CREDIT CARD DEFAULT PREDICTION

TECHNICAL DOCUMENTATION

GOPAL JI GUPTA

Abstract:

Businesses all over the world are growing every day. With the help of technology, they have access to a wider market and hence, a large customer base and a chance to know their customers.
Customer segmentation refers to categorizing into different groups with similar characteristics. Customer segmentation can help business focus on each customer group in a different way, in order to maximize benefits for customers as well as the business.

On the basis of the Recency, Frequency, and Monetary model, we have segmented customers of the business into various meaningful groups using the k-means clustering algorithm and Hierarchical clustering & the main characteristics of the consumers in each segment have been clearly identified.

I. PROBLEM STATEMENT

With advancements in technology, the world around us has been transformed and continues to evolve still at a rather fast pace. This is especially true for Businesses because technological advancements have made it possible to know a lot about one's customers and understand their behavior and interests and strategize around it. This is where customer segmentation comes into play.
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 them 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 and most frequent buyers. It can also help us identify customers who we are at the risk of losing.

II. OBJECTIVE

In this project, our 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. The main purpose of this project is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. Also, as a result of this segmentation, we can target different segments with different strategies like we can formulate better relationships with our best customers to keep them happily engaged with our business, we can make efforts to retain the customers that we are at risk of losing and use several other strategies for different segments in order to maximize our customer base and sales revenue.

III. INTRODUCTION

Smart marketers understand the importance of “know thy customer.” Instead of simply focusing on generating more clicks, marketers must follow the paradigm shift from increased CTRs (Click-Through Rates) to retention, loyalty, and building customer relationships.

Instead of analyzing the entire customer base as a whole, it’s better to segment them into homogeneous groups, understand the traits of each group, and engage them with relevant campaigns rather than segmenting on just customer age or geography.

One of the most popular, easy-to-use, and effective segmentation methods to enable marketers to analyze customer behavior is RFM analysis.

RFM stands for Recency, Frequency, and Monetary value, each corresponding to a key customer trait. **Recency** signifies the how recently order was placed, **Frequency** signifies the number of times a customer ordered and **Monetary** signifies the total amount spent on the orders. These RFM metrics are important indicators of a customer’s behavior because frequency and monetary value affects a customer’s lifetime value, and recency affects retention, a measure of engagement.

RFM factors illustrate these facts:

- The more recent the purchase, the more responsive the customer is to promotions
- The more frequently the customer buys, the more engaged and satisfied they are
- Monetary value differentiates heavy spenders from low-value purchasers

IV. WHY CUSTOMER SEGMENTATION?

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. Customer segmentation has a lot of potential benefits. It helps a company to develop an effective strategy for targeting its customers. This has a direct impact on the entire product development cycle, the budget management practices, and the plan for delivering targeted promotional content to customers.

Customer segmentation can also help a company to understand how its customers are alike, what is important to them, and what is not. Often such information can be used to develop personalized relevant content for different customer bases. Many studies have found that customers appreciate such individual attention and are more likely to respond and buy the product. They also come to respect the brand and feel connected with it. This is likely to give the company a big advantage over its competitors. In a world where everyone has hundreds of emails, push notifications, messages, and ads dropping into their content stream, no one has time for irrelevant content.

Insights from customer segmentation are used to develop tailor-made marketing campaigns and for designing overall marketing strategy and planning which add another dimension to sales and marketing efficiency of these businesses.

V. DATA DESCRIPTION

The data used in this project is a transactional dataset 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. The transaction dataset of this online retail store has 8 variables as shown below, and it contains all the information about orders placed and transactions for the aforementioned time period.

Attributes Information:

- Invoice No: 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.

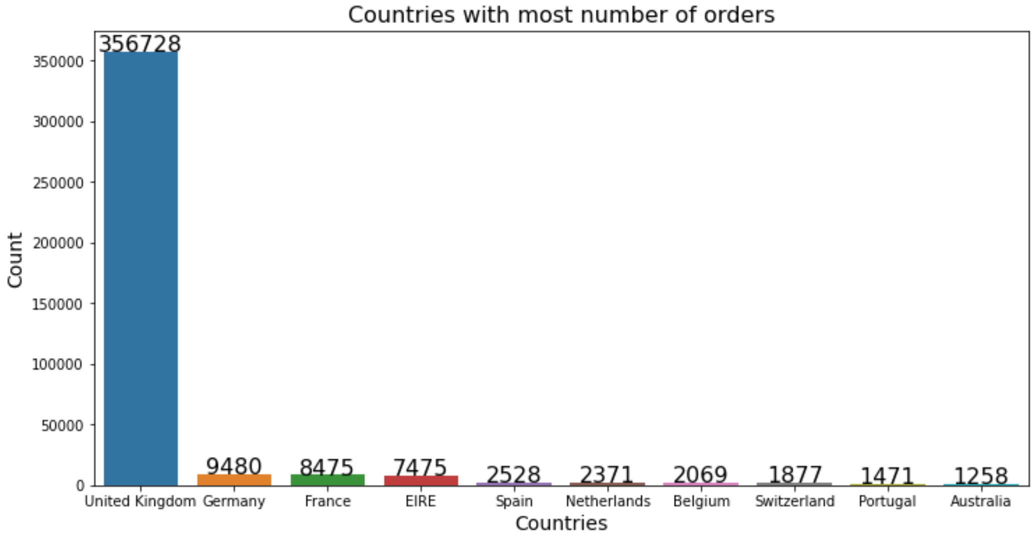
VI. EXPLORATORY DATA ANALYSIS

This dataset contains information on 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.

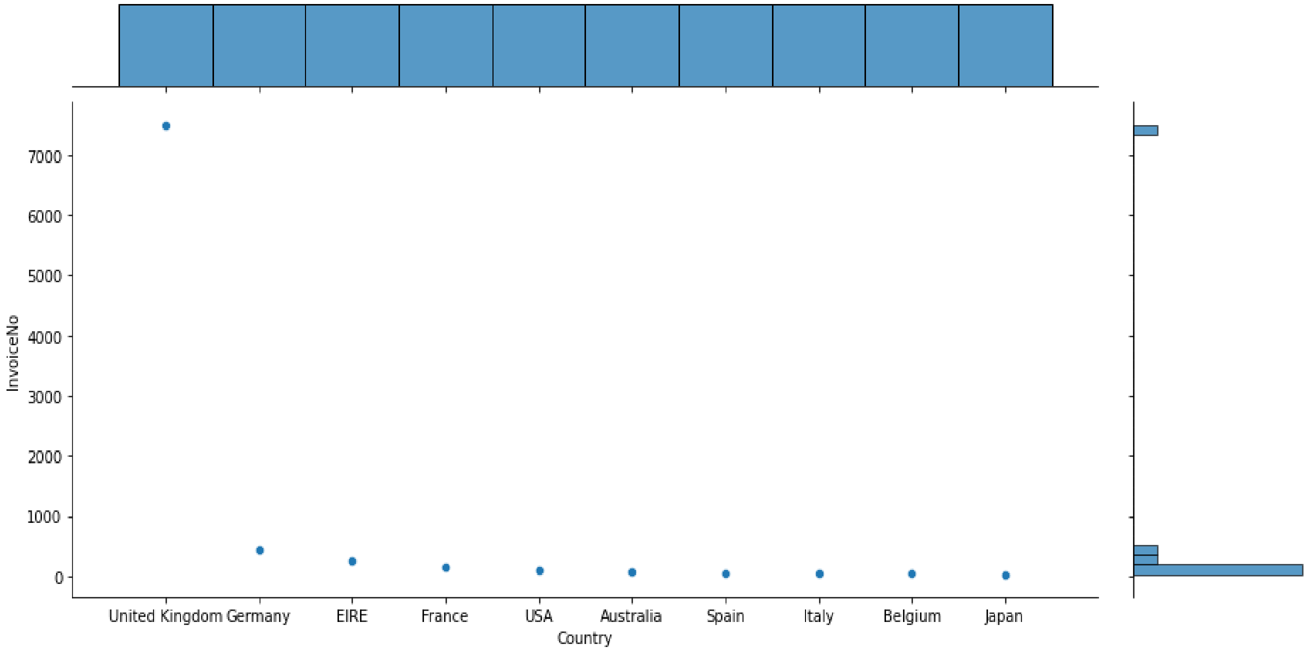
Overall, the dataset is very clean, but there are several null values and also some duplicate data. As a result, we had to employ some data cleaning processes.

The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable.

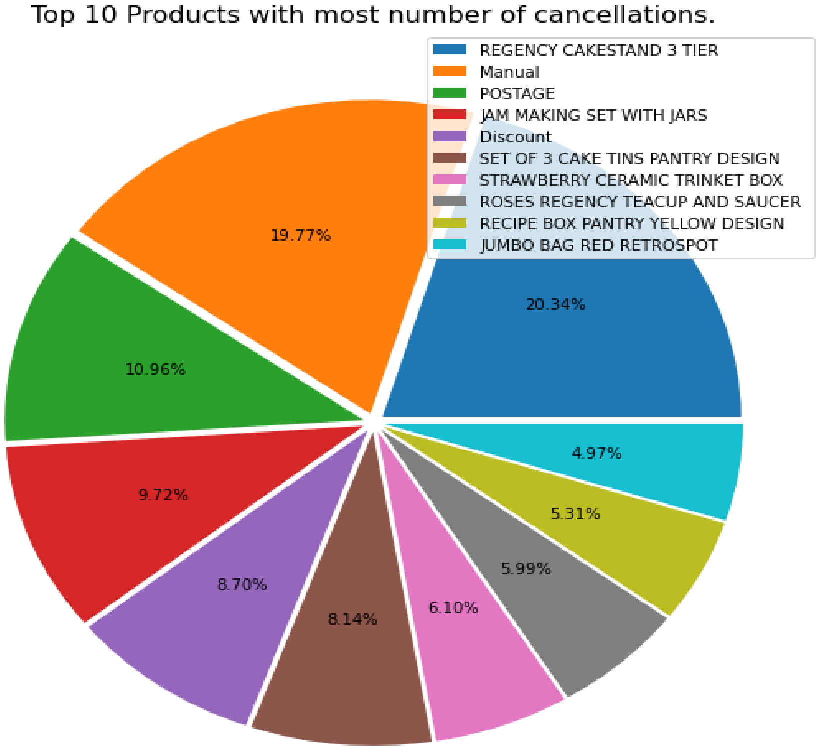
FEATURE RELATIONSHIP ANALYSES:-



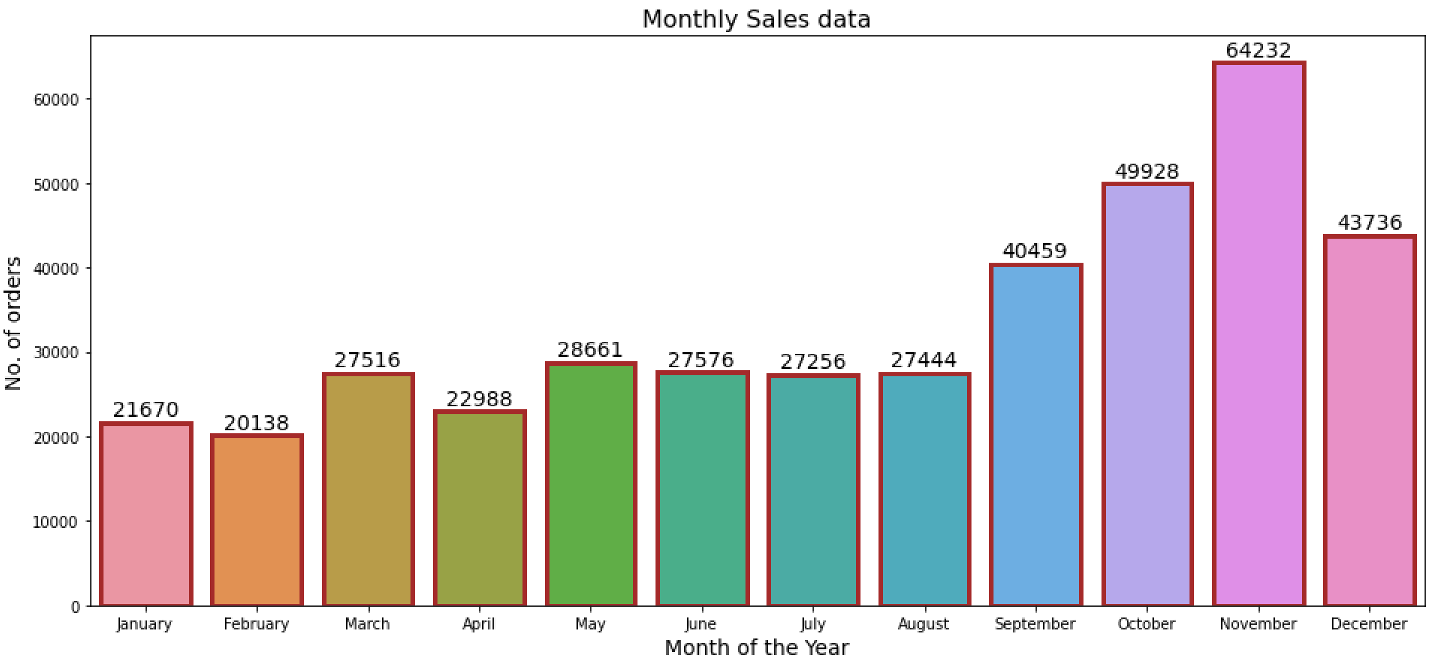
Here we have plotted the Countries whose residents have placed the most number of orders. From the graph above, we can clearly see that most of our orders were made by residents of the United Kingdom followed by Germany, France and EIRE (Ireland).



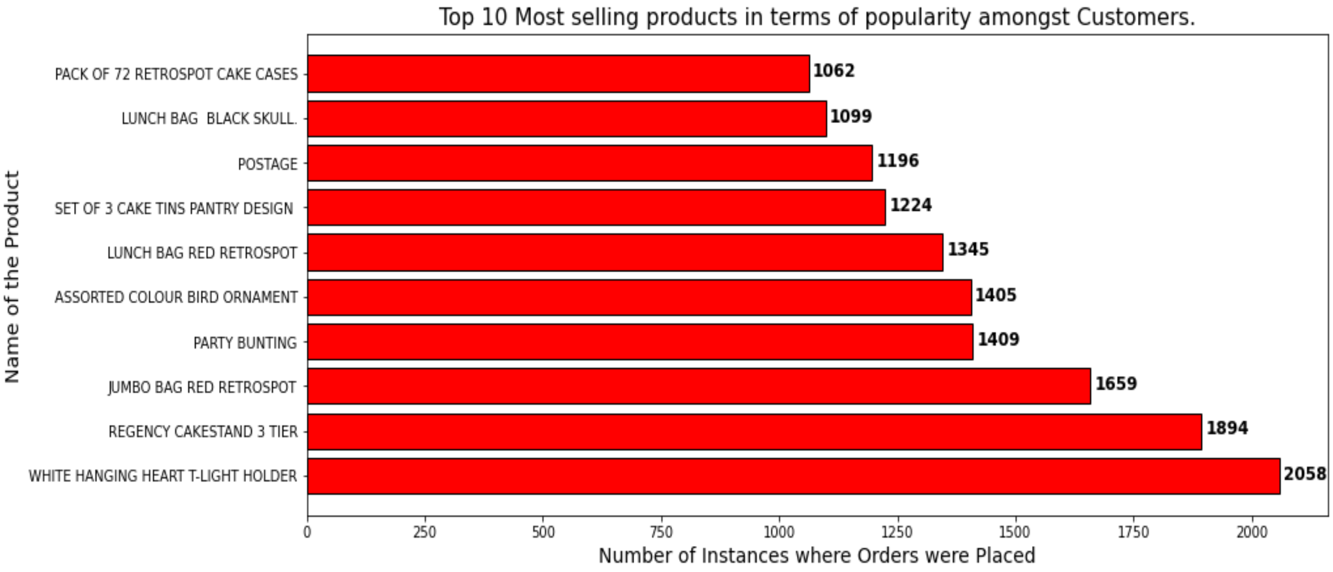
In the graph above, we have performed analysis of most canceled products versus the country of residence of customers. From the above graph, we can clearly see that most number of cancellations were made by the residents of United Kingdom.



Here we have plotted the top 10 most cancelled products. We can clearly see the products that were cancelled most with REGENCY CAKESTAND 3 TIER and Manual accounting for around 20.34 and 19.77 percent of cancellations out of the top 10.

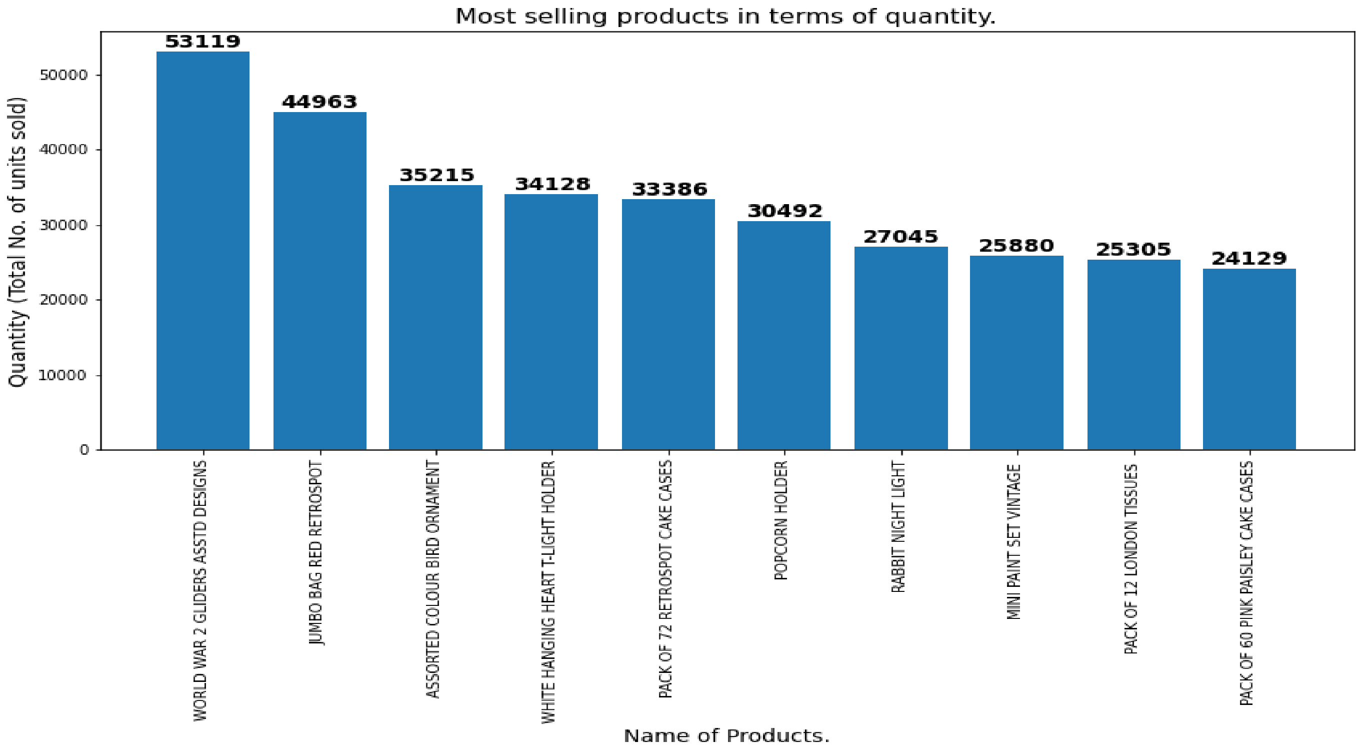


From the above graph depicting the total no. of orders placed every month, we can see that most orders are placed during the winter months. Which makes sense because the festive season in Europe occurs during the winter season.



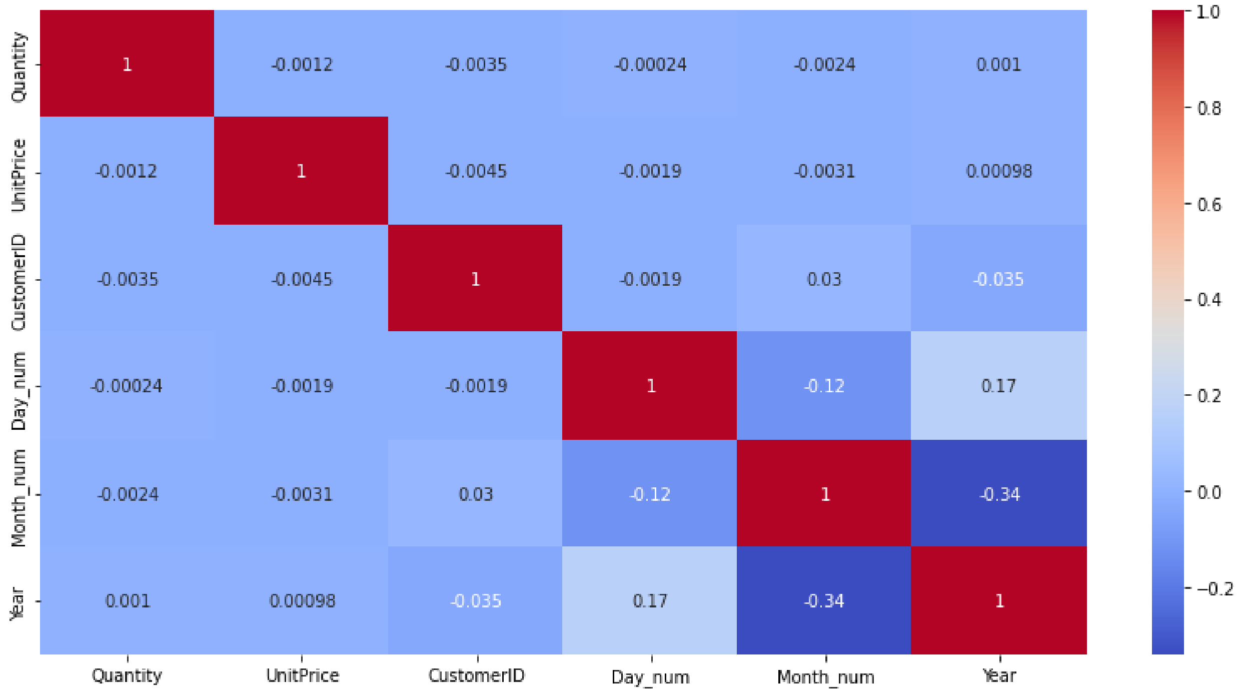
Above graph indicates the most selling products in terms of their popularity i.e. – how many customers have bought them. We can clearly see that most selling products are – WHITE HANGING HEART T-LIGHT HOLDER and REGENCY CAKESTAND 3 TIER with 2058 and 1894 instances of orders respectively.

The graph below, on the other hand 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.



CORRELATION ANALYSIS:-

Correlation heatmap is a type of plot that visualizes the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship. The values in the cells indicate the strength of the relationship, with positive values indicating a positive relationship and negative values indicating a negative relationship.



VII. MODEL IMPLEMENTATION

RFM ANALYSIS

Now, after EDA, we move on to using the RFM model and clustering algorithms such as K-means clustering and Hierarchical clustering. First we need to calculate recency, frequency and monetary scores and for that, we use our given dataset and do some data wrangling to find out how many days ago did every customer order, how many times did the customer order in a year's time and the total amount they spent on all their orders combined. These values are what we stored in recency, frequency and monetary features. Now using these RFM values, we can divide our customers into several categories – 4 in this case. Our best customer will be someone who is frequent, spends large sum on purchases and has bought with us recently.

After that, we plot the distribution of recency, frequency and monetary variables and we found that their distributions were positively skewed. As our models usually work best with normally distributed data so we applied log transformation to reduce skewness.

DATA PREPARATION

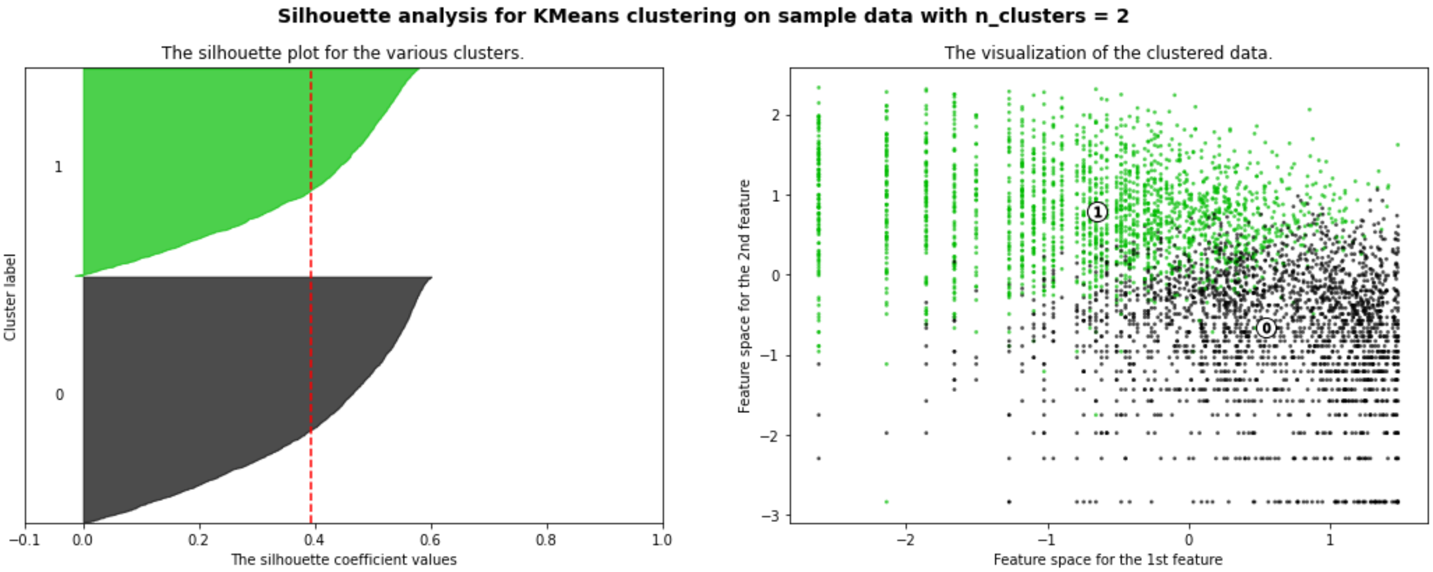
Data preparation includes feature engineering and feature selection. In feature engineering we came up with new features for our analysis – Recency, Frequency and Monetary in place of our old features to better suit our objective requirements.

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 from Sklearn library for scaling this data.

ALGORITHMS-

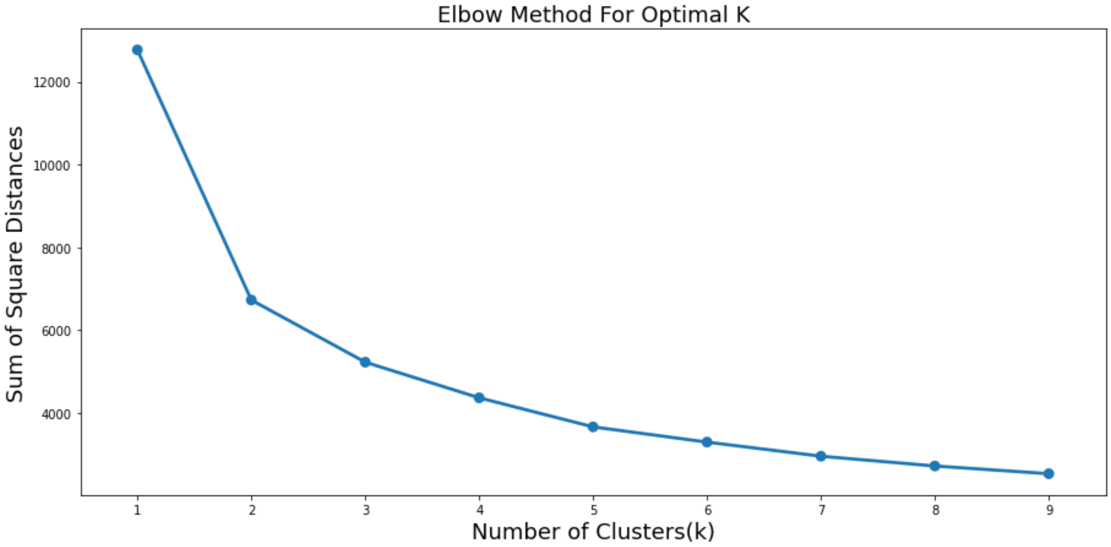
1. K MEANS CLUSTERING WITH SILHOUETTE SCORE:

K-means is a well-known clustering algorithm that is frequently used for unsupervised learning tasks. Silhouette analysis can be 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 when the number of clusters is 2 so optimal no. of clusters is 2.

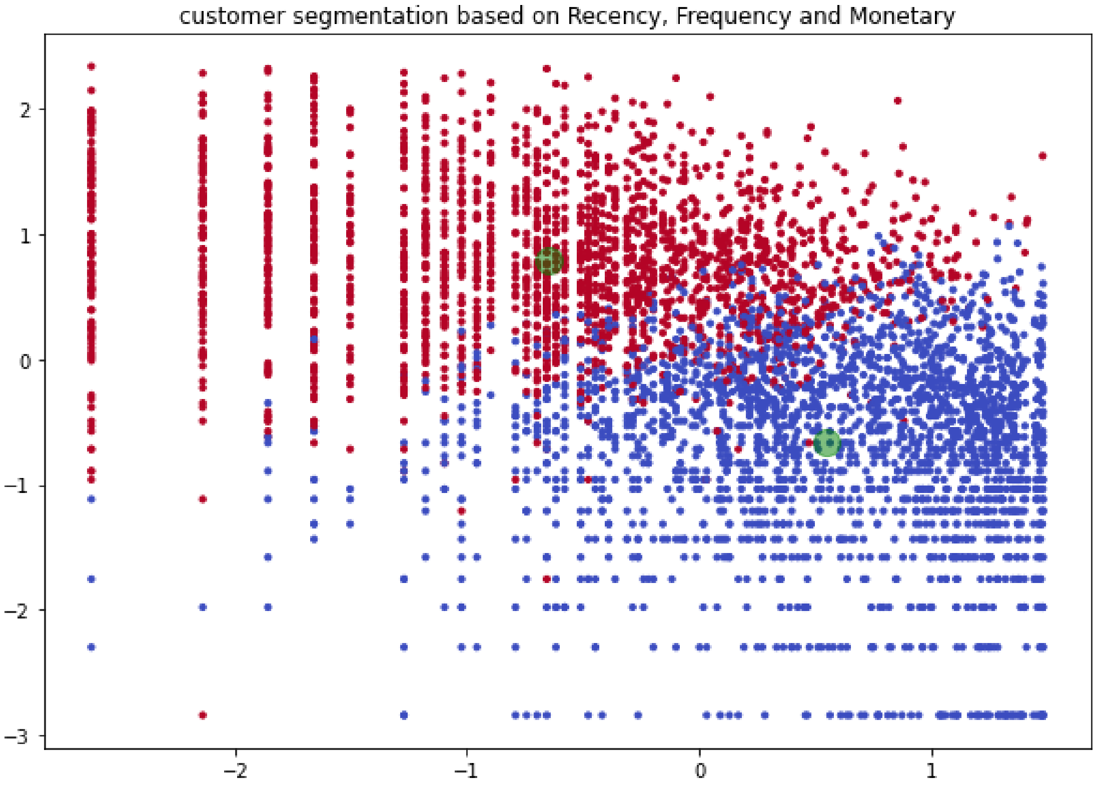


2. K MEANS CLUSTERING WITH ELBOW METHOD:

In this section, we built multiple clusters upon our normalized RFM data and tried to find out the optimal number of clusters in our data using the elbow method. 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.



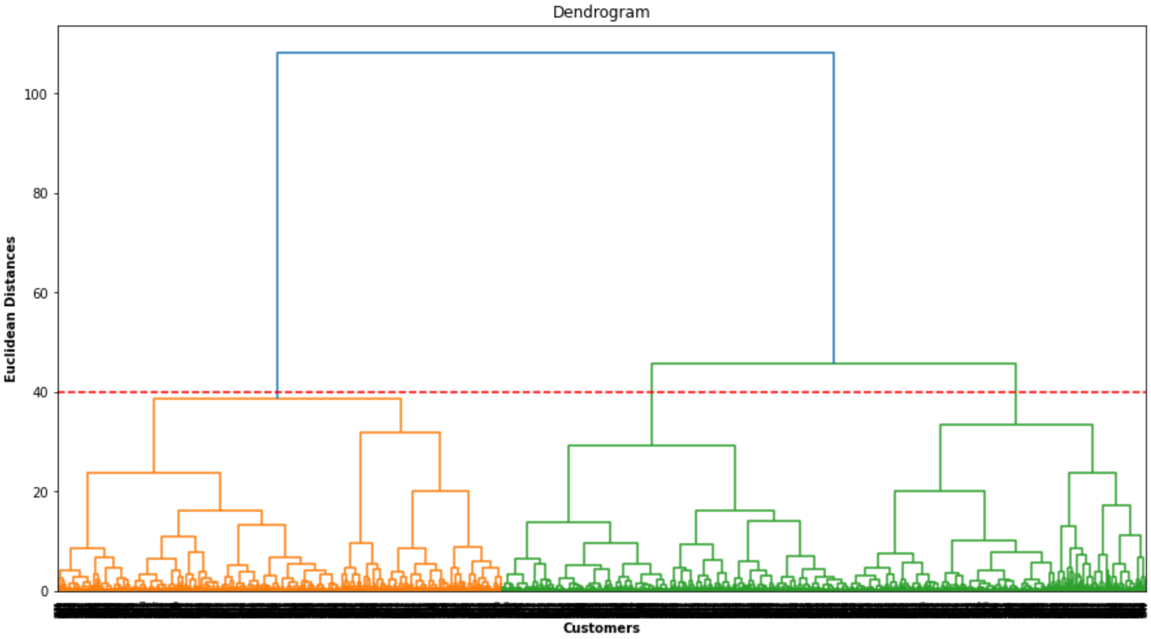
We can clearly see from the above graph that the optimal no. of clusters is 2 according to this analysis as the elbow is formed at that value.



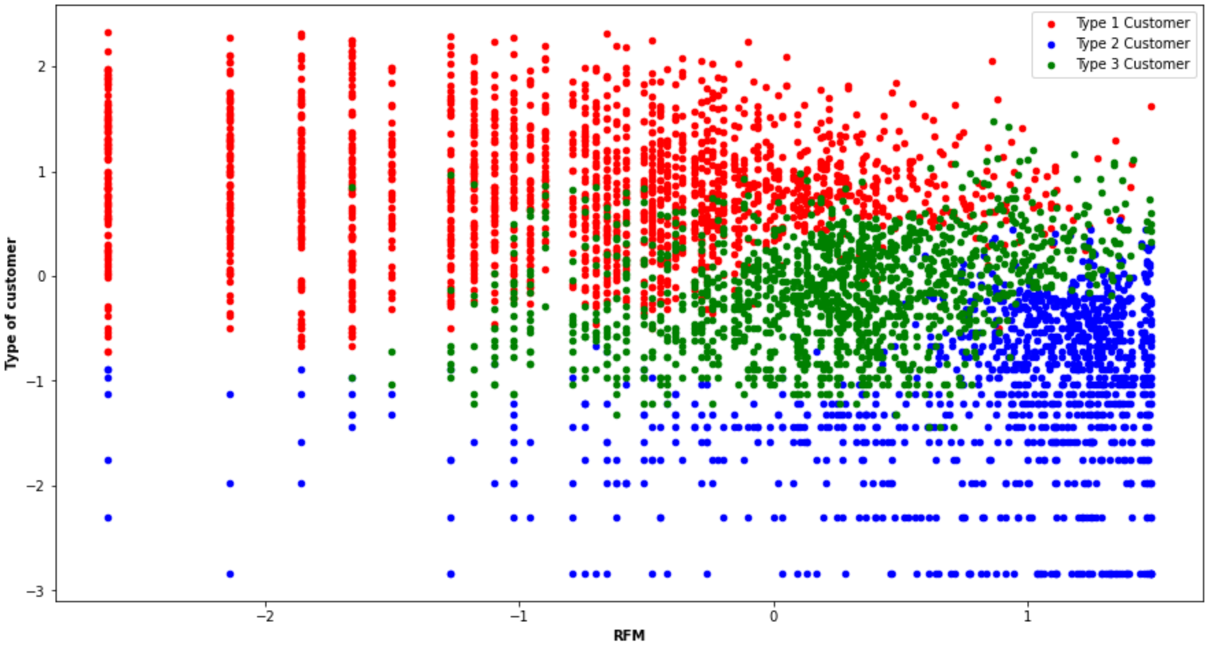
We can clearly see two separate clusters along with their cluster centers from the above graph.

3. AGGLOMERATIVE HIERARCHICAL CLUSTERING:

We used an Agglomerative Hierarchical Clustering algorithm but before that, we drew the dendrogram using the ward linkage to help us decide the number of clusters that we should use. We got the no. of optimal clusters as 3 from dendrogram and then we applied hierarchical clustering on our RFM data using clusters k=3.



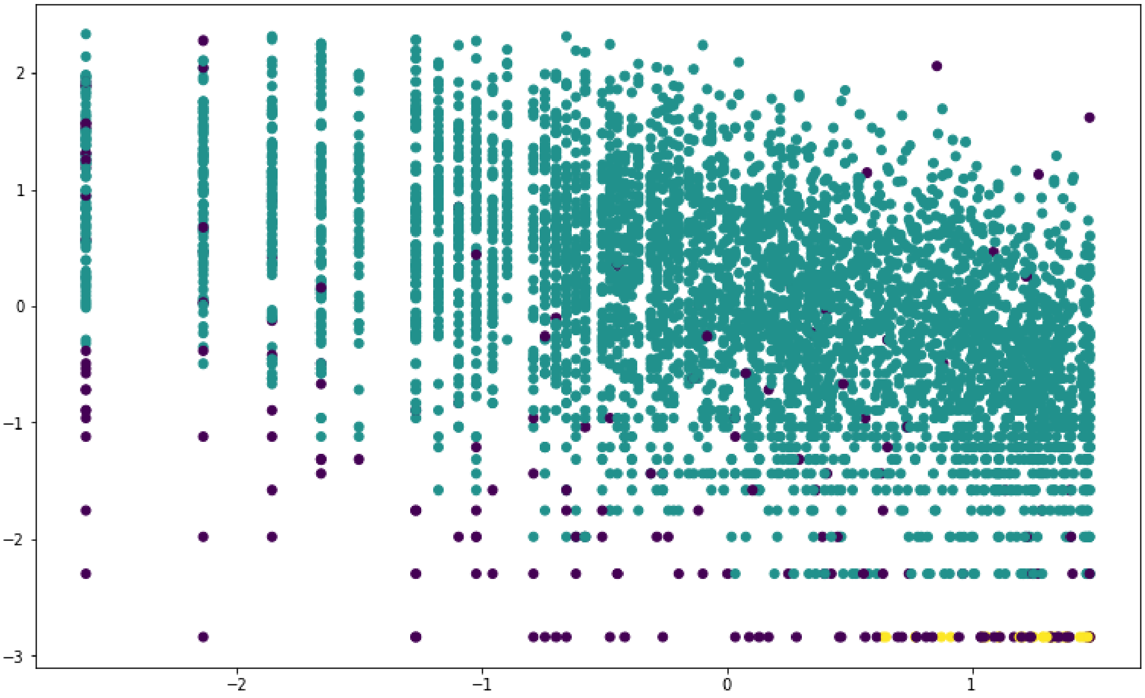
In the dendrogram, the optimal clusters is found by drawing a horizontal line such that it cuts the longest vertical line and the no. of vertical lines that this horizontal line cuts is the optimal no. of clusters. We can clearly see from the graph that it makes sense to draw the line at threshold value of 40 which gives us the no. of optimal clusters as 3.



We can clearly see three clusters forming in the above graph. Although there seems to be some overlap amongst some data points.

4. DBSCAN:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which may contain noise and outliers. DBSCAN algorithm identifies the dense region by grouping together data points that are close to each other based on distance measurement.



The optimal no. of clusters that we get from this model is 3.

We can clearly see three clusters in the above graph as can be identified by the colors – teal, purple and yellow although the yellow cluster is very small.

VII. CONCLUSIONS DRAWN

- We saw that most of the customers belong to the United Kingdom which makes sense as it is the country where our business is based out of.
- We saw that there are around 8872 instances where an order was canceled.
- We saw that most cancelled products according to our data are REGENCY CAKESTAND 3 TIER, Manual, POSTAGE, JAM MAKING SET WITH JARS etc.
- We also saw that most number of cancellations were made by residents of the United Kingdom which makes sense as UK residents have made the most orders as well.
- 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 that our most selling products in terms of how many separate instances of orders there were (how many customers bought them), are - WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, JUMBO BAG RED RETROSPOT, etc.
- WORLD WAR 2 GLIDERS ASSTD DESIGNS ,JUMBO BAG RED RETROSPOT, ASSORTED COLOUR BIRD ORNAMENT, WHITE HANGING HEART T-LIGHT HOLDER, etc. are our most sold products in terms of quantity (no. of units sold).
- 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.
- We have used Agglomerative Hierarchical Clustering, elbow method and silhouette score to find optimal no. of clusters.
- 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.