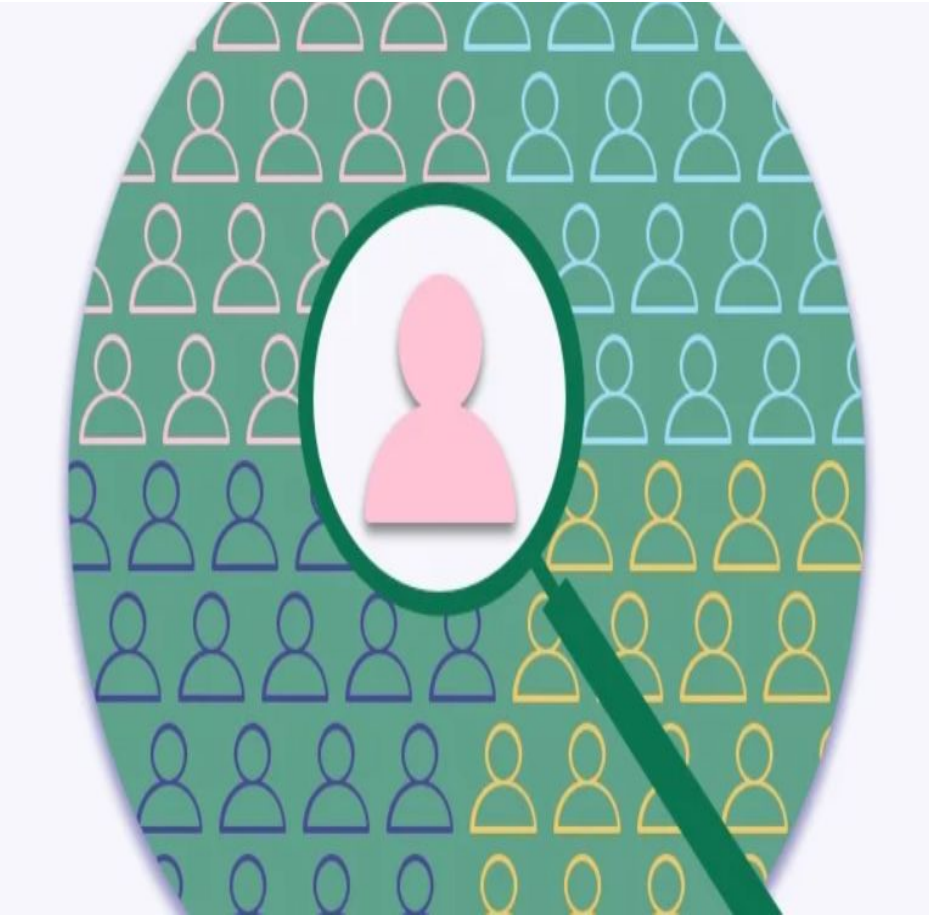
Capstone Project – 4

Online Retail Customer Segmentation

Presentation Outline :

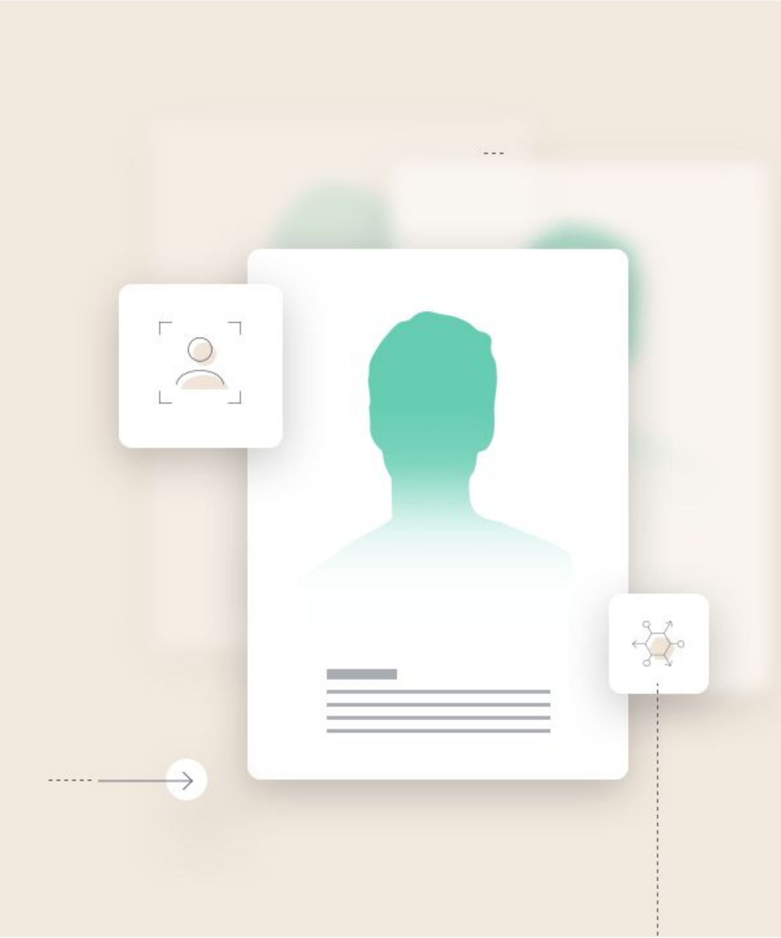
- Business Objective
- Exploring the Dataset
- Exploratory Data Analysis
- Data Cleaning
- Feature Engineering : RFM model
- ML Model Evaluation & Results
- Conclusion



Business Objective

Problem Statement :

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



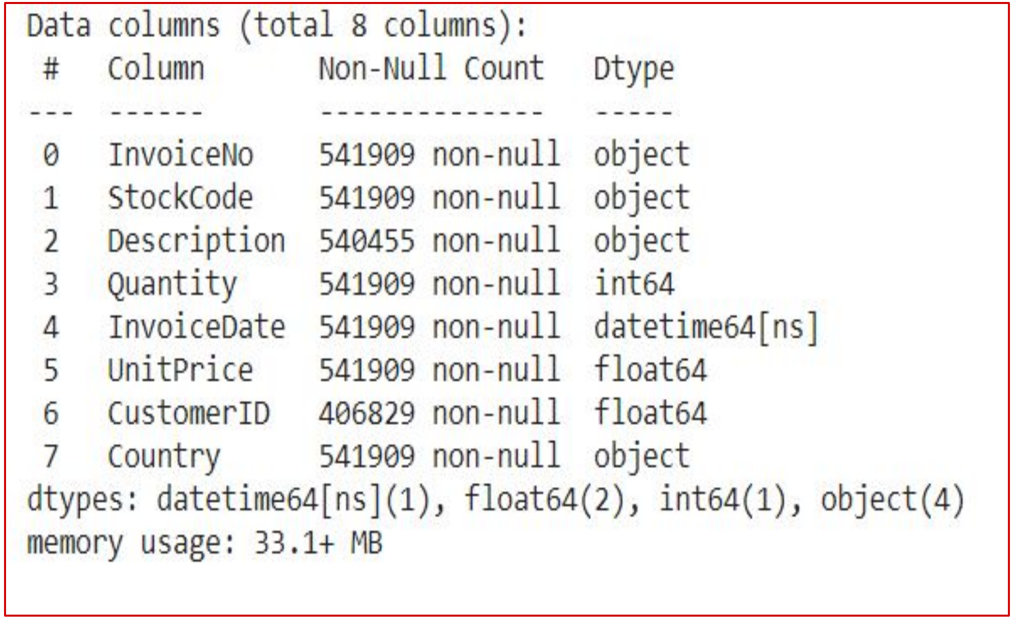
Data Summary :

Our dataset is composed of 8 columns and 541,909 rows.

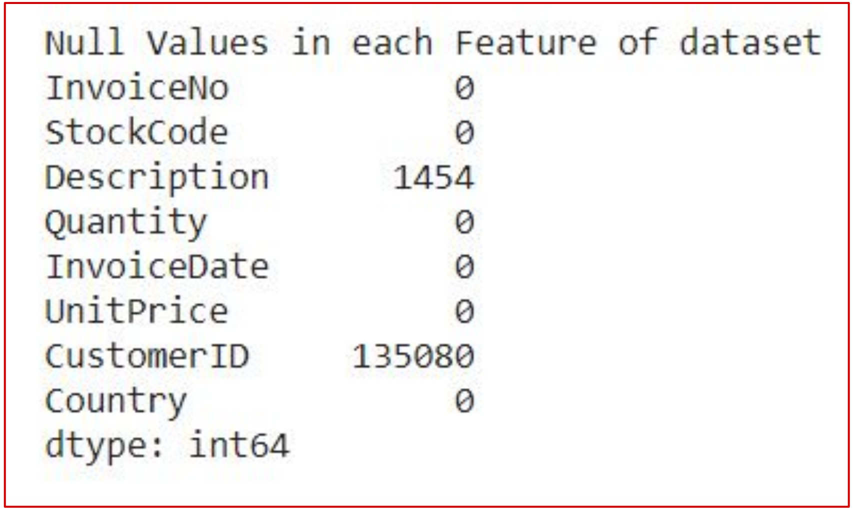
Attribute Information :

- **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- **Description**: Product (item) name. Nominal.
- **Quantity**: The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **UnitPrice**: Unit price. Numeric, Product price per unit in sterling.
- **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- **Country**: Country name. Nominal, the name of the country where each customer resides.

Attribute Information : Dtype & Null values



```
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  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

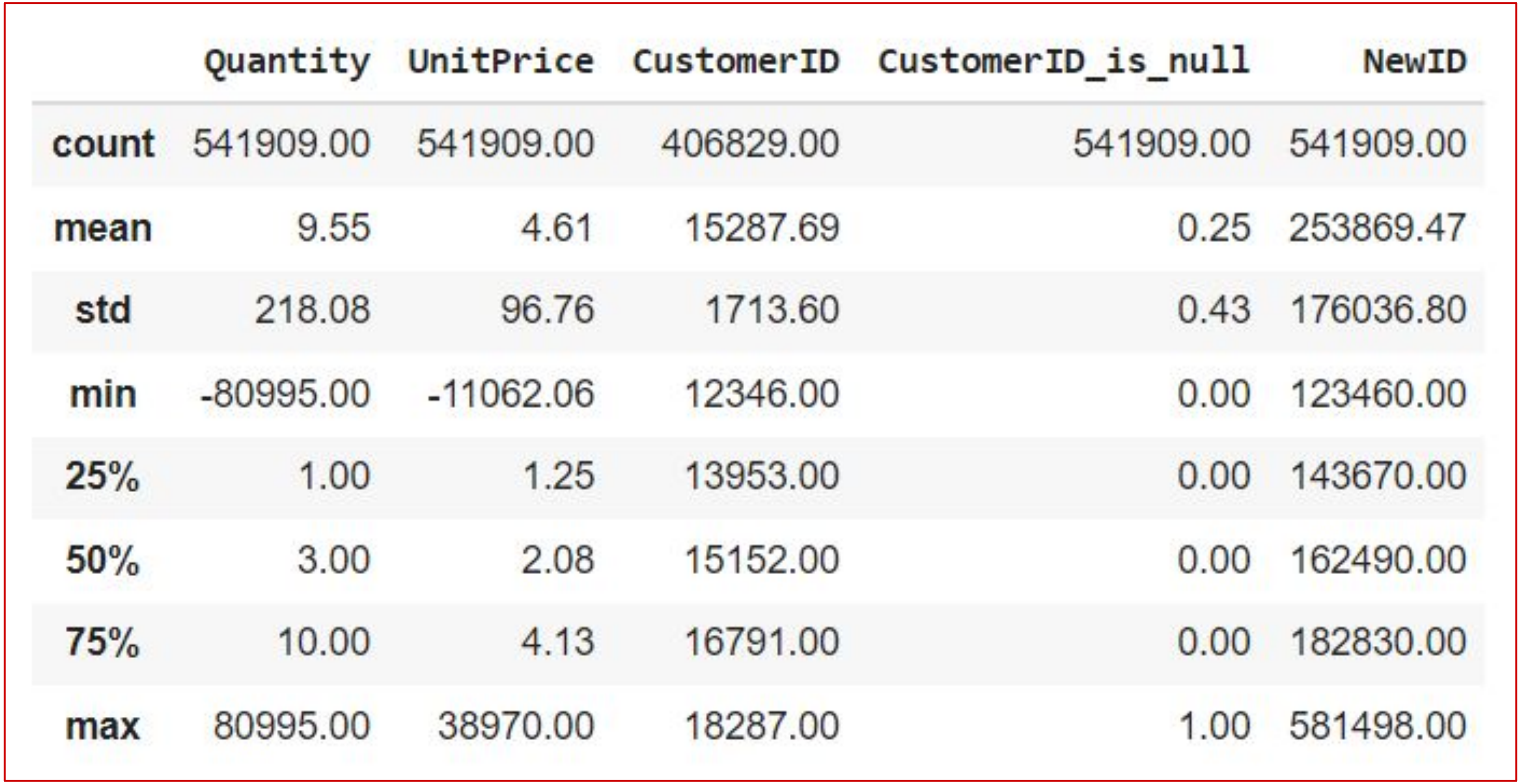


```
Null Values in each Feature of dataset
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

- We are missing values in the CustomerID and Description columns.
- Since 25% of the customer ID's are missing, we will create and fill a new column that has a 1 when customer ID is null and a 0 when it is not.
- We will investigate the records where the CustomerID field is null to determine whether to erase those rows or fill in the missing values.

- Percentage of customers missing: 24.93 %
- Percentage of Description missing: 0.27 %
- Number of duplicated records: 5268
- We will create a new customer ID column called **NewID** with the invoice numbers filling in for the missing values.

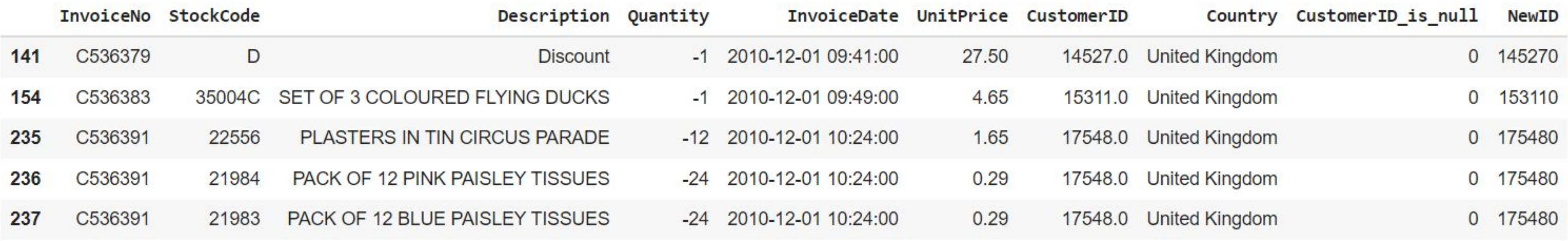
EDA : Descriptive Stats



	Quantity	UnitPrice	CustomerID	CustomerID_is_null	NewID
count	541909.00	541909.00	406829.00	541909.00	541909.00
mean	9.55	4.61	15287.69	0.25	253869.47
std	218.08	96.76	1713.60	0.43	176036.80
min	-80995.00	-11062.06	12346.00	0.00	123460.00
25%	1.00	1.25	13953.00	0.00	143670.00
50%	3.00	2.08	15152.00	0.00	162490.00
75%	10.00	4.13	16791.00	0.00	182830.00
max	80995.00	38970.00	18287.00	1.00	581498.00

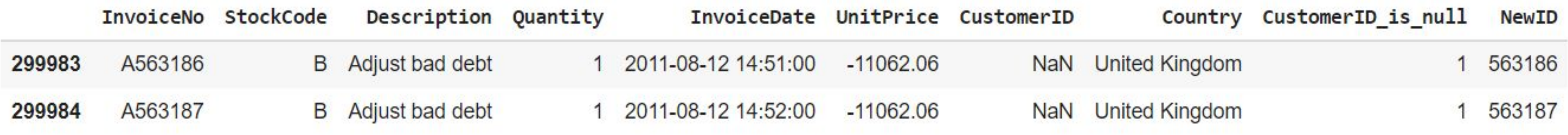
- Notice that there are negative values in the Quantity and UnitPrice columns. I am assuming these are orders that were cancelled and items that were returned, but let's make sure.
- We found in descriptive statistics above that customers buy an average quantity of about 10 per product

EDA(continued) : Negative Values



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	CustomerID_is_null	NewID
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom	0	145270
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom	0	153110
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom	0	175480
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom	0	175480
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom	0	175480

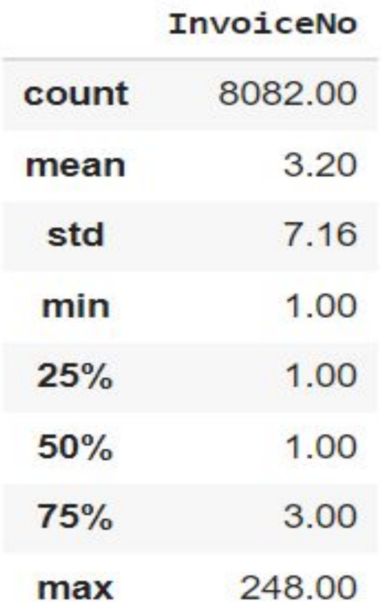
- Nothing came back when we filtered the cancelled orders by Quantity > 0, this confirms that the negative values mean the order was cancelled.
- There were 9288 cancelled orders out of 25900 Unique orders. & Percentage of orders cancelled is 35.86%



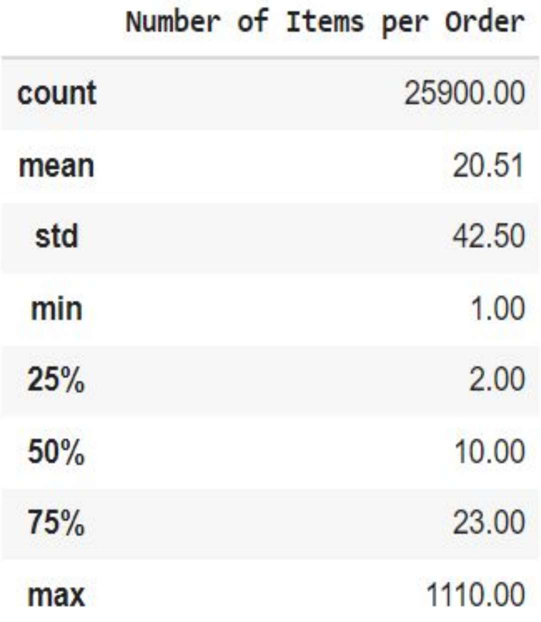
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	CustomerID_is_null	NewID
299983	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	NaN	United Kingdom	1	563186
299984	A563187	B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	NaN	United Kingdom	1	563187

- "Adjust bad debt" tells us that this is an adjustment for a customer with insufficient funds or an allowance for a customer who never paid for the order

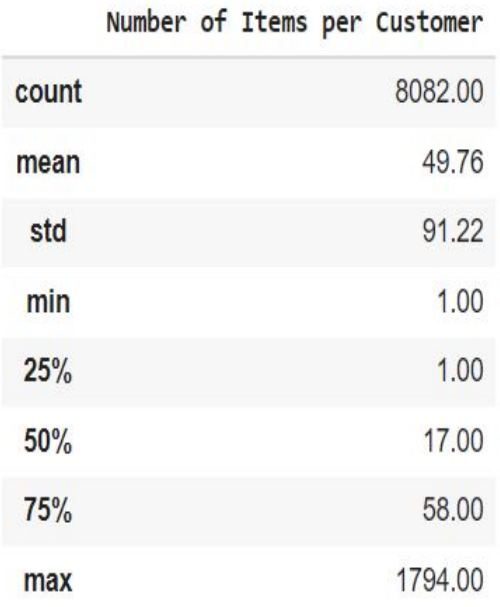
EDA(continued): Exploring the Orders



	InvoiceNo
count	8082.00
mean	3.20
std	7.16
min	1.00
25%	1.00
50%	1.00
75%	3.00
max	248.00



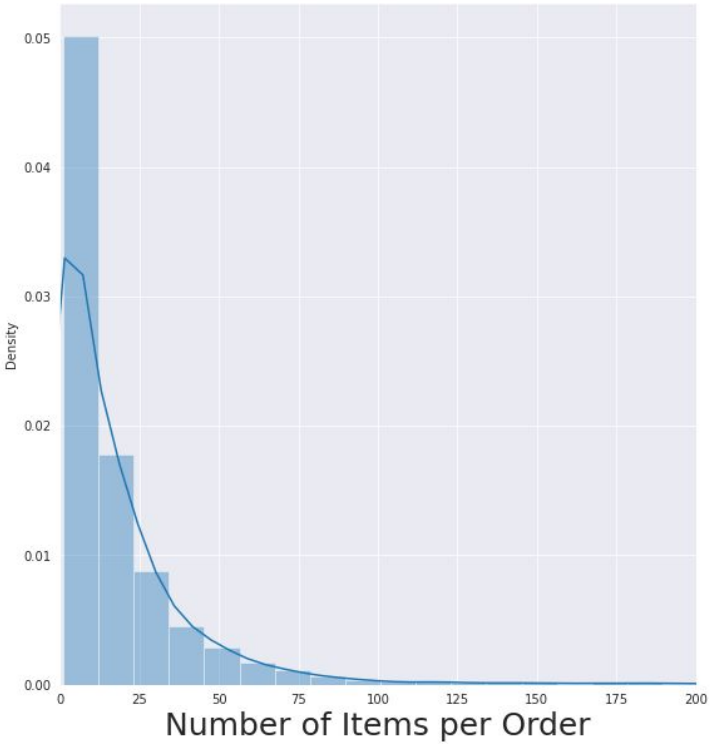
	Number of Items per Order
count	25900.00
mean	20.51
std	42.50
min	1.00
25%	2.00
50%	10.00
75%	23.00
max	1110.00

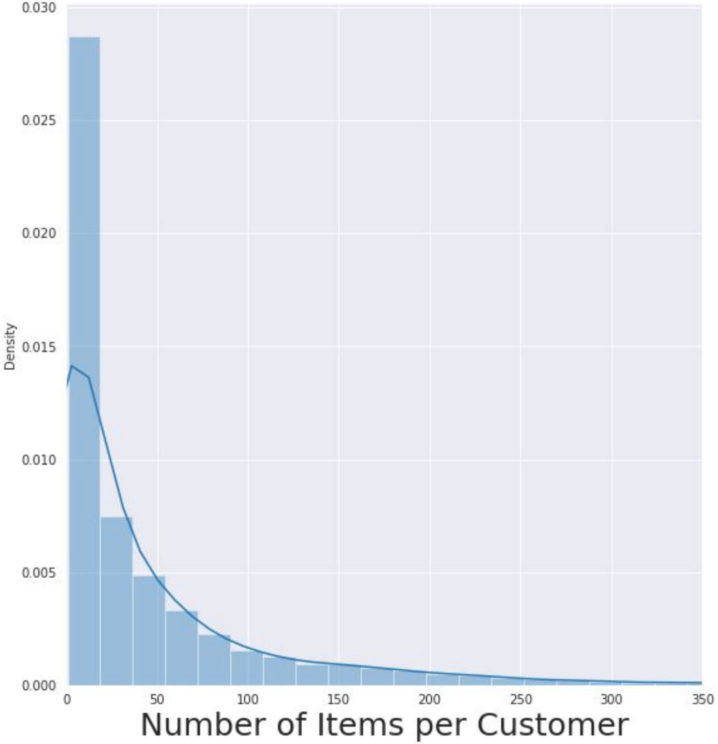


	Number of Items per Customer
count	8082.00
mean	49.76
std	91.22
min	1.00
25%	1.00
50%	17.00
75%	58.00
max	1794.00

- The average number of orders per customer is 3.
- The average number of items per order is 20.5
- The average number of items per customer is 50.

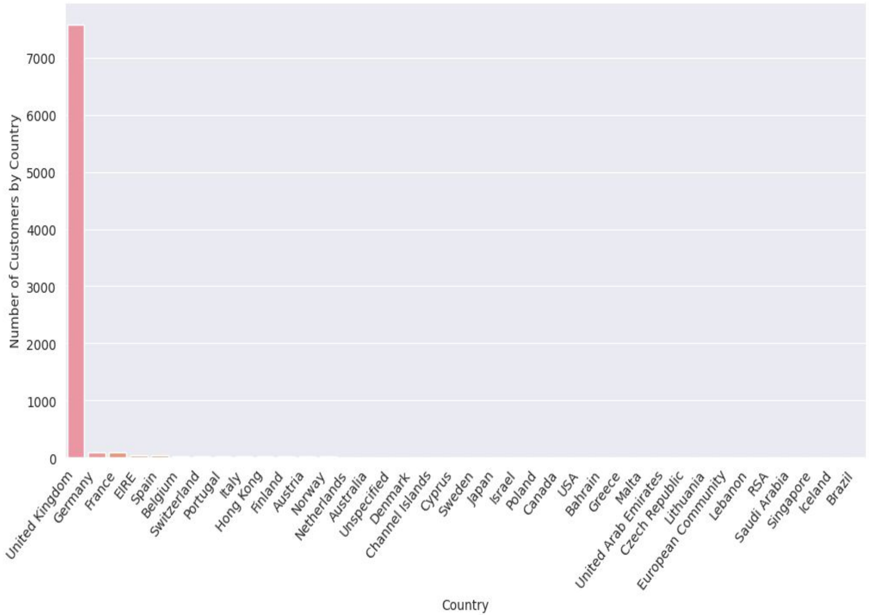
EDA(continued) : Visualize the Orders

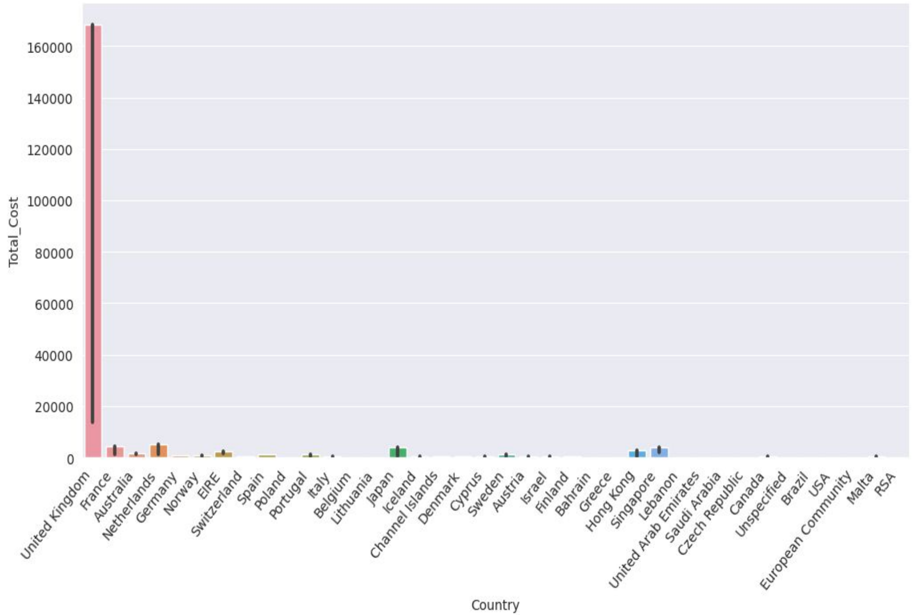




- We have skewed left distributions for both plots. The average number of items per order is 20.5 and the average number of items per customer is 50.

EDA(continued) : Total revenue per country



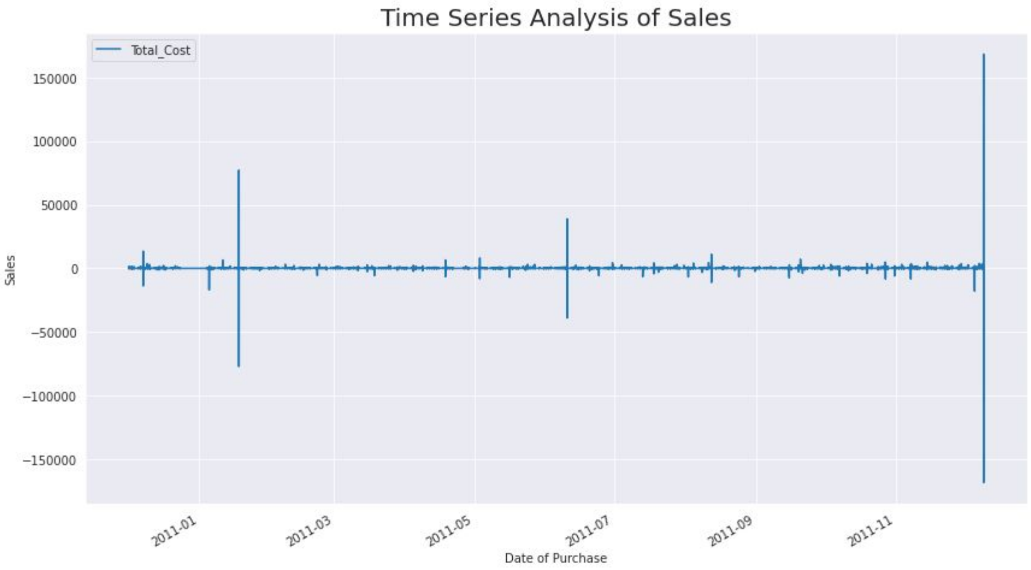


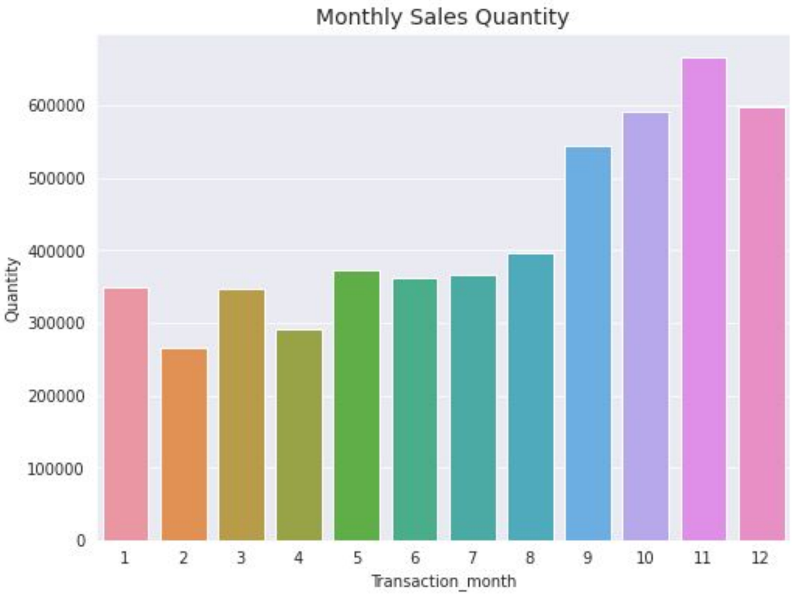
The United Kingdom has significantly more customers than the other countries

Also the UK has the most sales revenue

- Percentage of customers from the UK: 93.88 %
- Number of transactions: 23494
- Number of products Bought: 4065
- Number of customers: 7587

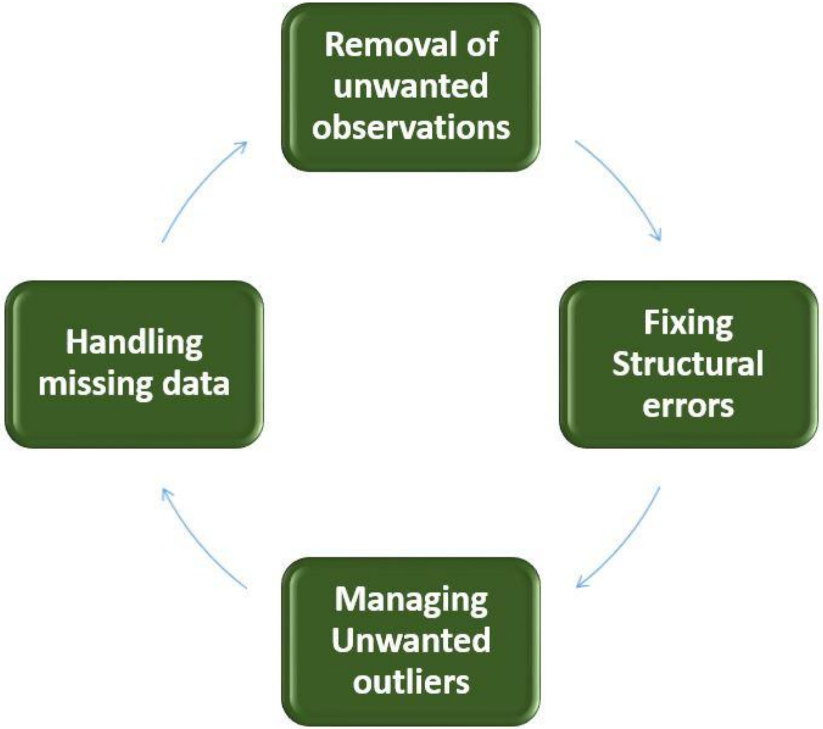
EDA(continued) : Total Sales





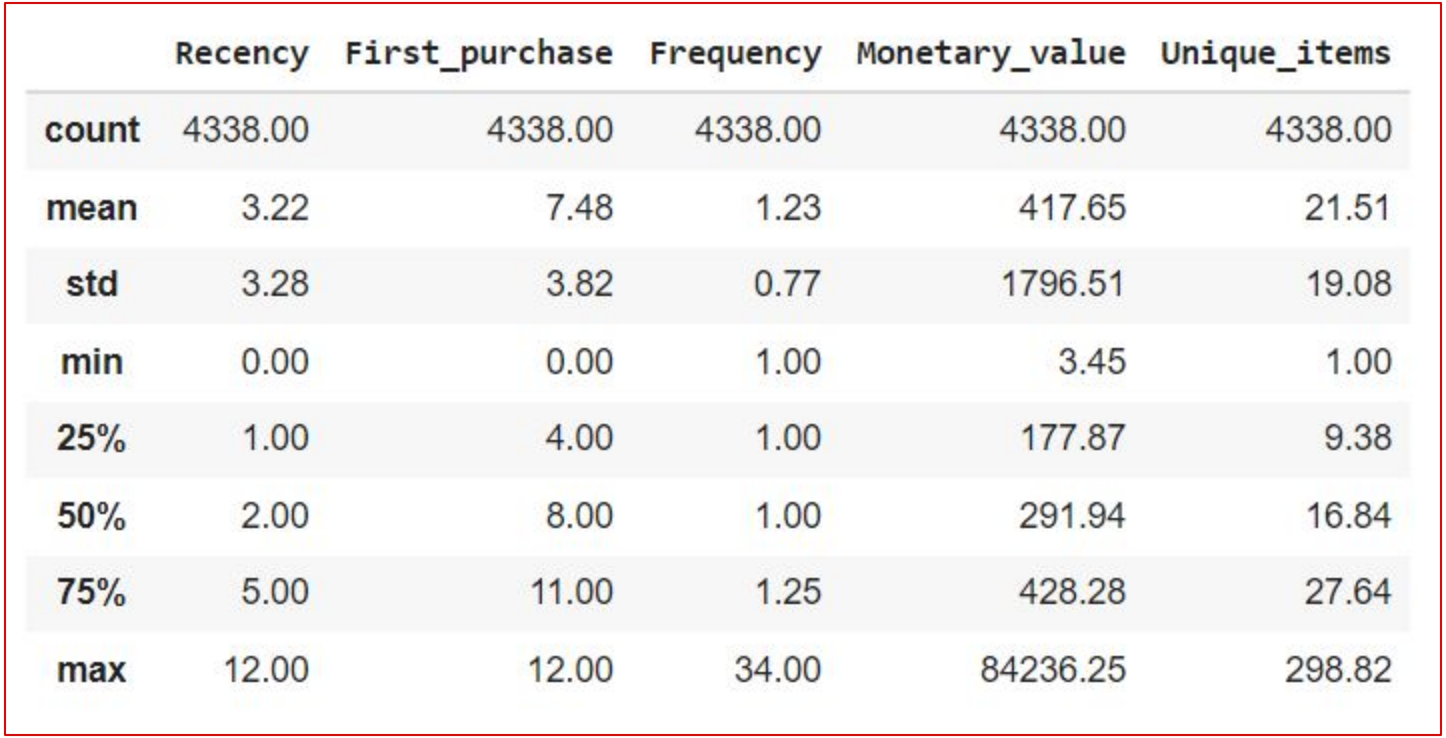
- Year End sales are high.
- Most of the sales quantity happened in between September and December month (Christmas festive sales)

Data Cleaning :



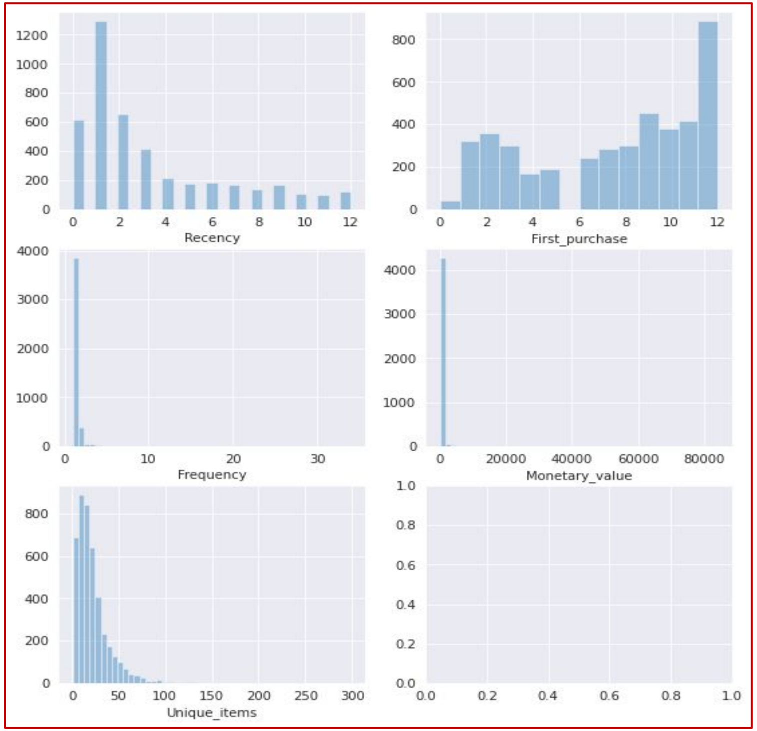
- No more need of newly created explored column
- Drop all rows that contain NaN
- Drop duplicated records.
- Drop transactions with sales price zero.
- Drop cancelled transactions

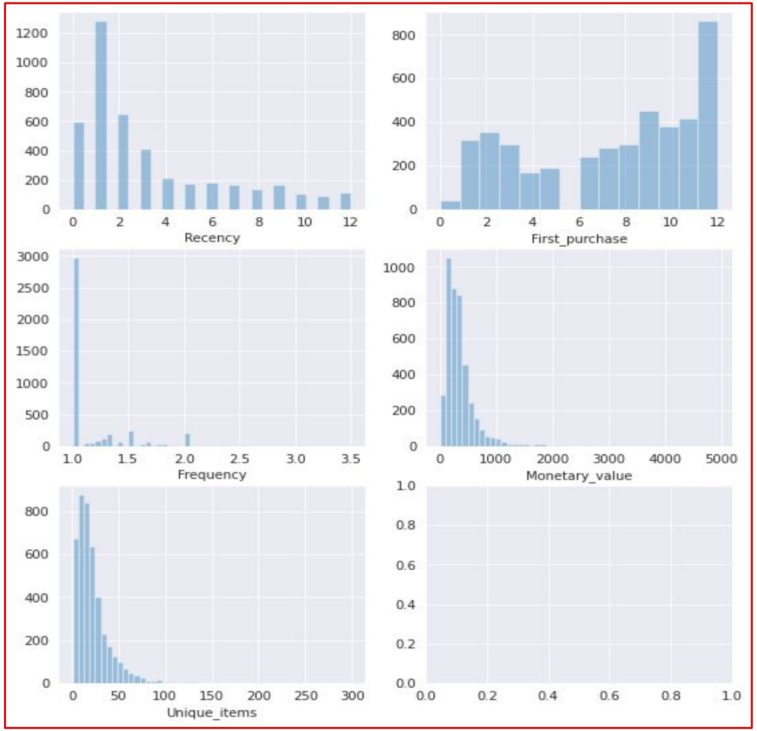
Feature Engineering : RFM Model



	Recency	First_purchase	Frequency	Monetary_value	Unique_items
count	4338.00	4338.00	4338.00	4338.00	4338.00
mean	3.22	7.48	1.23	417.65	21.51
std	3.28	3.82	0.77	1796.51	19.08
min	0.00	0.00	1.00	3.45	1.00
25%	1.00	4.00	1.00	177.87	9.38
50%	2.00	8.00	1.00	291.94	16.84
75%	5.00	11.00	1.25	428.28	27.64
max	12.00	12.00	34.00	84236.25	298.82

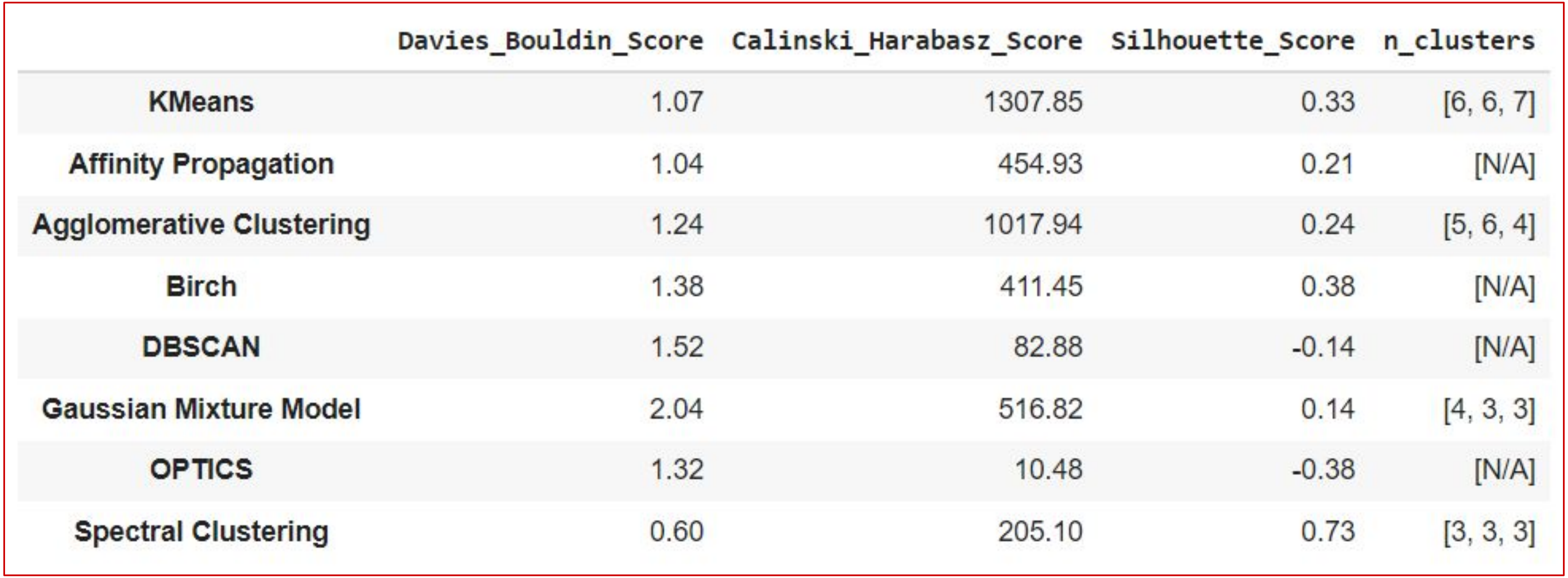
RFM Model : RFM Model Visualization





- The plots show some potential outliers in both frequency and Monetary value, so we will drop those customers from our dataset. Such that customers having Z-score > 3 will be dropped.
- After removing outliers, variables have very different scales and some of them are heavily skewed. So we will normalize all variables.

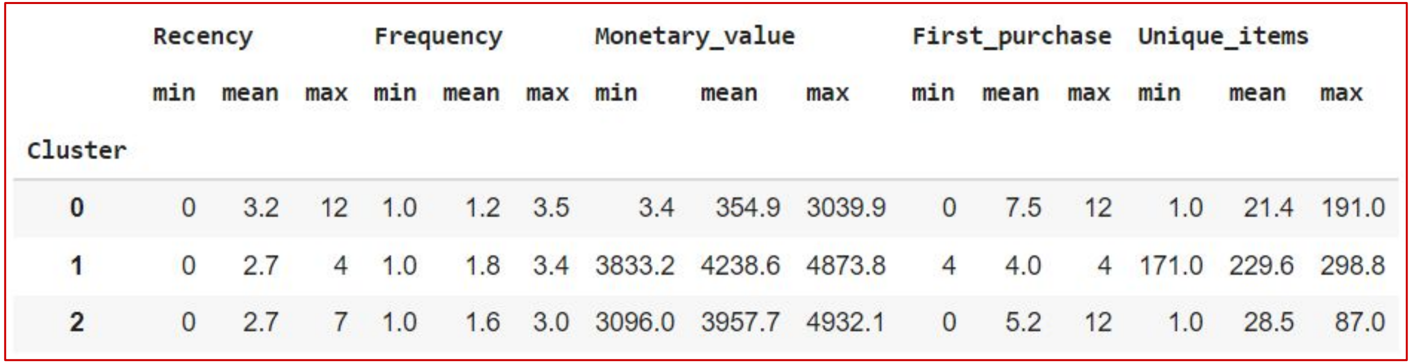
ML Model Evaluation : Promising model



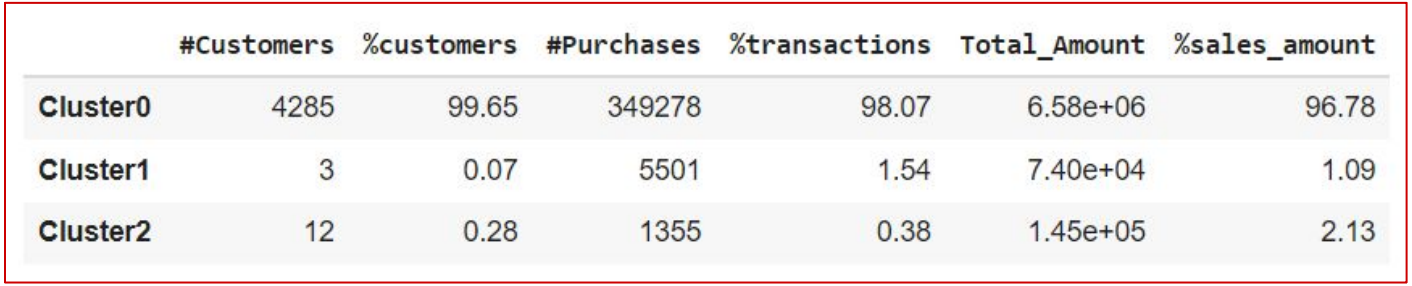
	Davies_Bouldin_Score	Calinski_Harabasz_Score	Silhouette_Score	n_clusters
KMeans	1.07	1307.85	0.33	[6, 6, 7]
Affinity Propagation	1.04	454.93	0.21	[N/A]
Agglomerative Clustering	1.24	1017.94	0.24	[5, 6, 4]
Birch	1.38	411.45	0.38	[N/A]
DBSCAN	1.52	82.88	-0.14	[N/A]
Gaussian Mixture Model	2.04	516.82	0.14	[4, 3, 3]
OPTICS	1.32	10.48	-0.38	[N/A]
Spectral Clustering	0.60	205.10	0.73	[3, 3, 3]

- We have 3 promising models, Spectral clustering with 3 clusters and Kmeans with 6 and 7 clusters. Next, we will find the best model with the best number of clusters that results in clearest interpretation of separation between clusters.

Model Evaluation Result : Spectral clustering



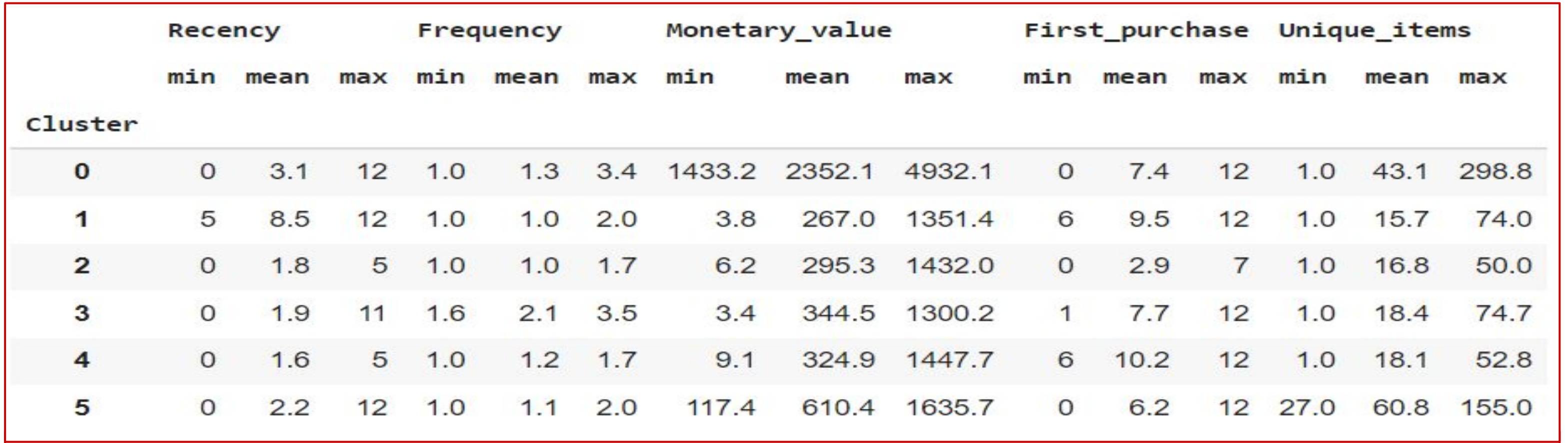
	Recency			Frequency			Monetary_value			First_purchase			Unique_items		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max
Cluster															
0	0	3.2	12	1.0	1.2	3.5	3.4	354.9	3039.9	0	7.5	12	1.0	21.4	191.0
1	0	2.7	4	1.0	1.8	3.4	3833.2	4238.6	4873.8	4	4.0	4	171.0	229.6	298.8
2	0	2.7	7	1.0	1.6	3.0	3096.0	3957.7	4932.1	0	5.2	12	1.0	28.5	87.0



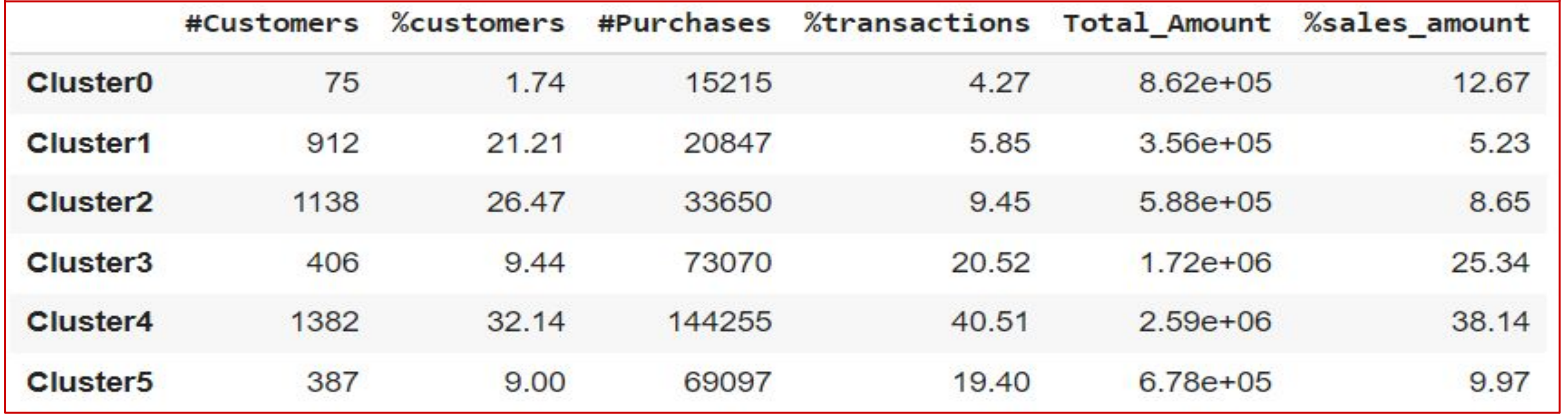
	#Customers	%customers	#Purchases	%transactions	Total_Amount	%sales_amount
Cluster0	4285	99.65	349278	98.07	6.58e+06	96.78
Cluster1	3	0.07	5501	1.54	7.40e+04	1.09
Cluster2	12	0.28	1355	0.38	1.45e+05	2.13

We can see that *Cluster0* alone contains 4285 which is 99.65 of the whole population accounting for 96.78 of the total sales amount! Clearly, this result in not meaningful.

Model Evaluation Result : K-means clustering



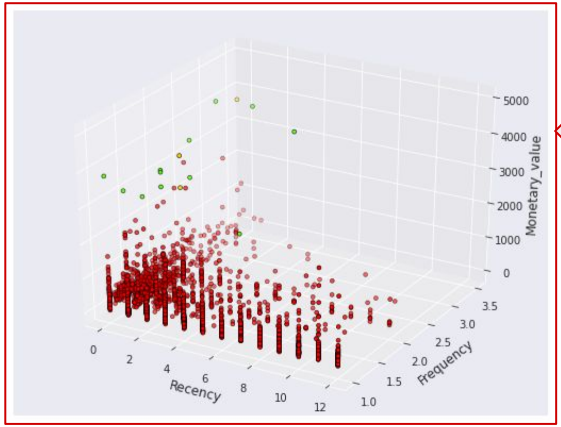
	Recency			Frequency			Monetary_value			First_purchase			Unique_items		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max
Cluster															
0	0	3.1	12	1.0	1.3	3.4	1433.2	2352.1	4932.1	0	7.4	12	1.0	43.1	298.8
1	5	8.5	12	1.0	1.0	2.0	3.8	267.0	1351.4	6	9.5	12	1.0	15.7	74.0
2	0	1.8	5	1.0	1.0	1.7	6.2	295.3	1432.0	0	2.9	7	1.0	16.8	50.0
3	0	1.9	11	1.6	2.1	3.5	3.4	344.5	1300.2	1	7.7	12	1.0	18.4	74.7
4	0	1.6	5	1.0	1.2	1.7	9.1	324.9	1447.7	6	10.2	12	1.0	18.1	52.8
5	0	2.2	12	1.0	1.1	2.0	117.4	610.4	1635.7	0	6.2	12	27.0	60.8	155.0



	#Customers	%customers	#Purchases	%transactions	Total_Amount	%sales_amount
Cluster0	75	1.74	15215	4.27	8.62e+05	12.67
Cluster1	912	21.21	20847	5.85	3.56e+05	5.23
Cluster2	1138	26.47	33650	9.45	5.88e+05	8.65
Cluster3	406	9.44	73070	20.52	1.72e+06	25.34
Cluster4	1382	32.14	144255	40.51	2.59e+06	38.14
Cluster5	387	9.00	69097	19.40	6.78e+05	9.97

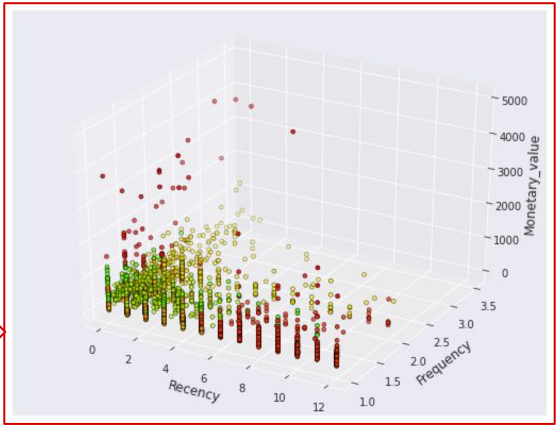
By examining cluster_summary dataframe and cluster_stats dataframe it is interesting to see that each cluster indeed contains a group of consumers that have certain distinct and intrinsic features.

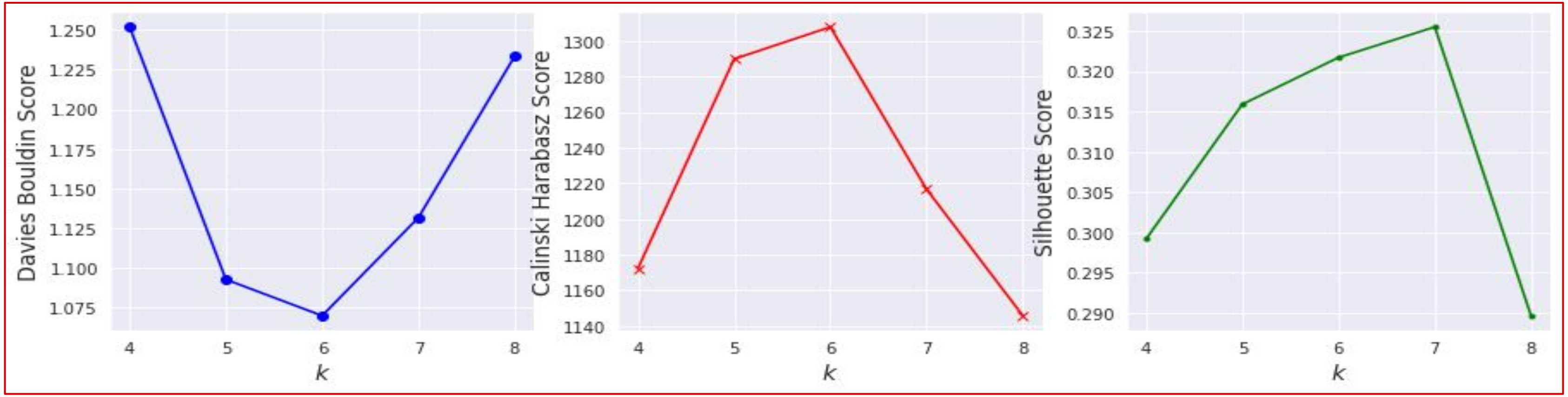
Model Evaluation : Visualization Plot



Spectral clustering

K-means clustering





Elbow plots of different evaluation metrics for K-means clustering used at different K values indicates that K=6 can yield a good solution.

Conclusion :

Understanding the clusters -

- Cluster0 contains 75 customers, composed of 1.7% of the whole population. This group seems to be the most profitable group as it accounts for 12.7% of the total sales amount. Most of the customers in this group have started shopping with the online retailer in the second quarter of the year with an average first_purchase of 7.4, and continued to the end of the year with average recency of 3.1 months since the last purchase. Also, customers in this group seem to shop frequently during the month with an average frequency of 1.3 transactions per month. Thus, this group can also be categorized as high recency and high frequency.

- In contrast, Cluster1 includes 912 customers, representing 21.2% of the whole population and accounts for only 5.2% of the total sales amount. This group seems to be the least profitable group as none of the customers in this group purchased anything in the last five months of the year. Even for the first seven months of the year, the consumers didn't shop often, and the average value of frequency was only 1 transaction per month.

- Cluster2 contains 1138 customers, composed of 26.5% of the whole population, and accounts for 8.7% of the total sales amount. This group includes new customers with an average first_purchase of 3 and average recency of 1.8.

- Cluster5 contains 387 customers, composing 9% of the whole population, and accounts for 10% of the total sales. This segment has fairly high profitability with an average monetary_value of £610 per transaction and moderate frequency with an average of 1.1 transactions per month. What is interesting about this cluster is the large average number of unique items in each transaction. This indicates that most of the customers in this segment are actually organizational customers, not individuals.

Conclusion :

- Further, Customers in Cluster4 are more recent than those in Cluster5. This segment is the largest one, including 32.1% of the whole population and accounts for 38.1% of the total sales. This segment includes the loyal customers who started shopping with the online retailer in the first quarter of the year with an average first_purchase of 10.2 and maintained a moderate purchase frequency and high recency with an average of 1.2 and 1.6.
- Finally, Cluster3 contains about 9.4% of the whole population and accounts for 25.3% of the total sales amount. Customers in this segment shop frequently with an average frequency of 2.1 transactions per month. They also have moderate monetary value, £344.5 per transaction. This segment can be considered the second most profitable segment.

Concluding Remarks -

- Customer segmentation based on the buying pattern of customers through strategically important is an equally challenging task.
- Customer retention is another major concern for both online and physical enterprises.
- In the present work, the RFM model is implemented for synthetic and real datasets, to analyze customer segmentation.
- Also, clusters are evaluated using silhouette score, calinski harabasz score, davies bouldin score for K-Means clustering algorithm with a different number of clusters.
- Based on the Silhouette Score, calinski harabasz score, davies bouldin score the Sales Recency, Sales Frequency and Sales Monetary can be analyzed and an optimal solution is found.
- It has been shown in this analysis that there are few steps in the whole data mining process that are very crucial and the most time-consuming: data preparation, model interpretation and evaluation.

Q & A