Customer Segmentation

Suraj Kumar

Shreya Ranjan

Data Science Trainee

Alma Better, Bangalore

Abstract:

Customer segmentation is a technique for dividing the customer base into groups based on shared traits. By using customer segmentation, marketers will be able to more effectively target certain audience subsets with their marketing. Our objective is to find the largest customer segments in a dataset of transactions from a UK-based, registered non-store internet retailer that took place between December 1, 2010, and December 9, 2011. The company primarily offers each gift for every occasion.

We will use an RFM model, one of the 6 different customer segmentation models currently available (Recency, Frequency and Monetary). RFM is a technique that's frequently used to identify customers based on when their last purchase was made, how many purchases they've made overall, and how much money they've spent. This is frequently used to determine High Value Customer base. We used the silhouette score method and the elbow method to calculate the ideal value of "K" and performed K-Means clustering to differentiate between the clusters or segments.

Problem Statement:

In this project, our goal is to identify the key consumer categories using a transactional data set that includes every transaction made by a UK-based, registered non-store internet retailer between December 1, 2010, and December 9, 2011. The company primarily offers each gift for every occasion. The company has a large number of wholesalers as clients.

Customer segmentation is a technique for classifying customers into groups based on shared characteristics. By using customer segmentation, marketers will be able to more effectively target certain audience subsets with their marketing. There are 541909 records of transactions with 8 features in the dataset.

Data Description:

The dataset contains 541909 records and 8 features which consists of:

- **Invoice No:** Invoice number. Nominal, a six-digit integral number assigned to each transaction specifically. This code denotes a cancellation if it begins with the letter "c.".
- **Stock Code:** Product (item) code. A 5-digit integral number known as the nominal is assigned to each unique product.
- **Description:** Name of the Product (Item). Nominal.
- **Quantity:** The number of each item (product) in each transaction. Numeric.
- **Invoice Date:** Invoice Time and date. The day and time that each transaction was created, represented by a number.
- **Unit Price:** Unit pricing. Numeric, Sterling unit price for the product.
- **CustomerID:** Customer number. Nominal, a five-digit integral number assigned to every customer uniquely.
- **Country:** Country name. Nominal, the name of the country in which each customer resides.

Introduction:

In this project, our goal is to identify the key consumer categories using a transactional data set that includes every transaction made by a UK-based, registered non-store internet retailer between December 1, 2010, and December 9, 2011. The company primarily offers each gift for every occasion.

Customer segmentation is the practice of grouping the consumers of a firm into categories that represent the similarities among the customers in each category. In order to optimize each customer's value to the company, it is important to segment customers in order to determine how to interact with them.
Customer segmentation may enable marketers to reach out to each customer in the most efficient manner. A customer segmentation analysis enables marketers to accurately identify distinct groups of customers based on demographic, behavioral, and other factors by utilizing the vast amount of customer (and potential customer) data accessible.

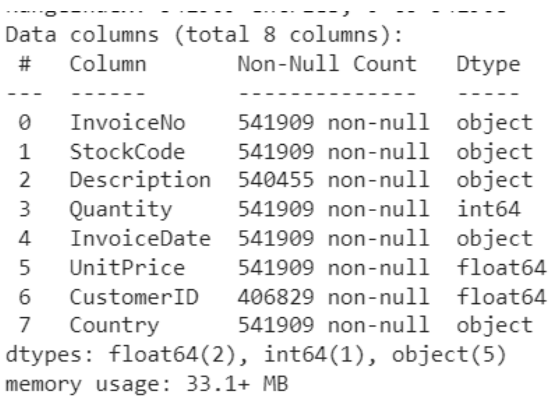
Marketers can better target different audience subgroups with their marketing efforts by segmenting their audiences. Both product development and communications might be a part of those efforts. Specifically, segmentation helps a company:

- Create and deliver targeted marketing communications that will connect to some customer segments but not to others (who will receive messages tailored to their needs and interests, instead).
- Depending on the segment, choose the most effective communication medium, which may be an email, a social media post, radio advertising, or another strategy.
- Determine possibilities for new or improved products or services.
- Establish better customer relationships.
- Test pricing options.
- Focus on the most profitable customers.
- Improve customer service.
- Cross-sell and upsell additional goods and services.

Exploratory Data Analysis:

- **Data Cleaning**:

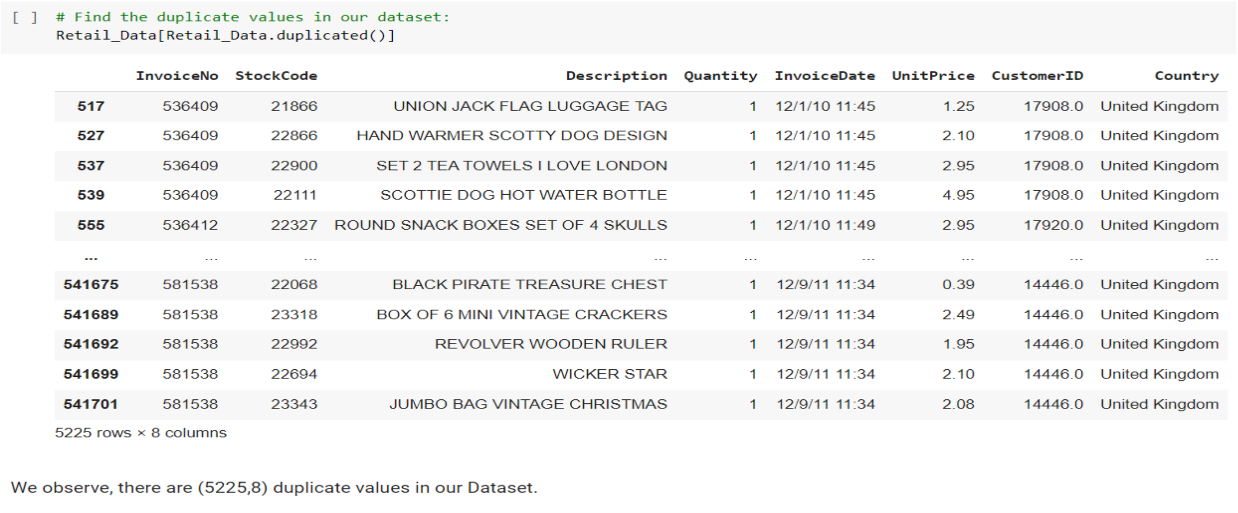
  One of the most time-consuming steps in data analysis is data cleaning. A dataset may contain formatting problems, missing numbers, duplicate rows, spelling errors, and other flaws. These issues make data analysis challenging and lead to inaccurate conclusions. A dataset may have missing or duplicate data for a variety of reasons. As we carry out cleaning and transformation processes including merging data, reindexing data, and reshaping data, missing or duplicate data is occasionally introduced. Other times, it is present in the original dataset due to difficulties with data conversion, storage, or user input.
  Information about different columns of the Dataset:



```
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Because most of the machine learning models, we are going to use will give us an error if we send them NaN values, we need to fill in the missing variables. The simplest solution is to simply fill them with 0, however doing so can significantly lower the accuracy of our model. There are numerous ways to fill in missing data, but before selecting the most effective one, we must understand the kind of missing item and its importance.

In the dataset we are holding, there are (5225,8) duplicate records, and two columns have a significant number of missing values. Therefore, we made the decision to discard these data before moving on.



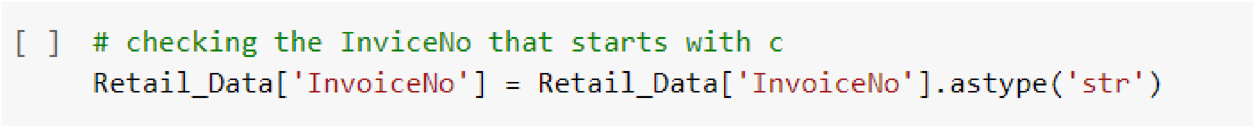
```
[ ] # Find the duplicate values in our dataset:
    Retail_Data[Retail_Data.duplicated()]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
517	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	12/1/10 11:45	1.25	17908.0	United Kingdom
527	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	12/1/10 11:45	2.10	17908.0	United Kingdom
537	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	12/1/10 11:45	2.95	17908.0	United Kingdom
539	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	12/1/10 11:45	4.95	17908.0	United Kingdom
555	536412	22327	ROUND SNACK BOXES SET OF 4 SKULLS	1	12/1/10 11:49	2.95	17920.0	United Kingdom
...	...	...	...	...	...	...	...	...
541675	581538	22068	BLACK PIRATE TREASURE CHEST	1	12/9/11 11:34	0.39	14446.0	United Kingdom
541689	581538	23318	BOX OF 6 MINI VINTAGE CRACKERS	1	12/9/11 11:34	2.49	14446.0	United Kingdom
541692	581538	22992	REVOLVER WOODEN RULER	1	12/9/11 11:34	1.95	14446.0	United Kingdom
541699	581538	22694	WICKER STAR	1	12/9/11 11:34	2.10	14446.0	United Kingdom
541701	581538	23343	JUMBO BAG VINTAGE CHRISTMAS	1	12/9/11 11:34	2.08	14446.0	United Kingdom

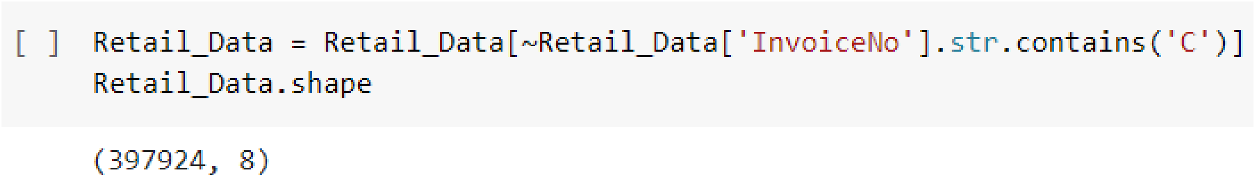
5225 rows × 8 columns

We observe, there are (5225,8) duplicate values in our Dataset.

Before going on to feature engineering or visualizations, we had to clean up a few more features in addition to the duplicate entries and missing value issues. The dataset contains 3680 records with cancelled orders and quantities that are negative, thus we will ignore these numbers and only take into account active orders for our segmentation needs.



```
[ ] # checking the InviceNo that starts with c
    Retail_Data['InvoiceNo'] = Retail_Data['InvoiceNo'].astype('str')
```



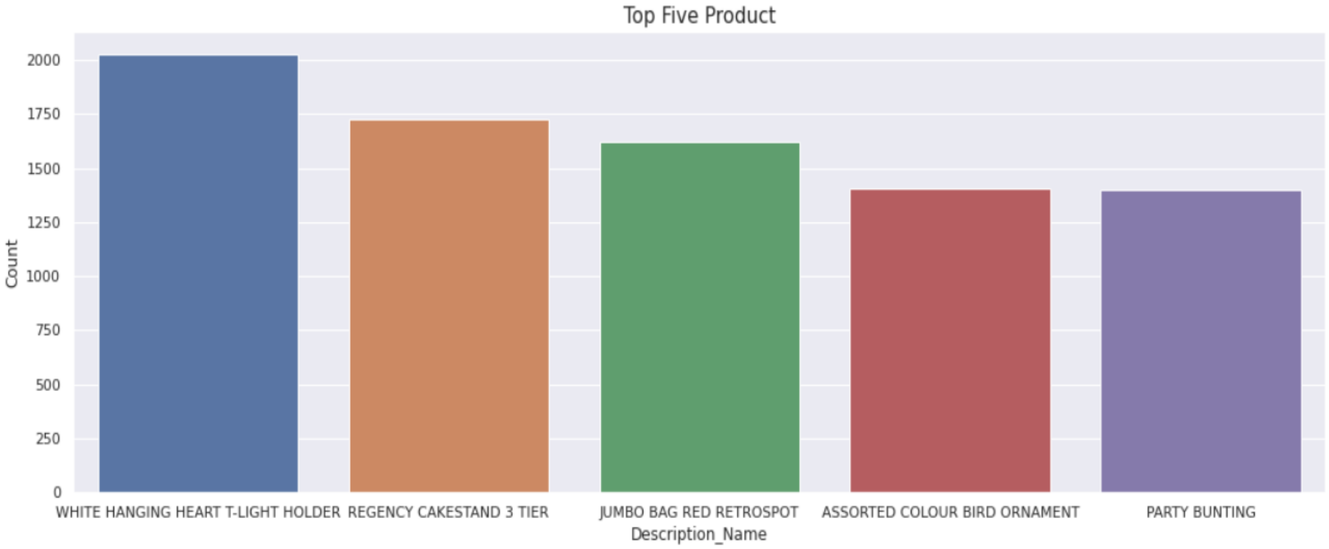
```
[ ] Retail_Data = Retail_Data[~Retail_Data['InvoiceNo'].str.contains('C')]
    Retail_Data.shape

    (397924, 8)
```

After looking into the describe

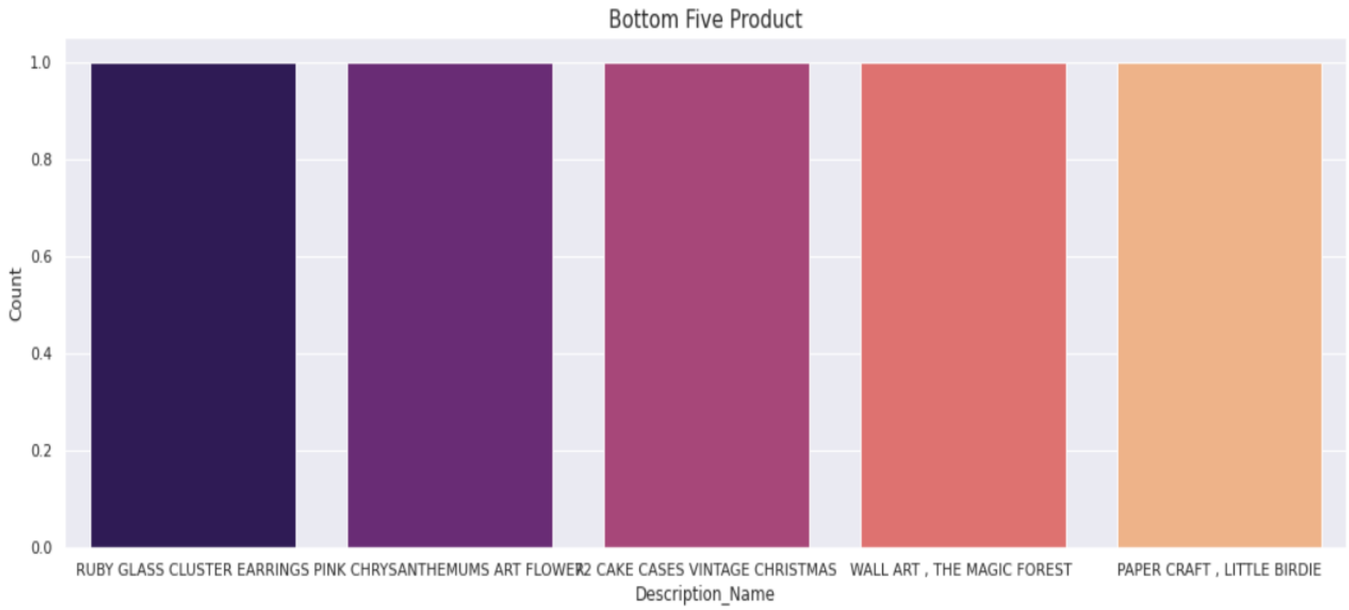
() output, we discovered that numerous records have unit prices of zero, which is not possible for any store or company. Since no business offers things for free, we looked into each record and chose to take into account those where the unit price is higher than zero for categorization.

- **Feature Engineering:**
  As soon as we got a clean and prepared dataset, we made the decision to build some new features that would aid in the segmentation of various client segments according to the segmentation model. The act of choosing, modifying, and transforming raw data into features that may be used in machine learning is known as feature engineering. It could be important to create and train better features in order to make machine learning effective on new tasks. A machine learning technique called feature engineering uses data to generate new variables that aren't present in the training set. With the aim of simplifying and optimizing data transformations while also improving model accuracy, it can generate new features for both supervised and unsupervised learning.

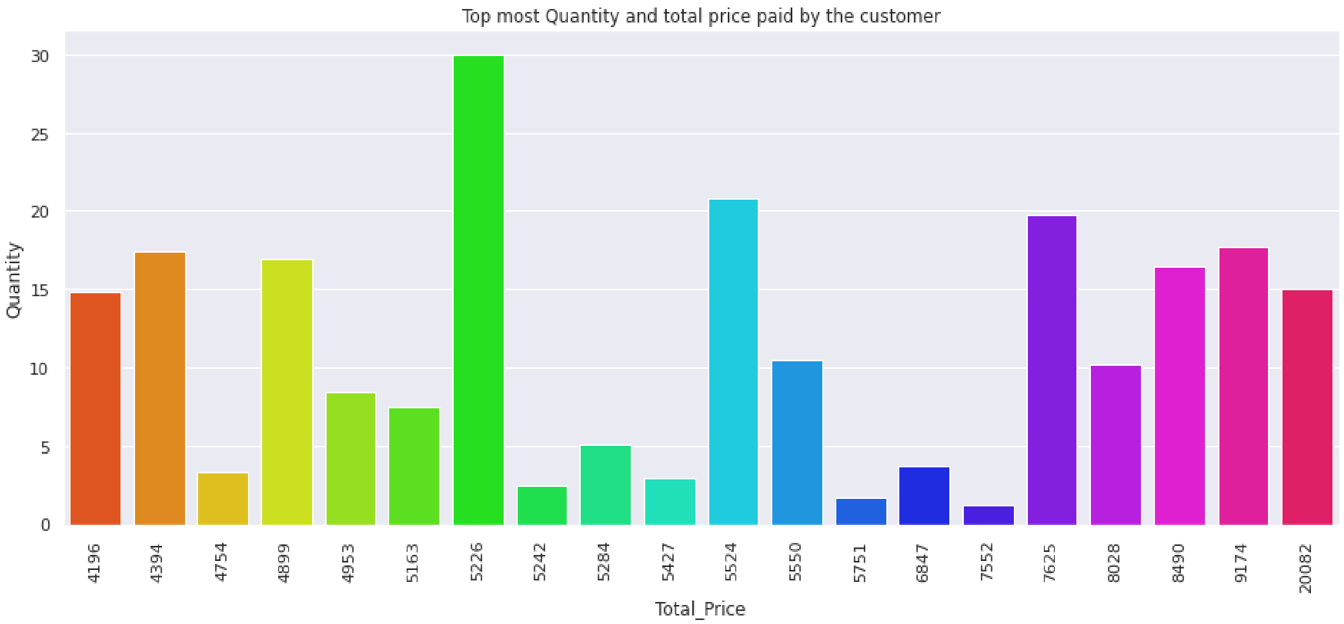
- **Data Visualization/Exploration**:
  - The top 5 most sold products are:



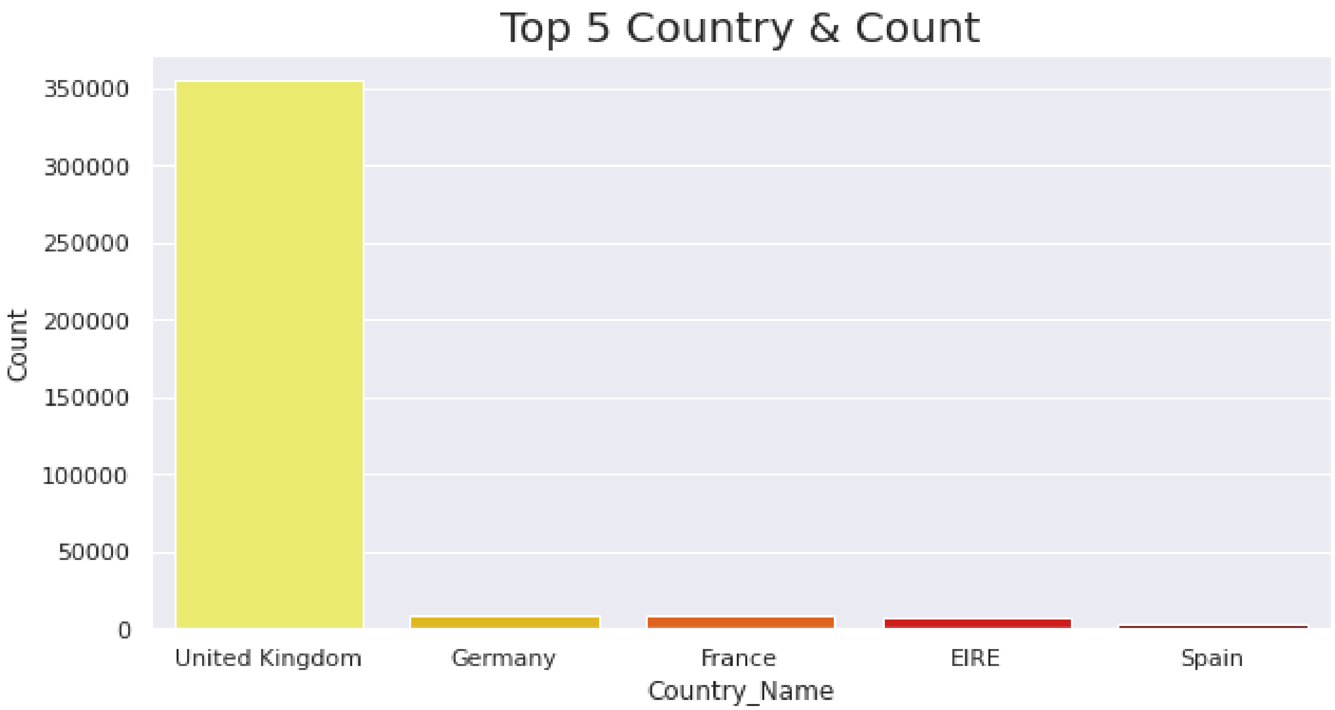
- The bottom 5 least sold products are:



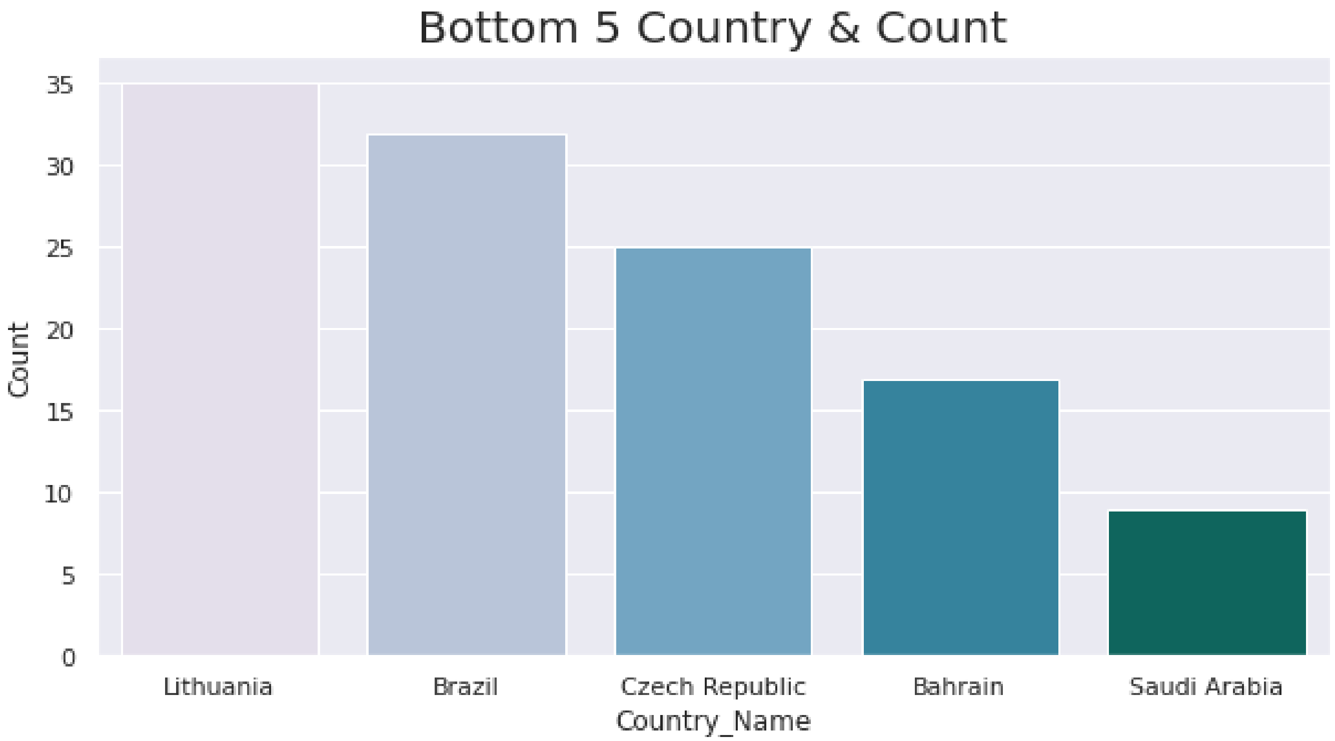
- Top most Quantity and total price paid by the customer:



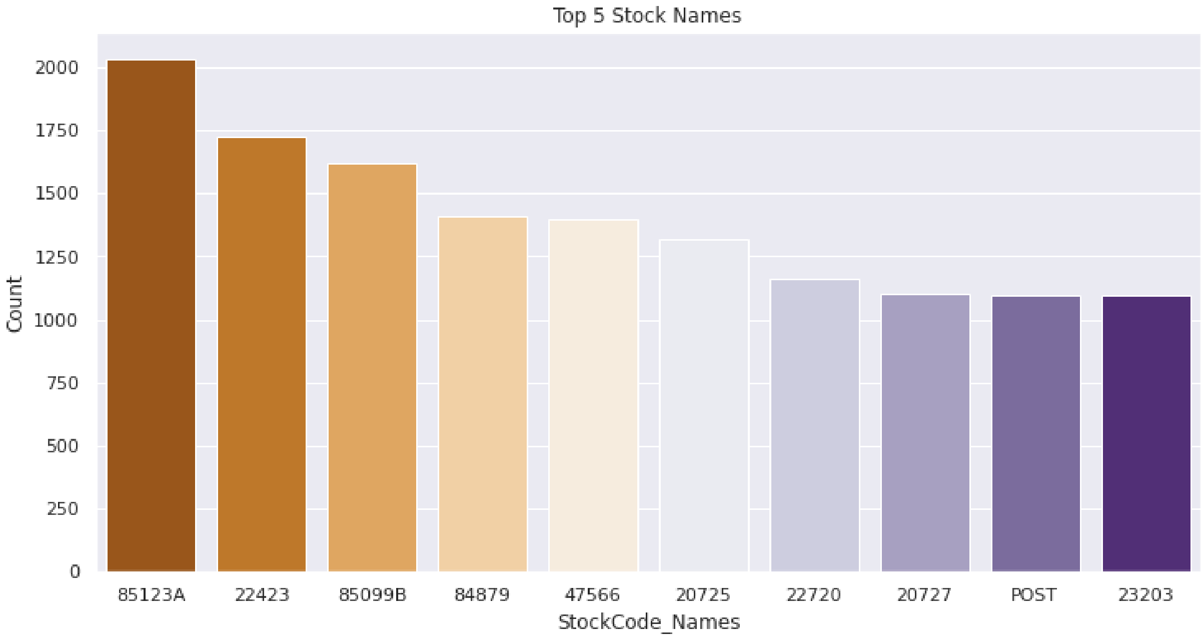
- The top 5 countries where the customers belong or reside:



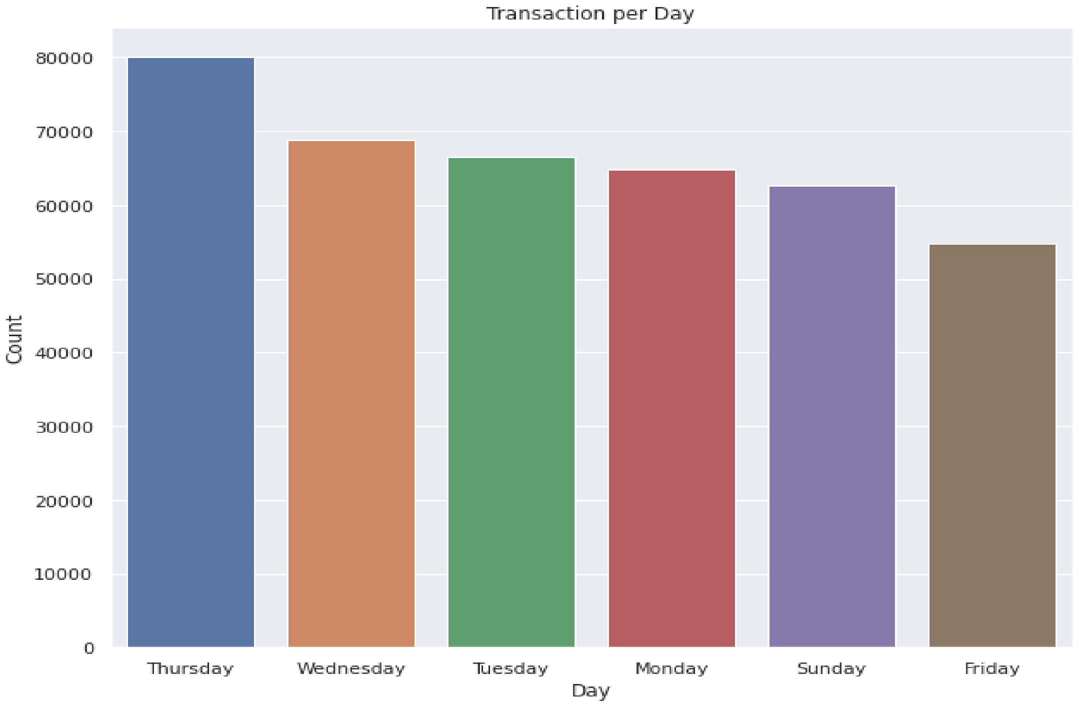
- The bottom 5 countries where the least customers belong or reside:



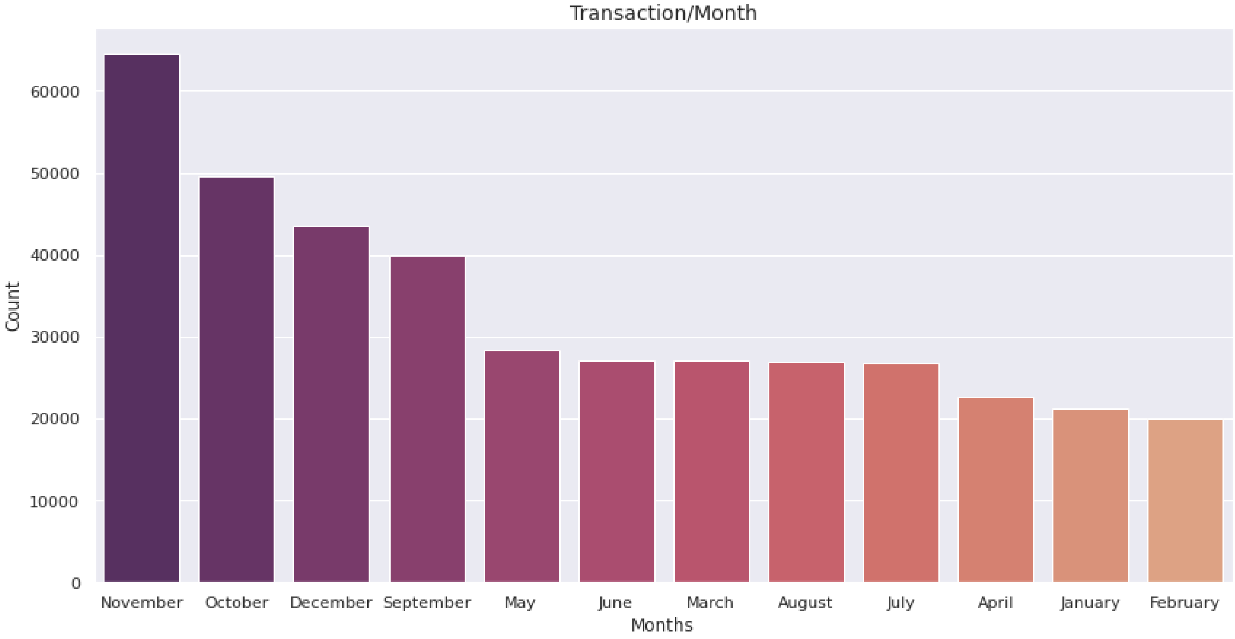
- Top 10 Stock name based on selling:



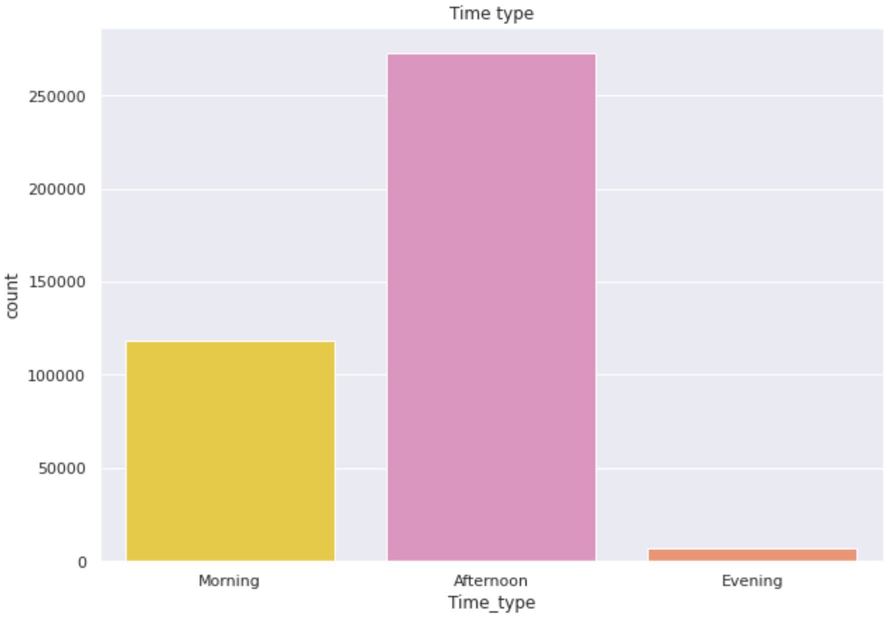
- Most of the customers had made a purchase on Thursday followed by Wednesday and the least number of purchases on Friday.



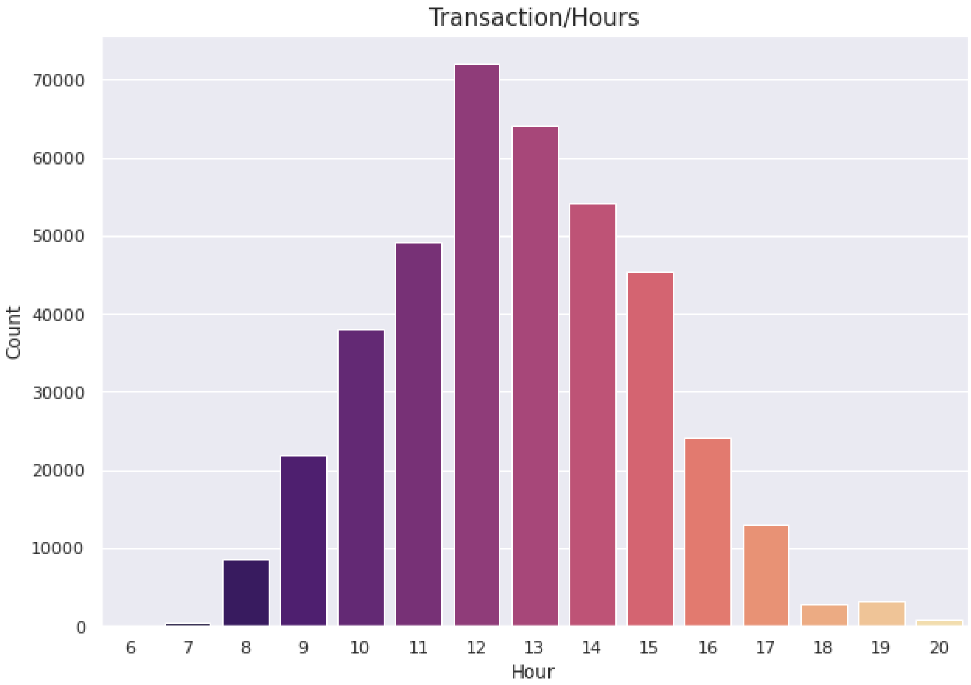
- The highest number of purchases has occurred during the festive months October to December and the least number of purchases has occurred during the initial months of January and February.



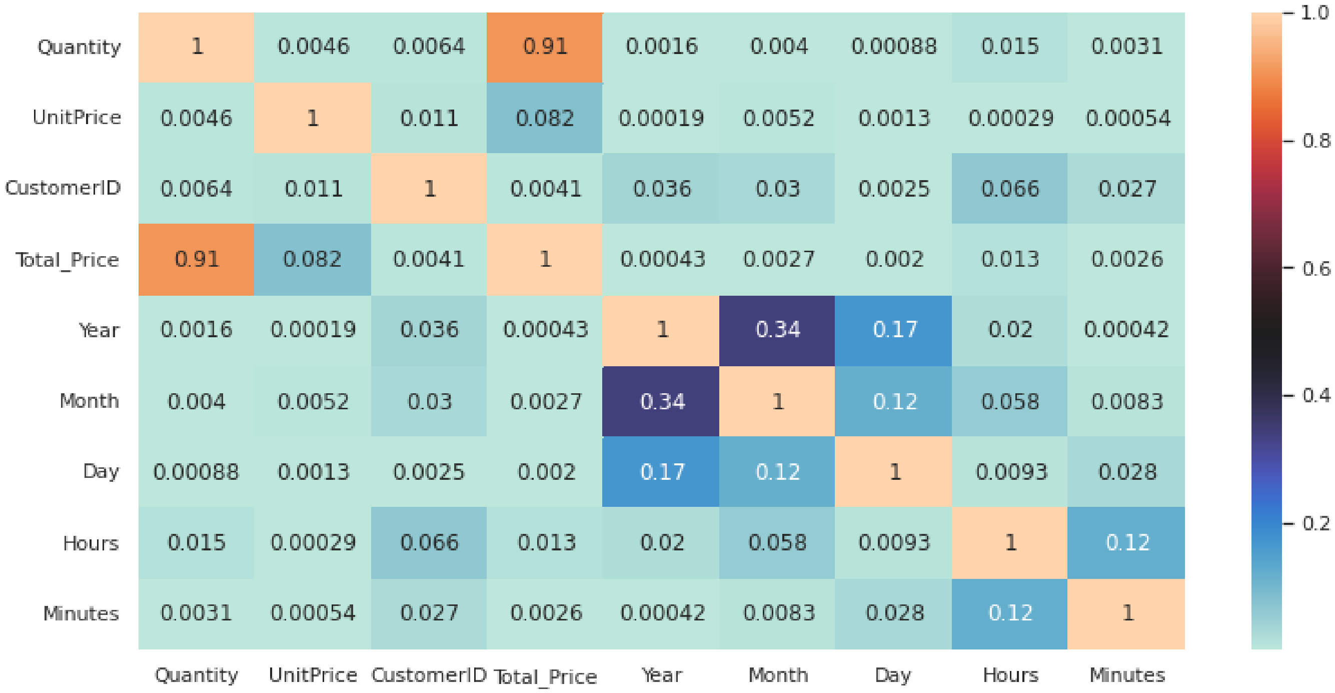
- Interestingly the time of the day in which most number of the purchases has taken place is during the afternoon and the least number of purchases during evening.



- People do most of their shopping between the hours of 11 and 14, therefore practically all of them are free at this time.

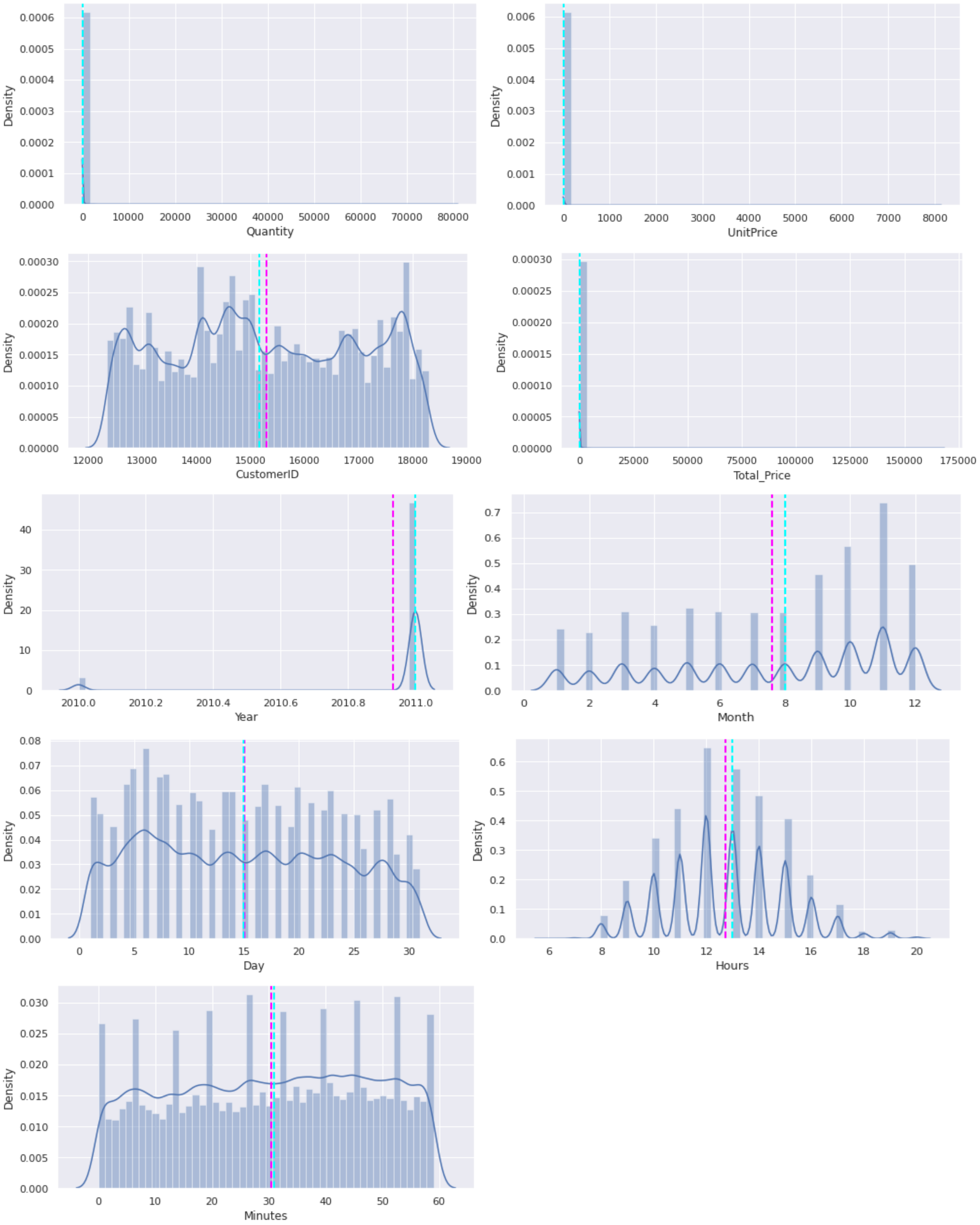


- Correlation matrix:



	Quantity	UnitPrice	CustomerID	Total_Price	Year	Month	Day	Hours	Minutes
Quantity	1	0.0046	0.0064	0.91	0.0016	0.004	0.00088	0.015	0.0031
UnitPrice	0.0046	1	0.011	0.082	0.00019	0.0052	0.0013	0.00029	0.00054
CustomerID	0.0064	0.011	1	0.0041	0.036	0.03	0.0025	0.066	0.027
Total_Price	0.91	0.082	0.0041	1	0.00043	0.0027	0.002	0.013	0.0026
Year	0.0016	0.00019	0.036	0.00043	1	0.34	0.17	0.02	0.00042
Month	0.004	0.0052	0.03	0.0027	0.34	1	0.12	0.058	0.0083
Day	0.00088	0.0013	0.0025	0.002	0.17	0.12	1	0.0093	0.028
Hours	0.015	0.00029	0.066	0.013	0.02	0.058	0.0093	1	0.12
Minutes	0.0031	0.00054	0.027	0.0026	0.00042	0.0083	0.028	0.12	1

- Distribution of all the numerical features are right skewed and thus we had to apply log transformation on these features to bring them near normal.



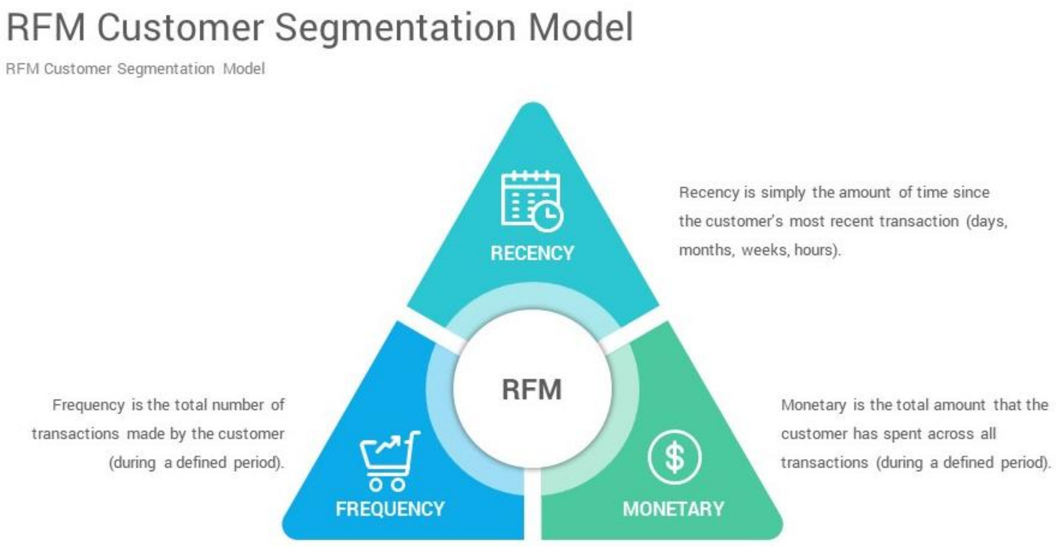
RFM Segmentation & Analysis:

It is always required to establish several quantitative factors on which the algorithm will execute segmentation prior to using any clustering techniques. These include characteristics like the total amount spent, the customer's activity level, their most recent visit, etc.
The RFM model is based on three quantitative factors:

- Recency: How recently a customer has made a purchase (Latest Date - Last Inovice Data)
- Frequency: How often a customer makes a purchase (Count of Invoice No. of transactions)
- Monetary: How much money a customer spends on purchases (Sum of Total Amount for each customer)

One of the steps is the RFM model, which stands for Recency, Frequency, and Monetary. For each customer, we calculate the recency—the number of days since the customer's last visit—the frequency—how frequently they make repeat purchases—and the monetary—their overall spending.

There are other processes where we break each of these attributes into the appropriate groups and figure out a score for each client. However, since segmentation may be done manually, this strategy does not need machine learning algorithms. As a result, we will skip the second stage and use the RFM characteristics directly as input for clustering algorithms.


RFM Customer Segmentation Model
RFM Customer Segmentation Model
RECENCY
RFM
FREQUENCY
MONETARY

Recency is simply the amount of time since the customer's most recent transaction (days, months, weeks, hours).

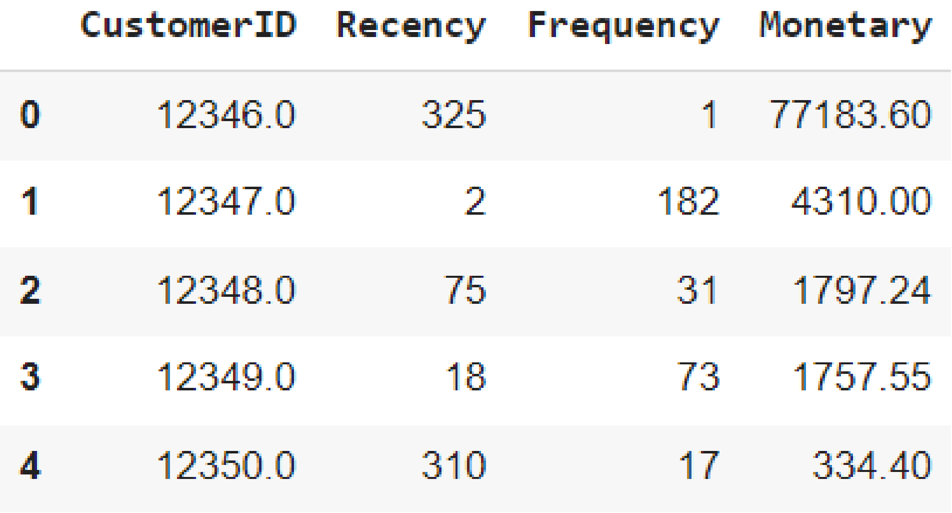
Frequency is the total number of transactions made by the customer (during a defined period).

Monetary is the total amount that the customer has spent across all transactions (during a defined period).

- Performing RFM segmentation, step by step:

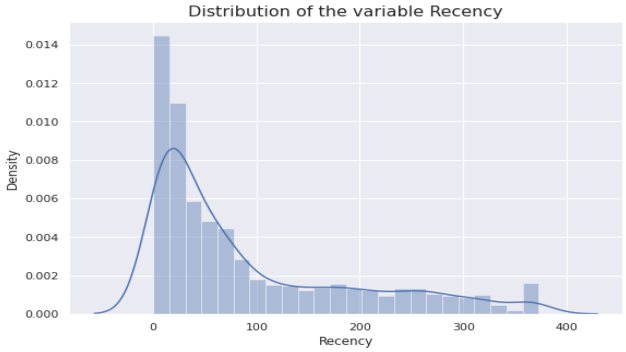
- **Step 1:** The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer. The necessary raw data can be assembled in an Excel spreadsheet or database, and should be easily accessible through the company's transactional or CRM databases.
- **Step 2:** The second step involves using various methods to divide the client list into tiered groups for each of the three dimensions (R, F, M). It is recommended to divide the customers into four tiers for each dimension, unless utilizing specialized software, so that each customer is assigned to one tier in each dimension.
- **Step 3:** In the third step, customer groups will be chosen based on the RFM segments in which they appear to receive particular types of communications.
- **Step 4:** The fourth step actually goes beyond RFM segmentation by creating custom messages for each consumer subgroup. RFM marketing enables marketers to communicate with customers in a far more successful way by focusing on the behavioral patterns of specific groups.

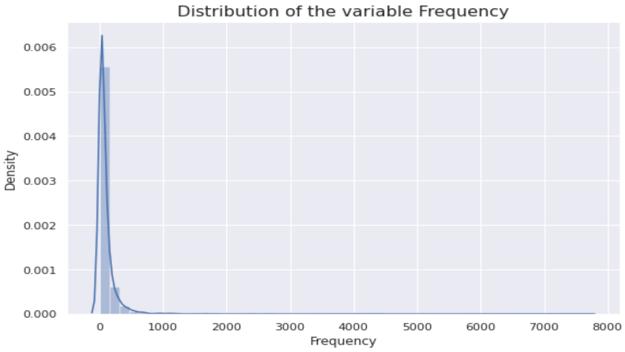
- We calculated the RFM scores and then the RFM rank for each customer as shown below:

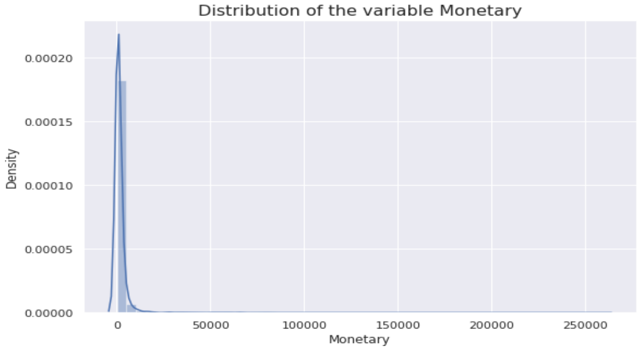


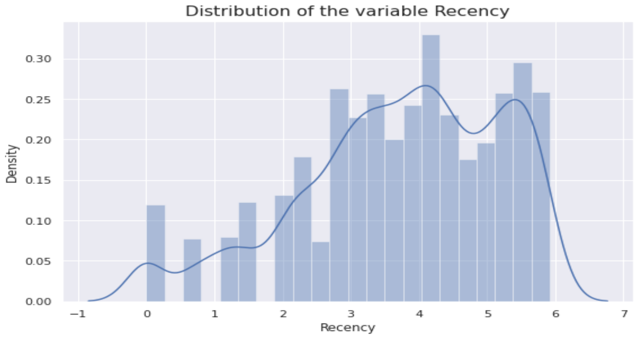
	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	18	73	1757.55
4	12350.0	310	17	334.40

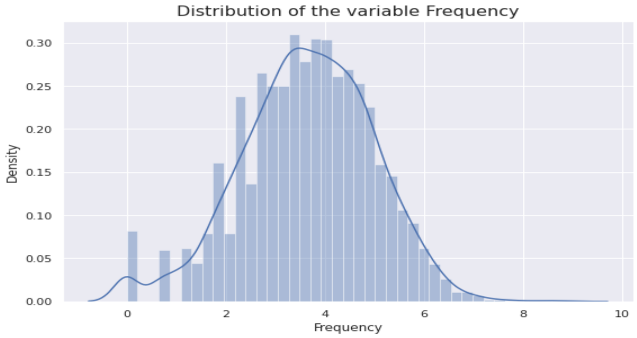
- The distribution of Recency, Frequency and Monetary is again right skewed and log transformation was applied to bring their distribution near normal.

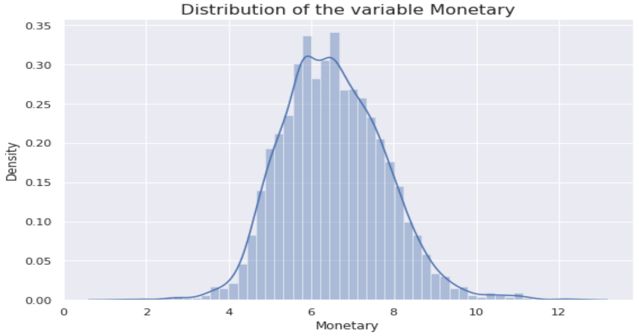












Clustering Models:

The most significant unsupervised learning problem is clustering, which, like all other problems of this type, involves identifying a structure in a set of unlabeled data. "The process of grouping objects into groups whose members are related in some way" could be a broad definition of clustering. Therefore, a cluster is a collection of things that are dissimilar from the objects in other clusters but comparable to one another.

- **K-Means Clustering:**

  The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It accomplishes this using a simple conception of what the optimal clustering looks like:

  - The arithmetic mean of all the points in the cluster makes up the cluster center.
  - In comparison to other cluster centers, each point is closer to its own cluster center. The K-means model is predicated on these two presumptions. It is presumptively known how many clusters there are. The flat clustering algorithm is yet another name for it. The letter "K" in K-means stands for the number of clusters that the algorithm identified from the data.

  According to this strategy, data points are assigned to the cluster so that the total of their squared distances from the centroid is as small as it can be. It is essential to remember that less variation within clusters results in more identical data points inside the same cluster. To tackle the issue, K-means employs the Expectation-Maximization approach. The closest cluster is chosen by the expectation step, and the centroid of each cluster is determined by the maximizing step.

- **Hierarchical Clustering:**

  By creating a hierarchy, hierarchical clustering determines cluster assignments. Either a top-down or bottom-up technique is used to do this:

  - The bottom-up strategy is agglomerative clustering. Up until every point has been combined into a single cluster, it merges the two that are the most comparable.

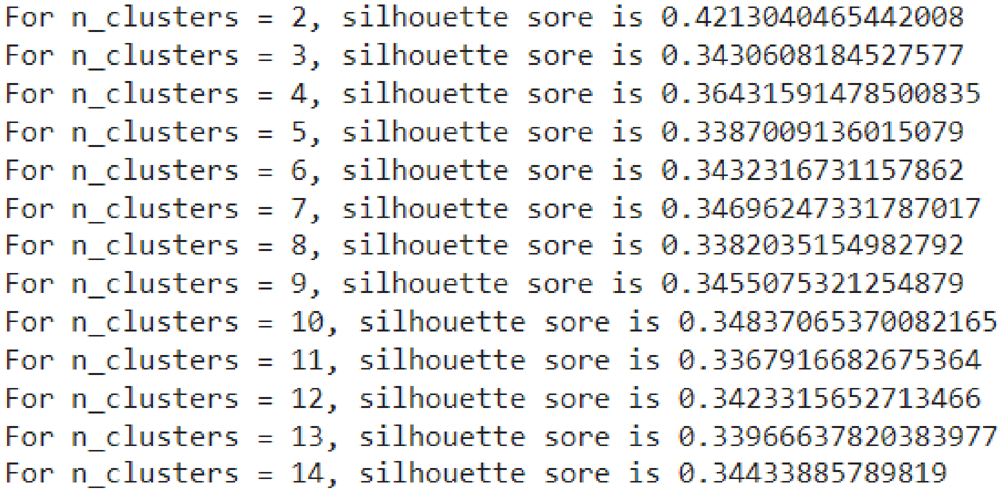
- The top-down strategy is called divisive clustering. It divides the least similar clusters at each stage until just a single data point is left. It starts with all points as a single cluster.

- ## Evaluation Metrics:

  - **Silhouette Method:**
    Silhouette score is used to evaluate the quality of clusters created using clusteringalgorithms such as K-means in terms of how well samples are clustered with othersamples that are similar to each other. The silhouette score is calculated for each sample of different clusters. To calculate the silhouette score for each.

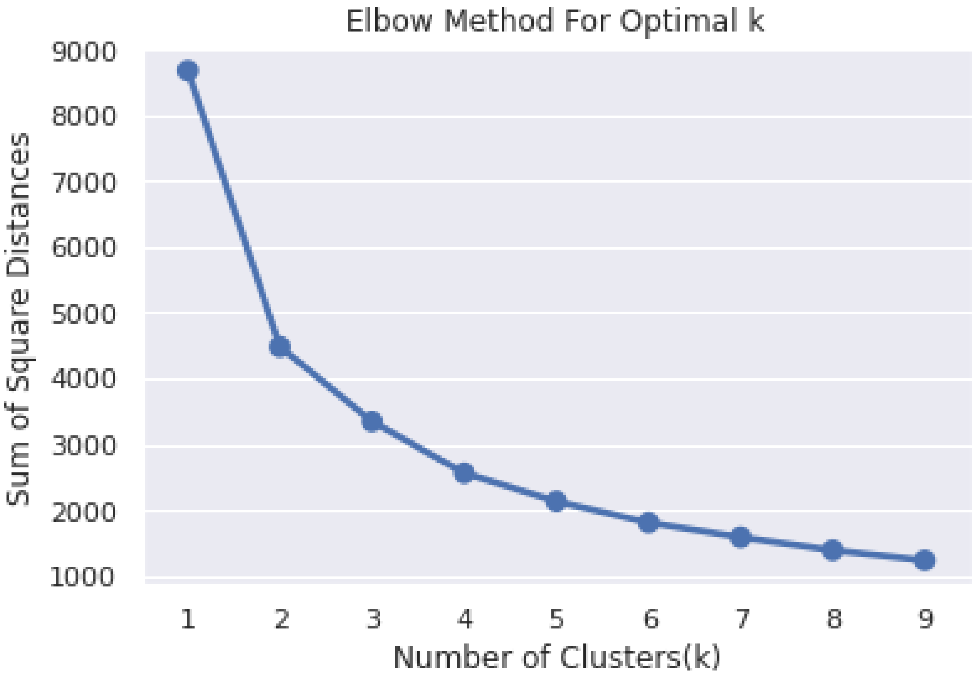
  - **Elbow Method:**
    The KElbowVisualizer implements the elbow method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. Ifthe line chart resembles an arm, then the elbow is a good indication that the underlying model fits best at that point. In the visualizer the elbow will be annotated with a dashed line. By default, the scoring parameter metric is set todistortion, which computes the sum of squared distances from each point to itsassigned center.

  - **Silhouette score method on Recency and Monetary:**

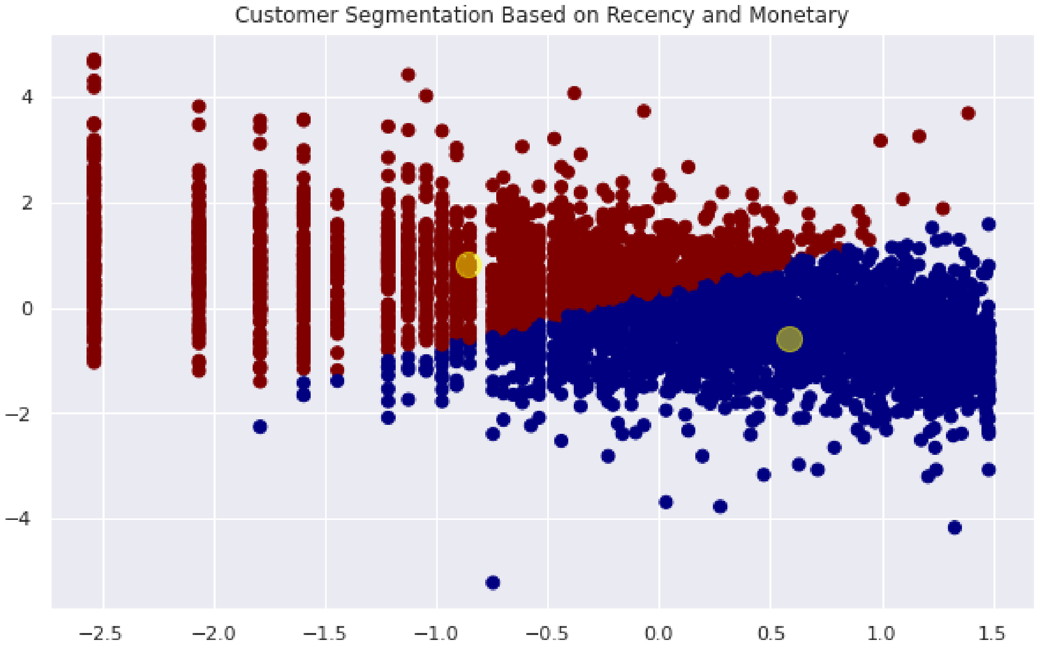


```
For n_clusters = 2, silhouette sore is 0.4213040465442008
For n_clusters = 3, silhouette sore is 0.3430608184527577
For n_clusters = 4, silhouette sore is 0.36431591478500835
For n_clusters = 5, silhouette sore is 0.3387009136015079
For n_clusters = 6, silhouette sore is 0.3432316731157862
For n_clusters = 7, silhouette sore is 0.34696247331787017
For n_clusters = 8, silhouette sore is 0.3382035154982792
For n_clusters = 9, silhouette sore is 0.3455075321254879
For n_clusters = 10, silhouette sore is 0.34837065370082165
For n_clusters = 11, silhouette sore is 0.3367916682675364
For n_clusters = 12, silhouette sore is 0.3423315652713466
For n_clusters = 13, silhouette sore is 0.33966637820383977
For n_clusters = 14, silhouette sore is 0.34433885789819
```

- **Applying Elbow Method on Recency and Monetary:**

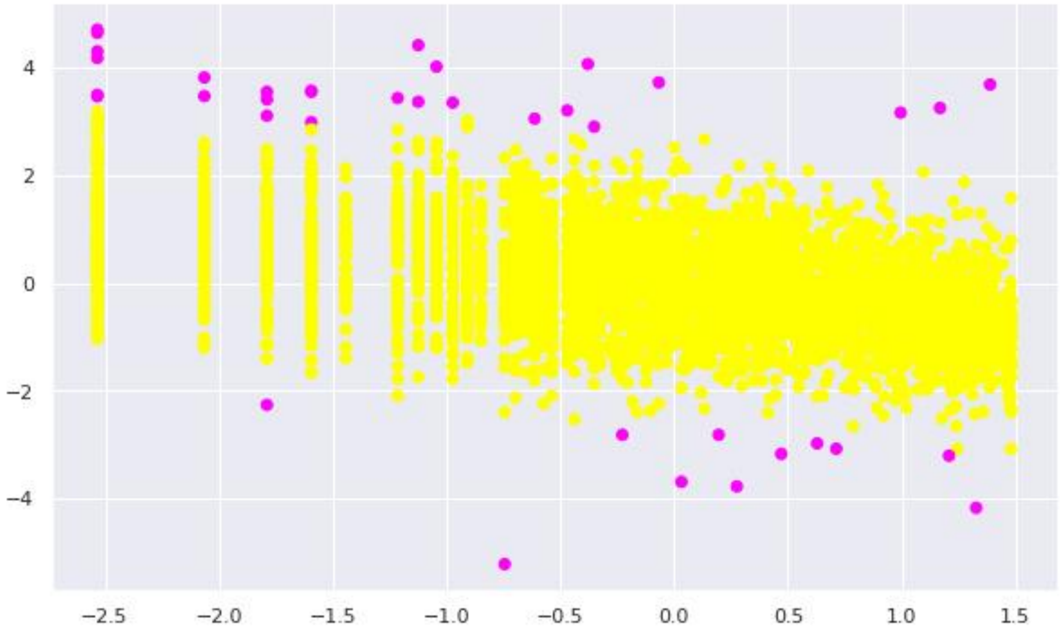


- **Hyperparameter Tuning for the Best Value of K:**

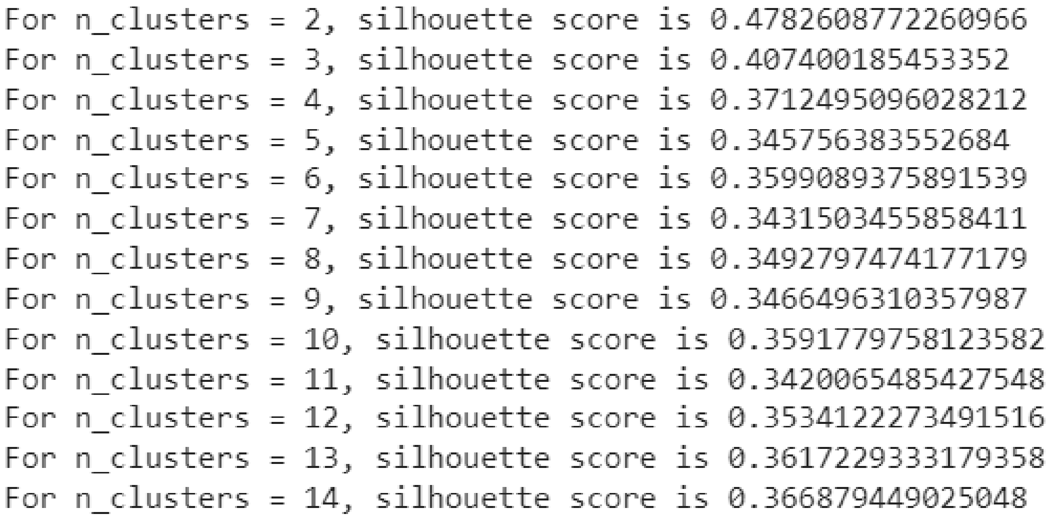


- **Implementation of Density Based Spatial Clustering of Applications with Noise (DBSCAN)**
    - Distance between nearest points.

- **DBSCAN on Recency and Monetary:**

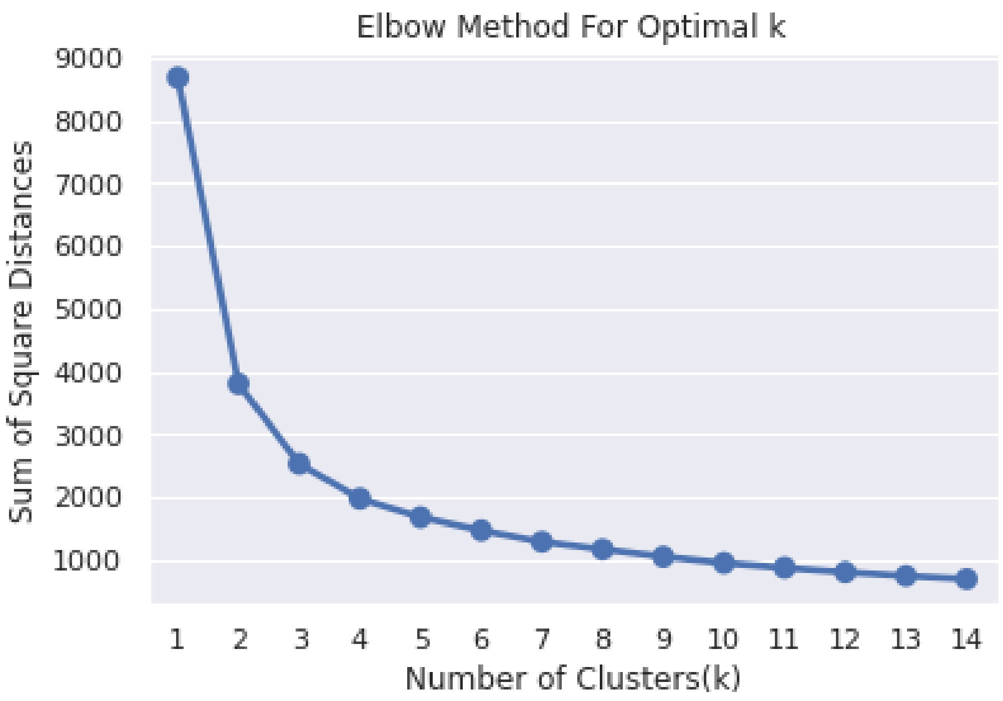


- **Silhouette score method on Frequency and Monetary:**

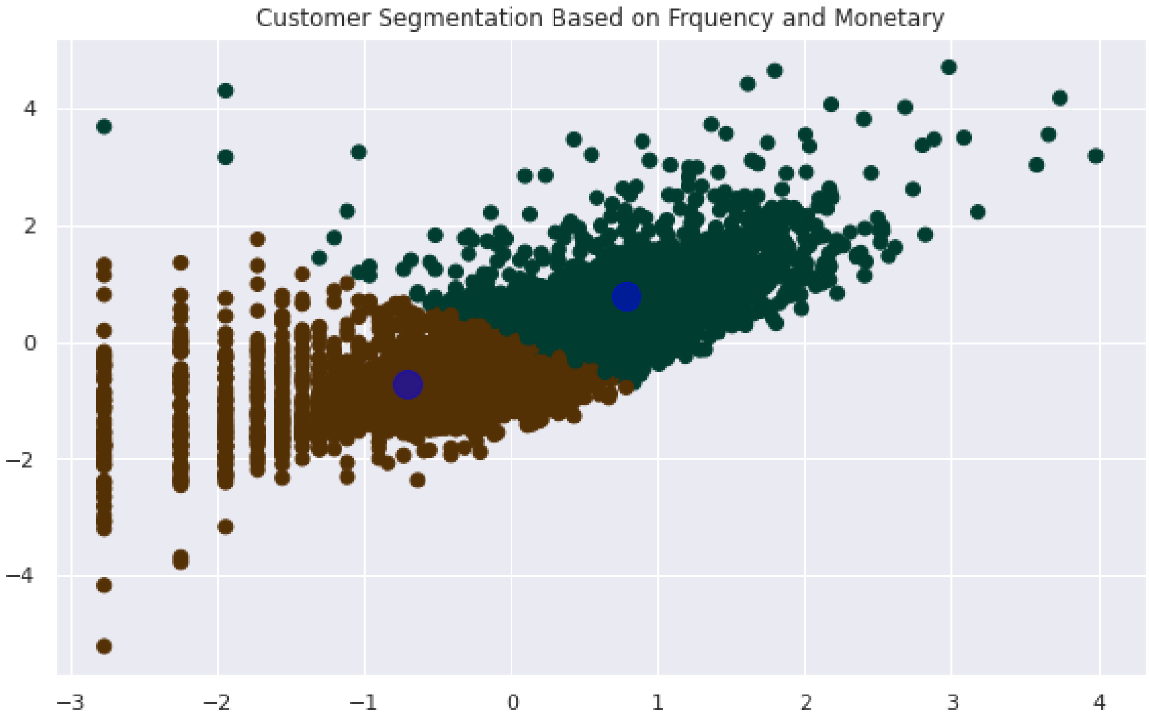


```
For n_clusters = 2, silhouette score is 0.4782608772260966
For n_clusters = 3, silhouette score is 0.407400185453352
For n_clusters = 4, silhouette score is 0.3712495096028212
For n_clusters = 5, silhouette score is 0.345756383552684
For n_clusters = 6, silhouette score is 0.3599089375891539
For n_clusters = 7, silhouette score is 0.3431503455858411
For n_clusters = 8, silhouette score is 0.3492797474177179
For n_clusters = 9, silhouette score is 0.3466496310357987
For n_clusters = 10, silhouette score is 0.3591779758123582
For n_clusters = 11, silhouette score is 0.3420065485427548
For n_clusters = 12, silhouette score is 0.3534122273491516
For n_clusters = 13, silhouette score is 0.3617229333179358
For n_clusters = 14, silhouette score is 0.366879449025048
```

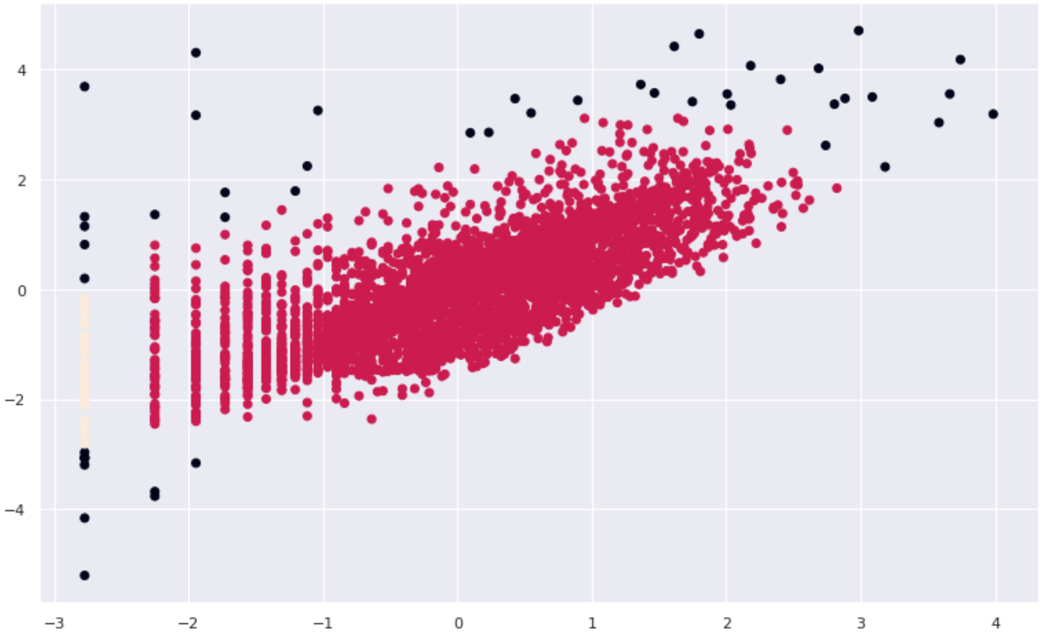
- **Applying elbow method on Frequency and Monetary:**



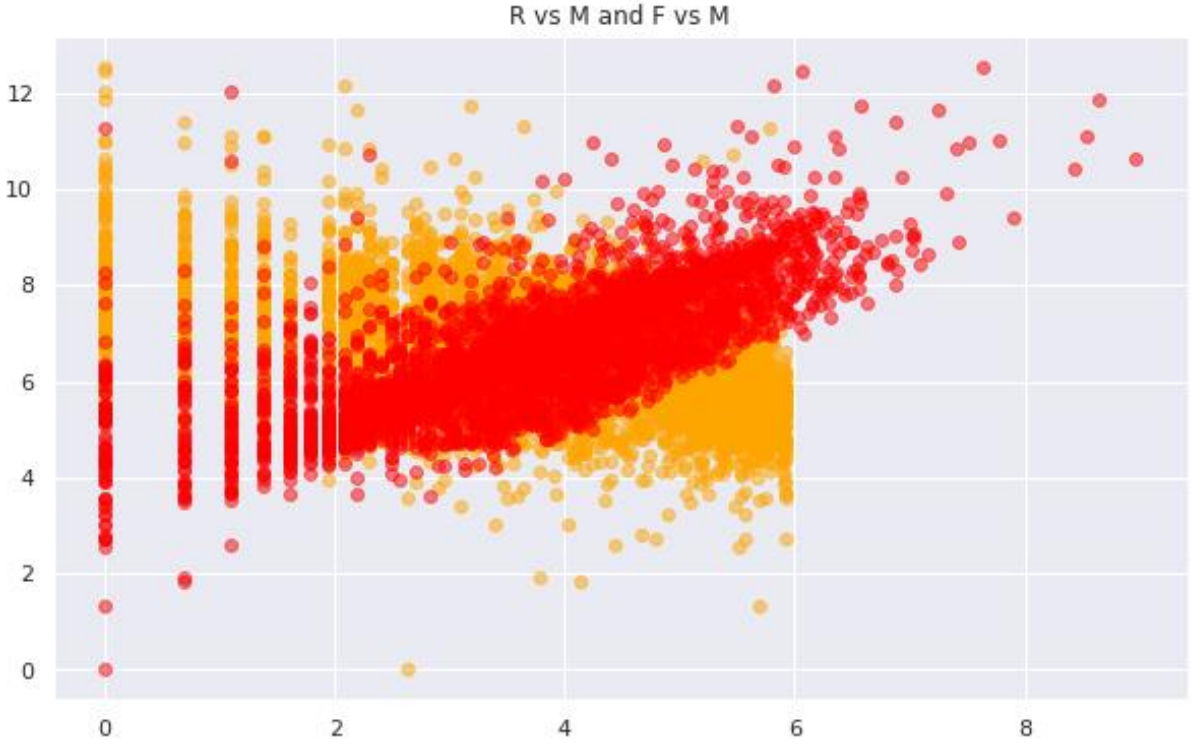
- **Hyperparameter Tuning for Best Value of K:**



- **DBSCAN on Frequency and Monetary:**



- **Comparison between R vs M and F vs M**



Challenges:

- There were many duplicated records and missing values present.
- The transactional information included cancelled orders, and a small number of records had 0 as the unit price, which isn't possible in real life.
- No feature was distributed normally.
- Features had to be created to calculate RFM scores.
- Finding the optimal number of clusters through silhouette method or elbowmethod.

Conclusion:

- We performed consumer segmentation using a variety of steps throughout the analysis. Starting with data wrangling, we tried to deal with duplicates, null values, and feature updates. We then performed some exploratory data analysis in an effort to derive observations from the dataset's features.
- Then, for each of the consumers, we developed some quantitative components, such as recency, frequency, and monetary data, known as the RFM model. On these features, we applied the KMeans clustering algorithm. To determine the ideal number of clusters, which was 2, we also performed silhouette and elbow method analyses.
- Customers with low frequency and high value transactions were part of one cluster, while those with low frequency and high value transactions were part of another cluster.
- There may be other adjustments made to this analysis, though. Depending on the goals and preferences of the firm, one may decide to cluster into a greater number. After clustering, the tagged feature can be put into supervised machine learning algorithms for classification that can forecast the classes for fresh sets of observations.
- The clustering can also be done on a new set of features, such segmenting customers based on the times of their visits, determining customer lifetime value (CLV), and many more.

References:

- Analytics Vidhya
- Machine Learning Mastery
- Scikit-learn
- Towards Data Science