DATA MINING PROJECT

Dinesh Yadav Mekala

Table of Contents

Problem 1: Clustering

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.

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

DataFrame Info :

RangeIndex: 210 entries, 0 to 209 Data
columns (total 7 columns):

Table 1 Bank Data info

Column	Non-Null Count	Dtype
spending	210 non-null	float64
advance_payments	210 non-null	float64
probability_of_full_payment	210 non-null	float64
current_balance	210 non-null	float64
credit_limit	210 non-null	float64
min_payment_amt	210 non-null	float64
max_spent_in_single_shopping	210 non-null	float64

DataFrame Null values

Table 2 Bank Data Null values

Columns	Null Values
spending	0
advance_payments	0
probability_of_full_payment	0
current_balance	0
credit_limit	0
min_payment_amt	0
max_spent_in_single_shopping	0

Observation

- 7 variables and 210 records.
- No missing record based on intial analysis.
- All the variables numeric type.

Table 3 Bank Data Description

	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

Observation:

- Looking at the summary data, overall the data looks good.
- Mostly for all the variable, mean/medium are almost equal.
- The data is almost evenly distributed.
- Standard Deviation is high for spending variable.

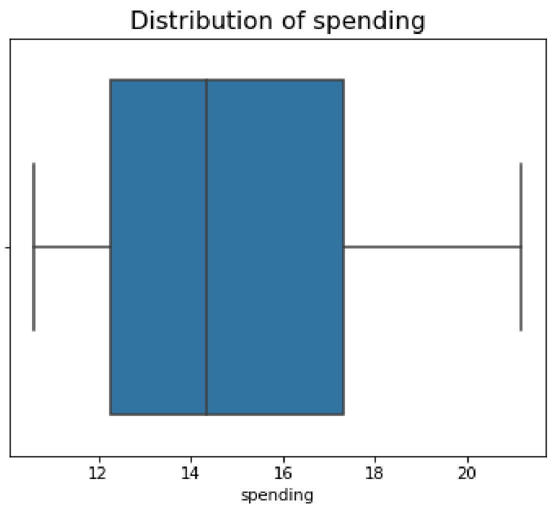
Univariate Analysis

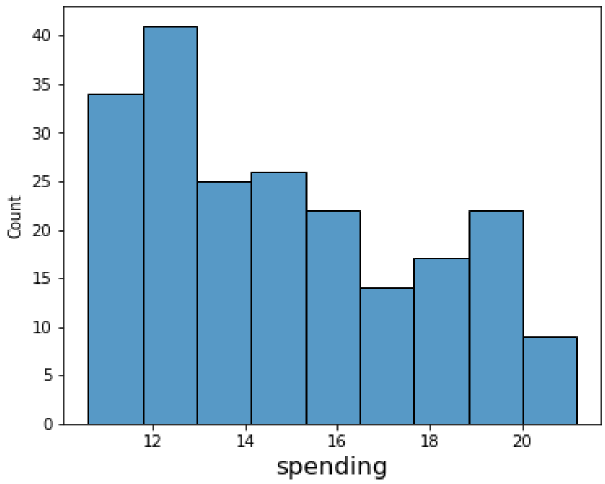
Spending

Table 4 Spending Description

Range of values	10.59
Minimum spending	10.59
Maximum spending	21.18
Mean value	14.847523809523818
Median value.	14.355
Standard deviation.	2.909699430687361
Null values.	False
spending - 1st Quartile (Q1)	12.27
spending - 3st Quartile (Q3)	17.305
Interquartile range (IQR) of spending	5.035
Lower outliers in spending	4.717499999999999
Upper outliers in spending	24.8575
Number of outliers in spending upper	0
Number of outliers in spending lower	0
% of Outlier in spending upper	0 %
% of Outlier in spending lower	0 %

Plots for Spending variable





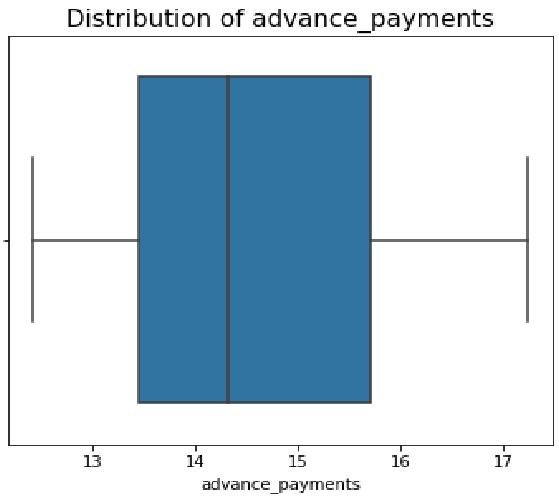
Advance Payments

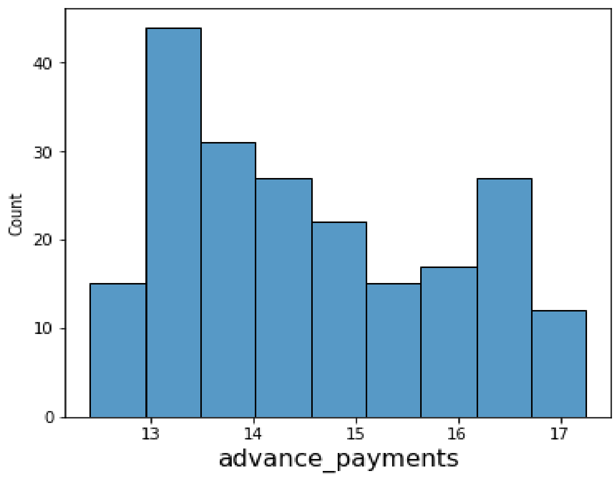
Table 5 ***Advance Payments Description***

Range of values	4.84
Minimum advance_payments	12.41
Maximum advance_payments	17.25
Mean value	14.559285714285727
Median value	14.32
Standard deviation	1.305958726564022

Null values	False
advance_payments - 1st Quartile (Q1)	13.45
advance_payments - 3st Quartile (Q3)	15.715
Interquartile range (IQR) of advance_payments	2.2650000000000006
Lower outliers in advance_payments	10.052499999999998
Upper outliers in advance_payments	19.1125
Number of outliers in advance_payments upper	0
Number of outliers in advance_payments lower	0
% of Outlier in advance_payments upper	0 %
% of Outlier in advance_payments lower	0 %

Plots for Advance Payments



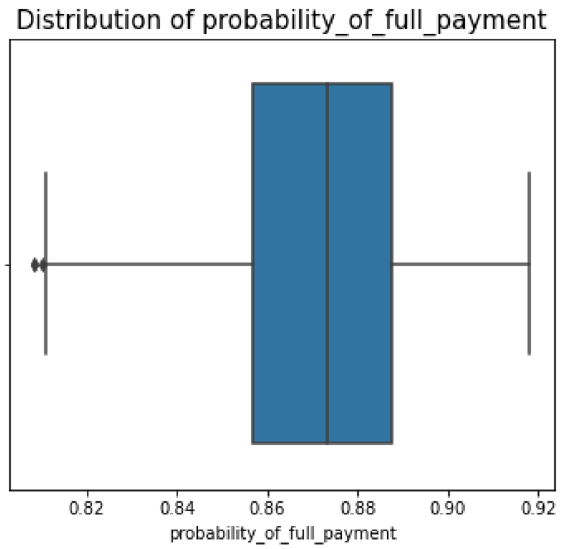


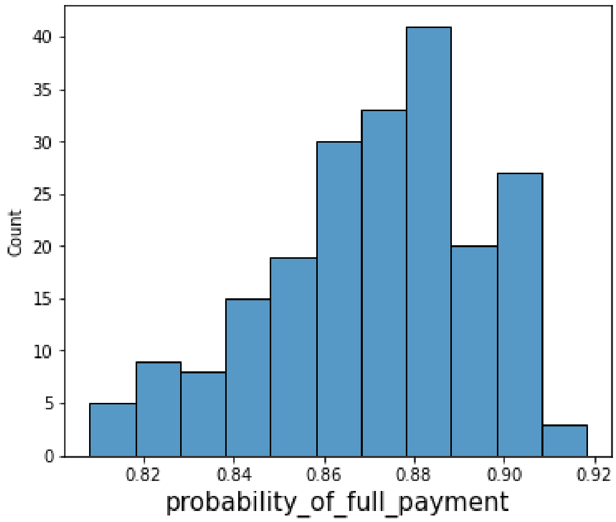
Probability of Full Payment

Table 6 ***Probability of Full Payment Description***

Range of values	0.11019999999999996
Minimum probability_of_full_payment 0.8081	
Maximum probability_of_full_payment	0.9183
Mean value	0.8709985714285714
Median value	0.8734500000000001
Standard deviation	0.0236294165838465
Null values	False
probability_of_full_payment - 1st Quartile (Q1)	0.8569
probability_of_full_payment - 3st Quartile (Q3)	0.887775
Interquartile range (IQR) of probability_of_full_payment	0.030874999999999986
Lower outliers in probability_of_full_payment	0.8105875
Upper outliers in probability_of_full_payment	0.9340875
umber of outliers in probability_of_full_payment upper	0
Number of outliers in probability_of_full_payment lower	3
% of Outlier in probability_of_full_payment upper	0 %
% of Outlier in probability_of_full_payment lower	1 %

Plots for Probability of full payment



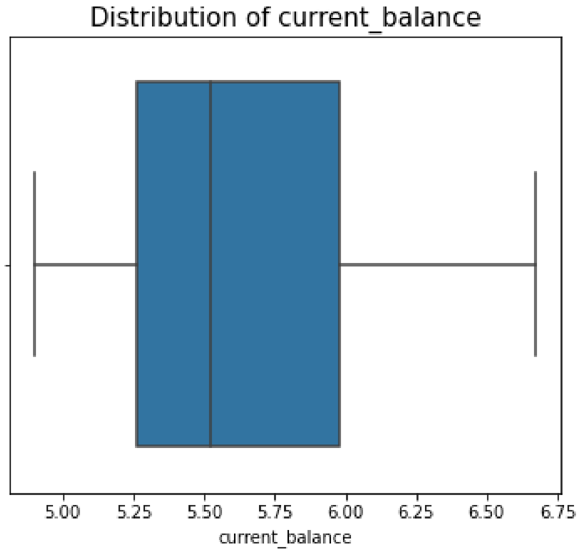


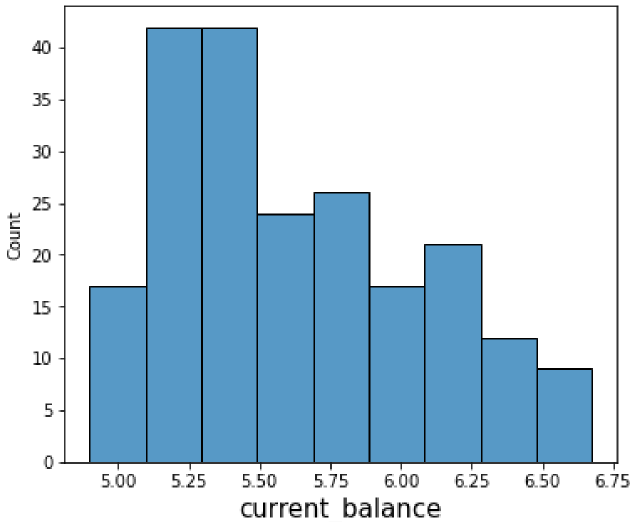
Current Balance

Table 7 Current Balance Description

Range of values	1.7759999999999998
Minimum current_balance	4.899
Maximum current_balance	6.675
Mean value	5.628533333333335
Median value	5.5235
Standard deviation	0.44306347772644944
Null values	False
current_balance - 1st Quartile (Q1)	5.26225
current_balance - 3st Quartile (Q3)	5.97975
Interquartile range (IQR) of current_balance	0.7175000000000002
Lower outliers in current_balance	4.186
Upper outliers in current_balance	7.056000000000001
Number of outliers in current_balance upper	0
Number of outliers in current_balance lower	0
% of Outlier in current_balance upper	0 %
% of Outlier in current_balance lower	0 %

Plots for Current Balance



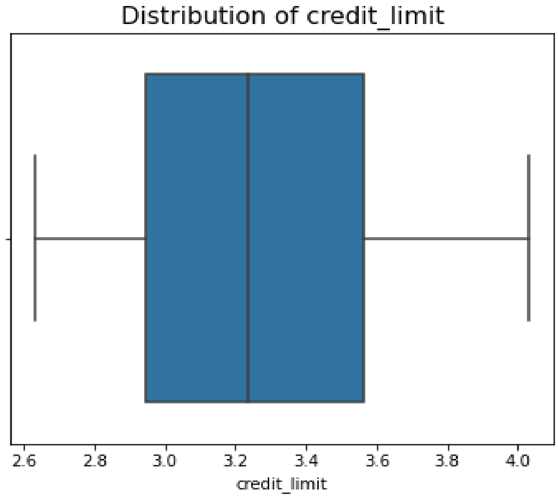


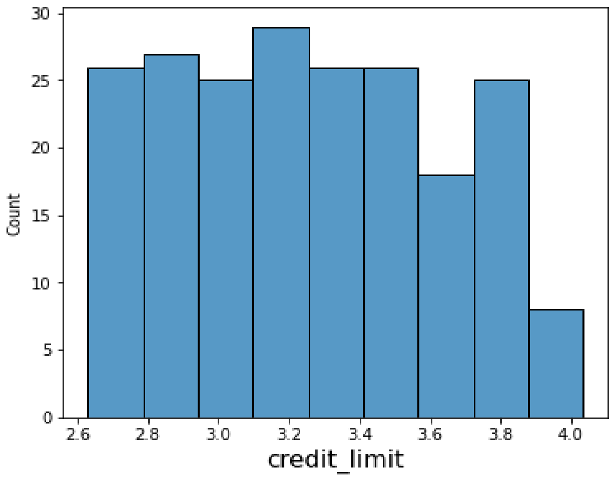
Credit Limit

Table 8 Credit Limit Description

Range of values	1.4030000000000005
Minimum credit_limit	2.63
Maximum credit_limit	4.033
Mean value	3.258604761904763
Median value	3.237
Standard deviation	0.37771444490658734
Null values	False
credit_limit - 1st Quartile (Q1)	2.944
credit_limit - 3st Quartile (Q3)	3.56175
Interquartile range (IQR) of credit_limit	0.61775
Lower outliers in credit_limit	2.017375
Upper outliers in credit_limit	4.488375
Number of outliers in credit_limit upper	0
Number of outliers in credit_limit lower	0
% of Outlier in credit_limit upper	0 %
% of Outlier in credit_limit lower	0 %

Plots for Credit Limit



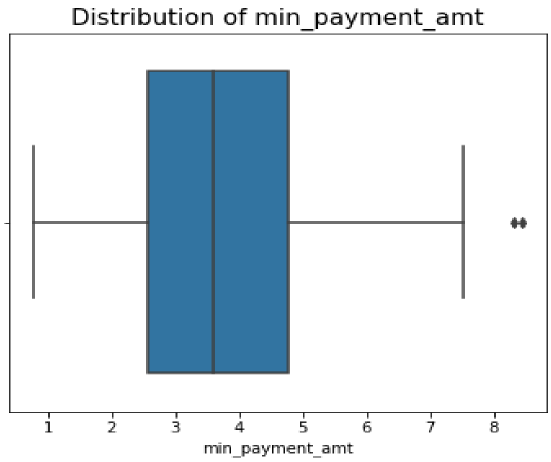


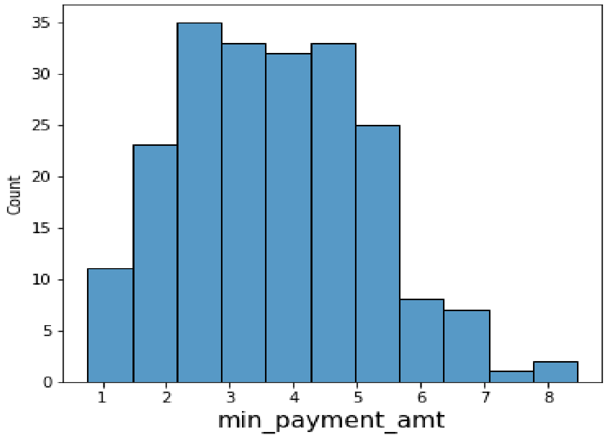
Minimum Payment Amount

Table 9 Minimum Payment Description

Range of values	7.690899999999999
Minimum min_payment_amt	0.7651
Maximum min_payment_amt	8.456
Mean value	3.7002009523809503
Median value	3.599
Standard deviation	1.5035571308217792
Null values	False
min_payment_amt - 1st Quartile (Q1)	2.5615
min_payment_amt - 3st Quartile (Q3)	4.76875
Interquartile range (IQR) of min_payment_amt	2.2072499999999997
Lower outliers in min_payment_amt	-0.7493749999999992
Upper outliers in min_payment_amt	8.079625
Number of outliers in min_payment_amt upper	2
Number of outliers in min_payment_amt lower	0
% of Outlier in min_payment_amt upper	1 %
% of Outlier in min_payment_amt lower	0 %

Plots for Minimum Payment Amount



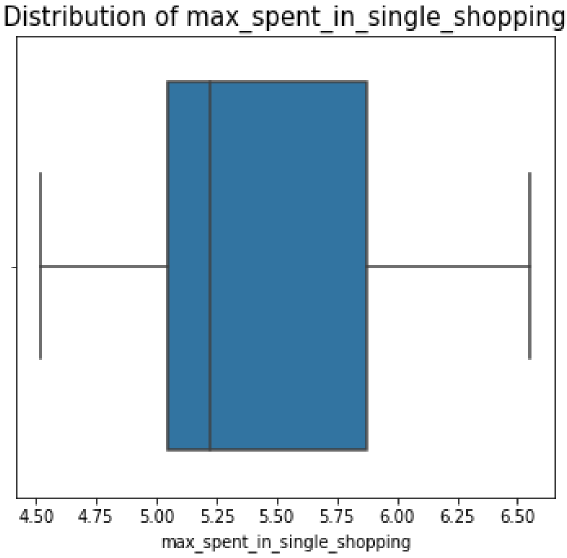


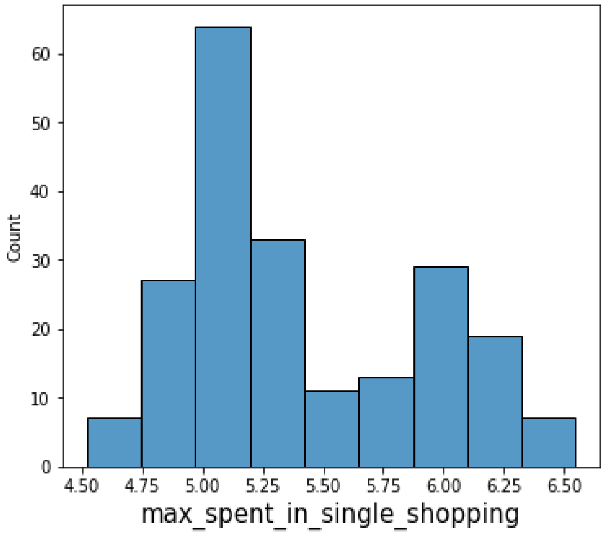
Max Spent in Single Shopping

Table 10 Spent in single shopping description

Range of values	2.0309999999999997
Minimum max_spent_in_single_shopping	4.519
Maximum max_spent_in_single_shoppings	6.55
Mean value	5.408071428571429
Median value	5.223000000000001
Standard deviation	0.49148049910240543
Null values	False
max_spent_in_single_shopping - 1st Quartile (Q1)	5.045
max_spent_in_single_shopping - 3st Quartile (Q3)	5.877
Interquartile range (IQR) of max_spent_in_single_shopping	0.8319999999999999
Lower outliers in max_spent_in_single_shopping	3.797
Upper outliers in max_spent_in_single_shopping	7.125
Number of outliers in max_spent_in_single_shopping upper	0
Number of outliers in max_spent_in_single_shopping lower	0
% of Outlier in max_spent_in_single_shopping upper	0 %
% of Outlier in max_spent_in_single_shopping lower	0 %

Plot for Max Spent in Single Shopping

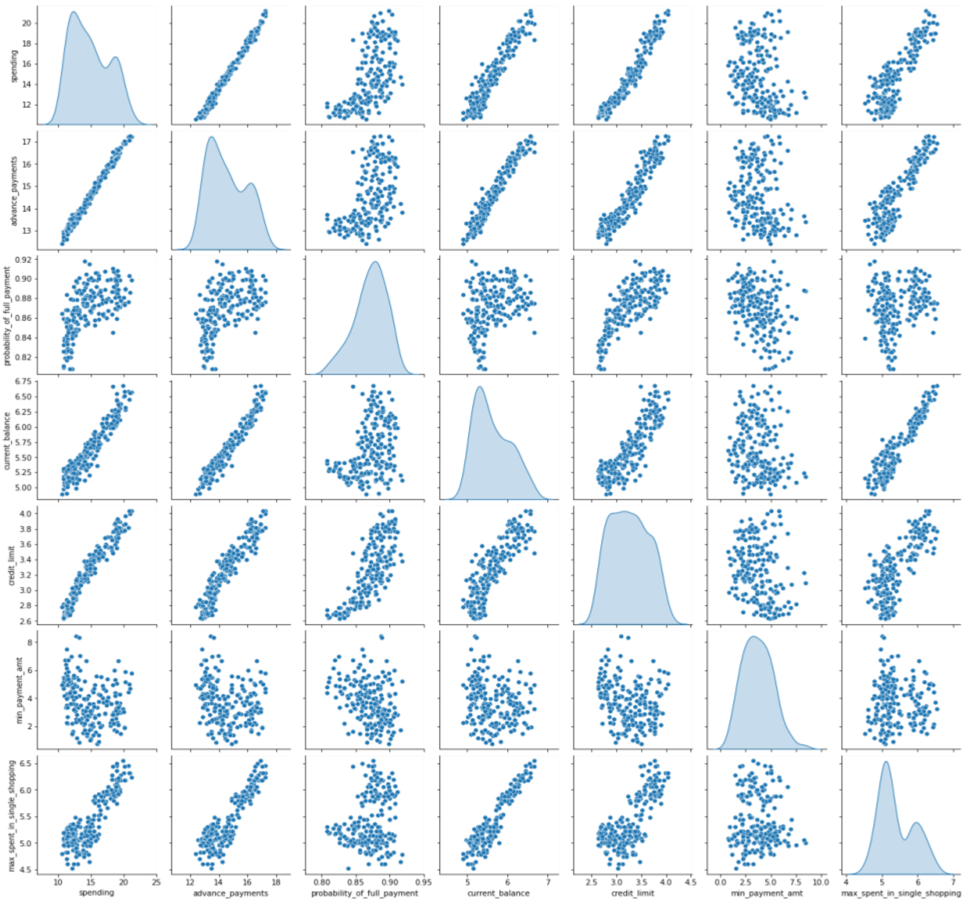




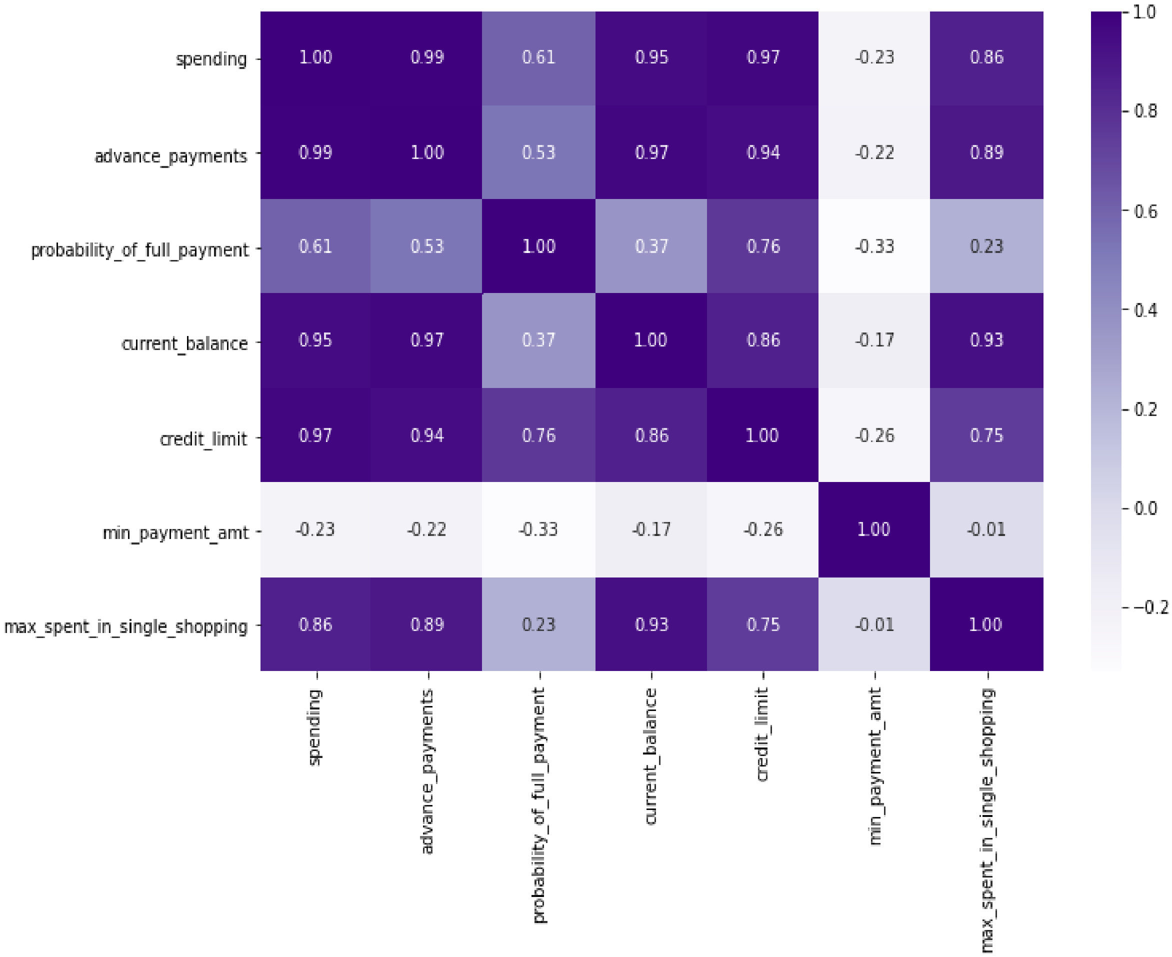
Observation:

- **Credit limit average is around $3.258(10000s)**
- **Distribution is skewed to right tail for all the variable except probability of full payment variable, which has left tail skew**

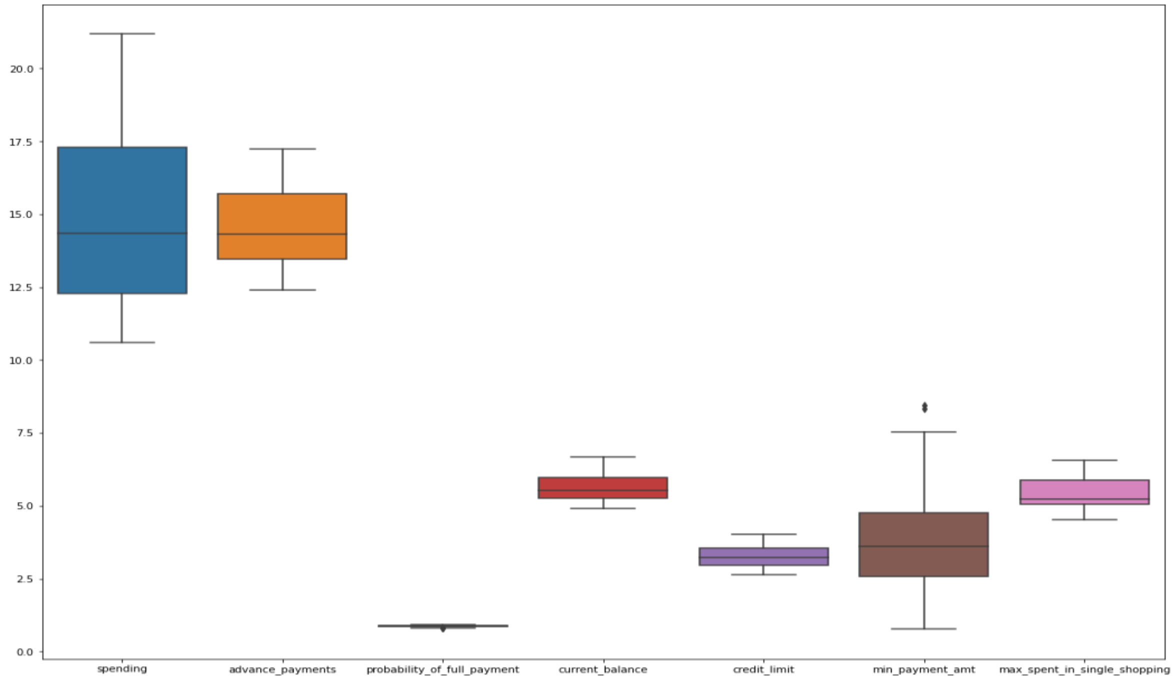
Multivariate Analysis



From the pair plot we can observe strong positive correlation between • Spending and Advance payments, • Advance payments and current_balance • Credit limit and spending • Spending and current_balance • Credit limit and advance payments • Max spent in single shopping current balance

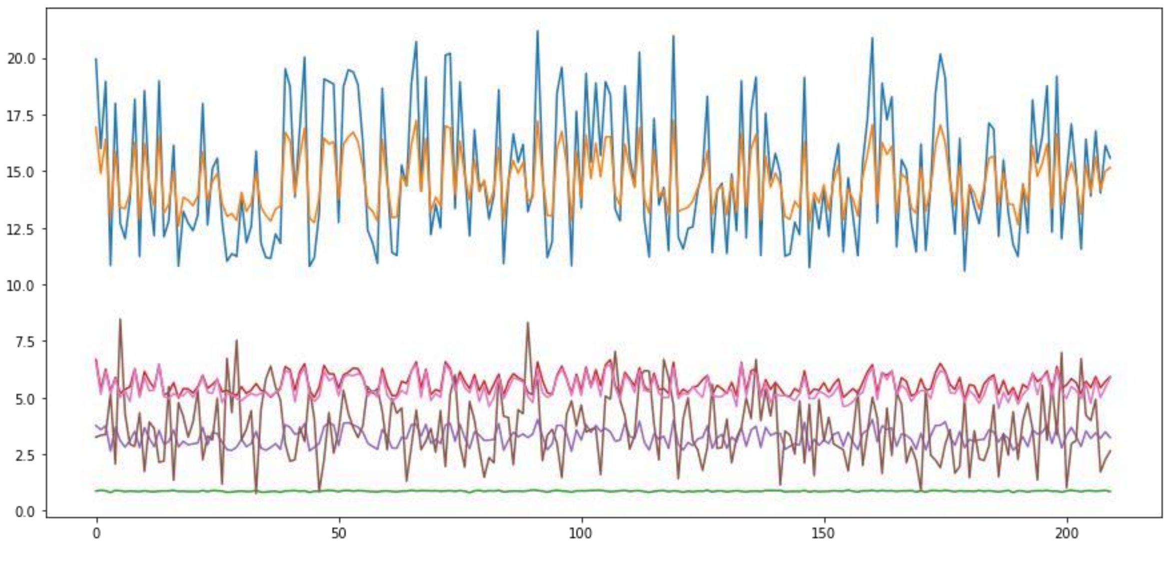


Outliers

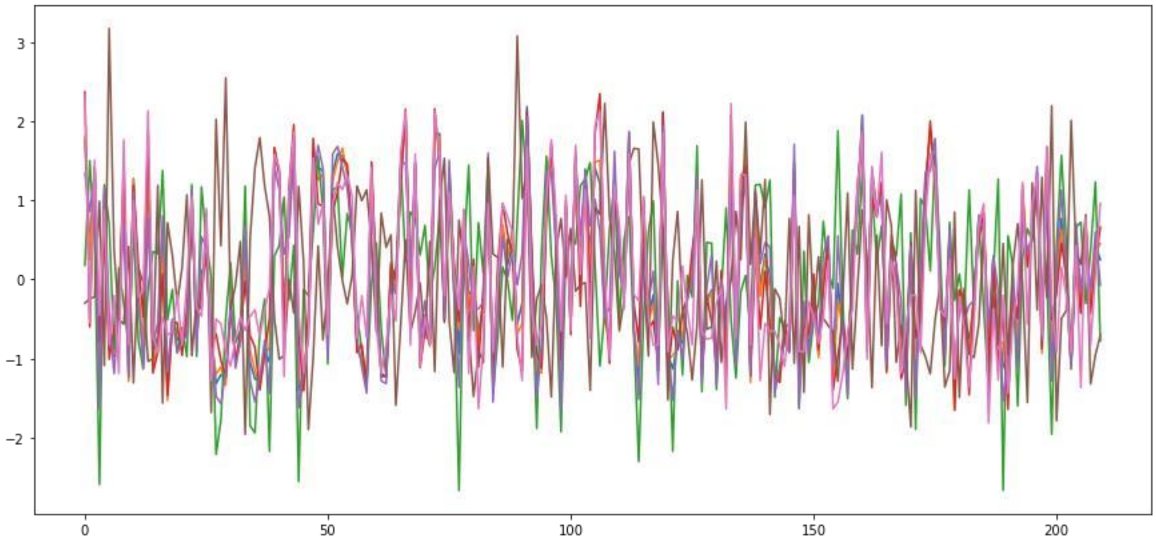


1.2 Do you think scaling is necessary for clustering in this case? Justify

Plot before scaling



Plot after scaling

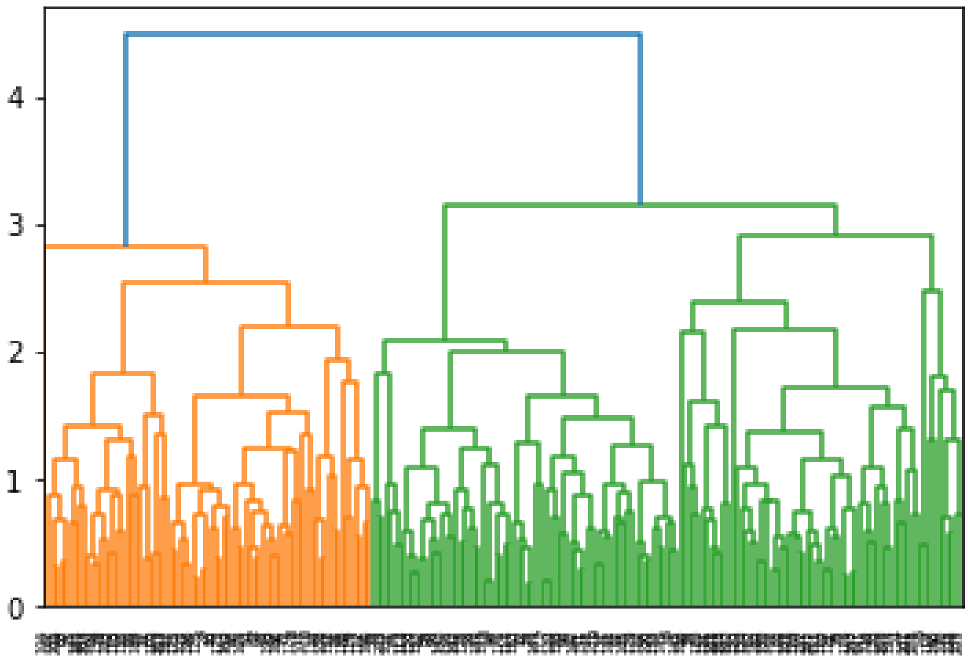


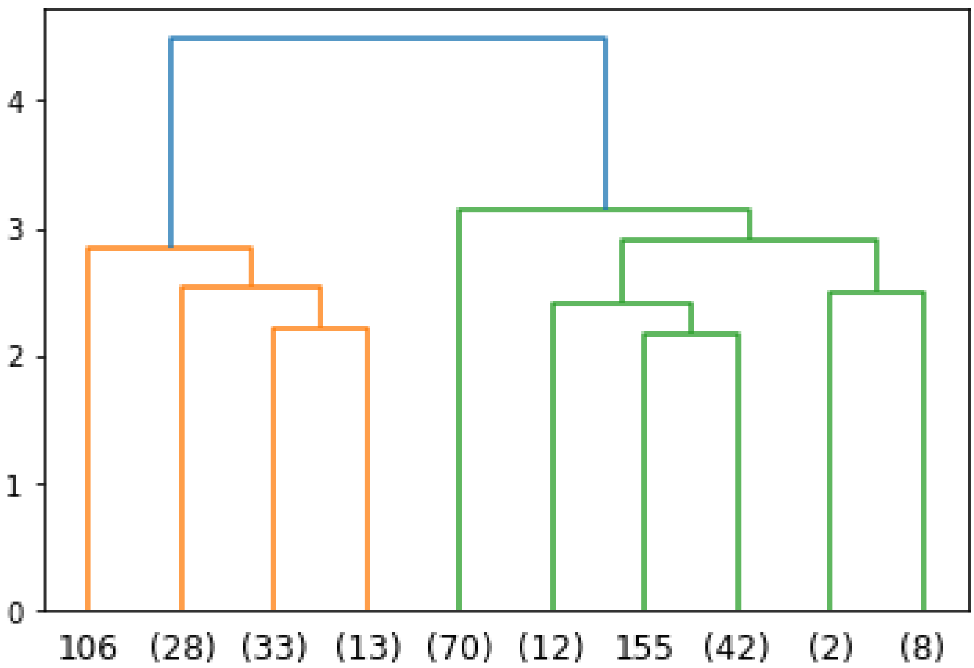
Scaling needs to be done as the values of the variables are different. spending, advance payments are in different values and this may get more weightage. Also have shown below the plot of the data prior and after scaling.

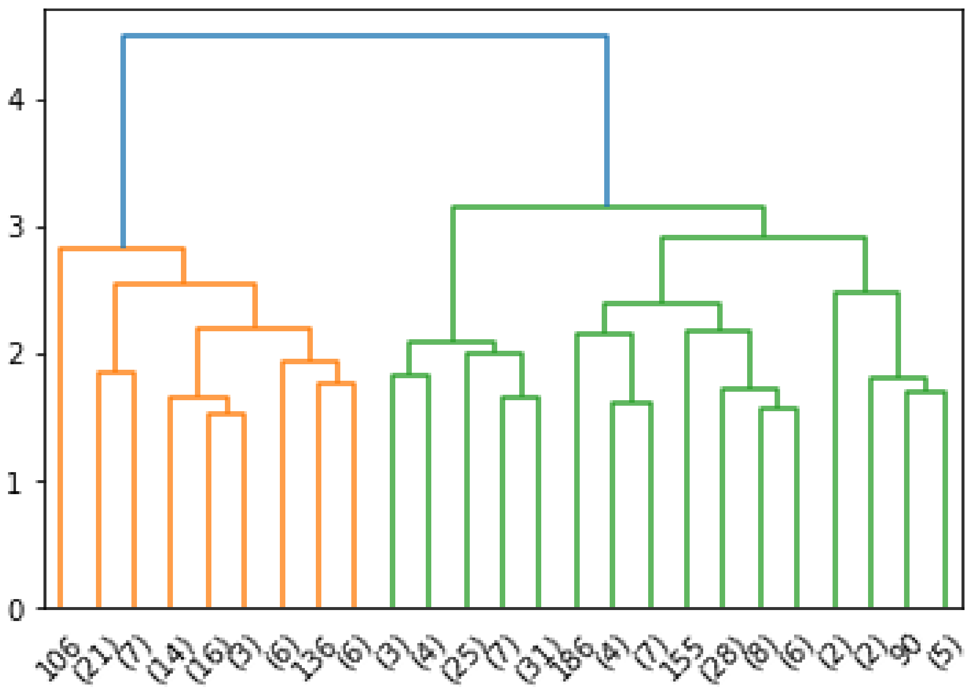
Scaling will have all the values in the relative same range. We use zscore to standardise the data to relative same scale -3 to +3.

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

Dendogram







Cluster

```
array([1, 3, 1, 2, 1, 3, 2, 2, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 2, 2, 2
,
       1, 2, 3, 1, 3, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 1, 1, 3, 1, 1
```

```
,
        2, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 3, 2, 2, 1, 3, 1
,
        1, 3, 1, 2, 3, 2, 1, 1, 2, 1, 3, 2, 1, 3, 3, 3, 3, 1, 2, 1, 1, 1
,
        1, 3, 3, 1, 3, 2, 2, 1, 1, 1, 2, 1, 3, 1, 3, 1, 3, 1, 1, 2, 3, 1
,
        1, 3, 1, 2, 2, 1, 3, 3, 2, 1, 3, 2, 2, 2, 3, 3, 1, 2, 3, 3, 2, 3
,
        3, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 2, 2, 1, 2, 3, 2, 3, 2, 3, 1
,
        3, 3, 2, 2, 3, 1, 1, 2, 1, 1, 1, 2, 1, 3, 3, 2, 3, 2, 3, 1, 1, 1
,
        3, 2, 3, 2, 3, 2, 3, 3, 1, 1, 3, 1, 3, 2, 3, 3, 2, 1, 3, 1, 1, 2
,
        1, 2, 3, 3, 3, 2, 1, 3, 1, 3, 3, 1], dtype=int32)
```

Table 11

Cluster	Frequency
1	75
2	70
3	65

Table 12

	0	**1**	**2**	**3**	**4**
spending	19.94	15.99	18.95	10.83	17.99
advance payments	16.92	14.89	16.42	12.96	15.86
Probability of full payment	0.8752	0.9064	0.8829	0.8099	0.8992
current balance	6.675	5.363	6.248	5.278	5.89
credit limit	3.763	3.582	3.755	2.641	3.694
min_payment_amt	3.252	3.336	3.368	5.182	2.068

Max spent in single shopping	6.55	5.144	6.-	5.185	5.837
cluster	1	3	1	2	1

For cluster grouping based on the dendrogram, 3 clusters or 4 clusters looks good. By doing the further analysis, it's clear that 3 group cluster solution is the ideal cluster based on the hierarchical clustering.

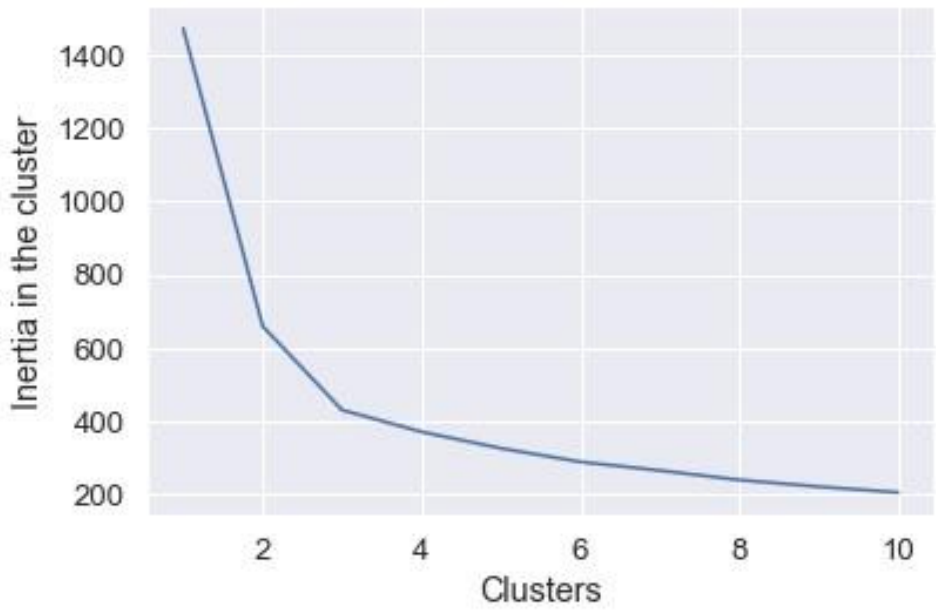
And 3 group cluster solution gives a pattern based on high/medium/low spending with max spent in single shopping and probability of full payment.

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

The optimum clusters is 3 clusters.

The K-mean inertia for 3 clusters is 430.65

Elbow Curve



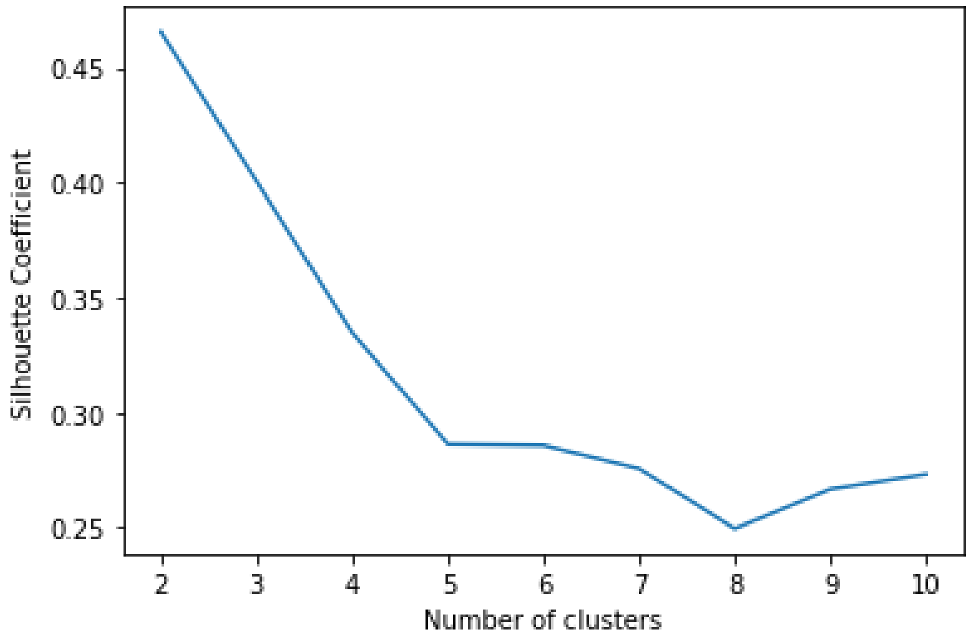
3 clusters in kmeans is better we ses that based on current dataset given, 3 cluster solution makes sense based on the high spending pattern, medium spending pattern and low spending pattern

Table 13

	0	**1**	**2**	**3**	**4**
spending	19.94	15.99	18.95	10.83	17.99
advance_payments	16.92	14.89	16.42	12.96	15.86
probability_of_full_payment	0.8752	0.9064	0.8829	0.8099	0.8992
current_balanc	6.675	5.363	6.248	5.278	5.89
credit_limit	3.763	3.582	3.755	2.641	3.694
min_payment_amt	3.252	3.336	3.368	5.182	2.068
max_spent_in_single_shopping	6.55	5.144	6.148	5.185	5.837
cluster	1	3	1	2	1
Clus_kmeans	0	2	0	1	0

The silhouette score for scaled data is 0.40072705527512986

Silhouette Coefficient Graph



Silhouette width

Table 14

	0	1	2	3	4
spending	19.94	15.99	18.95	10.83	17.99
advance_payments	16.92	14.89	16.42	12.96	15.86
probability_of_full_payment	0.8752	0.9064	0.8829	0.8099	0.8992
current_balance	6.675	5.363	6.248	5.278	5.89
credit_limit	3.763	3.582	3.755	2.641	3.694
min_payment_amt	3.252	3.336	3.368	5.182	2.068
max_spent_in_single_shopping	6.55	5.144	6.148	5.185	5.837
cluster	1	3	1	2	1
Clus_kmeans	0	2	0	1	0
sil_width	0.573699	0.366386	0.637784	0.512458	0.362276

Table 15

Cluster_Size	Cluster_Percentage
71	33.81
72	34.29
67	31.9

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

Table 16

cluster	1	2	3
spending	14.4	11.9	18.5
Advance payments	14.3	13.2	16.2
Probability of full payment	0.9	0.8	0.9
Current balance	5.5	5.2	6.2
Credit limit	3.3	2.8	3.7
Min payment amt	2.7	4.7	3.6
Max spent in single shopping	5.1	5.1	6.0

Cluster 1 : This cluster have customers who spend on purchases on a regular basis and pay their bills on a regular basis, but have a credit limit which is not high.

Cluster 2 : This cluster have customers who are low on purchases and doesn't have a good record in full payments. They have very less credit limit. They have a good percentage in Minimum Payment amount.

Cluster 3 : This cluster have customers who spend the most, they have good rate in advance payments. They are given very good credit limit as they have good probability in full payments.

Promotional Strategies for each cluster

Cluster 1 : Medium Spending Group

- They are potential target customers who are paying bills and doing purchases and maintaining comparatively good credit score. So we can increase credit limit or can lower down interest rate.
- Promote premium cards/loyalty cards to increase transactions.
- Increase spending habits by trying with premium ecommerce sites, travel portal, travel airline s/hotel, as this will encourage them to spend more

Cluster 2 : Low Spending Group

- customers should be given remainders for payments. Offers can be provided on early payment to improve their payment rate.
- Increase their spending habits by tying up with grocery stores, utilities

Cluster 3 : High Spending Group

- More reward points might increase their purchases.
- Maximum max spent in single shopping is high for this group, so can be offered discount/ offer on next transactions upon full payment.
- Increase their credit limit
- Give loan against the credit card, as they are customers with good repayment record.
- Tie up with luxury brands, which will drive more one time maximum spending link code.

Problem 2: CART-RF-ANN

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.

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

DataFrame info:

```
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
```

Table 17 Insurance Info

Column	Non-Null Count	Dtype
Age	3000 non-null	int64
Agency_Code	3000 non-null	object
Type	3000 non-null	object
Claimed	3000 non-null	object
Commision	3000 non-null	float64
Channel	3000 non-null	object
Duration	3000 non-null	int64
Sales	3000 non-null	float64
Product Name	3000 non-null	object
Destination	3000 non-null	object

Table 18 Insurance null values

Age	0
Agency_Code	0
Type	0
Claimed	0
Commision	0
Channel	0
Duration	0
Sales	0
Product Name	0
Destination	0

Observation

- 10 variables and 3000 records.
- No missing record based on intial analysis.
- All the variables are not numeric type.

DataFrame Description

Table 19 Insurance Description

	count	mean	std	min	25%	50%	75%	max
Age	3000	38.091	10.463518	8	32	36	42	84
Commision	3000	14.5292	25.481455	0	0	4.63	17.235	210.21
Duration	3000	70.00133	134.053313	-1	11	26.5	63	4580
Sales	3000	60.24991	70.733954	0	20	33	69	539

Observation:

- Looking at the summary data, overall the data looks good.
- Standard Deviation is high for Duration.

DataFrame Head

Table 20 Insurance Head

	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0.7	1	7	2.51	2	0
1	36	2	1	0	1	34	20	2	0
2	39	1	1	5.94	1	3	9.9	2	1
3	36	2	1	0	1	4	26	1	0
4	33	3	0	6.3	1	53	18	0	0

Duplicates

Number of duplicate rows = 139

The data shows there are 139 records, but it can be of different customers, there is no customer ID or any unique identifier, so we can’t drop duplicates.

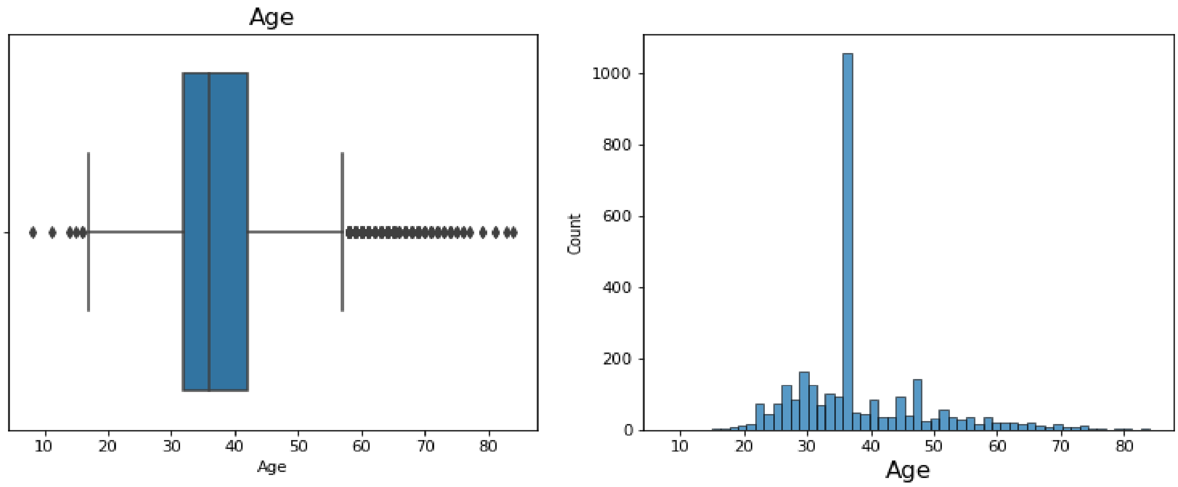
Univariate Analysis:

Age

Table 21 Age Description

Range of values	76
Minimum Age	8
Maximum Age	84
Mean value	38.091
Median value	36.0
Standard deviation	10.463518245377944
Null values	False
spending - 1st Quartile (Q1)	32.0
spending - 3st Quartile (Q3)	42.0
Interquartile range (IQR) of Age	10.0
Lower outliers in Age	17.0
Upper outliers in Age	57.0
Number of outliers in Age upper	198
Number of outliers in Age lower	6
% of Outlier in Age upper	7 %
% of Outlier in Age lower	0 %

Plot for Age



The Age variable is normally distributed and it has many outliers present on both the sides

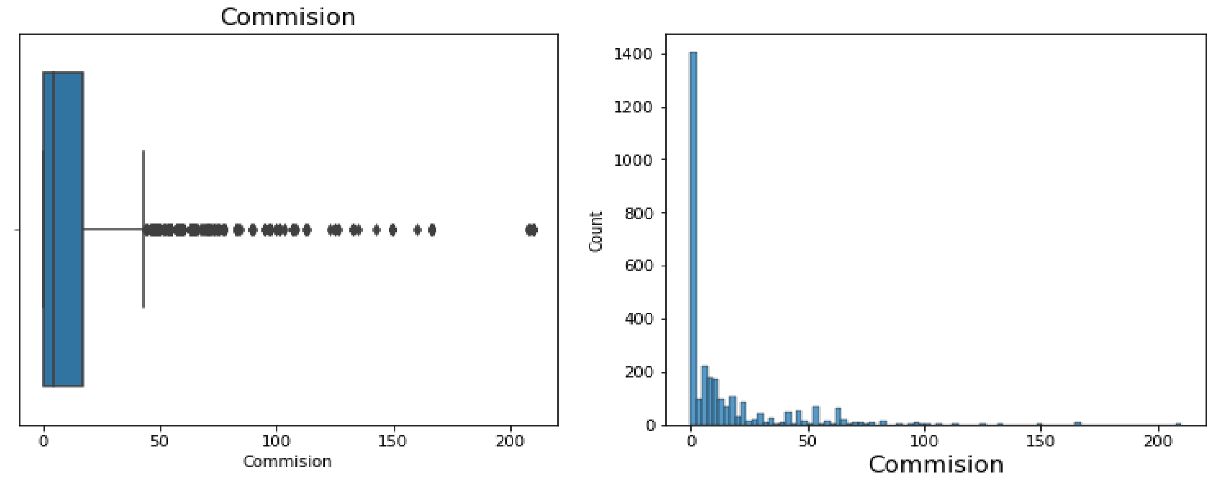
Commission

Table 22 Commission Description

Range of values	210.21
Minimum Commision	0.0
Maximum Commision	210.21
Mean value	14.529203333333266
Median value	4.63

Standard deviation	25.48145450662553
Null values	False
Commision - 1st Quartile (Q1)	0.0
Commision - 3st Quartile (Q3)	17.235
Interquartile range (IQR) of Commision	17.235
Lower outliers in Commision	-25.8525
Upper outliers in Commision	43.0875
Number of outliers in Commision upper	362
Number of outliers in Commision lower	0
% of Outlier in Commision upper	12 %
% of Outlier in Commision lower	0 %

Plot for Commission



The Commission variable is highly right skewed and has outliers on the upper side

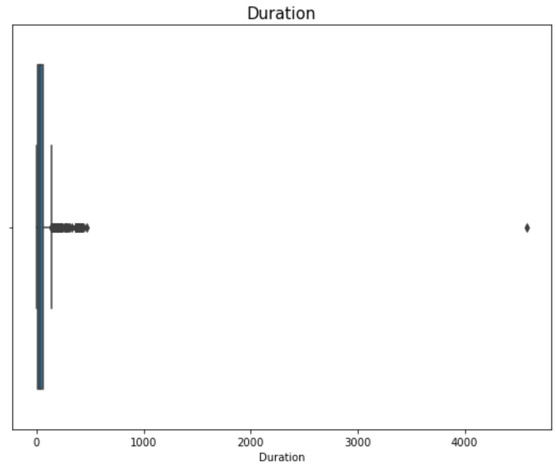
Duration

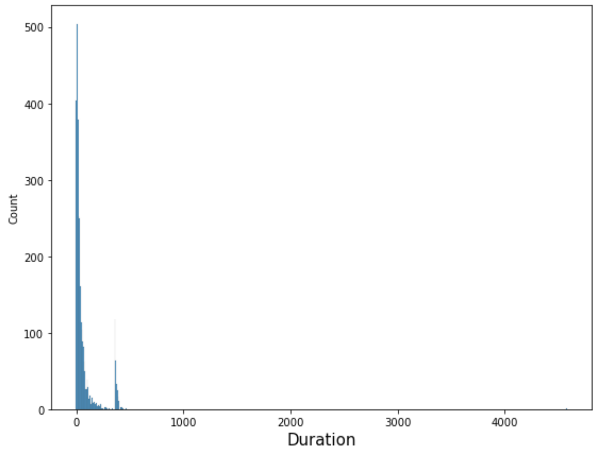
Table 23 Duration Description

Range of values	4581
Minimum Duration	-1
Maximum Duration	4580
Mean value	70.00133333333333
Median value	26.5
Standard deviation	134.05331313253495
Null values	False
Duration - 1st Quartile (Q1)	11.0
Duration - 3st Quartile (Q3)	63.0
Interquartile range (IQR) of Duration	52.0
Lower outliers in Duration	-67.0
Upper outliers in Duration	141.0
Number of outliers in Duration upper	382
Number of outliers in Duration lower	0
% of Outlier in Duration upper	13 %

% of Outlier in Duration lower	0 %

Plots for Duration





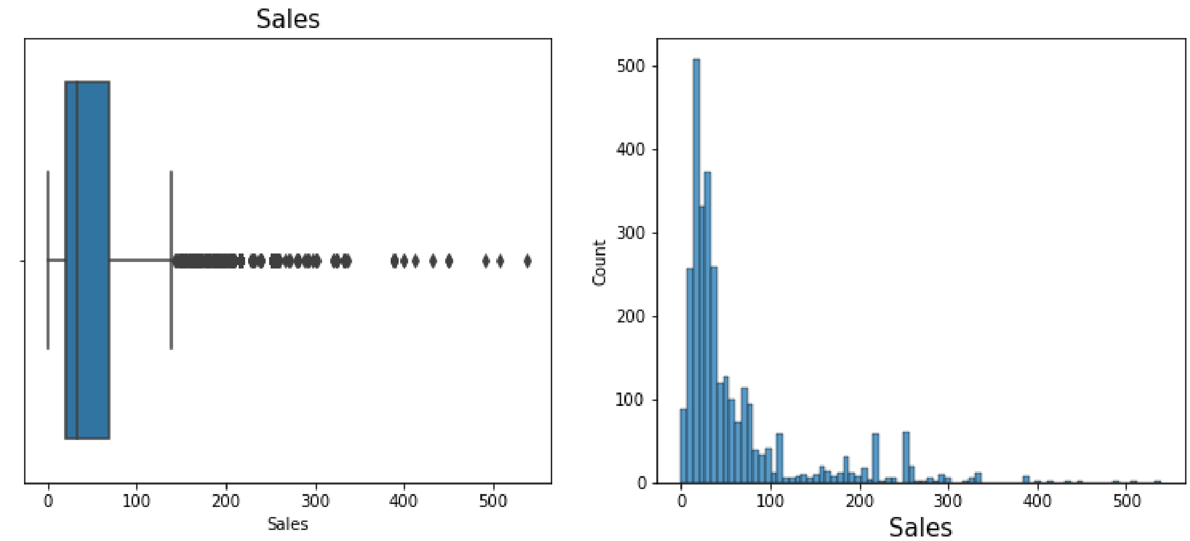
The Duration variable is highly right skewed and has outliers on the upper side.

Sales

Table 24 Sales Description

Range of values	539.0
Minimum Sales	0.0
Maximum Sales	539.0
Mean value	60.24991333333344
Median value	33.0
Standard deviation	70.73395353143047
Null values	False
Sales - 1st Quartile (Q1)	20.0
Sales - 3st Quartile (Q3)	69.0
Interquartile range (IQR) of Sales	49.0
Lower outliers in Sales	-53.5
Upper outliers in Sales	142.5
Number of outliers in Sales upper	353
Number of outliers in Sales lower	0
% of Outlier in Sales upper	12 %
% of Outlier in Sales lower	0 %

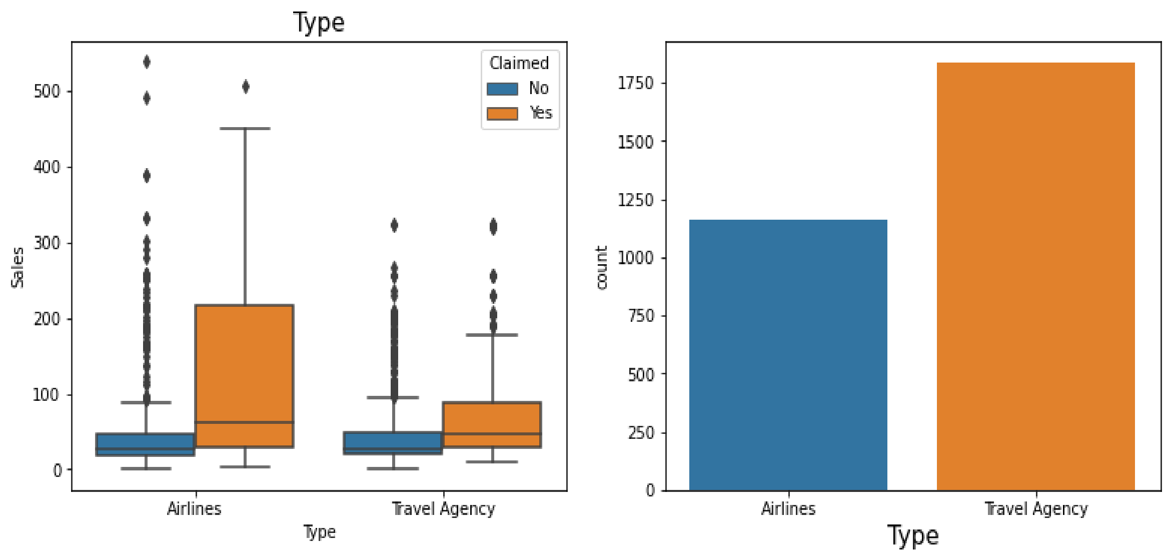
Plots for Sales



The Sales variable is right skewed and has outliers on the upper side.

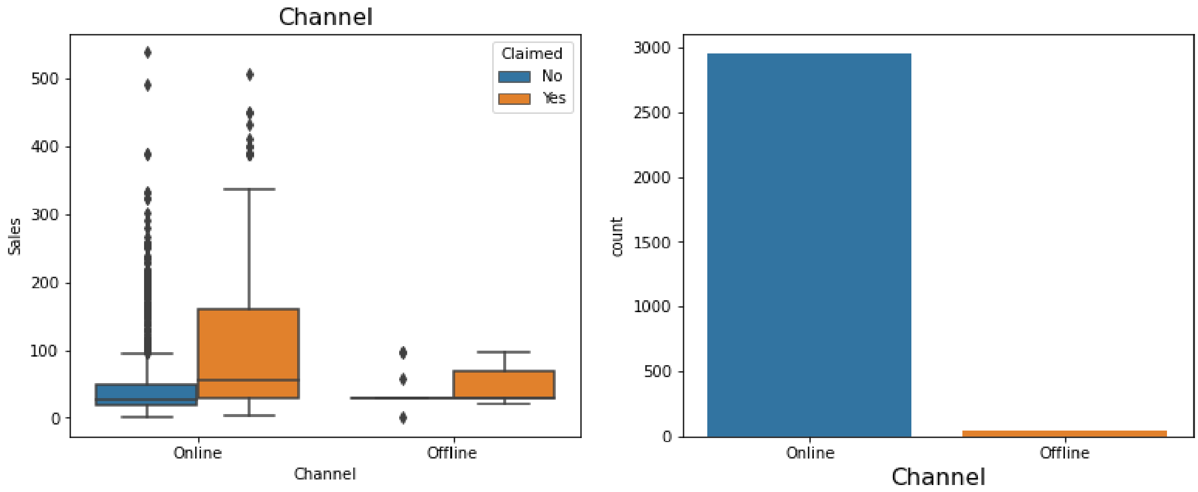
Categorical variables

Type



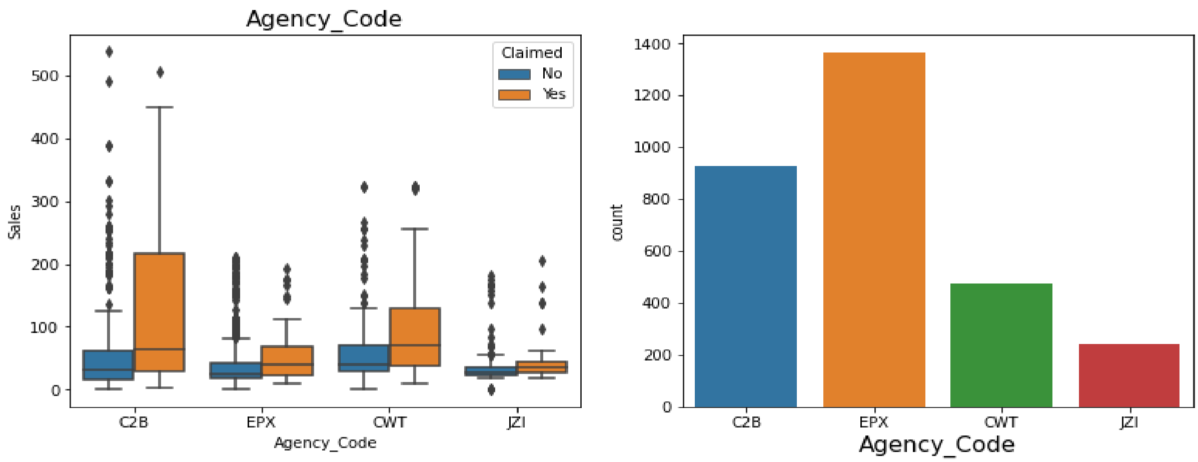
Insurance sales is high in Travel agency when compared to the Airlines. But if seen the claim rate it’s higher in Airlines.

Channel



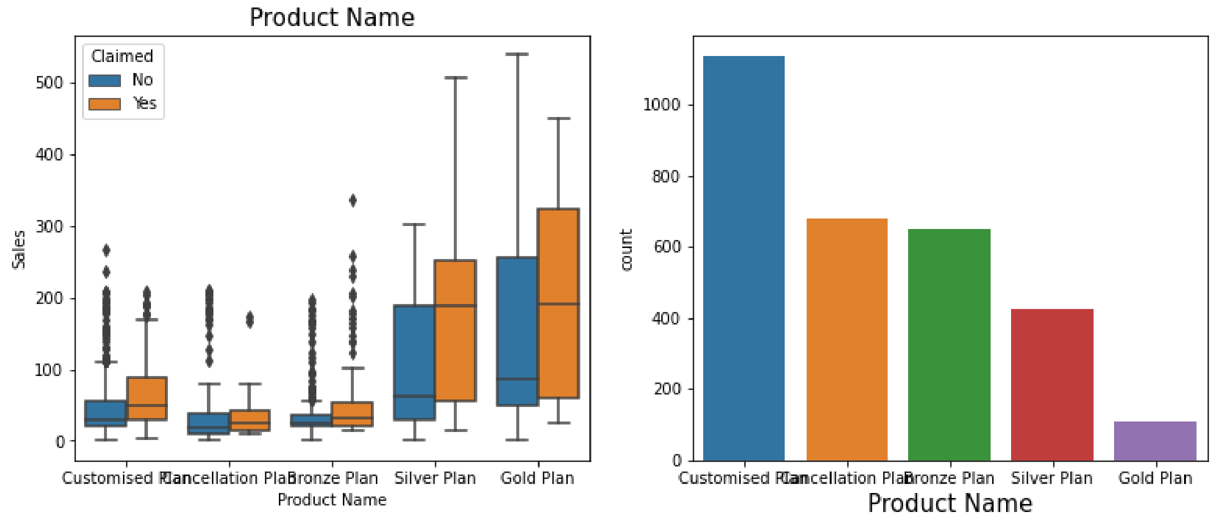
If seen the claim rate is higher in online bookings when compared to offline.

Agency Code



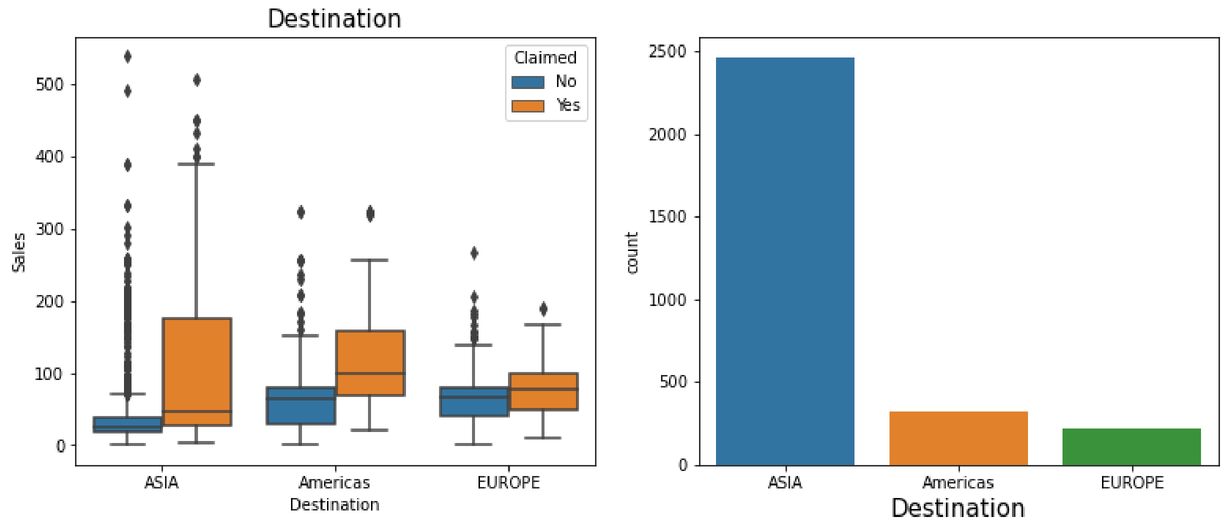
In count EPX Agency code has the highest number of sales. But if seen the claim rate is higher in C2B.

Product Name



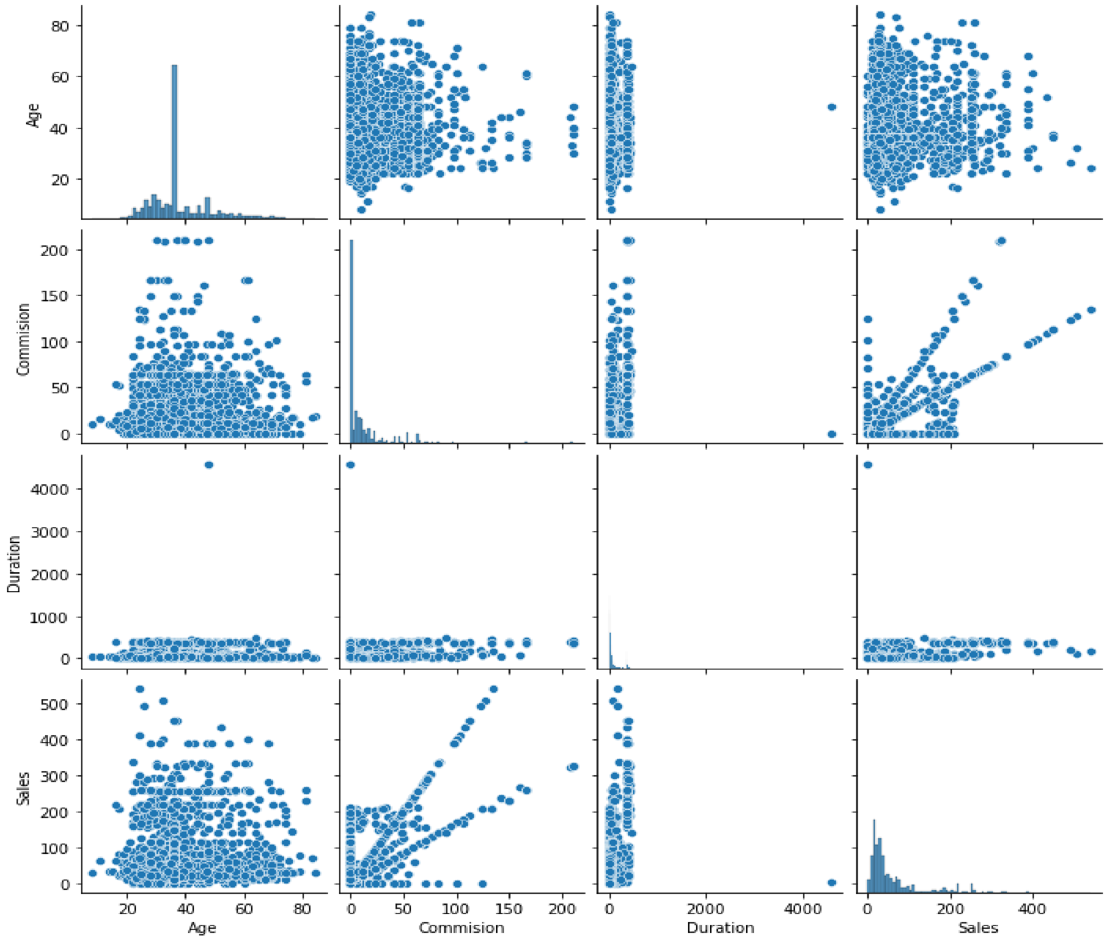
Customised Plan has the highest sales count when compared to all. But Gold plan has the highest claim rate.

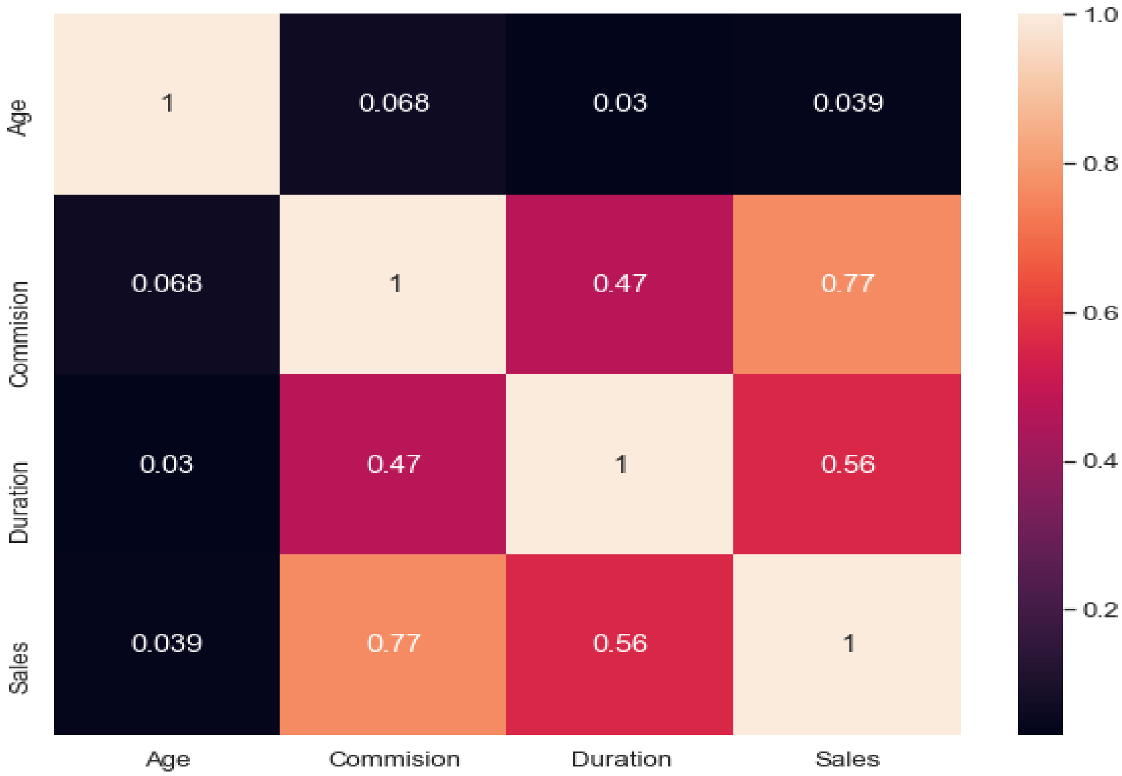
Destination



Customers travelling to Asia have the highest claim rate and insurance sales as well.

Multivariate Analysis





From the pair plot we can observe strong positive correlation between • Commission and Age • Commission and Sales

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

Decision Tree Classifier

Table 25 Feature Importance DT

	Imp
Agency_Code	0.674494
Sales	0.222345
Product Name	0.092149
Commision	0.008008
Duration	0.003005
Age	0.000000
Type	0.000000
Channel	0.000000
Destination	0.000000

Table 26 Predication Probability DT

	0	**1**
0	0.656751	0.343249
1	0.979452	0.020548
2	0.921171	0.078829
3	0.656751	0.343249
4	0.921171	0.078829

Random Forest

Table 27 Feature Importance RF

	Imp
Agency_Code	0.364408
Product Name	0.206559
Sales	0.159045
Commision	0.110955
Type	0.075465
Duration	0.050107
Age	0.023102
Destination	0.005485
Channel	0.004873

Table 28 Prediction Probability RF

	0	**1**
0	0.776949	0.223051
1	0.965672	0.034328
2	0.916237	0.083763
3	0.690907	0.309093
4	0.895991	0.104

Artificial Neural Network

Table 29 Prediction Probability NN

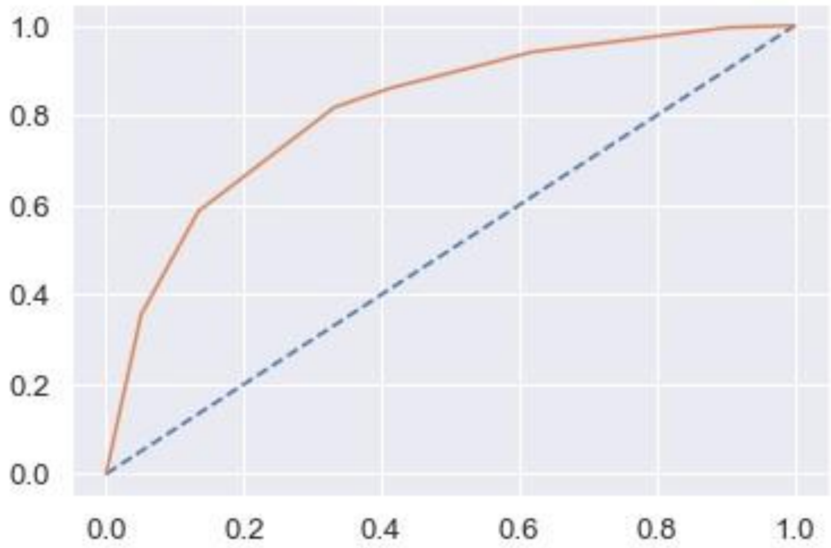
	0	**1**
0	0.822676	0.177324
1	0.933407	0.066593
2	0.918772	0.081228
3	0.688933	0.311067
4	0.913425	0.086575

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

Decision Tree Classifier

Train Data

AUC : 0.812



Confusion Matrix

array([[1258, 195],
       [268, 379]])

Table 30 Classification Report Train DT

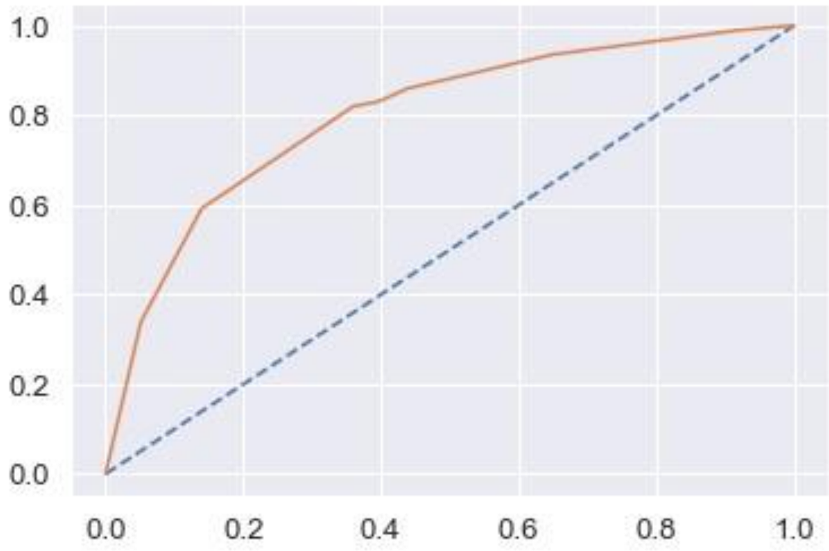
	precision	recall	f1-score	support
0	0.82	0.87	0.84	1453
1	0.66	0.59	0.62	647
accuracy			0.78	2100
macro avg	0.74	0.73	0.73	2100
weighted avg	0.77	0.78	0.78	2100

Table 31

Cart train precision	0.66
Cart train recall	0.59
Cart train f1	0.62
Cart train accuracy	0.779

Test Data

AUC: 0.800



Confusion Matrix

array([[536, 87],
       [113, 164]])

Table 32 Classification Report Test DT

	precision	recall	f1-score	support
0	0.83	0.86	0.84	623
1	0.65	0.59	0.62	277
accuracy			0.78	900
macro avg	0.74	0.73	0.73	900
weighted avg	0.77	0.78	0.77	900

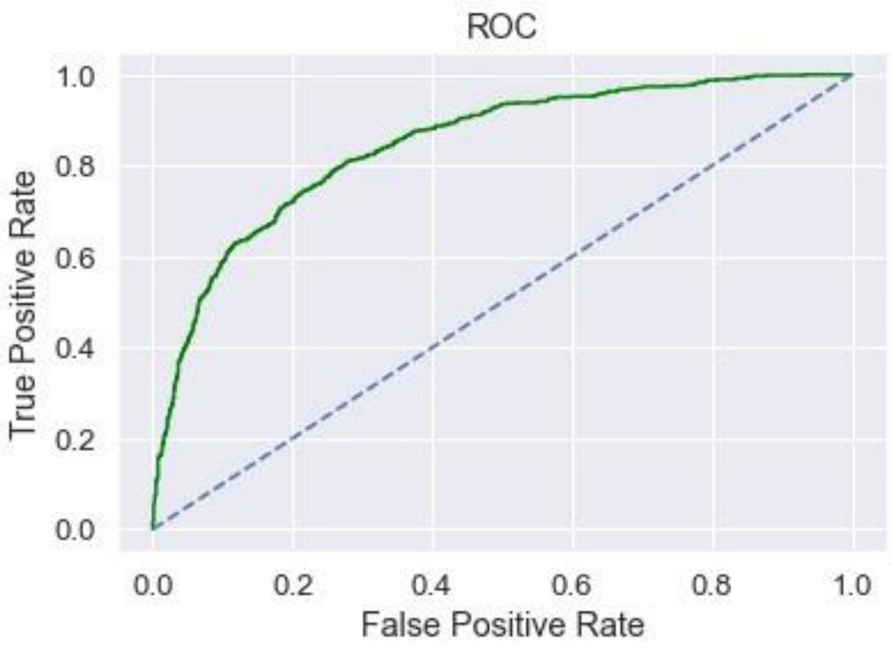
Table 33

Cart test precision	0.65
Cart test recall	0.59
Cart test f1	0.62
Cart test accuracy	0.77

Random Forest

Train Data

AUC : 0.84



Confusion Matrix

array([[1289, 164],
 [246, 401]])

Table 34 Classification Report Train RF

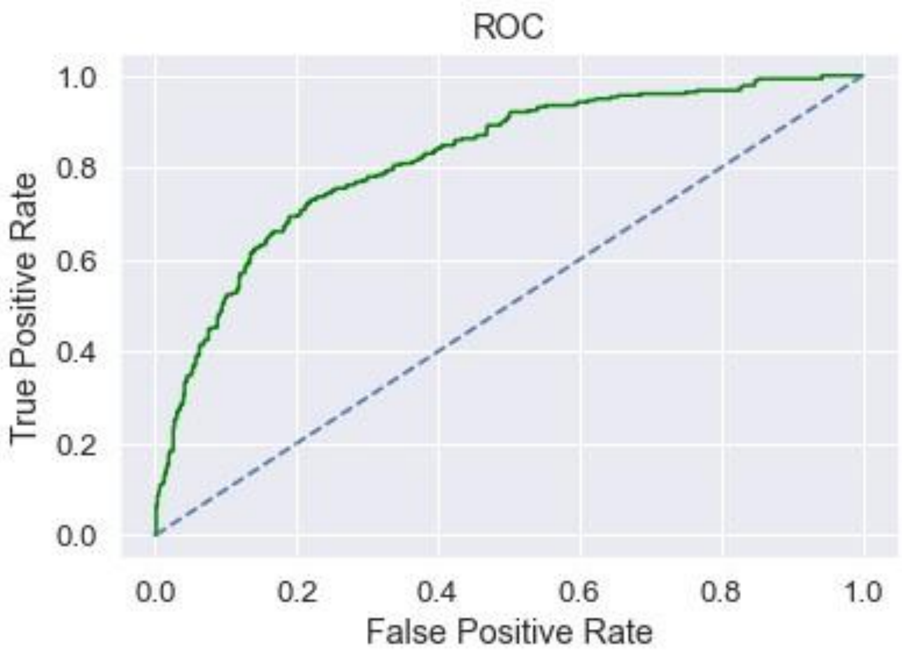
	precision	recall	f1-score	support
0	0.84	0.89	0.86	1453
1	0.71	0.62	0.66	647
accuracy			0.80	2100
macro avg	0.77	0.75	0.76	2100
weighted avg	0.80	0.80	0.80	2100

Table 35

Rf train precision	0.71
Rf train recall	0.62
Rf train f1	0.66
Rf accuracy	0.80

Test Data

AUC : 0.81



Confusion Matrix

array([[544, 79],
       [116, 161]])

Table 36 Classification Report Test RF

	precision	recall	f1-score	support
0	0.82	0.87	0.85	623
1	0.67	0.58	0.62	277
accuracy			0.78	900
macro avg	0.75	0.73	0.74	900
weighted avg	0.78	0.78	0.78	900

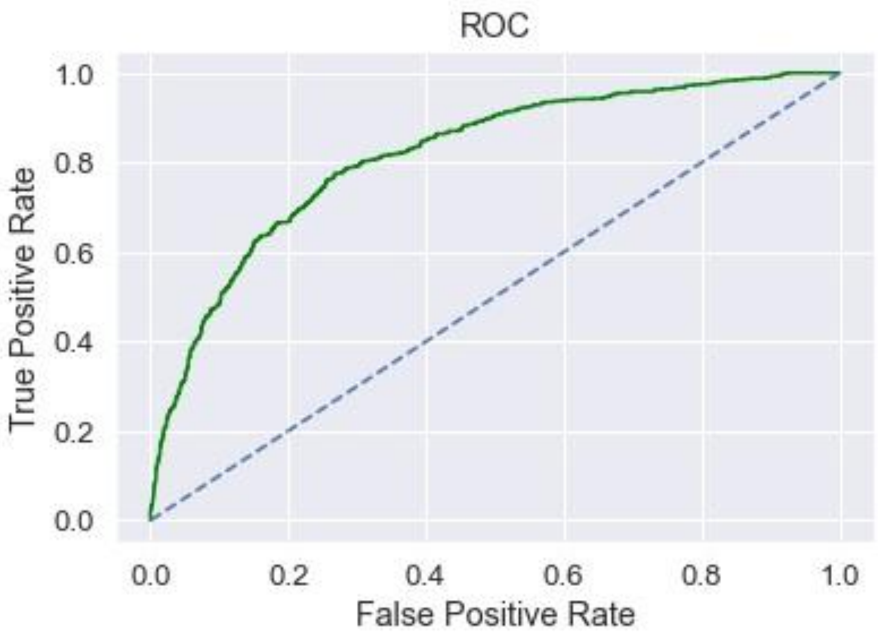
Table 37

Rf test precision	0.67
Rf test recall	0.58
Rf test f1	0.6
Rf accuracy	0.78

Artificial Neural Networks

Train Data

AUC : 0.81



Confusion Matrix

array([[1298, 155],
       [315, 332]])

Table 38 Classification Report Train NN

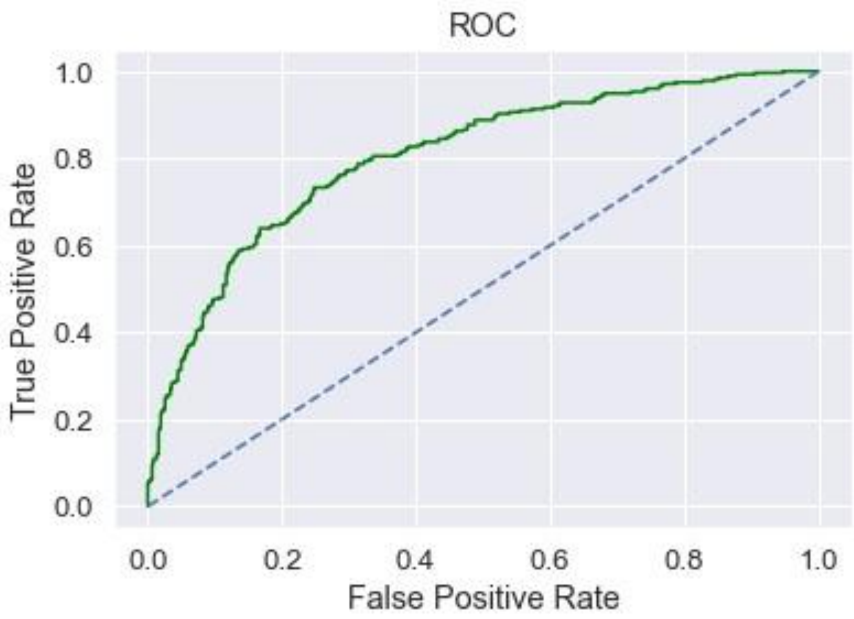
	precision	recall	f1-score	support
0	0.80	0.89	0.85	1453
1	0.68	0.51	0.59	647
accuracy			0.78	2100
macro avg	0.74	0.70	0.72	2100
weighted avg	0.77	0.78	0.77	2100

Table 39

Nn train_precision	0.68
Nn train_recall	0.51
Nn train_f1	0.59
Nn accuracy	0.77

Test Data

AUC : 0.80



Confudion Matrix

array([[553, 70],
       [138, 139]])

Table 40 Classification Report Test NN

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

Table 41

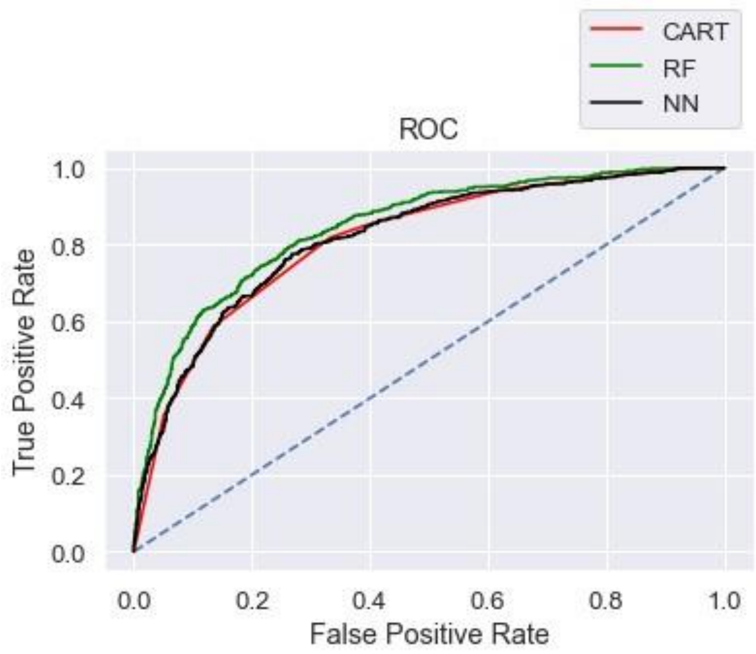
Nn test precision	0.67
Nn test recall	0.5
Nn test f1	0.57
Nn accuracy	0.77

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

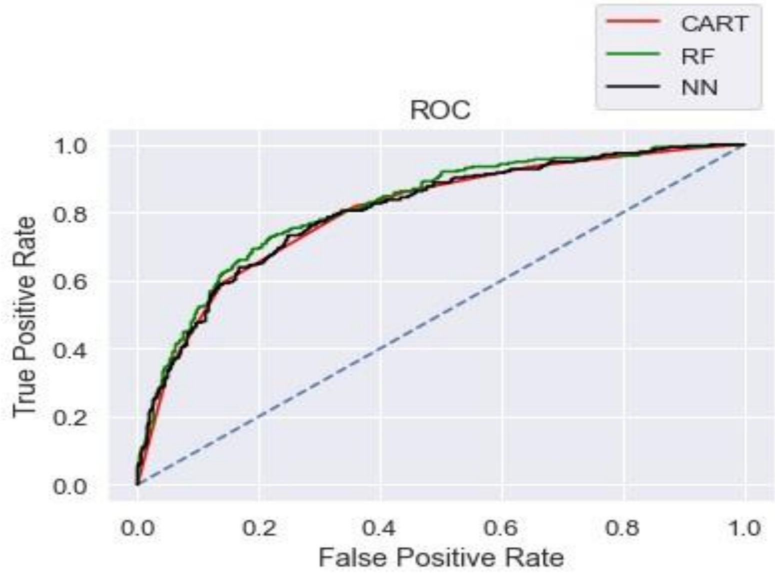
Table 42 Comparison Table of all models

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.78	0.78	0.80	0.78	0.78	0.77
AUC	0.81	0.80	0.84	0.82	0.82	0.80
Recall	0.59	0.59	0.62	0.58	0.51	0.50
Precision	0.66	0.65	0.71	0.67	0.68	0.67
F1 Score	0.62	0.62	0.66	0.62	0.59	0.57

Train ROC



Test ROC



We can select the RF model, as it has better accuracy, precision, recall, f1 score better than other two CART & NN.

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

Collecting more real time unstructured data and past data will be helpful.
This is understood by looking at the insurance data by drawing relations between different variables such as day of the incident, time, age group, and associating it with other external information such as location, behavior patterns, weather information, airline/vehicle types, etc.

Streamlining online experiences benefitted customers, leading to an increase in conversions, which subsequently raised profits. As per the data 90% of insurance is done by online channel. Other interesting fact, is almost all the offline business has a claimed associated, need to find why? Need to train the JZI agency resources to pick up sales as they are in bottom, need to run promotional marketing campaign or evaluate if we need to tie up with alternate agency Also based on the model we are getting 80%accuracy, so we need customer books airline tickets or plans, cross sell the insurance based on the claim data pattern. Other interesting fact is more sales happen via Agency than Airlines and the trend shows the claim are processed more at Airline.

Key performance indicators of insurance claims are:

- Reduce claims cycle time
- Increase customer satisfaction
- Combat fraud
- Optimize claims recovery

Reduce claim handling costs Insights gained from data and AI-powered analytics could expand the boundaries of insurability, extend existing products, and give rise to new risk transfer solutions in areas like a non-damage business interruption and reputational damage.