Capstone Project-4

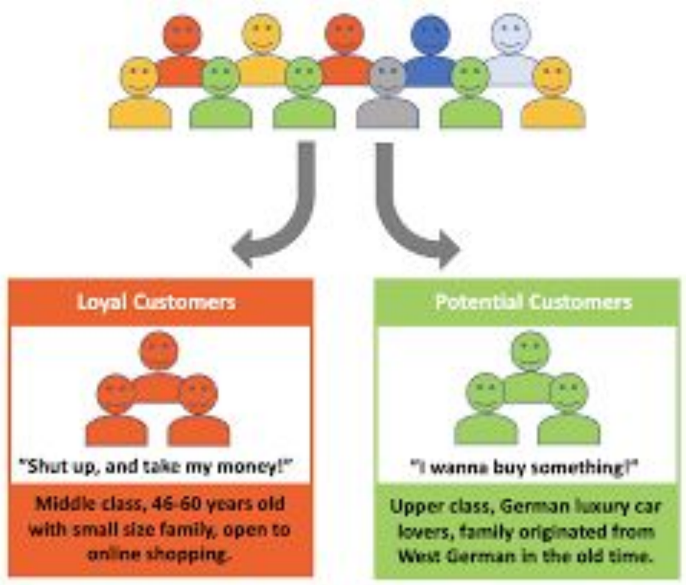
Online Retail Customer Segmentation

Team Members

Mohd Taufique
Sonica Sinha

Points for Discussion

- Problem Statement
- Introduction
- Data Summary
- Data cleaning
- Exploratory Data Analysis
- Feature Engineering
- RFM
- Clustering Analysis
- Summary
- Challenges
- Conclusion



Problem Statement

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

Introduction

- Businesses all over the world are growing every day. With the help of technology, they have access to a wider market and hence, a large customer base.
- Customer segmentation refers to categorizing customers into different groups with similar characteristics.
- Customer segmentation can help businesses focus on each customer group in a different way, in order to maximize benefits for customers as well as the business.
- This project mainly deals in segmenting customers of an online business store in the UK.

Data Summary

- A transnational data set with transactions occurring between 1st December 2010 and 9th December 2011 for a UK-based online retailer.
- Shape (rows- 541909, columns-8).
- The company mainly sells unique all-occasion gifts.
- Many customers of the company are wholesalers.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Attribute Summary

We are given the following columns in our data:

- **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**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **Unit Price**: 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

Data Cleaning

- In this dataset , we have null values present in the 'CustomerID' and 'Description' column. These have to be dropped as there is no way of filling them strategically.
- Cancelled orders exist in the data, these too have been removed.
- Date, month and year were extracted from the 'InvoiceDate' column.

Unique Values present in our dataset

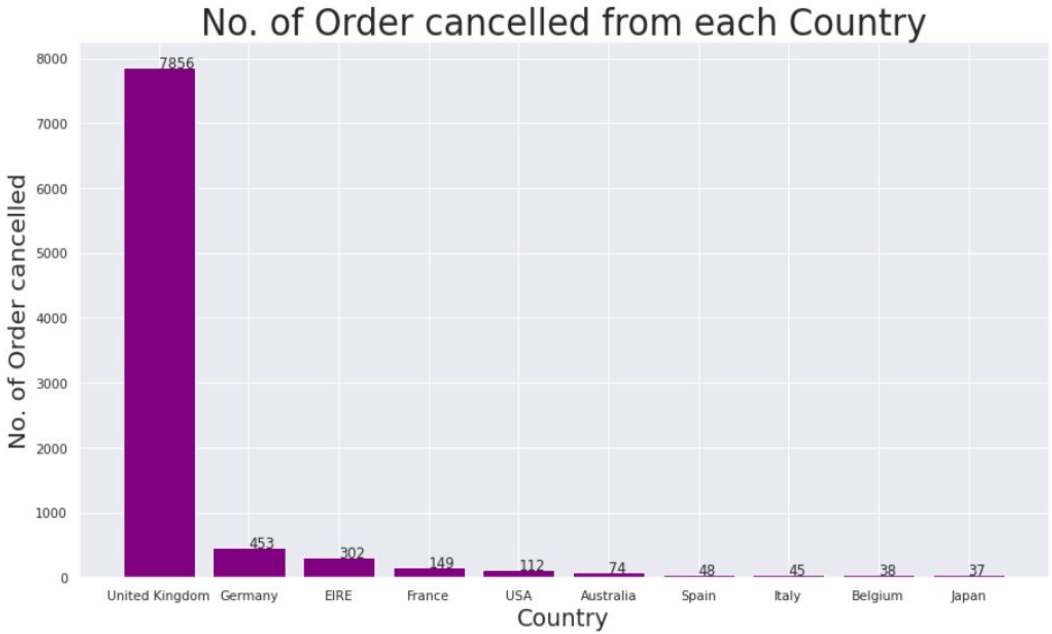
```
Total Unique Values in InvoiceNo - 25900
Total Unique Values in StockCode - 4070
Total Unique Values in Description - 4224
Total Unique Values in Quantity - 722
Total Unique Values in InvoiceDate - 23260
Total Unique Values in UnitPrice - 1630
Total Unique Values in CustomerID - 4373
Total Unique Values in Country - 38
```

Null Values present in our dataset

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

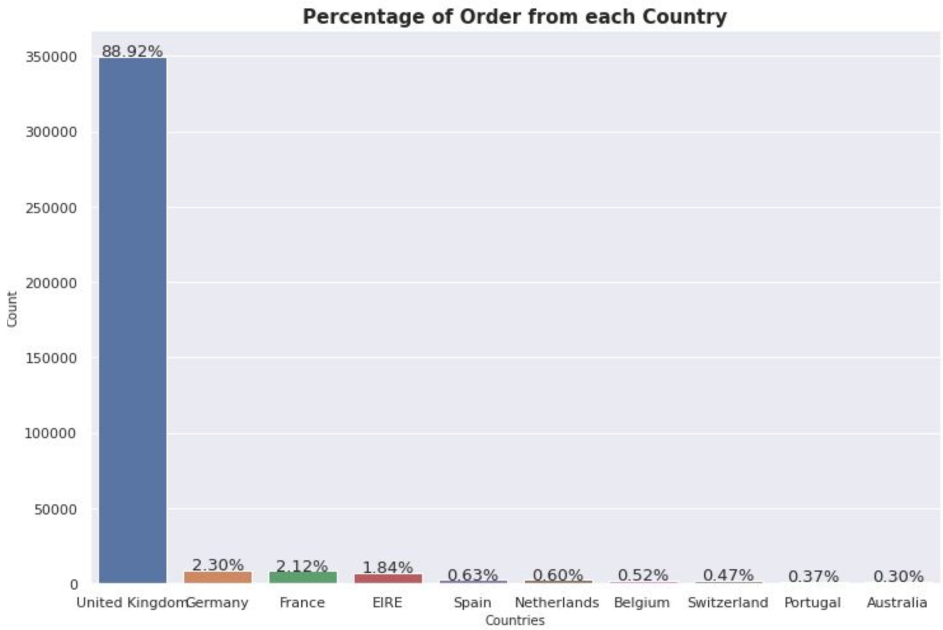
Checking InvoiceNo, how many order got cancelled ?

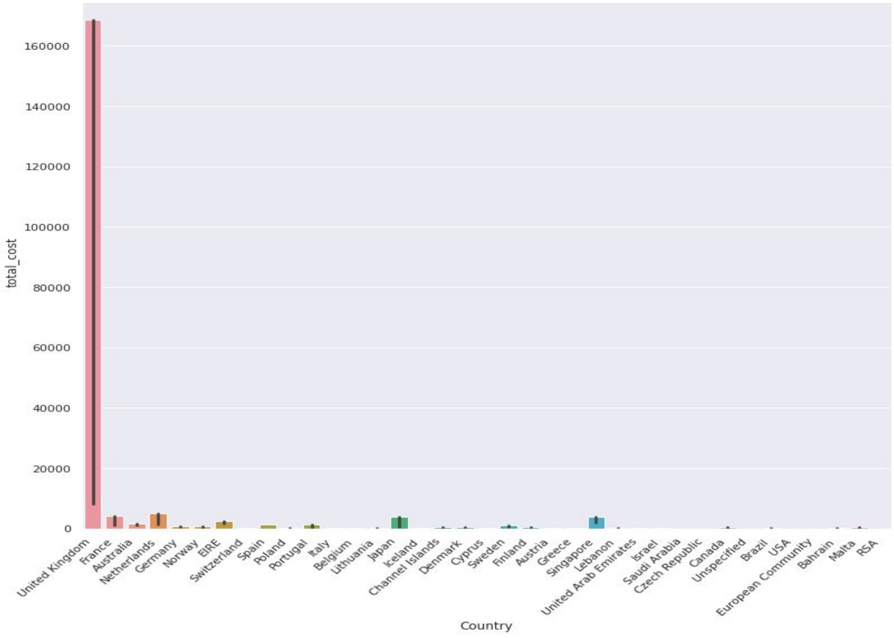
- As we can see most number of orders were cancelled from United Kingdom.
- Total number of orders cancelled were 9288 which equals to the **35.86%**.
- And also we analyzed that the average number of orders per customer was **5**.




```
We have  9288  cancelled orders.
Percentage of orders canceled: 9288/25900 (35.86%)
```

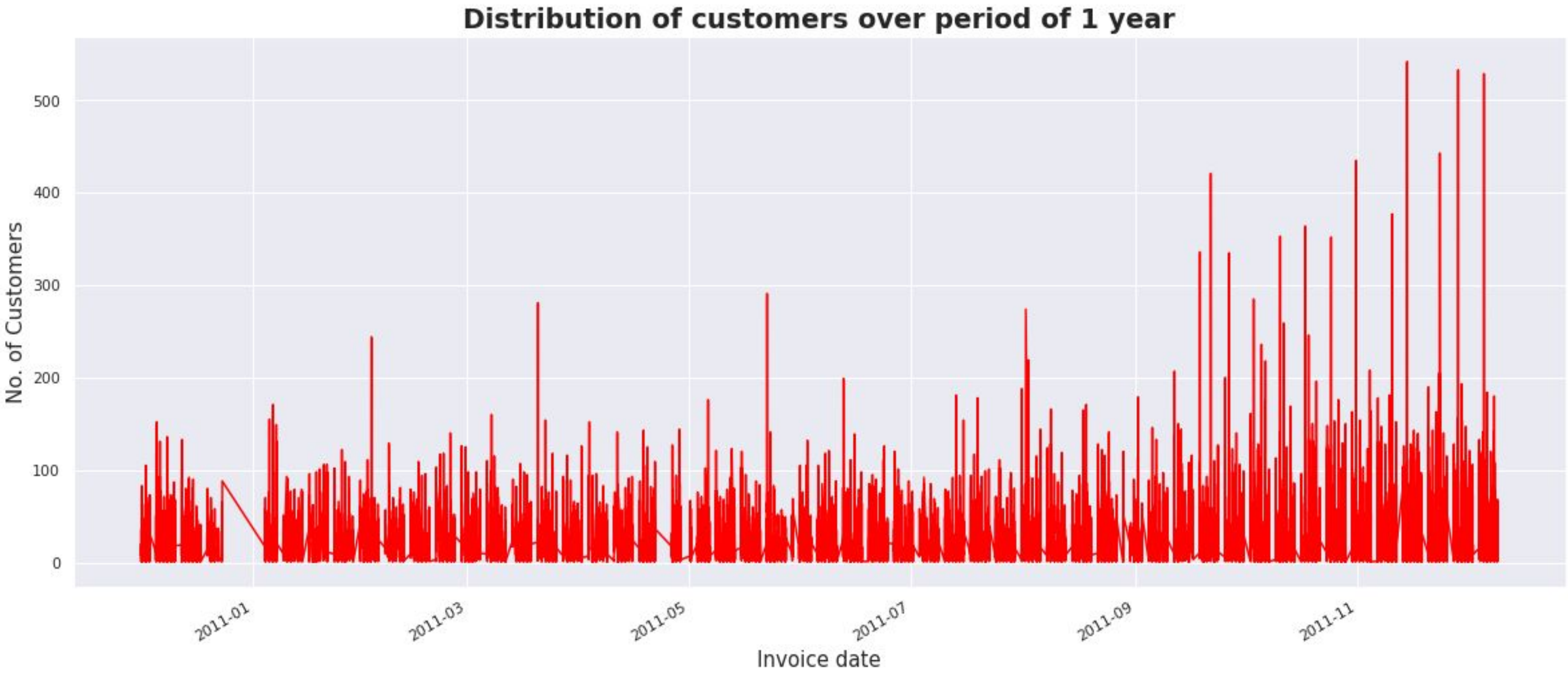
Customers by Country





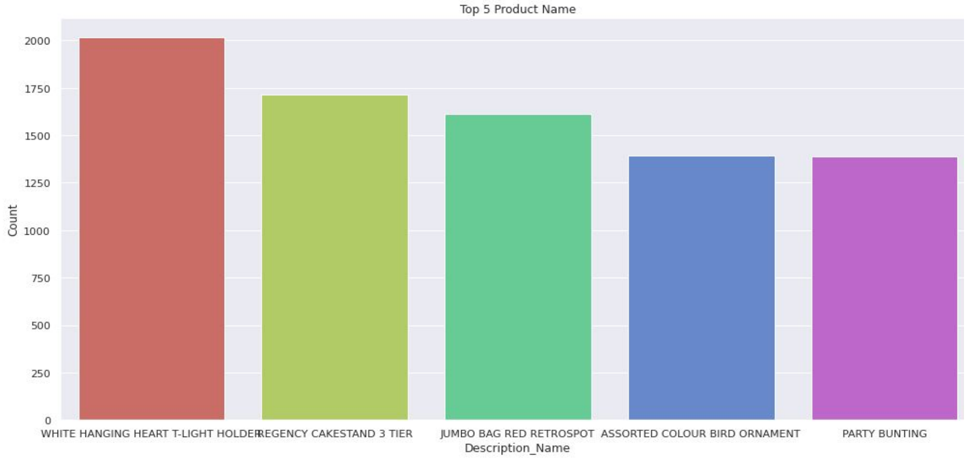
- In first count plot we can clearly see the maximum percentage of order has been placed from **UK** which is **88.92%** out of total 37 countries world wide.
- In another bar plot, we can see that from revenue point also **UK** showed maximum revenue generation.

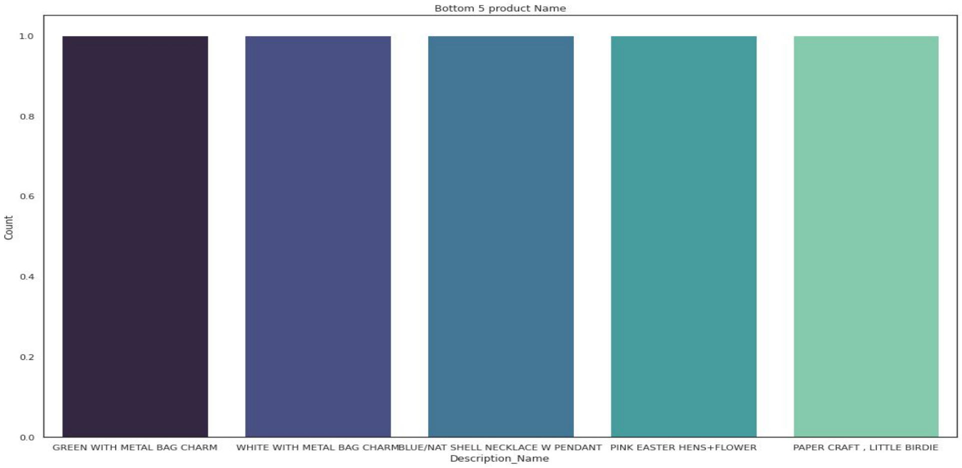
Distribution of Customers Over Period of 1 Year



Product purchased by Customers

- Count plot showing the top 5 product in which - **White Hanging Heart T-Light Holder** is most selling product.
- And the bottom 5 product, in which **Paper Craft , Little Birdie** is the least.





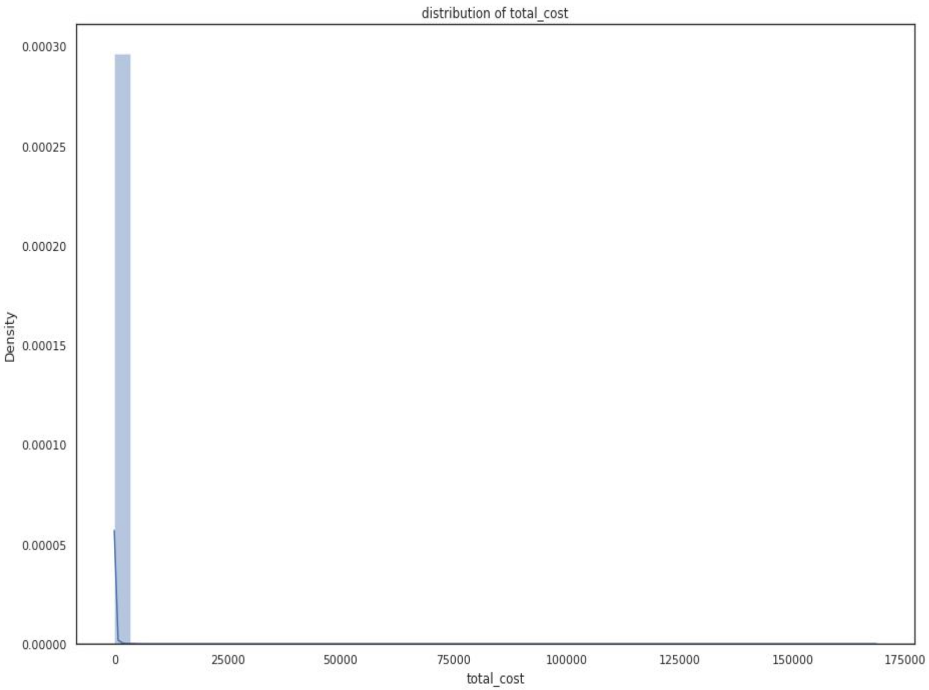
FEATURE ENGINEERING

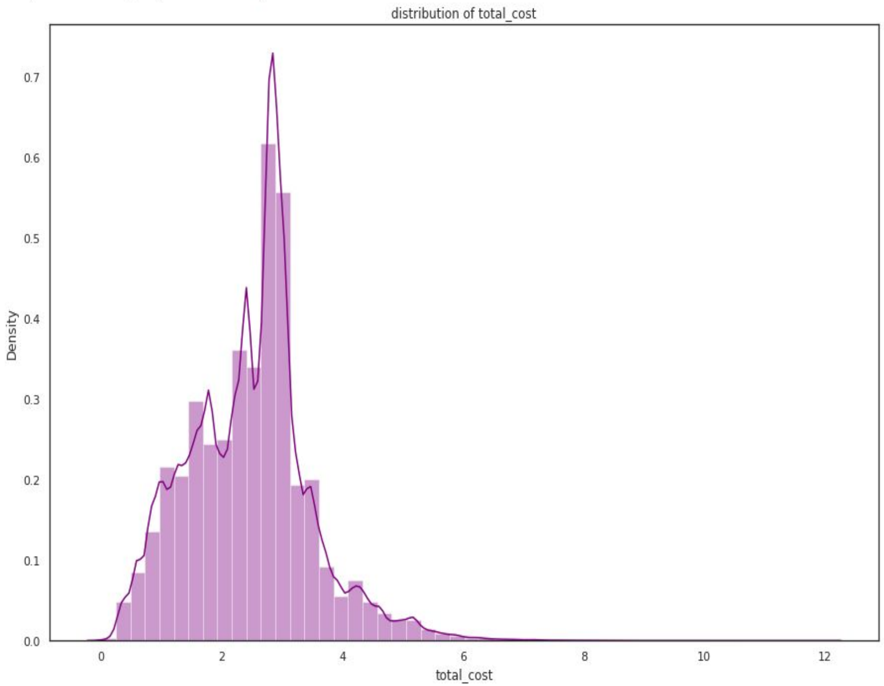
```
[ ]  # Creating new feature Day from Invoicedate
     retail_df_copy['Day']=retail_df_copy['InvoiceDate'].dt.day_name()
```

```
[ ]  # Creating some new features from Invoicedate like hours,year,month_num,day_num, month_name
     retail_df_copy["year"] = retail_df_copy["InvoiceDate"].apply(lambda x: x.year)
     retail_df_copy["month_num"] = retail_df_copy["InvoiceDate"].apply(lambda x: x.month)
     retail_df_copy["day_num"] = retail_df_copy["InvoiceDate"].apply(lambda x: x.day)
     retail_df_copy["hour"] = retail_df_copy["InvoiceDate"].apply(lambda x: x.hour)
     retail_df_copy["minute"] = retail_df_copy["InvoiceDate"].apply(lambda x: x.minute)
```

```
[ ]  retail_df_copy['Month']=retail_df_copy['InvoiceDate'].dt.month_name()
```

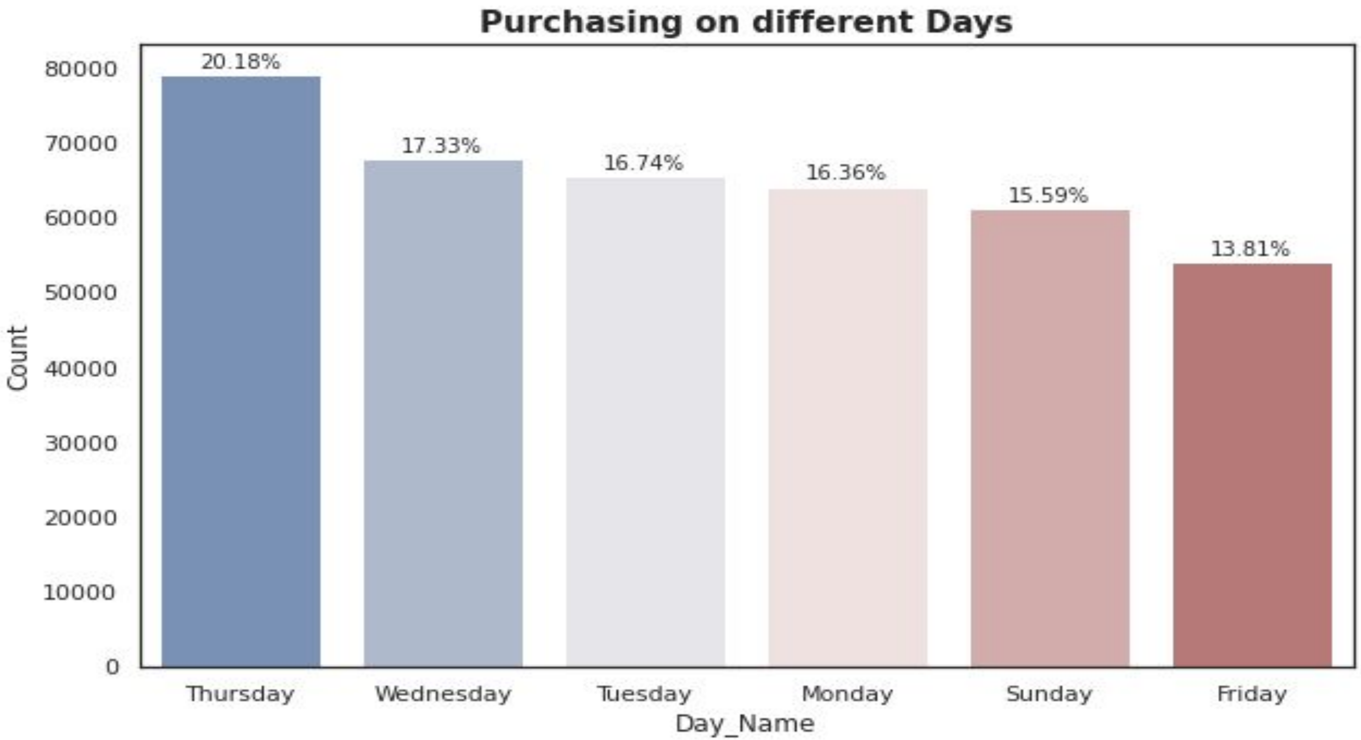
```
[ ] # Creating new column total_cost
    retail_df_copy['total_cost'] = retail_df_copy['Quantity'] * retail_df_copy['UnitPrice']
```





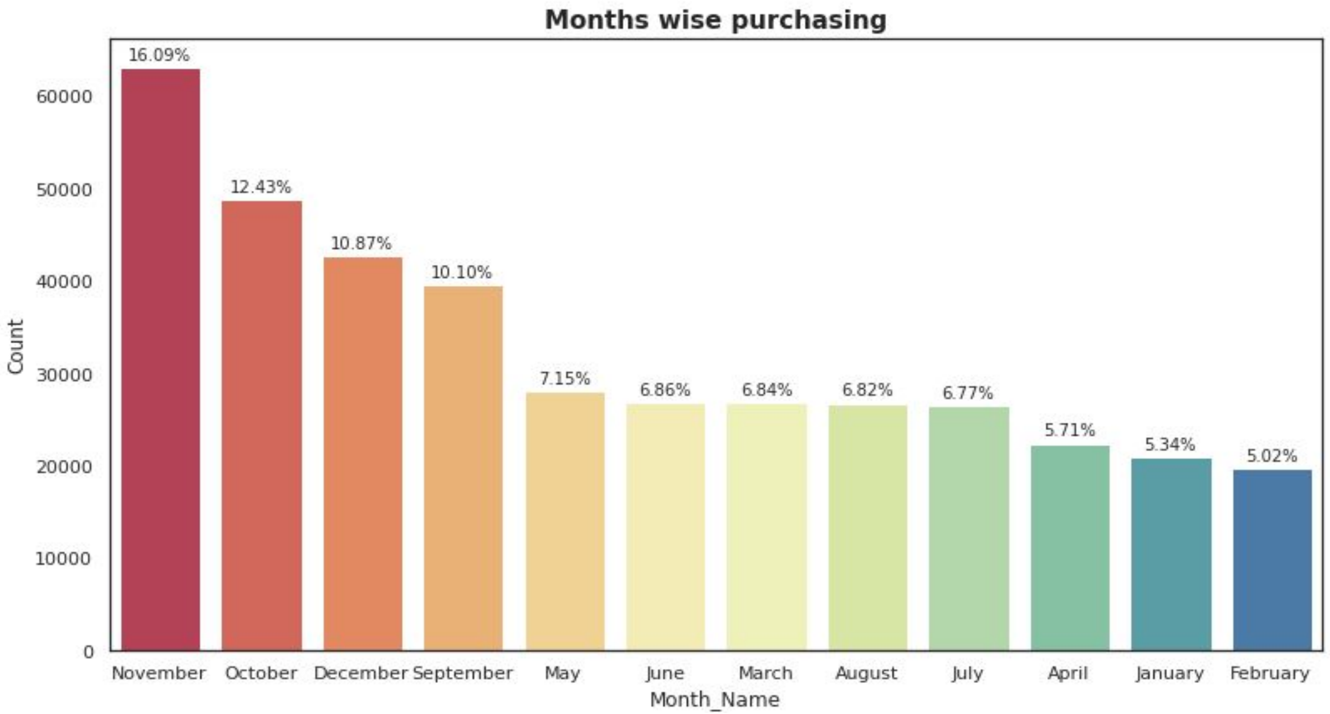
- Distribution of Total_Cost is highly positively skewed.
- Distribution of Total_Cost after log transformation.

Top days for purchasing



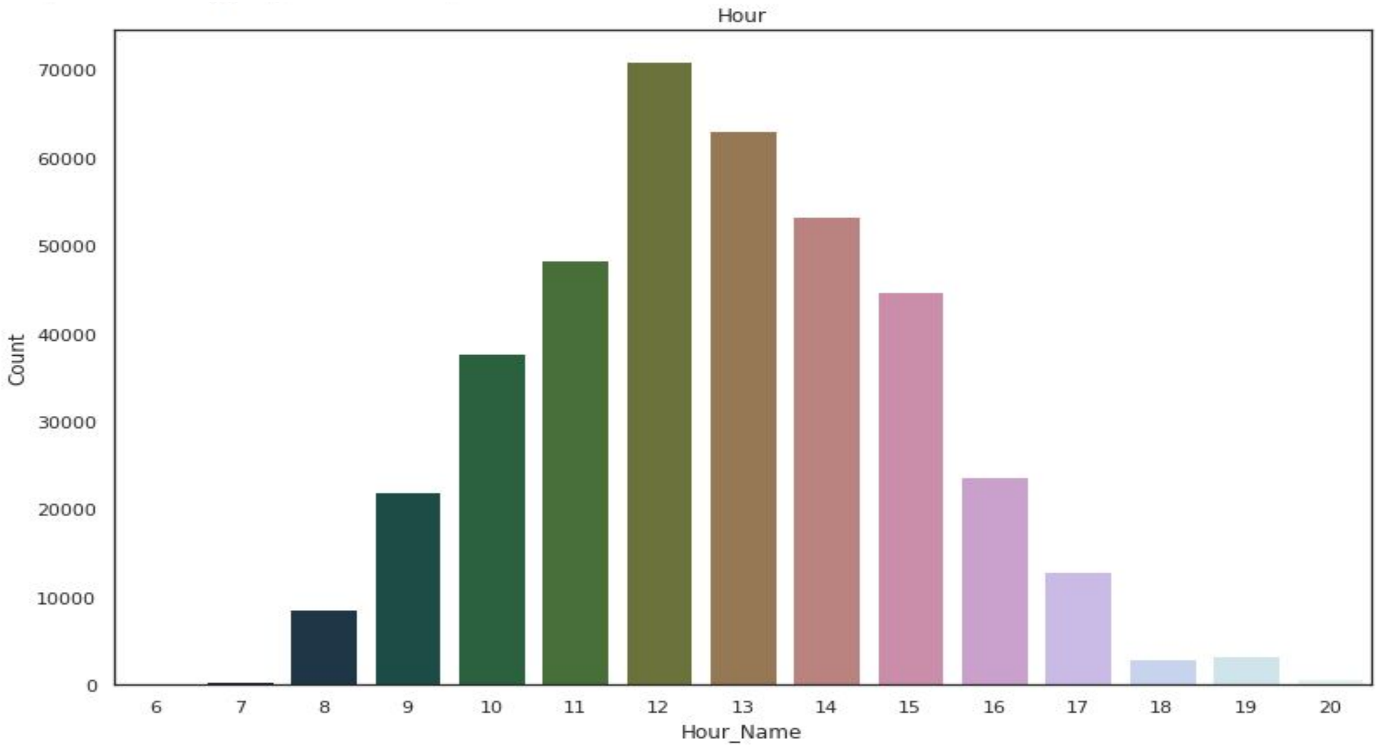
- Most of the customers have purchased the items in **Thursday ,Wednesday and Tuesday**.

Top Month for purchasing



- **Most** numbers of customers have purchased the gifts in the month of **November, October and December**.
- **Least** numbers of purchasing are in the month of **April and February**.

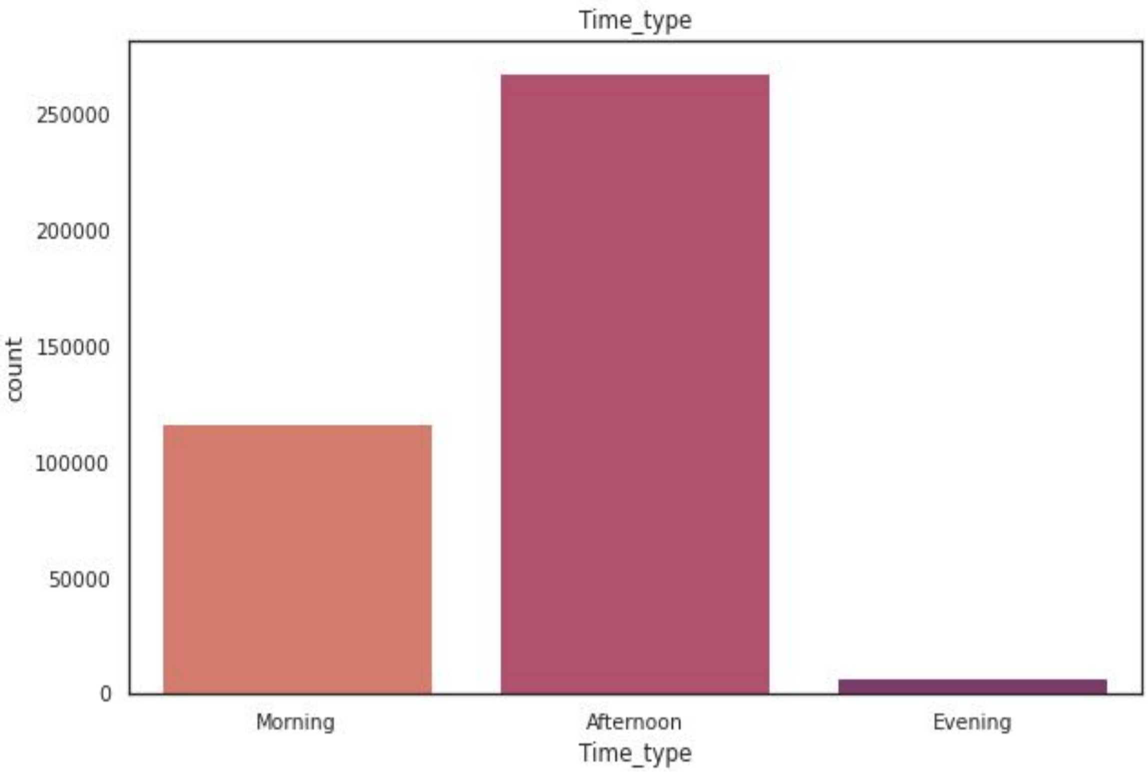
Top Hour for purchasing & transaction



- Most numbers of purchasing is done between **12pm to 2pm**.

Top day duration for purchasing

- Most of the customers have purchased the items in **Afternoon**.
- Moderate numbers of customers have purchased the items in **Morning** and least numbers of customers have purchased the items in **Evening**.



Recency, Frequency & Monetary (RFM)

RFM Metrics

RECENCY

The freshness of the customer activity, be it purchases or visits

E.g. Time since last order or last engaged with the product

FREQUENCY

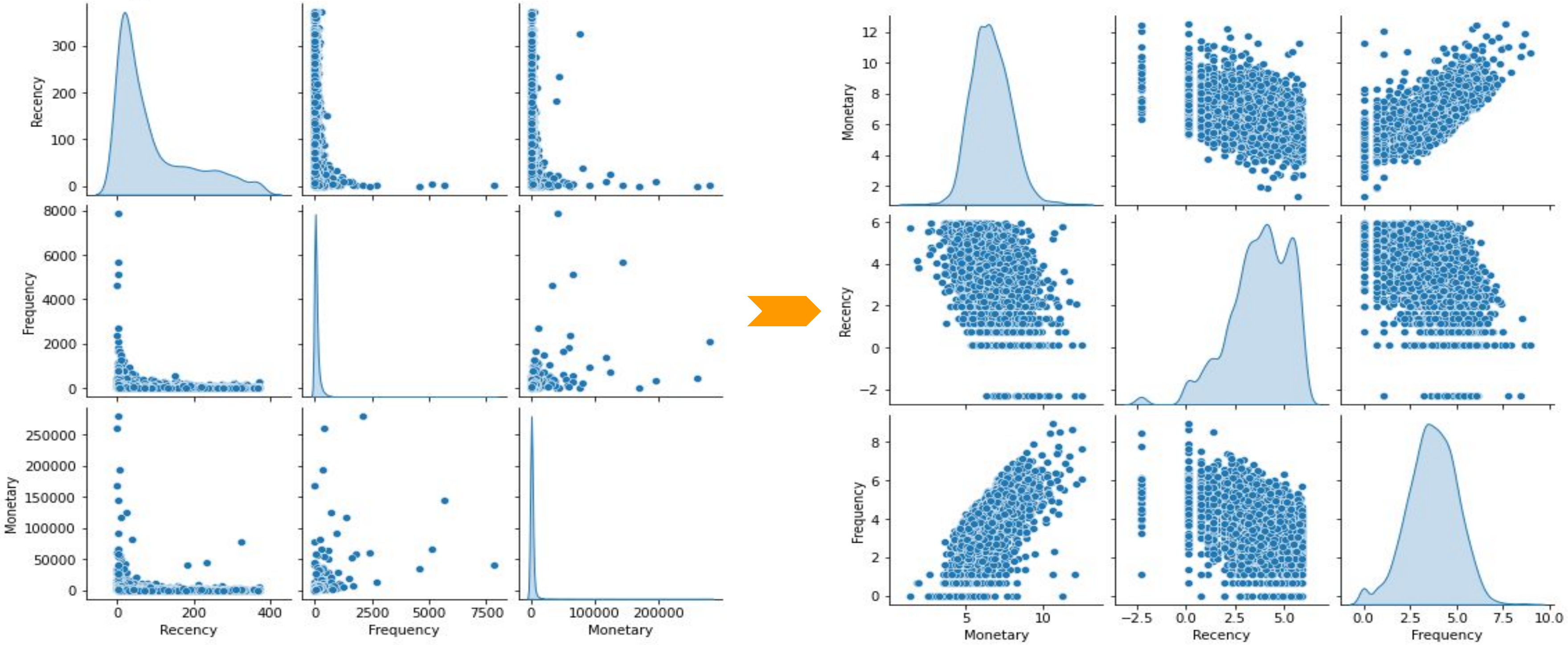
The frequency of the customer transactions or visits

E.g. Total number of transactions or average time between transactions/ engaged visits

MONETARY

The intention of customer to spend or purchasing power of customer

E.g. Total or average transactions value



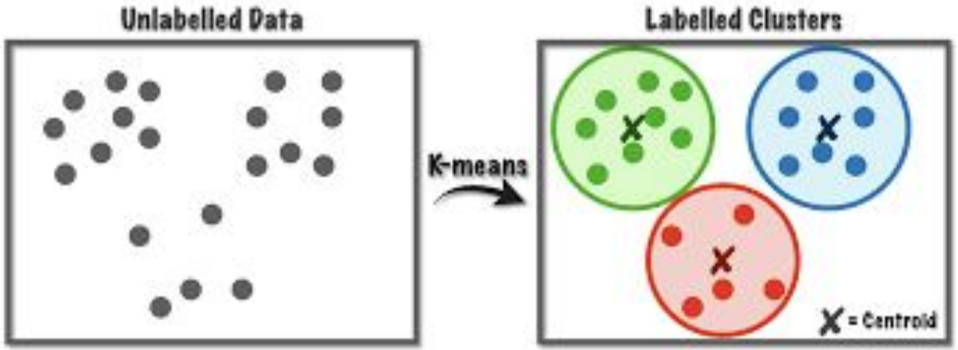
Pair plot showing the log transformation of the value generated by RFM model.

Clustering

Clustering is an unsupervised classification technique to understand the groups of classes in the data.

Models used for Clustering:

- K-Means Clustering
- DBSCAN
- Hierarchical Clustering



K-Means Clustering

- **K-means** algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.
- K-Means requires the number of clusters to be specified during the model building process. To know the right number of clusters, methods such as silhouette analysis and elbow method can be used. These methods will help in selection of the optimum number of clusters.

Methods to find optimal clusters

- **Silhouette score** : Silhouette score is used to evaluate the quality of clusters that ranges from -1 to1 , where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.
- **Elbow method** : a point from where the value of clusters starts decreasing suddenly. It calculates the sum of the square of the points and calculates the average distance.
- **DBSCAN** : DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It is basically a clustering algorithm based on density.

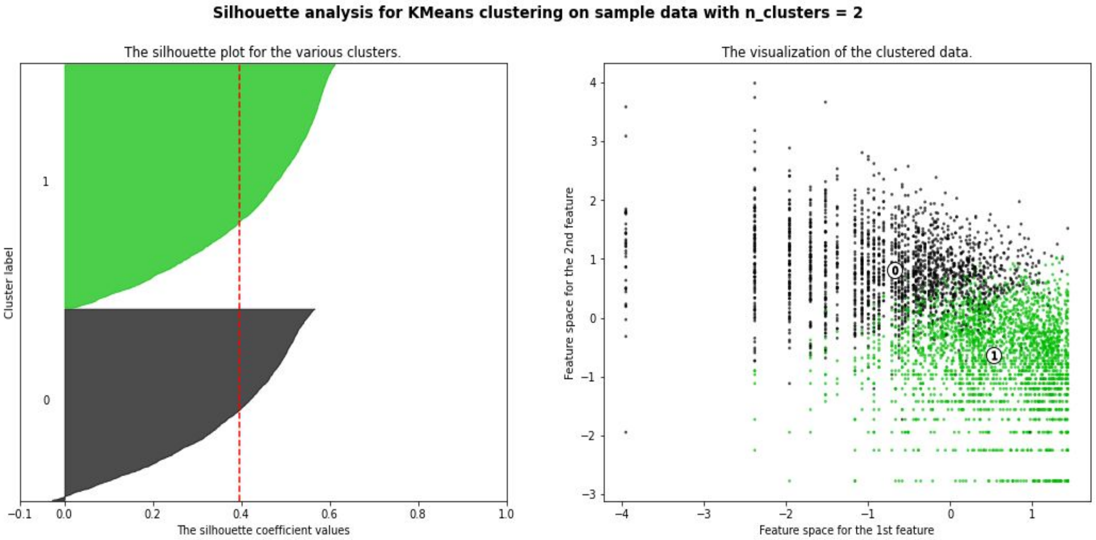
Hierarchical Clustering

- **Hierarchical clustering** is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom. To get the number of clusters for hierarchical clustering, we make use of an awesome concept called a Dendogram.

  Dendogram : A Dendogram is a type of tree diagram showing hierarchical relationships between different sets of data.

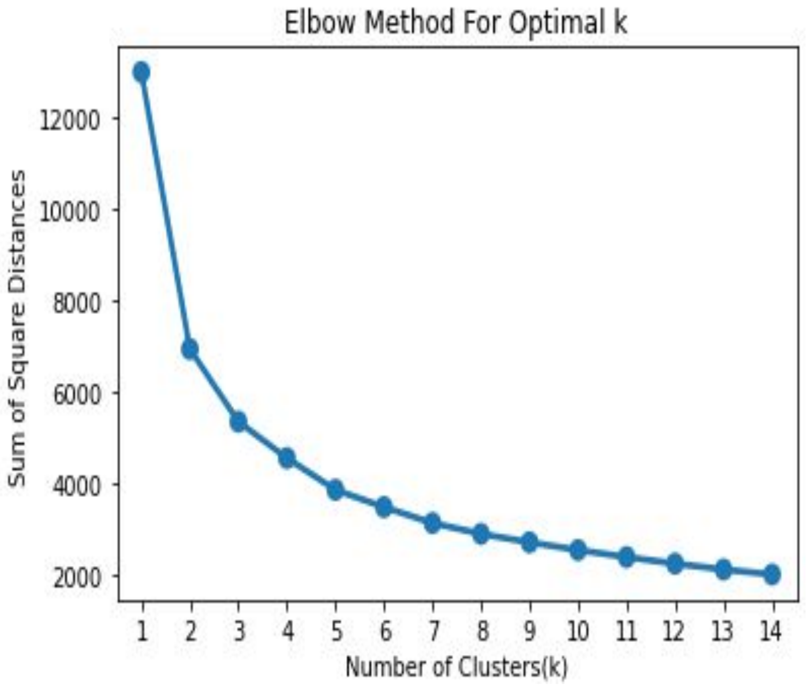
Silhouette analysis on RFM

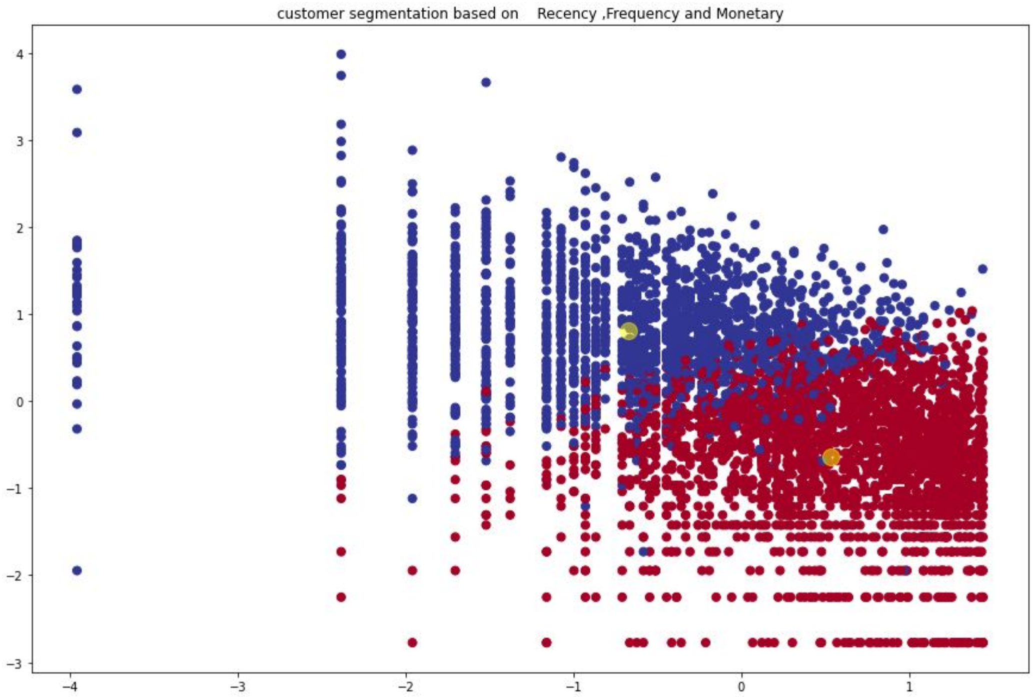
- The best silhouette score came up with **n_clusters = 2**.
- The average silhouette score is = **0.3955**.




```
For n_clusters = 2 The average silhouette_score is : 0.395520935854327
For n_clusters = 3 The average silhouette_score is : 0.30760934385372846
For n_clusters = 4 The average silhouette_score is : 0.299065893607 5084
For n_clusters = 5 The average silhouette_score is : 0.2776137265878769
For n_clusters = 6 The average silhouette_score is : 0.2765091669765864
For n_clusters = 7 The average silhouette_score is : 0.26673065111937905
For n_clusters = 8 The average silhouette_score is : 0.263281041567485
For n_clusters = 9 The average silhouette_score is : 0.24757930019808283
For n_clusters = 10 The average silhouette_score is : 0.26063427891209173
For n_clusters = 11 The average silhouette_score is : 0.2588527995000704
For n_clusters = 12 The average silhouette_score is : 0.26047649311366
For n_clusters = 13 The average silhouette_score is : 0.26003332882071595
For n_clusters = 14 The average silhouette_score is : 0.2615187898075716
For n_clusters = 15 The average silhouette_score is : 0.2602781282489402
```

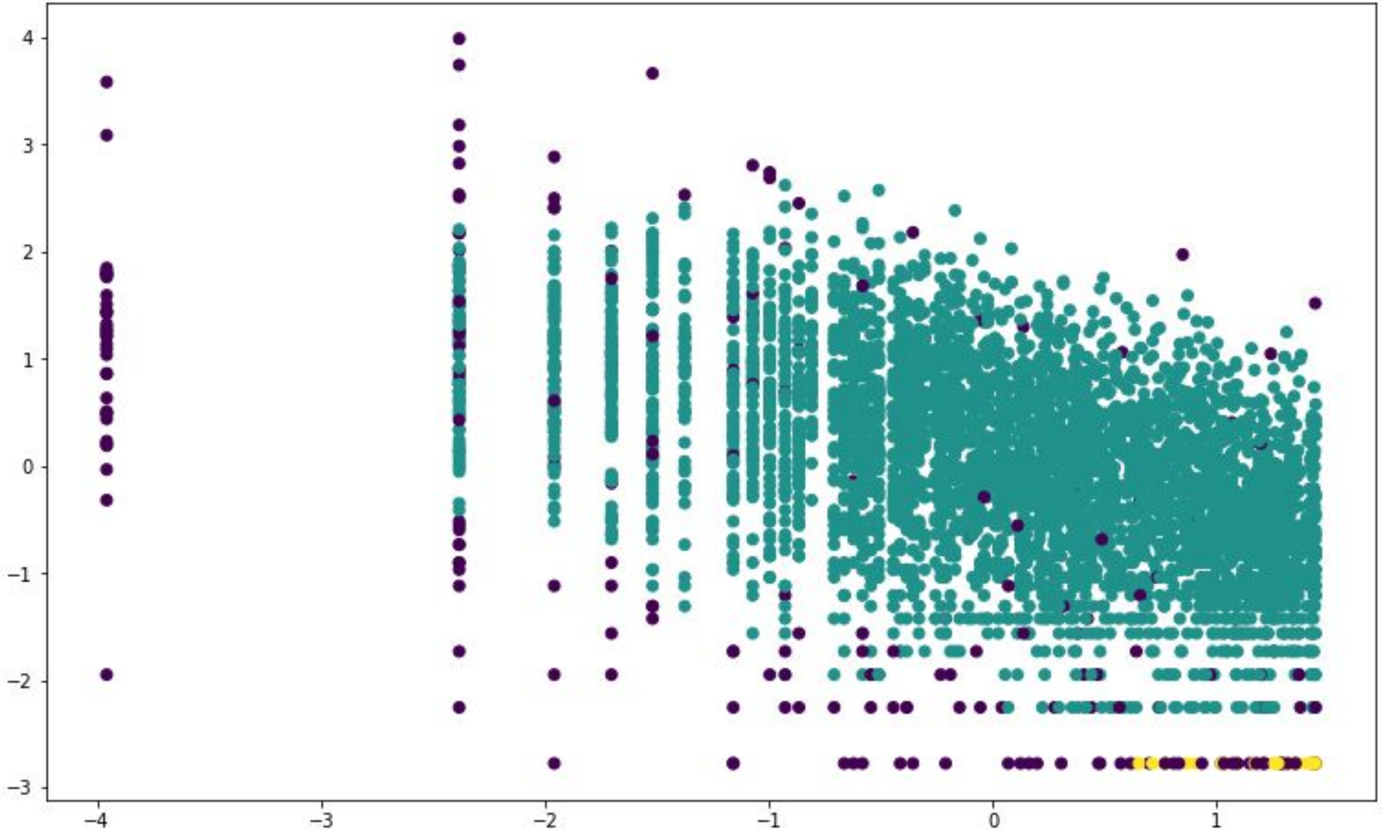
Elbow method and Cluster chart on RFM





DBSCAN on RFM

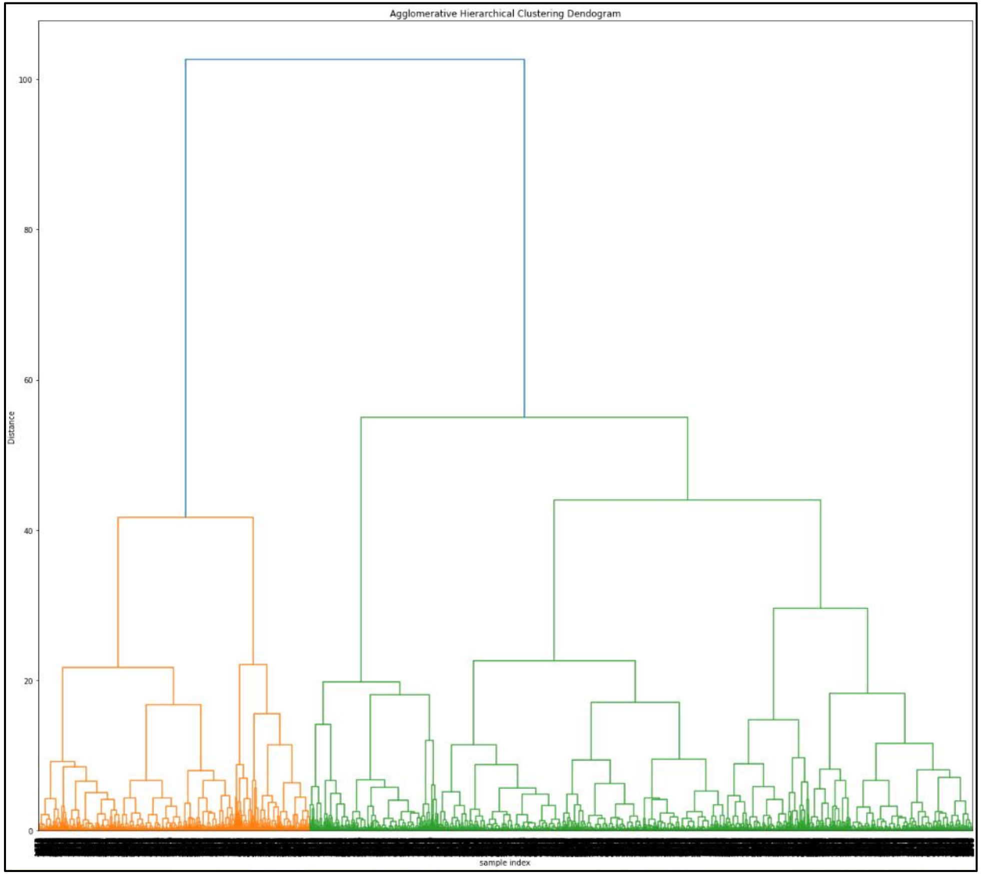
- No. of Clusters resulted is **3.**



Hierarchical Clustering

- The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold=90 degree.

- **No. of Clusters = 2**



RFM Analysis

AI

CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore	Cluster
12346.0	325	1	77183.60	1	1	4	114	1
12347.0	2	182	4310.00	4	4	4	444	0
12348.0	75	31	1797.24	2	2	4	224	1
12349.0	18	73	1757.55	3	3	4	334	0
12350.0	310	17	334.40	1	1	2	112	1
12352.0	36	85	2506.04	3	3	4	334	0
12353.0	204	4	89.00	1	1	1	111	1
12354.0	232	58	1079.40	1	3	3	133	1
12355.0	214	13	459.40	1	1	2	112	1
12356.0	22	59	2811.43	3	3	4	334	0

Summary

SL No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means with silhouette_score	RFM	2
2	K-Means with Elbow methos	RFM	2
3	Hierarchical clustering	RFM	2
4	DBSCAN	RFM	3

Challenges

- Huge dataset
- Null values Treatment
- Treatment of cancelled orders
- Right number of ‘k’ for clusters

Conclusion

- This project mainly focused on developing customer segments for a UK based online store, selling unique all occasion gifts.
- Top Five percentage of orders from Countries: **United Kingdom(88.95%)**, Germany(2.33%), France(1.84%), Ireland(1.84%) and Spain(0.62%).
- The month which give maximum business: **November**, October, December, September and May.
- Most of the customers usually purchase products in between **10:00 A.M to 2:00 P.M** and top time duration of a day for purchasing: **Afternoon** > Morning > Evening.
- Using a **recency, frequency and monetary** (RFM) analysis, the customers have been segmented into various clusters.
- By applying different clustering algorithm to our dataset , we get the optimal **number of cluster is equal to 2**.
- The business can focus on these different clusters and provide to customers of each sector in a different way, which would not only benefit the customer but also the business at large.


thank you