Capstone Project

Online Retail Customer Segmentation

Individual Contributor

Aamir Sohail

Content

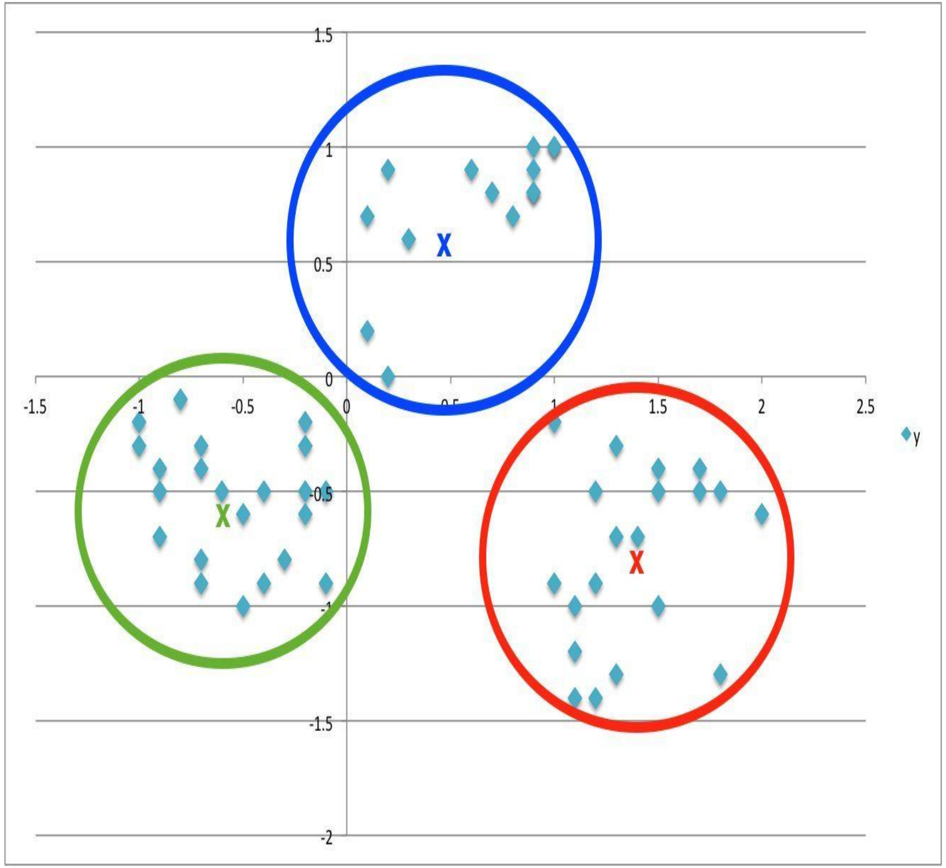
PROBLEM
STATEMENT DATA

SUMMARY

ANALYSIS

CHALLENGES

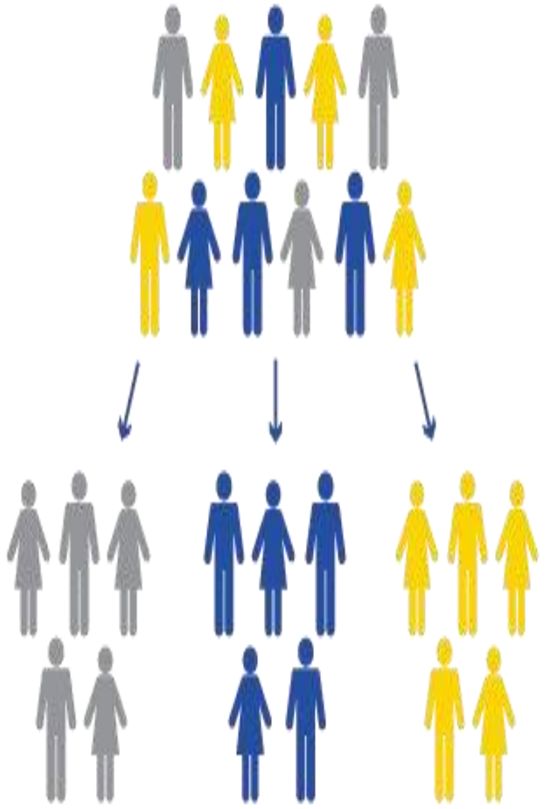
CONCLUSION



What is Customer Segmentation?

- **Practice of dividing a customer base into groups of individuals** that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

- Allows us to better understand our customers **helping us target these customers in a more efficient manner and improve the customer experience**.

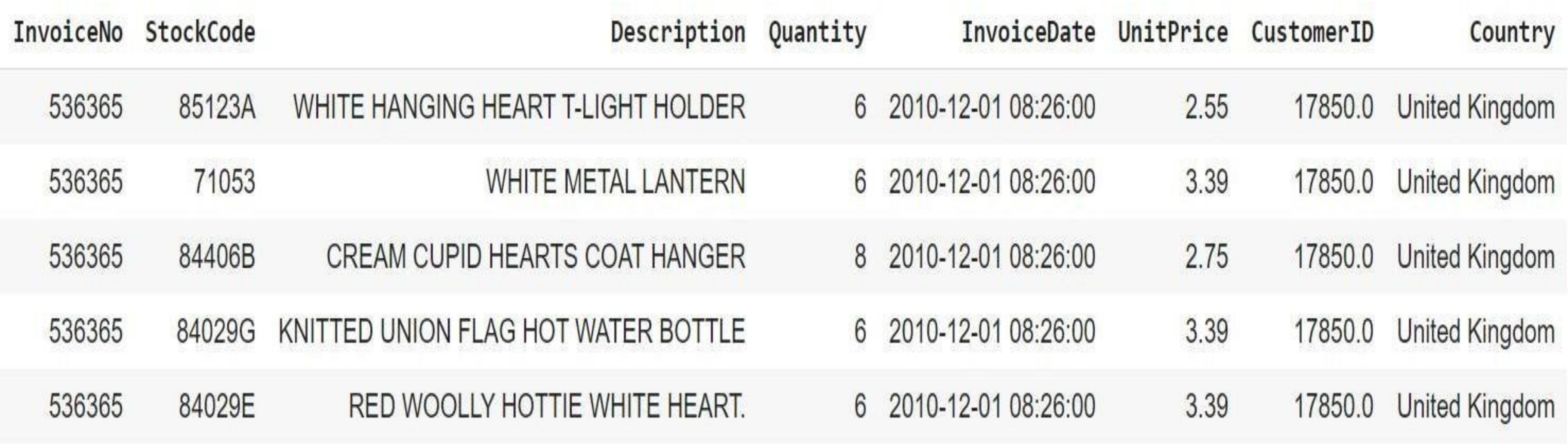


Problem Statement

Given a dataset related to a online retailer based out of the UK, we need to analyse and identify major customer segments using K Means algorithm and also using different verification method to confirm the result.

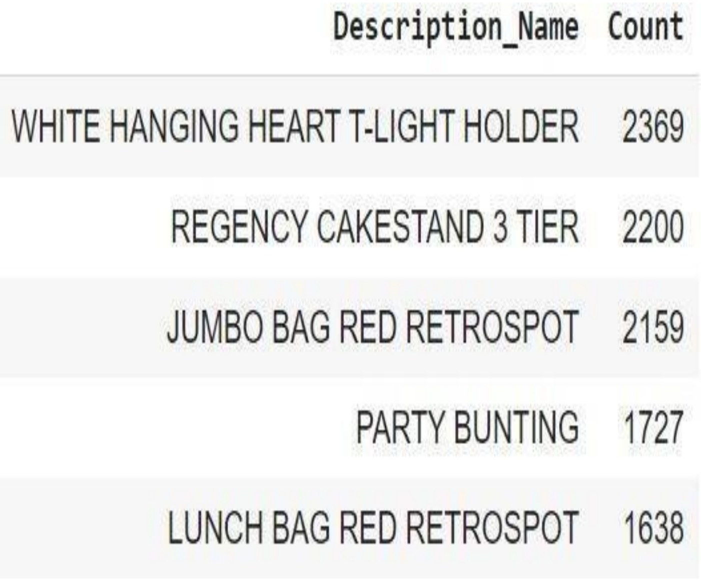
Data Summary

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

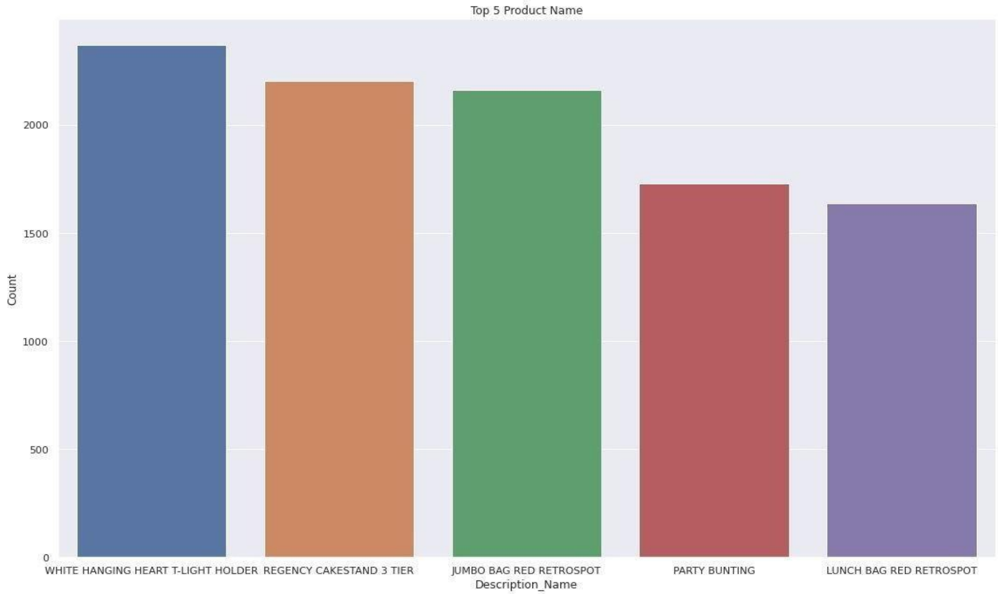


InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT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

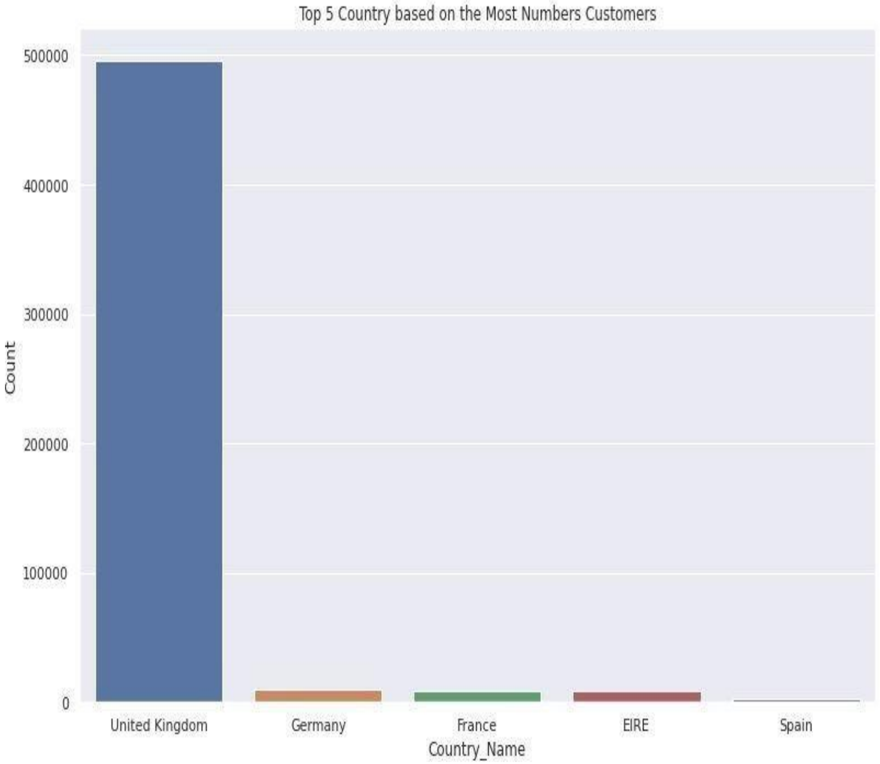
Finding the most Purchased Products

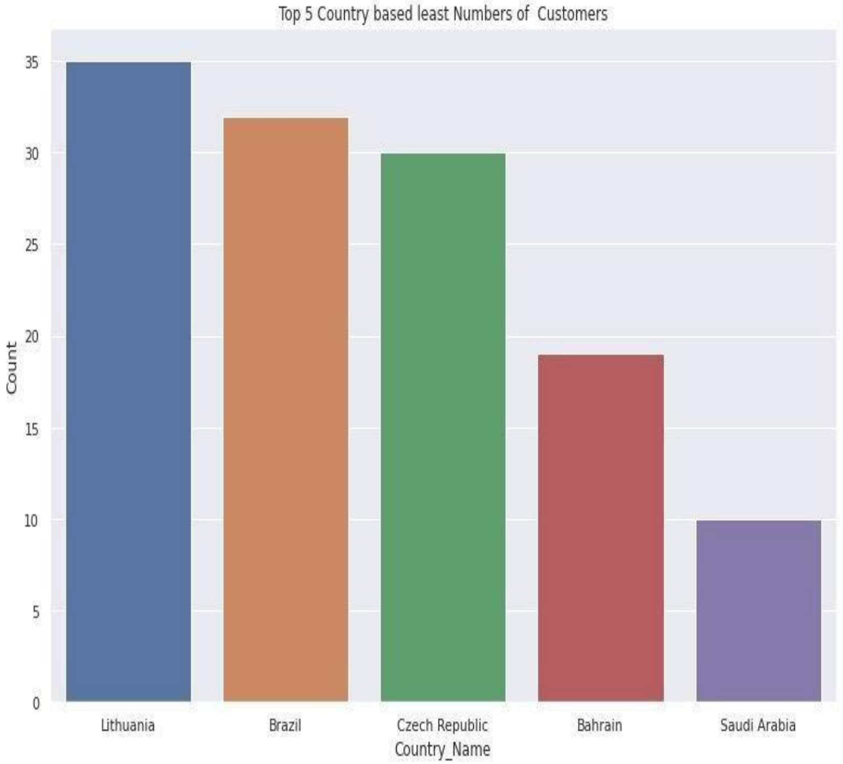


Description_Name	Count
WHITE HANGING HEART T-LIGHT HOLDER	2369
REGENCY CAKESTAND 3 TIER	2200
JUMBO BAG RED RETROSPOT	2159
PARTY BUNTING	1727
LUNCH BAG RED RETROSPOT	1638



Top 5 vs Bottom 5 countries

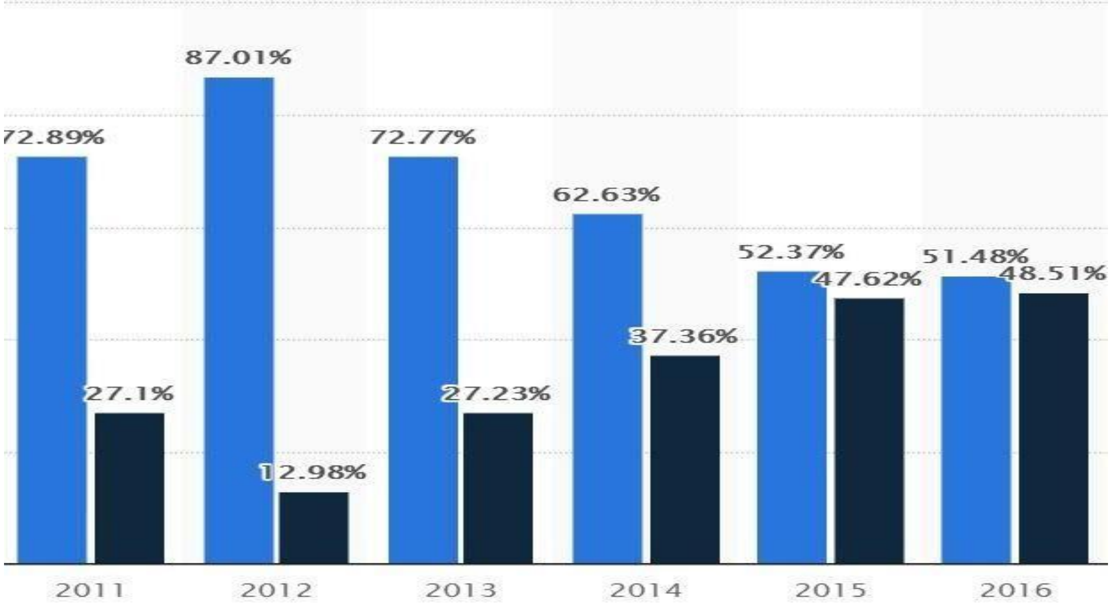




Analysis

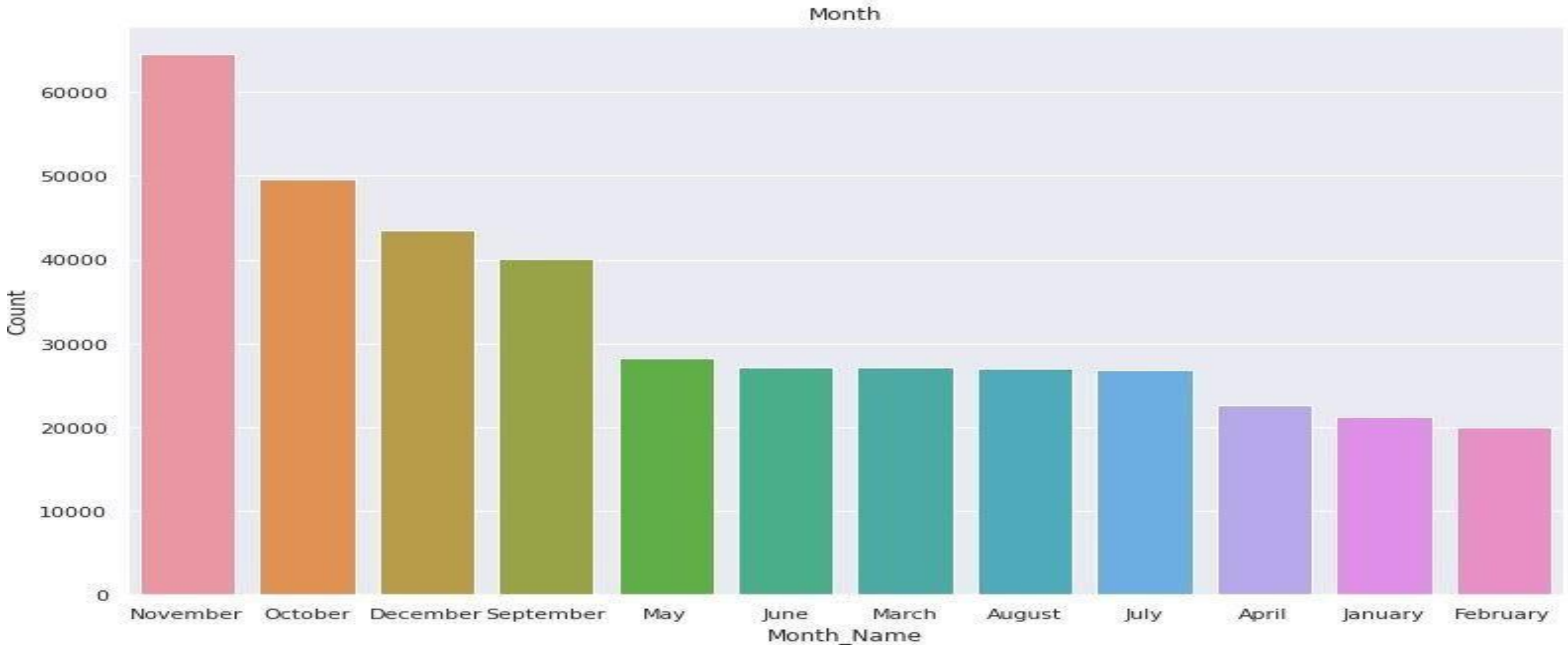
UK

Saudi



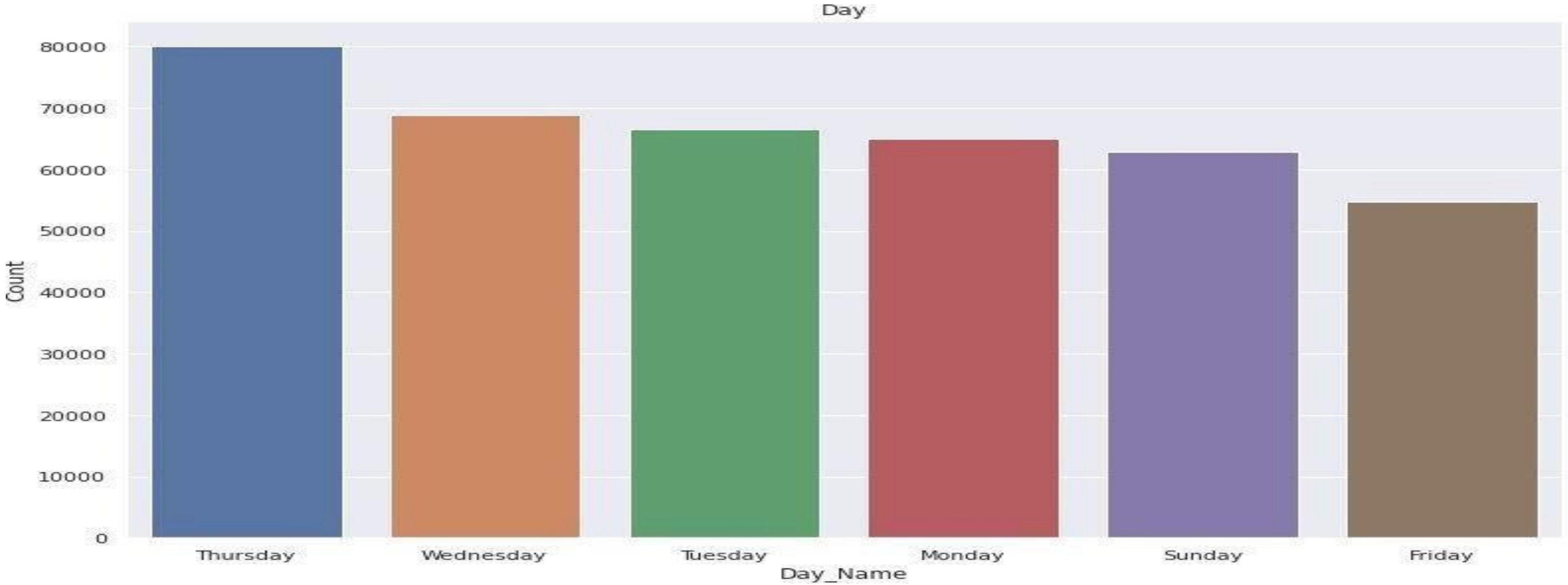
Source obtained from Statista comparing online purchases from 2011 to 2016

Month-wise analysis

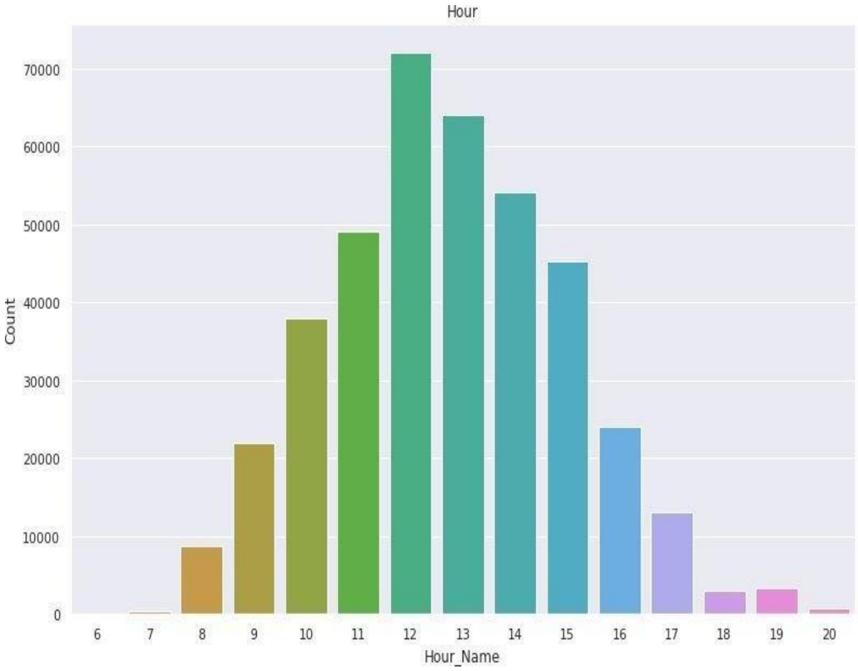


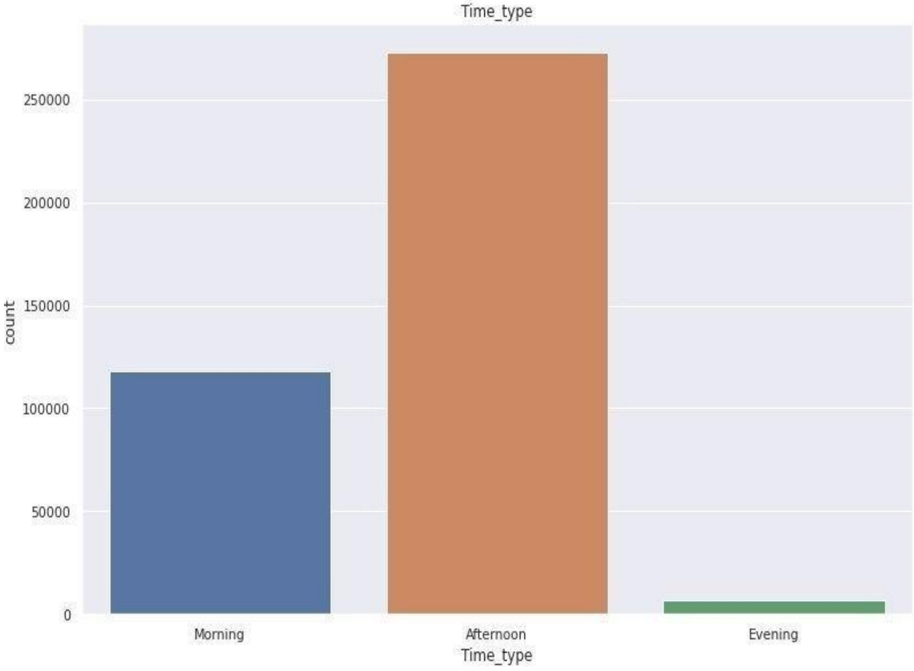
November and December could be the months with highest sales in anticipation of Christmas

Daywise analysis



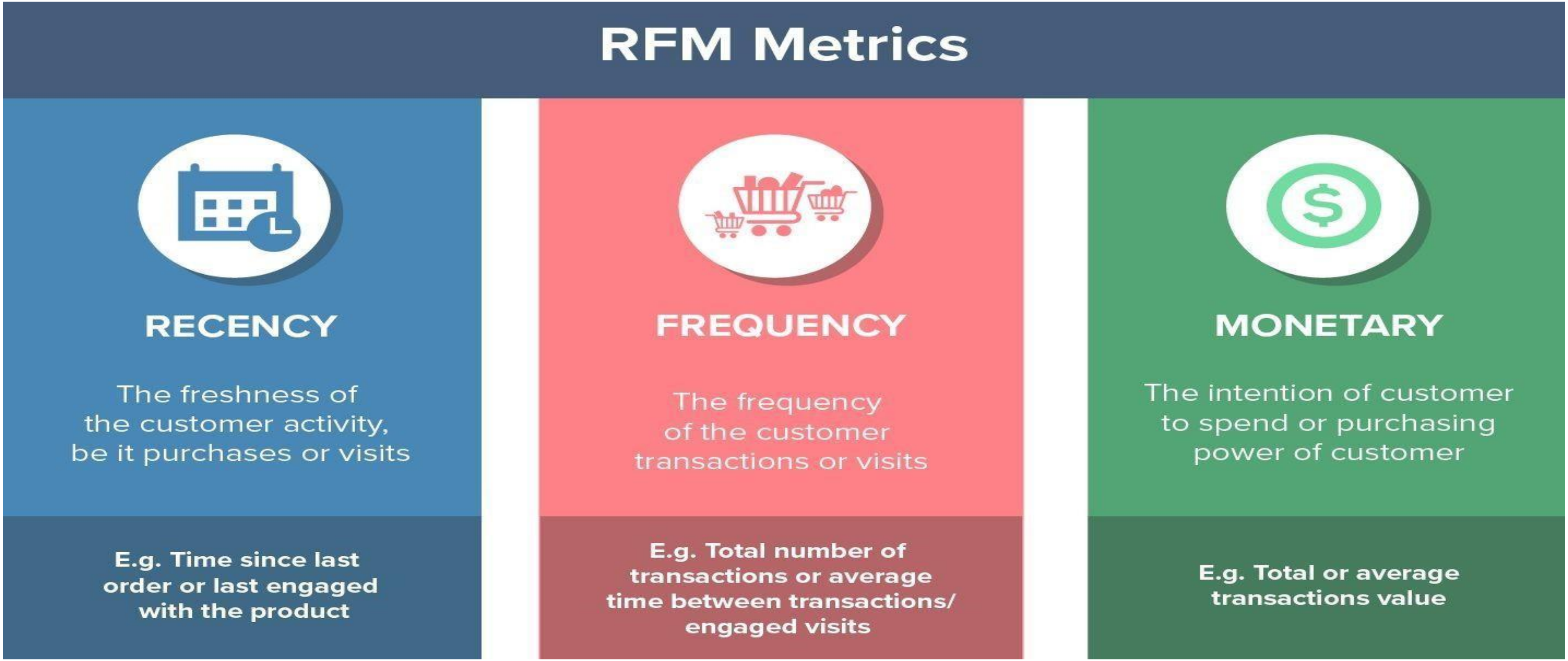
Hourwise analysis





Working hours witnessing the highest sales could be attributed to the fact that a large part of the dataset is Wholesalers' data

Recency,Frequency,Monetary values



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

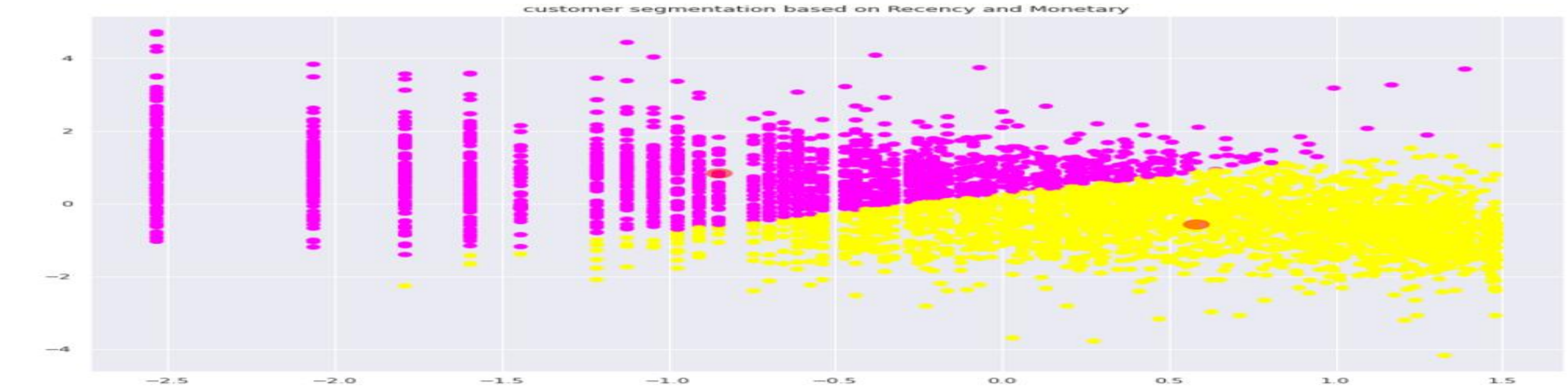
E.g. Total number of transactions or average time between transactions/ engaged visits

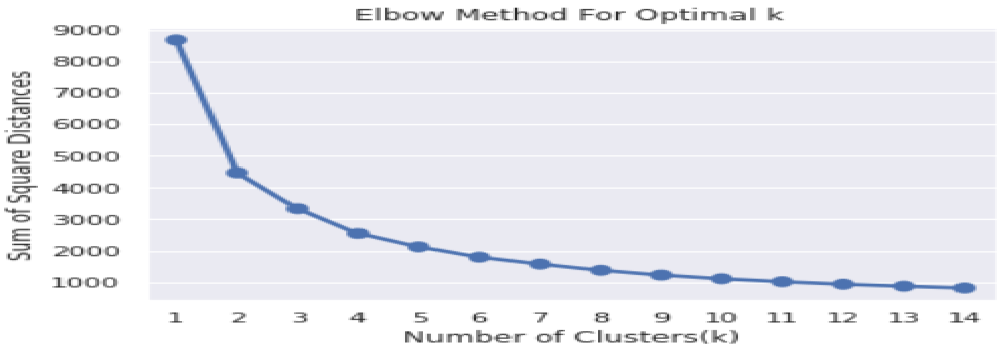
MONETARY

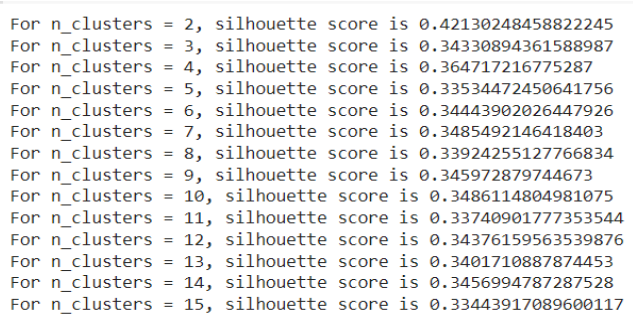
The intention of customer to spend or purchasing power of customer

E.g. Total or average transactions value

Silhouette score and Elbow method onR&M

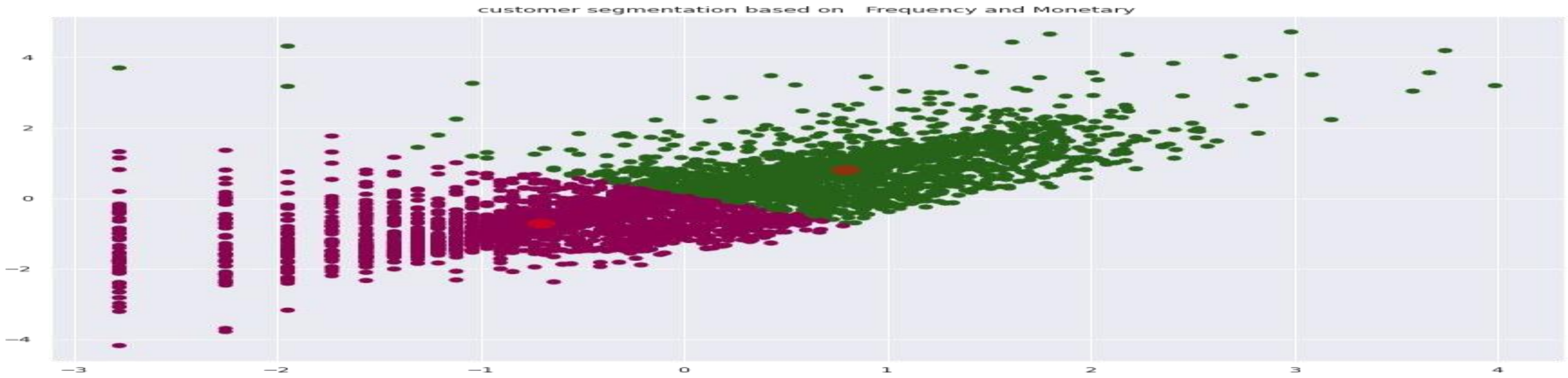


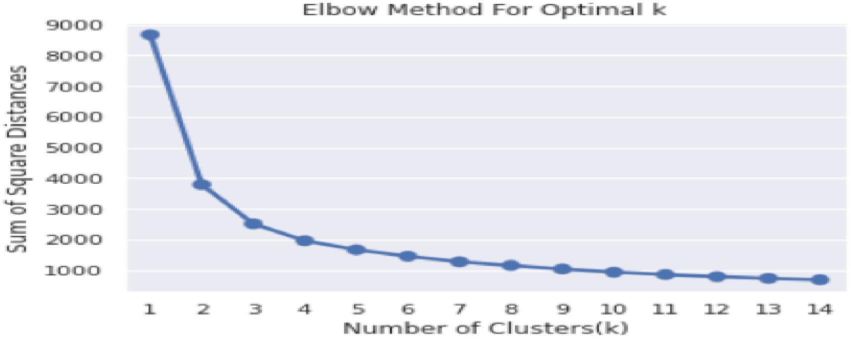


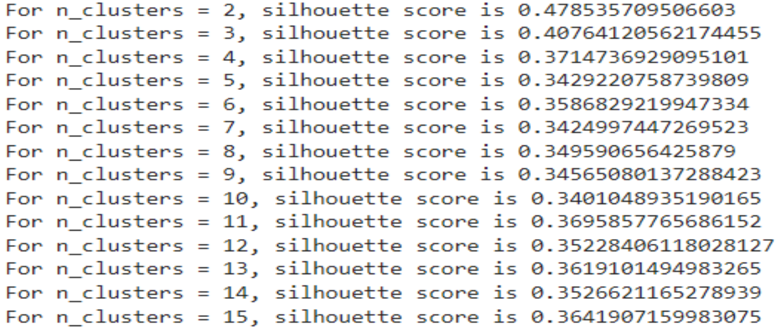


```
For n_clusters = 2, silhouette score is 0.42130248458822245
For n_clusters = 3, silhouette score is 0.34330894361588987
For n_clusters = 4, silhouette score is 0.364717216775287
For n_clusters = 5, silhouette score is 0.33534472450641756
For n_clusters = 6, silhouette score is 0.34443902026447926
For n_clusters = 7, silhouette score is 0.348549214641840З
For n_clusters = 8, silhouette score is 0.33924255127766834
For n_clusters = 9, silhouette score is 0.345972879744673
For n_clusters = 10, silhouette score is 0.3486114804981075
For n_clusters = 11, silhouette score is 0.33740901777353544
For n_clusters = 12, silhouette score is 0.34376159563539876
For n_clusters = 13, silhouette score is 0.3401710887874453
For n_clusters = 14, silhouette score is 0.3456994787287528
For n_clusters = 15, silhouette score is 0.33443917089600117
```

Silhouette score and Elbow method on F&M

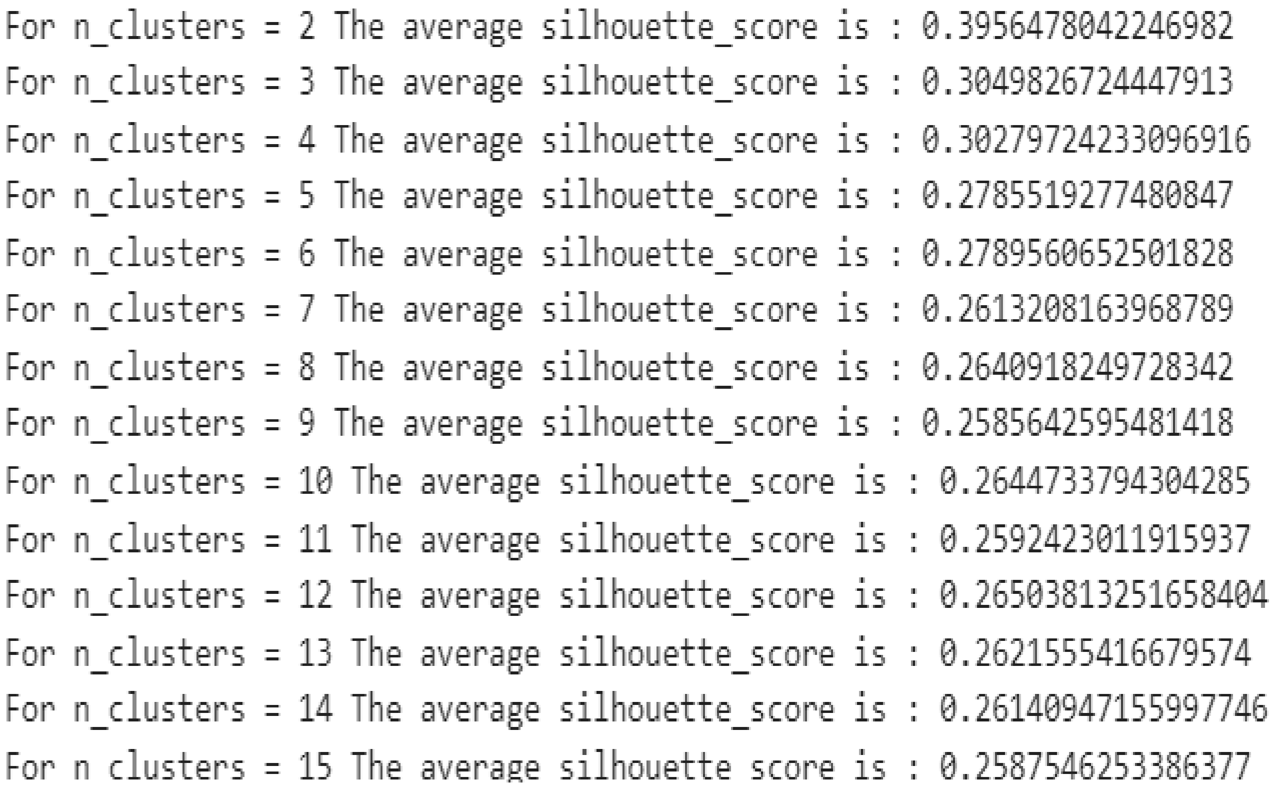






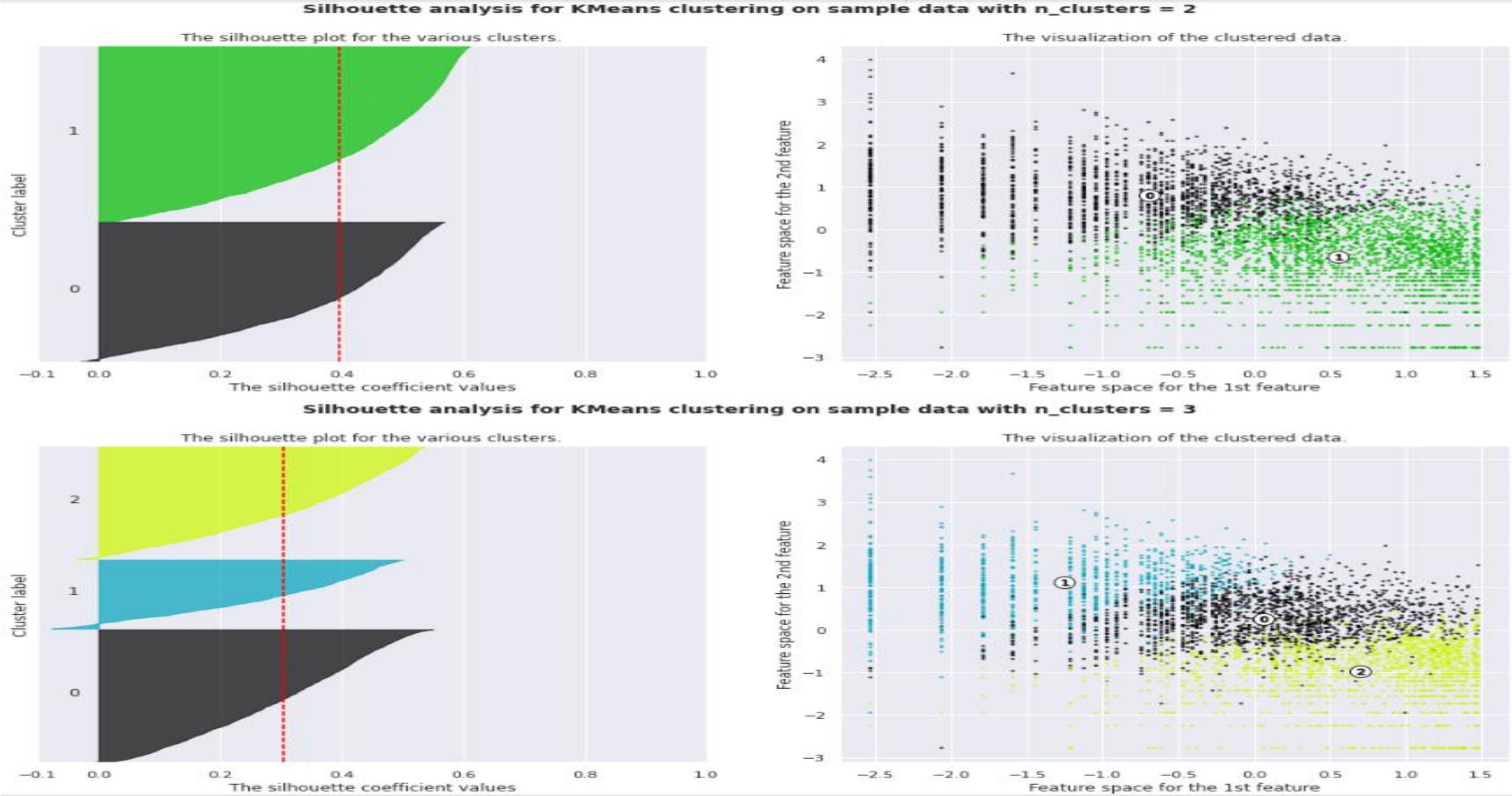
```
For n_clusters = 2, silhouette score is 0.478535709506603
For n_clusters = 3, silhouette score is 0.40764120562174455
For n_clusters = 4, silhouette score is 0.371473692909510l
For n_clusters = 5, silhouette score is 0.3429220758739809
For n_clusters = 6, silhouette score is 0.3586829219947334
For n_clusters = 7, silhouette score is 0.3424997447269523
For n_clusters = 8, silhouette score is 0.349590656425879
For n_clusters = 9, silhouette score is 0.34565080137288423
For n_clusters = 10, silhouette score is 0.3401048935190165
For n_clusters = 11, silhouette score is 0.3695857765686152
For n_clusters = 12, silhouette score is 0.35228406118028127
For n_clusters = 13, silhouette score is 0.3619101494983265
For n_clusters = 14, silhouette score is 0.3526621165278939
For n_clusters = 15, silhouette score is 0.3641907159983075
```

Silhouette analysis on R, F and M

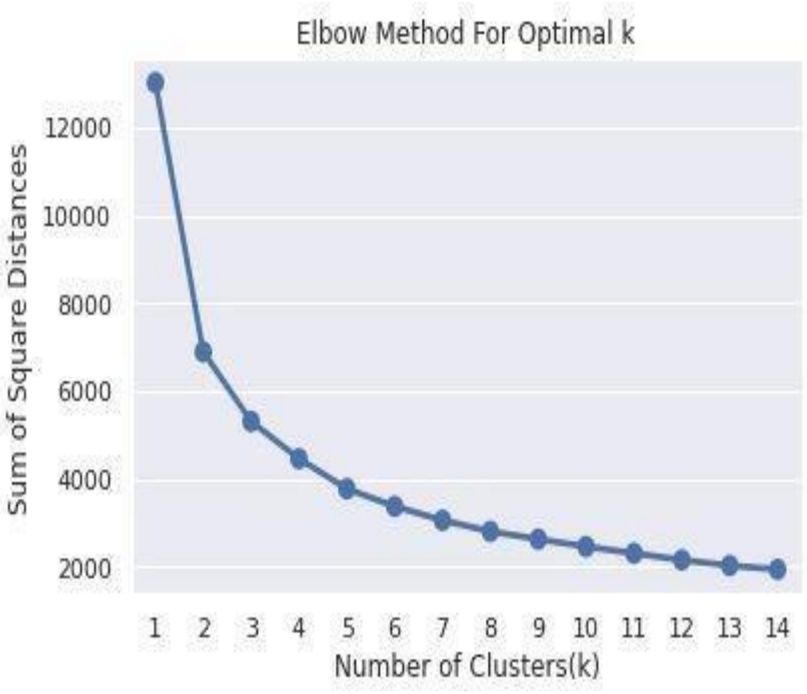


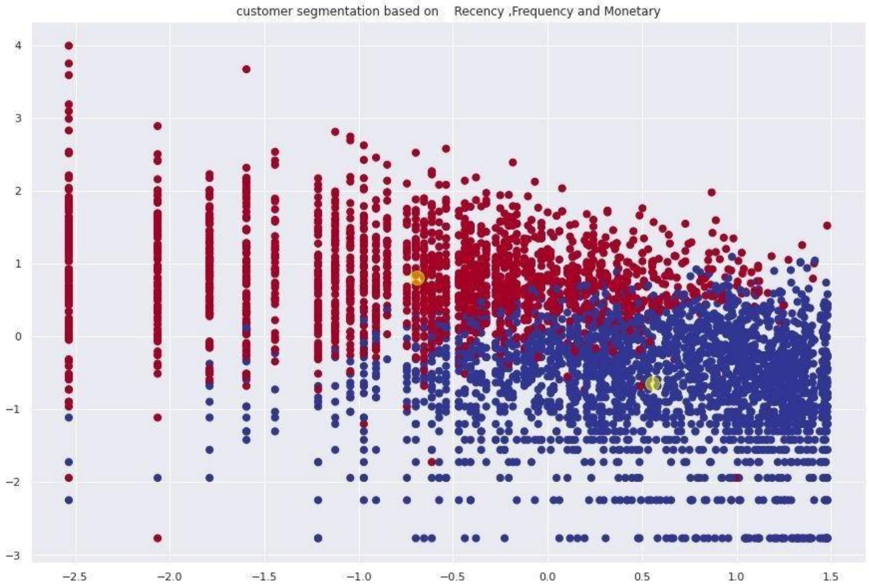
```
For n_clusters = 2 The average silhouette_score is : 0.3956478042246982
For n_clusters = 3 The average silhouette_score is : 0.3049826724447913
For n_clusters = 4 The average silhouette_score is : 0.30279724233096916
For n_clusters = 5 The average silhouette_score is : 0.2785519277480847
For n_clusters = 6 The average silhouette_score is : 0.2789560652501828
For n_clusters = 7 The average silhouette_score is : 0.2613208163968789
For n_clusters = 8 The average silhouette_score is : 0.2640918249728342
For n_clusters = 9 The average silhouette_score is : 0.2585642595481418
For n_clusters = 10 The average silhouette_score is : 0.2644733794304285
For n_clusters = 11 The average silhouette_score is : 0.2592423011915937
For n_clusters = 12 The average silhouette_score is : 0.26503813251658404
For n_clusters = 13 The average silhouette_score is : 0.2621555416679574
For n_clusters = 14 The average silhouette_score is : 0.26140947155997746
For n clusters = 15 The average silhouette score is : 0.258754625338637
```

Silhouette analysis on RFM

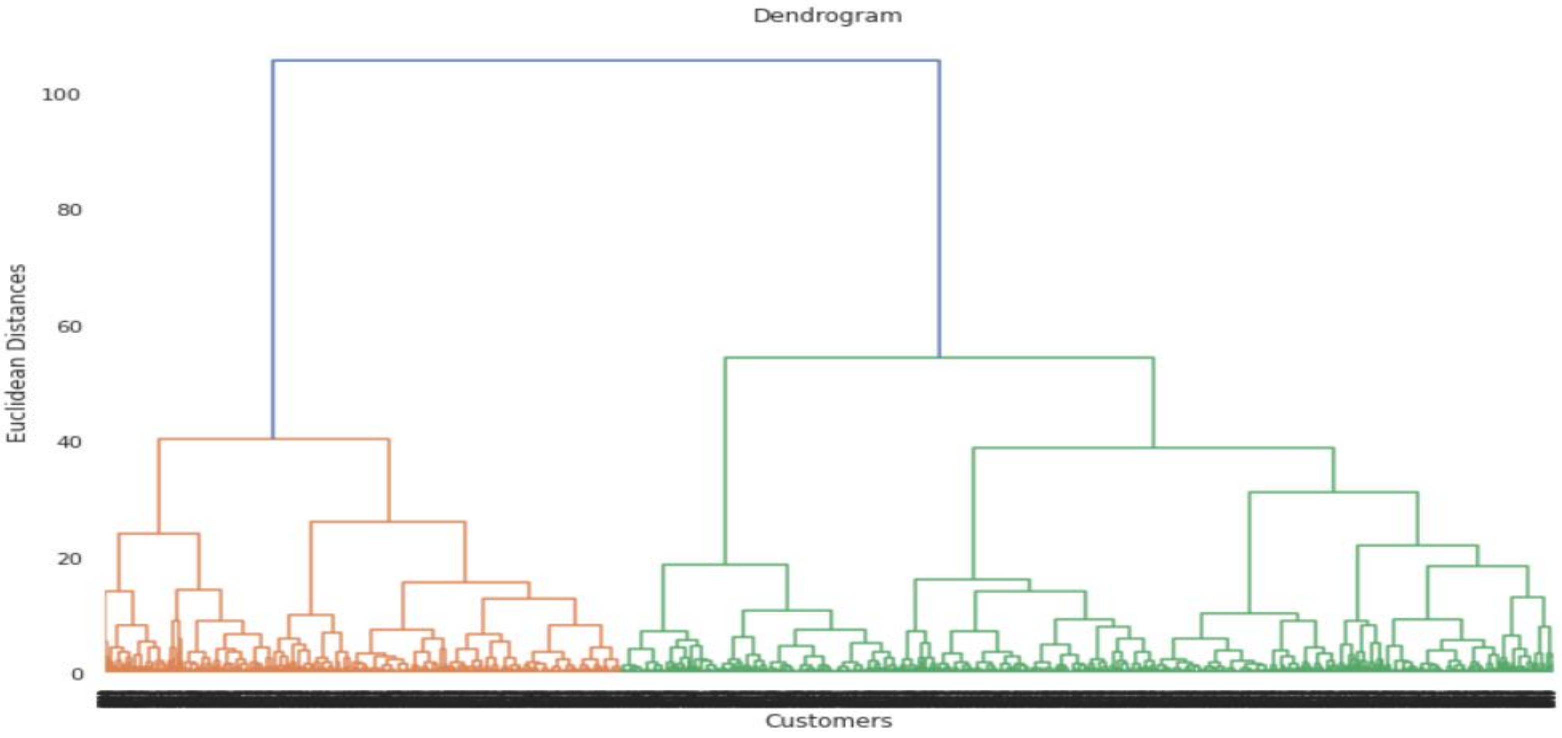


Elbow method and Cluster chart on RFM

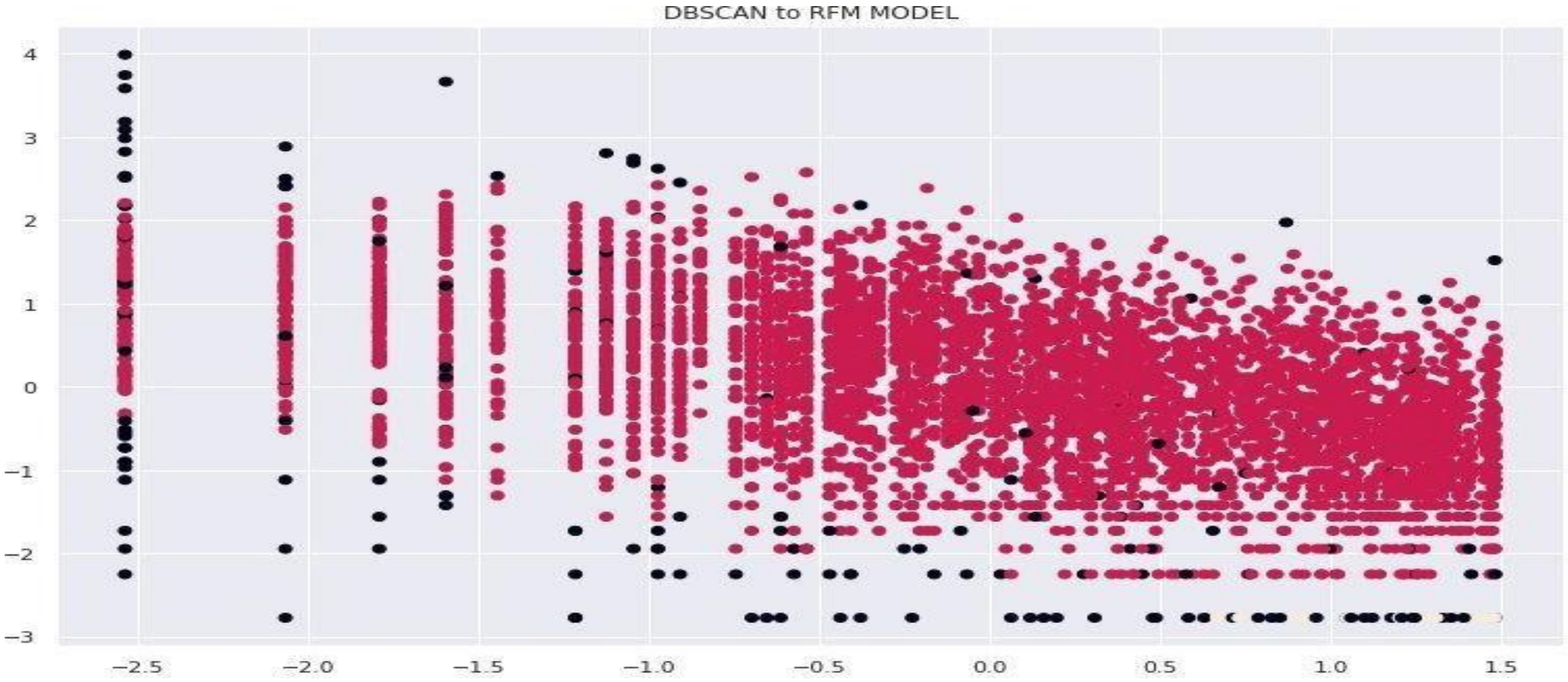




Dendrogram



DBSCAN

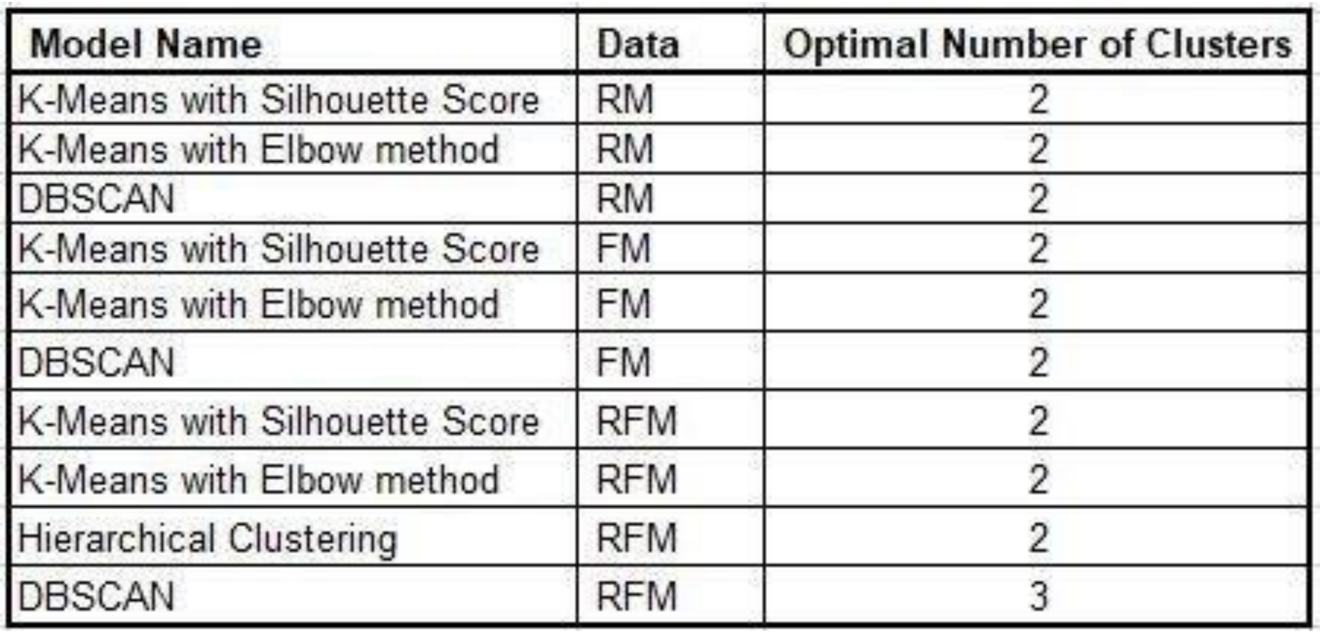


Challenges

- **Tackling refunds**

- **Right number of ‘k’ for clusters**

Conclusion



Model Name	Data	Optimal Number of Clusters
K-Means with Silhouette Score	RM	2
K-Means with Elbow method	RM	2
DBSCAN	RM	2
K-Means with Silhouette Score	FM	2
K-Means with Elbow method	FM	2
DBSCAN	FM	2
K-Means with Silhouette Score	RFM	2
K-Means with Elbow method	RFM	2
Hierarchical Clustering	RFM	2
DBSCAN	RFM	3