DATA MINING PROJECT BUSINESS REPORT

BY

RITUSRI MOHAN

CONTENTS

PROBLEM 1

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).

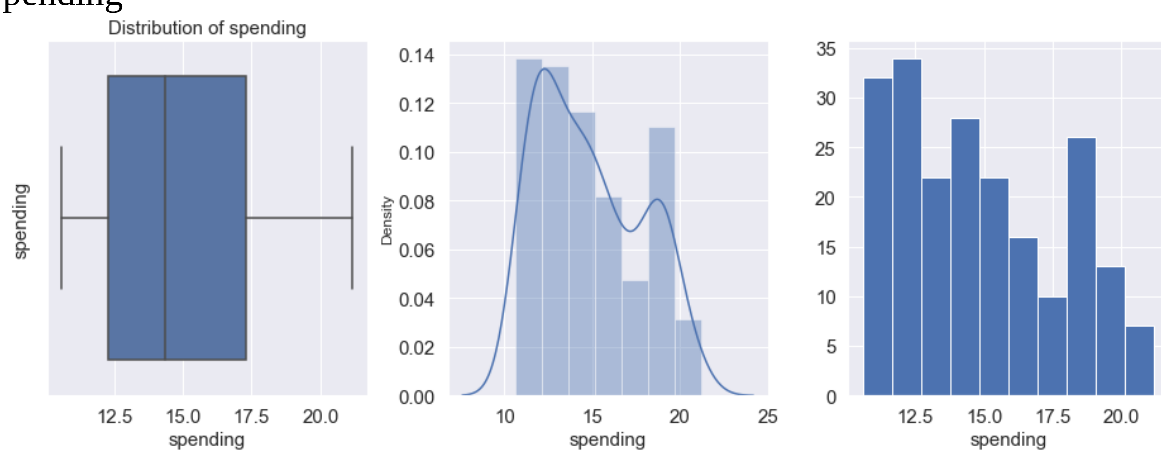
Summary statistic

	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

From the above table, we can see that for most of the variables, the mean and the median are very close values. The standard deviation is the highest for the 'Spending' variable.

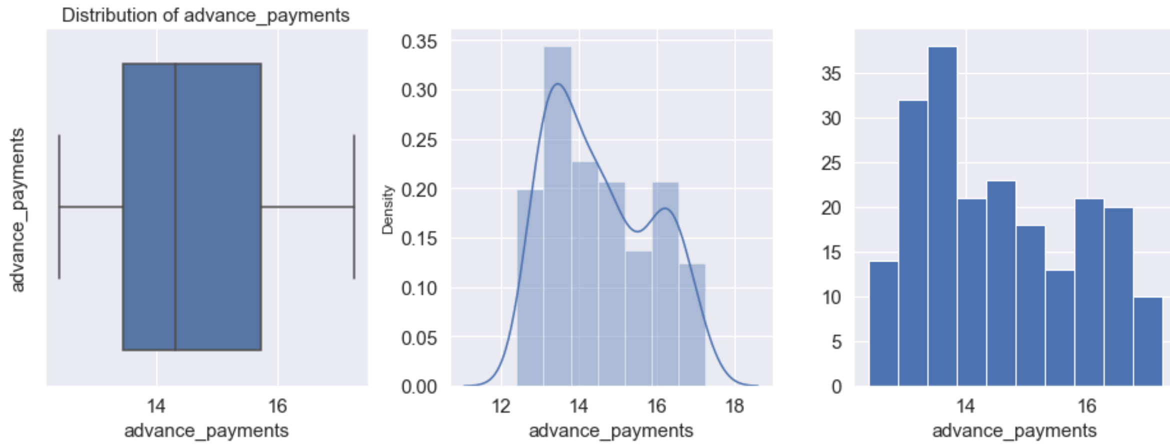
Univariate Analysis

- Spending



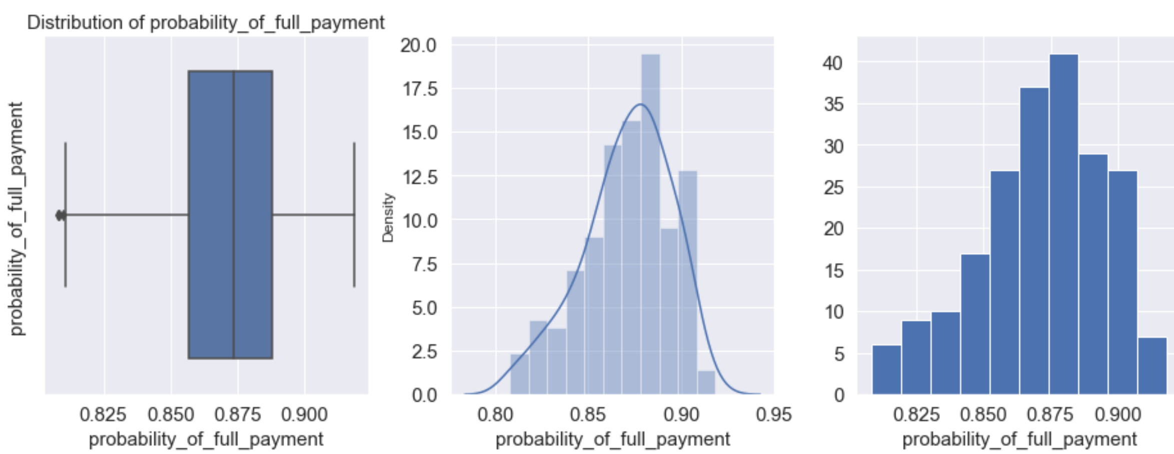
There are no outliers in the 'spending' variable. The distribution seems to be positively skewed.

• Advance payment



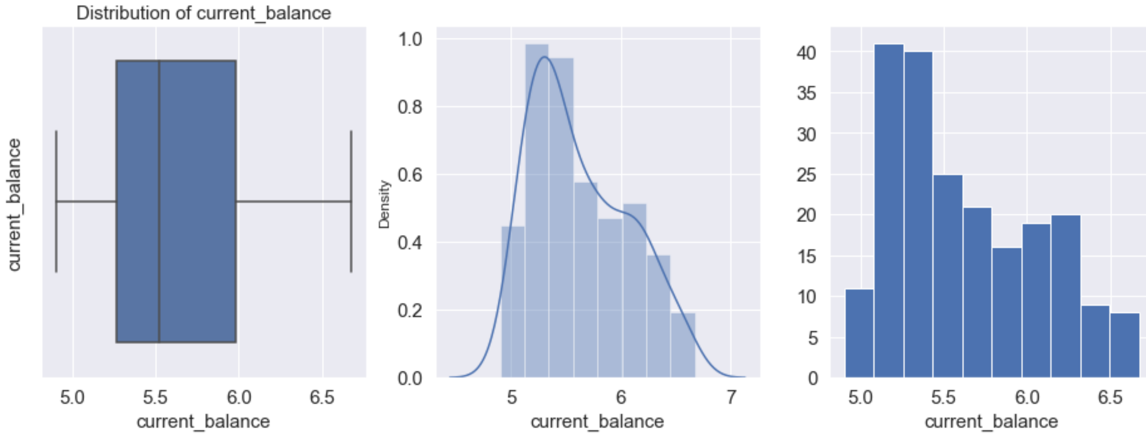
There are no outliers in the 'advance_payment' variable. The distribution seems to be bimodal.

• Probability of full payment



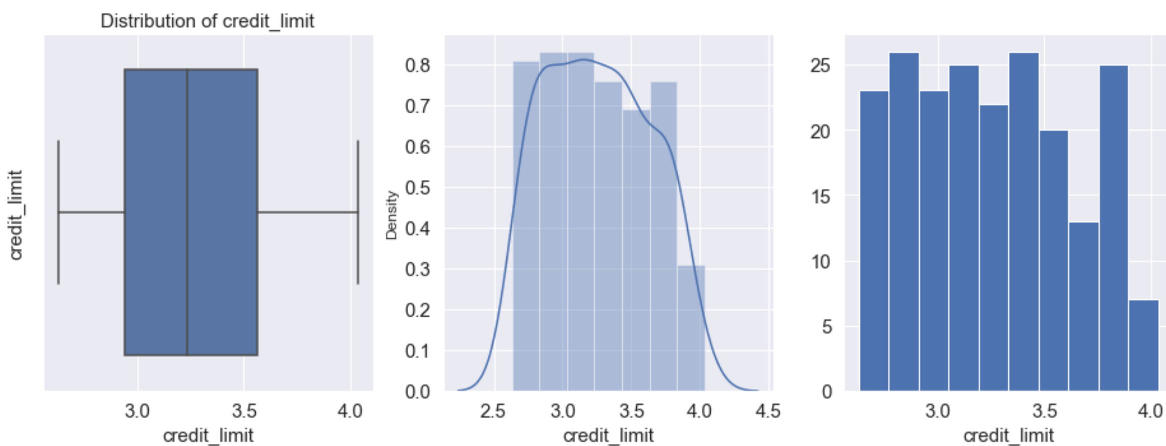
There are 3 outliers in the 'probability_of_full_payment' variable and the distribution seems to be negatively skewed.

• Current balance



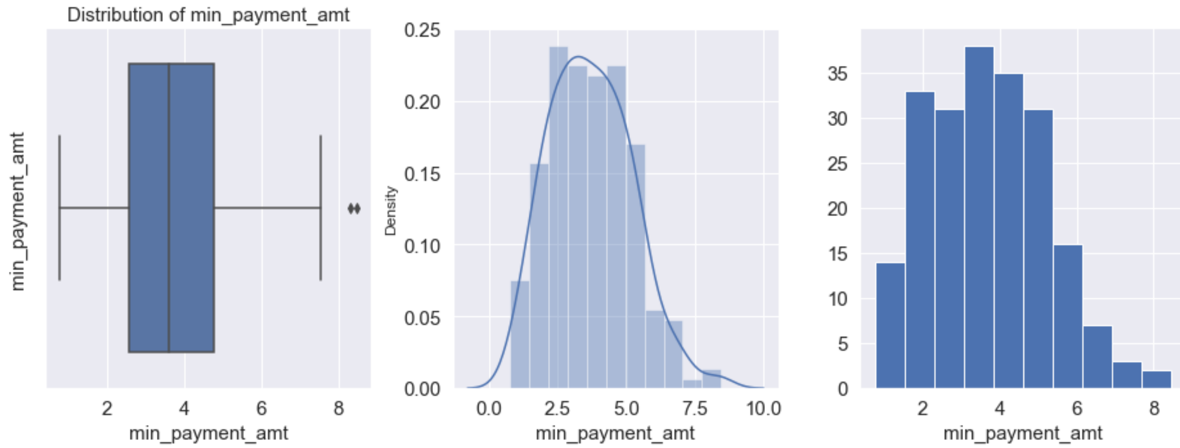
There are no outliers in the ‘current_balance’ variable. The distribution seems to be positively distributed.

• Credit limit



There are no outliers in the ‘credit_limit’ variable. The distribution seems to be normally distributed.

- Minimum payment amount



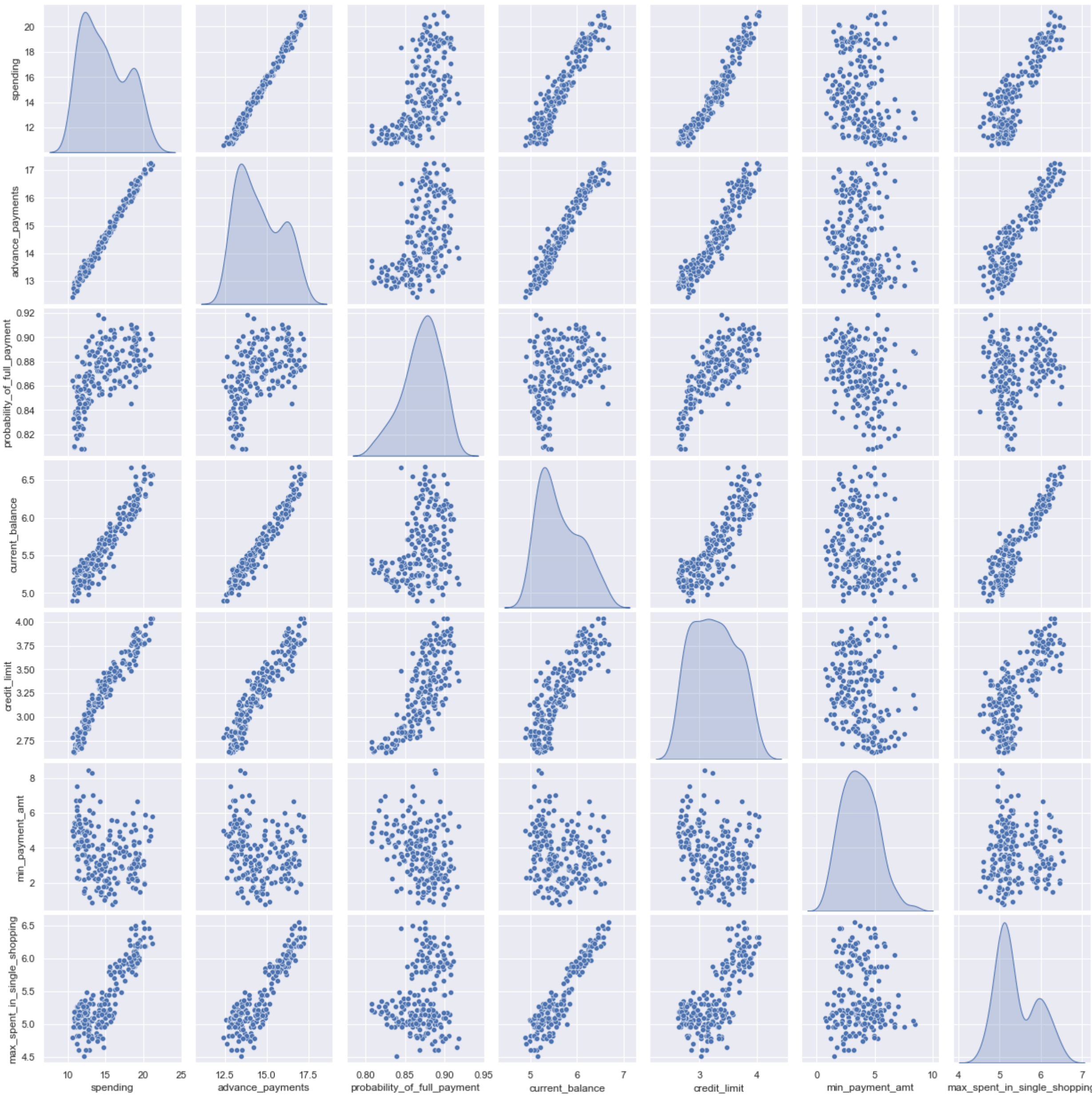
There are 2 outliers in the 'min_payment_amt' variable. The distribution seems to be normally distributed.

- Maximum spent in single shopping

There are no outliers in the 'max_spent_in_single_shopping' variable. The distribution seems to be bimodal.

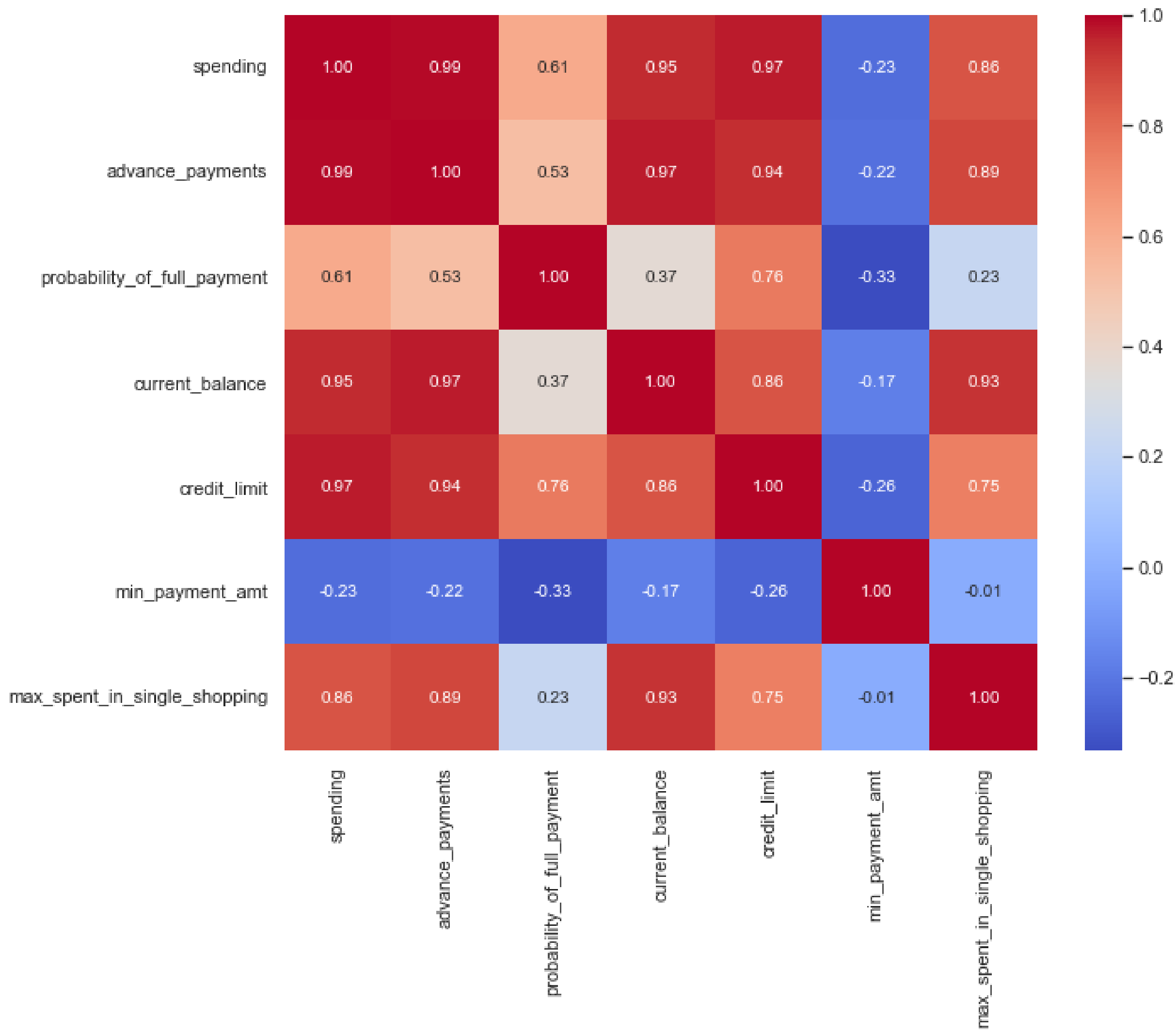
Multivariate Analysis

- Pairplot



This plot helps us to understand the relationship between all the numerical values in the dataset and establish the trends in the dataset

• Heatmap

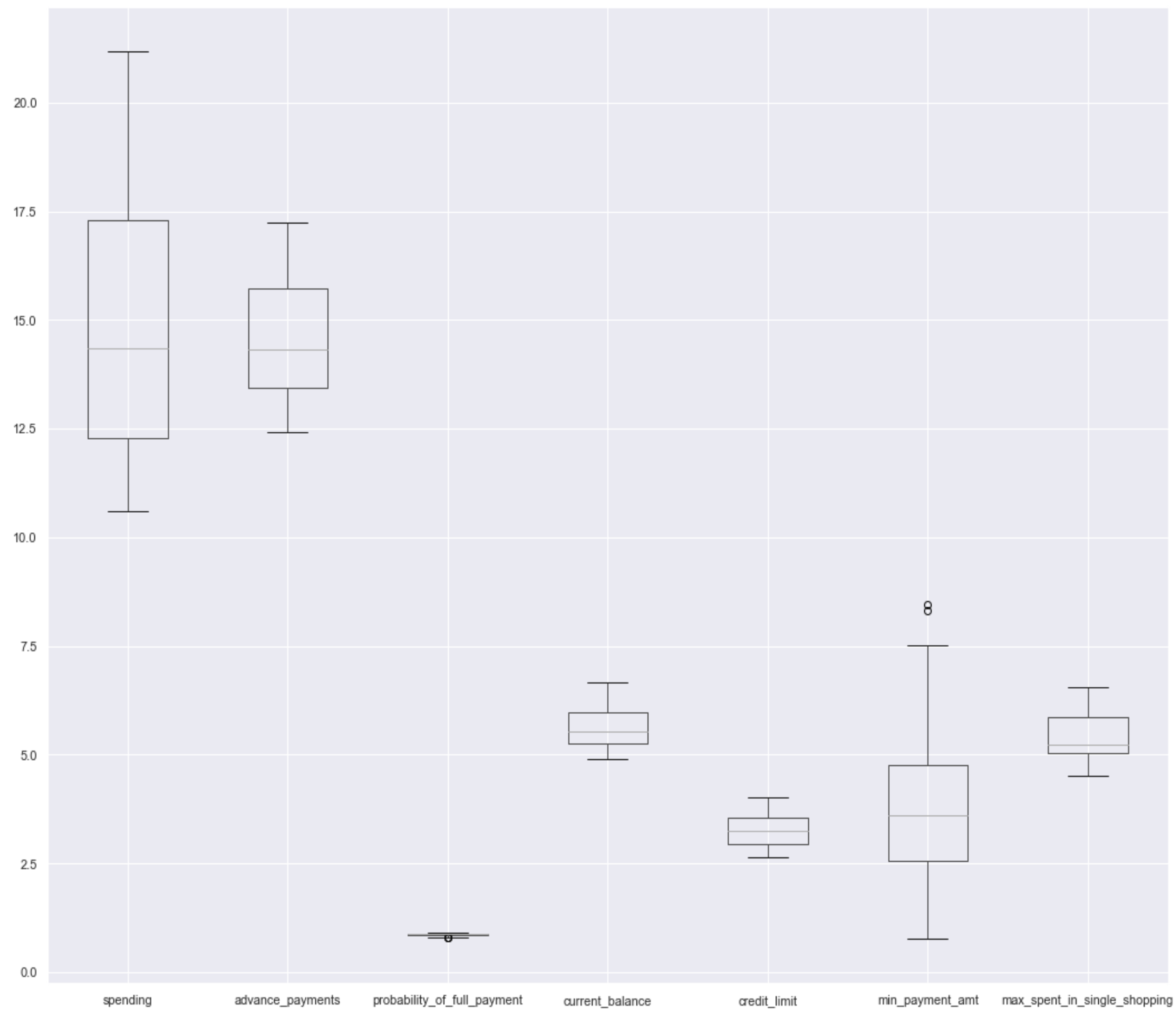


✓ From the above heatmap, it is seen that 'spending' variable is highly positively correlated with 'advance_payments' and 'credit_limit'. This means that-
  ➢ since there is advance payment of the purchases, the amount spent is higher.
  ➢ as the credit limit in the credit card is high, larger purchases are made.

✓ There strong positive correlation between 'advance_payments' and 'credit_limit'. This is because higher limit in the credit card allows one to make higher advance payments.

- ✓ There is a highly negative correlation between ‘min_payment_amt’ and ‘probability_of_full_payment’ as they are contradictory payment methods and a customer who uses one method of payment, doesn’t use the other.

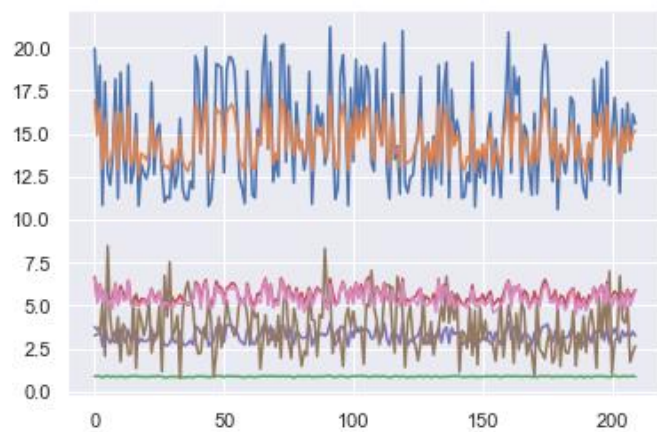
- Boxplot(outliers)



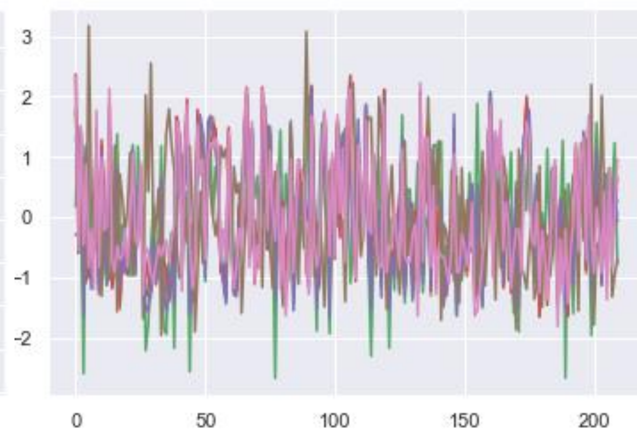
1.2 Do you think scaling is necessary for clustering in this case? Justify.

In the given dataset, the numeric values of variables: ‘Spending’ and ‘Advance_payments’ are considerably high. In order to bring about standardization in values, scaling is required. Z-score method has been used for this case study. This method tells us how much the standard deviation is away from the mean and in which direction.

The difference in the dataset values, before and after scaling can be seen below.

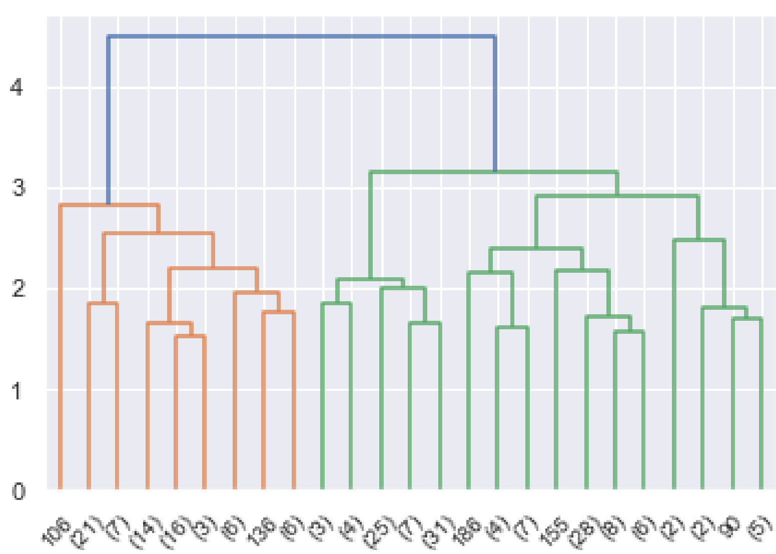


Before Scaling



After Scaling

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

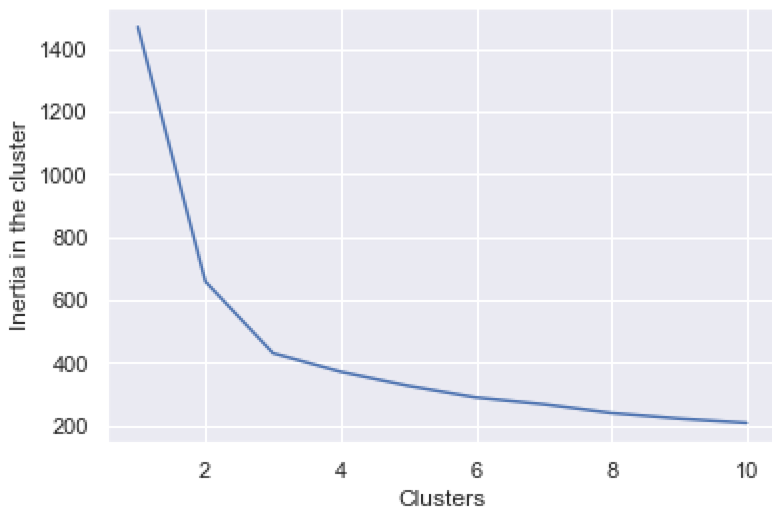


Dendrogram

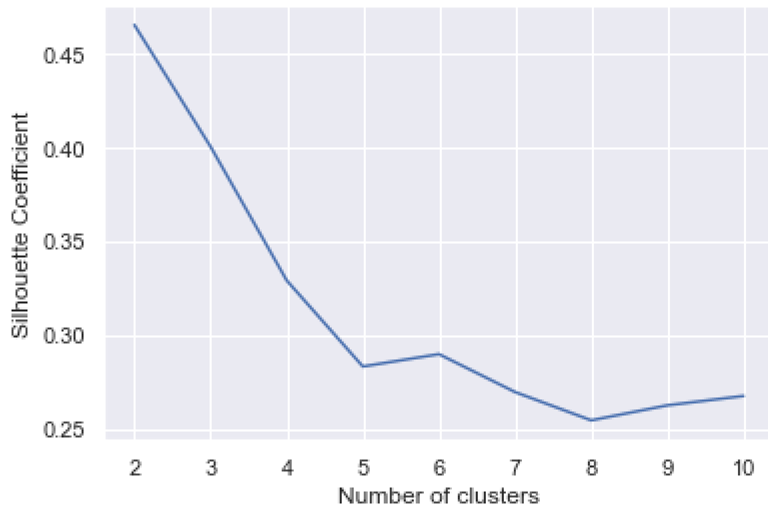
clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
1	18.129200	16.058000	0.881595	6.135747	3.648120	3.650200	5.987040	75
2	11.916857	13.291000	0.846766	5.258300	2.846000	4.619000	5.115071	70
3	14.217077	14.195846	0.884869	5.442000	3.253508	2.768418	5.055569	65

Based on the Dendrogram, 3 clusters look optimal. The 3 group clusters give a pattern of ‘spending’(high/medium/low) with ‘maximum spent in single shopping’ and ‘probability of full payment.’

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.



From the Inertia in cluster vs Clusters plot, it is seen that cluster value=3 (on the x-axis) is the elbow as it is the point where there is decrease in inertia.



In the Silhouette Coefficient Vs Number of clusters graph, the number of optimal clusters can be taken as 3 or 4. Based on the dataset, 3 cluster solution can provide a spending pattern (High/Medium/Low).

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

clusters-3	1	2	3
spending	18.129200	11.916857	14.217077
advance_payments	16.058000	13.291000	14.195846
probability_of_full_payment	0.881595	0.846766	0.884869
current_balance	6.135747	5.258300	5.442000
credit_limit	3.648120	2.846000	3.253508
min_payment_amt	3.650200	4.619000	2.768418
max_spent_in_single_shopping	5.987040	5.115071	5.055569
Freq	75.000000	70.000000	65.000000

Cluster Group Profiles:

- Group 1: High Spending
- Group 3: Medium Spending
- Group 2: Low Spending

Recommendations-

Group 1: High Spending-

- ✓ Providing the customers with reward points can increase sales.
- ✓ Discount can be offered as the maximum payment in single shopping is highest.
- ✓ Increase in credit limit will lead to larger purchases.
- ✓ Give loan against the credit card, as they are customers with good repayment record.
- ✓ Collaboration with luxurious brands can increase the maximum amount spent in one purchase

Group 3: Medium Spending-

- ✓ The customers in this group are the potential target customers as they do purchases and general payments and also have a decent credit score. For them, credit limit can be increased.
- ✓ Their spending habit can be increased by offering points on hotel stays, travel tickets and providing gift vouchers on ecommerce platforms.
- ✓ Introduction of loyalty cards can increase sales.

Group 2: Low Spending-

- ✓ The customers in this group spend the least. Therefore, reminder messages /mails can be sent.
- ✓ Payments can be improved by providing cashbacks.
- ✓ Tie up with departmental stores and companies providing household utilities like electricity, gas, phone services can increase spending.

PROBLEM 2

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.

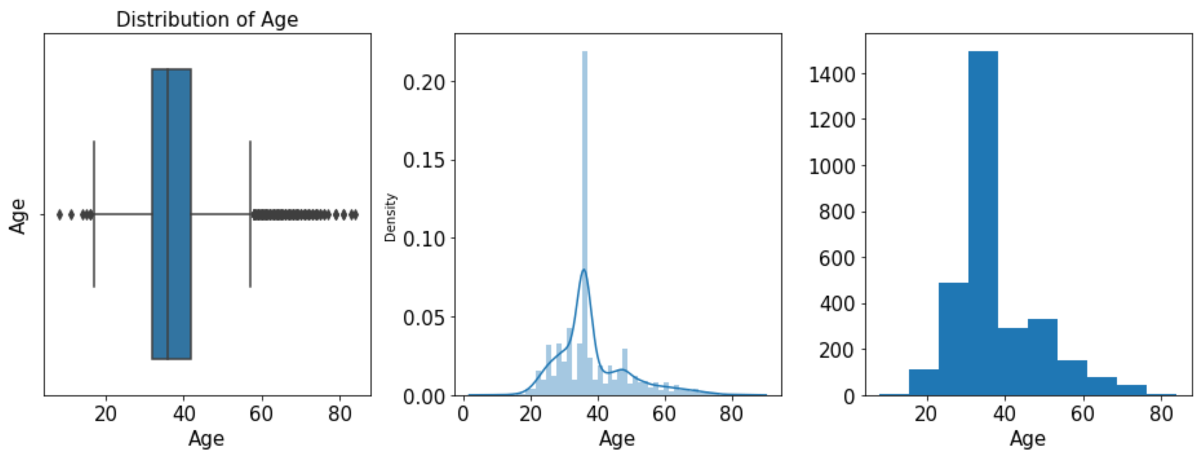
2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000	NaN	NaN	NaN	38.091	10.4635	8	32	36	42	84
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000	NaN	NaN	NaN	14.5292	25.4815	0	0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000	NaN	NaN	NaN	70.0013	134.053	-1	11	26.5	63	4580
Sales	3000	NaN	NaN	NaN	60.2499	70.734	0	20	33	69	539
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

From the above table, it is visible that for the ‘Commision’ & ‘Sales’ variables, the mean and median varies significantly. The minimum value for ‘Duration’ variable is -1 which is a wrong entry as a negative value is not possible.

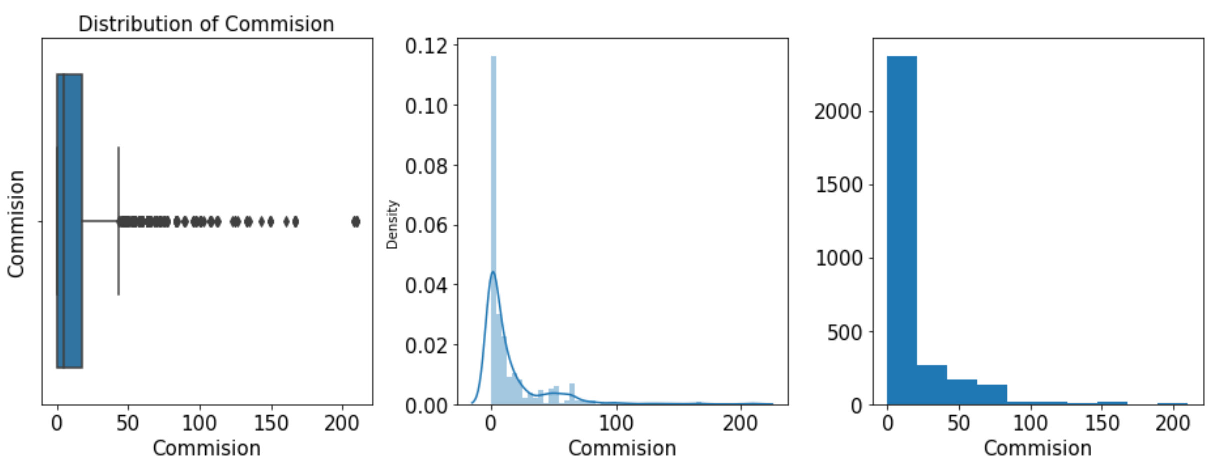
UNIVARIATE ANALYSIS

- Age



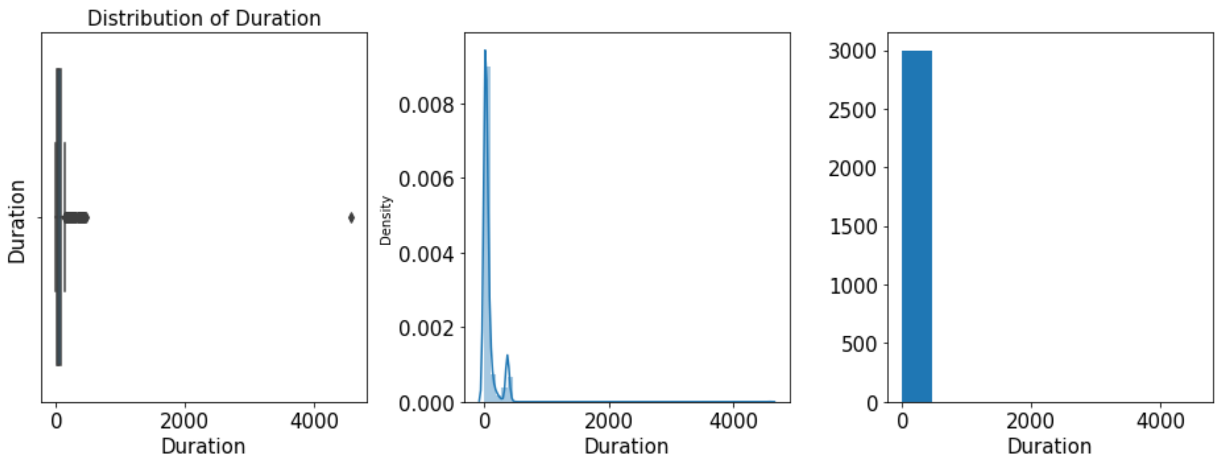
There are many outliers in the ‘Age’ variable and the distribution is positively skewed.

- Commision



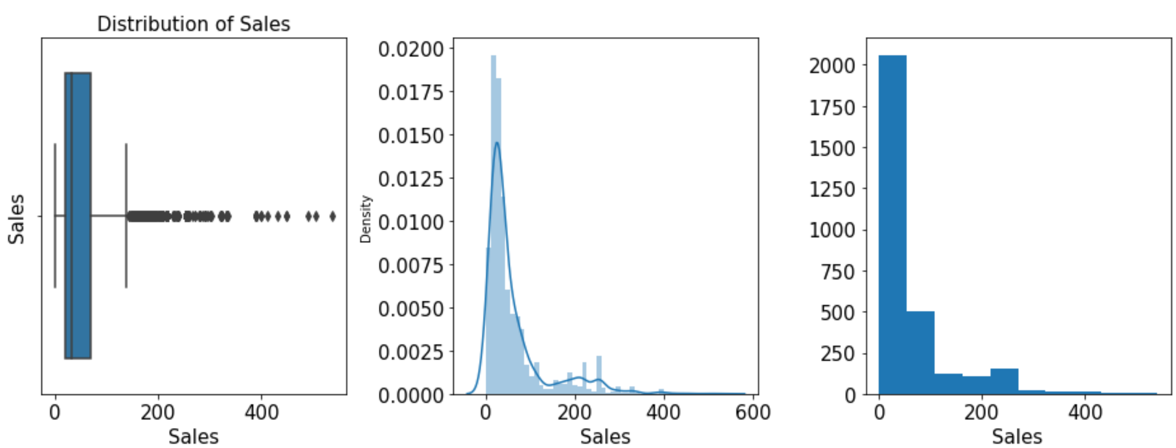
There are many outliers in the ‘Commision’ variable and the distribution is positively skewed.

- Duration



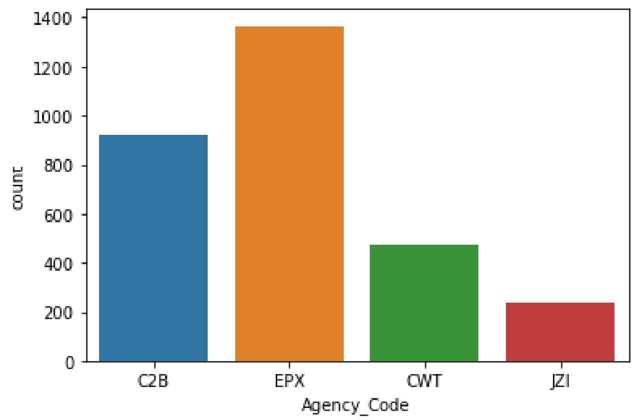
There are many outliers in the 'Duration' variable and the distribution is positively skewed

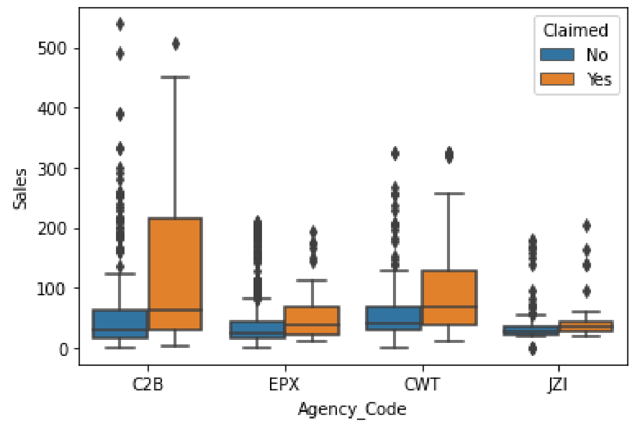
- Sales



There are many outliers in the 'Sales' variable and the distribution is positively skewed.

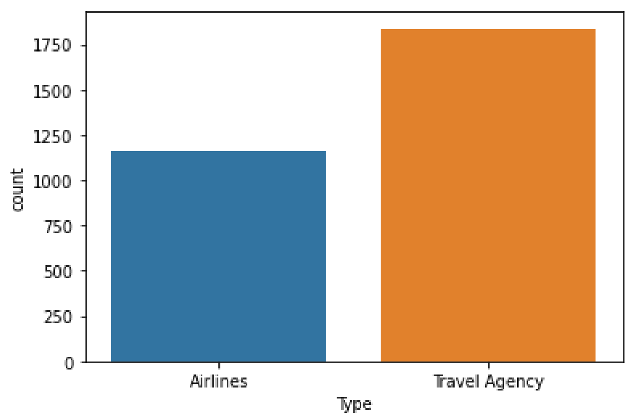
- Agency_Code

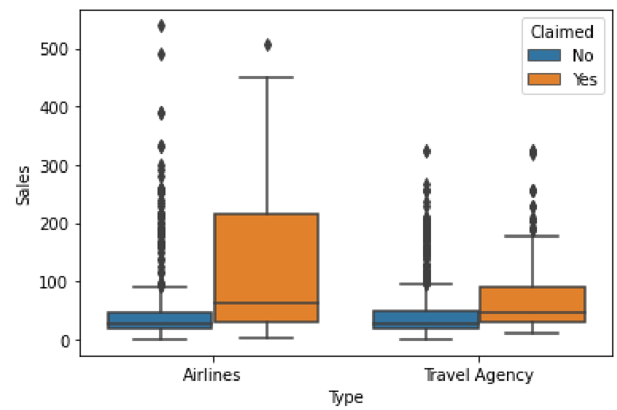




The frequency of 'EPX' in the 'Agency_Code' categorical variable is the highest when compared to other sub-categories with a value of almost 1400.

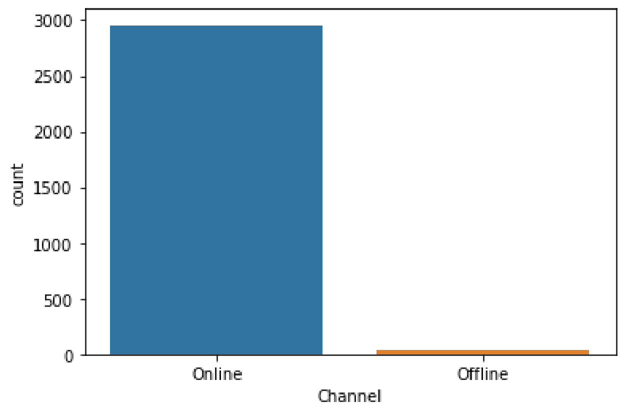
- Type

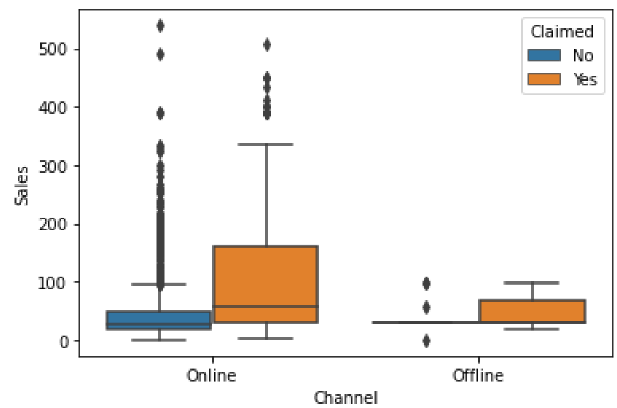




The frequency of 'Travel Agency' in the 'Type' categorical variable is the highest when compared to other sub-categories with a value of almost 1750.

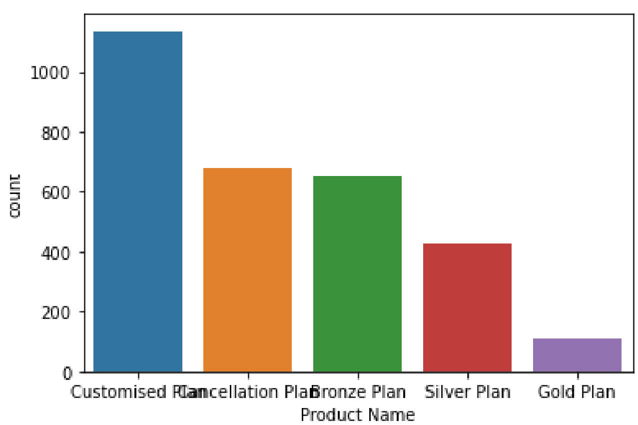
- Channel

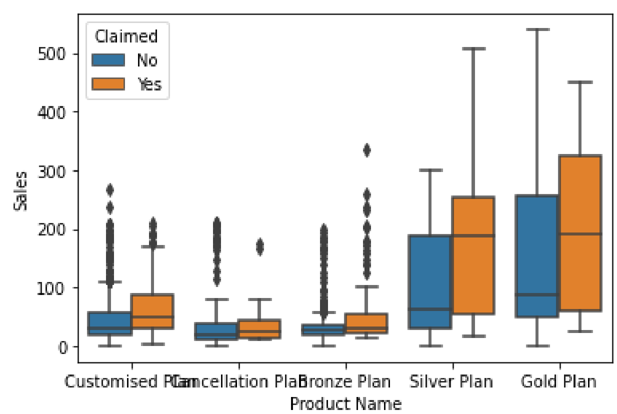




The frequency of ‘Online’ in the ‘Channel’ categorical variable is the highest when compared to other sub-categories with a value of almost 2800.

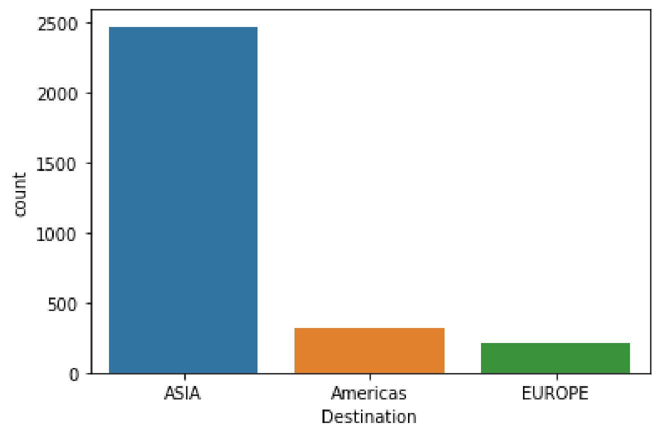
- Product Name

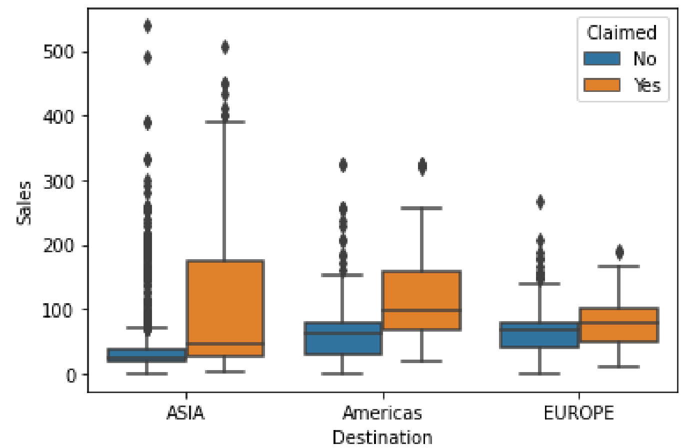




The frequency of ‘Customised’ in the ‘Product Name’ categorical variable is the highest when compared to other sub-categories with a value of almost 1200.

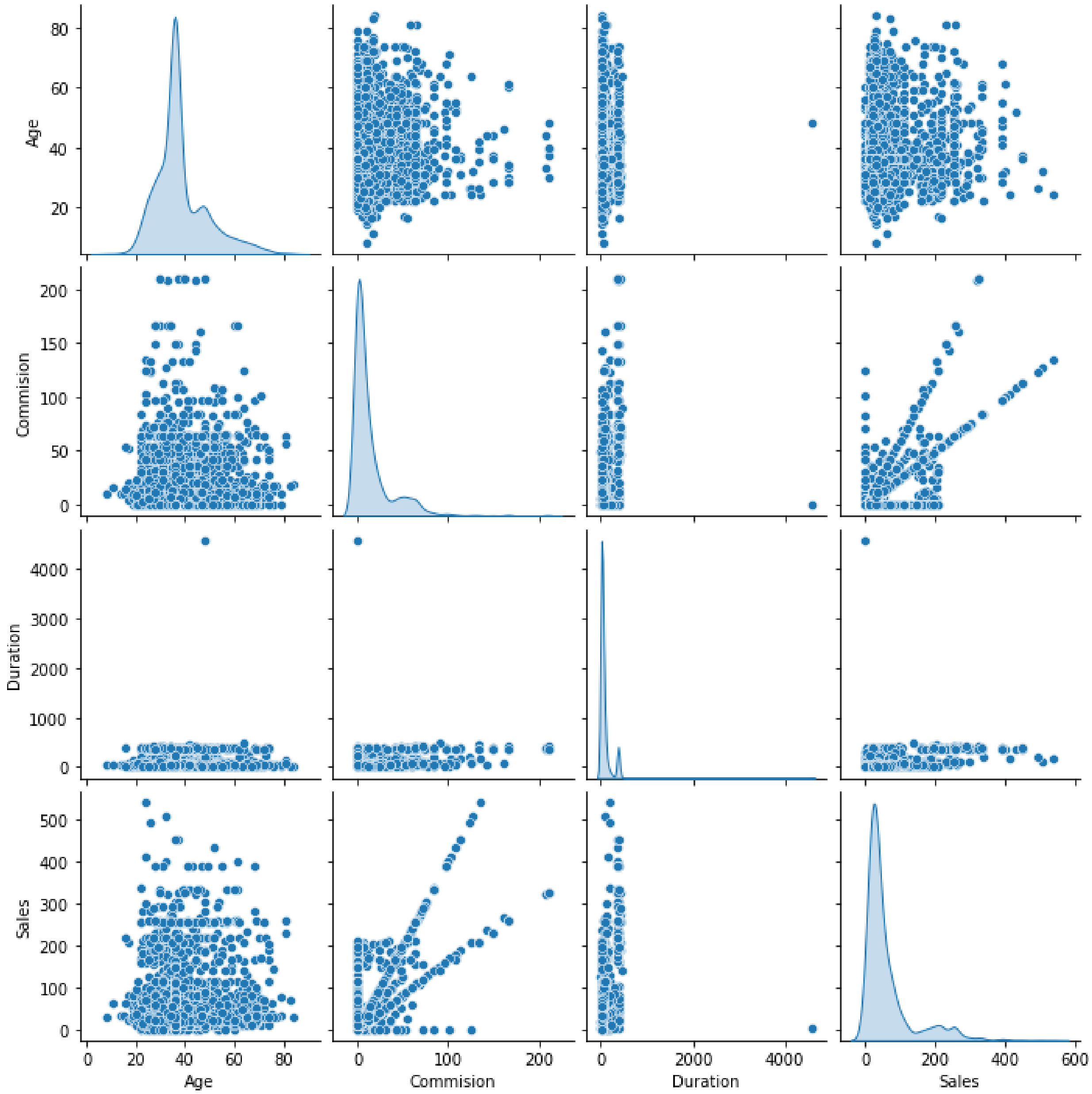
- Destination





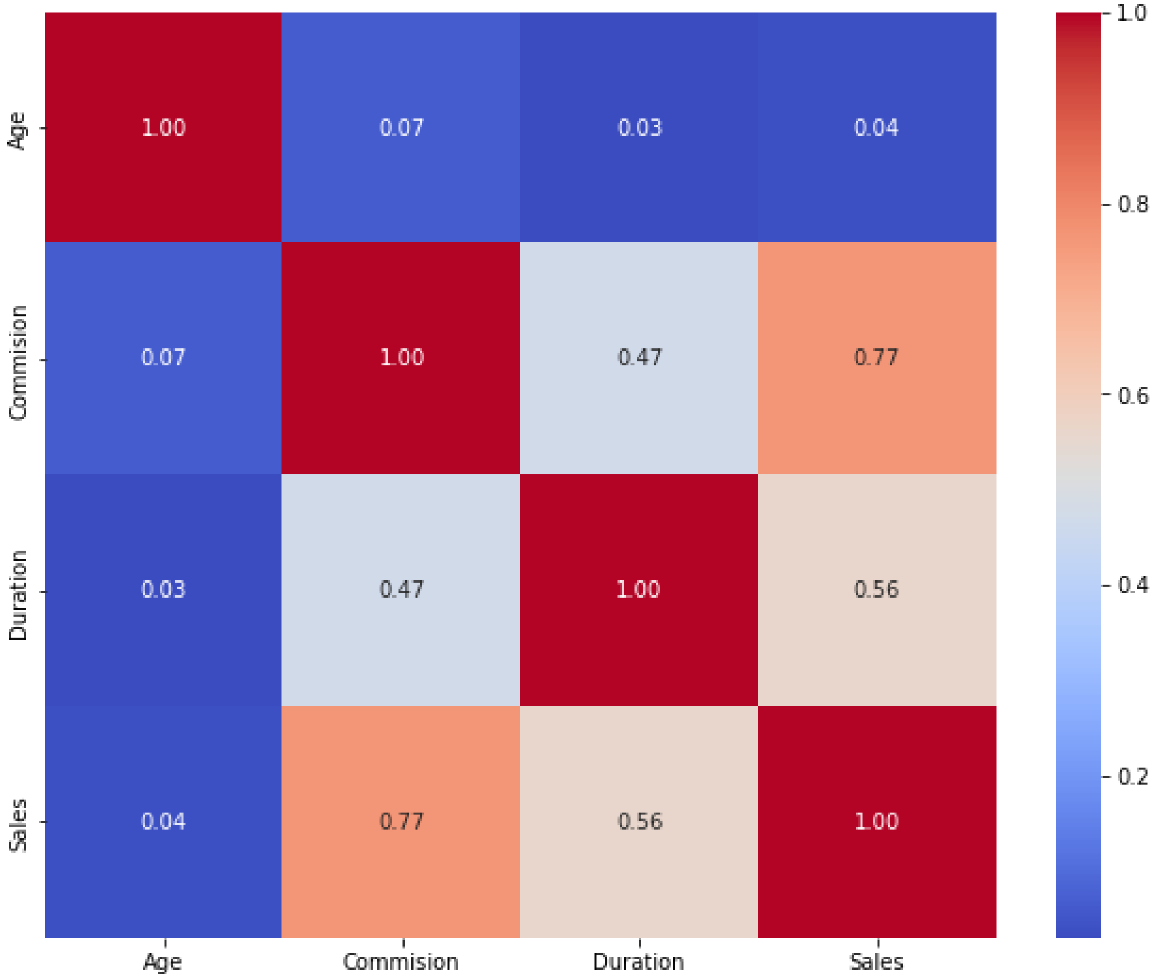
The frequency of ‘Asia’ in the ‘Destination’ categorical variable is the highest when compared to other sub-categories with a value of almost 2450.

MULTIVARIATE ANALYSIS



This plot helps us to understand the relationship between all the numerical values in the dataset and establish the trends in the dataset.

HEATMAP



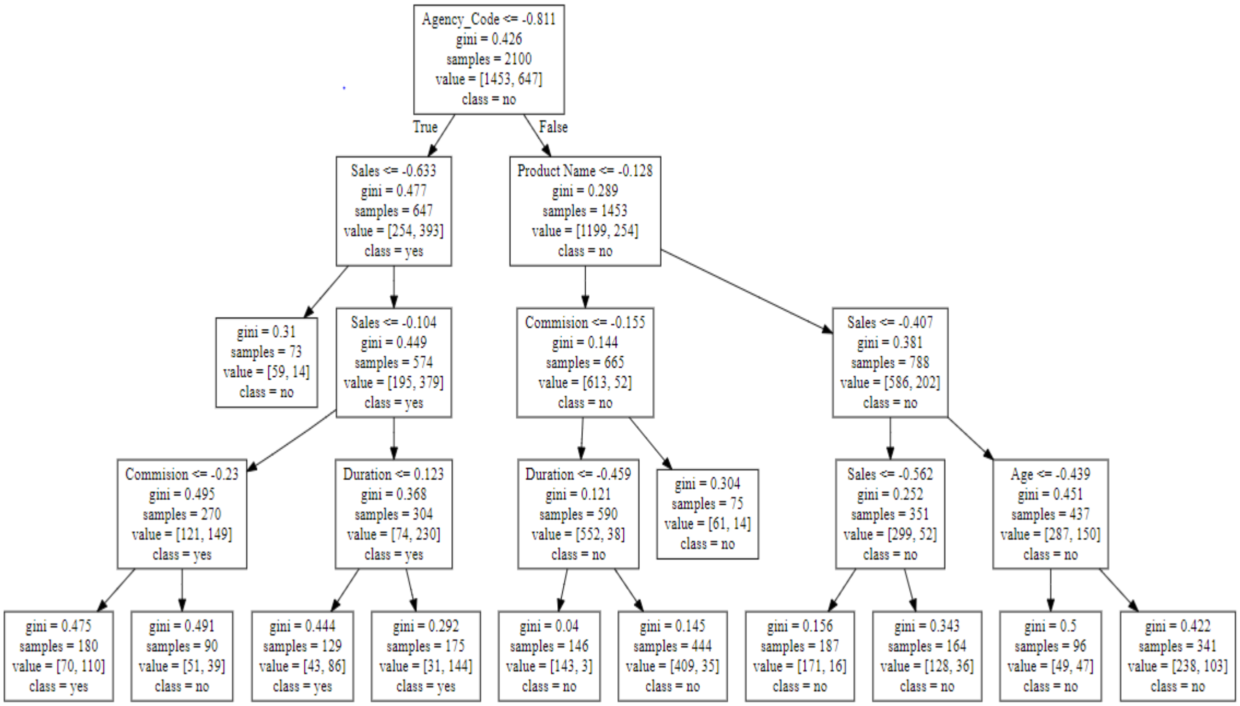
✓ There is a positive correlation between sales and commission.

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

The data was split into test and train and then the dimensions were checked for which the following output was obtained.

```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
```

- Decision Tree classifier-



Variable Importance :

	Imp
Agency_Code	0.634112
Sales	0.220899
Product Name	0.086632
Commision	0.021881
Age	0.019940
Duration	0.016536
Type	0.000000
Channel	0.000000
Destination	0.000000

Predicted Classes and Probabilities:

	0	1
0	0.697947	0.302053
1	0.979452	0.020548
2	0.921171	0.078829
3	0.510417	0.489583
4	0.921171	0.078829

- Random Tree classifier-

Variable Importance:

	Imp
Agency_Code	0.565184
Sales	0.210684
Product Name	0.105034
Duration	0.051381
Commision	0.034964
Age	0.029495
Destination	0.002392
Type	0.000866
Channel	0.000000

Predicted Classes and Probabilities:

	0	1
0	0.764425	0.235575
1	0.988304	0.011696
2	0.905682	0.094318
3	0.561293	0.438707
4	0.883504	0.116496

- Neural Network classifier-

Predicted Classes and Probabilities:

	0	1
0	0.758510	0.241490
1	0.800720	0.199280
2	0.796798	0.203202
3	0.710660	0.289340
4	0.731916	0.268084

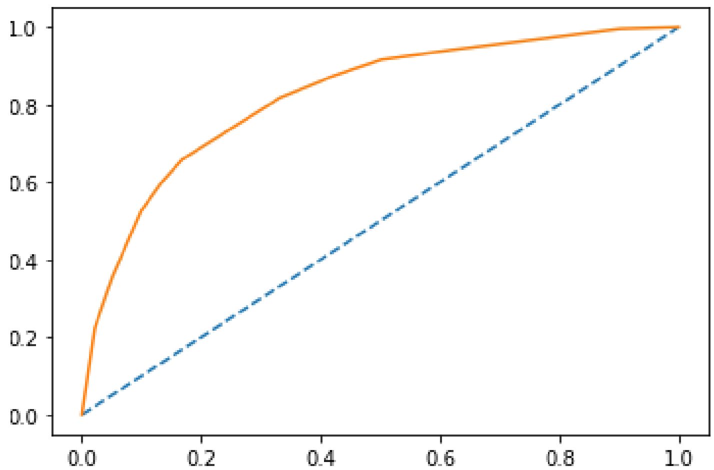
2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

- Decision Tree Classifier

  ➢ *Training data*

Area under the curve: 0.82

ROC:



Confusion matrix:

```
array([[1309,  144],
       [ 307,  340]], dtype=int64)
```

Classification report:

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1453
1	0.70	0.53	0.60	647
accuracy			0.79	2100
macro avg	0.76	0.71	0.73	2100
weighted avg	0.78	0.79	0.78	2100

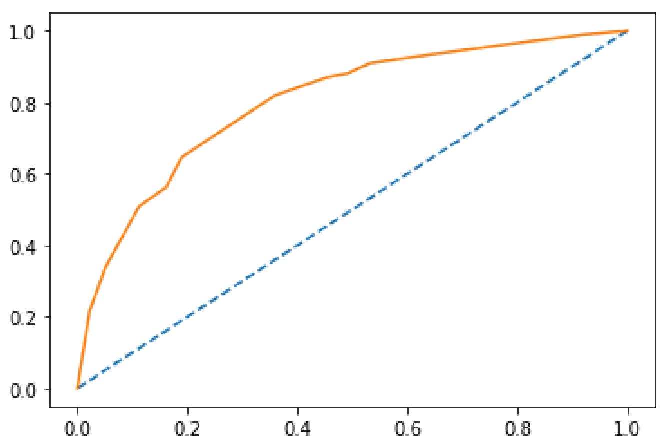
Cart Inference for training data –

- ✓ AUC: 82%
- ✓ Accuracy: 79%
- ✓ Precision: 70%
- ✓ f1-Score: 60%

➢ *Test data*

Area under the curve: 0.80

ROC:



Confusion matrix:

```
array([[553,  70],
       [136, 141]], dtype=int64)
```

Classification report:

```
              precision    recall  f1-score   support

           0       0.80      0.89      0.84       623
           1       0.67      0.51      0.58       277

    accuracy                           0.77       900
   macro avg       0.74      0.70      0.71       900
weighted avg       0.76      0.77      0.76       900
```

Cart Inference for test data –

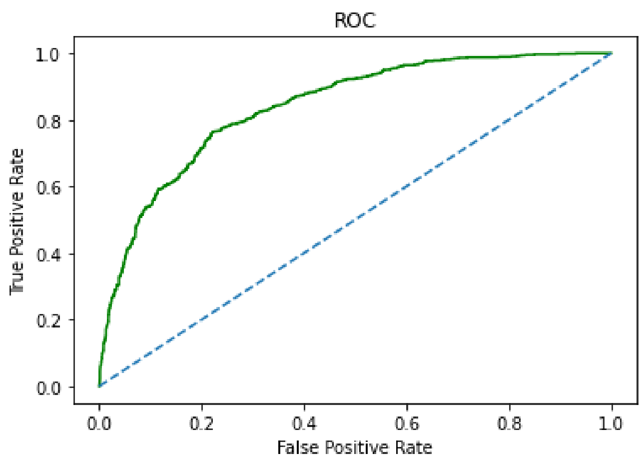
- ✓ AUC: 80%
- ✓ Accuracy: 77%
- ✓ Precision: 80%
- ✓ f1-Score: 84%

- Random Forest Classifier

  - *Training data*

Area under the curve: 0.84

ROC:



Confusion matrix:

```
array([[1294,  159],
       [ 276,  371]], dtype=int64)
```

Classification report:

	precision	recall	f1-score	support
0	0.82	0.89	0.86	1453
1	0.70	0.57	0.63	647
accuracy			0.79	2100
macro avg	0.76	0.73	0.74	2100
weighted avg	0.79	0.79	0.79	2100

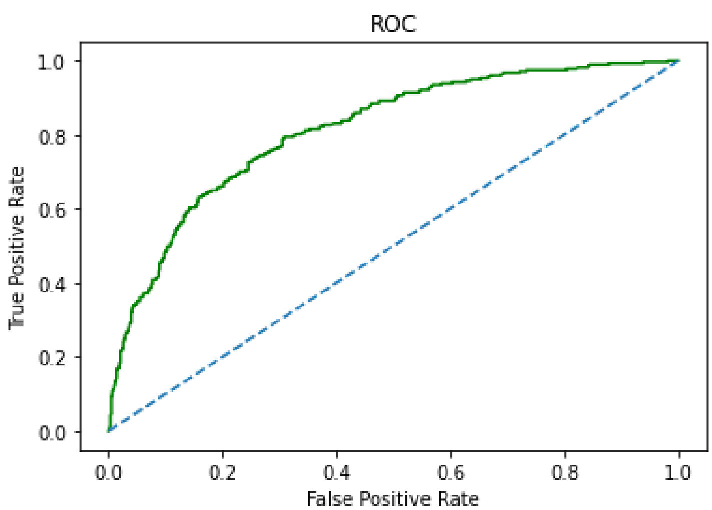
Cart Inference for training data –

- ✓ AUC: 84%
- ✓ Accuracy: 79%
- ✓ Precision: 70%
- ✓ f1-Score: 63%

➢ *Test data*

Area under the curve: 0.81

ROC:



Confusion matrix:

```
array([[548,  75],
       [125, 152]], dtype=int64)
```

Classification report:

```
              precision    recall  f1-score   support

           0       0.81      0.88      0.85       623
           1       0.67      0.55      0.60       277

    accuracy                           0.78       900
   macro avg       0.74      0.71      0.72       900
weighted avg       0.77      0.78      0.77       900
```

Cart Inference for test data –

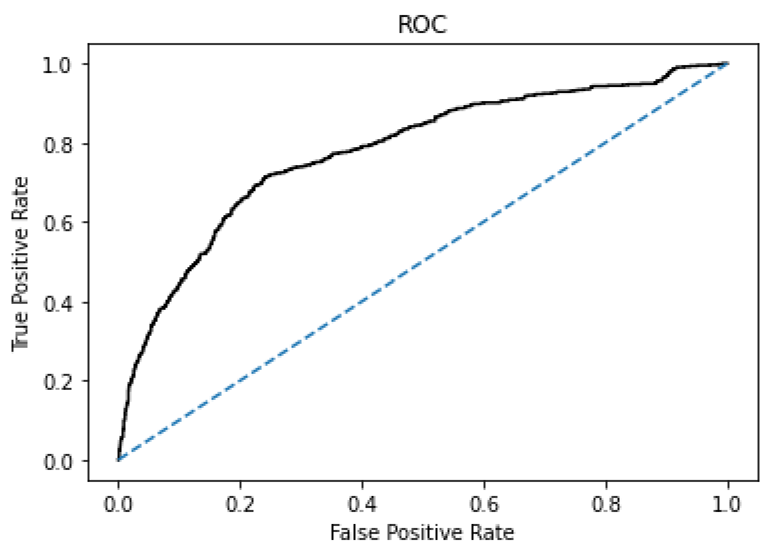
- ✓ AUC: 82%
- ✓ Accuracy: 78%
- ✓ Precision: 67%
- ✓ f1-Score: 60%

- Neural Network Classifier

➢ *Training data*

Area under the curve: 0.77

ROC:



Confusion matrix:

```
array([[1340,  113],
       [ 396,  251]], dtype=int64)
```

Classification report:

	precision	recall	f1-score	support
0	0.77	0.92	0.84	1453
1	0.69	0.39	0.50	647
accuracy			0.76	2100
macro avg	0.73	0.66	0.67	2100
weighted avg	0.75	0.76	0.73	2100

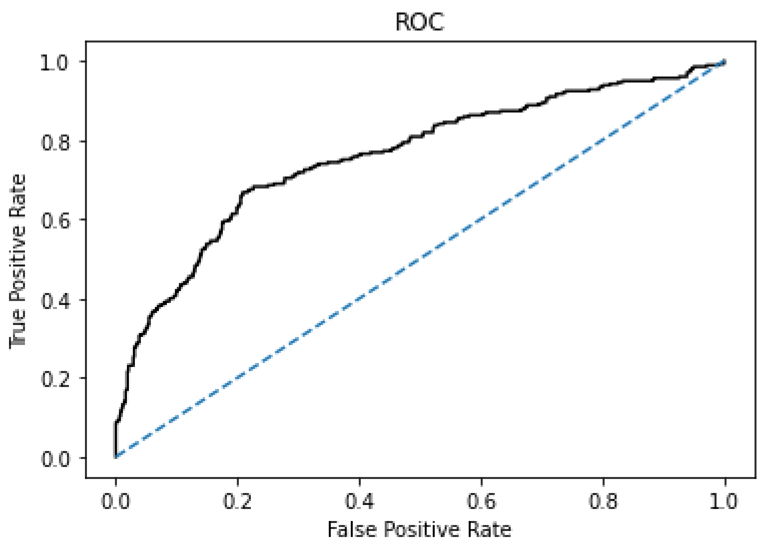
Cart Inference for training data –

- ✓ AUC: 77%
- ✓ Accuracy: 76%
- ✓ Precision: 69%
- ✓ f1-Score: 50%

- ➢ *Test data*

Area under the curve: 0.75

ROC:



Confusion matrix:

```
array([[576,  47],
       [171, 106]], dtype=int64)
```

Classification report:

	precision	recall	f1-score	support
0	0.77	0.92	0.84	623
1	0.69	0.38	0.49	277
accuracy			0.76	900
macro avg	0.73	0.65	0.67	900
weighted avg	0.75	0.76	0.73	900

Cart Inference for test data –

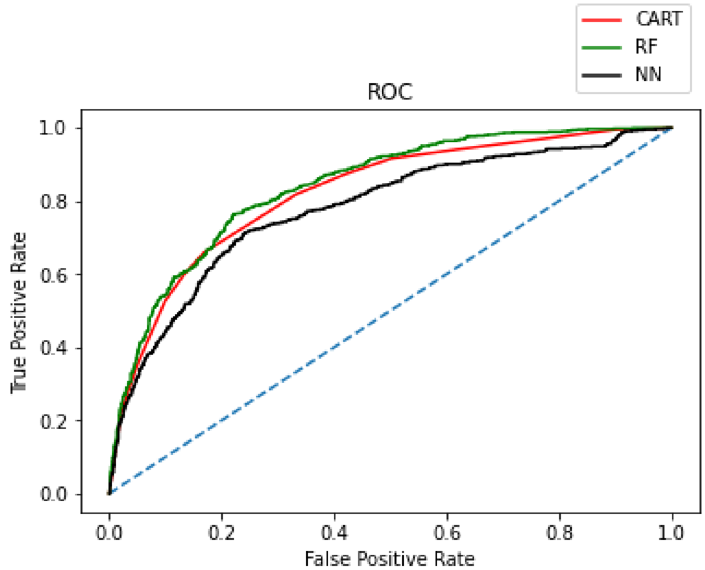
- ✓ AUC: 75%
- ✓ Accuracy: 76%
- ✓ Precision: 69%
- ✓ f1-Score: 49%

Since the Training and Test set results are almost similar, the model is a good model.

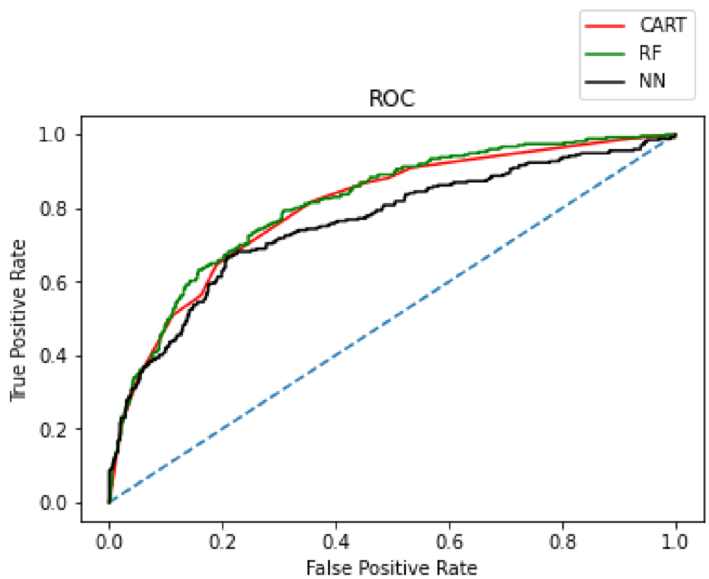
2.4 Final Model: Compare all the models and write an inference which model is best/optimized.

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.79	0.77	0.79	0.78	0.76	0.76
AUC	0.82	0.80	0.84	0.81	0.78	0.76
Recall	0.53	0.51	0.57	0.55	0.39	0.38
Precision	0.70	0.67	0.70	0.67	0.69	0.69
F1 Score	0.60	0.58	0.63	0.60	0.50	0.49

ROC for all models using *Training* data :



ROC for all models using *Test* data:



✓ The best model is the Random forest classifier model as it has better accuracy, precision, recall, f1 score when compared to CART and Neural network classifier.

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

- ✓ The JZI agency needs various means to pick up sales. This can be done by starting a promotional marketing campaign or try to tie up with alternate agency.
- ✓ The number of sales is more via Agency rather than Airlines. There needs to be a thorough check in the way in which the Agency works.
- ✓ Claim handling costs can be reduced.
- ✓ Claim cycle time can be reduced.