Capstone Project - 4

Online Retail Customer Segmentation

Presented By

Rupesh Sharma

Contents

Introduction

- With the rise in competitors within the same categories of the marketplace, there exists a fierce competition for new customers as well as retaining the existing ones. It requires exceptional customer service irrespective of the size of the company.
- The objective of my project is to segment the customers of an online retail store in order to develop business strategies for the store.

About the dataset

The online retail dataset includes transnational data from a registered online store of United Kingdom between December 1, 2010 and December 9, 2011. This dataset aids in customer segmentation, which in turn drives business strategies for the store.

It consists of 5,41,909 rows and 8 columns.

	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

Data Dictionary

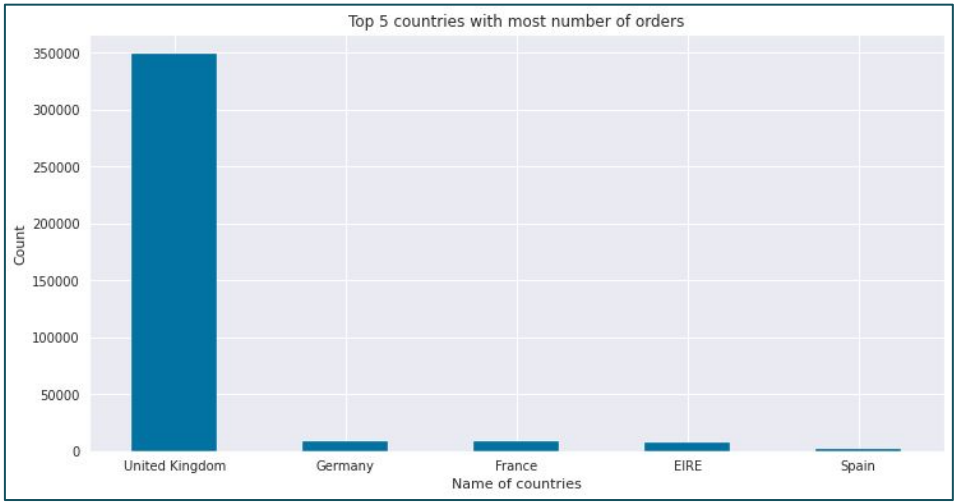
Column	Data Type	Description
InvoiceNo	Object	Invoice number of the product
StockCode	Object	Product code
Description	Object	Product name
Quantity	Integer	Quantities of each product per transaction
InvoiceDate	Datetime	Invoice date and time
UnitPrice	Float	Unit price of the product
CustomerID	Float	ID of the customer
Country	Object	Country of the customer

Data Cleaning

- The dataset contains duplicate rows, which rendered the dataset redundant, so I removed the rows with duplicate values.
- The dataset contains null values in CustomerID and Description so, the rows with null values in those columns were removed.
- There were also a few negative values in the Quantity and UnitPrice columns. As a result, I removed those rows, and further investigation of these columns revealed that negative Quantity reflects cancelled orders and negative UnitPrice could be free gifts.

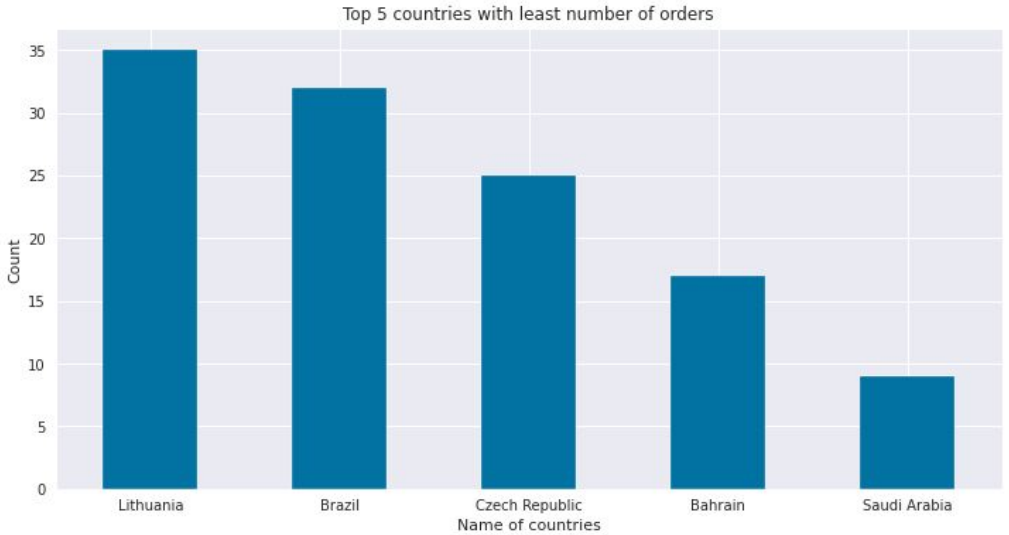
Data Analysis & Visualisation

- The countplot on the right shows the top five countries in terms of order volume.
- According to this visualisation, the majority of orders come from the United Kingdom, followed by Germany and France, indicating that online retail store is very popular in Europe.



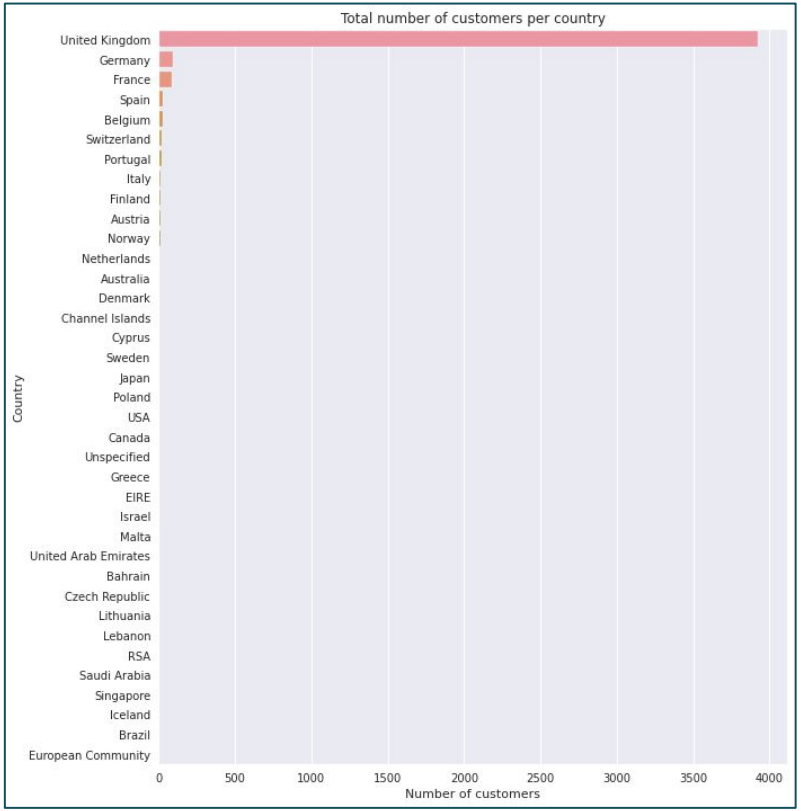
Data Analysis & Visualisation

- The countplot on the right shows the five countries with the lowest order volume.
- According to the visualisation, the least orders come from Saudi Arabia, followed by Bahrain, indicating that online retail store has very few Middle Eastern customers.



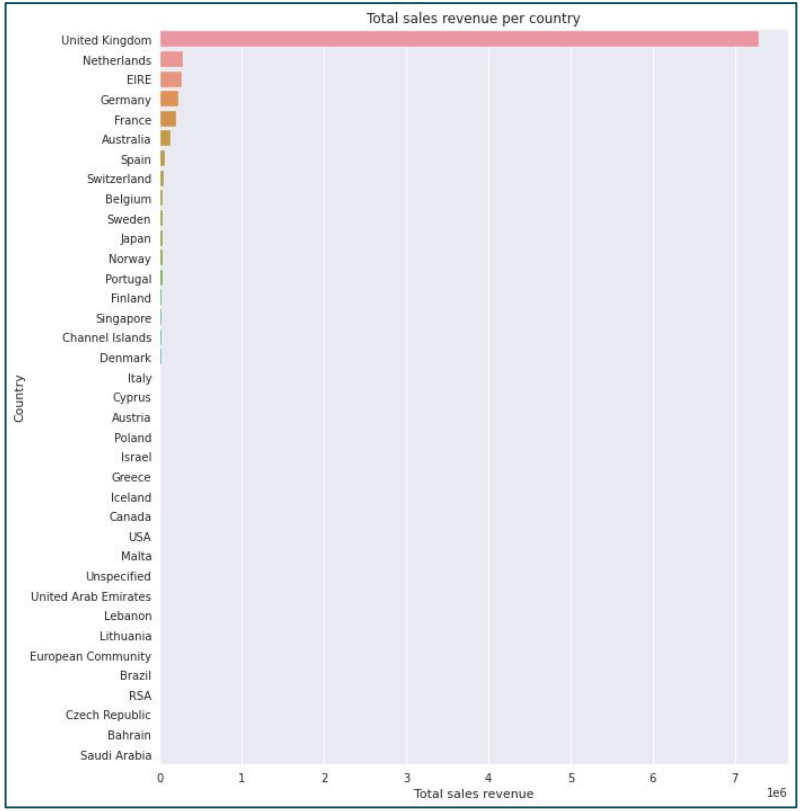
Data Analysis & Visualisation

- The right-hand barplot depicts the total number of customers by country.
- From the visualisation, it appears that the United Kingdom has the most customers.



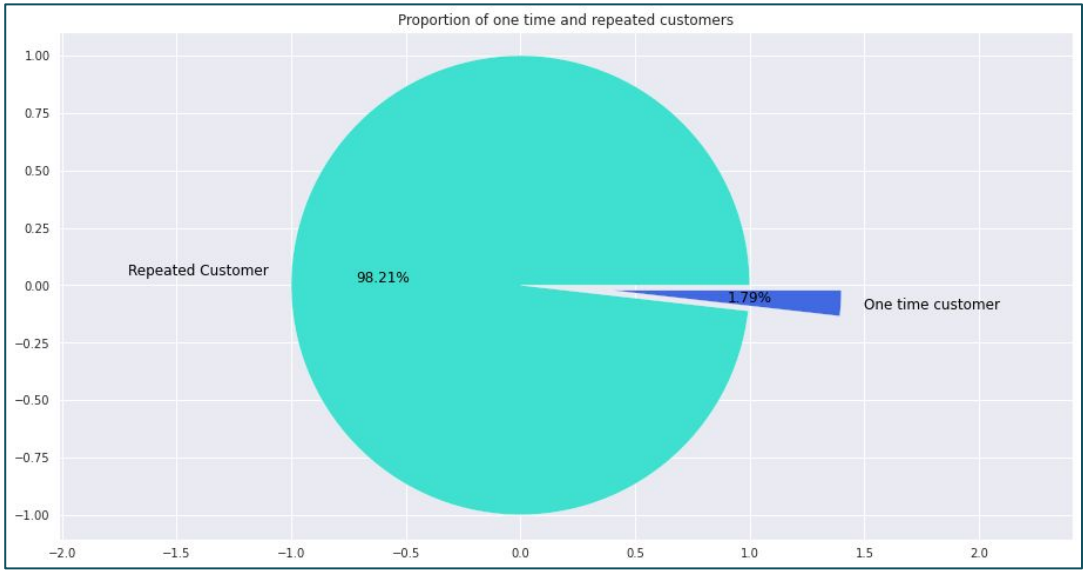
Data Analysis & Visualisation

- The right-hand barplot depicts total sales revenue by country.
- According to the graph, the United Kingdom has the highest sales revenue.



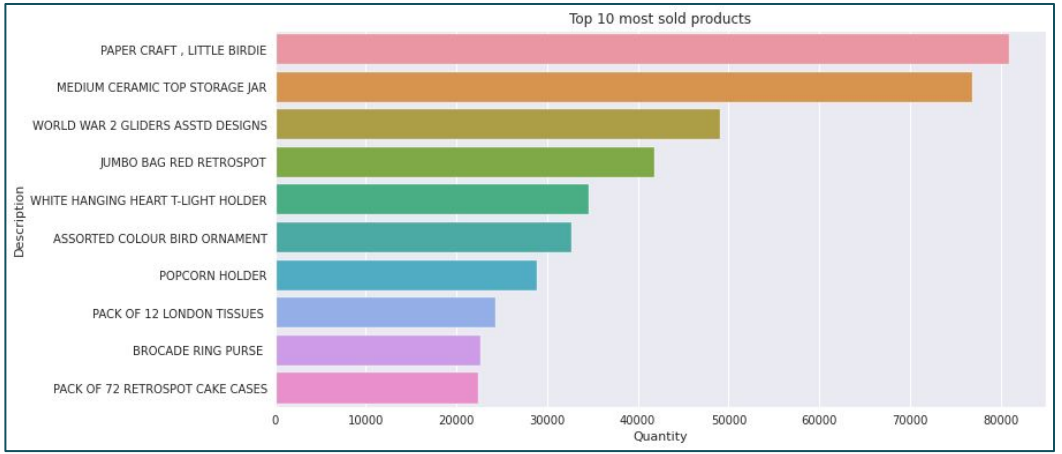
Data Analysis & Visualisation

- The right-hand pie chart depicts the proportion of one-time and repeat customers.
- The retail store appears to have 98 percent repeat customers. The majority of customers appear to be satisfied with the retail store.



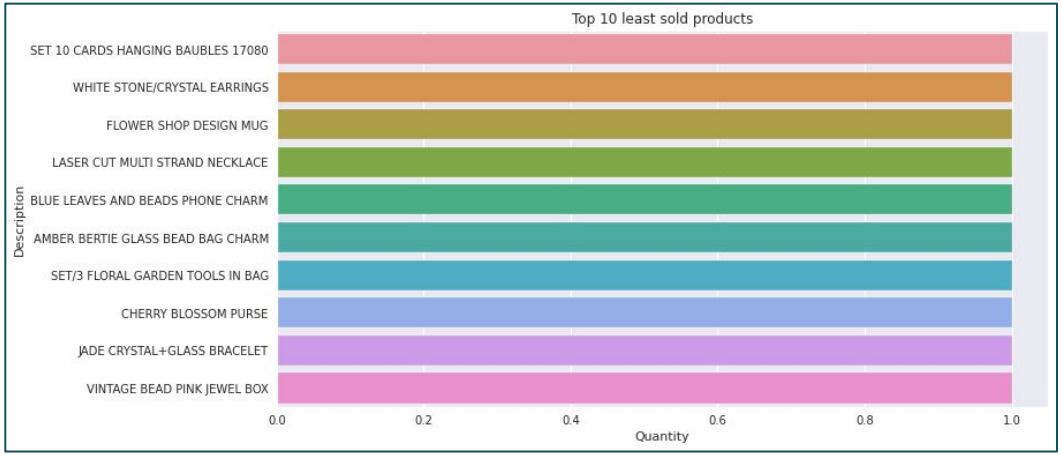
Data Analysis & Visualisation

- The visualisation on the right depicts the top ten most sold products.
- The graph shows that the store's best-selling items are paper craft and little birdie.



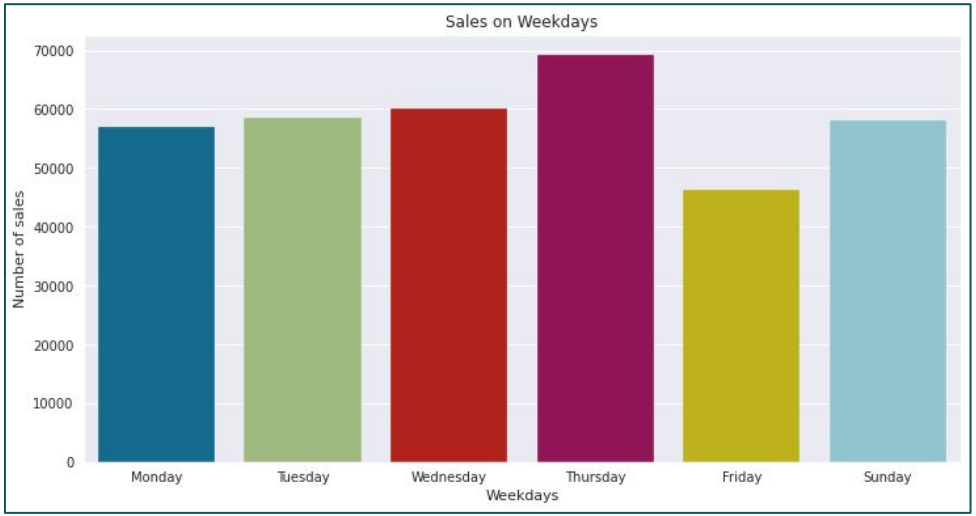
Data Analysis & Visualisation

- The visualisations on the right depict the top ten least products of the store.
- It appears that vintage bead pink jewel box is the least popular item.



Data Analysis & Visualisation

- The visualisation on the right depicts weekday sales.
- It appears that the majority of orders are placed on Thursday. It's also worth noting that no orders are placed on Saturdays, indicating that the online store is closed for maintenance or stock inspection on Saturdays.

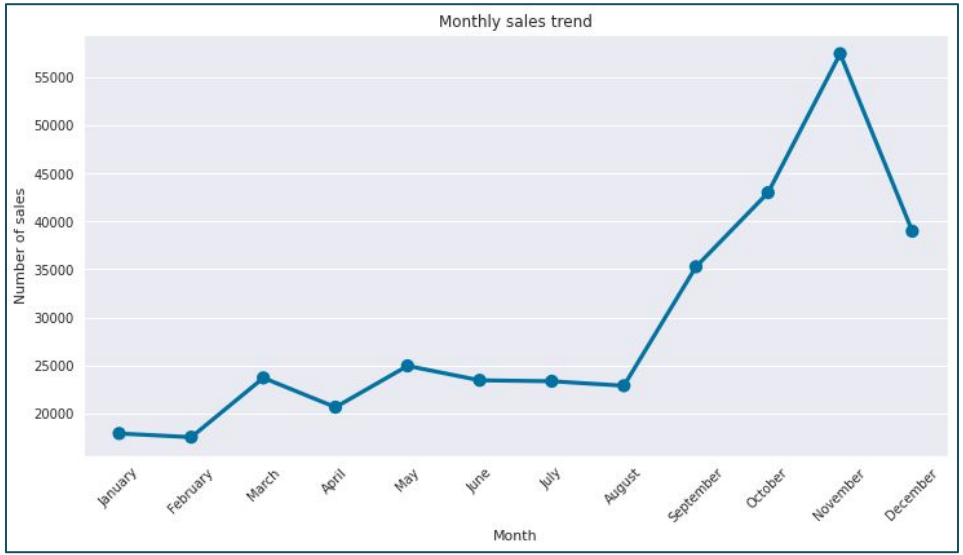


Data Analysis & Visualisation

- The line plot depicts sales at a specific time of day.
- It appears that the majority of sales occurred between 11 a.m. and 2 p.m.

Data Analysis & Visualisation

- The line plot depicts the monthly sales trend.
- The majority of sales appear to take place in the fourth quarter of the year, with November being the busiest shopping month.



RFM Segmentation

RFM segmentation (Recency, Frequency, Monetary) is a tried-and-true marketing model for behavior-based customer segmentation. It categorises customers based on their purchase history - how recently, how frequently, and how much they purchased. It helps in identifying customers who are more likely to respond to promotions as well as future personalisation services.

Following are the features used for RFM modelling:

1. Recency(R): The number of days since the last purchase. How recently did the customer make a purchase?
2. Frequency(F): The total number of times a transaction is performed. How many times has the customer bought something from the store?
3. Monetary(M): The total amount of money spent on purchases. How much money did the customer spend?

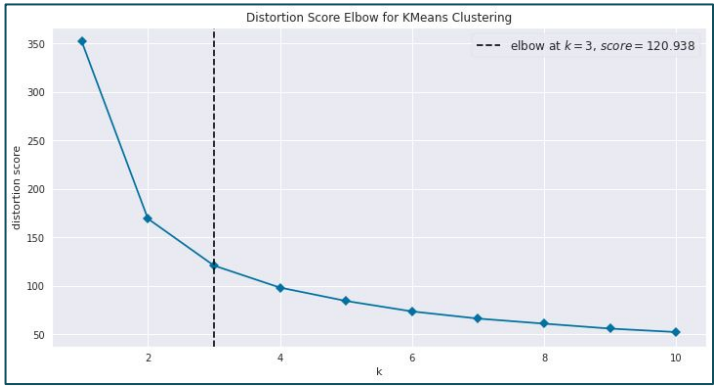
Feature Engineering

- Extracted new features from the InvoiceDate column for analysis.
- Created three new features Recency, Frequency and Monetary for RFM modelling.
- As RFM features are skewed in nature, cube root transformation on Recency and log transformation on Frequency and Monetary is applied.
- MinMax scaling is used to scale transformed RFM features.

Model Implementation

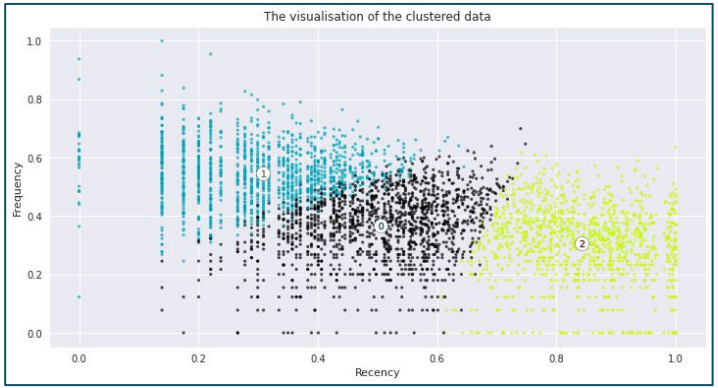
KMeans Clustering:

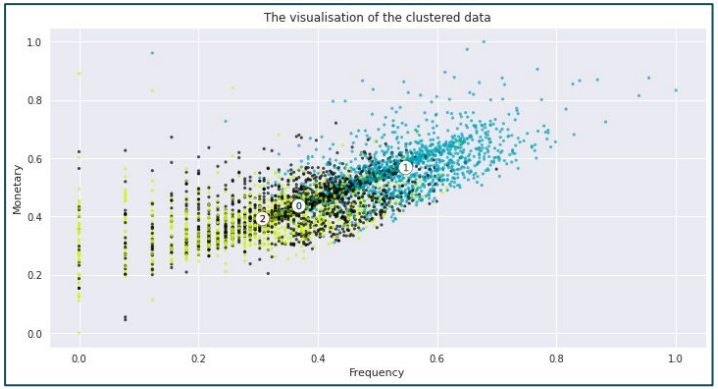
- Used elbow method to find the optimal number of clusters.
- The elbow of the curve is at K = 3. As a result, I have used Kmeans to group customers into three groups.

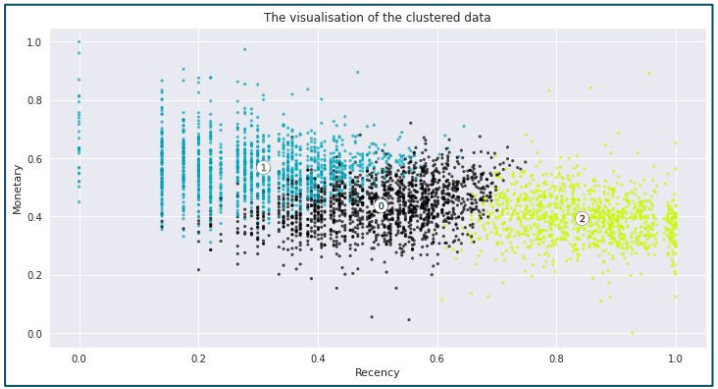


Model Implementation

KMeans Clustering:



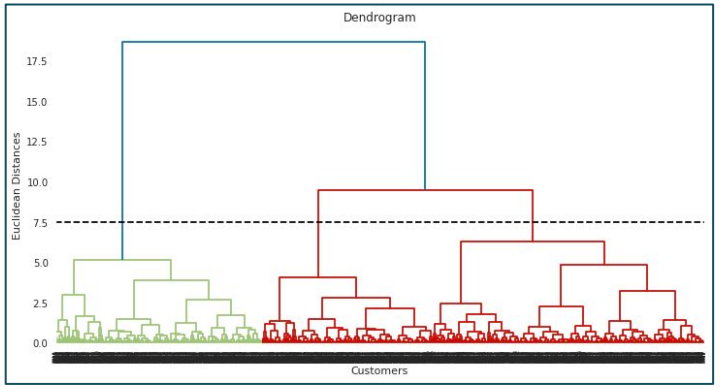




Model Implementation

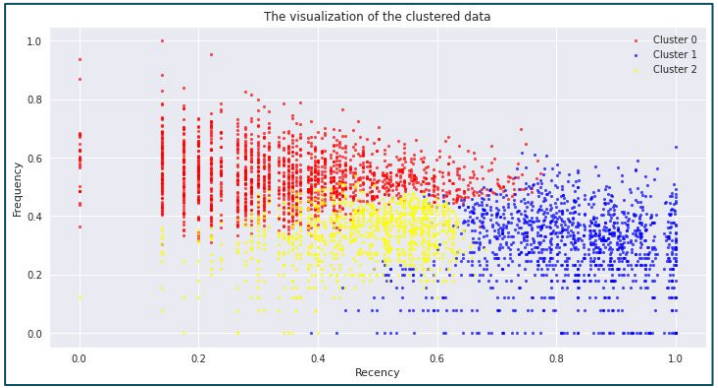
Hierarchical Clustering:

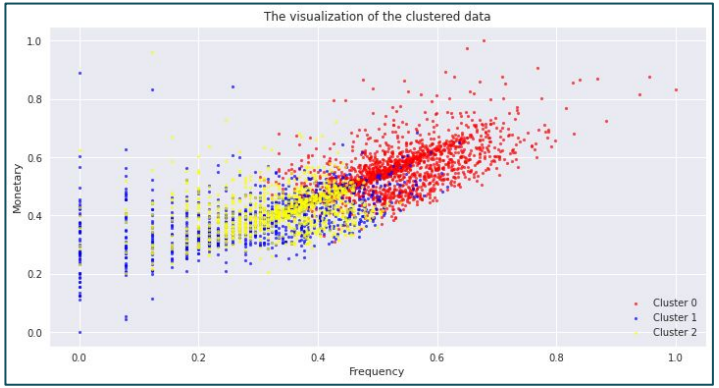
- Used dendrograms to find the optimal number of clusters.
- The Dendrogram indicates that the number of clusters should be three. As a result, I have partitioned the customers into three groups.

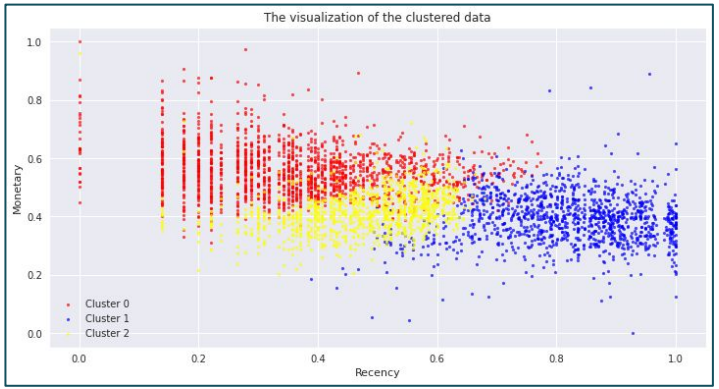


Model Implementation

Hierarchical Clustering:

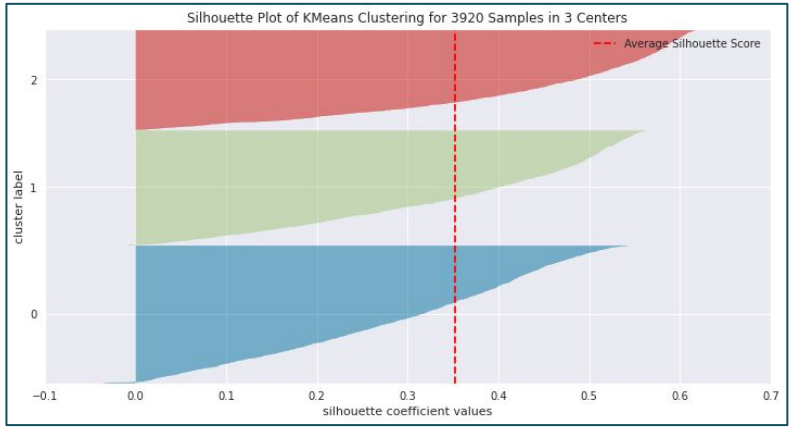






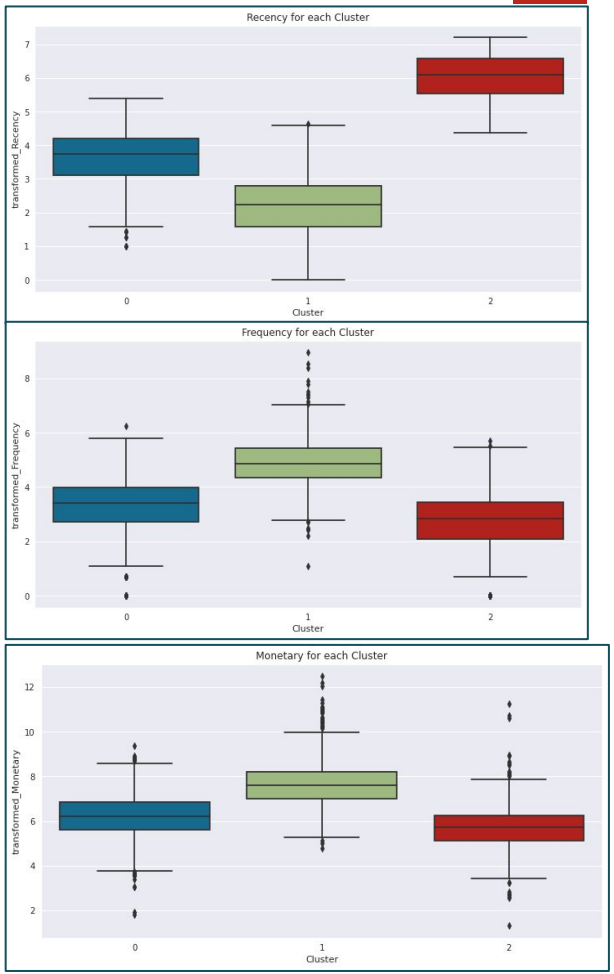
Comparing the clusters

- Silhouette score is used to compare the clusters of two clustering algorithms.
- The silhouette score of KMeans and Hierarchical clustering is 0.35 and 0.30 respectively.
- As the silhouette of Kmeans is higher, I have segmented the customers using Kmeans.



Analysing the clusters

- Tier 1 customers, who buy frequently from the store, spend a significant amount of money, and have recently purchased from the store, make up Cluster 1.
- Tier 2 customers make up Cluster 0, which are moderate-level customers who buy from the store on a regular basis, spend a reasonable amount of money, but haven't bought from the store in a while.
- Cluster 2 consists of Tier 3 customers who shop infrequently, spend less money, and haven't purchased from the store in a long time.



Challenges

- Handling the large dataset.
- Dealing with negative quantity values.
- Analysing the clusters and drawing conclusions.

Conclusion

- Customers in Tier 1 (i.e. Cluster 1) are the most valuable to the company. This group should feel valued and appreciated when communicating with them. These customers are likely to account for a disproportionately large portion of overall revenue, so keeping them happy should be a top priority. More in-depth analysis of their individual preferences and affinities will open up even more possibilities for more personalised marketing.
- Customers in Tier 2 (i.e. Cluster 0) can be considered active and loyal as they purchase products on a regular basis. This group should be targeted with campaigns that make them feel appreciated and encourage them to spend more. They should be rewarded with special offers in order for them to spread the word about the brand to their friends, through social media.
- Customers in Tier 3 (i.e. Cluster 2) have a high recency rate, indicating that they are about to be churned out. While re-engaging churned customers can be difficult, high-spending customers in this group are worth pursuing. It's critical to communicate with them based on their specific preferences, which can be determined by their previous transaction data.

Thank You