000011 01
Lorem ipsum dolor sit amet, consect adipiscing elit, sed diam nonummy nibh euismod.
101001 02
Lorem ipsum dolor sit amet, consect adipiscing elit, sed diam nonummy nibh euismod.
010110 03
Lorem ipsum dolor sit amet, consect adipiscing elit, sed diam nonummy nibh euismod.
001011 04
Lorem ipsum dolor sit amet, consect adipiscing elit, sed diam nonummy nibh euismod.
21%
5%
56%

DATA MINING PROJECT

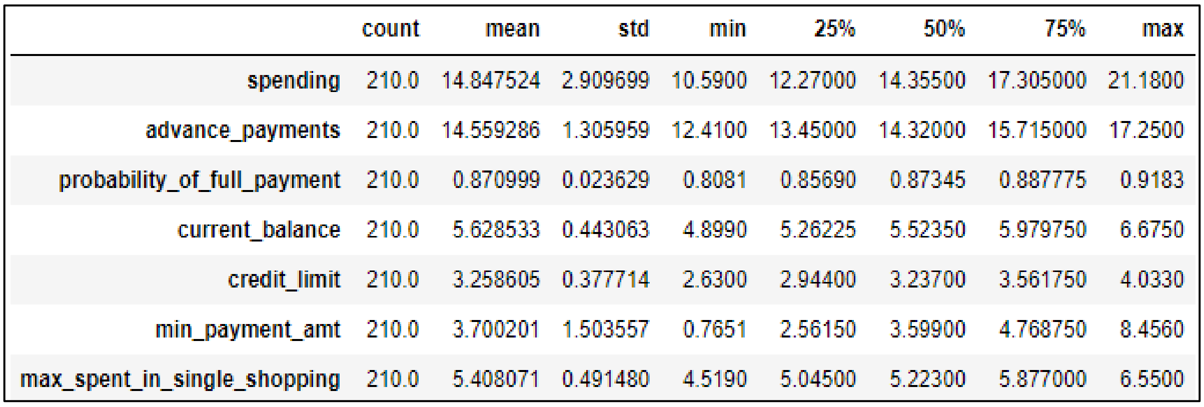
Balasubramaniyam. R

Table of Contents

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. Please note that it is a summarized data that contains the average values in all the columns considering all the months, and not for any particular month. 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).



	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

1.1.1. Data summary table

i. Spending of our customers is between Rs.10,590 to a maximum of Rs.21,180. 75% of the spending is above Rs.17,305.

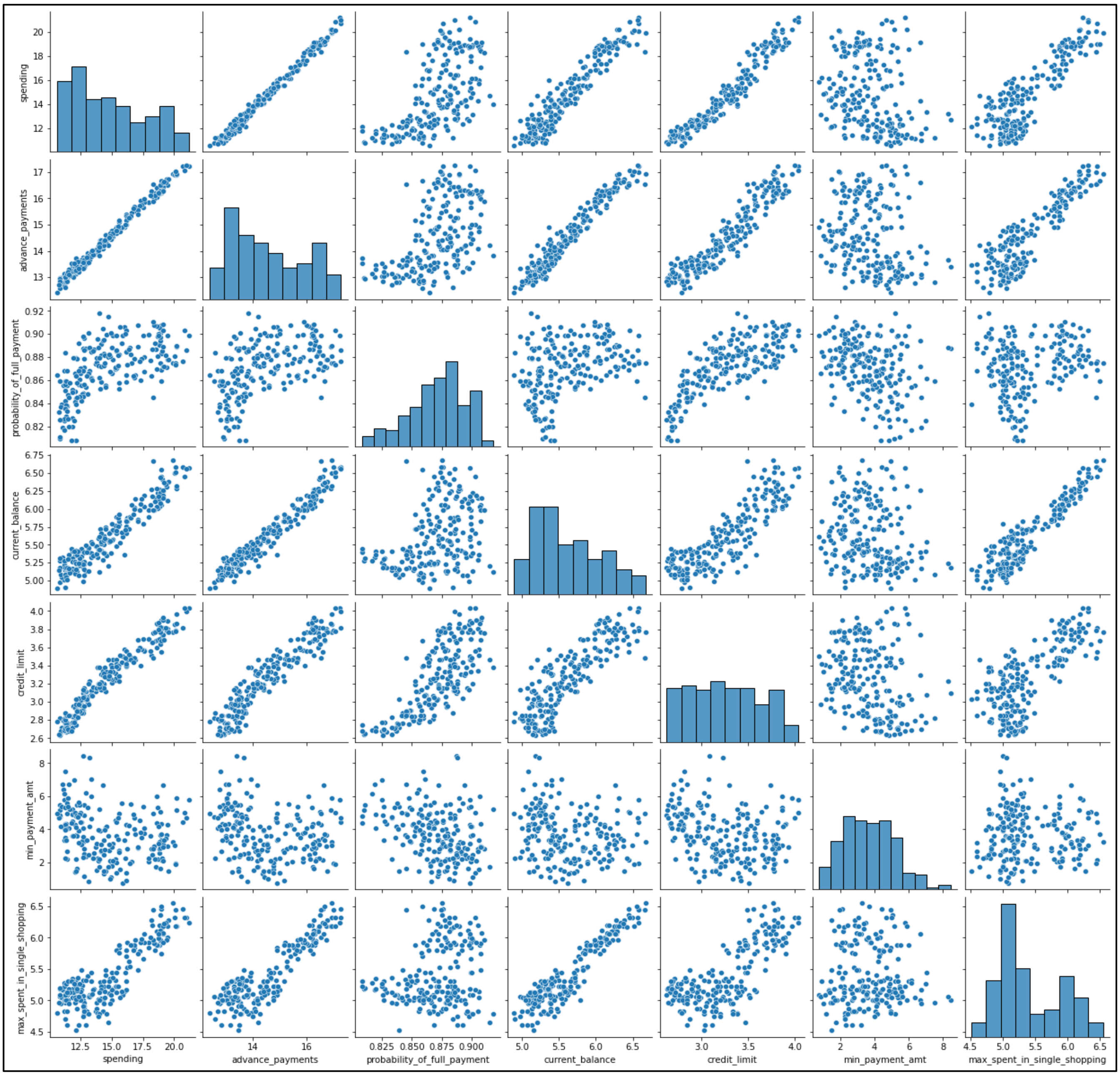
ii. The probability of the credit card payment done in full by the customer to the bank is between 80.8% to a maximum of 91.8%.

iii. Amount paid by the customer in advance by cash even before the credit card bill got generated for any month is between Rs.1241 to a maximum of Rs.1725. 75% of advance payment is above Rs.1571.

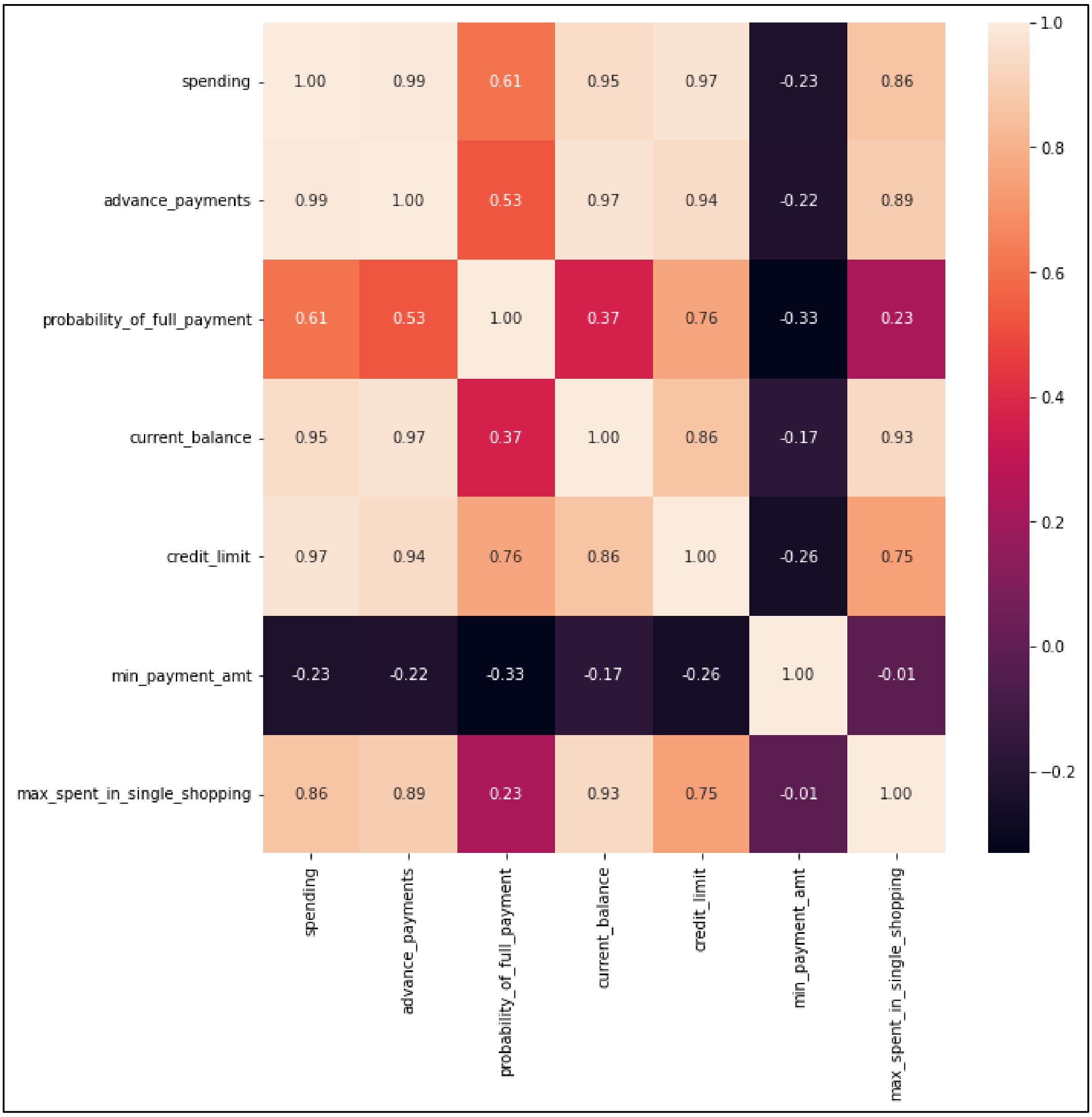
iv. Limit of the amount in credit card (10000s) sanctioned by the bank to the customer is between Rs.26,300 to a maximum of Rs.40,330.

v. The 75% of the balance amount left in the credit card account to make the future purchases is above Rs. 5,979 and the maximum is Rs. 6,675.

vi. The Maximum amount spent by the customer for a single transaction using the credit card is Rs. 6,550.



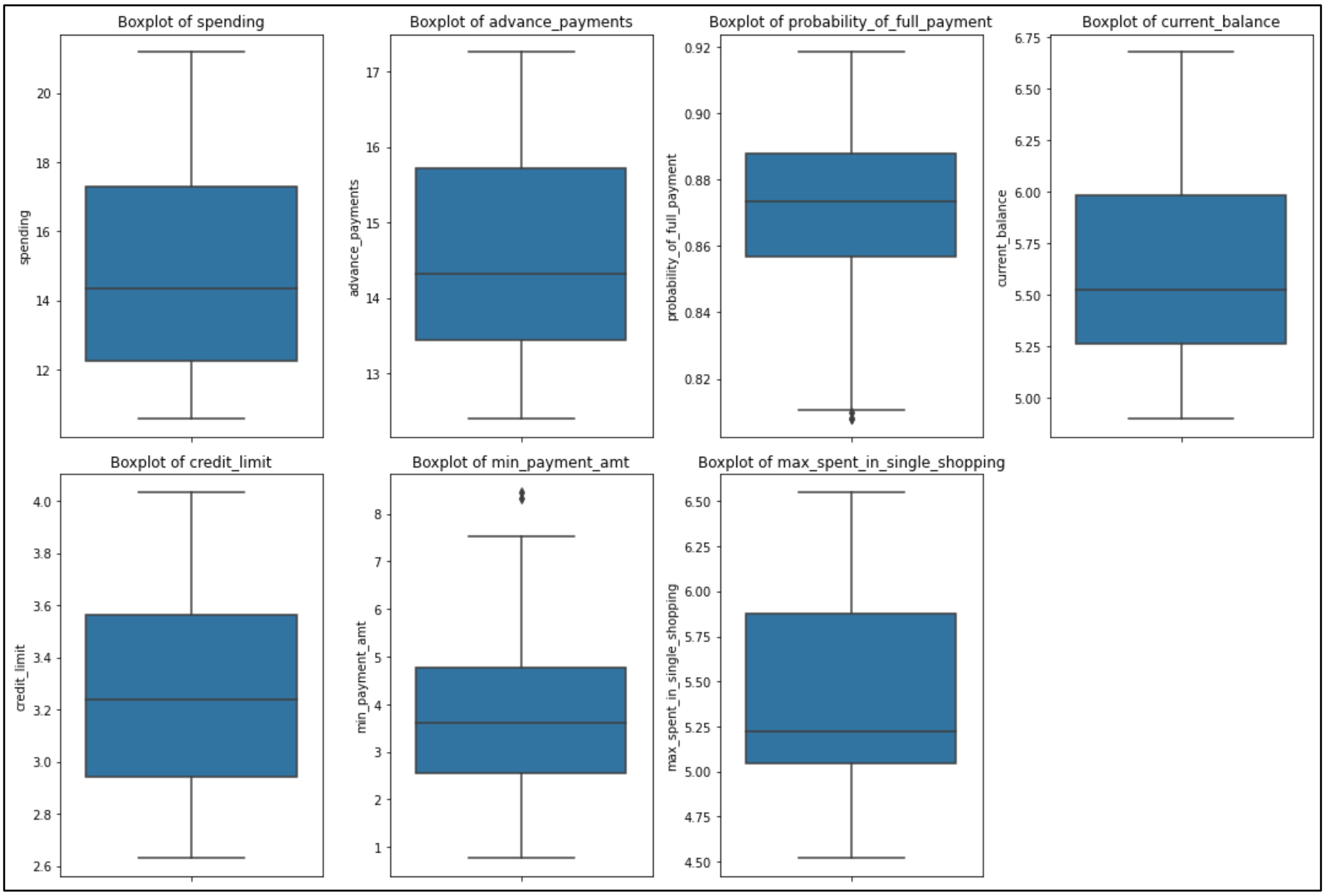
1.1.2. Pairplot



1.1.3. Correlation matrix

There is a strong positive correlation between the customer spending and Advance payments, Current balance, Credit limit, maximum amount spent in a single transaction.

The correlation between the customer spending and probability of full payment is at 0.61. Remaining 39% customers can be identified and provided with an option to use EMI and bank can make profits from the interest amount.

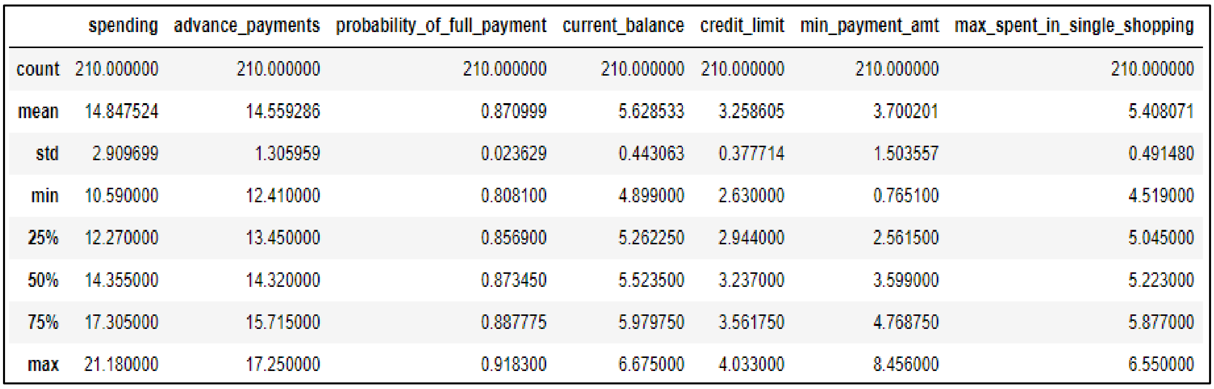


1.1.4. Boxplot of data

The dataset is having outliers in the probability of full payment and the min payment amt.

There is right skewed data in the spending, advance payment, current balance, and max spent in single shopping features. There is left skewed data in the probability of full payment feature.

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

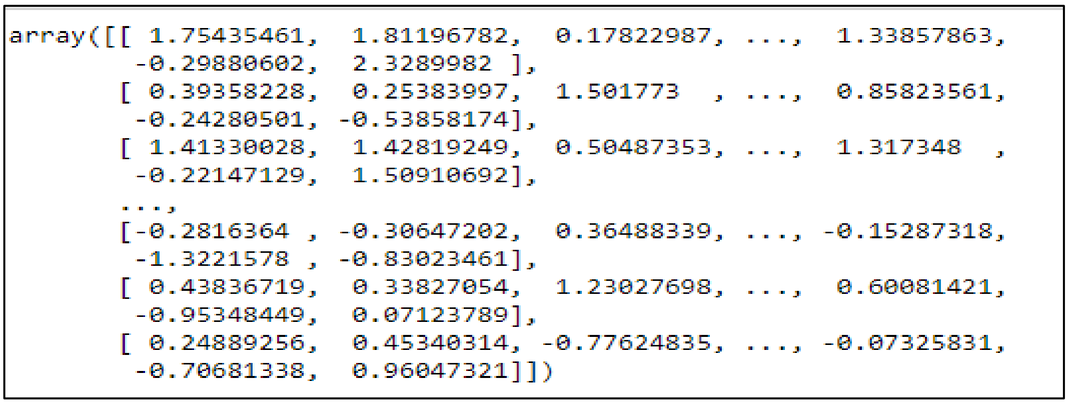


	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

1.2.1. Summary of data

By using the describe () function we can summarize our data set, from this summary the dataset is having datapoints which are far from each other in the available columns.

For example: the mean, std dev values in the probability of full payment is very far from the mean, std dev values in other columns of our dataset. Here our clustering algorithm may treat the features with different weightage and may give a nonoptimal output.

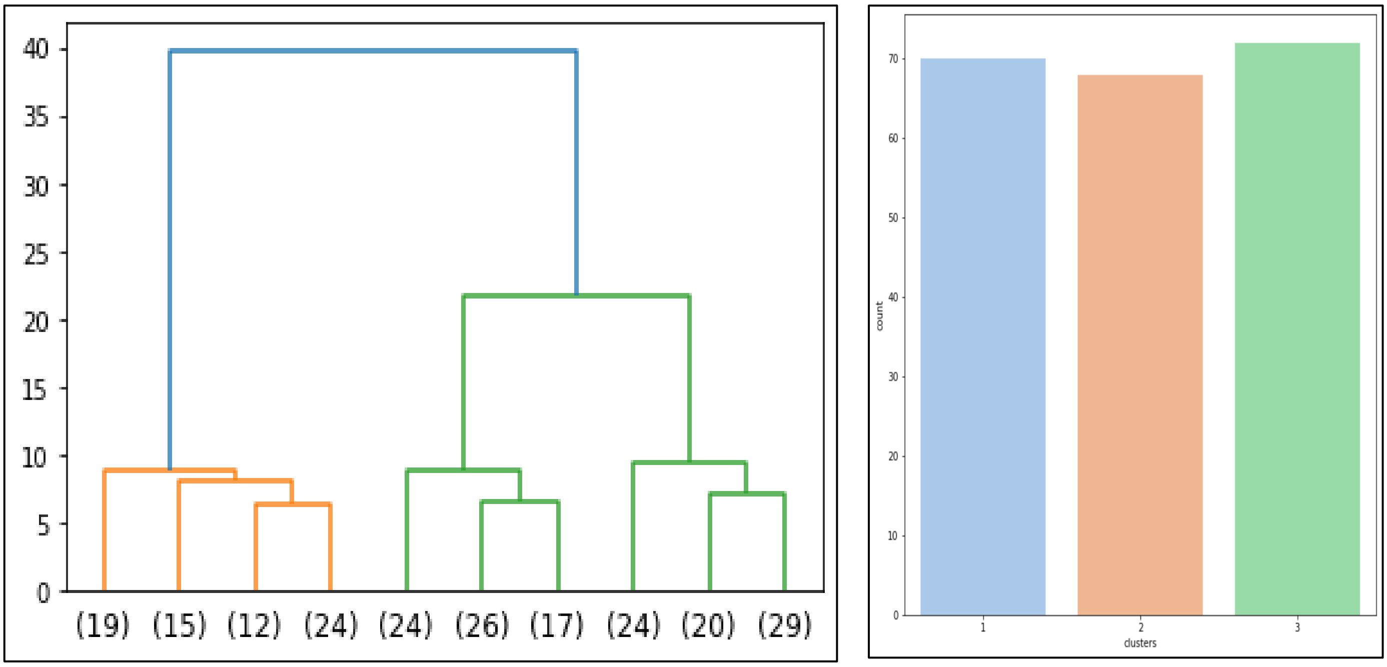


```
array([[ 1.75435461,  1.81196782,  0.17822987, ...,  1.33857863,
        -0.29880602,  2.3289982 ],
       [ 0.39358228,  0.25383997,  1.501773  , ...,  0.85823561,
        -0.24280501, -0.53858174],
       [ 1.41330028,  1.42819249,  0.50487353, ...,  1.317348  ,
        -0.22147129,  1.50910692],
       ...,
       [-0.2816364 , -0.30647202,  0.36488339, ..., -0.15287318,
        -1.3221578 , -0.83023461],
       [ 0.43836719,  0.33827054,  1.23027698, ...,  0.60081421,
        -0.95348449,  0.07123789],
       [ 0.24889256,  0.45340314, -0.77624835, ..., -0.07325831,
        -0.70681338,  0.96047321]])
```

1.2.2. Scaled data

Therefore, it is required to use the scaled data in our clustering algorithms to get optimum output. Standard scaler can be used to scale our dataset by applying the z-score to our datapoints.

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



1.3.1. Dendrogram, Clusters (1,2,3)

In the dendrogram locate the largest vertical difference between the nodes, and in the middle pass a horizontal line. From the above dendrogram we can see there are three clusters, and the green cluster is having more entries than the orange cluster.

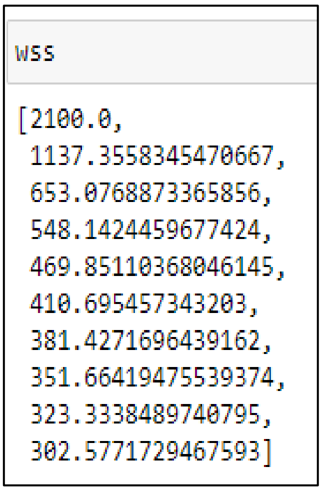
By analysing our data after clustering we found that there are three segments of customers,

i. The customers in cluster-1 are spending from Rs.15,380 to a maximum of Rs.21,180. It is 33.3% of the customers.

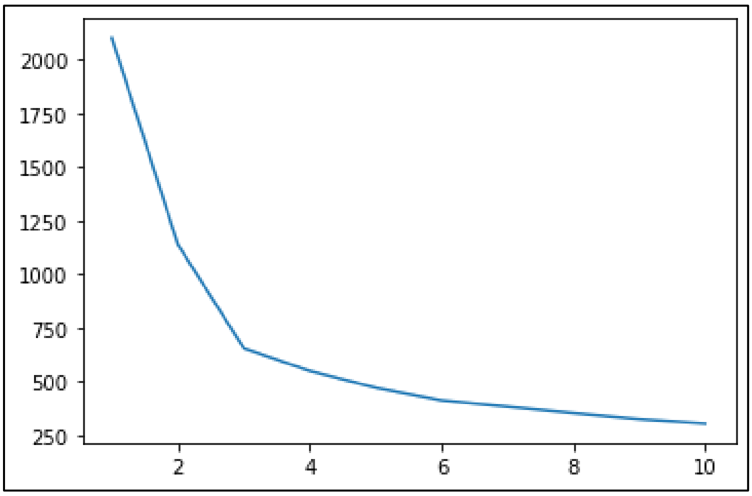
ii. The customers in cluster-2 are spending from Rs.12,080 to a maximum of Rs.16,630. It is 32.4% of the customers.

iii. The customers in cluster-3 are spending from Rs.10,590 to a maximum of Rs.13,340. It is 34.3% of the customers.

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.



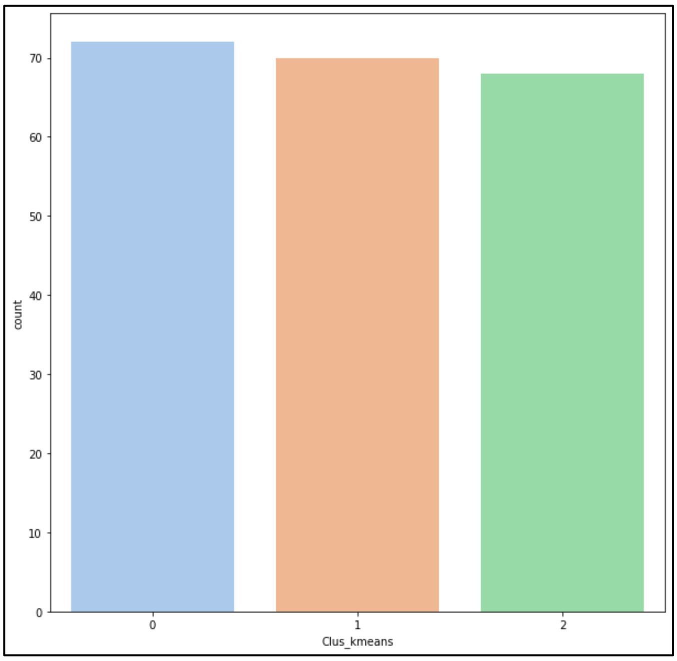
```
WSS

[2100.0,
 1137.3558345470667,
 653.0768873365856,
 548.142445967 7424,
 469.85110368046145,
 410.695457343203,
 381.427169643 9162,
 351.66419475539374,
 323.3338489740795,
 302.5771729467593]
```



1.4.1. WSS, Elbow curve

There is no significant difference in the WSS value after the elbow point 3 (no. of clusters). From this it is decided that the optimum number of clusters is three. The silhouette score is 0.48, this score is on the positive side.

1.4.2. K-means Cluster



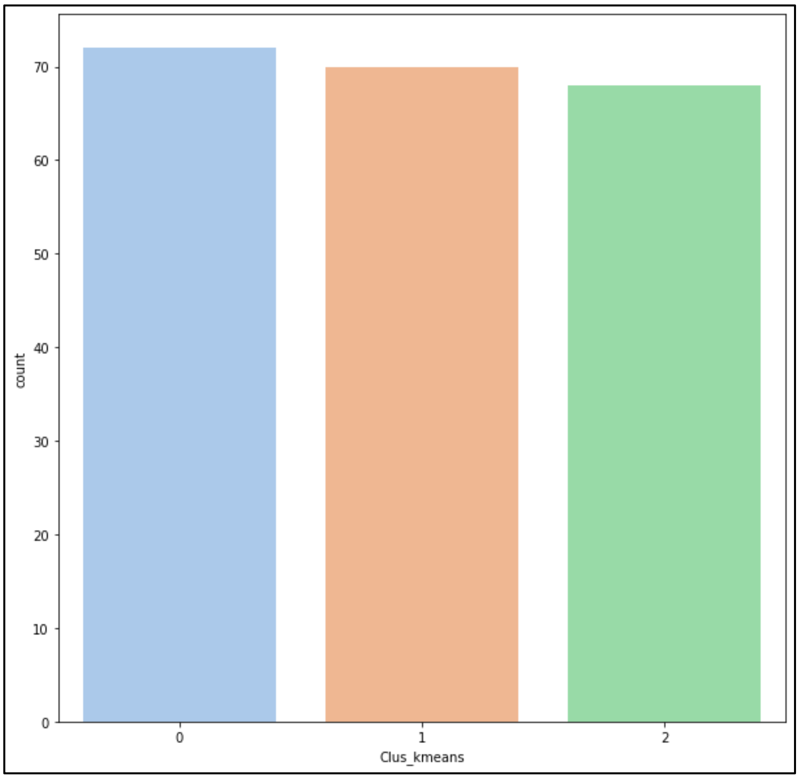
By analysing our data after clustering we found that there are three segments of customers,

i. The customers in cluster-0 are spending from Rs.10,590 to a maximum of Rs.13,340. It is 34.3% of the customers.

ii. The customers in cluster-1 are spending from Rs.15,380 to a maximum of Rs.21,180. It is 33.3% of the customers.

iii. The customers in cluster-2 are spending from Rs.12,080 to a maximum of Rs.16,630. It is 32.4% of the customers.

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



1.5.1.K-means cluster

The below recommendations are for cluster profiles created using the K-Means clustering,

- For the customers in cluster-0. Since they are the least spending and majority (34.3%) group of customers. Discount offers on the purchases can be introduced to encourage them to spend more.

- For the customers in cluster-1. These customers are the most spending and the second major (33.3%) group in our bank. They can be provided with loyalty points for each transaction they make using our credit card. This may encourage them to spend even more than their current usage.

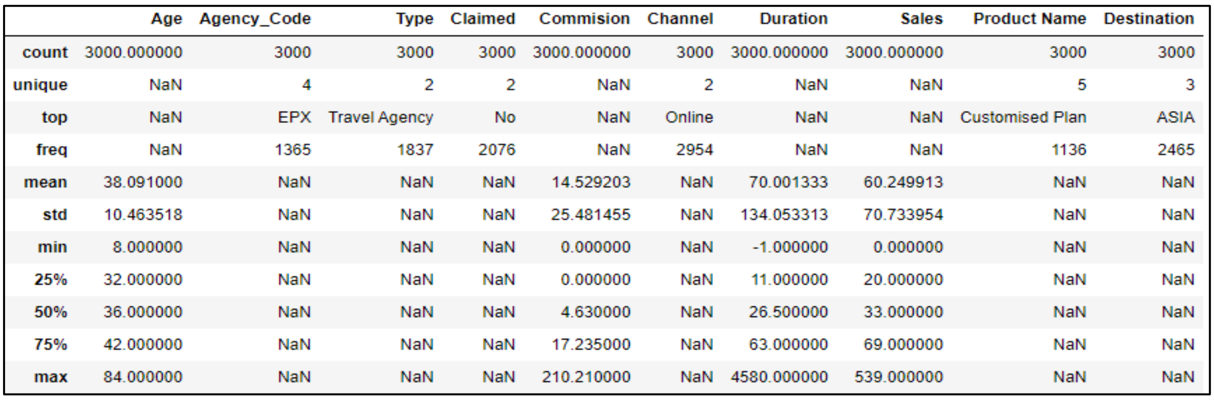
- The loyalty points can be used to make the customers to buy new products and reduce a certain percentage as a discount during the checkout.

- For the customers in cluster-2. These customers are in the centre of our spending customers. They need to be observed further to provide a optimum solution for encouraging them to spend more.

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.

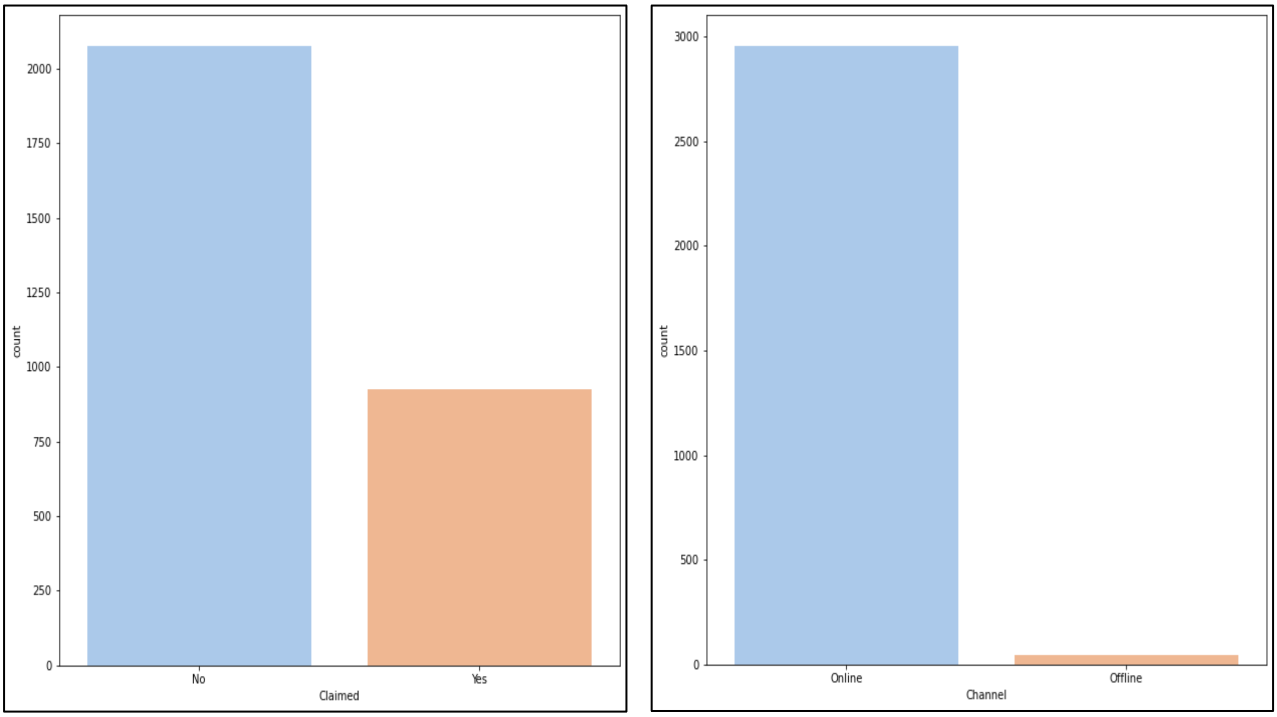
2.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.000000	3000	3000	3000	3000.000000	3000	3000.000000	3000.000000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	2954	NaN	NaN	1136	2465
mean	38.091000	NaN	NaN	NaN	14.529203	NaN	70.001333	60.249913	NaN	NaN
std	10.463518	NaN	NaN	NaN	25.481455	NaN	134.053313	70.733954	NaN	NaN
min	8.000000	NaN	NaN	NaN	0.000000	NaN	-1.000000	0.000000	NaN	NaN
25%	32.000000	NaN	NaN	NaN	0.000000	NaN	11.000000	20.000000	NaN	NaN
50%	36.000000	NaN	NaN	NaN	4.630000	NaN	26.500000	33.000000	NaN	NaN
75%	42.000000	NaN	NaN	NaN	17.235000	NaN	63.000000	69.000000	NaN	NaN
max	84.000000	NaN	NaN	NaN	210.210000	NaN	4580.000000	539.000000	NaN	NaN

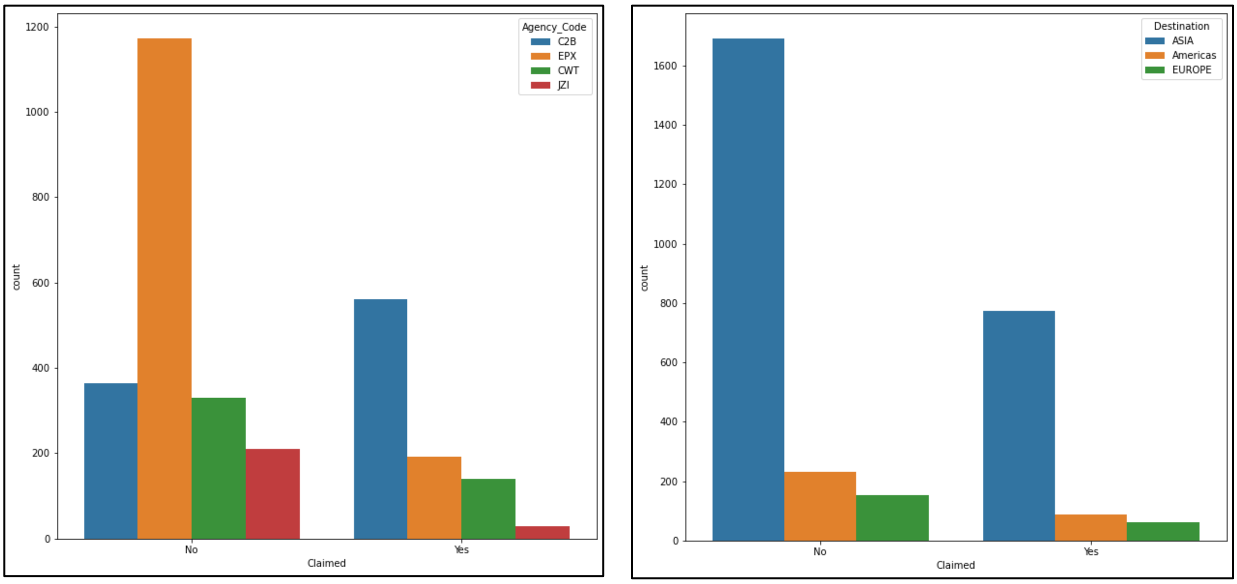
2.1.1. Data summary

From the above summary, the mean age of the insured is 38, and there are two distribution channel of tour insurance agencies, Travel agency type of tour insurance firm is selling more insurance.



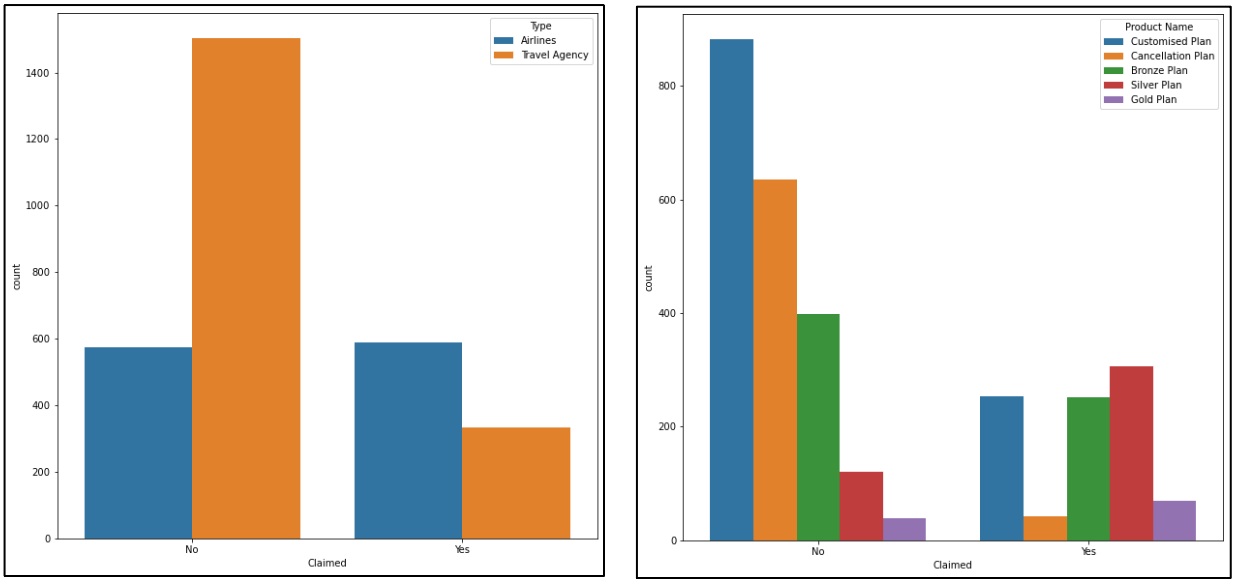
2.1.2. Claimed, Channel bargraph

The percentage of the insured who have claimed the insurance is 30.8% and the insured who did not claim the insurance is 69.2%. Most customers are preferring the online distribution channel of tour insurance agencies.



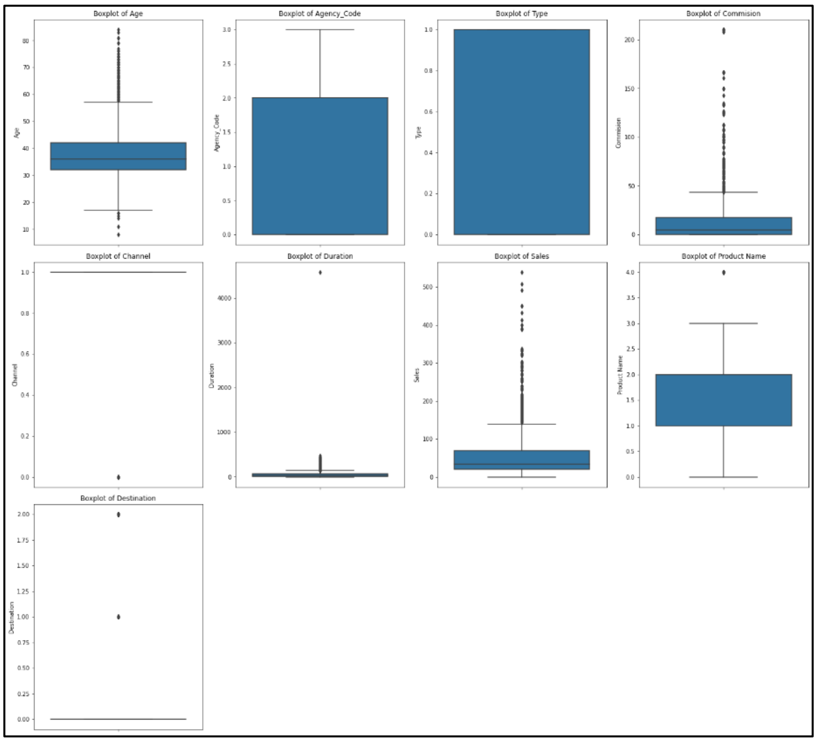
2.1.3. Claimed as per Agency code and destination

The most insurance claims are made by insured from C2B tour firm and from Asia. Least claim from JZI tour firm and Europe destination.



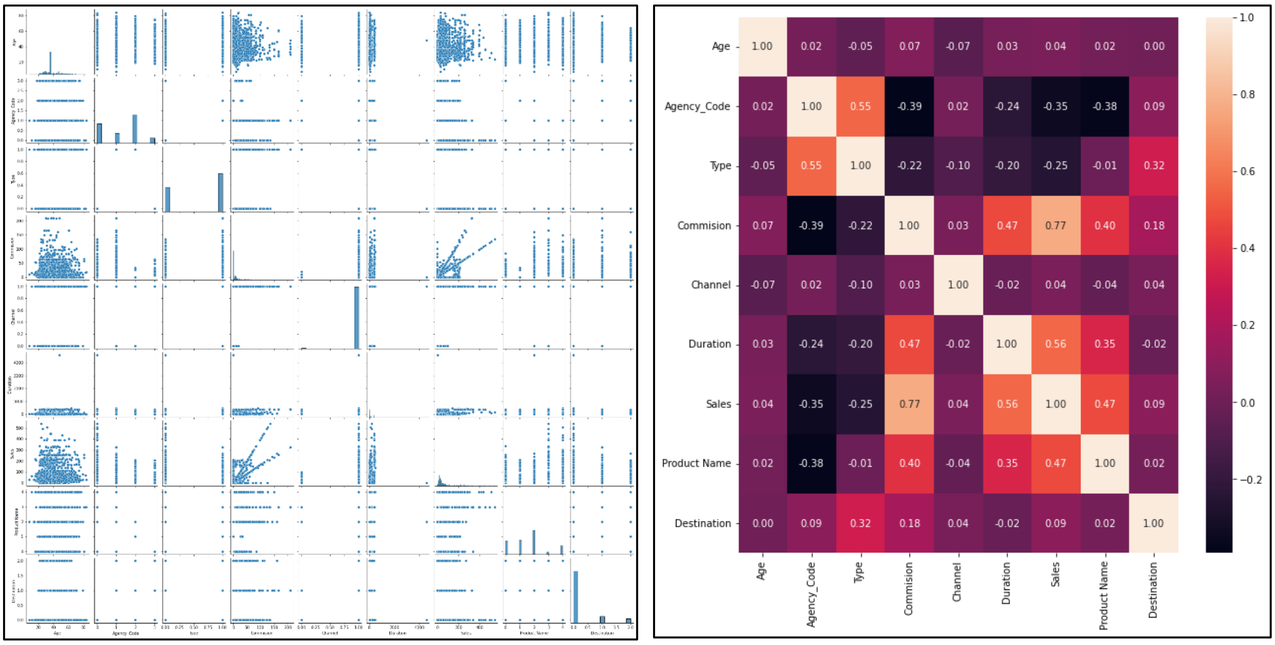
2.1.4. Claimed as per Type and Product name

The most insurance claims are made from insured by airline type of tour insurance firms and the silver plan of the tour insurance product. Least claim from the travel agency and Cancellation plan of the tour insurance product.



2.1.5. Boxplot of data

Outliers are present in Age, Commission, Channel, Duration, Sales, Product name, Destination. And data is skewed in age, agency code, commission, duration, sales, product name.



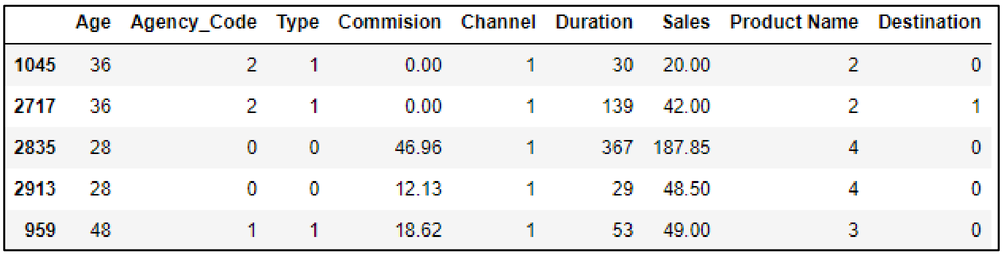
2.1.6.Pairplot, correlation matrix

There is a strong correlation between the sales and commission values of dataset.

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

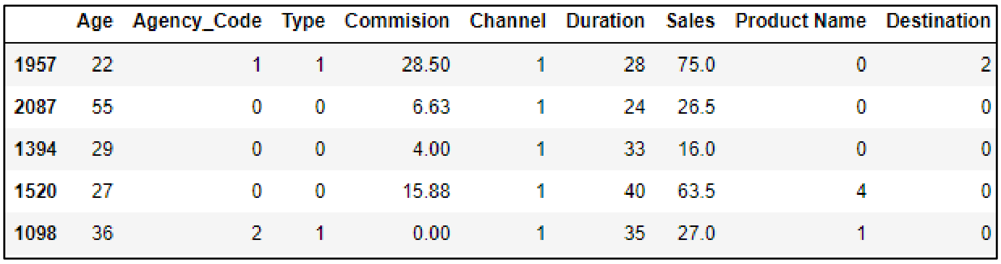
The dataset is having some object datatypes. They are converted to categorical and numerical datatypes. The train and test data are split using train_test_split, random_state=1 is used.

Table 2.2.1. X_train data,



	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
1045	36	2	1	0.00	1	30	20.00	2	0
2717	36	2	1	0.00	1	139	42.00	2	1
2835	28	0	0	46.96	1	367	187.85	4	0
2913	28	0	0	12.13	1	29	48.50	4	0
959	48	1	1	18.62	1	53	49.00	3	0

Table 2.2.2. X_test data,



	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
1957	22	1	1	28.50	1	28	75.0	0	2
2087	55	0	0	6.63	1	24	26.5	0	0
1394	29	0	0	4.00	1	33	16.0	0	0
1520	27	0	0	15.88	1	40	63.5	4	0
1098	36	2	1	0.00	1	35	27.0	1	0

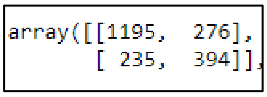
2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

CART Performance Metrics

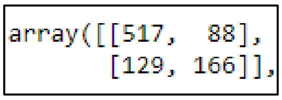
Accuracy: 75.6 (Train data)

75.8 (Test data)

Confusion matrix: Train data Test data

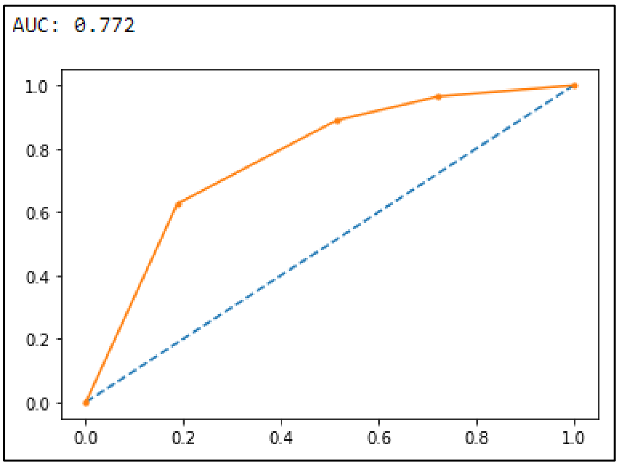


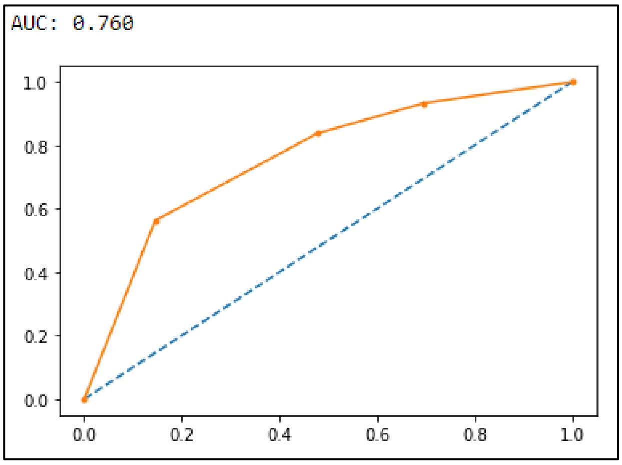
```
array([[1195,  276],
       [ 235,  394]],
```



```
array([[517,  88],
       [129, 166]],
```

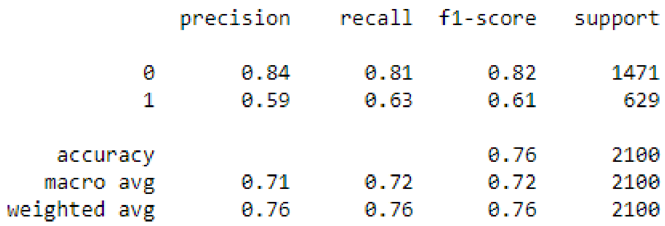
ROC Curve, AUC: Train data Test data





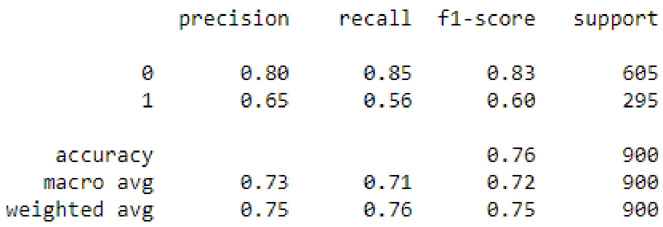
Classification reports:

Training data



```
              precision    recall  f1-score   support

           0       0.84      0.81      0.82      1471
           1       0.59      0.63      0.61       629

    accuracy                           0.76      2100
   macro avg       0.71      0.72      0.72      2100
weighted avg       0.76      0.76      0.76      2100
```

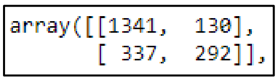
Testing data



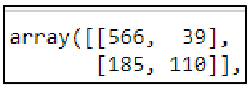
```
              precision    recall  f1-score   support

           0       0.80      0.85      0.83       605
           1       0.65      0.56      0.60       295

    accuracy                           0.76       900
   macro avg       0.73      0.71      0.72       900
weighted avg       0.75      0.76      0.75       900
```

Random Forest Performance Metrics

Confusion matrix: Train data Test data

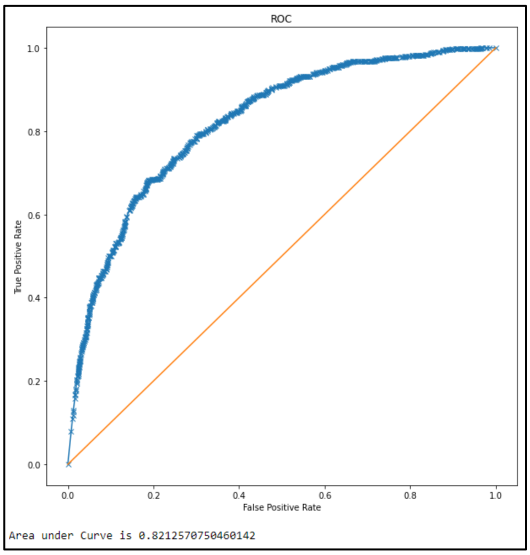


```
array([[1341,  130],
       [ 337,  292]],
```

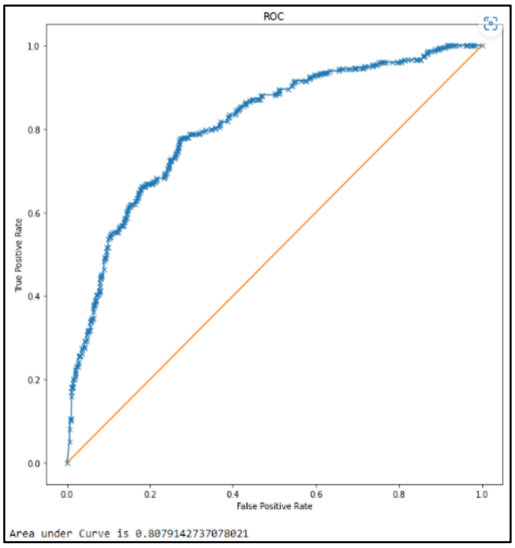


```
array([[566,  39],
       [185, 110]],
```

ROC Curve, AUC: Train data Test data



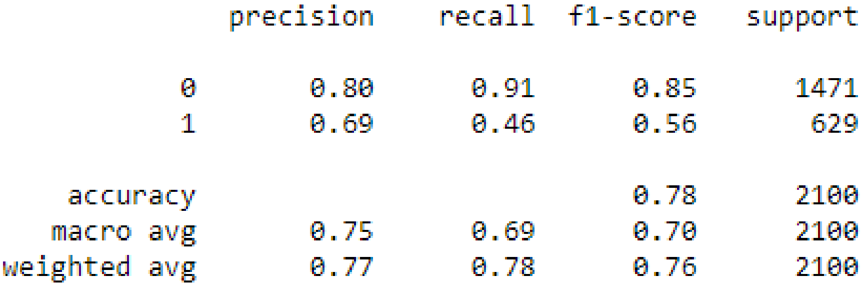
```
Area under Curve is 0.8212570750460142
```



```
Area under Curve is 0.8079142737078021
```

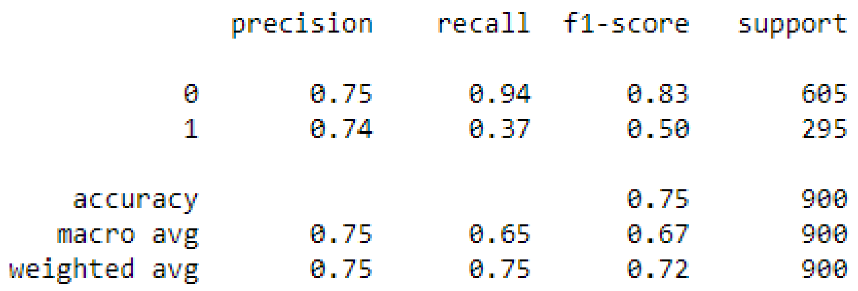
Classification reports:

Training data:



```
              precision    recall  f1-score   support

           0       0.80      0.91      0.85      1471
           1       0.69      0.46      0.56       629

    accuracy                           0.78      2100
   macro avg       0.75      0.69      0.70      2100
weighted avg       0.77      0.78      0.76      2100
```

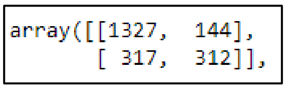
Testing data:



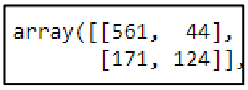
```
              precision    recall  f1-score   support

           0       0.75      0.94      0.83       605
           1       0.74      0.37      0.50       295

    accuracy                           0.75       900
   macro avg       0.75      0.65      0.67       900
weighted avg       0.75      0.75      0.72       900
```

Artificial Neural Network Performance Metrics

Confusion matrix: Train data Test data

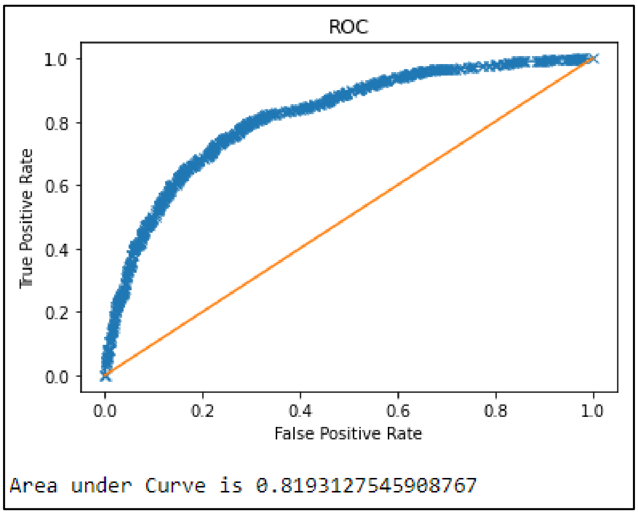


```
array([[1327,  144],
       [ 317,  312]],
```

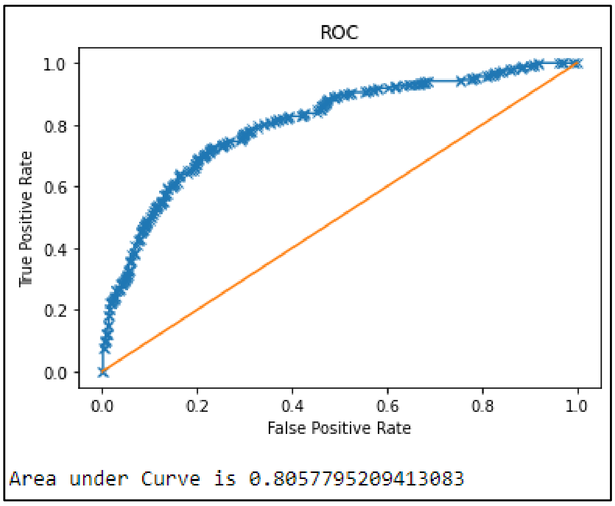


```
array([[561,  44],
       [171, 124]],
```

ROC Curve, AUC: Train data Test data

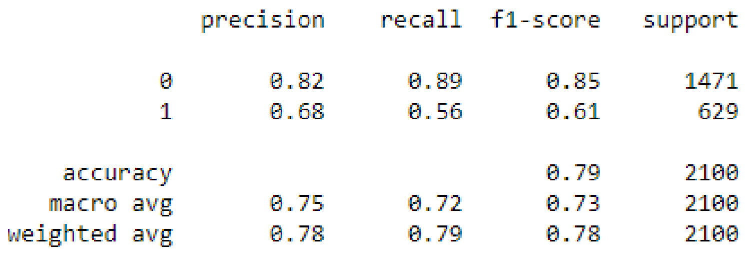


```
Area under Curve is 0.8193127545908767
```



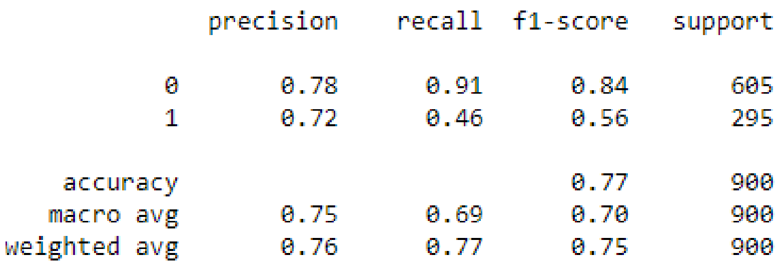
```
Area under Curve is 0.8057795209413083
```

Classification reports: Training data:



```
              precision    recall  f1-score   support

           0       0.82      0.89      0.85      1471
           1       0.68      0.56      0.61       629

    accuracy                           0.79      2100
   macro avg       0.75      0.72      0.73      2100
weighted avg       0.78      0.79      0.78      2100
```

Testing data:



```
              precision    recall  f1-score   support

           0       0.78      0.91      0.84       605
           1       0.72      0.46      0.56       295

    accuracy                           0.77       900
   macro avg       0.75      0.69      0.70       900
weighted avg       0.76      0.77      0.75       900
```

Precision is the ability of a classifier not to label an instance positive that is actually negative.

Recall is the ability of a classifier to find all positive instances.

The weighted average of F1 should be used to compare classifier models.

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

Based on the Precision, Recall, F1 score we can conclude that the CART model **is** best suited for our prediction.

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

The insured claims are higher in air lines and Asia destination. The insurance company need to investigate the reason for these claim requests.

The company may also consider making changes in its silver plan as it also have high claim rates.

The C2B agency need to make changes in the criteria to give insurance to the customers. This may help to reduce future claim from the insured.