Data Mining Project Report

NAME: SHOUNACK MANDAL

COURSE: PGP - DSBA Online Sep.

Date: 12/ December / 2021

Contents

LIST OF FIGURE

LIST OF TABLES

Problem 1: Clustering

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

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

Introduction

The purpose of this whole exercise is to explore the dataset. The data consists of activities of different customers based on their credit card usage. We will perform exploratory data analysis to understand what the given data has to say and then use clustering techniques to develop a customer segmentation so that the bank can give promotional offers to its customers based on the clusters we have identified.
Sample of the dataset shows the data of 7 variables with different usage of credit card.

Exploratory Data Analysis

Dataset Information

	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 1: Sample of the dataset.

Descriptive Statistics

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

Table 2: - Description of the original dataset

Inferences:-

1. On an average customers spends not more than 15 K per month and offers cash in advance on an average of 1456
2. The probability of full payment is very consistent by the customers, which is above 80 % and with an average of 87% and SD = 0.02.
3. A few outliers can be observed in the probability of full outliers and min payment amount.
4. The current balance with customers is on an average of 4899 and credit limit at 32586.

```
<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 3: Info of Data-Set

5. All the variables are in numeric data type.
6. There isn`t any null or duplicates records in the original dataset.
7. The dataset contains in total 210 observations of credit card activities of various customers and there are no nominal variables or identification columns in the data to be rectified, data looks clean and so we can start out exploratory analysis on the data-set.

Univariate Analysis

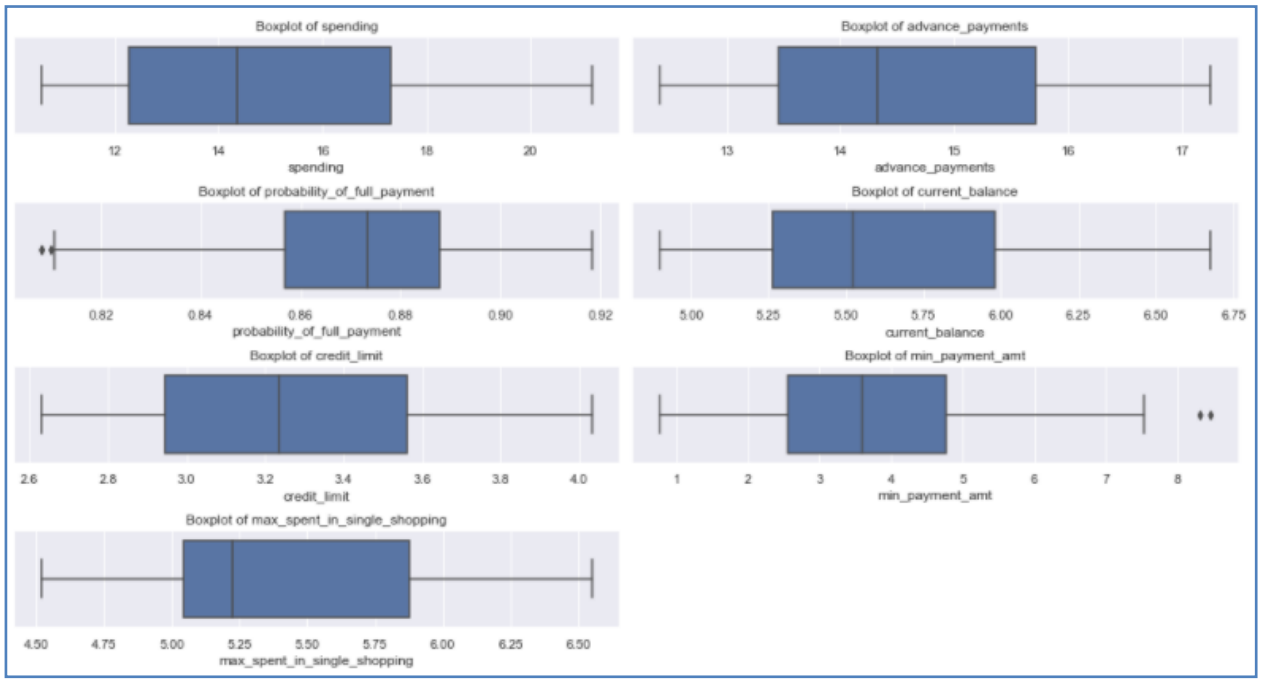
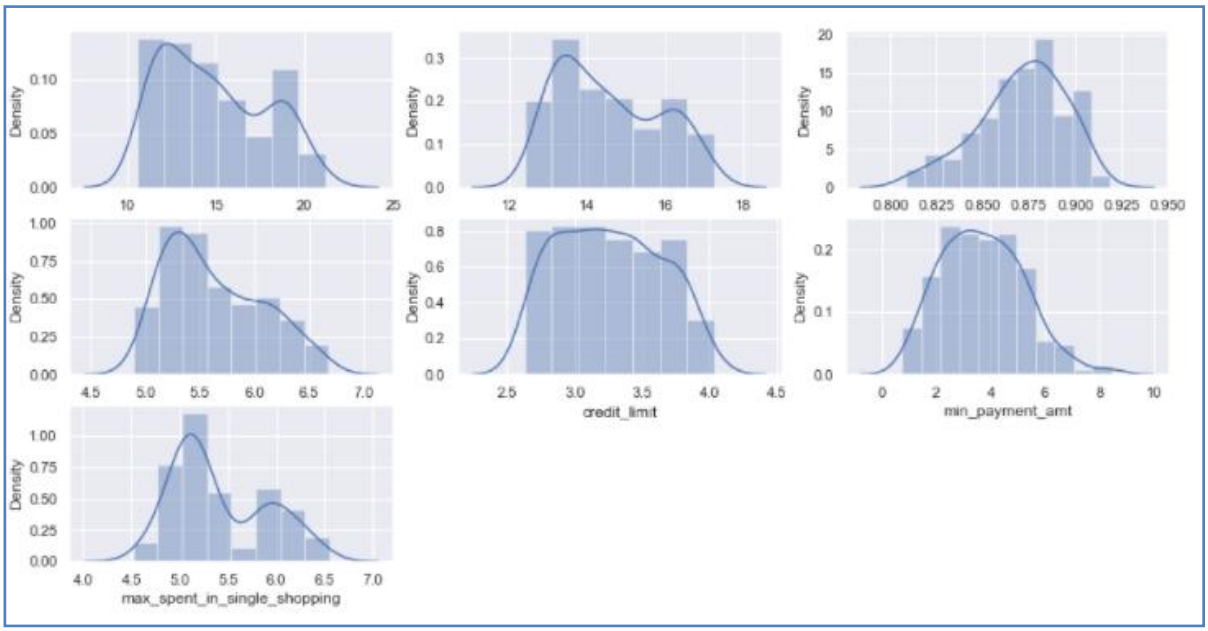


Figure 1 : - Box-plot for all variables

Bi-Variate Analysis

Distribution Plot



Inferences:-

1. Very few outliers so we Figure 2 : - Distribution plot for all variables. because of the high expenses made by the consumers as they had segment of high credit limit and hence higher minimum payment amount which causing the outlier here.
2. There is very less in differences in their mean and average values.

Multivariate analysis

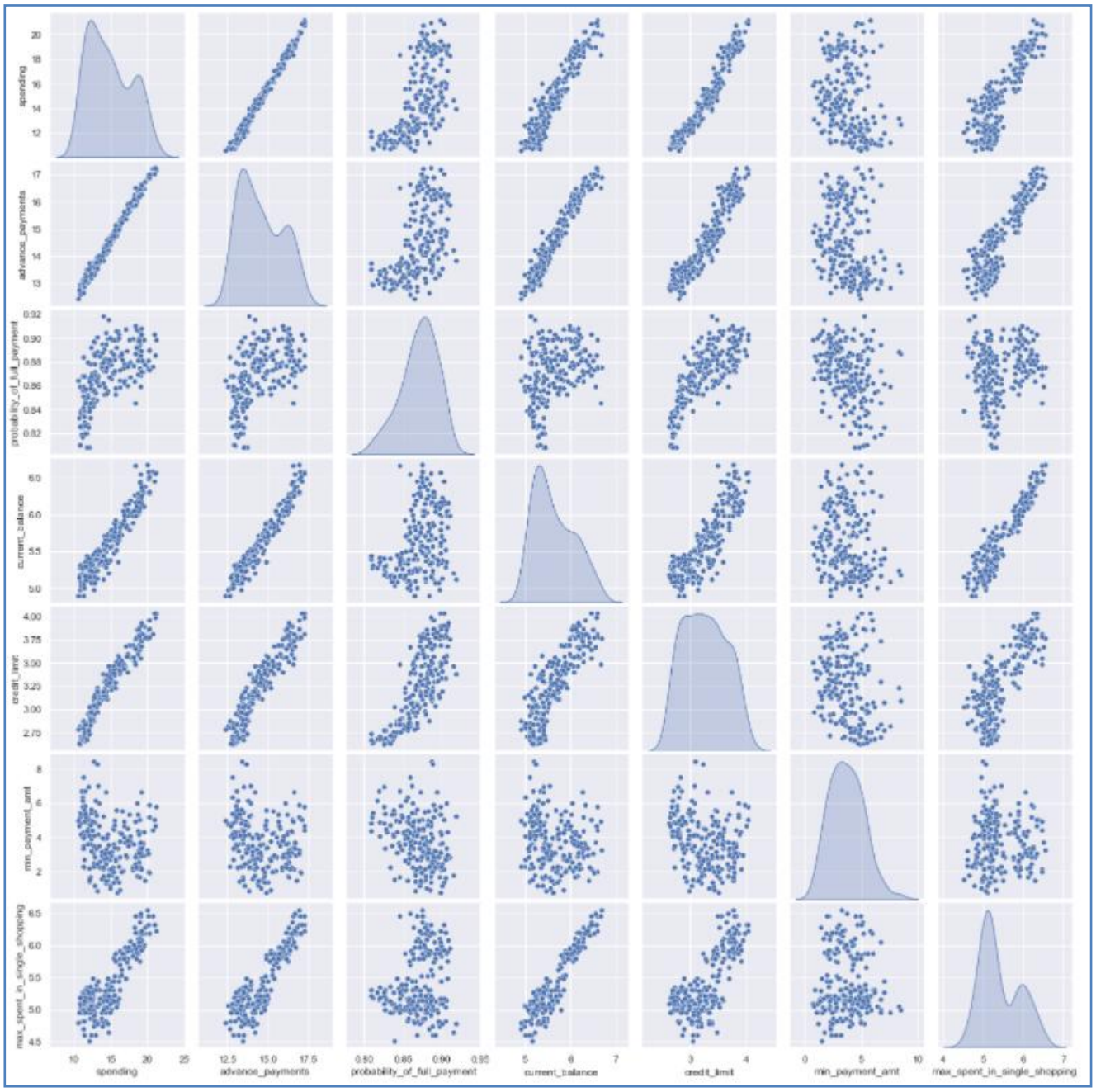


Figure 3 : - Pair-plot for all the variables.

Correlation Heat-map

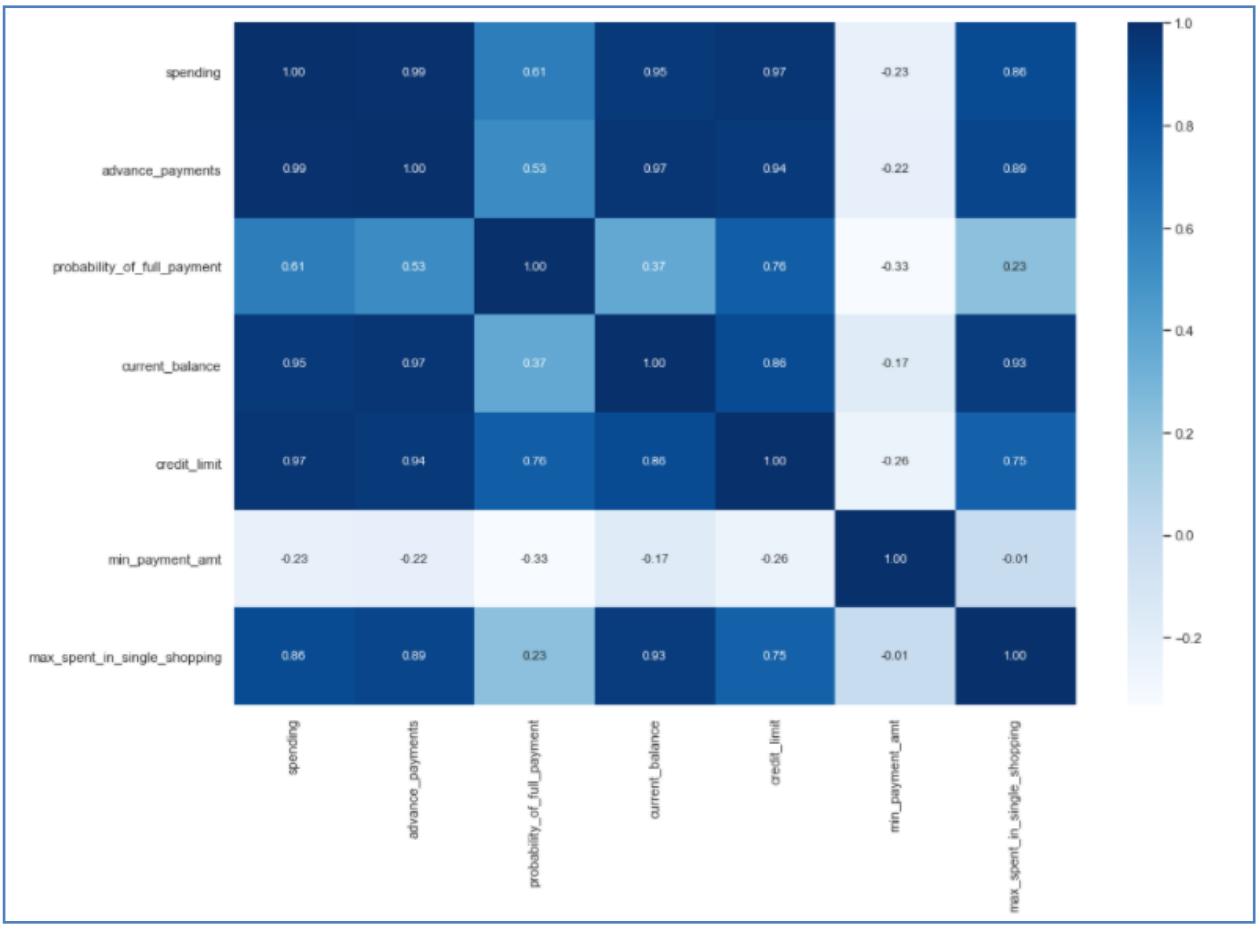


Figure 4 : - Correlation heat-map.

Inferences:-

1. Correlation values are always between 1 to – 1, values varies from positively correlated towards negatively correlated. In middle values with 0 denotes to no correlation.
2. Few as very high positive correlation are spending and advance payment with each of the other variables.
3. Except for ‘minimum payment amount’ all other variables are highly positive correlated to each other.
4. The best correlation can be seen in variables these are ‘advance_payment’, ‘current_balance’ and ‘credit_limit’.

1.2 Do you think scaling is necessary for clustering in this case? Justify 2

Let`s identify the variance for all variables in the given data-set.

```
spending                      8.466351
advance_payments              1.705528
probability_of_full_payment   0.000558
current_balance               0.196305
credit_limit                  0.142668
min_payment_amt               2.260684
max_spent_in_single_shopping  0.241553
dtype: float64
```

Table 3 : - Variance for all the variables.

Only the target variable has much high value as compared to other variables i.e.,. variance of 8.46 for spending and in other hand all other variable has variance less than 2.26.

The standard score can be calculated with the formulae: -

$$z = \frac{X - \overline{X}}{S}$$

Figure 3 : Z-Score formulae.

Where Z is the standard score, S = The standard deviation of a sample set, X = each value in the data-set, X bar = mean of all values in the data set.

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

Table 3 : - Scaled data.

The standard scalar function has converted the original data in the range of -2 to +3.

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them 7

Hierarchical Clustering

This analysis of clustering and performing hierarchy of clusters is called hierarchical clustering. Mainly two type of clusters:-

Linkage

There are different methods to measure the distance among the clusters once formed. These are many linkage types. For our dataset we will work upon:-

1. **Single linkage** (Distance between the clusters is defined as shortest distance between two points in each cluster)
2. **Complete linkage** (Distance between the clusters is defined as longest distance between two points in each cluster)
3. **Ward`s Method** (Join records and clusters together progressively to produce larger and larger clusters)

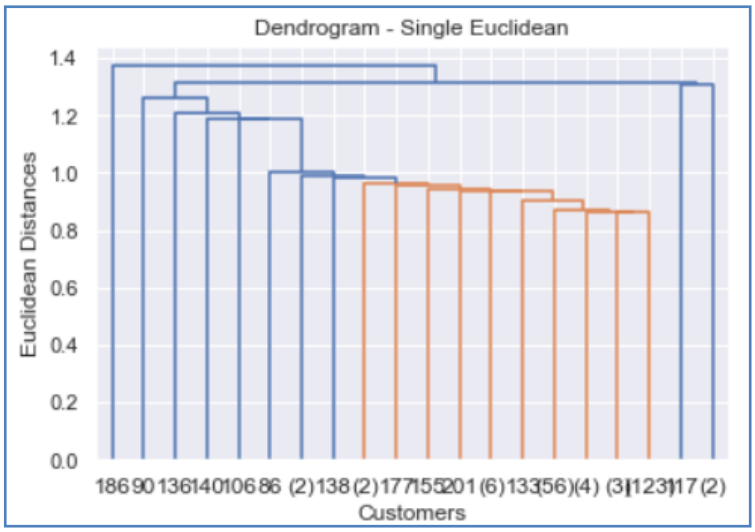


Figure :- Dendrogram – Single Euclidean

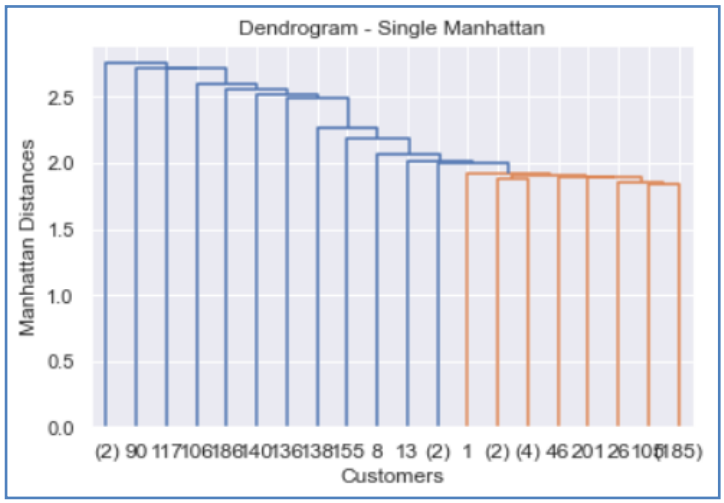


Figure : Dendrogram – Single Manhattan

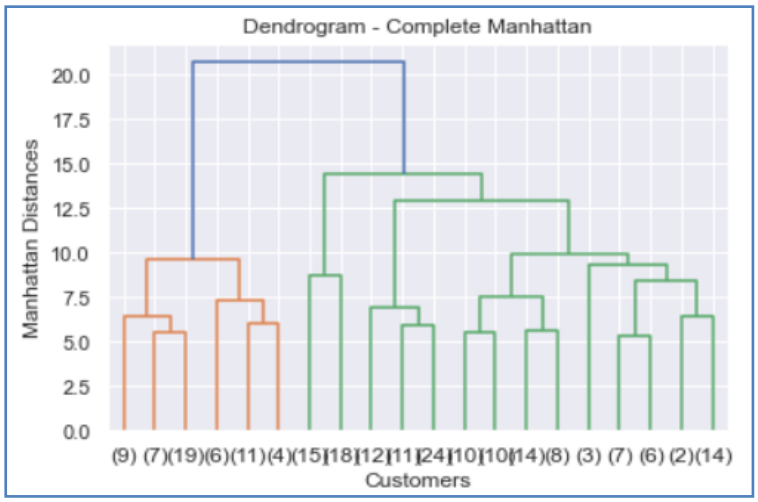


Figure : Dendrogram – Complete Manhattan

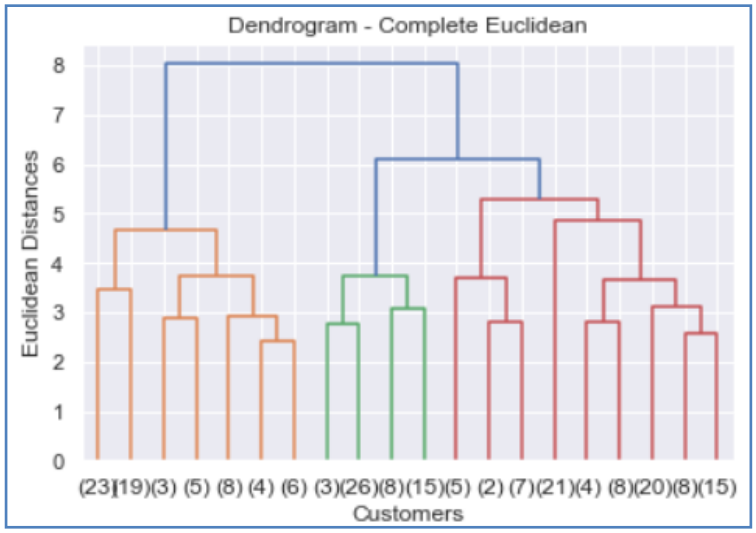


Figure : Dendrogram – Complete Euclidean

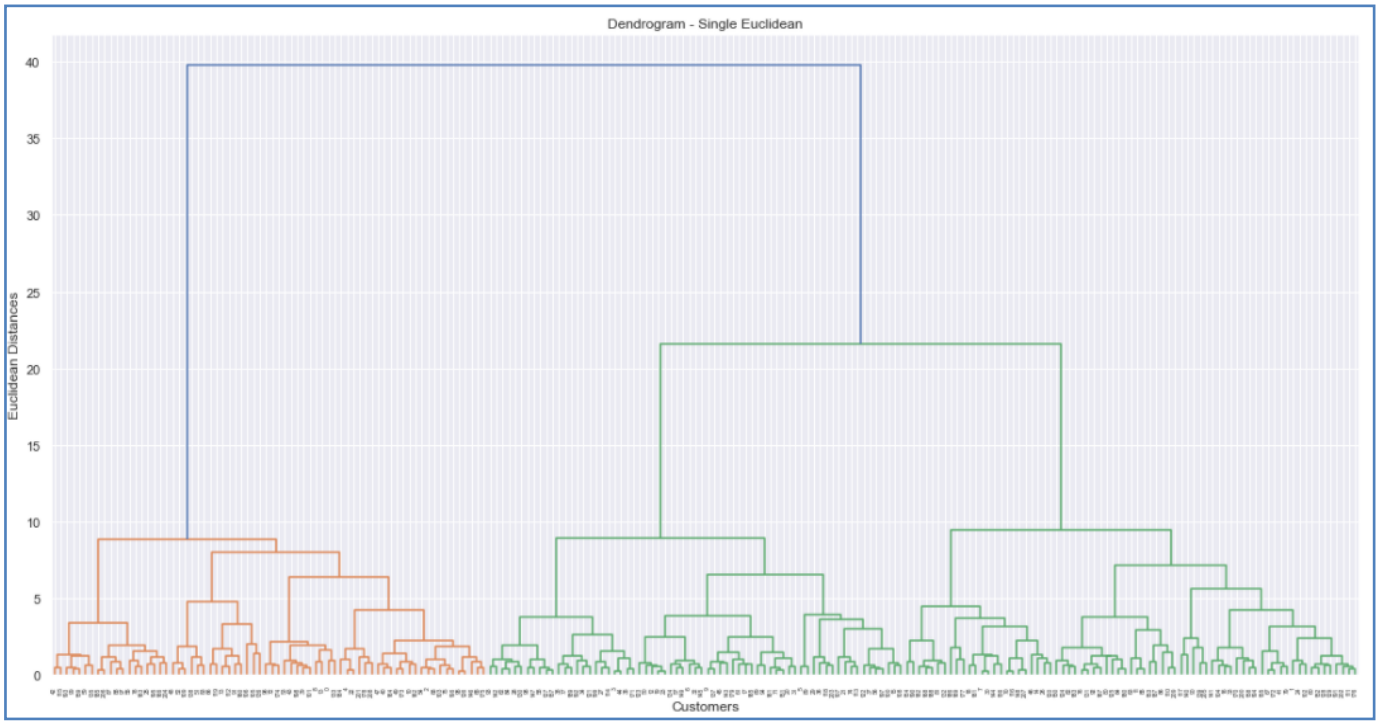


Figure : Dendrogram – Wards Method.

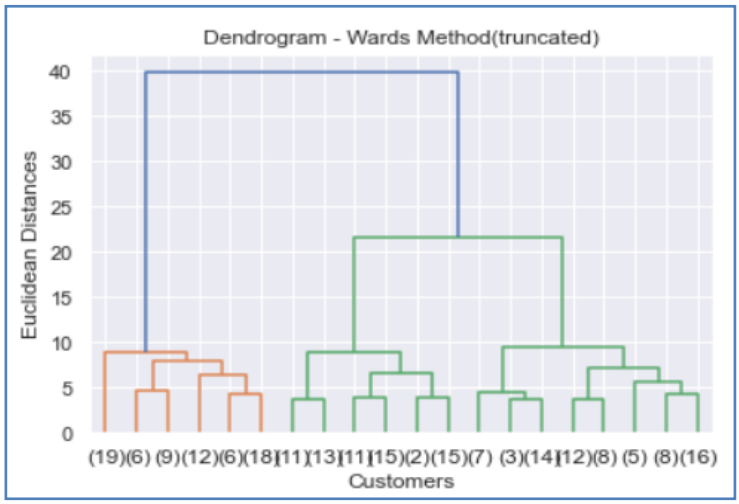


Figure : Dendrogram – Wards Method(truncated)

Euclidean complete linkage shows the 3 clusters with 3 different colors Similarly from the above wards method it formed 3 clusters with orange one and green two can be seen in the truncated wards. Henceforth, the optimum number of cluster can be achieved is 3 clusters.

Head of the data-set with added columnof cluster number for all the values:-

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

Figure : Dendrogram – Wards Method.

Frequency for the clusters aggregated column wise table:-

clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	frequency
1	1.213983	1.217445	0.568505	1.198256	1.130594	-0.040697	1.242686	70
2	-1.024932	-0.999559	-0.972589	-0.881418	-1.088249	0.832836	-0.583025	67
3	-0.223402	-0.250010	0.347508	-0.340041	-0.085328	-0.725360	-0.656511	73

Figure : Dendrogram – Wards Method.

Relationship between the clusters and variables using scatterplot:-

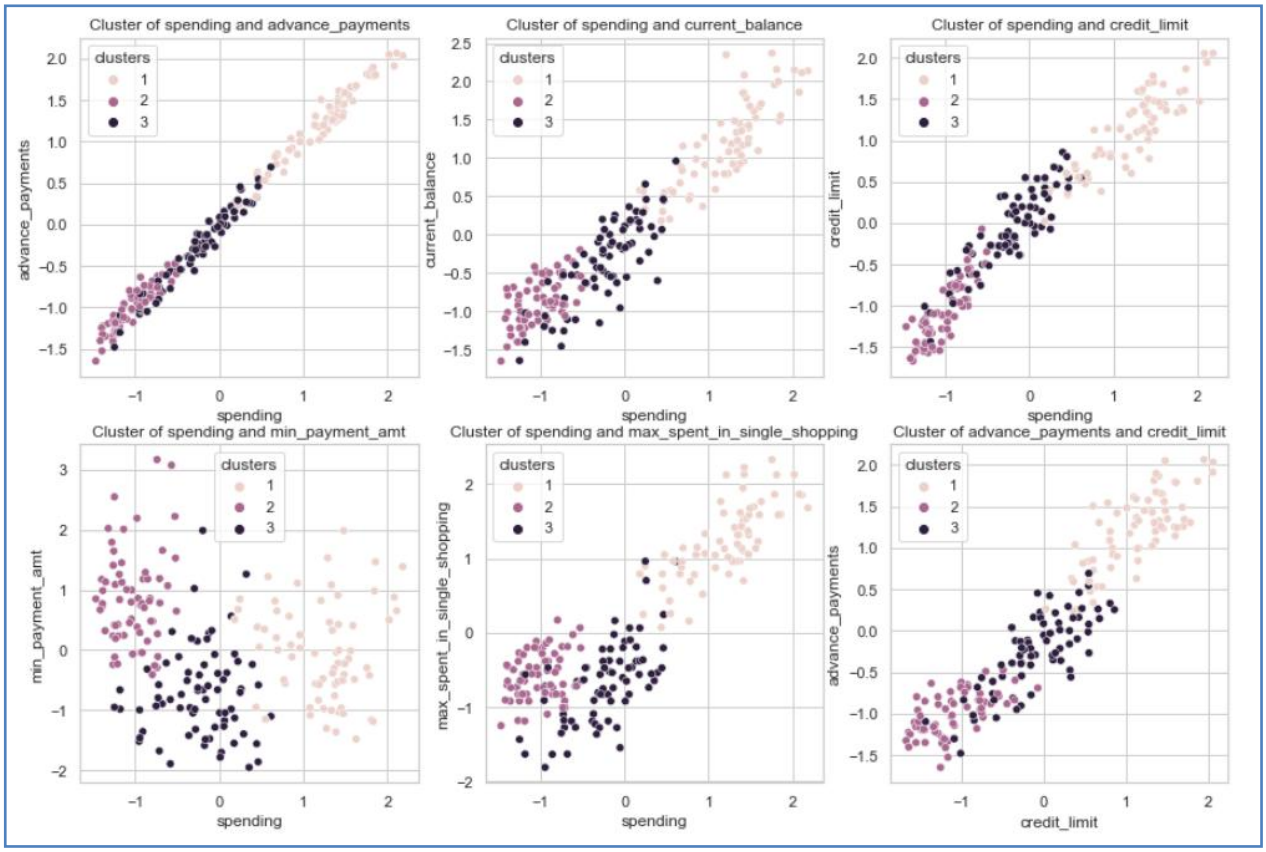


Figure : Hierarchical clusters

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. 7

K-Means Clustering

It is an iterative algorithm that tries to segregate the dataset into kpre-defined distinct non-overlapping sub-groups (clusters) where each data points belongs to one of the gloup or so called cluster. The benefits of it that it tries to form the intra-cluster as similar data points and also keeps the clusters as different or far as possible.

K-meaning the target variable which refers to the number of centroid in the given datset.

Applying the K-Means clustering on the scaled data which can be done using two main techniques and finding the K-Value as optimum number:-

1. The Elbow Method
2. The Silhouette Method

Firstly tried applying k-means clustering with number of clusters as 3. Below are the labels after applying k-means clustering.

```
array([1, 3, 1, 2, 1, 2, 2, 3, 1, 2, 1, 3, 2, 1, 3, 2, 3, 2, 3, 2, 2, 2,
       1, 2, 3, 1, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 1, 1, 3, 1, 1,
       2, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 3, 2, 2, 3, 3, 1,
       1, 3, 1, 2, 3, 2, 1, 1, 2, 1, 3, 2, 1, 3, 3, 3, 3, 1, 2, 3, 3, 1,
       1, 2, 3, 1, 3, 2, 2, 1, 1, 1, 2, 1, 2, 1, 3, 1, 3, 1, 1, 2, 2, 1,
       3, 3, 1, 2, 2, 1, 3, 3, 2, 1, 3, 2, 2, 2, 3, 3, 1, 2, 3, 3, 2, 3,
       3, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 3, 2, 1, 2, 3, 2, 3, 2, 3, 3,
       3, 3, 3, 2, 3, 1, 1, 2, 1, 1, 1, 2, 1, 3, 3, 3, 3, 2, 3, 1, 1, 1,
       3, 3, 1, 2, 3, 3, 3, 3, 1, 1, 3, 3, 3, 2, 3, 3, 2, 1, 3, 1, 1, 2,
       1, 2, 3, 1, 3, 2, 1, 3, 1, 3, 1, 3], dtype=int32)
```

Table: -K-Means Labels

From the below WSS scores, we can see that for cluster 1 the score is 1469.99 and the score for 2 clusters dropped to 659.17 which is a significant difference in the scores.

```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.38509060801107,
 327.0510614531657,
 290.1034759156124,
 265.8065121386725,
 245.06913618849228,
 225.16666685375478,
 212.36896604963718]
```

Table: - WSS Score

For cluster 3, the score is 430.65 whereas from cluster 3 to cluster 10 we see that there is no significant decrease in the scores. Hence we arrive at the optimum no of clusters as 3.

Elbow Method

Looking at the Elbow plot we can easily assume that there wouldn`t be any substantial decrease in the scores. Hence 3rd score can be choosen as the optimum one.

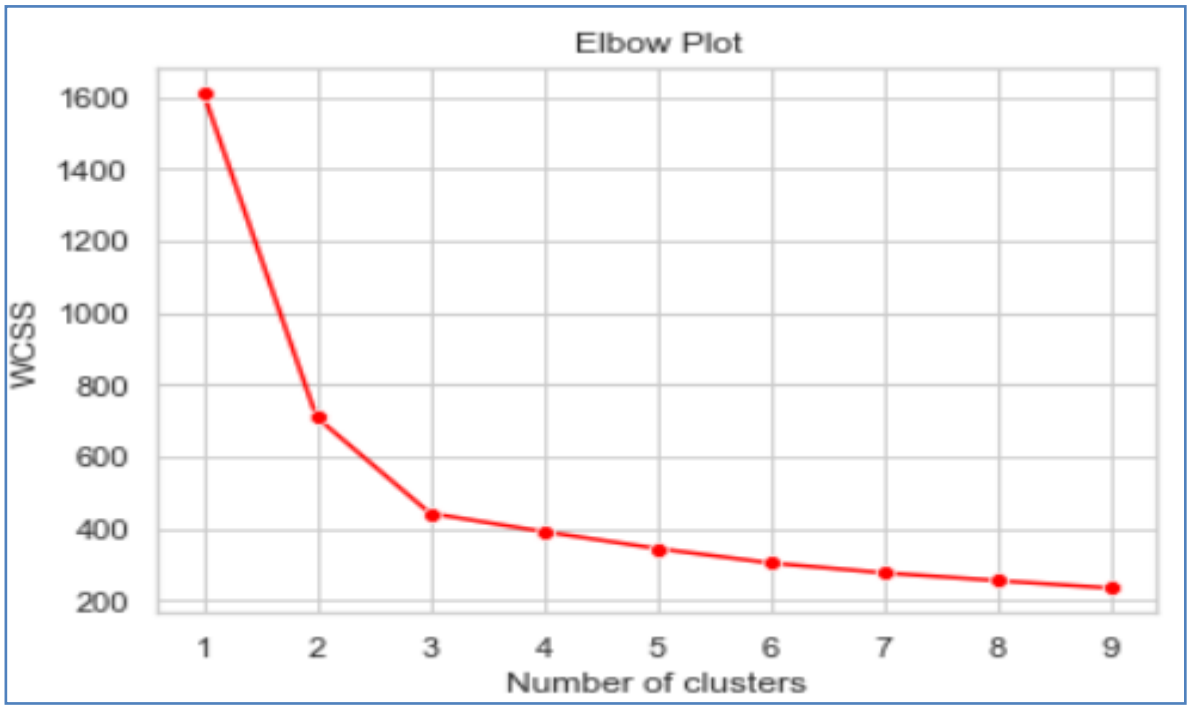


Figure : The Elbow Method

Silhouette Method

Silhouette is a different method to determine optimal number of clusters for given dataset. It defines as coefficient of measure of how similar an observation to its own cluster compared to that of other clusters. The range of silhouette coefficient varies between -1 to 1. Having the 3rd score in this method is good coefficient score for the given dataset which is 0.45 Approx.

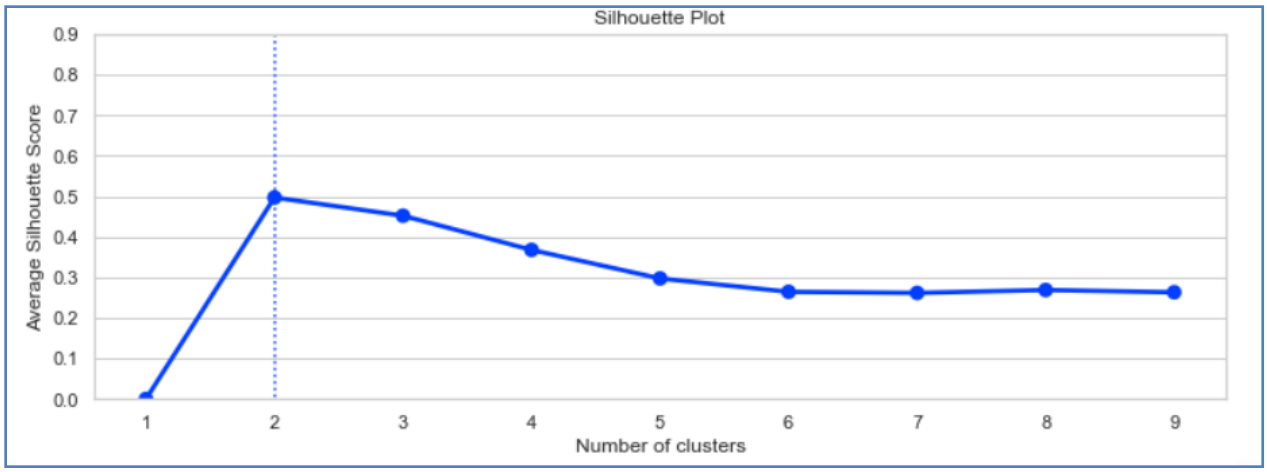


Figure : The Silhouette Method

Showing the frequency and K-means cluster appending on the original scaled dataset:-

```
0    72
1    67
2    71
Name: k_clusters
```

Figure : Frequency table of the clusters

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

Figure : K – Means cluster merged with original dataframe.

For easy understanding shown in the scatter plot for the 3 clusters separating each other with distinct colors (Black, Pink and White):-

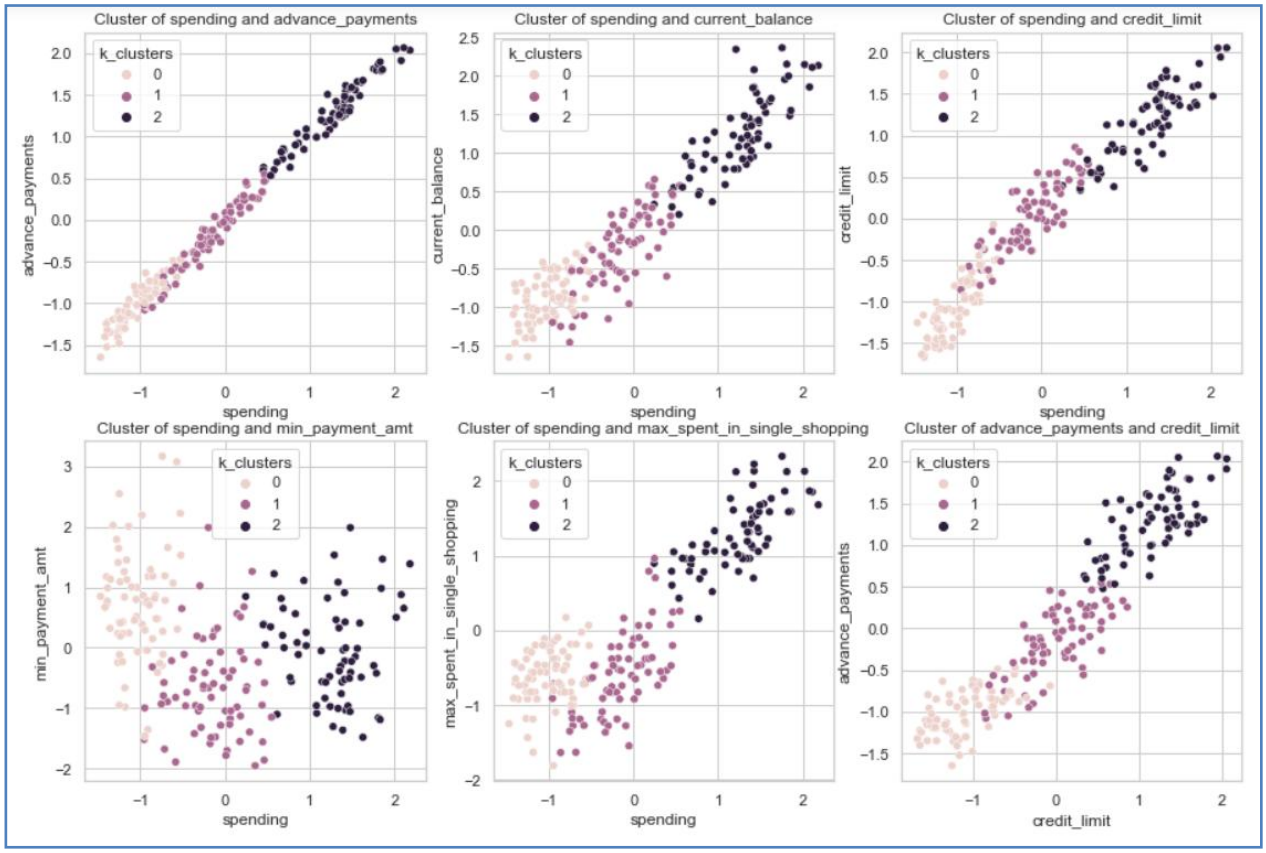


Figure : K – K-Means cluster scatter plot.

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

Below given table shows the levels of different customers differentiated based on clusters and later analysis has been given based on the table.

k_clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	frequency
0	-1.030253	-1.006649	-0.964905	-0.897685	-1.085583	0.694804	-0.624809	72
1	1.256682	1.261966	0.560464	1.237883	1.164852	-0.045219	1.292308	67
2	-0.141119	-0.170043	0.449606	-0.257814	0.001647	-0.661919	-0.585893	71

Figure : Frequency of K – Means based on 3 cluster.

1. Customers under K_cluster 0 i.e. 1, have a high spending, current balance, credit_limit and max_spent_in_single_shopping which clearly shows that they are premium high-net worth customers who make expensive purchases on their credit cards.
2. Customers under K_cluster 1 i.e. 2 have the least spending and credit_limits compared to other clusters. This signifies that they are customers who have recently bought credit cards or youths who have started working recently. Bank can provide customized offers to this segment to promote more spending on credit cards.
3. Customers under K_cluster 2 i.e. 3 have a relatively lesser spending, current balance, credit_limit and max_spent_in_single_shopping which indicate that they are upper middle class customers. The bank can provide promotional offers to this segment such that they increase their spending and are potential customers who can move into premium segments.

2. Problem 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).

Exploratory Data analysis

Sample of Dataset

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

Figure : Frequency of K – Means based on 3 cluster.

```
<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: - Info for the dataset

Duplicated in the dataset

There are 139 rows with duplicate values in the original dataset. Hence it needs to be removed from the dataset. The tables given below shows the value before and after the duplicated been removed so the total count dropped down to 2861 from 3000.

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

Table: - Original dataset description

	count	mean	std	min	25%	50%	75%	max
Age	2861.0	38.204124	10.678106	8.0	31.0	36.00	43.00	84.00
Commision	2861.0	15.080996	25.826834	0.0	0.0	5.63	17.82	210.21
Duration	2861.0	72.120238	135.977200	-1.0	12.0	28.00	66.00	4580.00
Sales	2861.0	61.757878	71.399740	0.0	20.0	33.50	69.30	539.00

Table: - Dataset after duplicates removed

Description of the data

From the above table below are the observations:

1. Age seems to be normally distributed. The minimum age is 8 yrs and max age is 84 yrs which shows that we might have the correct data on Age. Average age of customers in the dataset is 38 yrs.
2. Commission received for tour insurance is widely spread. The average commission is 14.52 however the minimum commission received is 0 and maximum commission received is 210.21 and 95% of data lies within 63.21 which indicates that there might be outliers.
3. Duration of the tour ranges from 134 days to 4580 days which means there are outliers. 95% of data lies within 367 days whereas maximum days shows 4580 days which is close to 12 years. It is surprising to see that someone would go for a tour for 12 years. Need to consult the business to validate the data. For now we will go with further analysis as is. The minimum duration shows -1 which cannot be the case in reality. Hence we might drop this row or impute it mean value to treat bad data.

	count	mean	std	min	25%	50%	75%	max
Age	2861.0	38.204124	10.678106	8.0	31.0	36.00	43.00	84.00
Commision	2861.0	15.080996	25.826834	0.0	0.0	5.63	17.82	210.21
Duration	2861.0	72.120238	135.977200	-1.0	12.0	28.00	66.00	4580.00
Sales	2861.0	61.757878	71.399740	0.0	20.0	33.50	69.30	539.00

Table: - Descriptive statistics

4. Sales figures ranges from 0 to 539 (in 100's). The average sales is amounted to 60.24. There might be outliers as well.
5. Except for Age the other variables seem to be skewed with outliers present.

Univariate analysis

From the figure below we can see all the numerical variables have outliers.

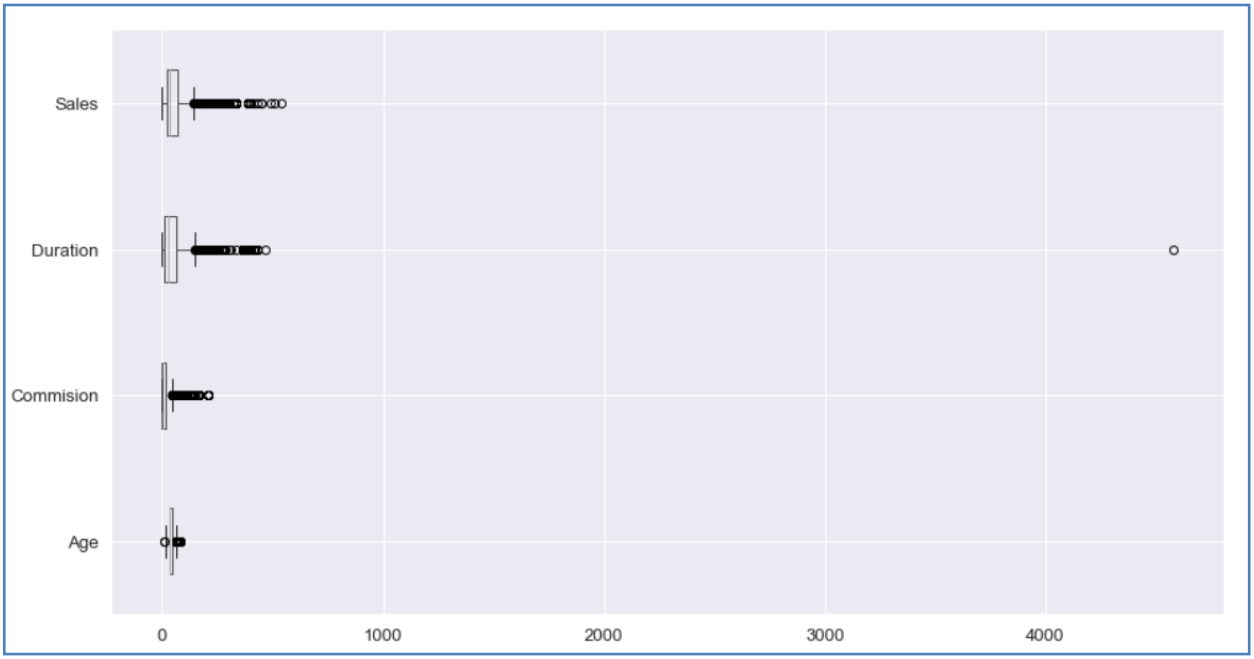


Figure : - Box plot

Distribution plot

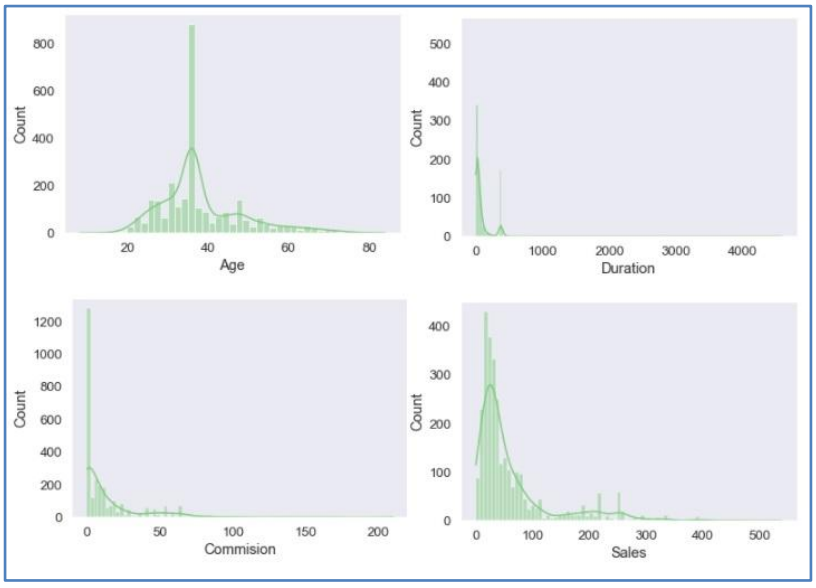


Figure : - Distribution plot

Plotting Numerical variable with claimed status

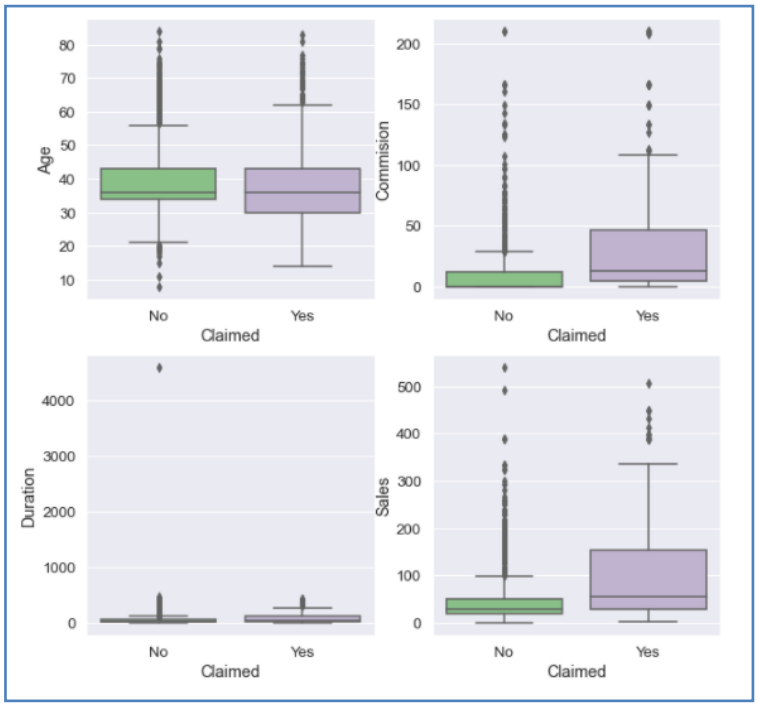


Figure : - Box plot

Except age all the variables showing high skewness. The commission, Duration and sales data points are left skewed and having outliers.

Bivariate Analysis

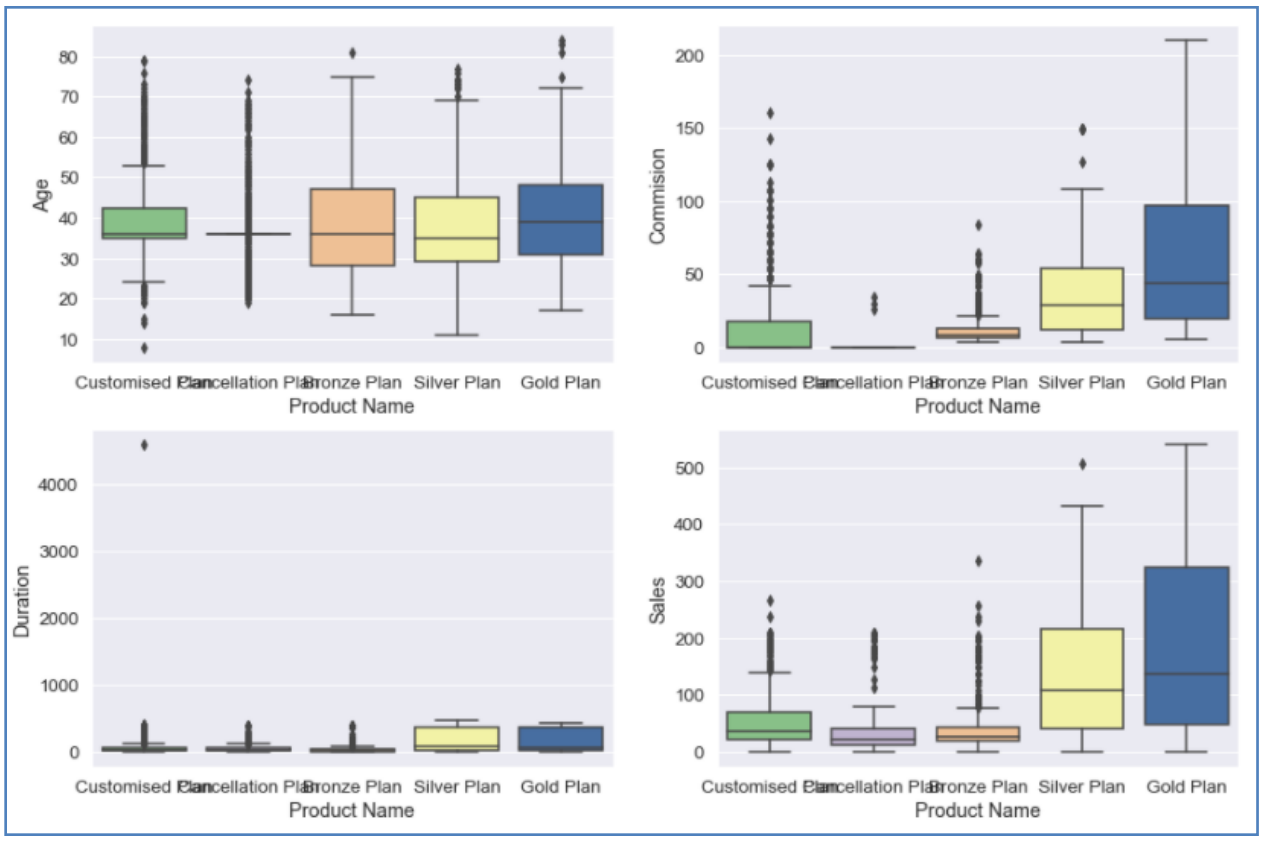


Figure : - Box plot for products and continuous variables

As we can see that box plot of product name and commission column, for the gold plan commission become larger than the other product. Even the sales is maximum for the golden plan may be because agency getting more commission to sale this particular product.

Boxplot age from 30 to 50 preferring airlines type more over travel agency. Rest of the boxplot not showing convincing information. From the boxplot, C2B and CWT agency_code getting more commission over other two agency_code. C2B Agency_Code having highest number of sales over other three agency_code

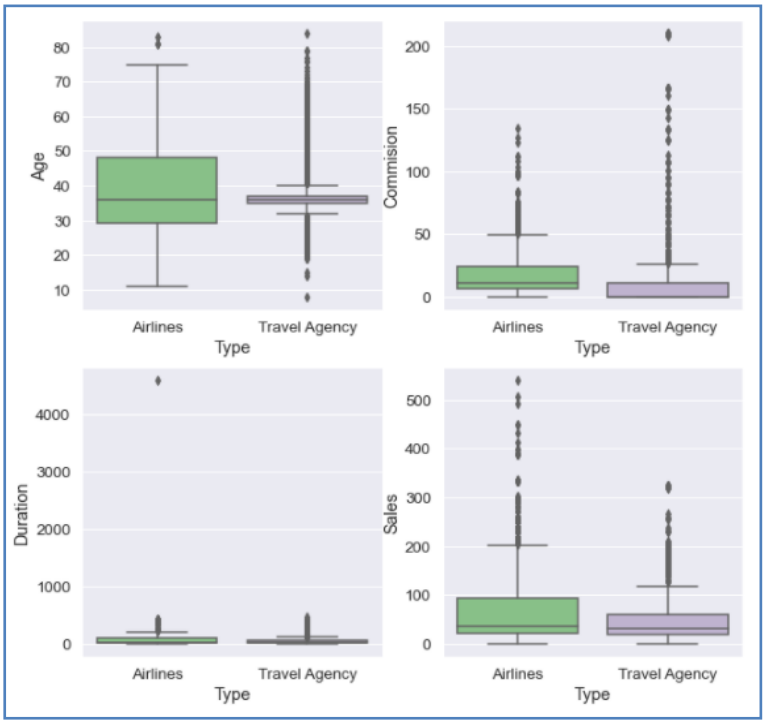


Figure : - Box plot for products and continuous variables

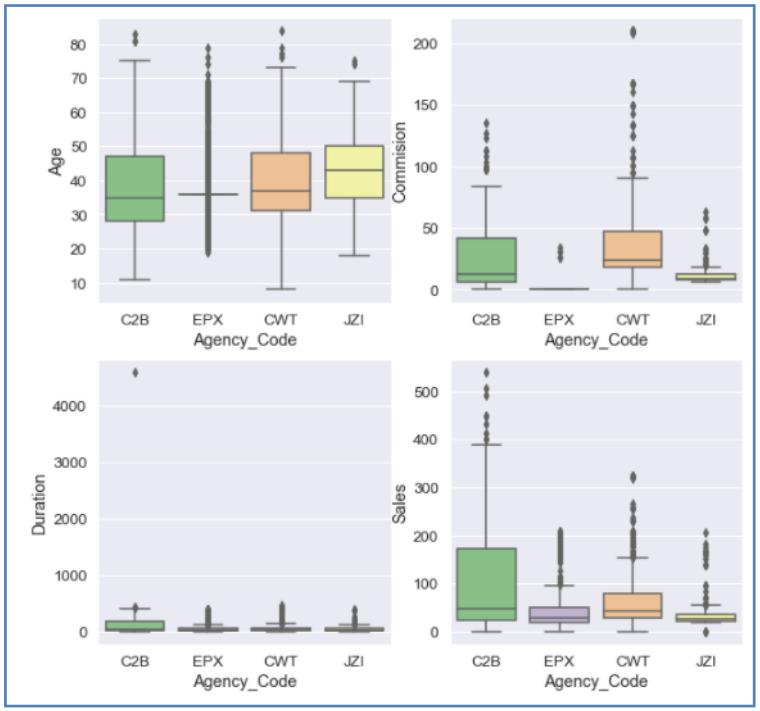


Figure : - Box plot for products and continuous variables

Multivariate Analysis

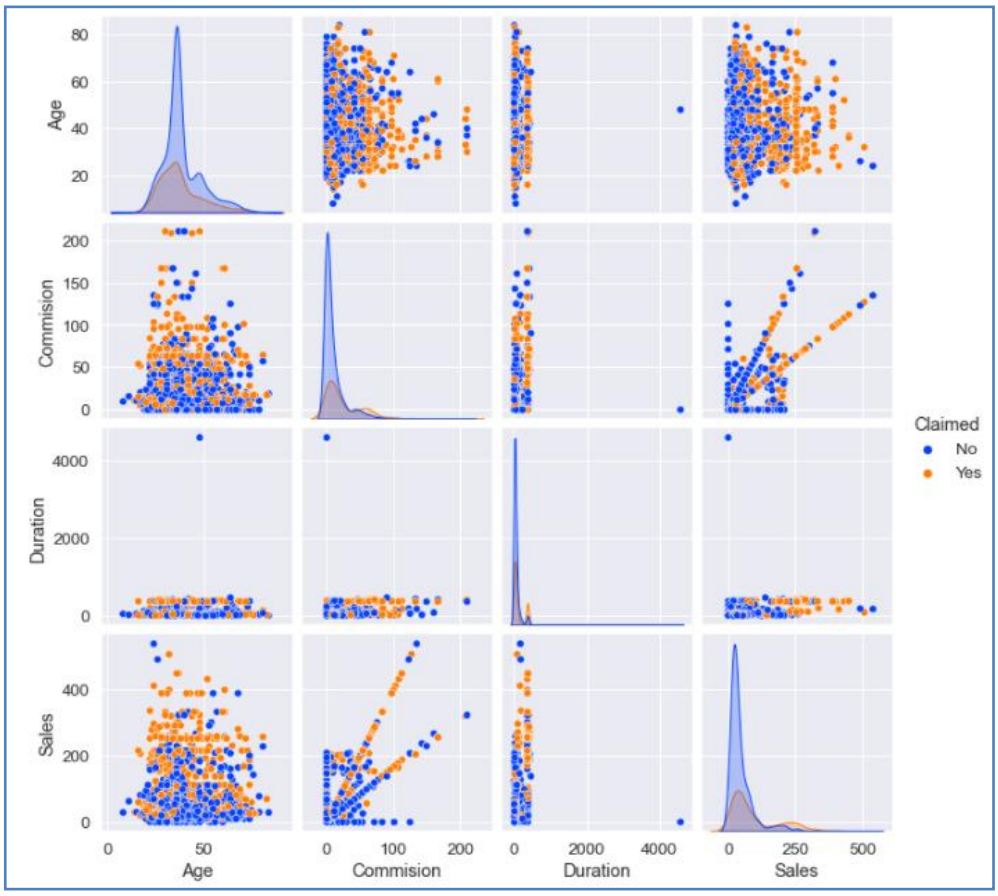


Figure : - Pair plot

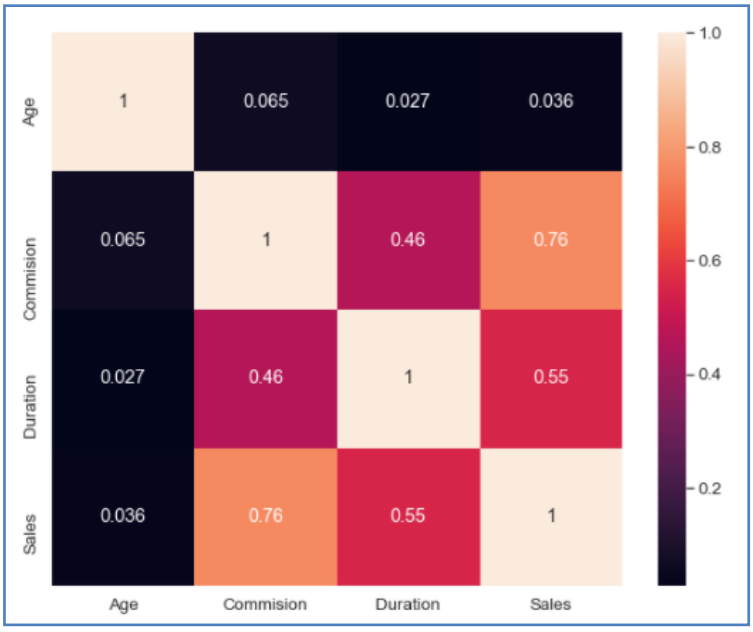


Figure : - Correlation

From the above correlation plot, Continuous columns showing less correlation with each other. Sales and commission column showing high correlation among them. As the sales increase commission increase. It can be figure out from pair plot also.

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

Data Split

Converted the categorical data into numerical form:-

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


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


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


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


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


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

Figure : - Conversion table

The data is split into training 70 % and test 30 % dataset:-

```
Train dataset:  (2002, 9)
Test dataset:  (859, 9)
Train labels:  (2002,)
Test labels:  (859,)
```

Figure : - Label Dimensions

CART

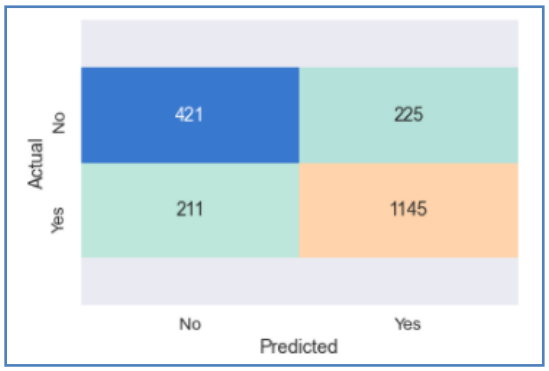


Figure : - Label Dimensions

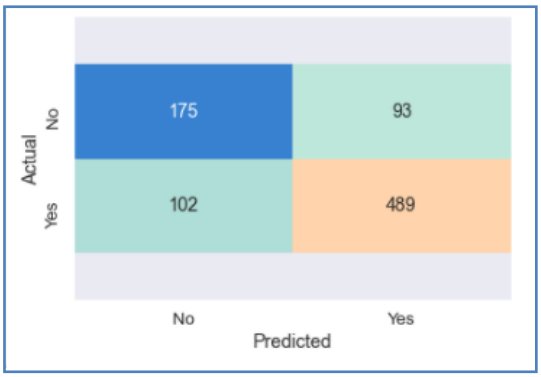


Figure : - Label Dimensions

	precision	recall	f1-score	support
0	0.84	0.84	0.84	1356
1	0.67	0.65	0.66	646
accuracy			0.78	2002
macro avg	0.75	0.75	0.75	2002
weighted avg	0.78	0.78	0.78	2002

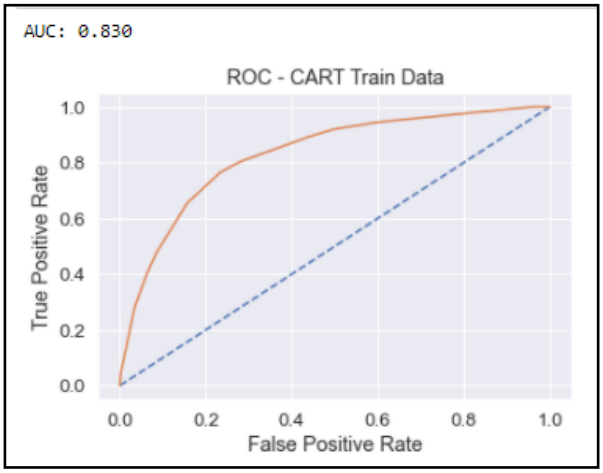
Figure : - MLP classifier training

	precision	recall	f1-score	support
0	0.84	0.83	0.83	591
1	0.63	0.65	0.64	268
accuracy			0.77	859
macro avg	0.74	0.74	0.74	859
weighted avg	0.78	0.77	0.77	859

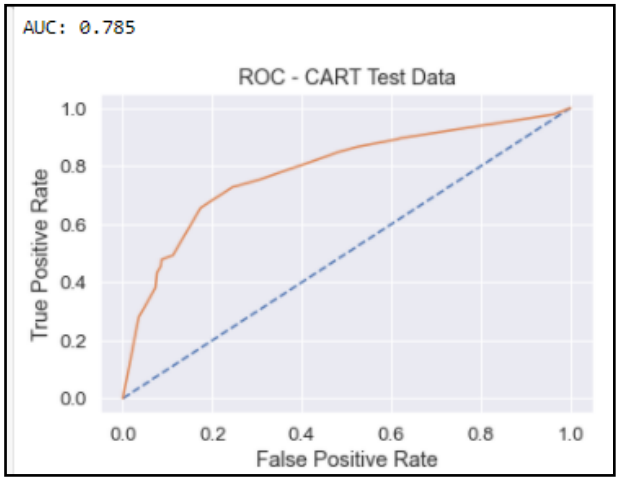
Figure : - MLP classifier test dataset

2.3Performance 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.

AUC and ROC for the training data:-



AUC and ROC for the test data : -



RANDOM FOREST

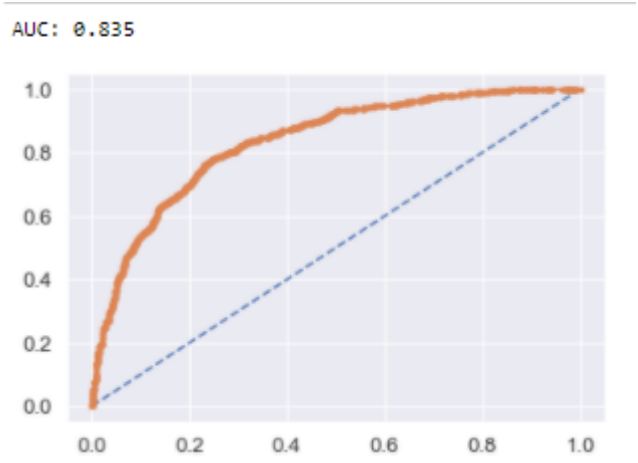


Figure : - Training dataset

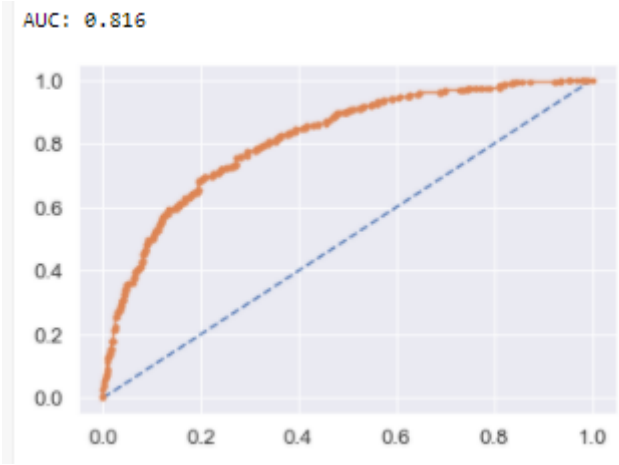


Figure : - Test Dataset

ANN

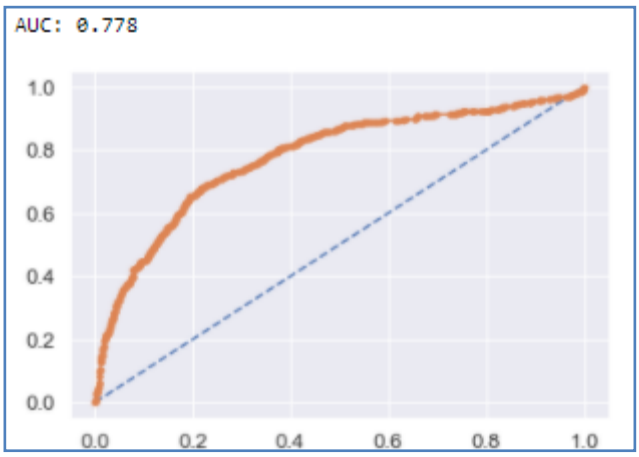


Figure : - AUC and ROC for the training

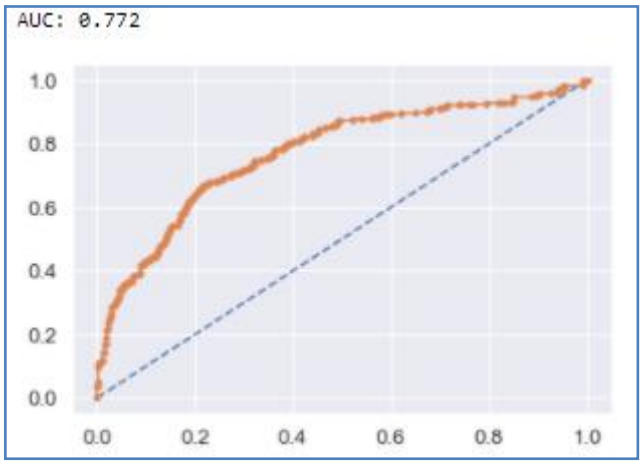


Figure : - AUC and ROC for the test data:-

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

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

```
<matplotlib.legend.Legend at 0x135527af088>
```