DATA MINIING PROJECT

BUSINESS REPORT

Date-02/10/2022

Contents

Problem 1

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. Please note that it is a summarized data that contains the average values in all the columns considering all the months, and not for any particular month. You are given the task to identify the segments based on credit card usage.

Data Description:

1. spending: Average spending on card.
2. advance_payments: Amount paid by the customer.
3. probability_of_full_payment: Probability of the credit card payment done in full.
4. current_balance: balance amount left in the credit card.
5. credit_limit: Limit of the amount in credit card.
6. min_payment_amt: average minimum amount paid by the customer on monthly card Bill.
7. max_spent_in_single_shopping: Maximum amount spent by the customer for a single Transaction.

Sample of the dataset:

	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

Exploratory Data Analysis:

Let us check the types of variables in the data frame.

Column	Dtype
Spending	float64
advance_payments	float64
probability_of_full_payment	float64
current_balance	float64
credit_limit	float64
min_payment_amt	float64
max_spent_in_single_shopping	float64

There are total 210 rows and 7 columns in the dataset and all columns are float data type.

Checking for missing values in the dataset:

```
 Column                        Non-Null Count
 ---------                     --------------
Spending                       210 non-null
advance_payments               210 non-null
current_balance                210 non-null
credit_limit                   210 non-null
min_payment_amt                210 non-null
max_spent_in_single_shopping   210 non-null
probability_of_full_payment    210 non-null
```

From the above results we can see that there is no missing value present in the dataset.

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

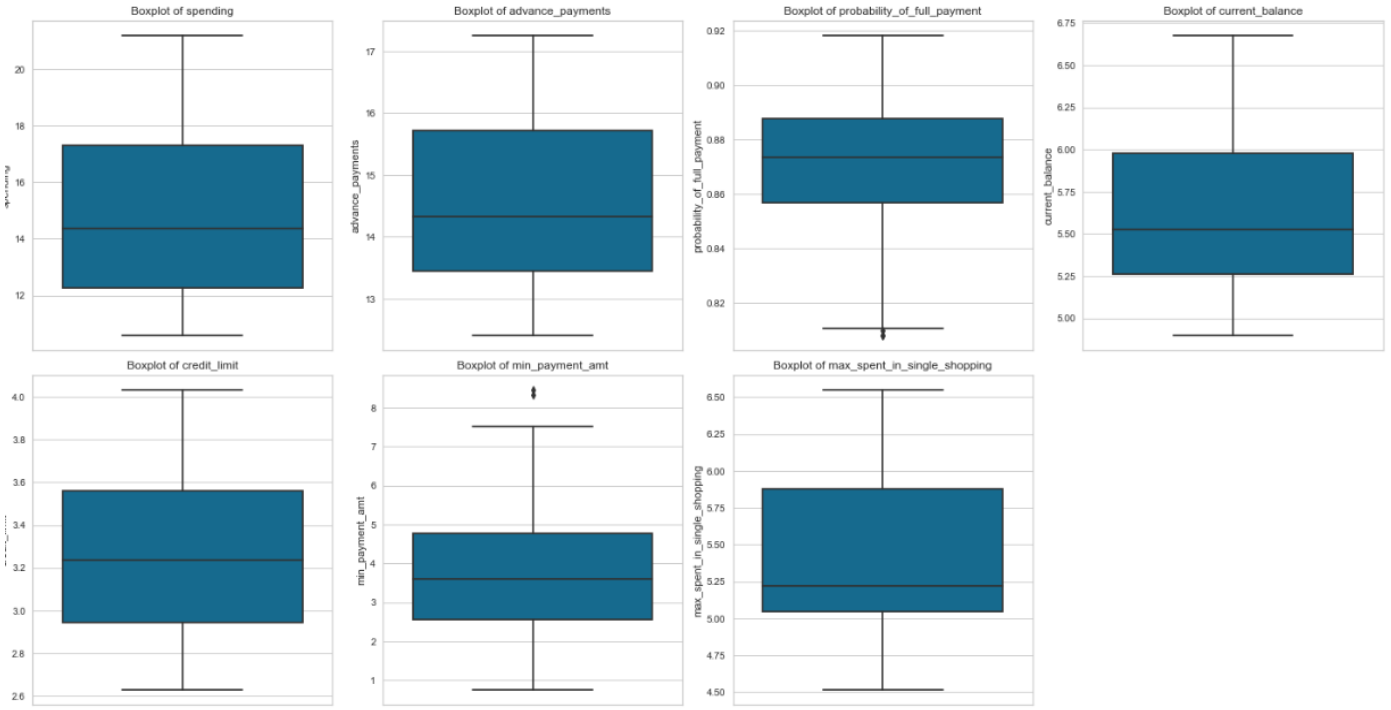
We have checked that there is no missing value present in the data and shape of the data is 210 rows with 7 columns & all float data types.

We are using describe function to get descriptive summary of the data and here is the sample.

	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

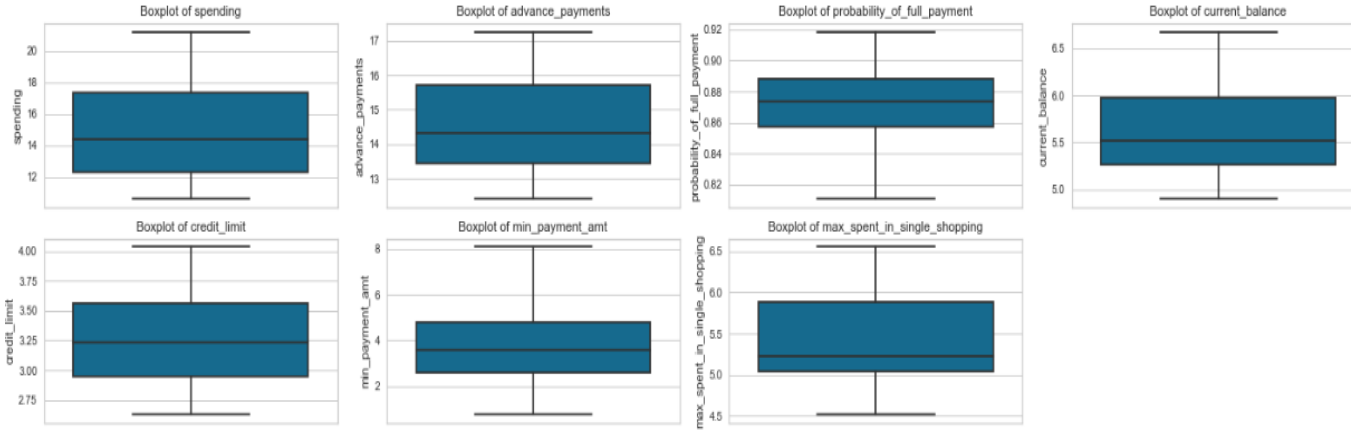
We have also checked for duplicate and no duplicate found.

Checking dataset for outliers:



We can see that outliers present in the data and it has to be treated as clustering results are affected by the presence of outliers.

Boxplots after treating the outliers:



Data Visualization:

We don't have any categorical variables present in the data.

Univariate Analysis:

Non visual representation:

In our data we have no object data type so using describe function.

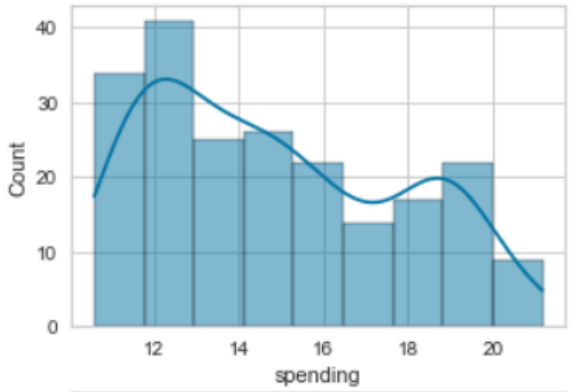
	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.590000	12.27000	14.35500	17.305000	21.180000
advance_payments	210.0	14.559286	1.305959	12.410000	13.45000	14.32000	15.715000	17.250000
probability_of_full_payment	210.0	0.871025	0.023560	0.810588	0.85690	0.87345	0.887775	0.918300
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

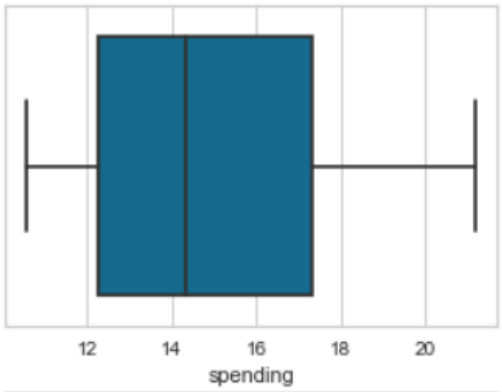
Insights:

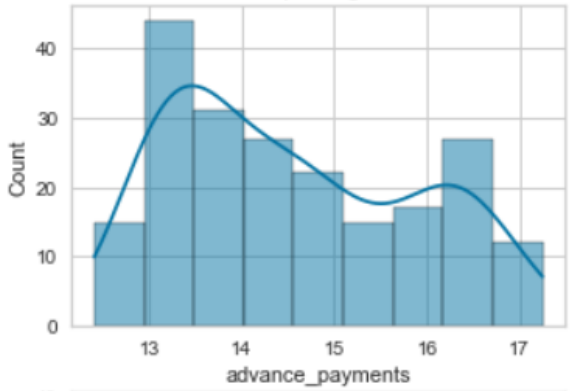
1. Average spending made by the customer is 14.84*1000 = 14,840Rs and highest spending is 21,180Rs.
2. minimum spending in single shopping is 4510Rs and maximum is 6550Rs so we can say that in single shopping customer spending good amount compare to maximum spending as different is not too much.
3. Maximum sanctioned credit limit given to customer by bank is Rs 40,300.

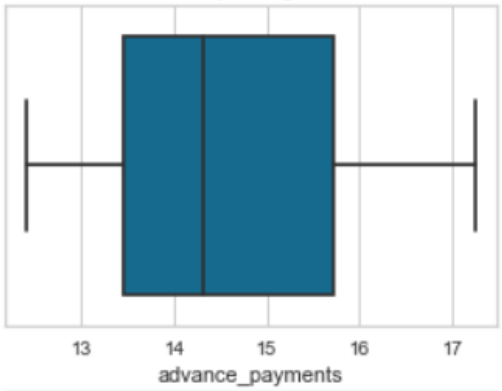
Visual representation:

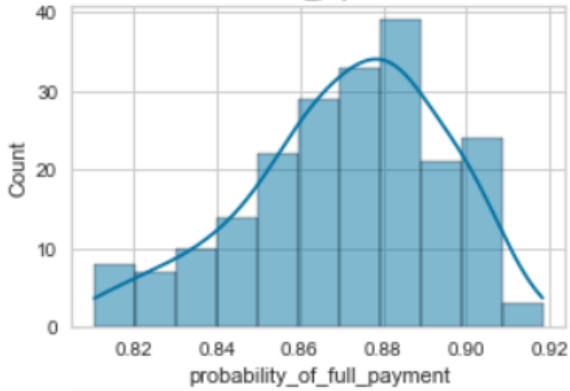
We are using Boxplots and histogram to visualize the distribution of the data.

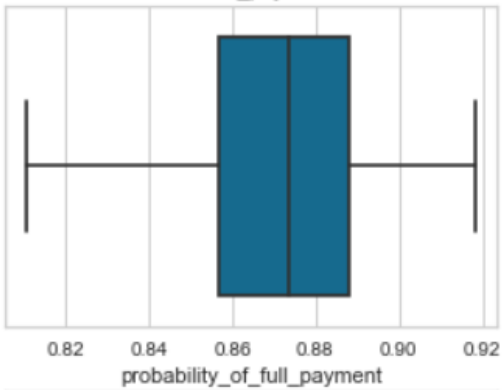

Count
spending

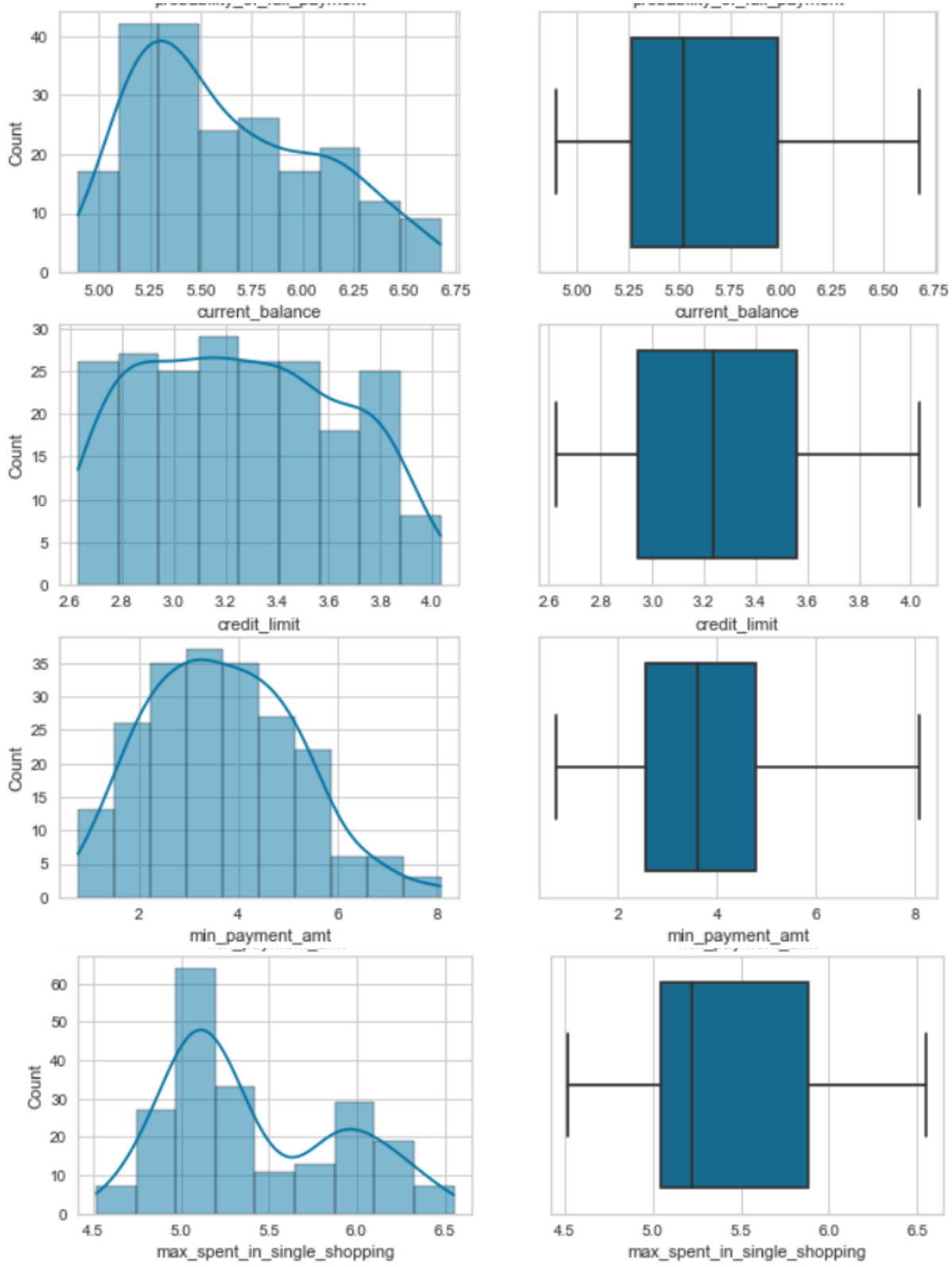

spending


Count
advance_payments


advance_payments


Count
probability_of_full_payment


probability_of_full_payment



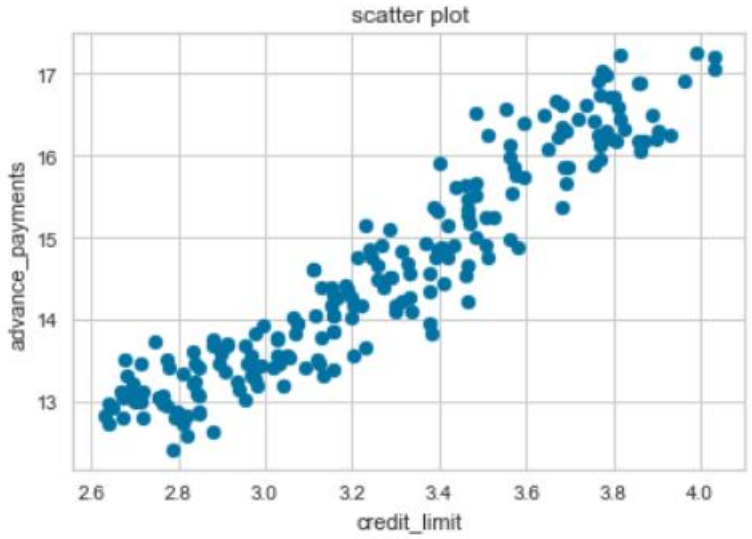
Insights:

1. From the above box plots we can say that there is no outliers.

2. From above figure, we can say that for variables 'probability_of_full_payment' distribution is slightly left tailed.

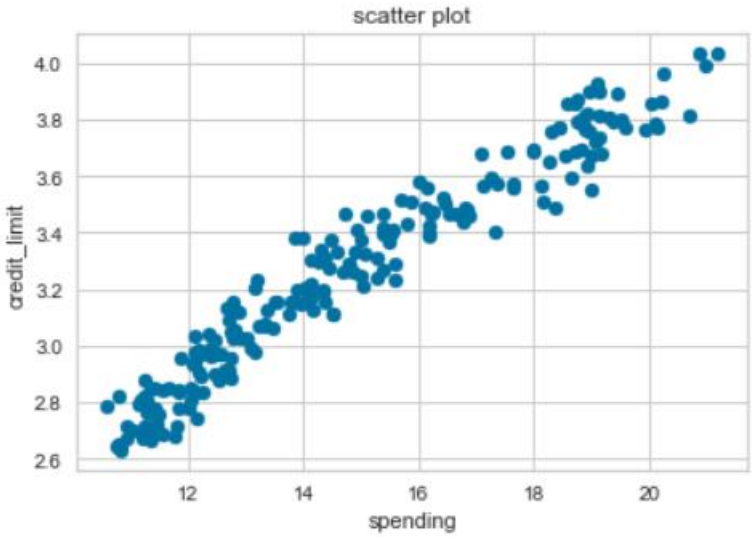
Bivariate Analysis:

We are using scatter plot to draw relation between two variables by using various combination of variables.



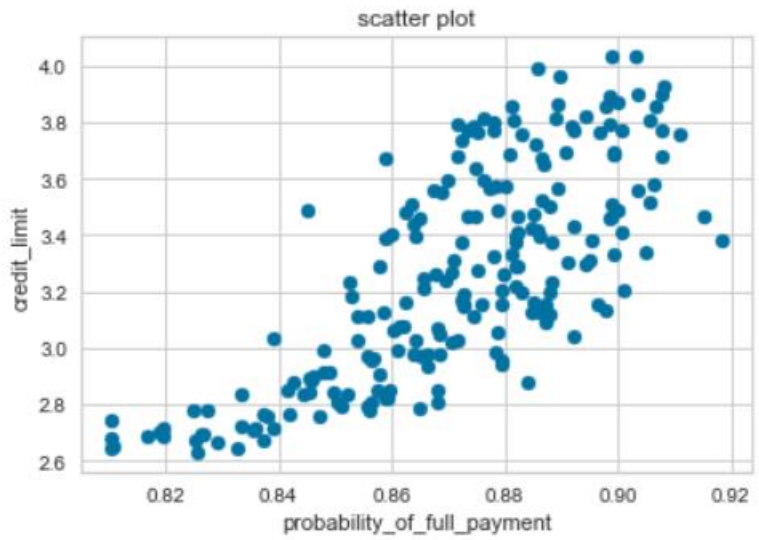
Insights:

1. we can see that both 'credit_limit' & 'advance_payments' are highly correlated that means those who have higher credit limit are supposed to pay in advance before bill gets generated.



Insights:

1. We can see that higher the credit limit means higher the spending.

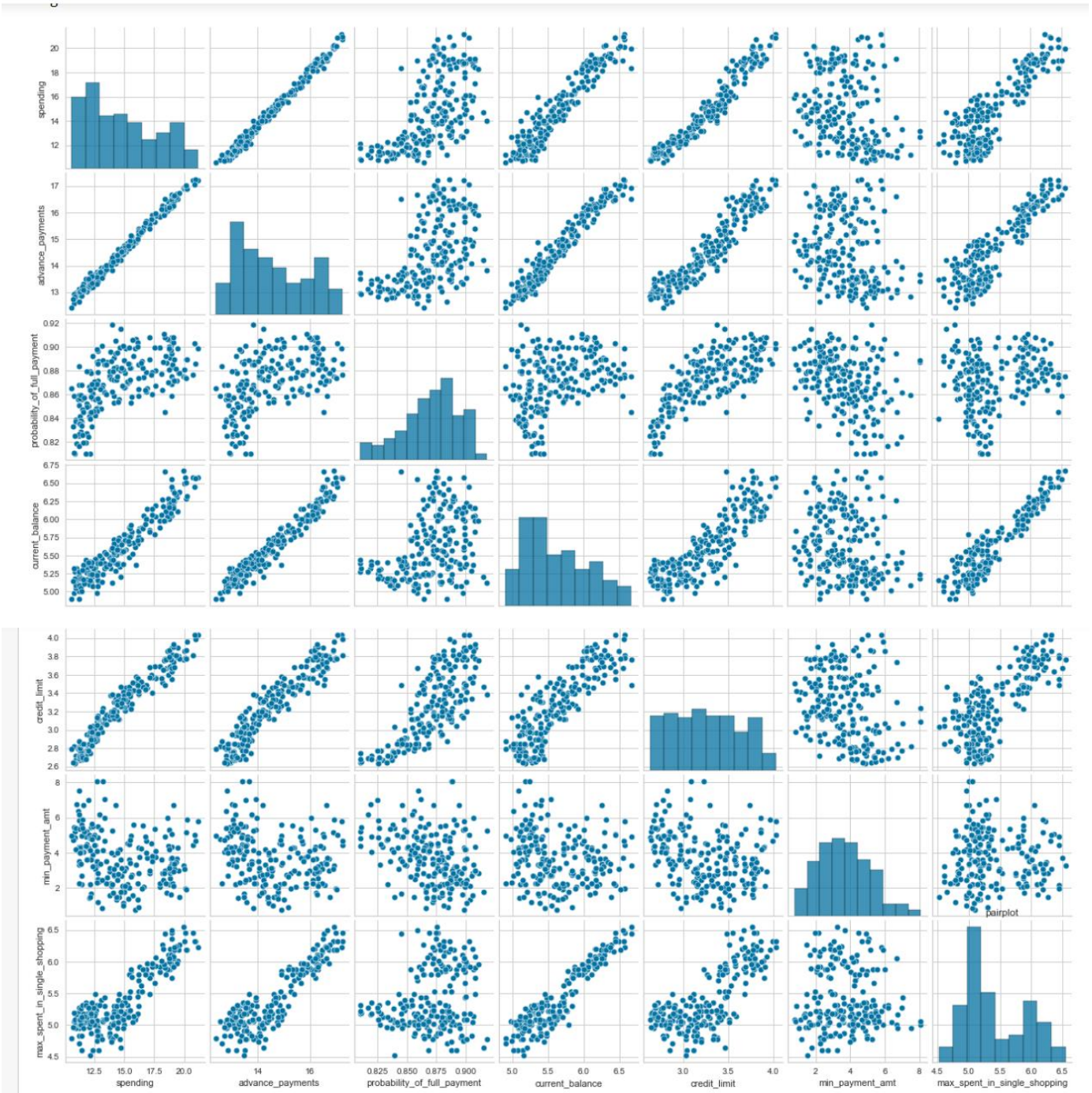


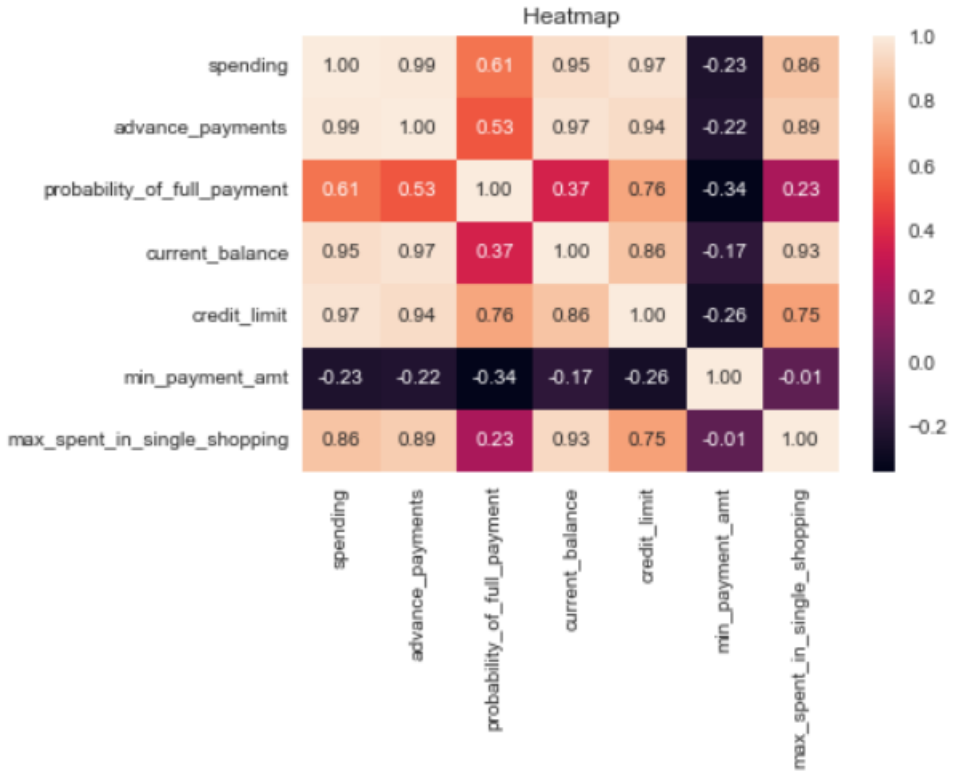
Insights:

1. We can see that those who have higher credit limit are expected to pay full amount and therefore, Credit limit seems to be a very important parameter which is obvious because credit limit assigned based on the customer transaction and relation with bank.

Multivariate Analysis:

We are using pairplot and heat map to find out correlation between variables.


spending
advance_payments
probability_of_full_payment
current_balance
credit_limit
min_payment_amt
max_spent_in_single_shopping
pairplot



Insights:

1. variable min_payment_amt has weak correlation with other variables.
2. variable spending has strong correlation with other variables expect min_payment_amt.
3. Variable 'spending' is highly correlated to advance_payments which indicate that when customer spent high amount while shopping, he also makes advance payment before bill gets generated.

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

	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

We can see that mean, std and median is not same for all features so if we go without scaling then one of the features will have higher impact on our final output.

From above output we can see that in our data there is a difference in the magnitude of the values therefore we need to bring the variables on the same scale and hierarchical clustering methos used distance-based computation so scaling is required so that all features have same weightage.

we are using standrdscaler which applied Z score and bring mean to 0 and standard deviation to 1

Sample of scaled data:

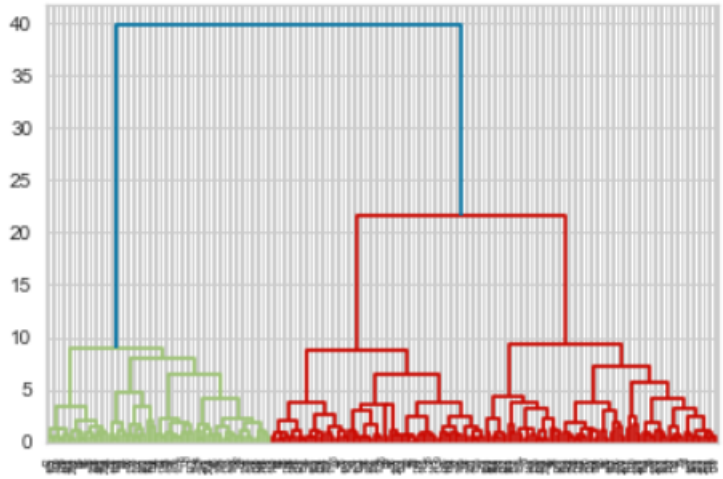
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.177628	2.367533	1.338579	-0.298625	2.328998
1	0.393582	0.253840	1.505071	-0.600744	0.858236	-0.242292	-0.538582
2	1.413300	1.428192	0.505234	1.401485	1.317348	-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

We have successfully scaled the data and now this data can be used for our clustering technique

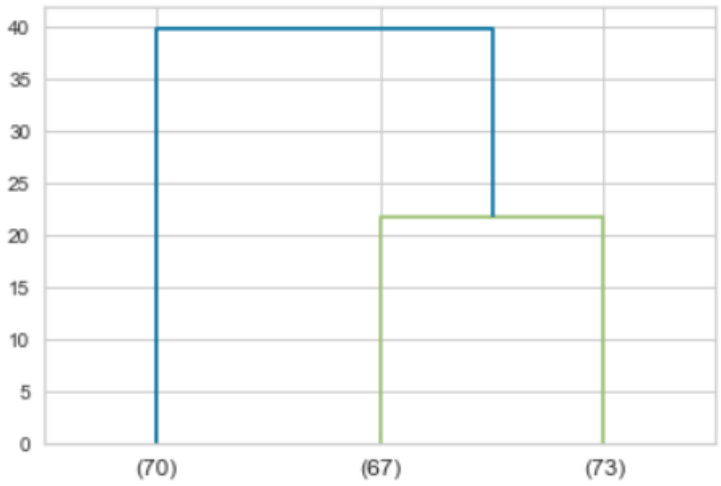
Q 1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them?

We will try different dendrogram with different linkage method and will select best linkage method which gives us best number of clusters.

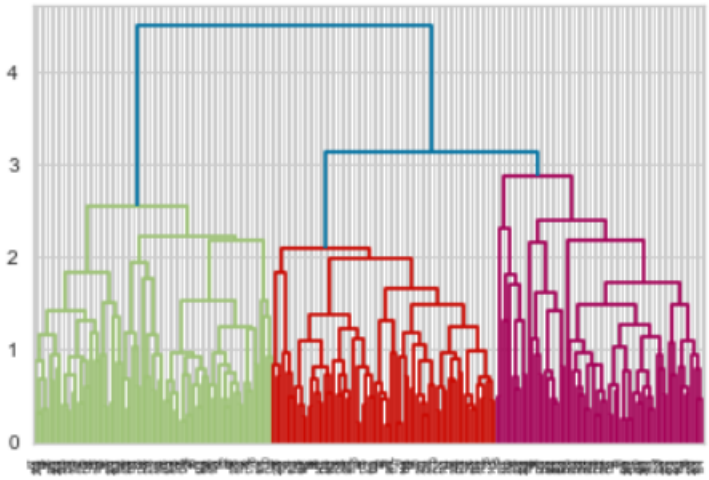
Ward Linkage method:



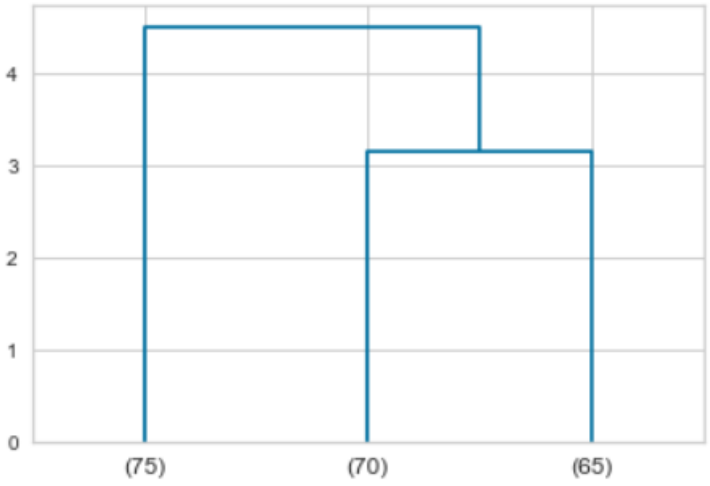
With only 3 number of clusters:



Average Linkage method:

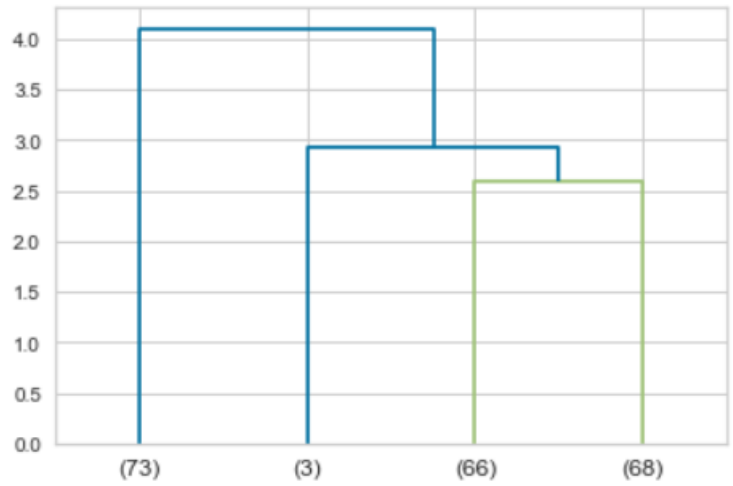


With only 3 number of clusters:

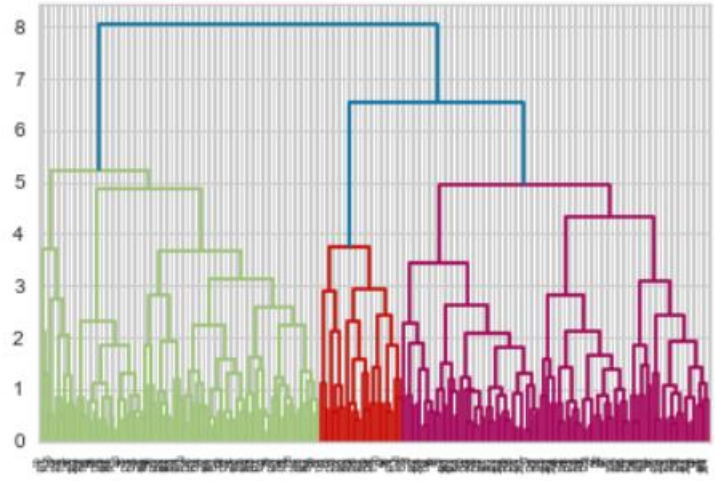


Centroid Linkage method:

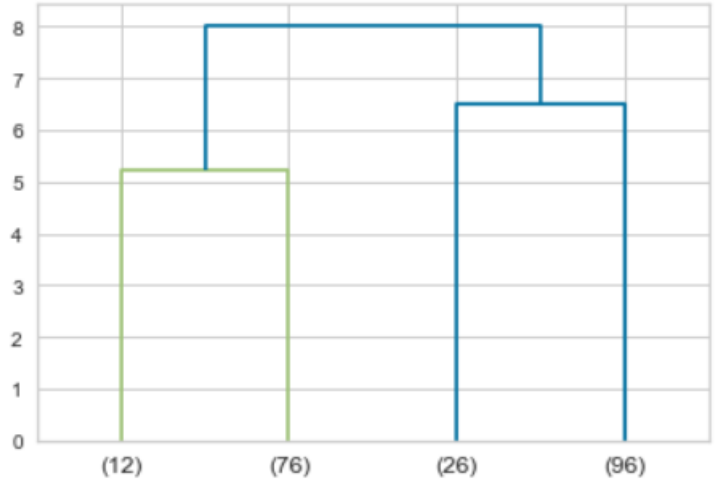
With only 4 number of clusters:



Complete Linkage method:



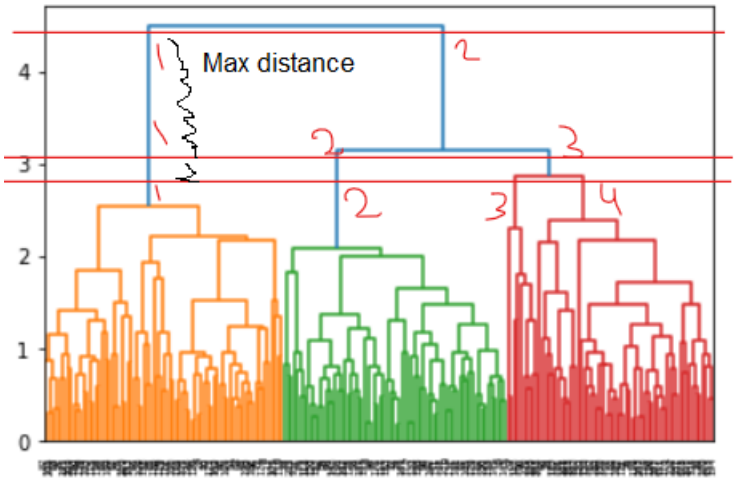
With only 4 number of clusters:



From above dendrograms with different linkage method we can see that dendrogram with method centroid giving 3 data points in 1 one of the clusters so it is not useful similarly in complete linkage method 1 cluster having 12 data points so these are not useful methos here.

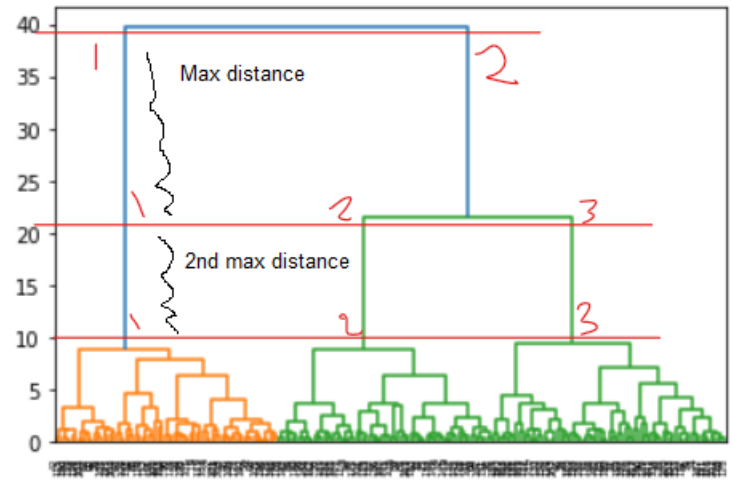
In dendrogram obtained from ward linkage method we can form 3 number of clusters after cutting the dendrogram as it covers maximum distance from 2 to 3 cluster so we will perform fcluster by using 3 number of clusters

Below we have attached images of cutting dendrogram to see visually that how we decide 3 number of clusters.



We can see here that we got max distance when we jump 2 cluster to 3 clusters but if we go ahead from 3 to 4 cluster jump is not significant

Fig: Dendrogram obtained from Average method



We can see here that we got max distance when we jump 2 cluster to 3 clusters and inside 3 cluster there is good distance but if we go ahead from 3 to 3+ clusters jump is not significant

Fig: Dendrogram obtained from ward method

Sample of dataset with cluster columns:

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

Q 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?

In KMeans clustering we need to assign cluster first then we evaluate the output with given clusters.

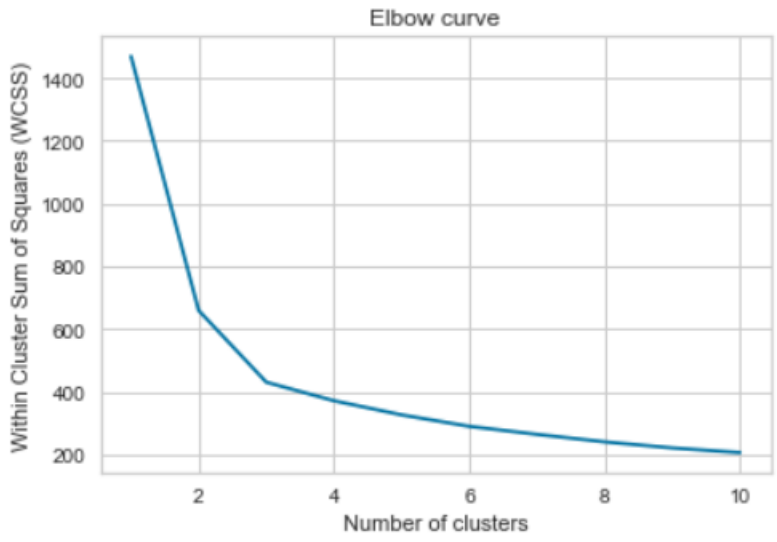
So first trying with K=2 clusters
And got WSS/inertia value as = 659.1474009548498

Now Calculating WSS for other values of K from 1 to 10 and got following WSS values.

```
[1469.9999999999995,
 659.1474009548498,
 430.29848175122294,
 371.0356644664014,
 325.97412847298756,
 289.45524862464816,
 263.859944426353,
 239.9444663501791,
 220.5935394610811,
 205.76334196787008]
```

It observed that WSS reduces as K keeps increasing.

Plotting these wss values on elbow curve.



From above graph we can see that wss is not dropping significantly after cluster 3 so we can go with 3 optimum clusters.

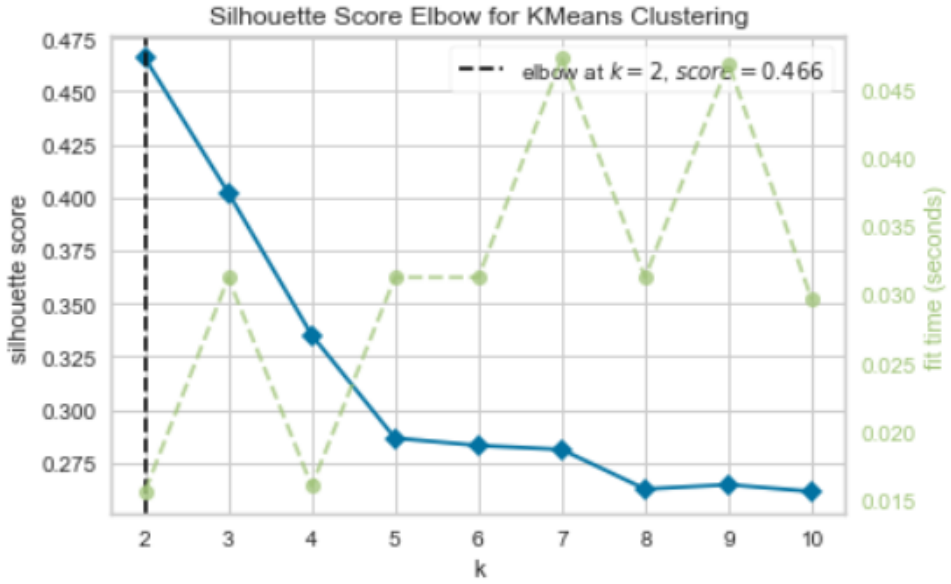
Cluster evaluation for 3 & 4 clusters the silhouette score:

silhouette score for 3 number of clusters = 0.4008059221522216

silhouette score for 4 number of clusters = 0.3373662527862716

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

We can also use K-Elbow Visualizer which can tells the best optimum cluster with best silhouette score.



From Above elbow graph we can see silhouette score for all clusters and according to it 2 is the best number of cluster for this data but 2 cluster does not make any sense to the business therefore we will go with 3 clusters as silhouette score is better in cluster 3 than cluster 4.

Sample of the data after Appending KMeans Clusters:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Clus_kmeans4
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	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

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

Cluster Profiling:

First, we will talk about clusters obtained from hierarchical clustering.

```
3    73
1    70
2    67
Name: Fclusters, dtype: int64
```

So, we have identified 3 optimum clusters using Dendrogram where cluster 3 has maximum 73 data points and cluster 1 has second highest 70 data points and cluster 2 has remaining 67 data points.

Getting mean of the variables with respect to the clusters.

Fclusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371
2	11.872388	13.257015	0.848072	5.238940	2.848537	4.949433	5.122209
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178

Cluster 1: Highest spending customers.

Cluster 2: lowest spending customers but spending in single shopping is good.

Cluster 3: moderate spending customers with high probability of paying full payment.

Recommendations:

1. we can see that in cluster 2 where 67 customers spent very less and their probability_of_full_payment is also low but there spending in single shopping is good despite having low credit limit so bank can give credit limit increase offers to around 13 customers who have probability_of_full_payment between 86-88% & out of 67 customers 9 customers have probability_of_full_payment between 81-82% so bank should focus more on these customers.

2. In cluster 1, 70 customer is the part of this cluster and their credit card usage is quite good as they have spent 83% of their credit card limit so bank can offer credit card limit increase OR other credit cards to these customers as offering loan to these customers won't make much difference as they already have a high balance.

3. In cluster 3 majority of the customer falls in this group and they spent moderate amount in shopping but their probability of paying full payment is good so bank can give some offer to these customers which lead customer to more spending may be card, they are using don't have much offers and discount so bank can look into this.

4. In clusters 2 customers despite having low credit limit spending good amount in single shopping that means these customers use credit card more in certain situation like festival or any sell, discount etc. giving loan to these customers can also be a good idea.

Cluster profile for KMeans clusters:

```
0    72
2    71
1    67
Name: Clus_kmeans4, dtype: int64
```

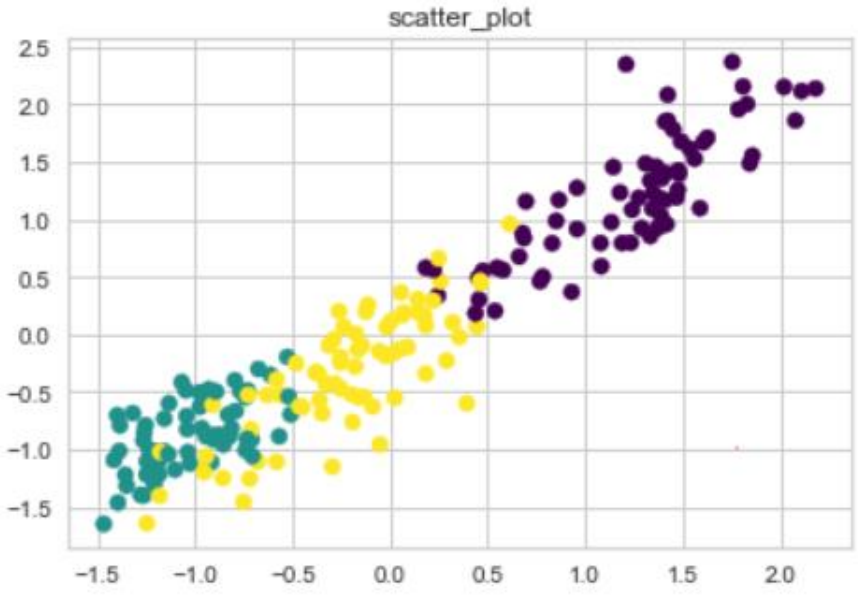
So, we have identified 3 optimum clusters using KMeans where cluster 0 has maximum 72 data points and cluster 2 has second highest 71 data points and cluster 1 has remaining 67 data points.

Getting mean of the variables with respect to the clusters:

Clus_kmeans4	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	11.856944	13.247778	0.848330	5.231750	2.849542	4.733892	5.101722
1	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701
2	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803

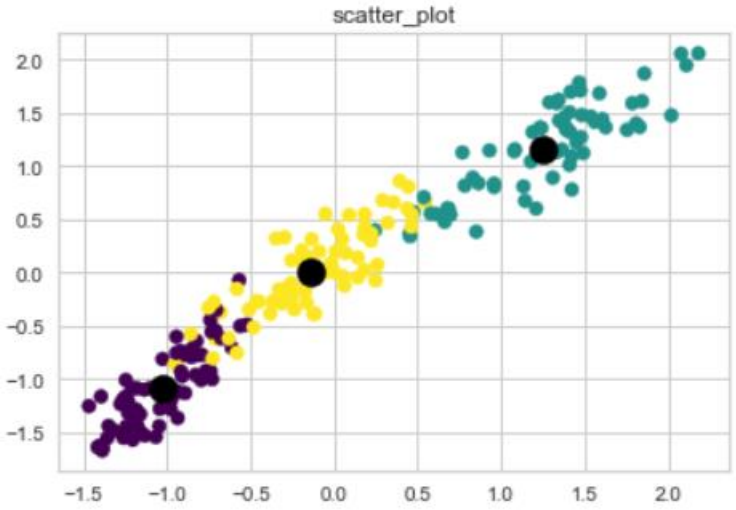
As we can see that there is no significant difference in the mean of variables from Fclusters to Clus_kmeans4 as number of clusters is same 3 and number of data points contained in these 3 clusters are also almost same so we can go with same Recommendations.

By using scatter plot Visualising Fclusters that how exactly it formed.



we can see that 3-cluster formed in the graph which we obtained in hierarchical clustering.

KMeans clusters visualization by using scatter plot:



we can see that 3-cluster formed in the graph which we obtained in KMeans clustering and dark black circle is centroid of clusters.

Problem 2

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.

Data Description:

1. Age: Age of insured in numbers
2. Agency_Code: Code of tour firm like - C2B, EPX, CWT, JZI
3. Type: Type of tour insurance firms like- Airlines, Travel Agency
4. Claimed: Claim submitted status- Yes or No
5. Commission: The commission received for tour insurance firm in numbers
6. Channel: Distribution channel of tour insurance agencies like- Online, Offline
7. Duration: Duration of the tour (Duration in days) in numbers
8. Sales: Amount worth of sales per customer in numbers
9. Product Name: Name of the tour insurance products like- Customised Plan, Bronze Plan, Gold Plan, Cancellation Plan, Silver Plan
10. Destination: Destination of the tour like- ASIA, Americas, EUROPE

Sample of the dataset:

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

Exploratory Data Analysis:

Let us check the types of variables in the data frame.

Column	Dtype
Age	int64
Agency_C	object
Type	object
Claimed	object
Commisio	float64
Channel	object
Duration	int64
Sales	float64
Product N	object
Destinatic	object

There are total 3000 rows and 10 columns in the dataset and out of 10, 6 columns are object data type, 2 columns are integer data types and rest is float data type.

Checking for missing values in the dataset:

```
 #   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)
```

From the above results we can see that there is no missing value present in the dataset.

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

As per checking there are 139 duplicate values found but I'm not removing these because there is no customer ID or any unique Id mentioned so these can be the different customers like same plan can be given to the other customer and we don't have any ID, unique number so cannot remove these values.

	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

we have seen minimum tour duration is -1 which is not correct and even 0 is also not correct why company will give insurance for 0 days in any plan it means no plan given so we need to compute these values and we are replacing with nearest value 1.

Sample data after replacing anomalies:

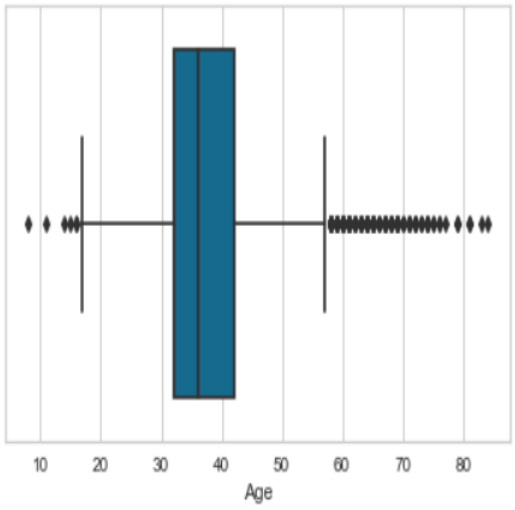
	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.002667	134.052619	1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

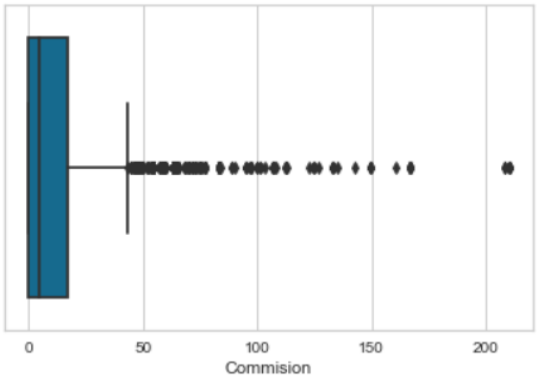
Checking unique values for categorical column:

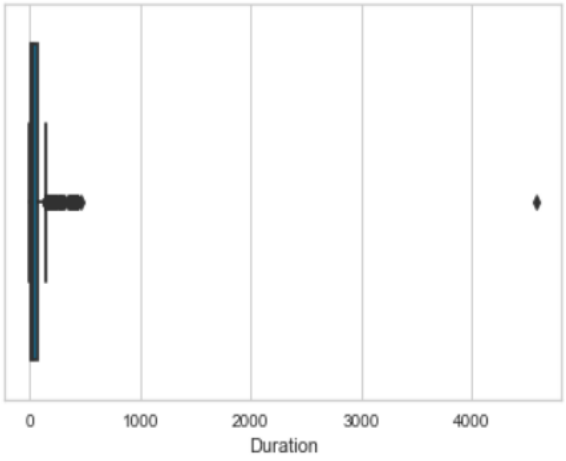
```
['C2B' 'EPX' 'CWT' 'JZI']
['Airlines' 'Travel Agency']
['No' 'Yes']
['Online' 'Offline']
['Customised Plan' 'Cancellation Plan' 'Bronze Plan' 'Silver Plan'
 'Gold Plan']
['ASIA' 'Americas' 'EUROPE']
```

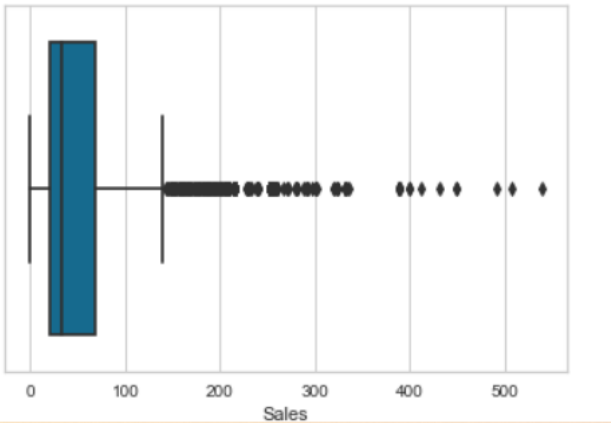
Unique values look good as no repetition found.

<u>Checking dataset for outliers:</u>









We can see that outliers present in the data for now we are not treating the outliers.

Data Visualization:

Univariate Analysis:

Non visual representation:

	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	70.002667	134.052619	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

Insights:

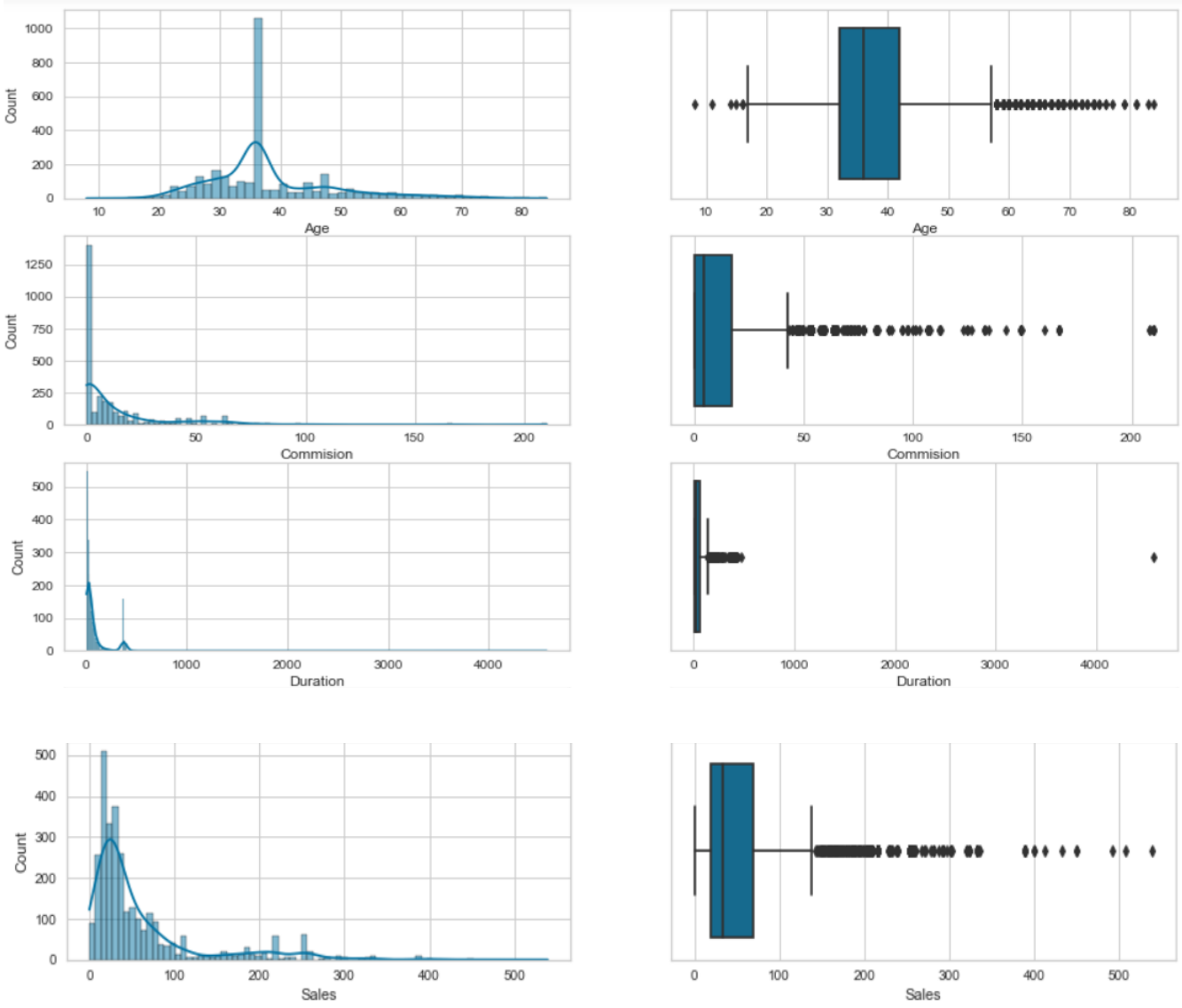
1. 50% of the customer's age is less than 36 and maximum age is 84 and average age is 38.

2. Average duration of the tour is 70 and minimum age duration offer by company is 1.

3. commission & Sales- mean and median varies significantly.

Visual representation:

We are using Boxplots and histogram to visualize the distribution of the data.



Insights:

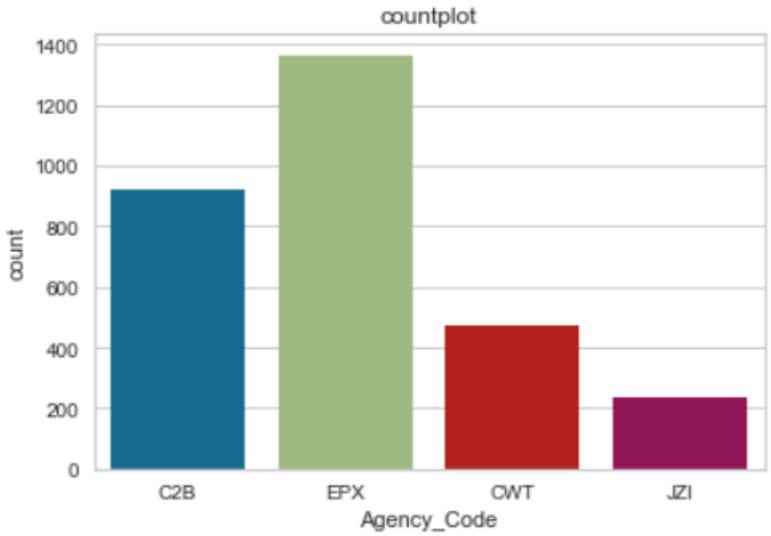
1. We can see that outliers present in the data for all continuous variables.

2. Almost symmetric distributions observed for column 'Age'.

3. Apart from variable 'Age' all 3-variable showing distribution of data is skewed to the Right.

For Categorical Variables:

We are plotting multiple point plot for multiple categorical variables.

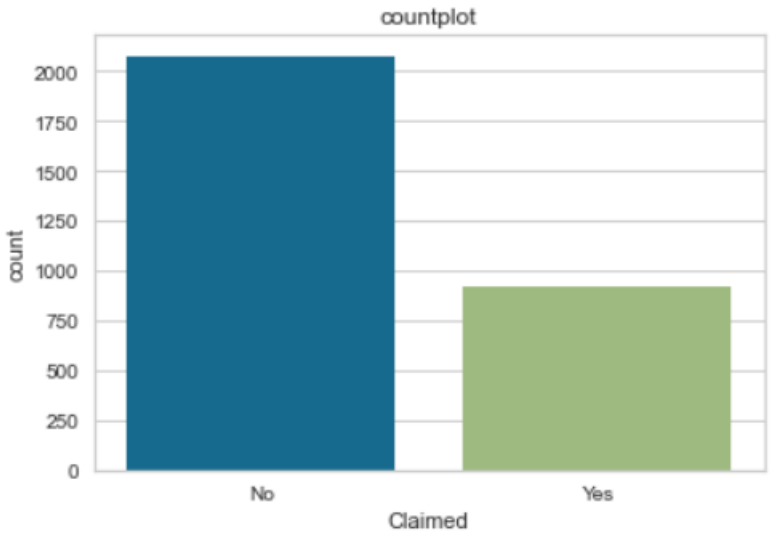


Insights:

1. Maximum Agency code 'EPX' assigned to the customers.

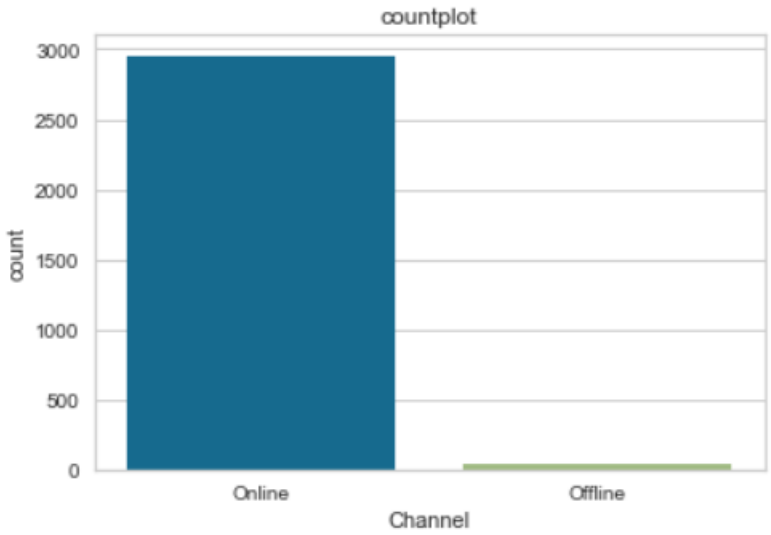
Insights:

1. Travel Agency is the most Type of tour insurance firms.



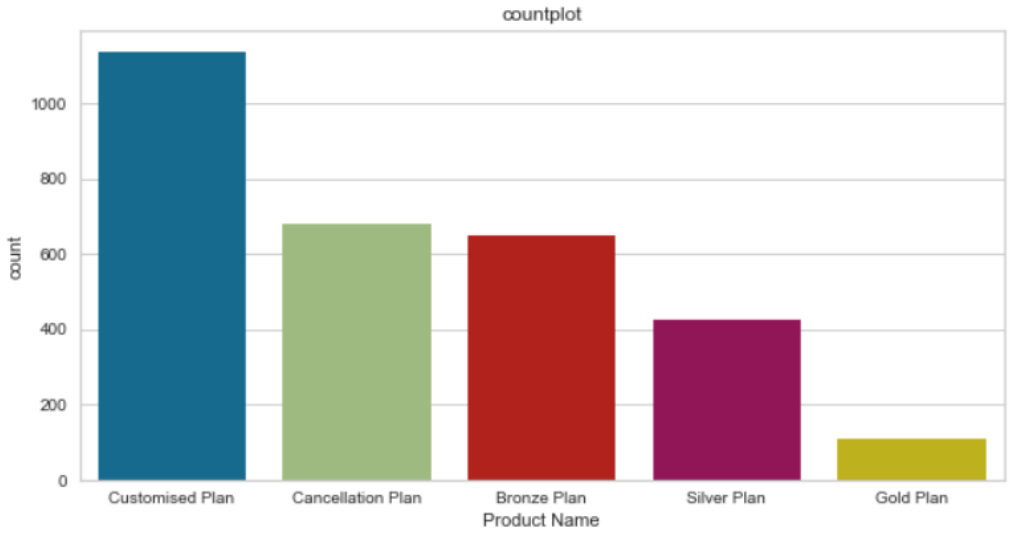
Insights:

1. Almost half of the customer submitting claims which is a concern.



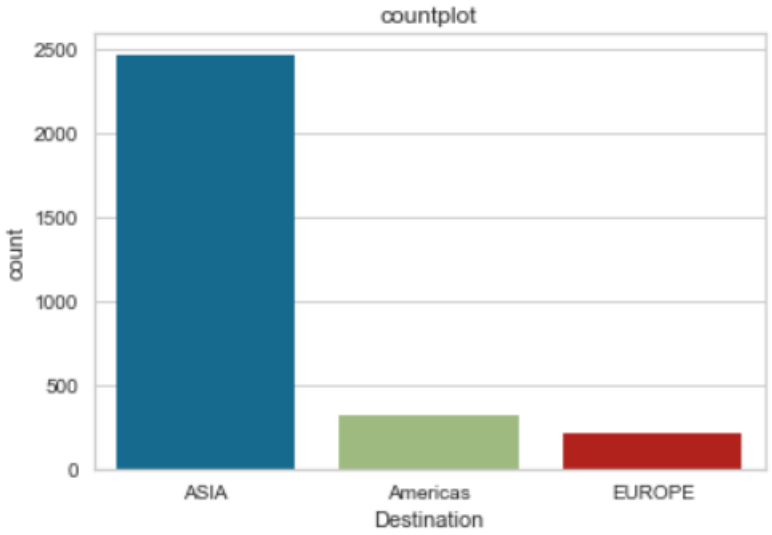
Insights:

1. mostly insurance distribution is through online.



Insights:

1. Most popular plan is Customised Plan and Gold plan is not so famous.

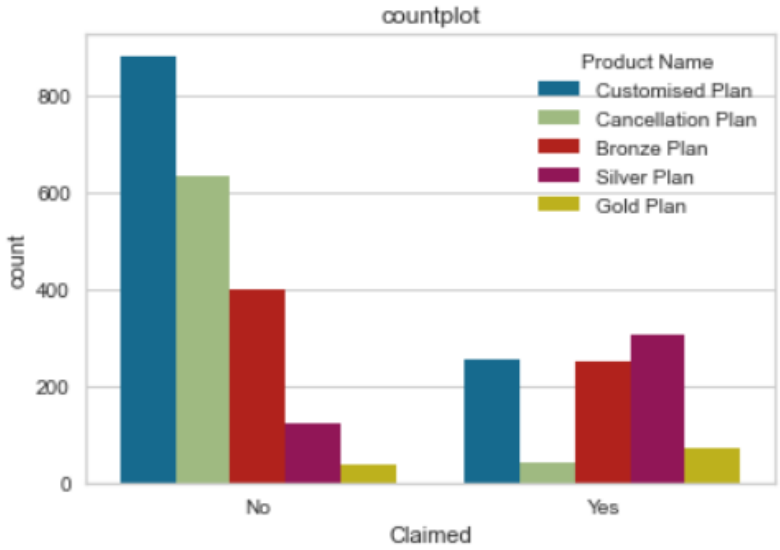


Insights:

1. Most of the customer prefer their destination as Asia.

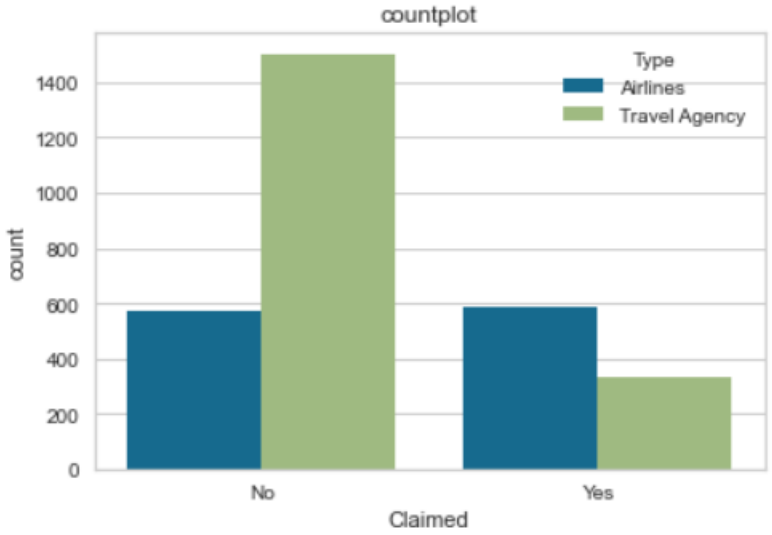
Bivariate Analysis:

We are plotting count plot and boxplot for set of 2 or more variables.



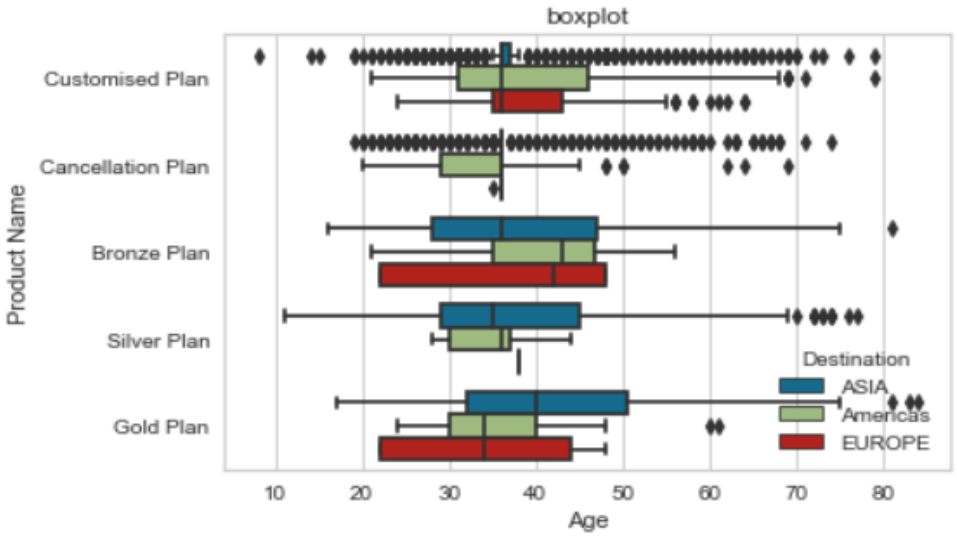
Insights:

1. Customer who has silver plan are the most who submitting claims so they might be not happy with the benefits offered within plan.



Insights:

1. Airlines has minimum number in terms of tour insurance firms in spite of that most of the customer who submitting claims belongs to type travel agency.



Insights:

1. Cancellation plan only chosen by the customer whose age is more than 18 and in silver plan there is only 1 customer whose destination is EUROPE.

Multivariate Analysis:

We are using pairplot and heat map to find out correlation between variables.

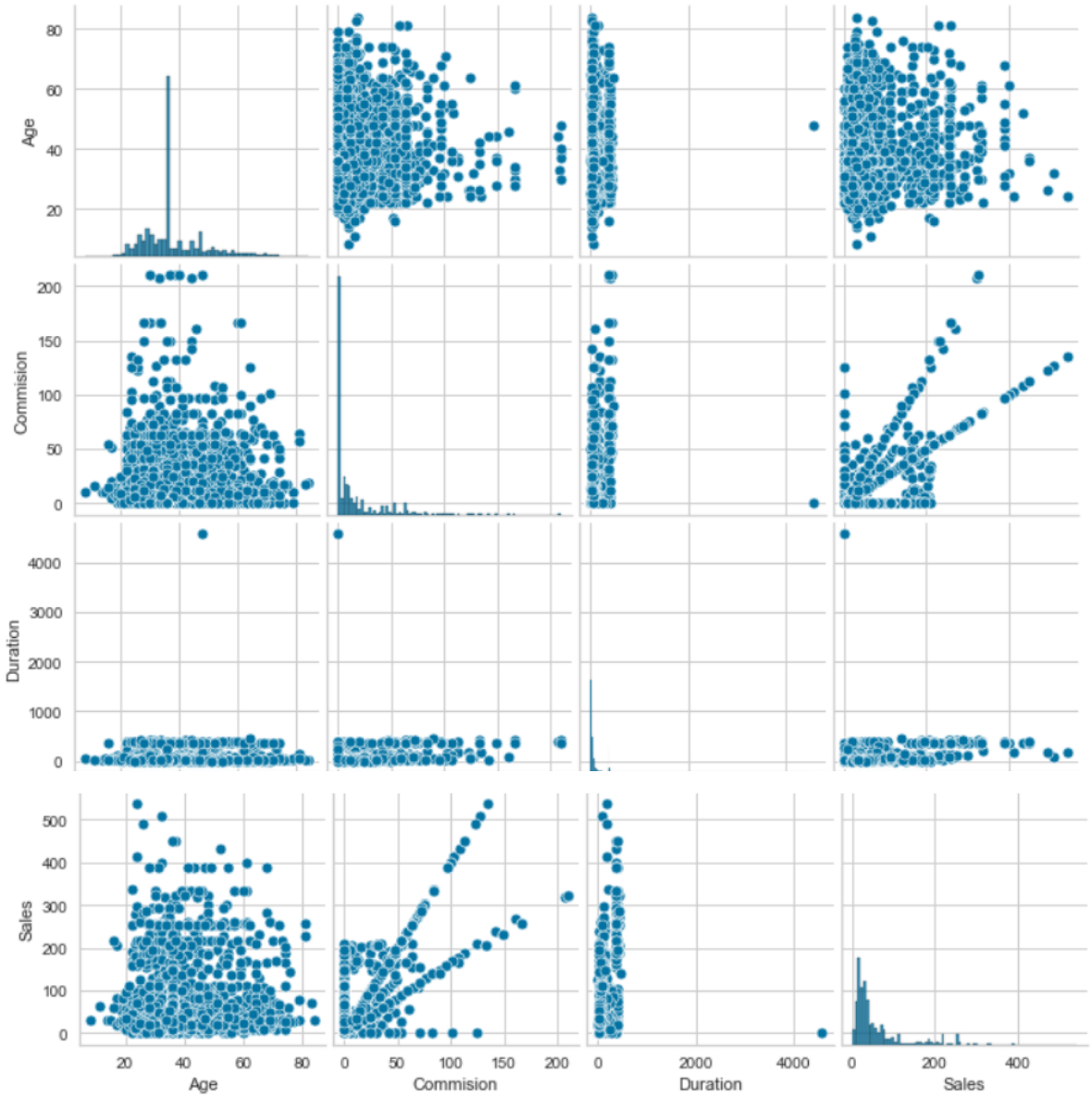
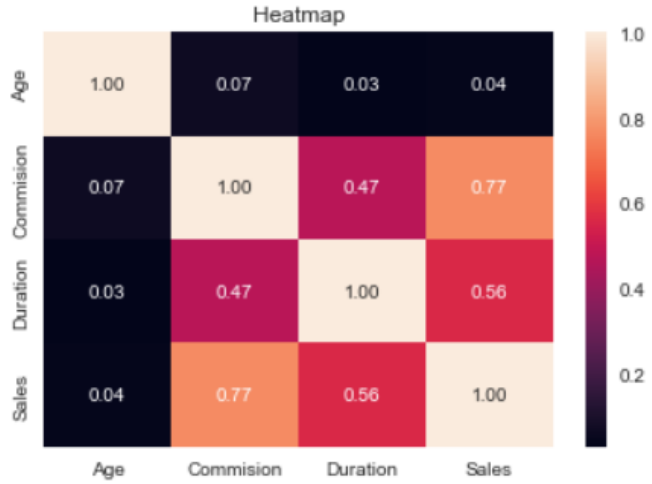


Fig: Pair plot

Now we are plotting Heatmap



Insights:

1. Based on the heatmap there is no negative relation found between variables.

2. Commission is highly correlated to sales as Commission is in percentage of sales so higher the sales mean higher the commission.

3. Age has weak correlation with other variables.

Q2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network?

We need to convert Object data type into categorical/numerical data to fit in the models and we are using encoding technique as (pd.categorical().codes()

And below is the output after converting data type into numerical data type.

```
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]
```

Proportion of 1s and 0s:

```
0    0.692
1    0.308
Name: Claimed, dtype: float64
```

Almost 30 % is YES and 70% is No in the data.

Extracting the target column into separate vectors for training set and test set:

Sample of the data which contain independent variables.

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

Target variable Claimed assign to variable Y.

Splitting data:

Splitting data into training and test set with 30% random value assigning to the test and remaining 70% is to the train set.

Shape of the train and test data.

```
X_train (2100, 9)
X_test (900, 9)
Y_train (2100,)
Y_test (900,)
```

We get 4 values from train test split.

Building a Decision Tree Classifier:

After fitting the train data to the model, we checked decision tree image on webgraphviz website & found that tree has overgrown/overfitted so we need to use Pruning to cut the decision tree from the middle.

Variable Importance:

```
                  Imp
Duration      22.662453
Agency_Code   21.637552
Sales         20.735783
Age           20.316413
Commision       7.986457
Product Name    4.056698
Destination     2.042921
Channel         0.338268
Type            0.223456
```

Variables Duration, Agency_Code, Sales, Age are the most important variable for our prediction.

Regularising the Decision Tree:

To avoid tree to be over grown we are using grid search to identify best parameters values for our model and after trying with multiple combination of max_depth, min_samples_leaf, min_samples_split we arrived at final parameter which is

criterion': 'Gini', 'max_depth': 4.85, 'min_samples_leaf': 10, 'min_samples_split': 150}
and we will build our model by using these parameters.

Variable Importance after generating new tree by using best parameters:

	Imp
Agency_Code	61.778648
Sales	21.505494
Product Name	7.840864
Duration	5.410453
Commision	3.056614
Age	0.407928
Type	0.000000
Channel	0.000000
Destination	0.000000

we can see with best parameters now important variables are only 2.

Predicting on Training and Test dataset:

Shape of the train and test predicting data.

```
ytrain_predict (2100,)
ytest_predict (900,)
```

Predicted Probabilities: we have added top 5 values below
For test Data-

	0	1
0	0.961538	0.038462
1	0.815057	0.184943
2	0.856589	0.143411
3	0.856589	0.143411
4	0.856589	0.143411

For train data-

	0	1
0	0.856589	0.143411
1	0.653333	0.346667
2	0.815057	0.184943
3	1.000000	0.000000
4	0.815057	0.184943

Building Random Forest Classifier:

We will use grid search to identify best parameters values for our model and after trying with multiple combination of max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators we arrived at final parameter which is

Best_grid_search = 'max_depth': 5, 'max_features': 8, 'min_samples_leaf': 8, 'min_samples_split': 46, ‘n_estimators': 350 and we will build our model by using these parameters.

Variable Importance:
After building RF model by using best grid search, we got following important variables.

	Imp
Agency_Code	0.495073
Sales	0.190351
Product Name	0.155380
Duration	0.059558
Commision	0.055168
Age	0.038734
Destination	0.004382
Channel	0.000969
Type	0.000385

Variables Agency_Code, Sales, Product Name are the important variables for prediction.

Predicting Probs*:* we have added top 5 values below

For test Data:

	0	1
0	0.870181	0.129819
1	0.818828	0.181172
2	0.810778	0.189222
3	0.817909	0.182091
4	0.909925	0.090075

For train data:

	0	1
0	0.783598	0.216402
1	0.690303	0.309697
2	0.712918	0.287082
3	0.736137	0.263863
4	0.741013	0.258987

Predicting on Training and Test dataset:
Shape of the train and test predicting data

```
ytrain_predict (2100,)
ytest_predict (900,)
```

Building a Neural Network Classifier:

we have to scale the data for ANN model and after applying scaling on the data X_trains and Y_trains are our scaled data.

We will use grid search to identify best parameters values for our model and after trying with multiple combination of hidden_layer_sizes, max_iteration, tolerance level, activation function we arrived at final parameter which is-

Best_grid_search = 'activation': 'relu', 'hidden_layer_sizes': 500, 'max_iter': 5000, 'solver': 'adam', 'tol': 0.01

NOTE: we cannot get Feature importance in ANN model we can say it's a disadvantage of ANN.

Predicting Probs*:* we have added top 5 values below

For test Data:

	0	1
0	0.880974	0.119026
1	0.795564	0.204436
2	0.859999	0.140001
3	0.882079	0.117921
4	0.903492	0.096508

For train data:

	0	1
0	0.812167	0.187833
1	0.690626	0.309374
2	0.727569	0.272431
3	0.425014	0.574986
4	0.675581	0.324419

Predicting on Training and Test dataset:

Shape of the train and test predicting data

```
ytrain_predict (2100,)
ytest_predict (900,)
```

Q2.3 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?

For CART Model

We are performing Confusion Matrix, Classification Report, AUC and ROC for the training data of CART model.

confusion_matrix :

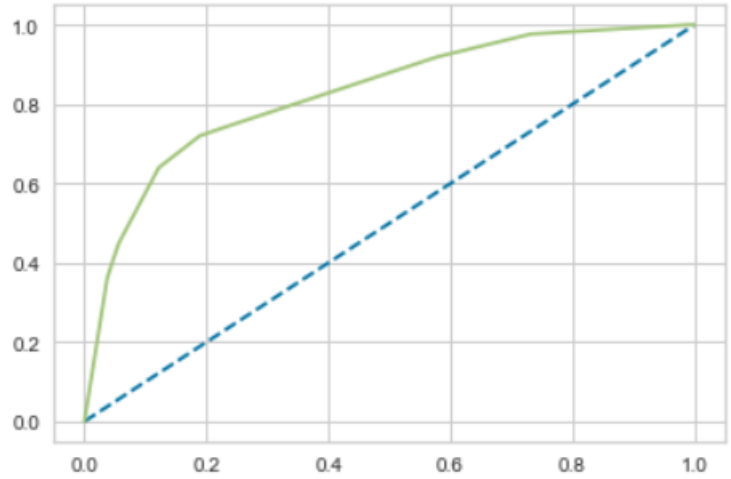
```
[1276,  177],
[ 233,  414]
```

classification_report:

	precision	recall	f1-score	support
0	0.85	0.88	0.86	1453
1	0.70	0.64	0.67	647
accuracy			0.80	2100
macro avg	0.77	0.76	0.77	2100
weighted avg	0.80	0.80	0.80	2100

AUC Score and ROC curve:

```
AUC: 0.826
[<matplotlib.lines.Line2D at 0x1b66e4174f0>]
```



We are performing Confusion Matrix, Classification Report, AUC and ROC for the test data of CART model.

confusion_matrix :
```
[537,  86],
[119, 158]
```

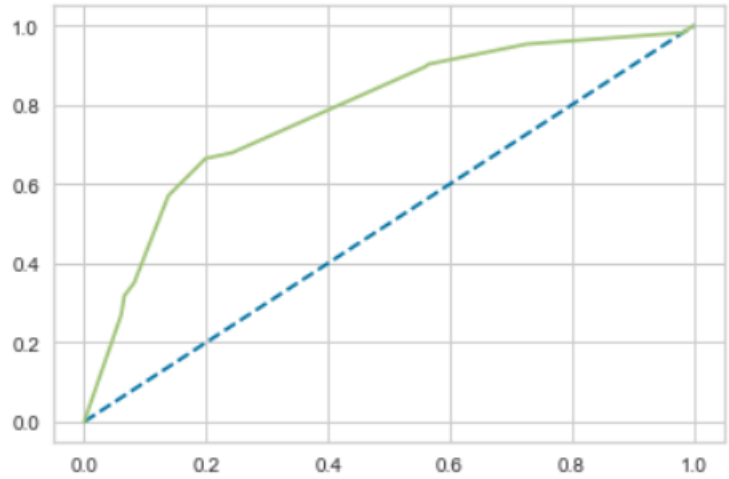
classification_report:

	precision	recall	f1-score	support
0	0.82	0.86	0.84	623
1	0.65	0.57	0.61	277
accuracy			0.77	900
macro avg	0.73	0.72	0.72	900
weighted avg	0.77	0.77	0.77	900

AUC Score and ROC curve:

```
AUC: 0.777

[<matplotlib.lines.Line2D at 0x1b6668d5c40>]
```



Cart Conclusion:

Train Data*:*

AUC- 82%

Accuracy- 80%

Precision- 70%

f1-Score- 67%

test Data*:*

AUC- 77%

Accuracy- 77%

Precision- 65%

f1-Score- 61%

Training and Test set results showing some changes, The Overall model performance is moderate enough to start predicting if any customer will submit claim or not.

Agency_Code, Sales is the most important variable for predicting claims.

For RF Model:

We are performing Confusion Matrix, Classification Report, AUC and ROC for the training data of RF model.

confusion_matrix :
```
[1309,  144],
[ 247,  400]
```

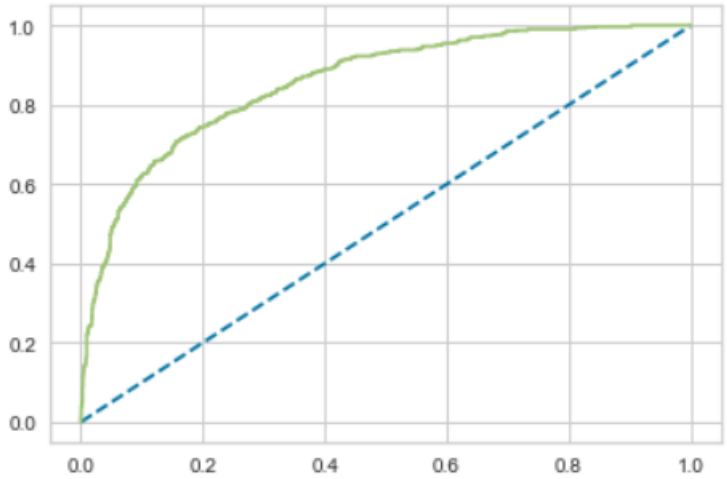
classification_report:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	1453
1	0.74	0.62	0.67	647
accuracy			0.81	2100
macro avg	0.79	0.76	0.77	2100
weighted avg	0.81	0.81	0.81	2100

AUC Score and ROC curve:

```
AUC: 0.857

[<matplotlib.lines.Line2D at 0x1b66ad55910>]
```



We are performing Confusion Matrix, Classification Report, AUC and ROC for the test data of RF model.

confusion_matrix :
```
[546,  77],
[131, 146]
```

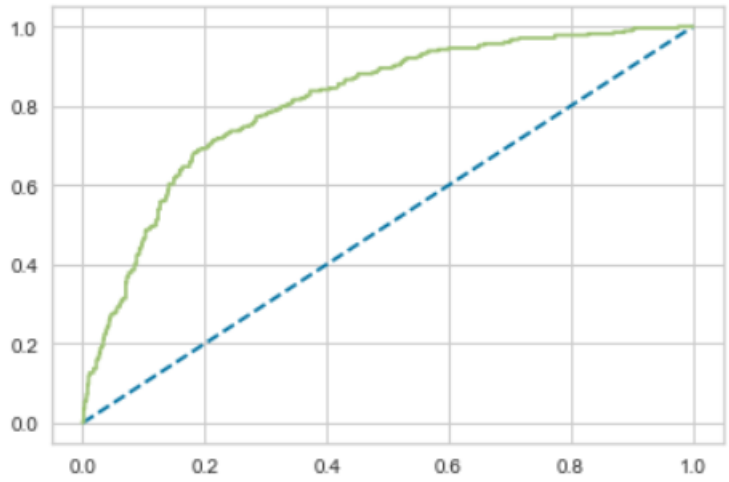
classification_report:

	precision	recall	f1-score	support
0	0.81	0.88	0.84	623
1	0.65	0.53	0.58	277
accuracy			0.77	900
macro avg	0.73	0.70	0.71	900
weighted avg	0.76	0.77	0.76	900

AUC Score and ROC curve:

```
AUC: 0.811

[<matplotlib.lines.Line2D at 0x1b66d9828b0>]
```



RF Conclusion:

Train Data:

AUC- 85%

Accuracy- 81%

Precision- 74%

f1-Score- 67%

test Data:

AUC- 81%

Accuracy- 77%

Precision- 65%

f1-Score- 58%

Training and Test set results showing some changes, The Overall model performance is moderate enough to start predicting if any customer will submit claim or not.

Agency_Code, Sales, Product Name are the most important variable for predicting claims.

For ANN Model:

We are performing Confusion Matrix, Classification Report, AUC and ROC for the training data of ANN model.

confusion_matrix :

```
[1289,  164],
[ 281,  366]
```

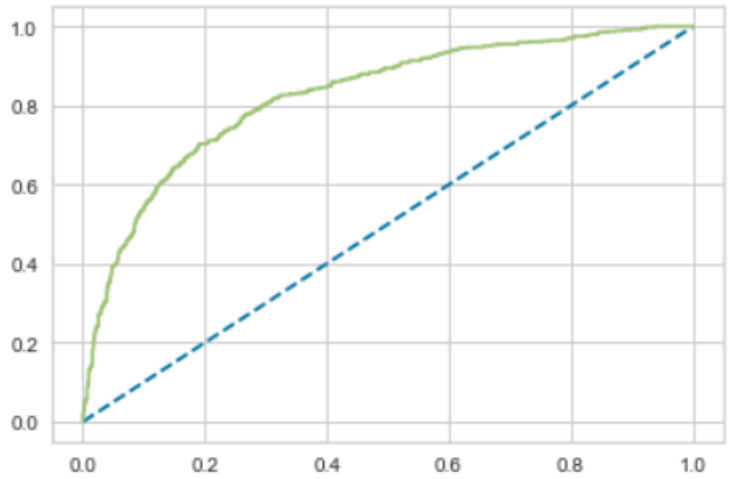
classification_report:

	precision	recall	f1-score	support
0	0.82	0.89	0.85	1453
1	0.69	0.57	0.62	647
accuracy			0.79	2100
macro avg	0.76	0.73	0.74	2100
weighted avg	0.78	0.79	0.78	2100

AUC Score and ROC curve:

```
AUC: 0.824

[<matplotlib.lines.Line2D at 0x1b66ddca580>]
```



We are performing Confusion Matrix, Classification Report, AUC and ROC for the test data of ANN model.

confusion_matrix :
```
[539,  84],
[141, 136]
```

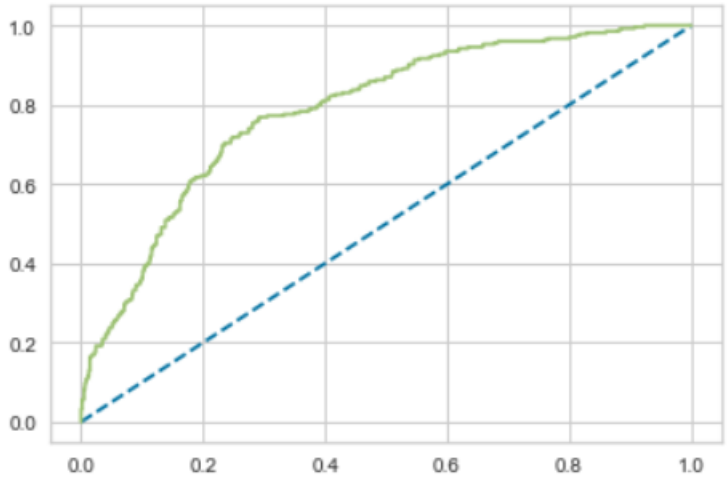
classification_report:

	precision	recall	f1-score	support
0	0.79	0.87	0.83	623
1	0.62	0.49	0.55	277
accuracy			0.75	900
macro avg	0.71	0.68	0.69	900
weighted avg	0.74	0.75	0.74	900

AUC Score and ROC curve:

```
AUC: 0.826

[<matplotlib.lines.Line2D at 0x1b66dd7e070>]
```



ANN Conclusion

Train Data*:*

AUC- 82%

Accuracy- 79%

Precision- 69%

f1-Score- 62%

test Data*:*

AUC- 78%

Accuracy- 75%

Precision- 62%

f1-Score- 55%

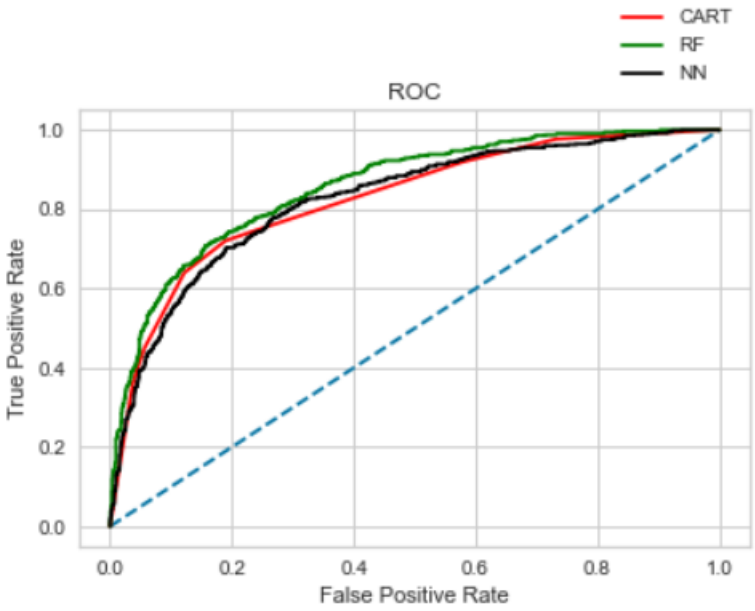
Q2.4 Final Model: Compare all the models and write an inference which model is best/optimized?

Comparison of the performance metrics from the 3 models:

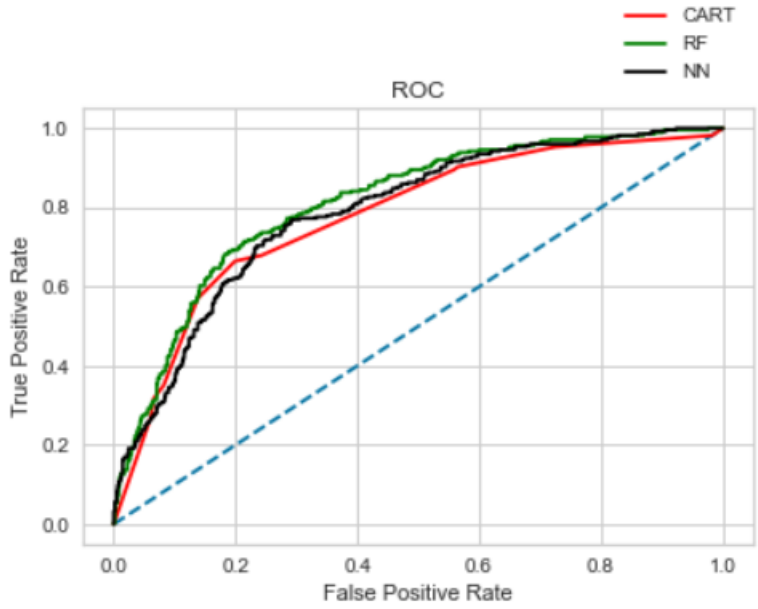
we have created a table which contain all score for train and test data for all models it will be easy to review the model performance.

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.80	0.77	0.81	0.77	0.79	0.75
AUC	0.83	0.78	0.86	0.81	0.82	0.83
Recall	0.64	0.57	0.62	0.53	0.57	0.49
Precision	0.70	0.65	0.74	0.65	0.69	0.62
F1 Score	0.67	0.61	0.67	0.58	0.62	0.55

ROC Curve for all 3 models on the Training data:



ROC Curve for all 3 models on the test data:



Out of the 3 models, Random Forest has slightly better performance than the Cart and Neural network model in terms of overall accuracy.

From Cart and Random Forest Model, the variable Agency_Code, Sales is found to be the most useful feature amongst all other features for predicting if a person will submit claim or not.

Q2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations?

1. As per the data most of insurance is done by online channel Other interesting fact, is almost all the offline business has a claimed associated, need to find why?

2. Agency JZI need to focus more as their sales are very low.

3. There is some moderate underfitting in train and test data so we can do better if we get more data.

4. More sales happen through Agency than Airlines and the trend shows that submitted claims are associated with Airline so need to check what is wrong with this firm.

5. As we have selected RF model as final model and as per checking classification report it obtained that recall is important considering the model has failed to predict 131(FN) customers who did submit claim with a recall of 0.57, so major focus can be upon improving the recall score which can provide some insights for the company to take prior steps in analysing those customers who might submit claims.

THE END.