Capstone Project

Online Retail Customer Segmentation

GOPAL JI GUPTA

Outline

1. Overview & Objective.
2. Data outline.
3. Exploratory data analysis
4. RFM segmentation
5. Model implementation
6. Conclusion

Overview & Objective

Overview

- Advancements in technology have made it possible for businesses to have a wider market and thus access a larger audience. Customer Segmentation refers to categorizing customers into different groups with similar characteristics .It can help businesses focus on each customer group in different way ,in order to maximize benefits for customers as well as for the business.

Objective

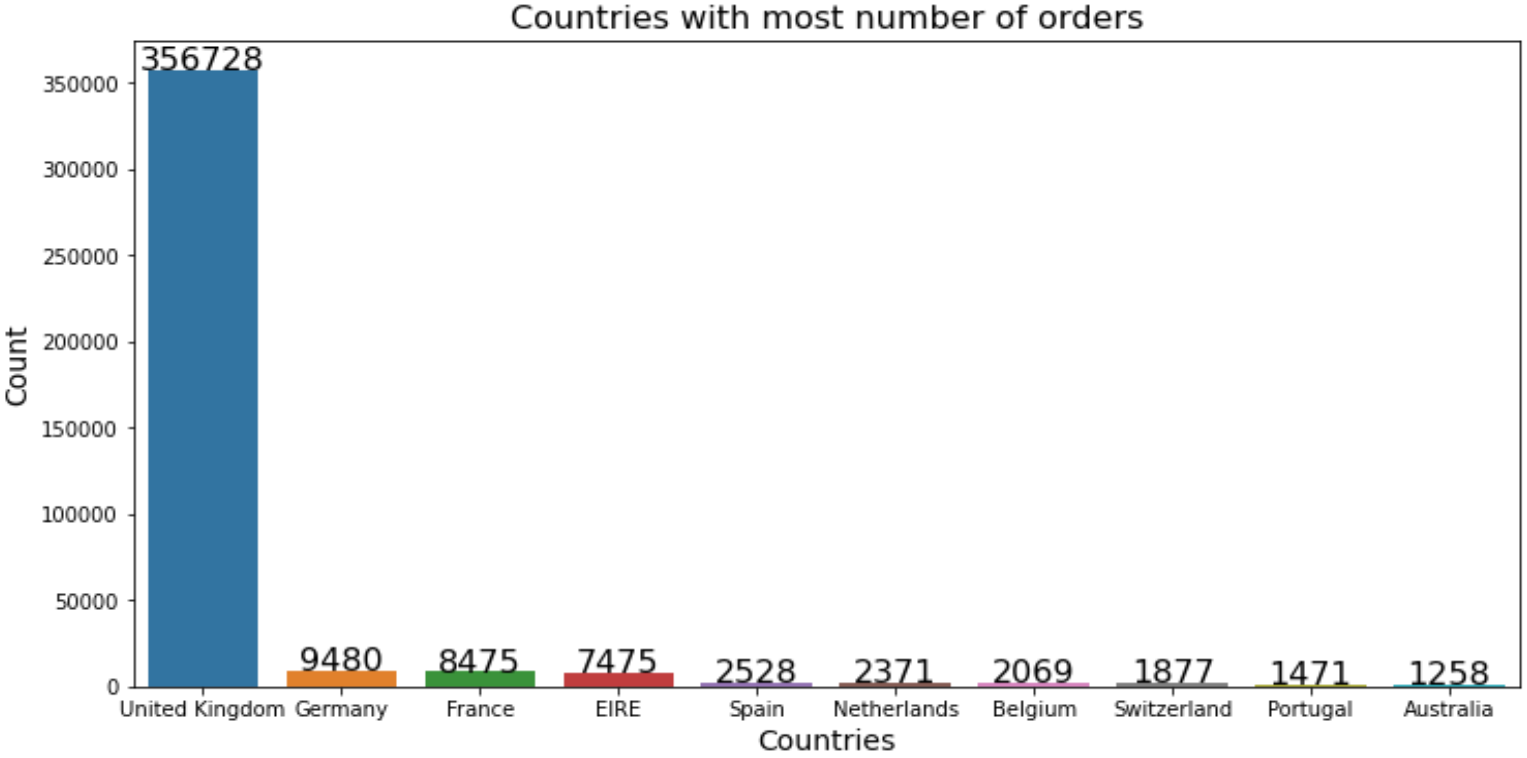
In this exercise, our task is to identify major customer segments on a transnational dataset which contains information regarding all the purchases and transactions between 01/12/2010 and 09/12/2011 for a UK based and registered non-store online retail. The company mainly sells unique all-occasion Gifts. Many customers of the company are wholesalers.

Data Outline

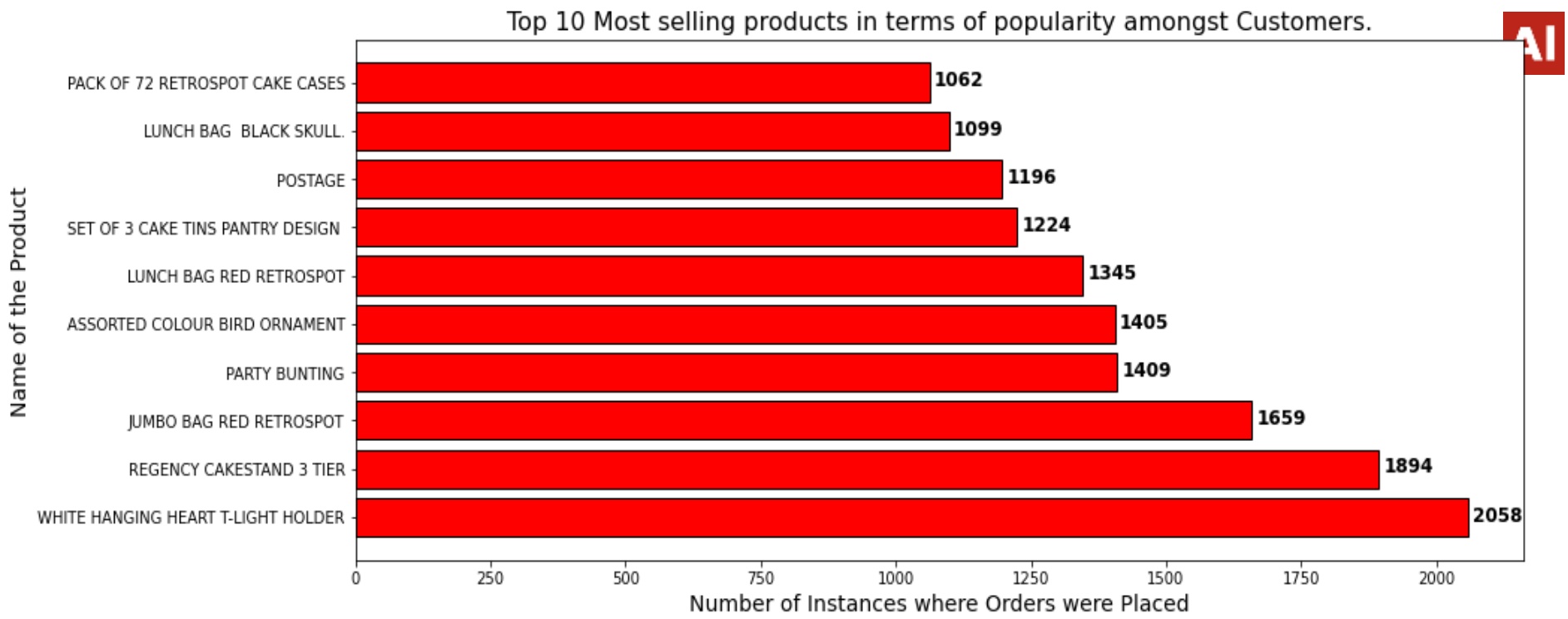
This dataset contain transactional data of online retail store. It contain 541908 rows and 8columns. Understanding the attributes of the dataset better:

- InvoiceNo : Invoice number – generated every time a transaction is made. If this code starts with letter ‘C’, it indicates a cancellation.
- StockCode: Product (item) code, a 5-digit unique number assigned to each distinct product.
- Description: Product (item) name.
- Quantity: The quantity of each item purchased in an order
- InvoiceDate: Invoice Date and time, the day and time when transaction was made
- UnitPrice: Unit price, Product price per unit in sterling.
- CustomerID: Unique customer identifier.
- Country: Country of residence of our Customers.

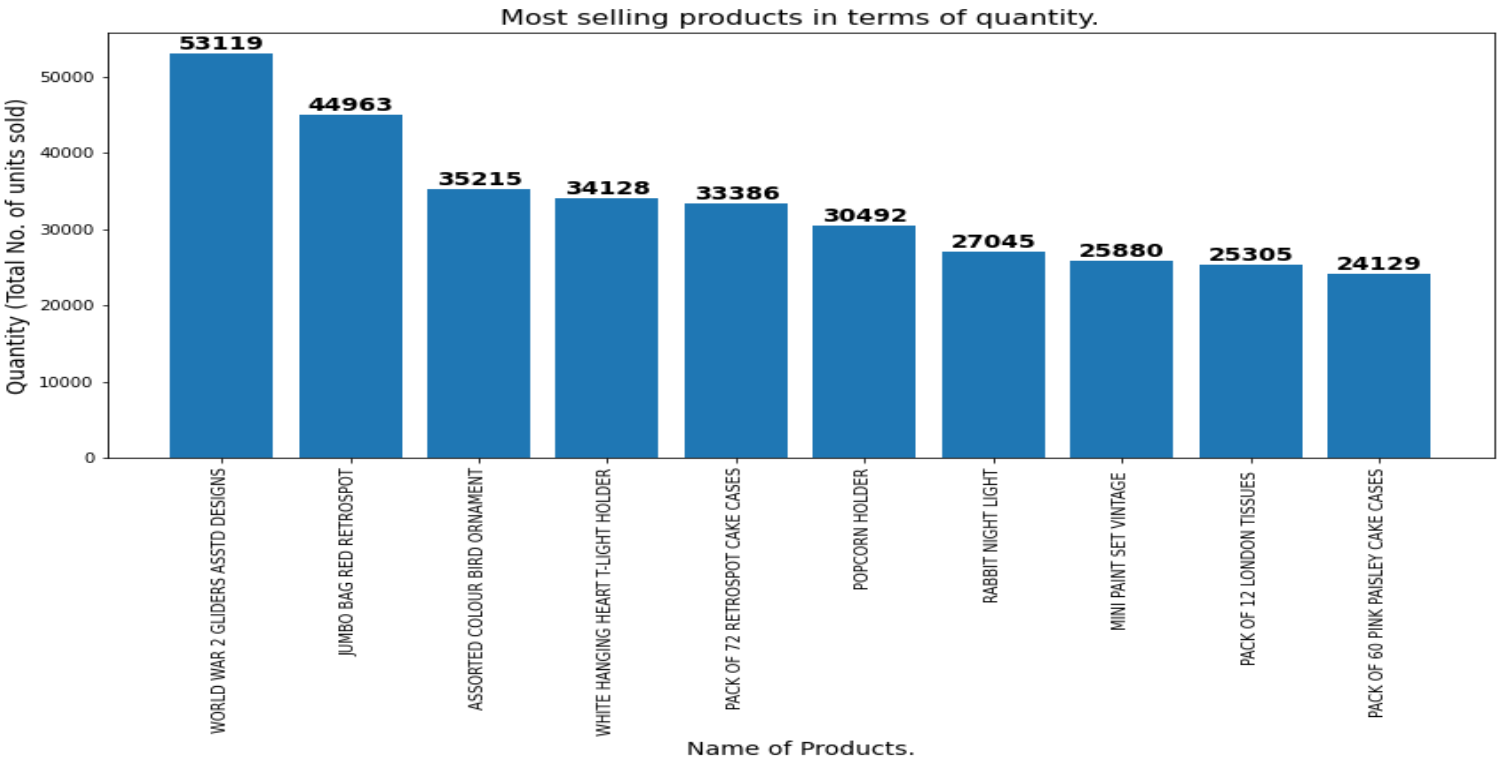
Exploratory Data Analysis



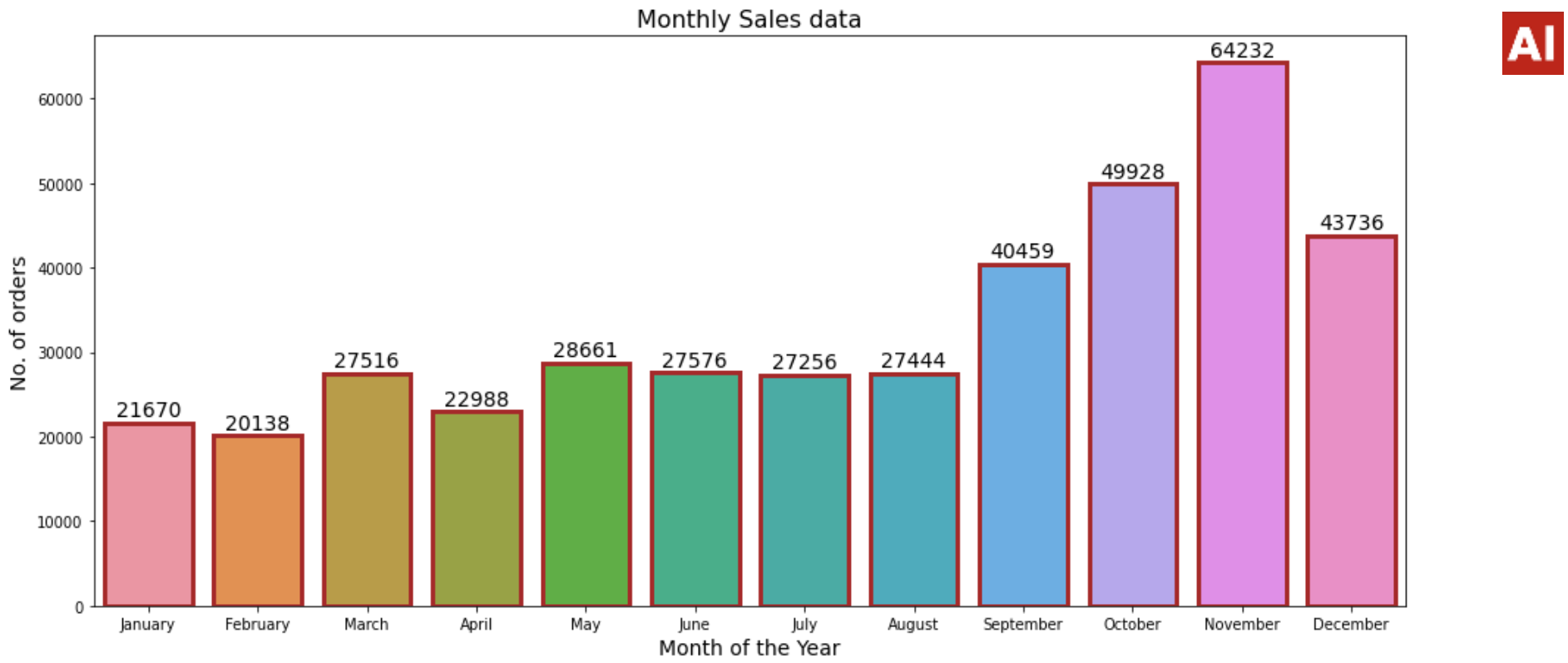
Here we have plotted the Countries whose residents have placed the most number of orders. From the graph above, we can clearly see that most of our orders were made by residents of the United Kingdom followed by Germany, France and EIRE (Ireland).



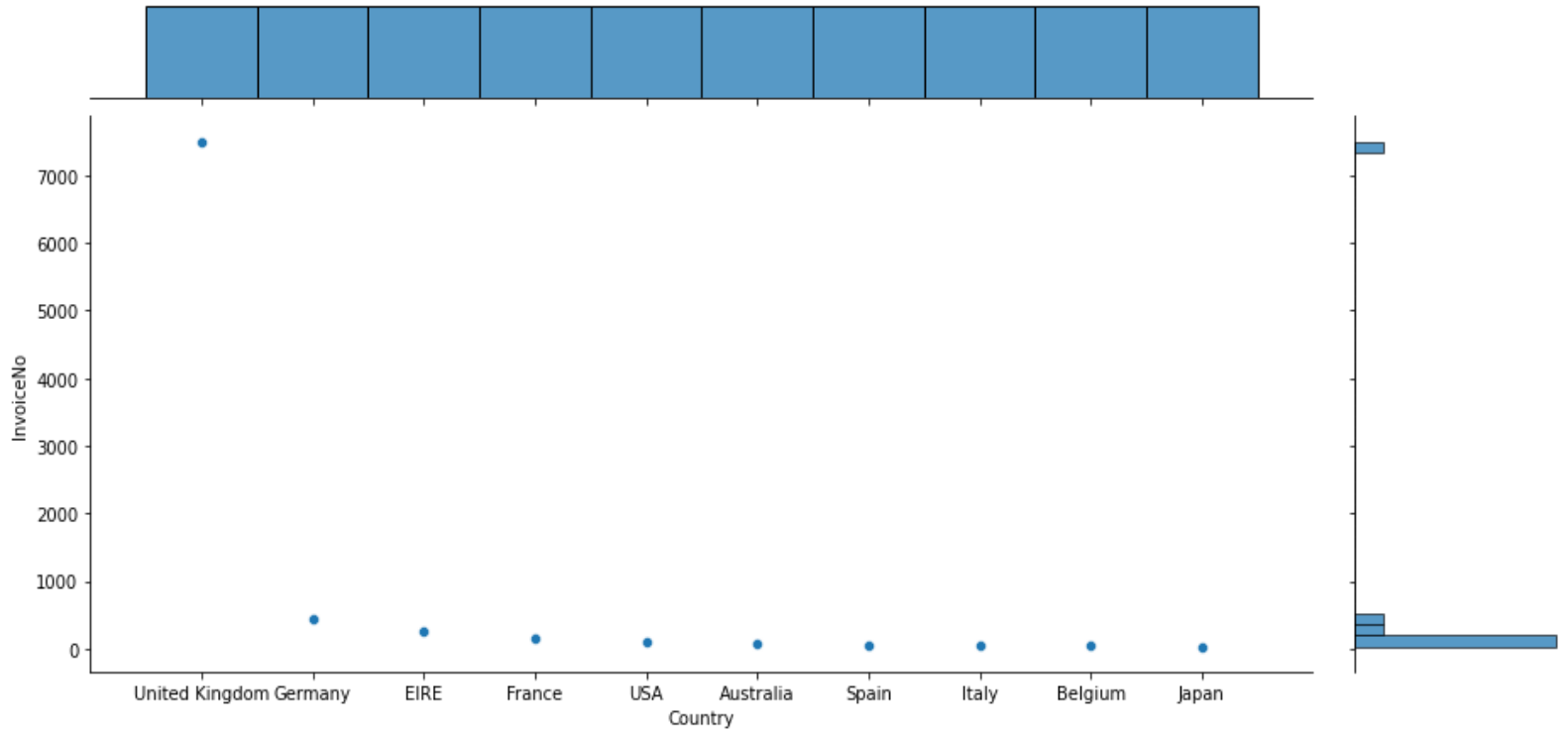
- Above graph indicates the most selling products in terms of their popularity i.e. – how many customers have bought them. We can clearly see that most selling products are – WHITE HANGING HEART T-LIGHT HOLDER and REGENCY CAKESTAND 3 TIER with 2058 and 1894 instances of orders respectively.



- The graph above, indicates the most selling products in terms of just sheer quantity i.e. total quantity of items purchased in total. We can clearly see from the graph below that the number of products that were sold in highest quantity are – WORLD WAR 2 GLIDERS ASSTD DESIGNS & JUMBO BAG RED RETROSPOT with 53119 and 44963 items sold respectively.

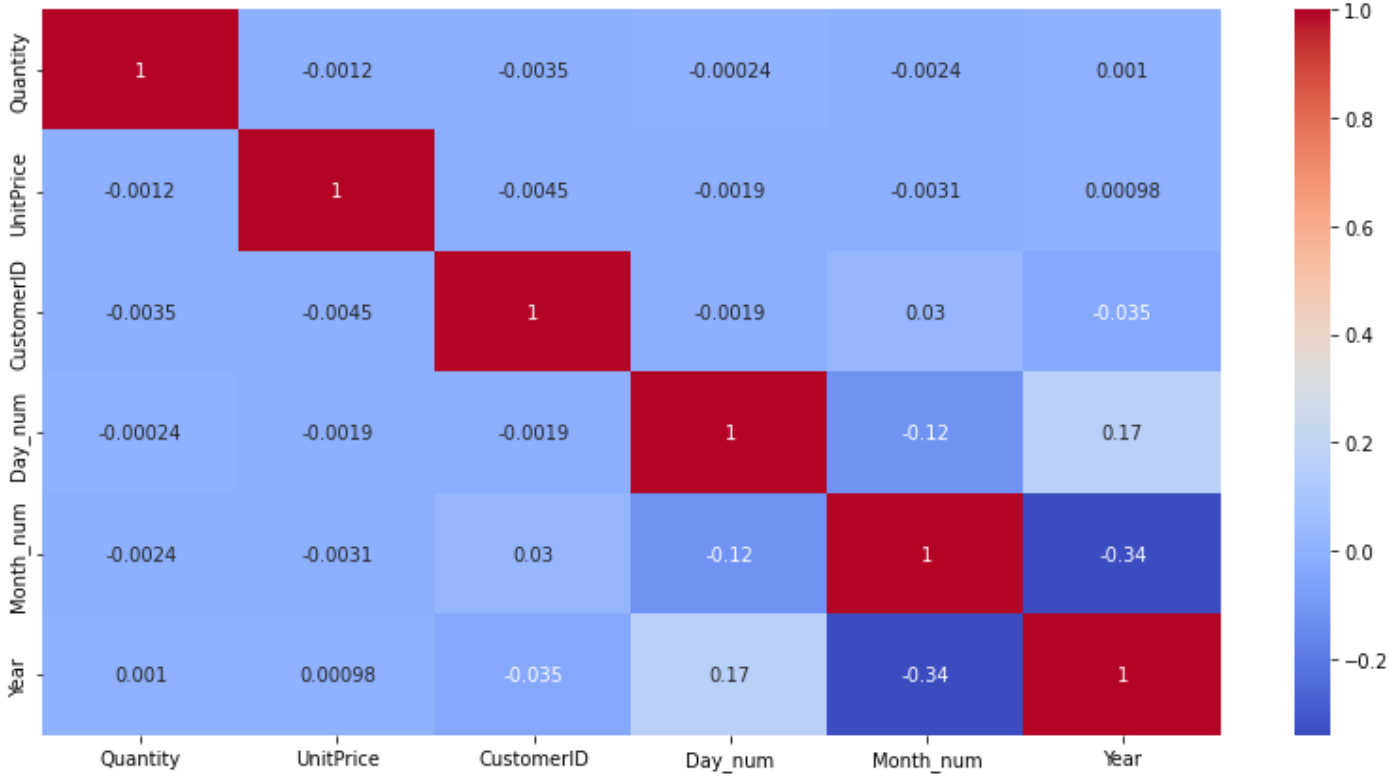


- From the above graph depicting the total no. of orders placed every month, we can see that most orders are placed during the winter months with November having highest orders at 64232 followed by October and December. Which makes sense because the festive season in Europe occurs during the winter season.



- In the graph above, we have performed analysis of number of cancellations versus the country of residence of customers. From the above graph, we can clearly see that most number of cancellations were made by the residents of United Kingdom which could be understood as they also made the most no. of orders.

Correlation Heatmap



Correlation heatmap visualizes the strength of relationships between numerical variables. The values in the cells indicate the strength of the relationship.

RFM segmentation

RFM stands for recency, frequency and monetary value

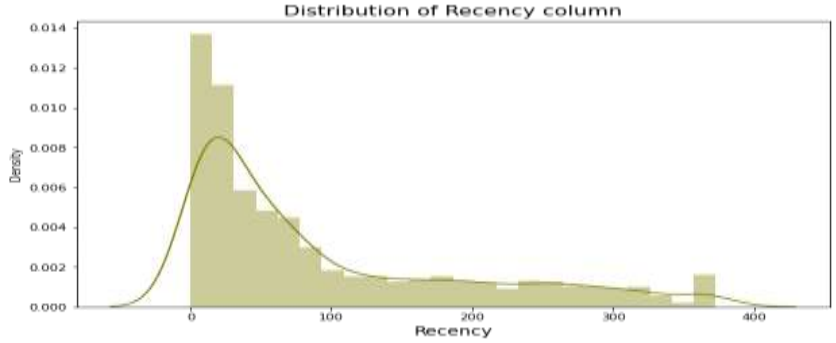
- • Recency signifies how recently has the customer ordered, frequency signifies the number of times the customer has ordered in a specific time frame and monetary value signifies the total amount spent by each customer for all their orders combined. We do these calculations for each customer and use these values to segment the customers in categories.
- • RFM based customer segmentation can help businesses focus on each customer group and enables them to use different strategies for separate segments in order to maximize benefits for customers and revenue for the business.
- • RFM segmentation is a powerful way to identify groups of customers for special treatment.
- •Then we perform scaling on our data because most clustering processes use distance between the datapoints as a measure of homogeneity and so that makes scaling very important for these processes otherwise, we would have biased results. So we use standardization using standardscaler for scaling this data.

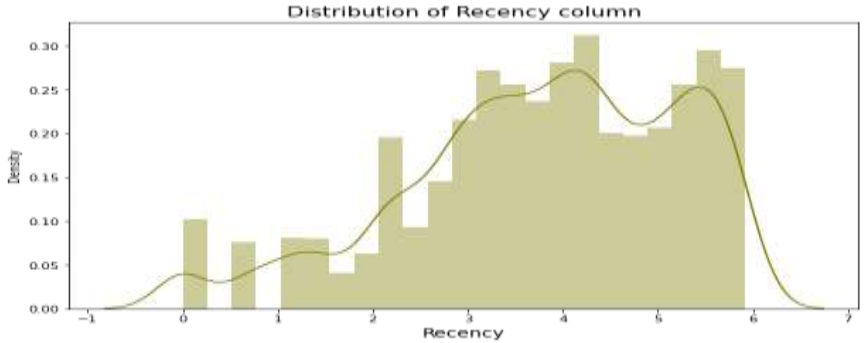
MODEL IMPLEMENTATION

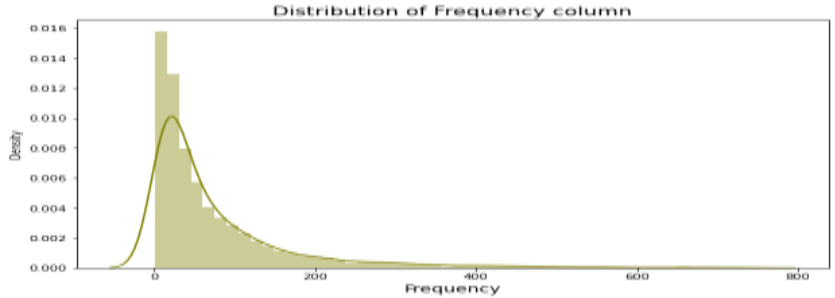
- Before starting with modelling, we check our RFM features distribution and found that they were positively skewed so we perform a log transformation on them and also we remove some outliers that we found.
- Next we move on to implementing ML algorithms. We have implemented four models :-

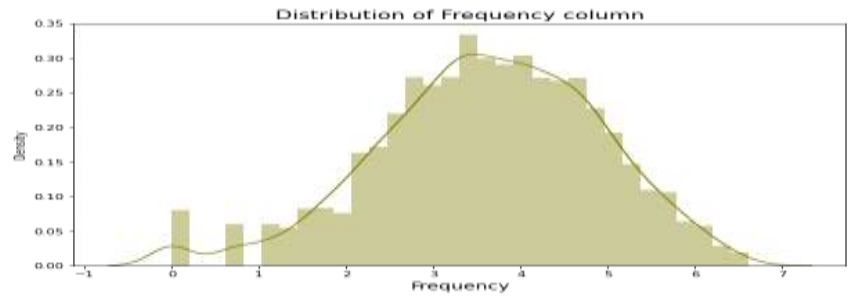
1. K means clustering with silhouette analysis.
2. K means clustering with elbow method
3. Agglomerative Hierarchical Clustering
4. DBSCAN

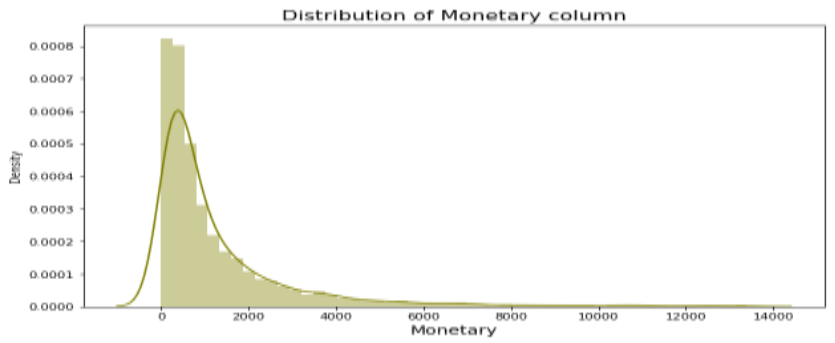
- Note – K means clustering is the algorithm used for clustering whereas silhouette score and elbow method are methods of choosing the optimal number of clusters.

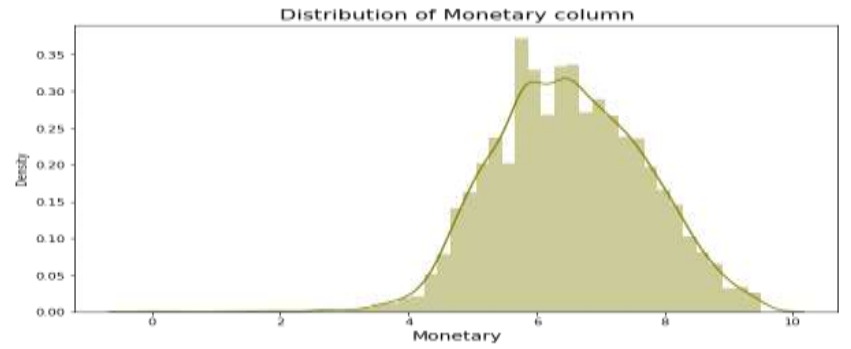




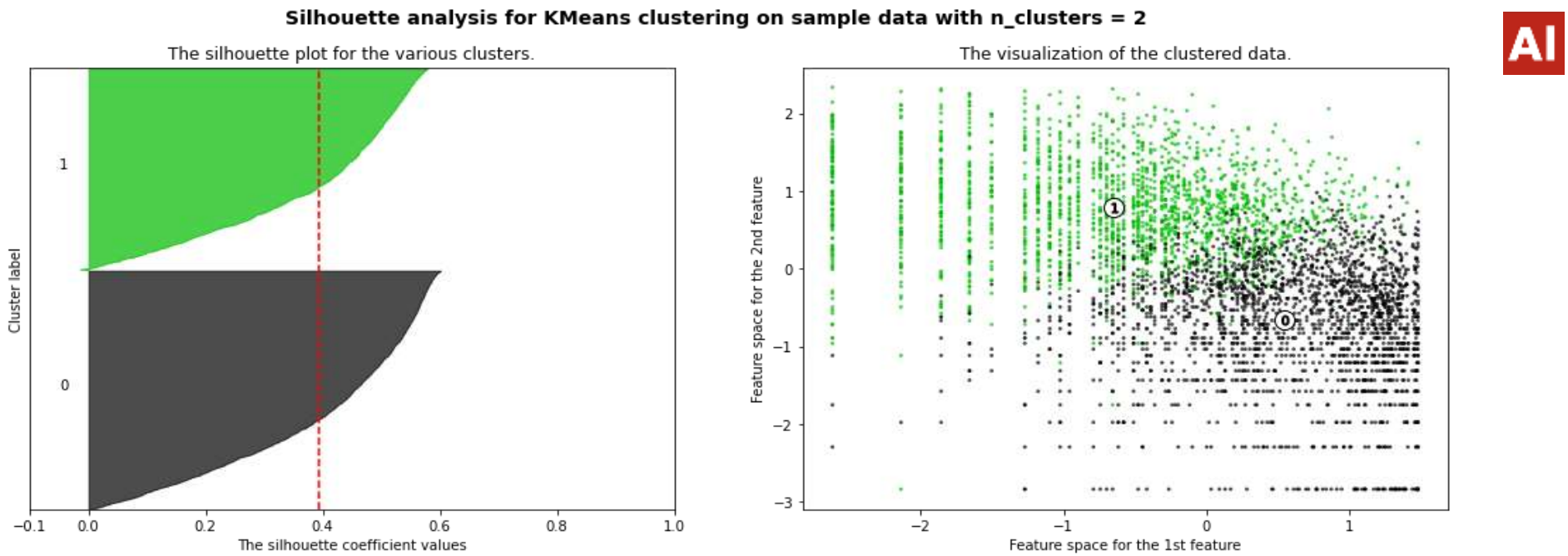






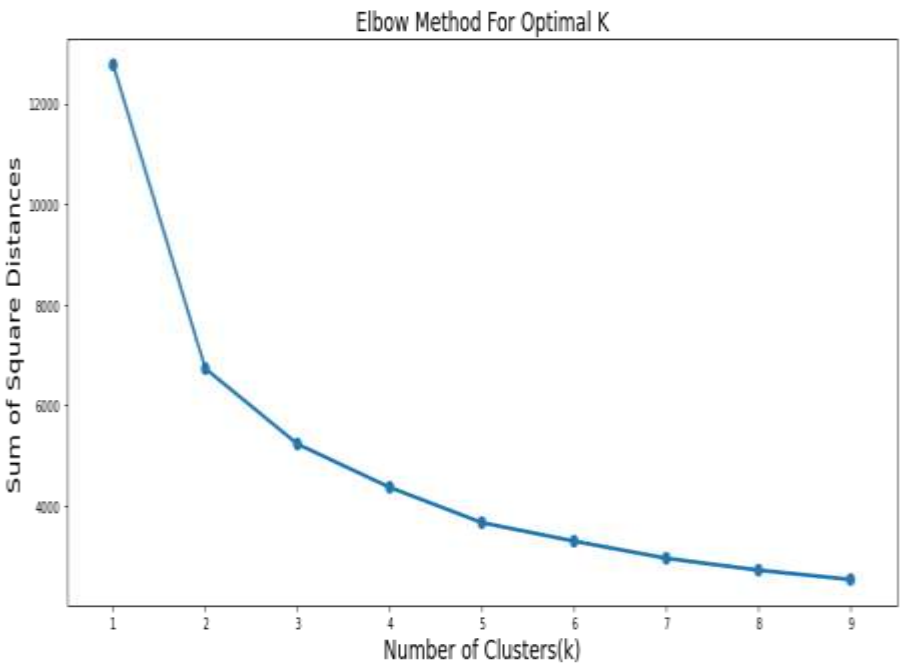


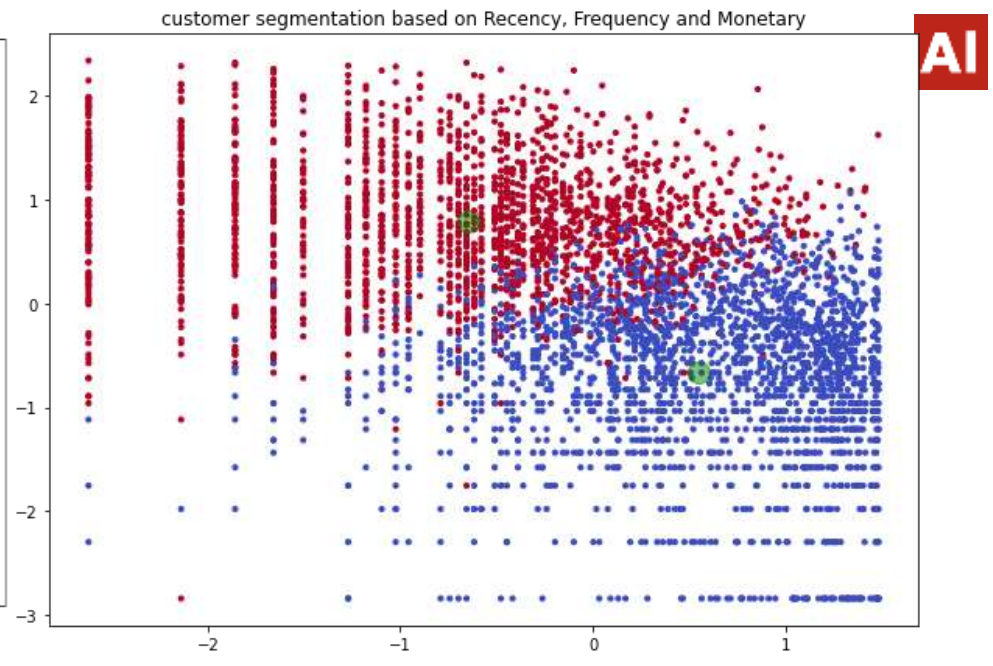
Log transformation of RFM variables – Before and After distributions.



Model Implementation: K means clustering with Silhouette analysis.

- K-means is a well-known clustering algorithm that is frequently used for unsupervised learning tasks. Silhouette analysis is used to study the separation distance between the resulting clusters . We checked the silhouette score for the number of clusters ranging from 2 to 8. The highest silhouette score obtained is 0.39 when K = 2 indicating that 2 is the optimal number of clusters.





Model Implementation: K means clustering with Elbow method.

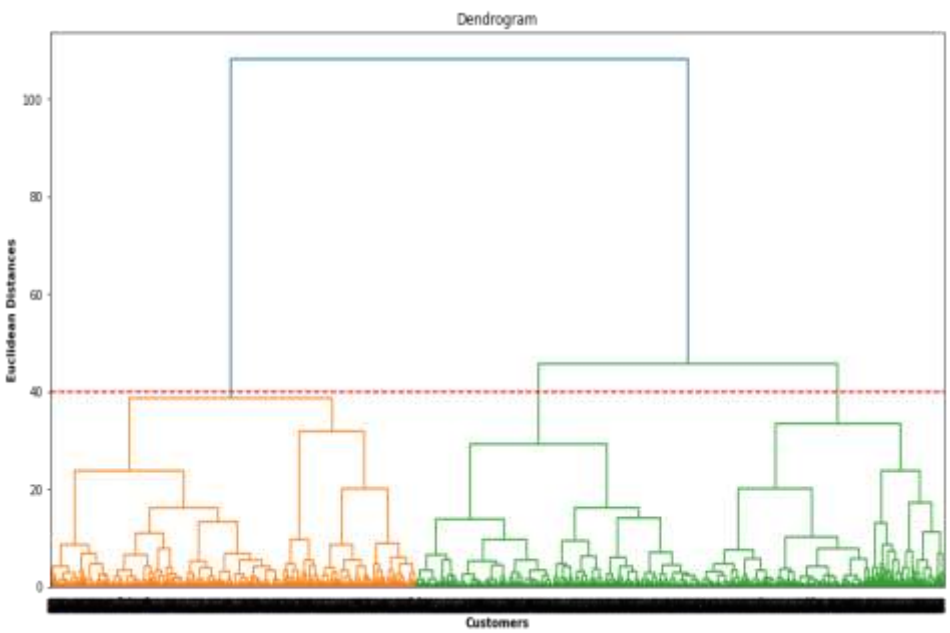
- Here, we built multiple no. of clusters upon our normalized RFM data .For each cluster, we have also extracted information about the sum of squared distances through which we built the elbow plot to find the desired number of clusters in our data. From the elbow graph, Clearly optimal no. of clusters is 2.
- We can clearly see the two clusters plotted in the 2nd graph. This is the result of using K-means clustering with K=2.

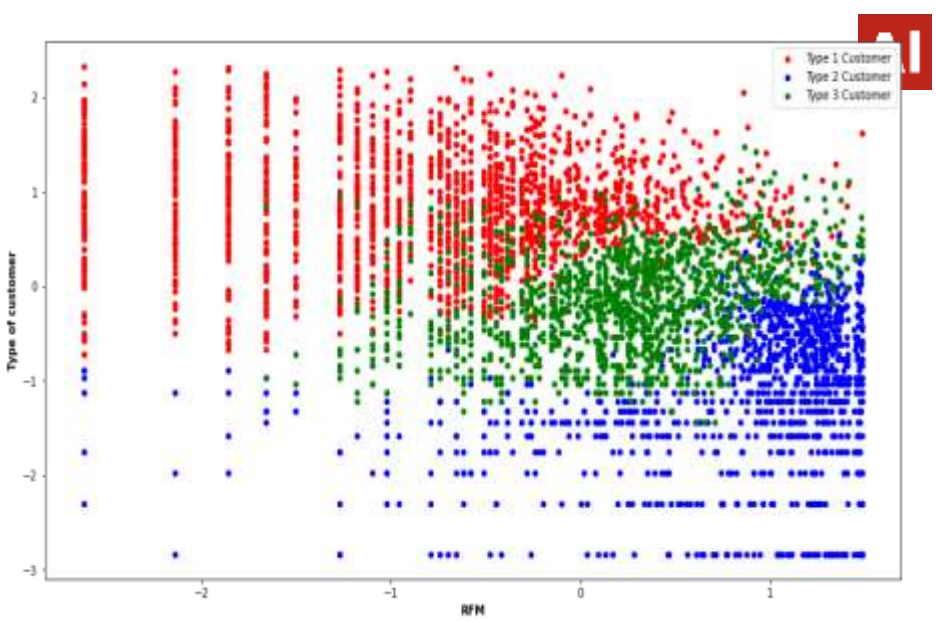
Cluster Profiling

Cluster	Recency	Frequency	Monetary	R	F	M	RFMScore	Recency_log	Frequency_log	Monetary_log
0	143.480694	23.199566	411.099827	3.101518	3.307592	3.278525	9.687636	4.570699	2.775255	5.724485
1	34.482829	137.630446	2365.316854	1.763711	1.574577	1.580215	4.918503	2.844539	4.632407	7.450798

Wholesale Customer: 'Cluster 1' is the high value customer segment as the customers in this group place orders with high monetary value and they order way more frequently than other members and they made a purchase recently. So we assume that these are the wholesale customers of retail store.

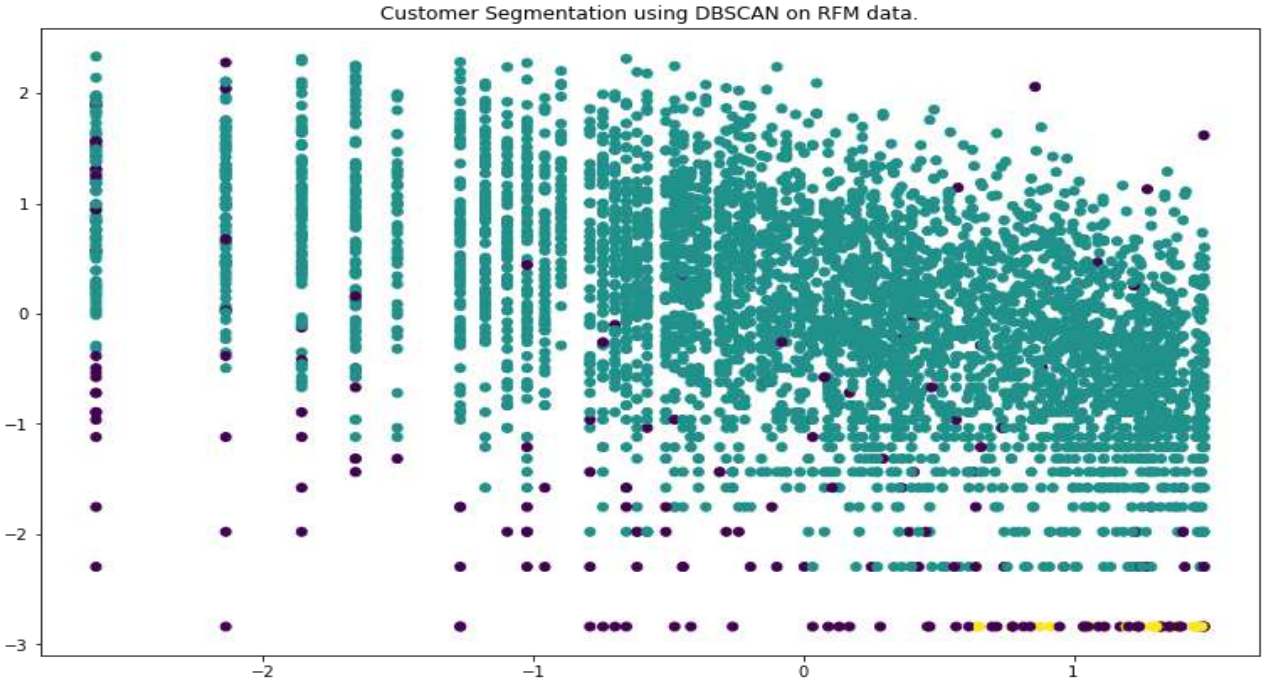
Average Customer: 'Cluster 0' is the average customer segment.These customers order less frequently as compared to wholesale customer and their orders do not have a high monetary value and its been a while since they last ordered with us. So these are our average customers.





Model Implementation: Agglomerative Hierarchical Clustering

- Here, we drew the dendrogram using the ward linkage to help us decide the number of clusters that we should use. The optimal no. of clusters is determined by drawing a horizontal line such that it cuts the longest vertical line. We can clearly see from the dendrogram that it makes sense to draw the line at threshold value of 40 which gives us the no. of optimal clusters as 3. Using K=3, we perform agglomerative hierarchical clustering and we can see the results – 3 separate clusters in the graph on the right.



Model Implementation: DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. DBSCAN algorithm identifies the dense region by grouping together data points that are close to each other based on distance measurement. It does not need to be initialized by providing the no. of clusters.
- From the graph above, we can see that there are 3 clusters formed.

Conclusions Drawn

- We saw that most number of orders were made by residents of United Kingdom which makes sense as it is the country where our business is based out of and we saw that UK residents have made the most no. of cancellations as well.
- We saw that there are around 8872 instances where an order was canceled.
- We saw that the winter months have the most sales with November, October and December having higher number of orders placed compared to the rest of the year. We also saw that most orders were placed on Thursdays and no orders were placed on Saturdays.
- We saw how we can segment our customers depending on our business requirements. We performed Recency, Frequency and Monetary value analysis for our entire customer base and used it to rank our customers.
- RFM analysis can help in answering many questions with respect to our customers and this can help companies to make marketing strategies for their customers, retaining their at risk of leaving customers and providing recommendations to their customers based on their interest.
- **The optimal no. of clusters in Agglomerative Hierarchical clustering is 2.**
- **The optimal no. of clusters in K-means with silhouette score and Elbow method is 2.**
- **The optimal no. of clusters in Agglomerative Clustering with threshold value 40 is 3.**
- **The optimal no. of clusters in DBSCAN clustering is 3.**

THANK YOU