24 October 2021

Data Mining Project

PGP-DSBA June-Batch

SOOREJ KUNNOOL BALAN

2021

Table of Contents

[1. Problem 1: Clustering 4](#_Toc86002889)

[1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 4](#_Toc86002890)

[Data Description 4](#_Toc86002891)

[Sample of the dataset: 4](#_Toc86002892)

[Exploratory Data Analysis: 5](#_Toc86002893)

[1.2 Do you think scaling is necessary for clustering in this case? Justify. 14](#_Toc86002894)

[1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. 16](#_Toc86002895)

[1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters. 18](#_Toc86002896)

[1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. 20](#_Toc86002897)

[2. Problem 2: CART-RF-ANN 22](#_Toc86002898)

[2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 22](#_Toc86002899)

[Data Description 22](#_Toc86002900)

[Sample of the dataset: 22](#_Toc86002901)

[Exploratory Data Analysis: 22](#_Toc86002902)

[2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network 35](#_Toc86002903)

[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. 38](#_Toc86002904)

[2.4 Final Model: Compare all the models and write an inference which model is best/optimized. 42](#_Toc86002905)

[2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations 42](#_Toc86002906)

[THE END 43](#_Toc86002907)

List of Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table-1.1 Dataset Sample | 4 |  |  | |  |
| Table-1.2 Concise data summary | 4 |  |  | |  |
| Table-1.3 Data Summary | 5 |  |  | |  |
| Table-1.4 Data Summary | 13 |  |  | |  |
| Table-1.5 Correlation matrix | 13 |  |  | |  |
| Table-1.6 Data summary of scaled data | 15 |  |  | |  |
| Table-1.7 Hierarchical cluster profile for 3 clusters | 17 |  |  | |  |
| Table-1.8 Hierarchical cluster profile for 4 clusters | 17 |  |  | |  |
| Table-1.9 Hierarchical cluster profile for 3 clusters | 18 |  |  | |  |
| Table-1.10 Minimum silhouette coefficient across clusters | 19 |  |  | |  |
| Table-1.11 K-Means cluster profile for 3 clusters | 19 |  |  | |  |
| Table-1.12 Hierarchical cluster profile for 3 clusters | 20 |  |  | |  |
| Table-1.13 K-Means cluster profile for 3 clusters | 20 |  |  | |  |
| Table-2.1 Dataset Sample | 22 |  |  | |  |
| Table-2.2 Concise data summary | 22 |  |  | |  |
| Table-2.3 Duplicate information | 23 |  |  | |  |
| Table-2.4 Data Summary | 23 |  |  | |  |
| Table-2.5 Data Summary with quartile values and range | 24 |  |  | |  |
| Table-2.6 records with duration < 0 | 28 |  |  | |  |
| Table-2.7 Updated record | 28 |  |  | |  |
| Table-2.8 Data summary | 29 |  |  | |  |
| Table-2.9 Categorical code mapping | 35 |  |  | |  |
| Table-2.10 Concise data summary | 36 |  |  | |  |
| Table-2.11 Training/Test Target variable summary | 36 |  |  | |  |
| Table-2.12 CART model feature importance | 37 |  |  | |  |
| Table-2.13 Random Forest model feature importance | 38 |  |  |  | |
| Table-2.14 CART – Confusion matrix – Training data | 39 |  |  |  | |
| Table-2.15 CART – Confusion matrix –Testing data | 39 |  |  |  | |
| Table-2.16 CART –Classification Report– Training data | 39 |  |  |  | |
| Table-2.17 CART –Classification Report–Testing data | 39 |  |  |  | |
| Table-2.18 RF – Confusion matrix – Training data | 40 |  |  |  | |
| Table-2.19 RF – Confusion matrix –Testing data | 40 |  |  |  | |
| Table-2.20 RF – Classification Report– Training data | 40 |  |  |  | |
| Table-2.21 RF – Classification Report–Testing data | 40 |  |  |  | |
| Table-2.22 ANN – Confusion matrix – Training data | 41 |  |  |  | |
| Table-2.23 ANN – Confusion matrix –Testing data | 41 |  |  |  | |
| Table-2.24 ANN – Classification Report– Training data | 42 |  |  |  | |
| Table-2.25 ANN – Classification Report–Testing data | 42 |  |  |  | |
| Table-2.26 Performance metrics across models | 42 |  |  |  | |

List of Figures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Figure-1.1 Histogram & Boxplot : spending | 6 |  |  | |  |
| Figure-1.2 Histogram & Boxplot : advance\_payments | 7 |  |  | |  |
| Figure-1.3 Histogram & Boxplot : probability\_of\_full\_payment | 8 |  |  | |  |
| Figure-1.4 Histogram & Boxplot : current\_balance | 9 |  |  | |  |
| Figure-1.5 Histogram & Boxplot : credit\_limit | 10 |  |  | |  |
| Figure-1.6 Histogram & Boxplot : min\_payment\_amt | 11 |  |  | |  |
| Figure-1.7 Histogram & Boxplot : max\_spent\_in\_single\_shopping | 12 |  |  | |  |
| Figure-1.8 Boxplot : probability\_of\_full\_payment & min\_payment\_amt | 12 |  |  | |  |
| Figure-1.9 Pairplot | 14 |  |  | |  |
| Figure-1.10 Original dataset plot | 15 |  |  | |  |
| Figure-1.11 Scaled dataset plot | 15 |  |  | |  |
| Figure-1.12 Dendrogram | 16 |  |  | |  |
| Figure-1.13 Truncated Dendrogram | 16 |  |  | |  |
| Figure-1.14 Inertia across clusters | 18 |  |  | |  |
| Figure-1.15 Silhouette scores across clusters | 19 |  |  | |  |
| Figure-2.1 Histogram & Boxplot : Age | 24 |  |  | |  |
| Figure-2.2 Count-plot & Pie-chart: Agency\_code | 25 |  |  | |  |
| Figure-2.3 Count-plot & Pie-chart: Type | 26 |  |  | |  |
| Figure-2.4 Count-plot & Pie-chart: Claimed | 26 |  |  | |  |
| Figure-2.5 Histogram & Boxplot : Commission | 27 |  |  | |  |
| Figure-2.6 Count-plot & Pie-chart: Channel | 28 |  |  | |  |
| Figure-2.7 Histogram & Boxplot : Duration | 29 |  |  | |  |
| Figure-2.8 Histogram & Boxplot : Sales | 30 |  |  | |  |
| Figure-2.9 Count-plot & Pie-chart: Product Name | 31 |  |  | |  |
| Figure-2.10 Count-plot & Pie-chart: Destination | 32 |  |  | |  |
| Figure-2.11 Pairplot | 32 |  |  |  | |
| Figure 2.12 Correlation matrix | 33 |  |  |  | |
| Figure 2.13 Stripplot: Agency\_Code vs Sales | 33 |  |  |  | |
| Figure 2.14 Stripplot: Product Name vs Sales | 34 |  |  |  | |
| Figure 2.15 Stripplot: Destination vs Sales | 34 |  |  |  | |
| Figure 2.16 Stripplot: Type vs Sales | 35 |  |  |  | |
| Figure-2.17 CART –ROC score/curve– Training data | 39 |  |  |  | |
| Figure-2.18 CART –ROC score/curve–Testing data | 39 |  |  |  | |
| Figure-2.19 RF – ROC score/curve– Training data | 40 |  |  |  | |
| Figure-2.20 RF – ROC score/curve–Testing data | 40 |  |  |  | |
| Figure-2.21 ANN – ROC score/curve– Training data | 41 |  |  |  | |
| Figure-2.22 ANN – ROC score/curve–Testing data | 41 |  |  |  | |

# 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



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

### Data Description

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)

### Sample of the dataset:

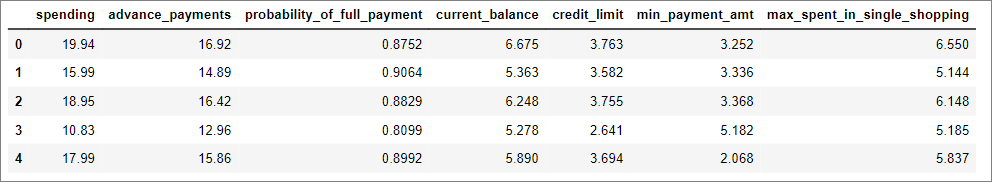


Table-1.1 Dataset Sample

### Exploratory Data Analysis:

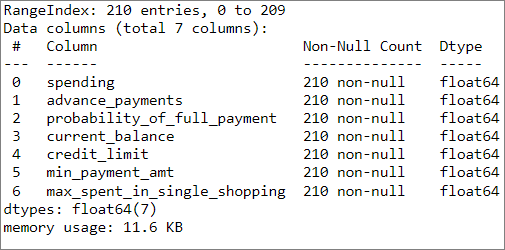


Table-1.2 Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 210 observations and 7 columns in the dataset. All the columns are of float type.

#### Check for missing values in the dataset

From Table-1.2 we can see that all the columns have 210 non-null values and hence we have no missing values in the dataset.

#### Check for duplicate observations in the dataset

There are no duplicate rows in the dataset.

#### Data summary

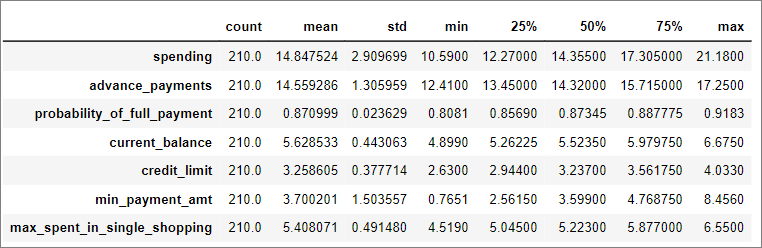


Table-1.3 Data Summary

The above data summary will be further explained in the univariate analysis section below.

#### Univariate Analysis:

Let’s check the central measures of tendency, quartiles, histogram, and boxplot of all 7 columns.

1. spending

Amount spent by customer per month is a continuous variable with the below stats (refer Table 1.3):

Mean = 14.847524

Standard Deviation = 2.909699

Min value in dataset = 10.59

Max value in dataset = 21.18

Range = Min – Max = 10.59

Q1(1st Quartile) = 12.27

Q2(2nd Quartile)/Median = 14.355

Q3(3rd Quartile) = 17.305

IQR(Inter-Quartile Range) = Q3- Q1 = 5.035

Quartile Min value = Q1 – 1.5 \* IQR = 4.7175

Quartile Max value = Q3 + 1.5 \* IQR = 24.8575 which is greater than max value, hence 21.18

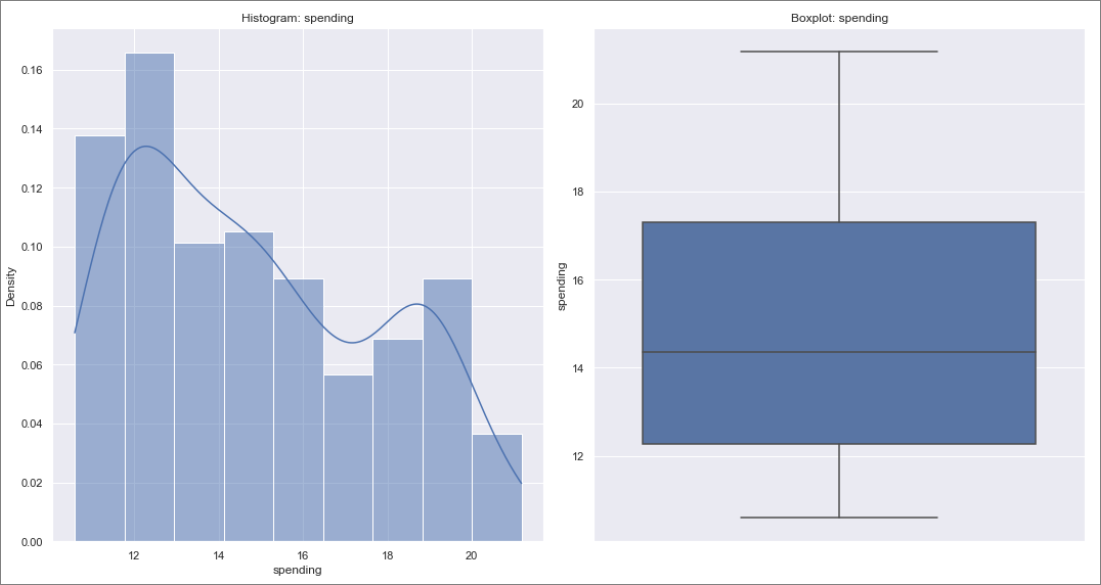


Figure-1.1 Histogram & Boxplot : spending

Figure-1.1 depicts the histogram and boxplot of “spending” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 2.948e-8 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “spending” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are no outliers present in ‘spending’.

1. advance\_payments

Amount paid by customer in advance is a continuous variable with the below stats (refer Table 1.3):

Mean = 14.559286

Standard Deviation = 1.305959

Min value in dataset = 12.4100

Max value in dataset = 17.2500

Range = Min – Max = 4.84

Q1(1st Quartile) = 13.45000

Q2(2nd Quartile)/Median = 14.32000

Q3(3rd Quartile) = 15.715000

IQR(Inter-Quartile Range) = Q3- Q1 = 2.265

Quartile Min value = Q1 – 1.5 \* IQR = 10.0525 which is lesser than min value, hence 12.41

Quartile Max value = Q3 + 1.5 \* IQR = 19.1125 which is greater than max value, hence 17.25

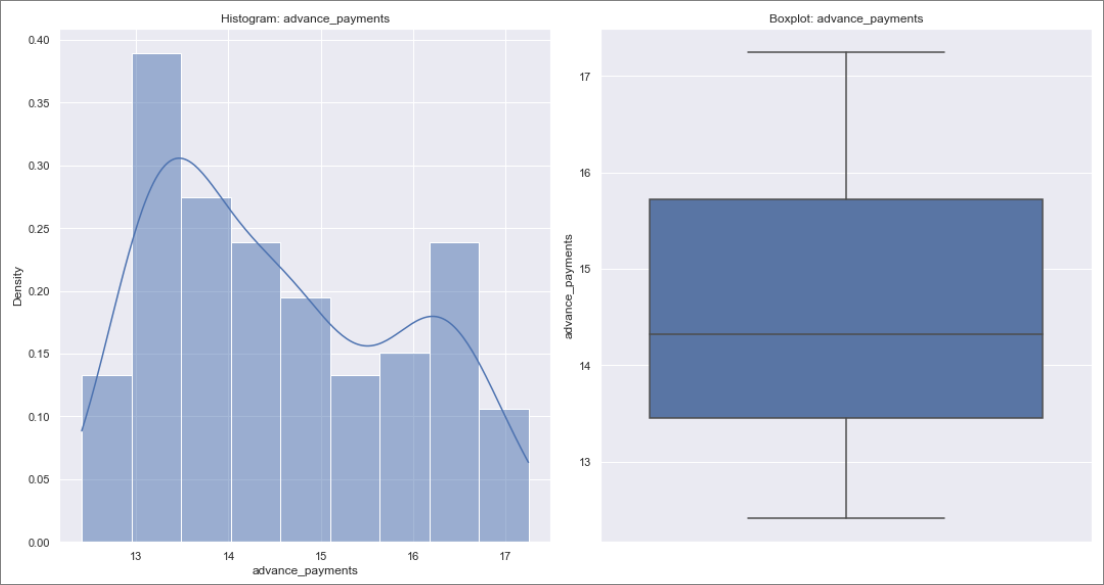


Figure-1.2 Histogram & Boxplot : advance\_payments

Figure-1.2 depicts the histogram and boxplot of “advance\_payments” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 5.902e-8 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “advance\_payments” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are no outliers present in ‘advance\_payments’.

1. probability\_of\_full\_payment

Probability of full payment to bank by customer is a continuous variable with the below stats (refer Table 1.3):

Mean = 0.870999

Standard Deviation = 0.023629

Min value in dataset = 0.8081

Max value in dataset = 0.9183

Range = Min – Max = 0.1102

Q1(1st Quartile) = 0.85690

Q2(2nd Quartile)/Median = 0.87345

Q3(3rd Quartile) = 0.887775

IQR(Inter-Quartile Range) = Q3- Q1 = 0.030875

Quartile Min value = Q1 – 1.5 \* IQR = 0.8106

Quartile Max value = Q3 + 1.5 \* IQR = 0.9341 which is greater than max value, hence 0.9183

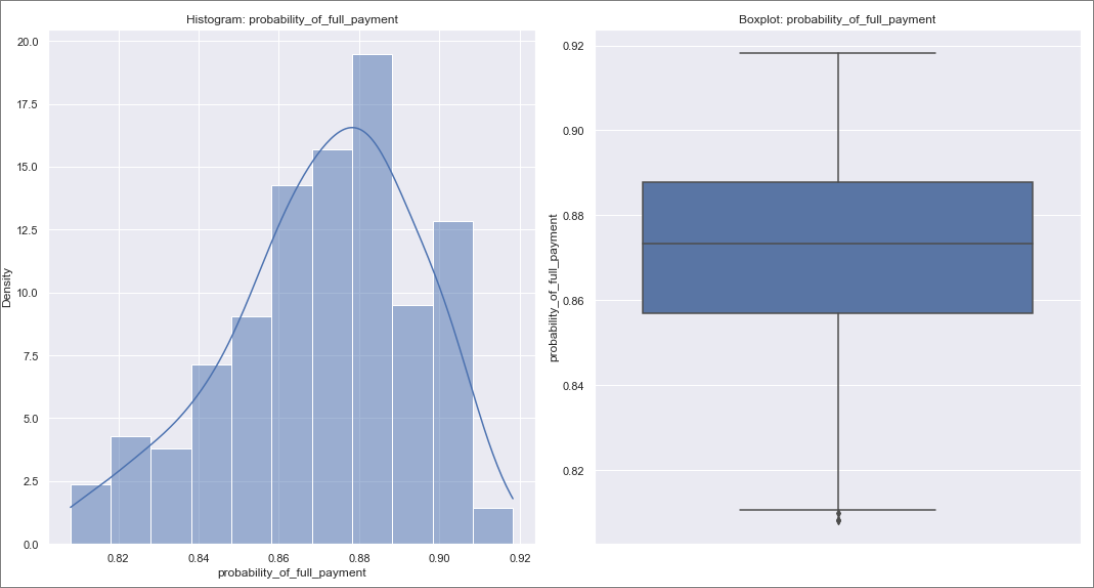


Figure-1.3 Histogram & Boxplot : probability\_of\_full\_payment

Figure-1.3 depicts the histogram and boxplot of “probability\_of\_full\_payment” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 0.00047 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “probability\_of\_full\_payment” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘probability\_of\_full\_payment’.

1. current\_balance

Balance amount left with customer is a continuous variable with the below stats (refer Table 1.3):

Mean = 5.628533

Standard Deviation = 0.443063

Min value in dataset = 4.8990

Max value in dataset = 6.6750

Range = Min – Max = 4.84

Q1(1st Quartile) = 5.26225

Q2(2nd Quartile)/Median = 5.52350

Q3(3rd Quartile) = 5.979750

IQR(Inter-Quartile Range) = Q3- Q1 = 0.7175

Quartile Min value = Q1 – 1.5 \* IQR = 4.186 which is lesser than min value, hence 4.8990

Quartile Max value = Q3 + 1.5 \* IQR = 7.056 which is greater than max value, hence 6.6750

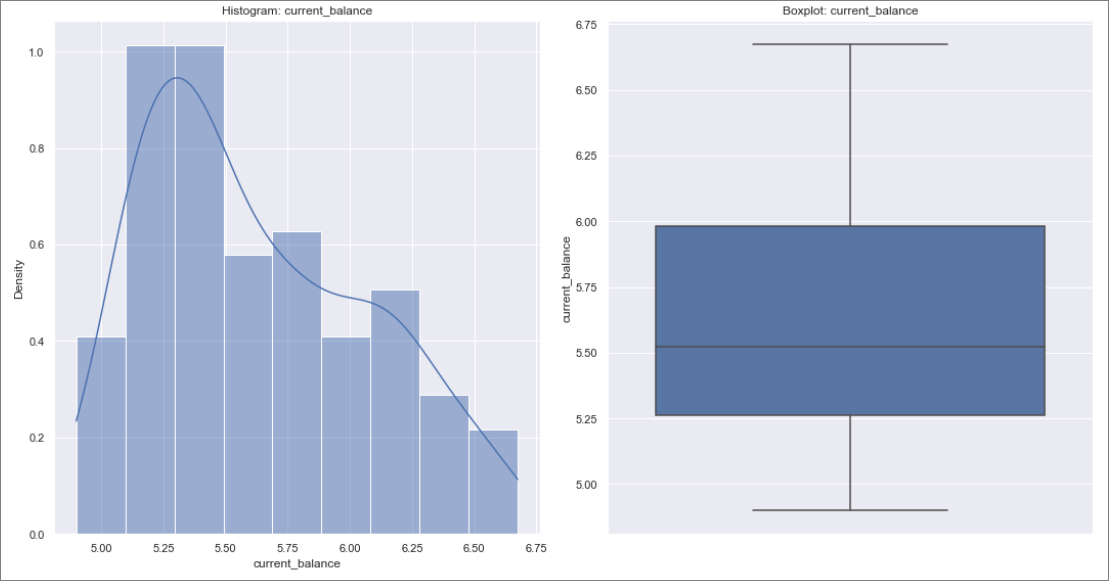


Figure-1.4 Histogram & Boxplot : current\_balance

Figure-1.4 depicts the histogram and boxplot of “current\_balance” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 2.828e-7 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “current\_balance” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are no outliers present in ‘current\_balance’.

1. credit\_limit

Credit limit of customer is a continuous variable with the below stats (refer Table 1.3):

Mean = 3.258605

Standard Deviation = 0.377714

Min value in dataset = 2.6300

Max value in dataset = 4.0330

Range = Min – Max = 1.403

Q1(1st Quartile) = 2.94400

Q2(2nd Quartile)/Median = 3.23700

Q3(3rd Quartile) = 3.561750

IQR(Inter-Quartile Range) = Q3- Q1 = 0.61775

Quartile Min value = Q1 – 1.5 \* IQR = 2.017 which is lesser than min value, hence 2.63

Quartile Max value = Q3 + 1.5 \* IQR = 4.488 which is greater than max value, hence 4.0330

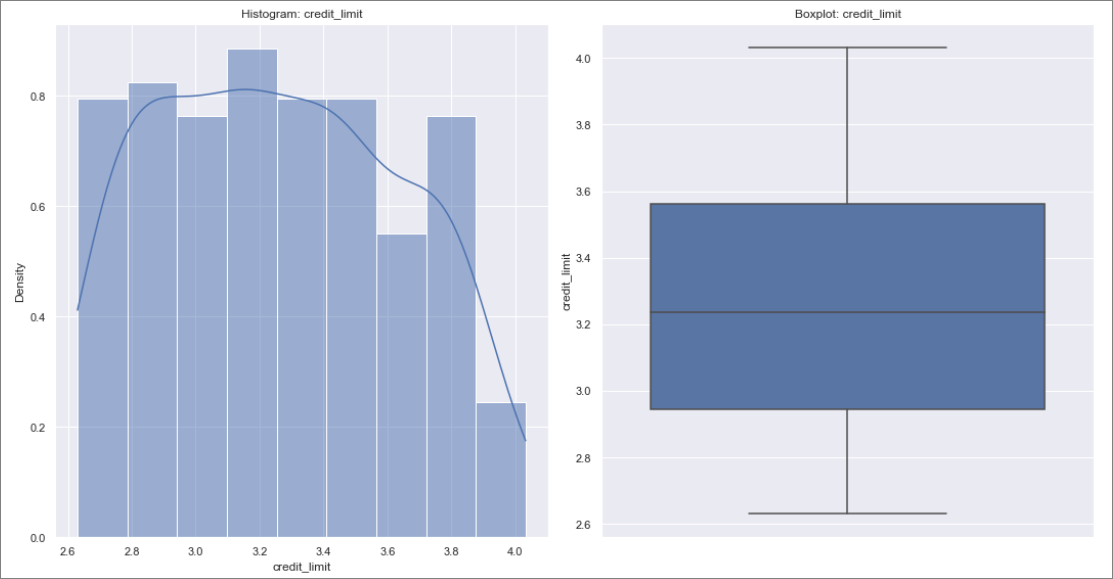


Figure-1.5 Histogram & Boxplot : credit\_limit

Figure-1.5 depicts the histogram and boxplot of “credit\_limit” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.444e-5 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “credit\_limit” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are no outliers present in ‘credit\_limit’.

1. min\_payment\_amt

Minimum monthly payment done by customer is a continuous variable with the below stats (refer Table 1.3):

Mean = 3.700201

Standard Deviation = 1.503557

Min value in dataset = 0.7651

Max value in dataset = 8.4560

Range = Min – Max = 7.6909

Q1(1st Quartile) = 2.56150

Q2(2nd Quartile)/Median = 3.59900

Q3(3rd Quartile) = 4.768750

IQR(Inter-Quartile Range) = Q3- Q1 = 2.20725

Quartile Min value = Q1 – 1.5 \* IQR = -0.7494 which is lesser than min value, hence 0.7651

Quartile Max value = Q3 + 1.5 \* IQR = 8.0796

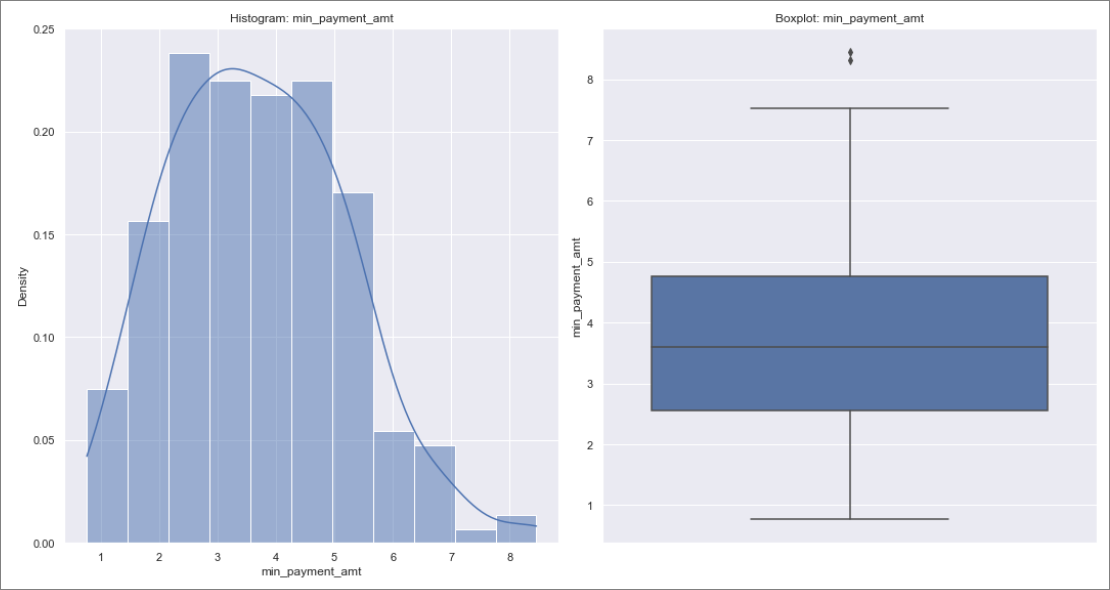


Figure-1.6 Histogram & Boxplot : min\_payment\_amt

Figure-1.6 depicts the histogram and boxplot of “min\_payment\_amt” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 0.0154 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “min\_payment\_amt” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘min\_payment\_amt’.

1. max\_spent\_in\_single\_shopping

Maximum amount spent by customer in one purchase is a continuous variable with the below stats (refer Table 1.3):

Mean = 5.408071

Standard Deviation = 0.491480

Min value in dataset = 4.5190

Max value in dataset = 6.5500

Range = Min – Max = 2.031

Q1(1st Quartile) = 5.04500

Q2(2nd Quartile)/Median = 5.22300

Q3(3rd Quartile) = 5.877000

IQR(Inter-Quartile Range) = Q3- Q1 = 0.832

Quartile Min value = Q1 – 1.5 \* IQR = 3.797 which is lesser than min value, hence 4.519

Quartile Max value = Q3 + 1.5 \* IQR = 7.125 which is greater than max value, hence 6.55

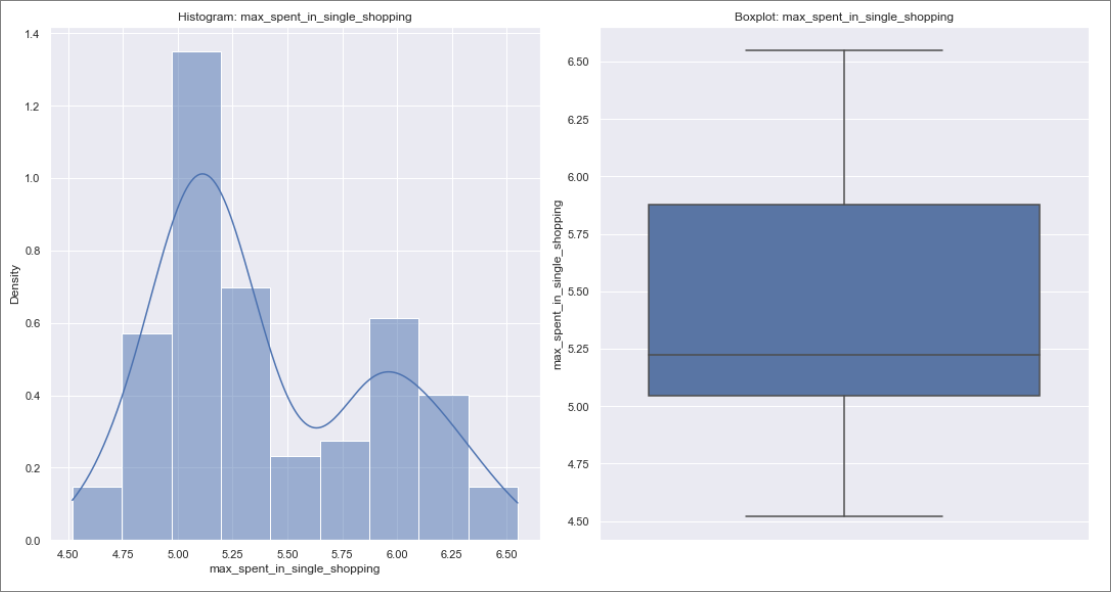


Figure-1.7 Histogram & Boxplot : max\_spent\_in\_single\_shopping

Figure-1.7 depicts the histogram and boxplot of “max\_spent\_in\_single\_shopping” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 7.141e-9 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “max\_spent\_in\_single\_shopping” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘max\_spent\_in\_single\_shopping’.

#### Outlier Treatment:

We have outliers present in 2 variables (probability\_of\_full\_payment & min\_payment\_amt) as seen in figures 1.3 and 1.6. We will impute the outliers with inter quartile min and maximum values for outliers below the quartile minimum and quartile maximum respectively.

Let’s check the boxplot now:

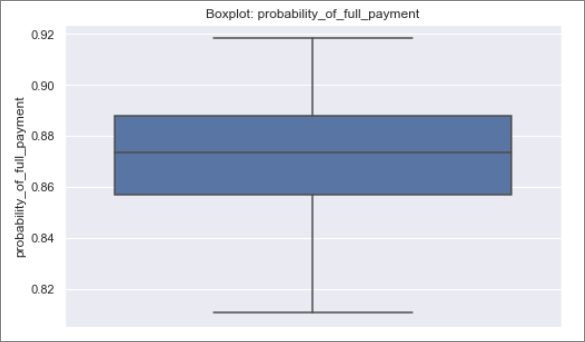
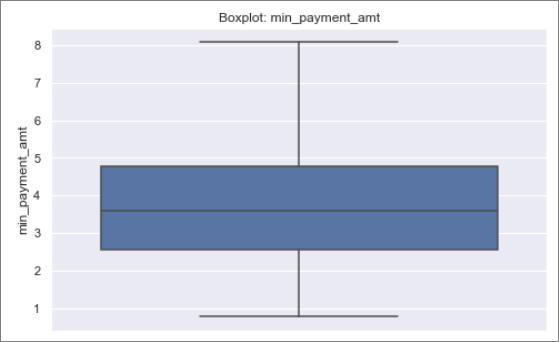
 

Figure-1.8 Boxplot : probability\_of\_full\_payment & min\_payment\_amt

We can see that there are no outliers in the above boxplots, after data has been imputed.

The updated data summary is shown below:

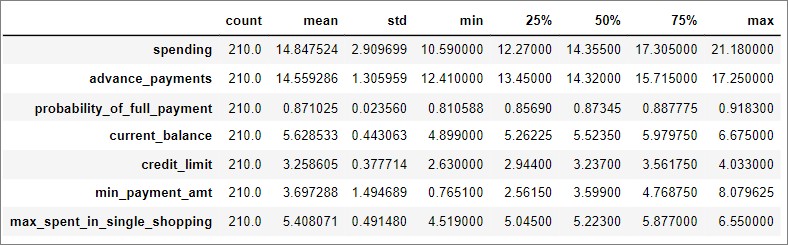


Table-1.4 Data Summary

#### Bivariate Analysis:

Let us plot a heat map for the correlation matrix of given data frame.

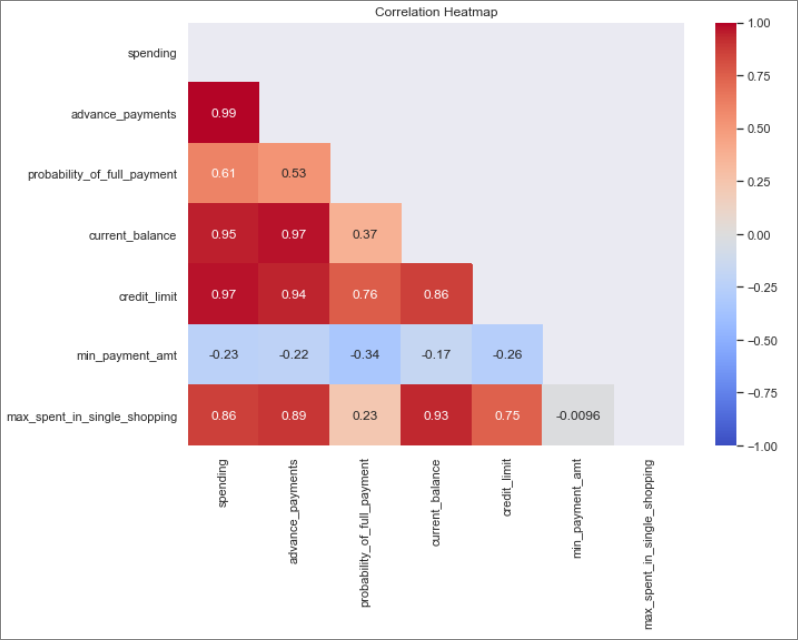


Table-1.5 Correlation matrix

We can see very strong positive correlation between the following variables:

* advance\_payments, spending, current\_balance and credit\_limit

This is expected as people with higher spending power are rewarded with higher credit limit and will do more advance payments and maintain higher balance in their accounts.

We can see weak negative correlation in min\_payment\_amt against all the other variables.

Let’s also check the Pairplot:

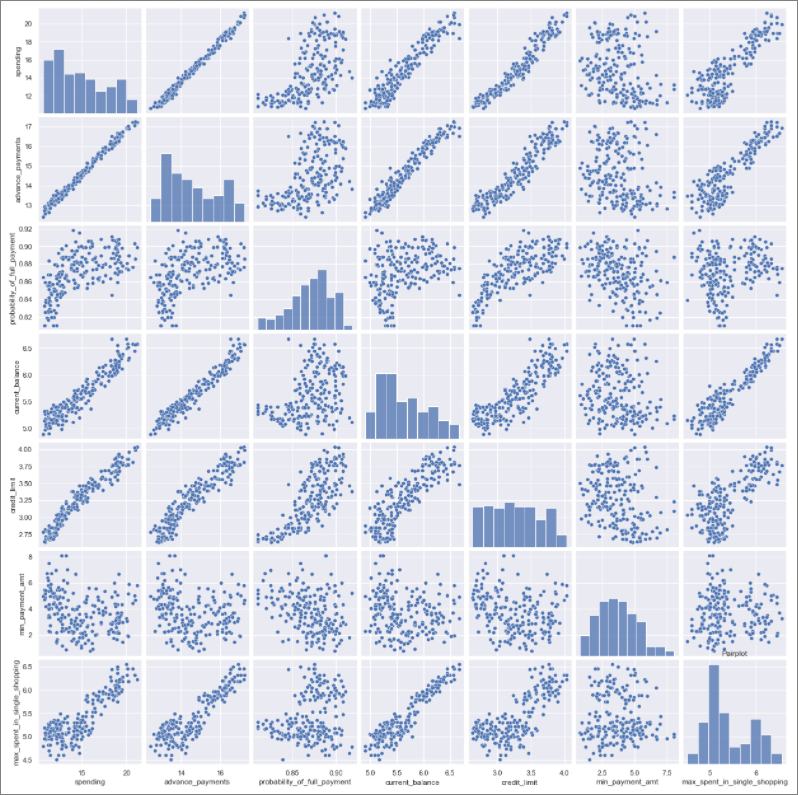


Figure-1.9 Pairplot

Pairplot substantiates the heatmap and displays the correlation between variables as stated above.

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

Since clustering is based on Euclidian distance to form the groups(clusters), we need to ensure that all the variables are in the same scale to justify creation of proper clusters. If we look at the means of variables in Table 1.4, they range from 0.87 to 14.85, wherein some variables are probabilities (value ranging 0 to 1), while variables such as spending is in the range of 10.59 to 21.18. Hence if we perform clustering without scaling, the groups would be more influenced by spending rather than probability\_of\_full\_payment.

If we were to plot the values,

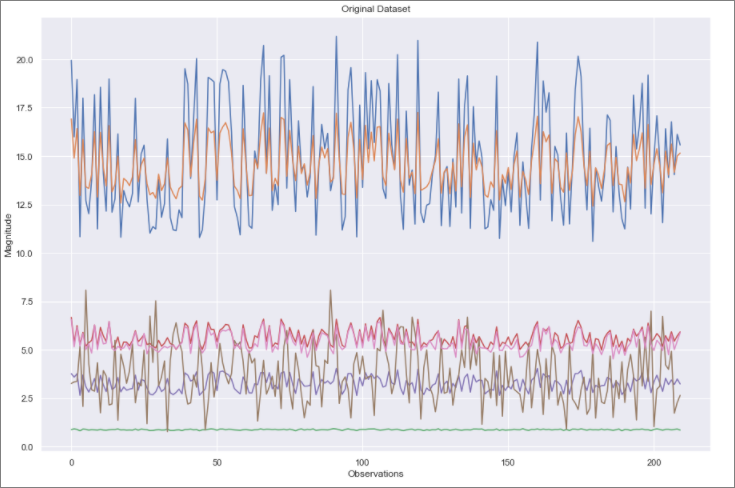


Figure-1.10 Original dataset plot

Let’s perform scaling and check the data summary.

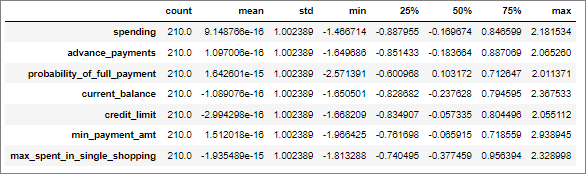


Table-1.6 Data summary of scaled data

From Table 1.6 we can see that all the means are close to 0 and standard deviation is close to 1.

We will check the plot of the scaled data:

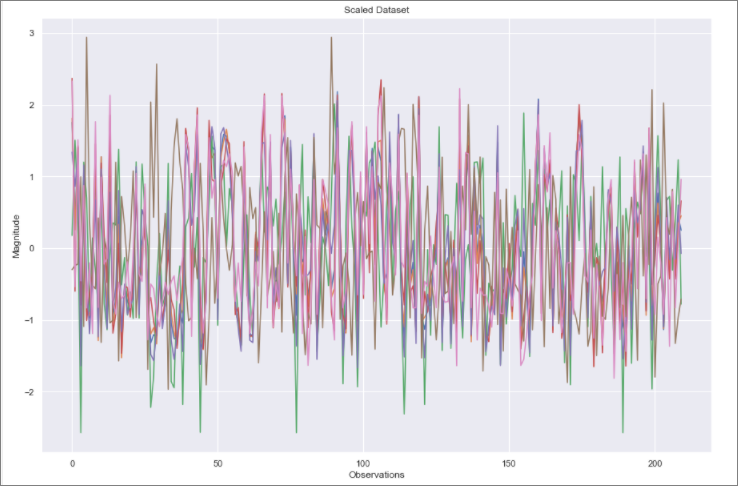


Figure-1.11 Scaled dataset plot

So, in the scaled dataset we have given equal weightage to all the variables, so that while clustering based on Euclidean distance, one variable wouldn’t have more influence on others just based on magnitude.

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

Let us use fcluster(flat cluster for hierarchical) with linkage method as average.

The dendrogram for the scaled data set is displayed below:

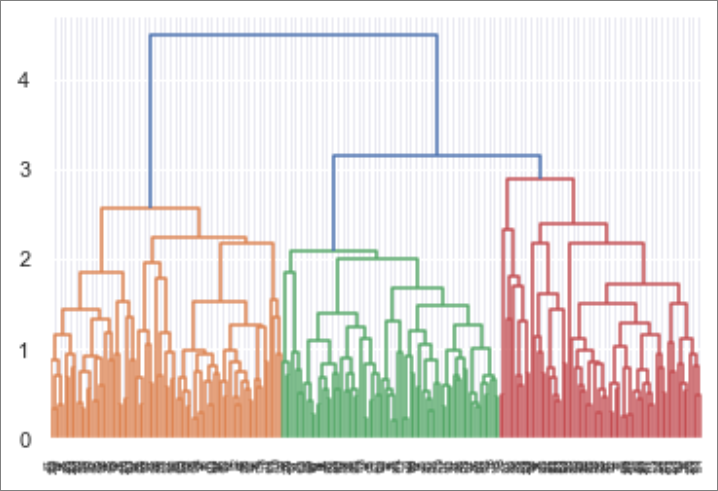


Figure-1.12 Dendrogram

Let us truncate the dendrogram to see the initial 10 clusters.

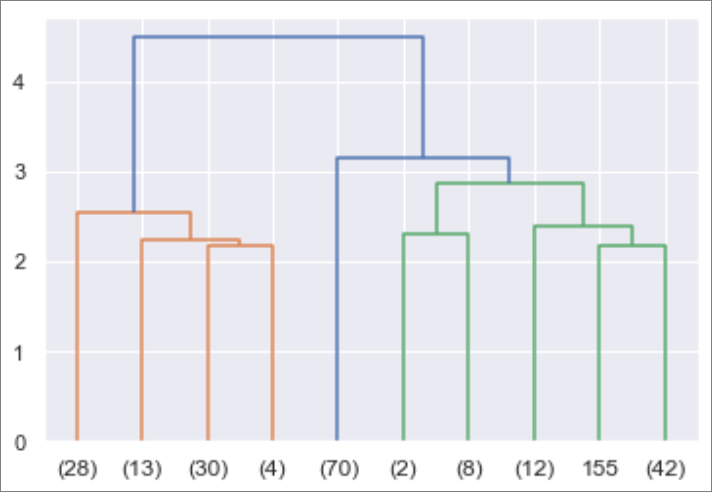


Figure-1.13 Truncated Dendrogram

From the above dendrogram we can see that at distance 3, we can have 3 clusters and at about 2.7 we can have 4 clusters. Since we are trying to classify basis on customer spend, a normal grouping could be High, Medium and Low, which amounts to 3 clusters. We will also try to see with 4 clusters and see if any meaningful cluster profile emerges.

#### Hierarchical clustering (fcluster) based on 3 clusters:

On forming flat clusters from the hierarchical clustering defined by the average linkage matrix, we have the following mean values across the variables for all 3 clusters.

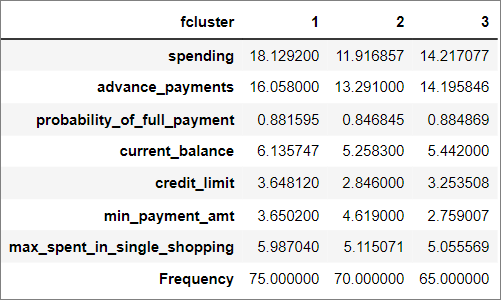


Table-1.7 Hierarchical cluster profile for 3 clusters

As initially surmised, we have 3 clusters with profile as follows:

Cluster 1 - High spending customers with higher credit limit and probability of full payments.

Cluster 2 - Low spending customers with lowest credit limit and lowest probability of full payment. They have the highest min\_payment\_amt, which could be because the frequency of shopping for these customers are less, and they prefer to buy in bulk.

Cluster 3 - Medium spending customers with mid-level credit limit and highest probability of full payment.

#### Hierarchical clustering (fcluster) based on 4 clusters:

On forming flat clusters from the hierarchical clustering defined by the average linkage matrix, we have the following mean values across the variables for all 4 clusters.

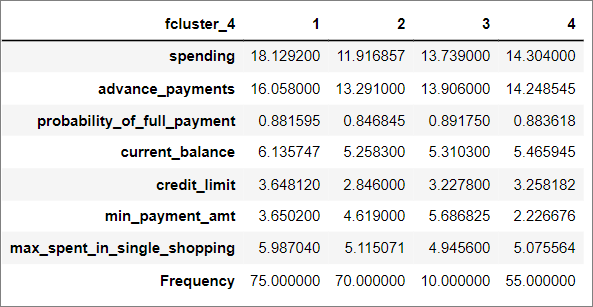


Table-1.8 Hierarchical cluster profile for 4 clusters

Clusters 1 and 2 retain the same characteristics as that seen above for 3 cluster profiling:

The medium spending customers seen in fcluster-3 has been further divided into 2 clusters

Cluster 3 - Lower of the medium spending customers, but with highest probability of full payment. They also have the highest min\_payment\_amt which indicates that the frequency of shopping is low.

# Cluster 4 - Higher of the medium spending customers. They have the least min\_payment\_amt across the clusters, which indicates that they are frequent shoppers.

Since the frequency of this new cluster is less than 5% of the total observations, it would be optimal to consider 3 clusters rather than 4 clusters.

Let’s re-arrange the cluster values as 1 for High spending, 2 for Medium spending and 3 for Low spending customers. The cluster profile will be as shown below:

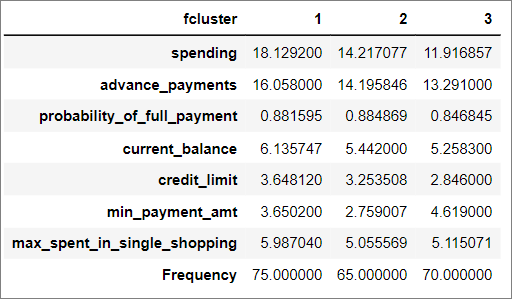


Table-1.9 Hierarchical cluster profile for 3 clusters

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

Let us apply K-Means clustering on the scaled dataset. We will run K-Means clustering for a range of clusters and look at the inertia across clusters (within cluster sum of squares) as shown below:

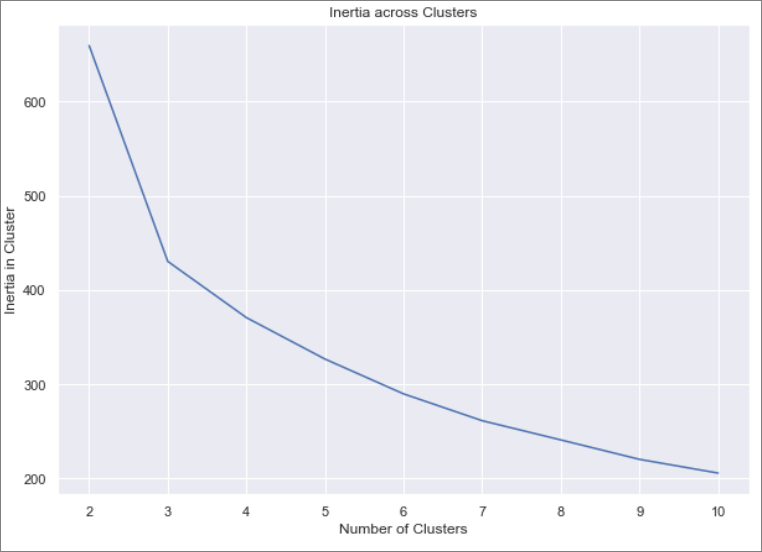


Figure-1.14 Inertia across clusters

The above elbow curve does not show a clear elbow, but the optimal cluster should be one of the clusters ranging from 3 to 6.

Let’s plot the silhouette scores for the range of clusters and see which cluster has the maximum silhouette score.

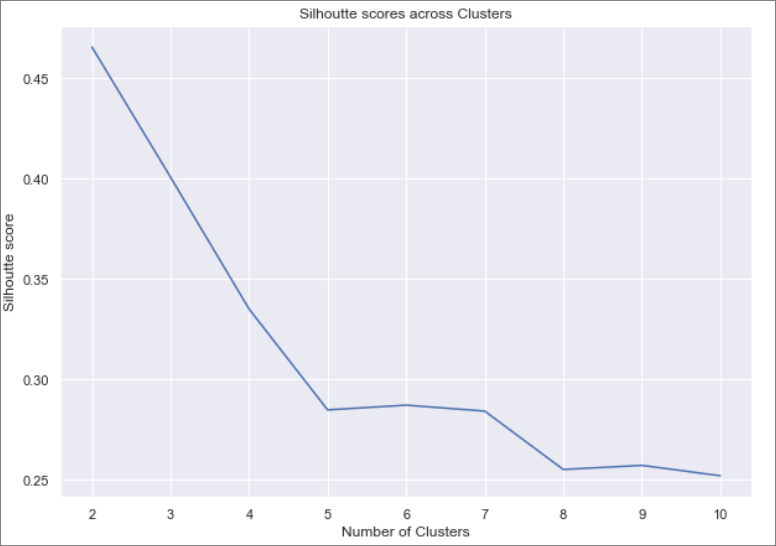


Figure-1.15 Silhouette scores across clusters

Ideal cluster should high silhouette score, in the above graph 2 is showing highest silhouette score, but this grouping will not be able to provide much context from a business perspective. So, let’s consider clusters 3,4,5 and 6 and check their minimum silhouette sample value.

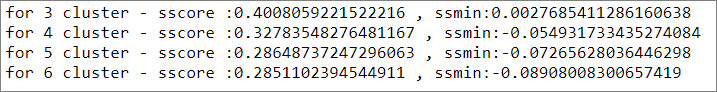


Table-1.10 Minimum silhouette coefficient across clusters

From the above minimum silhouette coefficient across clusters, we can surmise that cluster 3 is optimal as min value is greater than 0, showing that the clusters don’t overlap, whereas for other cluster values, there are some overlapping.

#### K-Means clustering based on 3 clusters:

On performing K-Means clustering for 3 clusters on the scaled dataset, we have the following mean values across the variables for all 3 clusters.

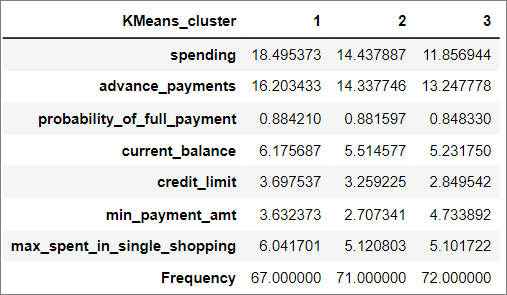


Table-1.11 K-Means cluster profile for 3 clusters

Of the 210 observations, we have 3 clusters of 67, 71 and 72 observations.

Cluster 1 – Highest spending customers with mid-level min\_payment\_amt and all other parameter values higher than other clusters.

Cluster 2 – Medium spending customers with lowest min\_payment\_amt (frequent shoppers) and all other parameter values at mid-level.

Cluster 3 – Lowest spending customers with highest min\_payment\_amt(lowest frequency of shopping) and all other parameter values at lowest level.

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

Let’s look at both the hierarchical and K-mean cluster profiles:

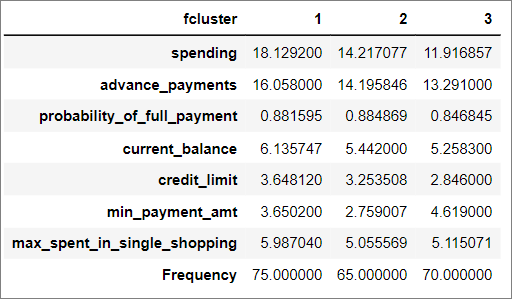
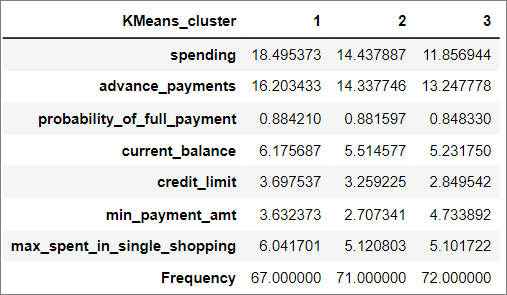
 

Table-1.12 Hierarchical cluster profile for 3 clusters Table-1.13 K-Means cluster profile for 3 clusters

The fcluster and K-Means cluster for 3 clusters have shown the same trend across the variables except for probability\_of\_full\_payment, where K-Means has shows that highest spending customers also has the highest probability of making a full payment, which stands to reason.

Hence we will consider the K-Means cluster profile for the description and promotional strategies the bank could consider.

#### Cluster profiling:

Cluster 1 – These are the customers with the highest spending capacity. They have excellent credit limit and retain highest probability for paying the full payment. They maintain higher balance and is in the forefront when it comes to advance payments. The minimum payment amount is average, but the frequency of shopping is higher than customers with least spending power. Comparing with other clusters these customers have higher spending power per purchase.

Cluster 2 – These are the customers with medium spending capacity. They have the least min\_payment\_amt which indicates these are frequent shoppers. They are at the mid-level in all other variables. Their probability of full payment is very near to Cluster 1 and hence are credit-worthy customers.

Cluster 3 – These are the customers with least spending power. They are at the low-level across variables, except min\_payment\_amt, which is the highest among clusters. This indicates these customers are the least frequent shoppers and must be doing bulk purchases towards monthly needs.

#### Promotional Strategies across Clusters:

Cluster 1

* Increase credit limit for these customers, as their probability of full payment is higher.
* Increase spending habit of these customers by having a rewards program, with points based on spending tiers as part of an elite group.
* Provide rebates on higher spend limits as these customers tend to have higher spend limit per purchase.
* Collaborate with luxury goods manufacturer and have offers, which would entice these customers.

Cluster 2

* Increase credit limit for these customers as they have comparable probability of full payment with cluster 1.
* These customers are frequent shoppers and hence have promotional offers across ecommerce sites and a robust rewards program.
* Provide loan with lower interest rates, as the chance of repayment is high.
* Have a rewards program for these customers, with points based on spending tiers.

Cluster 3

* Implement rewards program with brownie points for early payments.
* Have tie-ups with departmental stores and provide incentives for higher spends.
* Have a bill payments program with incentives in the point of rewards or rebates.

Please find below the output file with all input data and additionally with the fcluster and KMeans\_cluster data. Both these columns have below mentioned values:

1 – High spending customers

2 – Medium spending customers

3 – Lowest spending customers.



# 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.



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

### Data Description

1. Age of insured (Age)
2. Code of tour firm (Agency\_Code)
3. Type of tour insurance firms (Type)
4. Target: Claim Status (Claimed)
5. The commission received for tour insurance firm (Commission is in percentage of sales)
6. Distribution channel of tour insurance agencies (Channel)
7. Duration of the tour (Duration in days)
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)
9. Name of the tour insurance products (Product)
10. Destination of the tour (Destination)

### Sample of the dataset:

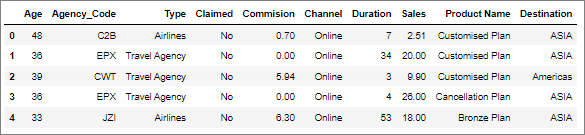


Table-2.1 Dataset Sample

### Exploratory Data Analysis:

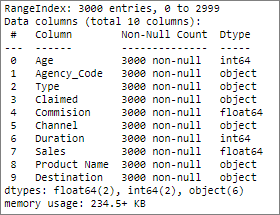


Table-2.2 Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 3000 observations and 10 columns in the dataset. We have 2 integer type columns, 6 object type columns and 2 float type columns.

#### Check for missing values in the dataset

From Table-2.2 we can see that all the columns have 3000 non-null values and hence we have no missing values in the dataset.

#### Check for duplicate observations in the dataset

There are 139 duplicate rows in the dataset as can be seen below:

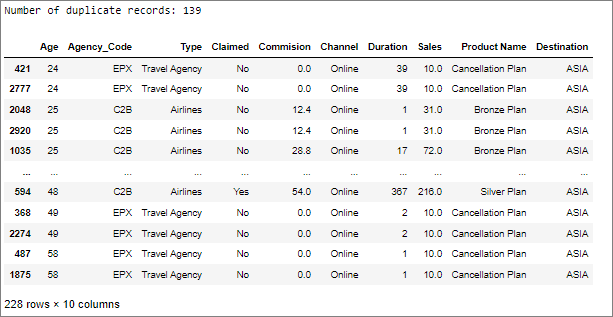


Table-2.3 Duplicate information

Since we do not have any unique customer identification column, the duplicates could have been raised by different customers of same age. With this assumption we will not delete any of the duplicate rows.

#### Data summary

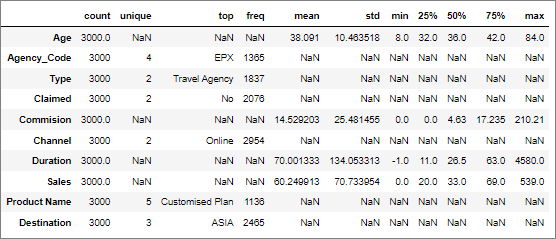


Table-2.4 Data Summary

The above data summary will be further explained in the univariate analysis section below.

#### Univariate Analysis:

Let’s check the central measures of tendency, quartiles, histogram and boxplot for integer/float columns.

For object columns we will find the count of categorical values via count plot.

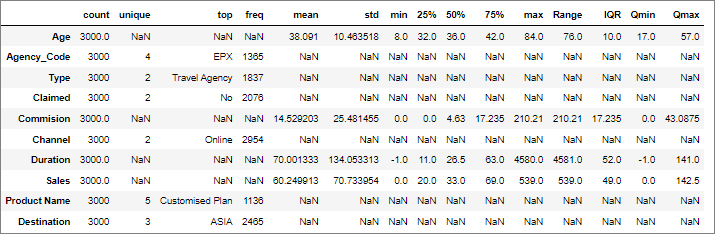


Table-2.5 Data Summary with quartile values and range

1. Age

Age of insured customer is a continuous variable with the below stats (refer Table 2.5):

Mean = 38.091

Standard Deviation = 10.464

Min value in dataset = 8

Max value in dataset = 84

Range = Max – Min = 76

Q1(1st Quartile) = 32

Q2(2nd Quartile)/Median = 36

Q3(3rd Quartile) = 42

IQR(Inter-Quartile Range) = Q3- Q1 = 10

Quartile Min value = max(Q1 – 1.5 \* IQR, min) = 17

Quartile Max value = min(Q3 + 1.5 \* IQR, max) = 57

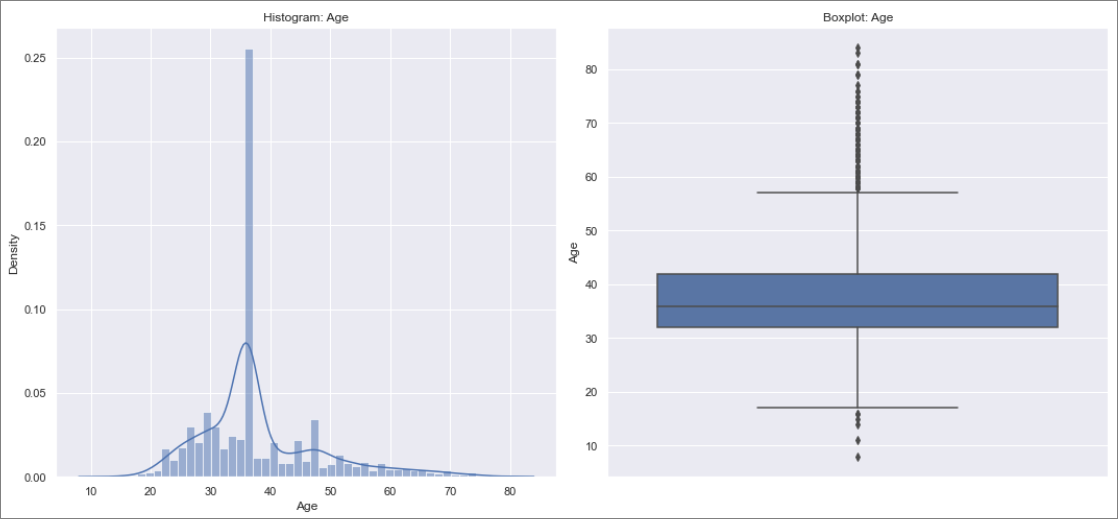


Figure-2.1 Histogram & Boxplot : Age

Figure-2.1 depicts the histogram and boxplot of “Age” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 1.07e-40 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “age” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘age’, but these are expected values of customer ages and hence we will not be treating this column for outliers.

1. Agency\_Code

Code of tour firm is a discrete variable with the below stats (refer Table 2.5):

There are 4 unique values and Agency\_code ‘EPX’ is maximum across the observations with a count of 1365.

The distribution of agency codes are as follows in the dataset:

EPX 1365

C2B 924

CWT 472

JZI 239

Let’s look at the count-plot and pie-chart:

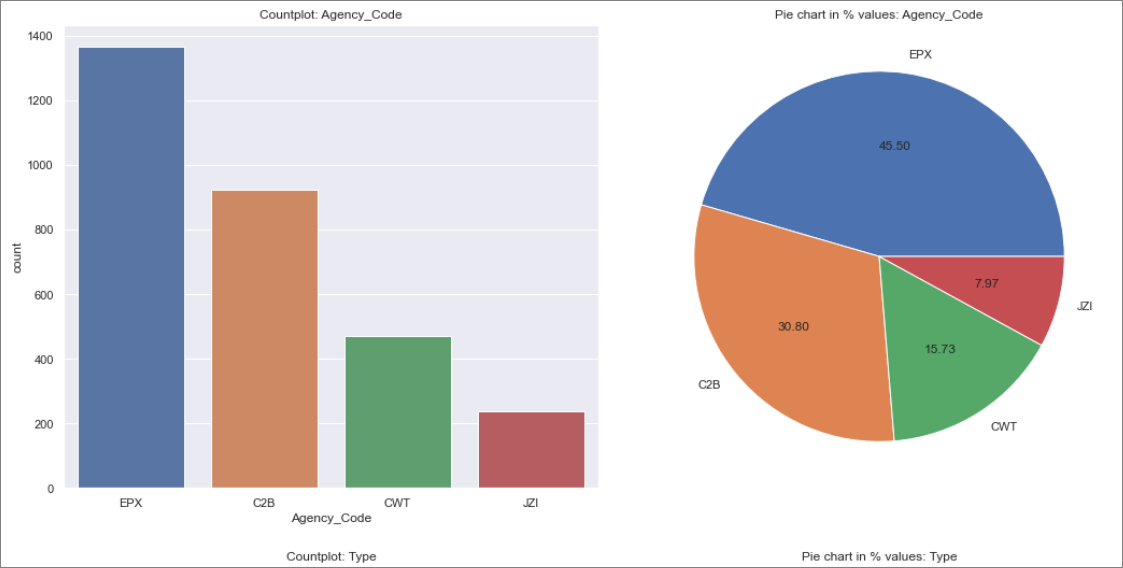


Figure-2.2 Count-plot & Pie-chart: Agency\_code

In the given dataset we can see that EPX conducts max of the tours with 1365 observations (45.50%) and JZI conducts the least at 239 observations (7.97%).

1. Type

Type of tour insurance firms is a discrete variable with the below stats (refer Table 2.5):

There are 2 unique values, and ‘Travel Agency’ is maximum across the observations with a count of 1837.

The distribution of type are as follows in the dataset:

Travel Agency 1837

Airlines 1163

Let’s look at the count-plot and pie-chart:

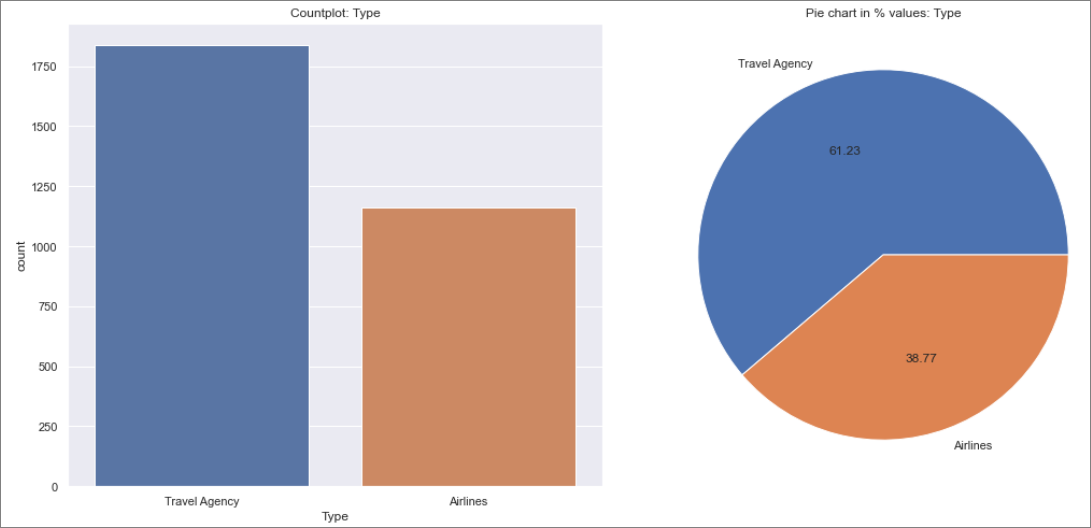


Figure-2.3 Count-plot & Pie-chart: Type

In the given dataset we can see that ‘Travel Agency’ conducts max of the tours with 1837 observations (61.23%) and Airlines conducts the least at 1163 observations (36.77%).

1. Claimed

Claim status (target variable) is a discrete variable with the below stats (refer Table 2.5):

There are 2 unique values and claimed ‘No’ is maximum across the observations with a count of 2076.

The distribution of claim status are as follows in the dataset:

No 2076

Yes 924

Let’s look at the count-plot and pie-chart:

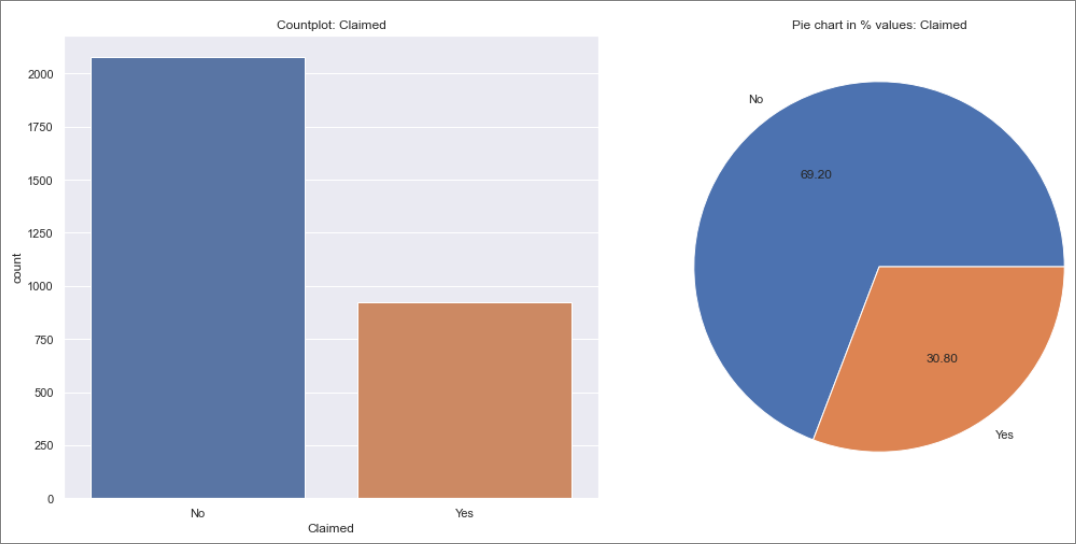


Figure-2.4 Count-plot & Pie-chart: Claimed

In the given dataset we can see that claim status “No” has highest frequency with 2076 observations (69.20%) and we can see number of claims are at 924 observations (30.80%).

1. Commission

The commission received for tour insurance firm is a continuous variable with the below stats (refer Table 2.5):

Mean = 14.529

Standard Deviation = 25.481

Min value in dataset = 0

Max value in dataset = 210.21

Range = Max – Min = 210.21

Q1(1st Quartile) = 0

Q2(2nd Quartile)/Median = 4.63

Q3(3rd Quartile) = 17.235

IQR(Inter-Quartile Range) = Q3- Q1 = 17.235

Quartile Min value = max(Q1 – 1.5 \* IQR, min value) = 0

Quartile Max value = min(Q3 + 1.5 \* IQR, max value) = 43.0875

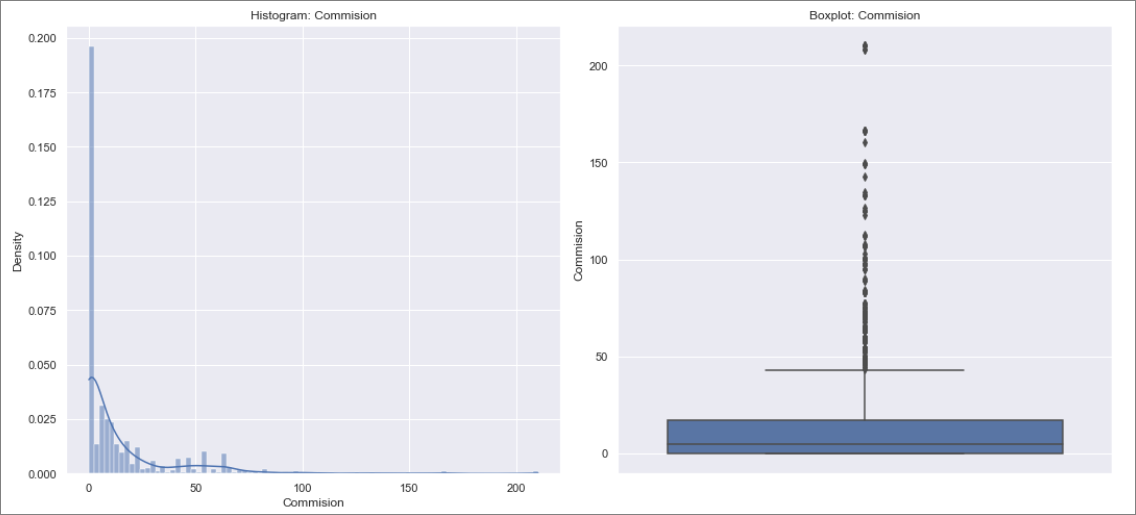


Figure-2.5 Histogram & Boxplot : Commission

Figure-2.5 depicts the histogram and boxplot of “Commission” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 0 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Commission” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘commission’ and looking at the number of outliers present, it does not look like erroneous data. We will not be treating these outliers.

1. Channel

Distribution channel of tour insurance agencies is a discrete variable with the below stats (refer Table 2.5):

There are 2 unique values and Online is maximum across the observations with a count of 2954.

The distribution of channels are as follows in the dataset:

Online 2954

Offline 46

Let’s look at the count-plot and pie-chart:

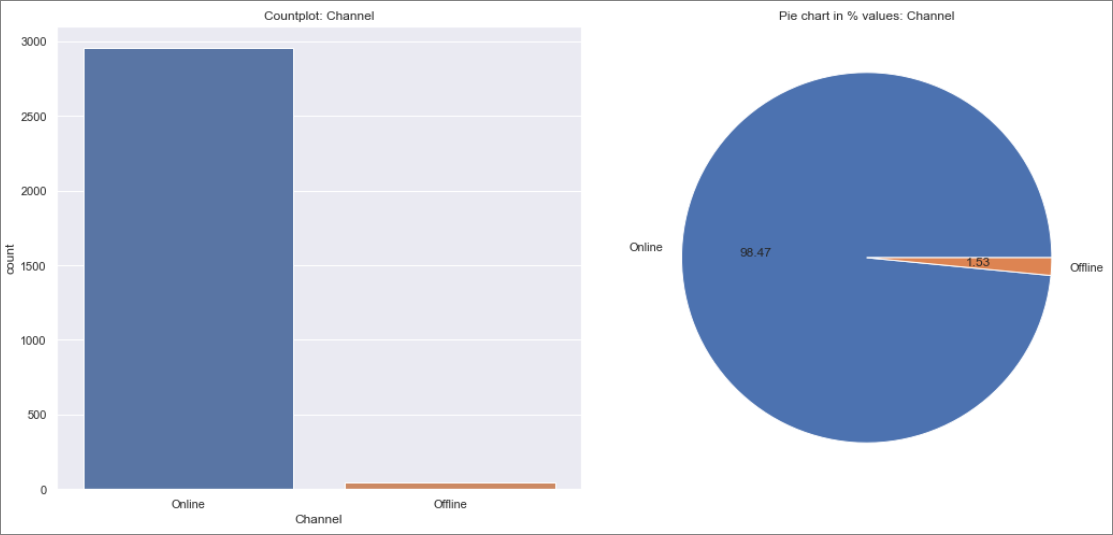


Figure-2.6 Count-plot & Pie-chart: Channel

In the given dataset we can see that preferred channel for customers is “Online” with highest frequency of 2954 observations (98.47%) and we can see “Offline” is in 46 observations (1.53%).

1. Duration

Duration of the tour (in days) is a continuous variable (refer Table 2.5):

We can see that duration has a min value of -1, which is not possible.

Let’s check the number of records with duration < 0

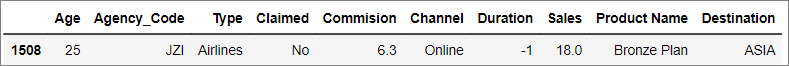


Table-2.6 records with duration < 0

Let’s impute the duration of -1, with the average duration of records which has Agency - JZI, Product - Bronze Plan, Destination -ASIA. The above observation has a sales value of 18, so let’s consider the average duration for sales value between 15 and 20, for all the mentioned parameters.

The average duration of the above criteria comes to 12 days.

Let’s impute the value and check the modified record and updated data summary:

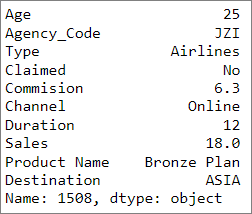


Table-2.7 Updated record

Let’s check the updated data summary table:

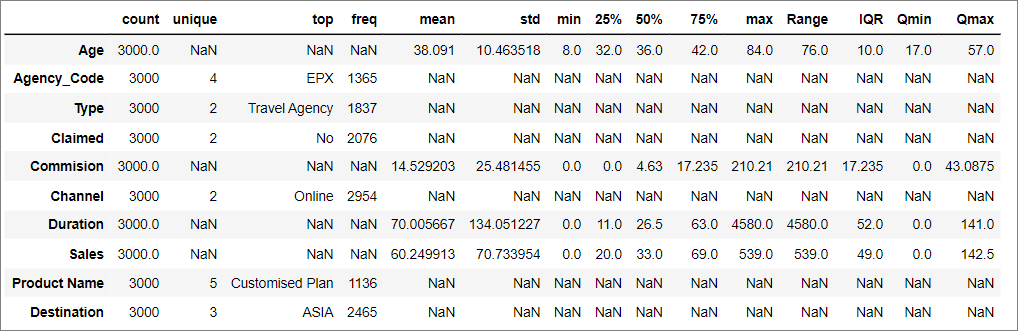


Table-2.8 Data summary

From table 2.8 we have the below stats for ‘Duration’ column:

Mean = 70

Standard Deviation = 134.05

Min value in dataset = 0

Max value in dataset = 4580

Range = Max – Min = 4580

Q1(1st Quartile) = 11

Q2(2nd Quartile)/Median = 26.5

Q3(3rd Quartile) = 63

IQR(Inter-Quartile Range) = Q3- Q1 = 52

Quartile Min value = max(Q1 – 1.5 \* IQR, min) = 0

Quartile Max value = min(Q3 + 1.5 \* IQR, max) = 141

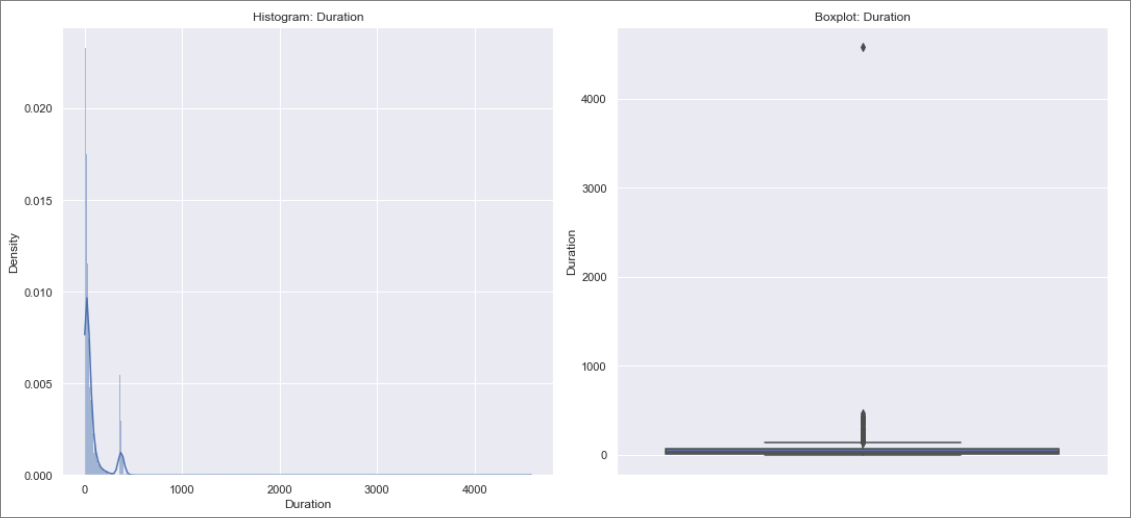


Figure-2.7 Histogram & Boxplot : Duration

Figure-2.7 depicts the histogram and boxplot of “Duration” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 0 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Duration” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘Duration’, but these don’t look like erroneous data.

1. Sales

Amount worth of sales per customer in procuring tour insurance policies is a continuous variable with the below stats (refer Table 2.8):

Mean = 60.25

Standard Deviation = 70.734

Min value in dataset = 0

Max value in dataset = 539

Range = Max – Min = 539

Q1(1st Quartile) = 20

Q2(2nd Quartile)/Median = 33

Q3(3rd Quartile) = 69

IQR(Inter-Quartile Range) = Q3- Q1 = 49

Quartile Min value = max(Q1 – 1.5 \* IQR, min value) = 0

Quartile Max value = min(Q3 + 1.5 \* IQR, max value) = 142.5

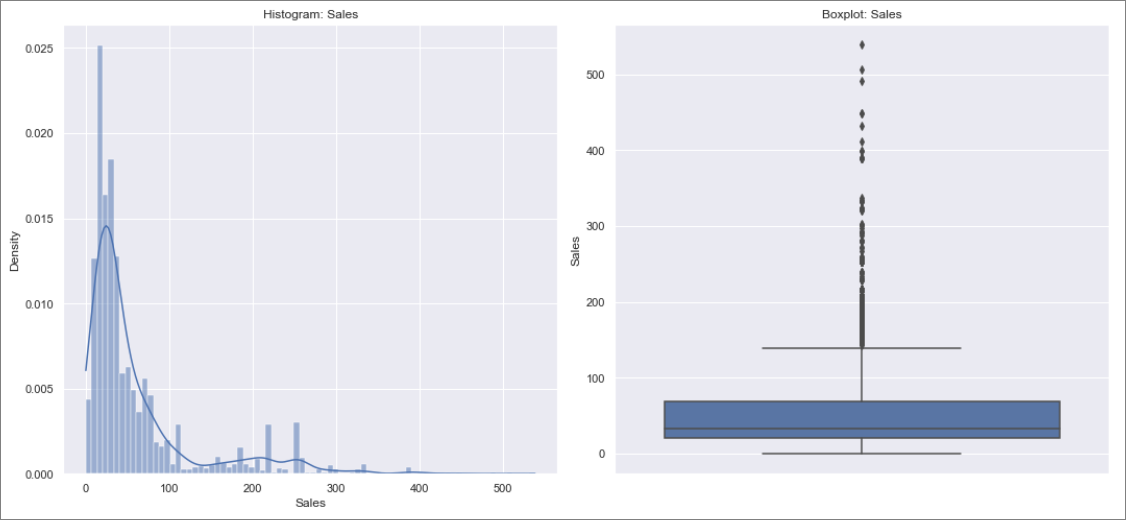


Figure-2.8 Histogram & Boxplot : Sales

Figure-2.8 depicts the histogram and boxplot of “Sales” which appears to be normal in nature but the Shapiro-Wilk test (can be performed on a continuous variable) returns a pvalue of 0 which is much lesser than 0.05, and hence we will have to reject the null hypothesis that the data is normally distributed. Hence the “Sales” column data does not have a normal distribution, as per the provided observations.

From the boxplot we can see that there are outliers present in ‘Sales’, but it does not look like erroneous data. We will not be treating these outliers.

1. Product Name

Name of the tour insurance products is a discrete variable with the below stats (refer Table 2.8):

There are 5 unique values, and ‘Customised Plan’ is maximum across the observations with a count of 1136.

The distribution of Product Names are as follows in the dataset:

Customised Plan 1136

Cancellation Plan 678

Bronze Plan 650

Silver Plan 427

Gold Plan 109

Let’s look at the count-plot and pie-chart:

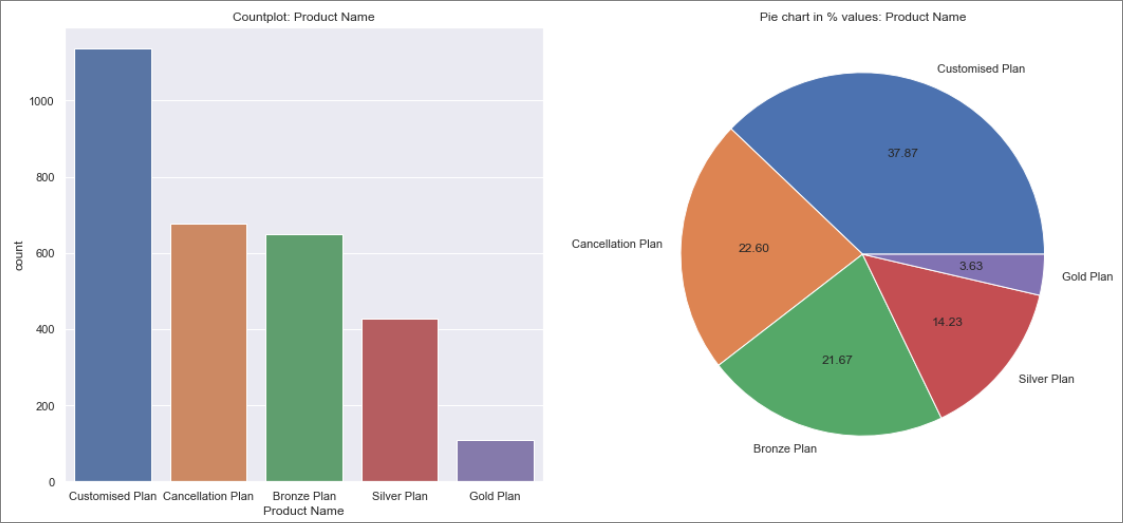


Figure-2.9 Count-plot & Pie-chart: Product Name

In the given dataset we can see that preferred product is ‘Customised Plan’ for customers with highest frequency of 1136 observations (37.87%) and we can see least opted product is ‘Gold Plan’ with 109 observations (3.63%).

1. Destination

Destination of the tour is a discrete variable with the below stats (refer Table 2.8):

There are 3 unique values, and ‘ASIA’ is maximum across the observations with a count of 2465.

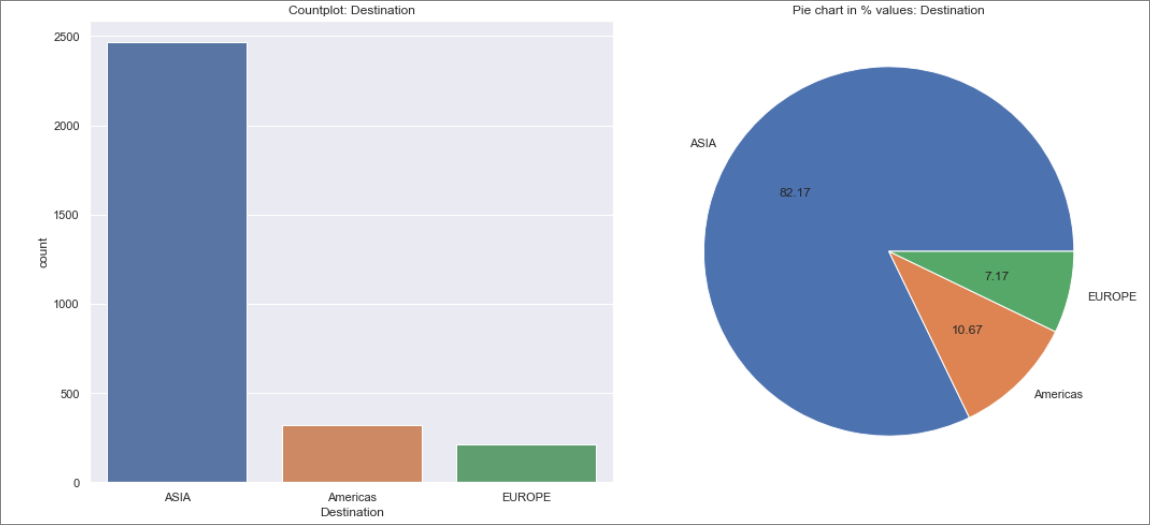
The distribution of destination are as follows in the dataset:

ASIA 2465

Americas 320

EUROPE 215

Let’s look at the count-plot and pie-chart:

 Figure-2.10 Count-plot & Pie-chart: Destination

In the given dataset we can see that preferred destination is ‘ASIA’ for customers with highest frequency of 2465observations (82.17%) and we can see least toured destination is ‘Europe’ with 215 observations (7.17%).

#### Bivariate Analysis:

Let us plot a Pairplot and check for correlation:

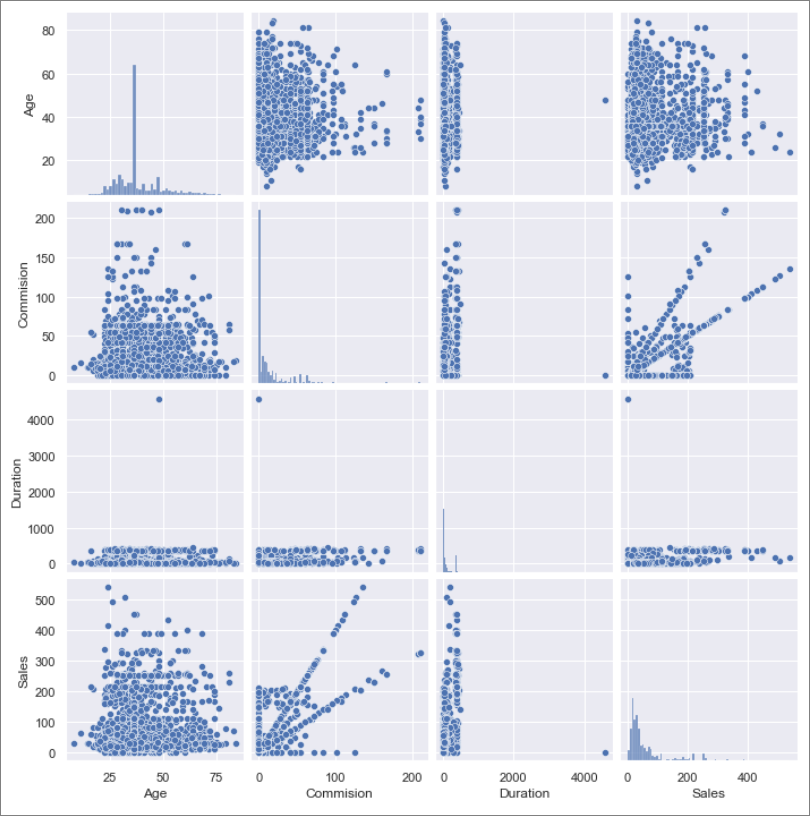


Figure-2.11 Pairplot

From the Pairplot we can see a positive correlation between sales and commission.

Let us plot a heat map for the correlation matrix of given data frame.

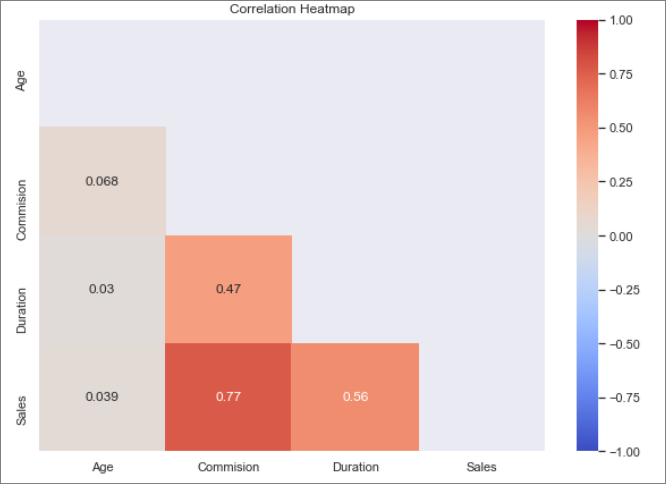


Figure 2.12 Correlation matrix

As seen from the Pairplot there is a strong positive correlation between sales and commission. Sales and Duration have a medium positive correlation. All other continuous variable has weak correlation with others.

#### Multivariate Analysis:

Let’s check the claim status against the tour agency:

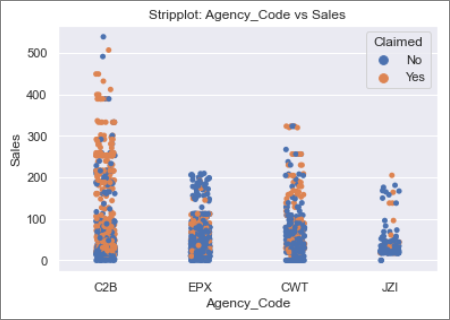
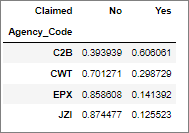
 

Figure 2.13 Stripplot: Agency\_Code vs Sales

From figure 2.13 we can see claim status has been raised in tours conducted by ‘C2B’ at maximum and least is at JZI.

Let’s check the claim status against Product Name:

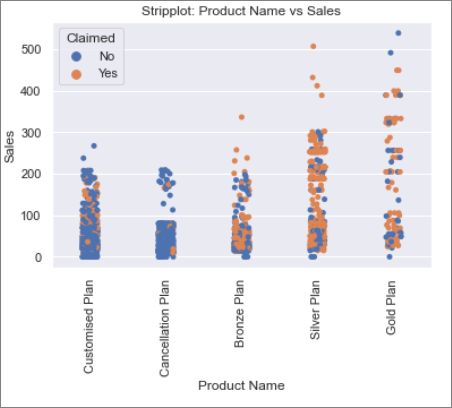
 

Figure 2.14 Stripplot: Product Name vs Sales

From the above plot we can see that the percent of claims are higher in the last 2 product names - Silver Plan and Gold Plan.

Let’s check the claim status against Destination:

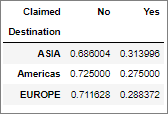
 

Figure 2.15 Stripplot: Destination vs Sales

F

From the above plot we can see that claims are raised more when destination is ASIA.

Let’s check the claim status against Type:

From the below plot we can see that more tours are conducted by Travel Agency, but we have higher number of claims that Travel Agency at Airlines.

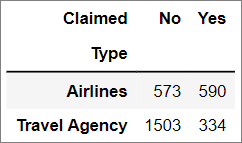
 

Figure 2.16 Stripplot: Type vs Sales

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

As a first step lets convert all object columns into category codes for the CART,RF and ANN models to be executed.

The codes are assigned as shown below:

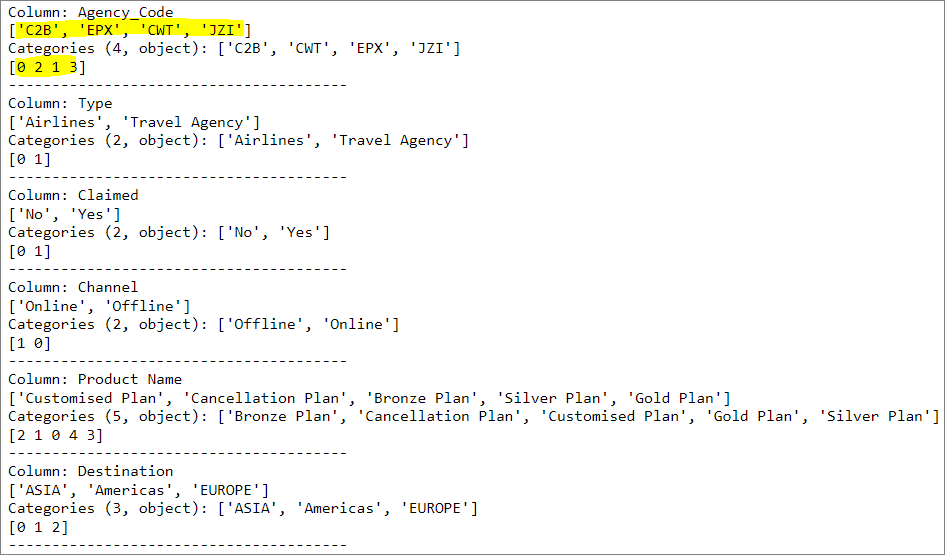


Table-2.9 Categorical code mapping

The above table can be inferred as follows:

For column Agency\_Code we have 4 categories as highlighted. The highlighted number 0,2,1 and 3 denote the highlighted categories respectively, i.e., ‘C2B’ is mapped to 0, ‘EPX’ to 1 and so on.

Our target variable ‘Claimed’ has 0 denoting ‘No’ and 1 denoting ‘Yes’.

Let’s check the concise data summary:

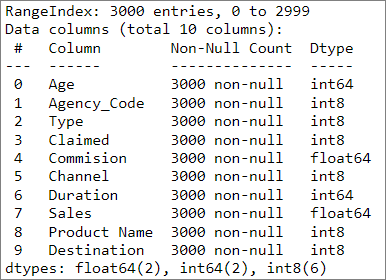


Table-2.10 Concise data summary

#### Data Split – Test & Train:

We will split the input dataset into training and test datasets in the ratio (70:30). This is an accepted convention where we will train the model with about 2/3rds of the data we have, and we will evaluate the model against the test data which is the remaining 1/3rds from the input data set.

After performing the above split, lets look at the frequency of data in training/test target variable.

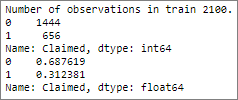
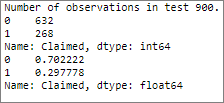
 

Table-2.11 Training/Test Target variable summary

From the above table we can see that training dataset has 2100 observations and test dataset has 900 (which is the 70:30 slit of the original 3000 observations).

We can see that the percentage of claimed in both training and dataset is ~30%, which is apt, as we do not want to have the percentage very different across both the datasets.

Let’s look at each of the classification models applied across the above training/test datasets.

#### CART model:

We will implement CART model via Decision Tree classifier present in sklearn library's cluster module.

We will not be scaling the data for Decision tree classifier as they are not sensitive to the variance in data.

We have 2100 observations in training set. So, we can have ~1% of 2100, 20 as min\_samples\_leaf and 3 times min\_sample\_leaf as min\_samples\_split, i.e., 60. We will start with max\_depth as 10. We will consider above values as mid points and run GridSearchCV for above mentioned values together with values lesser and greater. GridSearchCV will train the model based on the hyperparameters provided and will lets us know of the best value for each parameter. We will use these best values as reference points and repeat this whole cycle of GridSearchCV execution on training data set using the Decision tree classifier, until we reach the optimal values.

CART model for the training dataset with ideal parameters are as shown below:

DecisionTreeClassifier(max\_depth=5,

min\_samples\_leaf=11,

min\_samples\_split=40)

We had initialized the classifier with criteria as ‘gini’ and random\_state as 123.

Based on this model, the feature importance has been identified as shown in table 2.12.

We can see that Agency\_Code and Sales together have a combined importance of 77% towards determining the target variable. Type and Destination features do not contribute towards determining the target variable as per CART model.

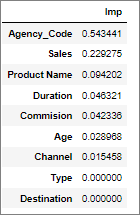


Table-2.12 CART model feature importance

#### Random Forest model:

We will implement Random Forest classifier model via RandomForestClassifier in sklearn library's ensemble module. We will not be scaling the data as ensemble models are not sensitive to the variance in data.

We have 2100 observations in training set. So, we can have ~1% of 2100, 20 as min\_samples\_leaf and 3 times min\_samples\_leaf as min\_samples\_split, i.e., 60. We will consider max\_depth as 10. Let’s set cross validation to 3, as random forest will take considerable time to run. Default value of n\_estimators is 100, we will try to check for 201 and 301 to see if it works better for this dataset. We can set start value of max\_features to be near to square root of number of features(so in this case number of features is 9, let’s start with value 4). We will consider above values as mid points and run GridSearchCV for above mentioned values together with values lesser and greater. GridSearchCV will train the model based on the hyperparameters provided and will lets us know of the best value for each parameter. We will use these best values as reference points and repeat this whole cycle of GridSearchCV execution on training data set using the Random Forest model, until we reach the optimal values.

Random Forest model for the training dataset with ideal parameters are as shown below:

RandomForestClassifier(max\_depth=9,

max\_features=4,

min\_samples\_leaf=20,

min\_samples\_split=60,

n\_estimators=201)

We had initialized the classifier with criteria as ‘gini’ and random\_state as 123.

Based on this model, the feature importance has been identified as shown in table 2.13.

We can see that Agency\_Code, Product Name and Sales together have a combined importance of 70% towards determining the target variable. Channel is the least important feature towards determining the target variable as per Random Forest model.

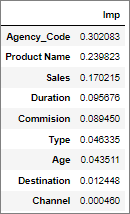


Table-2.13 Random Forest model feature importance

#### Artificial Neural Network model:

We will implement Artificial Neural Network(ANN) model via MLPClassifier in sklearn library's neural\_network module. We need to perform scaling in ANN. We will use standard scaler to fit and transform the training dataset and subsequently transform the training dataset. Default hidden layer size is 100, so we will also give value 50 to see how the grid search evaluates. Default activation is 'relu', we will also use 'logistic' to see which helps in converging the model against training dataset faster. We will evaluate both sgd and adam for gradient based optimization. default tolerance is 0.0001, we will also try for 0.001.

We will consider above values as mid points and run GridSearchCV for above mentioned values together with values lesser and greater. GridSearchCV will train the model based on the hyperparameters provided and will lets us know of the best value for each parameter. We will use these best values as reference points and repeat this whole cycle of GridSearchCV execution on training data set using the Artificial Neural Network model, until we reach the optimal values.

Artificial Neural Network model for the training dataset with ideal parameters are as shown below:

MLPClassifier(hidden\_layer\_sizes=50,

max\_iter=1400,

activation = ‘relu’,

solver = ‘adam’,

tol = 0.0001)

## 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 Model:

* Accuracy:

Training Data - 0.79

Test Data - 0.796

We can see that accuracy is at around 79% on training and test datasets. Accuracy in test dataset is slightly better than training dataset. The model does not seem to be overfitted for the training data.

* Confusion matrix:

Table-2.14 CART – Confusion matrix – Training data Table-2.15 CART – Confusion matrix –Testing data

From the data we can see that precision is higher than recall (for positive target variable) for both training and testing data. This might be desirable in this business scenario as the insurance company wouldn’t want to lose customers based on the model’s prediction that a customer might raise a claim.

* ROC\_AUC\_Score / ROC\_Curve:

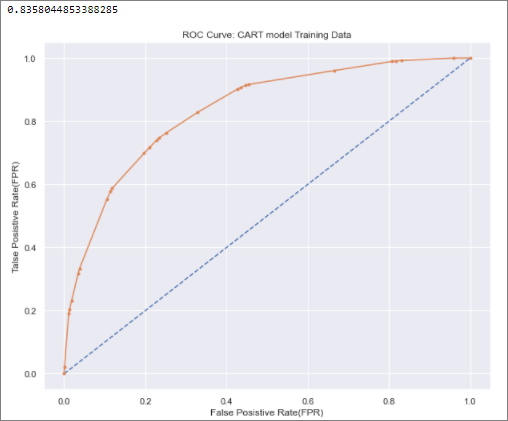
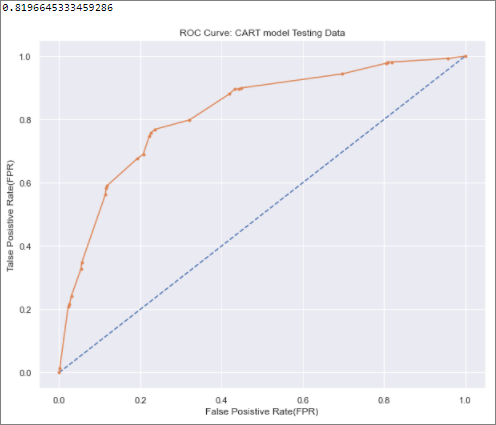
 

Figure-2.17 CART –ROC score/curve– Training data Figure-2.18 CART –ROC score/curve–Testing data

ROC score for training dataset is 0.836 against testing data 0.820, which is comparable. The ROC score is above 0.8 and will perform well at discriminating between target variable class.

* Classification report:

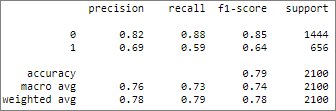
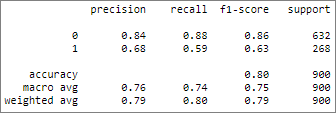
 

Table-2.16 CART –Classification Report– Training data Table-2.17 CART –Classification Report–Testing data

We can see that precision, recall ,f1-score and accuracy are around the same between testing and training datasets. Overall, the measures are high and hence the model is good.

#### Random Forest Model:

* Accuracy:

Training Data - 0.795

Test Data - 0.798

We can see that accuracy is at around 80% on training and test datasets. Accuracy in test dataset is slightly better than training dataset. The model does not seem to be overfitted for the training data.

* Confusion matrix:

Table-2.18 RF – Confusion matrix – Training data Table-2.19 RF – Confusion matrix –Testing data

From the data we can see that precision is higher than recall(for positive target variable) for both training and testing data. This might be desirable in this business scenario as the insurance company wouldn’t want to lose customers based on the model’s prediction that a customer might raise a claim.

* ROC\_AUC\_Score / ROC\_Curve:

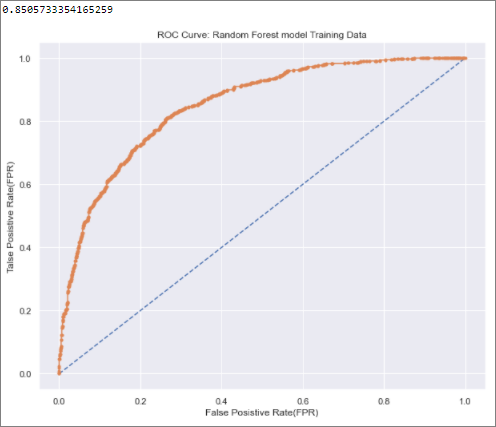
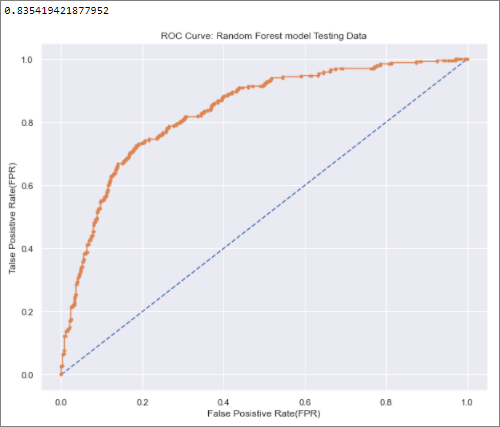
 

Figure-2.19 RF – ROC score/curve– Training data Figure-2.20 RF – ROC score/curve–Testing data

ROC score for training dataset is 0.851 against testing data 0.835, which is comparable. The score is high, and the model will perform well at discriminating target variable class.

* Classification report:

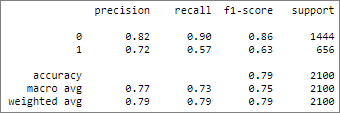
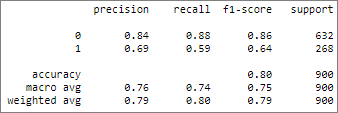
 

Table-2.20 RF – Classification Report– Training data Table-2.21 RF – Classification Report–Testing data

We can see that precision, recall ,f1-score and accuracy are around the same between testing and training datasets. Overall, the measures are high and hence the model is good.

#### Artificial Neural Network:

* Accuracy:

Training Data - 0.794

Test Data - 0.782

We can see that accuracy is at around 79% on training and test datasets. Accuracy in test dataset is slightly lower than training dataset. The model does not seem to be overfitted for the training data.

* Confusion matrix:

Table-2.22 ANN – Confusion matrix – Training data Table-2.23 ANN – Confusion matrix –Testing data

From the data we can see that precision is higher than recall (for positive target variable) for both training and testing data. This might be desirable in this business scenario as the insurance company wouldn’t want to lose customers based on the model’s prediction that a customer might raise a claim.

* ROC\_AUC\_Score / ROC\_Curve:

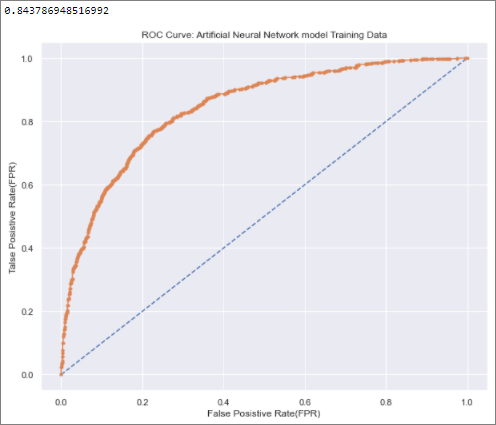
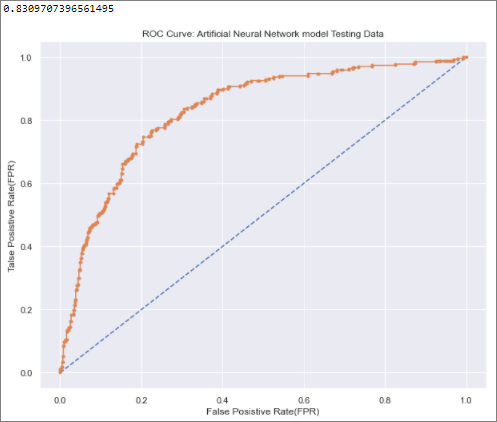
 

Figure-2.21 ANN – ROC score/curve– Training data Figure-2.22 ANN – ROC score/curve–Testing data

ROC score for training dataset is 0.844 against testing data 0.831, which is comparable. The score is high, and the model will perform well at discriminating target variable class.

* Classification report:

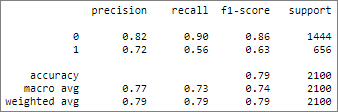
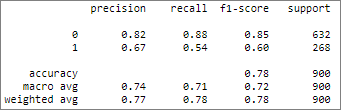
 

Table-2.24 ANN – Classification Report– Training data Table-2.25 ANN – Classification Report–Testing data

We can see that precision, recall ,f1-score and accuracy are around the same(within 5% points) between testing and training datasets. Overall, the measures are high and hence the model is good.

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

Let’s look at the below table which has the values of accuracies, precision, recall, roc\_auc\_score &

f1-score for both training and testing data sets across the 3 models:

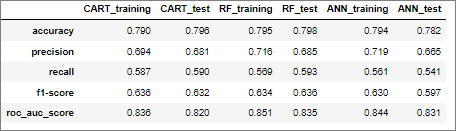


Table-2.26 Performance metrics across models

Although all the 3 models, have comparable metrics across the training and test data, we can see that the Random Forest model have the best performance metrics against the test data for all the metrics. The accuracy, precision, and roc\_auc\_score is highest for the Random Forest model on the training data compared to other models.

Hence we can consider the Random Forest model as the best among the 3 models compared for this dataset.

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

* More data is required to make informed decisions. Claims could be dependent on weather, travel mode, medical condition of the tourist, accident rates at the destination etc.
* 98% of all tours were booked via online channel, hence business needs to look for passing on promotional features on insurance to end customers more than dealing with agencies.
* Tours arranged by agency C2B has a very high claim rate of 61% (refer Figure 2.13). Business needs to investigate and change their strategy.
* Tour Products- Silver Plan & gold plan has very high claim rates of 71.66% and 64.22%. Business needs to investigate these products and change their strategy.
* More claims are raised at tours conducted by Airlines, even though the tours conducted by Travel Agencies are much higher. Business needs to investigate on the root cause.

## THE END