



**College: Engineering and Information Technology**  
**Department: Information Technology**  
**Program: Data Analytics**

**Data Mining – DAT401**

**Customer Segmentation Using RFM**

**Prepared by:**

<b>Student IDs</b>	<b>Names</b>
<b>202110844</b>	<b>Layan Ahmad</b>
<b>202110616</b>	<b>Haniyah Alzaben</b>
<b>202110613</b>	<b>Leena AlSalhi</b>

**Supervised by:**  
**Dr. Ghazi AlNaymat**

**Data of Submission: 2 – 5 – 2024**

**Academic Year 2023- 2024 – Spring semester**

## Table of Contents

<b>Abstract</b> .....	4
<b>Introduction</b> .....	5
<b>Literature Review</b> .....	6
<b>Methodology</b> .....	7
a. Data Sources and Collection Methods .....	7
b. Data Preprocessing (Data Cleaning) .....	7
b. Exploratory Data Analysis (EDA) .....	7
d. Data Preprocessing for Clustering .....	8
e. Clustering .....	8
f. Frequent Pattern Mining and Association Rule Mining .....	8
<b>Data Cleaning and Transformation</b> .....	8
a. Handling Duplicates .....	9
b. Handling Missing Values: .....	9
c. Removing Cancelled Transactions and Negative Quantities .....	9
e. Cleaning Description Column: .....	10
f. Treating Zero Unit Prices .....	11
g. Cleaning Country Anomalies .....	11
h. Date-Time .....	11
i. Detecting Outliers .....	11
<b>Exploratory Data Analysis (EDA)</b> .....	12
a. EDA with Temporal Variables .....	13
b. EDA with Geographic & Description .....	14
<b>Feature Engineering</b> .....	15
a. RFM Analysis .....	15
b. Product Diversity .....	16
c. Additional Behavioral Features .....	16
d. Geographic features .....	17
e. Seasonality and Trends .....	17
<b>Dimensionality Reduction</b> .....	21
Why did we apply Dimensionality Reduction? .....	21
Principle Component Analysis (PCA) method .....	21
<b>Clustering</b> .....	23
Clustering Model - K-means .....	23
Clustering Model - K-medoids .....	24
Clustering Model - Hierarchical .....	25
<b>Cluster Analysis and Profiling</b> .....	26
Radar Chart (results and interpretation) .....	26
<b>Frequent Pattern Mining and Association Rule Mining</b> .....	26

<b>Cluster 0 - Local Shoppers (Results and Interpretation)</b> .....	26
<b>Association rules of cluster 0</b> .....	27
<b>Cluster 1 - Regular Shoppers (Results and Interpretation)</b> .....	28
<b>Association Rules of Cluster 1</b> .....	28
<b>Cluster 2 - Frequent Buyers (Results and Interpretation)</b> .....	28
<b>Association Rules of Cluster 2</b> .....	29
<i>Conclusion</i> .....	30
<i>References</i> .....	31
<i>Appendices</i> .....	32

## ***Abstract***

*When it comes to businesses, understanding the behavior of the customer is a crucial step in order to thrive in competitive markets. Our project focuses on customer segmentation as a way to gain insights about the customers that the business deals with, with the purpose of understanding their behaviors, the purchases patterns and the engagement of the customers with the products that a specific brand/business sell. We aim to enhance the marketing strategies of the business and increase the growth of sales. We will be applying the K-Means clustering algorithm to segment our customers. Our goal is to extract intricate customer segments characterized by their recency, frequency, and monetary value. With these insights, we endeavor to enhance the marketing strategies and product recommendations to resonate with the needs and preferences of each segment of the customers.*

**Keywords—***Clustering, RFM, Market Basket Analysis.*

## ***Introduction***

When it comes to e-commerce, understanding the behavior of customers is the key to success, as by understanding their needs the business will be able to develop successful marketing strategies and will thrive in the competitive market. The development of online retail has marked the beginning of remarkable accessibility to data, which helped the business to gain insights about their business, customers, and products. Businesses can dig into this vast amount of data with the help of different data mining methods in order to discover patterns, trends and preferences that influence the behavior of customers.

The purpose of this project is to leverage different data mining methods to gain informative insights from transactional data.

Before data mining begins, there are several steps that are important when it comes to data analytics to ensure accurate results, our approach begins with these steps: Data Cleaning & Transformation, Data Preprocessing, EDA, RFM Analysis, Feature Engineering, PCA, Cluster Analysis, Frequent Pattern Mining as well as Association Rule Mining and lastly, the interpretation of the results that will aid in the decision making process and planning of our strategies.

## ***Literature Review***

In the referenced article, that was written by A. Joy Christy et al. the main discussed topic is the use of the RFM analysis which is a technique used to analyze customer behavior as we mentioned before. They analyzed transactional data to better understand the customer behavior. Other key proposed techniques were the use of K-Means and Fuzzy C-means methods. They have compared the results of both algorithms and provided meaningful insights that can help the marketing team to better target their customers.

Moreover, they evaluated the time taken, iterations and the silhouette score of the different clustering techniques they used. They highlighted how effective was their proposed RM K-means technique in reducing the computational time needed while focusing on providing helpful insights.

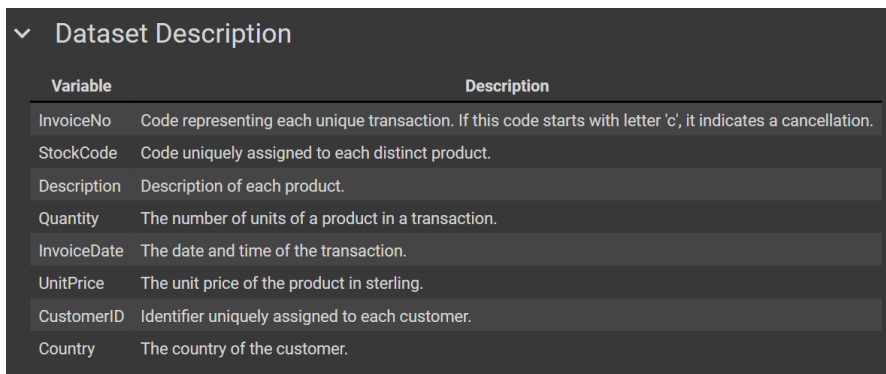
Overall, the article showed how the RFM based analysis was effective in offering meaningful insights about the customer segments.

# Methodology

Throughout this project, we have applied an organized approach which entailed of data cleaning, Exploratory Data Analysis, RFM analysis, dimensionality reduction using Principal Component Analysis, tested 3 clustering algorithms which we will be discussing them later in this report, and lastly, we have applied frequent pattern mining and association rule mining to each cluster independently to understand the patterns of each segmentation separately.

## a. Data Sources and Collection Methods

Our dataset was obtained from Kaggle, it captures a snapshot of different transactions from retailers in the UK from the years 2010 and 2011. The dataset contains 541,909 rows and 8 essential columns, which are: **InvoiceNo** which contains the number of invoices of every transaction where an invoice number signifies different products that are bought in one transaction. A customer's single transaction can be represented in different rows where each row represents a single item.



The image shows a screenshot of a 'Dataset Description' table. The table has two columns: 'Variable' and 'Description'. It lists eight variables: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country, each with a brief description of what the variable represents.

Variable	Description
InvoiceNo	Code representing each unique transaction. If this code starts with letter 'c', it indicates a cancellation.
StockCode	Code uniquely assigned to each distinct product.
Description	Description of each product.
Quantity	The number of units of a product in a transaction.
InvoiceDate	The date and time of the transaction.
UnitPrice	The unit price of the product in sterling.
CustomerID	Identifier uniquely assigned to each customer.
Country	The country of the customer.

## b. Data Preprocessing (Data Cleaning)

Before starting the analysis, we must ensure our data is in the appropriate structure, ready for analysis, there shouldn't be any missing values, no duplicate values or any other issue that affects the accuracy of the analysis. In this step, we handled all these problems, we have also handled canceled transactions, negative quantities, anomalies, and outlier detection. By addressing these problems, we have the data prepared for further analysis.

## b. Exploratory Data Analysis (EDA)

We have also performed EDA to observe the distribution, relationships between columns and patterns with the purpose of gaining insights into our data. We have visualized summary statistics, seeing which country is the most frequently using the word cloud, and many other plots that have helped us gain insights and context for the following steps.

#### **d. Data Preprocessing for Clustering**

Before applying the clustering methods, we employed a series of preprocessing steps to ensure the quality and accuracy of the clustering process. The preprocessing step included feature engineering, where we have computed the Recency, Frequency and Monetary for each customer based on their transactional behavior, we have also incorporated other relevant features. During this step, we have also applied outlier detection as outliers can significantly impact the clustering process. Moreover, we have assessed the correlation between the features to identify the redundant features. After that, we applied standardization to scale the numerical features. And the last step in data preprocessing was implementing PCA as a method to address the curse of dimensionality.

#### **e. Clustering**

To segment our customers, we have utilized 3 methods to their RFM values and additional features that we added, with the purpose of exploring different techniques to segmentation and evaluate their performance in identifying distinct points.

#### **f. Frequent Pattern Mining and Association Rule Mining**

The last step in this project is extract the frequent items and the meaningful rules to each cluster independently to understand the behavior of each segment and the relationships between items, we have also extracted the rules to understand the preferences of our consumers.

## ***Data Cleaning and Transformation***

We begin with data cleaning as a first step as it's an essential phase in any analytics process as it ensures the reliability, integrity, and quality of data. When our purpose is to extract meaningful insights from large datasets like the one that we're working on, the necessity of ensuring the quality of data cannot be overstated because clean data establishes the basis of accurate analysis which helps us as data analysts to derive accurate conclusions and make informed decisions.

We have employed different tasks, such as: we handled duplicates, missing values, cancelled transactions, anomalies and we have applied outlier detection. This is going to increase the trustworthiness of the findings.

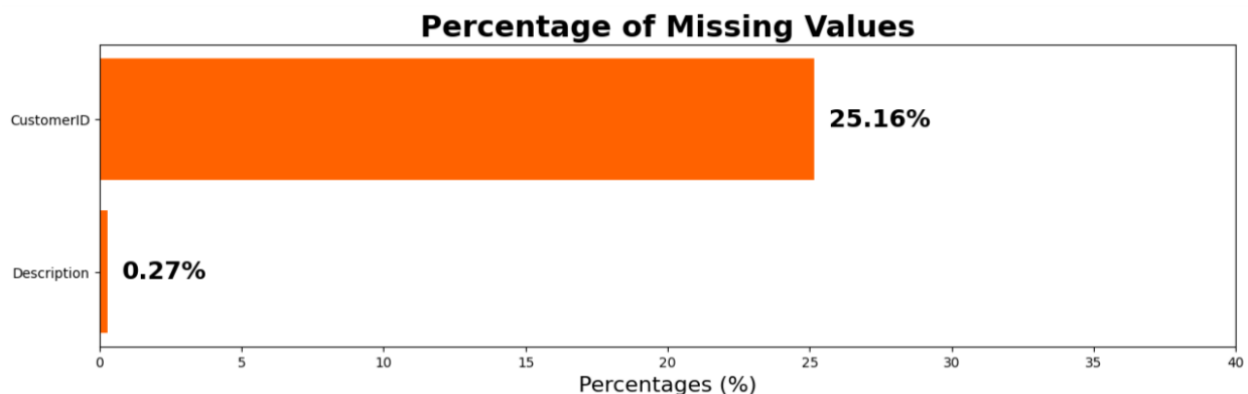


### a. Handling Duplicates

We started with handling duplicate rows, which is crucial to avoid redundancy. We had only 5268 duplicate rows that were dropped as they consume unnecessary storage space to maintain the integrity of data.

### b. Handling Missing Values:

After that, we checked if we had any missing values, and there were missing values in 2 columns which are the CustomerID and the Description columns. We decided to plot the missing values percentage in both columns, to see its proportion of the missing values to the whole dataset.



In the CustomerID column there were 135,037 missing values which make up 25.16% of the data. And for the Description column there were only 1454 missing values which make up 0.27% of the entire data. There are several reasons to why we have these missing values, these could be because of incomplete data entry, or maybe because there were anonymous transactions, and for the CustomerID column, it could be because of privacy considerations, perhaps some customers prefer to stay anonymous.

We extracted the rows that the Description as null and we dropped them since there isn't a way that we could use to fill them. However, for the CustomerID column, since a customer's transaction can be represented in multiple rows, we checked if the row with the missing customerID had a matching InvoiceID in a different row and filled it with that value. And if there isn't a matching InvoiceNo we generated new unique customer IDs.

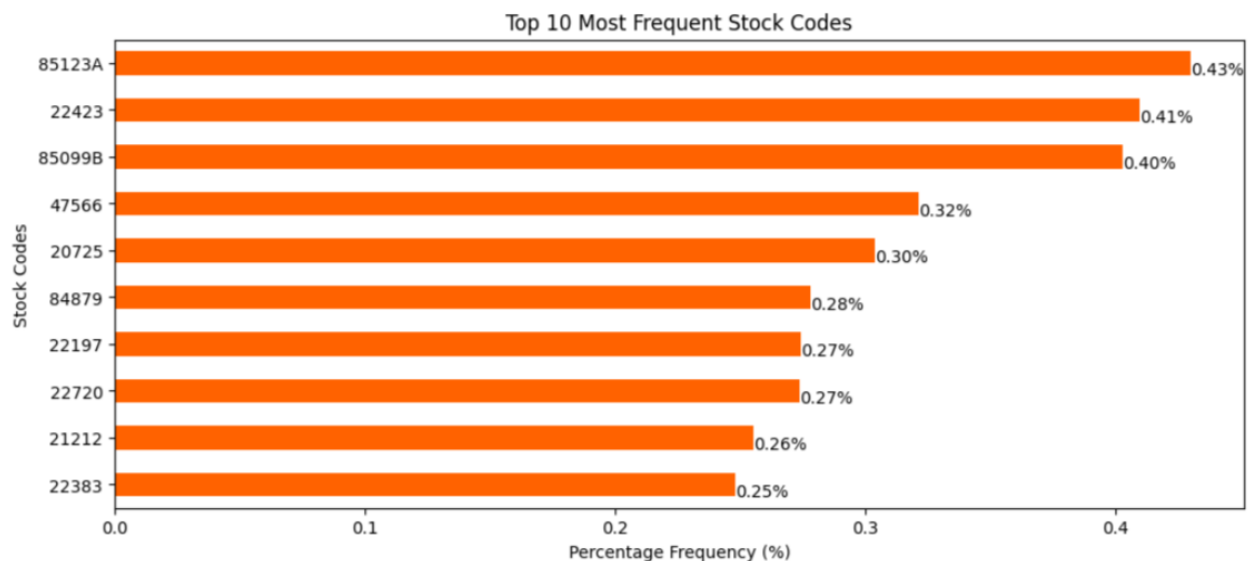
### c. Removing Cancelled Transactions and Negative Quantities

Consequently, we noticed that some invoice numbers have the letter 'C' in them which indicated that these represent transactions that were cancelled so we decided to drop these rows and understand their common

characteristics separately. The percentage of cancelled transactions in the entire dataset was only 1.73%. We have also noticed that there were quantities that were represented as negative numbers. We had different assumptions to why they're negative numbers, it's possible that they're items that were returned or that they are cancelled transactions or just an error. So, we have decided to drop them because they can introduce ambiguity and inconsistencies and having cancelled/returned transactions doesn't align with the objective of this project.

#### d. StockCode Anomalies

After that, we checked the number of unique stock codes in our dataset and we had 3925, which suggests that the online retail store has an extensive selection of products. We plotted the 10 most frequent stock codes to identify which items are the most popular, which gives us insights about the most frequent items that the customers buy.



We spotted different types of anomalies in this column that don't make sense, such as: 'BANK CHARGES', 'C2', 'DOT' and 'M'. These anomalies might represent services or non-product transactions (like postage fees) which is not the focus of our project which is clustering customers based on their product purchases, so we decided to remove them.

#### e. Cleaning Description Column:

When we started analyzing the 'Descriptions' column, we discovered that most of the descriptions were in the uppercase format. We noticed that all the descriptions are in uppercase, which might be a standardized format. We decided to extract the descriptions that contained lower case characters and a few of them contained lowercase format which we considered them to be anomalies, for example: 'Adjustment' and

‘wrongly coded 20713’, these are not considered products and are anomalies that need to be removed, and we fixed the products that were entered in the lower-case format in order to have a standardized format and to avoid any inconsistencies.

#### **f. Treating Zero Unit Prices**

We have also observed that there are 573 rows(items) with a price of ‘0’, this could occur because of data entry errors, or perhaps the customer that bought that item had an offer (like buy 1, get 1 free). Therefore, they got that item for free. We have decided to drop the rows with unit price of 0 as they could skew the data and give inaccurate results and we intend to preserve the data integrity.

#### **g. Cleaning Country Anomalies**

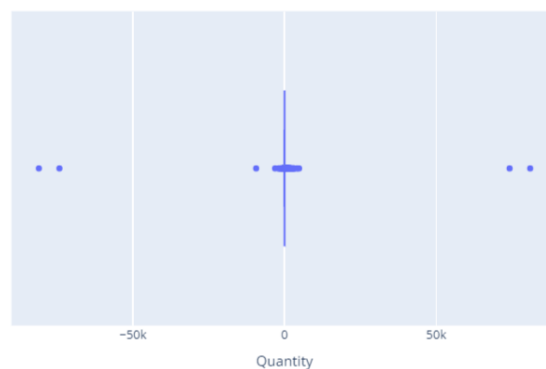
There were ‘unspecified’ countries in the country column so dropped them as well.

#### **h. Date-Time**

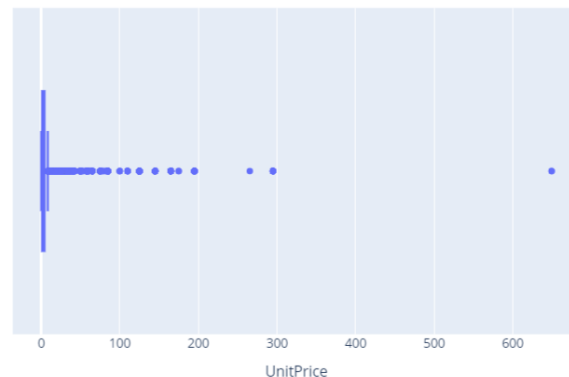
For the ‘InvoiceDate’ column, we have converted it to the datetime format which will help us in temporal analysis as using datetime objects can help in obtaining features like the year, month and day which help in analyzing the trends, patterns and seasonality in transactional data. Also, converting to the datetime format has helped in the visualization, such as time series and histograms.

#### **i. Detecting Outliers**

As a last step in the data cleaning process, we examined the dataset for outliers. It’s crucial to check for outliers because they can be errors from either in data entry or data collection, we addressed the outliers to ensure the quality of the data. First, we checked for the ‘Quantity’ column by visualizing it with a boxplot, and there were 2 extreme values which are 74.2K, 80.9K, -74.2K, and -80.9K.



We decided to drop these values to ensure the reliability of the data as they can significantly skew the data and the clustering algorithms which will lead to results that are not accurate. Moreover, our focus is on normal/typical transactions that will give us insights into the customer behavior. And for the 'UnitPrice' column, we have set a threshold of 250 and removed the values that were above the threshold and removed them as we considered them as extreme outliers.



## ***Exploratory Data Analysis (EDA)***

EDA is a significant phase for the first understanding of our data. We examined the structure of the data we are working on, as well as the distribution of our numerical data. By printing out the statistical summary, we can get to know more about the distribution and the 5-summary statistics: the minimum value, Q1, Q2, Q3 and the maximum value along with the count, the average and standard deviation. All the numerical features 'Quantity', 'UnitPrice', 'DayOfWeek', 'Month', 'DayOfMonth', 'Year' & 'Hour' have the same count which indicates the number of rows in the dataset.

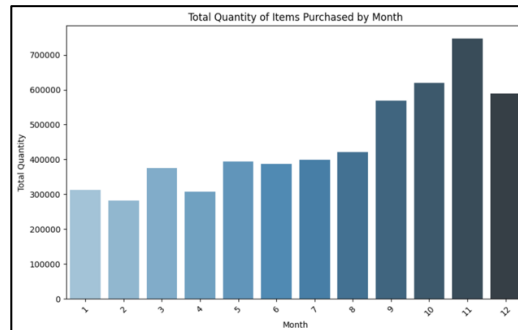
Digging deeper into these statistics, the mean of the 'Quantity' shows the average quantity of the items purchased per transaction which equals to 10.35. The minimum quantity of the items purchased per transaction is 1 item. The maximum quantity is 4,800 items. The median (Q2) is 4 items. The 25<sup>th</sup> percentile and the 75<sup>th</sup> percentile are 1, 12 respectively. Lastly, the standard deviation which explain the variability of the quantity across the transactions is approximately 37.87.

The mean, min, Q1, Q2, Q3, max and standard deviation of the 'UnitPrice' are 3.26, 0.04, 1.25, 2.08, 4.13, 195 and 4.05 respectively.

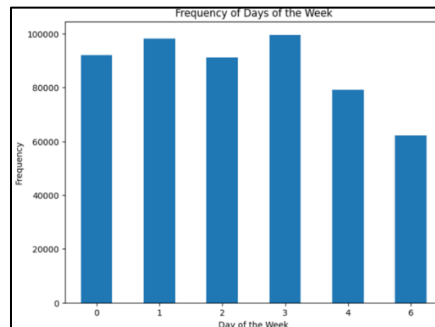
Finally, the mean, min, Q1, Q2, Q3, max and standard deviation of the 'Hour' are 13.07, 6, 11, 13, 15, 20 and 2.44 respectively.

### a. EDA with Temporal Variables

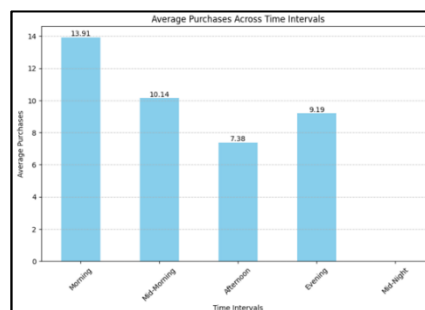
Considering the temporal variables in our EDA process is necessary as it help in identifying underlying trends and seasonality.



By just looking at the bar plot above, we can easily see that the quantities sold each month vary during the year. In February, which is the 2<sup>nd</sup> month of the year, this is usually when the least amount of items is purchased. However, in November we can see a peak indicating the largest amount of items purchased. We can also tell that during May and August the quantity sold is somehow stable and there isn't much of change.

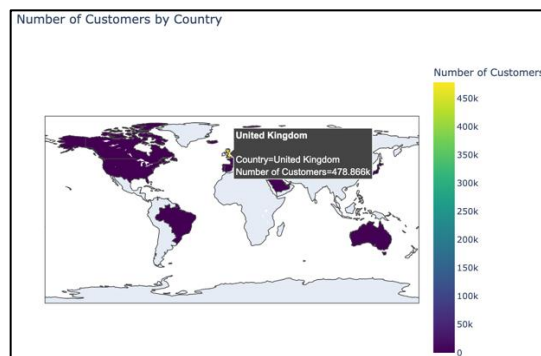


Digging deeper into the days of the week, we can't determine a pattern, but we can see that on Sunday, which is the last day of the week, the number of transactions decreased compared to the remaining days.



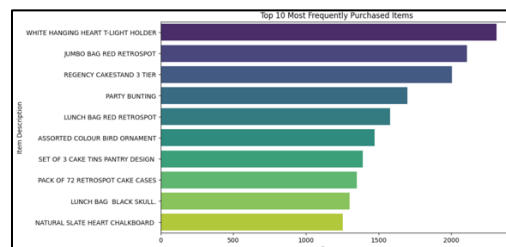
Moving further to these daytime intervals which are the following, Morning from 6 to 12 am, Mid-morning from 12 to 3 pm, Afternoon starts at 3 till 6 pm, Evening starts at 6 till 11pm and Midnight starts at 12pm until 6am. Keeping in mind that these intervals aren't equally separated, we can see the average purchases across the time intervals we mentioned earlier in the bar plot above. Most of the purchases happen during the morning time with an average of 13.91. The average purchase in mid-morning is still considered high compared to the morning time. However, there are no purchases happening during midnight time.

## b. EDA with Geographic & Description



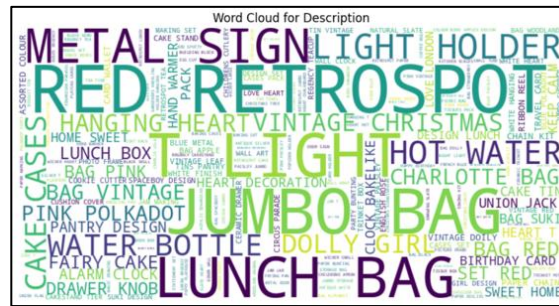
The analysis obtained from this graph makes it simple to understand that, out of the 38 countries available in the data the United Kingdom has a significant disparity in the number of customers as they are approximately 478.86k customers which is the majority of our customers. Meaning that the customers in the UK are the most important to consider later on in our pattern discovery journey.

Moreover, this significant difference may affect the marketing campaigns, the different preferences among all other customers and it's important to consider for future profitability and business growth.



Identifying the top 10 most frequently purchased items provides valuable insights regarding the customers preferences and the popularity of the products sold. The common item that's purchased the most is 'WHITE

HANGING HEART T-LIGHT HOLDER' and the 10<sup>th</sup> most frequent purchased item is 'NATURAL SLATE HEART CHALKBOARD'.



The size of the word is the description word cloud determines how frequent is the item. ‘RED RETROSPOT’ is the largest size in the cloud meaning it’s the most frequent among others. Moreover, ‘LUNCH BAG’, ‘JUMBO BAG’, ‘T-LIGHT’, ‘METAL SIGN’ and ‘LIGHT HOLDER’ are also frequent since they are shown in a large font size.

## Feature Engineering

### a. RFM Analysis

RFM analysis is a type of analysis where you study the R which is the recency, F the frequency, and M the monetary. These features help in understanding the customer behavior as we used them along with other engineered features to segment our customers. Recency indicates when the customer's last purchase was. We started by creating a new dataset called 'transactions' that consists of only unique "CustomerID" and "StockCode" which is represented in a list. Each customer has a list of stockcodes that they might have purchased previously.

**“Recency”** indicates when the customer's last purchase was. The lower recency means that this customer's last purchase was recent.

**“Frequency”** signifies the number of times a consumer makes a purchase in a specific time frame. Customers with higher frequency are the most valuable to our business as we don’t want to lose them.

**“Total\_Products\_Purchased”** feature shows the total number of the products purchased by a customer. This feature also helps in understanding the customer behavior as it shows the overall number of purchases regardless of the frequency of a specific item.

**“Monetary”** indicates the amount of money spent by that customer. Customers with a higher monetary are the most valuable customers as they contribute to the revenue of the business.

**“Average\_Transaction\_Value”** indicates the average spending per transaction for a customer. The higher the average gets the more the customer is interested in shopping from this business.

### **b. Product Diversity**

**“Unique\_Products\_Purchased”** shows the total number of the unique products purchased by a customer. This feature helps in understanding the customer behavior as it shows the diversity of the purchased products. If the value is high, then this means that this customer might be interested to try purchasing new items.

### **c. Additional Behavioral Features**

We decided to add some additional features to help us understand some customers behavior and the way they shop, like what is the avg number of days between a customer purchase or how often they buy, and what is their favorite shopping day/hour. These features may help us in segmenting customers based on their habits.

**“Average\_Days\_Between\_Purchases”** it tells us about the avg or how often customer buy things. This feature is helpful because we can know when they might make their next purchase. Furthermore, this can also help the market in making decisions for sending them offers on that specific day or even reminders. We started calculating the difference between each purchase for each customer using the “InvoiceDay” column which is derived from an existing “InvoiceDate” feature. Then the meaning is calculated from the differences that were previously calculated.

**“Favorite\_Shopping\_Day”** This feature tells us about the most favorable day that a customer may tend to purchase more things on this day. This can also mean to try to send promotions and discounts to a specific customer. It is calculated by getting the frequency or the count of each Day with the combination of ‘CustomerID’ and ‘DayOfWeek’. Then the favourite shopping day is selected by finding the day with the highest frequency.

**“Favorite\_Shopping\_Hour”** It shows us the specific hour within a day that is favourable to a customer. We started by getting the frequency using the “Hour” feature. Then we selected the hour with the highest frequency.

I will interpret the first row of our transactional data frame after adding these 3 features as an example:

- **‘Average\_Days\_Between\_Purchases’:** 0.000000, this means that the customer purchases very frequently, or even on daily basis.



- **'Favorite\_Shopping\_Day':** 2, meaning Tuesday might be their Favorite shopping day since they purchase a lot during that day.
- **'Favorite\_Shopping\_Hour':** 3, the customer tends to purchase more around 3 PM.

#### **d. Geographic features**

We added a geographical feature that tells us more about the customer, like where they are from. This feature is also helpful because knowing where our customers come from can mean they might have different shopping habits. This benefits the market as well in tailoring their products based on where their customers are from and helps better in grouping customers. We started by getting the percentage of customers from unique countries that are listed in our dataset. The results were that 91.6% of our customers are purchasing from or are from the United Kingdom, 1.7% of the customers are from Germany which is very low compared to the UK, 1.5% from France and EIRE, and 0.5% from Spain. After knowing where the customers come from. We decided to add a feature called `Is_UK`, which is a binary feature that assigns the value 1 if the customer is from UK, and 0 if the customer isn't from UK. Firstly, we will group the data by "CustomerID" and "Country" and we will calculate the transaction count per country for each customer. Then the country with the maximum number of transactions for each customer will be considered the main country for the customer. Finally, the "`Is_UK`" column is created indicated whether a customer is from UK or not.

#### **e. Seasonality and Trends**

In this part, we are adding features that will help us understand the seasonal trends and to extract patterns in the customers' purchase behavior. These features can give us interesting insights in defining market strategies and further enhance market decisions.

**"Monthly\_Spending\_Mean"** This feature will tell us about the average amount a customer spends each month. A customer who has a high avg means that this customer buys a lot, therefore, this can also indicate that the customer is interested in premium products.

**"Monthly\_Spending\_Std"** this feature tells us about how much a customer's spending varies from month to another. For example, a higher value means their spending varies and is not stable, goes up and down.

**"Spending\_Trend"** It shows if the customer spendings is increasing, stable or decreasing over time.

I will interpret the first row of our transactional data frame after adding these 3 features as an example:

- **‘Monthly\_Spending\_Mean’**: 1428.990, this customer spends around an avg of \$1428.990 per month.
- **‘Monthly\_Spending\_Std’**: 0.000000, this is an indication that the customer spends the same exact amount every month without any change in the amount spent.
- **‘Spending\_Trend’**: 0.00, there is no change in the customer spending habits.

Finally, we created a dataset called ‘transactions’ that focuses on our unique customer with a new feature that helps us gain a deeper knowledge of the customer behavior and it can help us cluster our customers better.

Our final transactional dataset features are:

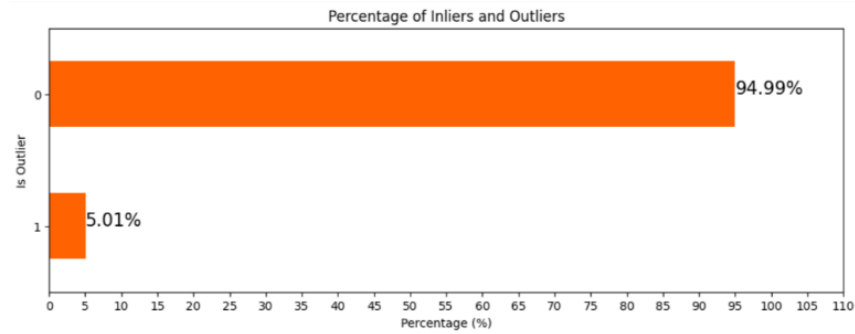
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5350 entries, 0 to 5349
Data columns (total 14 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                -
0   CustomerID                           5350 non-null   string
1   StockCode                            5350 non-null   object
2   Recency                              5350 non-null   Int64
3   Frequency                            5350 non-null   Int64
4   Total_Products_Purchased             5350 non-null   Int64
5   Monetary                             5350 non-null   Float64
6   Average_Transaction_Value            5350 non-null   Float64
7   Unique_Products_Purchased           5350 non-null   Int64
8   Average_Days_Between_Purchases       5350 non-null   Float64
9   Favorite_Shopping_Day                5350 non-null   Int32
10  Favorite_Shopping_Hour               5350 non-null   Int32
11  Monthly_Spending_Mean                5350 non-null   Float64
12  Monthly_Spending_Std                 5350 non-null   Float64
13  Spending_Trend                       5350 non-null   Float64
dtypes: Float64(6), Int32(2), Int64(4), object(1), string(1)
memory usage: 606.2+ KB
```

Variable	Description
CustomerID	Identifier uniquely assigned to each customer, used to distinguish individual customers.
Recency	The number of days that have passed since the customer's last purchase.
Frequency	The total number of transactions made by the customer.
Total_Products_Purchased	The total quantity of products purchased by the customer across all transactions.
Monetary	The total amount of money the customer has spent across all transactions.
Average_Transaction_Value	The average value of the customer's transactions, calculated as total spend divided by the number of transactions.
Unique_Products_Purchased	The number of different products the customer has purchased.
Average_Days_Between_Purchases	The average number of days between consecutive purchases made by the customer.
Favorite_Shopping_Day	The day of the week when the customer prefers to shop, represented numerically (0 for Monday, 6 for Sunday).
Favorite_Shopping_Hour	The hour of the day when the customer prefers to shop, represented in a 24-hour format.
Is_UK	A binary variable indicating whether the customer is based in the UK (1) or not (0).
Monthly_Spending_Mean	The average monthly spending of the customer.
Monthly_Spending_Std	The standard deviation of the customer's monthly spending, indicating the variability in their spending pattern.
Spending_Trend	A numerical representation of the trend in the customer's spending over time. A positive value indicates an increasing trend, a negative value indicates a decreasing trend, and a value close to zero indicates a stable trend.

## Outlier Detection

After the completion of the feature engineering step, it's crucial to check for the availability of outliers in these newly generated features.

To detect these outliers, we used the isolation forest model that will try to detect the number of outliers.

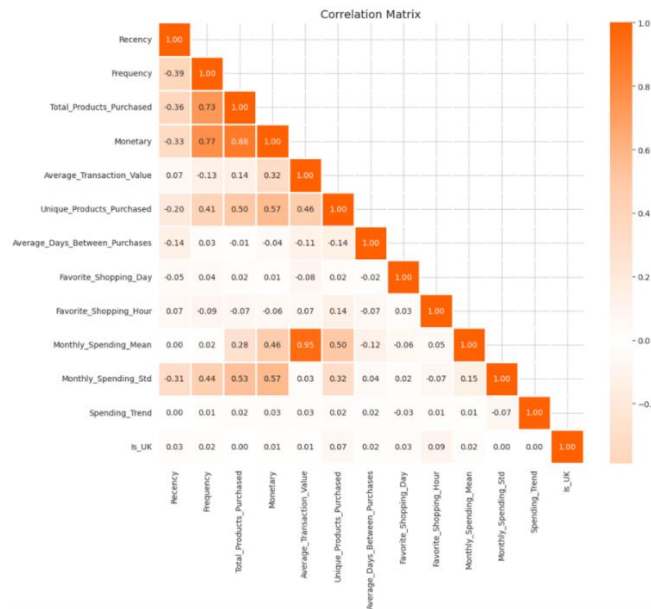


As we can see that there are approximately 5% of the customers that are considered as outliers in the dataset. The outlier percentage is reasonable, and it not considered to be too high so that we don't miss a lot of data. Our model has detected a moderate percentage of outliers in our dataset.

The model will work with a contamination parameter of 0.05 which means that we are expecting the percentage of the outlier in the data to be 5%. The model will try to isolate 5% of the data that might be considered as outliers. Then we created a column named 'Is\_Outlier' where 1 indicates that this row is an outlier and 0 means it's not. Although 5% isn't much, we decided to remove these outliers as they may affect our clustering results later on.

## Correlation Analysis

Studying the correlation between all the numerical features in our dataset will help us understand how each feature is related to the rest of the features. For instance, if there is a feature that is highly correlated with another then that indicates a strong relationship between them which is why we decided to drop “Monthly\_Spending\_Mean” as it’s highly correlated with “Average\_Transaction\_Value” with a correlation of 95%.



## Feature Scaling

Feature scaling is a preprocessing step used to transform numerical features into a similar scale. This also contributes to having a better clustering results. We used the standardization technique to scale all the numerical features so all the values will be centered around 0 which will be the mean in this case.

# *Dimensionality Reduction*

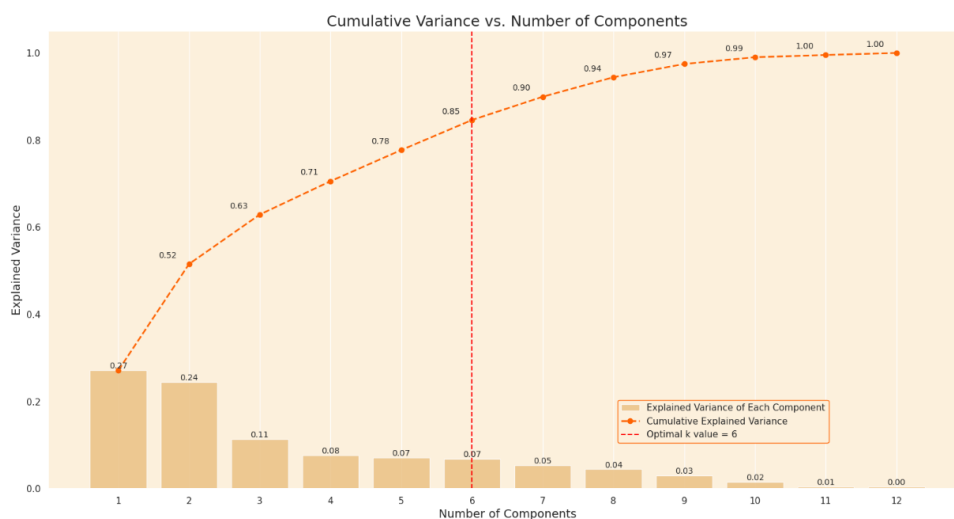
## Why did we apply Dimensionality Reduction?

- Since we will be applying a K-means clustering method that is based on distance, it can struggle with having many features. Therefore, we needed the clustering method to focus on the relevant features rather than keeping them all.
- If we focused on only the relevant features, we could reduce the amount of noise that is in our data, which can lead to more accurate clustering results.
- We will be interpreting the clustering results by visualizing them. However, visualizing them with all the features can be challenging, therefore visualizing the customer groups into two or three dimensions can be easier to understand and interpret.
- The computation speed will increase since we have reduced the number of features.

## Principle Component Analysis (PCA) method

We will be applying PCA methods. It is one of the most well-known techniques, we chose it because it allows us to remove irrelevant information while keeping a significant amount of the original information in our dataset. And it is computationally an efficient technique.

To identify the optimal k number of components we will start by visualizing all the components (features) that are available up to 12 features. On top of each component there is a variance value, which means how much this component can add extra information or value to our dataset. Those variances are cumulated with its previous cumulated component variance representing a curved line indicating how much of the total variance in the dataset is captured by each principal component.



So, for component number 1, it describes approximately 27% of the variance. For the second component it explains 52% of the variance, 63% of the variance is described by component number 3, 71% of the variance

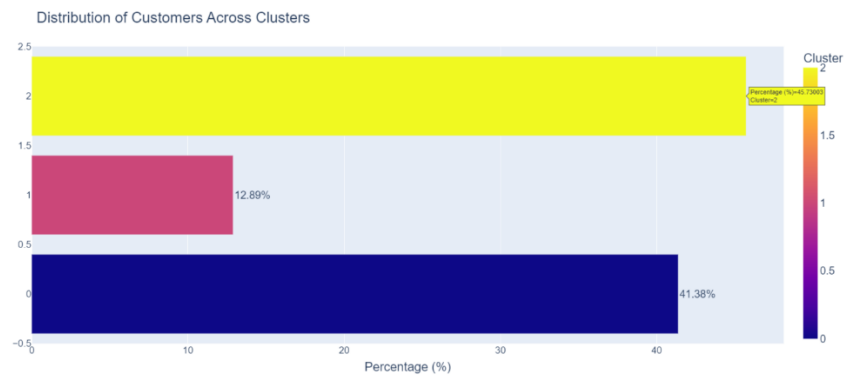
is described by component number 4, 78% of the variance is described in component number 5. Also, we have reached our optimal k components with the sixth component that explains 85% of the variance. After the 6<sup>th</sup> component adding another component doesn't significantly increase the cumulative variance which means it doesn't add extra value to the components. As we can see from the plot the increase in the cumulative variance starts stopping in the 6<sup>th</sup> component which captures approximately 85% of the total variance.

Our next step is to create a new dataset called 'transactions\_pca' which involves only the first 6 components to retain the relevant information to effectively group customers. Each row represents a unique customer and for the columns they represent different components from PC1 to PC6. So, for each customer the data has been compressed into these 6 components and the values mean how much each principle component is present for each customer.

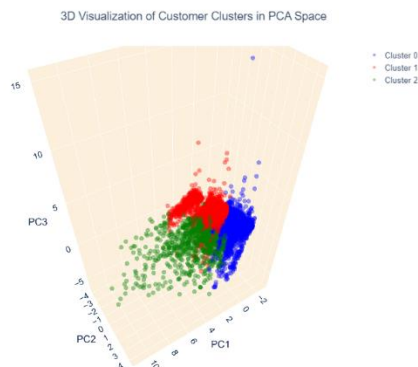
# Clustering

For the purpose of understanding our customer's behavior, we have decided to employ clustering as a way to segment them to gain valuable insights about their behavior, needs and preferences. Segmentation enables businesses to construct targeted marketing strategies to enhance the product offerings and the satisfaction of the customers. We have implemented 3 different clustering techniques. Starting off with K-Means, as we have a determined number of segments for our customers, which is 3, we haven't applied any specific approach which specifies the ideal number of k clusters. We have clustered the customers with similar features as this indicates that they have almost identical shopping behaviors.

## Clustering Model - K-means



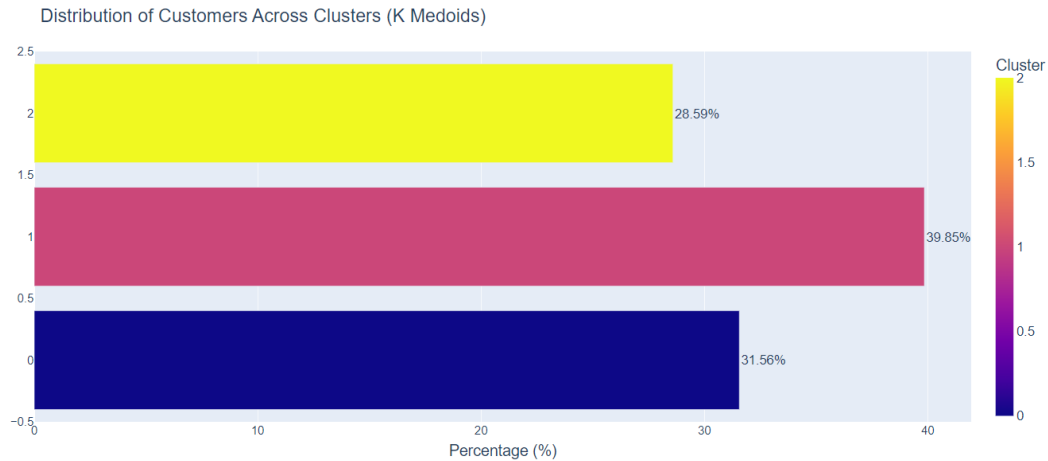
Cluster 0 had 2103 points, Cluster 1 had 655 points, and Cluster 2 had 2324 points indicating it has the highest portion of the population of the customers, 0.217 is the outcome of the silhouette score. This result signifies that the clusters are separated well. The image below shows the distribution of the data points across the clusters:



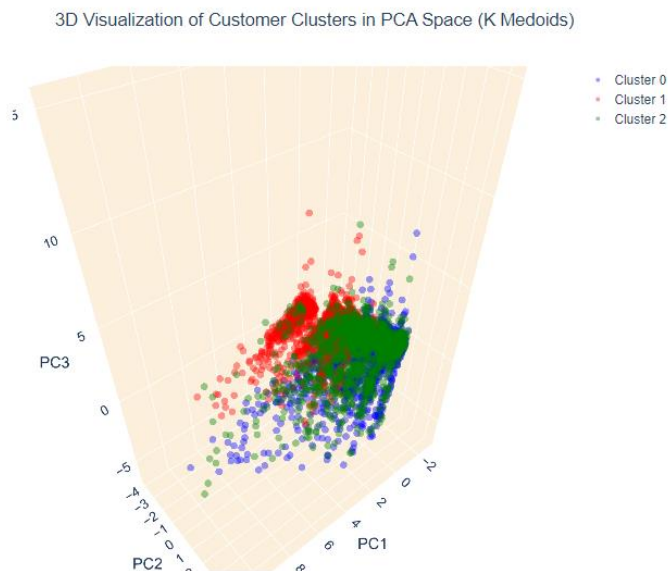
Metric	Value
Number of Observations	5082
Silhouette Score	0.21798565373564963
Calinski Harabasz Score	1214.3687026666817
Davies Bouldin Score	1.498730109194775

## Clustering Model - K-medoids

The second clustering method that we have applied is K-Medoids, with the same number of segments, which is three.



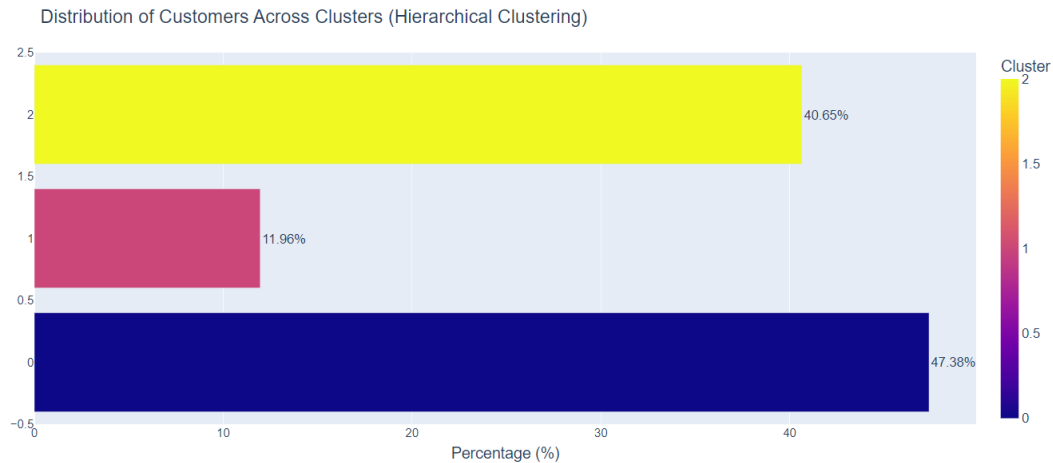
The outcome is that Cluster 0 had 1603 customers, Cluster 1 had 2025 customers and Cluster 2 had 1454 customers. 0.08 is the silhouette score, this suggests that there's a weak separation between the clusters. Therefore, we concluded that the K-Medoids isn't the appropriate method to use for clustering as it wasn't able to separate the clusters well.





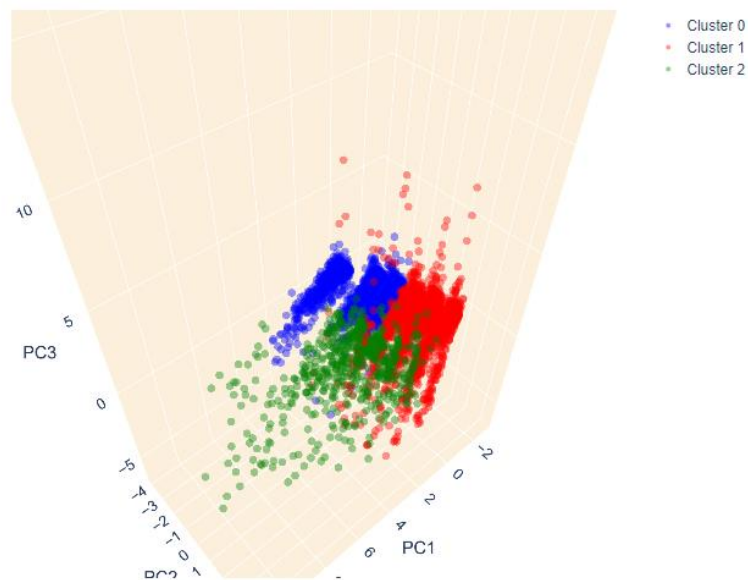
## Clustering Model - Hierarchical

The last clustering algorithm that we tried was Hierarchical clustering. The outcome of this method was that Cluster 0 had 2391 points, Cluster 1 had 608 points and Cluster 2 had 2066 points.



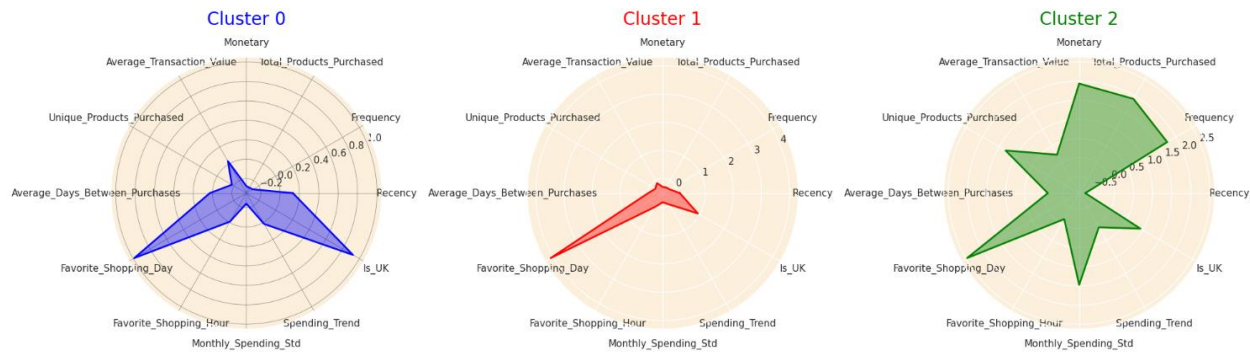
With a silhouette score of 0.211. At the end of this step, we have determined that both K-Means and Hierarchical clustering gave promising results. However, we have decided to proceed with the results of K-Means since the silhouette score is slightly higher which means it separated the points better.

3D Visualization of Customer Clusters in PCA Space (Hierarchical Clustering)



# Cluster Analysis and Profiling

## Radar Chart (results and interpretation)



Based on the radar chart results and the characteristics of each cluster, these are the names of our segments:

**Cluster 0:** Local Shoppers > customers are mostly located in the UK.

**Cluster 1:** Regular Shoppers > they are customers that tend to purchase at a moderate level but not frequently.

**Cluster 2:** Frequent Buyers > They tend to purchase a lot in terms of high monetary, frequency and total products purchased.

## Frequent Pattern Mining and Association Rule Mining

Frequent pattern mining is important in data mining to identify the most frequent patterns occurring in the dataset. We have implemented it to understand the behavioral patterns of the consumers. By analyzing the transactional data, it will enable us to identify patterns of the purchases and associations among purchases for each segment. By identifying the frequent items, businesses will be able to tailor their marketing campaigns and recommendations of the products. We selected a minimum support of 0.06 to extract frequent items.

### Cluster 0 - Local Shoppers (Results and Interpretation)

Cluster 0 had 119 frequent item sets which indicates that this segment has a basic or limited shopping patterns, they show consistent purchase habits, but only for specific items.

One of the frequent products was “**JUMBO BAG RED RETROSPOT**” with a support of 0.178, this item is a tote bag that has a retro spot pattern, this item was frequently bought with “**JUMBO BAG APPLES**“, which is another tote bag that has apples on the bag. We have noticed that the “**JUMBO BAG RED RETROSPOT**” item was

frequently bought with another bag that has a different pattern. This implies that local shoppers have a preference for tote bags, or because of a fashion trend, or perhaps there was a promotion in the shop, where if they buy a tote bag, they get the other one for free.

Another item that was frequent was the ‘**PINK REGENCY TEACUP AND SAUCER**’ which is a teacup and a saucer that are pink, this item was frequently bought with another item twice, each time with a teacup and saucer of different colors or patterns. This pattern reveals a similar association as observed before with the first item. Customers who have this behavior may have a desire for collecting things of different patterns to add diversity to their collection, or perhaps they like to mix-and-match, some people may like to have a diverse range of colors and patterns for their tableware.

By understanding the frequent items, the business can take advantage of the association between the frequent items to create promotions that persuade the customers to buy these items together. Also, the business can optimize their inventory management tactics as they should ensure that the items that are bought together should be sufficient stock levels of these products.

### **Association rules of cluster 0**

After extracting the frequent patterns, we should understand the rules of cluster. Cluster 0 had 23 association rules. Some of the interesting rules were:

**JUMBO BAG APPLES → JUMBO BAG RED RETROSPOT.**

**JUMBO BAG SCANDINAVIAN BLUE PAISLEY → JUMBO BAG RED RETROSPOT.**

We have observed that customers who buy any type of jumbo bag, tend to buy the jumbo bag red with the retro spot pattern afterwards, this means that this specific item is consistently being bought by other customers which is why they always include it in their transactions. This information can help the business by applying cross-selling, as they could place the jumbo bag with the red retro spot with other items as a bundle offer, so this encourages the customer to buy additional items.

Another set of interesting rules were:

**ROSES REGENCY TEACUP AND SAUCER → PINK REGENCY TEACUP AND SAUCER.**

**PINK REGENCY TEACUP AND SAUCER → ROSES REGENCY TEACUP AND SAUCER.**

**GREEN REGENCY TEACUP AND SAUCER → PINK REGENCY TEACUP AND SAUCER.**

These rules have revealed that local shoppers love to collect diverse colors/patterns of teacups and saucers, as whenever they buy a teacup and saucer in a specific color they tend to buy it in a different color, this indicates that these results could be influenced by seasons or occasions, where they buy different colors for particular occasions. Sellers can apply cross-selling for these products as well. Another thing the sellers can do is to offer customization options for the customer where they can pick a specific color for a teacup and a different color for the saucer, allowing the customers to make their purchases based on their preferences. Moreover, sellers can introduce limited edition sets featuring unique patterns and colors, this will drive the customers to buy them to add the limited-edition items to their collections.

## Cluster 1 - Regular Shoppers (Results and Interpretation)

When we first extracted the frequent item sets in cluster1 we found 93 of them. The itemset that had the highest support was 'CREAM HANGING HEART T-LIGHT HOLDER' with a support of 0.171. Which means that customers in cluster 1, the regular shoppers mostly buy this item. Other interesting frequent items were 'JUMBO BAG RED RETROSPOT' with a support of 0.139. This item is common with cluster 0 meaning that there is some common interest between the local shoppers and the regular shoppers or the some of the regular shoppers might be local as well.

Regular shoppers also tend a lot to buying cake and party related items such as 'REGENCY CAKESTAND 3 TIER', 'PARTY BUNTING', 'SET OF 3 CAKE TINS PANTRY DESIGN ' and '72 SWEETHEART FAIRY CAKE CASES' which might mean that they like having parties from time to time.

### Association Rules of Cluster 1

Studying the rules associated for cluster1 – the regular shoppers help in understanding the behavior of their shopping. What was interesting about the association rules generated for this cluster is that we only got 3 rules based on the thresholds chosen. The rules were:

**JUMBO BAG POLKADOT —→ JUMBO BAG RED RETROSPOT.**

**GREEN REGENCY TEACUP AND SAUCER —→ ROSES REGENCY TEACUP AND SAUCER.**

**ROSES REGENCY TEACUP AND SAUCER —→ GREEN REGENCY TEACUP AND SAUCER.**

We are 70% sure that regular shoppers who buy 'JUMBO BAG POLKADOT' are most likely going to buy 'JUMBO BAG RED RETROSPOT' as well. A similar rule was generated for customers in cluster 0, meaning that most of the customers in our data likes shopping for these types of bags. One good suggestion for the business besides the one mentioned for cluster 0 is recommending the 'JUMBO BAG RED RETROSPOT' whenever the customer views or buys 'JUMBO BAG POLKADOT'.

Moreover, we are 72% sure that regular shoppers who buy 'GREEN REGENCY TEACUP AND SAUCER' are most likely going to buy 'ROSES REGENCY TEACUP AND SAUCER' and 76% sure that shoppers who buy 'ROSES REGENCY TEACUP AND SAUCER' are going to buy 'GREEN REGENCY TEACUP AND SAUCER' which is a higher confidence. This indicates the interest of having matching sets of teacups and saucers.

## Cluster 2 - Frequent Buyers (Results and Interpretation)

We have extracted 89,044 frequent items, this cluster has the most frequent items this indicates that there's a wide range of purchasing behaviors, customers in this clusters who are considered frequent buyers like to engage in different product combinations, which is what lead to this number of frequent item sets.

A few of the most frequent items were: 'JUMBO BAG RED RETROSPOT', 'SET OF 3 CAKE TINS PANTRY DESIGN ', 'NATURAL SLATE HEART CHALKBOARD ' and 'HEART OF WICKER SMALL'. This suggests several things about the behavior of these customers. For example, since the most frequent item bought is the jumbo

bag, this indicates that these customers like practicality, as they are frequent buyers, they'd want a tote bag that's practical to put all the items they bought in. Another thing that we observed was that the customers in this segment are full of sentiment, they have consistently bought heart-shaped items that could be given as a gift in many different occasions.

Retailers can benefit from these insights to make tailored promotions for the frequent customers. For example, the retailers can offer discounts on large purchases of these items or exclusive offers for themed products that are decorative, since the consumers in this segment tend to buy lots of decorative items. By doing this, they encourage the customers to buy more and this is going to increase their sales.

## **Association Rules of Cluster 2**

We have extracted 2193864 rules from this clusters, the rules with the highest confidence and the most interesting ones were:

### **SET OF 3 CAKE TINS PANTRY DESIGN —> REGENCY CAKESTAND 3 TIER.**

This rule signifies that customers who are interested in buying cake tins that are pantry-themed tend to buy a cake stand to add it to their pantry. These customers like baking, perhaps they are chefs or have a hobby of cooking. They appreciate both design and functionality in kitchen items as they look for matching sets for their baking and presentation. Retailers can take advantage of these insights to arrange kitchen or baking related collections or give targeted promotions to the customers who consistently buy these items.

### **KNEELING MAT HOUSEWORK DESIGN —> GARDENERS KNEELING PAD CUP OF TEA.**

This rule implies that there's a correlation between the customers who buy housework designed kneeling mats and a gardening kneeling pad. Customers who bought this are likely to enjoy doing chores and take pleasure in sitting outdoors as well as gardening, since they bought gardening related items. The retailers can leverage this insight by making subscription services that deliver convenience and savings to those customers that make purchases in this category of items. By adding complementary items into the subscription packages, retailers will benefit from the long-term buyers relationship that they will establish with their customers.

## ***Conclusion***

To Conclude, in our data mining project we looked at a huge amount of data consisting of 541,909 rows to understand how customers behave and understand their purchasing behavior to formulate targeted marketing strategies. We decided to stick with the K-means clustering method and grouped our customers into 3 distinct segments which are: Local shoppers, Regular Shoppers, and Frequent Buyers. The silhouette score was 0.218 which was the highest amongst various clustering methods that we have tried. By leveraging these insights, markets can benefit from this information to tailor their decisions and strategies to fit each cluster's unique preferences. Furthermore, we generated rules to help delve deeper into understanding the customers in each segment and bring actionable rules. In essence, businesses can benefit from these rules and gain valuable insights into customer behavior which can drive sustainable growth.

## *References*

[1] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, “RFM Ranking – an Effective Approach to Customer Segmentation,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 10, Sep. 2018, doi: <https://doi.org/10.1016/j.jksuci.2018.09.004>.

[2] Ecosystem, E. (2022, May 17). Understanding K-means Clustering in Machine Learning. Medium. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

[3] G. (2023, August 31). Frequent Pattern Mining in Data Mining. GeeksforGeeks. <https://www.geeksforgeeks.org/frequent-pattern-mining-in-data-mining/>

[4] Twin, A. (2024, February 23). What Is Data Mining? How It Works, Benefits, Techniques, and Examples. Investopedia. <https://www.investopedia.com/terms/d/datamining.asp#:~:text=Data%20mining%20uses%20algorithms%20and,neural%20networks%2C%20and%20predictive%20analysis.>

[5] G. (2023, January 11). Association Rule. GeeksforGeeks. <https://www.geeksforgeeks.org/association-rule/>

# Appendices

## Raw Data:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

## IsolationForest Model

```
# Initializing the IsolationForest model with a contamination parameter of 0.05
model = IsolationForest(contamination=0.05, random_state=0)

# Fitting the model on our dataset (excluding 'CustomerID' and 'StockCode')
transactions['Outlier_Scores'] = model.fit_predict(transactions.drop(columns=['CustomerID', 'StockCode']).to_numpy())

# Creating a new column to identify outliers (1 for inliers and -1 for outliers)
transactions['Is_Outlier'] = [1 if x == -1 else 0 for x in transactions['Outlier_Scores']]

# Display the first few rows of the transactions dataframe
transactions.head()
```

## Generating Rules Function

```
# Define a function to generate association rules with a minimum confidence
def generate_rules(frequent_itemsets, metric="lift", min_threshold=1, min_confidence=0.5):
    rules = association_rules(frequent_itemsets, metric=metric, min_threshold=min_threshold)
    # Filter rules based on confidence
    rules = rules[rules['confidence'] >= min_confidence].reset_index(drop=True)
    return rules
```