PICTURE URL: HTTPS://WWW.MYGREATLEARNING.COM/BLOG/TOP-DATA-MINING-TOOLS

Project - Data Mining

Debsmita Chakraborty
BATCH-JULY C

Table of Contents:

PROBLEM 1: CLUSTERING- BANK MARKETTING

PROBLEM 1: CLUSTERING- INSURANCE

GRAPHS and CHARTS:

PROBLEM 1: CLUSTERING- BANK MARKETTING

PROBLEM 2: CLUSTERING- BANK MARKETTING

PROBLEM 1: CLUSTERING

BANK MARKETTING

URL-https://blog.vimarketingandbranding.com/5-ways-marketing-can-benefit-bank

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 idea of the project is to identify the segments that a bank should choose to give better benefits /promotional offers to its customers depending on the data given.

The data is of 210 rows and 7 columns with no null values.

The data looks like below:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837
5	12.70	13.41	0.8874	5.183	3.091	8.456	5.000
6	12.02	13.33	0.8503	5.350	2.810	4.271	5.308
7	13.74	14.05	0.8744	5.482	3.114	2.932	4.825
8	18.17	16.26	0.8637	6.271	3.512	2.853	6.273
9	11.23	12.88	0.8511	5.140	2.795	4.325	5.003

Table 1.1- Data Sample

Data Dictionary for Market Segmentation:

- The data columns are described as below and will be used with the initial assigned variables as stated:
- spending: Amount spent by the customer per month (in 1000s)
- advance payments: Amount paid by the customer in advance by cash (in 100s)
- probability_of_full_payment: Probability of payment done in full by the customer to the bank
- current balance: Balance amount left in the account to make purchases (in 1000s)
- credit limit: Limit of the amount in credit card (10000s)
- min_payment_amt: minimum paid by the customer while making payments for purchases made monthly (in 100s)
- max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s)

The data has the entries all as a float type:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Table 1.2- Data types and entries

Descriptive Analysis

We will first begin with UNIVARIATE ANALYSIS of data:

	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

Table 1.3- Descriptive Analysis

From the above table 1.3 we infer:

- The mean, values and median of spending and advance_payments are nearly equal.
- The mean, values and median for current_balance, credit_limit, min_payment_amt and max_spent_in_single_shopping is very near to each other.
- Standard deviation is high for spending variable which is 2.909
- The probability_of_full_payment has the smallest values in all the above mean, median and maximum amongst all categories.

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

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

We will analyse the data in detail now through UNIVARIATE ANALYSIS:

Spending

Minimum spending: 10.59
Maximum spending: 21.18
Mean value: 14.847523809523818
Median value: 14.355
Standard deviation: 2.909699430687361
Null values: False

credit_limit

Minimum credit_limit: 2.63
Maximum credit_limit: 4.033
Mean value: 3.258604761904763
Median value: 3.237
Standard deviation: 0.37771444490658734
Null values: False

Advance_payments

advance_payments: 12.41
advance_payments: 17.25
Mean value: 14.559285714285727
Median value: 14.32
Standard deviation: 1.305958726564022
Null values: False

min_payment_amt

Minimum min_payment_amt: 0.7651
Maximum min_payment_amt: 8.456
Mean value: 3.7002009523809503
Median value: 3.599
Standard deviation: 1.5035571308217792
Null values: False

probability_of_full_payment

Minimum probability_of_full_payment 0.8081
Maximum probability_of_full_payment: 0.9183
Mean value: 0.8709985714285714
Median value: 0.8734500000000001
Standard deviation: 0.0236294165838465
Null values: False

min_payment_amt

Minimum max_spent_in_single_shopping: 4.519
Maximum max_spent_in_single_shoppings: 6.55
Mean value: 5.408071428571429
Median value: 5.223000000000001
Standard deviation: 0.49148049910240543
Null values: False

current_balance

Minimum current_balance: 4.899
Maximum current_balance: 6.675
Mean value: 5.628533333333335
Median value: 5.5235
Standard deviation: 0.44306347772644944
Null values: False

Table 1.4- Descriptive Analysis

Inferences from above:

- Advance payment has the highest minimum value of 12.41 and the least is for probability of full payment at 0.8081.
- There are no null values in any of the data we are dealing with.
- Maximum spending has the highest maximum value of 21.18 and probability of full payment is the least at 0.91.83.
- Spending has the highest standard deviation 2.909 and probability of full payment the least 0.0236.
- Highest median value is of Advance payments at 14.32 and mean is the highest for spending 14.847.

Spending

```
spending - 1st Quartile (Q1) is:  12.27
spending - 3st Quartile (Q3) is:  17.305
Interquartile range (IQR) of spending is  5.035
```

Advance_payments

```
advance_payments - 1st Quartile (Q1) is:  13.45
advance_payments - 3st Quartile (Q3) is:  15.715
Interquartile range (IQR) of advance_payments is  2.2650000000000006
```

probability_of_full_payment

```
probability_of_full_payment - 1st Quartile (Q1) is:  0.8569
probability_of_full_payment - 3st Quartile (Q3) is:  0.887775
Interquartile range (IQR) of probability_of_full_payment is  0.030874999999999986
```

current_balance

```
current_balance - 1st Quartile (Q1) is:  5.26225
current_balance - 3st Quartile (Q3) is:  5.97975
Interquartile range (IQR) of current_balance is  0.7175000000000002
```

credit_limit

```
credit_limit - 1st Quartile (Q1) is:  2.944
credit_limit - 3st Quartile (Q3) is:  3.56175
Interquartile range (IQR) of credit_limit is  0.61775
```

min_payment_amt

```
min_payment_amt - 1st Quartile (Q1) is:  2.5615
min_payment_amt - 3st Quartile (Q3) is:  4.76875
Interquartile range (IQR) of min_payment_amt is  2.2072499999999997
```

max_spent_in_single_shopping

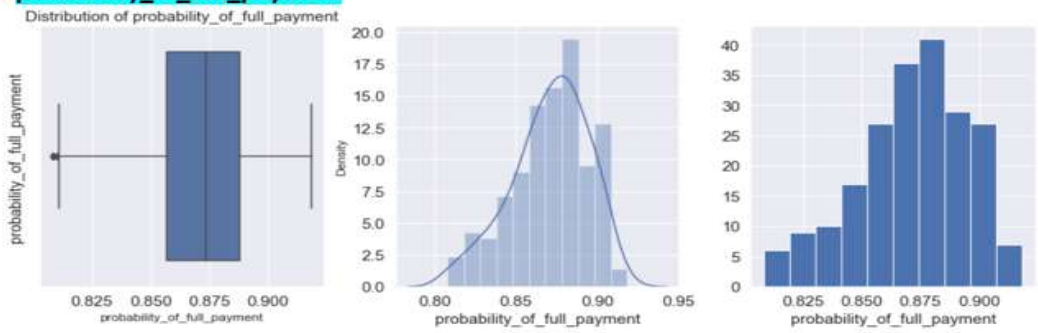
```
max_spent_in_single_shopping - 1st Quartile (Q1) is:  5.045
max_spent_in_single_shopping - 3st Quartile (Q3) is:  5.877
Interquartile range (IQR) of max_spent_in_single_shopping is  0.8319999999999999
```

Table 1.5- Quartiles for all the data

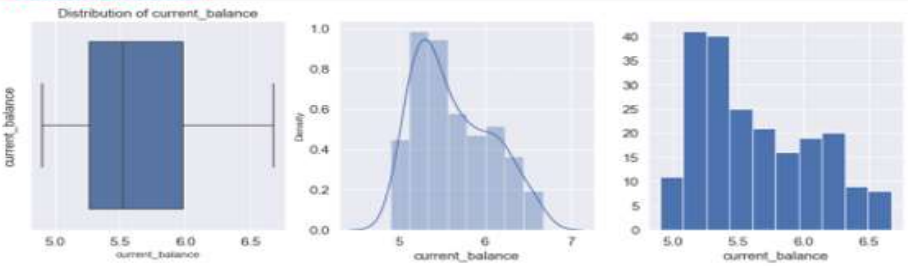
Inferences from above:

- Spending is the highest with 5.035.
- The probability_of_full_payment is the least at 0.0308.

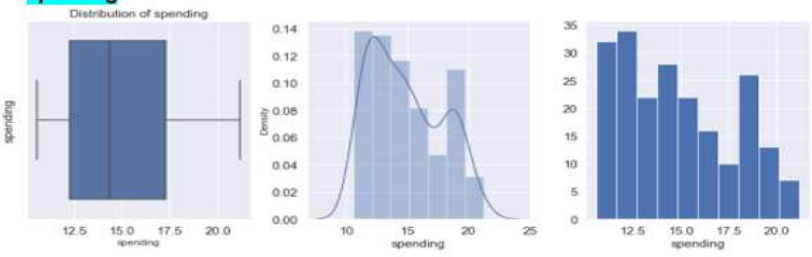
probability_of_full_payment



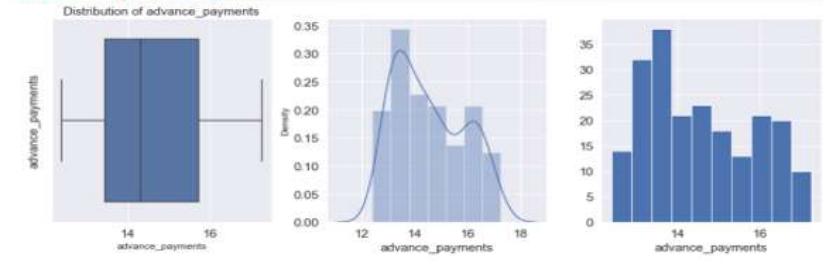
current_balance



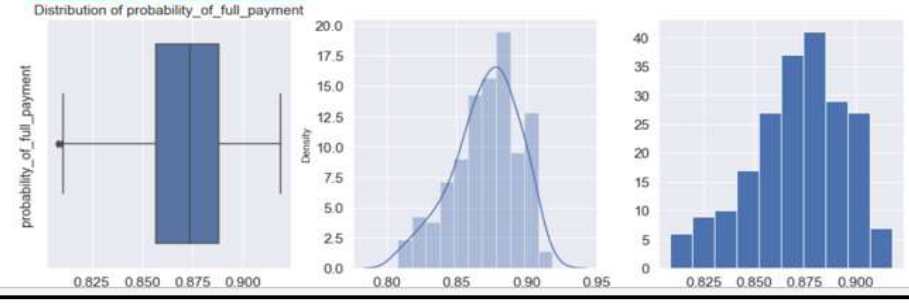
Spending

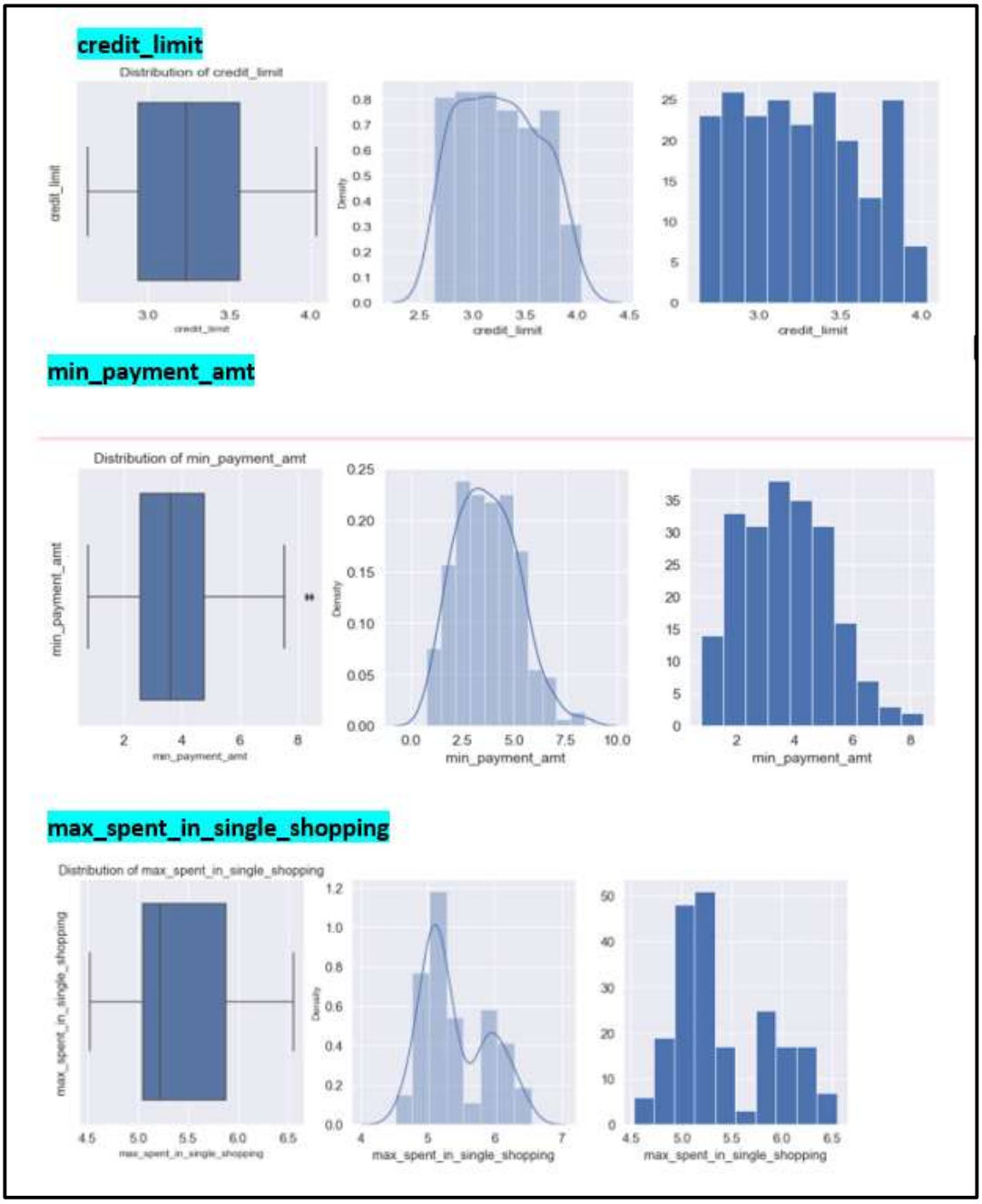


Advance_payments



probability_of_full_payment





1.6- Graphs/Plots

INFERENCE:

Advance_payments

- advance_payments - 1st Quartile (Q1) is: 13.45
- advance_payments - 3st Quartile (Q3) is: 15.715
- Interquartile range (IQR) of advance_payments is 2.2650000000000006
- Lower outliers in advance_payments: 10.052499999999998
- Upper outliers in advance_payments: 19.1125
- pending - 1st Quartile (Q1) is: 12.27

Spending:

- spending - 3st Quartile (Q3) is: 17.305
- Interquartile range (IQR) of spending is 5.035
- Number of outliers in spending upper: 0
- Number of outliers in spending lower: 0
- % of Outlier in spending upper: 0 %
- % of Outlier in spending lower: 0 %
-

probability_of_full_payment

- probability_of_full_payment - 1st Quartile (Q1) is: 0.8569
- probability_of_full_payment - 3st Quartile (Q3) is: 0.887775
- Interquartile range (IQR) of probability_of_full_payment is 0.030874999999999986
- Lower outliers in probability_of_full_payment: 0.8105875
- Upper outliers in probability_of_full_payment: 0.9340875
- Number of outliers in probability_of_full_payment upper: 0
- Number of outliers in probability_of_full_payment lower: 3
- % of Outlier in probability_of_full_payment upper: 0 %
- % of Outlier in probability_of_full_payment lower: 1 %

current_balance

- current_balance - 1st Quartile (Q1) is: 5.26225
- current_balance - 3st Quartile (Q3) is: 5.97975
- Interquartile range (IQR) of current_balance is 0.7175000000000002
- Lower outliers in current_balance: 4.186
- Upper outliers in current_balance: 7.056000000000001
- Number of outliers in current_balance upper: 0
- Number of outliers in current_balance lower: 0
- % of Outlier in current_balance upper: 0 %
- % of Outlier in current_balance lower: 0 %

Credit_limit

- credit_limit - 1st Quartile (Q1) is: 2.944
- credit_limit - 3st Quartile (Q3) is: 3.56175
- Interquartile range (IQR) of credit_limit is 0.61775
- Lower outliers in credit_limit: 2.017375
- Upper outliers in credit_limit: 4.488375
- Number of outliers in credit_limit upper: 0
- Number of outliers in credit_limit lower: 0
- % of Outlier in credit_limit upper: 0 %
- % of Outlier in credit_limit lower: 0 %

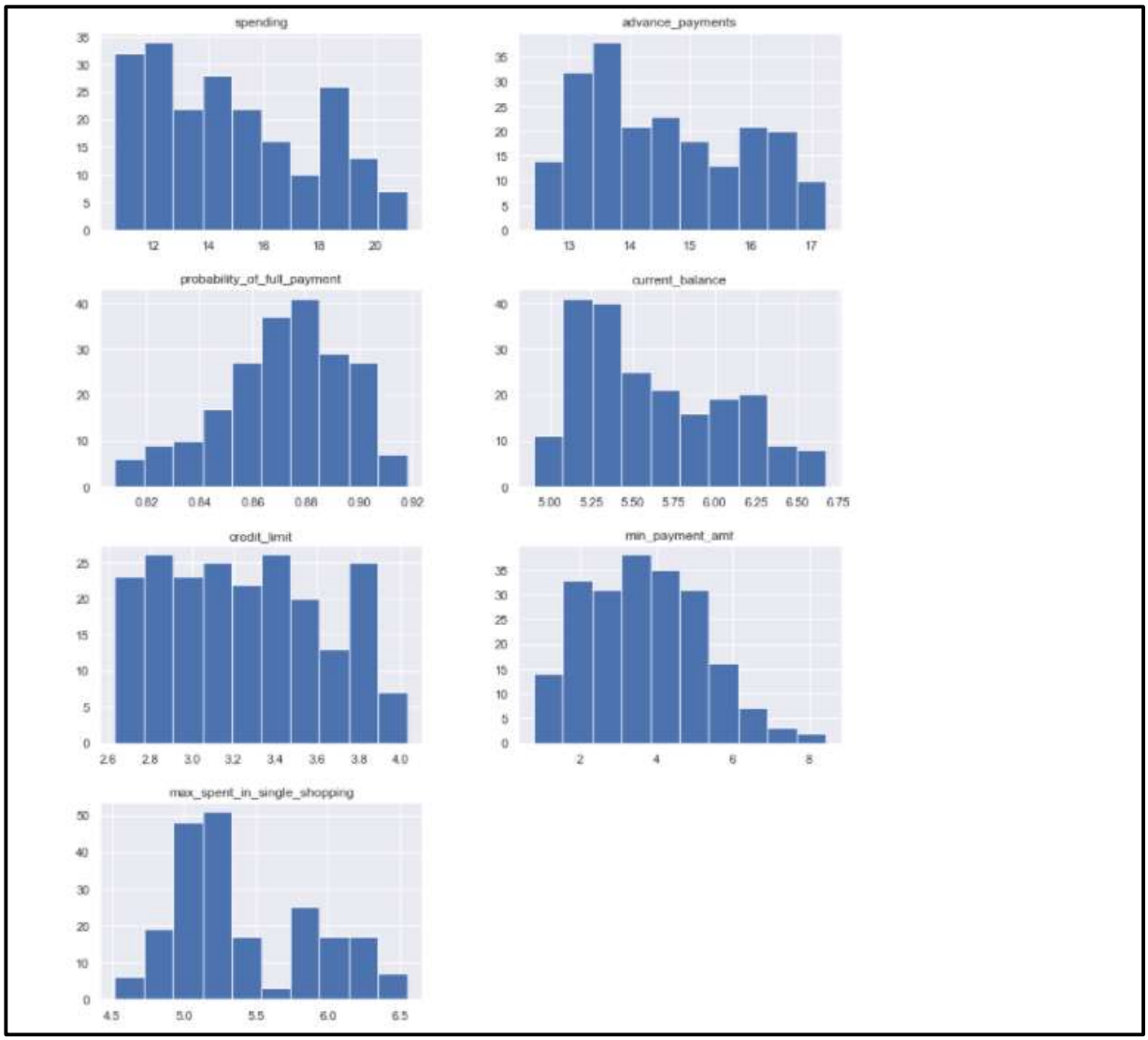
min_payment_amt

- min_payment_amt - 1st Quartile (Q1) is: 2.5615
- min_payment_amt - 3st Quartile (Q3) is: 4.76875
- Interquartile range (IQR) of min_payment_amt is 2.2072499999999997
- Lower outliers in min_payment_amt: -0.7493749999999992
- Upper outliers in min_payment_amt: 8.079625
- Number of outliers in min_payment_amt upper : 2
- Number of outliers in min_payment_amt lower : 0
- % of Outlier in min_payment_amt upper: 1 %
- % of Outlier in min_payment_amt lower: 0 %

max_spent_in_single_shopping

- max_spent_in_single_shopping - 1st Quartile (Q1) is: 5.045
- max_spent_in_single_shopping - 3st Quartile (Q3) is: 5.877
- Interquartile range (IQR) of max_spent_in_single_shopping is 0.8319999999999999
- Lower outliers in max_spent_in_single_shopping: 3.797
- Upper outliers in max_spent_in_single_shopping: 7.125
- Number of outliers in max_spent_in_single_shopping upper: 0
- Number of outliers in max_spent_in_single_shopping lower: 0
- % of Outlier in max_spent_in_single_shopping upper: 0 %
- % of Outlier in max_spent_in_single_shopping lower: 0 %

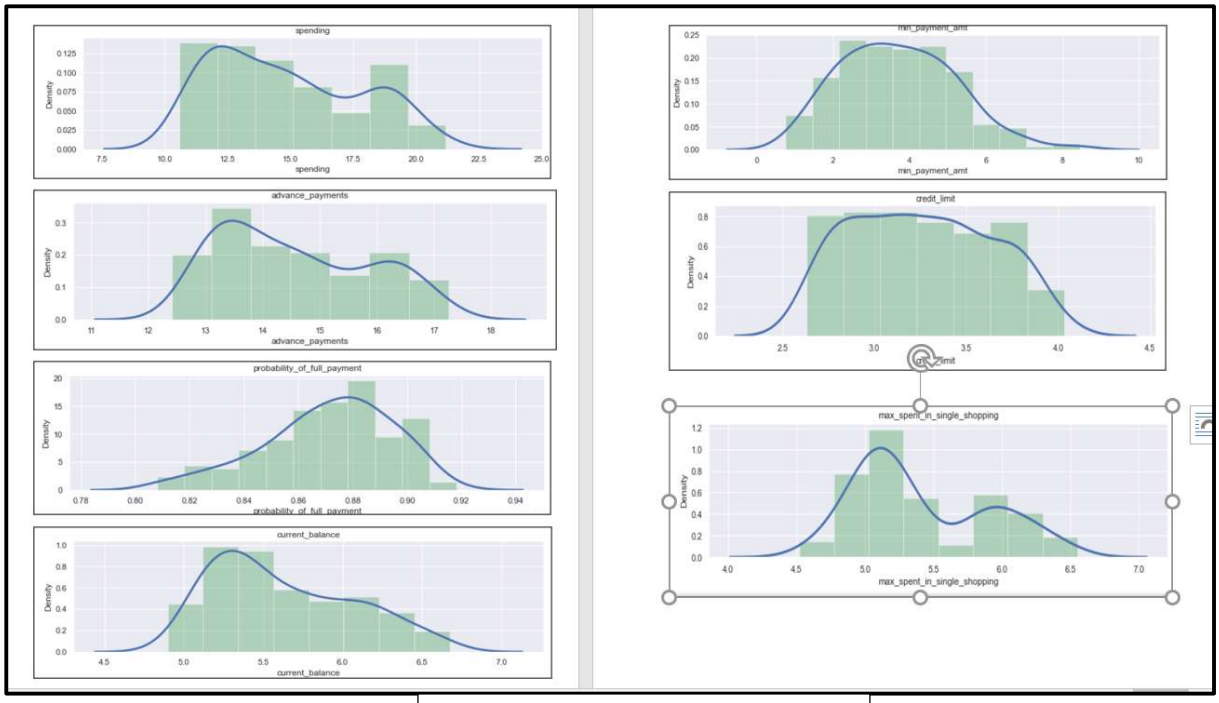
Plotting histogram to check independent variables:



1.7- Histogram

INFERENCE:

- Here we can see category wise maximum and minimum values.



1.8- KDE PLOT

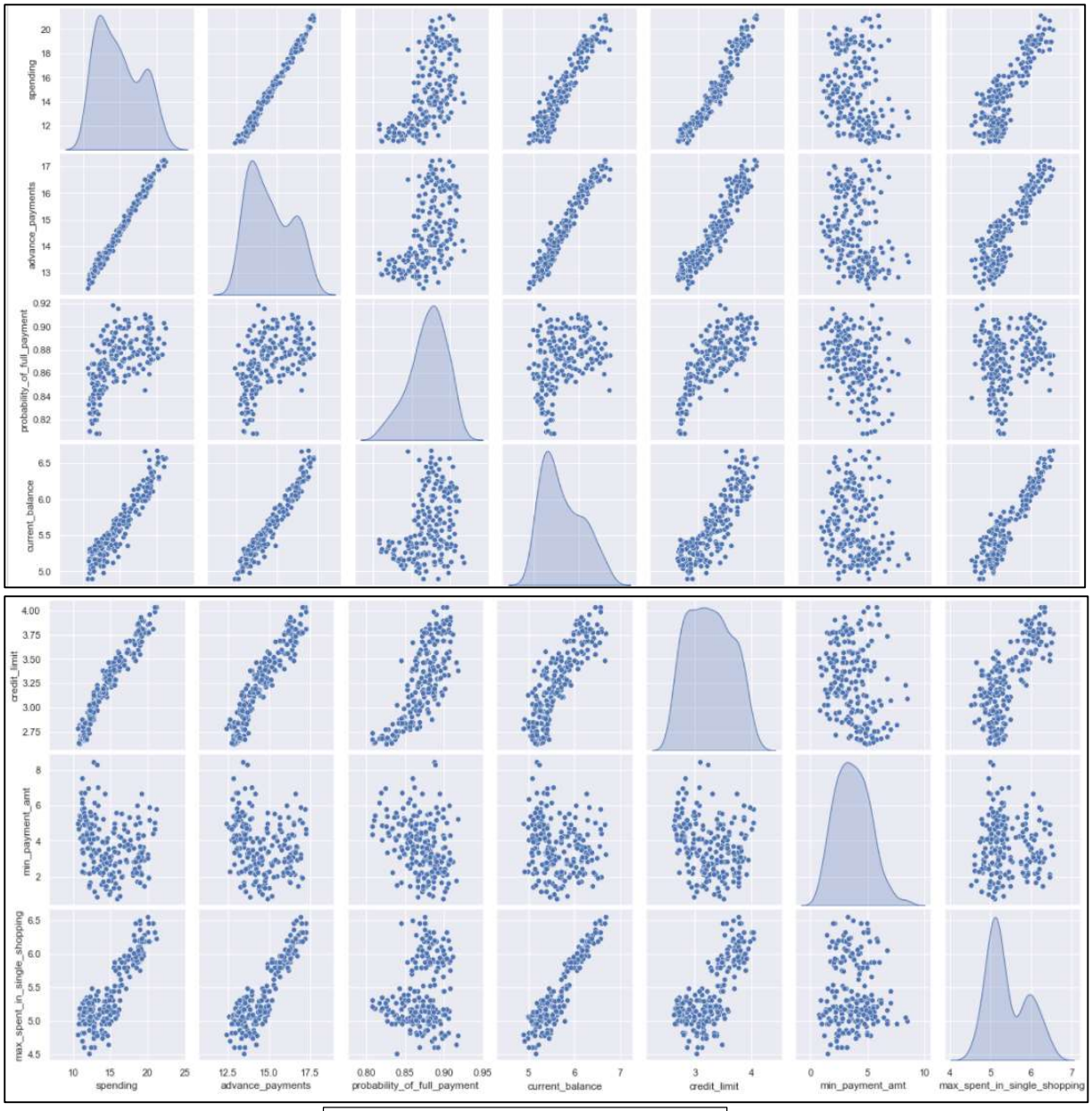
Inference:

The skewness is as below of the data:

- 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

We will analyse the data in detail now through MULTIVARIATE ANALYSIS:

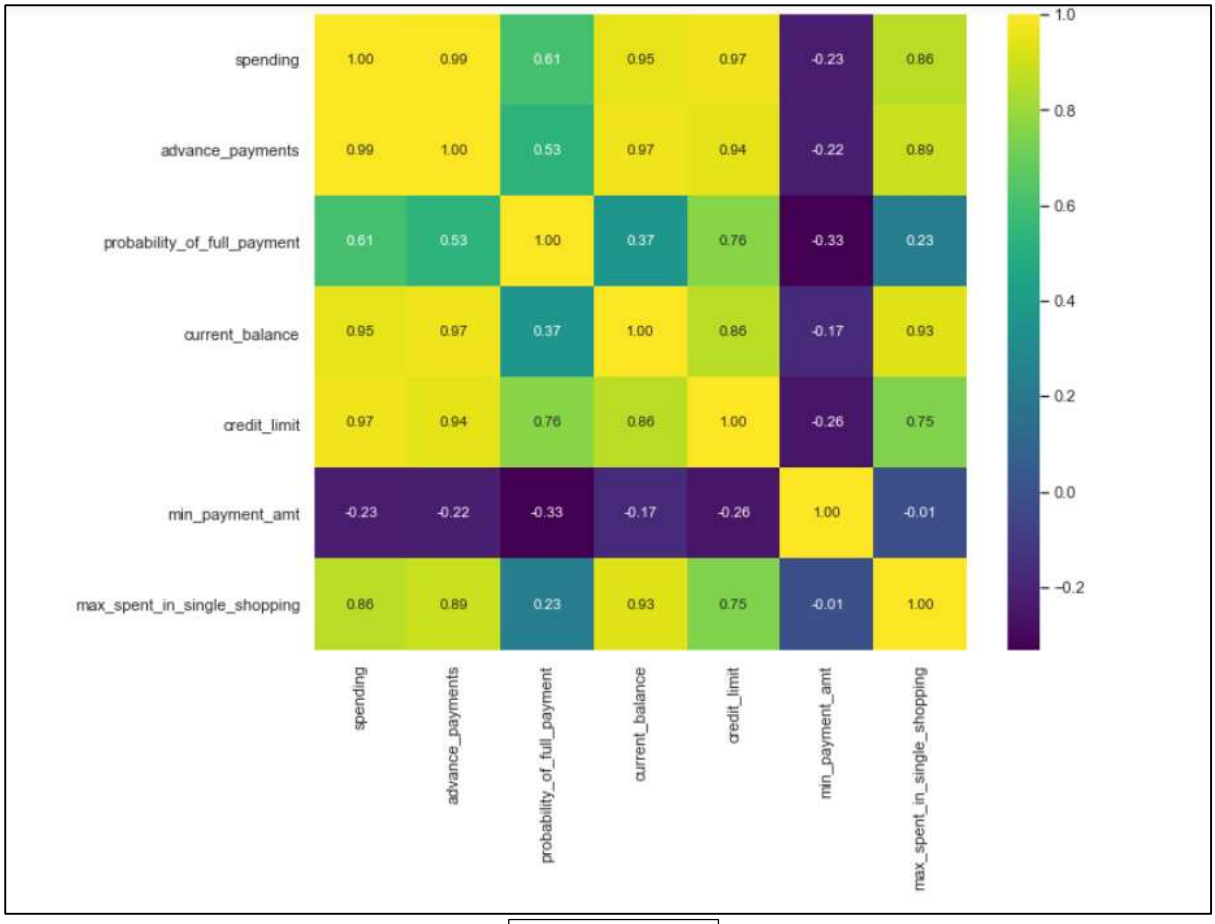
Multivariate analysis (MVA) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.



1.10- PAIR PLOT

Inference from below pair plot figure 1.10:

- 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

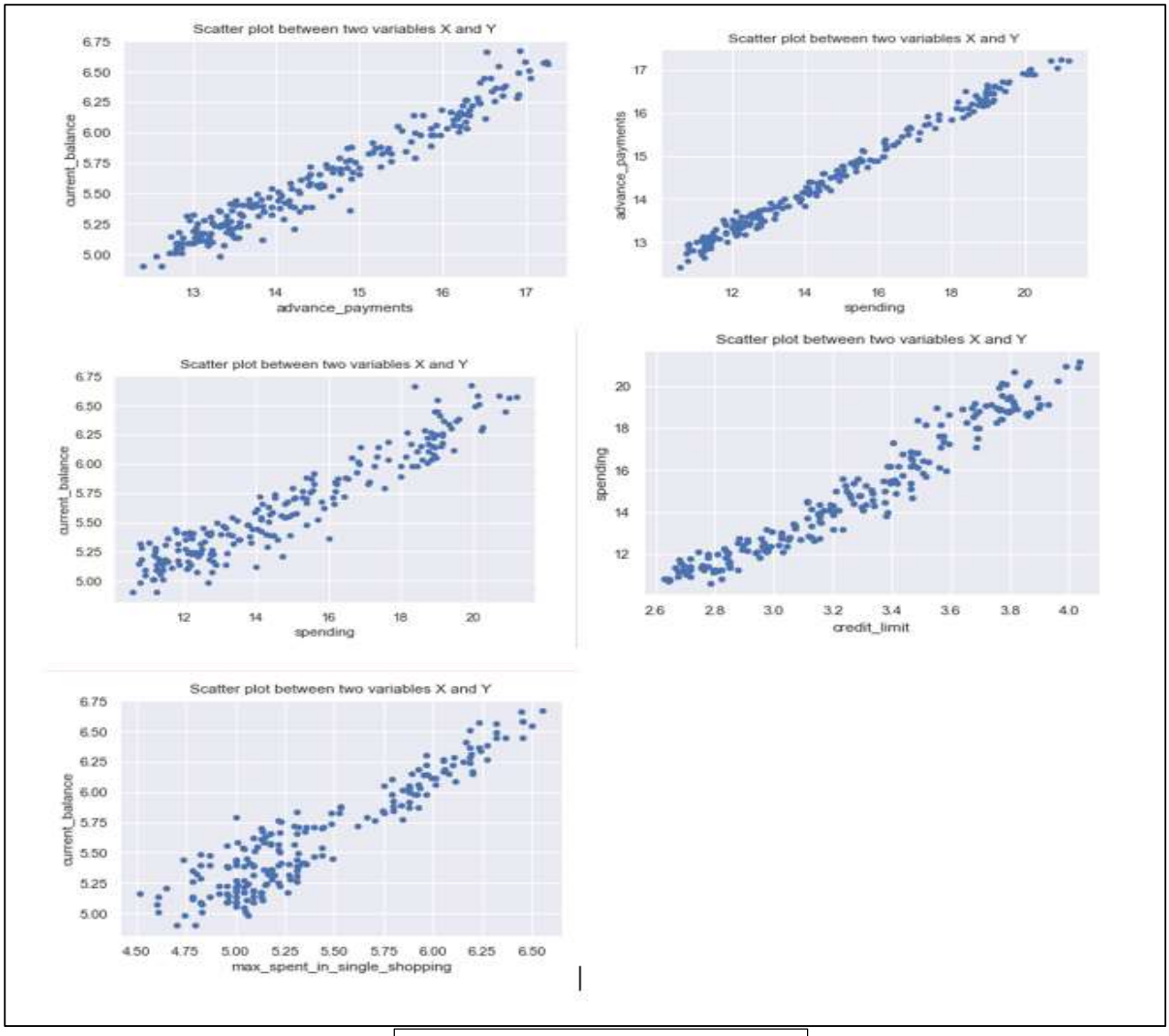


1.10- Heatmap

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

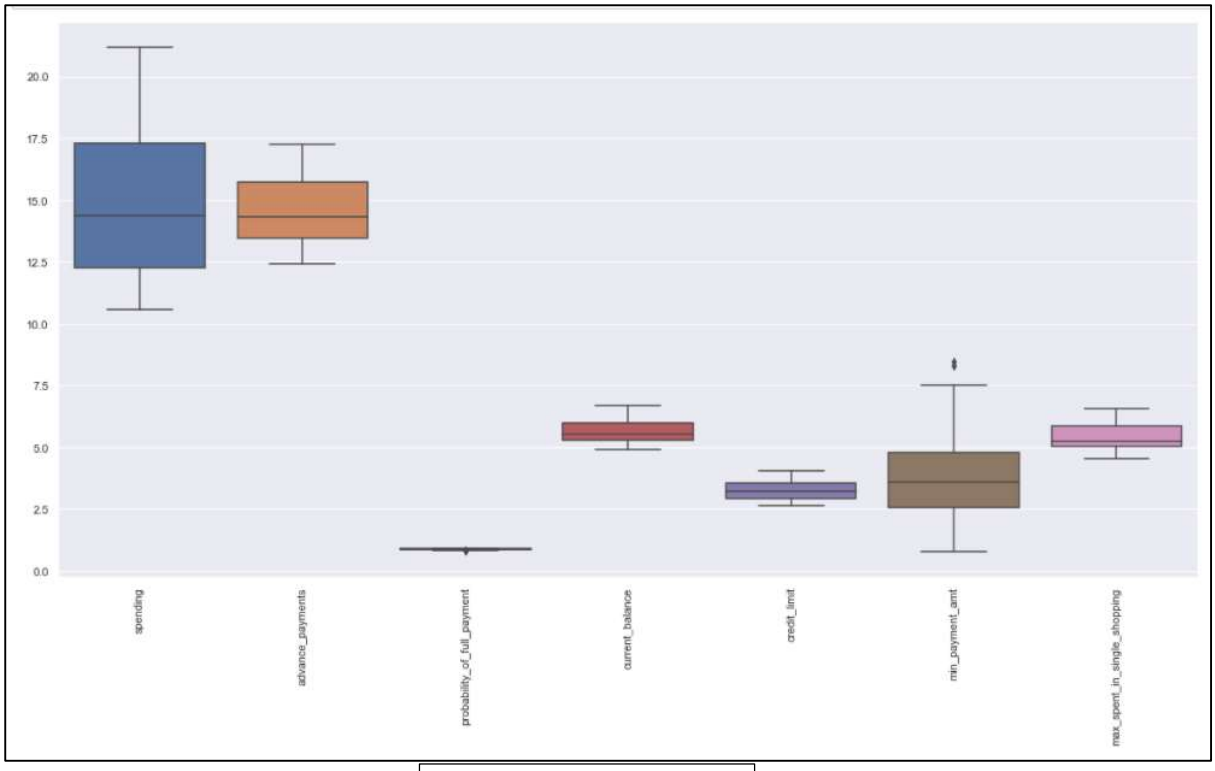
We will analyse the data in detail now through BIVARIATE ANALYSIS:



1.11- Scatter PLOT

Inference:

- The scatter plot shows relationship between the different variants in the data.
- The data used here is to show co-relation between the variants in the x and y axis.
- The figures have been represented in a way to show the co-relation between co-related categories.



1.12-Outliers

INFERENCE:

- Data has no visible outliers except for min_payment_amt.
- However, post coding we get the below result for outliers under each head
  - No. of outliers in probability_of_full_payment:3
  - No. of outliers in min_payment_amt: 2
  - No of attributes with outliers are: 2

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

Yes, it is necessary to normalize data before performing any further analysis as the data has variables ranging differently. As we see below after using the describe function, we see the data has huge variations of the mean, median and mode of one variable to the other:

	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

1.2.1-Describe function

- We will go to perform **ZScore** test which will help us scale and know how many standard deviations is the point away from the mean and also the direction.
- Brought data in the range +3 to -3.
- The covariance can also been seen having a lot of variations amongst the different columns in the data.

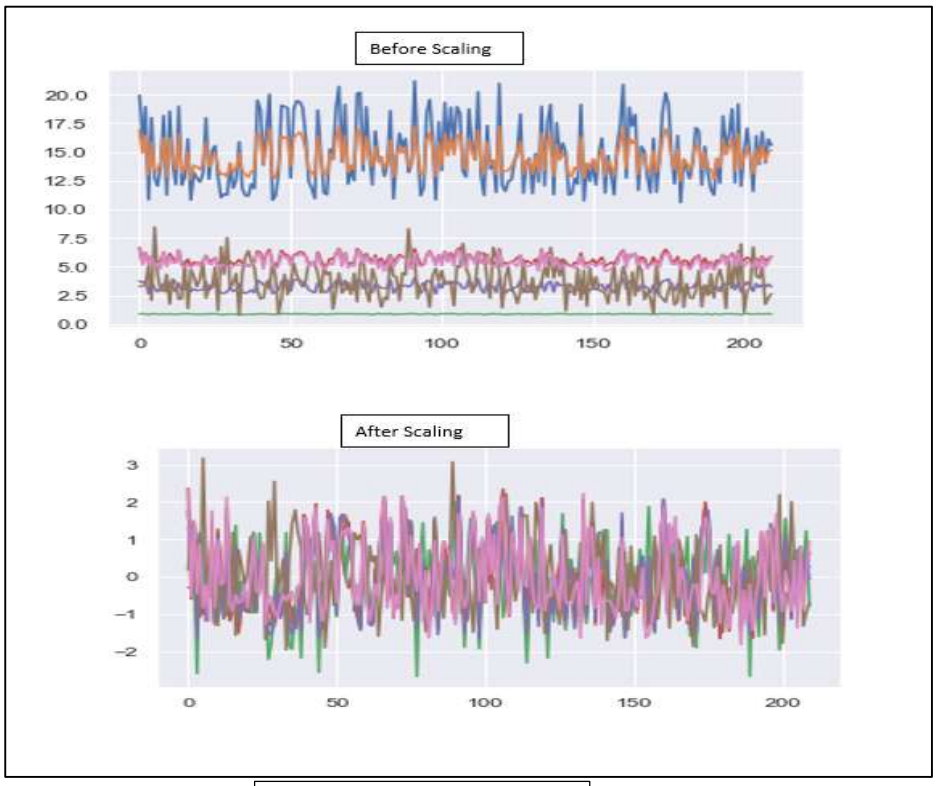
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
spending	8.466351	3.778443	0.041823	1.224704	1.066911	-1.004356	1.235133
advance_payments	3.778443	1.705528	0.016332	0.562666	0.466065	-0.426766	0.571753
ability_of_full_payment	0.041823	0.016332	0.000558	0.003852	0.006798	-0.011777	0.002634
current_balance	1.224704	0.562666	0.003852	0.196305	0.143992	-0.114290	0.203125
credit_limit	1.066911	0.466065	0.006798	0.143992	0.142668	-0.146543	0.139068
min_payment_amt	-1.004356	-0.426766	-0.011777	-0.114290	-0.146543	2.260684	-0.008187
ent_in_single_shopping	1.235133	0.571753	0.002634	0.203125	0.139068	-0.008187	0.241553

1.2.2-Covariance result

- In scaling the below data numbers did not change only the difference of units were bought to same scale.

	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

1.2.3-ZScore result



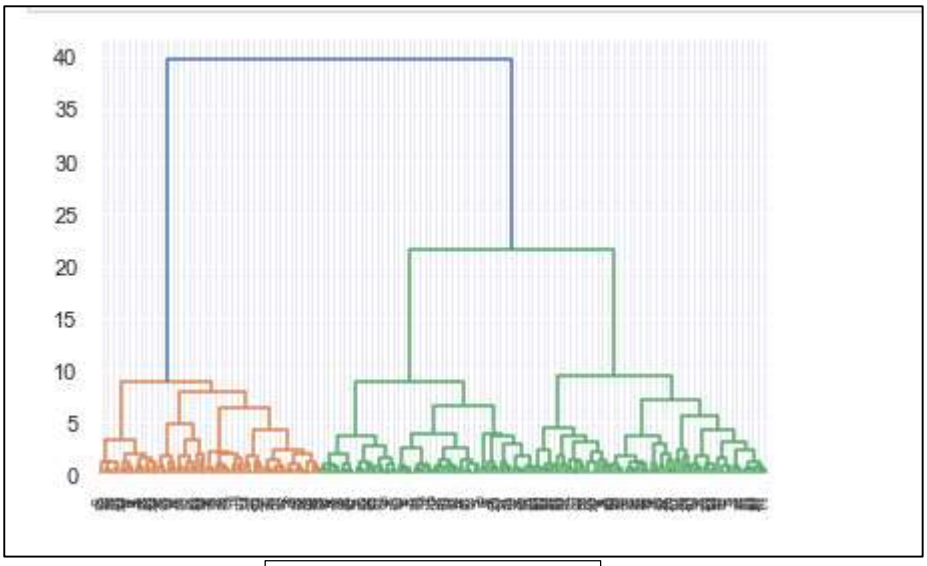
1.2.4-Scaling data

Inference:

- As we see in pre-scaled data, we have diversified variables shown in two bifurcations.
- Post scaling the data range is between -3 to +3 and a single graph represents the data with lesser variability. Scaling will have all the values in the relative same range.

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

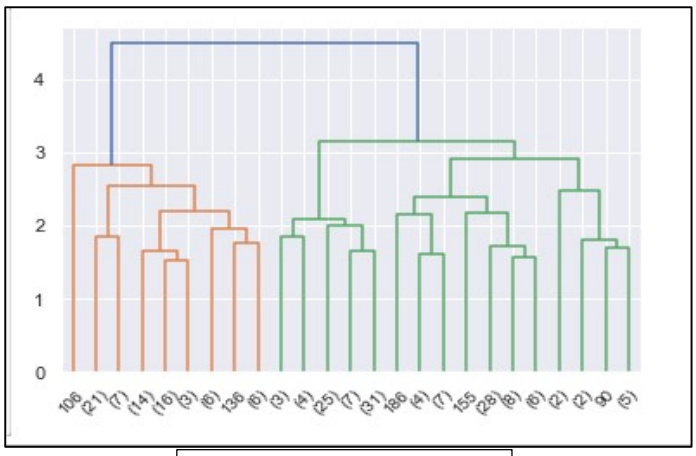
Hierarchical Clustering means we build a hierarchy of clusters. We are using WARD linkage here as the data has variations. WARD linkage specifies the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster. As ward method makes one cluster containing all objects. At each step, the process makes a new cluster that minimizes variance, measured by an index called E (also called the sum of squares index).



1.3.1-Dendrogram

Inference:

- From the above figure if we draw a line between y=10 horizontally we get three clusters. Hence this figure tells us that we need to have three clusters for work.
- Since the difference between y=10 to y=40 is high we will stick to drawing a line horizontally from y=10 axis.



1.3.2-Dendrogram-3 clusters

Inference: And three group cluster solution gives a pattern based on high/medium/low spending with max_spent_in_single_shopping (high value item) and probability_of_full_payment(payment made)

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.

The K-means algorithm **identifies k number of centroids**, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid. We will be using K-means and silhouette score for determining clusters and studying the data within them.

The WSS values for all the clusters is as below starting from cluster 1 to 10.

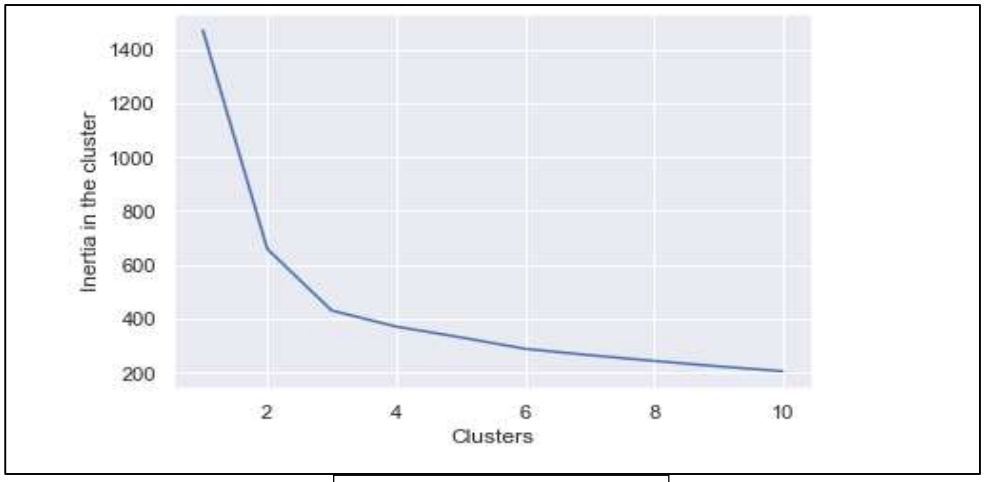
```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.18461253510196,
 331.2409459349326,
 288.76945770226405,
 265.23860056589234,
 243.96538449120453,
 223.7033873288709,
 205.38240664557657]
```

1.4 -Inertia in the cluster

Inference:

- K_means for cluster 3 is 430.658.
- K_means for cluster 4 is 371.184.
- As the value of K increases, **there will be fewer elements in the cluster**.

The technique to determine K, the number of clusters, is called the elbow method.



1.4.1-Inertia in the cluster

Inference:

- Figure 1.4.1 is the elbow method wherein we have taken inertia in clusters in the y-axis and clusters formed in the x axise.
- The figure above shows us after point 3 we have a huge fall in clusters. Hence point 3 is the opti mum clusters we can use.
- The K_means for cluster numbers used 3 is 430.65897315130064.

We have tried to take a look at 4 clusters for which the data looks like below:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Clus_kmeans
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	3
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	0
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	3
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	3

1.4.2-Cluster 4

However we will go with 3 clusters only as per our Kmeans:

cluster	1	2	3
spending	14.4	11.9	18.5
advance_payments	14.3	13.2	16.2
probability_of_full_payment	0.9	0.8	0.9
current_balance	5.5	5.2	6.2
credit_limit	3.3	2.8	3.7
min_payment_amt	2.7	4.7	3.6
max_spent_in_single_shopping	5.1	5.1	6.0

1.4.3-Cluster 3-Transposed values

```
array([[-1.03025257, -1.00664879, -0.9649051 , -0.89768501, -1.08558344,
         0.69480448, -0.62480856],
       [ 1.25668163,  1.26196622,  0.56046437,  1.23788278,  1.16485187,
        -0.04521936,  1.29230787],
       [-0.14111949, -0.17004259,  0.4496064 , -0.25781445,  0.00164694,
        -0.66191867, -0.58589311]])
```

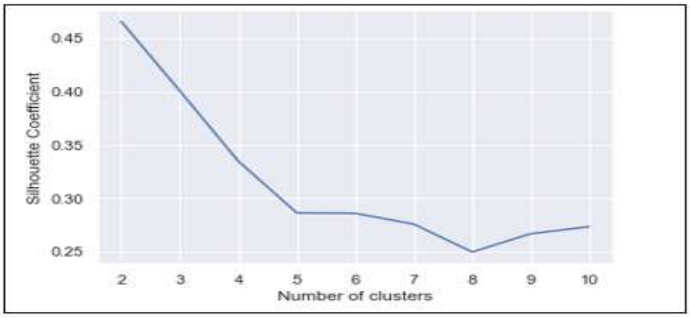
1.4.4-Cluster 3 array data(km_res)

Inference 1.4.3 and 1.4.4:

The array represent how the clusters 3 have been formed with the datas as listed in the data dictionary. We do see some negative values which means we cannot cluster data beyond three to maintain data accuracy.

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of.

The graphical representation of Silhouette Score:



Below is the Silhouette Score

```
[0.46577247686580914,
 0.40072705527512986,
 0.3347542296283262,
 0.28621461554288646,
 0.285726896652541,
 0.2756098749293962,
 0.24943558736282168,
 0.2666366921192433,
 0.2731288488219916]
```

Silhouette data head:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Clus_kmeans	sil_width
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	3	0.445327
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	0	0.049939
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	3	0.443575
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0.532008
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	3	0.081568

1.4.5-Silhouette Data representations

Interpretations:

- 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.
- All the scores are positive and well apart hence we can say all the clusters can be distinguished.

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

We have selected 3 clusters K_Means here.

cluster	1	2	3
spending	14.4	11.9	18.5
advance_payments	14.3	13.2	16.2
probability_of_full_payment	0.9	0.8	0.9
current_balance	5.5	5.2	6.2
credit_limit	3.3	2.8	3.7
min_payment_amt	2.7	4.7	3.6
max_spent_in_single_shopping	5.1	5.1	6.0

1.4.6-3 group via K_Means

clusters-3	1	2	3
spending	18.129200	11.916857	14.217077
advance_payments	16.058000	13.291000	14.195846
probability_of_full_payment	0.881595	0.846766	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	2.768418
max_spent_in_single_shopping	5.987040	5.115071	5.055569
Freq	75.000000	70.000000	65.000000

1.4.7-Heirarchial Clustering

- The clusters here are 3. Hence under each heading we can see the K_Means above.

Cluster Group Profiles

Group 1: High Spending
Group 3: Medium Spending
Group 2: Low Spending

- We can see from the clusters above:
  High spending which is cluster 1 have lowest results for probability_of_full_payment and the highest is the spending.
  Medium Spending which is cluster 3 we have lowest results for probability_of_full_payment and the highest is the spending.
  Low Spending which is cluster 2 we have the highest for advance payments and the lowest for probability_of_full_payment.

The promotional strategies suggested will be:

Group 1: High Spending

The group that spends the highest can be our target to improve more services for hence:

- We will propose vouchers for the groups from brands to attract them to pay on time without fail when they are paying in full.
- Give discounts of 2.5% to 5% when they are spending beyond a particular limit to attract more transactions.
- Increase their credit limit and this will let them come back to us as the limit spending is more.
- Increase spending habits
- Since they repay the loans quickly, we can give them cashbacks on repayment.

Group 2: Low Spending

- Since they are re-paying slowly, we can introduce more benefits and vouchers to attract them to re-pay at the given time.
- Motivate them by giving them updates about new cashbacks and promotional codes.
- The low spenders should be contacted quite often to understand their areas of lag or concern to repay the loan.

Group 3: Medium Spending

- Since this group is neither a quick payer nor a Defaulters, we need to keep them updated about their dues and date of return to make it effective for them to manage their expenses prior.
- They can be retained by giving less interest rate loans and more credit limit as bad debt forecasting ratio will be quite low.
- We can give them promotional or seasonal offers as and when one approaches for a loan to retain them with us.
- Increase spending habits by trying with premium ecommerce sites, travel portal, travel airlines /hotel, as this will encourge them to spend more
- Promote premium cards/loyality cars to increase transcations since the credit score is good.

PROJECT 2:

INSURANCE

https://www.thetravelmagazine.net/covid-19-crisis-will-travel-insurance-cover-me-for-covid-19-in-the-future.html

Introduction

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.

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)

Data Information

The data set has 3000 rows and 10 columns. The 10 columns are of the heading as described above in the attribute information. The data looks like below:

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

2.1.1- Data headings

The information in the data is of the below type:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Age           3000 non-null   int64
 1   Agency_Code   3000 non-null   object
 2   Type          3000 non-null   object
 3   Claimed       3000 non-null   object
 4   Commision     3000 non-null   float64
 5   Channel       3000 non-null   object
 6   Duration      3000 non-null   int64
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   object
 9   Destination   3000 non-null   object
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

2.1.2- Data Types

Inference from data type is:

- The data which is of int64 is having all numerical values, whole numbers
- The data types with object category have alphabetical data
- The float data type has decimal and whole numbers.

The data has no missing values which needs to be looked into.

	Age	Commision	Duration	Sales
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	38.091000	14.529203	70.001333	60.249913
std	10.463518	25.481455	134.053313	70.733954
min	8.000000	0.000000	-1.000000	0.000000
25%	32.000000	0.000000	11.000000	20.000000
50%	36.000000	4.630000	26.500000	33.000000
75%	42.000000	17.235000	63.000000	69.000000
max	84.000000	210.210000	4580.000000	539.000000

2.1.3- Descriptive Data

INFERENCE:

- The data has the mean median and quartiles for all numerical columns in dataset given to us.
- As we can see the mean of commission is the least at 14.529 and the Duration is the maximum with 70.0013.
- The data with the least STD (standard deviation) is Age and the maximum is duration.
- The Q1 of Commission is 0 whereas the highest is for Age.
- The Q3 is the maximum for Sales and the least for Commission.
- The negative value of Duration indicates something wrong with the data which needs analysis.
- The difference in the Commission and Sales is too high which needs to be analysed as the results for Mean, STD and Quartiles indicate.

	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

2.1.4- Described Data

INFERENCE:

- Categorial code variable maximum unique count is 5.

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

```
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
```

2.1.4- Nominal Data

Inference from NOMINAL data type is:

- The data column with their subcategories have been shown on the left.
- All categories with their counts and data type have been represented.

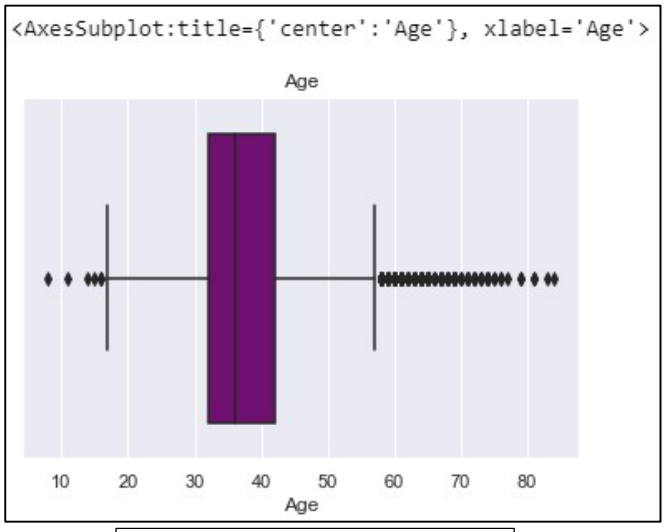
After performing the duplicate test, we found 139 rows* 10 columns as a duplicate. However, since it is a travel information with no unique ID's we shall not drop any column or rows thinking its unique for different customers.

Univariate Analysis

It takes data, summarizes that data and finds patterns in the data. Lets see below:

1. AGE DATA:

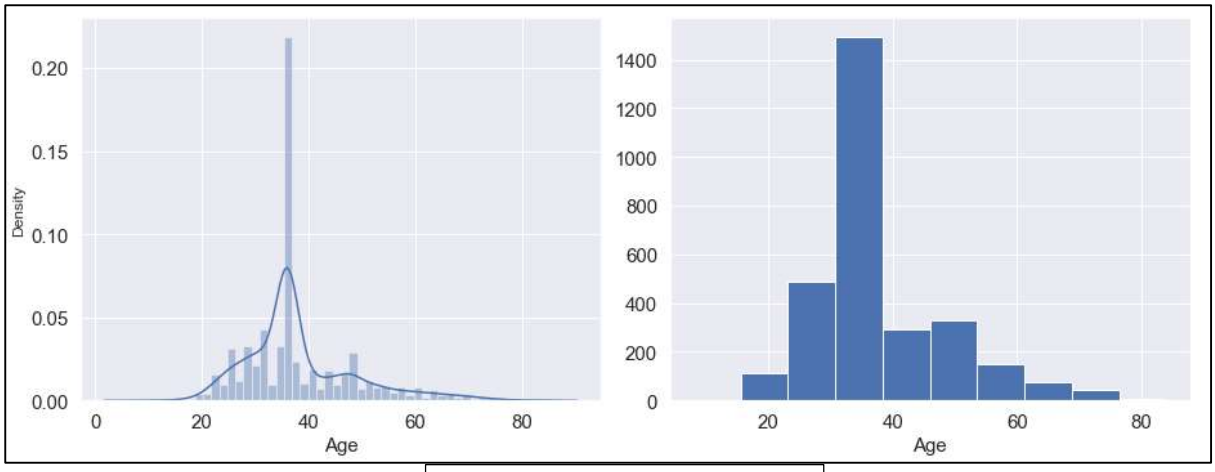
Minimum Age: 8
Maximum Age: 84
Mean value: 38.091
Median value: 36.0
Standard deviation: 10.463518245377944
Null values: False
spending - 1st Quartile (Q1) is: 32.0
spending - 3rd Quartile (Q3) is: 42.0
Interquartile range (IQR) of Age is 10.0



2.1.5- OUTLIER for AGE

Inference:

- Here we see a lot of outliers starting from 57years of age.
- As per our results of outliers:
  -Lower outliers in Age: 17.0
  -Upper outliers in Age: 57.0



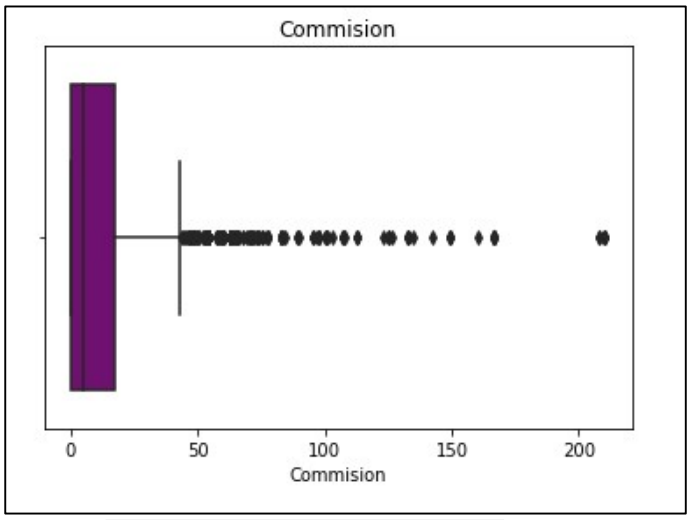
2.1.6- Dist. plot and Histogram(AGE)

Inference:

- The first Dist. Plot shows us huge counts in the age slab 35-40 which is also shown by histogram the second figure.
- As per the graphs the lowest is under the slab of 70-80 years of age.

2. COMMISSION:

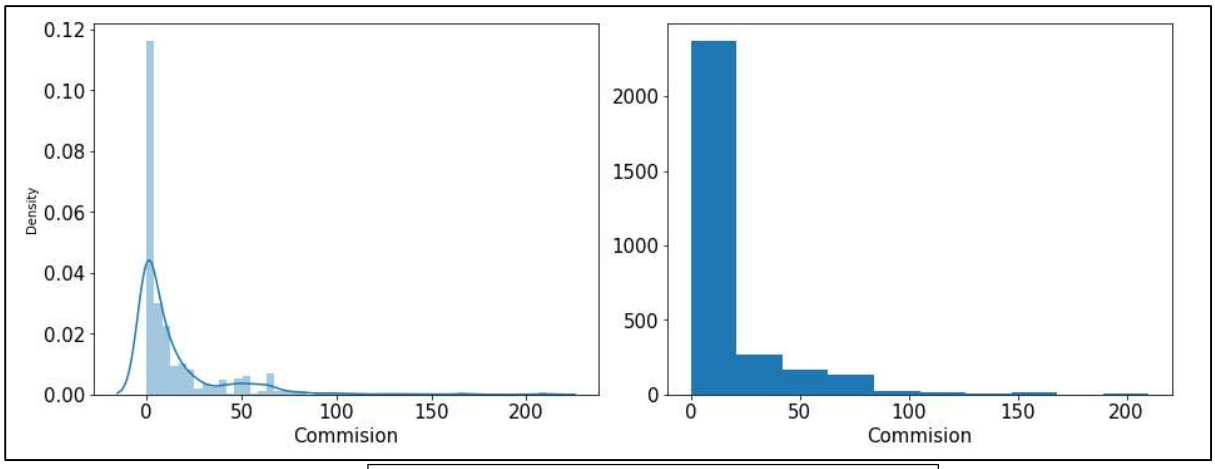
Minimum Commission: 0.0
Maximum commission: 210.21
Mean value: 14.529203333333266
Median value: 4.63
Standard deviation: 25.48145450662553
Null values: False
Commission - 1st Quartile (Q1) is: 0.0
Commission - 3rd Quartile (Q3) is: 17.235
Interquartile range (IQR) of commission is 17.235



2.1.7- Box Plot for Commission

INFERENCE:

- The box plot has outliers starting from -25.8525 and ending at 43.0875.
- As the graph above represents, we have a lot of outliers in the Commission.
- We haven't treated any outliers in the diagram.



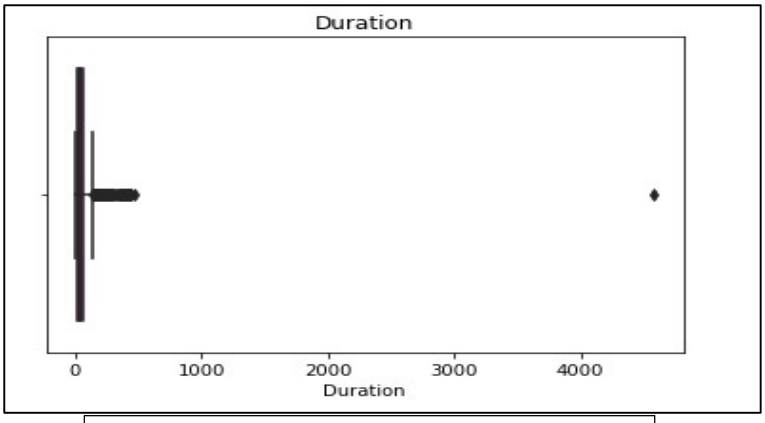
2.1.8- Dist. plot and Histogram (COMMISSION)

INFERENCE:

- As we see a lot of data lies on the left side of the diagram both in the dist. Plot and in the Histogram.
- The data range of values is 210.21.
- 0-50 range has a lot of values.

3. DURATION:

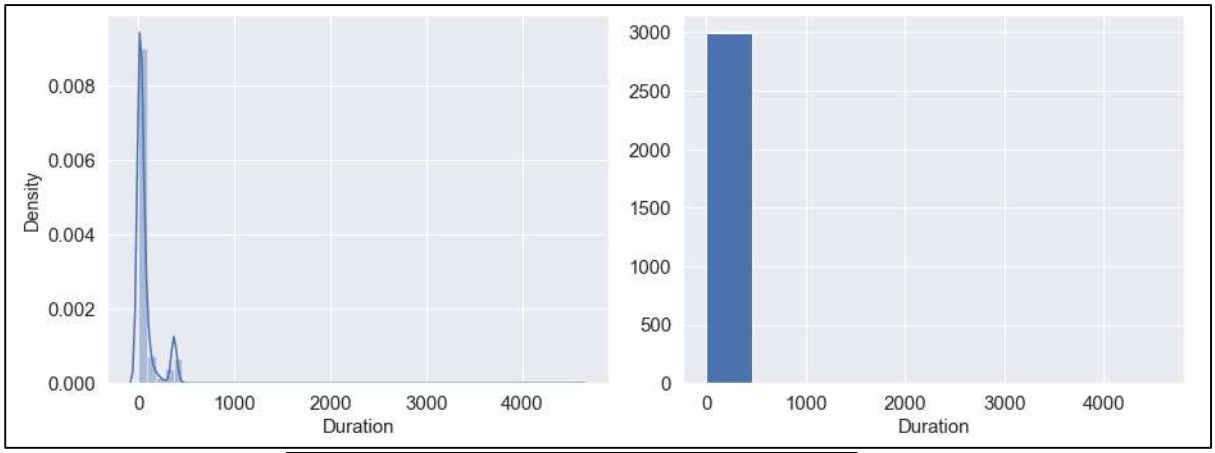
Minimum Duration: -1
Maximum Duration: 4580
Mean value: 70.00133333333333
Median value: 26.5
Standard deviation: 134.05331313253495
Null values: False
Duration - 1st Quartile (Q1) is: 11.0
Duration - 3rd Quartile (Q3) is: 63.0
Interquartile range (IQR) of Duration is 52.0



2.1.9- BOX Plot (Duration)

INFERENCE:

- As we can see above the outliers are too high in the DURATION data.
- Outliers range from:
  -Lower outliers in Duration: -67.0
  -Upper outliers in Duration: 141.0
- The range of values in the data is 539.0



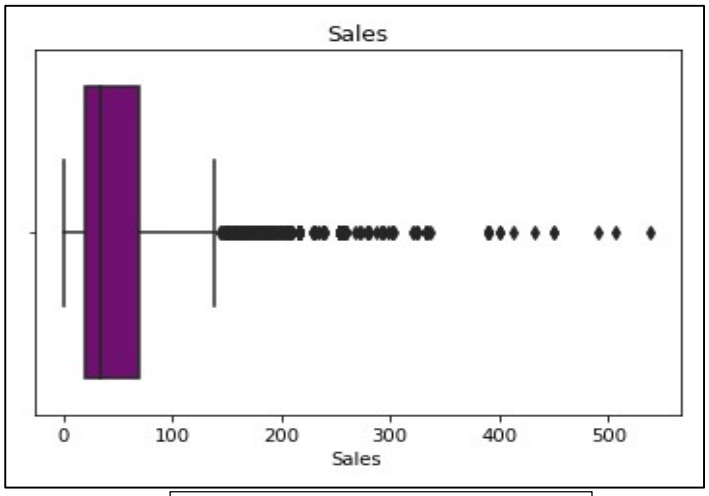
2.1.10- Dist. plot and Histogram (DURATION)

INFERENCE:

- The data has a lot of skewness in the left side of the range from 0-500.
- The maximum data as per the histogram is in the range 0-500.

4. Sales:

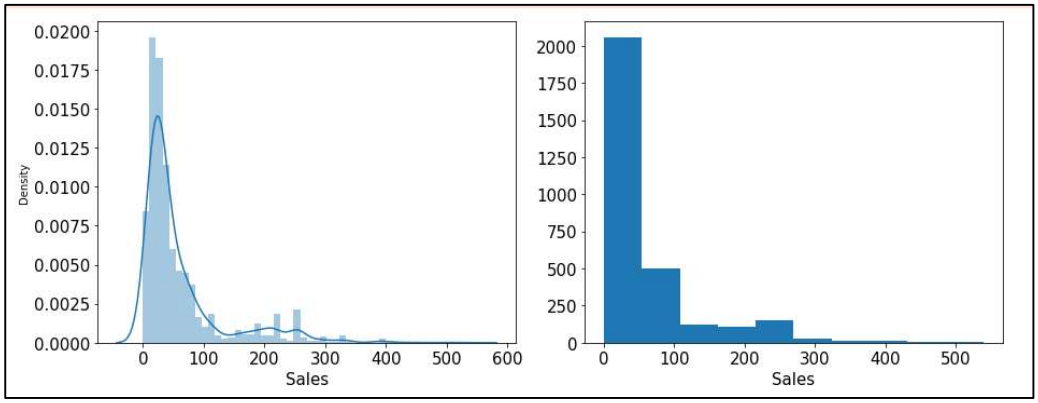
Minimum Sales: 0.0
Maximum Sales: 539.0
Mean value: 60.24991333333344
Median value: 33.0
Standard deviation: 70.73395353143047
Null values: False
Sales - 1st Quartile (Q1) is: 20.0
Sales - 3rd Quartile (Q3) is: 69.0
Interquartile range (IQR) of Sales is 49.0



2.1.11- Box Plot(Sales)

INFERENCE:

- Lower outliers in Sales: -53.5
- Upper outliers in Sales: 142.5
- We have a wide range in outliers spread densely in 100-200 sales range and mildly spread at the highest sales values near about 500.



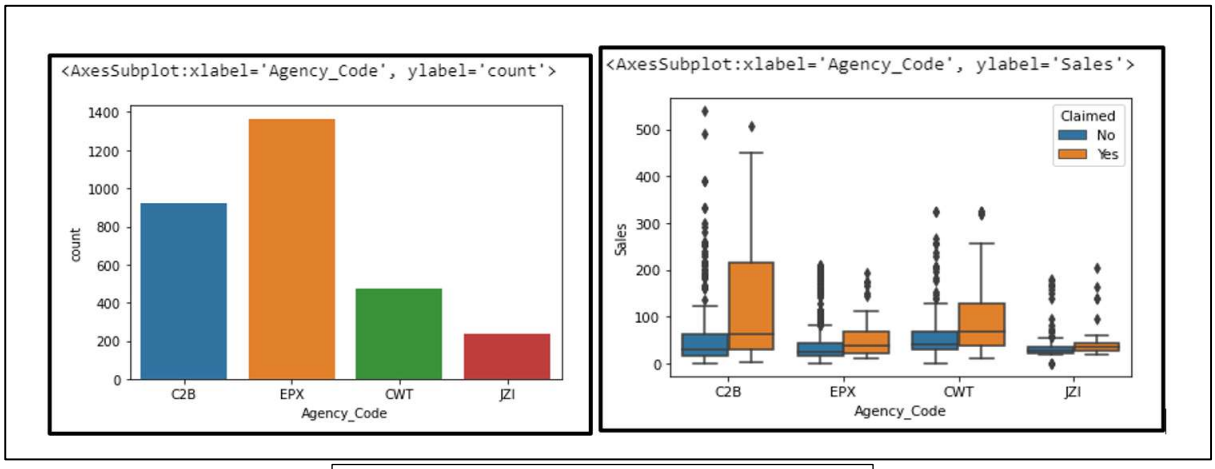
2.1.12- - Dist. plot and Histogram (Sales)

INFERENCE:

- We see the data having a lot of values in 0-100 which becomes quite consistent after 200 as per the dist. Plot.
- The data of Sales ranges between the values of 0-100.
- The range of values in Sales is 539.

CATEGORICAL VARIABLES:

5. Agency_Code:

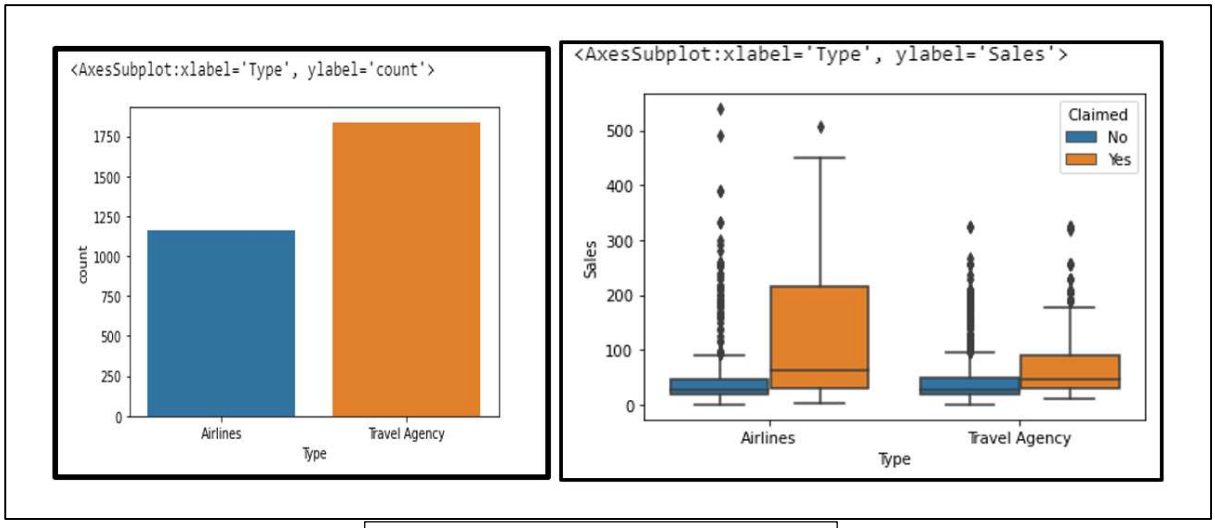


2.1.13- - Count PLOT and Box Plot (Agency Code)

INFERENCE:

- The data under Agency has 4 subcategories namely:
  C2B with 924 values
  CWT with 472 values
  EPX with 1365 values
  JZI with 239 values
- As we see from the count plot maximum is EPX and the least is JZI.
- The box plot on the right side shows us a lot of outliers for all the data however as per the image we can say that C2b has a lot of outliers for NO claims and the least is for the claimed values yes in CWT.

6. Type:



2.1.14- - Count PLOT and Box Plot (TYPE)

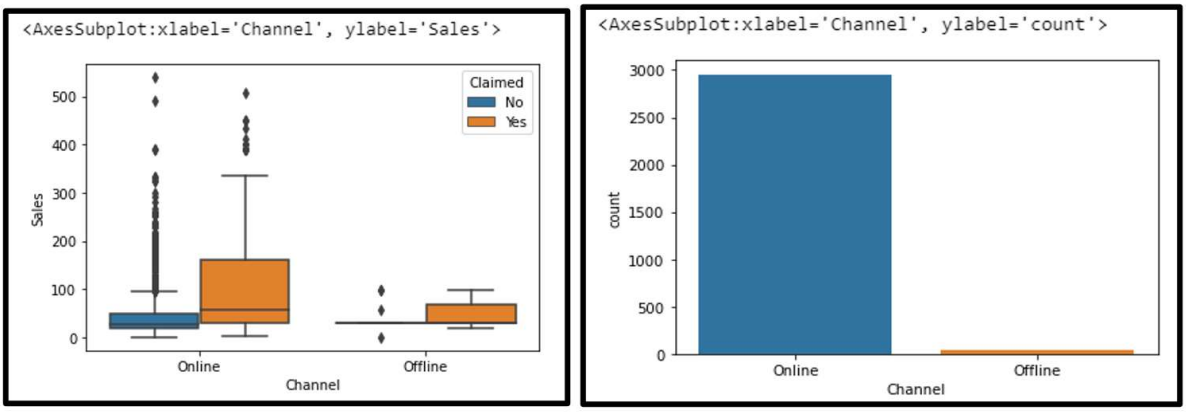
INFERENCE:

- The subheadings under Travel Agency are:

Sub-heading	Count of values
Airlines	1163
Travel Agency	1837
Grand Total	**3000**

- The travel agency has more data and values than airlines.
- The Outliers is the maximum in Airlines with category NO and the outliers with the Yes in Airlines sales is the least.

7. CHANNEL:



2.1.15 - Box Plot and Count Plot and (Channel)

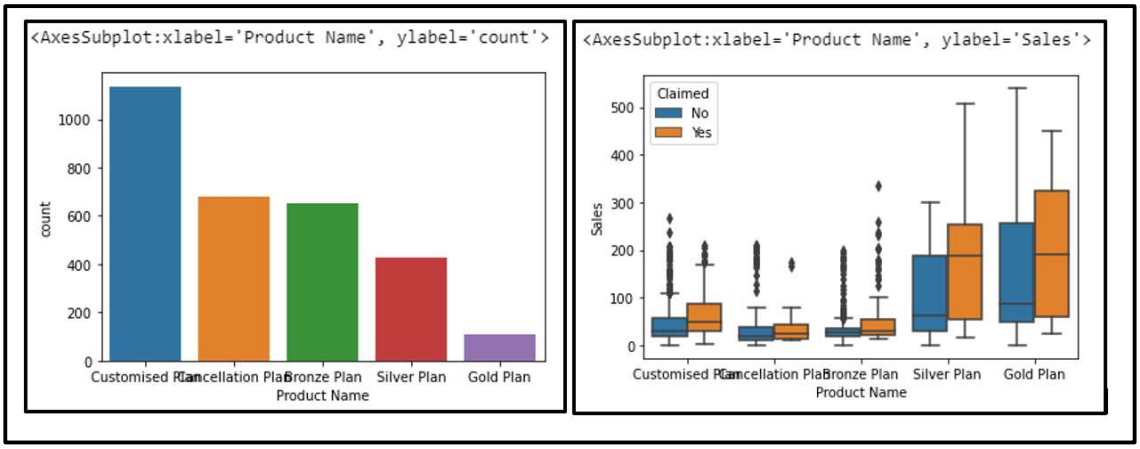
INFERENCE:

- The sales is the highest through the channel Online mode than Offline.

CHANNEL	Count
Offline	46
Online	2954
Grand Total	**3000**

- The data of online mode has a lot of outliers.

8. PRODUCT NAME:



2.1.16 - Count Plot and Box Plot (Product Name)

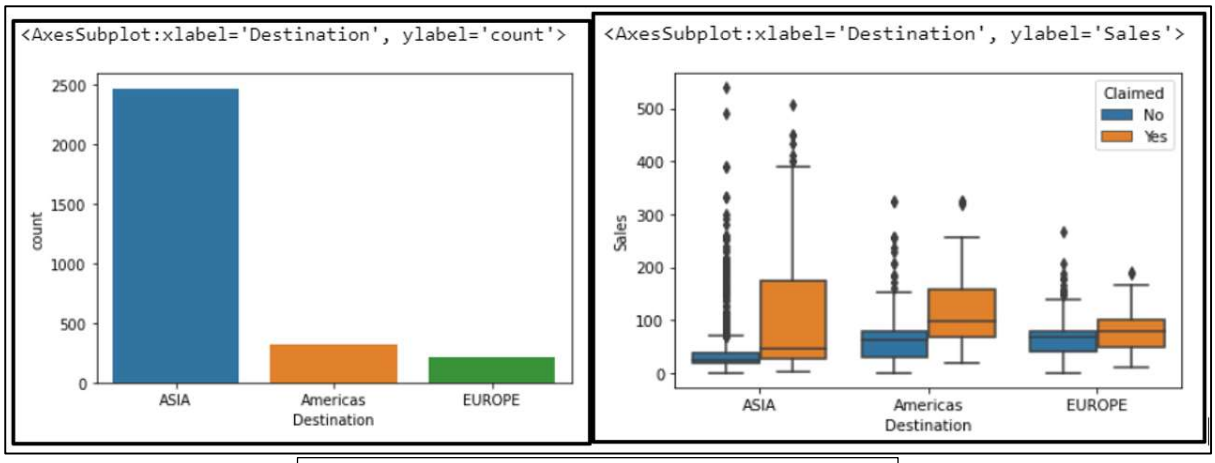
INFERENCE:

- The sub-headings as per the data is:

Product Name	Count
Bronze Plan	650
Cancellation Plan	678
Customised Plan	1136
Gold Plan	109
Silver Plan	427
Grand Total	**3000**

- The maximum data lies in Customised Plan and the minimum in Gold Plan.
- As we can see in the box plot, we have outliers in all the sub-heading schemes except for Silver and Gold Plan.

9. Destination



2.1.16 - Count Plot and Box Plot (Destination)

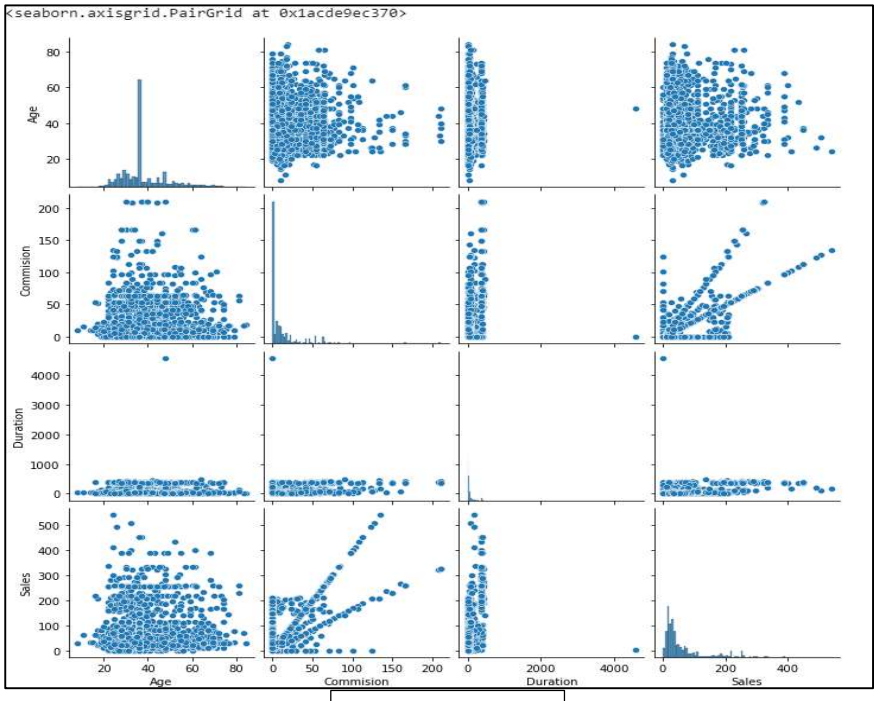
INFERENCE:

- The subheading with the count is as below:

Sub-headings	Count of Values
Americas	320
ASIA	2465
EUROPE	215
Grand Total	**3000**

- The maximum values in the Subcategory Asia and the least is in Europe.
- The outliers are maximum for Sales in the destination for Asia with the claims made NO.

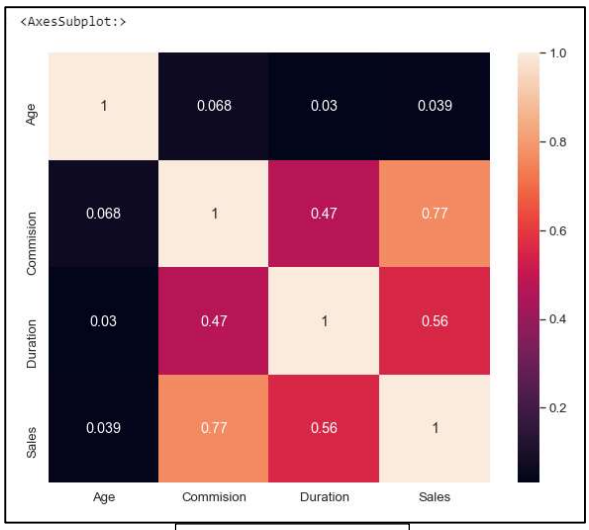
We will analyse the data in detail now through MULTIVARIATE ANALYSIS:



2.1.17 – Pair Plot

INFERENCE:

- The data shows that there is some relationship between the data.



2.1.18– Heatmap

INFERENCE:

- The heatmap has positive and negative values.
- The positive values represent a stronger corelation to that of the negative values.
- The relation between age and commision is the highest at 0.68 and the lowest is sales and age which is 0.039.
- Overall we see a positive co-relation between all values.

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

We have to split our data into Independent and Dependant variable with all Integer type values to perform CART model. Post converting the categorical (object data type) to numerical or integer data type we have .

```
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]
```

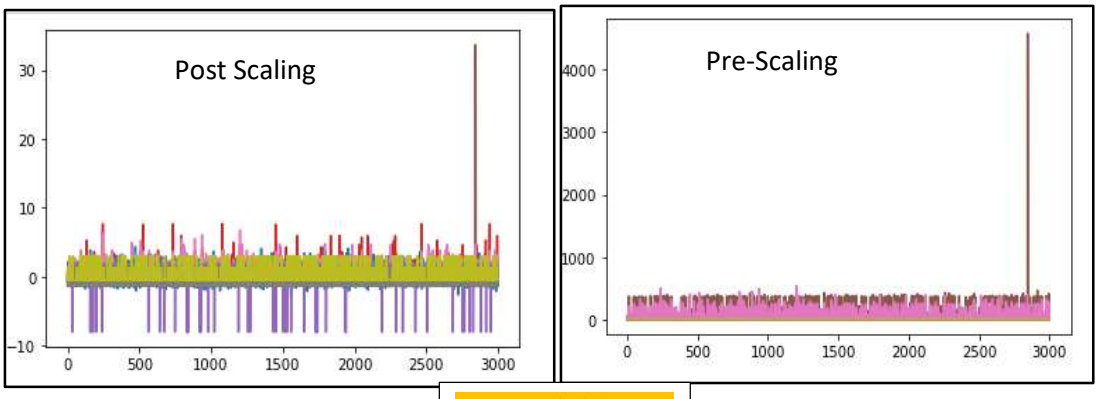
2.2.1– Data assigned values Categorical to Integer.

Post converting Data we have:

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

2.2.2– Data headings

Lets see graphically how different the variations in data are post and pre-scaling:



2.2.3– Scaled data

Inference from the above can be drawn that Pre-scaling data was not complete and well distributed with a lot of variations whereas in post we can make some sort of analysis which will be more clear from Z_SCORE results below:

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
0	0.947162	-1.314358	-1.256796	-0.542807	0.124788	-0.470051	-0.816433	0.268835	-0.434646
1	-0.199870	0.697928	0.795674	-0.570282	0.124788	-0.268605	-0.569127	0.268835	-0.434646
2	0.086888	-0.308215	0.795674	-0.337133	0.124788	-0.499894	-0.711940	0.268835	1.303937
3	-0.199870	0.697928	0.795674	-0.570282	0.124788	-0.492433	-0.484288	-0.525751	-0.434646
4	-0.486629	1.704071	-1.256796	-0.323003	0.124788	-0.126846	-0.597407	-1.320338	-0.434646

2.2.4– Z-Score Results

Inference:

- The values are in negative and positive mixed which can be looked at keeping in mind the below:
  - Positive z-score: The individual value is greater than the mean.
  - Negative z-score: The individual value is less than the mean.
- We can see above no values are lesser than -2 (low) and more than 2(high) which represents no maximum high and low values.

Training and Testing:

We have the below dimensions of rows and columns for training and testing data:

```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
```

2.2.5–Dimensions

Post performing a decision tree we see that the data in there is too high and its difficult to conclude or analyse since it is an overfitted decision tree. To have data in the same scale and uniformity we will be doing another train-test and then decision tree.

TREE BUILDING RESULTS of 1st Attempt

The below is the important significant numbers from each heading with GINI feature:

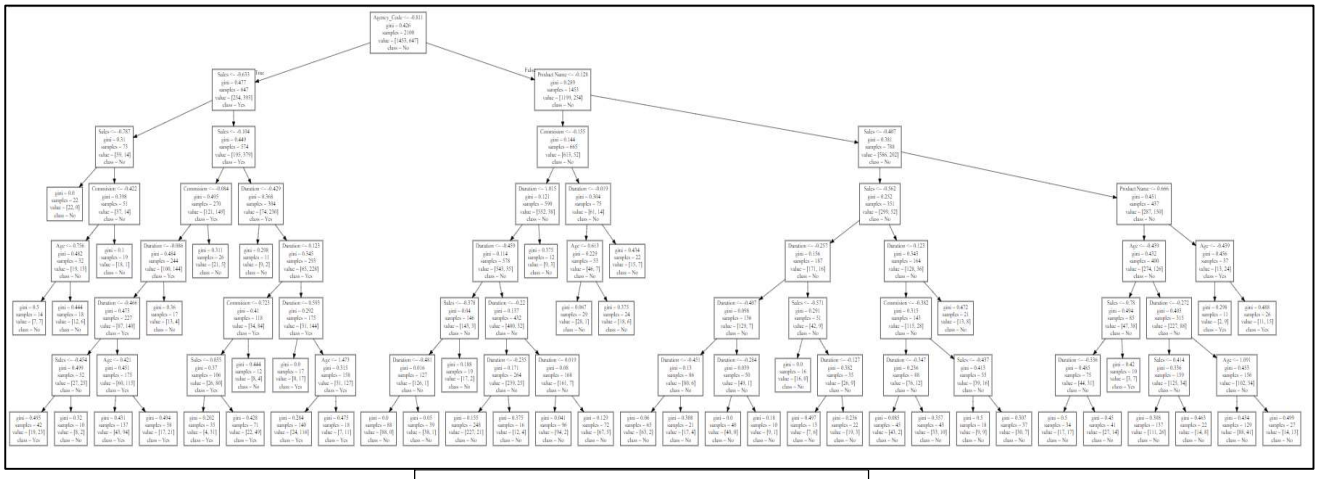
	Imp
Age	0.161897
Agency_Code	0.191766
Type	0.001507
Commision	0.100035
Channel	0.007032
Duration	0.258847
Sales	0.207893
Product Name	0.048289
Destination	0.022734

2.2.6–Gini importance Feature

INFERENCE:

- The above tells us which is the most significant variable from descending to ascending assigning Gini index to all.
- Age is the most important Variable.
- Business needs to focus on it according to the same values or order and can ignore the last value variable which is Destination.

The Decision model looks like in 2nd attempt:



2.2.6–Decision Tree with GINI Index

Inference:

- The decision tree has 2100 samples with values in the range of [1453,647] since the maximum is 1453 the parent node has NO as the class.
- When Agency code is less than equal to -0.81 for true it calculates gini index and nodes for Sales and false product name.
- The important values are the ones which is having a value of 0.1 and the ones with 0.01 can be ignored.
- The tree goes on until a uniform and accurate result is reached.
- The predictions have been as below for the classes and probabilities:

	0	1
0	0.697947	0.302053
1	0.979452	0.020548
2	0.921171	0.078829
3	0.510417	0.489583
4	0.921171	0.078829

Random Forest Classifier- Since a random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. We will be using the below:

```
RandomForestClassifier(max_depth=6, max_features=3, min_samples_leaf=8,
                       min_samples_split=46, n_estimators=350, random_state=1)
```

- Max_depth is the maximum depth of the tree which has been taken as 6. The max depth is in the data until all nodes are broken till pure or is less than min_samples_split.

- N_estimator is the number of trees in the forest with random state -1 which is only to maintain the purity of data.
- Min_Sample_leaf-The minimum number of samples required to be at a leaf node.
- max_features-The search for a split does not stop until at least one valid partition of the node samples is found, even if it requires to effectively inspect more than max_features features.
- Min_samples_split=The minimum number of samples required to split an internal node.The important features of the training and the testing model is:

	0	1
0	0.778010	0.221990
1	0.971910	0.028090
2	0.904401	0.095599
3	0.651398	0.348602
4	0.868406	0.131594

Artificial Neural Network-

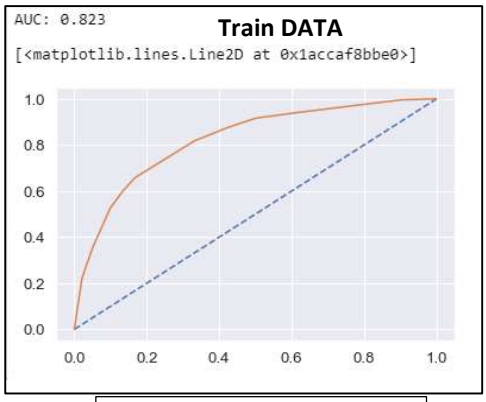
As predictions calculated by ANN is machine learning, the complete set of data is divided into 3 groups. Training set is used for teaching the model. Validation is used to determine when to stop learning (protect from overfitting). Test dataset is used to calculate the errors of the model built but on the data that was not seen by the machine before. Linear correlation between predicted values and the real values can be calculated for each data set separately or fore whole set.

	0	1
0	0.822676	0.177324
1	0.933407	0.066593
2	0.918772	0.081228
3	0.688933	0.311067
4	0.913425	0.086575

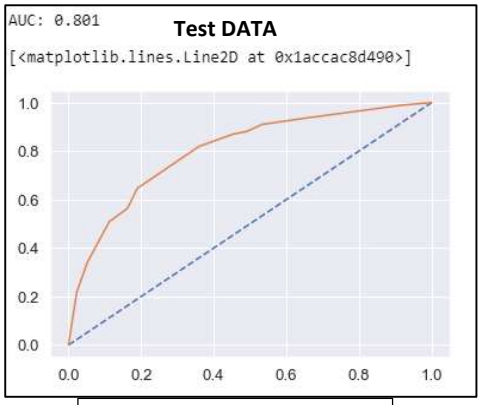
Inference—We have the probability predictions of the data as 0 and 1.

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model.

First we will go by **Classification and Regression Tree (CART) Model** and predict value of one variable to others:



2.3.1–AUC for Train data



2.3.2–AUC for Test data

The results for the CART Model are:

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1453
1	0.70	0.53	0.60	647
accuracy			0.79	2100
macro avg	0.76	0.71	0.73	2100
weighted avg	0.78	0.79	0.78	2100

	precision	recall	f1-score	support
0	0.80	0.89	0.84	623
1	0.67	0.51	0.58	277
accuracy			0.77	900
macro avg	0.74	0.70	0.71	900
weighted avg	0.76	0.77	0.76	900

2.3.3–Train data and Test Data scores

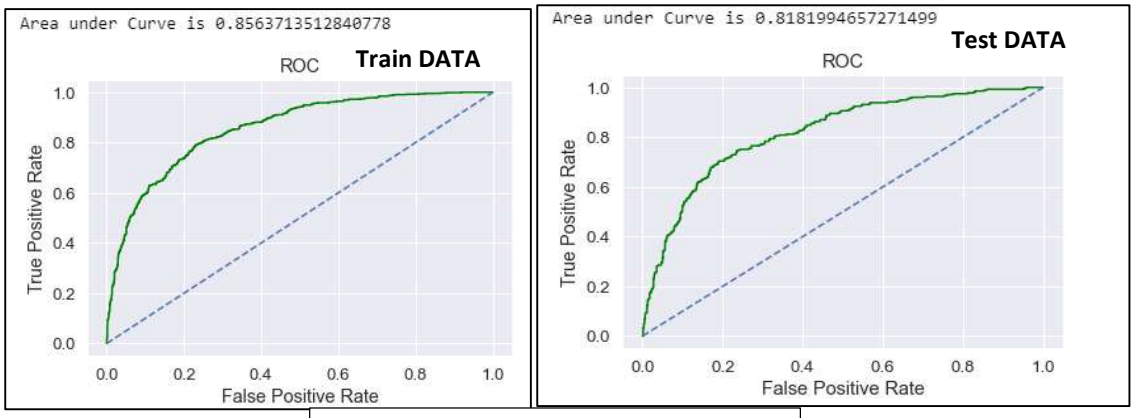
Test Data:
- AUC: 86%
- Accuracy: 80%
- Precision: 72%
- f1-Score: 66%

Train Data:
- AUC: 82%
- Accuracy: 78%
- Precision: 68%
- f1-Score: 62

INFERENCE: Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

- cart_test_precision 0.67
- cart_test_recall 0.51
- cart_test_f1 0.58

Random Forest Conclusion:



2.3.4–Train data and Test Data AUC curve

Train DATA	precision	recall	f1-score	support
0	0.84	0.89	0.86	1453
1	0.72	0.61	0.66	647
accuracy			0.80	2100
macro avg	0.78	0.75	0.76	2100
weighted avg	0.80	0.80	0.80	2100

Test DATA	precision	recall	f1-score	support
0	0.82	0.88	0.85	623
1	0.68	0.56	0.62	277
accuracy			0.78	900
macro avg	0.75	0.72	0.73	900
weighted avg	0.78	0.78	0.78	900

2.3.4–Train data and Test Data scores RFC model

Train Data:

- AUC: 86%
- Accuracy: 80%
- Precision: 72%
- f1-Score: 66%

Train Data:

- AUC: 86%
- Accuracy: 80%
- Precision: 72%
- f1-Score: 66%

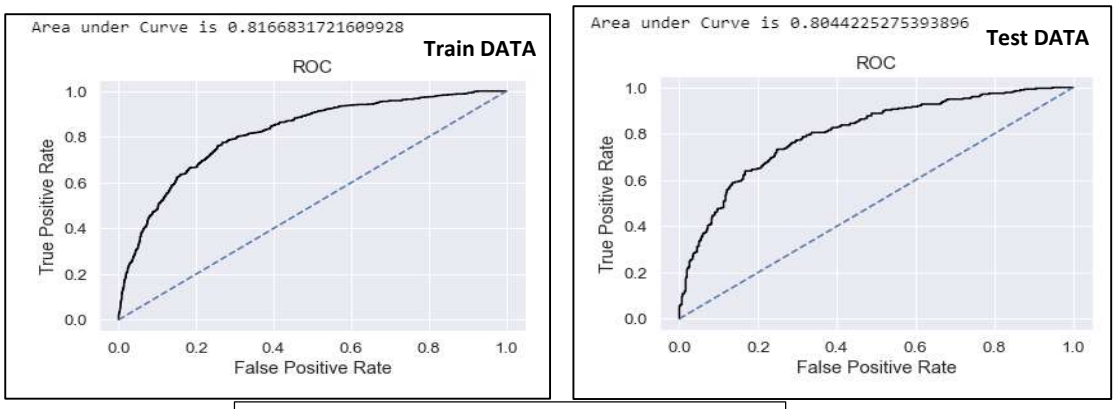
Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

NN Model Performance:

Train DATA	precision	recall	f1-score	support
0	0.80	0.89	0.85	1453
1	0.68	0.51	0.59	647
accuracy			0.78	2100
macro avg	0.74	0.70	0.72	2100
weighted avg	0.77	0.78	0.77	2100

Test DATA	precision	recall	f1-score	support
0	0.80	0.89	0.84	623
1	0.67	0.50	0.57	277
accuracy			0.77	900
macro avg	0.73	0.69	0.71	900
weighted avg	0.76	0.77	0.76	900

2.3.5–Train data and Test Data scores NN model



2.3.6–Train data and Test Data AUC model

Train Data:
- AUC: 82%
- Accuracy: 78%
- Precision: 68%
- f1-Score: 59

Test Data:

- AUC: 80%
- Accuracy: 77%
- Precision: 67%
- f1-Score: 57%

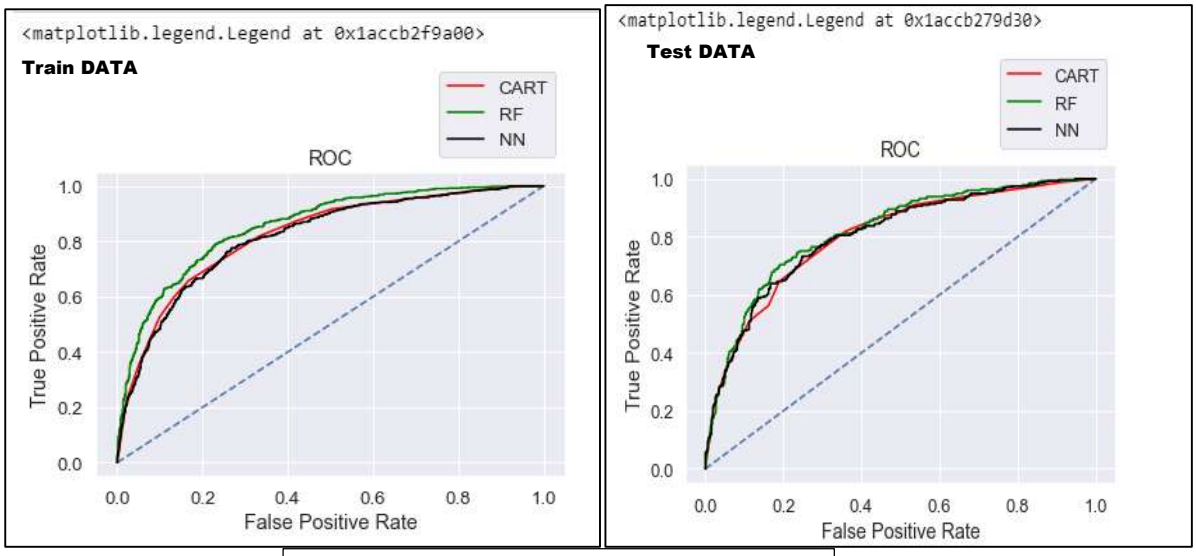
Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

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

The data of the results from all the models have been taken in a data frame and is as below:

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.79	0.77	0.80	0.78	0.78	0.77
AUC	0.82	0.80	0.86	0.82	0.82	0.80
Recall	0.53	0.51	0.61	0.56	0.51	0.50
Precision	0.70	0.67	0.72	0.68	0.68	0.67
F1 Score	0.60	0.58	0.66	0.62	0.59	0.57

2.3.7–Train and Test Outputs of all the models



2.3.8–Train and Test Outputs in AUC curve

INFERENCE:

- As we see from the table of data frame the results for RF curve are the best and we will be selecting the same to proceed with the exploring of this data.
- As we see the AUC curve the RF model is represented in green which shows the best results and accuracy is the best.

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

The inferences are:

- Offline sales are very low at 1796, more online schemes and vouchers shall be given to attract customers as 90% sales is through the same and is the core reason contributing to profits.
- We need to have better Products for Gold Plan as the sales is the least there.
- Agency_Code JZI needs to have better planning and sales as it is the least. More promotional campaigns and customer attraction competitions/Vouchers can be done here.
- 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.
- More analysis should be done on the below data wherein Claims haven't been made through agency is more whereas ticket booking made by agency is more.

Sum of Sales	Claims Made		
Agency_code	No	Yes	Grand Total
C2B	22603.89	65161.91	87765.8

CWT	18300.2	13245.85	31546.05
EPX	42905.28	9881.74	52787.02
JZI	7096.87	1554	8650.87
Grand Total	**90906.24**	**89843.5**	**180749.74**

- New objective should be set wherein travel insurance risk minimization and customer satisfaction should be the goal.
- Key performance indicators (KPI) The KPI's of insurance claims are:
- Reduce claims cycle time
- Increase customer satisfaction
- Combat fraud
- Optimize claims recovery
- Reduce claim handling costs Insights gained from data and AI-powered analytics could expand the boundaries of insurability, extend existing products, and give rise to new risk transfer solutions in areas like a non-damage business interruption and reputational damage.