**Project**

**Data Mining**

**-By R.Prakash**

**TABLE OF CONTENTS**

**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. **Please note that it is a summarized data that contains the average values in all the columns considering all the months, and not for any particular month**. 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……………………5

(Univariate, Bi-variate, and multivariate analysis).

**1.2** Do you think scaling is necessary for clustering in this case? Justify………………………………14

**1.3** Apply hierarchical clustering to scaled data. Identify the number of optimum clusters…..15

using Dendrogram and briefly describe them

**1.4** Apply K-Means clustering on scaled data and determine optimum clusters…………………..19   
 Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

**1.5** Describe cluster profiles for the clusters defined. Recommend different promotional…..22 strategies for different clusters.

**Problem 2: CART-RF-ANN**

An Insurance firm providing tour insurance is facing higher claim frequency……………………….23 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…………………..23 (Univariate, Bi-variate, and multivariate analysis).

**2.2** Data Split: Split the data into test and train, build classification model CART,………………..35 Random Forest, Artificial Neural Network

**2.3** Performance Metrics: Comment and Check the performance of Predictions on……………39 Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.

**2.4** Final Model: Compare all the models and write an inference which model is…………………48 best/optimized.

**2.5** Inference: Based on the whole Analysis, what are the business insights and………………….49 recommendations

**List of Tables:**

**Table-1.1-Head of the dataset…………………………………………………………………………………………..5**

**Table 1.2-Before scaling…………………………………………………………………………………………………..14**

**Table 1.3-After scaling……………………………………………………………………………………………………..15**

**Table-1.4-Cluster profile kmeans……………………………………………………………………………………..20**

**Table 2.1-Five point summary………………………………………………………………………………………….23**

**Table 2.2-Before Encoding……………………………………………………………………………………………….35**

**Table 2.3-After Encoding………………………………………………………………………………………………….36**

**Table 2.4-Independent variable………………………………………………………………………………………..36**

**List of figures:**

**Fig-1.1-Boxplot and histplot for first 3 variables…………………………………………………………………6**

**Fig-1.2-Boxplot and histplot for last 3 variables…………………………………………………………………8**

**Fig-1.3-Multivariate plot………………………………………………………………………………………………….10**

**Fig 1.4-Correlation plot……………………………………………………………………………………………………11**

**Fig 1.5-Outliers……………………………………………………………………………………………………………….12**

**Fig 1.6-Outliers probability………………………………………………………………………………………………13**

**Fig 1.7- Outliers after treatment………………………………………………………………………………………13**

**Fig 1.8-Dendogram………………………………………………………………………………………………………….15**

**Fig 1.9-Truncated dendogram………………………………………………………………………………………….16**

**Fig 1.10-Cluster profile Hierarchial………………………………………………………………………………….17**

**Fig 1.11-Elbow method……………………………………………………………………………………………………20**

**Fig 2.1-Boxplot………………………………………………………………………………………………………………..24**

**Fig-2.2-Count plot Agency……………………………………………………………………………………………….26**

**Fig-2.3-count plot agency target……………………………………………………………………………………..26**

**Fig2.4-count plot type……………………………………………………………………………………………………..27**

**Fig-2.5-count plot type target…………………………………………………………………………………………28**

**Fig 2.6-Count plot claimed………………………………………………………………………………………………29**

**Fig 2.7-Count plot claimed target…………………………………………………………………………………….30**

**Fig 2.7-Count plot channel………………………………………………………………………………………………30**

**Fig 2.9-Count plot Destination…………………………………………………………………………………………31**

**Fig 2.10-Count plot product name……………………………………………………………………………………31**

**Fig 2.11-Count plot product name……………………………………………………………………………………32**

**Fig-2.12-Plot for Age……………………………………………………………………………………………………….32**

**Fig-2.13-BoxPlot for Age…………………………………………………………………………………………………32**

**Fig-2.14-Plot for commission…………………………………………………………………………………………..33**

**Fig-2.15-Plot for sales……………………………………………………………………………………………………..33**

**Fig-2.16-Plot for sales claim…………………………………………………………………………………………….33**

**Fig-2.17-Plot for Duration……………………………………………………………………………………………….34**

**Fig-2.18-Plot for duration claim……………………………………………………………………………………….34**

**Fig-2.19-Pairplot……………………………………………………………………………………………………………..35**

**Fig-2.20-Correlation plot…………………………………………………………………………………………………36**

**Fig 2.21-ROC -Decision tree-Train…………………………………………………………………………………..38**

**Fig-2.22-ROC-Decision tree-Test…………………………………………………………………………………….39**

**Fig-2.23-ROC-Random forest Train…………………………………………………………………………………39**

**Fig-2.24-ROC-Random forest Test………………………………………………………………………………….40**

**Fig-2.25-ROC-ANN- Train……………………………………………………………………………………………….40**

**Fig-2.26-ROC-ANN Test………………………………………………………………………………………………….41**

**Fig-2.27-Decision tree confusion matrix- Train………………………………………………………………...41**

**Fig-2.28-Decision tree confusion matrix- Test………………………………………………………………….42**

**Fig-2.29-Random forest- confusion matrix- Train…………………………………………………………….42**

**Fig-2.30-Random forest- confusion matrix- Test………………………………………………………………43**

**Problem 1: Clustering**

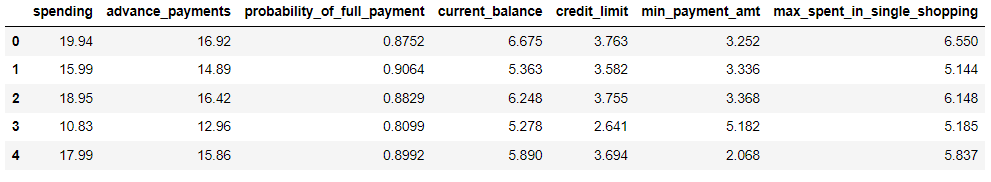
**A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage**.

**Data** **Dictionary** **for** **Market** **Segmentation:**

1. spending: Amount spent by the customer per month (in 1000s)
2. advance\_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability\_of\_full\_payment: Probability of payment done in full by the customer to the bank
4. current\_balance: Balance amount left in the account to make purchases (in 1000s)
5. credit\_limit: Limit of the amount in credit card (10000s)
6. min\_payment\_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. max\_spent\_in\_single\_shopping: Maximum amount spent in one purchase (in 1000s)
   1. **Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).**

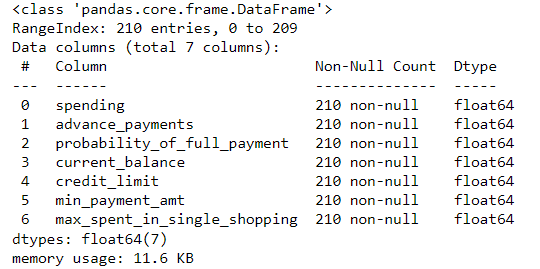
**Inference from the dataset:**

The first five rows of the data set are as below:



Tab-1.1-Head of the dataset

The inference from the dataset can be ascribed from below



The dataset has 210 rows and 7 columns. All the variables are given in **float** data type.

**Data pre-processing:**

**a) Null values:**

The dataset has **Nill** null values.

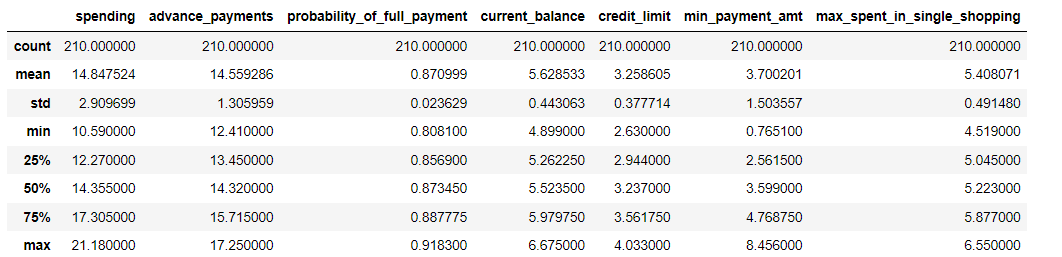
**b) Duplicates:**

The dataset has no duplicate rows.

**Exploratory Data Analysis:**

**a) Univariate analysis:**

**5-point summary**



**Boxplot and histplot for variables-spending,advance payment and probability**

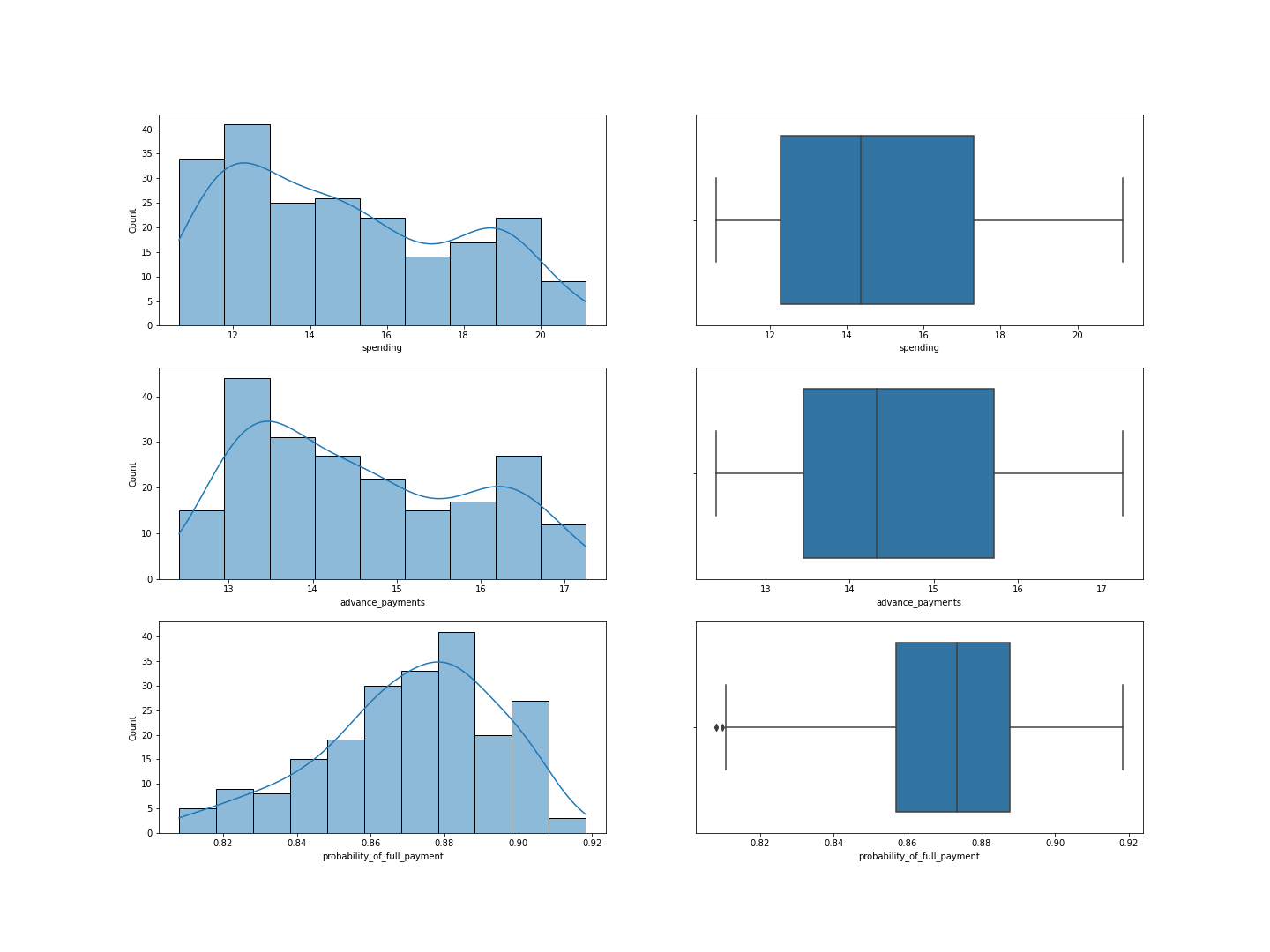


Fig-1.1-Boxplot and histplot for first 3 variables

**a)Spending:**

1. It tells the amount of money spent by a customer per month from credit card.
2. Average amount spent-14,847 rupees.
3. Minimum amount spent-10,590 rupees.
4. Maximum amount spent-21,180 rupees.
5. Median value-14,355 rupees.
6. 25% of the customers spent less than or equal to-12,270 rupees.
7. 75% of the customers spent less than or equal to-17,305 rupees
8. It is **right skewed**.
9. There are **no outliers** in the dataset.
10. The minimum amount spent by a customer per month is 21,180 rupees.

**b) advance\_payments**:

1. It tells the amount paid by a customer in advance before credit statement is generated by the bank for a particular moth.
2. The minimum amount paid by a customer in advance by cash -1,456 rupees.
3. The maximum amount paid by a customer in advance by cash -1,725 rupees.
4. 25% of the customers paid in advance less than or equal to-1,345 rupees.
5. Almost 75 percentage of the customers paid less than or equal to 1,571 rupees in advance.
6. It is **rightly skewed** and there are **no outliers**.

**c) Probability\_of\_full\_payment:**

1. It tells the probability that the customer will pay the full amount generated for the month.
2. It is expected that all customers have a probability of making due amount in full is at least 80 percent.
3. Average probability-87%.
4. A customer has a maximum probability to make payment is at 91.8%.
5. 25% of the customers have less than or equal to 85.6% probability to make full payment.
6. 75% of the customers have less than or equal to 88% probability to make full payment.
7. There are **few outliers** in the dataset.
8. Left skewed.

**Boxplot and histplot for variables-current balance,credit limit,minimum payment amount and maximum spent in single shopping.**

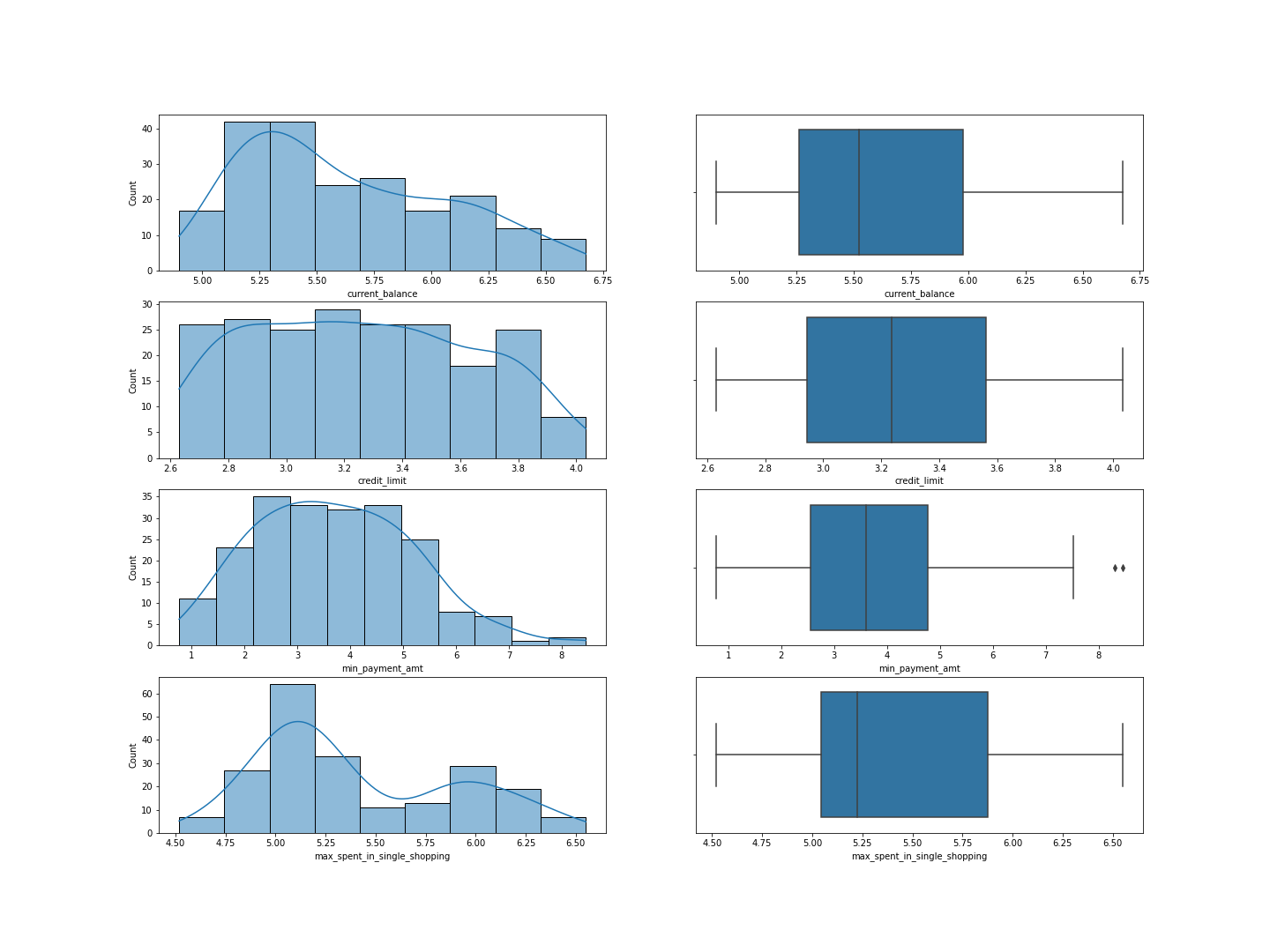


Fig-1.2-Boxplot and histplot for last 3 variables

**d)current\_balance:**

1. It denotes the amount that the customer has as credit card balance to make purchases.
2. All customers have at least 4,899 rupees as their balance.
3. Average balance-5,628 rupees.
4. Maximum balance held by a customer-6,675 rupees.
5. 25% of the customers have less than 5,262 rupees as credit card balance.
6. 75% of the customers have less than 5,979 rupees as credit card balance.

**e)credit\_limit:**

1. It denotes minimum credit limit sanctioned for a customer by the bank.
2. The lowest credit limit is 26,300 rupees.
3. The highest credit limit is 40,330 rupees.
4. Average credit limit is 32,586 rupees.
5. Almost 25 percent of the customers have credit limit less than or equal to 29,440 rupees.
6. Almost 75 percent of the customers have credit limit less than or equal to 35,617 rupees.

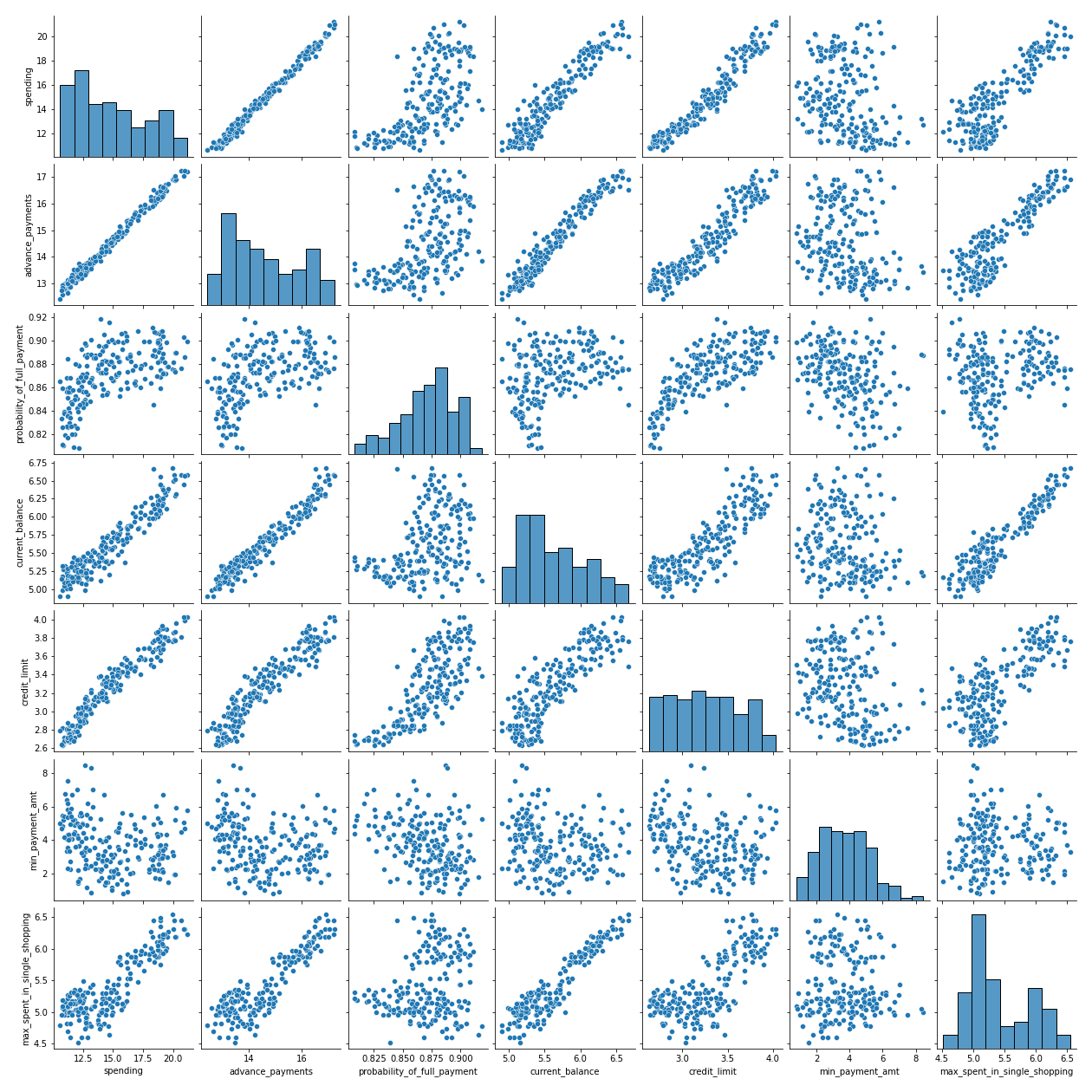
**f)min\_payment\_amt :**

1. It denotes the minimum payment paid by a customer instead of paying the entire credit card bill amount on an average per month.
2. Minimum amount paid by all customers is at least 76 rupees.
3. Average minimum amount paid-370 rupees.
4. 25 percent of the customers paid less than or equal to 256 rupees as minimum payment.
5. Maximum amount paid as minimum by a customer is 845 rupees.
6. There are **outliers** in this variable.

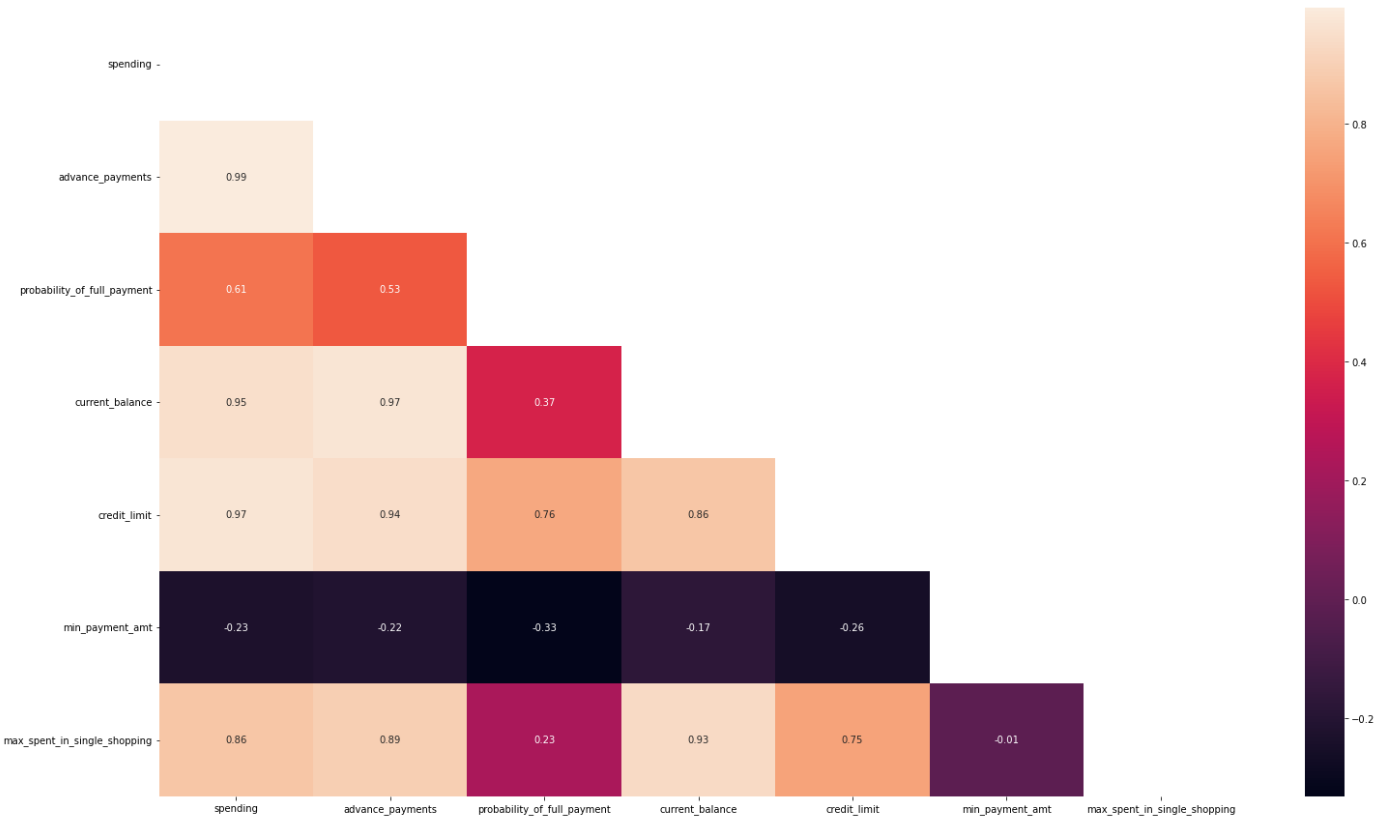
**g)max\_spent\_in\_single\_shopping:**

1. It denotes the maximum amount spent by the customer for a single transaction using the credit card.
2. Average maximum amount-5,408 rupees
3. Maximum amount-6,550 rupees.
4. Minimum amount-4,519 rupees.
5. It is **right skewed**.

**Multivariate analysis:**



**Fig-1.3-Multivariate plot**



**Fig 1.4-Correlation plot**

**Positive correlation:**

* + 1. Spending and advance\_payment-It means the customer who spends more also paid more credit in advance.
    2. Spending and current\_balance-When the customer spending increases, the current balance of the customer also increases.
    3. Spending and credit limit: The person who spends more in a month also has high credit limit which is obvious.
    4. Current balance and Advance payments: The customer who has high amount as current balance also made high payments for his credit card in advance.
    5. Current balance and maximum spent in single shopping: The customer who has high current balance also spent more amount in single shopping. It indicates the economic potential of the customer to spend more.

**Negative Correlation**:

1. There exists a slight negative correlation between probability of full payment and minimum payment amount.

As the dataset has **multicollinearity**, it reduces the reduces the statistical significance of independent variables. But, multicollinearity has little impact on clustering .

**Data Preparation:**

**a)Outlier detection and treatment:**

Outliers affects clustering process, so it needs to be treated.

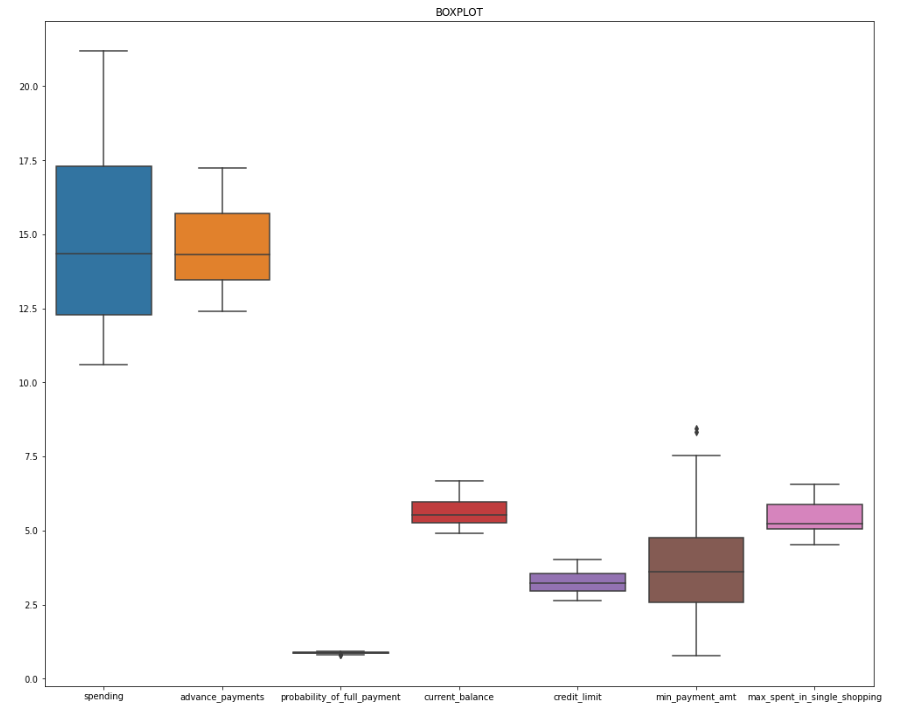
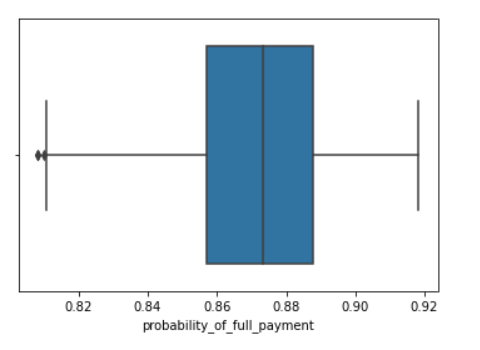


Fig1.5-Outliers

* From the box plot, it is clear that the variables ‘**probability\_of\_full\_payment’** and ‘**min\_payment\_amt** ‘ has few outliers.
* The presence of outliers in ‘min\_payment\_amt’ indicates that it may impact the overall prediction. So, the outliers need to be treated.
* We also need to treat the ‘probability’ variable.
* By using boxplot method, we treat the outliers.

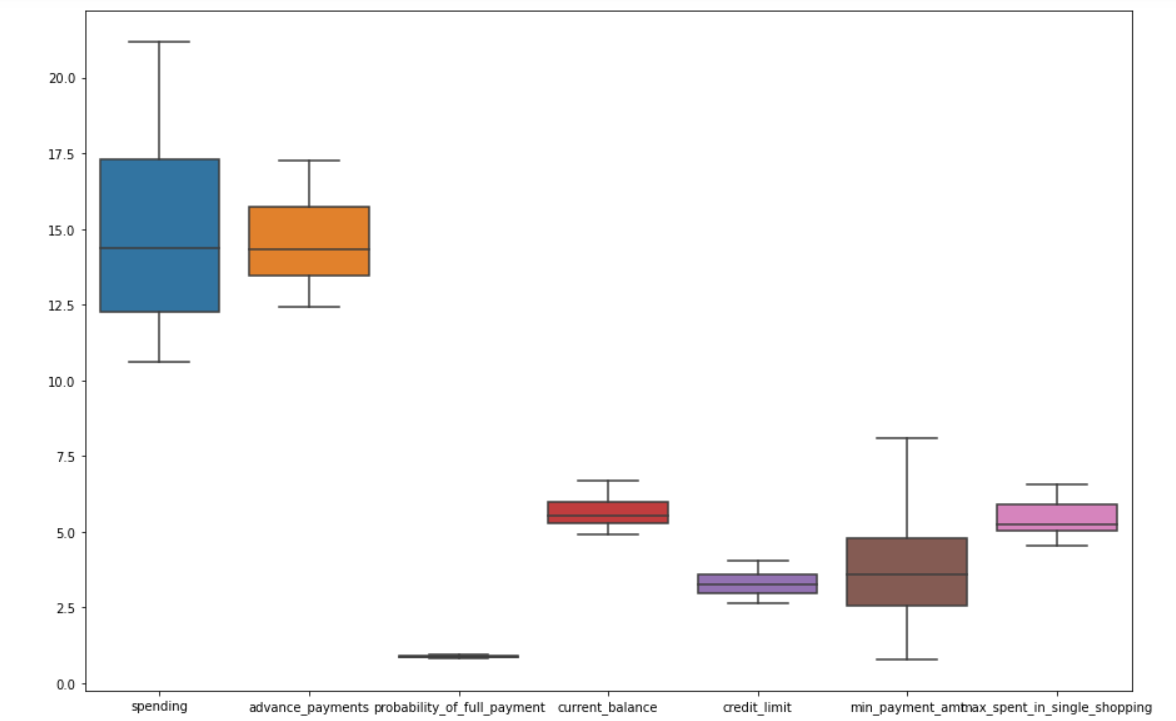
**Boxplot method or Winsorization:**

1. Any data point more than Q3+1.5\*IQR or less than Q1-1.5\*IQR is taken as an outlier
2. 50% of data points are within ±0.5 IQR of the median
3. In a normal distribution 68% are with ±1σ
4. So IQR (50%) is slightly less than ±1σ(68%)
5. In order to correspond ±3 σ range, ±1.5IQR (i.e. 3\* ±0.5 IQR) is taken as range to identify outliers.
6. Then the outlier is replaced with higher and lower range values.

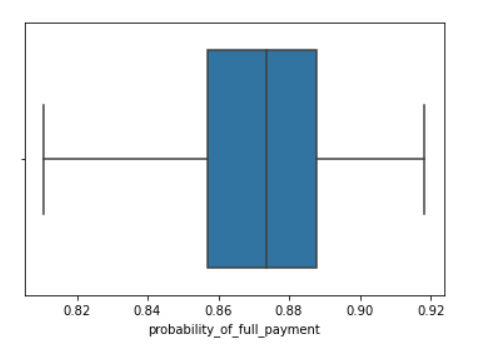


**Fig 1.6-Outliers probability**

**After outlier treatment:**

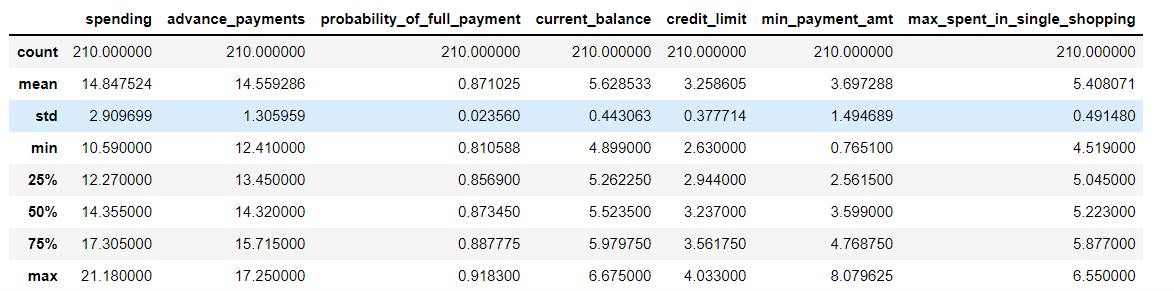


**Fig 1.7- Outliers after treatment**



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

**Before Scaling:**



**Table 1.2-Before scaling**

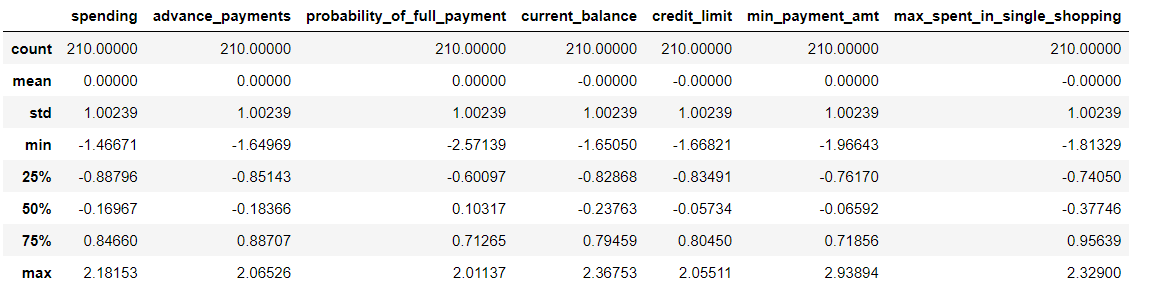
Scaling is the technique to bring the data points closer to each other. As ***clustering is a distance-based algorithm***, scaling is necessary. In this data set, scaling is done for three reasons:

1)For easier comparison among different measurement units. For example, in this dataset variable ‘Probability’ is given in percentage and others are given in absolute numbers.

2)High variance between variables. Variables with high variance may get higher weightage than the other variables thus influencing the model building. For instance, the standard deviation of variable ‘min\_payment\_amt’ is just 367 whereas for variable ‘credit\_limit’ it is 3,777.

3) It is also time-consuming to train and run the algorithm with this wide variation. So, it is necessary to scale down the data points around the mean as zero.

**After Scaling**

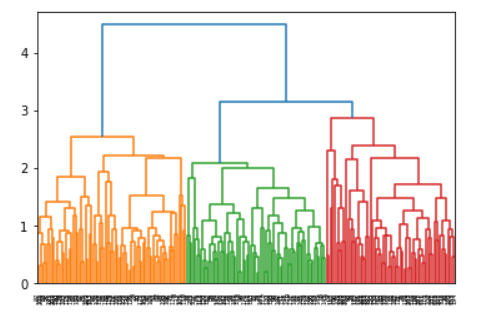


**Table 1.3-After scaling**

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

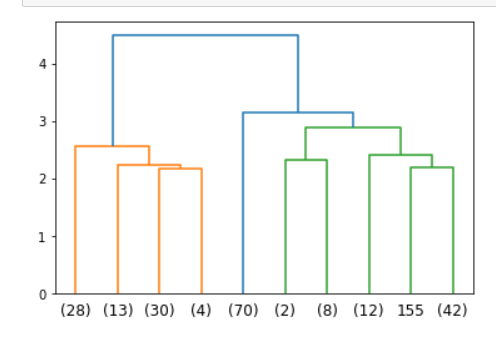
**Hierarchical clustering by divisive method:**

* In this clustering technique, records are sequentially grouped to create clusters based on distances between records and distances between clusters.
* It also produces a useful graphical display of the clustering process and results, called a dendrogram.
* Below is the dendrogram created by using **average linkage** method. The vertical lines indicate the distance between the clusters. **More the distance of the vertical lines in the dendrogram, more the distance between those clusters.**

****

**Fig 1.8-Dendogram**

As the dendrogram looks little clumsy we truncate it for better understanding. Truncation is used to condense the dendrogram.

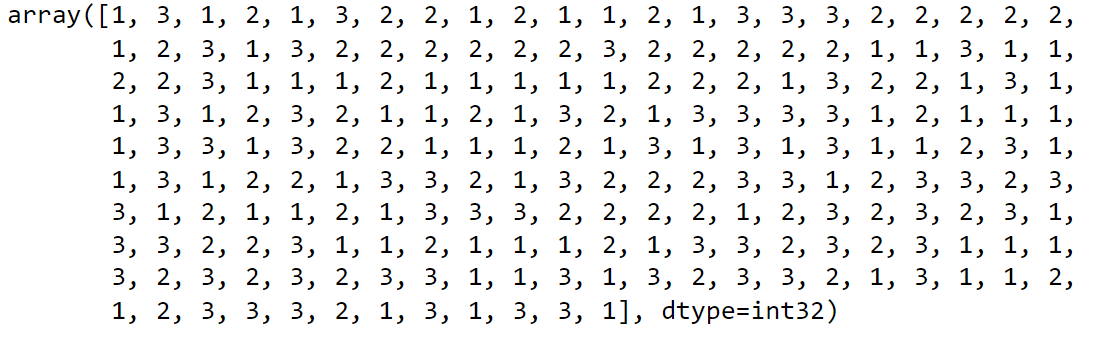
****

**Fig 1.9-Truncated dendogram**

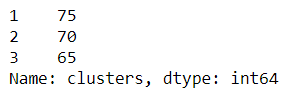
* The dendrogram is cut at appropriate level to determine the number of clusters.
* The value at the bottom indicates the number of observations at leaf nodes.

**F-Cluster:**

* By using **f cluster** and **criterion as distance,** we arrange the dataset into number of clusters.
* Now, we can set a threshold distance and draw a horizontal line (Generally, we try to set the threshold in such a way that it cuts the tallest vertical line). Let’s set this threshold as 2.9 and draw a horizontal line. We get the below 3 clusters.

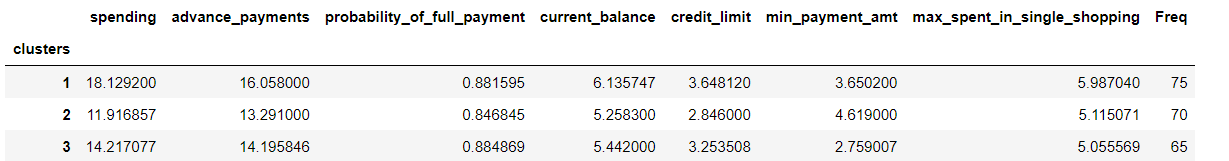


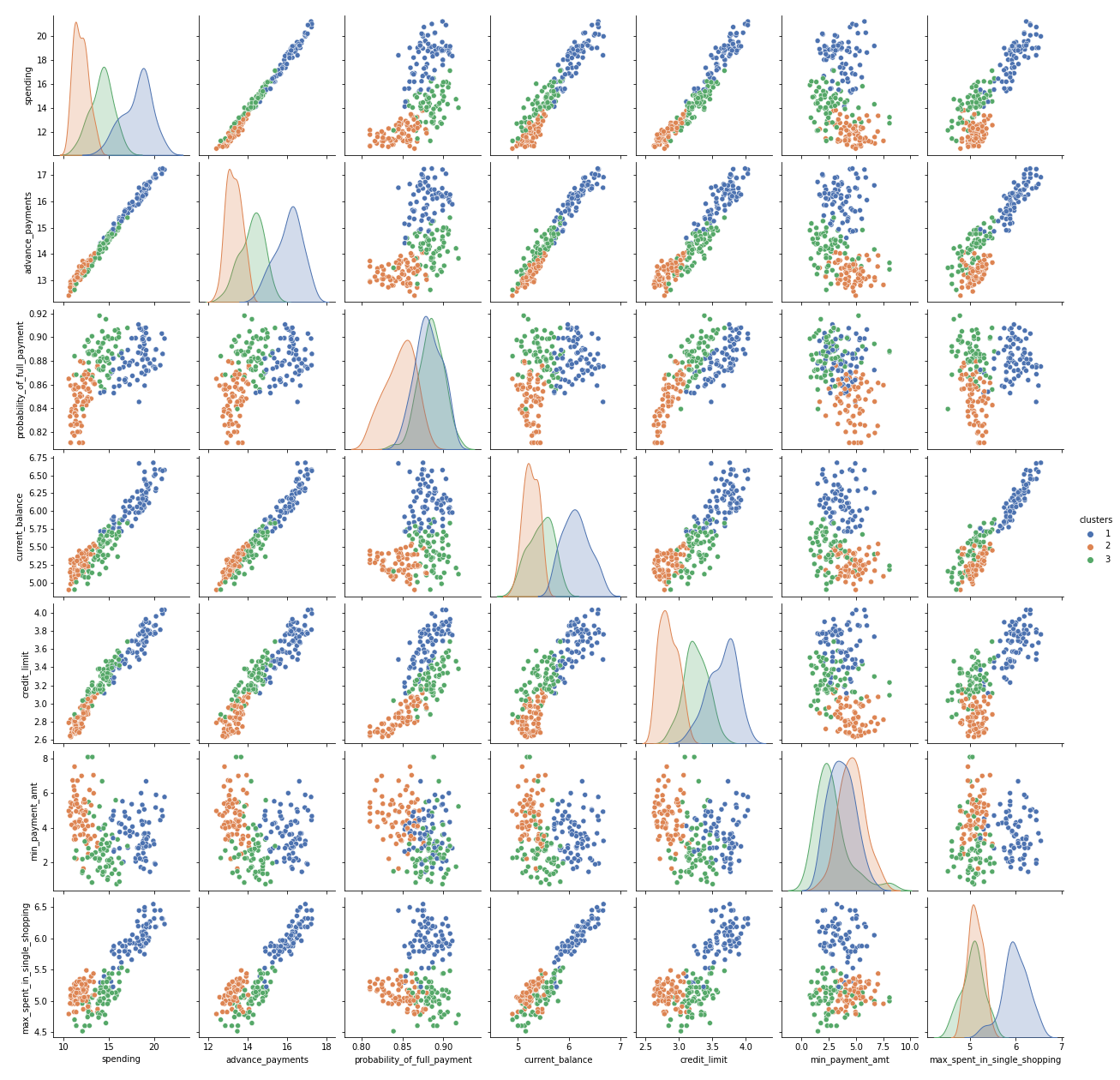
* Below is the frequency of the three clusters:



* It means that the first cluster has 75 customers, second cluster has 70 customers and third cluster has 65 customers.

**Cluster profiling:**



****

**Fig 1.10-Cluster profile Hierarchical**

**Cluster 1: High value customers**

* They are highest spending customers.
* They have higher probability of making payment in full.
* With 75 customers they are the topmost customers in the bank.
* They spend the highest in single shopping.
* They have less minimum payment amount.

**Cluster3: Medium level customers**

1. Among all the customers, medium level customers have higher chance to make payments in full. They may have good credit score.
2. They are moderate spenders.
3. There are 65 customers in this category.
4. They have less minimum payment amount.

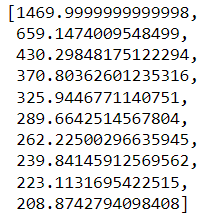
**Cluster 2: Economic customers**

* There are 70 customers in this category.
* They are low spending customers.
* They have lesser chance of making payments in full comparatively.
* Their credit limit is small.
* They have more minimum payment amount.

**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.**

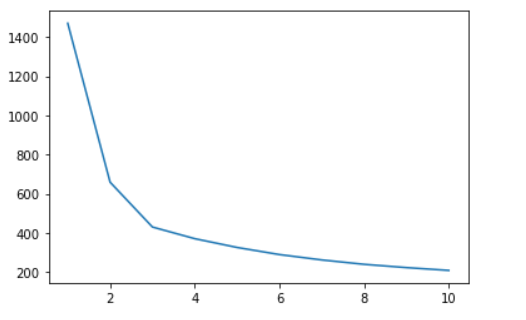
**K-Means clustering:**

* It is non-hierarchical approach to form good clusters by pre-specifying the desired number of clusters, k.
* The ‘means’ in the K-means refers to averaging of the data, that is, finding the centroid.
* K inertia tells the total within sum of squares for the given clusters. Below is the k inertia values for clusters from 1 to 10.



**Elbow method:**

* ***Elbow method***  is used to find the optimal number of clusters by plotting within sum of squares(WSS) as a graph.
* The value of k is chosen to be optimum, where addition of one more cluster does not lower the value of total WCSS appreciably.
* X-axis indicate the number of clusters and y-axis indicate the WSS.



**Fig 1.11-Elbow method**

The above elbow plot does not indicate any clear break in the elbow after k=3. Hence, we can take optimum number of clusters as 3.

**Silhouette score**

* Silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
* 1: Means clusters are well apart from each other and clearly distinguished.
* 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
* -1: Means clusters are assigned in the wrong way.

silhouette score when cluster is 4 is equal to 0.32943733699973826

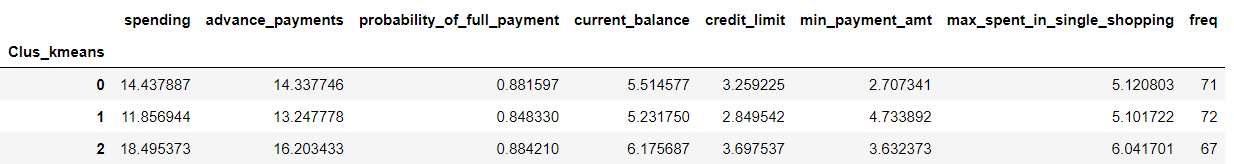
silhouette score when cluster is 3 is equal to 0.4008059221522216

So, **three** is the optimal number of clusters.

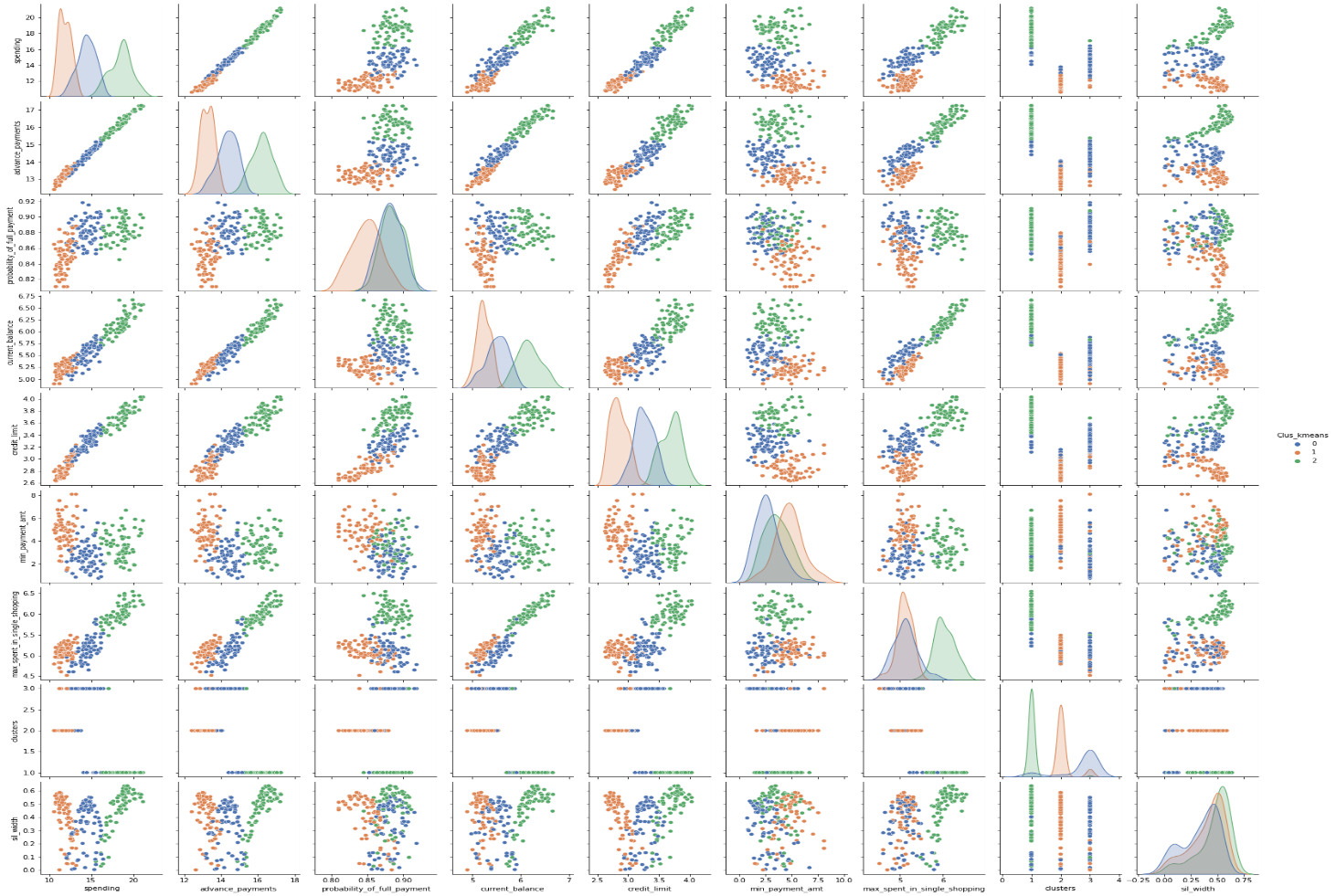
Below is the frequencies of the number of clusters



**Cluster Profiling**



**Table-1.4-Cluster profile kmeans**

****

Cluster 0: Medium level customers.

Cluster 1: Economic customers.

Cluster 2: High level customers.

**Cluster 2: High value customers**

* They are highest spending customers.
* They have higher probability of making payment in full.
* There are 67 customers
* They spend the highest in single shopping.
* They have less minimum payment amount.

**Cluster0: Medium level customers**

* Medium level customers have higher chance to make payments in full. They may have good credit score.
* There are 71 customers in this category.

**Cluster 1: Economic customers**

* There are 72 customers in this category.
* They are low spending customers.
* They have lesser chance of making payments in full comparatively.
* Their credit limit is small.

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

Cluster profiles-Already covered

**Recommendations**

**High level customers**:

1. As the probability of them making advance payments is higher, they are comparatively low risk customers. So, bank can bring offers targeting them such as unlimited airport longue access, discounted business class tickets.
2. By increasing the base of these customers, the bank can earn more revenue.
3. As they spend largest in single shopping, bank can tie-up with companies to offer membership for clubs, holiday resorts.

**Medium level Customers:**

1. Bank can increase their credit limit as they have higher probability of making payment in full.
2. Bank can make efforts to move them to higher spending category.
3. Bank can incentivise these customers by offering fuel cards and many others.

**Economic Customers:**

1. As, they are small spenders, bank can provide cashback offers, vouchers to increase their spending.
2. To increase their spending on daily basis, it can tie up with dining, food delivery apps.
3. They make higher minimum payment amount so their current debt will be carried forward. So, it is riskier.
4. As they pay minimum payment amount more often, they have higher chances of making defaults comparatively ,so banks should send them reminders often.

**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, Bi-variate, and multivariate analysis).**

**Data Dictionary:**

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)

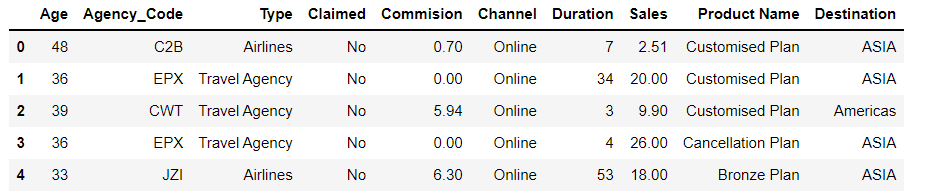
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)

9. The commission received for tour insurance firm (Commission is in percentage of sales)

10.Age of insured (Age).

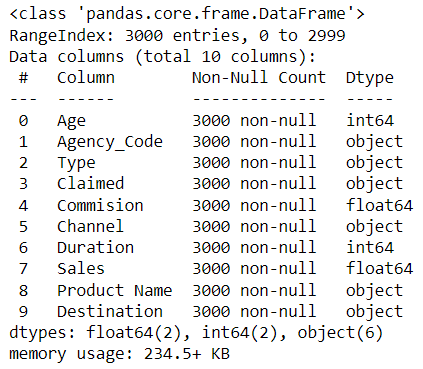
**Inference from the dataset:**

The top 5 rows of the dataset are as below



**Data pre-processing:**

* The dataset has 3000 rows and 10 columns.
* The dataset has 6 object datatypes,2 Integer datatypes and 2 float datatypes.



**Null values:**

There are no null values in the dataset

**Five point summary:**

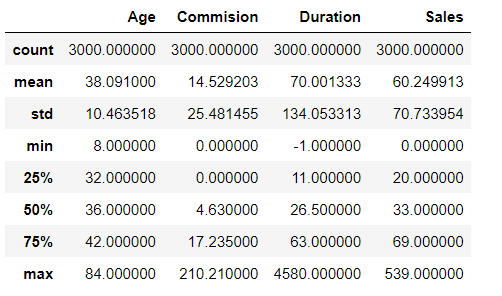


Table 2.1-Five point summary

* The commission received for tour insurance firm (Commission is in percentage of sales) varies from 0 to 210 percentage of sales.
* The age of insured varies from 8 to 84 years old.
* Duration of the tour (Duration in days) ranges from -1 to 4580 days.
* Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s) varies from 0 rupees to 53,900 rupees.

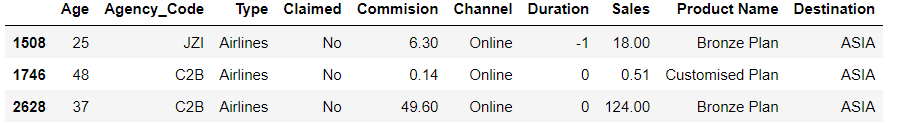
**Data preparation:**

**a) Duplicates:**

* There are **139 duplicates** in the dataset.
* The possibility of having duplicates is less as it has 10 variables.
* As the dataset is large, removal of 139 duplicates will have less impact. So, it is removed.

**b) Treating bad data:**

Duration variable has 3 samples with less than 1 days. It is impossible as it takes at least a day to make a trip.



So, it is replaced with median value.

**c)Outlier treatment:**

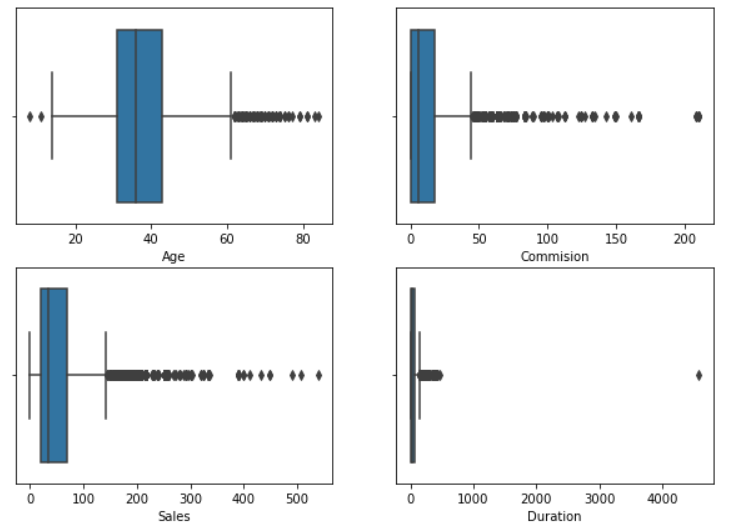


Fig 2.1-Boxplot

* Decision Trees and random forest are insensitive to noisy data or outliers since, extreme values or outliers never cause much reduction in **Residual Sum of Squares(RSS),** because they are never involved in the split.
* It will have little impact on ANN.
* Also, to avoid generalization outliers are left undisturbed.

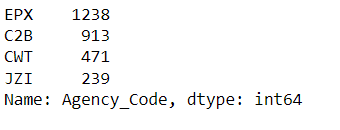
Exploratory Data Analysis:

**a) Univariate and bi-variate analysis:**

**Categorical variables:**

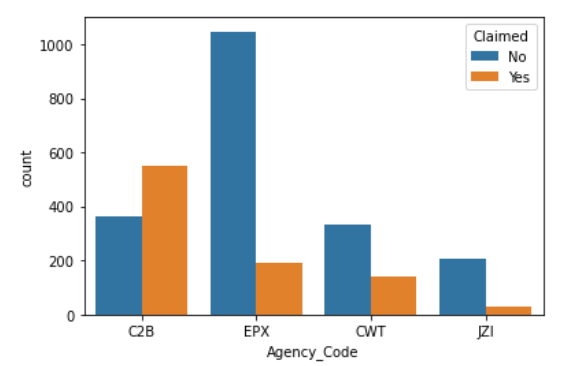
**1) Agency code:**

* The four Codes of the tour firm are EPX, C2B, CWT and JZI.
* The highest and lowest customers are in EPX and JZI agency codes respectively.
* Claim rate is highest at ‘**C2B’ agency**.
* The EPX agency has less claim rate despite higher sales.





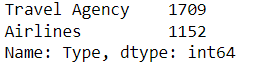
**Fig-2.2-Count plot Agency**



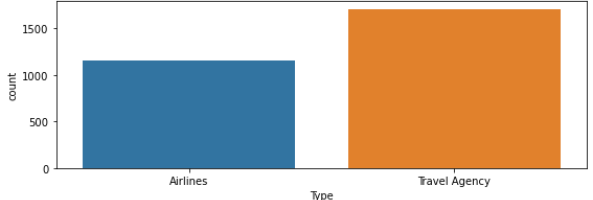
**Fig-2.3-count plot agency target**

2) **Type**:

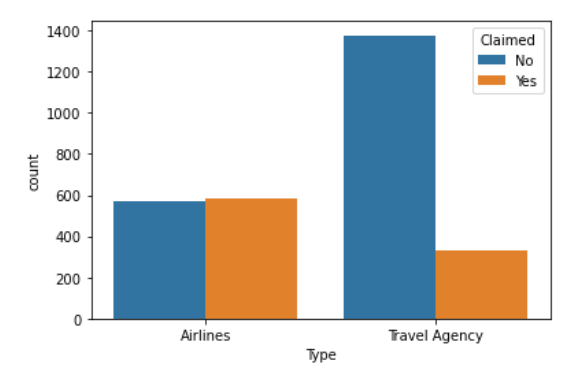
* There are two types of tour insurance firms (Type).
* They are Travel Agency and Airlines.
* The number of customers from these two types are given as below:



* The claim proportion is highest among airlines.
* The number of customers is highest at travel agency.

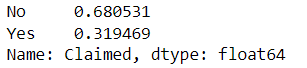
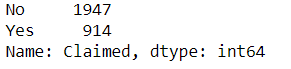


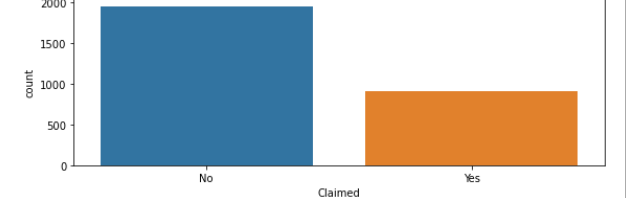
**Fig2.4-count plot type**



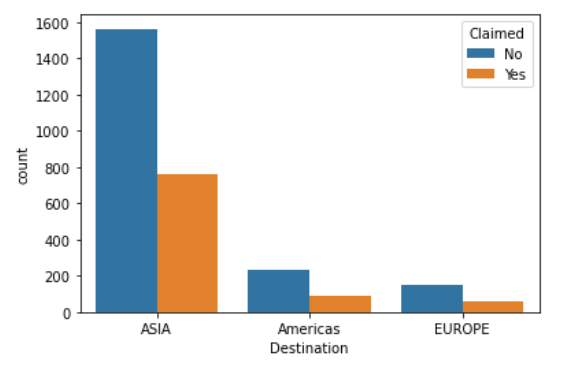
**Fig-2.5-count plot type target**

**3)Claimed:**





**Fig 2.6-Count plot claimed**

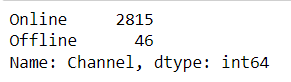


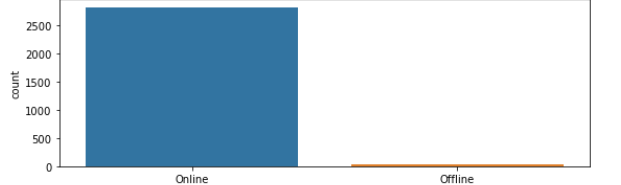
**Fig 2.7-Count plot claimed target**

* The target variable 'claimed' tells the number of persons who availed the insurance.
* It tells 914 out of 2861 customers claimed insurance benefits which is nearly 32 percent of insurers.
* Claim rate is more at ASIA destination customers.

**4)Channel**

* The number of customers from online and offline are as below

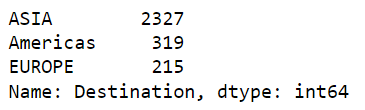


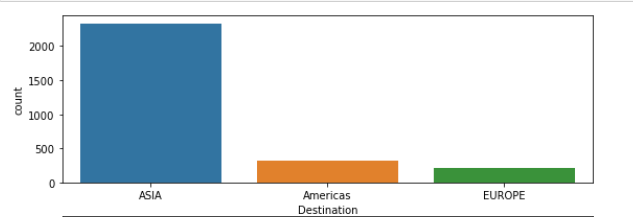


**Fig 2.8-Count plot channel**

* Almost 98 percent of the tour insurance agencies has distribution channel through ONLINE mode.

**5)Destination:**

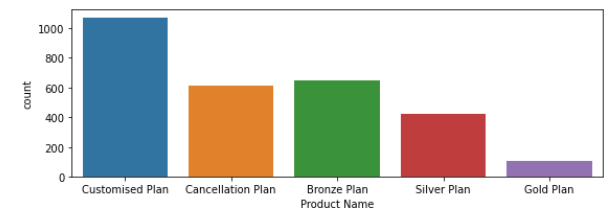




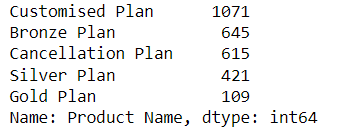
**Fig 2.9-Count plot Destination**

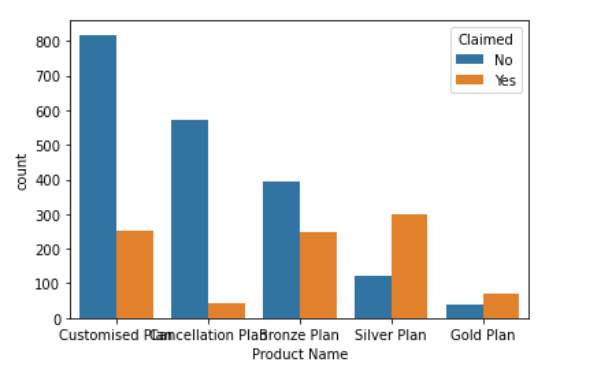
* The dataset contains three set of Destination of the tour namely ASIA, Americas and EUROPE.
* Most preferred destination is ASIA.

**6)Product Name:**



**Fig 2.10-Count plot product name**



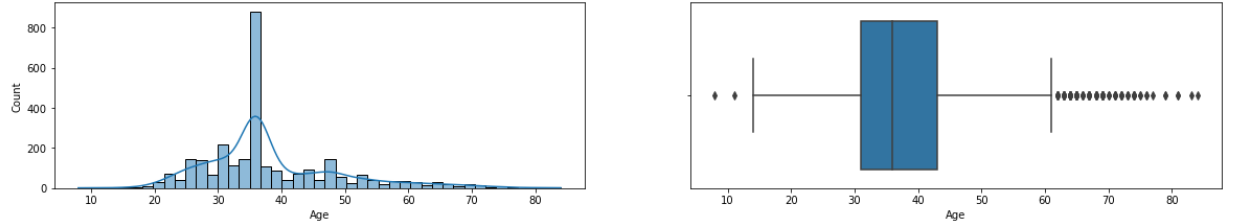


**Fig 2.11-Count plot product name**

* There are four variants of tour insurance products namely Customised Plan, Cancellation Plan, Bronze Plan, Silver Plan and Gold Plan.
* Out of this, Customised plan is widely preferred.
* Claim rate is more at silver and gold plan customers.

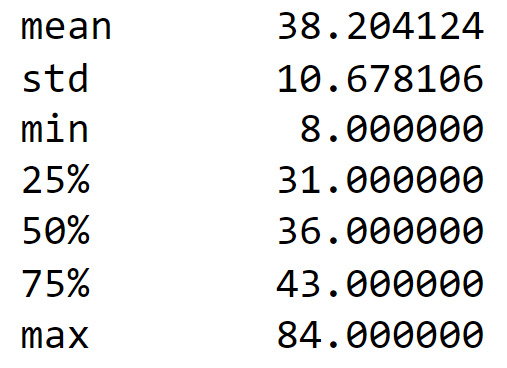
**Numerical variables**:

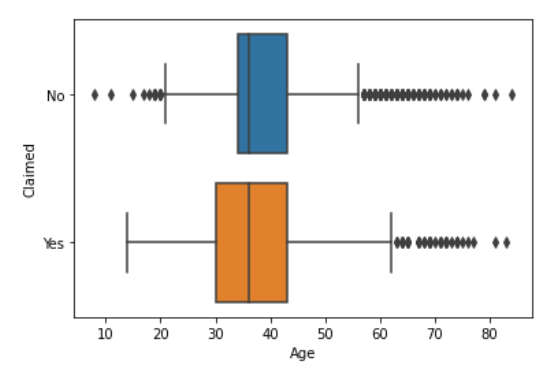
**7)Age:**



**Fig-2.12-Plot for Age**

* The five-point summary of the variable ‘Age’ is as below

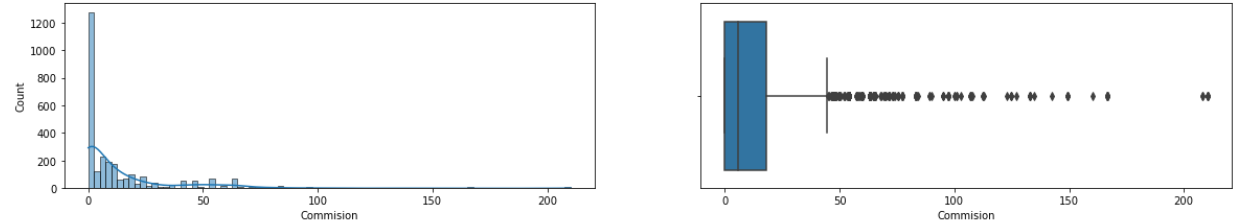




**Fig-2.13-BoxPlot for Age**

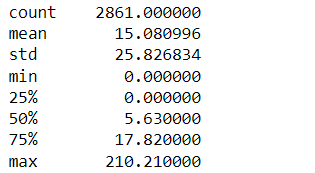
* Most of the customers were within 45 years of age.
* 50 percent of the claimant ranges between 30-43 years.
* There are many outliers in the dataset**.**

**8)Commission:**



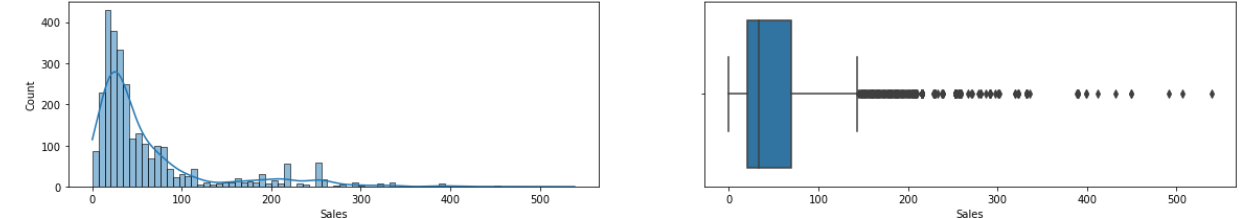
**Fig-2.14-Plot for commission**

* The five-point summary of the variable ‘commission’ is as below



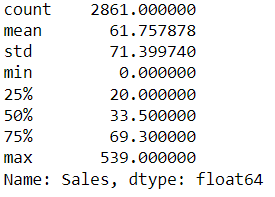
* There are few erroneous data as the commission charged is more than that of the sales amount.

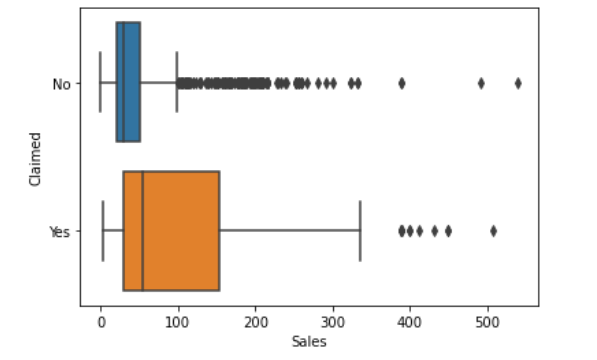
**9)Sales:**



**Fig-2.15-Plot for sales**

* The five-point summary of the variable ‘sales’ is as below

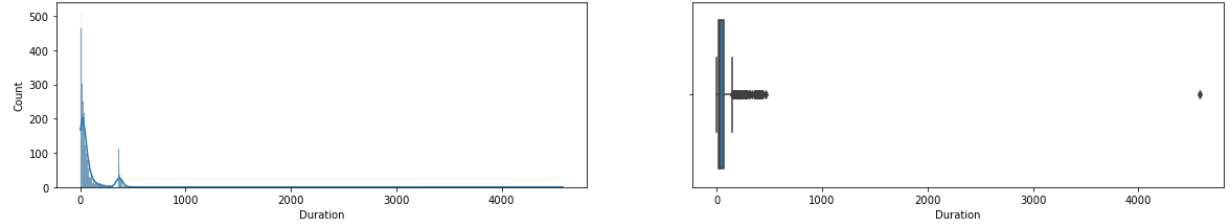




**Fig-2.16-Plot for sales claim**

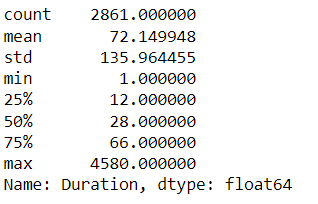
* The claim rate is high among the high value customers.
* The sales is right skewed.

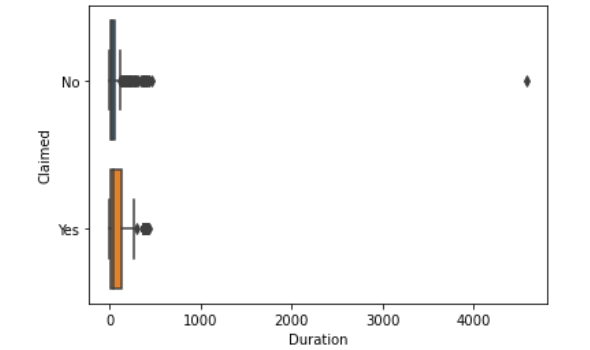
**10)Duration:**



**Fig-2.17-Plot for Duration**

* The five point summary of the variable ‘Duration’ is as below

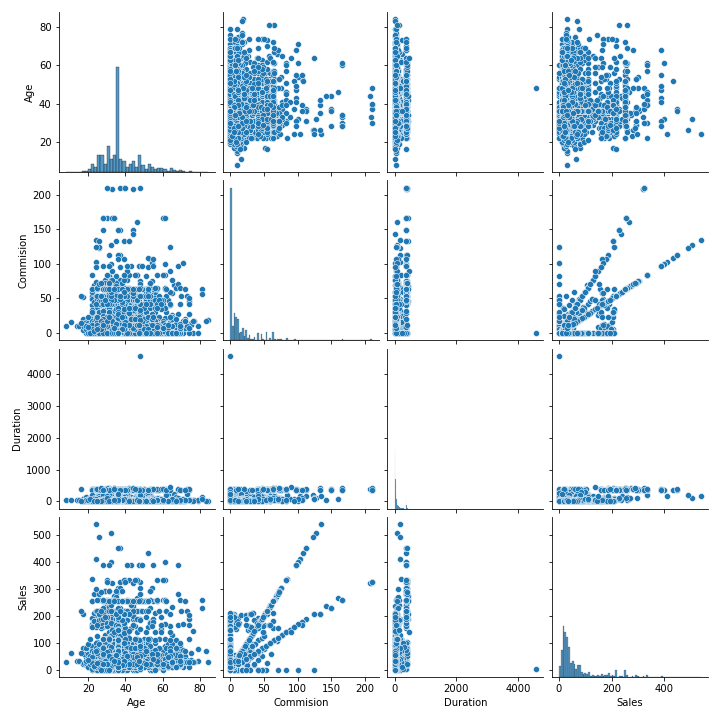




**Fig-2.18-Plot for duration claim**

* There are outliers in the dataset.

**Multivariate analysis:**

****

**Fig-2.19-Pairplot**

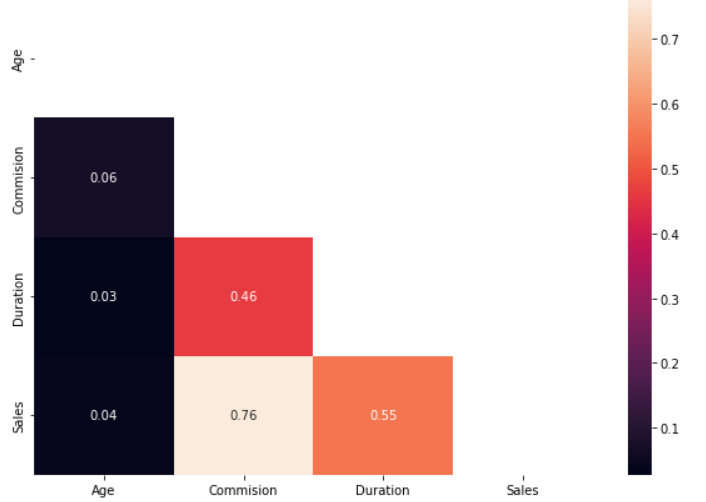


Fig-2.20-Correlation plot

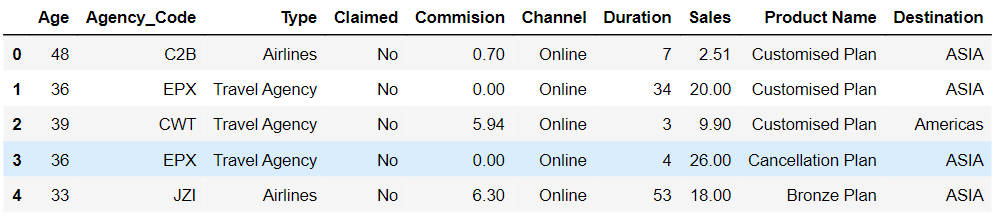
Moderate positive correlation exists between

* Commission and Sales
* Duration and Commission.
* Duration and sales.

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

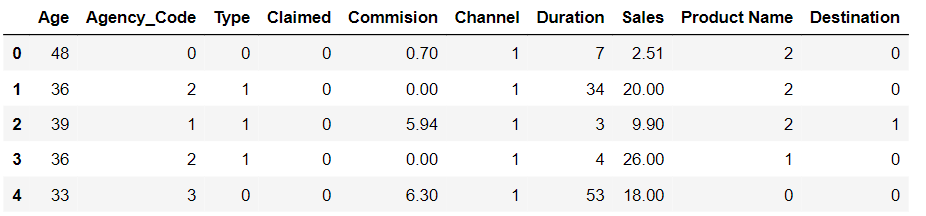
**Encoding:**

* For model building, it is necessary to convert the object datatype into numerical or categorical datatypes for the Algorithm to understand better.
* Our dataset contains **six object** datatypes which need to be converted.



**Table 2.2-Before Encoding**

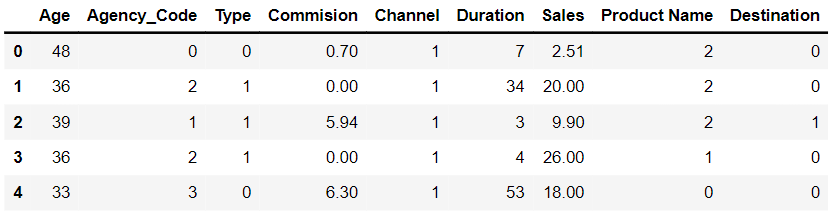
After conversion the first five rows of the dataset will look as below



**Table 2.3-After Encoding**

**Splitting the dataset:**

* Dataset is split into independent variables and dependent variable (Target variable).
* Independent variable-Includes all variables except ‘Claimed’ variable.



**Table 2.4-Independent variable**

* Dependent variable-It includes only the target variable ‘Claimed’.
* Then, it is split into training data and testing data.
* Training set-To train the model.
* Testing set-To evaluate whether the model can generalise well to new, unseen data. For instance, if the training accuracy is extremely high while the testing accuracy is poor then this is a good indicator that the model is probably overfitted.
* Testing data- 30 percent of the total observations is considered as optimal level.
* The dataset is split as such that training set contains 2002 observations and test data contains 859 variables.

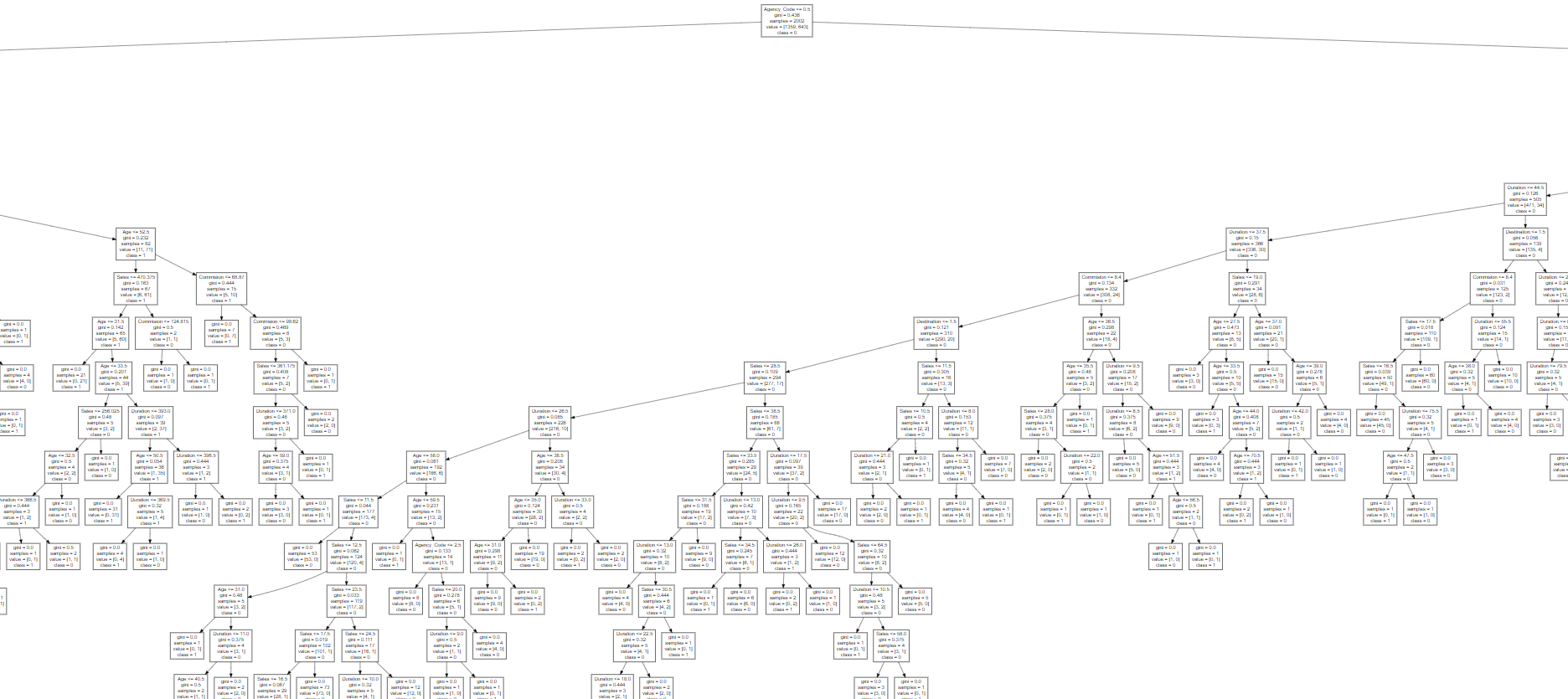
**Decision Tree Classifier**

**Scaling:**

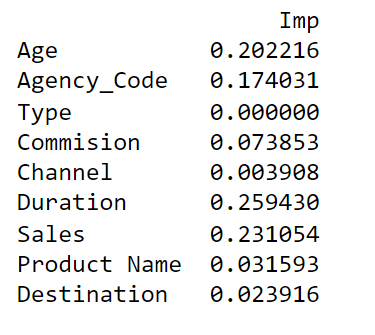
* Both decision Tree and Random Forest algorithms are scale-invariant.
* It is because it isn’t a distance-based algorithm and it just make a cut at a certain level.

**Building the model:**

* The model is built by using scikit learn by using ‘gini’ as criterion.
* Then the training data is fitted into the model.
* By creating a dot file, we are able to get the image of the decision tree.



* In building the model tree, the importance given to the variables is given below:

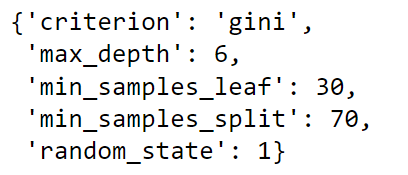


* From this, we can interpret that the ‘Duration’ variable is given utmost importance for splitting. The ‘Type’ variable was never considered for splitting the decision tree.
* Then, pruning(Cutting the decision tree) is done at optimal level by analysing the decision tree.
* By using grid search cross validation, we determine the acceptable level of parameters for the pruning process.

**Parameters:**

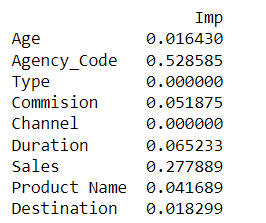
* Depth-It tells that the number of levels needed. It can predicted from the graph of decision tree generated.
* min\_samples\_leaf- It ensures that every terminal or leaf node has atleast ‘n’ number of samples.It is usually 1-3 percent of the total observation.
* min\_samples\_split-It tells the condition to split the node such that the split should happen only if it contains ‘n’ samples. It is taken as three times as of the min\_samples\_leaf.
* Criterion as ‘gini’

From running the grid sample, we get the best parameters for the model as below.



* Then the model is fitted with the training data.
* The model is used to predict for training data.

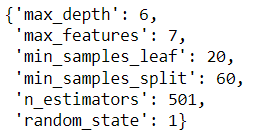
**Feature importance** taken for decision tree:



The variable **‘agency\_code’** is given utmost importance for building the model.

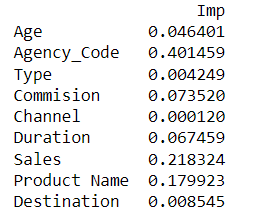
**Random forest classifier:**

* The model is built by using scikit learn.
* By using grid search, best value for the parameters are determined. Apart from parameters used in decision tree, new parameters such as n estimators which denotes the number of trees are included.
* The obtained best parameters are



* Then the training data is fitted into the model.

**Feature importance** given for random forest model building



In this model also, agency\_code is given utmost importance for building the model.

* The model is now used to predict the outcome for both testing and training data.

**Artificial Neural networks:**

**Scaling:**

As ANN is affected by variation in the dataset, the data is scaled using standard scaler.

We do it after splitting because the test data should not alter the mean and the standard deviation of the training data.

**Model building:**

* To determine the best value for the parameters. Grid search method is used.
* The values of the parameter are



1. Hidden\_layer\_sizes-It means creating a neural network with 100 neurons.
2. Max\_iter-It tells the number of iterations to update the values.
3. Solver- It is either Adam or sgd. Method to optimize the neural network.
4. Tol-Threshold level to iterate until we get the desired weights.

* Then the training data is fitted into the model and predicted.

**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.**

**ROC(Receiver operating characteristics)-AUC(Area under the curve) curve:**

**a)Decision Tree:**

ROC is used to visualize the model performance.

X-axis denotes False positivity rate and y- axis denotes the True positive rate.

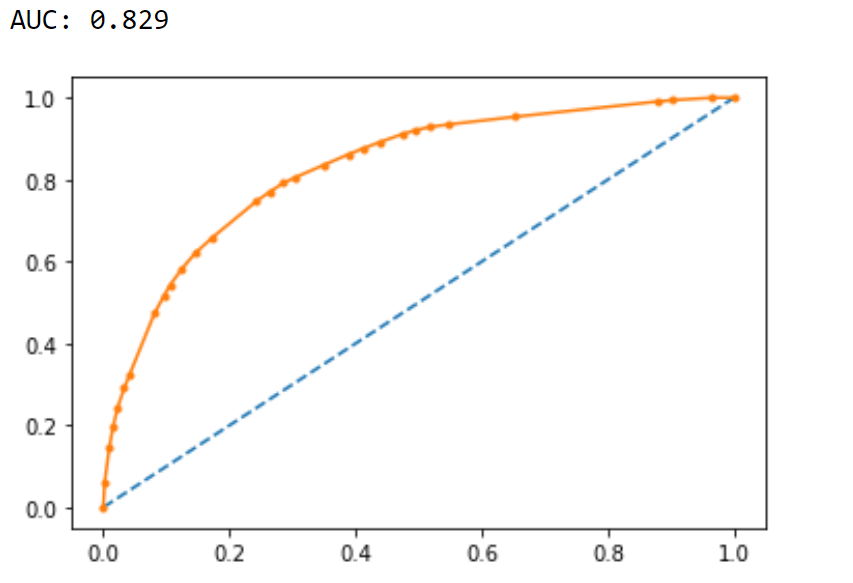
It helps us visualize how well our machine learning classifier is performing

AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. More area under the curve, better the model.

When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.

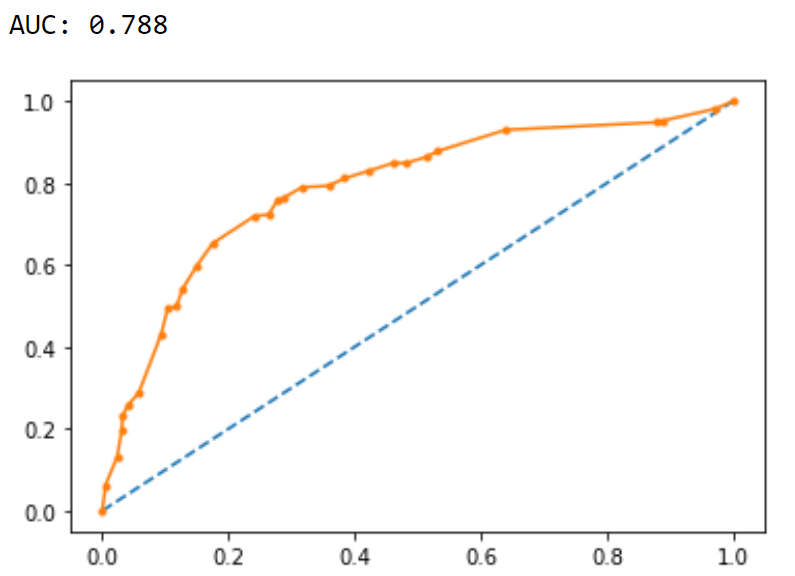
When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.

**The AUC for training data is 0.82**



**Fig 2.21-ROC -Decision tree-Train**

**AUC for testing data is 0.78.**

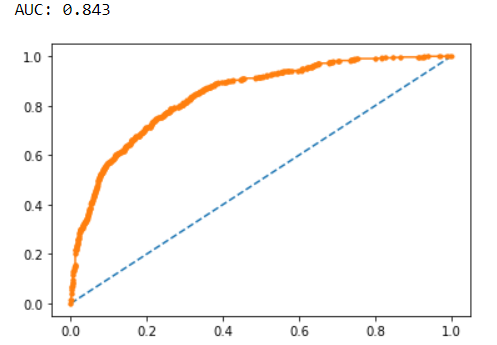


**Fig-2.22-ROC-Decision tree-Test**

**b)Random forest:**

Training data:

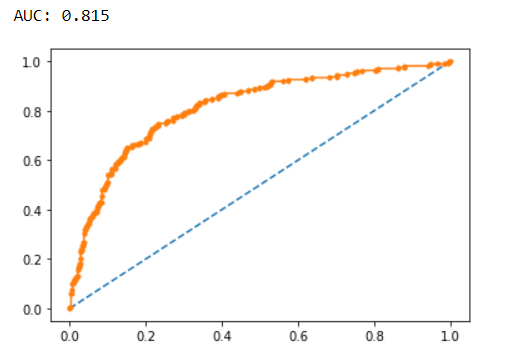
**The AUC for training data is 0.843**



**Fig-2.23-ROC-Random forest Train**

Testing data:

**The AUC for testing data is 0.815**

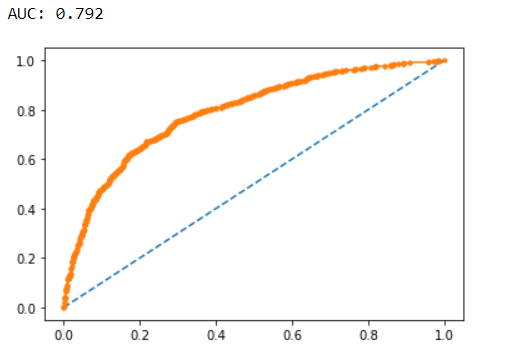


**Fig-2.24-ROC-Random forest Test**

**c)Artificial Neural Network:**

Training data:

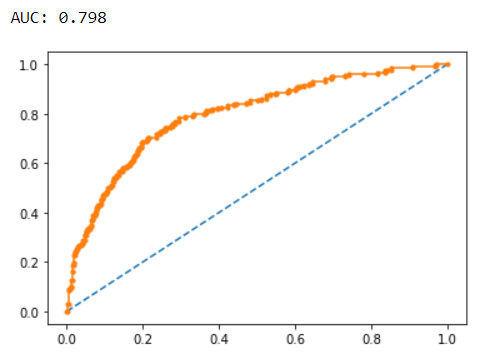
**The AUC score for training data is 0.792**



**Fig-2.25-ROC-ANN- Train**

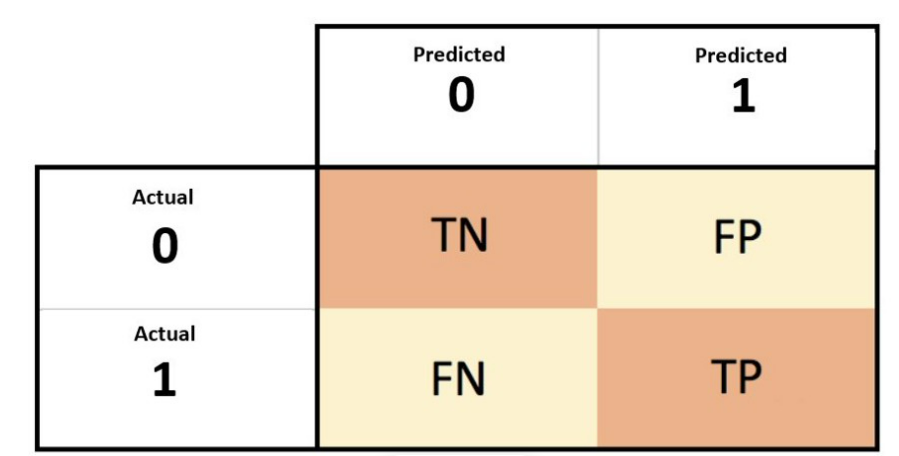
Test data:

**The AUC score for testing data is 0.798**



**Fig-2.26-ROC-ANN Test**

**Confusion matrix:**



True Negative (TN)-Actually not claimed and also predicted as not claimed.

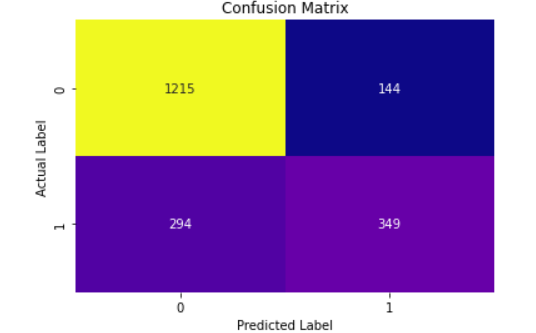
False Positive (FP)- Actually not claimed and but predicted as claimed.

False negative (FN)- Actually claimed and but predicted as not claimed.

True positive (TP)- Actually claimed and also predicted as claimed.

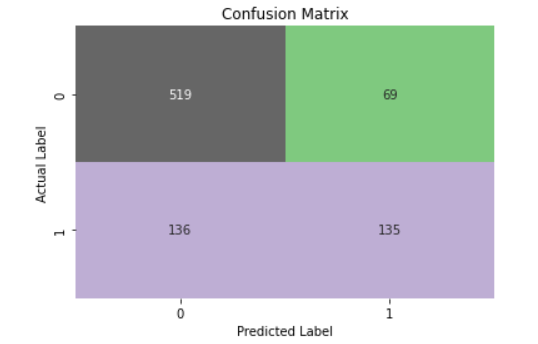
**a)Decision tree:**

**For training data:**



**Fig-2.27-Decision tree confusion matrix- Train**

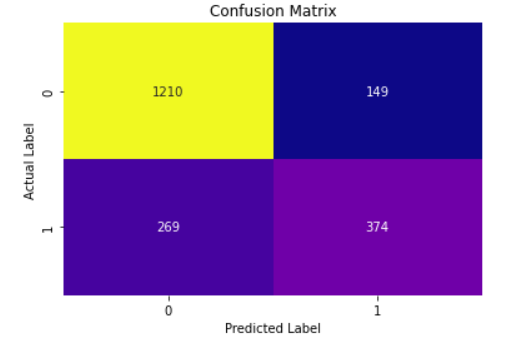
**For testing data:**



**Fig-2.28-Decision tree confusion matrix- Test**

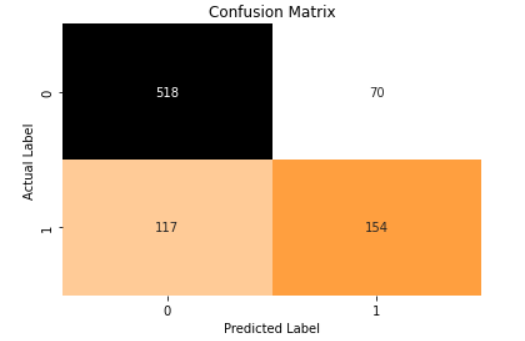
**b)Random Forest**

**Training data:**



**Fig-2.29-Random forest- confusion matrix- Train**

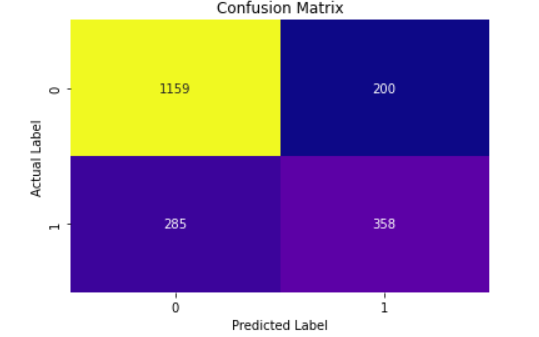
**Testing data:**



**Fig-2.30-Random forest- confusion matrix- Test**

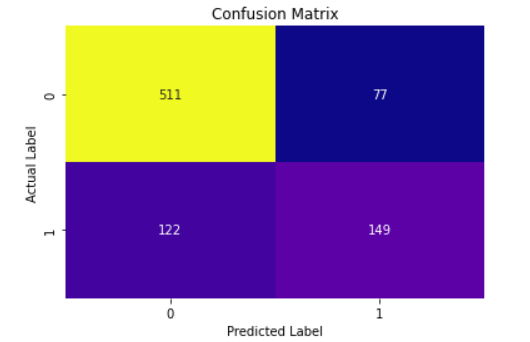
**C)Artificial Neural Networks:**

**Training Data**:



**Fig-2.30-ANN- confusion matrix- TRAIN**

**Testing Data:**

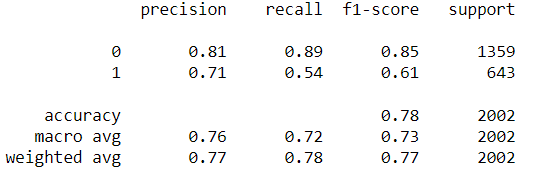


**Fig-2.30-ANN- confusion matrix- TEST**

**Classification report:**

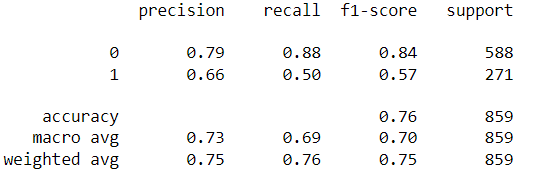
**a)Decision Tree**

**Training data:**



* The accuracy for training data is 0.78.
* The precision of the training data is 0.71
* The recall of the training data is 0.54.
* The f1 score of the training data is 0.61

**Testing data:**

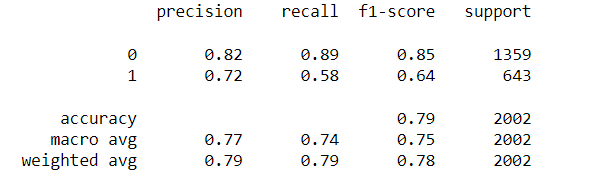


* The accuracy for testing data is 0.76.
* The precision of the testing data is 0.66
* The recall of the testing data is 0.50.
* The f1 score of the testing data is 0.57.

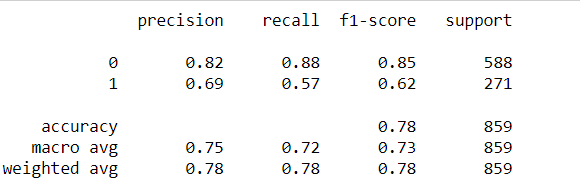
**b)Random Forest:**

**Classification report**:

**Training data:**

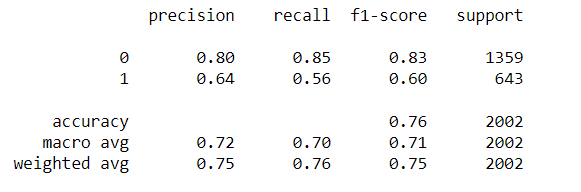


**Testing data:**

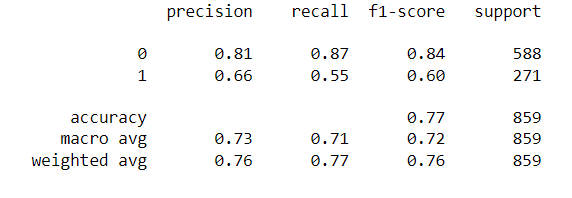


**c)Artificial Neural Networks:**

**Training data:**



**Test data:**



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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Decision tree (Train) | Decision tree (Test) | Random Forest (Train) | Random Forest (Test) | ANN(Train) | ANN(Test) |
| Accuracy | 0.78 | 0.76 | 0.79 | 0.78 | 0.76 | 0.77 |
| Recall | 0.54 | 0.50 | 0.58 | 0.57 | 0.56 | 0.55 |
| AUC | 0.829 | 0.788 | 0.84 | 0.81 | 0.792 | 0.798 |
| F1 score | 0.61 | 0.57 | 0.64 | 0.62 | 0.60 | 0.60 |
| Precision | 0.71 | 0.66 | 0.72 | 0.69 | 0.64 | 0.66 |

True Negative (TN)-Actually not claimed and also predicted as not claimed.

False Positive (FP)- Actually not claimed and but predicted as claimed.

False negative (FN)- Actually claimed and but predicted as not claimed.

True positive (TP)- Actually claimed and also predicted as claimed.

**Sensitivity/Recall:**

* In this data set, false negative (FN) is more important as it fails to identify the claimant. This error is most costly.
* As sensitivity/recall is being influenced by false negativity, it shall be our top most performance measure.
* Increase in sensitivity, decreases the false negativity.
* Our recall is quite moderate.

**Precision:**

* The second most important measure is Precision. It tells how sure one is of determining the true positive.
* That is, how precisely it finds the claimant.
* Here, the precision is good compared to recall.

**Accuracy:**

* It tells how accurately the model classifies the claimant and non-claimant.
* All model performs almost at the same level.
* As the data is imbalanced, accuracy is less important.

**F1 Score:**

* It is best when we need to take both recall and precision into account.
* F1 score is moderate for all the models.

**AUC**:

* The AUC score ranges well within the range of 0.5-1. It tells that the model classifies the claimant and non-claimant really well.

**Overall insights:**

* From all the models, we can say that our models have **high precision and low recall**. It means, if it predicts that a customer will claim insurance, you can trust it — it is positive. However, if it tells that the customer will not claim, you can’t trust it, the chances are that the customer can still claim.
* From the classification report, we can observe that the F1 score is moderate but on contrary the precision is high which could be because of the class imbalance as there are very few 1's.
* So, when taking into account all the performance metrics **random forest** outperforms the other models.
* The Overall model performance is moderate enough to start predicting if any new customer will claim or not.

**Recommendation**

a) silver plan and gold plan -high value customers category seems to have high claim rate. So, the reason for the same needs to be found out.

b) As the airlines type have more claim rate the reason has to be known.

c) **As the policies from Agency ‘C2B’ have highest claim rate, company should look for any irregularities. If found, strong action needs to be taken.**

d) As the age of 75 percent of the claimants are less than 43 years, the policies can be customized more effectively to suit them.

e) Here the metric recall is important considering the model has a recall rate of 0.57, so major focus can be upon improving the recall score which can provide some insights for the tour company to take proactive steps in analysing those customers who might claim and take necessary steps.