Data Mining Project

Submitted by : Rishab Singla

Table of Contents

List of Figures

PROBLEM 1: CLUSTERING

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

QUESTION 1.1

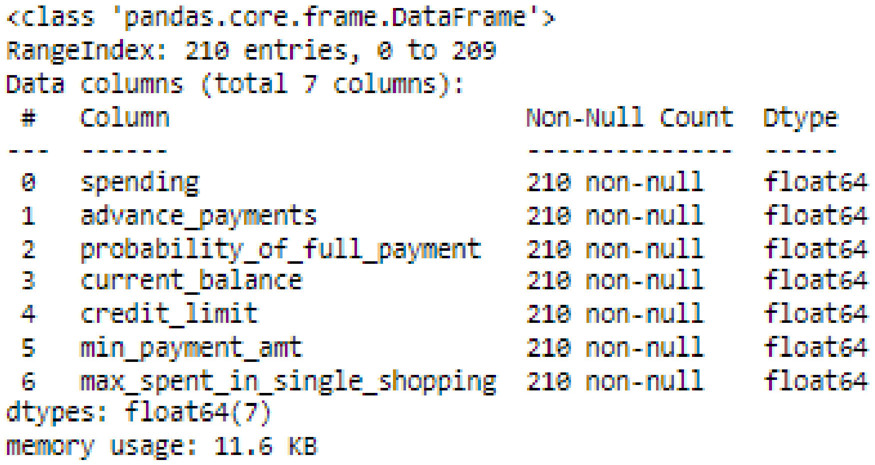
Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

ANSWER 1.1

Data Dictionary for Market Segmentation:

- 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)

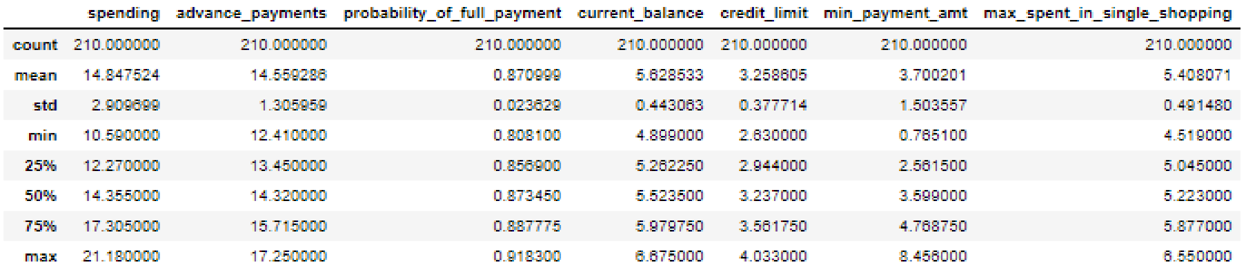
The dataset consists of 210 rows and 7 columns. All the columns are of datatype float which are spending, advance_payments, probability-of_full_payment , current_balance, credit_limit, min_payment_amt, max_spent_in_single_shopping. There are no null values and duplicates in the data.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Figure 1: Data set Info

The mean spending done by customers is 14.84 during past few months. The mean advance payments that are done are 14.55 which are a little less than mean spend but are almost inline. According to this data it can be said that most people are paying on time. This can also be confirmed with the mean probability of full payment which is 0.87. The mean of maximum amount spent in single shopping is 5.4. Looking at the maximum/minimum and median values of the data, it seems there are not a lot of outliers.



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

Figure 2: Dataset Description

Univariate Analysis of all variables

Spending

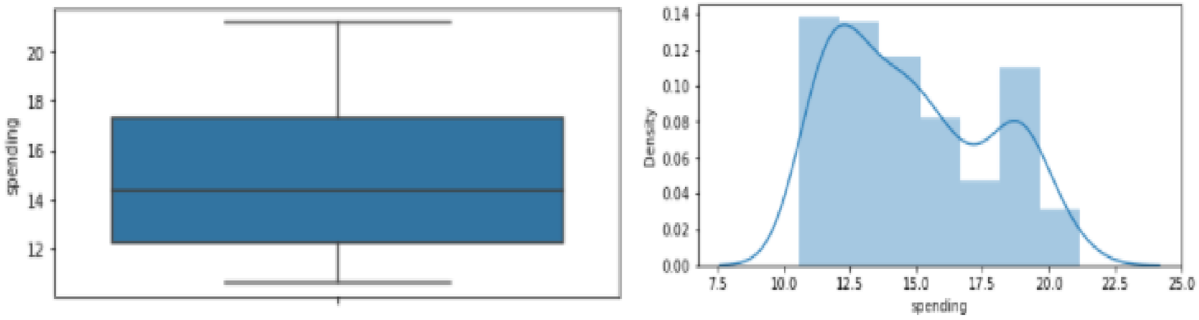


Figure 3: Variable "spending"

Observations: The boxplot of spending variable shows no outliers. The histogram shows the data is right skewed with 0.399 value of skewness.

Advance_Payments

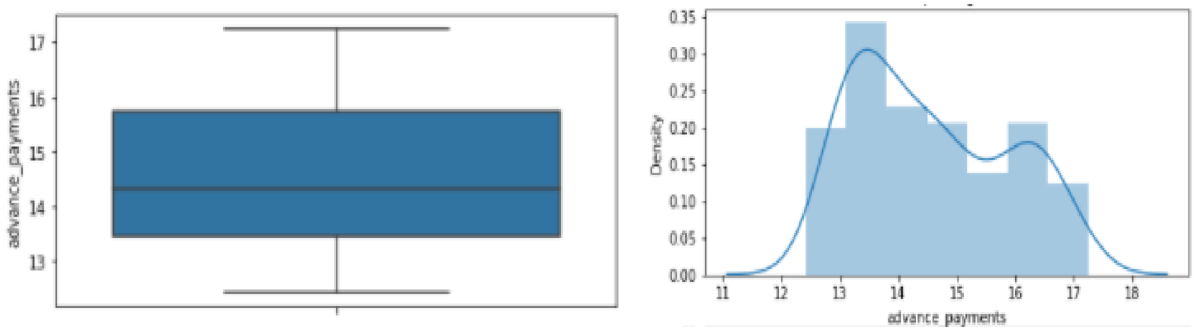


Figure 4: Variable "advance_Payments"

Observation: The boxplot of advance_payments variable shows no outliers. The histogram shows the data is right skewed with 0.386 value of skewness.

Probability_of_full_payment

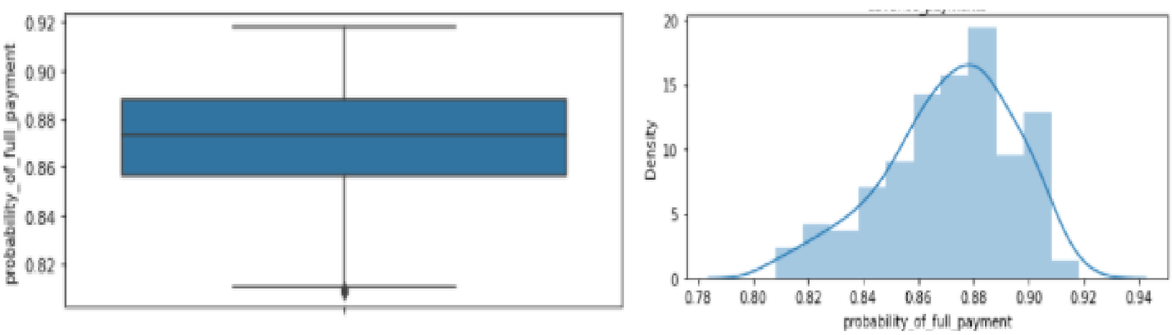


Figure 5: Variable "probability_of_full_payment"

Observation: The boxplot of probability_of_full_payments variable shows few outliers. The histogram shows the data is left skewed with -0.537 value of skewness.

Current_balance

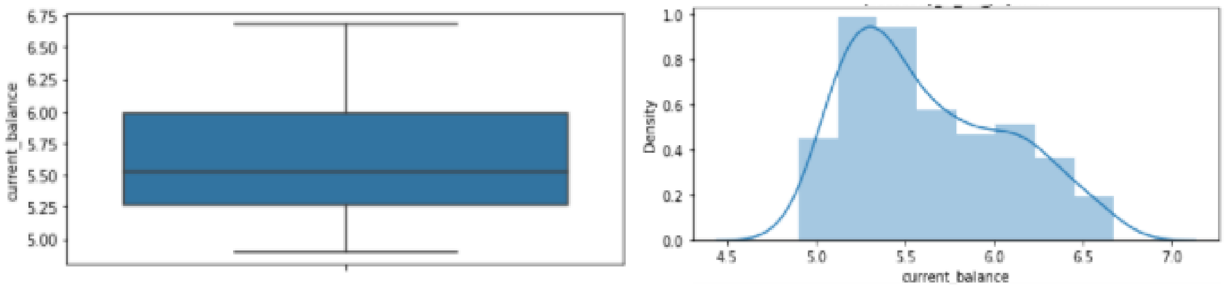


Figure 6: Variable "current_balance"

Observation: The boxplot of current_balance variable shows no outliers. The histogram shows the data is right skewed with 0.525 value of skewness.

Credit_limit

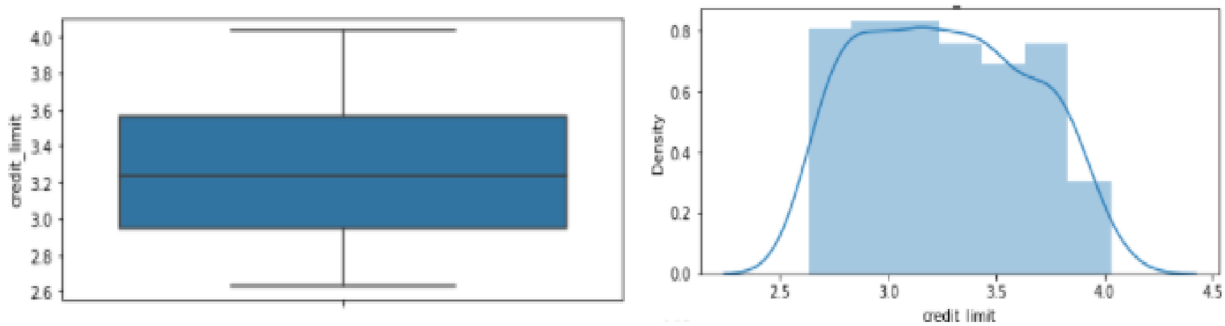


Figure 7: Variable "credit_limit"

Observation: The boxplot of credit_limit variable shows no outliers. The histogram shows the data is slightly right skewed with 0.134 value of skewness.

Min_payment_amount

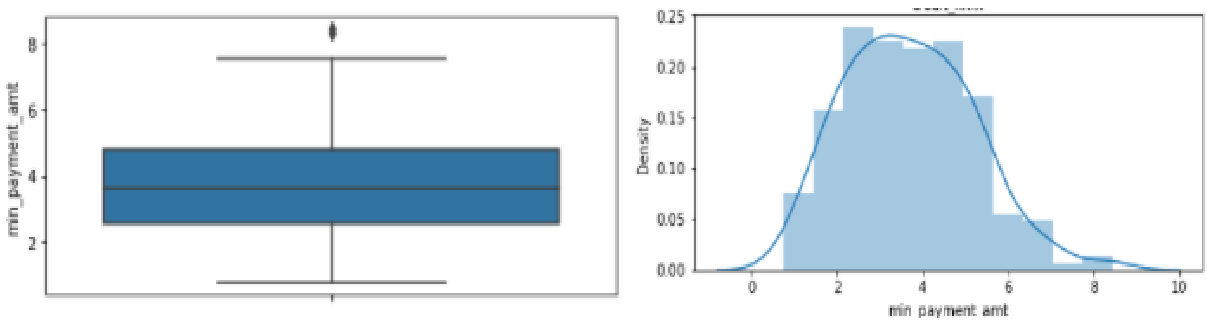


Figure 8: Variable "min_payment_amt"

Observation: The boxplot of Min_payment_amount variable shows few outliers. The histogram shows the data is slightly right skewed with 0.401 value of skewness.

Max_spent_in_single_shopping

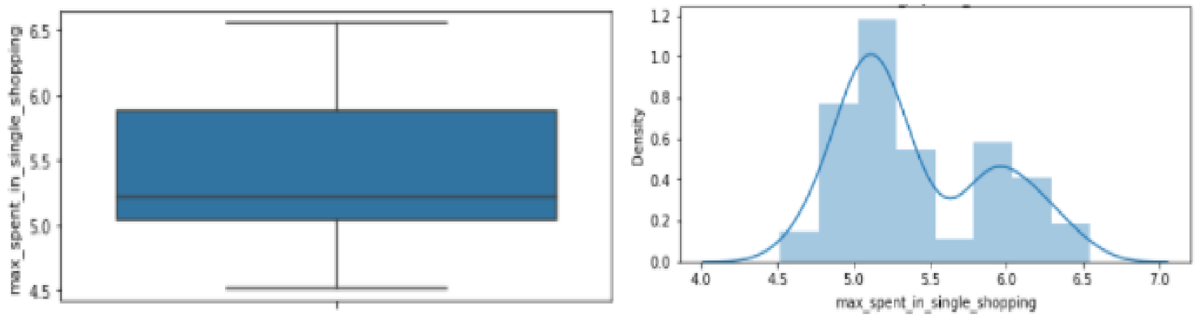
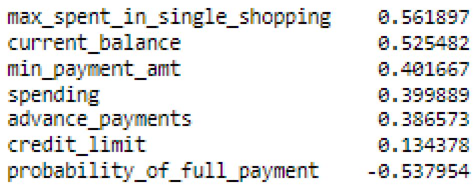


Figure 9: Variable "max_spent_in_single_shopping"

Observation: The boxplot of Max_spent_in_single_shopping variable shows no outliers. The histogram shows the data is slightly right skewed with 0.561 value of skewness.



```
max_spent_in_single_shopping    0.561897
current_balance                 0.525482
min_payment_amt                 0.401667
spending                        0.399889
advance_payments                0.386573
credit_limit                    0.134378
probability_of_full_payment    -0.537954
```

Figure 10: Skewness

Multivariate Analysis

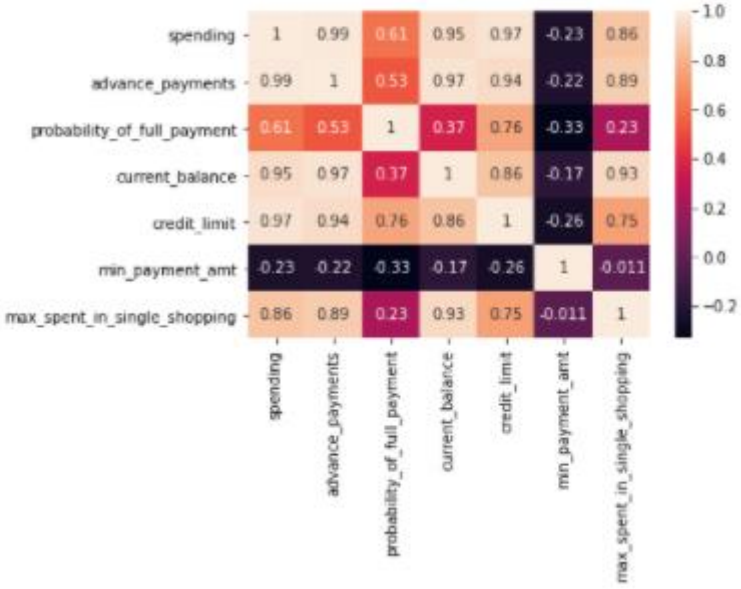


Figure 11: Heatmap

There is high correlation between

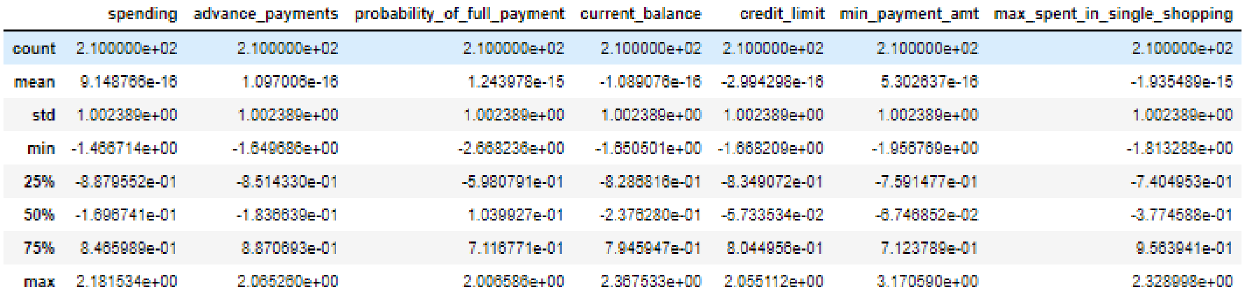
- Spending and advance payments
- Spending and current balance
- Spending and credit limit
- Advance payments and current balance
- Advance payment and credit limit
- Maxspent in singleshopping and current balance

QUESTION 1.2

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

ANSWER 1.2

Yes, Scaling is necessary to normalize the data. Clustering uses distance measure to determine whether the record belongs to the cluster or not. Therefore the clustering algorithm is affected by the scale of variables. Variables with high standard deviation will have a higher weight and variable with low standard deviation will have lower weight while performing clustering but this will not give the desired result. So all the variables in the dataset should have the same standard deviation so that while clustering all variables are weighted equally.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02
mean	9.148766e-16	1.097006e-16	1.243978e-15	-1.089076e-16	-2.994298e-16	5.302637e-16	-1.935489e-15
std	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00
min	-1.466714e+00	-1.649686e+00	-2.668236e+00	-1.650501e+00	-1.668209e+00	-1.956769e+00	-1.813288e+00
25%	-8.879552e-01	-8.514330e-01	-5.980791e-01	-8.286816e-01	-8.349072e-01	-7.591477e-01	-7.404953e-01
50%	-1.696741e-01	-1.836639e-01	1.039927e-01	-2.376280e-01	-5.733534e-02	-6.746852e-02	-3.774588e-01
75%	8.465989e-01	8.870693e-01	7.116771e-01	7.945947e-01	8.044956e-01	7.123789e-01	9.563941e-01
max	2.181534e+00	2.065260e+00	2.006586e+00	2.367533e+00	2.055112e+00	3.170590e+00	2.328998e+00

Figure 12: Scaled Data

Here we can see after scaling the standard deviation of all variables is same ensuring same weightage of all variables. The method used here for scaling is Standard Scalar.

QUESTION 1.3

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

Answer 1.3

Clustering is the patterns in the data are used to identify similar observations and group them together into a cluster.

On a mathematical front, distance between observations is calculated and observations nearest to each other form a cluster. Various methods of calculating distance are Euclidean distance, Manhattan distance and Chebyshev distance. The most widely used is the Euclidean distance which calculates distance using the Pythagoras theorem and calculates the hypotenuse distance. Hierarchical clustering also produces a useful graphical display of the clustering process known as dendrogram. Here using the ward linkage method which uses Euclidean distance a dendrogram is created to understand the clusters.

On x axis is observations and y axis is the distance.

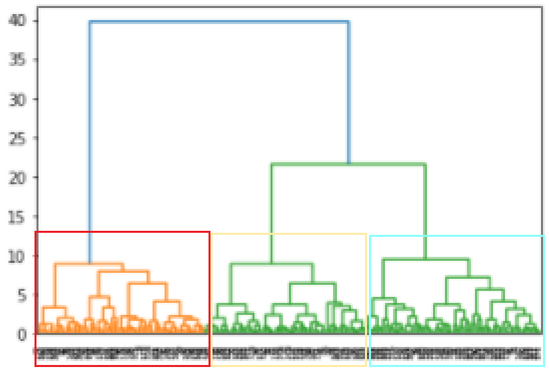


Figure 13: Dendrogram

The dendrogram has overgrown and created many small clusters but we can visually figure out there are major 3 clusters that are formed. To understand he clusters better and to have enough observations in each cluster the number of clusters has been defined to 10. The dendrogram created for this also gives us that there are 3 major clusters which can be differentiated visually.

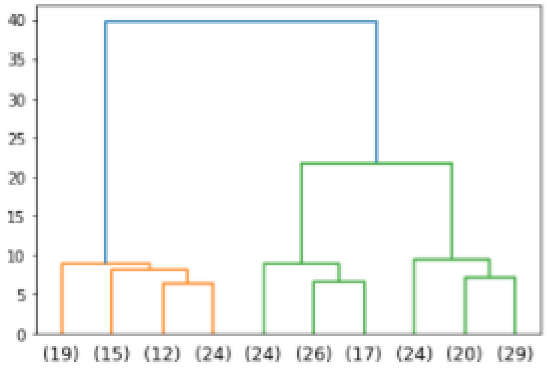


Figure 14: Dendrogram

Creating a dendrogram using another linkage method i.e. average

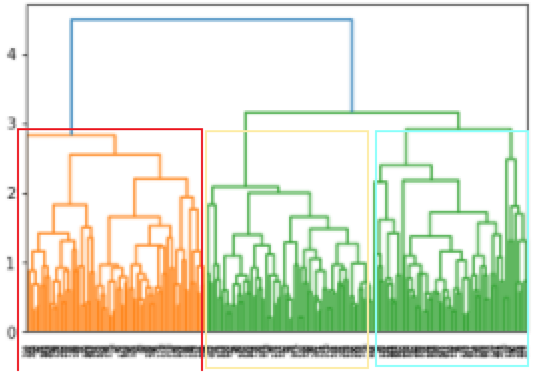


Figure 15: Dendrogram

Here in this method also we can visually differentiate between 3 major clusters so there is not much difference that can be seen in ward and average method.

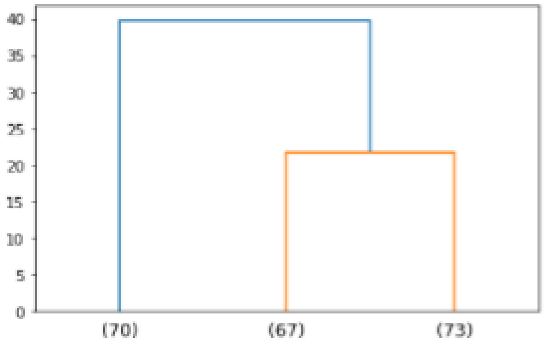
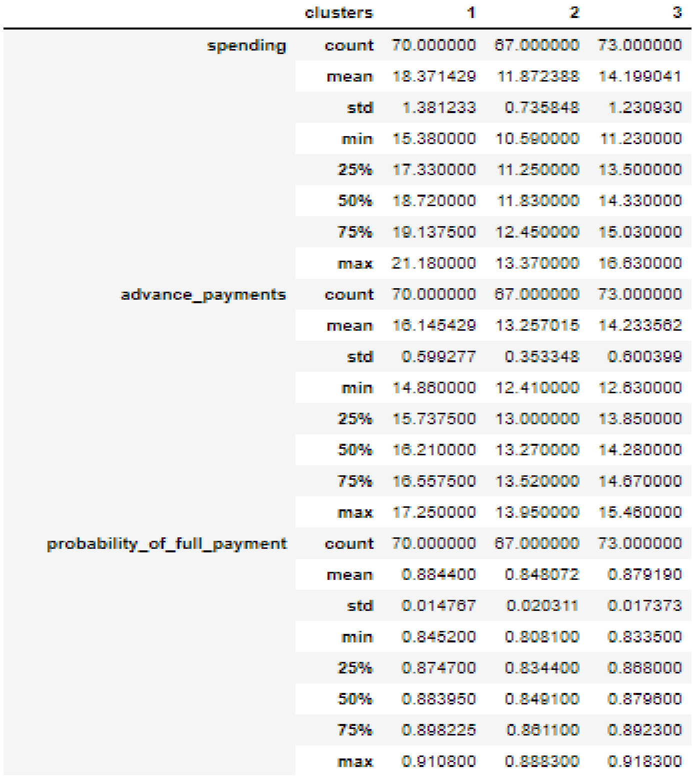


Figure 16: Final Dendrogram

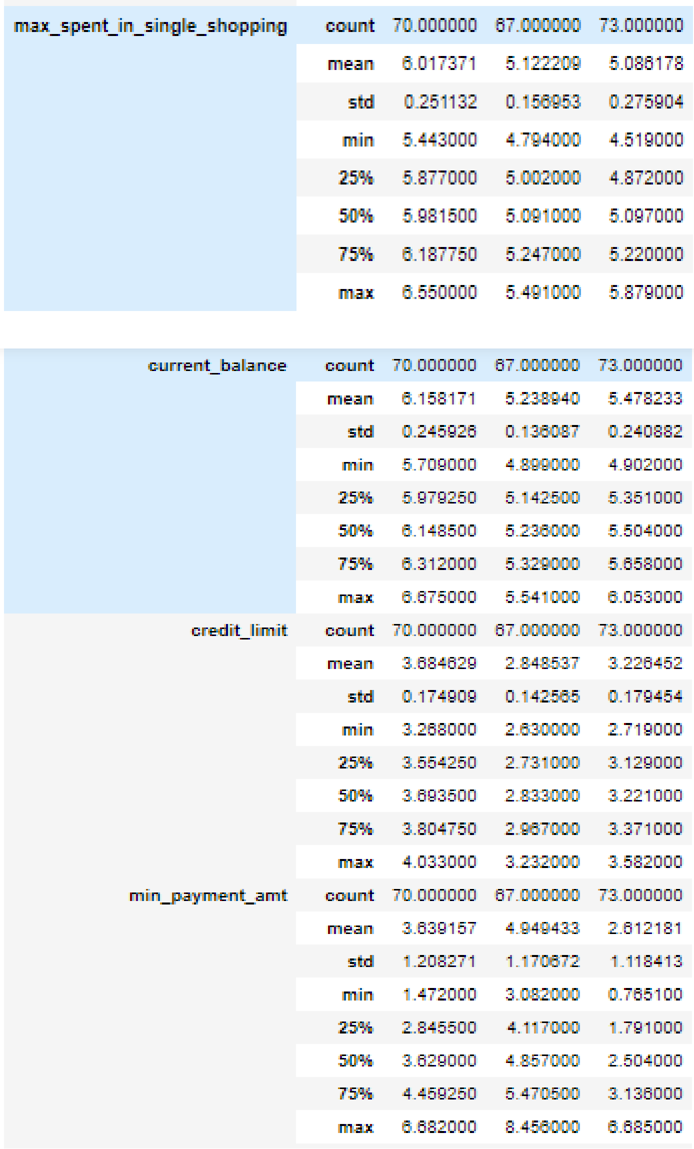
The 3 clusters have been formed each consisting of 70, 67, 73 records respectively using ward method.

Fclusters has been used to assign the cluster number to each record. The criterion used here is maxclust which allows giving the maximum number of clusters that need to be formed. The number of clusters formed has been preset as 3 within fclusters algorithm.

The below figure gives the description of clusters for each feature and different means and standard deviations defined for each cluster.



	clusters	1	2	3
spending	count	70.000000	67.000000	73.000000
	mean	18.371429	11.872388	14.199041
	std	1.381233	0.735848	1.230930
	min	15.380000	10.590000	11.230000
	25%	17.330000	11.250000	13.500000
	50%	18.720000	11.830000	14.330000
	75%	19.137500	12.450000	15.030000
	max	21.180000	13.370000	16.630000
advance_payments	count	70.000000	67.000000	73.000000
	mean	16.145429	13.257015	14.233562
	std	0.599277	0.353348	0.600399
	min	14.860000	12.410000	12.630000
	25%	15.737500	13.000000	13.850000
	50%	16.210000	13.270000	14.280000
	75%	16.557500	13.520000	14.670000
	max	17.250000	13.950000	15.460000
probability_of_full_payment	count	70.000000	67.000000	73.000000
	mean	0.884400	0.848072	0.879190
	std	0.014767	0.020311	0.017373
	min	0.845200	0.808100	0.833500
	25%	0.874700	0.834400	0.868000
	50%	0.883950	0.849100	0.879600
	75%	0.898225	0.861100	0.892300
	max	0.910800	0.888300	0.918300



max_spent_in_single_shopping	count	70.000000	67.000000	73.000000
	mean	6.017371	5.122209	5.086178
	std	0.251132	0.156953	0.275904
	min	5.443000	4.794000	4.519000
	25%	5.877000	5.002000	4.872000
	50%	5.981500	5.091000	5.097000
	75%	6.187750	5.247000	5.220000
	max	6.550000	5.491000	5.879000
current_balance	count	70.000000	67.000000	73.000000
	mean	6.158171	5.238940	5.478233
	std	0.245926	0.136087	0.240882
	min	5.709000	4.899000	4.902000
	25%	5.979250	5.142500	5.351000
	50%	6.148500	5.236000	5.504000
	75%	6.312000	5.329000	5.658000
	max	6.675000	5.541000	6.053000
credit_limit	count	70.000000	67.000000	73.000000
	mean	3.684629	2.848537	3.226452
	std	0.174909	0.142565	0.179454
	min	3.268000	2.630000	2.719000
	25%	3.554250	2.731000	3.129000
	50%	3.693500	2.833000	3.221000
	75%	3.804750	2.967000	3.371000
	max	4.033000	3.232000	3.582000
min_payment_amt	count	70.000000	67.000000	73.000000
	mean	3.639157	4.949433	2.612181
	std	1.208271	1.170672	1.118413
	min	1.472000	3.082000	0.765100
	25%	2.845500	4.117000	1.791000
	50%	3.629000	4.857000	2.504000
	75%	4.459250	5.470500	3.136000
	max	6.682000	8.456000	6.685000

Figure 17: Cluster Description

Cluster 1 denotes people who have high spending patterns, high current balance, maximum credit limit and they do highest amounts of advance payments.

Cluster 2 denotes people who have low spending patterns, low current balance, least credit limit and they do lowest amounts of advance payments.

Cluster 3 denotes people who have medium spending patterns, average current balance, medium credit limit and they do medium amounts of advance payments.

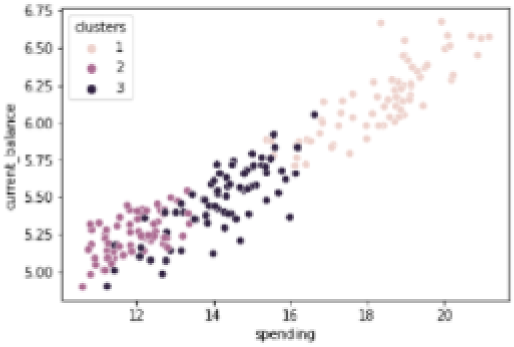


Figure 18: Clusters Scatter plot

The scatter plot between spending and current_balance identifies the different clusters that have formed. Both the features are highly correlated and same way the clusters are defined. Low spending with low current_balance is one cluster and same for other clusters. There are some datapoints that overlap but clusters formed depend on other features as well.

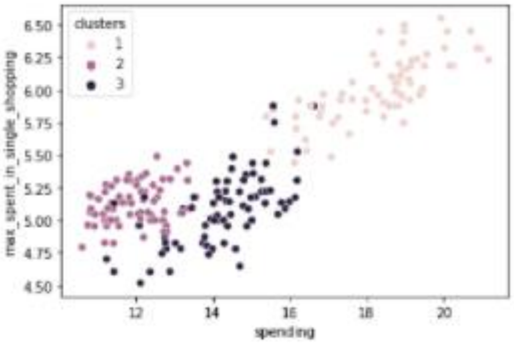


Figure 19: Clusters Scatter plot

The scatter plot between spending and max_spent_in_single_shopping identifies the different clusters that have formed. Both the features are highly correlated and same way the clusters are defined. Low spending with low max_spent_in_single_shopping is one cluster and same for other clusters. There are some datapoints that overlap but clusters formed depend on other features as well.

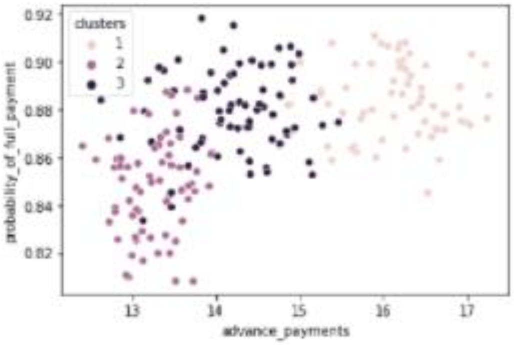


Figure 20: Clusters Scatter plot

The scatter plot between advance_payments and probability_of_full_payment identifies than clusters are more dependent on amount of advance_payment than the probability.

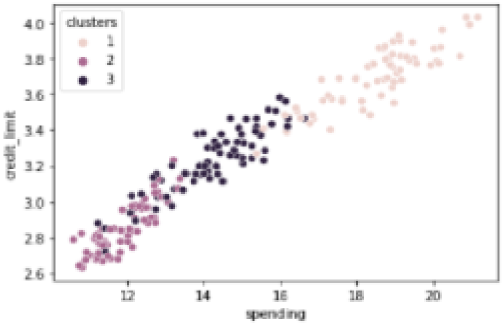


Figure 21: Clusters Scatter plot

The scatter plot between spending and credit_limit identifies the different clusters that have formed. Both the features are highly correlated and same way the clusters are defined. Low spending with low credit_limit is one cluster and same for other clusters. There are some data points that overlap but clusters formed depend on other features as well.

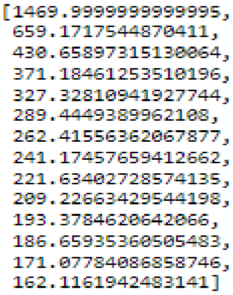
QUESTION 1.4

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.

ANSWER 1.4

Kmeans is a non hierarchical method to form clusters. This works on the concept of within sum of squares. Means in kmeans refers to finding the centroid of the data.

The below figure shows that within sum of squares (wss) for number of clusters =1 to 14. From this we can see there is a significant drop in wss value when clusters changed from 1 to 2. Similarly a drop is visible from 2 to 3 but going forward from here the drop is not significant enough when moving the number of clusters from 3 to 4 and so on. Therefore optimum number of clusters can be taken as 3 or 4.



```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.18461253510196,
 327.32810941927744,
 289.4449389962108,
 262.41556362067877,
 241.17457659412662,
 221.63402728574135,
 209.22663429544198,
 193.3784620642066,
 186.65935360505483,
 171.07784086858746,
 162.1161942483141]
```

Figure 22: WSS

The same can be understood by the elbow curve below where it is visible that there is no significant drop in wss values after 3 clusters.

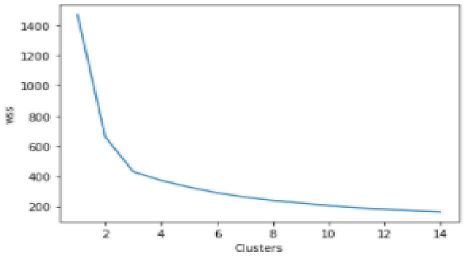
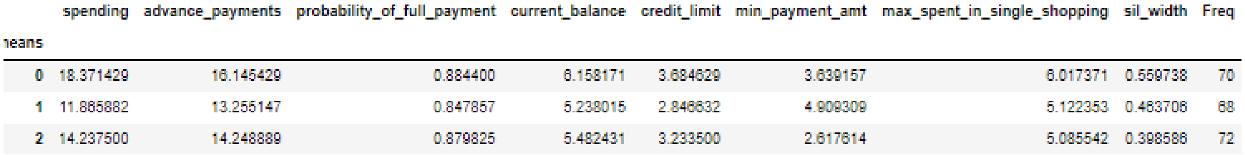


Figure 23: Elbow Curve

- The silhouette score for 2 clusters is 0.465
- The silhouette score for 3 clusters is 0.400
- The silhouette score for 4 clusters is 0.329

More the silhouette score better are the clusters but having only 2 clusters will not give any business implications so Kmeans clustering is done using 3 clusters.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	sil_width	Freq
ıeans									
0	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	0.559738	70
1	11.865882	13.255147	0.847857	5.238015	2.846632	4.909309	5.122353	0.463706	68
2	14.237500	14.248889	0.879825	5.482431	3.233500	2.617614	5.085542	0.398586	72

Figure 24: Cluster means

Cluster 1 denotes people who have high spending patterns, high current balance, maximum credit limit and they do highest amounts of advance payments.
Cluster 2 denotes people who have low spending patterns, low current balance, least credit limit and they do lowest amounts of advance payments
Cluster 3 denotes people who have medium spending patterns, average current balance, medium credit limit and they do medium amounts of advance payments

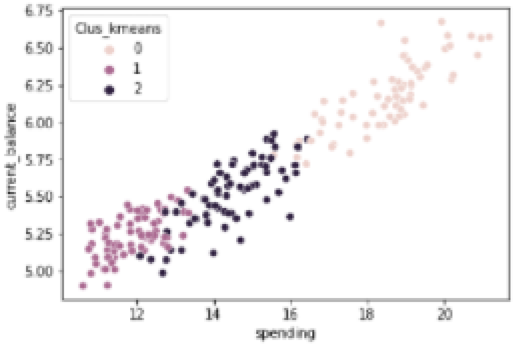


Figure 25: Clusters Scatter plot

The scatter plot between spending and current_balance identifies the different clusters that have formed. Both the features are highly correlated and same way the clusters are defined. Low spending with low current_balance is one cluster and same for other clusters. There are some datapoints that overlap but clusters formed depend on other features as well.

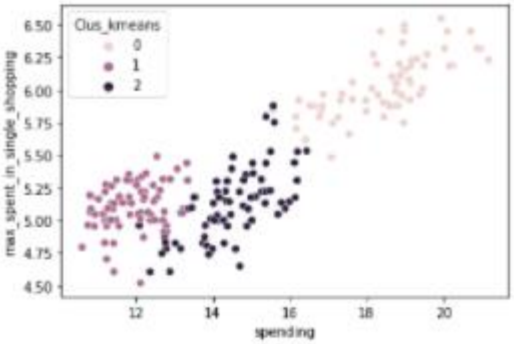


Figure 26: Clusters Scatter plot

The scatter plot between spending and max_spent_in_single_shopping identifies the different clusters that have formed. Both the features are highly correlated and same way the clusters are defined. Low spending with low max_spent_in_single_shopping is one cluster and same for other clusters. There are some datapoints that overlap but clusters formed depend on other features as well.

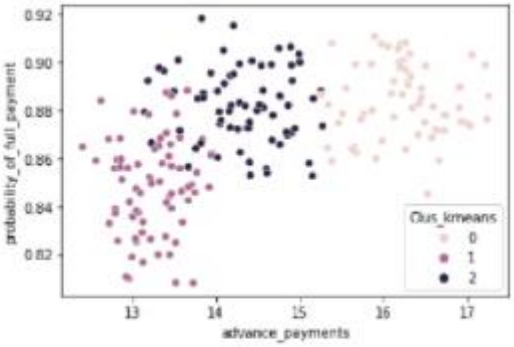


Figure 27: Clusters Scatter plot

The scatter plot between advance_payments and probability_of_full_payment identifies than clusters are more dependent on amount of advance_payment than the probability.

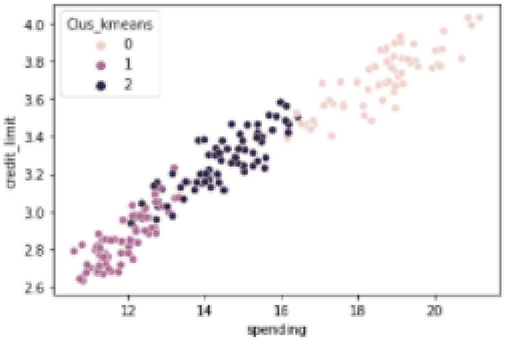


Figure 28: Clusters Scatter plot

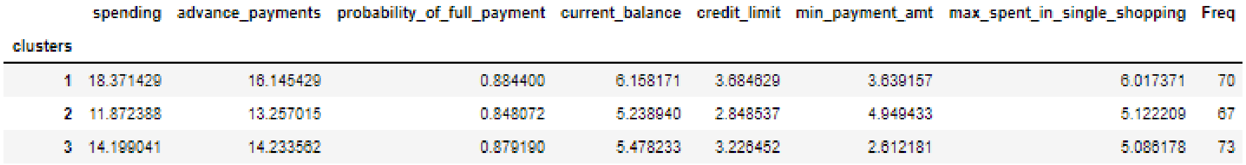
The scatter plot between spending and credit_limit identifies the different clusters that have formed. Both the features are highly correlated and same way the clusters are defined. Low spending with low credit_limit is one cluster and same for other clusters. There are some data points that overlap but clusters formed depend on other features as well.

QUESTION 1.5

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

ANSWERS 1.5

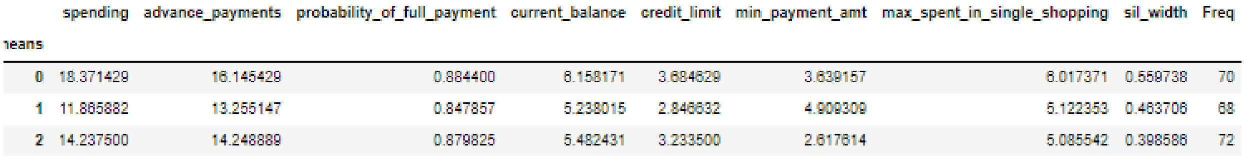
Hierarchical clusters



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

Figure 29: Cluster means

Kmeans clusters



neans	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	sil_width	Freq
0	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	0.559738	70
1	11.865882	13.255147	0.847857	5.238015	2.846632	4.909309	5.122353	0.463706	68
2	14.237500	14.248889	0.879825	5.482431	3.233500	2.617614	5.085542	0.398586	72

Figure 30: Cluster means

Promotional strategies for both hierarchical and kmeans clusters:-

High spenders:

- Initiate reward points on every purchase which will increase credit limit as well customers can use reward points for availing discounts. Validity of reward points should be quarterly or half yearly so customer should spend more to earn more rewards as current balance of these customers is also high. Extra reward points on advance payments
- The Bank should tie up with luxury brands ,so with every purchase both bank and spender benefits.
- Customers should get extra discount offers when paying with the bank credit/debit card.
- Credit limit and loans based on their spending and advance payments should be increased.

Medium Spenders:

- Bank should tie up with grocery shops to provide offers with their credit cards.
- Setting up payment targets. For example: if a customer does spend a certain amount in the current month, he gets loyalty bonus and increased credit limit.
- Promote premium cards with better benefits like discount on movies/petrol.

Low Spenders

- Tie up with grocery stores and utilities (things used in everyday life like electricity, sewerage, gas, phone bills) which will encourage them to do payments as a certain discount is provided to them which will even attract more customers of the same spending patterns.
- Reward points on advance payments.
- Low interest rate than other banks will help to improve loyalty among customers and high spending amounts.
- Cashback offers if they do a payment above a certain amount.

Problem 2: CART-RF-ANN

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

QUESTION 2.1

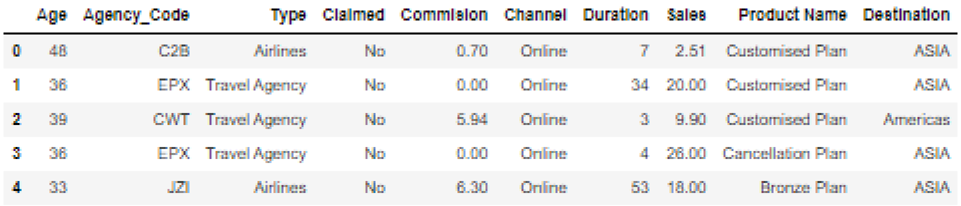
Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)

ANSWER 2.1

DATA DICTIONARY of the features

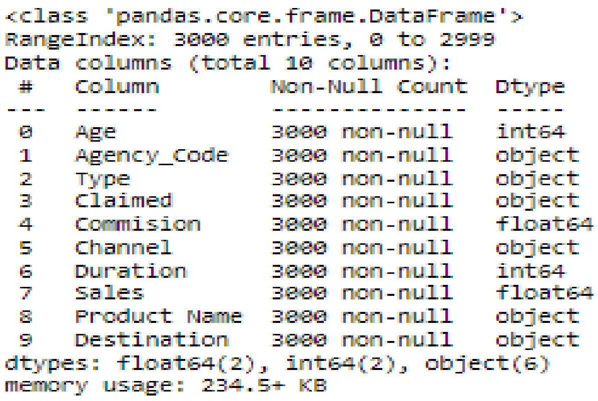
1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency_Code)
3. Type of tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)
6. Duration of the tour (Duration in days)
7. Destination of the tour (Destination)
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)(Sales)
9. The commission received for tour insurance firm (Commission is in percentage of sales)
10.Age of insured (Age)

The head of data to see if it has been loaded properly.



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

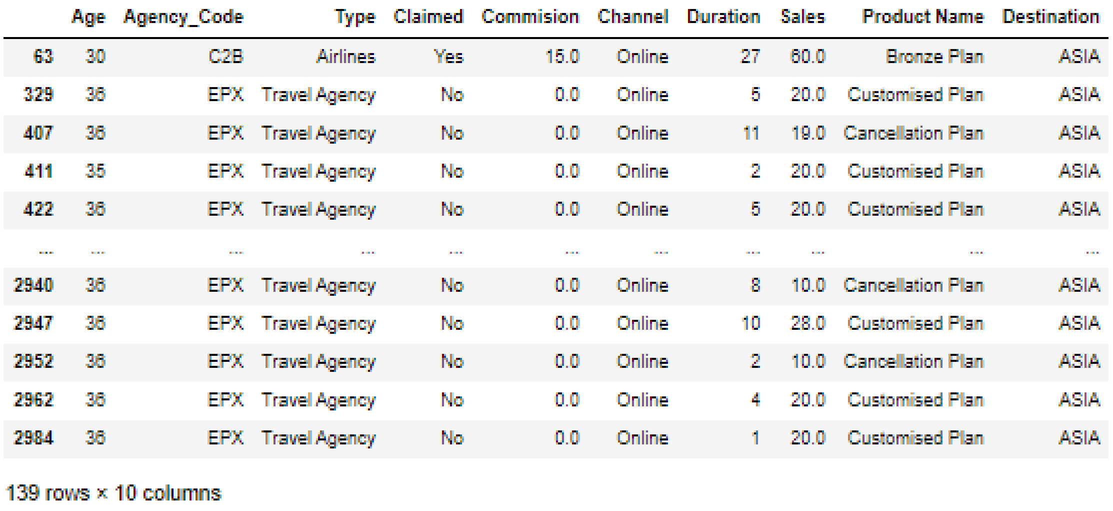
Figure 31: Head of data



```
<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
```

Figure 32: Dataset Information

The insurance dataset has 10 variables out of which 4 variables are numeric and 6 variables are categorical. There are no null values .There are 139 duplicated records. The below figure shows the duplicate record. The duplicates have not been dropped because there is no unique identifier (like Customer ID, Customer Name, Booking ID) to help understand these records are actually duplicates or these are for different customers.

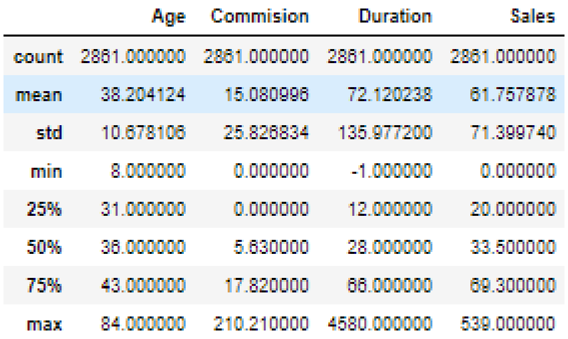


	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
63	30	C2B	Airlines	Yes	15.0	Online	27	60.0	Bronze Plan	ASIA
329	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
407	36	EPX	Travel Agency	No	0.0	Online	11	19.0	Cancellation Plan	ASIA
411	35	EPX	Travel 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

Figure 33: Duplicate records

Using describe function a basic understanding of the dataset can be drawn. The variables are in different scales and there is significant difference in 75% values and max values which signifies there are outliers.



	Age	Commision	Duration	Sales
count	2861.000000	2861.000000	2861.000000	2861.000000
mean	38.204124	15.080996	72.120238	61.757878
std	10.678106	25.826834	135.977200	71.399740
min	8.000000	0.000000	-1.000000	0.000000
25%	31.000000	0.000000	12.000000	20.000000
50%	36.000000	5.630000	28.000000	33.500000
75%	43.000000	17.820000	66.000000	69.300000
max	84.000000	210.210000	4580.000000	539.000000

Figure 34: Dataset Description

A count of the target variable 'Claimed' to understand how many customers have claimed insurance. Below is the count plot of the target variable Claimed.

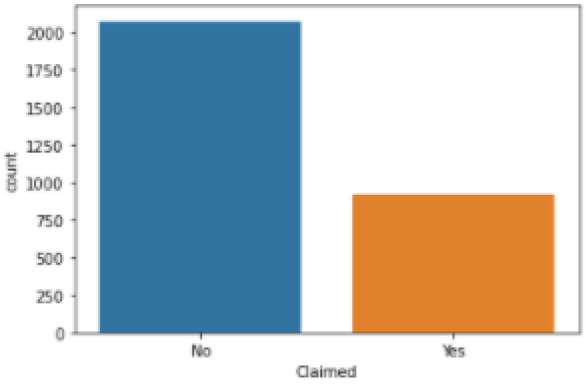
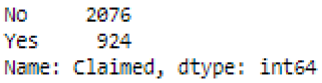


Figure 35: Variable "Claimed"



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

Figure 36: Value Count

Customers who have claimed are 924 and not claimed are 2076. This amounts to almost 33% of the records. The data is balanced between claimed and not claimed.

Univariate Analysis

Age

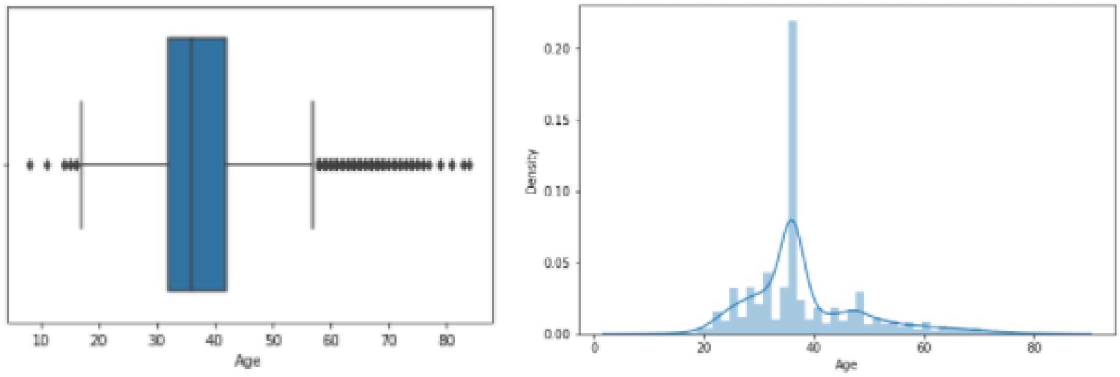


Figure 37: Variable "Age"

The boxplot shows a lot of outliers for variable Age. The distplot shows the distribution is right skewed withh a skewness value of 1.14

Duration

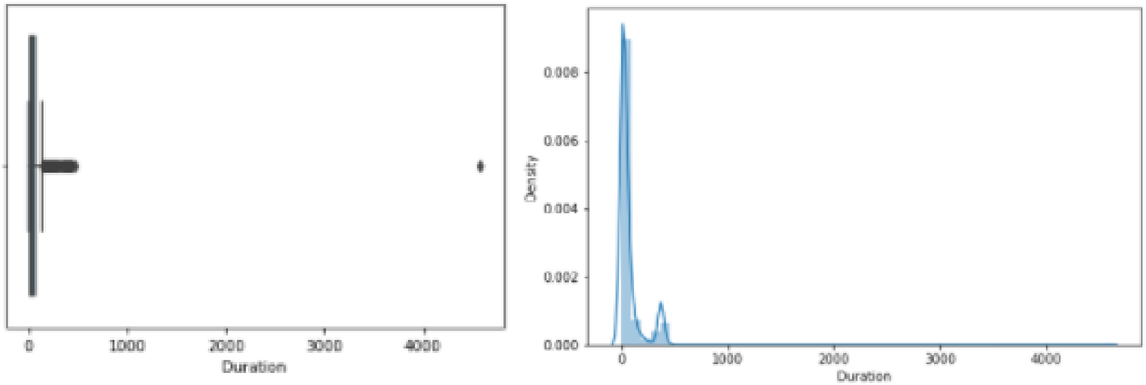


Figure 38: Variable "Duration"

- The boxplot shows there are a lot of outliers and one record has duration of over 4000.
- The distplot shows that distribution is highly right skewed with a value of skewness of 13.78 as most of the datapoints lie from 0 to 100 but the few exceeding this have a very high value, therefore a very high skewness value.
- There is a record with a value of duration 4580 which seems to be a data entry error but as all other outliers have not been treated, assuming this outlier is also a correct record this has not been dropped or treated. A way can be to treat this with median value when all the outliers are being treated.

Commission

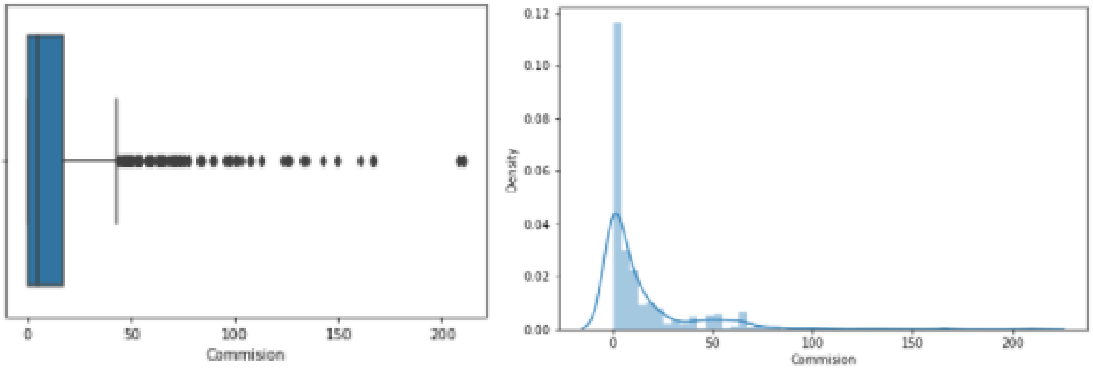


Figure 39: Variable "Commission"

The boxplot shows there are a lot of outliers. The distplot shows that distribution is right skewed with a skewness value of 3.14

Sales

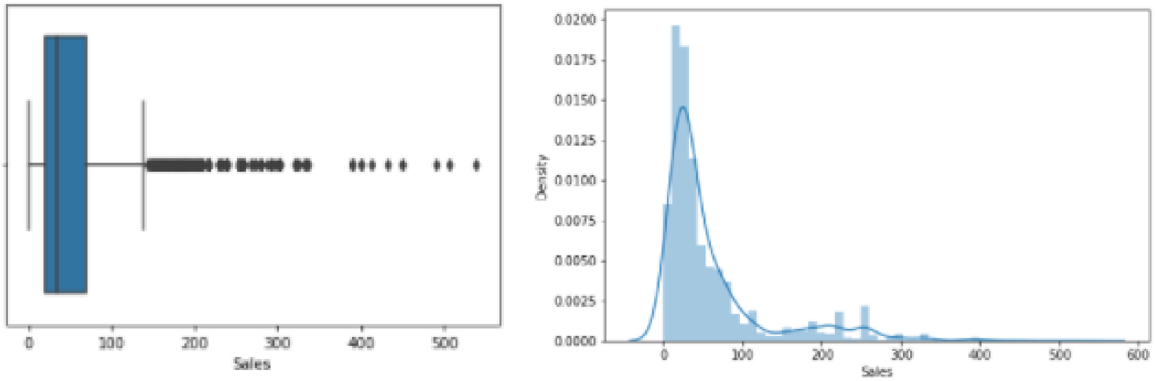
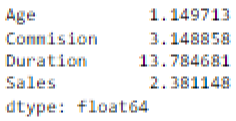


Figure 40: Variable "Sales"

The boxplot shows there are a lot of outliers. The distplot shows that distribution is right skewed with the skewness value of 2.38.



```
Age            1.149713
Commision      3.148858
Duration      13.784681
Sales          2.381148
dtype: float64
```

Figure 41: Skewness

Categorical variables

Product Name

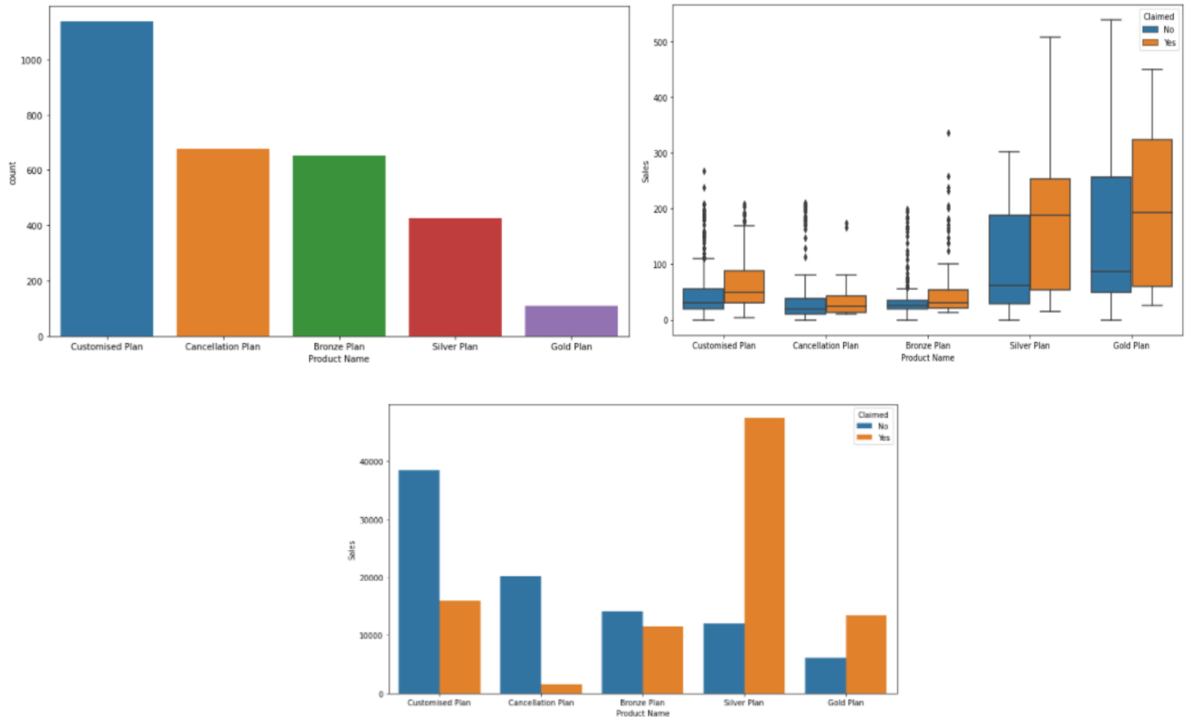
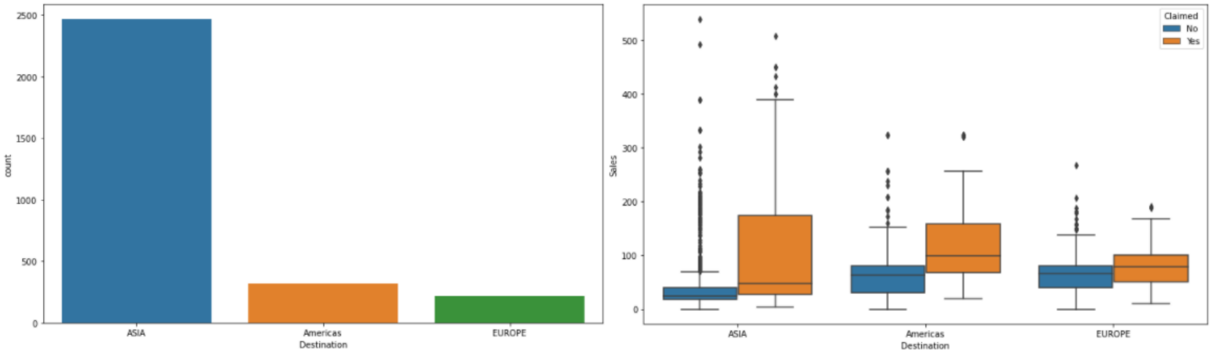


Figure 42: Variable "Product Name"

From above plots of product name it is identified that customized plan is the most preferred by customers and least preferred is gold plan. Insurance claims for silver plan and gold plan is very high. Cancellation plan is the best performing plan with low number of claims and good amount of sales. The boxplot between Product Name and Sales shows that the range for sales for gold plan is the maximum and claims in gold plan have also been high.

Destination



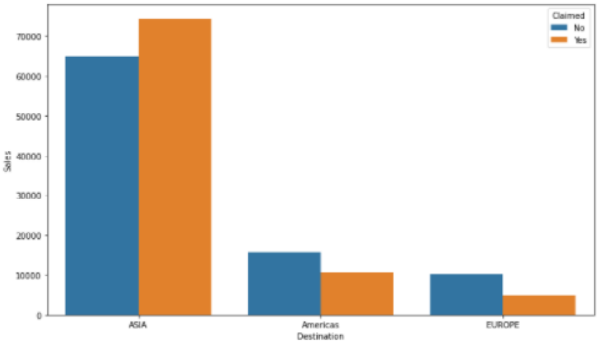


Figure 43: Variable "Destination"

Asia is the most frequent destination that customers go to followed by America destination and Europe. Insurance claims have been very high for ASIA destination although overall sales is highest for ASIA but the company is at a loss here because all these claims will affect the profits.

Channel

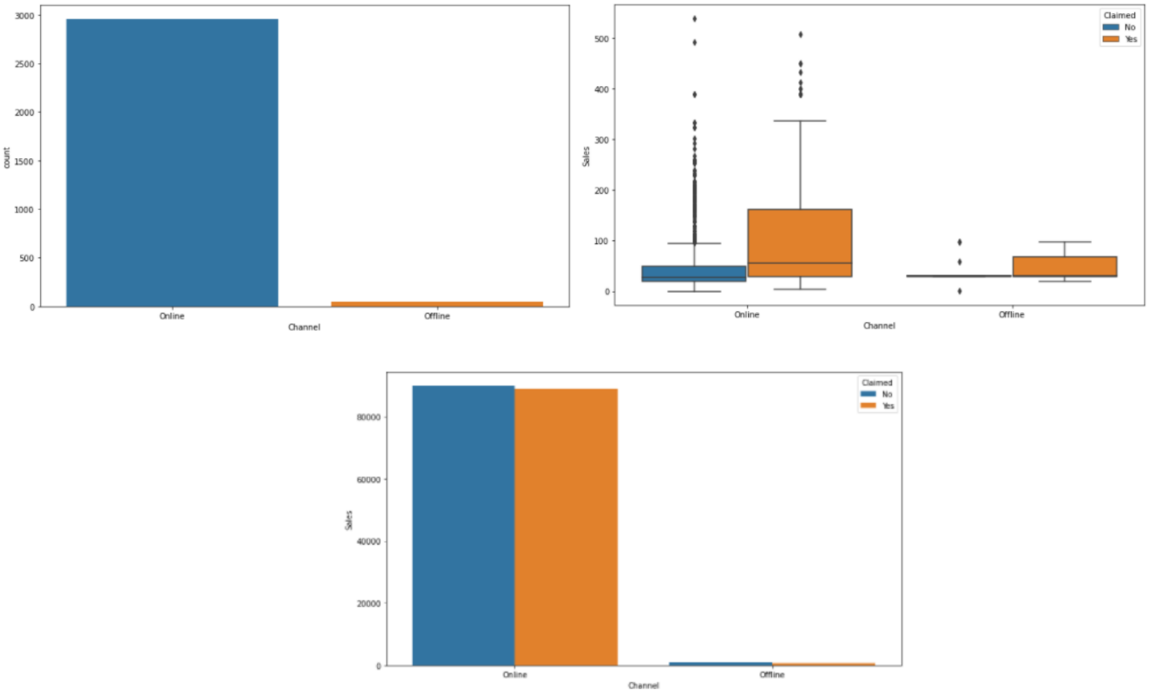


Figure 44: Variable "Channel"

The channel used by tour agencies to sell tour insurances is maximum through online and very few use offline. The ratio of claimed to not claimed in sales of online mode is almost same which is affecting the profits.

Agency_Code

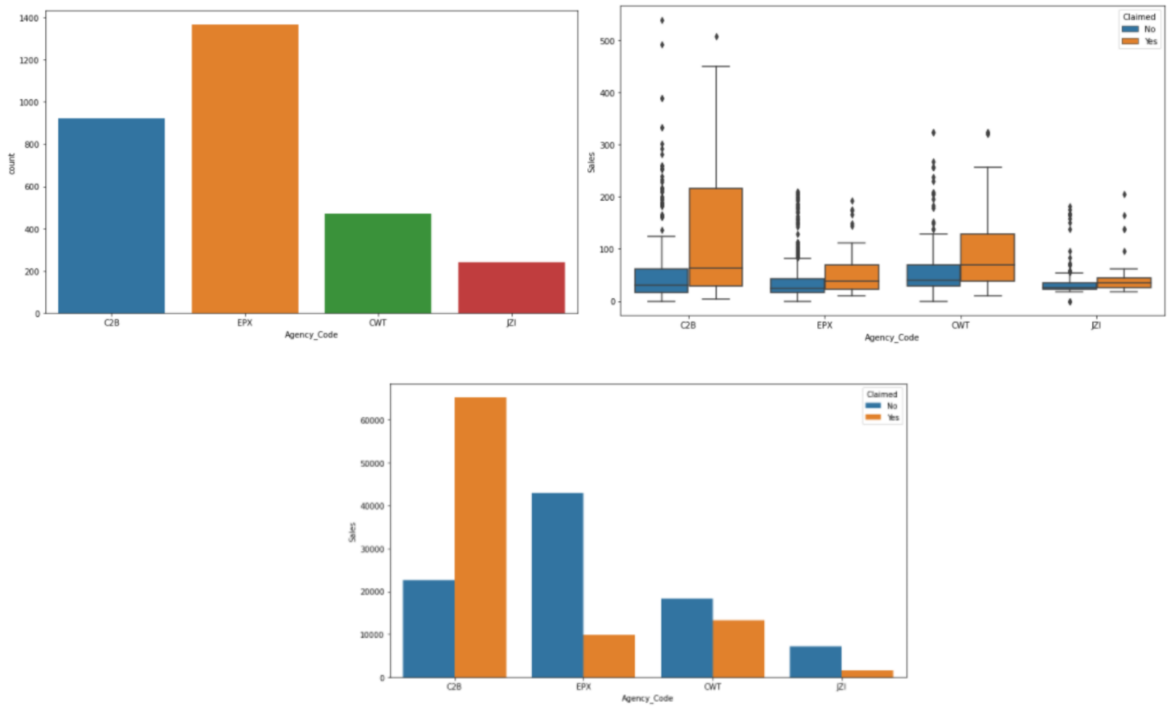
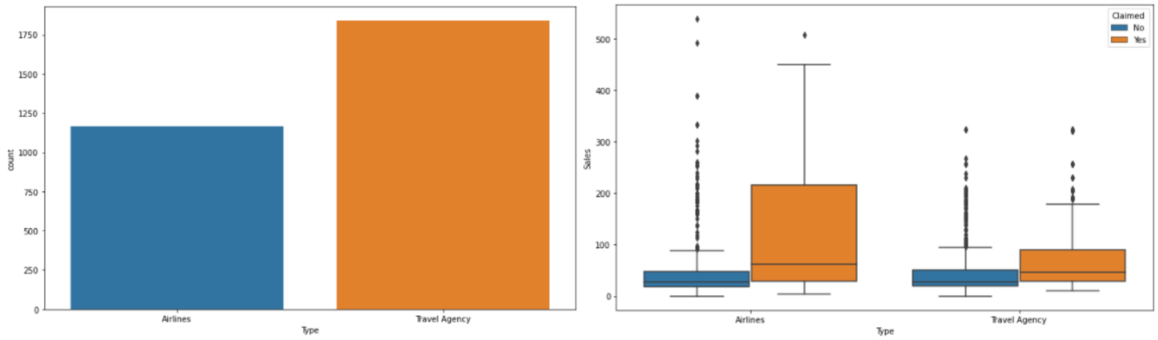


Figure 45: Variable “Agency_Code”

The agency which is frequently used is EPX followed by C2B. The boxplot between sales and agency shows that median of sales for claimed insurances is high than not claimed. The ratio of claimed to not claimed in sales for C2B is very high, therefore this agency is contributing to the most number of claims. EPX agency is performing really good with high sales and less number of claims.

Type



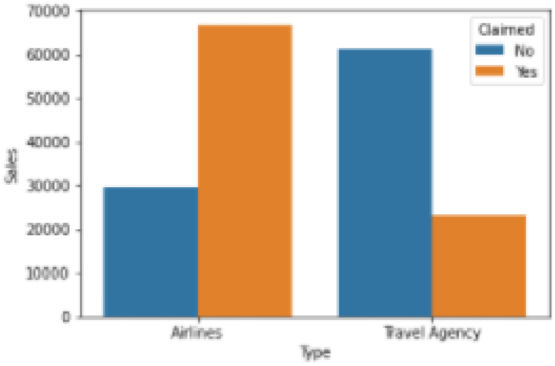


Figure 46: Variable "Type"

The customers have opted to book more from the travel agency than airlines. The boxplot between type and sales shows the sales range is higher for airlines and claims have also been high. Insurance claims in airlines have been very high compared to travel agency.

Multivariate Analysis:

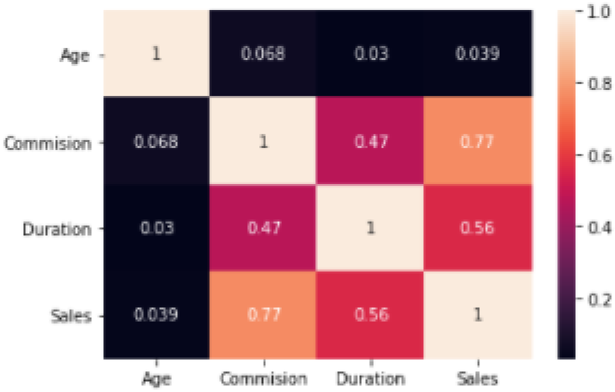


Figure 47: Heatmap

There is a strong correlation between Sales and Commission which is understood because higher the number of sales more is the commission that company will get.

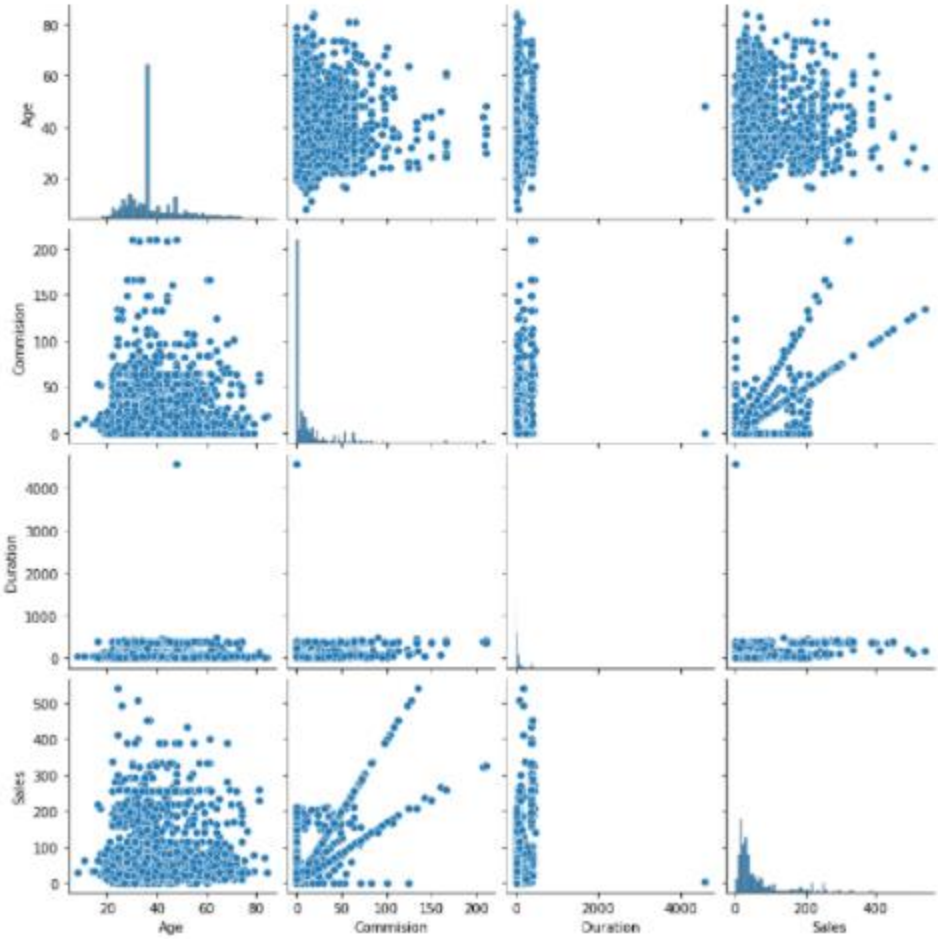


Figure 48: Pairplot

The pairplot above shows the same that there is high correlation between variables Sales and Commission.

QUESTION 2.2

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

ANSWER 2.2

To build the models the Categorical values have to be converted to numerical values. Below are the categorical code and the corresponding numeric code.

Variable: Agency_code

C2B, CWB, EPX, JZI : 0,1,2,3

Variable Type:

Airlines, Travel Agency:0,1

Variable Claimed:

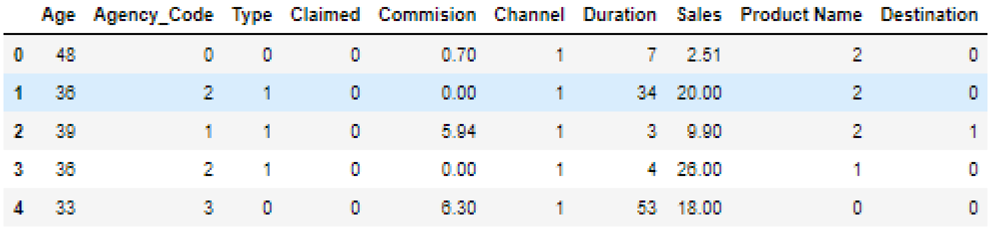
No,Yes: 0,1

Variable Channel:

Online, Offline: 0, 1

Variable Product Name:

Customized Plan, Cancellation Plan, Bronze Plan, Silver Plan, Gold Plan: 2,1,0,4,3



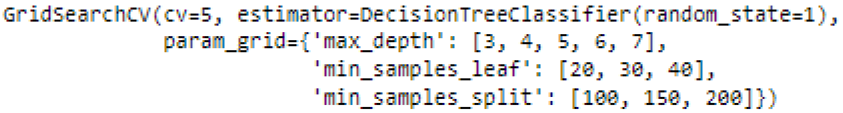
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0	0.70	1	7	2.51	2	0
1	36	2	1	0	0.00	1	34	20.00	2	0
2	39	1	1	0	5.94	1	3	9.90	2	1
3	36	2	1	0	0.00	1	4	26.00	1	0
4	33	3	0	0	6.30	1	53	18.00	0	0

Figure 49: Converted Dataset

The data has been split into two parts, one being the independent variables and other is the target variable ‘Claimed’. Using train_test_split function a function has been splited into 70% train and 30% test.

Decision Tree

A GridSearchCV has been used to determine the best parameters that can give a better precision, recall, f1score and accuracy for DecisionTreeCassifier.



```
GridSearchCV(cv=5, estimator=DecisionTreeClassifier(random_state=1),
             param_grid={'max_depth': [3, 4, 5, 6, 7],
                         'min_samples_leaf': [20, 30, 40],
                         'min_samples_split': [100, 150, 200]})
```

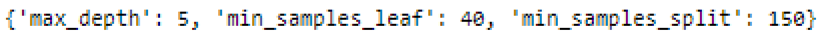
Figure 50: Grid Search CV

The best parameters for Decision Tree model are as:

Max_depth = 4 the maximum depth to which the decision tree should grow.

Min_samples_leaf = 30 denotes the minimum samples that have to be present after a node is split.

Min_samples_split = 150 denotes the minimum samples that need to be present in a node for the node to split.

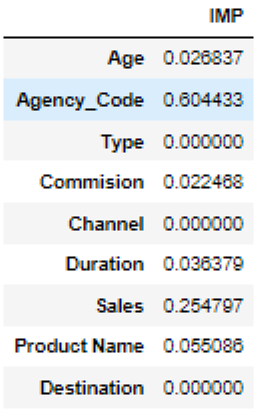


```
{'max_depth': 5, 'min_samples_leaf': 40, 'min_samples_split': 150}
```

Figure 51: Best Parameters

The most important features are Agency_code, Sales. The variables which do not have much effect on target variable are Age, Type, Channel, and Destination.

The features importance for a Decision Tree model is in the figure below.

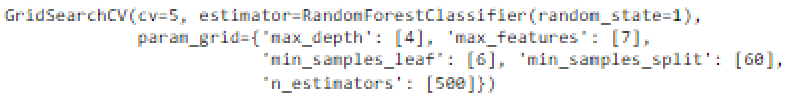


	IMP
Age	0.026837
Agency_Code	0.604433
Type	0.000000
Commision	0.022468
Channel	0.000000
Duration	0.036379
Sales	0.254797
Product Name	0.055086
Destination	0.000000

Figure 52: Feature Importance

Random Forest

GridSearchCV is done for the model Random Forest as well to identify the best parameters for the performance of the model.

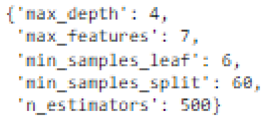


```
GridSearchCV(cv=5, estimator=RandomForestClassifier(random_state=1),
             param_grid={'max_depth': [4], 'max_features': [7],
                         'min_samples_leaf': [6], 'min_samples_split': [60],
                         'n_estimators': [500]})
```

Figure 53: Grid Search CV

After tuning the model by trying number of parameters, the parameters that were best for Random Forest are:

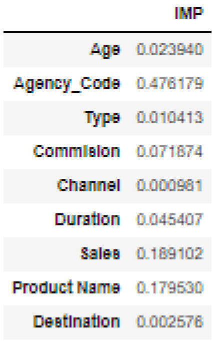
- Max_depth = 4 which denotes the maximum depth to which the decision tree should grow.
- Max_features = 7 which tells the model to take only 7 randomly selected features at a time to make a decision tree.
- Min_samples_leaf = 6 which denotes the minimum samples that have to be present after a node is split.
- Min_samples_split = 60 denotes the minimum samples that need to be present in a node for the node to split.
- N_estimators= 500 is number of decision trees that should be there in the random forest.



```
{'max_depth': 4,
 'max_features': 7,
 'min_samples_leaf': 6,
 'min_samples_split': 60,
 'n_estimators': 500}
```

Figure 54: Best Parameters

The feature importance for a Random Forest model is in the figure below.

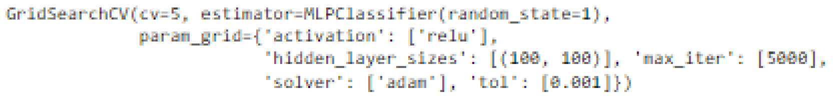


	IMP
Age	0.023940
Agency_Code	0.476179
Type	0.010413
Commision	0.071874
Channel	0.000981
Duration	0.045407
Sales	0.189102
Product Name	0.179530
Destination	0.002576

Figure 55: Feature Importance

Artificial Neural Network

GridSearchCV is done for the model Neural Networks as well to identify the best parameters for the performance of the model.

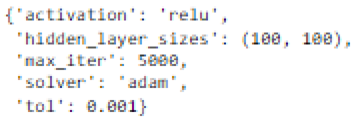


```
GridSearchCV(cv=5, estimator=MLPClassifier(random_state=1),
             param_grid={'activation': ['relu'],
                         'hidden_layer_sizes': [(100, 100)], 'max_iter': [5000],
                         'solver': ['adam'], 'tol': [0.001]})
```

Figure 56: Grid Search CV

After tuning the model by trying with several parameters, the parameters that were best for Neural Network were

- Activation = relu which is activation function used.
- Hidden_layer = (100,100) which tells the model to put 2 hidden layers with 100 nodes each.
- Max_iter = 5000 which is maximum amount of iterations to reach the tolerance level.
- Solver = adam is solver used to assign the weights to each nodes
- tol = 0.001 is tolerance level until which the weights have to get reassigned.



```
{'activation': 'relu',
 'hidden_layer_sizes': (100, 100),
 'max_iter': 5000,
 'solver': 'adam',
 'tol': 0.001}
```

Figure 57: Best Parameters

QUESTION 2.3

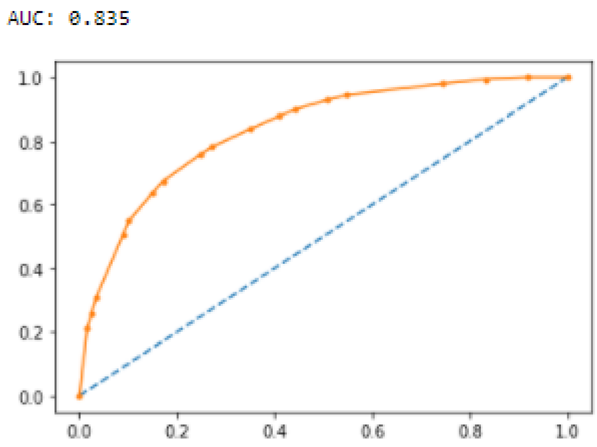
Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, and Plot ROC curve and get ROC_AUC score, classification reports for each model.

ANSWER 2.3

Model Performance of Decision Tree Classifier

Train data

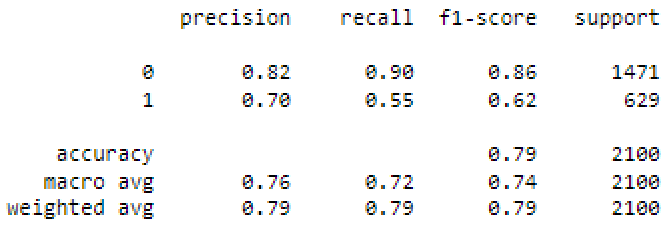
The Area under the curve is 0.835 for training dataset. Higher the AOC value better is the model so let’s understand all other performance metrics.



```
AUC: 0.835
```

Figure 58: ROC curve

The accuracy is 79% but the recall, precision and f1 score for 1 is average. The parameters for 1 are more important because it tells us about the customers that have claimed the insurance.



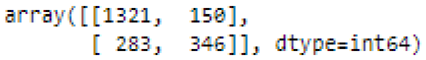
```
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1471
           1       0.70      0.55      0.62       629

    accuracy                           0.79      2100
   macro avg       0.76      0.72      0.74      2100
weighted avg       0.79      0.79      0.79      2100
```

Figure 59: Classification Report

The confusion matrix tells how many records have been predicted correctly by the model. 346 records are the ones predicted correctly for customers who have claimed. 283 records are the customers who had claimed but the model has predicted it wrong which is not good.

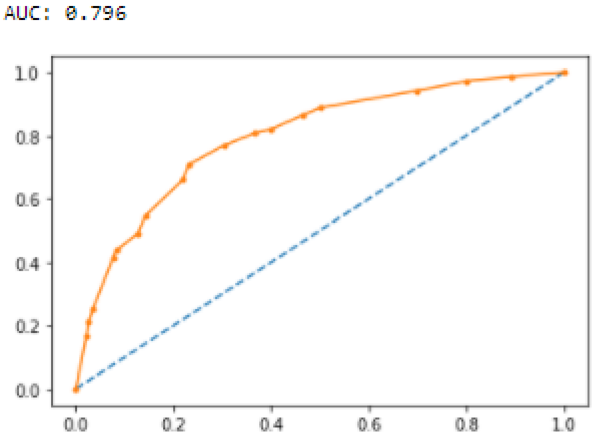


```
array([[1321,  150],
       [ 283,  346]], dtype=int64)
```

Figure 60: Confusion Matrix

Test data

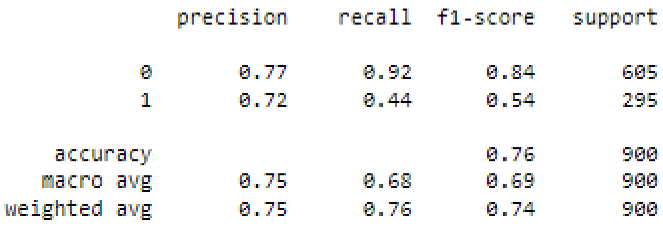
The area under the curve for testing dataset is 0.796 which in line with the train dataset



```
AUC: 0.796
```

Figure 61: ROC Curve

The precision, recall and f1 score for 1 are not in line with the train dataset as the difference is more than 10% between train and test data for recall. The accuracy for train model is 76% with the recall being very less.



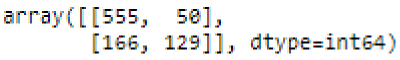
```
              precision    recall  f1-score   support

           0       0.77      0.92      0.84       605
           1       0.72      0.44      0.54       295

    accuracy                           0.76       900
   macro avg       0.75      0.68      0.69       900
weighted avg       0.75      0.76      0.74       900
```

Figure 62: Classification Report

The confusion matrix tells how many records have been predicted correctly by the model. 129 records are the ones predicted correctly for customers who have claimed. 166 records are the customers who had claimed but the model has predicted it wrong which is not good.



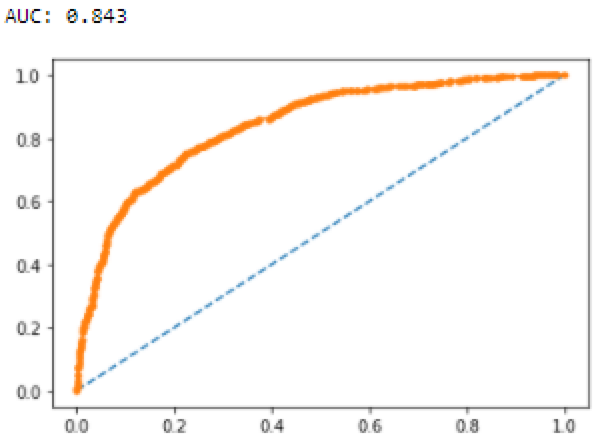
```
array([[555,  50],
       [166, 129]], dtype=int64)
```

Figure 63: Confusion Matrix

Model Performance of Random Tree Classifier

Train

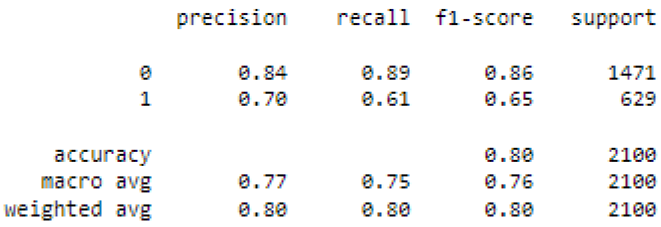
The Area under the curve is 0.843 for training dataset.



```
AUC: 0.843
```

Figure 64: ROC Curve

The accuracy is 80% and the recall, precision and f1 score for 1 is good.



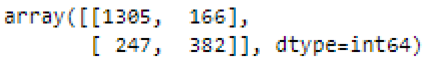
```
              precision    recall  f1-score   support

           0       0.84      0.89      0.86      1471
           1       0.70      0.61      0.65       629

    accuracy                           0.80      2100
   macro avg       0.77      0.75      0.76      2100
weighted avg       0.80      0.80      0.80      2100
```

Figure 65: Classification Report

The confusion matrix tells how many records have been predicted correctly by the model. 382 records are the ones predicted correctly for customers who have claimed. 247 records are the customers who had claimed but the model has predicted it wrong.



```
array([[1305,  166],
       [ 247,  382]], dtype=int64)
```

Figure 66: Confusion Matrix

TEST

The area under the curve for testing dataset is 0.816 .

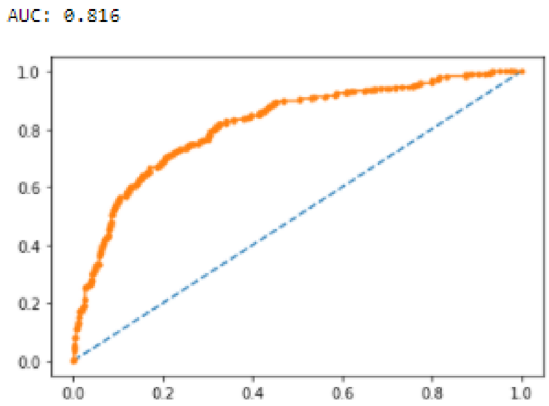
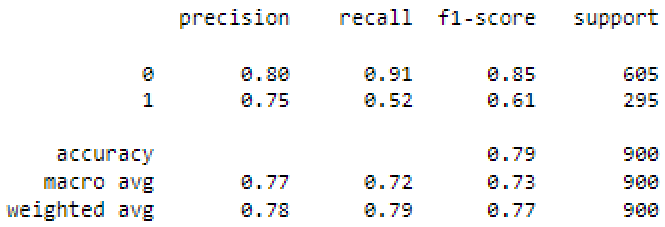


Figure 67: ROC Curve

The accuracy is 79% and the recall, precision and f1 score for 1 is good. Precision has improved for the test data whereas recall has gone done from the test data but the overall scores are in line with train data.



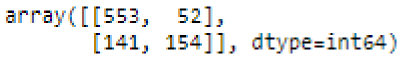
```
              precision    recall  f1-score   support

           0       0.80      0.91      0.85       605
           1       0.75      0.52      0.61       295

    accuracy                           0.79       900
   macro avg       0.77      0.72      0.73       900
weighted avg       0.78      0.79      0.77       900
```

Figure 68: Classification Report

The confusion matrix tells how many records have been predicted correctly by the model. 154 records are the ones predicted correctly for customers who have claimed. 141 records are the customers who had claimed but the model has predicted it wrong. The number has improved from the decision tree model.



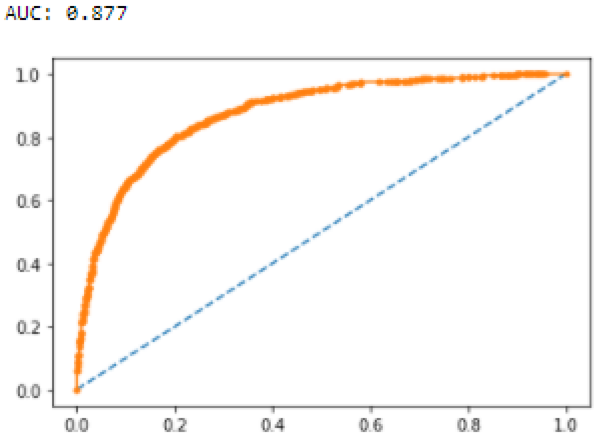
```
array([[553,  52],
       [141, 154]], dtype=int64)
```

Figure 69: Confusion Matrix

Model Performance of Artificial Neural Networks

Train

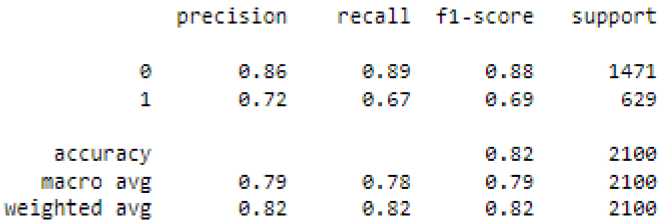
The Area under the curve is 0.877 for training dataset.



```
AUC: 0.877
```

Figure 70: ROC Curve

The accuracy is 82% and the recall, precision and f1 score for 1 is good. These have been the best values from all the three models.



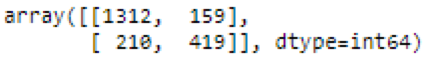
```
              precision    recall  f1-score   support

           0       0.86      0.89      0.88      1471
           1       0.72      0.67      0.69       629

    accuracy                           0.82      2100
   macro avg       0.79      0.78      0.79      2100
weighted avg       0.82      0.82      0.82      2100
```

Figure 71: Classification Report

The confusion matrix tells how many records have been predicted correctly by the model. 419 records are the ones predicted correctly for customers who have claimed. 210 records are the customers who had claimed but the model has predicted it wrong.

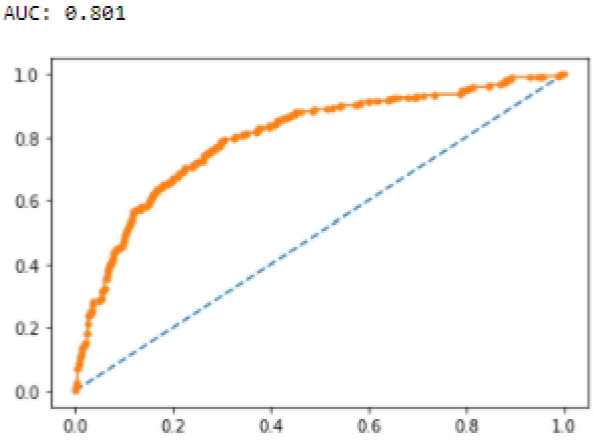


```
array([[1312,  159],
       [ 210,  419]], dtype=int64)
```

Figure 72: Confusion Matrix

TEST

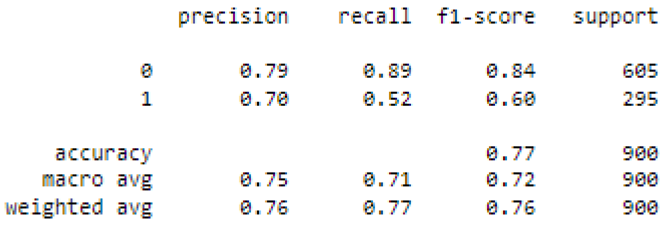
The area under the curve for testing dataset is 0.801 which is not in line with the train dataset. Let's have a look at other parameters if they are in line with train data.



```
AUC: 0.801
```

Figure 73: ROC Curve

The accuracy is 77% but the recall, precision and f1 score for 1 is average. There is a large difference in recall and f1 score from the test data which suggests than the model has been overfit. Thereby this model has learnt too much from training data and is not able to perform in the testing stage thereby this is not a good model.



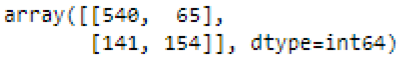
```
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       605
           1       0.70      0.52      0.60       295

    accuracy                           0.77       900
   macro avg       0.75      0.71      0.72       900
weighted avg       0.76      0.77      0.76       900
```

Figure 74: Classification Report

The confusion matrix tells how many records have been predicted correctly by the model. 154 records are the ones predicted correctly for customers who have claimed. 141 records are the customers who had claimed but the model has predicted it wrong.



```
array([[540,  65],
       [141, 154]], dtype=int64)
```

Figure 75: Confusion Matrix

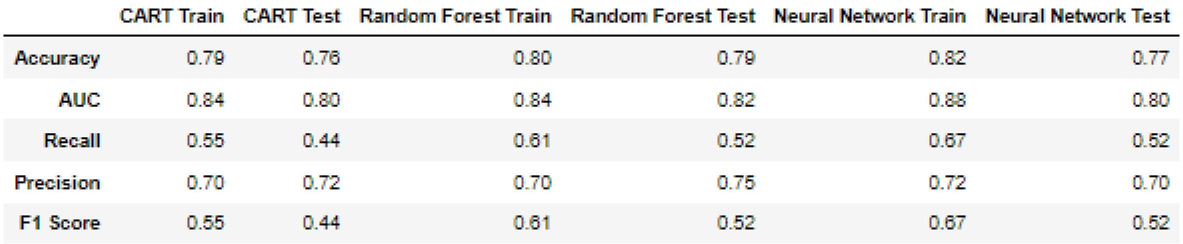
QUESTION 2.4

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

ANSWER 2.4

As the company is facing higher claims, the area of focus is to get a higher recall value as it focuses more on the customers who have claimed. The model should be able to predict more true positives which are more number of customers who have claimed the insurance. The false negatives have to be less in this case to get a higher recall. The dataset with recall will be the most suitable one.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$



	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.79	0.76	0.80	0.79	0.82	0.77
AUC	0.84	0.80	0.84	0.82	0.88	0.80
Recall	0.55	0.44	0.61	0.52	0.67	0.52
Precision	0.70	0.72	0.70	0.75	0.72	0.70
F1 Score	0.55	0.44	0.61	0.52	0.67	0.52

Figure 76: Model Comparison

Based on performance matrix of all models the model with best recall is Radom Forest.

The Decision Tree model has very low recall value which is not satisfying the requirement.

The recall for training dataset in ANN model is high but the model performance does not improve in test data. There seems to be over fitting of data. ANN has the same recall as that of Random Forest but the difference in train and dataset is significant which implies model is not good and has learnt too much from train data and is not able to perform the same in test data.

Random Forest performs better than both the other models in terms of accuracy, precision, recall and f1 score and test data is almost in line with the train data.

QUESTION 2.5

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

ANSWER 2.5

Business Insights on all of the above analysis.

- JZI agency has very low sales compared to other agencies so a marketing campaign can be held to improve customer reach and the plans can be improved.
- Increase customer satisfaction in each of the plans to help reduce claims.
- Take action against fraudulent claims as they are putting company at a loss.
- Silver and gold plan is facing higher claims so company should consider either cancelling these plans or a change to these plans.
- C2B agency is contributing the most to number of claims and EPX agency is performing very well with high sales and less number of claims. So C2B agency can adopt the plans similar to EPX agency.
- Airlines type insurance firms are facing a higher claim than travel agency firms so the business models of both should be compared to find what parameters are different that can help airline firms also to reduce their claims.
- ASIA has very high sales and also very high claims so this is the target audience where company should look for. As EPX agency is performing very well it can bring in more plans for ASIA destination which will help to increase the sales further and bring down the number of claims.