Capstone Project

Online Retail Customer Segmentation

Team Members

Saurabh Hase

Aakash Sharma

Problem Statement

In this project, our task is to identify major customer segments on a transnational dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

Data Description

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

Exploratory Data Analysis

Exploratory Data Analysis, EDA for short, is simply a ‘first look at the data’. It forms a critical part of the machine learning workflow and it is at this stage we start to understand the data we are working with and what it contains

Performed steps

1. Importing CSV file and analysing data

Using the pandas library we imported the CSV file to read the dataset. We used the delimiter function inside the read_csv to separate out the columns of the dataset.

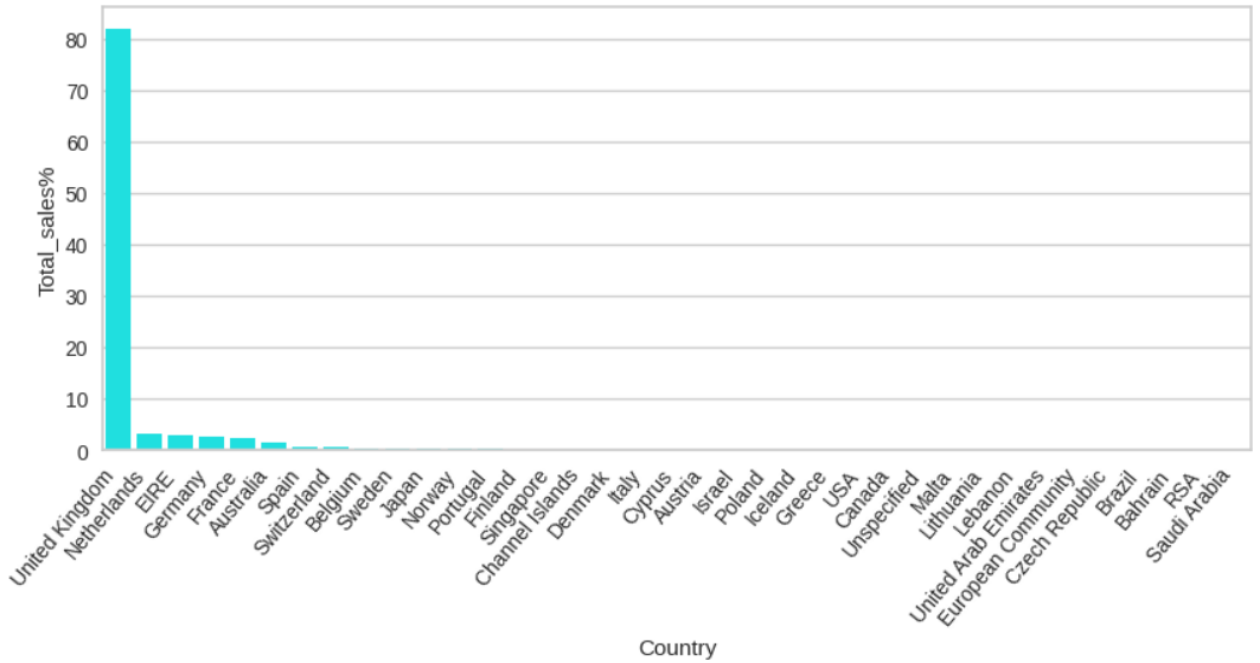
2. Data Cleaning

- Removing Null values from CustomerID and Description column
- Removing of duplicate rows
- Removing of negative values and Zero values from Quantity and Unit Price column
- Removing Cancellation invoices from InvoiceNo column

3. Top 5 items sold in UK

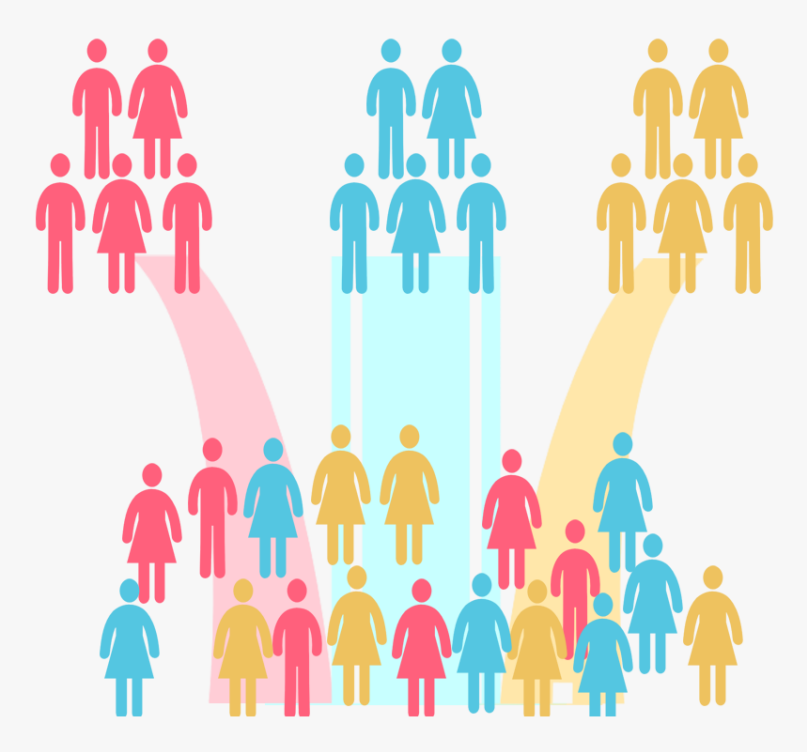
	StockCode	Description	Quantity
2602	23843	PAPER CRAFT , LITTLE BIRDIE	80995
2100	23166	MEDIUM CERAMIC TOP STORAGE JAR	77916
3020	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	54319
3444	85099B	JUMBO BAG RED RETROSPOT	46078
3459	85123A	WHITE HANGING HEART T-LIGHT HOLDER	36706

4. Country-wise customer Distribution



Customer Segmentation

Customer segmentation is the process by which you divide your customers up based on common characteristics - such as demographic or behaviours, so you can market to those customers more effectively.



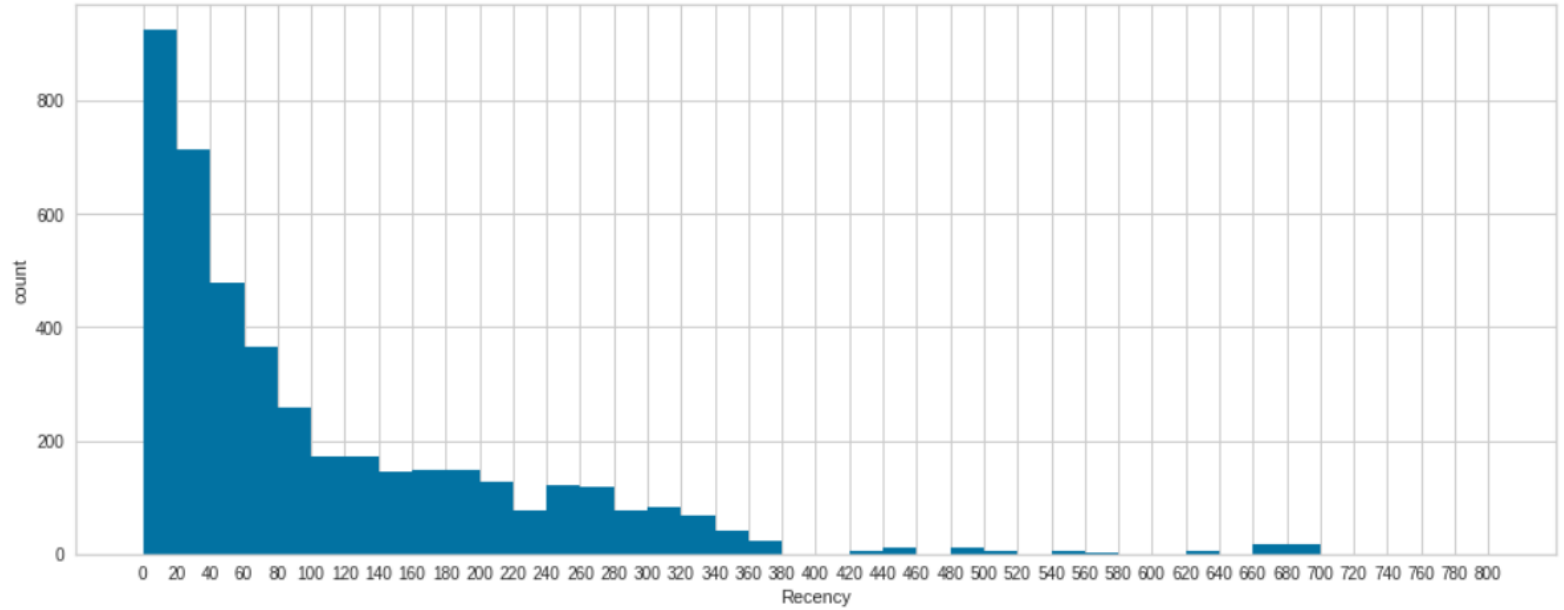
RFM Analysis

RFM analysis is a marketing technique used to quantitatively rank and group customers based on the Recency, Frequency and Monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns.

- **Your best customers** These are the customers who earn top scores in every category
- **Your big spenders:** Customers with top scores for Monetary value
- **Your loyal customers:** Customers with top scores for Frequency.
- **Your faithful customers:** Customers who score high for Frequency but low in Monetary value
- **Your at-risk customers:** Customers who have been in your top tier in the past (best, big spenders and/or loyal) but who now score low for Recency and Frequency.

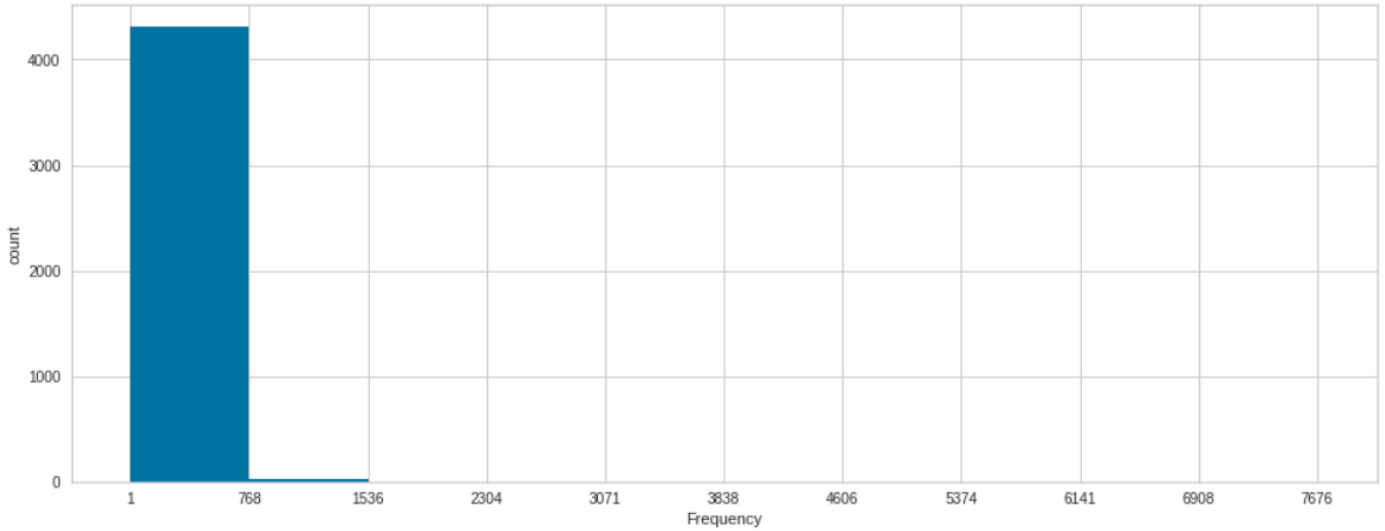
Recency

This refers to the amount of time since a customer's last interaction with a brand, which can include their last purchase, a visit to a website, use of a mobile app, a "like" on social media and more. Recency is a key metric because customers who have interacted with your brand more recently are more likely to respond to new marketing efforts.



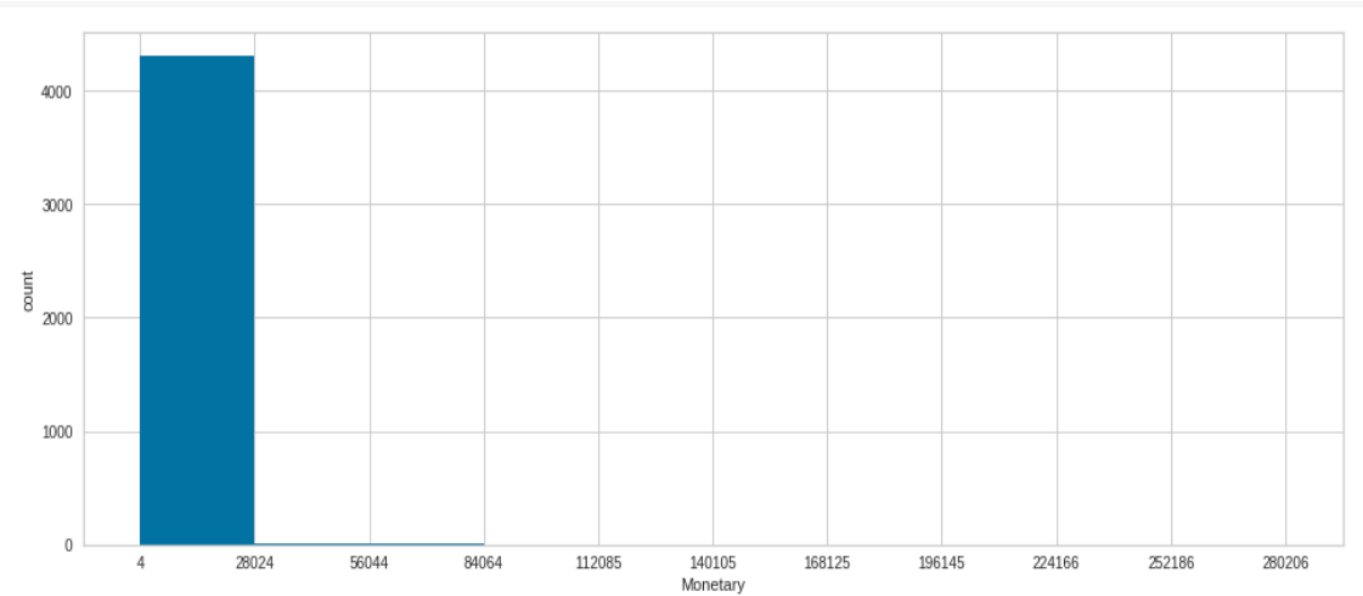
Frequency value

This refers to the number of times a customer has made a purchase or otherwise interacted with your brand during a particular period of time. Frequency is a key metric because it shows how deeply a customer is engaged with your brand



Monetary value

This refers to the total amount a customer has spent purchasing products and services from your brand over a particular period of time. Monetary value is a key metric because the customers who have spent the most in the past are more likely to spend more in the future.



Scaling The Data

Differences in the scales across input variables may increase the difficulty of the problem being modelled so we used scaling method to bring the values in the same scale

Standard scaler

Standardizing a dataset involves rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1.

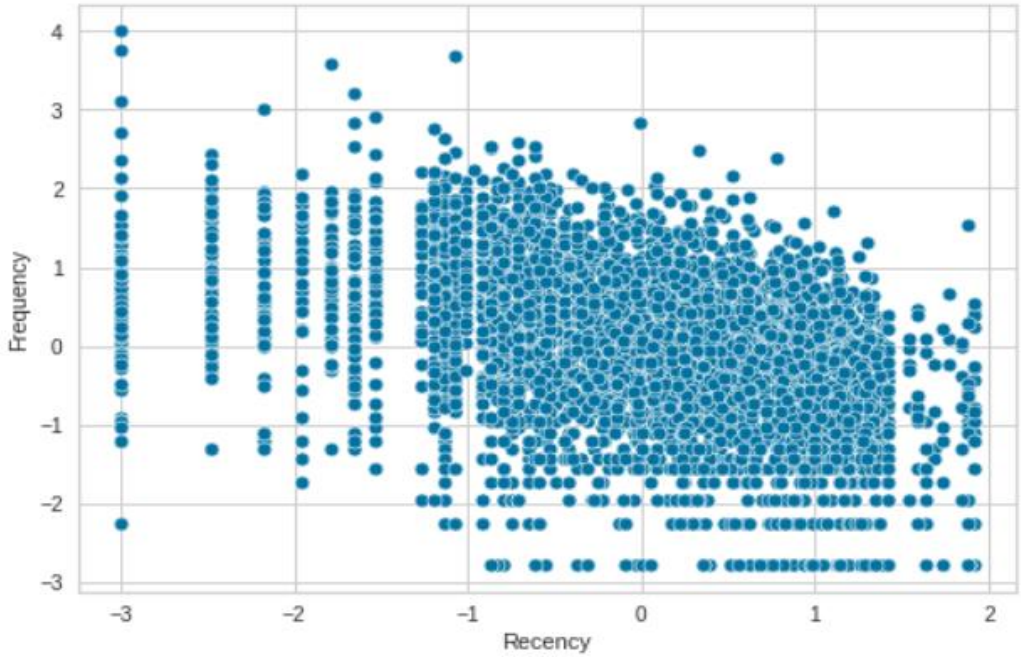
$$y = (x - \text{mean}) / \text{standard\_deviation}$$

$$\text{mean} = \text{sum}(x) / \text{count}(x)$$

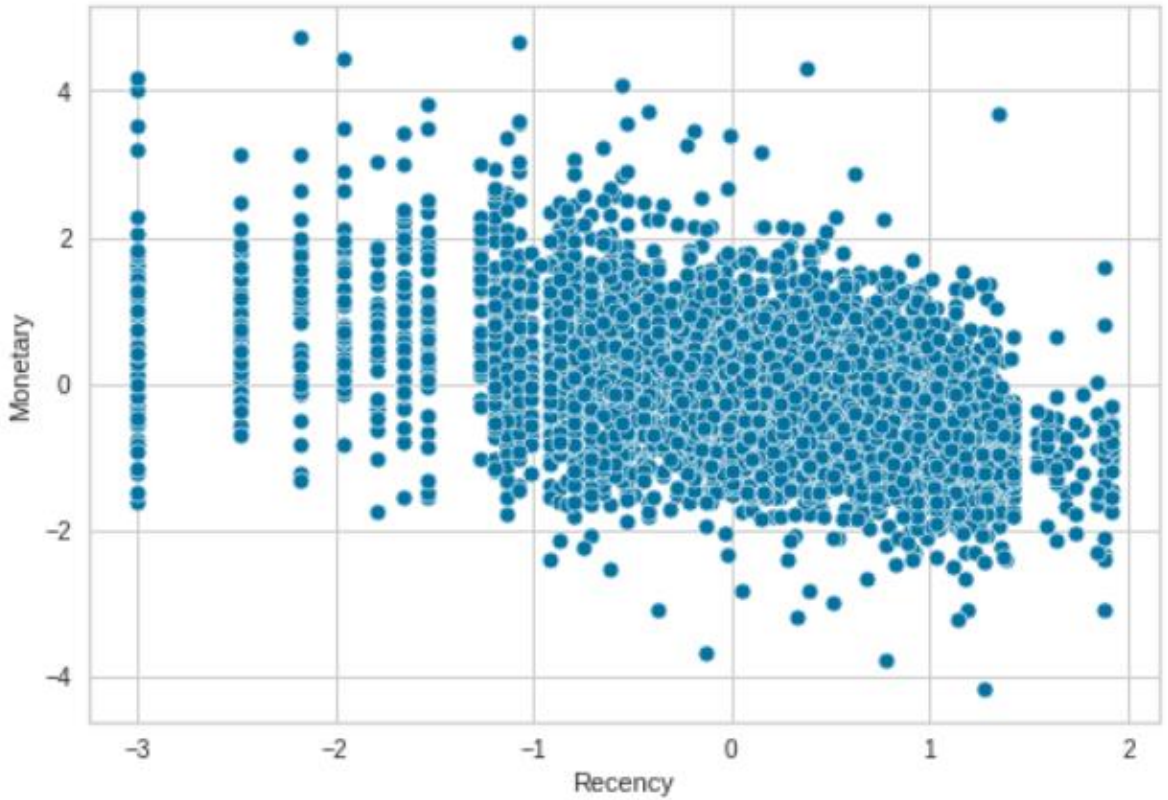
$$\text{standard\_deviation} = \text{sqrt}(\text{sum}((x - \text{mean})^2) / \text{count}(x)$$

Scatter Plot

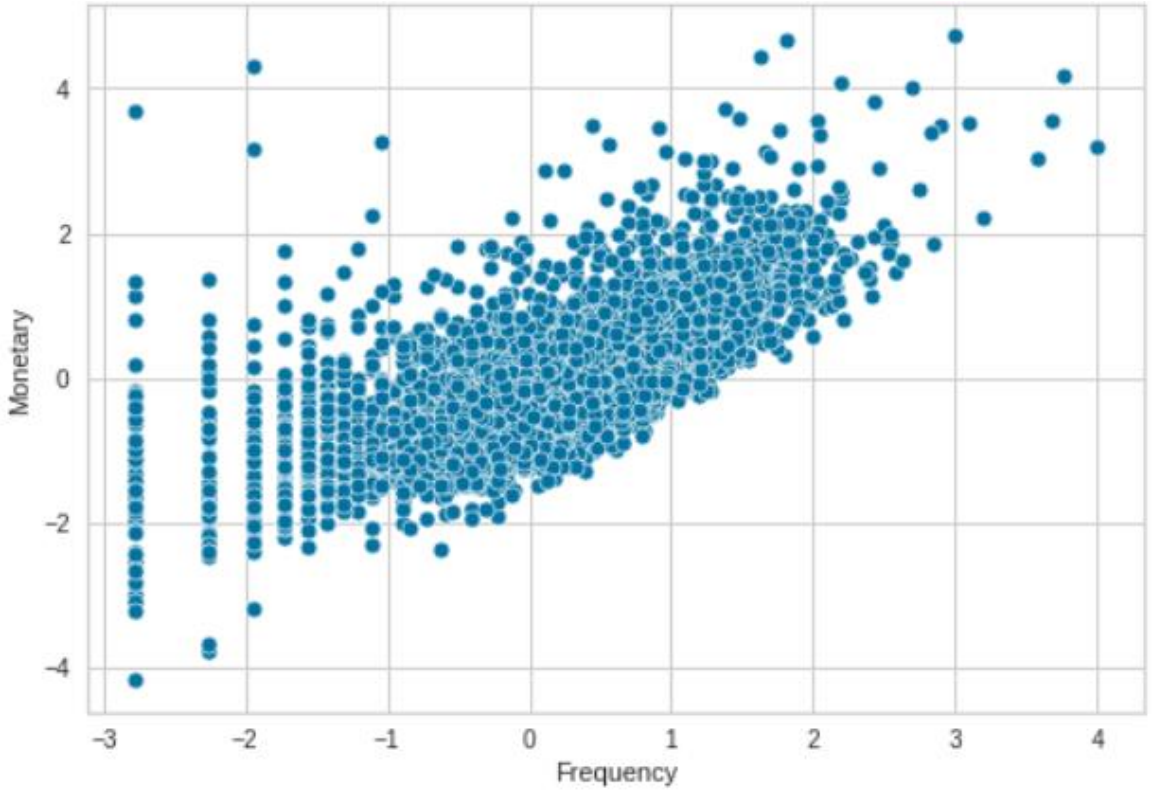
Recency Vs Frequency Scatter plot



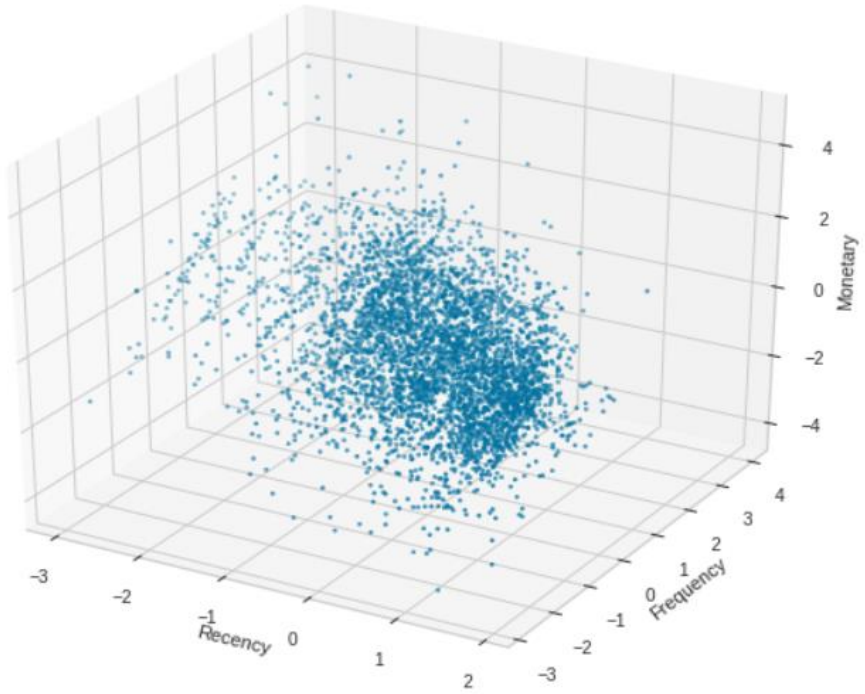
Recency vs Monetary Scatter Plot



Frequency vs Monetary Scatter Plot

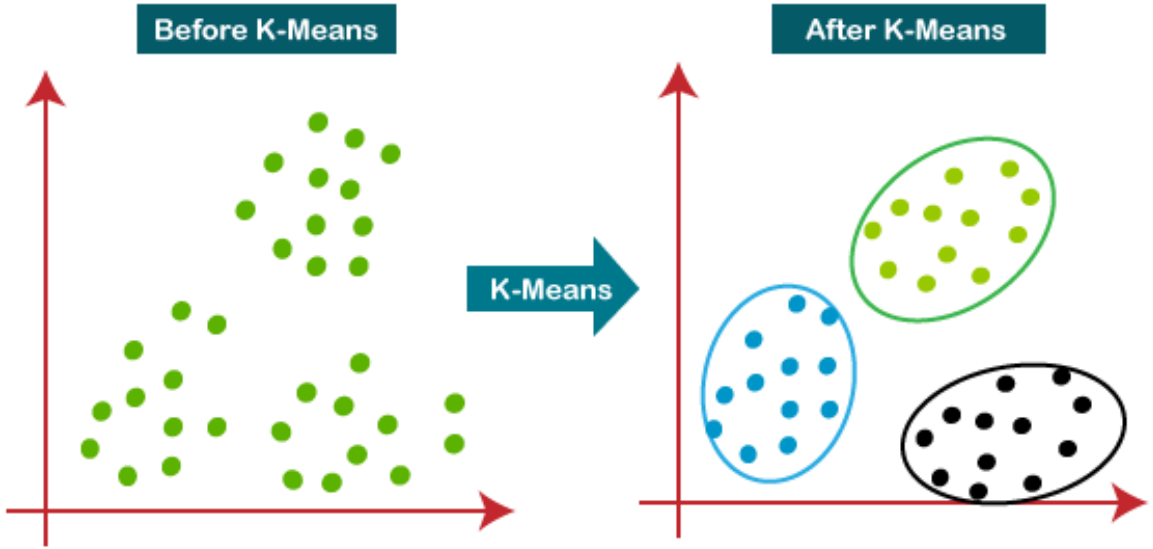


Recency, Frequency and Monetory 3D graph

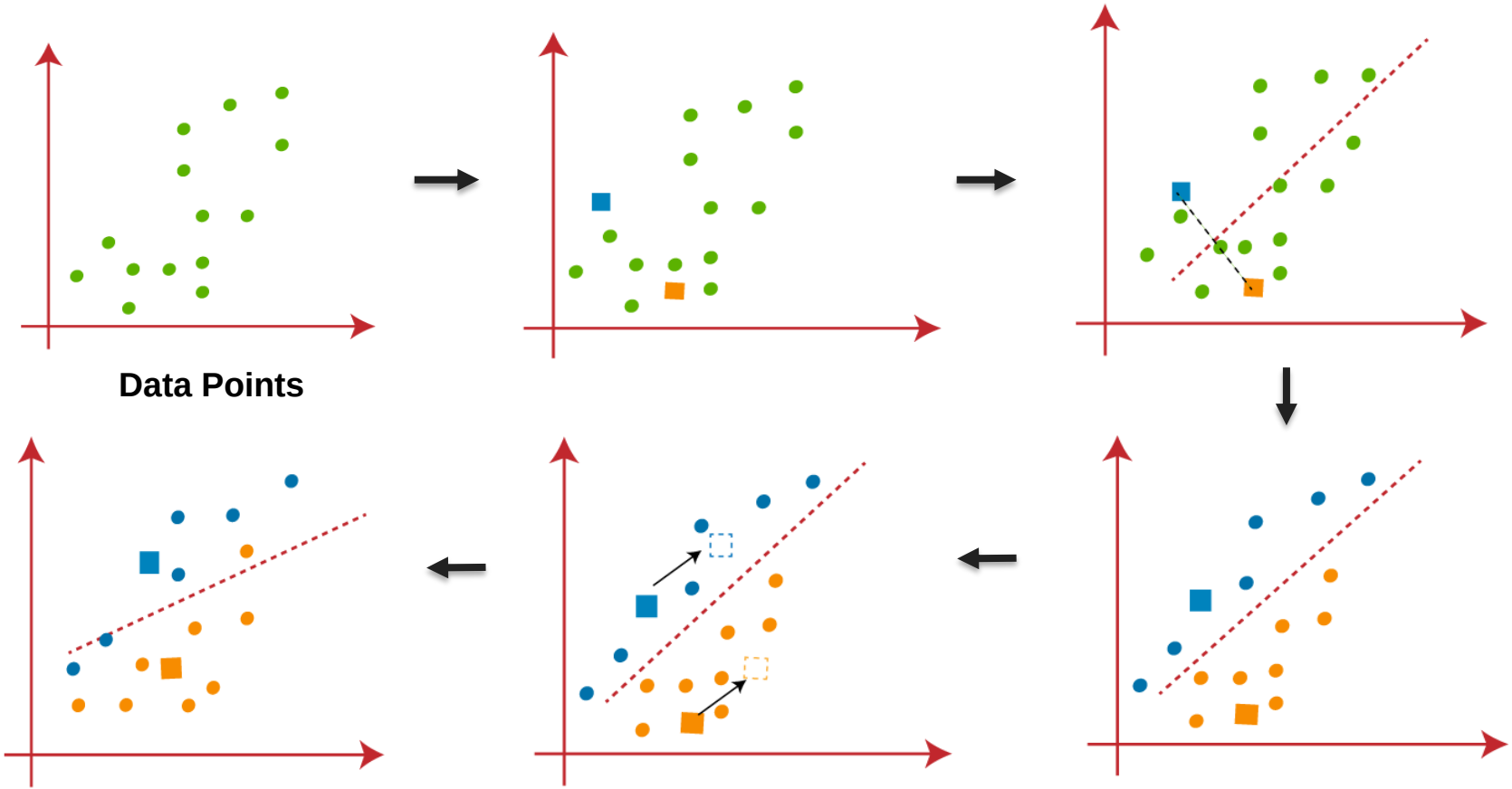


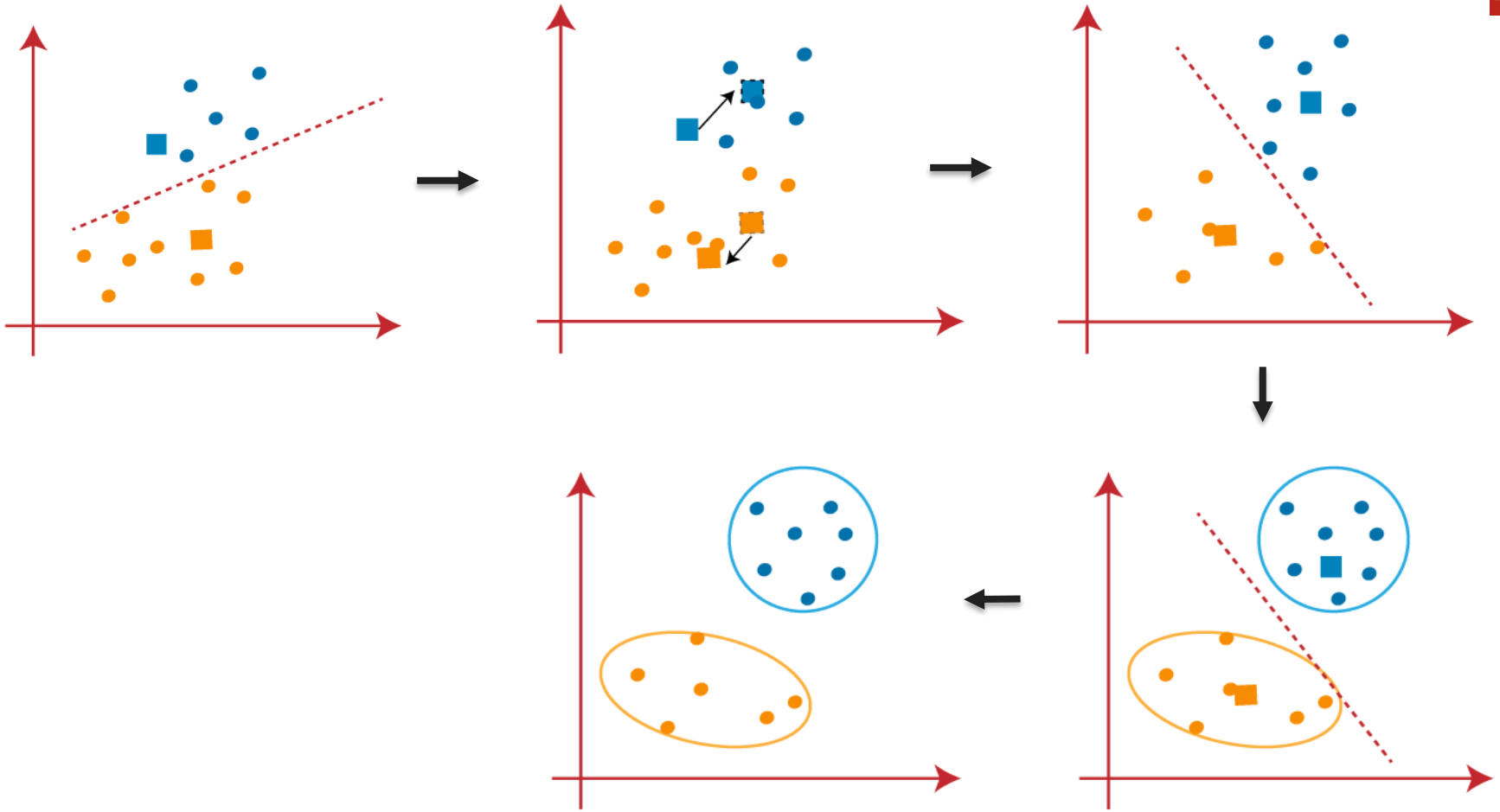
K-Means Clustering Model

K-Means Clustering is an Unsupervised learning iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.



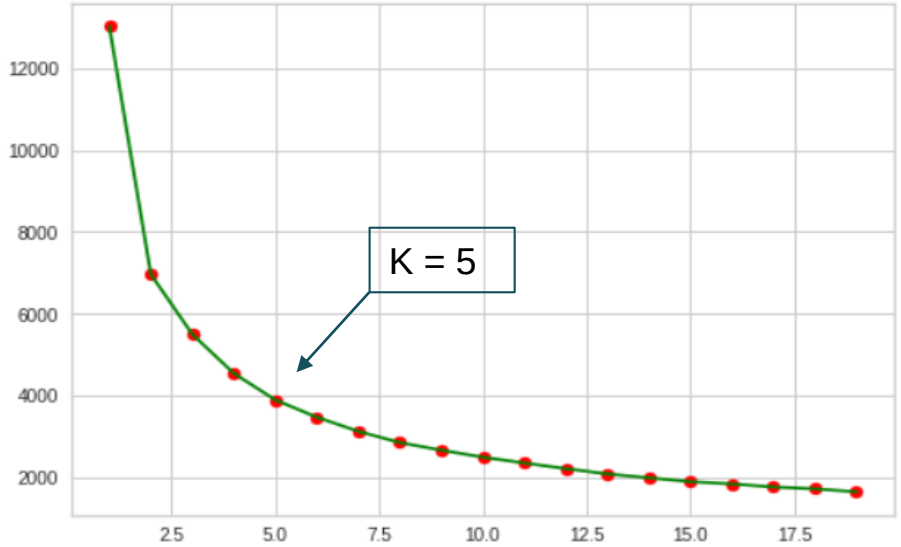
Working of K-Means Algorithm

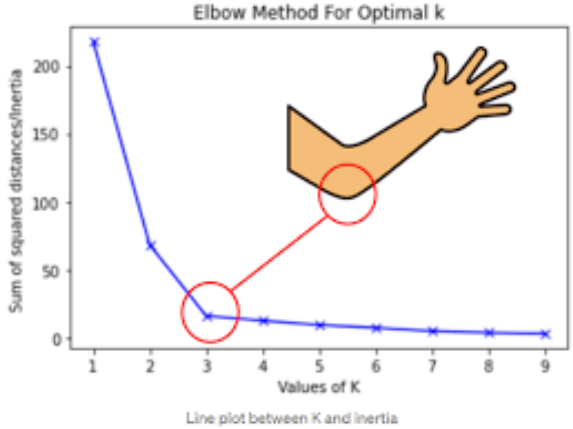



2 Clusters

Elbow Method

In cluster analysis, the **elbow method** is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

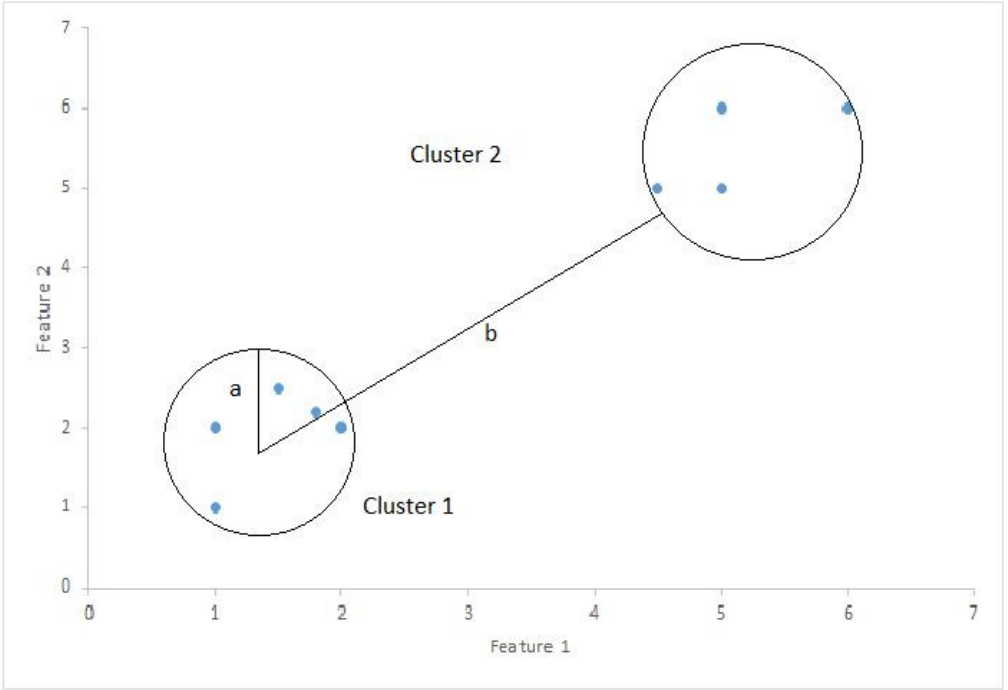




Silhouette Coefficient

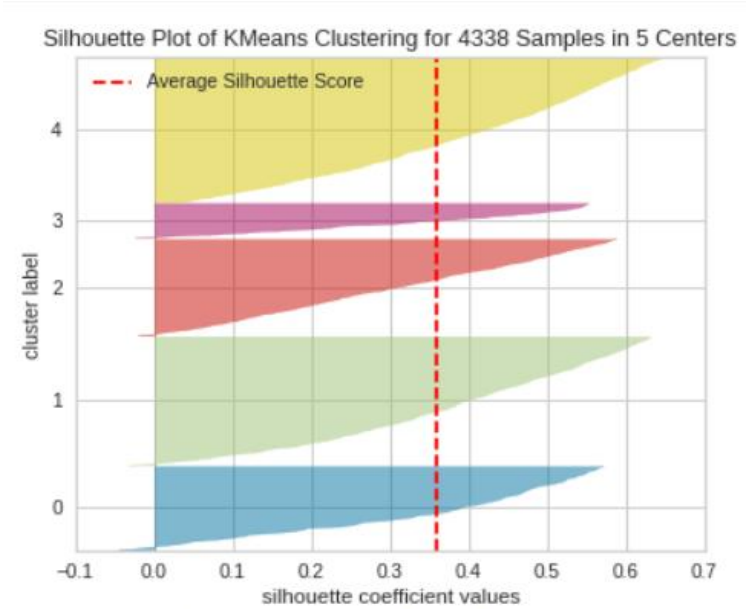
Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

- 1: Means clusters are well apart from each other and clearly distinguished.
- 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- -1: Means clusters are assigned in the wrong way.



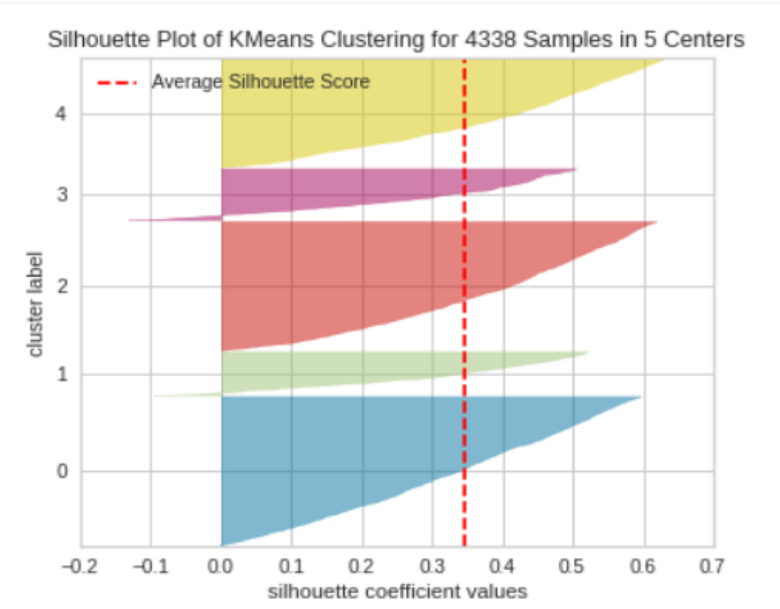
Silhouette Graphs

Recency and frequency Sillhouette Graph



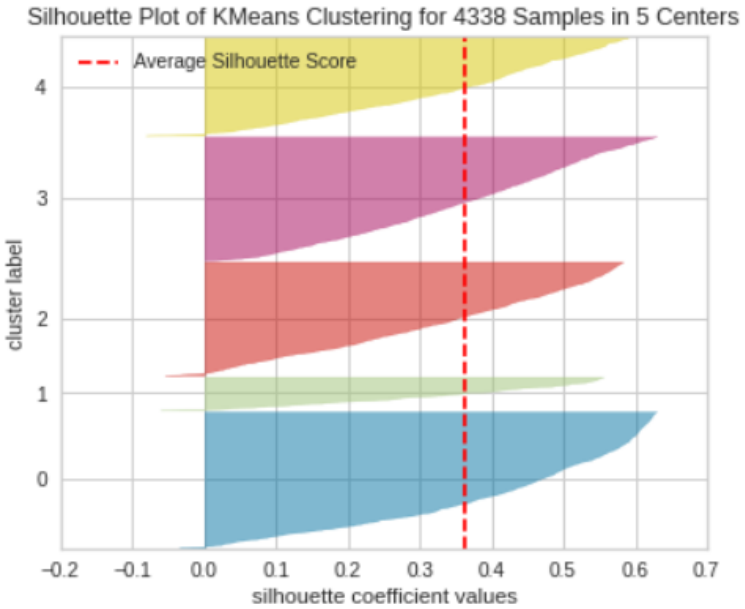
average silhouette score For RF is 0.36

Frequency and Monetary Silhouette Graph



average silhouette score For FM is 0.35

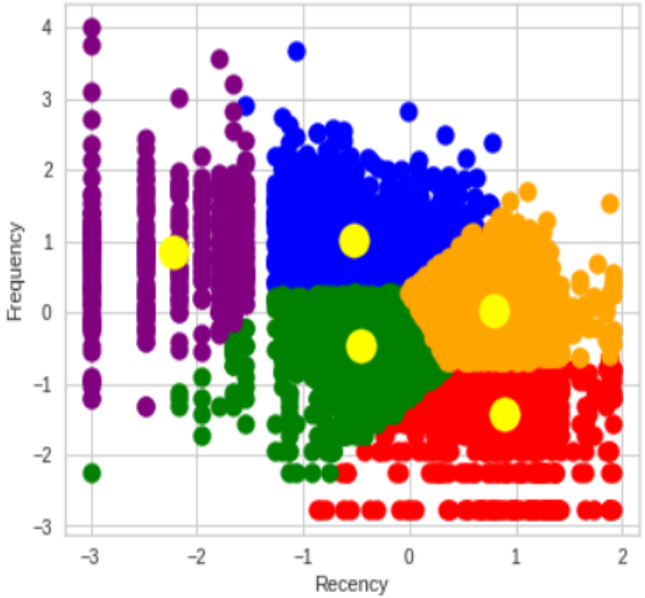
Monetary and Recency Silhouette Graph



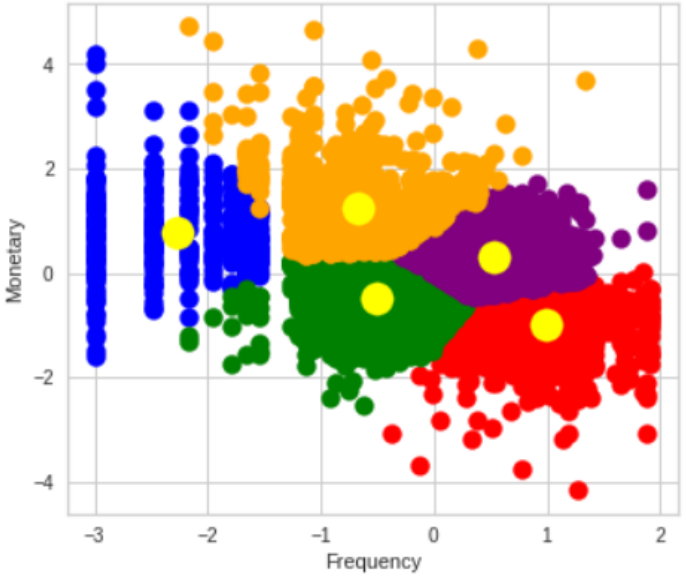
average silhouette score For MR is 0.36

Scatter Plots

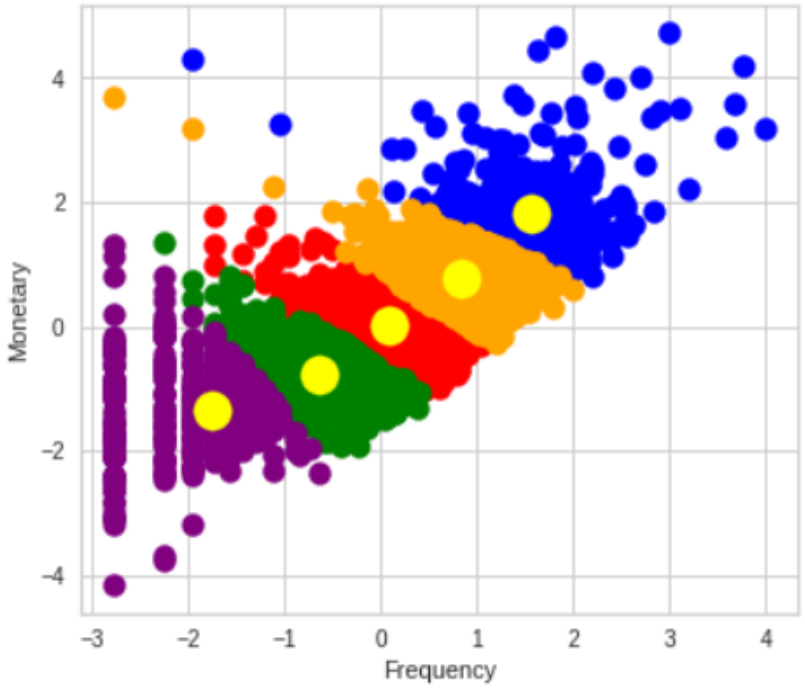
Recency vs Frequency Scatter plot



Frequency and Monetary Scatter Plot

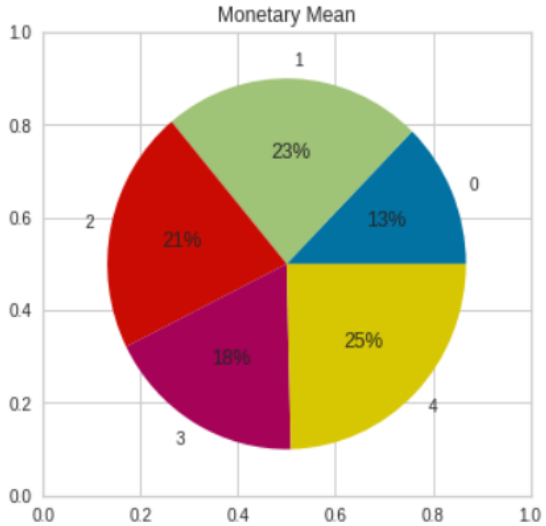


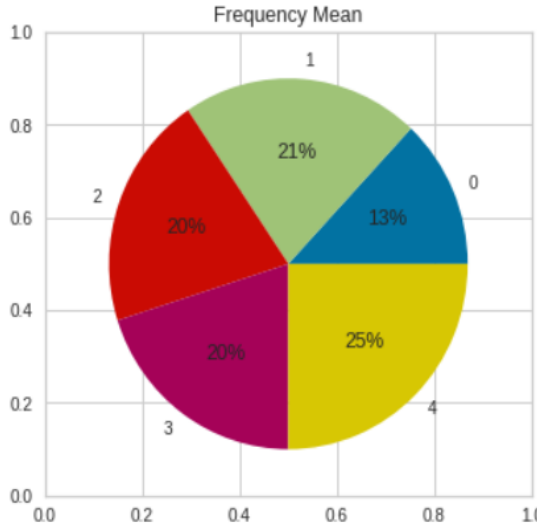
Frequency vs Monetary Scatter Plot

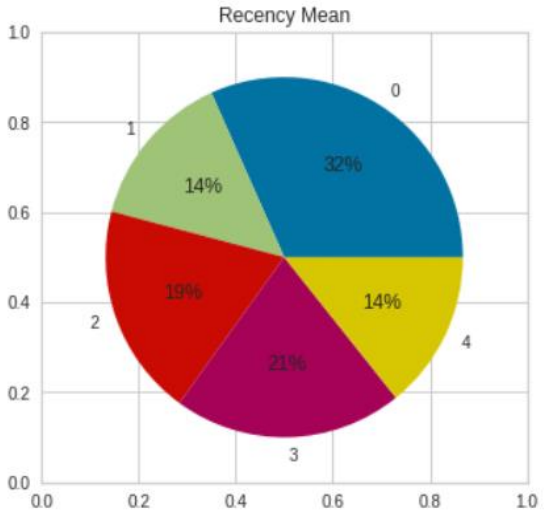


Final Customer Segmentation

After doing Recency, Frequency and Monetary distribution we take mean for each and segmented them in different group from 0 to 4 and ploted it on pie chart so we can make conclusion from it.







Conclusion

- **After competing the experimnt we have reached the following conclusion**

- Most purchases are from UK around 80%
- Top 5 Items sold in UK Are Paper Craft , Little Birdie - Quantity 80995, Medium Ceramic Top Storage Jar – Quantity 77916, World War 2 Gliders Asstd Designs- Quantity 54319, Jumbo Bag Red Retrospot – Quantity 46078, White Hanging Heart T-Light Holder- Quantity 36706
- Mean of Recency = 106.470954
- Mean of Frequency = 90.523744
- Mean of Monetary = 2048.688081
- Best number of cluster found is 5 by using Elbow Method
- Average sihouette score for Recency and Frequency =0.36
- Average sihouette score for Frequency and Monetary=0.35
- Average sihouette score for Monetary and Recency =0.36
- Group 0 has high Recency but Low Frequency and Low Monetary means they come recently but they are not frequent or they do not spend more.
- Group 4 has high Monetary and high Frequency customers spend the most and they are also requent


THANK
YOU!