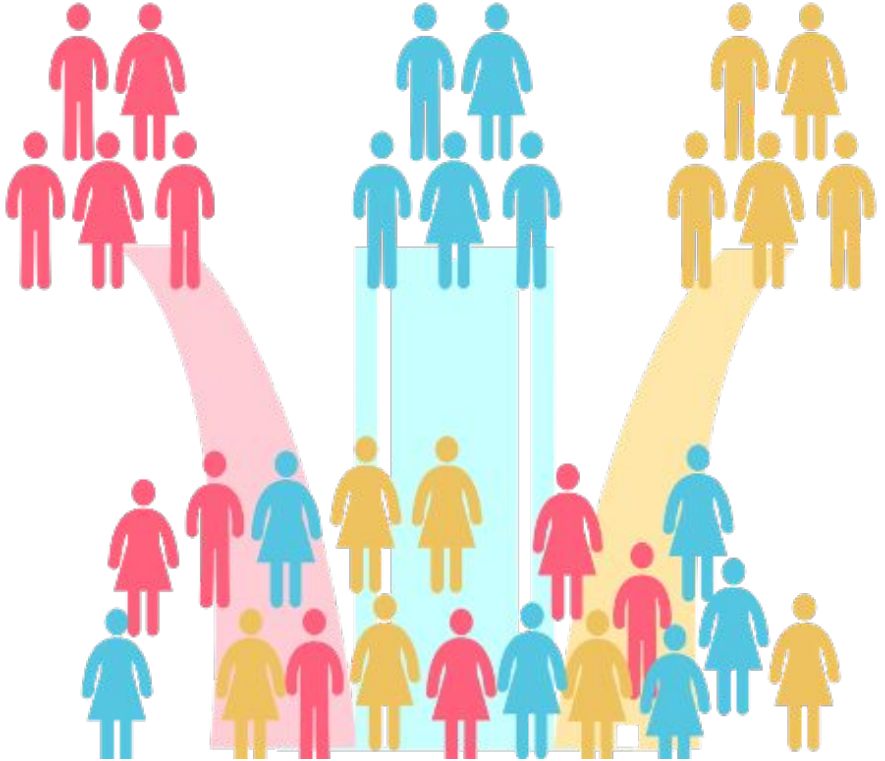
UNSUPERVISED LEARNING CAPSTONE PROJECT

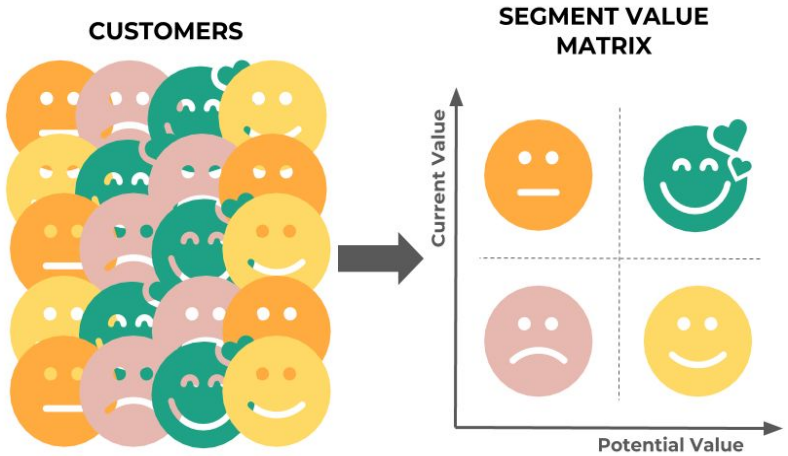
Customer Segmentation Analysis



Contents:

Introduction

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable.



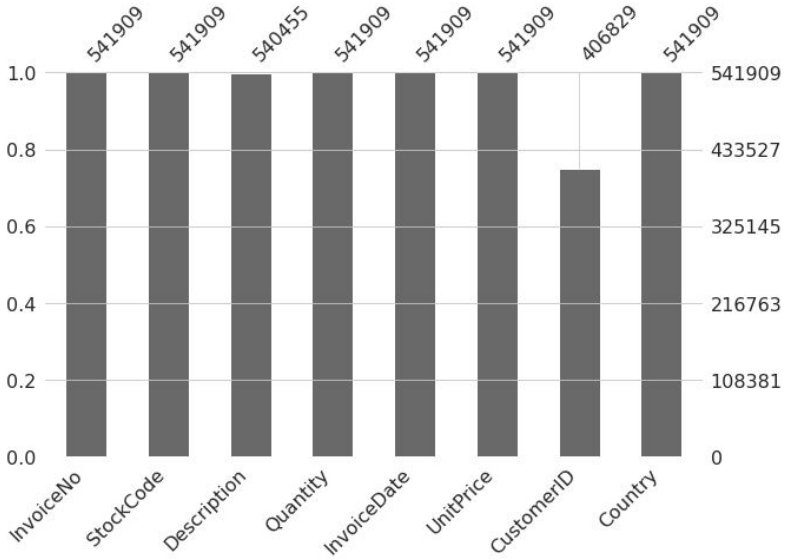
Problem Description

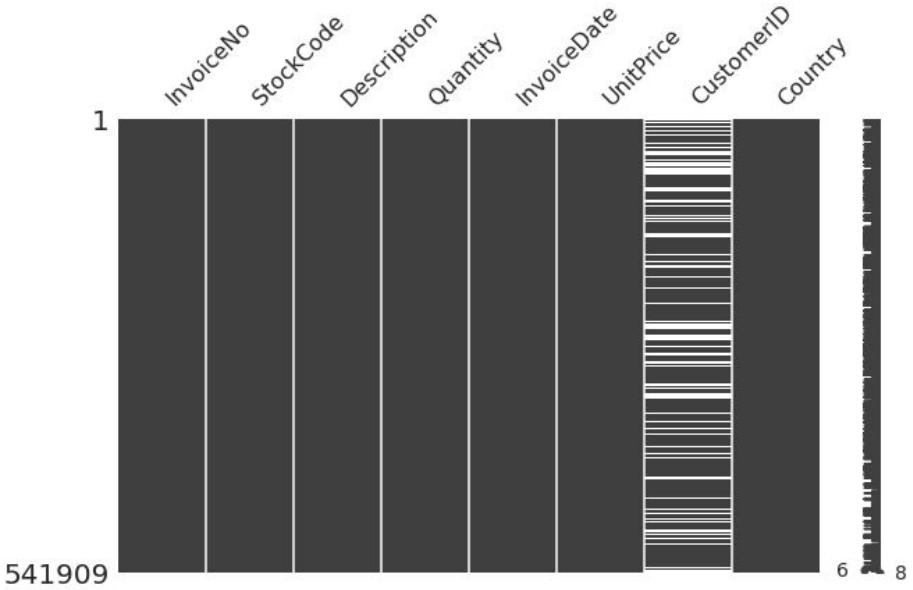
In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Description

- **InvoiceNo:** Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **Stock Code:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product (item) name. Nominal.
- **Quantity:** The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **Unit Price:** Unit price. Numeric, Product price per unit in sterling.
- **CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country:** Country name. Nominal, the name of the country where each customer resides.

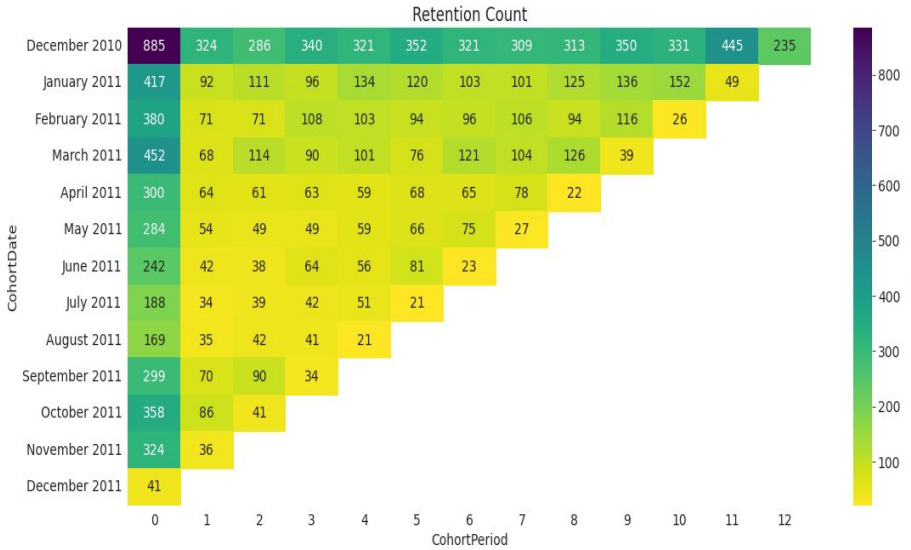
Looking for missing values

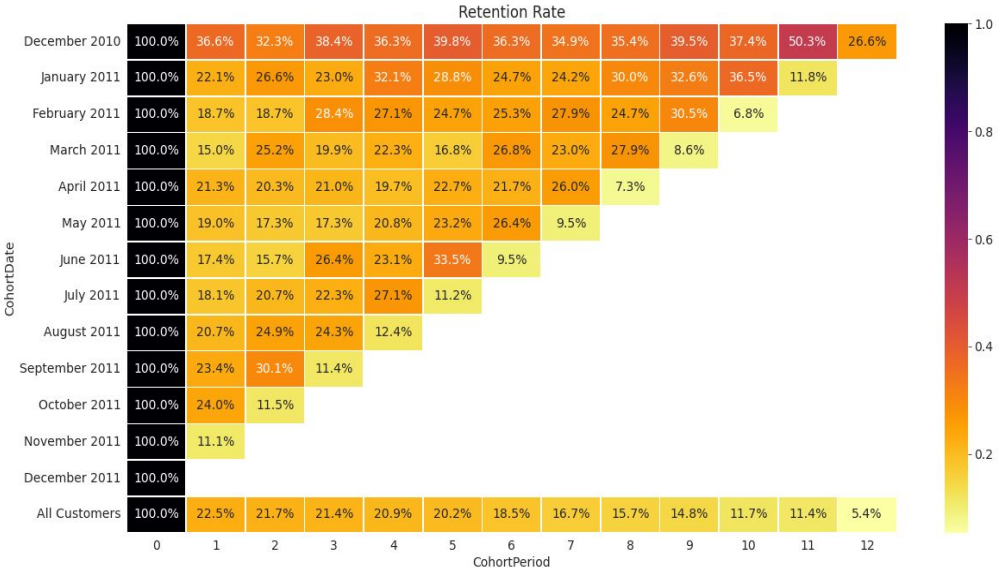




- White line in heatmap represent the position of null values in dataframe.
- Almost 25 percent of the data in Customer ID are missing and Description have only 0.27 percent of missing data.
- After removing the null values, there are 5225 duplicate observation in dataframe.

Cohort Retention Analysis

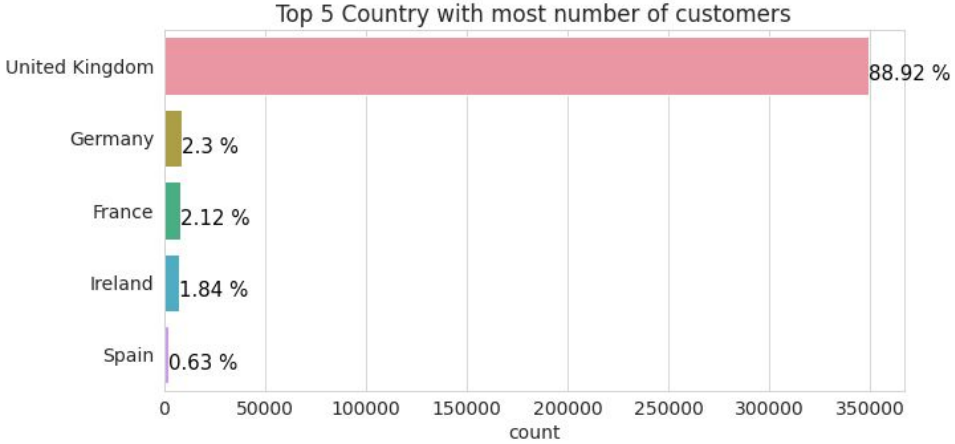


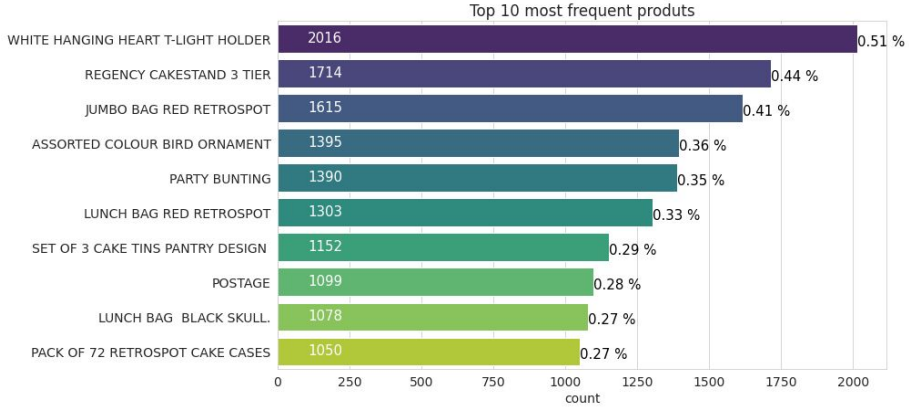


- On December 2010 there were 885 new customers out of which 334 customers which is 36.6% of 885 remains in next month and so on..
- Customer corresponds to December 2010 (as their first month of purchase) have highest retention rate among all customers.
- Retention rate for all users is monotonically decreasing, while the graph for cohort dates from December 2010 to September 2011 is not monotonic in nature

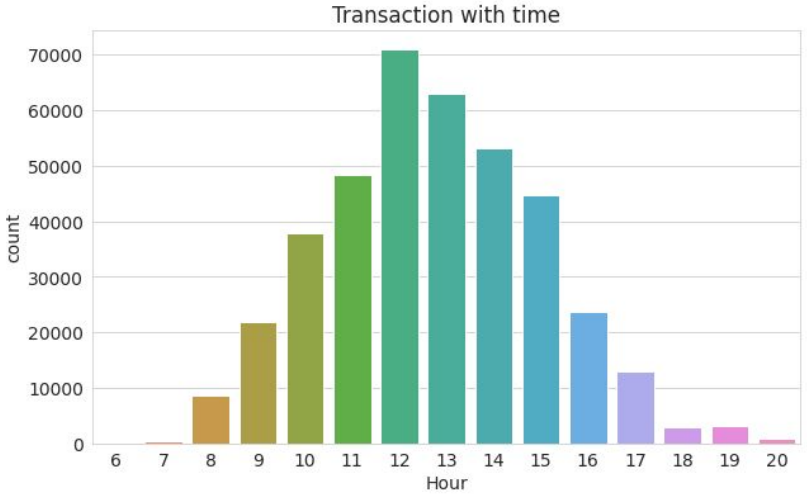
Exploratory Data Analysis

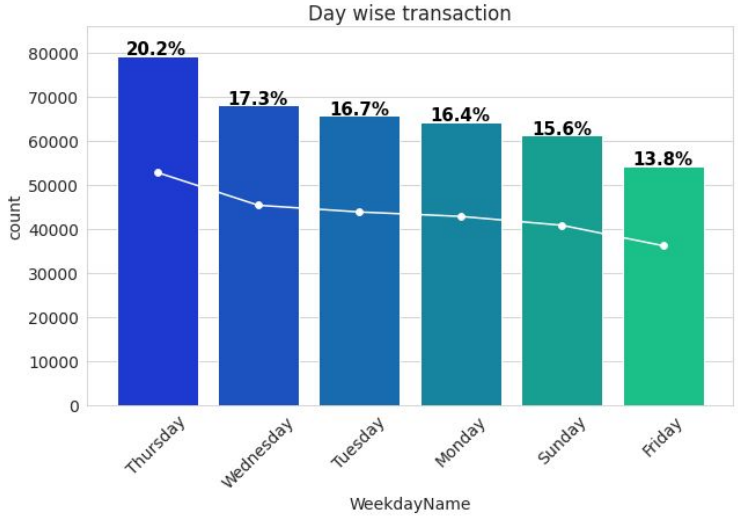
Top countries and Top popular products



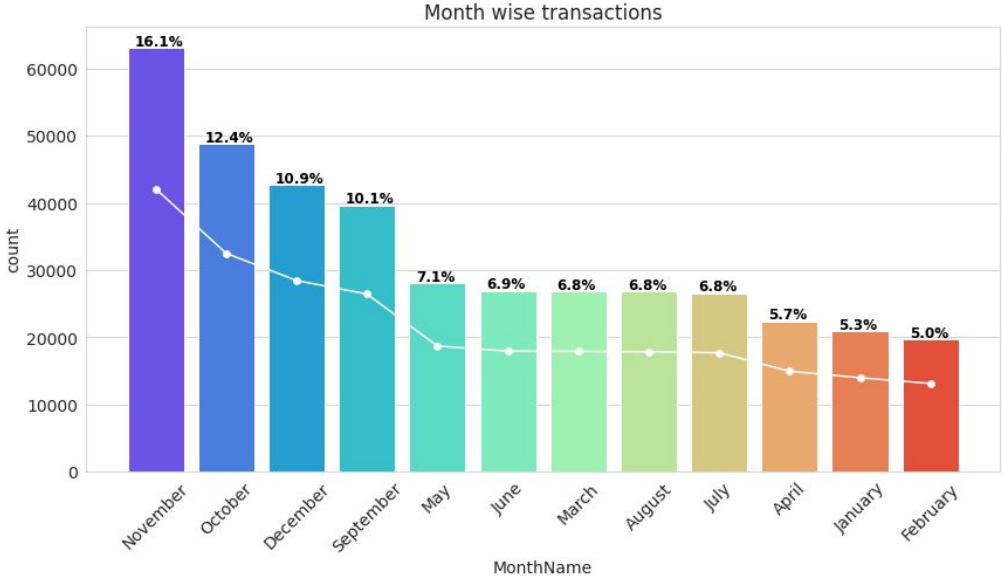


Since the company is based in the UK, most of its customers are from that country, followed by the top countries were also from Europe.

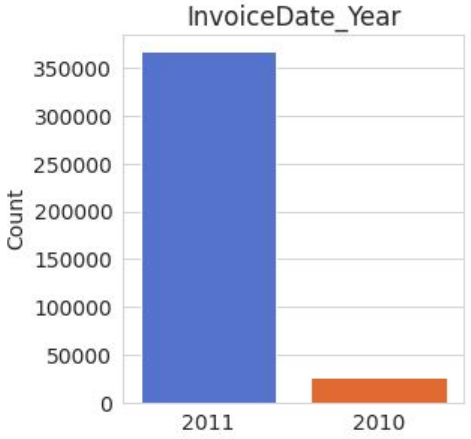


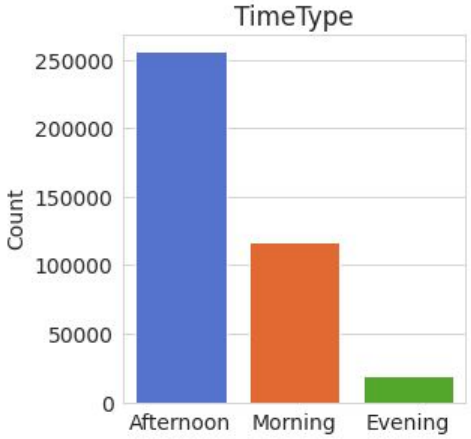


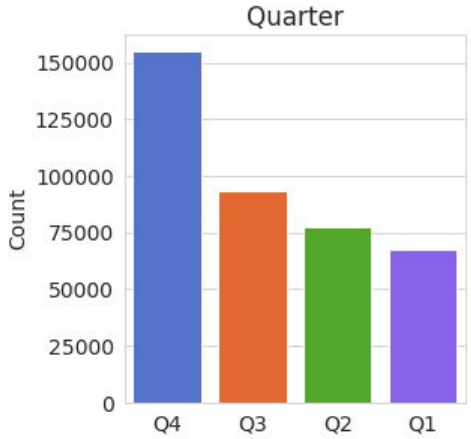
Month, Hour and Weekday

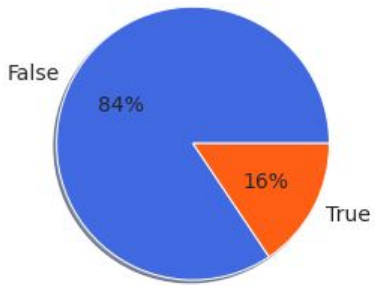


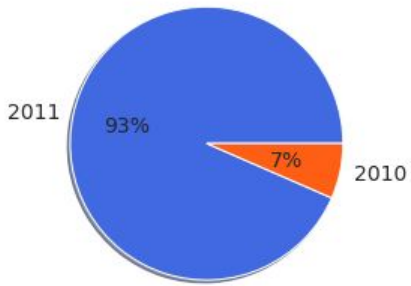
With other Categorical variable

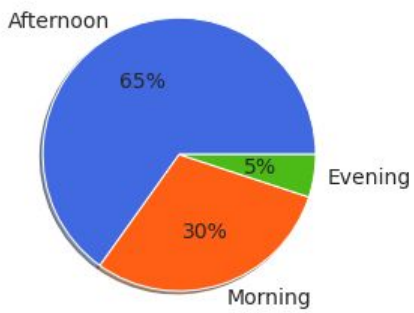


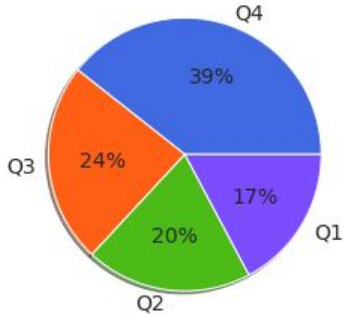




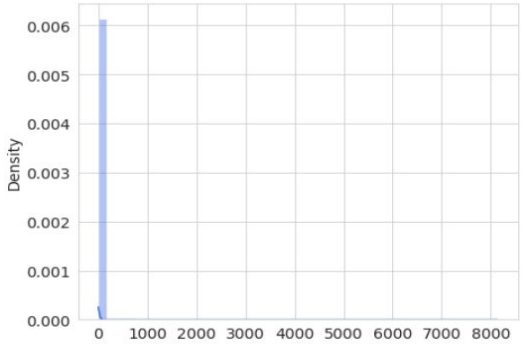




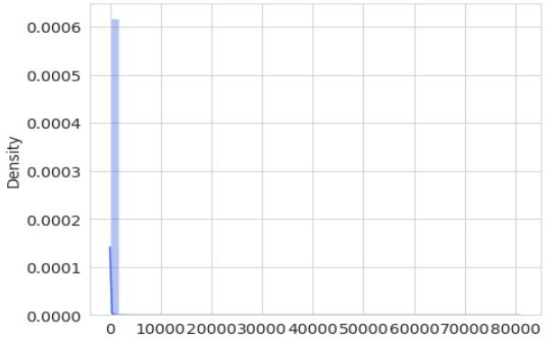




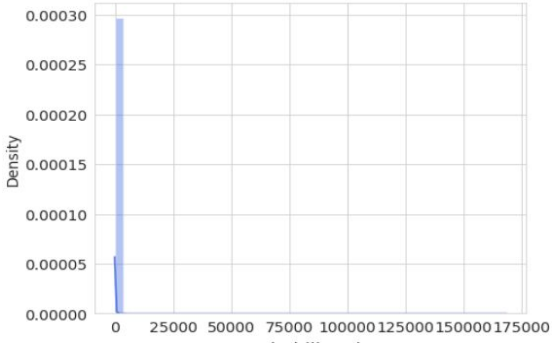
Quantity



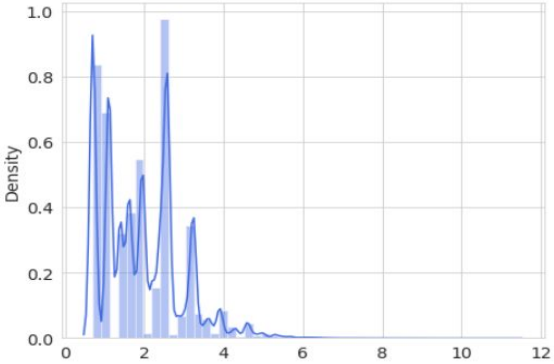
Unit Price

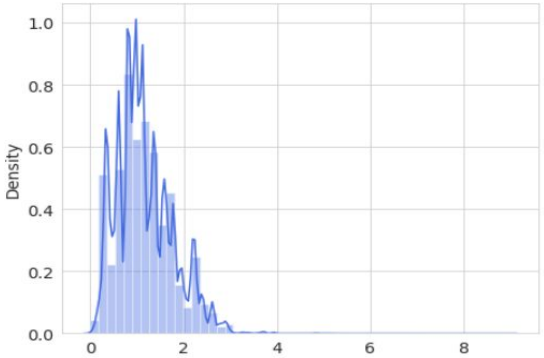


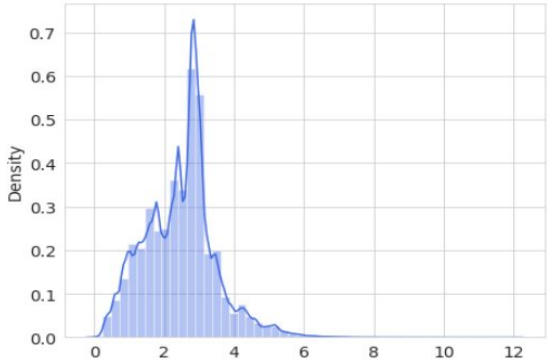
Total Amount

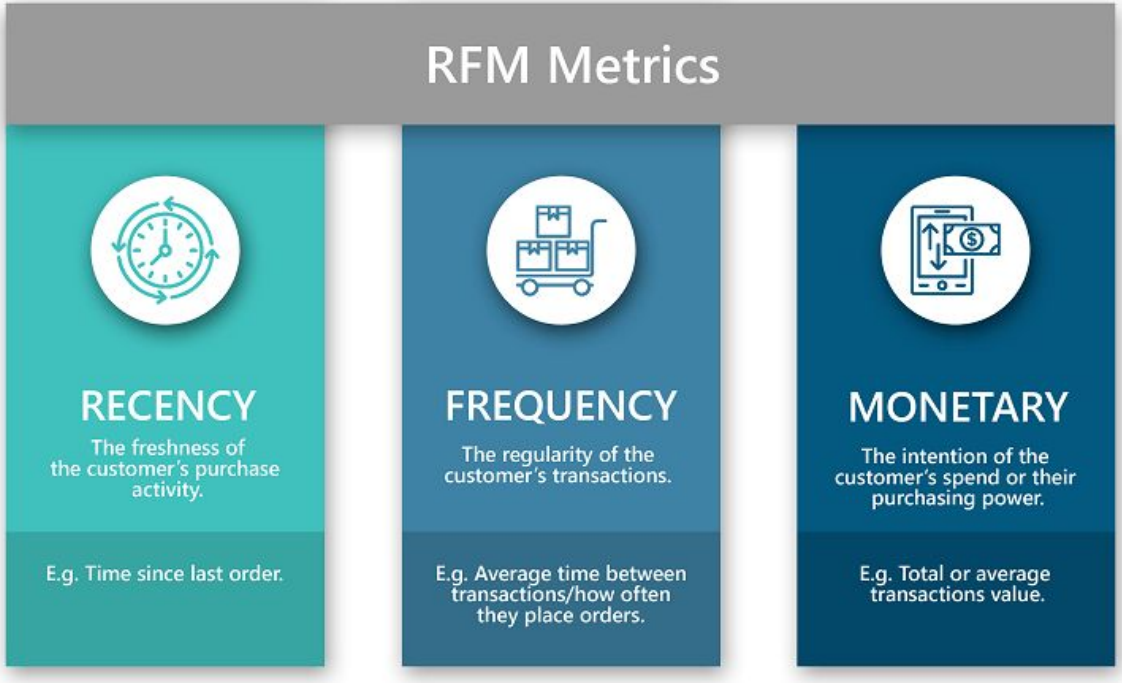


Logarithm





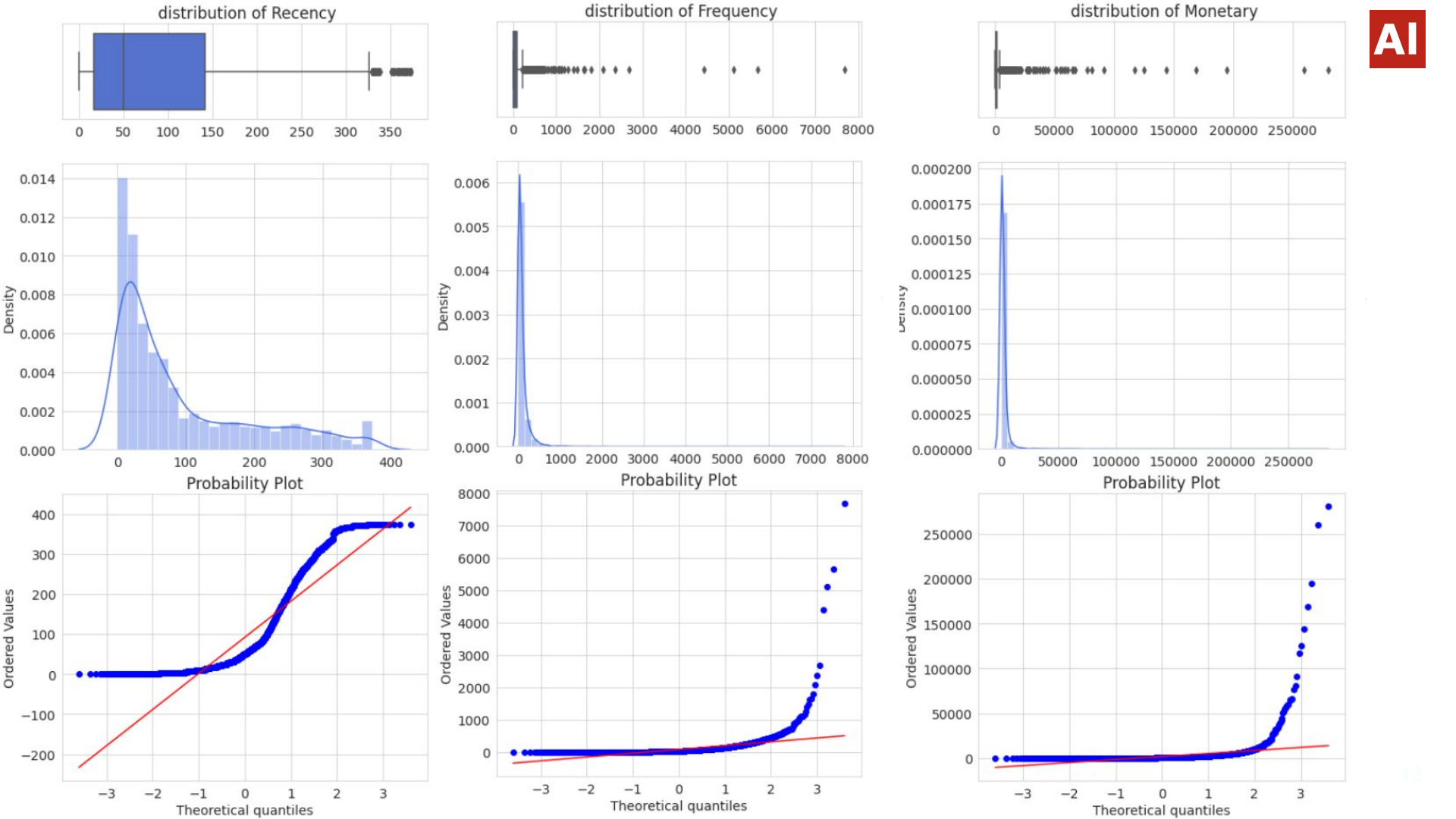


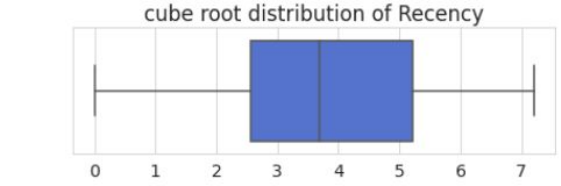


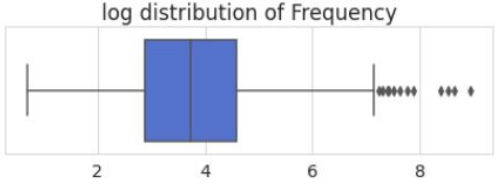
CustomerID	Recency	Frequency	Monetary
12346.0	325	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	18	73	1757.55
12350.0	310	17	334.40

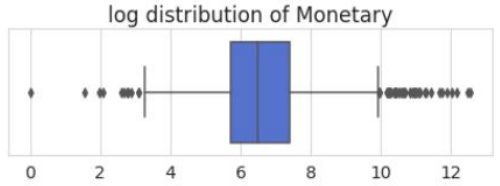
RFM is a method used for analyzing customer value. It is commonly used in database marketing and direct marketing and has received particular attention in retail and professional services industries. RFM stands for the three dimensions:

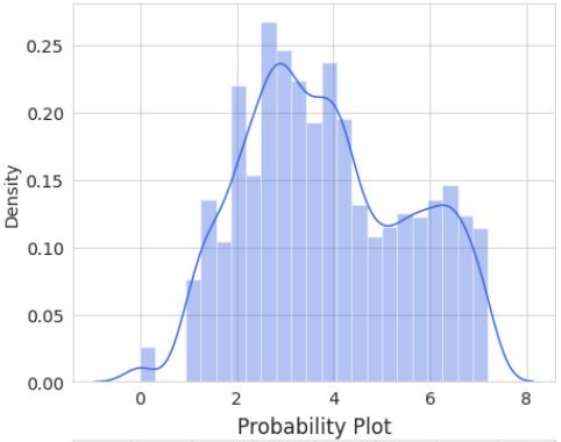
Recency – How recently did the customer purchase?
Frequency – How often do they purchase?
Monetary – How much do they spend?

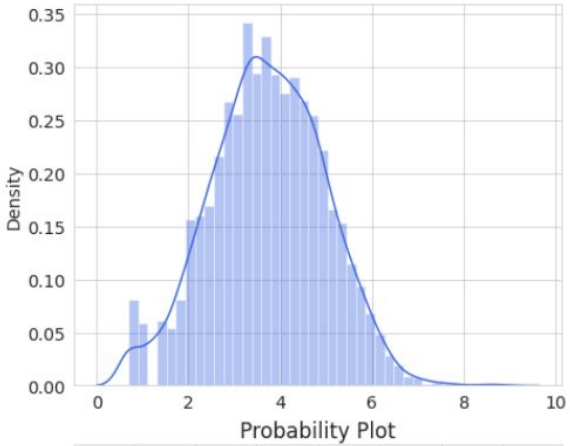

distribution of Recency
distribution of Frequency
distribution of Monetary
Density
Probability Plot
Ordered Values
Theoretical quantiles

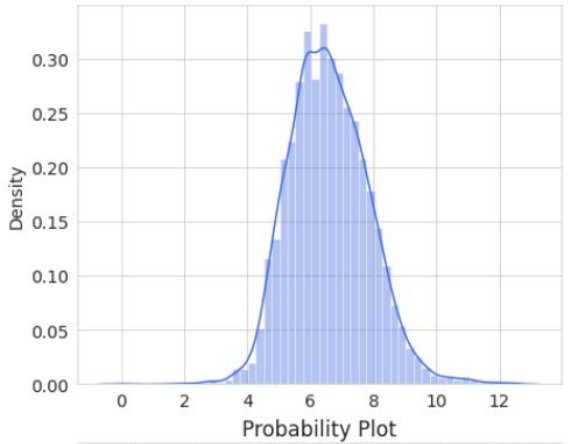

cube root distribution of Recency
0 1 2 3 4 5 6 7

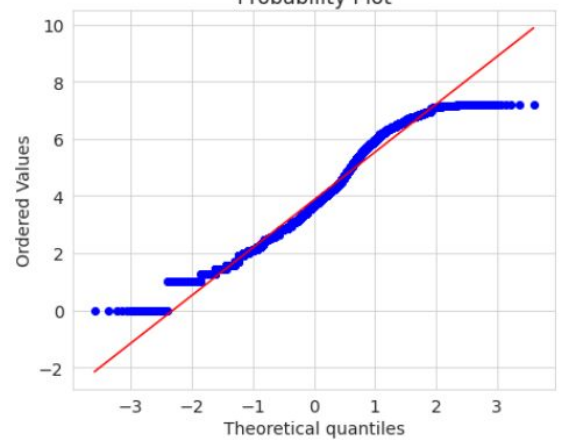

log distribution of Frequency
2 4 6 8

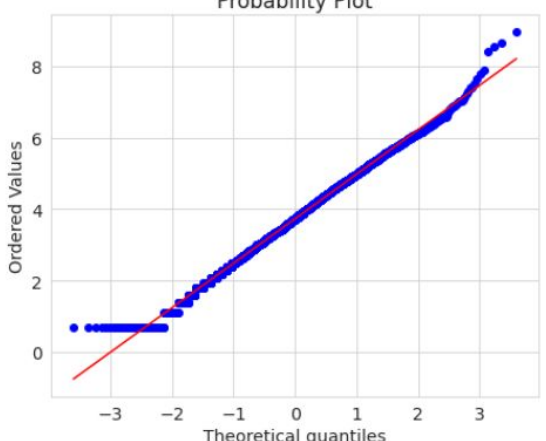

log distribution of Monetary
0 2 4 6 8 10 12

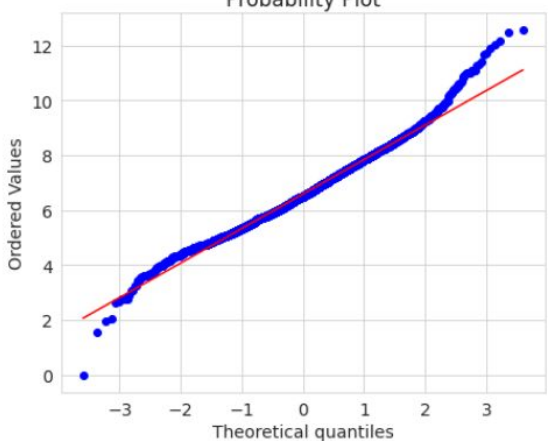

Density
0.25 0.20 0.15 0.10 0.05 0.00
0 2 4 6 8


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


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


Probability Plot
Ordered Values
10 8 6 4 2 0 −2
−3 −2 −1 0 1 2 3
Theoretical quantiles


Probability Plot
Ordered Values
8 6 4 2 0
−3 −2 −1 0 1 2 3
Theoretical quantiles

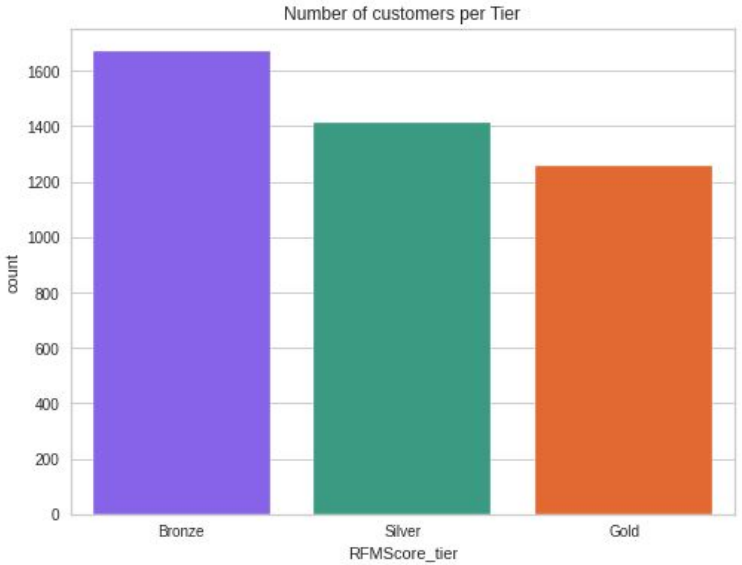

Probability Plot
Ordered Values
12 10 8 6 4 2 0
−3 −2 −1 0 1 2 3
Theoretical quantiles

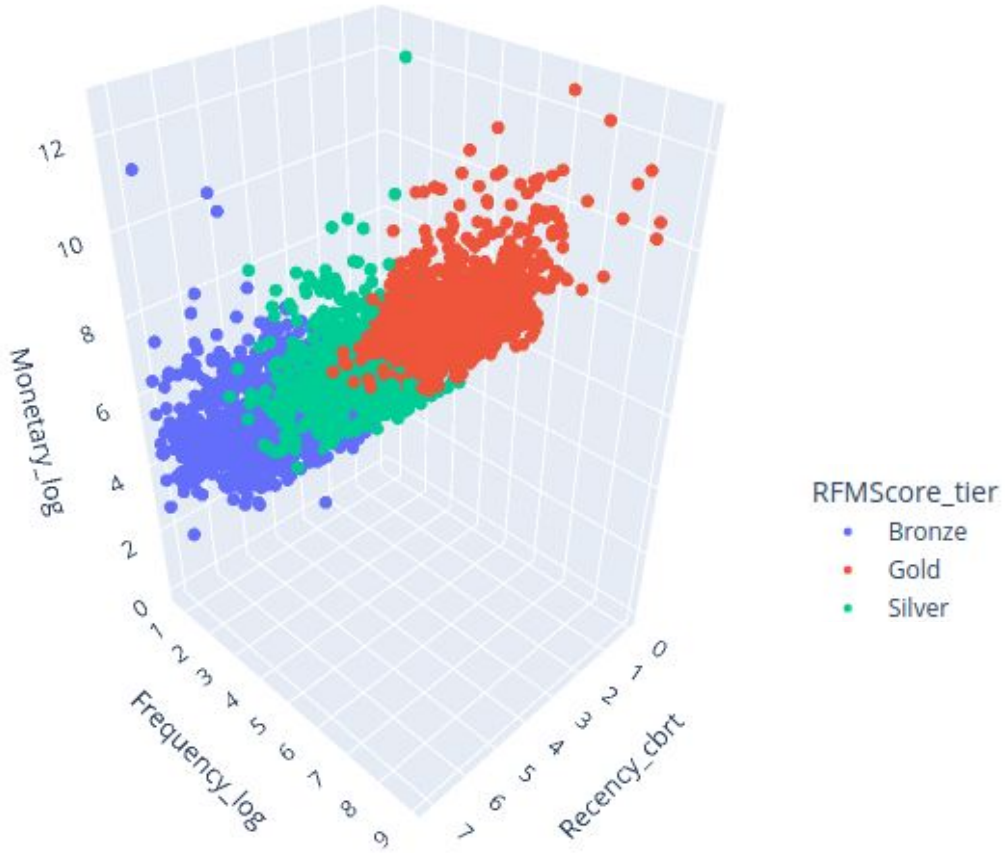
After getting the RFM values, a common practice is to create ‘quartiles’ on each of the metrics and assigning the required order. For example, suppose that we divide each metric into 5 cuts. For the recency metric, the highest value 5 will be assigned to the customers with the least recency value (since they are the most recent customers). For the frequency and monetary metric, the highest value, 5, will be assigned to the customers with the Top 20% frequency and monetary values, respectively. After dividing the metrics into quartiles, we can collate the metrics into a single column (like a string of characters {like ‘213’}) to create classes of RFM values for our customers. We can divide the RFM metrics into lesser or more cuts depending on our requirements

CustomerID	Recency	Frequency	Monetary	Recency_cbrt	Frequency_log	Monetary_log	R	F	M	RFMGroup	RFMScore	RFMScore_tier
12346.0	325	1	77183.60	6.875344	0.693147	11.253955	1	1	5	115	7	Bronze
12347.0	2	182	4310.00	1.259921	5.209486	8.368925	5	5	5	555	15	Gold
12348.0	75	31	1797.24	4.217163	3.465736	7.494564	2	3	4	234	9	Silver
12349.0	18	73	1757.55	2.620741	4.304065	7.472245	4	4	4	444	12	Gold
12350.0	310	17	334.40	6.767899	2.890372	5.815324	1	2	2	122	5	Bronze

Based on the RFM score I divide the customers into three tiers which help us to differentiate the customers.

- Gold
- Silver
- Bronze



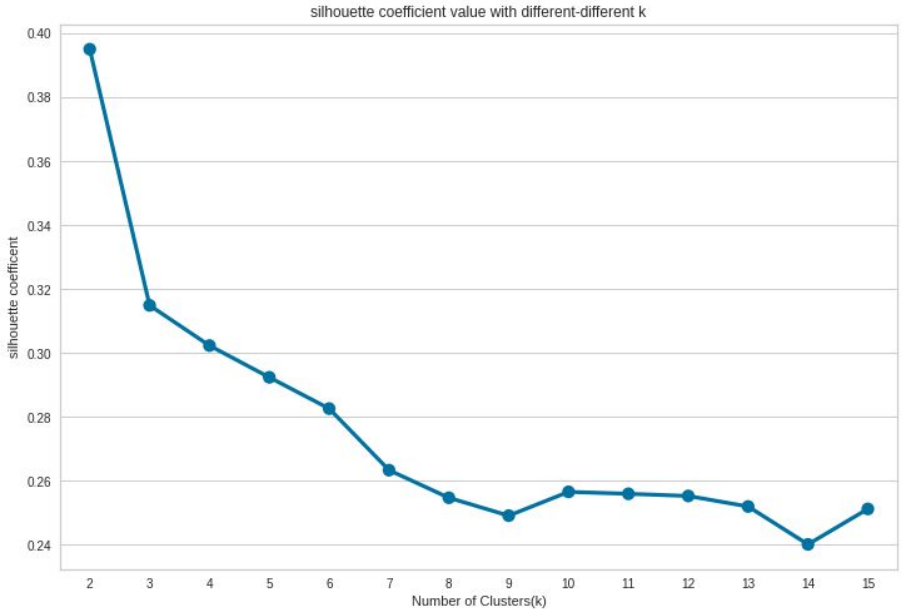


K-Means Clustering

Silhouette score method:

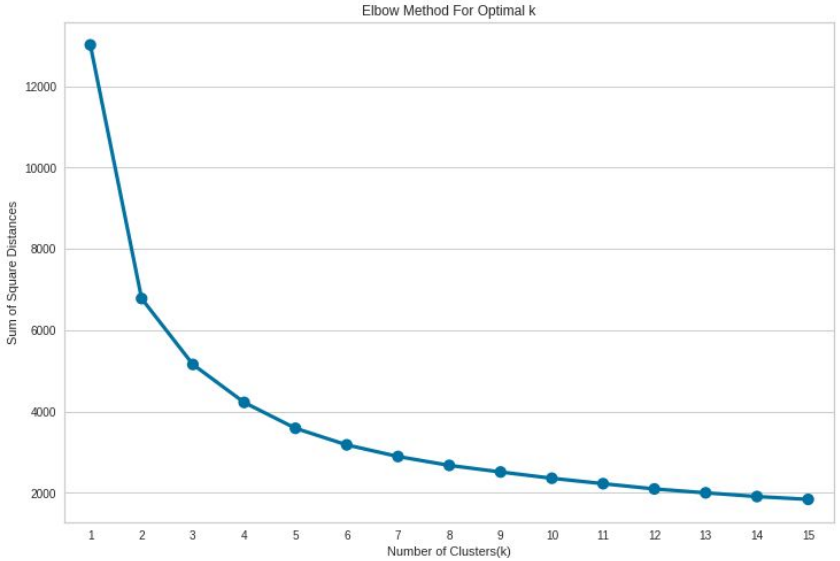
The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

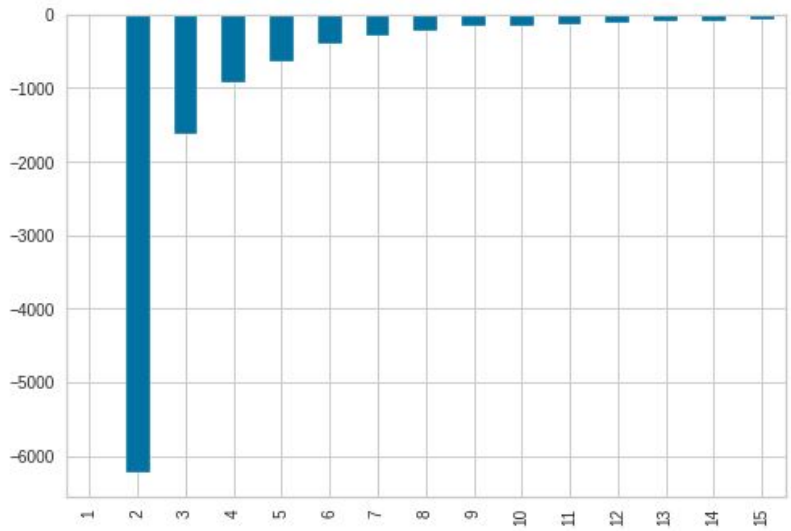
$$S = \frac{(b-a)}{max(a,b)}$$



for K = 2 the value of silhouette coefficient is maximum which is 0.3951

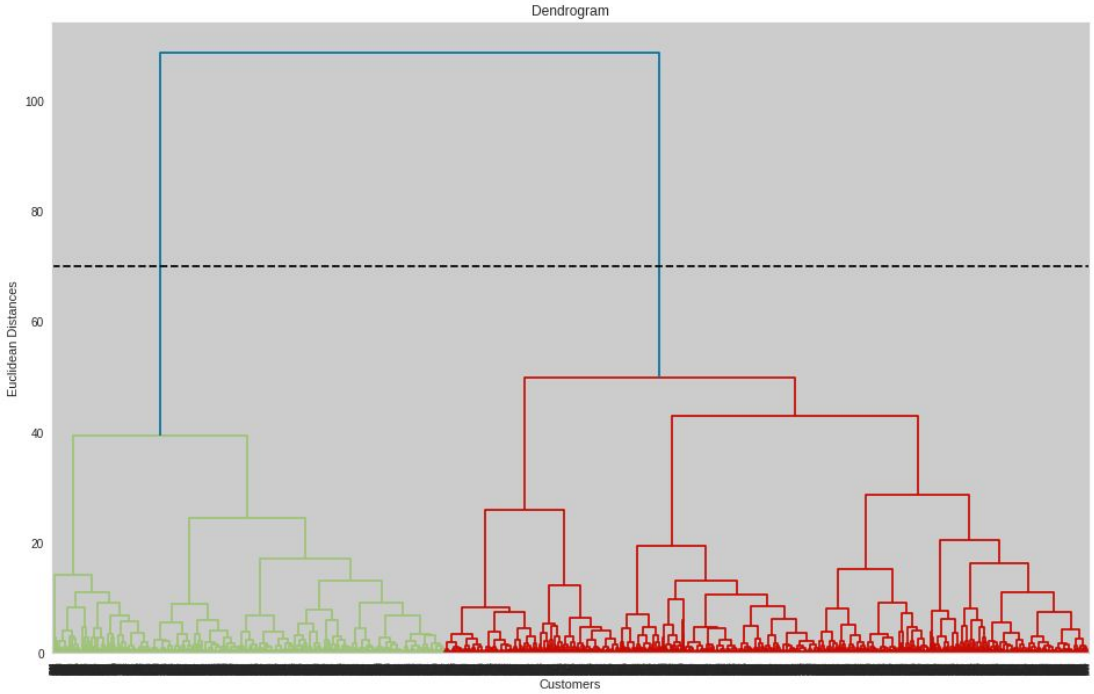
Elbow Method:





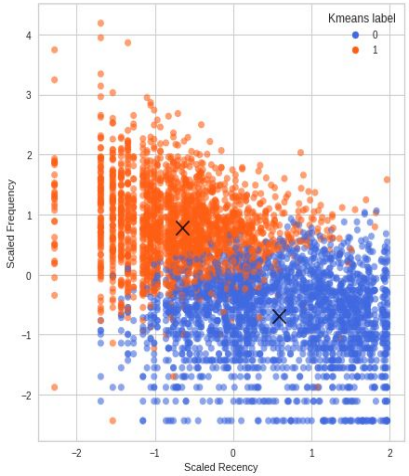
- *From the elbow method, it is clearly understood that, 2 clusters are performing best. Hence, 2 clusters will be selected to build the K Means model and classify the customers.*

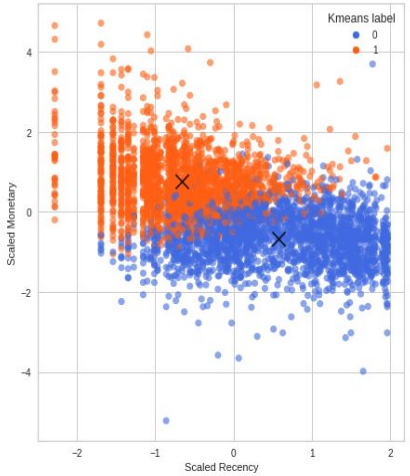
Hierarchical clustering

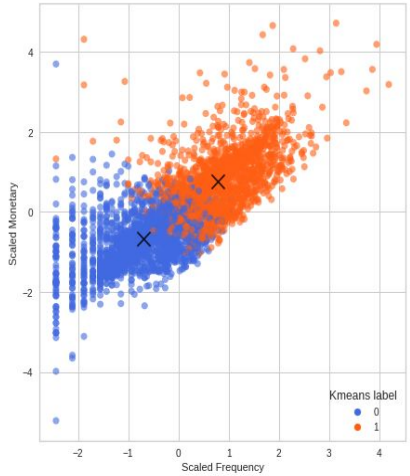


- Here, we can see No. of Clusters = 2 could be a good choice,
- 5 cluster would be my 2nd choice. 2 from green three from red.

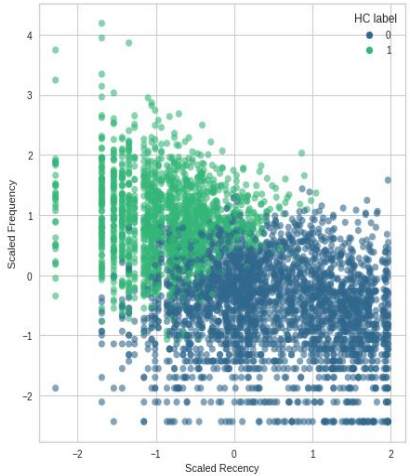
K-means cluster Analysis with 2 clusters

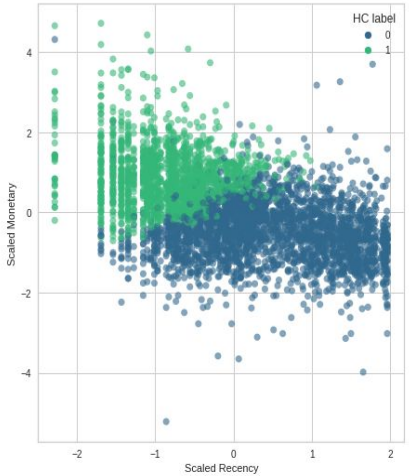


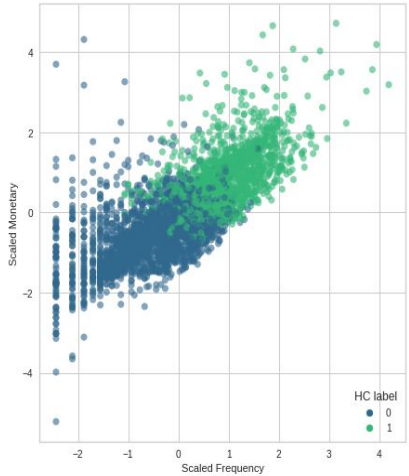




Hierarchical clustering cluster Analysis with 2 clusters







Conclusion

- This project mainly focused on developing customer segments for a UK based online store, selling unique all occasion gifts.
- Retention rate for all users is monotonically decreasing over time, while the retention rate for cohort dates from December 2010 to September 2011 is not monotonic in nature.
- Using a recency, frequency and monetary(RFM) analysis, the customers have been segmented into various clusters and got a silhouette score of 0.39 for two clusters
- By applying different clustering algorithm to our dataset, we get the optimal number of cluster is equal to 2.
- Since this is an unsupervised learning approach, there is no 100% correct answer, number of clusters will vary depending on the company's requirement.
- The business can focus on these different clusters and provide customer with services of each sector in a different way, which would not only benefit the customers but also the business at large.

Thank You!!!

Any questions?