1 Motivation

Shopping today encompasses more than just the act of making purchases and in the contemporary retail environment, gaining insight into customers extends beyond basic demographics. Customer segmentation is the practice that involves categorizing a company's customers into groups based on similarities. The aim is to tailor interactions with each group to maximize their value to the business. I believe a suitable customer segmentation model helps companies target specific customer groups, allowing for efficient allocation of marketing resources and maximizing selling opportunities.
The application domain is fundamental multivariate analysis and a simple model to categorize customer into specific groups. Therefore, I am going to utilize Unsupervised Machine Learning (K-means) which are theoretically taught in Multivariate Analysis and Machine Learning courses in combination with Recency Frequency Monetary (RFM) – a marketing analysis tool to fulfil the task. The structure of report follows: the second section presents problem formation including research questions, summary of dataset, exploratory insights from data; in the meantime, the third part mentions the method applied to make classification, particularly K-means and what insights we can draw from the clusters. The final section summarizes the result and improvement for better classification.

2 Problem Formation

2.1 Research questions

Concretely, this project aims at providing answers for the following questions:

- How can customer segmentation be achieved through analysis of their behavioral patterns, encompassing variables such as purchase frequency, recency of interactions, average order value?
- What are the key characteristics and behaviors of each customer segment identified through the segmentation analysis?

2.2 Dataset Information

The dataset is obtained via Online Retail. (2015). UCI Machine Learning Repository. This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. There are more than 500.000 data points with 8 features in the dataset.

Attribute information as below:

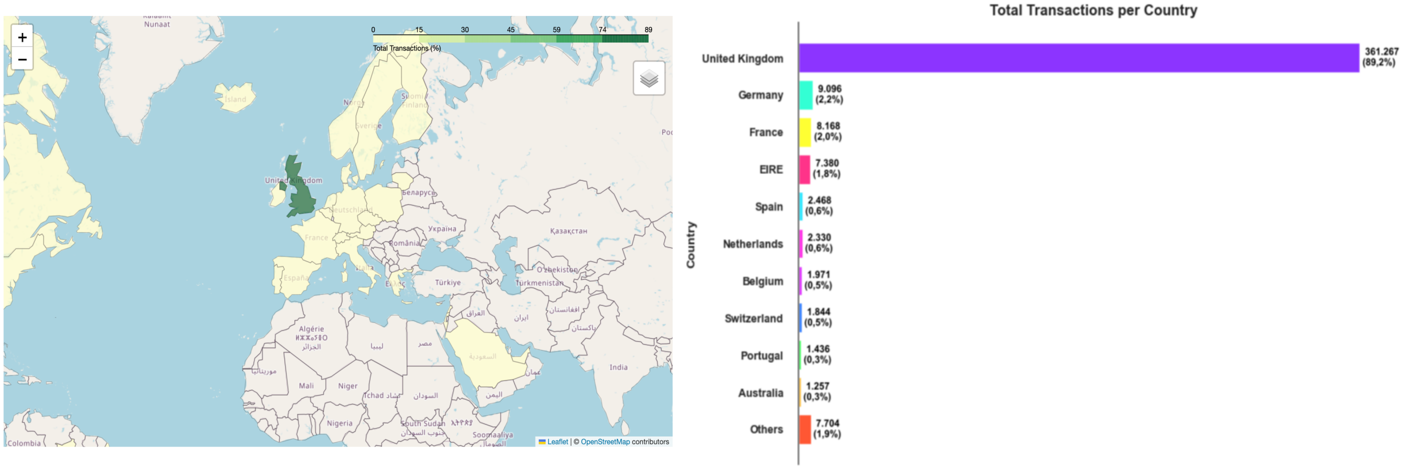
- **InvoiceNo**: Invoice number. Nominal. A 6-digit integral number is uniquely assigned to each transaction. If this code starts with the letter C, it indicates a cancellation.
- **StockCode**: Product (item) code. Nominal. A 5-digit integral number is uniquely assigned to each distinct product.
- **Description**: Product (item) name. Nominal.
- **Quantity**: The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate**: Invoice date and time. Numeric. The day and time when a transaction was generated.
- **UnitPrice**: Unit price. Numeric. Product price per unit.
- **CustomerID**: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer**.**
- **Country**: Country name. Nominal. The name of the country where a customer resides. [1]

2.3 Basic exploration

2.3.1 Most total transaction per country

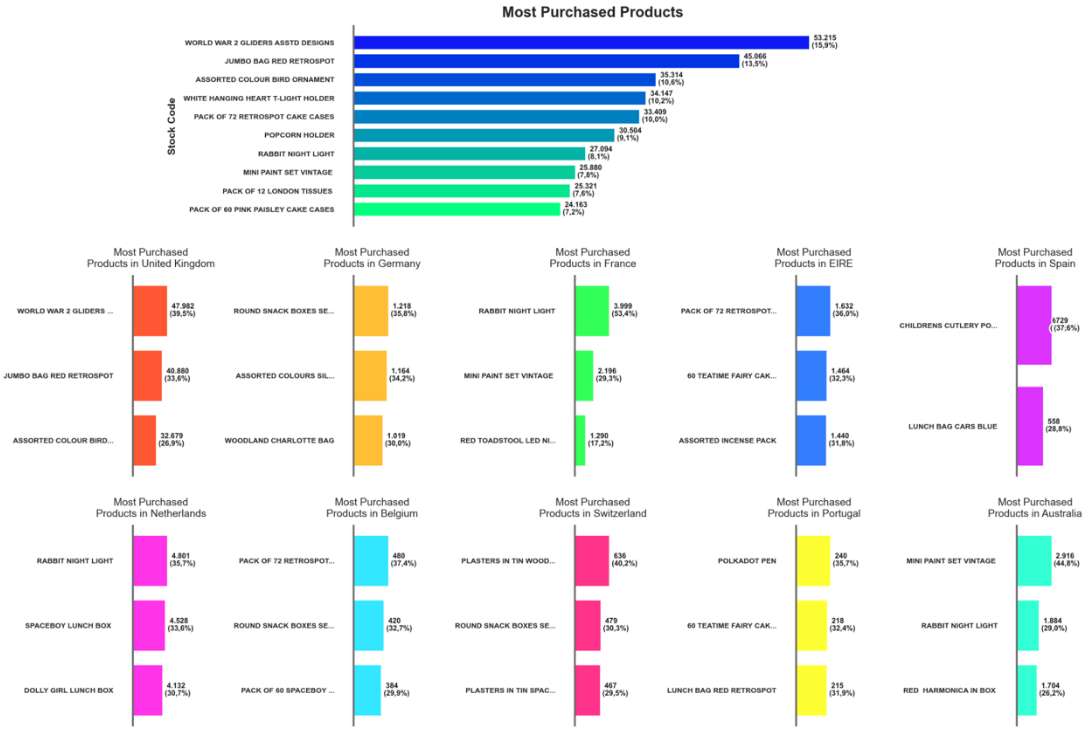
Transactions were made in 37 different countries in which there are 244 purchases where country is unspecified. Since they are a small number, I would dismiss them from the whole dataset. It is noticeable that the most sales make the United Kingdom with the largest number of orders (362.000 purchases) which account for approximately 89% of total transactions of the retailer. Germany, France and Eire come in second, third and fourth place with the proportion of 2.3%, 2.1% and 1.8% total orders respectively.

Figure 1. Total transaction per country



2.3.2 Most purchased products

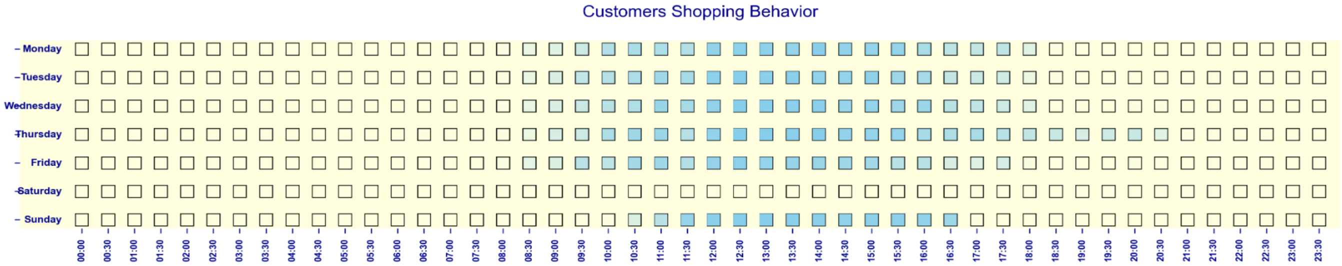
Figure 2. Most Purchased Products by Country



The top three favorite products in each country can vary. The most purchased products align with the following preferences: the top spot is held by World War 2 glider design, followed by the jumbo bag retrospot as the first runner-up, and the second runner-up is assorted ornament. These preferences are not surprising, given that England is the largest purchaser, indicating a strong correlation between the most purchased products and the preferences observed in the UK's purchasing patterns. Indeed, geographical conditions can play a significant role in shaping customer preferences to the types of products that are popular in a particular region.

2.3.3 Transaction period

Figure 3. Customer shopping behavior



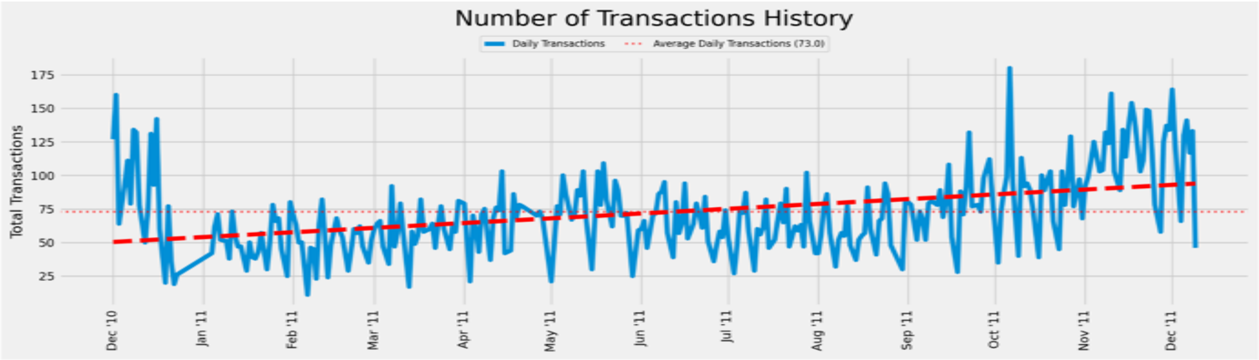


Figure 4. Number of Transaction over one year period

During the period from December 10, 2010 to December 9, 2011, the data suggests that people tend to make their purchasing activities towards the end of the year, likely in anticipation of holiday celebrations such as Christmas. Specifically, the highest number of transactions occurred during the month of November. In the meantime, consumer activity remains steady throughout the year. The average number of customer transactions during the period was 75 transactions. The trend line on the plot shows a rising trend, indicating that the number of transactions increased over the time. Besides, it seems that customer shopping behavior follows a distinct daily and weekly rhythm. Customers tend to make purchases during the day, with a peak around lunchtime. The absence of purchases on Saturdays indicates a trend of reduced customer activity on weekends. This could be due to businesses being closed, customers engaging in other weekend activities, or a decrease in demand for products and services during weekends. Simultaneously, transactions come to a halt between 9 pm and 8 am, suggesting a nighttime pause.

3 Data Analysis

3.1 Univariate Analysis

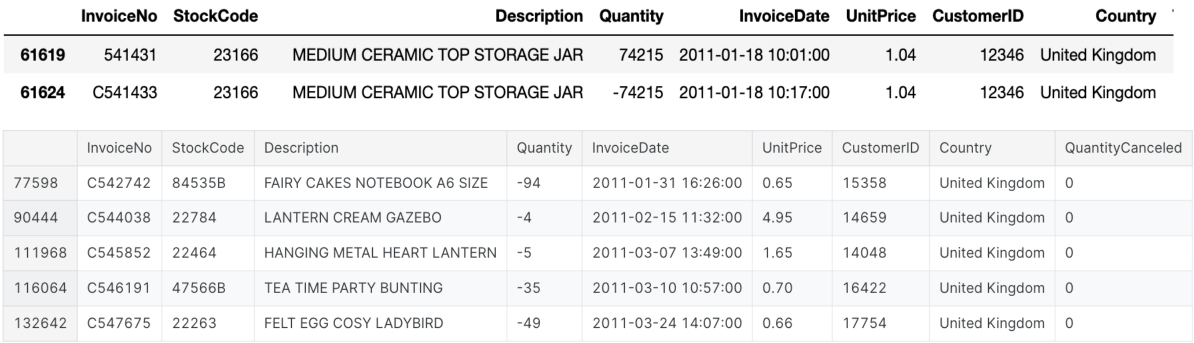
3.1.1 StockCode and Description

It appears that the "Stock Code" feature contains values such as "M", "D", "POST", and "CRUK", while the "Description" feature includes entries like "Discount", "Manual", "POSTAGE", and "CRUK Commission". These values likely correspond to fluctuations or changes in product stock and may not represent specific product names. As a result, it would be appropriate to remove those entries.

3.1.2 InvoiceNo

Entries with the prefix "C" for the InvoiceNo variable indicate canceled transactions, accounting for approximately 16.47% of the total number of transactions (3654 out of 22190). Upon closer examination, it is observed that cancellations do not necessarily correspond to orders that were previously made. Some canceled transactions without counterparts may stem from purchase orders executed before December 2010, the starting point of the database.

Figure 5. Quantity Canceled by prefix "C" InvoiceNo



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
61619	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	2011-01-18 10:01:00	1.04	12346	United Kingdom
61624	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346	United Kingdom

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	QuantityCanceled
77598	C542742	84535B	FAIRY CAKES NOTEBOOK A6 SIZE	-94	2011-01-31 16:26:00	0.65	15358	United Kingdom	0
90444	C544038	22784	LANTERN CREAM GAZEBO	-4	2011-02-15 11:32:00	4.95	14659	United Kingdom	0
111968	C545852	22464	HANGING METAL HEART LANTERN	-5	2011-03-07 13:49:00	1.65	14048	United Kingdom	0
116064	C546191	47566B	TEA TIME PARTY BUNTING	-35	2011-03-10 10:57:00	0.70	16422	United Kingdom	0
132642	C547675	22263	FELT EGG COSY LADYBIRD	-49	2011-03-24 14:07:00	0.66	17754	United Kingdom	0

3.1.3 Quantity

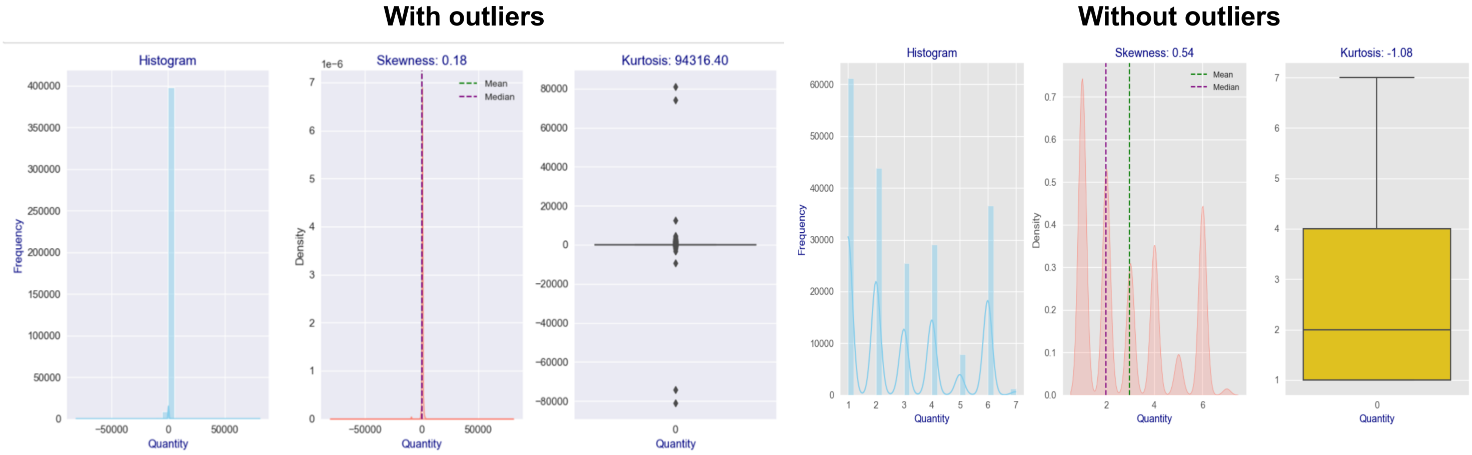


Figure 6. Quantity distribution with and without outliers

Approximately 9000 rows in the dataset exhibit negative values for the quantity variable. These negative quantities often correspond to Stock Code and Description values such as "D", "M", "discount", and "manual". While it is plausible that negative quantities could signify returned products, this assumption remains unverifiable. Notably, some irregular Description and Stock Code also feature negative quantity values.

After removing outliers, the resulting histogram shape is asymmetrical, indicating that Quantity does not follow a normal distribution. The skewness value of 0.54 suggests a positive skew, meaning that the data distribution has a longer tail to the right. In other words, there are more occurrences of high Quantity values compared to low ones. Additionally, the kurtosis value of -1.08 indicates a distribution that is flatter (platykurtic) than a normal distribution. This implies that the data tends to have fewer outliers and extreme values than a normally distributed dataset. The average Quantity is 2.9 and median is 2.00, which means if we arrange all the Quantity values in order, the middle value is 2.00. Mode of Quantity is 1.00, which is the most frequently occurring value in the Quantity dataset.

3.1.4 UnitPrice

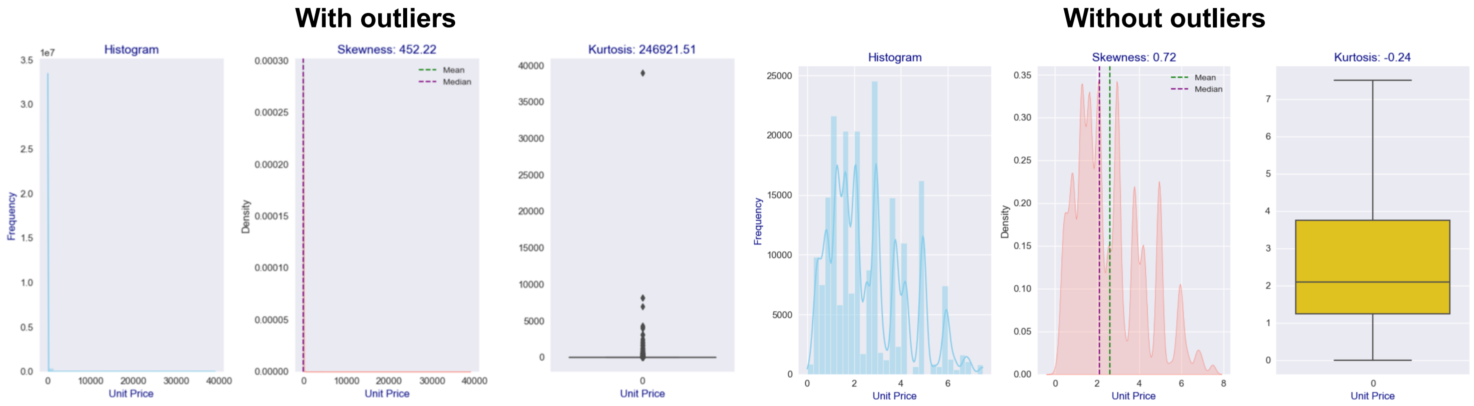


Figure 7. UnitPrice distribution with and without outliers

Similar to quantity, UnitPrice exhibits a wide range of values, with a very large maximum value and significant variability. The minimum value of 0 suggests that some items may be offered for free, which is somewhat unusual. These free items could be part of special offers or items provided for free to different customers, such as manuals. The average price per transaction is approximately 3,47. With a standard deviation of approximately 69.79, there is considerable variability or dispersion in the price across transactions. 25% of the transactions have a price of 1.25 or lower. This indicates that a quarter of the transactions involve relatively small price. In this case, the outliers in the unit price column are not something we want to dismiss, as they may correspond to very expensive ordered or returned products. However, it would be useful to examine the distribution of typical (outlier-free) price of products that were ordered.

The resulting histogram displays an asymmetrical shape, indicating a non-normal distribution. A skewness value of 0.72 suggests positive skewness, meaning the data distribution is skewed towards the right. This implies a longer tail on the right side of the graph, indicating a prevalence of high extreme values compared to low ones. Furthermore, a kurtosis value of -0.24 indicates a distribution with less taper (flatter) than a standard normal distribution. This means that the data distribution curve lacks a sharp peak typical of a normal distribution. With negative kurtosis, the distribution tends to be flatter in the center and has shorter tails than a normal distribution. This suggests that the data has fewer extreme values and is more concentrated around the mean value.

3.2 Bivariate Analysis

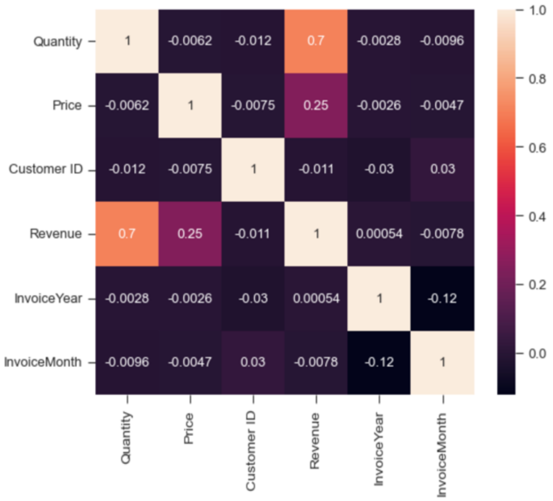


Figure 8. Correlation matrix

Correlations using Pearson's rank correlation coefficient. A new column Revenue which is the result of Quantity multiplying with UnitPrice is added.

There is a strong positive correlation (0.702) between Quantity and Revenue, indicating that higher quantities sold are associated with higher revenues. In the meantime, quantity has weak negative correlations with other variables, such as InvoiceDate (-0.007), Price (-0.006), Customer ID (-0.012), InvoiceYear (-0.003), and InvoiceMonth (-0.010), suggesting very little linear relationship.

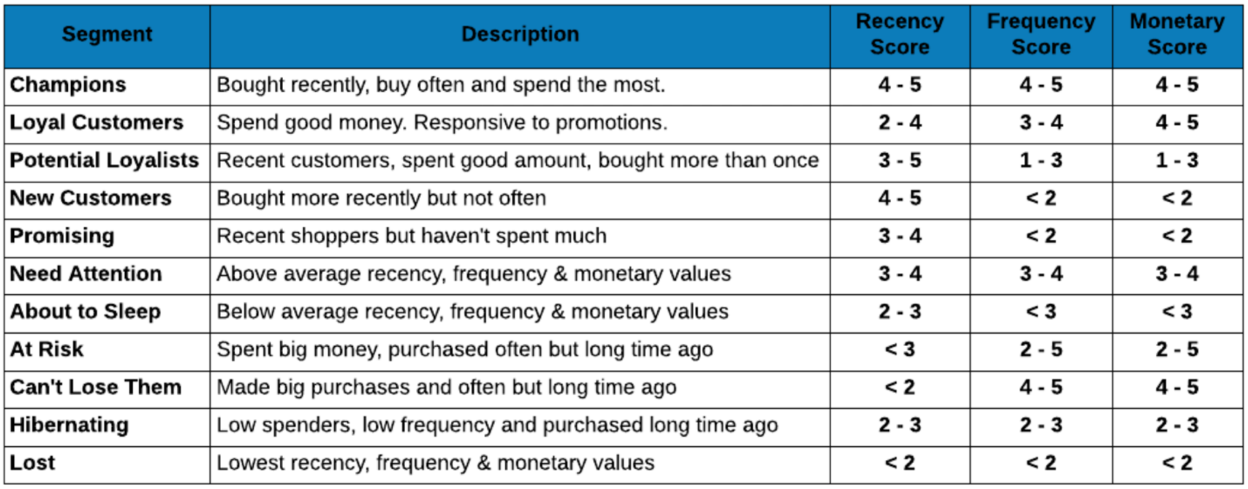
There is a weak positive correlation (0.250) between price and revenue, indicating that higher prices may lead to higher revenues.

4 Methodology

4.1 RFM segmentation

4.1.1 Theoretical Background

Recency, frequency, monetary value (RFM) is a model used in marketing analysis that segments a company's consumer base by their purchasing patterns or habits. In particular, it evaluates customers' *recency* (how long ago they made a purchase), *frequency* (how often they make purchases), and *monetary value* (how much money they spend). RFM analysis numerically ranks a customer in each of these three categories, generally on a scale of 1 to 5 (the higher the number, the better the result). The "best" customer would receive a top score in every category.[2]



Segment	Description	Recency Score	Frequency Score	Monetary Score
Champions	Bought recently, buy often and spend the most.	**4 - 5**	**4 - 5**	**4 - 5**
Loyal Customers	Spend good money. Responsive to promotions.	**2 - 4**	**3 - 4**	**4 - 5**
Potential Loyalists	Recent customers, spent good amount, bought more than once	**3 - 5**	**1 - 3**	**1 - 3**
New Customers	Bought more recently but not often	**4 - 5**	**< 2**	**< 2**
Promising	Recent shoppers but haven't spent much	**3 - 4**	**< 2**	**< 2**
Need Attention	Above average recency, frequency & monetary values	**3 - 4**	**3 - 4**	**3 - 4**
About to Sleep	Below average recency, frequency & monetary values	**2 - 3**	**< 3**	**< 3**
At Risk	Spent big money, purchased often but long time ago	**< 3**	**2 - 5**	**2 - 5**
Can't Lose Them	Made big purchases and often but long time ago	**< 2**	**4 - 5**	**4 - 5**
Hibernating	Low spenders, low frequency and purchased long time ago	**2 - 3**	**2 - 3**	**2 - 3**
Lost	Lowest recency, frequency & monetary values	**< 2**	**< 2**	**< 2**

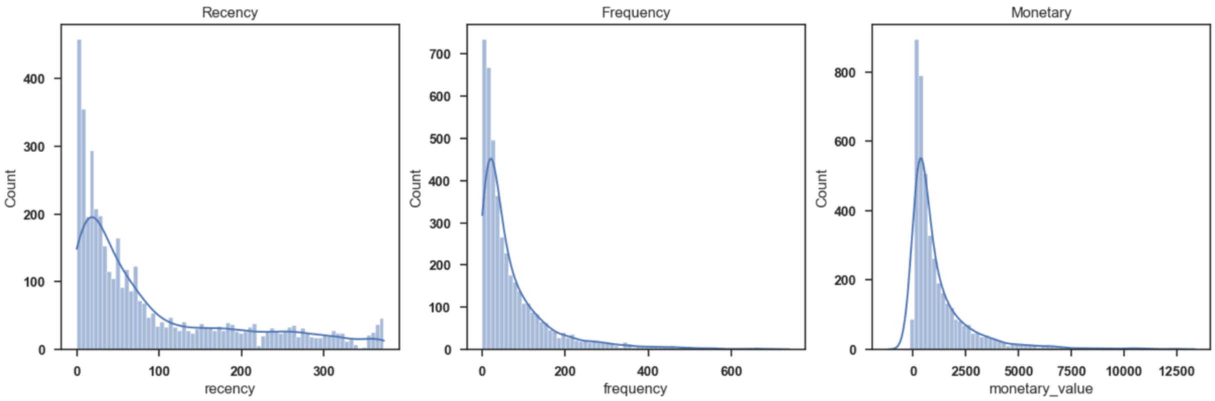
Table 1. RFM customer category

- The most important factor in identifying customers who are likely to respond to a new offer is recency. Customers who purchased more recently are more likely to purchase again than are customers who purchased further in the past.
- The second most important factor is frequency. Customers who have made more purchases in the past are more likely to respond than are those who have made fewer purchases.
- The third most important factor is total amount spent, which is referred to as monetary. Customers who have spent more (in total for all purchases) in the past are more likely to respond than those who have spent less.[2]

4.1.2 RFM analysis

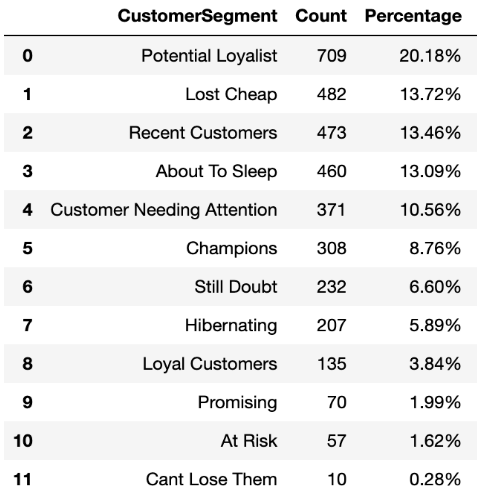
The recency of customer transactions typically follows a unimodal distribution and tends to be right-skewed. In the meantime, the frequency of customer transactions is commonly characterized by a unimodal distribution with a right-skewed shape. Customer spending patterns often exhibit a unimodal distribution with a tendency towards right-skewness. As mentioned on IBM Knowledge Center: ***It is not unusual for these histograms to indicate somewhat skewed distributions*** *rather than a normal or symmetrical distribution.* However, log transformation is applied to avoid bias.

Figure 9. Recency, Frequency, Monetary distribution

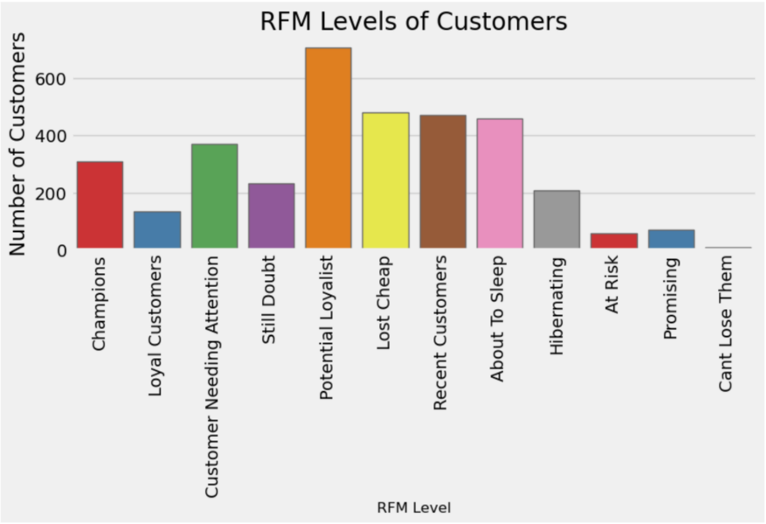


4.1.3 RFM Result

Figure 10. RFM customer categories



	CustomerSegment	Count	Percentage
0	Potential Loyalist	709	20.18%
1	Lost Cheap	482	13.72%
2	Recent Customers	473	13.46%
3	About To Sleep	460	13.09%
4	Customer Needing Attention	371	10.56%
5	Champions	308	8.76%
6	Still Doubt	232	6.60%
7	Hibernating	207	5.89%
8	Loyal Customers	135	3.84%
9	Promising	70	1.99%
10	At Risk	57	1.62%
11	Cant Lose Them	10	0.28%



Lost Cheap Customers: Approximately 14% (482) of customers fall into this category. These customers made few transactions but spent a significant amount of money. Re-engaging them could potentially yield substantial benefits. **Can't Lose Them**: A small percentage, around 0.28% (10) of customers, have been lost despite making numerous transactions and spending generously. Efforts should be made to win them back.

Recent Customers: This segment represents around 13% (473) of new customers. They are recent additions to the customer base and warrant attention to ensure they become loyal in the future.

Loyal Customers: Approximately 3.84% (135) of customers are identified as loyal customers. These customers consistently engage with the company and are valuable assets.

Champions: About 8.8% (308) of customers are considered champions. They make frequent transactions and spend generously, contributing significantly to the company's revenue.

Customers Needing Attention, About to Sleep, At Risk: This segment constitutes around 26% (926) of customers. They require special attention to prevent further attrition and retain their loyalty.

Promising, Still Doubtful: Roughly 9% (336) of customers fall into this category. Despite engaging frequently, they remain hesitant. Offering attractive incentives could help convert them into loyal customers**.**

Potential Loyalists: Approximately 21.4% (825) of customers show potential to become loyal customers. Rewarding their engagement could facilitate their transition into loyal patrons, aligning with the company's objectives.

4.2 K-means Clustering

K-means is selected in combination with RFM due to its simplicity and effectiveness in addressing clustering problems using unsupervised learning techniques. It particularly excels with large datasets. However, one challenge associated with K-means clustering is the necessity to predefine the number of clusters before running the algorithm. To determine the optimal number of clusters, the Elbow method is commonly employed. The results suggest that the optimal number of clusters is 2, as it corresponds to the point where the curve starts to bend or show a significant change in slope. Cluster visualizations are then presented based on this optimal number of clusters to provide insights into the underlying patterns and structures within the data.

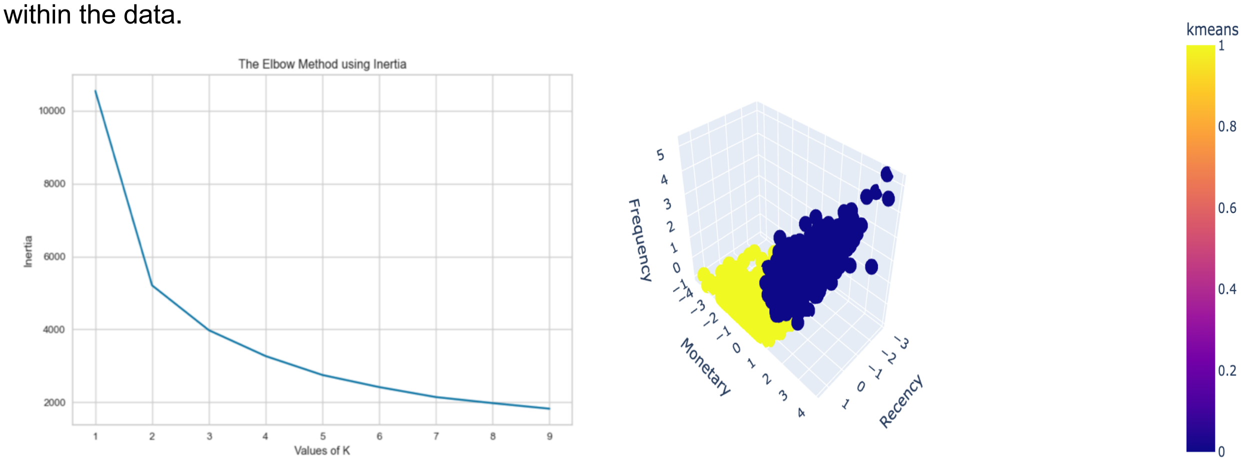
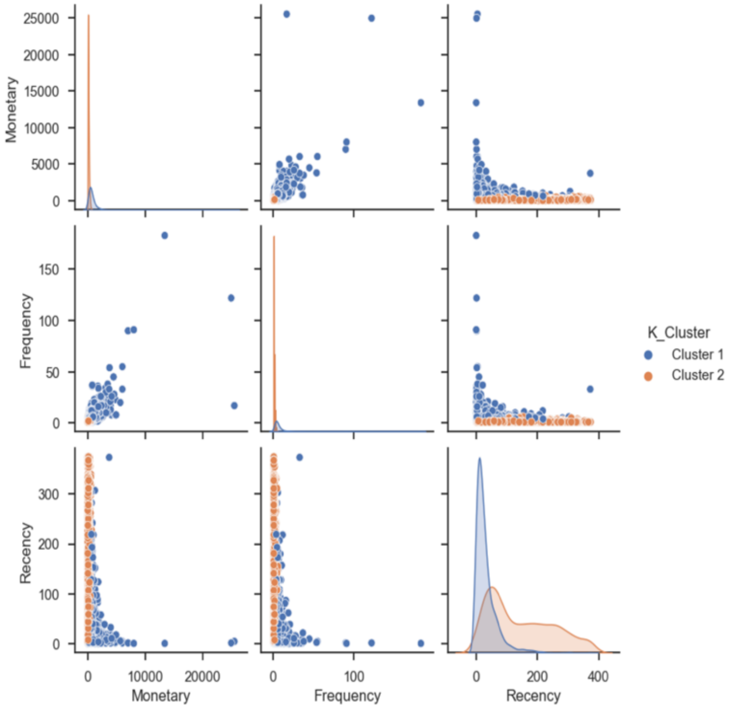
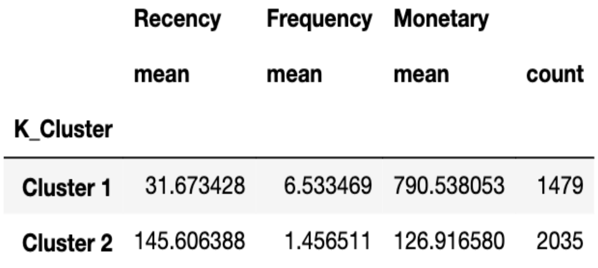


Figure 11. Elbow and cluster visualization

4.3 Cluster Interpretation

Cluster 1 made their most recent purchase around 31.67 days ago on average. The mean frequency value is approximately 6.53 and monetary value is approximately 790.54. This group represents a segment of customers who have made more recent purchases, higher frequency of purchases, and spend more money on average compared to Cluster 2. Cluster 1 is best described by some representatives such as Loyal Customers, Champions, or Potential Loyalists. They exhibit high purchasing frequency and significant total spending, making them the most valuable segment. Prioritizing marketing and retention efforts towards them is essential for maximizing returns and fostering long-term loyalty. Cluster 2, on the other hand, represents a segment of customers who have made less recent purchases with lower frequency of purchases, and spend less money on average. Conversely, the mean recency value is significantly higher at approximately 145.61 days which suggests that group 2 made their most recent purchase much later compared to Cluster 1. The mean frequency value is notably lower at approximately 1.46 for group 2 who made fewer purchases, the average monetary value is substantially lower at approximately 126.92. These customers may require targeted marketing strategies to re-engage them or encourage them to increase their spending. This group has a lower but reasonable value of monetary as the group includes newly registered consumers starting shopping with the retailer very recently. It seems to have represented ordinary consumers and therefore has a certain level of uncertainty in terms of profitability. In the long-term view, some of the consumers might be potentially very highly profitable or unprofitable at all.





	Recency	Frequency	Monetary	
	mean	mean	mean	count
K_Cluster				
Cluster 1	31.673428	6.533469	790.538053	1479
Cluster 2	145.606388	1.456511	126.916580	2035

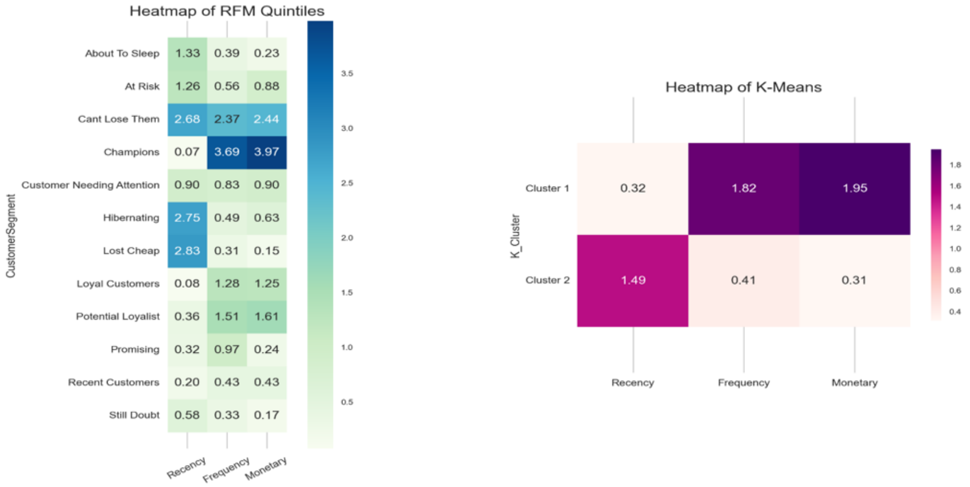
Figure 12. Recency, Frequency, Monetary by cluster

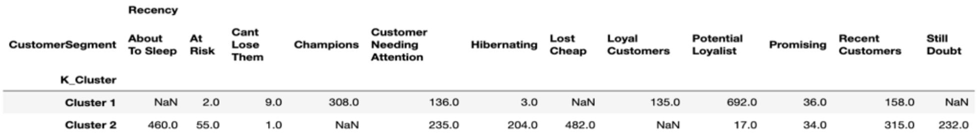
4.4 Comparison between K-means and RFM segmentation

In this particular practice, the goal is to understand customer behavior and to identify distinct clusters within the data and uncover patterns that may not be immediately apparent. K-means clustering doesn't require predefined labels; it automatically groups similar data points together based on feature similarity. However, the results are varied according to the initial choice of centroids. In the meantime, RFM segments are easily interpretable as they are based on simple and intuitive concepts like recency, frequency, and monetary value.

A combination of both RFM analysis and K-means clustering may yield the most comprehensive insights, leveraging the strengths of each approach to provide a deeper understanding of customer behavior and preferences. Institutively, customer segmentation from RFM result can be merged into 2 clusters:

Cluster 1: About To Sleep, At Risk, Hibernating, Lost Cheap, Recent Customers, Still Doubt
Cluster 2: Champions, Loyal Customers, Potential Loyalist





	Recency											
CustomerSegment	About To Sleep	At Risk	Cant Lose Them	Champions	Customer Needing Attention	Hibernating	Lost Cheap	Loyal Customers	Potential Loyalist	Promising	Recent Customers	Still Doubt
K_Cluster												
Cluster 1	NaN	2.0	9.0	308.0	136.0	3.0	NaN	135.0	692.0	36.0	158.0	NaN
Cluster 2	460.0	55.0	1.0	NaN	235.0	204.0	482.0	NaN	17.0	34.0	315.0	232.0

Figure 13. Heatmap of RFM Quintiles and K-means

5. Critical Evaluation

Customer segmentation poses a significant challenge, requiring thorough analysis of customer data to categorize them effectively. Utilizing a combination of RFM analysis and K-means clustering, customers are classified into two key groups: "Need Attention" and "Loyal Customers." The former group is at risk of leaving and requires targeted efforts to convert them into loyal patrons. Meanwhile, the latter, characterized by frequent and recent transactions, represent valuable assets for the retailer.

K-means clustering and RFM (Recency, Frequency, Monetary) analysis are both powerful tools utilized in this project for understanding customer segmentation based on their purchasing behavior. Rather than comparing which one is better, leveraging the strengths of both techniques offers a more comprehensive approach to gaining insights into customer behavior.

K-means clustering excels in identifying distinct clusters within the data based on similarity in purchasing behavior. It automatically groups customers with similar characteristics, providing insights into different customer segments. However, it requires predefining the number of clusters, which can be subjective and may not always result in the most optimal solution.

On the other hand, RFM analysis focuses on evaluating customers based on their recency, frequency, and monetary value of purchases. It offers intuitive insights into customer behavior and categorizes customers into segments based on their purchasing patterns. RFM segmentation is easily interpretable and provides actionable information for targeted marketing strategies and customer retention efforts.

The research encounters several limitations, predominantly due to outliers present in the collected data. Particularly, certain transactions within the dataset display abnormal monetary values, potentially undermining the effectiveness of both the RFM model and clustering algorithm. Moreover, the fact that the dataset is dominated solely by sales from the United Kingdom might not adequately reflect other countries' dynamics, given variations in lifestyle and consumer preferences. Additionally, the lack of key demographic variables, like customer age, could have impacted the clustering methodology.

References

1. **List of Figures**

2. List of tables

3. References

[1] https://archive.ics.uci.edu/dataset/352/online+retail

[2] https://blog.rsquaredacademy.com/customer-segmentation-using-rfm-analysis/

[3]Code reference:
https://github.com/adzict/online_retail_customer_segmentation/blob/main/Online%20Retail%20Customer%20Segmentation.ipynb