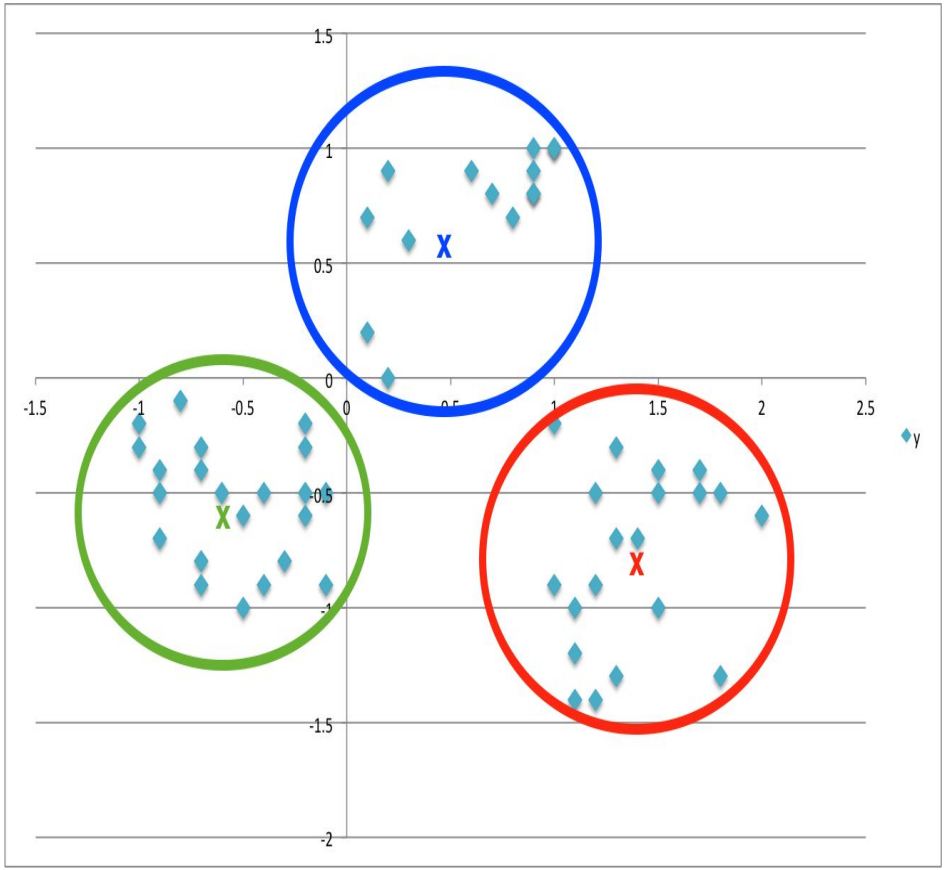
# Online Retail Customer Segmentation

# Content

- PROBLEM STATEMENT
- DATA SUMMARY
- ANALYSIS
- CHALLENGES
- CONCLUSION
- Q & A



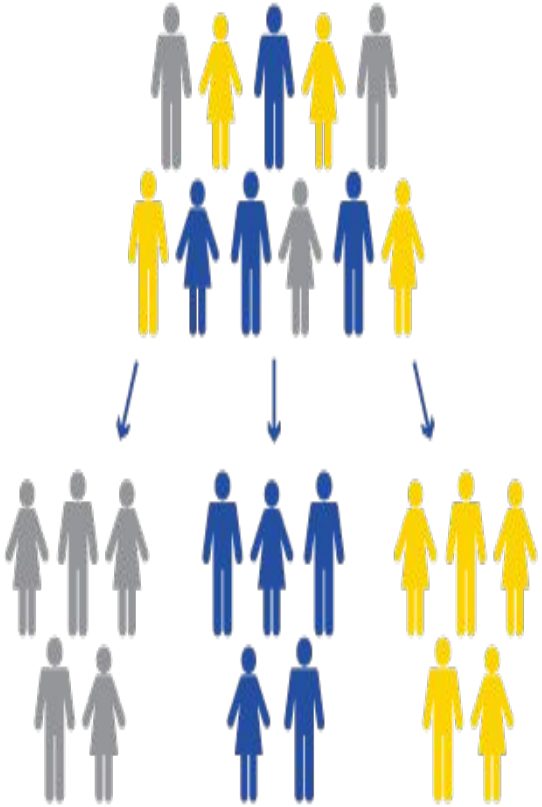
# Problem Statement

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.

# 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**.



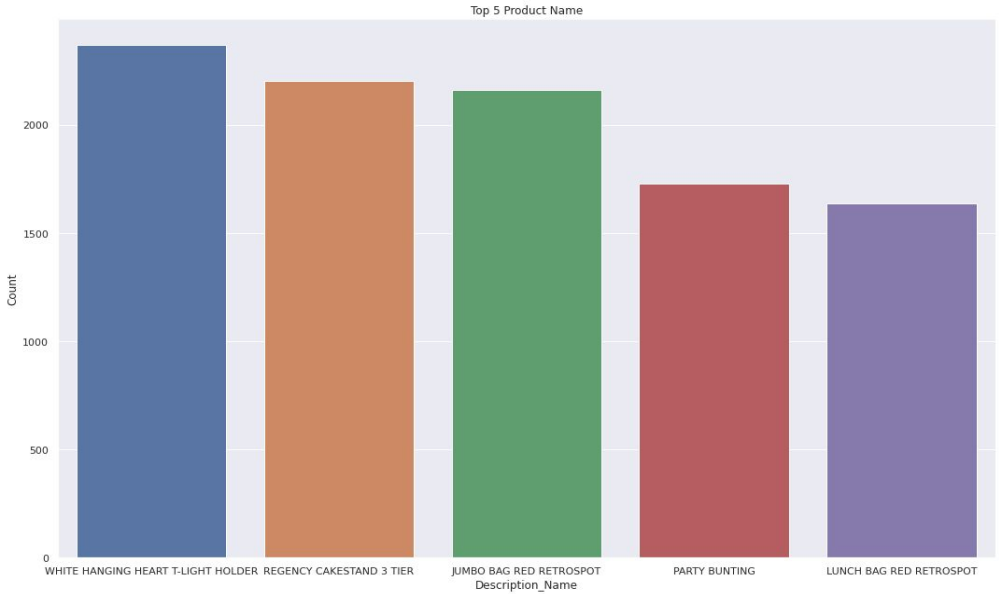
# Data Summary

- 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.**

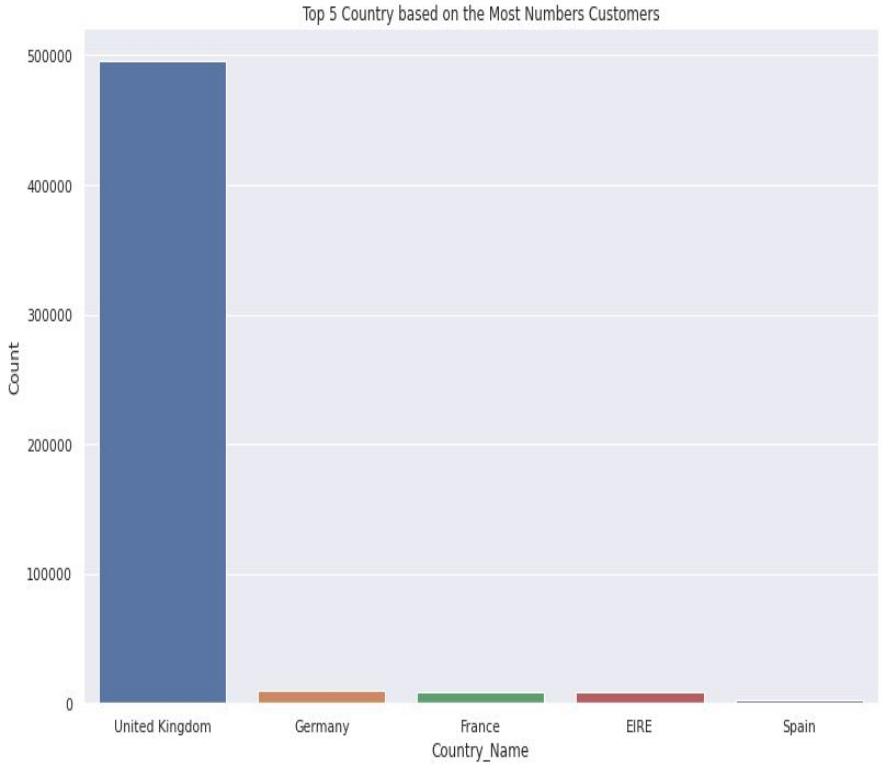
| InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|
| 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 2010-12-01 08:26:00 | 2.55 | 17850.0 | United Kingdom |
| 536365 | 71053 | WHITE METAL LANTERN | 6 | 2010-12-01 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 |

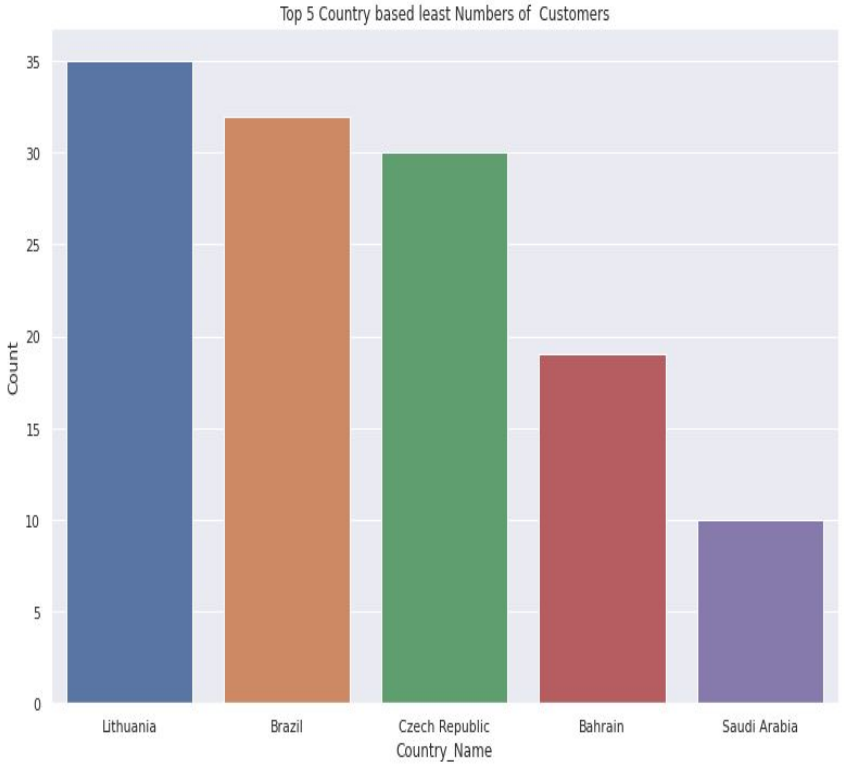
# Cleaning Data Finding the most Purchased Products

| 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

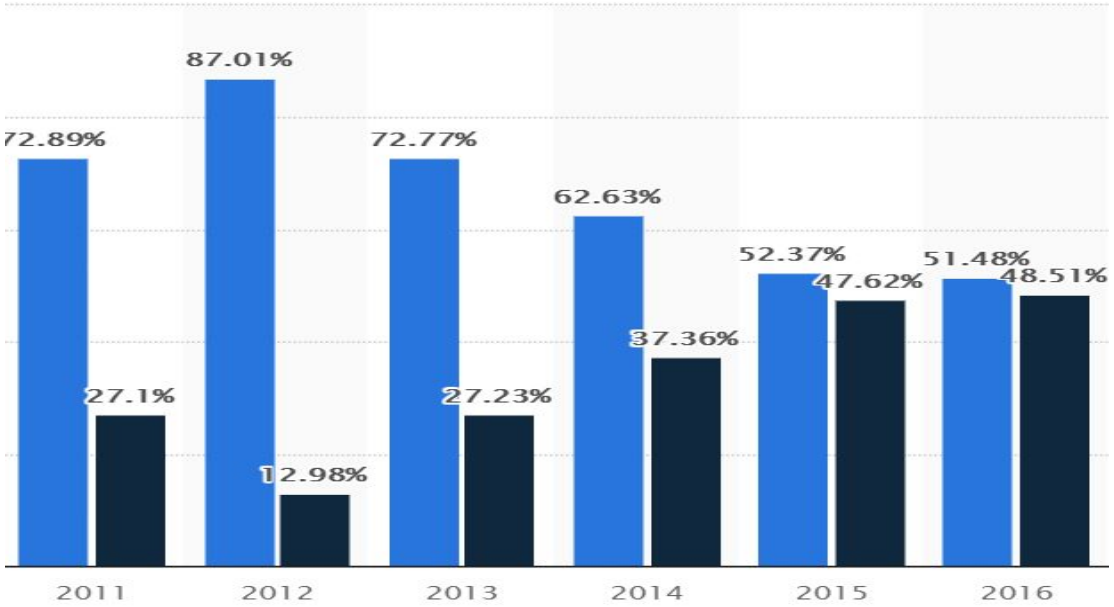


# Analysis

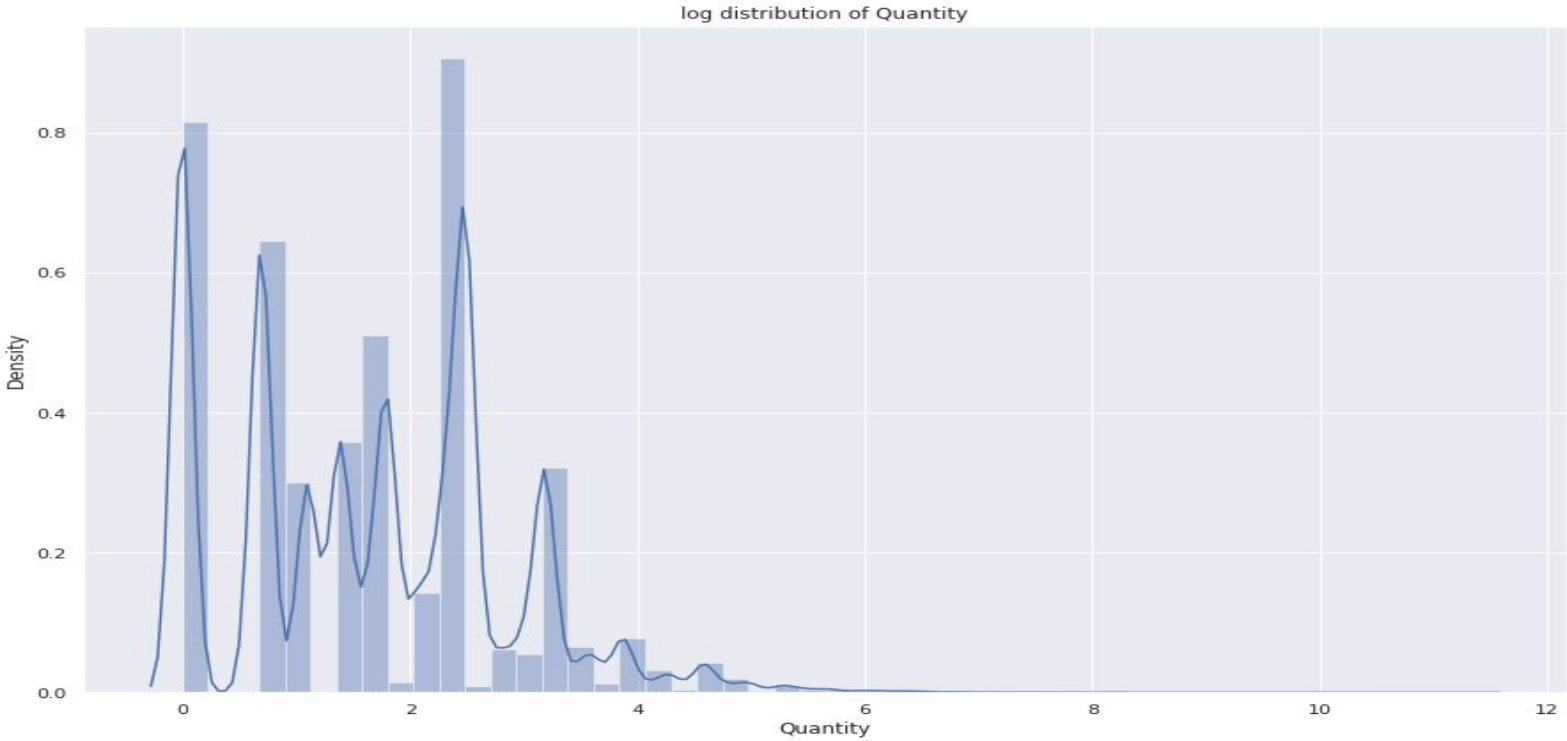
**UK**

**Saudi**

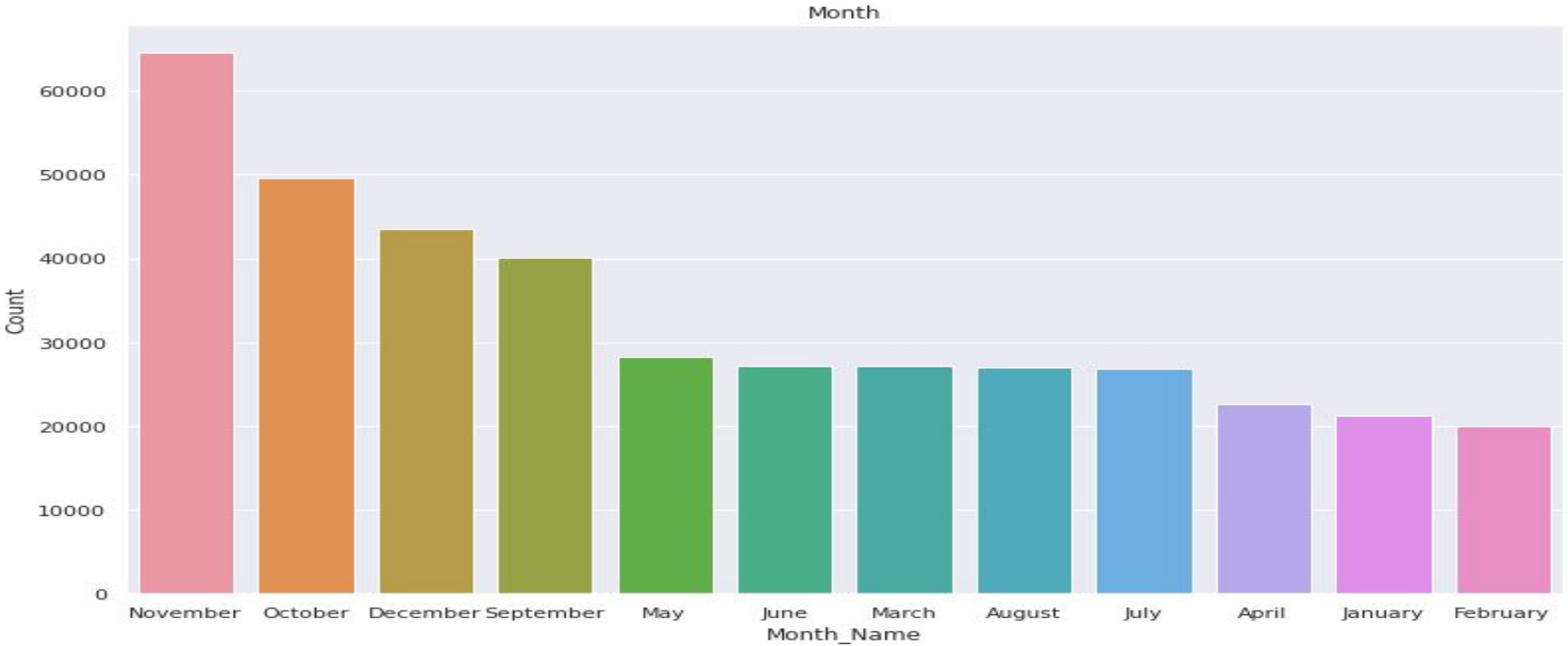



**Source obtained from Statista comparing online purchases from 2011 to 2016**

# Log Distribution of Quantity

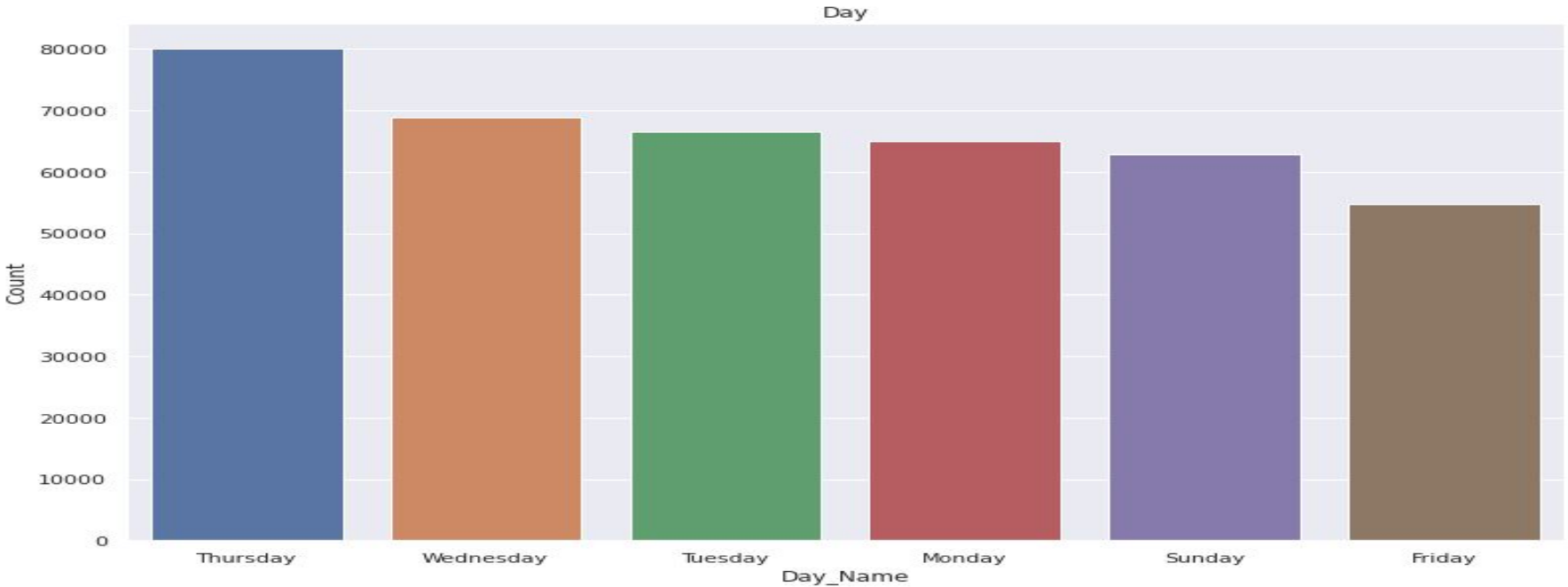


# Month-wise analysis

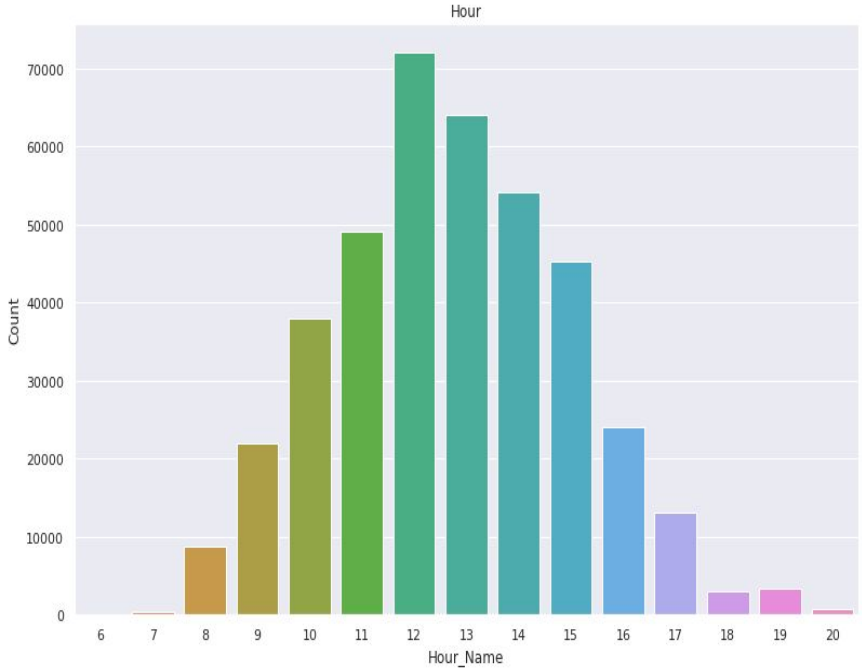


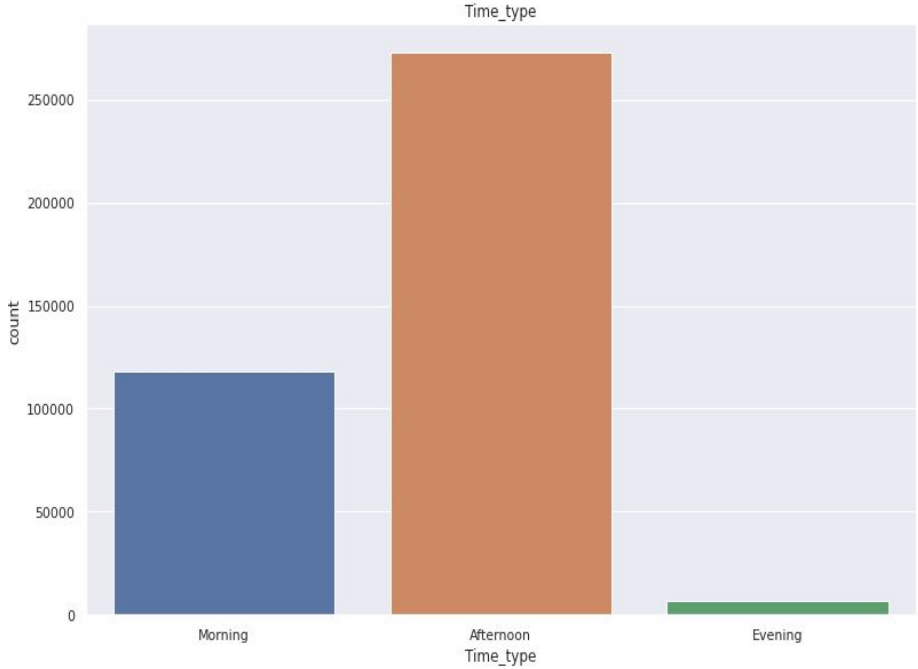

**November and December could be the months with highest sales in anticipation of Christmas**

# Daywise analysis



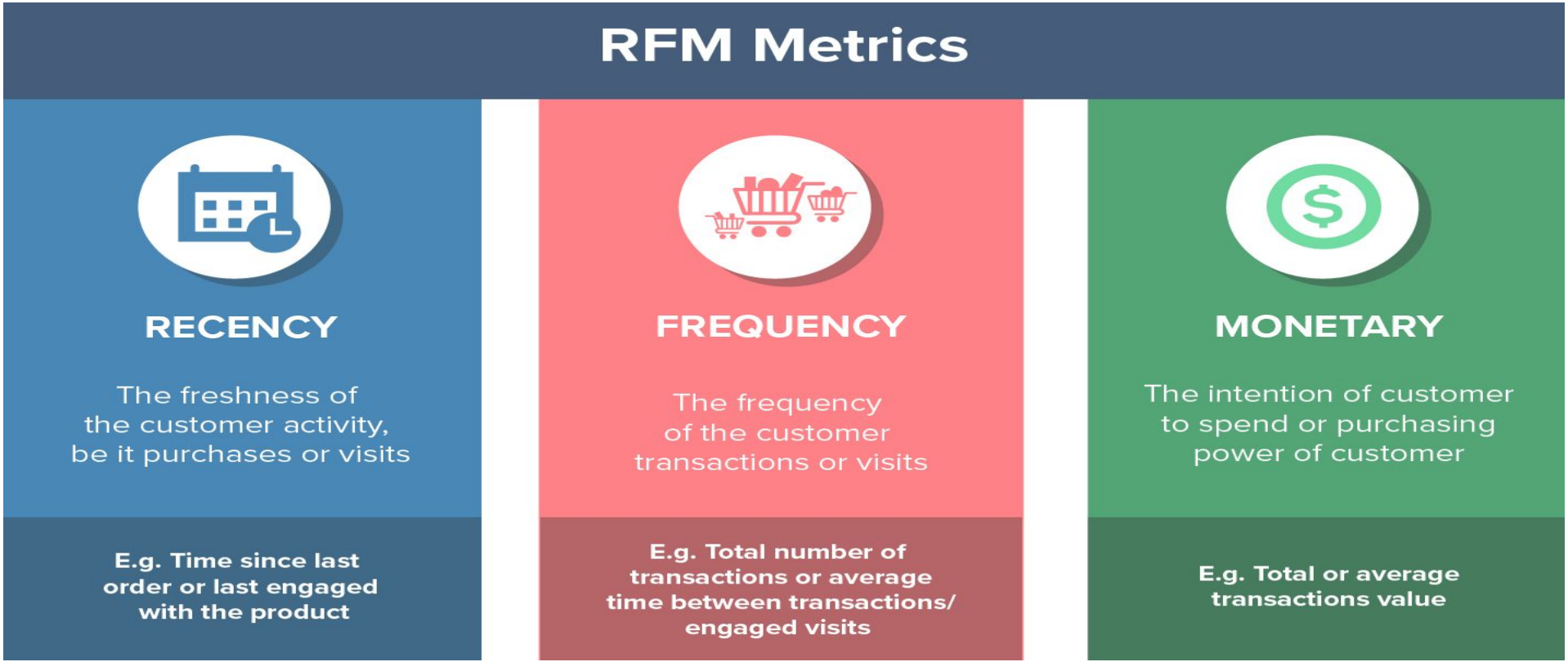
# Hourwise analysis



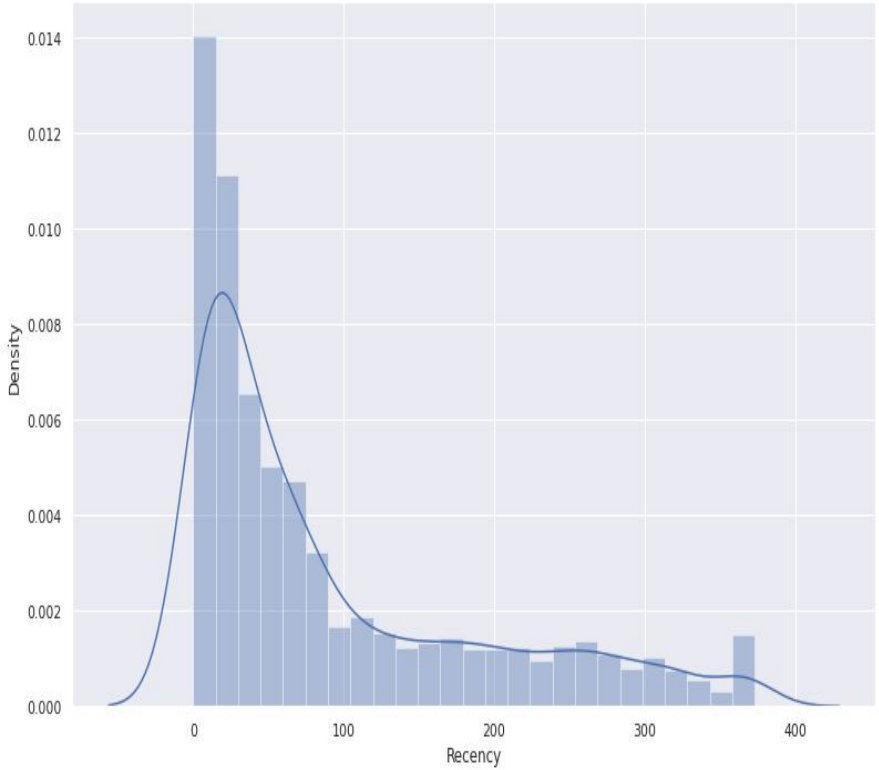



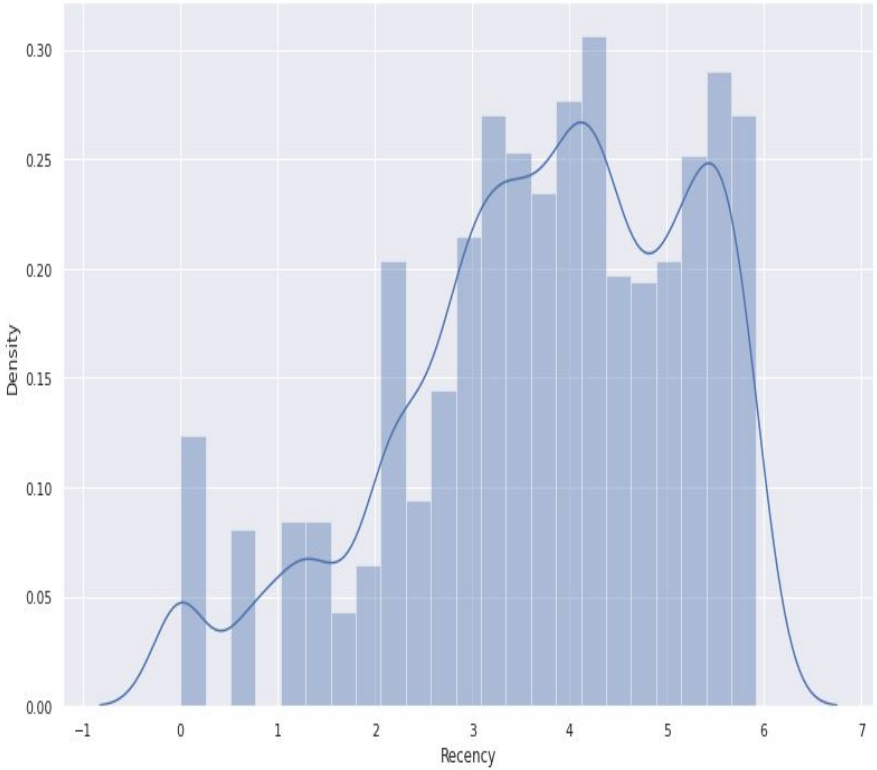

**Working hours witnessing the highest sales could be attributed to the fact that a large part of the dataset is Wholesalers' data**

# Recency,Frequency,Monetary values

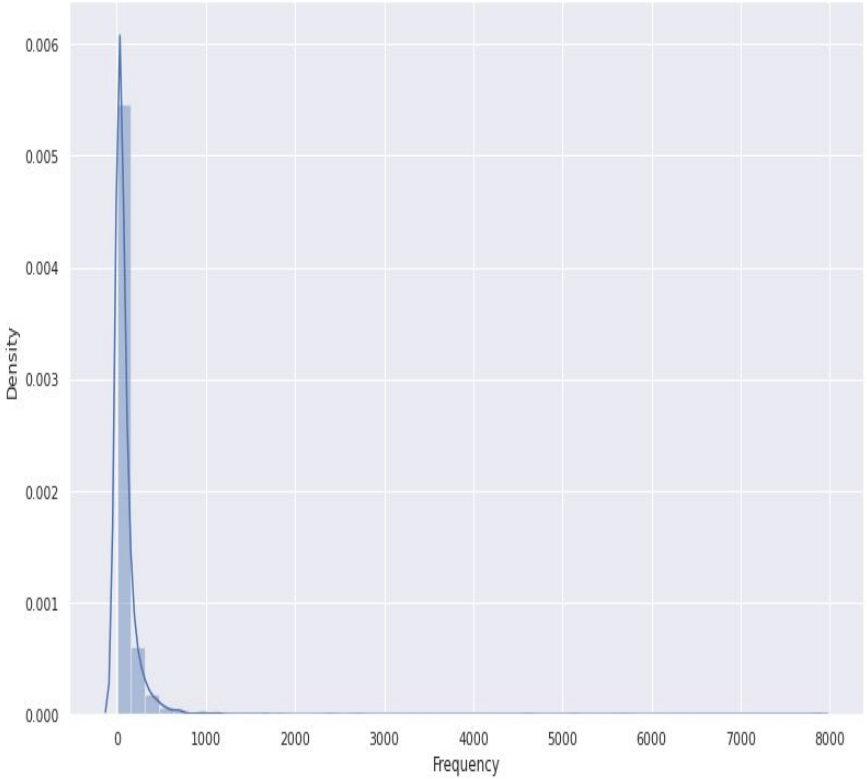


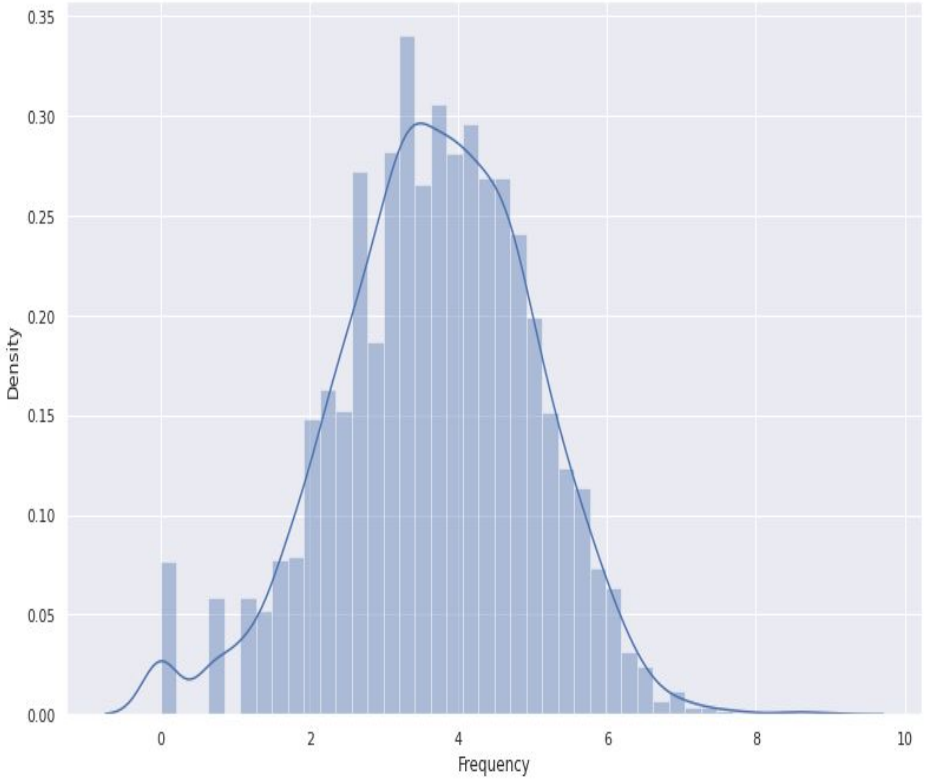
# Recency


0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
100
200
300
400
Recency

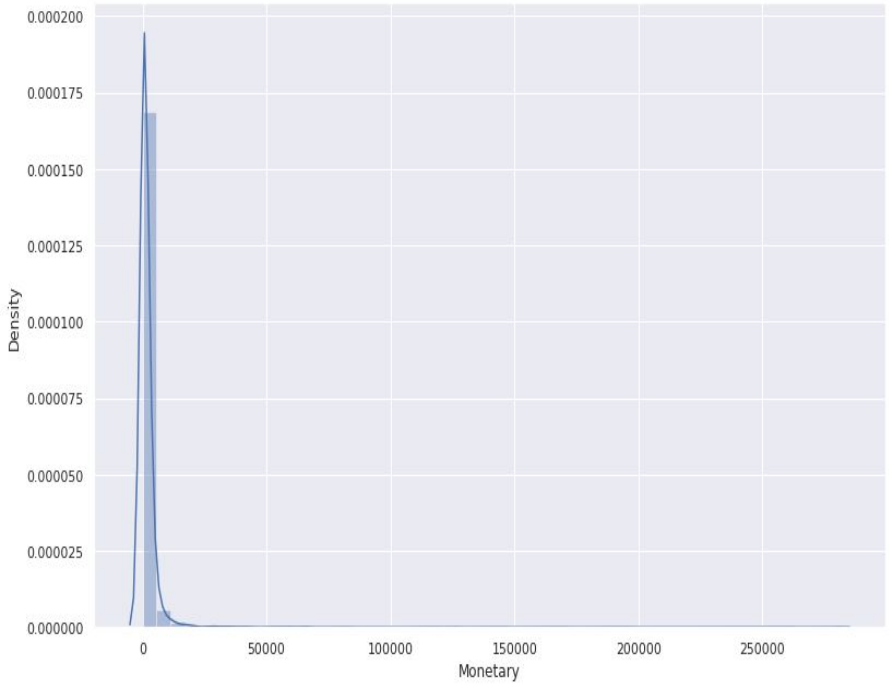

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
−1
0
1
2
3
4
5
6
7
Recency

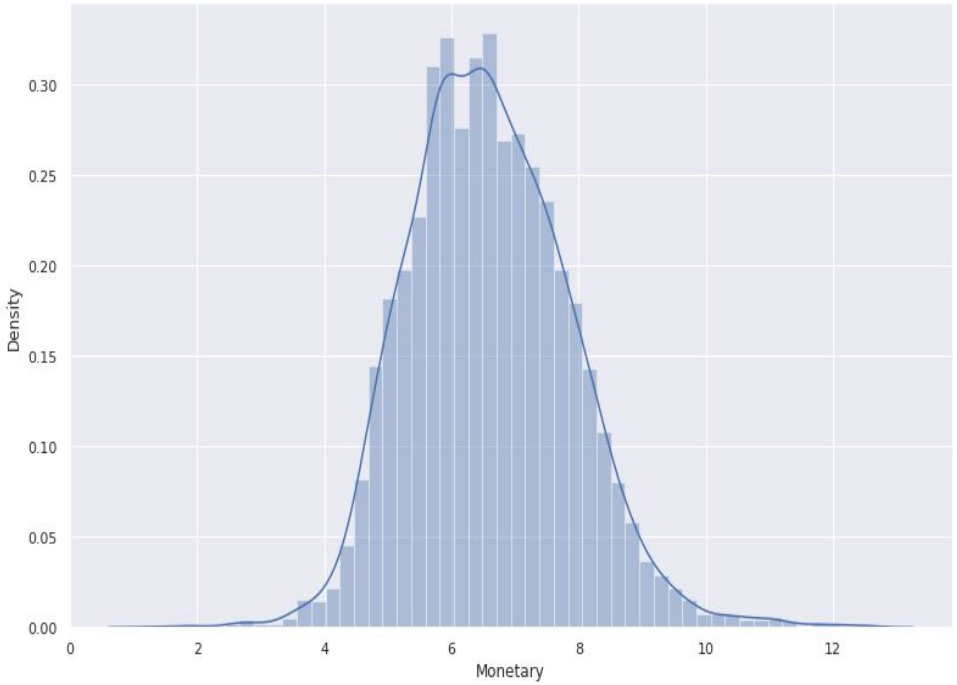
# Frequency


0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
10
Frequency

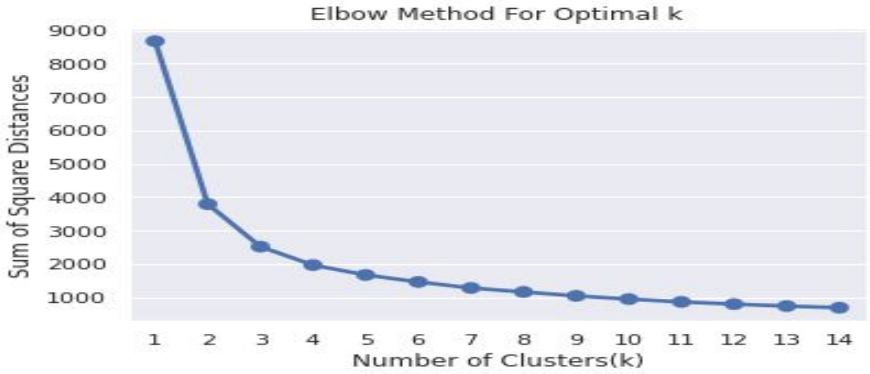
# Monetary

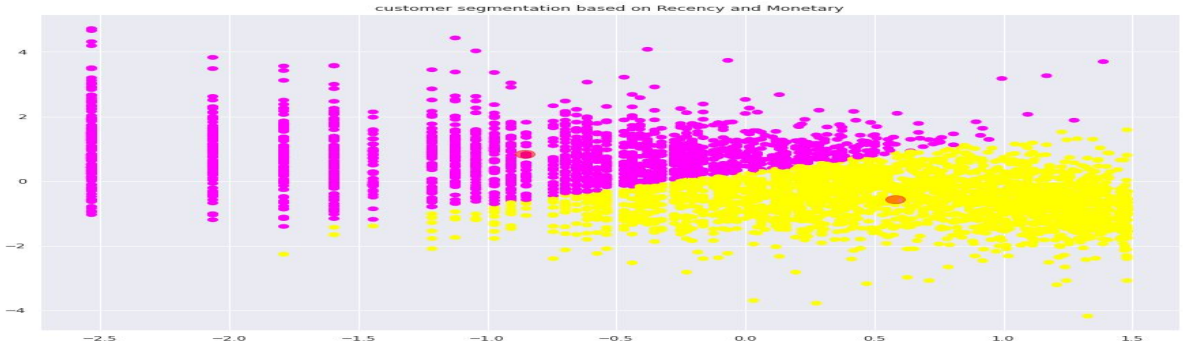

0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
Density
0
50000
100000
150000
200000
250000
Monetary


0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
0
2
4
6
8
10
12
Monetary

# Silhouette score and Elbow method on R&M

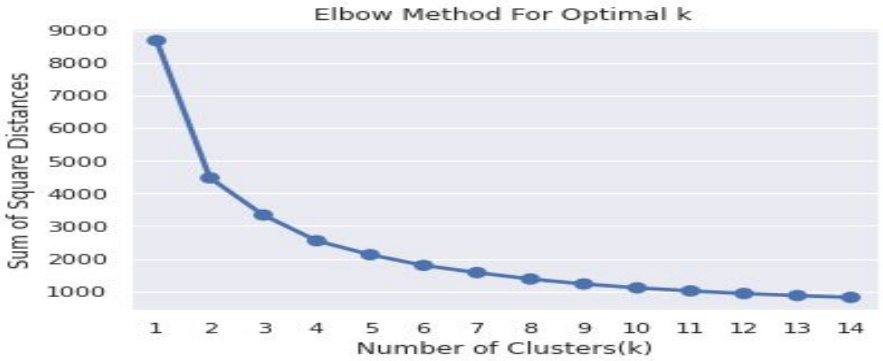
```
For n_clusters = 2, silhouette score is 0.4216081125935063
For n_clusters = 3, silhouette score is 0.3432957775914936
For n_clusters = 4, silhouette score is 0.36494104664274657
For n_clusters = 5, silhouette score is 0.33668503688485785
For n_clusters = 6, silhouette score is 0.34397809419193187
For n_clusters = 7, silhouette score is 0.3458567202377316
For n_clusters = 8, silhouette score is 0.33919727934627264
For n_clusters = 9, silhouette score is 0.3458423886312394
For n_clusters = 10, silhouette score is 0.34850666375861195
For n_clusters = 11, silhouette score is 0.3385166366909024
For n_clusters = 12, silhouette score is 0.3427649471441594
For n_clusters = 13, silhouette score is 0.34083950250492523
For n_clusters = 14, silhouette score is 0.3406096956008792
For n_clusters = 15, silhouette score is 0.34223526314989594
```

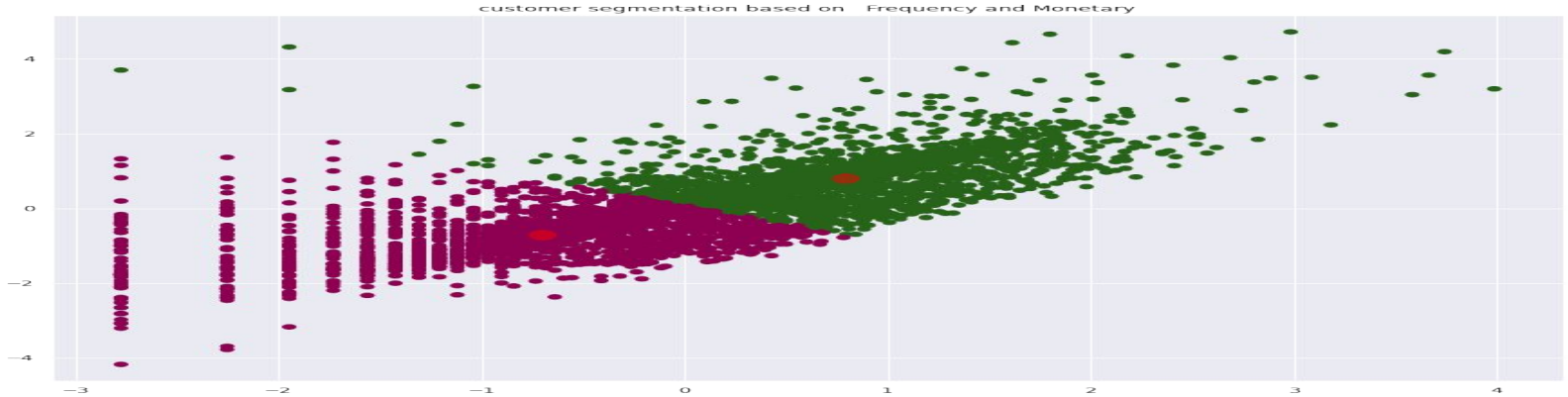


# Silhouette score and Elbow method on F&M

```
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 0.3713782596510203
For n_clusters = 5, silhouette score is 0.34479733808079405
For n_clusters = 6, silhouette score is 0.35974563779013946
For n_clusters = 7, silhouette score is 0.33835032540639154
For n_clusters = 8, silhouette score is 0.3519892091800133
For n_clusters = 9, silhouette score is 0.3460160650521864
For n_clusters = 10, silhouette score is 0.3619887930235607
For n_clusters = 11, silhouette score is 0.36822618560766546
For n_clusters = 12, silhouette score is 0.35460489785135785
For n_clusters = 13, silhouette score is 0.3624674157300161
For n_clusters = 14, silhouette score is 0.36520616987776316
For n_clusters = 15, silhouette score is 0.36101570873847355
```

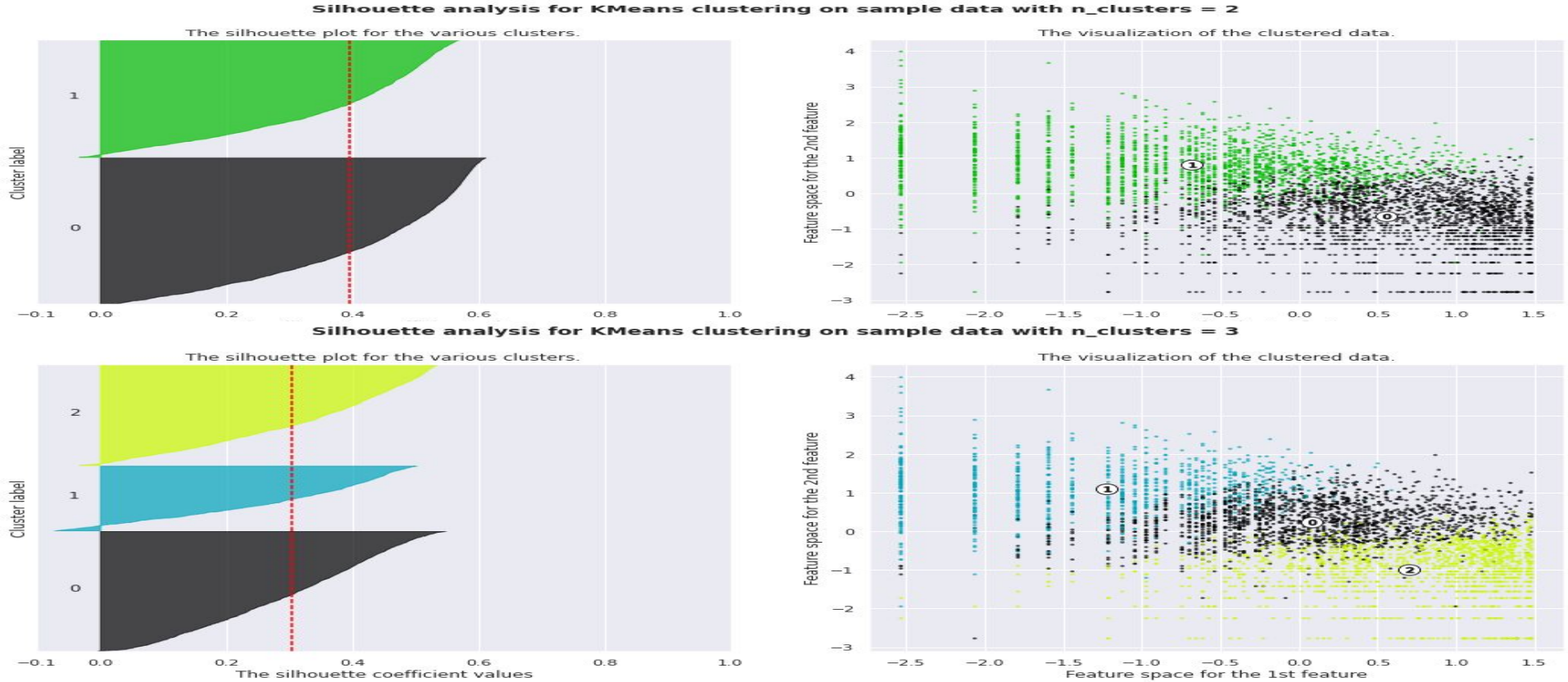


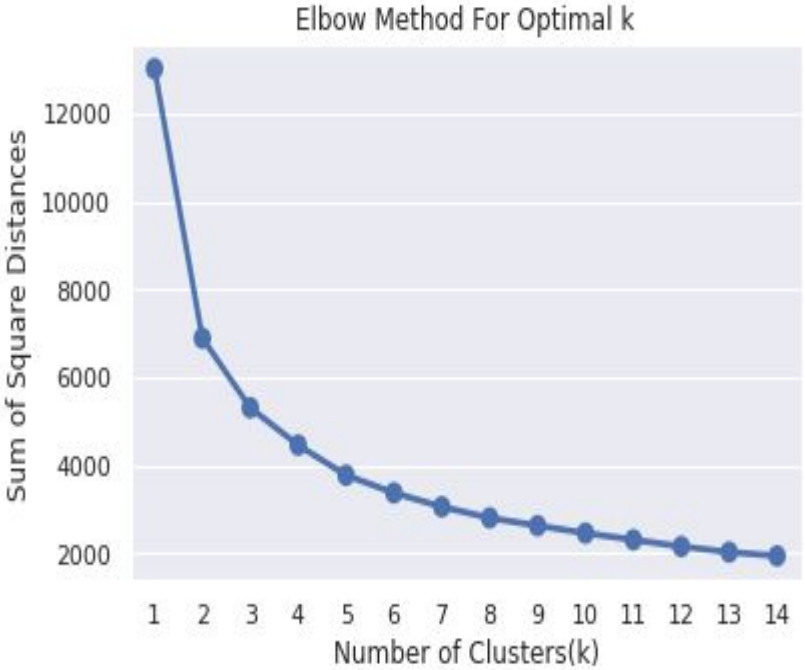
# Silhouette analysis on R, F and M

```
For n_clusters = 2 The average silhouette_score is : 0.395423791756615
For n_clusters = 3 The average silhouette_score is : 0.30310658681490 85
For n_clusters = 4 The average silhouette_score is : 0.302725517496819 86
For n_clusters = 5 The average silhouette_score is : 0.27880346166089 47
For n_clusters = 6 The average silhouette_score is : 0.27854318607070516
For n_clusters = 7 The average silhouette_score is : 0.2623613650755882
For n_clusters = 8 The average silhouette_score is : 0.2638608672365028
For n_clusters = 9 The average silhouette_score is : 0.25878886517568883
For n_clusters = 10 The average silhouette_score is : 0.259477127868534 05
For n_clusters = 11 The average silhouette_score is : 0.25946024250011 22
For n_clusters = 12 The average silhouette_score is : 0.263599810039632 45
For n_clusters = 13 The average silhouette_score is : 0.262169054485507 76
For n_clusters = 14 The average silhouette_score is : 0.26102008903605 79
For n_clusters = 15 The average silhouette_score is : 0.25496577320666 74
```

# Silhouette analysis on RFM



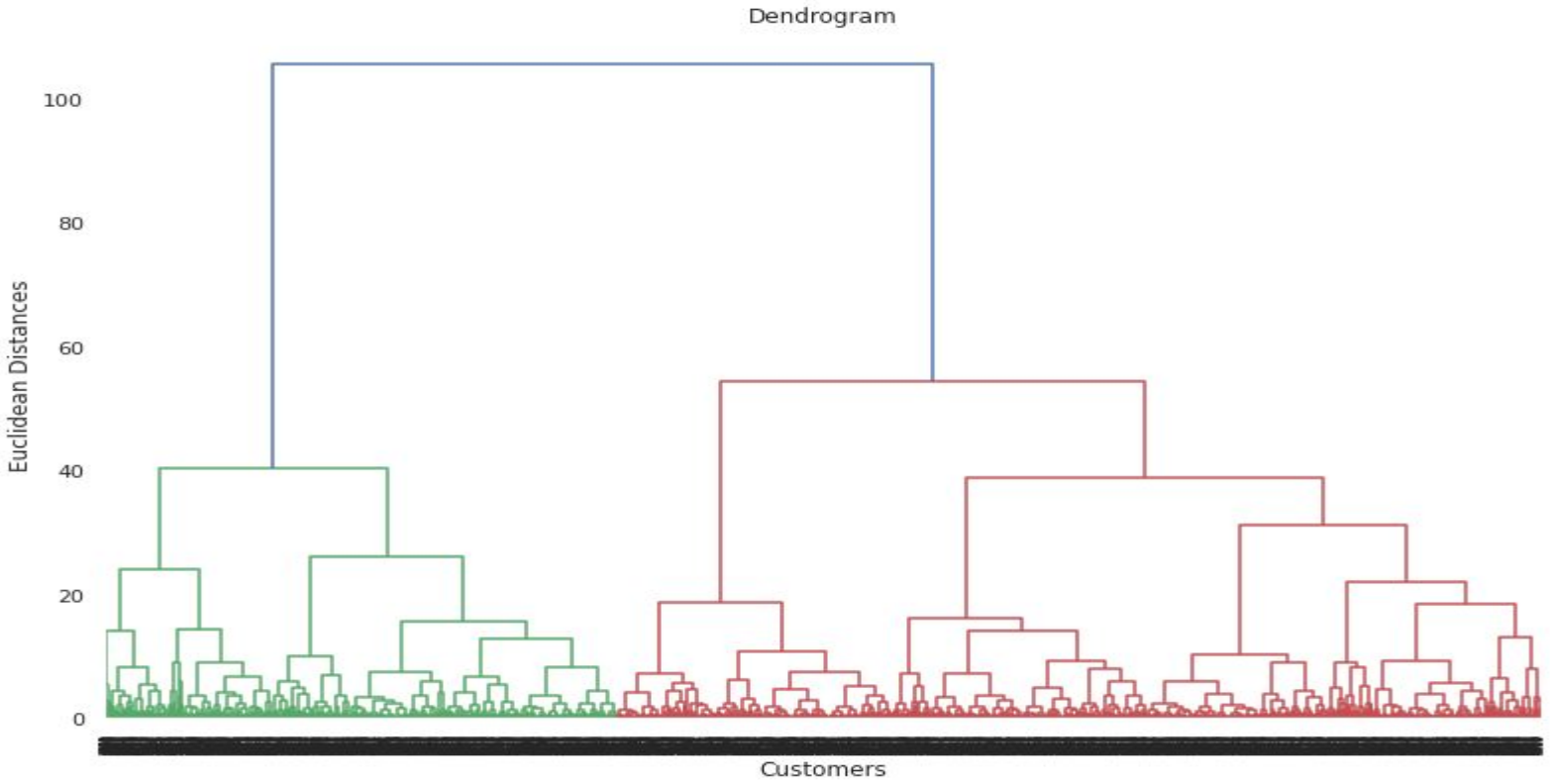
# Elbow method and Cluster chart on RFM



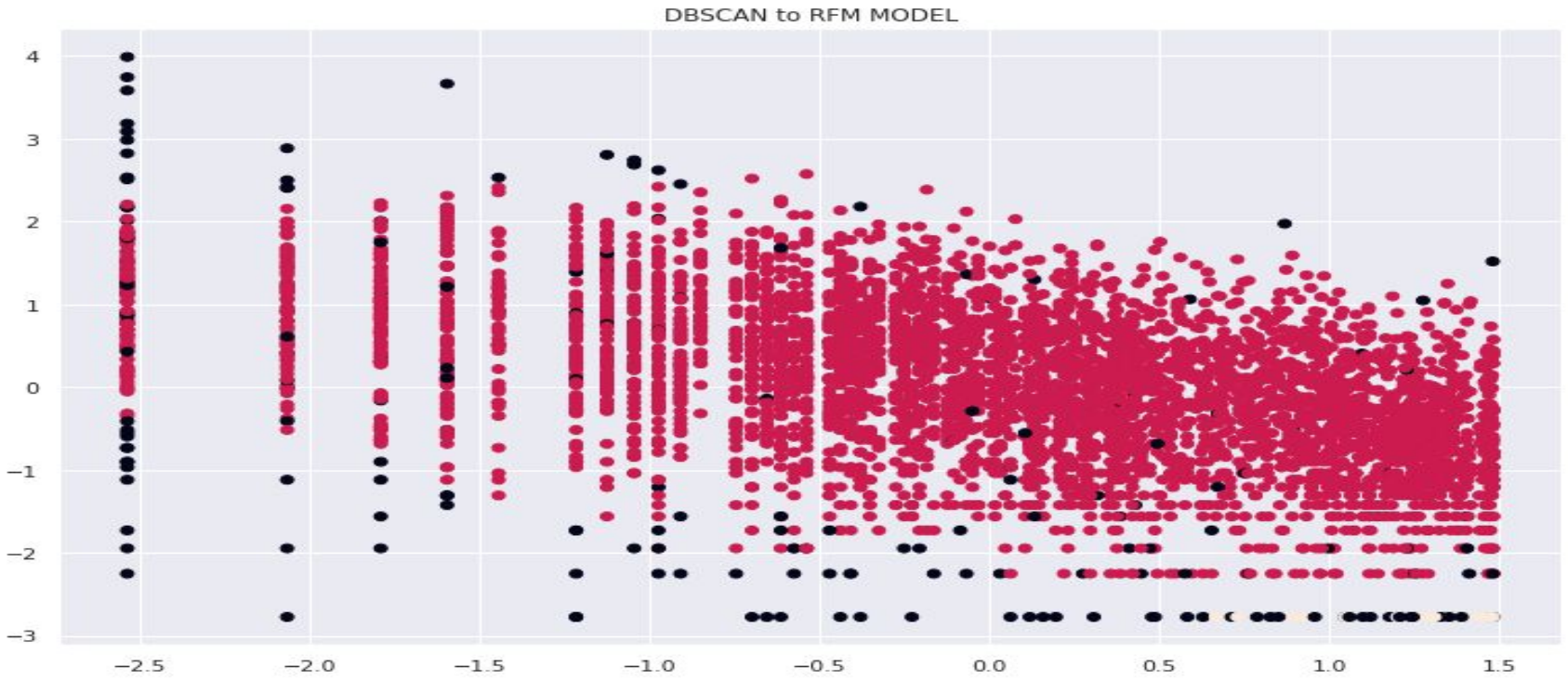

# RFM Analysis

| CustomerID | Recency | Frequency | Monetary | R | F | M | RFMGroup | RFMScore | Recency_log | Frequency_log | Monetary_log | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **12346.0** | 325 | 1 | 77183.60 | 4 | 4 | 1 | 441 | 9 | 5.783825 | 0.000000 | 11.253942 | 1 |
| **12347.0** | 2 | 182 | 4310.00 | 1 | 1 | 1 | 111 | 3 | 0.693147 | 5.204007 | 8.368693 | 0 |
| **12348.0** | 75 | 31 | 1797.24 | 3 | 3 | 1 | 331 | 7 | 4.317488 | 3.433987 | 7.494007 | 1 |
| **12349.0** | 18 | 73 | 1757.55 | 2 | 2 | 1 | 221 | 5 | 2.890372 | 4.290459 | 7.471676 | 0 |
| **12350.0** | 310 | 17 | 334.40 | 4 | 4 | 3 | 443 | 11 | 5.736572 | 2.833213 | 5.812338 | 1 |
| **12352.0** | 36 | 85 | 2506.04 | 2 | 2 | 1 | 221 | 5 | 3.583519 | 4.442651 | 7.826459 | 0 |
| **12353.0** | 204 | 4 | 89.00 | 4 | 4 | 4 | 444 | 12 | 5.318120 | 1.386294 | 4.488636 | 1 |
| **12354.0** | 232 | 58 | 1079.40 | 4 | 2 | 2 | 422 | 8 | 5.446737 | 4.060443 | 6.984161 | 1 |
| **12355.0** | 214 | 13 | 459.40 | 4 | 4 | 3 | 443 | 11 | 5.365976 | 2.564949 | 6.129921 | 1 |
| **12356.0** | 22 | 59 | 2811.43 | 2 | 2 | 1 | 221 | 5 | 3.091042 | 4.077537 | 7.941449 | 0 |

# 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 |