GREAT LEARNING

POST GRADUATE PROGRAM IN DATA SCIENCE & BUSINESS ANALYTICS

BUSINESS REPORT

CASE STUDIES ON:

- Clustering – Bank Credit Card
- CART-RF-ANN – Insurance Claim Status

Submitted By:

STEFFIN JOHN

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CASE STUDY

1. **Clustering – Bank Credit Card**
2. **CART-RF-ANN – Insurance Claim Status**

Case Study 1 - Clustering

Overview:

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.

Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

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

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

Table 1: Top five Rows

These are the top 5 rows of the data, with double digit values in Spending and advance_payments, single digit values in current_balance, credit_limit, min_payment_amt and max_spent_in_single_shopping, and point values in probability_of_full_payment.

The shape of the data is (210,7)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   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
dtypes: float64(7)
memory usage: 11.6 KB
```

Table 2: Information of Data

There are no null values in the dataset and all the values are of Float data type.

Since the difference between the 25th percentile and the minimum value of Probability_of_full_payment is so great, we may infer that there must be some outliers in this column after viewing the descriptive analysis of the data.

Additionally, because the Probability_of_full_payment data range has more weight from its 25th percentile to median than from its median to 75th percentile, the box plot will be right-skewed.

Similarly, we may infer that there will be outliers in the data column for min_payment_amt because the difference is quite high.

	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

Table 3: Description of Data

Before looking at the outliers in the dataset, we confirmed if we have any duplicate values in the data. There are no Duplicate values in the dataset.

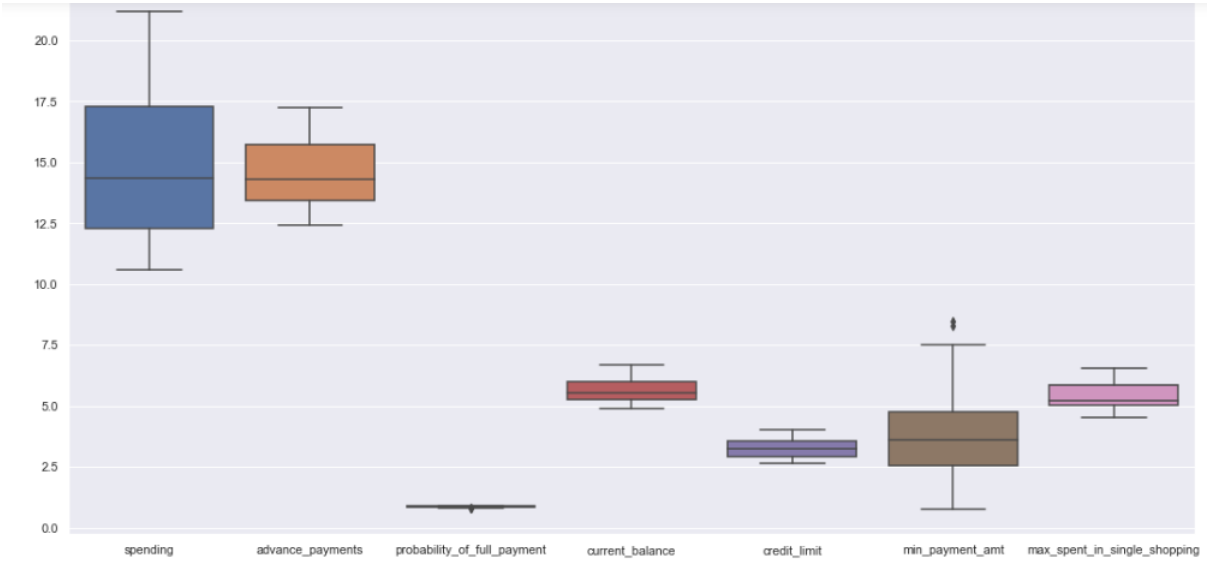

Figure 1: Boxplot of Data

As analysed in the description table of the data, there are outliers in both probability_of_full_payment and min_payment_amt. However, the number of outliers is quite less, therefore, I don't see a point in treating the outliers as treating them may disturb the data.

Apart from this, we can see that in the above figure that almost all the box plots seem to be positively skewed. Let's have a look at variable skewness in the next code.

```
max_spent_in_single_shopping    0.561897
current_balance                 0.525482
min_payment_amt                 0.401667
spending                        0.399889
advance_payments                0.386573
credit_limit                    0.134378
probability_of_full_payment    -0.537954
dtype: float64
```

Table 4: Skewness of Data

In the above data, we can see that apart from probability_of_full_payment, all the other variables are positively skewed. Also, we can see that the range of Skewness values lies between -0.5 and 0.5, which conveys that the distribution is approximately symmetric.

UNIVARIATE ANALYSIS

The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Spending

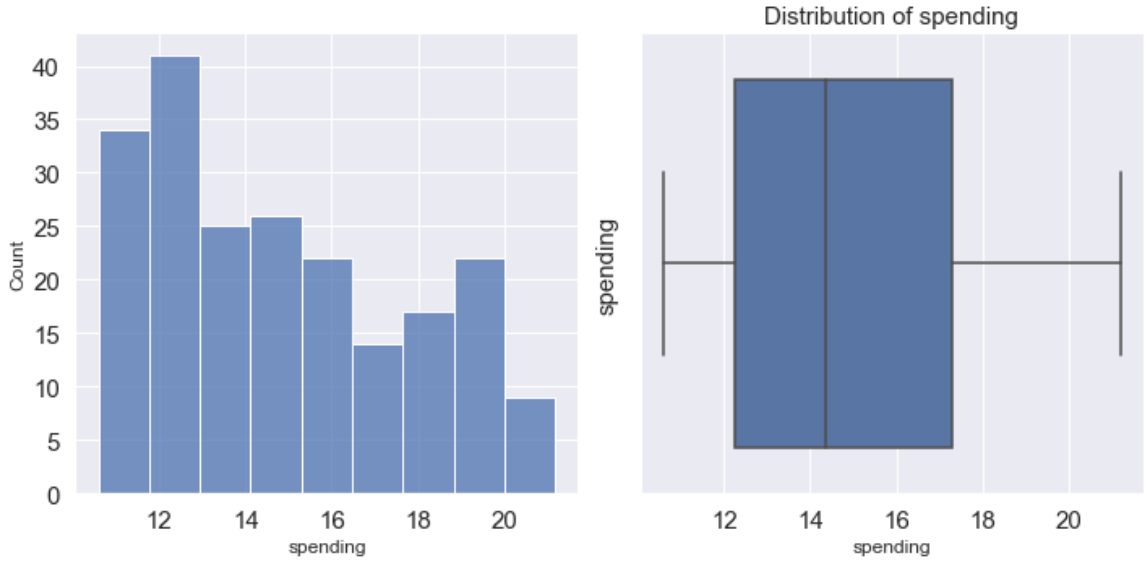

Figure 2: Histogram & Boxplot of Spending Data

Observation

1. Spending data is rightly skewed

2. No outliers in the dataset

advance_payments

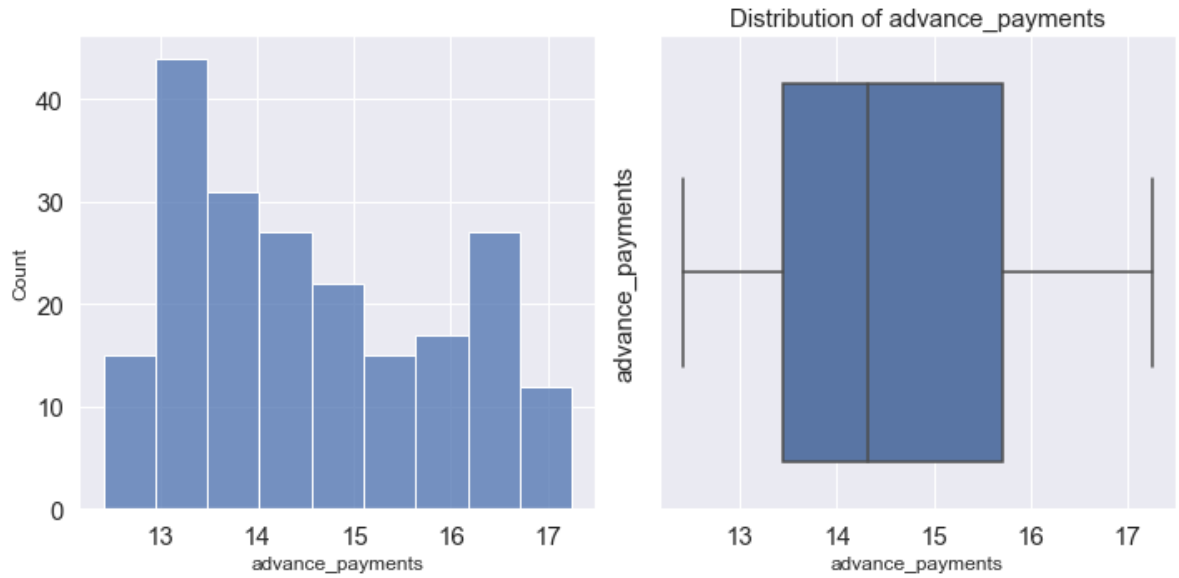

Figure 3: Histogram & Boxplot of Advance Payment Data

Observation

1. Spending data is rightly skewed

2. No outliers in the dataset

probability_of_full_payment¶

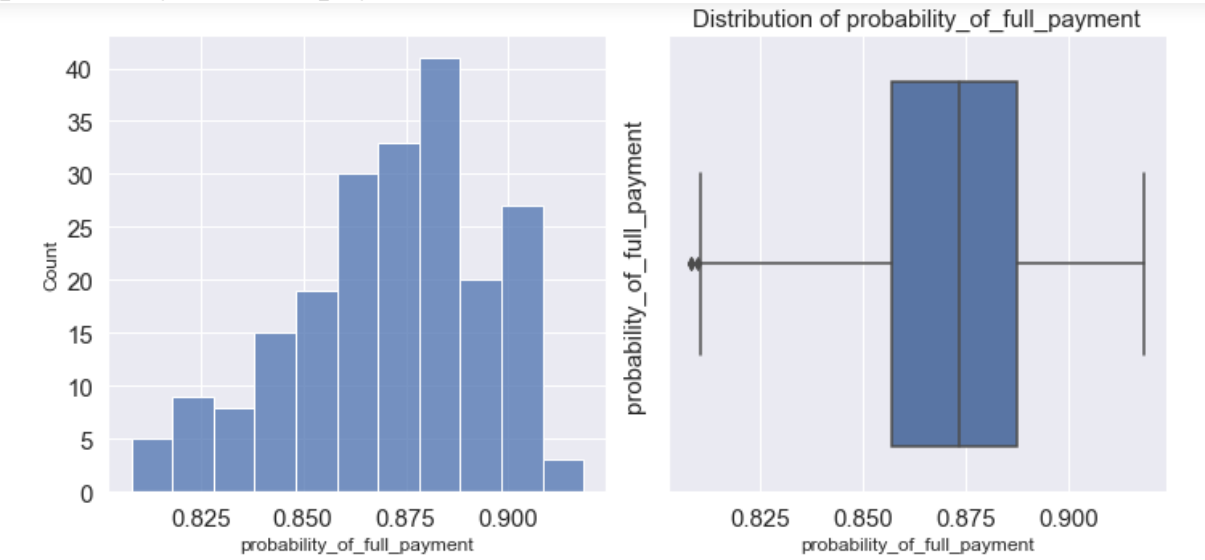

Figure 4: Histogram & Boxplot of Probability of full Payment Data

Observation

1. Spending data is slightly left skewed

2. There is 1% outliers in the dataset

current_balance

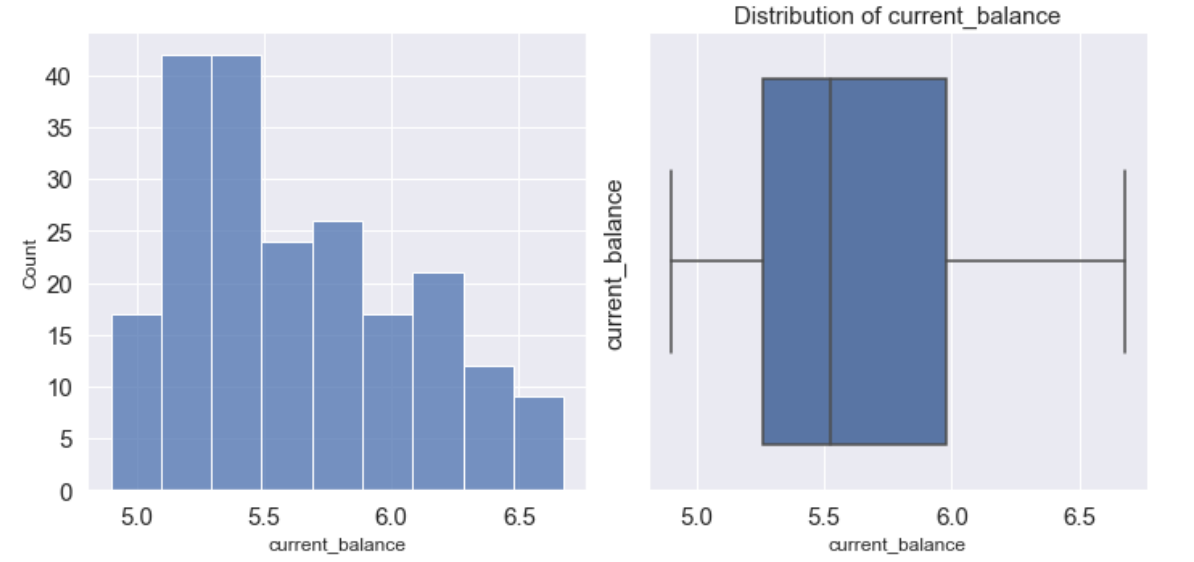

Figure 5: Histogram & Boxplot of Current Balance Data

Observation

1. Spending data is slightly rightly skewed

2. No outliers in the dataset

credit_limit

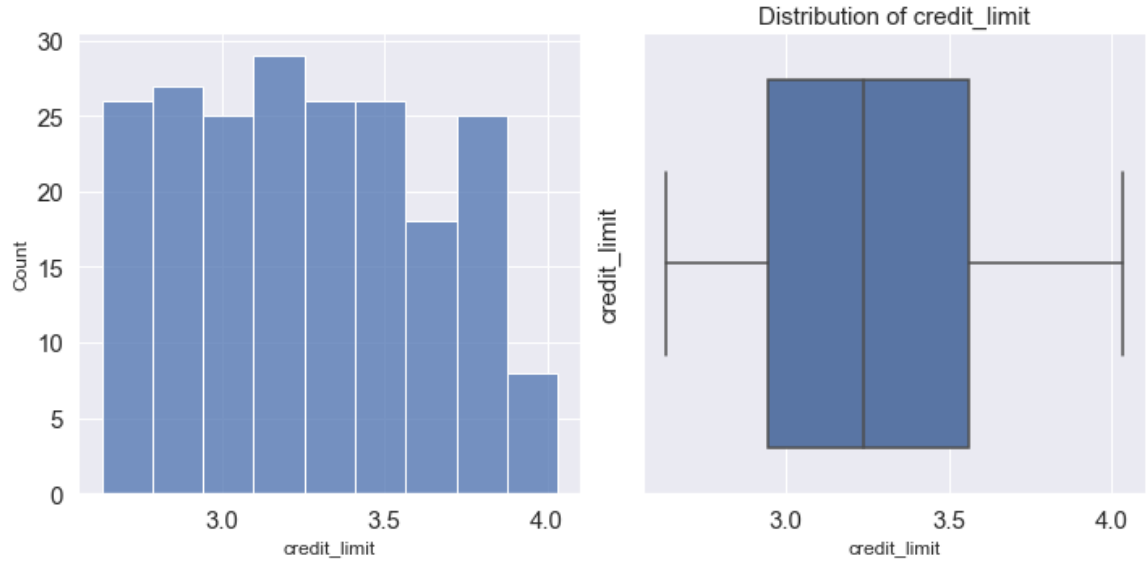

Figure 6: Histogram & Boxplot of Credit limit Data

Observation

1. No outliers in the dataset

min_payment_amt

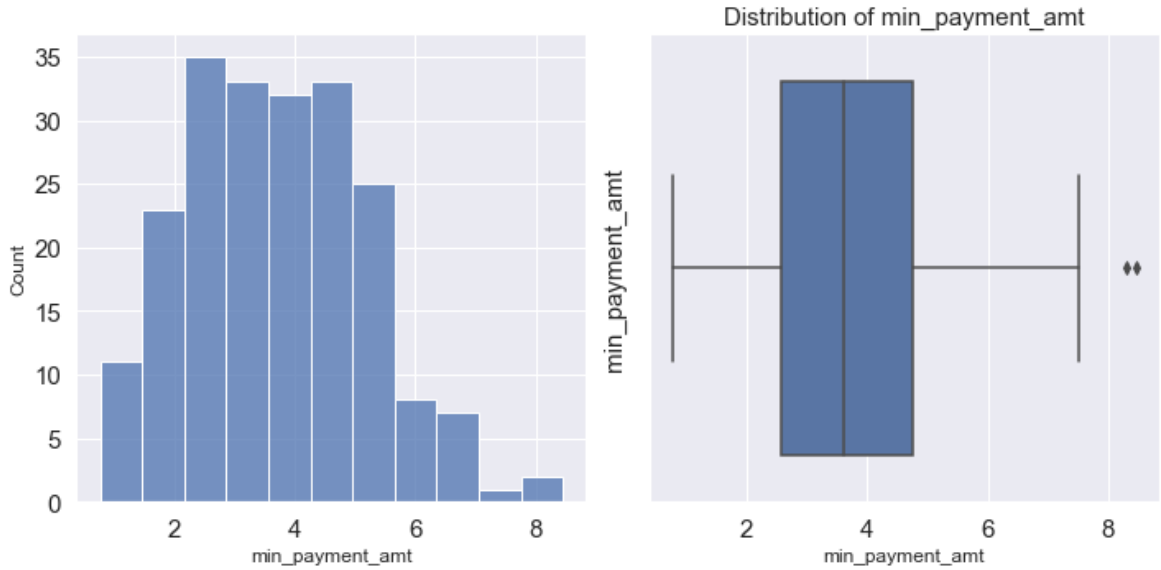

Figure 7: Histogram & Boxplot of Min Payment Amount Data

Observation

1. Spending data is slightly right skewed

2. There is 1% outliers in the dataset

max_spent_in_single_shopping

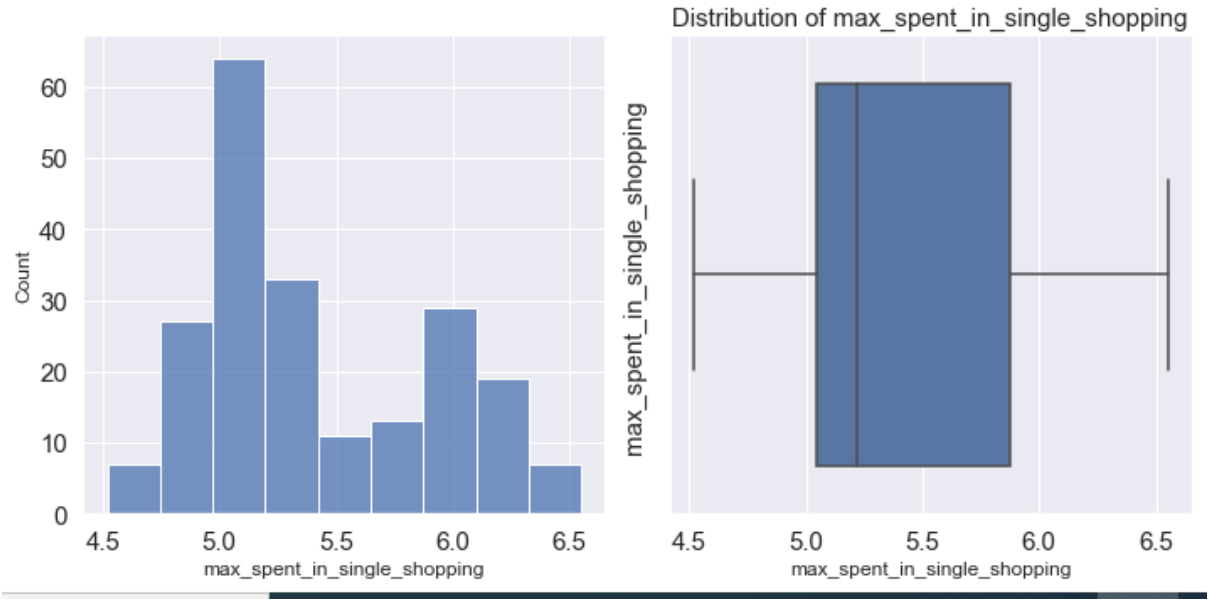

Figure 8: Histogram & Boxplot of Max Payment in Single Shopping Data

Observation

1. Spending data is slightly rightly skewed

2. No outliers in the dataset

MULTIVARIATE ANALYSIS

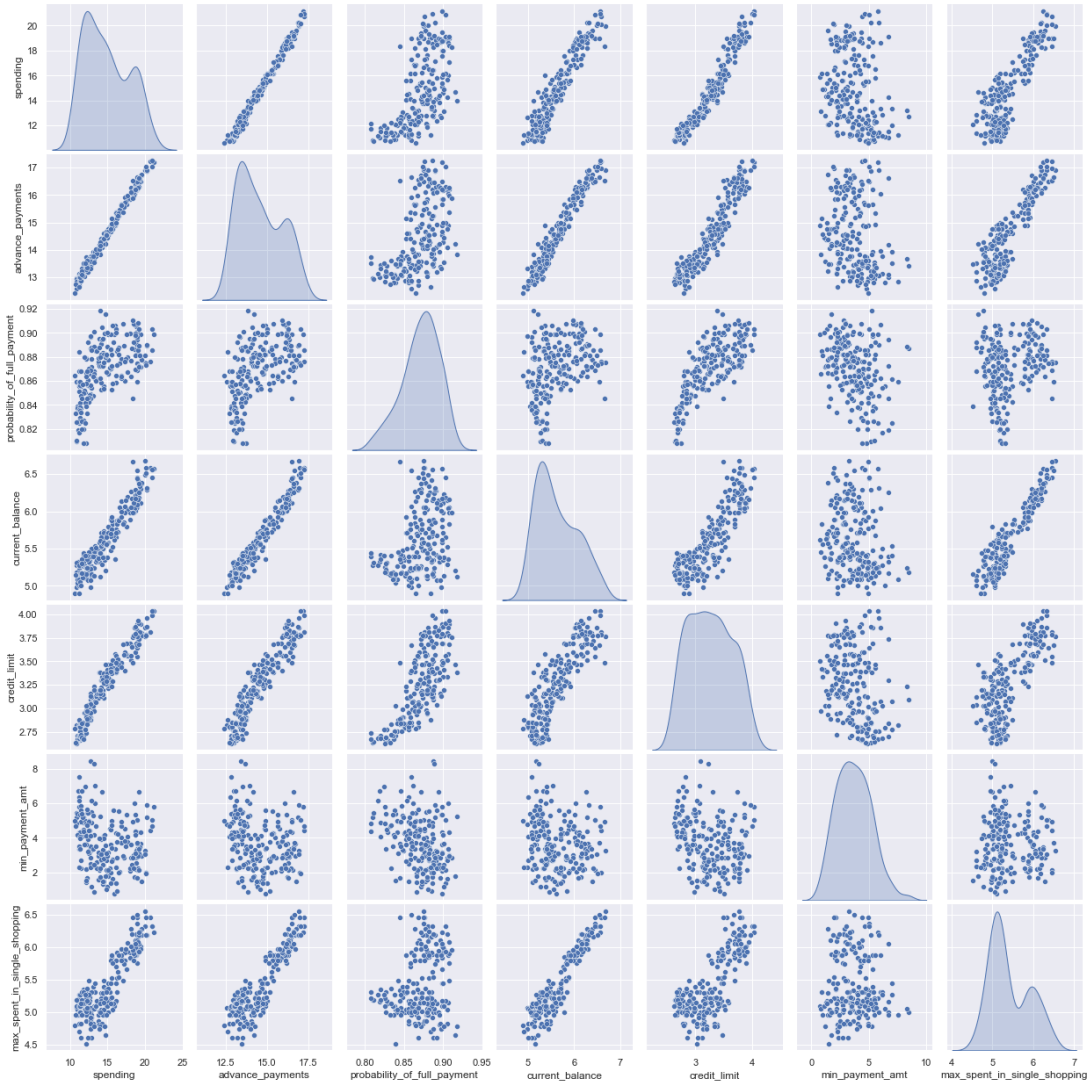

Figure 9: PairPlot for checking multi-Collinearity

We may infer utilising scatterplots for all the variables after performing multivariate analysis on the dataset's variables.

Many of the factors we can see in this have substantial correlations; for further information, let's look at the heatmap and the correlation table.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
spending	1.000000	0.994341	0.608288	0.949985	0.970771	-0.229572	0.863693
advance_payments	0.994341	1.000000	0.529244	0.972422	0.944829	-0.217340	0.890784
probability_of_full_payment	0.608288	0.529244	1.000000	0.367915	0.761635	-0.331471	0.226825
current_balance	0.949985	0.972422	0.367915	1.000000	0.860415	-0.171562	0.932806
credit_limit	0.970771	0.944829	0.761635	0.860415	1.000000	-0.258037	0.749131
min_payment_amt	-0.229572	-0.217340	-0.331471	-0.171562	-0.258037	1.000000	-0.011079
max_spent_in_single_shopping	0.863693	0.890784	0.226825	0.932806	0.749131	-0.011079	1.000000

Table 5: Correlation table

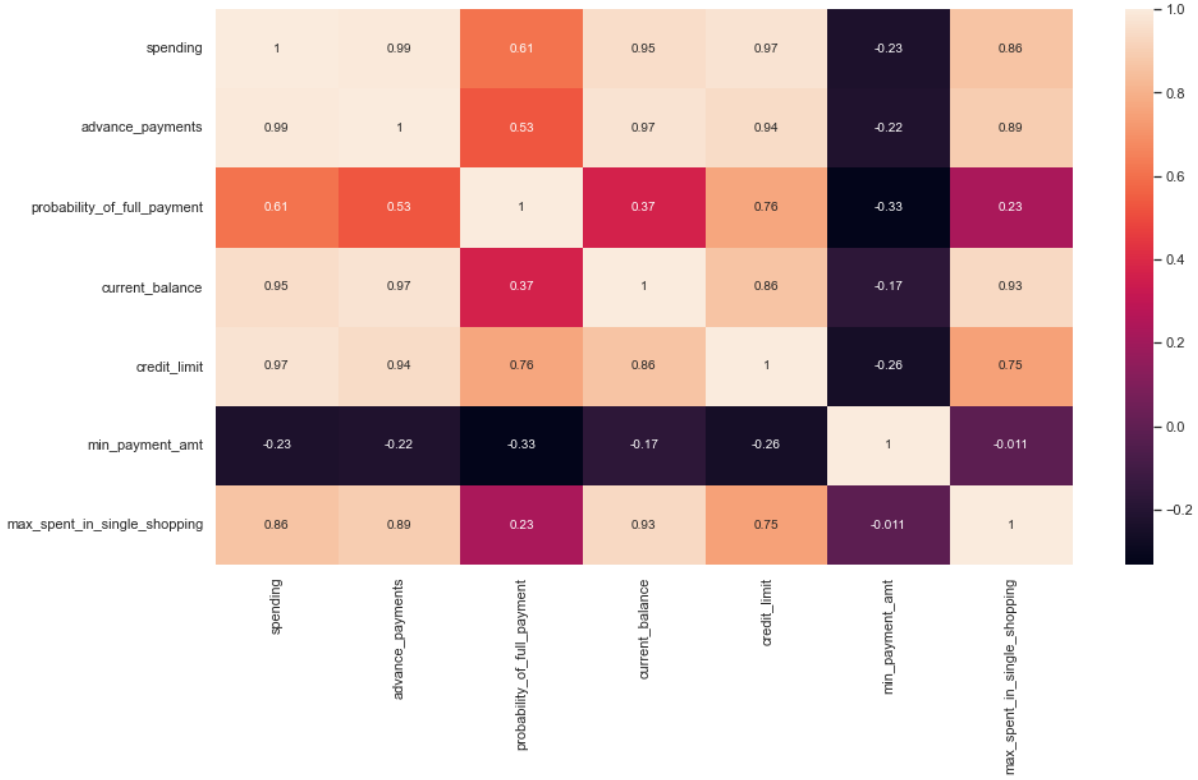

Figure 10: Heatmap

Observation

1. Strong positive correlation between:

- spending & advance_payments,
- advance_payments & current_balance,
- credit_limit & spending
- spending & current_balance
- credit_limit & advance_payments
- max_spent_in_single_shopping & current_balance

2. min_payment_amt is the weakest correlation

Observing the Correlations in the dataset, we infer that high multicollinearity exists within any two independent variables. However, we will not be treating multi-collinearity, as it does not impact the clustering process.

BI-VARIATE ANALYSIS

Let's check the Strong Positive Correlations

spending & advance_payments

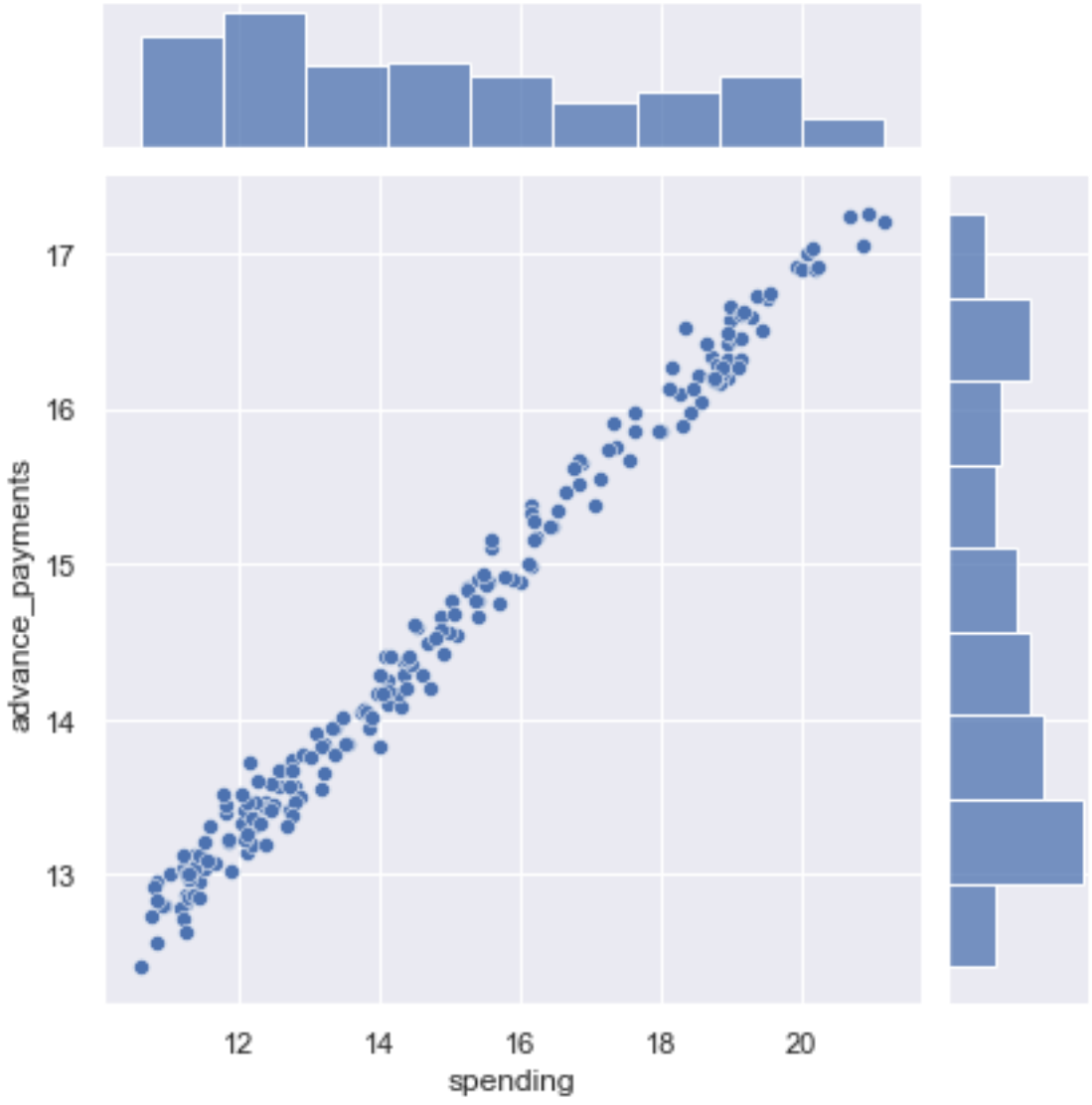

Figure 11: Bi-Variate Relationship b/w Spending & Advance Payment

advance_payments & current_balance

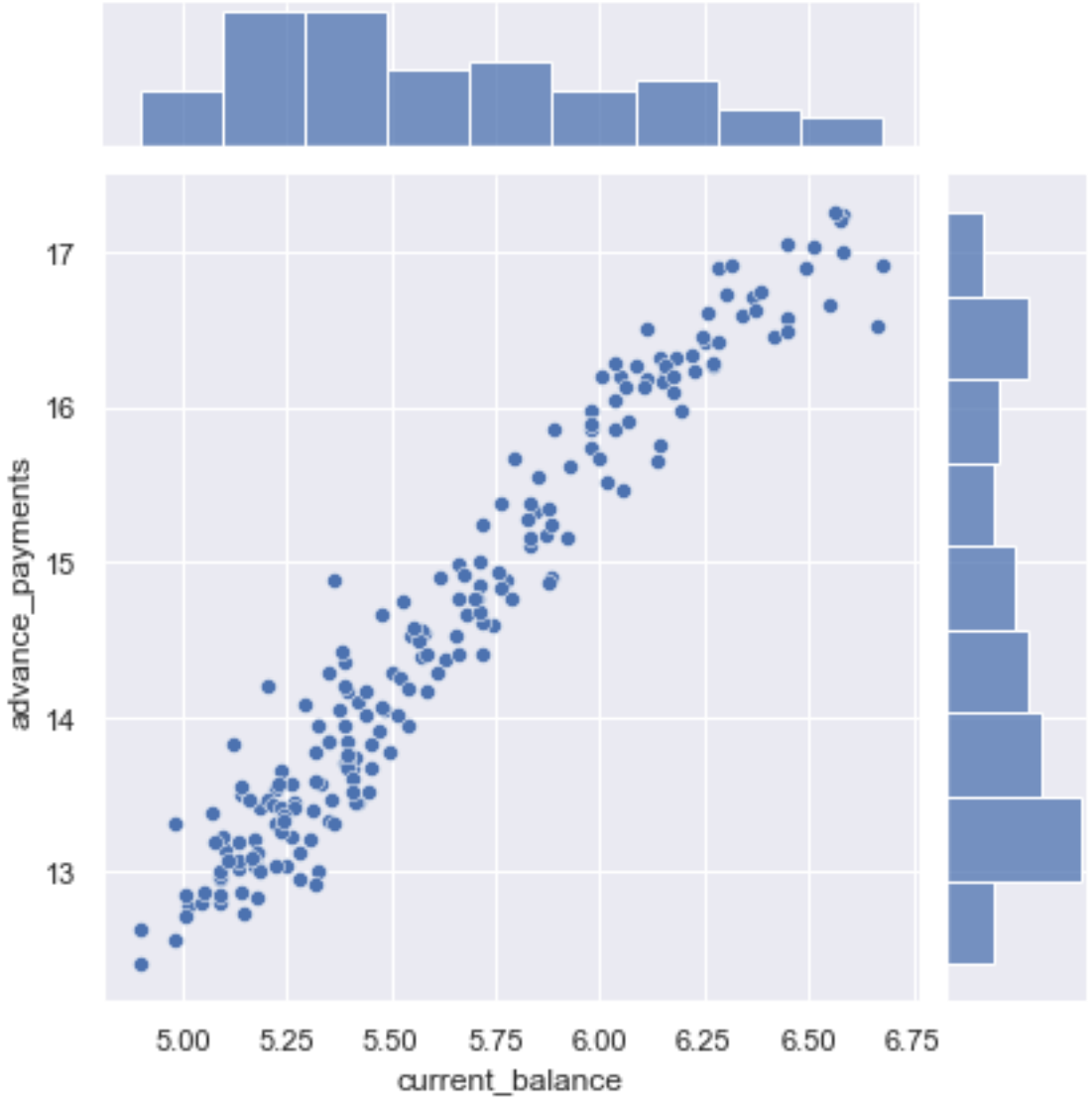

Figure 12: Bi-Variate Relationship b/w Current Balance & Advance Payment

credit_limit & spending

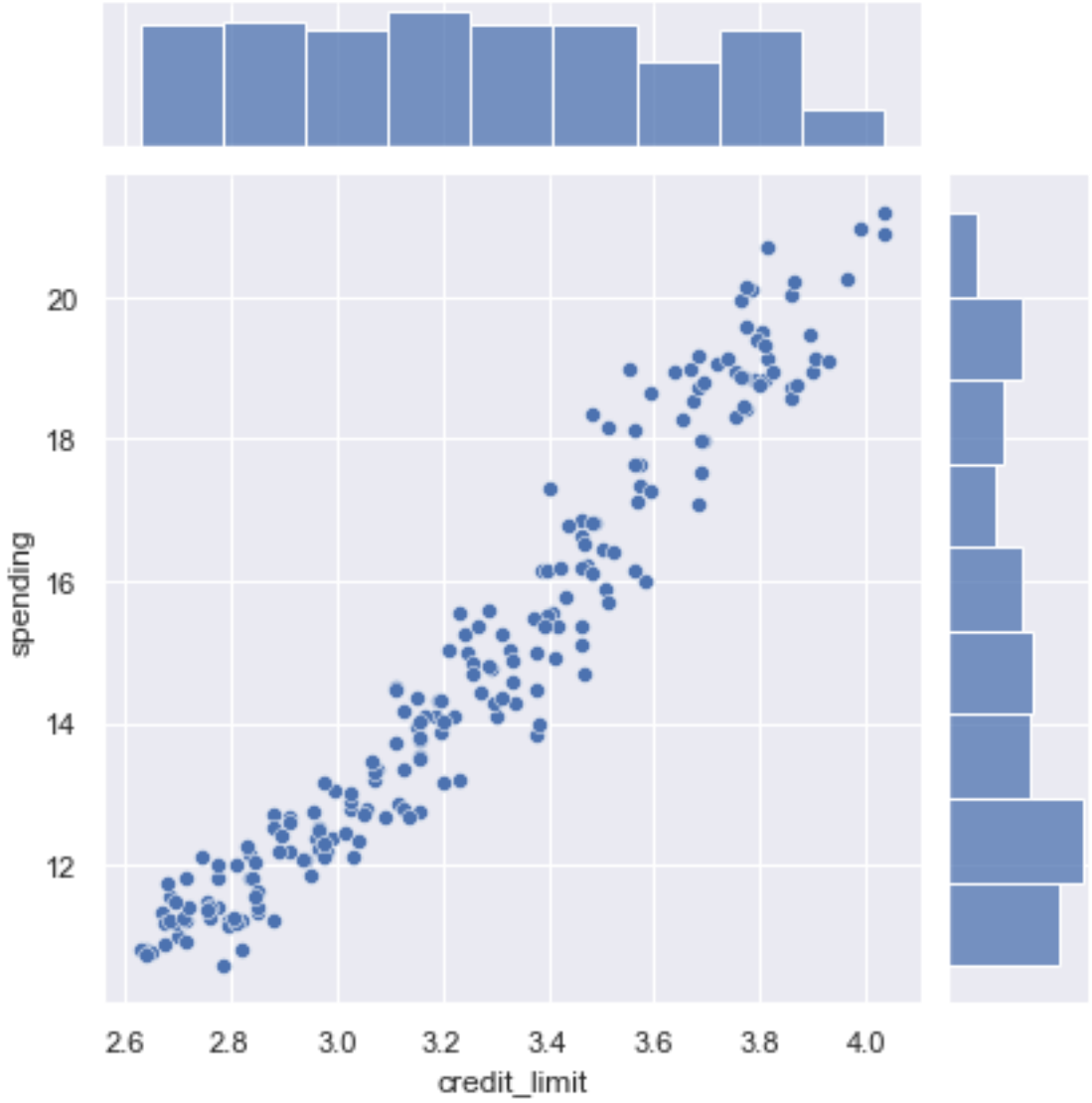

Figure 13: Bi-Variate Relationship b/w Credit Limit & Spending

spending & current_balance

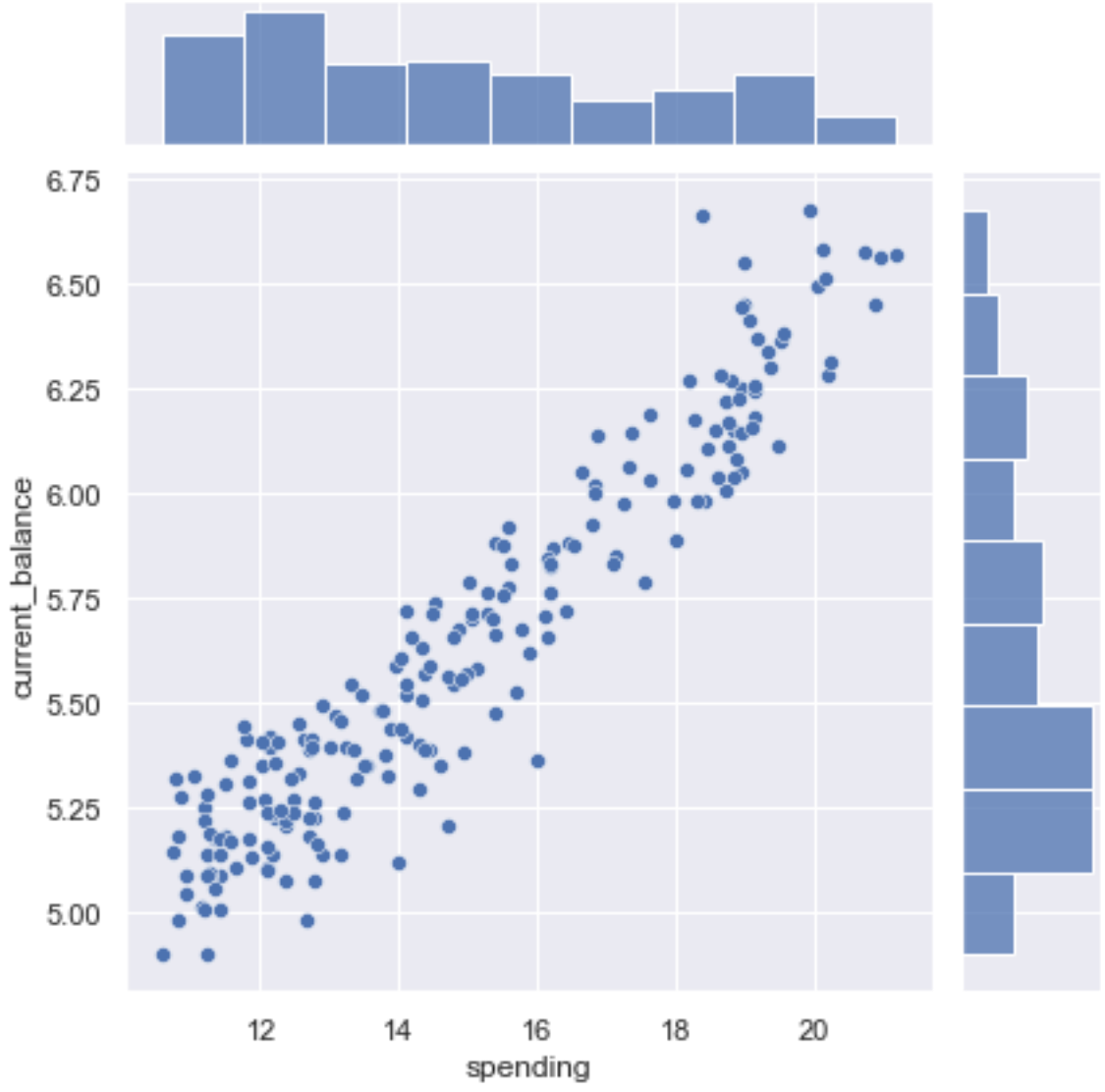

Figure 14: Bi-Variate Relationship b/w Current Balance & Spending

credit_limit & advance_payments

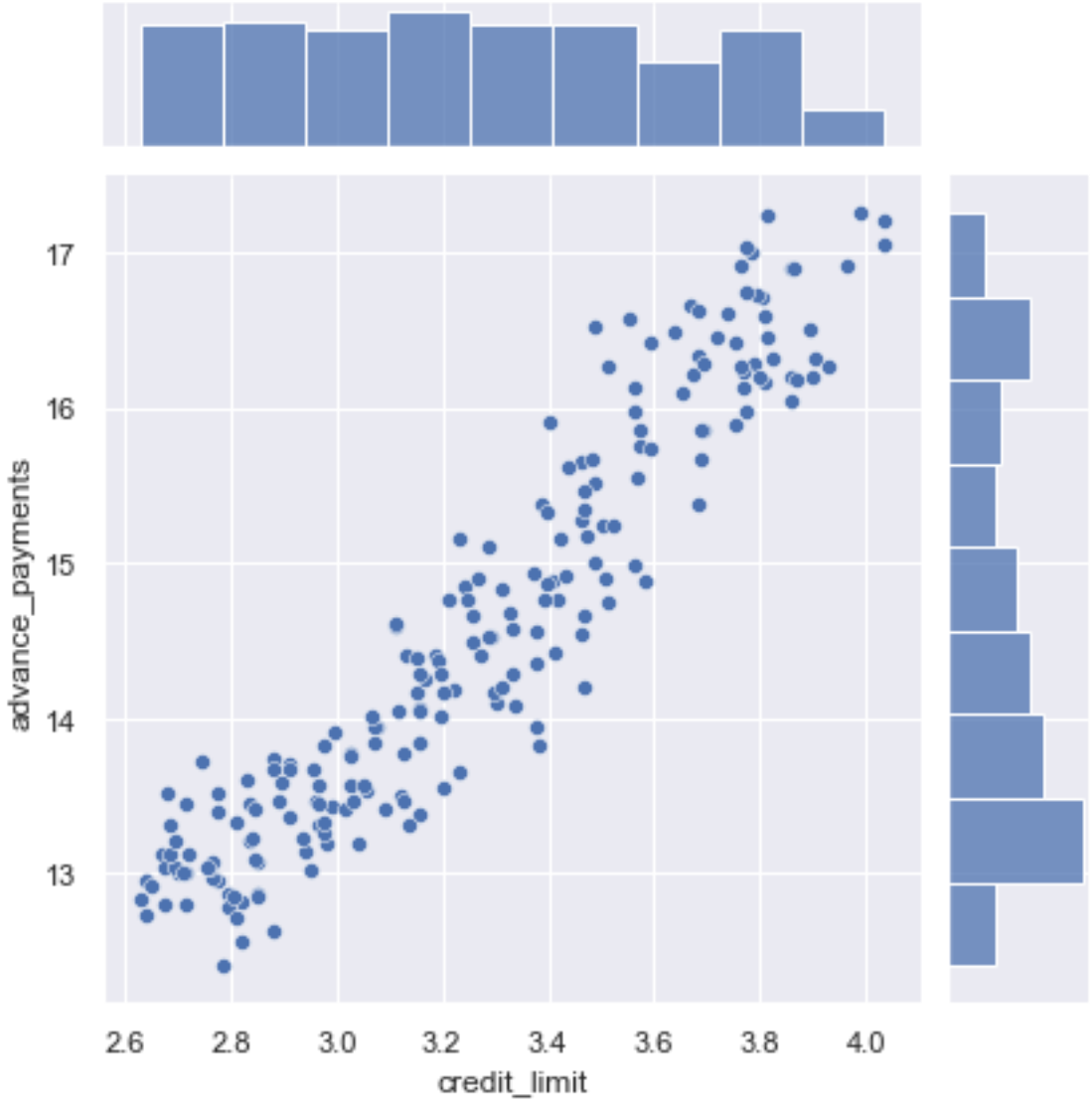

Figure 15: Bi-Variate Relationship b/w Credit limit & Advance Payment

max_spent_in_single_shopping & current_balance

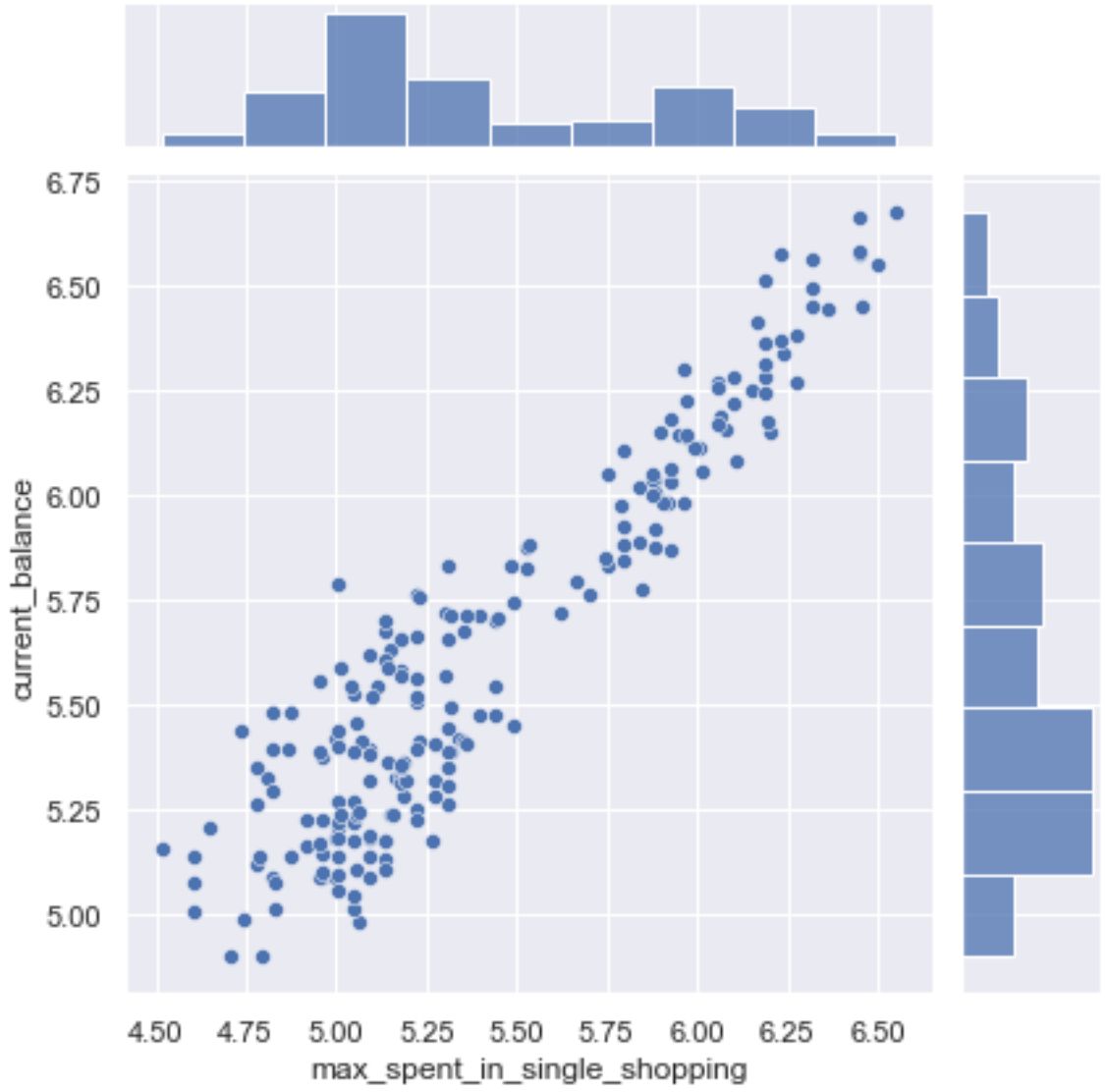

Figure 16: Bi-Variate Relationship b/w Current Balance & Max Spend in Single Shopping

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

Yes, I think that Scaling is necessary for clustering in this case as standardising the data prevents variables with larger scales from dominating the clustering process.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.178230	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	1.501773	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	0.504874	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-2.591878	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	1.196340	0.591544	1.155464	-1.088154	0.874813

Table 6: Scaled dataframe table

As we can see from the data spending, advance_payments are in different values and this may get more weightage.

Scaling will help keep the values in relatively same range.

Let's see how the data look before and after scaling in from of a plot graph.

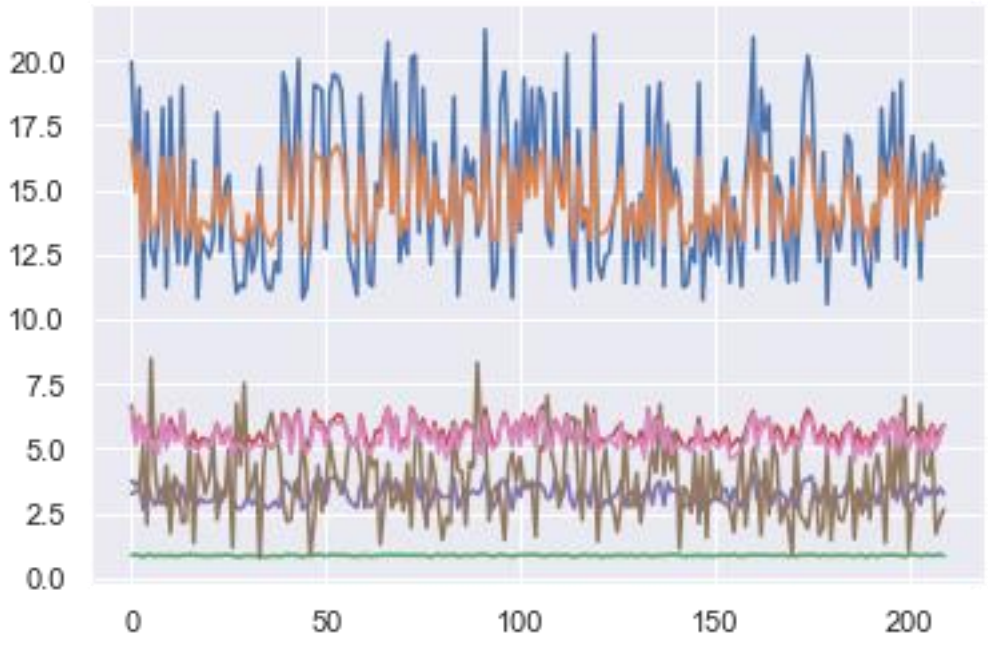

Figure 17: Graphical data before Scaling

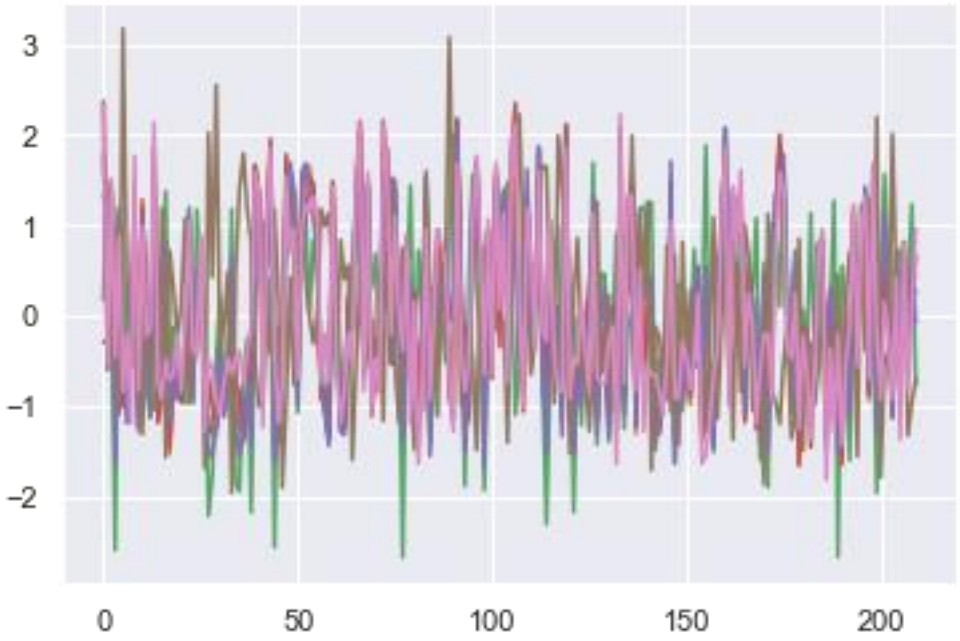

Figure 18: Graphical data after Scaling

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

Using various links and distance criterion, we have constructed a Dendrogram for the scaled data from which we received several cluster patterns. Following are analysis of all the dendrogram, the Ward approach had been selected:

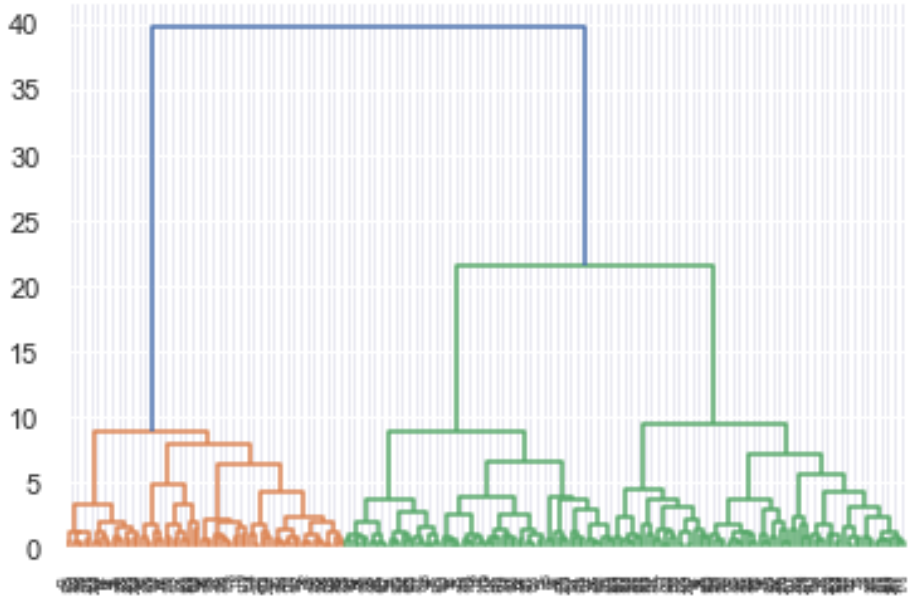

Figure 19: Dendrogram (Full)

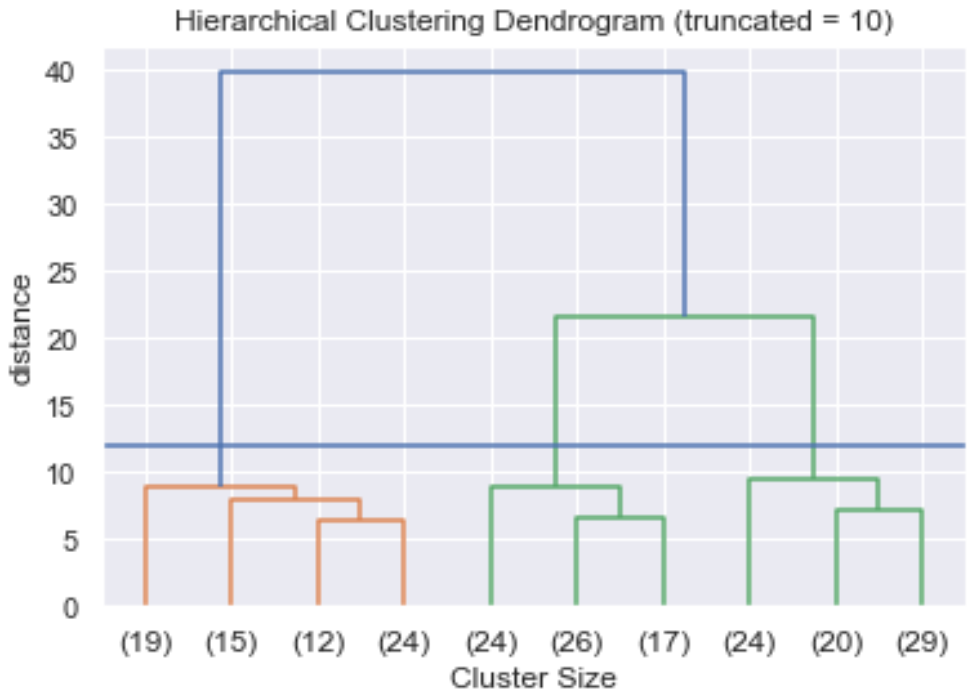

Figure 20: Dendrogram (truncated = 10)

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The result is a collection of clusters, each of which is distinct from the others and contains objects that are largely similar to one another.

After applying hierarchical clustering to a scaled dataset, we receive the following mean values for each of the three cluster formations:

clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	frequency
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	70
2	11.872388	13.257015	0.848072	5.238940	2.848537	4.949433	5.122209	67
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178	73

Table 7: Clusters table with Frequency

Let’s have a visual representation of the clustering applied on the dataset:

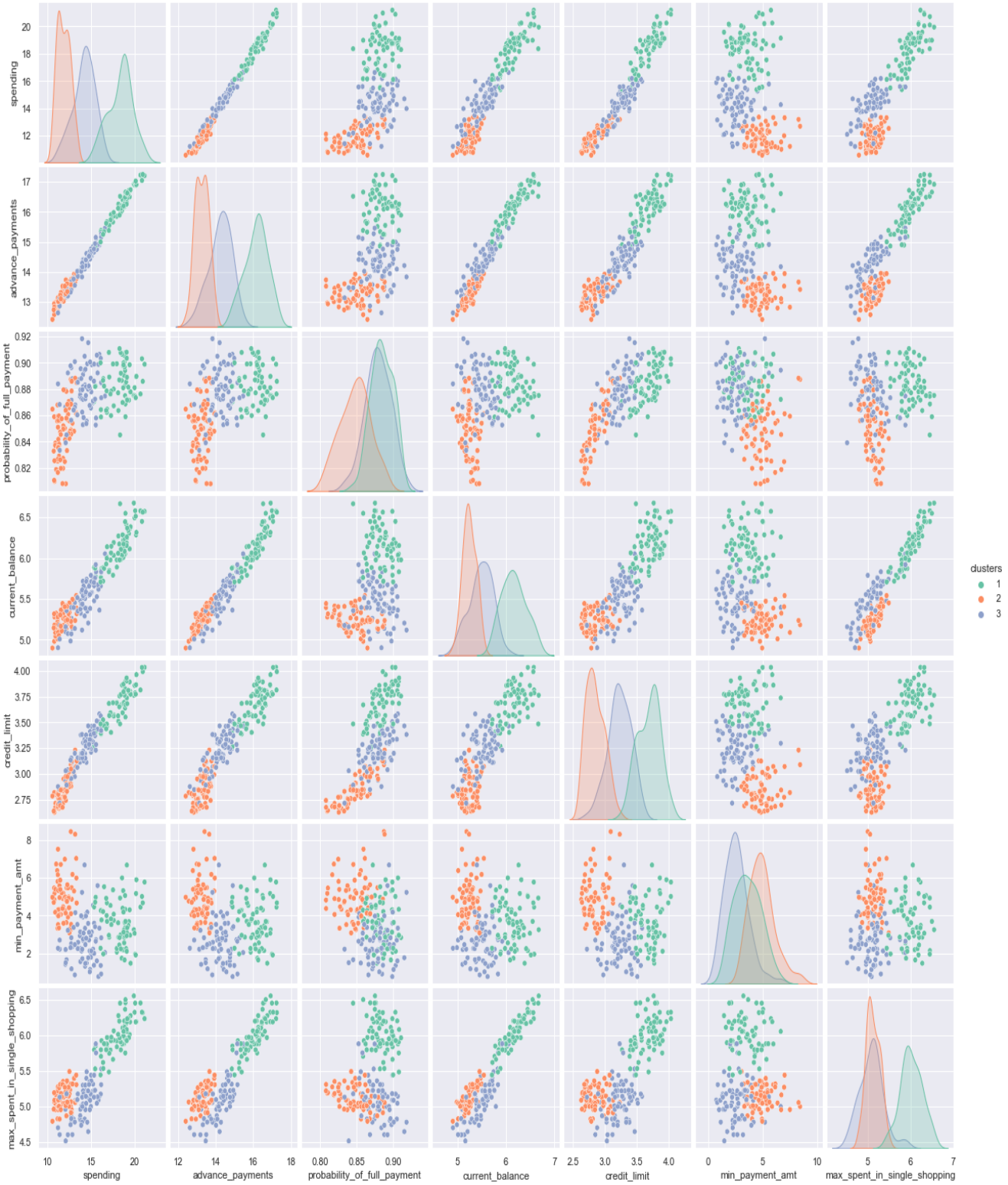

Figure 21: Visual representation of Clusters

For a more focused view of Probability_of_full_payment cluster:

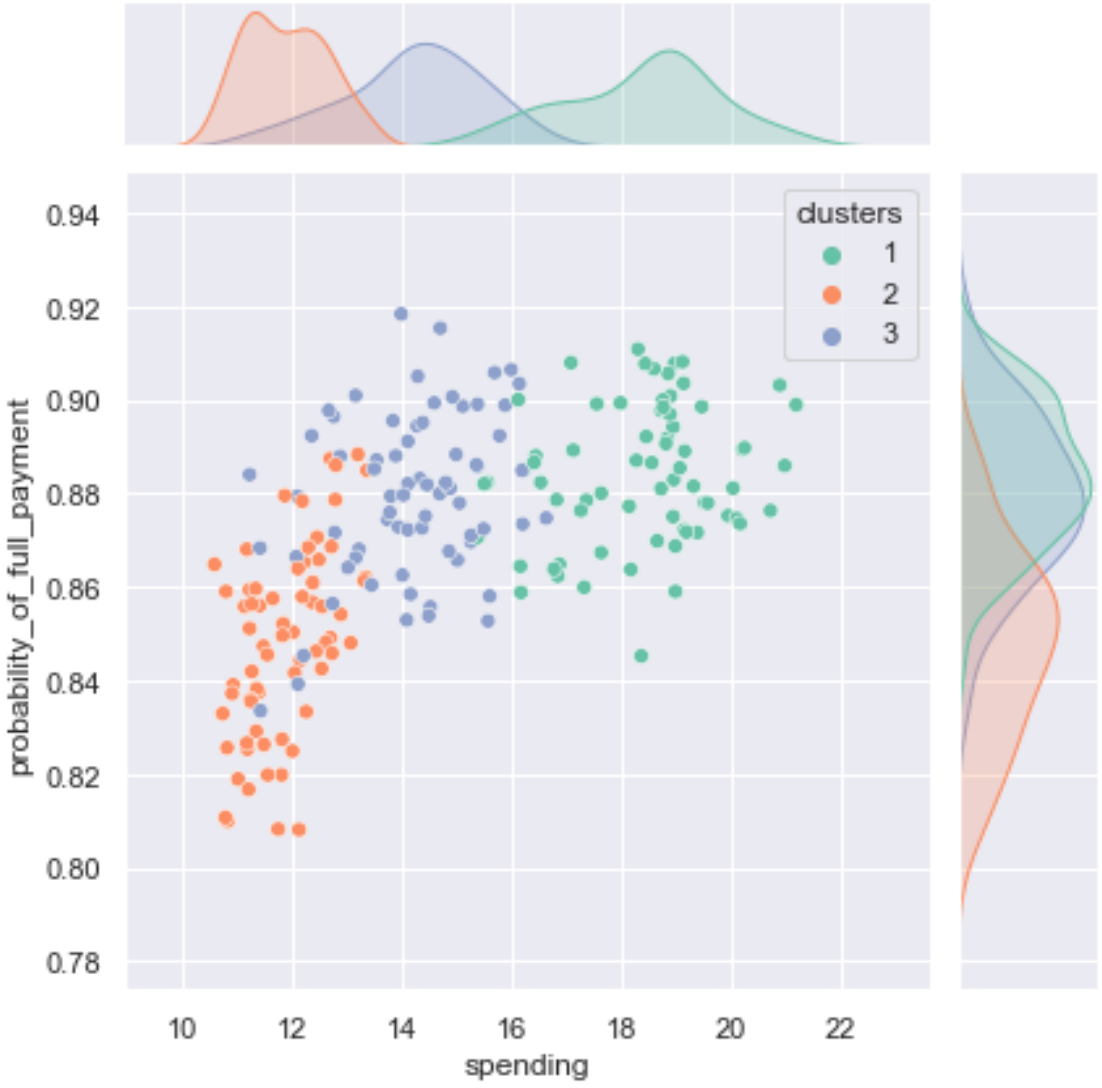

Figure 22: Visual representation of Cluster Spending & Probability of full payment

Observation

1. Based on current dataset given, 3 cluster solution makes sense based on the spending pattern

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

2. As we can observe from the 3 cluster segmentations, the customers under the high spender's cluster have higher valuations and probabilities across the various criteria mentioned except the min_payment_amt where the customers of the low spenders cluster have a higher bill value amount as their minimum amount that would have to be remitted.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve. Explain the results properly. Interpret and write inferences on the finalized clusters.

K-Means is a non-hierarchical approach to forming good clusters is to prespecify a desired number of clusters, k. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

This approach divides the groupings into non-overlapping ones with no hierarchical connections between them. K-means clustering is frequently utilised in applications involving huge datasets.

We fit the scaled data using the K-Means package from SKlearn, calculating the inertia and then adding up the within-cluster sum of squares. (WSS)

Now with the help of **Elbow Curve**, for a given number of clusters, the total within-cluster sum of squares (WCSS) is computed.

That value of k is chosen to be optimum, where addition of one more cluster does not lower the value of total WSS appreciably.

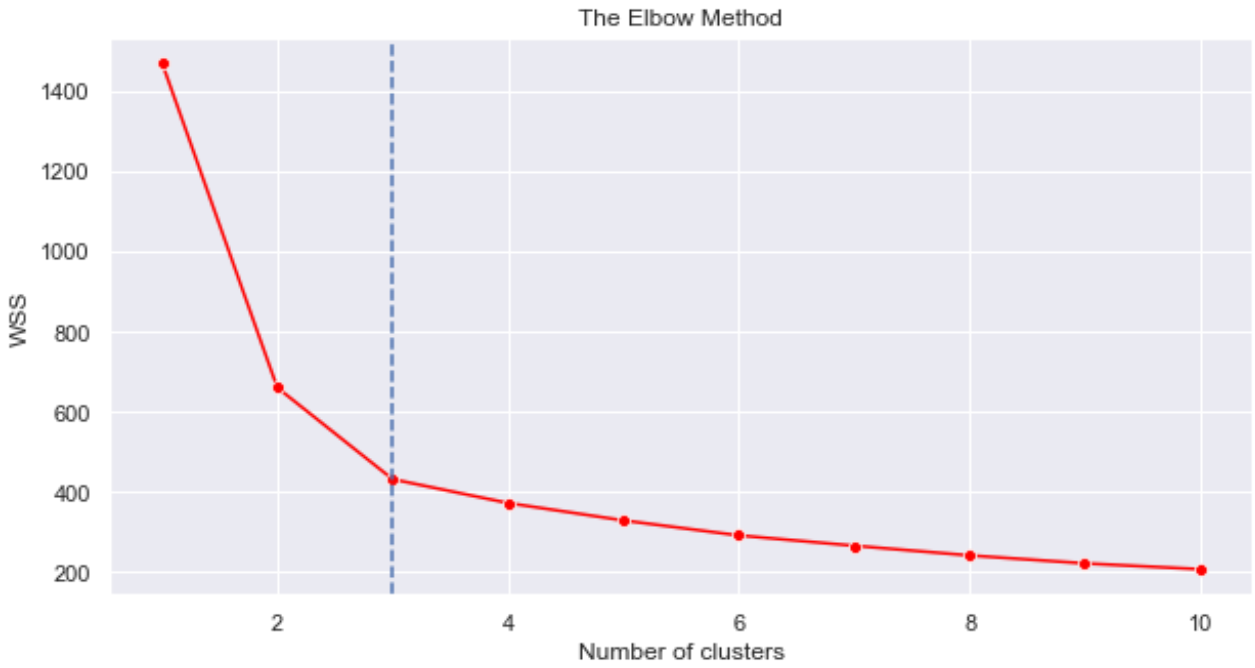

Figure 23: Elbow Curve of the Clusters

We should go with 3 cluster segmentation as per our business recommendation as we see an elbow at Cluster number = 3 after which the scree plot seems redundant.

Total 3 Cluster group makes sense based on the spending:

Group 1: High Spending

Group 2: Medium Spending

Group 3: Low Spending

Silhouette Score

This method measures how tightly the observations are clustered and the average distance between clusters. For each observation a silhouette score is constructed which is a function of the average distance between the point and all other points in the cluster to which it belongs, and the distance between the point and all other points in all other clusters, that it does not belong to. The maximum value of the statistic indicates the optimum value of k.
However, the Silhouette Score of 2 clusters was more appropriate, however, objective of this clustering effort is to devise a suitable recommendation system. It may not be practical to manage a very low number of tailor-made recommendations. Therefore, Cluster number = 3 serves the purpose of our requirement to produce valuable insights.

silhouette_score = 0.40072705527512986

Upon performing Non-Hierarchical clustering on scaled dataset, we obtain mean values within 3 cluster formations as follows:

Kmeans_Clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	k_frequency
0	11.856944	13.247778	0.848253	5.231750	2.849542	4.742389	5.101722	70
1	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701	67
2	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803	73

Table 8: Non-Hierarchical Clusters (K-Means)

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

Group 1: High Spending Group

• Offerings like higher reward points with a higher likelihood can boost their spending power.
• Including a free EMI option as part of a campaign with a bank's affiliated brands might be a terrific motivation for this demographic.
• The segmentation of this group's maximum max spent in single shopping is the highest, which explains the discounts or alluring offers made on subsequent transactions with full upfront payments.
• Regular evaluation and raising of credit limits
• Preferential customer treatment, which can encourage more extravagant spending
• Because it is obvious that the clients in this category are financially sound, it may be possible to develop appealing loan programmes just for them.
• The one-time maximum spending would increase as a result of collaborations with high end luxury goods and accessories.

Group 2 : Medium Spending Group

• Due to regular maintenance of a higher credit score and subsequent timely bill payment, it is proposed that the consumers in this segmentation cluster are the target customers with the highest potential.

• Keeping RBI requirements in mind, customers in this group might have their credit limits expanded, reviewed on a regular basis, and their interest rates significantly marginalised.

• The long-term growth in transactional values would result from the advertising and promotion of premium cards or loyalty cards of specific brand collaboration alliances.

• Once the aforementioned credit limits are raised, premium partners in e-commerce, travel portals, airlines, and hotels will automatically see an increase in spending patterns.

Group 3: Low Spending Group

• We can analyse which products and services this market segment spends the most money on and give discounts and offers on credit card usage in accordance.

• In order to prevent missed billing cycle due dates, customers in this category must receive prompt payment reminders.

• Small-scale campaigns might be undertaken to provide this segment's clients enticing incentives for making payments early. This would increase the rate of payments received and lower the default rates.

Problem 2: CART-RF-ANN

Overview:

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 and compare the models' performances in train and test sets.

Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

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

After importing the necessary libraries and data in the python notebook, below is the top 5 rows of the data.

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

Table 9: Top 5 Heads of Insurance Data

In the below image, we can see that there are no null values in the data. 2 of the 10 variables are of Data type Float, 2 variables are of Data type Integer and the remaining 6 are of Object Data type.

The shape of the data is (3000,10)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Age           3000 non-null   int64  
 1   Agency_Code   3000 non-null   object 
 2   Type          3000 non-null   object 
 3   Claimed       3000 non-null   object 
 4   Commision     3000 non-null   float64
 5   Channel       3000 non-null   object 
 6   Duration      3000 non-null   int64  
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   object 
 9   Destination   3000 non-null   object 
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

Table 10: Data Type of Insurance Data

Descriptive Analysis:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000.0	NaN	NaN	NaN	38.091	10.463518	8.0	32.0	36.0	42.0	84.0
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000.0	NaN	NaN	NaN	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000.0	NaN	NaN	NaN	70.001333	134.053313	-1.0	11.0	26.5	63.0	4580.0
Sales	3000.0	NaN	NaN	NaN	60.249913	70.733954	0.0	20.0	33.0	69.0	539.0
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 11: Descriptive Analysis of Insurance Data

For Object data type variables like, Agency_code, Type, Claimed, Channel, Product Name, and Destination, there are very less unique values.

The topmost frequent value of:

- Agency_code is **EPX** with a frequency of 1365
- Type is **Travel Agency** with a frequency of 1837
- Claimed is **No** with a frequency of 2076
- Channel is **Online** with a frequency of 2954
- Product Name is **Customised Plan** with a frequency of 1136
- Destination is **ASIA** with a frequency of 2465

For the float and integers data type values like: Age, Commision, Duration and Sales the difference between its 75th percentile and Max value is very large, indicating there will be large number of outliers in the data.

After this, let's check if we have any duplicates in the data; We have 139 duplicate rows in the data. Though it shows there are 139 records, but it can be of different customers, there is no customer ID or any unique identifier, so I am not dropping them off.

Then we analysed the Object data type variables for their count of unique values:

```
Agency Code:
EPX    1365
C2B     924
CWT     472
JZI     239
Name: Agency_Code, dtype: int64

Type:
Travel Agency    1837
Airlines         1163
Name: Type, dtype: int64

Channel:
Online     2954
Offline      46
Name: Channel, dtype: int64

Product Name:
Customised Plan      1136
Cancellation Plan     678
Bronze Plan           650
Silver Plan           427
Gold Plan             109
Name: Product Name, dtype: int64

Destination:
ASIA        2465
Americas     320
EUROPE       215
Name: Destination, dtype: int64

Claimed:
No     2076
Yes     924
Name: Claimed, dtype: int64
```

Table 12: Object Type Variable of Insurance Data

UNIVARIATE ANALYSIS:

The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Age Variable

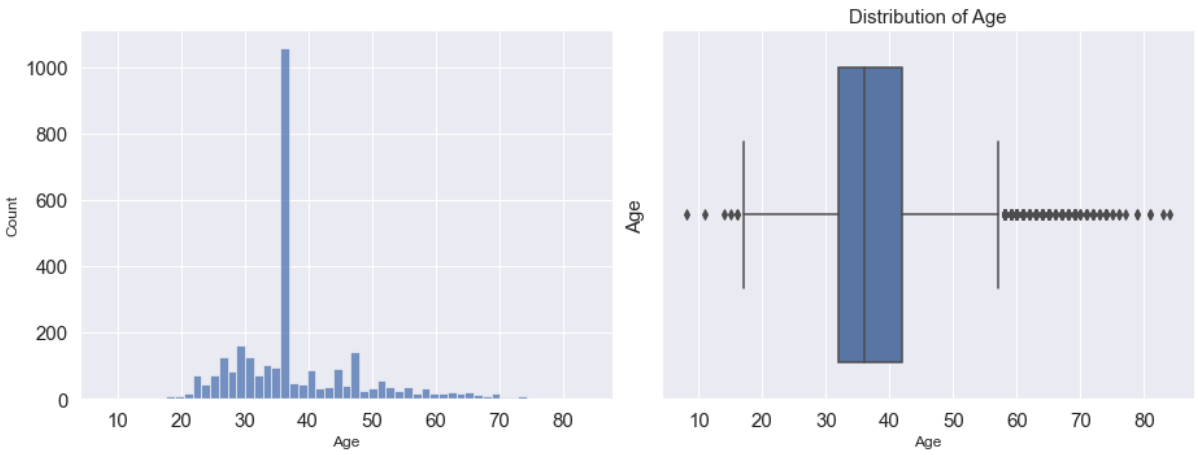

Figure 24: Histogram & Boxplot of Age Variable

Observation

1. Their 7% are outliers in the dataset

Commission Variable

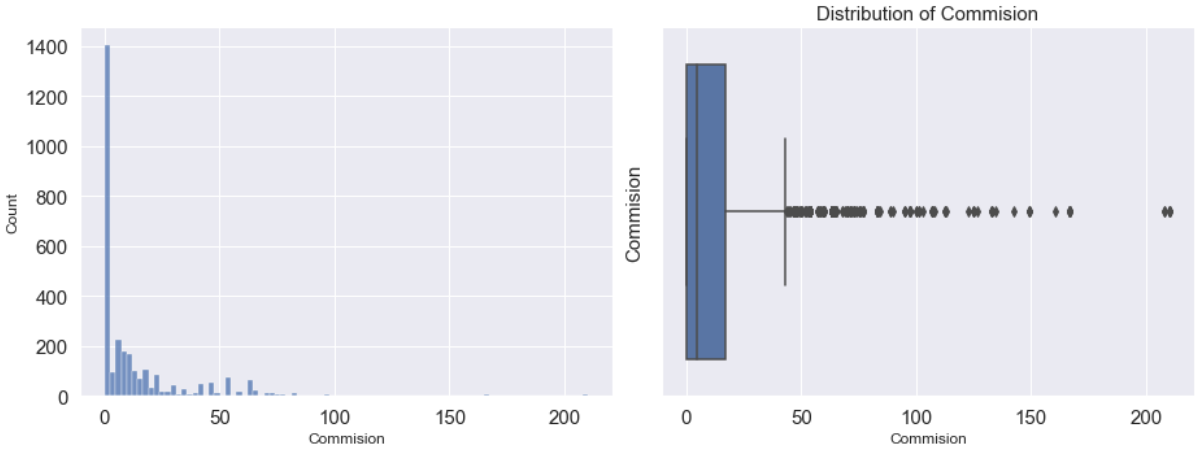

Figure 25: Histogram & Boxplot of Commission Variable

Observation

1. There are 12% outliers in the dataset

Duration variable

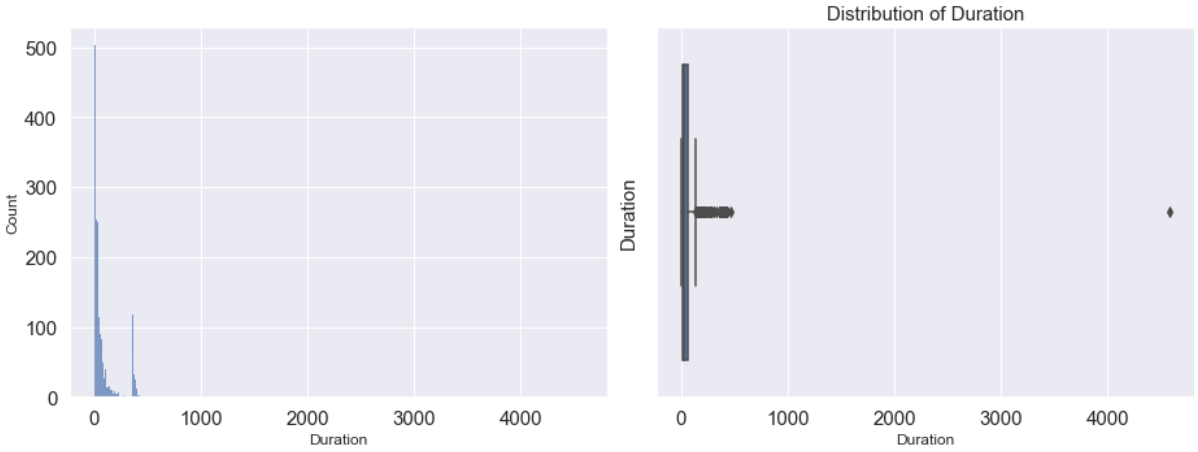

Figure 26: Histogram & Boxplot of Duration Variable

Observation

1. There are 13% outliers in the dataset

Sales variable

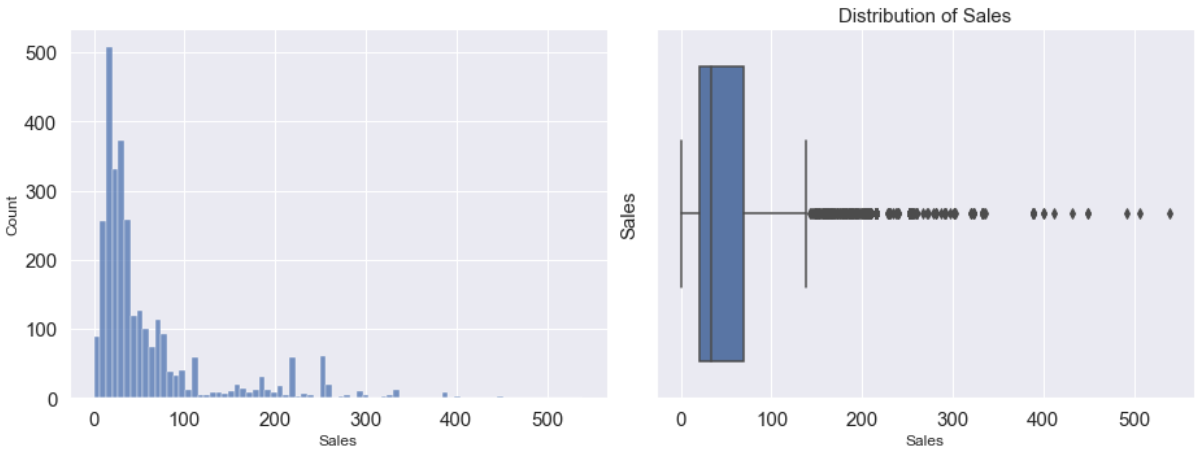

Figure 27: Histogram & Boxplot of Sales Variable

Observation

1. There are 12% outliers in the dataset

There are outliers in all the variables, but the sales and commission can be a genuine business value. Random Forest and CART can handle the outliers. Hence, Outliers are not treated for now, we will keep the data as it is.

Now let's analyse Categorical Variables:

Agency_Code:

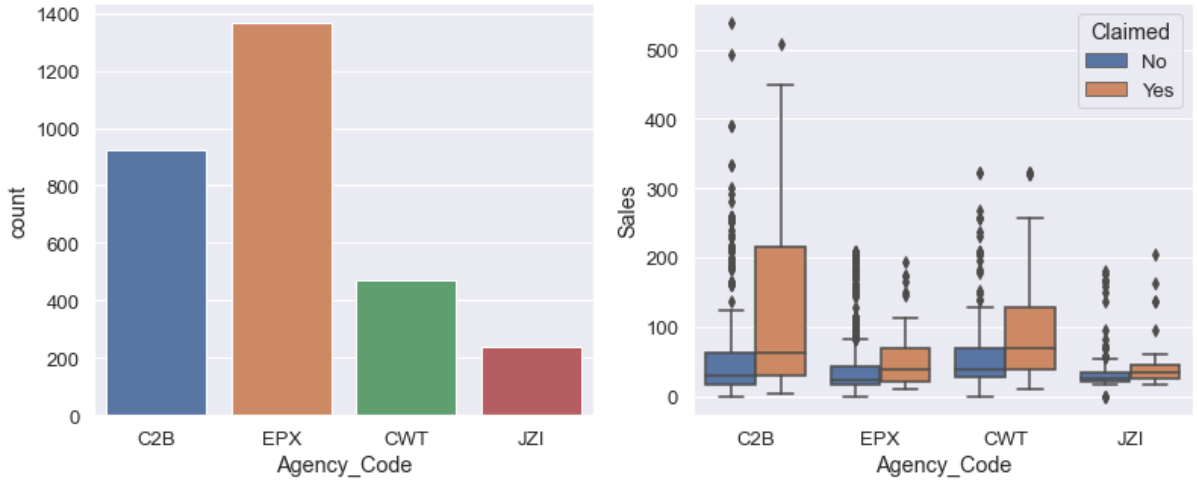

Figure 28: Countplot & Boxplot of Agency Code Variable

Type

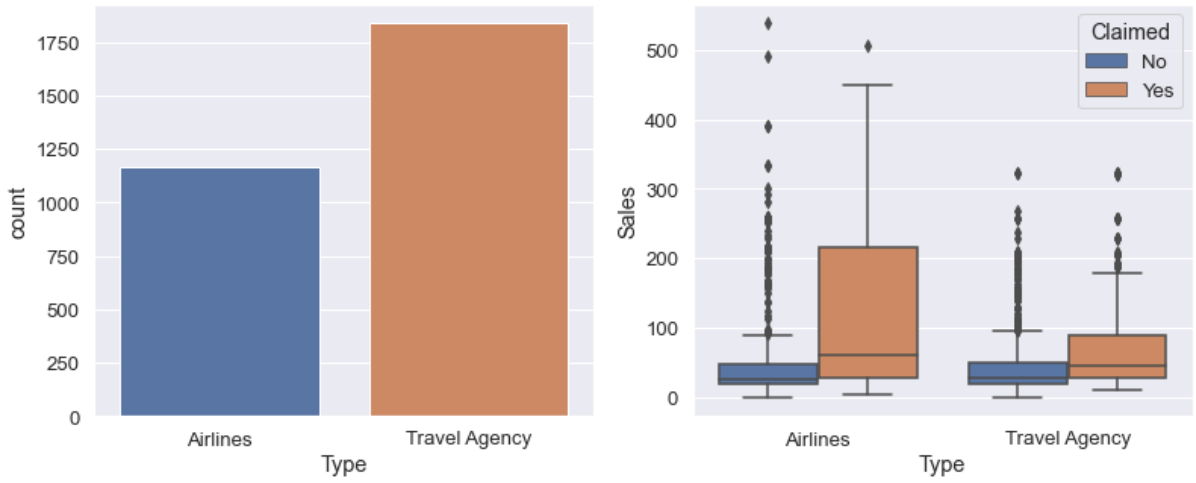

Figure 29: Countplot & Boxplot of Type Variable

Channel

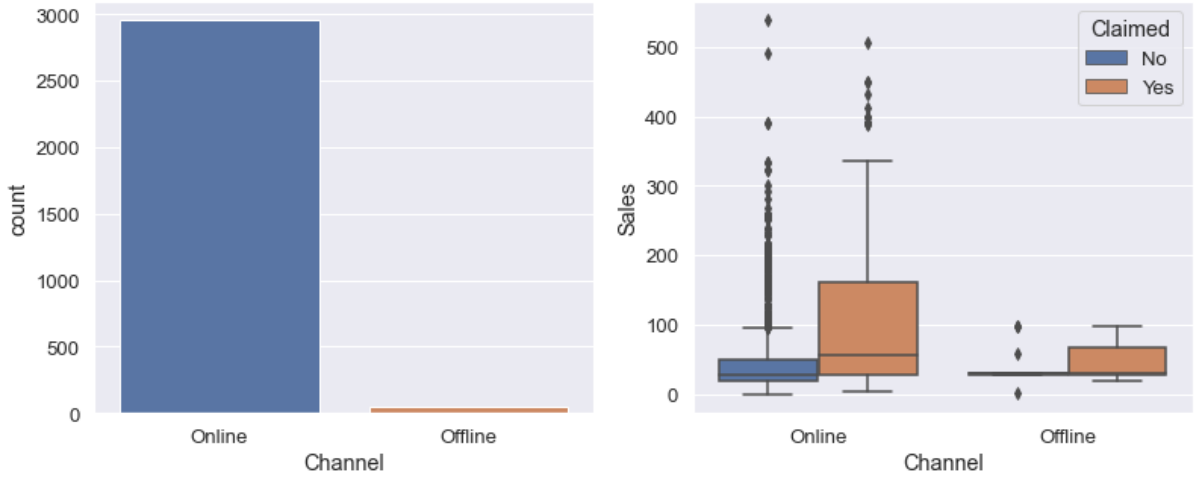

Figure 30: Countplot & Boxplot of Channel Variable

Product Name

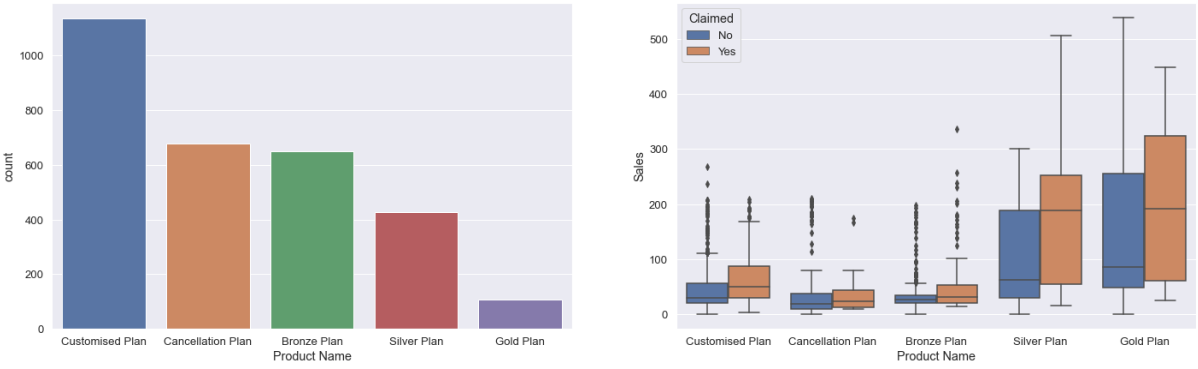

Figure 31: Countplot & Boxplot of Product Name Variable

Destination

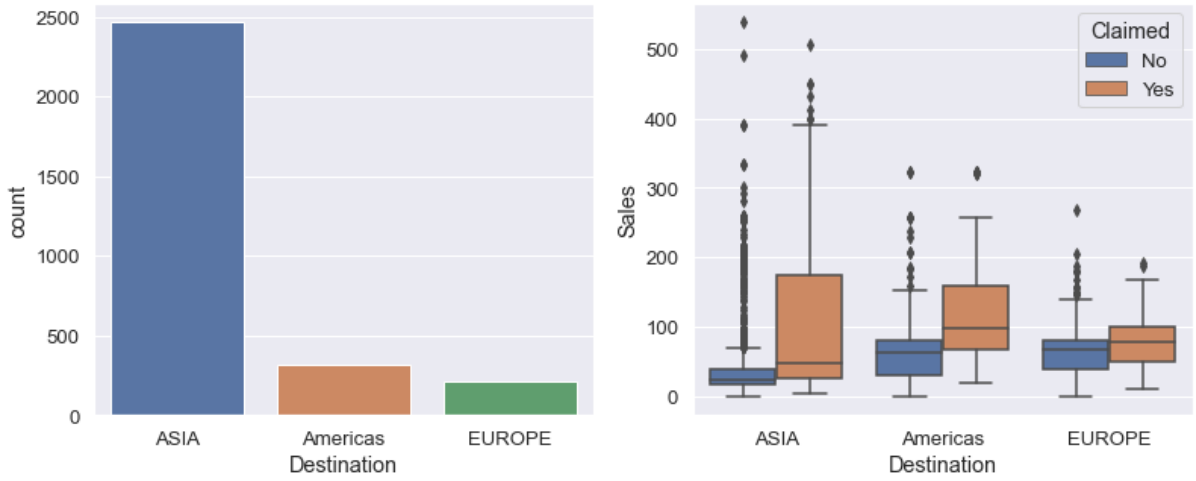

Figure 32: Countplot & Boxplot of Product Name Variable

MULTIVARIATE ANALYSIS

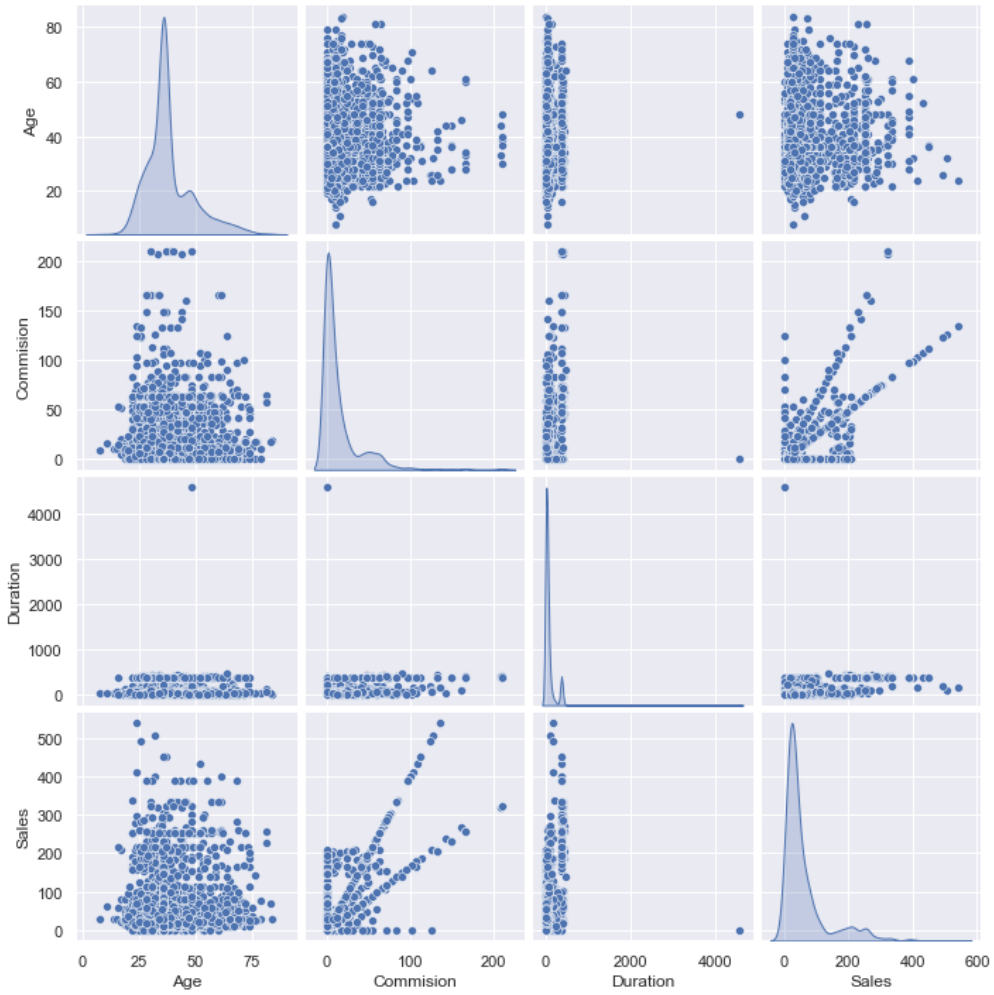

Figure 33: Pairplot of Insurance Data

We may infer utilising scatterplots for all the variables after performing multivariate analysis on the dataset's variables.

Many of the factors we can see in this have substantial correlations; for further information, let's look at the heatmap and the correlation table.

	Age	Commision	Duration	Sales
Age	1.000000	0.067717	0.030425	0.039455
Commision	0.067717	1.000000	0.471389	0.766505
Duration	0.030425	0.471389	1.000000	0.558930
Sales	0.039455	0.766505	0.558930	1.000000

Table 13: Correlation Matrix table of Insurance Data

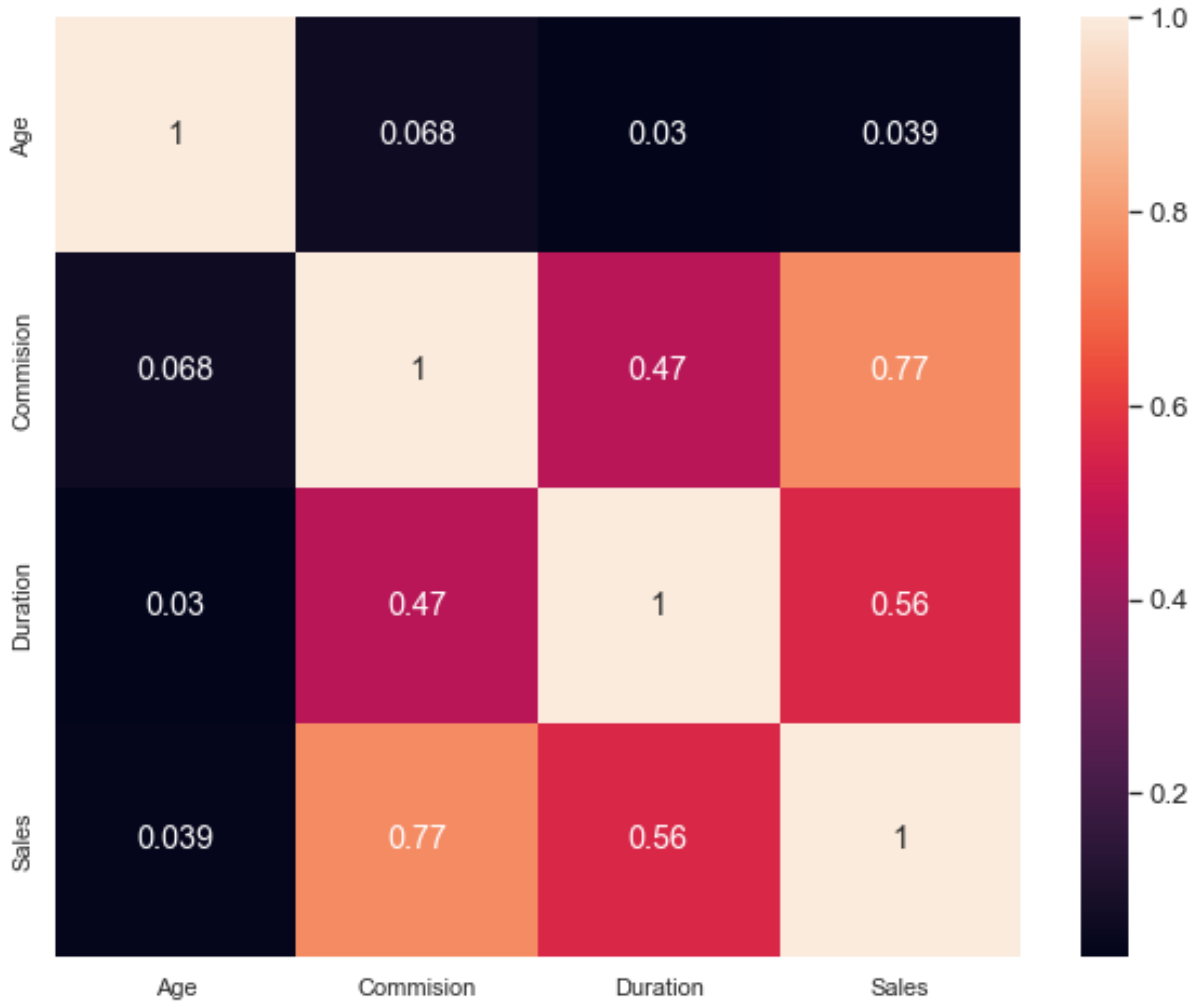

Figure 34: HeatMap of Insurance Data

Observation

- Highest Positive correlation can be found only between sales and commission

BI-VARIATE ANALYSIS:

Let's check the Strong Positive Correlations between Sales and Commission

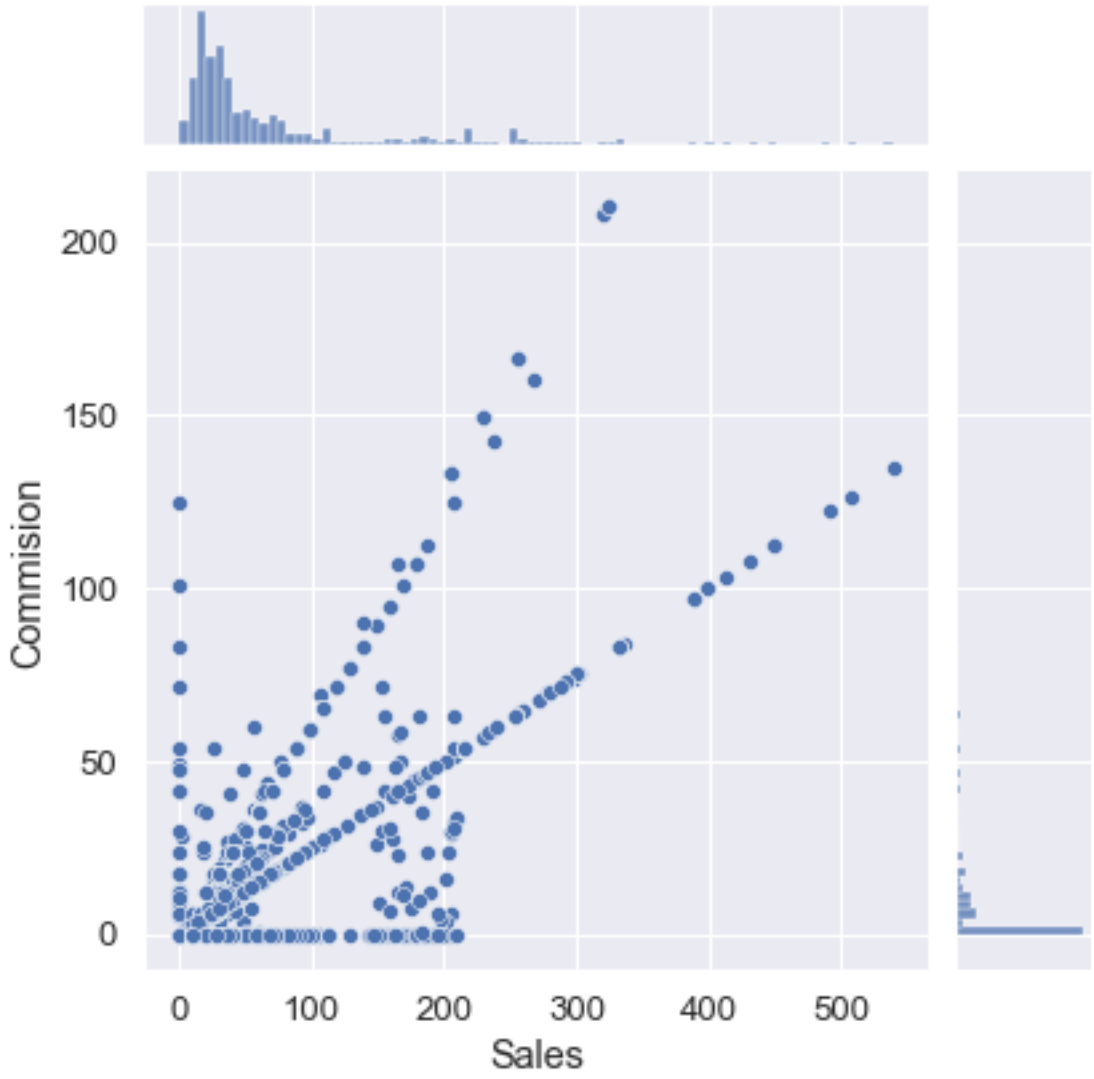

Figure 35: Bi-Variate Relationship b/w Sales and Commission Data

There is no major correlation in any of the two variables but in comparison, Sales and Commission has a correlation of 0.76 which is high in comparison with other variables.

On the next step, we are changing the data type of Object variables into Categorical data. After which, the all the data types of the data are either Integer or Float.

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

Table 14: Conversion Object variables into Categorical data of Insurance Data

Now let's see the Proportions of 0s and 1st of our target variable:

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

There is no issue of class imbalance here as we have reasonable proportions in both the classes.

2.2 Data Split: Split the data into test and train, built classification model CART, Random Forest

Let's see how the data look before and after scaling in from of a plot graph.

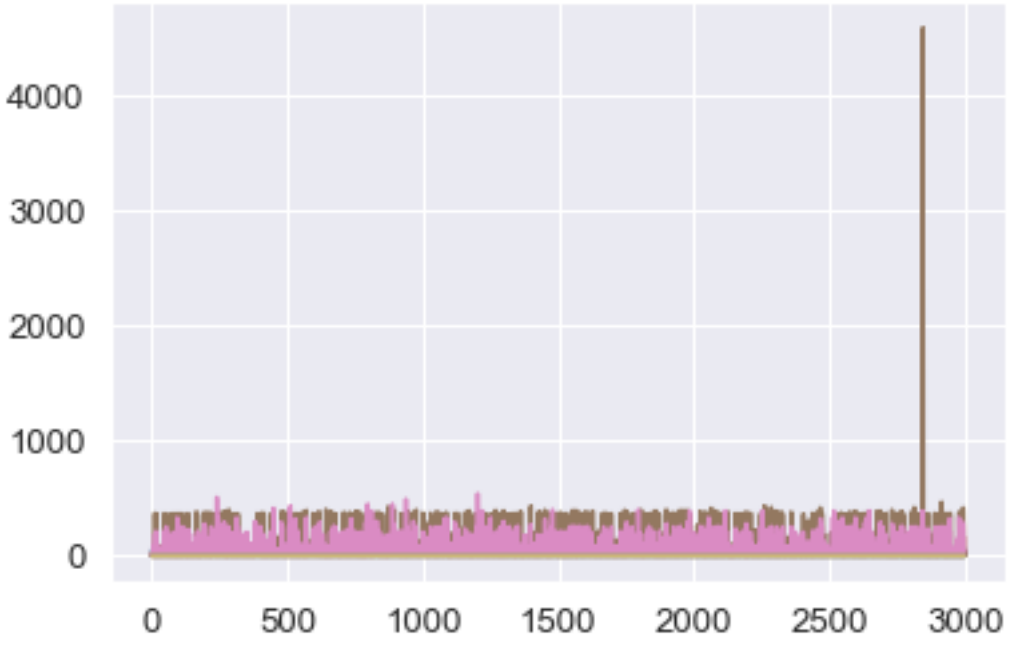

Figure 36: Graphical presentation of Insurance Data before Scaling

Figure 37: Graphical presentation of Insurance Data after Scaling

Firstly, splitting the data into Train and Test data.

Below is the data shape:

```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
```

CART MODEL

CART is a Binary Decision Tree model. I have used Gini Index as its Criteria. It is an attribute that Maximizes the reduction in impurity is chosen as the Splitting Attribute.

Using the Decision Tree Classifier and the Grid search method, I have identified the best grid:

- **'criterion'**: 'gini',
- **'max_depth'**: 4.0,
- **'min_samples_leaf'**: 46,
- '**min_samples_split**': 280

After looking at the decision tree, we extracted the variable importance shown below:

	Imp
Agency_Code	0.621974
Sales	0.257721
Product Name	0.057386
Commision	0.023406
Duration	0.023111
Age	0.016403
Type	0.000000
Channel	0.000000
Destination	0.000000

As per the above extract, Agency_code is the most important variable in the dataset, followed by Sales and Product Name.

Commision has comparatively very less importance, however Age, Type, Channel, Duration and Destination have no importance in the model building.

RANDOM FOREST

Random Forest Consists of many individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction. Class with most votes becomes model's prediction.

Using the random forest classifier and grid search function, we identified the best grid parameters:

- **'max_depth'**: 6,
- **'max_features':** 3,
- **'min_samples_leaf':** 8,
- **'min_samples_split':** 46,
- **'n_estimators':** 350

We extracted the variable importance as per RF:

	Imp
Agency_Code	0.263989
Product Name	0.219047
Sales	0.161785
Commision	0.143733
Type	0.076743
Duration	0.075790
Age	0.048086
Destination	0.009889
Channel	0.000937

Like CART, for RF as well Agency_code has the most importance in the model, however Sales and Product Name exchanged places. In this model, each of the variable pays a role in model building at some importance level but Channel Variable has the lowest importance of them all.

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

CART Performance Matrix

After predicting the test and train data, below is the head of ytest_predict_dtcl:

	0	1
0	0.887805	0.112195
1	0.432432	0.567568
2	0.432432	0.567568
3	0.208163	0.791837
4	0.937143	0.062857

Training data

AUC and ROC curve of CART:

AUC: 0.825

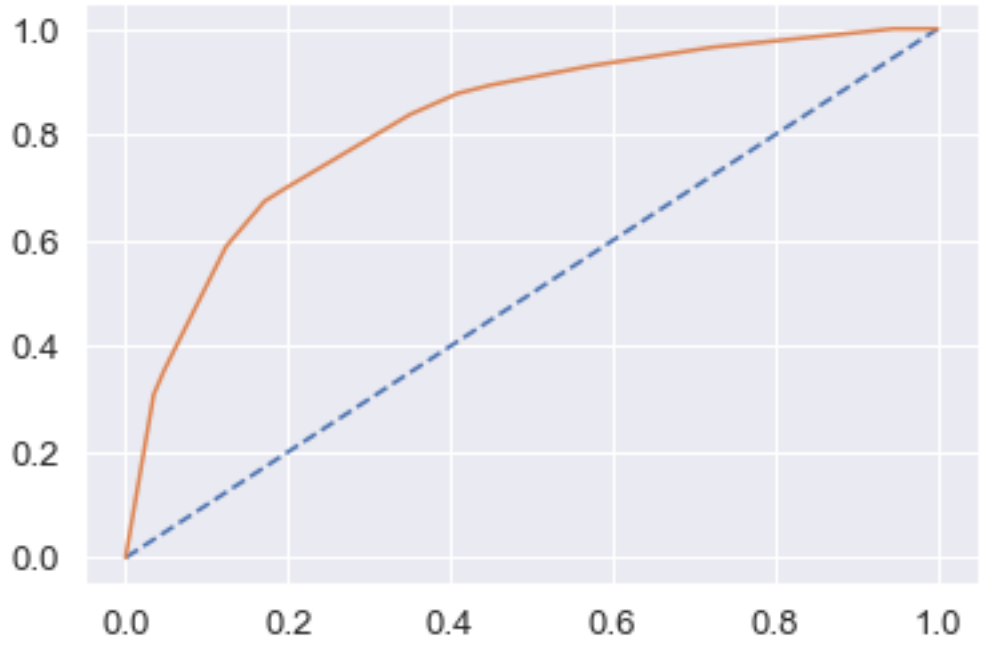

Figure 38: AUC & ROC Curve of CART (Training data)

Confusion Matrix on CART

```
array([[1289,  182],
       [ 259,  370]], dtype=int64)
```

Train Data Accuracy = **0.79**

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.88	0.85	1471
1	0.67	0.59	0.63	629
accuracy			0.79	2100
macro avg	0.75	0.73	0.74	2100
weighted avg	0.78	0.79	0.79	2100

Testing data

AUC and ROC curve of CART:

AUC: 0.792

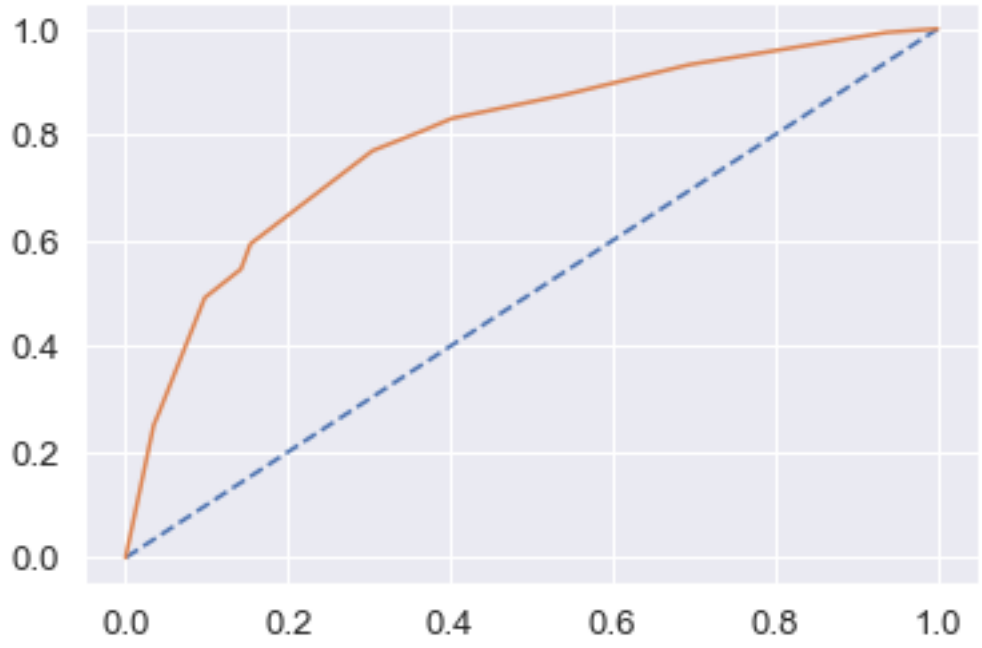

Figure 39: AUC & ROC Curve of CART (Testing data)

Confusion Matrix on CART:

```
array([[546,  59],
       [150, 145]], dtype=int64)
```

Test Data Accuracy = **0.7677777777777778**

Classification Report

	precision	recall	f1-score	support
0	0.78	0.90	0.84	605
1	0.71	0.49	0.58	295
accuracy			0.77	900
macro avg	0.75	0.70	0.71	900
weighted avg	0.76	0.77	0.75	900

Cart Conclusion

Cart Conclusion Train Data:

- AUC: 82.5%
- Accuracy: 79%
- Precision: 67%
- f1-Score: 63%

Test Data:

- AUC: 89.2%
- Accuracy: 76.7%
- Precision: 71%
- f1-Score: 58%

RANDOM FOREST Performance Matrix:

Training data

AUC and ROC curve of RF:

Area under Curve (AUC) is **0.853038986921499**

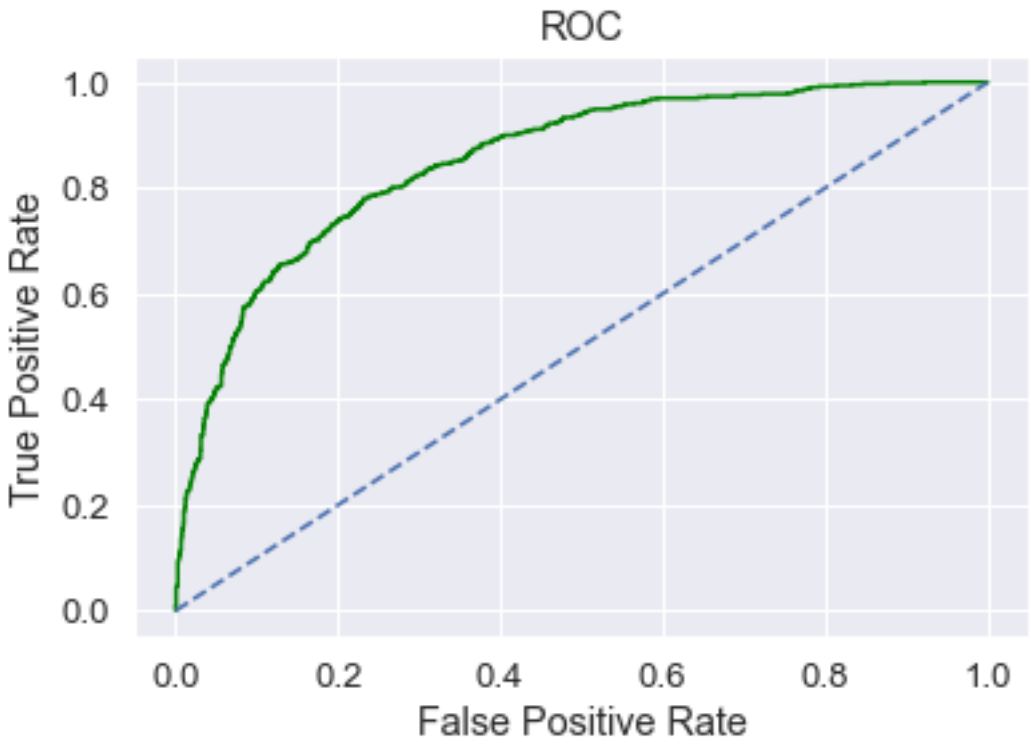

Figure 40: AUC & ROC Curve of RF (Training data)

Confusion Matrix on RF

```
array([[1331,  140],
       [ 258,  371]], dtype=int64)
```

Data Accuracy = **0.8104761904761905**

Classification Report

	precision	recall	f1-score	support
0	0.84	0.90	0.87	1471
1	0.73	0.59	0.65	629
accuracy			0.81	2100
macro avg	0.78	0.75	0.76	2100
weighted avg	0.80	0.81	0.80	2100

Testing data

AUC and ROC curve of RF:

Area under Curve (AUC) is **0.8199355652052107**

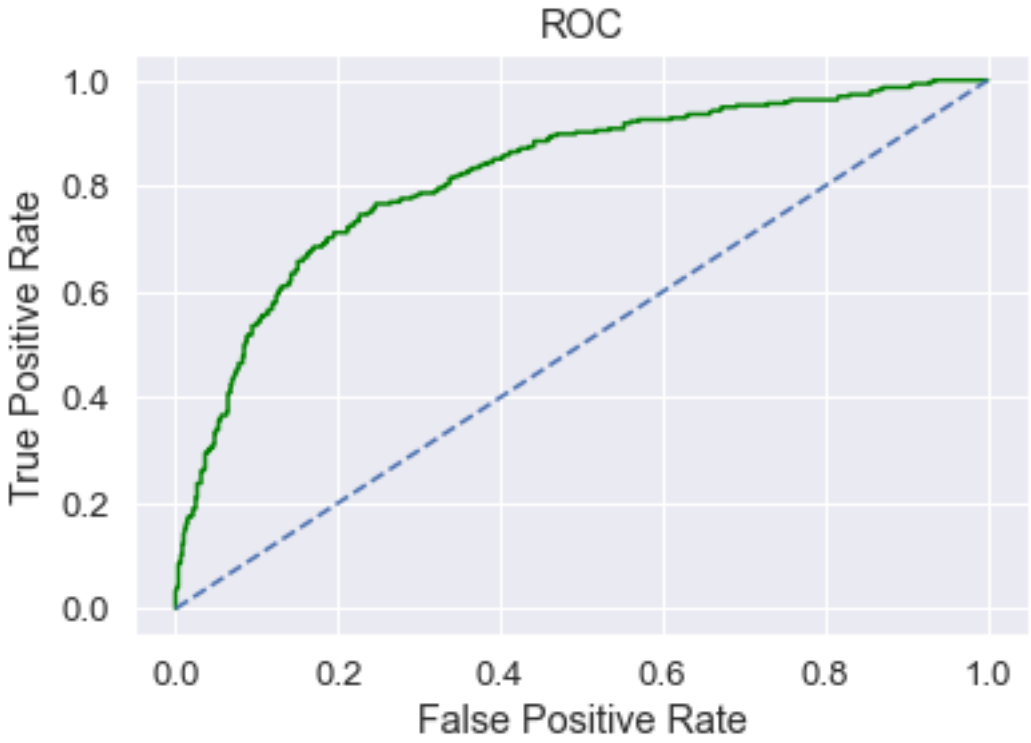

Figure 41: AUC & ROC Curve of RF (Testing data)

Confusion Matrix on RF:

```
array([[555,  50],
       [154, 141]], dtype=int64)
```

Data Accuracy 0.7733333333333333

Classification Report on Test Data:

	precision	recall	f1-score	support
0	0.78	0.92	0.84	605
1	0.74	0.48	0.58	295
accuracy			0.77	900
macro avg	0.76	0.70	0.71	900
weighted avg	0.77	0.77	0.76	900

Random Forest Conclusion

Train Data:

- AUC: 85%
- Accuracy: 81%
- Precision: 73%
- f1-Score: 65%

Test Data:

- AUC: 81%
- Accuracy: 77%
- Precision: 74%
- f1-Score: 58%

Observation

Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

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

Below we are comparing Accuracy, AUC, Recall, Precision and F1 score of all the models, where Target is 0, i.e. the claimed as NO.

The logic to choose Claimed as NO is that the model is calculating Claimed as No more accurately than Claimed as Yes. Also, this way we will be able to identify using the attributes that which policy will not be claimed with more than approx. 75% accuracy.

	CART Train	CART Test	Random Forest Train	Random Forest Test
Accuracy	0.79	0.77	0.81	0.77
AUC	0.83	0.79	0.85	0.82
Recall	0.59	0.49	0.59	0.48
Precision	0.67	0.71	0.73	0.74
F1 Score	0.63	0.58	0.65	0.58

Table 15: Comparison of the performance metrics of Insurance Data

ROC curve of the **Train data** of all the models:

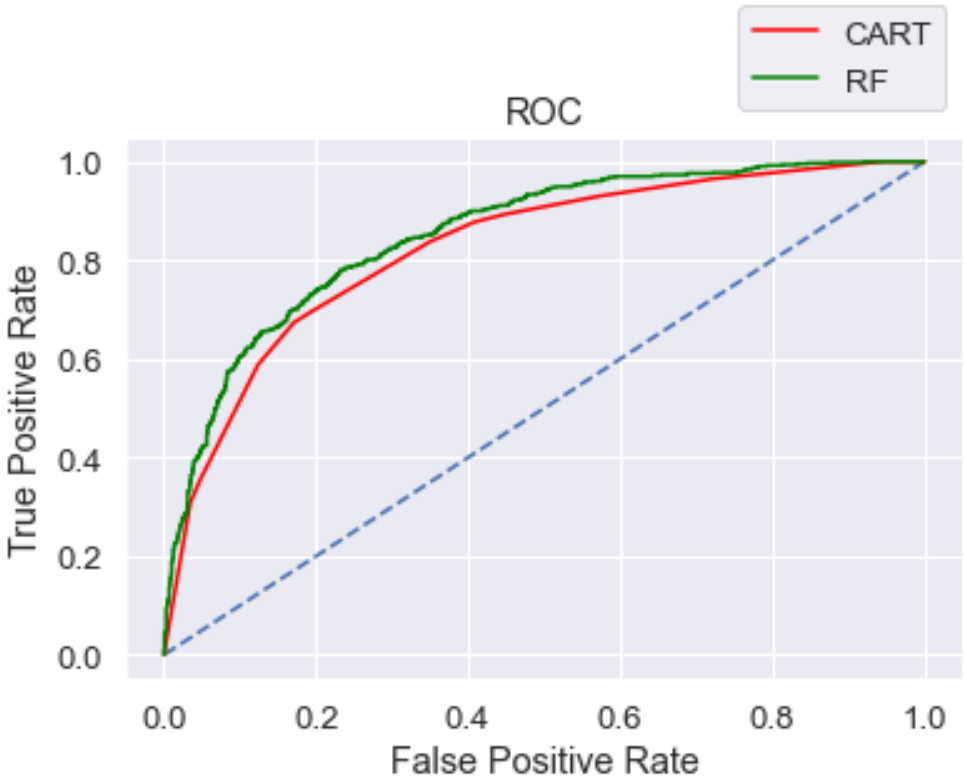

Figure 42: AUC & ROC Curve of all the models (Training data)

ROC curve of the **Test data** of all the models:

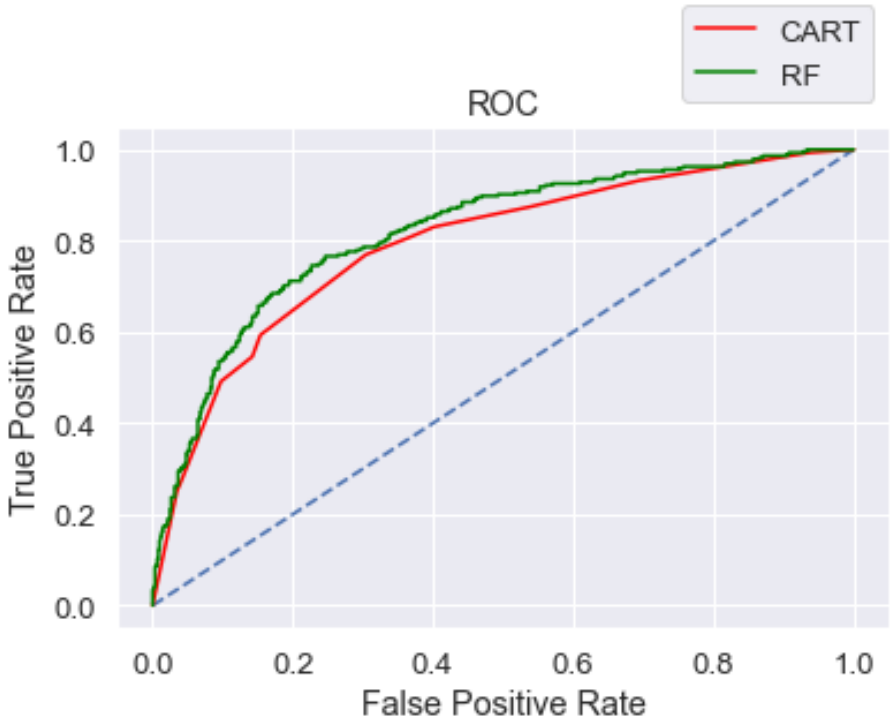

Figure 43: AUC & ROC Curve of all the models (Testing data)

Out of the 2 models, Random Forest has slightly better performance than the Cart. I am selecting the RF model, as it has better accuracy, precision, recall, f1 score.

Overall, all the 2 models are reasonably stable enough to be used for making any future predictions. From Random Forest Model, the variable change is found to be the most useful feature amongst all other features for predicting if a person will claim or not. If change is NO, then those policies have more chances of getting claimed.

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

Business insights:

• After applying the supervised learning algorithm, it appears that all of the models have excellent accuracy in identifying clients who will not make a claim for travel insurance.

• As far as we are aware, this insurance company had a higher-than-average claim frequency, and this methodology would undoubtedly aid in lowering the ratio.

• I advise the insurance company to partner with additional Agencies in order to grow its business, as the variable Agency code appears to be the most crucial component in determining the model.

• In order to reduce the frequency of claims, they should add some clauses to their policy's terms and conditions that would be advantageous to both the company and the clients.

• This insurance company can choose its profitable consumers with ease using this model and customer data.

• The team can easily focus on the clients who won't file claims for travel insurance. You only make money when returning consumers regard an organisation as being loyal and trustworthy, therefore as soon as the team receives customer data from individuals who, according to the model, have NO claim status, the team must forge strong relationships with such clients.

• In my opinion, the travel insurance provider ought to offer more options for Product Name. They now provide Bronze, Cancellation, Customized, Gold, and Silver plans, but adding a few more to the list will entice clients to select the best plan that works best for them and, as a result, reduce the frequency of claims for the business. Additionally, doing the same would increase sales for the tour insurance company.

• Customers that fall under the NO claim status can be advised to purchase a product plan with a higher commission rate.

• Customers benefited from the simplification of online interactions, which improved conversions and profitability.

• According to research, 90% of insurance transactions take place online. • Another fascinating aspect is that virtually every offline business claims to be related; researchers need to determine why.

• We must launch a promotional marketing campaign, train the JZI agency resources to pick up sales as they fall off, and assess whether we need to partner with another agency.

• Additionally, the model gives us an accuracy of 80%, therefore we need customers to buy plans or airline tickets and cross-sell insurance based on the claim data pattern.

• We must launch a promotional marketing campaign, train the JZI agency resources to pick up sales as they fall off, and assess whether we need to partner with another agency.

• Additionally, the model gives us an accuracy of 80%, therefore we need customers to buy plane tickets or plans and cross-sell insurance based on claim data patterns.

• Another noteworthy finding is that agencies generate more sales than airlines do, and that trend indicates that airlines handle more claims. To comprehend the workflow and why it is the way it is, we might need to dig deep into the process.

• Key performance metric (KPI) The insurance claim KPIs are:

1) Shorten the claims cycle

2) Boost client satisfaction

3) Prevent fraud

4) Increase claim reimbursement

- 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.

Recommendations

I strongly recommended we collect more real time unstructured data and past data if possible.

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, behaviour patterns, weather information, airline/vehicle types, etc.