

Datamining PROJECT REPORT

DM



April 16, 2022

Brijal Panchal

Contents

[Problem 1: Clustering 2](#_Toc101013442)

[1.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 2](#_Toc101013443)

[1.2. Do you think scaling is necessary for clustering in this case? Justify 5](#_Toc101013444)

[1.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them 6](#_Toc101013449)

[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. 9](#_Toc101013467)

[1.5. Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. 11](#_Toc101013481)

[2.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis) 13](#_Toc101013493)

[2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest 17](#_Toc101013494)

[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. 18](#_Toc101013514)

[2.4 Final Model: Compare all the models and write an inference which model is best/optimized. 22](#_Toc101013545)

[2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations? 22](#_Toc101013548)

# Problem 1: Clustering

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

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

Sample of the dataset:

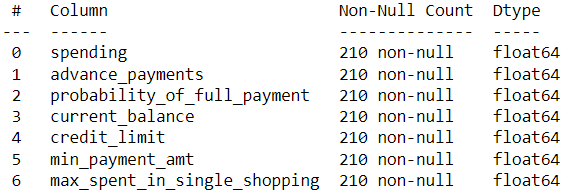


Data Discerption

* **spending**: Amount spent by the customer per month (in 1000s)
* **advance\_payments**: Amount paid by the customer in advance by cash (in 100s)
* **probability\_of\_full\_payment**: Probability of payment done in full by the customer to the bank
* **current\_balance**: Balance amount left in the account to make purchases (in 1000s)
* **credit\_limit**: Limit of the amount in credit card (10000s)
* **min\_payment\_amt** : minimum paid by the customer while making payments for purchases made monthly (in 100s)
* **max\_spent\_in\_single\_shopping**: Maximum amount spent in one purchase (in 1000s)

Exploratory Data Analysis

Let us check the types of variables in the data frame and missing values in dataset.



There are total 210 rows and 6 column in the dataset. all columns are continuous dataset having datatype float. There is no missing value and duplicates present in the dataset.

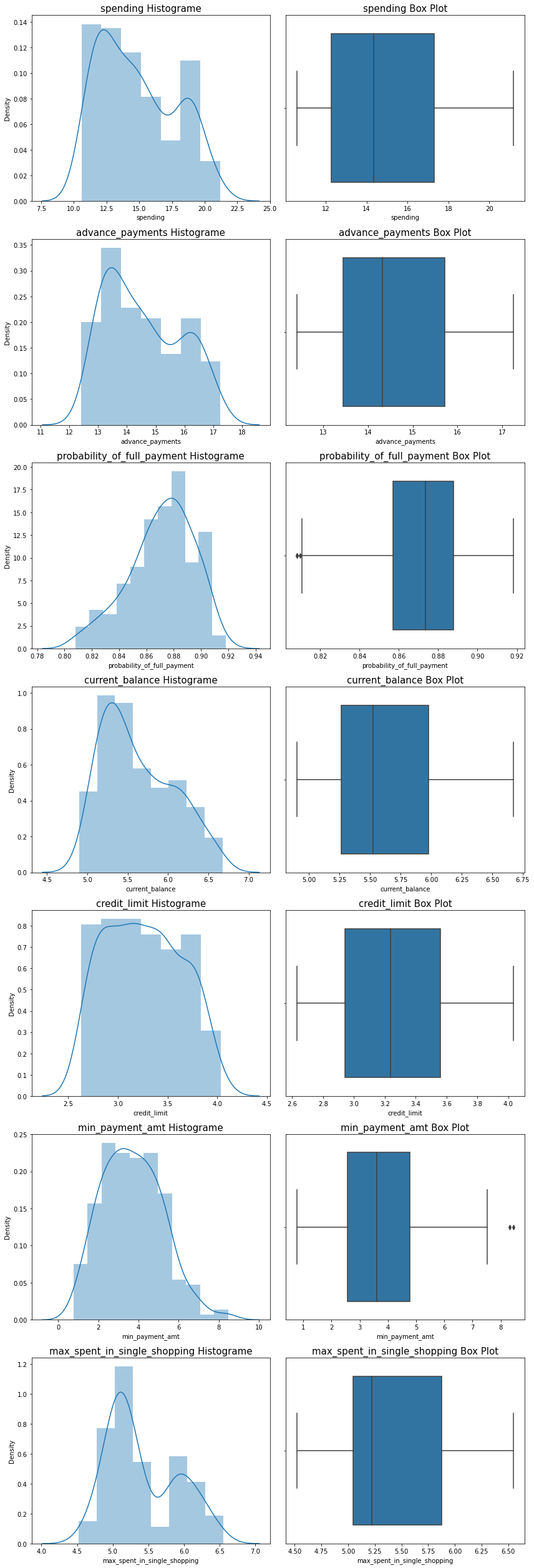
Summary of description

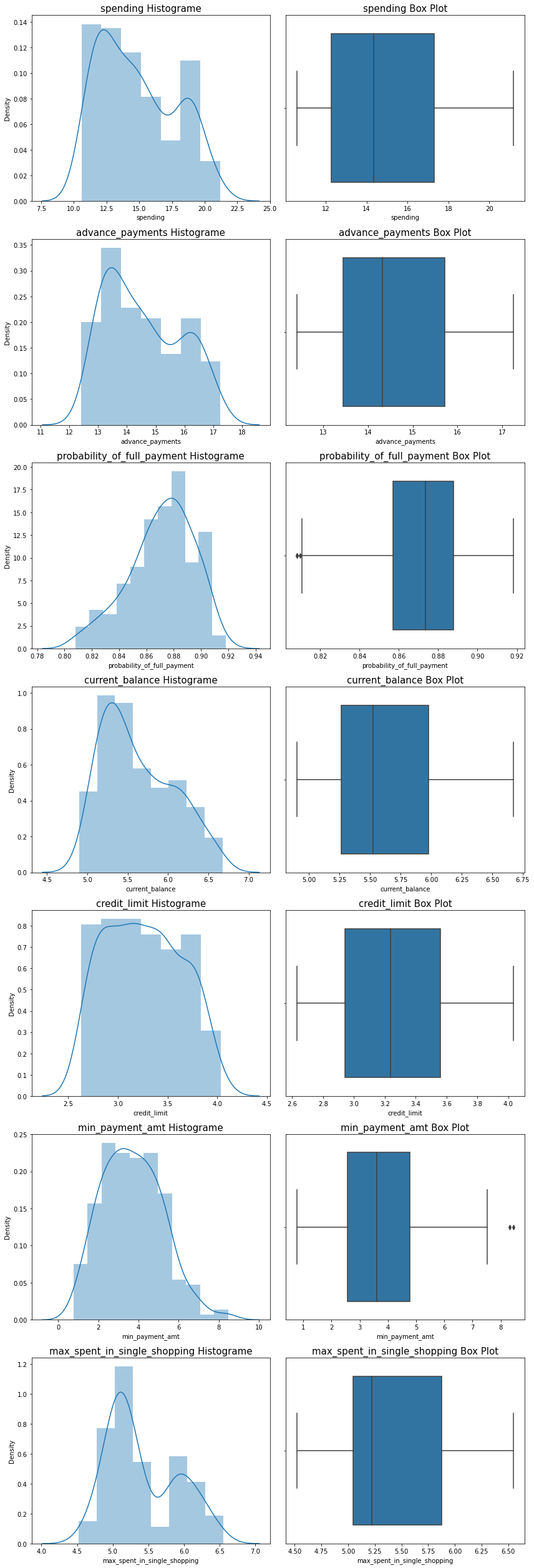


From description, we can visibly see that all columns have almost same mean and median value.

* Minimum spending from credit card per month is 10.59K rupees and maximum is 21.18K rupees while average spending per month is around 14.84K rupees
* Minimum amount paid by the customer in advance by cash is 1,241 rupees and maximum is 1,725 rupees while average spending per month is around 1,455.9 rupees
* Average Probability of payment done in full by the customer to the bank is around 87.09% where minimum is 80.8% and maximum is 91.83%
* Minimum balance amount left in the account to make purchases is 4,899 rupees and maximum is 6,672 rupees.
* Average credit card limit is 32,586 rupees where minimum is 26,300 rupees and maximum is 40,330 rupees
* Average minimum paid by the customer while making payments for purchases made monthly is 3,700 rupees where minimum is 765.1 rupees and maximum is 8,456 rupees
* Average maximum amount spent in one purchase is 5,408 rupees

Histogram and Box-Plot



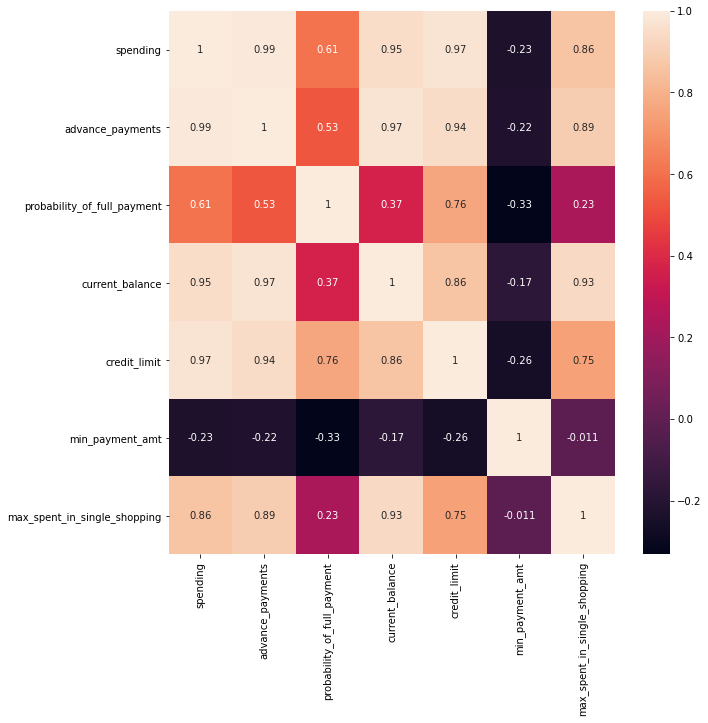


As we can see from histogram and box plot, probability\_of\_full\_payment and min\_payment\_amt has few outliers in the data. Spending, advance\_payments, current\_balance, max\_spent\_in\_single\_shopping are right skewed while probability\_of\_full\_payment is left skewed and credit\_limit , min\_payment\_amt are normally distributed.

Correlation matrix



Heat Map



As per heat-map, we can see that advance\_payments is highly correlated with spending, similarly current\_balance and credit\_limit is highly correlated with spending and advance\_payments. Max\_spent\_in\_single\_shopping is highly correlated with current\_balance.

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

**Summery description before scaling**:



As we can see from summery of description, scale of each variables are different as well as some variables have different units. Standardization is necessary in cluster analysis because groups are defined based on the distance between points in mathematical space.

**Summery description after scaling**:



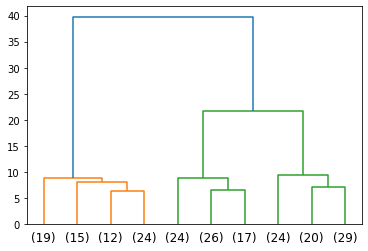
* 1. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

Step 1: Linkage

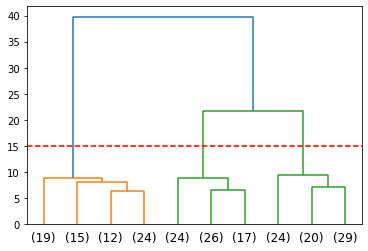
* **Ward** variance minimization algorithm has been used for calculating distance between points in mathematical space to identify the cluster in linkage. Ward’s method tends to create compact clusters of small size. It is a least squares method, so implicitly assumes a Gaussian model.

Step 2: Dendogram

* Drawing dendogram on ward linkage



* The x-axis contains the samples and y-axis represents the distance between these samples. The vertical line with maximum distance is the blue line and hence we can decide a threshold of 15 and cut the dendrogram
* Dendogram after cut-off the data at 15



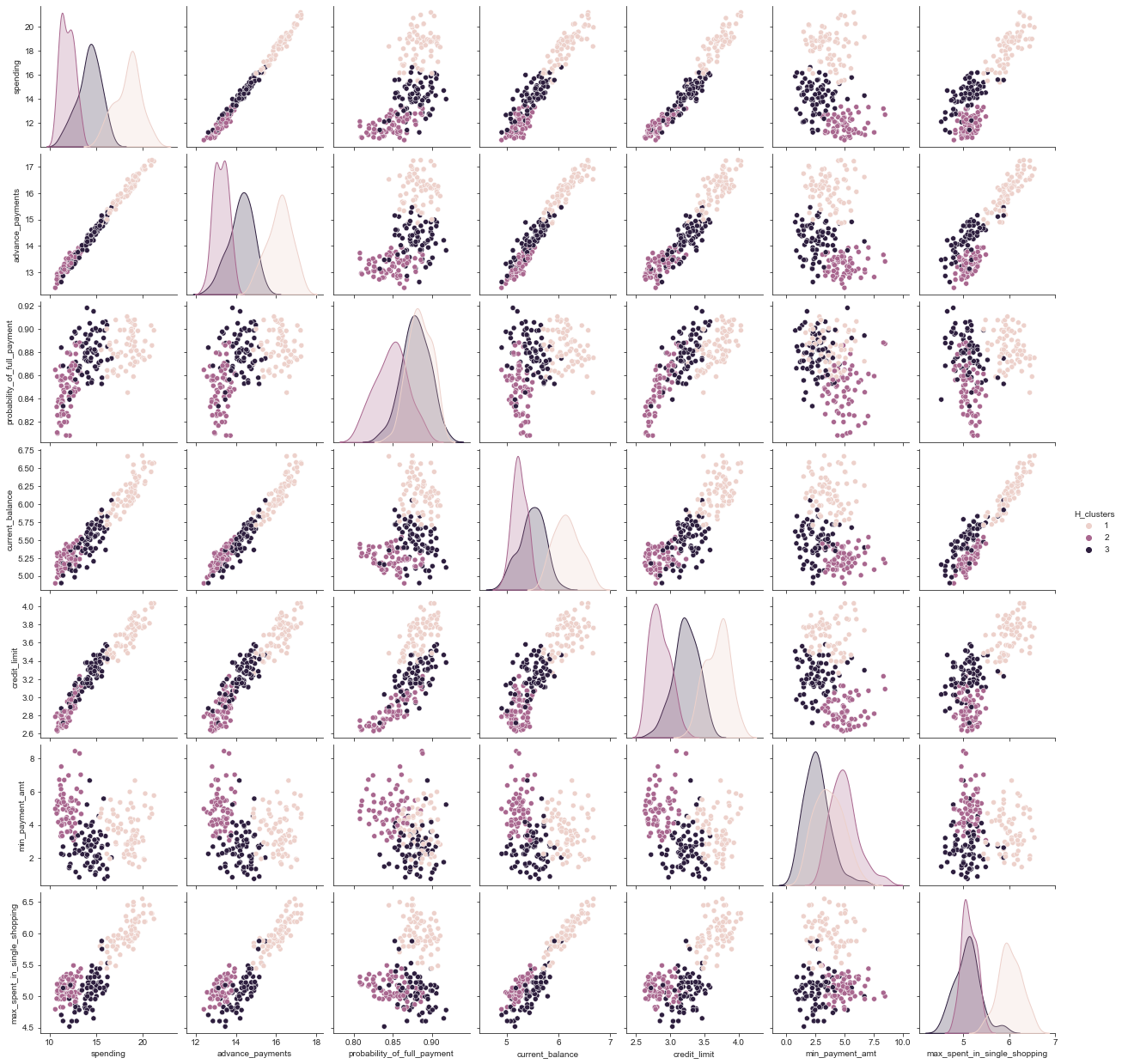
* We can see after cut-off, 3 clusters has been formed One cluster from (19) to (24), Second cluster from (24) to (17), finally (24) to (29) makes total 3 cluster,

Step 3: Hierarchical clustering

* **Fclusters** has been used on scaled data for hierarchical clustering on ward linkage data
* **Maxclust** criterion has used in which 3 max cluster count is assign to form
* **DataFrame** after applying cluster form in which 1,2,3 represent different clusters group



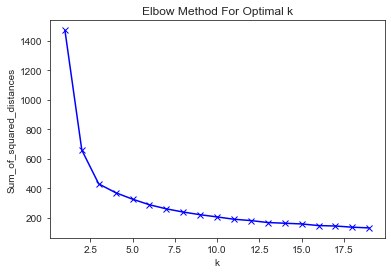
* **visualize** the Three clusters



We can clearly visualise the three cluster in all parameters.

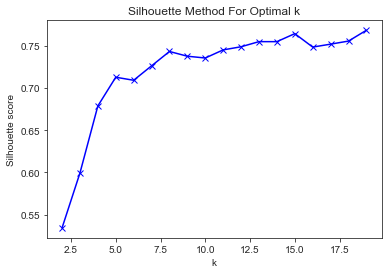
* 1. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve. Explain the results properly. Interpret and write inferences on the finalized clusters.
* K-mean Inertia after applying Kmean clustering for 2 cluster is 659.1717544870411
* K-mean inertia after applying Kmean clustering for 3 cluster is 430.65897315130064

Elbow Curve



In the plot above the elbow is at k=6 or k=7 indicating the optimal k for this dataset is 6 (which we can safely assume) after which distance is almost converged which is in agreement to the hierarchical clustering!

Silhouetted Method

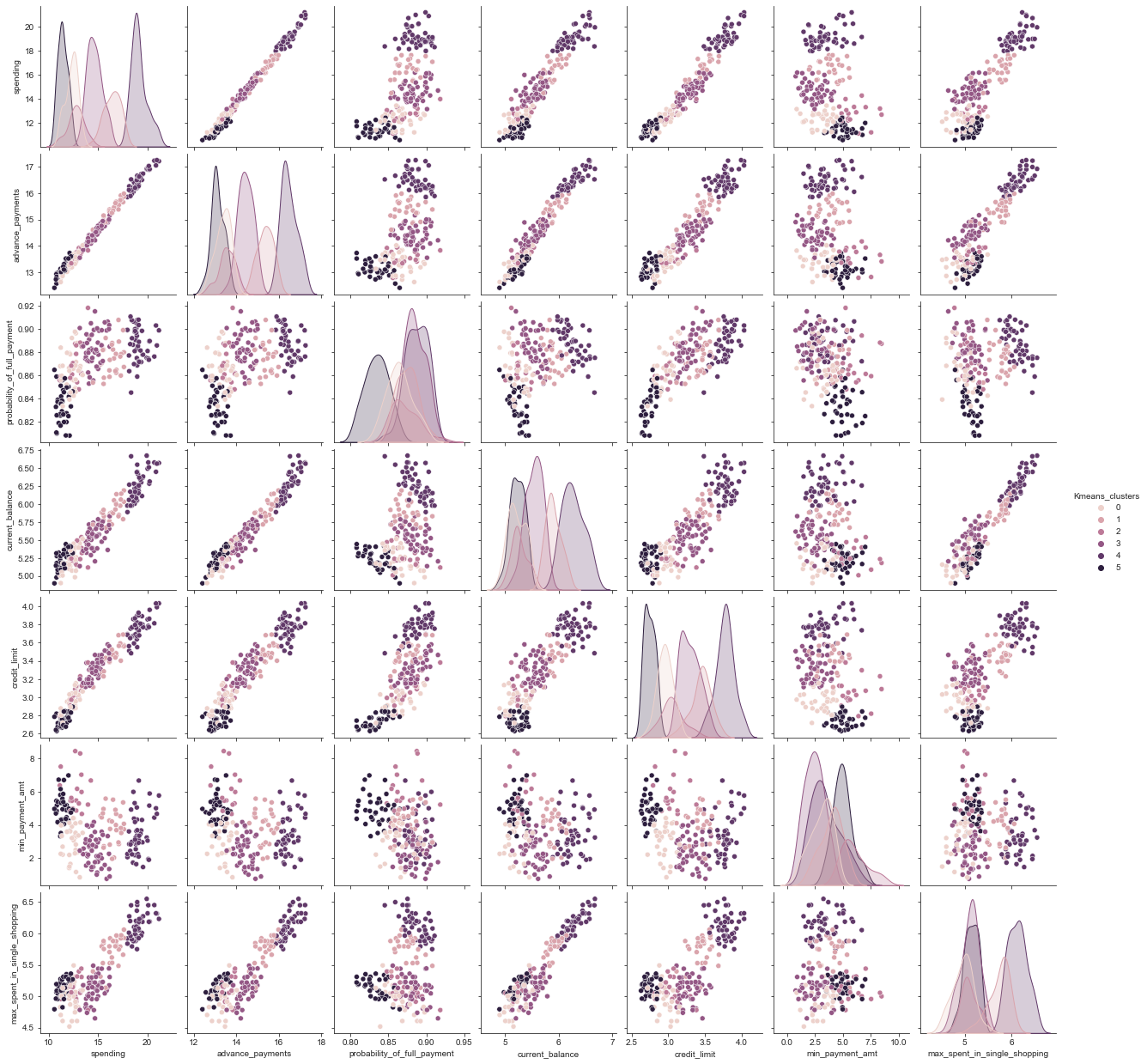


Here also we see that k=6 is the optimum number of clusters after which silhouette score is converged! So our finalised number of clusters are 6.

DataFrame after applying number of cluster as 6 in Kmean cluster method.



Visualize the six clusters



We can clearly visualise the six cluster in all parameters.

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

By use of K-mean cluster group bank can recommend bellow group

* Cluster 1 describes group having average in spending(1000) is 12.316875000000001 , advance payments(100) is 13.3759375 , probability of full payment is 0.8645968750000002 , current balance(1000) is 5.219281249999998 ,credit limit(10000) is 2.9548124999999996 , min payment amount(100) is 3.020253125000001 and max spent in single shopping(1000) is 4.9608750000000015 .
* Cluster 2 describes group having average in spending(1000) is 16.43428571428571 , advance payments(100) is 15.357857142857142 , probability of full payment is 0.8750785714285717 , current balance(1000) is 5.899714285714283 ,credit limit(10000) is 3.449035714285714 , min payment amount(100) is 3.958964285714285 and max spent in single shopping(1000) is 5.733785714285714 .
* Cluster 3 describes group having average in spending(1000) is 12.759444444444446 , advance payments(100) is 13.550555555555555 , probability of full payment is 0.8723055555555556 , current balance(1000) is 5.266555555555556 ,credit limit(10000) is 3.052777777777778 , min payment amount(100) is 6.031611111111111 and max spent in single shopping(1000) is 5.105500000000001 .
* Cluster 4 describes group having average in spending(1000) is 14.650624999999998 , advance payments(100) is 14.430833333333332 , probability of full payment is 0.88368125 , current balance(1000) is 5.5390625 ,credit limit(10000) is 3.297520833333334 , min payment amount(100) is 2.4228145833333334 and max spent in single shopping(1000) is 5.112729166666667 .
* Cluster 5 describes group having average in spending(1000) is 19.151041666666664 , advance payments(100) is 16.469166666666663 , probability of full payment is 0.8870895833333335 , current balance(1000) is 6.268854166666666 ,credit limit(10000) is 3.7729375000000007 , min payment amount(100) is 3.460416666666666 and max spent in single shopping(1000) is 6.12725 .
* Cluster 6 describes group having average in spending(1000) is 11.431388888888888 , advance payments(100) is 13.119166666666668 , probability of full payment is 0.8344972222222222 , current balance(1000) is 5.227916666666667 ,credit limit(10000) is 2.7457777777777768 , min payment amount(100) is 4.960527777777777 and max spent in single shopping(1000) is 5.138416666666668 .

By use of Hierarchical cluster group bank can recommend bellow group

* Cluster 1 describes group having average in spending(1000) is 18.371428571428567 , advance payments(100) is 16.14542857142857 , probability of full payment is 0.8844 , current balance(1000) is 6.15817142857143 ,credit limit(10000) is 3.684628571428572 , min payment amount(100) is 3.6391571428571434 and max spent in single shopping(1000) is 6.017371428571428 .
* Cluster 2 describes group having average in spending(1000) is 11.872388059701493 , advance payments(100) is 13.257014925373136 , probability of full payment is 0.8480716417910452 , current balance(1000) is 5.238940298507461 ,credit limit(10000) is 2.8485373134328356 , min payment amount(100) is 4.949432835820894 and max spent in single shopping(1000) is 5.122208955223879 .
* Cluster 3 describes group having average in spending(1000) is 14.199041095890408 , advance payments(100) is 14.233561643835612 , probability of full payment is 0.8791904109589038 , current balance(1000) is 5.478232876712326 ,credit limit(10000) is 3.22645205479452 , min payment amount(100) is 2.6121808219178084 and max spent in single shopping(1000) is 5.086178082191782 .

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

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

Sample of the dataset:

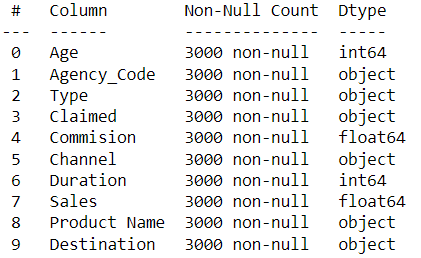


Data Discerption

* Age: Age of insured
* Agency\_Code: Code of tour firm
* Type: Type of tour insurance firms
* Claimed: Target: Claim Status
* Commission: The commission received for tour insurance firm(in percentage of sales)
* Channel: Distribution channel of tour insurance agencies
* Duration: Duration of the tour(days)
* Sales: Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)
* ProductName: Name of the tour insurance products
* Destination: Destination of the tour

Exploratory Data Analysis

Let us check the types of variables in the data frame and missing values in dataset.



There are total 3000 rows and 9 column in the dataset. Age and Duration column are integer, Commision and Sales are float while rest are object data type. There is 139 duplicate values present in data.

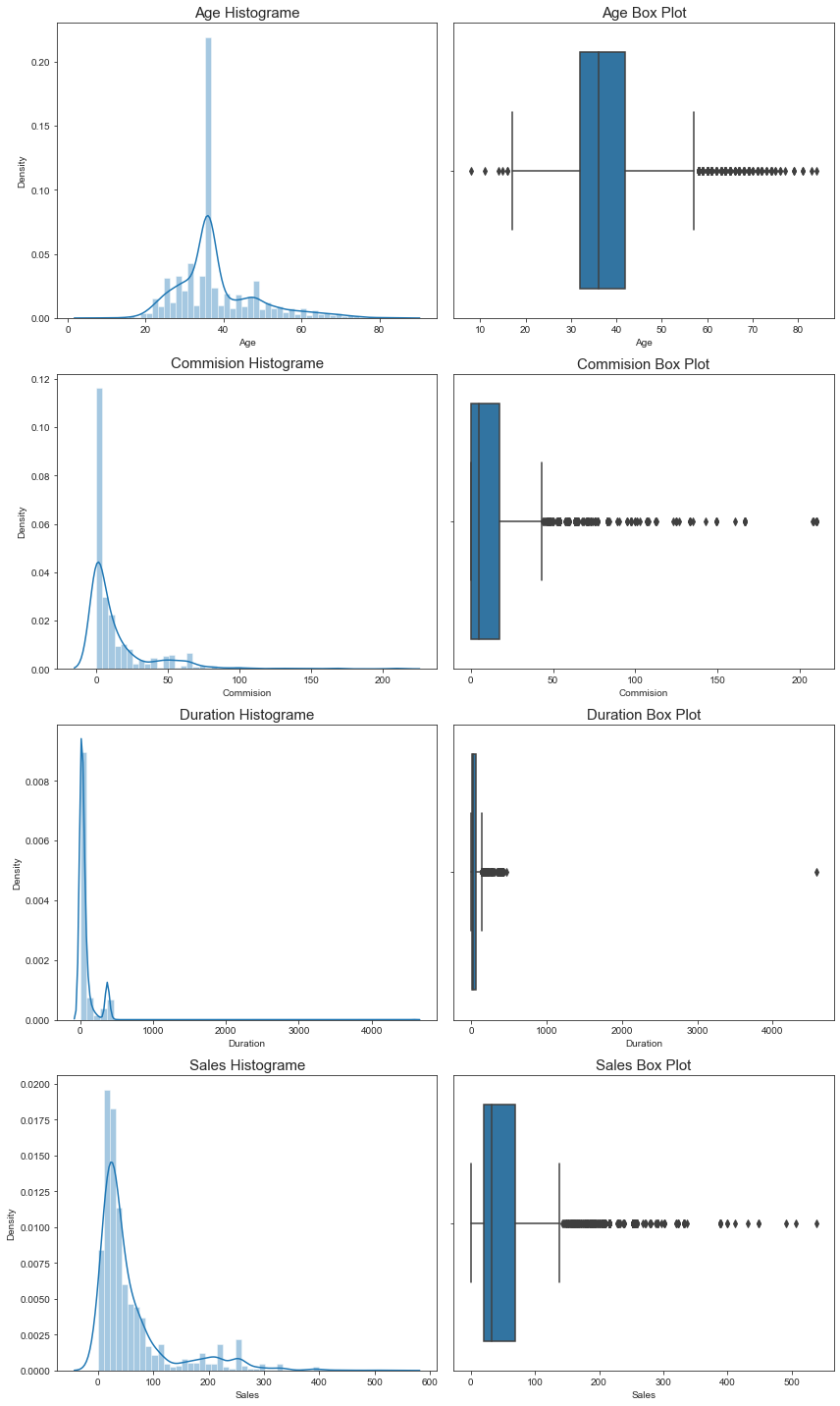
Summary of description

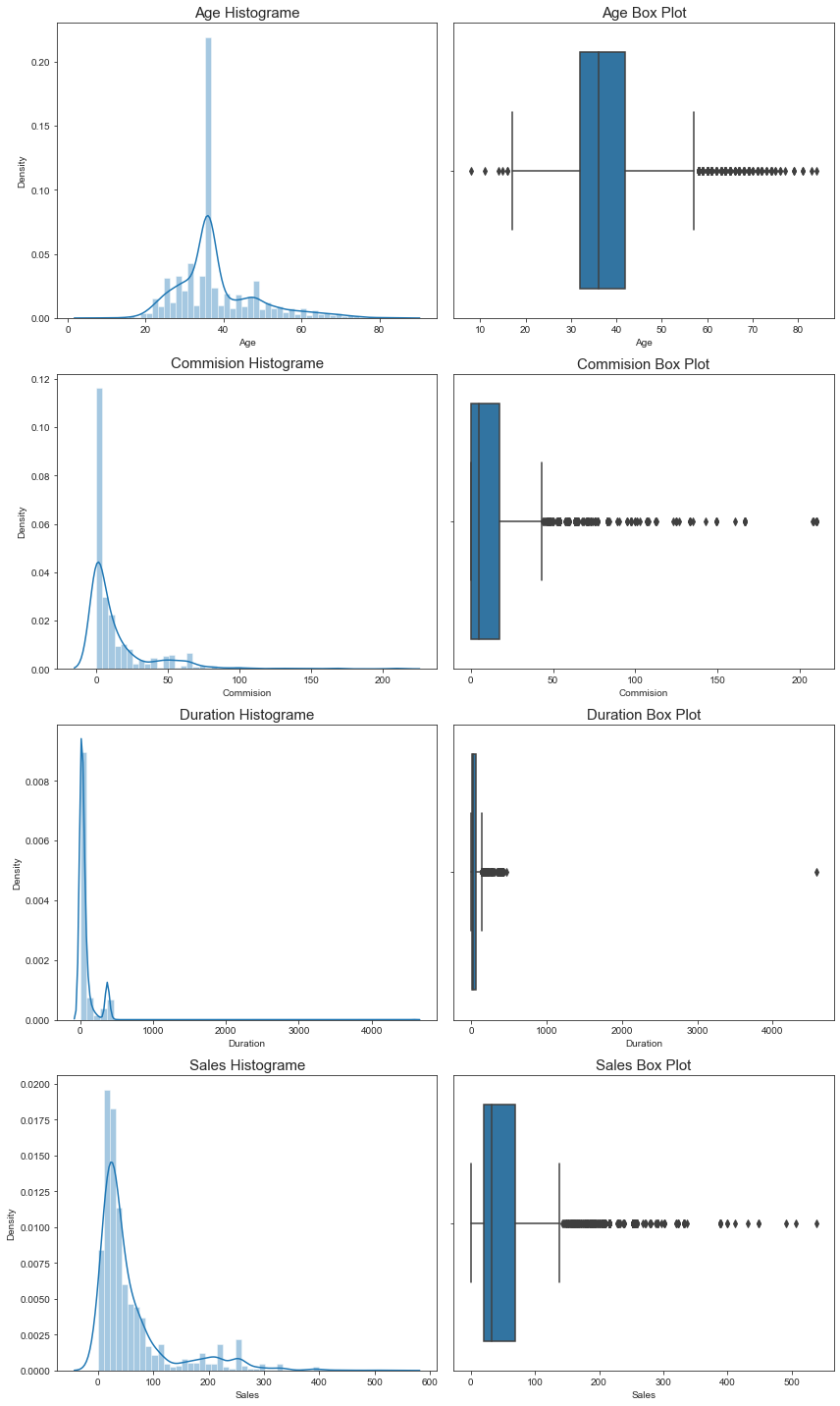


From description, we can visibly see that all columns have different mean and median value.

* Minimum age for insurance is 8 and maximum is 84 while average age is around 38. Standard deviation is 10.46.
* Minimum commission for insurance is 0 and maximum is 210.21 while average commission is around 14.52. Standard deviation is 25.48
* Minimum duration for insurance is -1 and maximum is 4580 while average duration is around 70. Standard deviation is 134.053. Duration is -1 which can’t be possible. Replacing duration as 0 where it is -1.
* Minimum sales for insurance is 0 and maximum is 539 while average commission is around 60.24. Standard deviation is 70

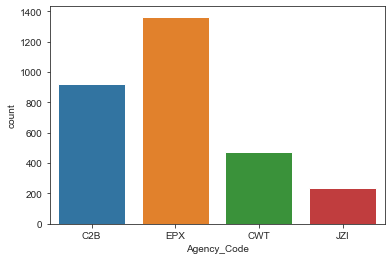
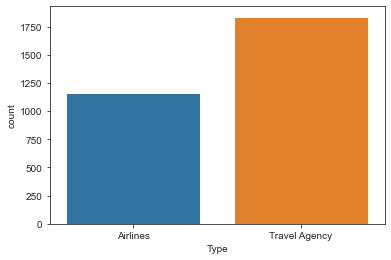
Histogram and Box-Plot

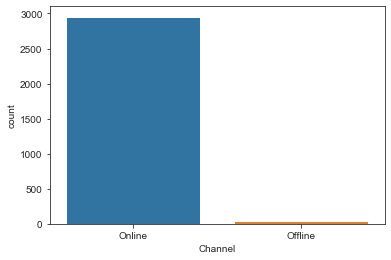
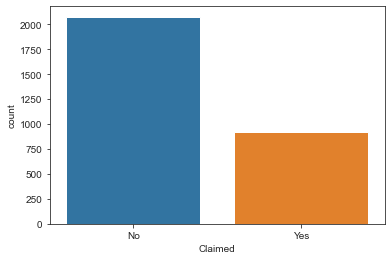


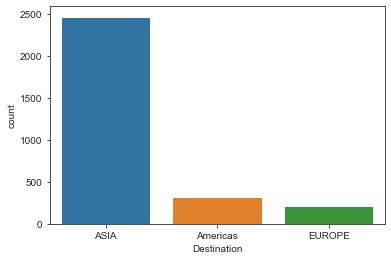
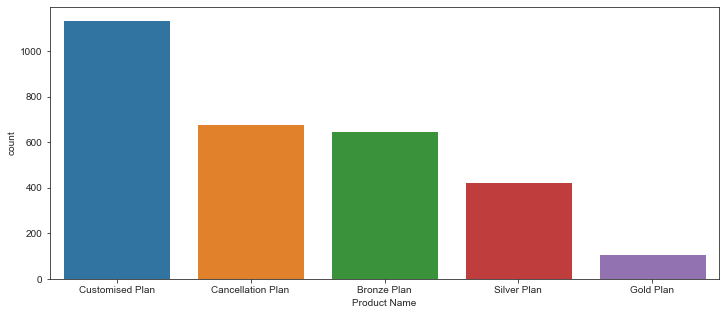


As we can see from histogram and box plot all are left skewed as well as have outliers in data.

Bar Plot

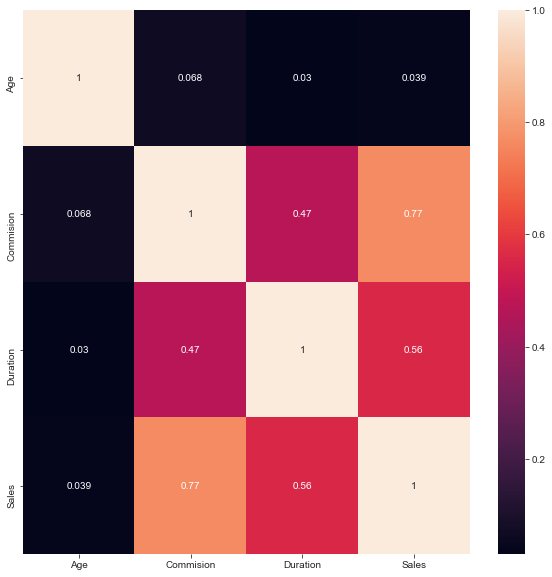




Correlation matrix



Heat Map



As per heat-map, we can see that sales and commission are highly correlated with each other.

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

Split data into training (70%) and test set (30%) and prepared the mode.

CART:

* After applying GridSearchCV, parameters for best CART model are

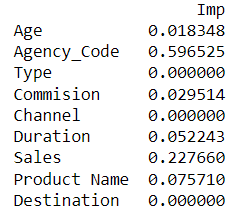
{'criterion': 'gini',

'max\_depth': 4,

'min\_samples\_leaf': 1,

'min\_samples\_split': 12}

* Train accuracy for above parameter generated CART model is 0.81%
* Test accuracy for above parameter generated CART model is 0.77%
* Gini index of all variables for above parameter generated CART model



Random Forest:

* After applying GridSearchCV, parameters for best Random forest model are

{'max\_depth': 5,

'max\_features': 6,

'min\_samples\_leaf': 50,

'min\_samples\_split': 20,

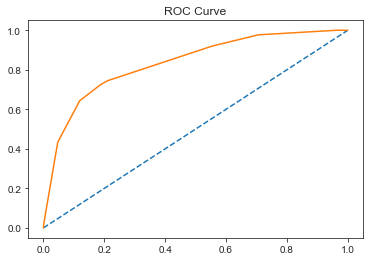
'n\_estimators': 701}

* Train accuracy for above parameter generated CART model is 0.80%
* Test accuracy for above parameter generated CART model is 0.75%

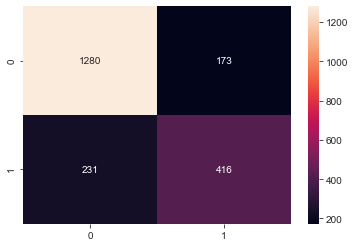
2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.

* CART:
* Train:

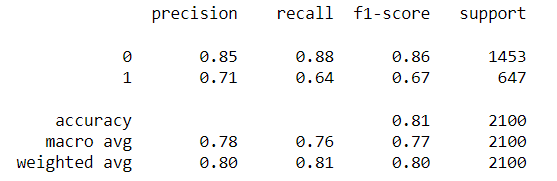
1. Accuracy : 0.807
2. AUC Train dataset: 0.834



1. Confusion matrix

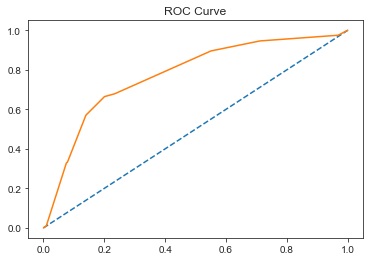


1. Classification report

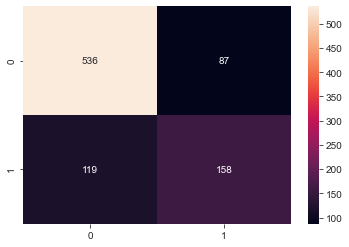


* Test:

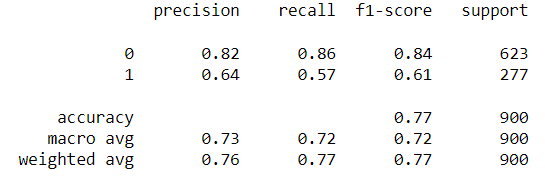
1. Accuracy : 0.771
2. AUC Train dataset: 0.775



1. Confusion matrix

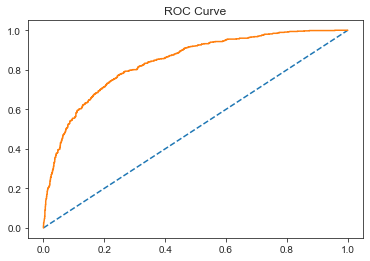


1. Classification report

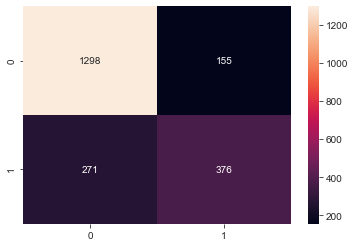


* Random Forest:
* Train:

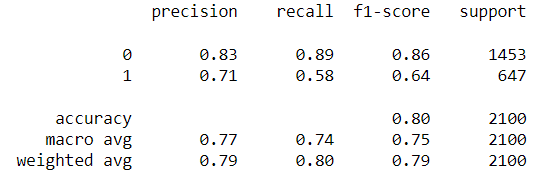
1. Accuracy : 0.797
2. AUC Train dataset: 0.839



1. Confusion matrix

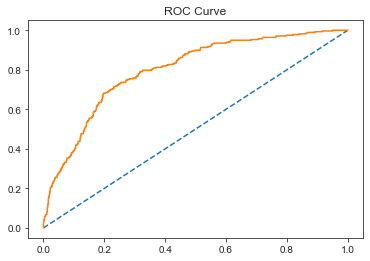


1. Classification report

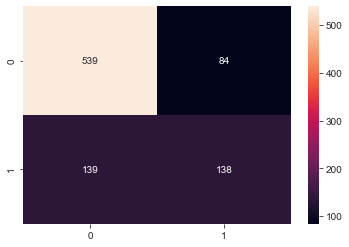


* Test:

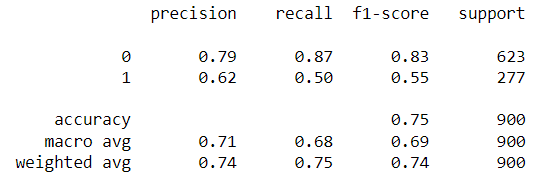
1. Accuracy : 0.752
2. AUC Train dataset: 0.799



1. Confusion matrix



1. Classification report



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



By comparing Accuracy and AUC of CART and Random forest model for train and test data Random forest is best model as it has better AUC then CART.

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

* By looking at the insurance data and when we find correlation of different variables such as day of the incident, time, age group, etc, we can conclude that more data is required
* Streamlining online experiences benefitted customers, leading to an increase in conversions, which subsequently raised profits.
* As per the data 90% of insurance is done by online channel and almost all the offline business has a claimed associated
* Also based on the model we are getting 80% accuracy, so we need customer books airline tickets or plans, cross sell the insurance based on the claim data pattern