CUSTOMER SEGMENTATION

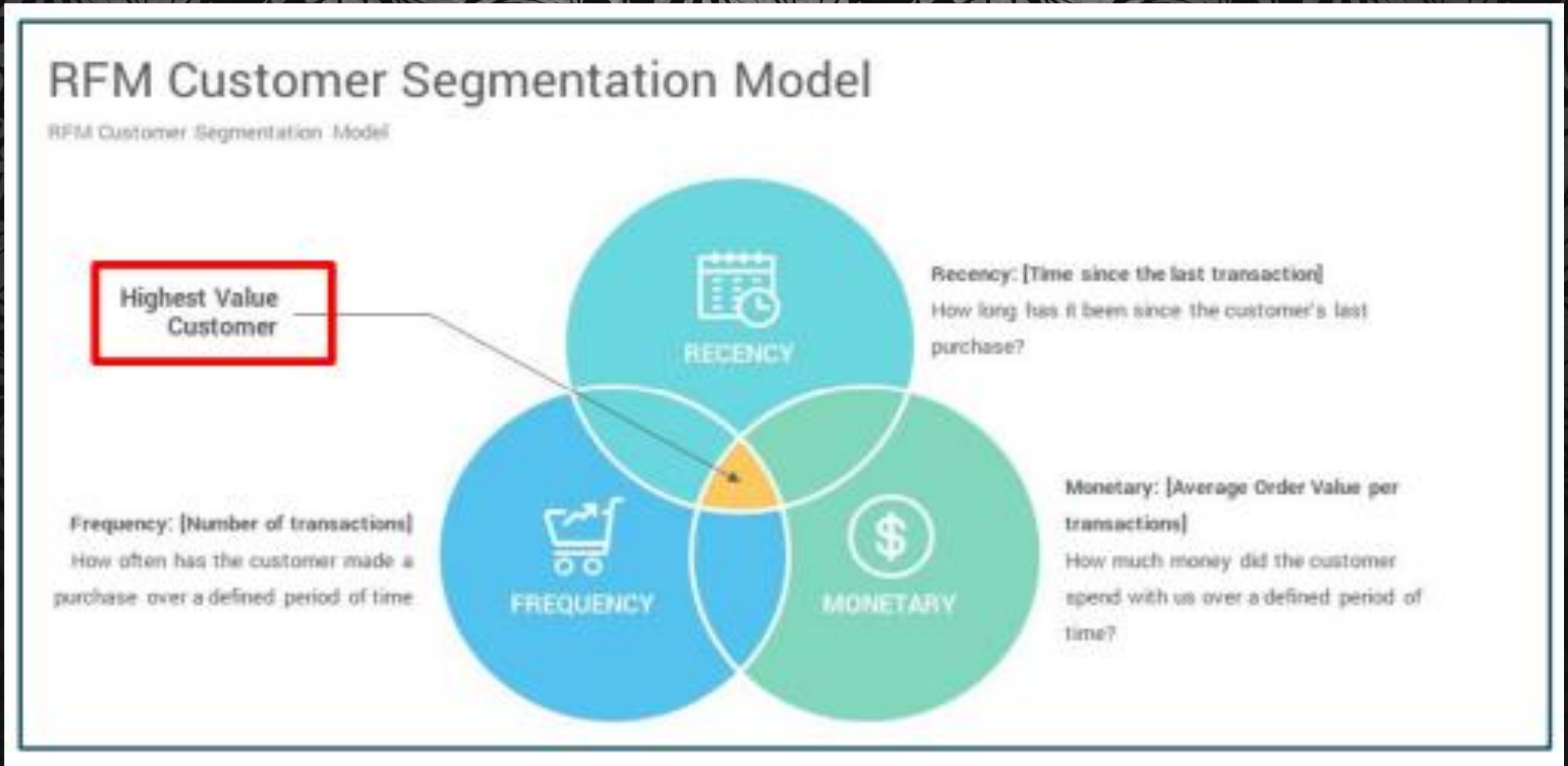
BY-CHARAN

POINTS TO DISCUSS

- problem statement
- data summary
- EDA / feature analysis
- data preparation
- creating RFM model
- implementing various clustering models
- challenges
- conclusion

PROBLEM STATEMENT

- This project aims to identify major customer segments on a transnational data set for a UK-based online retail.
- Create RFM table .
- analyze and identify major customer segmentation using k means algorithm and also different algorithms to confirm our result.



RFM Customer Segmentation Model

RFM Customer Segmentation Model

Highest Value Customer

Recency: [Time since the last transaction]

How long has it been since the customer's last purchase?

Frequency: [Number of transactions]

How often has the customer made a purchase over a defined period of time

Monetary: [Average Order Value per transactions]

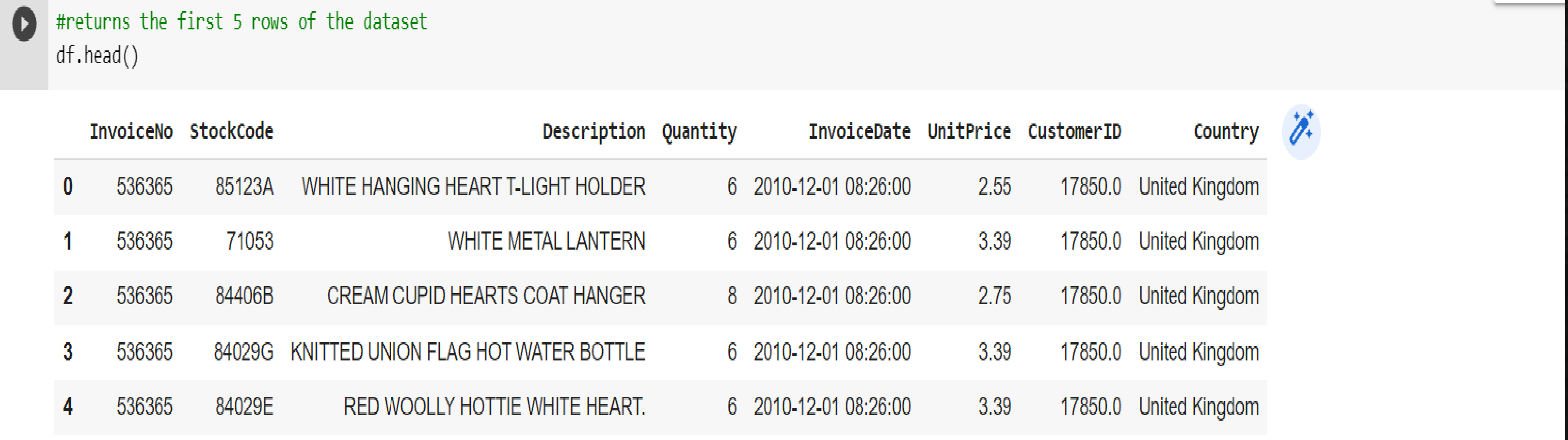
How much money did the customer spend with us over a defined period of time?

DATA SUMMARY

- Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction..
- Stock code: Product (item) code. 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name.
- Quantity: The quantities of each product (item) per transaction.
- Invoice Date: Invoice Date and time. The day and time when each transaction was generated.
- Unit price: Unit price. Product price per unit in sterling.
- Customer ID: Customer number.
- Country: Country name. Nominal, the name of the country where each customer resides.

BASIC DATA EXPLORATION

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



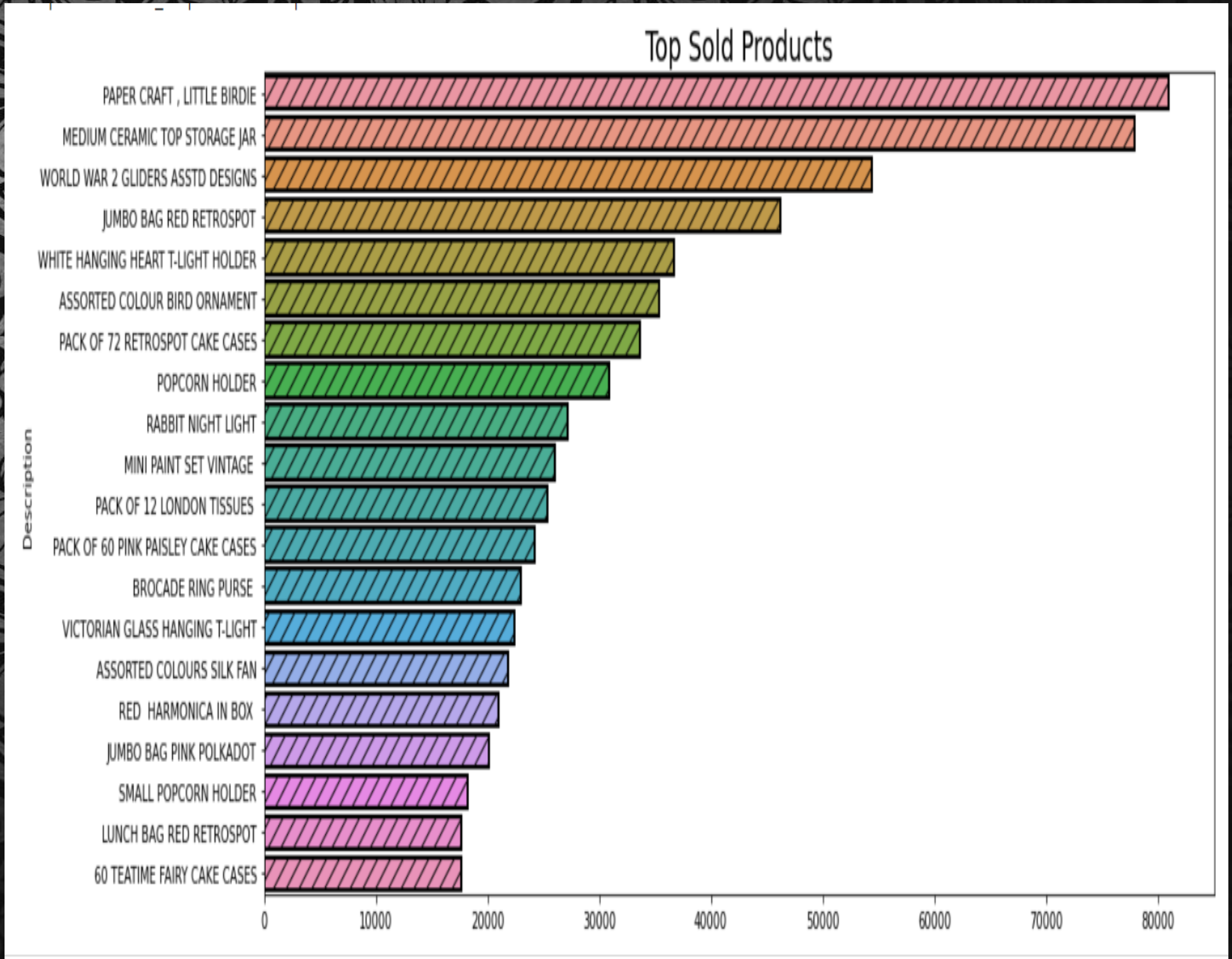
```
#returns the first 5 rows of the dataset
df.head()
```

	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

EDA(EXPLORATORY DATA ANALYSIS

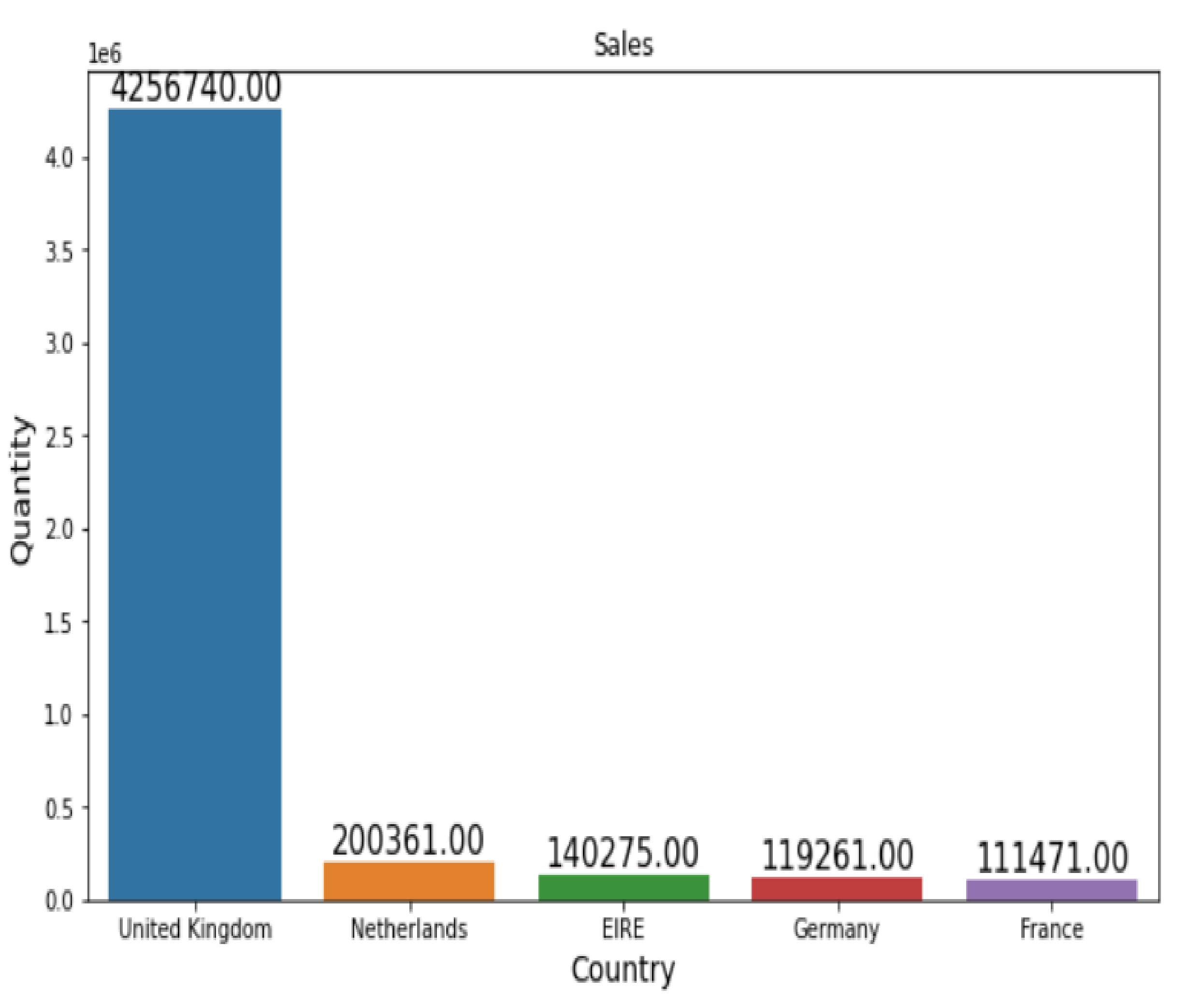
TOP SOLD PRODUCTS

There are so many product in the shop but here we have taken the top sold 20 Products.
From this bar chart we can see that paper craft and little birdie is the top sold product in the shop .

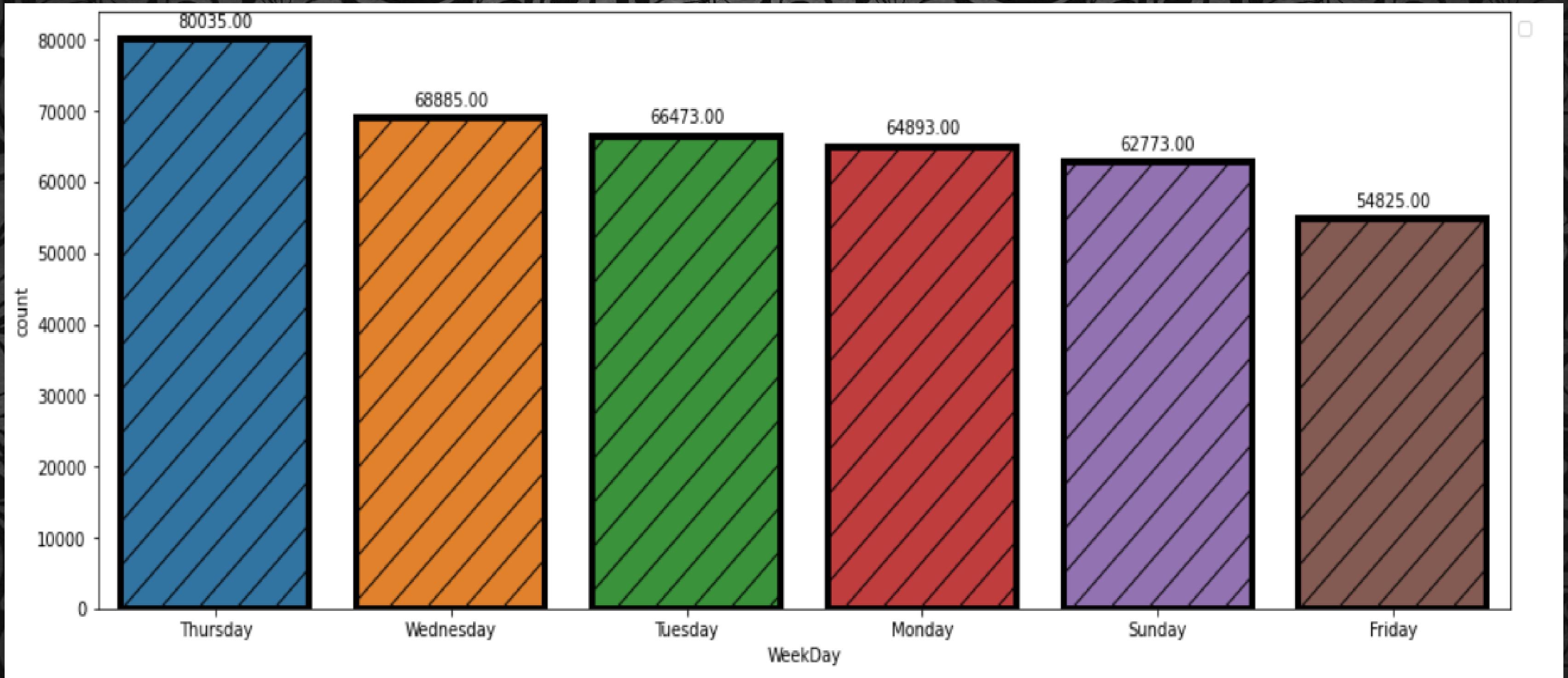


TOP 5 PURCHASED COUNTRY

In This graph, we can observe that most purchases are from the United Kingdom.
So we can tell that this is UK based company

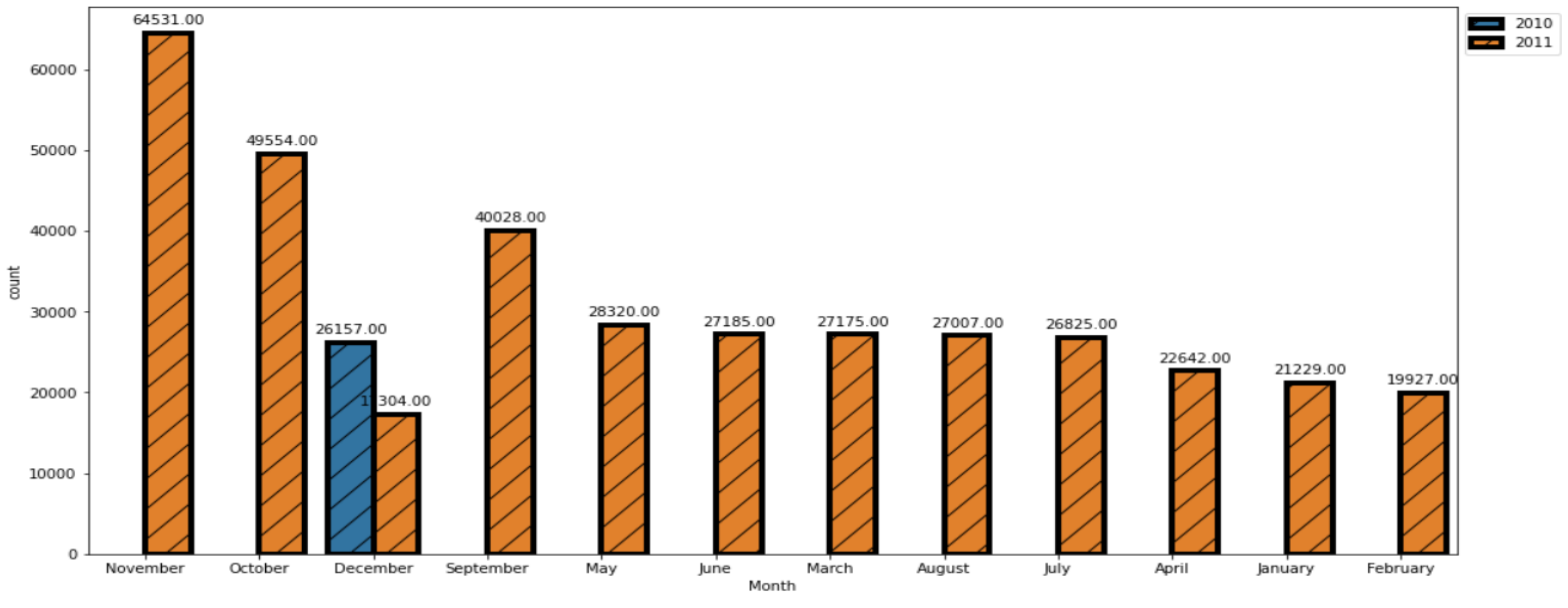


DAY WHICH HAS HIGHEST PURCHASE RATE



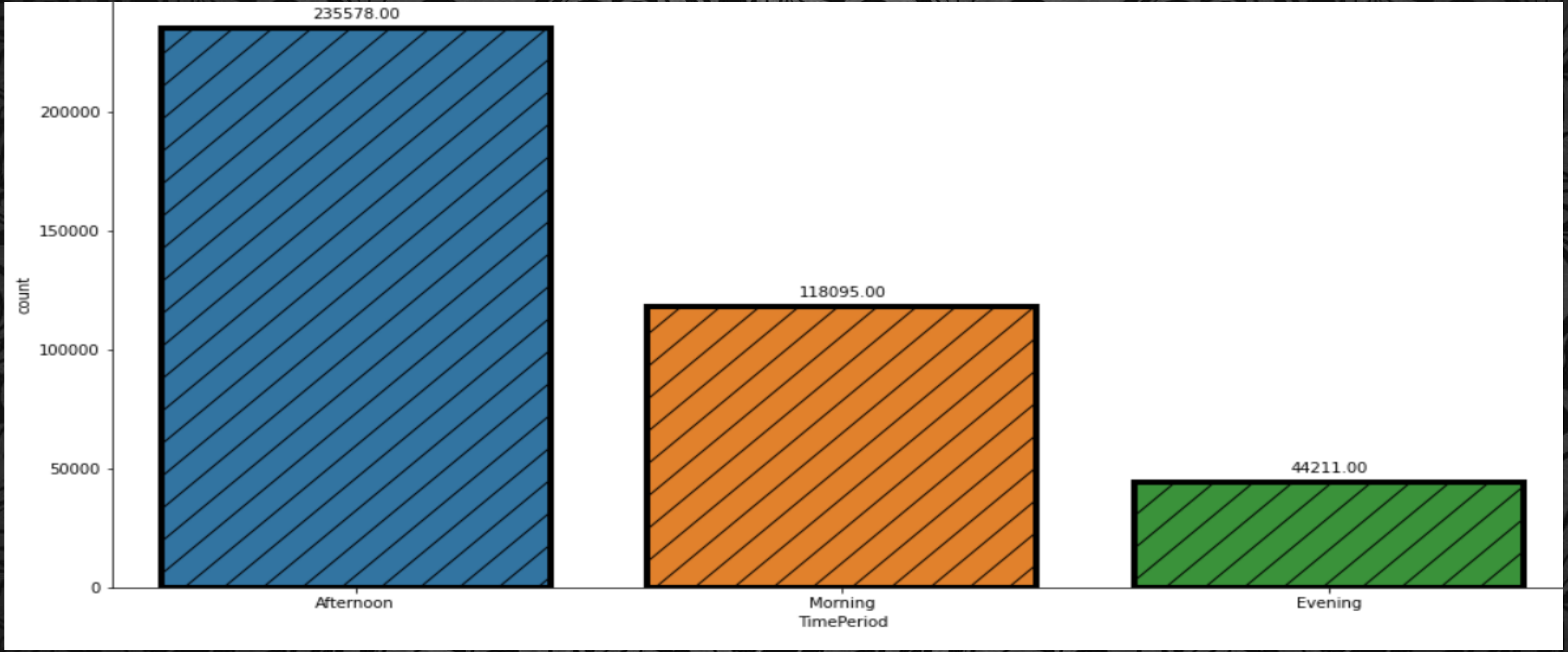
From the above bar graph we can conclude that Most of the customers have purchased items on Thursday, Wednesday, Tuesday

MONTH WHICH HAS HIGHEST PURCHASE RATE



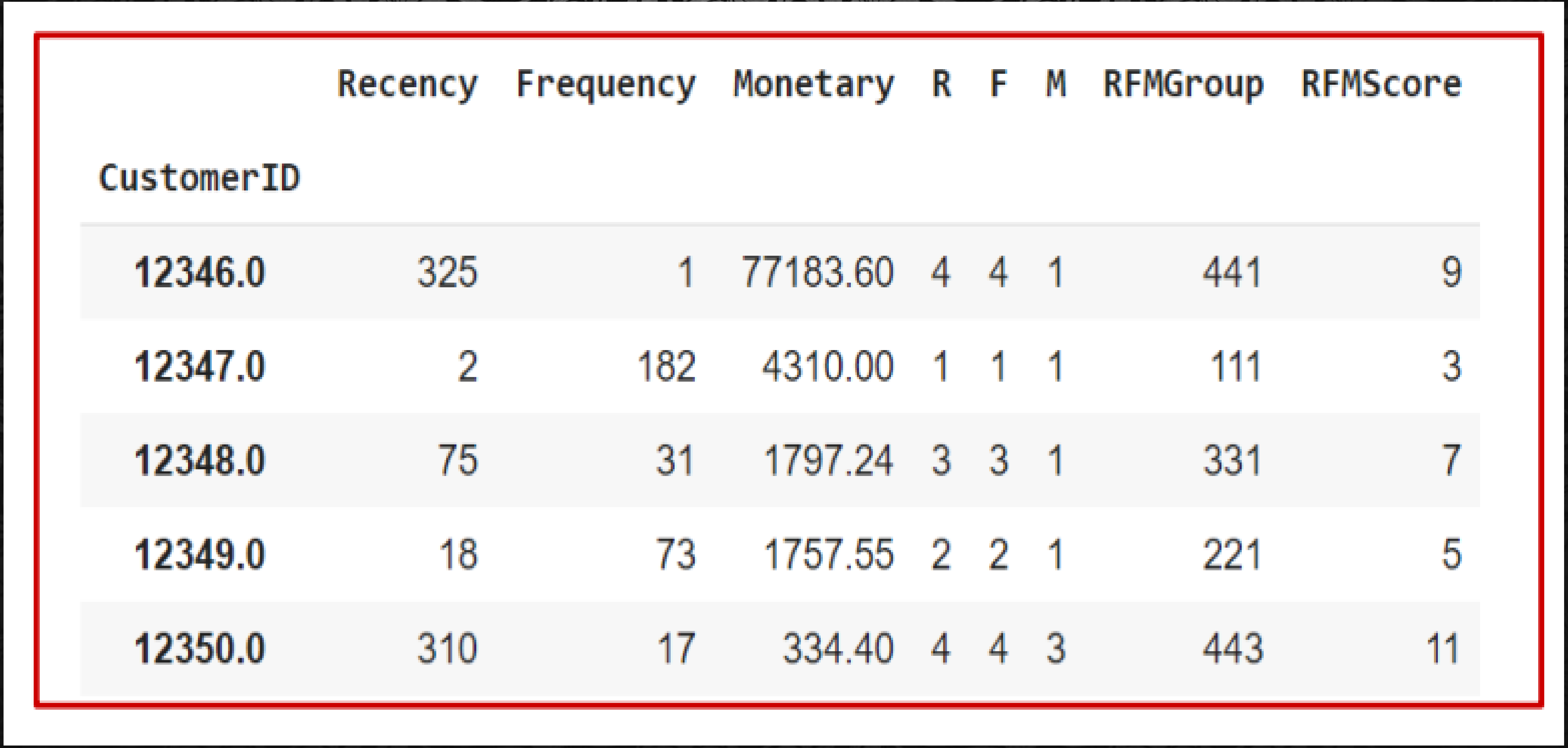
By looking at this bar chart we can conclude that
Most of the customers have purchased items in November, October, December, and the least number of purchases in April, January, February

TIME PERIOD WISE REVIEW



Most of the customers purchase in the afternoon time. The 12th hour of the day is a peak for purchasing items. Moderate numbers of customers have purchased the items in the Morning and the least numbers of customers have purchased the items in the Evening.

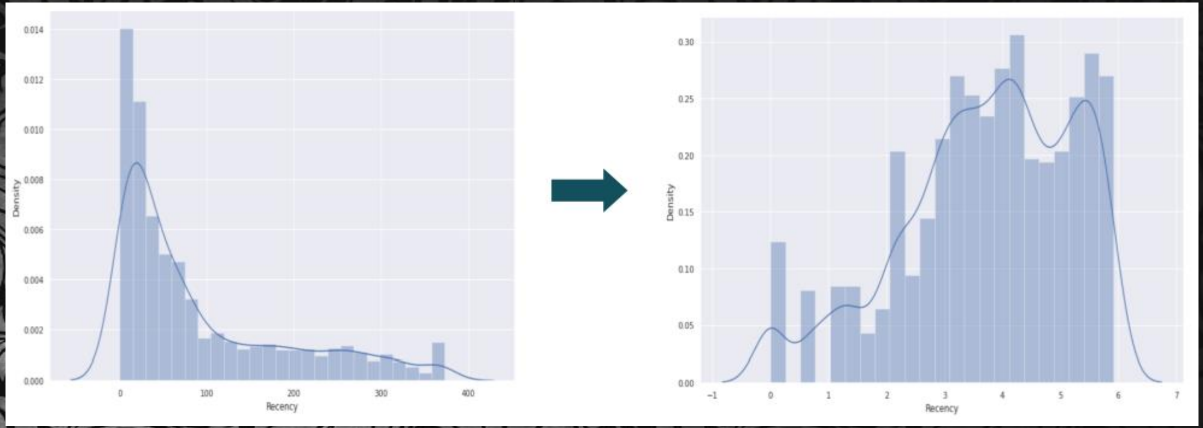
CREATE THE RFM MODEL (RECENCY, FREQUENCY, MONETARY VALUE)



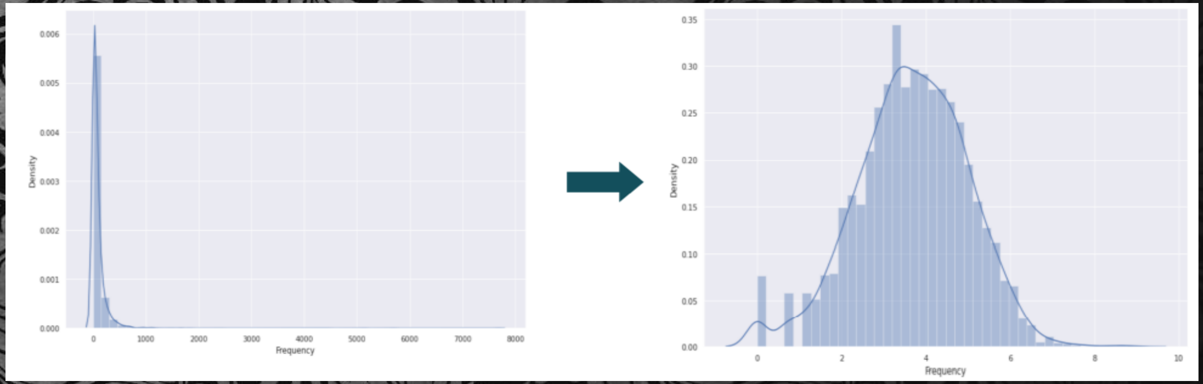
	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
CustomerID								
12346.0	325	1	77183.60	4	4	1	441	9
12347.0	2	182	4310.00	1	1	1	111	3
12348.0	75	31	1797.24	3	3	1	331	7
12349.0	18	73	1757.55	2	2	1	221	5
12350.0	310	17	334.40	4	4	3	443	11

LOG TRANSFORMATION

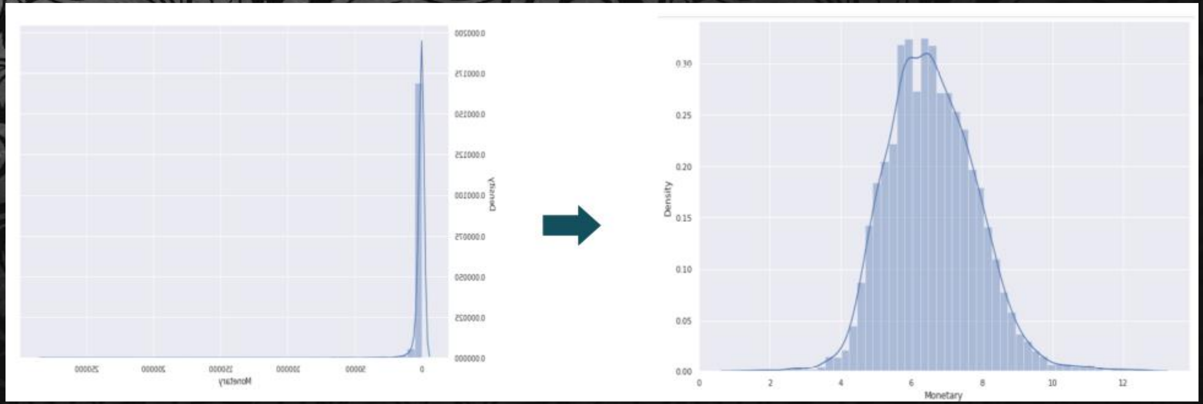
LOG TRANSFORMATION ON RECENCY



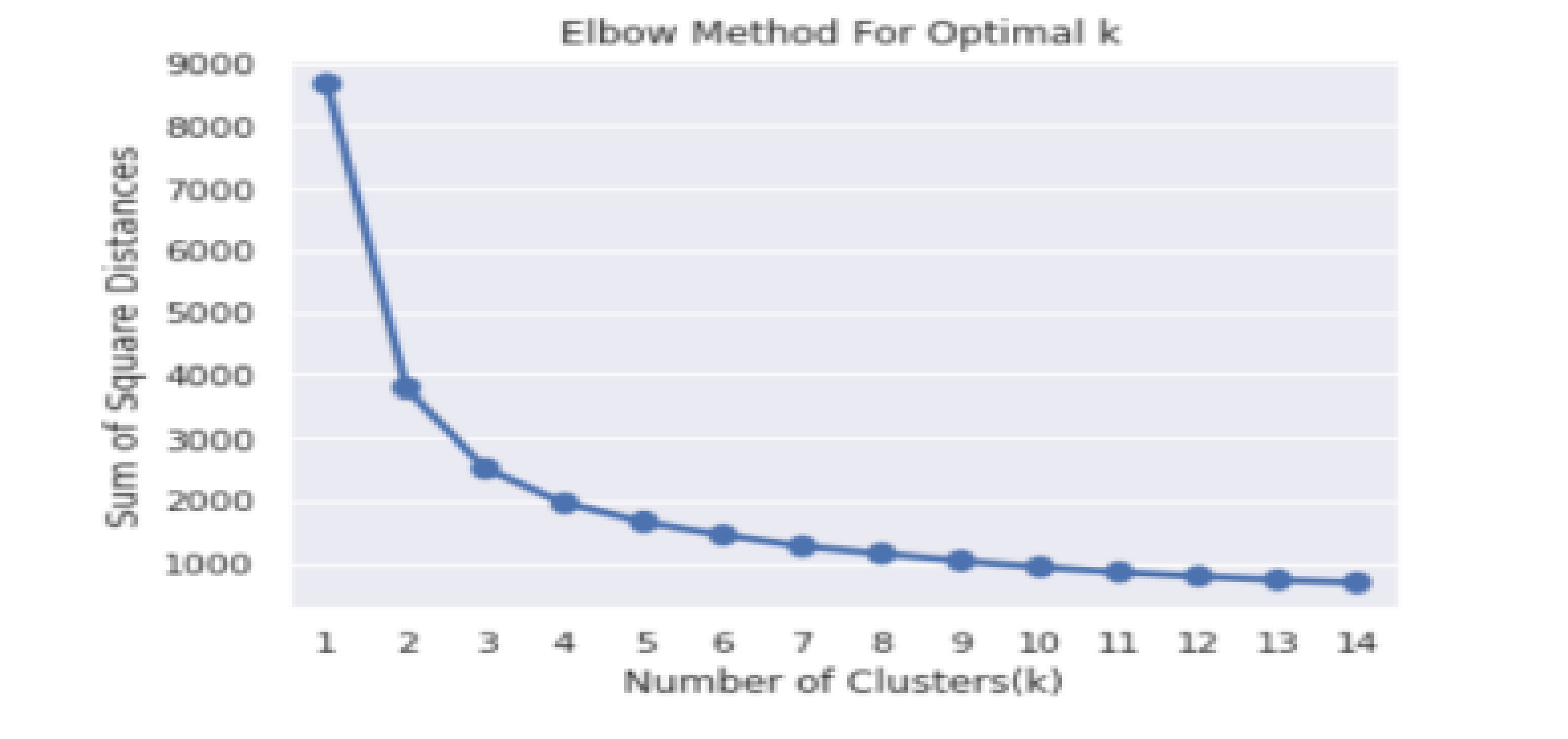
LOG TRANSFORMATION ON FREQUENCY



LOG TRANSFORMATION ON MONETARYVALUE

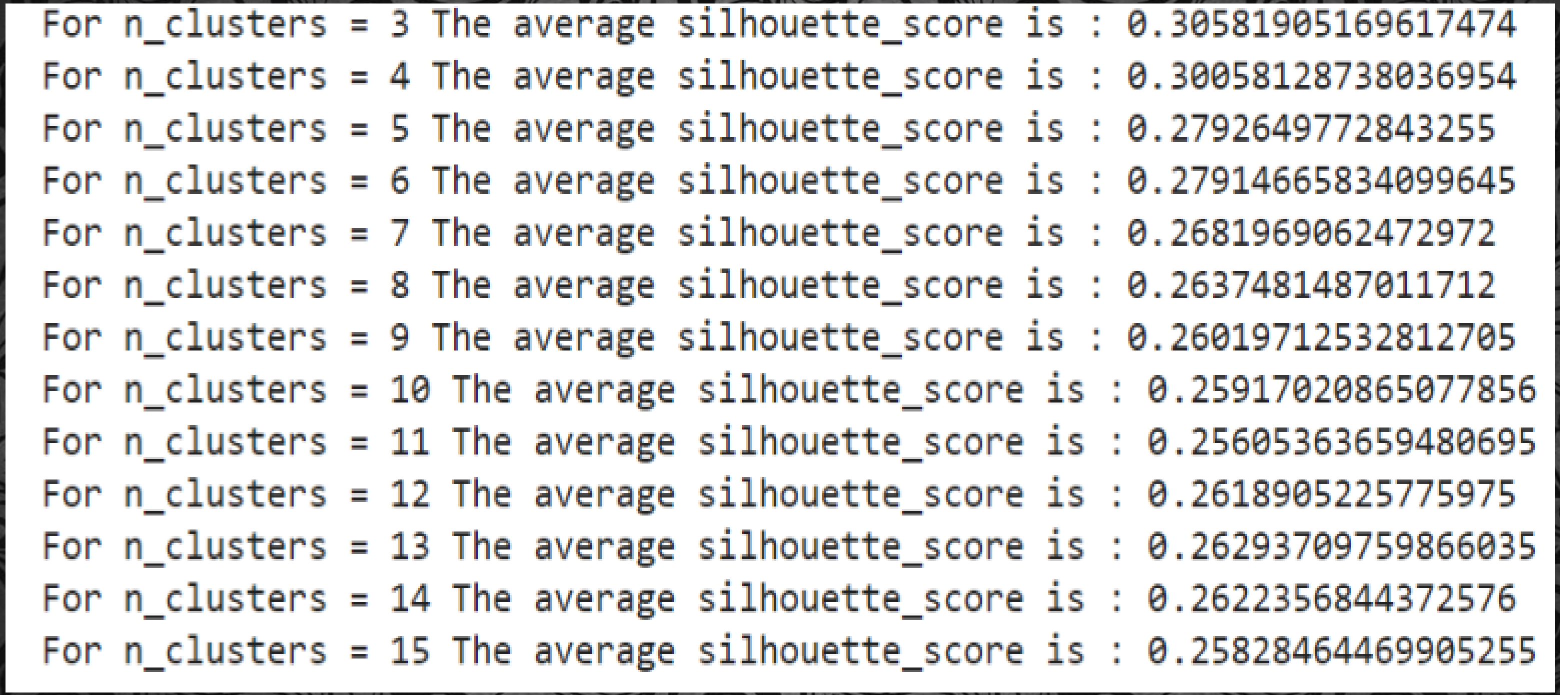


ELBOW METHOD



Elbow method : a point from where the value of clusters starts decreasing suddenly, indicates the optimal number of clusters.

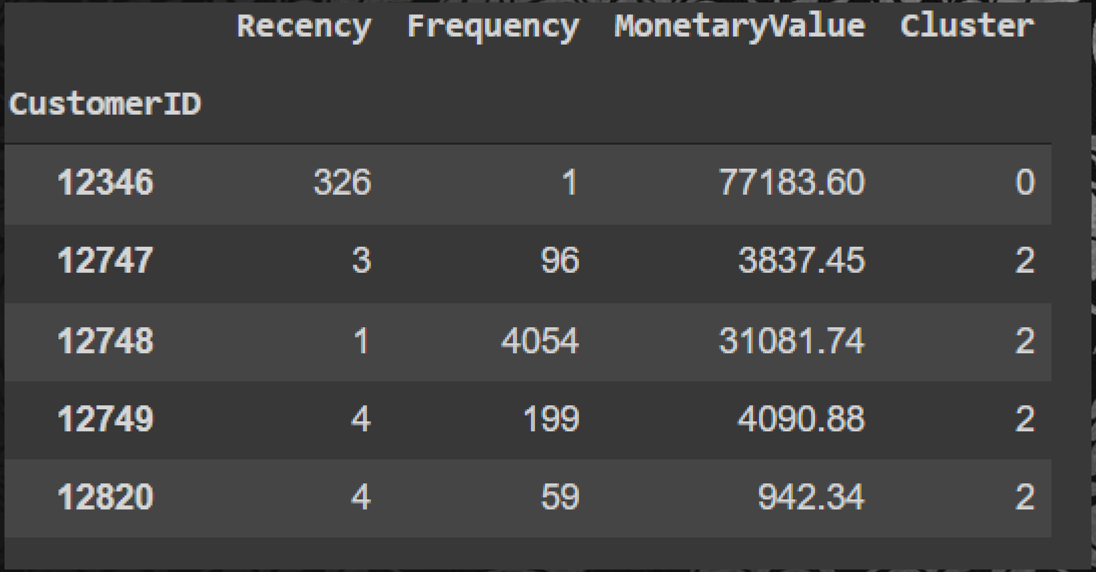
APPLYING SILHOUETTE METHOD ON RECENCY, FREQUENCY AND MONETARY



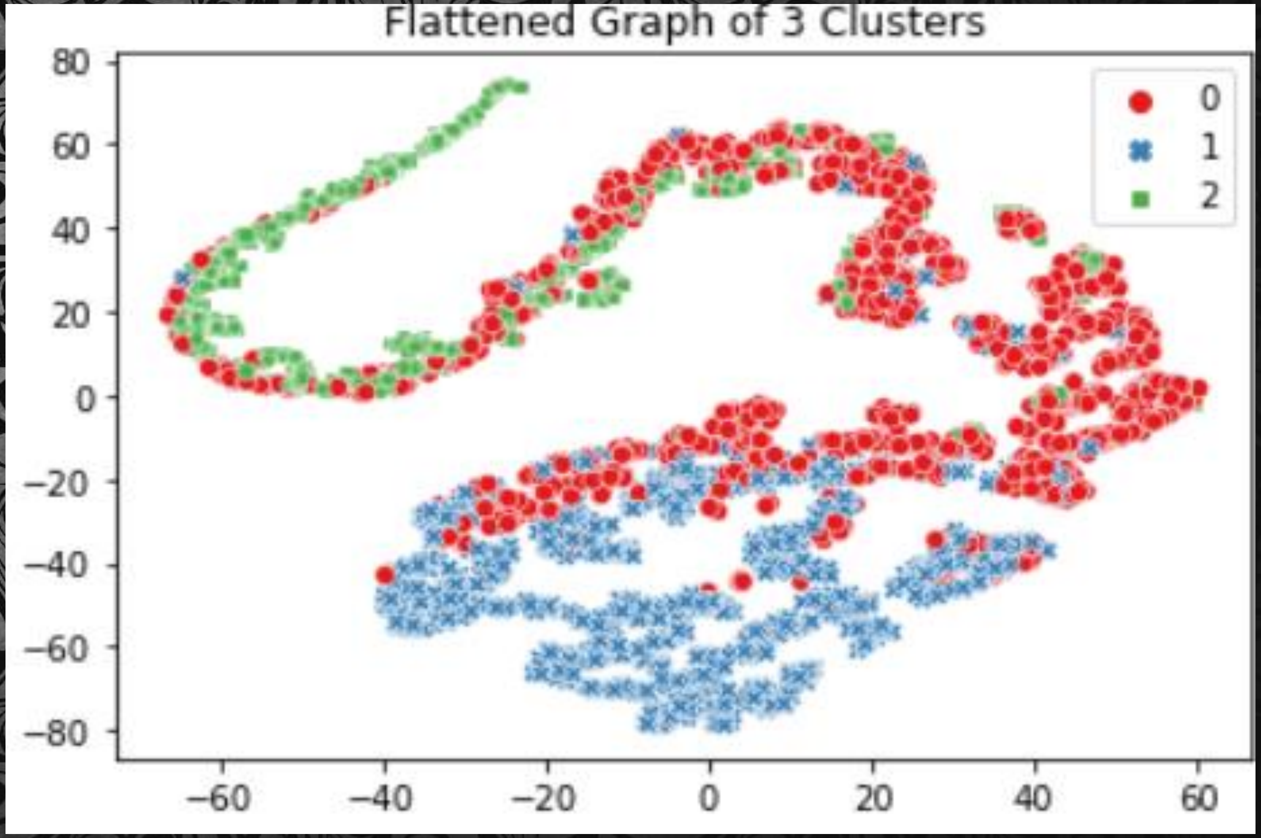
```
For n_clusters = 3 The average silhouette_score is : 0.30581905169617474
For n_clusters = 4 The average silhouette_score is : 0.30058128738036954
For n_clusters = 5 The average silhouette_score is : 0.279264977284325
For n_clusters = 6 The average silhouette_score is : 0.27914665834099645
For n_clusters = 7 The average silhouette_score is : 0.268196906247297
For n_clusters = 8 The average silhouette_score is : 0.263748148701171
For n_clusters = 9 The average silhouette_score is : 0.26019712532812705
For n_clusters = 10 The average silhouette_score is : 0.25917020865077856
For n_clusters = 11 The average silhouette_score is : 0.25605363659480695
For n_clusters = 12 The average silhouette_score is : 0.261890522577597
For n_clusters = 13 The average silhouette_score is : 0.26293709759866035
For n_clusters = 14 The average silhouette_score is : 0.262235684437257
For n_clusters = 15 The average silhouette_score is : 0.25828464469905255
```

Silhouette Coefficient or silhouette score is **a metric used to calculate the goodness of a clustering technique**. Its value ranges from -1 to 1. 1:

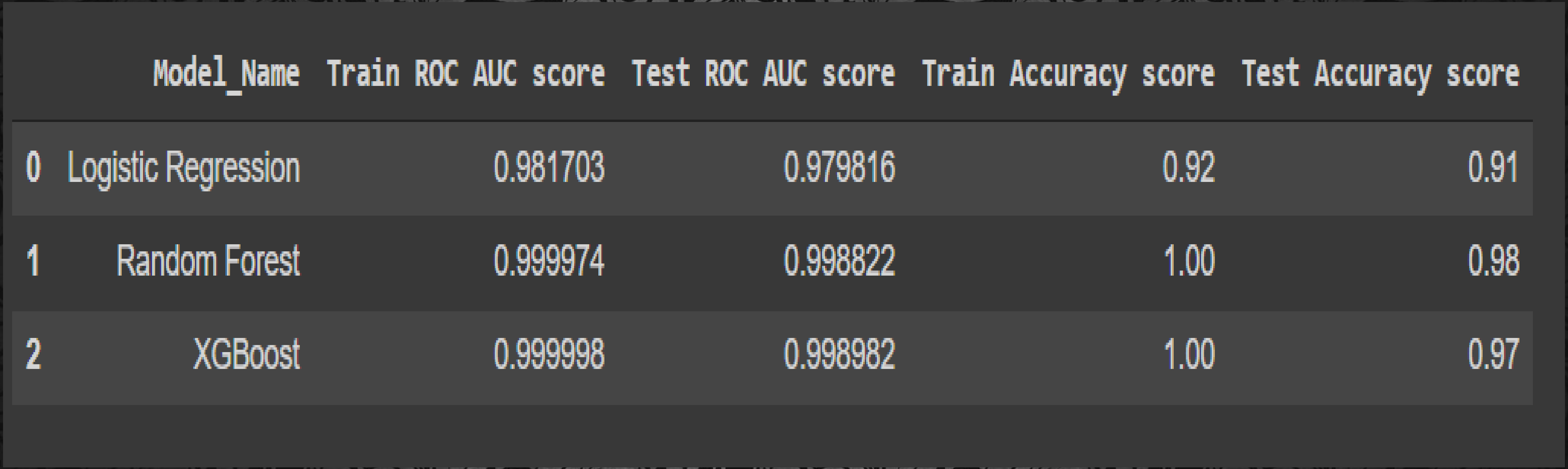
CLUSTER GRAPH



	Recency	Frequency	MonetaryValue	Cluster
CustomerID				
12346	326	1	77183.60	0
12747	3	96	3837.45	2
12748	1	4054	31081.74	2
12749	4	199	4090.88	2
12820	4	59	942.34	2



TRAIN TEST EVALUATION



	Model_Name	Train ROC AUC score	Test ROC AUC score	Train Accuracy score	Test Accuracy score
0	Logistic Regression	0.981703	0.979816	0.92	0.91
1	Random Forest	0.999974	0.998822	1.00	0.98
2	XGBoost	0.999998	0.998982	1.00	0.97

Model evaluation is **the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses**

CHALLENGES

- Understanding the problem statement.
- Figuring Out right Approach
- Dealing with Null And duplicate values
- Extracting Datetime Column Properly and creating RFM variables.
- Designing multiple visualizations to summarize the Data points
- Finding optimal number of clusters

CONCLUSION

THE DATA EXPLORATION OF ONLINE CUSTOMER SEGMENTATION DATASET SHOWS :

● MISSING AND DUPLICATE VALUES WERE FOUND.

● MOST OF THE PURCHASES ARE FROM THE UNITED KINGDOM.

● MOST OF THE CUSTOMERS HAVE PURCHASED ITEMS ON THURSDAY, WEDNESDAY, TUESDAY.

● MOST OF THE CUSTOMERS HAVE PURCHASED ITEMS IN NOVEMBER, OCTOBER, DECEMBER, AND THE LEAST NUMBER OF PURCHASES IN APRIL, JANUARY, FEBRUARY.

● MOST OF THE CUSTOMERS PURCHASE IN THE AFTERNOON TIME. THE 12TH HOUR OF THE DAY IS A PEAK FOR PURCHASING ITEMS.

CONCLUSION

CUSTOMER SEGMENTATION IS AN IMPORTANT MARKETING APPROACH THAT BUSINESSES SHOULD EMPLOY IN ORDER TO GAIN A BETTER UNDERSTANDING OF THE MARKET AND MAKE MORE INFORMED DECISIONS IN ORDER TO INCREASE SALES. K-MEANS CLUSTERING IS A BASIC BUT EFFECTIVE MACHINE LEARNING ALGORITHM THAT BUSINESSES CAN USE. FINALLY, IN ORDER TO OPTIMISE OUR MARKETING SUCCESS, WE MUST KEEP THE RFM CLIENT SEGMENTATION UP TO DATE.

THANK
YOU