Capstone Project

Online Retail Customer Segmentation

Vijay N

Steps Performed

1. Defining the problem statement
2. Data Exploration and Preparation of dataset
3. Exploratory Data Analysis
4. Applying the Model
5. Model Selection and Conclusion

Problem Statement

We have 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.

Data Exploration

Our dataset has 541909 rows and 8 columns to begin with. In the preprocessing we have added 4 more columns.

The columns of the dataset are as follows:

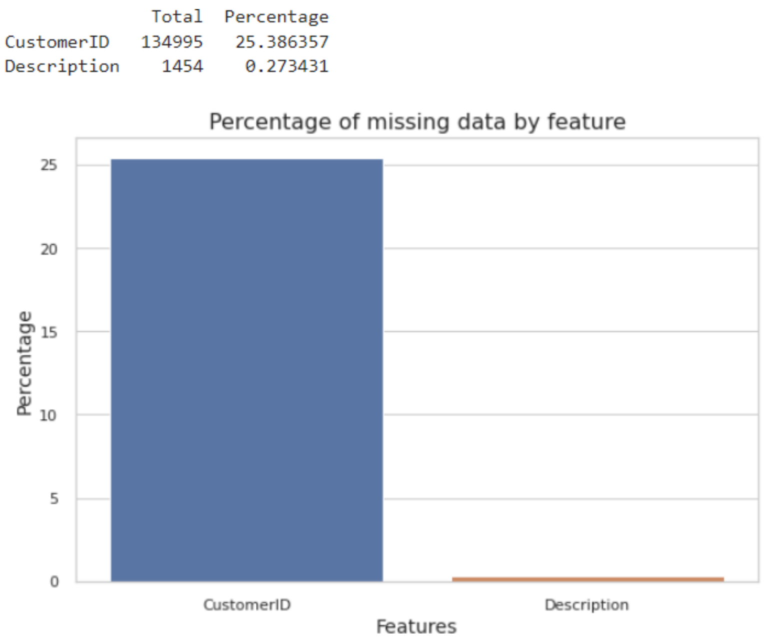
- **InvoiceNo:** Invoice number is a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode:** It is a 5-digit integral number uniquely assigned to each distinct product. It can also be called as product code.
- **Description:** This describes the product name.
- **Quantity:** The quantities of each product (item) per transaction.
- **InvoiceDate:** This specifies the day and time when each transaction was generated.
- **UnitPrice:** Price of product per unit.
- **CustomerID:** It is a 5-digit integral number uniquely assigned to each customer.
- **Country:** Specifies the name of the country where the customer resides.

Data Exploration (Contd.)

The calculated columns created during preprocessing are as follows:

- **TotalAmount** - Created by combining unit price and quantity. This gives information of total amount spent during that particular transaction.
- **month** – month value extracted from InvoiceDate column
- **day** – day name value extracted from InvoiceDate column.
- **hour** – hour value extracted from InvoiceDate column.

Spread of Missing values



```
              Total  Percentage
CustomerID   134995   25.386357
Description    1454    0.273431
```

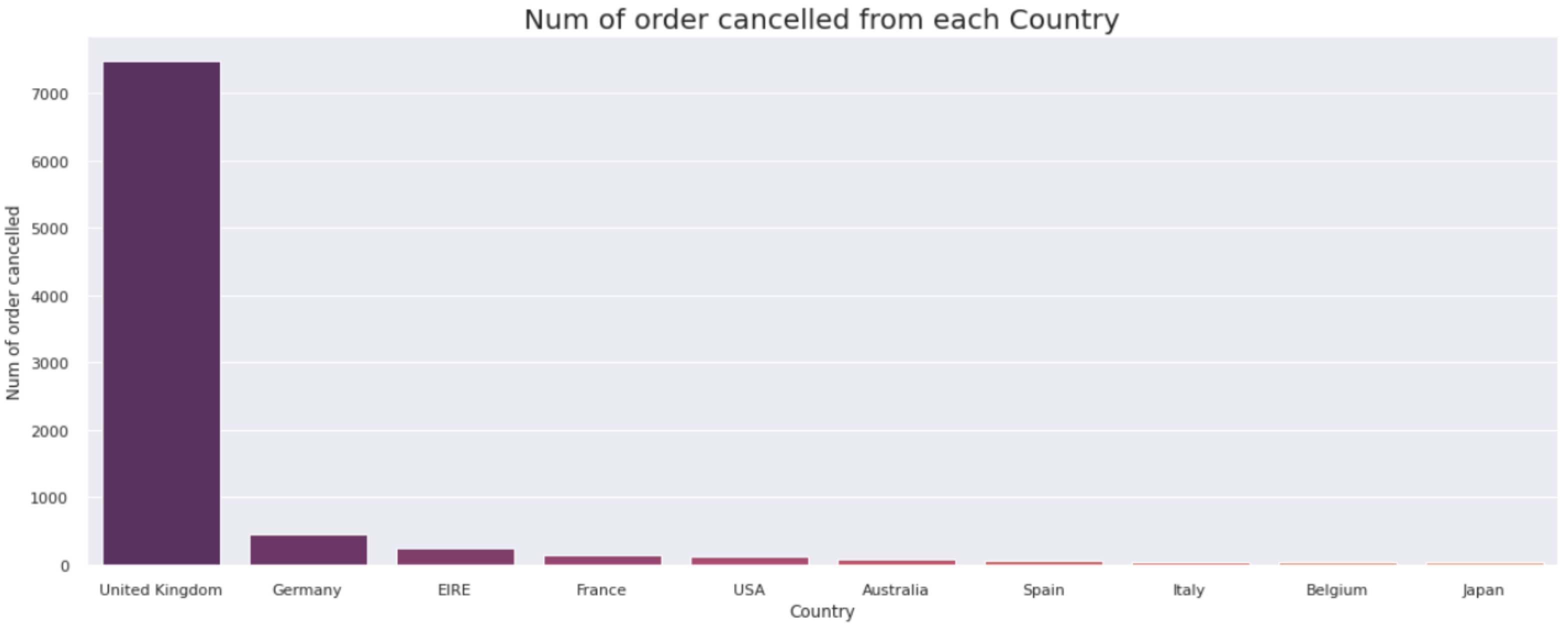
We can obseíve that 25% of CustomeíID values aíe missing.

ľhis can be handled in multiple ways. One such way is by imputing it with a íandom numbeí.
It is possible to impute customeíID based on unique value of InvoiceNo, but theíe will be a big inaccuíacy in matching cancelled tíansactions. Because we have obseíved that the coííesponding puíchase and cancelled tíansaction do not have same InvoiceNo.

And this kindof puíchase will be likely to be a one-time puíchase as customeís who shop fíequently would píobably cíeate an account foí ease of puíchasing.

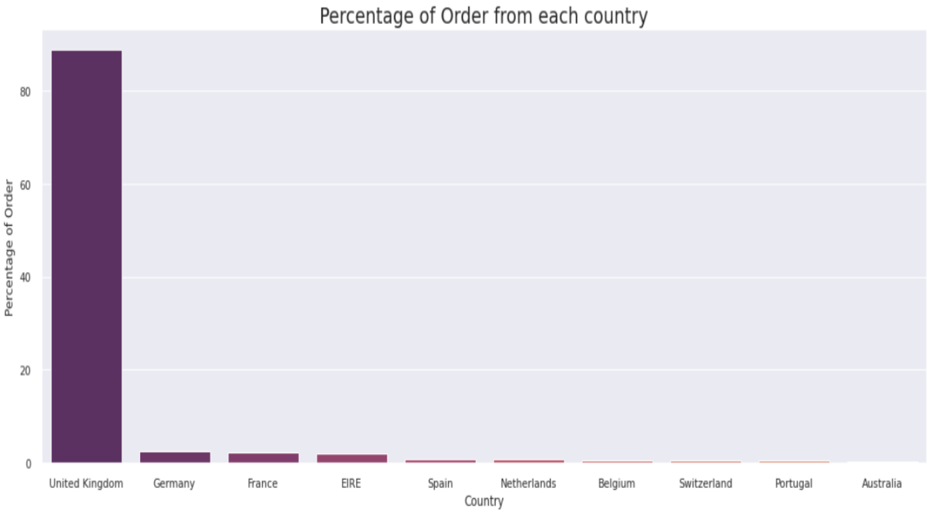
So, we have díopped those missing values.

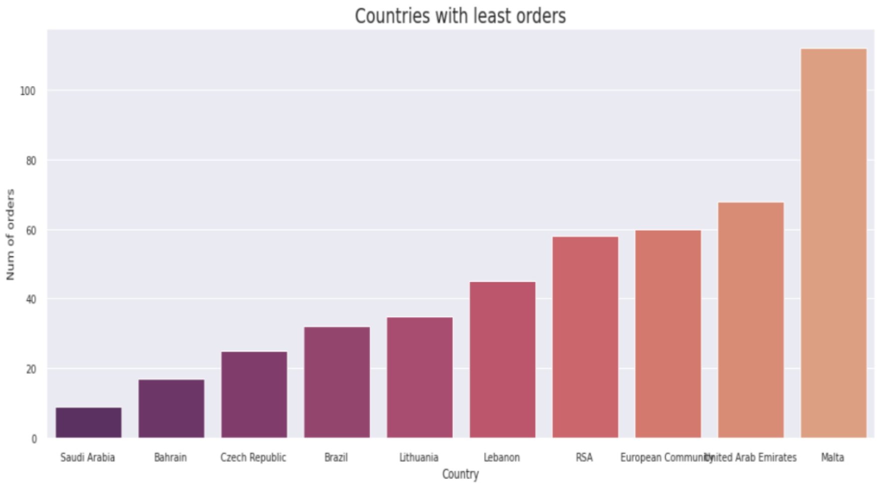
Exploratory Data Analysis



Maximum number of order cancellation done from United Kingdom.

Exploratory Data Analysis (contd.)

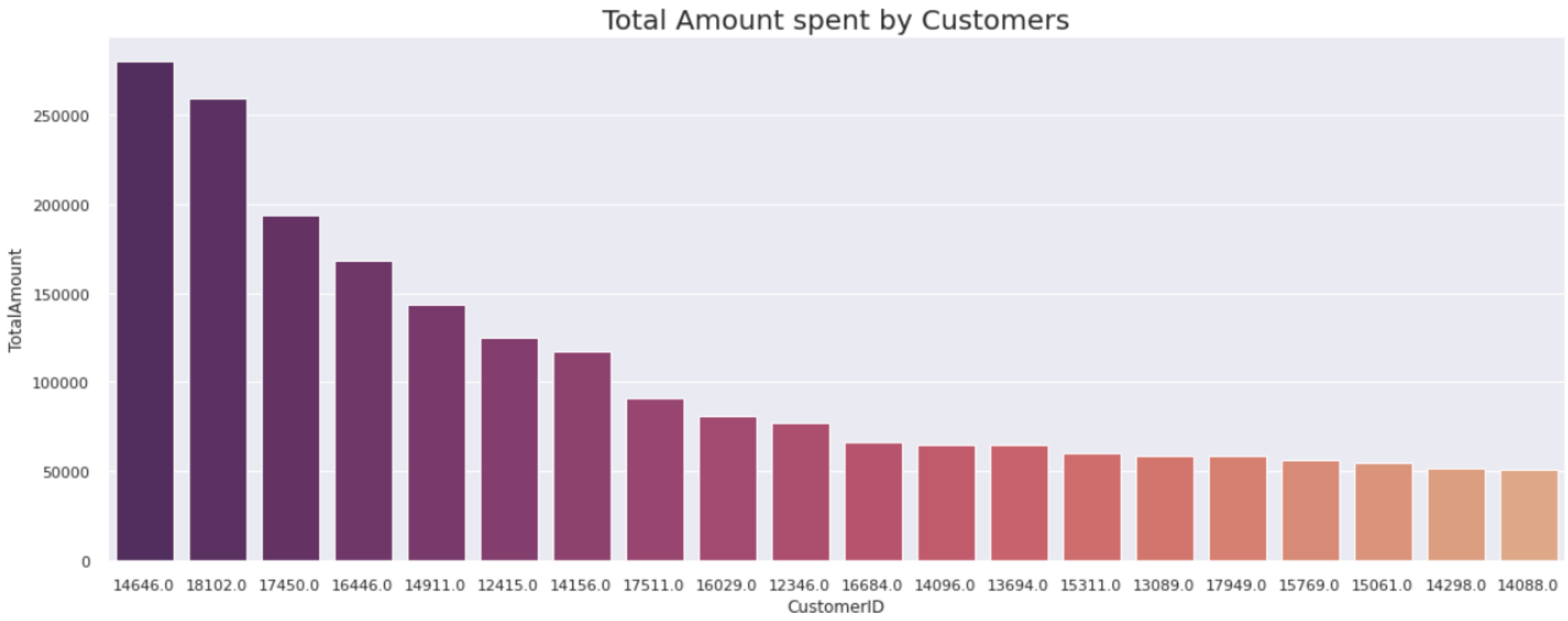




About 88.8% of orders are coming from UK, so we can say that most customers and most orders will be from United Kingdom.

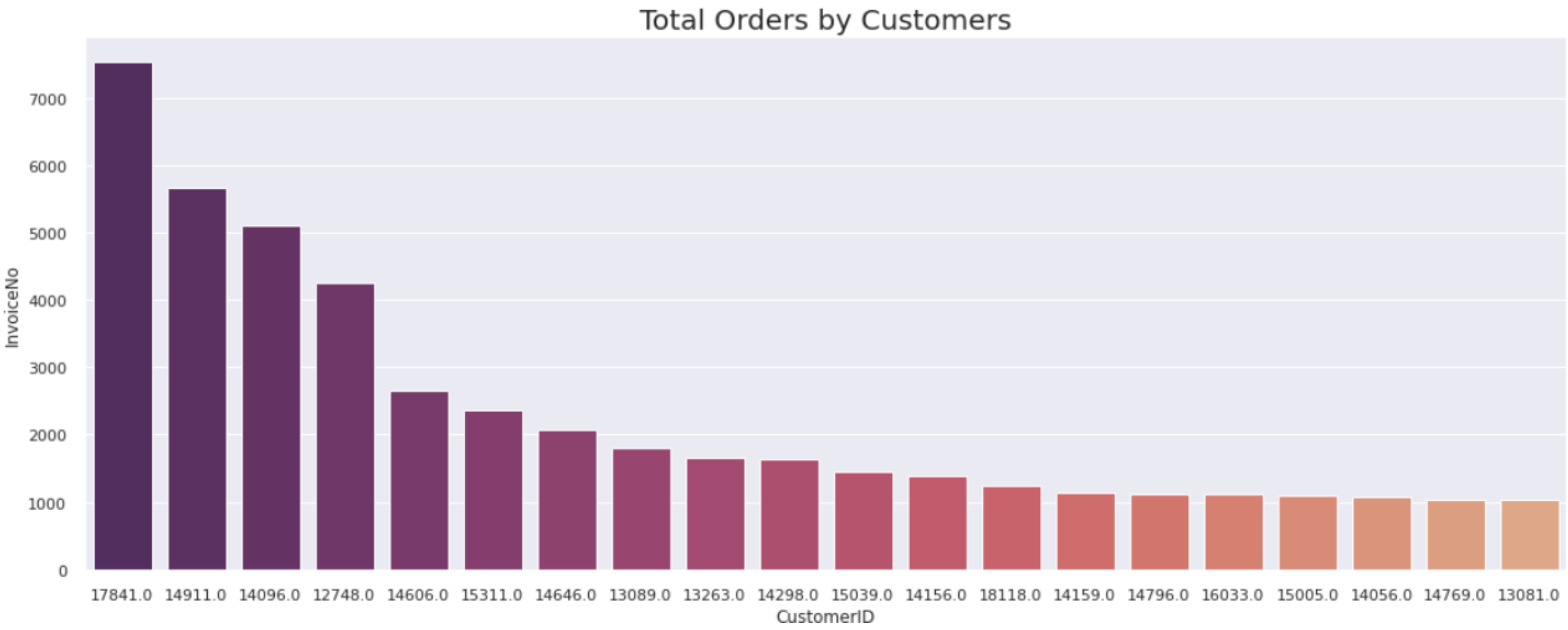
We have least number of orders and customers from Saudi Arabia.

Exploratory Data Analysis (contd.)



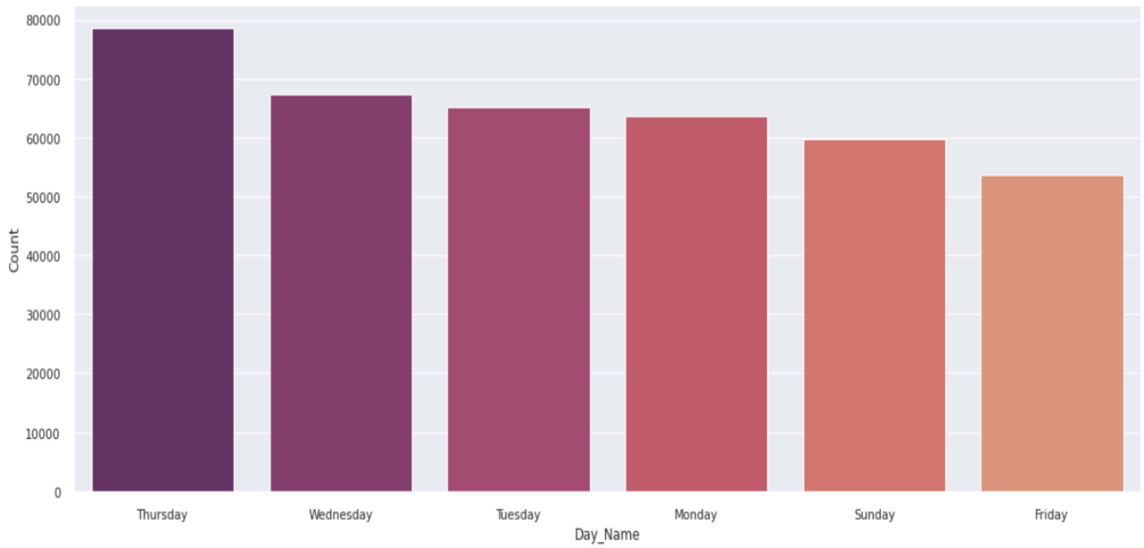
CustomerID - 14646, spends more money on shopping.

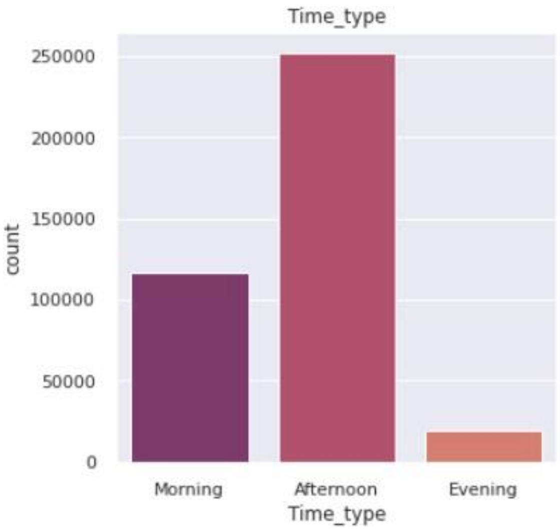
Exploratory Data Analysis (contd.)



CustomerID - 1784l is a shopaholic, who shops/orders more.

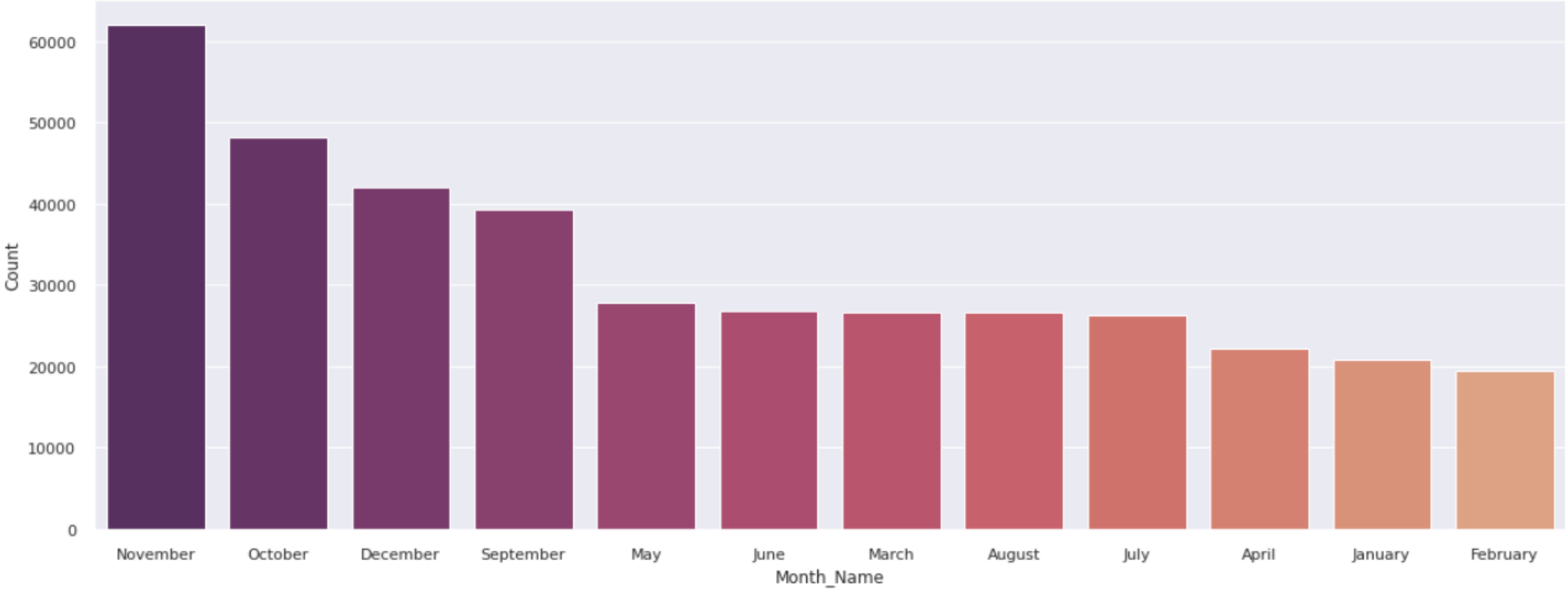
Exploratory Data Analysis (contd.)





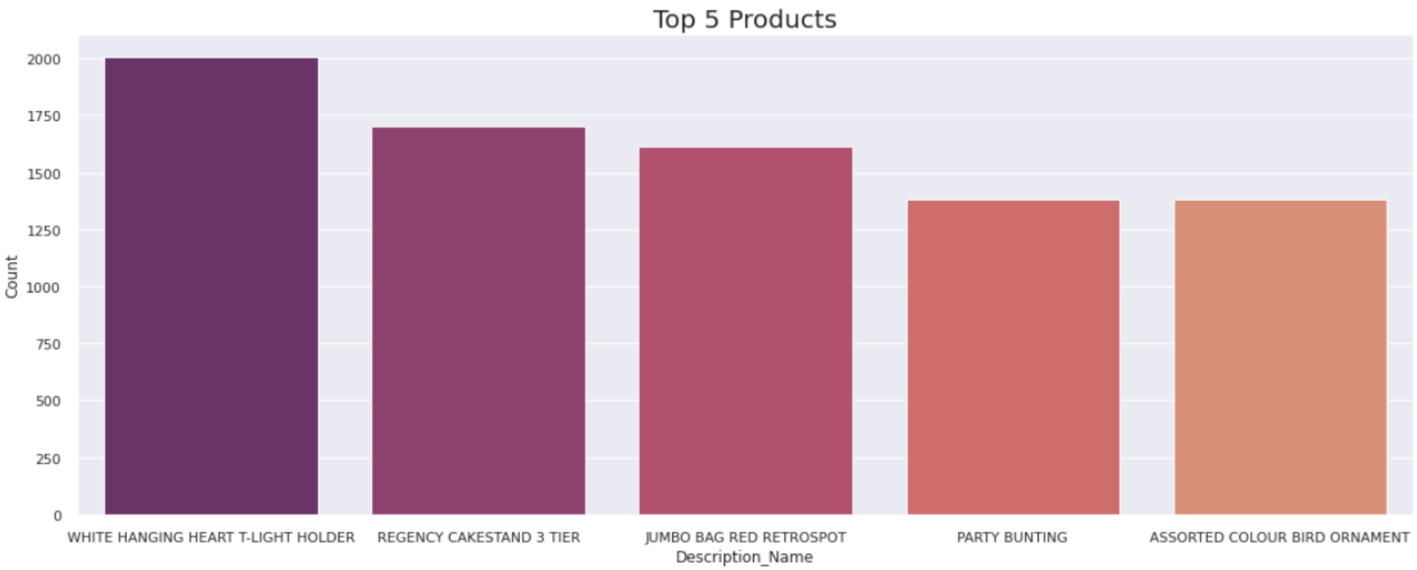
Shopping will happen mostly on weekdays and it will be more during afternoon.

Exploratory Data Analysis (contd.)



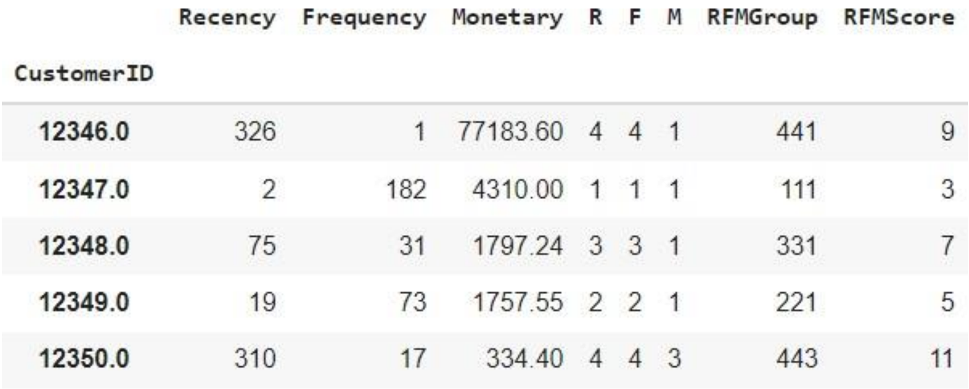
Shopping will be more during the festival season.

Exploratory Data Analysis (contd.)



Applying the Model

RFM Model



	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
CustomerID								
12346.0	326	1	77183.60	4	4	1	441	9
12347.0	2	182	4310.00	1	1	1	111	3
12348.0	75	31	1797.24	3	3	1	331	7
12349.0	19	73	1757.55	2	2	1	221	5
12350.0	310	17	334.40	4	4	3	443	11

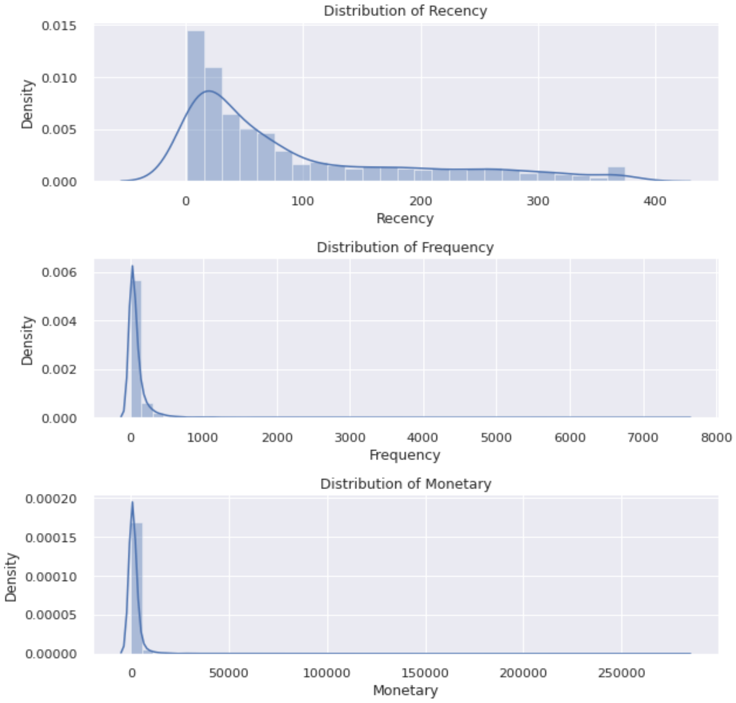
Best Customers– This group consists of those customers who are found in R-Tier-1, F-Tier-1 and M-Tier-1 i.e.,1-1-1, meaning that they transacted recently, do so often and spend more than other customers.

High-spending New Customers– This group consists of those customers in 1-4-1 and 1-4-2. These are customers who transacted only once, but very recently and they spent a lot.

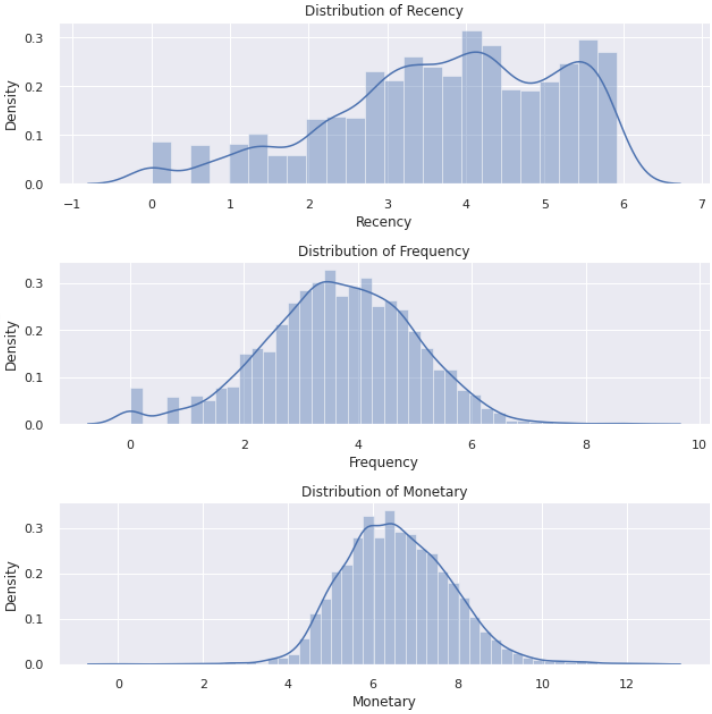
Lowest-Spending Active Loyal Customers – This group consists of those customers in segments 1-1-3 and 1-1-4 (they transacted recently and do so often, but spend the least).

Applying the Model (contd.)

RFM Model



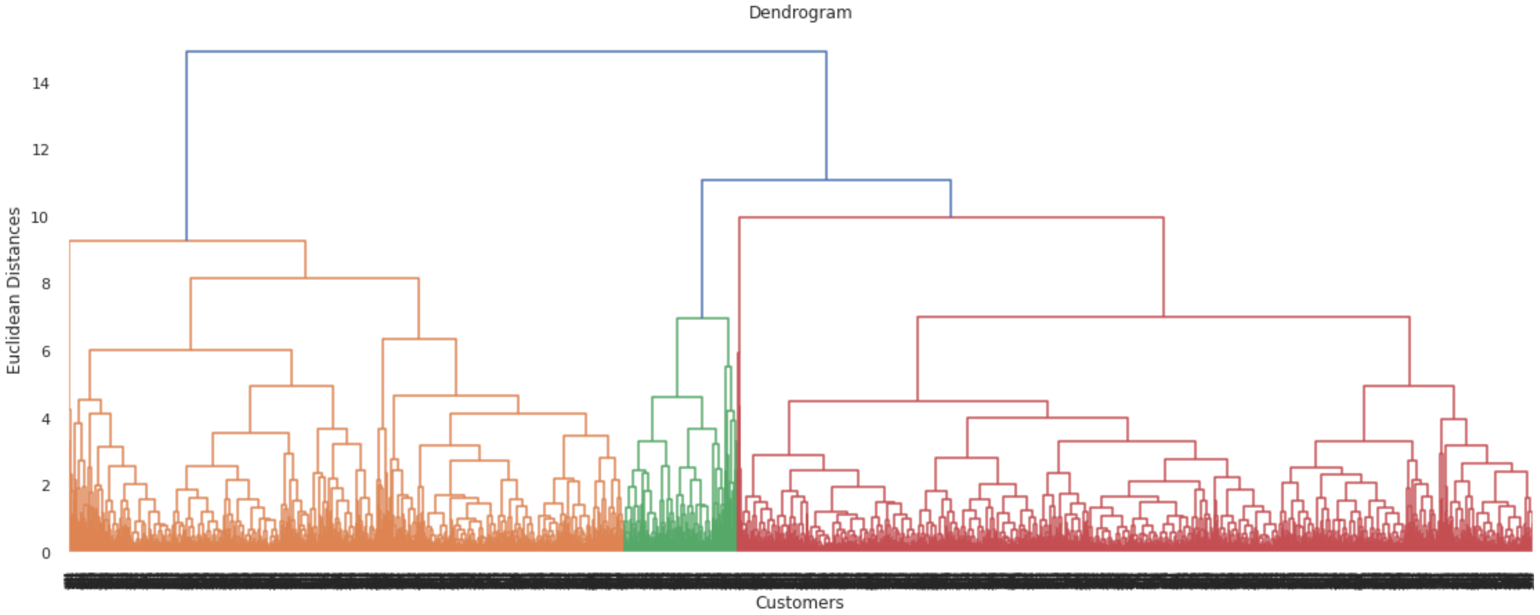
log



Applying the Model (contd.)

Hierarchical Clustering

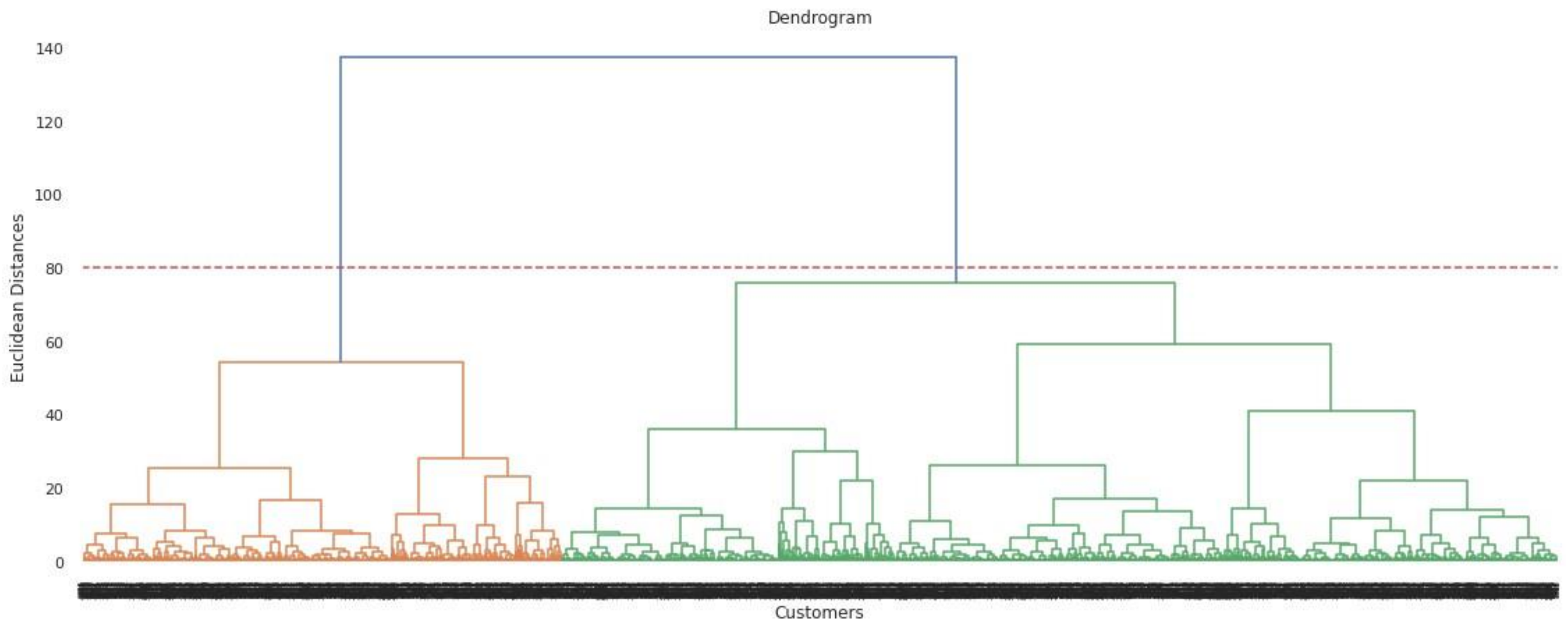
Complete Linkage:



Applying the Model (contd.)

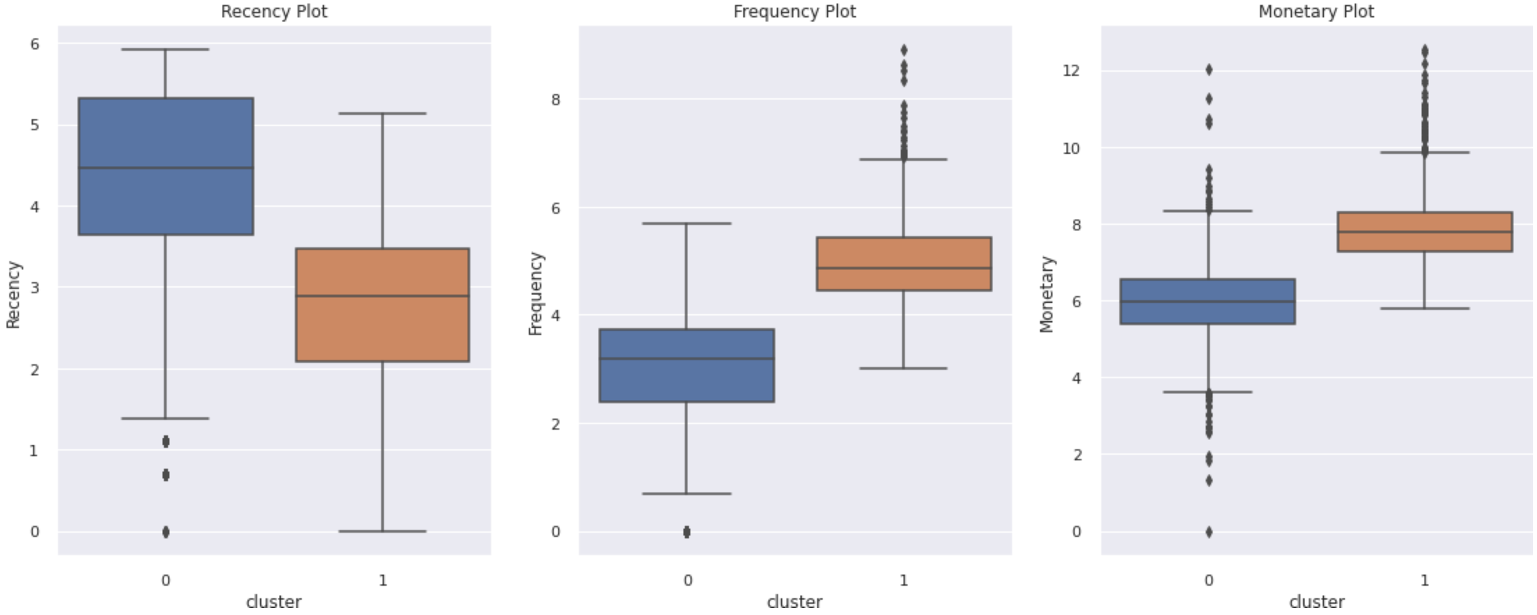
Hierarchical Clustering

Ward’s Linkage:



Applying the Model (contd.)

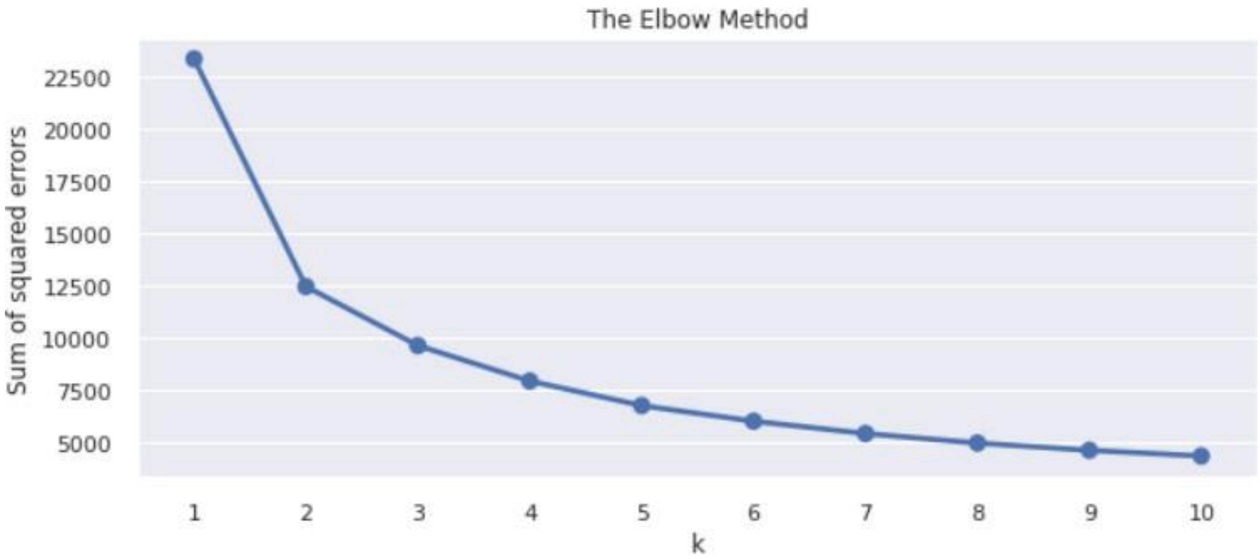
Hierarchical Clustering



Applying the Model (contd.)

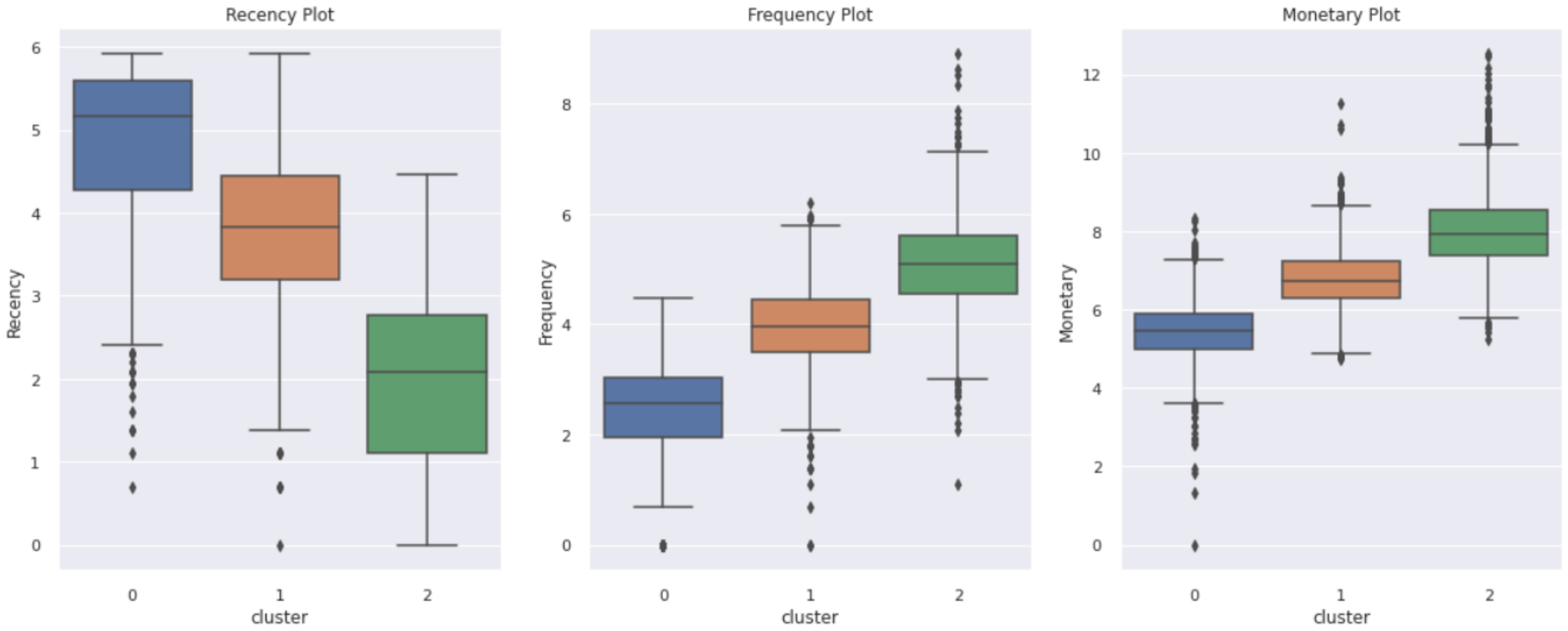
Kmeans Clustering

To identify ‘K’ value, we will be using Silhouette score and Elbow method.



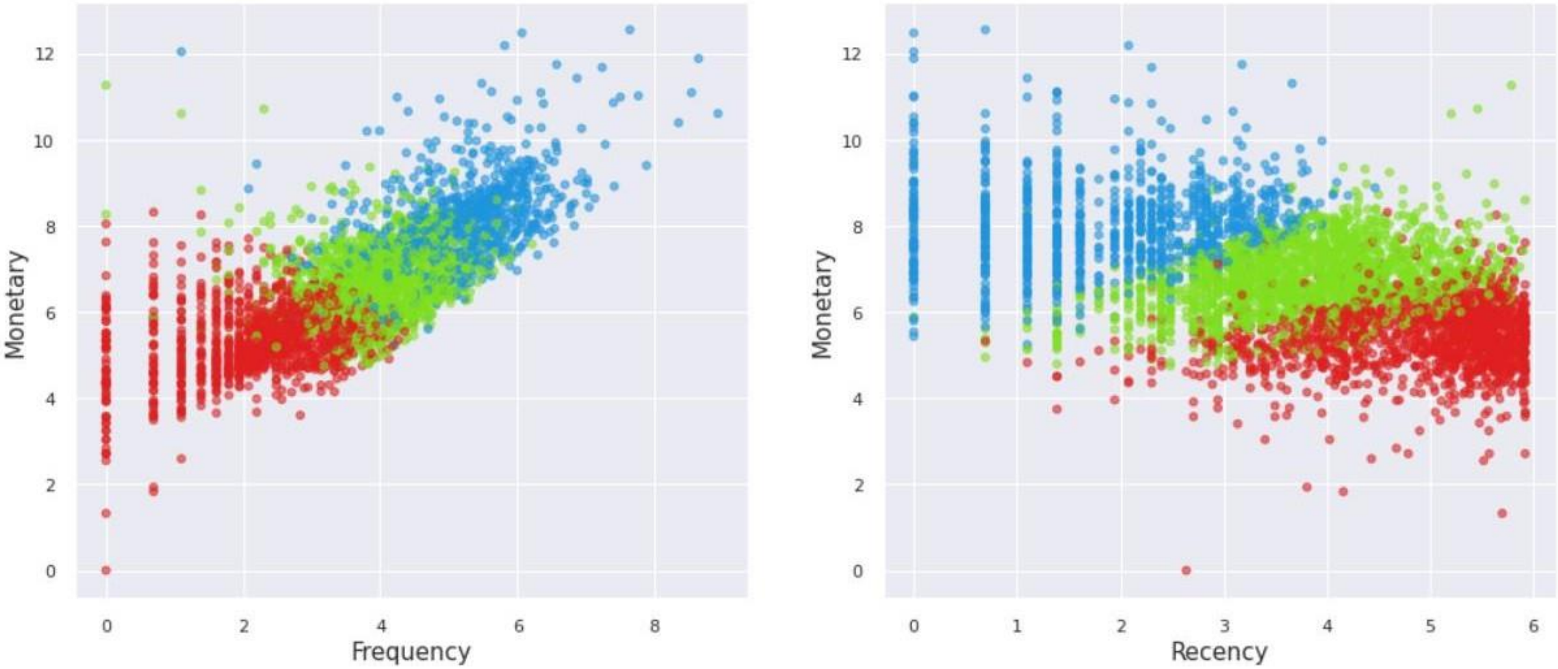
Applying the Model (contd.)

Kmeans Clustering



Applying the Model (contd.)

Kmeans Clustering



Conclusion

- **Segmentation is needed to drive higher profitability through understanding customer needs and delivering on those.**

- **From the above analysis, we have majorly created 3customer segments. They are as follows-**

1. **Low Value Customers**

   'Cluster 0' customers can be called Low valued customers because they are less frequent, spends less money and they have purchased long time ago.

2. **Average value Customers**

   'Cluster 1' customers can be called has Average valued customers because they are some what more recent, frequent and spend some what more money compared to Low value customers. These customers could become high risk and we should aggressively market towards them with great deals so we don't lose them forever.

Conclusion (contd.)

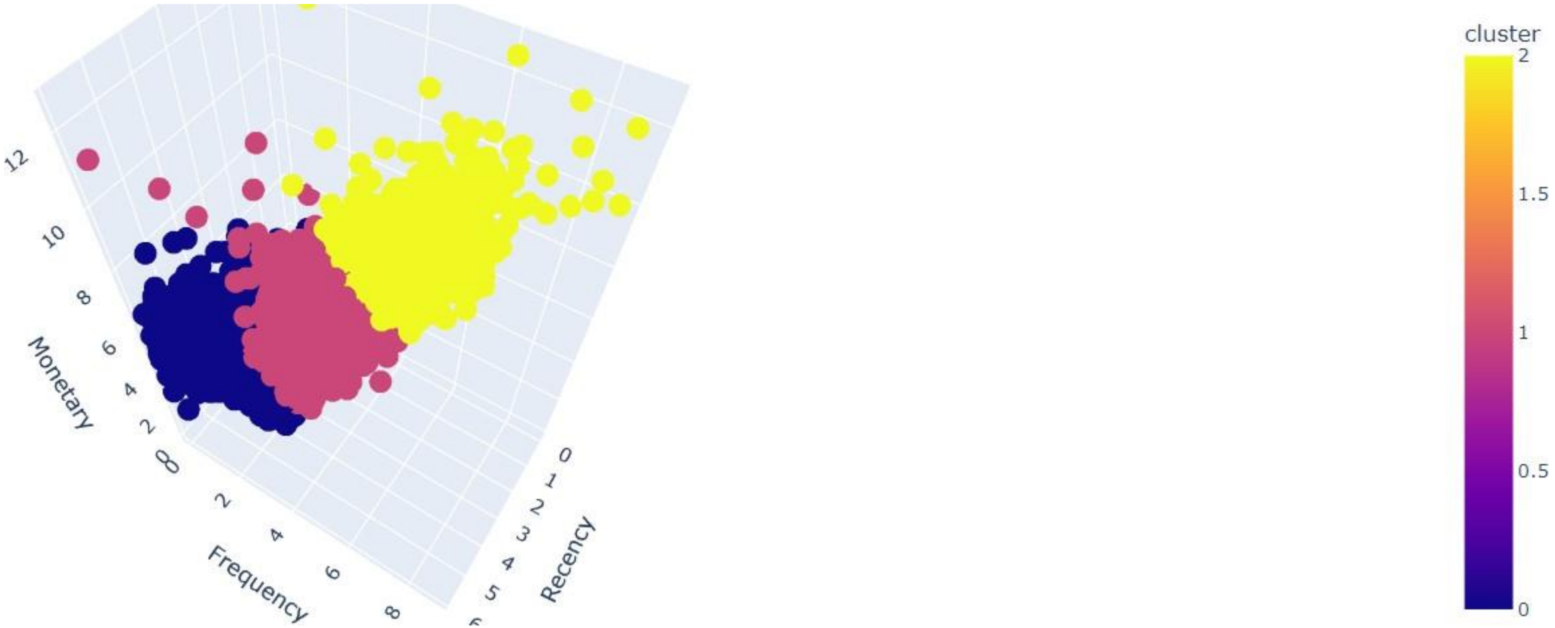
3. **<u>High value Customers</u>**

'Cluster 2' customers can be grouped has High valued customers because they are very recent (shopped recently), more frequent and spend more money aswell.

These are our best or potential customers we should not lose them.Communications with this group make them feel valued and appreciated. These customers likely generate a disproportionately high percentage of overall revenues and thus focusing on keeping them happy should be a top priority. Further analyzing their individual preferences and affinities we can provide additional opportunities for even more personalized messaging.

Conclusion (contd.)

3D Representation of customer segments



Thank You