06-DEC-2020

Data Mining

Clustering (Hierarchical and K-Means), CART, Random Forest, Artificial Neural Network

By **Chetan Dudhane**

PGP-DSBA July B Group 2
chetan.dudhane@gmail.com

INTRODUCTION

This report consists of two problem statements -

- PROBLEM 1 - Segmentation of Bank Marketing Data (Clustering)
- PROBLEM 2 - Classification of Insurance data using different models (CART, RF and ANN)

Please find the Jupyter Code Notebook in the Google Drive link below. Analysis code is in Python. Datasets used are in the same directory. - https://bit.ly/2KulOQj

PROBLEM 1 - Bank Marketing

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

1.A DATA EXPLORATION

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.88	6.68	3.76	3.25	6.55
15.99	14.89	0.91	5.36	3.58	3.34	5.14
18.95	16.42	0.88	6.25	3.76	3.37	6.15
10.83	12.96	0.81	5.28	2.64	5.18	5.19
17.99	15.86	0.90	5.89	3.69	2.07	5.84
12.70	13.41	0.89	5.18	3.09	8.46	5.00
12.02	13.33	0.85	5.35	2.81	4.27	5.31
13.74	14.05	0.87	5.48	3.11	2.93	4.83
18.17	16.26	0.86	6.27	3.51	2.85	6.27
11.23	12.88	0.85	5.14	2.80	4.33	5.00

Table 1.1 : Bank Data (First 10 Rows)

Data Dictionary for Bank Data:

- 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)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Table 1.2 : Summary Info of Bank Data

1.B BASIC UNDERSTANDING OF DATA

1. Total No. Of Customer Entries = 210
2. Total No. Of Variables = 7

   Data Type Float - All variables
3. There is NO Missing Data
4. There are NO Duplicate entries
5. We need to segment customers in different clusters based on their usage data
6. We'll use HIERARCHICAL and K-MEANS Clustering Algorithms

1.C DESCRIPTIVE ANALYSIS

	count	mean	std	min	0.25	0.50	0.75	max
spending	210.00	14.85	2.91	10.59	12.27	14.36	17.31	21.18
advance_payments	210.00	14.56	1.31	12.41	13.45	14.32	15.72	17.25
probability_of_full_payment	210.00	0.87	0.02	0.81	0.86	0.87	0.89	0.92
current_balance	210.00	5.63	0.44	4.90	5.26	5.52	5.98	6.68
credit_limit	210.00	3.26	0.38	2.63	2.94	3.24	3.56	4.03
min_payment_amt	210.00	3.70	1.50	0.77	2.56	3.60	4.77	8.46
max_spent_in_single_shopping	210.00	5.41	0.49	4.52	5.05	5.22	5.88	6.55

Table 1.3 : Bank Data - Statistical Description

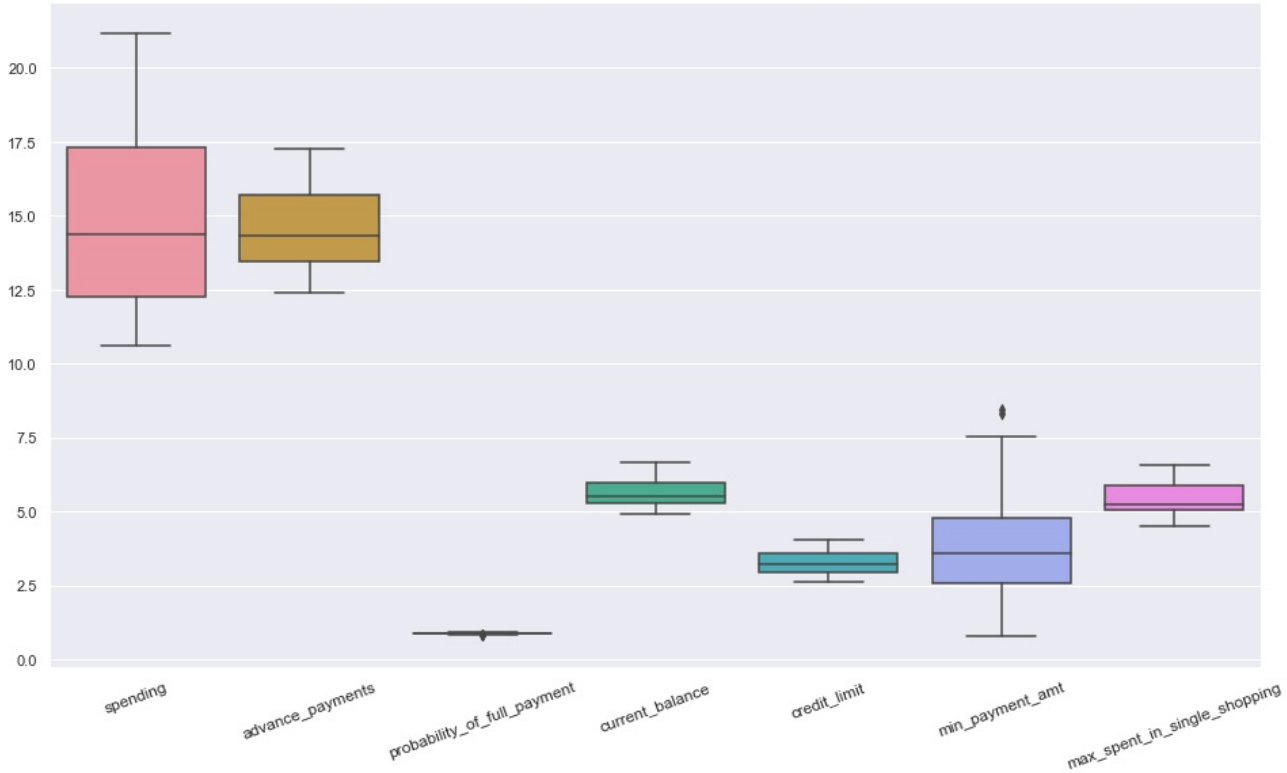
[Q 1.1] Read the data and do exploratory data analysis. Describe the data briefly.

- Read the data -
  - Data is read and stored as Pandas Data Frame for analysis
  - First 5 rows of the data is given below

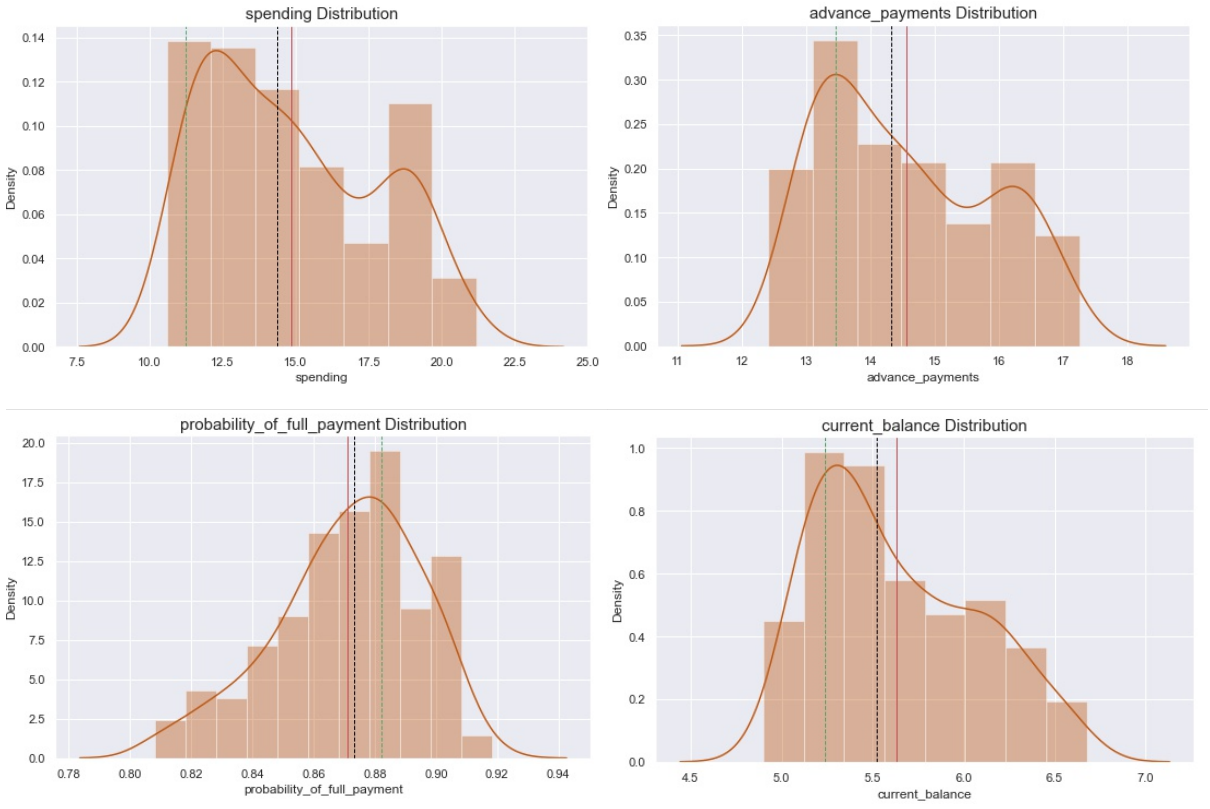
spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.88	6.68	3.76	3.25	6.55
15.99	14.89	0.91	5.36	3.58	3.34	5.14
18.95	16.42	0.88	6.25	3.76	3.37	6.15
10.83	12.96	0.81	5.28	2.64	5.18	5.19
17.99	15.86	0.90	5.89	3.69	2.07	5.84

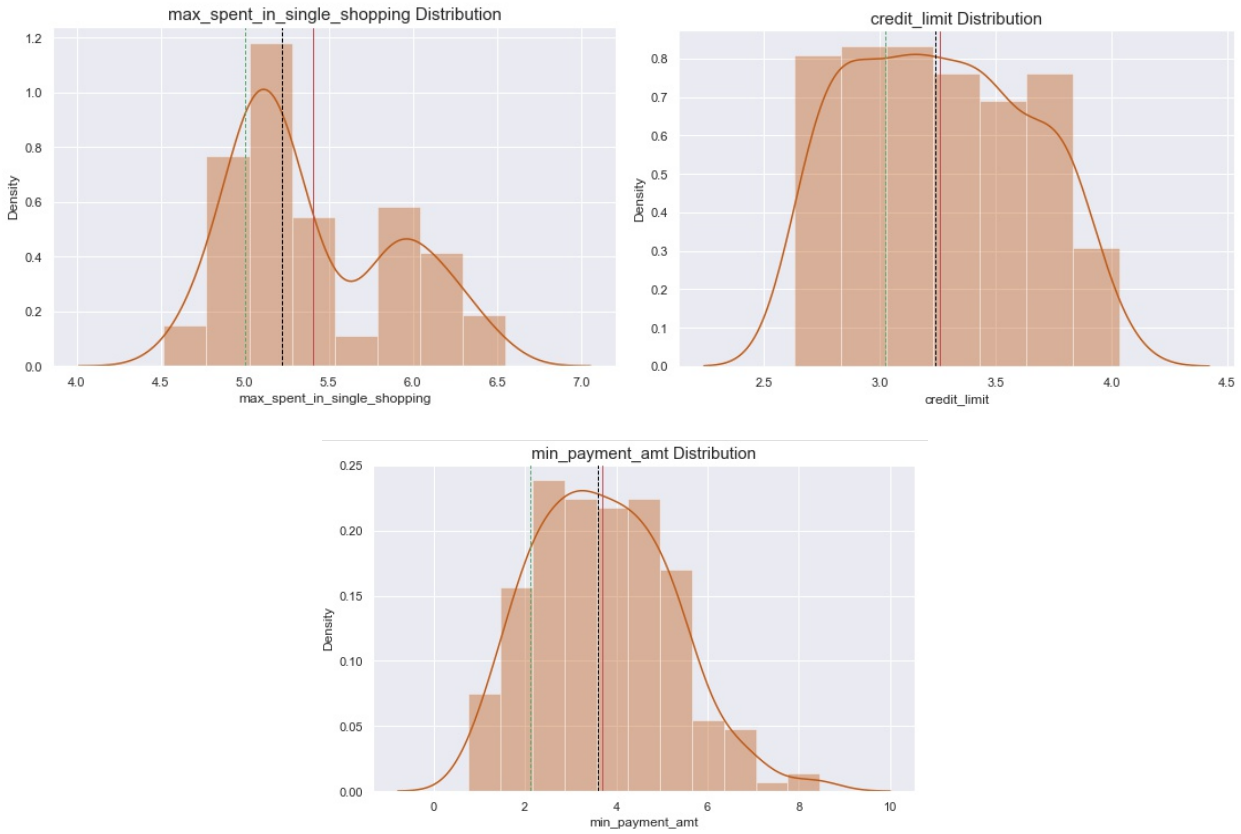
○ Exploratory Data Analysis -

- **Check Outliers** - Box-plot of each Variable is given below
  - We notice that there are NO significant Outliers

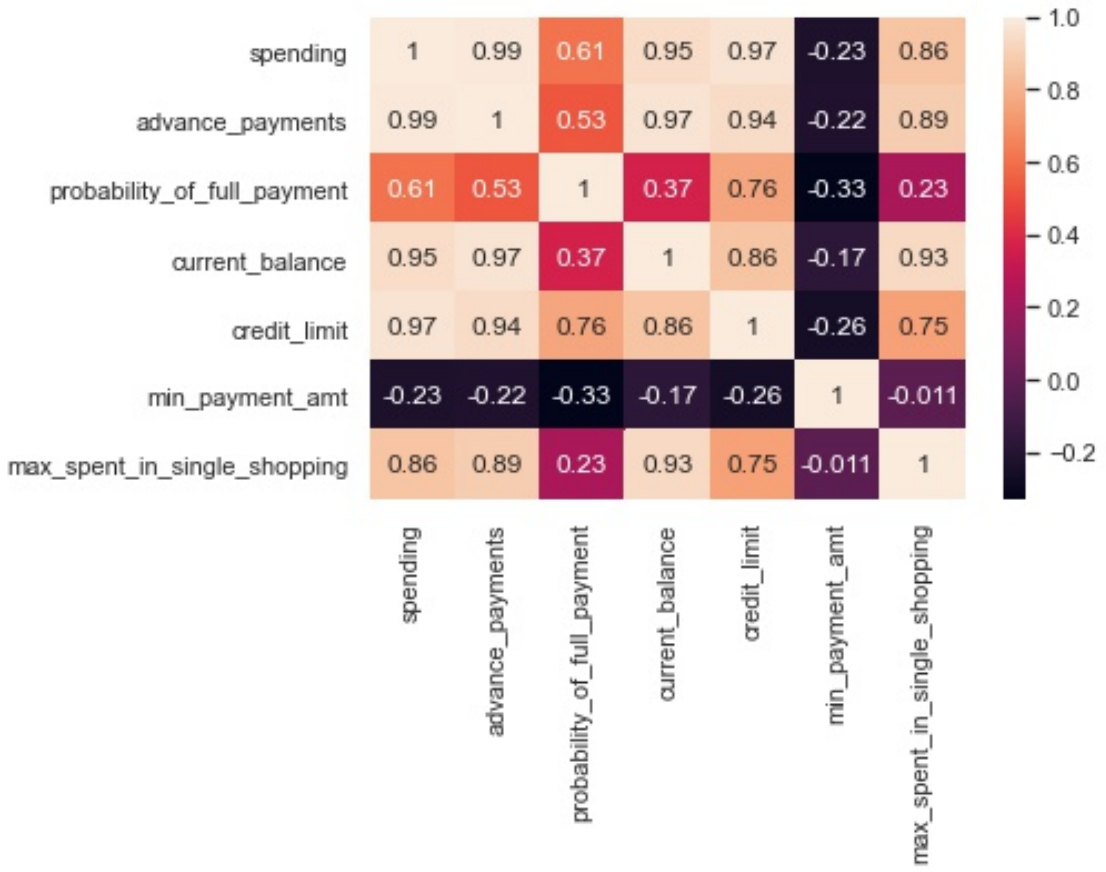


- **Check Distribution of each variable -**





- **Correlation Heat-map of Variables**



- **Description of Data -**

  - Total Records (Customers) in the data - 210
  - Total variables - 7 (All of Data Type - Float)
  - There are NO Missing Values
  - There are NO Duplicate entries
  - There is high **Correlation** between -

    - 'spending', 'advance_payments', 'current_balance', 'credit_limit' and 'max_spent_in_single_shopping'

    - means, persons with **High Bank Balance will have a High Credit Limit, who will in turn spend more in single shopping**

    - But, note that **persons spending max in single shopping have the least correlation with Probability of making Full payments**

  - Variables 'min_amount_due', 'probability_of_full_payment' and 'credit_limit' - are fairly well Normally Distributed
  - spending --------> Mean = 14.85, Median = 14.36, CV = 19.6
    advance_payments --------> Mean = 14.56, Median = 14.32, CV = 8.97
    probability_of_full_payment --------> Mean = 0.87, Median = 0.87, CV = 2.71
    current_balance --------> Mean = 5.63, Median = 5.52, CV = 7.87
    credit_limit --------> Mean = 3.26, Median = 3.24, CV = 11.59
    min_payment_amt --------> Mean = 3.7, Median = 3.6, CV = 40.63
    max_spent_in_single_shopping ----> Mean = 5.41, Median = 5.22, CV = 9.09
  - In the above, CV = Coefficient of Variation
  - There is a large Probability that customers will pay full Amount Due
  - There is a large variation in Customers paying just the Minimum Amount Due
    - maybe, because most of them are paying Full Dues
  - Variable 'probability_of_full_payment' has the least variation
    - means, maximum values hover around its Mean

[Q 1.2] Do you think scaling is necessary for clustering in this case? Justify.

- Clustering Algorithm clusters similar or homogeneous observations together.
- Similarity or Homogeneity of records is determined by distance metrics.
- Hence, **Clustering is a 'Distance' based Algorithm**
- If the units of all variables is not same, then the **larger values will dominate** the algorithm, which will be incorrect and would lead to biased output
- For example,

    - Here ‘advance_payments’ is in 100s and ‘credit_limit’ is in 10000s,

    - So in sheer magnitude, ‘advance_payments’ will have larger values than ‘credit_limit’

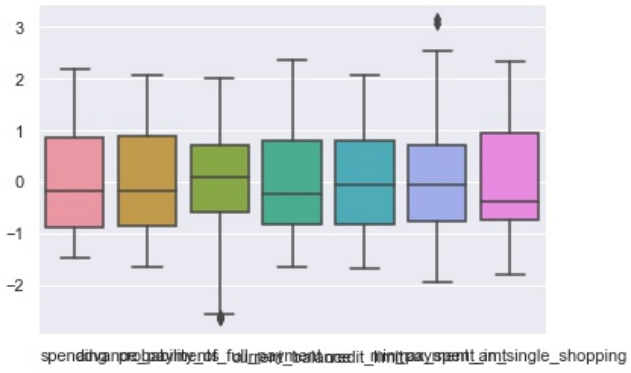
    - But in actual, it may not be true

    - In this case, in the absence of Scaling, Algorithm would incorrectly favour ‘advance_payments’

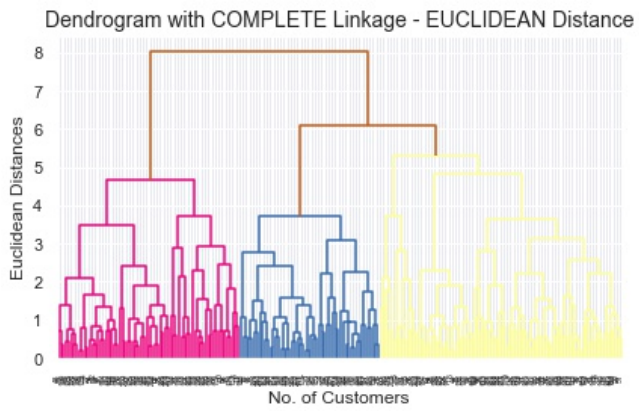
- **Standardising** or **Scaling takes care of this issue**
- Standardising converts all variables into standard scale with mean=0 and standard deviation=1.
- Here in this example, as variables have different units (multiples of 100s, 1000s, 10000s) - IT IS NECESSARY **TO SCALE**
- We'll apply Standard Scaling

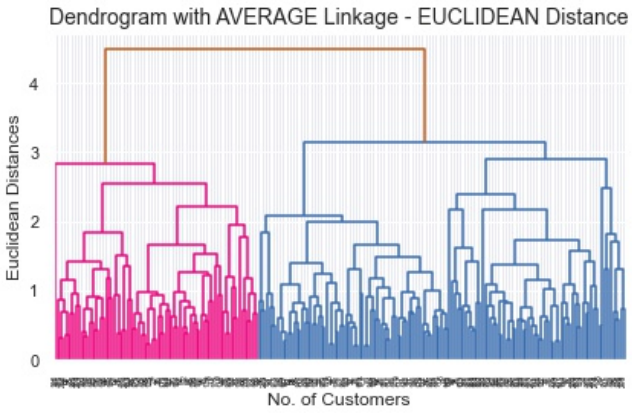
[Q 1.3] Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

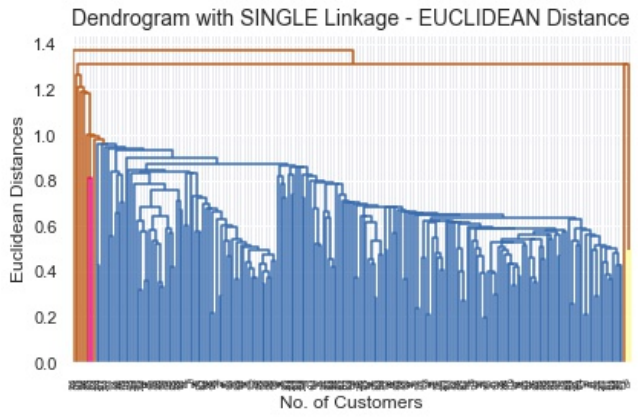
- Bank Data is Z-scaled using the function - StandardScaler
- Box-plot of Scaled Data given at the side

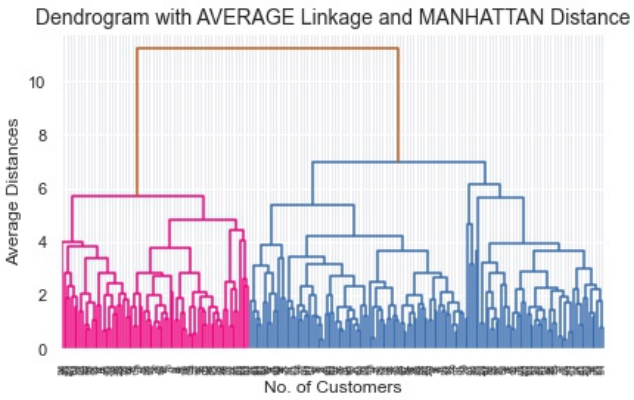


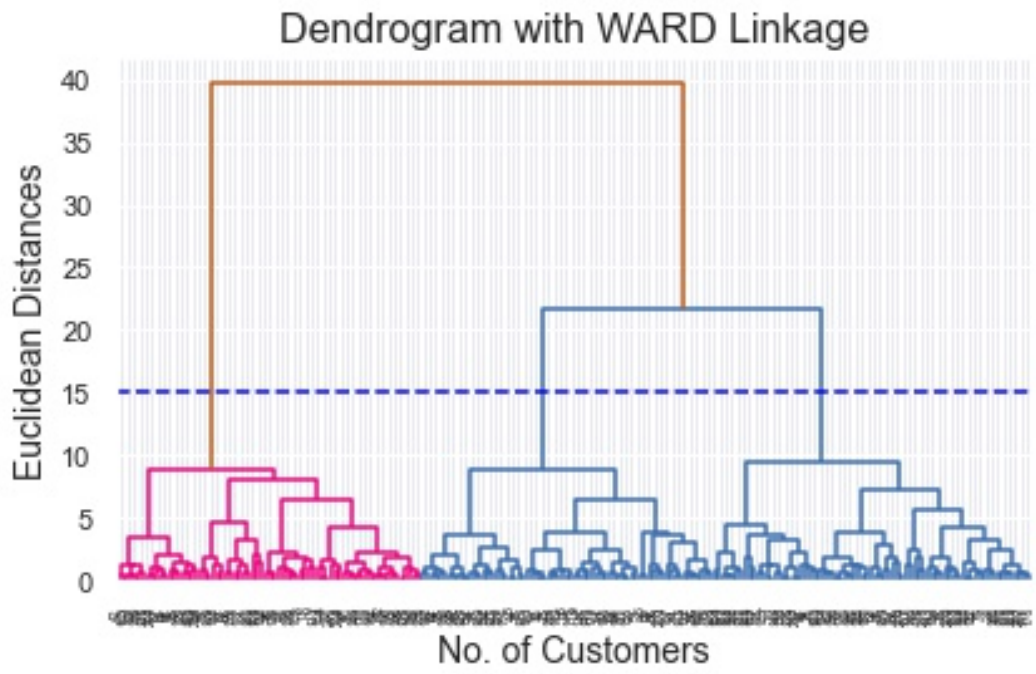
- Dendrograms are plotted with various combinations of Linkages and Distance Metrics











- Dendrogram is the visual representation of Cluster formations.
- The **vertical lines** represent the distance **between** the clusters
- So, **Higher the distance between the Clusters, More and Better separability between them**
- So, considering the above point - **WARD Linkage** gives Better **Separability** between Clusters (refer figures given above)

  (Note : Ward linkage considers distance by only Euclidean method)
- Choosing the exact combination of linkages and distances is a subjective choice depending on domain, no. of observations, cluster size created, etc
- WE CHOOSE WARD LINKAGE
- **Choosing Optimal Number of Clusters using Dendrogram** -

  - Choose the tallest lines between Clusters (max separation between clusters)

  - Draw a horizontal line crossing these vertical lines

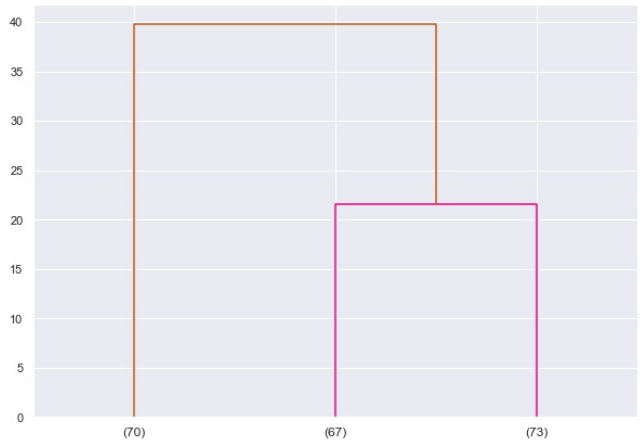
  - Number of vertical lines crossed by this horizontal line

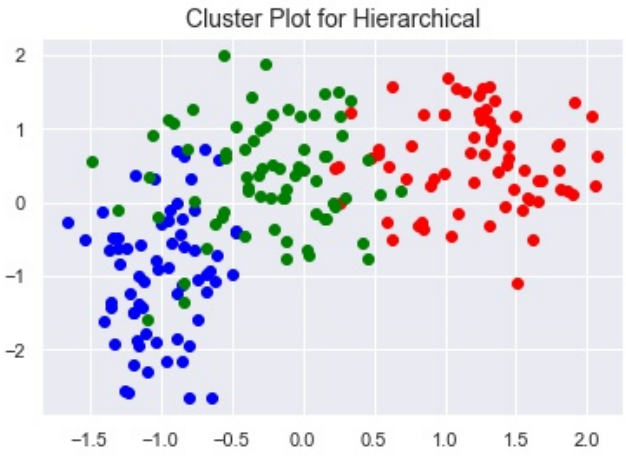
    = Number of Optimal Clusters

  - Here, it will be 2 or 3 clusters

  - WE CHOOSE 3 CLUSTERS

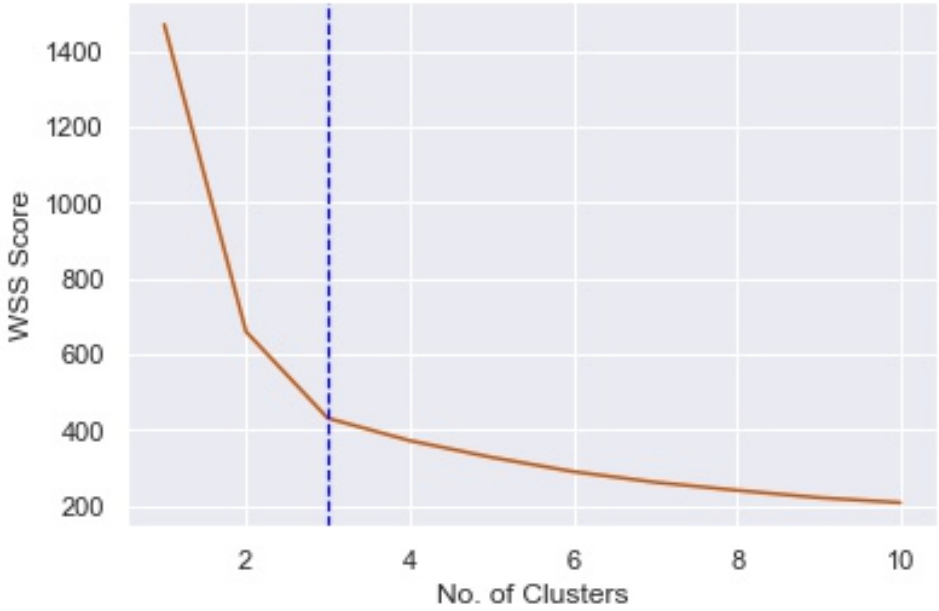
  - Refer the WARD Linkage figure above
- Final Truncated Ward Linkage Dendrogram and Cluster Plot with 3 Clusters given below





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

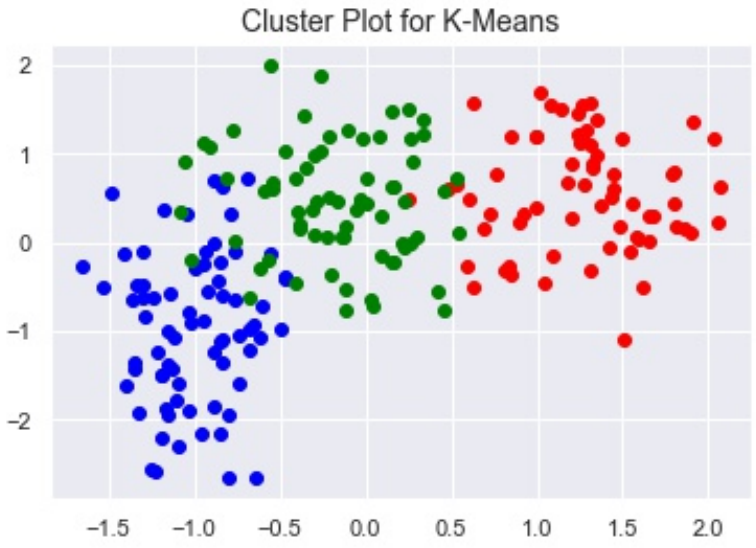
- Bank Data is scaled
- <u>Choosing Optimal Number of Clusters by WSS Elbow Curve</u> -
  - WSS - stands for Within Sums of Squares - i.e. Intra-Cluster Sums of Squares
  - **WSS is the Sum of distance squared of every point from the centroid of its own cluster**
  - We run a loop to find WSS for 2 to 10 clusters and plot it as given below



  - The location of the bend or an elbow like shape is considered as the definitive indicator of optimal point
  - The Elbow in the above plot is at k=3 (marked by vertical blue dashed line)
  - So, **the Optimal Number of Clusters is taken as k = 3**

- <u>Choosing Optimal Number of Clusters by Silhouette Score method</u> -
  - **Silhouette Score** is the **Average of Silhouette Widths** of all points for the considered k number of clusters
  - Silhouette Score of Clusters explains separability of clusters.
  - Values range from (-1, 1) ---> where higher value indicates a better separation between clusters. An higher value is desired.

- We run a loop to find Silhouette Score of 2 to 10 clusters
- The maximum values are for k=2 and then k=2 -
  Sil Score = 0.4657 for k=2
  Sil Score = 0.4007 for k=3
- Even though 2 clusters have maximum Silhouette score, 2 clusters are too less for profiling, **WE CHOOSE 3 CLUSTERS** which have a high Sil Score too.
- Cluster plot for K-Means Clustering for k=3, given below



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

- HIERARCHICAL CLUSTERING - Means of all variables by clusters

Cluster	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1	18.370	16.150	0.880	6.160	3.680	3.640	6.020
2	11.870	13.260	0.850	5.240	2.850	4.950	5.120
3	14.200	14.230	0.880	5.480	3.230	2.610	5.090

Means of All Variables by Clusters (Hierarchical)

- K-MEANS CLUSTERING - Means of all variables by clusters (Cluster Means)

Cluster	spending	advance_ payments	probability_of_ full_payment	current_ balance	credit_ limit	min_paym ent_amt	max_spent_in_s ingle_shopping
0	18.50	16.20	0.88	6.18	3.70	3.63	6.04
1	11.86	13.25	0.85	5.23	2.85	4.74	5.10
2	14.44	14.34	0.88	5.51	3.26	2.71	5.12

Means of All Variables by Clusters (K-Means)

- Cluster Sizes
  - K Means
    Cluster 0 - 67
    Cluster 1 - 72
    Cluster 2 - 71
  - Hierarchical
    Cluster 1 - 70
    Cluster 2 - 67
    Cluster 3 - 73
  - Cluster sizes and metrics are almost the same for both. Lets consider **K-means**
- Cluster 0 - are financial elite
  - are the highest spenders with highest credit Limit with highest bank balance
  - This Cluster should be given enhanced Credit Limits and upgrades
  - Every offer, discount and festival plans should be individually informed
  - Priority Club and Lounge memberships can be offered for a small fee
- Cluster 1 - are the least Spenders with least Credit Limit and least bank balance
  - Consumer Loans should be offered with long term EMI offers
  - Temporary Enhanced Credit limit on festivals can be offered to aid big spends
  - Offer free annual charges on upgrades if they register 2 or more monthly billers
- Cluster 2 - are the Middle Spenders but they pay Minimum Amount the least
  - they are fence sitters, should be given high incentives along-with strategies Suggested for Cluster 1
  - No Cost EMI for big budget items can be offered
  - Offer reduced interest rate during festivals if they pay only Minimum Amount due

PROBLEM 2 - Insurance Claim Modelling

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

2.A DATA EXPLORATION

Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
36	EPX	Travel Agency	No	0.0	Online	34	20.0	Customised Plan	ASIA
39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
36	EPX	Travel Agency	No	0.0	Online	4	26.0	Cancellation Plan	ASIA
33	JZI	Airlines	No	6.3	Online	53	18.0	Bronze Plan	ASIA
45	JZI	Airlines	Yes	15.75	Online	8	45.0	Bronze Plan	ASIA
61	CWT	Travel Agency	No	35.64	Online	30	59.4	Customised Plan	Americas
36	EPX	Travel Agency	No	0.0	Online	16	80.0	Cancellation Plan	ASIA
36	EPX	Travel Agency	No	0.0	Online	19	14.0	Cancellation Plan	ASIA
36	EPX	Travel Agency	No	0.0	Online	42	43.0	Cancellation Plan	ASIA

Insurance Data (First 10 rows)

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)

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

Summary Info of Insurance data

2.B BASIC UNDERSTANDING OF DATA

1. Total No. Of Customer Entries = 3000
2. Total No. Of Variables = 10 [9 Dependent - 1 Target Variable (Claimed)]
3. There is NO Missing Data
4. There are 139 Duplicate entries

    - Customer ID is not present so having similar profile of tourists is common

    - But, We decide to drop these duplicate rows

5. Final dataset entries for Modelling - 2861

6. We need to create different models to predict if a tourist taking insurance would 'Claim' or 'Not Claim' the insurance

[Q 2.1] Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

- Data is read into a Pandas Data Frame. First 5 rows of data below-

Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
36	EPX	Travel Agency	No	0.0	Online	34	20.0	Customised Plan	ASIA
39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
36	EPX	Travel Agency	No	0.0	Online	4	26.0	Cancellation Plan	ASIA
33	JZI	Airlines	No	6.3	Online	53	18.0	Bronze Plan	ASIA

- Null Value Check -

  - There are **NO Missing Values** in the dataset
  - Variable 'Duration' - has 3 entries of -1 & 0 (which seems falsely mentioned) - But its just 3 entries, so Ignoring it

- Descriptive Statistics and Data Exploration -

	count	unique	top	freq	mean	std	min	0.25	0.50	0.75	max
Age	2861				38.20	10.68	8.00	31.00	36.00	43.00	84.00
Agency_Code	2861	4	EPX	1238							
Type	2861	2	Travel Agency	1709							
Claimed	2861	2	No	1947							

	count	unique	top	freq	mean	std	min	0.25	0.50	0.75	max
Commision	2861				15.08	25.83	0.00	0.00	5.63	17.82	210.21
Channel	2861	2	Online	2815							
Duration	2861				72.12	135.98	-1.00	12.00	28.00	66.00	4580.00
Sales	2861				61.76	71.40	0.00	20.00	33.50	69.30	539.00
Product Name	2861	5	Customised Plan	1071							
Destination	2861	3	ASIA	2327							

- Number of **Continuous Variables - 4**

  Age --------> Mean = 38.09, Median = 36.0, CV = 27.47

  Commision --------> Mean = 14.53, Median = 4.63, CV = 175.38

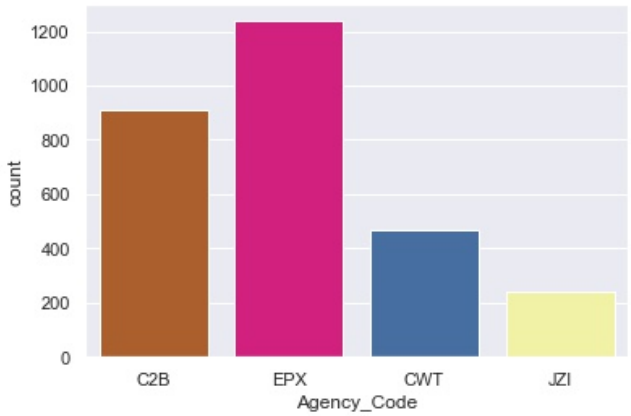
  Duration --------> Mean = 70.0, Median = 26.5, CV = 191.5

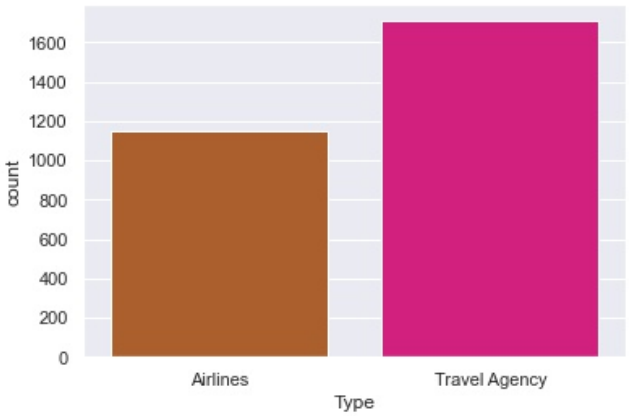
  Sales --------> Mean = 60.25, Median = 33.0, CV = 117.4

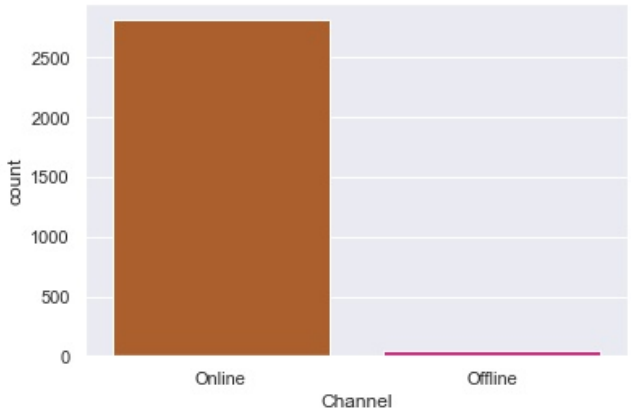
- Variable 'Age —-> has the least variation - so max tourists are aged around 35-40
- Variable 'Commision' —-> has max variation with min=0 to max=210.21
- Number of **Categorical variables = 5**

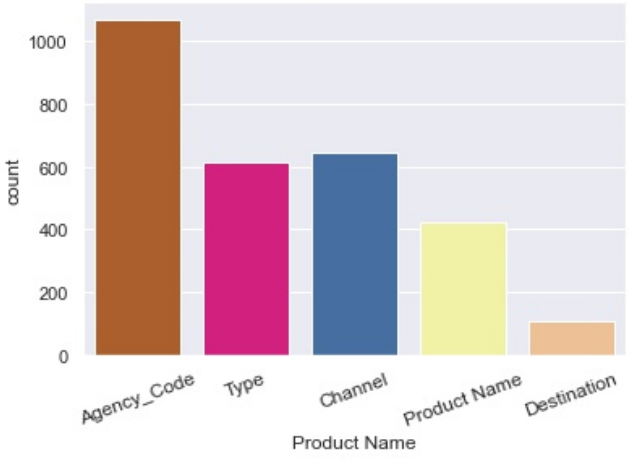
  ['Agency_Code', 'Type', 'Channel', 'Product Name', 'Destination']

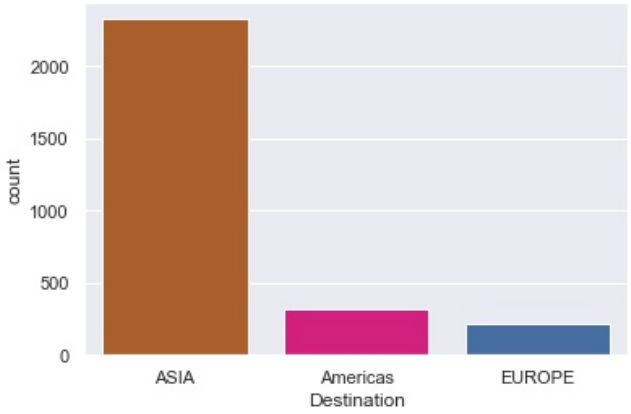
  - Count-plots of all Categorical variables given below



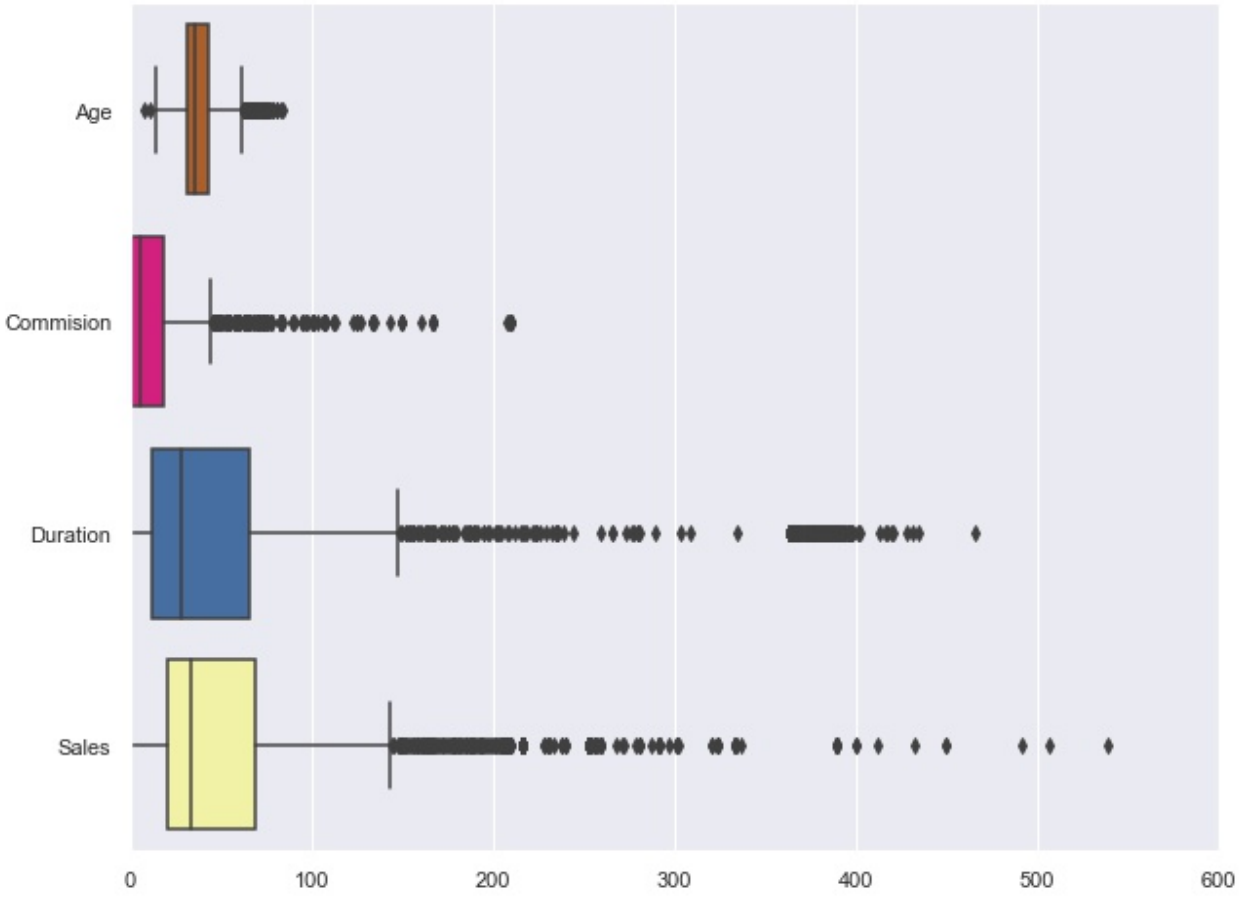








- Check Outliers - Box-plot of all Continuous variables given below



- There are many outliers in the data

- As, CART and RF is not sensitive to Outliers

- **We'll Treat Outliers only for Neural Network Modelling**

- **<u>Check Correlation</u> - Correlation matrix given below**

- Sales and Commission have significant correlation

- It can cause some issues of Multi-Collinearity

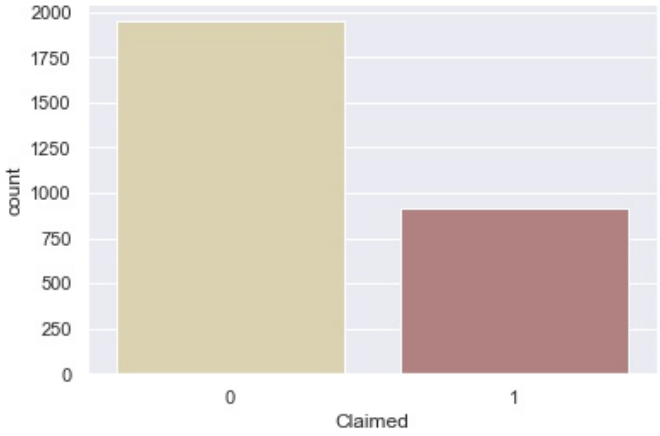
- But correlation is not high enough to take a definitive decision on it

- So, We'll keep both Variables for now

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

TRAIN AND TEST SPLIT

- Final dataset for modelling has **2861** observations and **10** variables
- We encode the data for modelling
- We assign **9 variables except 'Claimed' to dependent variable x and 'Claimed' to y**
- We perform a **70:30 - Train:Test Split**
- Train data - 2002 observations
- Test data - 859 observations



- As seen above, there is a reasonable proportion of 0s and 1s in the target variable - 'Claimed' ['Not Claimed' - 0 and 'Claimed' = 1]

CART MODEL

- We call DecisionTreeClassifier with random state = 0
- We perform a GridSearch on all the following parameters over 10 folds -

```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [2, 4, 6, 10, 15],
    'min_samples_leaf': [10, 15, 20, 30],
    'min_samples_split': [200, 300, 400],
}
```

- Best Parameters for CART were found to be as follows -

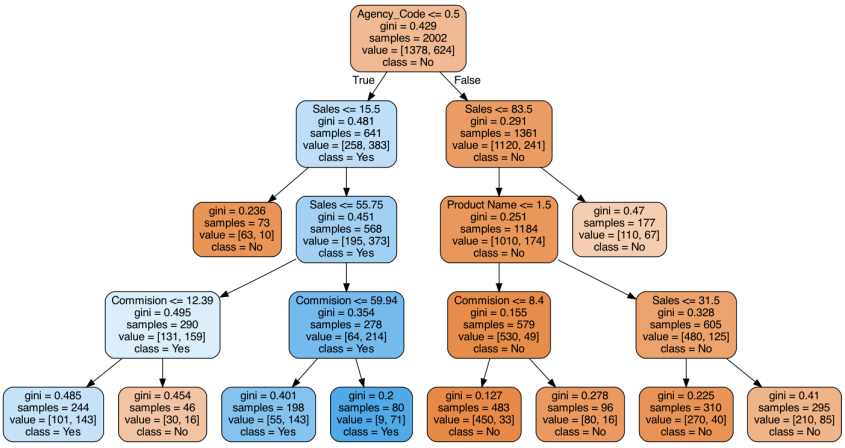
'criterion': 'gini',

'max_depth': '4',

'min_samples_leaf': '20',

'min_samples_split': '200'

- We build CART model using these Best Parameters to get given Decision Tree -



RANDOM FOREST MODEL

- We call RandomForestClassifier with random state = 0
- We perform a GridSearch on all the following parameters over ***10 folds*** -

```
param_grid = {
    'max_depth': [6, 8, 10],              ## 4, 6, 8, 10, 12,
    'max_features': [2, 3, 4],            ## 2, 3, 4, 5, 6
    'min_samples_leaf': [5 ,10, 15],      ## 5, 10 ,15, 20, 30, 50, 60,100
    'min_samples_split': [30, 50, 100],   ## 30, 50, 60,70, 100
    'n_estimators': [100, 200, 300]       ## 100,200, 300
}
```

- Best Parameters for RANDOM FOREST were found to be as follows -

```
{'max_depth': 8,
'max_features': 2,
'min_samples_leaf': 5,
'min_samples_split': 30,
'n_estimators': 200}
```

- WE build the model using the above Best Parameters

ARTIFICIAL NEURAL NETWORK MODEL

- **Treat Outliers -**
  - We treat the Outliers as Neural Networks are sensitive to Outliers (IQR Method with whisker length = 1.5 * IQR)
- **Scale Data -**
  - We use Z-Scaling using the function StandardScaler
  - **Note** : We fit the Train data and only Transform the Test Data
- We call MLPClassifier with random state = 0
- We perform a GridSearch on all the following parameters over ***10 folds*** -

```
param_grid = {
    'hidden_layer_sizes': [(100,100,100), (300,300,300), 300, 500],
                          ## (100,100,100), 300, 500, 100, (300,300,300)
    'activation': ['relu'],                    ## logistic, relu
```

```
    'solver': ['sgd'],                    ## sgd, adam
    'tol': [0.001, 0.0001],               ## 0.01, 0.001, 0.0001
    'max_iter' : [10000]                  ## 5000, 10000
}
```

- Best Parameters for NEURAL NETWORK were found to be as follows -

  {'activation': 'relu',
  'hidden_layer_sizes': (300, 300, 300),
  'max_iter': 10000, 'solver':
  'sgd', 'tol': 0.0001}

- WE build the model using the above Best Parameters

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

CART

- TRAIN Data:

  AUC: 81.3%

  Accuracy: 78.42%

  Precision: 68%

  Recall: 57%

  f1-Score: 62%

- TEST Data

  AUC: 78.7%

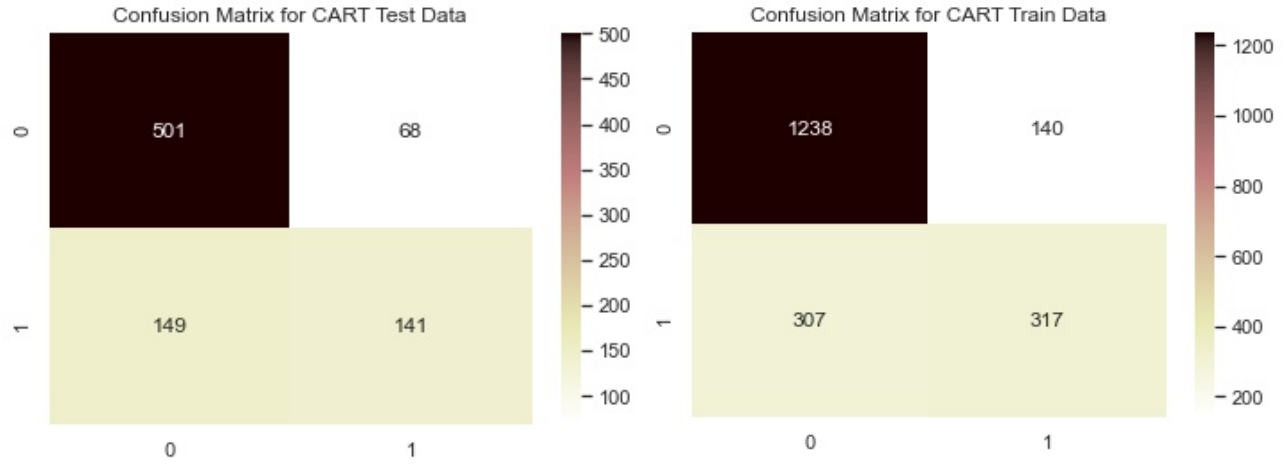
  Accuracy: 75.32%

  Precision: 67%

  Recall: 53%

  f1-Score: 59%

- CONFUSION MATRIX



- CLASSIFICATION REPORT

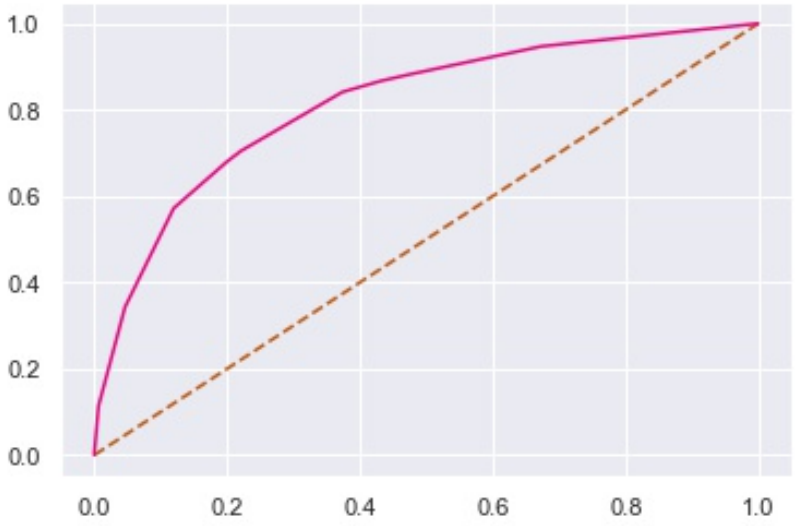
	precision	recall	f1-score	support
0	0.80	0.90	0.85	1378
1	0.69	0.51	0.59	624
accuracy	0.78	0.78	0.78	1
macro avg	0.75	0.70	0.72	2002
weighted avg	0.77	0.78	0.77	2002

Classification Report for TRAIN Data

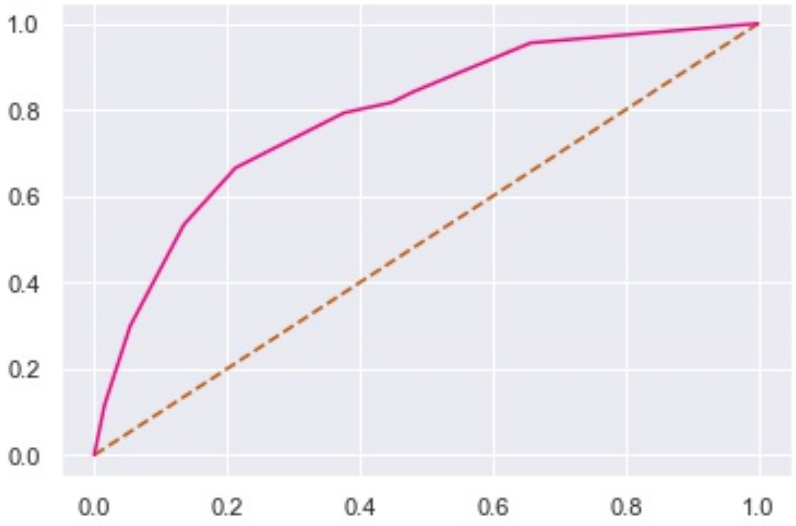
	precision	recall	f1-score	support
0	0.77	0.88	0.82	569
1	0.67	0.49	0.57	290
accuracy	0.75	0.75	0.75	1
macro avg	0.72	0.68	0.69	859
weighted avg	0.74	0.75	0.74	859

Classification Report for TEST Data

- ROC CURVE FOR TRAIN DATA (AUC = 81.3%)



- ROC CURVE FOR TEST DATA (AUC = 78.7%)



- Training and Test set results are almost similar - hence consistent
- But the overall numbers are not great
- This is not such a good model to predict if Tourists would Claim or Not Claim Insurance
- Agency_Code i.e which Agency books the tour matters the most - This is the most important variable for prediction

RANDOM FOREST

- TRAIN DATA:

  AUC: 86.6%

  Accuracy: 80.97%

  Precision: 74%

  Recall: 60%

  f1-Score: 66%

- TEST DATA:

  AUC: 79.6%

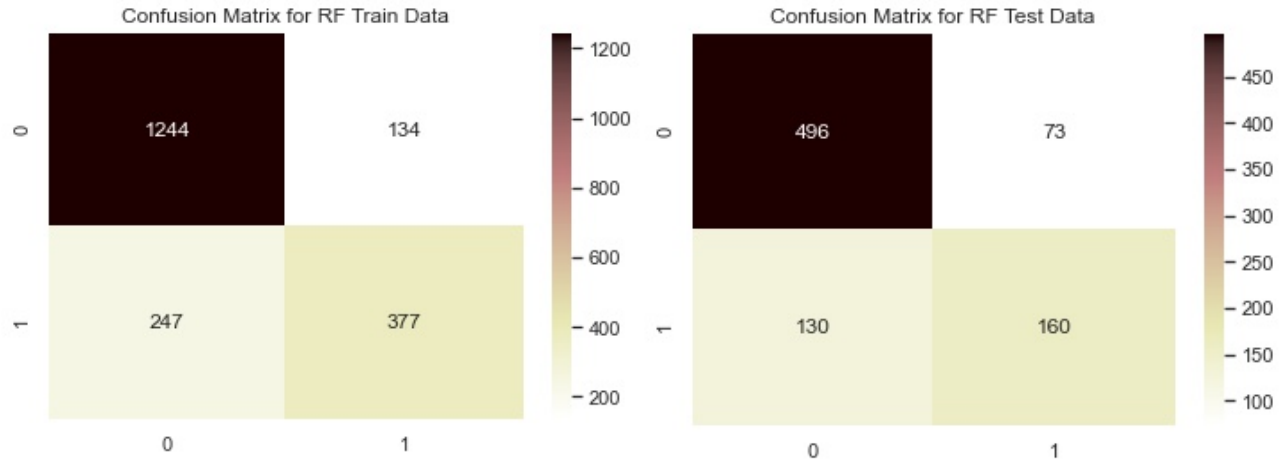
  Accuracy: 76.37%

  Precision: 69%

  Recall: 55%

  f1-Score: 61%

- CONFUSION MATRIX



- CLASSIFICTION REPORT

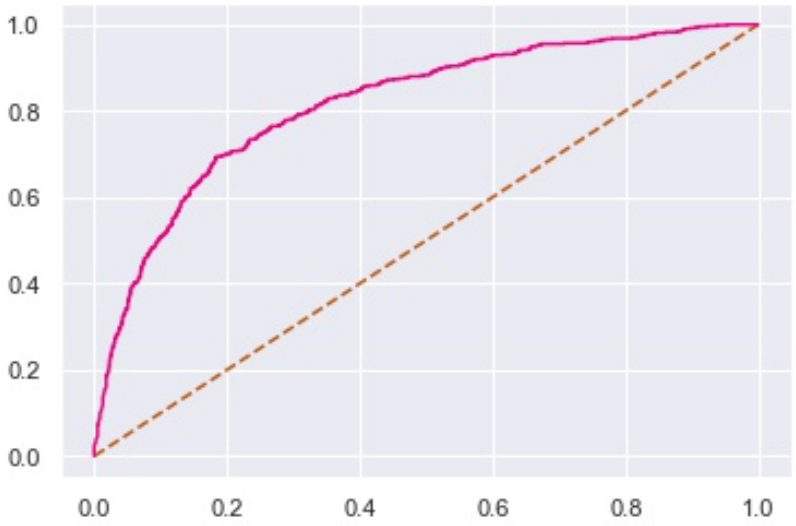
	precision	recall	f1-score	support
0	0.80	0.90	0.85	1378
1	0.69	0.51	0.59	624
accuracy	0.78	0.78	0.78	1
macro avg	0.75	0.70	0.72	2002
weighted avg	0.77	0.78	0.77	2002

Classification Report of TRAIN Data

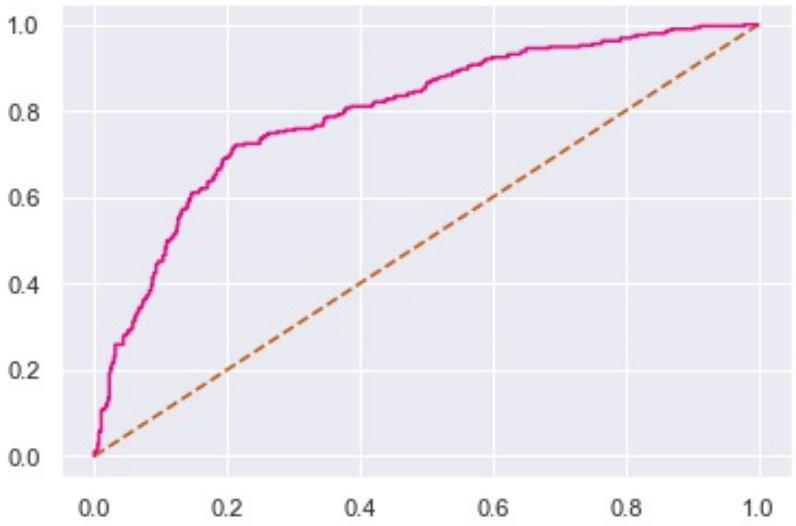
	precision	recall	f1-score	support
0	0.79	0.87	0.83	569
1	0.69	0.55	0.61	290
accuracy	0.76	0.76	0.76	1
macro avg	0.74	0.71	0.72	859
weighted avg	0.76	0.76	0.76	859

Classification Report of TEST Data

- ROC CURVE OF TRAIN DATA (AUC = 86.6%)



- ROC CURVE OF TEST DATA (AUC = 79.6%)



- Training and Test set results are almost similar - hence consistent
- Overall numbers could be better - Recall for 1(Claimed) could be better
- This is a fairly good model to predict if Tourists would Claim or Not Claim Insurance
- Agency_Code i.e which Agency books the tour, again matters the most - This is the most important variable for prediction
- Product_Name & Sales are also close 2nd and 3rd most Important Feature.
- This is clearly a better model than CART

ARTIFICIAL NEURAL NETWORK

- TRAIN DATA:

  AUC: 82.9%

  Accuracy: 78.87%

  Precision: 66%

  Recall: 62%

  f1-Score: 64%

- TEST DATA:

  AUC: 79.8%

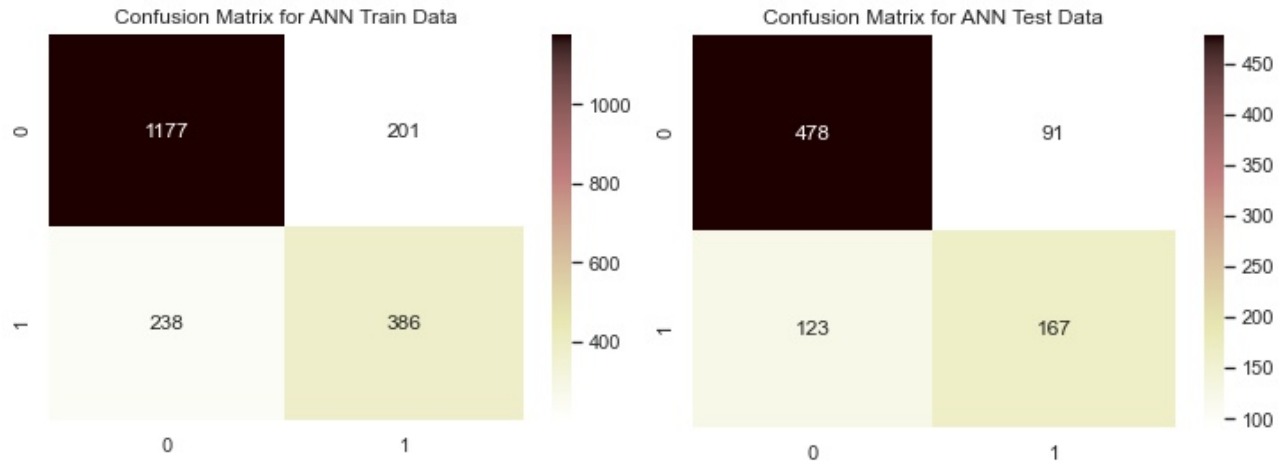
  Accuracy: 75.09%

  Precision: 65%

  Recall: 58%

  f1-Score: 61%

- CONFUSION MATRIX



- CLASSIFICATION REPORT

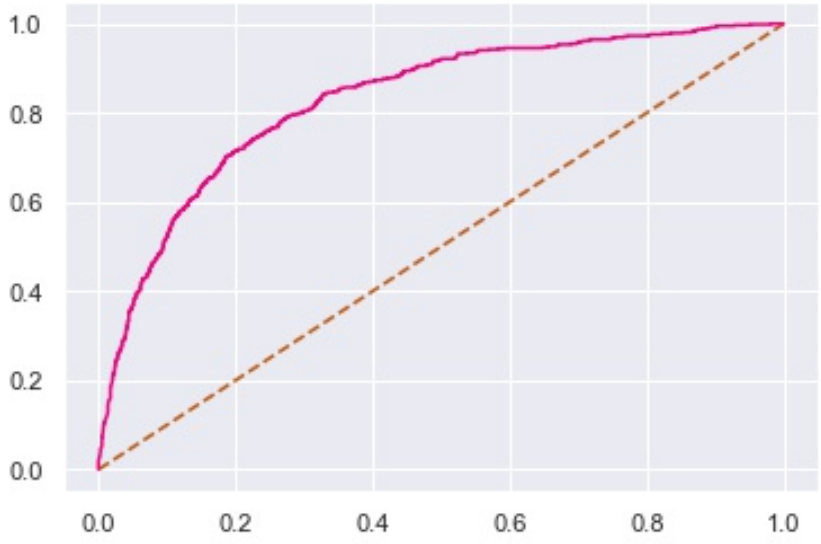
	precision	recall	f1-score	support
0	0.83	0.85	0.84	1378
1	0.66	0.62	0.64	624
accuracy	0.78	0.78	0.78	1
macro avg	0.74	0.74	0.74	2002
weighted avg	0.78	0.78	0.78	2002

Classification Report of TRAIN data

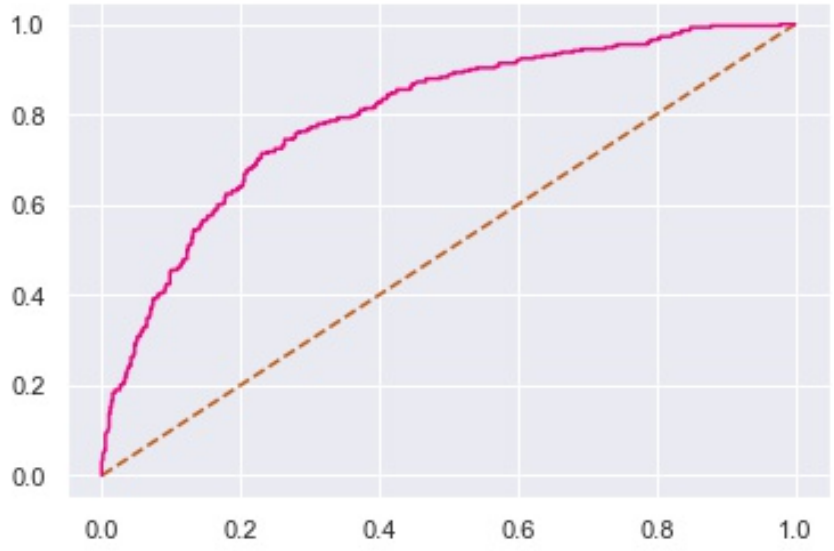
	precision	recall	f1-score	support
0	0.80	0.84	0.82	569
1	0.65	0.58	0.61	290
accuracy	0.75	0.75	0.75	1
macro avg	0.72	0.71	0.71	859
weighted avg	0.75	0.75	0.75	859

Classification Report of TEST data

- ROC CURVE OF TRAIN DATA (AUC = 82.9%)



- ROC CURVE OF TEST DATA (AUC = 79.8%)



- Training and Test set results are almost similar - hence consistent
- Overall numbers could be better - Recall for 1(Claimed) could be better
- This is a fairly good model to predict if Tourists would Claim or Not Claim Insurance

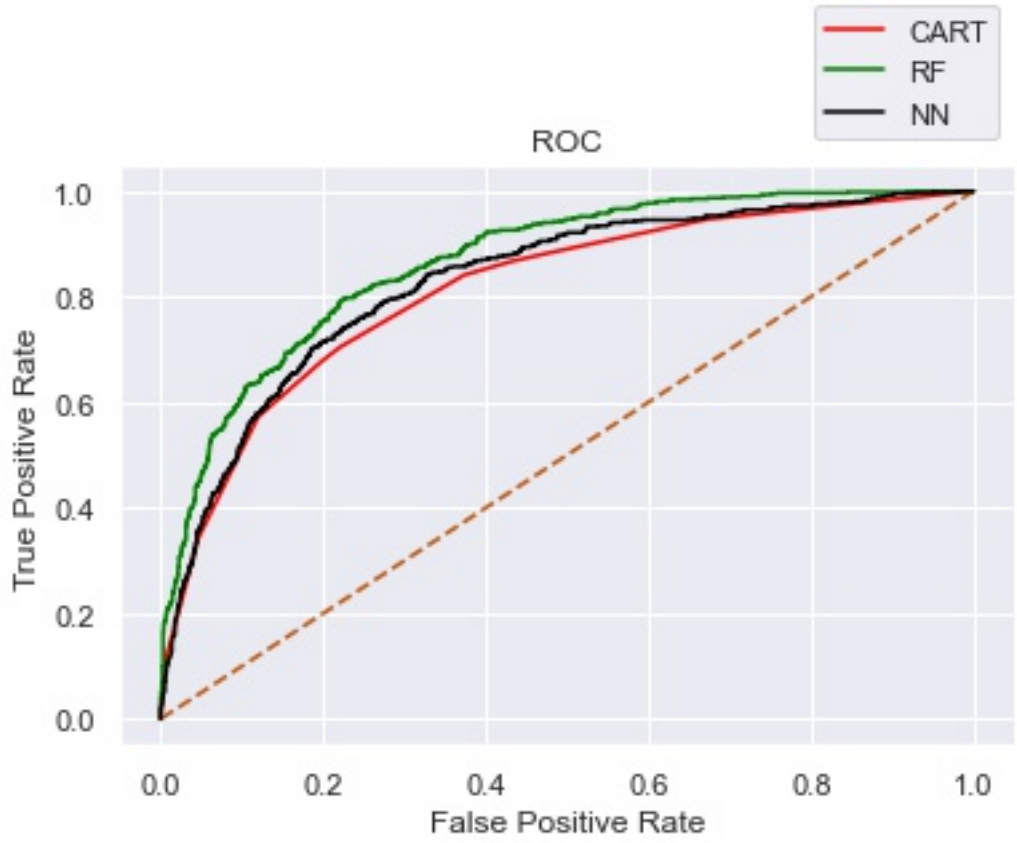
[Q 2.4] Final Model: Compare all the model and write an inference which model is best / optimized.

- PERFORMANCE METRICS OF ALL 3 MODELS -

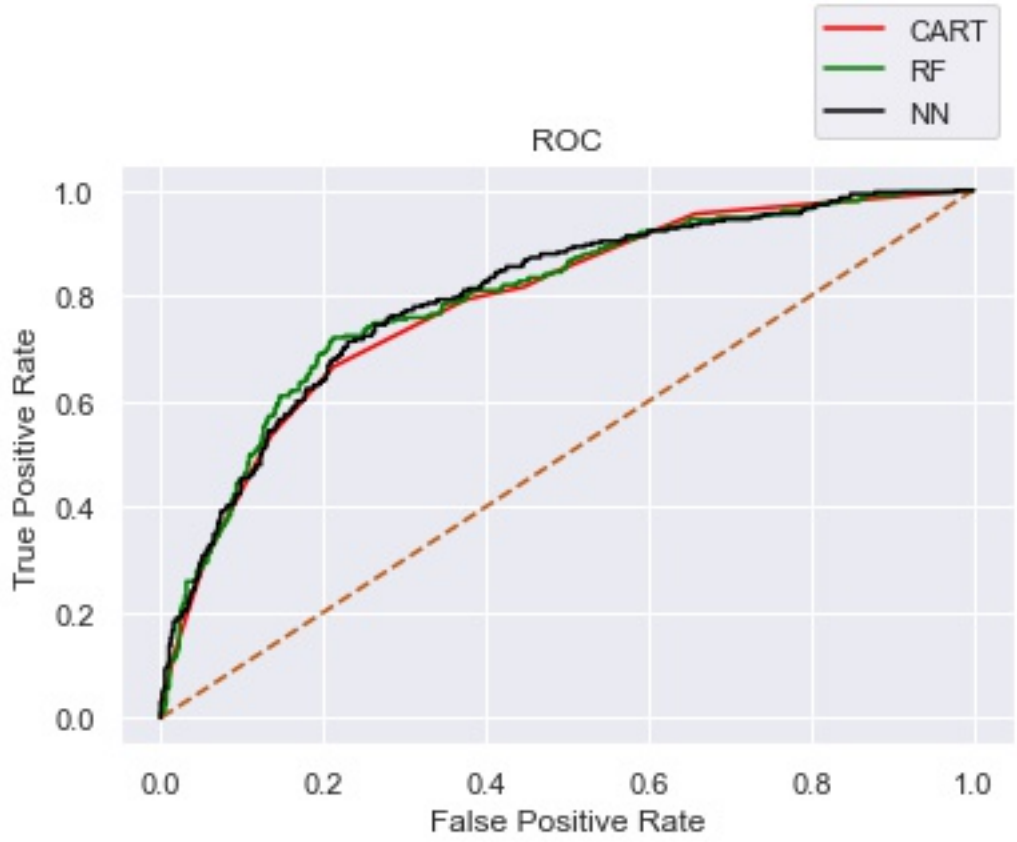
	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.784	0.753	0.810	0.764	**0.781**	0.751
AUC	0.813	0.787	0.866	0.796	**0.829**	0.798
Recall	0.510	0.490	0.600	0.550	**0.620**	0.580
Precision	0.690	0.670	0.740	0.690	**0.660**	0.650
F1 Score	0.590	0.570	0.660	0.610	**0.640**	0.610

Performance Metrics of All 3 Models

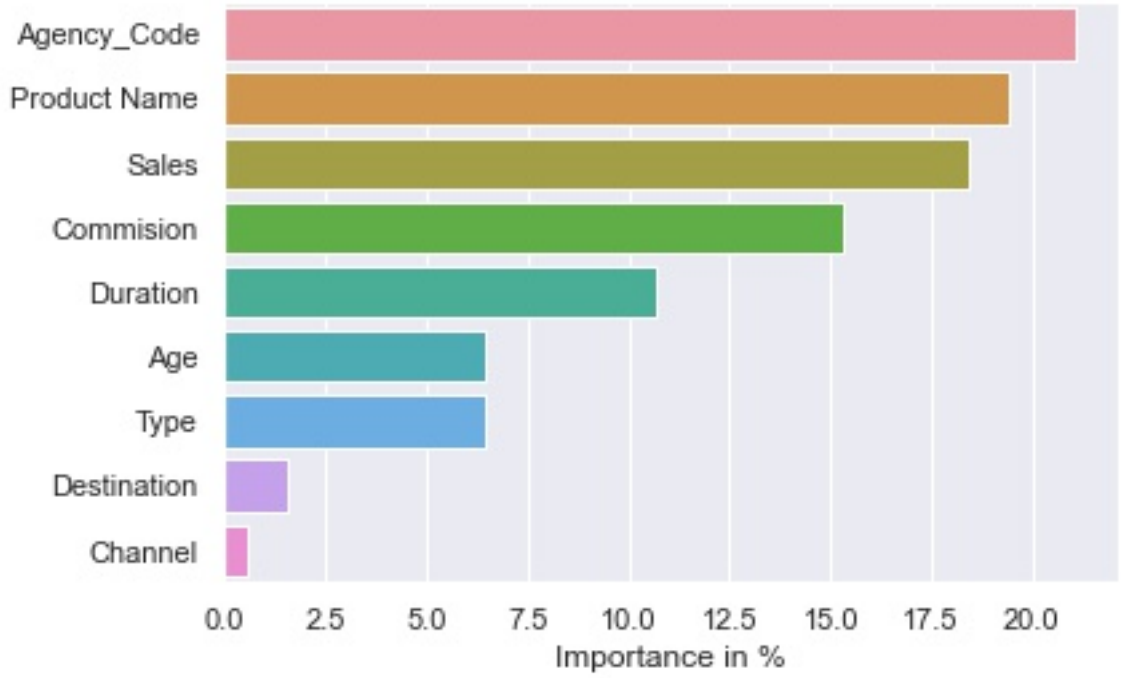
- ROC CURVE OF ALL 3 MODELS ON TRAIN DATA -



- ROC CURVE OF ALL 3 MODELS ON TEST DATA -



- INFERENCE OF ALL 3 MODELS
    - Out of the 3 models, Random Forest has slightly better performance than the CART and Neural network model
    - **Overall all the 3 models are reasonably stable enough to be used for making any future predictions.**
    - **But, because of the simplicity and its ability to extract Feature Importance - WE SHOULD PREFER TO USE RANDOM FOREST MODEL FOR FUTURE PREDICTIONS**
    - Feature Importance chart of RANDOM FOREST -

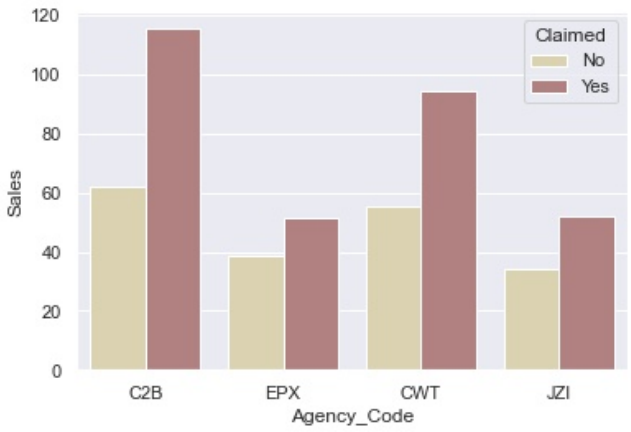


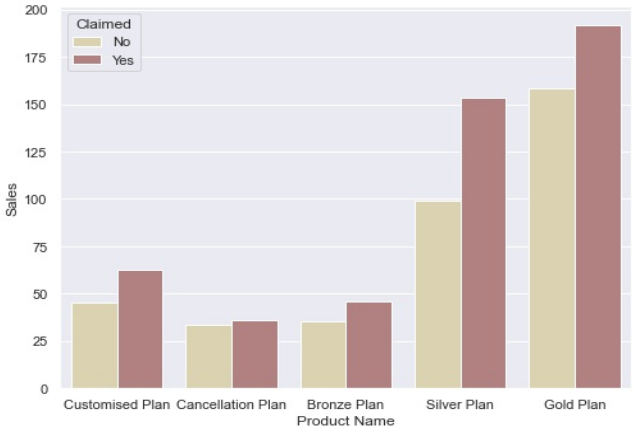
    - From Feature Importance chart above, it is evident that the variable '**Agency_Code**' is found to be the **most useful feature** amongst all other features for predicting if a person has claimed Insurance or not.
    - Variables '**Product Name**' and '**Sales**' are close 2nd and 3rd

        - showing they along with '**Agency_Code**' play very important role in determining whether a tourist would claim insurance or not

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

- Top 3 features which account for 60% of importance in Predictions are -
    - Agency - Tour Firm
    - Product Name - Insurance plan chosen
    - Sales - Sales of Tour Insurance Policies
- It is important that Company focusses on these 3 to improve claim ratio
- Note the charts below -





- Its clear from the charts above that -
    - Agencies - C2B & CWT and
    - Products - Gold & Silver plans

  They have maximum Sales but also very high claims
- These 2 agencies should be audited for any malpractice or frauds
- Their clienteles may include high risk individuals - More information and details about their clients should be sourced
- Agencies should be encouraged to profile their high risk customers and customise plans for them accordingly
- Insurance firm should restructure their Gold Plan - maybe with increased premiums
- Agencies selling Gold and Silver plans should have an added layer of Incentives for every ‘No Claim’ per plan sold

— — — END OF PROJECT — — —