

DATA MINING PROJECT REPORT

Akshaya Nallathambi

25th July, 2021



Table Of Contents

Problem 1

Problem statement	7
Data Description	7
Sample of the dataset	8
Types of variables in the data frame	9
1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).	10
1.2 Do you think scaling is necessary for clustering in this case? Justify.	20
1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.	22
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.	24
1.5 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.	28

Problem 2

Problem statement	30
Data Description	30
Sample of the dataset	31
Types of variables in the data frame	31

2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).	33
2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network	45
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.	45
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	58

List of Figures

FIGURE 1	8
FIGURE 2	10
FIGURE 3	10
FIGURE 4	11
FIGURE 5	14
FIGURE 6	21
FIGURE 7	21
FIGURE 8	23
FIGURE 9	25
FIGURE 10	26
FIGURE 11	27

FIGURE 12	28
FIGURE 13	28
FIGURE 14	31
FIGURE 15	33
FIGURE 16	33
FIGURE 17	34
FIGURE 18	37
FIGURE 19	46
FIGURE 20	46
FIGURE 21	48
FIGURE 22	48
FIGURE 23	50
FIGURE 24	50
FIGURE 25	51
FIGURE 26	52
FIGURE 27	53
FIGURE 28	53
FIGURE 29	54
FIGURE 30	55

List of Tables

TABLE 1	9
TABLE 2	13
TABLE 3	31
TABLE 4	36

List of Graphs

GRAPH 1	11
GRAPH 2	15
GRAPH 3	17
GRAPH 4	19
GRAPH 5	23
GRAPH 6	24
GRAPH 7	25
GRAPH 8	27
GRAPH 9	35
GRAPH 10	38
GRAPH 11	39
GRAPH 12	42
GRAPH 13	43
GRAPH 14	44
GRAPH 15	47

GRAPH 16	49
GRAPH 17	51
GRAPH 18	52
GRAPH 19	54
GRAPH 20	55

Problem 1

Problem statement-

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.

Data Description-

spending: Amount spent by the customer per month (in 1000s)

advance_payments: Amount paid by the customer in advance by cash (in 100s)

probability_of_full_payment: Probability of payment done in full by the customer to the bank

current_balance: Balance amount left in the account to make purchases (in 1000s)

credit_limit: Limit of the amount in credit card (10000s)

min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)

max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s)

Sample of the dataset-

FIGURE 1

	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
5	12.70	13.41	0.8874	5.183	3.091	8.456	5.000
6	12.02	13.33	0.8503	5.350	2.810	4.271	5.308
7	13.74	14.05	0.8744	5.482	3.114	2.932	4.825
8	18.17	16.26	0.8637	6.271	3.512	2.853	6.273
9	11.23	12.88	0.8511	5.140	2.795	4.325	5.003

There are 7 variables and all are float values. The data given is for 210 individuals. There are no null values.

Types of variables in the data frame-

TABLE 1

Spending	float64	Continuous
Advance_payments	float64	Continuous
Probability_of_full_payment	float64	Continuous
Current_balance	float64	Continuous
Credit_limit	float64	Continuous
Min_payment_amt	float64	Continuous
Max_spent_in_single_shopping	float64	Continuous

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

The below table gives the first 5 rows of sample data.

FIGURE 2

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.8752	6.675	3.763	3.252	6.550
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	5.890	3.694	2.068	5.837

The image below gives the basic information of the data set. It is clear that all the variables are of type float with 7 columns and 210 rows. There are no null values. The memory usage is 11.6 KB.

FIGURE 3

```
<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
```

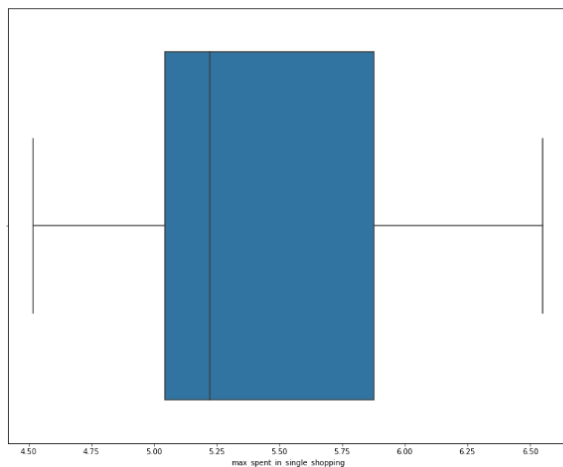
The below image gives the five point summary of the continuous variables in the data set. It is clear that the data needs scaling as the numbers are of different magnitude.

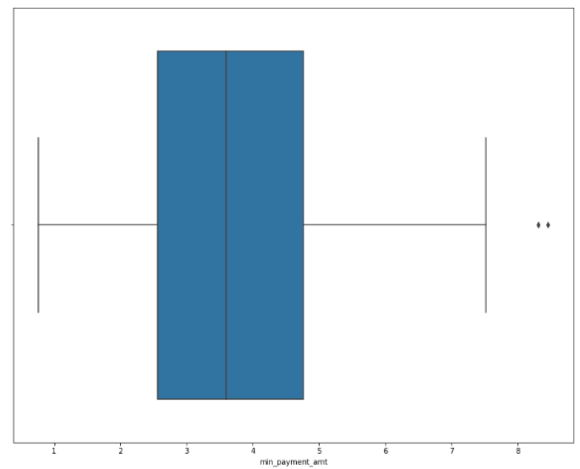
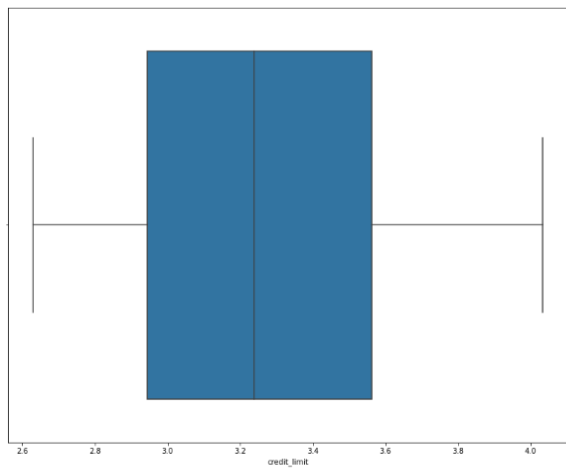
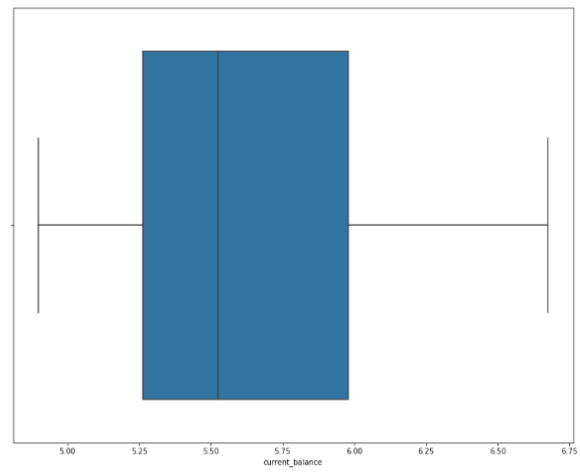
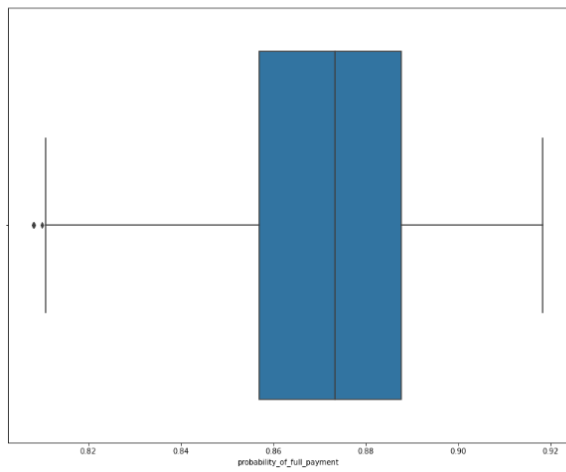
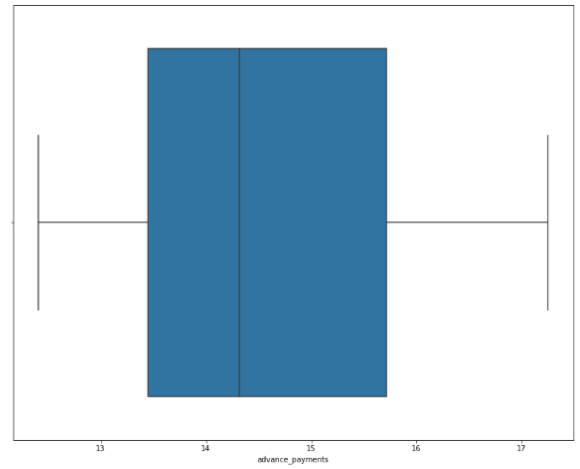
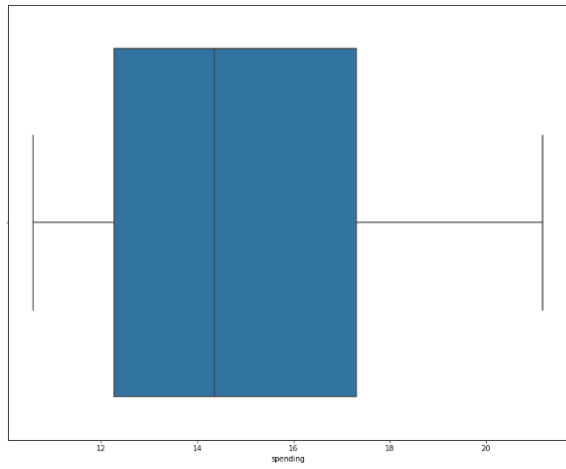
FIGURE 4

	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

There are no outliers in all the variables except “min_payment_amt”. This is evident from the box plots below,

GRAPH 1





There are no categorical variables in the given data set. So univariate and bivariate analysis is done only in the numerical variables of the data.

Univariate analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. ^[1]

The histograms are used for numerical variables to perform univariate analysis.

It is clear from the graph (Graph) that all the numerical variables are skewed.

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. ^[2]

TABLE 2

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ± 1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.

FIGURE 5

	Skewness
spending	0.397027
advance_payments	0.383806
probability_of_full_payment	-0.534104
current_balance	0.521721
credit_limit	0.133416
min_payment_amt	0.398793
max_spent_in_single_shopping	0.557876

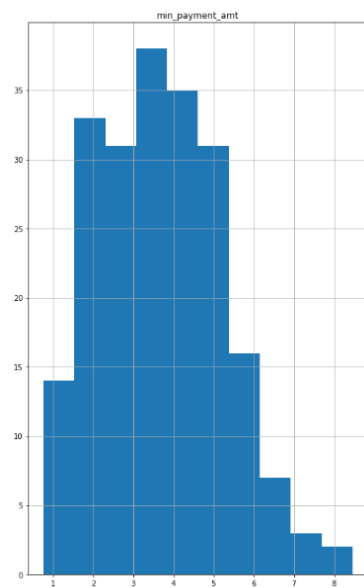
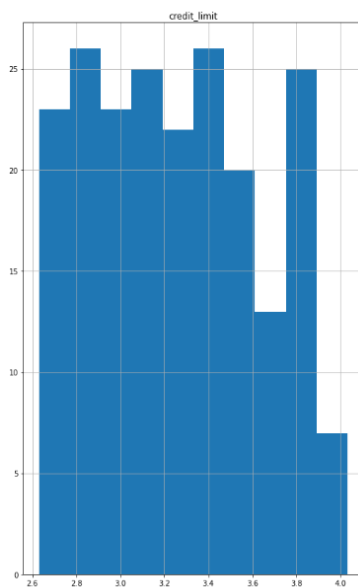
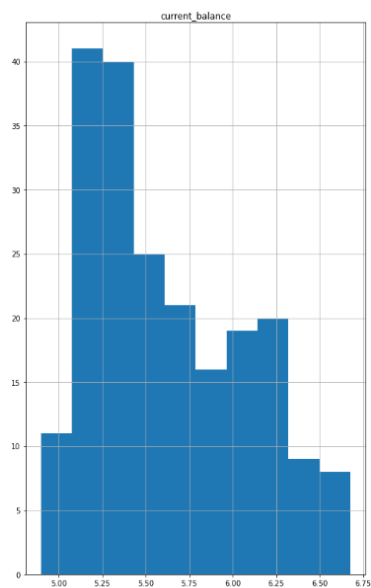
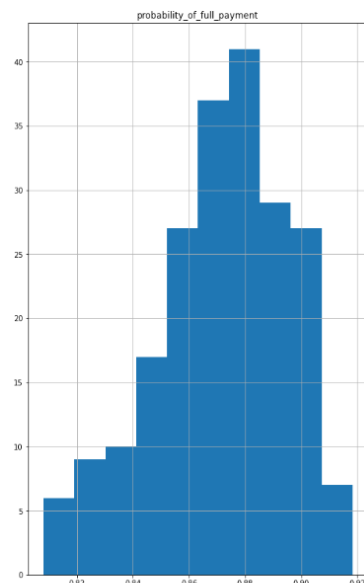
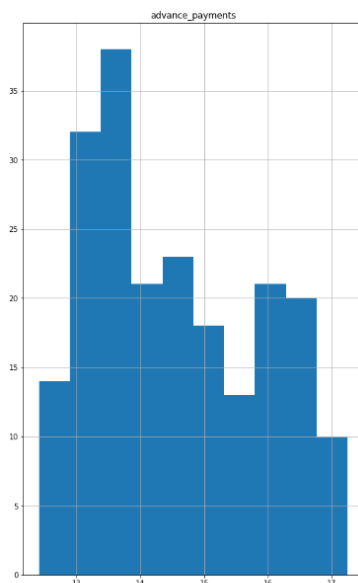
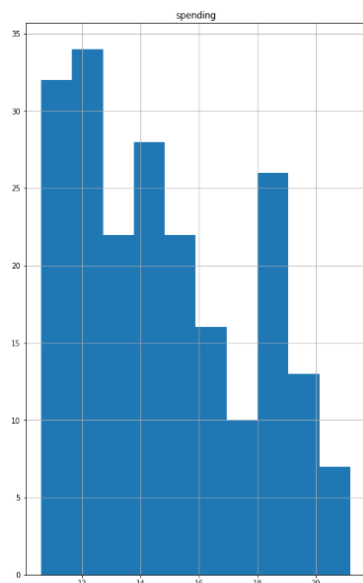
Right skewed variables:

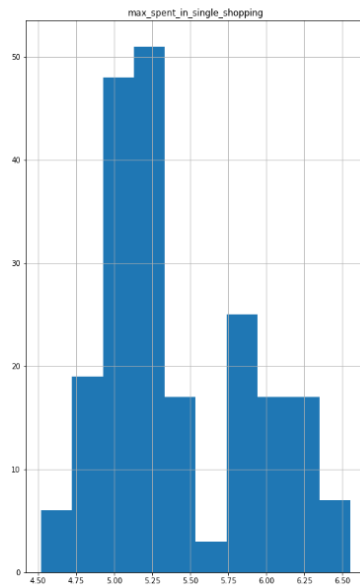
Spending, Advance_payments, Current_balance, Credit_limit, Min_payment_amt, Max_spent_in_single_shopping

Left skewed variables:

Probability_of_full_payment

GRAPH 2





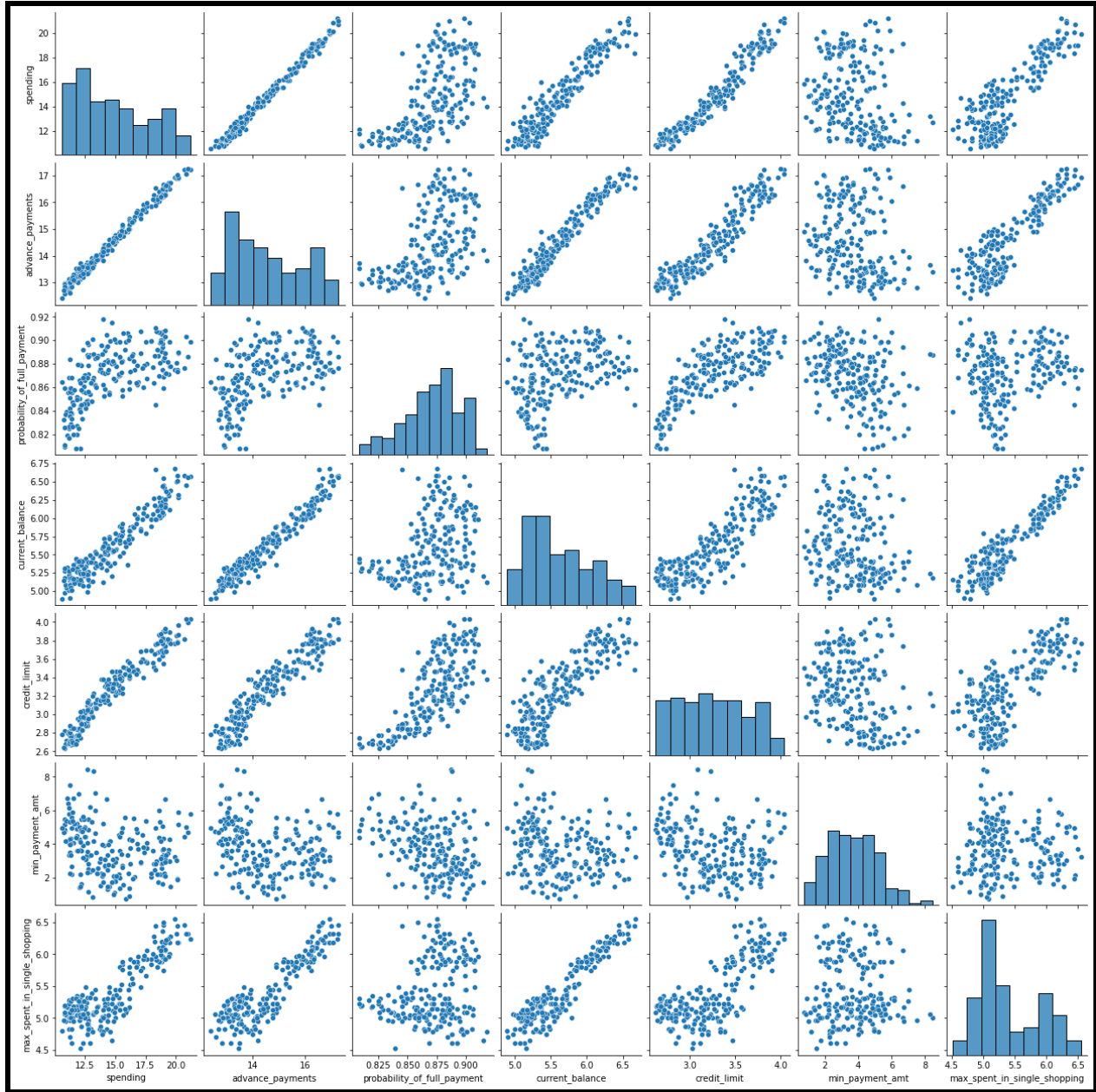
Bivariate analysis:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. ^[3]

The pairplot is generally used for numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. ^[4]

GRAPH 3

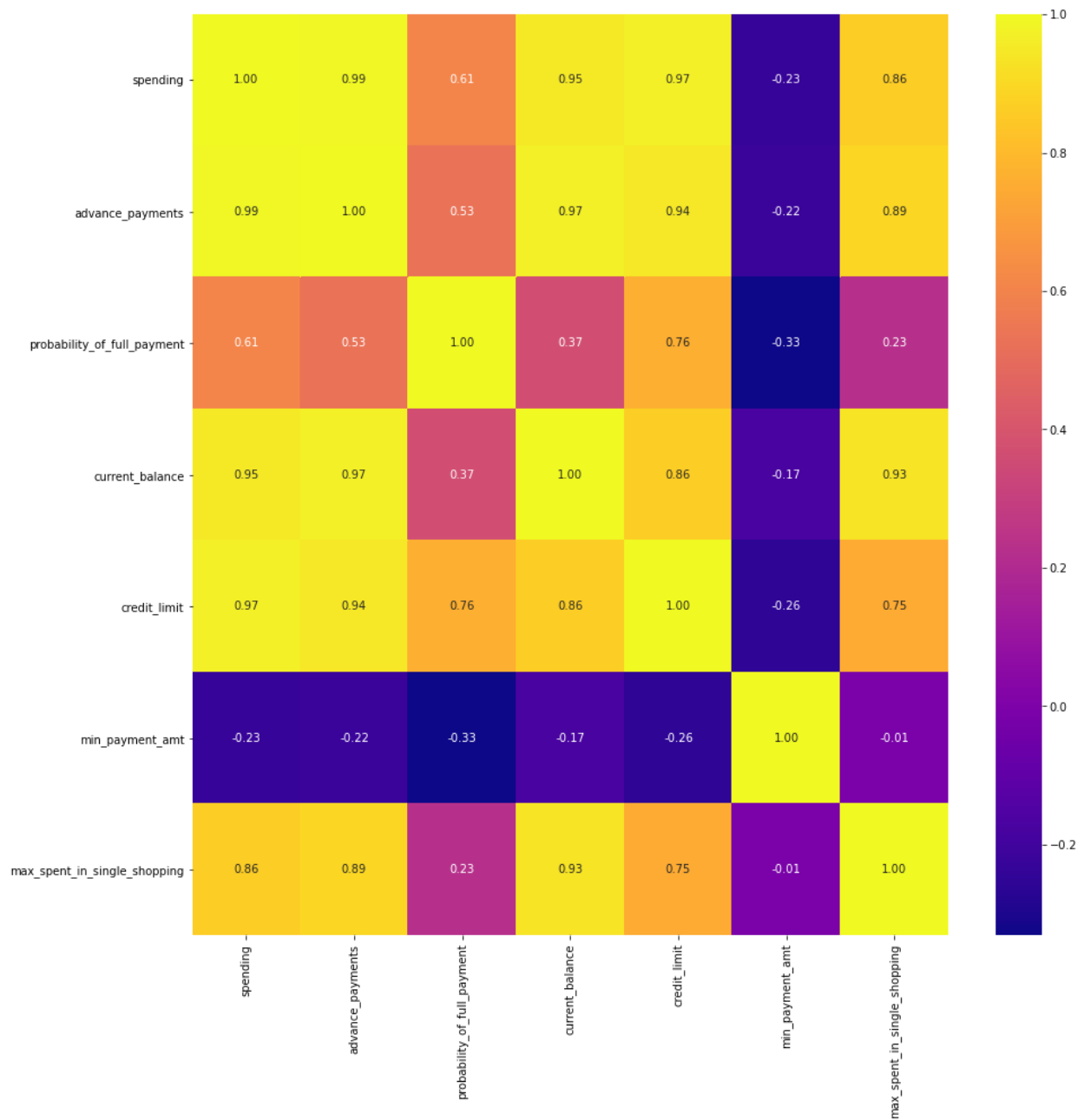


The below are the findings from the pairplot generated -

- The variable Spending is highly correlated with the variables Advance_payments, Current_balance, Credit_limit and Max_spent_in_single_shopping.
- The variable Advance_payments is highly correlated with the variables Current_balance, Credit_limit and Max_spent_in_single_shopping.
- The variable Current_balance is highly correlated with the variables Credit_limit and Max_spent_in_single_shopping.

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. The heat map for all the numerical variable is below,

GRAPH 4



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

Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. [5]

Scaling converts variables with different scales of measurements into a single scale. This is done only for the numerical variables.

The data is scaled using the formula $\frac{X-\mu}{\sigma}$.

μ : Mean

σ : Standard deviation

The process of scaling is necessary in the given data set as the variables of the data set are of different scales i.e. one variable has two digit numbers and other has only one digit number. For e.g. in our data set “Spending” has values in two digits and “Probability_of_full_payment” has only decimal values that start with zero on the left side of the decimal point. Since the data in these variables are of different scales, it is tough to compare these variables. Therefore scaling is done in the given data set.

Before Scaling:

FIGURE 6

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.8752	6.675	3.763	3.252	6.550
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	5.890	3.694	2.068	5.837
12.70	13.41	0.8874	5.183	3.091	8.456	5.000
12.02	13.33	0.8503	5.350	2.810	4.271	5.308
13.74	14.05	0.8744	5.482	3.114	2.932	4.825
18.17	16.26	0.8637	6.271	3.512	2.853	6.273
11.23	12.88	0.8511	5.140	2.795	4.325	5.003

After Scaling:

FIGURE 7

```
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]])
```

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

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. ^[6]

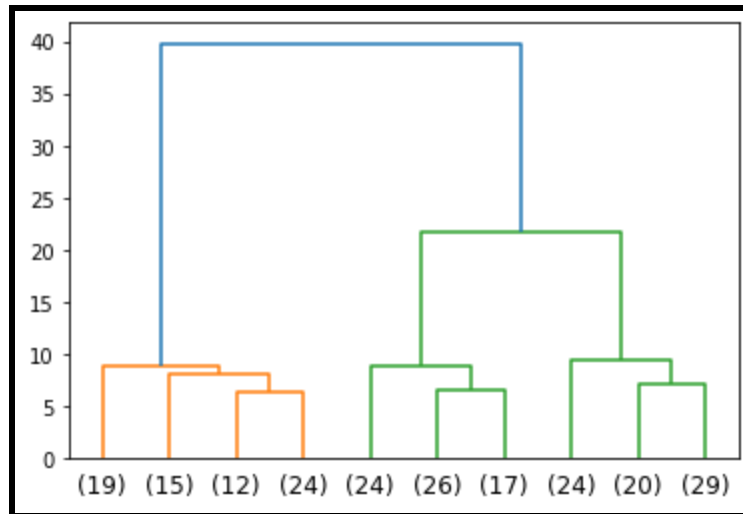
Ward's linkage is a method for hierarchical cluster analysis . The idea has much in common with analysis of variance (ANOVA). The linkage function specifying the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster. Ward's Method seeks to choose the successive clustering steps so as to minimize the increase in ESS at each step. ^[7]

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters. ^[8]

I have used the Ward's linkage method for the dendrogram in the given data set. The below image is the dendrogram generated for the data.

There are majorly three visible clusters in the data set. Even though the dendrogram has only two colors in it, we can see three clear divisions. The green color branches have two clusters and the orange color branches indicate one cluster.

GRAPH 5

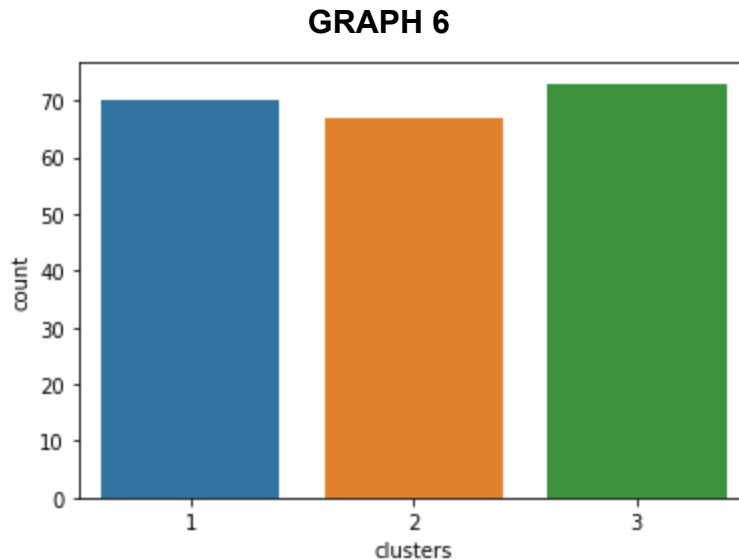


The below table is the final output, after appending the respective cluster number as a new column in the data set.

FIGURE 8

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
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	3
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

There are more customers in cluster 3 than cluster 1 and 2. There are 73 customers who fall under cluster 3. This is evident from the below count plot. The cluster 1 has 70 customers and cluster 2 has 67 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.

K Means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. ^[9]

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k. To determine the optimal number of clusters, we have to select the value of k at the “elbow” i.e. the point after which the distortion starts decreasing in a linear fashion. ^[10]

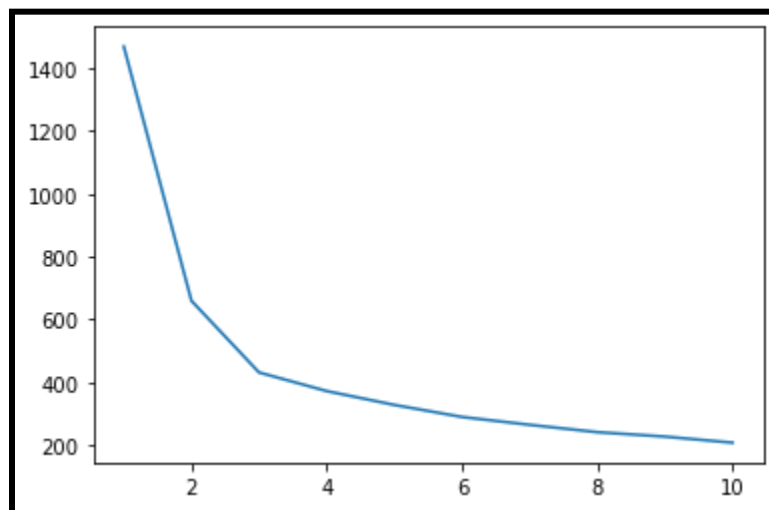
WSS means the sum of distances between the points and the corresponding centroids for each cluster. The WSS for different values of k is shown below.

FIGURE 9

```
[ 1469.9999999999995,  
  659.1717544870411,  
  430.65897315130064,  
  371.301721277542,  
  327.18981108824903,  
  288.76945770226405,  
  263.7681846966533,  
  240.34201335605746,  
  226.15552343181116,  
  206.78724537511738]
```

The below image is the elbow curve for the inertia of the various values of k. It is visibly clear that there is no steep drop in the curve after k = 3.

GRAPH 7



Therefore, the data set is segregated into three different clusters. The below table is the final output, after appending the respective cluster number as a new column in the data set.

FIGURE 10

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Clus_kmeans
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	0
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Silhouette Score = $(b-a)/\max(a,b)$

where,

a: average intra-cluster distance i.e the average distance between each point within a cluster.

b: average inter-cluster distance i.e the average distance between all clusters. ^[11]

The silhouette score for the given data set is 0.40072705527512986, which is close to 1. Thus the clustering is done right.

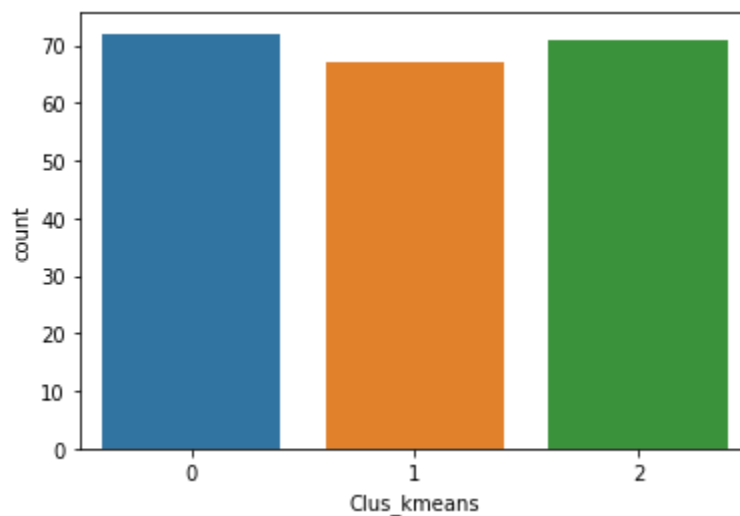
The below table is the sample of the given data set along with the silhouette width of each row.

FIGURE 11

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Clus_kmeans	sil_width
19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1	0.573699
15.99	14.89	0.9064	5.363	3.582	3.336	5.144	0	0.366386
18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1	0.637784
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0.512458
17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1	0.362276

There are more customers in cluster 0. Almost 72 customers fall under cluster 0, 67 customers under cluster 1 and 71 customers in cluster 2. This is evident from the below count plot.

GRAPH 8



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

Hierarchical clustering profile and recommendations:

FIGURE 12

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
clusters								
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

The average spending of cluster 1 seems to be higher than cluster 2. All the columns of cluster 1 have a higher number except for the “min_payment_amt” and “Freq”.

The bank can focus more on the cluster 1 customers for any retail shopping related offers as they have a higher spending average. They can be considered the highest spending group.

Next to cluster 1, the customers in cluster 2 have a higher probability value to pay the full payment. So they can also be considered for the promotional offers.

The customers of cluster 3 can be attracted more by giving them offers that have higher discounts. This group has a good balance left in their account and they seem to spend less than the other groups.

K-Means clustering profile and recommendations:

FIGURE 13

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	sil_width	freq
kmeans									
0	11.856944	13.247778	0.848253	5.231750	2.849542	4.742389	5.101722	0.397473	72
1	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701	0.468772	67
2	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803	0.339816	71

The average spending of cluster 1 seems to be higher than other clusters. All the columns of cluster 1 have a higher number except for the “min_payment_amt” and “Freq”.

The bank can focus more on the cluster 1 and cluster 2 customers for any kind of promotional offers as they have a higher spending average.

The customers of cluster 0 can be attracted more by giving them offers that have high discounts.

The customers of cluster 1 have a good current balance and credit limit. So this group can also be considered for any special offers other than the existing ones.

Problem 2

Problem statement-

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.

Data Description-

Claimed: Target (Claim Status)

Agency_Code: Code of tour firm

Type: Type of tour insurance firms

Channel: Distribution channel of tour insurance agencies

Product: Name of the tour insurance products

Duration: Duration of the tour

Destination: Destination of the tour

Sales: Amount of sales of tour insurance policies

Commission: The commission received for tour insurance firm

Age: Age of insured

Sample of the dataset-

FIGURE 14

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

There are 10 variables, out of which 6 variables are categorical and 4 are continuous. The data given is for 1000 individuals. There are no null values.

Types of variables in the data frame-

TABLE 3

Age	int64	Continuous
Agency_Code	object	Categorical
Type	object	Categorical
Claimed	object	Categorical

Commision	float64	Continuous
Channel	object	Categorical
Duration	int64	Continuous
Sales	float64	Continuous
Product Name	object	Categorical
Destination	object	Categorical

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

The below table gives the first 5 rows of sample data.

FIGURE 15

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

The image below gives the basic information of the data set. It is clear that there are three types of variables namely float, int and object with 10 columns and 3000 rows. There are no null values. The memory usage is 234.5 KB.

FIGURE 16

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
#   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
```

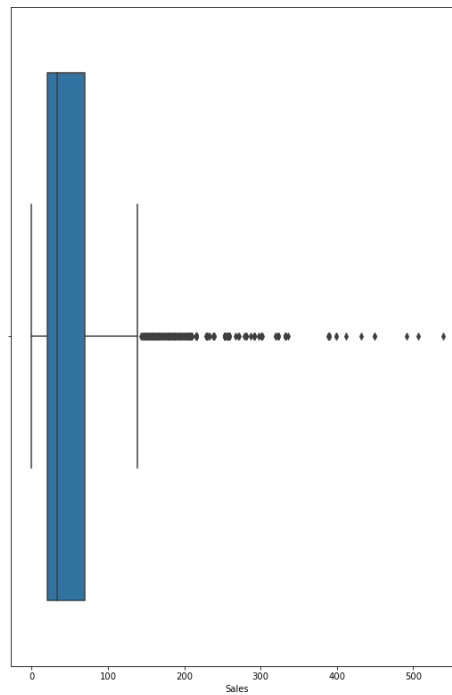
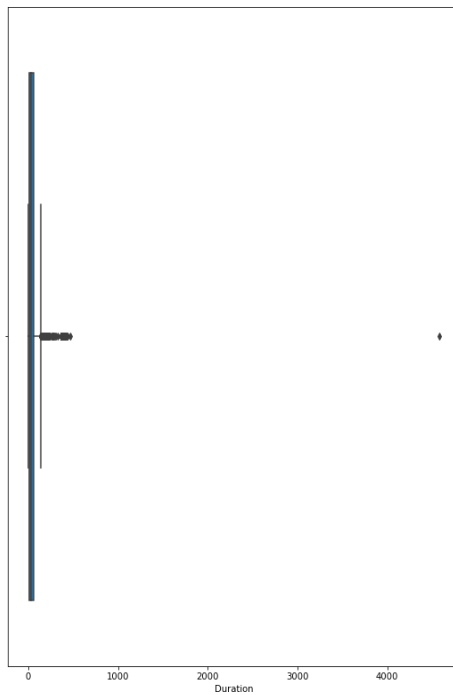
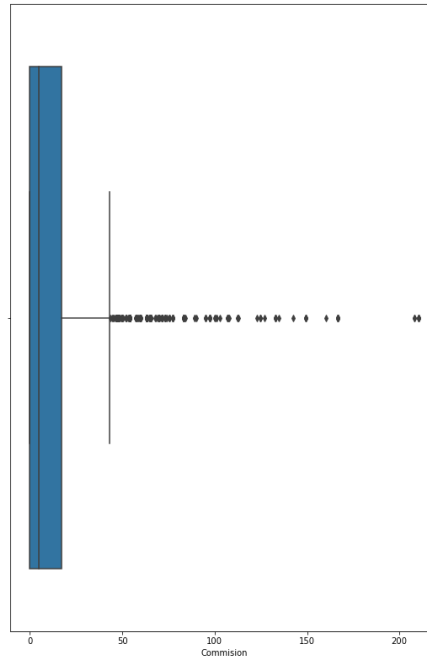
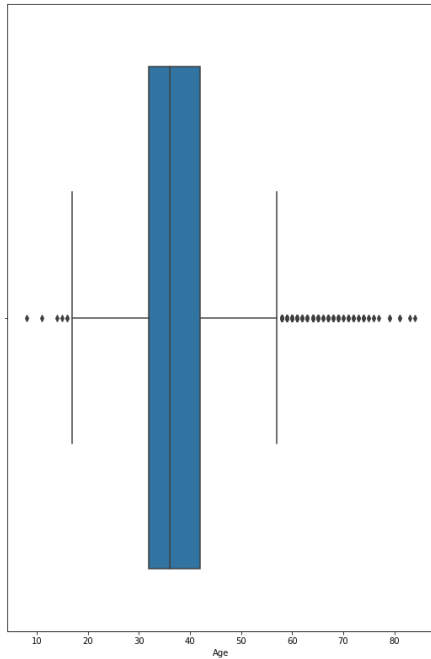
The below image gives the five point summary of the continuous variables in the data set.

FIGURE 17

	Age	Commision	Duration	Sales
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	38.091000	14.529203	70.001333	60.249913
std	10.463518	25.481455	134.053313	70.733954
min	8.000000	0.000000	-1.000000	0.000000
25%	32.000000	0.000000	11.000000	20.000000
50%	36.000000	4.630000	26.500000	33.000000
75%	42.000000	17.235000	63.000000	69.000000
max	84.000000	210.210000	4580.000000	539.000000

There are outliers in all the continuous variables. This is evident from the box plots below,

GRAPH 9



There are both categorical variables and continuous variables in the given data set. So univariate and bivariate analysis can be done in both.

Univariate analysis:

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. ^[1]

The histograms are used for numerical variables and count plots are used for categorical variables to perform univariate analysis.

It is clear from the graph (Graph) that all the numerical variables are skewed.

Skewness:

Skewness is a number that indicates to what extent a variable is asymmetrically distributed. The below table represents the level of skewness and the respective skewness value. ^[2]

TABLE 4

Skewness level	Value
Symmetrical or Not Skewed	0
Less Skewed Data	± 0.5 to 1
Highly Skewed Data	Greater than ± 1

When the skewness value is positive it is considered as right skewed data and when the skewness value is negative it is considered as left skewed data.

The table below shows the skewness value corresponding to each variable in the given data set.

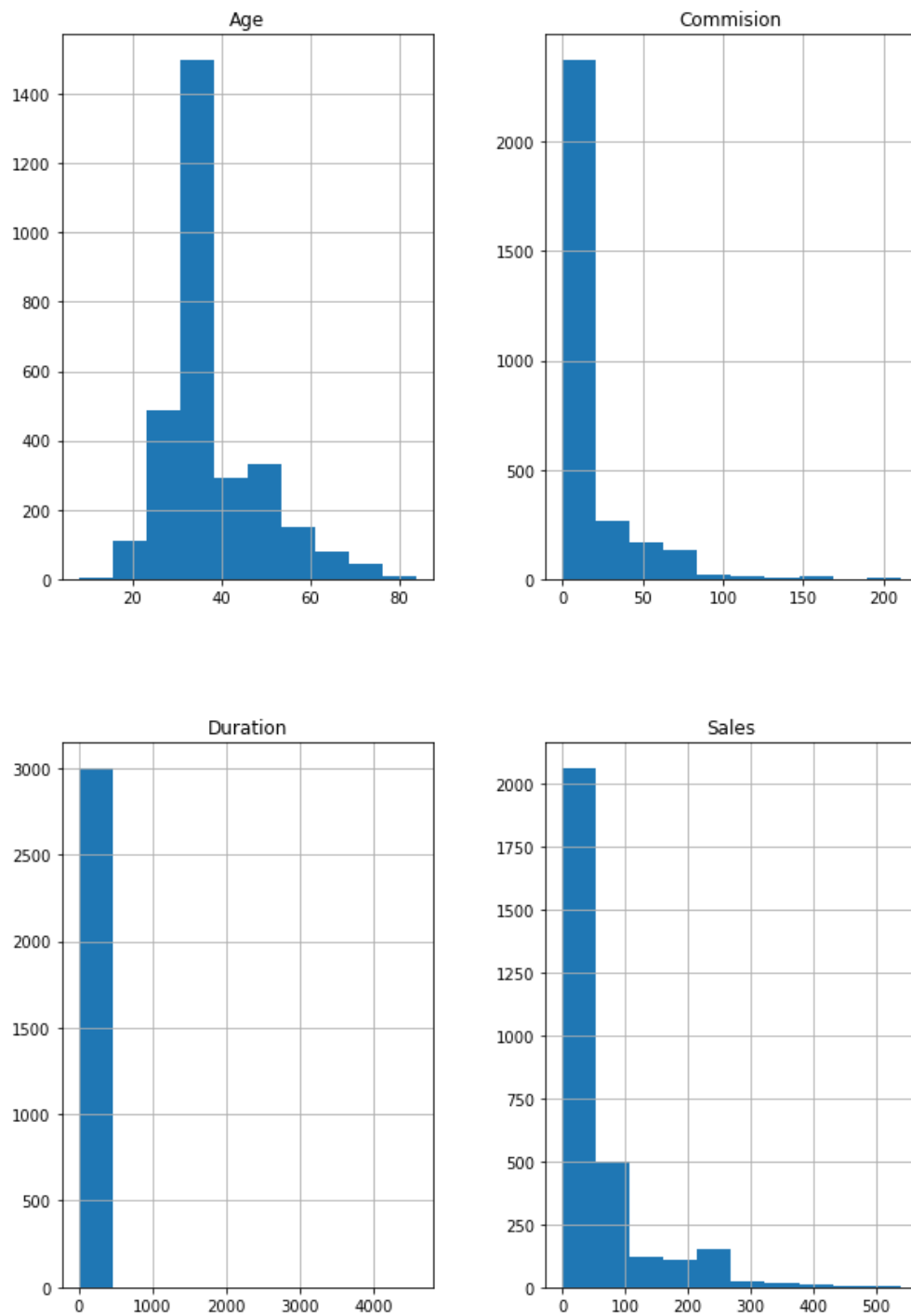
FIGURE 18

Skewness	
Age	1.149138
Commision	3.147283
Duration	13.777788
Sales	2.379958

All the variables have a positive skew value. Therefore, all four variables above are right skewed variables.

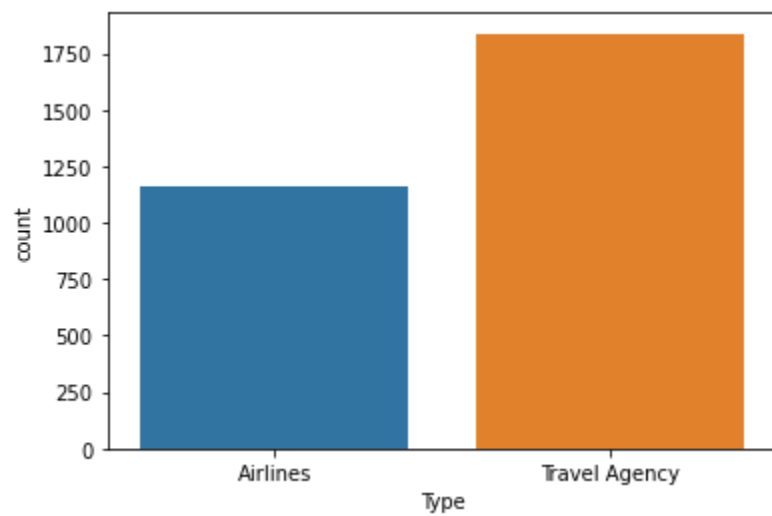
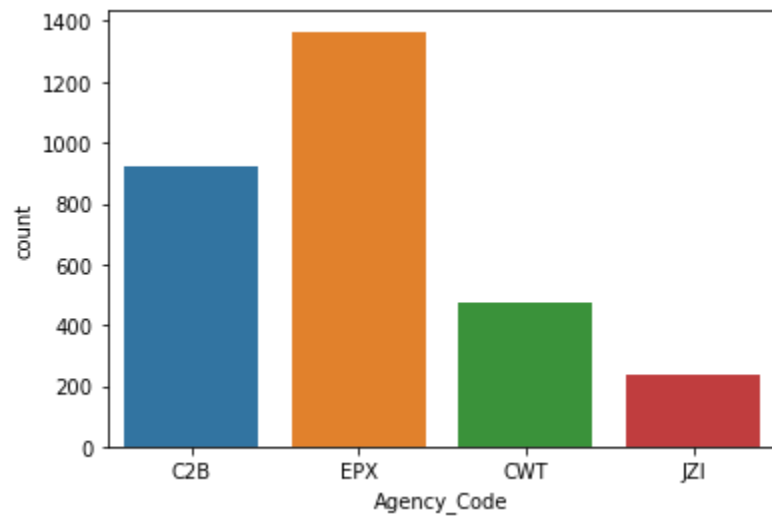
Univariate Analysis for Continuous variables-

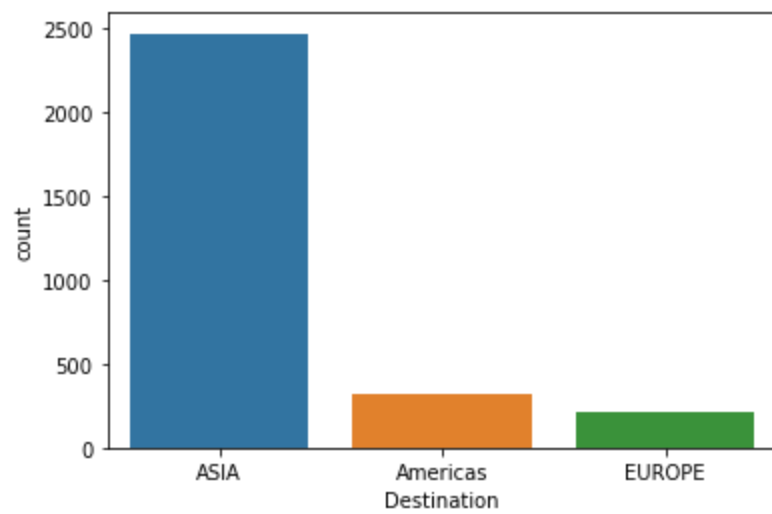
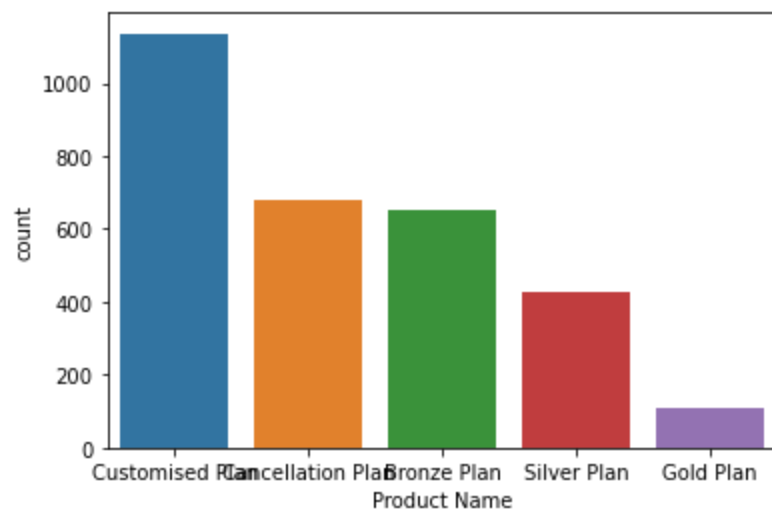
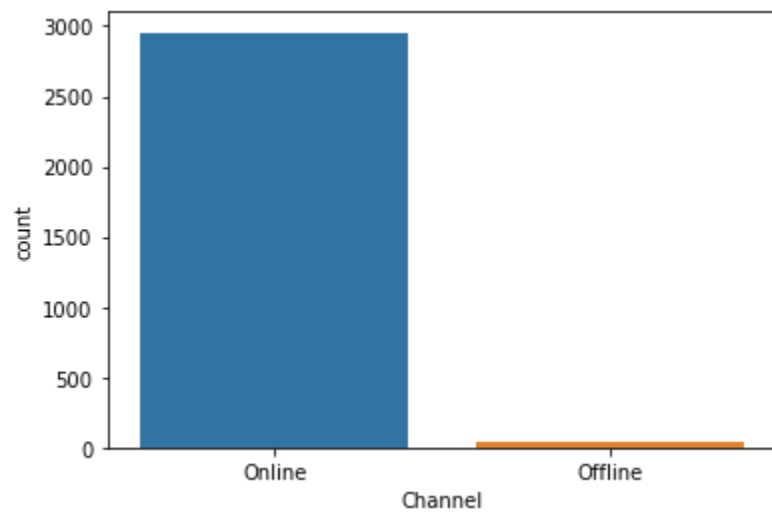
GRAPH 10



Univariate Analysis for Categorical variables-

GRAPH 11





Bivariate analysis:

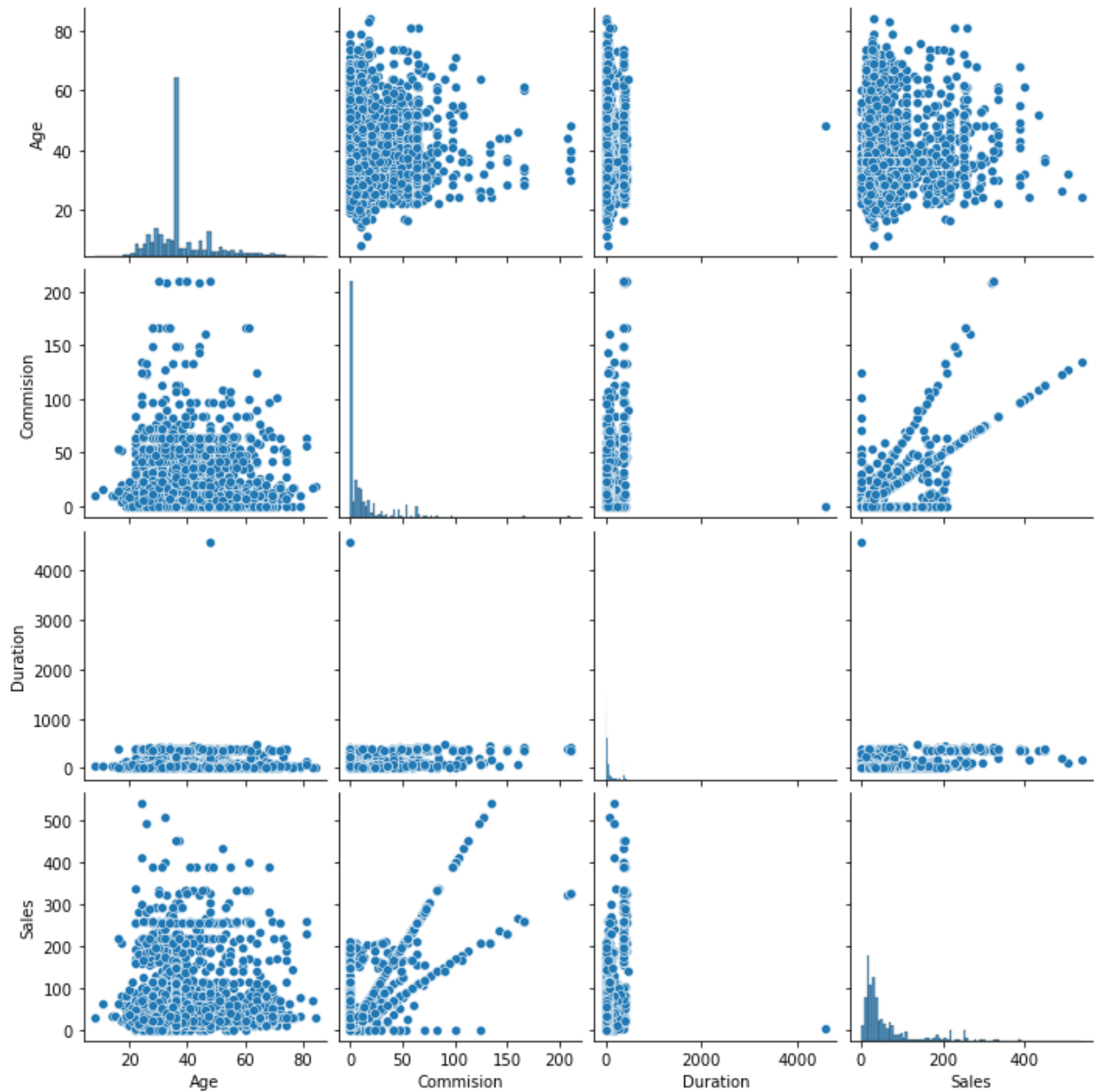
Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. ^[3]

The pairplot is generally used for numerical variables and box plots are used for categorical with numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. ^[4]

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed. ^[12]

GRAPH 12

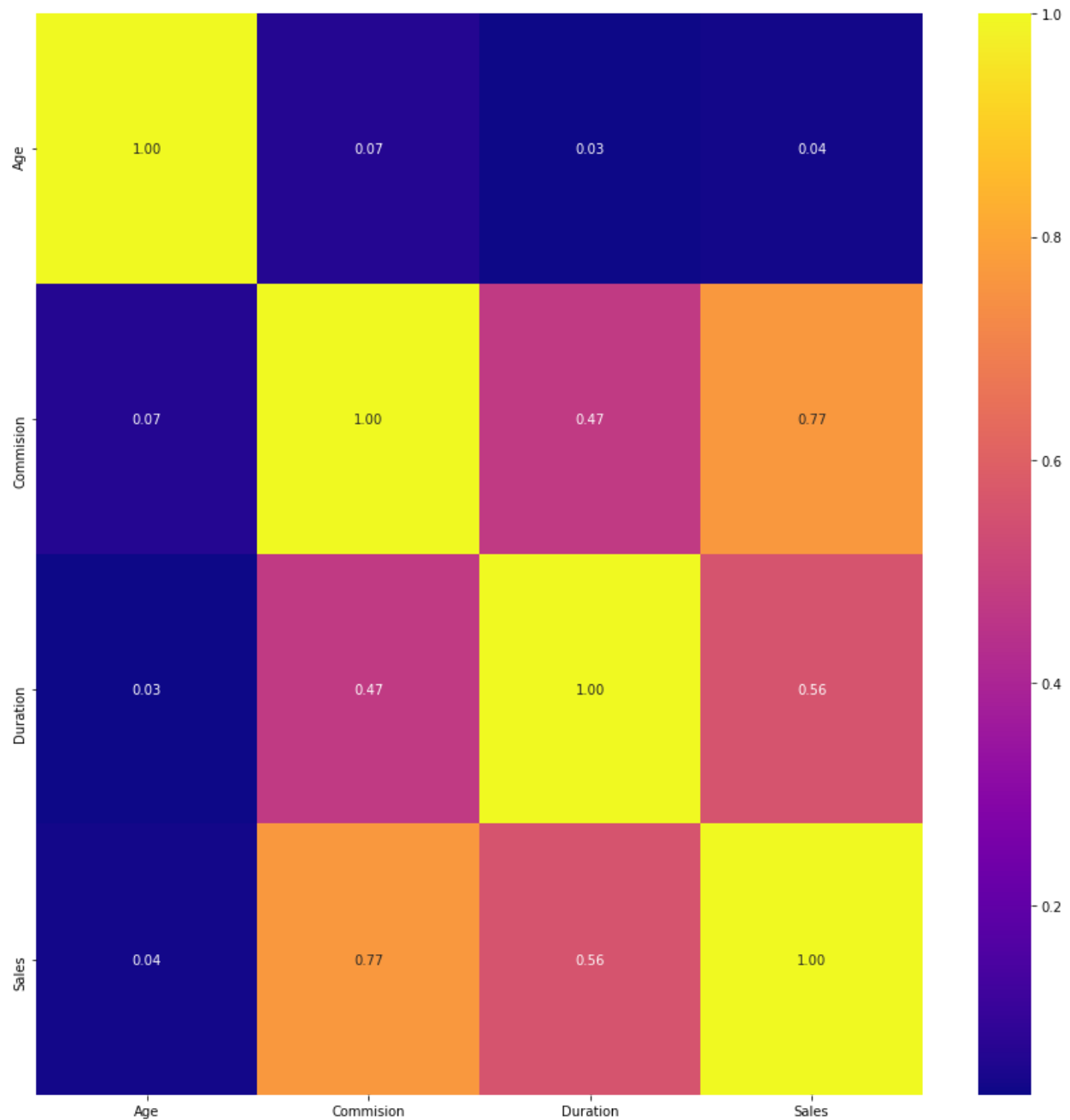


The below are the findings from the pairplot generated -

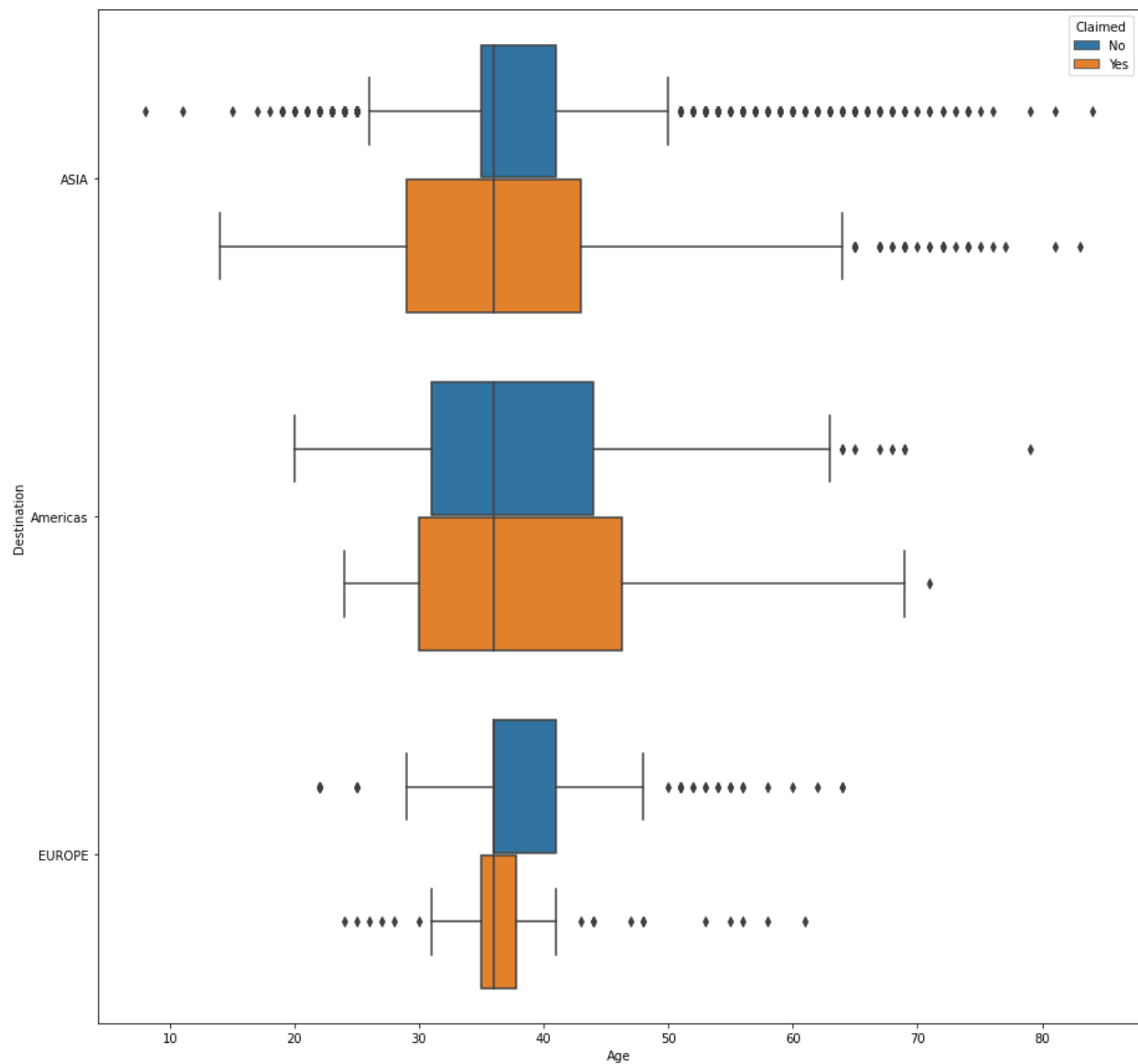
- The variable Sales and Commission are the variables that are correlated to each other.

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. But in the given data set none of the variables have a value 0.8 or more. The heat map for all the numerical variable is below,

GRAPH 13



GRAPH 14



The below are the findings from the boxplot generated -

- The people who visited Asia and America are the ones who have mostly claimed the insurance and they all are in their late 30s.
- The number of claims is low for the people who visited Europe.

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

CART:

A Classification And Regression Tree (CART), is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable. ^[13]

Random Forest:

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. ^[14]

Neural Networks:

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems. ^[15]

From the given data, 30 percent is taken for the test size and 70 percent is taken for the training. The target variable here is "Claimed". This is the same for all the three models taken into consideration.

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.

Accuracy – How accurately / cleanly does the model classify the data points. Lesser the false predictions, more the accuracy.

Sensitivity / Recall – How many of the actual True data points are identified as True data points by the model . Remember, False Negatives are those data points which should have been identified as True.

Specificity – How many of the actual Negative data points are identified as negative by the model

Precision – Among the points identified as Positive by the model, how many are really Positive

CART:

The below is the classification report of the training data-

FIGURE 19

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1471
1	0.69	0.51	0.58	629
accuracy			0.78	2100
macro avg	0.75	0.70	0.72	2100
weighted avg	0.77	0.78	0.77	2100

The below is the confusion matrix for training data-

FIGURE 20

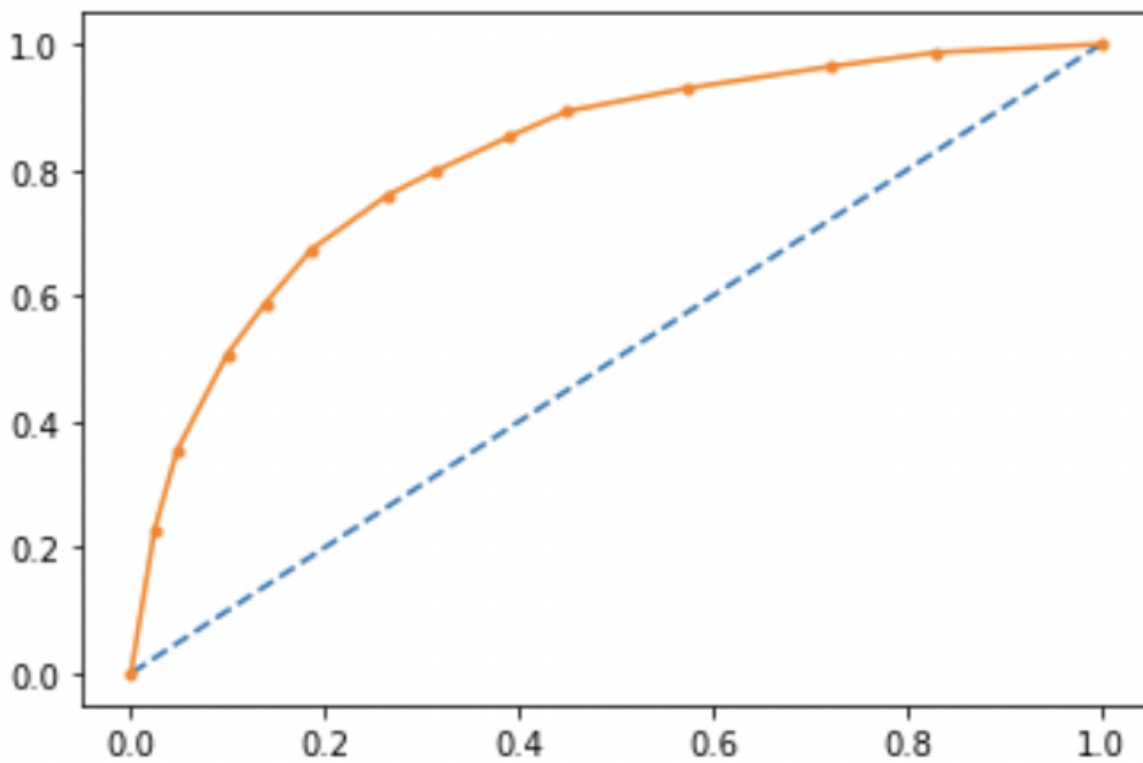
```
array([[1325, 146],
       [ 310, 319]])
```

The training data accuracy is 78.28%.

ROC curve and AUC score for training data-

GRAPH 15

AUC: 0.821



The below is the classification report of the testing data-

FIGURE 21

	precision	recall	f1-score	support
0	0.77	0.92	0.84	605
1	0.72	0.43	0.54	295
accuracy			0.76	900
macro avg	0.74	0.67	0.69	900
weighted avg	0.75	0.76	0.74	900

The below is the confusion matrix for testing data-

FIGURE 22

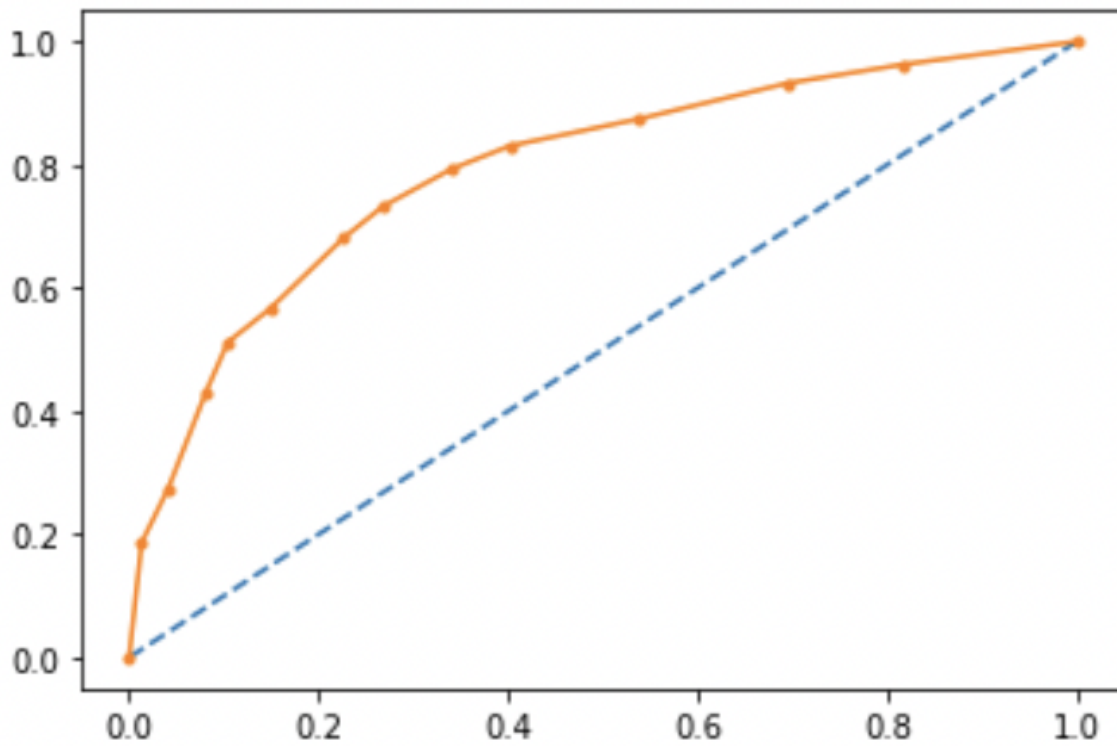
```
array([[556,  49],
       [168, 127]])
```

The training data accuracy is 75.88%.

ROC curve and AUC score for testing data-

GRAPH 16

AUC: 0.793



RANDOM FOREST:

The below is the classification report of the training data-

FIGURE 23

	precision	recall	f1-score	support
0	0.80	0.91	0.85	1471
1	0.70	0.46	0.56	629
accuracy			0.78	2100
macro avg	0.75	0.69	0.71	2100
weighted avg	0.77	0.78	0.76	2100

The below is the confusion matrix for training data-

FIGURE 24

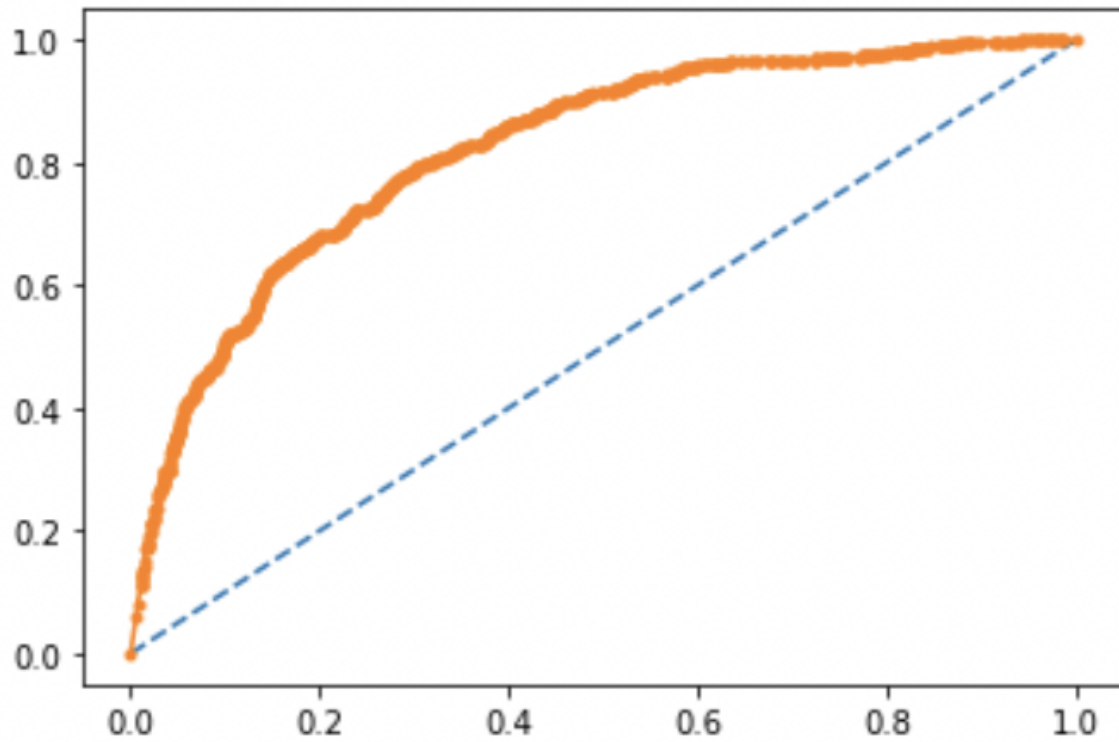
```
array([[1344, 127],
       [ 337, 292]])
```

The training data accuracy is 77.9%.

ROC curve and AUC score for training data-

GRAPH 17

AUC: 0.823



The below is the classification report of the testing data-

FIGURE 25

	precision	recall	f1-score	support
0	0.75	0.93	0.83	605
1	0.73	0.36	0.48	295
accuracy			0.75	900
macro avg	0.74	0.65	0.66	900
weighted avg	0.74	0.75	0.72	900

The below is the confusion matrix for testing data-

FIGURE 26

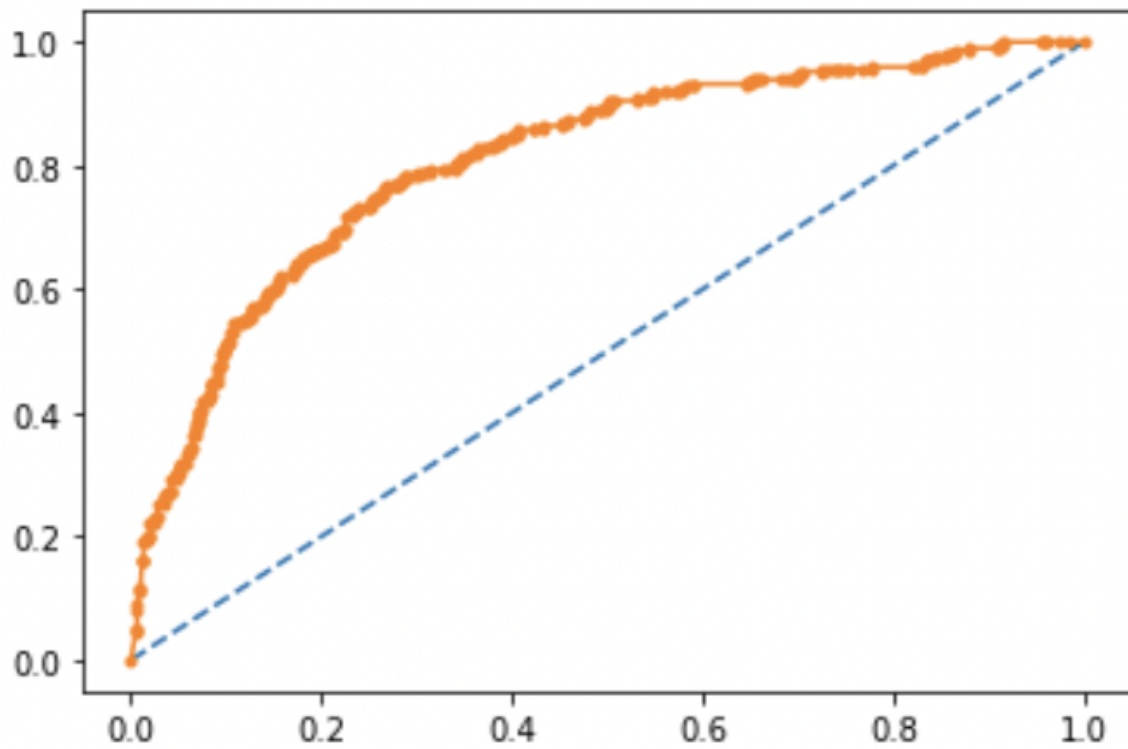
```
array([[565,  40],
       [188, 107]])
```

The testing data accuracy is 74.6%.

ROC curve and AUC score for testing data-

GRAPH 18

AUC: 0.810



NEURAL NETWORKS:

The below is the classification report of the training data-

FIGURE 27

	precision	recall	f1-score	support
0	0.76	0.96	0.85	1471
1	0.75	0.30	0.43	629
accuracy			0.76	2100
macro avg	0.75	0.63	0.64	2100
weighted avg	0.76	0.76	0.72	2100

The below is the confusion matrix for training data-

FIGURE 28

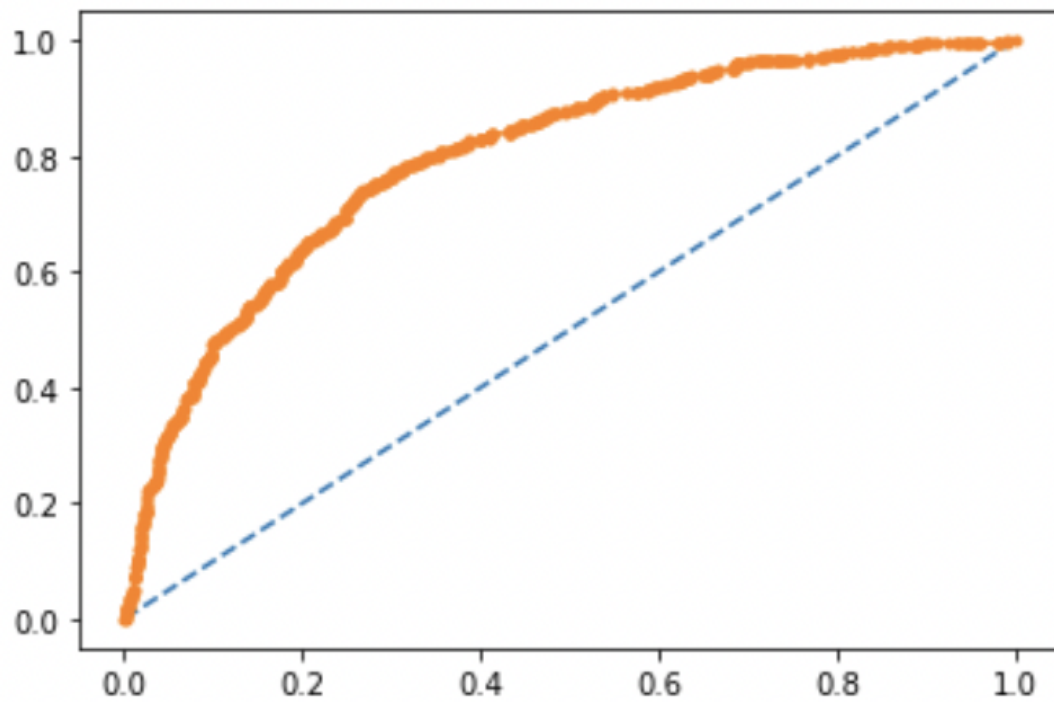
```
array([[1407, 64],
       [ 441, 188]])
```

The training data accuracy is 75.9%.

ROC curve and AUC score for training data-

GRAPH 19

AUC: 0.799



The below is the classification report of the testing data-

FIGURE 29

	precision	recall	f1-score	support
0	0.73	0.97	0.83	605
1	0.81	0.26	0.40	295
accuracy			0.74	900
macro avg	0.77	0.62	0.62	900
weighted avg	0.76	0.74	0.69	900

The below is the confusion matrix for testing data-

FIGURE 30

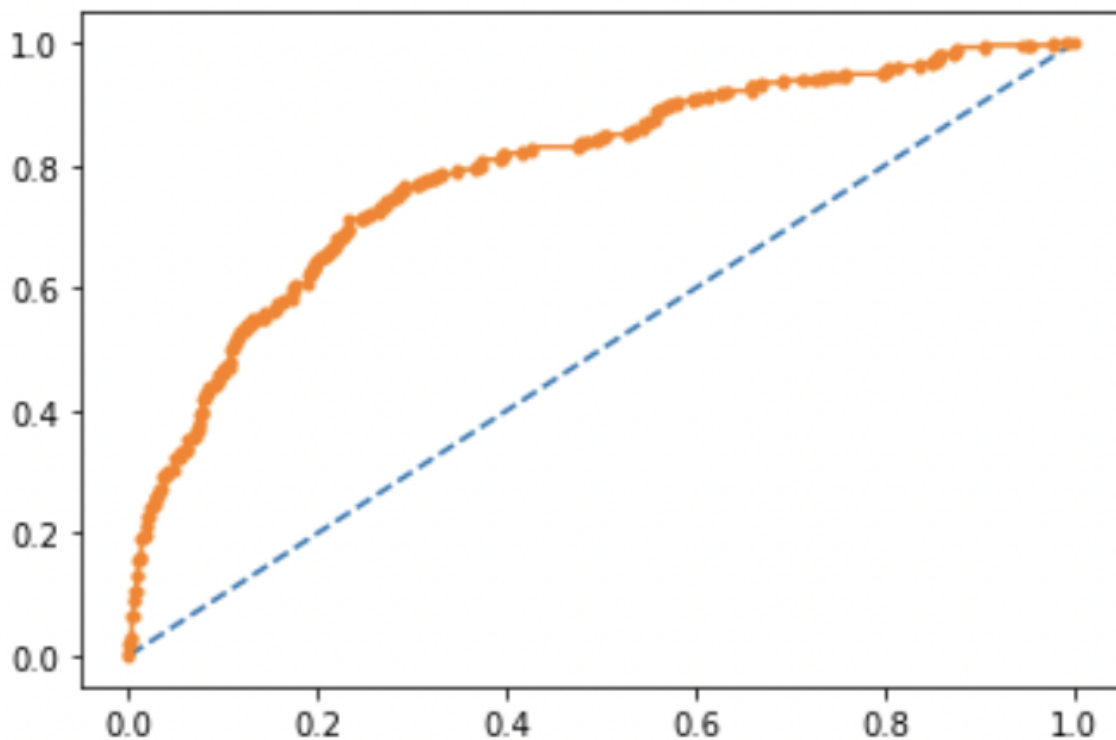
```
array([[587, 18],  
       [217, 78]])
```

The testing data accuracy is 73.8%.

ROC curve and AUC score for testing data-

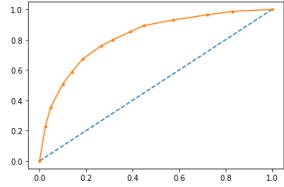
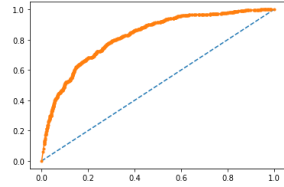
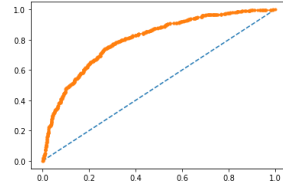
GRAPH 20

AUC: 0.792

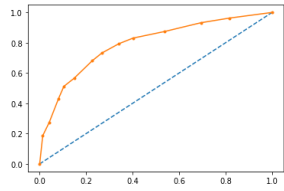
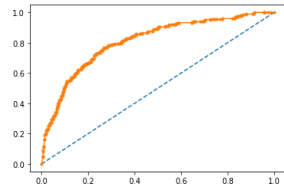
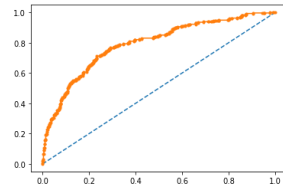


2.4 Final Model: Compare all the models and write an inference which model is best/optimized.

Training data-

	CART	Random Forest	Neural Network
Precision	0.69	0.70	0.75
Recall	0.51	0.46	0.30
F1 score	0.58	0.56	0.43
Accuracy	78.28%	77.9%	75.9%
AUC score	0.821	0.823	0.799
ROC curve			

Testing data-

	CART	Random Forest	Neural Network
Precision	0.72	0.73	0.81
Recall	0.43	0.36	0.26
F1 score	0.54	0.48	0.40
Accuracy	75.88%	74.6%	73.8%
AUC score	0.793	0.810	0.792
ROC curve			

Inference:

Based on the table of comparison displayed above, it is clear that the CART model has the best accuracy in both training and testing data. The F1 score is high which means both type 1 and type 2 errors are reduced. Also, the recall score is high in both training and testing which means type 2 error is low.

The Neural network has the highest Precision value which means type 1 error is low in this.

The Random forest has the highest Area under the curve (AUC).

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

These models can be used to predict the number of clients that will claim the insurance.

The people who visited Asia and America are the ones who have mostly claimed the insurance and they all are in their late 30s. In this case, we can focus more on the people who leave for these two continents.

The number of claims is low for the people who visited Europe so the company need not worry much about this group.

Majority of the claims are done in gold, silver and bronze plan. The age group also belongs to the late 30s and the 40s. The company should focus on these groups as well.

Also, the clients who have applied through offline channels should be given more focus as their claim is also high.

References

Websites-

- [1] <https://www.statisticshowto.com/univariate/>
- [2] <https://www.spss-tutorials.com/skewness/>
- [3] https://en.wikipedia.org/wiki/Bivariate_analysis
- [4] <https://pythonbasics.org/seaborn-pairplot/>
- [5] https://en.wikipedia.org/wiki/Feature_scaling
- [6] <https://www.displayr.com/what-is-hierarchical-clustering/>
- [7] <https://www.statistics.com/glossary/wards-linkage/>
- [8] <https://www.displayr.com/what-is-dendrogram/>
- [9] <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>
- [10] <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>
- [11] <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>
- [12] <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd5>
- [13] [https://wiki.q-researchsoftware.com/wiki/Machine_Learning_-_Classification_And_Regression_Trees_\(CART\)](https://wiki.q-researchsoftware.com/wiki/Machine_Learning_-_Classification_And_Regression_Trees_(CART))
- [14] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm
- [15] <https://www.investopedia.com/terms/n/neuralnetwork.asp>

End of Project