Capstone Project - 4

Online Retail Customer Segmentation

Team Members

Huzaifa Khan

Arbaaz Malik

Table of contents:-

- Introduction and Problem statement
- Data Description
- Data Cleaning
- Exploratory Data Analysis
- Data Transformation
- Model Building (Clustering)
- Clustering Profiling
- Conclusion

Introduction and Problem statement

Introduction

Businesses all over the world are growing every day. With the help of technology, they have access to a wider market and hence, alarge customer base. Customer segmentation refers to categorizing customers into different groups with similar characteristics. Customer segmentation can help businesses focus on each customer group in a different way, in order to maximize benefits for customers as well as the business. This project mainly deals in segmenting customers of an online business store in the UK.

Problem Statement

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

Understanding attributes of dataset better:-

The data being used here is a transnational data of an online store based in the UK, which mainly sells unique all-occasion gifts.

The data has 5,41,908 rows and 8 columns.

1. **InvoiceNo. :** Invoice number. Nominal, a6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

2. **StockCode:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

3. **Description:** Product (item) name. Nominal.
4. **Quantity:** The quantities of each product (item) per transaction. Numeric.

Continued...

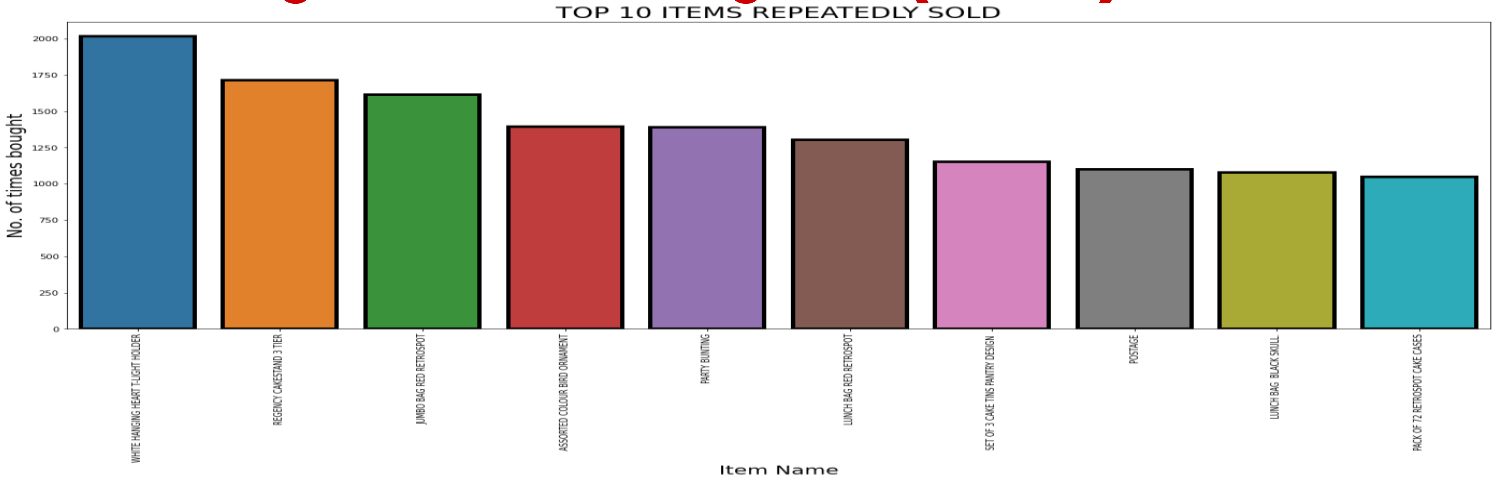
5. **InvoiceDate:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. **UnitPrice:** Unit price. Numeric, Product price per unit in sterling.
7. **CustomerID:** Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. **Country:** Country name. Nominal, the name of the country where each customer resides.

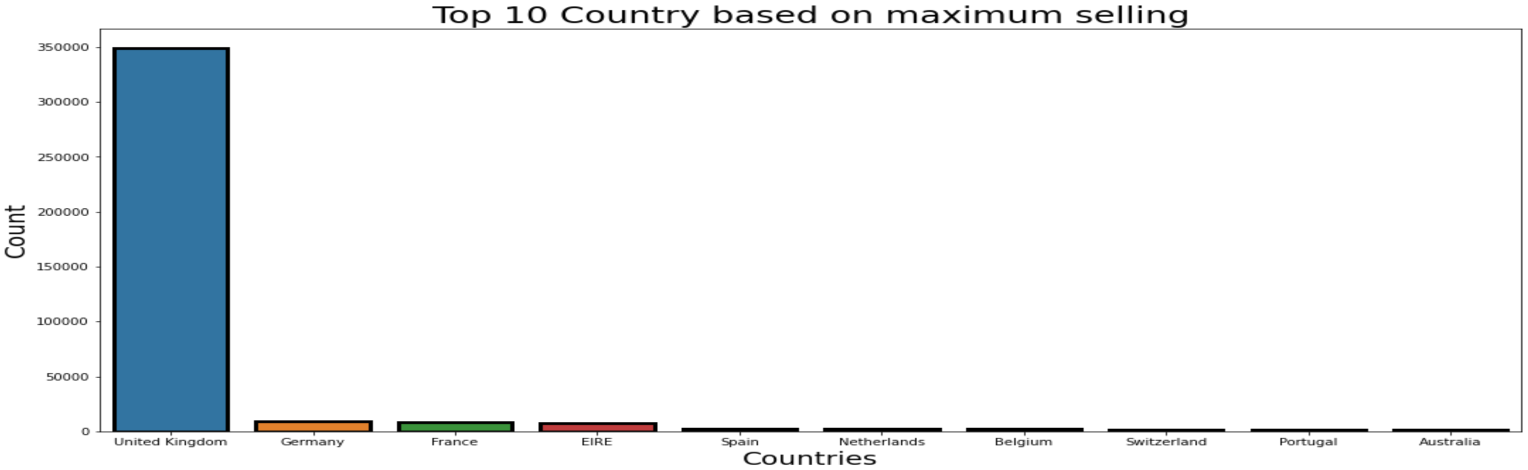
Each row represents a different item purchased by the customer.

Data Cleaning

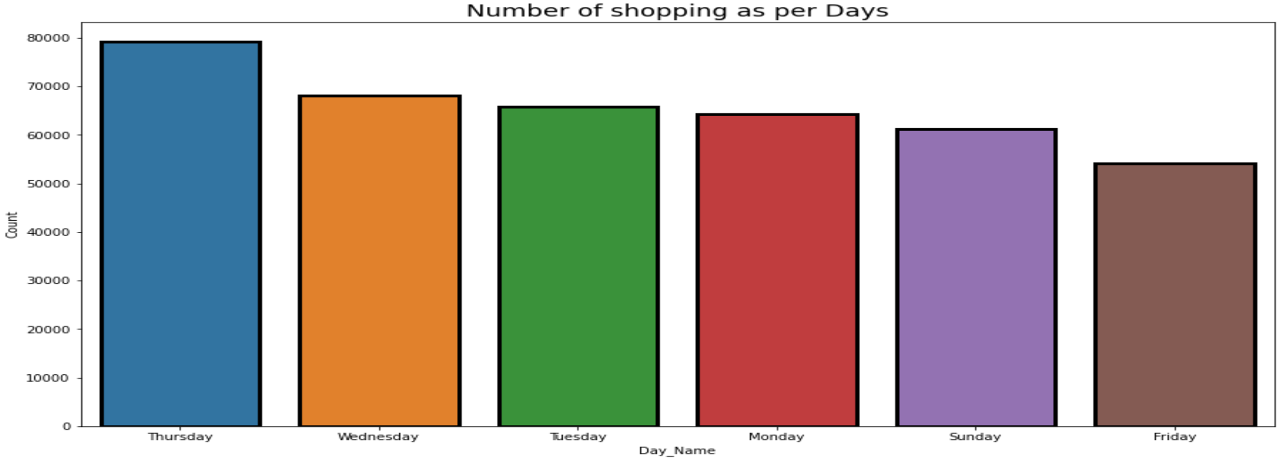
- After importing the data, the data must be cleaned.
- In this case, there are null values present in the ‘CustomerID’ and ‘Description’ column. These have to be dropped as there is no way of filling them strategically.
- Cancelled orders exist in the data, these too have been removed.
- Date, month and year were extracted from the ‘InvoiceDate’ column.
- Outliers in the ‘Quantity’ and ‘UnitPrice’ columns have been removed.
- ‘StockCode’ and ‘InvoiceDate’ column have been removed.

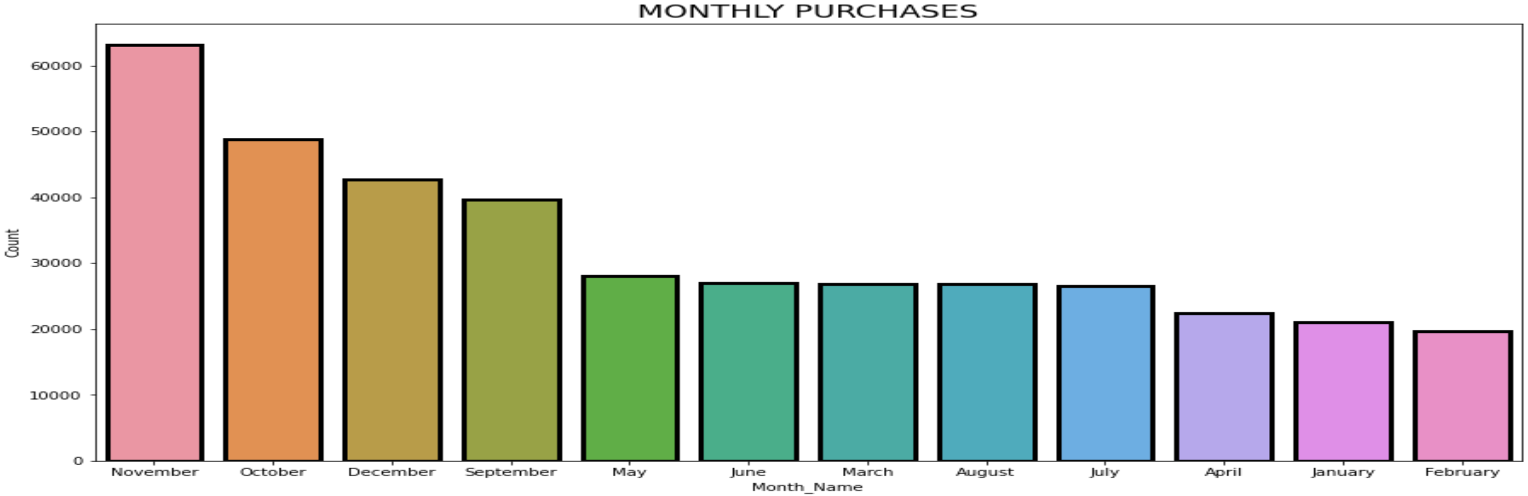
Exploratory Data Analysis (EDA)



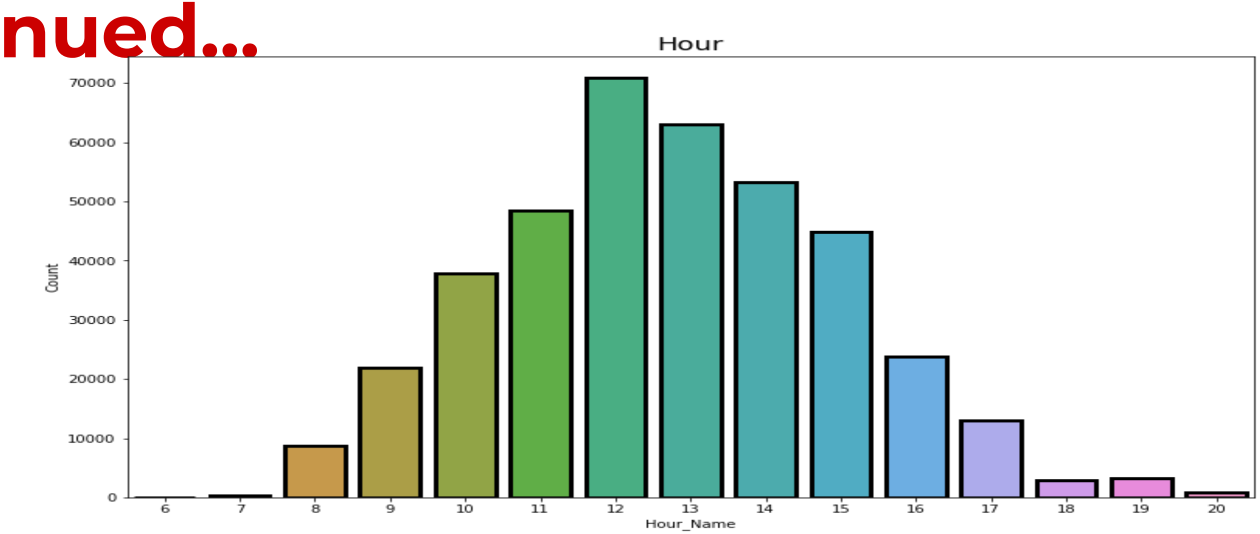


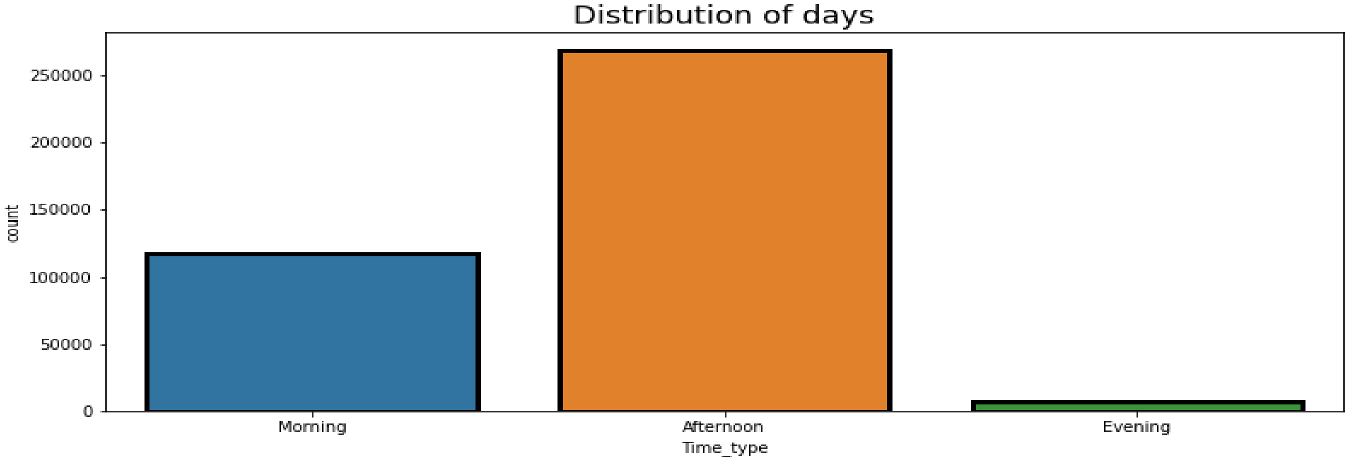
Continued...


Number of shopping as per Days
Count
0
10000
20000
30000
40000
50000
60000
70000
80000
Thursday
Wednesday
Tuesday
Monday
Sunday
Friday
Day_Name


MONTHLY PURCHASES
Count
0
10000
20000
30000
40000
50000
60000
November
October
December
September
May
June
March
August
July
April
January
February
Month_Name

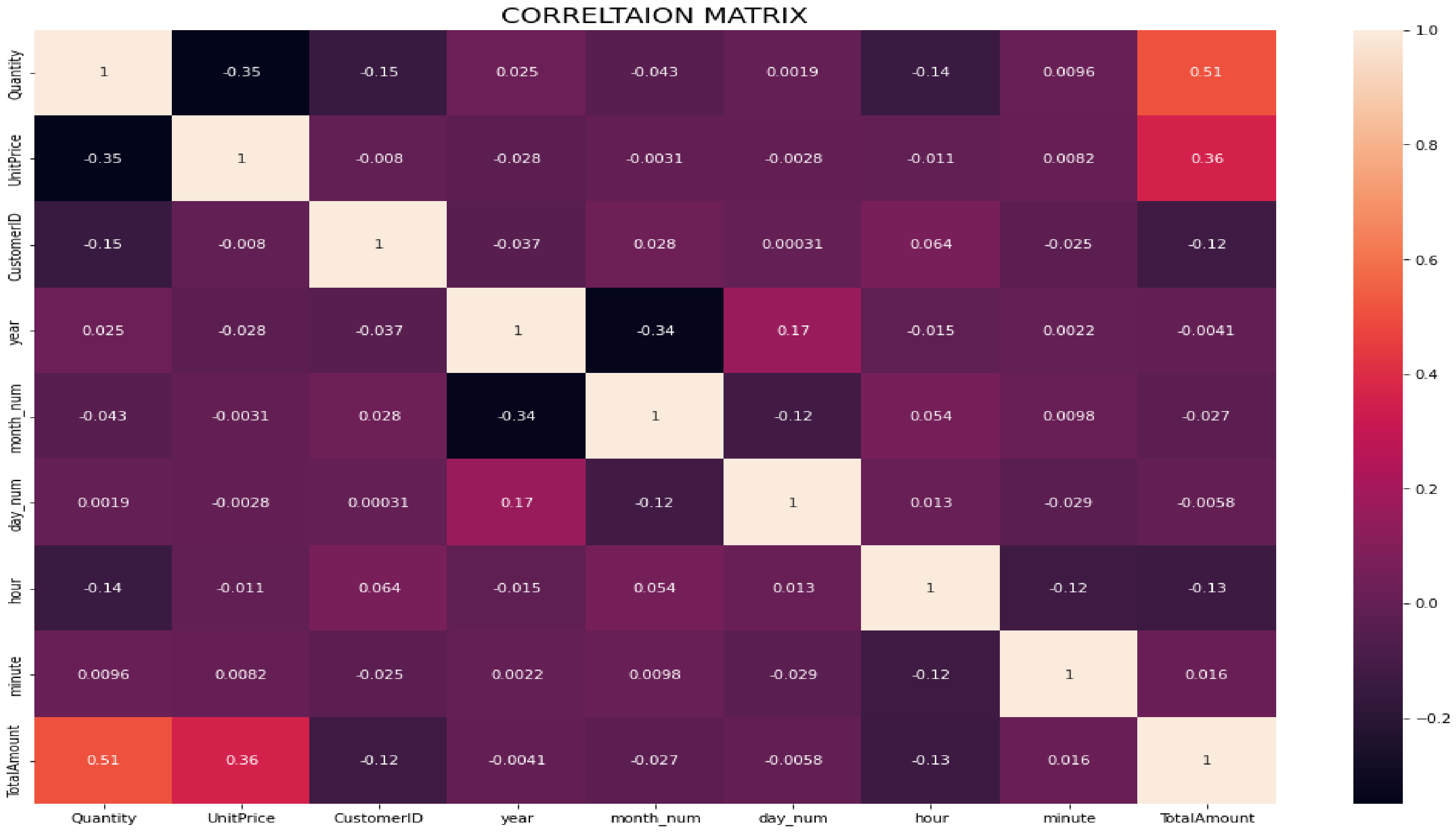
Continued...


Hour
Count
0
10000
20000
30000
40000
50000
60000
70000
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Hour_Name


Distribution of days
count
0
50000
100000
150000
200000
250000
Morning
Afternoon
Evening
Time_type

Continued...

CORRELTAION MATRIX



	Quantity	UnitPrice	CustomerID	year	month_num	day_num	hour	minute	TotalAmount
Quantity	1	-0.35	-0.15	0.025	-0.043	0.0019	-0.14	0.0096	0.51
UnitPrice	-0.35	1	-0.008	-0.028	-0.0031	-0.0028	-0.011	0.0082	0.36
CustomerID	-0.15	-0.008	1	-0.037	0.028	0.00031	0.064	-0.025	-0.12
year	0.025	-0.028	-0.037	1	-0.34	0.17	-0.015	0.0022	-0.0041
month_num	-0.043	-0.0031	0.028	-0.34	1	-0.12	0.054	0.0098	-0.027
day_num	0.0019	-0.0028	0.00031	0.17	-0.12	1	0.013	-0.029	-0.0058
hour	-0.14	-0.011	0.064	-0.015	0.054	0.013	1	-0.12	-0.13
minute	0.0096	0.0082	-0.025	0.0022	0.0098	-0.029	-0.12	1	0.016
TotalAmount	0.51	0.36	-0.12	-0.0041	-0.027	-0.0058	-0.13	0.016	1

Insights from EDA

- Quantities of products sold, mostly range from 1-12 units, also many purchases are in 24 units as well. This signifies units sold are in a dozen or two.
- Unit prices of products are mainly less than 3 pounds, there are some products with higher values as well.
- ‘White hanging heart T-light holder’ is most repeated order.
- The UK has the highest sales, this is logical as the company is UK based.
- Months of October, November and December repeat most frequently.
- Customer with the ‘ID 14911’ has purchased the most quantities.
- 'Pack of 72 retrospot cake cases' were sold the most in terms of quantity, around 15000 units.

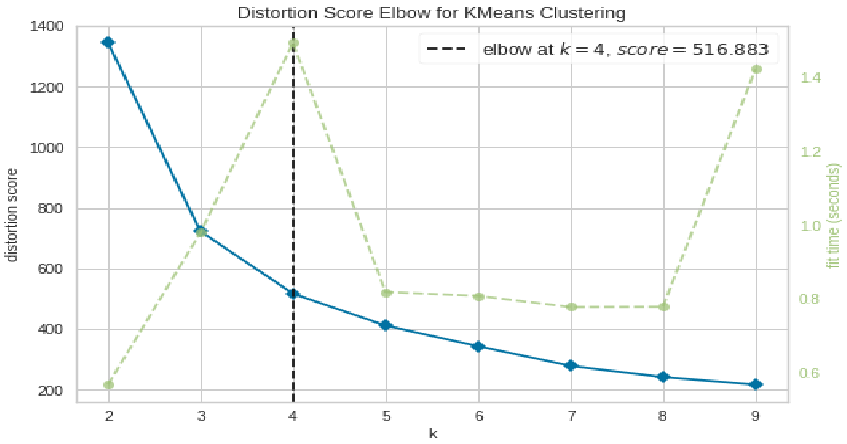
Data Transformation

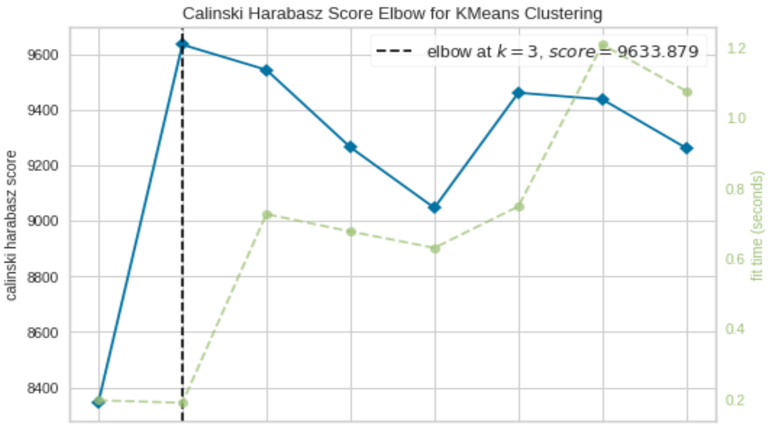
- In this section, a Recency, Frequency and Monetary (RFM) analysis about the data is done.
- Recency signifies the days since order, frequency signifies the number of times the customer is been billed and monetary signifies the sales each customer has provided.
- It can be seen that, frequency and monetary variables have a linear trend and frequency of orders have been high recently.
- The RFM dataframe is grouped on the basis of customer ID. The data now contains 4192 rows or customers.
- The ranges in the data differ, hence, the data is also scaled using a Standard Scaler.

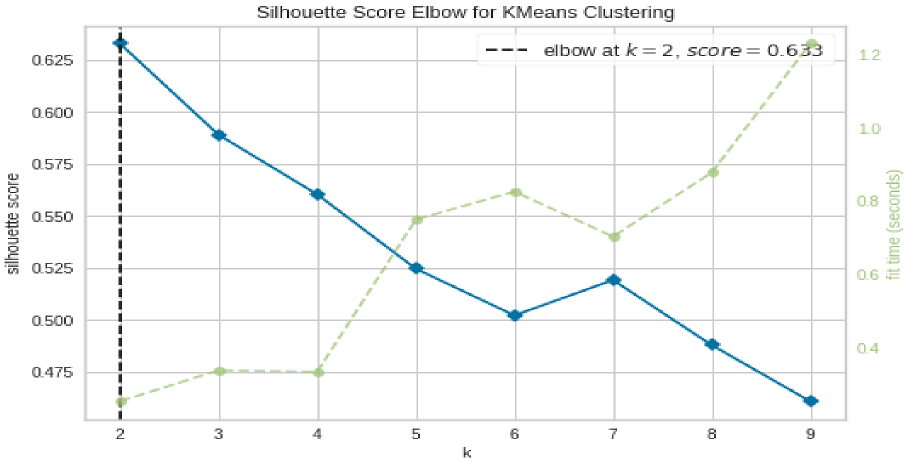
Model Building (Clustering)

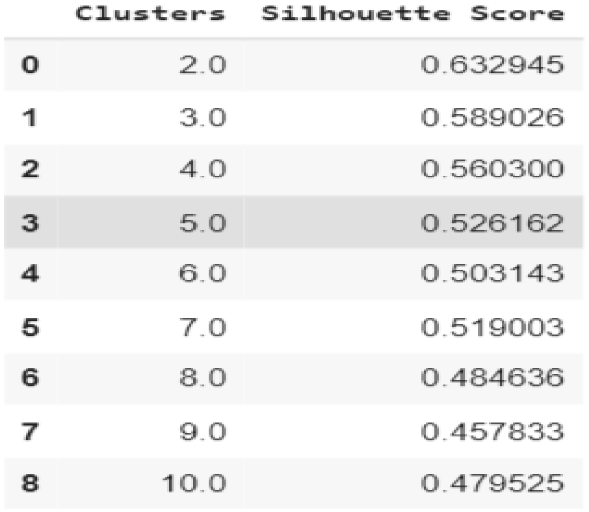
- In this section, we use K-Means algorithm to cluster the customers into different segments.
- To identify the optimum number of clusters, we use the elbow method and silhouette analysis.
- With both the methods, 3 clusters is optimum in this case.
- A K-Means model with 3 clusters is developed and customers are segmented into different clusters.

Continued (Elbow method and Silhouette Score



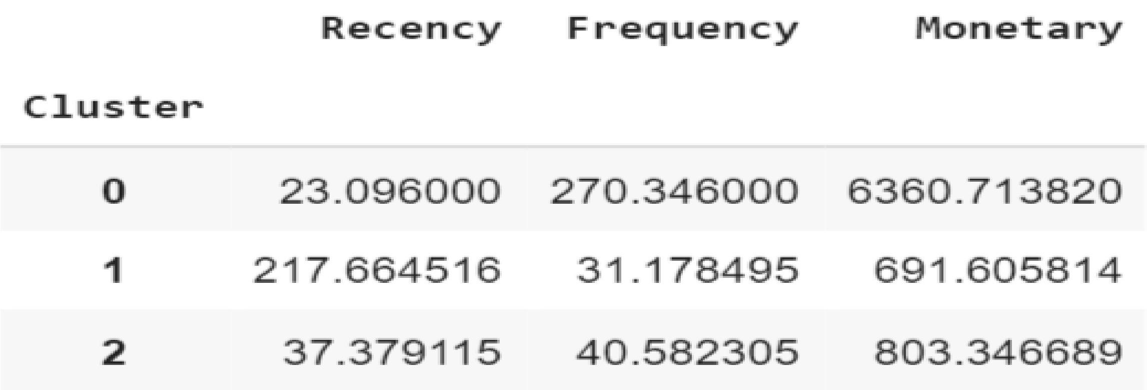






	Clusters	Silhouette Score
0	2.0	0.632945
1	3.0	0.589026
2	4.0	0.560300
3	5.0	0.526162
4	6.0	0.503143
5	7.0	0.519003
6	8.0	0.484636
7	9.0	0.457833
8	10.0	0.479525

Cluster Profiling



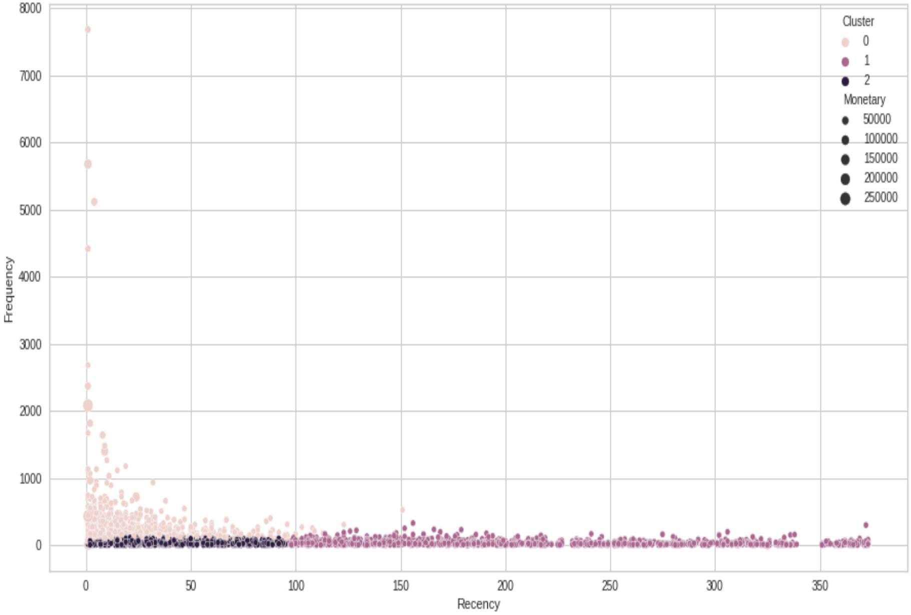
	Recency	Frequency	Monetary
Cluster			
0	23.096000	270.346000	6360.713820
1	217.664516	31.178495	691.605814
2	37.379115	40.582305	803.346689

➢ After grouping the clusters, using the mean, each cluster can be named as follows :-

1. cluster 0 :- comprises of customers who are very recent, frequent and also contribute largely to the sales.
2. Cluster 1 :- comprises of customers who are moderately recent, frequent and contribute an average amount to sales
3. Cluster 2 :- comprises of customers who made purchaces a long time ago and purchase infrequently and contribute the least towards the sales of the company.

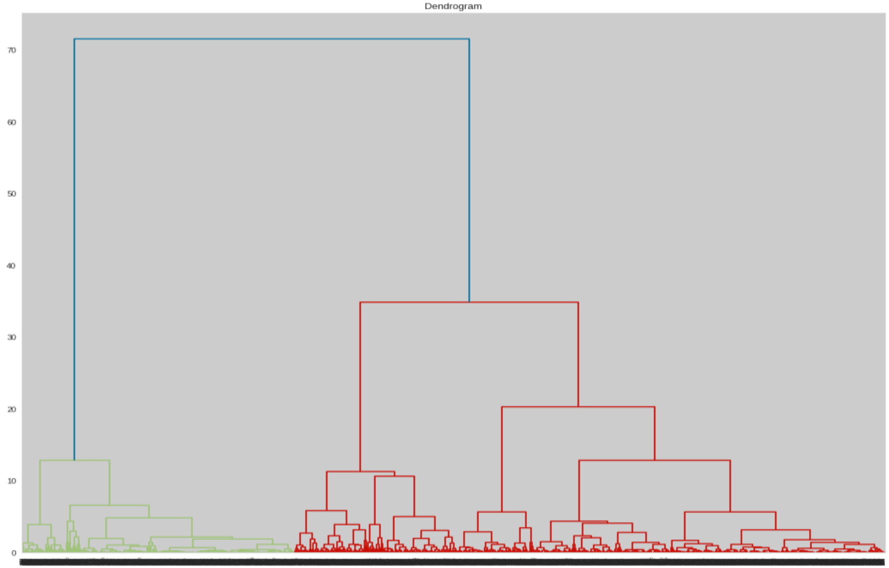
Model (Continued...)

Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.



Dendogram

A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data.



Conclusion:

- Throughout the analysis we went through various steps to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next, we did some exploratory data analysis and tried to draw observations from the features we had in the dataset.
- we saw how we can segment our customer depending on our business requirements. We perform RFM for our entire customer base.
- RFM analysis can help in answering many questions with respect to their customers and this can help companies to make marketing strategies for their customers, retaining their slipping customers and providing recommendations to their customer based on their interest.
- Using cluster profiling the average of recency, frequency and monetary values for each customer segment was identified.
- We used the K-means algorithm to segment our customer in various clusters having similar similarity. K-means did a pretty good job here, Also we remember that the more the number of cluster we take the better the result we get (seperation of multiple cluster).

THANK YOU!!!