Capstone Project 4

Customer Segmentation

Team Member

Suraj Kumar

Shreya Ranjan

CUSTOMER SEGEMENTATION:

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.

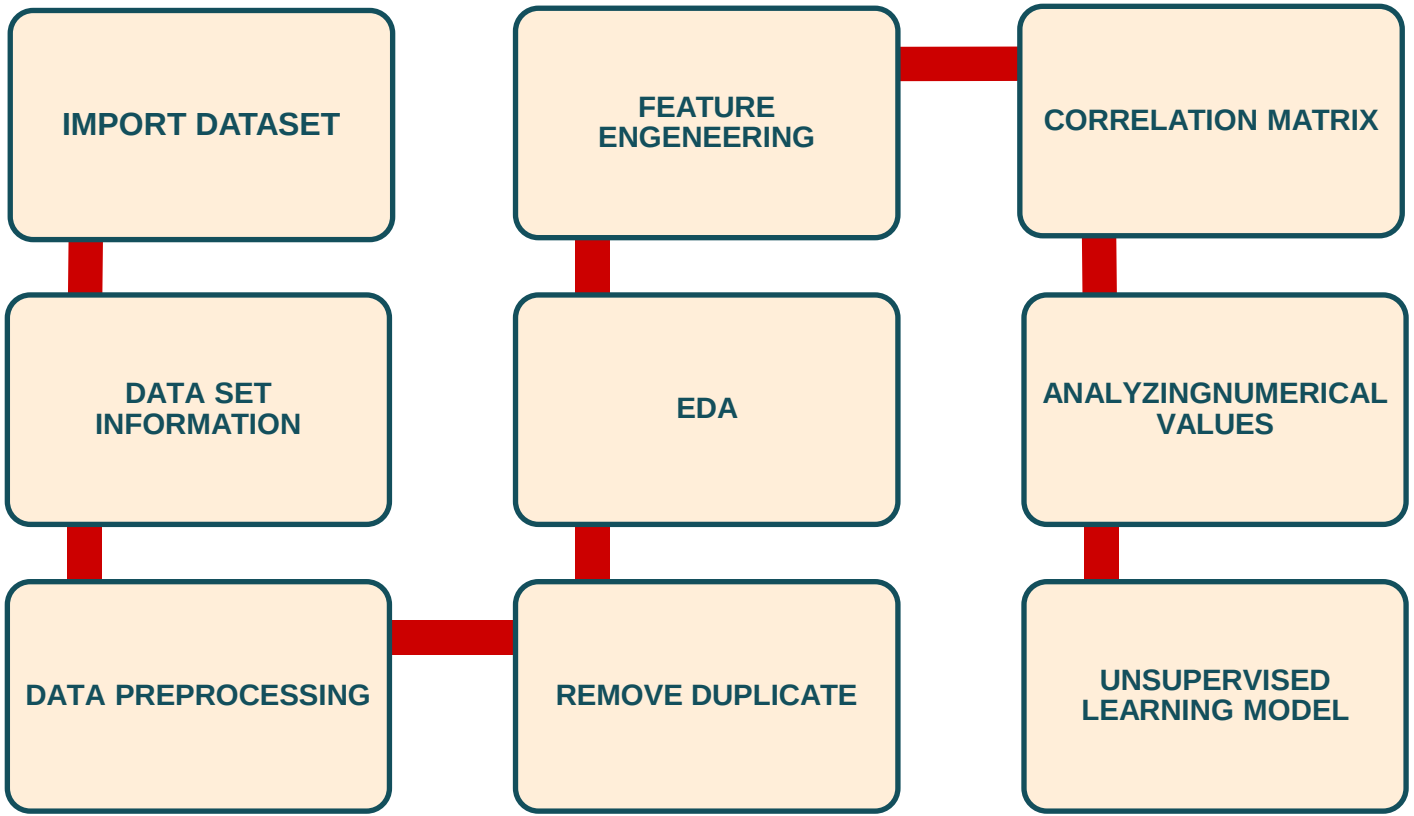
Problem description:

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data Set:

1}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.".
2}Stock Code: Product (item) code. A 5-digit integral number known as the nominal is assigned to each unique product.
3}Description: Name of the Product (Item). Nominal.
4}Quantity: The number of each item (product) in each transaction. Numeric.
5}Invoice Date: Invoice Time and date. The day and time that each transaction was created, represented by a number.
6}Unit Price: Unit pricing. Numeric, Sterling unit price for the product.
7}CustomerID: Customer number. Nominal, a five-digit integral number assigned to every customer uniquely.
8}Country: Country name. Nominal, the name of the country in which each customer resides.

Data Pipeline



Libraries:

1}Pandas
2}Numpy
3}Seaborn
4}Matplotlib
5}Datetime
6}Sklearn
7}Scripy
8}Pretty table

Data Wrangling:

Data Set

```
Retail_Data.shape
```

```
(541909, 8)
```

Unique Data

```
Retail_Data.nunique()
```

```
InvoiceNo      25900
StockCode       4070
Description     4223
Quantity         722
InvoiceDate    23260
UnitPrice       1630
CustomerID      4372
Country           38
dtype: int64
```

Data Describe

```
Retail_Data.describe()
```

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Data Info

```
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
```

Data Wrangling(cont..)

Data Head

```
Retail_Data.head(10)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/10 8:26	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/10 8:26	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	12/1/10 8:28	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/10 8:28	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/10 8:34	1.69	13047.0	United Kingdom

Data wrangling(cont..)

Data Tail

```
Retail_Data.tail(10)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
541899	581587	22726	ALARM CLOCK BAKELIKE GREEN	4	12/9/11 12:50	3.75	12680.0	France
541900	581587	22730	ALARM CLOCK BAKELIKE IVORY	4	12/9/11 12:50	3.75	12680.0	France
541901	581587	22367	CHILDRENS APRON SPACEBOY DESIGN	8	12/9/11 12:50	1.95	12680.0	France
541902	581587	22629	SPACEBOY LUNCH BOX	12	12/9/11 12:50	1.95	12680.0	France
541903	581587	23256	CHILDRENS CUTLERY SPACEBOY	4	12/9/11 12:50	4.15	12680.0	France
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/11 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/11 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/11 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/11 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/11 12:50	4.95	12680.0	France

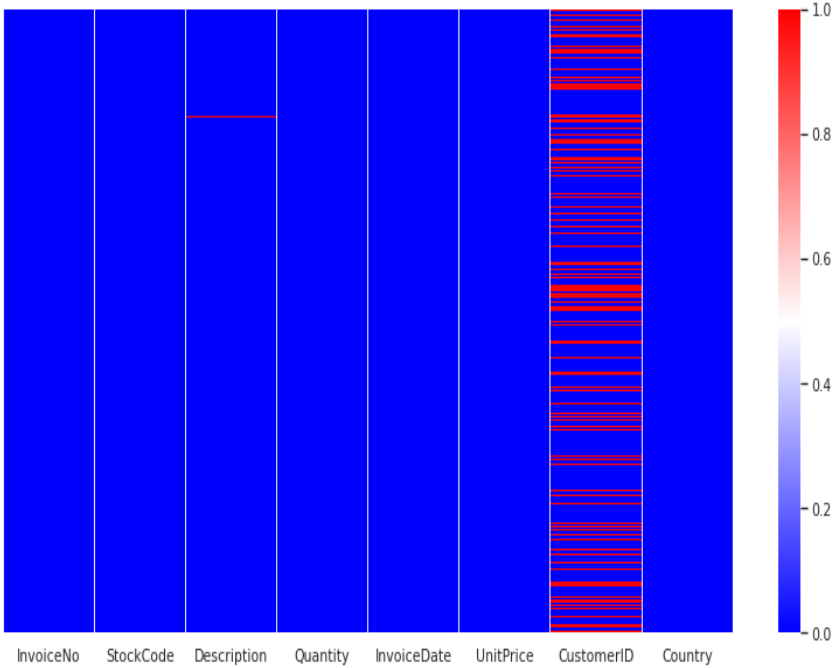
Data Pre-processing:

Null values

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

Null values present in **Description** and **CustomerID** column includes exactly 1454 and 135080 null values.

Heat Map



Shining red lines indicate null values.

Data Pre-processing:

Data info after dropping null values.

```
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    406829 non-null  object
 1   StockCode    406829 non-null  object
 2   Description  406829 non-null  object
 3   Quantity     406829 non-null  int64
 4   InvoiceDate  406829 non-null  object
 5   UnitPrice    406829 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      406829 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 27.9+ MB
```

Duplicate data:

	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

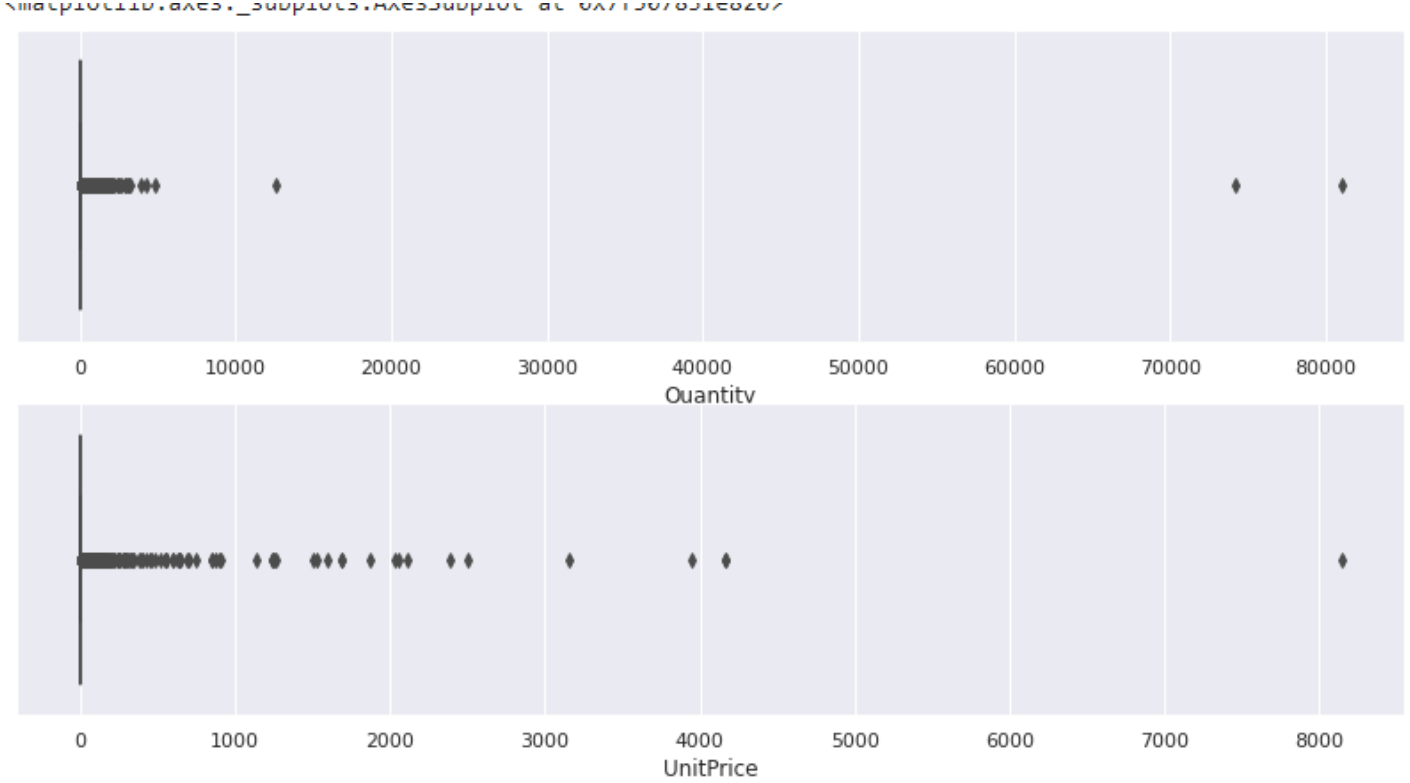
Dropping duplicate values:

After removing the duplicate values the shape of the dataset changes to(401604, 8)

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/11 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/11 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/11 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/11 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/11 12:50	4.95	12680.0	France

401604 rows × 8 columns

Outliners:


0
10000
20000
30000
40000
50000
60000
70000
80000
Quantity
0
1000
2000
3000
4000
5000
6000
7000
8000
UnitPrice

Exploratory data analysis(EDA):

Top 5 product based on maximum selling are :

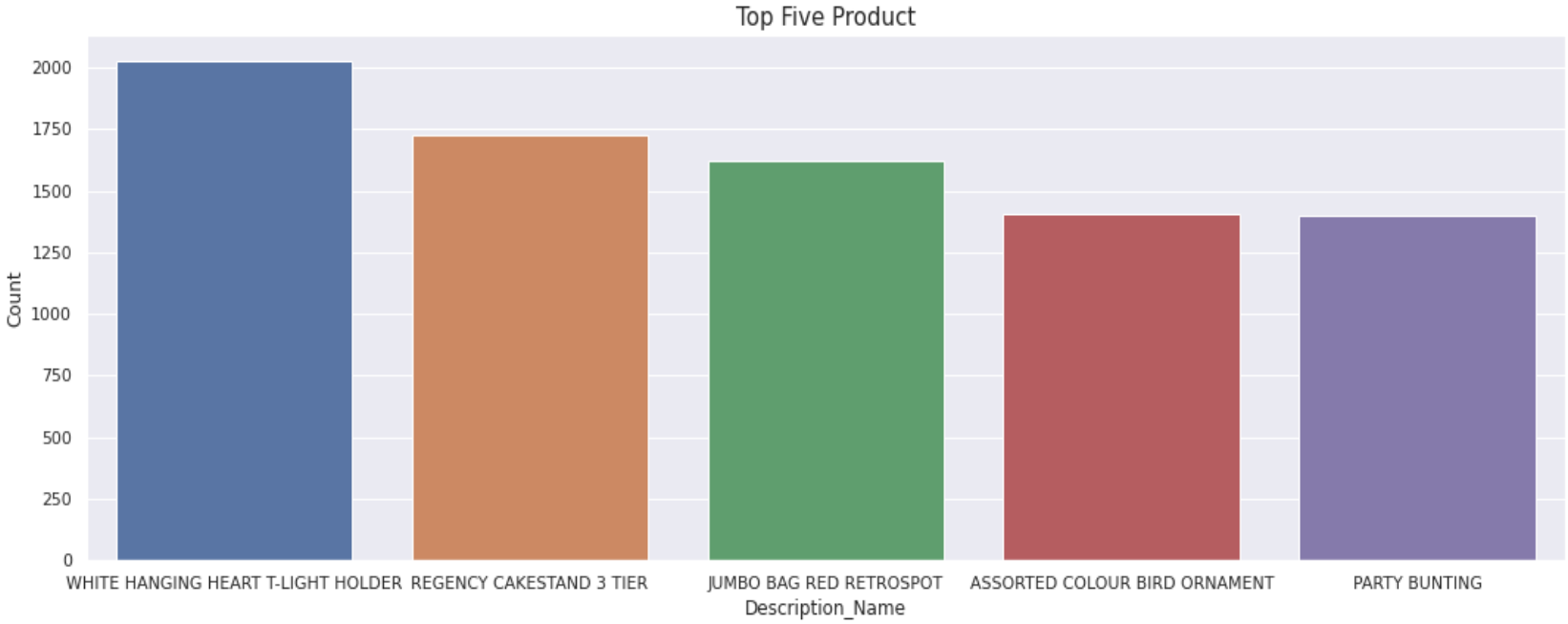
- WHITE HANGING HEART T-LIGHT HOLDER,
- REGENCY CAKESTAND 3 TIER
- JUMBO BAG RED RETROSPOT
- PARTY BUNTING
- LUNCH BAG RED RETROSPOT

Bottom 5 Product based on the selling are

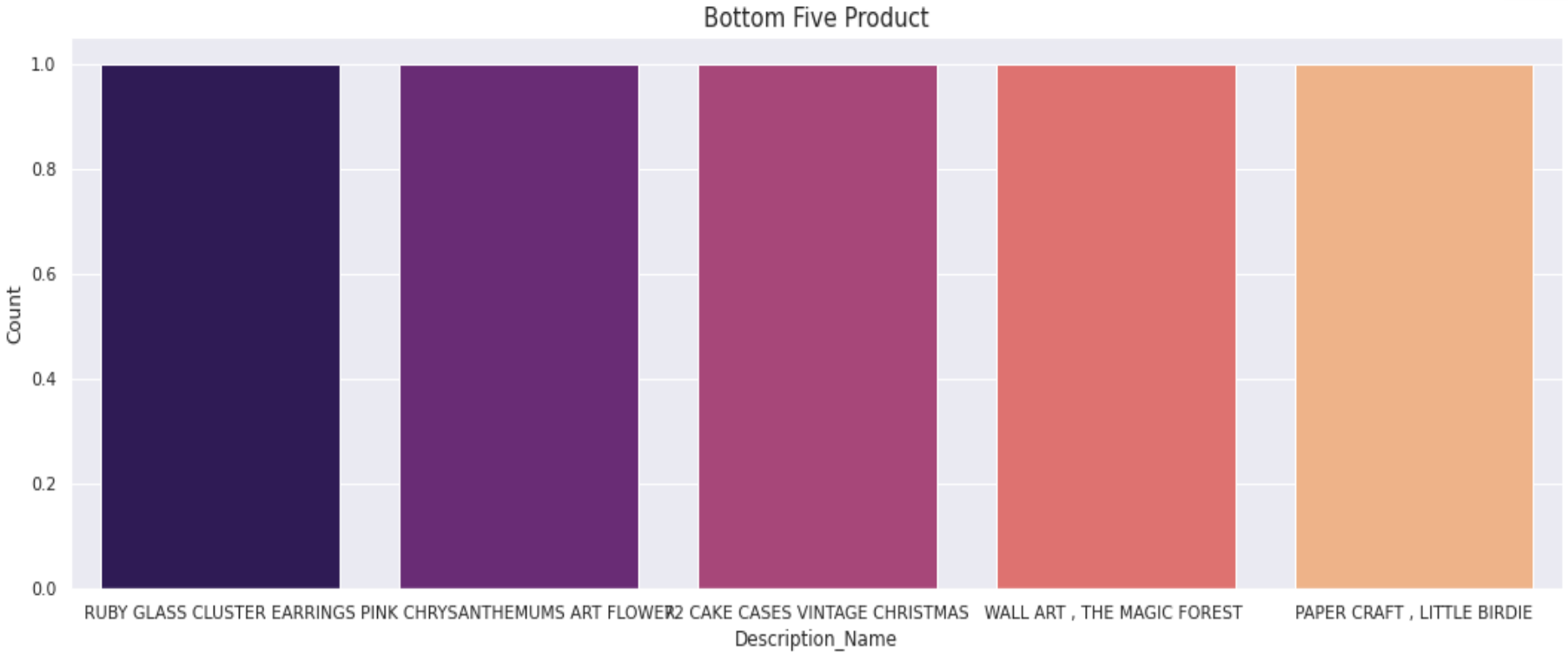
- RUBY GLASS CLUSTER EARRINGS
- PINK CHRYSANTHEMUMS ART FLOWER
- 72 CAKE CASES VINTAGE CHRISTMAS
- WALL ART , THE MAGIC FOREST
- PAPER CRAFT , LITTLE BIRDIE

Visualization of Description:

Top 5 product:



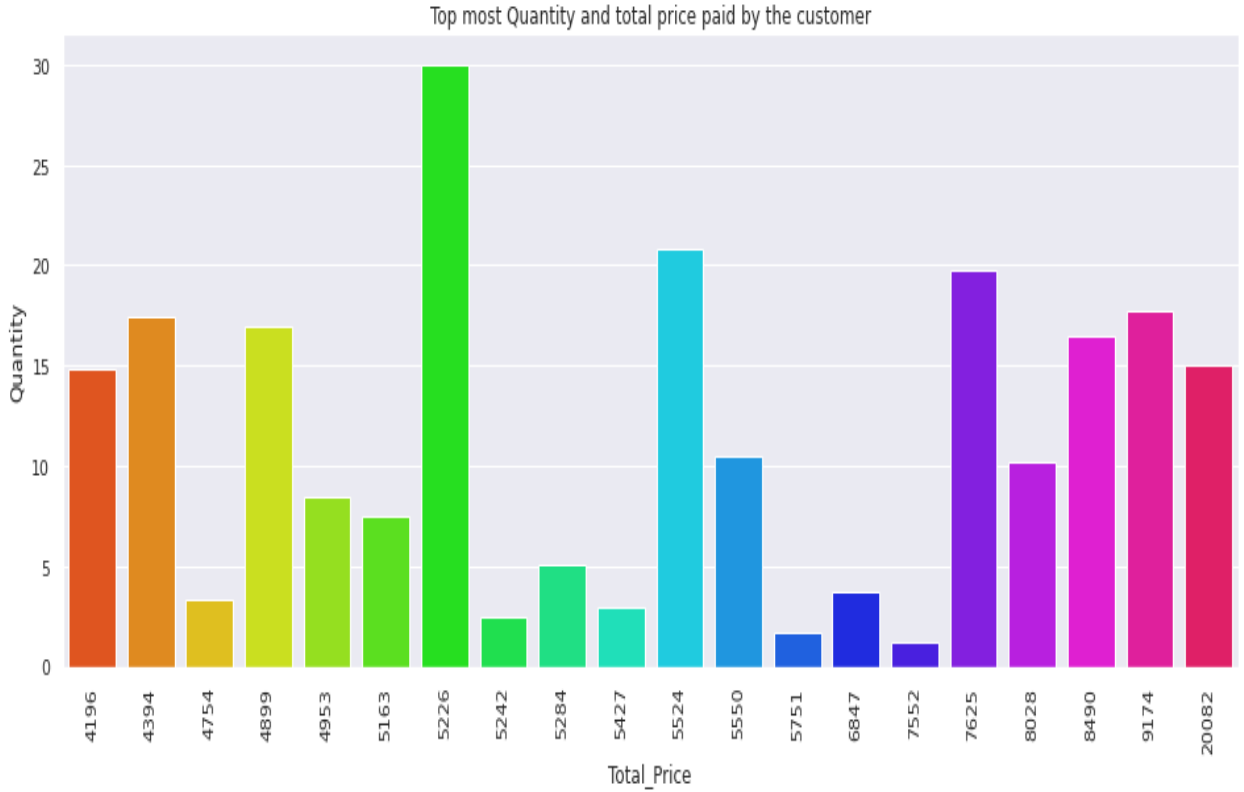
Bottom 5 product:



New column with the Total_Price paid by the customer:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Total_Price
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	United Kingdom	20.34

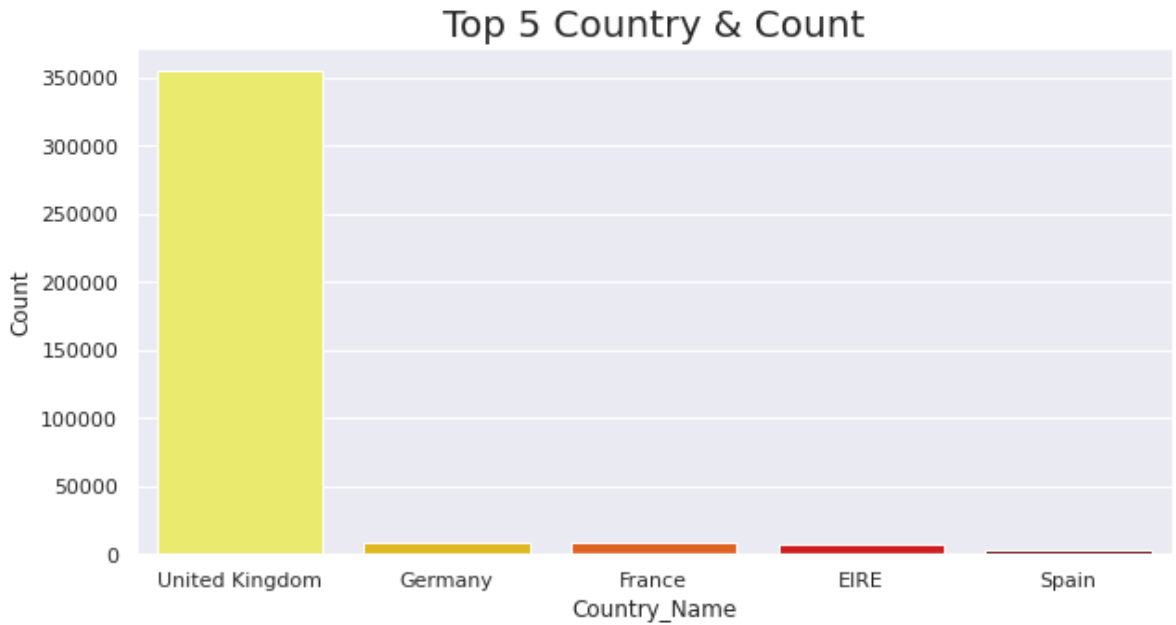
Top most Quantity and total price paid by the customer:



	Quantity	Total_Price
0	15.00	20082
1	17.70	9174
2	16.50	8490
3	10.20	8028
4	19.80	7625
5	1.25	7552
6	3.75	6847
7	1.65	5751
8	10.50	5550
9	20.80	5524
10	2.95	5427
11	5.04	5284
12	2.50	5242
13	30.00	5226
14	7.50	5163
15	8.50	4953
16	17.00	4899
17	3.30	4754
18	17.40	4394
19	14.85	4196

Country Analysis:

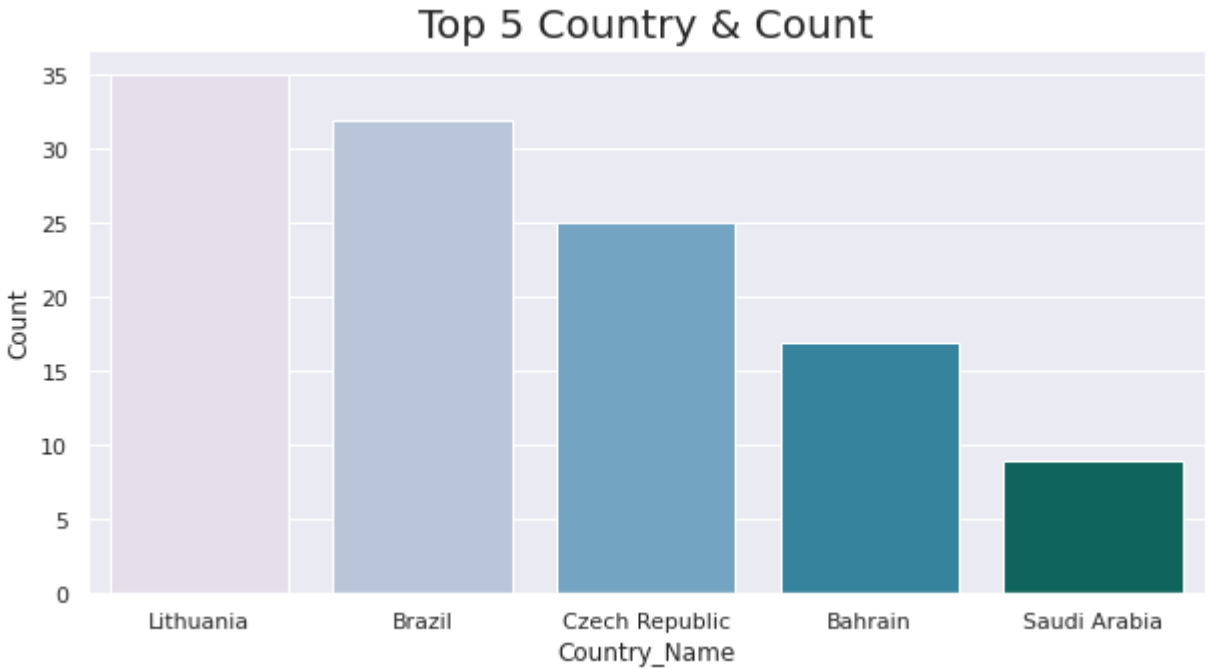
5 Highest transactions Countries:



	Country_Name	Count
0	United Kingdom	354345
1	Germany	9042
2	France	8342
3	EIRE	7238
4	Spain	2485

From the above graph, it is clear that the United Kingdom has more transactions than other nations, indicating that it has a larger likelihood of making a purchase than Germany, France, Ireland, and Spain.

The Lowest 5 countries

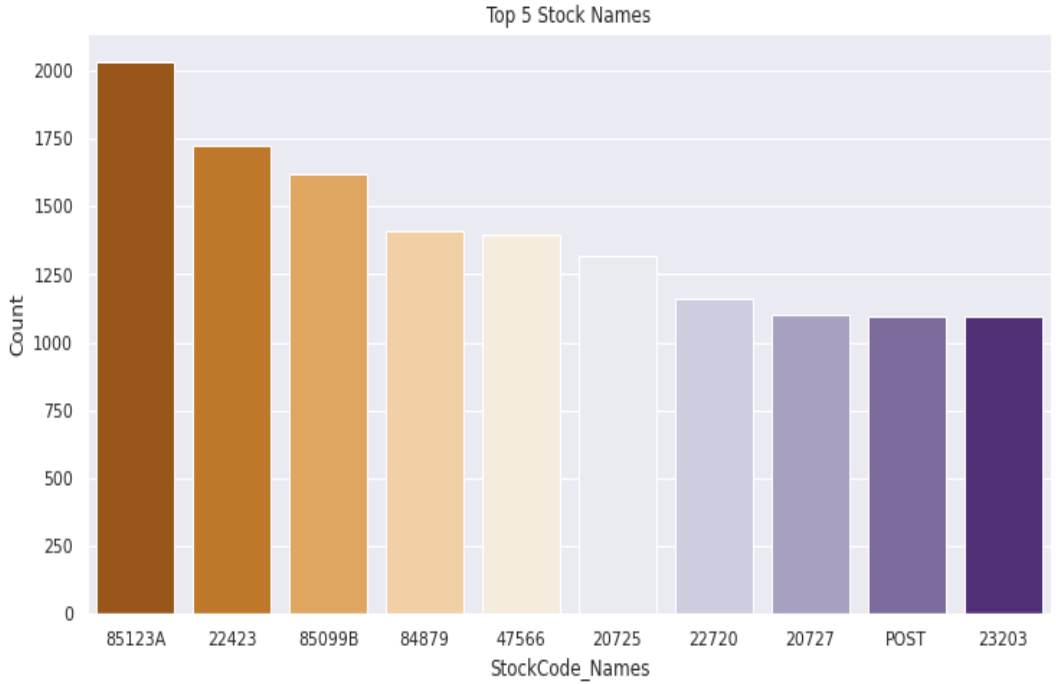


	Country_Name	Count
32	Lithuania	35
33	Brazil	32
34	Czech Republic	25
35	Bahrain	17
36	Saudi Arabia	9

The Saudi Arabia has the lowest purchasing history, so we won't concentrate more on these five nations while analyzing the consumer.

Stock Analysis:

StockCodes_Name with highest selling:

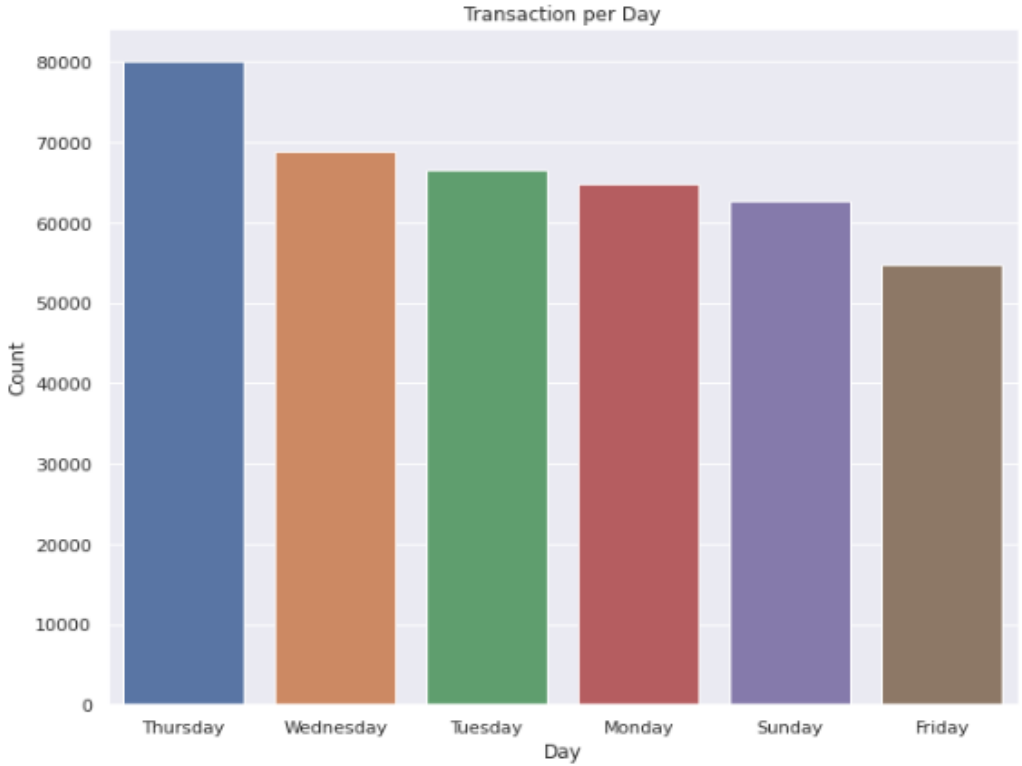


	StockCode_Names	Count
0	85123A	2035
1	22423	1724
2	85099B	1618
3	84879	1408
4	47566	1397
5	20725	1317
6	22720	1159
7	20727	1105
8	POST	1099
9	23203	1098

Top 10 Stock name based on selling

Feature engineering

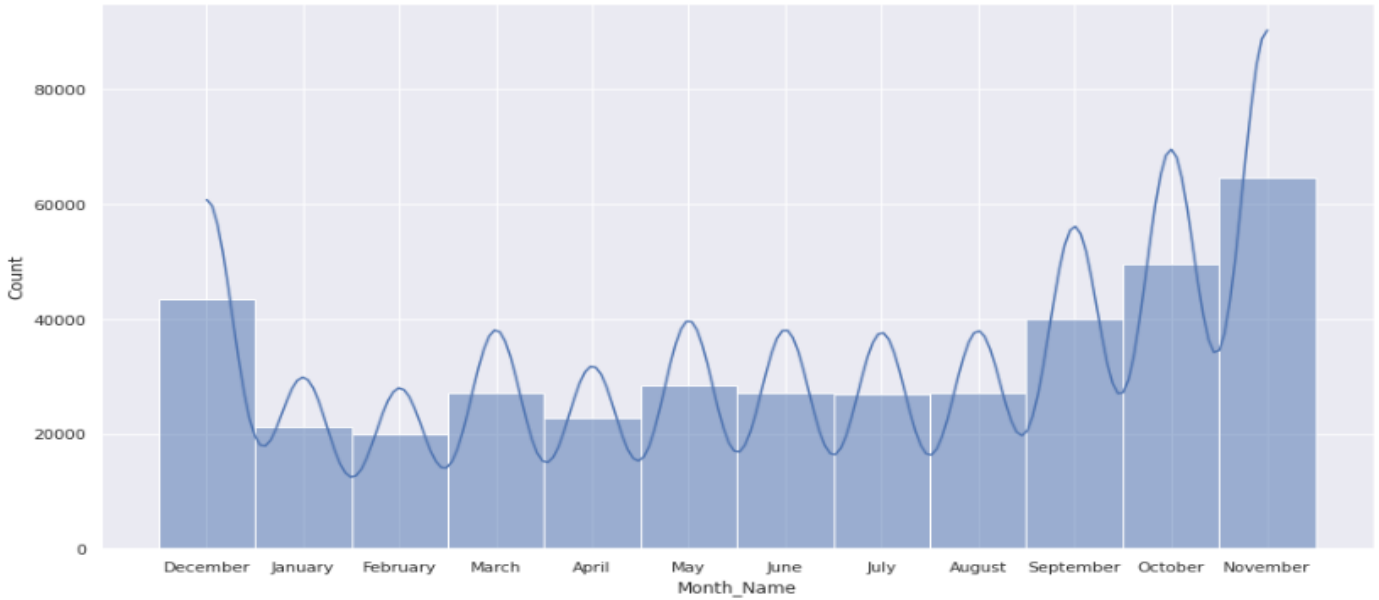
Transaction per day:



	Day	Count
0	Thursday	80052
1	Wednesday	68888
2	Tuesday	66476
3	Monday	64899
4	Sunday	62775
5	Friday	54834

From the above graph, it is clear that Thursday has higher transaction volume than the other days, possibly because clients are more available, and Friday has lower transaction volume.

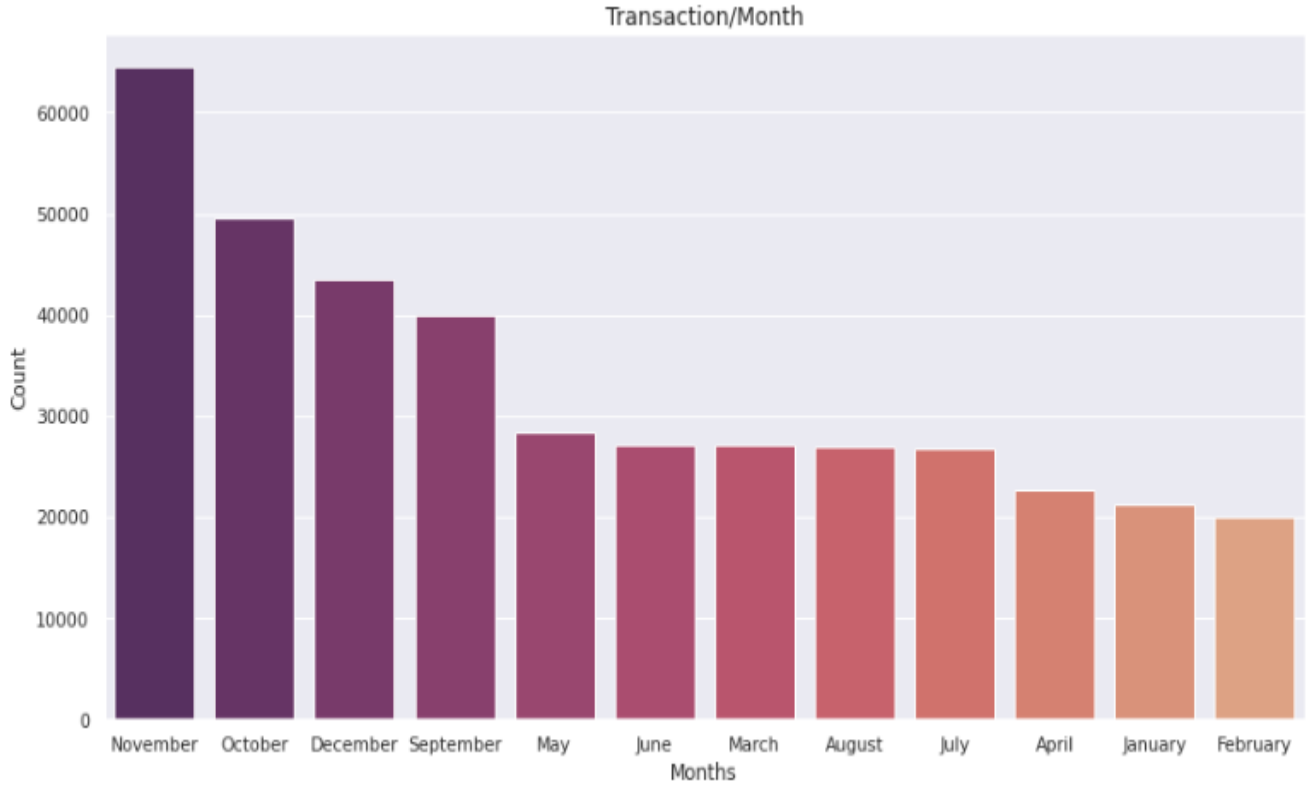
Purchasing Stats:



The above graph makes it clearly transparent that buyers are purchasing winter clothing and possibly even Diwali decorations in November, which is winter season.

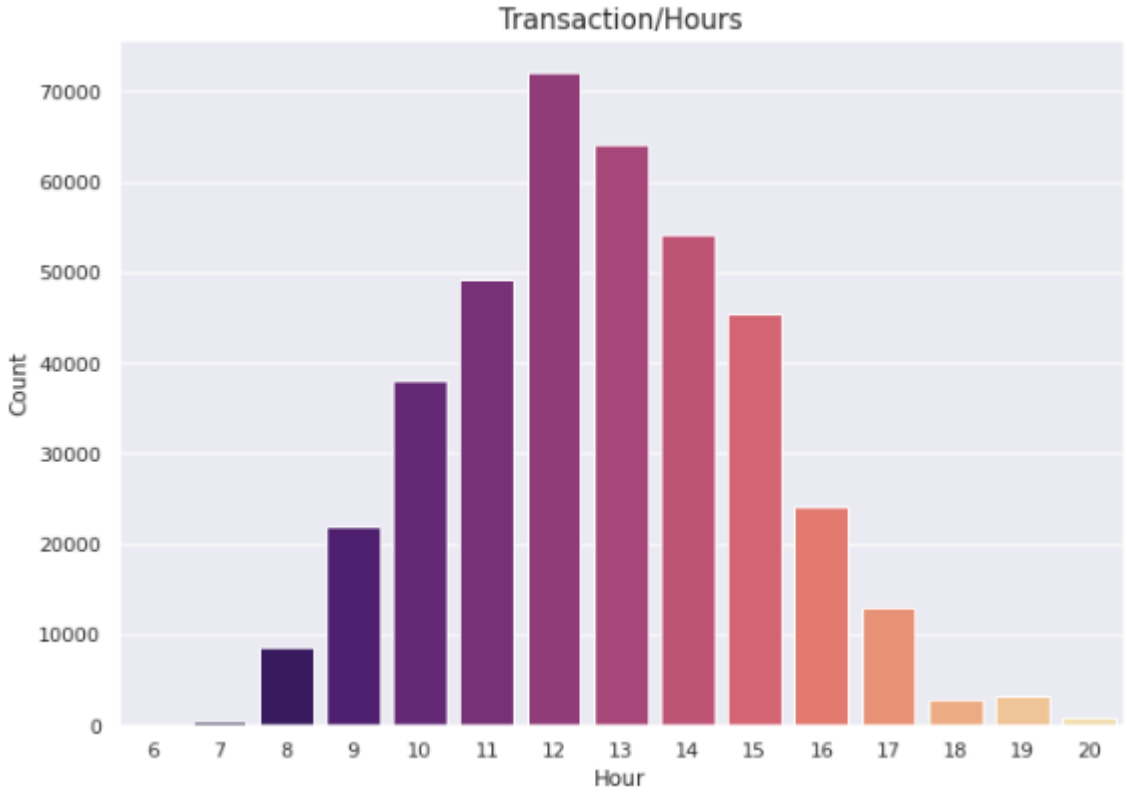
Given that customers make fewer purchases in January and February and that the winter season is coming to an end, it is obvious that customers make more purchases during the winter.

Transaction Per Month:



Months	Count
November	64545
October	49557
December	43464
September	40030
May	28322
June	27185
March	27177
August	27013
July	26827
April	22644
January	21232
February	19928

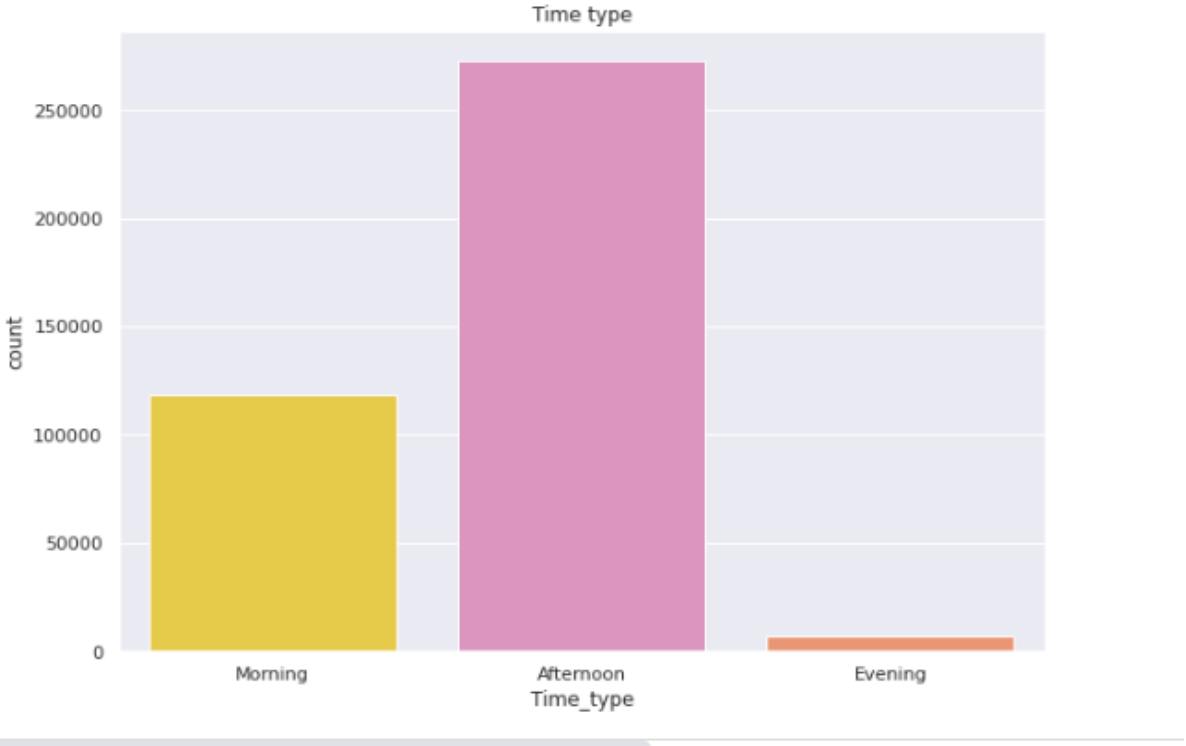
Transaction Per Hours:



	Hour	Count
0	12	72069
1	13	64031
2	14	54127
3	11	49092
4	15	45372
5	10	37999
6	16	24093
7	9	21945
8	17	13072
9	8	8691
10	19	3322
11	18	2929
12	20	802
13	7	379
14	6	1

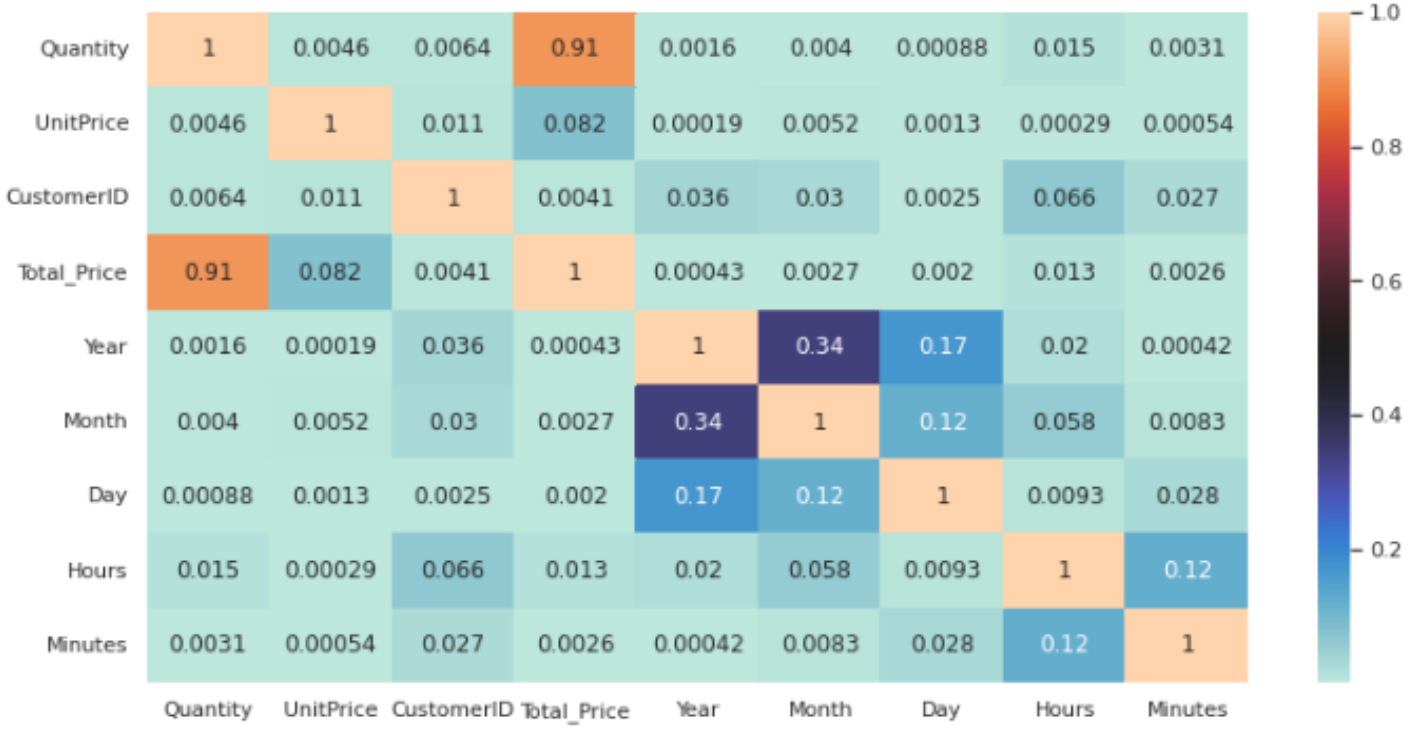
According to hour, people do most of their shopping between the hours of 11 and 4, therefore practically all of them are free at this time.

Distributing the day in Morning Afternoon and Evening



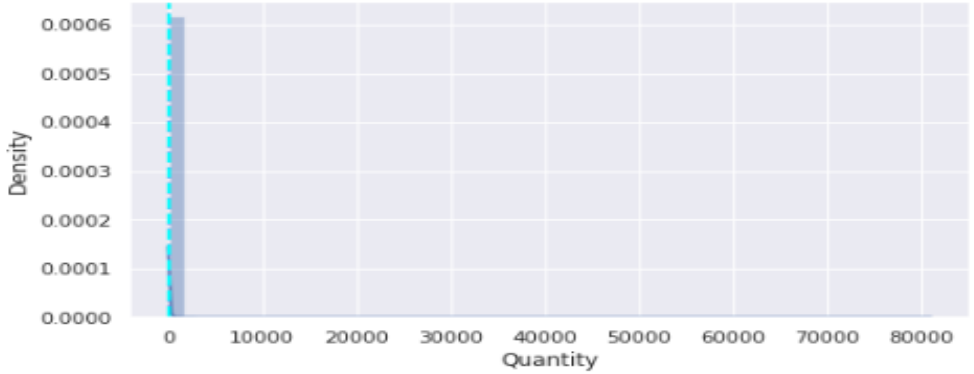
Most customers purchase things in the afternoon, followed by modest numbers of customers in the morning, and the smallest numbers of customers in the evening.

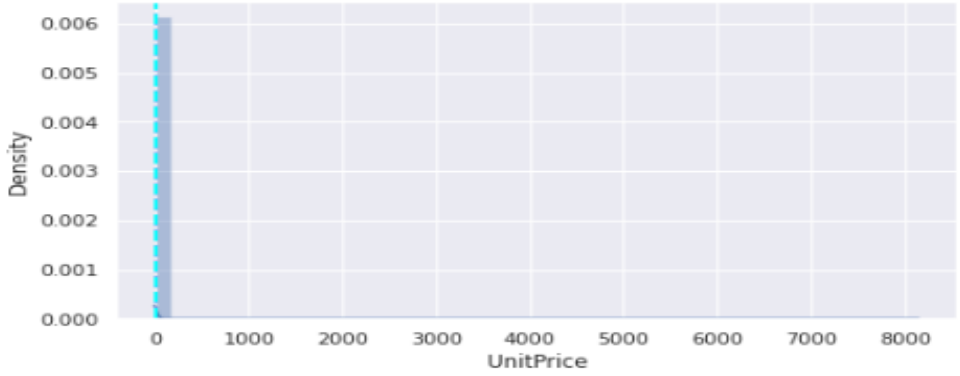
Correlation matrix:

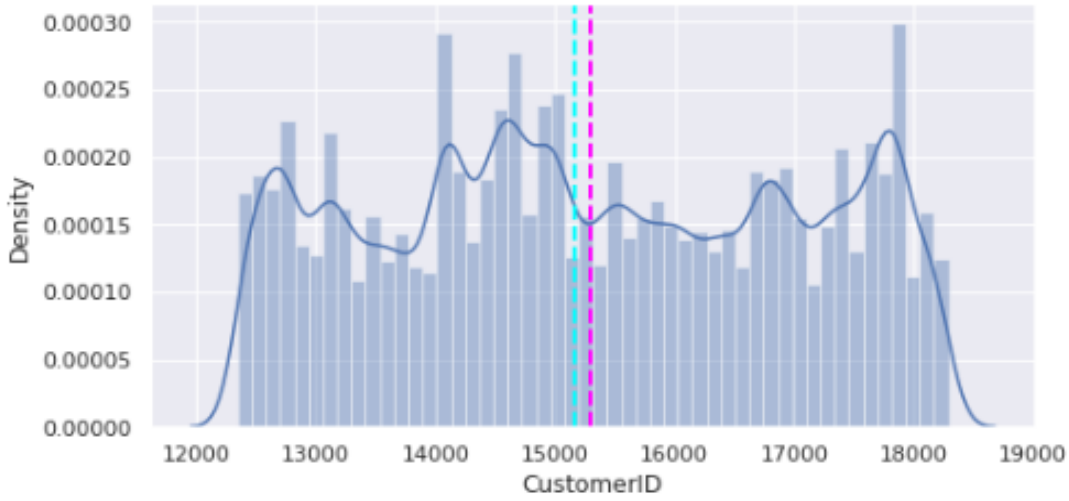


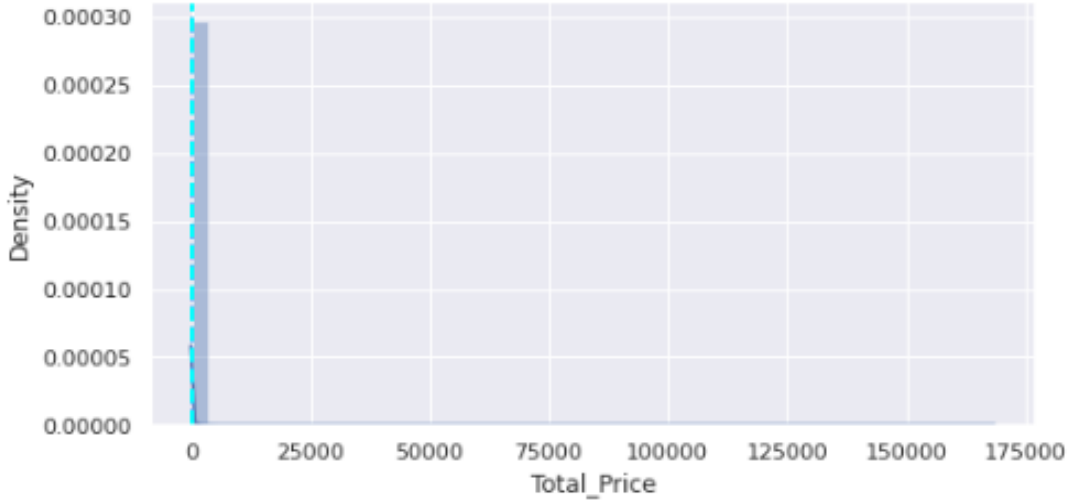
Numerical Features:

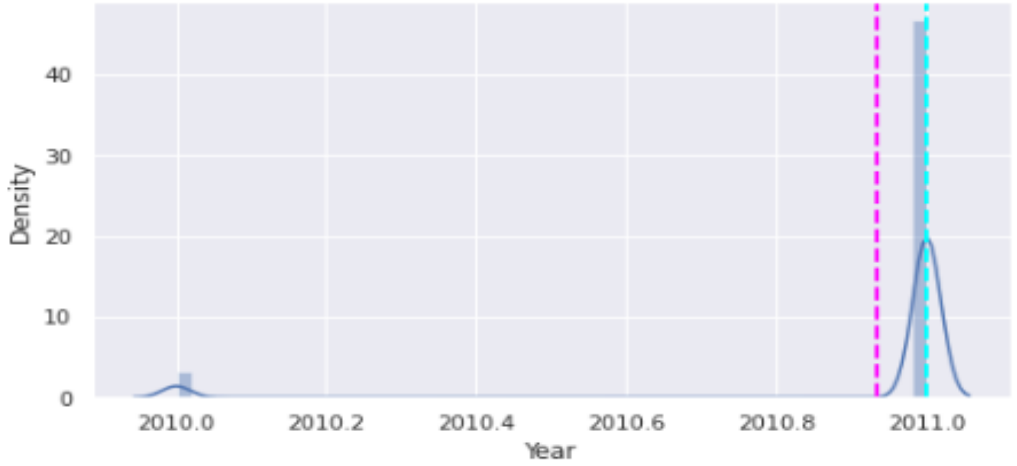
Distribution of all numerical variables:

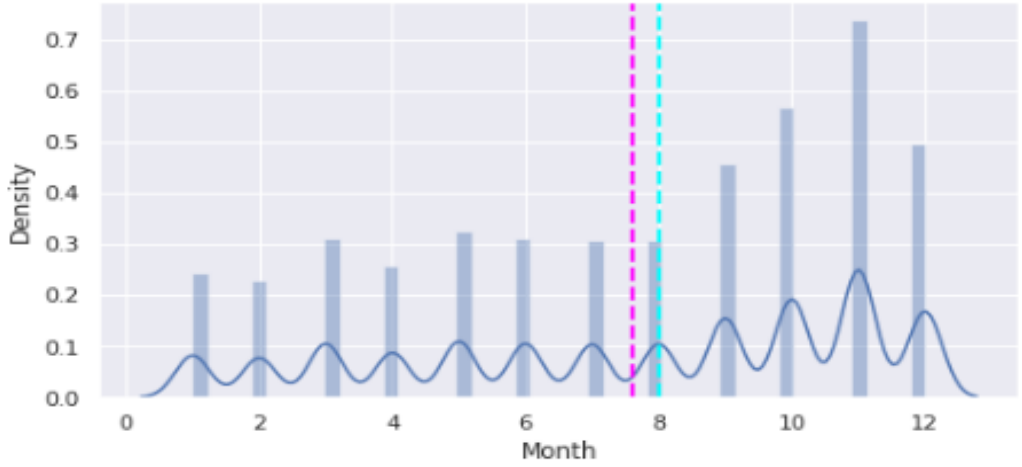


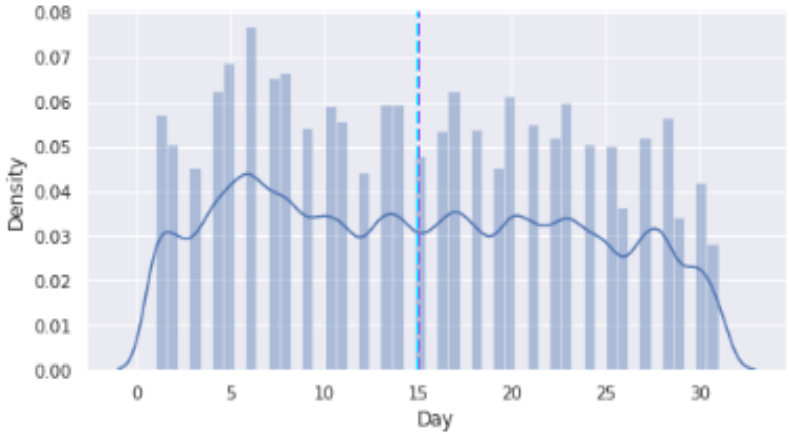


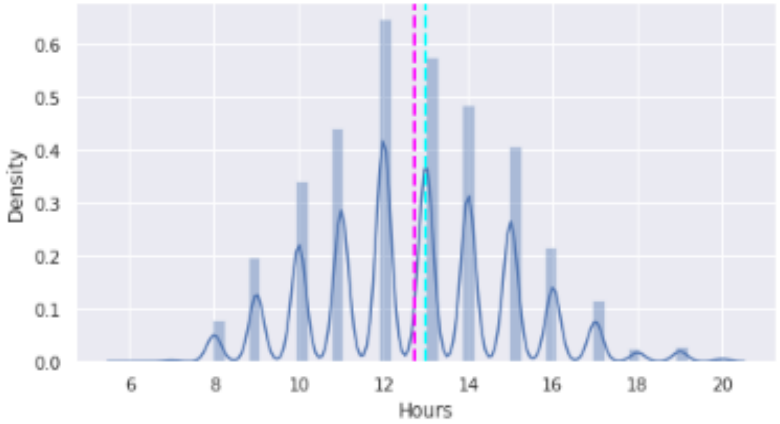

0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Density
12000
13000
14000
15000
16000
17000
18000
19000
CustomerID

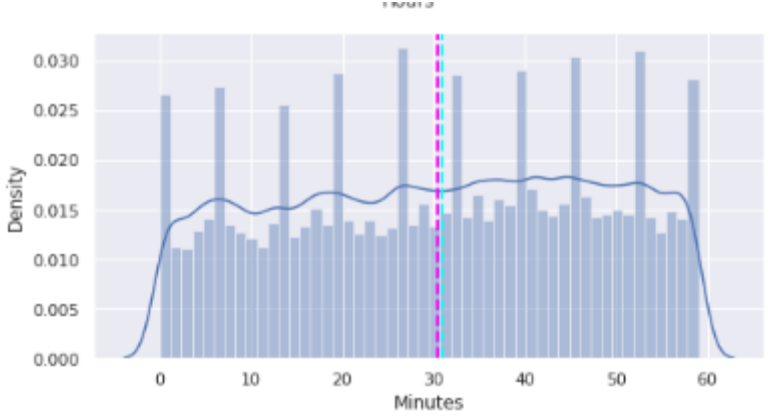

0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Density
0
25000
50000
75000
100000
125000
150000
175000
Total_Price


Density
40
30
20
10
0
2010.0
2010.2
2010.4
2010.6
2010.8
2011.0
Year

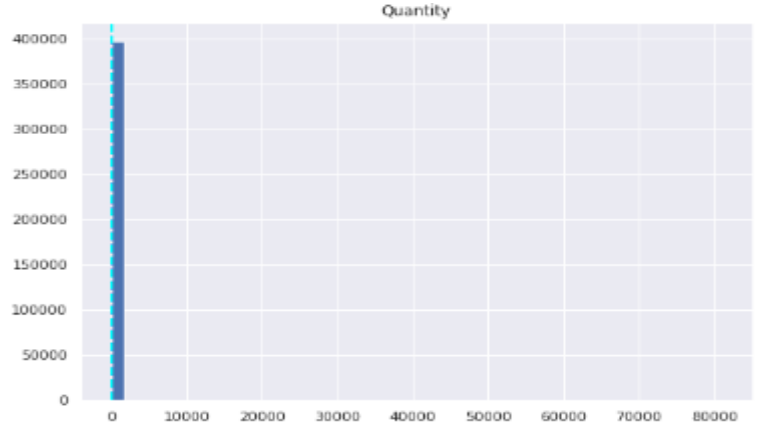

Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12
Month


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
0
5
10
15
20
25
30
Day


0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
6
8
10
12
14
16
18
20
Hours

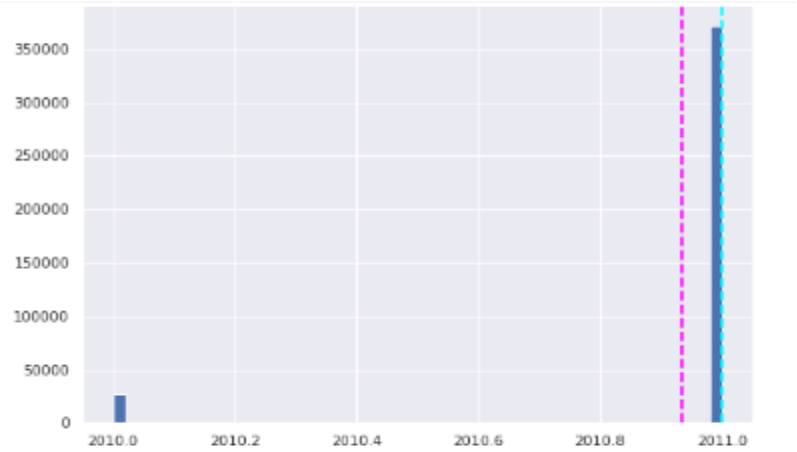

0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
10
20
30
40
50
60
Minutes

All numerical feature using histogram

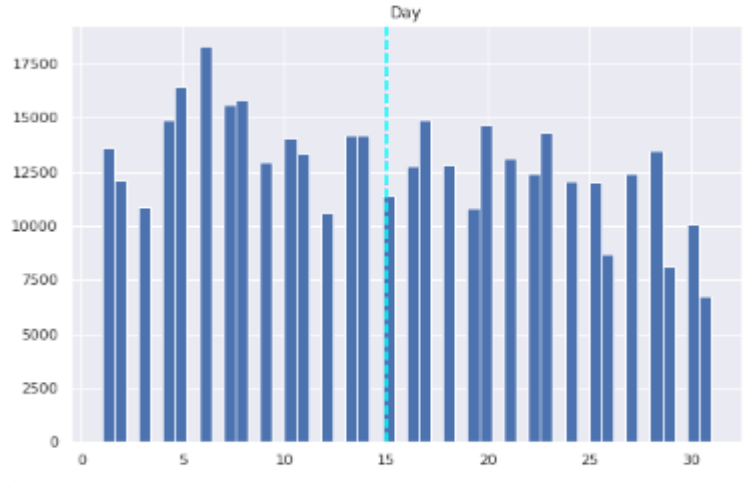


Skewness : 403.31943081039486
Kurtosis : 173965.71516660746

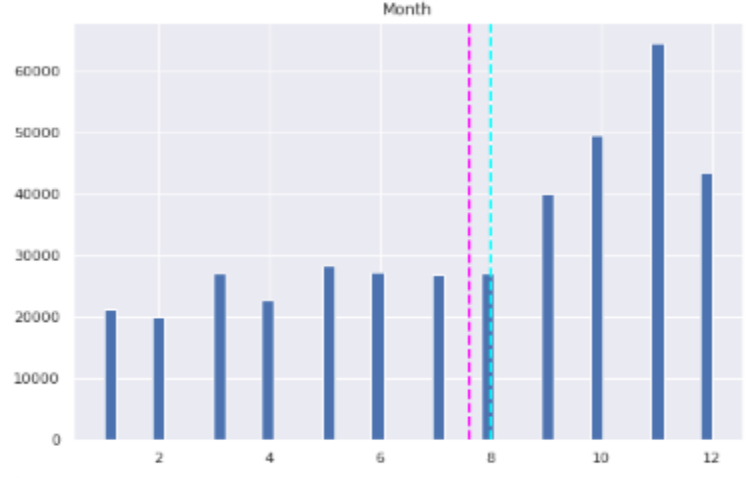
Skewness : 0.02577629847429845
Kurtosis : -1.1808302151571712

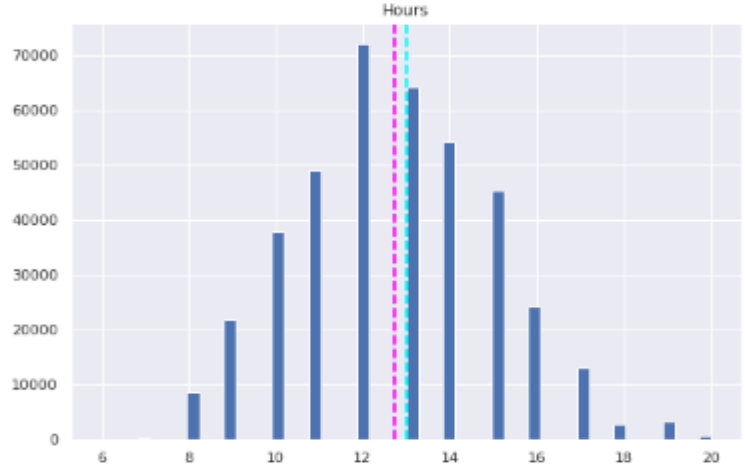


Skewness : -3.5045145019499735
Kurtosis : 10.281673570919017

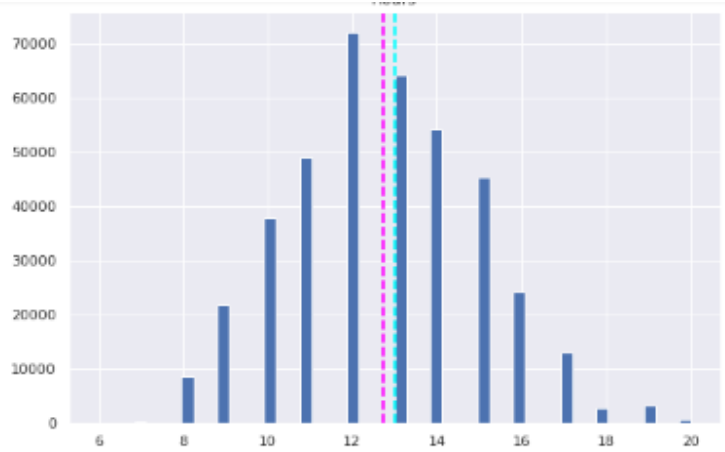


Skewness : 0.1144792789730314
Kurtosis : -1.1728727400772625

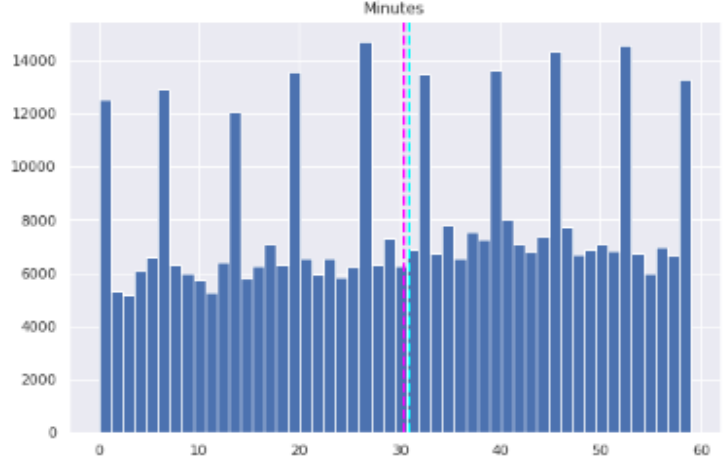




Skewness : 0.18903743045332783



Skewness : 0.18903743045332783
Kurtosis : -0.20968488890482462



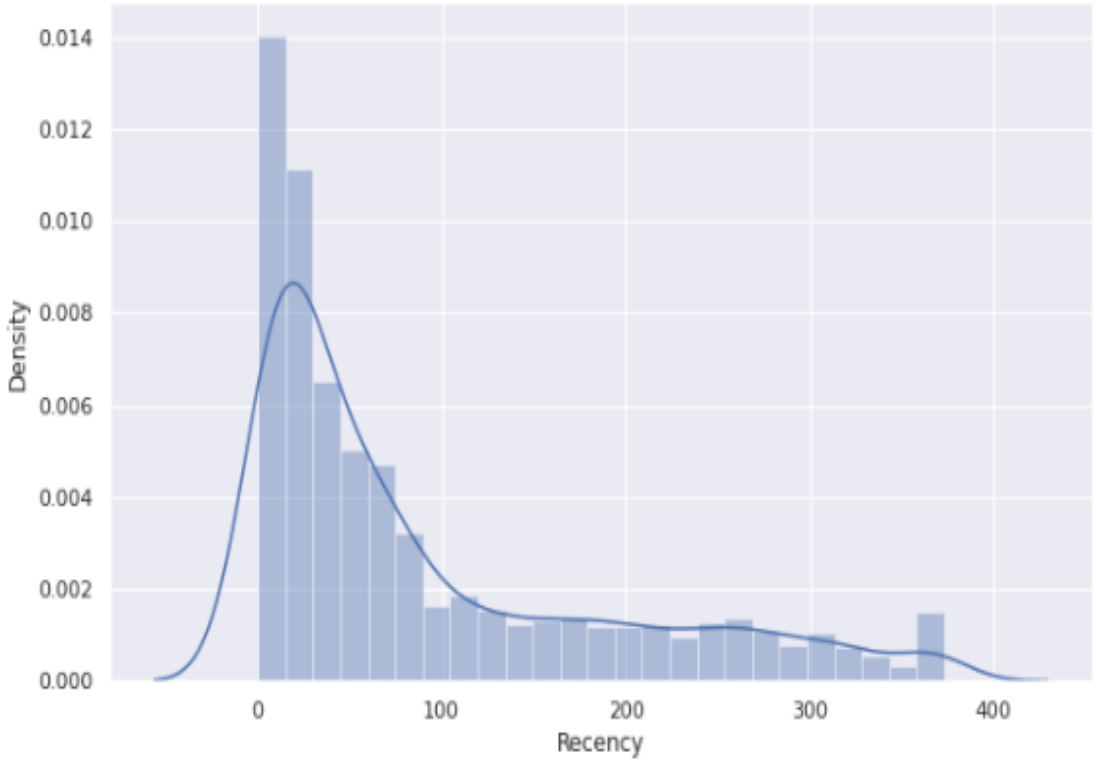
Skewness : -0.08092206133583811
Kurtosis : -1.1655389341755757

Recency, Frequency, Monetary (RFM) Model:

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.
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.

- **Recency :-** Latest Date - Last Inovice Data
- **Frequency :-**Count of Invoice No. of transactions
- **Monetary:-** Sum of Total Amount for each customer

Descriptive Statistics:

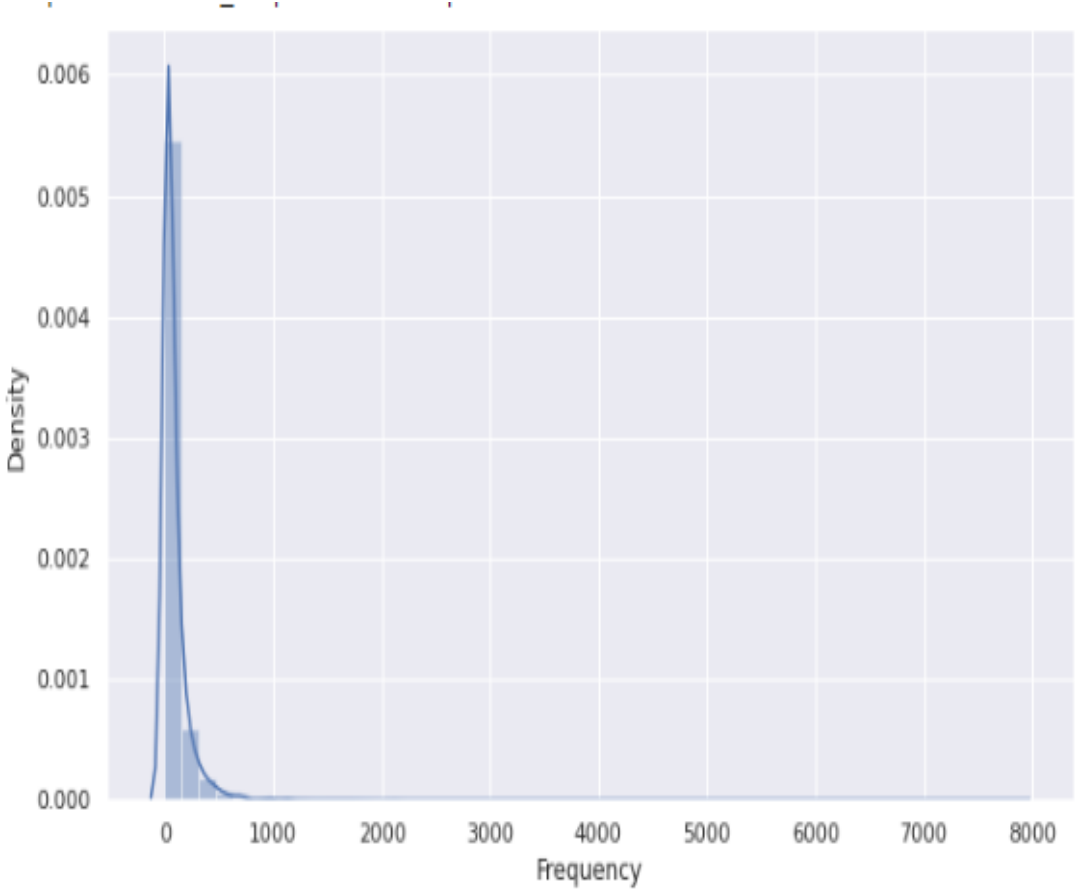


Distribution Plot

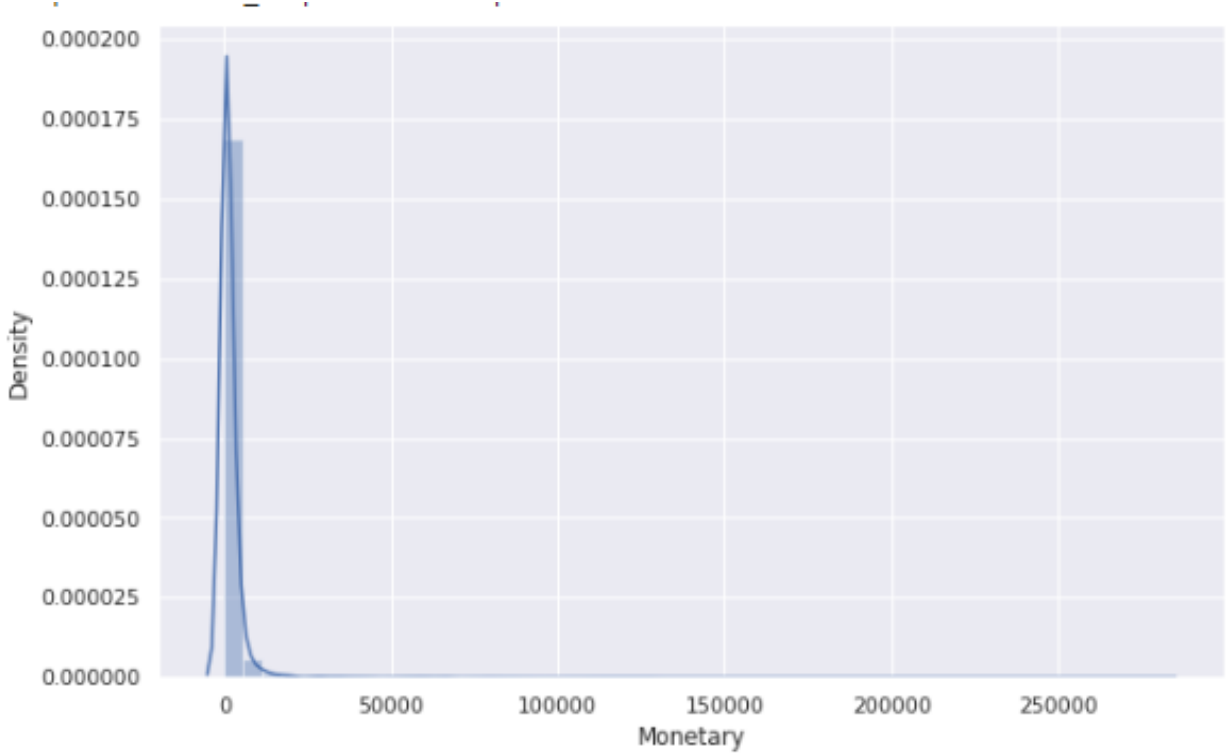
	CustomerID	Recency	Frequency	Monetary
0	12346.0	325	1	77183.60
1	12347.0	2	182	4310.00
2	12348.0	75	31	1797.24
3	12349.0	18	73	1757.55
4	12350.0	310	17	334.40

```
count    4339.000000
mean       92.041484
std       100.007757
min         0.000000
25%        17.000000
50%        50.000000
75%       141.500000
max       373.000000
Name: Recency, dtype: float64
```

Descriptive Statistics (Frequency):



Descriptive Statistics (Monetary)



Computing Quantile of RFM values:

Customers are categorized as top customers based on their frequency, amount spent, and recency. qcut() is a discretization function based on quantiles. Data is binned by qcut using sample quantiles. For instance, a categorical object representing quantile membership for each client would be produced from 1000 values for 4 quantiles.

Calculate & Add R, F and M segment value columns in the existing dataset to show R, F andM segment values:

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

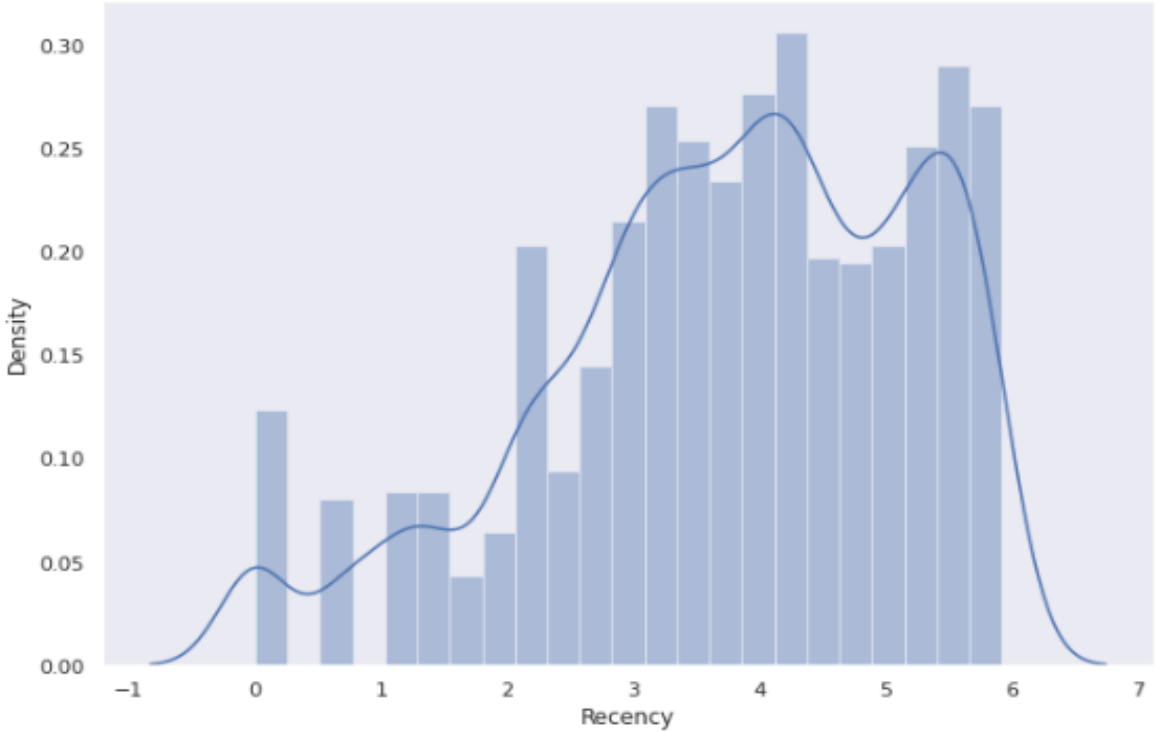
RFM Interpretation result:

Calculate and Add column showing total sum of RFMGroup values:

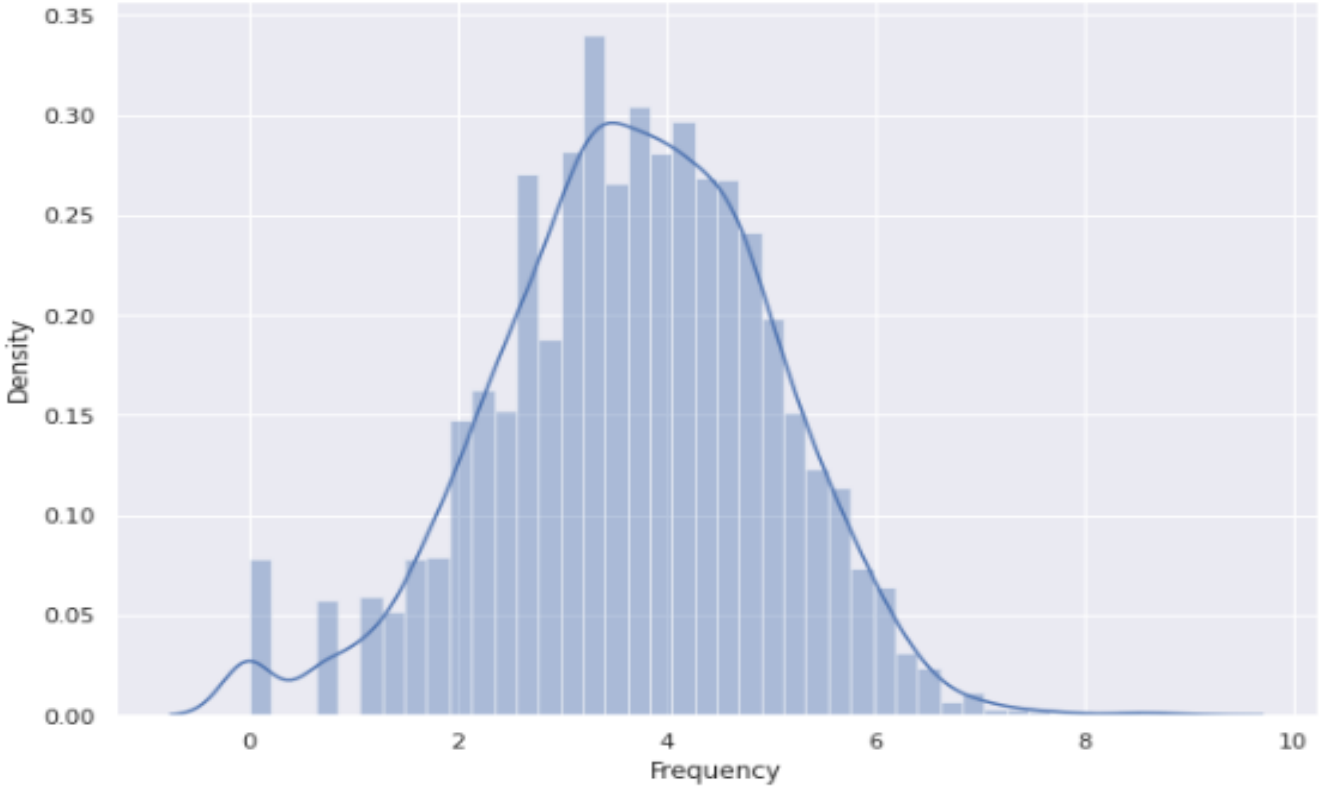
CustomerID	Recency	Frequency	Monetary	R_quantile	F_quantile	M_quantile	RFM_Score
12346.0	325	1	77183.60	4	4	1	441
12347.0	2	182	4310.00	1	1	1	111
12348.0	75	31	1797.24	3	3	1	331
12349.0	18	73	1757.55	2	2	1	221
12350.0	310	17	334.40	4	4	3	443

Log Transformation:

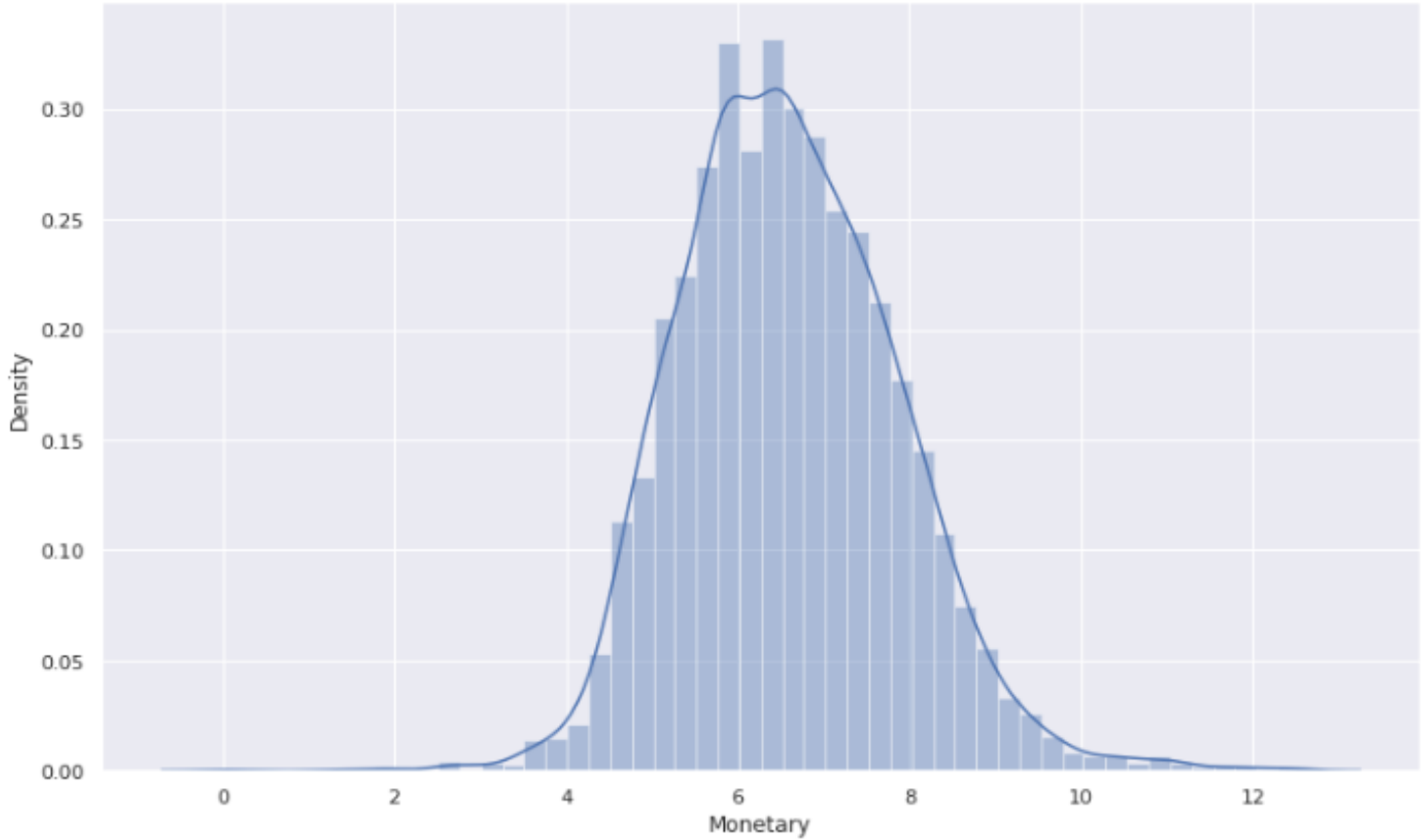
Normalization for Recency, the distribution of data is as follows:



Normalization for Frequency, the distribution of data is as follows:



Normalization for Monetary, the distribution of data is as follows:



K- Means Clustering Implementation:

It can be summarized as the process of finding data subgroups where data points in the same subgroup (cluster) are extremely similar and other data points in other clusters are very different.

Finding Optimal Number of Clusters

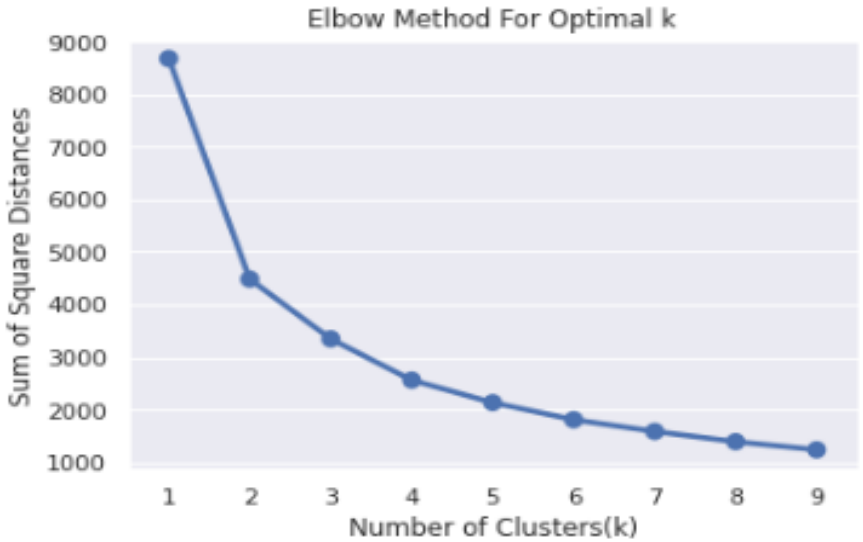
- There are two primary methods to define number of clusters:
  - Silhouette Score (math method)
    - Measures intra- and inter-cluster distance
  - Elbow criterion (visual method)
    - Plot number of clusters against within-cluster sum-of- squared-errors (SSE) - sum of squared distances from every data point to their cluster center

Silhouette Score:

```
For n_clusters = 2, silhouette sore is 0.42096647717466357
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 0.3401262629237664
For n_clusters = 6, silhouette sore is 0.34349645790029015
For n_clusters = 7, silhouette sore is 0.3424681104938937
For n_clusters = 8, silhouette sore is 0.33717186315551134
For n_clusters = 9, silhouette sore is 0.3449377656733416
For n_clusters = 10, silhouette sore is 0.34800869969584486
For n_clusters = 11, silhouette sore is 0.33805099910434905
For n_clusters = 12, silhouette sore is 0.3430661353191348
For n_clusters = 13, silhouette sore is 0.33933479417634393
For n_clusters = 14, silhouette sore is 0.34475828866054914
```

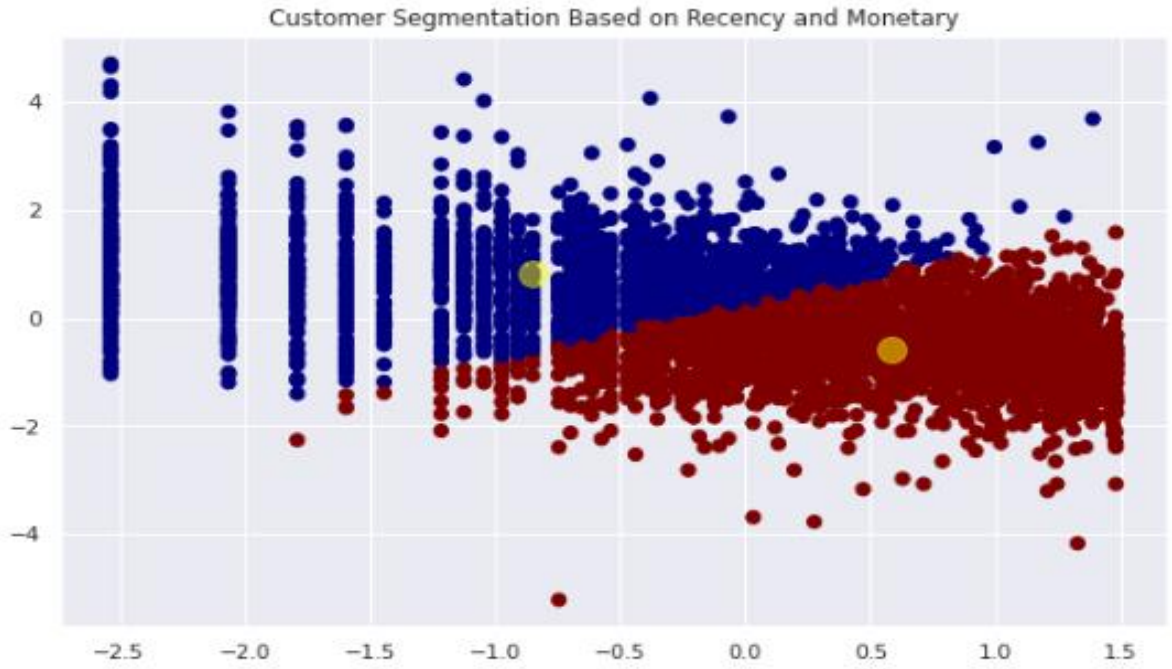
Elbow Method on Recency and Monetary:

The graph for the sum of square distance values and Number of Clusters



Hyperparameter Tuning For the Best Value of K

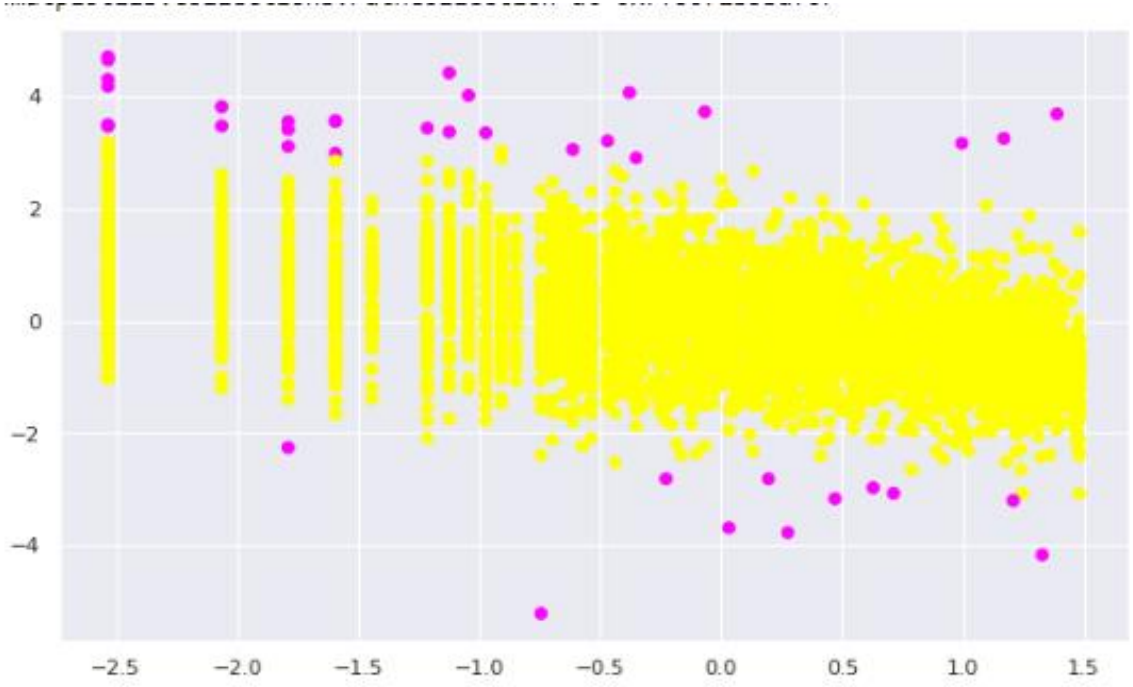
Customer segmentation by taking k=2



When we separate customers by Recency and Monetary value, we can observe that they are well-separated.

Implementation of Density Based Spatial Clustering of Applications with Noise (DBSCAN):

DBSCAN on Recency and Monetary

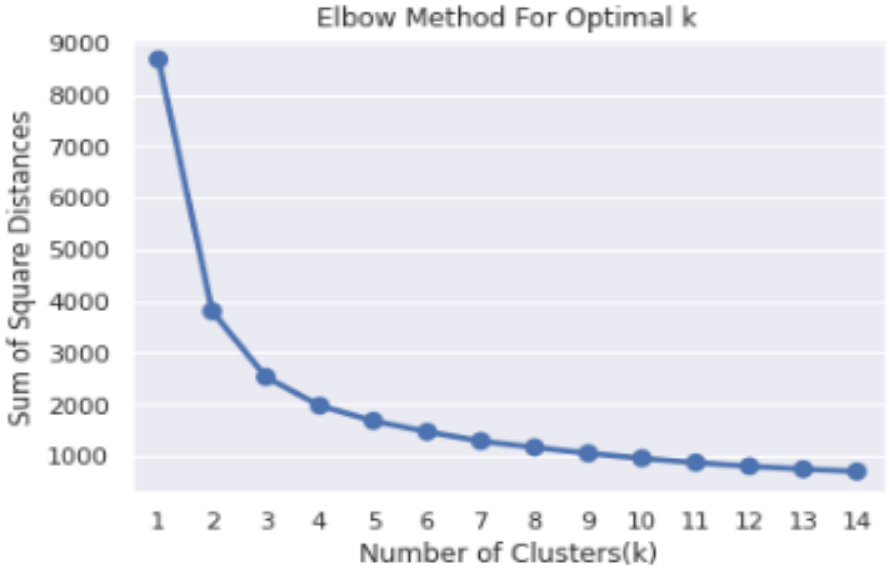


Silhouette score method on Frequency and Monetary:

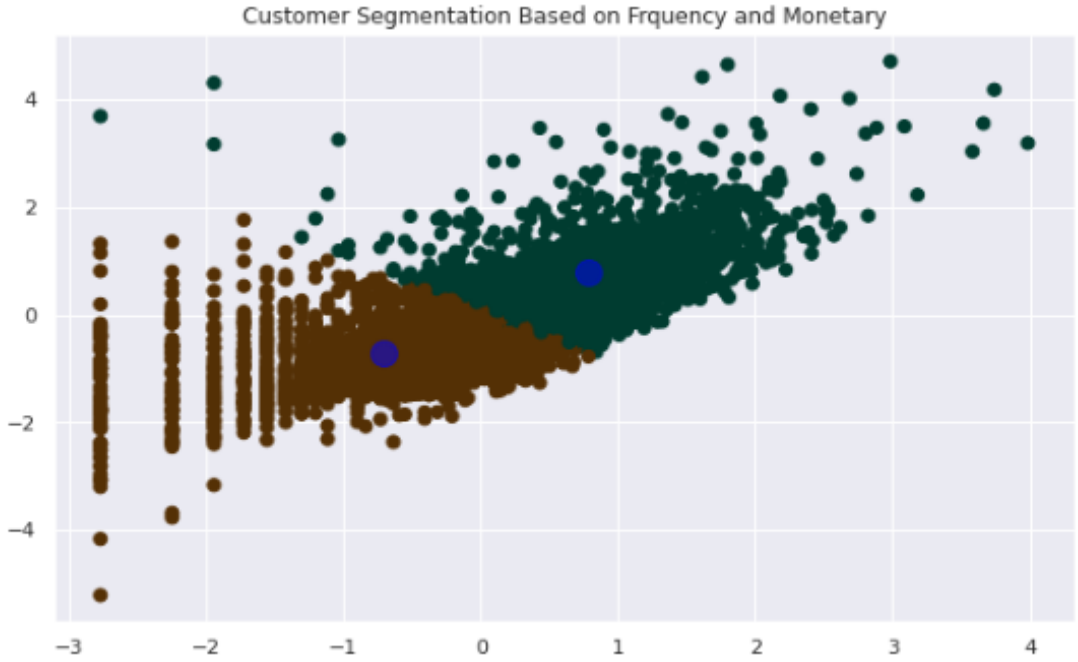
Silhouette score method on Frequency and Monetary

```
For n_clusters = 2, silhouette score is 0.4782608772260966
For n_clusters = 3, silhouette score is 0.4074267531808067
For n_clusters = 4, silhouette score is 0.37172714433473125
For n_clusters = 5, silhouette score is 0.3438018664611139
For n_clusters = 6, silhouette score is 0.35968013778064417
For n_clusters = 7, silhouette score is 0.3385630688823618
For n_clusters = 8, silhouette score is 0.3533303318815764
For n_clusters = 9, silhouette score is 0.3462295705010296
For n_clusters = 10, silhouette score is 0.34368766835247655
For n_clusters = 11, silhouette score is 0.3673040361128039
For n_clusters = 12, silhouette score is 0.35393726205051923
For n_clusters = 13, silhouette score is 0.36133414280576015
For n_clusters = 14, silhouette score is 0.3694740460402445
```

The graph for the sum of square distance values and Number of Clusters

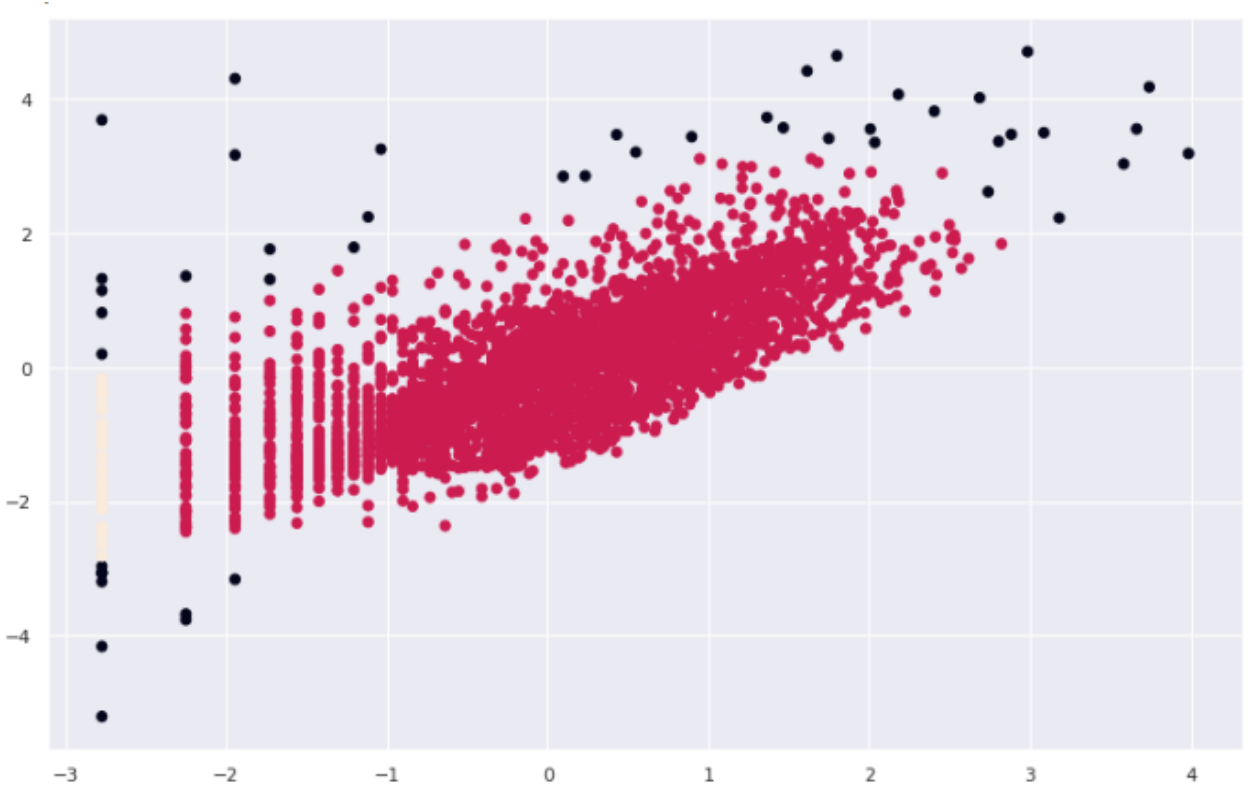


Hyperparameter Tuning For Best Value of K



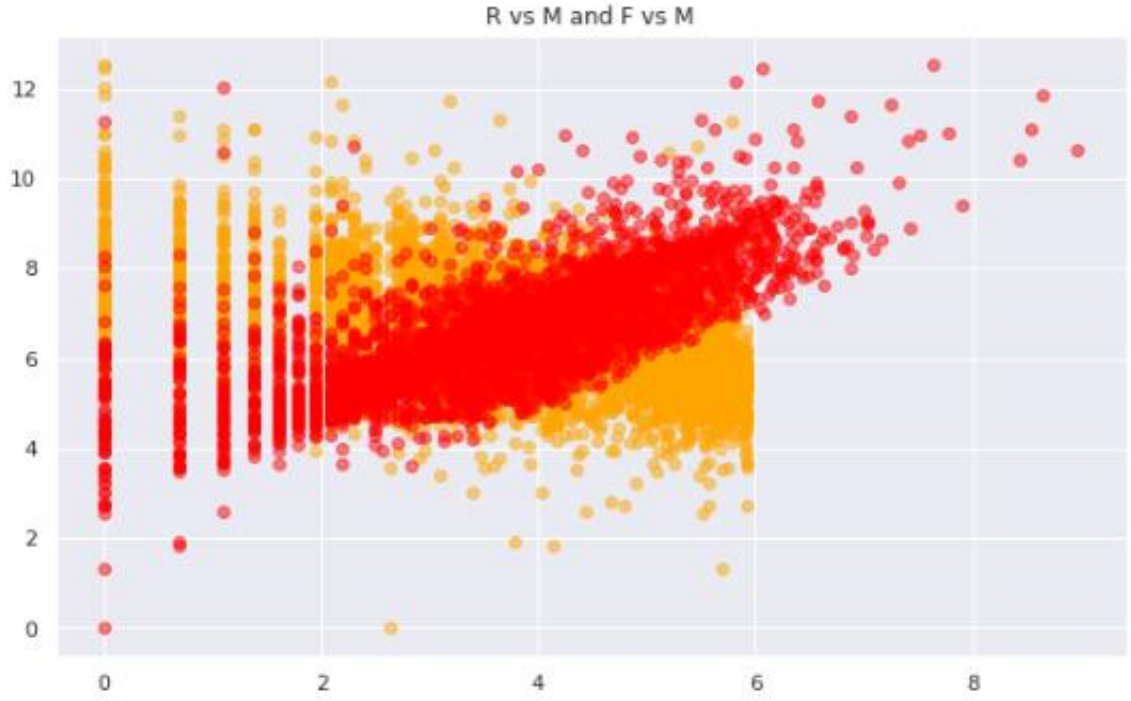
DBSCAN on Frequency and Monetary

DBSCAN method on Frequency and Monetary



Comparison between R vs M and F vs M

Plot R vs M and F vs M



Conclusion:

1}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.

2}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.

3}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.

4}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.

5}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.

Thank you