DATA MINING

PROJECT

SUMMARY ABOUT TWO DIFFERENT DATAS.

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.

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.

Yashveer Kothari. A
POST GRADUATE PROGRAM IN DATA SCIENCE AND BUSINESS ANALYTICS

TABLE OF CONTENTS		
CHAPTER/QUESTION #	DESCRIPTION	PAGE #

LIST OF TABLES AND CHARTS

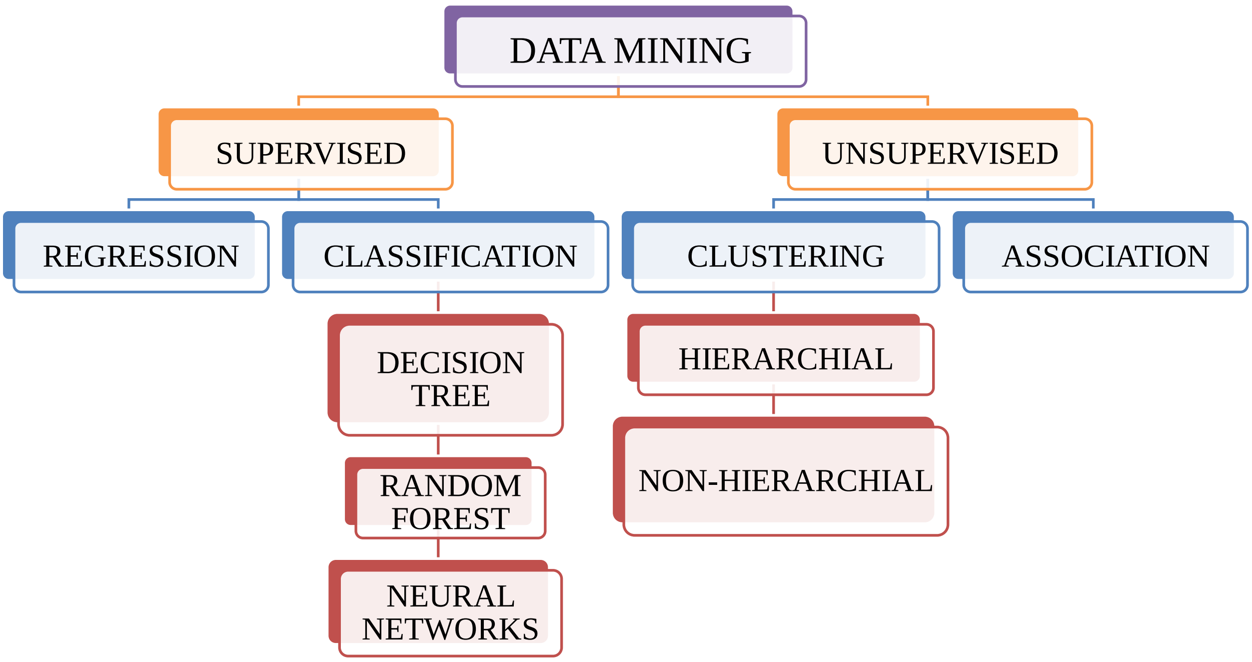
PROBLEM 1: CLUSTERING

LIST OF FIGURES		
PROBLEM 1: CLUSTERING		
Tables #	Topic	Page #

LIST OF FIGURES		
PROBLEM 2: CART-RF-ANN		
Tables #	Topic	Page #

ABOUT DATA MINING

- Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.



- **SUPERVISED CLASSIFICATION:**

Refers to learning algorithms that are used in classification and prediction, the variables are defined clearly. In Supervised classification Train and test cycles and Model evaluation methods are available, helps in decision making.

- **UNSUPERVISED CLASSIFICATION:**

Unsupervised learning uses machine learning algorithms to analyse and cluster unlabelled data sets. These algorithms discover hidden patterns in data without the need for human intervention.

PROBLEM 1

- **CLUSTERING:**

1. It is a part of Unsupervised Learning. It is a technique of the grouping object with heterogeneity between groups and homogeneity within the groups.
2. It can follow Agglomerative, Divisive or partitioning approach. Distance calculations is done to find similarity and dissimilarity in clustering problems.
3. Types:

   I. Hierarchical:

   A. Agglomerative: It has a bottom-top approach, it starts with objects forming separate group. Keeps merging the objects or groups that are close to one another. Identifies even the small size clusters.

   B. Divisive: It has top-bottom approach, starts with all object in the same clusters. Splits up on small clusters.

   II. Partitioning:

   A. K-Means: Constructs “K” partitions and each partition will represent a cluster where K<=n.

4. Measuring Distances:

I. Euclidean distance = $d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$.

II. Manhattan Distance = $| x_2 - x_1 | + | y_2 - y_1 |$

III. Chebyshev Distance = max (y2 - y1, x2 - x1).

IV. Minkowski Distance = (sum for i to N (abs(v1[i] — v2[i])) ^p) ^(1/p)

- **HIERARCHICAL CLUSTERING:**

1. Hierarchical clustering produces useful graphical display of the clustering process and results called Dendrogram.
2. Records are grouped sequentially to created clusters
3. Based on the distance between the records clusters are made.

INTRODUCTION

The dataset contains data about there customers or users during the past few months, to understand the customer's activities based on their credit card usage from which they plan to do a customer segmentation to give out offers. From this analysis our aim is to explore the data set, perform clustering using Hierarchical and K-means clustering, Extract the dendrogram, extract the Silhouette score and width and understand the different ways to promote the offers to various customer segments.

DATA DICTIONARY:

1. Spending: Amount spent by the customer per month (in 1000s)
2. Advance payments: Amount paid by the customer in advance by cash (in 100s)
3. Probability of full payment: Probability of payment done in full by the customer to the bank
4. current balance: Balance amount left in the account to make purchases (in 1000s)
5. credit limit: Limit of the amount in credit card (10000s)
6. Min payment amt: minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. Max spent in single shopping: Maximum amount spent in one purchase (in 1000s)

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

TABLE 1: TOP 5 SAMPLES

	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

TABLE 2: LAST 5 SAMPLES

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
205	13.89	14.02	0.8880	5.439	3.199	3.986	4.738
206	16.77	15.62	0.8638	5.927	3.438	4.920	5.795
207	14.03	14.16	0.8796	5.438	3.201	1.717	5.001
208	16.12	15.00	0.9000	5.709	3.485	2.270	5.443
209	15.57	15.15	0.8527	5.920	3.231	2.640	5.879

TABLE 3: INFORMATION ABOUT THE DATASET:

```
 #   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
```

TABLE 4: DESCRIPTION OF THE DATASET:

	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

TABLE 5: MISSING RECORDS IN THE DATA SET

```
spending                        0
advance_payments                0
probability_of_full_payment     0
current_balance                 0
credit_limit                    0
min_payment_amt                 0
max_spent_in_single_shopping    0
dtype: int64
```

There are no missing values in the dataset

- Based on the above tables, the following can be inferred:

1. There are a total of 7 variables and 210 records in Data set
2. No missing record based on initial analysis.
3. All the variables are numeric type
4. Data shape is 210 rows and 7 columns.
5. Based on descriptive summary, the data looks good.
6. We see for most of the variable, mean/medium are nearly equal
7. Include a 90% to see variations and it looks distributed evenly
8. Std Deviation is high for spending variable.

CHECKING FOR ANY DUPLICATE RECORDS IN DATA SET

- There are no duplicate records in the dataset

UNIVARIATE ANALYSIS FOR ALL THE VARIABLES IN THE DATA SET

- Univariate analysis is defined as analysis carried out on only one variable to

TABLE 6: SPENDING -DESCRIPTION

summarize or describe the variable

```
Description of spending
------------------------
count    210.000000
mean      14.847524
std        2.909699
min       10.590000
25%       12.270000
50%       14.355000
75%       17.305000
max       21.180000
```

FIGURE 1: SPENDING -DISTRIBUTION PLOT

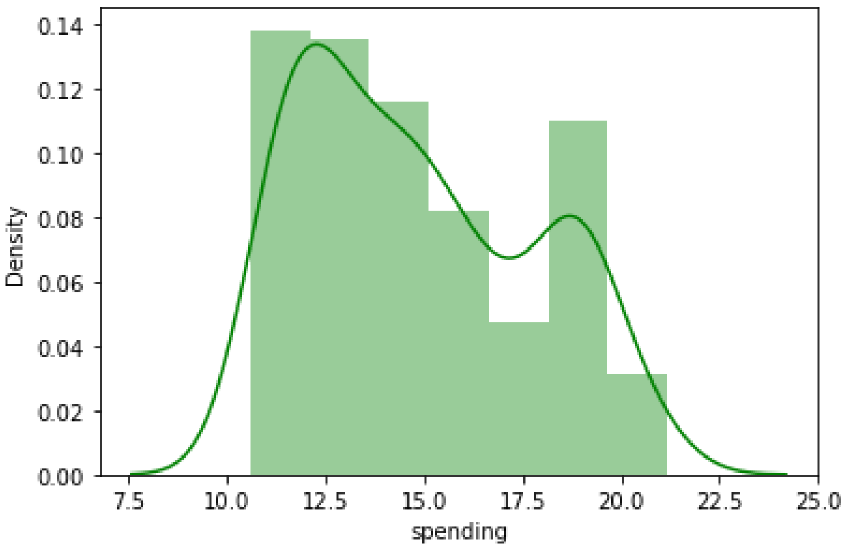
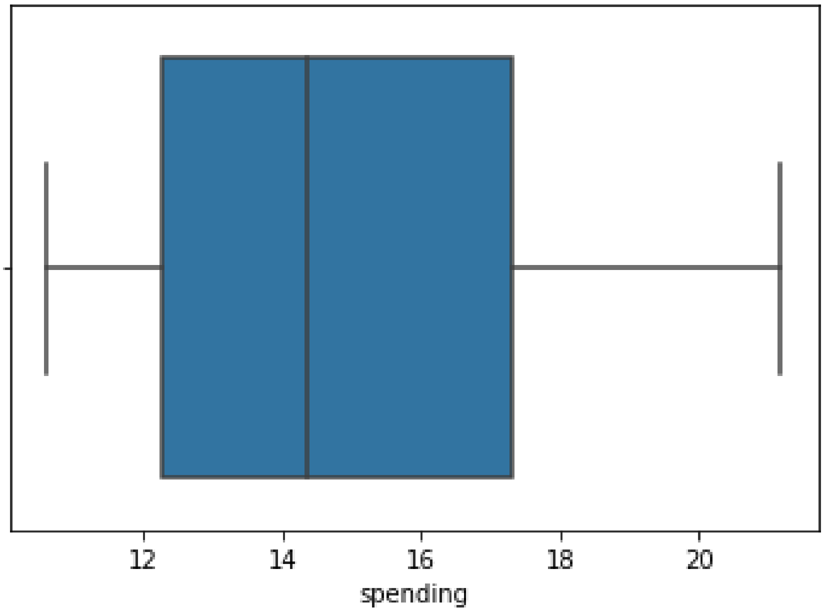


FIGURE 2: SPENDING- BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
spending - 1st Quartile (Q1) is:  12.27
spending - 3st Quartile (Q3) is:  17.305
Interquartile range (IQR) of spending is  5.035

Extracting the Outliers
Lower outliers in spending:  4.717499999999999
Upper outliers in spending:  24.8575

Extracting the number of outliers in spending Variable
Number of outliers in spending upper :  0
Number of outliers in spending lower :  0
% of Outlier in spending upper:  0 %
% of Outlier in spending lower:  0 %
```

INFERENCE FOR SPENDING

1. Range of values in spending variable is 10.592 (Max-Min = 1.18-10.59)
2. Minimum spending: 10.59
3. Maximum spending: 21.18
4. Mean value: 14.84
5. Median value: 14.35 (Q2)
6. Standard deviation: 2.90
7. There are no outliers in spending variable as per the boxplot and the above output.
8. The distribution is not distributed normally.
9. The Inter Quartile range as per the output is 5.03 for spending variable

TABLE 7: ADVANCE PAYMENTS- DESCRIPTION

```
Description of advance_payments
--------------------------------
count    210.000000
mean      14.559286
std        1.305959
min       12.410000
25%       13.450000
50%       14.320000
75%       15.715000
max       17.250000
```

FIGURE 3: ADVANCE PAYMENTS- DISTRIBUTION PLOT

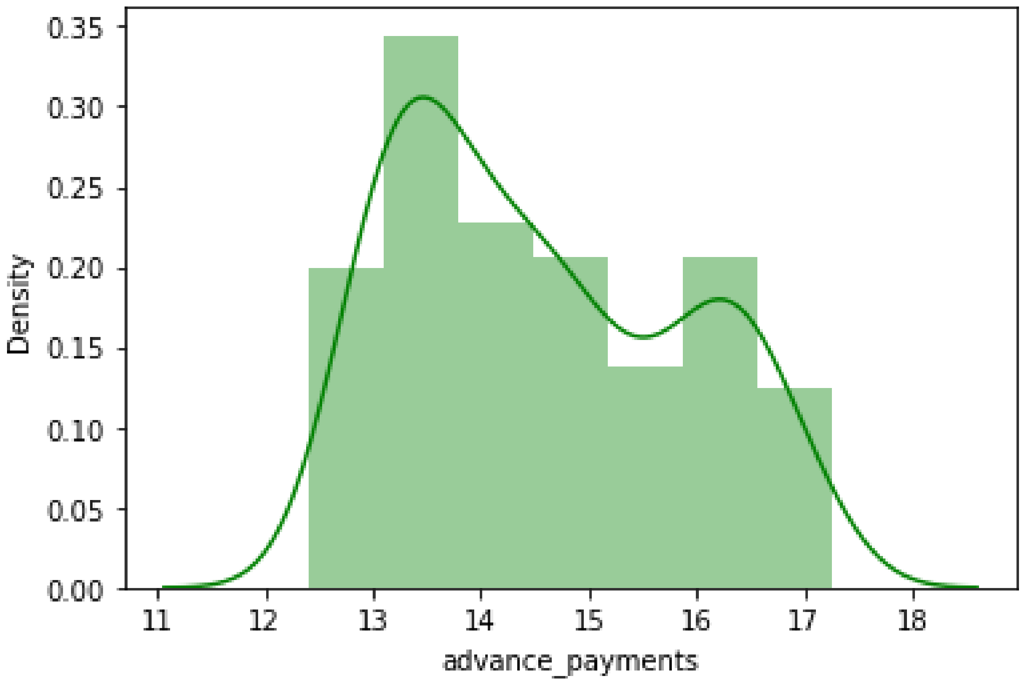
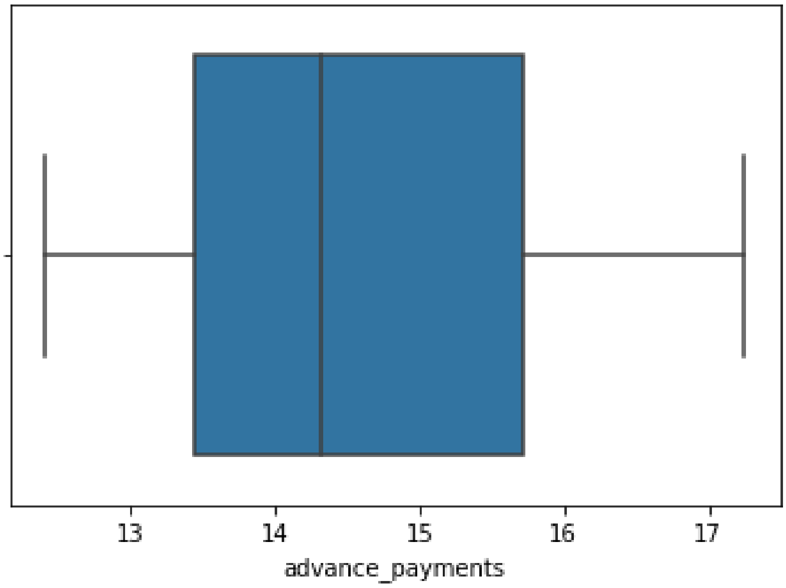


FIGURE 4: ADVANCE PAYMENTS- BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
advance_payments  - 1st Quartile (Q1) is:  13.45
advance_payments  - 3st Quartile (Q3) is:  15.715
Interquartile range (IQR) of advance_payments  is  2.2650000000000006

Extracting the Outliers
Lower outliers in advance_payments :  10.052499999999998
Upper outliers in advance_payments :  19.1125

Extracting the number of outliers in advance_payments Variable
Number of outliers in advance_payments  upper :  0
Number of outliers in advance_payments  lower :  0
% of Outlier in advance_payments  upper:  0 %
% of Outlier in advance_payments  lower:  0 %
```

INFERENCE FOR ADVANCE PAYMENTS

1. Range of values: 4.84
2. Minimum advance payments: 12.41
3. Maximum advance payments: 17.25
4. Mean value: 14.559285714285727
5. Median value: 14.32
6. Standard deviation: 1.30
7. The Interquartile range for advance payments as per the above output is 2.26.

8. The advance payments variable does not have any outliers as per the box plot and the above output.
9. The variable is slightly not distributed normally.

TABLE 8: PROBABILITY OF FULL PAYMENTS- DESCRIPTION

```
Description of probability_of_full_payment
------------------------------------------
count    210.000000
mean       0.870999
std        0.023629
min        0.808100
25%        0.856900
50%        0.873450
75%        0.887775
max        0.918300
```

FIGURE 5: PROBABILITY OF FULL PAYMENTS- DISTRIBUTION PLOT

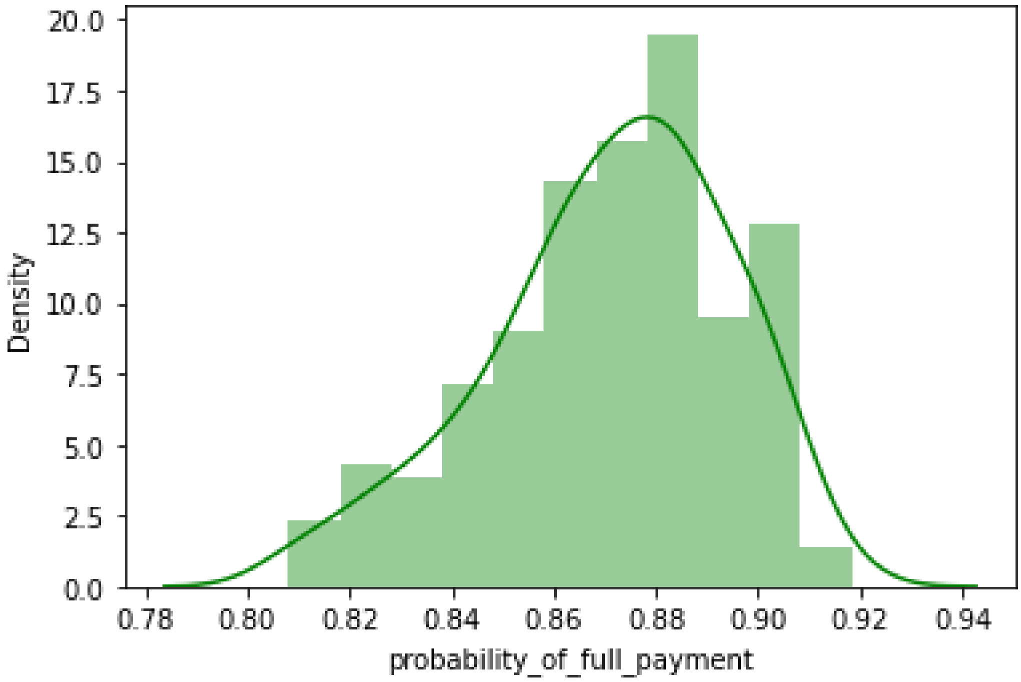
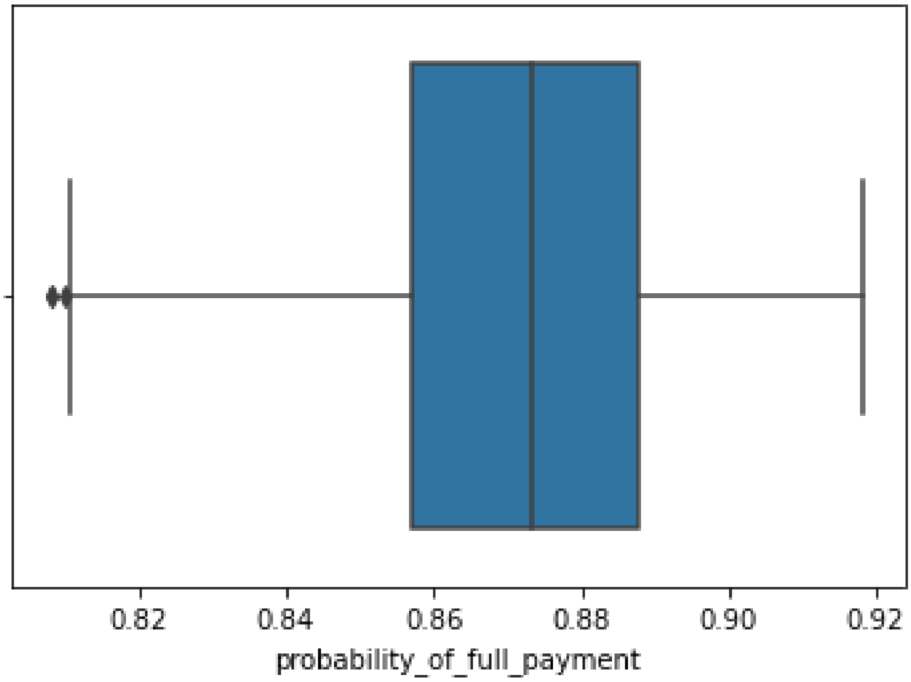


FIGURE 6: PROBABILITY OF FULL PAYMENTS- BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
probability_of_full_payment  - 1st Quartile (Q1) is:  0.8569
probability_of_full_payment  - 3st Quartile (Q3) is:  0.887775
Interquartile range (IQR) of probability_of_full_payment  is  0.030874999999999986

Extracting the Outliers
Lower outliers in probability_of_full_payment :  0.8105875
Upper outliers in probability_of_full_payment :  0.9340875

Extracting the number of outliers in probability_of_full_payment Variable
Number of outliers in probability_of_full_payment  upper :  0
Number of outliers in probability_of_full_payment  lower :  3
% of Outlier in probability_of_full_payment  upper:  0 %
% of Outlier in probability_of_full_payment  lower:  1 %
```

INFERENCE FOR PROBABILITY FOR FULL PAYMENT

1. Range is 0.11
2. Minimum probability of full payment:0.8081
3. Maximum probability of full payment:0.9183
4. Mean value: 0.87
5. Median value: 0.87

6. Standard deviation:0.02
7. Interquartile range (IQR) of probability of full payment is 0.03
8. The distribution plot shows that the variable is normally distributed forming a bell cur ve.
9. There are outliers in the variable as per the boxplot and above output, there are 3 outli ers behind the lower whisker.

TABLE 9: CURRENT BALANCE- DESCRIPTION

Description of current_balance

count	210.000000
mean	5.628533
std	0.443063
min	4.899000
25%	5.262250
50%	5.523500
75%	5.979750
max	6.675000

FIGURE 7: CURRENT BALANCE- DISTRIBUTION PLOT

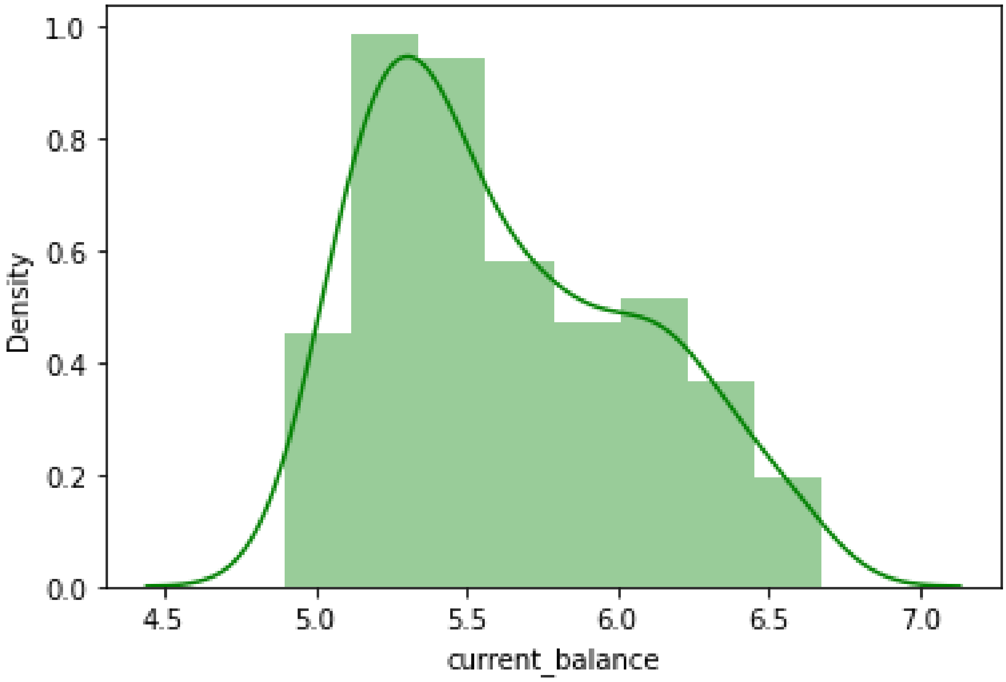
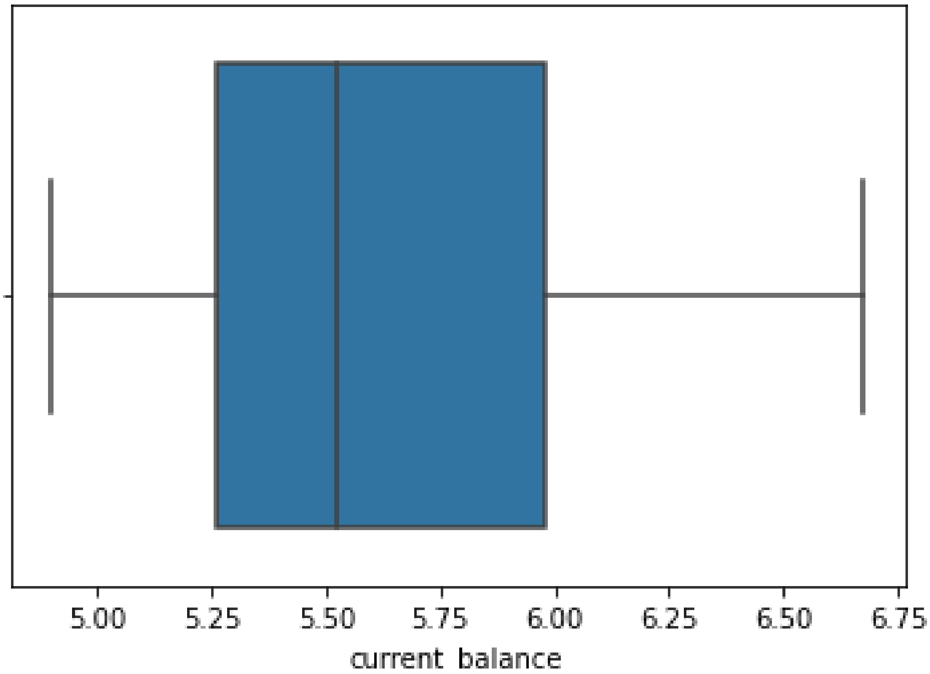


FIGURE 8: CURRENT BALANCE - BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
current_balance  - 1st Quartile (Q1) is:  5.26225
current_balance  - 3st Quartile (Q3) is:  5.97975
Interquartile range (IQR) of current_balance  is  0.7175000000000002

Extracting the Outliers
Lower outliers in current_balance :  4.186
Upper outliers in current_balance :  7.056000000000001

Extracting the number of outliers in current_balance Variable
Number of outliers in current_balance  upper :  0
Number of outliers in current_balance  lower :  0
% of Outlier in current_balance  upper:  0 %
% of Outlier in current_balance  lower:  0 %
```

INFERENCE FOR CURRENT BALANCE

1. Range is: 1.77
2. Minimum current balance: 4.899
3. Maximum current balance: 6.675
4. Mean value: 5.628533333333334
5. Median value: 5.5235
6. Standard deviation: 0.44306347772644 93
7. There are no outliers in variable current balance, as per the boxplot and output above.

8. The Inter Quartile range for variable Current balance is 0.71
9. There is a slight deviation in the distplot otherwise the variable mis normally distribut ed.

TABLE 10: CREDIT LIMIT- DESCRIPTION

```
Description of credit_limit
---------------------------
count    210.000000
mean       3.258605
std        0.377714
min        2.630000
25%        2.944000
50%        3.237000
75%        3.561750
max        4.033000
```

FIGURE 9: CREDIT LIMIT – DISTRIBUTION PLOT

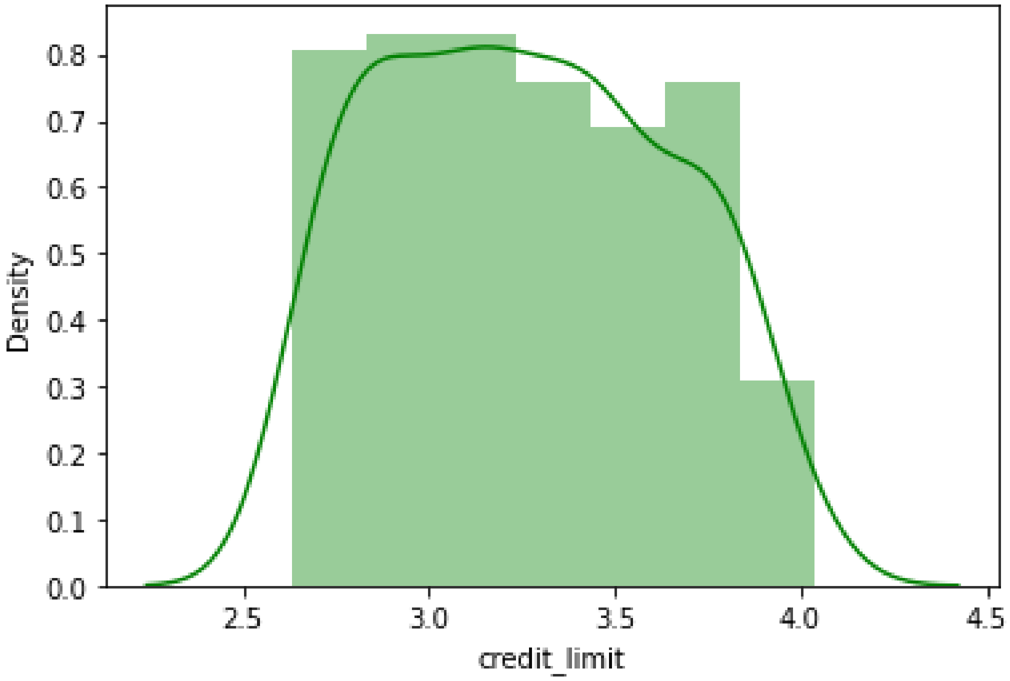
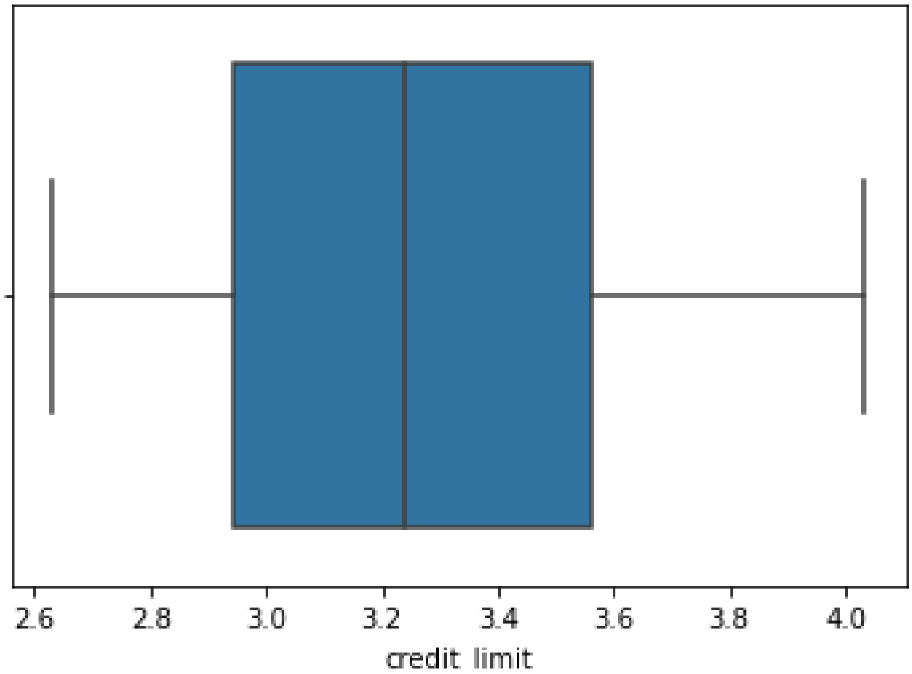


FIGURE 10: CREDIT LIMIT - BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
credit_limit  - 1st Quartile (Q1) is:  2.944
credit_limit  - 3st Quartile (Q3) is:  3.56175
Interquartile range (IQR) of credit_limit  is  0.61775

Extracting the Outliers
Lower outliers in credit_limit :  2.017375
Upper outliers in credit_limit :  4.488375

Extracting the number of outliers in credit_limit Variable
Number of outliers in credit_limit  upper :  0
Number of outliers in credit_limit  lower :  0
% of Outlier in credit_limit  upper:  0 %
% of Outlier in credit_limit  lower:  0 %
```

INFERENCE FOR CREDIT LIMIT

From the above, regarding the credit limit variable we infer the following:

1. Range is: 1.43
2. Minimum credit limit: 2.63
3. Maximum credit limit: 4.033
4. Mean value: 3.258604761904763
5. Median value: 3.237

6. Standard deviation: 0.3777144449065874
7. There are no outliers in this variable as per the box plot and the above output.
8. Interquartile range (IQR) of credit limit is 0.61775
9. The variable is normally distributed as per the Distribution Plot.

TABLE 11: MIN PAYMENT AMT- DESCRIPTION

Description of min_payment_amt

count	210.000000
mean	3.700201
std	1.503557
min	0.765100
25%	2.561500
50%	3.599000
75%	4.768750
max	8.456000

FIGURE 11: MIN PAYMENT AMT – DISTRIBUTION PLOT

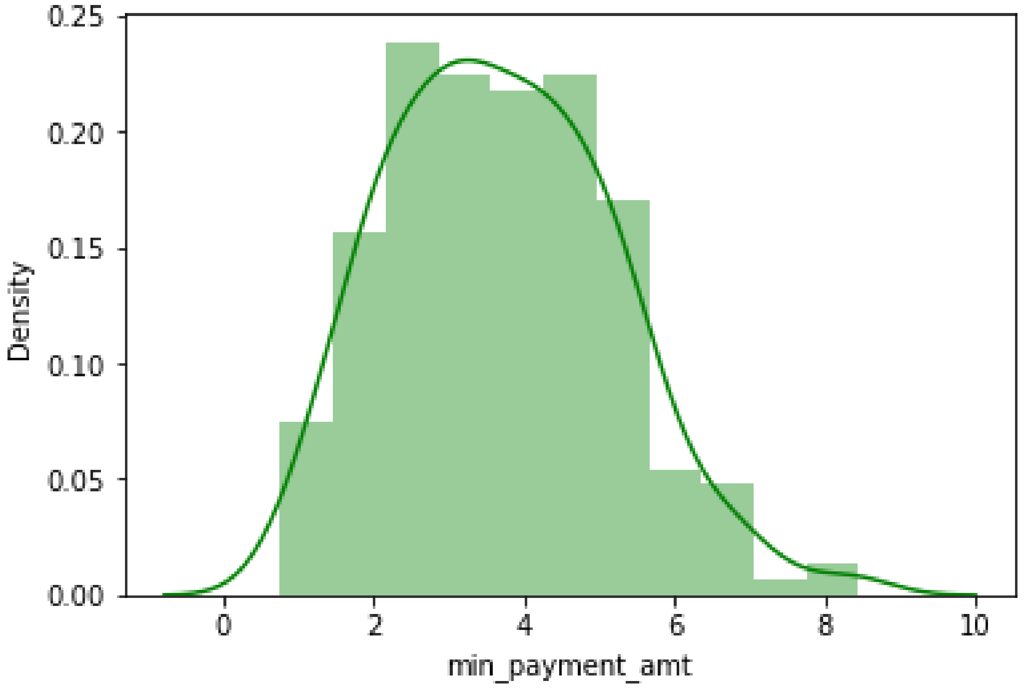
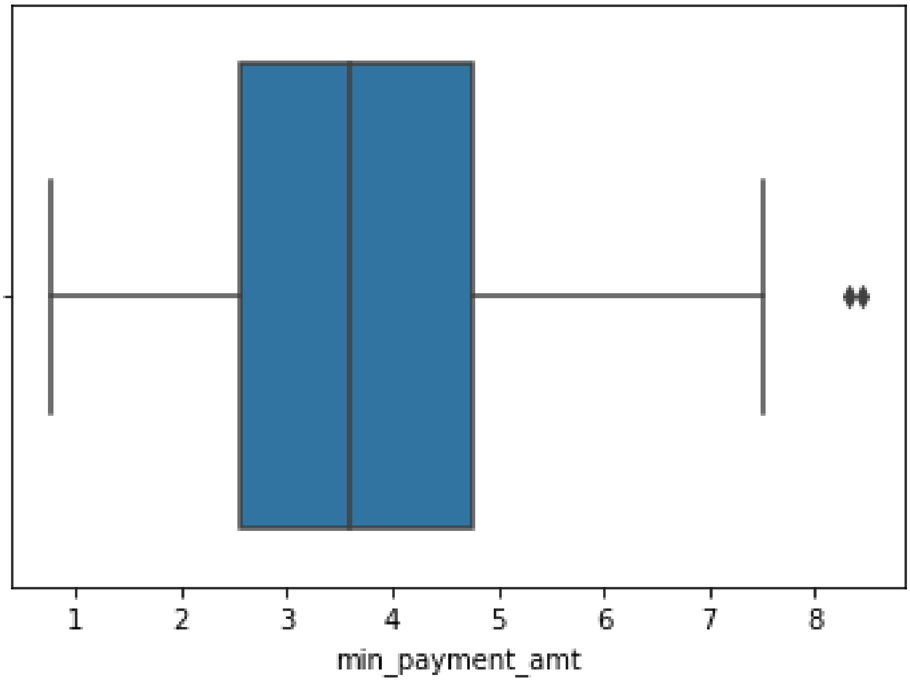


FIGURE 12: MIN PAYMENT AMT – BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
min_payment_amt  - 1st Quartile (Q1) is:  2.5615
min_payment_amt  - 3st Quartile (Q3) is:  4.76875
Interquartile range (IQR) of min_payment_amt  is  2.2072499999999997

Extracting the Outliers
Lower outliers in min_payment_amt :  -0.7493749999999992
Upper outliers in min_payment_amt :  8.079625

Extracting the number of outliers in min_payment_amt Variable
Number of outliers in min_payment_amt  upper :  2
Number of outliers in min_payment_amt  lower :  0
% of Outlier in min_payment_amt  upper:  1 %
% of Outlier in min_payment_amt  lower:  0 %
```

INFERENCE FOR MIN PAYMENT AMT

From the above, min payment amt variable we infer the following:

1. Range is 7.69
2. Minimum min payment amt: 0.7651
3. Maximum min payment amt: 8.456
4. Mean value: 3.7002009523809507

5. Median value: 3.599
6. Standard deviation: 1.5035571308217792
7. Interquartile range (IQR) of min payment amt is 2.20725
8. There are two outliers after the top whisker as per the output, and we also infer the same from the boxplot.
9. There is negligible deviation in the distribution plot, hence the data in the variable is normally distributed as it forms a bell curve.

TABLE 12: MAX SPENT IN SINGLE SHOPPING- DESCRIPTION

```
Description of max_spent_in_single_shopping
-------------------------------------------
count    210.000000
mean       5.408071
std        0.491480
min        4.519000
25%        5.045000
50%        5.223000
75%        5.877000
max        6.550000
```

FIGURE 13: MAX AMT SPENT IN SINGLE SHOPPING – DISTRIBUTION PLOT

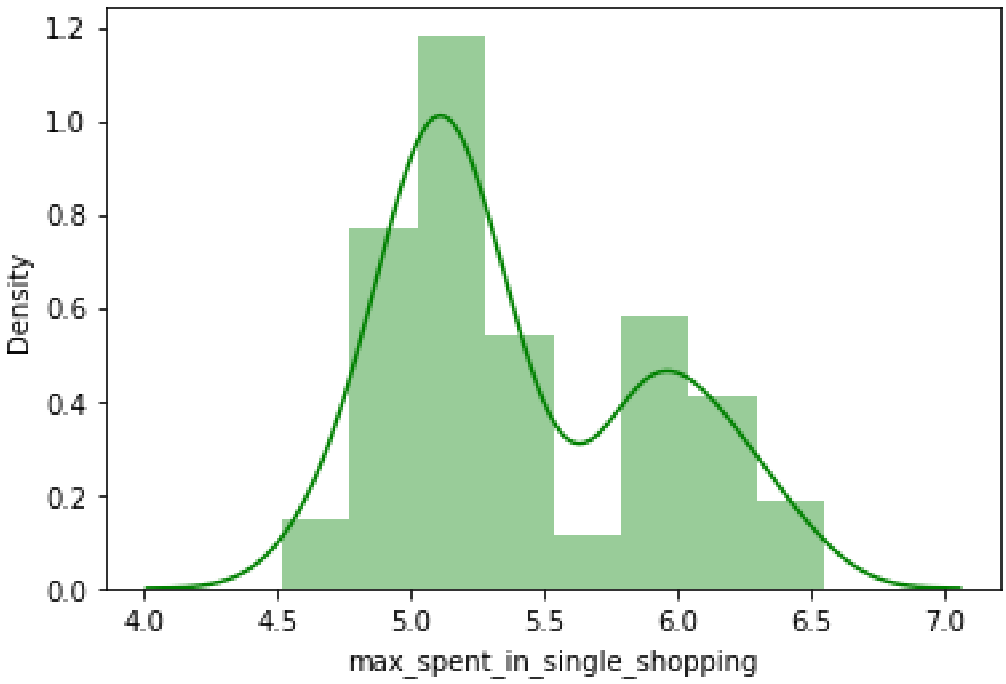
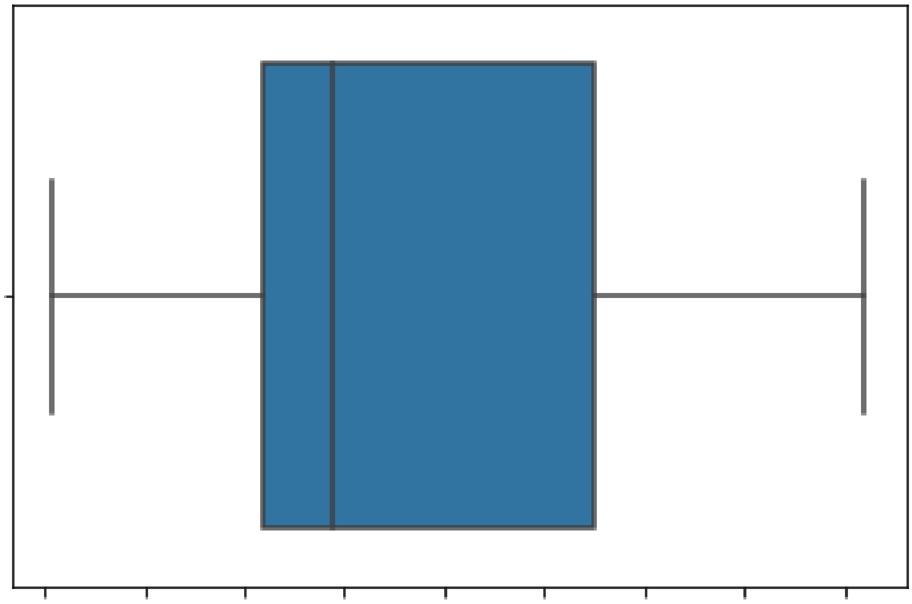


FIGURE 14: MAX AMT SPENT IN SINGLE SHOPPING– BOXPLOT



OUTPUT: IQR AND OUTLIER EXTRACTION

```
Extracting the Inter Quartile Range
max_spent_in_single_shopping  - 1st Quartile (Q1) is:  5.045
max_spent_in_single_shopping  - 3st Quartile (Q3) is:  5.877
Interquartile range (IQR) of max_spent_in_single_shopping  is  0.8319999999999999

Extracting the Outliers
Lower outliers in max_spent_in_single_shopping :  3.797
Upper outliers in max_spent_in_single_shopping :  7.125

Extracting the number of outliers in max_spent_in_single_shopping Variable
Number of outliers in max_spent_in_single_shopping  upper :  0
Number of outliers in max_spent_in_single_shopping  lower :  0
% of Outlier in max_spent_in_single_shopping  upper:  0 %
% of Outlier in max_spent_in_single_shopping  lower:  0 %
```

INFERENCE FOR MAX SPENT IN SINGLE SHOPPING

1. Range is: 7.69
2. Minimum max spent in single shopping: 4.519
3. Maximum max spent in single shopping: 6.55
4. Mean value: 5.408071428571429
5. Median value: 5.223000000000001
6. Standard deviation: 0.4914804991024054
7. Interquartile range (IQR) of min payment amt is 0.83
8. There are no outliers in this variable asper the output and the boxplot above

9. The variable is not normally distributed as there is no bell curve shape in the distribution plot, there are deviation in the plot.

FIGURE 15: HISTOGRAM FOR ALL VARIABLESIN THE DATASET

- A histogram can be used whenever there's a need to display a comparison of the distribution of certain numerical data in various ranges of intervals. Histogram examples can help an to see and understand quickly and easily essential meanings and patterns related to a large amount of data. They can be a benefit to a company's or organization's process of decision-making in various departments.

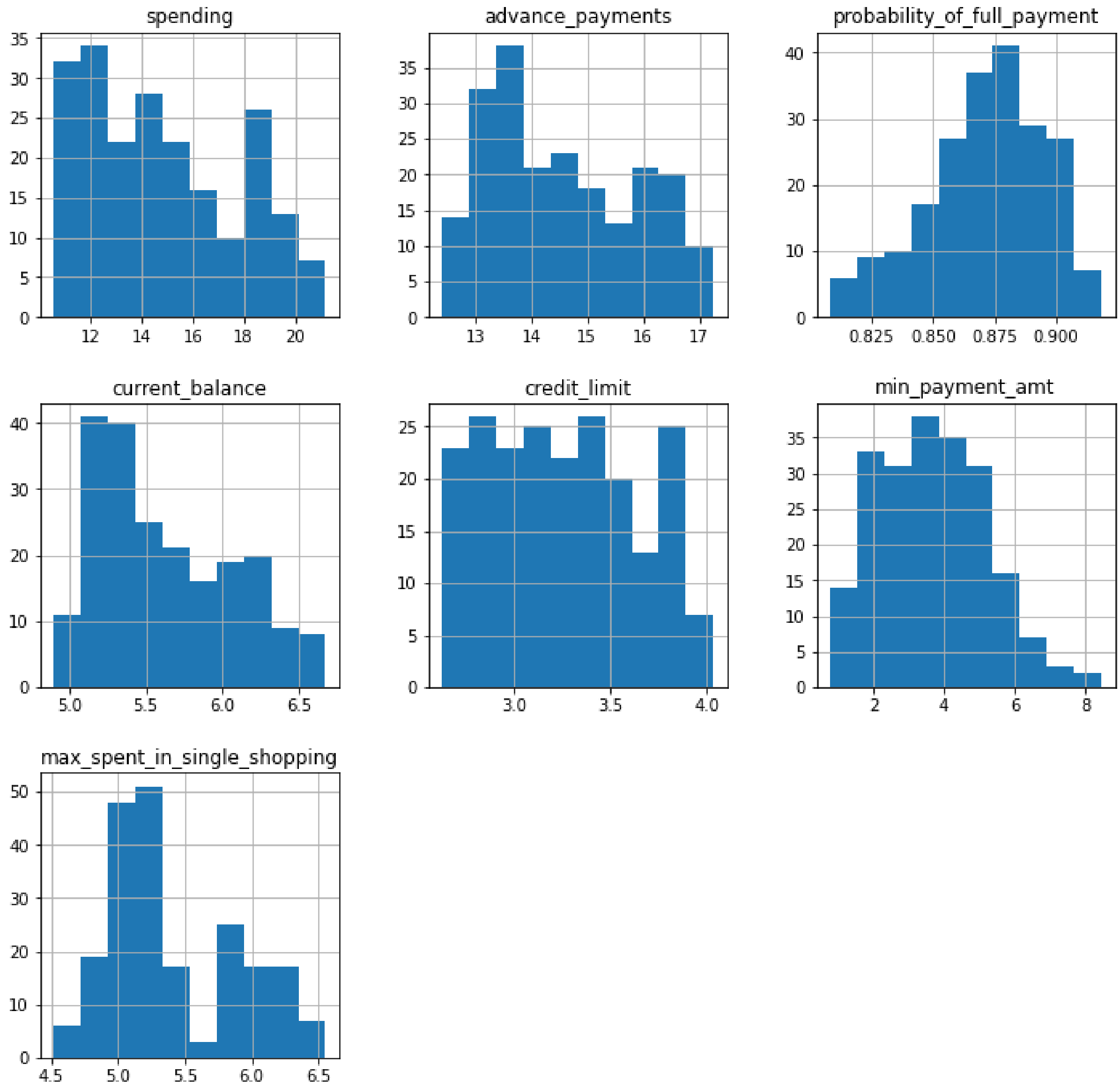
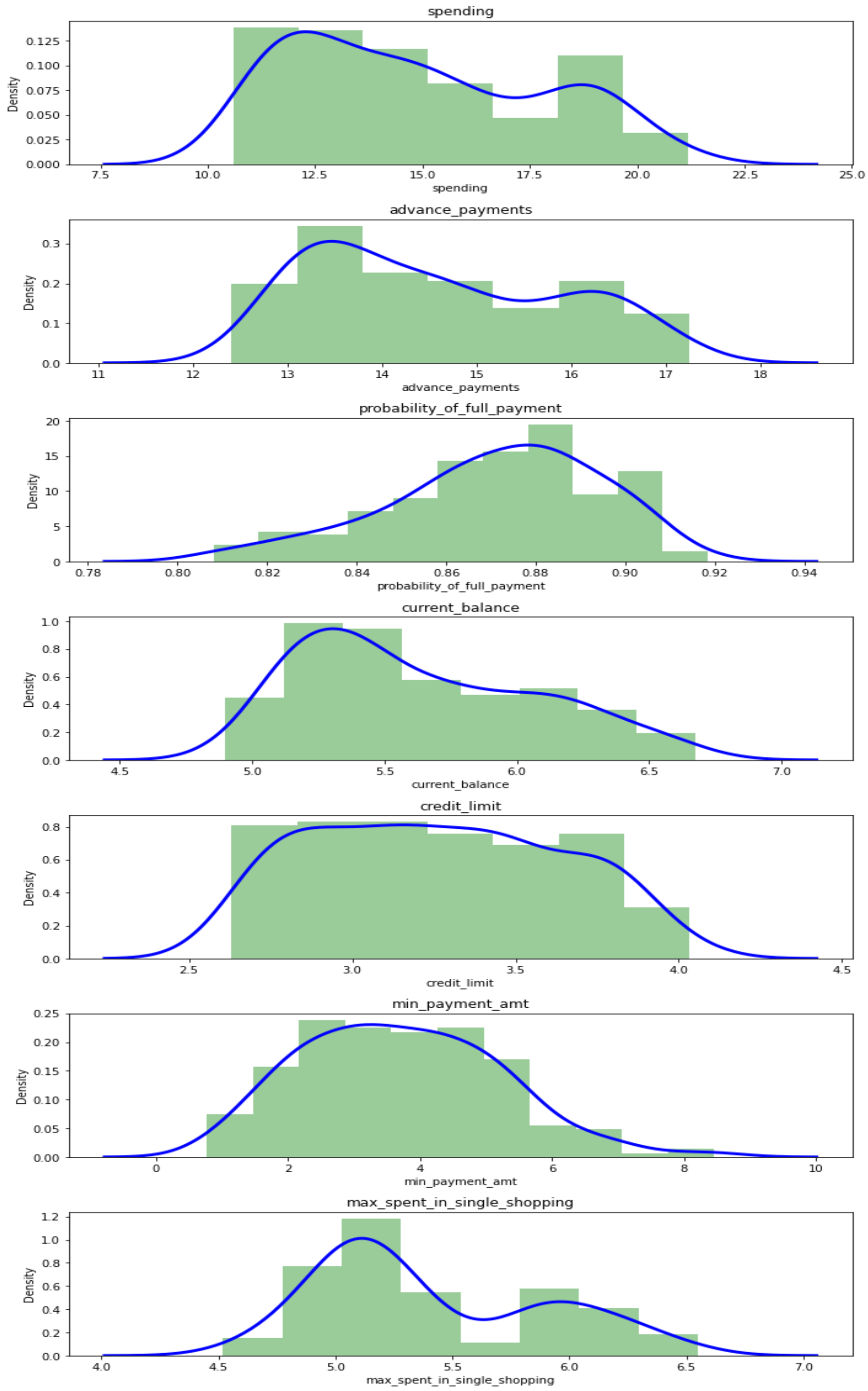


FIGURE 16: SKEWNESS OF THE VARIABLES

- Skewness, in statistics, is the degree of asymmetry observed in a probability distribution. Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness.



- The right skew or positive skew means that the most values are clustered around the left tail of the distribution whereas the right tail of the distribution is longer and vice-versa for negatively skewed or left skewed

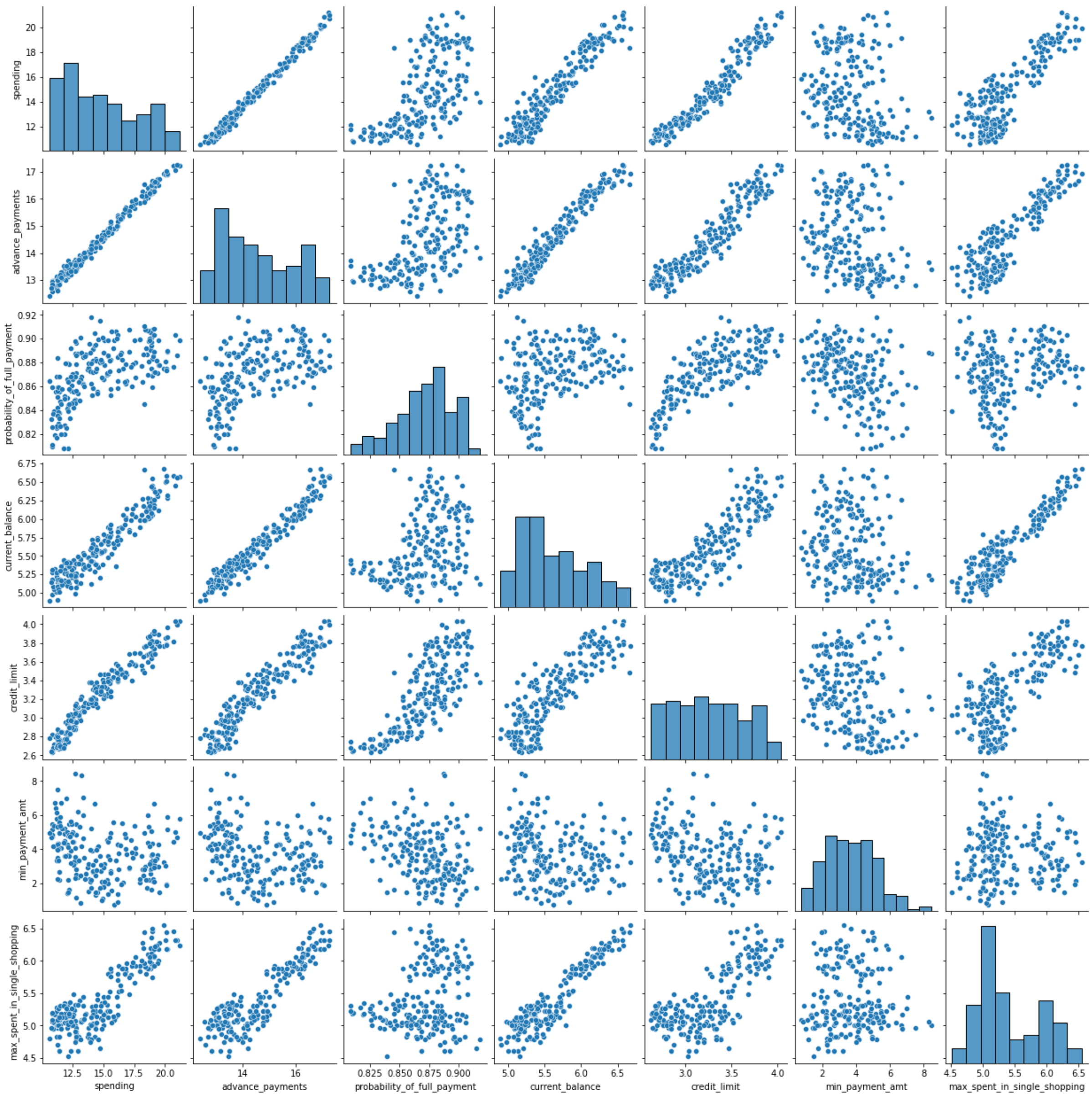
INFERENCE -SKEWNESS OF THE DATA SET

1. “Spending” variable has a slight deviation in the deviation, it slightly right skewed i.e., positively skewed.
2. The “advance payments” variable too has a similar inference as spending variable, it is right skewed or positively skewed.
3. The “probability of Full payment variable” is left skewed i.e., it is negatively skewed.
4. The “current balance” variable is right skewed or positively skewed, it has a slight deviation.
5. The “credit limit” variable has zero skewness i.e., it is normally distributed as it forms the bell curve shape, making it symmetrically skewed.
6. The “Min payment amt” variable has zero skewness i.e., it is normally distributed as it forms the bell curve shape, making it symmetrically skewed.
7. The “max shopping in single day” variable is not normally distributed. It is right skewed or positively skewed.

MULTI / BI VARIATE ANALYSIS FOR THE DATASET

- Multivariate analysis is based in observation and analysis of more than one statistical outcome variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

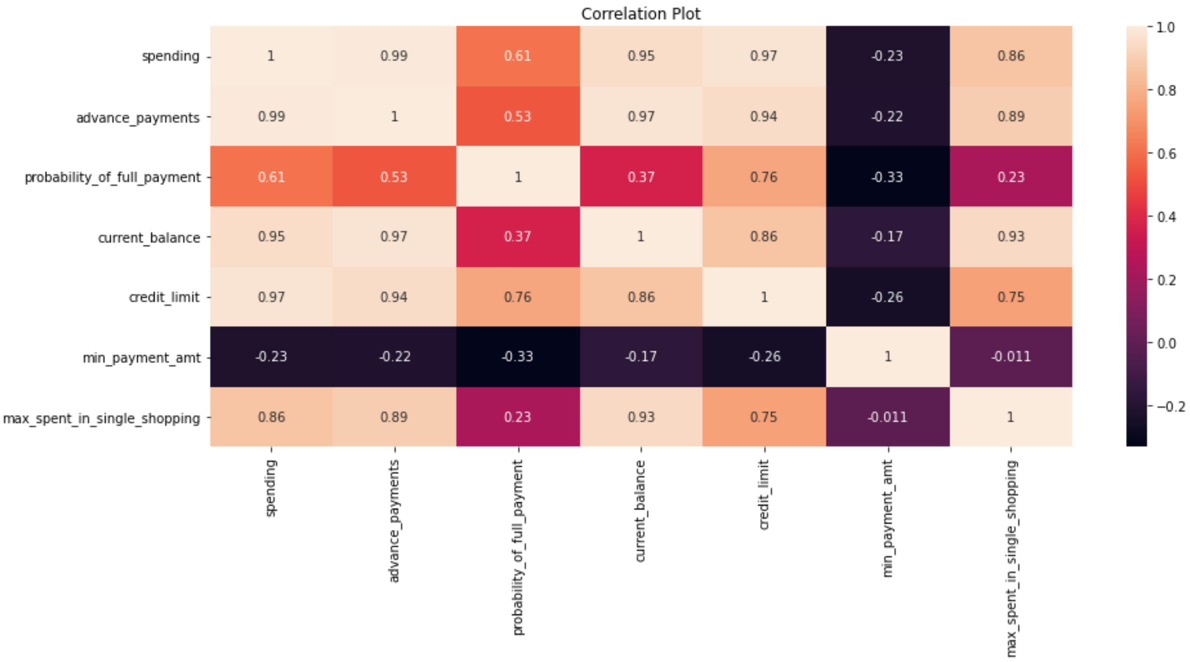
FIGURE 17: PAIR PLOT



INFERENCE PAIR PLOT

- From the pair plot we can see that, there is a very high correlation between Spending and advance payments, credit limit, current balance. The logic behind such a correlation may be because the customer is spending on paying the advance for his purchases. Similarly, when he spends he uses his credit card and therefore he spends within that credit limit and not further. His expenses from his credit card also affects his bank account balance, therefore there is a correlation between spending current balance.
- We can also infer that there is a strong correlation between Advance payments and current balance, credit limit. The reason being the payment of such amount from the credit card affects the bank account when bill is due, hence such a high correlation between Advance payments and current balance. Between credit limit it is because payments made through credit card can be within the customer's credit limit only.
- Maximum spent in single shopping and current balance also have strong correlation, logic behind it may be that due to the purchase/money spent it directly affects the balance in the account of the customer.

FIGURE 18: HEAT MAP / CORRELATION PLOT



INFERENCE FOR HEAT MAP / CORRELATION PLOT

- There is very high correlation between the following

1. “Spending” and “advance payments” (0.99)
2. “Advance payments” and “current balance” (0.97)
3. “Credit limit” and “spending” (0.97)
4. “Spending” and “current balance” (0.94)
5. “Credit limit “and “advance payments” (0.94)
6. “Max spent in single shopping” and “Current balance” (0.93)

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

1. Scaling is usually done when the values/units in the dataset are different and therefore

2. spending, advance payments are in different values/units and thus get higher weightage, Thus Scaling is necessary for this dataset.

3. Scaling will have all the values in the same range relatively.

4. Z score can be used for scaling the dataset.

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

TABLE 13: SCALED DATA TOP 5 SAMPLES

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.178230	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	1.501773	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	0.504874	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-2.591878	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	1.196340	0.591544	1.155464	-1.088154	0.874813

FIGURE 19: DENDROGRAM

- A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. Here Linkage method is being used.

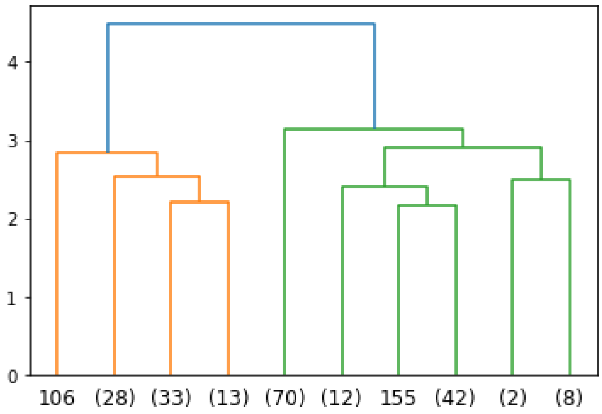


TABLE 14: CLUSTERING

	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

TABLE 15: ADDING FREQUENCY TO THE DATASET

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

INFERENCE FOR Q.1.3

1. Performing clustering using Dendrogram, we understand that there are two optimal cluster 3 and 4, performing with 3 cluster on further analysis looks good based on the hierarchical clustering.
2. With 3 clusters it provides with High, Medium, Low spending and also similar with probability of full payment and max shopping in single day.
3. Frequency of cluster 1 is 75 times/records, Frequency of cluster 2 is 70 times/records, Frequency of cluster 3 is 65 times/records.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

OUTPUT: KMEANS CLUSTERING

- K-Means Clustering: K-Means clustering is an unsupervised learning algorithm. There is no labelled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.6531439995162,
 326.6639378916672,
 289.2457367203014,
 265.43027192046856,
 239.49765708705579,
 221.66639706594844,
 204.8712753824312]
```

FIGURE 20: WEIGHTED SUM OF SQUARES CURVE / ELBOW CURVE

- Weighted Sum of Squares curve: It helps to know how many clusters are needed as output in K-means Clustering

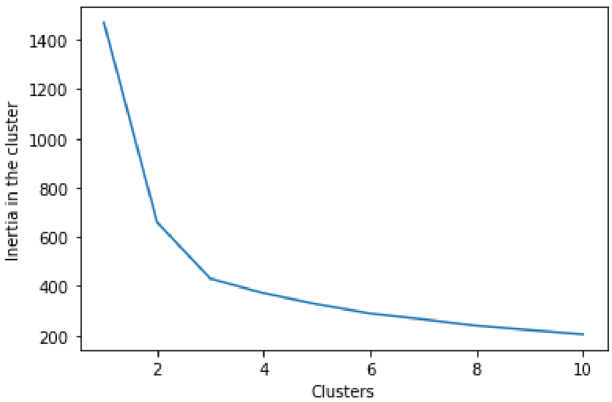


TABLE 16: KMEANS CLUSTERING DATASET

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

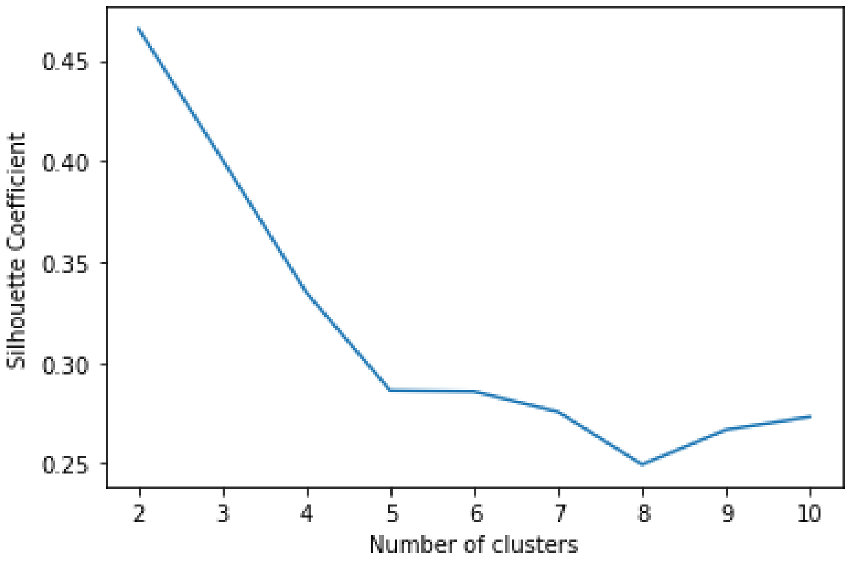
OUTPUT: EXTRACTING THE SILHOUETTE SCORES

- Silhouette Score: It is used to study the separation distance between the resulting clusters.

```
[0.46577247686580914,
 0.40072705527512986,
 0.3347542296283262,
 0.28621461554288646,
 0.285726896652541,
 0.2756098749293962,
 0.24943558736282168,
 0.266366921192433,
 0.2731288488219916]
```

FIGURE 21: SILHOUETTE PLOT

- The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually.



OUTPUT: SILHOUETTE WIDTH AND MINIMUM SILHOUETTE SCORE

- Silhouette Width: It is the average of the silhouette scores, higher the average Silhouette width higher is the quality of the clustering.
- The array list of Silhouette width cannot be included in the business report so kindly refer the Jupyter notebook.
- The minimum Silhouette Score is: 0.0027

```
0.002713089347678376
```

INFERENCE FOR Q.1.4.

1. On applying K-Means clustering on the scaled data we get 10 inertias which are then plotted in a graph called the Elbow curve or WSS curve.
2. From the Elbow curve we infer that the optimum clusters are 3. Because after 3 clusters the difference between the inertias are small or negligible, therefore making it 3 optimised clusters.
3. The K-Means clusters are fitted in the scaled dataset (Table 15)
4. The silhouette score is 0.4007 and minimum silhouette score is 0.0027 as per the output, from this we can infer that the clusters are not separable and are equidistant between the centroids. There is no blunder created in the clustering.

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

TABLE 17: CLUSTER PROFILES

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

CLUSTER 1: HIGH

1. Giving any reward points might increase their purchases.
2. maximum max spent in single shopping is high for this group, so can be offered discount/offer on next transactions upon full payment.
3. Increase their credit limit and
4. Increase spending habits.
5. Give loan against the credit card, as they are customers with good repayment record.
6. Tie up with luxury brands, which will drive more one time maximum spending.

CLUSTER 2: MEDIUM

1. They are potential target customers who are paying bills and doing purchases and maintaining comparatively good credit score. 2. So we can increase credit limit or can lower down interest rate.
2. Promote premium cards/loyalty cars to increase transactions.
3. Increase spending habits by trying with premium ecommerce sites, travel portal, travel airlines/hotel, as this will encourage them to spend more.

CLUSTER 3: LOW

1. Customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate.
2. Increase their spending habits by tying up with grocery stores, utilities (electricity, phone, gas, others)

PROBLEM 2 -CART-RF-ANN

CART-CLASSIFICATION AND REGRESSION TECHNIQUES

- Classification and regression trees (CART) are a set of techniques for classification and prediction. The technique is aimed at producing rules that predict the value of an outcome (target) variable from known values of predictor (explanatory) variables.

RF-RANDOM FOREST

- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

ANN-ARTIFICIAL NEURAL NETWORK

- Neural networks are often used for effective data mining, turning raw data into viable information. They look for patterns in large batches of data, allowing businesses to learn more about their customers, which can inform their marketing strategies, increase sales, and lower costs.

INTRODUCTION

- The dataset contains past year details of an insurance company. The main aim is to understand the why the claim frequency is higher and what measures can be taken to improve business. In this problem we will be doing exploratory data analysis, do classification using techniques, Apply Random Forest classifier (Decision tree) to improve prediction accuracy and also apply artificial neural network to provide measures using the same.

DATA DICTIONARY

1. Target: Claim Status (Claimed)
2. Agency code: Code of tour firm
3. Type: Type of tour insurance firms
4. Channel: Distribution channel of tour insurance agencies
5. Product: Name of the tour insurance products
6. Duration in days: Duration of the tour
7. Destination: Destination of the tour

8. Sales: Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100's)
9. Commission: The commission received for tour insurance firm (Commission is in percentage of sales)
10. Age: Age of insured (Age)

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

TABLE 18: TOP 5 SAMPLES

	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

TABLE 19: LAST 5 SAMPLES

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
2995	28	CWT	Travel Agency	Yes	166.53	Online	364	256.20	Gold Plan	Americas
2996	35	C2B	Airlines	No	13.50	Online	5	54.00	Gold Plan	ASIA
2997	36	EPX	Travel Agency	No	0.00	Online	54	28.00	Customised Plan	ASIA
2998	34	C2B	Airlines	Yes	7.64	Online	39	30.55	Bronze Plan	ASIA
2999	47	JZI	Airlines	No	11.55	Online	15	33.00	Bronze Plan	ASIA

TABLE 20: INFORMATION ABOUT THE DATASET

```
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)
```

TABLE 21: DESCRIPTION ABOUT THE DATASET

	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

TABLE 22: MISSING DATA IN THE DATASET

```
Age             0
Agency_Code     0
Type            0
Claimed         0
Commision       0
Channel         0
Duration        0
Sales           0
Product Name    0
Destination     0
dtype: int64
```

TABLE 23: DUPLICATE DATA IN THE DATASET

```
Number of duplicate rows = 139
```

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
63	30	C2B	Airlines	Yes	15.0	Online	27	60.0	Bronze Plan	ASIA
329	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
407	36	EPX	Travel Agency	No	0.0	Online	11	19.0	Cancellation Plan	ASIA
411	35	EPX	Travel Agency	No	0.0	Online	2	20.0	Customised Plan	ASIA
422	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
...	...	...	...	...	...	...	...	...	...	...
2940	36	EPX	Travel Agency	No	0.0	Online	8	10.0	Cancellation Plan	ASIA
2947	36	EPX	Travel Agency	No	0.0	Online	10	28.0	Customised Plan	ASIA
2952	36	EPX	Travel Agency	No	0.0	Online	2	10.0	Cancellation Plan	ASIA
2962	36	EPX	Travel Agency	No	0.0	Online	4	20.0	Customised Plan	ASIA
2984	36	EPX	Travel Agency	No	0.0	Online	1	20.0	Customised Plan	ASIA

139 rows × 10 columns

TABLE 24: UNIQUE VALUE IN THE COLUMNS OF THE DATASET

```
Agency_Code : 4
JZI     239
CWT     472
C2B     924
EPX    1365
Name: Agency_Code, dtype: int64


Type : 2
Airlines         1163
Travel Agency    1837
Name: Type, dtype: int64


Claimed : 2
Yes     924
No     2076
Name: Claimed, dtype: int64


Channel : 2
Offline      46
Online     2954
Name: Channel, dtype: int64
```

```
Product Name : 5
Gold Plan             109
Silver Plan           427
Bronze Plan           650
Cancellation Plan     678
Customised Plan      1136
Name: Product Name, dtype: int64


Destination : 3
EUROPE       215
Americas     320
ASIA        2465
Name: Destination, dtype: int64
```

INFERENCE FOR THE ABOVE

1. The Dataset has 3000 records and 10 variables
2. The shape of the dataset is (3000,10)
3. Age, Commission, Duration, Sales are numeric variable rest are categorial variables
4. There are no missing values in the dataset
5. There are around 139 duplicate records., regarding duplicates, I am not removing the duplicates as I assume that these data can be of different customers having the same type of business, product, destination.
6. Variable Agency code has 4 values namely JZI, CWT, C2B, EPX with 239, 472, 924, 1365 counts / records respectively
7. Variable Type has 2 values namely Airlines and Travel Agency with 1163 and 1837 records respectively.
8. Variable Claimed has 2 values namely Yes and No with 924 and 2076 counts / records respectively.
9. Variable Channel has 2 values namely offline and online with 46 and 2954 counts/ records respectively.

10. Product Name Variable has 5 Values namely gold plan, Silver Plan, Bronze Plan, Cancellation Plan and Customised Plan with 109, 427, 650, 678, 1136, counts / records respectively.
11. In Table 20, Duration variable has a minimum value in negative (-1)
12. Commission and sales Mean and Median Value varies significantly.

UNIVARIATE ANALYSIS FOR ALL THE VARIABLES IN THE DATASETS

TABLE 25: AGE VARIABLE DESCRIPTION

Description of Age

count	3000.000000
mean	38.091000
std	10.463518
min	8.000000
25%	32.000000
50%	36.000000
75%	42.000000
max	84.000000

FIGURE 22: AGE DISTRIBUTION PLOT

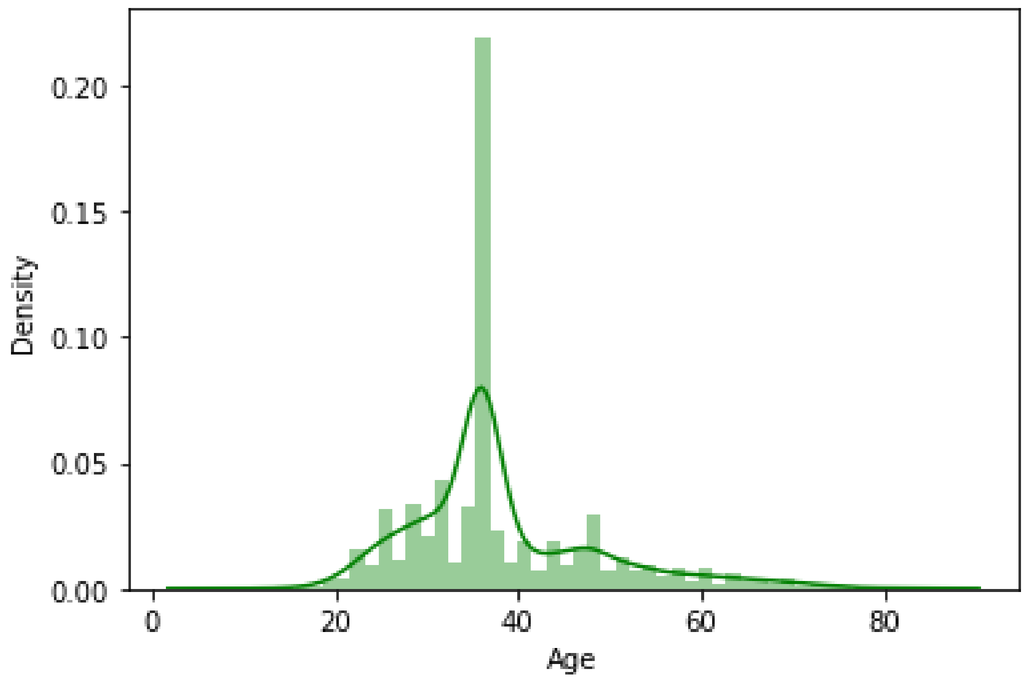
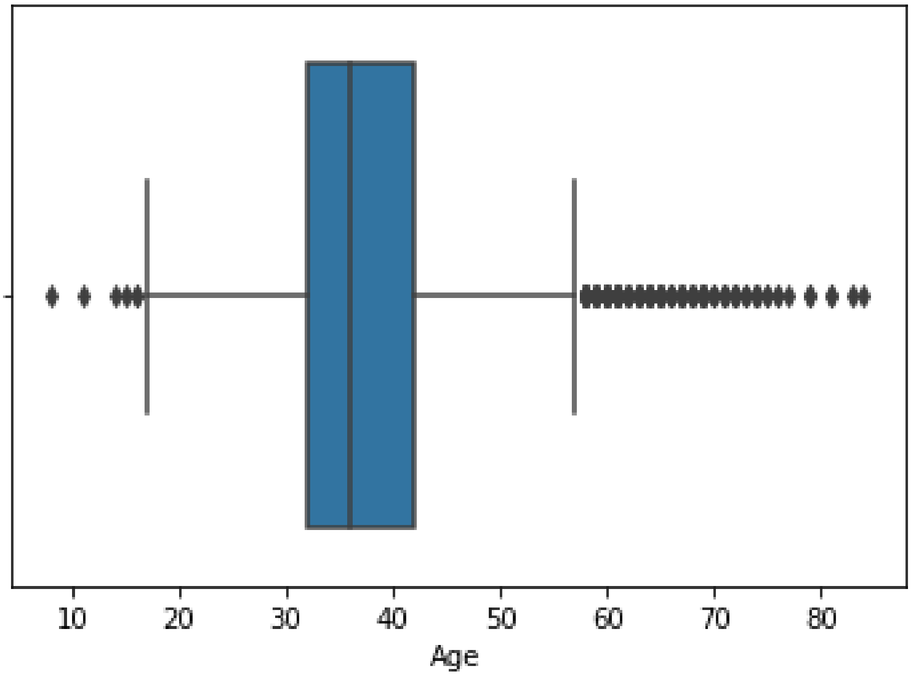


FIGURE 23: AGE BOXPLOT



OUTPUT: INTERQUARTILE RANGE AND OUTLIER DETECTION

```
1st Quartile (Q1) is:  32.0
3rd Quartile (Q3) is:  42.0
Interquartile range (IQR) of Age is  10.0
```

```
Number of outliers in Age upper :  198
Number of outliers in Age lower :  6
```

INFERENCE FOR AGE VARIABLE

1. Minimum Age: 8.00
2. Maximum Age: 84.00
3. Mean value: 38.09
4. Median value: 36.00
5. Standard deviation: 10.46
6. Range: 76.00
7. The Inter Quartile range is 10
8. The age variable has outliers, as observed in the box plot and the output above which says there are 198 outliers in upper whisker and 6 outliers in the lower whisker.

9. The Distribution plot is not normally distributed, it is right/positively skewed.

TABLE 26: COMMISSION DESCRIPTION

```
Description of Commision
--------------------------
count    3000.000000
mean       14.529203
std        25.481455
min         0.000000
25%         0.000000
50%         4.630000
75%        17.235000
max       210.210000
```

FIGURE 24: COMMISSION DISTRIBUTION PLOT

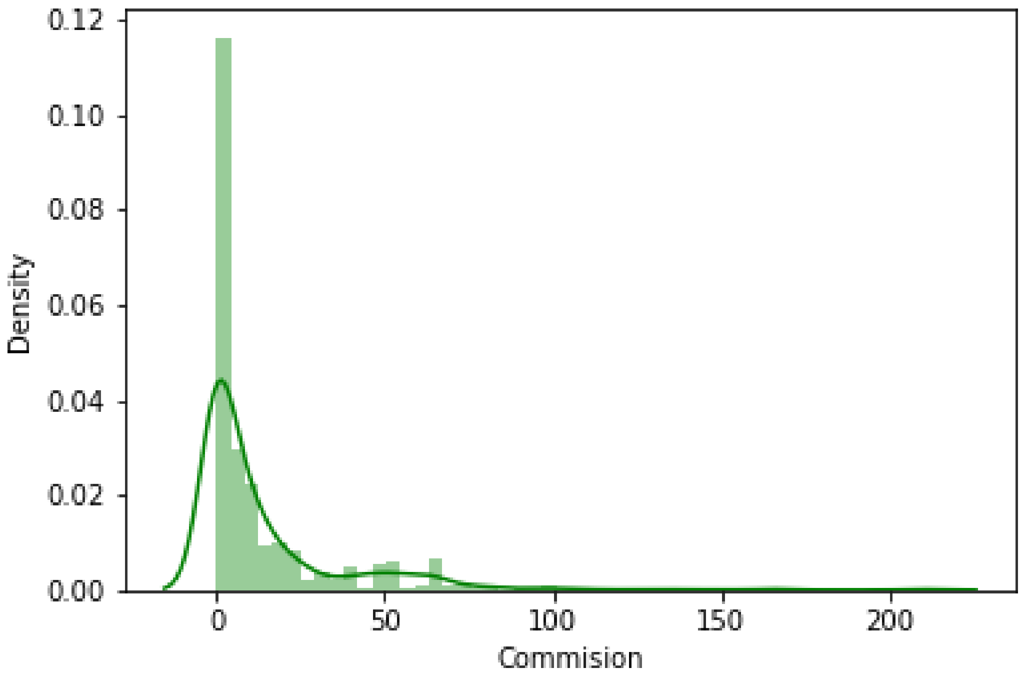
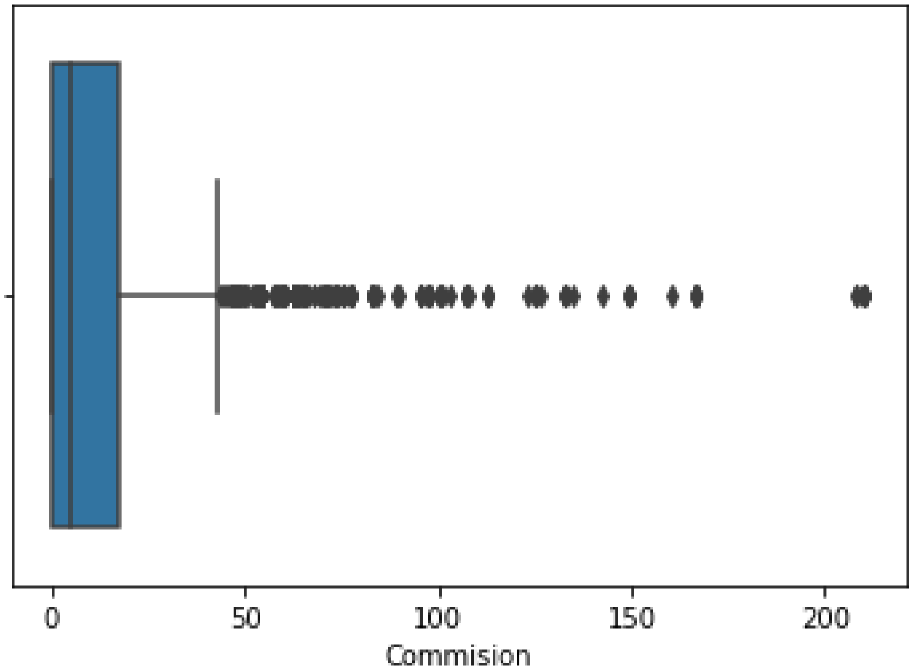


FIGURE 25: COMMISSION BOXPLOT



OUTPUT: INTERQUARTILE RANGE AND OUTLINE EXTRACTION

```
1st Quartile (Q1) is:  0.0
3rd Quartile (Q3) is:  17.235
Interquartile range (IQR) of Commision is  17.235

Upper outliers in Commision:  43.0875
Lower outliers in Commision:  -25.8525

Number of outliers in Commision upper :  362
Number of outliers in Commision lower :  0
```

INFERENCE FOR COMMISSION VARIABLE

1. Minimum Commission: 0.0

2. Maximum Commission: 210.21

3. Mean value: 14.52

4. Median value: 4.63

5. Standard deviation: 25.48

6.Inter quartile range: 17.35

7. Range: 210.21

8. There are Outliers in the commission variable with 362 outliers, as per the boxplot and the above output.

9. The distribution plot shows that the variable is normally distributed and is positively skewed.

TABLE 27: DURATION DESCRIPTION

Description of Duration

count	3000.000000
mean	70.001333
std	134.053313
min	-1.000000
25%	11.000000
50%	26.500000
75%	63.000000
max	4580.000000

FIGURE 26: DURATION DISTRIBUTION PLOT

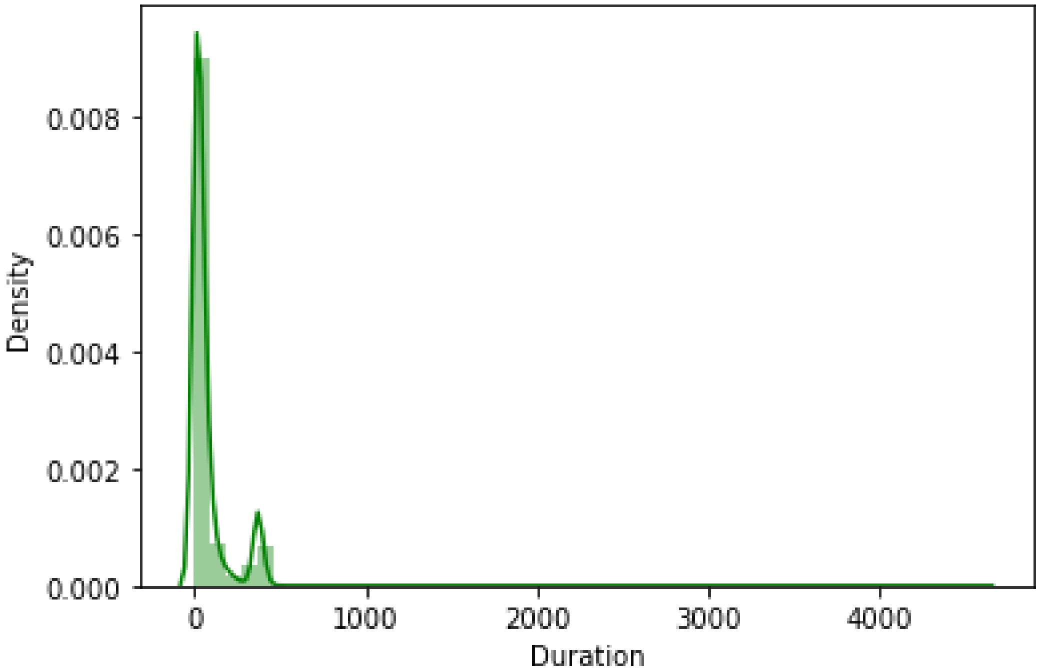
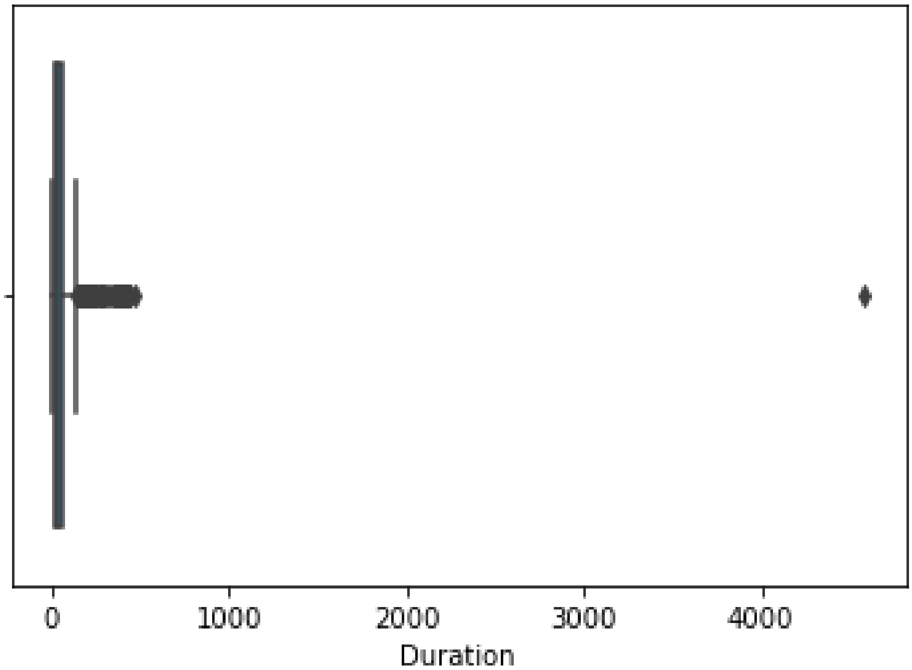


FIGURE 27: DURATION DISTRIBUTION PLOT



OUTPUT: INTERQUARTILE RANGE AND OUTLIER DETECTION

```
1st Quartile (Q1) is:  11.0
3rd Quartile (Q3) is:  63.0
Interquartile range (IQR) of Duration is  52.0

Upper outliers in Duration:  141.0
Lower outliers in Duration:  -67.0

Number of outliers in Commision upper :  382
Number of outliers in Commision lower :  0
```

INFERENCE FOR DUARTION VARIABLE

1. Minimum Duration: -1

2. Maximum Duration: 4580

3. Mean value: 70.00

4. Median value: 26.5

5. Standard deviation: 134.05

6. Range is: 4581

7. There are outliers in the duration variable as per the boxplot and the output above, there are 382 outliers in the variable.

8. The Interquartile range is 52.00

9. The distribution plot show that the variable is distributed normally and is right skewed or positively skewed.

TABLE 28: SALES VARIABLE DESCRIPTION

Description of Sales

count	3000.000000
mean	60.249913
std	70.733954
min	0.000000
25%	20.000000
50%	33.000000
75%	69.000000
max	539.000000

FIGURE 28: SALES VARIABLE DISTRIBUTION PLOT

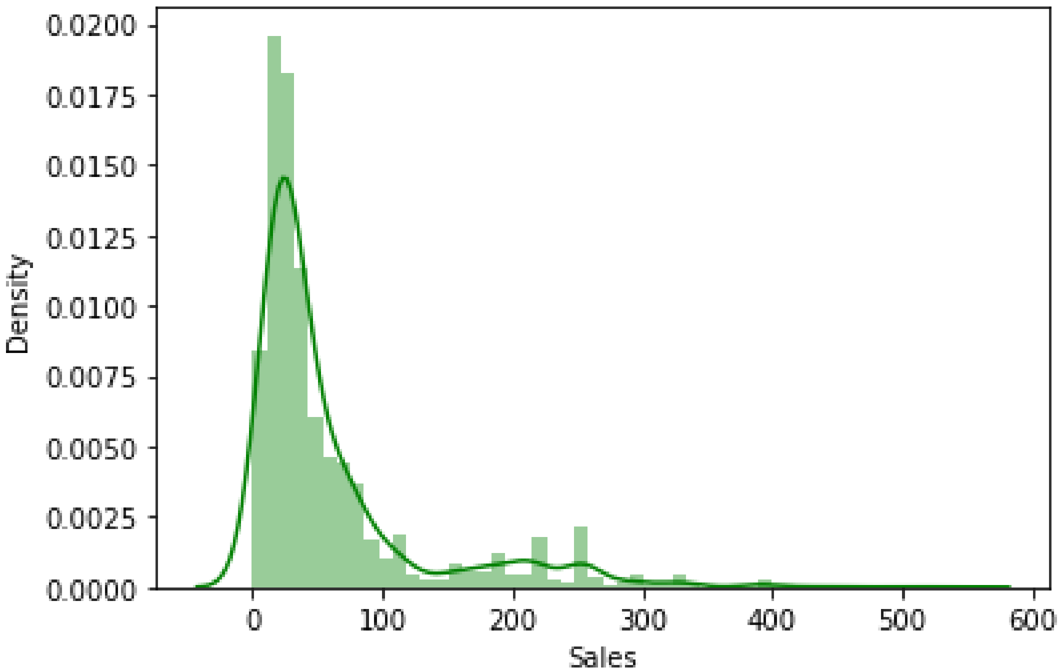
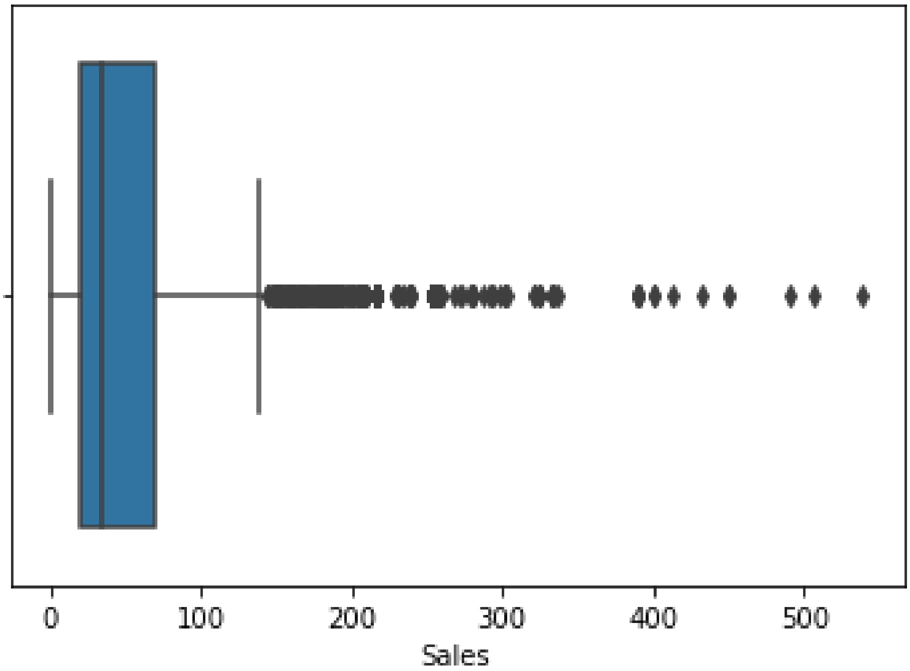


FIGURE 29: SALES VARIABLE BOXPLOT



OUTPUT: INTER QUARTILE RANGE AND OUTLIER EXTRACTION

```
1st Quartile (Q1) is:  20.0
3rd Quartile (Q3) is:  69.0
Interquartile range (IQR) of Sales is  49.0

Upper outliers in Sales:  142.5
Lower outliers in Sales:  -53.5

Number of outliers in Commision upper :  353
Number of outliers in Commision lower :  0
```

INFERENCE FOR SALES VARIABLE

1. Range: 539.00
2. Minimum Sales: 0.0
3. Maximum Sales: 539.0
4. Mean value: 60.24
5. Median value: 33.0
6. Standard deviation: 70.73
7. The interquartile range is 49.00
8. There are outliers in the variable as per the boxplot and output given above, there are 353 outliers in the variable.

9. The distribution plot is distributed normally and is right or positively skewed.

TABLE 29: UNIQUE VALUES

Age	70
Agency_Code	4
Type	2
Claimed	2
Commision	324
Channel	2
Duration	257
Sales	380
Product Name	5
Destination	3

FIGURE 30: CHANNEL VARIABLE BAR GRAPH

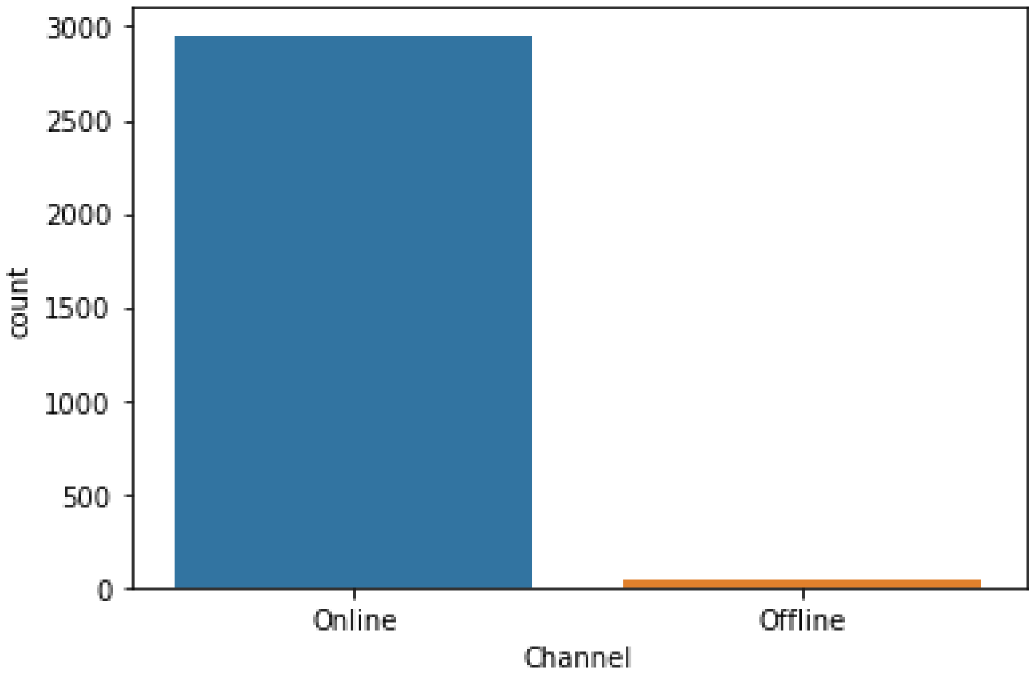
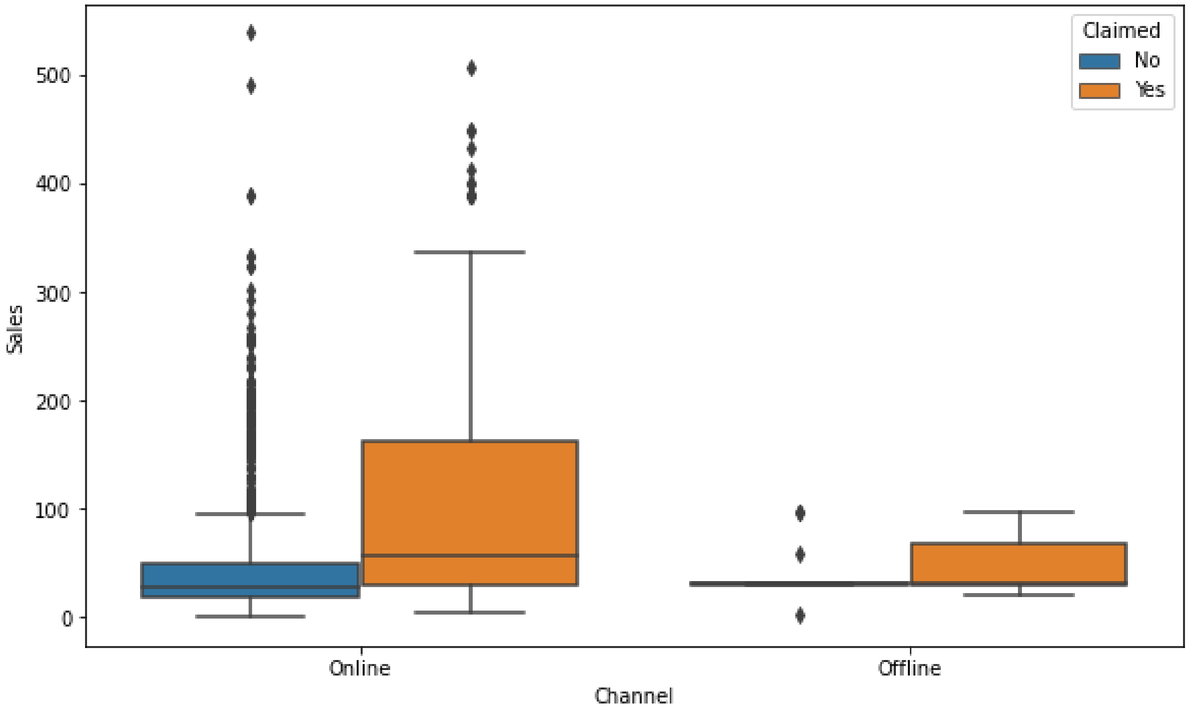


FIGURE 31: CHANNEL VARIABLE BOXPLOT



INFERENCE FOR CHANNEL VARIABLE

1. As the bar graph, the online channel is used more as a distribution channel approximately around 2900 records, whereas the offline is less than 200 approximately.
2. From the boxplot, it is clear that the interquartile range for Online channel which has claimed is higher than all the channels and claimed variable.
3. The online channel which has not claimed has more outliers comparatively.
4. The offline channel which is not claiming has only 3 outliers and very low inter quartile range.
5. The offline channel which is claimed does not have outliers.

FIGURE 32: BAR GRAPH FOR AGENCY CODE VARIABLE

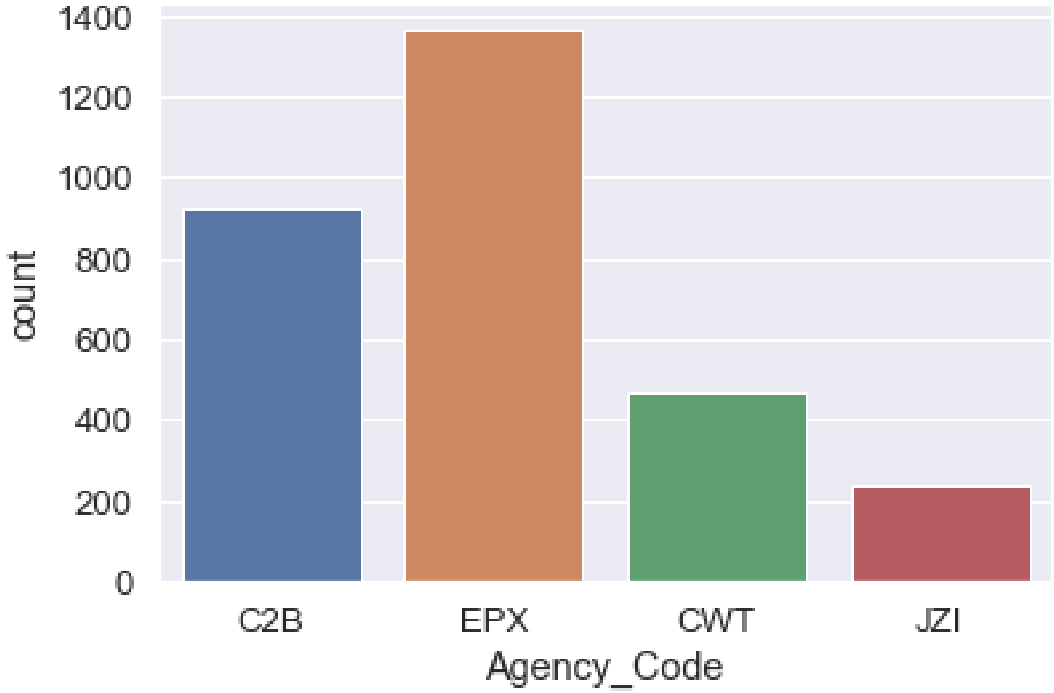
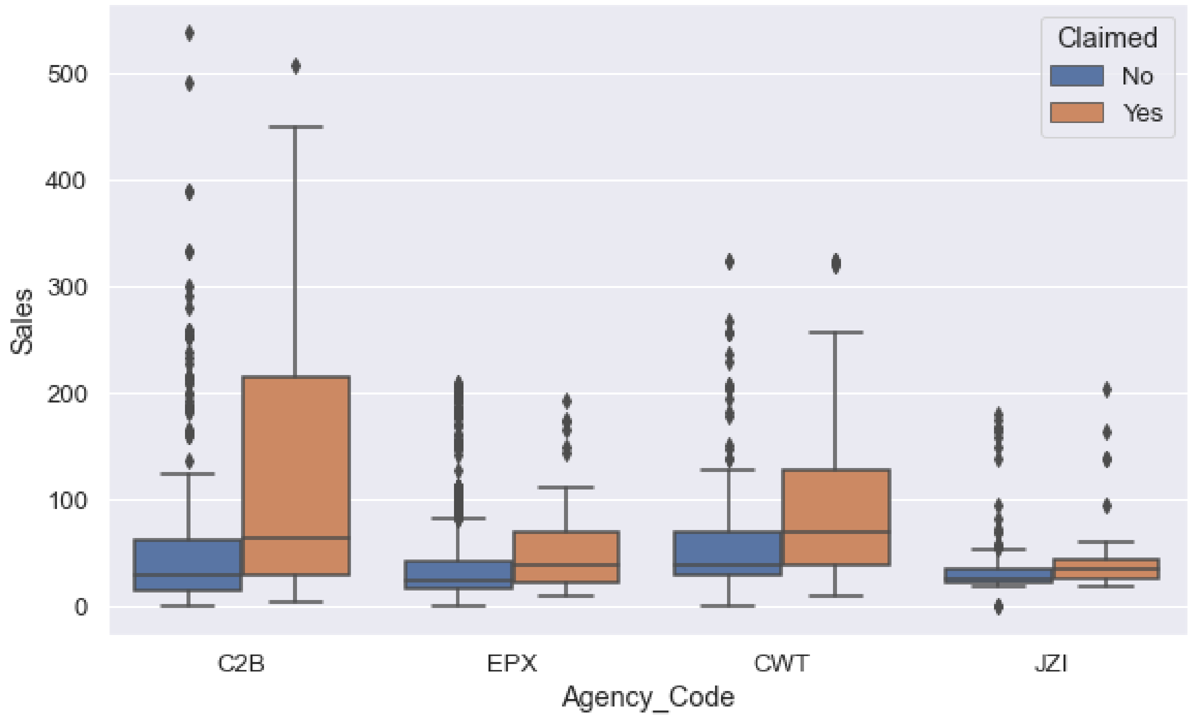


FIGURE 33: BOXPLOT FOR AGENCY CODE VARIABLE



INFERENCE FOR AGENCY CODE

1. As per the bar graph we infer that agent code EPX has highest sales done and agency code JZI has the lowest sale of insurance tour.
2. There are outliers in all the agency code records,
3. C2B claimed has the highest inter quartile range among all the inter quartiles and JZI not claimed has the lowest Inter quartile range.

FIGURE 34: BAR GRAPH FOR TYPE VARIABLE

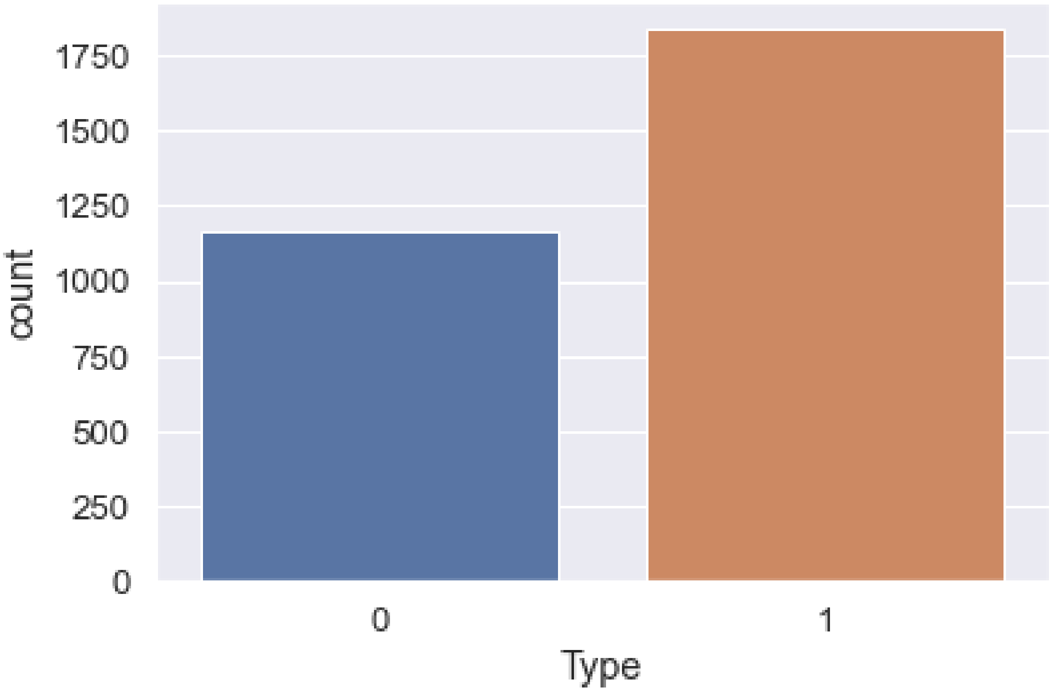
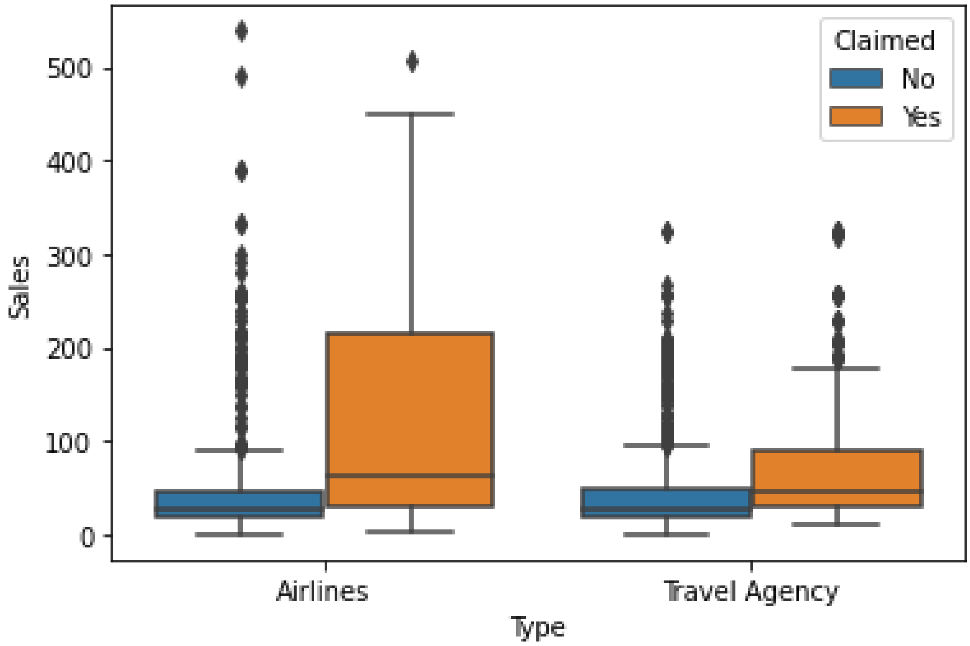


FIGURE 35: BOXPLOT FOR TYPE VARIABLE



INFERENCE FOR TYPE VARIABLE

1. Here 0 is Airlines and 1 is Travel Agency, as per the bar graph the Travel agency has had higher number of sales,
2. The boxplot shows that there are outliers in the variables and Airlines claimed has a greater number of outliers comparatively.
3. The Interquartile range is also higher for Airlines claimed comparatively.
4. Whereas the travel agency not claimed interquartile range is very low compared to travel agency claimed, but it is similar to Airlines not claimed.

FIGURE 36: BAR GRAPH FOR PRODUCT NAME VARIABLE

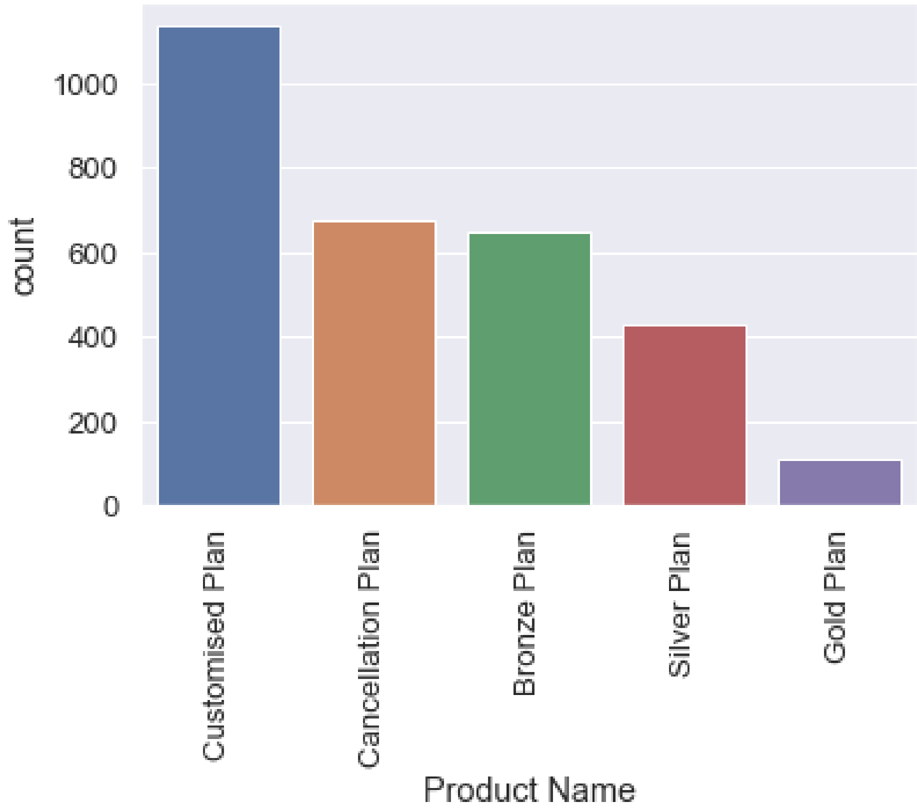
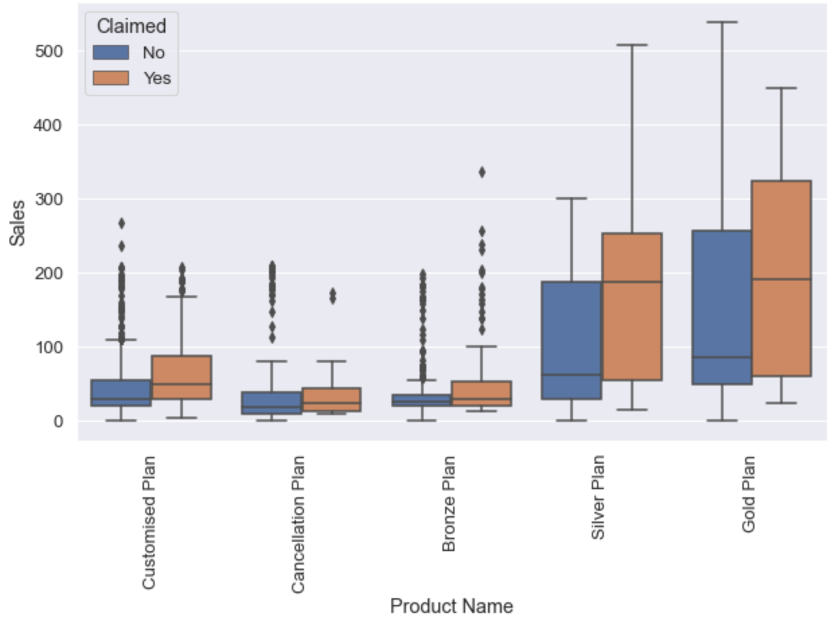


FIGURE 37: BOXPLOT FOR PRODUCT NAME VARIABLE



INFERENCE FOR PRODUCT NAME VARIABLE

1. The sale for Customised plan has been higher comparatively approx.1600., the lowest sale is for Gold Plan approx. less than 200.
2. The boxplot shows that there are outliers in Customized Plan, cancellation plan and bronze plan and there are no outliers in silver and gold plan.
3. The median is almost similar for Cancellation plan and bronze plan.
4. The Gold and Silver plan have a higher inter quartile range.

FIGURE 38: BARGRAPH FOR DESTINATION VARIABLE

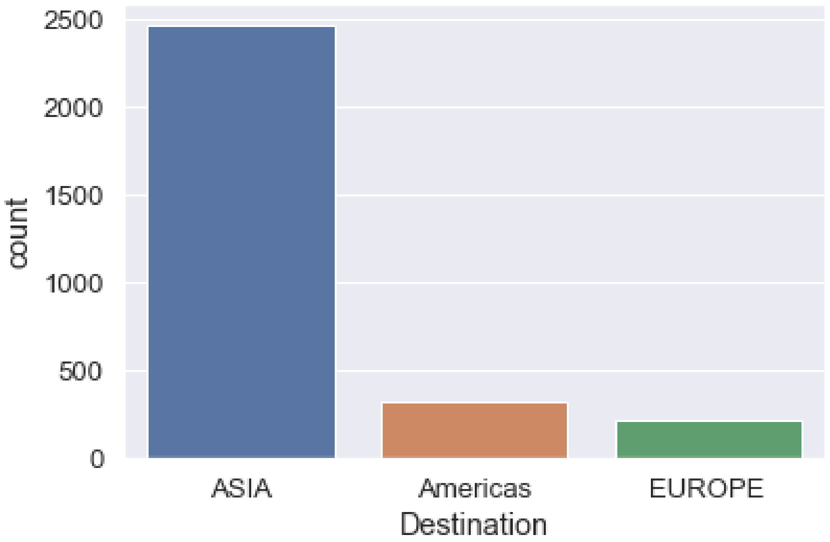
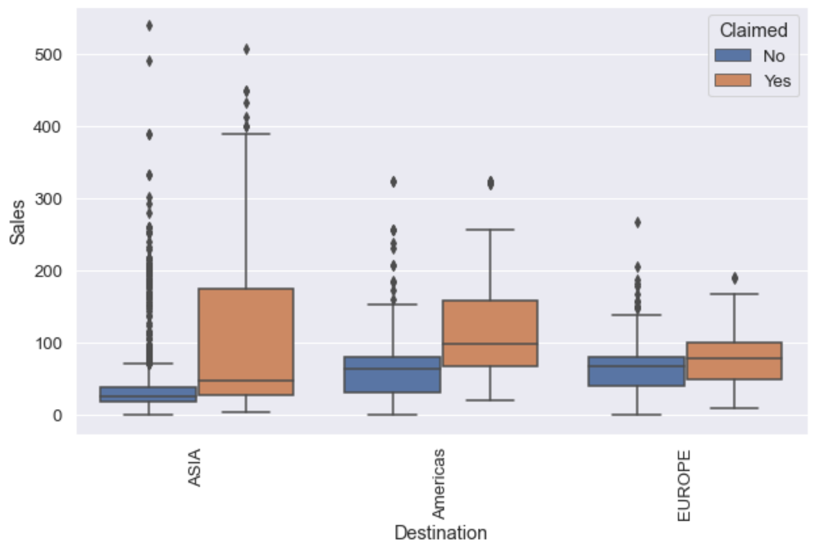


FIGURE 39: BOXPLOT FOR DESTINATION VARIABLE

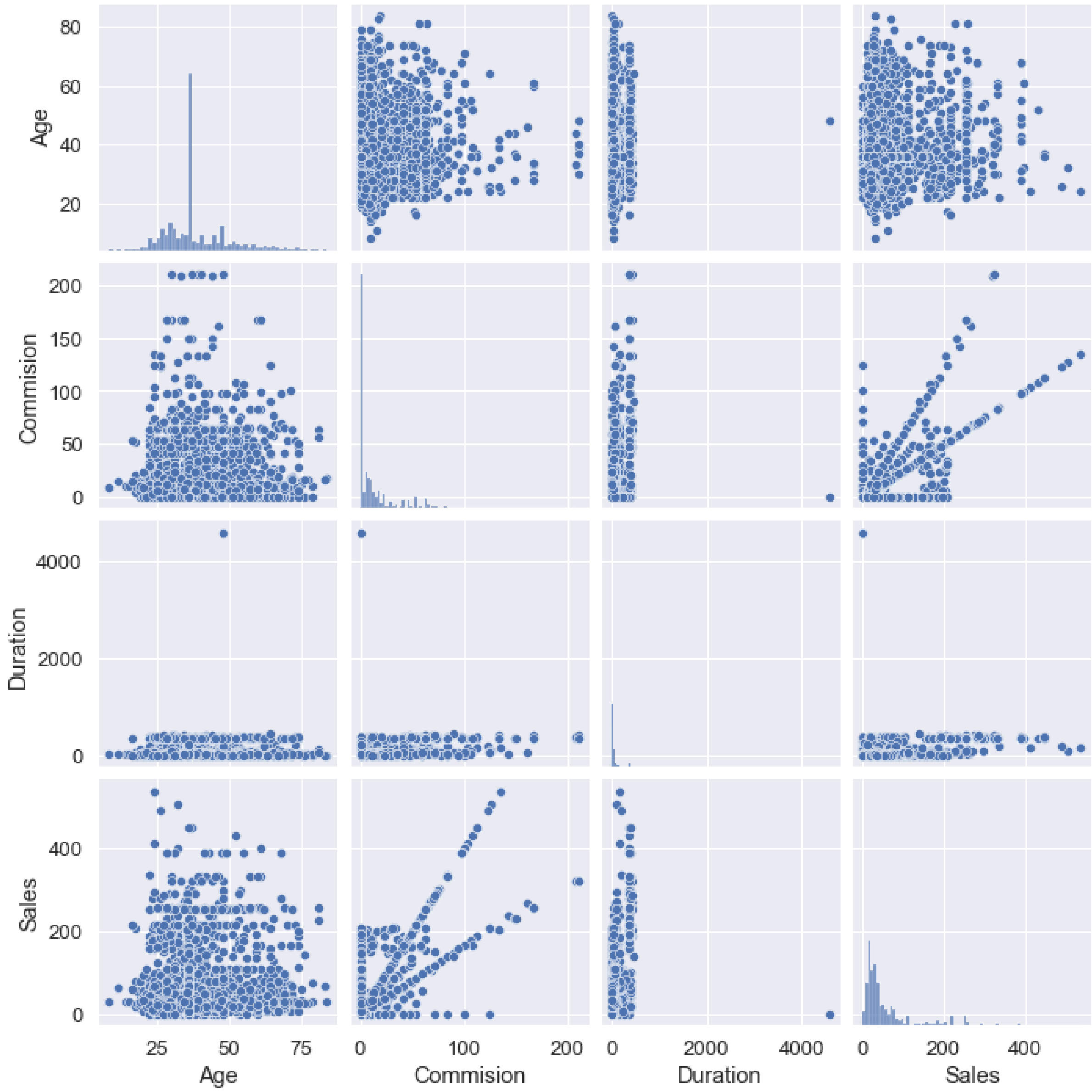


INFERENCE FOR DESTINATION VARIABLE

1. The sale in Asia is the highest amongst all the destination and the lowest in Europe, but Europe and America do not have significant difference they have negligible difference.
2. The boxplot shows that there are outliers in all destination records.

MULTIVARIATE ANALYSIS FOR ALL THE VARIABLES IN THE

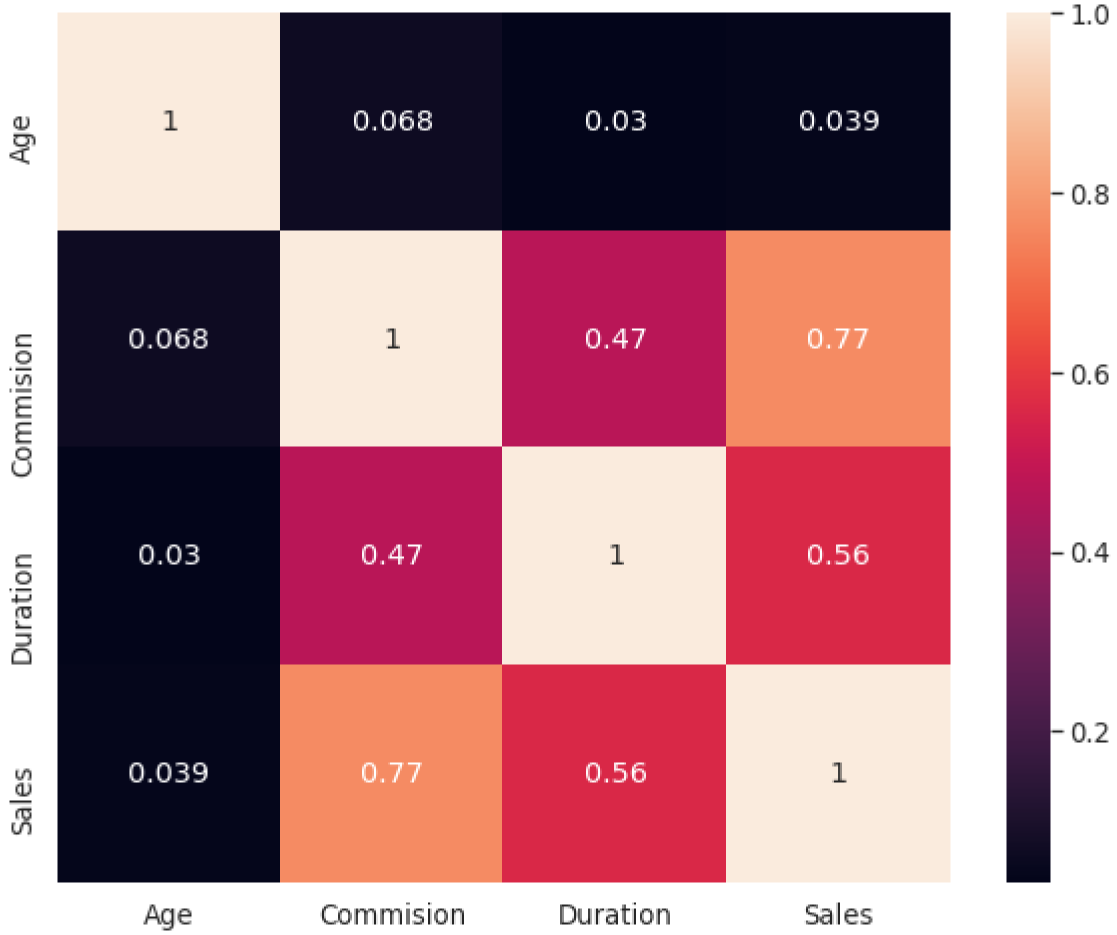
FIGURE 40: PAIRPLOT



INFERENCE PAIR PLOT

- There is no correlation between variables. Therefore, plotting the heatmap to get a better understanding about the correlation between the variables

FIGURE 41: HEATMAP / CORRELATION PLOT



INFERENCE HEATMAP /CORRELATION PLOT

There is no strong correlation between variables.

The highest correlation is between Sales and Commission but it is indeed not very strong it is only 0.77. The logic behind this would be that higher the sales more the commission and vice versa i.e., there is direct proportion between both the variables.

The reason of less correlation between variables can also be because of duplicate data and outliers, but duplicate data cannot be removed or imputed as the customers can have similar plans or destination or other variables.

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

TABLE 30: CONVERTING OBJECTS TO CATEGORICAL INFO TABLE

```
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   int8   
 2   Type          3000 non-null   int8   
 3   Claimed       3000 non-null   int8   
 4   Commision     3000 non-null   float64
 5   Channel       3000 non-null   int8   
 6   Duration      3000 non-null   int64  
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   int8   
 9   Destination   3000 non-null   int8   
```

TABLE 31: CONVERTED DATA TOP 5 SAMPLE

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0	0.70	1	7	2.51	2	0
1	36	2	1	0	0.00	1	34	20.00	2	0
2	39	1	1	0	5.94	1	3	9.90	2	1
3	36	2	1	0	0.00	1	4	26.00	1	0
4	33	3	0	0	6.30	1	53	18.00	0	0

OUTPUT: ALLOCATION OF VALUES AFTER CONVERSION

```
feature: Agency_Code
['C2B', 'EPX', 'CWT', 'JZI']
Categories (4, object): ['C2B', 'CWT', 'EPX', 'JZI']
[0 2 1 3]


feature: Type
['Airlines', 'Travel Agency']
Categories (2, object): ['Airlines', 'Travel Agency']
[0 1]


feature: Claimed
['No', 'Yes']
Categories (2, object): ['No', 'Yes']
[0 1]


feature: Channel
['Online', 'Offline']
Categories (2, object): ['Offline', 'Online']
[1 0]


feature: Product Name
['Customised Plan', 'Cancellation Plan', 'Bronze Plan', 'Silver Plan', 'Gold Plan']
Categories (5, object): ['Bronze Plan', 'Cancellation Plan', 'Customised Plan', 'Gold Plan', 'Silver Plan']
[2 1 0 4 3]
```

```
feature: Destination
['ASIA', 'Americas', 'EUROPE']
Categories (3, object): ['ASIA', 'Americas', 'EUROPE']
[0 1 2]
```

TABLE 32: SPLITTING INTO TRAIN AND TEST DATA-SAMPLE

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0.70	1	7	2.51	2	0
1	36	2	1	0.00	1	34	20.00	2	0
2	39	1	1	5.94	1	3	9.90	2	1
3	36	2	1	0.00	1	4	26.00	1	0
4	33	3	0	6.30	1	53	18.00	0	0

TABLE 33: SCALING THE DATASET- SAMPLE

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
0	0.947162	-1.314358	-1.256796	-0.542807	0.124788	-0.470051	-0.816433	0.268835	-0.434646
1	-0.199870	0.697928	0.795674	-0.570282	0.124788	-0.268605	-0.569127	0.268835	-0.434646
2	0.086888	-0.308215	0.795674	-0.337133	0.124788	-0.499894	-0.711940	0.268835	1.303937
3	-0.199870	0.697928	0.795674	-0.570282	0.124788	-0.492433	-0.484288	-0.525751	-0.434646
4	-0.486629	1.704071	-1.256796	-0.323003	0.124788	-0.126846	-0.597407	-1.320338	-0.434646

TABLE 34: SHAPE OF DATA AFTER SPLITTING THE DATA

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

- The Train data has 2100 records and the test data has 900 records.

FIGURE 42: DECISION TREE

- A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes)

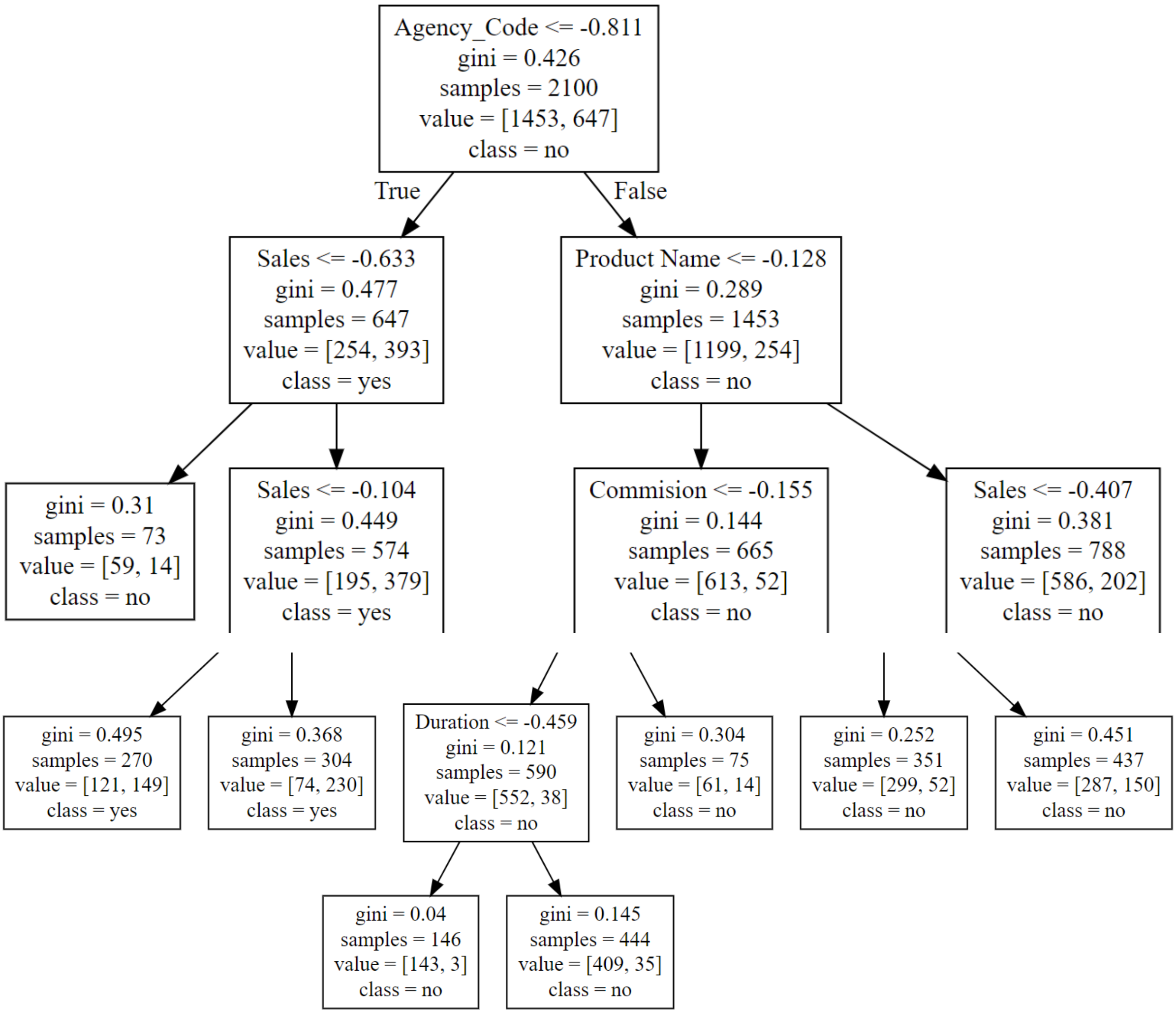


TABLE 35: IMPORTANCE VARIABLE

- Variable importance is determined by calculating the relative influence of each variable: whether that variable was selected to split on during the tree building process, and how much the squared error (over all trees) improved (decreased) as a result.

	Imp
Age	0.000000
Agency_Code	0.676527
Type	0.000000
Commision	0.008032
Channel	0.000000
Duration	0.000000
Sales	0.223015
Product Name	0.092427
Destination	0.000000

TABLE 36: PREDICTING ON TEST AND TRAIN DATASET

	0	1
0	0.656751	0.343249
1	0.935593	0.064407
2	0.935593	0.064407
3	0.656751	0.343249
4	0.935593	0.064407

INFERENCE FOR DECISION TREE CLASSIFICATION

(Step by step approach is provided in Jupyter Notebook).

1. The variable importance table show that the most important variable which contributed to the prediction is the Agency code and sales variable.

TABLE 37: PREDICTING MODELS FOR RANDOM FOREST

	0	1
0	0.763145	0.236855
1	0.957879	0.042121
2	0.899905	0.100095
3	0.667343	0.332657
4	0.851731	0.148269

TABLE 38: VARIABLE IMPORTANCE FOR RANDOM FOREST

	Imp
Age	0.007379
Agency_Code	0.391832
Type	0.072731
Commision	0.093041
Channel	0.000000
Duration	0.025512
Sales	0.127552
Product Name	0.278012
Destination	0.003941

- The variable importance table shows that the most important variable contributed to predict is the Agency code, Sales and Product name for Random Forest.

TABLE 39: PREDICTION DATA USING ANN

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

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

FIGURE 43: ROC CURVE FOR TRAIN DATA-DECISION TREE

AUC: 0.810

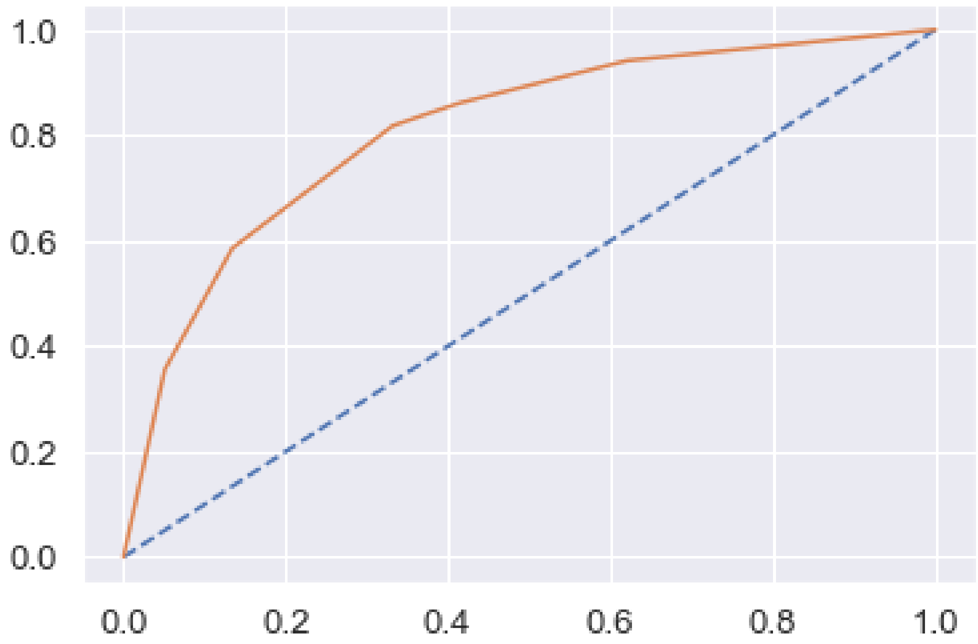


TABLE 40: CONFUSION MATRIX FOR TRAIN DATA-DECISION TREE

```
array([[1258,  195],
       [ 268,  379]], dtype=int64)
```

TABLE 41: CLASSIFICATION REPORT FOR TRAIN DATA-DECISION TREE

	precision	recall	f1-score	support
0	0.82	0.87	0.84	1453
1	0.66	0.59	0.62	647
accuracy			0.78	2100
macro avg	0.74	0.73	0.73	2100
weighted avg	0.77	0.78	0.78	2100

INFERENCE FOR TRAIN DATA-DECISION TREE

1. The accuracy for train data tunes from 65% to 78% after applying random forest prediction model.
2. Train precision: 0.66
3. Train recall: 0.59
4. Train f1: 0.62
5. AUC is: 81%

FIGURE 44: ROC CURVE FOR TEST DATA-DECISION TREE

AUC: 0.799

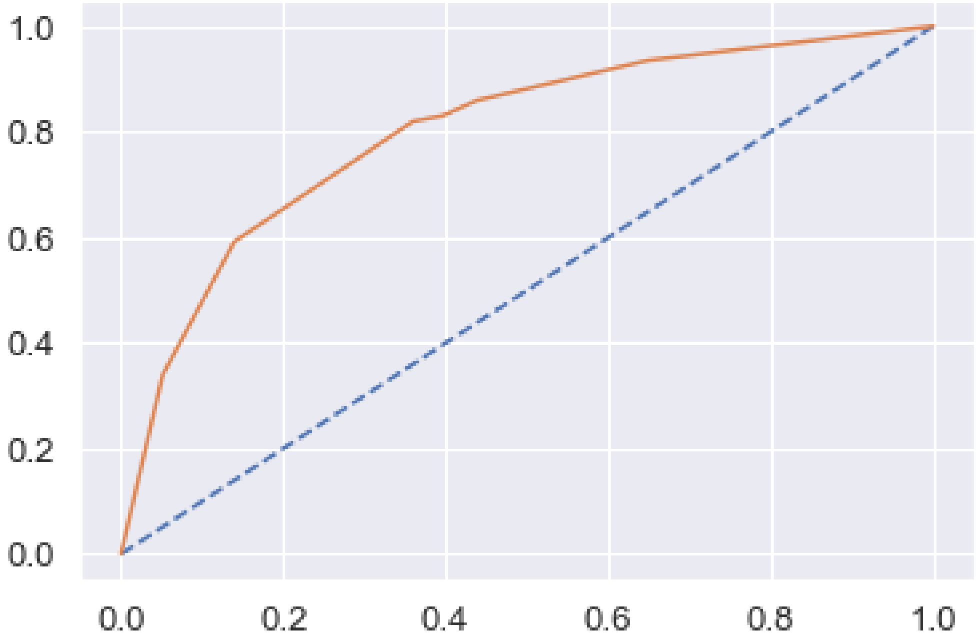


TABLE 42: CONFUSION MATRIX FOR TEST DATA-DECISION TREE

```
array([[1258,  195],
       [ 268,  379]],
```

TABLE 43: CLASSIFICATION REPORT FOR TEST DATA-DECISION TREE

	precision	recall	f1-score	support
0	0.83	0.86	0.84	623
1	0.65	0.59	0.62	277
accuracy			0.78	900
macro avg	0.74	0.73	0.73	900
weighted avg	0.77	0.78	0.77	900

INFERENCE FOR TEST DATA-DECISION TREE

1. Test precision: 0.65
2. Test recall:0.59
3. Test f1: 0.62
4. Accuracy is: 0.78
5. AUC IS: 80%

FIGURE 45: ROC FOR TRAIN DATA-RANDOM FOREST

Area under Curve is 0.8233458250318321

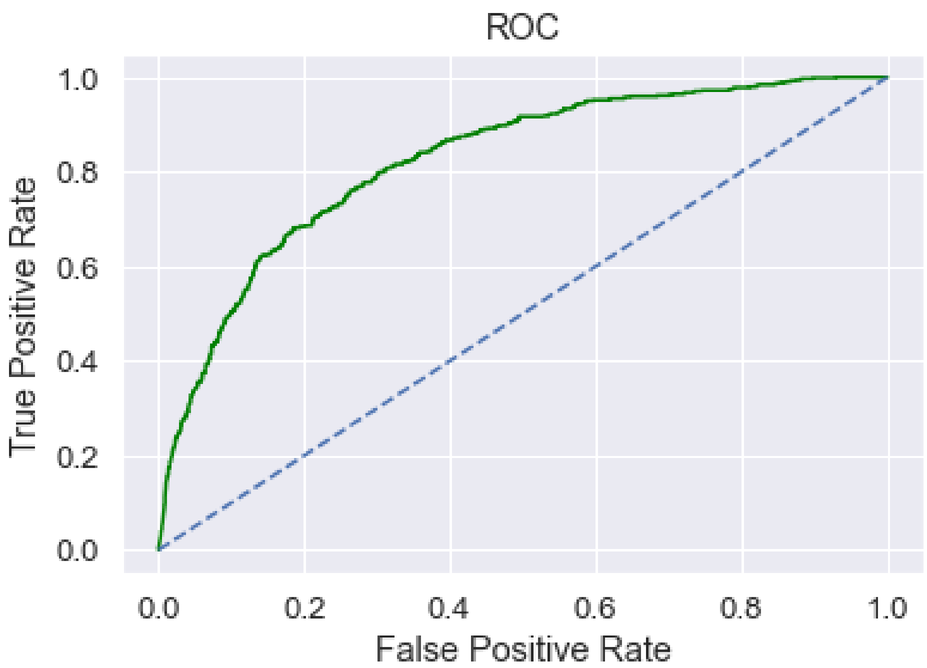


TABLE 44: CONFUSION MATRIX FOR TRAIN DATA-RANDOM FOREST

```
array([[1308,  145],
       [ 323,  324]])
```

TABLE 45: CLASSIFICATION REPORT FOR TRAIN DATA-RANDOM FOREST

	precision	recall	f1-score	support
0	0.80	0.90	0.85	1453
1	0.69	0.50	0.58	647
accuracy			0.78	2100
macro avg	0.75	0.70	0.71	2100
weighted avg	0.77	0.78	0.77	2100

INFERENCE FOR TRAIN DATA-RANDOM FOREST

1. rf test precision: 0.69

2. rf test recall: 0.50

3. rf test f1: 0.58

4. AUC: 82%

5. Accuracy: 0.78

FIGURE 46: ROC FOR TEST DATA-RANDOM FOREST

Area under Curve is 0.8107126921672818

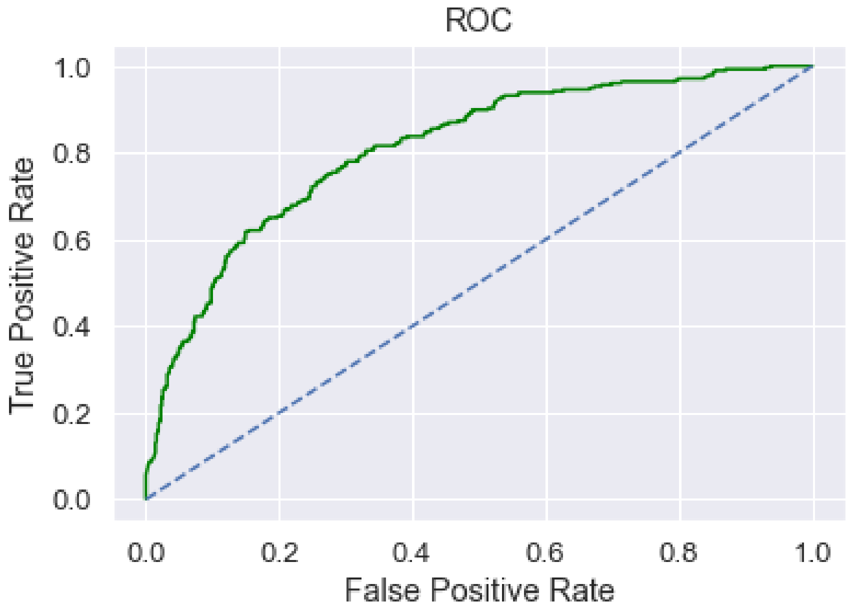


TABLE 46: CONFUSION MATRIX FOR TEST DATA-RANDOM FOREST

```
array([[558,  65],
       [138, 139]],
```

TABLE 45: CLASSIFICATION REPORTS FOR TEST DATA-RANDOM FOREST

	precision	recall	f1-score	support
0	0.80	0.90	0.85	623
1	0.68	0.50	0.58	277
accuracy			0.77	900
macro avg	0.74	0.70	0.71	900
weighted avg	0.76	0.77	0.76	900

INFERENCE FOR TEST DATA-RANDOM FOREST

1. rf test precision: 0.68

2. rf test recall: 0.50

3. rf test f1: 0.58

4. AUC: 0.81%

5. Accuracy: 0.77

6. Training and Test set results are almost similar, and with the overall measures high, the model is a good. Agency code is again the most important variable for predicting customer insurance claim.

FIGURE 47: ROC FOR TRAIN DATA-ANN

Area under Curve is 0.7790697921796932

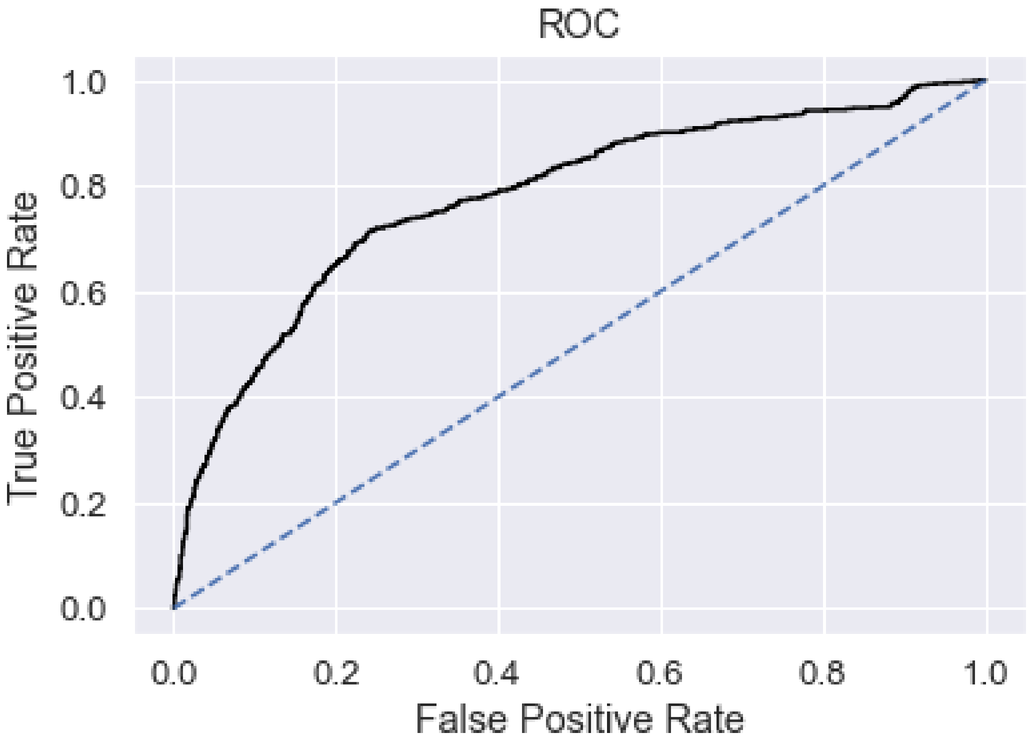


TABLE 48: CONFUSION MATRIX FOR TRAIN DATA-ANN

```
array([[1340,  113],
       [ 396,  251]]).
```

TABLE 49: CLASSIFICATION REPORTS FOR TRAIN DATA-ANN

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

INFERENCE FOR TRAIN DATA-ANN

1. train precision:0.69

2. train recall:0.39

3. train f1: 0.50

4. AUC: 0.78

5. Accuracy: 0.76

FIGURE 48: ROC FOR TEST DATA-ANN

Area under Curve is 0.7587891360657353

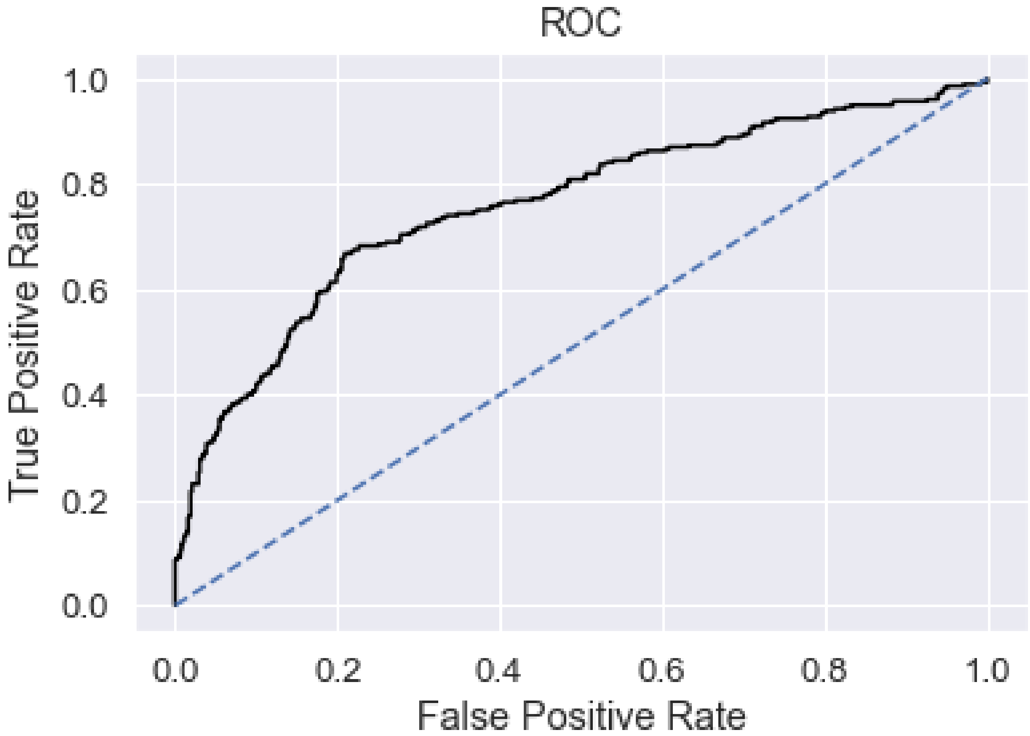


TABLE 50: CONFUSION MATRIX FOR TEST DATA-ANN

```
array([[576,  47],
       [171, 106]],
```

TABLE 51: CLASSIFICATION REPORTS FOR TEST DATA-ANN

```
              precision    recall  f1-score   support

           0       0.77      0.92      0.84       623
           1       0.69      0.38      0.49       277

    accuracy                           0.76       900
   macro avg       0.73      0.65      0.67       900
weighted avg       0.75      0.76      0.73       900
```

INFERENCE FOR TEST DATA-ANN

1. train_precision:0.69

2. train_recall:0.38

3. train_f1: 0.49

4. Accuracy: 0.76

5. AUC: 75%

6. Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

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

TABLE 52: COMPARISION OF ALL MODELS

	CART Train	Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.78	0.78	0.78	0.77	0.76	0.76
AUC	0.81	0.80	0.82	0.81	0.78	0.76
Recall	0.59	0.59	0.50	0.50	0.39	0.39
Precision	0.66	0.65	0.69	0.68	0.69	0.69
F1 Score	0.62	0.62	0.58	0.58	0.50	0.50

FIGURE 49: ROC COMPARISION OF ALL MODELS- TRAIN DATA

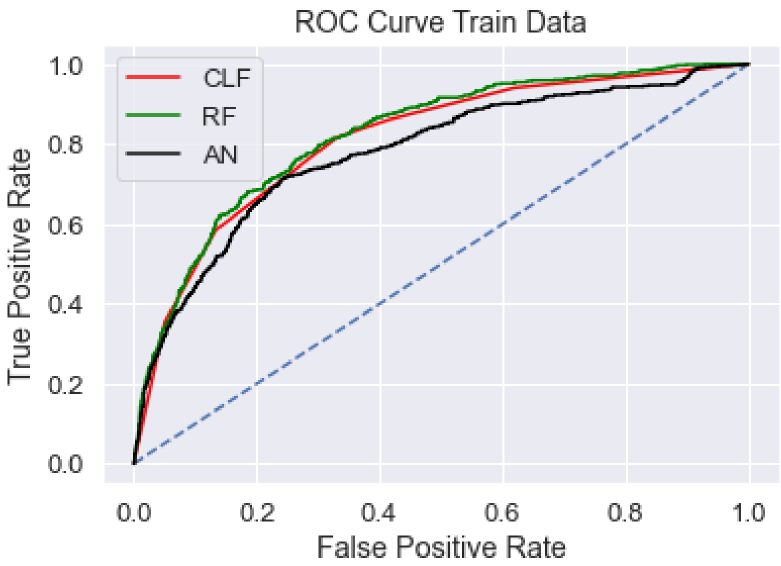
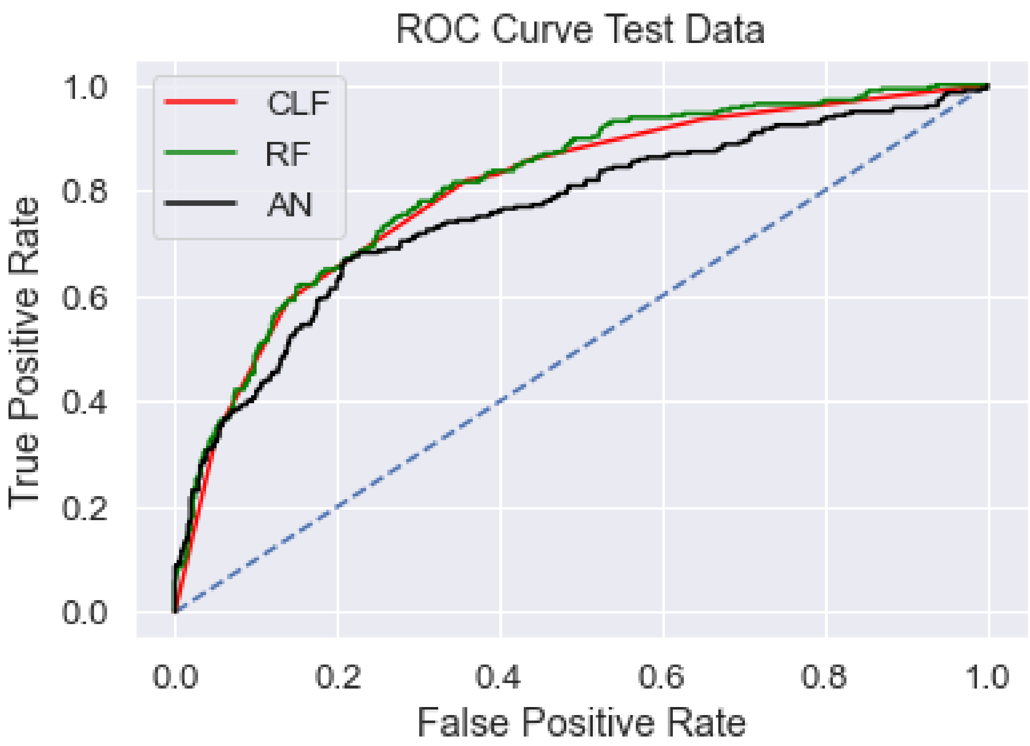


FIGURE 50: ROC COMPARISION OF ALL MODELS- TEST DATA



CONCLUSION: I am selecting the RF model, as it has better accuracy, precision, recall, f1 score better than other two DTF & AN.

INFERENCE FOR COMPARISION OF ALL MODELS

1. From the above corelation and analysis using the three models, I think there should be further more data provided by the insurance company to get an accurate analysis on the dataset.
2. Most conversions to business happened through the Online channel, which has increased the profit over the years, this can be interpreted as online insurance sale is 90% when compared to offline sale.
3. There should be training given or an audit to check on how strategies are made by JZI agency to sell insurance, as they very low sales comparatively. They can also work on merging few more agencies to enhance their business conversions or do some more marketing.
4. The reason for why all offline businesses is claimed is unclear, it can be understood with further data provided and further analysis.

5. The Travel agency sales is high compared to the airline ticketing, the only reason can be because the travel agent takes care of all process and the customer get a hassle-free experience which is a common reason, the main reason for such an increase cannot be interpreted from the data. To increase the sales of the airlines, we need to analyse the working of the travel agency which will help in mimicking the process and increase the sales there also.
6. The company can reduce it handling cost and expand its territory, boundary, products with the use of the above data analysis.
7. They can also work on outsourcing or transfer out few of there solutions to reduce the risk of high claims by reselling the insurance etc.
8. The company should reduce the claim processing time, so that the customer is satisfied.
9. By doing the above the insurance company will have a higher retention rate among customers.
10. The company should also reduce the fraud or other malpractices which can affect the company’s reputation and the business in large by using various technologies.

THE END