Capstone Project

Online Retail Customer Segmentation

By,
Rochak P V
rochakr4@gmail.com
cohort Hardeol

Problem Statement

This project aims to identify major customer segments on a transnational data set that contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

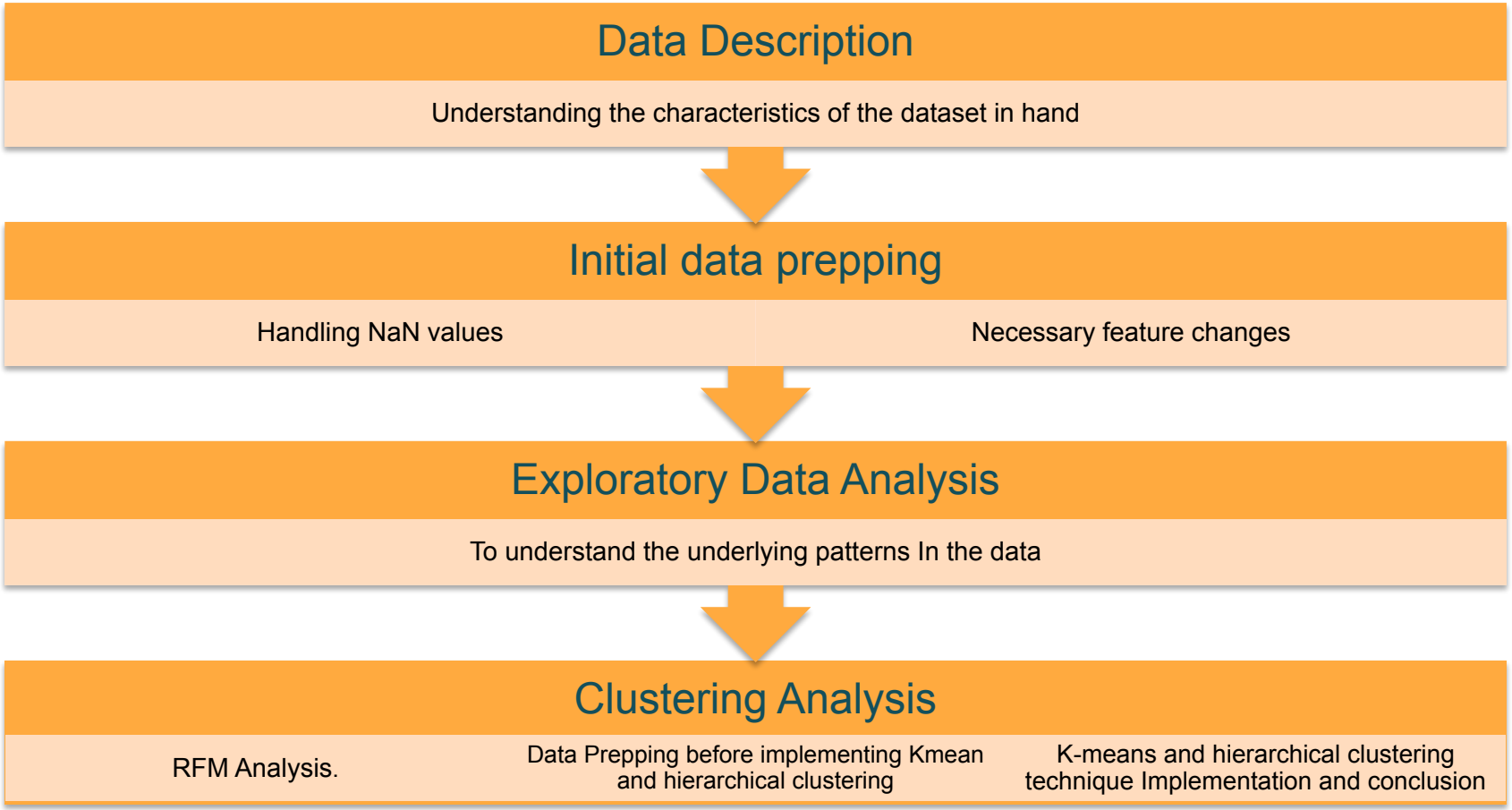
Objective

As mentioned in the problem statement, the objective of this project is to identify major customer segments using unsupervised machine learning techniques.

Methodology

Clustering is a data mining technique that groups unlabeled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information. Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic

Steps involved



Data Description: Understanding the characteristics of the dataset in hand

The dataset contains information pertaining to 5,41,909 purchases occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retailer.

Column Information:

1. **InvoiceNo:** An invoice number, nominal, is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter "c", it indicates a cancellation.
2. **StockCode:** Product (item) code. Nominal, is a 5-digit integral number uniquely assigned to each distinct product.
3. **Description:** Product (item) name.
4. **Quantity:** The quantities of each product (item) per transaction.
5. **InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. **UnitPrice:** Product price per unit in sterling.
7. **CustomerID:** Customer number. Nominal, is a 5-digit integral number uniquely assigned to each customer.

Libraries used:

NumPy, Pandas, Seaborn, Plotly, Matplot and Scikit Learn:

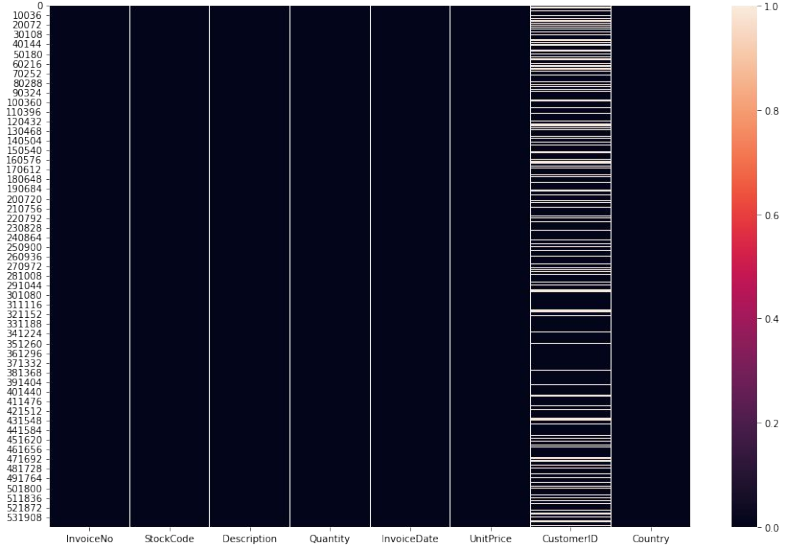
Data Description: Understanding the characteristics of the dataset in hand

A brief statistics description of all numerical features is given below,

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Brief information regarding the dataset such as NaN value count and data type is given below,

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```



Initial data prepping

Handling NaN values

- A lot of Nan values are found in the 'CustomerID' feature, which in turn is a unique identifier. So, implementing imputation techniques doesn't make sense. For further analysis NaN value rows are dropped.

Feature alteration

- Three new features i.e., 'Month', 'Day', 'Day_name' and 'year' are introduced using the 'invoice date' feature.
- There are 8,905 cancellation entries in the dataset. For better analysis, I am dropping all the canceled invoice rows because there are chances that it may act as noise or bias.
- Here all the transactions related to the cancellation are not canceled because the customer had an urge or intention of purchasing the item.
- For better understanding and analysis, a new feature is introduced i.e., 'Total sales' which is the arrived by multiplying quantity and unit price

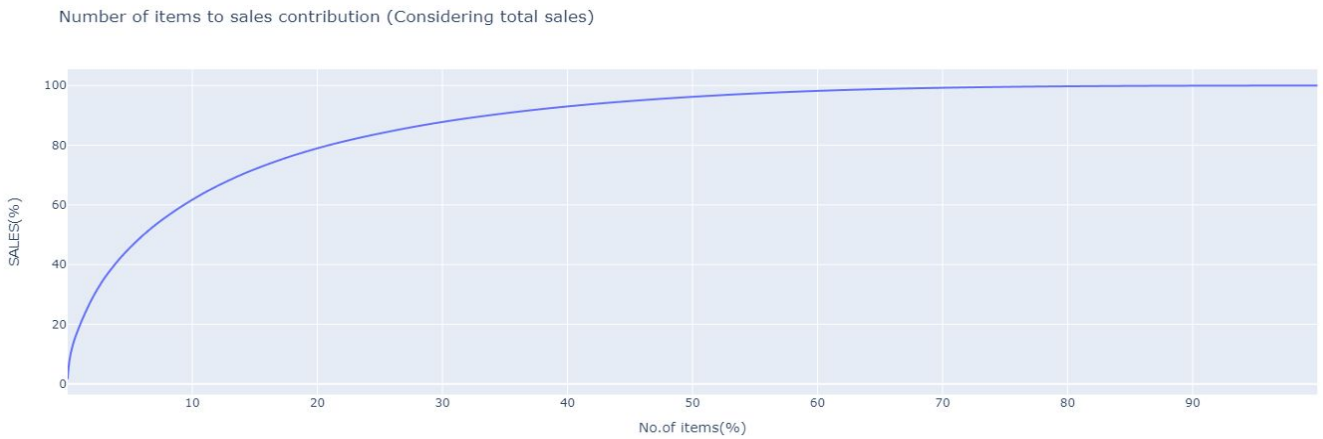
Exploratory Data Analysis

Top five and last five items in terms of number of items sold

Description	Quantity
PAPER CRAFT , LITTLE BIRDIE	80995
MEDIUM CERAMIC TOP STORAGE JAR	77916
WORLD WAR 2 GLIDERS ASSTD DESIGNS	54415
JUMBO BAG RED RETROSPOT	46181
WHITE HANGING HEART T-LIGHT HOLDER	36725
...	...
BLACK VINT ART DEC CRYSTAL BRACELET	1
FLOWER SHOP DESIGN MUG	1
SET 36 COLOURING PENCILS DOILEY	1
HEN HOUSE W CHICK IN NEST	1
AMBER BERTIE GLASS BEAD BAG CHARM	1

Top five and last five items in terms of unit price

Description	Quantity	UnitPrice
POSTAGE	3120	8142.750
Manual	7179	4161.060
DOTCOM POSTAGE	16	1599.260
PICNIC BASKET WICKER 60 PIECES	61	649.500
VINTAGE BLUE KITCHEN CABINET	26	295.000
...	...	...
POPART WOODEN PENCILS ASST	8900	0.120
FOLDING CAMPING SCISSOR W/KNIF & S	30	0.120
PORCELAIN BUDAH INCENSE HOLDER	1501	0.100
WRAP BAD HAIR DAY	700	0.100
PADS TO MATCH ALL CUSHIONS	4	0.001

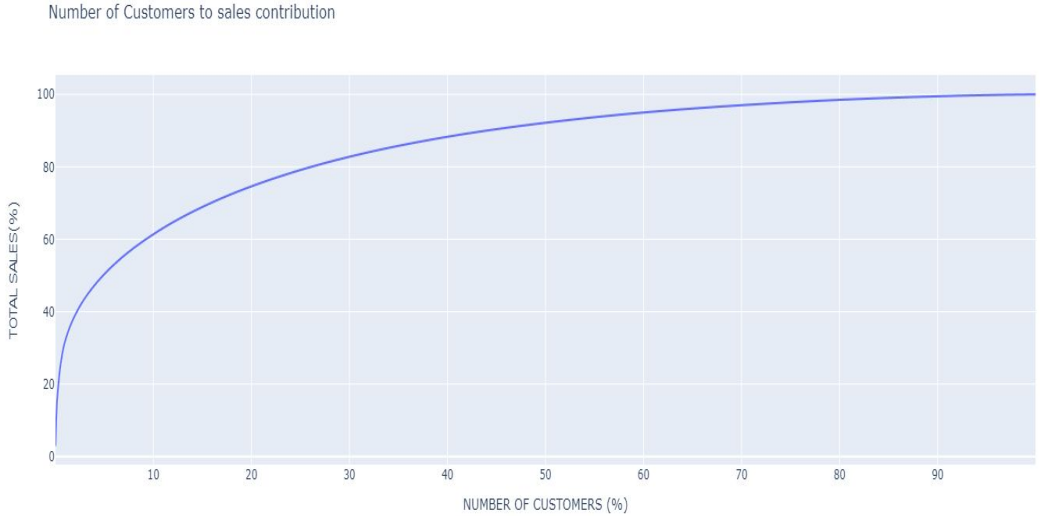


- **There are 3,877 types of items sold from 01/12/2010 to 09/12/2011**
- **The unit price of items is ranging from 0.001 to 8142.75.**
- **Unit price of 75% of items sold is less than 4.**
- **Top 237 sold items (considering total sale) i.e., 6.11% of total items for sale constitute 50.04% of total sales.**
- **Top 1306 sold items (considering total sale) i.e., 33.69% of total items for sale constitute 90.0% of total sales**

Exploratory Data Analysis

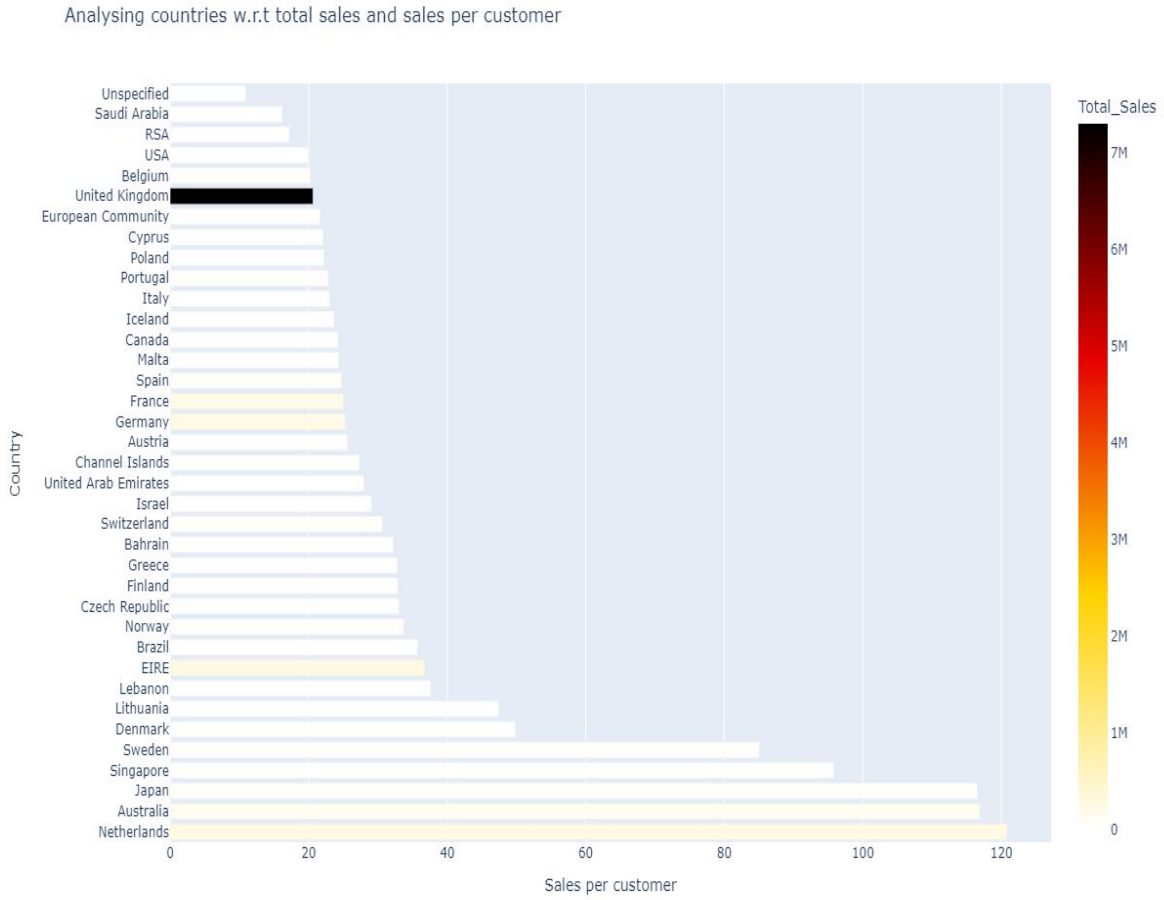
Top five and last five customerIDs in terms of total sales

CustomerID	Total_Sales
14646.0	280206.02
18102.0	259657.30
17450.0	194550.79
16446.0	168472.50
14911.0	143825.06
...	...
17956.0	12.75
16454.0	6.90
14792.0	6.20
16738.0	3.75
13256.0	0.00



- **There are 4,339 customers or customer IDs involved in this dataset.**
- **Top 212 customers (considering total sales) i.e., 4.89% of total customers constitute 50.05% of total sales.**
- **Top 1908 customers (considering total sales) i.e., 43.97% of total customers constitute 90.0% of total sales.**

Exploratory Data Analysis

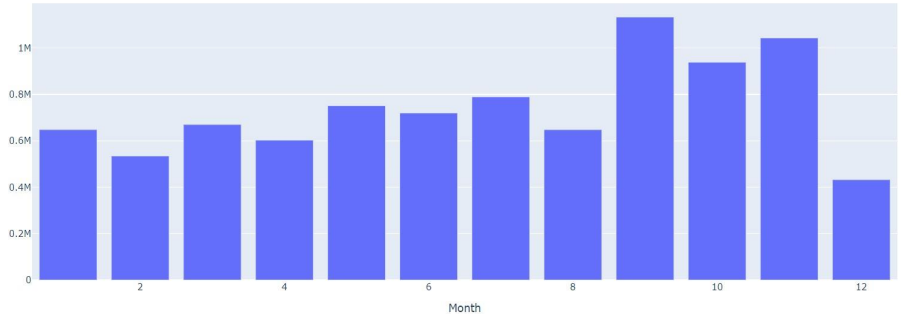


- **The United Kingdom has the highest sales and customers.**

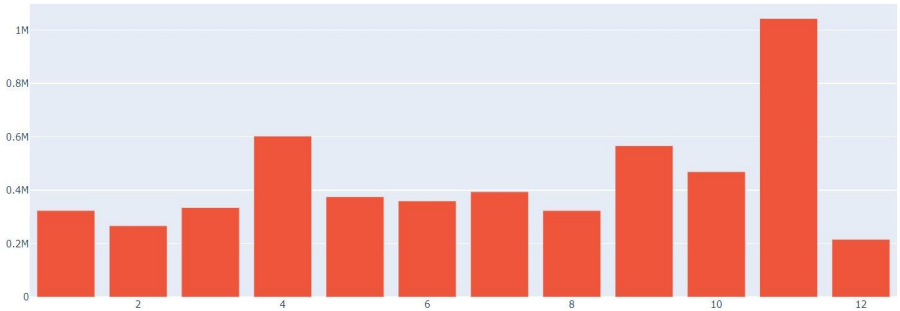
- **While considering the sales per customer ratio, the Netherlands, Australia and Japan are at the top.**

Exploratory Data Analysis

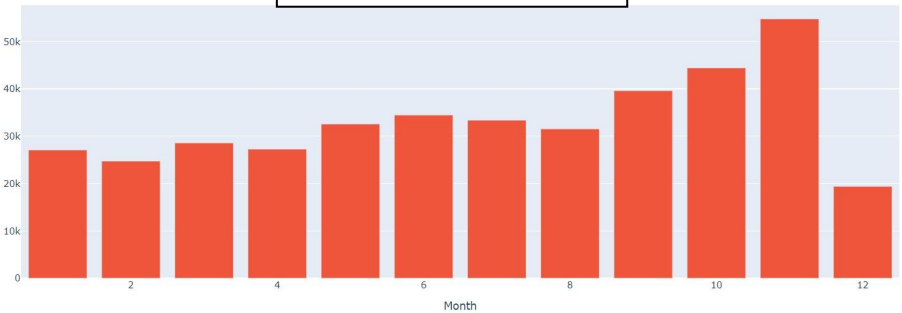
Month-wise total sales registered



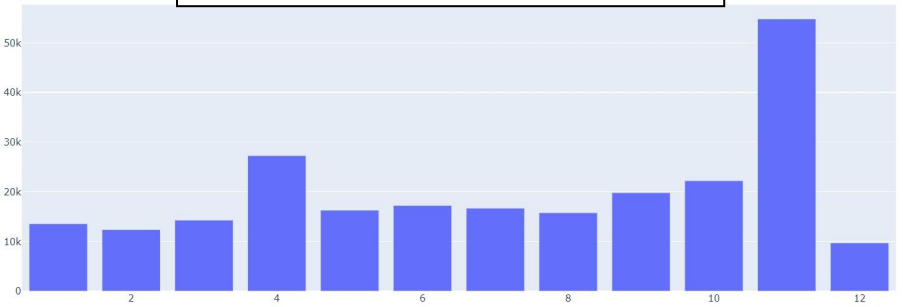
Month-wise Number of customers



Month-wise average sales



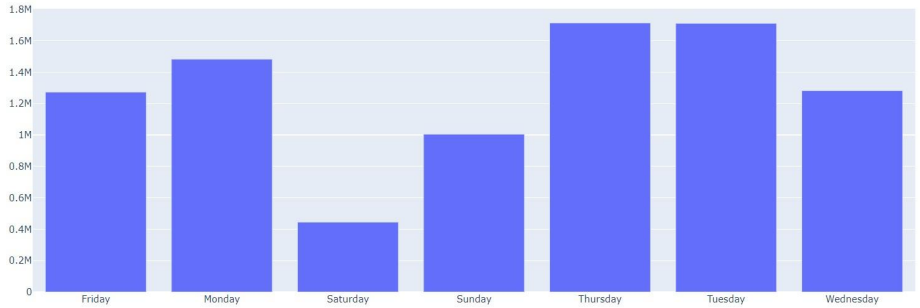
Month-wise average Number of customers



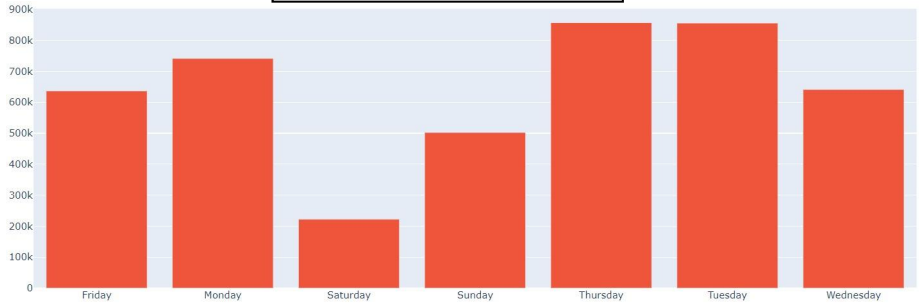
- **September month had the highest sales and November had the highest average sales.**
- **November month had the highest number of customers and average customer visits.**

Exploratory Data Analysis

Day-wise total sales registered

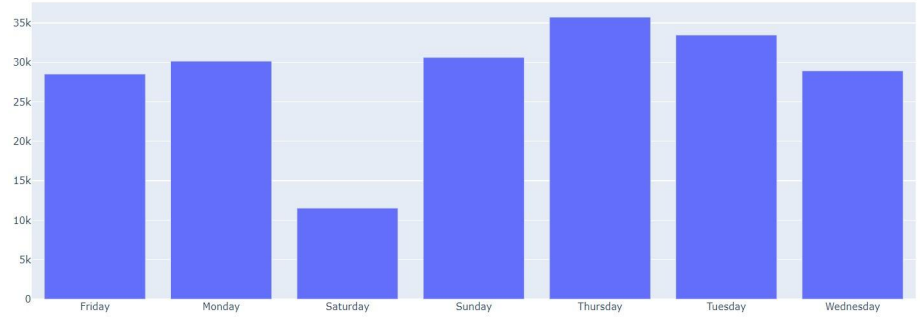


Day-wise average sales



Day-wise Number of customers

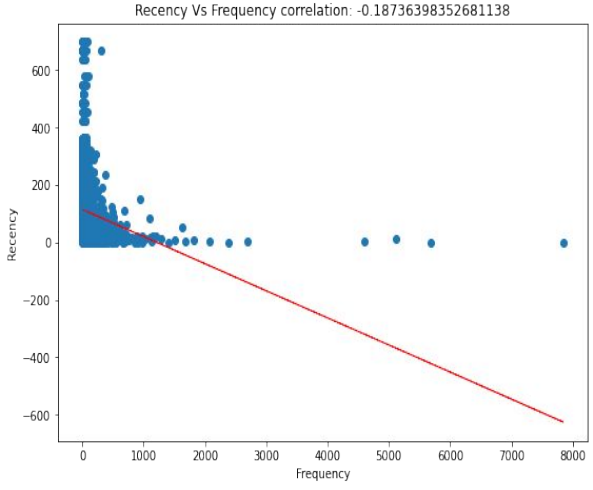
Day-wise average Number of customers

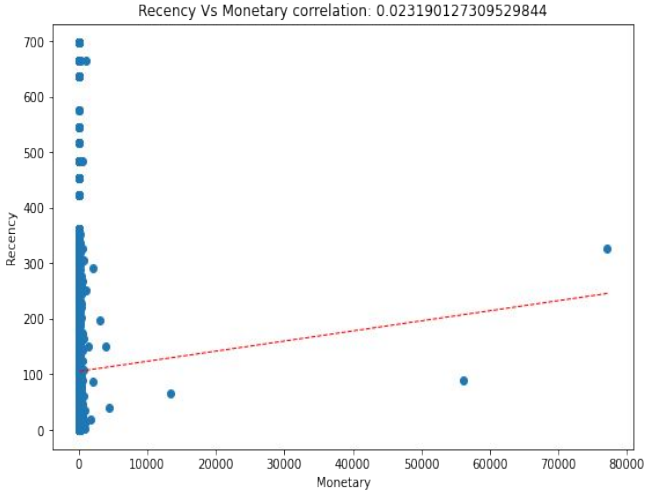


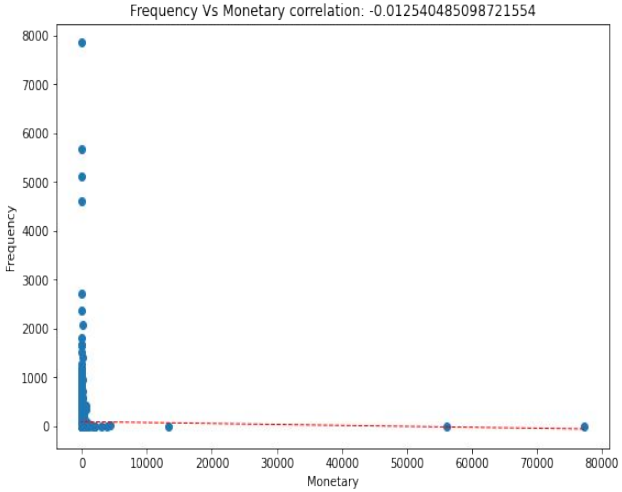
Thursday, followed by Tuesday, has the highest sales, customer visits, average sales and average customer visits.

Clustering Analysis

The RFM (recency, frequency, and monetary) model is a behaviour-based model used to analyze the behaviour of a customer and then make predictions based on the behaviour in the database. Moreover, recency represents the length of a time period since the last purchase, while frequency denotes the number of purchases within a specified time period. For monetary purposes, customers are coded by the total or average amount of money spent during a specified period of time



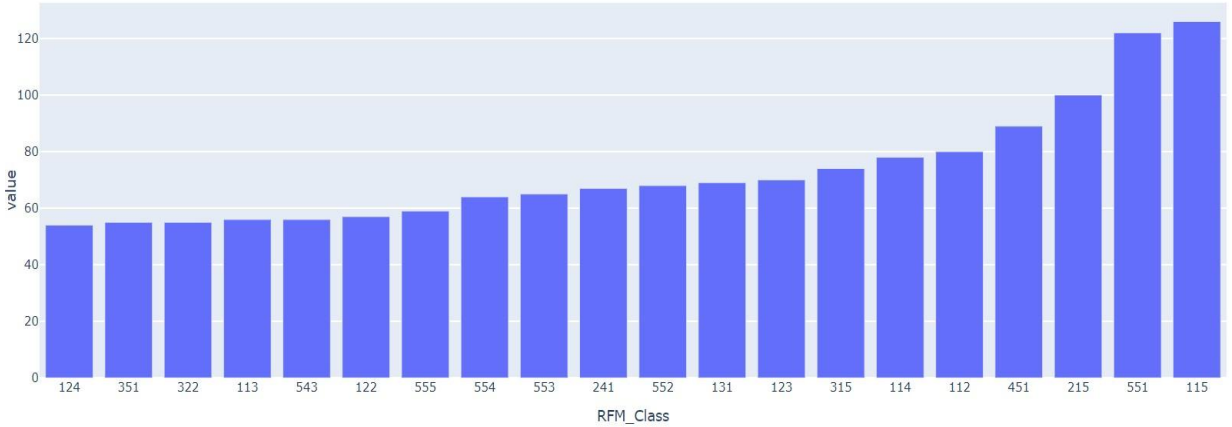


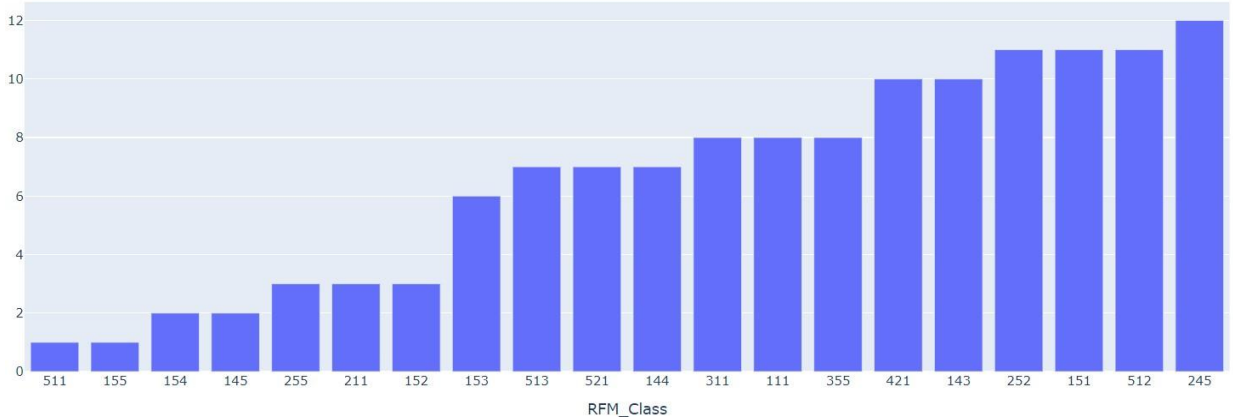


- **High correlation is not observed between 'Recency', 'Frequency' and 'Monetary'.**
- **The highest correlation value observed is 0.187 which is between Recency' and 'Frequency'.**

Clustering Analysis

The quintile method was used to compute R, F, and M scores, which divide recency, frequency, and monetary values into 5 groups (denoted from 1 to 5). Then, by combining R, F, and M scores, a three-digit RFM cell code was created and a composite score, which is the weighted sum of R, F, and M scores, was created.





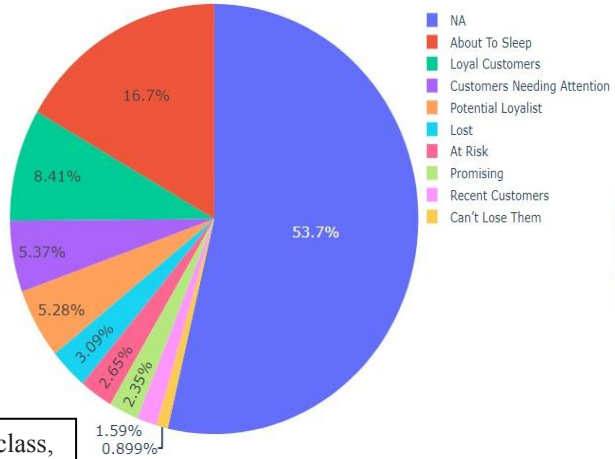
- **115, 551, 215, 451 and 112 are the top 5 RFM classes in terms of the number of customers.**

- **511, 155, 154, 145 and 255 are the last 5 RFM classes in terms of the number of customers.**

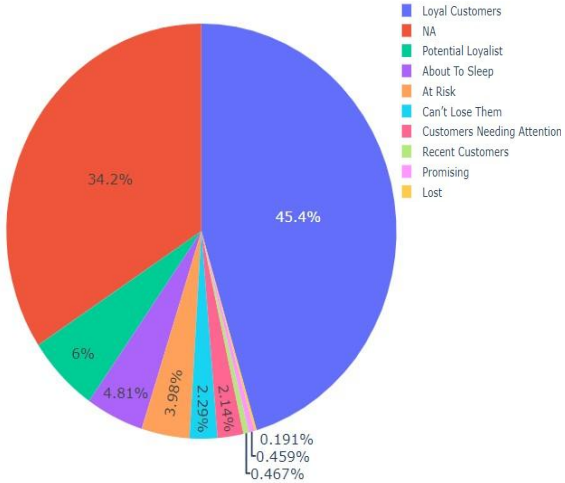
Clustering Analysis

SEGMENT	DESCRIPTION
Loyal Customers	Bought recently, buy often and spend the most ['555', '545', '455', '445','454','554', '444', '544']
Potential Loyalist	Recent customers, but spent a good amount and bought more than once. ['524', '525', '535', '534','551']
Recent Customers	Bought most recently, but not often. ['515', '514','522','511']
Promising	Recent shoppers, but haven't spent much. ['513', '523', '512', '533', '414']
Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though. ['333', '323', '334', '324','433','343']
About To Sleep	Below average recency, frequency and monetary values. ['222', '221', '212', '122', '223', '322', '232', '123', '131', '132', '311', '411', '215', '213', '113', '312']
At Risk	Spent big money and purchased often. But long time ago. ['154', '254', '354','344','244','144']
Can't Lose Them	Made biggest purchases, and often. But haven't returned for a long time. ['145', '155','255','245','135','144', '354','355']
Lost	Lowest recency, frequency and monetary scores. ['111','112', '121', '211']

Customer category percentage distribution w.r.t number of customers,



Customer category percentage distribution w.r.t total amount,



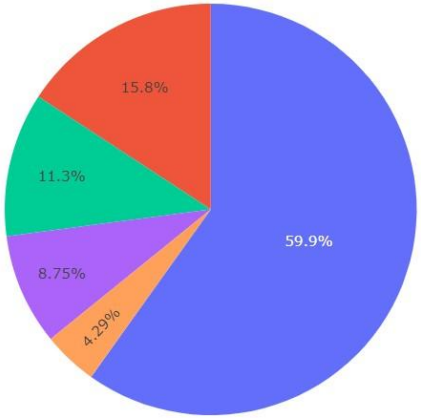
3D visualisation of a customer segment based on RFM class,

- **Around 50% of customers were able to categorise based on RFM class.**
- **Loyal customers (8.41%) are contributing 45.4% of total revenue**

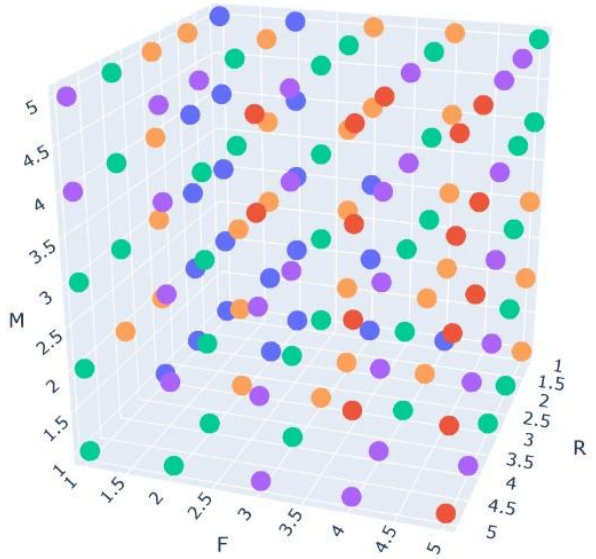
Clustering Analysis

A new feature, "RFM_Score," which is the sum or weighted sum of R, F, and M scores, has been added for further analysis.
In this project, a weighted sum is used, i.e., RFM_Score = 3R + 2F + M.
Based on the RFM composite score, customers were classified into 5 customer segments (A+, A, B+, B, C)

Customer category percentage distribution w.r.t total amount,

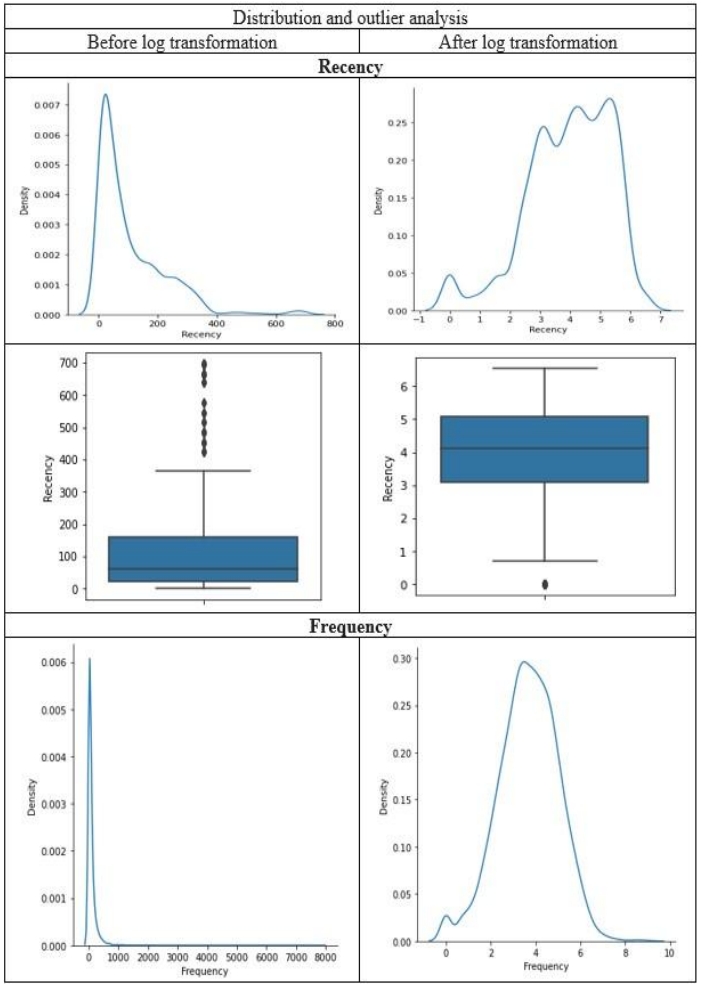


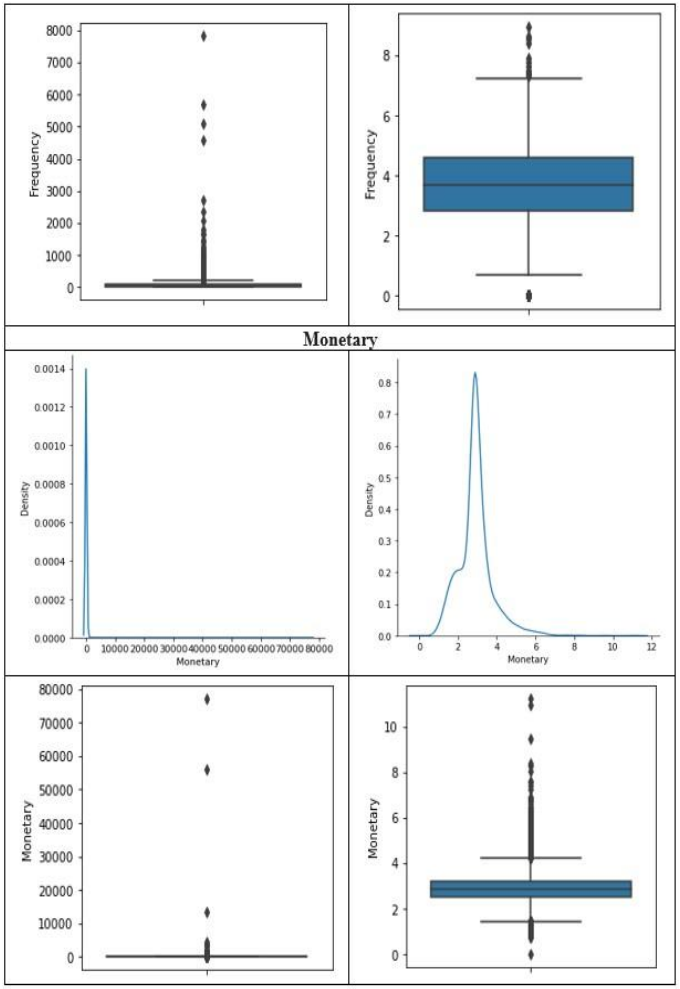
3D visualisation of a customer grade based on RFM class,



- **A+ graded customers, i.e., 18% of total customers, are responsible for 60% of total sales.**

Clustering Analysis: Data Prepping before implementing Kmean and hierarchical clustering





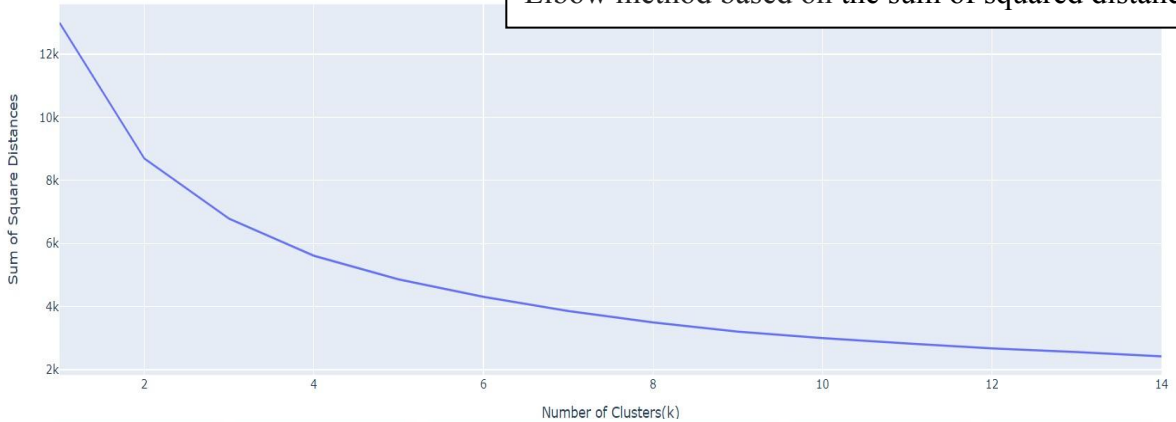
- **Frequency and monetary features have high skewness, and all features have outliers. In order to handle this, log transformation is implemented.**

- **After log transformation, a good number of outliers are observed in the monetary feature. In order to reduce those, capping has been implemented. Rows with a monetary value greater than 2,000 are omitted from the dataset**

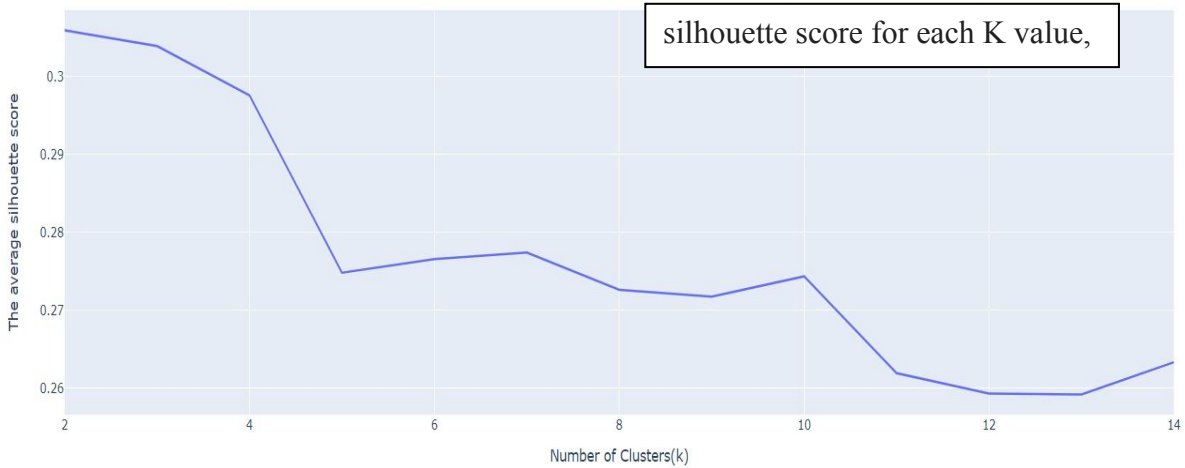
Clustering Analysis: K-mean clustering

In order to get the optimum K value, the elbow method based on the sum of squared distance and silhouette analysis is carried out,

Elbow method based on the sum of squared distance,



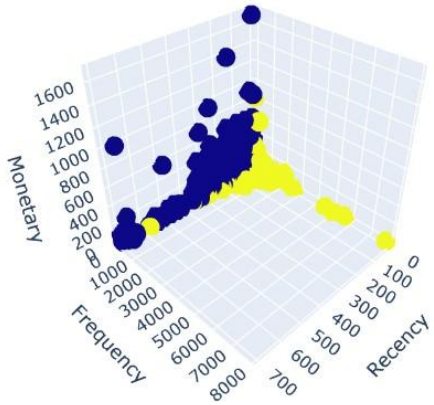
silhouette score for each K value,



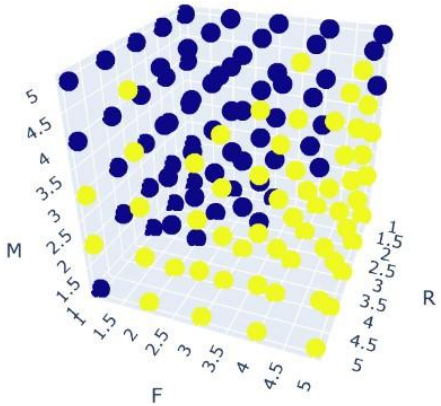
- **Both the silhouette and elbow methods prefer k value 2, then 3. But for further analysis, both K values are being used**

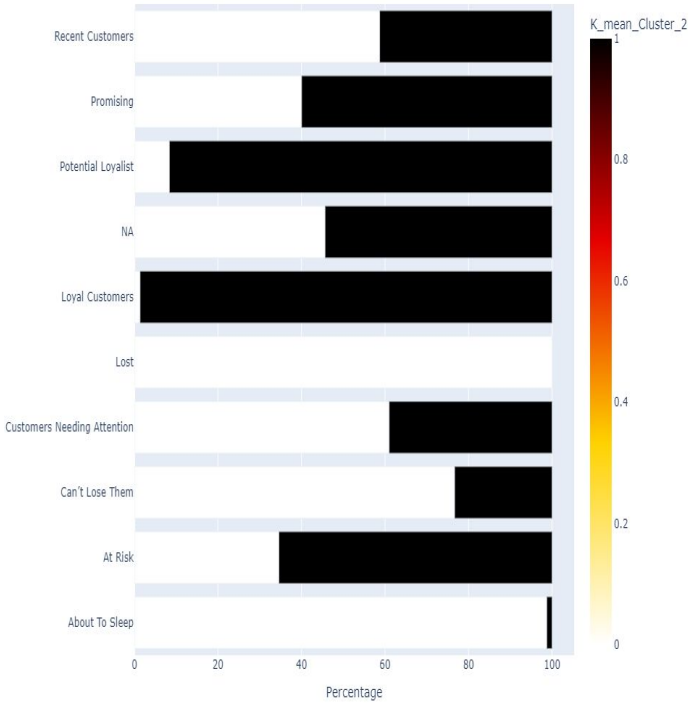
Clustering Analysis: K-mean clustering(K=2)

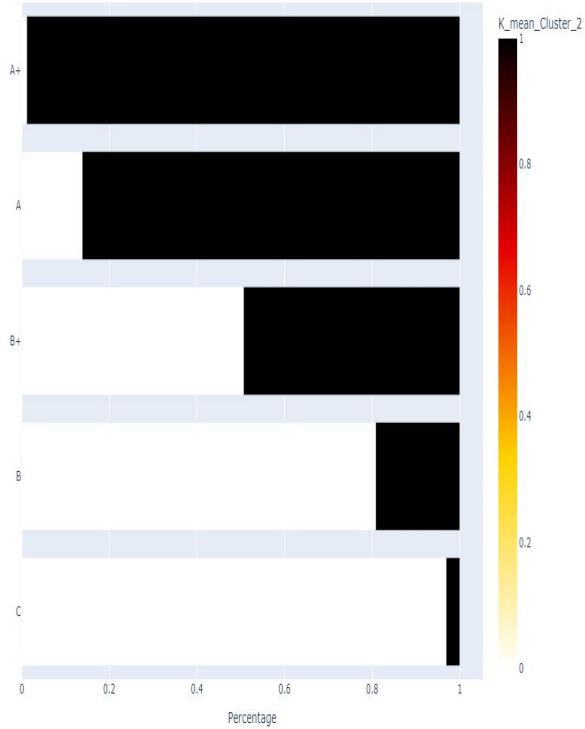
3D visualisation of clusters,



3D visualisation of clusters using the RFM class,





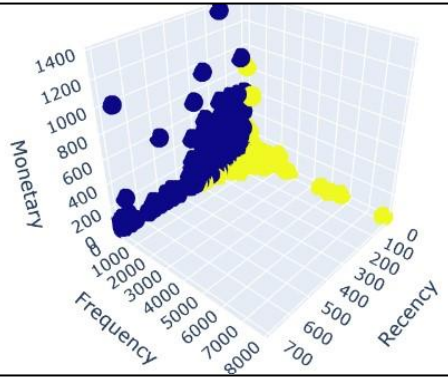


- **K means using K=2 is able to distinguish clearly between 'lost' and 'Loyal Customer' category.**
- **The first cluster contains all customers classified as 'Lost', 76% of customers classified as 'Can't Lose them' and 98% of customers classified as 'About to Sleep'.**
- **The second cluster contains 98% of all customers classified as 'Loyal customer' and 91% of customers classified as 'Potential Loyalist'.**
- **Considering customer grade, the first cluster mainly contains high-graded customers and the second cluster contains mostly low-grade customers.**

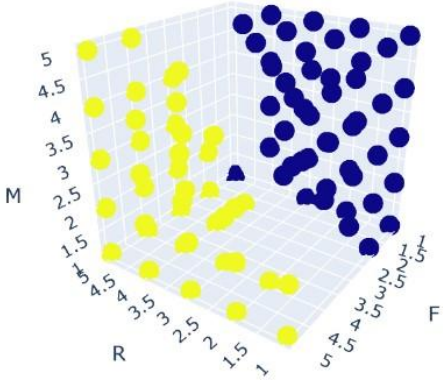
Clustering Analysis: K-mean clustering(K=2)

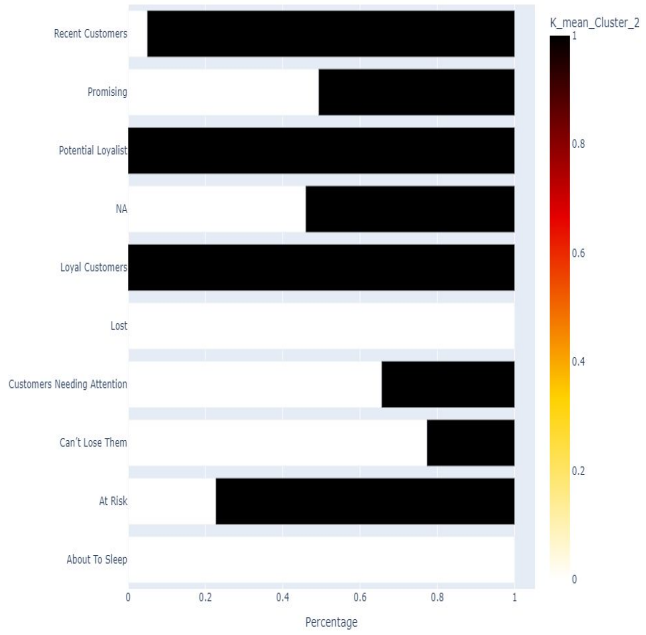
For the subsequent analysis, RFM classes that are distributed between both clusters are dropped. The intention of doing this was to comprehend the clusters a bit better. After dropping those RFM classes, out of 124 RFM classes, 88 RFM classes are present only in one cluster

3D visualisation of clusters after RFM class filtering,



3D visualisation of clusters using the RFM class after RFM class filtering,



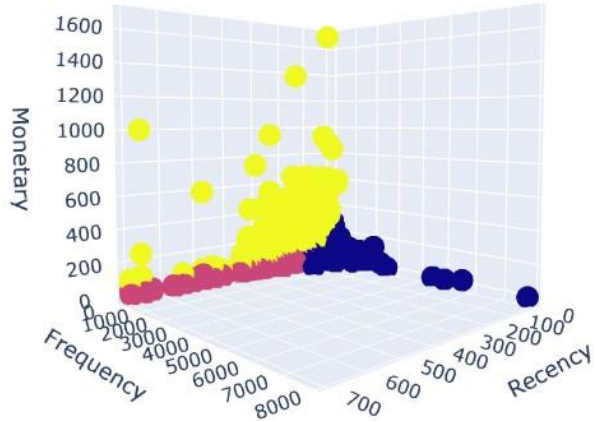


After removing those RFM classes that are distributed between both clusters,

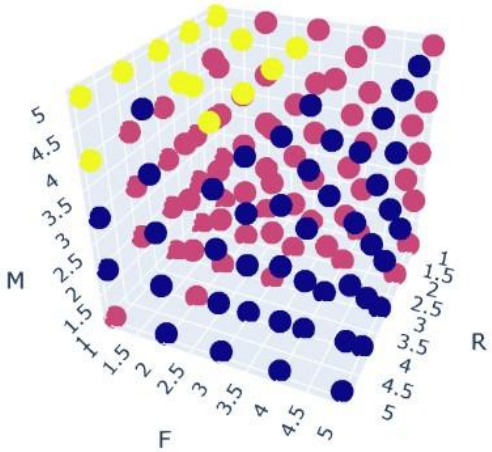
- **The first cluster contains all customers classified as 'Lost' and 'About to Sleep'.**
- **The second cluster contains all customers classified as 'Loyal customer' and 'Potential Loyalist'.**
- **Considering customer grade, the first cluster mainly contains high-graded customers and the second cluster contains mostly low-grade customers.**

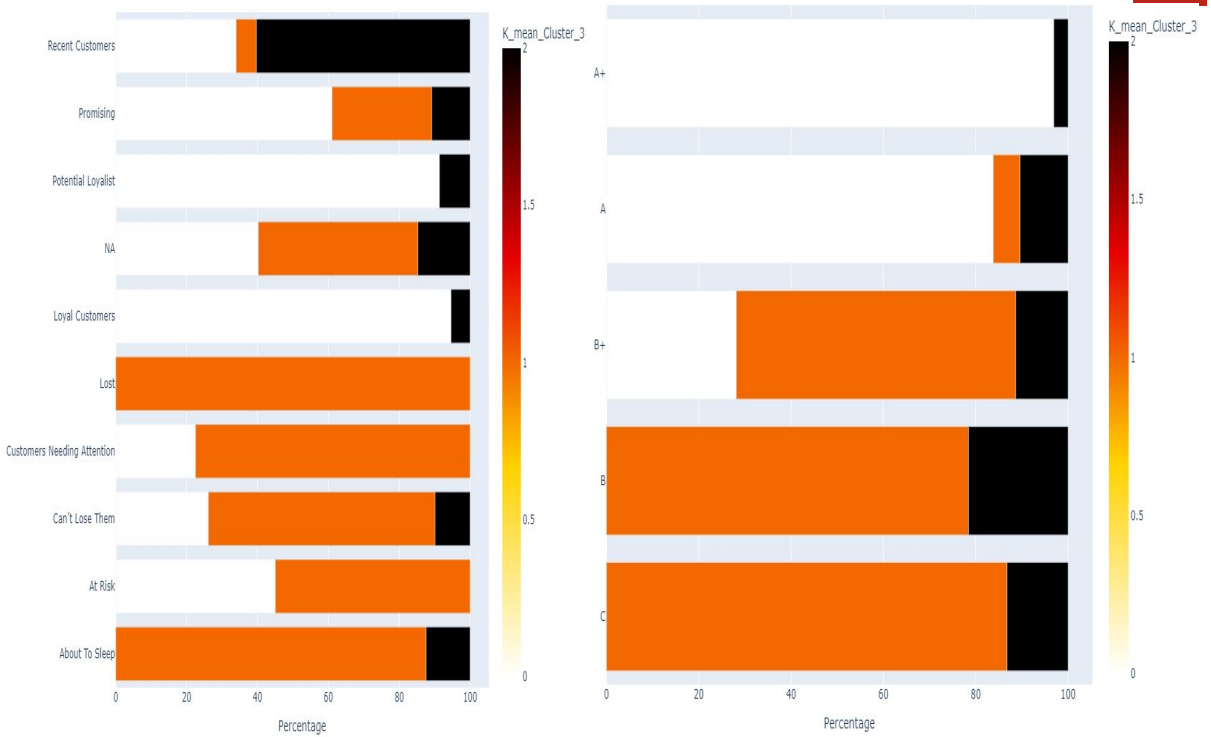
Clustering Analysis: K-mean clustering(K=3)

3D visualisation of clusters,



3D visualisation of clusters using the RFM class,



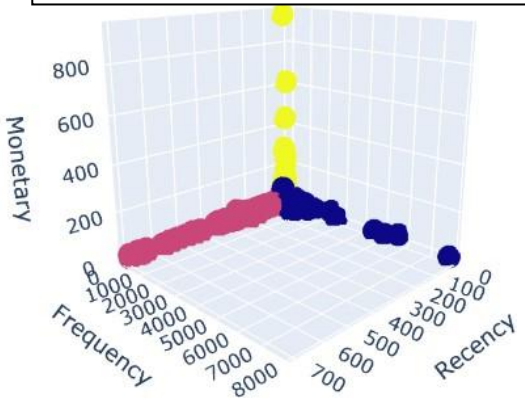


- **The first cluster contains all customers classified as 'Lost' and 87% of customers classified as 'About to Sleep'.**
- **The second cluster contains 94% of customers classified as 'Loyal customer' and 91% of customers classified as 'Potential Loyalist'.**
- **No such major distribution is observed in the third cluster. One of the reasons may be because around 50% of total customers were able to classify based on RFM classes.**
- **Considering customer grade, the first cluster mainly contains high-graded customers and the second cluster contains mostly low-grade customers.**

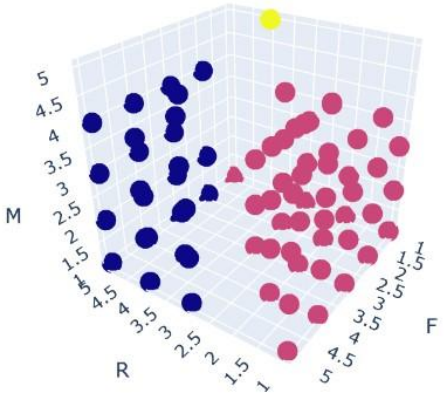
Clustering Analysis: K-mean clustering(K=3)

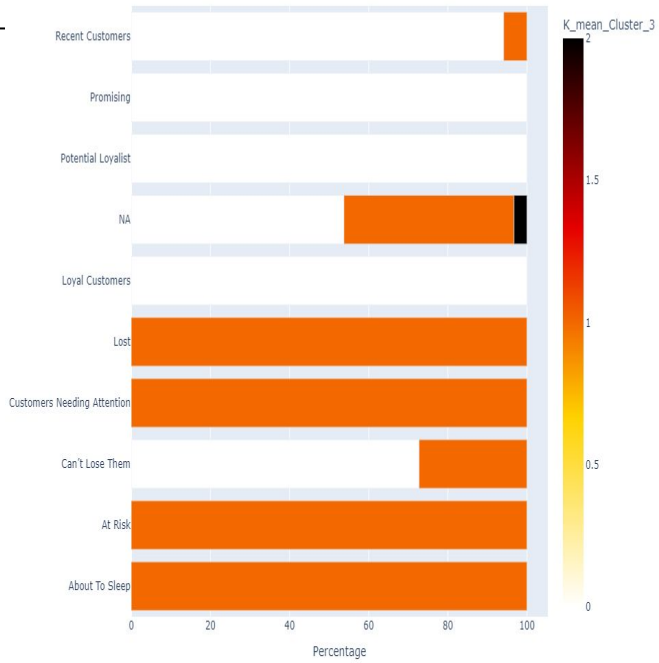
For the subsequent analysis, RFM classes that are distributed between both clusters are dropped. The intention of doing this was to comprehend the clusters a bit better. Out of 124 RFM classes, 74 RFM classes belong only to one cluster

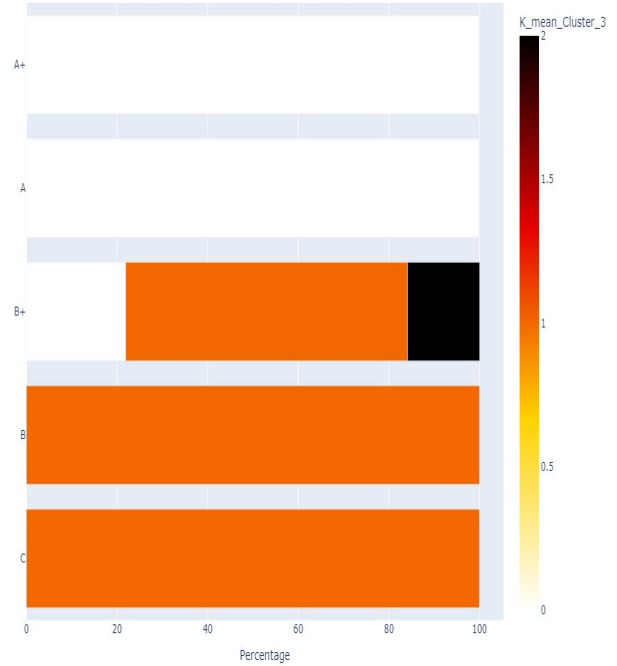
3D visualisation of clusters after RFM class filtering,



3D visualisation of clusters using the RFM class after RFM class filtering,





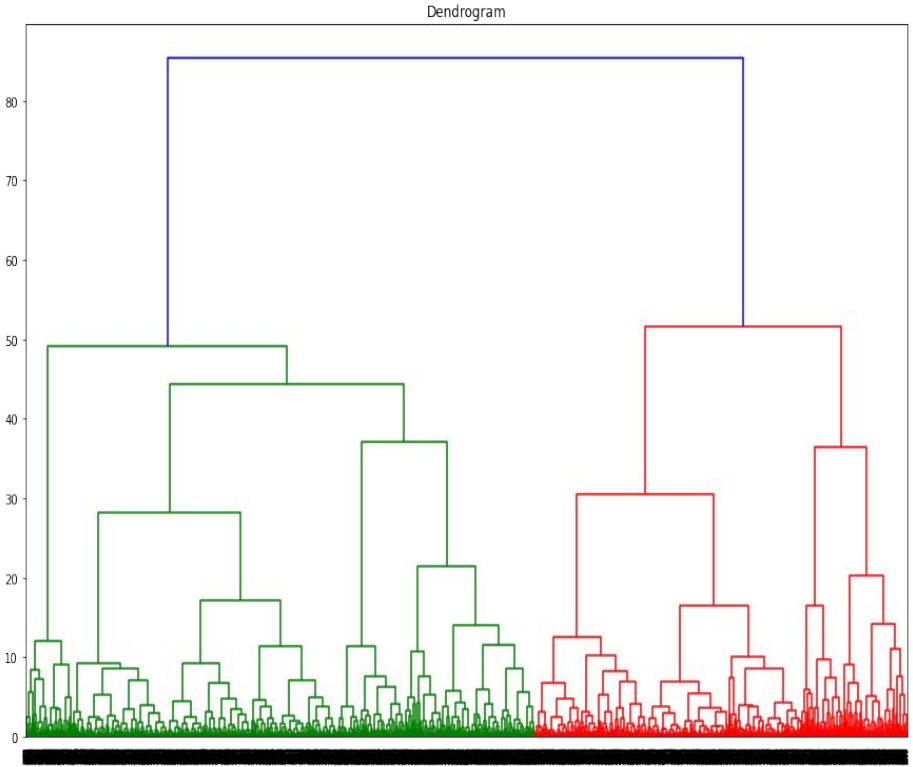


After removing those RFM classes that are distributed between both clusters,

- **The first cluster contains all customers categorized as 'potential Loyalist', 'Loyal Customers' and 'Promising'.**
- **The second cluster contains all customers categorized as 'Lost', 'customer needing attention', 'At Risk' and 'About to sleep'.**
- **Considering customer grade, the first cluster mainly contains high-graded customers and the second cluster contains mostly low-grade customers.**

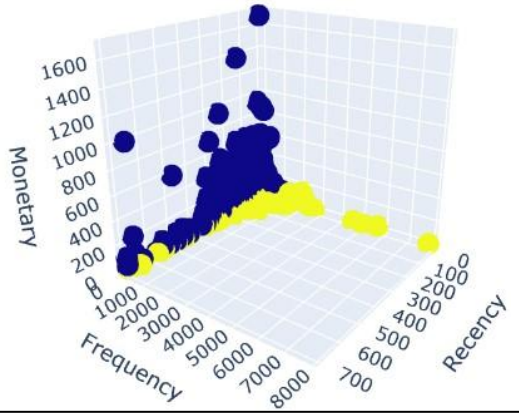
Clustering Analysis: Hierarchical Clustering

In hierarchical clustering, the agglomerate clustering method using ‘ward’ as a linkage and 'euclidean' as affinity is used.
The dendrogram method is used in order to get the optimal number of clusters,

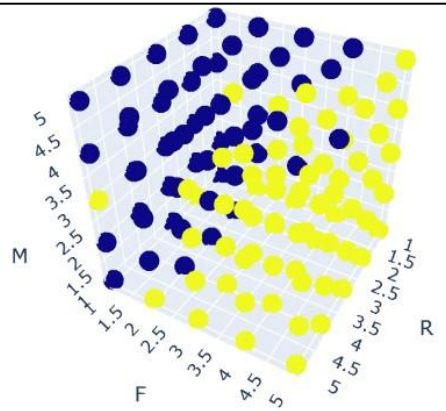


- The Dendrogram suggests that there are two clusters.

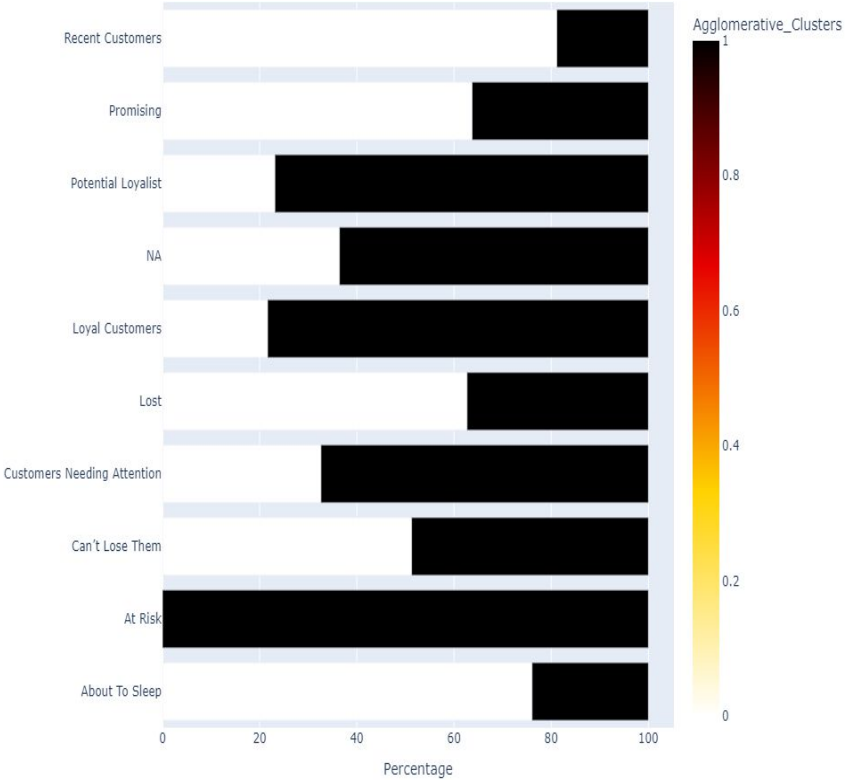
3D visualisation of clusters,

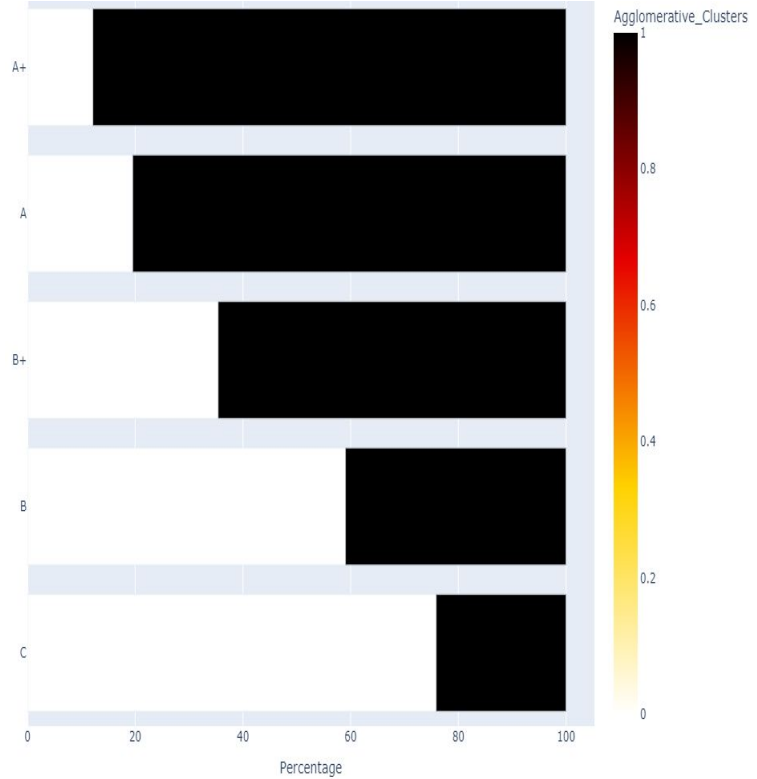


3D visualisation of clusters using the RFM class,



Clustering Analysis: Hierarchical Clustering



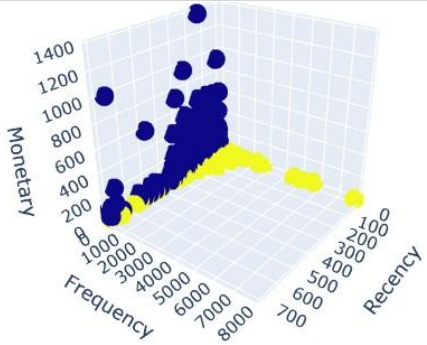


- **The first cluster contains 81% of customers classified as 'Recent customer' and 76% of customers classified as 'About to sleep'.**
- **The second cluster contains all customers classified as 'At Risk' and 78% of customers classified as 'Loyal customer'.**
- **The first cluster contains 87% of customers classified as 'A+' and 80% of customers classified as 'A'.**
- **The second cluster contains 75% of customers classified as 'C'.**

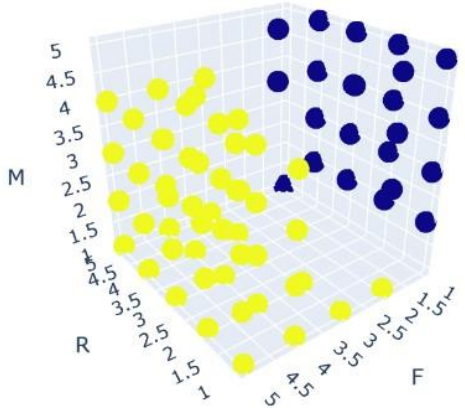
Clustering Analysis: Hierarchical Clustering

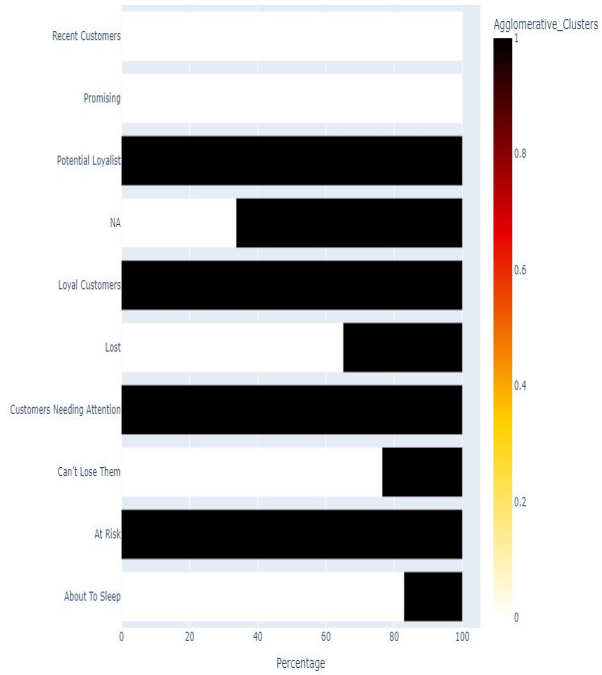
For the subsequent analysis, RFM classes that are distributed between both clusters are dropped. The intention of doing this was to comprehend the clusters a bit better. Out of 124 RFM classes, 73 RFM classes belong only to one cluster

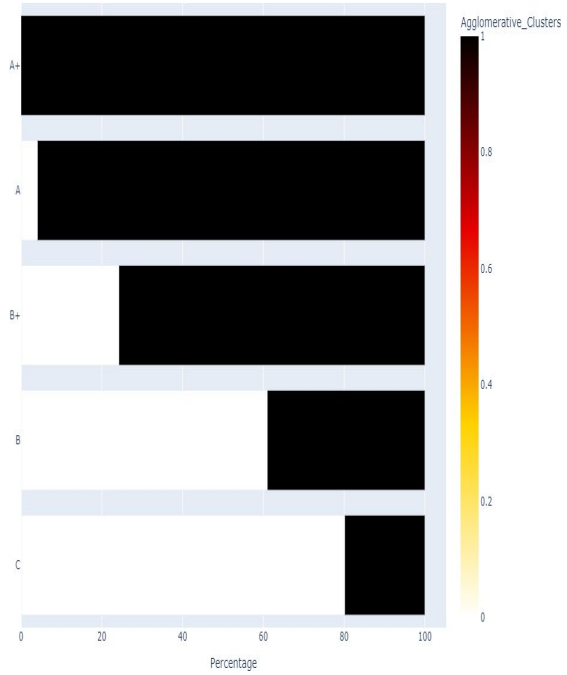
3D visualisation of clusters after RFM class filtering,



3D visualisation of clusters using the RFM class after RFM class filtering,



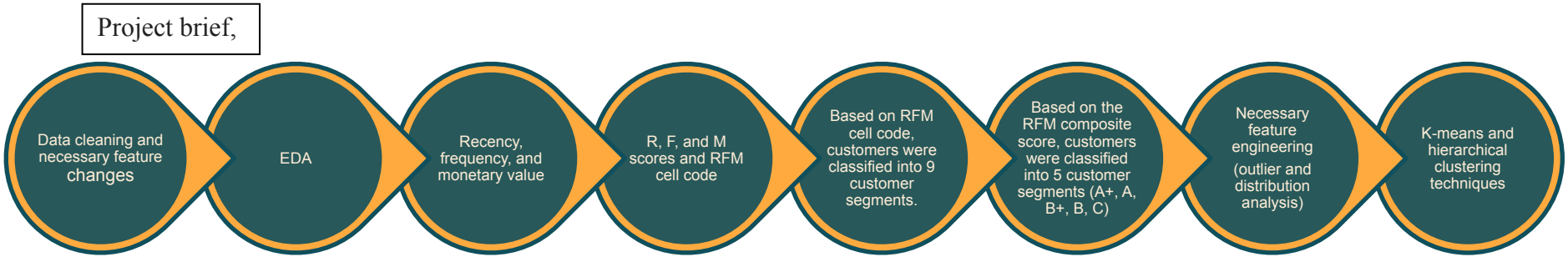




After removing those RFM classes that are distributed between both clusters,

- **The first cluster contains all customers categorized as 'potential Loyalist', 'Loyal Customers' and 'customer needing attention' and 'At Risk.**
- **The second cluster contains 82% of customers categorized as 'About to sleep', 76% of customers categorized as 'Can't Lose Them' and 64% of customers categorized as 'Lost'.**
- **Considering customer grade, the first cluster mainly contains high-graded customers and the second cluster contains mostly low-grade customers.**

Conclusion



Future Scope:

The future scope of this project includes inculcating data related to,

- Demographic (data includes age, sex, marital status, family size, occupation, education level, income, race, nationality and religion).
- Psychographic (psychographic characteristics include personality traits, interests, beliefs, values, attitudes and lifestyles).