Capstone Project-4

ONLINE RETAIL CUSTOMER SEGMENTATION

Submitted by- Shreesha K

A brief introduction

- In this era of a highly competitive business environment, a company must make effective marketing policies.
- Even though mass marketing strategies can get results, it is naive to assume 'one-size-fits-all', i.e., assuming that everyone will be interested in buying whatever a company sells.
- Depending on their lifestyle, preferences, and earning capacity, different customers have different purchase priorities.
- When the company successfully segments the customers into different groups, which share some common behavioral properties, the company can identify the needs of each group of customers, and adopt various marketing strategies to retain the customers, thereby can increase its market share.
- Moreover, customer segmentation also helps to determine new market opportunities, improve distribution strategies, and prevent customer churn.

Contents

- Describing the problem.
- About the dataset
- Exploratory Data Analysis
- Data Wrangling
- Feature Engineering
- Modelling
- Conclusion

Problem statement

Our objective is to build an ML model that can identify major customer segments on a transnational data set that contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. It is a company that sells unique all-occasion gifts. Many customers of the company are wholesalers.

About the Dataset

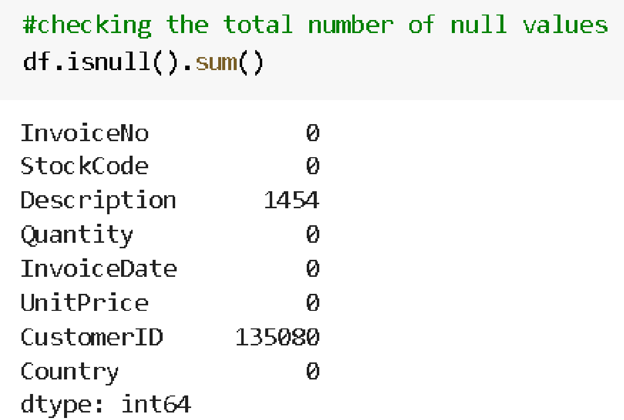
- The dataset contains 541909 rows and 8 columns.
- The columns in the dataset were: 'InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', 'Country’.
- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: 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.

About the Dataset

- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: 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.

Handling NULL and Duplicate entries

- In the ‘customerID’ column, almost 135080 values were and in the 'Description' columns, almost 1454 values were missing (Null). Since these values can't be replaced with any other values, the only option was to drop them.
- There were around 5225 duplicated entries and I removed them also.

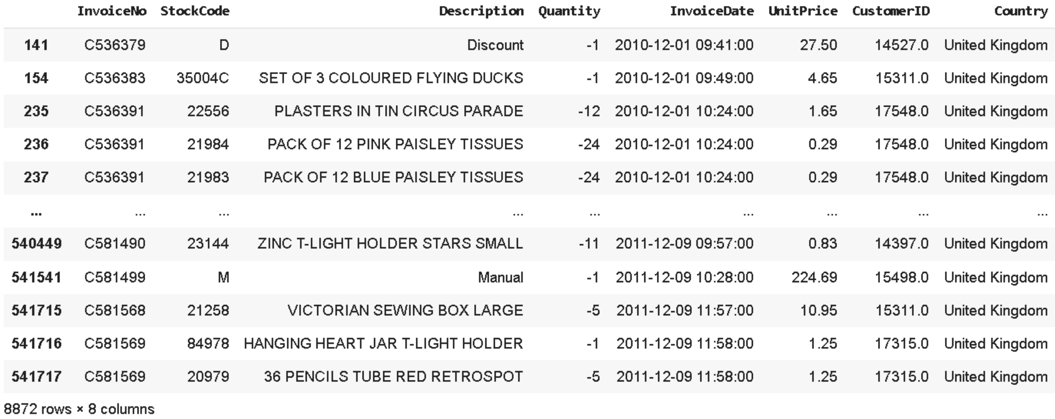


```
#checking the total number of null values
df.isnull().sum()
```

```
InvoiceNo            0
StockCode            0
Description       1454
Quantity             0
InvoiceDate          0
UnitPrice            0
CustomerID      135080
Country              0
dtype: int64
```

Data Wrangling

- **Some orders had negative unit quantities. When checked, the IDs corresponding to these orders had the prefix 'C' indicating that the order was canceled. These orders were also dropped.**

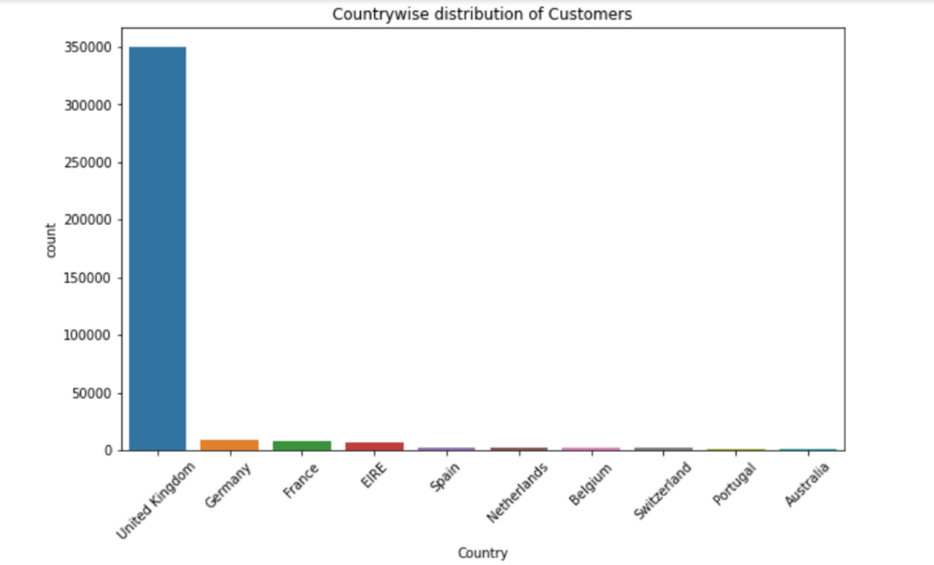


	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1.25	17315.0	United Kingdom

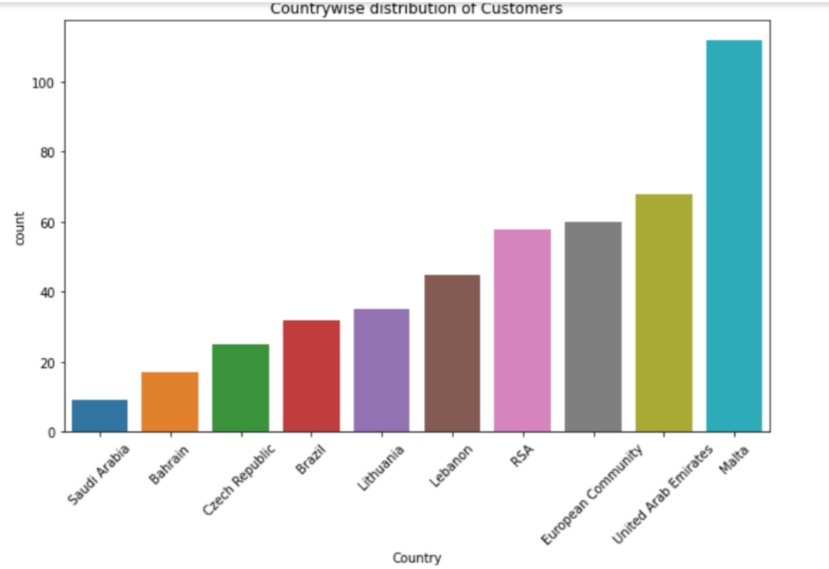
8872 rows × 8 columns

- **After removing the null, duplicate, and negative orders, there were 392732 rows left in the dataset.**

Exploratory Data Analysis



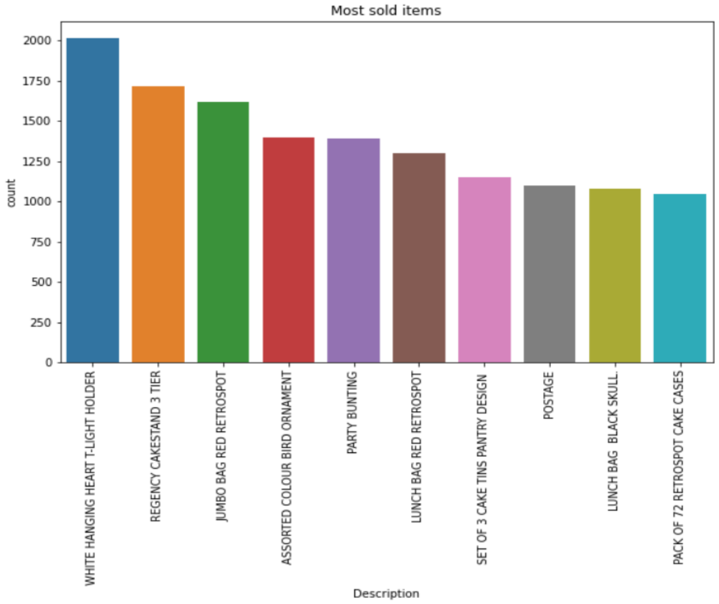
Countrywise distribution of Customers

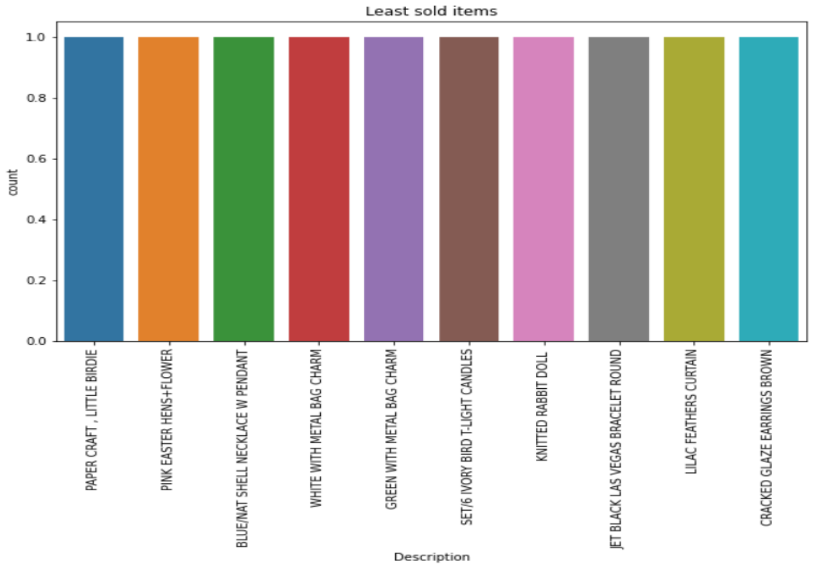


Countrywise distribution of Customers

- Most of the customers are from the United Kingdom and Saudi Arabia is the country with least number of customers.
- The proportion of customers from other countries is too less in comparison with that of the UK.

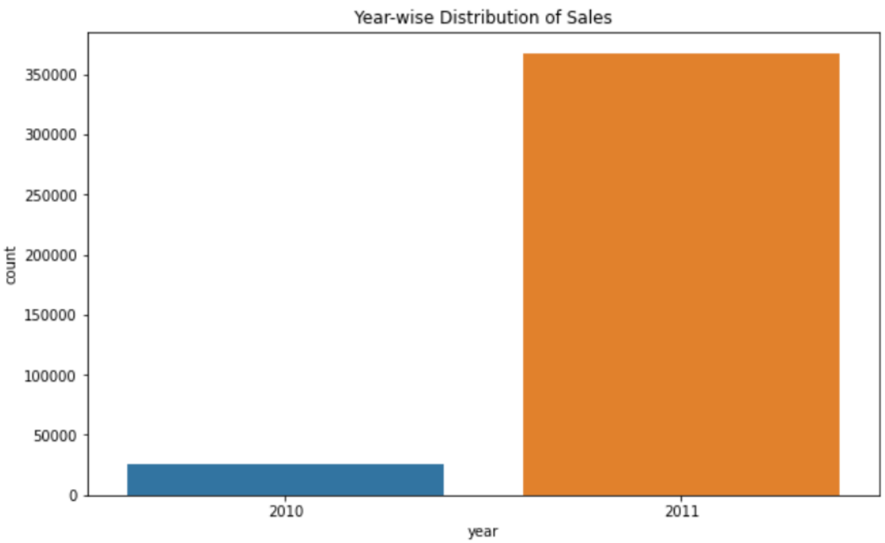
Exploratory Data Analysis

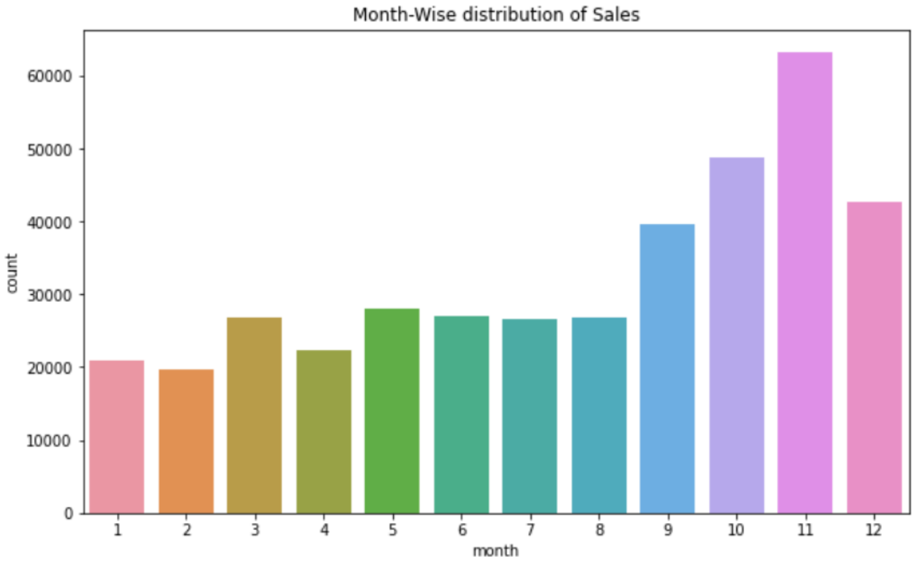




- 'WHITE HANGING HEART T-LIGHT HOLDER' is the highest sold item.
- We have many items whose only one unit has been sold.

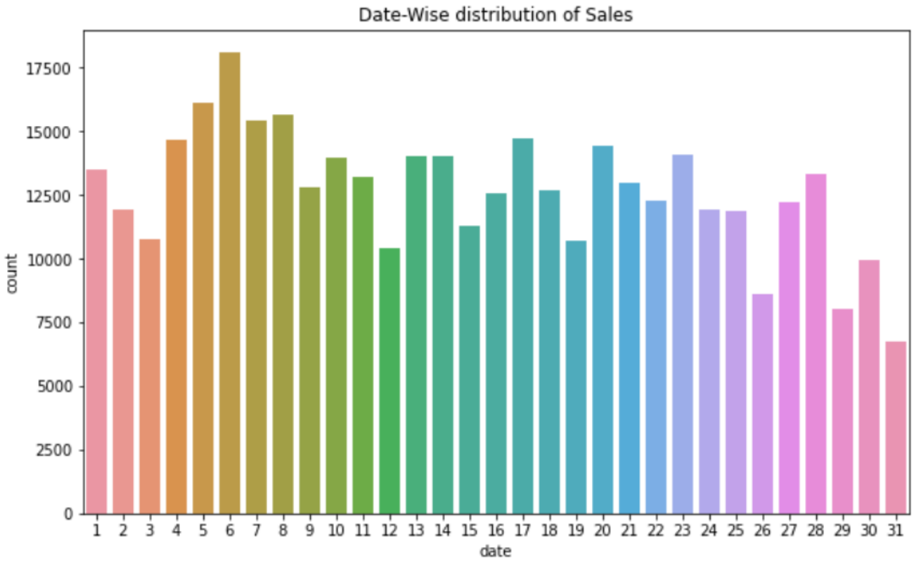
Exploratory Data Analysis

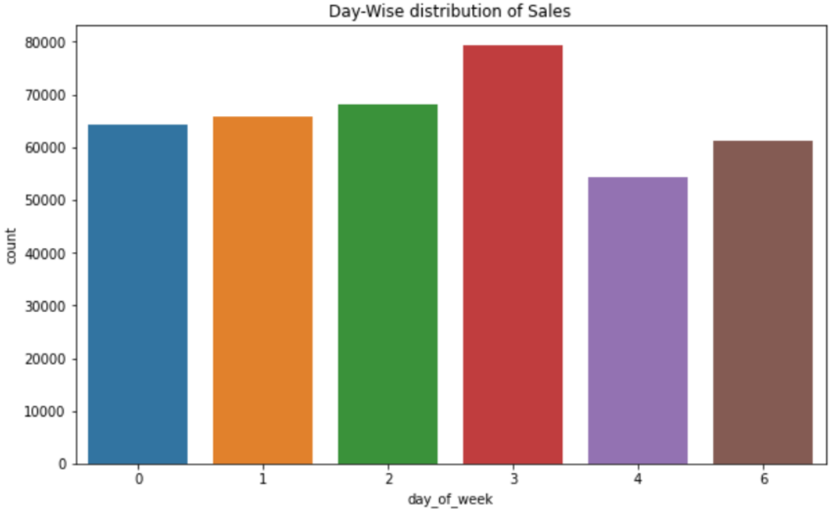




- 2011 was the year with the highest number of sales.
- November was the month with the maximum number of sales. We can also observe that most sales occurred in the ending months of the year.

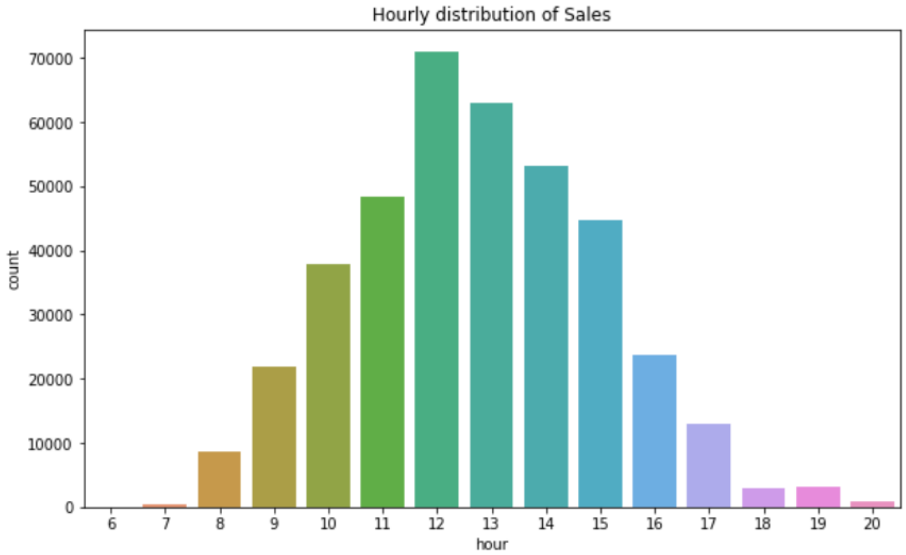
Exploratory Data Analysis

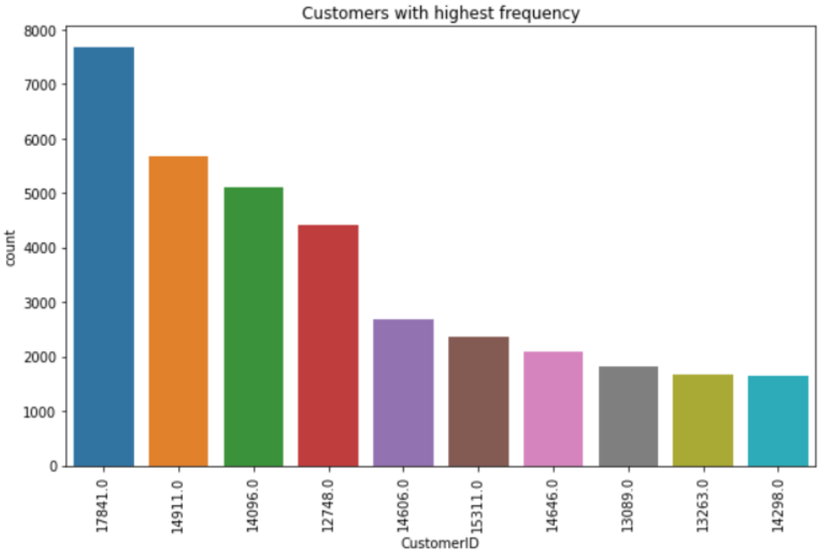




- Most sales were recorded at the beginning of the month.
- The highest sales occurred on Thursday, while on Friday, the least sales were recorded. We can note that there were no sales on Saturday.

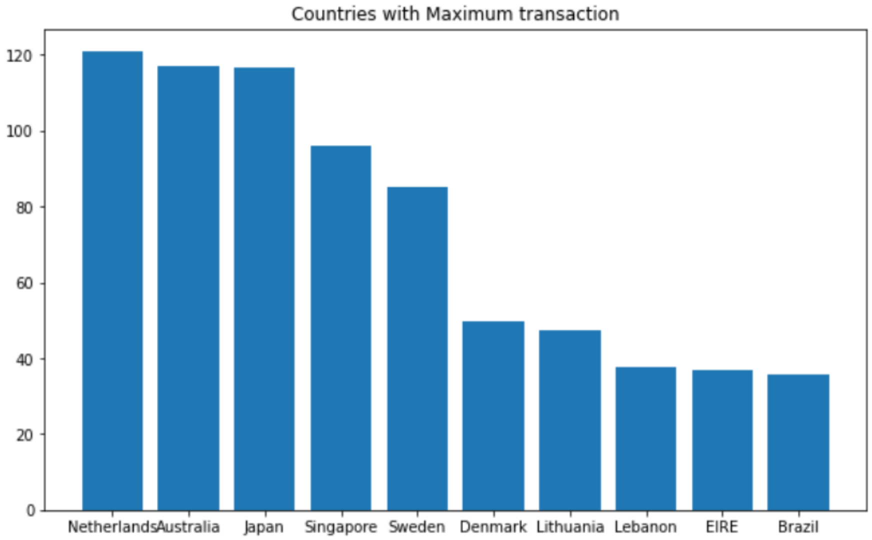
Exploratory Data Analysis

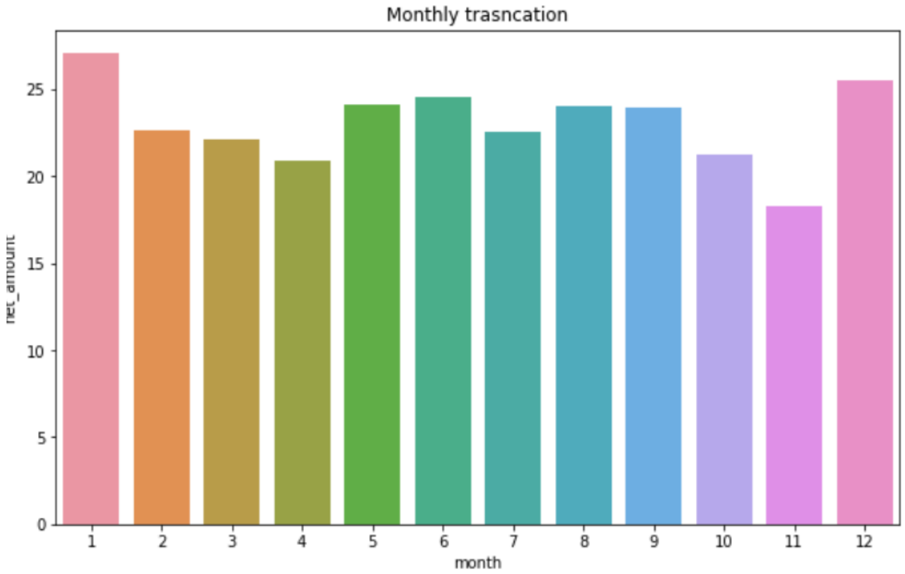




- In the afternoon, more sales were recorded. There were no sales before 6 AM and after 8 PM.
- Customer with customer ID 17841 has the highest frequency.

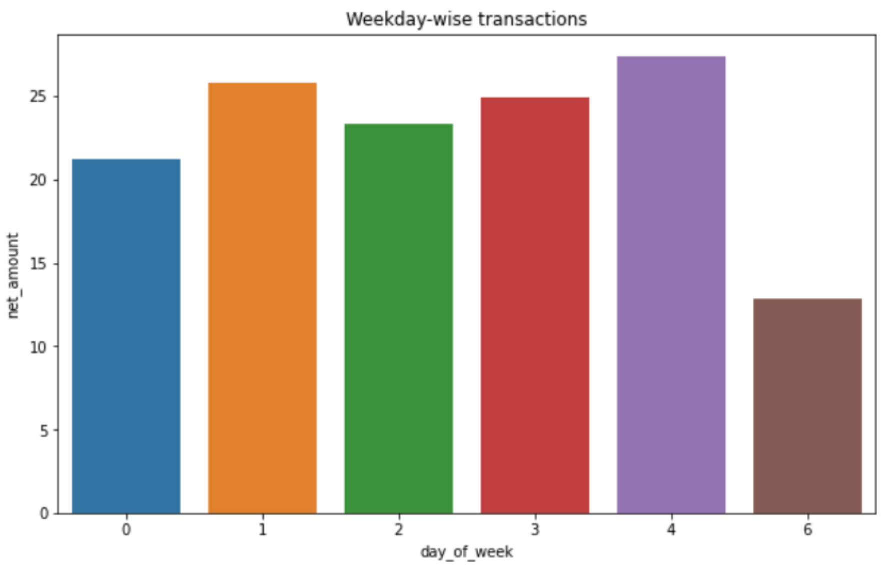
Exploratory Data Analysis

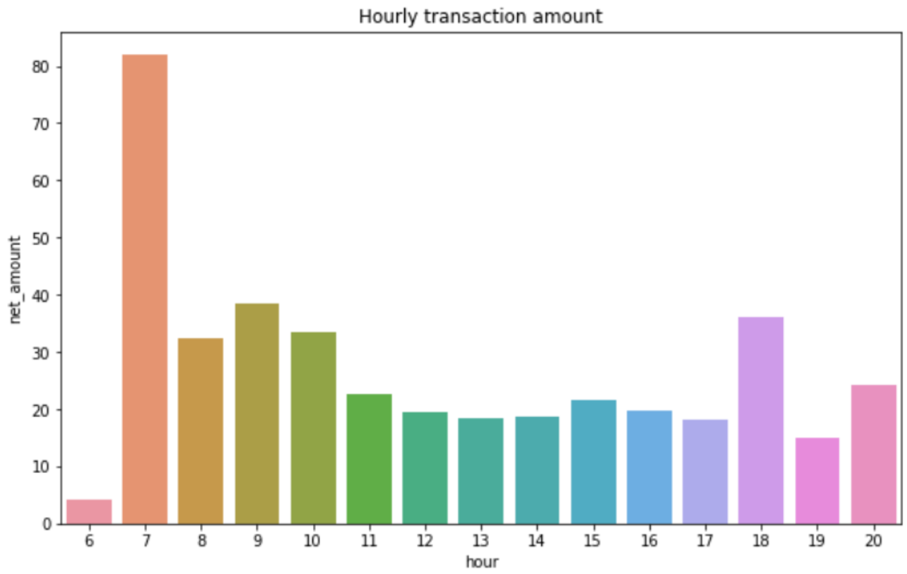




- The total transacted amount in an order can be calculated by multiplying the Unit Price by the Quantity (Transacted amount= Quantity * Unit Price).
- The Netherlands is the country that generates the highest revenue for the company.

Exploratory Data Analysis





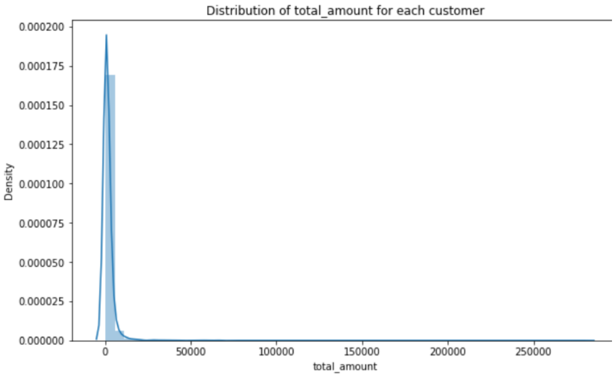
- On Friday, net transaction was the highest. Sunday is the day in which net transaction was the least.
- We can see that at 7AM, net transaction was the highest.

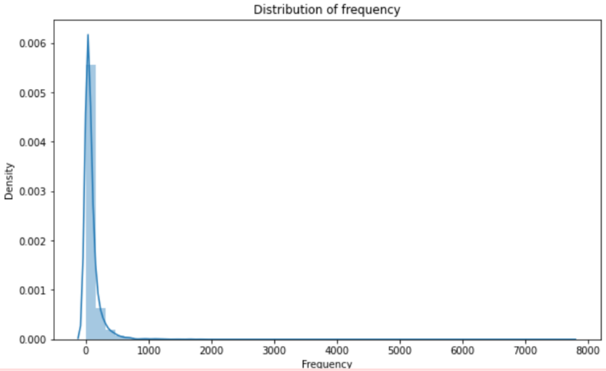
Feature Engineering

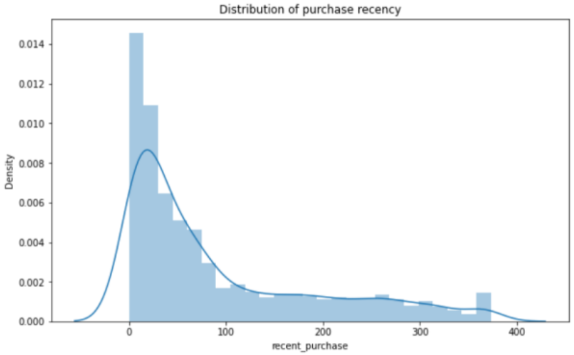
For our further analysis, a new dataframe was formed having the following columns.

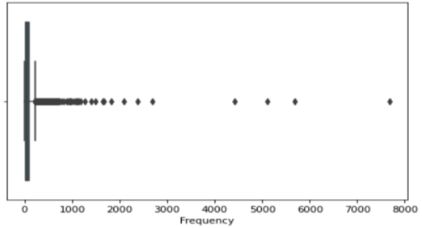
- Frequency: How often do the customers visit or how often do they purchase? It can be calculated by counting the Invoice no. of transactions.
- Total Amount: It is the revenue the company gets from each customer or it is the amount spent by the customer. We can calculate it by summing up the net transacted amount for each customer.
- Recent_purchase: It indicates how recently the customer visit the website or how recently did a customer purchase. It is calculated as the Latest date- the last invoice date.

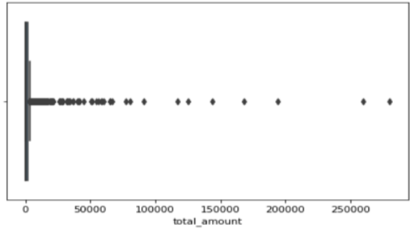
Feature Engineering

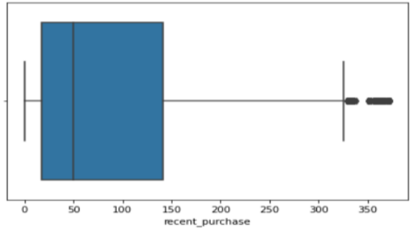

Distribution of total_amount for each customer


Distribution of frequency


Distribution of purchase recency

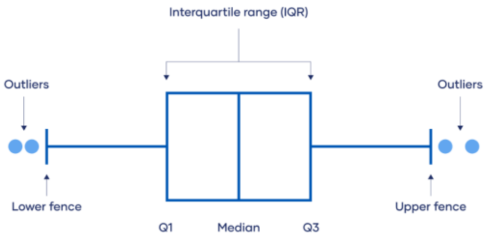




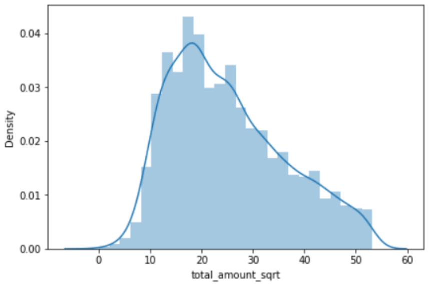


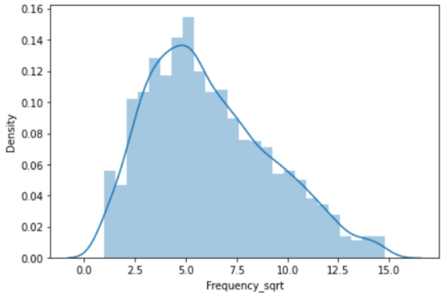
Feature Engineering

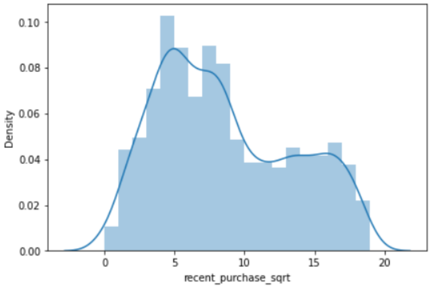
- **Outlier Removing:** Outliers were removed using the quantile method.



- Took the square root of each feature to reduce the skewness.





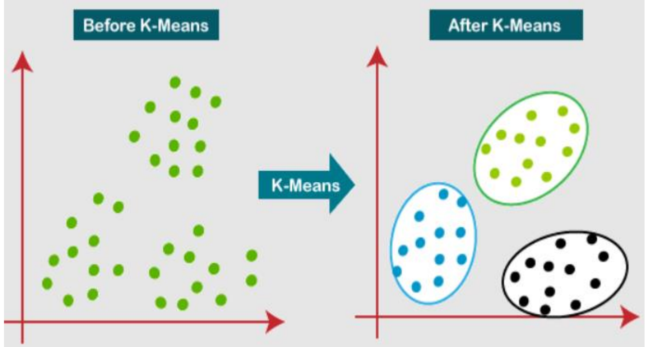


- Using the MinMaxScaler, the values were standardized.

Model building

KMeans Clustering

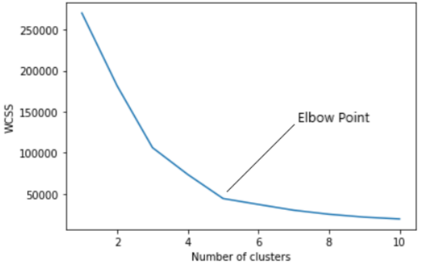
KMeans clustering is an unsupervised ML algorithm that groups the unlabeled dataset into different clusters. The user must specify the number of clusters to be formed. It is a centroid-based algorithm, which indicates that each cluster is associated with a centroid. The algorithm minimizes the sum of squared distances (SSD) between a given data point and its corresponding centroid.



Model building

The Elbow Method

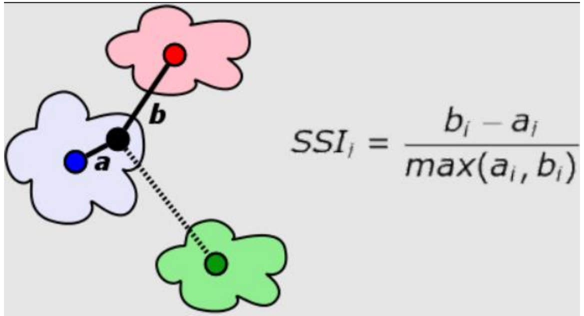
To find the optimum number of clusters, we use the elbow method. We first set several cluster sizes (K). For each value of K, we will fit the model and find the value of WCSS Score (Within Cluster Sum of Squares, also known as SSD). If we plot the values of K with their corresponding WCSS Score, the curve looks like an elbow. The WCSS value will rapidly decrease till a particular value of K, then after the curve becomes almost parallel to the X-axis. The K value corresponding to this transition point is the optimum value for K and further analysis is carried out.



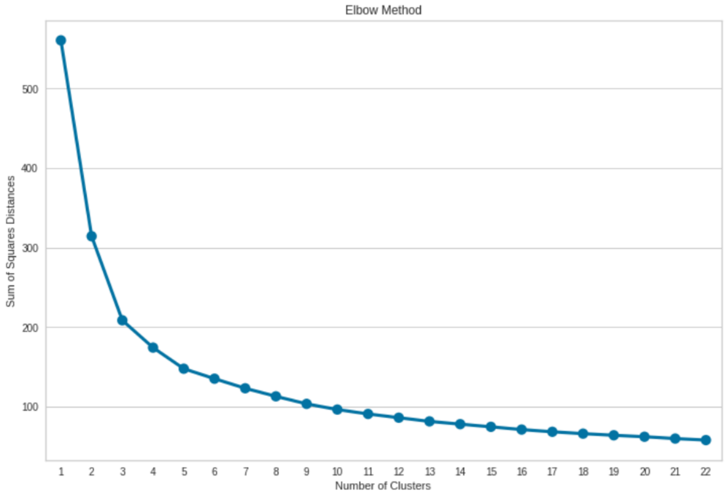
Model building

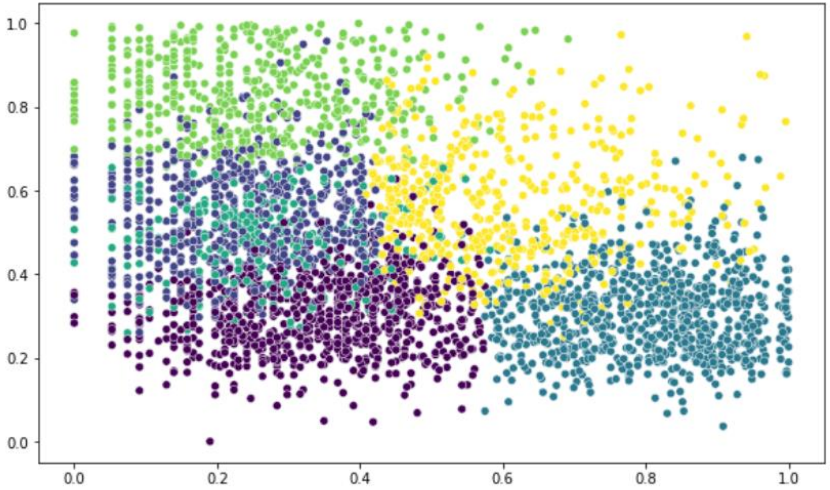
The Silhouette Score

The Silhouette Score is a performance parameter used to find to what extent the clustering model has performed. It is a score ranging between -1 to 1. A silhouette score of 1 means the clusters are well apart from one other and distinguished clearly, while -1 indicates the clustering was performed in the wrong way.



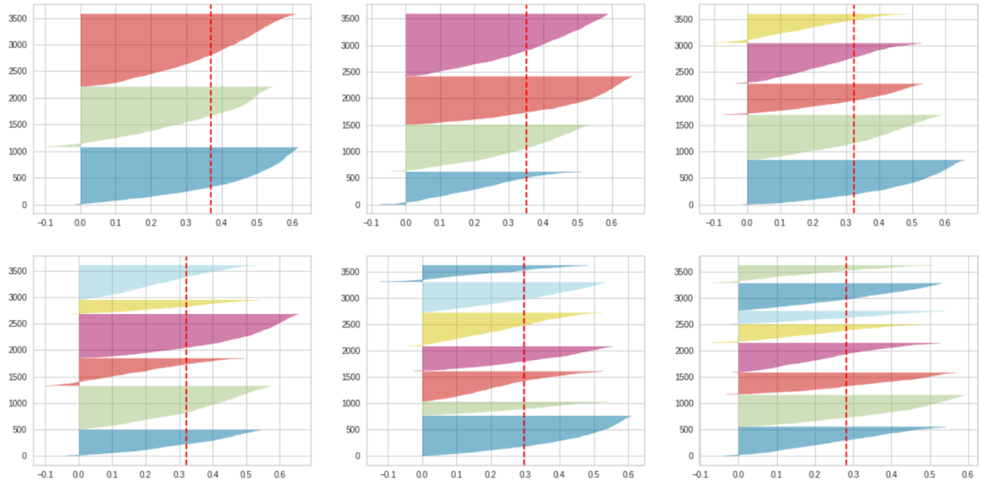
Model building





- We can see that when the number of clusters is 3, the shape of the curve suddenly changes and it can be considered as the transition point.
- We randomly checked for 6 clusters. The performance was not too great.

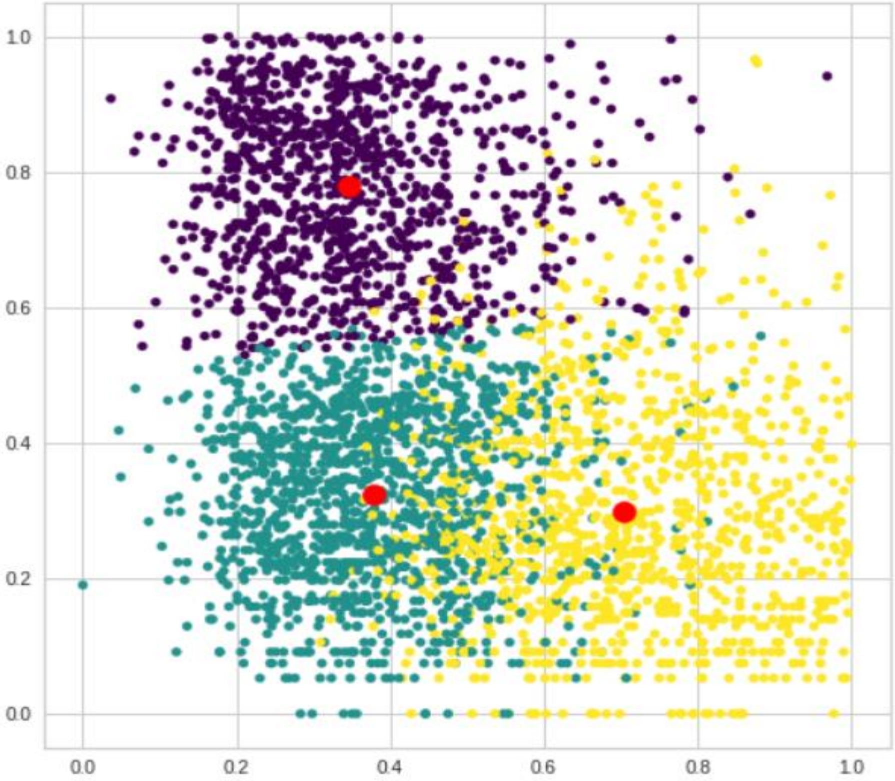
Model building



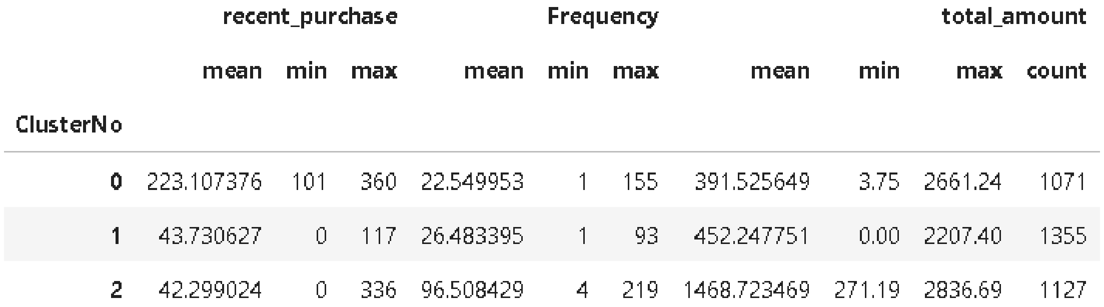
```
silhoutte score is 0.3702875239938246 for n clusters=3
silhoutte score is 0.352802105294061 for n clusters=4
silhoutte score is 0.32473039932052883 for n clusters=5
silhoutte score is 0.3196529178920309 for n clusters=6
silhoutte score is 0.2905181565496759 for n clusters=7
silhoutte score is 0.2940063368673435 for n clusters=8
silhoutte score is 0.2769665225736871 for n clusters=9
silhoutte score is 0.2691083531030274 for n clusters=10
```

- **From the elbow plot, we can observe that the optimum number of clusters might be between 3 to 10.**
- **After calculating the silhouette score for each number of clusters, we found that for n=3, the score was the highest.**

Model building



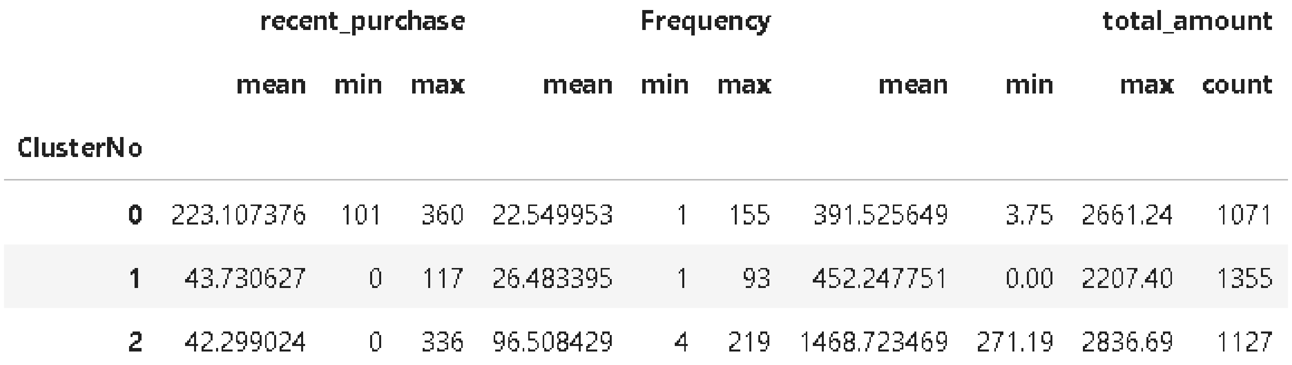
Summary And Conclusion



	recent_purchase			Frequency			total_amount			
	mean	min	max	mean	min	max	mean	min	max	count
ClusterNo										
0	223.107376	101	360	22.549953	1	155	391.525649	3.75	2661.24	1071
1	43.730627	0	117	26.483395	1	93	452.247751	0.00	2207.40	1355
2	42.299024	0	336	96.508429	4	219	1468.723469	271.19	2836.69	1127

- **After forming the clusters, we observed the behavior of customers in each cluster.**
- **'Cluster 0' Customers- Their recent purchase date is too old. They do not purchase very frequently and the amount they spend is not high. We can classify them as Low Valued Customers. There were 1071 customers in this cluster.**
- **'Cluster 1' Customers (1355 customers)- These customers have placed an order more recently, but the frequency with which they order is not that great. They spend comparatively more than cluster 0 customers. We classify them as Medium Valued Customers.**

Summary And Conclusion



	recent_purchase			Frequency			total_amount			
	mean	min	max	mean	min	max	mean	min	max	count
ClusterNo										
0	223.107376	101	360	22.549953	1	155	391.525649	3.75	2661.24	1071
1	43.730627	0	117	26.483395	1	93	452.247751	0.00	2207.40	1355
2	42.299024	0	336	96.508429	4	219	1468.723469	271.19	2836.69	1127

- 'Cluster 2' Customers- These customers are the most loyal to the company. They place the orders more frequently and recency is also high. The average amount they spend on an order is also high. They are 'High Valued Customers’. We have 1127 customers in this cluster.

Summary And Conclusion

- We can reward Cluster 2 customers. They can adapt to new products. Moreover, they help to promote the company.
- For cluster 1 customers, we can offer membership or loyalty programs such as discounts and special incentives. We can recommend them related products to upsell the products. This initiative helps to make them loyal customers of the company.
- In general, the customers in cluster zero do not have much impact on the company. But we have some customers who spent heavily in the past but not recently. We need to identify them and initiate a reactivation campaign. We can also offer them promotions and conduct surveys to know what went wrong. In this way, we can avoid losing them to competitors.

Summary And Conclusion

This brings us to the end of the project. We can perform some more robust analysis by considering other metrics like demography, lifestyle, or product features.

Thank you