Problem 1: Clustering

Problem Statement:

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

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)

Questions:

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

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

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

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

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

Question 1:

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

Answer:

As the first step we will multiply by the respective factors and convert data to their original values (i.e. values multiplied by their scales mentioned).
The first 10 lines of entries are:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19940	1692	0.8752	6675	37630	325.2	6550
1	15990	1489	0.9064	5363	35820	333.6	5144
2	18950	1642	0.8829	6248	37550	336.8	6148
3	10830	1296	0.8099	5278	26410	518.2	5185
4	17990	1586	0.8992	5890	36940	206.8	5837

The data set has 210 customer (rows) and 7 activities (columns). It can also be observed that all the columns have float data type with no missing values.

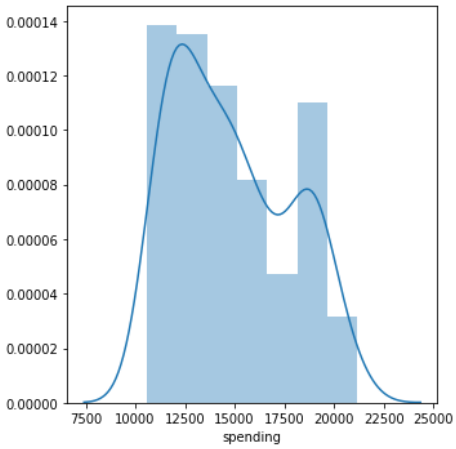
Sr No	Column	Non-Null Count	Dtype
0	spending	210 non-null	float64
1	advance_payments	210 non-null	float64
2	probability_of_full_payment	210 non-null	float64
3	current_balance	210 non-null	float64
4	credit_limit	210 non-null	float64
5	min_payment_amt	210 non-null	float64
6	max_spent_in_single_shopping	210 non-null	float64

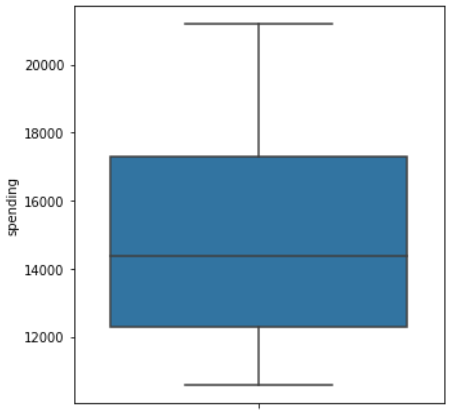
Five number summary of the data set:

	count	mean	std	min	25%	50%	75%	max
spending	210	14847.52381	2909.699431	10590	12270	14355	17305	21180
advance_payments	210	1455.928571	130.595873	1241	1345	1432	1571.5	1725
probability_of_full_payment	210	0.870999	0.023629	0.8081	0.8569	0.87345	0.887775	0.9183
current_balance	210	5628.533333	443.063478	4899	5262.25	5523.5	5979.75	6675
credit_limit	210	32586.04762	3777.144449	26300	29440	32370	35617.5	40330
min_payment_amt	210	370.020095	150.355713	76.51	256.15	359.9	476.875	845.6
max_spent_in_single_shopping	210	5408.071429	491.480499	4519	5045	5223	5877	6550

Univariate Analysis:

spending

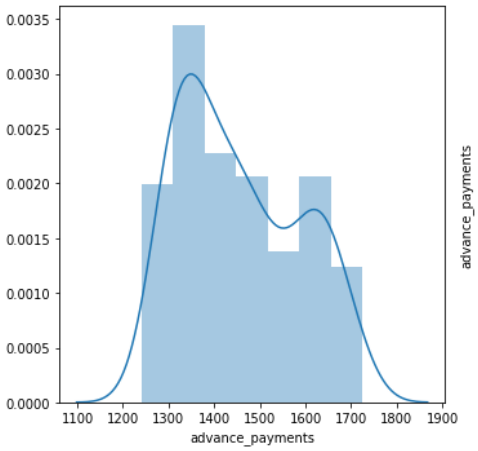


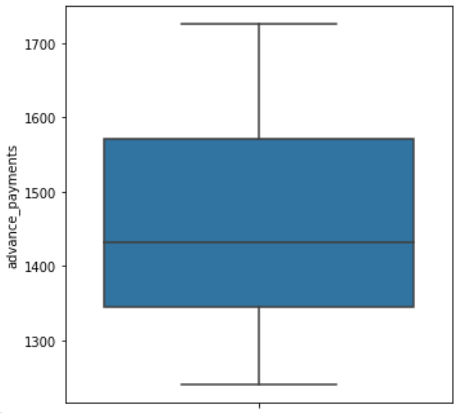


spending is Positive or Right skewed.
The number of outliers in spending is 0

---

advance_payments

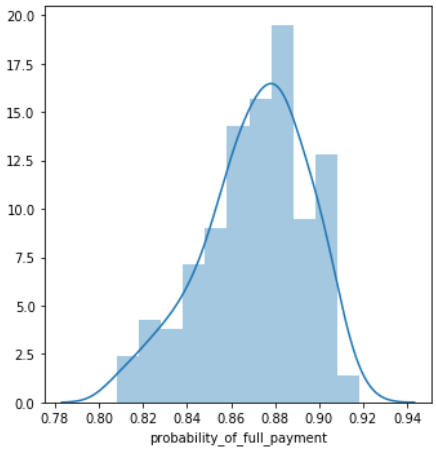


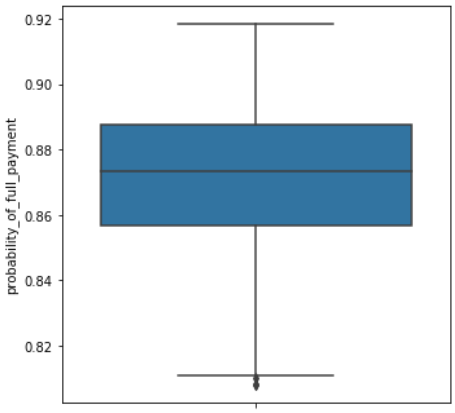


advance_payments is Positive or Right skewed.
The number of outliers in advance_payments is 0

---

probability_of_full_payment

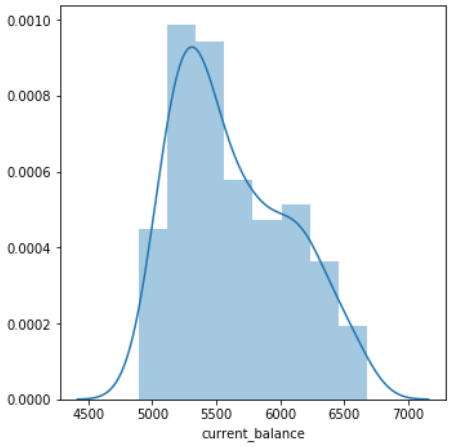


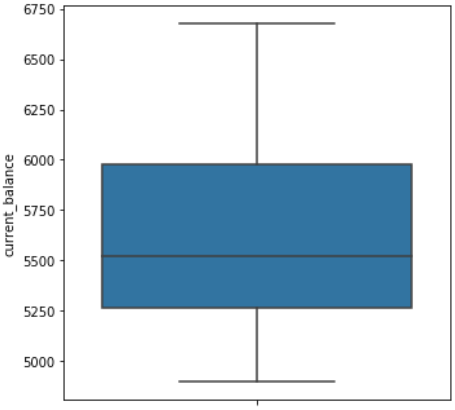


probability_of_full_payment is Negative or Left skewed.
The number of outliers in probability_of_full_payment is 3

---

current_balance

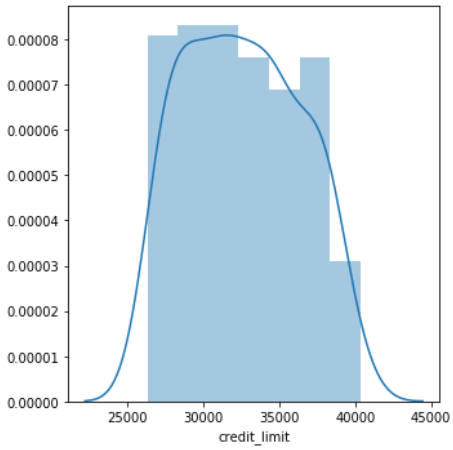


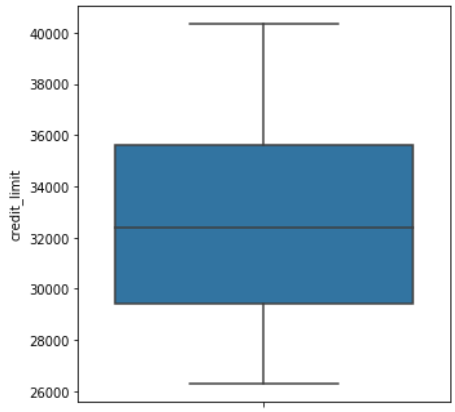


current_balance is Positive or Right skewed.
The number of outliers in current_balance is 0

---

credit_limit

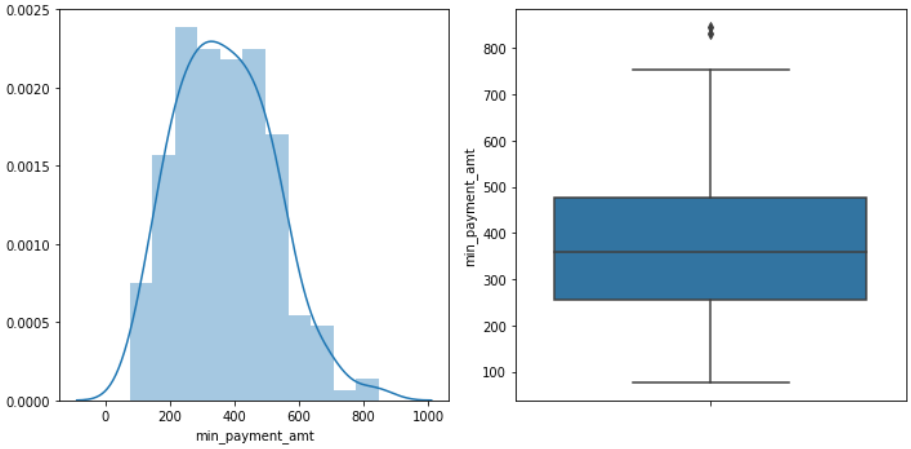




credit_limit is Positive or Right skewed.
The number of outliers in credit_limit is 0

---

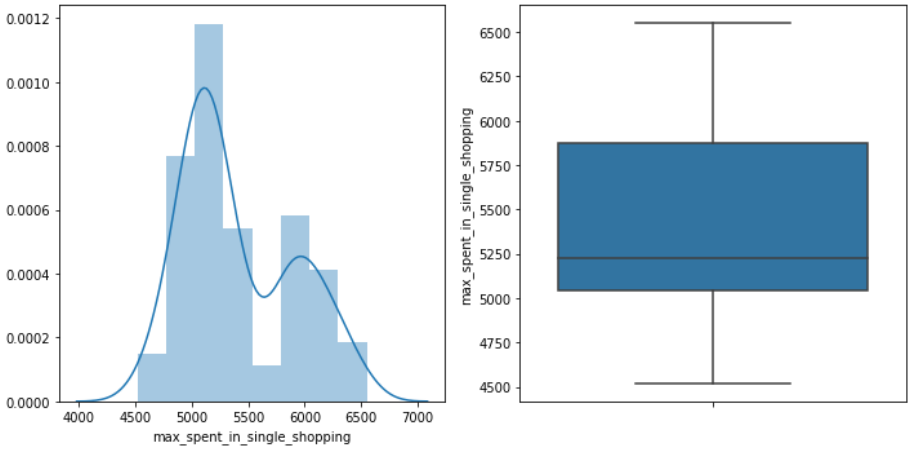
min_payment_amt



min_payment_amt is Positive or Right skewed.
The number of outliers in min_payment_amt is 2

---

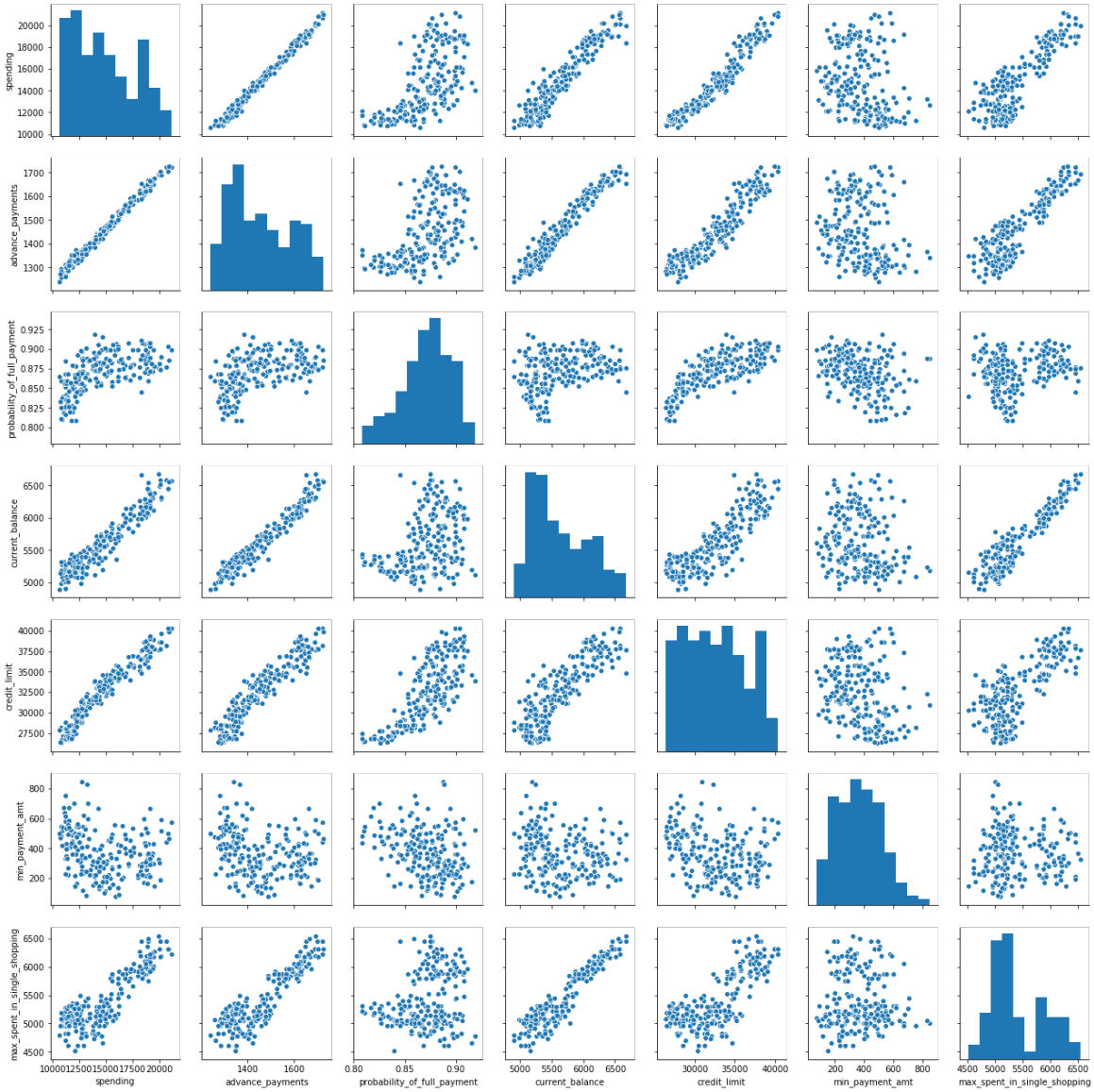
max_spent_in_single_shopping



max_spent_in_single_shopping is Positive or Right skewed.
The number of outliers in max_spent_in_single_shopping is 0

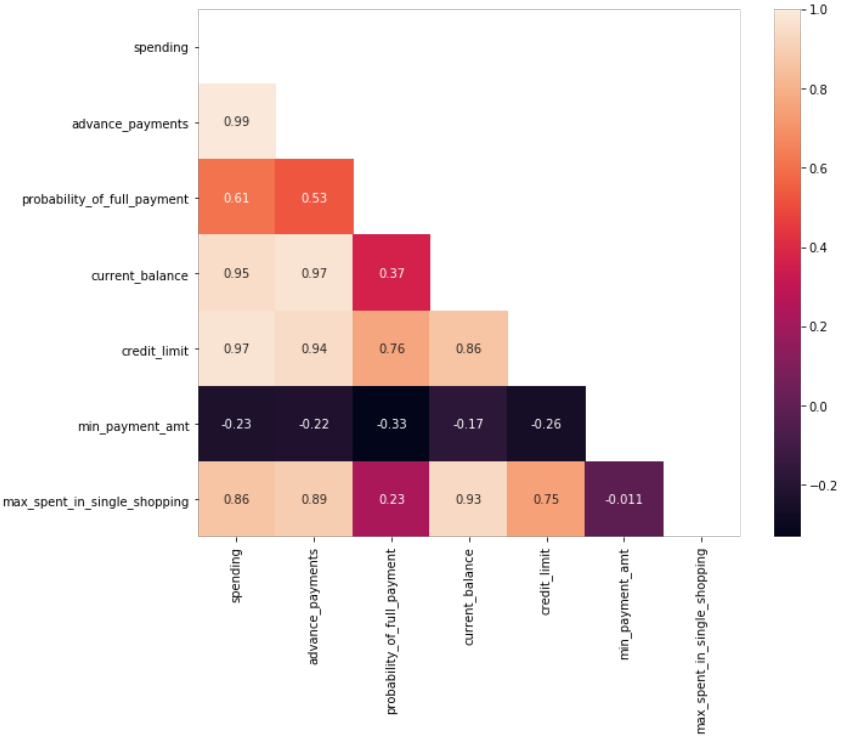
Multivariate Analysis:

Pair Plot



Since there seems to be a good correlation between the variables let us compute the correlation coefficients and plot the heatmap.

Heatmap:



Findings:

1. The data has 210 entries and 7 variables.
2. The columns in the data are:
    - 'spending'
    - 'advance_payments'
    - 'probability_of_full_payment'
    - 'current_balance'
    - 'credit_limit'
    - 'min_payment_amt'
    - 'max_spent_in_single_shopping'
3. All the columns are of 'float' data type.
4. There are 2 variables that have outliers are 'probability_of_full_payment' and 'min_payment_amt'.
5. It can also be observed that the variables are having different magnitudes and so there could be a requirement to scale when we perform weight/distance based models. Ex: Spends go up to 21000 while probability vary from 0.8 to 0.9. This may cause issues in distance/weight based models.
6. The number of outliers in the above 2 variables are:
    - probability_of_full_payment:3
    - min_payment_amt: 2

Since this is a very small portion of the data we may proceed without outlier treatment.

7. The variable, 'probability_of_full_payment', is very slightly left skewed and all other varaibles are very slightly right skewed.
8. There are no null/missing values in the given data set.

9. Variables like 'spending', 'advance_payments' and 'max_spent_in_single_shopping' are multi-modal.
10. From the scatter plot it can be observed that many variables are correlated. To get a detailed picture of the same, let us compute the correlation coefficients and plot the heatmap.
11. The following can be proposed from the heat map:
    - min_payment_amt has a slightly negative correlation with all the variables
    - advance_payments and spending has the highest correlation (close to ~1)
    - current_balance and advance_payments are highly positively correlated
    - current_balance and spending are highly positively correlated
    - credit_limit and advance_payments are highly positively correlated
    - credit_limit and spending are highly positively correlated
    - credit_limit and max_spent_in_single_shopping are highly positively correlated

On the whole the variables are highly positively correlated except for min_payment_amt and the other variables

Question 2:

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

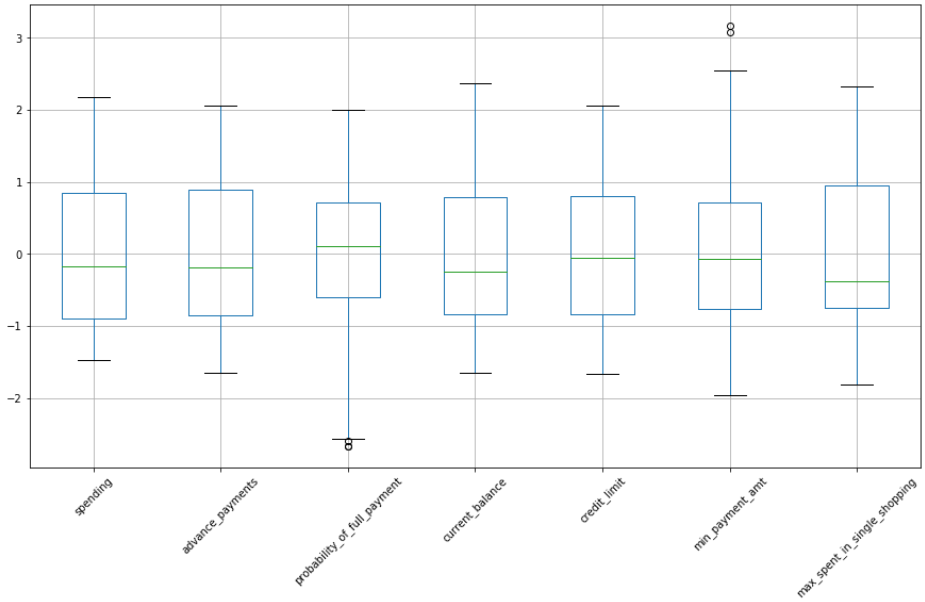
Answer:

Yes, scaling is necessary. We need to scale the data because clustering methods is a distance based model and if the values are taken on different magnitudes then the one with the higher magnitude will affect the distance more than the one with lower. Some variables are in the range 0.8-0.9 and some in the range 10,000-21,000 so scaling is the best way we can give equal weightage to all variables.

Ex:

If we have a data of weight (in lbs) and height (in cms) the two will be comparable. However, if we have weight (in lbs) and height (in feet) then the weight will affect the distance more than the height. To avoid this problem we have to scale.

In this case we will apply standard scaling and proceed.
Post scaling the data looks as shown below:

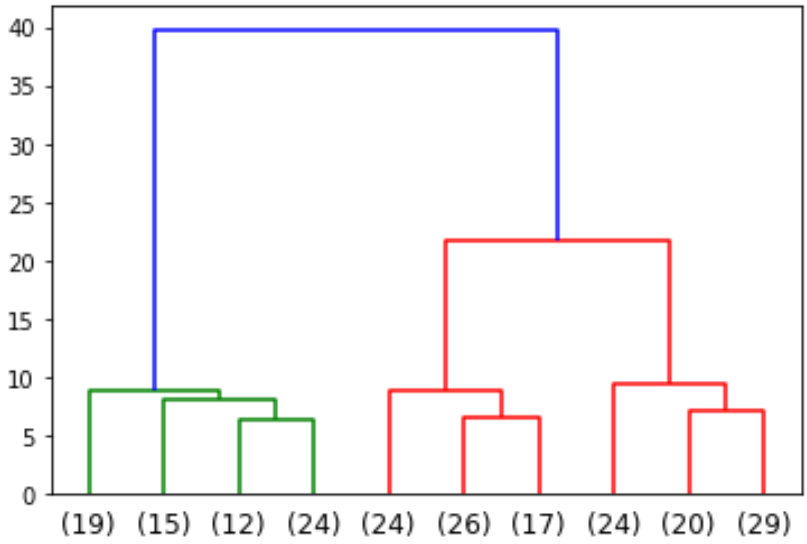


Question 3:

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

Answer:

On applying hierarchical clustering on the scaled data. The following dendrogram was obtained for the last 10 splits in the clustering:



Though the system directs us to proceed with 2 clusters, it is to be noted that the vertical line between 10 and 20 is long enough. So visually we can propose to have 3 clusters.

Cluster profiling we can observe the following data:

cluster_hierarchical	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Count
1	18371.42857	1614.542857	0.8844	6158.171429	36846.28571	363.915714	6017.371429	70
2	11872.38806	1325.701493	0.848072	5238.940299	28485.37313	494.943284	5122.208955	67
3	14199.0411	1423.356164	0.87919	5478.232877	32264.52055	261.218082	5086.178082	73

The clusters can be classified on the basis of current_balance or credit_limit:

- Cluster 1 : High
- Cluster 3 : Medium
- Cluster 2 : Low

However, on trying with clusters as 4 we get an interesting perspective of the data.

cluster_hierarchical	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Count
1	18371.429	1614.542857	0.8844	6158.171429	36846.28571	363.915714	6017.371429	70
2	11872.388	1325.701493	0.848072	5238.940299	28485.37313	494.943284	5122.208955	67
3	12798.75	1356.541667	0.873313	5254.458333	30400	237.112917	4894	24
4	14884.898	1456.081633	0.882069	5587.836735	33177.7551	273.024694	5180.306122	49

The clusters can be classified on the basis of spending:

- Cluster 1 : High
- Cluster 4 : Medium
- Cluster 3 : Low 2
- Cluster 2 : Low 1

$$\text{Low 1}_{\text{Spending}} < \text{Low 2}_{\text{Spending}} < \text{Medium}_{\text{Spending}} < \text{High}_{\text{Spending}}$$

Some promotions are:

- The clusters low 1 and low 2 are almost the same in terms of current_balance etc but the probability_of_full_payment and min_payment_amount are a lot different. Since low 2 has a good probability we need to increase their spends by giving offers on single large spends as that is also less in this cluster when compared to low 1. This will boost the total spends.
- Low 2 also has a very small fraction of customers in spite of the exceptionally good probability and credit limit.
- Medium needs to be carefully gauged as there are high single spends and the current_balance is quite comparable to the low clusters. So the bank must look out for defaulters.
- The advance_payment for high is more and that could be a reason for less customers in that cluster.

Question 4:

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

Answer:

Unlike the hierarchical clustering where we figure out the number of clusters based on the dendrogram, in k-means clustering we have to mention the number of clusters even before hand. So we will plot the WSS plot to find the optimum number of clusters.

The WSS plot for the first 10 clusters is shown below:

The WSS value and the silhouette scores corresponding to the clusters is shown below:

Number of Clusters	WSS	Silhouette Score
1	1470	NA
2	659.171754	0.465772
3	430.658973	0.400727
4	371.301721	0.327574
5	327.949158	0.279494
6	289.266818	0.291719
7	262.267556	0.284217
8	240.940047	0.268771
9	221.203492	0.258967
10	208.664615	0.236058

From the WSS plot we can say that 3 or 4 clusters is an optimum split. However, we can use the WSS score to confirm the same and pick one amongst 3 and 4. From the table the drop in WSS value from 3 to 4 is less and hence, the split may not be very meaningful.

Hence, we will proceed with 3 clusters.

Question 5:

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

Answer:

The cluster profiling for 3 clusters is shown below:

cluster_kmean	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	sil_width	Count
0	18495.37313	1620.343284	0.88421	6175.686567	36975.37313	363.237313	6041.701493	0.468772	67
1	14437.88732	1433.774648	0.881597	5514.577465	32592.25352	270.734085	5120.802817	0.339816	71
2	11856.94444	1324.777778	0.848253	5231.75	28495.41667	474.238889	5101.722222	0.397473	72

The clusters can be classified on the basis of current_balance or credit_limit:

- Cluster 0 : High
- Cluster 2 : Medium
- Cluster 1 : Low

The min silhouette width in the data is 0.00271, since it is positive, we can conclude that there is no data that is wrongly split into a different cluster.

The silhouette score for the data when split into 3 clusters is 0.4007.

The following business recommendations can be provided for the target customers:

- The medium cluster and the high cluster have almost same probability_of_full_payment but have a very huge difference in min_payment_amt. Since we know that high cluster will pay back full payment, we can decrease the min payment somewhere close to that of medium cluster so that there is a larger number of enrolments to get the card.
- The max_spent_in_single_shopping is almost same for the low and medium cluster even though there is a huge difference in credit limit. We can promote offers on payments that exceed a certain huge number in order to increase the max spent in single shopping. This will eventually increase the spending in return as they both are highly correlated.
- The low spending cluster has tendencies to swipe on huge single shopping and they also have a comparatively low probability of full payment. There is a risk of defaulters in the cluster that the bank must watch out for. To counter this risk we can promote them to invest in other investment plans so that we have a financial relation more than just the credit, this will also increase the cash flow for the bank.
- The high cluster has the highest credit_limit and also has a good probability_of_full_payment. However it has the lowest count of customers in this cluster. We need to focus more on marketing for this section of customers as they make huge transactions and they have a very good probability_of_full_payment as well. An easy way to do so is to increase the advance_payments as spending is highly correlated to spends.
- The medium and high cluster's spend is almost 3 times the amount spent for max_spent_in_single_shopping. However the low cluster spends only twice. The reason for this could be that the local/daily vendor do not have the facility to use the cards. So we will have to promote the small vendors to start using POS machines to increase these petty swipes, which will eventually increase the total spends.
- Adding to the above point. The low cluster is the cluster that pays the highest min_payment_amt and they can be rewarded with cash backs or reward points for this so that they spend more.

- All other factors of high are in line with low and medium clusters. However one of the reasons for less customers in this cluster could be because of the out of proportion high advance_payments. A decrease in this may lead to good increase in the number of customers in this cluster.

Note:

The data can be clustered into 3 or 4 meaningful clusters. For this situation we will stick to 3 clusters. However if the bank wants to focus more on the low spending clusters then we can go with 4 clusters and provide a more detailed insight on the low and medium spending clusters.

Problem 2: CART-RF-ANN

Problem Statement:

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

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)

Questions:

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

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

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

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

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

Question 1:

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

Answer:

The data contains 3000 entries (rows) and 10 variables (columns). The first 5 entries of the data is shown below:

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0	Online	34	20	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0	Online	4	26	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.3	Online	53	18	Bronze Plan	ASIA

The data type of the data set is:

#	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

The summary of the categorical variables is shown below:

	count	unique	top	freq
Agency_Code	3000	4	EPX	1365
Type	3000	2	Travel Agen	1837
Claimed	3000	2	No	2076
Channel	3000	2	Online	2954
Product Name	3000	5	Customised	1136
Destination	3000	3	ASIA	2465

The summary of the continuous variables is shown below:

	count	mean	std	min	25%	50%	75%	max
Age	3000	38.091	10.463518	8	32	36	42	84
Commision	3000	14.529203	25.481455	0	0	4.63	17.235	210.21
Duration	3000	70.001333	134.053313	-1	11	26.5	63	4580
Sales	3000	60.249913	70.733954	0	20	33	69	539

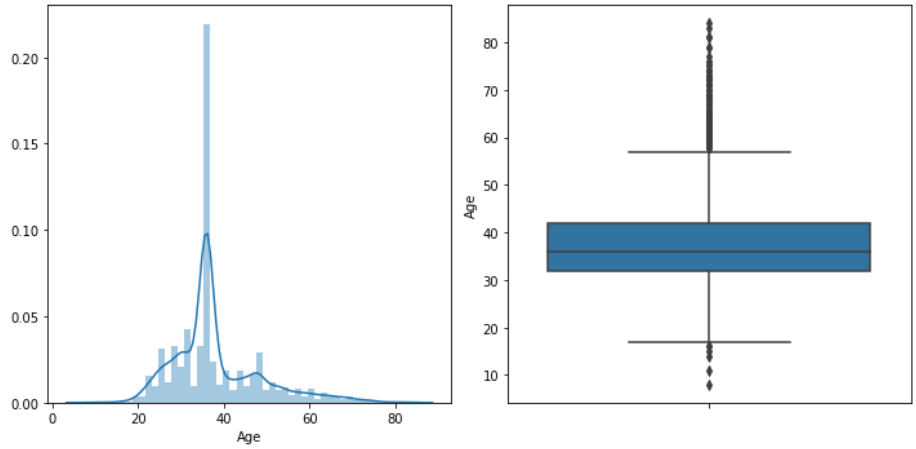
From the above table it can be observed that Duration (in days) variable has a negative value which is not possible. Similarly there is an entry which is of 4580 days (approx. 12 years), this too is practically not possible so we will drop this value as well.

We have 286 duplicate rows (repetitive entries). Though these values will assign a weightage to the repeated values, since we don't have a unique column like Customer name or Customer ID etc we cannot conclude if these are erroneously duplicated. So we will proceed retaining these values.

On the whole we have lost 2 rows out of 3000 rows (less than ~0.06%).

Univariate Analysis:
(Continuous Variables)

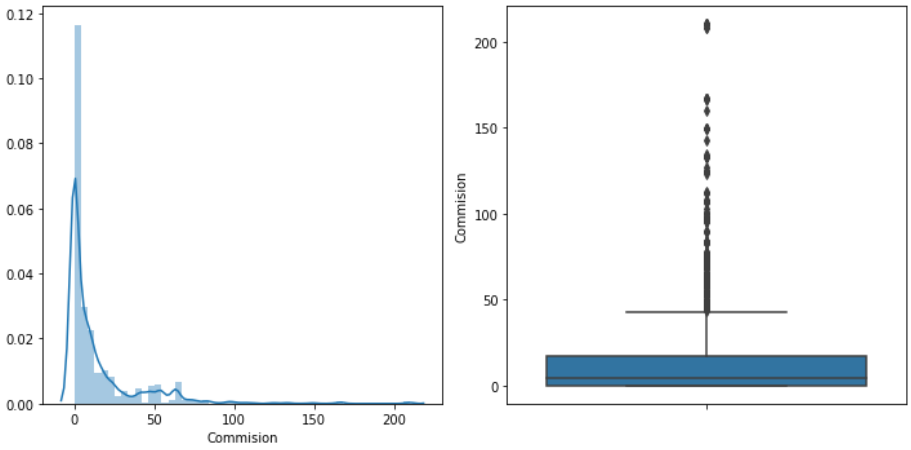
Age



Age is Positive or Right skewed.

The number of outliers in Age is 204 (6.80% of total data)

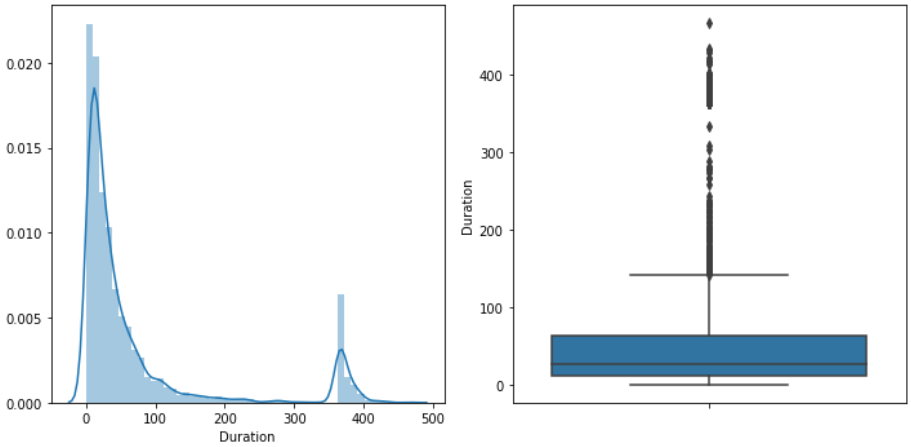
Commision



Commision is Positive or Right skewed.

The number of outliers in Commision is 362 (12.07% of total data)

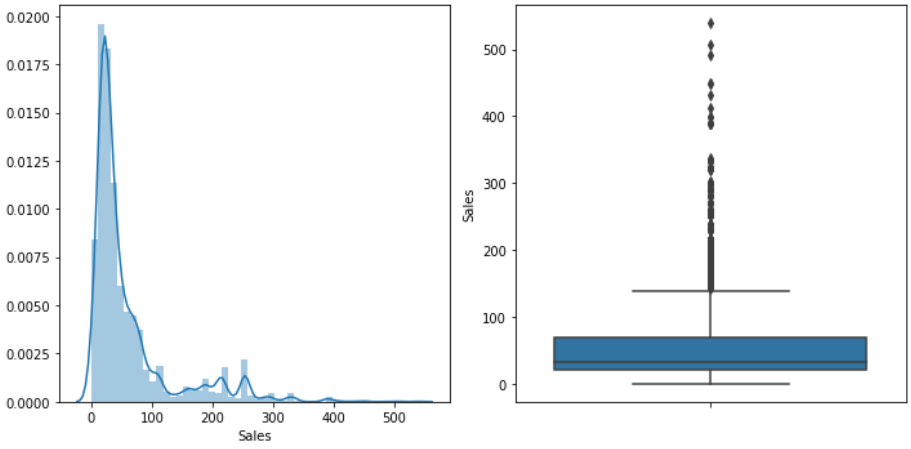
Duration



Duration is Positive or Right skewed.

The number of outliers in Duration is 381 (12.71% of total data)

Sales



Sales is Positive or Right skewed.

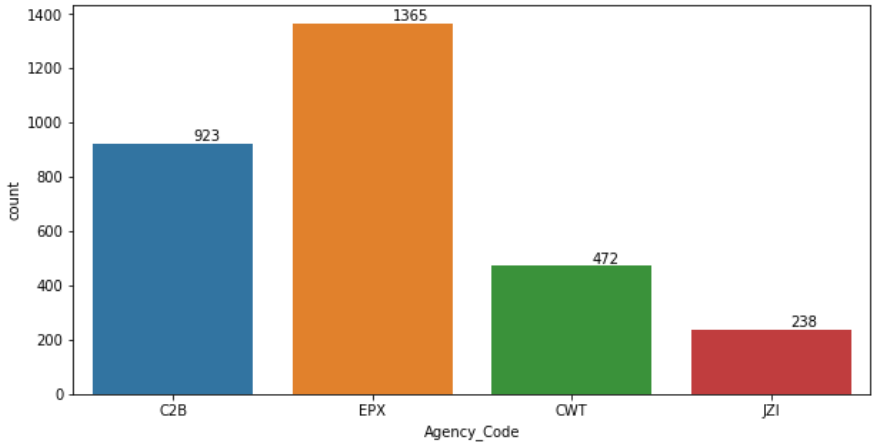
The number of outliers in Sales is 353 (11.77% of total data)

(Categorical Variables)

Agency_Code

```
The number of unique entries in the column Agency_Code : 4

The entry with the highest frequency in Agency_Code : EPX
```

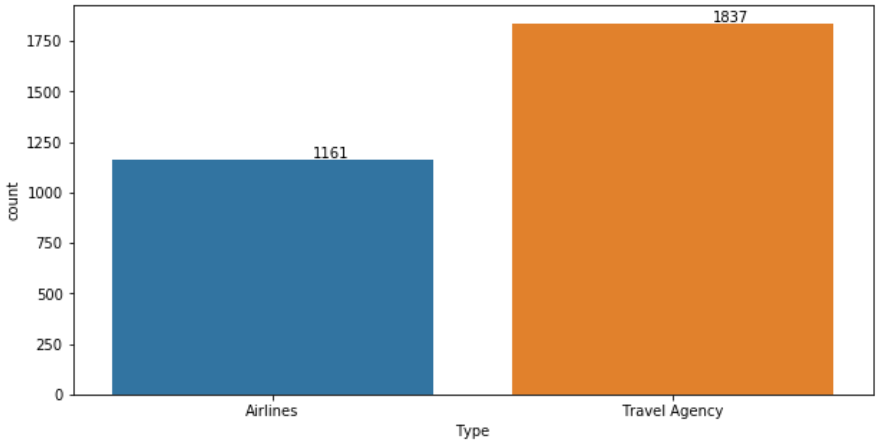



```
Percentage share:
EPX    45.530354
C2B    30.787191
CWT    15.743829
JZI     7.938626
```

Type

```
The number of unique entries in the column Type : 2

The entry with the highest frequency in Type : Travel Agency
```

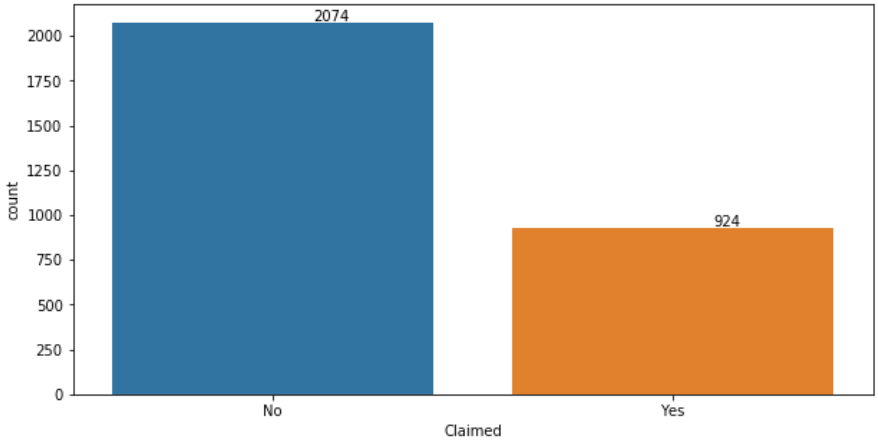



```
Percentage share:
Travel Agency    61.274183
Airlines         38.725817
```

Claimed

The number of unique entries in the column Claimed : 2

The entry with the highest frequency in Claimed : No

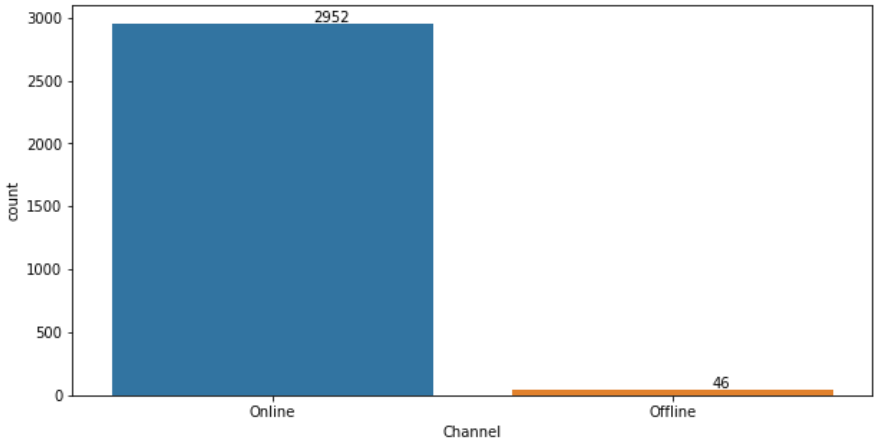


Percentage share:
No 69.179453
Yes 30.820547

Channel

The number of unique entries in the column Channel : 2

The entry with the highest frequency in Channel : Online

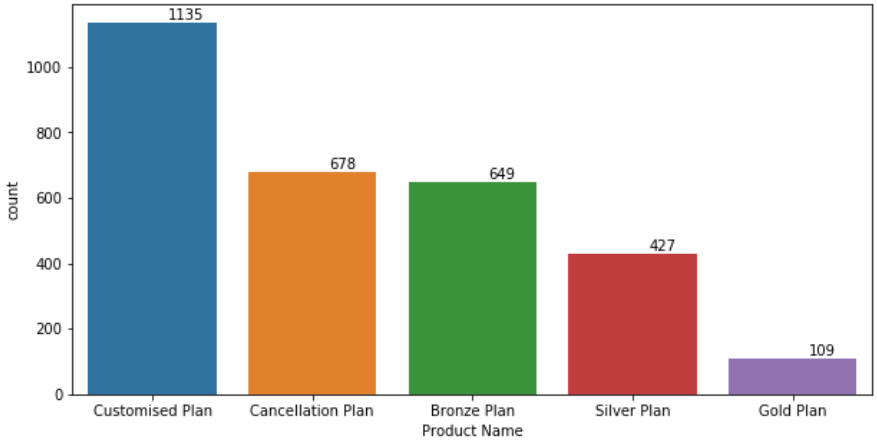


Percentage share:
Online 98.465644
Offline 1.534356

Product Name

The number of unique entries in the column Product Name : 5

The entry with the highest frequency in Product Name : Customised Plan



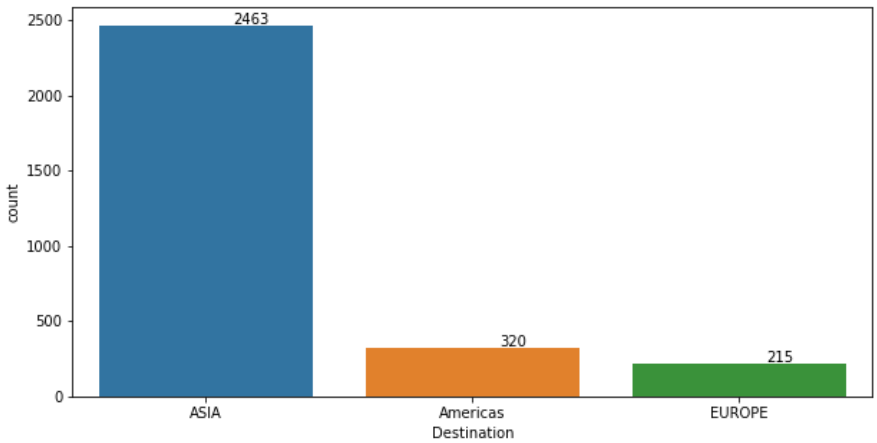
Percentage share:

Customised Plan	37.858572
Cancellation Plan	22.615077
Bronze Plan	21.647765
Silver Plan	14.242829
Gold Plan	3.635757

Destination

The number of unique entries in the column Destination : 3

The entry with the highest frequency in Destination : ASIA

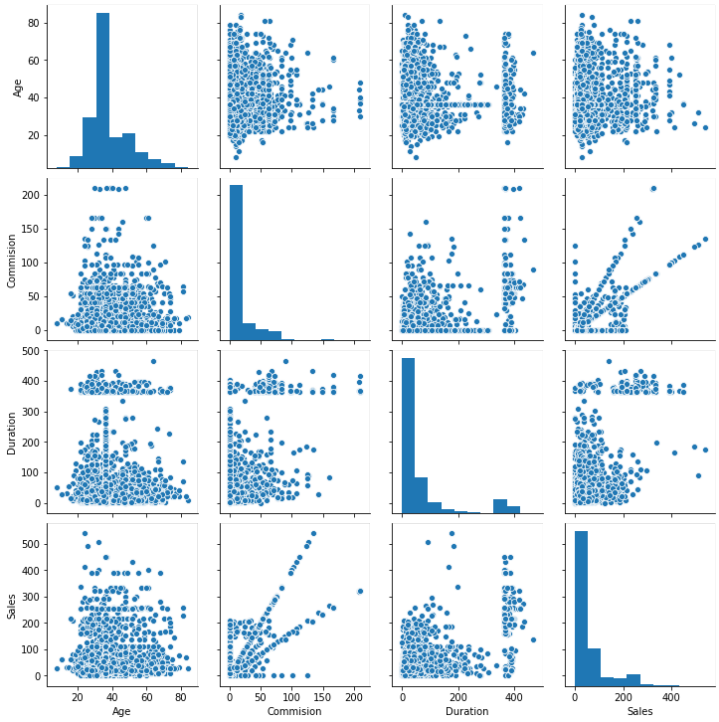


Percentage share:

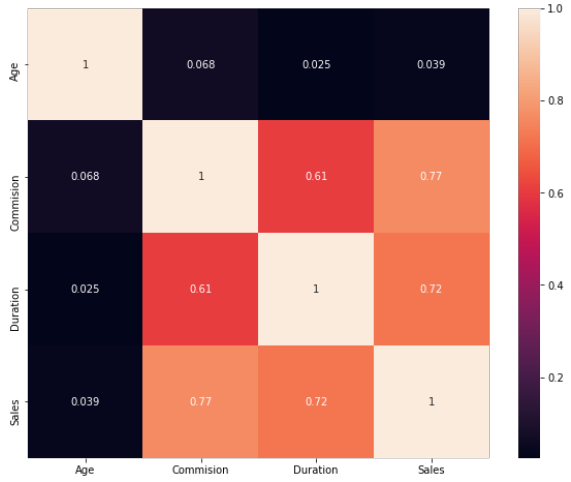
ASIA	82.154770
Americas	10.673783
EUROPE	7.171448

Multivariate Analysis:

Scatter plot for the continuous variables:

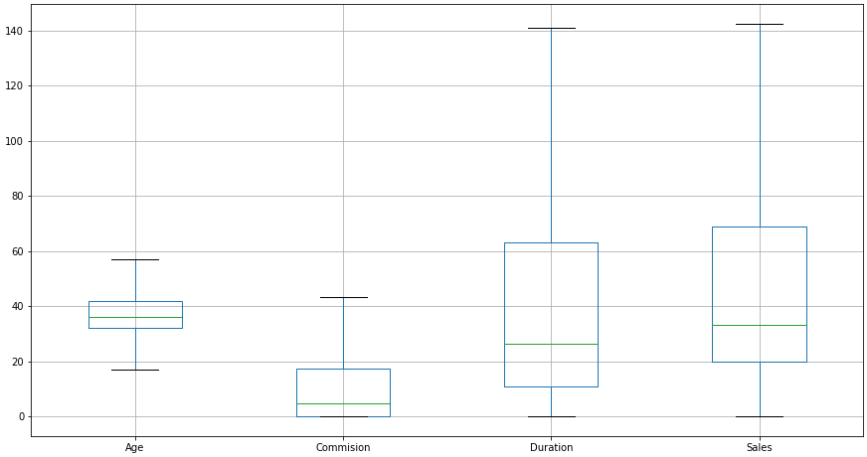


Heat map to visualize correlation between variables:



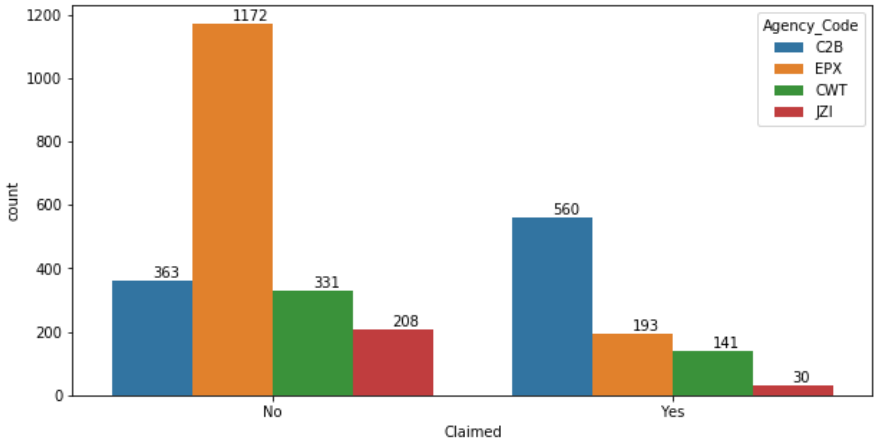
Since we have many outliers we will treat outliers and then proceed with model building.

Post outlier treatment the box plot for the data looks as shown below:

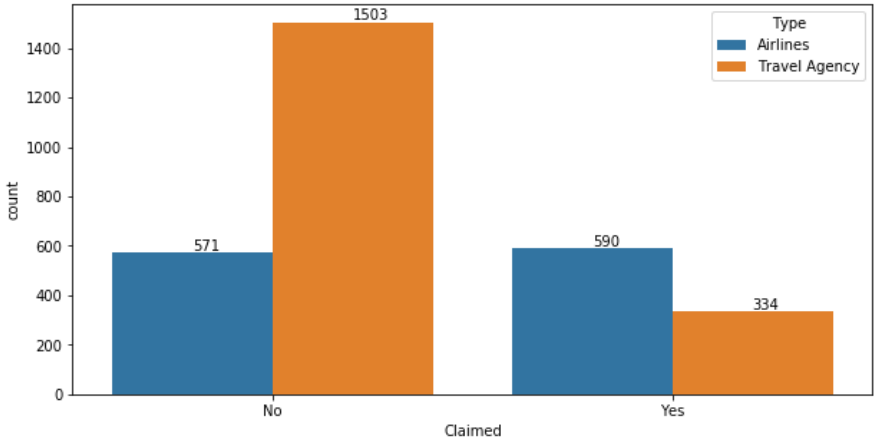


Since ‘Claimed’ is the dependent variable we will compare other variables against this to obtain some interesting insights.

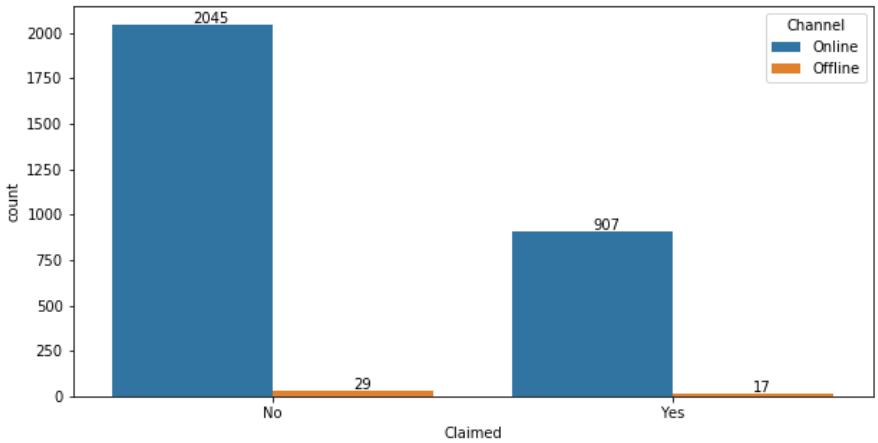
Agency_Code



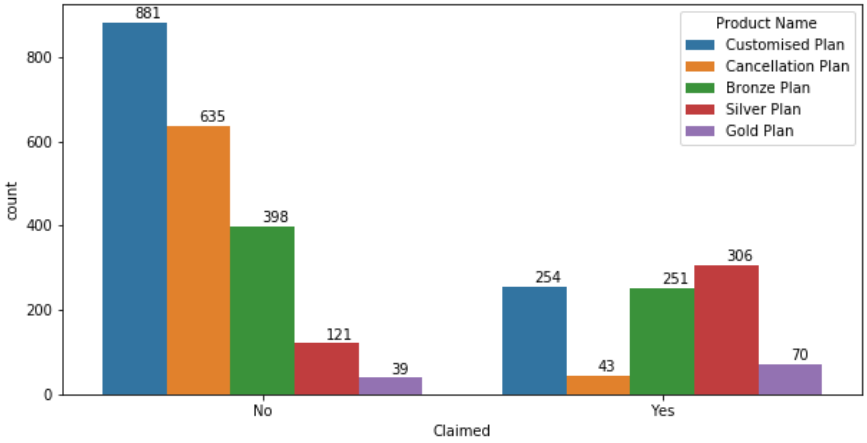
Type



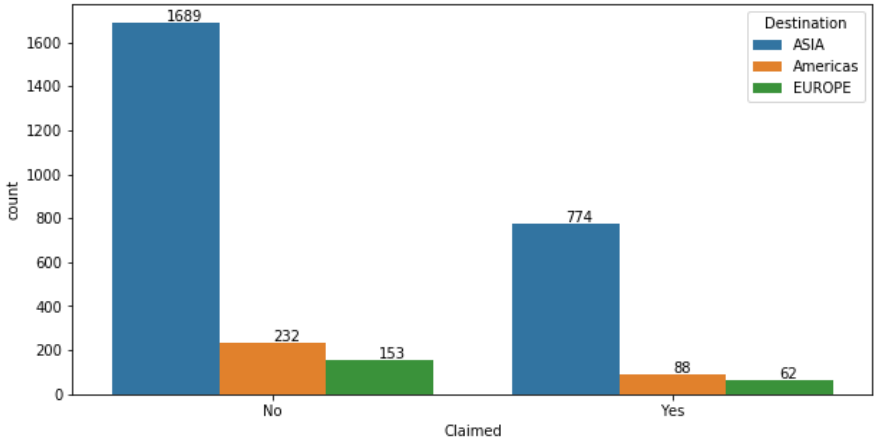
Channel



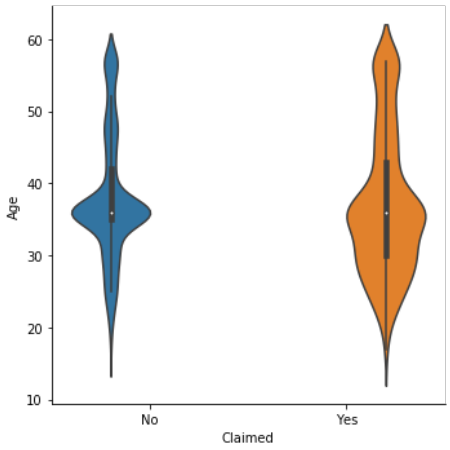
Product Name



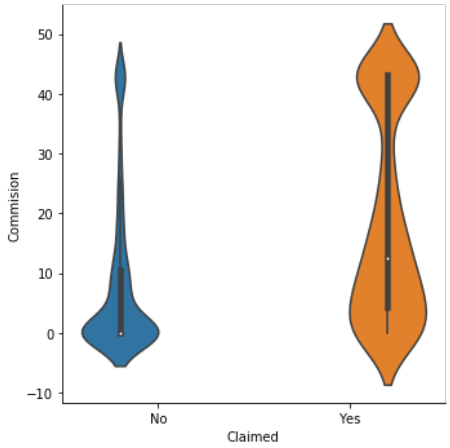
Destination



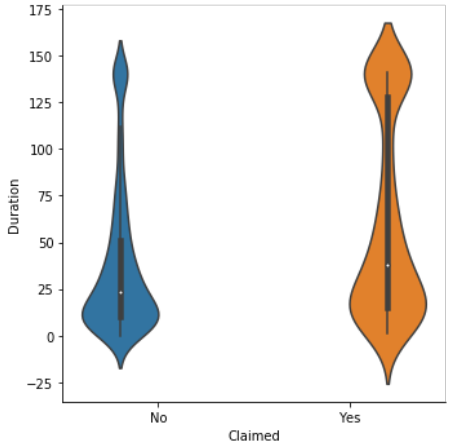
Age



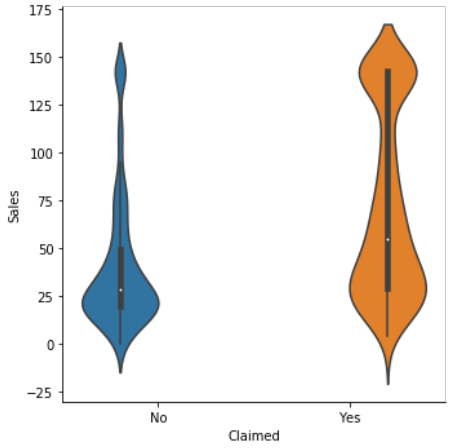
Commision



Duration



Sales



Findings:

- The data consists considerably high number of observations of customers who are in their 30's. This could be mainly due to the fact that this group travels the most compared to all other age groups.
- The data consists of many observations that collect low commission. The reason for this could be because the correlation of commission with duration of travel or/and

sales, as these 2 variables also have the same characteristics. These 3 variables are severely right skewed.

- The Agencies EPX and C2B contribute to more than 75% of the total observations.
- The ratio of claimed to not claimed in the data is approximately 3:7.
- Almost 98.5% of the total bookings are done online.
- Cancellation plan is the highest bought product. It covers almost 38% of the total products.
- Asia is the most travelled destination in the provided data set.
- From the plot of categorical data with the Claimed variable, the following can be inferred:
    - Though EPX has a high market share, they have the highest not claimed. Whereas, CB2 has a lower market share compared to EPX, but still has a higher number of claims.
    - The travel agency type insurance have higher possibility of not claimed that the airline type. Airline type has a higher chance of being claimed.
    - When a plan is customized there is a high number of no claim. However the Silver plan has a greater claim compared to the other products.
    - Since Asia is the most travelled location, the claims are also proportionally high. The similar case with other regions as well. So we can say that the destination of travel doesn't play a very pivotal role in the claims.
- From the plot of continuous data with the Claimed variable, the following can be inferred:
    - The age group of 30's have a high numbers of not claiming and not claiming. This is mainly because of the high number of observations present in that age group. However, customers in the early 30's and lesser have an unusual practice of claiming. The same is observed for late 30's and early 40's.
    - For larger values of Sales, Duration and Commission, there is a sudden increase in the number of claims observed.

Question 2:

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

Answer:

Since the CART, Random Forest and ANN models cannot interpret the nominal codes we will have to convert them into numeric equivalents.

Column : Agency_Code

Value	Code
C2B	0
EPX	2
CWT	1
JZI	3

Column : Claimed

Value	Code
No	0
Yes	1

Column : Product Name

Value	Code
Customised Plan	2
Cancellation Plan	1
Bronze Plan	0
Silver Plan	4
Gold Plan	3

Column : Type

Value	Code
Airlines	0
Travel Agenc	1

Column : Channel

Value	Code
Online	1
Offline	0

Column : Destination

Value	Code
ASIA	0
Americas	1
EUROPE	2

We will proceed with splitting the data into train and test at 7:3 ratio respectively.

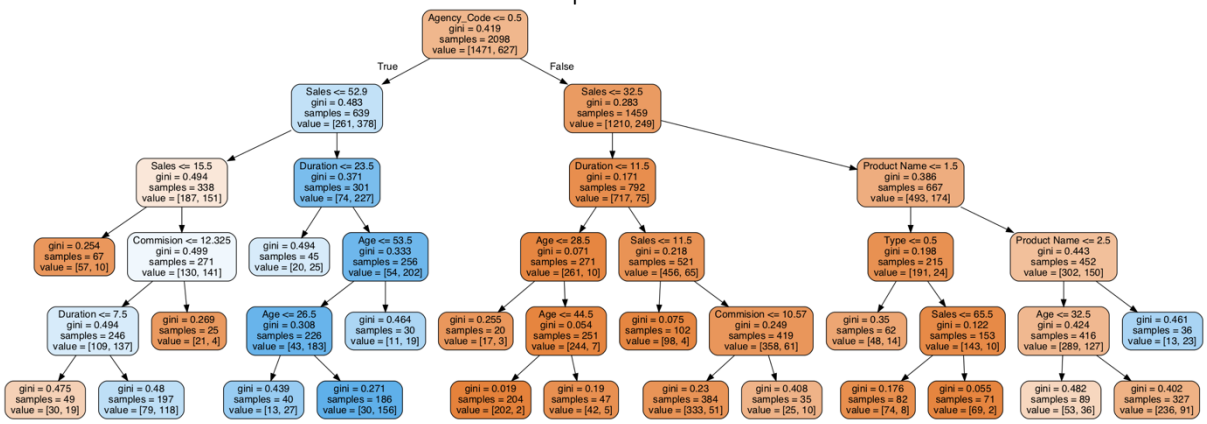
For all the models we need to perform model tunning by passing in some hyper parameters. These hyper parameters are many in number and different combinations. We will pass a few values and run the grid search. If for any parameter the lower or upper value is met then we will go in that direction to further explore values until we hit an equilibrium point.

CART Model:

Upon grid search the best hyper parameters for the CART model obtained is as shown below:

```
{     'criterion': 'gini',
      'max_depth': 5,
      'min_samples_leaf': 20,
      'min_samples_split': 150}
```

The tree thus obtained from the above best parameters is shown below:



We will create a CART classifier with the above hyper parameters.

Random Forest Model:

Upon grid search the best hyper parameters for the random forest model obtained is shown below:

```
{     'max_depth': 5,
      'max_features': 4,
      'min_samples_leaf': 20,
      'min_samples_split': 150,
      'n_estimators': 100}
```

We will create a random forest classifier with the above hyper parameters.

Artificial Neural Network:

Upon grid search the best hyper parameters for the artificial neural network obtained is shown below:

```
{     'hidden_layer_sizes': 50,
      'max_iter': 20,
      'solver': 'adam',
      'tol': 0.01}
```

We will create a neural network classifier with the above hyper parameters.

Question 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

Answer:

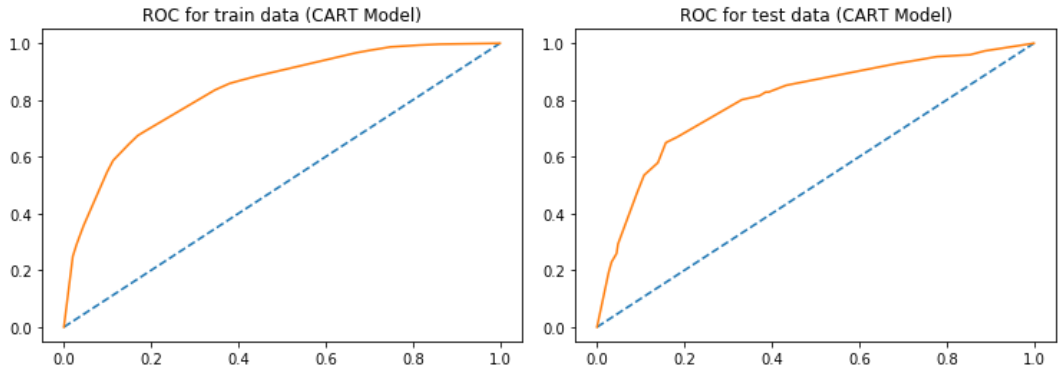
We will first predict the output for train data as input and the we will use the test data. We will compare the predicted outputs to the actual outputs. We obtain the confusion matrix and the classification report. We will plot the ROC to view the area under the plot.

CART Model:

Feature Importance

	Imp
Agency_Code	0.573956
Sales	0.244626
Product Name	0.078409
Duration	0.038015
Commision	0.031535
Age	0.025179
Type	0.008282
Channel	0
Destination	0

ROC



Confusion Matrix

Train Data

		Predicted	
		0 (Not Claimed)	**1 (Claimed)**
Actual	**0 (Not Calimed)**	1305	166
	1 (Claimed)	259	368

Test Data

		Predicted	
		0 (Not Claimed)	**1 (Claimed)**
Actual	**0 (Not Claimed)**	538	65
	1 (Claimed)	138	159

Classification Report

	CART (Train)	CART (Test)
Accuracy	80	77
AUC	83	80
Recall	59	54
Precision	69	71
F1 Score	63	61

Remarks:

- Agency plays a very pivotal role in the prediction of claim status
- The ROC is quite blunt and AUC is around 80% which is not a very bad score but neither very good.

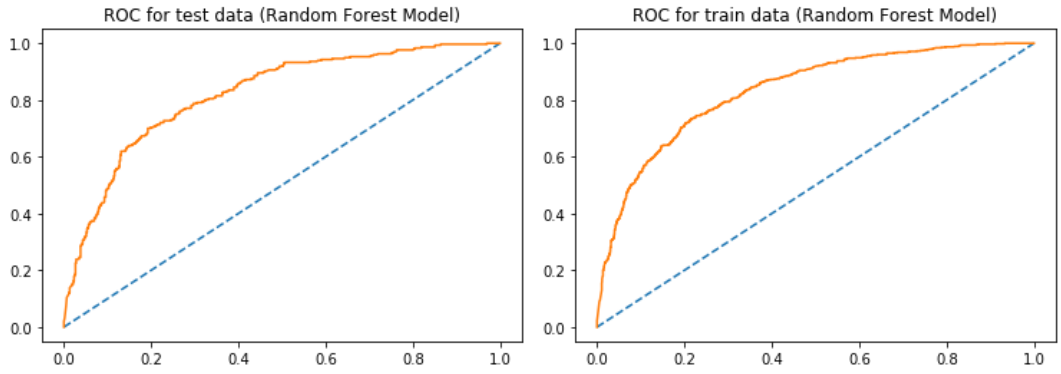
- The accuracy is pretty good for train and test predictions.
- The recall values are extremely low (roughly 55%)

Random Forest Model:

Feature Importance

	Imp
Agency_Code	0.359942
Product Name	0.206549
Sales	0.18202
Commision	0.112921
Type	0.06426
Duration	0.0467
Age	0.021696
Destination	0.00555
Channel	0.000362

ROC



Confusion Matrix

Train Data

		Predicted	
		0 (Not Claimed)	**1 (Claimed)**
Actual	**0 (Not Claimed)**	1343	128
	1 (Claimed)	307	320

Test Data

		Predicted	
		0 (Not Claimed)	**1 (Claimed)**
Actual	**0 (Not Claimed)**	545	58
	1 (Claimed)	156	141

Classification Report

	Random Forest (Train)	Random Forest (Test)
Accuracy	79	76
AUC	83	82
Recall	51	47
Precision	71	71
F1 Score	60	57

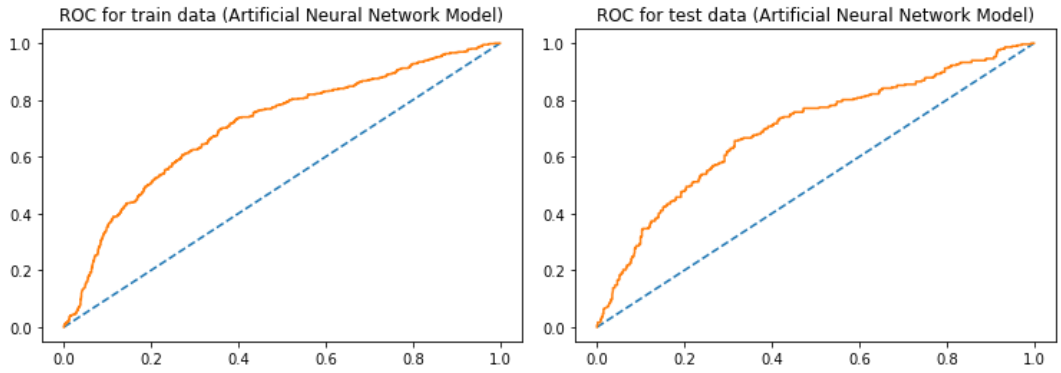
Remarks:

- Agency code and product name play a pivotal role in the prediction of claim status
- The ROC is quite blunt, and the score is around 82% for test data which is fairly good.
- The accuracy is 76% for test and 79% for train which means the model is not over fitted.
- The Precision is 71% and the value is fairly good.
- The recall rate however remains to be very low.

Artificial Neural Network (ANN):

We do not get feature importance in ANN model.

ROC



Confusion Matrix

Train Data

		Predicted	
		0 (Not Claimed)	1 (Claimed)
Actual	0 (Not Claimed)	1162	309
	1 (Claimed)	298	329

Test Data

		Predicted	
		0 (Not Claimed)	1 (Claimed)
Actual	0 (Not Claimed)	481	122
	1 (Claimed)	154	143

Classification Report

	Artificial Neural Network (Train)	Artificial Neural Network (Test)
Accuracy	71	77
AUC	71	69
Recall	52	54
Precision	52	71
F1 Score	52	61

Remarks:

- The model is not over fitted.
- The ROC is very blunt and flat, the AUC score is also low, 71% for train and 69% for test.
- The accuracy for test is at 77% which is quite low.
- The important features in this situation like recall and precision are very low.

Question 4:

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

Answer:

The summary of the parameters of all the models is shown below:

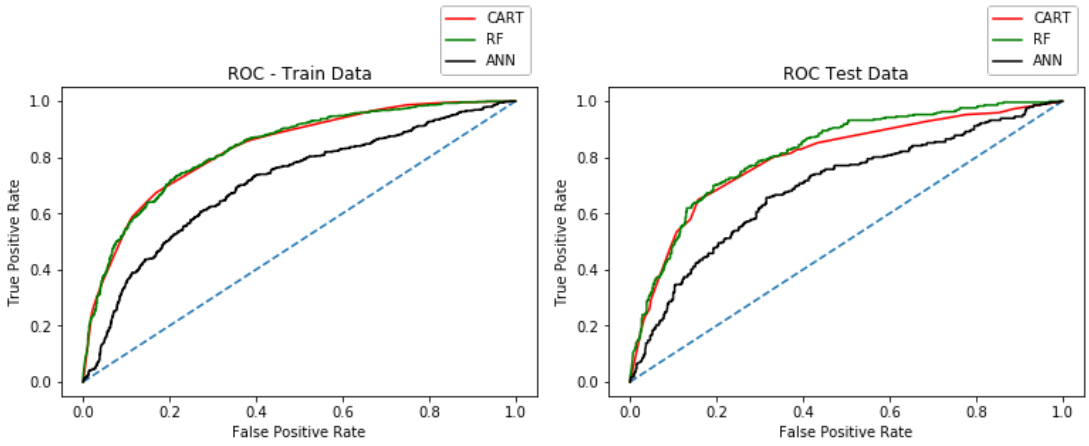
	CART (Train)	CART (Test)	Random Forest (Train)	Random Forest (Test)	Artificial Neural Network (Train)	Artificial Neural Network (Test)
Accuracy	80	77	79	76	71	77
AUC	83	80	83	82	71	69
Recall	59	54	51	47	52	54
Precision	69	71	71	71	52	71
F1 Score	63	61	60	57	52	61

Let us split the train and test and view them separately:

Train Data	CART (Train)	Random Forest (Train)	Artificial Neural Network (Train)
Accuracy	80	79	71
AUC	83	83	71
Recall	59	51	52
Precision	69	71	52
F1 Score	63	60	52

Test Data	CART (Test)	Random Forest (Test)	Artificial Neural Network (Test)
Accuracy	77	76	77
AUC	80	82	69
Recall	54	47	54
Precision	71	71	71
F1 Score	61	57	61

ROC



From the performance on train data, we can clearly see that the CART model has performed well compared to the other two.

From ROC, we can see that the CART model and Random Forest model perform equally good for the train data. However, for the test data we can see the area under the curve is more for Random forest.

However, the recall rate in the CART model is high and that is the desirable character in this case.

All the three models performed almost equally well on the test data but we can notice that the CART model has performed well in train and test data. Hence, we will proceed with CART model for the classification in this case. CART, when compared to the other two models, has performed well in train data and the drop in parameters when it comes to test data is not too much. This makes CART all the more better.

Question 5:

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

Answer:

Since the problem statement states that there is an increasing number of claims. The primary intention of this analysis was done to provide recommendations on flags that the company can look out for and a model to predict is a customer will claim or no. Here, we will need a high accuracy so as to rightly classify a customer who will claim and the one who won't. At the same time, for the company it is a lesser expensive mistake to predict a customer who will not claim as claimed rather than a predicting that a customer will not claim when he will claim. To encounter this we need to maintain the best possible recall rate as well.

Recommendations:

- There is an exceptionally large chunk of customers in the their mid/late 30's. It would be advisable if the company normalized this a little to have a balanced risk portfolio. This can be executed by increasing the commission for agents for the other age groups.
- There is a sudden spike in the duration of travel at 350+ days. This will skew the portfolio a little. This increases the probability of a claim .
- From feature importance, we have analysed that Agent plays a very important role in the prediction as well. Keeping this in mind. We can say that EPX agent has large number of customers that do not claim while so the company can relax around with EPX. On the other hand,C2B agent the highest number of customers who have claimed. So company needs to be watchful of C2B agent.
- The next important feature is Product. The company can encourage more customized plans (provided the customization is done calculatedly as it is done now) as there are a large number of customers who don't claim if they took a Customized plan. Similarly, the company needs to look into Silver Plan as it has the highest number of claims.
- For 100 policies, using our prediction tool the company will be able to predict if a policy will be claimed or no accurately for 80 policies. This way there is only 20% uncertainty/risk (based on accuracy).
- If we run new customers through our model and target on writing 100 customers for whom the model predicts that they will not claim. Then we can be sure that 80 of them will surely not claim. Whereas in earlier case we would have to assume that 70 of them will not claim (based on the precision for not claimed).
- By historical means if we found out that 100 customers will claim then we would set aside the amount to pay these 100 customers as claim reserve. However if the model predicts that 100 customers will claim we can be sure that only 54 of them will claim. The remaining amount (46 customer's claim money) can be invested elsewhere to make profits as claims reserve is considered to be a stagnated fund in the insurance industry.