Capstone Project-IV

Online Retail Customer Segmentation

By- Ajay Tiwari

CONTENTS

1. Introduction and Problem Statement
2. Data Description
3. Data Pipeline
4. Data Cleaning
5. Exploratory Data Analysis
6. Data Transformation
7. Modeling
8. Conclusion

Introduction and Problem Statement

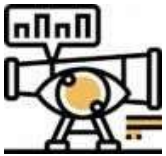
- Customer segmentation is one of the key aspects of business decision support system. In order to grow the business intelligently in competitive market, identification of potential customer should be done timely. All over the world business is growing in every field With the help of online platform and new technologies it strengthen its root more deeper, having access to wider market and large customers Online Retail business is also one of them, Customer segmentation refers to categorizing customers into different groups with similar characteristics. It can help to each customer group in a different way, in order to maximize their business and deliver true services who actually deserve.

- 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 DESCRIPTION

We have been provided with UK-based and registered online retail company which contains transaction between 01/12/2010 and 09/12/2011 with 541909 instances and 8 features.

1. **InvoiceNo:** A 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.
2. **StockCode**: It is a 5-digit number assigned to each distinct product.
3. **Description:** Name of each product.
4. **Quantity:** Number of a particular product sold on per transaction.
5. **InvoiceDate:** It holds the information of Date and time when transaction was generated.
6. **UnitPrice:** Price per unit of a particular product sold
7. **CustomerID:** A 5-digit number uniquely assigned to each customer.
8. **Country:** The name of the country where customer resides



DATA PIPELINE

1. **Exploratory Data Analysis (EDA):** In this part we have done some EDA on the features to see the trend.

2. **Data Processing**: In this part we went through each attributes and encoded the categorical features.

3. **Data Transformation:** In this the Recency, Frequency and Monetary (RFM) of the given data set is created.

4. **Modeling**: Finally in this part, created various clustering models. The optimal clusters obtained from each models were analyzed.

DATA CLEANING

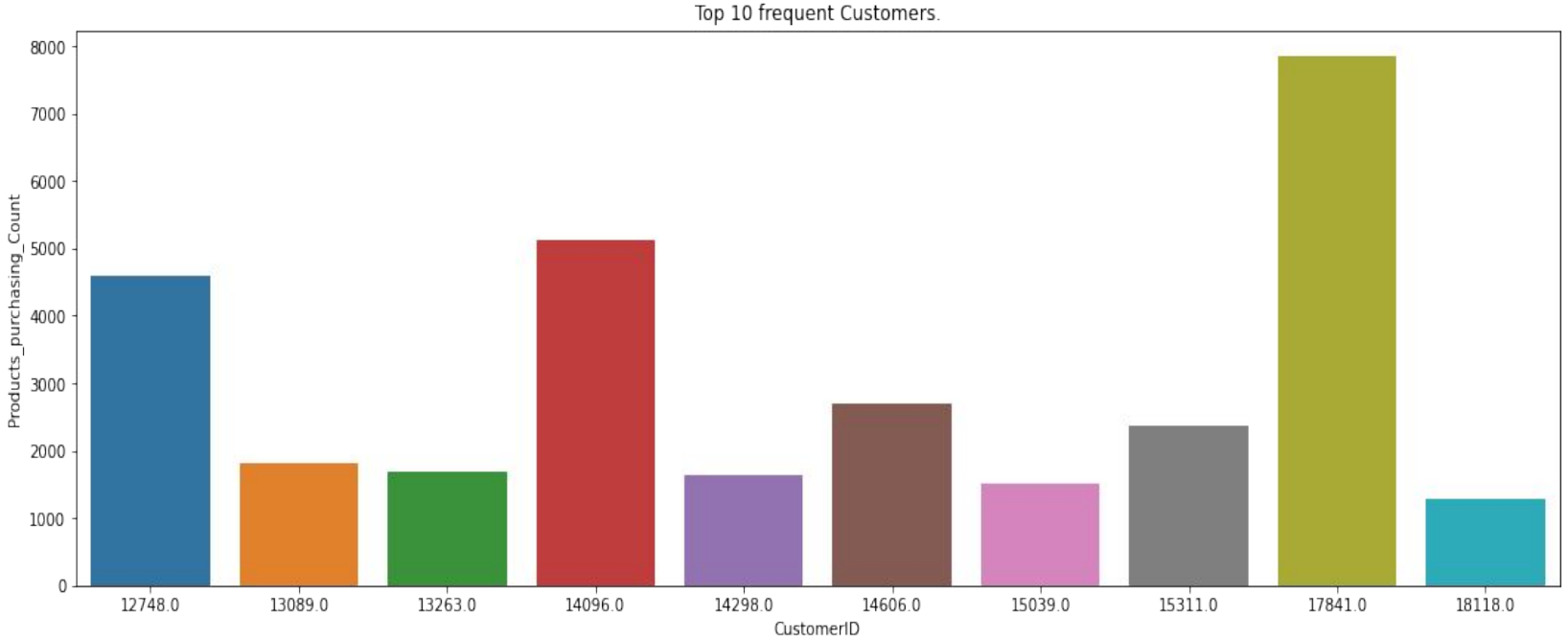
- The given dataset has 541909 rows and 8 features(columns).
- The dataset possess high number of null values in Description and CustomerID columns:

Variable	# of nulls	% of nulls
InvoiceNo	0	0
StockCode	0	0
Description	1454	0.27
Quantity	0	0
InvoiceDate	0	0
UnitPrice	0	0
CustomerID	135080	24.93
StockCode	0	0

Dropped all the null values because each customer IDs are uniquely assigned to a customer. If it is missing from the dataset, then we can impute it with other values, but it does not make any sense and if we do then we end up with biased results.

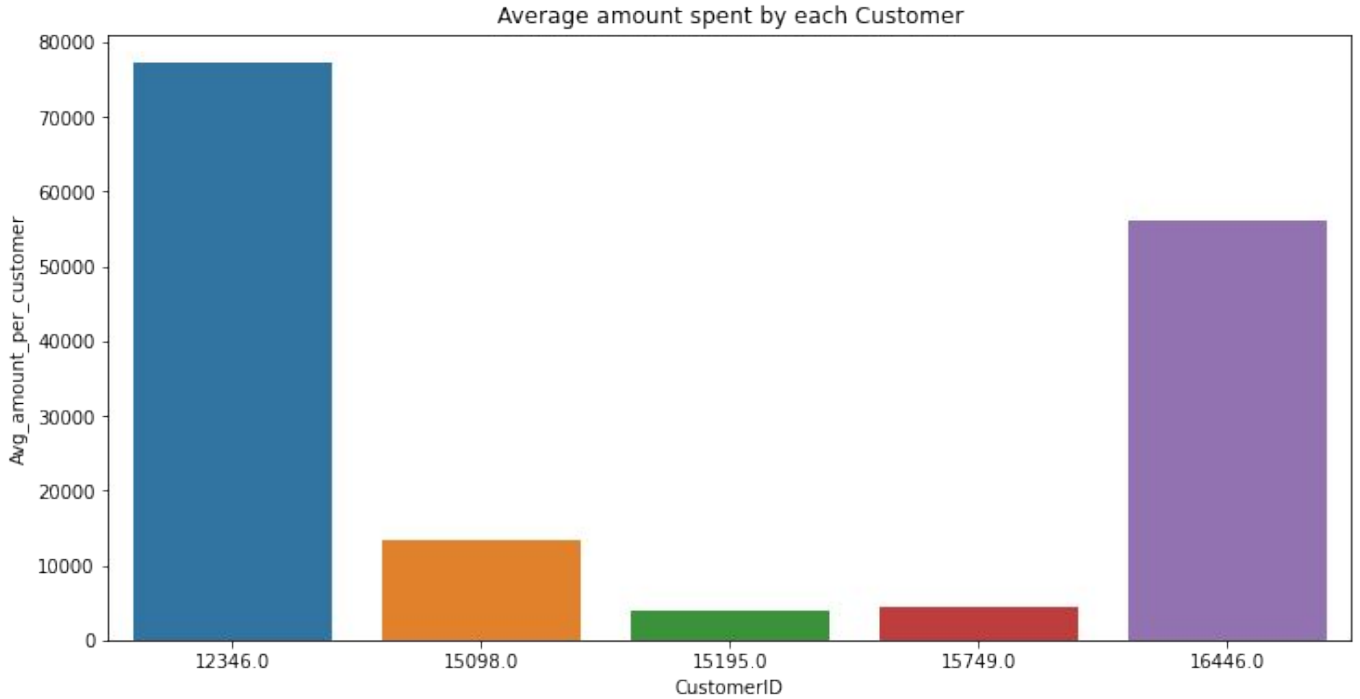
EDA (Visualizations)

TOP 10 FREQUENT CUSTOMERS BASED ON PURCHESING COUNT



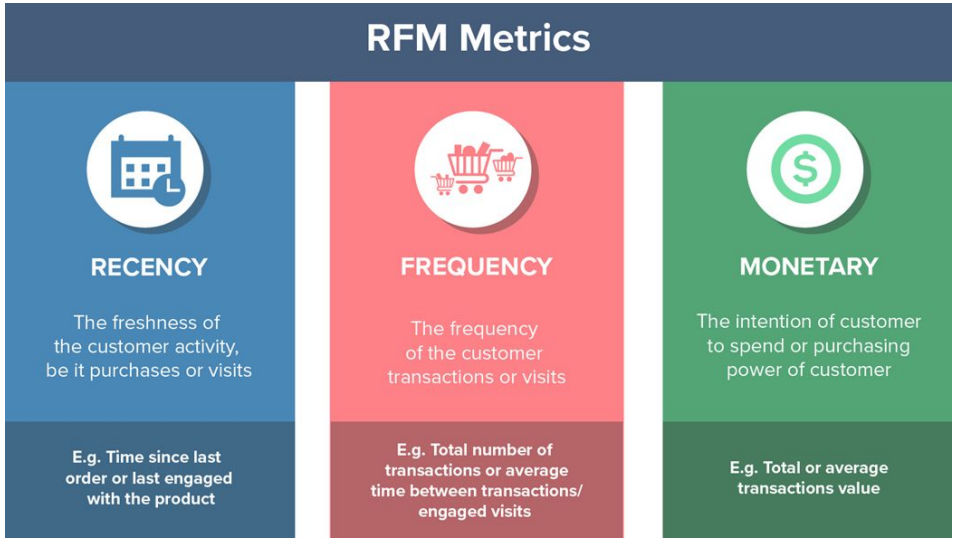
EDA (Visualizations)

Average amount spent by each customer



DATA TRANSFORMATION

In this **Recency, Frequency and Monetary (RFM) analysis** about the data is done. **Recency** signifies the days since order, **Frequency** signifies the number of times the customer is been billed and **Monetary** signifies the sales each customer has provided.

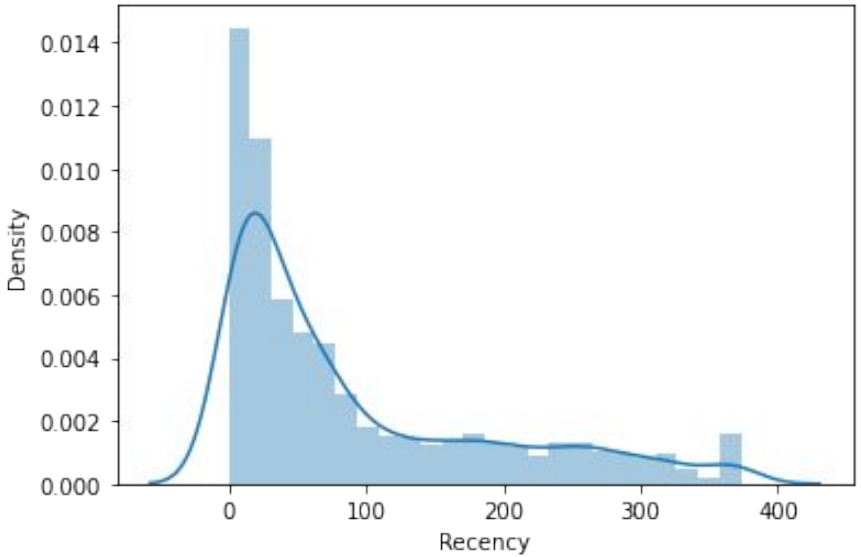


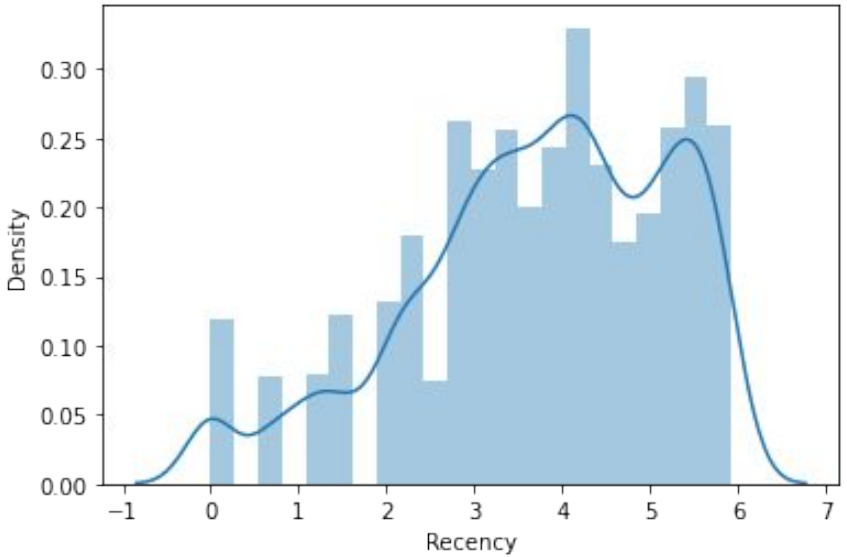
RFM OF THE GIVEN DATASET

	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12747.0	2	103	4196.01
2	12748.0	0	4596	33719.73
3	12749.0	3	199	4090.88
4	12820.0	3	59	942.34
5	12821.0	214	6	92.72
6	12822.0	70	46	948.88
7	12823.0	74	5	1759.50
8	12824.0	59	25	397.12
9	12826.0	2	91	1474.72

RECENCY

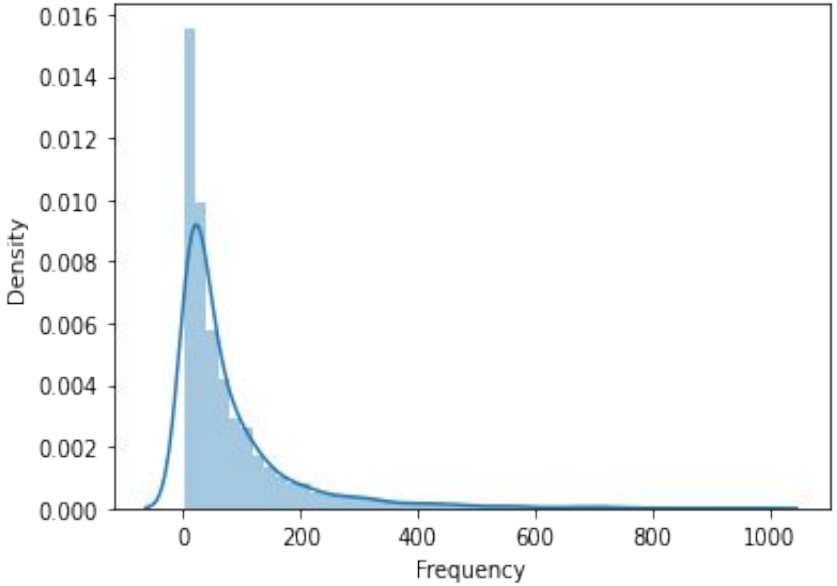
Distribution before and after normal distribution with log transformation

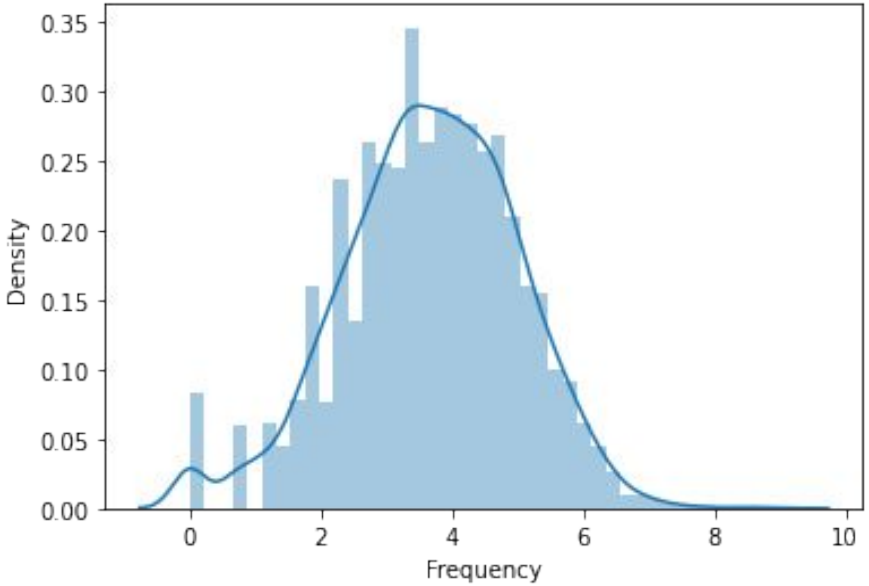




FREQUENCY

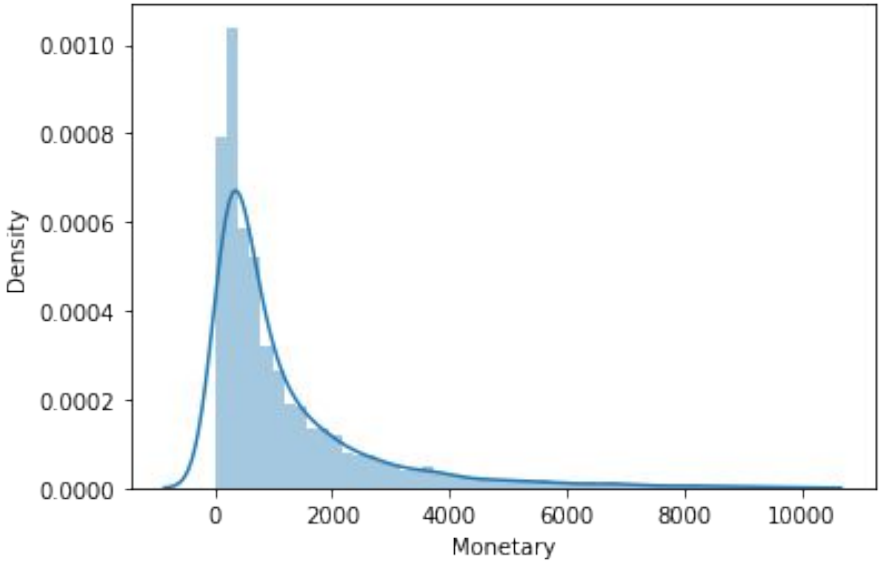
Distribution before and after normal distribution with log transformation

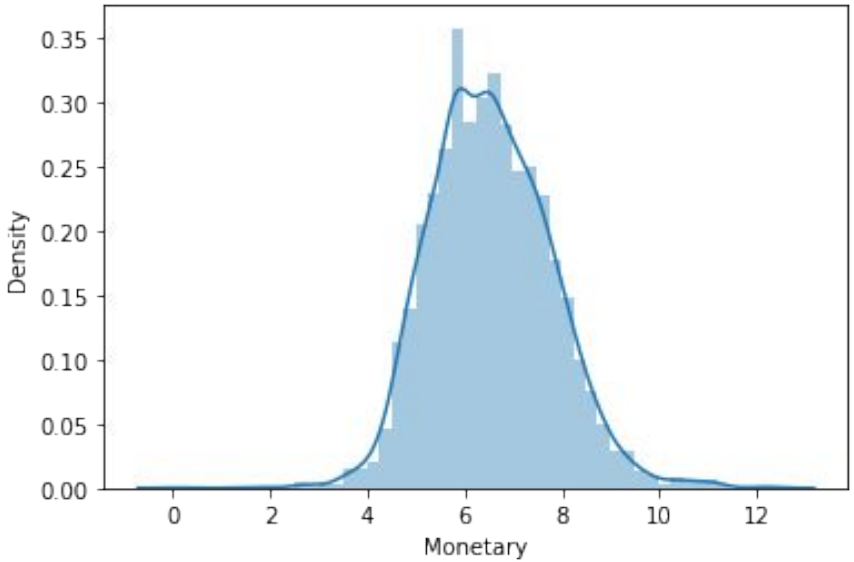




MONETARY

Distribution before and after normal distribution with log transformation





Loyalty Level to each customer

AI

	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	12346.0	325	1	77183.60	4	4	1	441	9	Silver
1	12747.0	2	103	4196.01	1	1	1	111	3	Platinum
2	12748.0	0	4596	33719.73	1	1	1	111	3	Platinum
3	12749.0	3	199	4090.88	1	1	1	111	3	Platinum
4	12820.0	3	59	942.34	1	2	2	122	5	Platinum

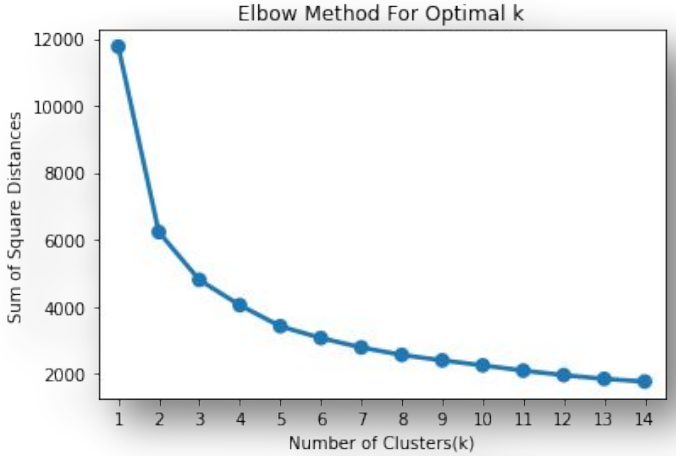
Modelling

Model selection

The model in use are:

- K-MEANS CLUSTERING

Data Modelling

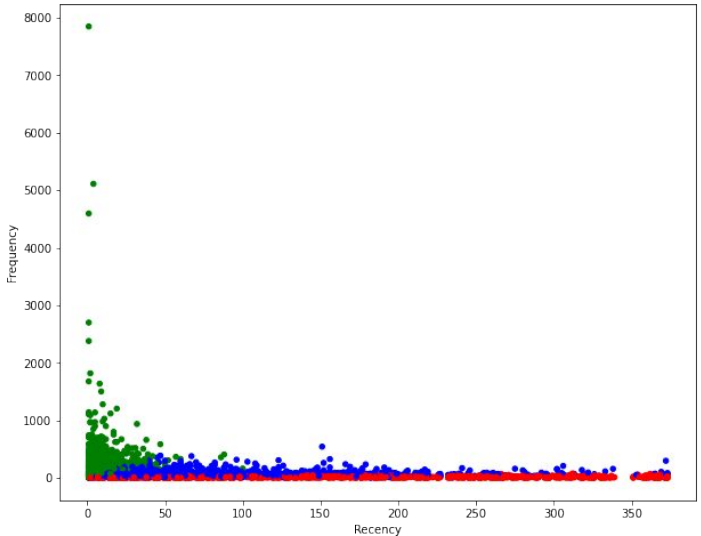



```
For n_clusters=2, the silhouette score is 0.39562494982602087
For n_clusters=3, the silhouette score is 0.30546474589014033
For n_clusters=4, the silhouette score is 0.2980445141762671
For n_clusters=5, the silhouette score is 0.2785167472954507
For n_clusters=6, the silhouette score is 0.27726461404219555
For n_clusters=7, the silhouette score is 0.26066118271648536
For n_clusters=8, the silhouette score is 0.2583990050322177
```

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	Silver	2
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	1
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	1
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	1
12820.0	3	59	942.34	1	2	2	122	5	Platinum	1

Data Modelling

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster	Color
12346.0	325	1	77183.60	4	4	1	441	9	Silver	2	blue
12747.0	2	103	4196.01	1	1	1	111	3	Platinum	1	green
12748.0	1	4596	33719.73	1	1	1	111	3	Platinum	1	green
12749.0	3	199	4090.88	1	1	1	111	3	Platinum	1	green
12820.0	3	59	942.34	1	2	2	122	5	Platinum	1	green



Challenges

- Identify the highly imbalance large dataset and manage it carefully.
- Ensure model is treating all groups fairly.
- Lot of NaN values.
- Before deploy model set the proper number of clusters.
- After deploying understand model behavior on real data.
- Tough to make prediction analysis to end users.

Conclusion

- Throughout the analysis we went through various steps to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.

- Next we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented K-Means clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 3.

- We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster.
- We saw higher values of frequency, monetary and low values of recency is deciding one class and low values of frequency, monetary and high values of recency is deciding other class.

References

I. Stack overflow
II. GeeksforGeeks
III. Almabetter
IV. Analytics Vidhya
V. Towards data science
VI. github

THANK YOU