Business Report
Project – Data Mining
Created by Amit Jain

Table of Contents

List of Figure

1. CLUSTERING

Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

Problem Statement:

"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

To understand the problem, Bank has given sample of 210 bank customer data in the bank_marketing_part1_Data.csv randomly collected , which have pattern of the payments, dues and other payment related patterns.

Assumption:

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Step of understanding the data:

- Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:

	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

Data dictionary:

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).

We have loaded the data into **bank_df** data frame and checked for the shape of the data:

Data have 210 Rows and 7 Columns.

Check column values:

```
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)
```

We have data have in all fields. And there is no NULL value present in data set.

Data types for all data is Float , which is Numeric in nature

Nullability and Duplicity Check:

```
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
```

There are no NULL values in the data, all fields have some values in it.

We checked for the Duplicate data, and there are not Duplicate rows in the data set.

Data description:

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

Insights:

spending: Amount spent by the customer have average of 14.84, with Standard deviation of 2.9 and minimum value is 10.59 and maximum value is 21.1, Data values seems normal, Treatment of data not required with this analysis.

advance_payments: Amount paid by the customer in advance have average of 14.55, with Standard deviation of 1.3 and minimum value is 12.41 and maximum value is 17.25, Data values seems normal, Treatment of data not required with this analysis.

probability_of_full_payment: Probability of payment done in full by the customer have average of 87%, with Standard deviation of 2.3% and minimum value is 80.8% and maximum value is 91.8%, Data values seems normal, Treatment of data not required with this analysis.

current_balance: Balance amount left in the account to make purchases have average of 5.62, with Standard deviation of 0.44 and minimum value is 4.89 and maximum value is 6.67, Data values seems normal, Treatment of data not required with this analysis.

credit_limit: Limit of the amount in credit card have average of 5.62, with Standard deviation of 0.37 and minimum value is 2.63 and maximum value is 4.03, Data values seems normal, Treatment of data not required with this analysis.

min_payment_amt : minimum paid by the customer while making payments for purchases made monthly have average of 3.7, with Standard deviation of 1.50 and minimum value is 0.76 and maximum value is 8.45, Data values seems normal, Treatment of data not required with this analysis.

max_spent_in_single_shopping: Maximum amount spent in one purchase have average of 5.40, with Standard deviation of 0.49 and minimum value is 4.51 and maximum value is 6.55, Data values seems normal, Treatment of data not required with this analysis.

Univariate analysis:

Univariate Analysis should be done for each data columns, to understand the data pattern within each column. There are multiple graphs analysis available for different data types of columns.

We usually go for Histogram ,Boxplot and QQplot for Data , which are Numeric in nature.

And Count Plot, Pie Chart for data , which are Categorical in nature.

Graph Usage:

Histogram gives beautiful distribution chart for the data

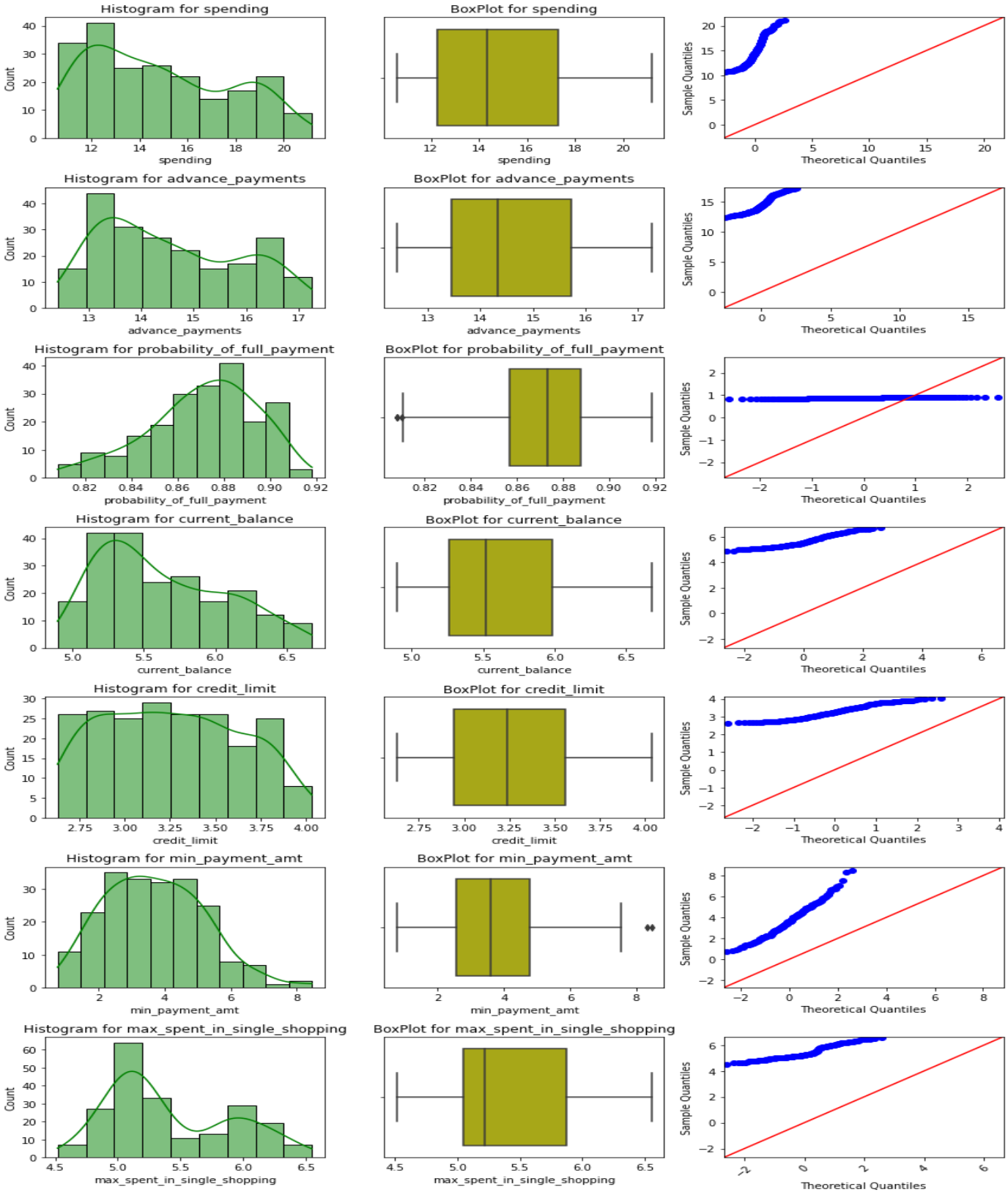
Boxplot is used for median and outliers of the data and where majority of the data present.

QQplot is used for checking the normality of the data.

Count Plot shows the Counts of each data segments for that column

Pie chart shows the proportion of the fields in complete data set.

Since we have all data available as of now are Continuous (Numeric), we will plot Histogram and Boxplot for this. Fig1: Combined HistPlot/BoxPlot/QQPlot



Observations:

- According to QQ Plot, we can see that all data points are distributed on full X axis, not on lying to each other on any one point on x axis, that's why all fields are normally distributed.
- We have outliers in data "**probability_of_full_payment**" and "**min_payment_amt**", which needs to be treated.
- Mean and Median is varying for "**Spending**" and "**Advance_payments**" fields, as compare to other fields of the data, so data scaling might be required , based on what method of analysis we are using.

Check for Skewness:

```
probability_of_full_payment    -0.522793
credit_limit                    0.134378
min_payment_amt                 0.360001
advance_payments                0.386573
spending                        0.399889
current_balance                 0.525482
max_spent_in_single_shopping    0.561897
```

Insights:

Probability of full payment is Left skewed

Max Spent in Single Shopping is Right Skewed .

Multivariate Analysis:

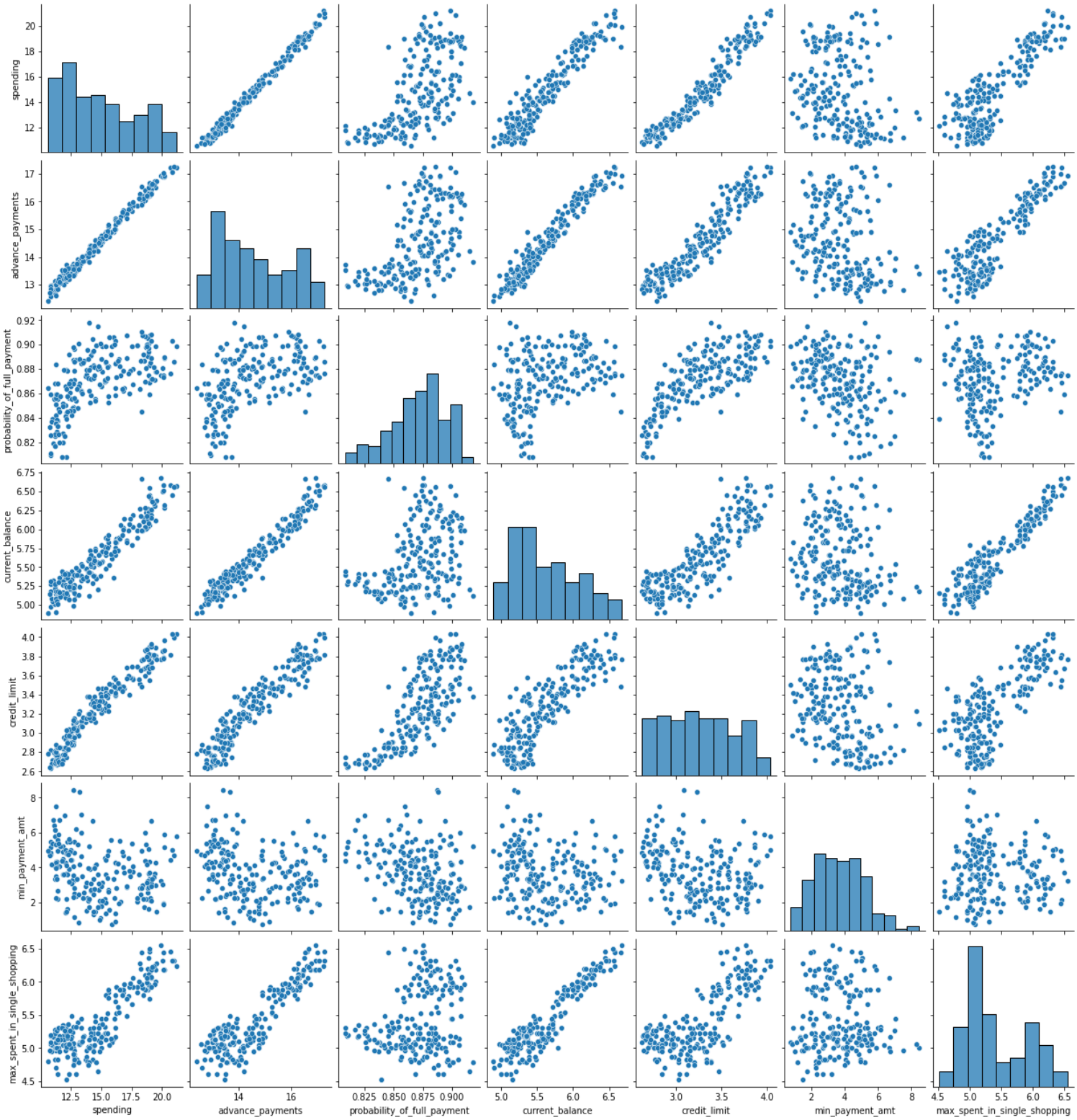
There are multiple ways for checking multivariate analysis .

Pair Plot: it gives comparison for each field of the data set

Heat Map: we can generate Heat map for the co-relations of the data elements. It gives a beautiful color-coded comparison of the data . and Strong to Light colors shows the Strong to low relationships between Fields.

Co-relation chart Metrics can also be build, which shows the co-relation numbers in a Tabular format.

Pair Plot: Fig 2: Pair Plot



Heat Map:

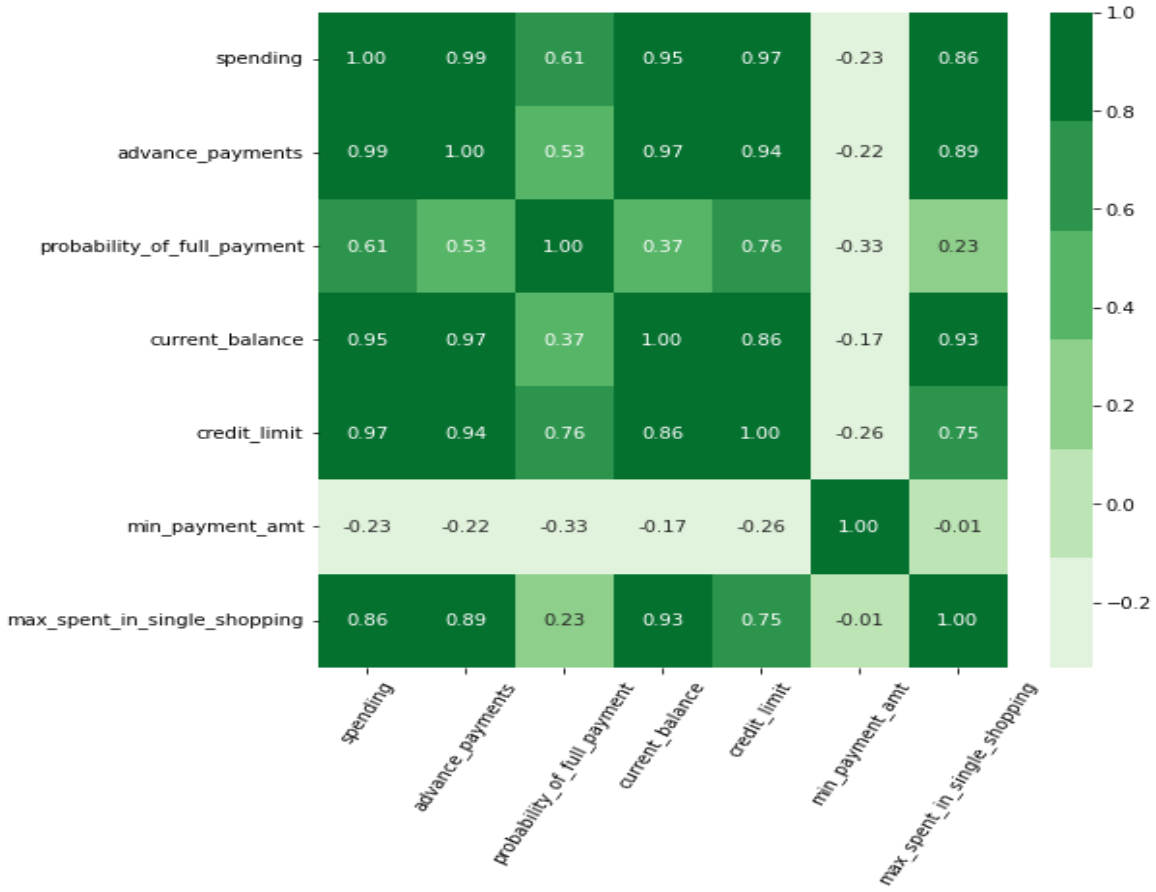


Fig 3: Heat Map

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
spending	1.000000	0.994341	0.608288	0.949985	0.970771	-0.229572	0.863693
advance_payments	0.994341	1.000000	0.529244	0.972422	0.944829	-0.217340	0.890784
ity_of_full_payment	0.608288	0.529244	1.000000	0.367915	0.761635	-0.331471	0.226825
current_balance	0.949985	0.972422	0.367915	1.000000	0.860415	-0.171562	0.932806
credit_limit	0.970771	0.944829	0.761635	0.860415	1.000000	-0.258037	0.749131
min_payment_amt	-0.229572	-0.217340	-0.331471	-0.171562	-0.258037	1.000000	-0.011079
in_single_shopping	0.863693	0.890784	0.226825	0.932806	0.749131	-0.011079	1.000000

Insights:

Positive Strong co-relations between:

- spending and current_balance
- Spending and advance_payments
- spending and credit_limit
- spending and max_spent_in_single_shopping

Which means as Current balance increases, Spending also increases.

Customer with good credit limits do more Spending and do maximum Advance_payments ,

Customers with Maximum current balance, they spent maximum in Single shopping as well.

Slight Negative co-relation between:

min_payment_amt have Negative co-relation with all other fields, though relationship is not too Strong

Box Plot for data comparisons:

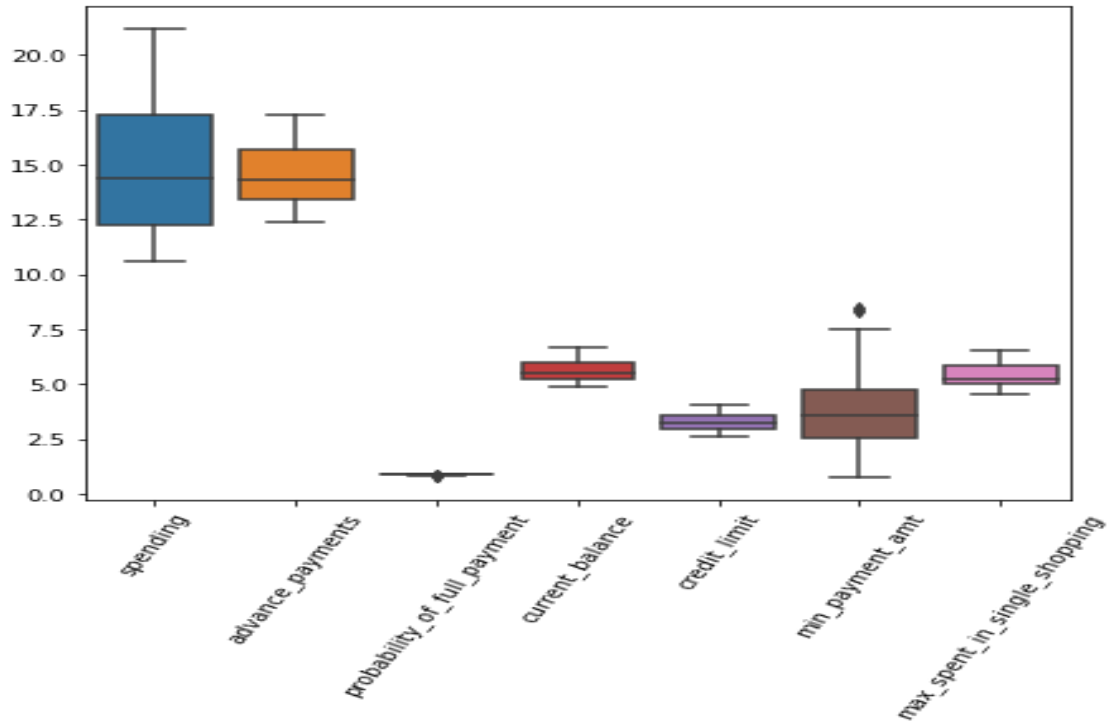


Fig 4: Compare Box Plot

Insights:

- Following Data have outliers: Probability_of_full_payment and min_payment_amount. Outliers has to be treated.
- Mean and Median is different for all the fields, so data is not in same scale.

Outlier Treatment:

Following Data have outliers: Probability_of_full_payment and min_payment_amount. Outliers has to be treated.

After treatment, we see following boxplot graph :

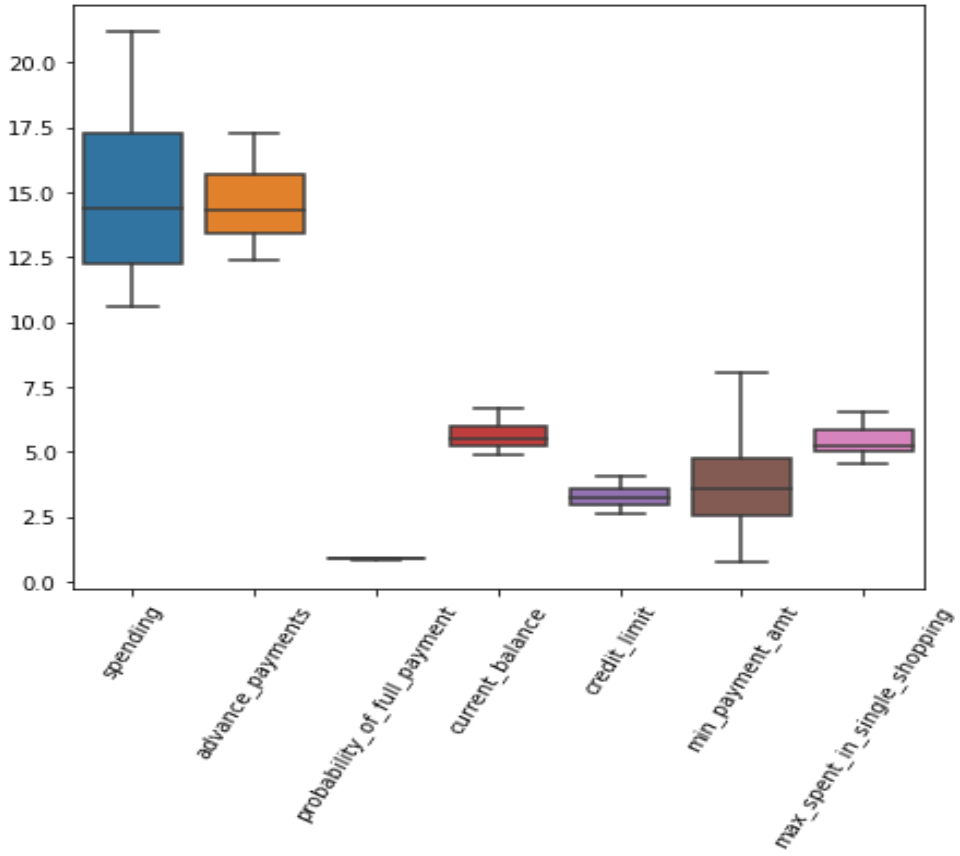


Fig 5: Compare Box Plot

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

For checking , if scaling is required or not, we need to check for the data description, it's mean, median, and its values ranging between minimum and maximum.

	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.590000	12.27000	14.35500	17.305000	21.180000
advance_payments	210.0	14.559286	1.305959	12.410000	13.45000	14.32000	15.715000	17.250000
probability_of_full_payment	210.0	0.871025	0.023560	0.810588	0.85690	0.87345	0.887775	0.918300
current_balance	210.0	5.628533	0.443063	4.899000	5.26225	5.52350	5.979750	6.675000
credit_limit	210.0	3.258605	0.377714	2.630000	2.94400	3.23700	3.561750	4.033000
min_payment_amt	210.0	3.697288	1.494689	0.765100	2.56150	3.59900	4.768750	8.079625
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.519000	5.04500	5.22300	5.877000	6.550000

Based on the following description of the data:

Mean and Median is varying for "Spending" and "Advance_payments" fields, as compare to other fields of the data, so data scaling is required to bring all fields on same range.

There are multiple ways and scaler available for data scaling. We used z-score scaling for the data scaling. And this is how data looks like 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	0.177628	2.367533	1.338579	-0.298625	2.328998
1	0.393582	0.253840	1.505071	-0.600744	0.858236	-0.242292	-0.538582
2	1.413300	1.428192	0.505234	1.401485	1.317348	-0.220832	1.509107
3	-1.384034	-1.227533	-2.571391	-0.793049	-1.639017	0.995699	-0.454961
4	1.082581	0.998364	1.198738	0.591544	1.155464	-1.092656	0.874813

Lets describe the data one more time:

	count	mean	std	min	25%	50%	75%	max
spending	210.0	9.148766e-16	1.002389	-1.466714	-0.887955	-0.169674	0.846599	2.181534
advance_payments	210.0	1.097006e-16	1.002389	-1.649686	-0.851433	-0.183664	0.887069	2.065260
probability_of_full_payment	210.0	1.642601e-15	1.002389	-2.571391	-0.600968	0.103172	0.712647	2.011371
current_balance	210.0	-1.089076e-16	1.002389	-1.650501	-0.828682	-0.237628	0.794595	2.367533
credit_limit	210.0	-2.994298e-16	1.002389	-1.668209	-0.834907	-0.057335	0.804496	2.055112
min_payment_amt	210.0	1.512018e-16	1.002389	-1.966425	-0.761698	-0.065915	0.718559	2.938945
max_spent_in_single_shopping	210.0	-1.935489e-15	1.002389	-1.813288	-0.740495	-0.377459	0.956394	2.328998

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

Clustering: Clustering is a mechanism, same as buckets, which gives an idea in which segment our data reside, We used clustering for creating segments on the data, so that we can apply different methods based on Clusters, we can apply targeted campaign, give different levels of advertisements, promotional offers or understand, the pattern of the data.

We have multiple Linkage method, like "ward" , "average", "single", "complete" etc. we have chosen ward linkage method.

In the "ward" Linkage method, it is similar to Average and Centroid Linkage method, In this method, we take Centroid of the data and Treat it as mean, take distance of each data point to the Centroid and the Summation of Sum(square of distance) / std.

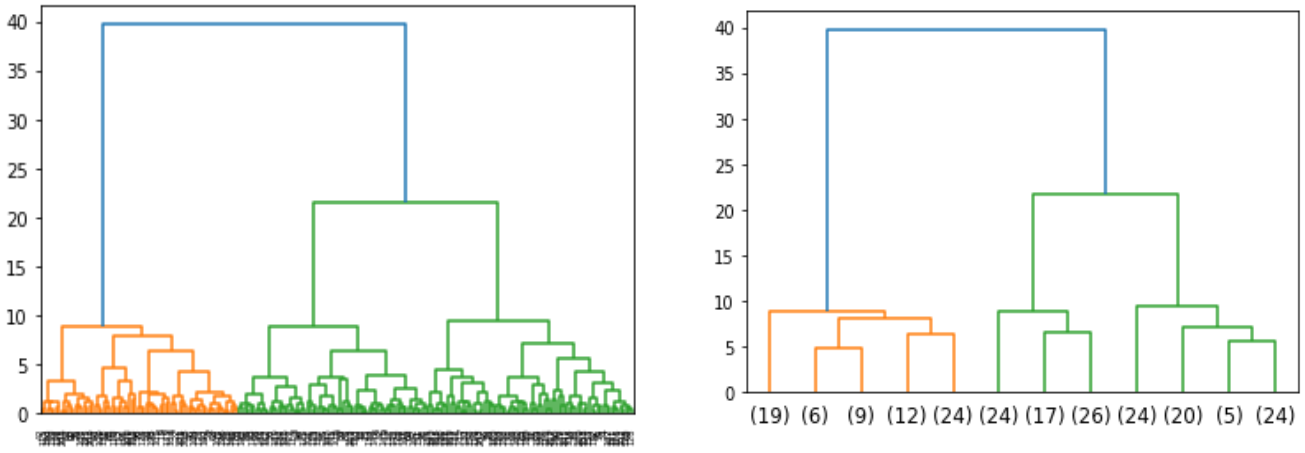


Fig 6: Dendrogram

We clearly see 3 color coding generated for the data, Blue, Orange and Green. And many other small buckets near to X-axis, which are visible properly.

So we choose **distance** mechanism to take a Last P clusters, and found this Dendrogram

And optimize the clusters. We found in the left hand side Dendrogram chart, that data is majorly classified into 3 clusters.

We append this cluster information back into main data set and found data like this:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	19.94	16.92	0.875200	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
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1

Lets try to understand the **Mean (average)** of each field of the data for each cluster:

	clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371
1	2	11.872388	13.257015	0.848155	5.238940	2.848537	4.940302	5.122209
2	3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178

Maximum values for each cluster:

	clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1	21.18	17.25	0.9108	6.675	4.033	6.682000	6.550
1	2	13.37	13.95	0.8883	5.541	3.232	8.079625	5.491
2	3	16.63	15.46	0.9183	6.053	3.582	6.685000	5.879

Minimum values for each cluster:

	clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1	15.38	14.86	0.845200	5.709	3.268	1.4720	5.443
1	2	10.59	12.41	0.810588	4.899	2.630	3.0820	4.794
2	3	11.23	12.63	0.833500	4.902	2.719	0.7651	4.519

Insights:

Cluster 1 has the highest Mean in fields like spending, “advance_payments”,” current_balance”. It also have the highest values in almost every field “spending , credit limit, advance_payment, probability_of_full_payment , current_balance, credit_limit”. So this can be consider as Richest group of people, as compared to other 2 groups.

cluster 2 has minimum average in “spending”, “advance_payments”, “current_balance.” It also has minimum actual values as well in almost every field, and Maximum in "min_payment_amt" which seems, they are not good in payments on time have pile of amount in their account.

cluster 3 has maximum mean values (after Cluster 1) in “spending”, “advance_payments”, “current_balance” . it also has middle Income category with 2nd highest values in almost all fields After Cluster 1.

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.

For applying Kmean Clustering , we need to begin with we import the libraries necessary for clustering and try to understand the WSS (within sum of squares) scores and plot an elbow curve.

This elbow curve will give us a rough estimate of the optimum number of cluster and we can further refine our inference by checking the silhouette score and silhouette visualizer and come to a conclusion that what should be our optimum cluster .

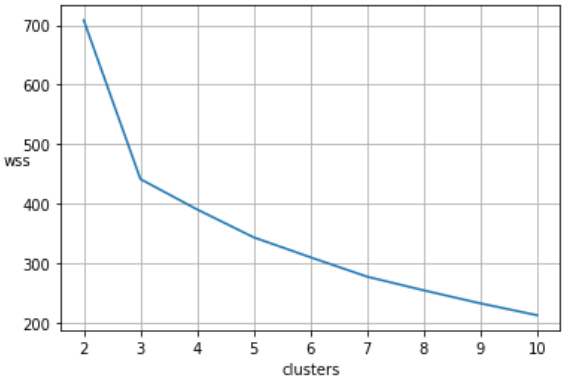


Fig 7 Elbow Chart

Actual WSS values in sorted order:

```
[708.4997372679238,
 441.1272533834272,
 390.09867501115224,
 343.085993111705,
 309.5848904184662,
 276.9504287893686,
 253.91486198354457,
 232.0123108155976,
 212.05173648777759]
```

As per the graph Slop of the WSS values from 2 to 3 is maximum steep , while from 3 to 4 and 4 to , Slop is not that Steep, so we will Stop at 3, and consider that number of optimum cluster will **be 3.**

Let's Fit the clusters into actual Data set.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Cluster_kmeans
)	19.94	16.92	0.875200	6.675	3.763	3.252	6.550	1
I	15.99	14.89	0.906400	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	0
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1

Also calculate the Silhouette score for each Row for their Clusters, add them as part of actual data set. and understand the data.

This is how the data look like:

	0	1	2	3	4
spending	19.940000	15.990000	18.950000	10.830000	17.990000
advance_payments	16.920000	14.890000	16.420000	12.960000	15.860000
probability_of_full_payment	0.875200	0.906400	0.882900	0.810588	0.899200
current_balance	6.675000	5.363000	6.248000	5.278000	5.890000
credit_limit	3.763000	3.582000	3.755000	2.641000	3.694000
min_payment_amt	3.252000	3.336000	3.368000	5.182000	2.068000
max_spent_in_single_shopping	6.550000	5.144000	6.148000	5.185000	5.837000
Cluster_kmeans	1.000000	2.000000	1.000000	0.000000	1.000000
sil_width	0.598554	0.421779	0.673243	0.536619	0.487637

Check for Minimum and maximum of Silhouette score:

Minimum value is : `-0.05981227785517841`

Maximum value is : `0.706373623438926`

Insights:

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Silhouette Score = (b-a)/max(a,b)

Where

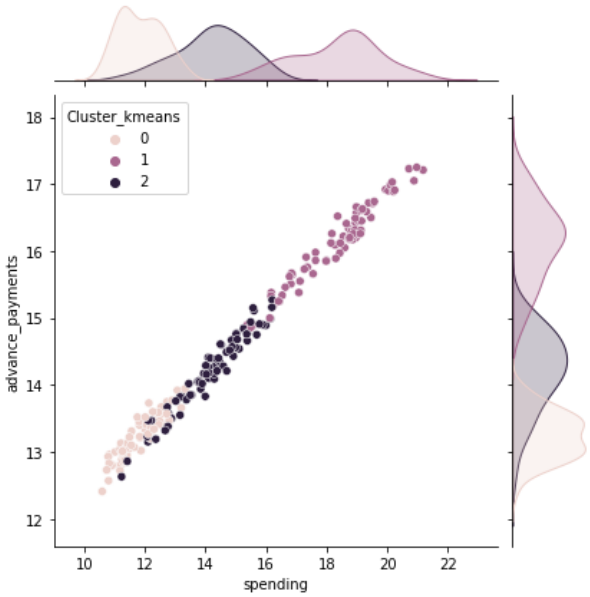
a= average intra-cluster distance i.e the average distance between each point within a cluster.

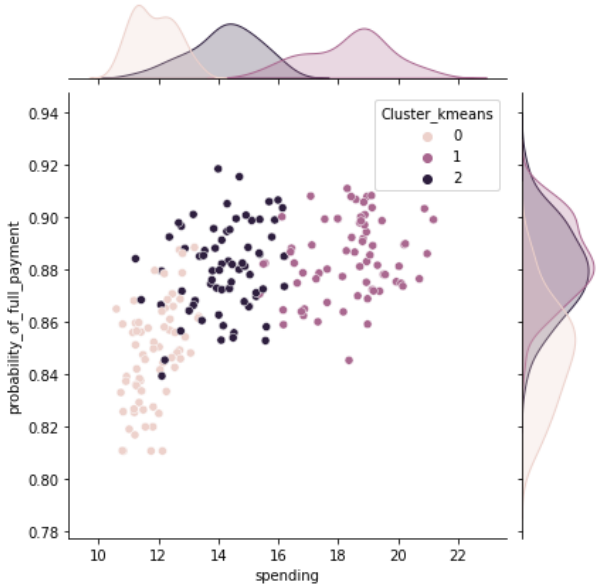
b= average inter-cluster distance i.e the average distance between all clusters.

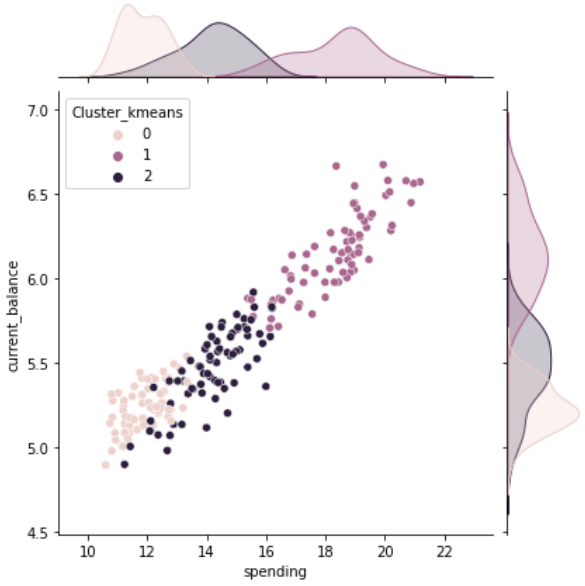
So in our case, we can see that Min and Max values are well ranging between -1 and +1

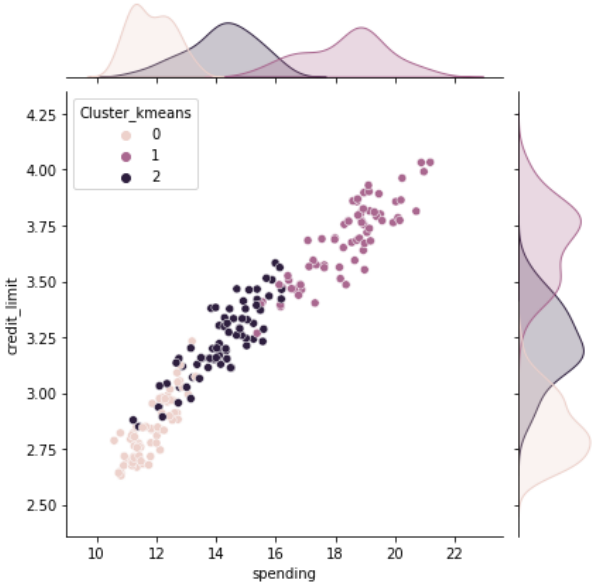
1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

For understanding the Segments of the data based on newly created Clusters, we will use Joint Plot :









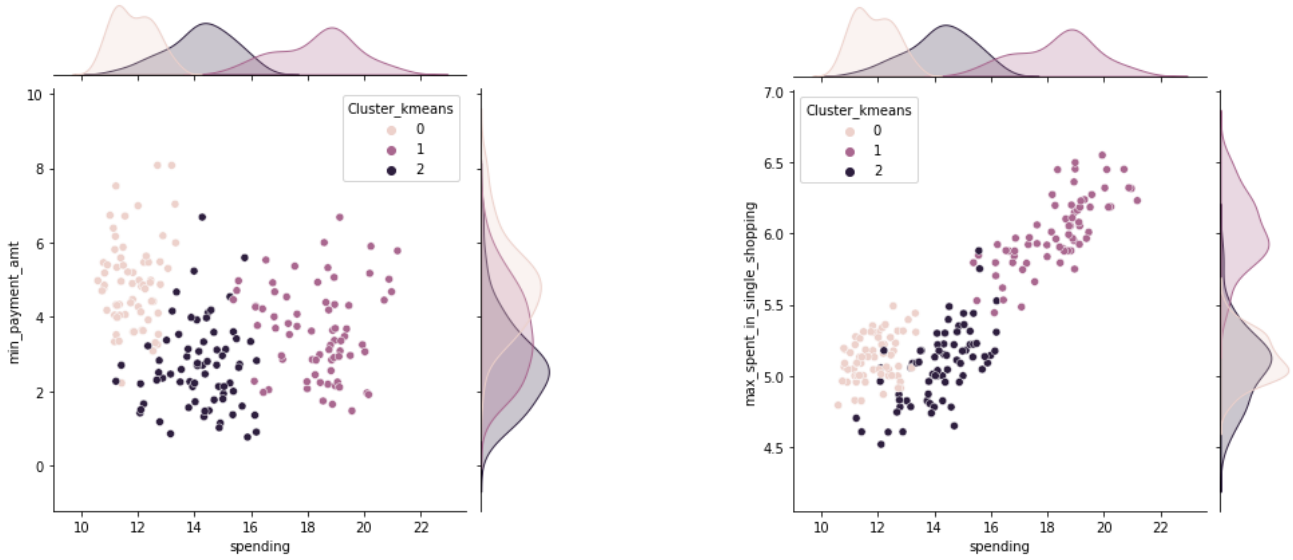


Fig 8: joint Plot

Check the Sum of the Spending for each Cluster and plot them :

	spending
0	793.51
1	1302.63
2	1021.84

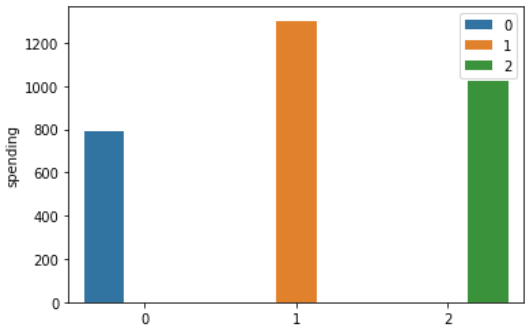


Fig 9: Bar Plot

Insights:

As we applied two different clustering algorithms to do the customer segmentation. Dendrogram and Kmeans algorithm. The clusters formed were nearly identical. We can classify the cluster into 3 different profiles, namely:

Cluster 0 - High risk customers with low advance payments and probability_of_full_payment. These customers have low current_balance,credit_limit ,max_spent_in_single_shopping,but high min_payment_amt.

Cluster 2 - Low risk customers with high advance_payments and probability_of_full_payment. These customers have high spending,current_balance,credit_limit, max_spent_in_single_shopping but less min_payment_amt.

Cluster 1 - Lowest risk customers among all other profiles with high advance_payments , we can say them Elite group of people and probability_of_full_payment as compared to the cluster 0 customers.

These customers have high spending, current_balance, credit_limit, max_spent_in_single_shopping than cluster 0 Customers.

2. CART-RF-ANN

Problem Statement:

"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.csv

To understand the problem, Insurance Company provided has given sample randomly collected data from multiple tour agencies, customer information, their pattern of Insurance claimed information, tour Location etc.

Assumption:

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Step of understanding the data:

- Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:

	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

Data dictionary:

1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency_Code)
3. Type of Tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)

6. Duration of the tour (Duration in days)
7. Destination of the tour (Destination)
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100's)
9. The commission received for tour insurance firm (Commission is in percentage of sales)
10. Age of insured (Age)

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

We have loaded the data into **insurance_df** data frame and checked for the shape of the data:

Data have 3000 Rows and 10 Columns.

Check column values:

```
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)
```

Data have all values in all fields. There are no NULL values in the data, all fields have some values in it.

We have mixed data type, Some are Integer, whole number , some are Float and some are Character in nature (Object). So we have both Continuous and Categorical data set.

Duplicity Check:

We checked for the Duplicate data, and there are not Duplicate rows in the data set.

Data description:

:

	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

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

Insights:

- when we check Continuous data set, there are difference in Mean, and Standard Deviation of the data, so Scaling might be required based on the requirement.
- There are bad values found in the dataset Duration and Commission , Example:

Duration minimum value is -1 which is not possible.

Commission is in the percentage form & we found here max value of commission is 210.21 and values more than 100.

Unique values for Categorical fields:

- unique values for Agency_Code are ['C2B' 'EPX' 'CWT' 'JZI']
- unique values for Type are ['Airlines' 'Travel Agency']
- unique values for Claimed are ['No' 'Yes']
- unique values for Channel are ['Online' 'Offline']
- unique values for Product Name are ['Customised Plan' 'Cancellation Plan' 'Bronze Plan' 'Silver Plan' 'Gold Plan']
- unique values for Destination are ['ASIA' 'Americas' 'EUROPE']

Values Counts and data proportion for all the categorical values:

- unique values for Agency_Code are
  EPX 1365
  C2B 924
  CWT 472
  JZI 239

- unique values for Type are
  Travel Agency 1837
  Airlines 1163

- unique values for Claimed are
  No 2076
  Yes 924

- unique values for Channel are
  Online 2954
  Offline 46

- unique values for Product Name are
  Customized Plan 1136
  Cancellation Plan 678
  Bronze Plan 650
  Silver Plan 427
  Gold Plan 109
- unique values for Destination are
  ASIA 2465
  Americas 320
  EUROPE 215

Data Insights:

- There are 4 Agency_Code present in the data set named as 'EPX' , 'C2B' , 'CWT' , 'JZI'
- 1365 customers have Agency_Code 'EPX' which is the max among all 4 Agency_Code present in the data.
- 239 customers have Agency_Code 'JZI' which is the min among all 4 Agency_Code present in the data.
- Most of the customers prefer Travel Agency as their tour insurance firm.
- 924 customers Claim their insurance.
- 2076 didn't Claim their insurance.
- 2954 customers choose online channel.
- Only 46 customers choose offline channel.
- Customized Plan is the most purchased insurance plan by the customers with a value count of 1136.
- Gold Plan is the least purchased insurance plan by the customers with a value count of 109.
- Asia is most preferred Destination of the tour.
- Europe is least preferred Destination of the tour.

Treatment of Bad Values

We saw above that Duration min value is -1, which is not possible. This has to be cleaned.

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
1508	25	JZI	Airlines	No	6.3	Online	-1	18.0	Bronze Plan	ASIA

We will convert this Duration with the Median of entire Duration field.

Also, we saw above that Commission is Percentage value and can never be more than 100%, so we need to treat them as well. we have 42 rows with Commission field as more than 100% and we can impute them with median values as well.

Check for Duplicate Records

We checked and found that there are 139 Duplicate records. But since it is a Tour Package data , and there are no Customer ID, or Employee IDs associated with any rows, we cannot say, that these records belong to same customer or firm, so we will not consider them as Duplicate, because same kind of Tour package can be offered to multiple customers as well So we will not drop the Duplicate records

Univariate analysis for continuous data:

First of all, we will divide data into Continuous and Categorical fields separately and then we would analyze them based on Graphs suitable for that data types.

Univariate Analysis should be done for each data columns, to understand the data pattern within each column. There are multiple graphs analysis available for different data types of columns.

We usually go for Histogram ,Boxplot and QQplot for Data , which are Numeric in nature.

And Count Plot, Pie Chart for data , which are Categorical in nature.

Graph Usage:

Histogram *gives beautiful distribution chart for the data*

Boxplot *is used for median and outliers of the data and where majority of the data present.*

QQplot *is used for checking the normality of the data.*

Count Plot *shows the Counts of each data segments for that column*

Pie chart *shows the proportion of the fields in complete data set.*

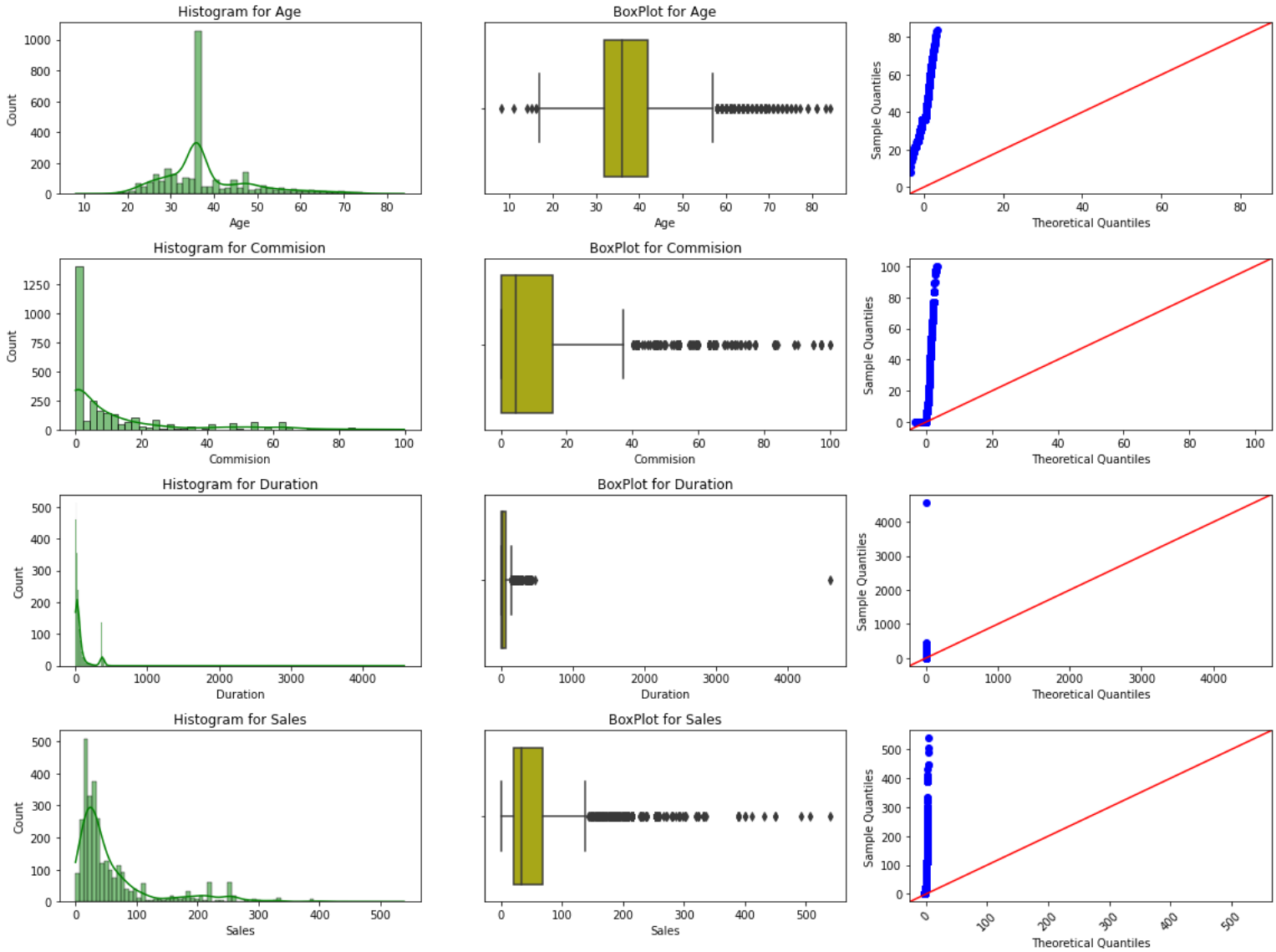


Fig 10: Combined HistPlot/BoxPlot/QQPlot

Observations:

- there are 4 Numeric fields in the data set Age, Commission, Duration and Sales
- Except Age, remaining other data fields are **not** 100% normally distributed,
- we have Commission, Duration and Sales , Right skewed
- all the data have outliers

Check for Skewness:

```
Age           1.149713
Commision     1.952623
Sales         2.381148
Duration     13.785722
```

Insights: We have Commission, Duration and Sales , Right skewed.

Description of the data:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000.0	NaN	NaN	NaN	38.091	10.463518	8.0	32.0	36.0	42.0	84.0
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000.0	NaN	NaN	NaN	12.537687	19.598981	0.0	0.0	4.63	15.6	99.9
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000.0	NaN	NaN	NaN	70.0105	134.049397	0.0	11.0	26.75	63.0	4580.0
Sales	3000.0	NaN	NaN	NaN	60.249913	70.733954	0.0	20.0	33.0	69.0	539.0
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Insights:

- Age:
  Age of insured ranges from a minimum of 8 to maximum of 84.
  The average Age of insured is around 38.1.
  The standard deviation of the Age of insured is 10.46.
  25% , 50% (median) and 75 % of the Age of insured are 32 , 36 and 42.

- Commision:
  The commission received for tour insurance firm ranges from a minimum of 0 to maximum of 99.9.
  The average Commision received for tour insurance firm is around 12.53.
  The standard deviation of the Commision received for tour insurance firm is 19.6.
  25% , 50% (median) and 75 % of the Commision received for tour insurance firm are 0 , 4.63 and 15.6.

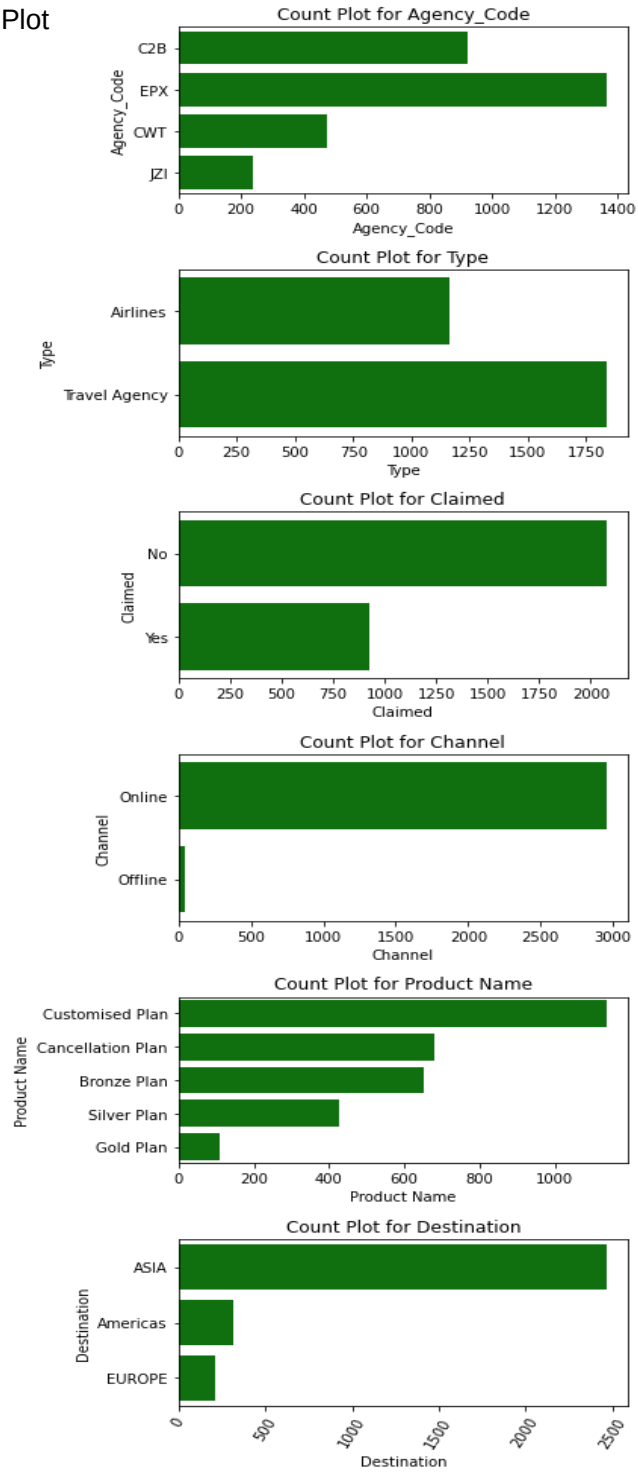
- Duration:
  Duration of the tour ranges from a minimum of 0 to maximum of 4580.
  The average Duration of the tour is around 70.
  The standard deviation of the Duration of the tour is 134.
  25% , 50% (median) and 75 % of the Duration of the tour are 11, 26.75 and 63.

- Sales:
  Amount of sales of tour insurance policies ranges from a minimum of 0 to maximum of 539.
  The average Sales of tour insurance policies is around 60.24.
  The standard deviation of the Sales of tour insurance policies is 70.73.
  25% , 50% (median) and 75 % of the Sales of tour insurance policies are 20 , 33 and 69.

Univariate analysis for Categorical data:

We use count plot for categorical variables, and it gives a bar graph like chart.

Fig 11: Count Plot



Proportion of the categorical values:

Name: Agency_Code,
EPX 45.5%
C2B 30.8%
CWT 15.7%
JZI 07.9%

Name: Type,
Travel Agency 61.2%
Airlines 38.7%

Name: Claimed,
No 69.2%
Yes 30.8%

Name: Channel,
Online 98.4%
Offline 01.5%

Name: Product Name,
Customised Plan 37.8%
Cancellation Plan 22.6%
Bronze Plan 21.6%
Silver Plan 14.2%
Gold Plan 03.6%

Name: Destination,
ASIA 82.1%
Americas 10.6%
EUROPE 07.1%

Insights:

Agency_Code
- There are 4 Agency_Code present in the data set named as 'EPX' , 'C2B' , 'CWT' , 'JZI'.
- Maximum no of 45% customers have Agency_Code 'EPX'
- Minimum 7% customers have Agency_Code 'JZI'

Type
- 61.2% customers prefer Travel Agency as their tour insurance firm.
- 38.7% customers prefer Airlines as their tour insurance firm.

Claimed

69.2 % didn't Claim their insurance.

30.8 % Claim their insurance.

Channel

98.4% customers choose online channel.

1.53% customers choose offline channel.

Product Name

maximum 37.86% customers purchased Customized Plan.

Minimum Only 3.6% customers purchased Gold Plan.

Destination

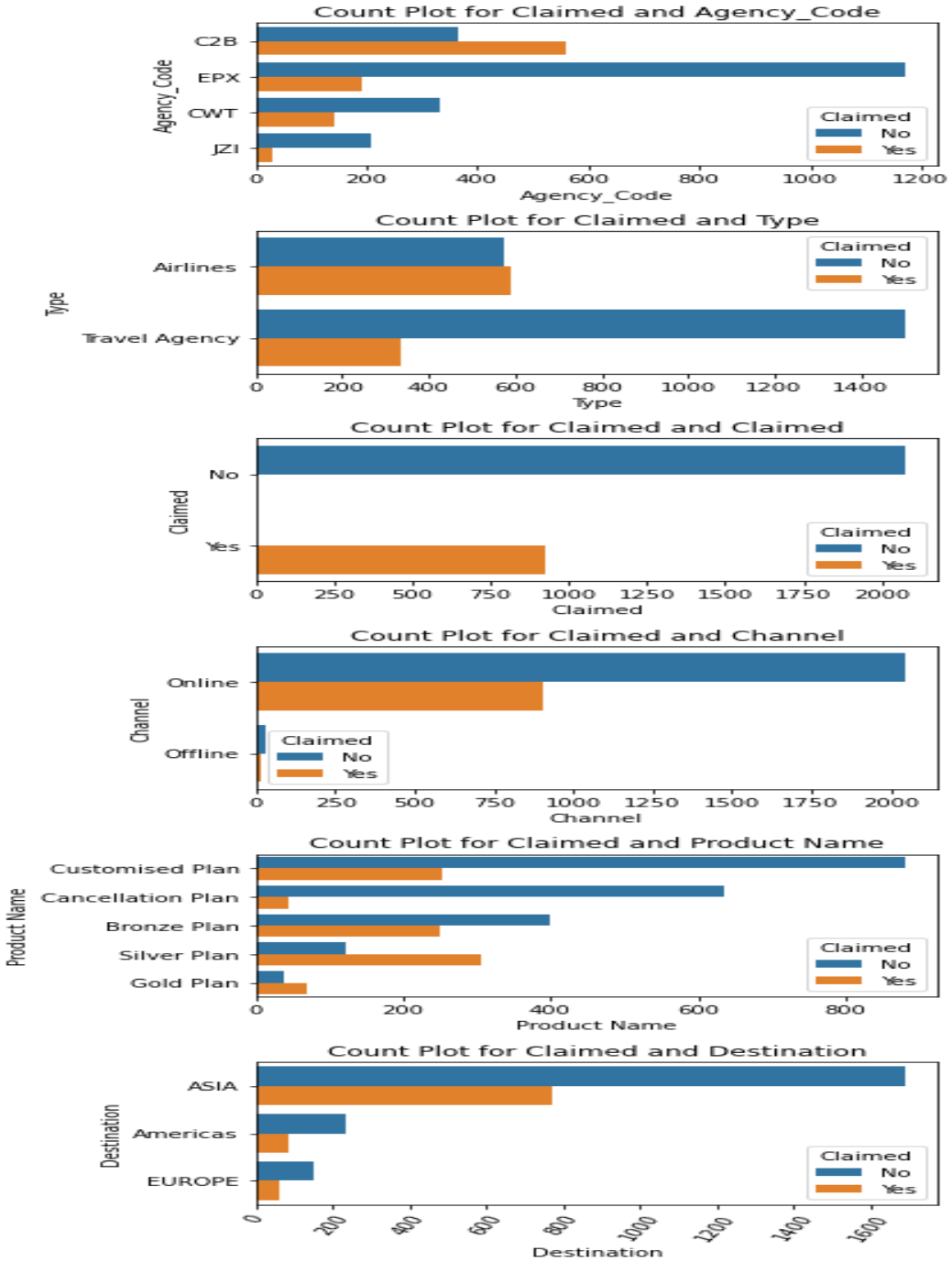
About 82 % customers, maximum choose Asia as Destination of the tour.

Only 7.1% customers choose Europe as Destination of the tour.

Bivariate Analysis:

Bivariate analysis is used for comparing 2 fields of the data set and it's impact on each other. There are multiple ways for doing the bivariate analysis. We will use Countplot with Hue for comparing 2 different variables of both Categorical types.

Fig 12 Count Plot



We also took Counts for each 2 Categorical fields , so that we can understand the data pattern. Based on counts and comparing the data, following Some of important Insights can be understand:

- Customers with Agency Code C2B claimed more insurance. 560 Claimed
- Most of Customers with Agency Code EPX didn't claimed insurance. 1172 no claimed vs 193 Claimed
- In online channel no claimed status is more than claimed Status. 2047 No Claimed vs 907 Claimed
- Silver Plan has maximum number of Claimed as 306
- Agency EPX sold maximum Products as Customized Plan (687) and Cancellation Plan (678)
- Travel Agency sold maximum Customized Plan as 1076
- Offline Gold Plan were sold least times Count is 2
- Online Customized Plan were sold maximum as 1092
- Travel Agency name EPX has done maximum number of bookings Count is 1365 and they have done maximum booking For Asia Destination Count is 1128
- There is no offline Booking for Airlines Types
- no one has done offline booking for Europe Destination

Multivariate Analysis

There are multiple ways for checking multivariate analysis .

Pair Plot: it gives comparison for each field of the data set

Heat Map: we can generate Heat map for the co-relations of the data elements. It gives a beautiful color-coded comparison of the data . and Strong to Light colors shows the Strong to low relationships between Fields.

Co-relation chart Metrics can also be built, which shows the co-relation numbers in a Tabular format.

Pair Plot for Numeric (Continuous) fields:

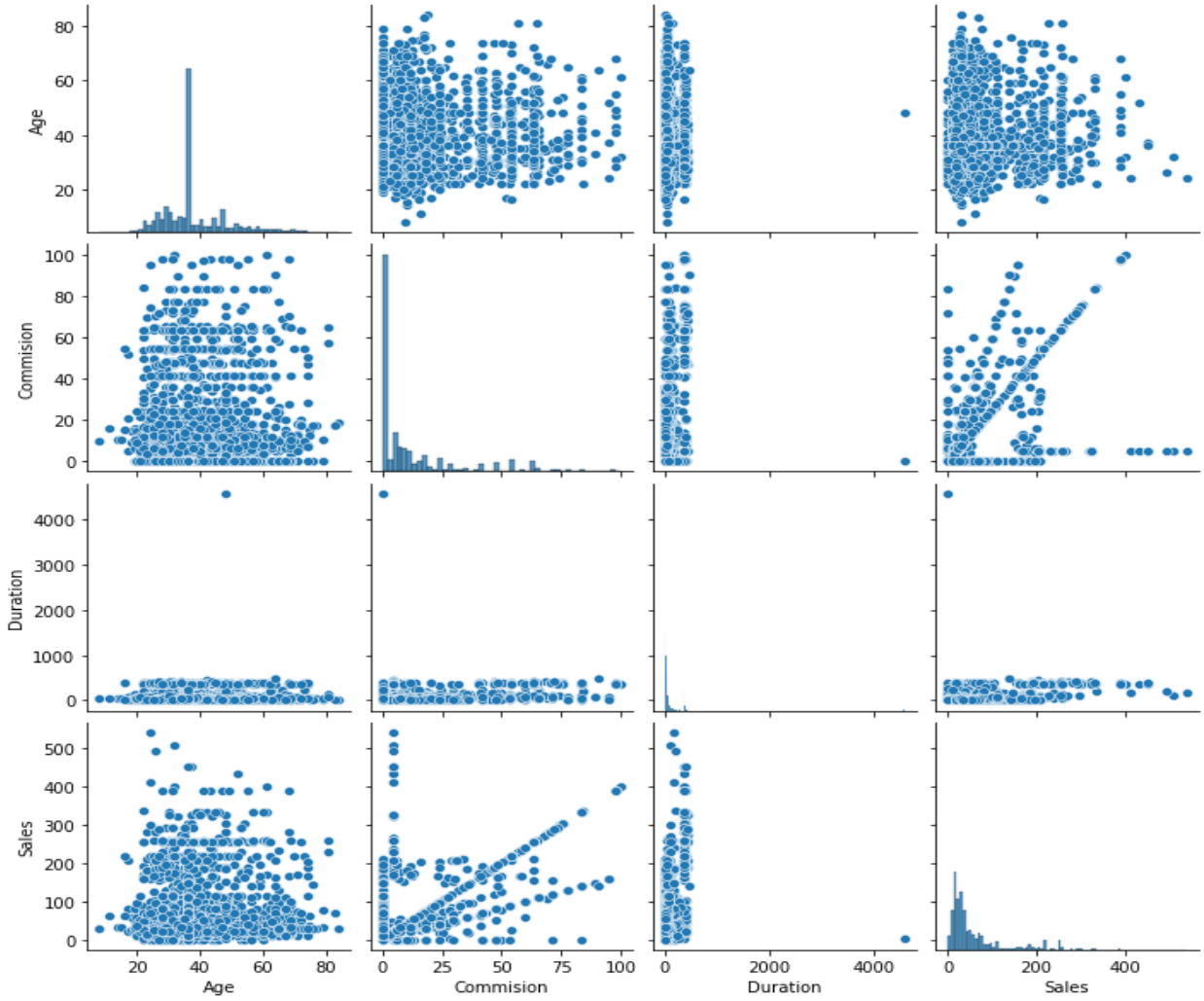


Fig 13: Pair Plot for Insurance data set

Heat Map:

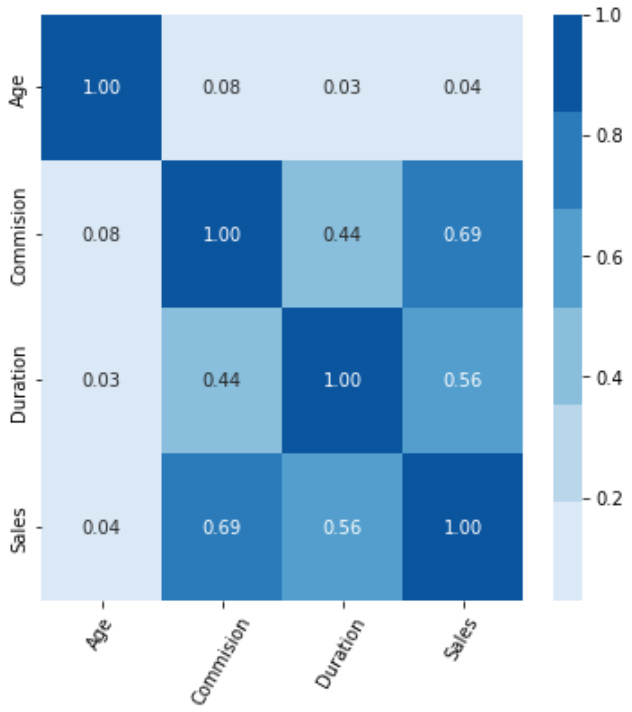


Fig 14: Heat Map

Insights:

Commission with Sales show strong correlation with 0.69.

Duration with Sales also show strong correlation i.e. 0.56.

Duration with Age and Age with Sales shows less positive correlation among all.

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

Before we Split the data and for any predictive analysis, we need to Convert all Categorical fields into Continuous/Discrete fields (Numeric).

Encoding:

We have total 6 Categorical fields in the data set. We will use general Codes mechanism to convert their values into appropriate Codes. Following are their unique values and Codes given for those values are as follows:

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

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

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

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

Columns is : 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]

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

Now every field looks Like Numeric in nature. Let's see , how the data looks like, after Encoding conversion of the data

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0	0.70	1	7.0	2.51	2	0
1	36	2	1	0	0.00	1	34.0	20.00	2	0
2	39	1	1	0	5.94	1	3.0	9.90	2	1
3	36	2	1	0	0.00	1	4.0	26.00	1	0
4	33	3	0	0	6.30	1	53.0	18.00	0	0

Description of the new data set:

	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.0	84.0
Agency_Code	3000.0	1.306333	0.994060	0.0	0.0	2.00	2.0	3.0
Type	3000.0	0.612333	0.487299	0.0	0.0	1.00	1.0	1.0
Claimed	3000.0	0.308000	0.461744	0.0	0.0	0.00	1.0	1.0
Commision	3000.0	12.537687	19.598981	0.0	0.0	4.63	15.6	99.9
Channel	3000.0	0.984667	0.122895	0.0	1.0	1.00	1.0	1.0
Duration	3000.0	70.010500	134.049397	0.0	11.0	26.75	63.0	4580.0
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.0	539.0
Product Name	3000.0	1.661667	1.258726	0.0	1.0	2.00	2.0	4.0
Destination	3000.0	0.250000	0.575277	0.0	0.0	0.00	0.0	2.0

After converting the data, we will drop the target variable which is "Claimed" in our data set. We split the data into training and test set labelled as "x" and "y" where "x" is data without target variable i.e., claimed and "y" is the target variable.

Check for shape of the train and test data set:

```
x_train (2100, 9)
x_test (900, 9)
y_train (2100,)
y_test (900,)
```

Once the split is done, we will build Decision tree (CART) model, Random Forest classifier (RFCL) model and Artificial neural network (ANN) model.

Decision Tree Classifier:

In CART model we use Decision tree classifier and fit our train and test data into the dt_model (Decision Tree model). Generate the decision tree and see , how it looks like, what is the Depth of the tree.

Without performing pruning the decision tree, this is how the Decision Tree would look like:

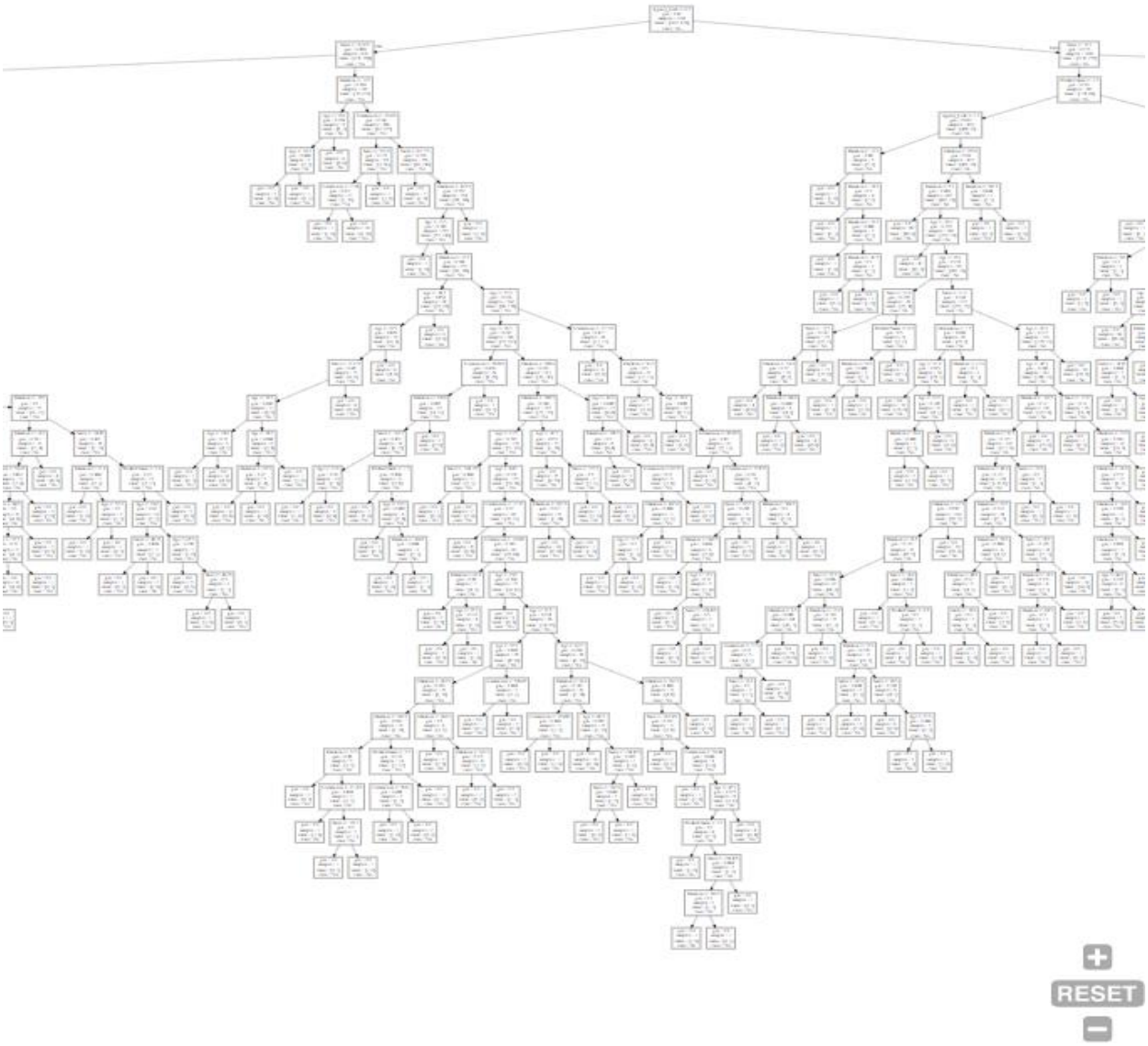


Fig 15 : Decision Tree, non-normalized

We can see that data is over fitted and there are more than 25 levels of the decision tree. This is not suitable for the modeling. So we will use grid search mechanism for best Possible search criteria for DTree, and this will also avoid DTree to over grown.

By trying various combinations of the grid parameters, we conclude the following best parameters for our decision tree model

```
{'criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 100, 'min_sample
s_split': 300}
```

After applying above listed Best parameters, this is how the tree would look like :

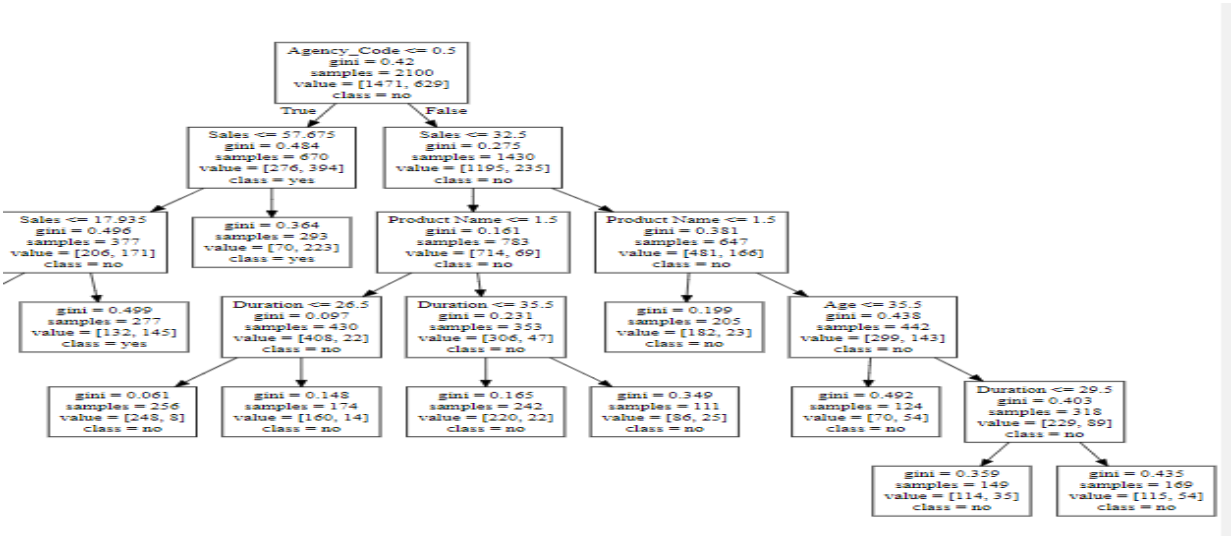


Fig 16 : Decision Tree, normalized

This is more suitable for data analysis , as it is not over grown and our model will not be overfitted. Following are the weighted chart for each fields of the data set:

```
                 Imp
Agency_Code   0.657599
Sales         0.246799
Product Name  0.060673
Duration      0.017588
Age           0.017342
Type          0.000000
Commision     0.000000
Channel       0.000000
Destination   0.000000
```

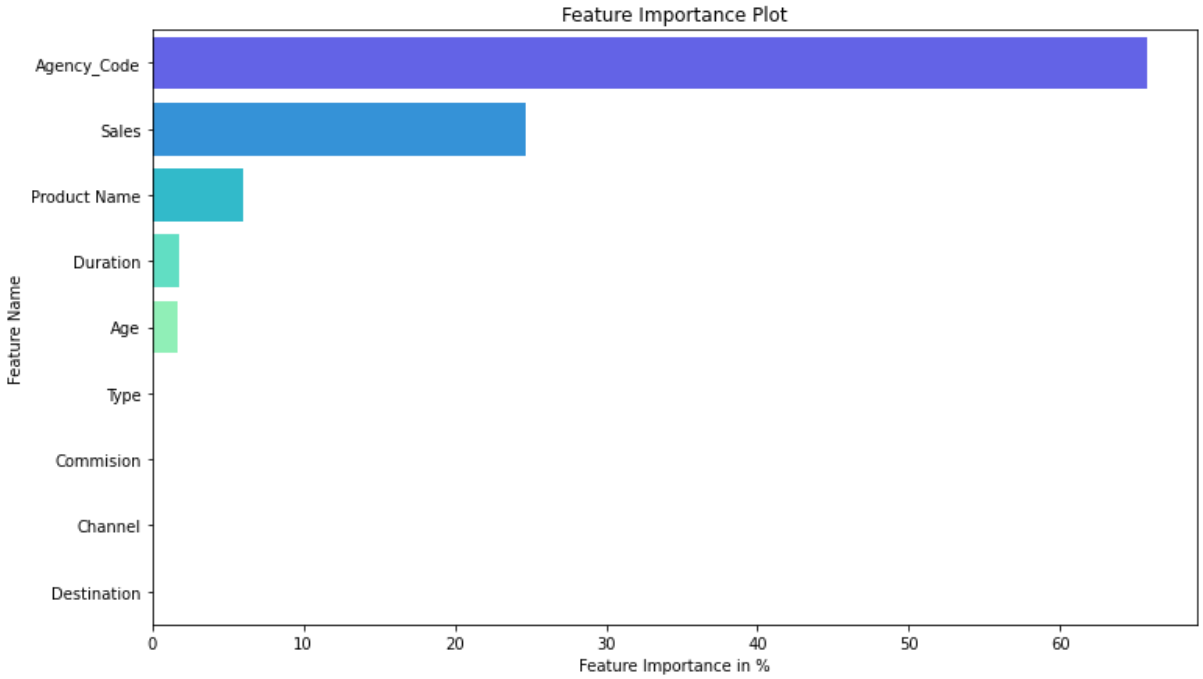


Fig 17 : Bar Plot

Insights:

- Agency_Code is the most important feature of our model followed by Sales, Product Name , Duration and Age.
- Type, Commision, Channel, Destination have about none affect in decision tree.
- The best parameters are max depth: 8, min sample leaf: 100, min sample split: 300.

We are ready to perform accuracy of the train and test samples.

Random Forest Classifier:

We will use random forest mechanism as a fresh classifier and will use fresh set of parameters for better observations. We will use Grid search for performing similar best parameter searching and not use best parameters from Decision tree. We have chosen multiple parameters, sample are:

```
GridSearchCV(cv=5, estimator=RandomForestClassifier(random_state=1),
             param_grid={'max_depth': [8, 10], 'max_features': [3, 4, 5],
                         'min_samples_leaf': [30, 50, 80],
                         'min_samples_split': [100, 150, 200],
                         'n_estimators': [100, 200]})
```

And following best parameters were found , after analysis of the data:

```
{'max_depth': 10,
 'max_features': 3,
 'min_samples_leaf': 30,
 'min_samples_split': 100,
 'n_estimators': 100}
```

WSS weightage of the data after applying above parameters to Random forest:

	Imp
Agency_Code	0.299520
Product Name	0.208599
Sales	0.162774
Commision	0.152713
Type	0.079664
Duration	0.055534
Age	0.031392
Destination	0.009806
Channel	0.000000

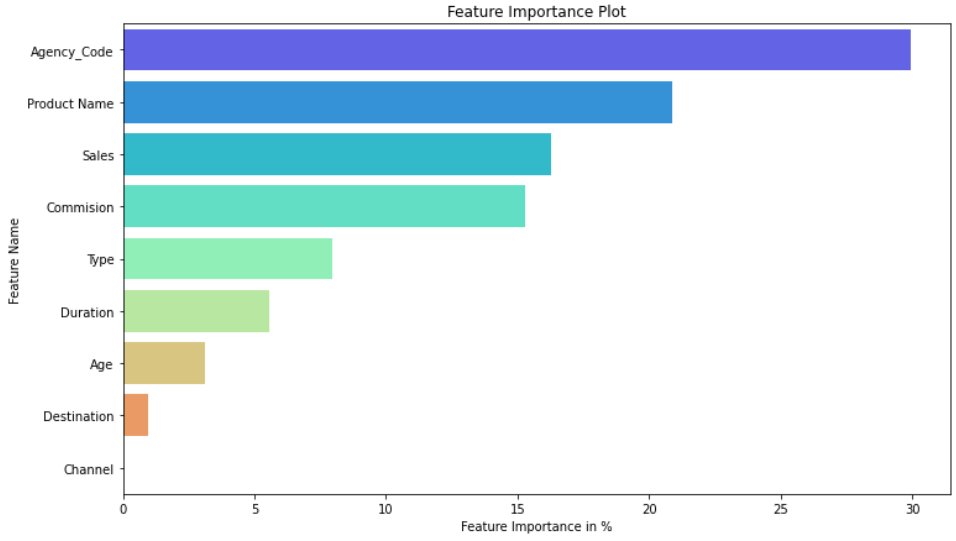


Fig 18 : Bar Plot

Artificial Neural Network Classifier:

Artificial neural network needs Scaling of the data, so we will use Standard Scalar Method for scaling the data set.

After applying scaling, we have performed ANN model. And we also used same grid search for searching the best parameters for building the ANN model.

We have found following list of parameters are best suited for the ANN model generations:

```
MLPClassifier(hidden_layer_sizes=100, max_iter=2500, random_state=1, tol=0
.01)
```

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.

CART Model

Train data FPR/TPR Chart and Confusion Matrices:

```
AUC: 0.819
Text(0, 0.5, 'True Positive Rate')
```

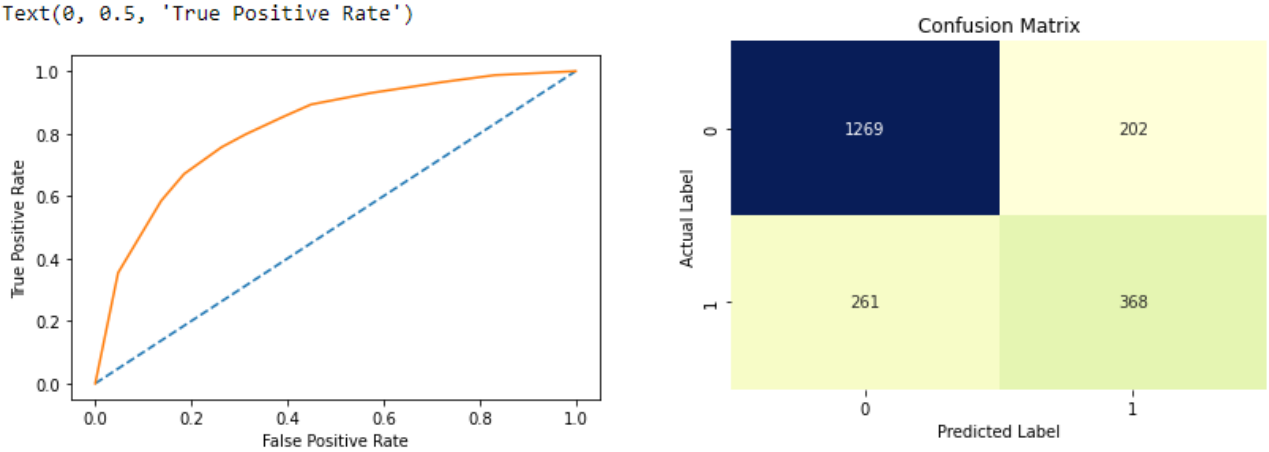


Fig 19 : AUC Curve and Classification report CART Training data

Train Data Print Classification Report:

	precision	recall	f1-score	support
0	0.83	0.86	0.85	1471
1	0.65	0.59	0.61	629
accuracy			0.78	2100
macro avg	0.74	0.72	0.73	2100
weighted avg	0.77	0.78	0.78	2100

Test data FPR/TPR Chart and Confusion Matrices:

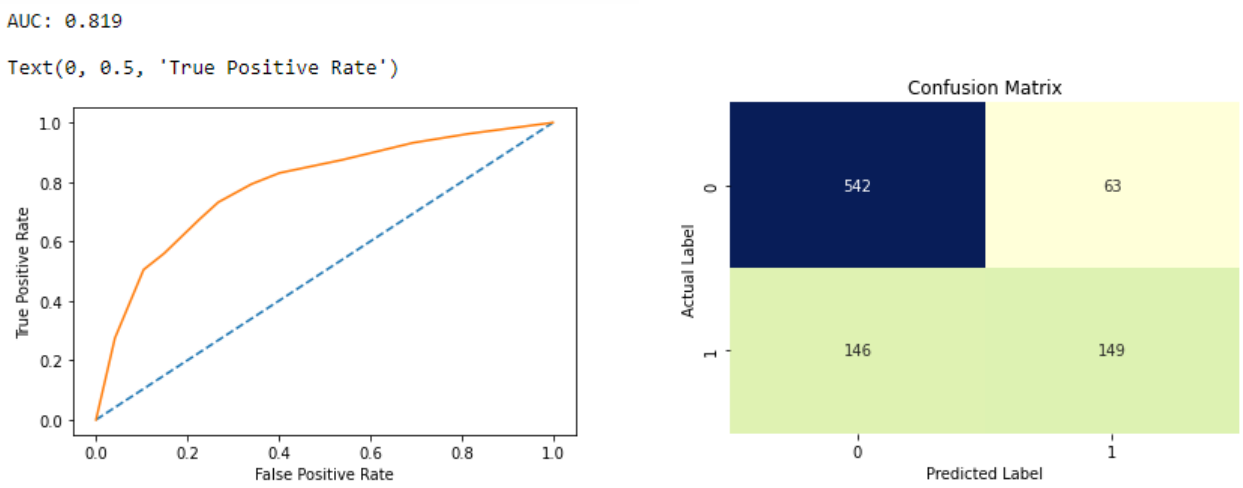


Fig 20 : AUC Curve and Classification report CART Testing data

Print Classification report for Test data set:

	precision	recall	f1-score	support
0	0.79	0.90	0.84	605
1	0.70	0.51	0.59	295
accuracy			0.77	900
macro avg	0.75	0.70	0.71	900
weighted avg	0.76	0.77	0.76	900

Insights:

- Train Data: AUC: 81.9% Accuracy: 78% Precision: 65% Recall:59% f1-Score: 61%
- Test Data: AUC: 81.9% Accuracy: 77% Precision: 70% Recall:51% f1-Score: 59%
- Training and Test set results are almost similar.
- Weightage Agency_Code 0.657599 Sales 0.246799 Product Name 0.060673 Duration 0.017588 Age 0.017342
- Agency_Code is the most important variable for predicting claim status.

Random Forest Classifier Model

Train data FPR/TPR Chart and Confusion Matrices:

```
AUC: 0.841
: Text(0, 0.5, 'True Positive Rate')
```

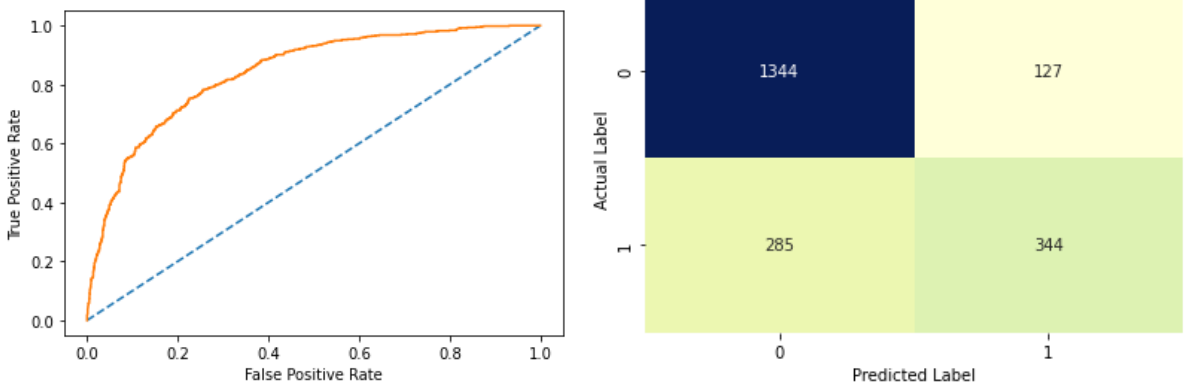


Fig 21 : AUC Curve and Classification report RF Training data

```
Classification report for Training data set is as follows :

              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1471
           1       0.73      0.55      0.63       629

    accuracy                           0.80      2100
   macro avg       0.78      0.73      0.75      2100
weighted avg       0.80      0.80      0.79      2100
```

Test data FPR/TPR Chart and Confusion Matrices:

```
AUC: 0.815
Text(0, 0.5, 'True Positive Rate')
```

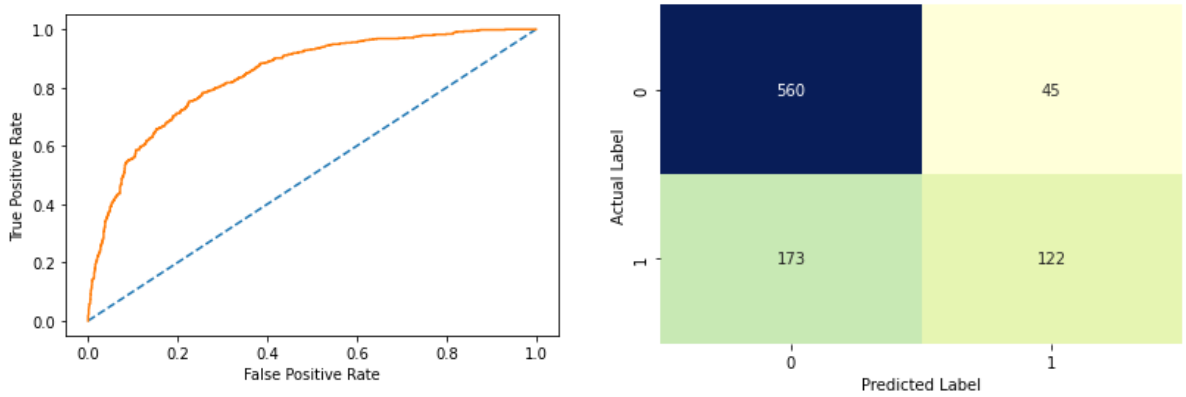


Fig 21 : AUC Curve and Classification report RF Testing data

```
Classification report for test data set is as follows :

              precision    recall  f1-score   support

           0       0.76      0.93      0.84       605
           1       0.73      0.41      0.53       295

    accuracy                           0.76       900
   macro avg       0.75      0.67      0.68       900
weighted avg       0.75      0.76      0.74       900


Accuracy Score for testing data using Randon Forest  76.0 %
Area Under Curve  for testing data using Randon Forest is 81.0 %
```

Insights:

- Train Data: AUC: 84.1% Accuracy: 80% Precision: 73% Recall:55% f1-Score: 63%
- Test Data: AUC: 81.5% Accuracy: 76% Precision: 73% Recall:41% f1-Score: 53%
- Training and Test set results are almost similar.
- Weightage Imp Agency_Code 0.299520 Product Name 0.208599 Sales 0.162774 Commision 0.152713 Type 0.079664
- Agency_Code is the most important variable for predicting claim status.

Artificial Neural Network Model

Train data FPR/TPR Chart and Confusion Matrices:

```
AUC: 0.731

Text(0, 0.5, 'True Positive Rate')
```

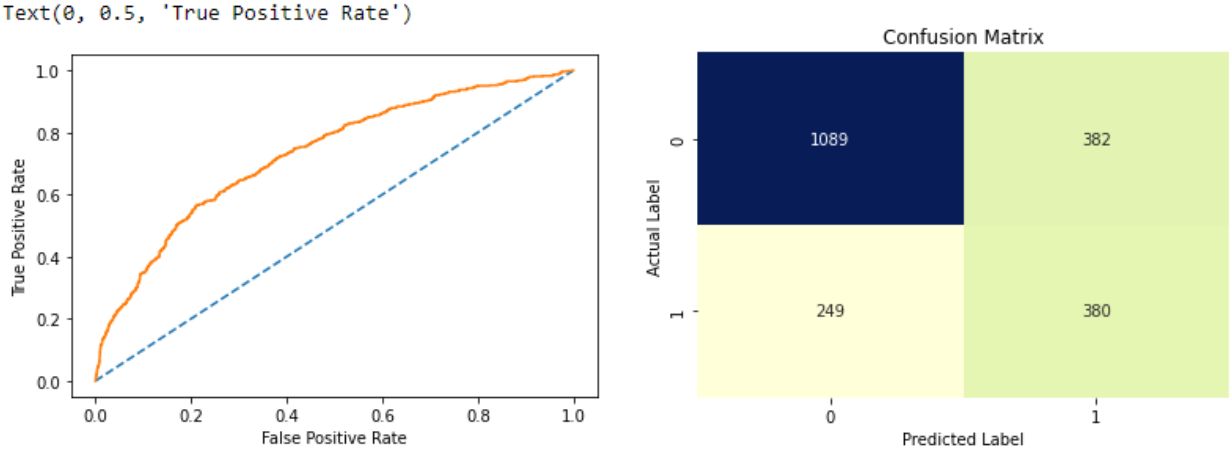


Fig 22 : AUC Curve and Classification report ANN Training data

Classification report for Training data set using ANN is :

	precision	recall	f1-score	support
0	0.81	0.74	0.78	1471
1	0.50	0.60	0.55	629
accuracy			0.70	2100
macro avg	0.66	0.67	0.66	2100
weighted avg	0.72	0.70	0.71	2100

Test data FPR/TPR Chart and Confusion Matrices:

```
AUC: 0.701

Text(0, 0.5, 'True Positive Rate')
```

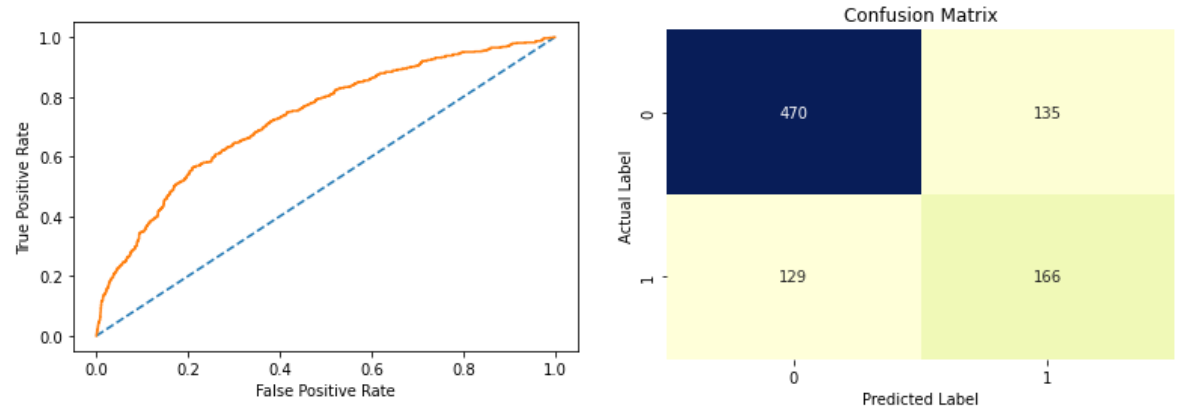


Fig 23 : AUC Curve and Classification report ANN Testing data

Classification report for test data set using ANN is :

	precision	recall	f1-score	support
0	0.78	0.78	0.78	605
1	0.55	0.56	0.56	295
accuracy			0.71	900
macro avg	0.67	0.67	0.67	900
weighted avg	0.71	0.71	0.71	900

Artificial neural network Method Insights:

Train Data: AUC: 73.1 % Accuracy: 70% Precision: 50% Recall:60% f1-Score: 55%

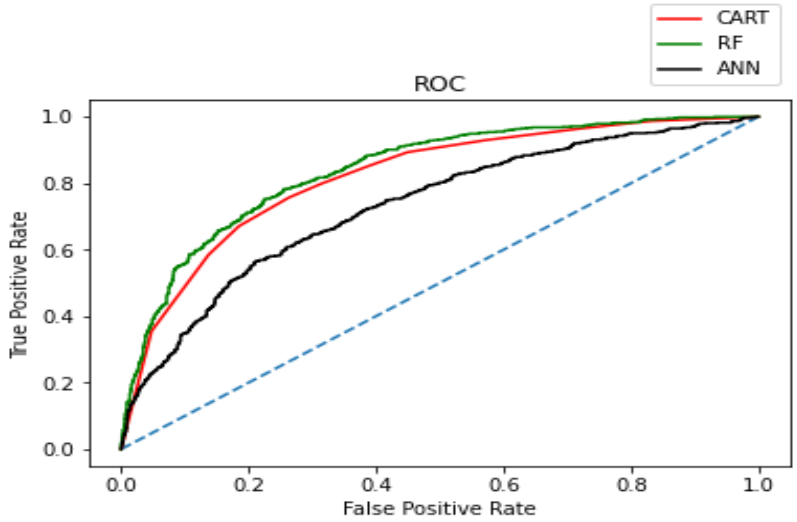
Test Data: AUC: 70.1% Accuracy: 71% Precision: 55% Recall:56% f1-Score: 56%

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

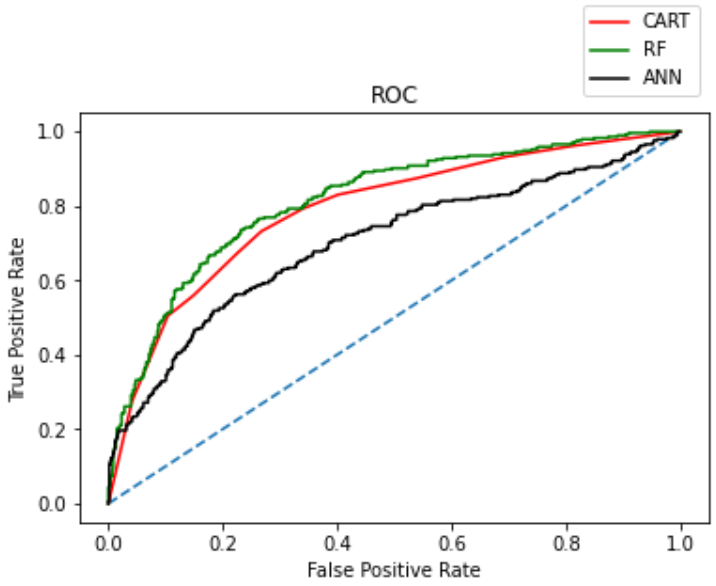
Comparison of Performance metrics from all the Models:

	Accuracy	AUC	Recall	Precision	F1 Score
CART DTree Train	0.65	0.59	0.61	0.78	0.79
Random Forest Train	0.73	0.55	0.63	0.80	0.81
Neural Network Train	0.50	0.60	0.55	0.70	0.70
CART DTree Test	0.70	0.51	0.59	0.77	0.82
Random Forest Test	0.73	0.41	0.53	0.76	0.84
Neural Network Test	0.55	0.56	0.56	0.71	0.73

Compare the AUC Curve Training data set for all three Models:



Compare the AUC Curve Testing data set for all three Models:



Insights:

- Based on Training and testing performance Matrices , we come to conclusion that, Out of the 3 models, Random Forest has slightly better performance than the Cart and Artificial Neural network model.
- All the 3 models can be used for making any future predictions.
- From Cart and Random Forest Model, the field Agency_Code is found to be the most useful feature amongst all other features for

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

Data Insights :

- Age:
  Age of insured ranges from a minimum of 8 to maximum of 84.
  The average Age of insured is around 38.1.
  The standard deviation of the Age of insured is 10.46.
  25% , 50% (median) and 75 % of the Age of insured are 32 , 36 and 42.

- Commision:
  The commission received for tour insurance firm ranges from a minimum of 0 to maximum of 99.9.
  The average Commision received for tour insurance firm is around 12.53.
  The standard deviation of the Commision received for tour insurance firm is 19.6.
  25% , 50% (median) and 75 % of the Commision received for tour insurance firm are 0 , 4.63 and 15.6.

- Duration:
  Duration of the tour ranges from a minimum of 0 to maximum of 4580.
  The average Duration of the tour is around 70.
  The standard deviation of the Duration of the tour is 134.
  25% , 50% (median) and 75 % of the Duration of the tour are 11, 26.75 and 63.

- Sales:
  Amount of sales of tour insurance policies ranges from a minimum of 0 to maximum of 539.
  The average Sales of tour insurance policies is around 60.24.
  The standard deviation of the Sales of tour insurance policies is 70.73.
  25% , 50% (median) and 75 % of the Sales of tour insurance policies are 20 , 33 and 69.

- Agency_Code
  There are 4 Agency_Code present in the data set named as 'EPX' , 'C2B' , 'CWT' , 'JZI'.
  Maximum no of 45% customers have Agency_Code 'EPX'
  Minimum 7% customers have Agency_Code 'JZI'

- Type
  61.2% customers prefer Travel Agency as their tour insurance firm.
  38.7% customers prefer Airlines as their tour insurance firm.

- Claimed
  69.2 % didn't Claim their insurance.
  30.8 % Claim their insurance.

- Channel
  98.4% customers choose online channel.
  1.53% customers choose offline channel.

- Product Name
  maximum 37.86% customers purchased Customized Plan.
  Minimum Only 3.6% customers purchased Gold Plan.

- Destination
  About 82 % customers, maximum choose Asia as Destination of the tour.
  Only 7.1% customers choose Europe as Destination of the tour.

Business Insights:

- Customers with Agency Code C2B claimed more insurance. 560 Claimed
- Most of Customers with Agency Code EPX didn't claimed insurance. 1172 no claimed vs 193 Claimed
- In online channel no claimed status is more than claimed Status. 2047 No Claimed vs 907 Claimed
- Silver Plan has maximum number of Claimed as 306
- Agency EPX sold maximum Products as Customized Plan (687) and Cancellation Plan (678)
- Travel Agency sold maximum Customized Plan as 1076
- Offline Gold Plan were sold least times Count is 2
- Online Customized Plan were sold maximum as 1092
- Travel Agency name EPX has done maximum number of bookings Count is 1365 and they have done maximum booking For Asia Destination Count is 1128
- There is no offline Booking for Airlines Types
- no one has done offline booking for Europe Destination
- Most of the customers choose Asia as Destination of the tour.
- Agency_Code is found to be the most useful feature amongst all other features for predicting if a customer has claim insurance or not.
- Maximum customers have Agency_Code 'EPX', which means customer prefer to book their Travel using EPX Agency.
- Least number of customers use Agency_Code 'JZI'
- most of the customers prefer Travel Agency as their tour insurance firm.
- Most of the customers prefer Airlines as their tour insurance firm.
- Most of the customers choose online channel for doing their insurance.

- Very few customers choose offline channel for doing their insurance.
- Commision with Sales show strong correlation with 0.69. So As comission increases, Sales also increased.

Recommendations:

- Need to train the JZI agency resources to pick up sales as they are least preferred channel of Tour booking,

- need to run promotional marketing campaign or evaluate if we need to tie up with alternate agency in order to attract more customers.

- Also based on the model we are getting 75-78% accuracy, so we need customer books airline tickets or plans, cross sell the insurance based on the claim data pattern.

- Other interesting fact is more sales happen via Travel Agency than Airlines and the trends shows the claim are processed more at Airline. So we need to understand the why we have more claims from Airlines and how to increase sales from Airlines.