DM PROJECT

DATA MINING

RHEA.S.M

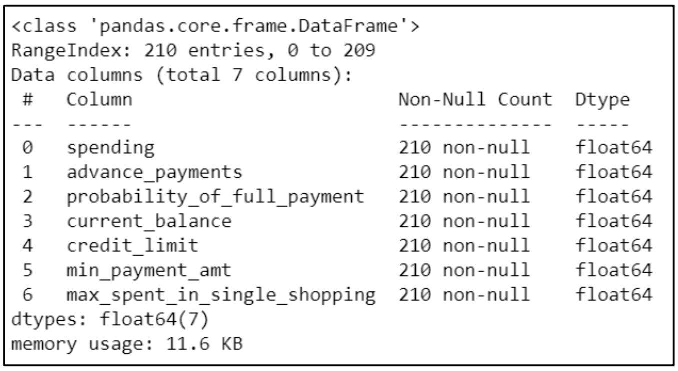
PGPDSBA Online Sep_B 2021

CONTENT

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.

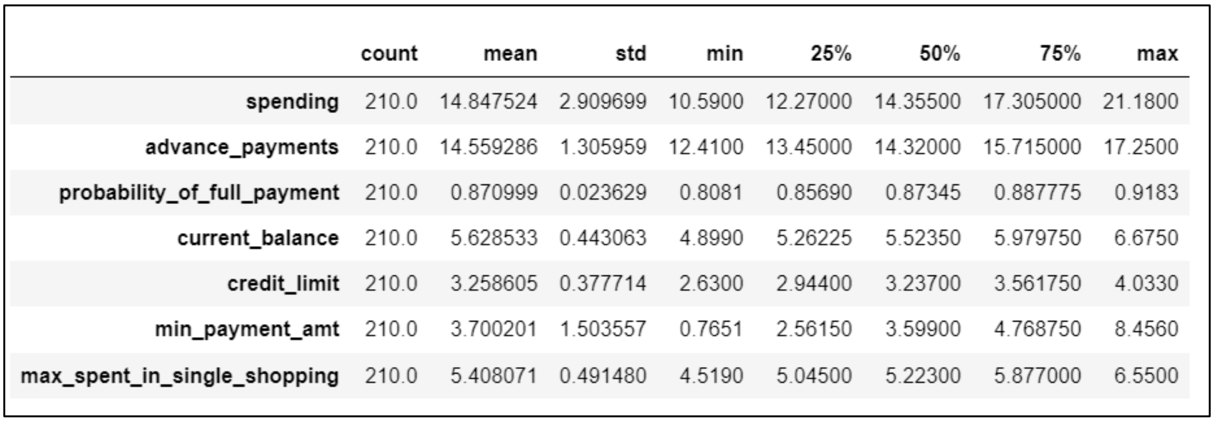
1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bivariate, and multivariate analysis).



```
<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
```

Fig 1.1 Information of data collected by the bank

From the above Fig 1.1 it is evident that the data consists of 210 entries with 7 columns in total. All these columns are numerical in nature. The different variables in this dataset are spending representing the amount spent by the customers, advance_payments representing the amount paid by the customer in advance by cash, probability_of_full_payment representing probability of full payment done by the customer to the bank, current_balance representing the balance amount left in the bank to make purchases, credit_limit representing the limit of amount in the credit card, min_payment_amt representing the minimum payment made by customers on monthly purchases and max_spent_in_single_shopping representing the maximum amount spent in one go. There are no non-null values present in 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 1.1 Describe function performed on data collected by the bank

The above Table 1.1 has been derived using descriptive statistics to summarise the data. The function used is "df.describe(include='all')". We can see the mean, count, median, standard deviation, quartile values, etc with this function. When a check for duplicates were done, there were none. The dataset looks good to go.

DATA VISUALISATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

A. UNIVARIATE ANALYSIS:

Univariate Analysis is the most basic form of statistical data analysis. It helps us individually asses the variables available to us in the dataset. With regards to data given, a histogram and boxplot has been individually created for each of the variables.

- SPENDING:

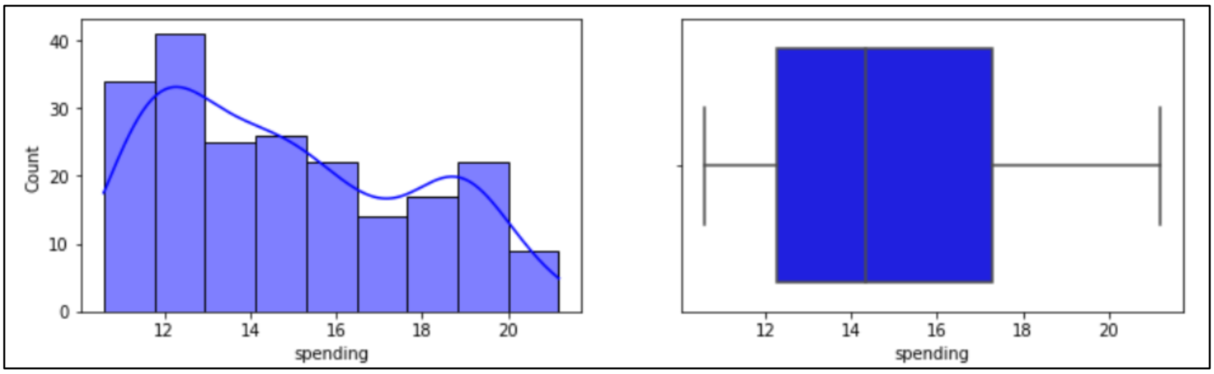


Fig 1.2 Spending pattern of customers

Fig 1.2 clearly tells us about the spending patterns of the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed.

- ADVANCE PAYMENTS:

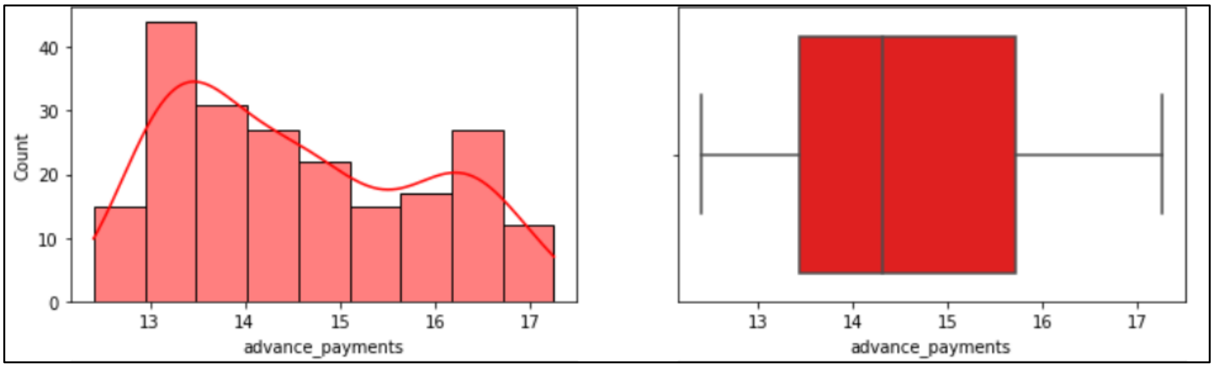


Fig 1.3 Advance Payments made by customers

Fig 1.3 tells us about the advance payments made by the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed.

- PROBABLITY OF FULL PAYMENT:

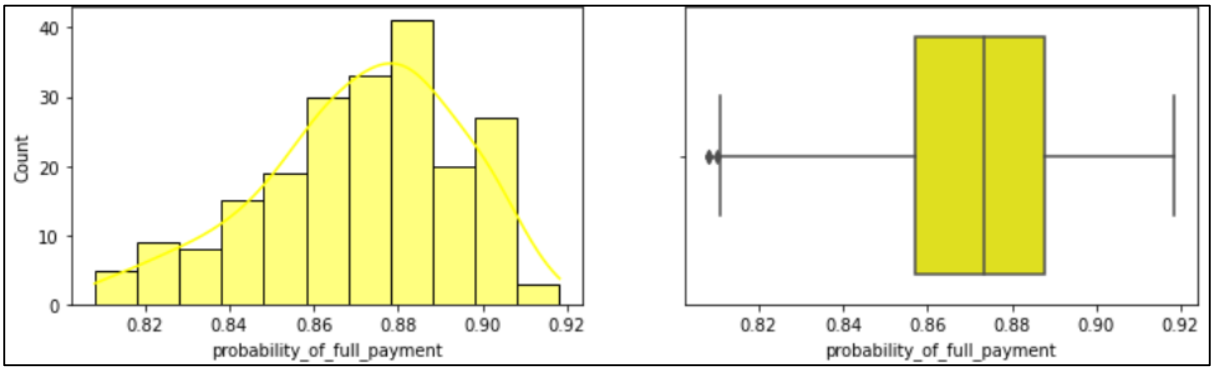


Fig 1.4 Probability of full payment by customers

Fig 1.4 clearly tells us about the probability of full payment made by the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. There are outliers present in this variable.

- CURRENT BALANCE:

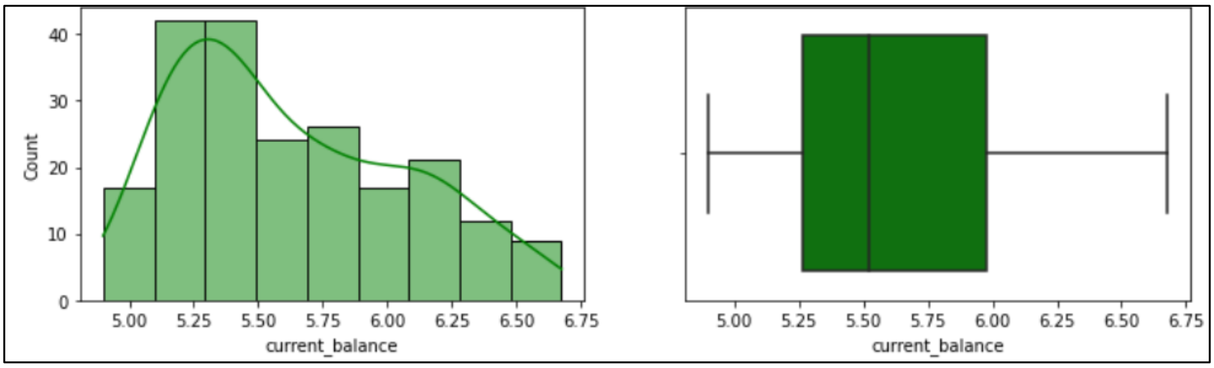


Fig 1.5 Current Balance of customers

Fig 1.5 tells us about the current balance of the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed.

- CREDIT LIMITS:

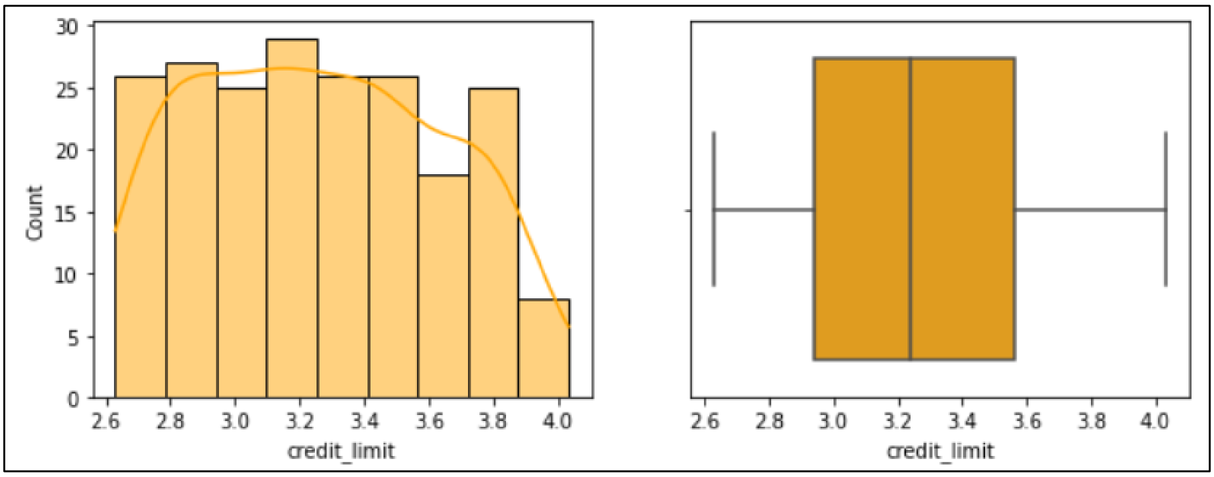


Fig 1.6 Credit limits of customers

Fig 1.6 tells us about the credit limits of the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median.

- MINIMUM PAYMENT AMOUNT:

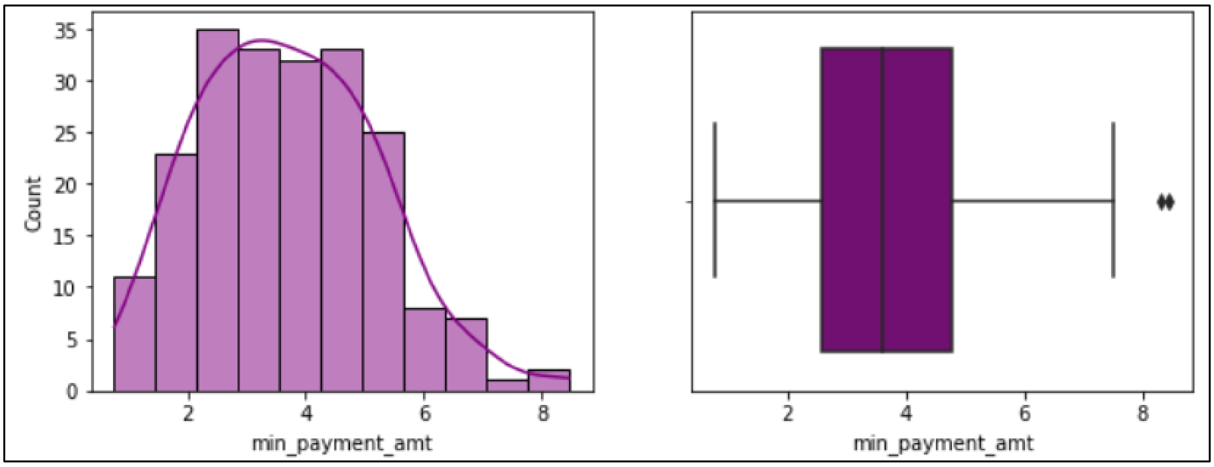


Fig 1.7 Minimum Payment amount by customers

Fig 1.7 tells us about the minimum payment amounts made by the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed. There are outliers present in this variable.

- MAXIMUM SPENT IN A SINGLE SHOPPING:

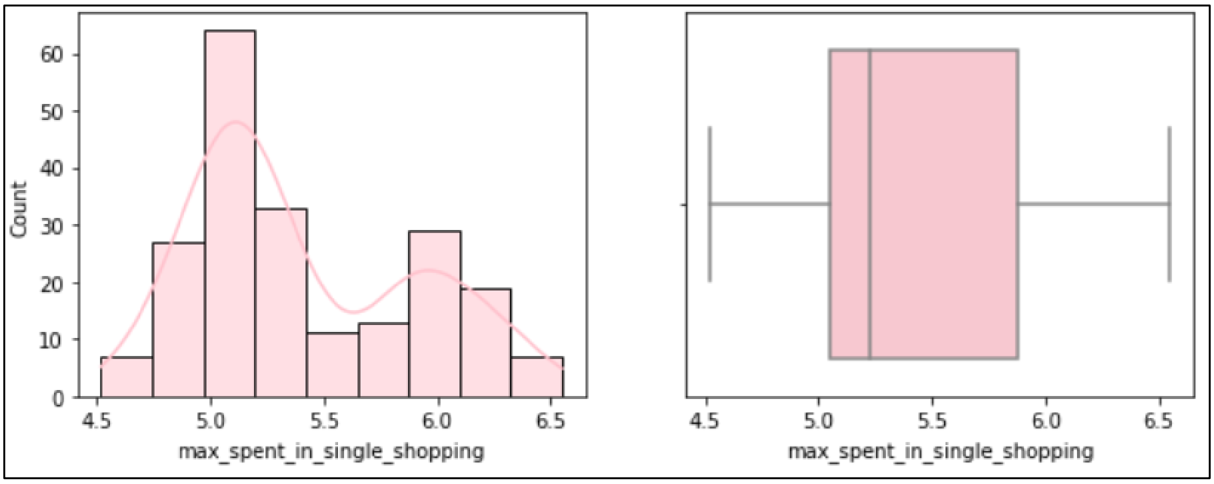


Fig 1.8 Maximum spent in single shopping amount by customers

Fig 1.8 clearly tells us about the maximum amount spent in single shopping experience by the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed.

B. MULTIVARIATE ANALYSIS

Multivariate Analysis is a Statistical procedure for analysis of data involving more than one type of variable(s). A Pair plot and Heat map has been created to fulfil the Multivariate Analysis of the data provided.

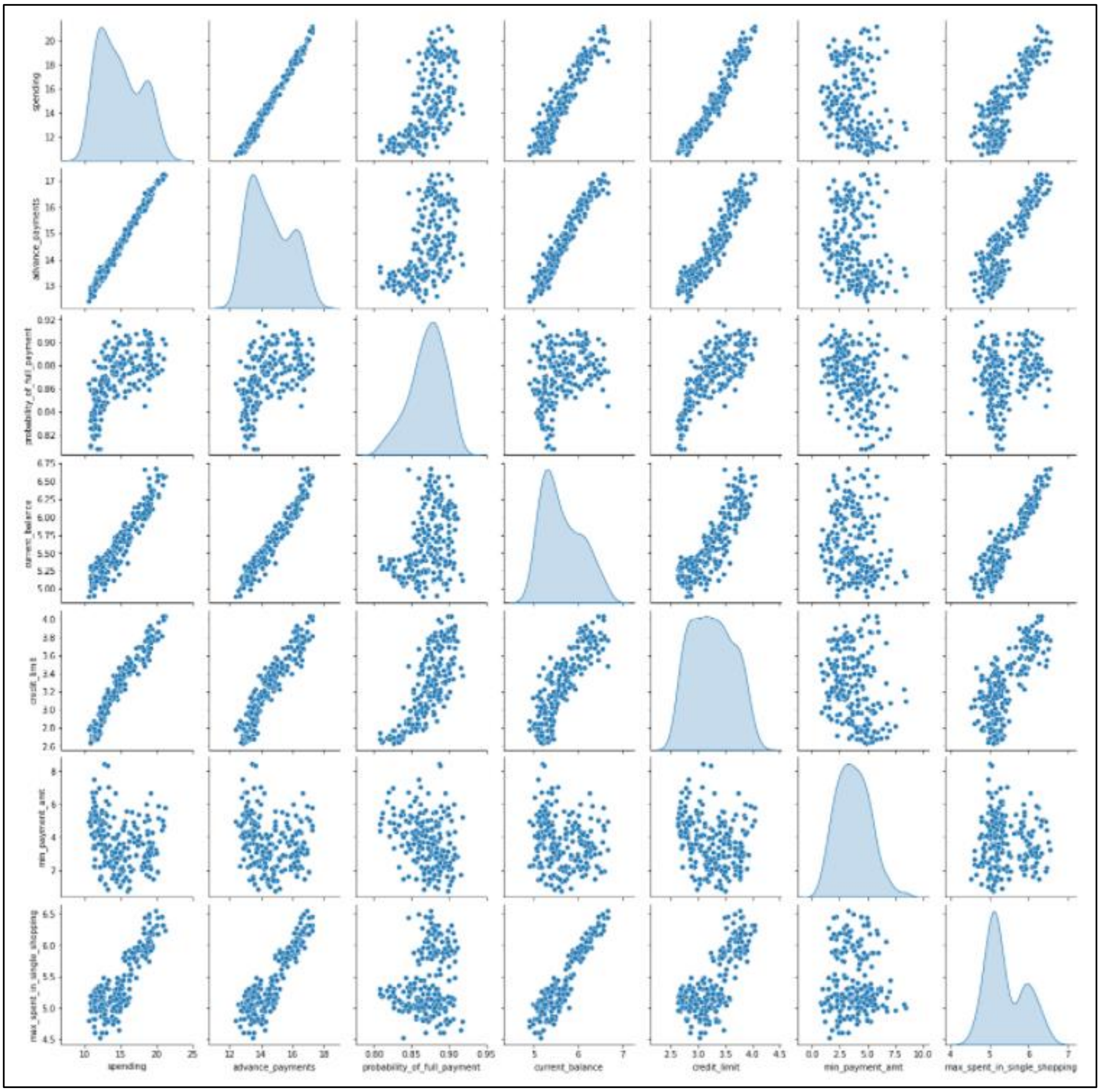


Fig 1.9 Multivariate analysis using a Pair plot

The above Fig 1.9 of a Pair plot essentially helps us understand the relationship between all the numerical values in the dataset. It helps us compare the variables to one another in a way that we are made to deeply understand the trends and patterns of the dataset provided.

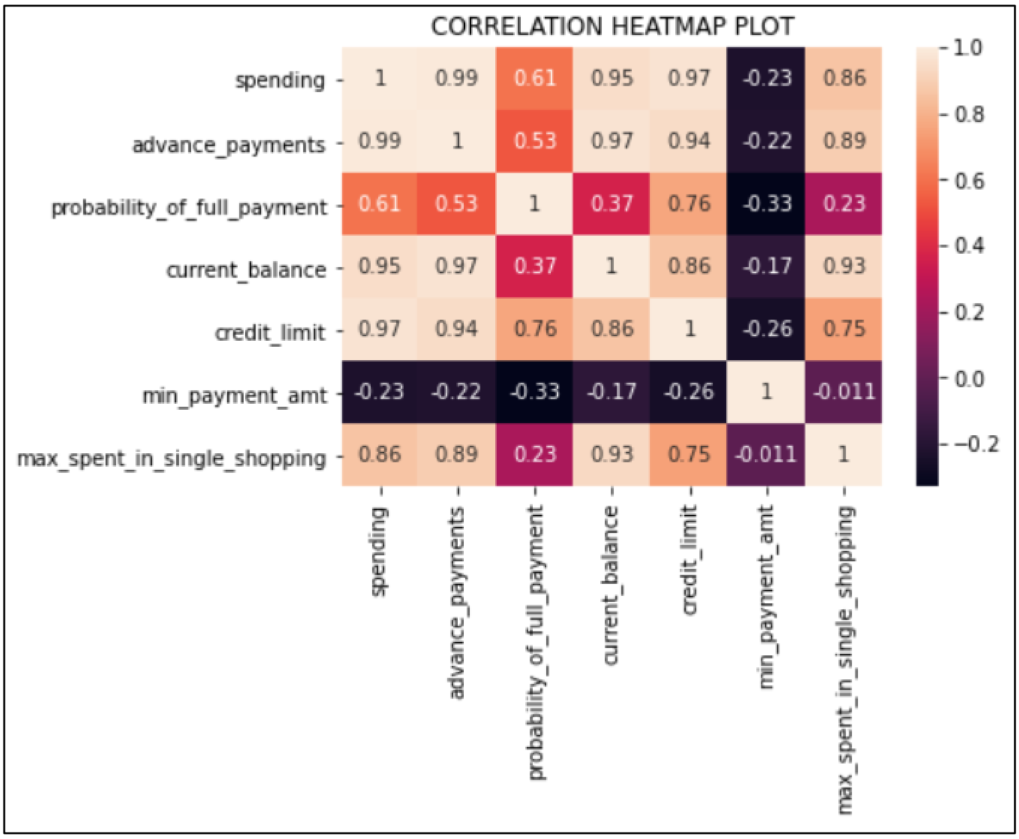


Fig 1.10 Multivariate analysis using a Correlation Heat map

The above Heat map (Fig 1.10) essentially shows us the correlation between pairs of different variables. It clearly shows us that the spending variable is highly positively correlated with the advance_payments, current_balance, credit_limit and max_spent_in_single_shopping. Thus telling us about the purchasing power of a customer. The more they spends, the more their credit limit and the more capable they are to make advance payments.

It also tells us that advance payments is highly positively correlated with current_balance, credit_limit and max_spent_in_single_shopping. While min_payment_amt is negatively correlated with every variable.

C. TREATMENT OF OUTLIERS

One of the biggest downfall for any model performance is the outliers present in the data. Outliers ideally are the extreme values for the specific column which affects the generalization of the data and model.

We choose to treat the outlier values by checking with their respective lower ranges and upper ranges. We did the same by creating a user defined function as can be seen in the attached Jupyter notebook. There are outliers are present only in two variables of the seven.

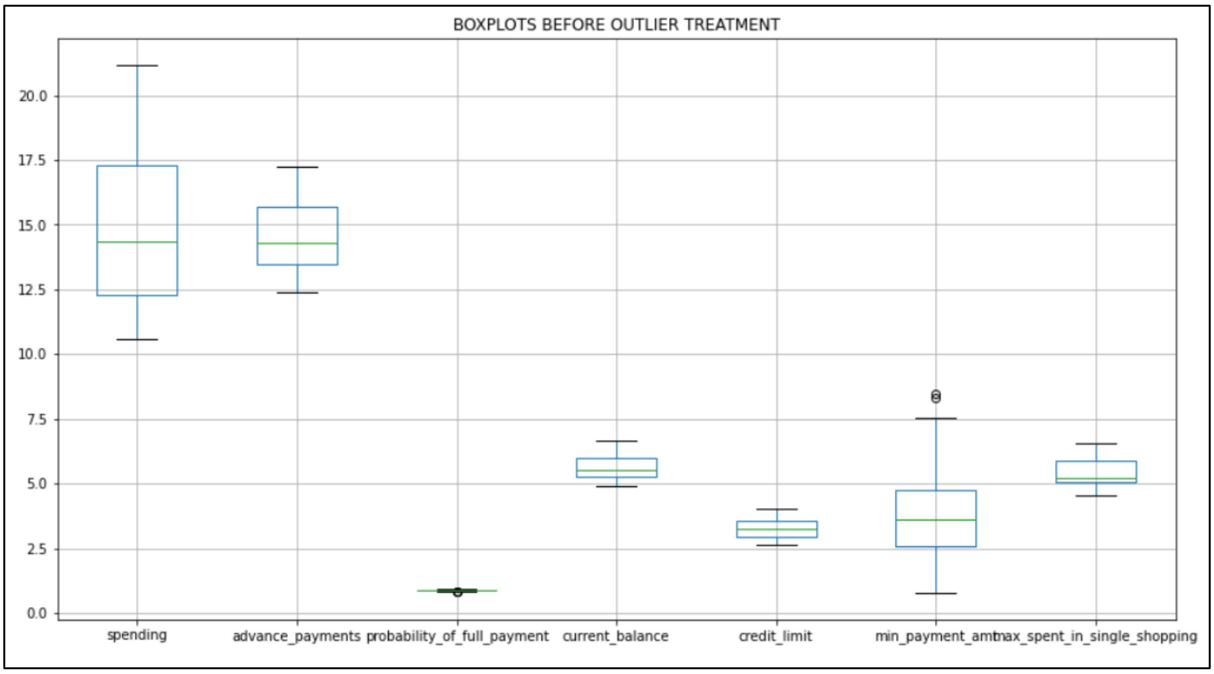


Fig 1.11 Boxplots before Outlier treatment

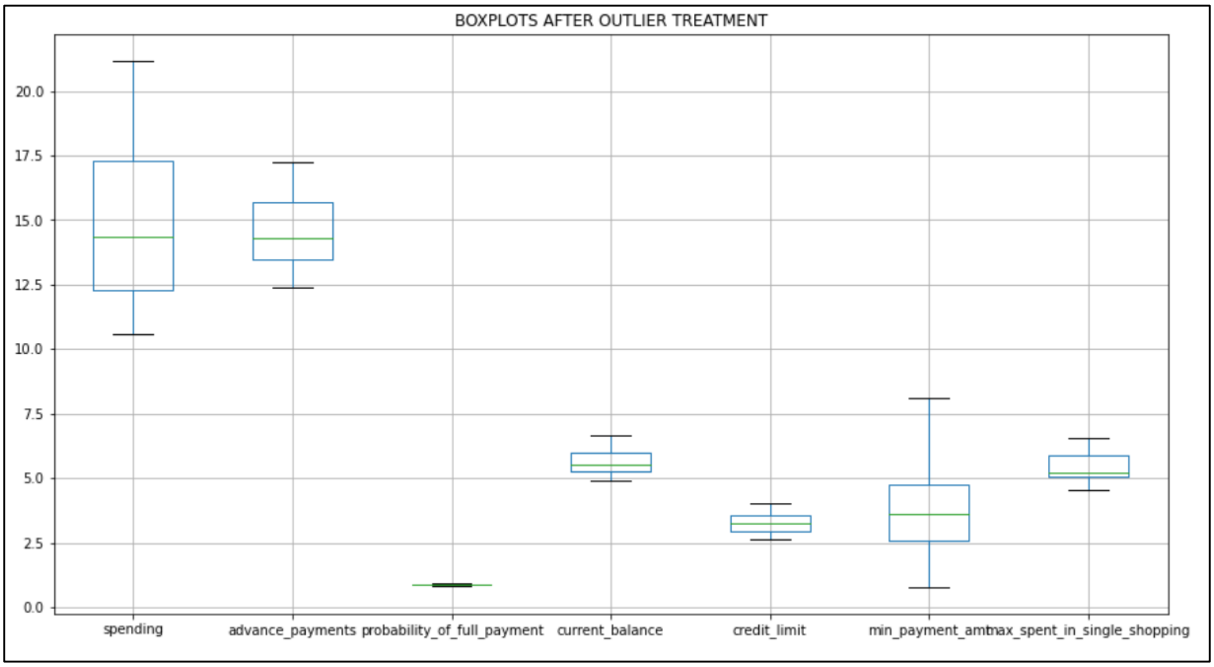


Fig 1.12 Boxplots after Outlier treatment

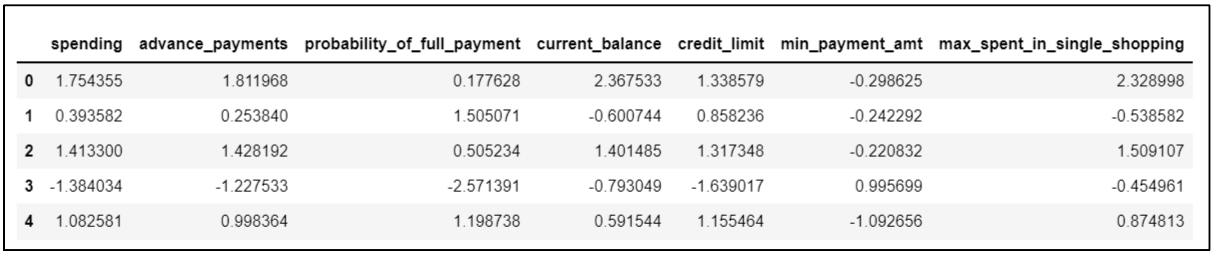
The above figures are a visualisation of both boxplots before (Fig 1.11) and after (Fig 1.12) Outlier treatment to show us the difference. Now that the outliers have been treated, we are good to move on to the scaling part.

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

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set spending is having values in thousands while advance_payments have values in hundreds and credit_limit has value in ten thousands. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

From the many ways there are to scale the data, I have used Z-score. A snippet of the scaled dataset can be seen below in Table 1.2.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.177628	2.367533	1.338579	-0.298625	2.328998
1	0.393582	0.253840	1.505071	-0.600744	0.858236	-0.242292	-0.538582
2	1.413300	1.428192	0.505234	1.401485	1.317348	-0.220832	1.509107
3	-1.384034	-1.227533	-2.571391	-0.793049	-1.639017	0.995699	-0.454961
4	1.082581	0.998364	1.198738	0.591544	1.155464	-1.092656	0.874813

Table 1.2: A snippet of the scaled dataset using Z-score

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

After scaling the data, using Z-score we were instructed to apply hierarchical clustering to the scaled data and did the same.

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. They are of two types:

- •Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- •Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

The linkage method used her is ‘average’, the truncate mode used is ‘lastp’ with a p of 25 to help us understand the Dendrogram better. The Average linkage method is defined as the distance between two clusters is defined as the average distance between each pair of points, one from each cluster.

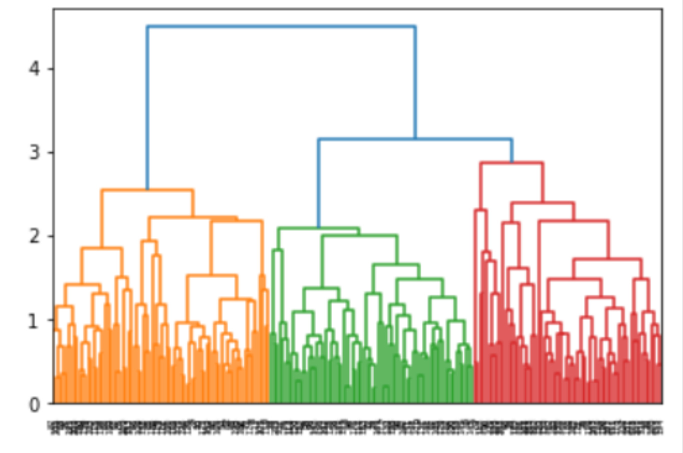


Fig 1.13 Dendrogram using Average Linkage method

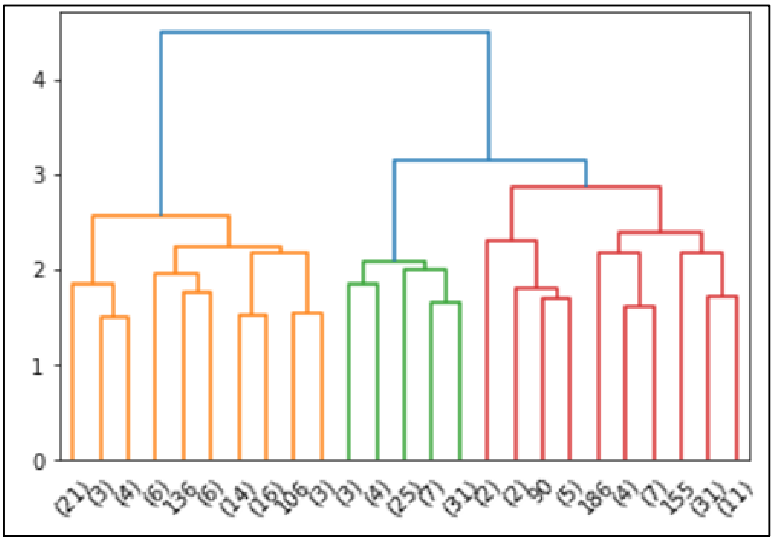
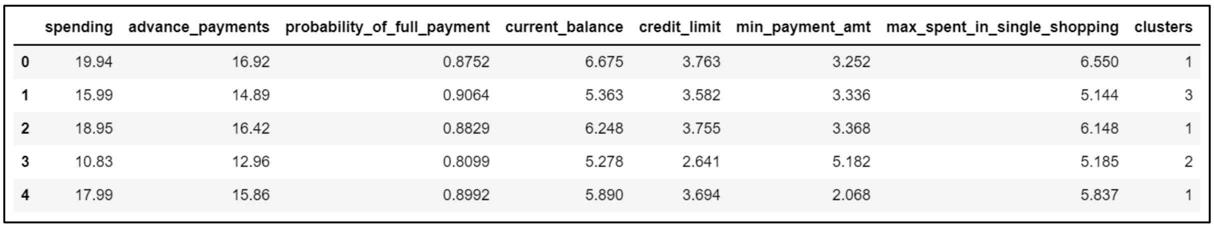


Fig 1.14 Dendrogram truncated

The above Fig 1.13 is a Dendrogram formed using the Average Linkage Method, while on the other hand Fig 1.14 is a truncated version of the same Dendrogram. We can see that Fig 1.13 doesn't look very clear and doesn't show us any numbers within any cluster, hence it is important to truncate the Dendrogram as it gives us an idea of the numbers within clusters.

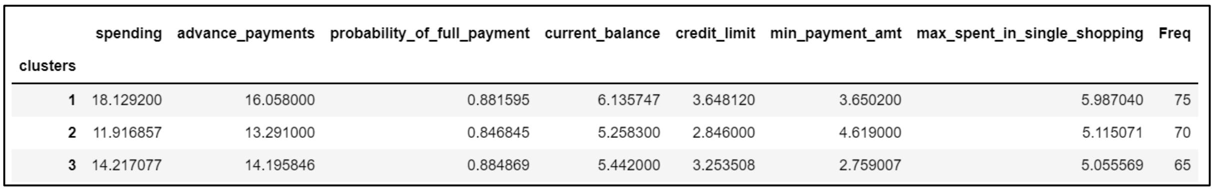
For cluster grouping based on the Dendrogram, 3 seems to look most optimal. Hence fcluster was performed with 3 clusters and criterion as 'maxclust'.



	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 1.3 Hierarchical clustering- Head of dataset with clusters column added

The above Table 1.3 is a snippet of the table with the head of the data divided into 3 clusters and the cluster column clearly showcased.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
clusters								
1	18.129200	16.058000	0.881595	6.135747	3.648120	3.650200	5.987040	75
2	11.916857	13.291000	0.846845	5.258300	2.846000	4.619000	5.115071	70
3	14.217077	14.195846	0.884869	5.442000	3.253508	2.759007	5.055569	65

Table 1.4 Hierarchical clustering- Dataset divided based on the 3 Clusters

The above Table 1.4 is a clear representation of the 3 clusters divided and the count by the column name of frequency of each cluster added at the end. Cluster 1 has a count of 75, Cluster 2 has a count of 70 and lastly Cluster 3 has a count of 65. Cluster 1 consists of the Highest spenders while Cluster 2 consists of the Lowest spenders and Cluster 3 consists of Medium level spenders.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

K-means clustering is the most used non-hierarchical clustering technique. It aims to partition n observations into k clusters in which each observation belongs to the cluster whose mean (centroid) is nearest to it, serving as a prototype of the cluster. It minimizes within-cluster variances (squared Euclidean distances).

There are many methods that are recommended for determination of an optimal number of partitions/clusters. The 2 main methods used here are Elbow method and Silhouette Method.

- The Elbow Method: For a given number of clusters, the total within-cluster sum of squares (WCSS) is computed. The Elbow method looks at the total WCSS as a function of the number of clusters.

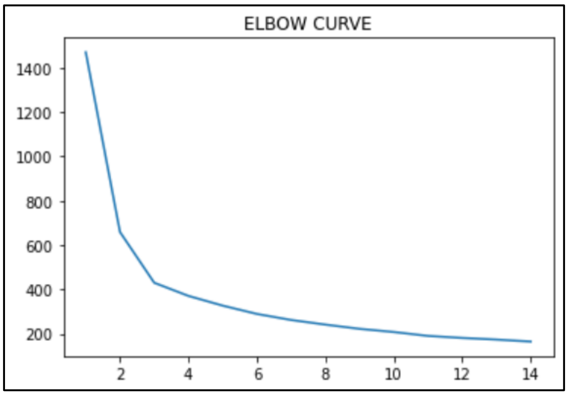


Fig 1.15 Elbow Curve

In the above Elbow curve (Fig 1.15) it is clear that 3 is the optimal number of clusters that needs to be made with regards to the dataset at hand. There is prominent dip shown there at 3.

- The Silhouette score Method: This method measures how tightly the observations are clustered and the average distance between clusters. The maximum value of the statistic indicates the optimum value of k.

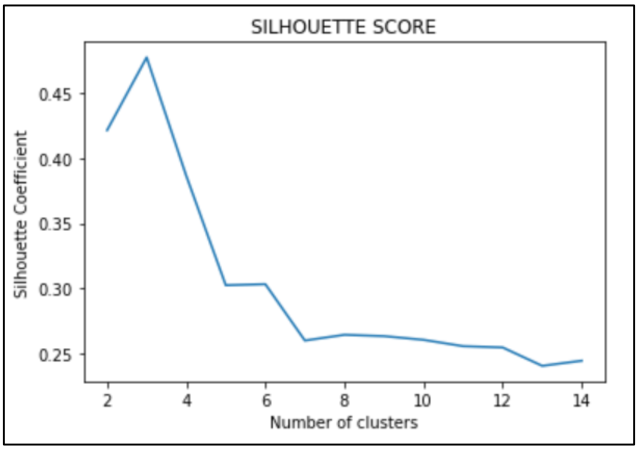
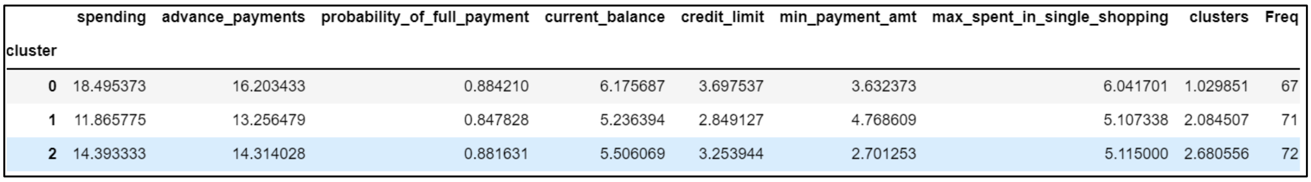


Fig 1.16 Silhouette Score

In the above Silhouette Score curve (Fig 1.16) it is clear that 3 is the optimal number of clusters that needs to be made with regards to the dataset at hand. It is determined through the highest point on the graph.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	Freq
cluster									
0	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701	1.029851	67
1	11.865775	13.256479	0.847828	5.236394	2.849127	4.768609	5.107338	2.084507	71
2	14.393333	14.314028	0.881631	5.506069	3.253944	2.701253	5.115000	2.680556	72

Table 1.5 K-Means clustering- Dataset divided based on the 3 Clusters

The above Table 1.5 is a clear representation of the 3 clusters divided and the count by the column name of frequency of each cluster added at the end. Cluster 0 has a count of 67, Cluster 1 has a count of 71 and lastly Cluster 3 has a count of 72. Cluster 0 consists of the Highest spenders while Cluster 1 consists of the Lowest spenders and Cluster 2 consists of Medium level spenders.

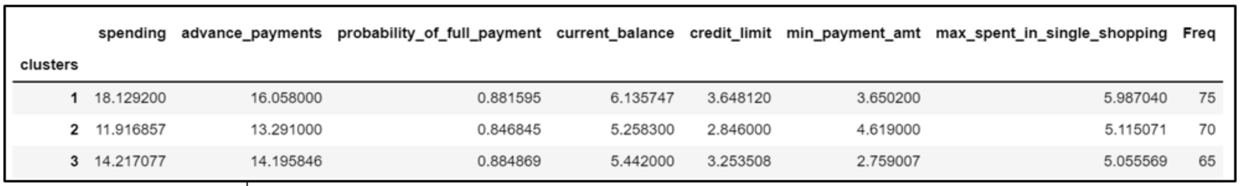
1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

Cluster Group Profiles:

Group 1: High Spending - Hierarchical (Cluster 1), K-means (Cluster 0). They determine the highest spenders.

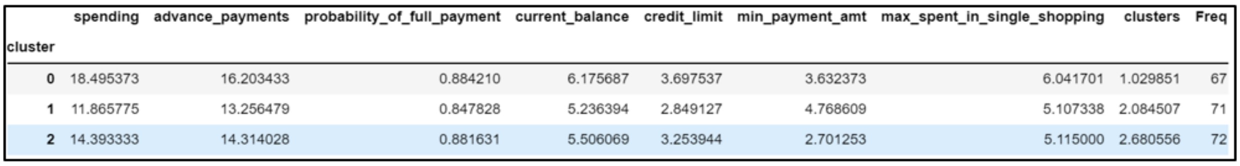
Group 2: Low Spending- Hierarchical (Cluster 3), K-means (Cluster 2). They determine the lowest spenders.

Group 3: Medium Spending- Hierarchical (Cluster 3), K-means (Cluster 2). They determine the medium spenders.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
clusters								
1	18.129200	16.058000	0.881595	6.135747	3.648120	3.650200	5.987040	75
2	11.916857	13.291000	0.846845	5.258300	2.846000	4.619000	5.115071	70
3	14.217077	14.195846	0.884869	5.442000	3.253508	2.759007	5.055569	65

Hierarchical clustering- Dataset divided based on the 3 Clusters



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	Freq
cluster									
0	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701	1.029851	67
1	11.865775	13.256479	0.847828	5.236394	2.849127	4.768609	5.107338	2.084507	71
2	14.393333	14.314028	0.881631	5.506069	3.253944	2.701253	5.115000	2.680556	72

K-Means clustering- Dataset divided based on the 3 Clusters

Fig 1.17 Tables showing difference between the two types of clustering

The Fig 1.17 represents a snippet from above of the difference in Hierarchical and K-Means clustering. While the clusters in Hierarchical are 1, 2 and 3; the clusters in K-Means are 0,1 and 2. The distribution among clusters can clearly be seen in the Frequency column at the end. While almost similarly distributed there is a very little difference in the clustering processes followed hence the difference in frequencies or count in each cluster.

Promotional Strategies for each group:

Group 1: High Spending:

Their purchases are already highest in all the 3 clusters to make it higher, they don't spend on only needs and necessities, they may be spend thrifts and have a real good purchasing power:

- Better reward points may help.
- If we increase their credit limits they would be lured to spending more.
- Offer loans to them since they seem to be customers with good payment records too.
- Tie ups with Luxury brands and offering exclusive discounts or cashbacks.

Group 2: Low Spending:

Their purchases are the lowest in all the 3 clusters to make it higher, since they spend on only needs and necessities:

- These customers should be continuously reminded of making payments.
- Incentivise them when they make payments on time or earlier.
- Tie ups with renowned daily essential grocery store may drive them towards spending a little more on seeing attractive discounts or offers.
- Offers on dining and take-away can be made available for them, everyone loves food.

Group 3: Medium Spending:

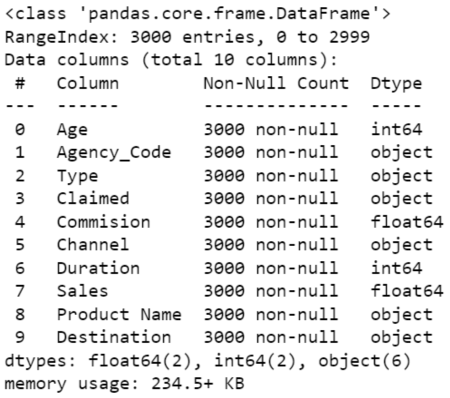
Their purchases are the medium and to make it higher, this group is usually a mix of both spe nding on necessities as well as on a few luxuries:

- Their credit scores are usually pretty good, so that can be incentivised by either increa sing their credit limits or by reducing the interest rates.
- With this group loyalty cards will work best, they must be promoted.
- Since they like a little luxury, lure them with additional benefits like lounge access.

Problem 2: CART-RF-ANN

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

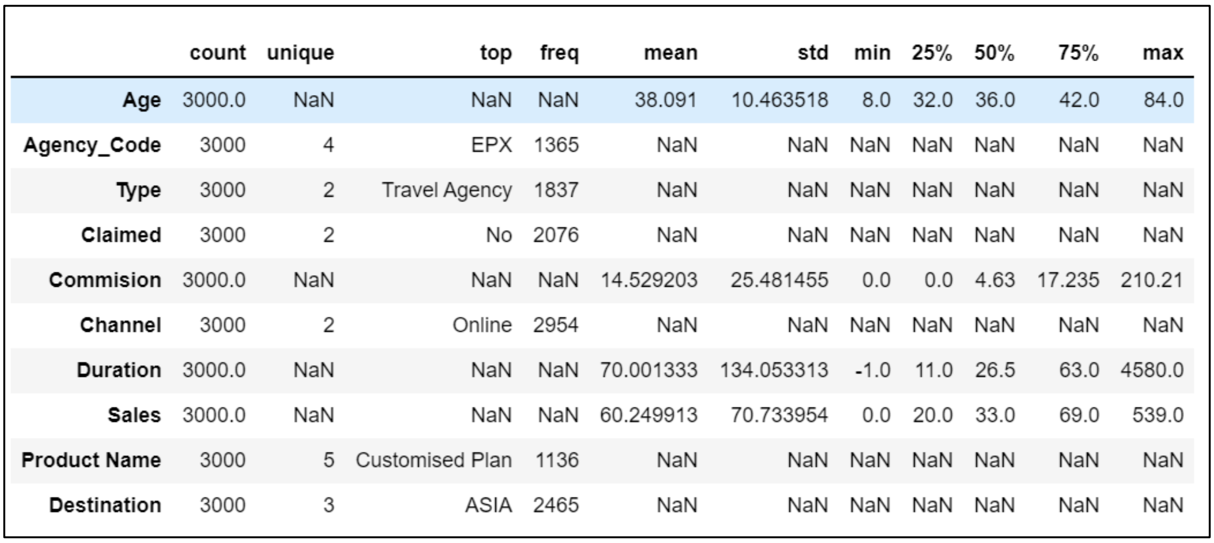
2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bivariate, and multivariate analysis).



```
<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
```

Fig 2.1 Information of data collected by the Insurance companies

There are 10 variables in total out of which Age, Commission, Duration and Sales are numeric in nature while the remaining 6 variables are Categorical in nature. These Categorical variables include Agency_code, Type, Claimed, Channel, Product Name and Destination. These will later be encoded as we proceed. There are no missing values in the data. There are 9 independent variables and 1 dependent variable, ie: Claimed. Since we are asked to investigate on the higher claim frequency of tour insurance. The dependent variable here is also called the target variable.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000.0	NaN	NaN	NaN	38.091	10.463518	8.0	32.0	36.0	42.0	84.0
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000.0	NaN	NaN	NaN	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000.0	NaN	NaN	NaN	70.001333	134.053313	-1.0	11.0	26.5	63.0	4580.0
Sales	3000.0	NaN	NaN	NaN	60.249913	70.733954	0.0	20.0	33.0	69.0	539.0
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2.1 Describe function performed on data collected by the agencies

The above Table 2.1 has been derived using descriptive statistics to summarize the data. The function used is “df.describe(include=’all’)”. We can see the mean, count, median, standard deviation, quartile values, etc with this function. When a check for duplicates were done, there were none. The dataset looks good to go.

The Duration variable shows the minimum value as -1 which is clearly not how time can be measured, hence this may possibly be a wrong entry. Not treating it, just mentioning the same. The rest of the data looks pretty good.

When checking for duplicates, while they are seen to be present in the data there does not seem to be any unique identifier to differentiate the entries. And since this is an Insurance business there is a possibility for the package designed for 2 or more people to be exactly the same. Hence the duplicate rows have not been removed from the dataset.

DATA VISUALISATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

A. UNIVARIATE ANALYSIS/ BIVARIATE ANALYSIS:

Univariate Analysis is the most basic form of statistical data analysis. It helps us individually asses the variables available to us in the dataset. With regards to data given, the numerical variables are represented by a histogram and boxplot. While the categorical variables are represented by a count plot and boxplot.

- AGE:

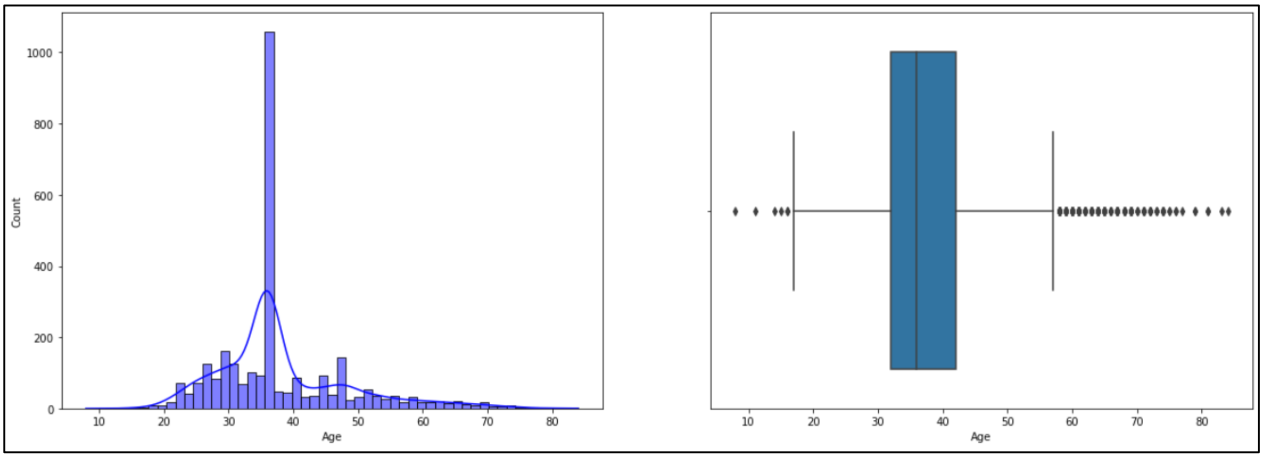


Fig 2.2 Age of customers

The above Fig 2.2 clearly tells us about the age of the different customers. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the

Mean > Median. It is also clear through the histogram that the data is right skewed. There are multiple Outliers present in this variable.

- SALES:

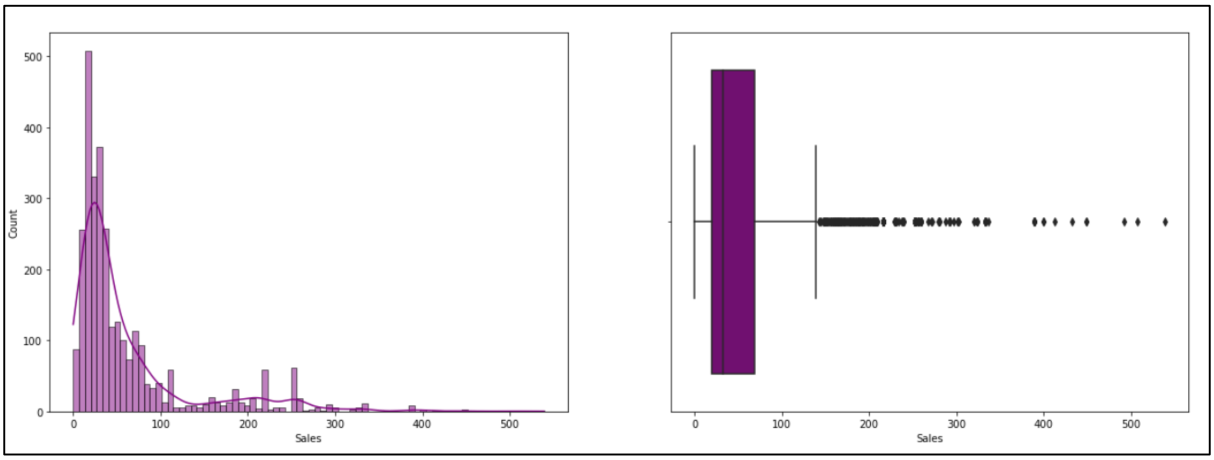


Fig 2.3 Sales

The above Fig 2.3 clearly talks about the sales point of view. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed. There are multiple Outliers present in this variable.

- COMMISION:

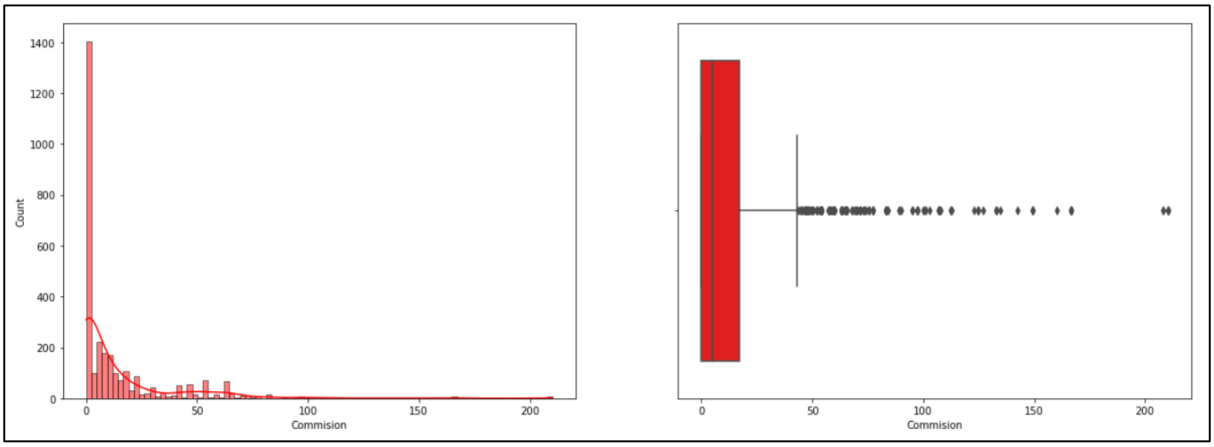


Fig 2.4 Commission

The above Fig 2.4 clearly talks about the commission point of view. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean

> Median. It is also clear through the histogram that the data is right skewed. There are multiple Outliers present in this variable.

- DURATION:

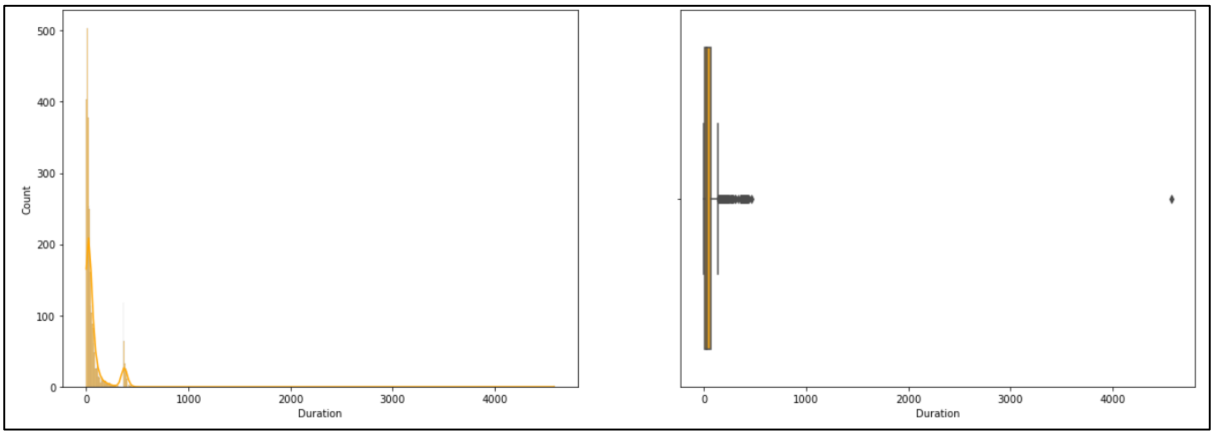


Fig 2.5 Duration

Fig 2.5 talks about the duration of policy. The box plot is seen to be positively skewed, we can clearly see this by the describe function performed where the Mean > Median. It is also clear through the histogram that the data is right skewed. There are multiple Outliers present in this variable.

- AGENCY CODE:

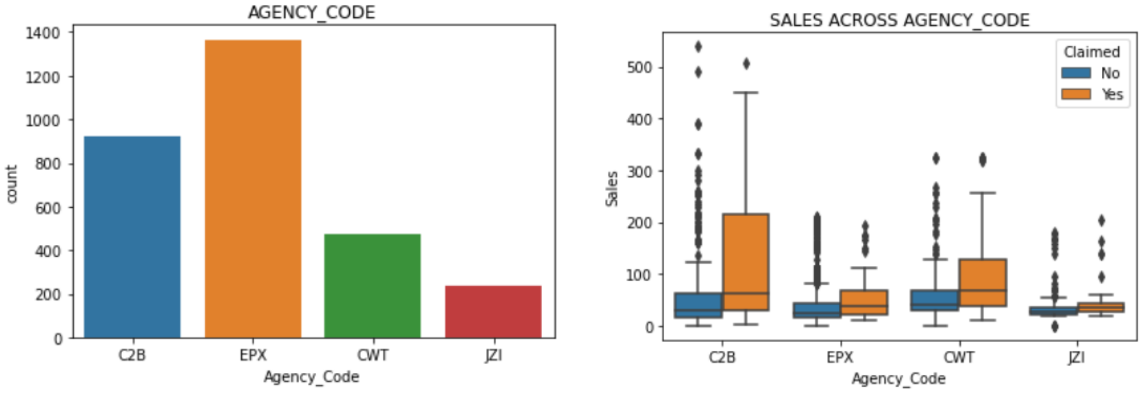


Fig 2.6 Agency Code

Now moving on to the categorical variable. Fig 2.6 talks about the agency code. There are 5 different agencies data that has been taken into consideration. The box plots have been created with respect to the Sales and whether or not the customer has claimed insurance. There seem to be a lot of outliers in the data.

- TYPE:

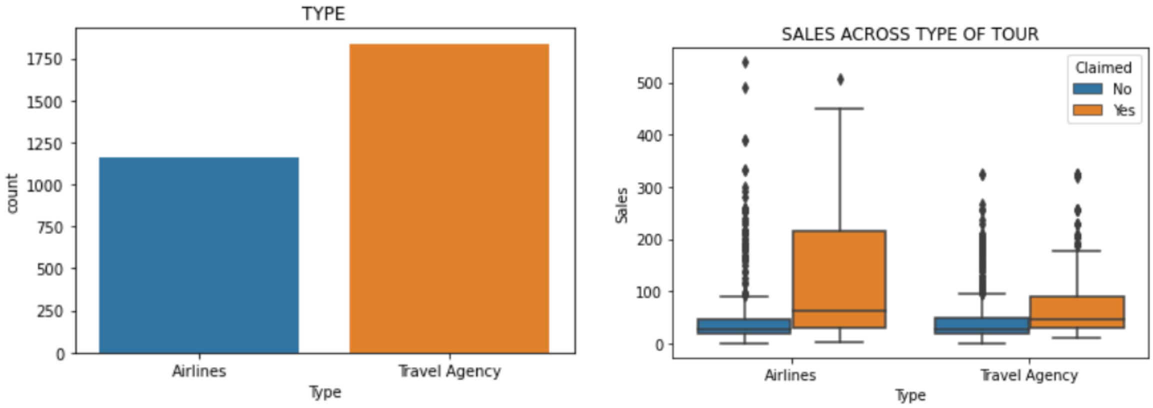


Fig 2.7 Type of Tour

Now moving on to the categorical variable. Fig 2.7 talks about the type. There are 2 different types ie: Airlines and Travel Agency. The box plots have been created with respect to the Sales and whether or not the customer has claimed insurance. There seem to be a lot of outliers in the data.

- CLAIM STATUS:

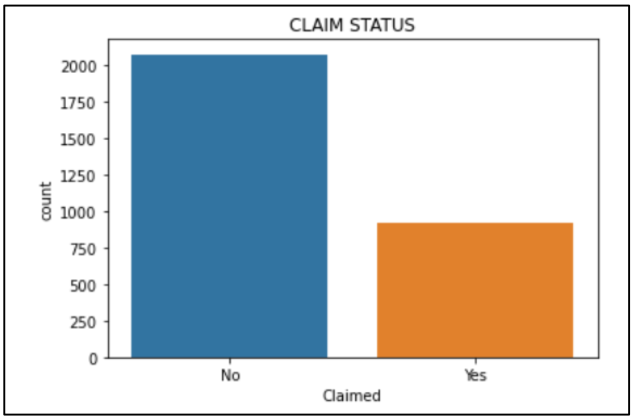


Fig 2.8 Claim Status

We can see that the above graph Fig 2.8 shows us the claim status or the number of tour insurance policies claimed. It is evident that not many still prefer a tour insurance based on the count plot presented above. It is still possible that there is an increase in claims than the years prior.

- CHANNEL:

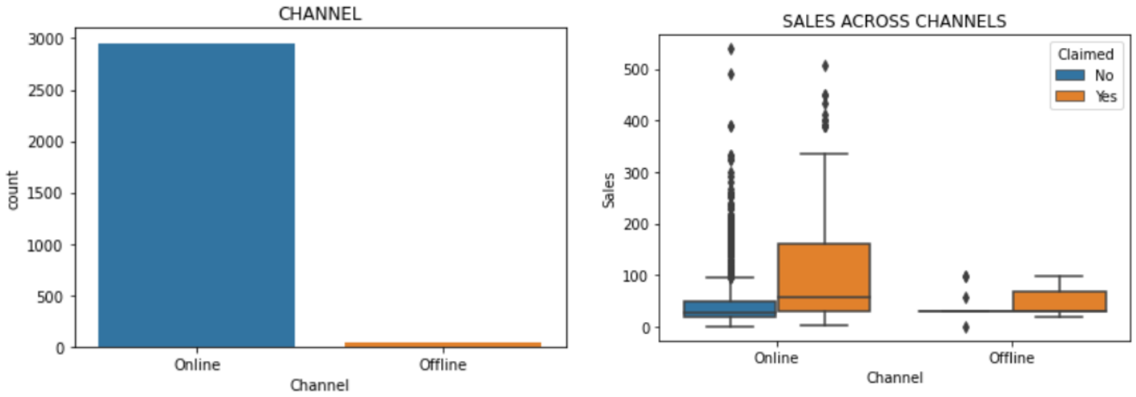


Fig 2.9 Channel

Now moving on to the categorical variable. Fig 2.9 talks about the Channel. There are 2 different types ie: Online and Offline. Clearly online channel sees most traffic. The box plots have been created with respect to the Sales and whether or not the customer has claimed insurance. There seem to be a lot of outliers in the data.

- INSURANCE PRODUCT NAME:

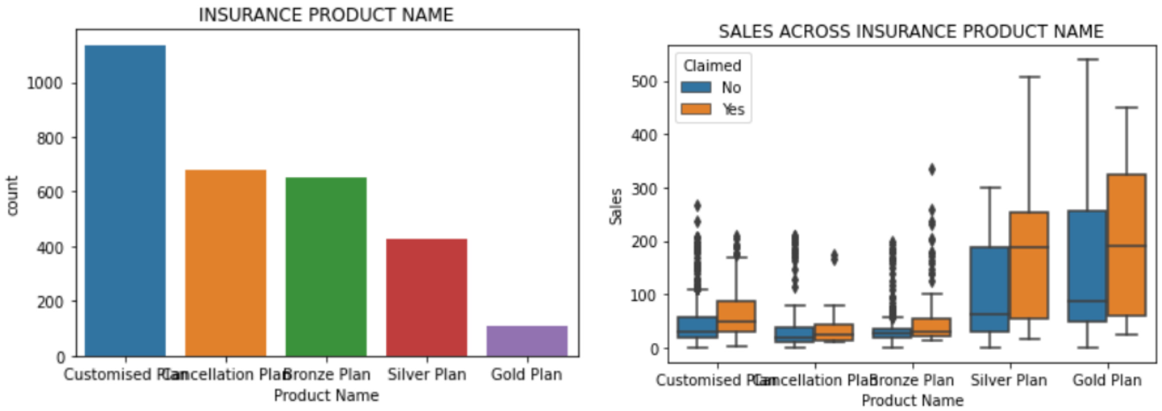


Fig 2.10 Insurance Product Name

Now moving on to the categorical variable. Fig 2.10 talks about the Product Name. There are 5 different Products offered. The most preferred plan is the Customized Plan. The box plots have been created with respect to the Sales and whether or not the customer has claimed insurance. There seem to be some outliers in the data, Silver and Gold plan do not seem to have any Outliers.

- DESTINATION:

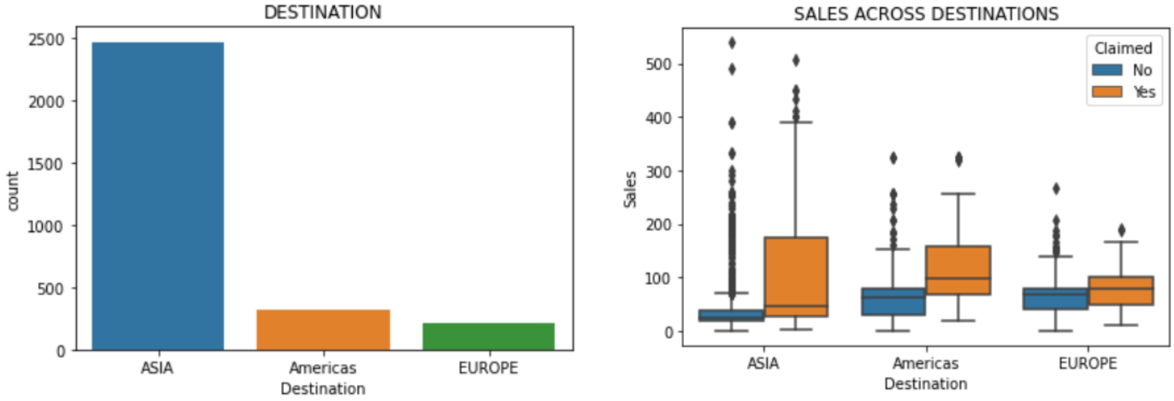


Fig 2.11 Destination

Fig 2.11 talks about the Destination. There are 3 different Destinations ie: Asia, America and Europe. The box plots have been created with respect to the Sales and whether or not the customer has claimed insurance. There seem to be some outliers in the data.

A. MULTIVARIATE ANALYSIS

Multivariate Analysis is a Statistical procedure for analysis of data involving more than one type of variable(s). A Pair plot and Heat map has been created to fulfil the Multivariate Analysis of the data provided.

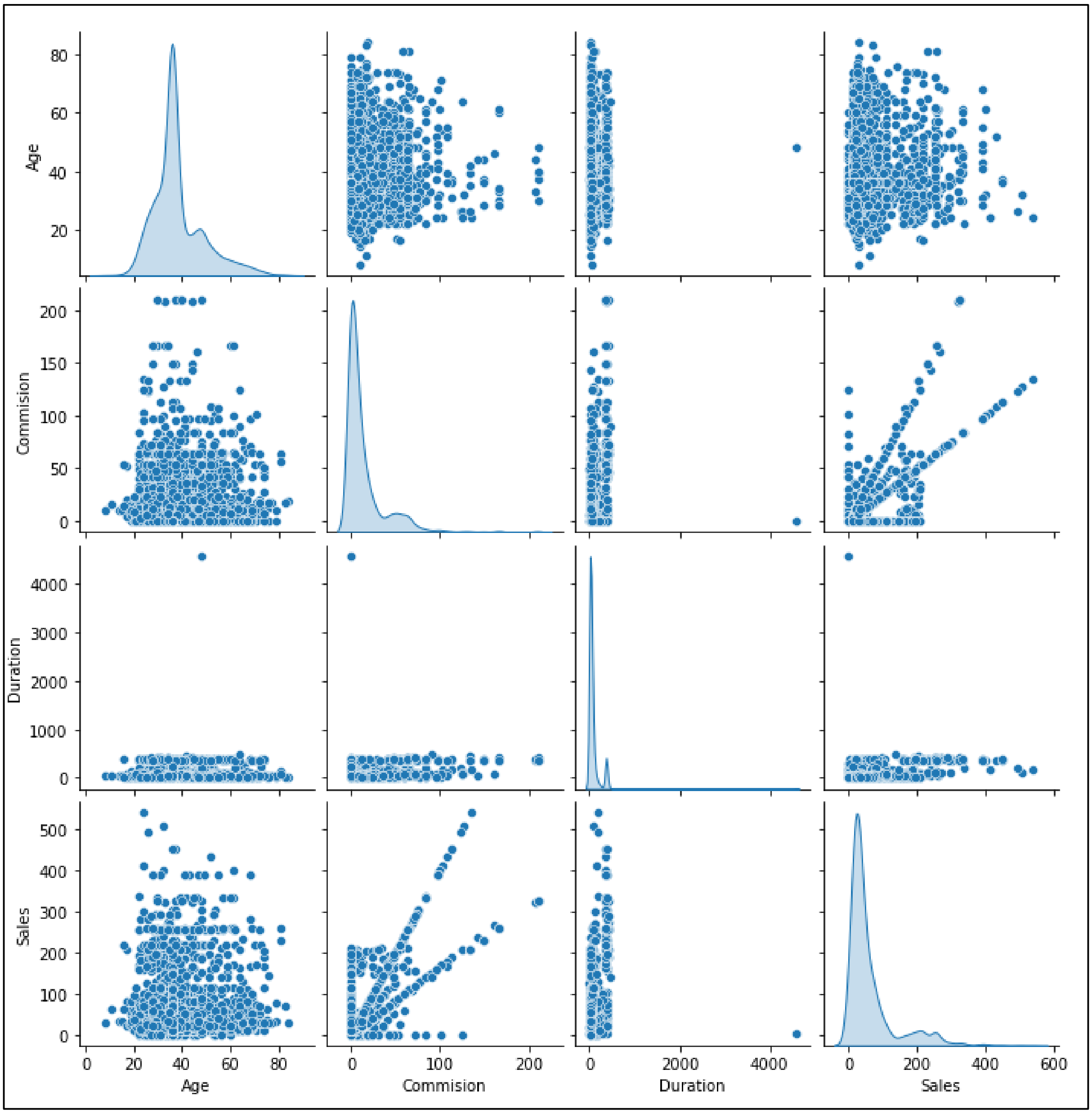


Fig 2.12 Multivariate analysis using a Pair plot

The above Fig 2.12 of a Pair plot essentially helps us understand the relationship between all the numerical values in the dataset. It helps us compare the variables to one another in a way that we are made to deeply understand the trends and patterns of the dataset provided.

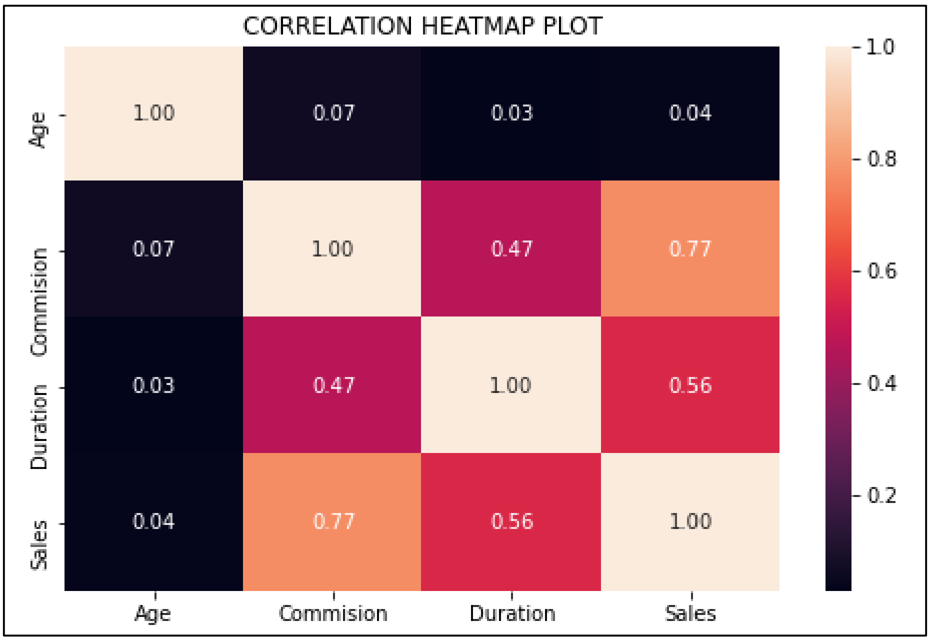


Fig 2.13 Multivariate analysis using a Correlation Heat map

The above Heat map (Fig 2.13) essentially shows us the correlation between pairs of different variables. It clearly shows us that the Sales is highly positively correlated with the commission as should be.

Once EDA through data visualisation is done we move on to converting all the Categorical data into numerical entries. This is done through Encoding.

One-Hot-Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.

This is required to do since the machine learning algorithms only works on the numerical data. That is why there is a need to convert the categorical column into numerical one.

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

We know that the target variable is ‘Claimed’. This target variable is also the independent variable and will at this step be split from the data in Y. While all the other dependent variables stay in X.

X = Agency_Code , Product Name, Sales, Type, Commission, Duration, Age, Destination, Channel

Y = “Claimed”

After X and Y have been split, it is now essential to split the data further into Train and Test data. According to our dataset we have taken a train-test split of 70% and 30%. As this is considered ideal in most scenarios. We now know that the training dataset consists of 2100 entries while the testing dataset consists only 900 out of the 3000 total.

Next we move on to build the 3 different types of models:

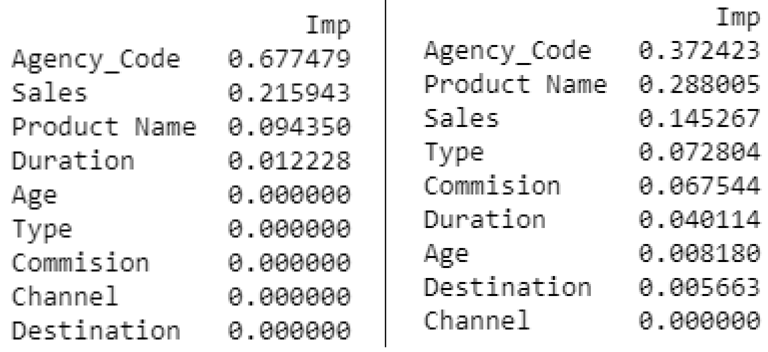
- CART for Decision Tree
- Random Forest
- Artificial Network (data scaled)

These models are built using parameter grids to choose the best parameters for each model and give us the best results possible.

The feature importance for:

Decision Tree Classifier

Random Forest



```
                   Imp
Agency_Code   0.677479
Sales         0.215943
Product Name  0.094350
Duration      0.012228
Age           0.000000
Type          0.000000
Commision     0.000000
Channel       0.000000
Destination   0.000000
```

```
                   Imp
Agency_Code   0.372423
Product Name  0.288005
Sales         0.145267
Type          0.072804
Commision     0.067544
Duration      0.040114
Age           0.008180
Destination   0.005663
Channel       0.000000
```

Fig 2.14 Feature Importance DTC vs RF

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

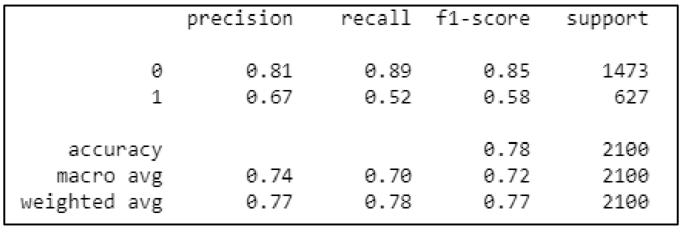
Once the models are built. There is a step where we are to fit the model on the training data and then go on to predict the training data and evaluate the training models performance. Then we go on to predict the test data and start comparing the model performances.

Below are the different Performance metrics:

CLASSIFICATION REPORTS:

1) Decision Tree

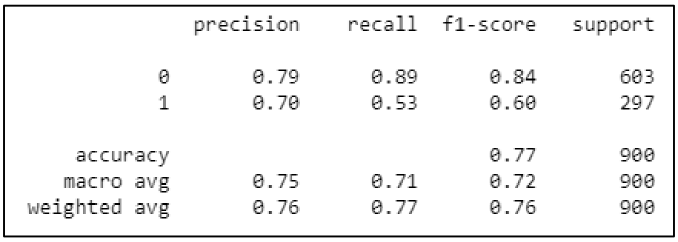
- Train data



```
              precision    recall  f1-score   support

           0       0.81      0.89      0.85      1473
           1       0.67      0.52      0.58       627

    accuracy                           0.78      2100
   macro avg       0.74      0.70      0.72      2100
weighted avg       0.77      0.78      0.77      2100
```

Fig 2.15(a) DTC- Classification Report of Train Data

- Test Data



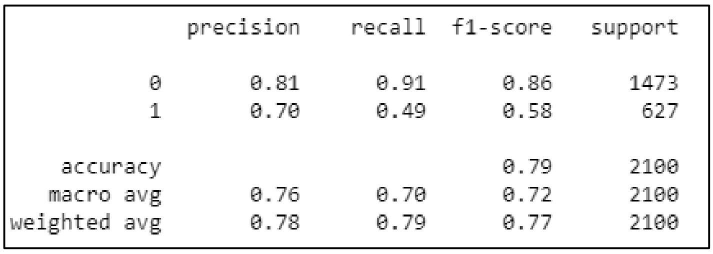
```
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       603
           1       0.70      0.53      0.60       297

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

Fig 2.15(b) DTC- Classification Report of Test Data

From the above classification report we see that the accuracy is almost similar. The test accuracy is at 77%. The precision and recall value as we see has improved on the test dataset.

2) Random Forest

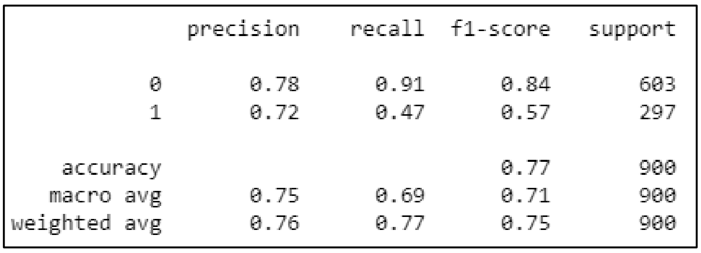
- Train Data



```
              precision    recall  f1-score   support

           0       0.81      0.91      0.86      1473
           1       0.70      0.49      0.58       627

    accuracy                           0.79      2100
   macro avg       0.76      0.70      0.72      2100
weighted avg       0.78      0.79      0.77      2100
```

Fig 2.16(a) RF- Classification Report of Train Data

- Test Data



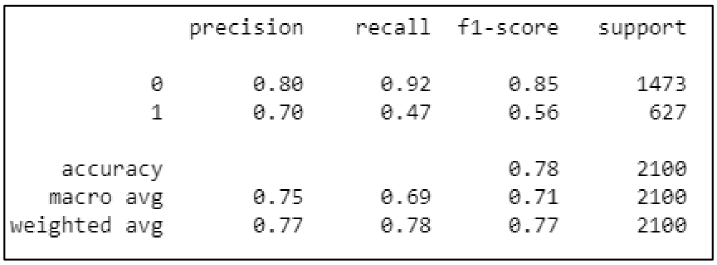
```
              precision    recall  f1-score   support

           0       0.78      0.91      0.84       603
           1       0.72      0.47      0.57       297

    accuracy                           0.77       900
   macro avg       0.75      0.69      0.71       900
weighted avg       0.76      0.77      0.75       900
```

Fig 2.16(b) RF- Classification Report of Test Data

From the above classification report we see that the accuracy is almost similar. The test accuracy is at 77%. The precision and recall value as we see has decreased on the test data set.

3) Artificial Network

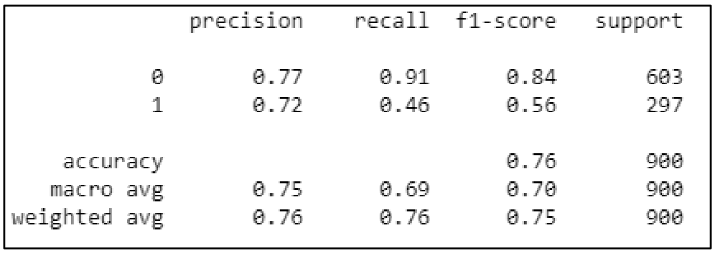
- Train Data



```
              precision    recall  f1-score   support

           0       0.80      0.92      0.85      1473
           1       0.70      0.47      0.56       627

    accuracy                           0.78      2100
   macro avg       0.75      0.69      0.71      2100
weighted avg       0.77      0.78      0.77      2100
```

Fig 2.17(a) ANN- Classification Report of Train Data

- Test Data



```
              precision    recall  f1-score   support

           0       0.77      0.91      0.84       603
           1       0.72      0.46      0.56       297

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

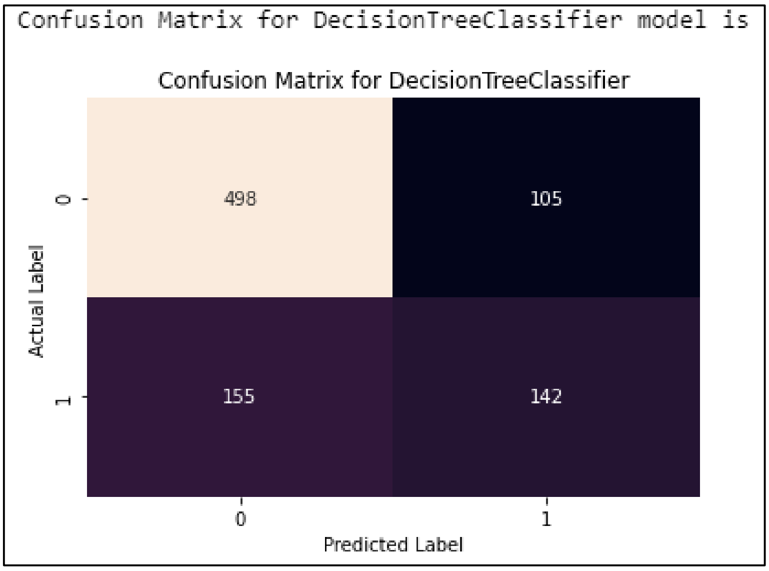
Fig 2.17(b) ANN- Classification Report of Test Data

From the above classification report we see that the accuracy is almost similar. The test accuracy is at 76%, which is a little lesser than the train data accuracy of 78%. The precision value has increased while, the recall value as we see has decreased on the test data set.

CONFUSION MATRIX:

1) CART for Decision Tree

- Test Data



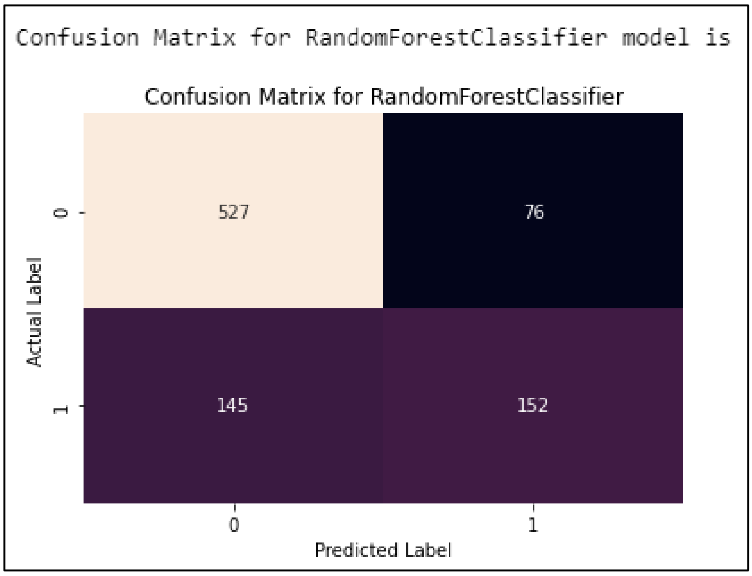
```
Confusion Matrix for DecisionTreeClassifier model is
```

Fig 2.18 DTC- Confusion Matrix

The True Negative value is at 498 while the True Positive value is at 142. There are a significant number of False Positives and False Negatives.

2) Random Forest

- Test Data



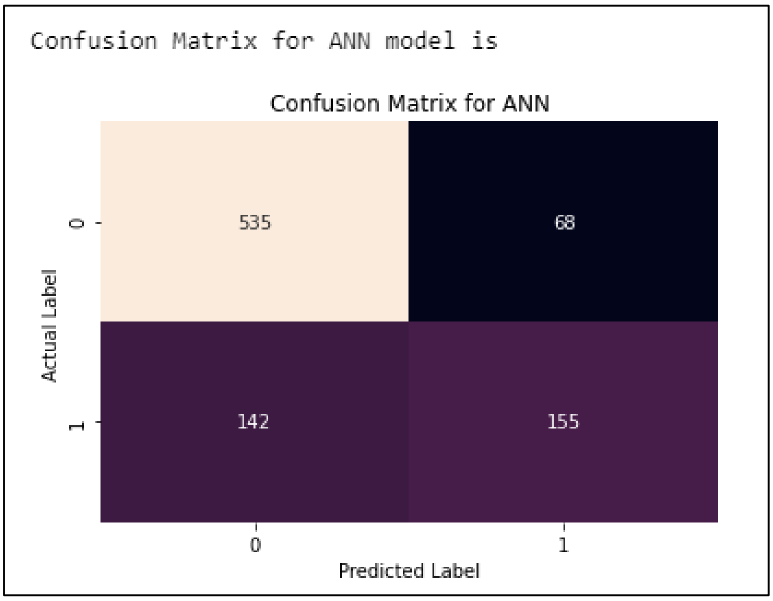
```
Confusion Matrix for RandomForestClassifier model is
```

Fig 2.19 RF- Confusion Matrix

The True Negative value is at 527 while the True Positive value is at 152. There are a significant number of False Positives and False Negatives.

3) Artificial Network (data scaled)

- Test Data



```
Confusion Matrix for ANN model is
```

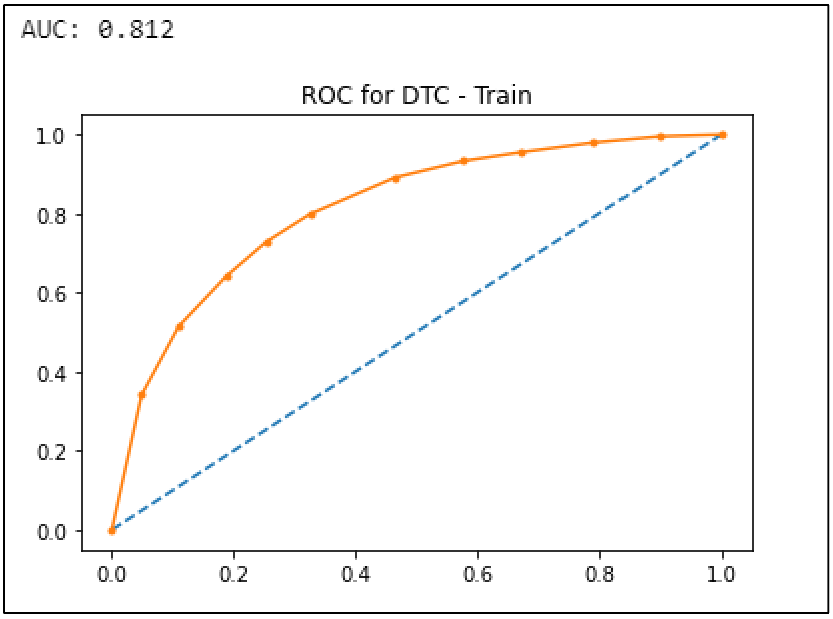
Fig 2.20 ANN- Confusion Matrix

The True Negative value is at 535 while the True Positive value is at 155. There are a significant number of False Positives and False Negatives.

ROC CURVE AND ROC_AUC CURVE:

1) CART for Decision Tree

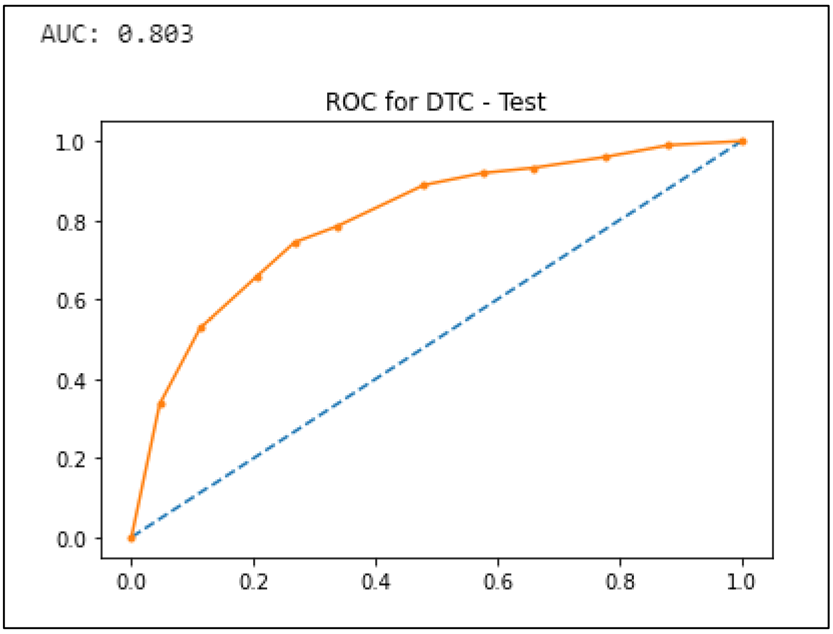
- Train Data



```
AUC: 0.812
```

Fig 2.21(a) DTC train- ROC Curve

- Test Data



```
AUC: 0.803
```

Fig 2.21(b) DTC test- ROC Curve

The Area Under the Curve (AUC) for the test data is 80.3% while the AUC of the train data is 81.2%.

2) Random Forest

- Train Data

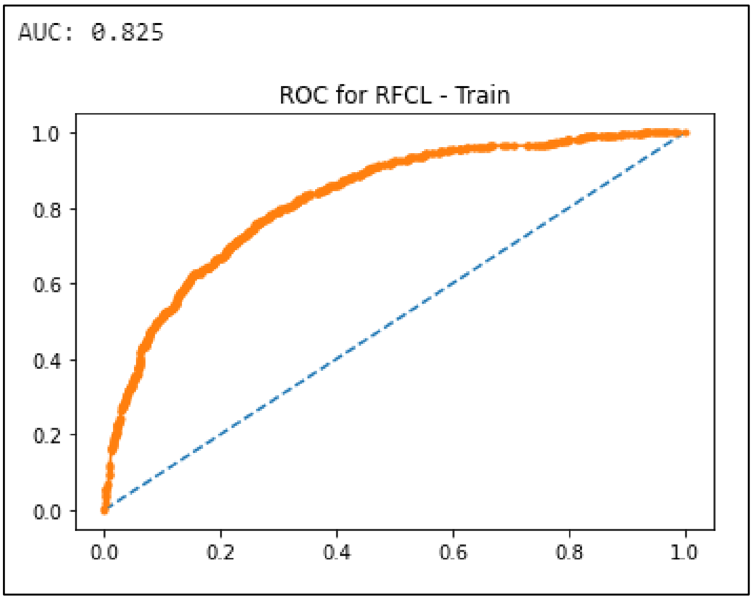


Fig 2.22(a) RF train- ROC Curve

- Test Data

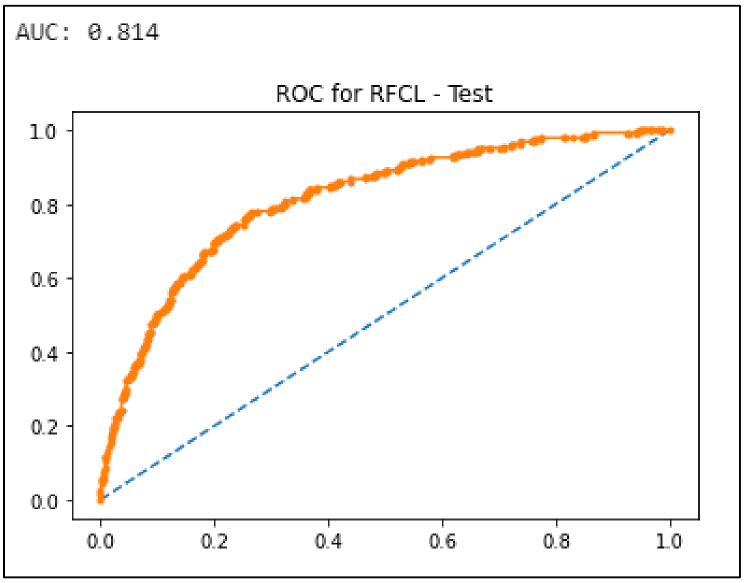
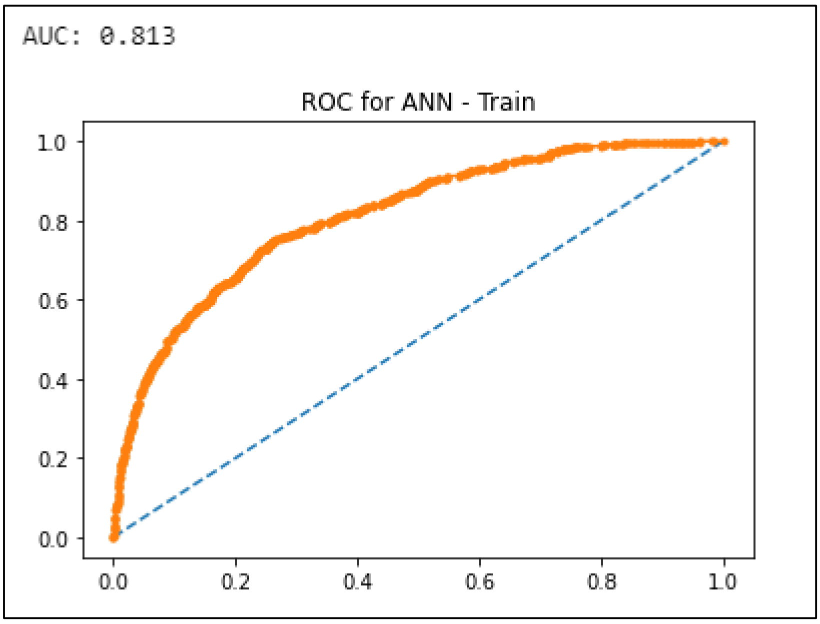


Fig 2.22(b) RF test- ROC Curve

The Area Under the Curve (AUC) for the test data is 81.4% while the AUC of the train data is 82.5%.

3) Artificial Network (data scaled)

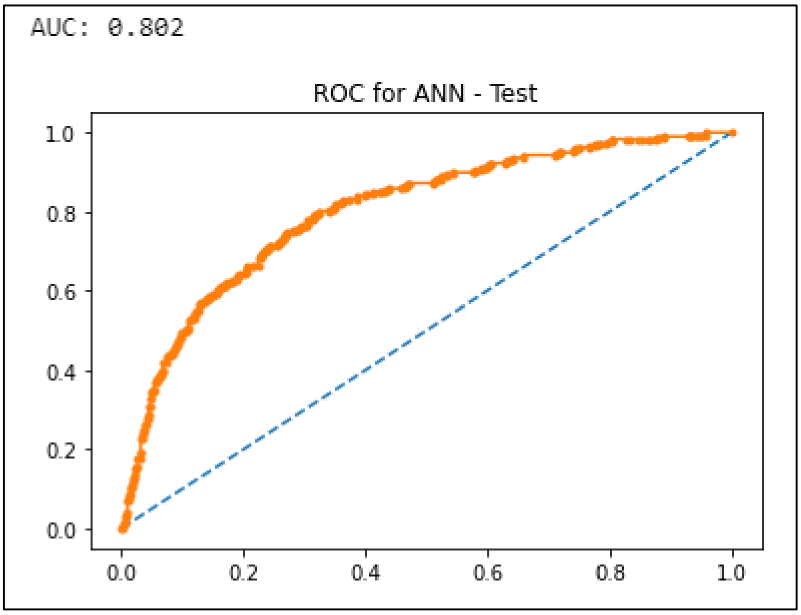
- Train Data



```
AUC: 0.813
```

Fig 2.23(a) ANN train- ROC Curve

- Test Data



```
AUC: 0.802
```

Fig 2.23(b) ANN test- ROC Curve

The Area Under the Curve (AUC) for the test data is 80.2% while the AUC of the train data is 81.3%.

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

Performance Metrics	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.78	0.77	0.79	0.77	0.78	0.76
AUC	0.81	0.80	0.82	0.81	0.81	0.80
Recall	0.52	0.53	0.49	0.47	0.47	0.46
Precision	0.67	0.70	0.70	0.72	0.70	0.72
F1 Score	0.58	0.60	0.58	0.57	0.56	0.56

Table 2.2 Comparison between all the models

From the above table we can clearly infer that the Accuracy, Area Under the Curve and Precision is best for the Random Forest model in comparison to the other two models(CART and ANN). Though the Recall and F1 Score is not the best of the lot, but since it is almost on the similar value lines for all the models minute differences can be ignored. Majorly out of 6 parameters, 4 best is still a very helpful in determining the optimal choice. Hence I have chosen the Random Forest model as the best one with respect to the data provided.

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

- There are 10 variables in total out of which Age, Commission, Duration and Sales are numeric in nature while the remaining 6 variables are Categorical in nature. These Categorical variables include Agency_code, Type, Claimed, Channel, Product Name and Destination. These will later be encoded as we proceed. There are no missing values in the data. There are 9 independent variables and 1 dependent variable, ie: Claimed. Since we are asked to investigate on the higher claim frequency of tour insurance. The dependent variable here is also called the target variable.

- The data collected seems to be limited, if the data were more we would be able to maybe find a better model fit and there would have been more data to test on, hence to some extent helping us increase the recall, precision, accuracy, etc.

- From the above table we can clearly infer that the Accuracy, Area Under the Curve and Precision is best for the Random Forest model in comparison to the other two models(CART and ANN).

- The above Heat map essentially shows us the correlation between pairs of different variables. It clearly shows us that the Sales is highly positively correlated with the commission as should be.

- There is more traffic on the Online channel than there is on the offline, the customers seemed to like the online experience better. And most customers chose to Customize their plans according to their needs, telling us prominently that in today’s day and age people value customizations over anything and that is a big chuck of where are revenue comes from.

- There was another interesting find people prefer going through Travel Agencies rather than the Airlines directly. And those who went through Agencies were to some extent convinced to claim tour insurance for their safety. There was a 3rd party push involved here.

- The agencies should aim to make the Online process of claiming insurance more easier and safer for the customers.

- The agencies have to get a hold of the footfall and convert them into potential tour insurance customers. There can be follow ups if need be, or options to take up the insurance within a stipulated period of time after purchase of a travel package.