, data points are divided into non-overlapping clusters. Each instance can be a member of only cluster and each cluster must have at least one object. Similar way to hierarchical clustering, we need to preset a “k”, to define number of clusters. This type of clustering works well with big datasets and clusters that have spherical shapes.

### 2.2.4 K Means Clustering

This is the part that we will give more details since this paper focuses on K means as a tool to analyze Customer behaviors.

In the world of unsupervised learning algorithms K-means is highly recognized algorithm works to divide data space into clusters as we will call it intra-class, data clustered based on their similarity; and inter-class, when data points are least similar from each other and therefore placed in different clusters.

K means Algorithm theory is based on Euclidian geometry. “K-Means algorithm minimizes the distance from the values in the cluster to the centroids based on the theory of vector distances in Euclidean space.” (Thanh et al., 2023) The algorithm identifies a k centroid from the data and assigns non- overlapping data instances to each nearest cluster, and it is guaranteed to converge in a finite number of steps which is generally quite small. Centroid here represents a center of cluster which the mean value of the cluster and it may not be a member of the dataset.

“This way algorithm works iteratively until each data point clustered and closer to its own centroid rather than other clusters centroids, minimizing intra-cluster distance at each step” (Lopez Yse, 2023).

K-means is an ideal instrument compared in all Clustering algorithms with its ability of efficient partitions of data, which optimizes the process and resource use in Engineering application and specific optimization problems (Zangana, Abdulazeez, 2023).

When we think of Euclidean geometry between each data point, it is easier to imagine and formulate. Below is the mathematical formulation if this algorithm, minimizing the distance between the instances and the centroids based on Euclidian space. (Thanh et.al., 2023)

$$d(x, c_i) = \sqrt{\Sigma_{j=1}^{m}(x_i - c_{ij})} \quad (1)$$

One obvious challenge for K means model, since it is an unsupervised learning there are no ready-labelled data and no testing around hyperparameters as we can do in supervised models, and it is the algorithm job in the background which finds patterns for each data points and define the clusters and it is rather organic process with finding centroids.

## 2.3 Recency Frequency Monetary Model

"The RFM model is famous for dividing customers into segments based on analyzing past transactional data" (Thanh Ho, Suong Nguyen et al). Initials of model name RFM stands for most important KPIs/key parameters to track down customers spending behaviors. Recency of a customer's purchases, the Frequency of their buying activity, and the monetary value of their spending (Wu et al., 2019).

Table 1 RFM features, after preprocessing

| **Initial data** | **Transformed data** |
|---|---|
| Last transaction (date) | Recency (numeric) |
| Count unique order/transaction | Frequency |
| Total money | Monetary (total money/ count purchase) |

Table 2:Recency Frequency Monetary Model (RFM) features

| **Metric** | **Description** |
|---|---|
| Recency (R) | How recently customer made the purchase? |
| Frequency (F) | How often made the customer a purchase |
| Monetary (M) | How much did the customer spend for the purchase(s) on average |

The aggregation calculation made for Recency Frequency Monetary model will be explained in Model Implementation (Chapter 3.4). Before applying K means it is important to preprocess data and create the required features for RFM Analysis.

# 3. Methodology

In this Chapter the methods chosen for this analysis will be investigated. The phases will include stages of data collection and exploratory Data analysis phases of the research.

## 3.1 Data Collection

For this study I wanted to study people's purchasing behaviors but also their screen-time and gender details so I could derive more information. But the challenge was to find all these observations at the same time and there was no such real-world data both with transaction dates, price and quantity information and screen times of customers, gender and locations and all these metadata. Therefore, I decided to work with a dataset where I can get the transactions, amounts of their transactions and frequency of customer purchases. The aim of this selection was to create Customer segments using K-Means algorithms based on Recency, Frequency and Monetary data of customers.

Data for this analysis derived from Kaggle by Atthoriq Putra Pangestu[1]. This is a sales transaction of a UK origin E-commerce business. The London based shop has a wide range of decorative products, accessories, houseware, gifts and they have been selling their products online since 2007. The business model is both B2B and B2C as they sell both direct consumer and other outlet channels.

The raw dataset contained more than 500 thousand rows and 8 columns. Each columns data type and descriptions are as below:

[1] Link to my source data https://www.kaggle.com/code/atthpp/e-commerce-analysis-rfm-clustering-with-k-means#Closing:-Business-Recommendations

1. Transaction No (categorical): a six-digit unique number that defines each transaction. The letter “C” in the code indicates a cancellation.
2. Date (numeric): the date when each transaction was generated.
3. Product No (categorical): a five or six-digit unique character used to identify a specific product.
4. Product (categorical): product(item) name.
5. Price (numeric): the price of each product per unit in pound sterling.
6. Quantity (numeric): the quantity of each product per transaction. Negative values related to cancelled transactions.
7. Customer No (categorical): a five-digit unique number that defines each customer.
8. Country (categorical): name of the country where the customer resides.

## 3.2 Exploratory Data Analysis

This part of the analysis is the initial exploration phase of the data. Using pandas and NumPy libraries of Python, a general idea of patterns in dataset is discovered.
How many unique customers are there? What is the shape of the data frame? Are there any missing or null values in any rows? How many duplicates and so on, all these questions are generally answered in the exploratory data analysis.

According to my initial investigation Customer E-commerce data based in UK has 536.295 observations and 8 columns (as named in 3.1 Data Collection), 38 different countries. Over 4718 unique customers and equal number of Customer numbers. There were 55 missing values in the Customer No, which represented 0.01 of all *Customer No* column, so those rows with missing values were dropped. Sales transactions in this dataset cover a one-year period, from November 2018 to November 2019.

Both unique number of product and product names were 3753 pieces. There was no product category column. It was slightly challenging to observe Product type of customers in this dataset because when looked closely products text-heavy and wordy hard to categorize.

Negative digits of Quantity columns were discovered and eliminated from the data frame for further calculations, causing negative average and negative values for revenue column which was something undesired.

## 3.3 Data Pre-Processing

For K means algorithms to work, data must be encoded in numerical format. For this purpose, I used One hot encoding. This is one of the most used approaches while handling categorical data. Each dimension of the matrix is the number of states in this feature, and each dimension represents a specific state. All other state dimensions are zero due to this processing, and just one feature matrix dimension is typically asserted for each state (Yu et al., 2022).

As a common practice of preprocessing, missing and negative values were removed. Reducing and re-editing some variables names and types was also part of this phase. All these applications are Processing some of the column according to K means and clustering algorithms.

*Revenue* column generated for future references and model implementations, for this purpose Unit *Price* and *Quantity* columns were used.

$$Revenue = QTY * Unit\ price \quad (1)$$

Invoice Date, Transaction No and Product No were object type of data, so they were returned into date type and numeric type of data respectively.

### 3.3.1 Standardisation with Standard Scaler

Part of the process before running the K means model is to indeed scale the data to have the best results for this approach. For this purpose, I used stand scaled then I fed scaled data to the model.

Standardization of the data is important to prepare the dataset for the K means model, by which the means of zero[2] and standard deviations is 1. Taking this step of scaling the data helps the algorithm to converge faster than usual.

## 3.4 Model Implementation

After data is explored and preprocessed and K means algorithms were implemented based on this new data frame of RFM. In the following paragraph I explain how "rfm_data" is generated. The brief methodology of what RFM model is given in Chapter.2.3.

For RFM analysis aggregated function is used. This aggregation is done after calculating the Recency columns first. Recency metric was based on setting reference day as one day after the latest purchase of each Customer. Frequency metric calculates by counting each customer as number of occurrences in each customer segment they fall in to. Monetary value sums up customers' purchases and categorizes and spreads them into their segments according to the Revenue they generated.

[2] For each future x in the dataset: $z = \frac{x-u}{\sigma}$ where z is standardized value, x is original value, μ is the means of x, σ standard deviation of x. (Source: Statistics How to)

After this aggregation, the new data frame was saved as "*rfm_data"* and I applied the K Means model based on this data frame.

Before applying K means it is important to decide on the k parameter, which represents the number of clusters. For this analysis I use the methods of Inertia curve (or known as elbow curve) and Silhouette coefficient score analysis.

**Model Performance Metrics**

As we apply unsupervised algorithms of K means there is a level of ambiguity around k hypermeter. But it can be overcome by running two different methods and visualizations, namely Inertia (Elbow score) and Silhouette score analysis.

Inertia as defined here, within-cluster of sums of squares is an indication of how accurate each cluster is.

## 3.5 Model Visualization: Inertia and Cohesion Score

Determining ideal number of cluster (k) with Elbow method. (Figure 3) Later on, Distortion score of data points is also visualized in addition to the elbow analysis as seen on Figure 4.

When plotting Inertia for different values of k (number of clusters) you observe a decreasing trend of k. This is due to the increased number of k creating a distortion which means that data points of one cluster and neighboring clusters distances centroids are getting closer. Therefore, borders of segmentation sphere are getting blurry, which is undesired. We need a lower value of distortion for well-defined segmentation.

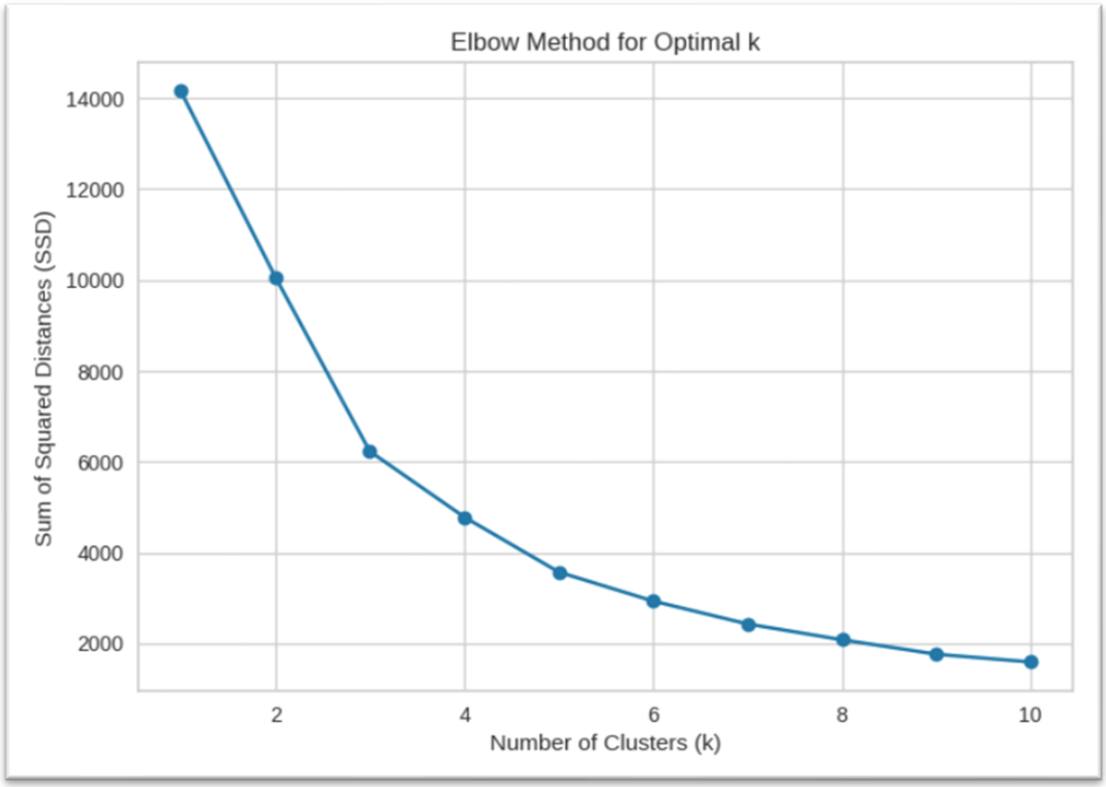



Figure 3 Elbow Method for decision of number of clusters (Source: authors own image)

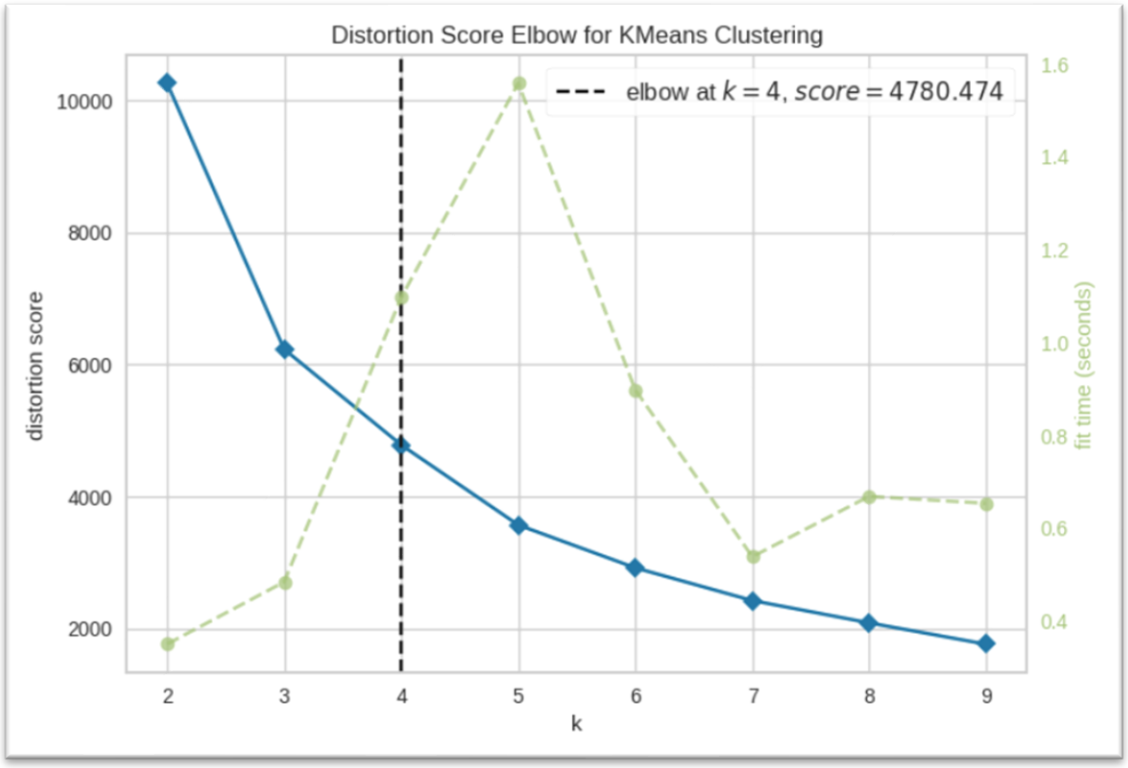



Figure 4 Distortion Score on Elbow for K Means (Source: authors own image)

# 4. Results and Discussion

For this analysis Google Colab, Scikit learn and K means++ were used. Relevant code for analysis was pushed to Git hub. The results of analysis are as below and metrics and KPIs are being explained in Chapter 4.2.

## 4.1 K-Means Performance Metrics and Comparison

$$\sum_{i=1}^{N}(x_i - C_k)^2 \quad (2)$$

N: number of samples within dataset

C: Center of cluster(centroid)

X: each instance in dataset

k: number of clusters

**Silhouette Coefficient**

Silhouette coefficient is another commonly used approach to decide how well K-means clusters are generated.

$$\mathrm{K} = \{\{(\mathrm{p1}, \mathrm{q1}),\ (\mathrm{p2}, \mathrm{q2}), \ldots (\mathrm{px}, \mathrm{qx})\}\} \quad (3)$$

K = number of clusters

(p, q): objects in a cluster

Calculating average distances based on Euclidian principal:

$$\sum_{i=1}^{n}(p1 - p_i) + (q1 - q\mathrm{i})^2$$

$$Si = (b_i - ai) \,/\, \max(a_i, b_i) \quad (4)$$

for $a_i > b_i$

$Si$ : Silhouette coefficient

$a_i$: average distance from the ith instance to all other instances in same cluster

$b_i$: average distance from the ith object to nearest another cluster

Ideally $a_i$ << $b_i$ condition should be obtained, we defined $a_i$ as average distance of the that data point to other data points in the same cluster. Logically, it should be a much smaller distance. If not:

$a_i$ >> $b_i$ it is highly possibly that somethings over there is misclassified.

## 4.2 Silhouette Analysis

After calculating silhouette coefficient (Si) formulated and explained in the previous chapter (4.1), we plot it to get a visual representation of each cluster number (k) and how well dataset is spread between k clusters. The more the data spread almost equally between and the less is the negative score, better the silhouette analysis is for that k cluster representation. The measure has a range of [-1, 1]. As mentioned above, the more positive on the score range is better the analysis interpretation. (Kumar, 2020)

## 4.3 Model Evaluation

After applying K means clustering on RFM Model for Customer segmentation, end results are as shown in Table 3.

Table 3 Customer Clusters after initial RFM Model

| **Clusters** | Number of customers | Recency (R) | Frequency (F) | Monetary (M) |
|---|---|---|---|---|
| C1 | 3427 | 42.2 | 126.055 | 12497.98 |
| C2 | 1274 | 245.36 | 46.33 | 3795.26 |
| C3 | 6 | 2.16 | 4689.33 | 247336.46 |
| C4 | 11 | 35.90 | 781.90 | 889918.30 |

**Cluster 1:**

- On average, Customers in this segment made their last purchase around 42 days ago.
- They made 126 purchases on average.
- Average spending pattern £12.497.
- This cluster represents 3427 customers with similar statistics.

**Cluster 2:**

- On average, Customers in this segment made their last purchase around 245 days ago. They took longer time compared to Cluster 1 (C1)
- They made 46 purchases on average.
- The average spending pattern is around £3800.
- This cluster represents 1274 customers with similar statistics.

**Cluster 3:**

- On average, Customers in this segment made their last purchase around 2 days ago. Here we see frequent-buyer characteristics for example.
- They made 4.6K purchases on average, which seems very high compared to other clusters.
- The average spending pattern is around £247K.
- This cluster represents 6customers with similar statistics. They are a very small group of people. High-income generators?

**Cluster 4:**

- On average, Customers in this segment made their last purchase around 35days ago.
- They made 780 purchases on average.
- Average spending pattern is around £890K.
- This cluster represents 11 customers with similar statistics. They are also a small group of people. We probably observe a niche group of customers here.

## 4.4 Challenges and Limitations

Two factors were decisive while doing my research. Time constraint and trying to pick an optimum model for customer segmentation was the hardest part. After formulating my research question, I quickly realized that it does not fit my dataset. Therefore, I had to reformulate my questions and method according to the data I had.

If I could start over, I would try to find a dataset both to analyze more features of e-commerce consumers such as gender, location, screen times during their purchase and customer segmentation. For RFM analysis I needed unique identifiers for transaction and customer as well timestamps of these transactions. To keep it simple and focused I chose to move forward with only segmentation and RFM analysis for this paper.

## 4.5 Model Improvements

Model needs improvement for sure. I had a relatively big negative Silhouette score when k is bigger than 2, which was not ideal. But it could be me as human error, something I skipped probably, I will do analysis again. On the other hand, Elbow score was normal and pointing around k values between 3 and 4.

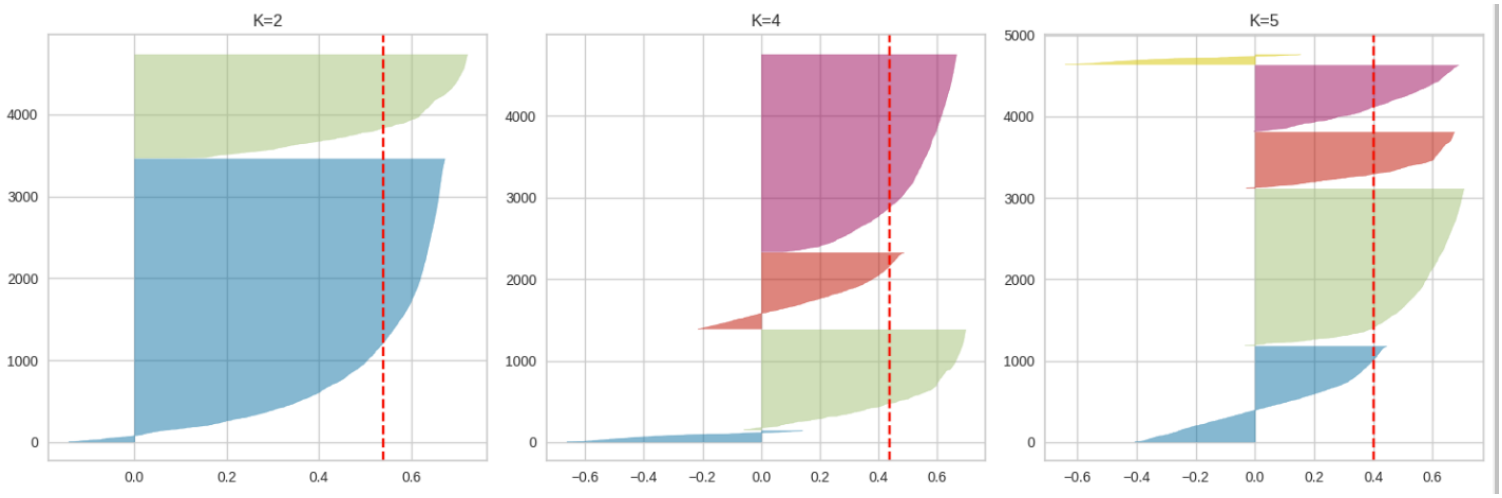



Figure 5 Silhouette analysis comparing different k values (Source: authors image)

# 5. Conclusions

## 5.1 Key Findings

Findings of the study are being discussed as referenced to different Cluster groups in Chapter 4.2.

**Customer Retention and Weighting Methods**

I wanted to analyze other metrics of RFM analysis such as Customer loyalty and customer retention to find answers to the third research question (*How can we identify Customer retention and measure Customer loyalty?* ). Nevertheless, with the given data, I realized it will be challenging, since the analysis required some other data and further methodology I have never applied before. Therefore RQ3, was eliminated later phase of the research.

Based on previous literature in the field, customer loyalty value after RFM based analysis makes it possible to evaluate as ordinal values of loyalty attribute. There are examples of such implementations in the literature. The most common statistical method is weighting based on *Entropy Method* for Multi criteria decision making. This could be further analysis based on initial Clustering segments and RFM feature metrics. Wu, Shi, Yangs (2021) article on User-Value Identification based on improved RFM And K means++ analysis is one of those studies.

Subjective weighting for the Customer Loyalty value was the reason that I ended the study after creating customer segments. My initial aim was to also detect the customer segment which would have a relatively higher *loyalty score.* But after some thinking and discussions around the topic made me decide that Customer loyalty is rather and subjective topic itself. At this moment with this dataset and features of customers, only based on their spending habits and frequencies, it would not add more so I ended analysis there.

## 5.2 Future Improvements

The future trajectory for this study is to implement an entropy analysis as part of subjective weighting of the features. Only then it would be healthy to interpret the customer retention values.

Another point to consider is that a dataset with given features and also churns data, membership, gender, location of customer will be even better for multi-attribute analysis such as RFM model.

# 6. List of Figures and Tables

Table 1 RFM features, after preprocessing ..... 7
Table 2:Recency Frequency Monetary Model (RFM) features ..... 8
Table 3 Customer Clusters after initial RFM Model ..... 17

Figure 1 Agglomerating and Divisive Hierarchical Clustering stages (Source: QuantDare) ..... 4
Figure 2 DBSCAN Clustering images, clustered density basis in a closed sphere. (Source: Primo.ai)..... 5
Figure 3 Elbow Method for decision of number of clusters (Source: authors own image) ..... 14
Figure 4 Distortion Score on Elbow for K Means (Source: authors own image) ..... 14
Figure 5 Silhouette analysis comparing different k values (Source: authors image) ..... 20

# 7. Bibliography


Dawane Vinit, Waghodekar Prajakta, Pagare, Jayshri. (2021). RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention . *International Conference on Smart Data Intelligence (ICSMDI 2021).* SSRN.

Diego Lopez Yse. (2023). *Introduction to K-Means Clustering.* Pinecone. Retrieved from https://www.pinecone.io/learn/k-means-clustering/#K-means-clustering

Fabio M. Miranda, Niklas Köhnecke, Bernhard Y. Renard. (2023). HiClass: A Python Library for Local Hierachichal Classification Compatible with Scikit-Leran. *Journal of Machine Learning Research 24*, 1-17. Retrieved from http://jmlr.org/papers/v24/21-1518.html.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow Concepts, Tools and Tecqniques to Build Intelligent Systems.* O'Reilly Media,.

Gil, D. (den 23 September 2023). Mastering Customer Segmentation with LLM. *Towards Data Science*.

Hewa Majeed Zangana, Adnan M Abdulazeez. (2023). Developed Clustering Algorithms for Engineering Applications: A Review. *International Journal of Informatics, Information System and Computer Engineering, 4*(2), 147-169. doi:https://doi.org/10.34010/injiiscom.v4i2.11636

Jun Wu, Li Shi,Liping Yang, Xiaxia Niu ,2 Yuanyuan Li, Xiaodong Cui, Sang-Bing Tsai, Yunbo Zhang5. (den 3 May 2021). User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Anlaysis. *Hindawi, 2021*, 8. doi:https://doi.org/10.1155/2021/9982484

Leandro Mirenda, José Viterbo, Flaviá Bernadini. (2020). Towards the Use of Clustering Algorithms in Recommender Systems. *AI and Semantic Technologies for Intelligent Information Systems(SIGODIS).* AMCIS.

Mahboubeh Khajvand, K. Z. (2010). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study. *Elsevier*, 57-63.

Matthias Carnein, Heike Trauttman. (2019). Optimizing Data Stream Representation: An Extensive Survey on Stream Clustering. *Springer*, 277-197. doi:https://doi.org/10.1007/s12599-019-00576-5

P. Anitha, Malini M. Patil. (2022, May). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences, 34*, 1785-1792. doi:https://doi.org/10.1016/j.jksuci.2019.12.011

Patrick Brus. (2018). *Clustering: How to Find Hyperparameters using Inertia.* Toward Data Science. Retrieved from https://towardsdatascience.com/clustering-how-to-find-hyperparameters-using-inertia-b0343c6fe819

Practicus AI. (2018). *The 5 Clustering Algorithms Data Scientists Need to Know.* Towards Data Science.

Satyam Kumar. (2020). *Silhouette Method — Better than Elbow Method to find Optimal Clusters.* Towards Data Science .

Talaat, F.M.; Aljadani, A.; Alharthi, B.; Farsi, M.A.; Badawy, M.; Elhosseini, M. (2023). Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing. *Mathematics, 11*, 1-26. doi:https://doi.org/10.3390/math11183930

Thanh Ho, Suong Nguyen, Huong Nguyen, Ngoc Nguyen, Dac-Sang Man, Thao-Giang Le. (2023). An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. *Business Systems Research, 14*, 26-53.

Zengyuan Wu, Lingmin Jin, Jiali Zhao, Lizheng Jing, Liang Chen. (2022). Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm. (M. Versaci, Ed.) *Hindawi, Computational Intelligence and Neuroscience*. doi:https://doi.org/10.1155/2022/9930613