DATA MINING

AMRITA JENA

PGP - Data Science and Business Analytics

PGPDSBA Online May_B 2021

Executive Summary

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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. Apply hierarchical clustering to scaled data and identifying the number of optimum clusters using Dendrogram and also to apply K-Means clustering on scaled data and determine optimum clusters.

.

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).

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

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

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.

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

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

	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

Table1-Dataset (first 5 records)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
spending                      210 non-null float64
advance_payments              210 non-null float64
probability_of_full_payment   210 non-null float64
current_balance               210 non-null float64
credit_limit                  210 non-null float64
min_payment_amt               210 non-null float64
max_spent_in_single_shopping  210 non-null float64
dtypes: float64(7)
memory usage: 11.6 KB
```

```
spending                      0
advance_payments              0
probability_of_full_payment   0
current_balance               0
credit_limit                  0
min_payment_amt               0
max_spent_in_single_shopping  0
dtype: int64
```

Table 2- Null value and missing value analysis

- The dataset contains 7 variables and 210
- No missing records were present in the dataset.
- All the variables are numeric type.

	count	mean	std	min	25%	50%	75%	90%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	18.9880	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	16.4540	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.8993	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.2733	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	3.7865	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	5.5376	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.1850	6.5500

Table3- Summary Statistic

From summary descriptive, the data looks good.

- For most of the variable, mean/medium are nearly equal

- I included a 90% to see variations and it looks distributely even.

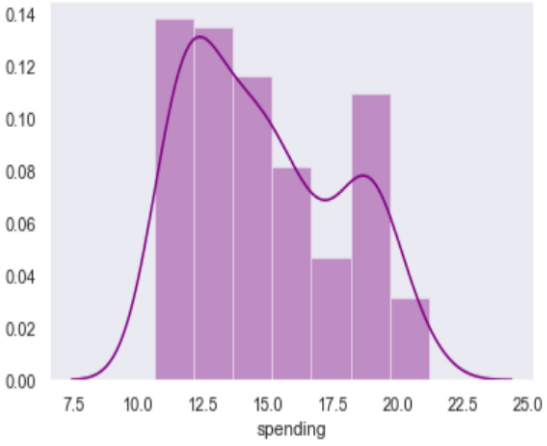
- Std Deviation is observed to be high for spending variable

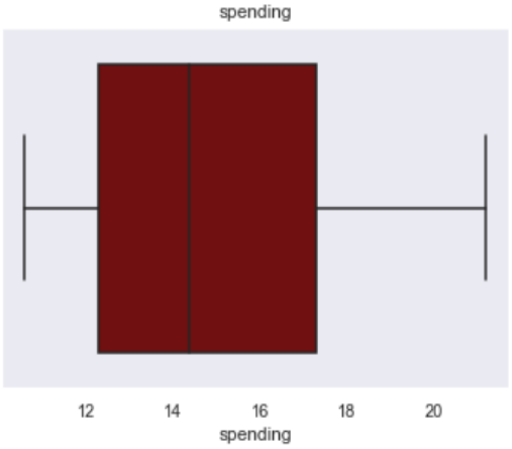
Exploratory Data Analysis Univariate / Bivariate/Multivariate Analysis-

EDA analysis helps us to understand the distribution of data in the dataset. With univariate analysis we can find patterns and we can summarize the data and have understanding about the data to solve our business problem-

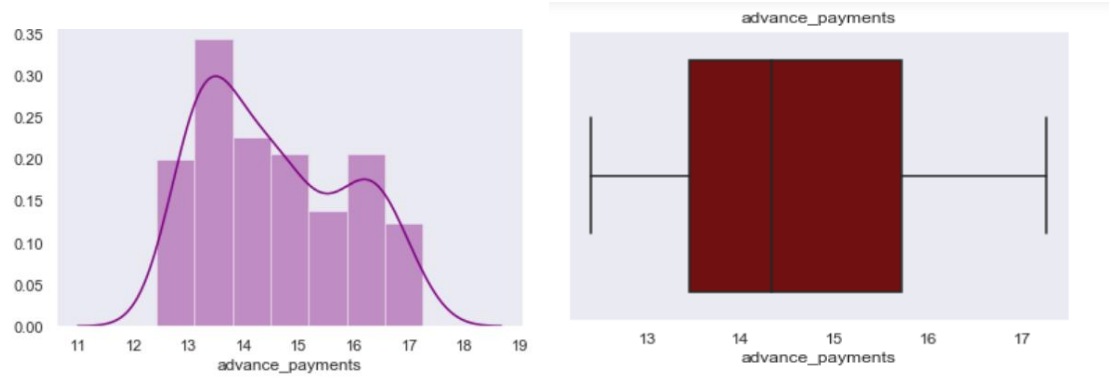
```
max_spent_in_single_shopping     0.561897
current_balance                  0.525482
min_payment_amt                  0.401667
spending                         0.399889
advance_payments                 0.386573
credit_limit                     0.134378
probability_of_full_payment     -0.537954
dtype: float64
```

Table 4- Skewness of variables

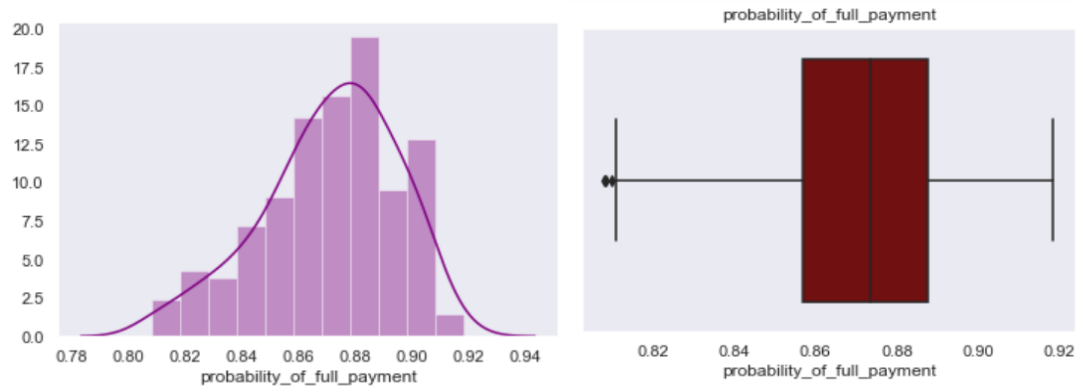




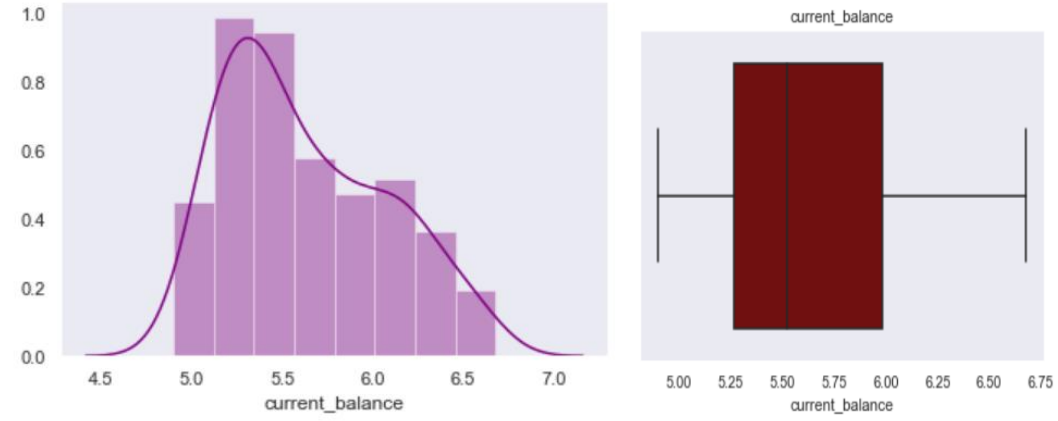
The box plot of the spending variable shows no outliers.
Spending is positively skewed - 0.399889.
The dist plot shows the distribution of data from 10 to 22



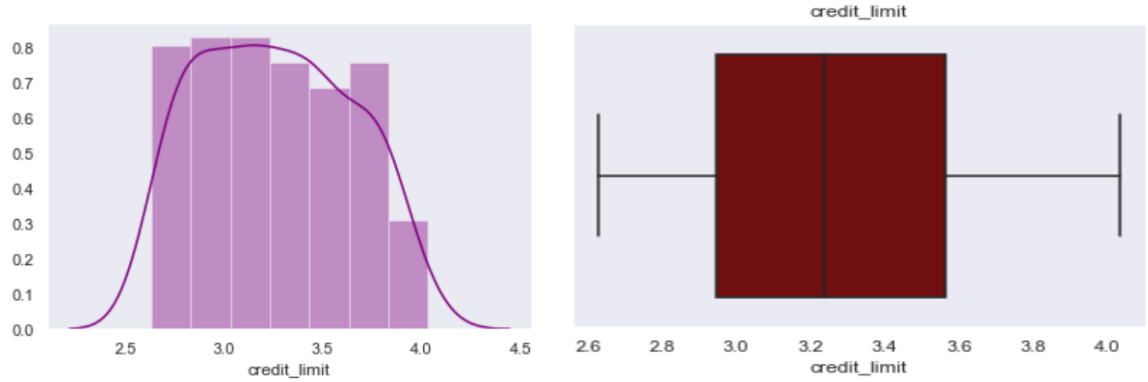
The box plot of the advance payments variable shows no outliers.
Advance payments is positively skewed - 0.386573.
The dist plot shows the distribution of data from 12 to 18



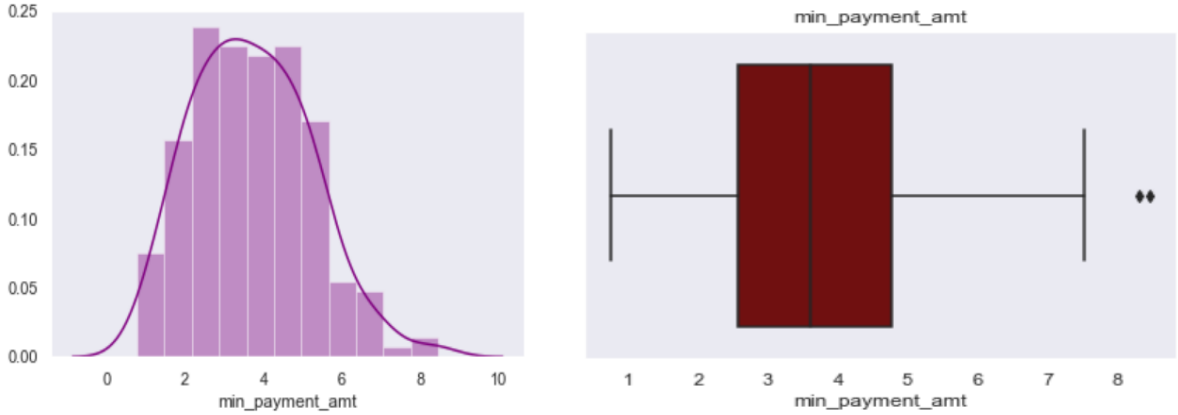
The box plot of the probability of full payment variable shows few outliers.
Probability of full payment is negatively skewed -0.537954
The dist plot shows the distribution of data from 0.80 to 0.92.
The Probability values is good above 80%



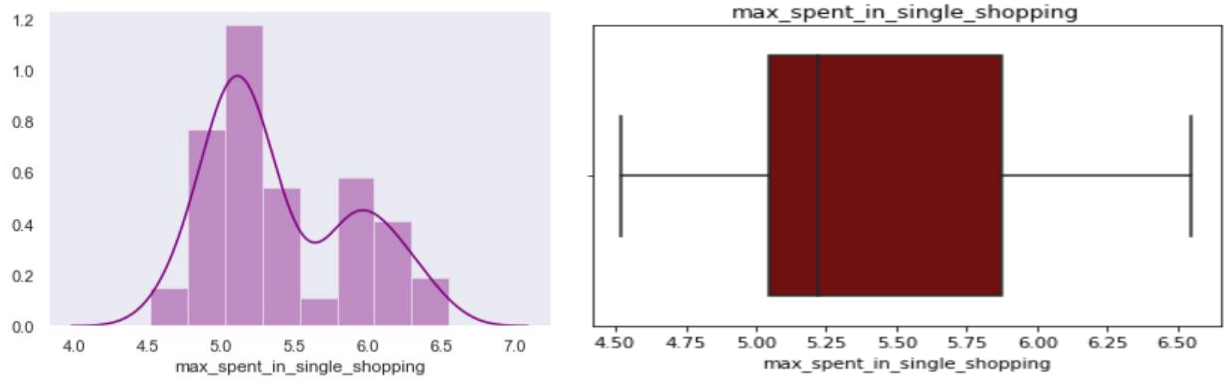
The box plot of the current balance variable shows no outliers.
Current balance is positively skewed - 0.525482
The dist plot shows the distribution of data from 5.0 to 6.5.



The box plot of the credit limit variable shows no outliers.
Credit limit is positively skewed - 0.134378
The dist plot shows the distribution of data from 2.5 to 4.0



The box plot of the min payment amount variable shows few outliers.
Min payment amount is positively skewed - 0.401667
The dist plot shows the distribution of data from 1 to 8



The box plot of the max spent in single shopping variable shows no outliers.
Max spent in single shopping is positively skewed - 0.561897
The dist plot shows the distribution of data from 4.5 to 6.5

Figure-1-Distribution of Data and Outliers

We could also visualize distribution of variables by using KDE Plot represents the Kernel Density Estimate KDE is used for visualizing the Probability Density of a continuous variable.
KDE demonstrates the probability density at different values in a continuous variable

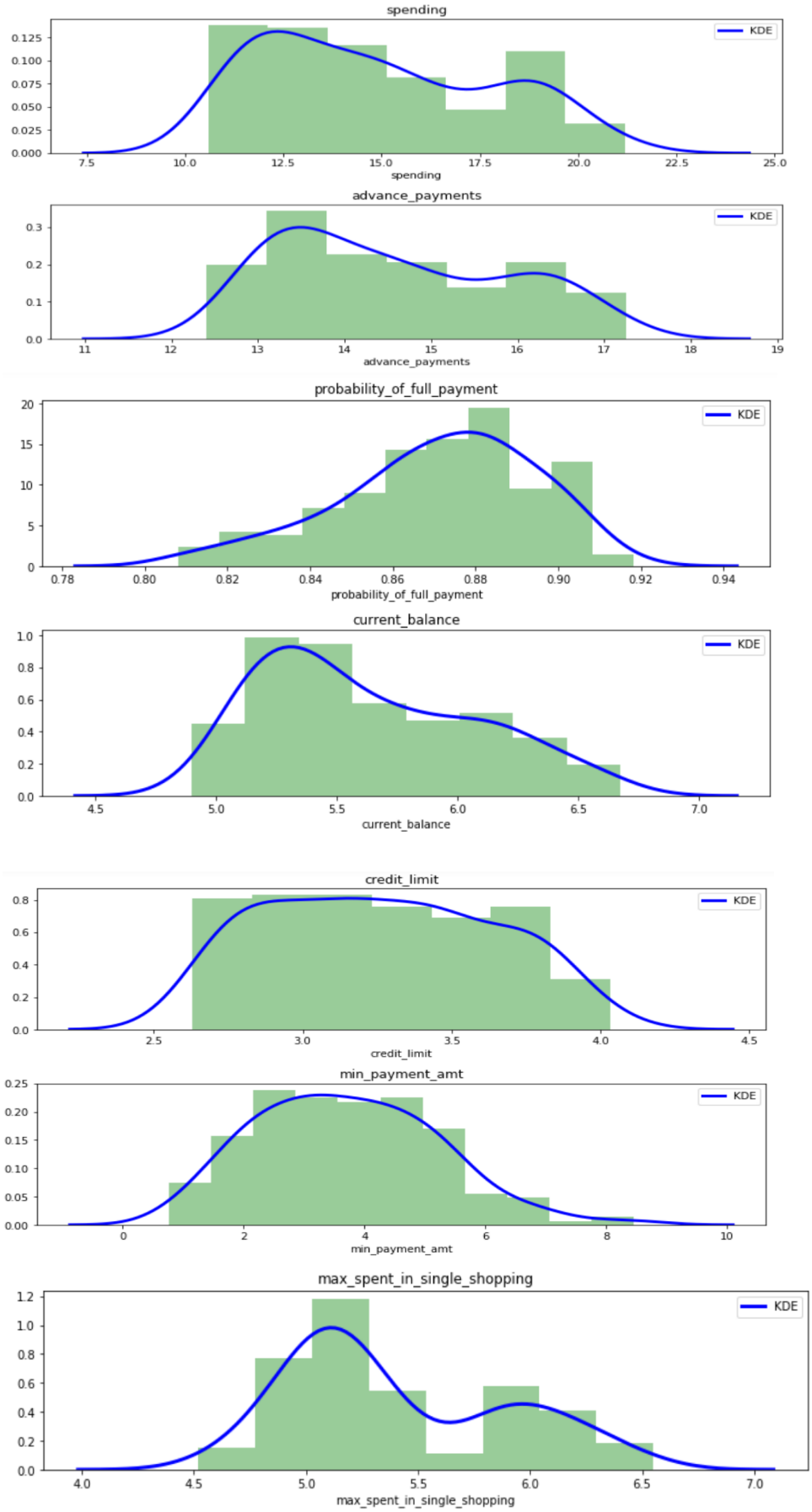
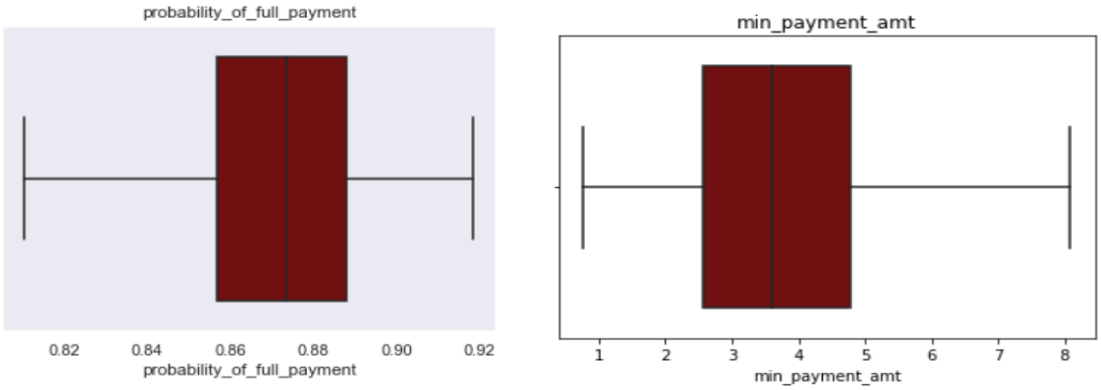


Figure.2- KDE Plot Distribution of Data

- Distribution is skewed to right tail for all the variables execpt for the probability_of_full_payment variable, which has left tail.

After treating outliers for probability_of_full_payment and Min_payment_amt variable we could observe theat the outliers have been removed.

Figure-3 Treated Outliers

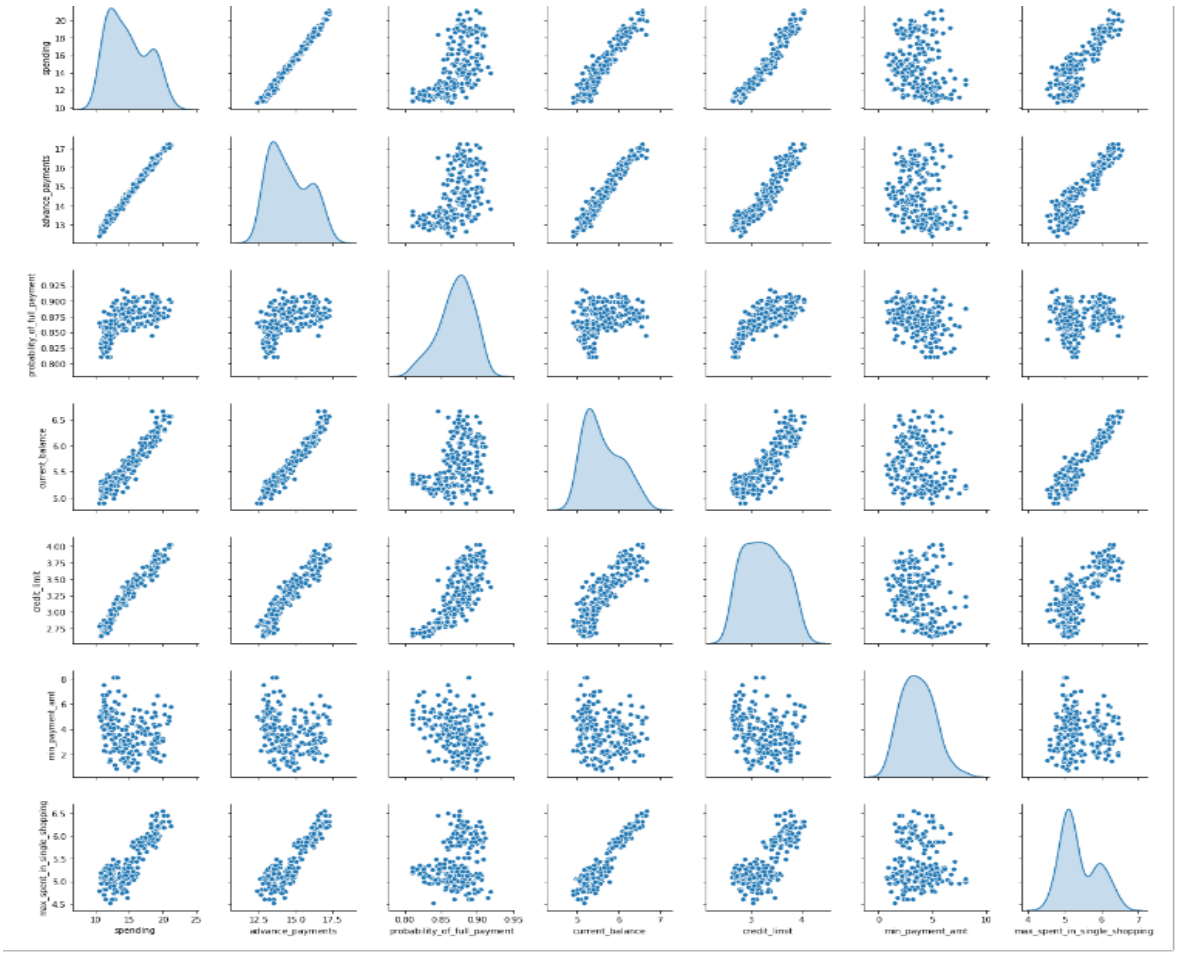


Multivariate Analysis-

- Strong positive correlation between
- spending & advance_payments,
- advance_payments & current_balance,
- credit_limit & spending
- spending & current_balance
- credit_limit & advance_payments
- max_spent_in_single_shopping current_balance are observed.

Table-6- Multivariate Analysis-

		correlation
spending	advance_payments	0.994341
advance_payments	current_balance	0.972422
credit_limit	spending	0.970771
spending	current_balance	0.949985
credit_limit	advance_payments	0.944829
max_spent_in_single_shopping	current_balance	0.932806
advance_payments	max_spent_in_single_shopping	0.890784
spending	max_spent_in_single_shopping	0.863693
current_balance	credit_limit	0.860415
probability_of_full_payment	credit_limit	0.761635
max_spent_in_single_shopping	credit_limit	0.749131
spending	probability_of_full_payment	0.608288
advance_payments	probability_of_full_payment	0.529244
current_balance	probability_of_full_payment	0.367915
probability_of_full_payment	min_payment_amt	0.331471



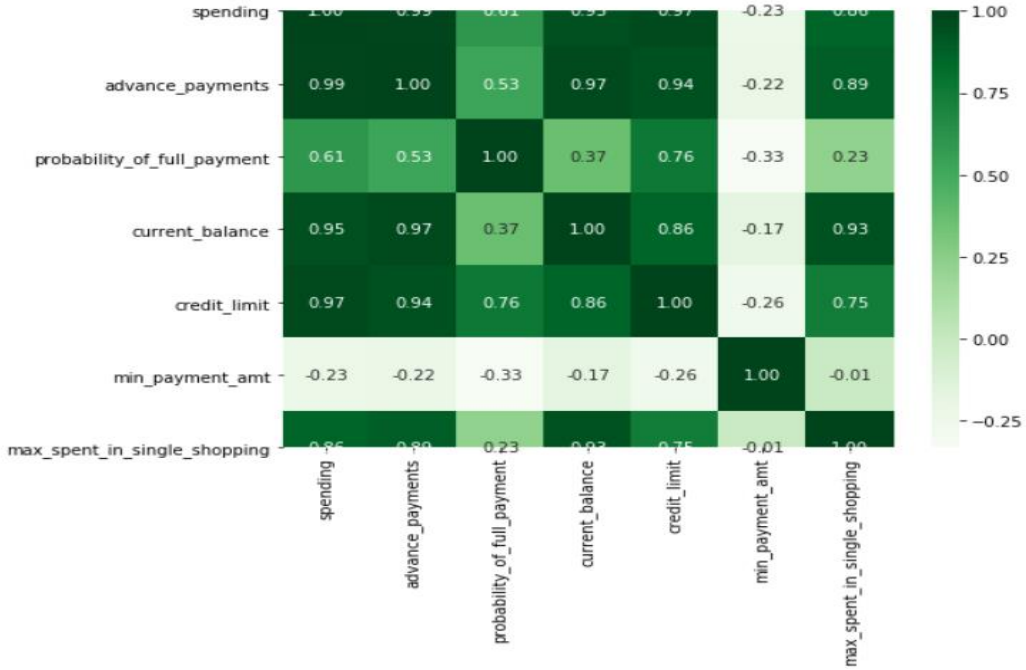


Figure-4-Pairplot and heatmap

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

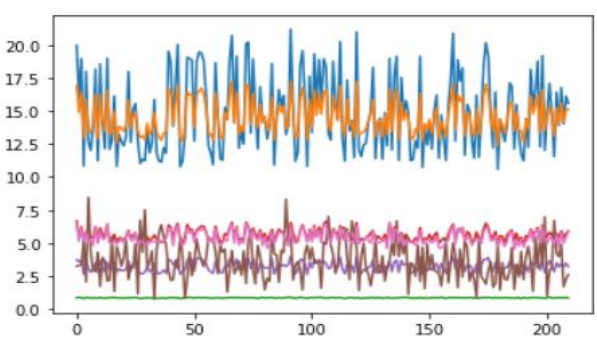
Yes Scaling is necessary as often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set spending, advance_payments are in different values and this may get more weightage. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of

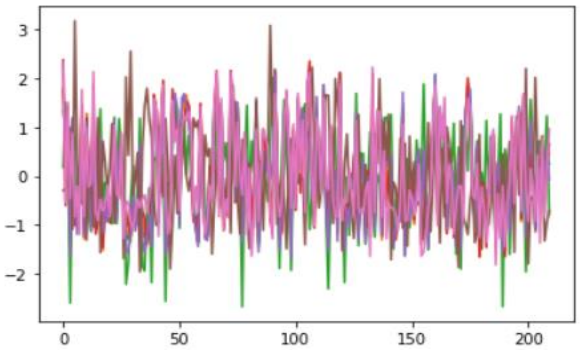
data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale.

Figure-4-Scaled Data



Data before scaling



Data after scaling

	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

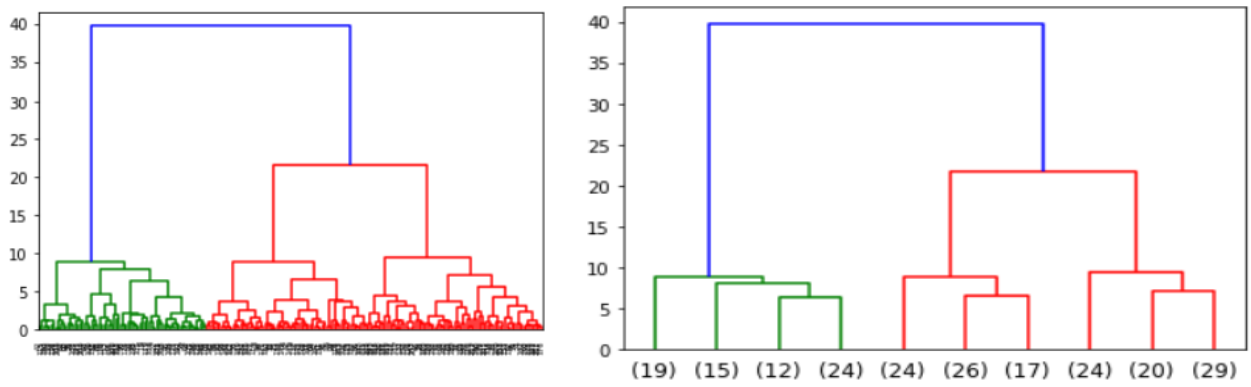
Table 6- Scaled Dataset

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

We have to perform Hierarchical clustering b applying both – ward's method & average method . By choosing ward's method to the scaled data.

WARDS METHOD-

Figure-5-Ward Method Dendrogram1 and Dendrogram2



Dendrogram 1 indicates all the data points have clustered to different clusters by wards method. To find the optimal number cluster through which we can solve our business objective we use truncate mode = lastp. Wherein we can give last p = 10 according to industry set base value and we get dendrogram 2.

Now, we can understand all the data points have clustered into 3 clusters. Next to map these clusters to our dataset we can use fclusters Criterion we can give "maxclust".

```
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)
```

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

Now, we can look at the cluster frequency in our dataset,

```
1    70
2    67
3    73
```

wardlink	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	70
2	11.872388	13.257015	0.848155	5.238940	2.848537	4.940302	5.122209	67
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178	73

Table7-Ward Cluster Result

AVERAGE METHOD-

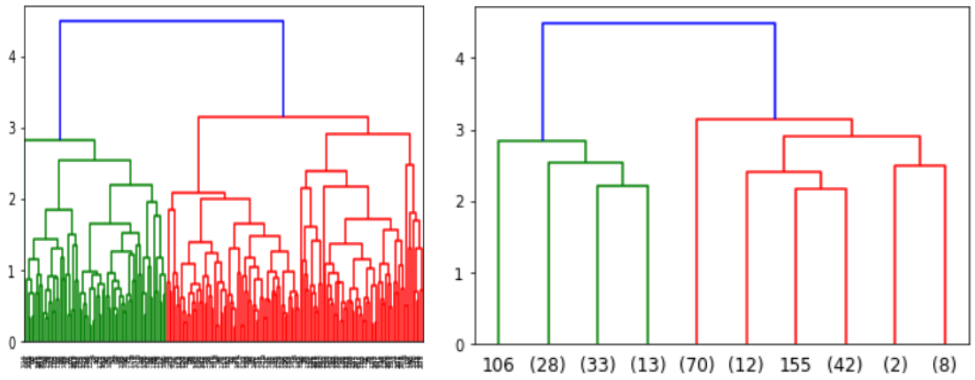


Figure-6-Average Method Dendrogram1 and Dendrogram2

Dendrogram 3 indicates all the data points have clustered to different clusters by wards method. To find the optimal number cluster through which we can solve our business objective we use truncate mode = lastp. Wherein we can give last p = 10 according to industry set base value and we get dendrogram 4.

Now, it could be observed all the data points have clustered into 3 clusters. Next to allot these clusters to the dataset I used fclusters Criterion and gave “maxclust”.

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

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	wardlink	link_average
19.94	16.92	0.875200	6.675	3.763	3.252	6.550	1	1
15.99	14.89	0.906400	5.363	3.582	3.336	5.144	3	3
18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1	1
10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2	2
17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1	1

Now, we can look at the cluster frequency in our dataset,

```
1    75
2    70
3    65
```

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	wardlink	Freq
link_average									
1	18.129200	16.058000	0.881595	6.135747	3.648120	3.650200	5.987040	1.213333	75
2	11.916857	13.291000	0.846845	5.258300	2.846000	4.619000	5.115071	2.114286	70
3	14.217077	14.195846	0.884869	5.442000	3.253508	2.759007	5.055569	2.830769	65

Table 8-Average Cluster Result

Inference:

Both the method are almost similar means, minor variation occurred which is obious.
There was not too much variations from both methods Cluster grouping based on the dendrogram, 3 or 4 looked good after doing further analysis, and based on the dataset had gone for 3 group cluster And three group cluster solution gives a pattern based on high/medium/low spending with max_spent_in_single_shopping and probability_of_full_payment .

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.

As per industrial standard I decide to give n_clusters = 3 and looked at the distribution of clusters according to the n clusters.
I applied K-means technique to the scaled data and calculated WSS for other values of K and applied Elbow Method.

```
[1469.9999999999998,
 659.171754487041,
 430.6589731513006,
 371.38509060801096,
 327.21278165661346,
 289.31599538959495,
 262.98186570162267,
 241.81894656086033,
 223.91254221002725,
 206.39612184786694]
```

WSS reduces as K keeps increasing

Table8- WSS

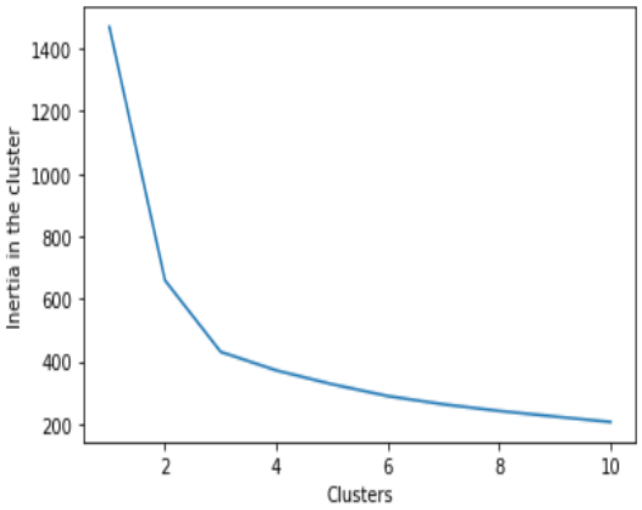


Figure-7-Elbow curve Method

Elbow method is one of the most famous methods by which we can select the right value of k and boost our model performance. Elbow plot was generated and it could be observed that the cut off point could be 3 or 4.

So, I generated the silhouette score for k=3 and k=4 to decide on the no of clusters to pick.

K=3 - 0.4007270552751299 k=4- 0.32757426605518075

silhouette score is better for 3 clusters than for 4 clusters. So, final clusters will be 3.

Then we did 3 Cluster Solution-

Table-9-K Means cluster profiling

```
array([0, 2, 0, 1, 0, 1, 1, 2, 0, 1, 0, 2, 1, 0, 2, 1, 2, 1, 1, 1, 1, 1,
       0, 1, 2, 0, 2, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 0, 0, 2, 0, 0,
       1, 1, 2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 2, 1, 1, 2, 2, 0,
       0, 2, 0, 1, 2, 1, 0, 0, 1, 0, 2, 1, 0, 2, 2, 2, 2, 0, 1, 2, 0, 2,
       0, 1, 2, 0, 2, 1, 1, 0, 0, 0, 1, 0, 2, 0, 2, 0, 2, 0, 0, 1, 1, 0,
       2, 2, 0, 1, 1, 0, 2, 2, 1, 0, 2, 1, 1, 1, 2, 2, 0, 1, 2, 2, 1, 2,
       2, 0, 1, 0, 0, 1, 0, 2, 2, 2, 1, 1, 2, 1, 0, 1, 2, 1, 2, 1, 2, 2,
       1, 2, 2, 1, 2, 0, 0, 1, 0, 0, 0, 1, 2, 2, 2, 1, 2, 1, 2, 0, 0, 0,
       2, 1, 2, 1, 2, 2, 2, 2, 0, 0, 1, 2, 2, 1, 1, 2, 1, 0, 2, 0, 0, 1,
       0, 1, 2, 0, 2, 1, 0, 2, 0, 2, 2, 2])
```

Proportion of labels classified and cluster profiling-

0	67
1	72
2	71

kmean_cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	wardlink	link_average	cluster	Frequency
0	18.5	16.2	0.9	6.2	3.7	3.6	6.0	1.0	1.0	0.0	67
1	11.9	13.2	0.8	5.2	2.8	4.7	5.1	2.1	2.1	1.0	72
2	14.4	14.3	0.9	5.5	3.3	2.7	5.1	2.9	2.7	2.0	71

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

wardlink	1	2	3
spending	18.371429	11.872388	14.199041
advance_payments	16.145429	13.257015	14.233562
probability_of_full_payment	0.884400	0.848155	0.879190
current_balance	6.158171	5.238940	5.478233
credit_limit	3.684629	2.848537	3.226452
min_payment_amt	3.639157	4.940302	2.612181
max_spent_in_single_shopping	6.017371	5.122209	5.086178
Freq	70.000000	67.000000	73.000000

link_average	1	2	3
spending	18.129200	11.916857	14.217077
advance_payments	16.058000	13.291000	14.195846
probability_of_full_payment	0.881595	0.846845	0.884869
current_balance	6.135747	5.258300	5.442000
credit_limit	3.648120	2.846000	3.253508
min_payment_amt	3.650200	4.619000	2.759007
max_spent_in_single_shopping	5.987040	5.115071	5.055569
wardlink	1.213333	2.114286	2.830769
Freq	75.000000	70.000000	65.000000

cluster	0	1	2
spending	18.5	11.9	14.4
advance_payments	16.2	13.2	14.3
probability_of_full_payment	0.9	0.8	0.9
current_balance	6.2	5.2	5.5
credit_limit	3.7	2.8	3.3
min_payment_amt	3.6	4.7	2.7
max_spent_in_single_shopping	6.0	5.1	5.1
wardlink	1.0	2.1	2.9
link_average	1.0	2.1	2.7
Frequency	67.0	72.0	71.0

Table-10- cluster profiles for all 3 Clustering method

From the clustering analysis three groups namely-
Group 1: High Spending Group
Group 2: Low Spending Group and
Group 3: Medium Spending Group are formed

Group 1: High Spending Group -
Maximum max_spent_in_single_shopping is high for this group.
Perks like awarding reward points, offering discount or offer on next transactions upon full payment could increase their excitement to spend more.
Increasing their credit limit may also help to encourage push their spending nature.

Group 2: Low Spending Group –
They are the lowest spending group. They should be properly reminded for their payment with regular reminders . Bank could also provide some perks to them to make up their mind to spend more.

Group 3: Medium Spending Group –
Customers of this group should also be awarded with reward points as well discount to boost them up for spending more.
In addition some superlative offers could be provided on early payments to encourage them to spend more.

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).

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

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.

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

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

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).

	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 1-Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
Age             3000 non-null int64
Agency_Code     3000 non-null object
Type            3000 non-null object
Claimed         3000 non-null object
Commision       3000 non-null float64
Channel         3000 non-null object
Duration        3000 non-null int64
Sales           3000 non-null float64
Product Name    3000 non-null object
Destination     3000 non-null object
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

- 10 variables
- Age, Commision, Duration, Sales are numeric variable
- rest are categorial variables
- 3000 records, no missing one
- 9 independant variable and one target variable - Clami

Table 2-Data information

	count	mean	std	min	25%	50%	75%	90%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	53.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	48.300	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	224.200	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	172.025	539.00

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000	NaN	NaN	NaN	38.091	10.4635	8	32	36	42	84
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000	NaN	NaN	NaN	14.5292	25.4815	0	0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000	NaN	NaN	NaN	70.0013	134.053	-1	11	26.5	63	4580
Sales	3000	NaN	NaN	NaN	60.2499	70.734	0	20	33	69	539
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 3-Descriptive Summary of dataset

From the summary it could be observed that the dataset has 4 numeric values and 6 categorical values
Agency code EPX has a frequency of 1365.
The most preferred type seems to be travel agency Channel is online
Customized plan is the most sought plan by customers and ASIA seems to be most sought destination place by customers

- As the duplicate values could contain information for more customers I did not dropped of the duplicate values and the outliers were not treated.

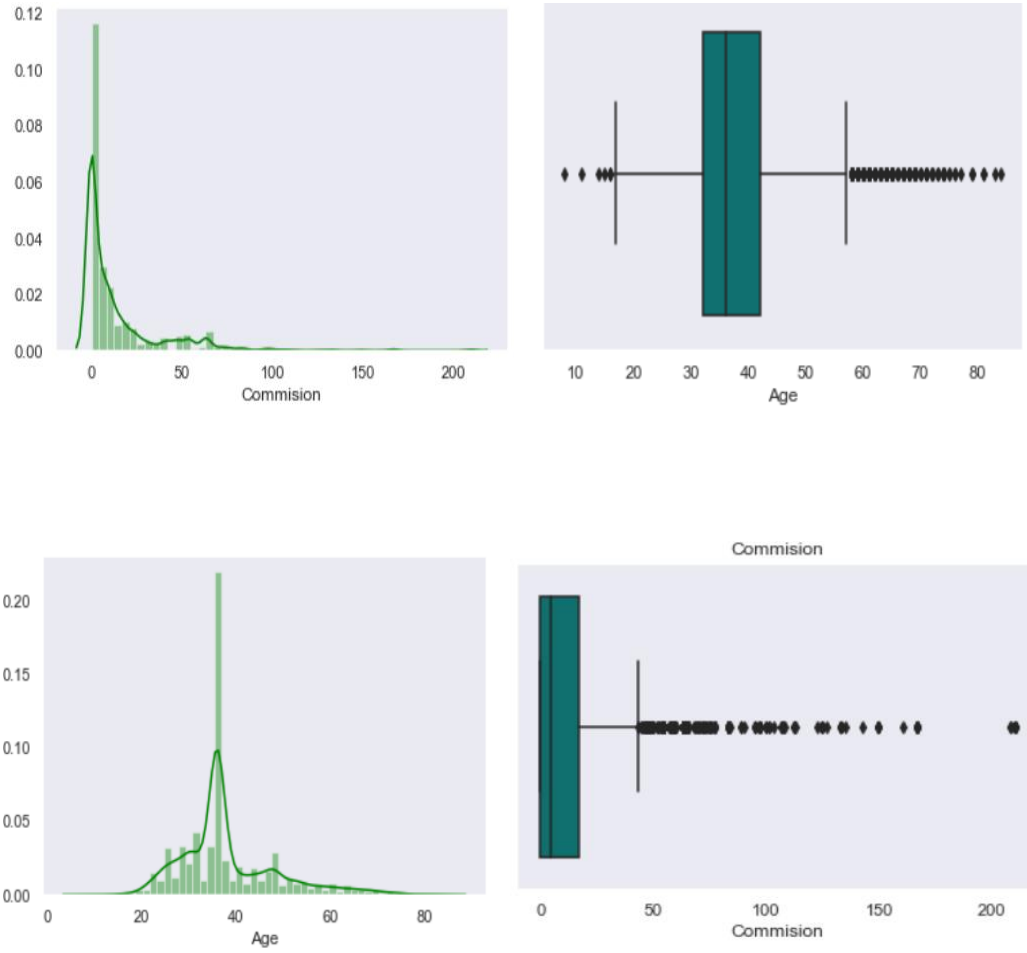
```
AGENCY_CODE :  4
JZI      239
CWT      472
C2B      924
EPX     1365
Name: Agency_Code, dtype: int64


TYPE :  2
Airlines         1163
Travel Agency    1837
Name: Type, dtype: int64


CLAIMED :  2
Yes      924
No      2076
Name: Claimed, dtype: int64


CHANNEL :  2
Offline       46
Online      2954
Name: Channel, dtype: int64
```

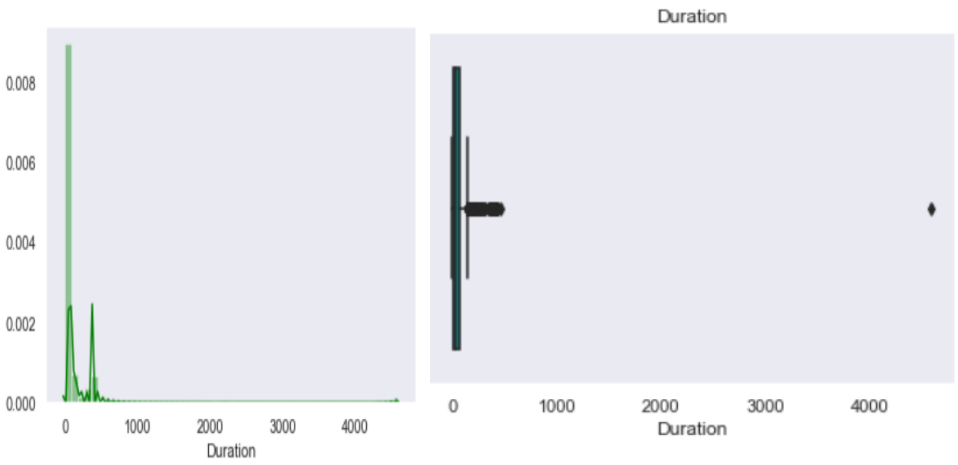
```
PRODUCT NAME :  5
Gold Plan               109
Silver Plan             427
Bronze Plan             650
Cancellation Plan       678
Customised Plan        1136
Name: Product Name, dtype: int64


DESTINATION :  3
EUROPE         215
Americas       320
ASIA          2465
Name: Destination, dtype: int64
```

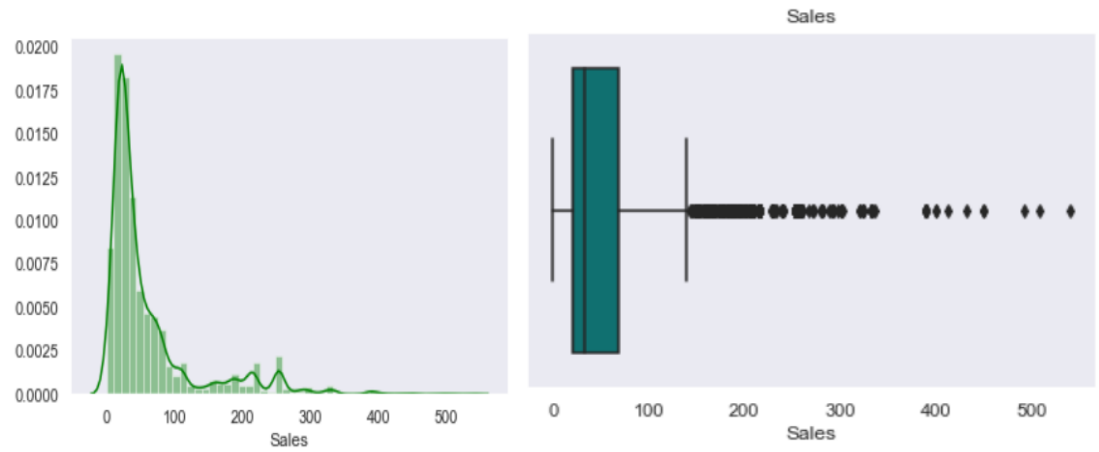
Table-4Unique counts of nominal values

Univariate /Multivariate Analysis-

Fig.1-Distribution of continuous data



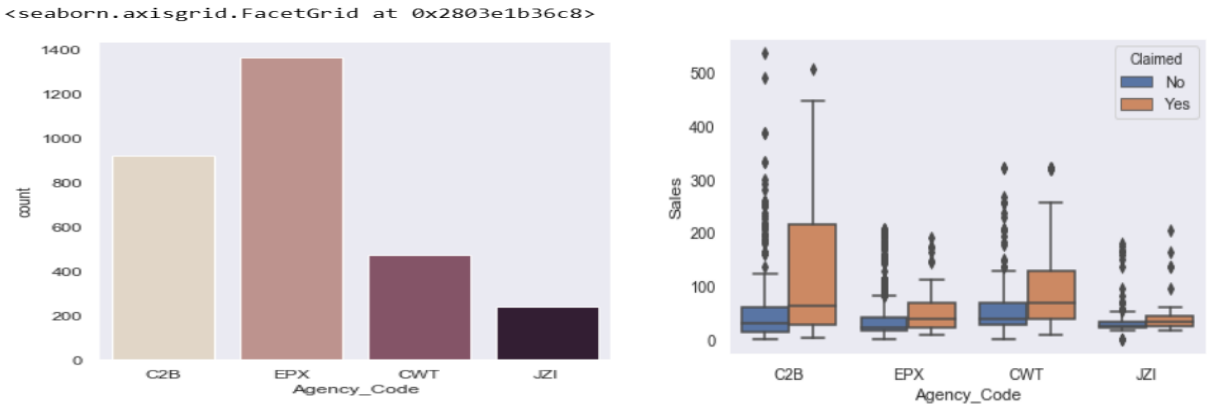




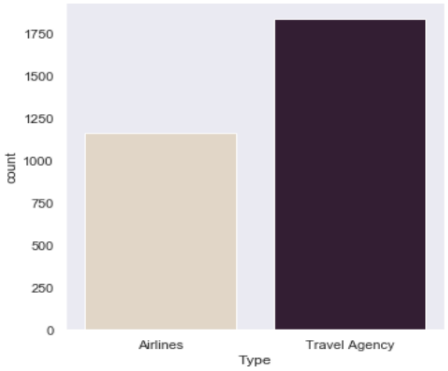
-All the continuous variables are right skewed and contains outliers.

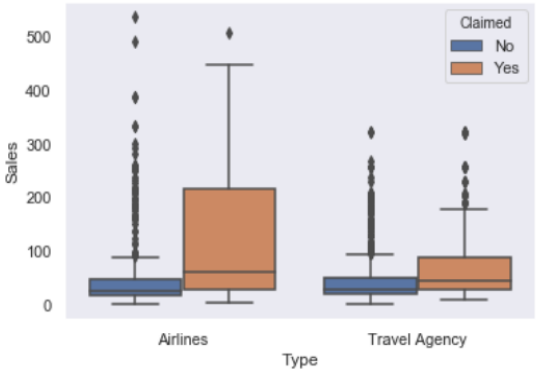
Distribution of categorical variable

Fig.2-Distribution of **categorical** data

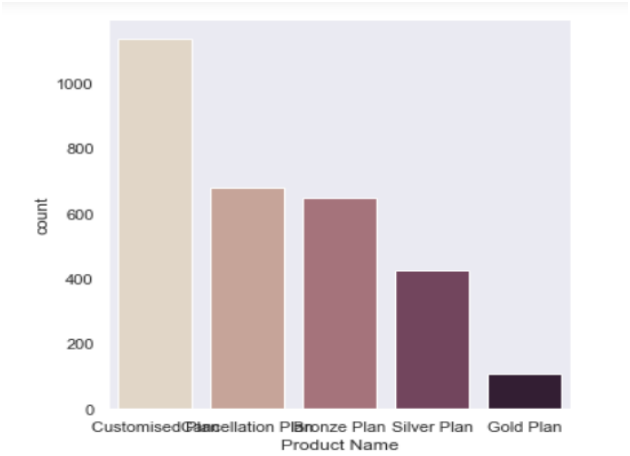


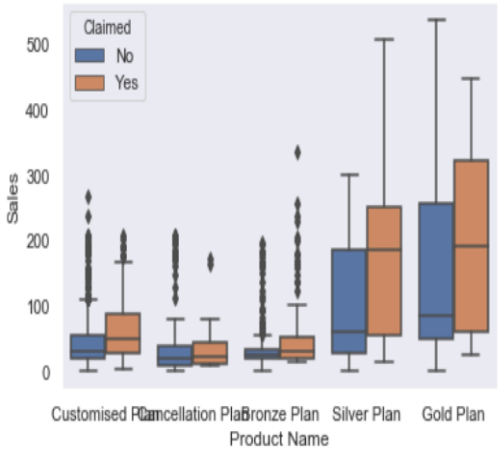
The distribution of the agency code shows EPX with maximum frequency and C2B have claimed more claims than other agency.



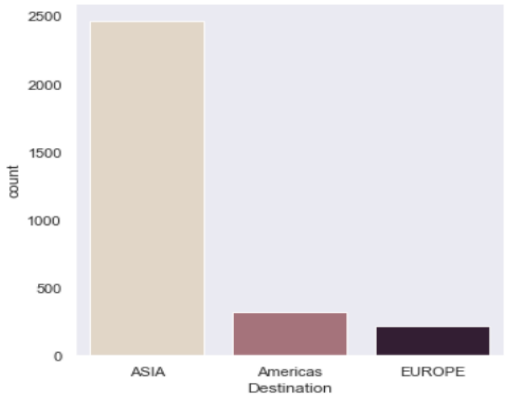


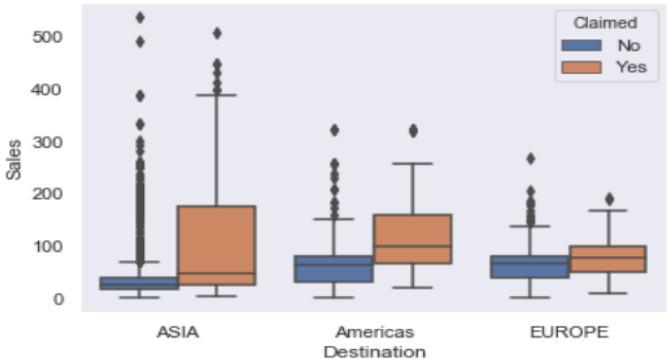
Airlines type has more claims.



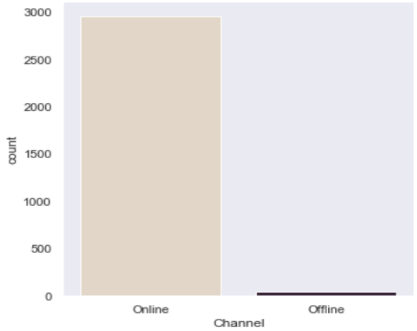


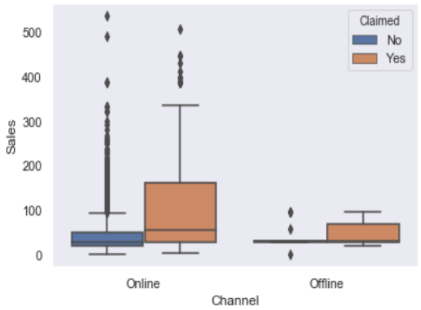
Customized plan seems to be most liked plan by customers when compared to all other plans.



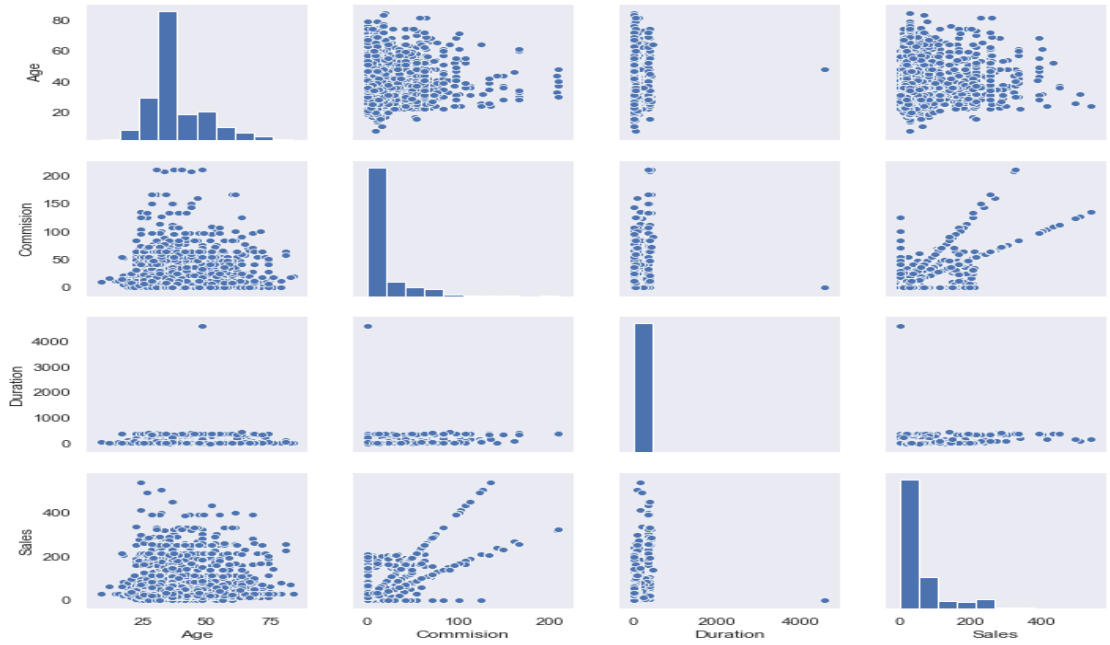


Asia is where customers choose when compared with other destination places





The majority of customers have used online medium, very less with offline medium



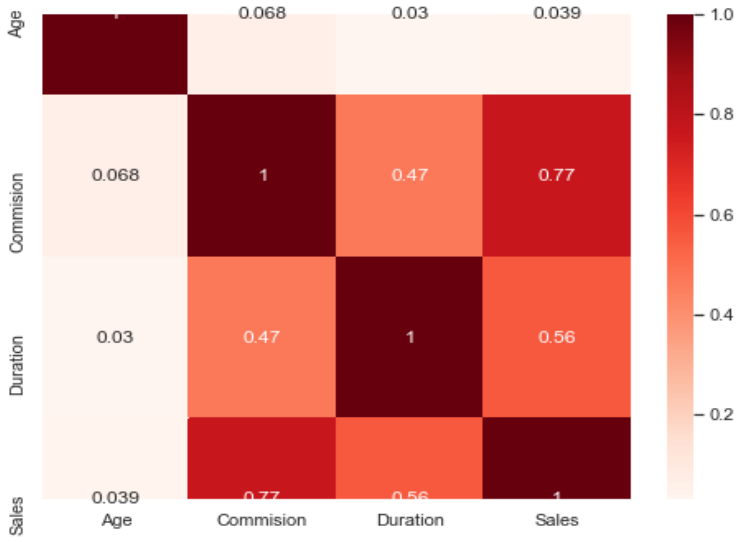


Fig.2-Distribution of pairplot and heatmap

Not much of multi collinearity is observed, no negative correlation is present the plot shows only positive Correlation.

```
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]
```

Table5- converting all object codes to categorical codes

Proportion of 1s and 0s-

```
0    0.692
1    0.308
```

2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural NetworkExtracting the target column into separate vectors for training set and test set¶.

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

Checking the dimensions of the training and test data

```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test labels (900,)
```

Link - http://webgraphviz.com/

Variable Importance-

	Imp
Agency_Code	0.599363
Sales	0.255785
Product Name	0.056555
Duration	0.037945
Age	0.030261
Commision	0.012676
Type	0.007416
Channel	0.000000
Destination	0.000000

FITTING THE OPTMAL VALUES TO THE TRAINING DATASET-

```
{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 50, 'min_samples_split': 150}

DecisionTreeClassifier(max_depth=10, min_samples_leaf=50, min_samples_split=150,
                       random_state=1)
```

Choosing best grid for Random Forest-

Grid Search for finding out the optimal values for the hyper parameters Due to large volume of data, trying for different parameter values in the gridsearch with higher cv value will have higher execution time, so the best values that came after the search are directly put in Param_grid.

```
GridSearchCV(cv=5, estimator=RandomForestClassifier(random_state=1),
             param_grid={'max_depth': [10, 12], 'max_features': [4, 8],
                         'min_samples_leaf': [35], 'min_samples_split': [100],
                         'n_estimators': [101]})

RandomForestClassifier(max_depth=10, max_features=8, min_samples_leaf=35,
                       min_samples_split=100, n_estimators=101, random_state=1)
```

Choosing best grid for ANN mode

```
{'hidden_layer_sizes': 8, 'max_iter': 2500, 'solver': 'sgd', 'tol': 0.001}


MLPClassifier(hidden_layer_sizes=8, max_iter=2500, random_state=1, solver='sgd',
              tol=0.001)
```

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.

Performance Metrics For CART Model-

Training data ROC CURVE

Testing data ROC CURVE

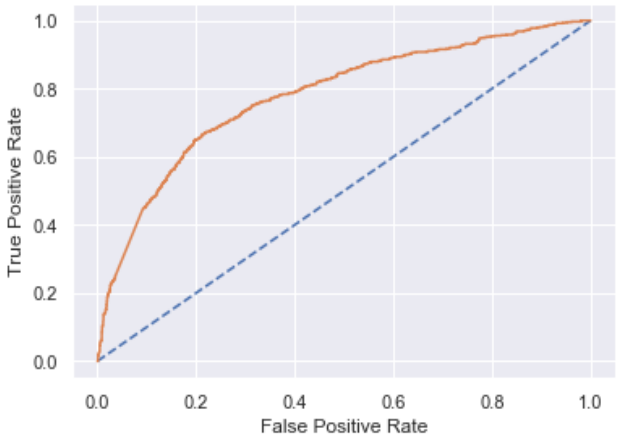


Fig3a- AUC: 0.779

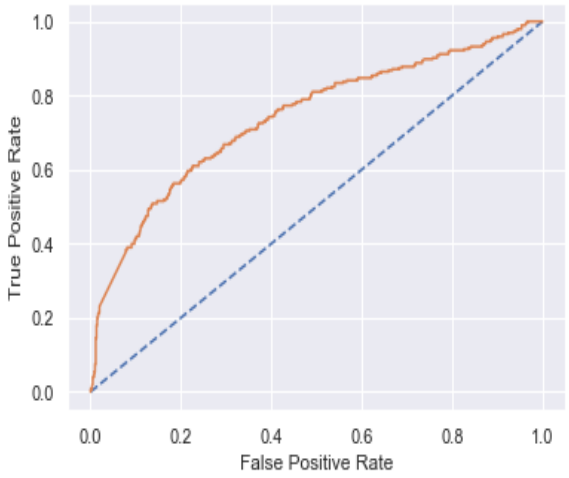


Fig3b- AUC: 0.740

Confusion Matrix for the training data -

```
array([[1418,   53],
       [ 472,  157]], dtype=int64)
```

Training Data Accuracy - 0.75

Training data matrix -

	precision	recall	f1-score	support
0	0.75	0.96	0.84	1471
1	0.75	0.25	0.37	629
accuracy			0.75	2100
macro avg	0.75	0.61	0.61	2100
weighted avg	0.75	0.75	0.70	2100

Confusion Matrix for the testing data -

```
array([[593,  12],
       [227,  68]], dtype=int64)
```

Testing Data Accuracy- 0.73

Testing Data matrix -

	precision	recall	f1-score	support
0	0.72	0.98	0.83	605
1	0.85	0.23	0.36	295
accuracy			0.73	900
macro avg	0.79	0.61	0.60	900
weighted avg	0.76	0.73	0.68	900

-

Random Forest Performance Matrix-

Training data ROC CURVE

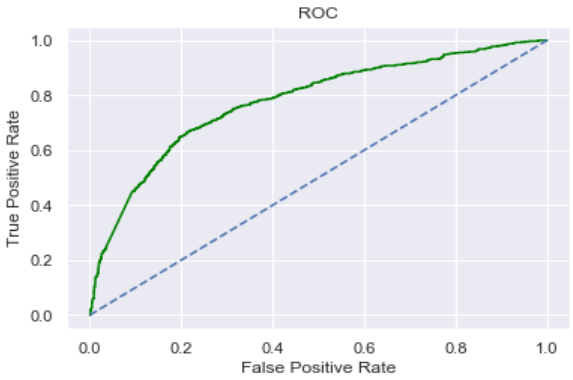


Fig4a- Area under Curve is 0.77

Testing data ROC CURVE

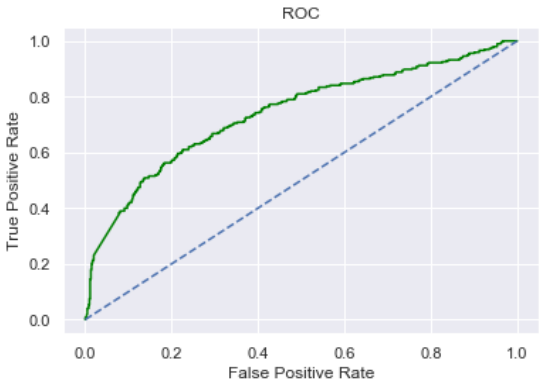


Fig- 4b- Area under Curve is 0.74

```
array([[1418,   53],
       [ 472,  157]], dtype=int64)
```

Confusion matrix for Training data -

Training Data Accuracy - 0.75

	precision	recall	f1-score	support
0	0.75	0.96	0.84	1471
1	0.75	0.25	0.37	629
accuracy			0.75	2100
macro avg	0.75	0.61	0.61	2100
weighted avg	0.75	0.75	0.70	2100

Training data matrix -

```
array([[593,  12],
       [227,  68]], dtype=int64)
```

Confusion Matrix for the testing data-

Testing Data Accuracy - 0.73

	precision	recall	f1-score	support
0	0.72	0.98	0.83	605
1	0.85	0.23	0.36	295
accuracy			0.73	900
macro avg	0.79	0.61	0.60	900
weighted avg	0.76	0.73	0.68	900

Testing data Matrix-

Neural Network Performance Matrix-

Training data ROC CURVE

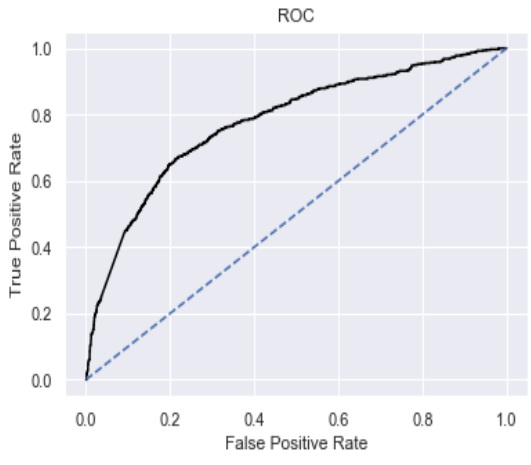


Fig-5a-AUC=0.77

Testing data ROC CURVE

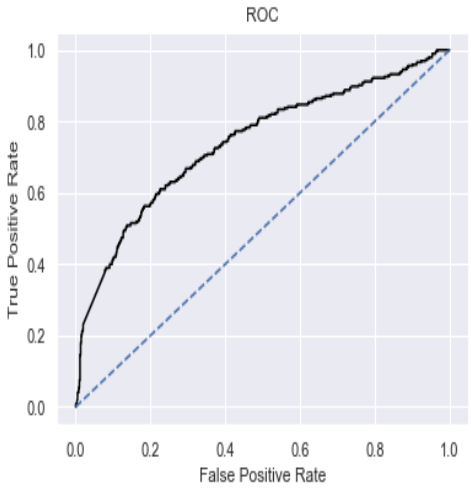


Fig-5b-AUC=0.74

```
array([[1418,   53],
       [ 472,  157]], dtype=int64)
```

Confusion Matrix Training Data-

Training Data Accuracy- 0.75

	precision	recall	f1-score	support
0	0.75	0.96	0.84	1471
1	0.75	0.25	0.37	629
accuracy			0.75	2100
macro avg	0.75	0.61	0.61	2100
weighted avg	0.75	0.75	0.70	2100

Training data matrix -

```
array([[593,  12],
       [227,  68]], dtype=int64)
```

Confusion Matrix for the testing data -

Testing Data Accuracy- 0.73

	precision	recall	f1-score	support
0	0.72	0.98	0.83	605
1	0.85	0.23	0.36	295
accuracy			0.73	900
macro avg	0.79	0.61	0.60	900
weighted avg	0.76	0.73	0.68	900

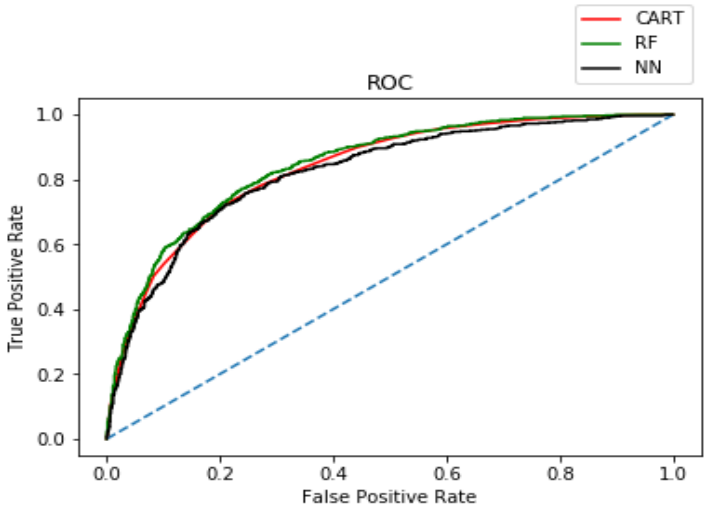
Testing Data Matrix -

I am selecting the RF model, as it has better accuracy, precision, recall, f1 score better than other two CART & NN

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

Table-1

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.75	0.73	0.75	0.73	0.75	0.73
AUC	0.78	0.74	0.78	0.74	0.78	0.74
Recall	0.25	0.23	0.25	0.23	0.25	0.23
Precision	0.75	0.85	0.75	0.85	0.75	0.85
F1 Score	0.37	0.36	0.37	0.36	0.37	0.36



2.5 Inference: Basis on these predictions, what are the business insights and recommendations

By manifesting the model, we could infer that streamlining online experiences has provided superficial welfare to customers, leading to an increase in conversions, which positively did raised profits.

Online channel contributes 90% of insurance claim and almost all the offline business has a claimed associated

JZI agency is entailed to have a little push up of resources to pick up sales as they are in bottom, need to run promotional marketing campaign or evaluate if we need to tie up with alternate agency

Also based on the model we are getting 80%accuracy, so we need customer books airline tickets or plans, cross sell the insurance based on the claim data pattern.