DATA MINING REPORT

CLUSTERING, CART, RF, ANN

Contents

Problem 1..5
1.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)...5
1.2. Do you think scaling is necessary for clustering in this case? Justify....……………………..….13
1.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.……...14
1.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters………………………………………..……………………………………..……....20
1.5. Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters...23
Problem 2..25
2.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)..25
2.2. Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network……………………………….....…………………….......…....37
2.3. Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model….………………………………………………………………….…...47
2.4. Final Model: Compare all the models and write an inference which model is best/optimized.......:..56
2.5. Inference: Based on the whole Analysis, what are the business insights and recommendations……………………………………………….………..................................57

List of Figures:

Fig.1–EDA - Univariate Analysis: Dist and boxplot……………………………………………………7
Fig.2-Bivariate and Multivariate Analysis - Correlation Heatmap……………………………………10
Fig.3–Pair Plot:………………………………………………………………………………11
Fig.4-Probability_of_full_payment and Min_payment_amt with Outliers BoxPlot……………………12
Fig.5–Treated outlier Boxplot……………………………………………………………………12
Fig.6-Dendrogram for visualization:…………………………………………………………………17
Fig.7-Dendrogram after Trimming:…………………………………………………………………17
Fig.8–Fclusters Cluster Plots:………………………………………………………………………20
Fig.9-Elbow curve for K-mean:……………………………………………………………………21
Fig.10–Cluster Plots:……………………………………………………………………………23
Fig.11-Outliers: All the four continuous variables have outliers:……………………………………29
Fig.12-EDA - Univariate Analysis: Categorical variables:……………………………………………30
Fig.13–Bivariate Analysis: Categorical variables:……………………………………………………32
Fig.14–Numerical Variable:Univariate Analysis………………………………………………………33
Fig.15-Bivariate Analysis……………………………………………………………………………35
Fig.16–Multivariate Analysis - Correlation Heatmap:…………………………………………………36
Fig.17-Pair Plot: ……………………………………………………………………………………37
Fig.18–CART: Confusion Matrix Train data…………………………………………………………39
Fig.19-Classification_report…………………………………………………………………………39
Fig.20–AUC and ROC curve for the training data:……………………………………………………40
Fig.21-Confusion Matrix Test data: …………………………………………………………………40
Fig.22–Classification_report…………………………………………………………………………40
Fig.23-AUC and ROC curve for the testing data………………………………………………………41
Fig.24–Random Forest: Confusion Matrix Train data::…………………………………………………42
Fig.25-Classification_report:…………………………………………………………………………43
Fig.26-AUC and ROC curve for the testing data:………………………………………………………43
Fig.27-Confusion Matrix Test data: …………………………………………………………………43
Fig.28–Classification_report…………………………………………………………………………44
Fig.29-AUC and ROC curve for the testing data………………………………………………………44
Fig.30–Neural Network Classifier: Confusion Matrix Train data………………………………………46
Fig.31-Classification_report…………………………………………………………………………46
Fig.32–AUC and ROC curve for the training data:……………………………………………………46
Fig.33-Confusion Matrix Test data: …………………………………………………………………47
Fig.34–Classification_report…………………………………………………………………………47
Fig.35-AUC and ROC curve for the testing data………………………………………………………47

Fig.17–CART: Confusion Matrix Train data…..………….…………………………………………………48
Fig.18-Classification_report…….......................................:..............................………………………..49
Fig.19–AUC and ROC curve for the training data:………..…………..………...……………………………..49
Fig.20-Confusion Matrix Test data: …………………………………………….:………..……………....…49
Fig.21–Classification_report…...……….……………………………………………………….…………50
Fig.22-AUC and ROC curve for the testing data…….............................………………………….………….50
Fig.23–Random Forest: Confusion Matrix Train data::……….……...……………………………………..51
Fig.24-Classification_report:..……………………………………….:………..……….…………....…51
Fig.25-AUC and ROC curve for the testing data:..………………………….:………..……………..……....…52
Fig.26-Confusion Matrix Test data: …………………………………………….:………..……………....…52
Fig.27–Classification_report…...……….……………………………………………………….…………52
Fig.28-AUC and ROC curve for the testing data…….............................………………………….………….53
Fig.29–Neural Network Classifier: Confusion Matrix Train data…..………………………….....………...54
Fig.30-Classification_report……............................………………………………...………..…...…………..54
Fig.28–AUC and ROC curve for the training data:…………………...……...…………………………..……..55
Fig.29-Confusion Matrix Test data: ……………………………………...…….……………………....…55
Fig.30–Classification_report…...……….……………………………………………………….…………55
Fig.31-AUC and ROC curve for the testing data…….............................………………………….………….56
Fig.32–ROC curves: CART, RF, ANN Training model:.……...............................………………….…………..57
Fig.33--ROC curves: CART, RF, ANN Testing model:.……..............................……………………….…………..58
Fig.34-ROC-Neural network Test & Train:...……..........................……………………….……………….…..58

List of Tables:

Table 1. Read Data first 5 row……………………………………………….…………………...….…..6
Table2. Describe the data:...……………..……...…..6
Table 3. 5-Point summary stat of scaled data: ………………………………………..…………………..14
Table 4. Fclusters Cluster Added to database ………………………………………………….……………17
Table5. Fclusters Cluster Profiling: ………………....................................……………………………….……17
Table 6. Agglomerative Clustering Cluster Profiling: ………………………….…...………………..……...18
Table 7. Append cluster labels from K-means into the original data: first 5 rows…………………...………...21
Table 8. Cluster Profiling: (K_mean cluster data that we appended to the original data).……….……………23
Table 9. Read Data first 5 row:............…………….....................................……………………………..26
Table 10. Describe the data: Numeric data: …………………..……………….…...……………………...…….27
Table 11. Describe the data: Categorical data:………………..………………..……………...…...…………27
Table 12. 3 Model Compare table……….………………………………………….……………….…………56

Problem 1: **Clustering**

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

Read Data first 5 row:

Spending	*Advance_ payments*	*Probability_ of_full_payment*	*Current_ balance*	*Credit_ limit*	*Min_payment _amt*	*Max_spent_in_s ingle_shopping*
19.94	16.92	0.8752	6.675	3.763	3.252	6.55
15.99	14.89	0.9064	5.363	3.582	3.336	5.144
18.95	16.42	0.8829	6.248	3.755	3.368	6.148
10.83	12.96	0.8099	5.278	2.641	5.182	5.185
17.99	15.86	0.8992	5.89	3.694	2.068	5.837

Data information:

RangeIndex: 210 entries, 0 to 209

Data columns (total 7 columns):

#	Column	Non-Null Count	Dtype
0	spending	210 non-null	float64
1	advance_payments	210 non-null	float64
2	probability_of_full_payment	210 non-null	float64
3	current_balance	210 non-null	float64
4	credit_limit	210 non-null	float64
5	min_payment_amt	210 non-null	float64

6 max_spent_in_single_shopping 210 non-null float64

dtypes: float64(7)

memory usage: 11.6 KB

Interpretation:

- ***The data has 210 rows and 7 columns with Float as the data type for each.***
- ***We have continuous, non-null data with 7 variables.***

Describe the data:

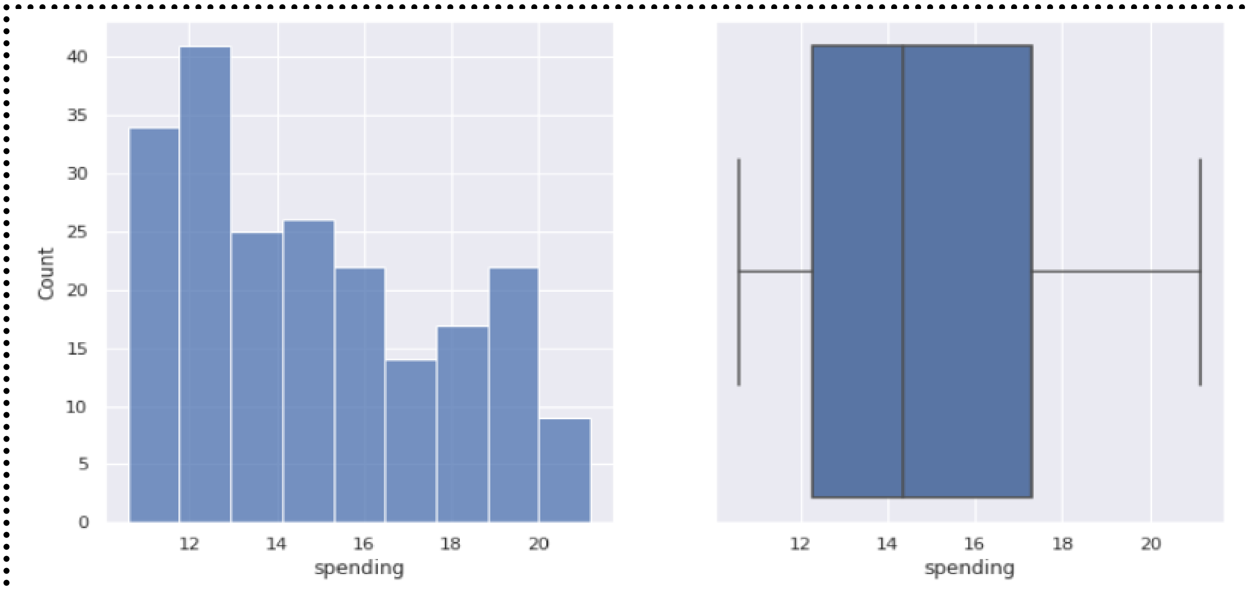
	Count	Mean	STD	MIN	25.00%	50.00%	75.00%	MAX
Spending	210	14.847524	2.909699	10.59	12.27	14.355	17.305	21.18
Advance_payments	210	14.559286	1.305959	12.41	13.45	14.32	15.715	17.25
Probability_of_full_payment	210	0.870999	0.023629	0.8081	0.8569	0.87345	0.887775	0.9183
Current_balance	210	5.628533	0.443063	4.899	5.26225	5.5235	5.97975	6.675
Credit_limit	210	3.258605	0.377714	2.63	2.944	3.237	3.56175	4.033
Min_payment_amt	210	3.700201	1.503557	0.7651	2.5615	3.599	4.76875	8.456
Max_spent_in_single_shopping	210	5.408071	0.49148	4.519	5.045	5.223	5.877	6.55

- ***The summary stats: average spending is approximate 14800/-***
- ***The average advance payment done by a customer is approximate 1400/- which is almost 10% of the monthly spend.***
- ***The average max spend in one purchase is approximate 5400/- and the minimum paid amount by customer during purchase is approximate 370/-***

EDA - Univariate Analysis

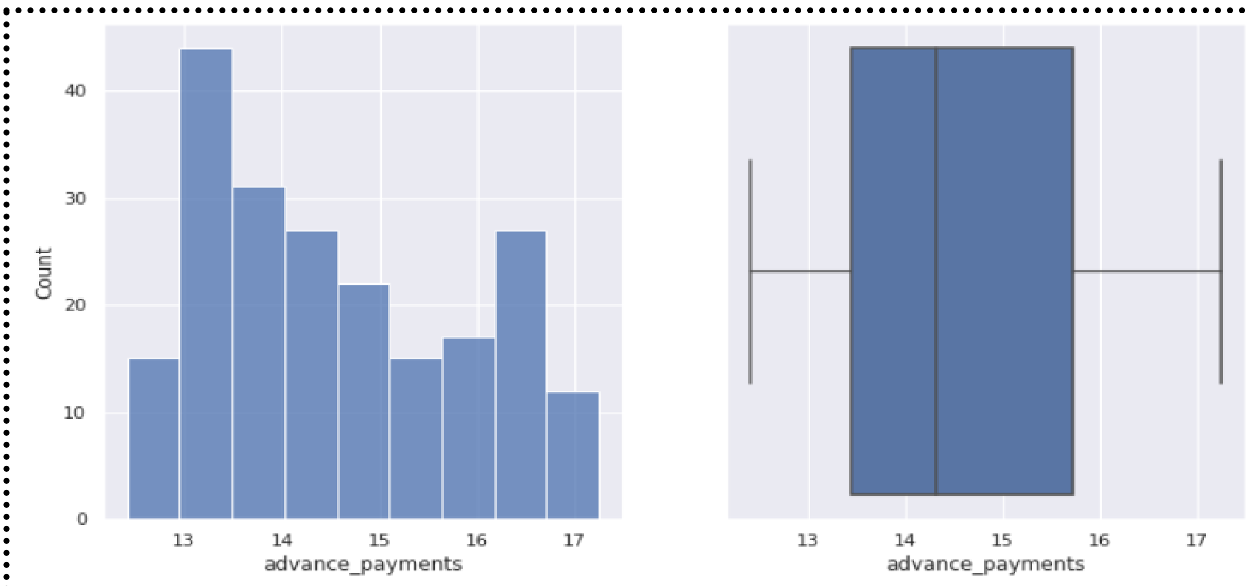
- ***Our objective is to derive the data, define and analyze the pattern present in each variable separately.***

Dist and boxplot of all variables: SPENDING



- ***The Boxplot tells us there are no outliers Spending distribution.***
- ***The distplot distribution can be said to be slightly left skewed. Skewness(spending) is 0.397. The distribution ranges between 11 to 20.***

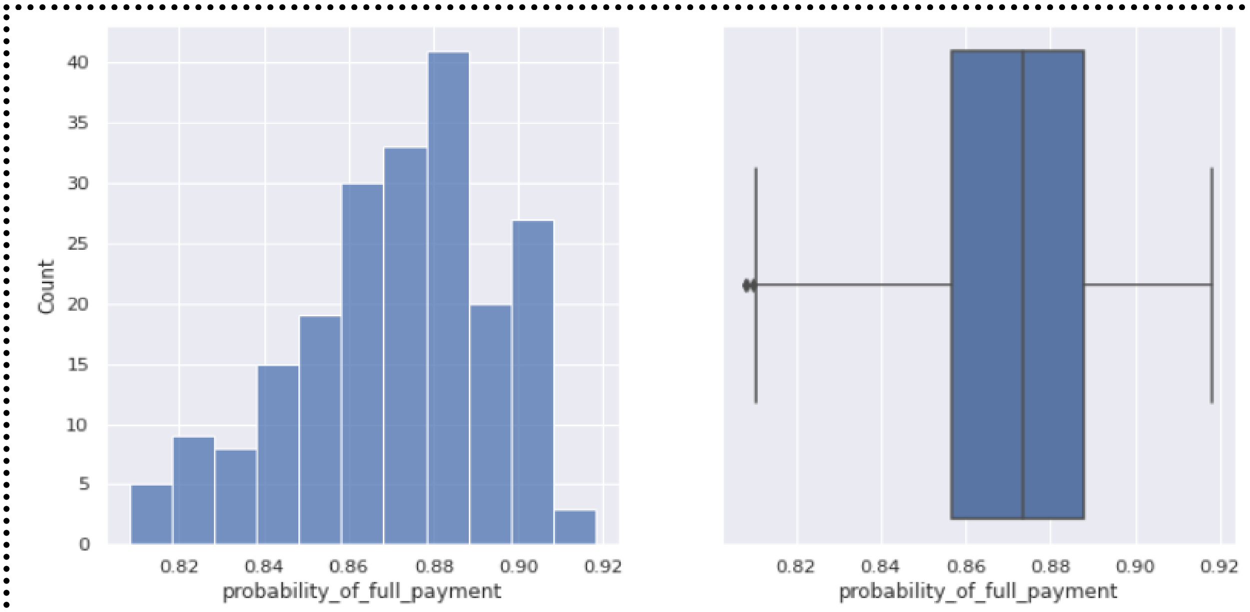
ADVANCE_PAYMENTS



- ***The Boxplot tells us there are no outliers advance_payments distribution.***

➢ ***The distplot distribution can be said to be slightly left skewed. Skewness(advance_payments) is 0.384. The distribution ranges between 12 to 17(100s).***

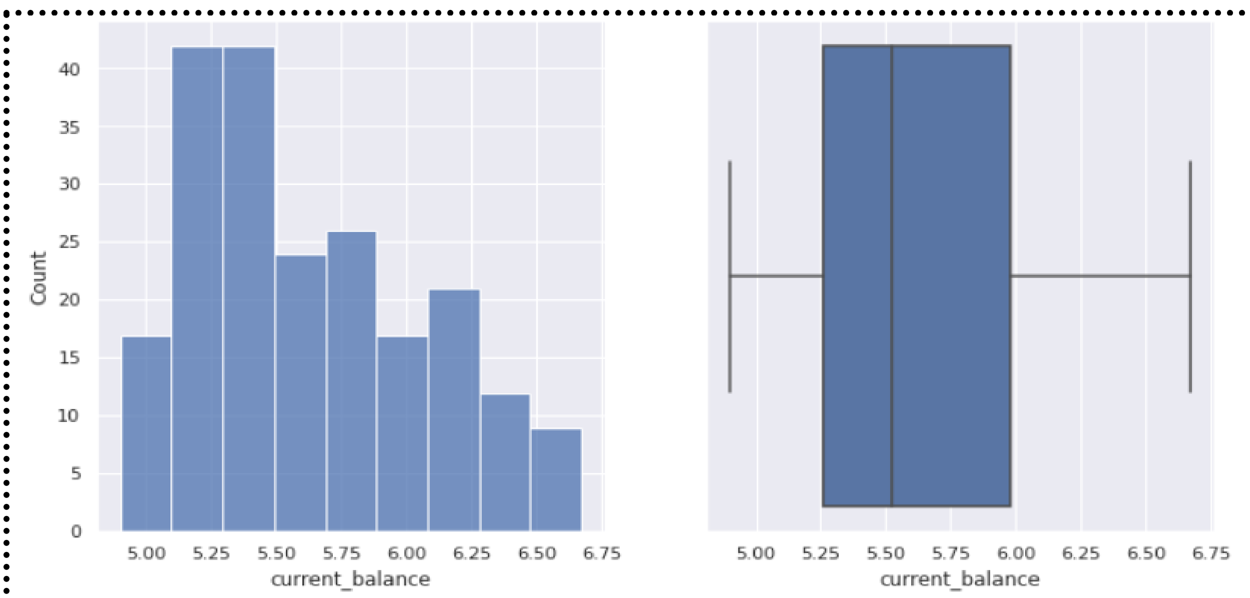
PROBABILITY OF FULL PAYMENT



➢ ***The Boxplot tells us there are few outliers for probability_of_full_payment distribution.***

➢ ***The distplot distribution can be said to be slightly right skewed. Skewness (probability_of_full_payment) is -0.534. The distribution ranges between 0.80 to 0.92.***

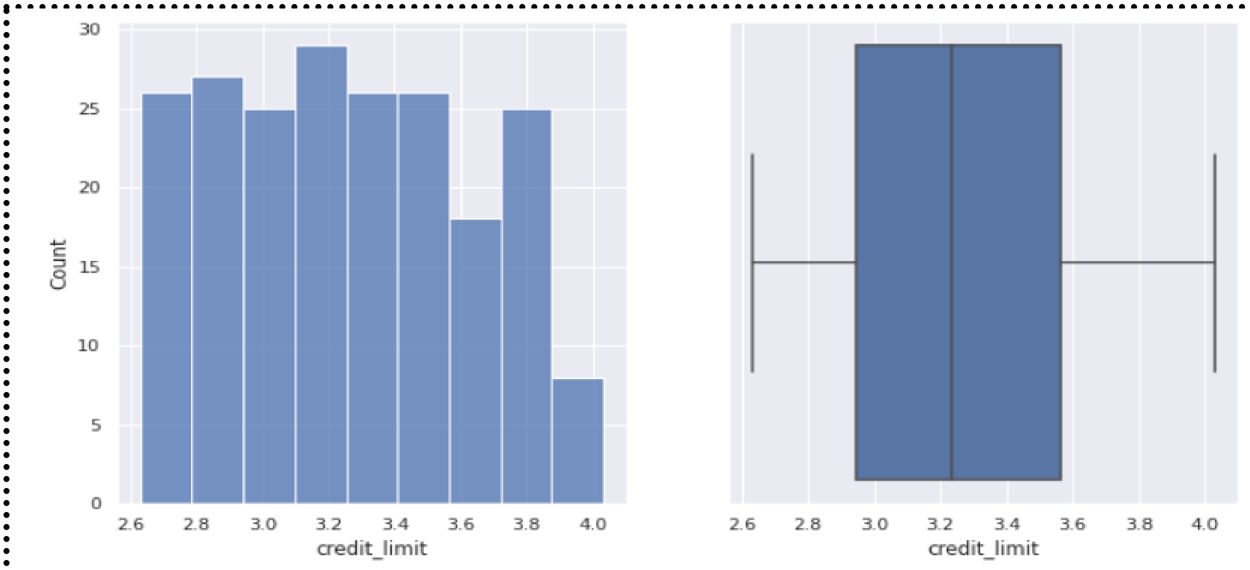
CURRENT BALANCE



➢ ***The Boxplot tells us there are no outliers for current_balance distribution.***

➢ ***The distplot distribution can be said to be slightly left skewed. Skewness (current_balance) is 0.522. The distribution ranges between 5 to 6.7(1000s).***

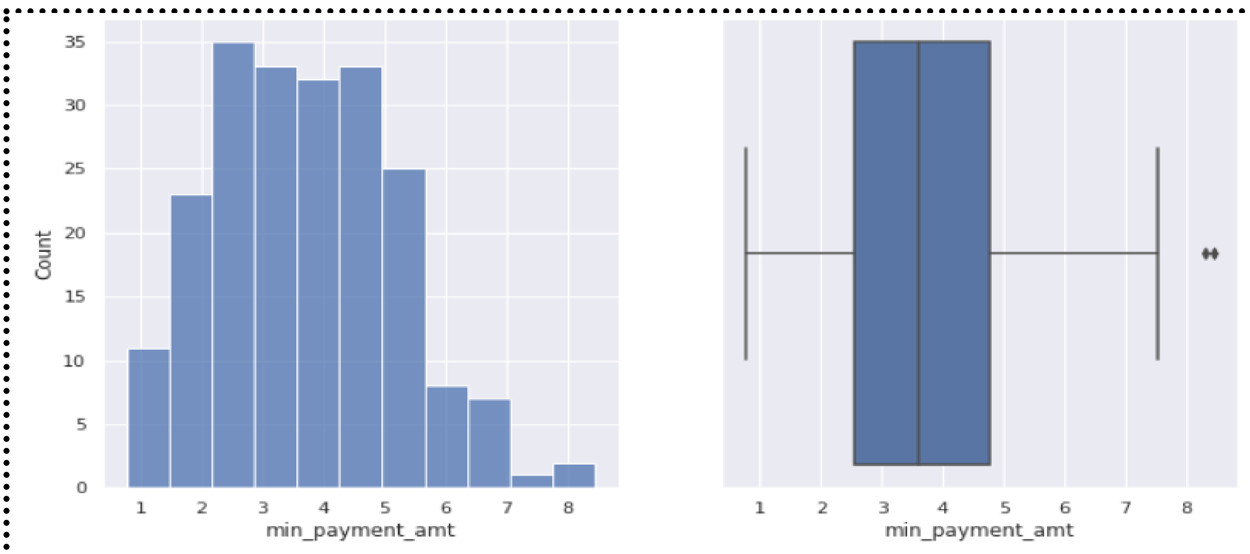
CREDIT LIMIT



➢ ***The Boxplot tells us there are no outliers for credit_limit distribution.***

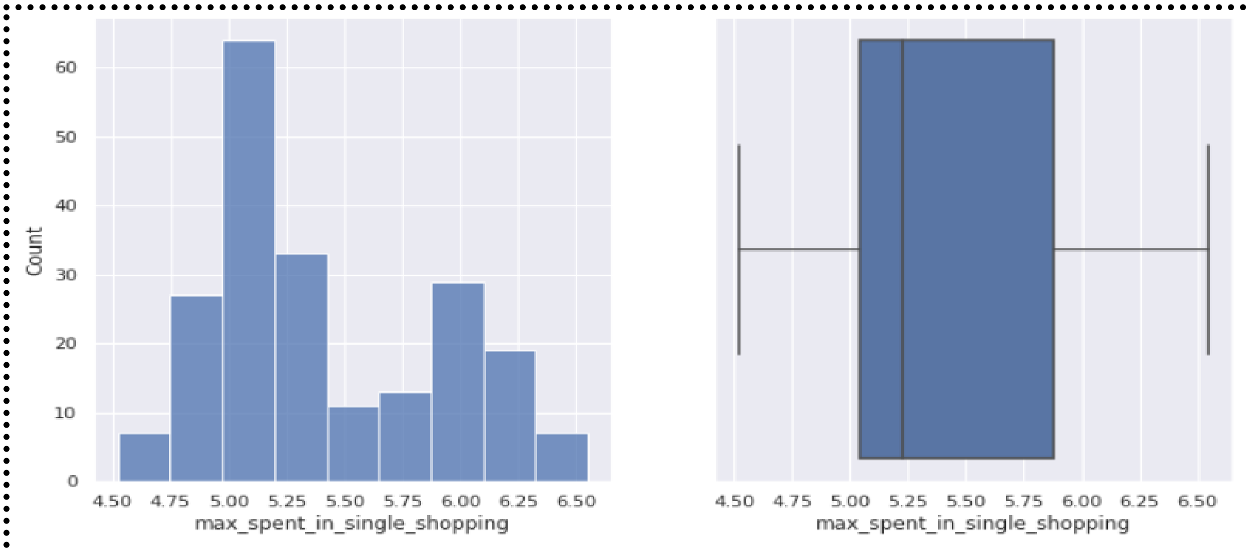
➢ ***The distribution can be said to be normally distributed. Skewness (credit_limit) is 0.133. The distribution ranges between 2.6 to 4 (10000s)***

MIN PAYMENT AMT



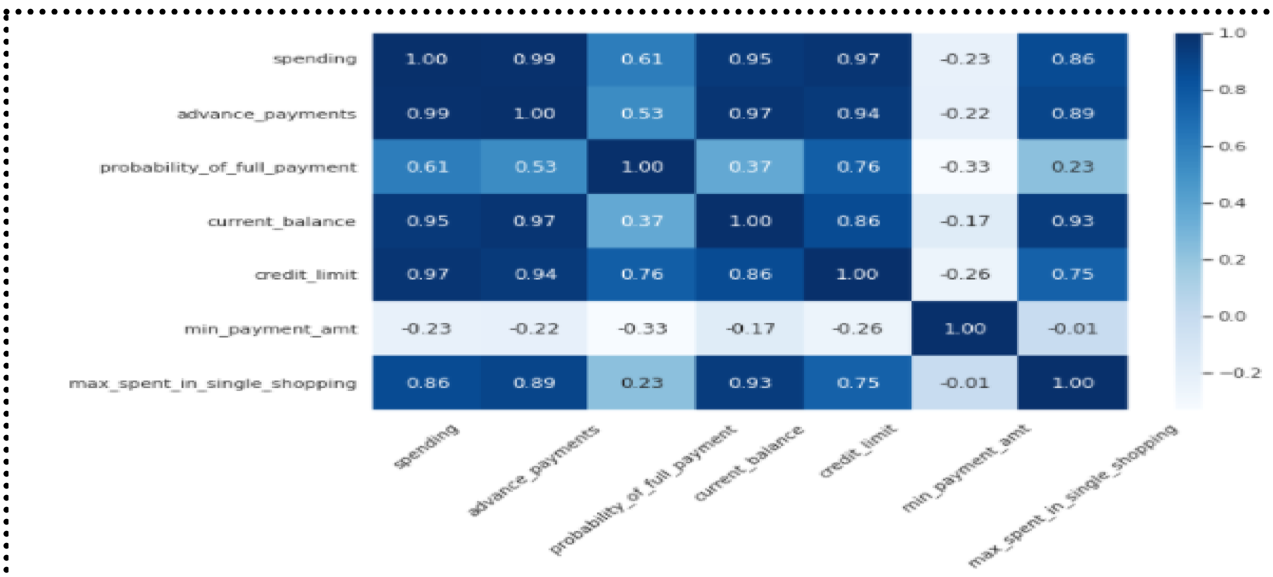
- ➢ ***The Boxplot tells us there are few outliers for min_payment_amt distribution.***
- ➢ ***The distribution can be said to be slightly left skewed. Skewness (min_payment_amt) is 0.399. The distribution ranges between 1 to 8 (100s)***

MAX SPEND IN SINGLE SHOPPING



- ➢ ***The Boxplot tells us there are no outliers for max_spent_in_single_shopping distribution.***
- ➢ ***The distribution can be said to be slightly left skewed. Skewness (max_spent_in_single_shopping) is 0.558. The distribution ranges between 4.5 to 6.5 (1000s)***

Bivariate and Multivariate Analysis - Correlation Heatmap:



➢ ***The relation between pairs of numeric variables is given by the heatmap.***

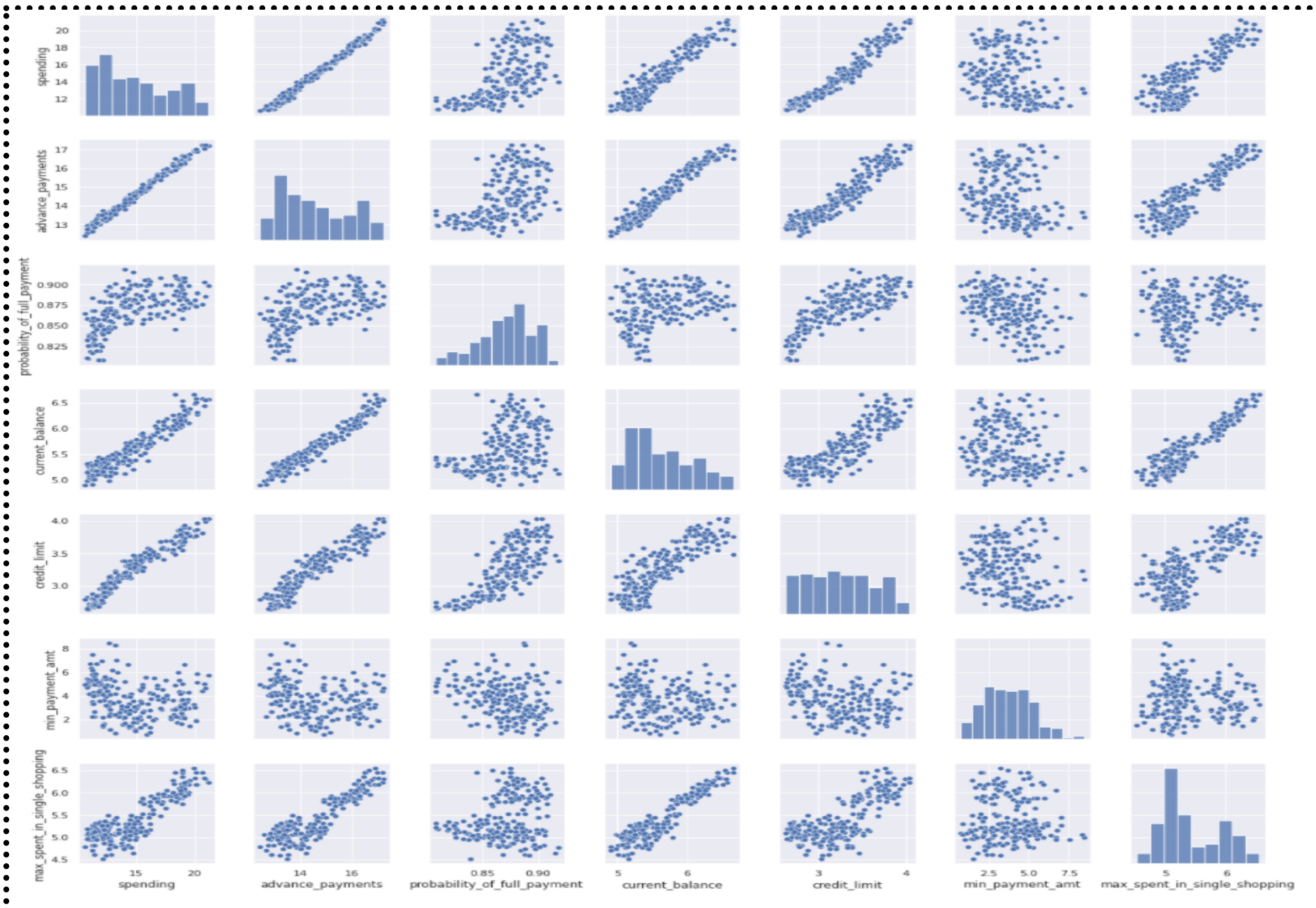
➢ ***The correlation between the following variables are highly positive: (variable are directly proportional)***

- ***Spending and Advance Payment***
- ***Current Balance and Advance Payment***
- ***Spending and Credit Limit***
- ***Spending and Current Balance***

➢ ***The correlation between the following variables are negative: (variable are inversely proportional)***

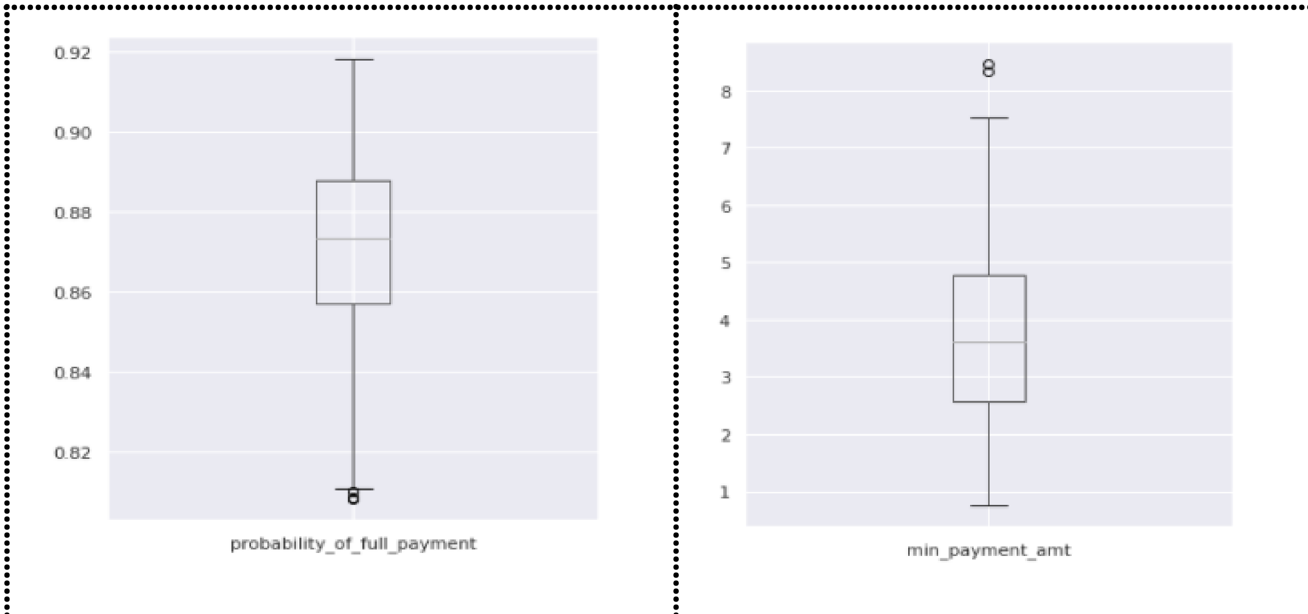
- ***Min Payment Amt and Probability Of Full Payment***
- ***Min Payment Amt and Credit Limit***
- ***Min Payment Amt and Spending***
- ***Min Payment Amt and Advance Payment***

Pair Plot:



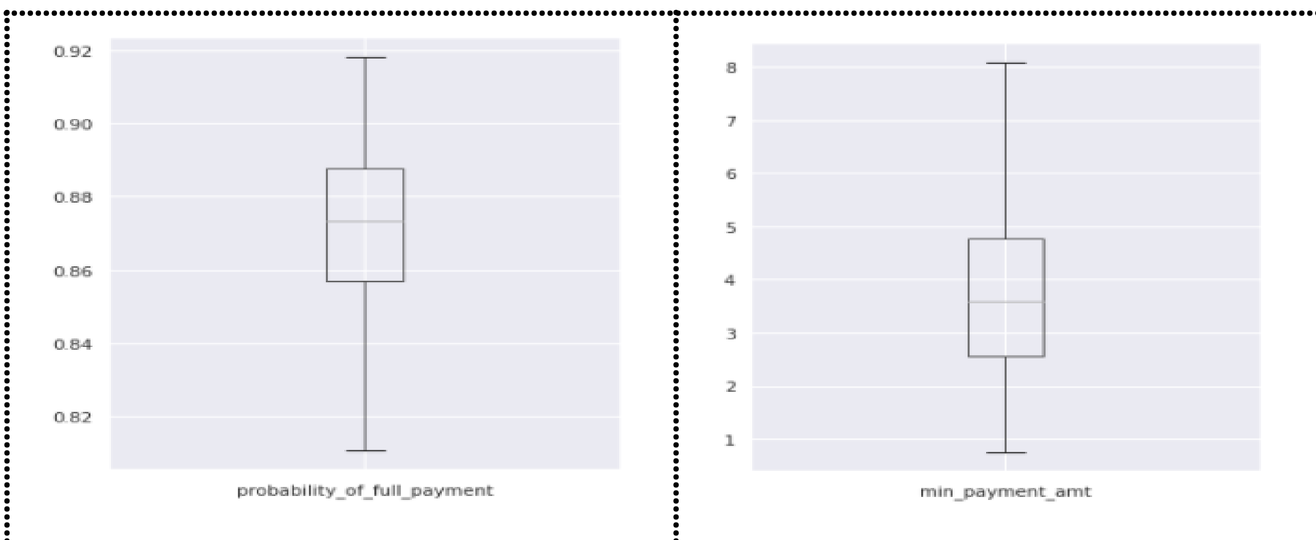
- ➢ ***We can see the same relation pattern(Strong correlation) between the variables in the Pair Plot above.***
- ➢ ***Spending & Advance payment by cash increases with credit card amount limit and current balance amount left in the account to make purchases.***
- ➢ ***We can see outliers in only two variables: Probability_of_full_payment and Min_payment_amt.***

Probability_of_full_payment and Min_payment_amt with Outliers:



- ➢ ***As we have few outliers, we will replace the outlier values for each variable using the IQR(Q1 and Q3) respectively.***

Outcome: No more outliers in the data:



1.2 Do you think scaling is necessary for clustering in this case? Justify.

- ***Scaling is used to eliminate redundancy in data during clustering and ensures that good quality clusters are generated.***
- ***As in this case, we see that though numeric, our variables vary in dimension (having different weight), which would lead some variables to have more weightage on the outcome than others.***
- ***For example if we look at our data summary we see:***
  - ***Standard Deviation(std) of Spending and Advance_Payments and Min_payment_amt is high (2.91, 1.30 and 1.50) when compared to others.***
  - ***Probability_of_full_payment(std) being the lowest. (0.023)***
- ***Same if we look at the variance:***
  - ***spending 8.43***
  - ***advance_payments 1.70***
  - ***probability_of_full_payment 0.00***
  - ***current_balance 0.20***
  - ***credit_limit 0.14***
  - ***min_payment_amt 2.22***
  - ***max_spent_in_single_shopping 0.24***
- ***As we do have large variance data, we do scale our variables. So as to avoid(creating a bias), the variables with largest variance have disproportionately more influence on cluster outcome.***
- ***Hence before clustering all the variables should be scales to have the same weight(unit). Can be done by:***
  - ***Z-score method***
  - ***Min-Max method***
- ***Using Min-Max method method to scale the 7 variables here:***
- ***Min-Max method has now brought the data closer and decreased the variance between them: We can see the change in std and variance after scaling. (between 0 & 1)***
- ***We look at the variance after scaling:***

- ***spending*** ***0.08***
- ***advance_payments*** ***0.07***
- ***probability_of_full_payment*** ***0.05***
- ***current_balance*** ***0.06***
- ***credit_limit*** ***0.07***
- ***min_payment_amt*** ***0.04***
- ***max_spent_in_single_shopping*** ***0.06***

5-Point summary stat of scaled data:

	Count	Mean	STD	MIN	25.00%	50.00%	75.00%	MAX
Spending	210	0.4	0.27	0	0.16	0.36	0.63	1
Advance_payments	210	0.44	0.27	0	0.21	0.39	0.68	1
Probability_of_full_payment	210	0.56	0.22	0	0.43	0.58	0.72	1
Current_balance	210	0.41	0.25	0	0.2	0.35	0.61	1
Credit_limit	210	0.45	0.27	0	0.22	0.43	0.66	1
Min_payment_amt	210	0.4	0.2	0	0.25	0.39	0.55	1
Max_spent_in_single_shopping	210	0.44	0.24	0	0.26	0.35	0.67	1

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Clustering is a method of grouping similar objects into classes.

It is used to convert data into structures that can be easily understood and manipulated.

There are two types:

- ***Hierarchical clustering***
- ***Partitioning clustering***

We will be using Hierarchical clustering.

- *It groups similar objects into groups called clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.*

It has a Bottom Up approach: We start by treating each object as a separate cluster. Then, repeatedly executes the following two steps:

(1) Identify the two clusters that are closest to each other

(2) Merge the two most similar clusters.

For the 1st step we measure distance - hierarchical clustering, including the following:

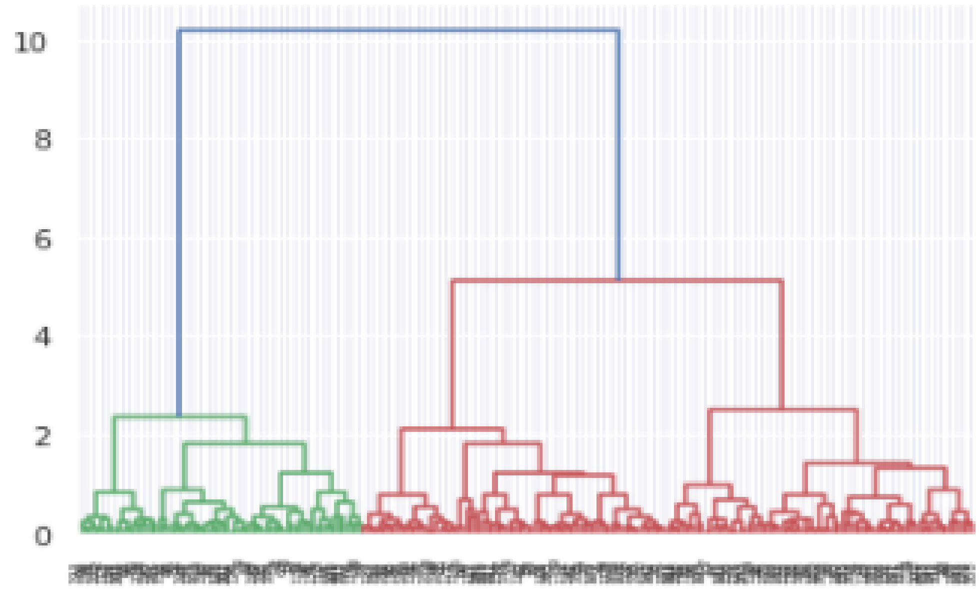
- *Euclidean Distance*
  - *It follows pythagoras theorem to find shortest distance between two points: x1(i, j) and x2(v,u), then distance = √(i-v)^2+(j-u)^2*
- *Manhattan distance*
  - *It is a distance between the projection of points(modulus): x1(i, j) and x2(v,u), then distance = \i-v\+\j-u*

For the 2nd step we merge clusters by similarity - Using linkage methods:

- *Single Linkage*
  - *Two clusters with the closest minimum distance are merged*
- *Complete Linkage*
  - *Two clusters with the closest maximum distance are merged*
- *Centroid Linkage*
  - *Two clusters with the lowest centroid distance are merged.*
- *Ward's Linkage*
  - *Two clusters are merged based on their error sum of square values.*
- *Average Linkage*
  - *Average linkage method uses the average pairwise proximity among all pairs of objects in different clusters. Clusters are merged based on their lowest average distances.*

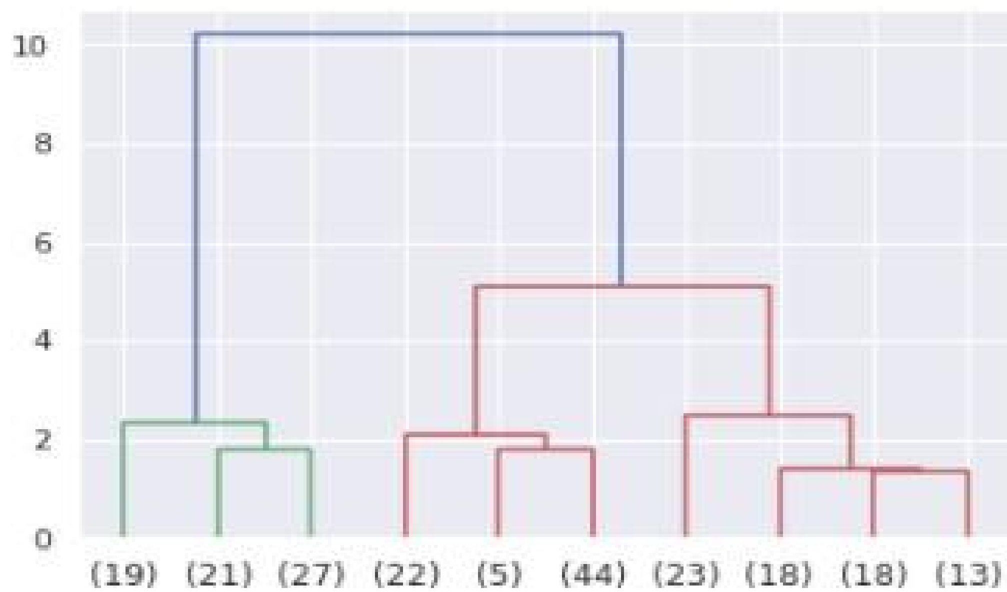
I will be using **Ward's Linkage** *method for merge on scaled data. And* **Euclidean** *to find the shortest distance, as that is one of the commonly used and default.*

Here is the Dendrogram for visualization:



- ***Next step would be to choose the number of clusters to go with.***
- ***By using '`lastp`' where p=10, we get a clear picture of our structure above:***

Dendrogram after Trimming:



- ***We can conclude by getting 2 main clusters, however as 2 clusters are not optimal for industrial use, we go with 3 clusters here as our optimal clusters count.***

❖ ***Proceeding with Hierarchical Clustering implementation:***

➢ ***Using Fclusters to proceed:***

➢ ***Number of clusters: 3, Affinity: ‘Euclidean’ and Linkage: ‘Ward’***

Fclusters Cluster Added to database:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	0.88	0.93	0.60	1.00	0.81	0.34	1.00	1
1	0.51	0.51	0.89	0.26	0.68	0.35	0.31	3
2	0.79	0.83	0.67	0.76	0.80	0.36	0.80	1
3	0.02	0.11	0.00	0.21	0.01	0.60	0.33	2
4	0.70	0.71	0.82	0.56	0.76	0.18	0.65	1

Fclusters Cluster Frequency:

1 67

2 71

3 72

Name: clusters, dtype: int64

Fclusters Cluster Profiling:

Clusters	***Spending***	***Advance_payments***	***Probability_of_full_payment***	***Current_balance***	***Credit_limit***	***Min_payment_amt***	***Max_spent_in_single_shopping***	***Freq***
1	0.74	0.78	0.68	0.72	0.76	0.4	0.75	67
2	0.13	0.19	0.36	0.2	0.17	0.56	0.29	71
3	0.35	0.39	0.65	0.34	0.43	0.24	0.29	72

- ***For banks Tier 1 users are high spenders, as well as their Maximum amount spent in one purchase, along with their probability of payment done in full by the customer to the bank is high.***
- ***For banks Tier 2 users are low monthly spend, their Credit limit, probability of payment done in full by the customer to the bank is also on the lower end, when compared to other clusters.***
- ***For banks Tier 3 users are medium monthly spend, their probability of payment done in full by the customer to the bank is high.***
- ***Following the same for Agglomerative Clustering we get equal outcome:***

Cluster Frequency:

0 72

1 67

2 71

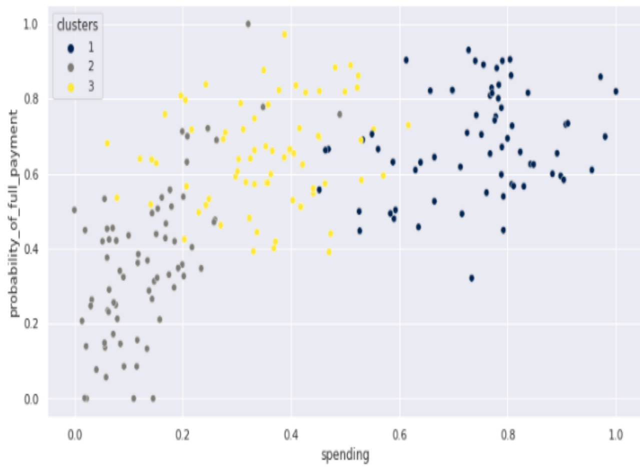
Name: Agglo_CLusters, dtype: int64

Agglomerative Clustering Cluster Profiling:

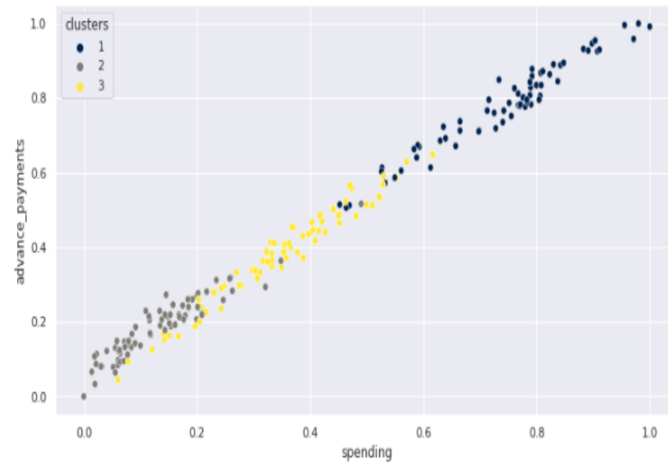
Agglo_CLusters	*Spending*	*Advance_payments*	*Probability_of_full_payment*	*Current_balance*	*Credit_limit*	*Min_payment_amt*	*Max_spent_in_single_shopping*	*Freq*
0	0.13	0.19	0.36	0.2	0.17	0.56	0.29	71
1	0.74	0.78	0.68	0.72	0.76	0.4	0.75	67
2	0.35	0.39	0.65	0.34	0.43	0.24	0.29	72

Fclusters Cluster Plots:

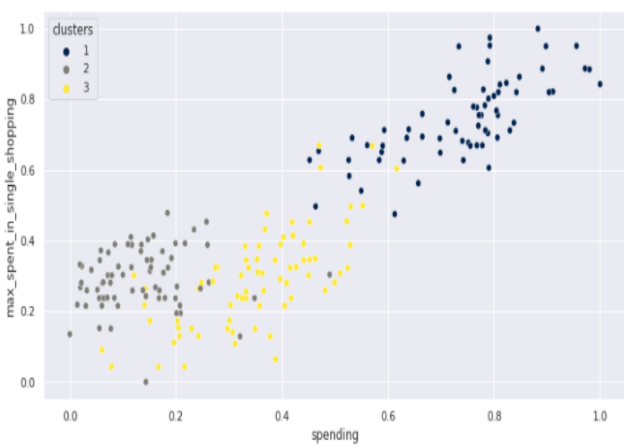
Plot:1



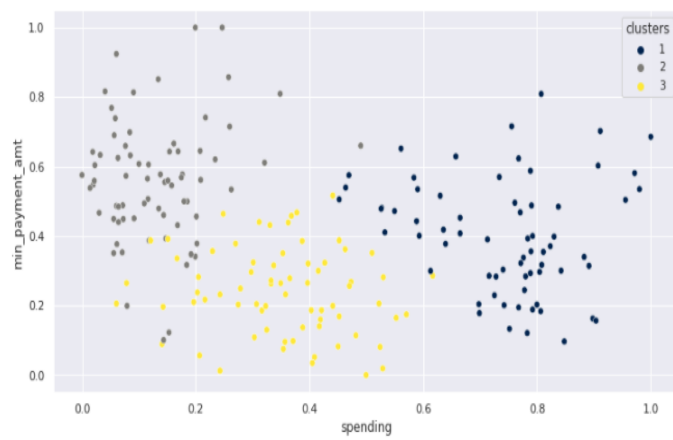
Plot:2



Plot:3



Plot:4

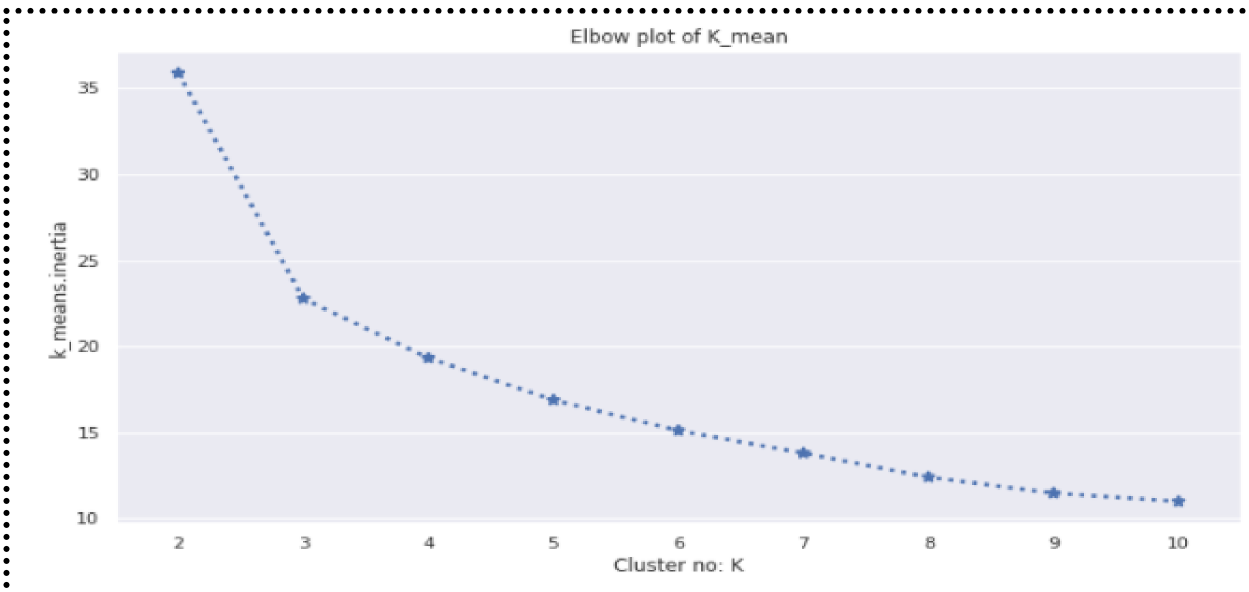


- ***We are able to make out the variance in each cluster plot and identify the cluster groups.***
- ***The cluster follows the same path as all the plots, 1 being on the higher end of the spectrum and 2 on the lower end.***
- ***This can be used by the bank to narrow down user behavior in each cluster group.***

1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

- ***K-means clustering uses "centroids", K different randomly-initiated points in the data, and assigns every data point to the nearest centroid.***
- ***In this case, will start with K=3 to proceed.***
  - ***It's Within Cluster Sum of Squares as :*** *22.770013026036565*
- ***Calculating WSS for other values of K :***
  - ***K=2 k_mean inertia is 35.86521074048986***
  - ***K=3 k_mean inertia is 22.770013026036565***
  - ***K=4 k_mean inertia is 19.323702051792594***
  - ***K=5 k_mean inertia is 16.88519118531759***
  - ***K=6 k_mean inertia is 15.1045576826233***
  - ***K=7 k_mean inertia is 13.802943437508269***
  - ***K=8 k_mean inertia is 12.40244683471637***
  - ***K=9 k_mean inertia is 11.480004867341448***
  - ***K=10 k_mean inertia is 10.995733610492227***
- ***To find the optimum K value I have created an elbow curve.***

Elbow curve for K-mean:



- ➢ *We can see that after 3, there is an almost constant decrease in K_mean inertia value.*
- ➢ *Hence will take 3 as the optimum number of clusters for the given data.*

❖ ***Silhouette Score:***

- ➢ *Silhouette Score is a metric used to calculate the goodness of a clustering technique.*
- ➢ *Its value ranges from -1 to 1.*
- ➢ *Silhouette Score for our optimal K=3 is: 0.41888372435617*

❖ ***Cluster evaluation for all above k values, the silhouette score:***

- ➢ `K=2 silhouette_score is 0.5000644323521964`
- ➢ `K=3 silhouette_score is 0.41888372435617`
- ➢ `K=4 silhouette_score is 0.336795122525894`
- ➢ `K=5 silhouette_score is 0.28630579552105506`
- ➢ `K=6 silhouette_score is 0.2894420745589243`
- ➢ `K=7 silhouette_score is 0.26322890706145396`
- ➢ `K=8 silhouette_score is 0.25966810176636784`
- ➢ `K=9 silhouette_score is 0.26177716346826413`
- ➢ `K=10 silhouette_score is 0.24801301630025002`

❖ ***As 2 clusters with the highest score, cannot be used for industrial application, will select the next high score value as we got above.***

Append cluster labels from K-means into the original data: first 5 rows:

Spending	*Advance_payments*	*Probability_of_full_payment*	*Current_balance*	*Credit_limit*	*Min_payment_amt*	*Max_spent_in_single_shopping*	*KMCluster*
19.94	16.92	0.8752	6.675	3.763	3.252	6.55	1
15.99	14.89	0.9064	5.363	3.582	3.336	5.144	0
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2
17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1

Cluster Frequency:

0 69

1 64

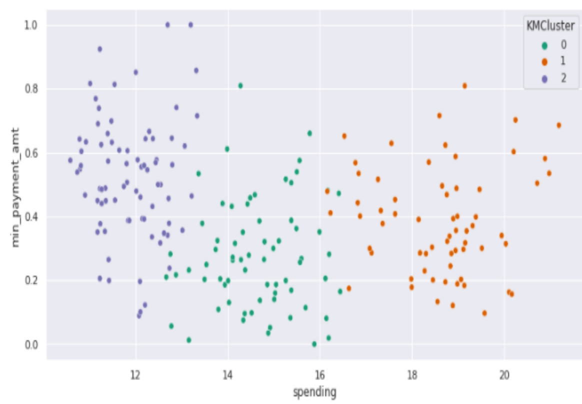
2 77

Name: KMCluster, dtype: int64

- ***K mean has divided our data into 3 clusters.***
- ***The data is distributed almost equally. Cluster 2 having more data compared to the other two.***

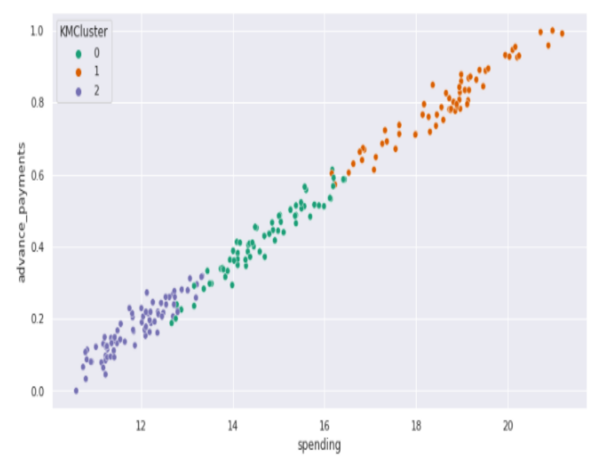
Cluster Plots:

Plot:1



Plot:2

Plot:3



Plot:4

Inference:

- ***We get three cluster and if we look at the cluster profiling we can see that:***
  - ***Tier 0 is has 69 users***
  - ***Tier 1 is has 64 users***
  - ***Tier 2 is has the highest 77 users***
- ***The graph results are similar from our earlier model.***
- ***The above plots show a pattern format followed in each pair graph, with Tier 2 on the lower end and Tier end at the high end.***
- ***We can conclude the data is properly clustered, benefiting further analysis.***

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

- ***Now performing Cluster Profiling for our K_mean cluster data that we appended to the original data.***

Cluster Profiling: (*K_mean cluster data that we appended to the original data*)

KMCluster	Spending	Advance_payments	Probability_of_full_payment	Current_balance	Credit_limit	Min_payment_amt	Max_spent_in_single_shopping	Freq
0	14.65	14.44	0.88	5.55	3.29	2.8	5.17	69
1	18.61	16.25	0.88	6.2	3.71	3.59	6.06	64
2	11.9	13.26	0.85	5.23	2.86	4.59	5.09	77

- ***We have three groups with their frequencies displayed in the above table.***
  - ***Tier 0 is (medium)***
  - ***Tier 1 is (high)***
  - ***Tier 2 is (low)***

- ***By viewing the above data and analyzing it mean there is a pattern, which is similar to the pattern one sees during Fcluster.***

Inferences:

- ***The Group 1 or Tier 1 data has of Kmean:***
    - ***High Spending, High Advance_payments, High Probability_of_full_payment, High Current_balance, High Credit_limit, Medium Min_payment_amt and High Max_spent_in_single_shopping.***
    - ***This pattern matched Agglomerative Cluster's Group 1 or Tier 1 data.***
- ***The Group 0 or Tier 0 data has of Kmean:***
    - ***Medium Spending, Medium Advance_payments, High Probability_of_full_payment, Medium Current_balance, Medium Credit_limit, Low Min_payment_amt and Medium Max_spent_in_single_shopping.***
    - ***This pattern matched Agglomerative Cluster's Group 2 or Tier 2 data.***
- ***The Group 2 or Tier 2 data has of Kmean:***
    - ***Low Spending, Low Advance_payments, Low Probability_of_full_payment, Low Current_balance, Low Credit_limit, High Min_payment_amt and Low Max_spent_in_single_shopping.***
    - ***This pattern matched Agglomerative Cluster's Group 0 or Tier 0 data.***
- ***Kmean Group 0 = Agglomerative Group 2***
- ***Kmean Group 1 = Agglomerative Group 1***
- ***Kmean Group 2 = Agglomerative Group 0***

Recommendation:

- **For Group 1 -**
    - ***High Spenders - give purchase credit points(Loyal customer base)***
    - ***Medium Min_payment_amt, High Advance_payments - offer premium on credit card.***
    - ***High Current_balance - offer higher credit limits, would encourage purchases.***
    - ***High Probability_of_full_payment - offer membership & exclusive promotion on loyalty.***
    - ***High Max_spent_in_single_shopping - pitching more brands and products would benefit, as they have higher conversion rate.***

- ❖ **For Group 0**
  - ➢ ***Medium Spending, Medium Advance_payments - offering promotions, price match offers.***
  - ➢ ***Medium Credit_limit, Low Min_payment_amt - offering membership deals, cashbacks and low interest loans, credit purchases.***
  - ➢ ***Medium Max_spent_in_single_shopping.***
  - ➢ ***High Probability_of_full_payment - rewarding loyalty points, keeping them updates on upcoming offers and detail, would encourage customers to make more purchases.***
- ❖ ***For Group 2***
  - ➢ ***High Min_payment_amt and low on every other variable, shows the customer is not spending though the bank much.***
  - ➢ ***Need to maintain timely engagements with the customer to let them know about new offers and deals.***
  - ➢ ***Reward on purchase and repayment of the credit would improve engagements.***
  - ➢ ***Customers here in this group have to be given extra attention, as they are more likely to churn.***

Problem 2: **CART-RF-ANN**

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

Data Shape:

Rows and columns: (3000, 10)

Data information:

```
#    Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Age            3000 non-null   int64
 1   Agency_Code    3000 non-null   object
 2   Type           3000 non-null   object
 3   Claimed        3000 non-null   object
 4   Commision      3000 non-null   float64
 5   Channel        3000 non-null   object
 6   Duration       3000 non-null   int64
 7   Sales          3000 non-null   float64
 8   Product Name   3000 non-null   object
 9   Destination    3000 non-null   object
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

Read Data first 5 row:

Age	Agency_Code	Type	Claimed	Commis s-ion	Chann -el	Durati -on	Sales	Product Name	Destination
48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
36	EPX	Travel Agency	No	0	Online	34	20	Customised Plan	ASIA
39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
36	EPX	Travel Agency	No	0	Online	4	26	Cancellation Plan	ASIA
33	JZI	Airlines	No	6.3	Online	53	18	Bronze Plan	ASIA

Interpretation:

- ***The data has 3000 rows and 10 columns with 2 floats, 2 int and 6 object as the data type.***
- ***There are no null values.***
- ***Independent variable:***

- ➢ *We have 4 continuous variables (Age, Commision, Duration, Sales)*
- ➢ *5 categorical variables (Agency_Code, Type, Channel, Product Name, Destination).*

Describe the data: Numeric data:

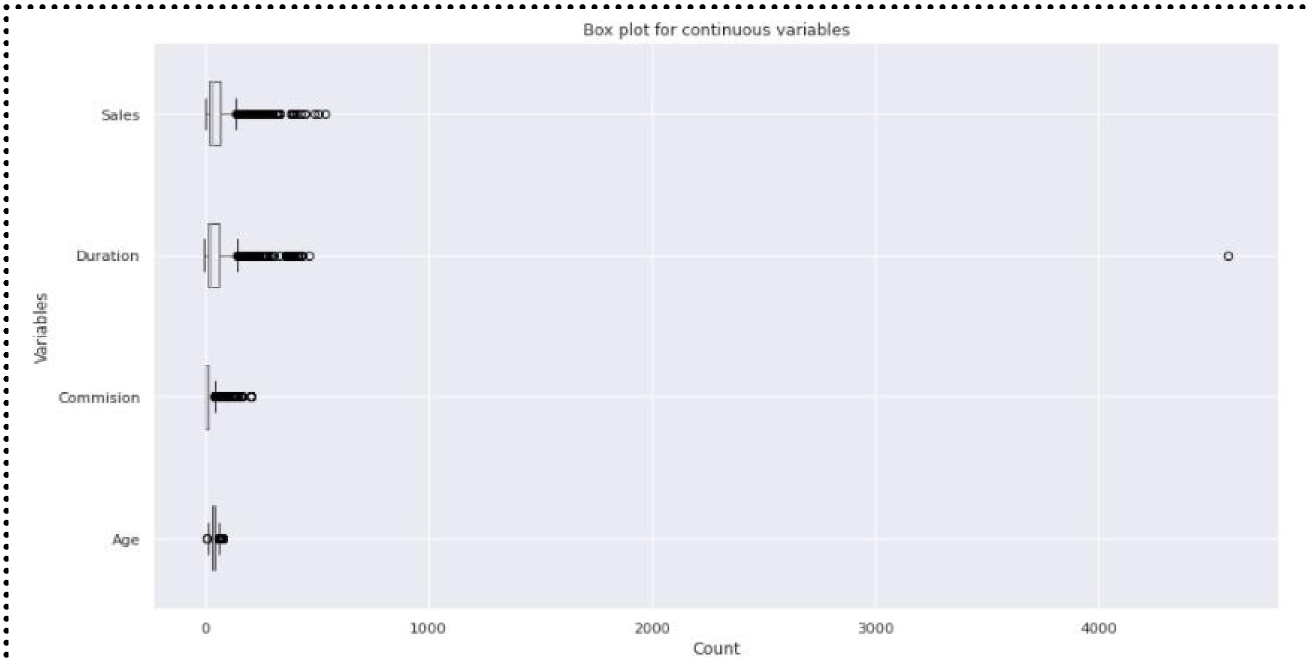
	Count	Mean	STD	MIN	25.00%	50.00%	75.00%	MAX
Age	3000	38.091	10.463518	8	32	36	42	84
Commision	3000	14.529203	25.481455	0	0	4.63	17.235	210.21
Duration	3000	70.001333	134.053313	-1	11	26.5	63	4580
Sales	3000	60.249913	70.733954	0	20	33	69	539

Describe the data: Categorical data:

	Count	Unique	Top	Freq
Agency_Code	3000	4	EPX	1365
Type	3000	2	Travel Agency	1837
Claimed	3000	2	No	2076
Channel	3000	2	Online	2954
Product Name	3000	5	Customised Plan	1136
Destination	3000	3	ASIA	2465

- ❖ *Claimed is our Dependent variable :*
  - ➢ *We see majority is 'No' and it's frequence is 2076*
- ❖ *Most types are Travel Agency frequency : 1365 and most prefer Channel is Online.*
- ❖ *There are 139 duplicate rows. Will be dropping them, there is a possibility that the data would be from different customers. As there is no unique Identifier to check this and we have 3000 rows, for this case. I will be dropping the duplicate.*
- ❖ *The data shape is now:* `(2861, 10)`.

Outliers: All the four continuous variables have outliers:



- ❖ ***Would need to treat outliers during Neural Networks.***
- ❖ ***Value counts of Categorical Variables***
  - ➢ ***AGENCY_CODE : 4***

    JZI 239

    CWT 471

    C2B 913

    EPX 1238

    Name: Agency_Code, dtype: int64
  - ➢ ***TYPE : 2***

    Airlines 1152

    Travel Agency 1709

    Name: Type, dtype: int64
  - ➢ ***CLAIMED : 2***

    Yes 914

    No 1947

    Name: Claimed, dtype: int64
  - ➢ ***CHANNEL : 2***

    Offline 46

Online 2815

Name: Channel, dtype: int64

- ***PRODUCT NAME : 5***

  Gold Plan 109

  Silver Plan 421

  Cancellation Plan 615

  Bronze Plan 645

  Customised Plan 1071

  Name: Product Name, dtype: int64

- ***DESTINATION : 3***

  EUROPE 215

  Americas 319

  ASIA 2327

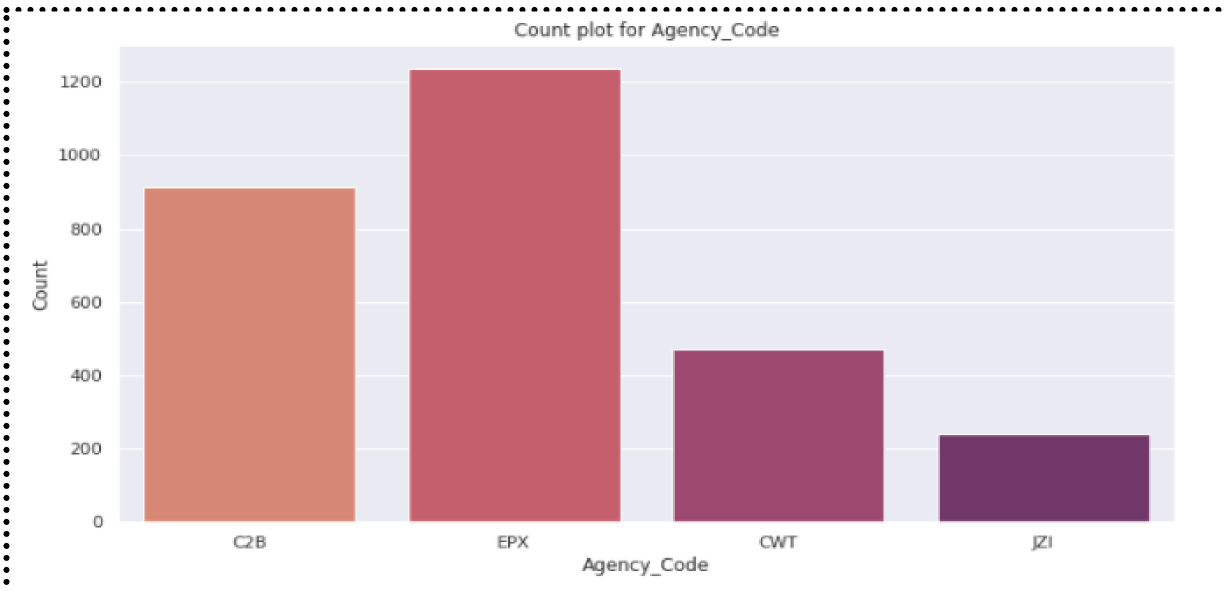
  Name: Destination, dtype: int64

❖ ***The data is okay to proceed. Target Data: Spread is 30 & 70%, it no unbalanced.***

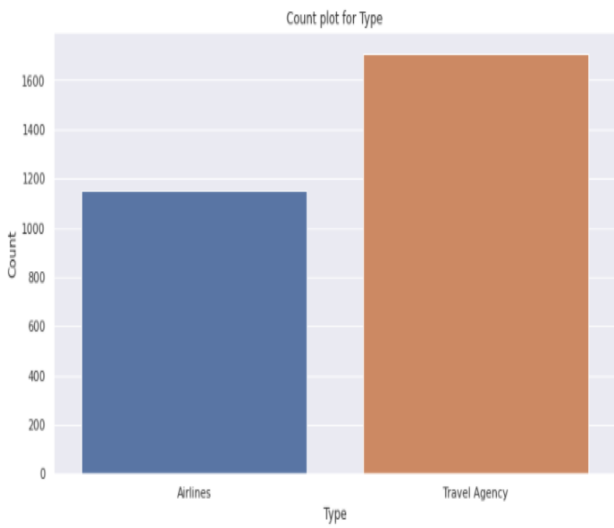
- ***No 0.680531***
- ***Yes 0.319469***

  Name: Claimed, dtype: float64

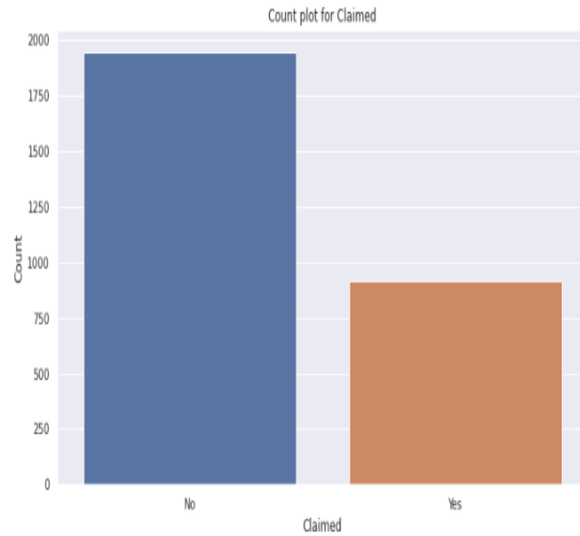
EDA - Univariate Analysis: Categorical variables:



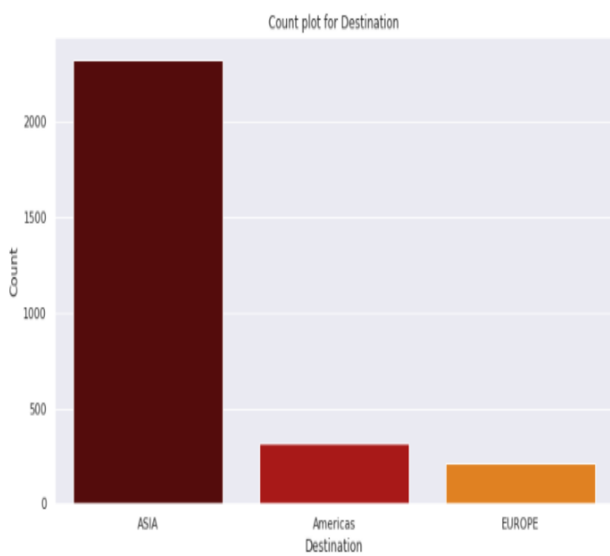
❖ ***From the above countplot it is clear that the data count of EPX > C2B > CWT and the smallest is JZI.***



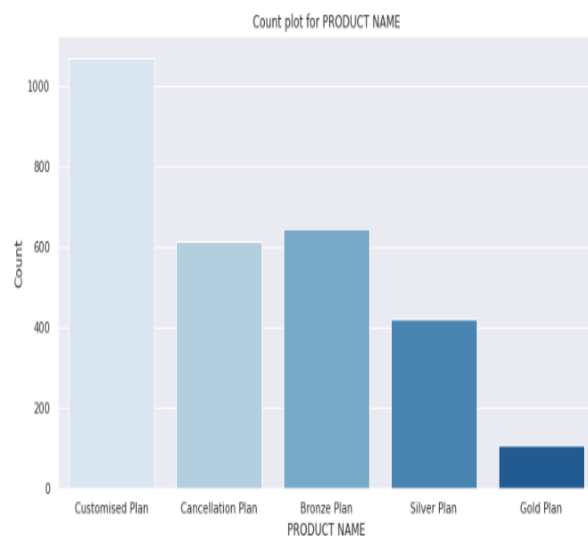
❖ ***We can see based on this data Travel Agency is preferred to 'Airlines'***



❖ ***We see that no of claims made a c***
❖
❖ ***omparatively less to total count.***

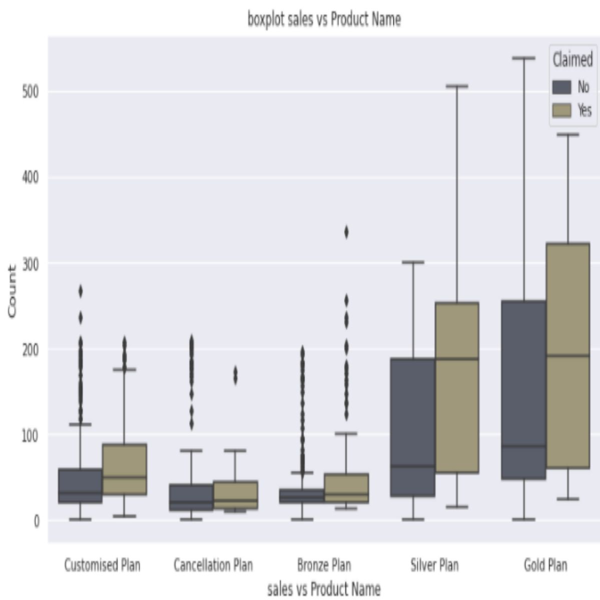


❖ ***ASIA has the highest count > American Destination > Europe***

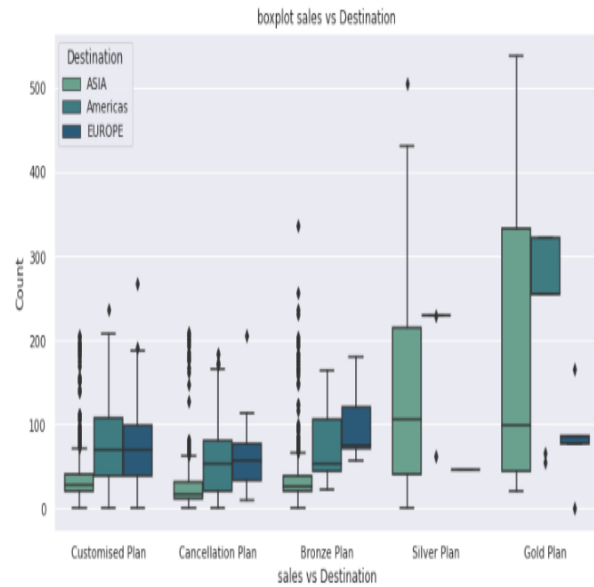


❖ ***It is evident the highest count is for Customised plan and the lowest is for Gold plan.***

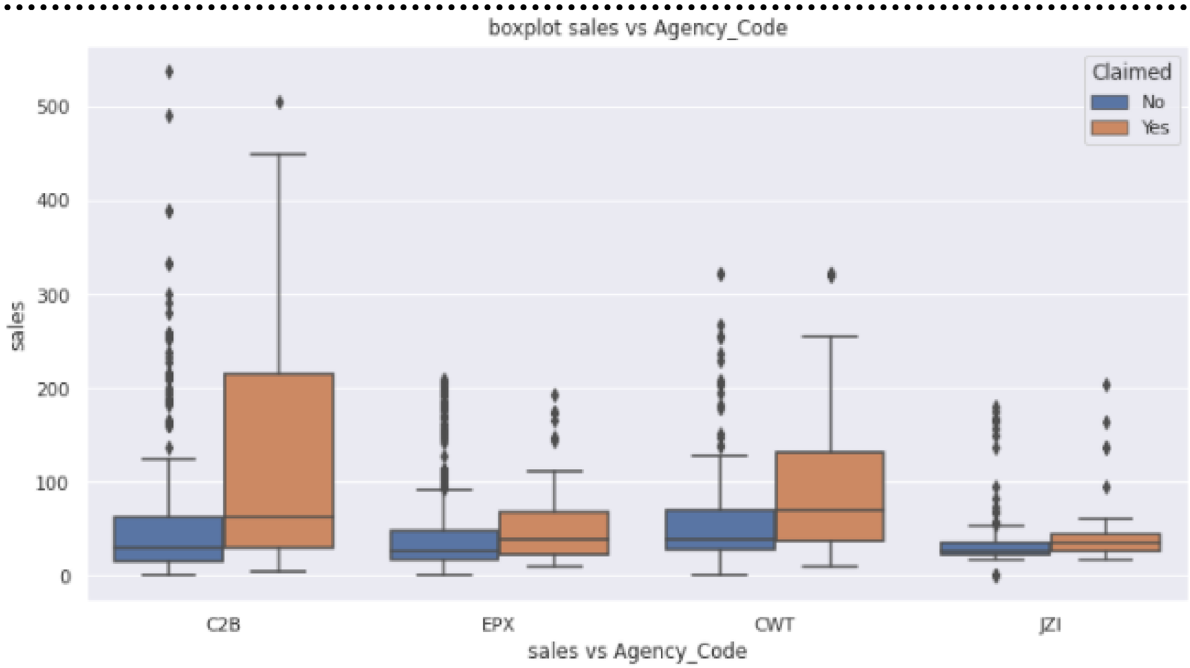
Bivariate Analysis: Categorical variables:



❖ ***Based on the figure we can see that most sales and claims are made for Gold and Silver plans.***

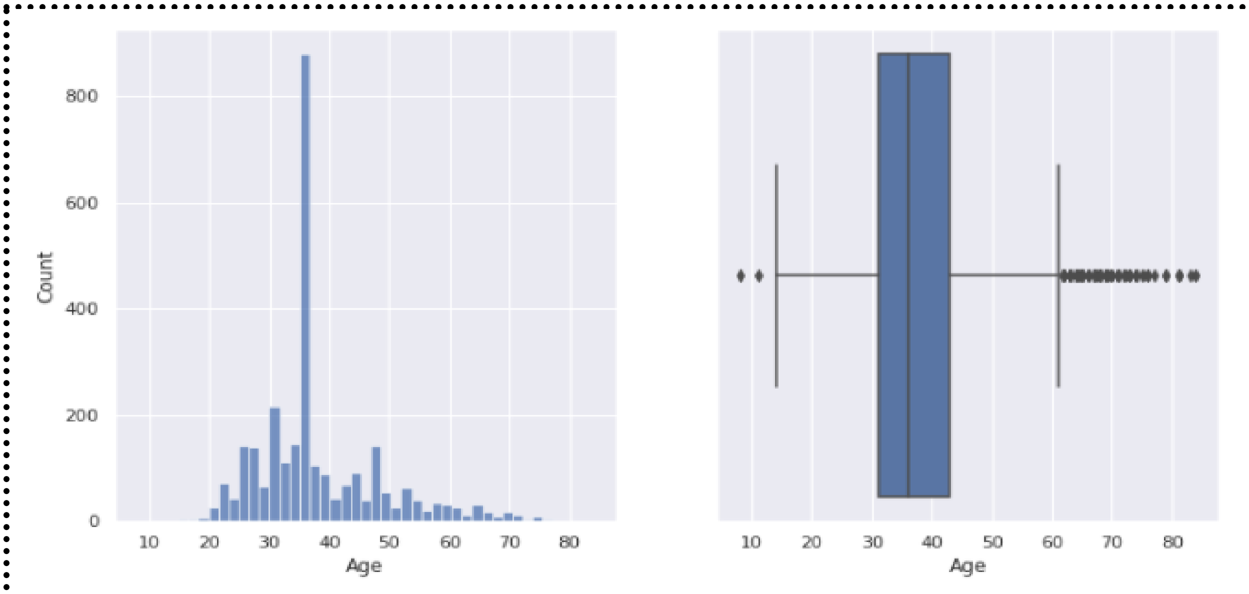


❖ ***We can conclude that Gold and Silver plan sale are highest in ASIA***



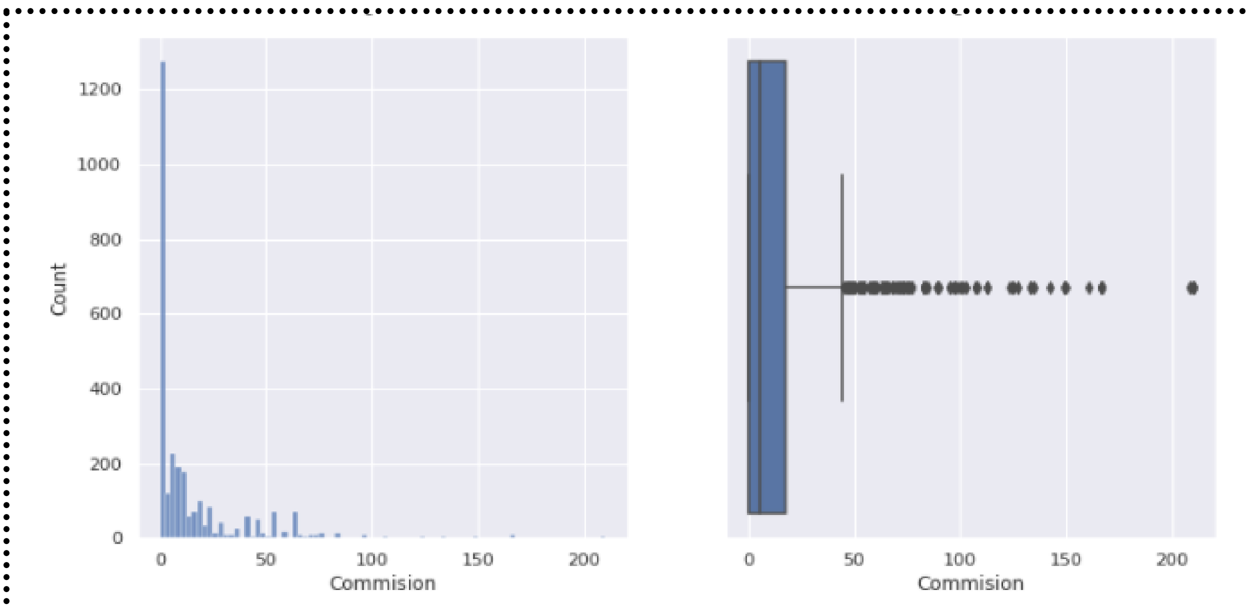
❖ ***The highest number of Claims are recorded at C2B agency for sales.***

Numerical Variable: Univariate Analysis
AGE:



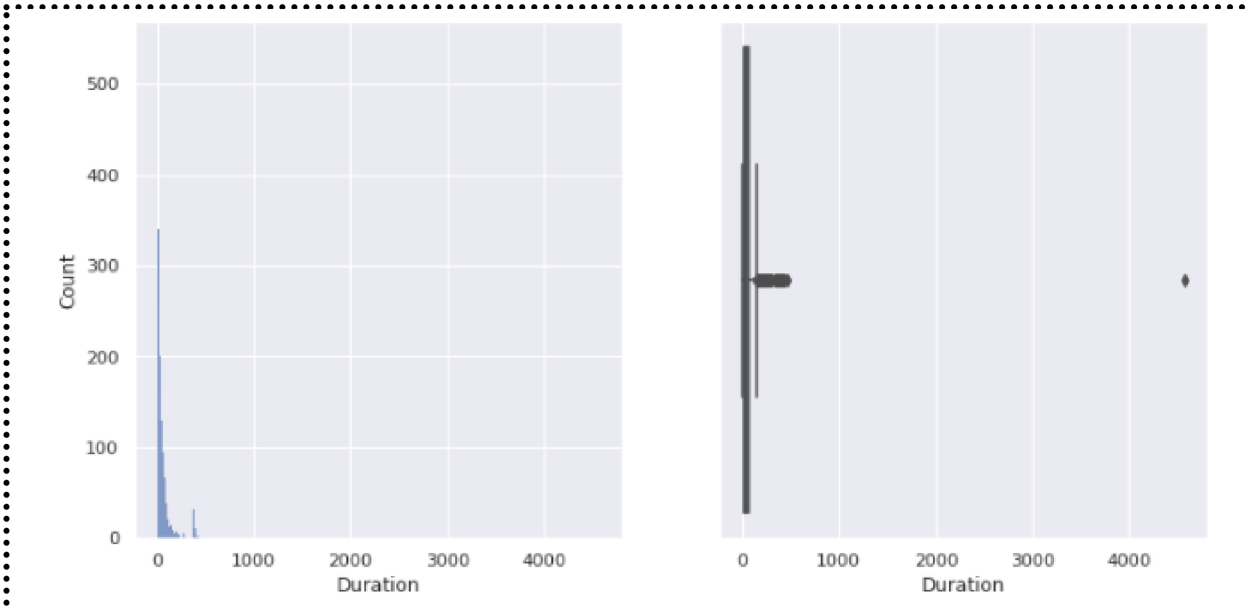
- ***The Boxplot tells us there are outliers for Age distribution.***
- ***The distribution can be said to be normally distributed. Skewness (Age) is 1.103. The distribution ranges between 20 to 75***

COMMISION:



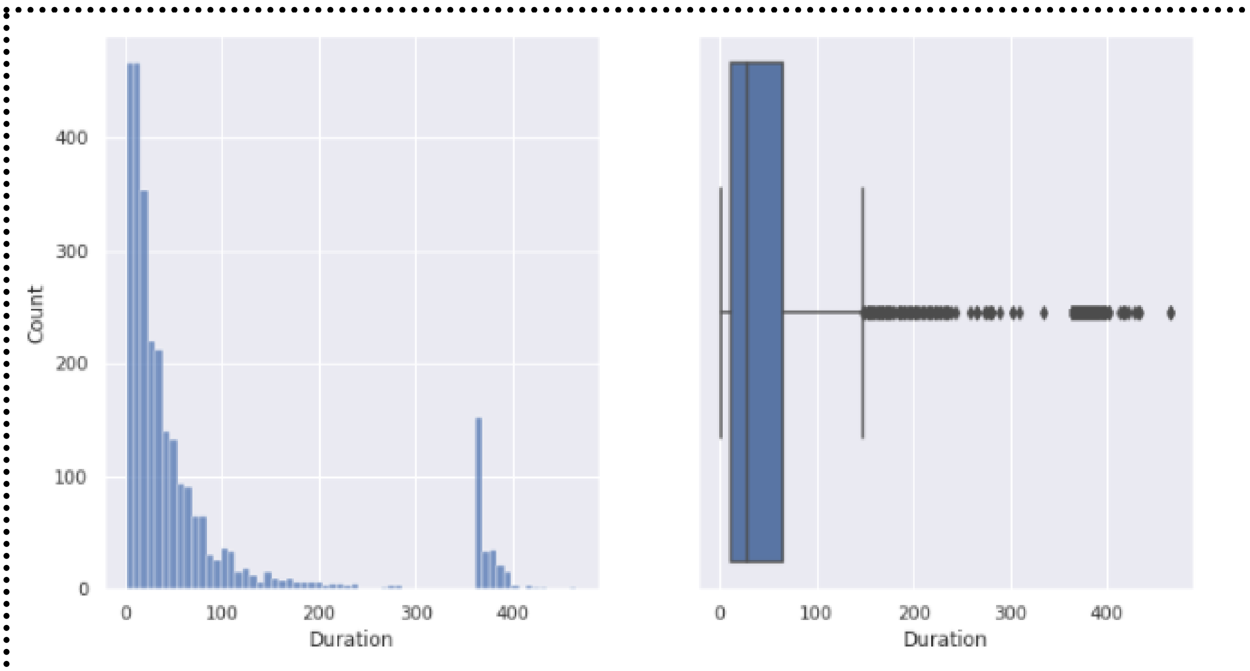
- ***Boxplot tells us there are outliers for Commision distribution.***
- ***The distribution can be said to be normally distributed. Skewness (Commision) is 3.103. The distribution ranges between 0 to 50***

DURATION:



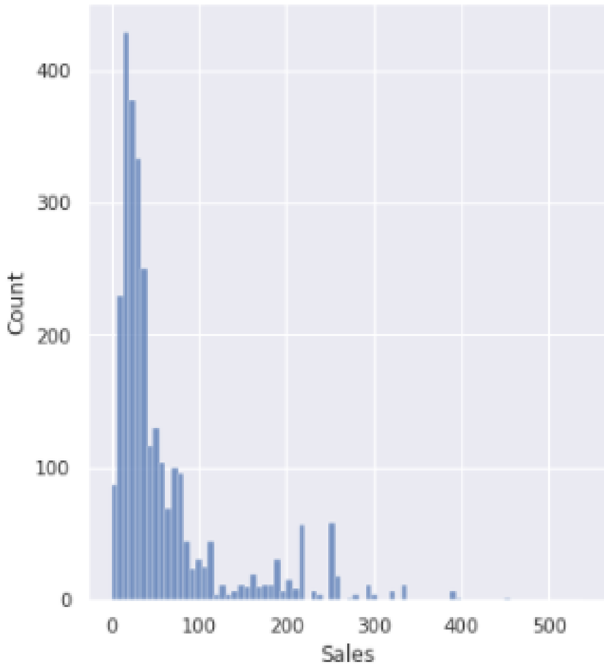
- ❖ ***The Boxplot tells us there are outliers for Duration distribution.***
- ❖ ***The distribution can be said to be normally distributed. Skewness (Duration) is 13.779.***
- ❖ ***As the min Duration shows -1 and time can't be negation or 0, we will treat it.***

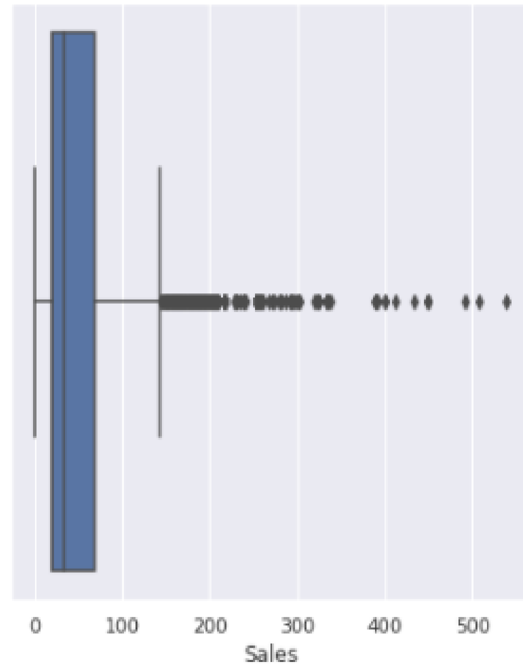
DURATION:



- ❖ ***The new Duration boxplot after removing -1, 0 and the single entry of*** `4580`.
- ❖ ***Skewness (Duration) is reduced to 2.19.***

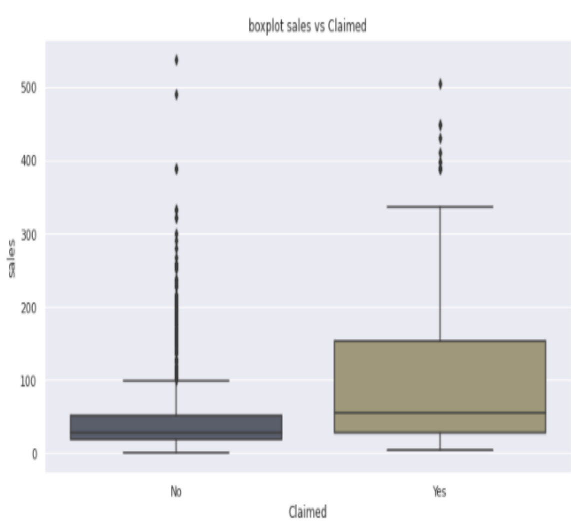
SALES



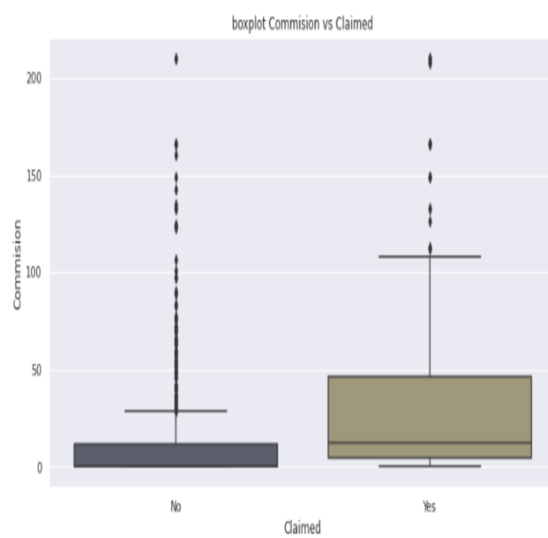


- ❖ ***The Boxplot tells us there are outliers for Sales distribution.***
- ❖ ***The distribution can be said to be normally distributed. Skewness (Sales) is 2.343. The distribution ranges between 0 to 500***

Bivariate Analysis:

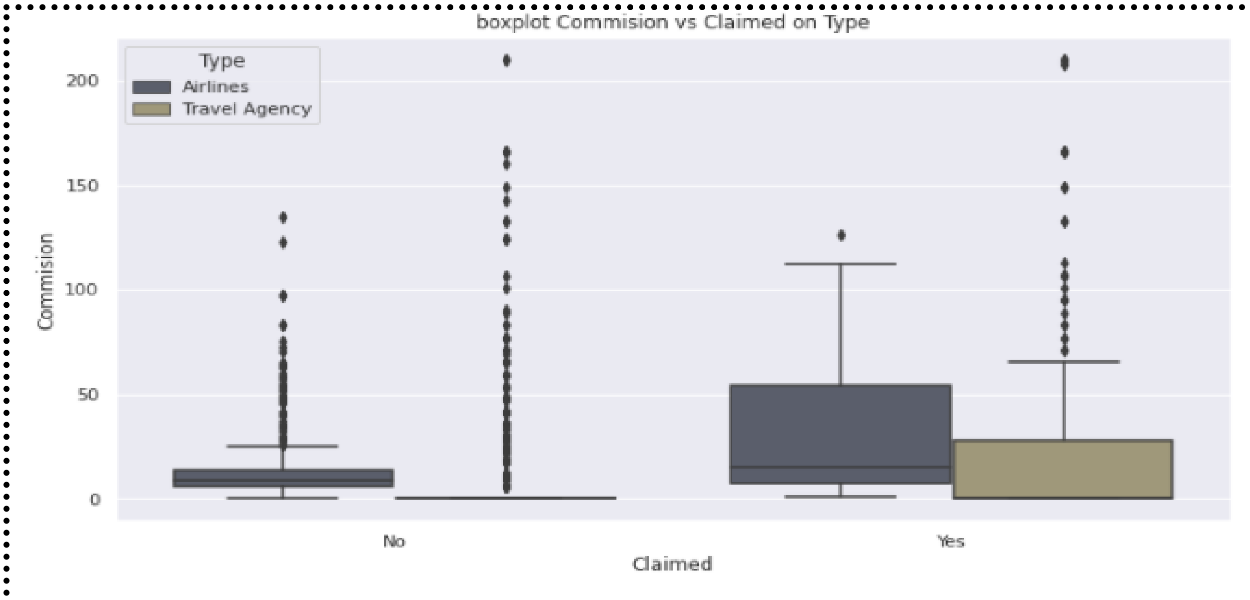


- ❖ **We can see that for sales the claims recorded are higher.**

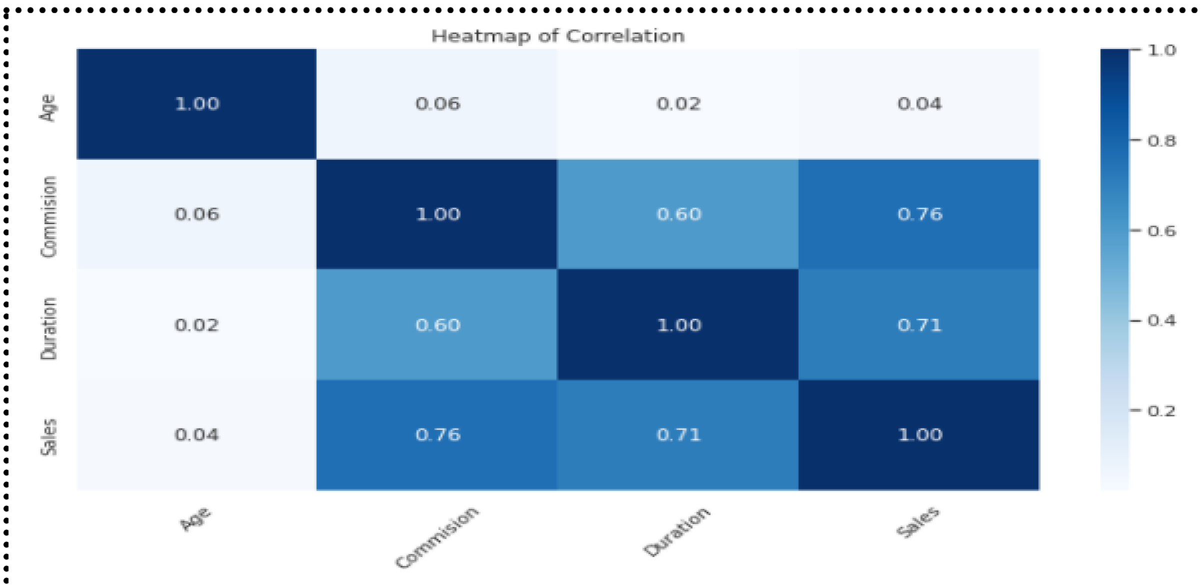


- ❖ **We can see that for *Commision* the claims are higher.**

❖ *We see that Airlines has more claims than Travel Agency.*



Multivariate Analysis - Correlation Heatmap:



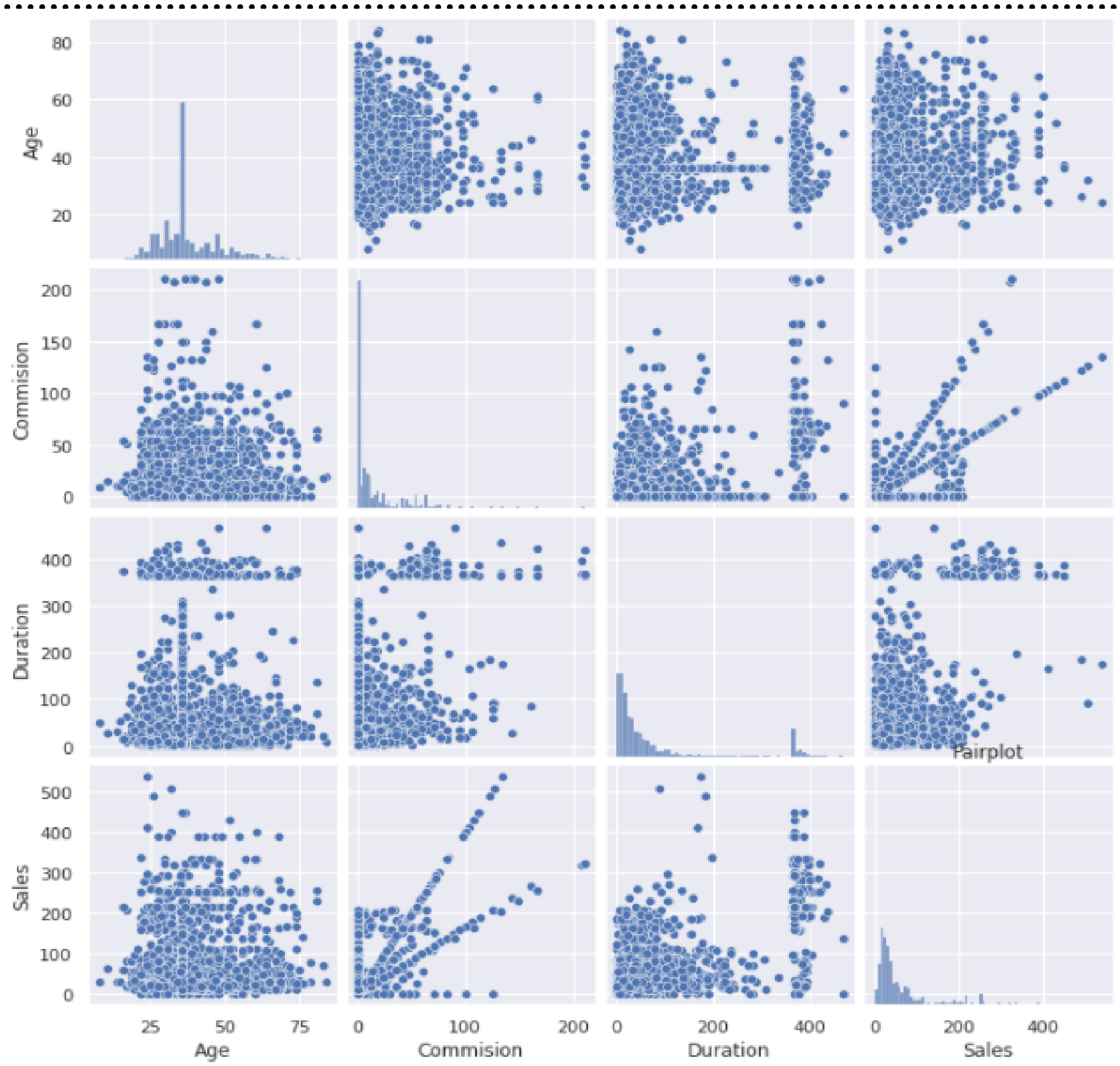
➢ ***The relation between pairs of numeric variables is given by the heatmap.***

➢ ***The correlation between the following variables are highly positive: (variable are directly proportional)***

- ***Sales and Commision (0.76)***
- ***Duration and Commision (0.60)***
- ***Sales and Duration(0.71)***

- ➢ ***We can conclude that these three variables are related to each other. Whereas we have Age that has a weak relation with all variables.***
- ➢ ***Sales and Commission have the highest value of correlation, as sales increase the commision also increases.***
- ➢ ***As for Age we can understand that is not a major factor affecting or interacting with the other variables so far.***

Pair Plot:



❖ ***In this plot we can see direct relations between Duration, Commision and Sales.***

2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.

- ***Splitting data into Train and Test at the default radio of 30:70% with random state as 123.***
- ***We have object data type, which we will convert to categorical codes:***

Data info after conversion:

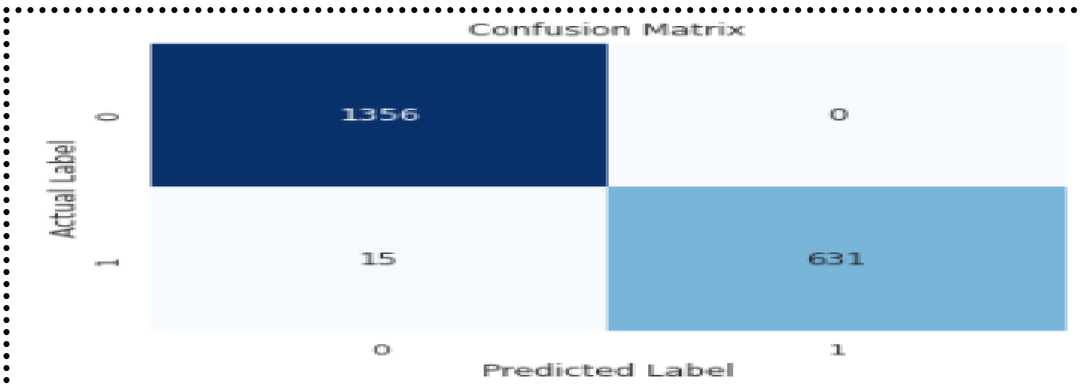
```
#   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Age           2861 non-null   int64
 1   Agency_Code   2861 non-null   int8
 2   Type          2861 non-null   int8
 3   Claimed       2861 non-null   int8
 4   Commision     2861 non-null   float64
 5   Channel       2861 non-null   int8
 6   Duration      2861 non-null   int64
 7   Sales         2861 non-null   float64
 8   Product Name  2861 non-null   int8
 9   Destination   2861 non-null   int8
dtypes: float64(2), int64(2), int8(6)
memory usage: 208.5 KB
```

- ***Target variable class distribution.***
  - ***Yes: 31.95%***
  - ***No: 68.05%***
- ***After splitting the data into test and train, distribution is:***
  - `X_train (2002, 9)`
  - `X_test (859, 9)`
  - `train_labels (2002,)`
  - `test_labels (859,)`
- ***Variable Importance:***
  - `Duration        0.273646`

- ➢ *Sales* 0.210878
- ➢ *Agency_Code* 0.179736
- ➢ *Age* 0.167237
- ➢ *Commision* 0.091713
- ➢ *Product Name* 0.049420
- ➢ *Destination* 0.023208
- ➢ *Channel* 0.004161
- ➢ *Type* 0.000000

Decision Tree Classifier

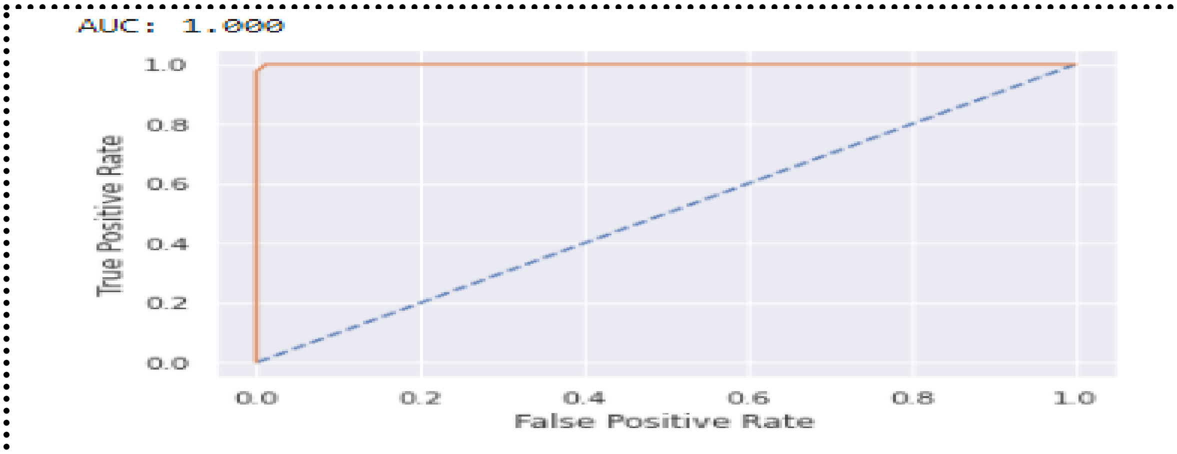
- ❖ ***Creating a model with default feature value with criterion as Gini (It is calculated by subtracting the sum of squared probabilities of each class from one.)***
- ❖ ***The parameters are: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, random_state=123}***
- ❖ ***CART: Confusion Matrix Train data:***



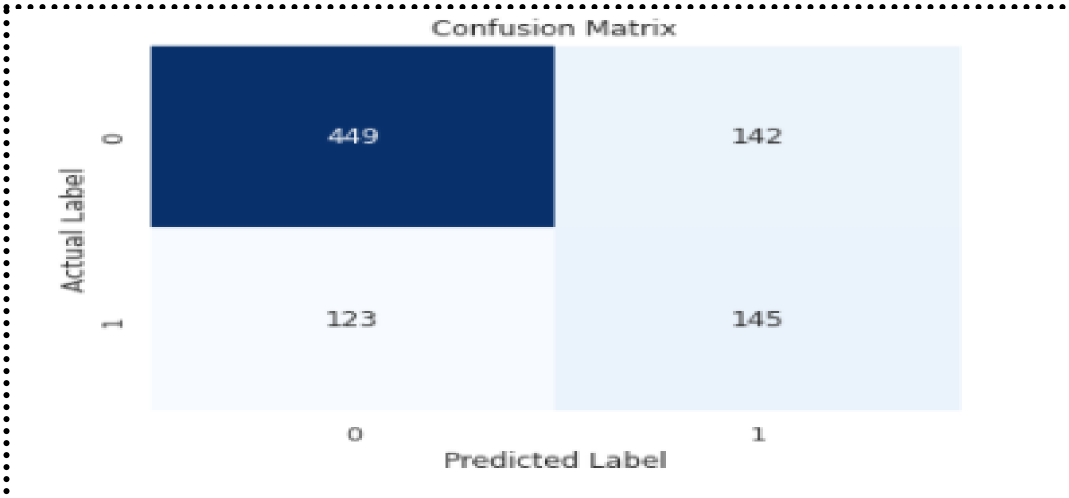
- ❖ ***Classification_report:***

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1356
1	1.00	0.98	0.99	646
accuracy			0.99	2002
macro avg	0.99	0.99	0.99	2002
weighted avg	0.99	0.99	0.99	2002

- ***AUC and ROC curve for the training data:***



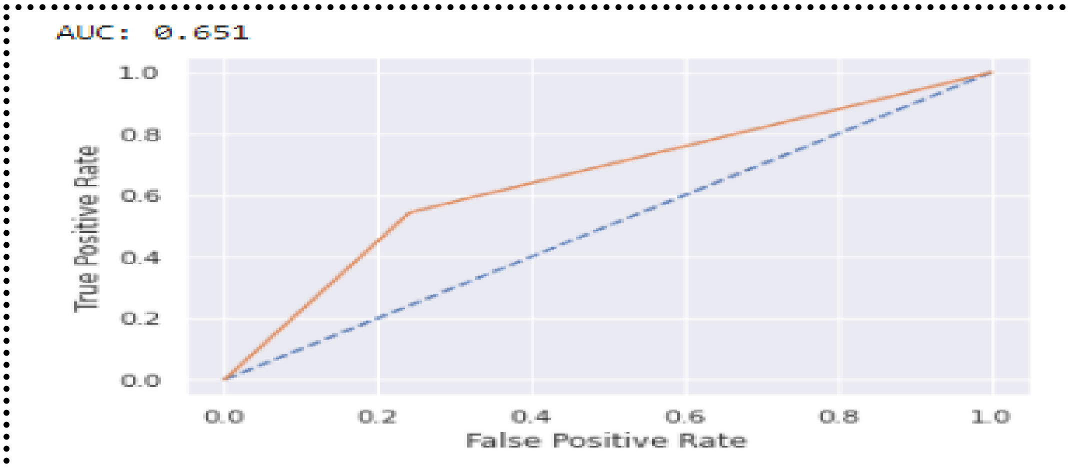
- ***Training data Accuracy Score:***
  - ***0.9925074925074925***
- ***Confusion Matrix Test data:***



- ***Classification_report:***

	precision	recall	f1-score	support
0	0.78	0.76	0.77	591
1	0.51	0.54	0.52	268
accuracy			0.69	859
macro avg	0.65	0.65	0.65	859
weighted avg	0.70	0.69	0.69	859

❖ *AUC and ROC curve for the testing data:*



❖ *Testing data Accuracy Score:*

- ➢ ***0.6915017462165308***

❖ ***From the data above we can say that the model with default values is over fitted:***

- ➢ ***Accuracy from the Training data is: 1.0***
- ➢ ***Accuracy from the Test data is : 0.65***

❖ ***As we see, the accuracy of training and test results are very different. Training data being 1, concludes that the model is*** **Overfitted.**

❖ ***Next we will perform a Grid search that will help us in Tuning the model and address overfitting.***

❖ ***The Tuning Parameters used are:***

- ➢ ***Max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure.***
- ➢ ***Min_samples_leaf: The minimum number of samples required to be at a leaf node.***
- ➢ ***Min_samples_split: The minimum number of samples required to split an internal node.***

❖ ***The best parameters for Grid search are: DecisionTreeClassifier(max_depth=5, min_samples_leaf=20, min_samples_split=200, random_state=123)***

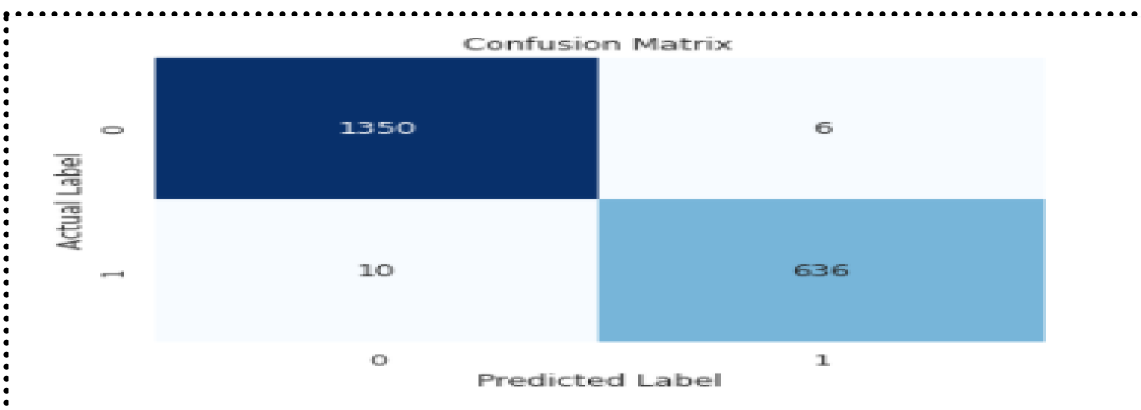
❖ ***The most important variables are:***

- ➢ ***Agency_Code 0.587848***
- ➢ ***Sales 0.219399***

- ➢ `Product Name          0.109473`
- ➢ `Age                   0.032647`
- ➢ `Duration              0.021835`
- ➢ `Commision             0.020183`
- ➢ `Destination           0.008615`
- ➢ `Type,Channel          0.000000`

Random Forest:

- ❖ *We start with Default parameters to implement random forest to our split data of 30:70*
- ❖ *The parameters are: {'criterion': 'gini', 'n_estimators'=100, 'max_depth': None, 'Min_samples_leaf': 1, 'min_samples_split': 2, max_features="auto", random_state=123}*
- ❖ *Variable Importance:*
  - ➢ `Duration        0.262404`
  - ➢ `Sales           0.203511`
  - ➢ `Age             0.172830`
  - ➢ `Commision       0.121100`
  - ➢ `Agency_Code     0.100321`
  - ➢ `Product Name    0.095428`
  - ➢ `Destination     0.022865`
  - ➢ `Type            0.014934`
  - ➢ `Channel         0.006606`
- ❖ *Random Forest: Confusion Matrix Train data:*



❖ *Classification_report:*

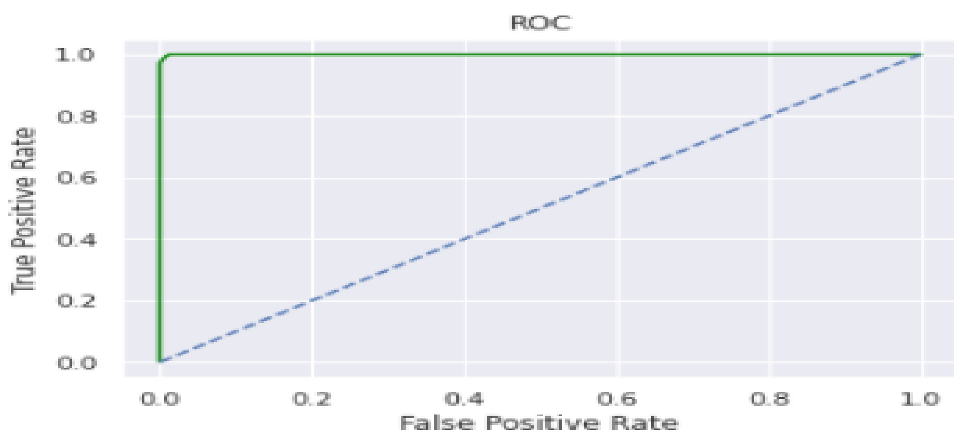
	precision	recall	f1-score	support
0	0.99	1.00	0.99	1356
1	0.99	0.98	0.99	646
accuracy			0.99	2002
macro avg	0.99	0.99	0.99	2002
weighted avg	0.99	0.99	0.99	2002

❖ *Training data Accuracy Score:*

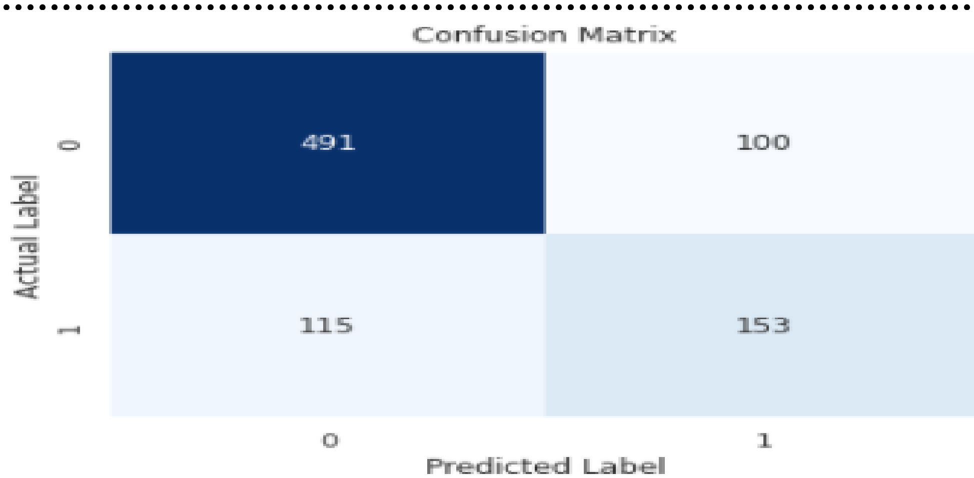
➢ *0.9920079920079921*

❖ *AUC and ROC curve for the training data:*

Area under Curve is 0.9998561604427518



❖ *Confusion Matrix Test data:*



❖ *Classification_report:*

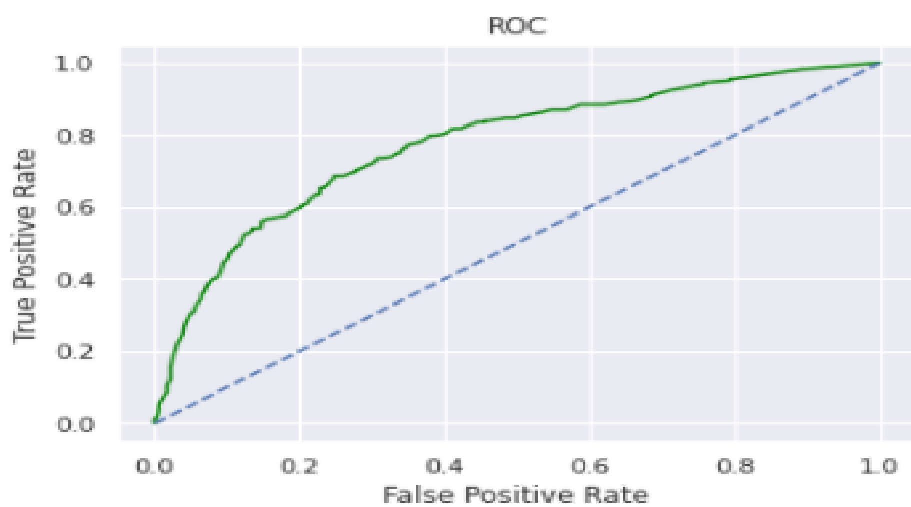
	precision	recall	f1-score	support
0	0.81	0.83	0.82	591
1	0.60	0.57	0.59	268
accuracy			0.75	859
macro avg	0.71	0.70	0.70	859
weighted avg	0.75	0.75	0.75	859

❖ *Testing data Accuracy Score:*

➢ *0.7497089639115251*

❖ *AUC and ROC curve for the testing data:*

Area under Curve is 0.7747935449655277



❖ ***From the data above we can say that the model with default values is over fitted:***

➢ ***Accuracy from the Training data is: 0.99***

➢ ***Accuracy from the Test data is : 0.77***

❖ ***As we see, the accuracy of training and test results are very different. Training data being 0.99 almost 1, concludes that the model is* Overfitted.**

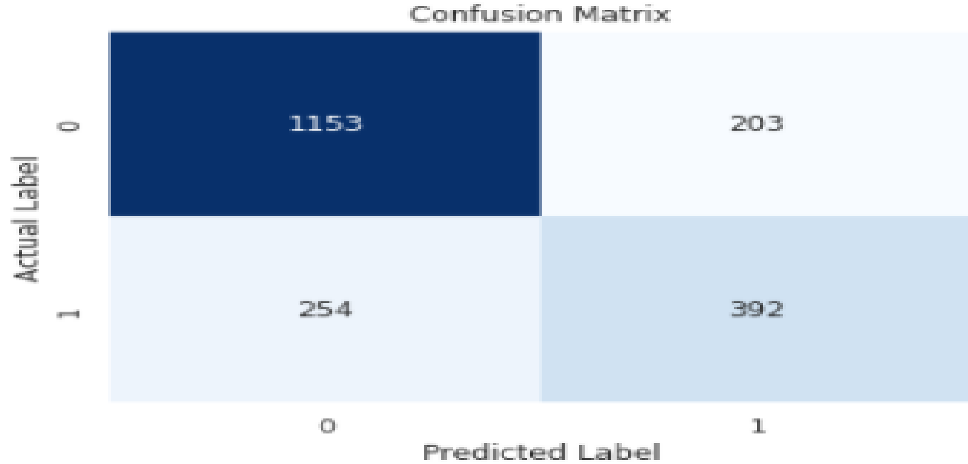
❖ ***Next we will perform a Grid search that will help us in Tuning the model and address overfitting.***

- ***The Tuning Parameters used are:***
  - ***Max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure.***
  - ***Min_samples_leaf: The minimum number of samples required to be at a leaf node.***
  - ***Min_samples_split: The minimum number of samples required to split an internal node.***
  - ***n_estimators : The number of trees in the forest.***
  - ***max_features : default="auto" The number of features to consider when looking for the best split.***
- ***The best parameters for Grid search are: RandomForestClassifier(max_depth=10, max_features=5, min_samples_leaf=50, min_samples_split=20, n_estimators=50, random_state=123)***
- ***The most important variables are:***
  - ***Agency_Code 0.416913***
  - ***Product Name 0.256784***
  - ***Sales 0.169993***
  - ***Commision 0.062347***
  - ***Duration 0.042298***
  - ***Age 0.024639***
  - ***Type 0.022198***
  - ***Destination 0.004829***
  - ***Channel 0.000000***

Artificial Neural Network:

- ***We start with Default parameters to implement artificial neural network to our split data of 30:70***
- ***The parameters are: {'hidden_layer_sizes': 100, 'max_iter': 200, 'solver': 'adam', 'Tol': 0.1, random_state=123}***

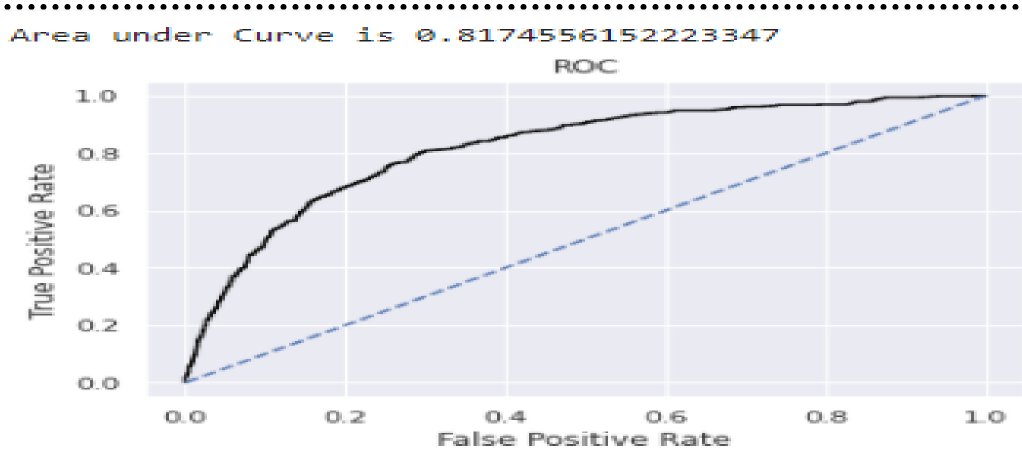
- ❖ **Artificial Neural Network: Confusion Matrix Train data:**



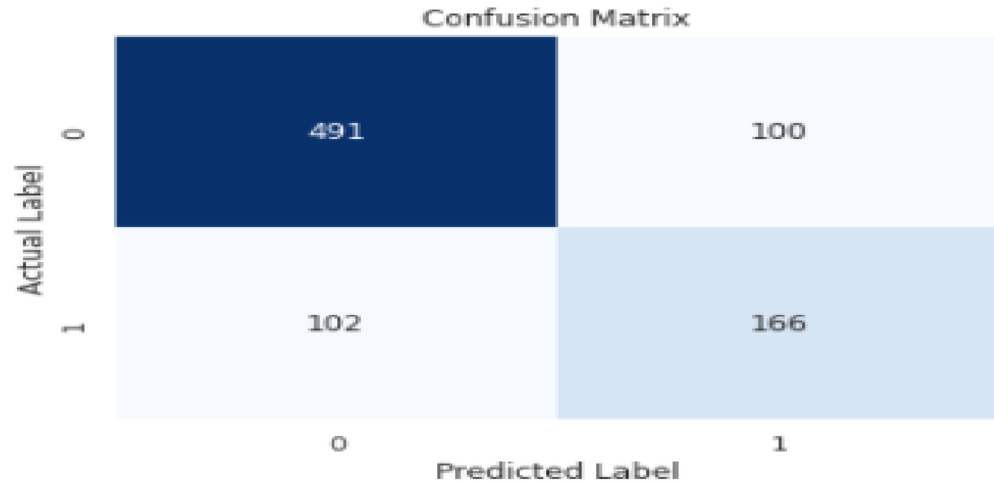
- ❖ *Training data Accuracy Score:*
  - ➢ *0.7717282717282717*
- ❖ *Classification_report:*

	precision	recall	f1-score	support
0	0.82	0.85	0.83	1356
1	0.66	0.61	0.63	646
accuracy			0.77	2002
macro avg	0.74	0.73	0.73	2002
weighted avg	0.77	0.77	0.77	2002

- ❖ *AUC and ROC curve for the training data:*



❖ ***Confusion Matrix Test data:***



❖ ***Testing data Accuracy Score:***

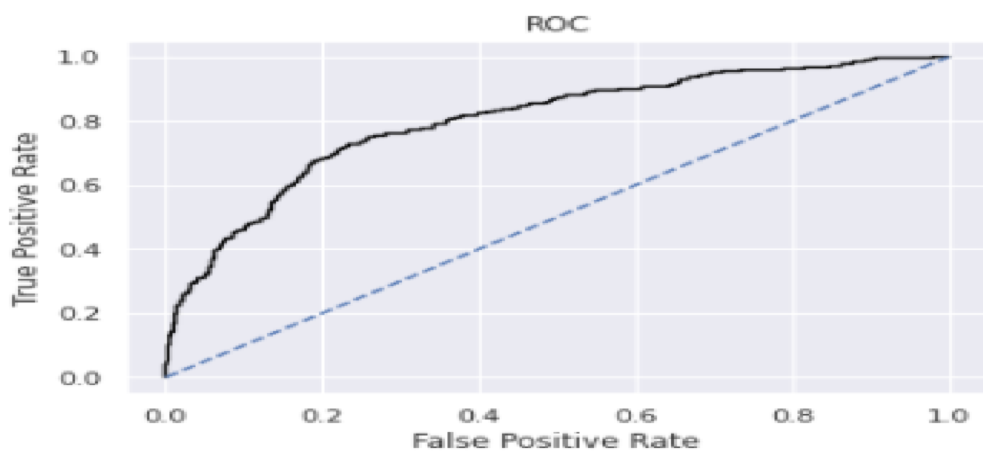
➢ ***0.7648428405122235***

❖ ***Classification_report:***

```
              precision    recall  f1-score   support

           0       0.83      0.83      0.83       591
           1       0.62      0.62      0.62       268

    accuracy                           0.76       859
   macro avg       0.73      0.73      0.73       859
weighted avg       0.76      0.76      0.76       859
```

❖ ***AUC and ROC curve for the testing data:***

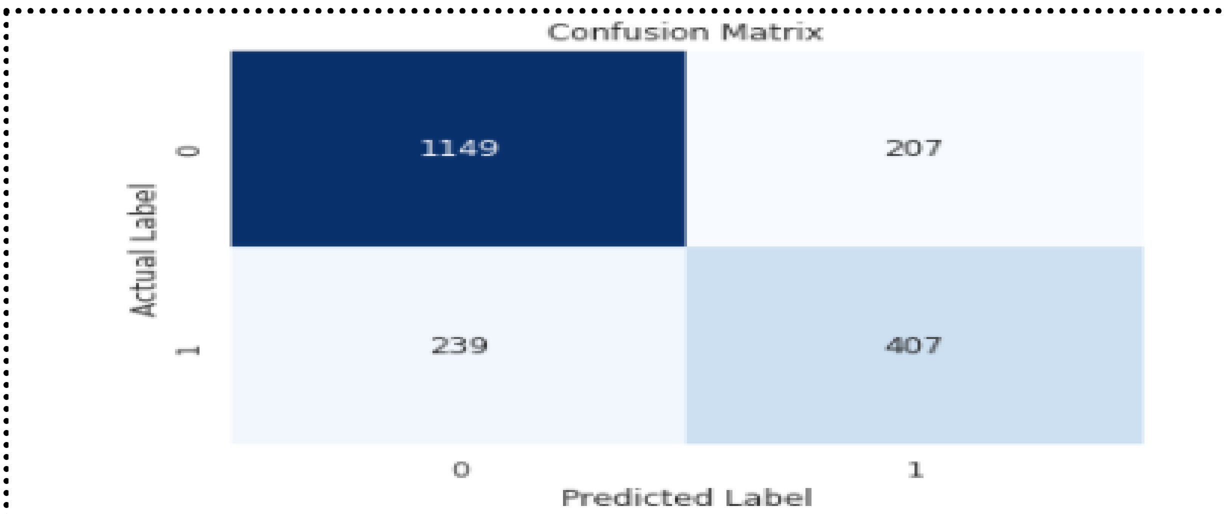
Area under Curve is 0.7996502260272242



- ***From the data above we can say that the model with default values is over fitted:***
  - ***Accuracy from the Training data is: 0.77***
  - ***Accuracy from the Test data is : 0.76***
- ***As we see, the accuracy of training and test results are very close.***
- ***Next we will perform a Grid search that will help us in Tuning the model and address overfitting.***
- ***The Tuning Parameters used are:***
  - ***Hidden_layer_sizes: The ith element represents the number of neurons in the ith hidden layer.***
  - ***Solver: The solver for weight optimization.***
  - ***tol: Tolerance for the optimization.***
  - ***Max_iter: Maximum number of iterations.***
- ***The best parameters for Grid search are: MLPClassifier(hidden_layer_sizes=200, max_iter=2500, random_state=123, tol=0.001)***

2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

- **CART: Confusion Matrix Train data:**



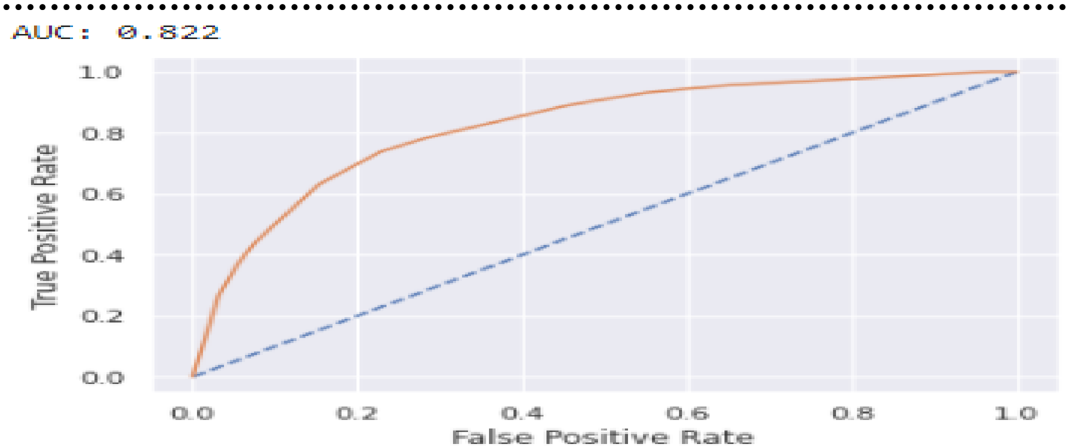
❖ **Classification_report:**

```
              precision    recall  f1-score   support

           0       0.83      0.85      0.84      1356
           1       0.66      0.63      0.65       646

    accuracy                           0.78      2002
   macro avg       0.75      0.74      0.74      2002
weighted avg       0.77      0.78      0.78      2002
```

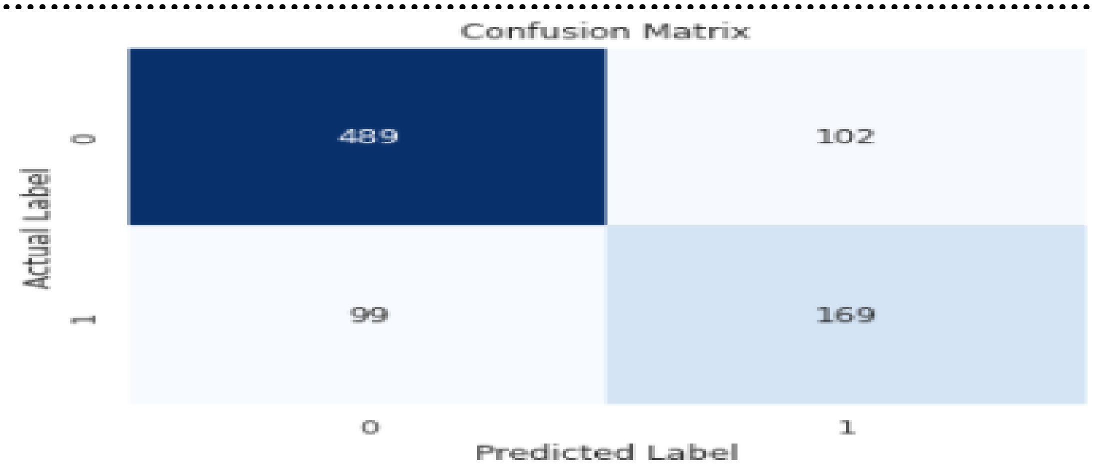
❖ **Training data Accuracy Score:**

➢ ***0.7772227772227772***

❖ **AUC and ROC curve for the training data:**



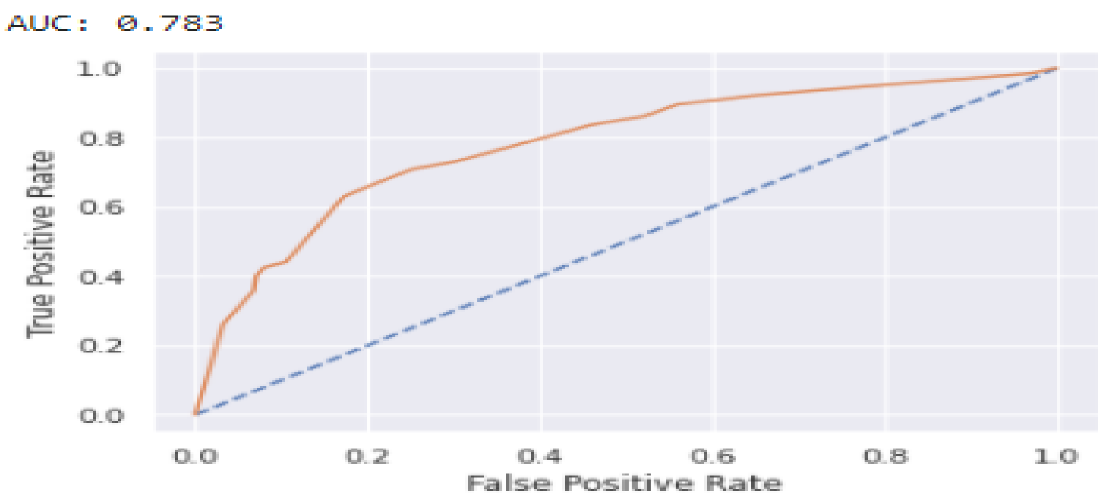
❖ **Confusion Matrix Test data:**



- **Classification_report:**

	precision	recall	f1-score	support
0	0.83	0.83	0.83	591
1	0.62	0.63	0.63	268
accuracy			0.77	859
macro avg	0.73	0.73	0.73	859
weighted avg	0.77	0.77	0.77	859

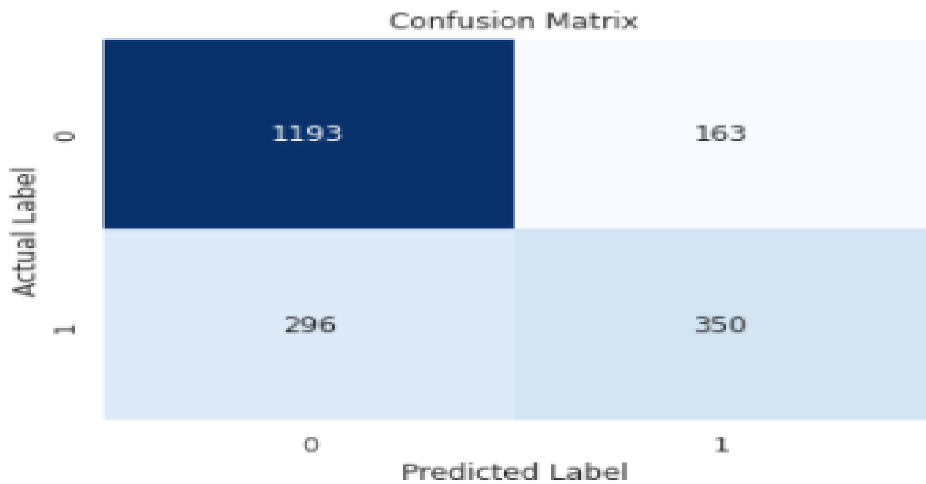
- ***Testing data Accuracy Score:***
    - ***0.7660069848661234***
- ***AUC and ROC curve for the testing data:***



- ***The best parameters are: DecisionTreeClassifier(max_depth=5, min_samples_leaf=20, min_samples_split=200, random_state=123)***
- ***From the data above we can say we were able to build a good model:***
    - ***Accuracy from the Training data is: 0.77***
    - ***Accuracy from the Test data is : 0.76***
    - ***Which is close, we had addressed the over-fitting issue scenario here.***
    - ***There is still room for improvement.***
    - ***Though the model can be used.***

- ***The most important variables are:***
    - *Agency_Code 0.587848*
    - *Sales 0.219399*
    - *Product Name 0.109473*
    - *Age 0.032647*
    - *Duration 0.021835*
    - *Commision 0.020183*
- ***Training:*** *cart_train_precision 0.66, cart_train_recall 0.63, cart_train_f1 0.65*
- ***Testing:*** *cart_test_precision 0.62, cart_test_recall 0.63, cart_test_f1 0.63*

- **Random Forest: Confusion Matrix Train data:**

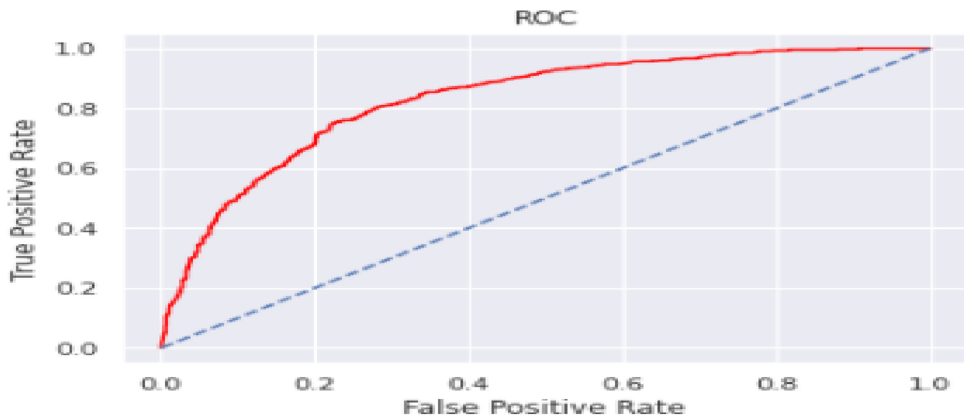


- **Classification_report:**

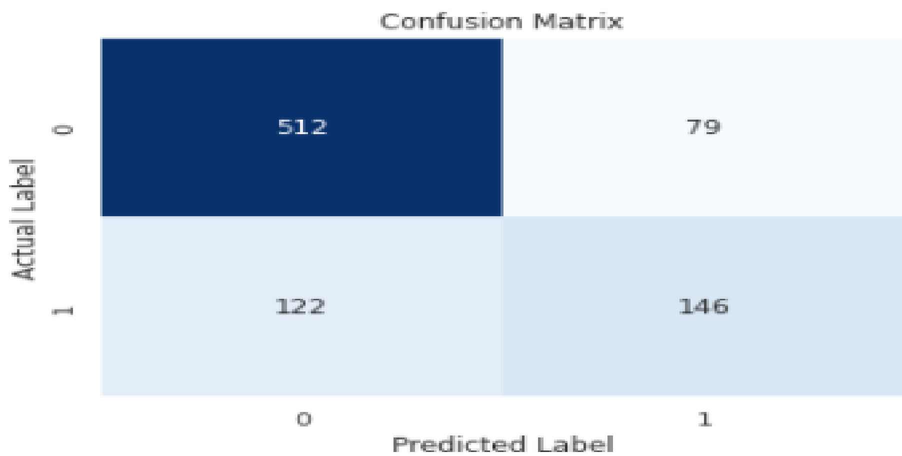
	precision	recall	f1-score	support
0	0.80	0.88	0.84	1356
1	0.68	0.54	0.60	646
accuracy			0.77	2002
macro avg	0.74	0.71	0.72	2002
weighted avg	0.76	0.77	0.76	2002

- ***Training data Accuracy Score:***
    - ***0.7707292707292708***
- **AUC and ROC curve for the training data*:***

```
Area under Curve is 0.8298024146780276
```



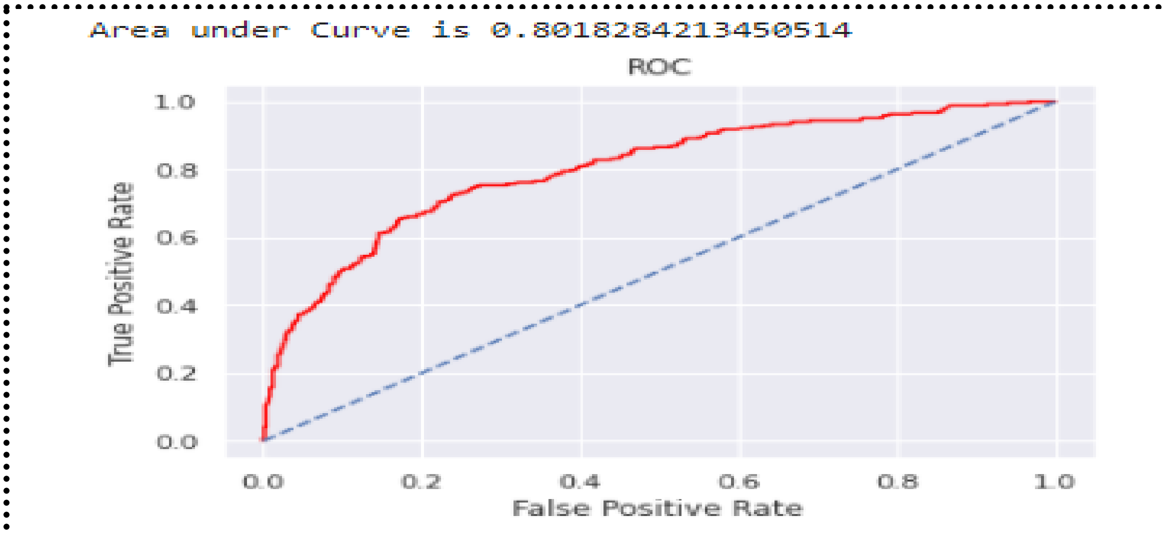
- **Confusion Matrix Test data*:***



- ***Classification_report:***

	precision	recall	f1-score	support
0	0.81	0.87	0.84	591
1	0.65	0.54	0.59	268
accuracy			0.77	859
macro avg	0.73	0.71	0.71	859
weighted avg	0.76	0.77	0.76	859

❖ *AUC and ROC curve for the testing data:*



❖ *Testing data Accuracy Score:*

➢ *0.7660069848661234*

Inference -

❖ **There was an overfitting issue, after Hyperparameter tuning. We can further work on it to improve.**

❖ *The best parameters are:* **RandomForestClassifier(max_depth=10, max_features=5, min_samples_leaf=50, min_samples_split=20, n_estimators=50, random_state=123)**

❖ *From the data above we can say we were able to build a good model:*

➢ *Accuracy from the Training data is: 0.77*

➢ *Area under Curve is 0.8174556152223347*

➢ *Accuracy from the Test data is : 0.76*

➢ *Area under Curve is 0.8018284213450514*

➢ *Which is close, we had addressed the over-fitting issue scenario here.*

➢ *There is still room for improvement.*

➢ *Though the model can be used.*

❖ *The most important variables are:*

➢ `Agency_Code     0.416913`

➢ `Product Name    0.256784`

➢ `Sales           0.169993`

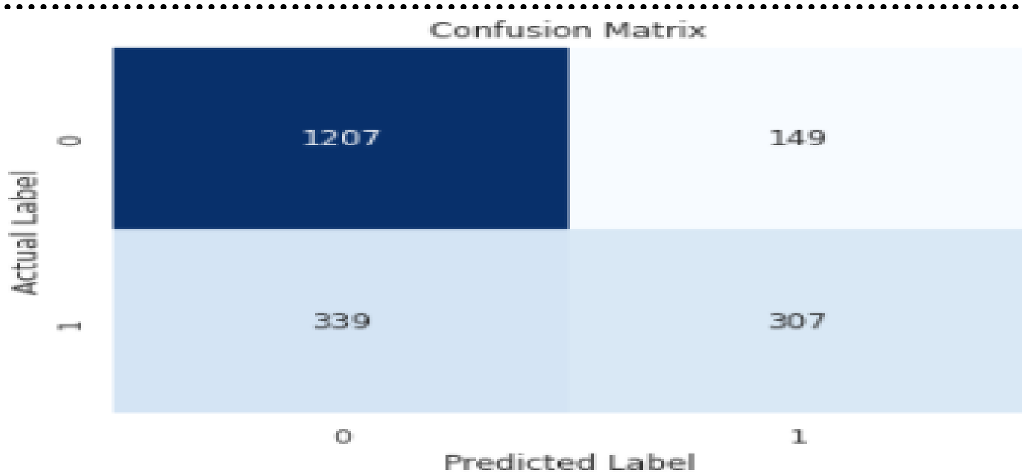
➢ `Commision       0.062347`

- ➢ *`Duration      0.042298`*
- ➢ *`Age           0.024639`*
- ➢ *`Type          0.022198`*
- ➢ *`Destination   0.004829`*
- ➢ *`Channel       0.000000`*

❖ *Training*: *`rf_train_precision 0.68, rf_train_recall 0.54, rf_train_f1 0.6`*

❖ *Testing*: *`rf_test_precision 0.65, rf_test_recall 0.54, rf_test_f1 0.59`*

Neural Network Classifier: Confusion Matrix Train data:

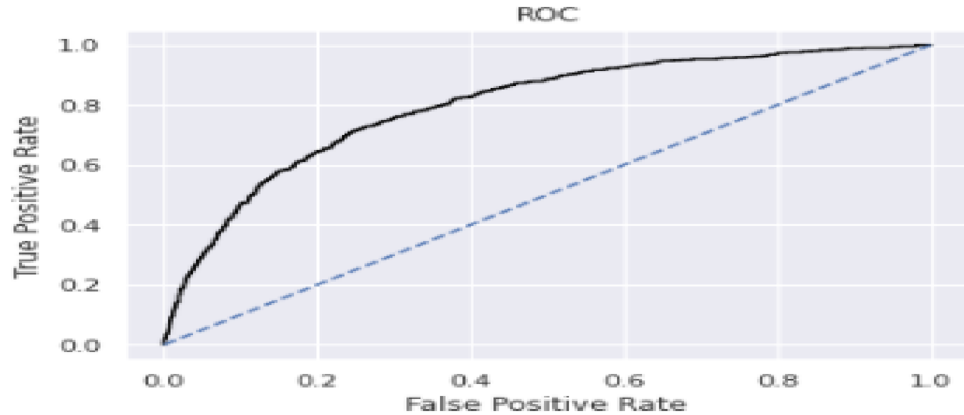


❖ **Classification_report*:***

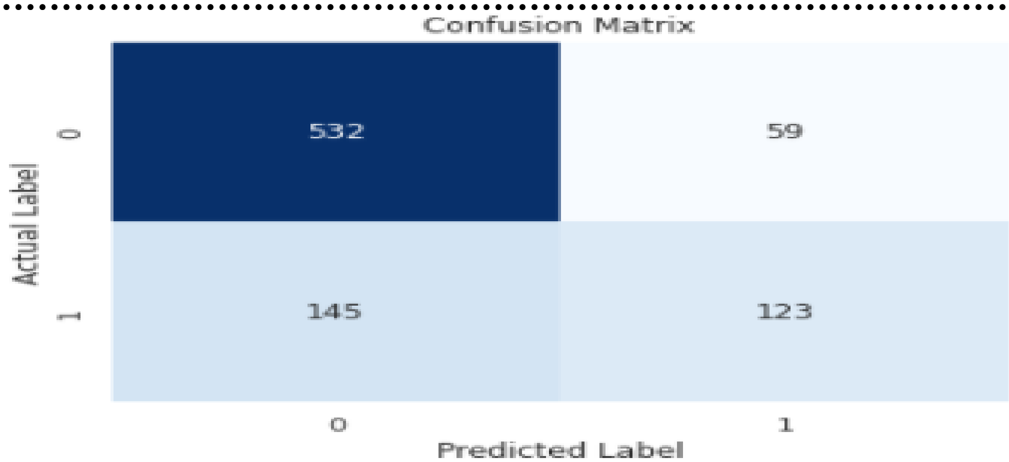
	precision	recall	f1-score	support
0	0.78	0.89	0.83	1356
1	0.67	0.48	0.56	646
accuracy			0.76	2002
macro avg	0.73	0.68	0.69	2002
weighted avg	0.75	0.76	0.74	2002

- **AUC and ROC curve for the training data:**

Area under Curve is 0.7987416321908363



- **Training data Accuracy Score:**
  - ***0.7562437562437563***
- **Confusion Matrix Test data*:***

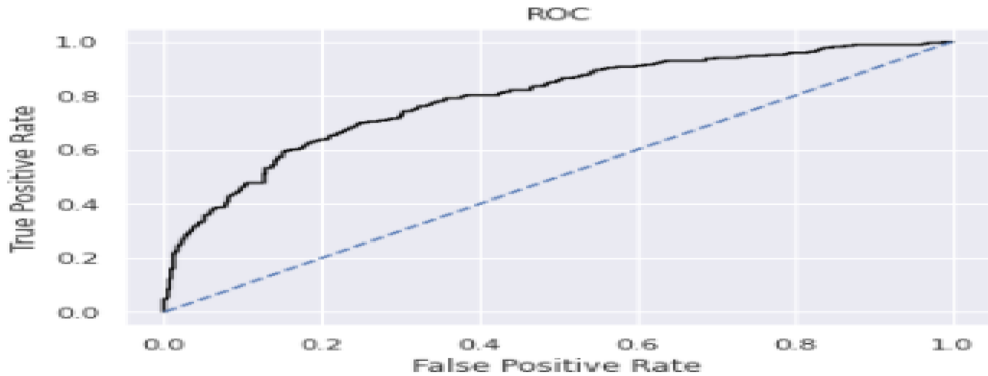


- **Classification_report*:***

	precision	recall	f1-score	support
0	0.79	0.88	0.83	591
1	0.64	0.48	0.55	268
accuracy			0.75	859
macro avg	0.72	0.68	0.69	859
weighted avg	0.74	0.75	0.74	859

❖ **AUC and ROC curve for the testing data:**

```
Area under Curve is 0.7897631133671742
```



❖ ***Testing data Accuracy Score:***

➢ ***0.7543655413271245***

Inference:

❖ **The model behaviour was the same before and after tuning. We can further work on it to improve.**

❖ ***The best parameters are:*** **Neural Network Classifier >> MLPClassifier(hidden_layer_sizes=100, max_iter=2500, random_state=123, tol=0.01)**

❖ ***From the data above we can say we were able to build a good model:***

➢ ***Accuracy from the Training data is: 0.77***

➢ ***Area under Curve is 0.7987416321908363***

➢ ***Accuracy from the Test data is : 0.76***

➢ ***Area under Curve is 0.7897631133671742***

➢ ***The accuracy is very close to each other suggesting this model can be used for further analysis.(Same as we saw above in default parameter)***

➢ ***Of Course there is still room for improvement.***

➢ ***Though the model can be used.***

❖ ***Training:***

➢ ***nn_train_precision 0.67***

➢ ***nn_train_recall 0.48***

➢ ***nn_train_f1 0.56***

- ***Testing:***
  - ***nn_test_precision 0.64***
  - ***nn_test_recall 0.48***
  - ***nn_test_f1 0.55***

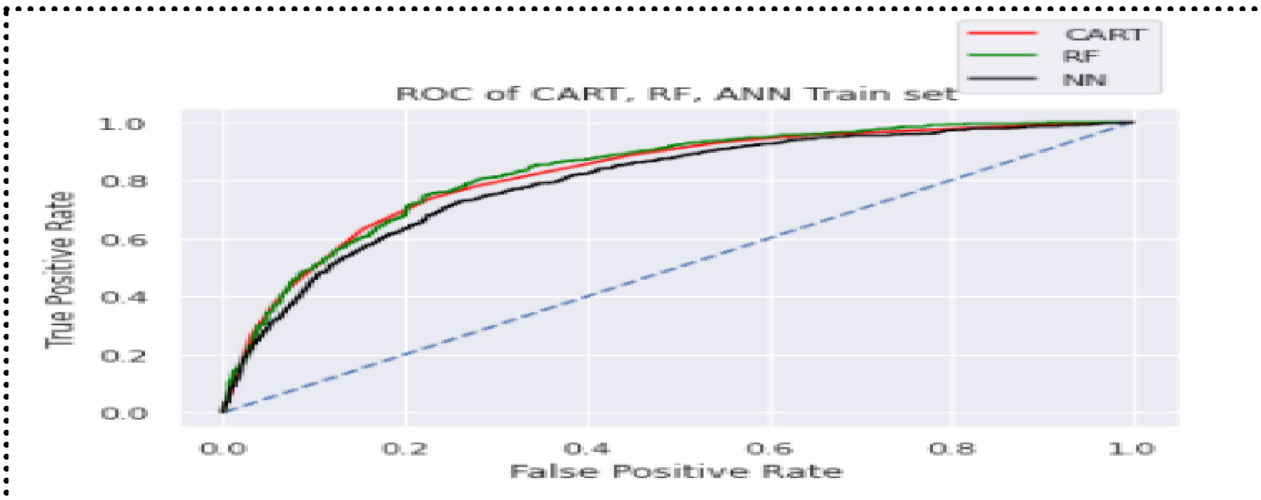
2.4 Final Model: Compare all the models and write an inference which model is best/optimized.

3 Model Tuned table: The overall Tuned Models output table..

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.78	0.77	0.77	0.77	0.76	0.75
AUC	0.82	0.78	0.83	0.8	0.8	0.79
Recall	0.63	0.63	0.54	0.54	0.48	0.48
Precision	0.66	0.62	0.68	0.65	0.67	0.64
F1 Score	0.65	0.63	0.6	0.59	0.56	0.55

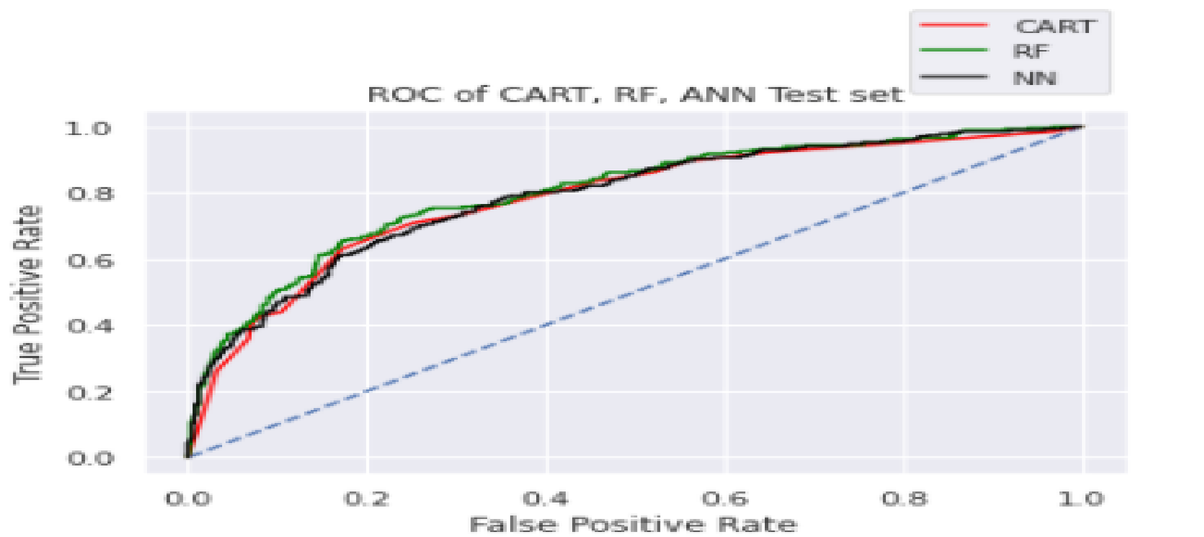
- ***We can see that the 'Artificial neural network' model gives us the best result, both AUC and neural network test set result at the tolerance rate of 0.01.***

ROC curves: CART, RF, ANN Training model:



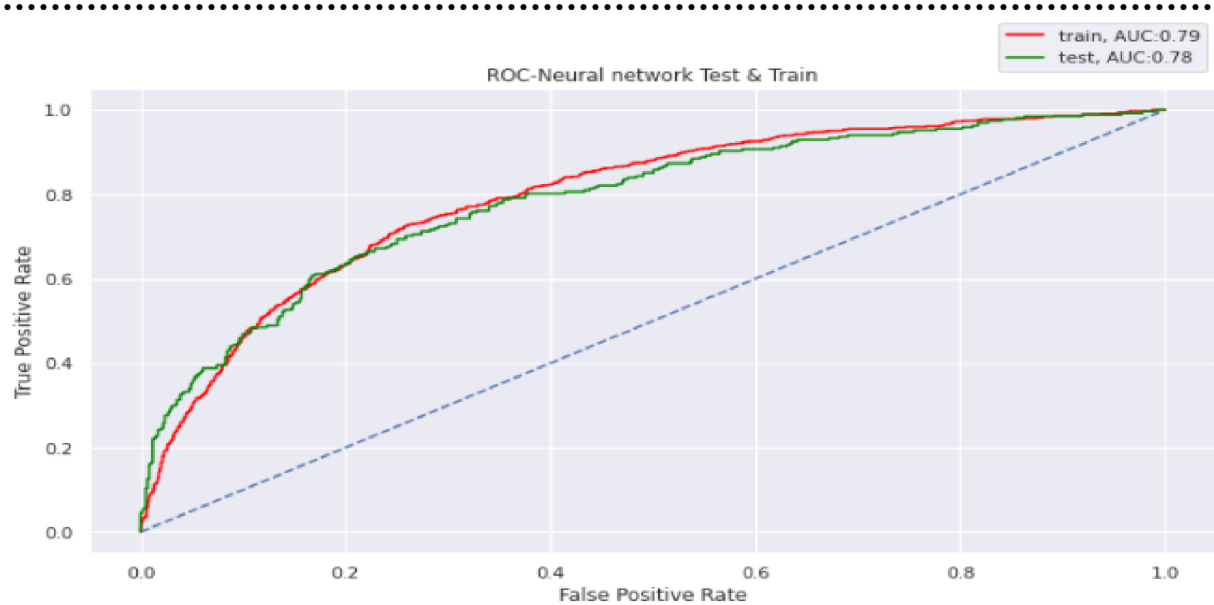
❖ ***From the above graph we can conclude that the Neural Network Classifier is the best model, when compared with other two.***

ROC curves: CART, RF, ANN Testing model:



❖ ***From the above graph we can conclude that the Neural Network Classifier is the best model, when compared with other two.***

2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations



- ***Looking at the Artificial neural network model data we were able to predict better, 80% accuracy for test sets.***
- ***The models give us around 80 or close to 80% accuracy for both test and train model, though it can still be improved, with tuning.***
- ***We see that more sales happen in Travel Agency than Airlines.***
- ***We see that the Airlines has more claims, more attention has to be given.***
- ***Tuning and making changes to the model gave better results.***
- ***As per the data we can see that insurance is done online and so are the claims.***
- ***Increasing customer satisfaction will help me revenue and also less claims.***
- ***Reducing claim handling cost.***
- ***Need to pay more attention to the JZI agency as the most claims were reported there, and also need more sales.***
- ***Improvement in Customer Service requests will reduce he claims.***

END!