Question 1.1

Read the data and do exploratory data analysis. Describe the data briefly.

Distribution:

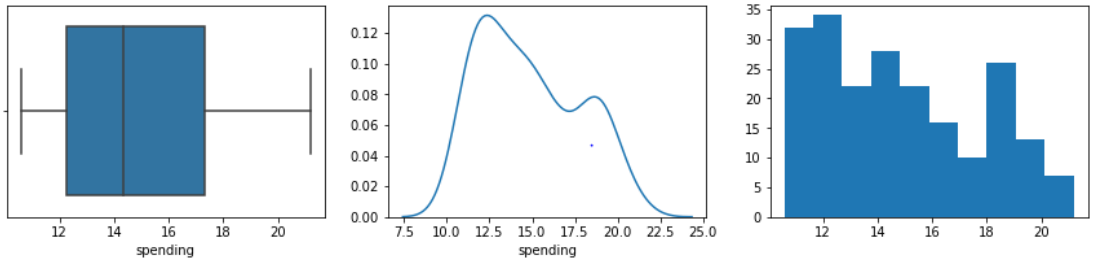
Except for probability_of_full_payment which is somewhat normally distributed none of the other variables have normal distribution.

Outliers:

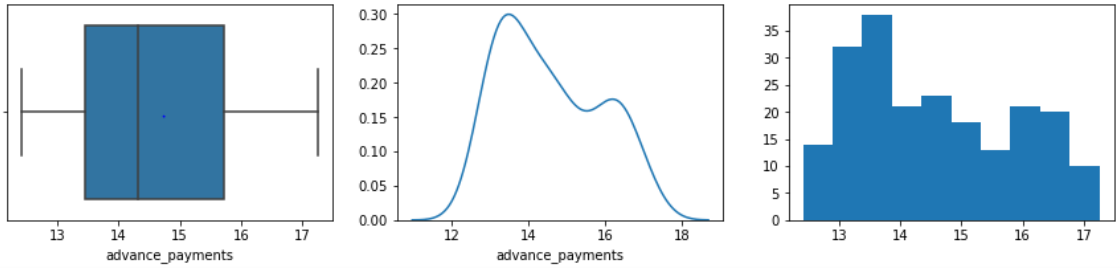
Except for probability_of_full_payment and min_payment_amt which has few outliers none of the other variables have outliers.

Please find below the plots that depict above observations .

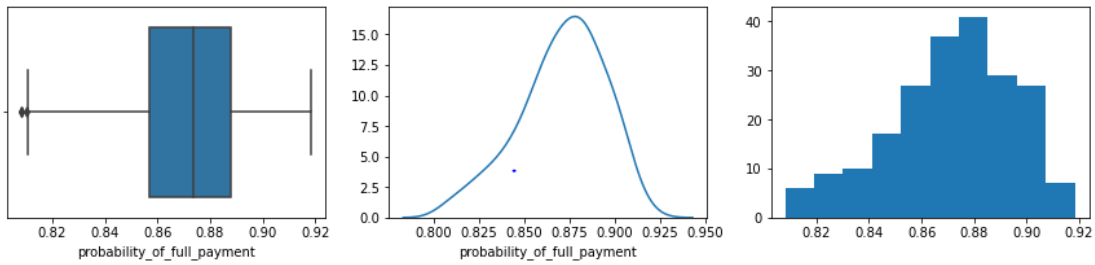
```
1. Univariate analysis for spending

Mean is 14.847524, Median is 14.355000, Mode(s) are 11.2300
Column spending does not have outliers
```




```
Column spending is not normally distributed

 2. Univariate analysis for advance_payments

Mean is 14.559286, Median is 14.320000, Mode(s) are 13.4700
Column advance_payments does not have outliers
```



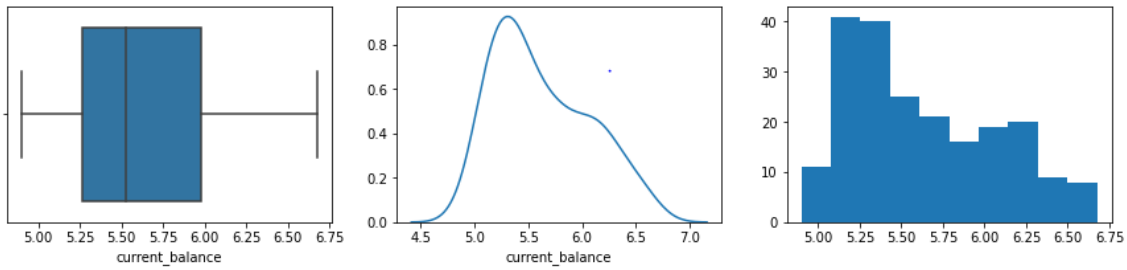

```
Column advance_payments is not normally distributed

 3. Univariate analysis for probability_of_full_payment

Mean is 0.870999, Median is 0.873450, Mode(s) are 0.8823
Column probability_of_full_payment has outliers
```



Column probability_of_full_payment is not normally distributed

4. Univariate analysis for current_balance

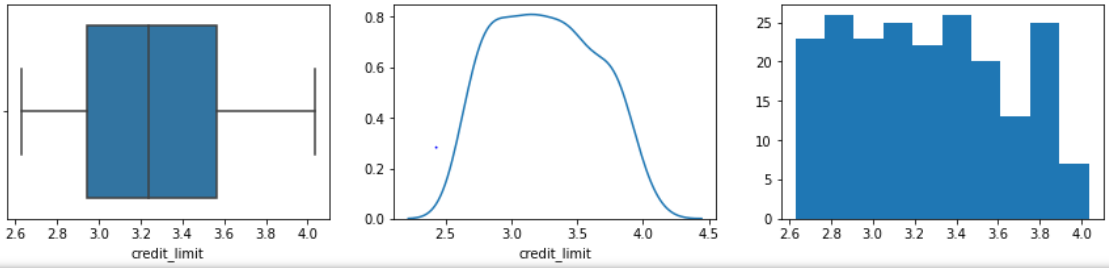
Mean is 5.628533, Median is 5.523500, Mode(s) are 5.2360
Column current_balance does not have outliers



Column current_balance is not normally distributed

5. Univariate analysis for credit_limit

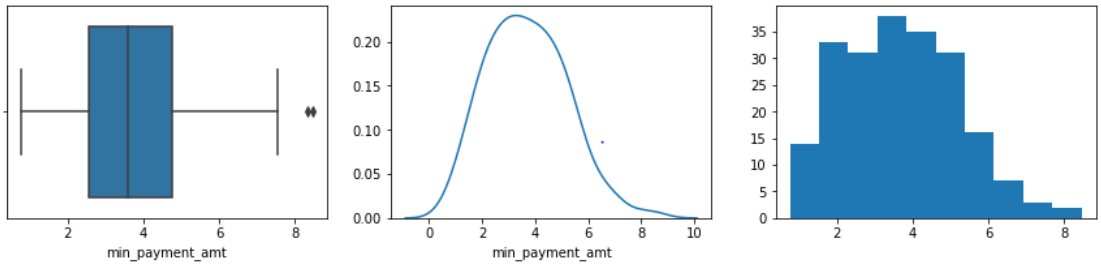
Mean is 3.258605, Median is 3.237000, Mode(s) are 3.0260
Column credit_limit does not have outliers



Column credit_limit is not normally distributed

6. Univariate analysis for min_payment_amt

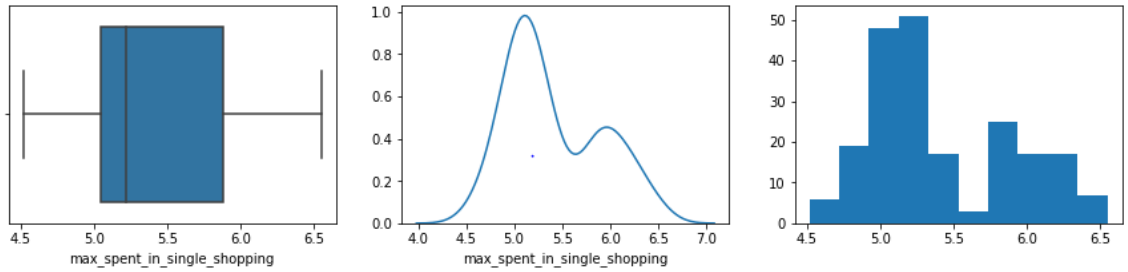
Mean is 3.700201, Median is 3.599000, Mode(s) are 2.1290
Column min_payment_amt has outliers



```
Column min_payment_amt is not normally distributed
```

```
 7. Univariate analysis for max_spent_in_single_shopping
```

```
Mean is 5.408071, Median is 5.223000, Mode(s) are 5.0010
Column max_spent_in_single_shopping does not have outliers
```




```
Column max_spent_in_single_shopping is not normally distributed
```

Correlation:

Positive correlations listed incrementally without overlaps among them as below:

- Spending has high positive correlation with max_spent_in_single_shopping, credit_limit, current_balance and advance_payments
- Advance_payments has high positive correlation with max_spent_in_single_shopping, credit_limit, current_balance
- Probability_of_full_payment has positive correlation with credit_limit
- Current_balance has high positive correlation with max_spent_in_single_shopping, credit_limit
- Credit_limit has positive correlation with max_spent_in_single_shopping

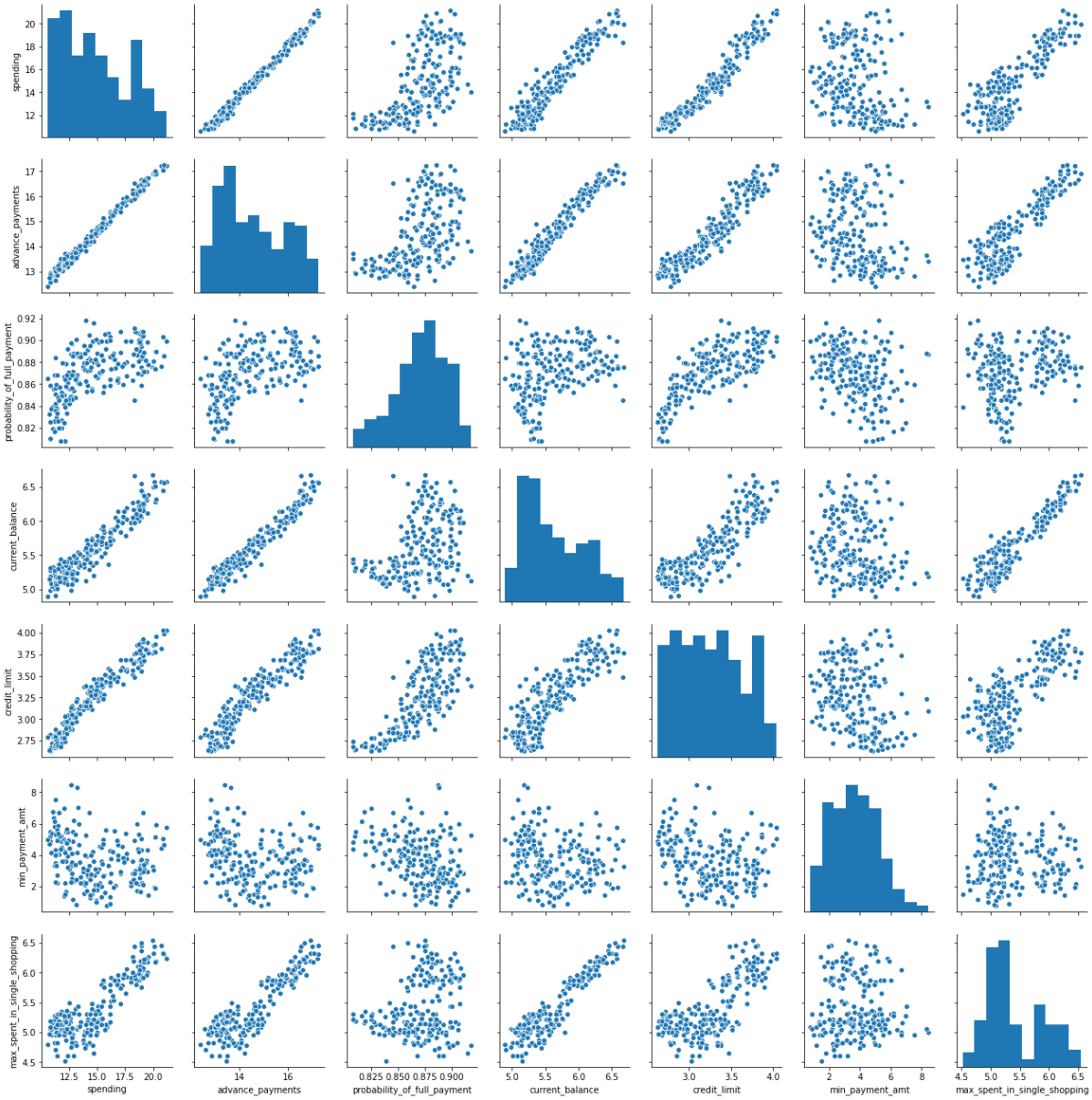
Negative correlation

- No negative correlation across variables has been observed.

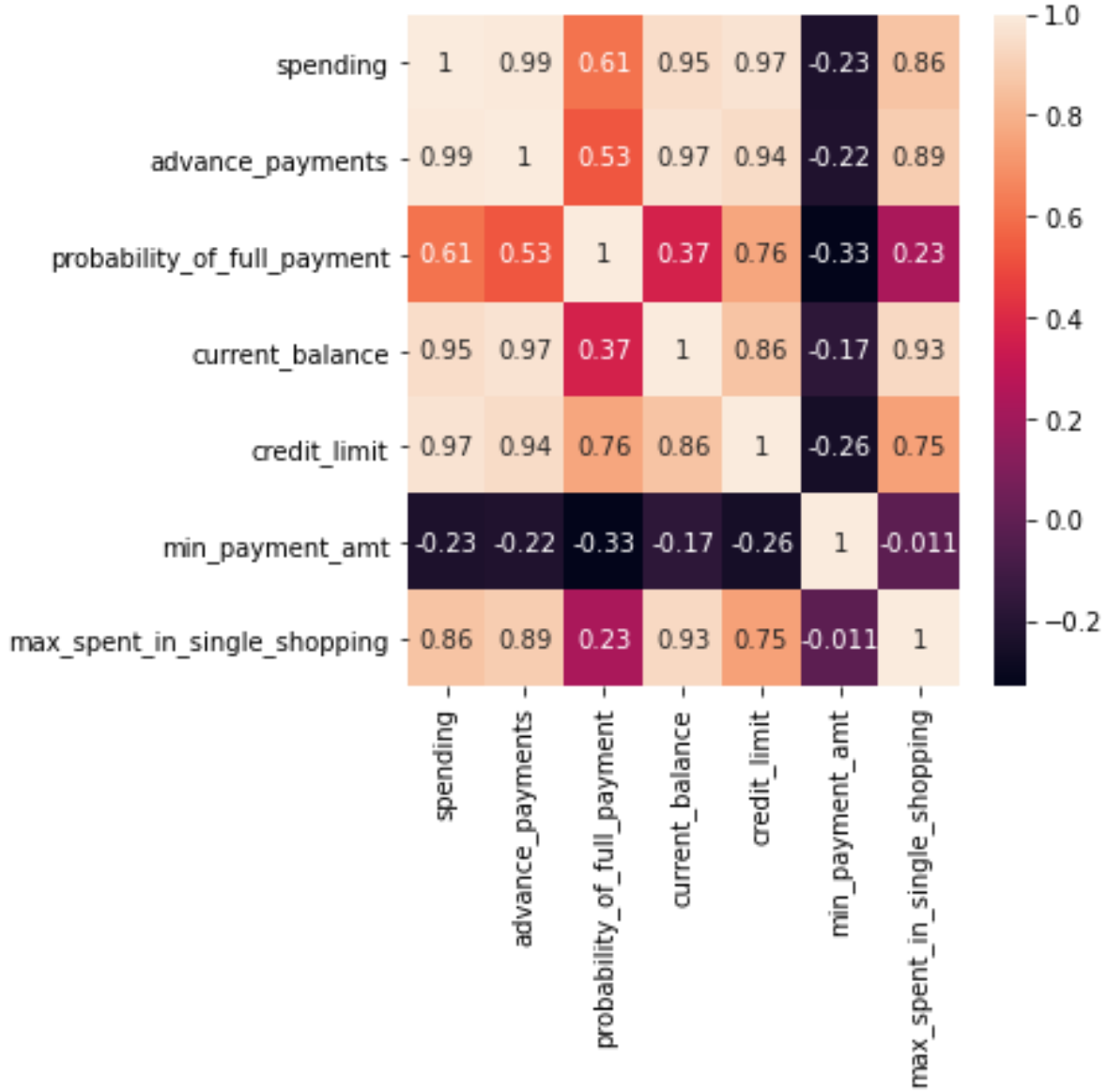
Note: Min_payment_amt does not show noticeable correlation with other attributes.

Please find below the pair plot and the related heat map depicting the above observations.

Pair plot:


spending
advance_payments
probability_of_full_payment
current_balance
credit_limit
min_payment_amt
max_spent_in_single_shopping

Heat map:



Question 1.2

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

Yes, scaling is necessary for clustering.

Scaling the dataset is one of the important data pre processing steps required for hierarchical clustering especially because of its distance based calculation approach during the processing of clustering the data.

Looking at the mean values across the variables in the given dataset, it is observed that there is a difference in magnitude of data across different columns. Hence given that there isn't significant variance among the data we could go for scaling the data through standardization.

Please find below the summary of statistics for the dataset depicting the mean and standard deviation (a derivative of variance) for the dataset.

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

Question 1.3

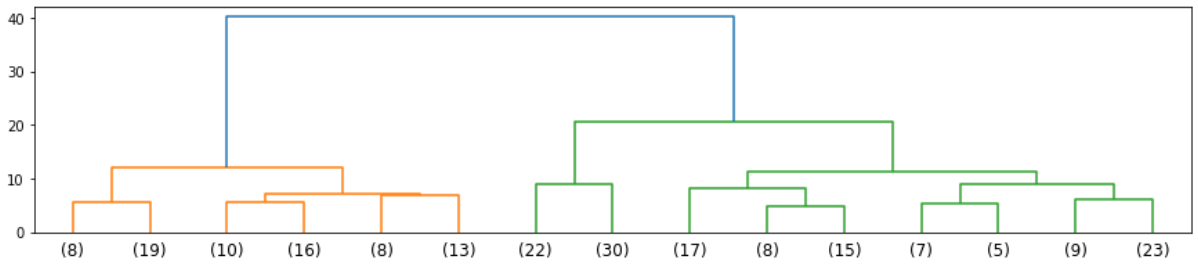
Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

Below is the dendrogram of the scaled dataset for top 15 clusters following ward linkage method. The dendrogram is depicting the hierarchical clustering outcome based on the agglomerative clustering approach.

However based on the below image we can clearly identify that the data makes two differentiated clusters as suggested by orange and green colours.

We could have pretty much created this dendrogram for top 10 clusters instead. However elongating it further to 15 clusters reassures the inference that the datasets remains clearly divided into two clusters only despite increasing the depth.

Hence 2 clusters will be optimal for the given dataset.



Please find below the 2 clusters created with the respective mean values across all the attributes.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
clusters							
1	18.37143	16.14543	0.8844	6.158171	3.684629	3.639157	6.017371
2	13.08557	13.76621	0.864298	5.363714	3.045593	3.730723	5.103421

Total count of records across clusters:

Cluster 1: 70 records
Cluster 2: 140 records.

Question 1.4

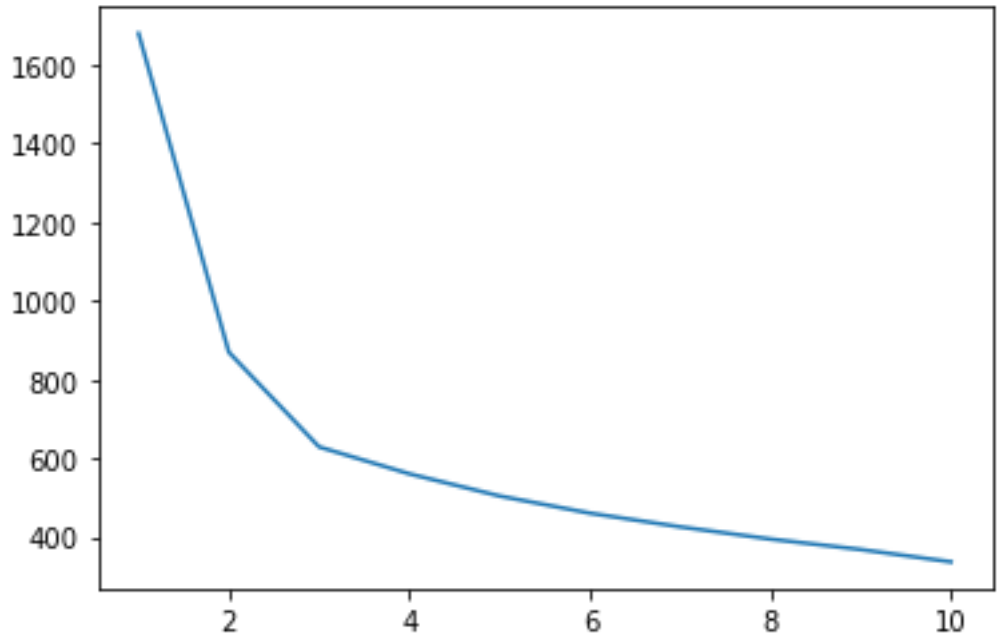
Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

By definition, K-means as name suggests refers to averaging the data which in other words is about finding the relevant centroid across K clusters by iterating the clustering technique for the given data set until we could end up with K clusters whose constituent data(s) are at the closest distance to their centroids when compared to other centroids adopting eucledian distance approach.

Determining the optimal cluster count: Once the clusters are created the model can be evaluated using WSS plot. WSS plot/distortion plot helps to know how many clusters are needed as output in K-means clustering. WSS plot stands for within sum of squares plot. also known as distortion plot or error plot. This operates with the notion that within sum of squares (variance) will be high for K=1 but when k begins to increase the sum of within sum of squares across clusters will start reducing from K=1. However as number of clusters increases (i.e incrementing K) there could be a point where the drop of WSS may not be significant.

Please find below the WSS plot for the given dataset for 1 to 10 clusters. The elbow suggests that the recommended number of plots could mostly be 2 and may be 3 clusters too. Exact clusters can be further recommended based on model evaluation subsequently as below.

WSS plot:



WSS (Within sum of squares) for K=1 to 10 as below

1 cluster	1470
2 clusters	659.1717545
3 clusters	430.6589732
4 clusters	371.2834477
5 clusters	326.4050987
6 clusters	289.576081
7 clusters	262.0073251
8 clusters	241.7067984
9 clusters	222.4015998
10 clusters	206.0987182

Model evaluation:

Hence to narrow down the most optimal number of clusters we could use techniques to validate if the mapping of observations to the clusters are valid or not. Silhouette scoring technique is one such. This includes computation of distance between each observations i.e distance between itself and every established centroid (i.e Silhouette width) to determine if there are any other centroids outside current cluster whose distance could be the shortest for the given observation. The average of such silhouette width will be referred to as silhouette score whose values would be within the range of -1 to +1. A positive score that yield a score closest to +1 will qualify the optimal number of clusters suggesting that the clusters are well seperated.

Please find below the silhouette scores for various clusters.

2 clusters	0.465772477
3 clusters	0.400727055
4 clusters	0.327574266
5 clusters	0.286214616
6 clusters	0.2878518
7 clusters	0.271660926
8 clusters	0.246766236
9 clusters	0.251625898
10 clusters	0.260362904
11 clusters	0.257961455

Above suggests creating 2 clusters on the given datasets will be the optimal one as it has the highest silhouette score suggesting well separated clusters making clear distinction them for the data it carries.

Please find below the clusters created using K means clustering (2 clusters) with respective mean values across all the attributes.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
clusters							
1	18.158571	16.05481	0.883817	6.127429	3.660519	3.480417	5.97174
0	12.930602	13.69346	0.863577	5.339699	3.025917	3.827444	5.081737

Question 1.5

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

Cluster 1 above indicates customers in high spending category end up paying in full with high probability. This segment also suggests customers probably are maximizing the card utilization given that the current balance is comparatively lesser while making their payment in full by month end.

Recommendation 1: So bank can consider offering balance transfer for these customers at a competitive interest rate to improve interest income as well as with the reduced risk of bad debts from this customer group who is expected to pay on time while having the spending ability.

Cluster 0 suggests that these customers probably are conservative spenders while enjoying staggered payments resulting in interest income to the bank.

Recommendation 2: So bank can consider offering 0% interest rate for new purchases towards increasing their spending ability while being able to increase the market share in terms of volume of transaction. This would also fetch transaction based commission income for the bank. However the risk of defaulting customers needs to be analyzed.

Question 2.1

Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

Please note that the object type categorical variables such as Agency_Code, Type, Claimed, Channel, Product_Name and Destination needs to be converted to integer values as a part of data pre processing and th and the Sales.

Distribution:
Following are the continuous variables in the dataset: Age, Commision, Duration and Sales.
None of the variables are normally distributed while age is somewhat observed to be closer to normal distribution.

Outliers:
Except Duration which has minimal outlier all other variables namely Age, Commision and Sales have good amount of outliers. All the outliers are beyond the maximum whisker.

Null Check:
There are no nulls across all the columns.

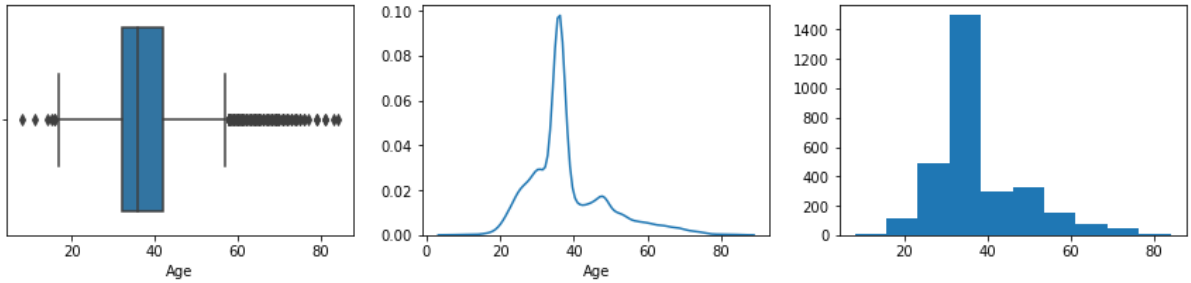
PFB the five number summary;

	Age	Commision	Duration	Sales
count	3000	3000	3000	3000
mean	38.091	14.529203	70.001333	60.24991
std	10.463518	25.481455	134.05331	70.73395
min	8	0	-1	0
25%	32	0	11	20
50%	36	4.63	26.5	33
75%	42	17.235	63	69
max	84	210.21	4580	539

The above summary suggests that across all the variables outliers are significantly affecting the mean given the extent of difference between median and the max especially the Duration column.

PFB plots from univariate analysis.

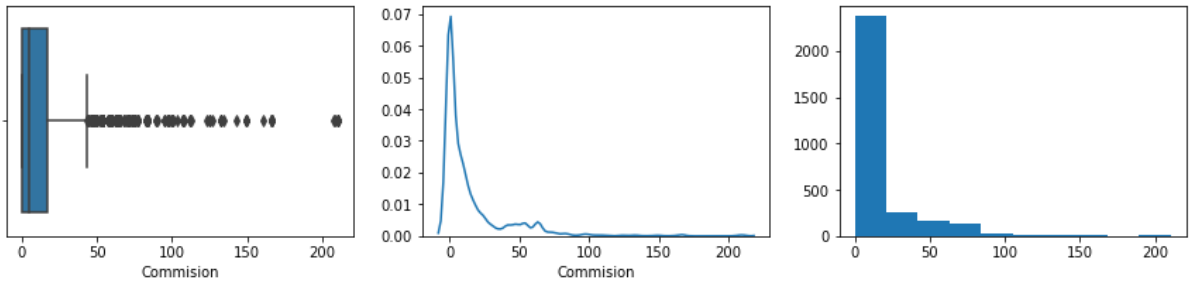
```
1. Univariate analysis for Age

Mean is 38.091000, Median is 36.000000, Mode(s) are 36.0000
Column Age has outliers
```



Column Age is not normally distributed

2. Univariate analysis for Commision

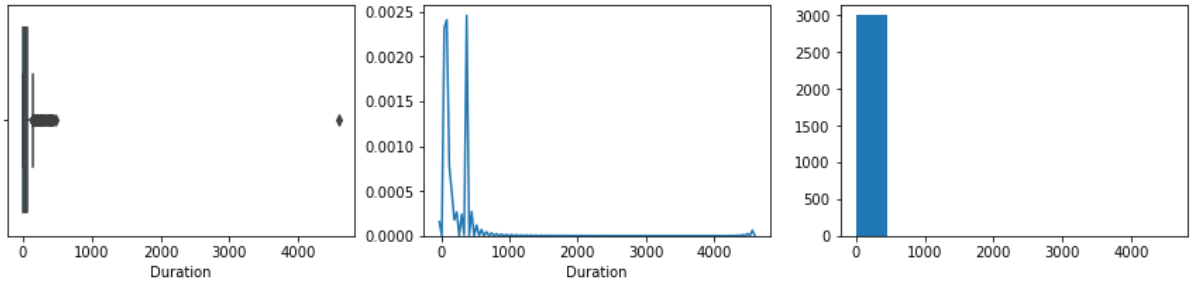
Mean is 14.529203, Median is 4.630000, Mode(s) are 0.0000
Column Commision has outliers



Column Commision is not normally distributed

3. Univariate analysis for Duration

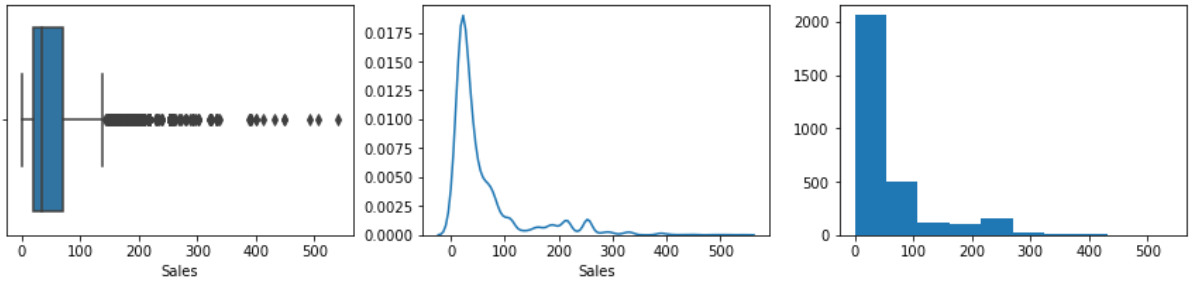
Mean is 70.001333, Median is 26.500000, Mode(s) are 8.0000
Column Duration has outliers



Column Duration is not normally distributed

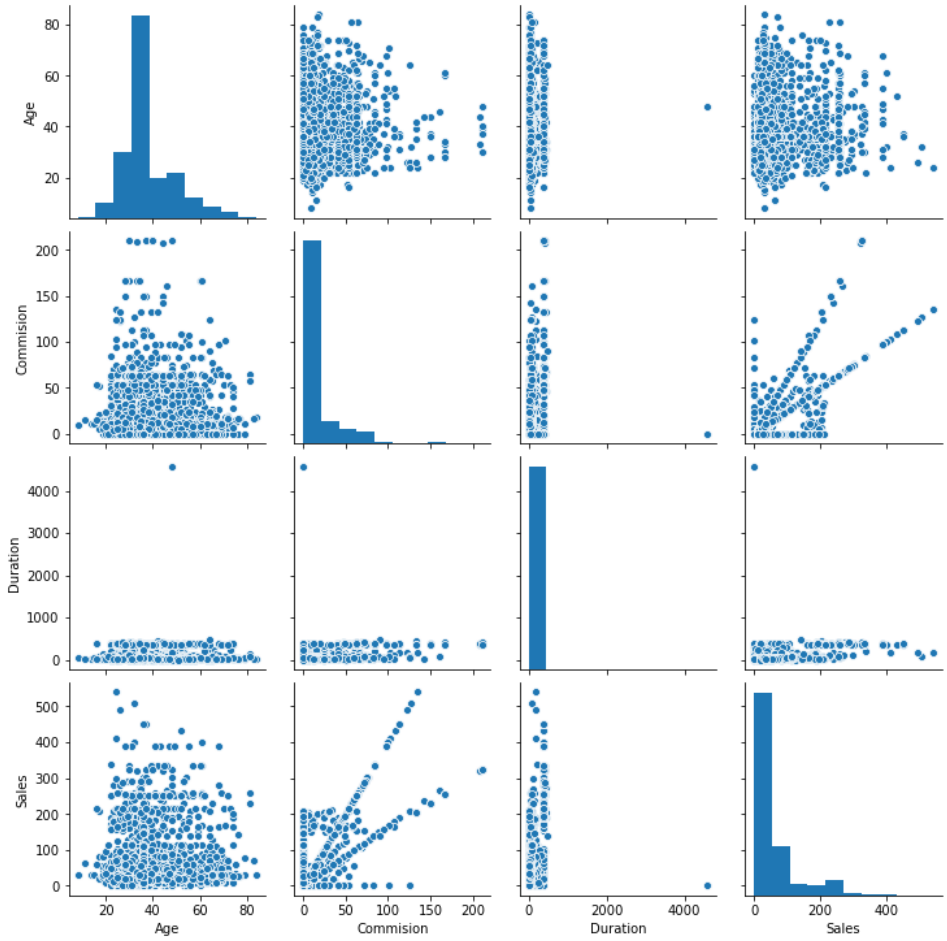
4. Univariate analysis for Sales

Mean is 60.249913, Median is 33.000000, Mode(s) are 20.0000
Column Sales has outliers

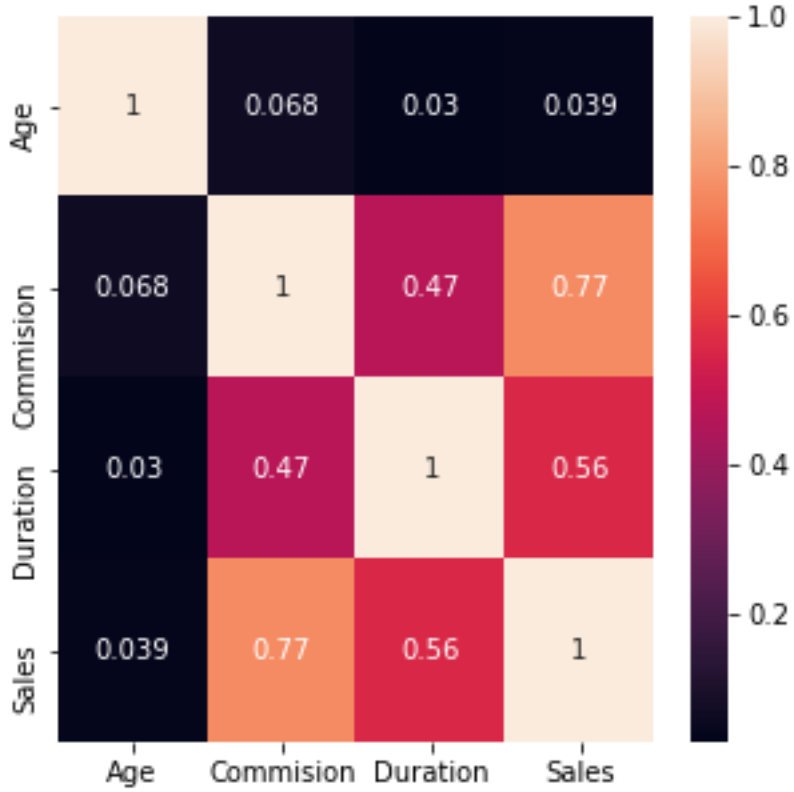


Column Sales is not normally distributed.

Pair plot as below:



Heat map:



Above pair plot and heatmap suggests that Commision is highly corelated with Sales.
The pair plot also suggests that some of the maximum sales and commission tend to decline along with age of the insured.

Apart from above statistics we also have following profiling of the categorical columns.

```
Value counts as below for each Feature:
Feature: Agency_Code
EPX    :1365
C2B    :924
CWT    :472
JZI    :239

Feature: Type
Travel Agency :1837
Airlines      :1163

Feature: Claimed
No     :2076
Yes    :924

Feature: Channel
Online     :2954
Offline    :46

Feature: Product Name
```

```
Customised Plan   :1136
Cancellation Plan :678
Bronze Plan       :650
Silver Plan       :427
Gold Plan         :109

Feature: Destination
ASIA        :2465
Americas    :320
EUROPE      :215
```

Question 2.2

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

Model building considerations:

Split the dataset into independent and dependent variables.

Following would be the independent variables on which we will be building all the 3 models (i.e CART, Random Forest, Artificial Neural Network)

- Age,
- Agency_Code
- Type,
- Commision,
- Channel,
- Duration,
- Sales,
- Product Name,
- Destination

Following would be the dependent/target variable.

- Claimed

All of the categorical/object type independent variables needs to be converted to numeric for any of the model building and hence following variables to be converted with numeric encoding to replace the textual values accordingly.

- Agency_Code
- Type
- Channel
- Product Name
- Destination

Neural networks perform better with scaled data and hence scaling to be applied for neural networks model building.

Divide the data into training and test data with 70:30 ratio. Follow same random state across all 3 models.

PFB the summary on the training data:

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
count	2100	2100	2100	2100	2100	2100	2100	2100	2100
mean	37.491429	1.276667	0.606667	11.60438	0.985714	45.58286	50.44487	1.689524	0.255714
std	8.962014	0.993609	0.488606	15.30746	0.118694	45.75714	42.74753	1.260226	0.582286
min	17	0	0	0	0	0	0	0	0
25%	32	0	0	0	1	11	20	1	0
50%	36	2	1	5	1	27	33	2	0
75%	42	2	1	17.82	1	64	69.3	2	0
max	57	3	1	43.0875	1	141	142.5	4	2

PFB the summary on the test data:

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
count	900	900	900	900	900	900	900	900	900
mean	37.764444	1.375556	0.625556	10.11378	0.982222	44.22556	48.3303	1.596667	0.236667
std	9.405349	0.992218	0.484248	14.30139	0.132216	44.73821	40.92977	1.253511	0.558671
min	17	0	0	0	0	-1	0	0	0
25%	32	0	0	0	1	11	20	1	0
50%	36	2	1	4	1	26	30	2	0
75%	42	2	1	13.2825	1	60.25	66.25	2	0
max	57	3	1	43.0875	1	141	142.5	4	2

Parameters to be considered for each model:

CART:

Splitting criteria: Gini (Decision tree uses gini index and gini gain to pick the best independent variable to split the data at each node)

Based on the decision tree visualization built using export_graphviz API it could be narrowed down that CART can be built with following parameters.

maximum depth : 5 - This parameter controls the depth of hierarchy from the root node at an optimal point beyond which it could result in over fitment of the model while training the data.

max_sample_leaf : 60 – This parameter control minimum observations require to create a new terminal or decision node.

max_samples_split : 180 – This parameter controls minimum observations from the dataset required to split a particular decision node/parent node into further nodes.

Based on the model build below are the feature importances:

Feature	Importance
Duration	0.240804
Sales	0.223446
Agency_Code	0.203293
Age	0.182107
Commision	0.07257
Product Name	0.043855

Destination	0.023909
Channel	0.007948
Type	0.002067

PFB the prediction on the training data set. Listed below are the counts by target class.

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
Claim_Prediction									
0	1095	1095	1095	1095	1095	1095	1095	1095	1095
1	368	368	368	368	368	368	368	368	368

PFB the prediction on the test data set. Listed below are the counts by target class.

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
Claim_Prediction									
0	216	216	216	216	216	216	216	216	216
1	59	59	59	59	59	59	59	59	59

Random Forest:

Based on the grid search below are the best parameters on which this model is built.

```
max_depth          : 7 – This parameter controls the depth of hierarchy from
the root node at an optimal point beyond which it could result in over
fitment of the model while training the data.
max_features       : 4 – This feature which is part of bootstrapping in
ensemble technique. Bootstrapping or bagging enables cross validation technique
to ensure we train the model for the unseen data.
min_samples_leaf   : 10 – This parameter control minimum observations require
to create a new terminal or decision node.
min_samples_split  : 30 – This parameter controls minimum observations from
the dataset required to split a particular decision node/parent node into
further nodes.
n_estimators       : 201 – This parameter controls how many decision trees
needs to be created as a part of random forest model.
```

Note: Feature importances for Random forest matches with the CART model.

PFB the prediction on the training data set. Listed below are the counts by target class.

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
Claim_Prediction									
0	1095	1095	1095	1095	1095	1095	1095	1095	1095
1	368	368	368	368	368	368	368	368	368

PFB the prediction on the test data set. Listed below are the counts by target class.

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
Claim_Prediction									
0	216	216	216	216	216	216	216	216	216
1	59	59	59	59	59	59	59	59	59

Neural network:

In order for neural network to perform better the data set can be normalized.

Based on the various trial and error approach consider some of the observations from grid search inclusive below were the most optimal hyper parameters considered for this model build.

- hidden_layer_sizes : (201,101,51) – Number of neurons in a single hidden layer. Please note three hidden layer considered in this model build post optimization exercise.
- max_iter :15 - number of epocs end to end across the layers between input and output. Also the iterations stopped even before reaching 15 at 13th iteration as the training loss reached global minima.
- activation :'relu' - It's a Mechanism by which the artificial neuron processes incoming information and passes it throughout the network.
- solver :'adam' – This parameter specifies the loss function used towards updating the weights during the iterations. This enables to progress along the parabolic curve by deriving error values optimally to reach the point of global minima.
- tol :0.1 – This parameter specifies tolerance for optimization which means that when the loss or score is not improving by at least the specified measure model shall be considered to be converged and hence the iterations can be stopped.

PFB the prediction on the training data set. Listed below are the counts by target class.

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
Claim_Prediction									
0	954	954	954	954	954	954	954	954	954
1	509	509	509	509	509	509	509	509	50

PFB the prediction on the testing data set. Listed below are the counts by target class.

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
Claim_Prediction									
0	192	192	192	192	192	192	192	192	192
1	83	83	83	83	83	83	83	83	83

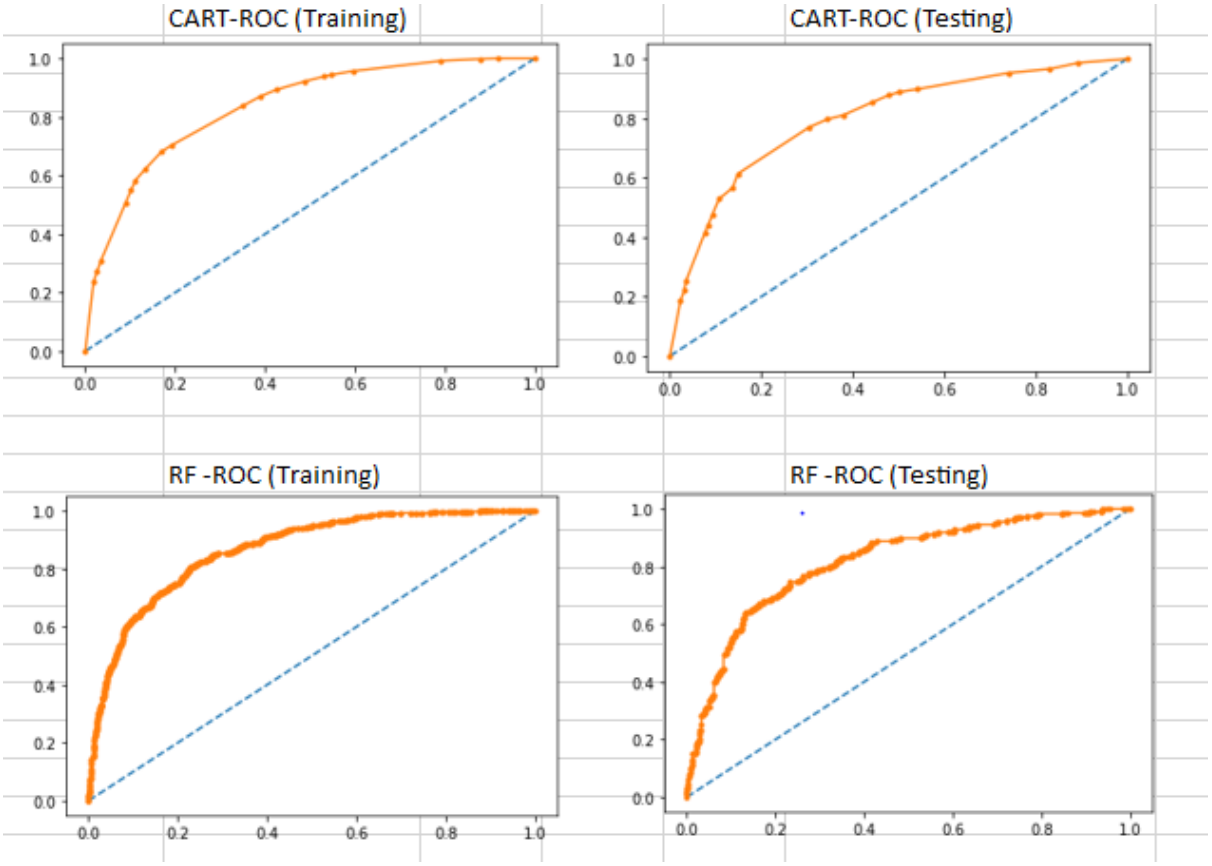
Question 2.3

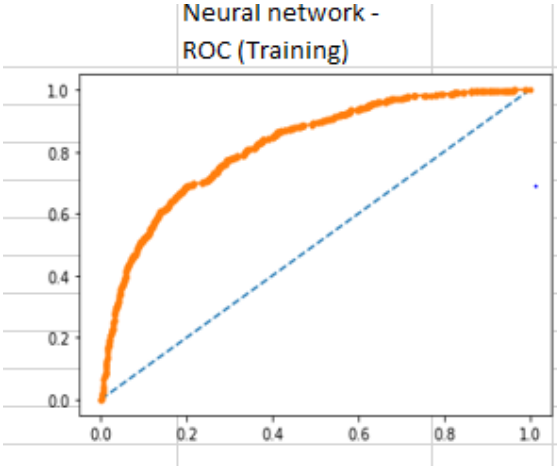
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model

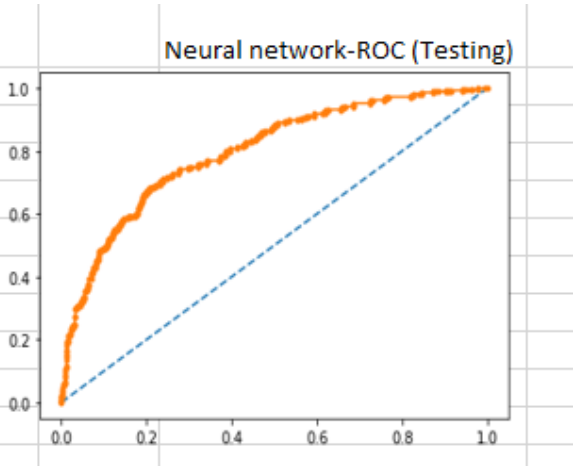
Model performance comparison:
AUC:

	CART	Random Forest	Neural network
AUC Training data	0.837	0.868	0.82
AUC Testing data	0.801	0.821	0.80

ROC:







Confusion matrix across all 3 models for training and testing data as below side by side:

CART Training		Predicted	
		Negative	Positive
actual	Negative	1307	164
	Positive	262	367

CART Testing		Predicted	
		Negative	Positive
actual	Negative	549	56
	Positive	155	140

Random Forest (Training)		Predicted	
		Negative	Positive
actual	Negative	1341	130
	Positive	301	328

Random Forest (Testing)		Predicted	
		Negative	Positive
actual	Negative	556	49
	Positive	177	118

Neural Network (Training)		Predicted	
		Negative	Positive
actual	Negative	58	1413
	Positive	66	328

Neural Network (Testing)		Predicted	
		Negative	Positive
actual	Negative	556	49
	Positive	177	118

Classification report across all models for training and testing data depicted below in next few tables.

CART (Training)	precision	recall	f1-score	support
0	0.83	0.89	0.86	1471
1	0.69	0.58	0.63	629
accuracy			0.8	2100
macro avg	0.76	0.74	0.75	2100
weighted avg	0.78	0.78	0.78	2100

CART (Testing)	precision	recall	f1-score	support
0	0.78	0.91	0.84	605
1	0.71	0.47	0.57	295
accuracy			0.77	900
macro avg	0.75	0.69	0.7	900
weighted avg	0.76	0.77	0.75	900

Random Forest (Training)	precision	recall	f1-score	support
0	0.85	0.91	0.88	1471
1	0.74	0.62	0.67	629
accuracy			0.82	2100
macro avg	0.79	0.76	0.77	2100
weighted avg	0.81	0.82	0.81	2100

Random Forest (Testing)	precision	recall	f1-score	support
0	0.79	0.92	0.85	605
1	0.74	0.49	0.59	295
accuracy			0.78	900
macro avg	0.77	0.71	0.72	900
weighted a	0.77	0.78	0.76	900

Neural network (Training)	precision	recall	f1-score	support
0	0.85	0.81	0.83	1471
1	0.6	0.68	0.64	629
accuracy			0.77	2100
macro avg	0.73	0.74	0.73	2100
weighted avg	0.78	0.77	0.77	2100

Neural network (Testing)	precision	recall	f1-score	support
0	0.81	0.84	0.83	605
1	0.65	0.59	0.62	295
accuracy			0.76	900
macro avg	0.73	0.72	0.72	900
weighted a	0.76	0.76	0.76	900

Out of all the 3 models Random forest has resulted in better accuracy but other two models have come closer to that with minor difference falling below the Random forest model. It can also be noted that given both classes are equally important apart from accuracy even the weighted average has performed marginally better in the random forest and hence we could narrow down on this model for productionizing the model.

Question 2.4

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

Out of all the 3 models Random forest has resulted in better accuracy but other two models have come closer to that with minor difference falling below the Random forest model. It can also be noted that given both classes are equally important apart from accuracy even the weighted average has performed marginally better in the random forest and hence we could narrow down on this model for productionizing the model.

Why go by accuracy?
Given the fact that predicting both classes of customer's who will end up claiming and those who will not claim correctly are important towards being able to come up with appropriate recommendation, accuracy becomes an important factor here. This way the model that best controls both type 1 and type 2 error would enable efficient overall classification from prediction perspective.

Accordingly, Random forest model fairs better comparatively here with 78% accuracy on the test data.

Question 2.5

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

Based on the feature importances as below it is observed that “Duration” plays the highest role among the variables along with other variables such as Sales, Agency Code and the Age. Hence the classification shall be analyzed towards the profit and loss of current packages accordingly to restrategize it for maximum benefit for the customers and insurance firm.

Feature	Importance
Duration	0.240804
Sales	0.223446
Agency_Code	0.203293
Age	0.182107
Commision	0.07257
Product Name	0.043855
Destination	0.023909
Channel	0.007948
Type	0.002067

Exploring the data further based on the Random forest outcome, top two Duration that yields to higher claim Include 141 and 2. Please find below the part of the count table in descending order.

Claim_Prediction	Duration	Count
1	141	8
	2	5
	3	3
	4	3
	5	3
	6	3
	7	3
	12	3
	22	3
	32	3
	8	2
	9	2

Firm can evaluate the kind of service providers for the tour or the nature of the tour to assess these segments to re evaluate the clauses for underwriting or revise the package configuration to make it an reasonably affordable package for a win win situation with the customer. If required some of these segments such as Duration 2 which has almost 83% of the insurance being claimed may not be a sustainable model at all and may need a revisit to reconfigure the plan to minimize risk while retaining benefit proposition to the customer.