DATA MINING
PROJECT BUSINESS REPORT

Sanjana M
PGP-DSBA Online

Table of Contents

List of Figures 2

List of Tables 3

Problem 1 6

1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 6

1.2 Do you think scaling is necessary for clustering in this case? Justify 10

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them 12

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. 13

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. 15

Problem 2 17

2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 17

2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network 21

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 22

2.4 Final Model: Compare all the models and write an inference which model is best/optimized. 28

2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations 29

List of Figures

Figure 1: Histogram and boxplot of variables of datatset-1 ... 8
Figure 2: Visualization for categorical variables of datatset-1... 8
Figure 3: Heatmap for correlation of dataset-1 ... 9
Figure 4: Pairplot of dataset-1 ... 10
Figure 5: Dendrogram ... 12
Figure 6:Truncated dendrogram ... 12
Figure 7: Elbow curve ... 14
Figure 8: K-Means cluster value counts ... 14
Figure 9: Histogram and Boxplot graphs... 19
Figure 11: Pairplot of dataset-2 ... 20
Figure 10: Heatmap for correlation of dataset-2 ... 21
Figure 12: DTC-Import features ... 23
Figure 13: DTC – Training data ROC ... 23
Figure 14: DTC- Testing data ROC ... 23
Figure 15: DTC- Training data Classification report ... 23
Figure 16: DTC- Testing data Classification report... 24
Figure 17: DTC- Training data Confusion matrix ... 24
Figure 18: DTC- Testing data Confusion matrix ... 24
Figure 19: RF-Important features ... 24
Figure 20: RF- Training data ROC ... 25
Figure 21: RF-Testing data ROC... 25
Figure 22: RF-Training data Classification report... 25
Figure 23: RF-Testing data Classification report ... 25
Figure 24: RF-Training data Confusion matrix ... 26
Figure 25: RF-Testing data Confusion matrix... 26
Figure 26: NN-Training data Confusion matrix ... 26
Figure 27: NN-Testing data Confusion matrix... 26
Figure 28: NN-Training data Classification report... 27
Figure 29: NN-Testing data Classification report ... 27
Figure 30: NN-Training data ROC ... 27
Figure 31: NN-Testing data ROC ... 27
Figure 32: Training data ROC from the 3 models... 28
Figure 33: Testing data ROC from the 3 models ... 29

List of Tables

Table 1: Info of datatset-1 6
Table 2: First 5 entries of dataset-1 7
Table 3: Unscaled dataset-1.......... 11
Table 4: Five point summary of dataset-1 before scaling.......... 11
Table 5: Scaled dataset-1 11
Table 6: Five point summary of dataset-1t after scaling 11
Table 7: Dataset with cluster_kMeans and sil_width.......... 14
Table 8: Cluster data with average variables’ values.......... 15
Table 9: Dataset-1 info.......... 17
Table 10: Head of dataset-1.......... 18
Table 11: Dataset-2 categorical variables value counts and unique values 18
Table 12: Comparison of the performance metrics from the 3 models.......... 28

Problem 1

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.

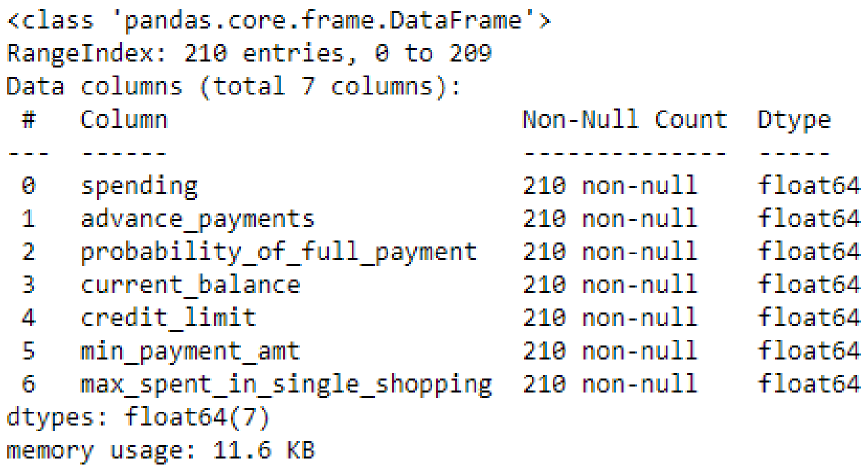
Dataset for Problem 1: bank_marketing_part1_Data.csv

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)

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

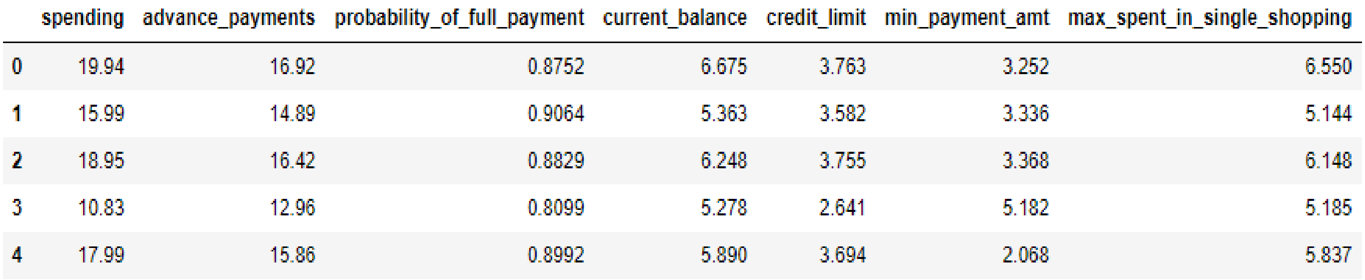
The data set has 210 rows and 7 columns.
The columns names are : 'spending', 'advance_payments', 'probability_of_full_payment', 'current_balance', 'credit_limit', 'min_payment_amt','max_spent_in_single_shopping'
There are no null or duplicate values in the dataset.
The datatype of all the variables is float.



```
<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
```

Table 1: Info of datatset-1

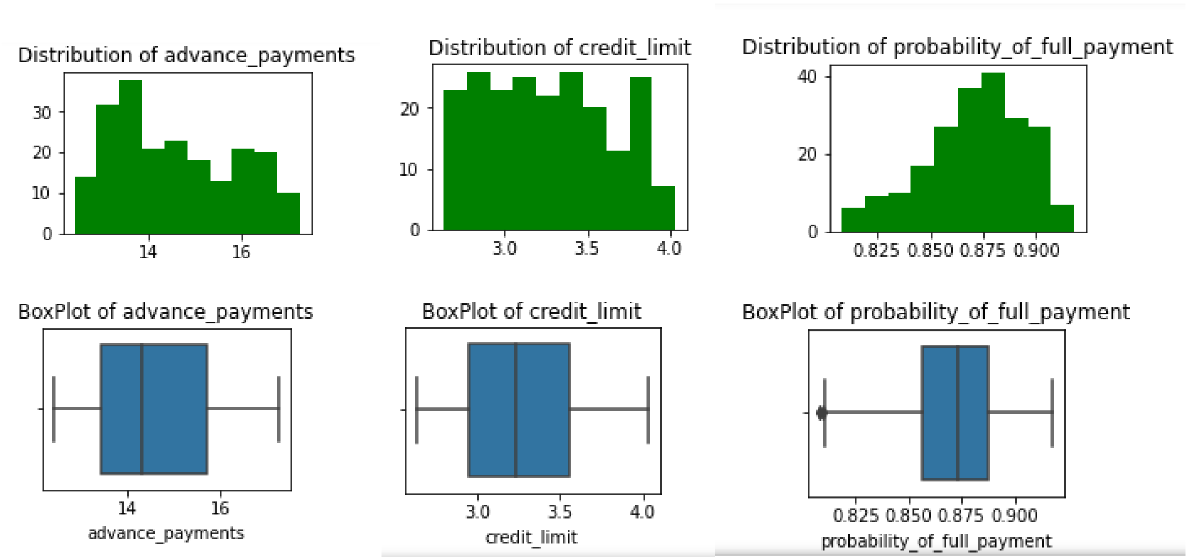
The first five entries of the dataset looks as below:

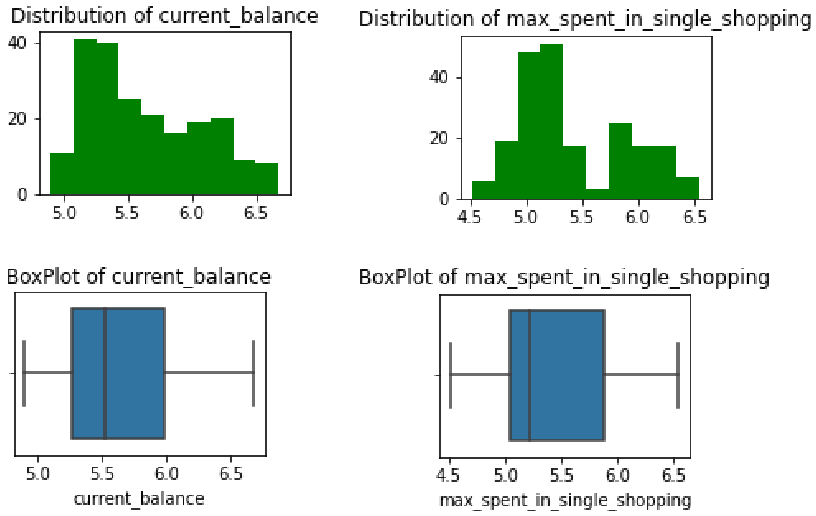


	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182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

Table 2: First 5 entries of dataset-1

Histogram and boxplot for each variable of the dataset is given below.
Probability_of_full_payment has outliers at lower end, min_payment_amount has outliers at higher end and others have no outliers.
Spending, Advance_payments, Current_balance, max_spent_in_single_shopping are right skewed.
Probability_of_full_payment and min_payment_amt left skewed





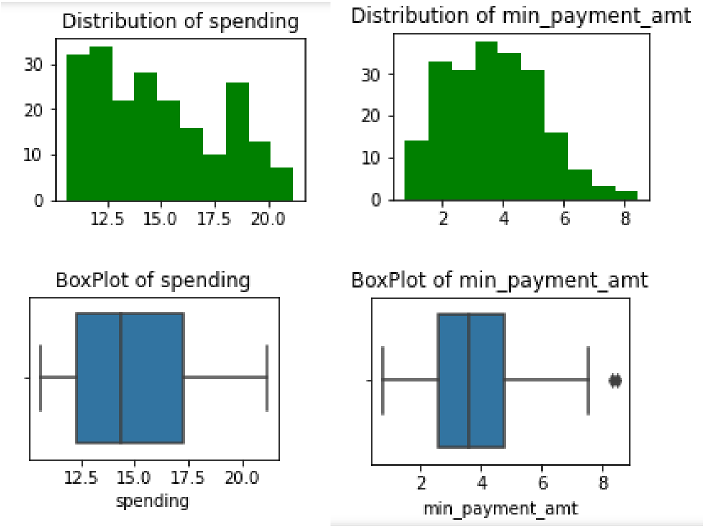


Figure 1: Histogram and boxplot of variables of datatset-1

Visualization for categorical variables:

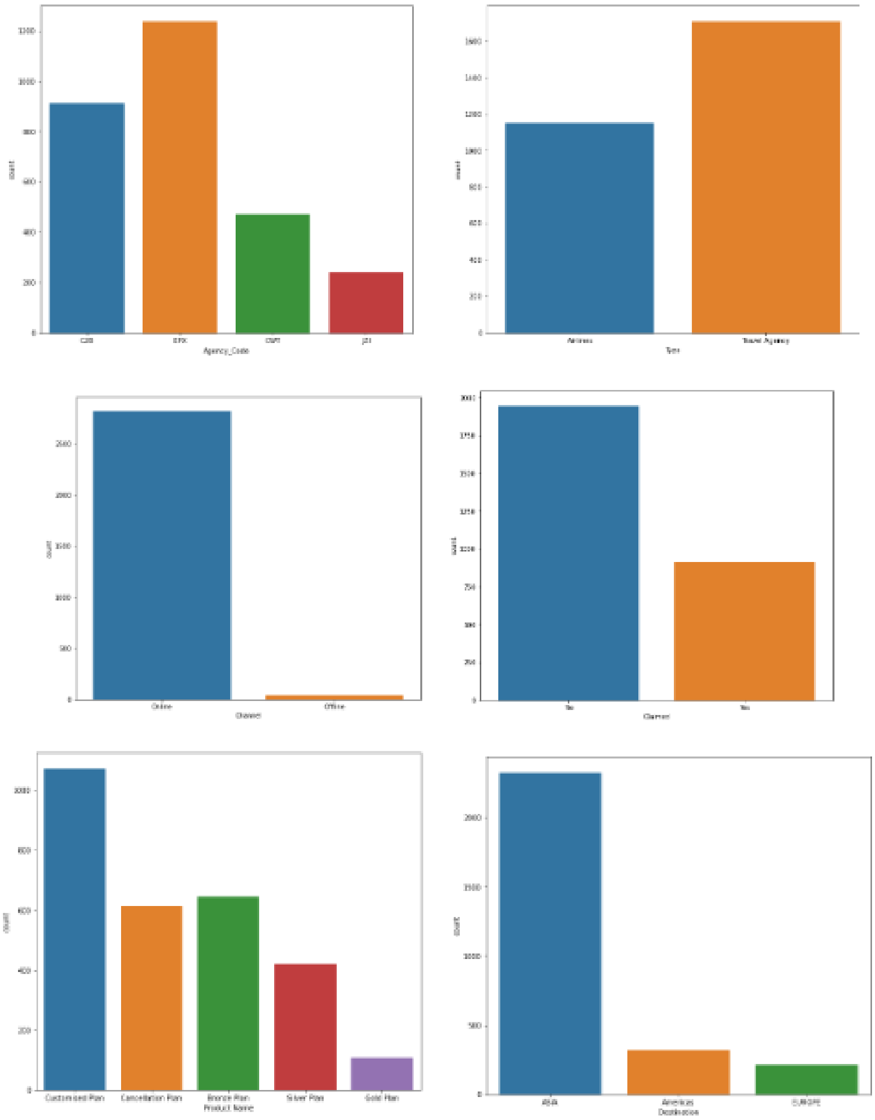


Figure 2: Visualization for categorical variables of datatset-1

Below is a heatmap, it is a type of a plot that visualizes the strength of relationships between numerical variables present in the dataset. A variable has highest correlation (=1) with itself.

For the given data Min_payment_amt has lowest correlation with other variables/factors. Spending and Advance payments, Spending and credit_limit, Current balance and advance payments have higest correlation.

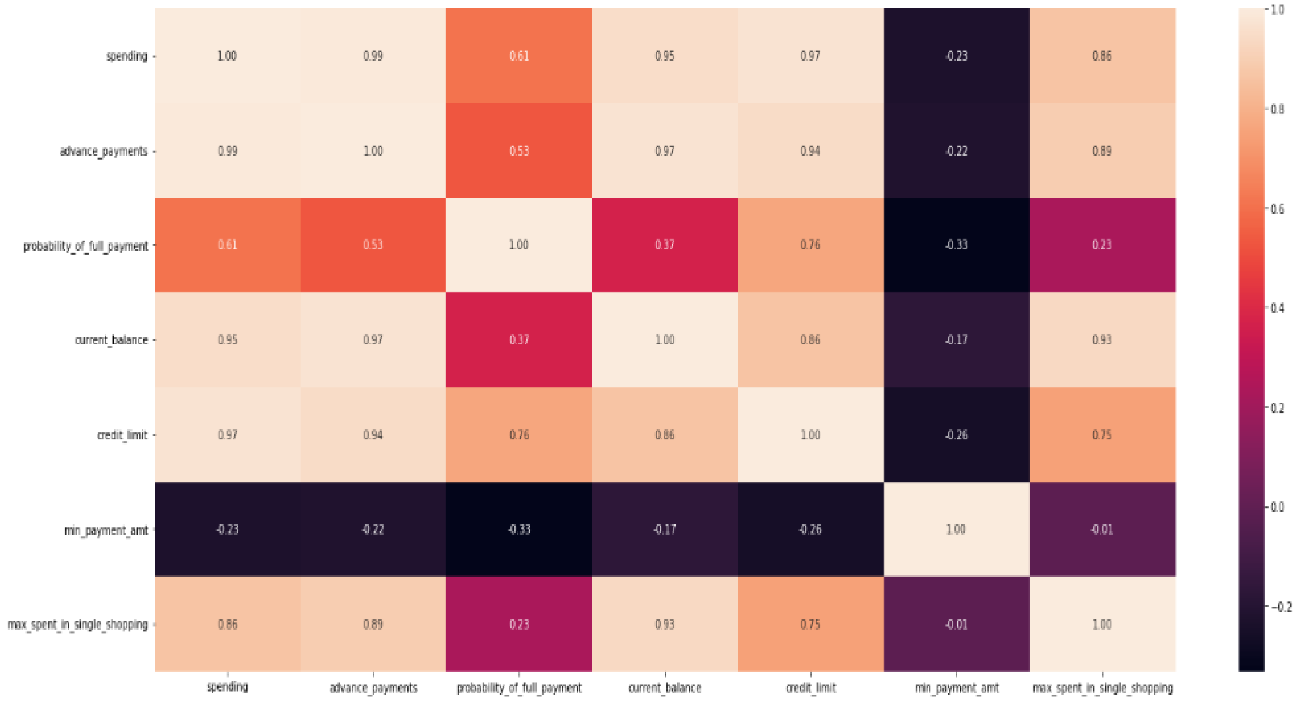


Figure 3: Heatmap for correlation of dataset-1

Observations from correlation Heat Map:
1. There is a strong correlation between Spending and Advance Payment.
2. There is a strong correlation between Current Balance and Spending.
3. There is a strong correlation between Current Balance and Advance Payments.
4. There is a strong correlation between credit limit and Spending
5. There is a strong correlation between credit limits and advance payments.
6. There is a strong correlation between credit limit and current balance.
7. There is a strong correlation between Maximum spending in single shopping and spending, advance payments, current balance

Give below is the pairplot of the dataset. A pairplot visualizes the relationship between 2 numeric variables.

From the pairplot we can see that Spending-advance_payment has one of the strongest positive relationships along with Spending-current_balance, Spending-credit_limit, advance_payment-credit_limit, advance_payment-current_balance, current_balance-max_spent_in_single_shopping.

Some of weaker relationships are seen between credit_limit-probability_of_full_payment, credit_limit-max_spent_in_single_shopping.

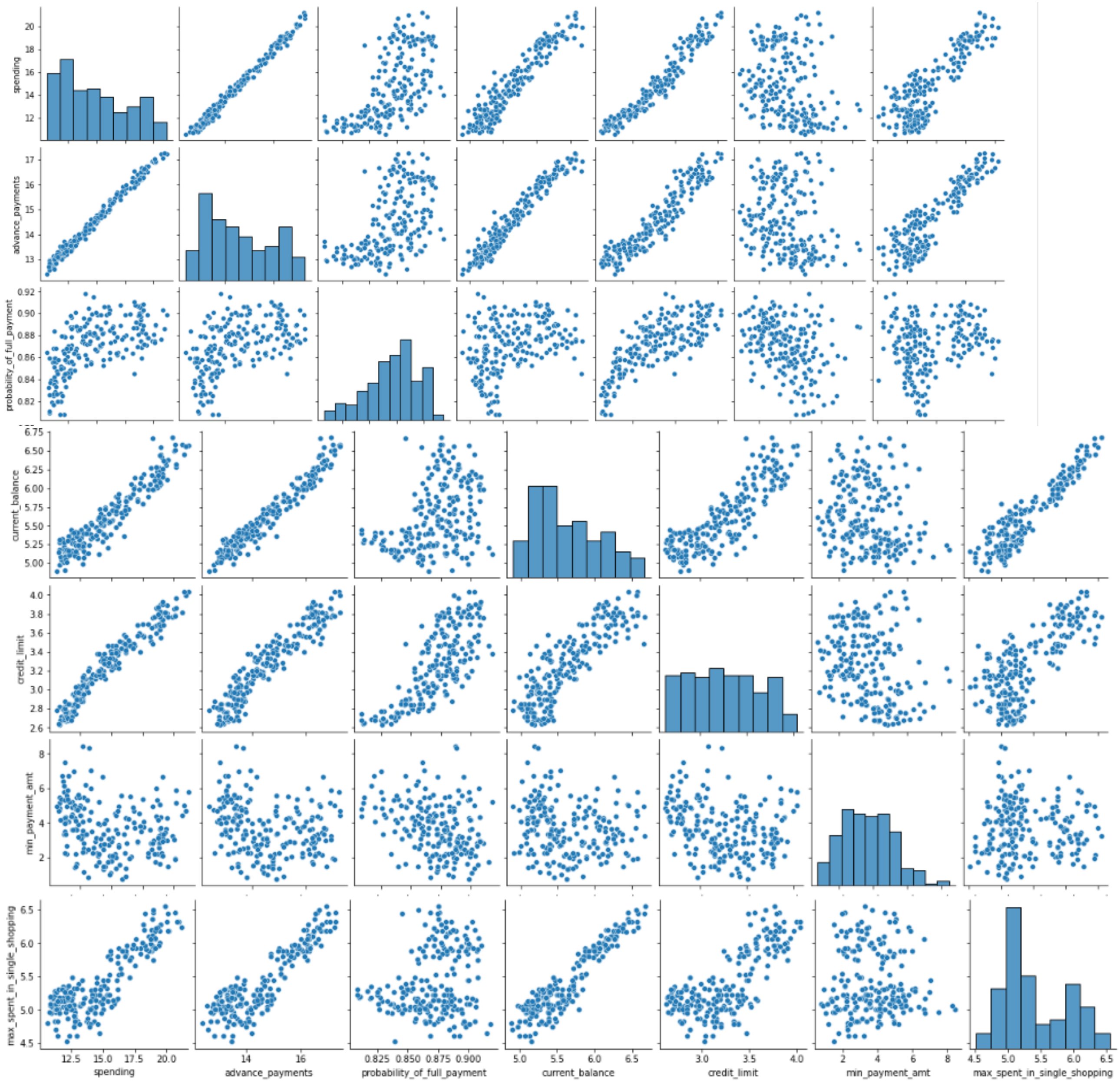


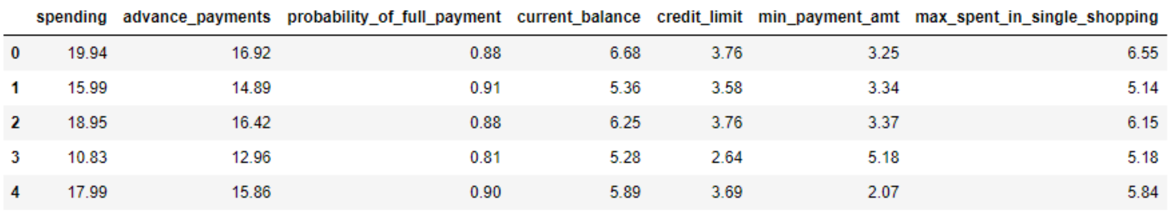
Figure 4: Pairplot of dataset-1

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

For the given dataset mean value of spending=14.85, adv payments=14.56, probability_of_full_payment=0.87, current_balance=5.63, credit_limit=3.26, min_payment_amt=3.70, max_spent_in_single_shopping=5.41.
Each of them are further in terms of 100s, 1000s, 10000s.
Scaling of the features is necessary for this case as not all the features in this dataset are of the same scale of measurement and the range of the features are not same.
Hence, it is necessary to scale the dataset.

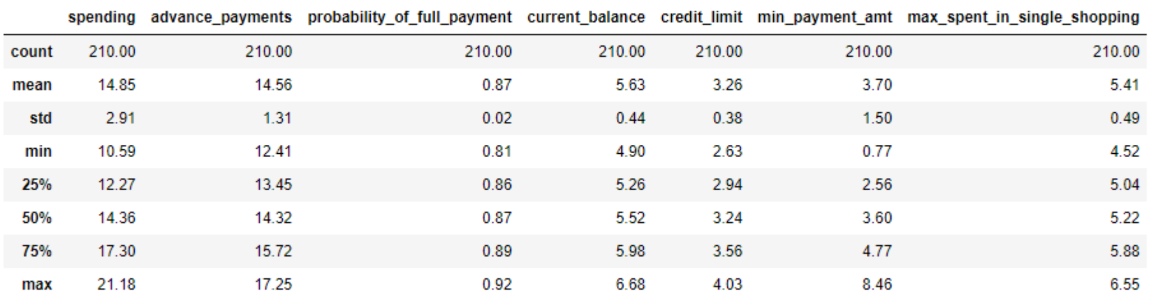
Scaling is done to make the data points closer to each other by reducing the distance. This will put the datapoints within a certain bracket and ease the processing and understanding of data. Scaling can be done using zscore from scipy package or StandardScaler from sklearn package can be used. Standard scaling makes the means of the data points=0, and standard deviation=1.

Data scaling is necessary before clustering as it controls the variability of the dataset, it converts data into specific range using a linear transformation which generate good quality clusters and improve accuracy of clustering algorithms



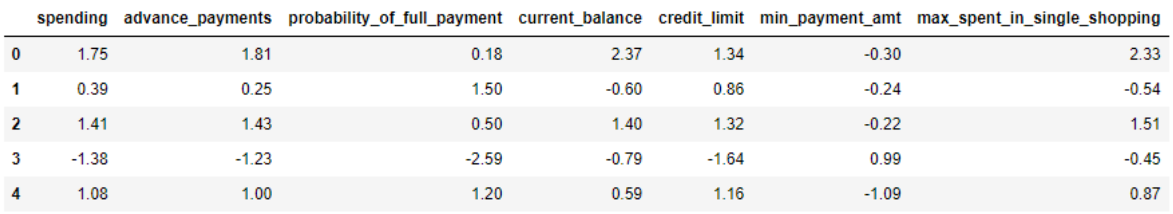
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.88	6.68	3.76	3.25	6.55
1	15.99	14.89	0.91	5.36	3.58	3.34	5.14
2	18.95	16.42	0.88	6.25	3.76	3.37	6.15
3	10.83	12.96	0.81	5.28	2.64	5.18	5.18
4	17.99	15.86	0.90	5.89	3.69	2.07	5.84

Table 3: Unscaled dataset-1



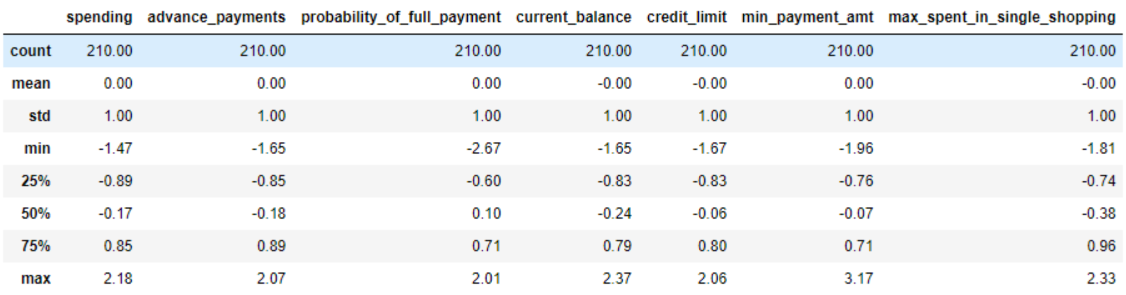
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.00	210.00	210.00	210.00	210.00	210.00	210.00
mean	14.85	14.56	0.87	5.63	3.26	3.70	5.41
std	2.91	1.31	0.02	0.44	0.38	1.50	0.49
min	10.59	12.41	0.81	4.90	2.63	0.77	4.52
25%	12.27	13.45	0.86	5.26	2.94	2.56	5.04
50%	14.36	14.32	0.87	5.52	3.24	3.60	5.22
75%	17.30	15.72	0.89	5.98	3.56	4.77	5.88
max	21.18	17.25	0.92	6.68	4.03	8.46	6.55

Table 4: Five point summary of dataset-1 before scaling



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.75	1.81	0.18	2.37	1.34	-0.30	2.33
1	0.39	0.25	1.50	-0.60	0.86	-0.24	-0.54
2	1.41	1.43	0.50	1.40	1.32	-0.22	1.51
3	-1.38	-1.23	-2.59	-0.79	-1.64	0.99	-0.45
4	1.08	1.00	1.20	0.59	1.16	-1.09	0.87

Table 5: Scaled dataset-1



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.00	210.00	210.00	210.00	210.00	210.00	210.00
mean	0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00
min	-1.47	-1.65	-2.67	-1.65	-1.67	-1.96	-1.81
25%	-0.89	-0.85	-0.60	-0.83	-0.83	-0.76	-0.74
50%	-0.17	-0.18	0.10	-0.24	-0.06	-0.07	-0.38
75%	0.85	0.89	0.71	0.79	0.80	0.71	0.96
max	2.18	2.07	2.01	2.37	2.06	3.17	2.33

Table 6: Five point summary of dataset-1t after scaling

We can see that the scaled dataset and its 5 point summary are in a consistent format that adheres to the standards. This helps create a single view, enables cross-functional analytics to drive improved insights, enhances productivity, ensures clean, decluttered & trusted database that can be governed, harmonizes data, provides data transparency and reduces cost. After applying standardisation we can see that mean=0 and std=1 as expected.

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

In hierarchical clustering records are sequentially grouped to create clusters based on distances between records and distances between clusters.
It produces a useful graphical display of the clustering process and results, called a "dendrogram"
Its advantageous as no assumptions on number of clusters is required at output and can be obtained by 'cutting'.
However, they are sensitive to outliers and have time complexity which gets evident for a larger dataset.

On applying hierarchical clustering on the given dataset, the below dendrogram is obtained:

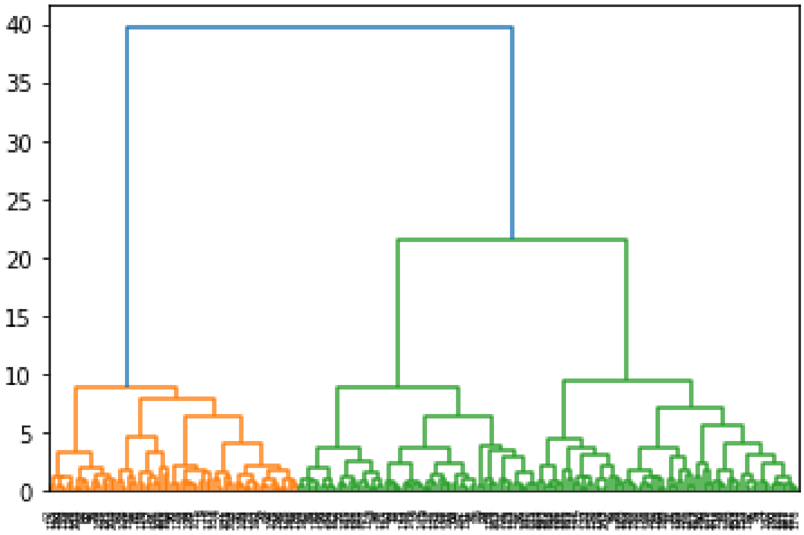


Figure 5: Dendrogram

On truncating the above for last 10 clustering, the below figure is obtained:

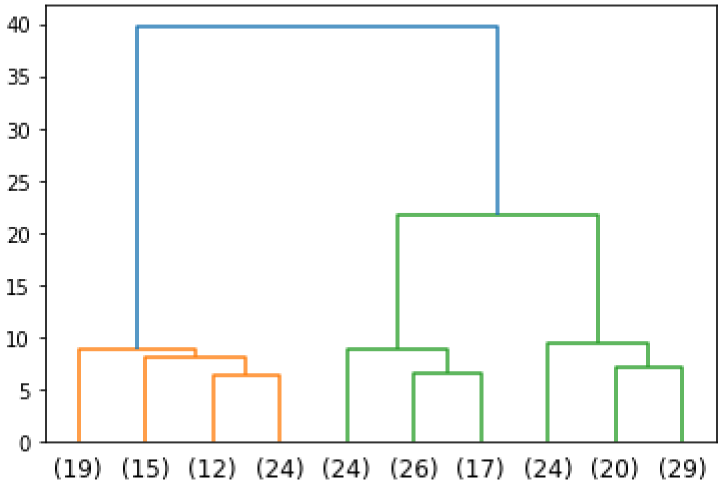


Figure 6:Truncated dendrogram

It is observed that the entire dataset can be clustered into 2 distinct groups.
1st cluster, orange colored, has (19+15+12+24) 70 number of items in total, 2nd cluster, green colored, has (24+26+17+24+20+29) 140 number of items in total. Together they make 210 records, which is that total records of the dataset.

However,generally speaking, 2 clusters really do not make much business impact as it is kind of implicit. SO we take optimal number of clusters to be 3.

Agglomerative clustering is a bottom up approach where each object starts off as a separate group. closer and similar the groups are merged. This helps identify small clusters. Dendrograms follow this approach, it is used to find the optimal number of clusters in a dataset by a treelink diagram that summarizes the process of clustering. x-axis - Records, y-axis-distance between records Similar records are joined by lines whose vertical length reflects the distance between the records.

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 a non-hierarchical clustering technique which the value of number of clusters (k) beforehand.
It assigns each record to one of the 'k' number of clusters according to their distance from each cluster. By default, Eucleidian distance is considered.
This minimises the measure of dispersion within the clusters.
The 'Means' in the K-Means refers to averaging of the data i.e. finding the centroid.
This method is used to larger dataset.

The k-Means inertia is found for k value 1 to 10 as below, an elbow curve is also plotted for the same.

1470.0000000000002, 659.1717544870406, 430.6589731513006, 371.30172127754213, 326.2289168297265, 290.0962461660115, 262.43855218665294, 240.1179962275735, 221.18781104234662, 206.94248345086626

WSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WSS with the K-value, the plot looks like an Elbow. As the number of clusters increases, the WSS value decreases. WSS value is the largest at k=1.
From the graph, it is clearly seen that the WSS value takes a major drop at k=2 (point of inflection) further which the drop is low. Generally speaking, 2 clusters really do not make much business impact as it is kind of implicit. Hence, k=3 is chosen.
Optimum number of clusters is 3.

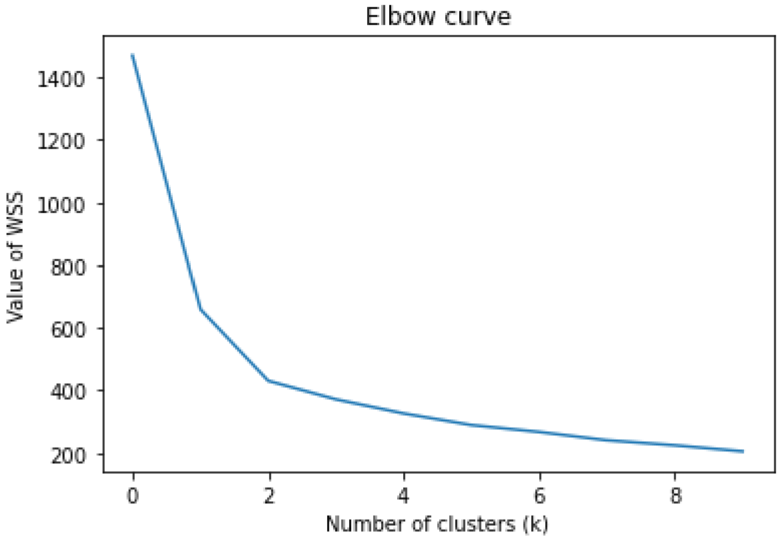
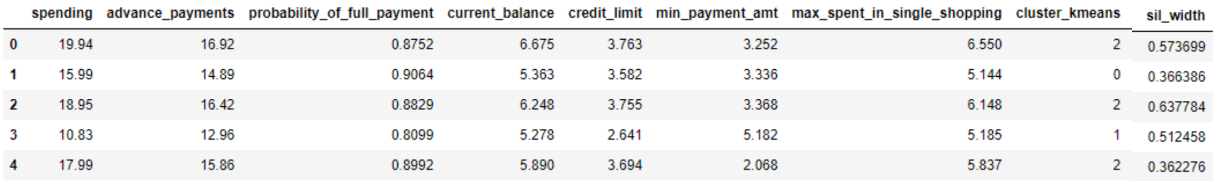


Figure 7: Elbow curve

K-Means inertia for k=3 is 430.66
Silhouette score is the average of all the silhouette widths, and is calculated to be 0.4
Silhouette score is less than 0.5, hence can be said that the cluster distinction is not great
Minimum silhouette width is 0.003, positive value says no record wrongly mapped to a cluster.
Maximum silhouette width is 0.639

After adding each record mapping to a cluster(cluster_kmeans) and adding its silhouette width (sil_ width), the data set looks as below:



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	cluster_kmeans	sil_width
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	2	0.573699
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	0	0.366386
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	2	0.637784
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	1	0.512458
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	2	0.362276

Table 7: Dataset with cluster_kMeans and sil_width

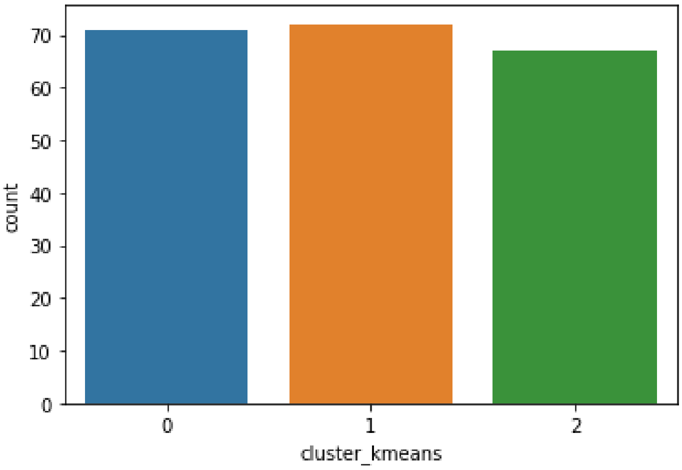
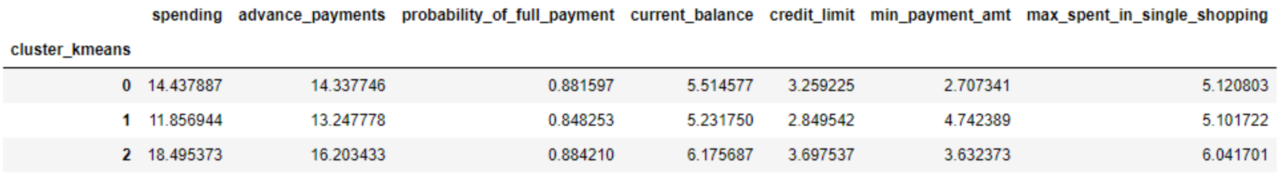


Figure 8: K-Means cluster value counts

The optimum number of clusters for the dataset is 3.
The first cluster,Cluster-0, blue coloured, has 72 records mapped to it
The second cluster, Cluster-1, orange coloured, has 71 records mapped to it.
The third cluster, Cluster-2,green coloured, has 67 records mapped to it
Totally, they have 210 records as expected.

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

The records are grouped by their cluster and average of variables are taken.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
cluster_kmeans							
0	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803
1	11.856944	13.247778	0.848253	5.231750	2.849542	4.742389	5.101722
2	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701

Table 8: Cluster data with average variables' values

We can see three clusters: Cluster-0, Cluster-1 & Cluster-2

From the table we can observe that Cluster-2 has maximum mean values for variables spending, advance_payments, probability_of_full_payment, current_balance, credit_limit, and max_spent_in_single_shopping.

Cluster-1 has the least spending, advance_payments, probability_of_full_payemnt,current_balance and credit_limit. However, this cluster has the highest mean for min_payment_amt.

Cluster-0 has least min_payment_amt, a very promising probability_of_full_payment and current_balance.

Spending to credit_limit ration is as follow:
Custer-0 = 4.37
Cluster-1 = 4.16
Cluster-2 = 5.44

The risk is least with Cluster-0 customers has they have least chances of default since their full payment of credit card bill is the highest.
It is safe to say that an increase in their credit limit will have a direct and positive impact on the business growth.
Their current balance is average, boosting their expenditure to lower the current balance and improve business can be achieved by giving discounts and offers on usage of credit cards.

Cluster-1 customers come with highest risk factor as their probability of full payment and advance payemnts is the least, and have high minimum payment amount. They have least current balance which further adds to the risk quotient.
The customers belonging to this group need to be pushed towards bill payments to reduce risk factor.

This can be achieved by cutting down on offers and discounts on usage of credit cards, by applying/increasing interest rates

Cluster-2 customers seem to be safe players. Their credit limit and current balance is high, indicting lower expenditure. They are showing potential with highest spent during single shopping spree. They seem to favour advance payments.
Business from this group can further be enhanced by encouraging them on advance payments with offers and perks.
Their credit card usage can be ramped up with offers and promotions.

Problem 2

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.

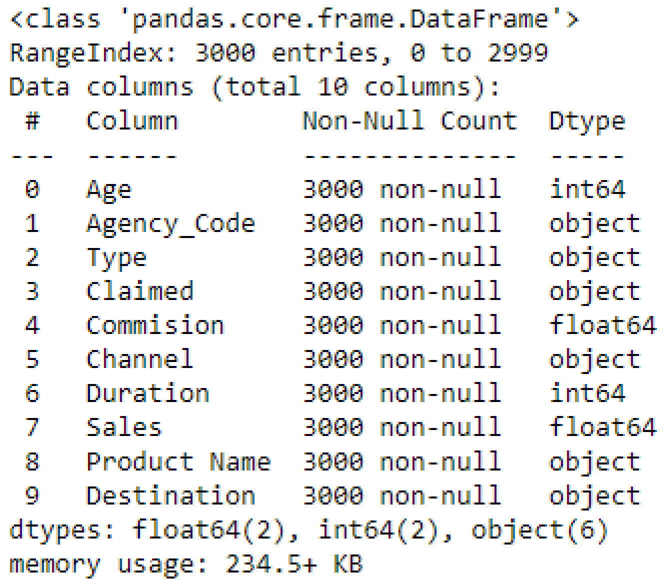
Dataset for Problem 2: insurance_part2_data-1.csv

Attribute Information:

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)

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

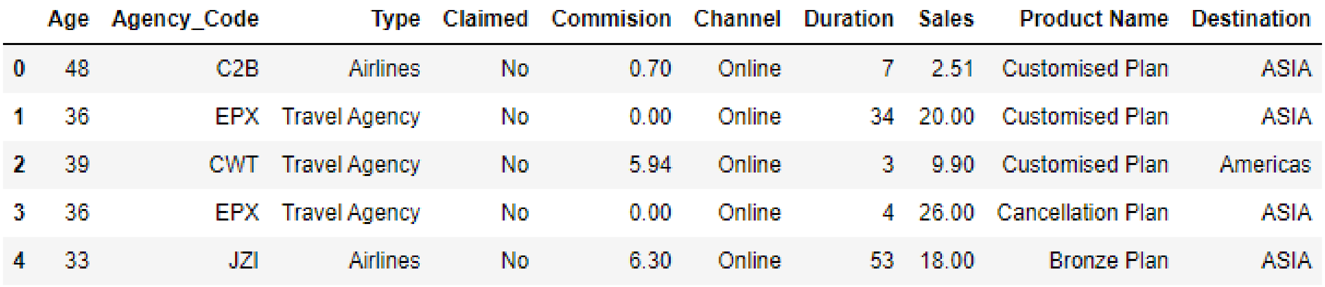
The data set has 3000 rows and 10 columns.
The columns names are : 'Age', 'Agency_Code', 'Type','Claimed', 'Commision','Channel', 'Duration', 'Sales', 'Product Name', Destination'
There are no null values in the datatset.
However, 139 duplicates are found.
The datatypes present are int, float and object.



```
<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
```

Table 9: Dataset-1 info

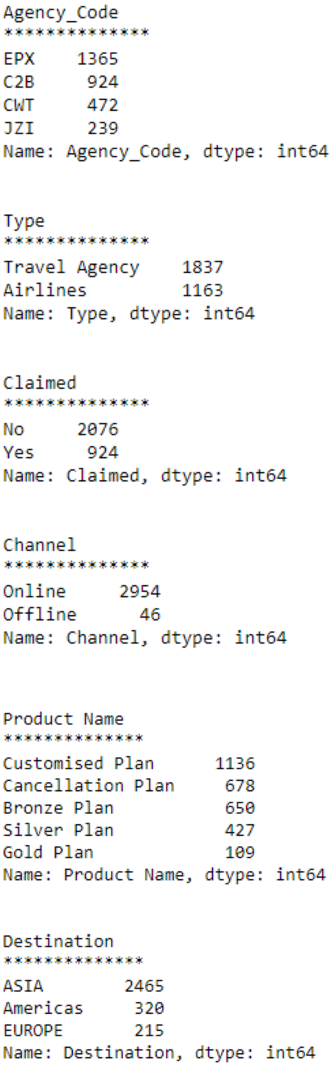
The first five entries of the dataset looks as below:



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

Table 10: Head of dataset-1

Categorical vaiables and their value counts along with theor unique values is given below:



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


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


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


Channel
**************
Online     2954
Offline      46
Name: Channel, dtype: int64


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


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

Table 11: Dataset-2 categorical variables value counts and unique values

Histogram and boxplot for each variable of the dataset is given below, outliers are seen in all of the variable.

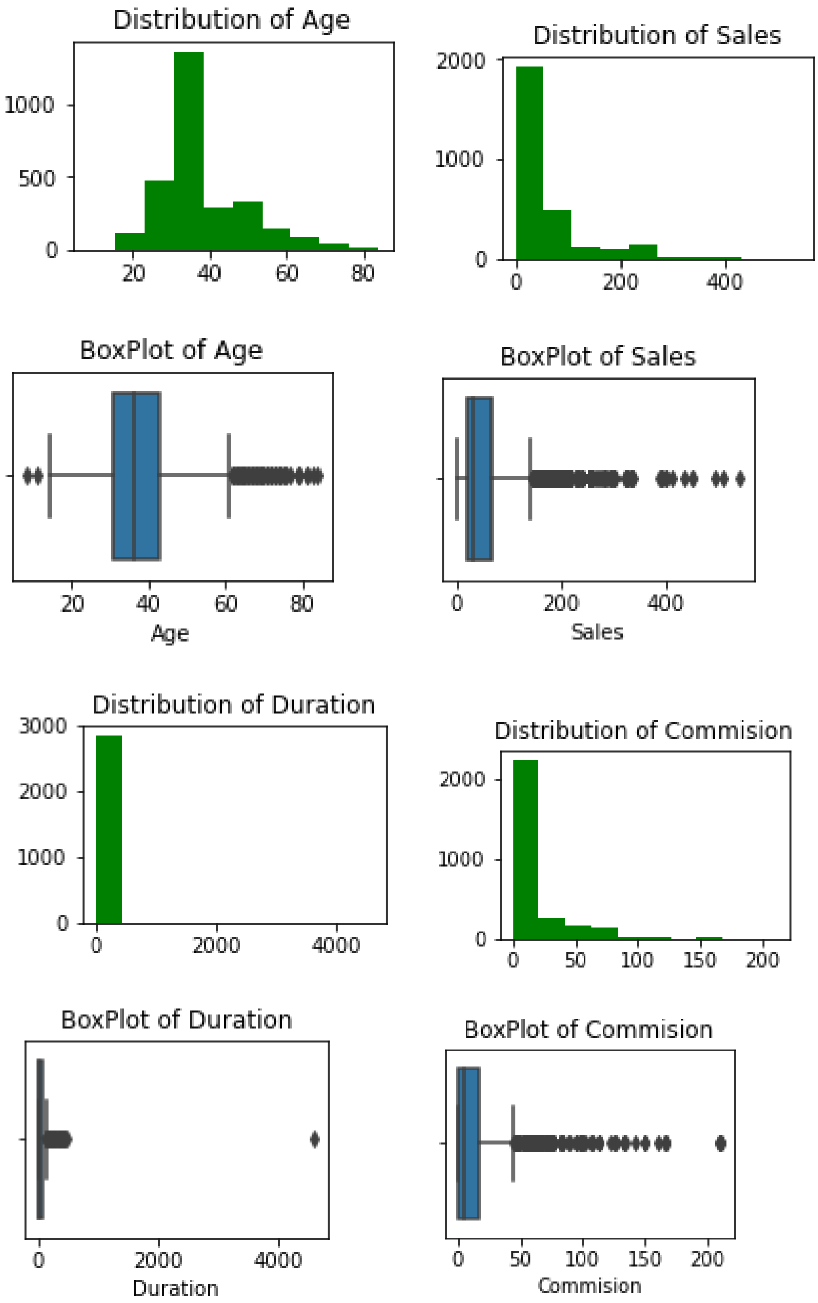


Figure 9: Histogram and Boxplot graphs

Give below is the pairplot of the dataset. A pairplot visualizes the relationship between 2 numeric variables.

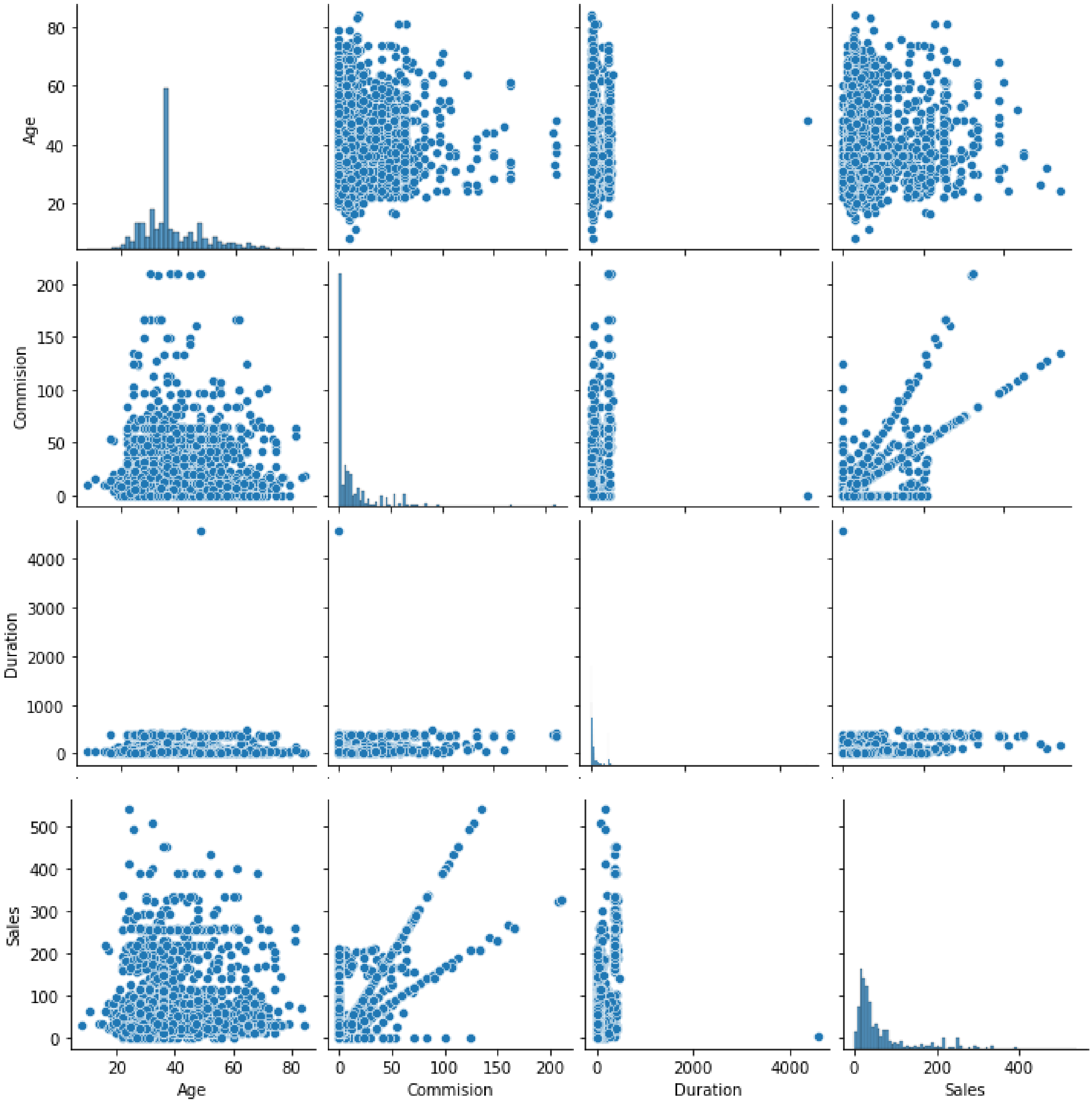


Figure 10: Pairplot of dataset-2

Below is a heatmap, it is a type of a plot that visualizes the strength of relationships between numerical variables present in the dataset. A variable has highest correlation (=1) with itself. We can see that Commision-Sales has the highest correlation, whereas Age has the lowest correlation with the rest of the independent variables.

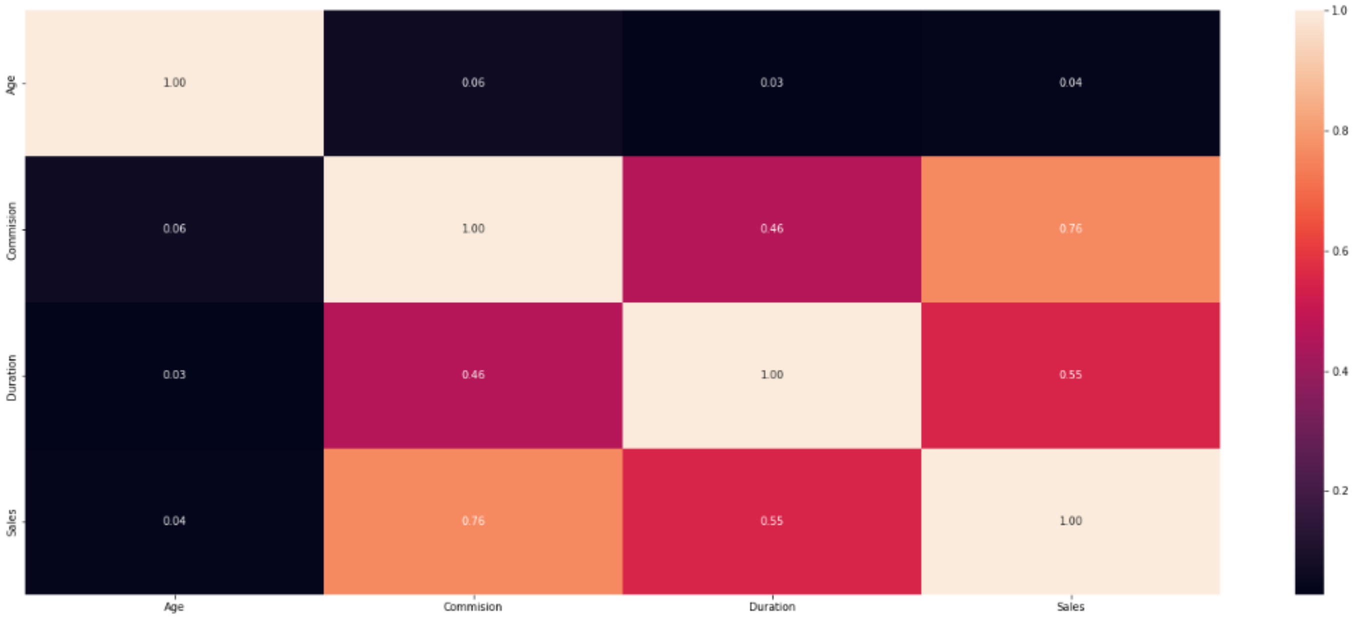


Figure 11: Heatmap for correlation of dataset-2

After dropping the duplicate values, the dataset has 2861 rows and 10 columns.
Object datatype is converted to integer.

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

The whole dataset is split into test data and train data. Training data is the subset of original data tha t is used to train the machine learning model, whereas testing data is used to check the accuracy of t he model.
In this case, 70% of the data is used for training and 30% is used for testing.
Training set contains 2002 records
Testing set contains 859 records

Decision tree learning is a supervised learning approach. In this technique, a classification or regression decision tree is used as a predictive model to draw conclusions about a set of observations. The parent node gets into split into child nodes and pruning is done to avoid aver growing of sub-trees/branches.

Multiple Decision trees together form a Random forest.

Random forest is an ensemble technique wherein we construct multiple models and take the average output of all models to take final decision/make predictions.

Artificial neural networks is a black-box technique, fairly modern and its extension is deep learning. Its design is roughly modelled around what is currently known about human brain functions.
It can learn, generalise and adapt. It has 3 layers: input, hidden, output.

Various input parameters can be passed for each of these models. The combination is calculated and suggested by using model searching from GridSearchCV.

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.

Import features tells, of all the available variables which one have the highest impact on the dependent element.

Receiver Operating Characteristics (ROC) Curve is a graphical plot that illustrates the ability or accuracy of a classifier system when its threshold is varied.
It gives the trade-off between sensitivity (TPR) and specificity (1-FPR).
Classifiers that give curves closer to the top-left corner indicate a better performance.

Classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1-score and support of the tested classified model.

Confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data.

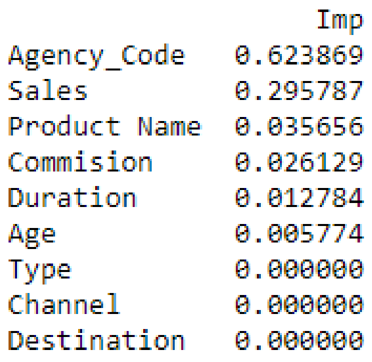
Accuracy is an evaluating metric for a classification model. Informally, it is the fraction of predictions our model got right.
Accuracy= Number of right predictions/Total number of predictions

Random state used is 0.

(a) For DecisionTreeClassifier:

Training data-
Precision: 1.0
Recall: 0.98
F1 Score: 0.99
Grid score: 0.80
Accuracy: 78%

Testing data-
Precision: 0.54
Recall: 0.48
F1 Score: 0.51
Grid score: 0.74
Accuracy: 75%



```
                     Imp
Agency_Code     0.623869
Sales           0.295787
Product Name    0.035656
Commision       0.026129
Duration        0.012784
Age             0.005774
Type            0.000000
Channel         0.000000
Destination     0.000000
```

Figure 12: DTC-Import features

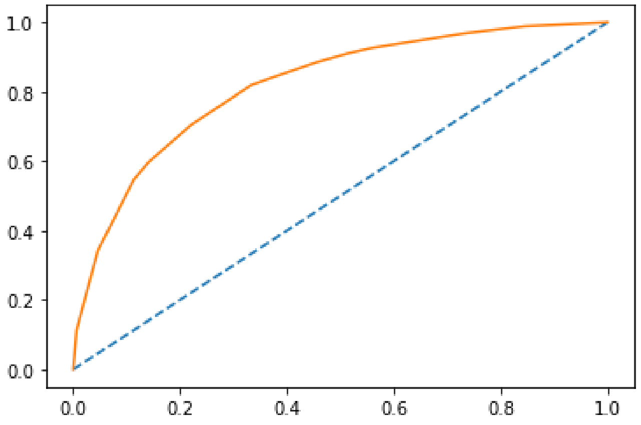


Figure 13: DTC – Training data ROC

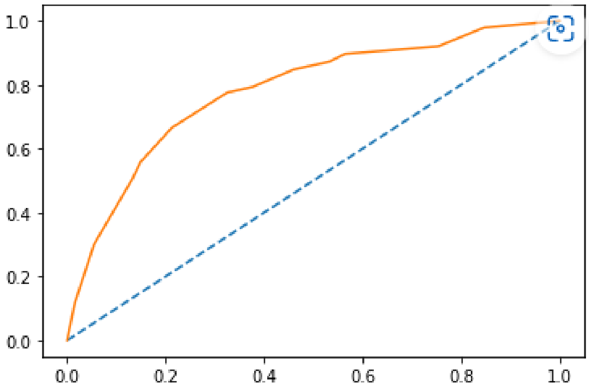
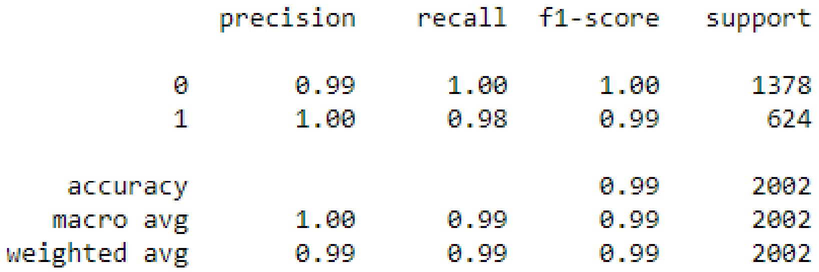
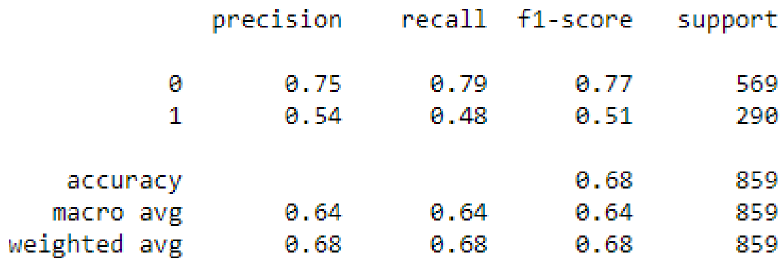


Figure 14: DTC- Testing data ROC



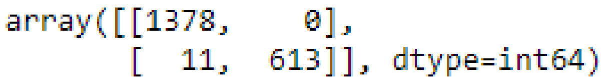
```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      1378
           1       1.00      0.98      0.99       624

    accuracy                           0.99      2002
   macro avg       1.00      0.99      0.99      2002
weighted avg       0.99      0.99      0.99      2002
```

Figure 15: DTC- Training data Classification report



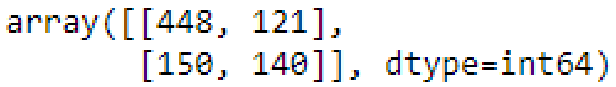
```
              precision    recall  f1-score   support

           0       0.75      0.79      0.77       569
           1       0.54      0.48      0.51       290

    accuracy                           0.68       859
   macro avg       0.64      0.64      0.64       859
weighted avg       0.68      0.68      0.68       859
```

Figure 16: DTC- Testing data Classification report



```
array([[1378,    0],
       [  11,  613]], dtype=int64)
```

Figure 17: DTC- Training data Confusion matrix



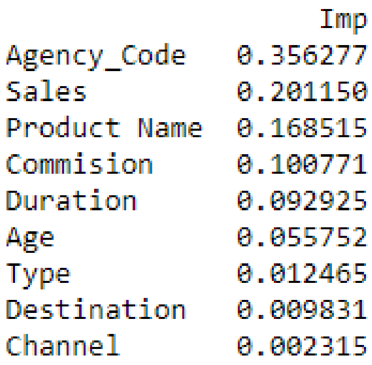
```
array([[448, 121],
       [150, 140]], dtype=int64)
```

Figure 18: DTC- Testing data Confusion matrix

(b) For RandomForestClassifier:

Training data-
Precision: 0.73
Recall: 0.61
F1 Score: 0.66
Grid score: 0.80
Accuracy: 81%

Testing data-
Precision: 0.67
Recall: 0.57
F1 Score: 0.62
Grid score: 0.70
Accuracy: 76%



```
                   Imp
Agency_Code   0.356277
Sales         0.201150
Product Name  0.168515
Commision     0.100771
Duration      0.092925
Age           0.055752
Type          0.012465
Destination   0.009831
Channel       0.002315
```

Figure 19: RF-Important features

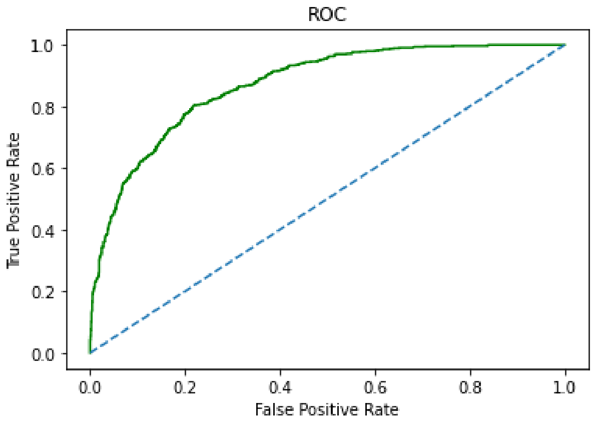


Figure 20: RF- Training data ROC

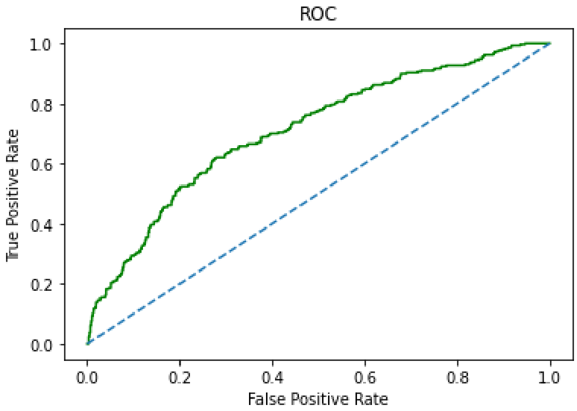
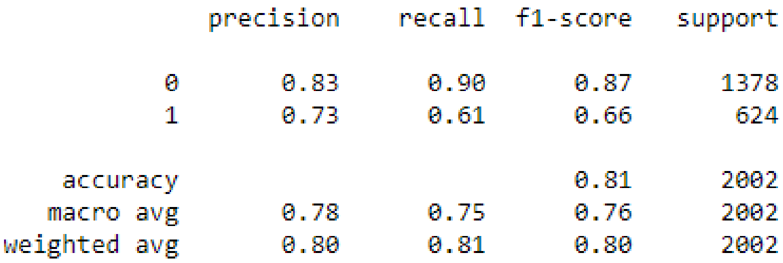
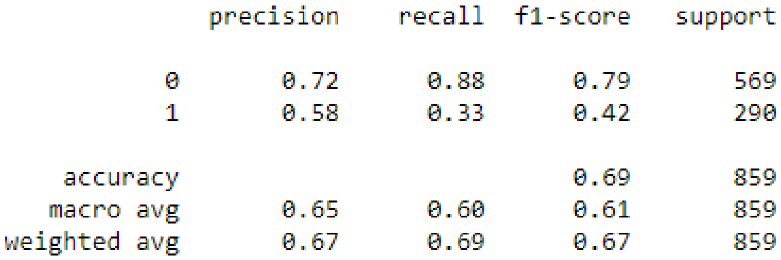


Figure 21: RF-Testing data ROC



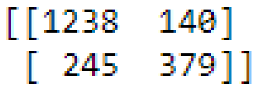
```
              precision    recall  f1-score   support

           0       0.83      0.90      0.87      1378
           1       0.73      0.61      0.66       624

    accuracy                           0.81      2002
   macro avg       0.78      0.75      0.76      2002
weighted avg       0.80      0.81      0.80      2002
```

Figure 22: RF-Training data Classification report



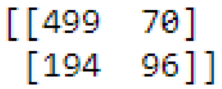
```
              precision    recall  f1-score   support

           0       0.72      0.88      0.79       569
           1       0.58      0.33      0.42       290

    accuracy                           0.69       859
   macro avg       0.65      0.60      0.61       859
weighted avg       0.67      0.69      0.67       859
```

Figure 23: RF-Testing data Classification report



```
[[1238  140]
 [ 245  379]]
```

Figure 24: RF-Training data Confusion matrix



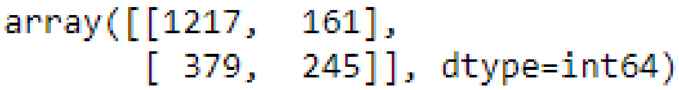
```
[[499  70]
 [194  96]]
```

Figure 25: RF-Testing data Confusion matrix

(c) For NeuralNetworks:

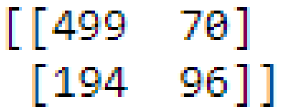
Training data-
Precision: 0.60
Recall: 0.39
F1 Score: 0.48
Grid score: 0.73
Accuracy: 73%

Testing data-
Precision: 0.58
Recall: 0.33
F1 Score: 0.42
Grid score: 0.70
Accuracy: 69 %



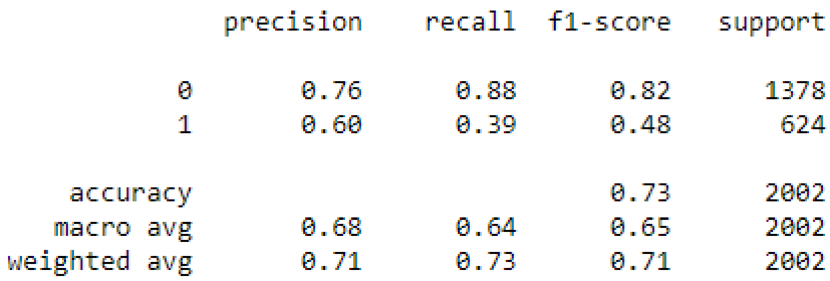
```
array([[1217,  161],
       [ 379,  245]], dtype=int64)
```

Figure 26: NN-Training data Confusion matrix



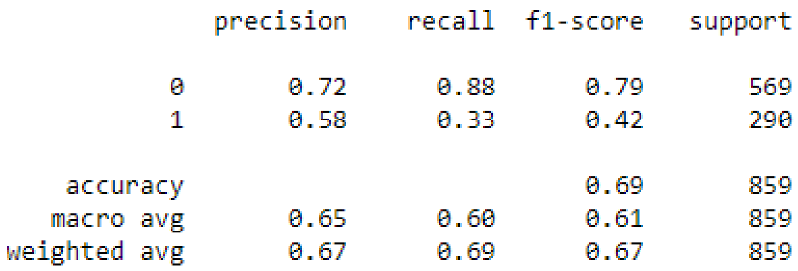
```
[[499  70]
 [194  96]]
```

Figure 27: NN-Testing data Confusion matrix



```
              precision    recall  f1-score   support

           0       0.76      0.88      0.82      1378
           1       0.60      0.39      0.48       624

    accuracy                           0.73      2002
   macro avg       0.68      0.64      0.65      2002
weighted avg       0.71      0.73      0.71      2002
```

Figure 28: NN-Training data Classification report



```
              precision    recall  f1-score   support

           0       0.72      0.88      0.79       569
           1       0.58      0.33      0.42       290

    accuracy                           0.69       859
   macro avg       0.65      0.60      0.61       859
weighted avg       0.67      0.69      0.67       859
```

Figure 29: NN-Testing data Classification report

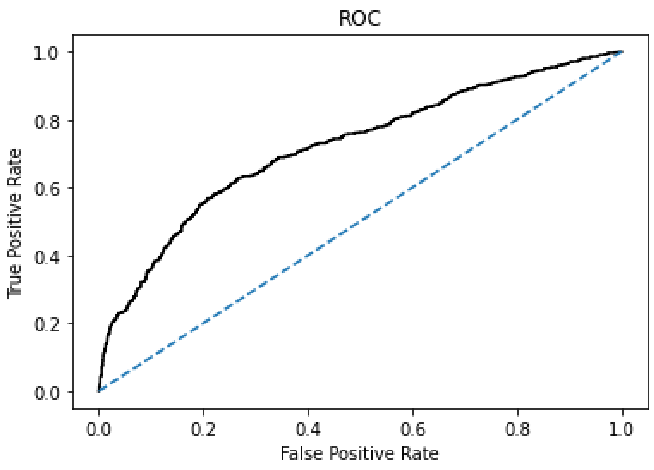


Figure 30: NN-Training data ROC

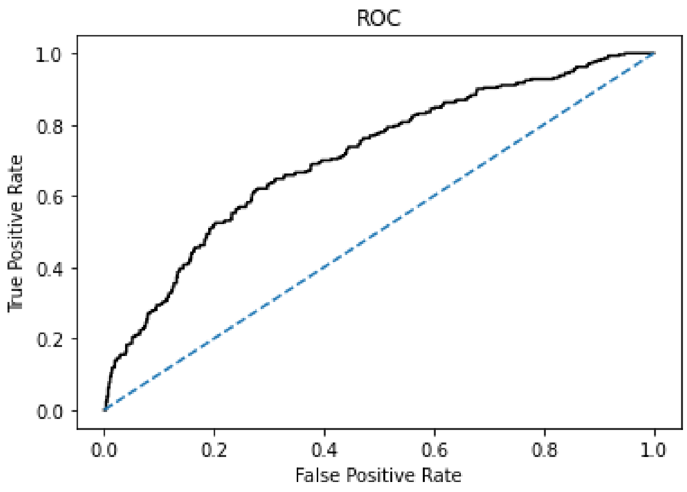
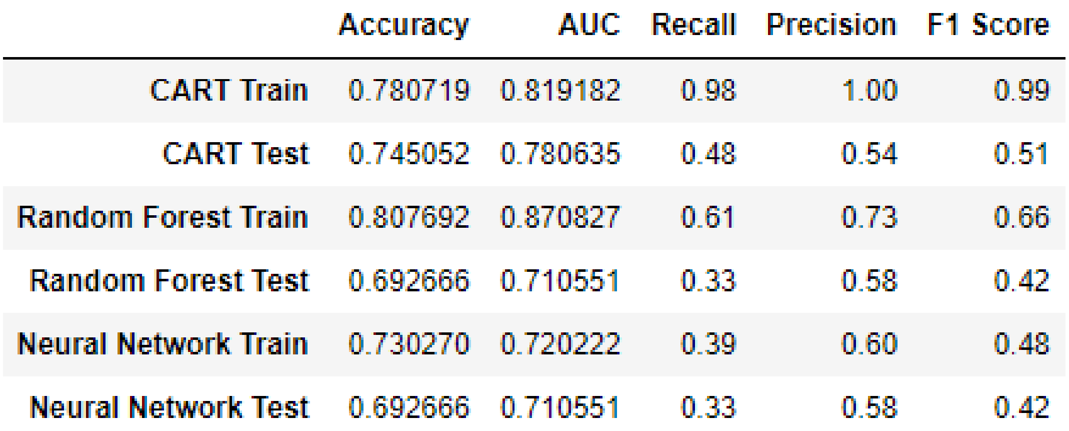


Figure 31: NN-Testing data ROC

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

Comparison between all the 3 models:



	Accuracy	AUC	Recall	Precision	F1 Score
CART Train	0.780719	0.819182	0.98	1.00	0.99
CART Test	0.745052	0.780635	0.48	0.54	0.51
Random Forest Train	0.807692	0.870827	0.61	0.73	0.66
Random Forest Test	0.692666	0.710551	0.33	0.58	0.42
Neural Network Train	0.730270	0.720222	0.39	0.60	0.48
Neural Network Test	0.692666	0.710551	0.33	0.58	0.42

Table 12: Comparison of the performance metrics from the 3 models

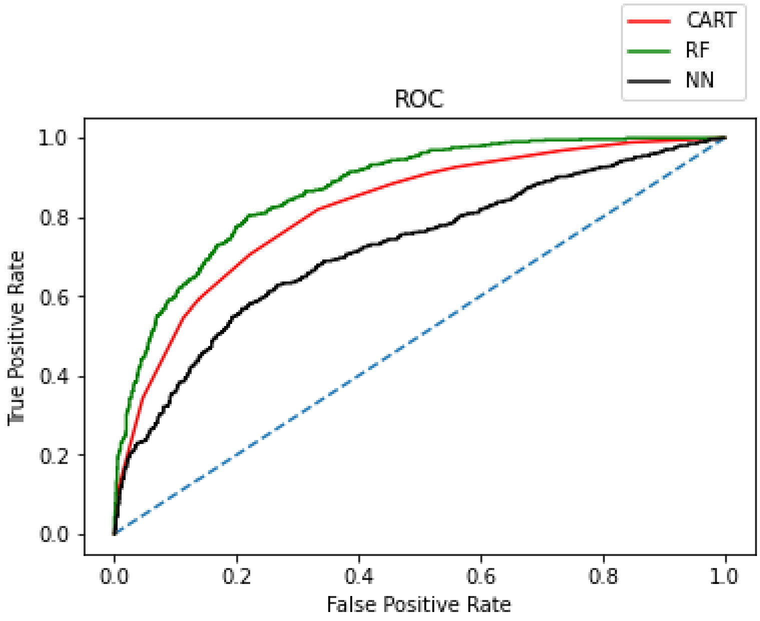


Figure 32: Training data ROC from the 3 models

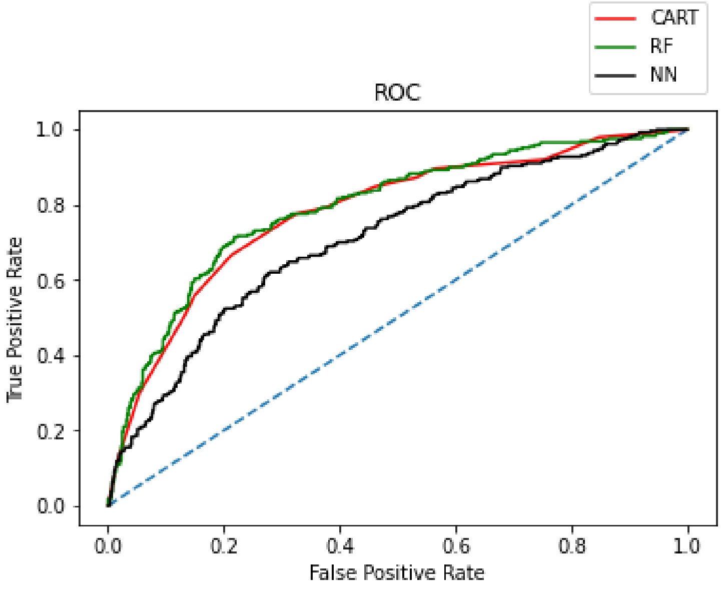


Figure 33: Testing data ROC from the 3 models

From the comparison table we can observe that the DecisionTreeClassifier (CART) is a better model than the rest. Accuracy for both Train and Test datas are close to each other. Their recall rates are also pretty good in comparison.
Neural Networks model stands next with equally good accuracy and good recall rates

If the business is looking to reduce their false positives, precision should be high and choosing CART model for it makes sense.
If the business is looking to reduce their false negatives, recall should be high, and again choosing CART model for it makes sense.

Overall, CART model can to chosen has the most optimum one.

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

Agency_Code C2B have the highest claims , most of who's distribution channel is online.
Maximum claimed is for product type Silver Plan

Commission for C2B agency can be reduced.
The price for tour package i.e. Silver plan for agency can be increased