DATA MINING PROJECT

Chart, funnel chart

Description automatically generated

Prepared by Tribid Maji

CONTENT

**Problem 1: Clustering**

* 1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).
  2. Do you think scaling is necessary for clustering in this case? Justify
  3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them
  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.
  5. Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

**Problem 2: CART-RF-ANN**

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

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

**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.

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

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

**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.

Dataset for Problem 1: [bank\_marketing\_part1\_Data.csv](https://olympus.greatlearning.in/courses/32928/files/2545009/download?verifier=S3J9GJEqixM1ahRxArYqyTODbCIXjzjKlWYz7EfN&wrap=1)

**Data** **Dictionary** **for** **Market** **Segmentation:**

1. spending: Amount spent by the customer per month (in 1000s)
2. advance\_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability\_of\_full\_payment: Probability of payment done in full by the customer to the bank
4. current balance: Balance amount left in the account to make purchases (in 1000s)
5. credit\_limit: Limit of the amount in credit card (10000s)
6. min\_payment\_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. max\_spent\_in\_single\_shopping: Maximum amount spent in one purchase (in 1000s)
   1. **Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).**

**Ans:**

So, we will import all the necessary libraries for cluster analysis,

Import NumPy as np

Import pandas as pd

Import matplotlib.pyplot as plt

Import seaborn as sns

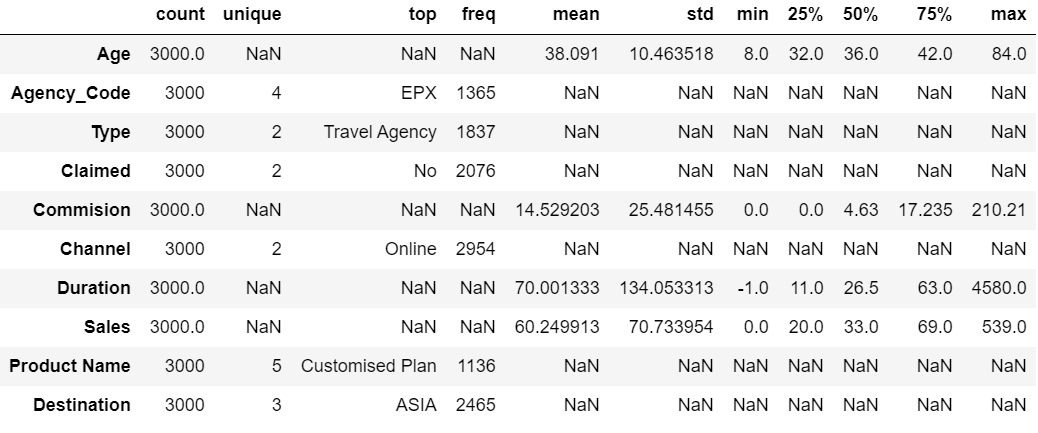
From sklearn.cluster import KMeans

From sklearn.metrics import silhouette samples, silhouette score

Reading the data,



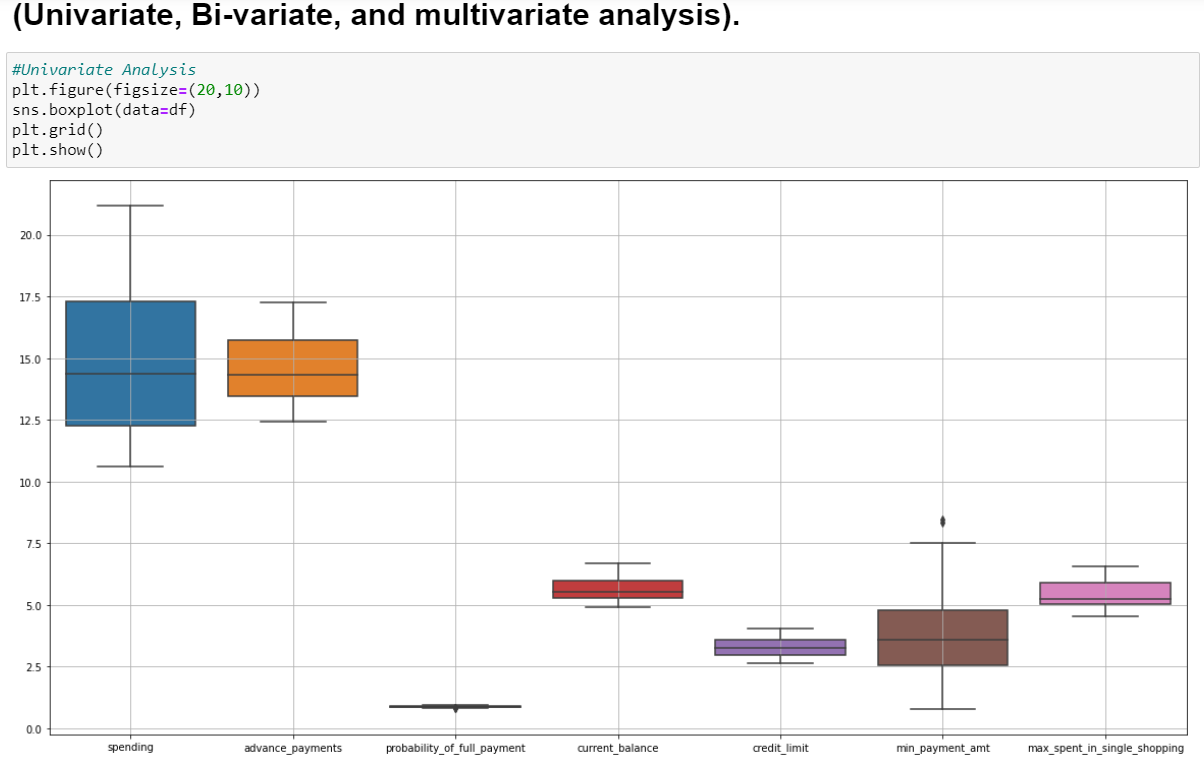
* The data seems to be perfect
* The shape of the data is (210, 7)
* The info of the data indicates that all values are float
* No Null values in the data
* No missing values in the data
* **Description of the Data**



* No null values present in any variables.
* The mean and median values seems to be almost equal.
* The standard deviation for spending is high when compared to other variables.
* No duplicates in the dataset

**Exploratory Data Analysis**

**Univariate / Bivariate analysis**

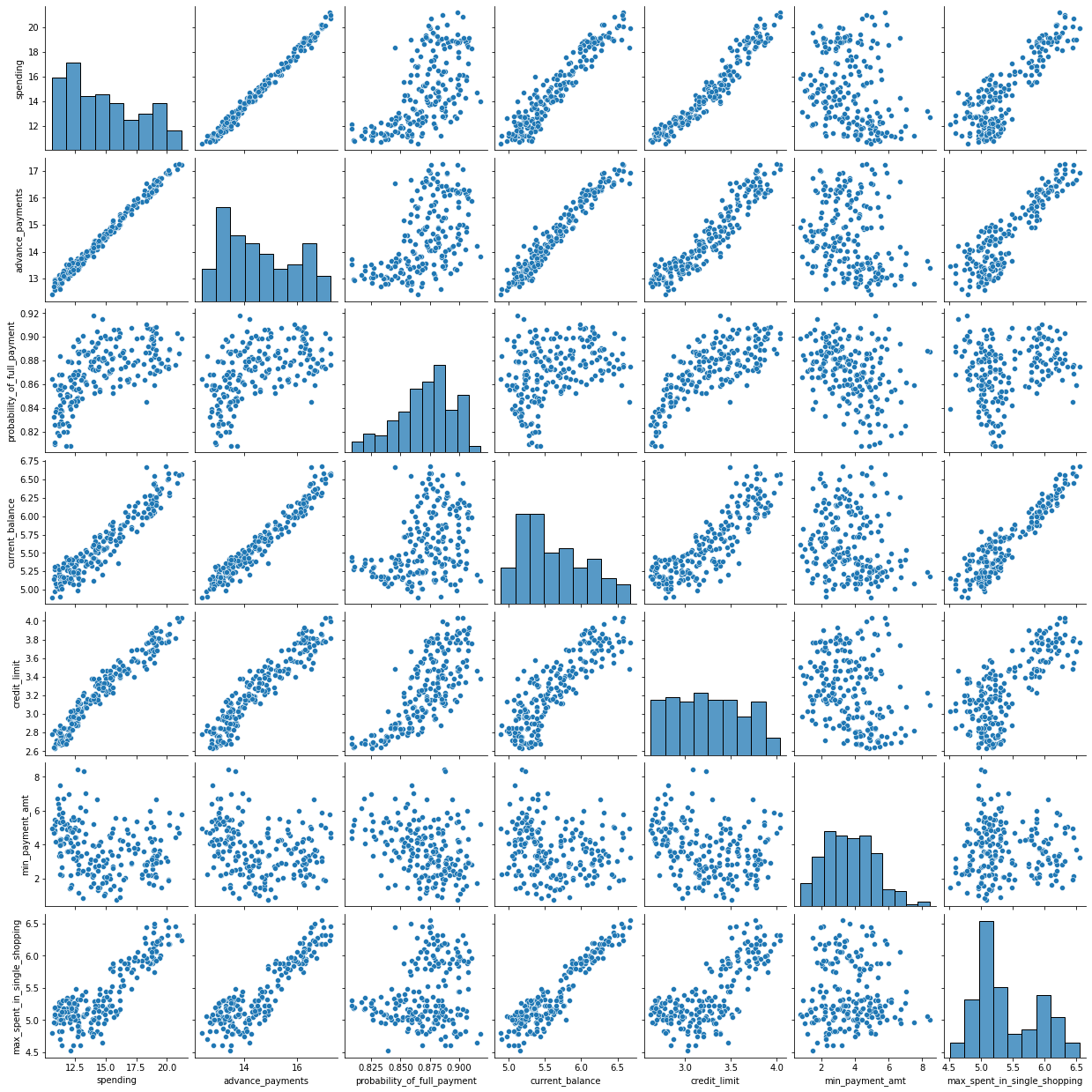


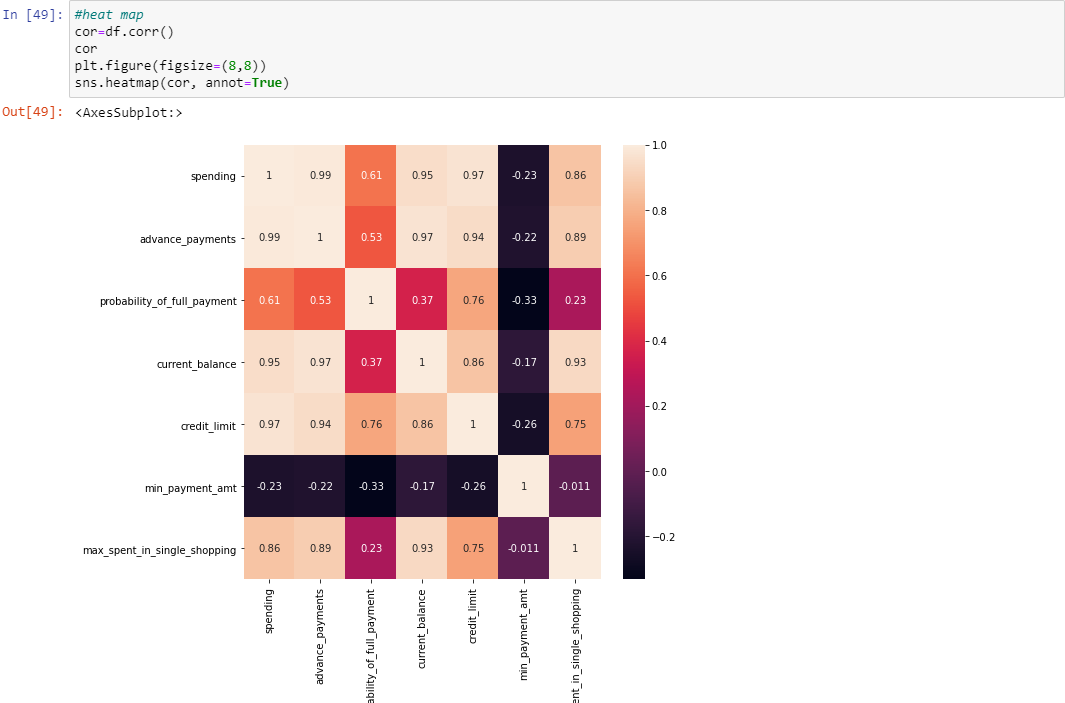
# The box plot of the variable shows no outliers:Spending,Advace Payment,Current Balance, Credit Limit,

# Max Spent in single shopping

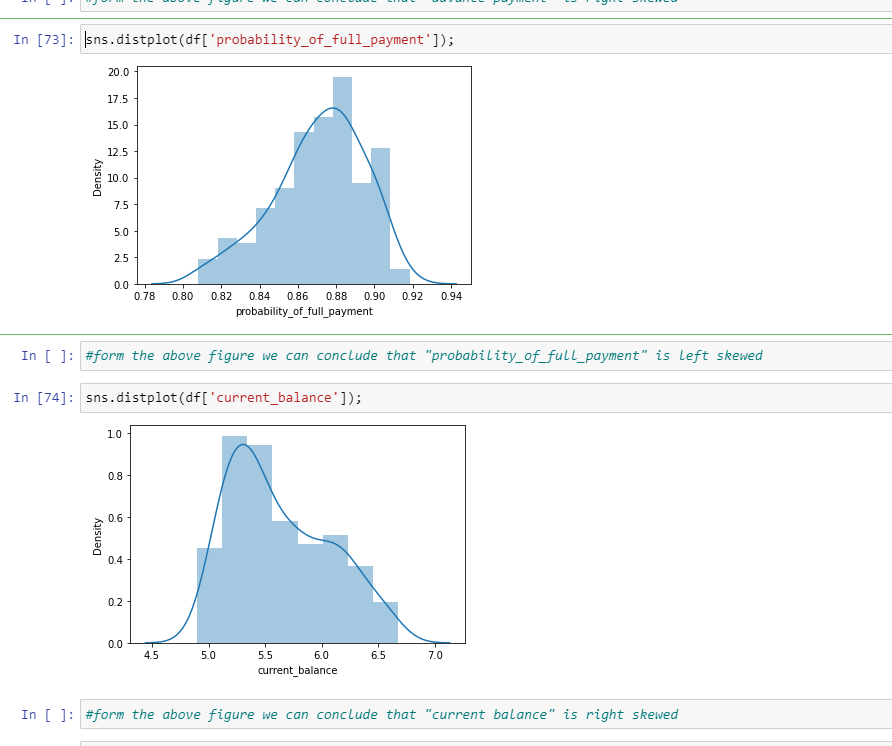
#The box plot of the variable show outliers :Min Payment amount & Probabilty of full payment

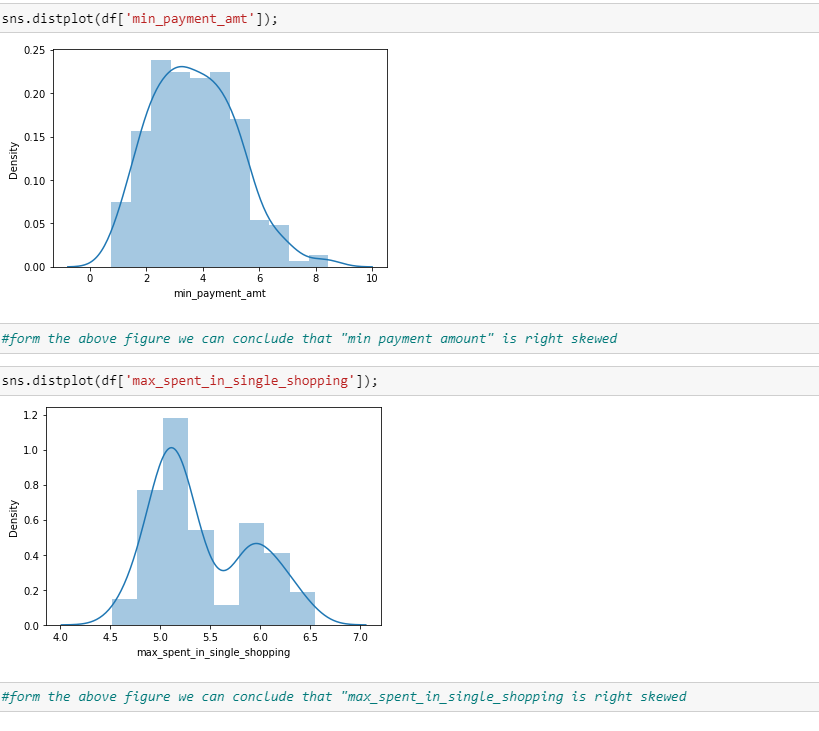
Pair-Plot











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

Ans :

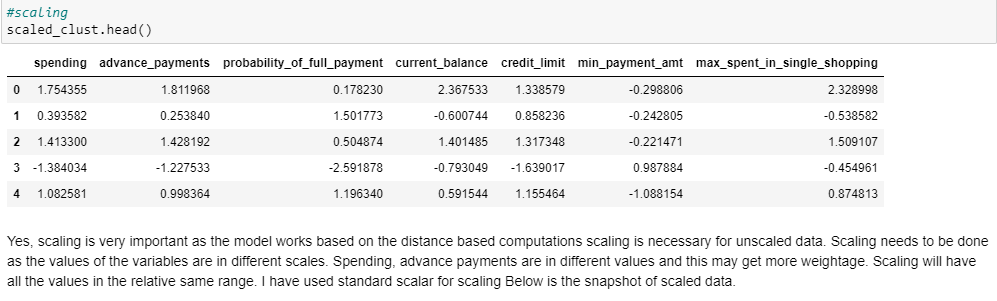
Yes, scaling is very important as the model works based on the distance based computations scaling is necessary for unscaled data.

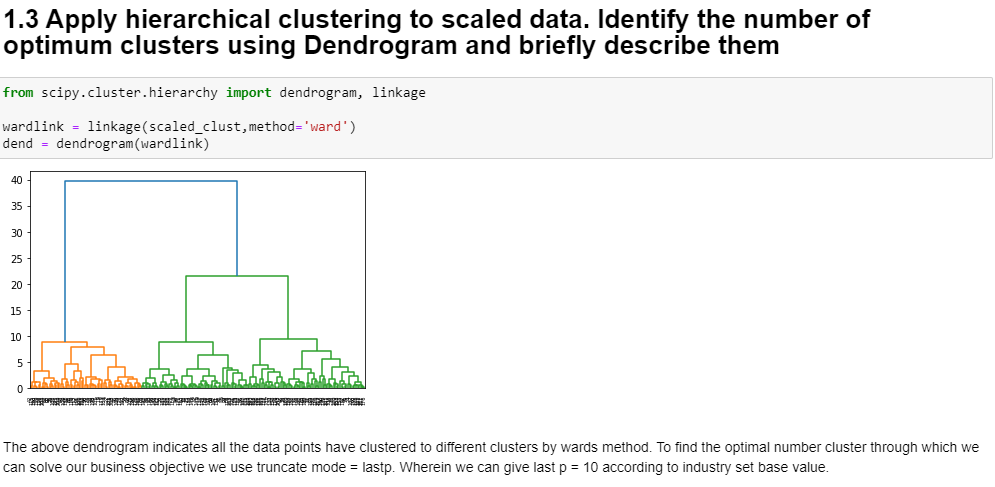
Scaling needs to be done as the values of the variables are in different scales.

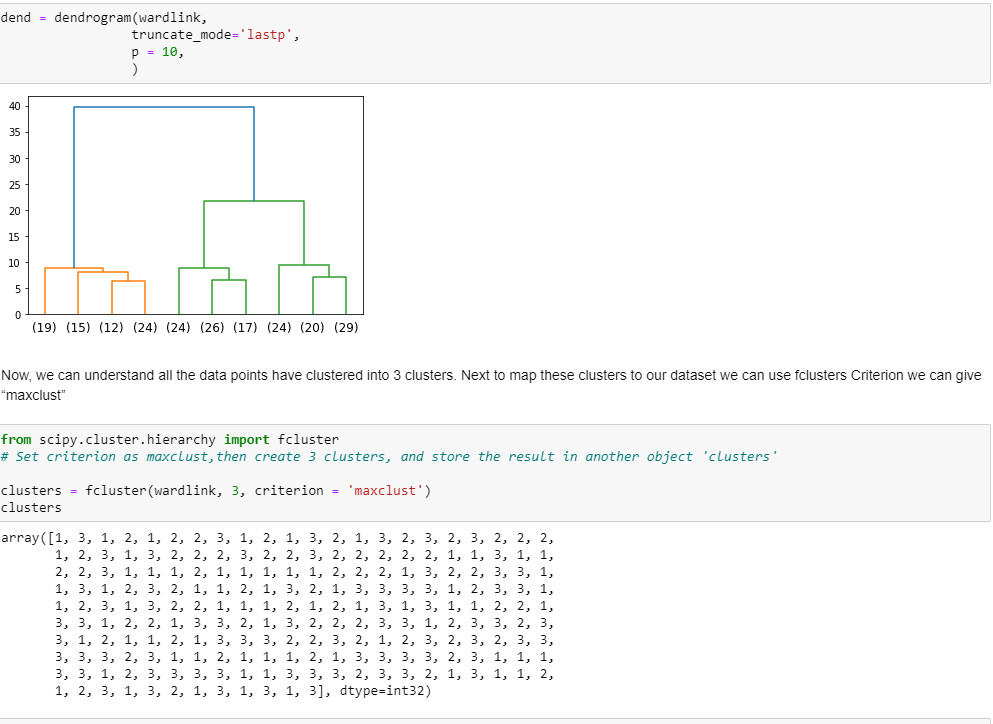
Spending, advance payments are in different values and this may get more weightage.

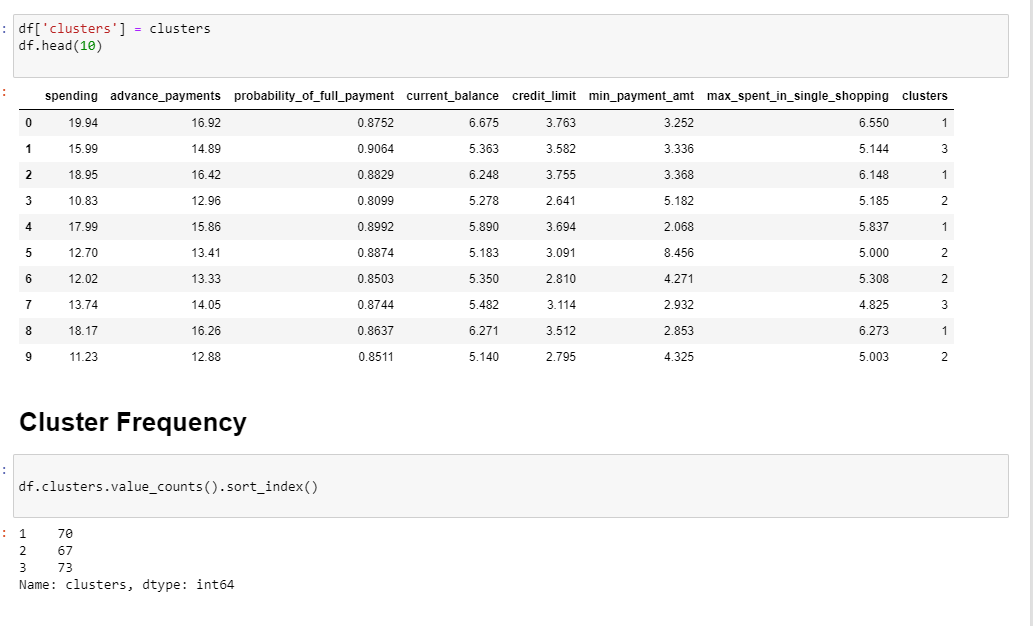
Scaling will have all the values in the relative same range.

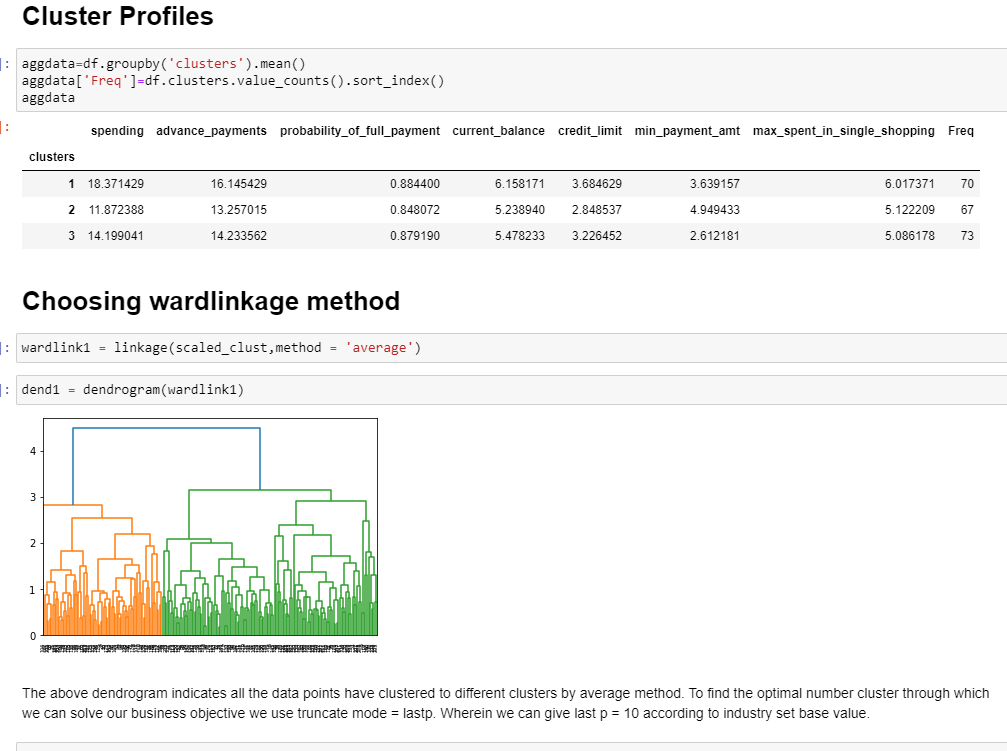
I have used standard scalar for scaling .Below is the snapshot of scaled data**.**

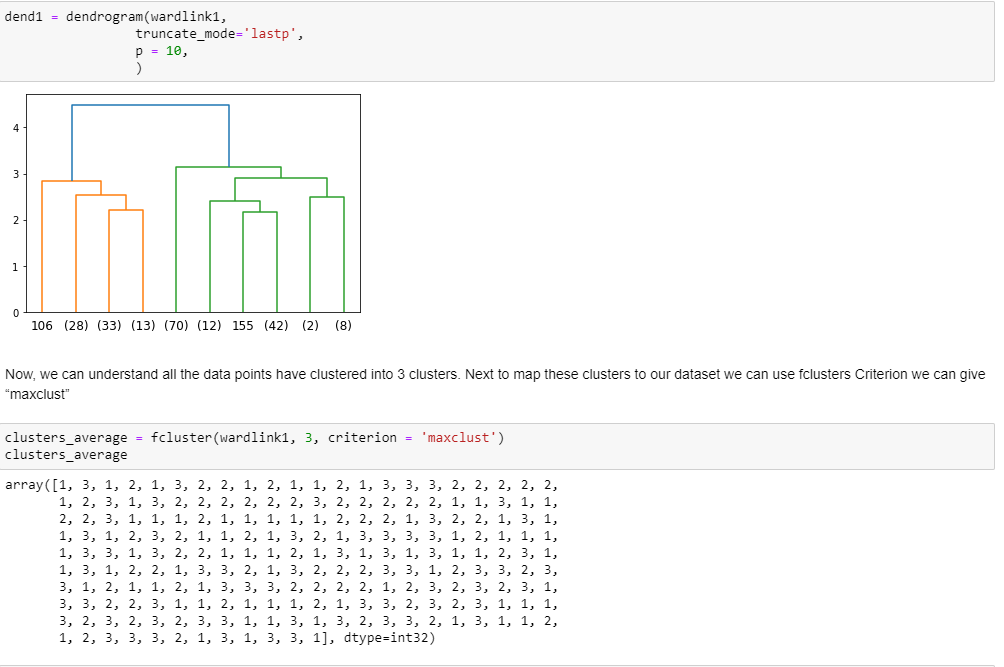


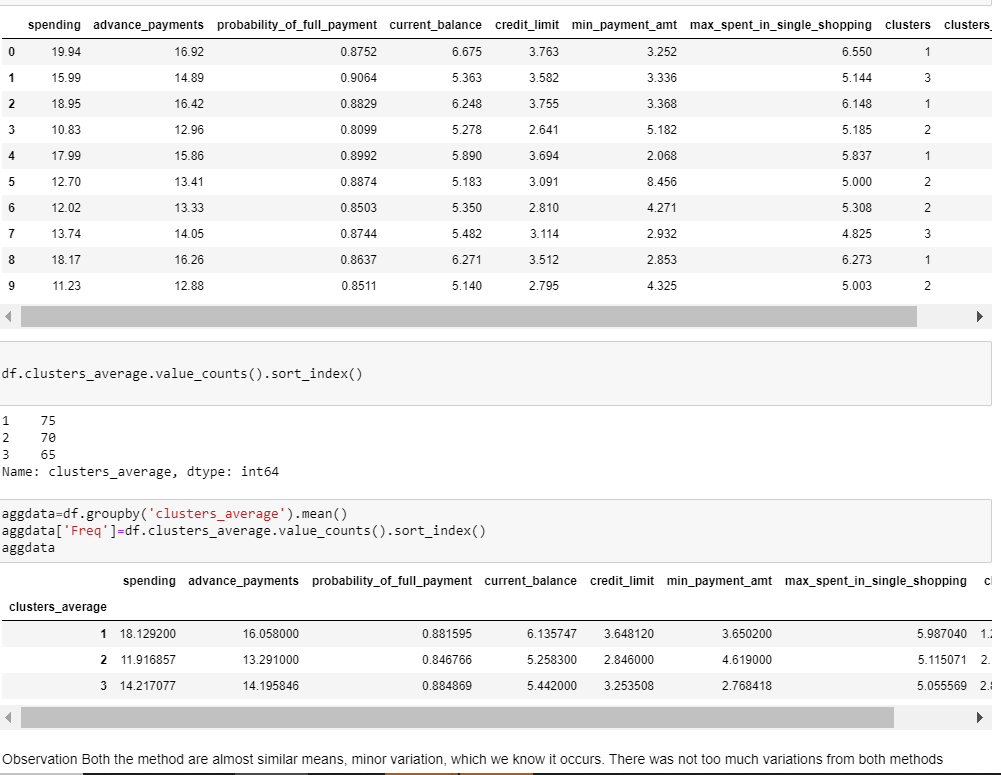










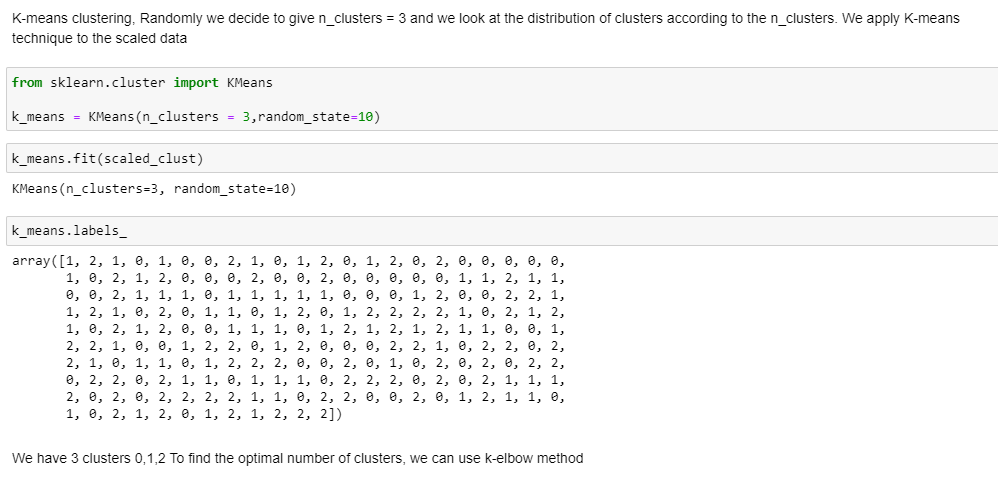


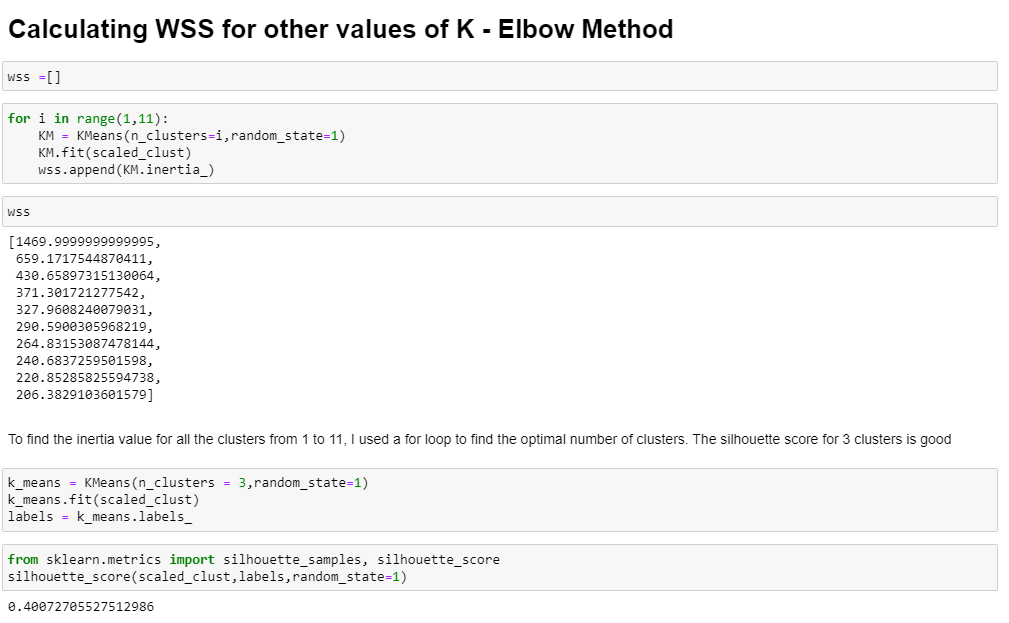
**Observation**

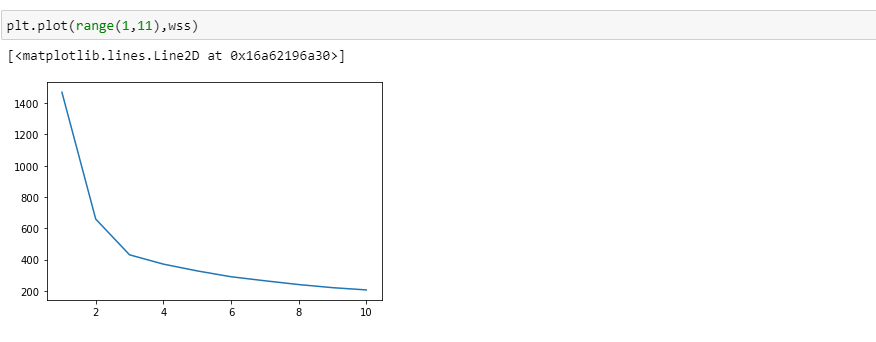
Both the method are almost similar means, minor variation, which we know it occurs. There was not too much variations from both methods Cluster grouping based on the dendrogram, 3 or 4 looks good. Did the further analysis, and based on the dataset had gone for 3 group cluster And three group cluster solution gives a pattern based on high/medium/low spending with max\_spent\_in\_single\_shopping (high value item) and probability\_of\_full\_payment (payment made).

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

Ans:







**Observation**

By K- Mean’s method we can at cluster 3 we find it optimal after there is no huge drop in inertia values. Also the elbow curve seems to show similar results. The silhouette width score of the K – means also seems to very less value that indicates all the data points are properly clustered to the cluster. There is no mismatch in the data points with regards to clustering Cluster grouping based on the dendrogram, 3 or 4 looks good. Did the further analysis, and based on the dataset had gone for 3 group cluster And three group cluster solution gives a pattern based on high/medium/low spending with max\_spent\_in\_single\_shopping (high value item) and probability\_of\_full\_payment (payment made).

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

**Ans:**

**Group 1**: High Spending Group –

Giving any reward points might increase their purchases. –

Maximum max\_spent\_in\_single\_shopping is high for this group, so can be offered discount/offer on next transactions upon full payment –

Increase their credit limit and –

Increase spending habits –

Give loan against the credit card, as they are customers with good repayment record. –

Tie up with luxury brands, which will drive more one\_time\_maximun spending.

**Group 2:** Low Spending Group - customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate. - Increase their spending habits by tying up with grocery stores, utilities (electricity, phone, gas, others)

**Group 3**: Medium Spending Group - They are potential target customers who are paying bills and doing purchases and maintaining comparatively good credit score. So we can increase credit limit or can lower down interest rate. - Promote premium cards/loyalty cars to increase transactions. - Increase spending habits by trying with premium ecommerce sites, travel portal, travel airlines/hotel, as this will encourage them to spend more.

**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.

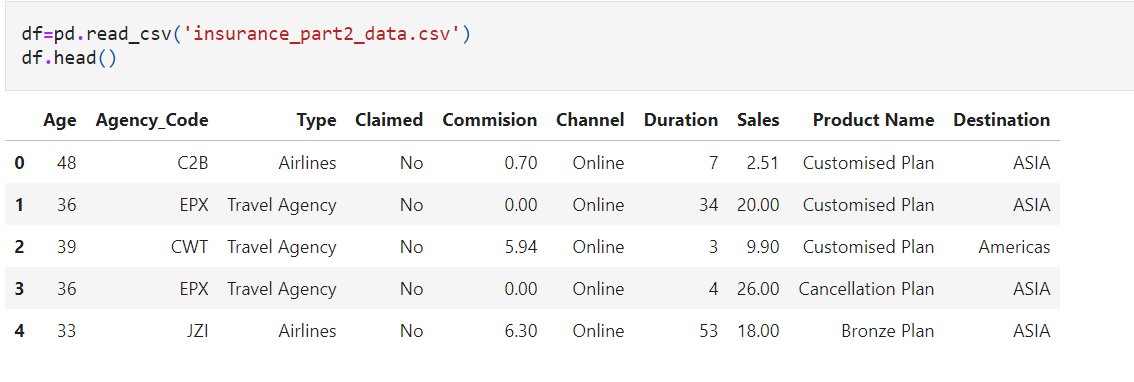
Dataset for Problem 2: [insurance\_part2\_data-1.csv](https://olympus.greatlearning.in/courses/32928/files/2545008/download?verifier=0zyUyBIShfSZhxGHFskuWdSjbe2FPooZ9DVr77Qt&wrap=1)

**Attribute** **Information:**

1. Target: Claim Status (Claimed)  
   2. Code of tour firm (Agency\_Code)  
   3. Type of tour insurance firms (Type)  
   4. Distribution channel of tour insurance agencies (Channel)  
   5. Name of the tour insurance products (Product)  
   6. Duration of the tour (Duration)  
   7. Destination of the tour (Destination)  
   8. Amount of sales of tour insurance policies (Sales)  
   9. The commission received for tour insurance firm (Commission)  
   10. Age of insured (Age)

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

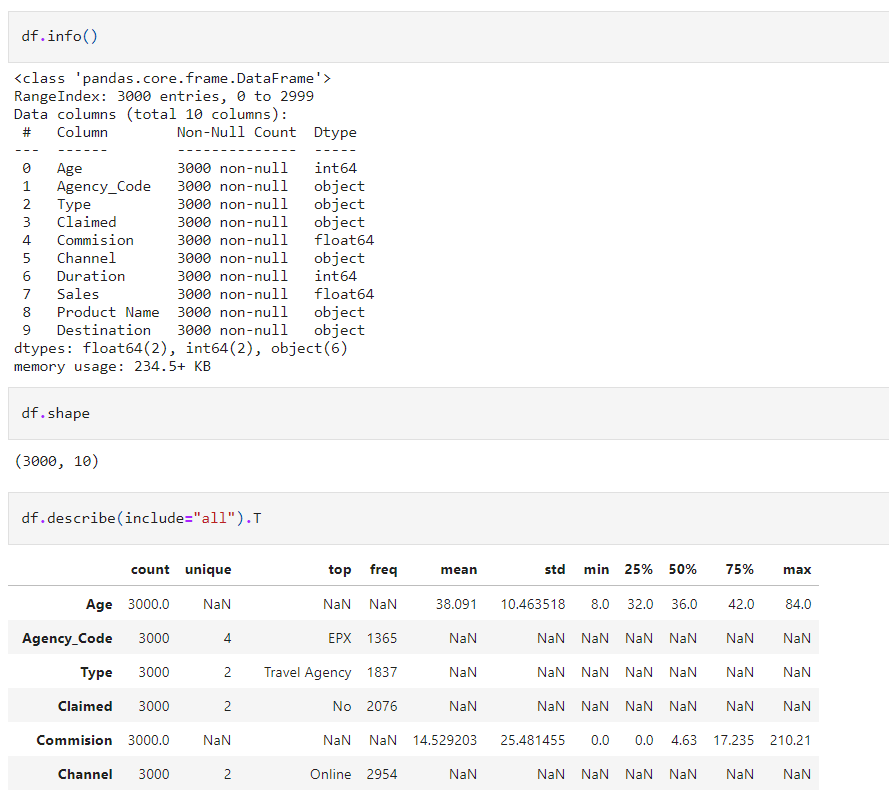
**Ans: Reading Data set**



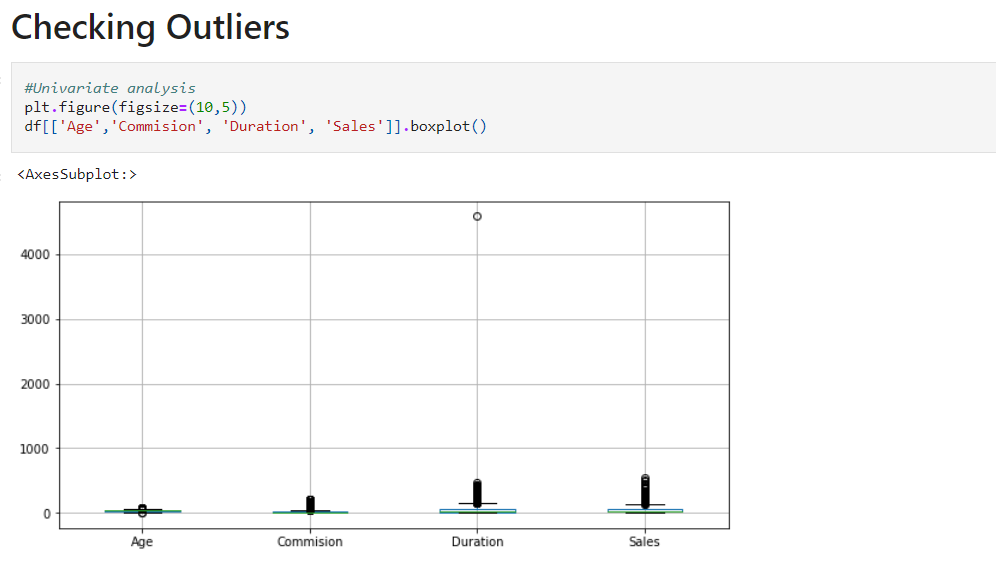
The data has read successfully,

The shape of the dataset is (3000, 10)

Info function clearly indicates the dataset has object, integer and float so we have to change the object data type to numeric value.

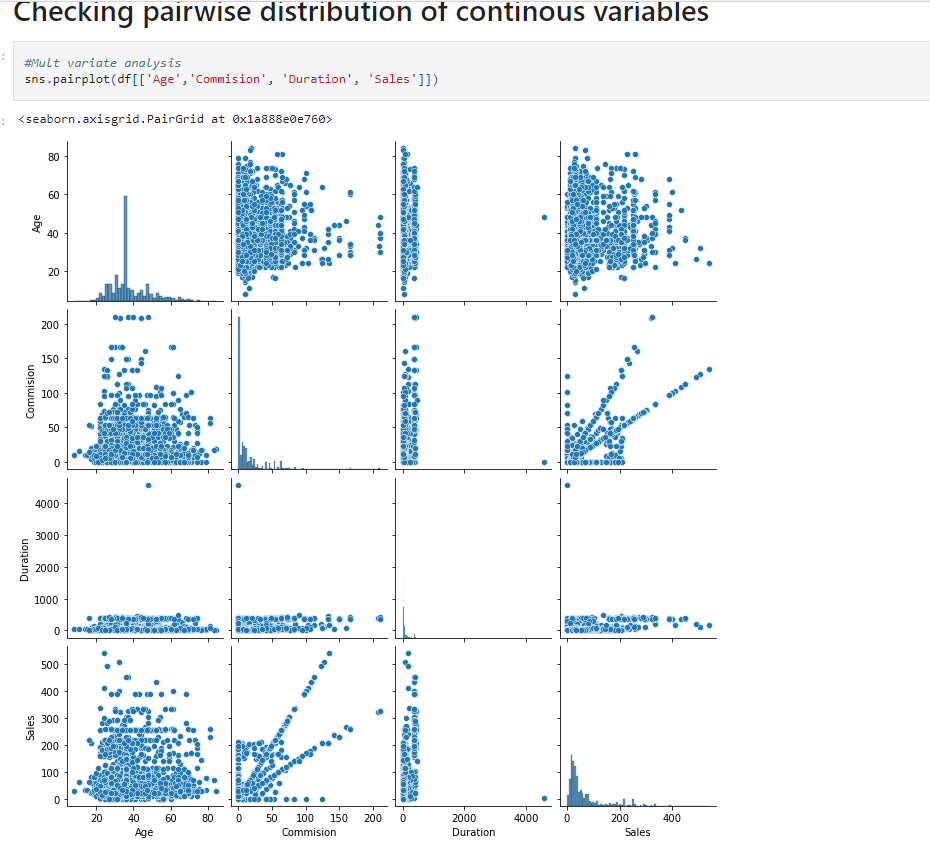


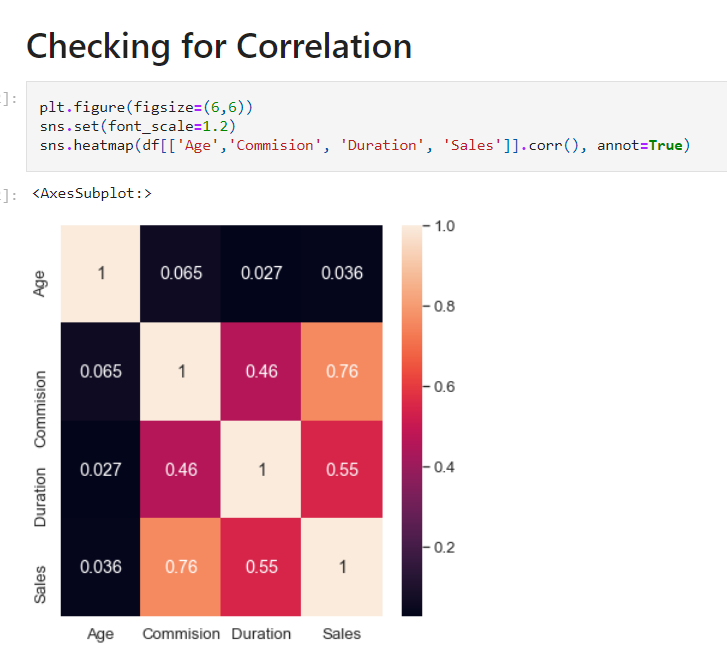


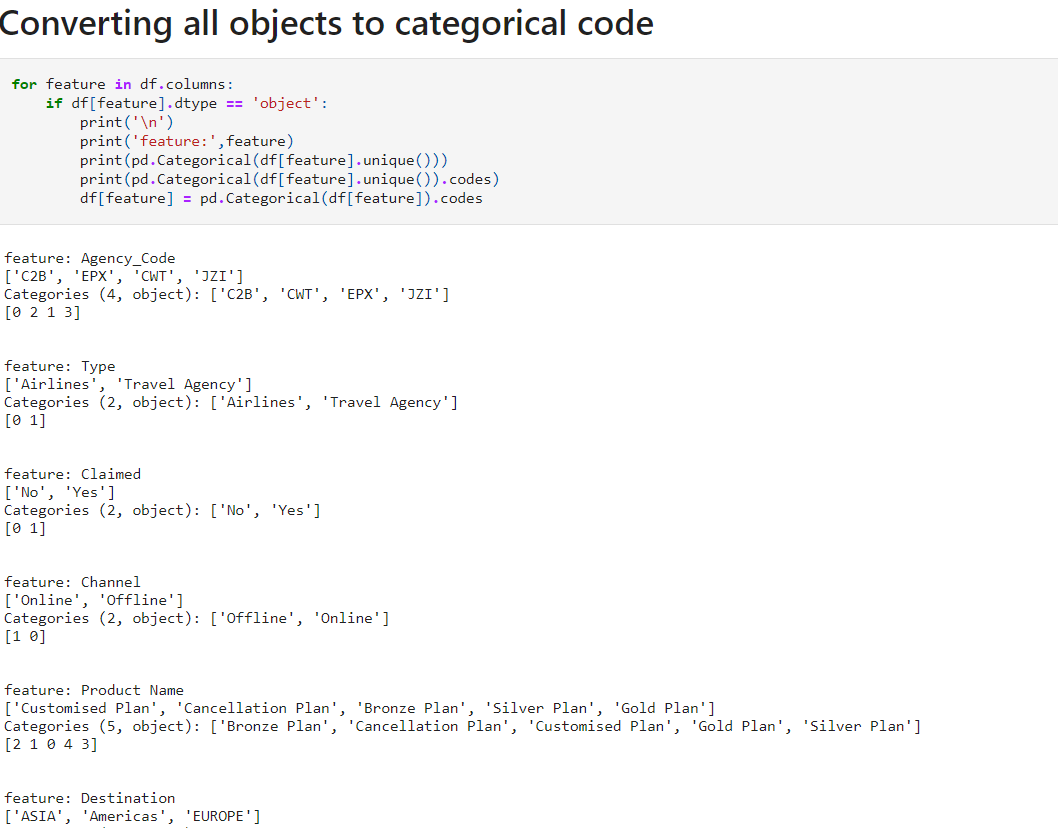


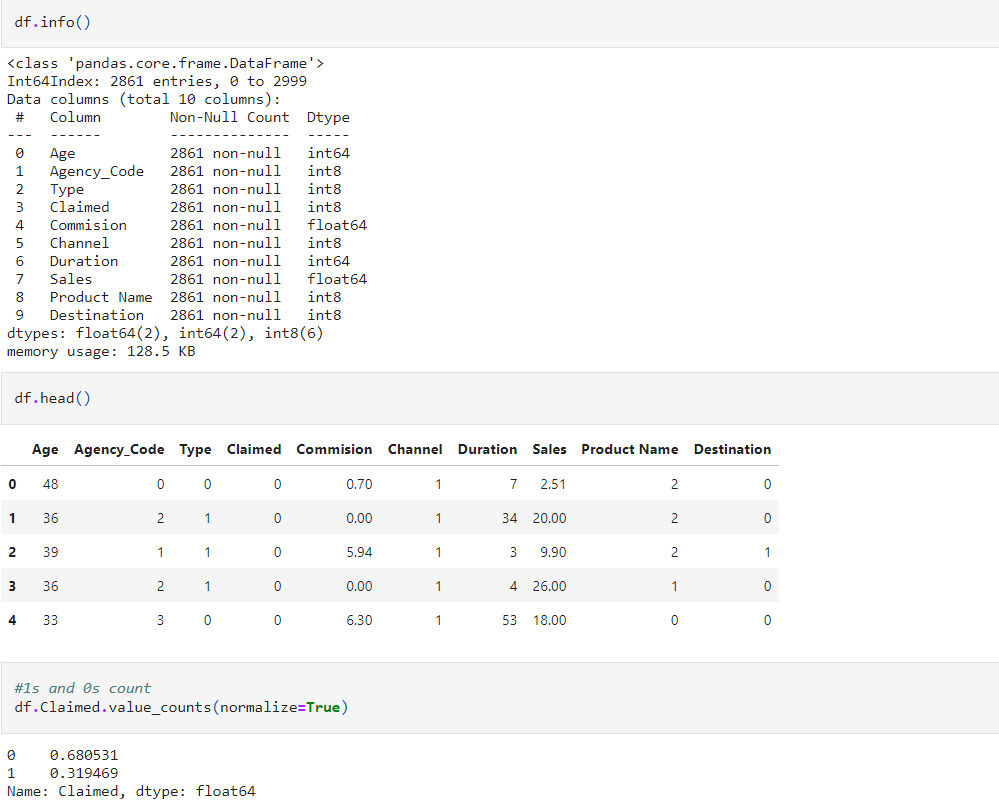
**from the above graph we can conclude that the below variables are with outliers:**

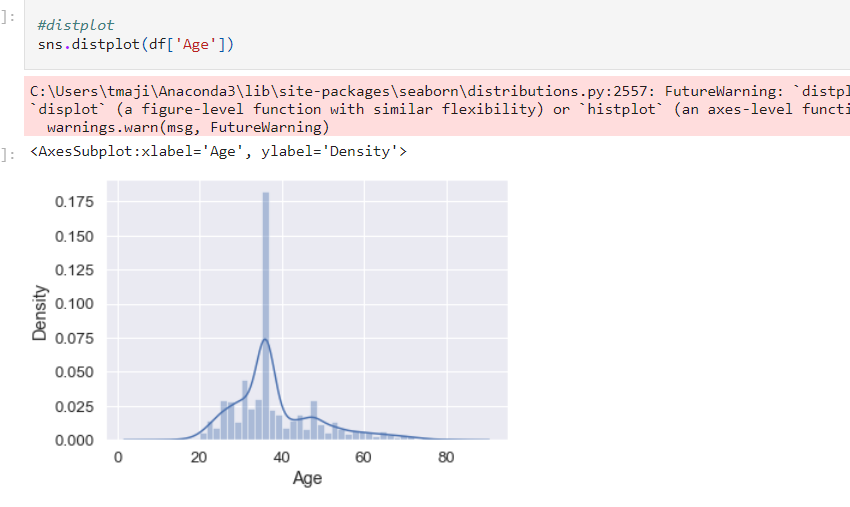
Age, Commision, Duration, Sales

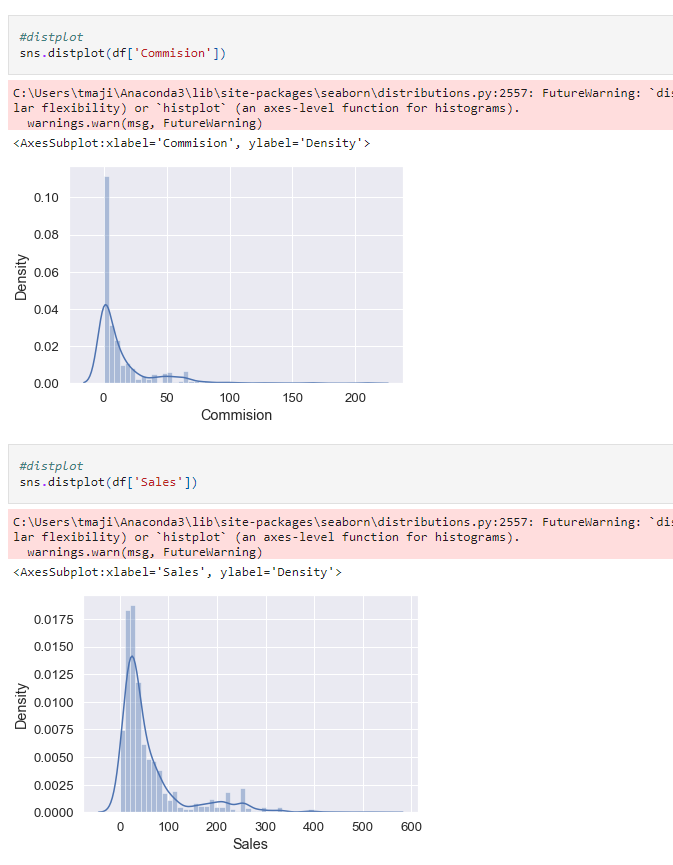


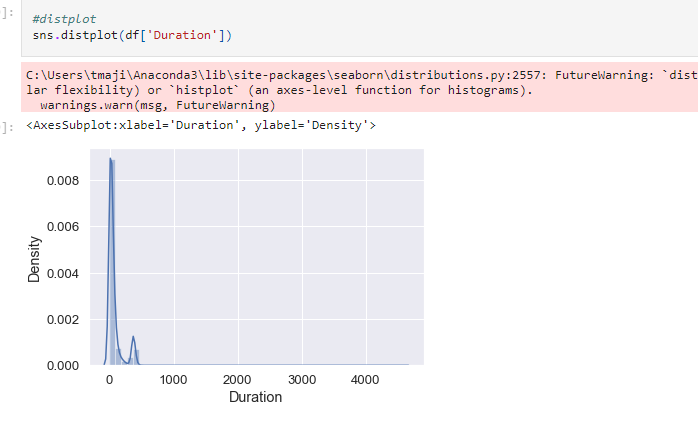










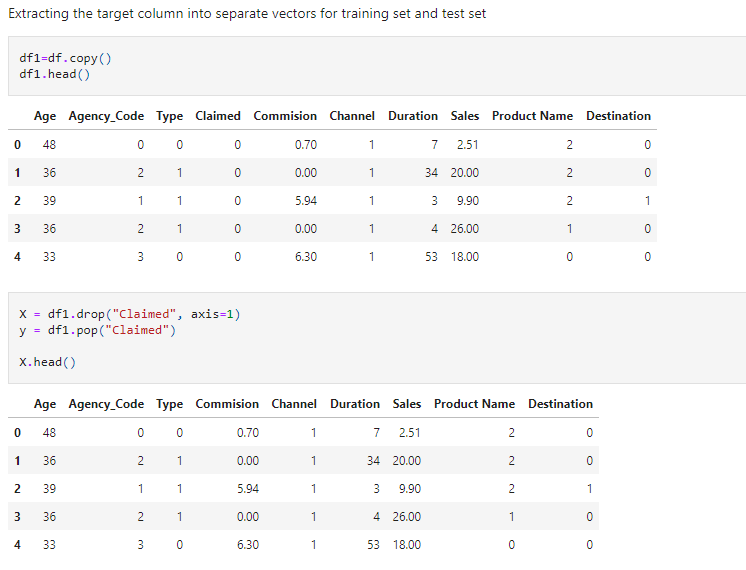


**from the above graph we can conclude that:below vaibles are right skewed**

Age, Duration, Sales, Commision

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

Ans:







# Model 1 # Building decision tree classifier

# 

# 

# 

# Model - 2 Ensemble Random Forest Classifier

# 

# Model 3 Neural network Classifier

# 

# 

# 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.

# Ans:

# 

# 

# 

# 

# 

# 

# 

# 

# 

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

# Ans:

# 

# 

# 

**CONCLUSION:**

I am selecting the RF model, as it has better accuracy, precision, recall, f1 score better than other two CART & NN.

**2.5 Inference: Basis on these predictions, what are the business insights and recommendations**

**Ans:**

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

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

Key performance indicators (KPI) The KPI’s of insurance claims are: • Reduce claims cycle time • Increase customer satisfaction • Combat fraud • Optimize claims recovery • Reduce claim handling costs Insights gained from data and AI-powered analytics could expand the boundaries of insurability, extend existing products, and give rise to new risk transfer solutions in areas like a non-damage business interruption and reputational damage.