DATA MINING

Jyoti Ranjan Padhiary

Contents

DATA MINING

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).

```
bank.head()
```

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

Figure 1

Figure 1 shows the data frame of banking dataset.

```
bank.describe(include='all')
```

	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

Figure 2

Figure 2 shows the description of the data.

```
bank.info()
```

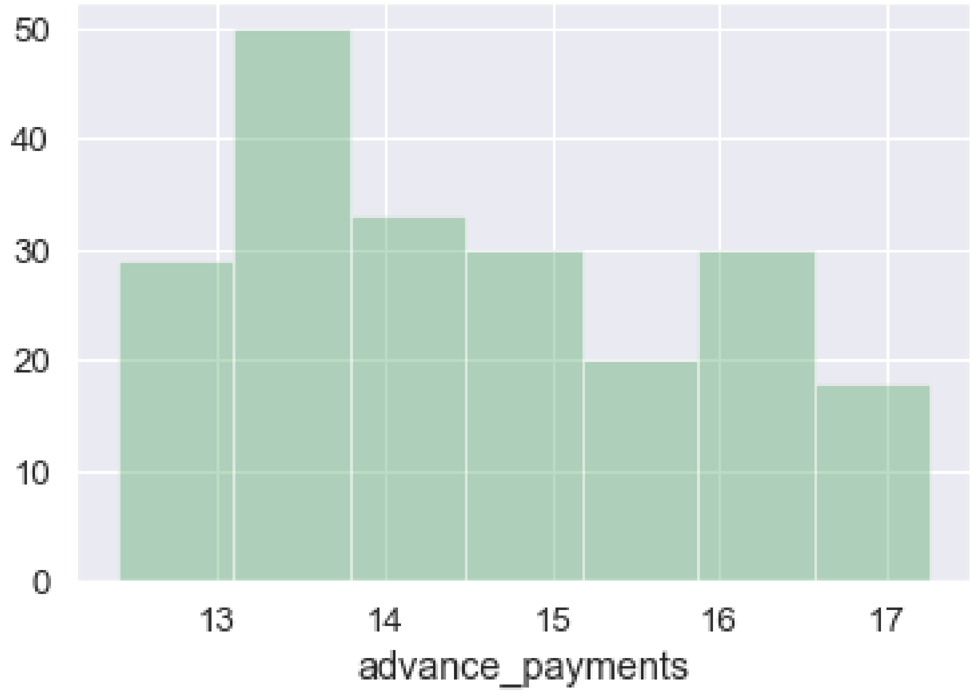
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              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
```

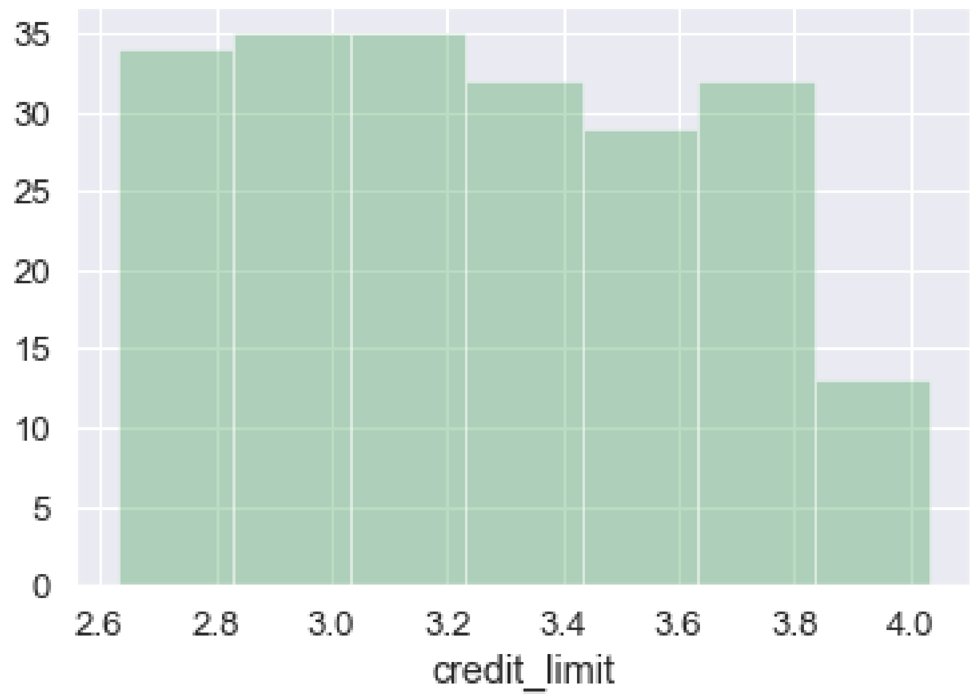
Figure 3

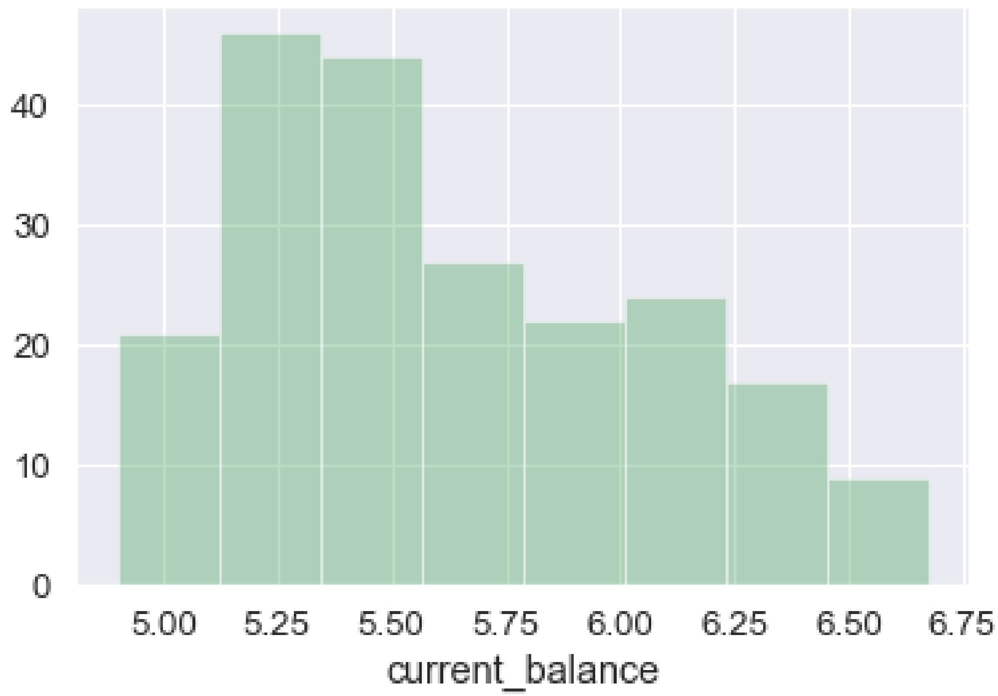
Figure 3 shows the information of the dataset, there are total 7 columns and 210 rows in the dataset. All variables are of float 64 data type. There are no duplicate data or missing values present.

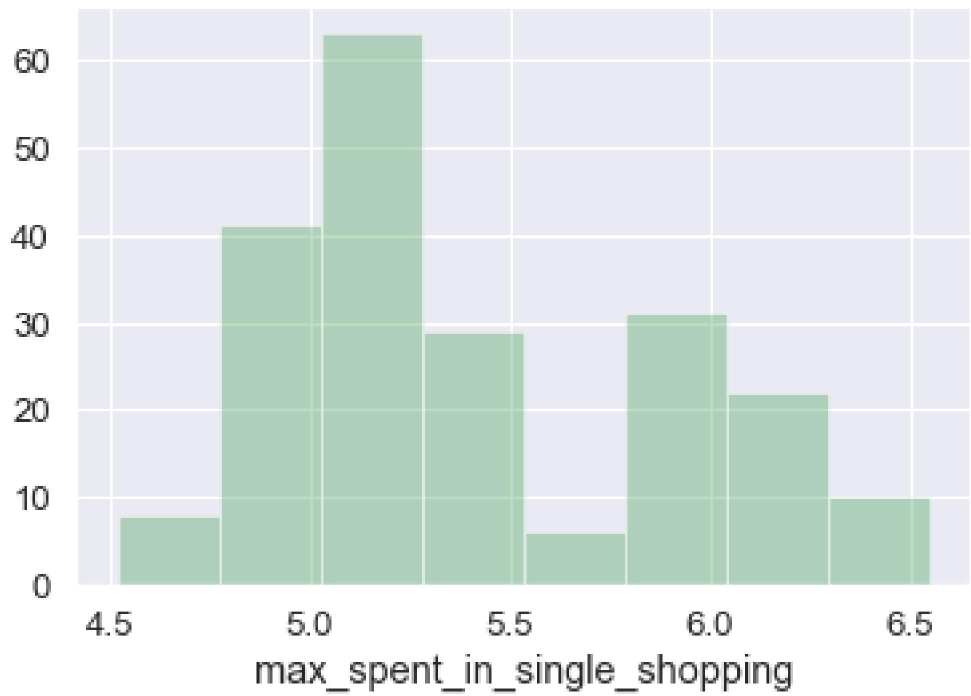
Univariate Analysis

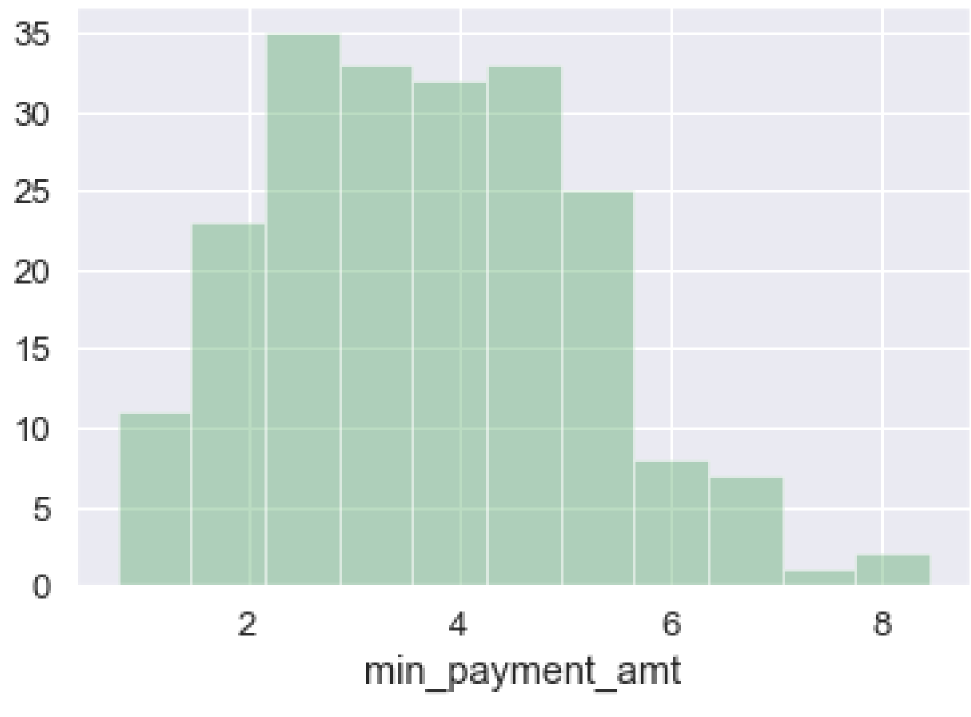
Distribution Plots

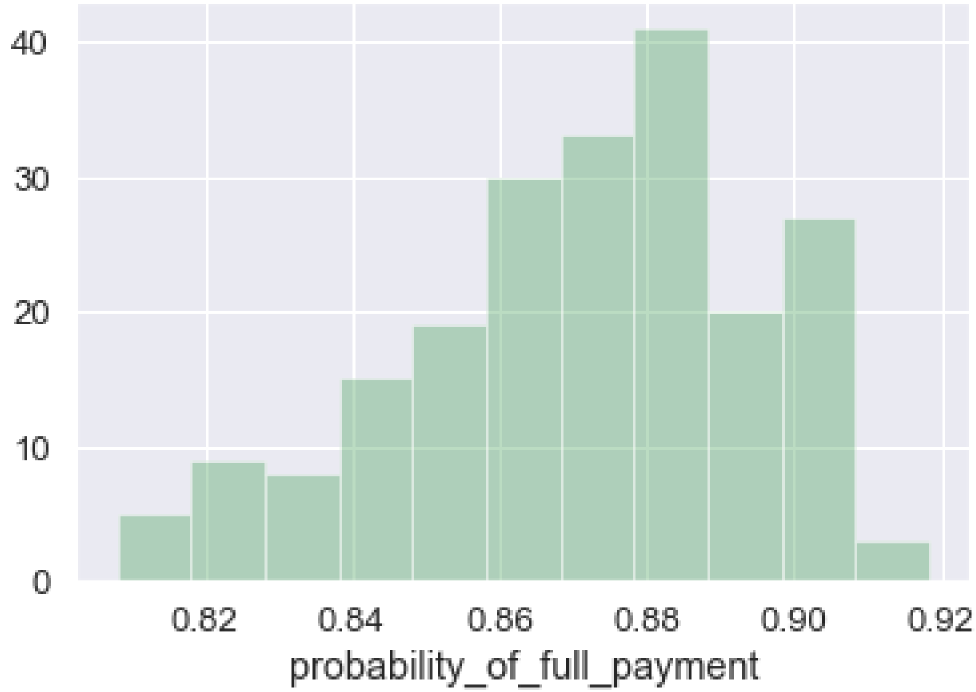


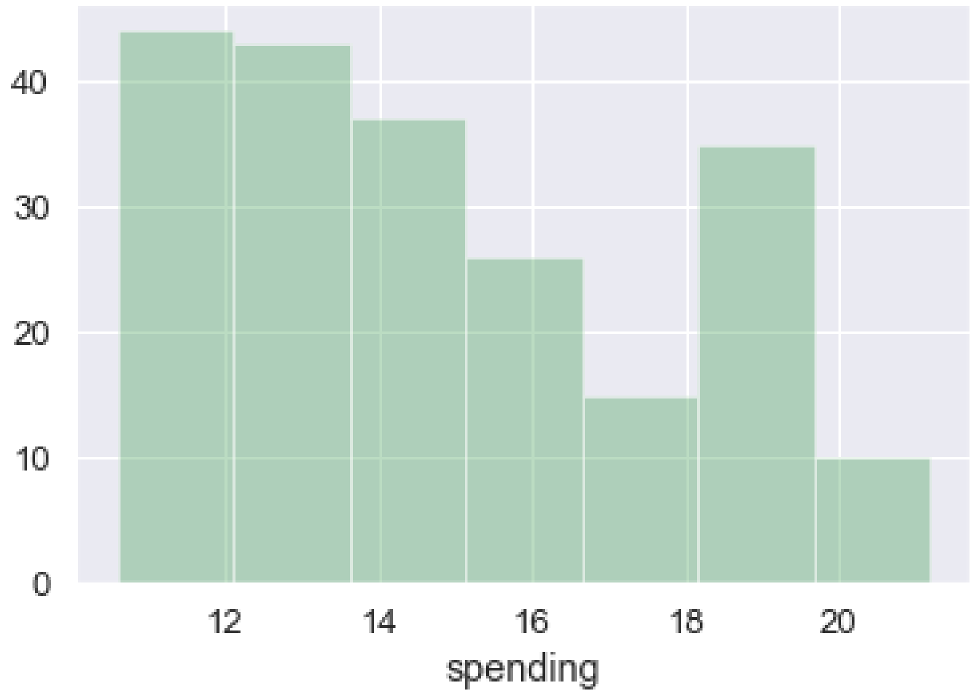

35
30
25
20
15
10
5
0
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
credit_limit


40
30
20
10
0
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
current_balance


60
50
40
30
20
10
0
4.5
5.0
5.5
6.0
6.5
max_spent_in_single_shopping


35
30
25
20
15
10
5
0
2
4
6
8
min_payment_amt





Inference of distribution plots of all variables are as below:

Advance payment - it is observed that most of the people have made advance payment of amount 13.5k

Credit limit – this is more or less equally distributed and equal no of data points in each credit limit level is there.

Current balance – this is positively skewed and most of the current balances are in the range of 5.25k to 5.75k

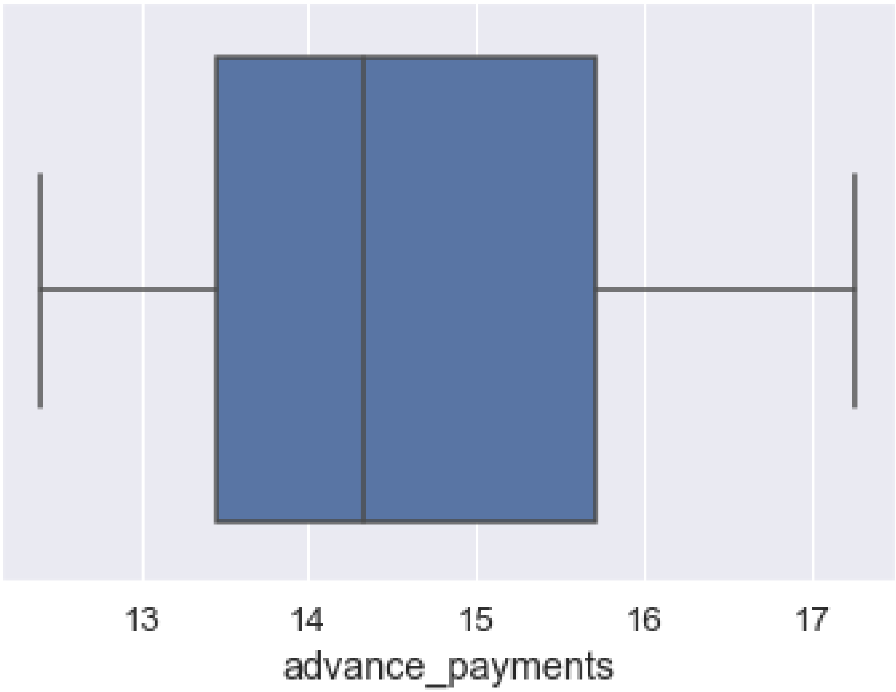
Max spend in single shopping – it is positively skewed and people have spent max in the range of 4.75k to 5.5k in single shopping.

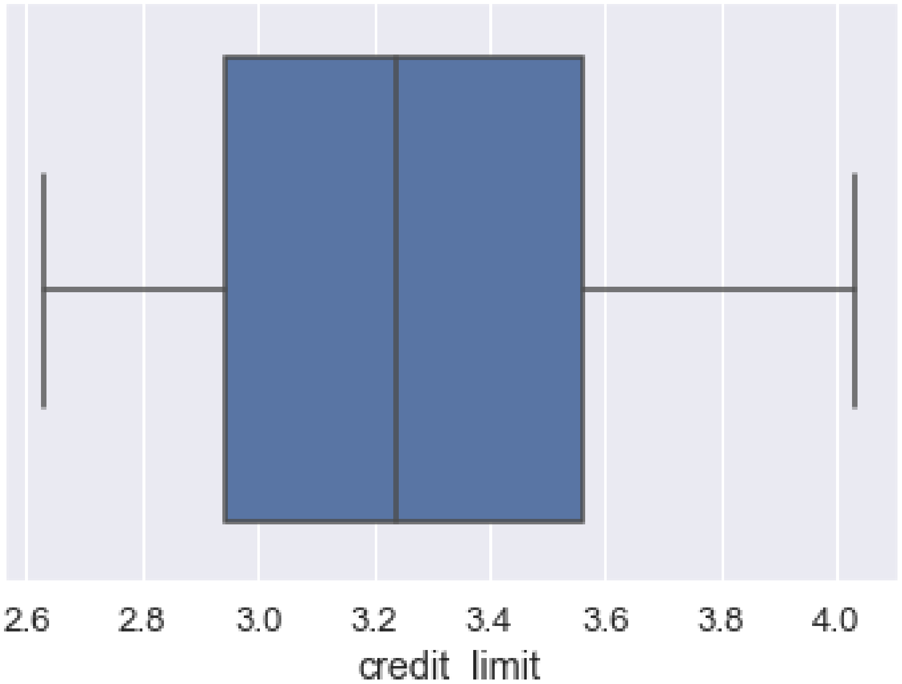
Min payment amount – it can be observed that the min payment amount by most of the people is Rs. 200 to Rs. 5500

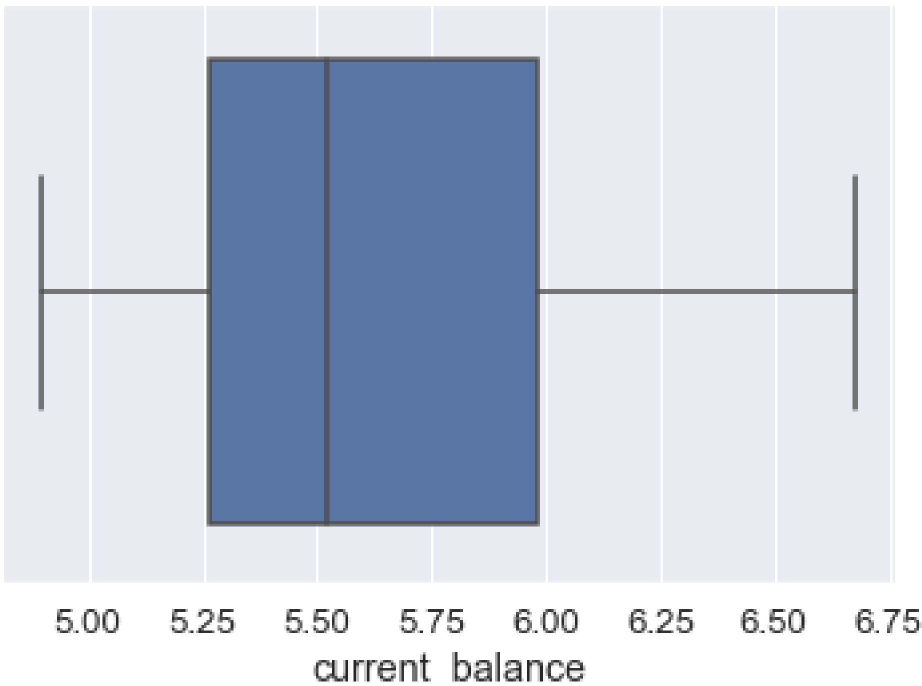
Probability of full payment – this is negatively skewed and there is a probability of 86% to 90% that people will pay full amount.

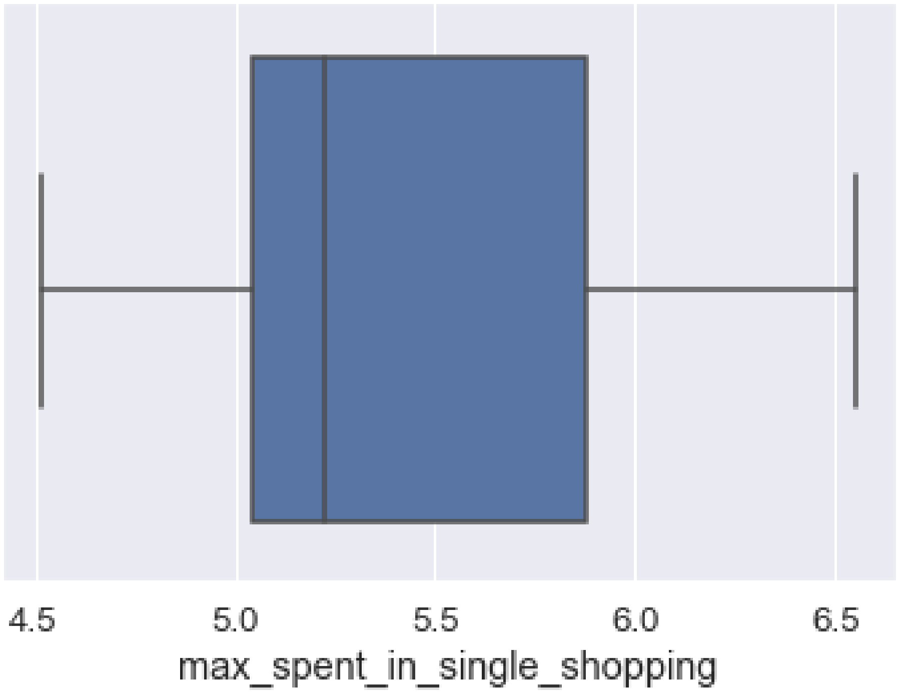
Spending - Distribution is not normal, the data is skewed towards right and most of spending is in between 12k to 14k.

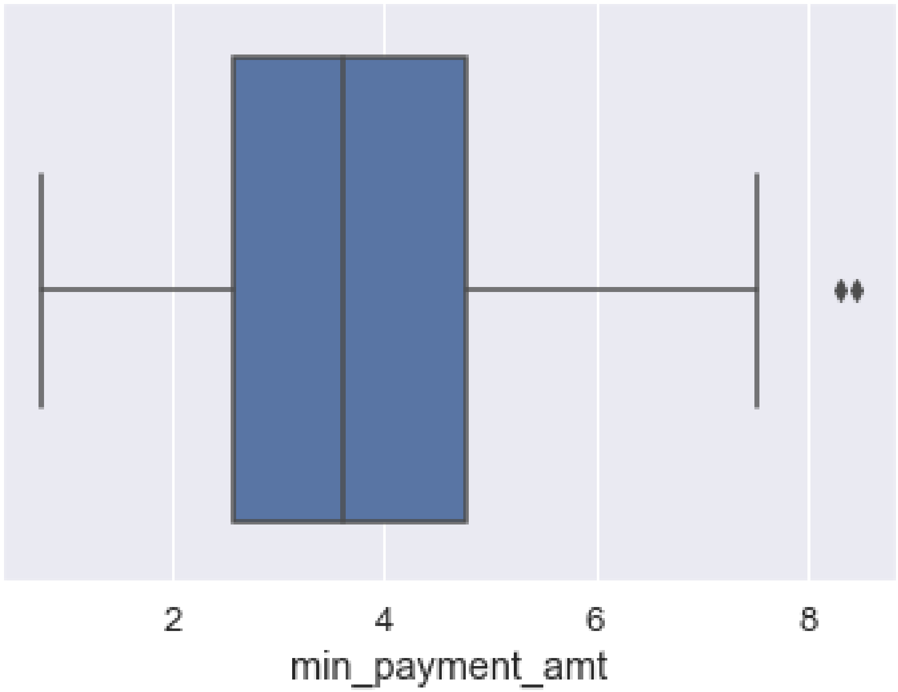
Box Plots

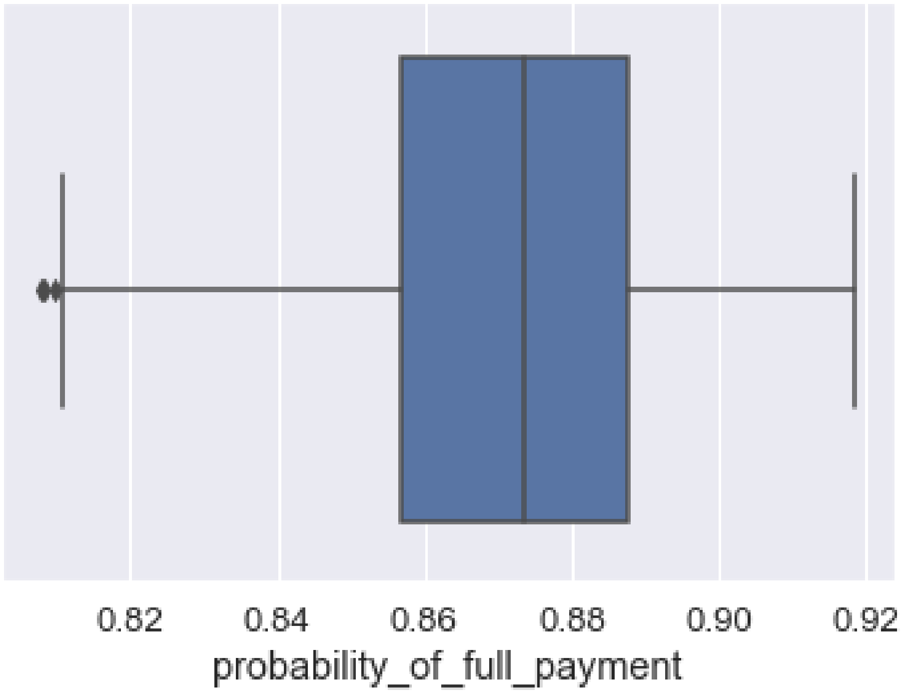


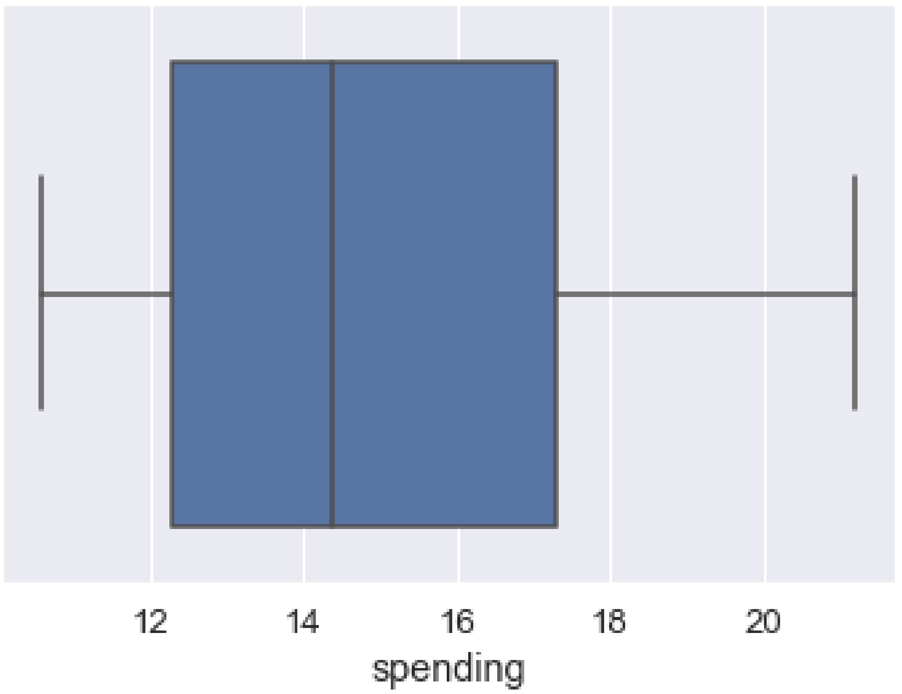

2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
credit_limit


5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
current_balance


4.5
5.0
5.5
6.0
6.5
max_spent_in_single_shopping


2
4
6
8
min_payment_amt





Inferences of box plots of each variable are as below:

Skewness -

> Positively skewed – the following variables are positively skewed Spending, advance payment, current balance, max spent in single shopping
>
> Negatively skewed – Probability of full payment is negatively skewed

Normal distribution – No variable is normally distributed

Outliers – min payment amount is having outlier.

Multivariate analysis

Correlation plot

Figure 4

Inference: from fig 4 it can be observed that there are some correlations exist in the data set between the variables which are (advance payment & current balance), (advance payment & credit limit), (credit limit & spending), (advance payment & spending)

Pair plot

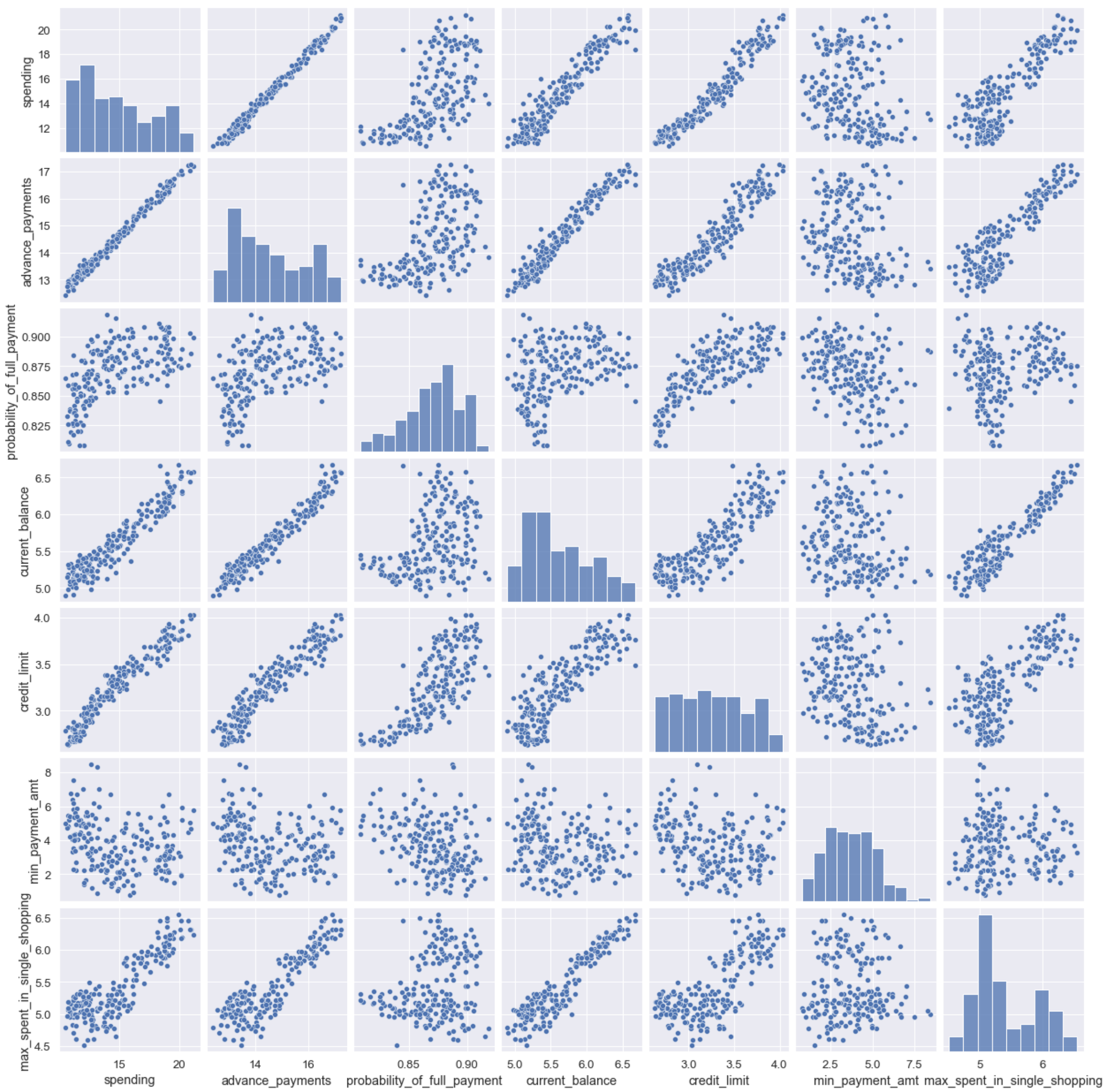


Figure 5

Inference - from fig 5 it can be observed that many variables are linearly correlated to each other.

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

```
bank.std()
```

```
spending                        2.909699
advance_payments                1.305959
probability_of_full_payment     0.023629
current_balance                 0.443063
credit_limit                    0.377714
min_payment_amt                 1.503557
max_spent_in_single_shopping    0.491480
dtype: float64
```

Figure 6

```
bank.var()
```

```
spending                        8.466351
advance_payments                1.705528
probability_of_full_payment     0.000558
current_balance                 0.196305
credit_limit                    0.142668
min_payment_amt                 2.260684
max_spent_in_single_shopping    0.241553
dtype: float64
```

Figure 7

From fig 6 and 7 it is observed that there is high standard deviation and high variance exist in the variables, also certain instances where the difference between min value and max value of the variables are very high. So scaling is necessary here.

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

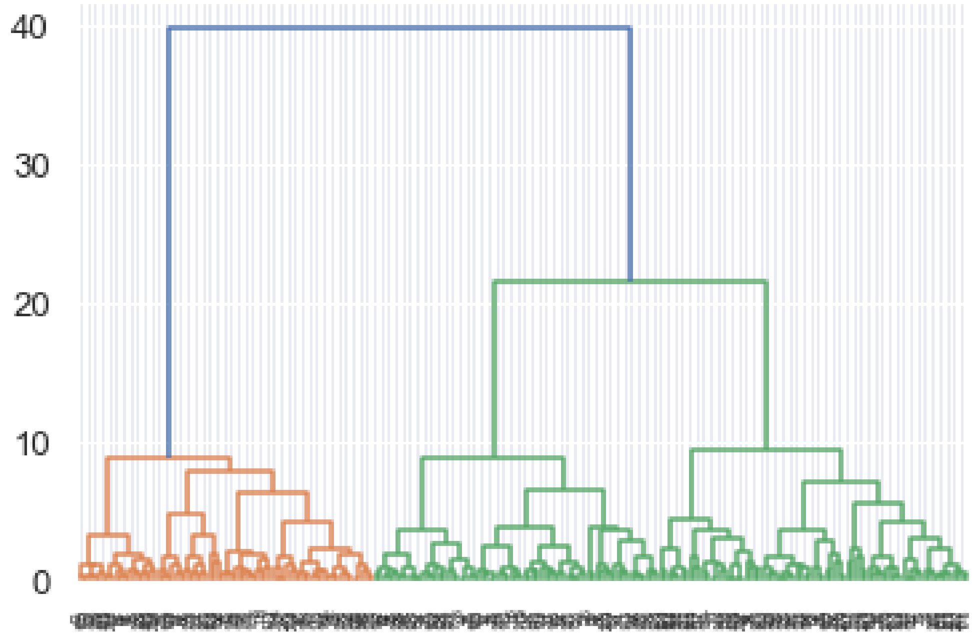


Figure 8

Fig 8 represents there are 3 clusters in the dataset, let us visualize them clearly by looking at the truncated dendogram.

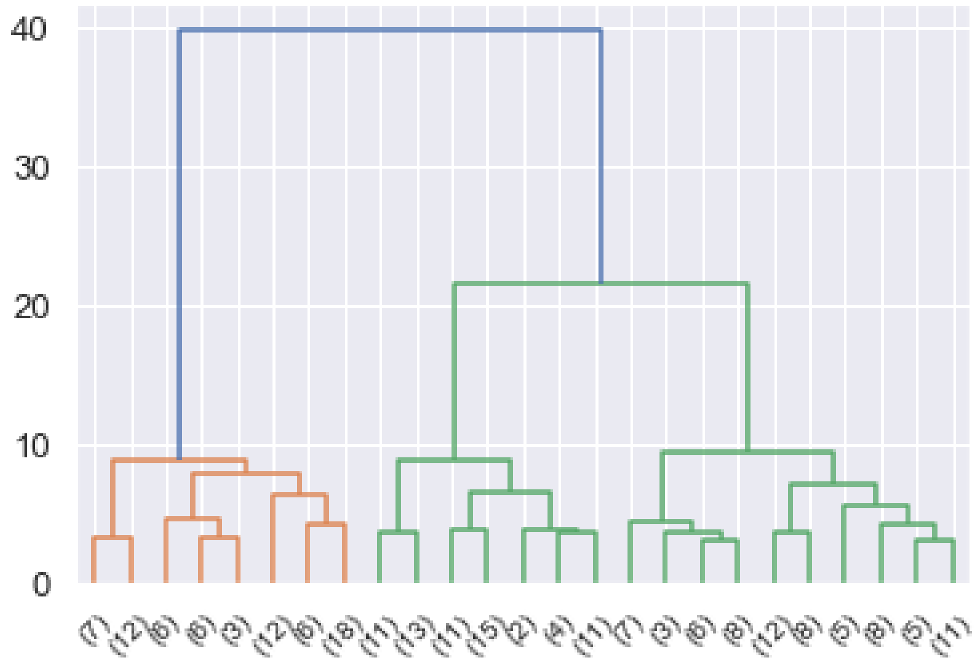


Figure 9

Fig 9 shows the truncated dendogram.

We will now include these clusters in the data frame and display below.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

Figure 10

clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	70
2	11.872388	13.257015	0.848072	5.238940	2.848537	4.949433	5.122209	67
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178	73

Figure 11

Inference – fig 11 describes the mean value of each cluster profile in each variable,

Cluster 1 – Highest values in all the parameters except min payment amount this class can be called the upper class consumers.

Cluster 2 – lowest values in all the parameters except min payment amount where it is highest and 5.12k is max amount spent in single shopping, so this can be called the lower class consumers.

Cluster 3 – the values are in between the cluster 1 and cluster 2 profile and min payment amount, max amount spent in single shopping is lowest so this class can be called the middle class consumers.

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 applied and wss values have been calculated for different no of clusters starting from 1 to 15 . wss values are displayed below in fig 12.

```
wss
```

```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.2941183362073,
 327.300123768427,
 289.46717056412893,
 262.94722991166003,
 241.12014842566663,
 221.6579778431169,
 211.2648737557329,
 195.11638715917454,
 179.1944068689924,
 172.62531020631653,
 164.37420228693057]
```

Figure 12

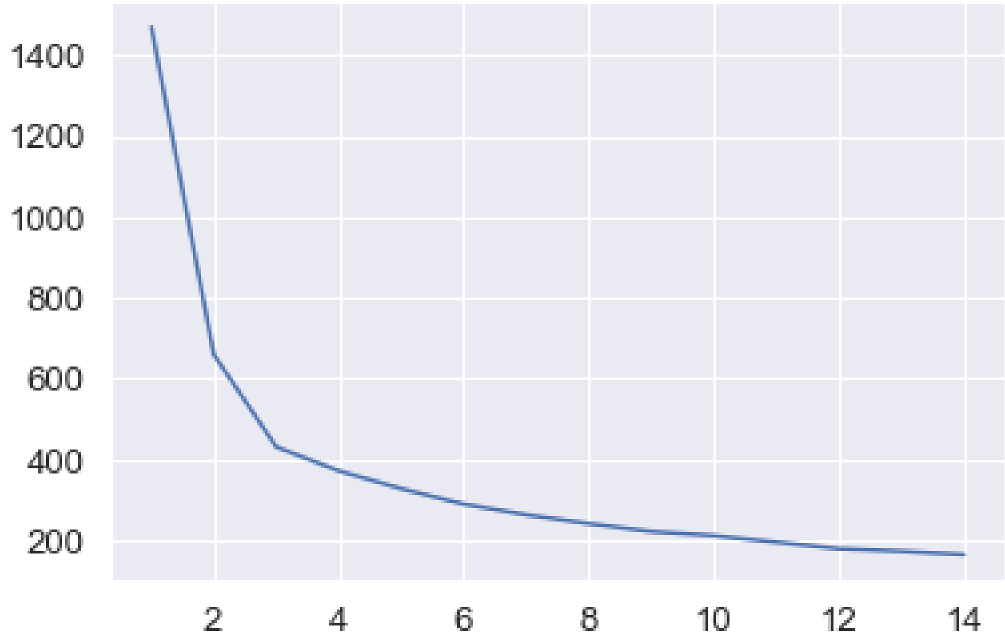


Figure 13

The elbow curve is plotted in fig 13 using the wss values, and from the figure it can be observed that after cluster 3 there is a drastic reduction in the wss score so we will choose 3 clusters now.

We will now add the k-means cluster column in the data frame and display below

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	Clus_kmeans
19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1	1
15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3	2
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1	1
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0
17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1	1

Figure 14

Fig 14 is showing the k-means clusters against each data, we have calculated the silhouette score which is coming 0.400.

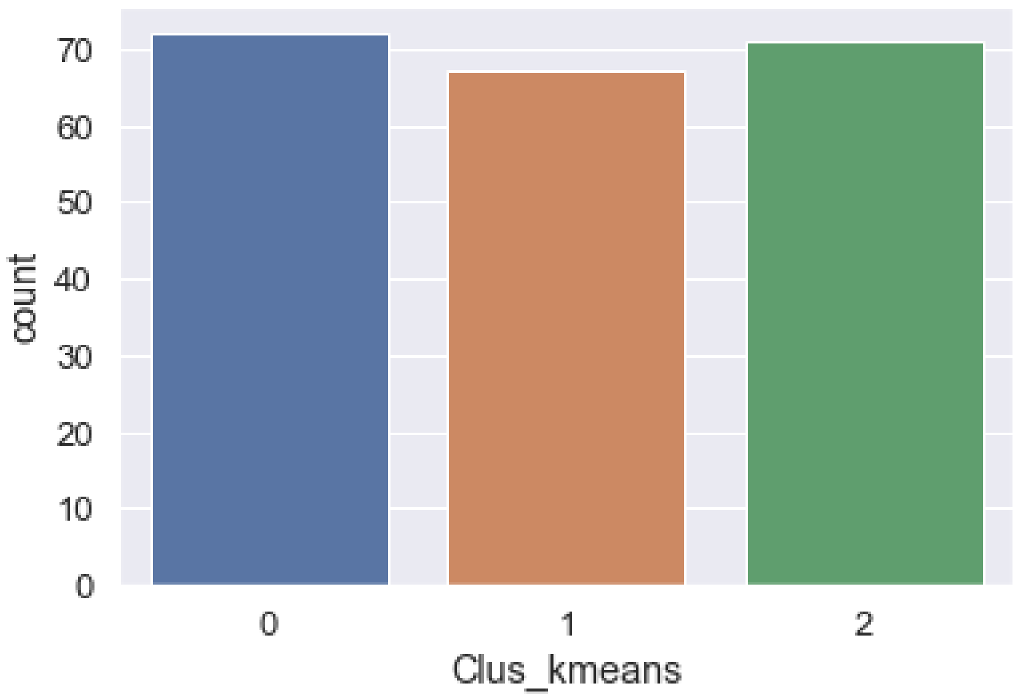


Figure 15

Fig 15 shows that the data set has been divide into 3 clusters 0, 1 and 2

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

Clus_kmeans	spending	advance_payments	probability_of_full_payment	current_balance
0	11.856944	13.247778	0.848253	5.231750
1	18.495373	16.203433	0.884210	6.175687
2	14.437887	14.337746	0.881597	5.514577

Figure 16

credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	Freq
2.849542	4.742389	5.101722	2.083333	NaN
3.697537	3.632373	6.041701	1.029851	70.0
3.259225	2.707341	5.120803	2.873239	67.0

Figure 17

Fig 16 and 17 together showing the cluster profile of the dataset.

Inferences – looking at the different values of each clusters we can now label them as high class, low class and medium class consumers for the bank.

Cluster 1 – high class consumers who are spending high with high credit limit, highest advance payment amount and highest probability of making payment in full.

Cluster 0 – low class consumers who are spending the lowest with lowest current balance in the account but spending more than their level in single shopping.

Cluster 2 - medium class consumers who are spending in between the high class and low class with more current balance than the lower class and least minimum amount of payment.

Promotional strategy:

Cluster 1 (high class consumers) – These are high class consumers who have high current balance in their account and high credit limit, so assuming they are business man we can promote CASA products to them. Along with that working capital loan, high premium LIC for tax saving and general insurance mutual funds are also suitable to promote to this class.

Cluster 0 – (low class consumers) These are low class consumers who have less current balance in their account and they are spending more in single shopping so bank should promote SIP, RD, FD to them. Assuming these are small vendors/shoppers we can also promote personal loan, business loan and MUDRA loan to them.

Cluster 2- (middle class consumers) – These are middle class people and mostly they are employed by any private company so we can promote car loan, home loan and personal loan to them as per their requirement.

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).

```
ins.head()
```

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0.00	Online	4	26.00	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.30	Online	53	18.00	Bronze Plan	ASIA

Figure 18

Fig 18 shows the data set of insurance firm.

	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

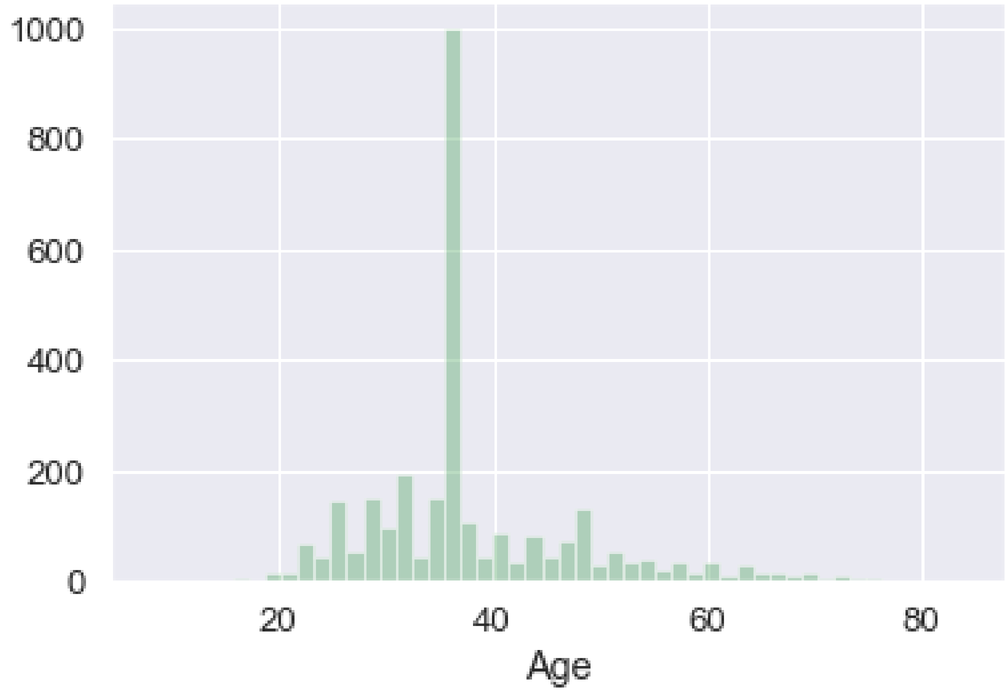
Figure 19

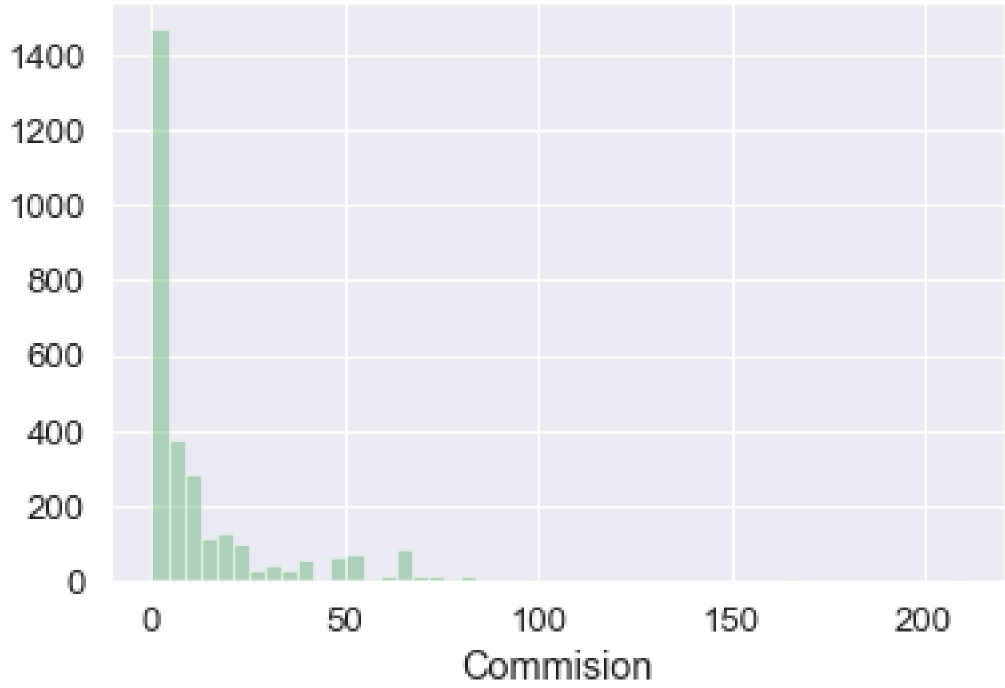
Fig 19 shows us the description of the dataset

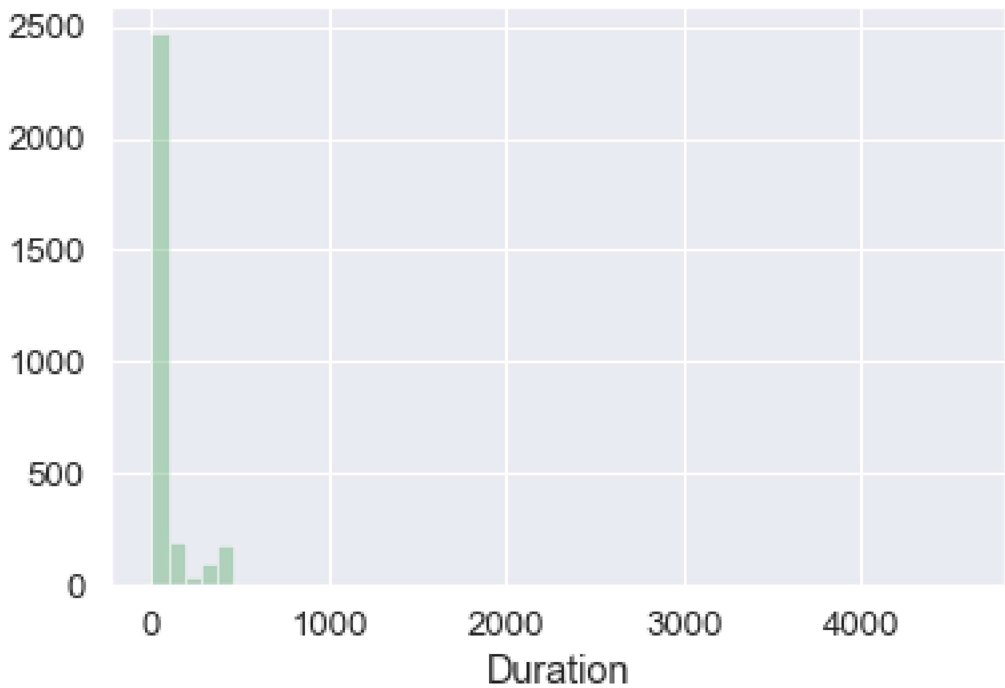
There are 3000 rows and 10 columns. There are some variables of object data type which has been converted in to numeric. There are no missing values. Some duplicate values are there so it is obvious in this case that same package is taken by same demography so we won’t impute them.

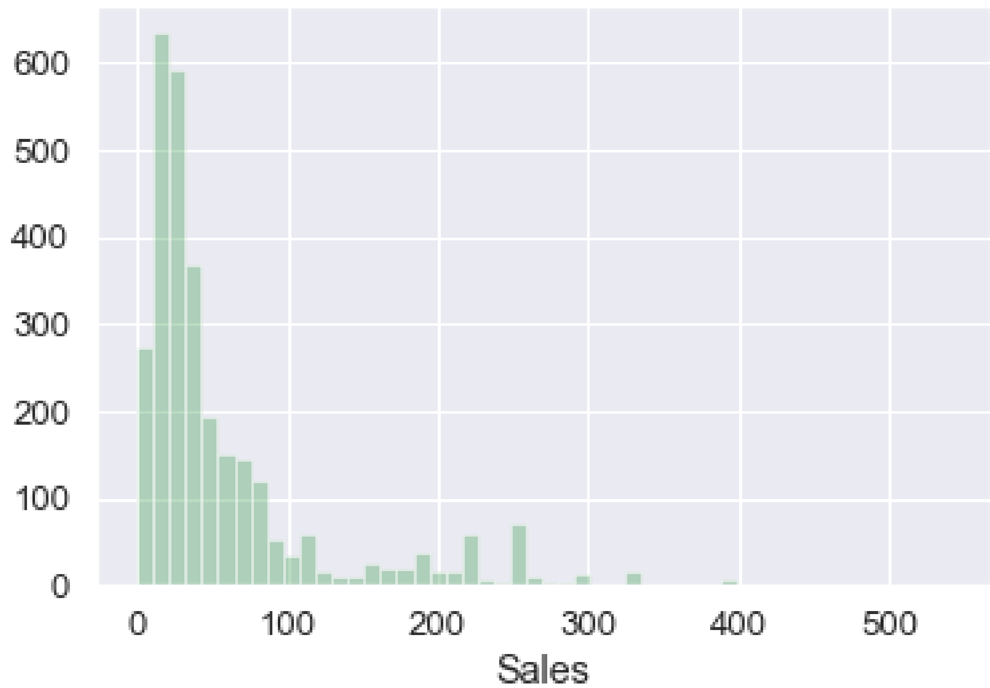
Univariate analysis

Distribution Plots





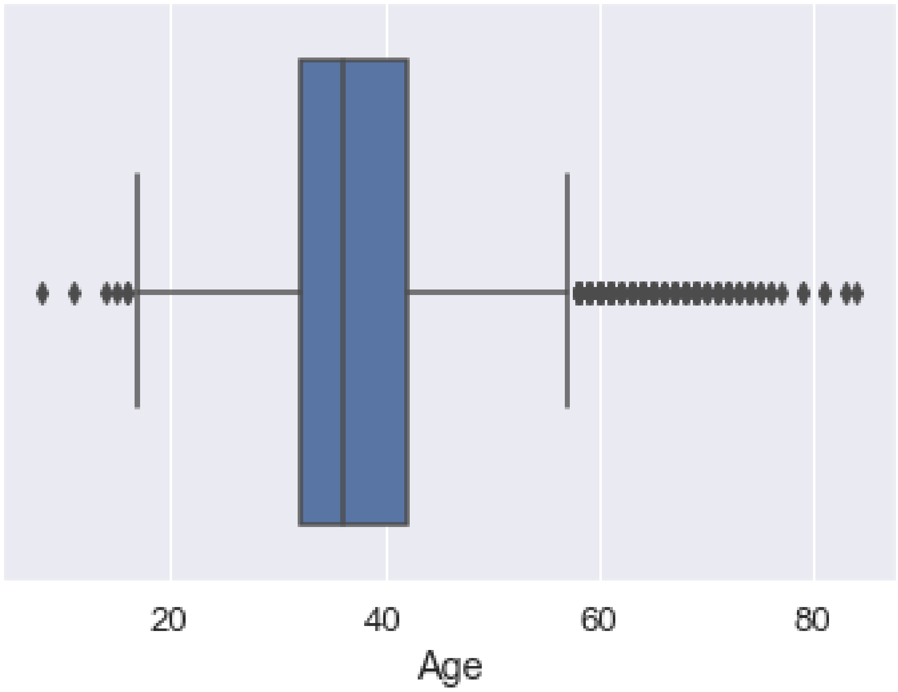


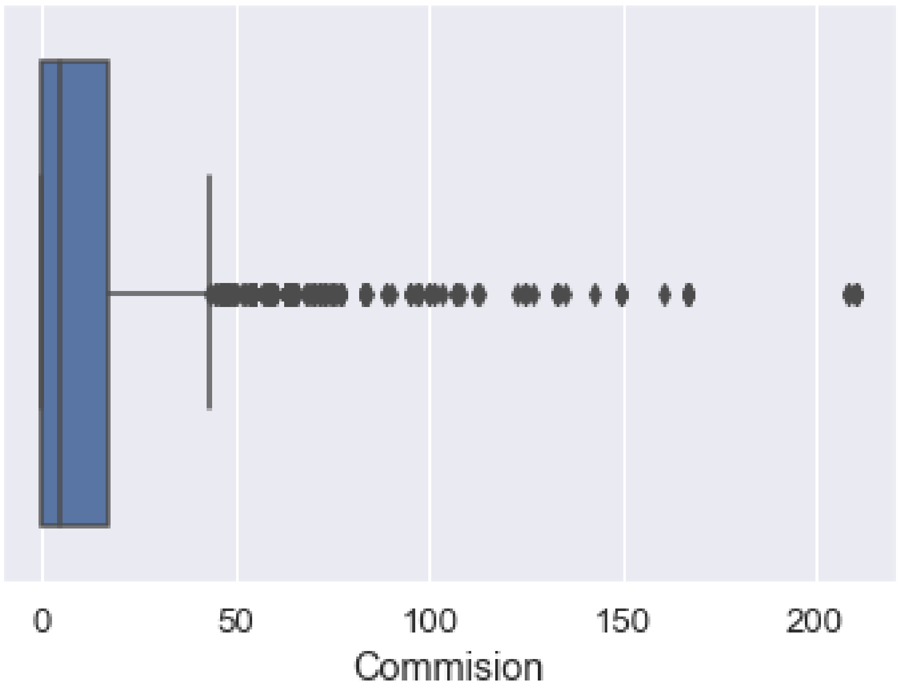


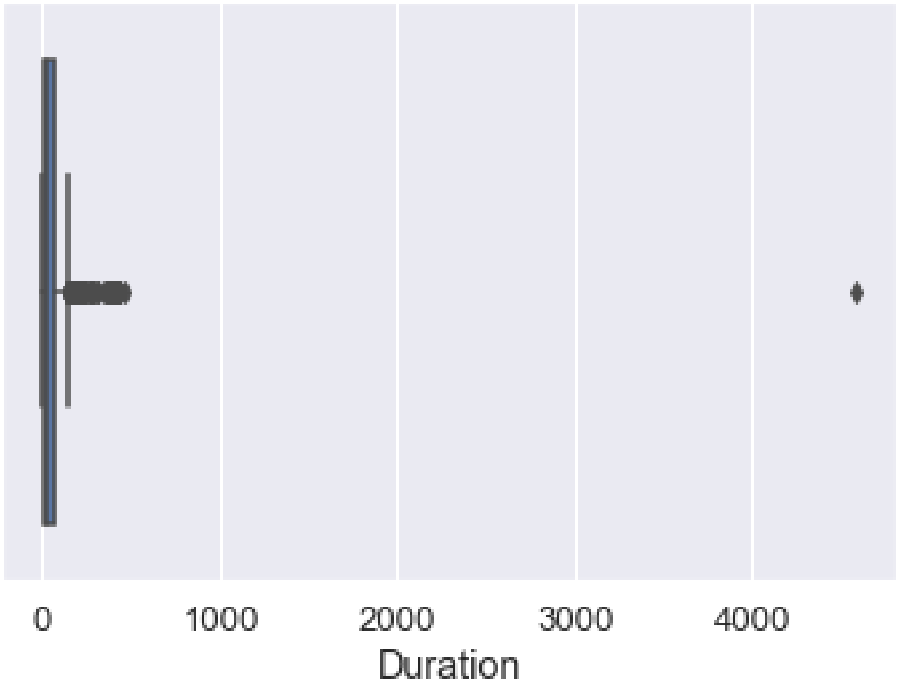
Inferences from distribution plots are as per below:

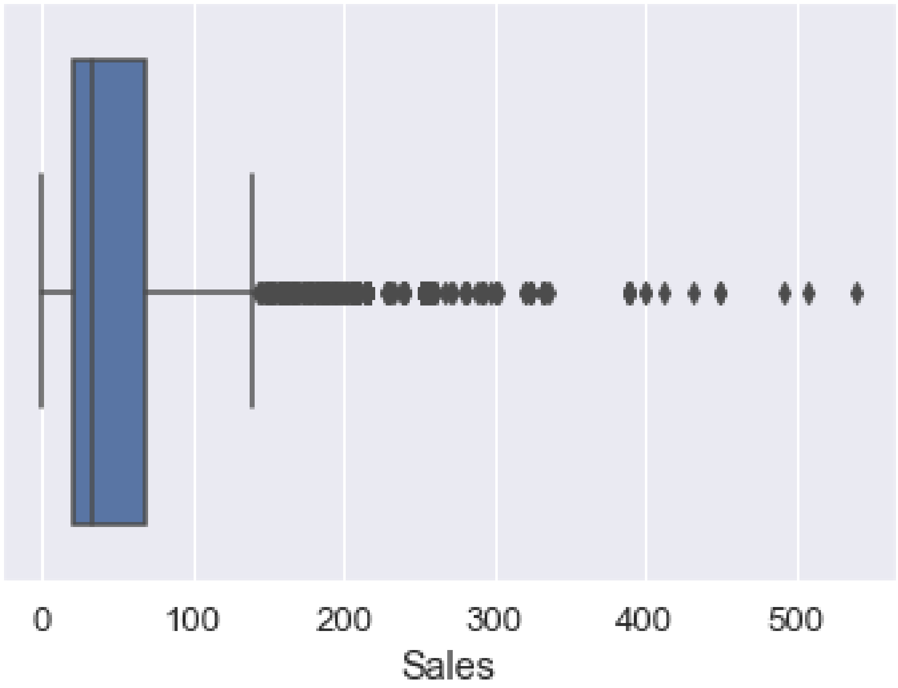
No plot is normally distributed, age, commission and sales are right skewed.

Box plots









Inferences from the box plots are as per below:

Outliers - All the variables are having outliers

Skewness – All are right skewed

Multivariate analysis

Correlation plot

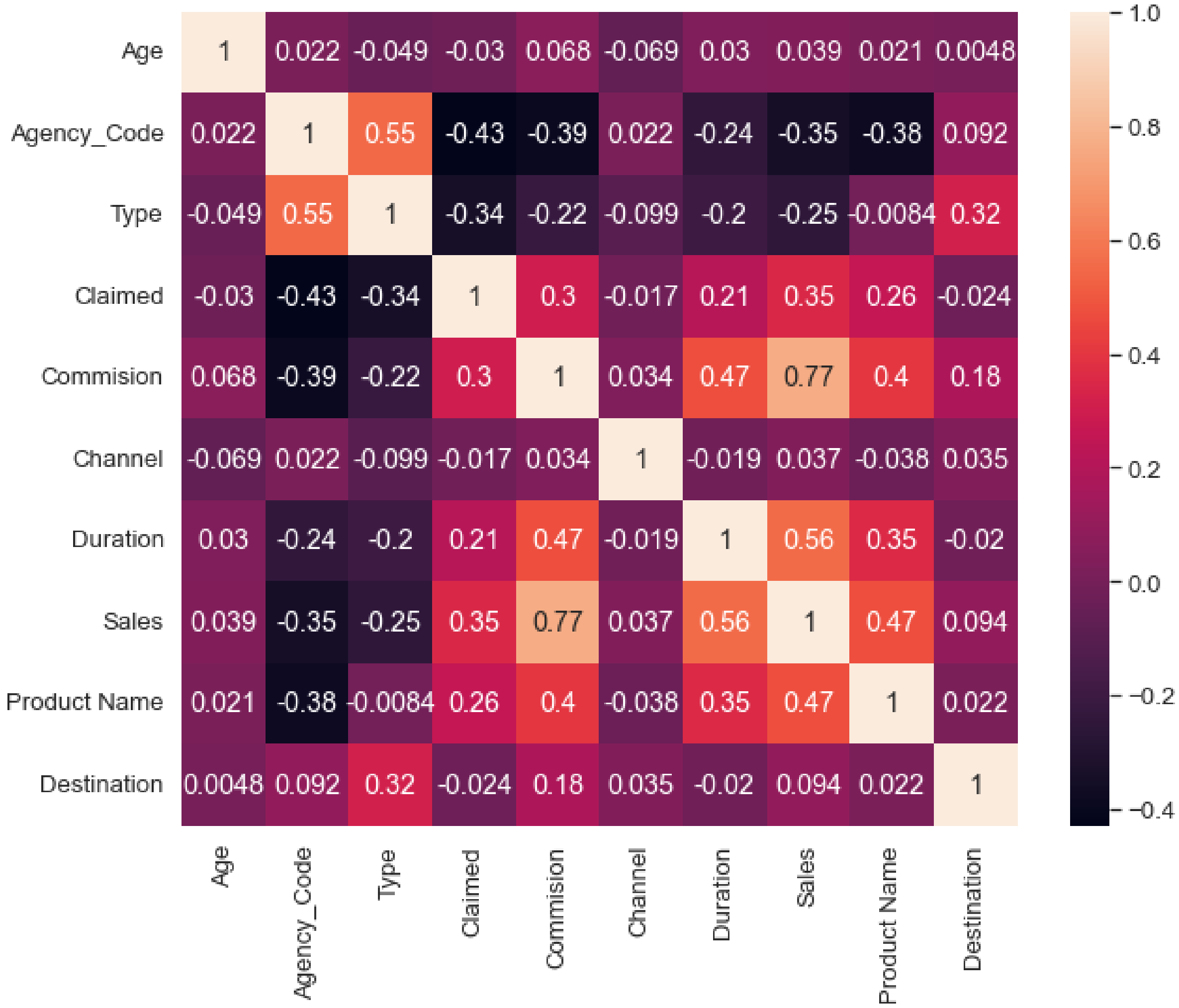


Figure 20

There is no correlation between variables except sales and commission.

Pair plot

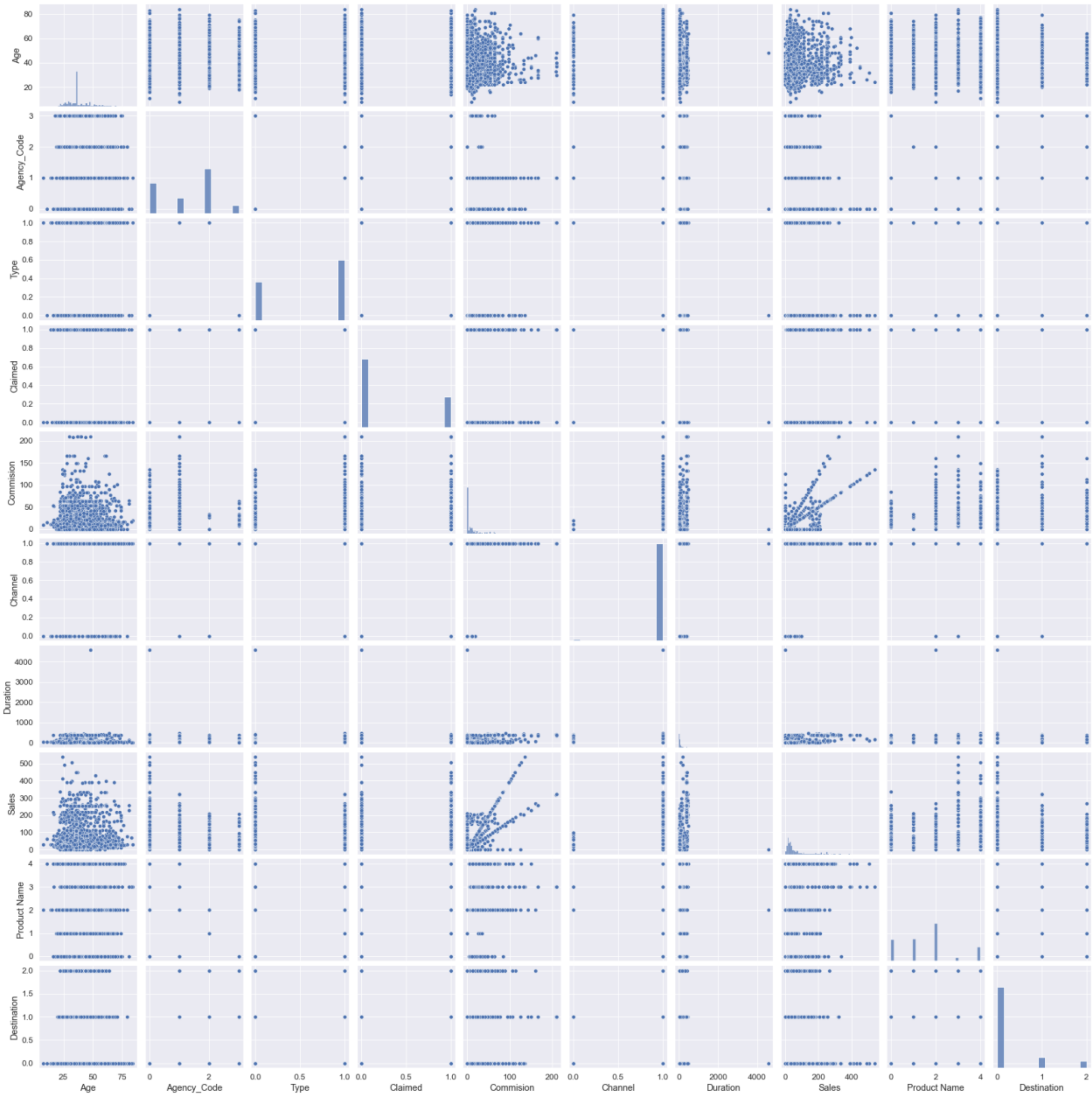


Figure 21

There is no linear relationship between variables

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

The data is split into two parts, train and test with 30% proportion for test data and 70% proportion for train data.

X= independent variables – "Age", "Type", "Commission", "Channel", "Duration", "Sales", "Product Name", "Destination"

Y= dependent variable – "Claimed"

```
shape of x_train is  (2100, 8)
shape of x_test is  (900, 8)
shape of y_train is  (2100,)
shape of y_test is  (900,)
```

Figure 22

Fig 22 shows us the break up train data set and test data set for dependent and independent variables.

Decision Tree Model

We have built the model and found feature importance below

	Imp
Duration	0.247517
Sales	0.196578
Age	0.192237
Product Name	0.178857
Commision	0.151090
Destination	0.022854
Channel	0.008256
Type	0.002612

Figure 23

Fig 23 shows the feature importance from decision tree model where we can see that most important variables are duration of tour, sales for the company and age of insured.

Now we have done a grid search cv to find the best parameters for our model which are as per below

```
{'max_depth': 7,
 'min_samples_leaf': 200,
 'min_samples_split': 150,
 'random_state': 1}
```

Figure 24

Fig 24 shows the best parameters of our decision tree model using which we will build the model and predict the target variable.

Random Forest Model

We have built the model and found feature importance below

	Imp
Duration	0.270305
Sales	0.221317
Age	0.180620
Commision	0.158601
Product Name	0.091396
Type	0.049163
Destination	0.022441
Channel	0.006157

Figure 25

Fig 25 shows the feature importance as per random forest model, here most important features are duration of the tour, sales of company and age of insured.

Now we have used grid search cv here and found the best parameters for our model which are as per below

```
{'bootstrap': True,
 'max_depth': 10,
 'max_features': 6,
 'min_samples_leaf': 155,
 'min_samples_split': 200,
 'n_estimators': 250,
 'random_state': 1}
```

Figure 26

Fig 26 shows the best parameters for our model using which we will build our model and predict the target variable.

Artificial Neural Network

We have built the ANN model and here there is no feature importance.

After the grid search cv the best parameters are as per below

```
{'hidden_layer_sizes': 500,
 'max_iter': 4500,
 'random_state': 1,
 'solver': 'sgd',
 'tol': 0.001}
```

Figure 27

Fig 27 shows the best parameters for ANN model using which we will predict the dependent variable.

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.

Performance evaluation of Decision Tree model

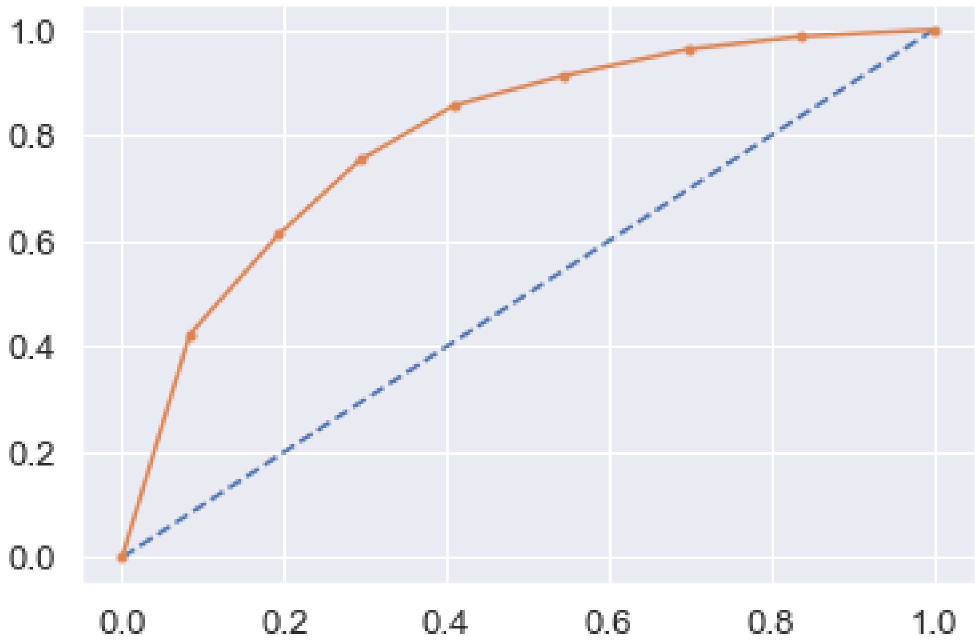


Figure 28

Fig 28 shows the ROC curve for train data for decision tree model

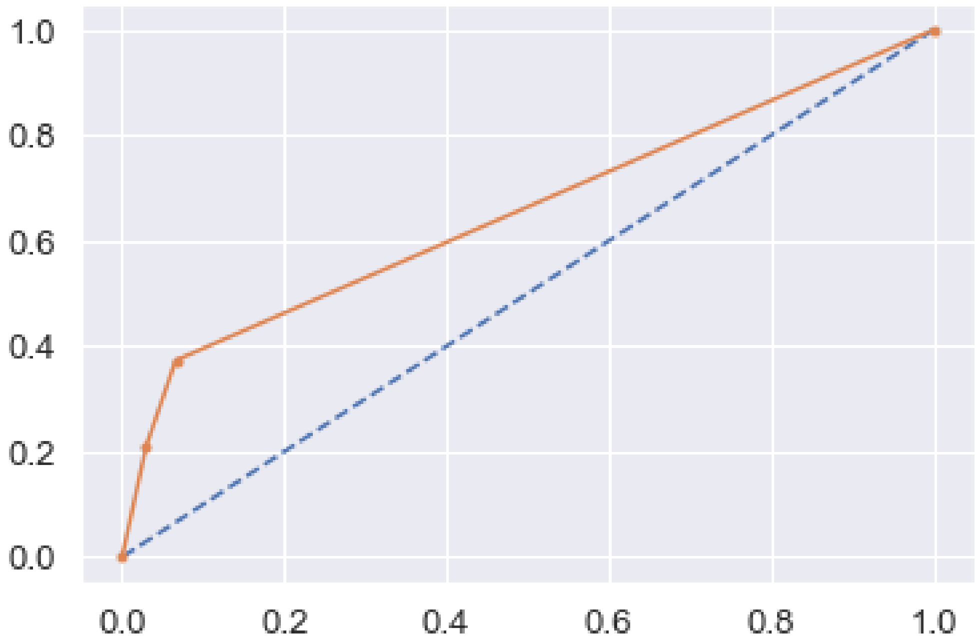


Figure 29

Fig 29 shows the ROC curve of test data for decision tree model.

Here the train data ROC curve is much better than the test data curve.

	precision	recall	f1-score	support
0	0.79	0.92	0.85	1471
1	0.68	0.42	0.52	629
accuracy			0.77	2100
macro avg	0.74	0.67	0.68	2100
weighted avg	0.76	0.77	0.75	2100

Figure 30

Fig 30 shows the classification report of decision tree model in train data.

	precision	recall	f1-score	support
0	0.67	1.00	0.80	605
1	0.00	0.00	0.00	295
accuracy			0.67	900
macro avg	0.34	0.50	0.40	900
weighted avg	0.45	0.67	0.54	900

Figure 31

Fig 31 shows the classification report of decision tree model in test data.

```
array([[1285,  186],
       [ 292,  337]], dtype=int64)
```

Figure 32

Fig 32 is the confusion matrix of decision tree model in train data.

Here TP = 1285, TN= 337, FN= 186, and FP = 292

```
array([[548,  57],
       [162, 133]], dtype=int64)
```

Figure 33

Fig 33 shows the confusion matrix of decision tree model in test data.

Here TP = 548, TN = 133, FN = 57, and FP = 162

Performance evaluation of Random Forest Model

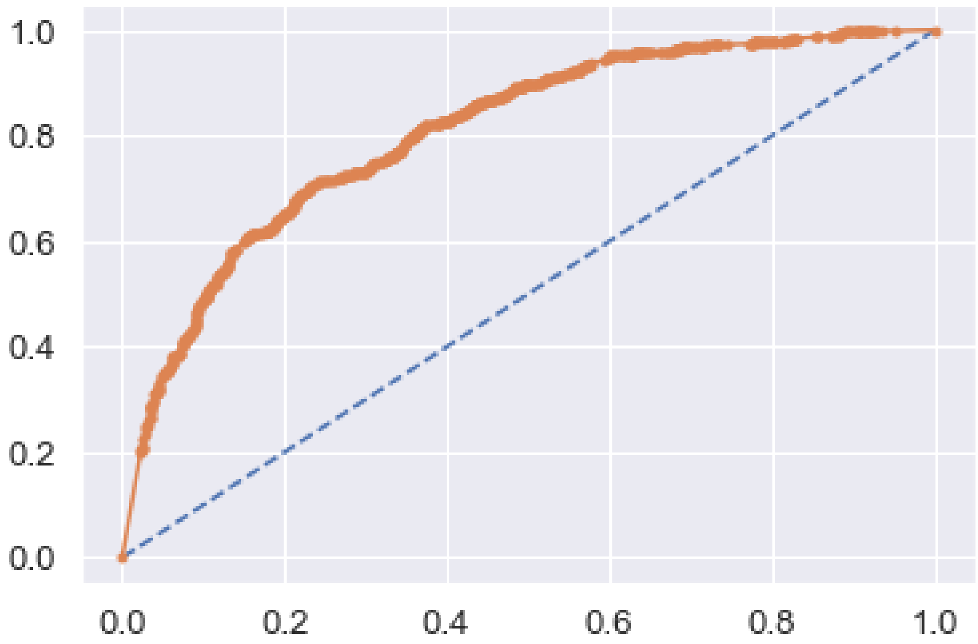


Figure 34

Fig 34 shows the ROC curve of Random forest (RF) model in train data

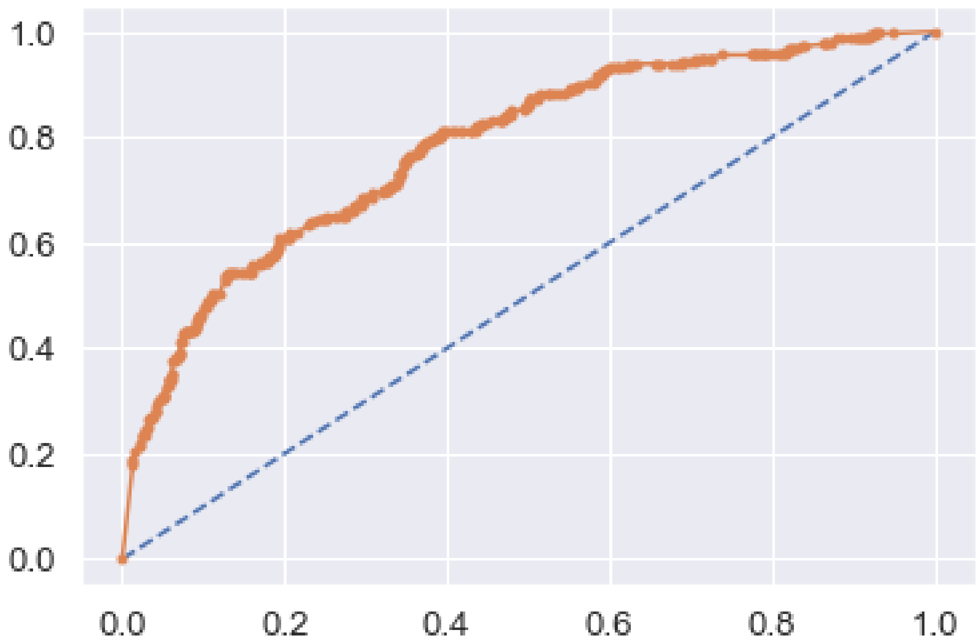


Figure 35

Fig 35 shows the ROC curve of RF model in test data.

	precision	recall	f1-score	support
0	0.79	0.92	0.85	1471
1	0.68	0.42	0.52	629
accuracy			0.77	2100
macro avg	0.74	0.67	0.68	2100
weighted avg	0.76	0.77	0.75	2100

Figure 36

Fig 36 shows the classification report of RF model in train data.

	precision	recall	f1-score	support
0	0.75	0.94	0.84	605
1	0.75	0.37	0.50	295
accuracy			0.75	900
macro avg	0.75	0.66	0.67	900
weighted avg	0.75	0.75	0.73	900

Figure 37

Fig 37 shows the classification report of RF model in test data.

	precision	recall	f1-score	support
0	0.75	0.94	0.84	605
1	0.75	0.37	0.50	295
accuracy			0.75	900
macro avg	0.75	0.66	0.67	900
weighted avg	0.75	0.75	0.73	900

Figure 38

Fig 38 shows the confusion matrix for RF model in train data

```
array([[1348,  123],
       [ 363,  266]], dtype=int64)
```

Figure 39

Fig 39 shows the confusion matrix for RF model using test data.

```
array([[568,  37],
       [185, 110]], dtype=int64)
```

Performance evaluation of ANN model

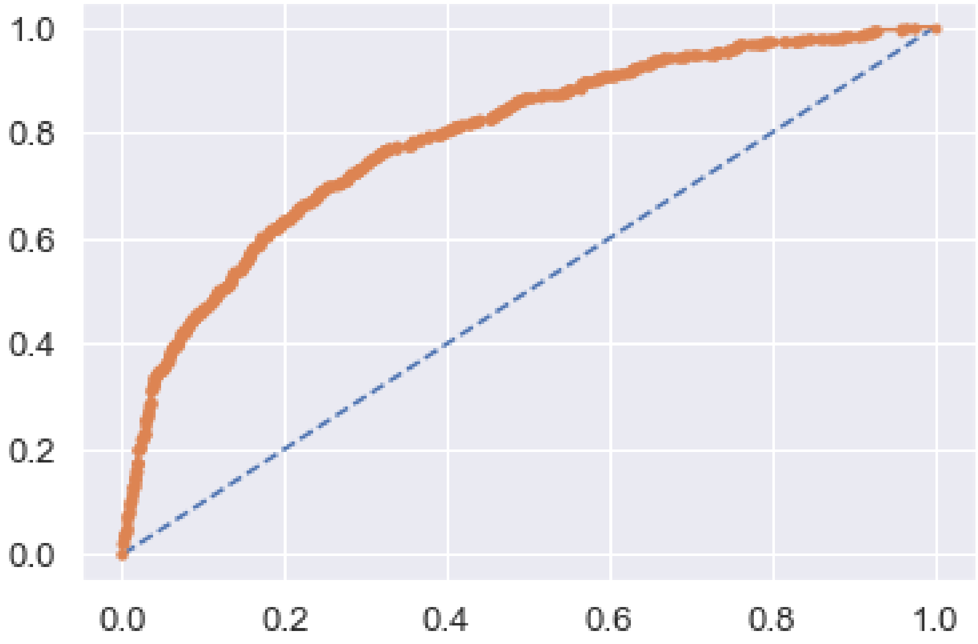


Figure 40

Fig 40 shows the ROC curve of ANN model in train data.

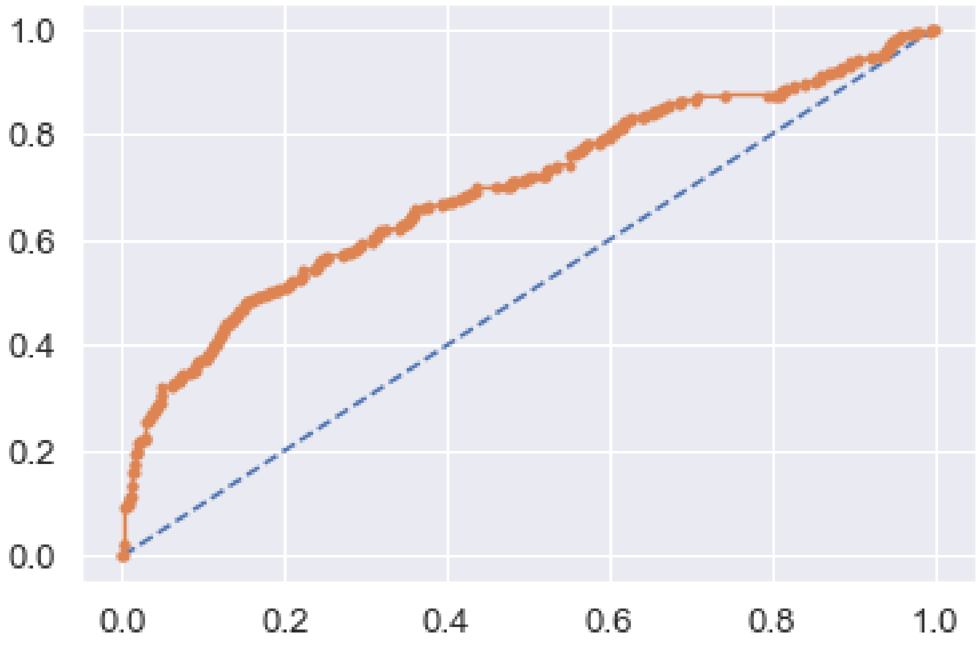


Figure 41

Fig 41 shows the ROC curve of ANN model in test data.

```
              precision    recall  f1-score   support

           0       0.77      0.95      0.85      1471
           1       0.76      0.35      0.48       629

    accuracy                           0.77      2100
   macro avg       0.77      0.65      0.67      2100
weighted avg       0.77      0.77      0.74      2100
```

Figure 42

Fig 42 shows the classification report of ANN model in train data.

```
              precision    recall  f1-score   support

           0       0.74      0.94      0.83       605
           1       0.72      0.32      0.44       295

    accuracy                           0.74       900
   macro avg       0.73      0.63      0.63       900
weighted avg       0.73      0.74      0.70       900
```

Figure 43

Fig 43 shows the classification report of ANN model in test data.

```
array([[1400,   71],
       [ 407,  222]], dtype=int64)
```

Figure 44

Fig 44 shows the confusion matrix of ANN model in train data.

```
array([[569,  36],
       [201,  94]], dtype=int64)
```

Figure 45

Fig 45 shows the confusion matrix of ANN model in test data.

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

	DecisionTreeModel	RandomForestModel	ANN Model
Accuracy for Train Data	0.77	0.77	0.77
Accuracy for Test Data	0.67	0.75	0.74

Figure 46

Fig 46 shows us the accuracy score of each model in train and test data. Here we can observe that RF model is having best accuracy score in both train and test data.

	DecisionTreeModel	RandomForestModel	ANN Model
AUC score for Train Data	0.79	0.80	0.79
AUC score for Test Data	0.65	0.78	0.69

Figure 47

Fig 47 shows the AUC score of each model in train and test data. Here we can observe that the RF model is having comparatively high AUC score in both train and test data.

	DecisionTreeModel	RandomForestModel	ANN Model
Precision score of 0 for train data	0.79	0.79	0.77
Precision score of 1 for train data	0.68	0.68	0.76
Precision score of 0 for test data	0.67	0.75	0.74
Precision score of 1 for test data	0.00	0.75	0.72

Figure 48

Fig 48 shows the precision scores of each model in train and test data. Here the scores of 1 is of interest to us as we want to predict those who will claim in future. So RF model is giving us a good score as compared to other two model.

	DecisionTreeModel	RandomForestModel	ANN Model
Recall score of 0 for train data	0.92	0.92	0.95
Recall score of 1 for train data	0.42	0.42	0.35
Recall score of 0 for test data	1.00	0.94	0.94
Recall score of 1 for test data	0.00	0.37	0.32

Figure 49

Fig 49 shows the recall values of each model in train and test data. Recall shows how many correct predictions happened positively, Here the score of 1 is of interest to us so RF model is giving us the best recall score in 1.

Best model is Random Forest model.

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

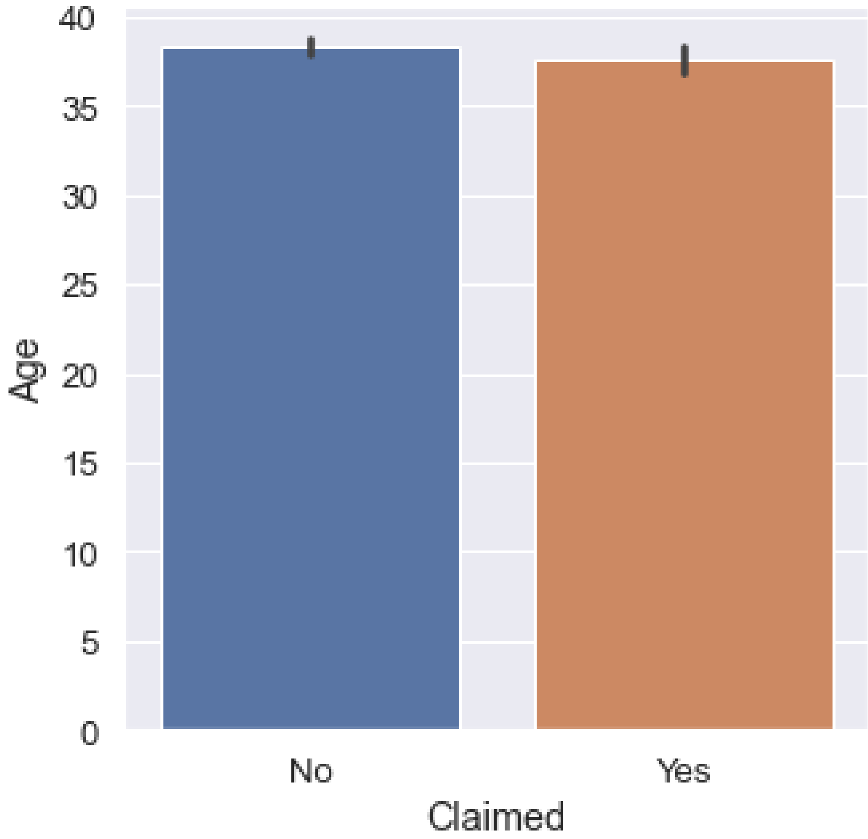


Figure 50

Fig 50 shows that there’s not much effect of age on claim status.

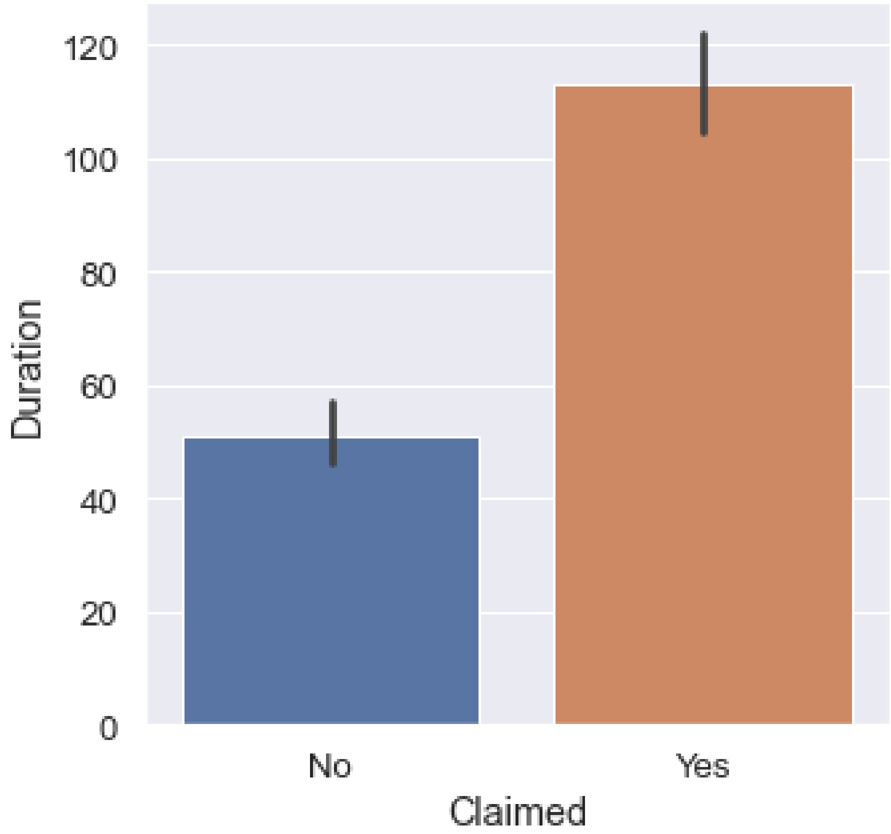


Figure 51

Fig 51 shows that for the less duration tour claim frequency is high.

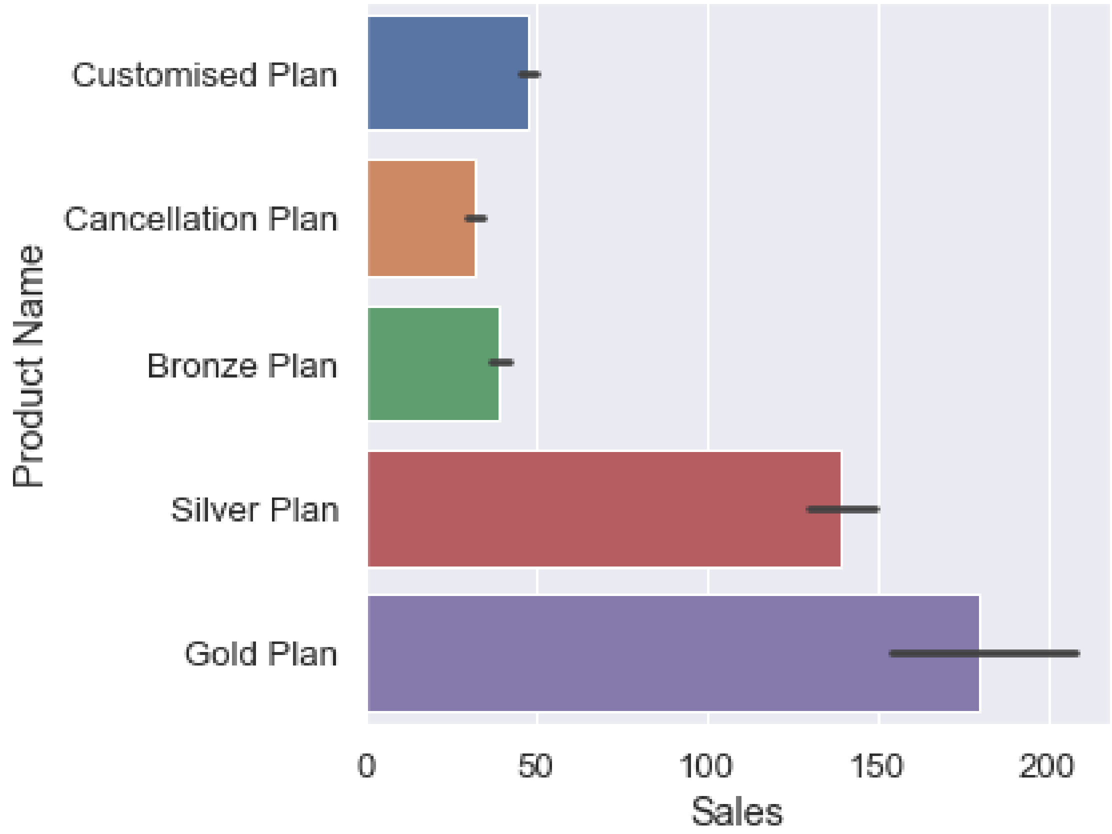


Figure 52

Fig 52 shows that sales for Gold plan is highest.

Destination	ASIA	Americas	EUROPE
Claimed			
No	1691	232	153
Yes	774	88	62

Figure 53

Fig 53 shows that for the destination ASIA claim is highest.

Product Name	Bronze Plan	Cancellation Plan	Customised Plan	Gold Plan	Silver Plan
Claimed					
No	399	635	882	39	121
Yes	251	43	254	70	306

Figure 54

Fig 54 shows that for silver plan the claim frequency is highest.

Type	Airlines	Travel Agency
Claimed		
No	573	1503
Yes	590	334

Figure 55

Fig 55 shows that for Airlines the frequency of claim is highest.

Destination	ASIA	Americas	EUROPE
Product Name			
Bronze Plan	622	16	12
Cancellation Plan	558	72	48
Customised Plan	777	210	149
Gold Plan	87	17	5
Silver Plan	421	5	1

Figure 56

Fig 56 shows that silver plan is given for ASIA destination mostly.

Recommendations:

1. Cancellation plan should be given for ASIA destination more. Also if we can increase its sales in other two continents then there will be improvement in claim frequency.

2.Customised plan has low claim frequency so company should increase its sales for ASIA destination.

3.Insurance for short duration tour should be increased more as claim frequency is less for this.

4.For Europe destination claim frequency is lowest, and company should sell more silver plan there.