greatlearning

PGP-DSE June 2020 Batch

Capstone Project

Final Report

Analysis of Customer Purchase Intentions in an E-Commerce Sector

Group 7

Arun V
Nikhil G
Nivetha Sekar
R Mathana Gopal
Sharavanan Rajaram

Mentor: Dr. Dipanjan Goswami
Domain: E-Commerce

Table of Contents

Introduction **2**

1.1 Understanding the Business 2

1.2 Aim of the Project 2

Literature Survey **3**

2.1 Proposed Methodology and Related Works 3

Dataset Description **5**

3.1 Attribute Information 5

3.2 Variable Categorization 6

Data Preprocessing **7**

Exploratory Data Analysis **10**

5.1 Relationship between variables 10

Feature Engineering **18**

6.1 RFM Analysis 18

K-Means Clustering **24**

7.1 KMeans Algorithm: 24

7.2 K-Means clustering on RFM variables 27

7.3 Principal Component Analysis (PCA) 31

Validation of Model **34**

8.1 Random Forest Classifier 34

8.2 Support Vector Machines (SVM) 35

Customer Segmentation from KMeans Clustering **37**

Business Recommendations **45**

Summary and Future Works **49**

References **55**

Appendix **56**

Customer Segmentation using RFM analysis 56

Additional Analysis 57

1. Introduction

The E-Commerce industry has enormously grown in the last couple of years capitalizing on the Internet penetration and Digital financial services, which has in turn affected the sales for many brick-and-mortar stores. Online Retail platforms, with their easy accessibility and convenience, allow customers to purchase products directly through their websites or mobile applications. Customers now easily buy, cancel and return the products from the comfort of their home.

E-Commerce websites provide details about the products, real-time reviews and recommendations. Online store prices are comparatively lesser than the offline stores. In most of the cases, we can observe that online platforms drive offline sales through provisions to automatically manage the product inventory. Online Retail provides opportunities for small to large stores and has proven to be one of the most trusted platforms by both buyers and sellers.[1]

1.1 Understanding the Business

The UK-Based Online Retail non-store business is involved in the sale of unique gifts for all occasions. They have recently launched an E-Commerce website to sell their products in order to expand their business in the B2B sector and since its launch they have been maintaining a steady number of customers across all parts of the United Kingdom. The Online Retail company deals with mostly wholesalers but also has some regular customers.

According to a survey conducted by the Interactive Media in Retail Group (IMRG), Online shopping has increased drastically about 5000% when compared to a decade ago in the United Kingdom. The consumer's way of shopping and expenditure has changed a lot in the past 10 years. The exponential increase indicates the interest shown by the consumers in the online retail when compared to in-store purchases.[2]

1.2 Aim of the Project

We propose to provide a business solution for this dataset through the following steps:

- Performing RFM Analysis to identify clusters based on customer purchase behavior so that the business is sustainable for a long time.
- Exploring the most purchased products, high-valued and least-valued customers.
- Identifying the best business strategy for improving the sales in the online retail business.
- Suggesting customized marketing campaigns for each customer segment to maximize sales and strengthen customer-business relationship.

2. Literature Survey

Due to the dynamic nature of the retail industry and technological advancements in online e-commerce platforms, identifying the factors contributing to the growth of a business has become the need of the hour. In recent days, identifying patterns in consumer activity has proven to be useful in targeting the right customers for a product. Customer transaction data along with the items purchased can be analyzed to identify the product buying rate and generate other valuable information about consumer behavior.

Understanding the customer purchase behavior is a key factor in determining the success or failure of a business. As a result of which many companies are now spending their time and money in formulating new business strategies to strengthen customer relations. One of the widely used techniques in segmenting the customers based on their purchase behavior is through Market Basket Analysis and RFM (Recency, Frequency, Monetary Value) analysis.

2.1 Proposed Methodology and Related Works

Building clusters based on customer purchasing patterns to identify best marketing strategies

The Online Retail dataset that we have chosen for analysis could be used to identify patterns in customer purchase intentions which could be materialistic in designing customer-centric business/marketing strategies. We will be building models to identify clusters having similar purchase behavior and segment the customers based on the type of products purchased, region-wise segmentation as well as frequency of purchases made. This could be achieved through K-Means Clustering and Agglomerative Clustering.

Related Works

P. Anitha, et.al., [3] proposes a methodology to use K-means clustering using Euclidean distance metric in order to segment the consumers based on the calculated RFM values. In this paper, KMeans clustering is implemented twice for analyzing the total transaction amount received for both Recent as well as Frequent transactions by partitioning customers based on I) Recency Vs Monetary value II) Frequency Vs Monetary value. Analysis of Silhouette scores for both transactions further resulted in model optimization.

Singh et.al., [4] presents a methodology to perform market segmentation using RFM analysis on a Big EFTPOS data. After the identification of RFM values for each retailer, KMeans Clustering and Agglomerative Hierarchical Clustering models were developed in order to identify active and inactive retailers. This helps in identifying the risk of attrition as well as provides insights to modify the marketing strategies used.

In Chen, D. et.al., [5] , an RFM based customer segmentation model has been implemented to identify similar characteristics among the customers and using K-means clustering and further improvements have been made using Decision Tree Classifier. This paper also provides recommendations based on the findings from RFM analysis in order to design customer-centric marketing strategies. Using the online transaction data of a retail company, customers were segmented into five clusters having both existing and new customers. In order to improve the performance of the model, the Decision Tree algorithm was incorporated to further classify the clusters into ‘Existing’ and ‘New’ consumer categories.

3. Dataset Description

The Online Retail dataset contains information about all transactions made between a time period of 1 year for an UK based non-store completely online retail franchise which sells gifts for all the occasions/festivals. The dataset also contains transactions made from other countries apart from the UK. But we could see that the majority of transactions belong to one country, the UK. Also, most of the customers are wholesalers.

There are nearly 541909 records and 8 features in this dataset.

The link for the dataset is given below:

https://archive.ics.uci.edu/ml/datasets/online+retail

3.1 Attribute Information

InvoiceNo	The Invoice no is a 6-digit numerical number which was generated at the time of the transaction. The Invoice number could precede with a character 'C' which denotes that the order was cancelled.
StockCode	Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description	Description consists of the Name of the product which is actually Nominal in nature.
Quantity	Quantity describes the number of quantities the product was purchased in a single transaction.
InvoiceDate	Invoice number consists of a Datetime format value inducted into this column. It tells about the year, month, date and time of the purchase.
UnitPrice	UnitPrice describes the price of the product for a single quantity.
CustomerID	CustomerID is a unique 5 digit numeric and nominal entry that is assigned to a customer and is linked with the purchase's invoice and quantity.
Country	Country shows which country/region the customer belongs.

3.2 Variable Categorization

The Numerical Features are:

- Quantity
- UnitPrice
- CustomerID

The Categorical Features are:

- InvoiceNo
- StockCode
- Description
- Country

The Datetime Features are:

- InvoiceDate

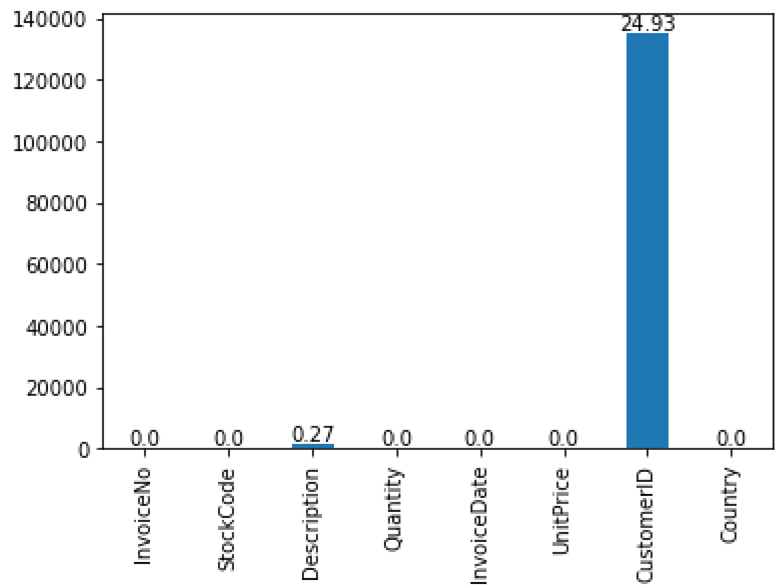
4. Data Preprocessing

After understanding the features of the given dataset, the next step is to prepare our data for modelling.

Checking the datatype of the features:

It was observed that CustomerID has a float data type which needs to be converted to int. And the InvoiceDate is in DateTime format which can be further used to extract only the month for seasonality classification.

Identifying Missing Values:



Nearly 24.9% of CustomerID and 0.268% of Description are missing. Since, we are only interested in Customer-centric behaviour, we can drop the missing values.

Checking Duplicate Entries:

In total there are 5225 duplicate entries, which can be removed from the dataset.

Data Type Conversion:

Convert CustomerID column to Integer data type from float.

Descriptive Statistics:

- We can observe that the mean quantity of products purchased by the customers is 12.
- We can observe a cancelled order for a high quantity of 80995 units.

	Quantity	UnitPrice
count	401604.000000	401604.000000
mean	12.183273	3.474064
std	250.283037	69.764035
min	80995.000000	0.000000
25%	2.000000	1.250000
50%	5.000000	1.950000
75%	12.000000	3.750000
max	80995.000000	38970.000000

Analysis of negative values in UnitPrice & Quantity:

- Negative values indicate a cancelled order.
- We have 8872 cancelled items, which is nearly 2.2% of total orders.
- This will have a significant impact on sales, which can be dropped from the data set.

Additional Observations:

- There are 4372 existing customers.
- Total number of unique orders: 22190
- We identified a few Non-Product entries in Description.
    - Bank Charges, CARRIAGE, Discount, DOTCOM POSTAGE, Manual, Next Day Carriage, PACKING CHARGE, POSTAGE
    - There 1991 entries with non-product items

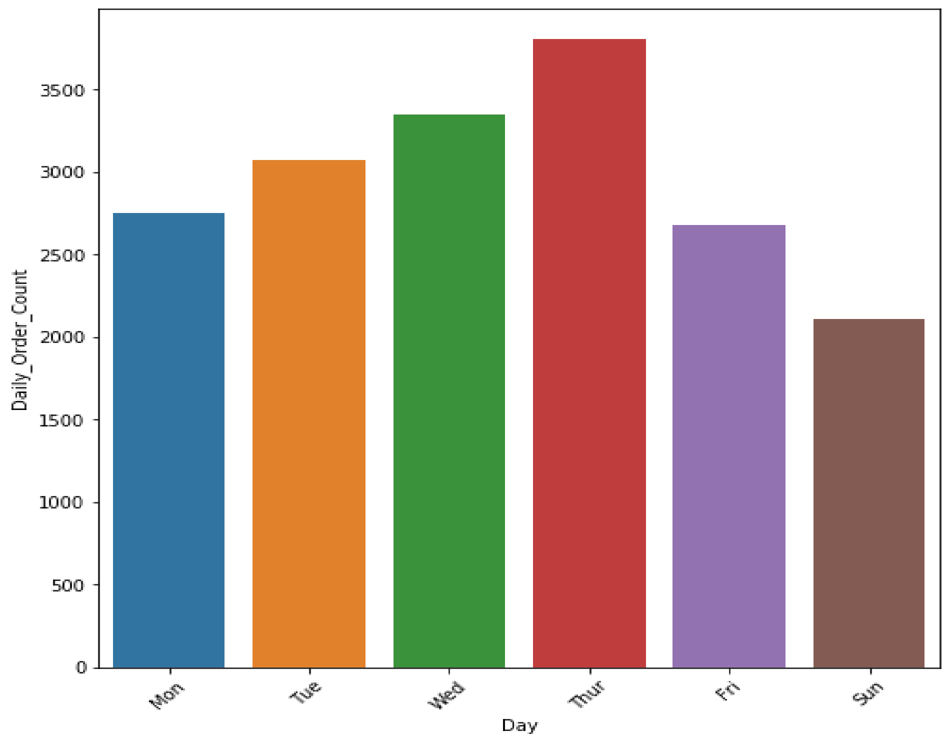
- We created new columns for total sales (Quantity*UnitPrice)
- We extracted date, year, month, day, quarter and hour from the Invoice date column.
- Nearly 88.95% of sales are based out of the UK.
- Not even 0.1% of sales are from Denmark, Japan, Poland, USA, Israel, Singapore, Iceland, Canada, Greece, Malta, United Arab Emirates, European Community, RSA, Lebanon, Lithuania, Brazil, Czech Republic, Bahrain, Saudi Arabia.
- And few countries are 'Unspecified' which can be dropped (only 0.059% sales)
- Since there are few countries with very less sales, we are categorizing them into "Others".
- There are 3665 unique StockCode entries.
- There are 3877 unique Description entries.
- As we can see here there are 3665 unique StockCode entries but 3877 unique Description values. Therefore, further analysis has to be done to identify duplicate entries for Description.
- The top 5 customers are 17841, 14911, 14096, 12748 and 14606.
- In the top 20 Customers, most of them belong to the United Kingdom, 2 of them are from EIRE and 1 from the Netherlands.
- Bottom Customers are mostly from the United Kingdom.
- Most of them are from the United Kingdom, 2 from EIRE, 1 from the Netherlands and Australia.
- The customer 14646 has spent the most amount of money (280206.02 Euros)
- All the customers are from the United Kingdom.
- Customer 13256 from the United Kingdom has spent 0 euros which shows that he is not a valuable customer.

5. Exploratory Data Analysis

5.1 Relationship between variables

Bivariate Analysis:

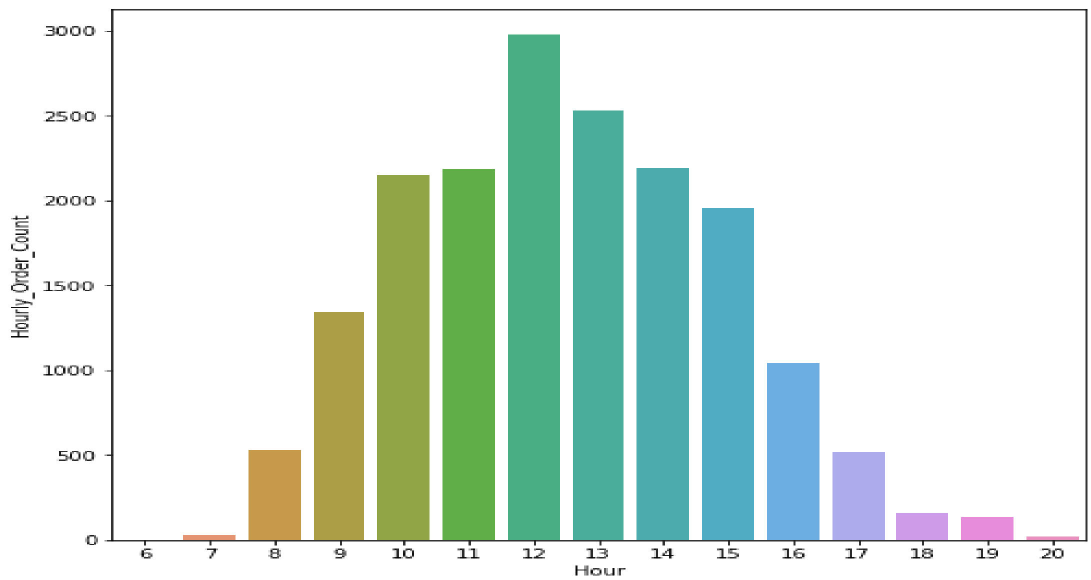
Daily Order Count



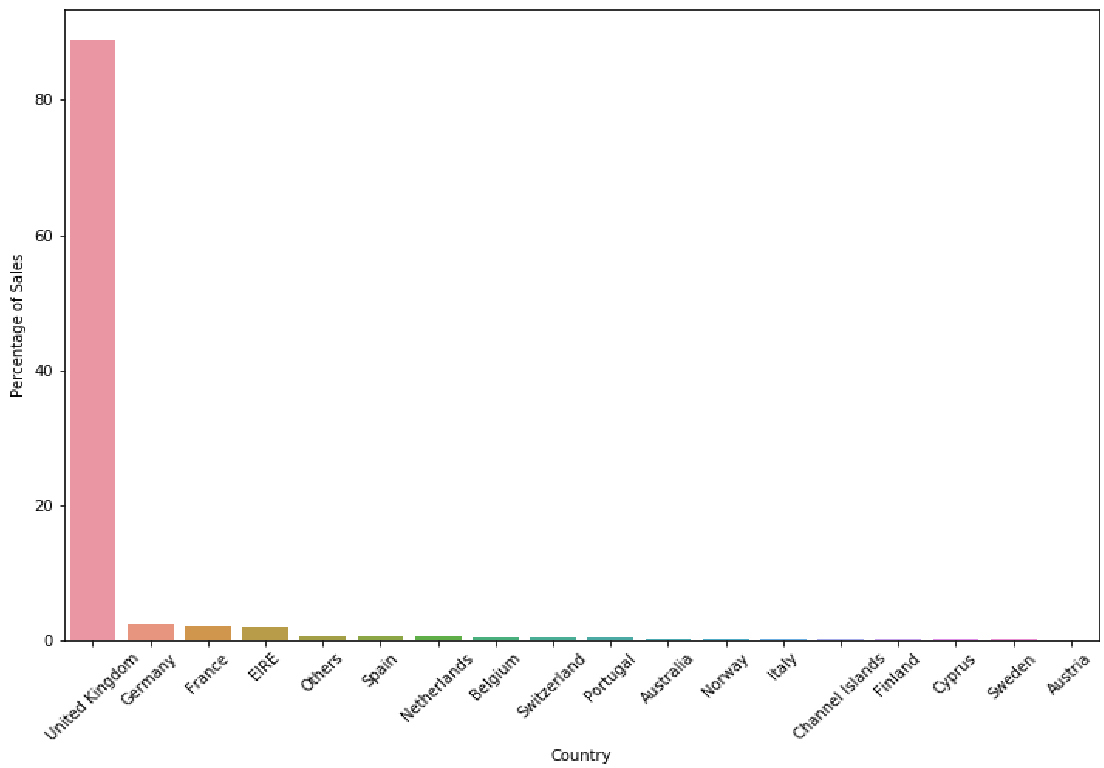
There are no orders placed on Saturdays and the highest number of orders were placed on Thursdays.

Hourly Order Count

- Mostly the orders are placed during 9AM to 4PM (Business Hours) with maximum no. of orders placed around 12PM.
- No orders are placed from 9PM to 6AM.



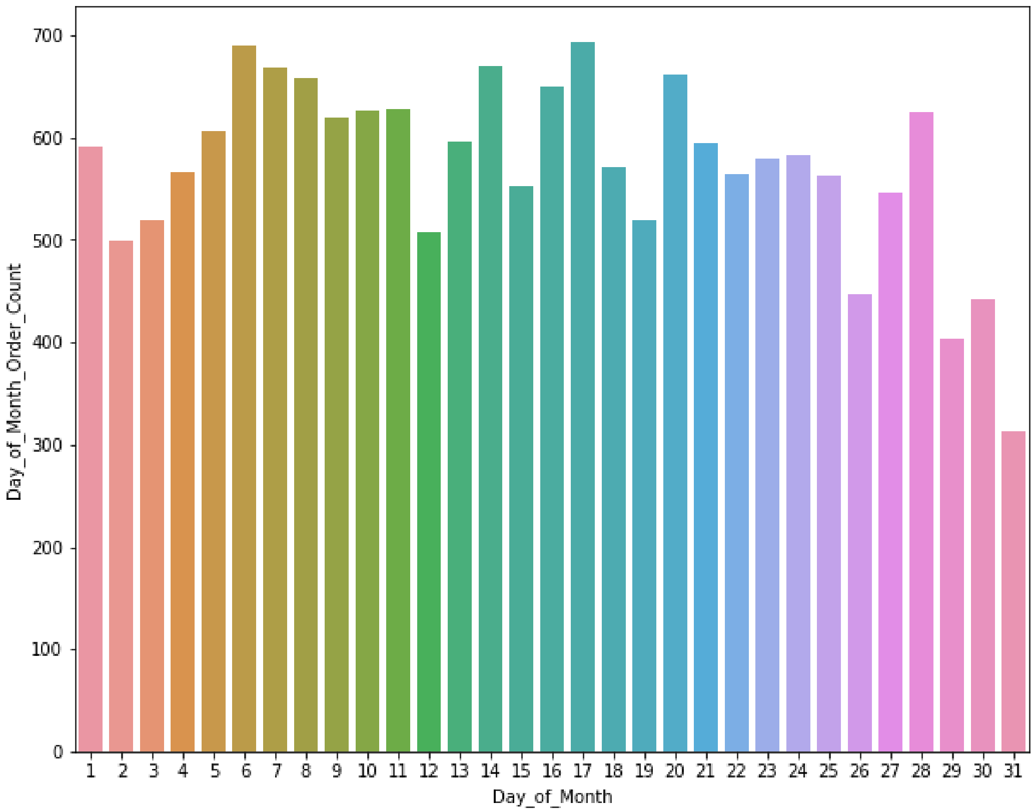
Countries with high Sales



The United Kingdom has high sales (88.9%) followed by Germany, France and Eire with more than 1% of sales remaining countries having less than 1% of sales.

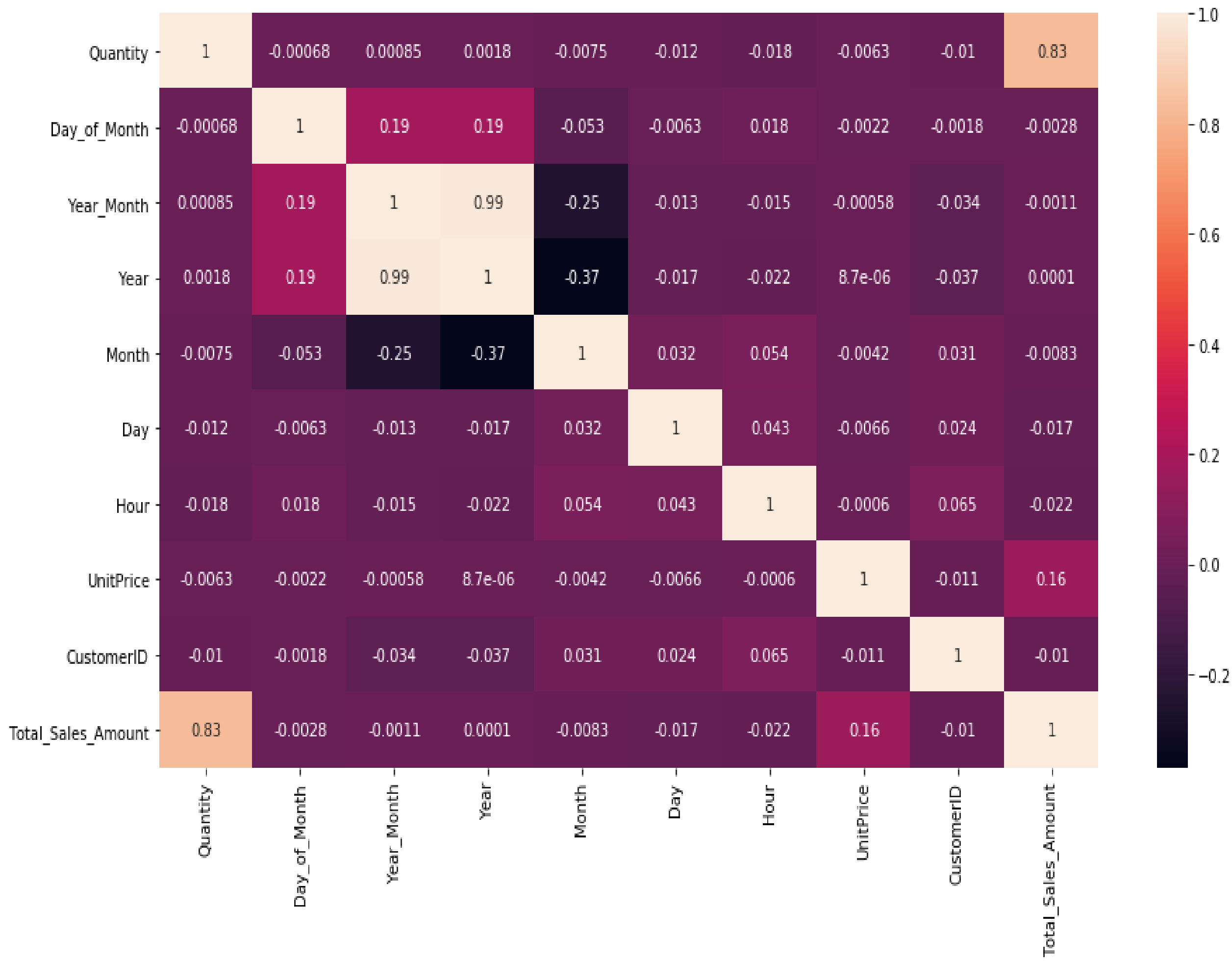
Day of Month Order Count

- As we can observe from the below graph, there is no significant difference in the number of orders placed each day of the month.
- There is a slight increase in the number of orders placed during the first two weeks and the last two weeks shows a slight dip in sales with least number of orders on the last day of the month.



Multivariate Analysis

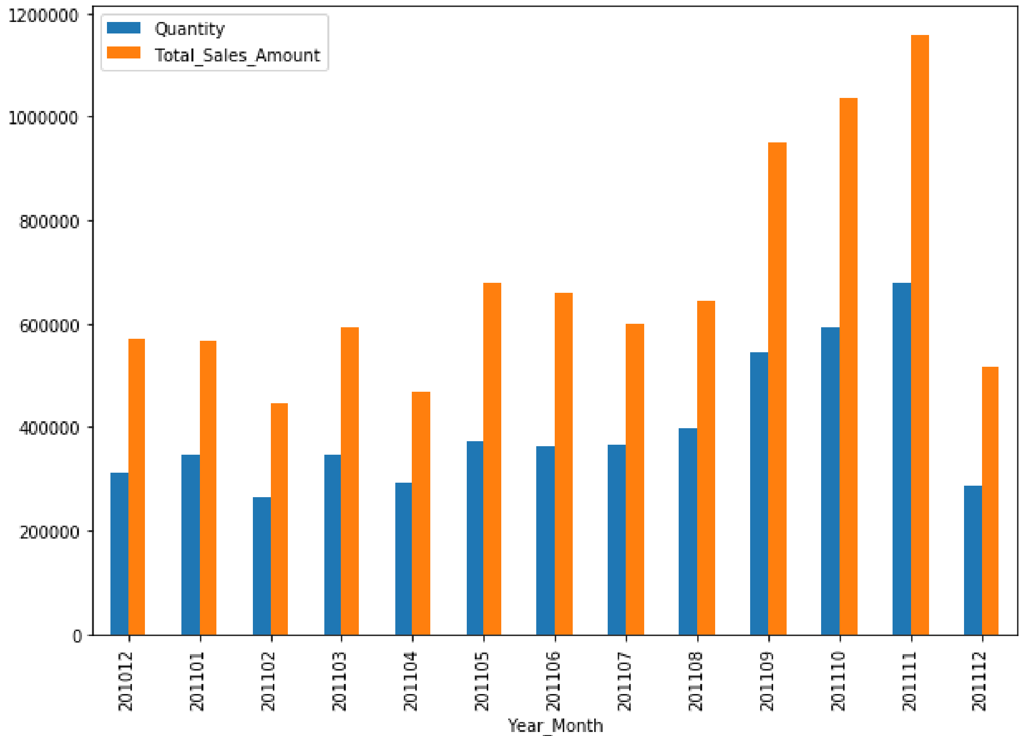
- Year_Month and Year (0.99)
- Quantity and Total_Sales_Amount (0.83)
- Need to do Scaling after we do RFM analysis to further analyze the correlation.



Quantity and Total_Sales_Amount Vs Year_Month

- The monthly revenue is highest for November 2011 followed by October and September.
- Revenue generation from September 2011 till November 2011 has increased significantly.

The number of quantities sold made in Dec 2011 is unusually low.

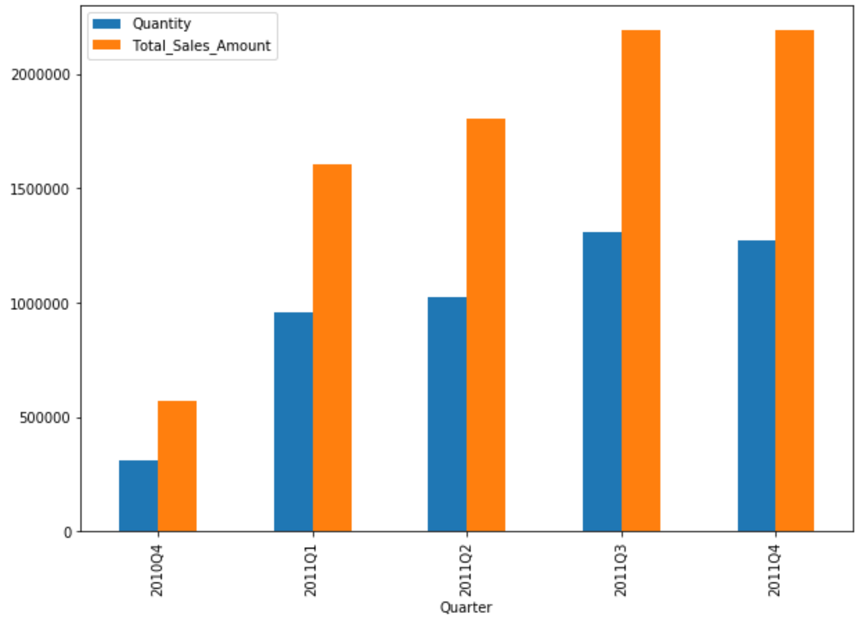


Checking orders in Dec 2011

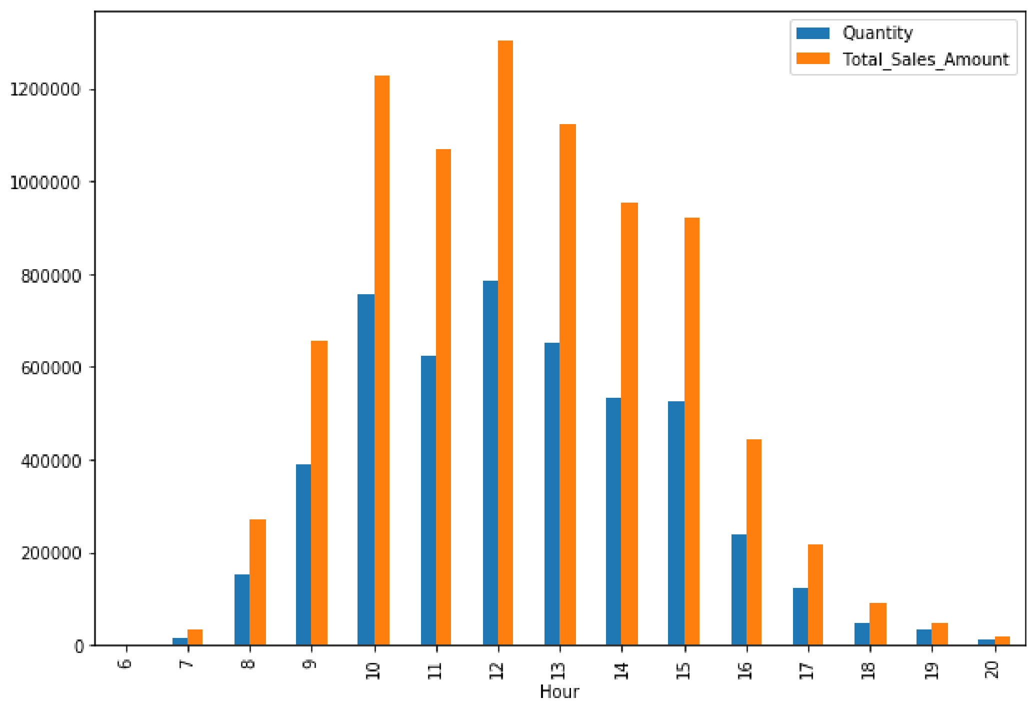
- We can observe that the order details in December are from 2011-12-01 08:33:00 to 2011-12-09 12:50:00 which is incomplete for analysis. Therefore, we can drop this year_month.
- High Sales is Observed for November month.
- This could be due to Black Friday Sales (25th November 2011).

Quantity and Total_Sales_Amount Vs Quarter

- 2011Q3 (3 months) and 2011Q4 (2 months) - since we dropped December records.
- Maximum quantity of items was sold during 2011Q3 and very less quantity was sold during 2010Q4 (as there is just one month in that year)
- Maximum revenue was generated during 2011Q3 and 2011Q4 compared to all other Quarters.



Quantity and Total_Sales_Amount Vs Hour



- Most quantities are sold during 9AM to 4PM (Business Hours) with maximum no. sold around 12PM.
- And revenue generation during business hours is high as well.

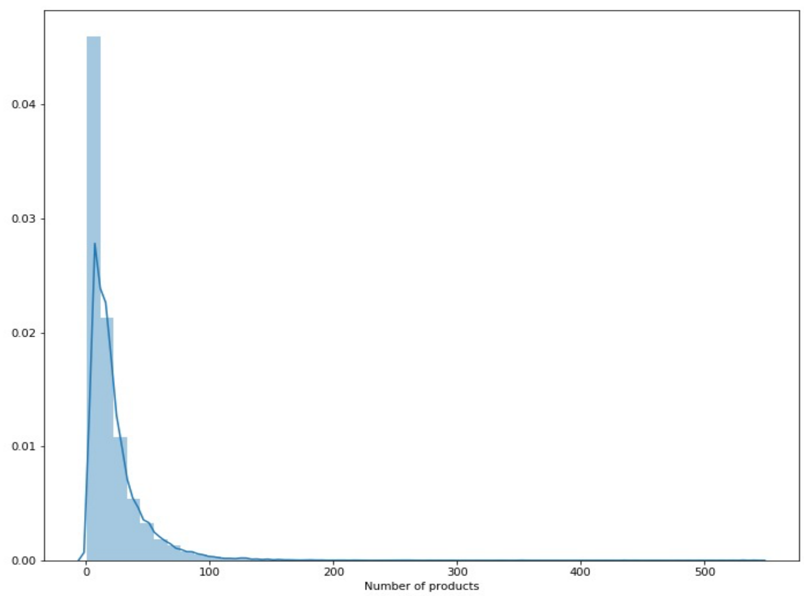
Multicollinearity

- In our model, we will be implementing K-Means clustering to identify the clusters in customer purchase patterns.
- K-Means is a distance-based algorithm which calculates Euclidean distance between data points.
- As a result, Collinearity between the variables will pose an issue with results as it will bring the data points closer together and reduce the Euclidean distance.
- As we can observe, there is not much of correlation between the variables except for Quantity and Total_Sales_Amount (0.83)
- Also, there are many categorical variables which will be converted into numerical features prior to K-Means model building (RFM metrics).

Distribution of variables

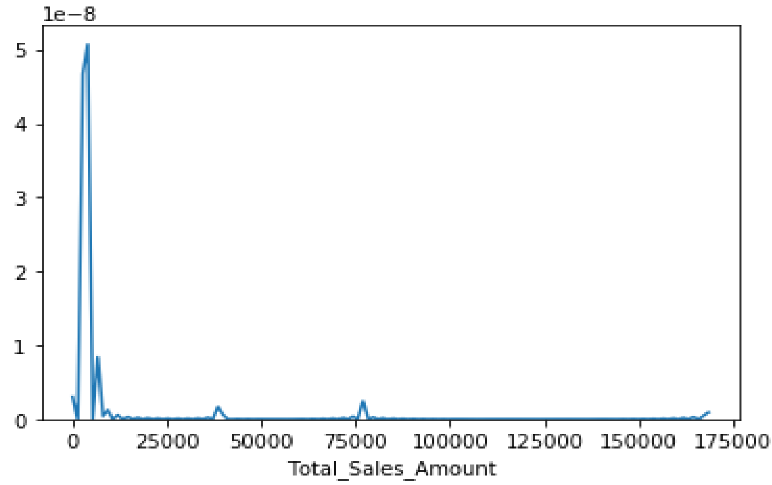
Skewness and Kurtosis

1. **Number of Products**



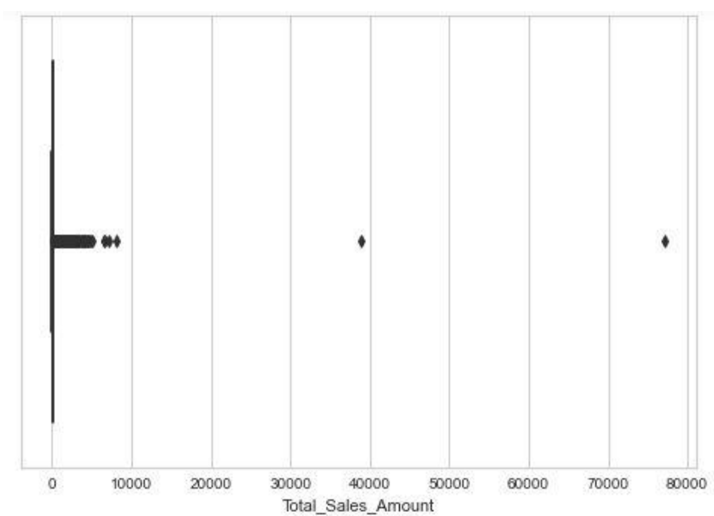
- We have a skewed distribution of products. Most people buy less than 25 items.
- Skewness = 5.042
- Kurtosis = 61.053
- We have right-skewed data with a sharp peak.

2. Total Sales Amount



- The total sale is right skewed which shows a positive sign for business, since there is an increase in the number of sales over the years.
- Majority of the customers purchase many products and price at lower ones.
- So that's why it causes right skewness.
- Only a few purchase high value products and also those with high values have less sales but still good for business as it produces profit anyways.

Presence of outliers and its treatment



- As we can observe, there are outliers in the Total Sales Amount.
- These outlier values need to be treated before building our cluster model.

6. Feature Engineering

6.1 RFM Analysis

RFM analysis (**Recency, Frequency, Monetary Value**) is a customer segmentation method which analyses customers' past purchase patterns in order to group the customers based on their behavioral patterns. This is a widely used model to identify the target customer base in order to design customer-centric marketing strategies.

- **RECENCY (R)**: Number of days since the last purchase was made.
- **FREQUENCY (F)**: Total number of purchases made.
- **MONETARY VALUE (M)**: Total money spent by each customer.

In our approach, we will be implementing K-Means clustering to identify clusters based on the above-mentioned metrics.

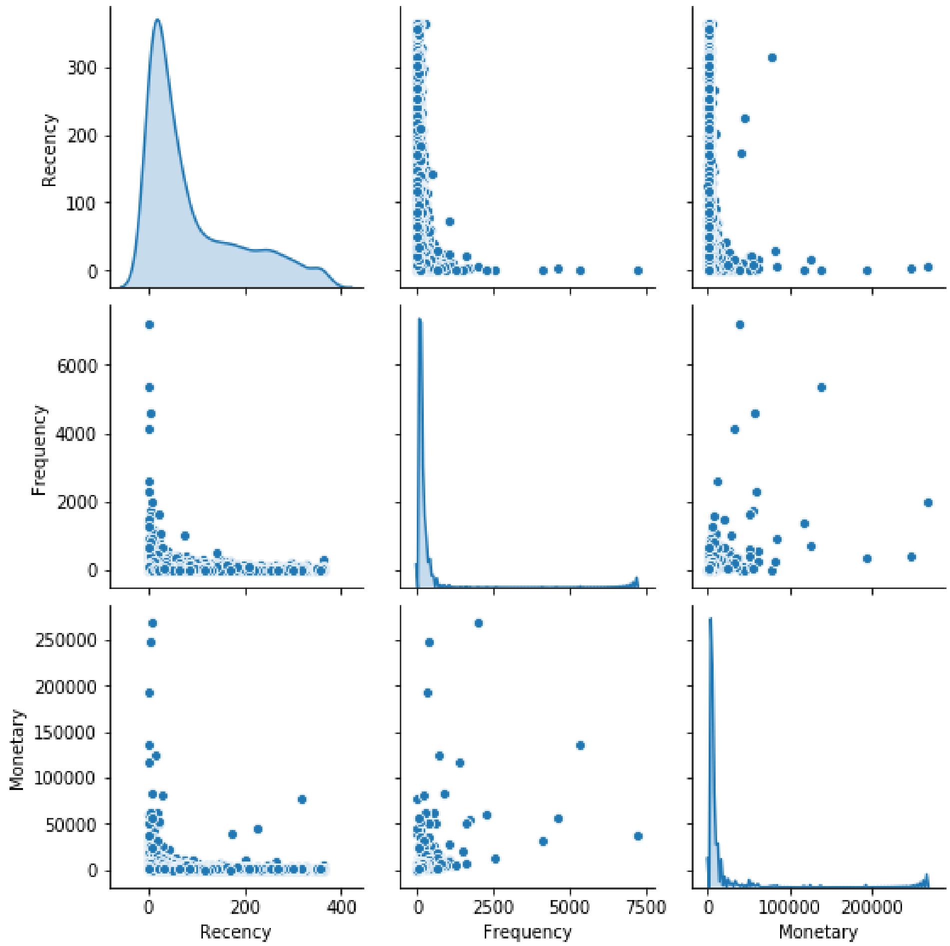
RFM Table:

CustomerID	Recency	Frequency	Monetary
12346	316	1	77183.60
12347	30	171	4085.18
12348	66	31	1797.24
12349	9	73	1757.55
12350	301	17	334.40

- Customer with ID = 12346 has recency: 316 days, frequency:1, and monetary: 77183,60 £.

Check for Outliers in the data:

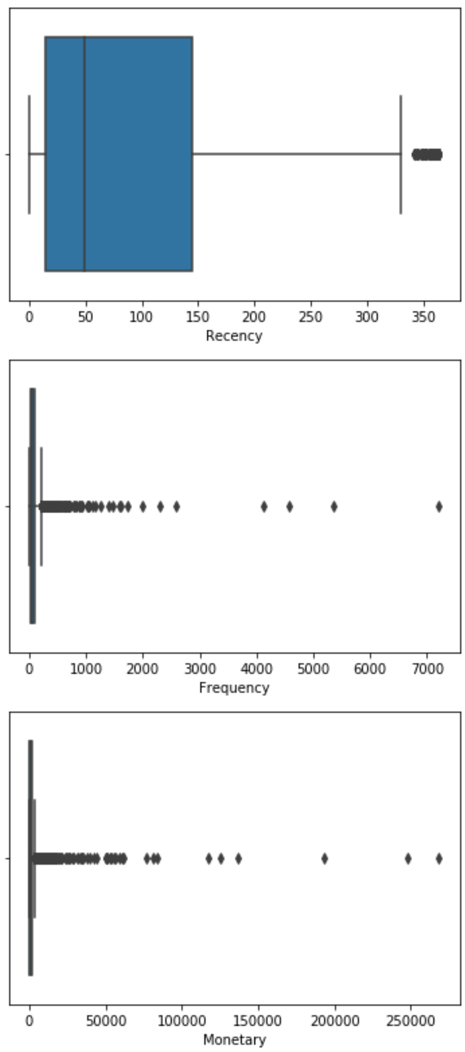
The below pair plot shows the relationship among the variables: Recency, Frequency and Monetary.



We can observe that the data is right-skewed. Therefore, we need to perform transformation in order to apply KMeans Clustering.

Boxplots to check for outliers:

We have illustrated the outliers in the data using Boxplots as shown below. We can observe that there are many outliers in Frequency and Monetary values. This is because we can find customers who have extreme values for those features.

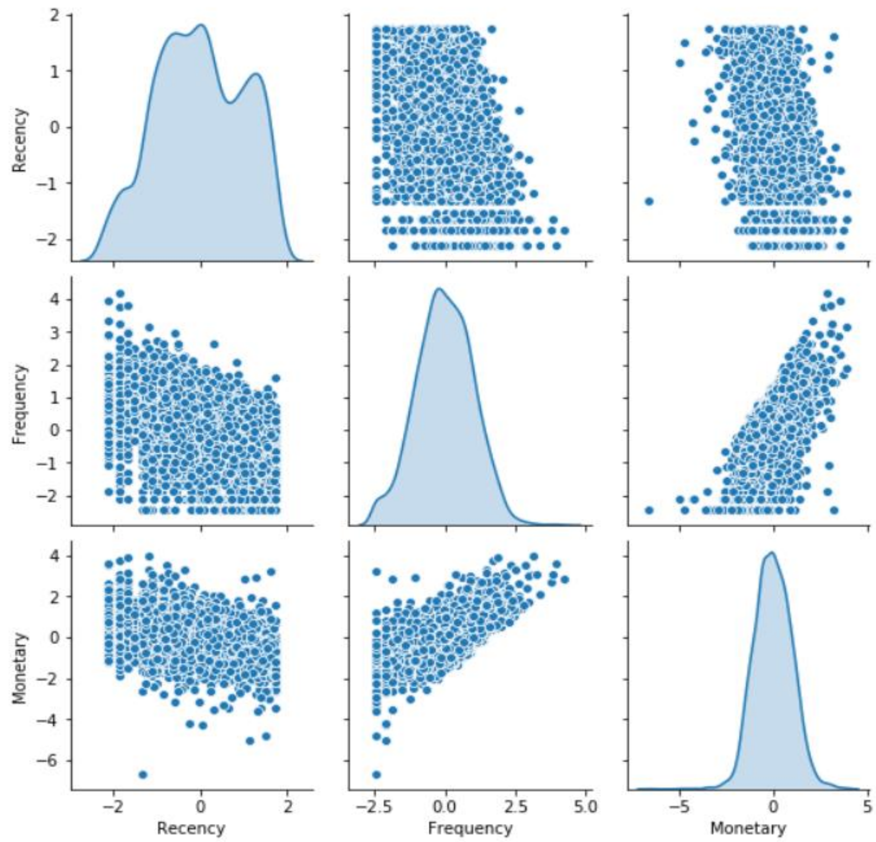


Outlier Treatment using Power Transformer:

We have used PowerTranformer( ) from sklearn.preprocessing package to transform the RFM data. PowerTransformer( ) provides non-linear transformations in which data is mapped to a normal distribution to stabilize variance and minimize skewness (treats heteroscedasticity). Outliers get transformed and distance between outliers reduces which in turn reduces skewness. One advantage of PowerTransformer( ) is that it takes the outliers into consideration and it will do scaling also so we don't have to perform scaling using MinMaxScaler( ) or StandardScaler( ) separately. Also, for our business case, outliers have crucial data (extreme values are needed) so we went for scaling instead of outlier removal.

RFM transformed data:

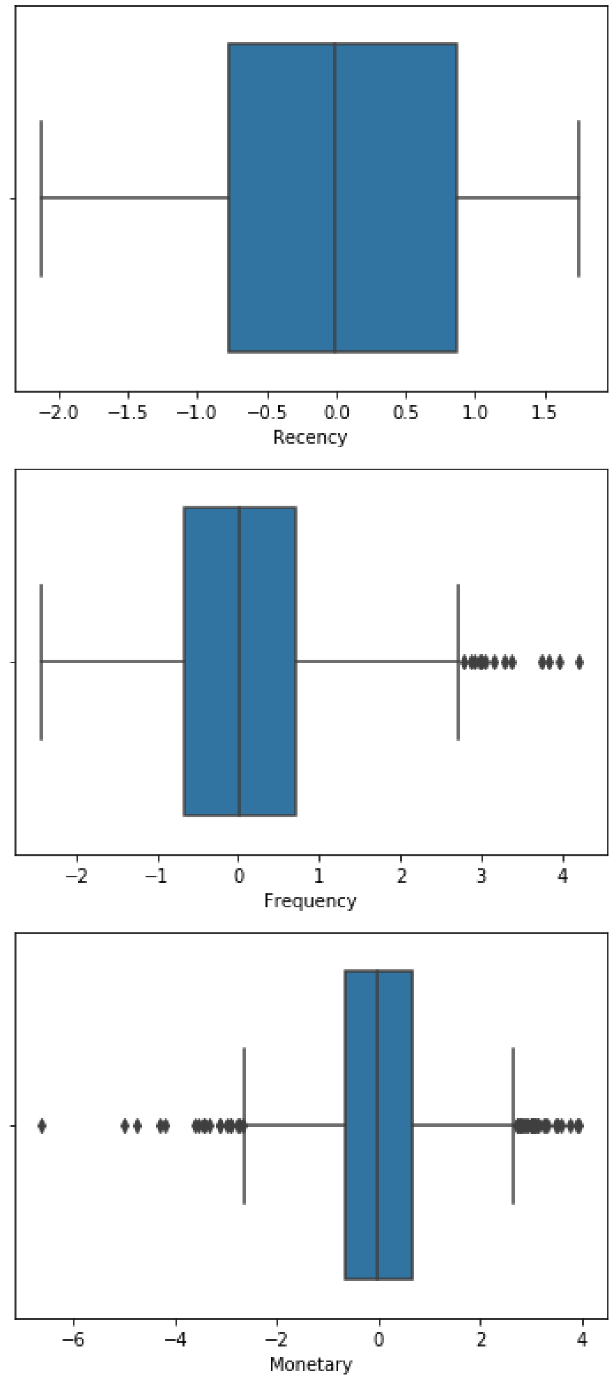
CustomerID	Recency	Frequency	Monetary
12346	1.603704	-2.433416	3.257663
12347	-0.348071	1.167772	1.369843
12348	0.215211	-0.195306	0.769547
12349	-1.051570	0.483645	0.752738
12350	1.554599	0.660685	0.571179



After performing transformation, the data has become close to normal. Skewness has reduced.

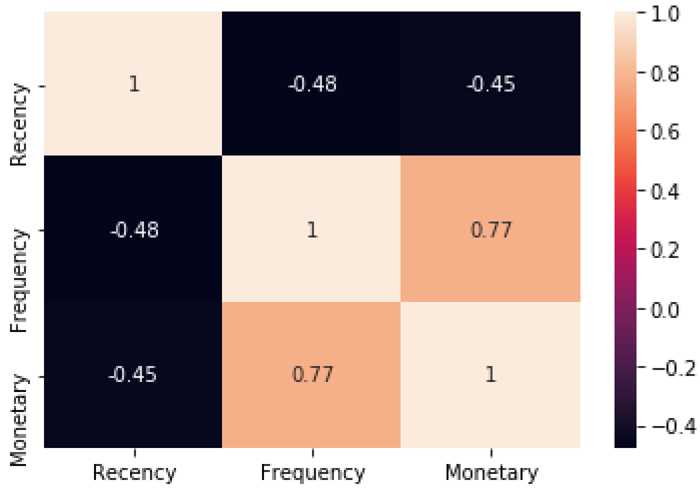
Boxplots after Transformation:

We can observe from the below boxplots that the outliers have now been transformed without much data loss. As we also need few extreme values in our data for real-time business models.



Checking Correlation among the RFM variables:

We have analyzed the correlation among the RFM variables by plotting a heatmap with correlation coefficients. The heatmap is illustrated in the below figure:



On one hand, we have a negative correlation between:

- Recency and Frequency
- Recency and Monetary

On the other hand, the correlation between **Monetary and Frequency** is positive compared to negative ones but still not that strong (<0.8).

7. K-Means Clustering

KMeans Algorithm:

KMeans is a clustering algorithm (unsupervised learning). Clustering generally works on the basis of two factors: Distance and Similarity.

We will be using Distance-Based KMeans Clustering in our model.

Distance-Based Clustering:

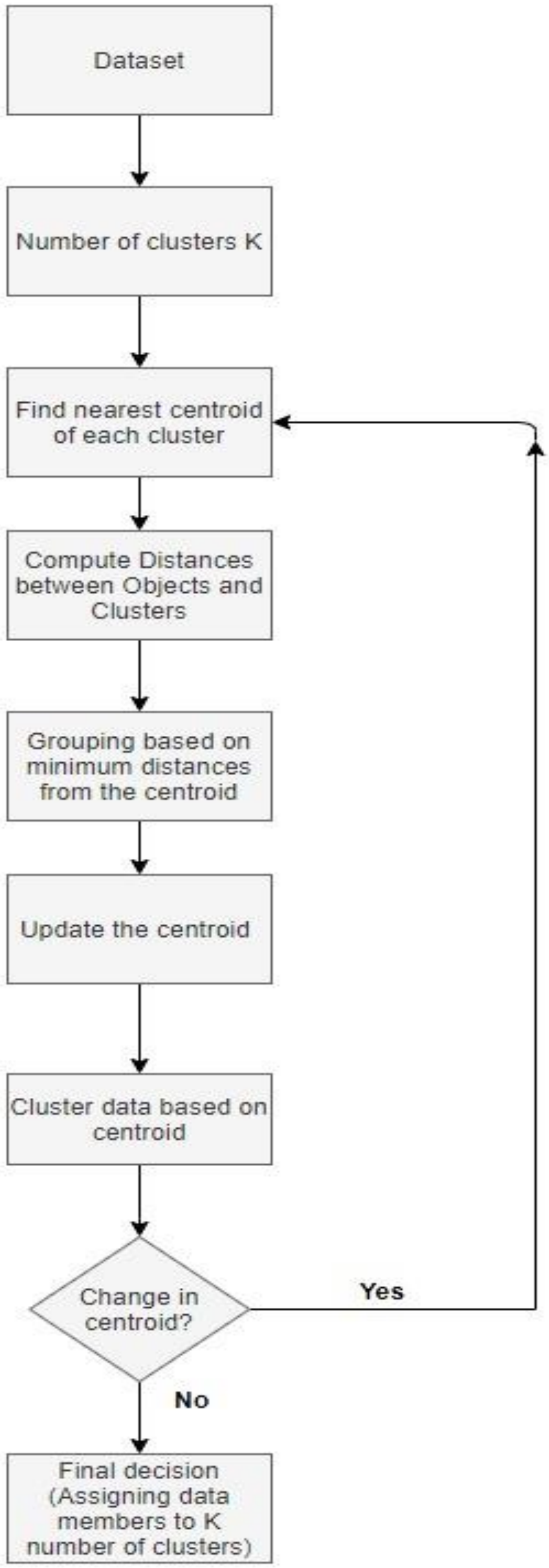
- If the distance between two observations is small, then they belong to the same cluster.
- If the distance is large, then they belong to different clusters.

KMeans is a type of Hard clustering technique. In Hard clustering technique, the probability that a sample belongs to a particular cluster is 1.

Steps Involved in KMeans Clustering:

1. In KMeans algorithm, initially we may not know the number of clusters needed for the problem statement and so we take random clusters and build the model. Without the assumptions of this cluster number, modelling is not possible.
2. KMeans also requires the random centroid values in the beginning of the algorithm. But eventually in the end, we get centroid values from KMeans after the model is built.
3. KMeans keeps calculating the distance and updates the centroids accordingly. This process is repeated until the consecutive centroids have the same values. But one challenge with this process in the bigger datasets, is that it will take a lot of time to converge.

The complete flowchart of steps involved in KMeans clustering is depicted below.


Dataset
Number of clusters K
Find nearest centroid of each cluster
Compute Distances between Objects and Clusters
Grouping based on minimum distances from the centroid
Update the centroid
Cluster data based on centroid
Change in centroid?
Yes
No
Final decision (Assigning data members to K number of clusters)

Finding Optimum value of clusters, K:

There are two methods in finding optimal cluster value. They are as follows:

1. **Elbow Curve**

- Elbow curve can be plotted by using the number of clusters against the inertia value (WCSSE - Within Cluster Sum of Square Error)
- Rate of change of error is rapid only up to a point and after that point the drop in error is not rapid as before. That point is called the elbow point.
- If we get a proper elbow point and do the cluster based on that 'k' value, then we would get proper discriminate characteristics in the cluster.

2. **Silhouette Score**

- Silhouette score is a statistical method to validate the quality of the clustering model. Higher the silhouette score better the model.
- The score is calculated for each cluster.

Based on the Elbow curve we can validate if the clusters where we found the elbow point have the highest Silhouette score. If both are satisfied, then we would take that particular value of K as the optimal value and build the final model.

Assumptions for Clustering Algorithms

- Data should be scaled (Mean=0, Standard Deviation = 1) to make the distribution more normal.
- Number of centroids have to be initialized at the beginning and further build models for different values of centroids to find the optimum value which results in low cluster inertia and Silhouette score.
- Size of clusters assumed helps in making decisions on the cluster boundaries, which in turn helps in calculating the number of data points within each cluster.

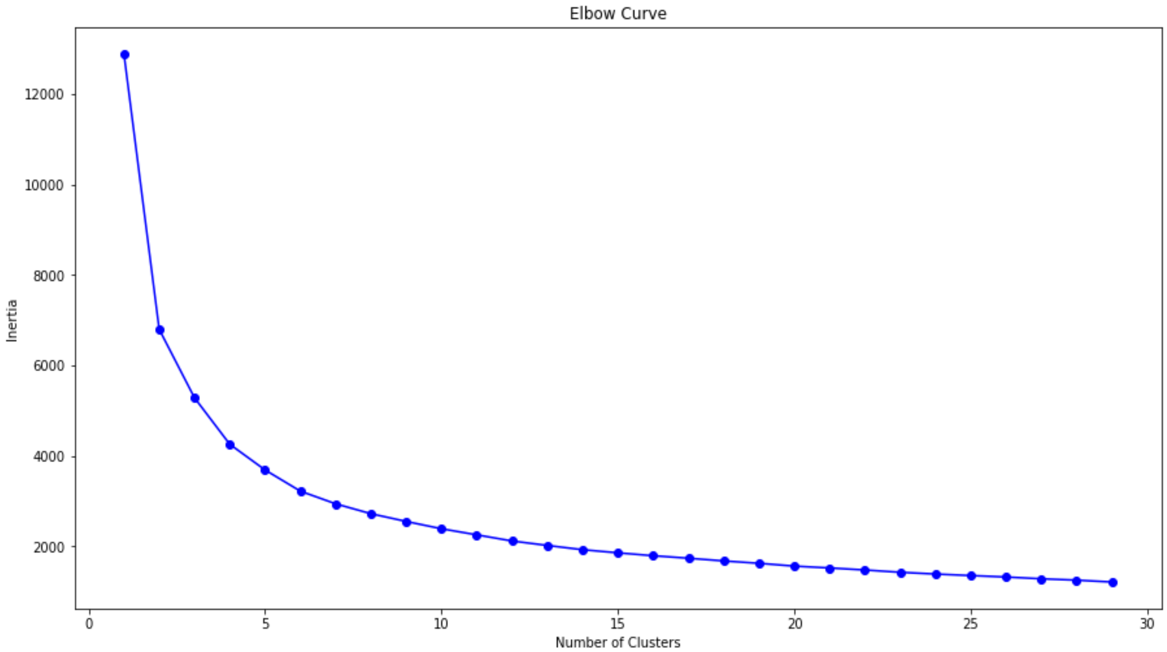
K-Means clustering on RFM variables

- We applied K-Means clustering on the RFM variables

WCSSE Score:

Value of k	WCSSE Score
1	12894.00000
2	6798.017695
3	5285.741912
4	4257.982589
5	3690.258186
6	**3222.055880**
7	2942.095458
8	2722.983243
9	2552.144943
10	2390.022240

Elbow Curve:



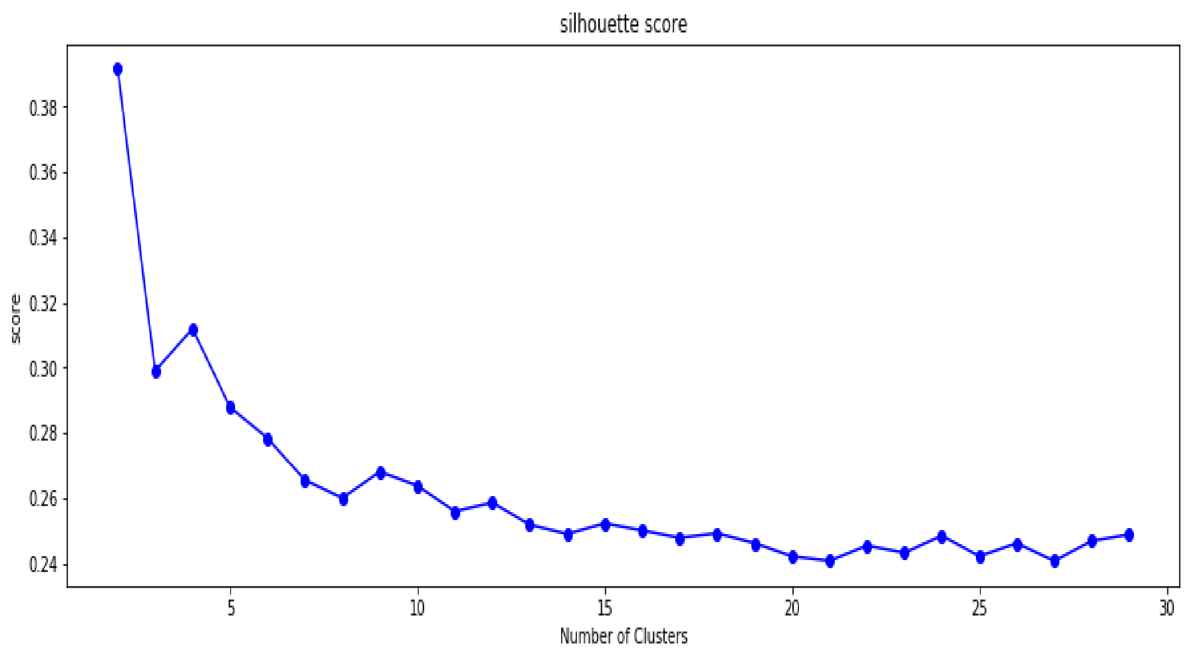
- As we can observe, there is a dip in Inertia at K=6. Therefore, we can take the optimum value of **K=6** for initial K-Means modelling.

We can check with Silhouette Analysis as well.

Silhouette Score Analysis:

- The silhouette coefficient can vary between -1 and +1. A coefficient close to +1 means that the instance is well inside its own cluster and far from other clusters, while a coefficient close to 0 means that it is close to a cluster boundary.
- Coefficient close to -1 means that instance may have been assigned to the wrong cluster.
- We have plotted Silhouette Plot for different values of K (from 2 to 30)
- We observed that K=6 was the optimum value.
- Below are the graphs for K=6.

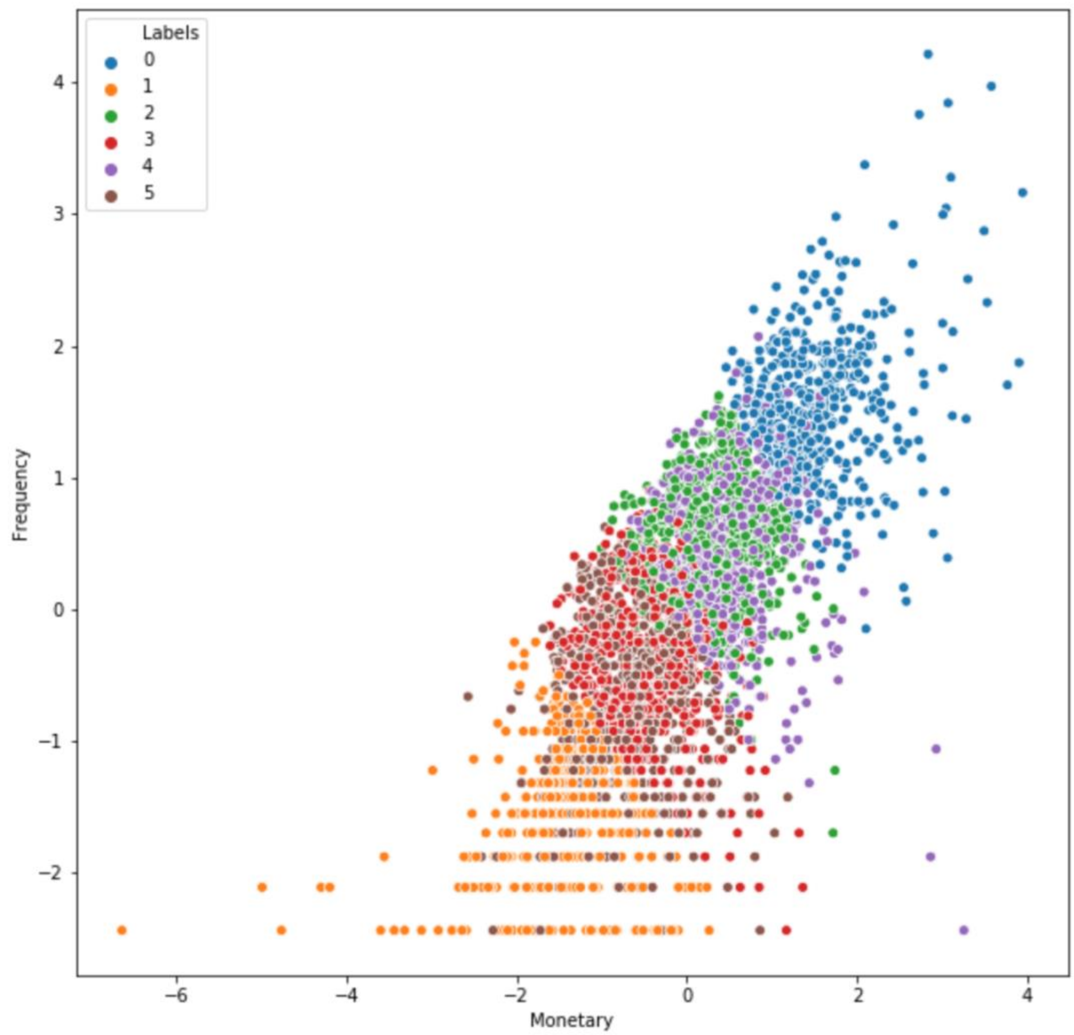
Value of k	Silhouette Score
2	0.391593
3	0.299047
4	0.311823
5	0.287876
6	**0.278471**
7	0.265460
8	0.260023
9	0.268064
10	0.263765



The above plot confirms that the optimal number of clusters, **K = 6**.

The below table depicts the first 5 rows with the cluster labels we got after implementing KMeans clustering model with optimum value of K=6.

Recency	Frequency	Monetary	Labels
1.603704	-2.433416	3.257660	4
-0.348071	1.167772	1.369843	0
0.215211	-0.195306	0.769547	4
-1.051570	0.483645	0.752738	2
1.554599	-0.660685	-0.571179	3



From the above scatterplot, we can observe that the features: Frequency and Monetary are having high correlation between them, this could lead to Multicollinearity effect. KMeans will not function well in the presence of Multicollinearity which can be confirmed from the overlapping of clusters from the above graph. Therefore, it is advisable to perform Principal Component Analysis (PCA) to reduce the effect of Multicollinearity.

Principal Component Analysis (PCA)

PCA is an unsupervised non-parametric learning technique which is usually used for dimensionality reduction. Primary problem with high dimensionality in the dataset is that it causes overfitting of the model.

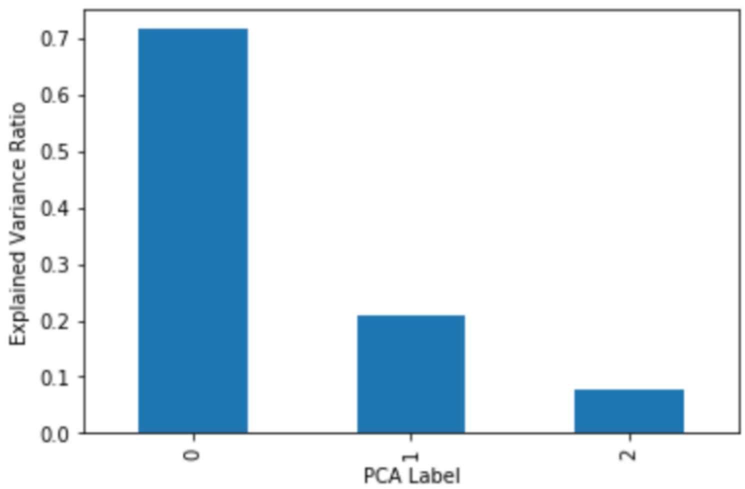
In our dataset, we can observe multicollinearity among the features: Frequency and Monetary. Therefore, we can perform PCA to deal with the multicollinearity problem.

Working Principle of PCA:

- In PCA technique, instead of dropping correlated dimensions, we engineer a new composite dimension to represent the original features and replace them with new derived features.
- PCA makes the data to be centered around the origins by standardizing the data and the data on all the dimensions are subtracted from their means to shift the data points to the origin.
- PCA generates the covariance matrix / correlation matrix for all the dimensions.
- PCA computes the eigen vectors which are called principal components and the corresponding eigen values which are the magnitude of variance captured.
- Sort the eigen values in descending order and select the one with the largest value. The first principal component covers most of the data's original characteristics/features.

PCA in general is affected by the outliers so necessary treatments need to be performed to make it normal before building the model. The interpretation of the PCA model will get more complex when the number of features increases.

In the graph below, we can observe that the first two PCA components (PC1 and PC2) can explain about the majority of the data's features which can be calculated using inbuilt explained variance ratio. Based on our business model and the requirements, we can take the first two components which explain nearly 90% of the variance ratio.

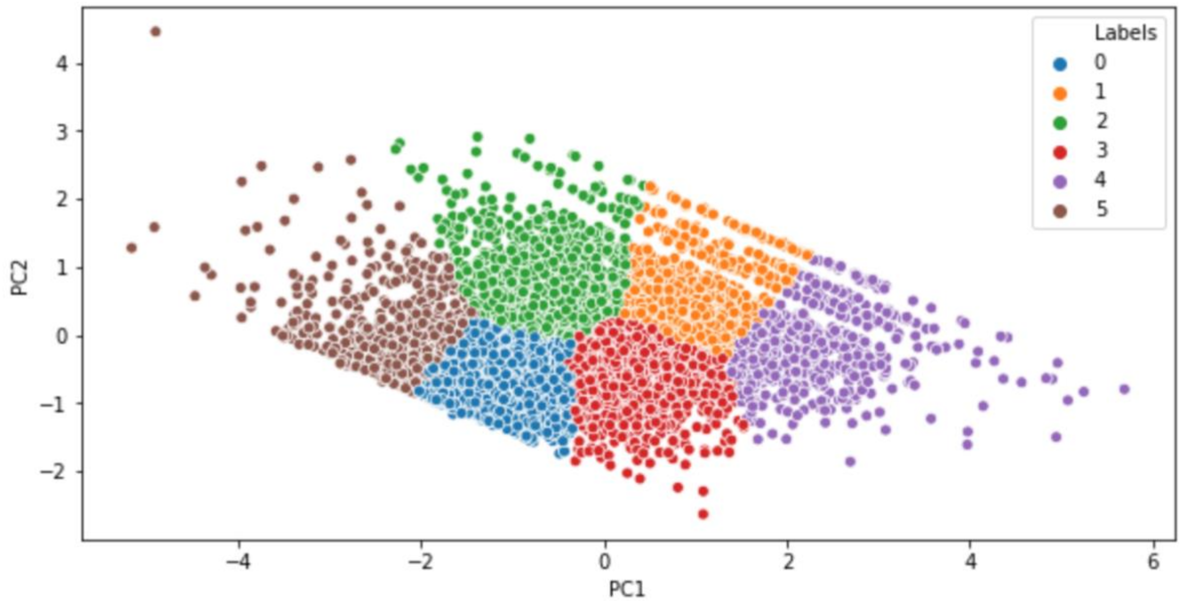


KMeans Clustering using PCA data:

The table below depicts the first 5 rows of the cluster labels we got after implementing KMeans clustering model with optimum value of K=6 using PCA transformed data.

PC1	PC2	PC3	Labels
-0.309292	-1.846721	3.949644	3
1.729960	-0.593668	0.126930	4
0.242446	-0.416260	0.666791	3
1.279800	0.470820	0.216261	1
-1.527807	-0.919250	0.017722	0

The figure below shows the distribution of clusters formed using KMeans clustering on PCA data. We can observe a clear differentiation among the cluster characteristics. The segments are separated from each other without any overlap.



WCSSE and Silhouette Score for KMeans with PCA:

	K=6
WCSSE	3221.89
Silhouette Score	0.278471

8. Validation of Model

The Model Quality can’t be validated using Clustering methods. We need to use classification models for checking our cluster quality. There are different metrics from scikit-learn that we can use If we need to validate the model.

Random Forest Classifier

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. The decision at the end is taken from multiple opinions in Ensemble methods.

Random forest is a supervised learning ensemble algorithm. In the Decision Tree, the decision is taken only from one tree which can cause overfitting and bias but in Random Forest multiple trees are constructed and from that the results are classified based on 'vote'. Based on the vote, the data which belongs to the majority vote class will be classified into that particular class. Random Forest can be used for both Regression and Classification. Random Forest efficiently runs on larger datasets. It has the ability to deal with outliers’ data in classification models. It has an effective method for estimating missing data and maintains accuracy even when some proportions of the data are missing.

Random Forest uses the following two methods. They are as follows:

1. Row Sampling (Bootstrapping):

As the name suggests the rows are only taken in random but all the columns are considered for this sampling process. Here we take a sample out of the data. For example, taking a sample of 60 rows and using these 60 samples we construct multiple decision trees (5 Decision Trees) as this is a random forest.

The test input is applied on all the decision trees. It has to be noted that all the trees are independent of each other in the Random Forest. The majority class of the result is taken into account and the test input is classified accordingly. If there are 3 results pointing to class 0 and 2 results points to class 1 then the input test data is classified into the class 0. This process is also called the Bagging Algorithm.

2. Column Sampling:

One change from the above process is that both rows and columns are considered and selected in random. The randomness in this is why it's called Random Forest. The row and column samples which were formed in random are called Bags. The row samples and columns samples for the random forest can be controlled in the hyper parameters using the row samples, column samples. The Bag is also considered as a metrics score for model validation.

Example from a 100 data a sample of 60 is taken and a model is constructed. The 60 samples are one bag. 40 samples which were selected is called out of bag. The remaining 40 samples are taken into account as testing data and is applied to the model which was constructed using 60 samples. If the OOB (Out of Back) score is good then the precision, recall scores will be good.

We implemented Random Forest Classifier in order to validate our KMeans cluster model.

The table below shows the results that we got in terms of accuracy score for both with and without PCA models.

	With PCA	**Without PCA**
Accuracy Score	0.975%	0.963%

Reasons for choosing Random Forest Classifier:

- This method has hyperparameters of Decision Tree Classifier and Bagging Classifier techniques in order to control the tree's growth as well as ensemble respectively.
- This method also avoids overfitting of the model but selecting best features out of random subsets of features thereby increasing the diversity.
- Allows higher bias and lower variance.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised machine learning model which can be used for both Regression and Classification. In our case, we have used SVM for classification.

The main aim of using SVM is to create more discriminate clusters as the hyperplanes in the SVM model can give more discriminate values. The data points on the dotted hyperplane are called the support vectors.

The distance between the two dotted hyperplanes is called the marginal distance. Higher the distance between the margins, more discriminate the clusters. When we have a new observation coming into the model, if the data point is above the original hyperplane it classifies them into the positive side's label, else if it lies below the original hyperplane it will be classified as -ve side's label.

SVM works in a way that it chooses the best marginal plane which has the highest Margin distance. The main aim is to make sure the margin is as large as possible. SVM allows some misclassification to happen to avoid the overfitting problem. If the number of support vectors in the hyperplanes are large in number, then the model will reduce the overfitting.

There are two types of model clusters which need to be considered.

1. Linearly Separable
2. Non-Linearly separable.

If the clusters can be separated by linear method, then it is linearly separable but in the case of non-linear, data will be non-linearly spaced and cannot be split.

In non-linear cases, we need to convert the data from lower dimensions to higher dimensions. For example, from 2-D data to 3-D data.

We implemented Support Vector Machine - Classifier in order to validate our KMeans cluster model in order to reduce overfitting problems.

The table below shows the results that we got in terms of accuracy score for both with and without PCA models.

	With PCA	**Without PCA**
Accuracy Score	0.865%	0.932%

9. Customer Segmentation from KMeans Clustering

The table below shows the segments of customers based on their purchase behavior obtained from KMeans Clustering technique. We can analyze these customer segments in order to identify the best marketing strategy to improve business and strengthen the customer-business relationship.

CustomerID	Recency	Frequency	Monetary	Labels
12346	316.0	1.0	77183.60	0
12347	30.0	171.0	4085.18	4
12348	66.0	31.0	1797.24	0
12349	9.0	73.0	1757.55	5
12350	301.0	78.0	334.40	2

We are going to segment the customers based on three values:

1. **Mean** value of Recency, Frequency and Monetary
2. **Median** value of Recency, Frequency and Monetary
3. **Month-wise Total Sales Amount** for each cluster

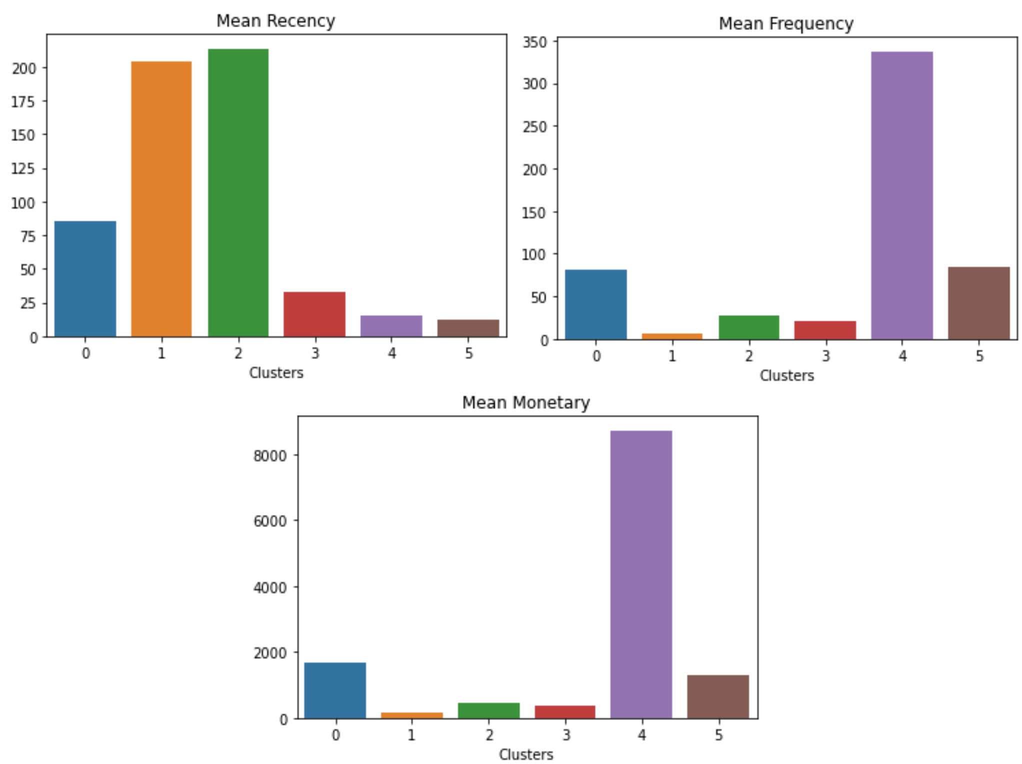
Comparison of Overall Mean values of Recency, Frequency and Monetary with individual Mean value of Clusters:

	Recency	Frequency	Monetary
Overall Mean	90.461610	87.414146	1947421697

	Mean Recency	Mean Frequency	Mean Monetary
Cluster 0	85.848837	80.52907	1693.73555
Cluster 1	204.439024	6.595528	158.79061
Cluster 2	213.644025	27.354717	462.000516
Cluster 3	32.719018	20.878528	372.35096
Cluster 4	15.196995	337.295492	8720.93307
Cluster 5	11.90502	84.671642	1276.37117

Mean RFM values:

The bar graphs shown below illustrates the Mean Value of Recency, Frequency and Monetary for each Customer Segment.



Since, there are outliers in the RFM variables, we also observed the median values of RFM to study the variation in customer behavior.

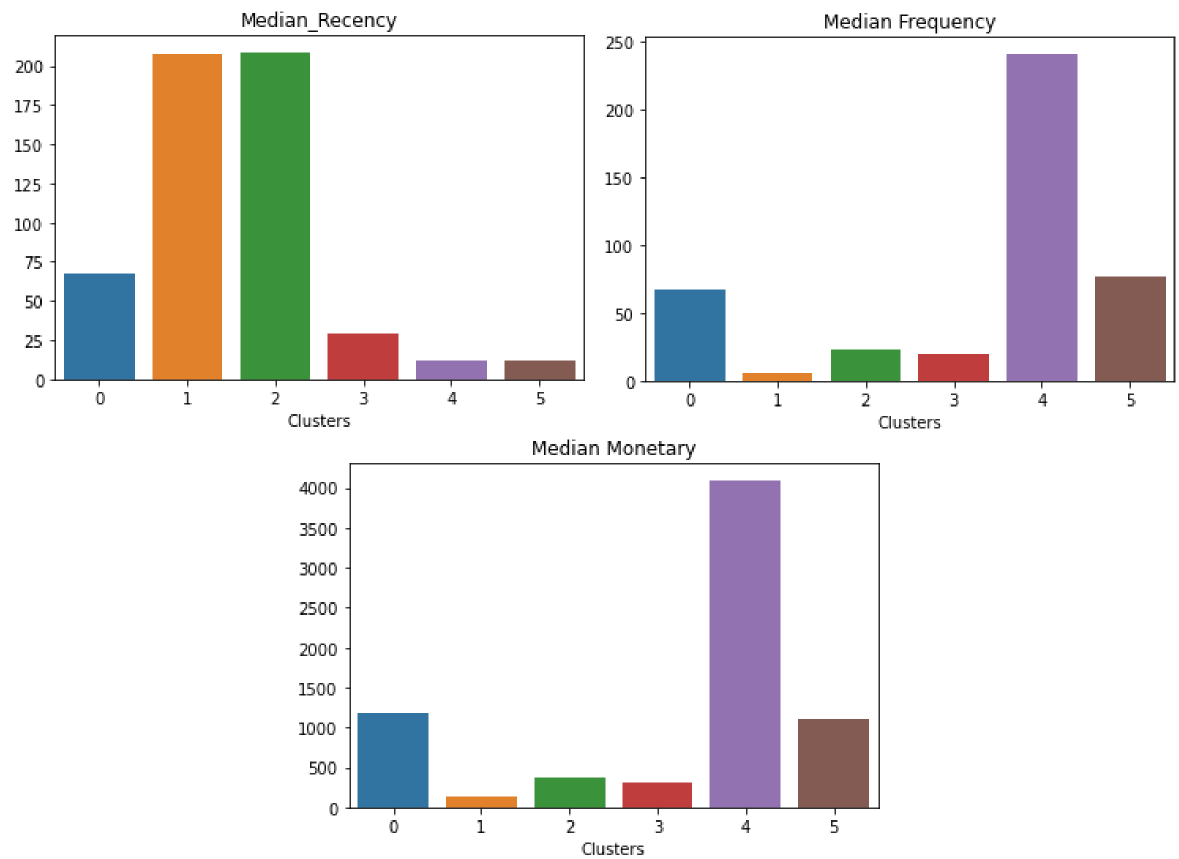
Comparison of Overall Median values of Recency, Frequency and Monetary with individual Median value of Clusters:

	Recency	Frequency	Monetary
Overall Median	49	40	650.59

	Median Recency	Median Frequency	Median Monetary
Cluster 0	67	67	1182.465
Cluster 1	208	6	134.625
Cluster 2	209	23	379.35
Cluster 3	29	19	320.46
Cluster 4	12	241	4097.37
Cluster 5	12	77	1105.78

Median RFM values:

The bar graphs shown below illustrates the Median Value of Recency, Frequency and Monetary for each Customer Segment.



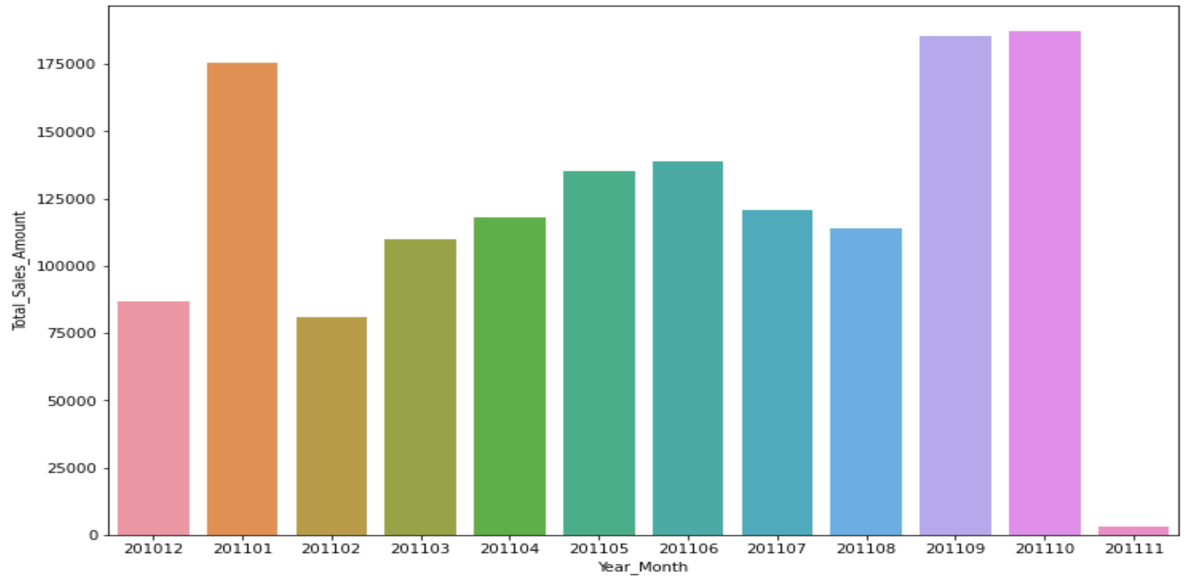
From the above bar plots, we can clearly observe a difference in Frequency and Monetary values of each cluster as there were many outliers in those variables. Therefore, we compared the median RFM values of each cluster with the overall median value to segment them into different types.

Month-wise Total Sales Amount for each cluster

The following are the characteristics/purchase behaviors of each Cluster:

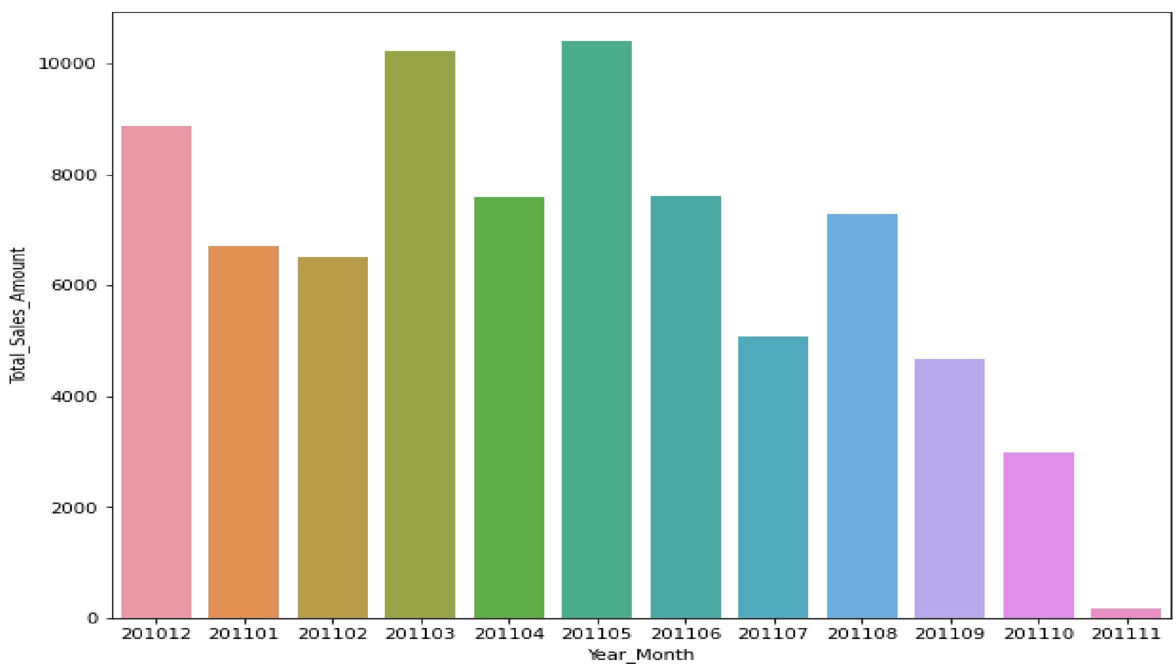
Cluster 0:

- Total 860 customers.
- They are the **Average-valued Customers** - Moderate Recency, Moderate Frequency and Moderate Monetary values.
- These customers have the potential for becoming regular customers.
- Efforts should be made to turn them into regular customers.



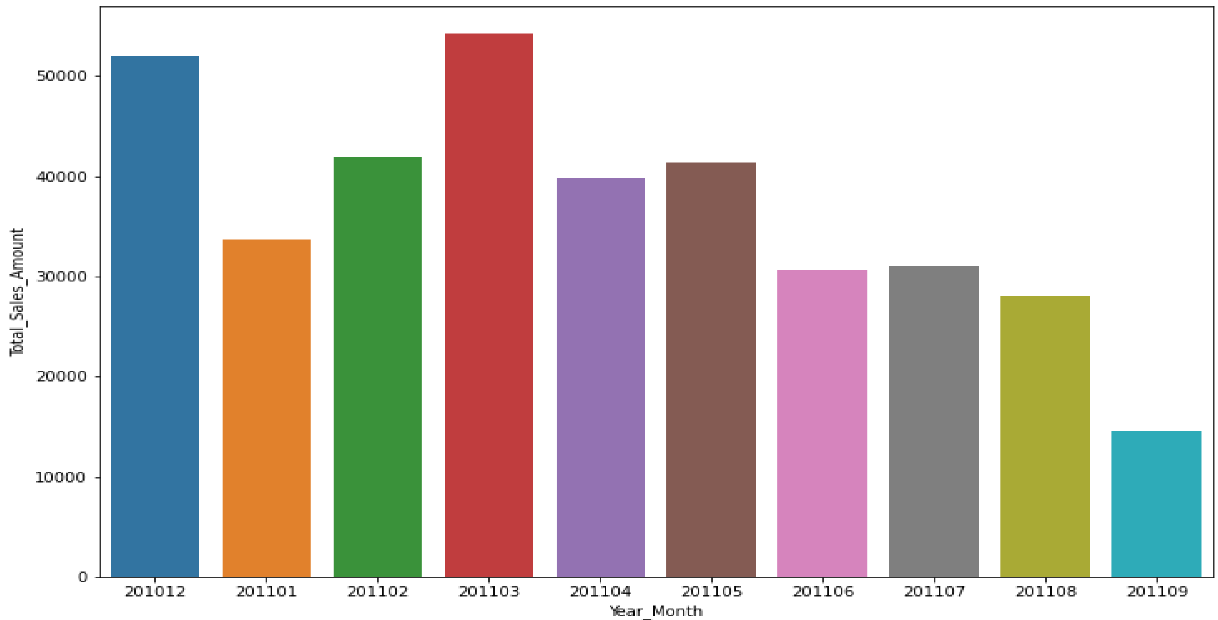
Cluster 1:

- Total 492 customers.
- They are the **Customers at Risk** having High Recency, Low Frequency and low Monetary values.
- We should focus on these customers to make them Loyal to our business.



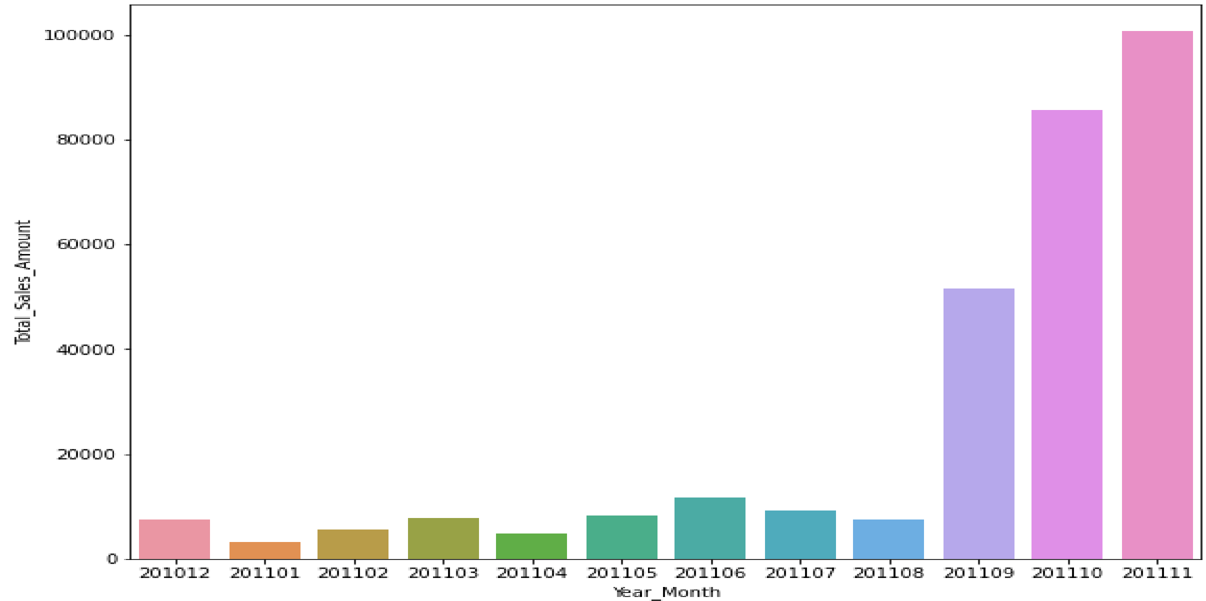
Cluster 2:

- Total 795 customers.
- They are the **Lost Customers** having High Recency, Low Frequency, Low Monetary values.
- These customers have not purchased any products for the past two months.
- These are the customers who are no longer in business with us hence we need to strategize marketing plans to make them buy again.



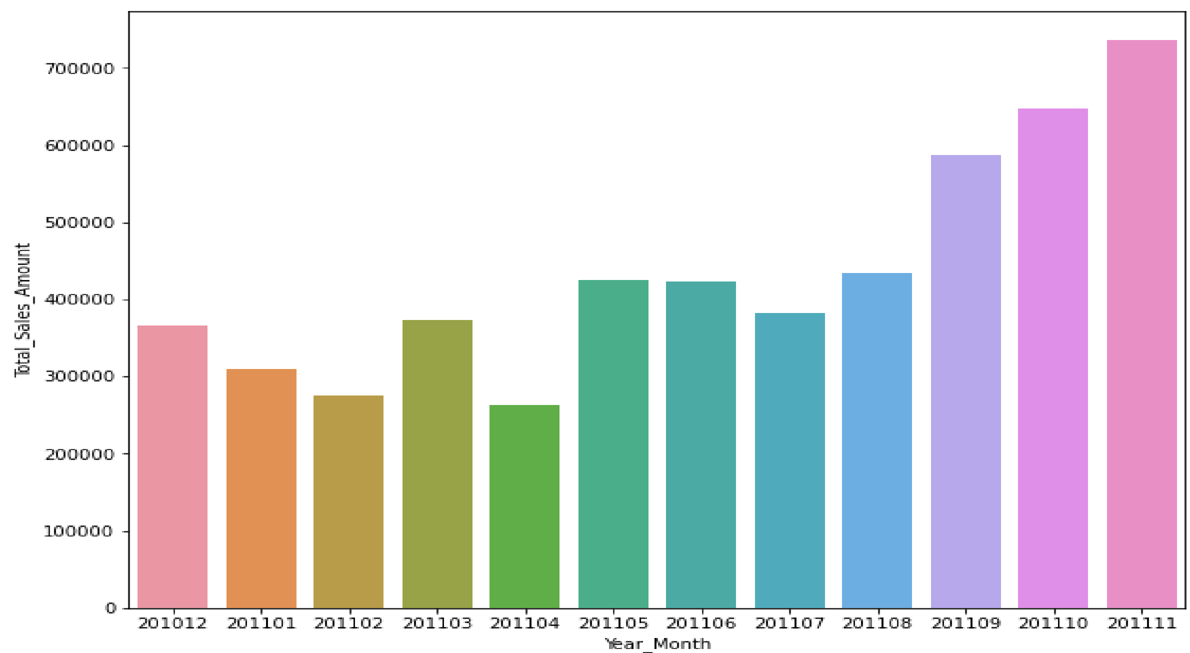
Cluster 3:

- Total 815 customers.
- They are the **Recent Active Buyers** having Low Recency and Low Frequency with Low Monetary values.
- These are the customers who have started to buy recently, and we need to focus on them to make them regular buyers.



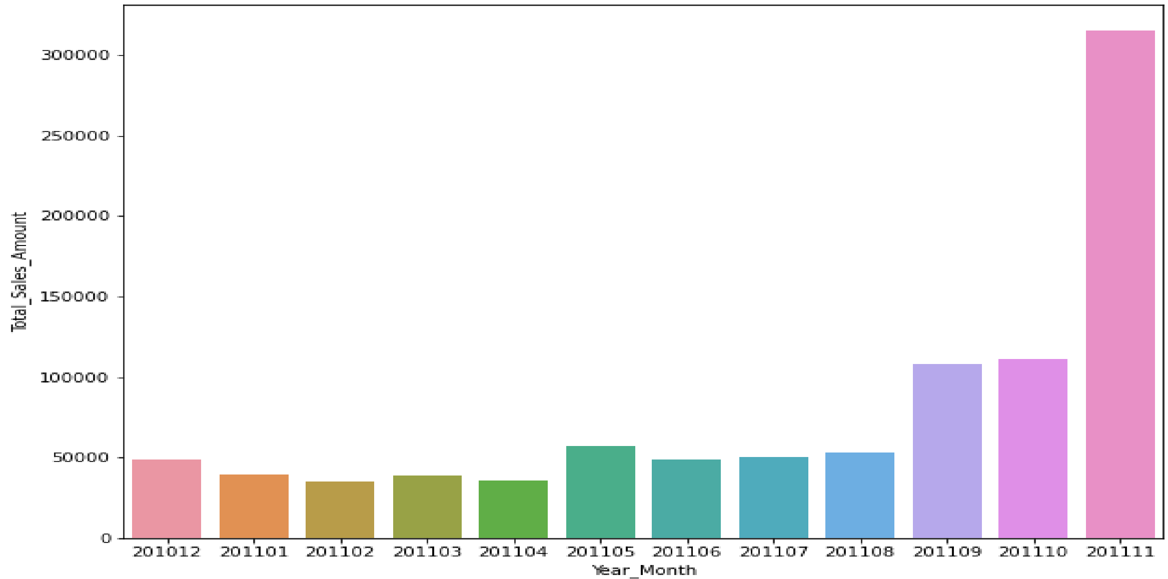
Cluster 4:

- Total 599 customers.
- They are the **High-valued Frequent Buyers** having Low Recency, High Frequency and High Monetary value.
- We should try to retain these customers and make them loyal to our business.



Cluster 5:

- Total 737 customers.
- They are the **Regular Customers** having Low Recency, Moderate Frequency and Moderate Monetary values.
- We should focus on these customers to make them purchase more as they are steady customers and are good for business.



Assigning Segment Type to each Cluster:

Customer Segmentation

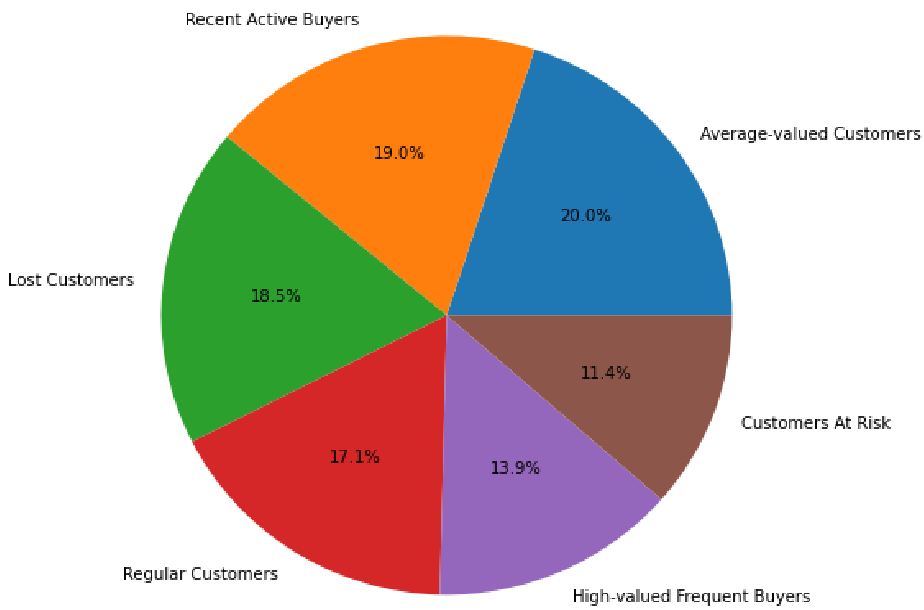
Labels	Customer Segment	Number of Customers
0	Average-valued Customers	860
1	Customers at risk	492
2	Lost Customers	795
3	Recent Active Buyers	815
4	High-valued Frequent Buyers	599
5	Regular Customers	737

10. Business Recommendations

Revenue Analysis for each Customer Segment

We have analyzed the revenue contribution from each customer segment and observed that each segment contributes almost equally in terms of revenue. Therefore, it is important to focus on each customer segment for improvement in sales.

The pie chart shown below illustrates the percentage of contribution from each customer segment towards revenue.



Customer Segment	% of Revenue Contribution
Average-valued customers	20.0%
Customers at Risk	11.4%
Lost Customers	18.5%
Recent Active Buyers	19.0%
High-valued Frequent Buyers	13.9%
Regular Customers	17.1%

- Customers at Risk contribute to nearly 11.5% of revenue which is lowest among the segment types.
- Average-valued Customers contribute to nearly 20.0% towards the revenue.
- The revenue lost through churned customers is nearly 18.5% which is alarming.

This seems to be good for business as the maximum revenue is coming from steady customers and only a comparatively less percentage of revenue is at risk. On the other hand, we should also focus on the high (18.5%) of revenue lost over the years by persuading the lost customers to buy again.

Business Recommendations:

Average-valued Customers

- Out of 860 total lost customers, we have 752 customers from the UK.
- Average-valued Customers spend on almost all months except in November when their purchases are low.
- These types of customers show sudden spikes in purchasing behaviors or drop in purchases occasionally.
- They tend to stock up the products in advance for seasonal sales especially in the months of January, September and October so that they can make a good return on Investment (ROI).
- They are usually frugal spenders. Therefore, we can design a mix of **Emotion-driven and metric-driven marketing campaigns** such that we should try to persuade them by messaging the upcoming offers, convince them of right choices by providing them with product descriptions and negotiable rates through emotional messages.
- Through this we can win over their loyalty for a longer time.

Customers at Risk

- Out of 492 total Lost customers, we have 463 customers from the UK.
- We need to identify these customers well in advance before they churn by monitoring their purchase activity regularly.
- If their spending frequency is at an alarming rate, we need to send them personalized messages and provide them with customer support.
- We can design a **Personalized Customer Engagement Campaign** to deepen the relationship with such customers and regularly monitor their purchase behaviors.

Lost Customers

- Out of the total 795 customers, there are 723 Low-valued customers from the UK.
- This segment of customers are the ones we have lost over the years.
- According to a study by Marketing Metrics, there is only a 20 to 40 percent chance of winning back an ex-customer.
- One of the best strategies to win back lost customers is by offering a new deal.
- From the graph shown below, we can observe that customers in this segment don't usually buy in months of “Off seasons”.
- This could be because no discounts will be available during those months.
- So, in order to attract them we can introduce **End of Season Sale campaigns**.

Recent Active Buyers

- Out of the total 815 customers, there are 761 Recent Active Buyers from the UK.
- We can observe a drastic increase in purchases from these customers in the last three months of the year from the bar graph shown below.
- These customers are making purchases during the holiday seasons when discounts are high.
- We need to focus on how to make them purchase more during off seasons.

- We can plan a **Replenishment Marketing Campaigns** for these customers.
    - Remind the customer several days before the season begins.
    - Provide all product details and offers up front.
    - We can make them regular subscribers for the seasons.
    - If needed, follow-up with an incentive.

High-valued Frequent Buyers

- Out of the total 599 customers, there are 514 High-valued Frequent Buyers from the UK.
- High-valued frequent buyers spend on almost all months and bring in maximum revenue to the business.
- we should try to retain these customers and make them loyal to our business.
- We can design a **Loyalty Marketing Campaign** such that we should try to provide them with maximum reward benefits and encourage them to continue buying from us.
- They are the most important segment of customers who contribute to the business growth. Therefore, we need to provide them with rewards and improve their lifetime value and also to strengthen their relationship with the business.

Regular Customers

- Out of the total 737 customers, there are 674 Regular Customers from the UK
- We can observe a gradual increase in purchases from these customers.
- We need to focus on how to make them purchase more as they are steady customers and are good for business.
- We can analyze purchase behavior (most bought products) to identify valuable insights about what the customers are looking for, then deliver personalized recommendations right after they make their purchase in future.
- This is known as **Dynamic Product Recommendation** through cross-selling campaigns.

11. Summary and Future Works

1. In our study, we built a model using K-Means clustering technique and identified 6 clusters based on the purchase patterns and then segmented the customers.
2. Principal Component Analysis (PCA) was performed in order to remove multicollinearity effect and clustering was done.
3. The model was validated by implementing the Random Forest as well as Support Vector Machine (SVM) Classifier.
4. The summary of our findings are given in the table below:

Cluster Model Summary:

	Inertia	Silhouette Score	Number of Clusters, K
KMeans Clustering (Without PCA)	**3222.0558**	**0.2784**	**6**
KMeans Clustering (With PCA)	**3221.8848**	**0.2785**	**6**

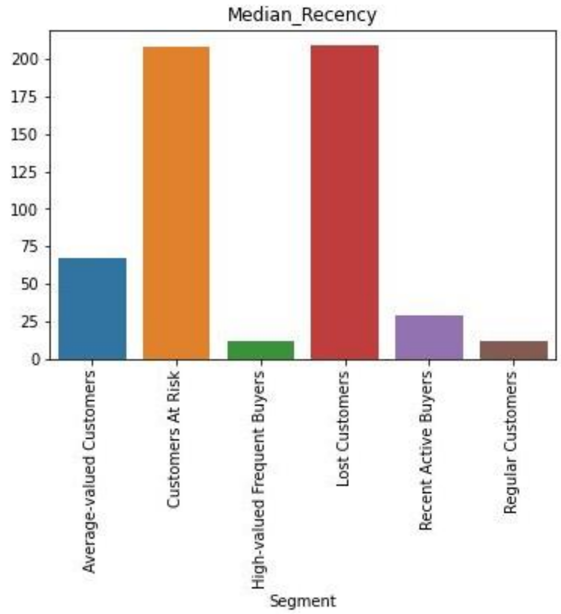
Classification Models Summary:

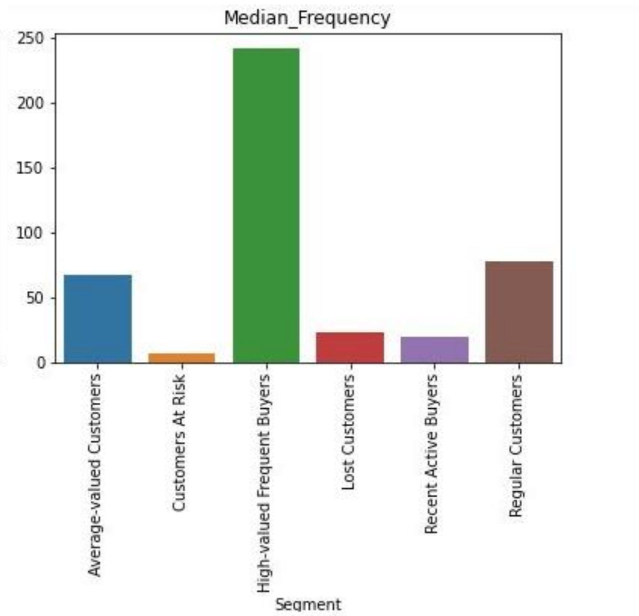
Random Forest Classifier	**With PCA**	**Without PCA**
Accuracy Score	**0.975%**	**0.963%**
Support Vector Machine - Classifier	**With PCA**	**Without PCA**
Accuracy Score	**0.865%**	**0.932%**

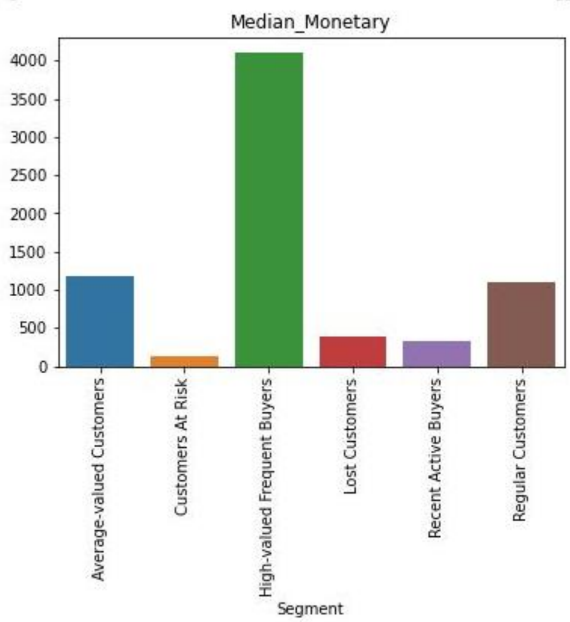
Customer Purchase Intentions and Marketing Suggestions:

Customer Segment	Purchase Behavior	Marketing Suggestions
Average-valued customers	Moderate Recency, Moderate Frequency, Moderate Monetary and Average Purchase pattern each month.	Emotion-driven and Metric-driven marketing campaigns
Customers at Risk	High Recency, Low Frequency, Low Monetary and Low purchases made in the last few months.	Personalized Customer Engagement Campaign
Lost Customers	High Recency, Low Frequency, Low Monetary and no purchases made for the past two months.	End of Season Sale Campaigns
Recent Active Buyers	Low Recency, Low Frequency, Low Monetary and sudden spike in purchases made in the past 3 months	Replenishment Marketing Campaigns
High-valued Frequent Buyers	Low Recency, High Frequency, High Monetary and frequent buying pattern through the year with maximum contribution towards revenue.	Loyalty Marketing Campaign
Regular Customers	Low Recency, Moderate Frequency, Moderate Monetary and purchases made through the year with an increase in spending for the past 3 months.	Dynamic Product Recommendation

Customer Behavioral Insights:







1. **Average-valued Customers**

   - These customers have Moderate Recency, Moderate Frequency, and Moderate Monetary values.

   - They also purchase through the year with average spending in each month.

- Based on this purchase intention, we have suggested Emotion-driven and Metric-driven marketing campaigns through which we can try to persuade them to buy more by sending them personalized messages with details on offers, new products etc.

2. **Customers at Risk**

    - These customers have High Recency, Low Frequency, and Low Monetary values.
    - They have made low purchases in the last few months.
    - Based on this purchase intention, we have suggested Personalized Customer Engagement campaigns, wherein we will be constantly in touch with these customers to monitor their interests and purchase patterns.

3. **Lost Customers**

    - These customers have High Recency, Low Frequency, and Low Monetary values.
    - They have not made any purchases in the last two months.
    - Based on this purchase intention, we have suggested End of Season Sale Campaigns because by their purchase pattern we observed that they usually prefer to buy only when there are offers/discounts.

4. **Recent Active Buyers**

    - These customers have Low Recency, Low Frequency, and Low Monetary values.
    - There is a sudden spike in purchases made by these customers in the past 3 months.
    - Based on this purchase intention, we have suggested Replenishment Marketing Campaigns to encourage them to buy more through regular emails giving coupons etc.

5. **High-valued Frequent Buyers**

    - These customers have Low Recency, High Frequency, High Monetary values.
    - They show a frequent buying pattern through the year with maximum contribution towards revenue.
    - Based on this purchase intention, we have suggested a Loyalty Marketing Campaign to retain them by offering Reward points and Loyalty cards.

6. **Regular Customers**

    - These customers have Low Recency, Moderate Frequency, Moderate Monetary values.
    - They have made purchases throughout the year with an increase in spending for the past 3 months.
    - Based on this purchase intention, we have suggested Dynamic Product Recommendation by suggesting them to buy more by sending them customized product suggestions.

Business Recommendation Summary:

1. Average-valued Customers have the potential for becoming regular customers. Therefore, efforts should be made to turn them into loyal customers.

2. Customers at Risk are the customers who are at a risk of churning. We should focus on these customers to make them Loyal to our business.

3. Lost Customers are the customers who are no longer in business with us hence we need to strategize a marketing plan to make them buy again.

4. Recent Active Buyers are the customers who have started to buy recently, and we need to focus on them to make them regular buyers.

5. High-Valued Frequent Buyers are the customers who are our best customers who buy regularly and bring in high revenue to the business. We should try to retain these customers and make them loyal to our business.

6. Regular Customers are the customers who have shop regularly throughout the year and are also contributing significantly to the revenue. We should focus on these customers to make them purchase more as they are steady customers and are good for business.

The future scope of this project would be to:

- Identify the scope for further Hyperparameter tuning.
- Fish bone diagram (cause and effect) to illustrate the relationship among variables which lead to low customer satisfaction.
- Perform extensive literature survey on the working of Perturbation, sensitivity, and the type of sampling technique which could be used for a dataset with less features and more data points.
- Develop a recommendation system on the basis of customer purchase behaviors for an E-Commerce sector to suggest products that users in the same cluster usually prefer.
- Using pipelines, build the final model for deployment so that the model performance could be validated in production.

References

[1] https://ecommerceguide.com/guides/what-is-ecommerce/

[2] Chen, D., Sain, S. & Guo, K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. J Database Mark Cust Strategy Manag 19, 197–208 (2012). https://doi.org/10.1057/dbm.2012.17

[3] P. Anitha, Malini M. Patil, RFM model for customer purchase behavior using K-Means algorithm, Journal of King Saud University - Computer and Information Sciences, 2019, ISSN 1319-1578, DOI: https://doi.org/10.1016/j.jksuci.2019.12.011

[4] Ashishkumar Singh, Grace Rumantir, Annie South, and Blair Bethwaite. 2014. Clustering Experiments on Big Transaction Data for Market Segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing (BigDataScience '14). Association for Computing Machinery, New York, NY, USA, Article 16, 1–7. DOI: https://doi.org/10.1145/2640087.2644161

[5] Chen, D., Sain, S. & Guo, K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. J Database Mark Cust Strategy Manag 19, 197–208 (2012). DOI: https://doi.org/10.1057/dbm.2012.17

[6] Kurniawan, Fachrul & Umayah, Binti & Hammad, Jehad & Nugroho, Supeno & Hariadi, Mohamad. (2017). Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data. Knowledge Engineering and Data Science. 1. 20. DOI: 10.17977/um018v1i12018p20-25

[7] Olena Piskunova and Rostyslav Klochko. Classification of e-commerce customers based on Data Science techniques
http://ceur-ws.org/Vol-2649/paper2.pdf

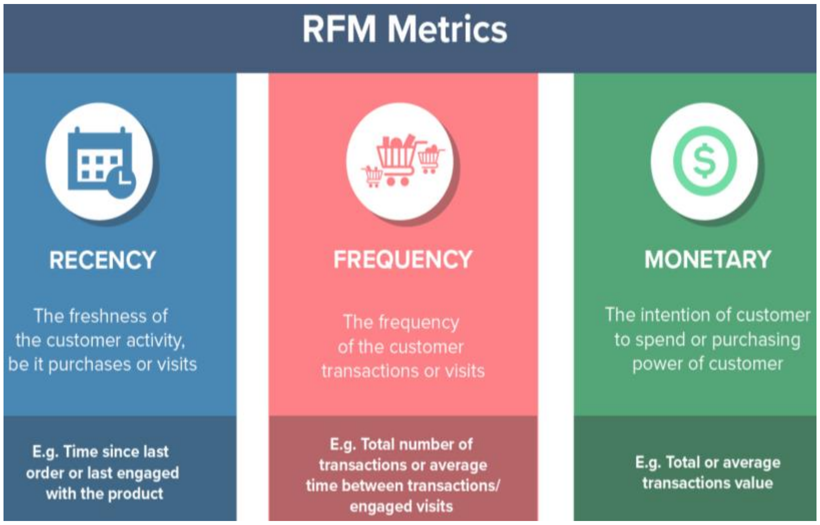
[8] https://www.zaius.com/customer-lifetime-value-cltv/

Appendix

Customer Segmentation using RFM analysis

RFM stands for Recency, Frequency and Monetary value. It is a three-dimensional approach to identify specific target groups based on the three metrics mentioned thereby quantitatively determining best-valued customers based on the recent purchases, purchase frequency and how much the customer spends on these purchases.

Based on the insights from RFM analysis, business firms segment their customer base and design marketing strategies for better profitability. The main disadvantage of this technique is that during the RFM analysis, we may fail to consider certain quantifiable /non-quantifiable factors that could define customer behaviors and how they will respond to marketing campaigns.



Source: https://clevertap.com/blog/rfm-analysis/

Early Findings

- This company sells gifts for all the occasions/festivals. So, as per analysis, there could be seasonality in sales in this dataset as festive seasons could be there at different times of an year.

- When we have multiple quantities, the overall purchase value could be calculated by multiplying the quantity and unit price.

- Another observation is that this dataset provides additional information that the majority of customers belong to the wholesaler’s category. Based on the amount of quantity purchased, we can identify which category the customer belongs to (Small, Medium, or Large-scale wholesaler).

- There could also be customers who are not wholesalers whose purchase behaviors needs to be analyzed.
- Extracting InvoiceNos preceding with character ‘C’, we could find out the total proportion of invoices that were cancelled.
- The Country information of the customers could be utilized to build region-wise segmentation.
- As per earlier analysis we have found the majority of them belong to the UK while there is a significant number of customers from other countries.

Additional Analysis

The scope for further analysis is as follows:

- **Bestseller** - Identifying the best-selling products using StockCode

  Grouping the products based on stockCode and aggregating the sum to find which are the products that are best sellers in this Dataset. There is also scope for analyzing which is the product that has been Bestseller on particular seasons.

- **Buyer Frequency** - Identifying buyer frequency for each StockCode

  Based on the customer data and the history of his purchases we could find the estimate of the frequency at which he could buy the same products.

- **Total Sales** - Calculating the total sales (UnitPrice * Quantity) for each customer/StockCode

  This could be added as a new feature to the dataset which could tell about the value of a purchase. We can also estimate which customer has the highest monetary value in this dataset.

- **Country-wise Sales** - Identify the product pricing in each country and suggesting best pricing for low-sales countries

  There are other countries' data in the dataset in which the customer from different countries could also have purchased the same product. But the pricing of the products could be different in different countries. We can do the analysis if the product pricing in different countries contributes to the purchase of the product.

Customers who have placed the highest number of orders:

	CustomerID	Country	InvoiceNo
4018	17841.0	United Kingdom	7847
1887	14911.0	EIRE	5675
1297	14096.0	United Kingdom	5111
334	12748.0	United Kingdom	4595
1669	14606.0	United Kingdom	2700

Customers who have placed the least number of orders:

	CustomerID	Country	InvoiceNo
0	12346.0	United Kingdom	1
2794	16144.0	United Kingdom	1
1741	14705.0	United Kingdom	1
643	13185.0	United Kingdom	1
2752	16093.0	United Kingdom	1

Calculation of the missing values in percentage for this dataset:

```
## Missing values
missing = df.isnull().sum()/df.shape[0] * 100
missing[missing > 0]
```

```
Description     0.268311
CustomerID     24.926694
dtype: float64
```

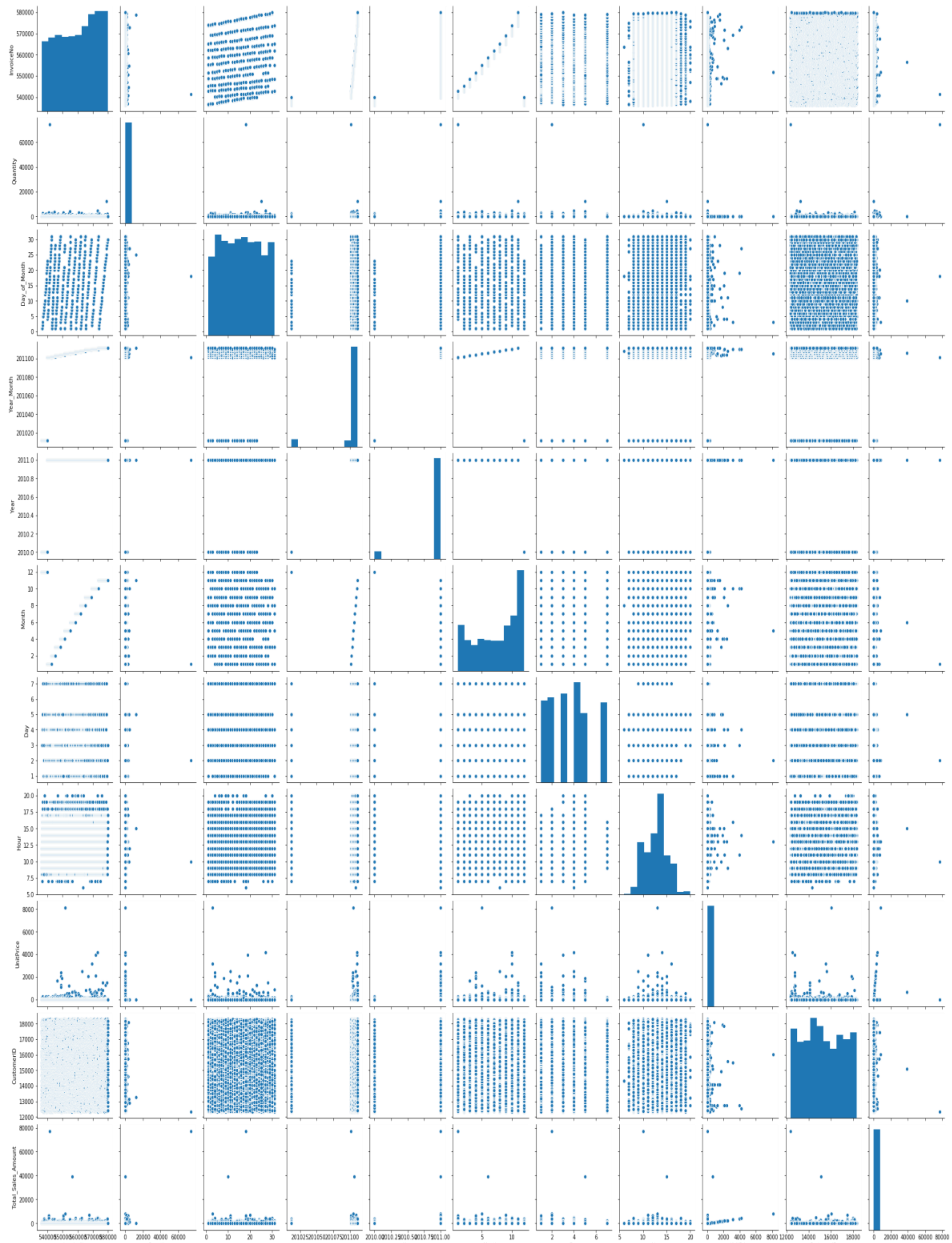
Quarterly Revenue

	Quantity	Total_Sales_Amount
Quarter		
2010Q4	311063	570422.730
2011Q1	961188	1608267.990
2011Q2	1027331	1805775.531
2011Q3	1309216	2193704.143
2011Q4	1557088	2709038.500

Countries with Maximum Percentage of Sales

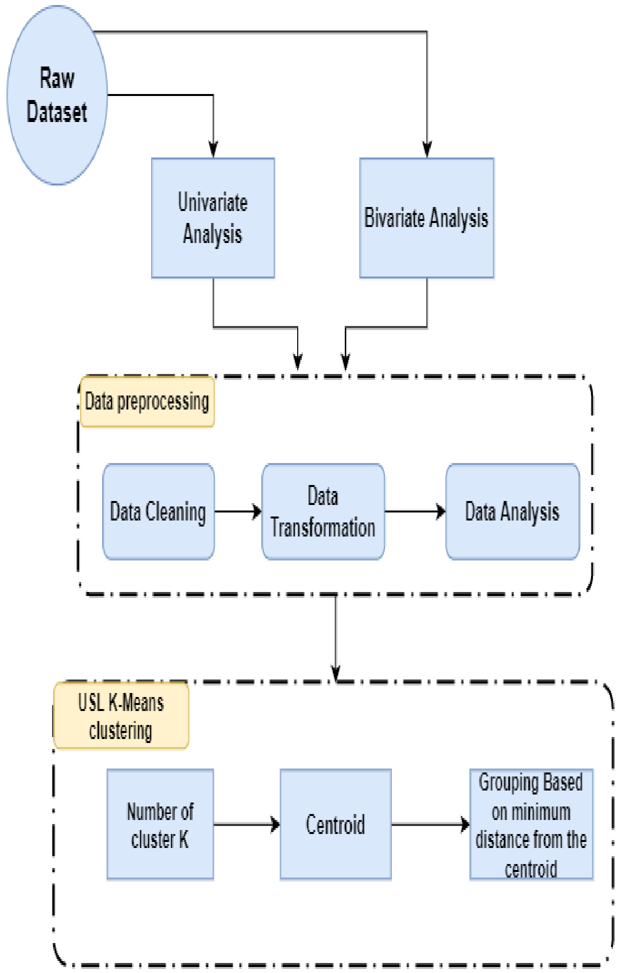
	Country	Percentage of Sales
0	United Kingdom	88.922471
1	Germany	2.298514
2	France	2.120275
3	EIRE	1.840441
4	Others	0.727468
5	Spain	0.631474
6	Netherlands	0.601683
7	Belgium	0.517147
8	Switzerland	0.469022
9	Portugal	0.369972
10	Australia	0.301478
11	Norway	0.272960
12	Italy	0.193007
13	Channel Islands	0.190206
14	Finland	0.174419
15	Cyprus	0.153540
16	Sweden	0.114582
17	Austria	0.101341

Pair Plot



InvoiceNo
Quantity
Day_of_Month
Year_Month
Year
Month
Day
Hour
UnitPrice
CustomerID
Total_Sales_Amount

Process Flow Diagram:

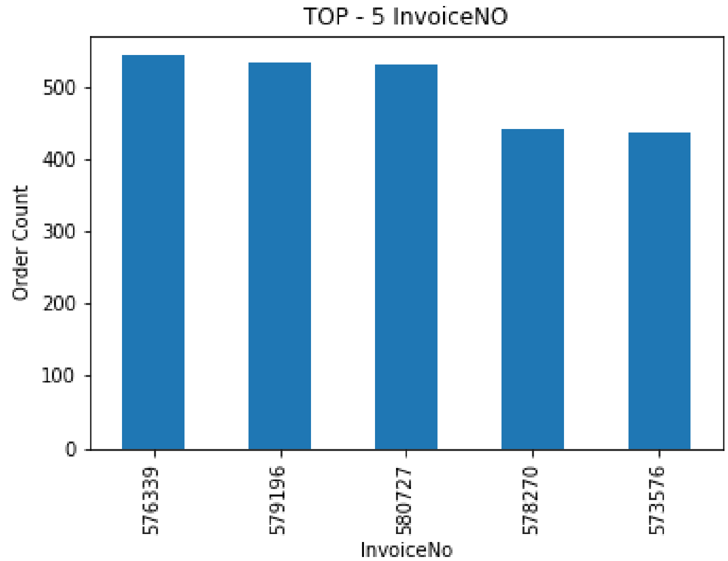


Univariate Analysis:

Feature-1: InvoiceNo

- The Invoice no is a 6-digit number numerical number which was generated at the time of the transaction. The Invoice number could precede with a character ‘C’ which denotes that the order was cancelled which is beyond the scope of our analysis. So, we have dropped those records already.
- There are 18536 unique InvoiceNo.

Top-5 InvoiceNo with maximum orders:



- CustomerID 14096 has the greatest number of orders.

Feature 2: StockCode

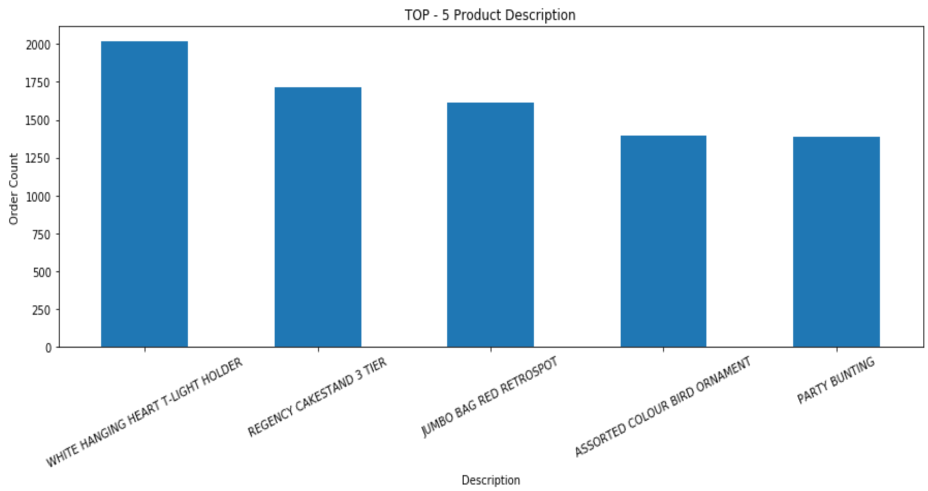
- Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Top-5 StockCode with maximum orders

* From the above plot, we can infer that StockCode 85123A holds the highest order count and the product is WHITE HANGING HEART T-LIGHT HOLDER is frequently bought.

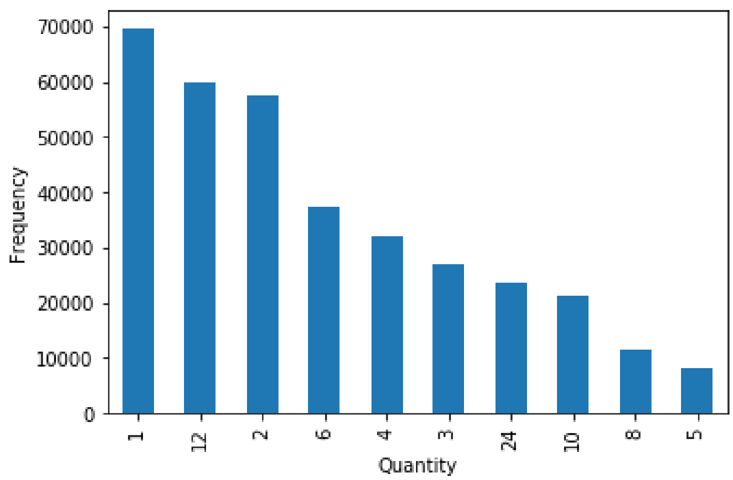
Feature 3: Description

* Description consists of the Name of the product which is actually Nominal in nature.
* There are 3877 distinct product values.



Feature 4: Quantity

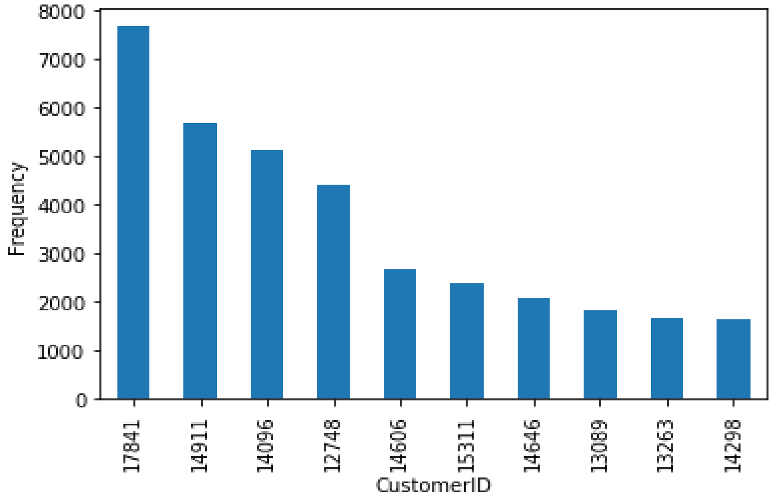
* Quantity describes the number of quantities the product was purchased in a single transaction.



* Most customers have made only single quantity purchase

Feature 5: CustomerID

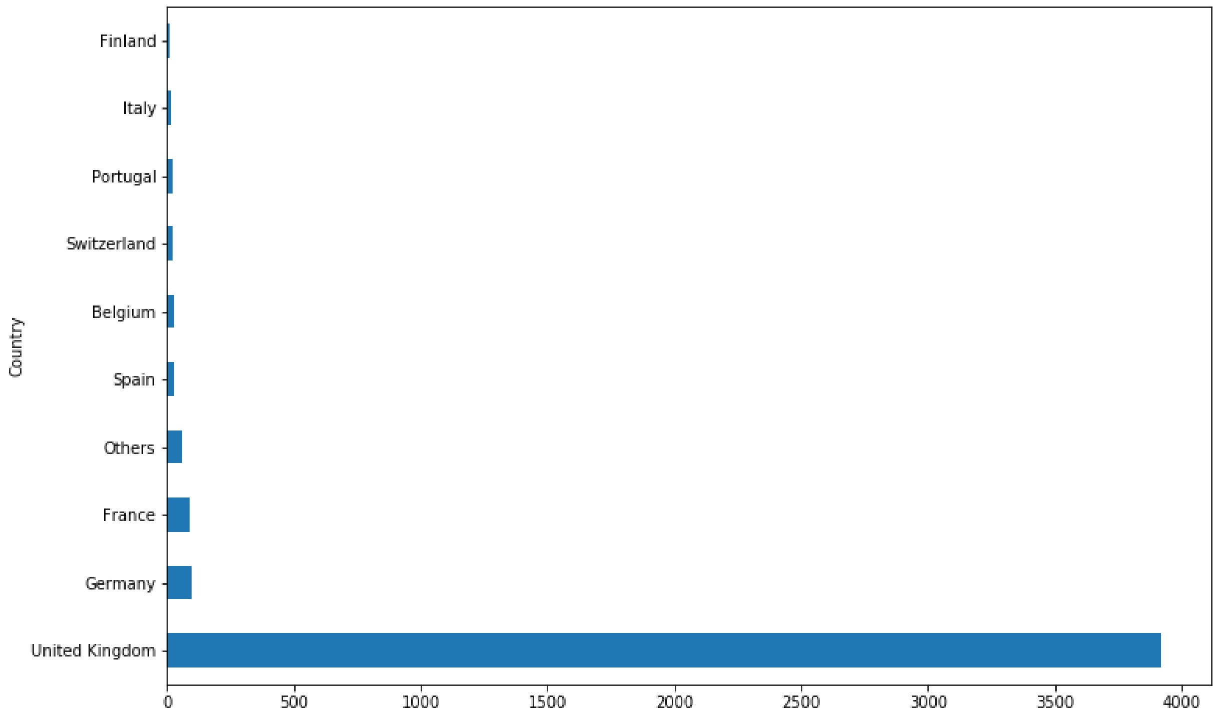
- CustomerID is a unique 5 digit numeric and nominal entry that is assigned to a customer.
- There are a total of 4339 customers.



- The purchase frequency of top 10 customers in the year 2010 to 2011 ranges from 7847 to 1637.
- There are 4267 customers who shopped more than once.
- There are only 72 one-time shoppers.
- There are 4039 customers who purchased more than 5 times.

Feature 6: Country

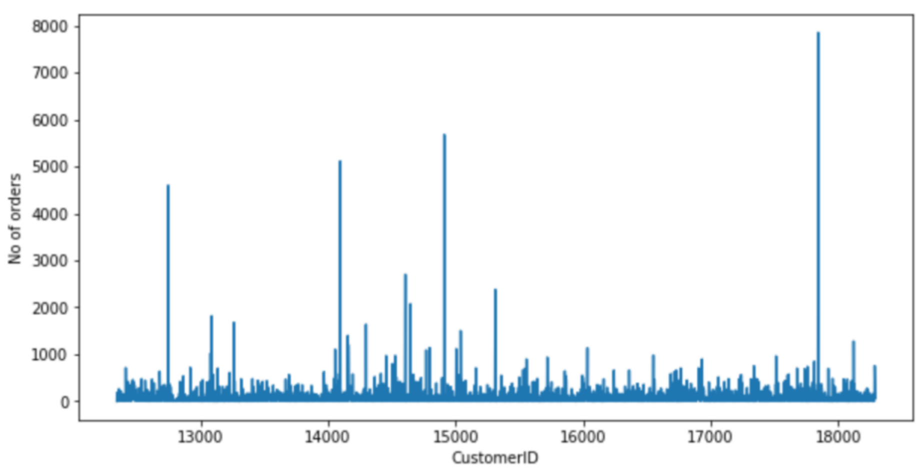
- Country shows which country/region the customer belongs.
- Total customers from the top 10 cities are 4280.



Out of total 4339, 4280 customers are from Top 10 selling Cities which is around 98.64%

- The United Kingdom, Germany and France are the top 3 countries which have more customers.
- There are zero countries where shopping is done only once, which is a good sign for business because customers are shopping more than once from all the countries.

Identifying the best-valued and least-valued customers



- From the above graph, we can see that the majority of the customers placed less than 1000 orders.
- Top 5 customers have placed more than 5000 orders and bottom customers have a minimum order of 1.

Top 20 customers

	CustomerID	Country	Total Number of Orders Placed
4019	17841	United Kingdom	7676
1888	14911	EIRE	5672
1298	14096	United Kingdom	5111
334	12748	United Kingdom	4413
1670	14606	United Kingdom	2677
2185	15311	United Kingdom	2366
1698	14646	Netherlands	2080
570	13089	United Kingdom	1814
699	13263	United Kingdom	1667
1443	14298	United Kingdom	1637
1973	15039	United Kingdom	1477
1342	14156	EIRE	1395
4222	18118	United Kingdom	1263
1345	14159	United Kingdom	1175
1806	14796	United Kingdom	1132
2714	16033	United Kingdom	1128
1946	15005	United Kingdom	1112
1271	14056	United Kingdom	1088
1788	14769	United Kingdom	1062
566	13081	United Kingdom	1028

Bottom 20 customers

	CustomerID	Country	Total Number of Orders Placed
0	12346	United Kingdom	1
4197	18084	United Kingdom	1
705	13270	United Kingdom	1
3226	16738	United Kingdom	1
2442	15657	United Kingdom	1
3228	16742	United Kingdom	1
4218	18113	United Kingdom	1
3249	16765	United Kingdom	1
2451	15668	United Kingdom	1
522	13017	United Kingdom	1
3225	16737	United Kingdom	1
3935	17715	United Kingdom	1
1295	14090	United Kingdom	1
728	13302	United Kingdom	1
3334	16881	United Kingdom	1
731	13307	United Kingdom	1
2923	16323	United Kingdom	1
3732	17443	Others	1
3389	16953	United Kingdom	1
4263	18174	United Kingdom	1

Top 20 spenders

	CustomerID	Country	Total Money Spent
1698	14646	Netherlands	280206.02
4210	18102	United Kingdom	259657.30
3737	17450	United Kingdom	194390.79
3017	16446	United Kingdom	168472.50
1888	14911	EIRE	143711.17
57	12415	Australia	124914.53
1342	14156	EIRE	117210.08
3780	17511	United Kingdom	91062.38
2711	16029	United Kingdom	80850.84
0	12346	United Kingdom	77183.60
3185	16684	United Kingdom	66653.56
1298	14096	United Kingdom	65164.79
1005	13694	United Kingdom	65039.62
2185	15311	United Kingdom	60632.75
570	13089	United Kingdom	58762.08
4102	17949	United Kingdom	58510.48
2526	15769	United Kingdom	56252.72
1992	15061	United Kingdom	54534.14
1443	14298	United Kingdom	51527.30
1293	14088	United Kingdom	50491.81

Bottom 20 spenders

	CustomerID	Country	Total Money Spent
693	13256	United Kingdom	0.00
3226	16738	United Kingdom	3.75
1802	14792	United Kingdom	6.20
3023	16454	United Kingdom	6.90
4107	17956	United Kingdom	12.75
3332	16878	United Kingdom	13.30
3969	17763	United Kingdom	15.00
731	13307	United Kingdom	15.00
2567	15823	United Kingdom	15.00
2753	16093	United Kingdom	17.00
3389	16953	United Kingdom	20.80
4131	17986	United Kingdom	20.80
3495	17102	United Kingdom	25.50
4333	18268	United Kingdom	25.50
2442	15657	United Kingdom	30.00
594	13120	United Kingdom	30.60
3705	17408	United Kingdom	32.65
3249	16765	United Kingdom	34.00
2506	15744	United Kingdom	34.80
4012	17831	United Kingdom	35.40