PGP-DSBA

DATA MINING PROJECT REPORT

JULY 2020 // PREPARED BY JOTINDER SINGH MATTA

QUESTION - 1

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.

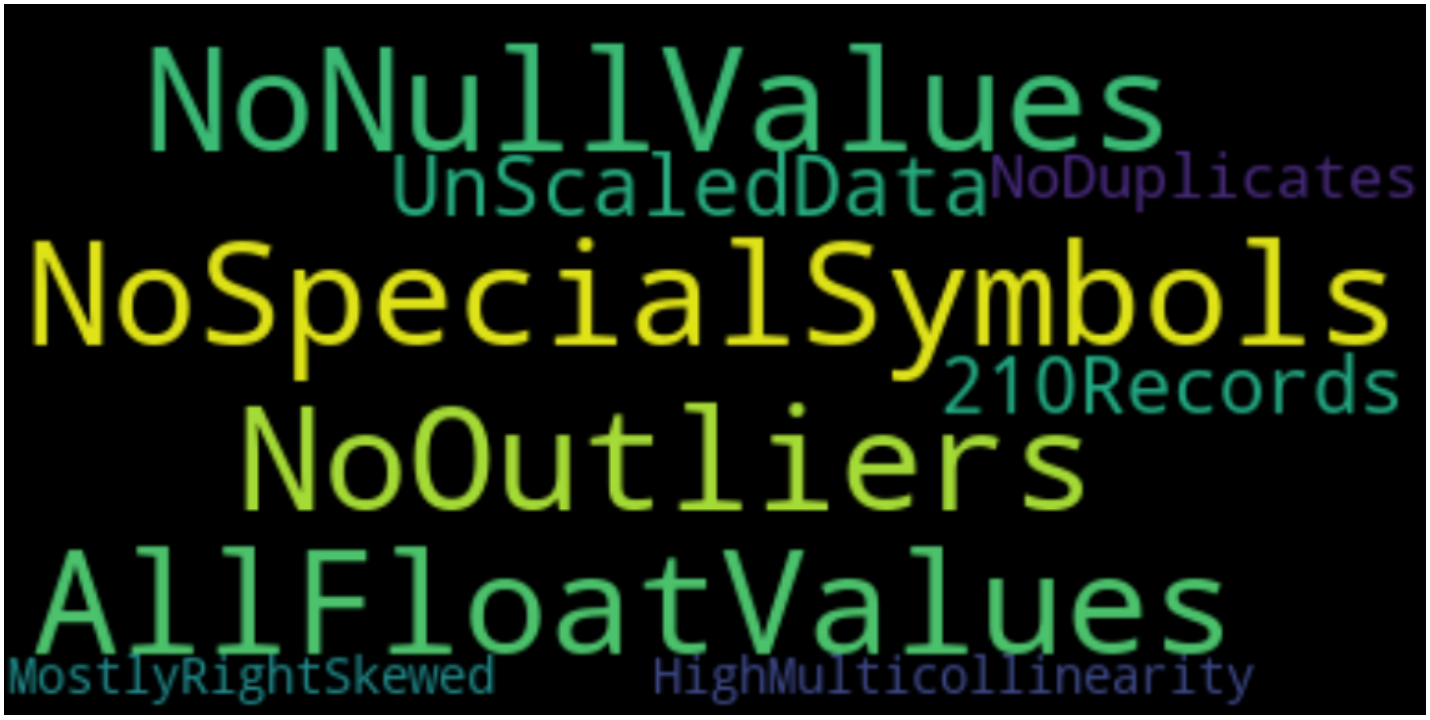
Data Dictionary for Market Segmentation:

1. **spending**: Amount spent by the customer per month (in 1000s)
2. **advance_payments**: Amount paid by the customer in advance by cash (in 100s)
3. **probability_of_full_payment**: Probability of payment done in full by the customer to the bank.
4. **current_balance**: Balance amount left in the account to make purchases (in 1000s)
5. **credit_limit**: Limit of the amount in credit card (10000s)
6. **min_payment_amt** : minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. **max_spent_in_single_shopping**: Maximum amount spent in one purchase (in 1000s)

Question 1.1

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

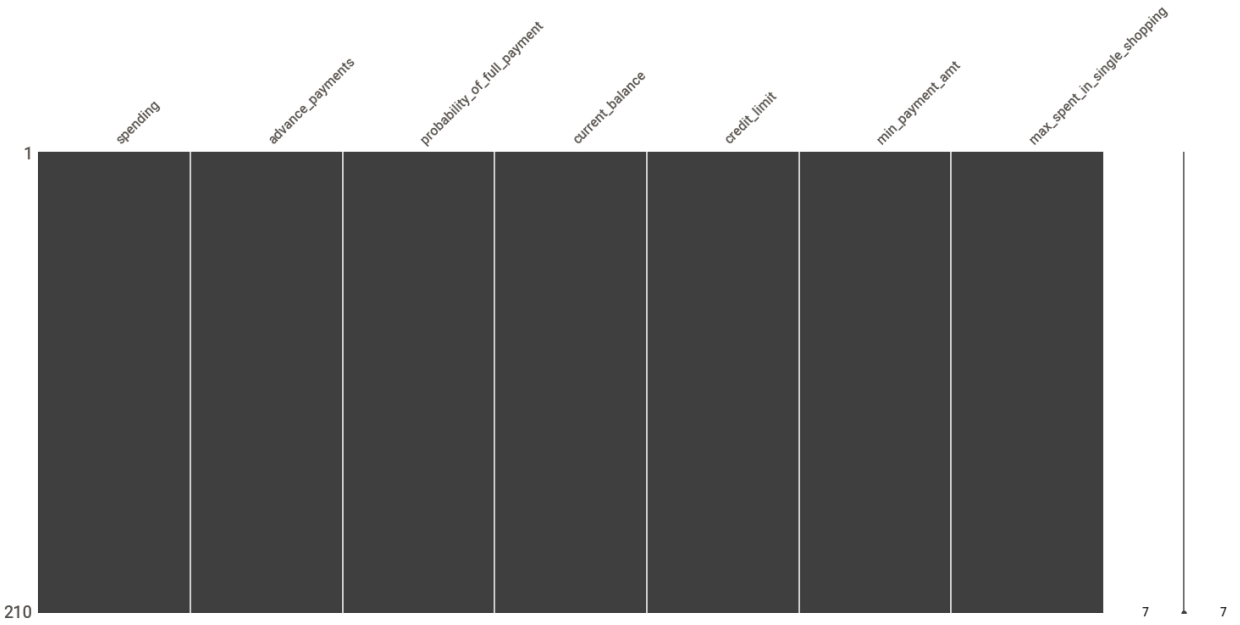
Bank credit card usage data set for 210 users has been shared with us. Below are some of the salient features of this data set.



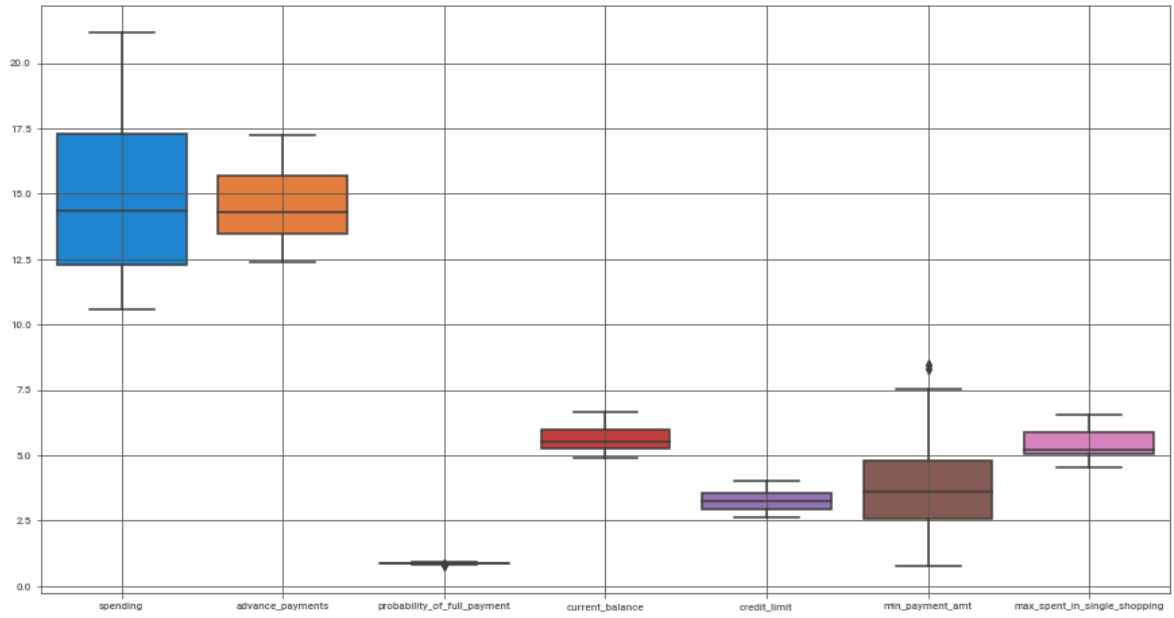
No Outliers are present with the exception of 2 outlier (1.5 times the IQR) values in "min_payment_amt" column, hence it **did not** warrant outlier treatment.

Mostly right skewed, with an exception of "probability_of_full_payment", which is left skewed.

No Null Values



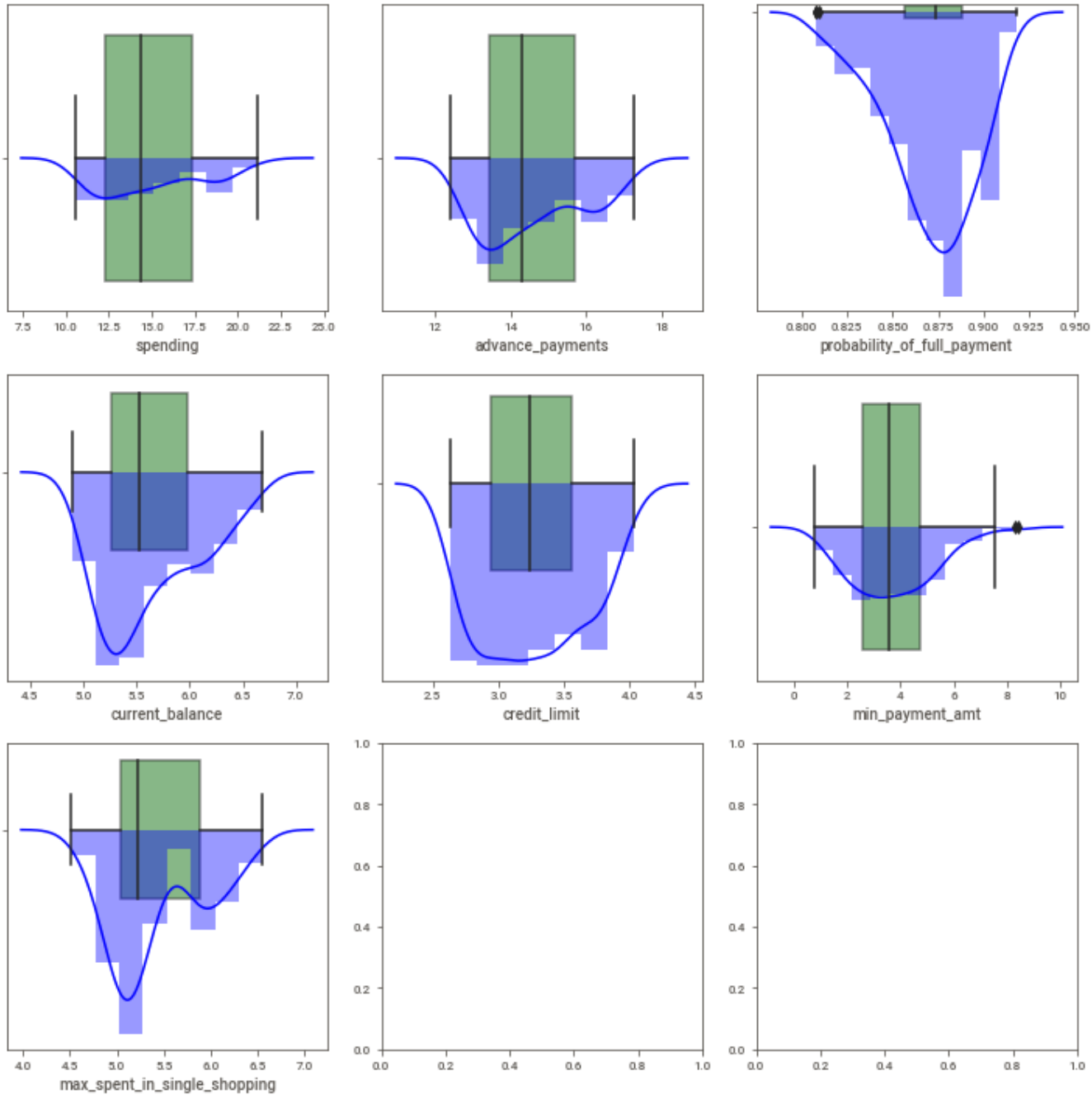
No Outliers



5 Point Summary

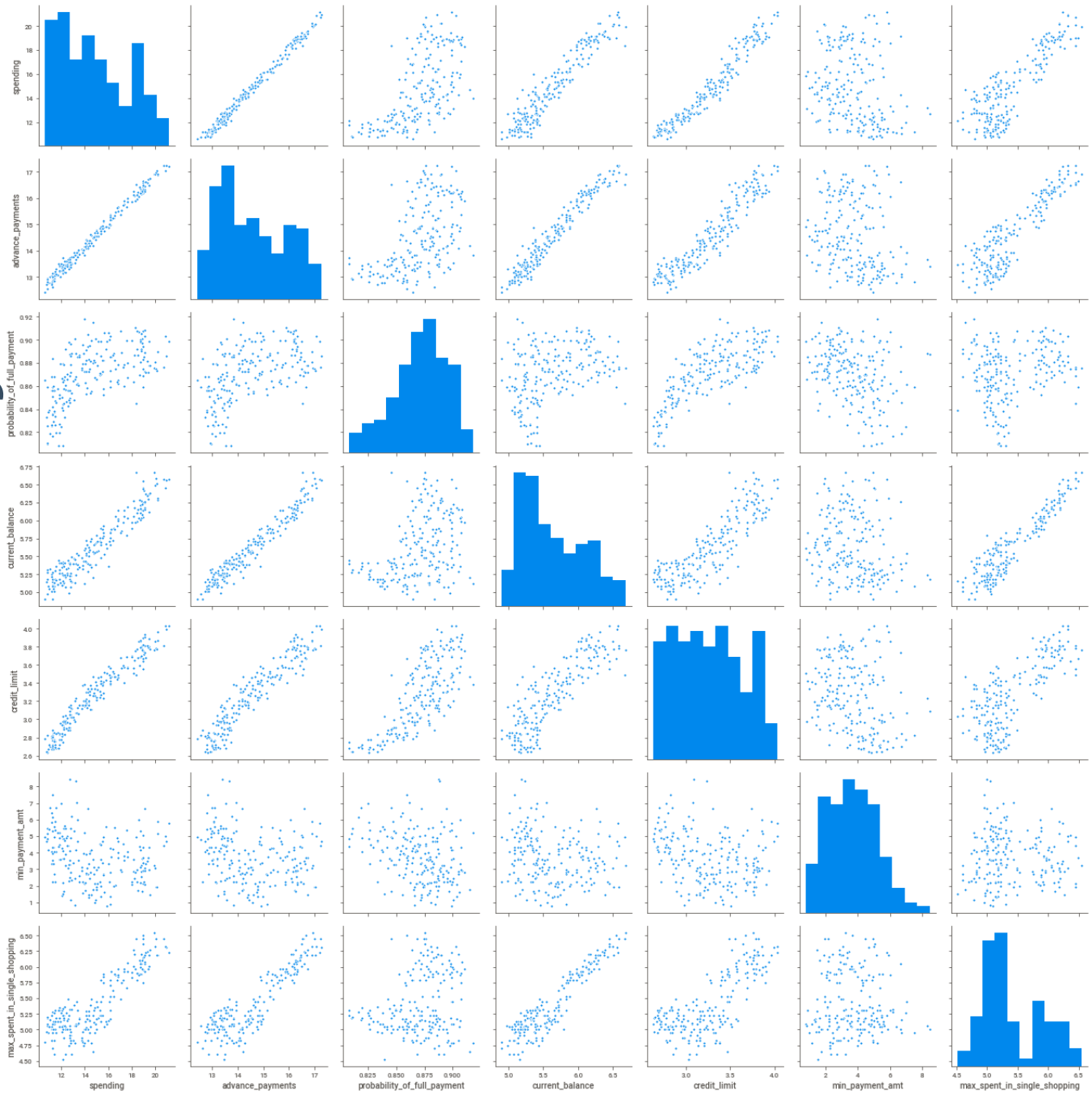
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	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

UniVariate Analysis



Mostly right skewed, with an exception of "probability_of_full_payment", which is left skewed.

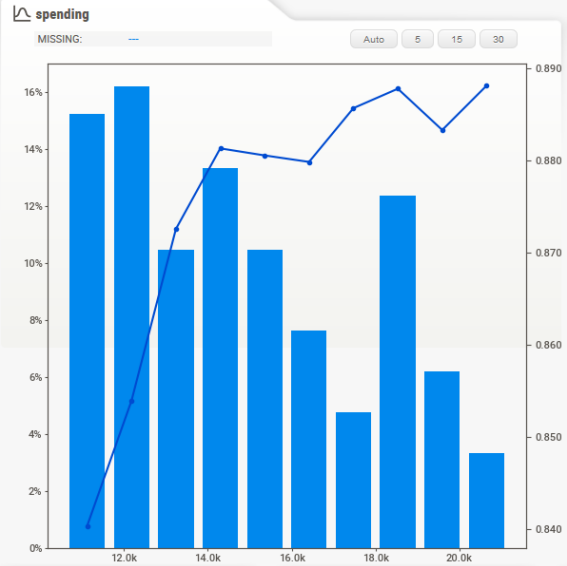
BiVariate Analysis

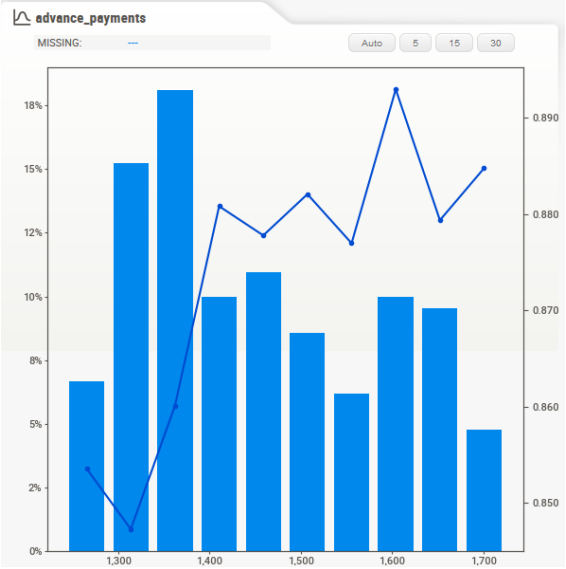


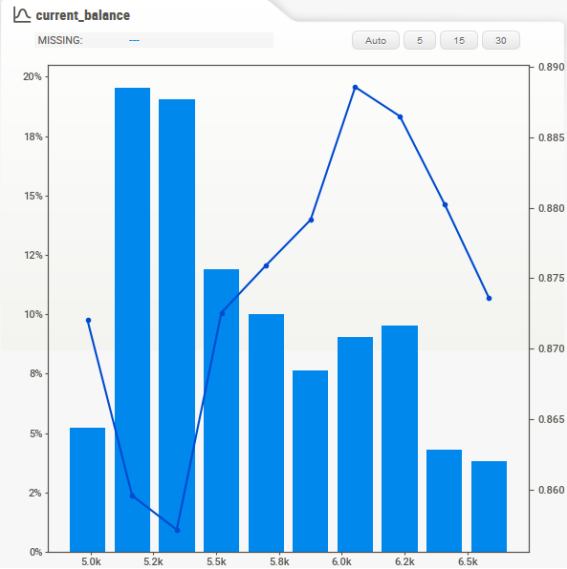
Multiple linear relationships are observed using the above bivariate analysis using matplotlib's pairplot.

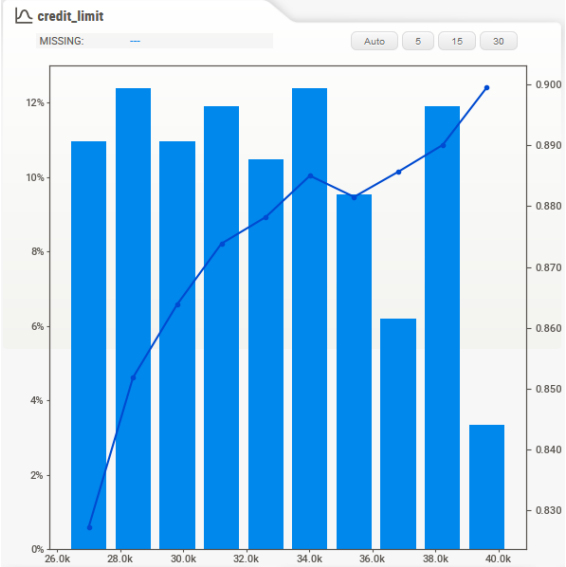
E.g. credit_limit & spending, advance_payments & current_balance etc.

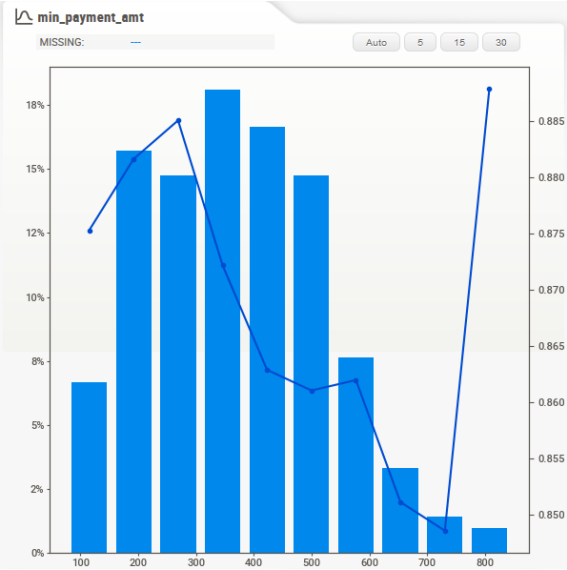
BiVariate Analysis - probability_of_full_payment

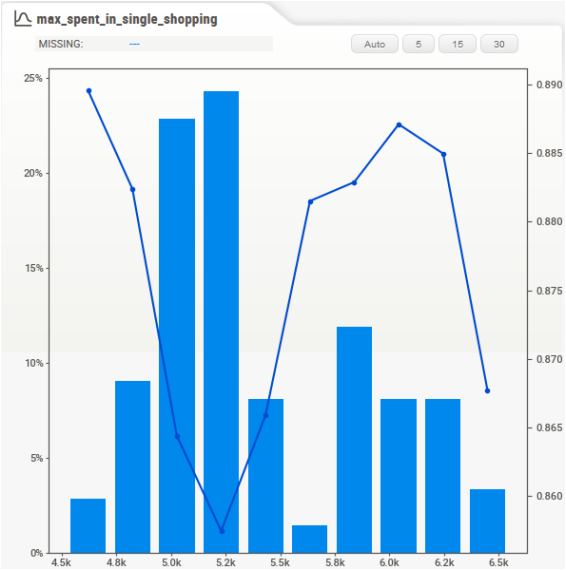












BiVariate Analysis - probability_of_full_payment

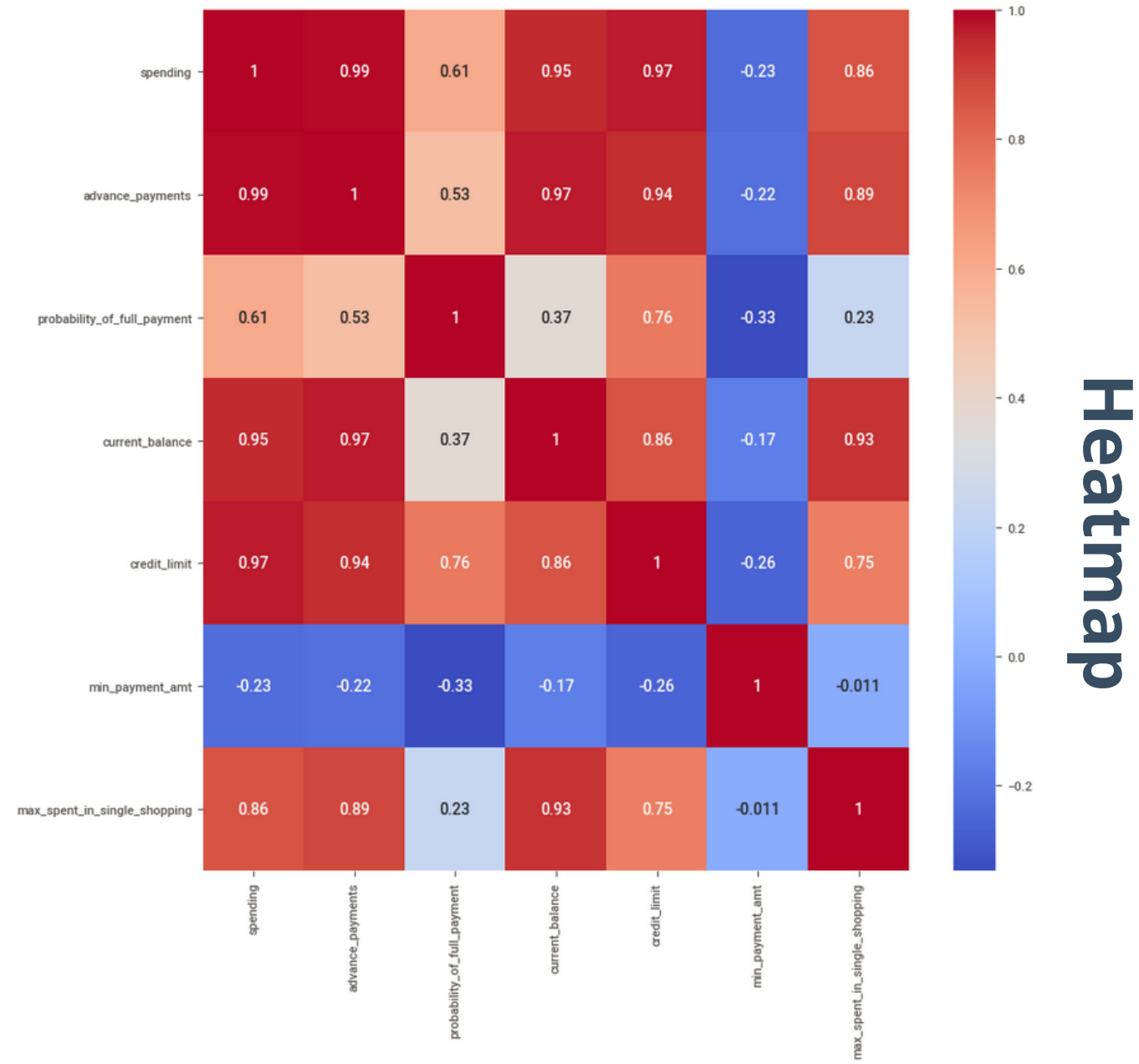
Doing BiVariate analysis of "probability_of_full_payment" with all the other columns, reveals some very interesting facts about the data set.

1. **spending:** As the customer spending increases, so does the probability of full payment.
2. **advance_payments:** As the amount of advance payments by the a customer increases, so does the probability of full payment.
3. **current_balance:** The probability of full payment increases with the current balance before reaching its peak at approx. 6K and then declines sharply.
4. **credit_limit:** Customers having higher credit limit had higher probability of full payment.
5. **min_payment_amt:** On an average as the minimum payment amount increases, probability of full payment decreases, with an exception of probability at minimum payment amount of 800.
6. **max_spent_in_single_shopping:** No clear trend was observed here.

Various important factors like **spending** & **credit_limit** were observed which had very high co-relation to the **probability_of_full_payment.**

We could see the same co-relation in the pairplot earlier as well.

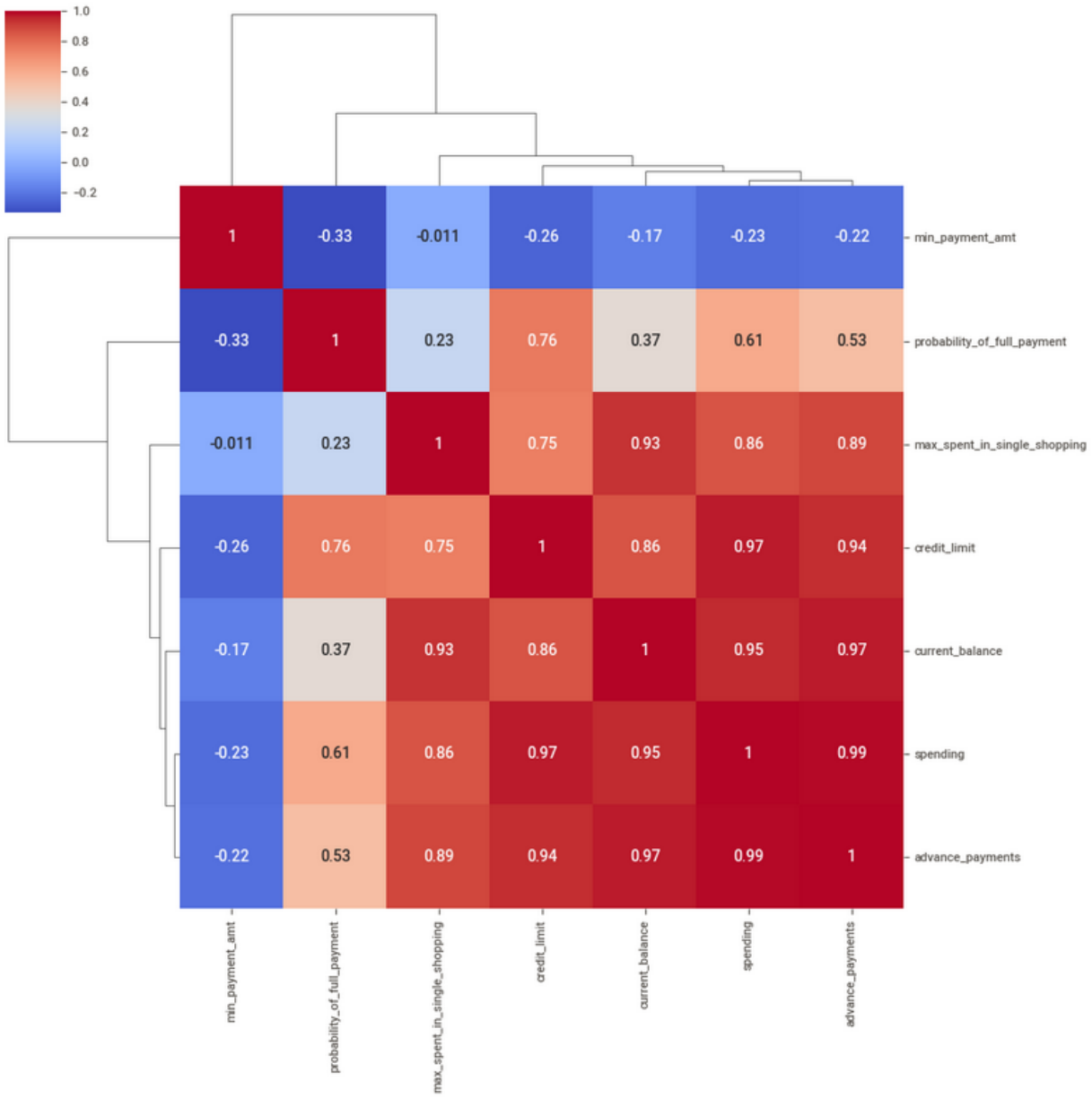
Also the same would be clearer when we look at the clustermap next, which clearly groups together the variables with high co-relation.



Heatmap

Strong co-relations are represented by red/ orange color boxes, while the weak co-relations are represented by blueish color boxes.

ClusterMap



Looking at the cluster map above, we can see high co-relations amongst below variables, represented by 5x5 red/ orange colored squared on the bottom right of the above cluster map.

1. advance_payments
2. spending
3. current_balance
4. credit_limit
5. max_spent_in_single_shopping

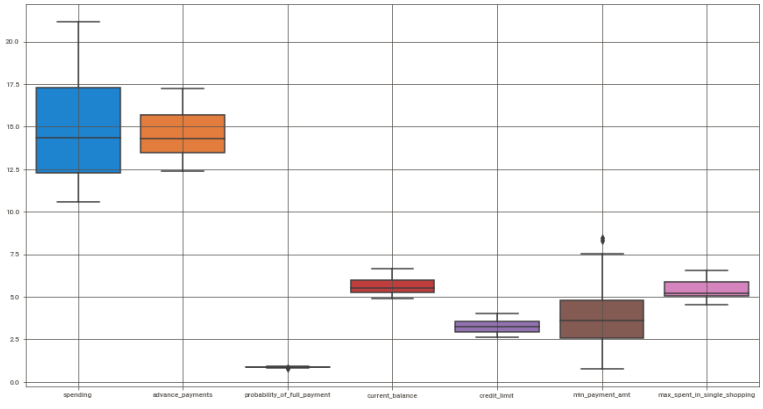
Question 1.2

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

Yes, scaling is required for this data set as the original data given to us is not scaled properly.

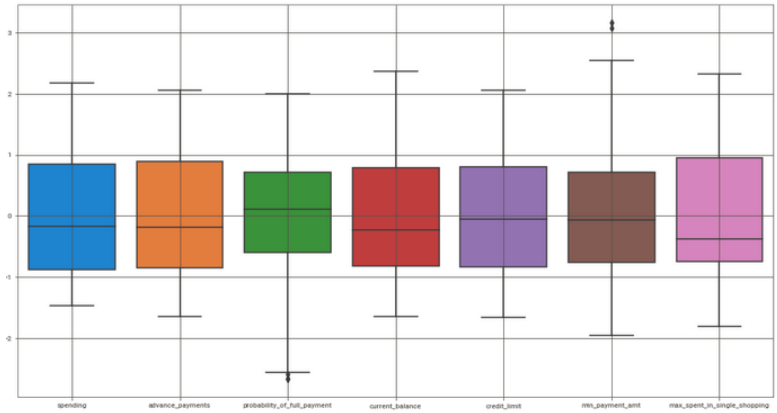
Although the data has been scaled, however it has not been scaled as per a standard scale across different columns. We can see some of the columns are scaled to the hundreds (100s), some are scaled to thousands (1000s) while some other are scaled to ten thousands (10000s).

Hence it becomes very important to first convert these columns into the absolute values and then apply standard scaling hence moving all the values to a scale of -1 to +1.



Before Scaling

After Scaling



Question 1.3

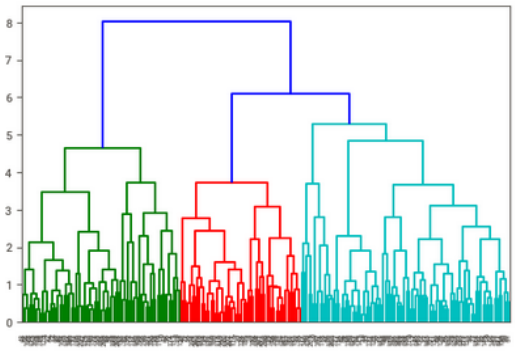
Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

We applied hierarchical clustering on the scaled data using 3 different linkage methods. We used following linkage methods and compared the clustering patterns for all of them.

Linkage methods used:-

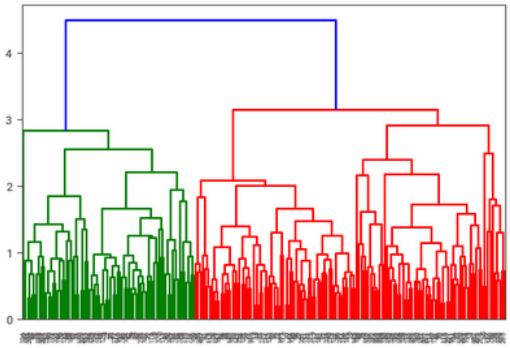
1. Complete Linkage
2. Average Linkage
3. Ward Linkage

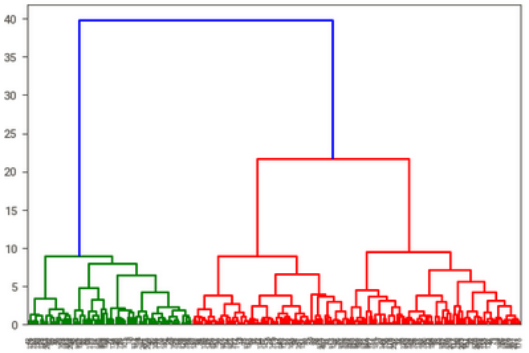
Below are the dendograms obtained for each of these methods.



Complete Linkage

Average Linkage

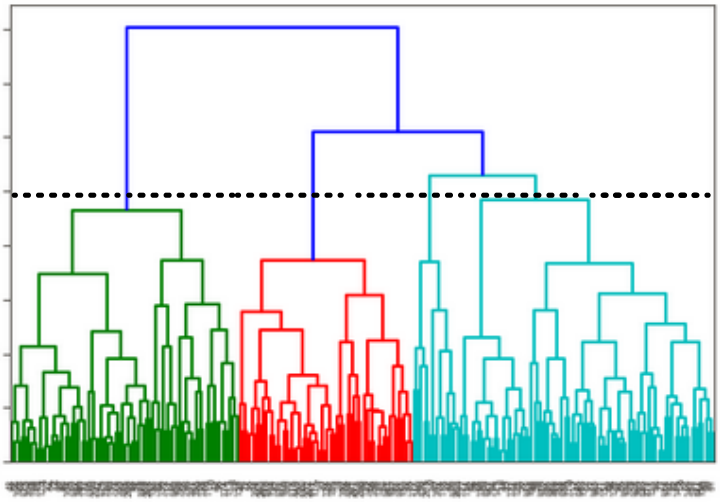




Ward Linkage

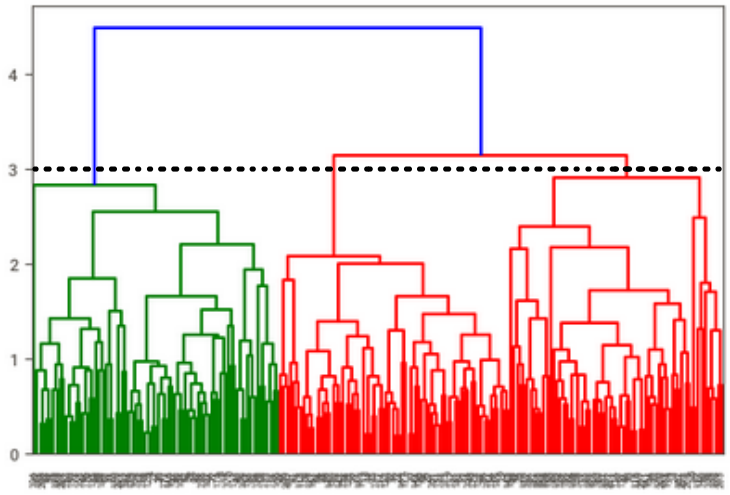
Finding optimum clusters - using dendogram

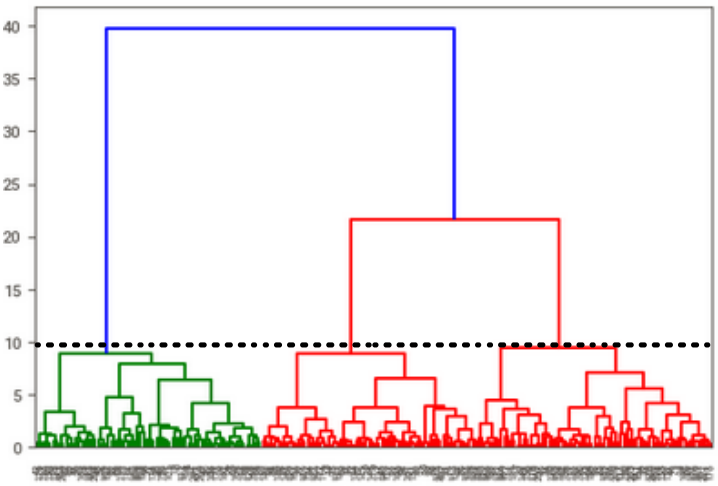
We look for the largest euclidean distance that we can vertically without crossing any horizontal line to determine the optimal number of clusters. Please note this is subjective and we might have to change the criteria based on the scenario, hence this is not a thumb rule per se. Hence in addition to the above we also take into account the visual cues that we can get just by looking at the dendogram.



Complete Linkage
4 Clusters

Average Linkage
3 Clusters

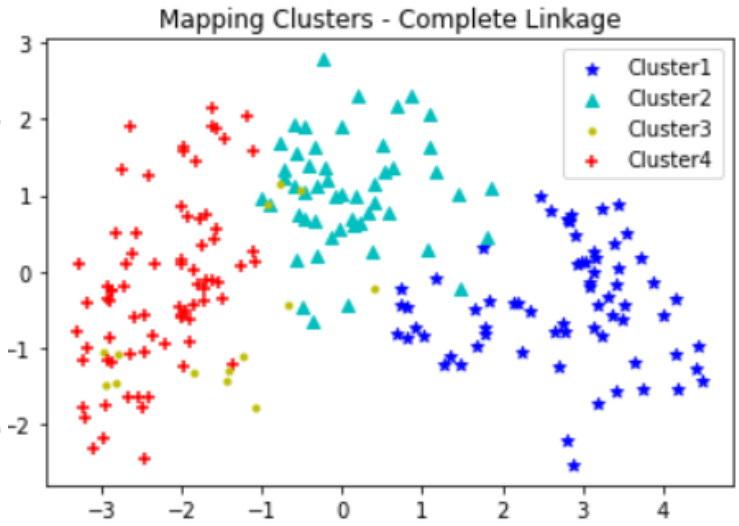




Ward Linkage
3 Clusters

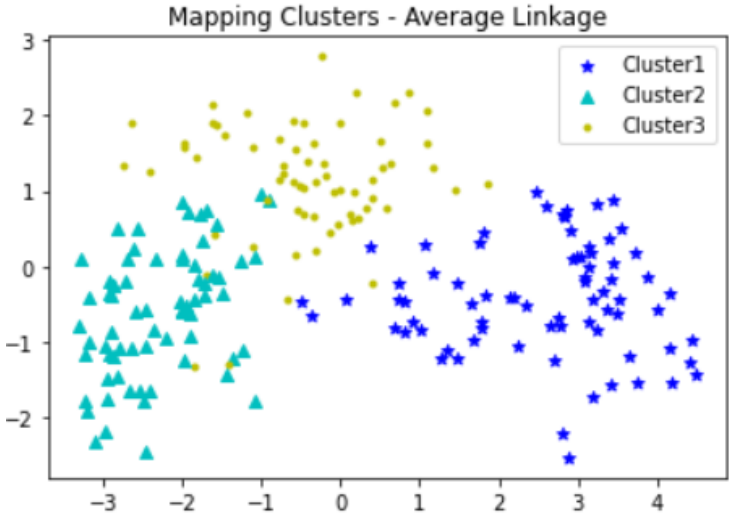
Now we visualize the clustered data for each linkage method, using the number of clusters determined above. We converted the data set into 2D using PCA and then mapped the different clusters using color codes.

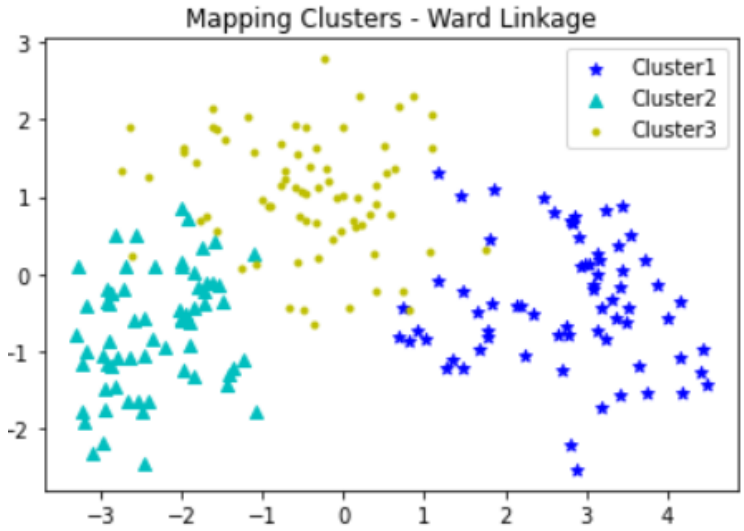
Cluster Visualization - using scatterplot



Complete Linkage
4 Clusters

Average Linkage
3 Clusters





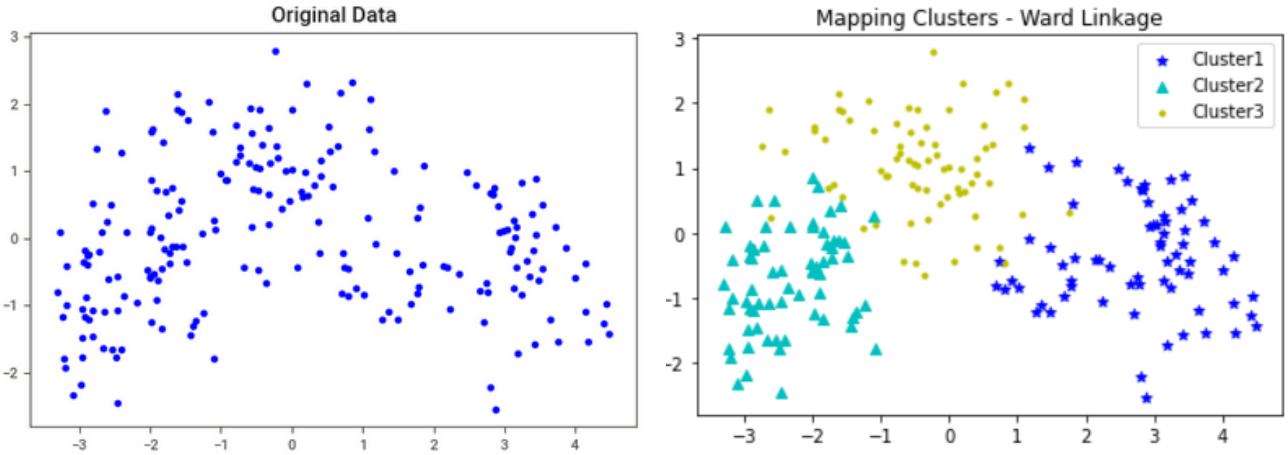
Ward Linkage
3 Clusters

Conclusion

Looking at the above scatter plots, we can clearly see the 3 clusters we get using the ward linkage method, depicts the most clear picture. The clusters are not overlapping and are distinctively different from each other when we use ward linkage method.

Other methods also provided with good clustering, however ward method was tad bit more efficient than the other linkage methods tried above.

Hence we conclude by saying that ward linkage provides us with 3 optimal clusters in order to do customer segmentation on the banking data set provided to us.



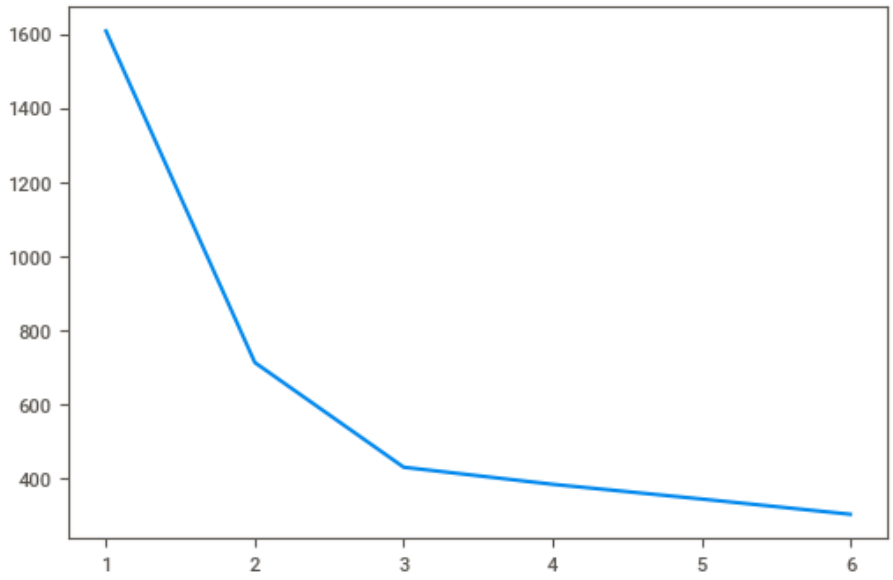
Question 1.4

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

We applied K-means clustering algorithm on our banking data set, starting with 1 cluster going up to 6 clusters. we then evaluated these different clusters using elbow curve and silhouette score methods.

We first calculated the WSS or Within Sum of Squares for different number of clusters, which were stored in a list and then plotted to form the elbow curve below.

Elbow Curve



Looking at the elbow curve plot above, we could see the Within Sum of Squares or the WSS decreased very rapidly from 1 to 3, however after 3 the decrease in WSS was not that significant.

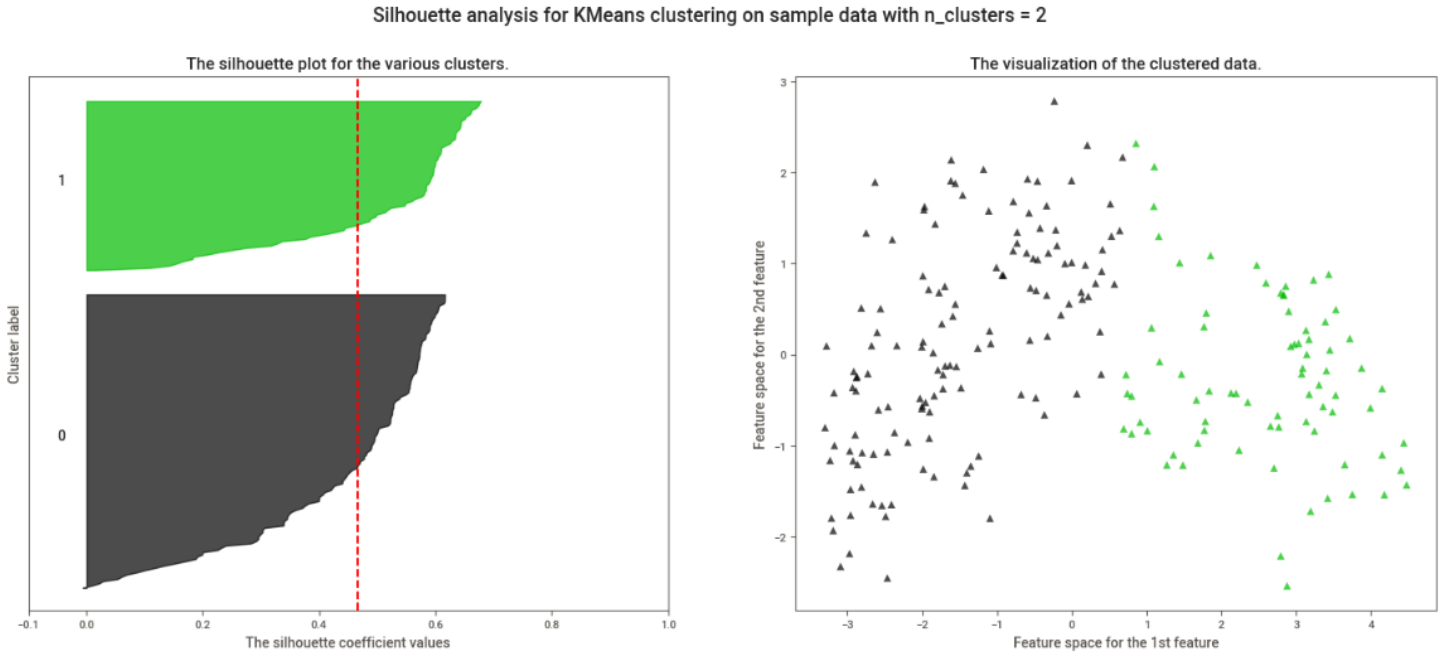
Hence we could conclude from this graph that increasing the clusters beyond 3 will not be beneficial to our cause of having well separated and meaningful clusters.

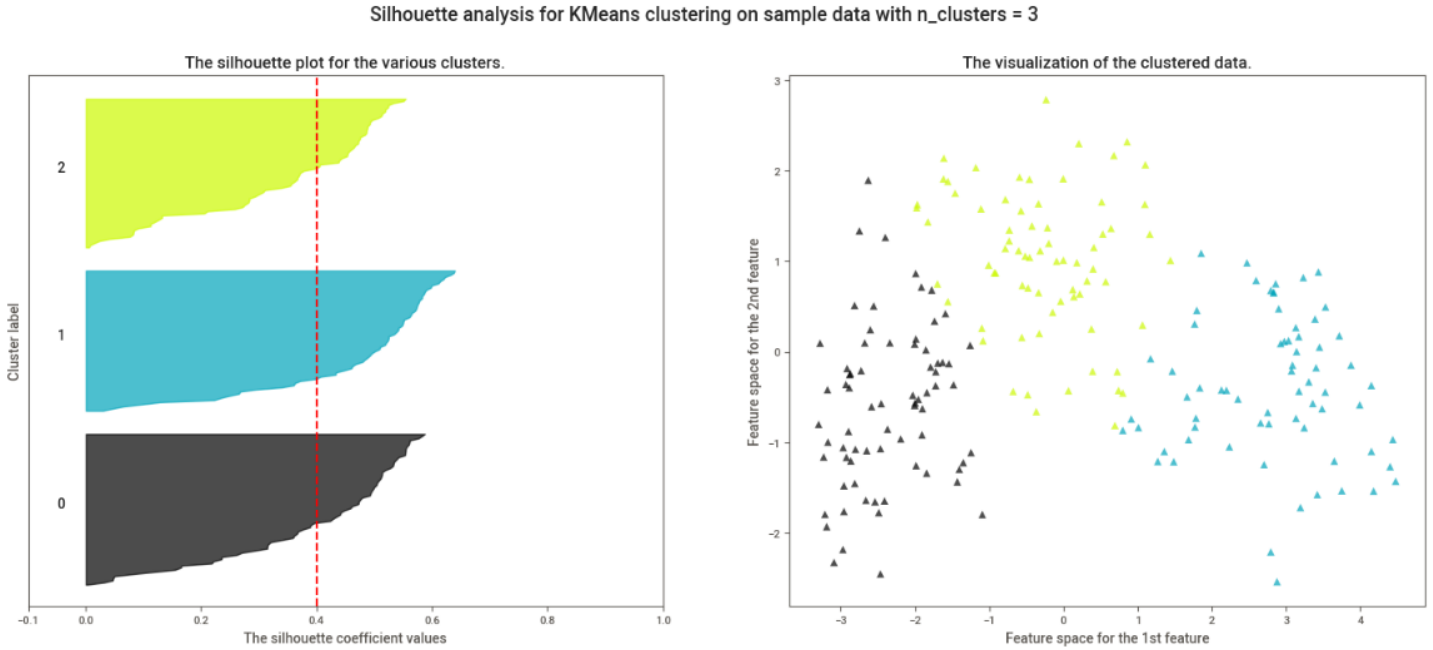
Next we used the Average Silhouette Score method to evaluate the number of clusters that would be optimal for our data set.

We potted the silhouette score for individual clusters as well as their average score (red dotted line) on the left and the scatter plot of 2d PCA representation of the clusters on the right hand side.

Below are some of the observations.

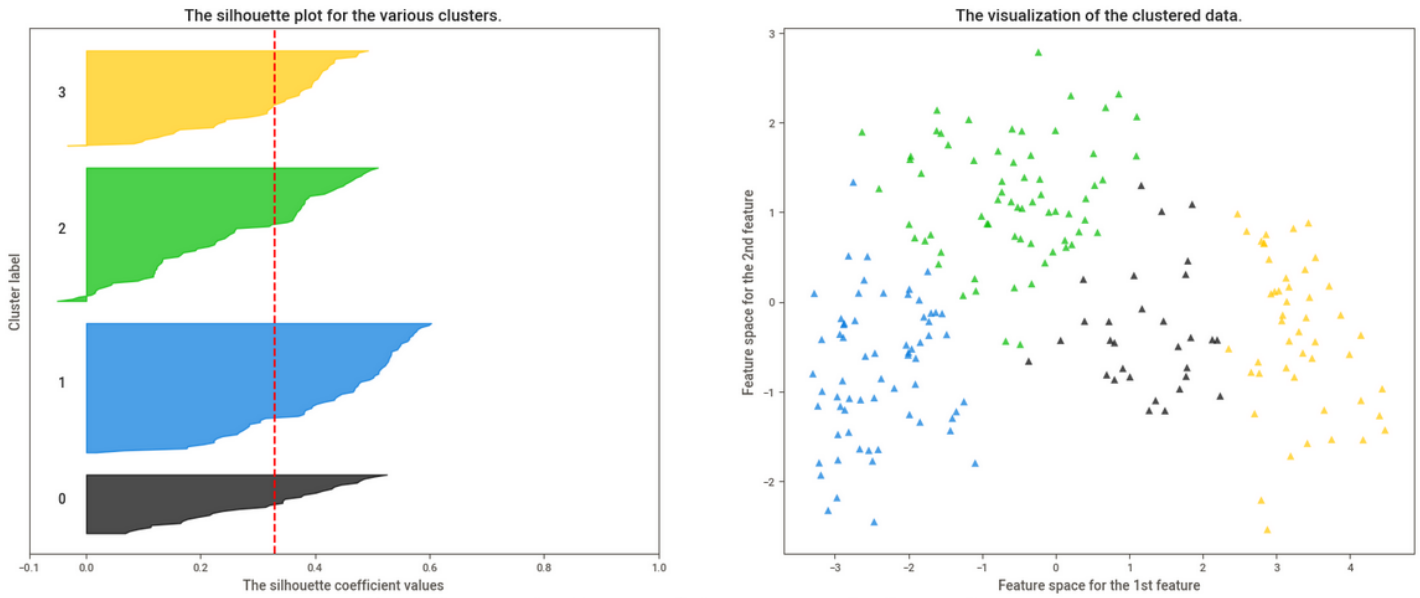
Average Silhouette Score



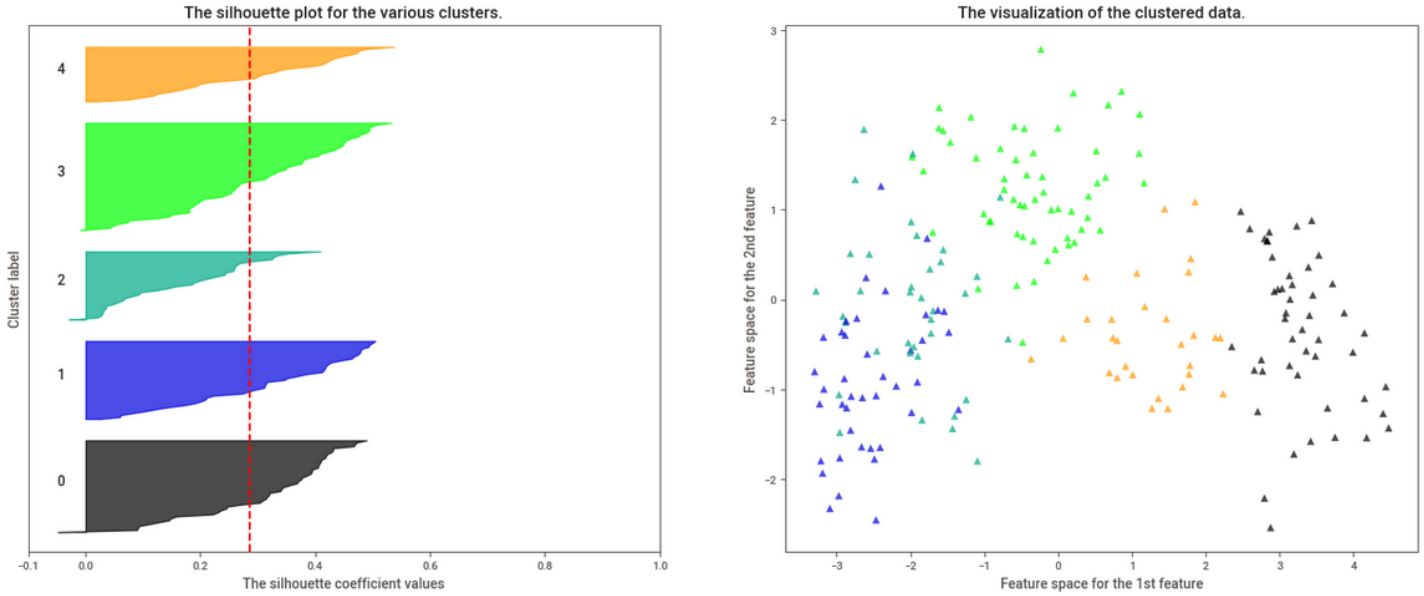


Average Silhouette Score

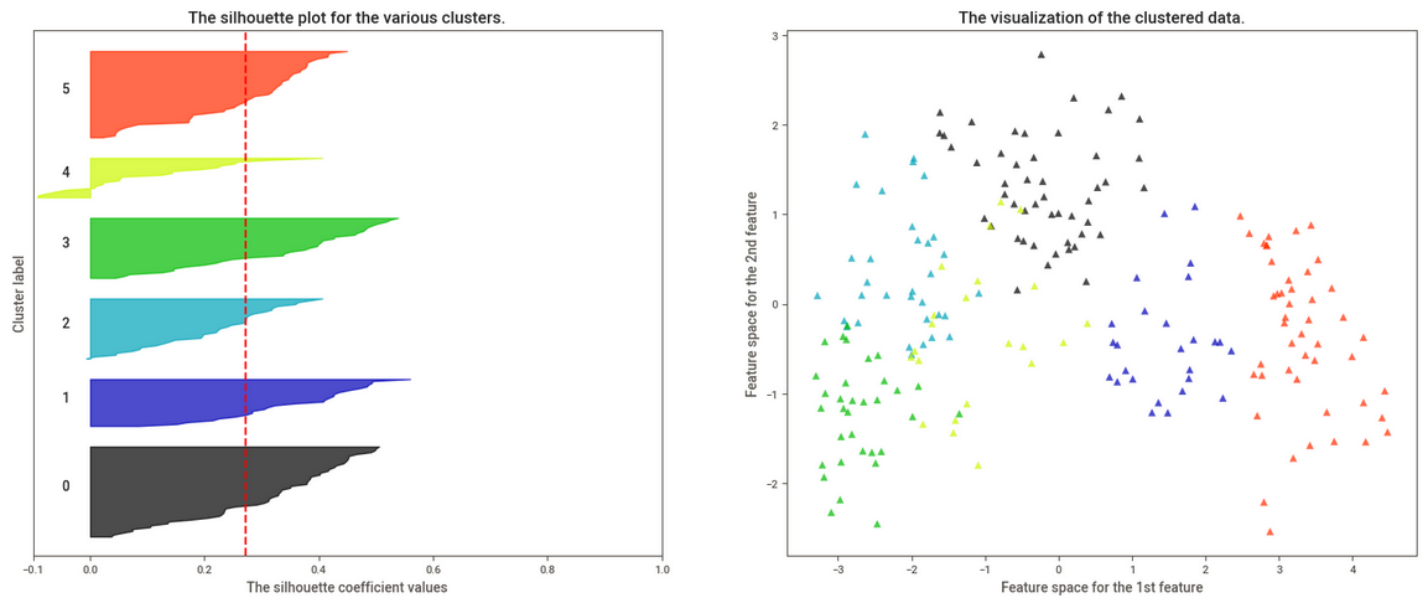
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



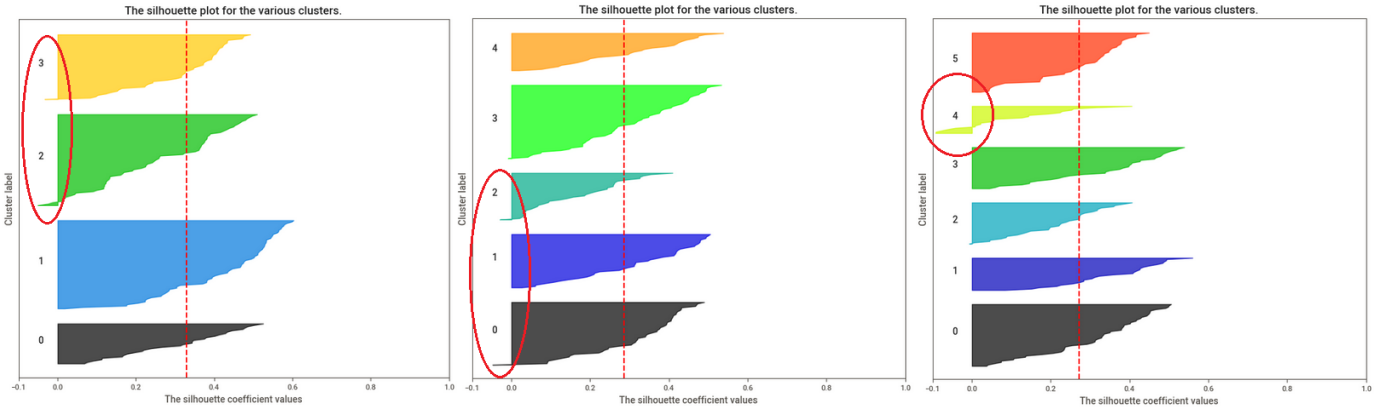
Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



Average Silhouette Score

Now analyzing the above silhouette graphs, we can make a few observations.

We can see in some of the graphs on the left there is a negative silhouette width indicated, even if it is only for a few points within a cluster. These are shown below.



A negative silhouette width, indicates that the point might have been clustered incorrectly as the distance between its own centroid is more compared to the centroid of the nearest neighboring cluster. Hence we eliminate all such clusters.

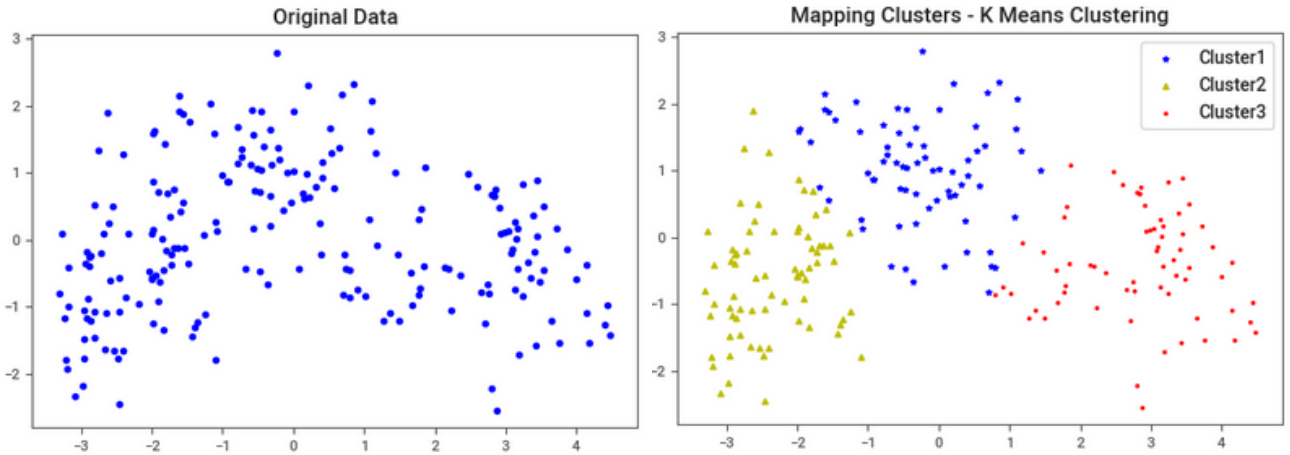
With this we removed options with clusters 4,5 & 6, leaving us with just 2 options i.e. 2 clusters or 3 clusters.

Now for options with 2 or 3 clusters, we can see the average silhouette score is also comparable and both the options have no negative sil-width issue. Also for both the options we can see the clusters are well separated visually as well, looking at the right hand side scatter plot.

Conclusion

The call here is a subjective one, since having 3 clusters will give us a much better market segmentation and opportunity to target specific customers, which is not possible to the same extent with 2 clusters. Also 3 clusters co-incide with our finding from elbow curve above.

Hence we finally choose the option with 3 clusters, after evaluating various graphs above, including the elbow curve and the average silhouette score plots.

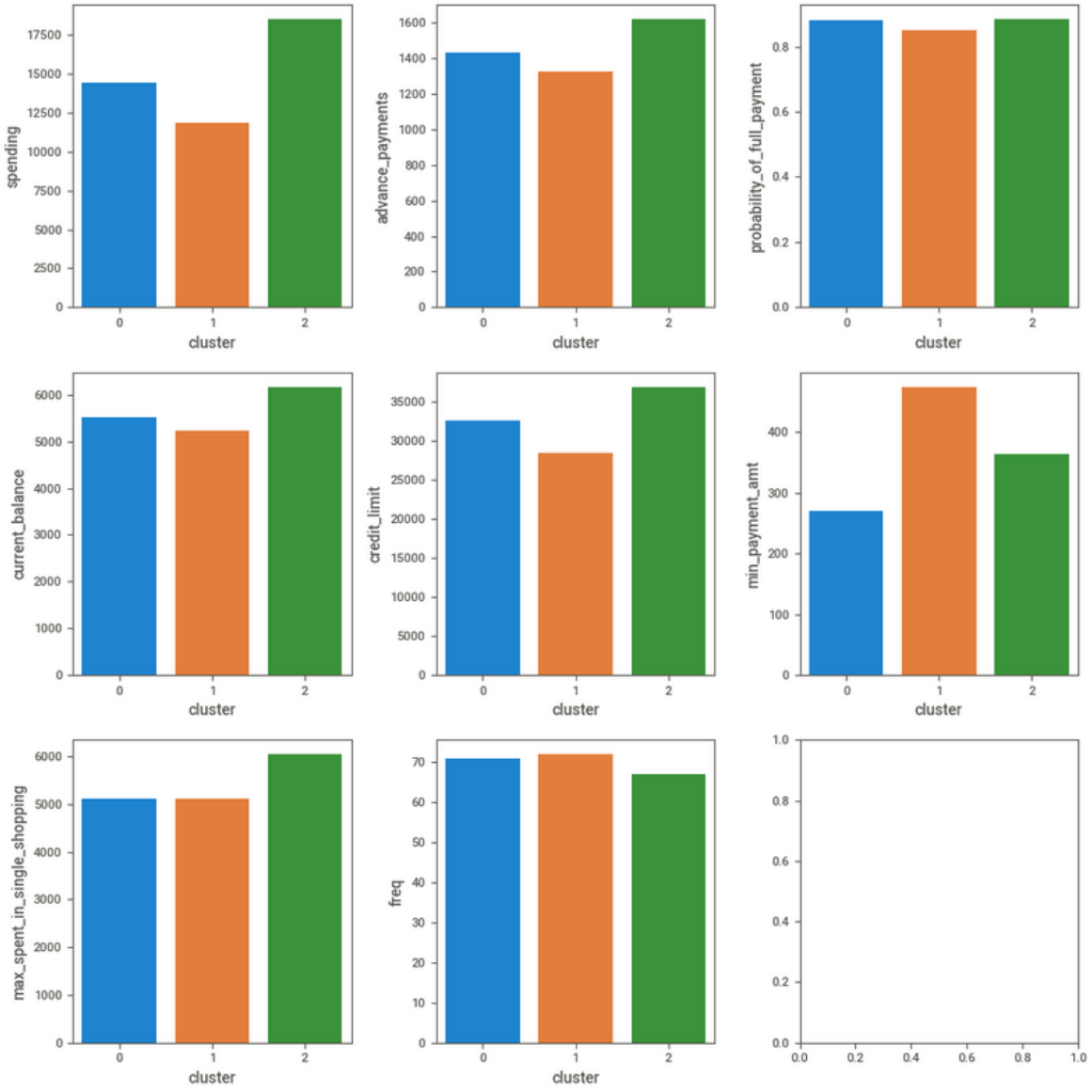


Question 1.5

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

We have defined 3 clusters using the K-means clustering algorithm above. We will describe the cluster profiles for these 3 clusters and the corresponding promotional strategies that can be applied on these 3 clusters.

Cluster Profiles

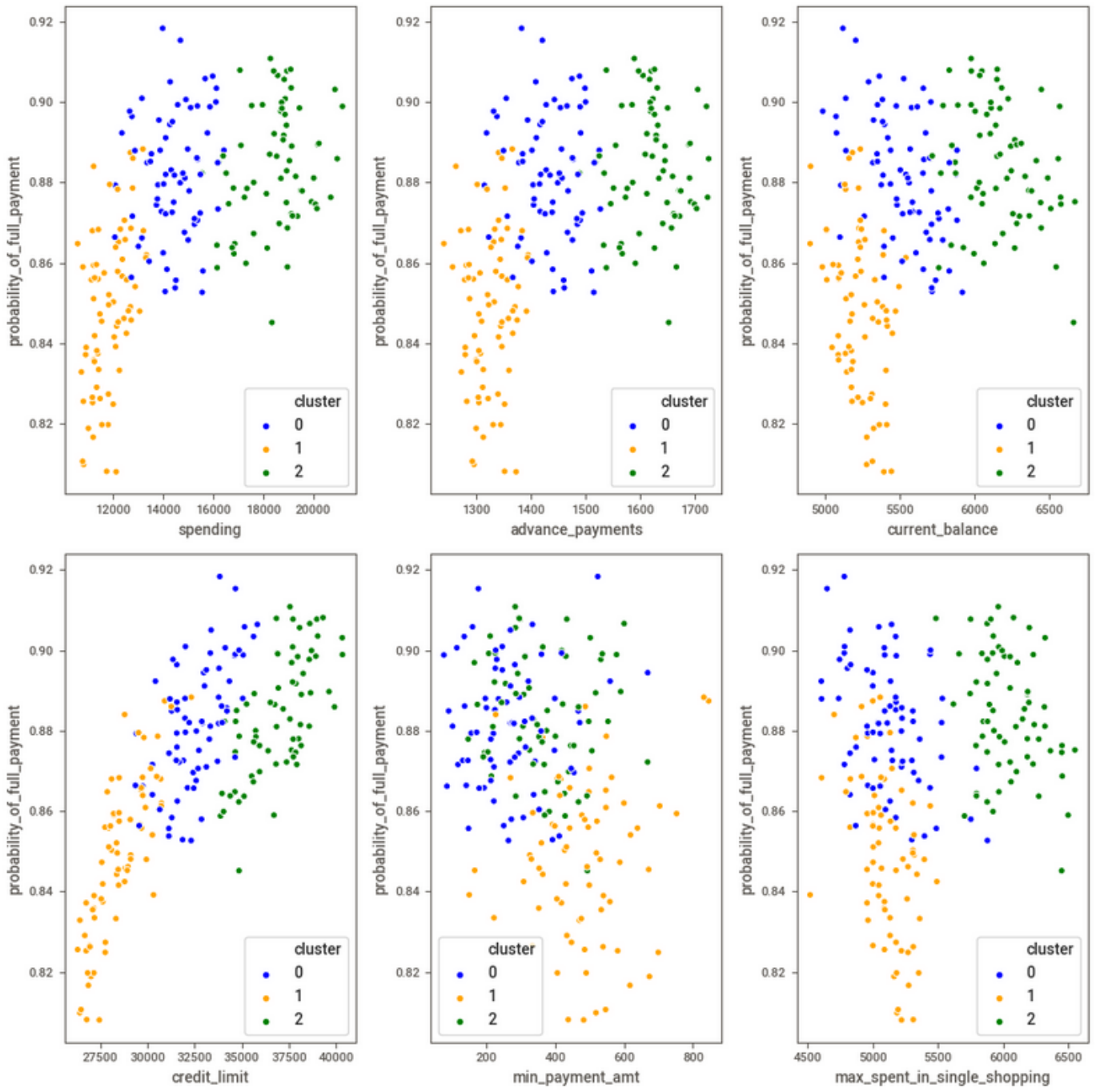


Cluster Profiles

Taking an average of data in each column for the three different clusters, we get the below cluster profile.

cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	freq
0	14437.887324	1433.774648	0.881597	5514.577465	32592.253521	270.734085	5120.802817	71
1	11856.944444	1324.777778	0.848253	5231.750000	28495.416667	474.238889	5101.722222	72
2	18495.373134	1620.343284	0.884210	6175.686567	36975.373134	363.237313	6041.701493	67

We then make a scatter plot for each column with the "probability_of_full_payment" using cluster column as hue to get the below.



Cluster Profiles

Based on above analysis, we can thus segment the customers into 3 different categories.

1. **Regular Customers (Cluster = 1, Color = Orange)**
2. **Premium Customers (Cluster = 0, Color = Blue)**
3. **VIP Customers (Clusters = 2, Color = Green)**

One thing which is apparent from the above graphs is that, VIP customers have the highest probability to make a full payment, followed by the Premium customers, followed by Regular customers.

VIP Customers tend to have the highest **spending,advance_payment,current_balance** as well as **credit_limit**.

Regular Customers tend to have the lowest **spending,advance_payment,current_balance** as well as **credit_limit**.

Premium Customers lie in between VIP and Regular Customers on most of the criterion.

Promotional Strategy

VIP Customers - Since these customers are obviously the cash cows for the bank, it makes sense to **offer higher credit limits, higher bonus points based on spending and other incentives** to these customers as they spend more and generally pay the full amount on time.

Regular Customers - These customers perform the worst on almost all metrics, except one i.e. max_spent_in_single_shopping where they are almost at par with Premium Customers. Since these customers spend more in single go and have lower probability of making a full payment, bank can **provide EMI offers to such customers**. This would enable the customers to cover their expenses in a better manner while earning interest for the bank.

Premium Customers These are the safe players. They have good enough credit limits, and have higher probability to make full payment. Bank needs to push such customers to spend more, as they less risk customers. max_spent_in_single_shopping for Premium customers is at par with Regular Customers, whilst Premium customers can perform much better on this front. Bank should **introduce promotional discounts and bonuses to boost spend in single go for such customers**. Also higher credit limit might be allowed for such customer.

QUESTION - 2

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**)
7. Destination of the tour (**Destination**)
8. Amount of sales of tour insurance policies (**Sales**)
9. The commission received for tour insurance firm (**Commission**)
10. Age of insured (**Age**)

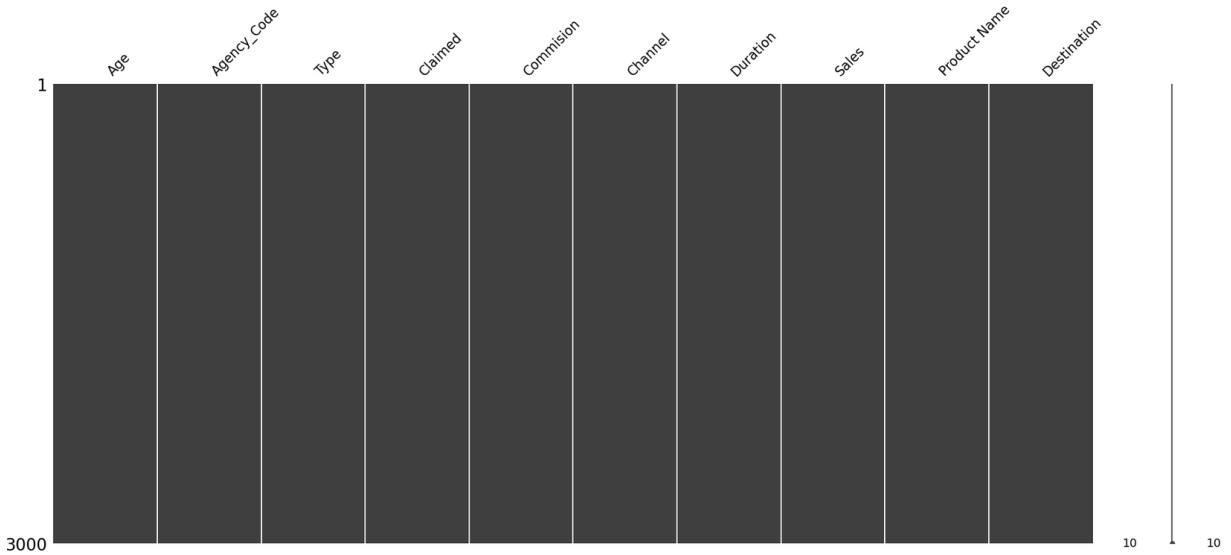
Question 2.1

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

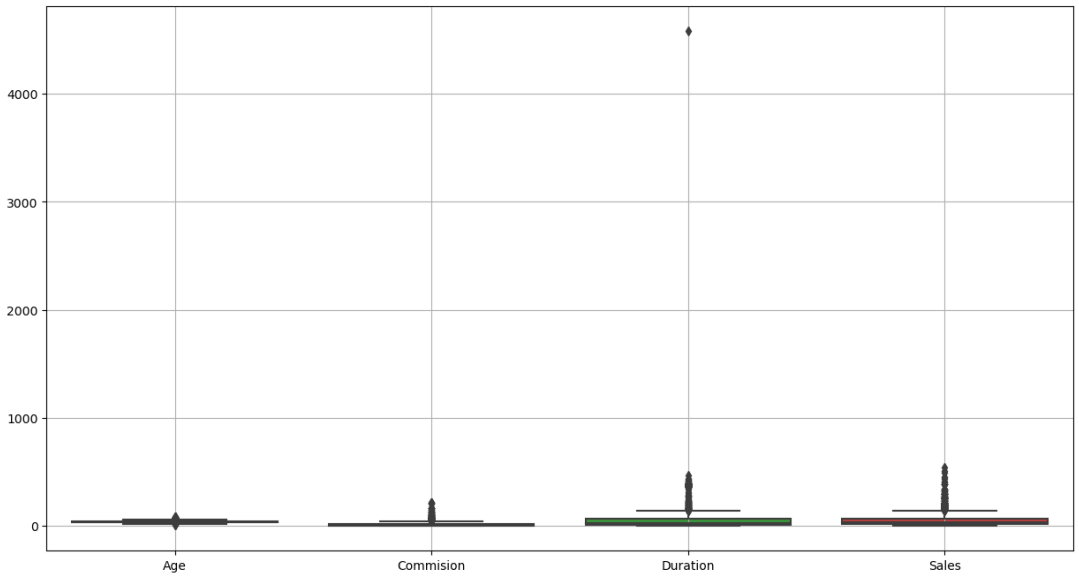
We imported the dataset using pandas read_csv function and performed descriptive analysis. Below were few of the observations that we were able to make.

Salient Features

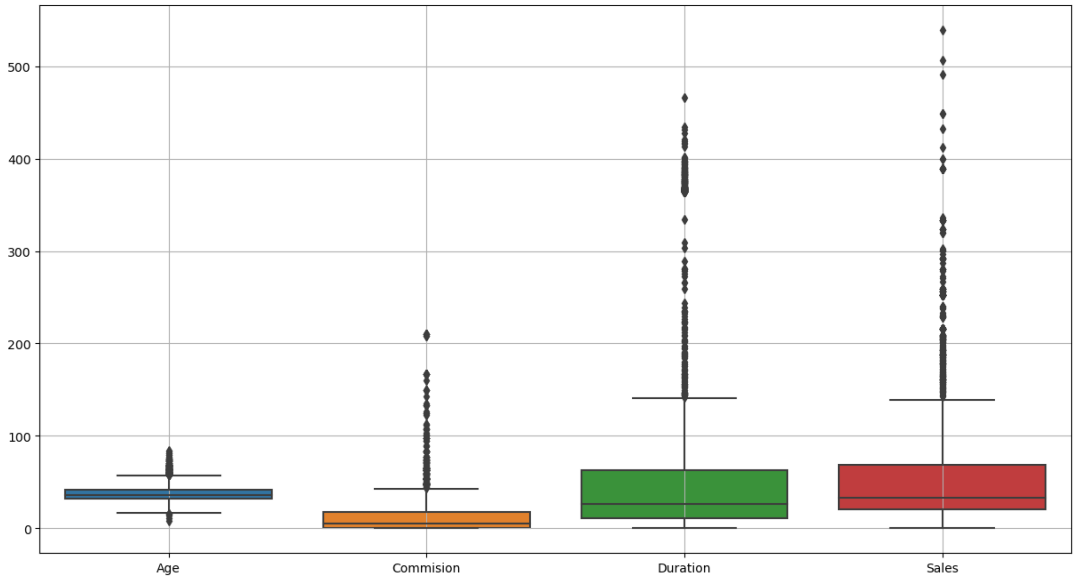
No Null Values



Before removing extreme outliers



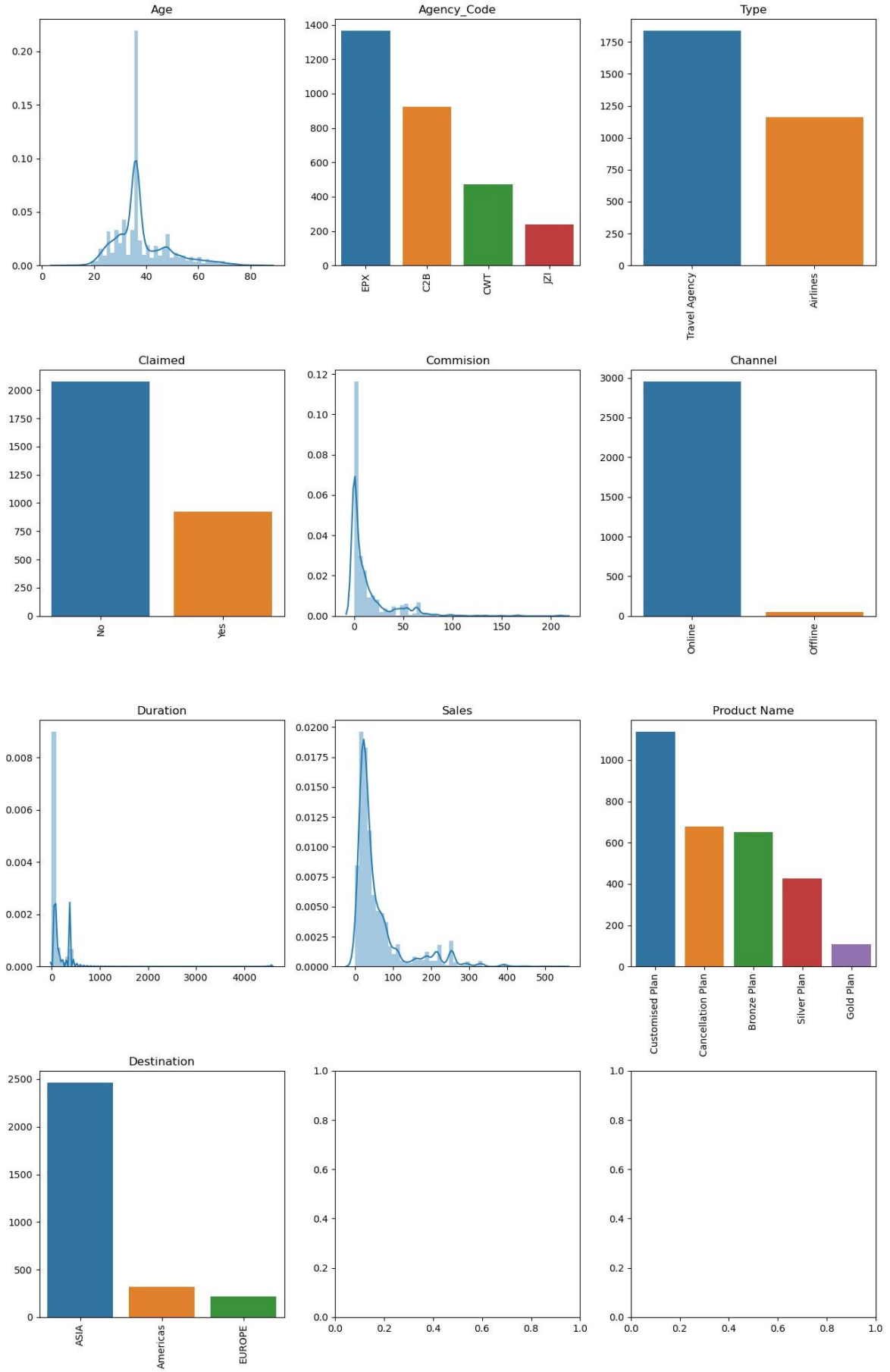
After removing extreme outliers



5 Point Summary

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	2998.000000	2998	2998	2998	2998.000000	2998	2998.000000	2998.000000	2998	2998
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2074	NaN	2952	NaN	NaN	1135	2463
mean	38.092061	NaN	NaN	NaN	14.536765	NaN	68.520680	60.283996	NaN	NaN
std	10.462712	NaN	NaN	NaN	25.488146	NaN	105.790319	70.744865	NaN	NaN
min	8.000000	NaN	NaN	NaN	0.000000	NaN	0.000000	0.000000	NaN	NaN
25%	32.000000	NaN	NaN	NaN	0.000000	NaN	11.000000	20.000000	NaN	NaN
50%	36.000000	NaN	NaN	NaN	4.630000	NaN	26.500000	33.000000	NaN	NaN
75%	42.000000	NaN	NaN	NaN	17.245000	NaN	63.000000	69.000000	NaN	NaN
max	84.000000	NaN	NaN	NaN	210.210000	NaN	466.000000	539.000000	NaN	NaN

Univariate Analysis



Few Inferences from Univariate analysis:-

1) Age - 36 is the most frequently occuring age. It has been observed all the records for EPX agency have the age. This seems to be an anamoly with the data, however for the lack of evidence and for the purpose of this report, we will consider all the records to be genuine and hence we do not modify this column whatsoever.

2) Agency_code - We can see the top most agency selling the insurance is EPX, followed by C2B, then CWT and finally JZI.

3) Type - More sales are made via the Travel agencies compared to the airlines.

4) Claimed - More records with claimed status as "No" are present. Only 30% of overall records are present with Claimed status as "Yes".

5) Commission - Average commission is 14.5, while the maximum commission being 210

6) Channel - Almost all the sales come from the Online channel, with Offline channel contributing just 1.5 % of the total sales.

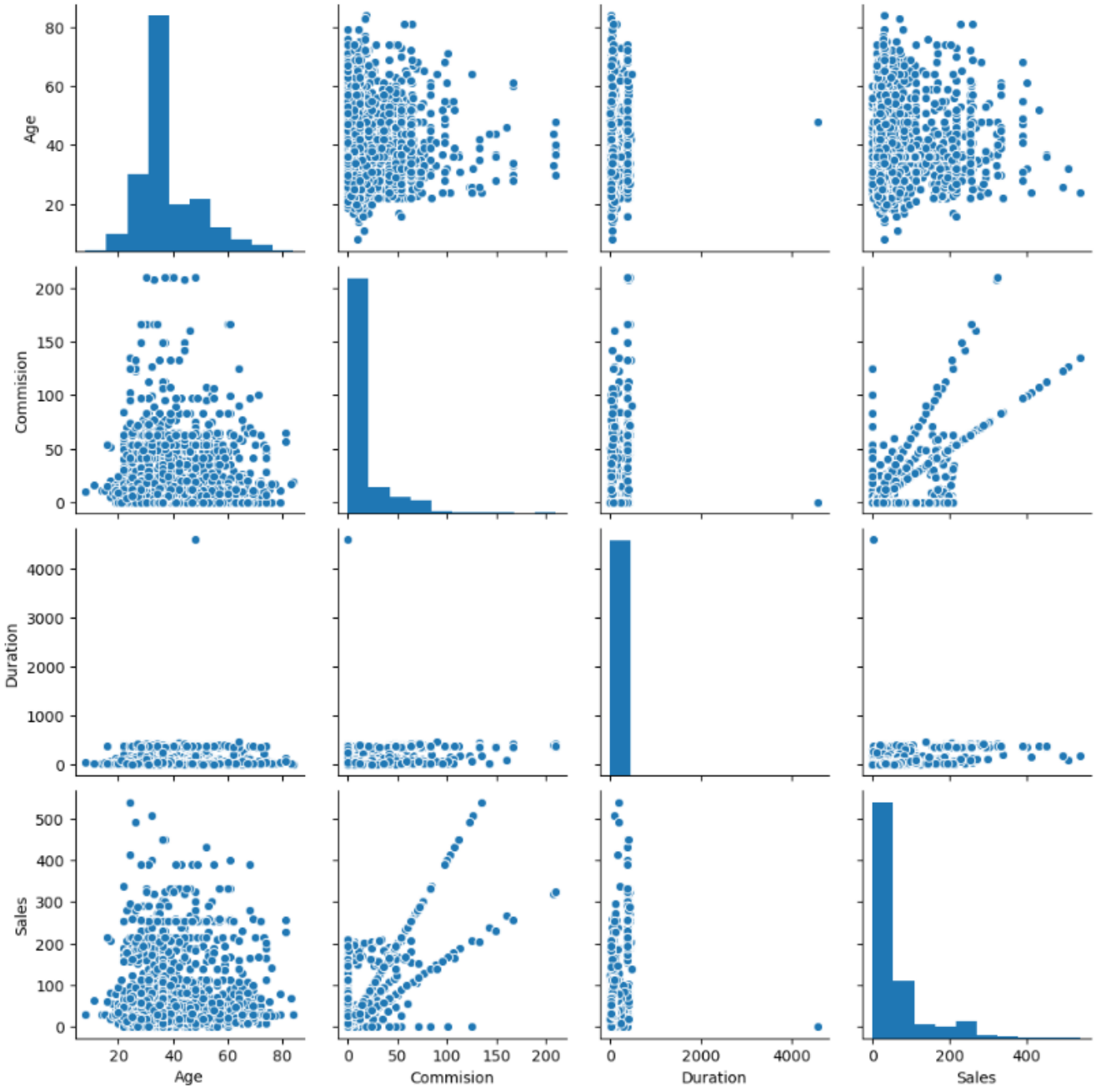
7) Duration - Average duration after removing the extreme outliers is 68 days.

8) Sales - Average sales figure is 60.24, while the maximum is 539.

9) Product_Name - Most customers prefer a customized plan, however from the standard pack "Cancellation Plan" and "Bronze Plan" are pretty popular as well.

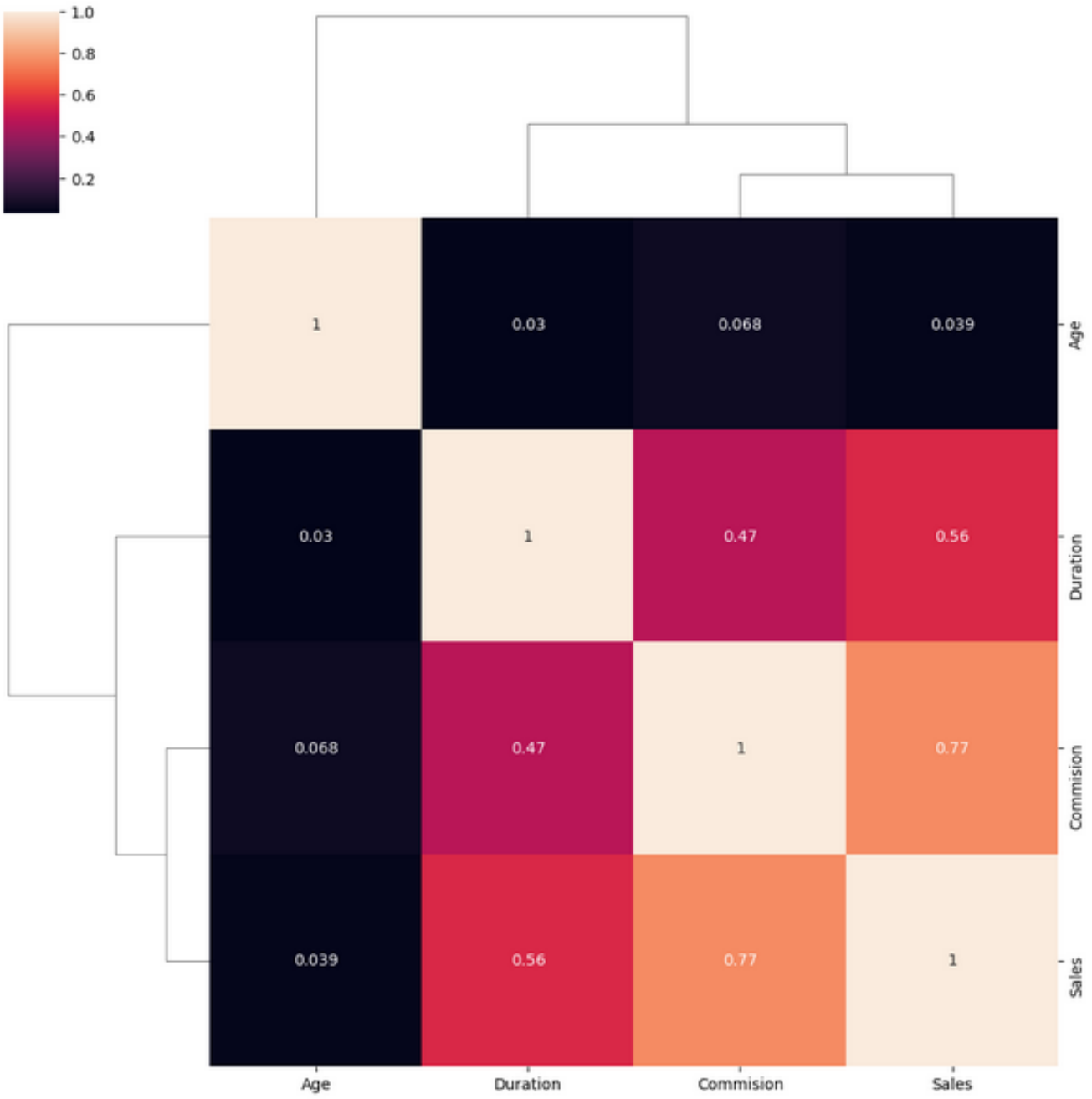
10) Destination - Most popular destination is Asia, followed by Americas and Europe.

BiVariate Analysis



We do not see any strong co-relations between the different variables. The strongest co-relation present in the data set is between "Sales" and "Commission".

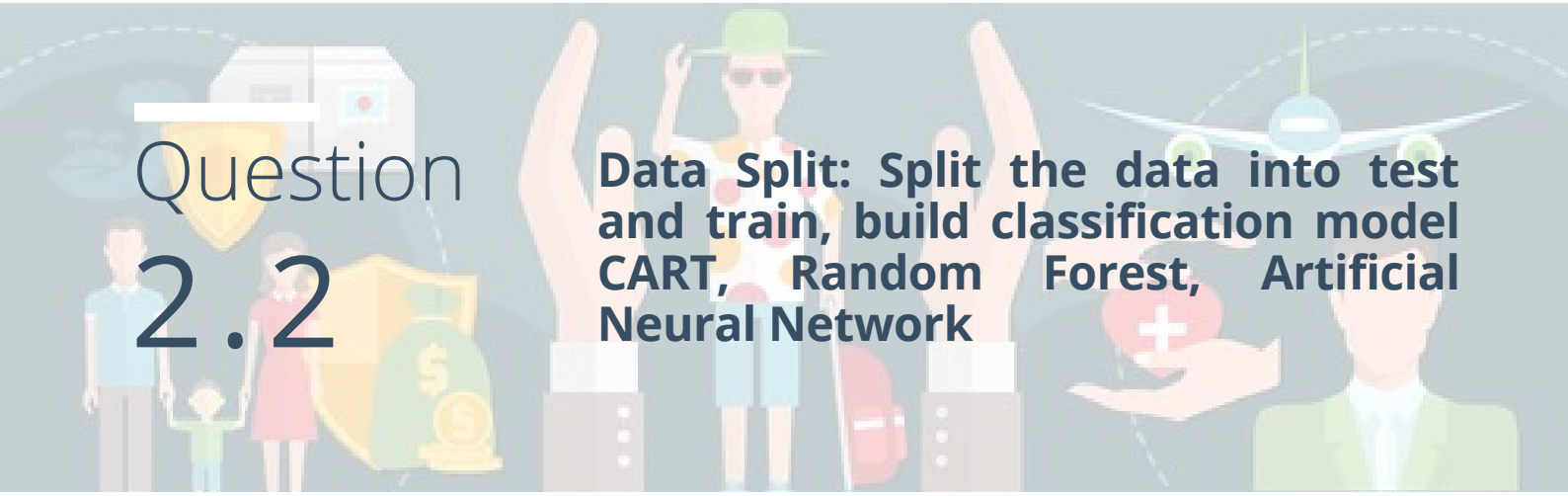
BiVariate Analysis



No strong co-relations observed from the above cluster map.

Strongest co-relation present between "**Sales**" and "**Commission**", which makes sense, as higher the sales, higher would be the commission earned.

This is followed by the co-relation between "**Duration**" and "**Sales**", which again makes sense and the revenue earned would be higher for longer duration insurance cover.



Question 2.2

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

Before applying any of the models, we first followed below steps as part of data pre-processing.

1) Remove Extreme Outliers - We removed two values from **Duration** variable, one was 4850 while the other one was -1. We dropped both the rows as it did not make significant impact on the overall size of the data set.

2) Check for Duplicates - We checked for duplicate records and found 139 duplicate values, however it seemed like genuine data, since multiple customers can opt for a similar plan of similar duration and hence these duplicates did not seem to be due to any error.

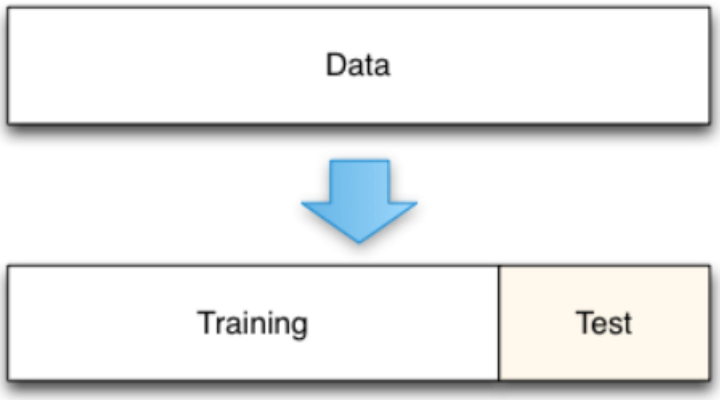
Moreover we wanted our model to be able to generalize things in a better manner, hence we chose to stick with the duplicate values. Also the duplicates did not form a significant portion of our overall data set of 3000 records.

3) Converting Categorical Data to Continuous - In order for us to use any of the supervised machine learning algorithms a.k.a CART, Random Forest and ANN, we first converted our categorical data into numerical data using one hot encoding.

4) Standardizing our data - Standardization of the data was performed only before applying ANN. CART and RF do not require our data to be standardized hence we did not standardize our data before applying CART and RF.

We then moved to the task of splitting the data into training and testing data sets, so that we could train our model on the training set, while checking various performance metrics like accuracy etc. on the testing set.

We first removed the **Claimed** variable from the data frame to form our **X** or the predictor variable set.
Then we popped the **Claimed** variable to form our **Y** or the Target variable set.



After this we split both **X & Y** in the ratio of 70:30, i.e. 70% training data while the rest 30% is our testing data set.

Data was randomized during the process of splitting, to ensure data is well distributed amongst the two data sets. This ensures there are no unexpected patterns in either of our data sets, which can lead to incorrect training or predictions at the end of the day.

We used random_state=1 to make the split. Also the shuffle parameter was left to its default value of "True" to ensure the data is split randomly.

We then proceeded with the task of building various prediction models including - CART, Random Forest and Artificial Neural Networks (ANN) which are discussed in detail in the following pages.

Applying CART ML Model

We first applied CART model using **DecisionTreeClassifier** from sklearn library, without any pruning methods and got a huge tree as a result which yielded an accuracy of just 68% on the test data.

Then we tried various pruning parameter combinations using GridSearchCV and calculating accuracy score on test data after each iteration.

We performed a total of 4 **GridSearchCV** iterations, each time narrowing down further on the parameters to optimize our model as much as possible.

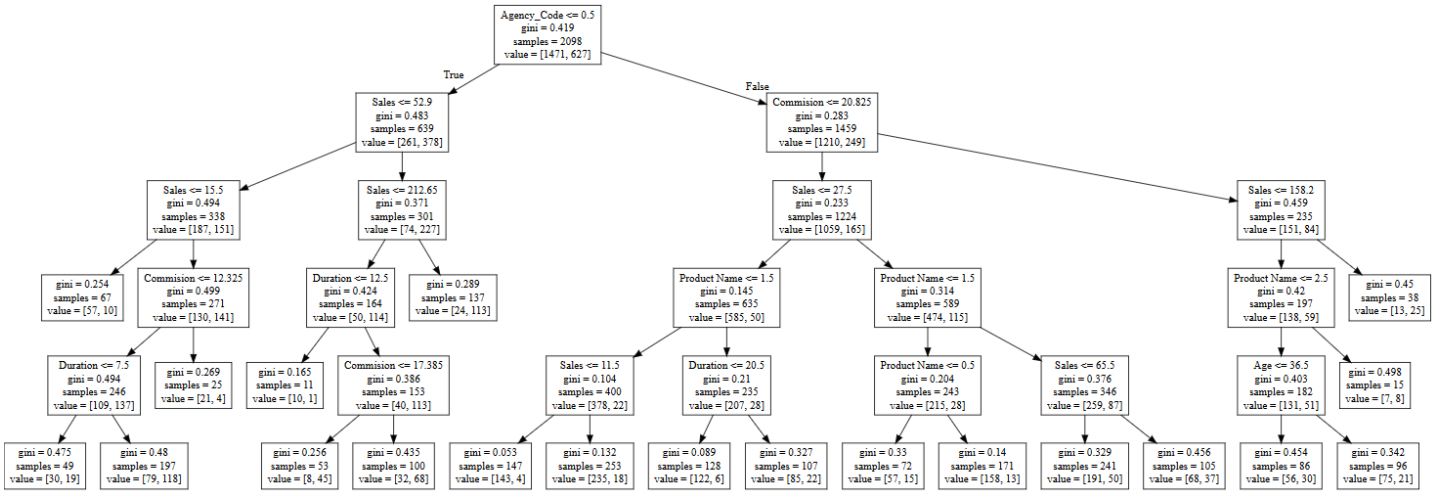
Cross Validation parameter value cv=10 was used for each of the iterations. The best accuracy after 4 iterations that we were able to elicit was **77.77%** on the test data. Other performance metrics will be discussed later.

Default criterion of gini gain was used for decision nodes.
Best params and the feature importance that was obtained after optimizing the model were as below.

```
{'max_depth': 5,
 'max_features': 6,
 'min_samples_leaf': 6,
 'min_samples_split': 142,
 'random_state': 27}
```

	Imp
Agency_Code	0.5621
Sales	0.2403
Commision	0.1024
Duration	0.0537
Product Name	0.0359
Age	0.0055
Type	0.0000
Channel	0.0000
Destination	0.0000

Below is the tree that we obtained after 4 iterations of GridSearchCV and implementing various pruning techniques.



Applying Random Forest ML Model

After CART we proceeded with the Random Forest ML model using **RandomForestClassifier** from sklearn library, which is an ensemble model based on CART model, in effect having multiple trees as part of the model and in the end voting happening in a democratic manner between the different trees to mark the final Target variable value.Concept of wisdom of the crowd is used here to ensure a better fit model.

Again number of pruning techniques were tried along with different number of trees in the ensemble using the **GridSearchCV**. We performed a total of 3 iterations, each time fine tuning the param grid to further optimize the model.

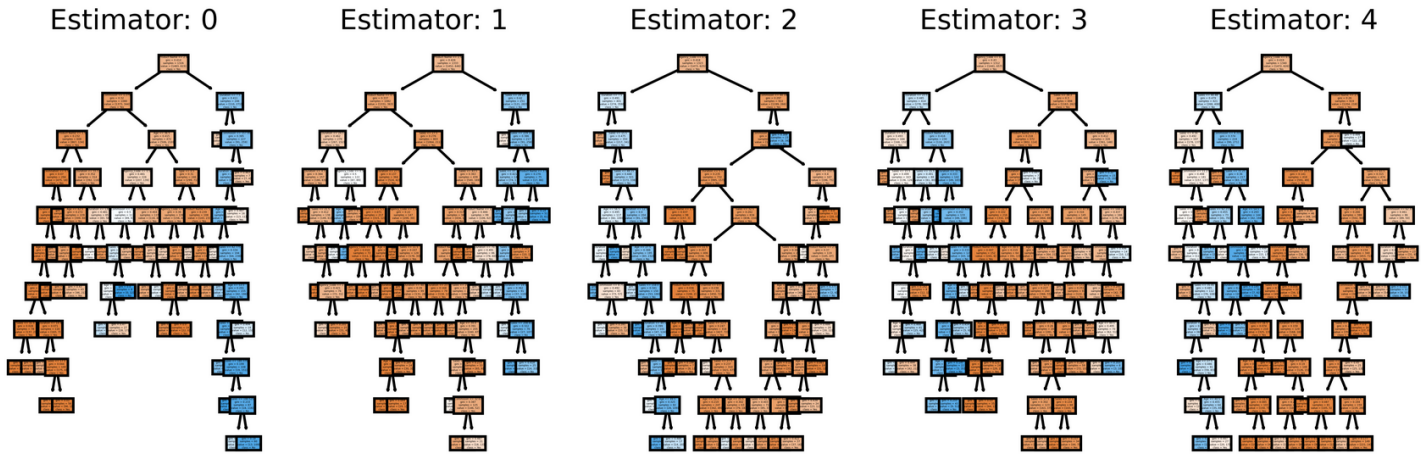
The best accuracy that we were able to achieve w.r.t the test data was **78%** after 3 iterations. Other performance metrics will be discussed later.

Default criterion of gini gain was used for decision nodes.
Best params and the feature importance that was obtained after optimizing the model were as below.

```
{'max_depth': 10,
 'max_features': 4,
 'min_samples_leaf': 7,
 'min_samples_split': 63,
 'n_estimators': 22}
```

Feature	Imp
Agency_Code	0.2972
Product Name	0.2123
Sales	0.1850
Duration	0.1077
Commision	0.0951
Age	0.0569
Type	0.0376
Destination	0.0070
Channel	0.0011

Below are the first 5 trees from our ensemple of 22 trees that we obtained after 3 iterations of GridSearchCV and implementing various pruning techniques.



Applying Artificial Neural Networks (ANN) Model

After Random Forest we proceeded with the Artificial Neural networks (ANN) model using **MLPClassifier** from sklearn library, also we used **Keras** sequential model for one of the iterations.

9 iterations of GridSearchCV were performed on MLPClassifier, each time fine tuning below parameters.

activation - ReLu, Tanh, logistic & finally LeakyReLU (Using Keras). Note that to implement LeakyReLU activation function using Keras, we sequentially added a layer after each hidden layer with alpha = 0.01.

hidden_layers_sizes - Various combinations of 5,(5,5),10,(10,10),15,(15,15),(20,20), (30,30),(40,40),50,(50,50),(80,80),100,(100,100),200,300 were tried over multiple iterations.

max_iter - Various values of iteration varying from 1000 to 6000 with the step size of 1000 were tried over multiple iterations. This is a stopping criteria for the ANN model.

solver - We tried using two optimizer techniques i.e. sgd (Stochastic Gradient Descent) and adam (Adaptive Movement) were used over multiple iterations.

tol - tol which is short for tolerance is the variable which defines the stopping criteria for the ANN algorithm. We tried multiple tolerance values ranging from 0.001,0.0001,0.00001 to 0.000001.

n_iter_no_change - number of iterations with no change was kept to its default value of 10 for all the GridSearchCV iterations.

random_state - random state was kept 1 throughout all the iterations to get a consistent idea w.r.t the accuracy of all the other parameters present inside the param grid.

verbose - verbose = 100 was kept to track the progress of the model after each fit. Since each iteration was time and resource intensive, this helped us keep track of the things at a finer level.

Applying Artificial Neural Networks (ANN) Model

The best accuracy that we were able to achieve w.r.t the test data was **77.22%** after 9 iterations. Other performance metrics will be discussed later.

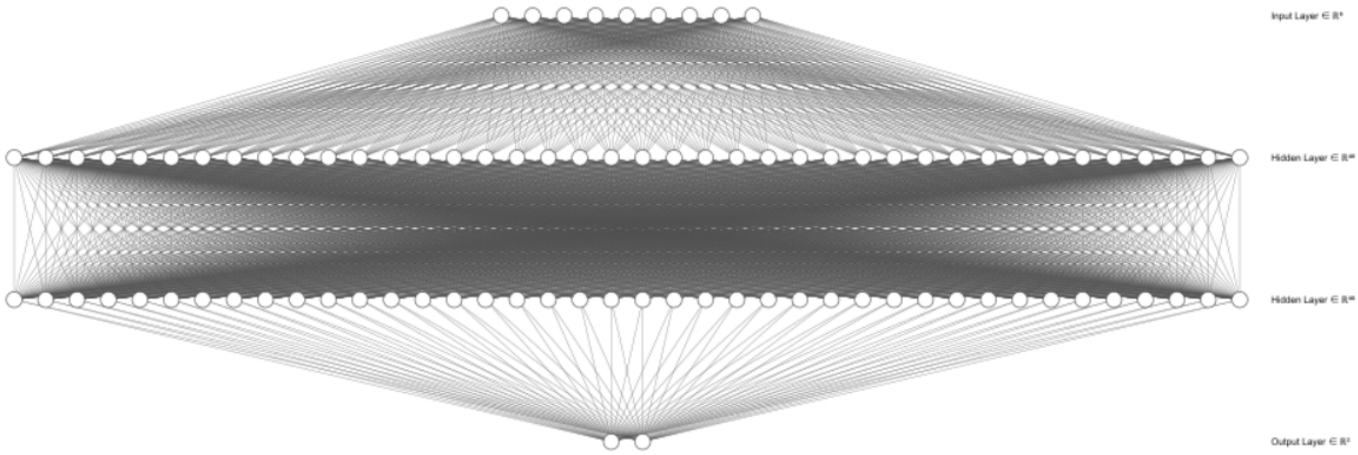
Best params that were obtained after optimizing the model were as below.

```
{'hidden_layer_sizes': (40, 40),
 'max_iter': 6000,
 'random_state': 1,
 'solver': 'sgd',
 'tol': 1e-05,
 'verbose': 10}
```

Our ANN model hence had total 4 layers, 1 input, 2 hidden and 1 output layer. These were as below.

Input Layer - 9 Neurons
Hidden Layer 1 - 40 Neurons
Hidden Layer 2 - 40 Neurons
Output Layer - 2 Neurons

Below is a visual representation of our ANN model.



Question 2.3

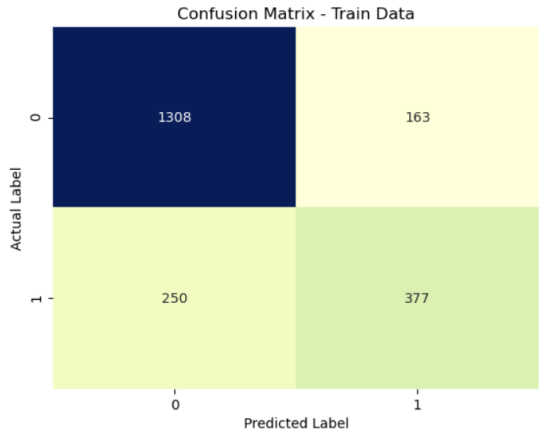
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model

Below are some of the other performance metrics that we calculated for all our ML models discussed above.

CART ML Model - Performance

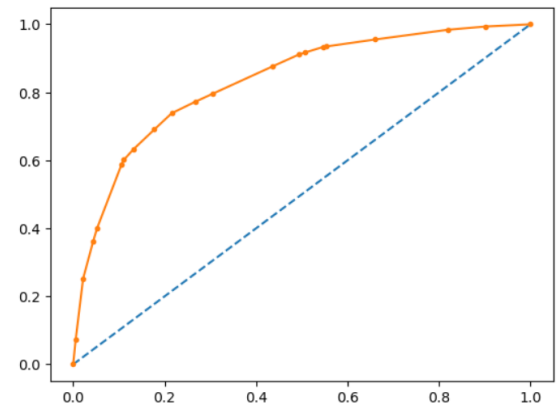
Train Data

Accuracy Score (Train Data) - 80.31%



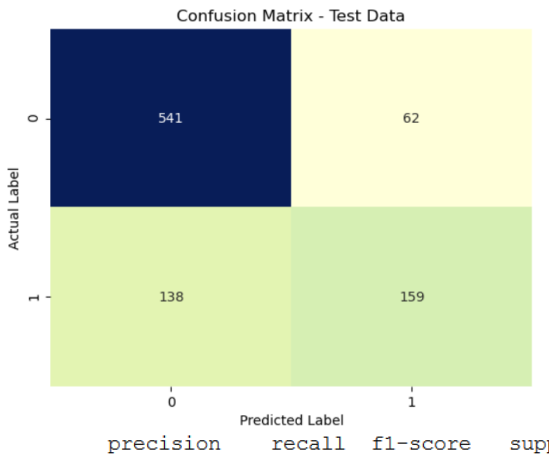
	precision	recall	f1-score	support
0	0.84	0.89	0.86	1471
1	0.70	0.60	0.65	627
accuracy			0.80	2098
macro avg	0.77	0.75	0.75	2098
weighted avg	0.80	0.80	0.80	2098

AUC: 0.832



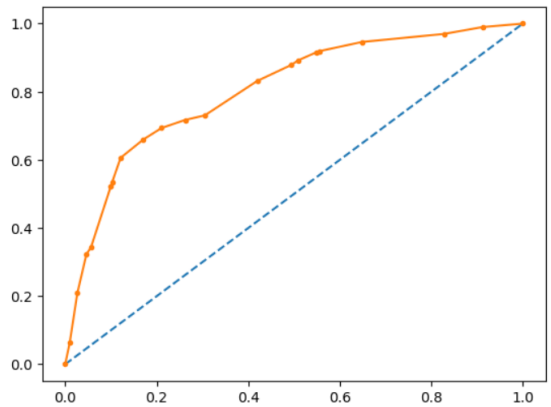
Test Data

Accuracy Score (Test Data) - 77.77%



	precision	recall	f1-score	support
0	0.80	0.90	0.84	603
1	0.72	0.54	0.61	297
accuracy			0.78	900
macro avg	0.76	0.72	0.73	900
weighted avg	0.77	0.78	0.77	900

AUC: 0.805

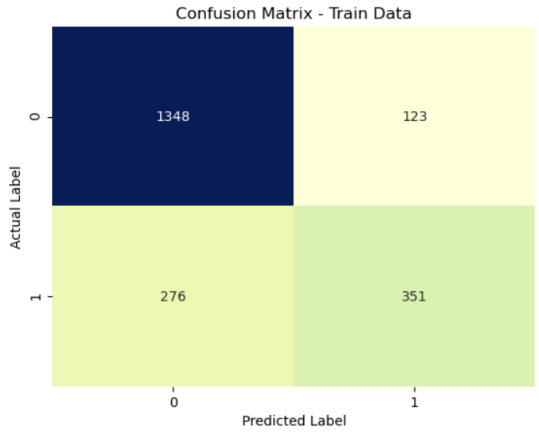


Random Forest ML Model - Performance

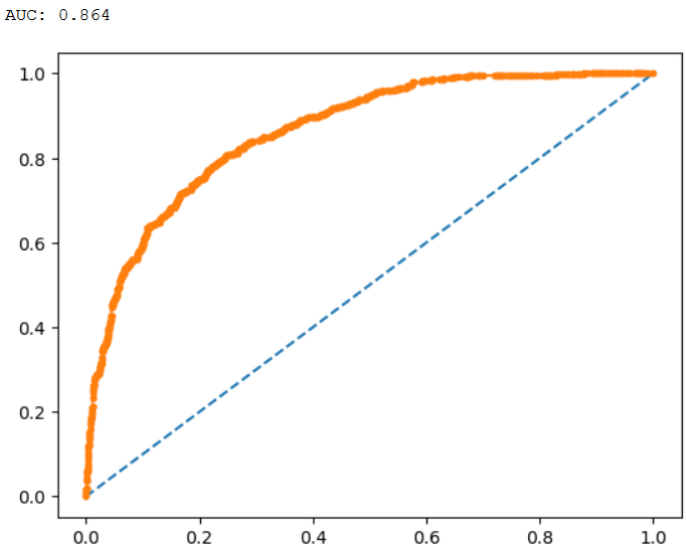
Below are some of the other performance metrics that we calculated for Random Forest ML model.

Train Data

Accuracy Score (Train Data) - 80.98%

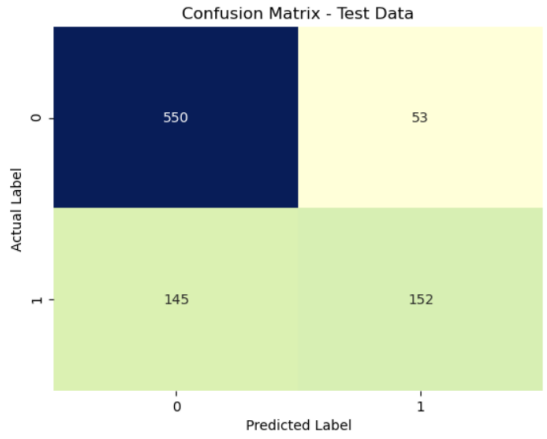


	precision	recall	f1-score	support
0	0.83	0.92	0.87	1471
1	0.74	0.56	0.64	627
accuracy			0.81	2098
macro avg	0.79	0.74	0.75	2098
weighted avg	0.80	0.81	0.80	2098

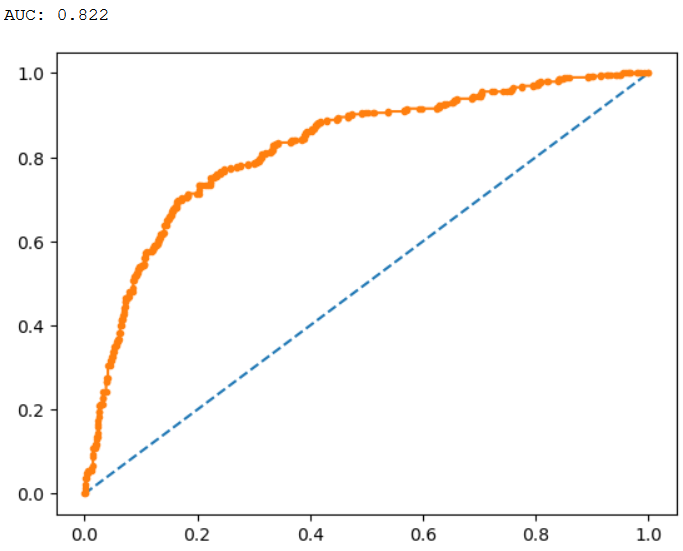


Test Data

Accuracy Score (Test Data) - 78.00%



	precision	recall	f1-score	support
0	0.79	0.91	0.85	603
1	0.74	0.51	0.61	297
accuracy			0.78	900
macro avg	0.77	0.71	0.73	900
weighted avg	0.77	0.78	0.77	900

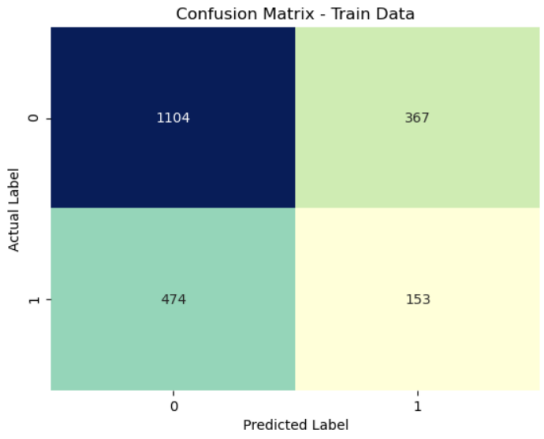


Artificial Neural Network Model - Performance

Below are some of the other performance metrics that we calculated for Artificial Neural Network (ANN) ML model.

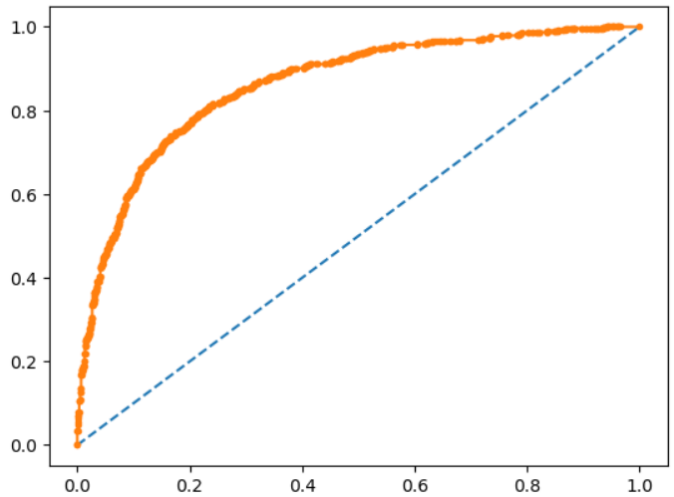
Train Data

Accuracy Score (Train Data) - 81.31%



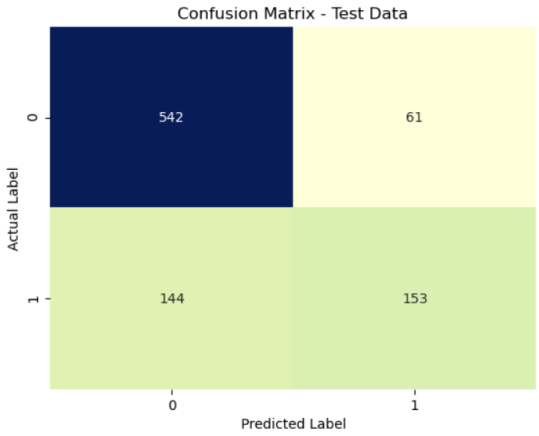
	precision	recall	f1-score	support
0	0.83	0.92	0.87	1471
1	0.74	0.57	0.65	627
accuracy			0.81	2098
macro avg	0.79	0.74	0.76	2098
weighted avg	0.81	0.81	0.81	2098

AUC: 0.862



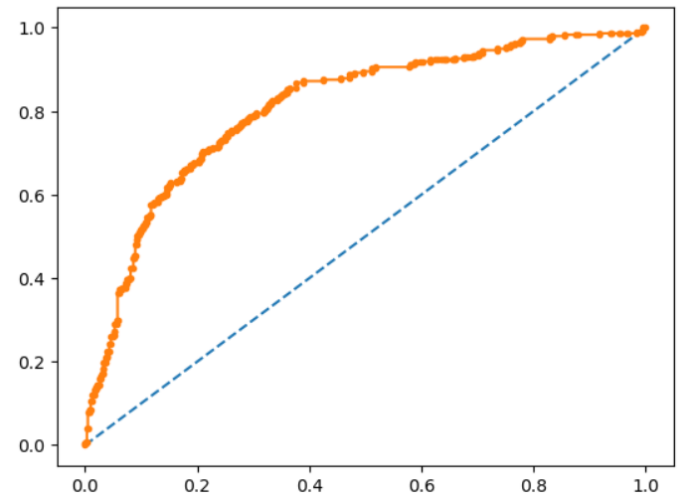
Test Data

Accuracy Score (Test Data) - 77.22%



	precision	recall	f1-score	support
0	0.79	0.90	0.84	603
1	0.71	0.52	0.60	297
accuracy			0.77	900
macro avg	0.75	0.71	0.72	900
weighted avg	0.77	0.77	0.76	900

AUC: 0.809



Question 2.4

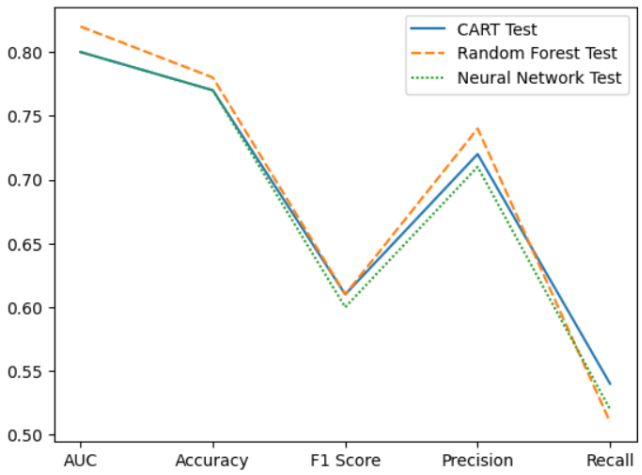
Final Model: Compare all the model and write an inference which model is best/optimized.

Final Model

So far we have built multiple models using supervised machine learning techniques and have evaluated them individually to check for their overall performance.

Taking a collective look at the performance metrics for all the models and making a final selection of the model to be used.

	CART Test	CART Train	Random Forest Test	Random Forest Train	Neural Network Test	Neural Network Train
Accuracy	0.77	0.80	0.78	0.80	0.77	0.81
AUC	0.80	0.83	0.82	0.86	0.80	0.86
Recall	0.54	0.60	0.51	0.56	0.52	0.57
Precision	0.72	0.70	0.74	0.74	0.71	0.74
F1 Score	0.61	0.65	0.61	0.64	0.60	0.65



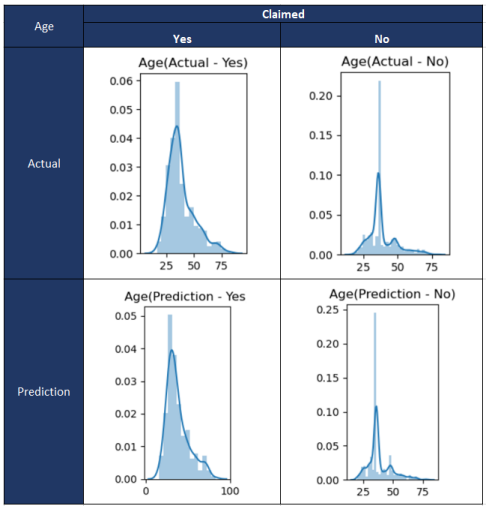
Looking at above figures and the lineplot, we can see Random Forest wins on almost all the metrics, except the recall where it lags behind by a tiny margin.

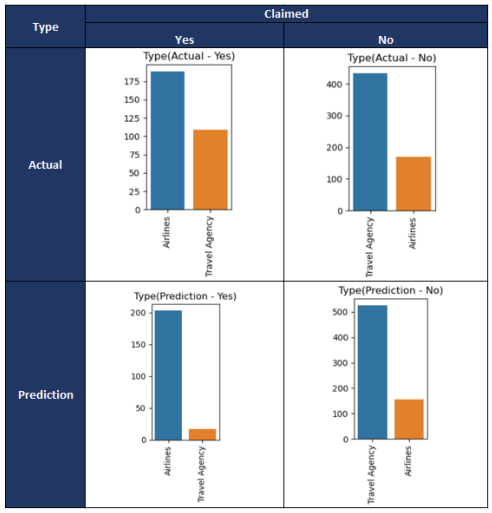
Hence looking at the overall performance, we select **Random Forest** as the model of our choice here, as it is the most optimized of all the models we have tried.

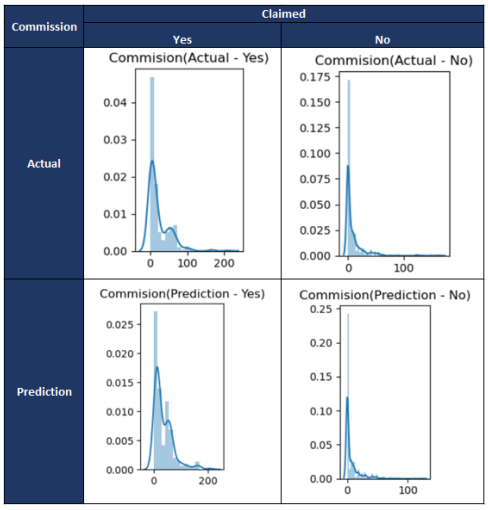
Question 2.5

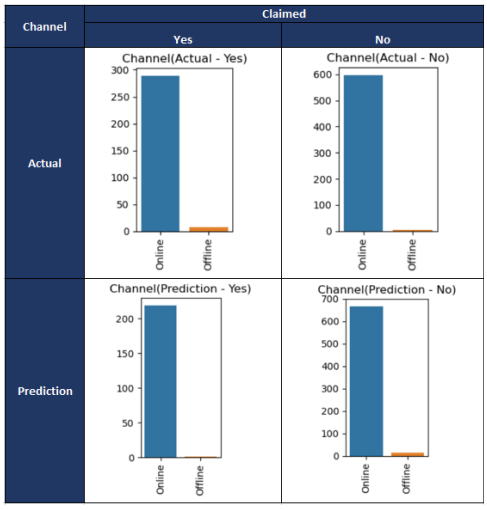
Inference: Basis on these predictions, what are the business insights and recommendations

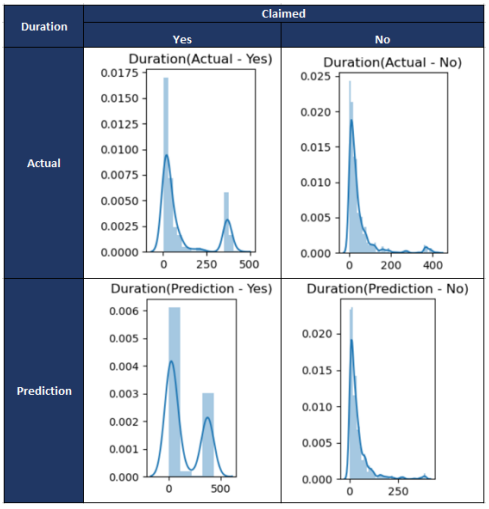
Looking at below figures, we can conclude that the predictions though not very accurate, still help us identify broad trends, like customers from which travel agency tend to claim the most, customers traveling to which destination tend to claim the most, what is the average duration of the customers who tend to file a claim etc.

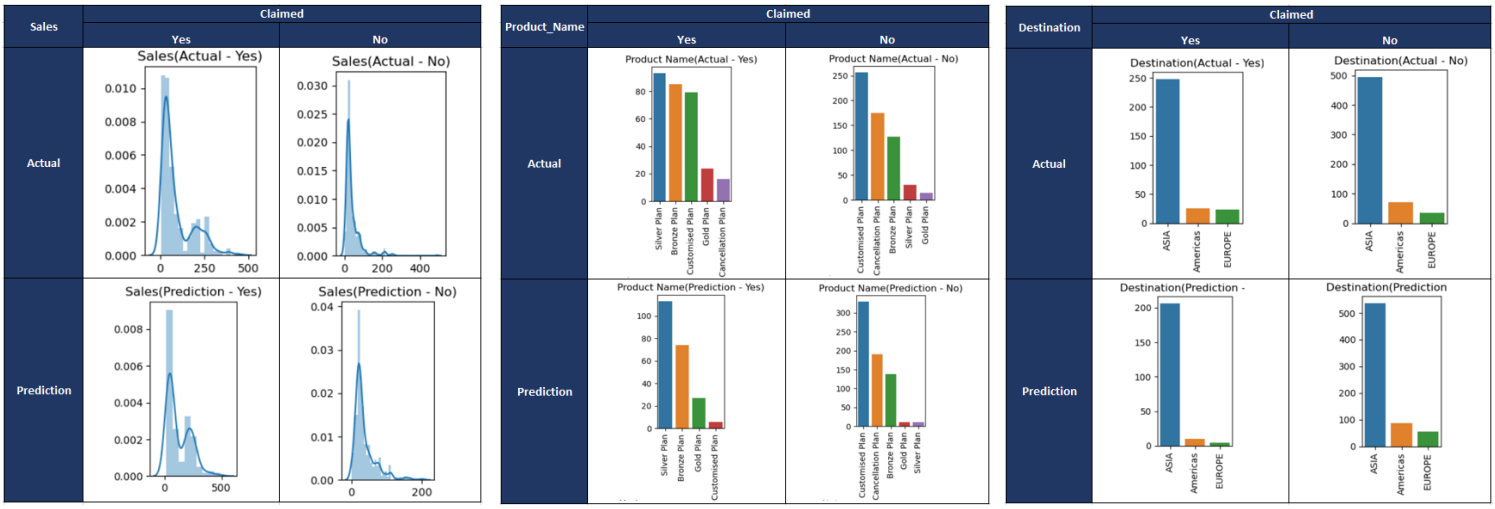












We can clearly see that with the help of our prediction model i.e. Random Forest model, we can to a very large extent predict many crucial decision making factors for the insurance firm.

If we can predict people from which age group tend to claim more, we can then accordingly modify the premium for the insurance policies.

If we are able to predict, customers from which travel agency tend to claim more, and if we find severe irregularities, we can initiate an audit to check if fraudulent claims are being made by a particular agency.

If we are able to predict people traveling to which destination tend to claim more, we can accordingly modify our business strategy to either modify the premiums or provide discounts for locations where claims are low.

If we can predict whether or not a customer will claim an insurance policy we can then also predict the amount of sales that we anticipate to make from people who will claim and from those who won't. We can see the sales pattern between actual and predicted are pretty much similar, hence if we plot the sales chart using the predicted values we can then safely predict what our sales will be.

Similarly multiple usable information can be derived from the production data, based on which business strategies can be modified in real time and for future customers as well.

Business Insights

Recommendations

A few recommendations looking at the predictions done above.

1) We can see claims from agency C2B are very high in number compared to other agencies, it might be in part due to the fact that C2B is mostly involved in insurance for ASIA location where overall claims are anyway on the higher side. However we will still recommend to audit the claims from this agency to rule out any foul play.

2) We can also see, even though the number of overall insurance issued by airlines is much less than the Travel agencies, however when it comes to the claims, more number of people who purchased their insurance from airlines tend to claim. Reason behind such anomaly needs to be investigated.

3) We can see their are two bumps in the graph for duration where people have actually claimed the insurance. We can ignore the short term chunk since the same pattern is observed for people who have not claimed the insurance. Also in general there are more people purchasing short term insurance hence the bump. However the second bump i.e. for long term is abnormal for people who have claimed, no such bump is present for people who have not claimed. Hence people who are buying long term insurance need to scrutinized more by the insurance firm compared to others.

4) Insurance claims are higher from certain locations like ASIA, hence the company should modify the premium accordingly for such locations.

Thank You!