Data Mining Project

Nandha Keshore Utti

PG-DSBA Online

March' 22

Date: 31/07/2022

Contents

Problem 1 (Clustering) 5

1.1.Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). ….. 5

1.2. Do you think scaling is necessary for clustering in this case? Justify. ………... 11

1.3.Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. 13

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. ………………….. 15

1.5.Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. ………………………………………………………17

Problem 2 (CART-RF-ANN) 20

2.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). ……...……………..…………... 20

2.2. Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network ..……………………………………… 30

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 …..………………………34

2.4. Final Model: Compare all the models and write an inference which model is best/optimized. ……………………………………………………………………….40

2.5. Inference: Based on the whole Analysis, what are the business insights and recommendations. ……………………………………………………………………..41

List of figures

Fig.1 – Problem-1: Hist plot of all numeric variables ……….…………………….…………..… 7
Fig.2 – Problem-1: Box plot of all numeric variables ……………………………………… 8
Fig.3 – Problem-1: Pair plot of all numeric variables ……….…………………….………… 10
Fig.4 – Problem-1: Heat map of all numeric variables …………..……………………….. 11
Fig.5 – Problem-1: Complete dendogram ………………..….… 17
Fig.6 – Problem-1: Truncated (at last 10) dendogram ……….......……….. 18
Fig.7 – Problem-1: Array of assigned fclusters for all the records ………..………... 19
Fig.8 – Problem-1: Scatter plot of 3 fclusters ……………….……........……….... 19
Fig.9 – Problem-1: k-means WSS inertia values ……….………........…………. 20
Fig.10 – Problem-1: Elbow curve plot ……….………........…………. 20
Fig.11 – Problem-1: Array of assigned k-means clusters for all the records …………. 20
Fig.12 – Problem-1: Scatter plot of 3 k-means clusters …………. 20
Fig.13 – Problem-1: List of silhouette score values ……………………. 20
Fig.14 – Problem-1: Bar diagram of silhouette scores ……………………. 20
Fig.15 – Problem-2: Hist plot of all numeric variables …………………….23
Fig.16 – Problem-2: Box plot of all numeric variables before outlier treatment …….23,24
Fig.17 – Problem-2: Box plot of all numeric variables after outlier treatment ………….24
Fig.18 – Problem-2: Pair plot of all numeric variables …………………….25
Fig.19 – Problem-2: Heatmap of all numeric variables …………………….26
Fig.20 – Problem-2: Box plot of ‘Agency code’ vs ‘Sales’ …………………….27
Fig.21 – Problem-2: Box plot of ‘Type’ vs ‘Sales’ …………………….27
Fig.22 – Problem-2: Box plot of ‘Channel’ vs ‘Sales’ …………………….28
Fig.23 – Problem-2: Box plot of ‘Destination’ vs ‘Sales’ …………………….28
Fig.24 – Problem-2: Box plot of ‘Destination’ vs ‘Duration’ …………………….29
Fig.25 – Problem-2: Scatter plot of sales vs commission on different products ….29
Fig.26 – Problem-2: Train and Test data set shapes ……………………..30
Fig.27 – Problem-2: Root node of main decision tree diagram………………………….31
Fig.28 – Problem-2: Set parameters for GridSearch of CART model ……………………31
Fig.29 – Problem-2: Best parameters obtained for GridSearch of CART model ……..31
Fig.30 – Problem-2: Set parameters for RF model ……..323
Fig.31 – Problem-2: Set parameters for GridSearch of RF model ……………………..32
Fig.32 – Problem-2: Best parameters obtained for GridSearch of RF model ……..32
Fig.33 – Problem-2: Train scaled data ……..33
Fig.34 – Problem-2: Test scaled data ……..333
Fig.35 – Problem-2: Set parameters for ANNf model ……..33
Fig.36 – Problem-2: Set parameters for GridSearch of ANN model ……………………34
Fig.37 – Problem-2: Best parameters obtained for GridSearch of ANN model ……..34
Fig.38 – Problem-2: Confusion matrix of train data set from CART model ……..34
Fig.39 – Problem-2: ROC curve of train data set from CART model ……..35
Fig.40 – Problem-2: Classification report of train data set from CART model ……..35
Fig.41 – Problem-2: Confusion matrix of test data set from CART model ……..35
Fig.42 – Problem-2: ROC curve of test data set from CART model ……..36
Fig.43 – Problem-2: Classification report of test data set from CART model ……..36
Fig.44 – Problem-2: Confusion matrix of train data set from RF model ……..36
Fig.45 – Problem-2: ROC curve of train data set from RF model ……..37
Fig.46 – Problem-2: Classification report of train data set from RF model ……..37
Fig.47 – Problem-2: Confusion matrix of test data set from RF model ……..37

Fig.48 – Problem-2: ROC curve of test data set from RF model ……..38
Fig.49 – Problem-2: Classification report of test data set from RF model ……..38
Fig.50 – Problem-2: Confusion matrix of train data set from ANN model ……..38
Fig.51 – Problem-2: ROC curve of train data set from ANN model ……..39
Fig.52 – Problem-2: Classification report of train data set from ANN model ……..39
Fig.53 – Problem-2: Confusion matrix of test data set from ANN model ……..39
Fig.54 – Problem-2: ROC curve of test data set from ANN model ……..40
Fig.55 – Problem-2: Classification report of test data set from ANN model ……..40

List of tables

Table.1 – Problem-1: Data loaded with first five records ……………………………. 5
Table.2 – Problem-1: Data information table ………..…………………………….... 5
Table.3 – Problem-1: Data types table ………………………..…………………….... 6
Table.4 – Problem-1: Data description table ………………..……………................6
Table.5 – Problem-1: Skew values table ….………………………………………..... 7
Table.6 – Problem-1: Outlier records of ‘probability_of_full_payment’ ……………. 9
Table.7 – Problem-1: Outlier records of ‘min_payment_amt’ ……………………… 9
Table.8 – Problem-1: Scaled dataset ………………………………………………... 12
Table.9 – Problem-1: Data frame added with k-means clusters column ……………… 16
Table.10 – Problem-1: Description table cluster-1 customers ……………………........ 18
Table.11 – Problem-1: Description table cluster-2 customers …………………... 18
Table.12 – Problem-1: Description table cluster-3 customers ….…………… 18
Table.13 – Problem-2: Data loaded with first five records …………..……... 20
Table.14 – Problem-2: Data types table …………………..…….……... 20
Table.15 – Problem-2: Data information table ……………..… 21
Table.16 – Problem-2: Data description table ……………………………….…… 21
Table.17 – Problem-2: Data frame of duplicated records ……………………………..…… 22
Table.18 – Problem-2: Skew values table ……………………..…… 23
Table.19 – Problem-2: Data frame after data encoding ……………………… 30
Table.20 – Problem-2: Feature importances of CART model ……………………… 31
Table.21 – Problem-2: Feature importances of RF model ……………………… 32
Table.22 – Problem-2: First 5 sample probabilities obtained using CART model ………… 34
Table.23 – Problem-2: First 5 sample probabilities obtained using RF model … 36
Table.24 – Problem-2: First 5 sample probabilities obtained using ANN model ……… 38
Table.25 – Problem-2: Comparison of all parameters of all three models …………….. 40

Problem 1 (Clustering)

Problem Statement:

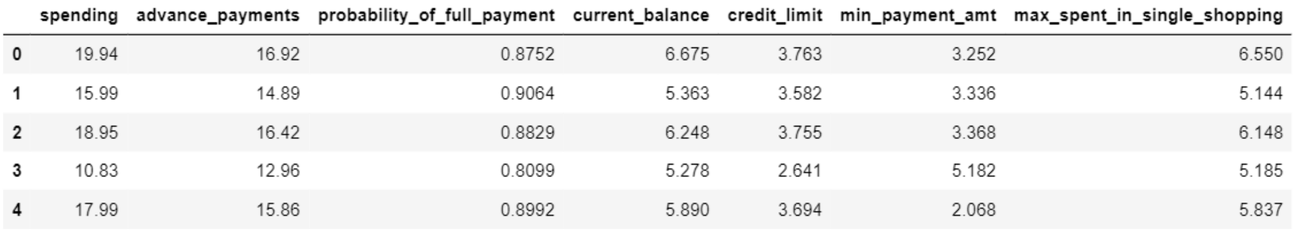
A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

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

Exploratory Data Analysis:

- **Data description:**

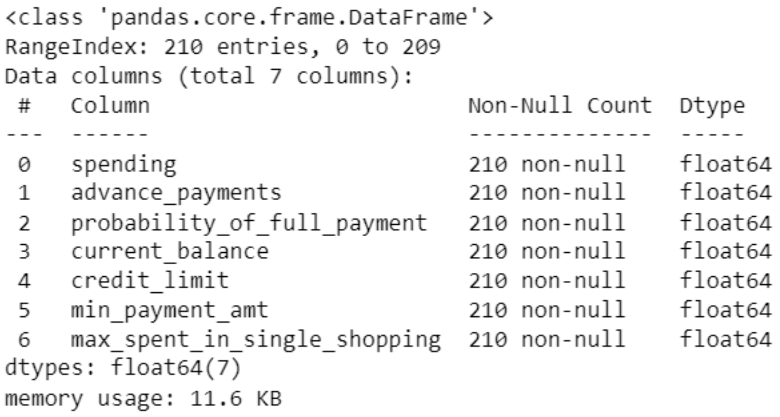
Reading the data file and loading first five records:



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

Table. 01

Dataset information:

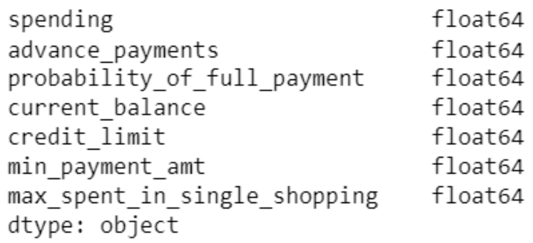


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              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. 02

- There are no null values.
- Total 210 records and 7 features are in the given dataset.
- There are no duplicated records.

Data types:

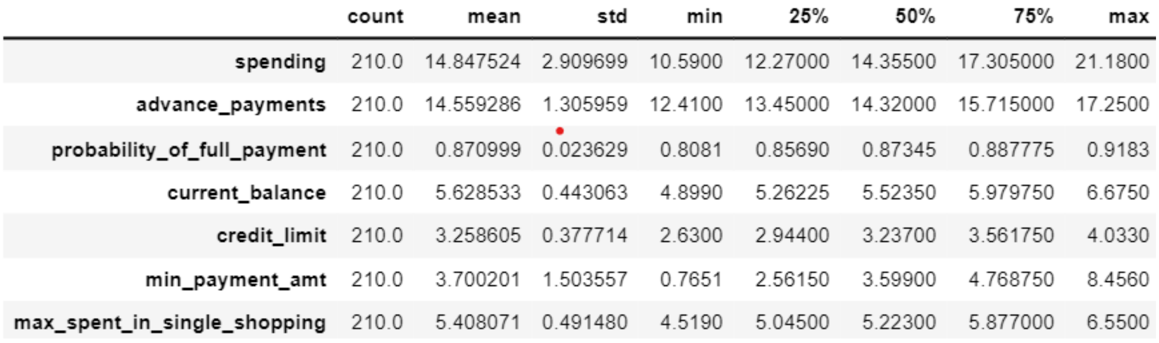


```
spending                        float64
advance_payments                float64
probability_of_full_payment     float64
current_balance                 float64
credit_limit                    float64
min_payment_amt                 float64
max_spent_in_single_shopping    float64
dtype: object
```

Table.3

- All the features are numeric, no object type variable is present.

Dataset description:



	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

Table.4

Insights:

- Average spending is ~14.8 (1000s) per month by the customers
- Average credit_limit is ~3.25 (10000s) for all 210 customers
- Average percentage of spendings per month for the available credit limit for the 210 customers is 45.56405239168971
- On an average, 87.09% of customers are doing the full payment
- It shows that average current balance is less on comparing to the average credit_limit, which indicates customers' spendings are more
- Average percentage of current_balance comparing to average credit_limit is 17.272832222964258
- Maximum amount spent by a customer in single shopping is 6.55 (in 1000s)
- On an average, minimum amount paid by customers while making purchases made monthly is ~3.70 (100s)
- On an average, ~14.55 (100s) amount paid by the customers in advance by cash
- Variation in the data is very minimum for all the features

- **Data pre-processing:**

- There are no null values.
- All the features are numeric, no categorical variable present
- Total 210 records and 7 features are there in the given dataset

- There are no duplicated records in the dataset
- Anamolies are not observed in the dataset

- ➢ **Data visualization:**

Univariate analysis:

- Let’s visualize all the numeric columns using hist plot and check the distribution natur e of the features.

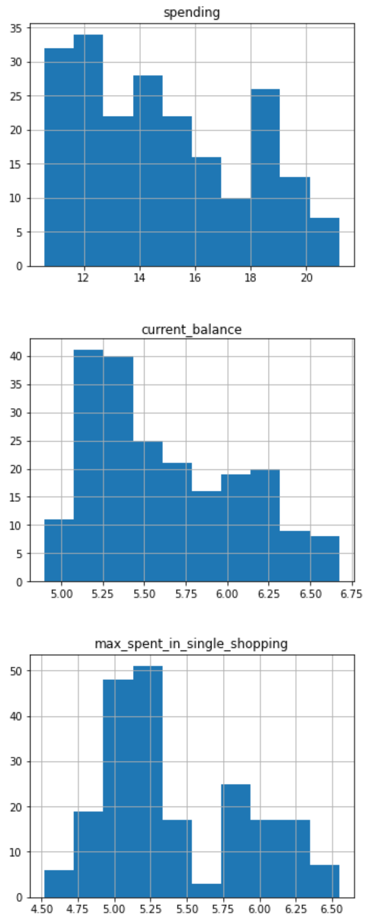
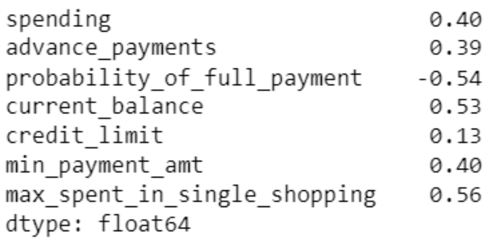


Fig. 01

Checking skewness:



```
spending                         0.40
advance_payments                 0.39
probability_of_full_payment     -0.54
current_balance                  0.53
credit_limit                     0.13
min_payment_amt                  0.40
max_spent_in_single_shopping     0.56
dtype: float64
```

Table. 05

Interpretations:

- 'Spending', 'advance_payments', 'credit_limit', 'min_payment_amt' are normally distributed features
- 'probability_of_full_payment' is slightly left skewed distribution
- 'current_balance', 'max_spent_in_single_shopping' are slightly right skewed distributions

Checking for outliers and its porportions:

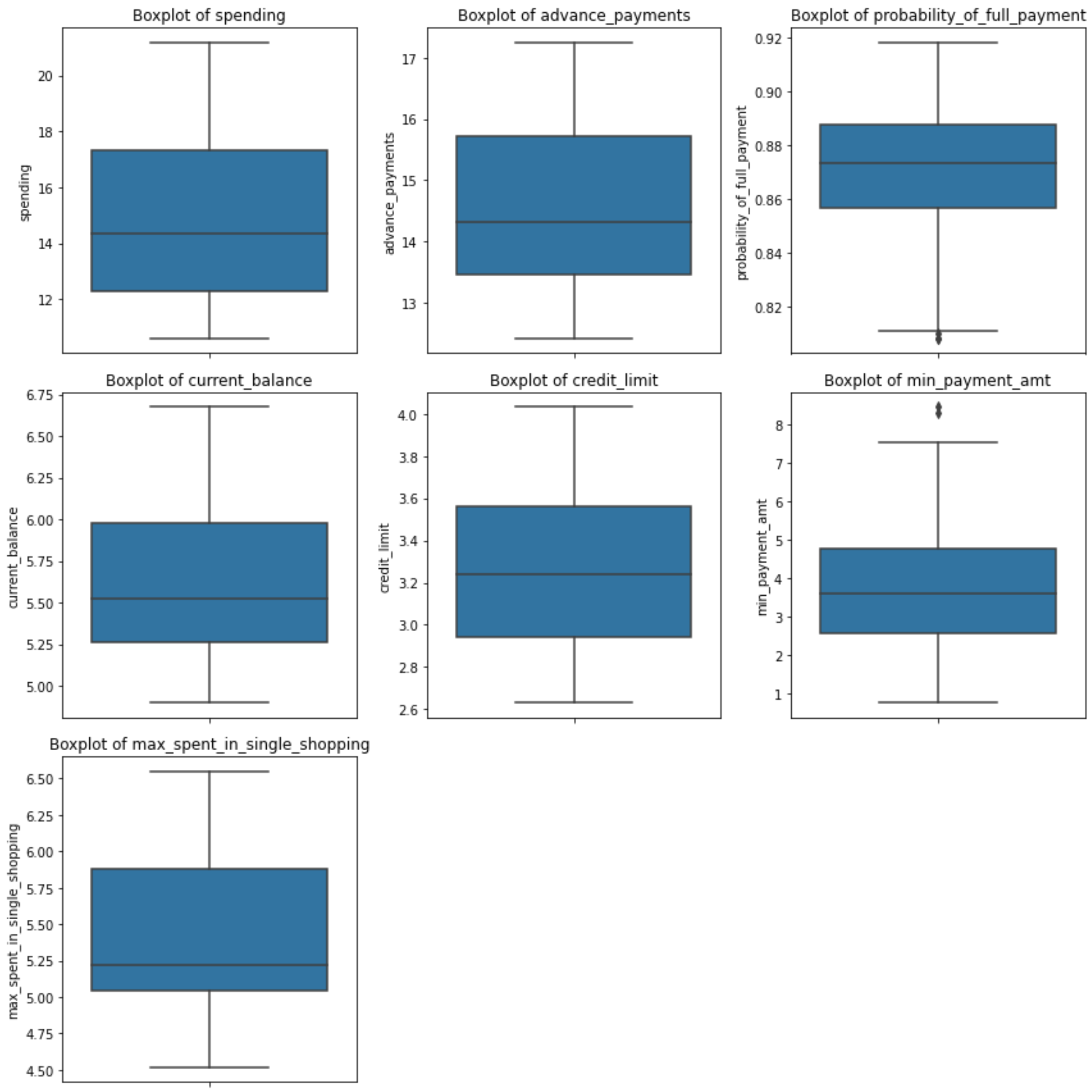


Fig. 02

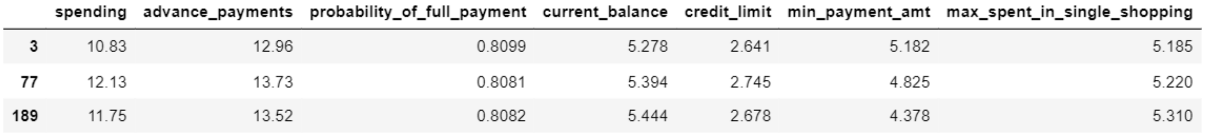
Interpretations:

- Outliers are absent in all the features except for 'probability_of_full_payment', 'min_payment_amt' with very minimum outliers.

- 'probability_of_full_payment' has outliers in lower range and 'min_payment_amt' has outliers in upper range
- Let us check how much proportion outliers are present in 'probability_of_full_payment', 'min_payment_amt'.

'probability_of_full_payment':

- There are 3 outlier records out of 210 records in 'probability_of_full_payment' feature. Outlier proportion for 'probability_of_full_payment' feature is 1.43%

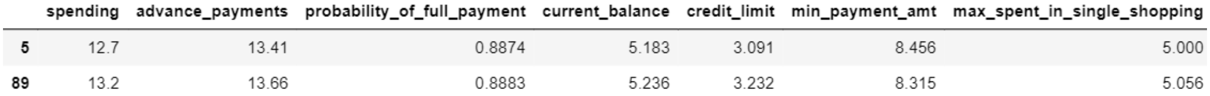


	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
77	12.13	13.73	0.8081	5.394	2.745	4.825	5.220
189	11.75	13.52	0.8082	5.444	2.678	4.378	5.310

Table. 06

'min_payment_amt':

- There are 2 outlier records out of 210 records in 'probability_of_full_payment' feature. Outlier proportion for 'min_payment_amt' feature is 0.95%



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
5	12.7	13.41	0.8874	5.183	3.091	8.456	5.000
89	13.2	13.66	0.8883	5.236	3.232	8.315	5.056

Table. 07

- As we can see, there is very less proportion of outliers in the above discussed two features. So, we can skip the outlier treatment for this dataset.

Bivariate analysis:

- Let's plot the pair plot and heatmap to check correlation b/w the data features

Pair plot:

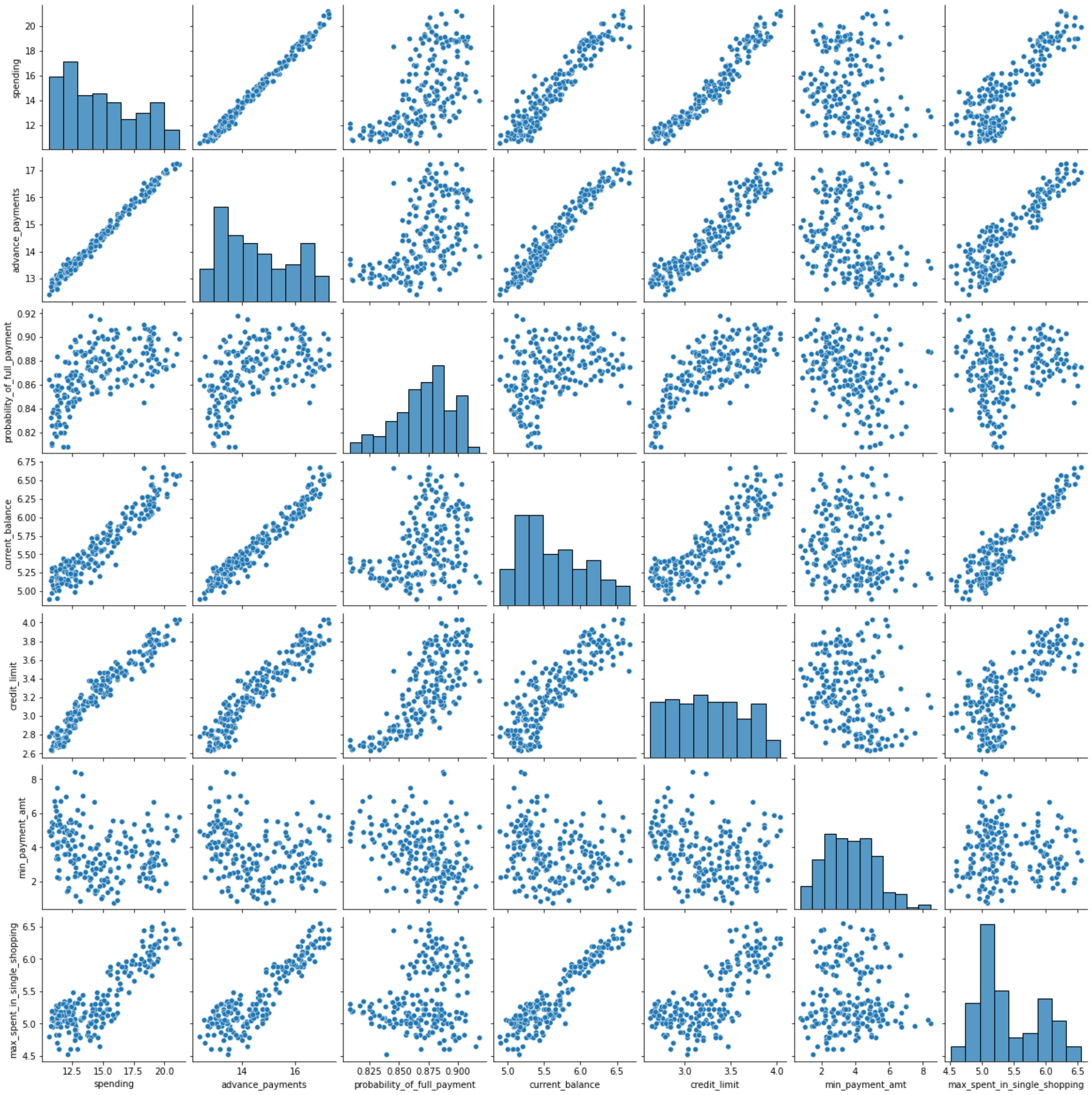


Fig.3

Heatmap:

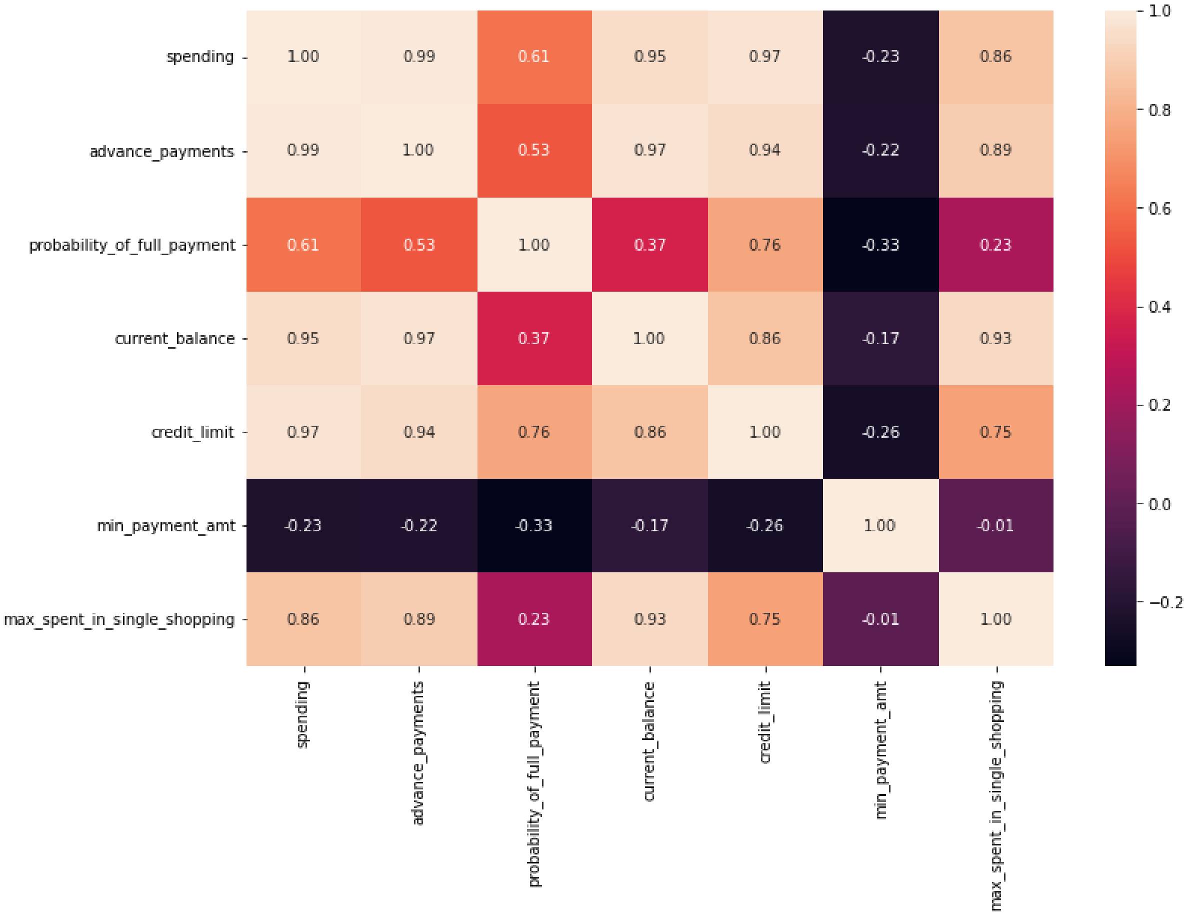


Fig.4

Insignts (From both pairplot and heatmap):

There is very good correlation between all the features except for below mentioned pairs:

- **‘min_payment_amt’** has negative correlation with the all features

- **'probability_of_full_payment'** is,

  -Moderately correlated with 'spending', 'advance_payments', 'credit_limit',
  -Weakly correlated with 'current_balance', 'max_spent_in_single_shopping'
  -Negatively correlated with 'min_payment_amt'

Multivariate analysis:

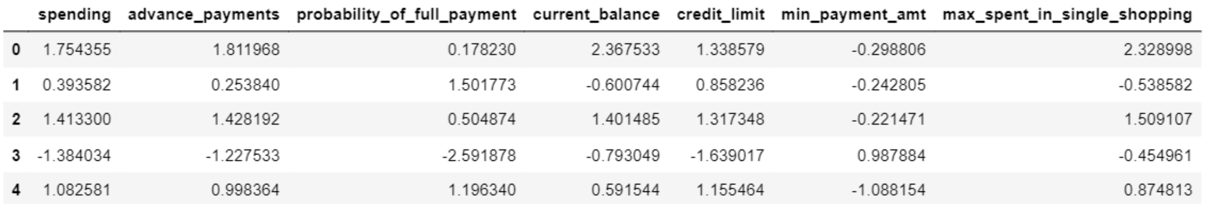
- Multivariate analysis is not possible for given dataset as we do not have any object ty pe variable

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

- Scaling is necessary for clustering in this case because all of the numerical features are not on same weight. So, we need to transform the features onto same scale.

For example,

- Features like 'spending', 'current_balance', 'max_spent_in_single_shopping' are on 1000s scale
- Features like 'advance_payments', 'min_payment_amt' are on 100s scale
- credit_limit is on 10000s scale
- 'probability_of_full_payment' feature is in range of 0 to 1
- So, we should do the scaling for this dataset.
- Let us scale the data using 'z-scale' method.
- This method scales the data in such a way that the mean value of the features tends to 0 and the standard deviation tends to 1.



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

Table. 08

- Dataset has been scaled successfully.

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

- Hierarchical clustering is applied on the scaled dataset.
- Clusters are created between the variables ‘spending’ and ‘current balance’ with a consideration that based on the spending and balance available, business can identify the customer profile from segmentation perspective.
- Let’s create a dendrogram using the scaled data and ‘ward-linkage’ method.

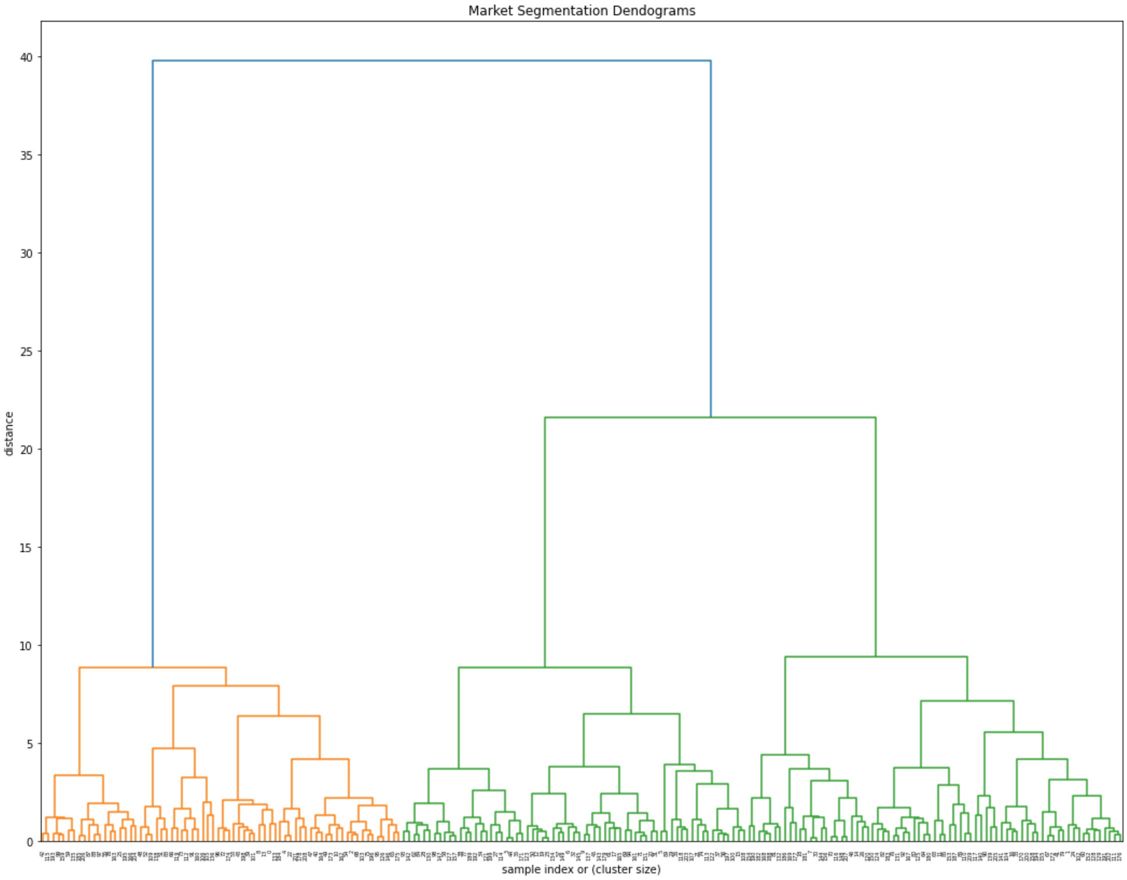


Fig. 05

- Let’s observe the dendogram by truncating the distance at 10.

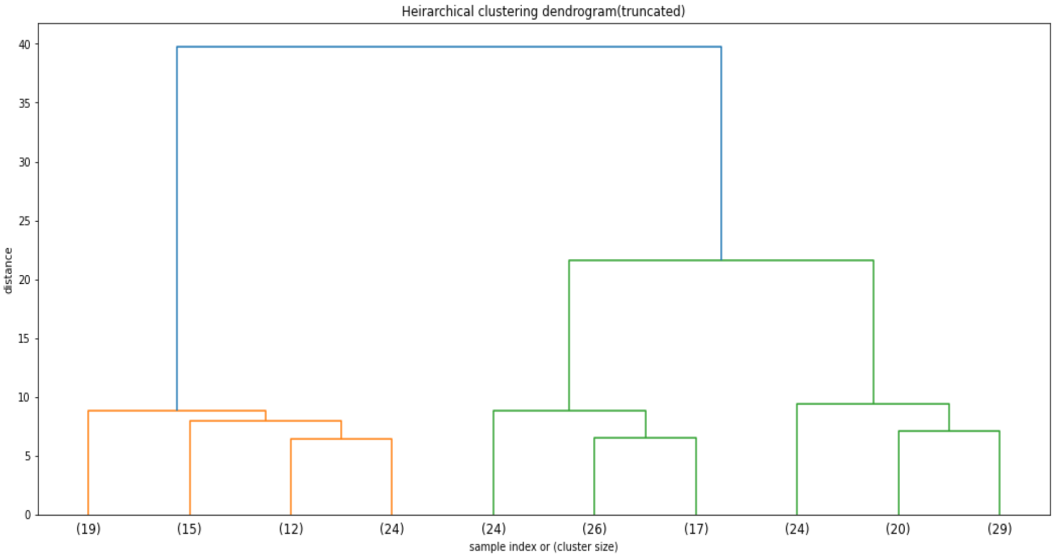
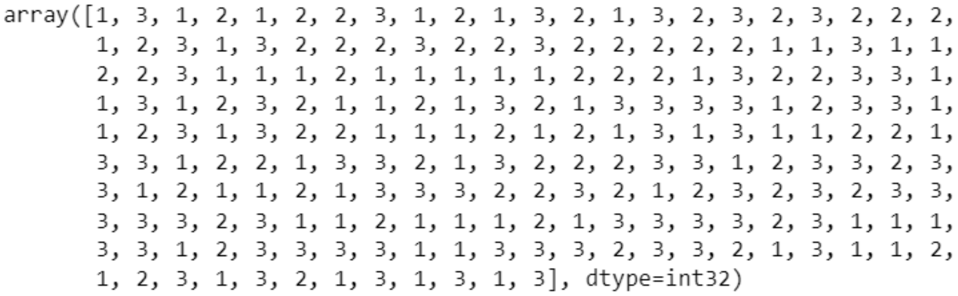


Fig. 06

- However as normally to understand the dataset, normally 2 clusters are not preferred because in most of the cases, business is already aware about the 2 classes in the dataset and hence to generate some more insights, segmentation with more than 2 clusters is preferred.

- As an example, for a bank dataset, bank would like to know more than ‘good’ and ‘not so good’ customers and hence more insight we are able to generate with more than 2 clusters, better it is for business.
- Hence let’s consider 3 clusters and plot the clusters to confirm if the desired clusters are providing the required segmentation details.
- Let us use SciPy’s F-cluster method:



```
array([1, 3, 1, 2, 1, 2, 2, 3, 1, 2, 1, 3, 2, 1, 3, 2, 3, 2, 3, 2, 2, 2,
       1, 2, 3, 1, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 1, 1, 3, 1, 1,
       2, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 3, 2, 2, 3, 3, 1,
       1, 3, 1, 2, 3, 2, 1, 1, 2, 1, 3, 2, 1, 3, 3, 3, 3, 1, 2, 3, 3, 1,
       1, 2, 3, 1, 3, 2, 2, 1, 1, 1, 2, 1, 2, 1, 3, 1, 3, 1, 1, 2, 2, 1,
       3, 3, 1, 2, 2, 1, 3, 3, 2, 1, 3, 2, 2, 2, 3, 3, 1, 2, 3, 3, 2, 3,
       3, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 3, 2, 1, 2, 3, 2, 3, 2, 3, 3,
       3, 3, 3, 2, 3, 1, 1, 2, 1, 1, 1, 2, 1, 3, 3, 3, 3, 2, 3, 1, 1, 1,
       3, 3, 1, 2, 3, 3, 3, 3, 1, 1, 3, 3, 3, 2, 3, 3, 2, 1, 3, 1, 1, 2,
       1, 2, 3, 1, 3, 2, 1, 3, 1, 3, 1, 3], dtype=int32)
```

Fig.07

- We can see, all records are allotted to 3 different cluster.

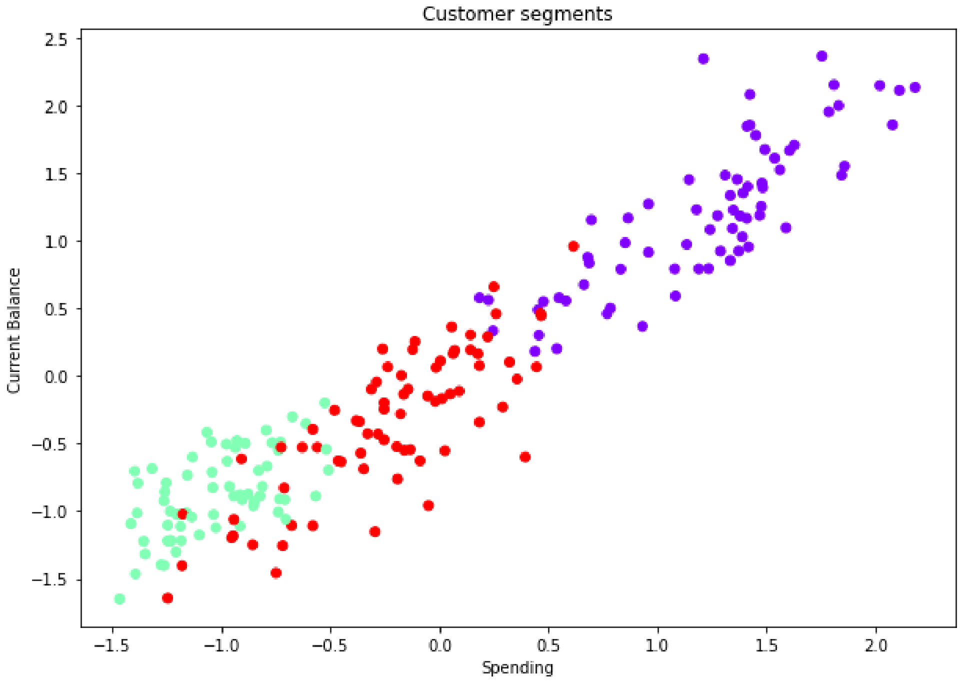
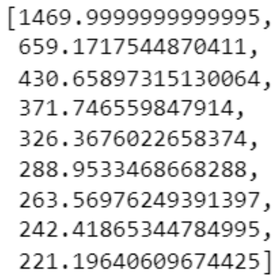


Fig. 08

- From above scatter plot, we can also see visually 3 cluster segmentation.

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.

- Let’s perform k-means clustering on the scaled data.
- Let’s generate k-means inertia values from 2 to 10 clusters. These values are as shown below:



```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.746559847914,
 326.3676022658374,
 288.9533468668288,
 263.56976249391397,
 242.41865344784995,
 221.19640609674425]
```

Fig. 09

- **Elbow curve:**

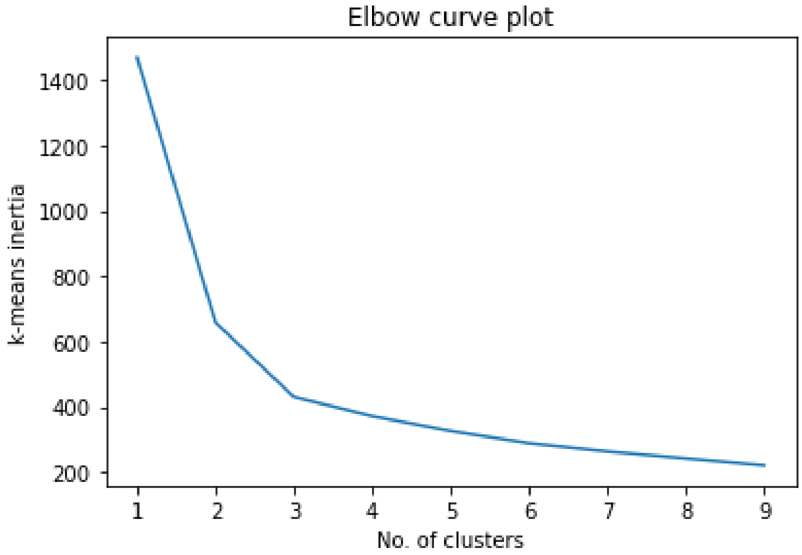
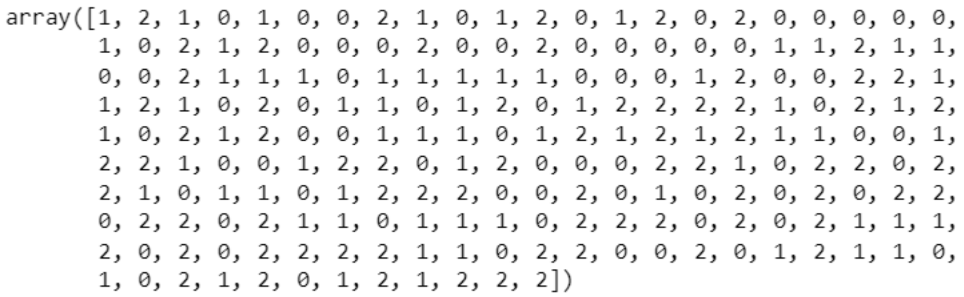


Fig. 10

- From both elbow curve and k-means inertia methods, we can choose 3 clusters as optimal number because there is reasonable inertia drop for both 2 & 3 cluster, but as discussed earlier 3 cluster segmentation study is efficient business study.
- Cluster segmentation using k-means method is as shown below:



```
array([1, 2, 1, 0, 1, 0, 0, 2, 1, 0, 1, 2, 0, 1, 2, 0, 2, 0, 0, 0, 0, 0,
       1, 0, 2, 1, 2, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 1, 1, 2, 1, 1,
       0, 0, 2, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 2, 0, 0, 2, 2, 1,
       1, 2, 1, 0, 2, 0, 1, 1, 0, 1, 2, 0, 1, 2, 2, 2, 2, 1, 0, 2, 1, 2,
       1, 0, 2, 1, 2, 0, 0, 1, 1, 1, 0, 1, 2, 1, 2, 1, 2, 1, 1, 0, 0, 1,
       2, 2, 1, 0, 0, 1, 2, 2, 0, 1, 2, 0, 0, 0, 2, 2, 1, 0, 2, 2, 0, 2,
       2, 1, 0, 1, 1, 0, 1, 2, 2, 2, 0, 0, 2, 0, 1, 0, 2, 0, 2, 0, 2, 2,
       0, 2, 2, 0, 2, 1, 1, 0, 1, 1, 1, 0, 2, 2, 2, 0, 2, 0, 2, 1, 1, 1,
       2, 0, 2, 0, 2, 2, 2, 2, 1, 1, 0, 2, 2, 0, 0, 2, 0, 1, 2, 1, 1, 0,
       1, 0, 2, 1, 2, 0, 1, 2, 1, 2, 2, 2])
```

Fig. 11

- Let's check this segmentation visually using scatter plot.

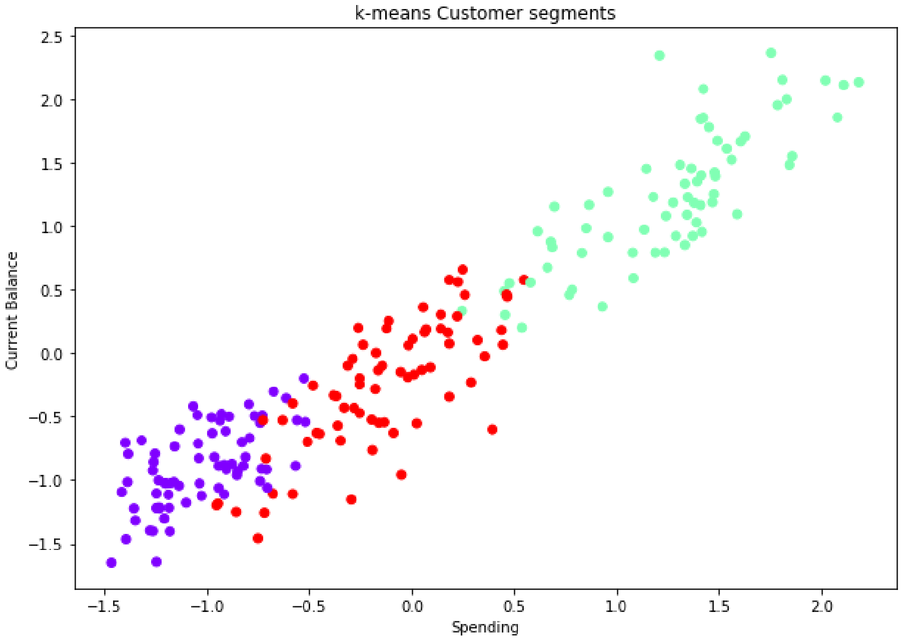
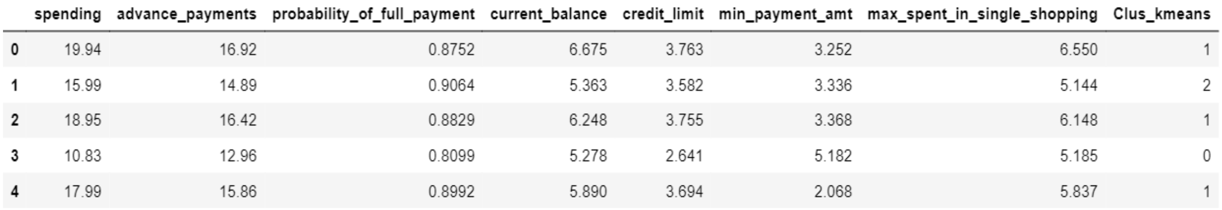


Fig. 12

- From above scatter plot, we can also see visually 3 cluster segmentation. And it also has better clustering than hierarchical f-cluster.
- So, let's analyse the data using k-means clusters.
- Let's create a column in original data frame assigning each record to that particular cluster.

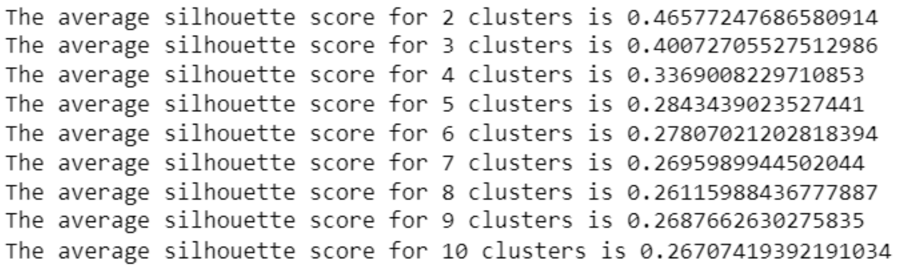


	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Clus_kmeans
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.8099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

Table. 09

Silhouette analysis:

- Let’s check the silhouette score for 2 to 10 clusters:



```
The average silhouette score for 2 clusters is 0.46577247686580914
The average silhouette score for 3 clusters is 0.40072705527512986
The average silhouette score for 4 clusters is 0.3369008229710853
The average silhouette score for 5 clusters is 0.2843439023527441
The average silhouette score for 6 clusters is 0.27807021202818394
The average silhouette score for 7 clusters is 0.2695989944502044
The average silhouette score for 8 clusters is 0.26115988436777887
The average silhouette score for 9 clusters is 0.2687662630275835
The average silhouette score for 10 clusters is 0.26707419392191034
```

Fig. 13

Silhouette score bar diagram:

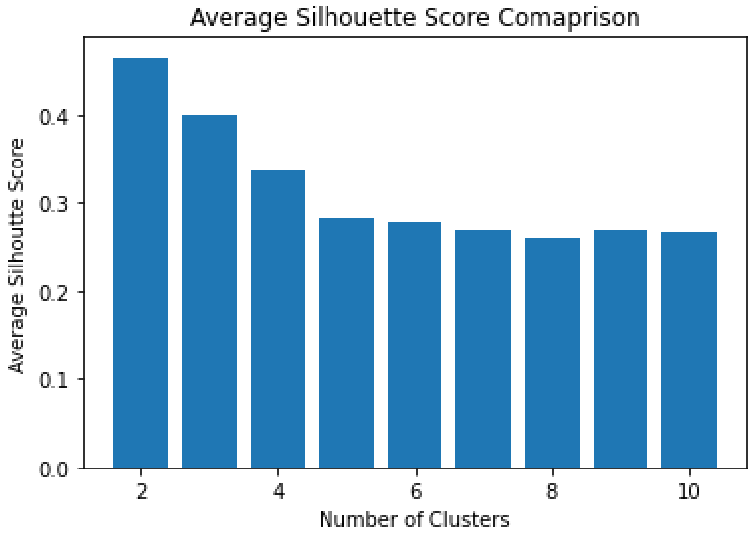


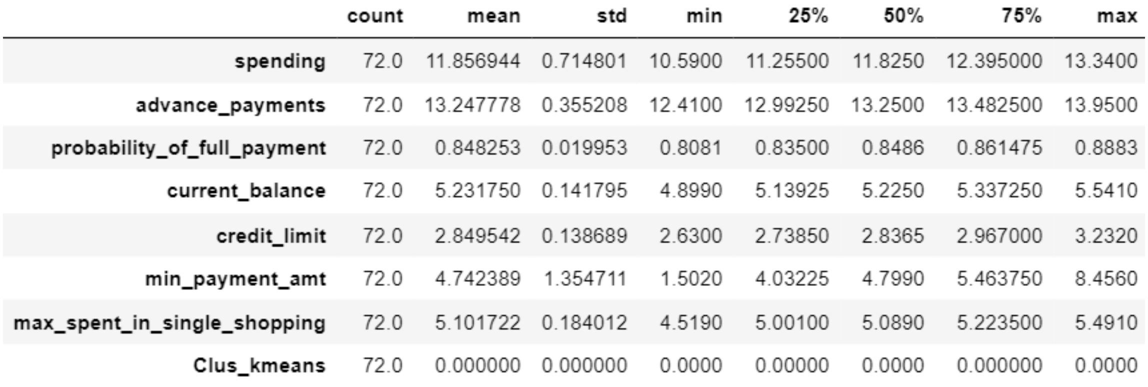
Fig. 14

- As per the silhouette score, optimal number of clusters are 2.
- However as mentioned earlier, 2 is not preferred way of profiling the dataset and hence the ideal number of clusters to be considered is 3 with a silhouette score of ~0.40.

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

- Let’s describe each cluster that we segmented in terms of ‘spendings’ and ‘current_balance’.

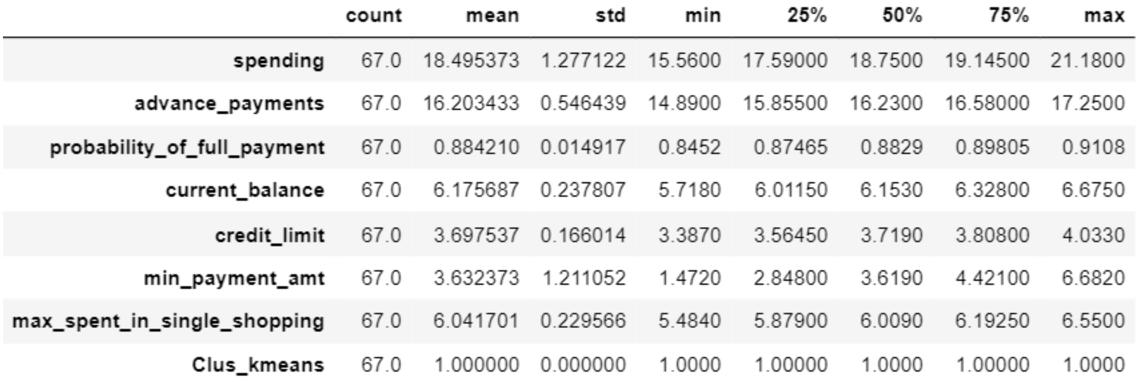
Cluster-1 customers:



	count	mean	std	min	25%	50%	75%	max
spending	72.0	11.856944	0.714801	10.5900	11.25500	11.8250	12.395000	13.3400
advance_payments	72.0	13.247778	0.355208	12.4100	12.99250	13.2500	13.482500	13.9500
probability_of_full_payment	72.0	0.848253	0.019953	0.8081	0.83500	0.8486	0.861475	0.8883
current_balance	72.0	5.231750	0.141795	4.8990	5.13925	5.2250	5.337250	5.5410
credit_limit	72.0	2.849542	0.138689	2.6300	2.73850	2.8365	2.967000	3.2320
min_payment_amt	72.0	4.742389	1.354711	1.5020	4.03225	4.7990	5.463750	8.4560
max_spent_in_single_shopping	72.0	5.101722	0.184012	4.5190	5.00100	5.0890	5.223500	5.4910
Clus_kmeans	72.0	0.000000	0.000000	0.0000	0.00000	0.0000	0.000000	0.0000

Table. 10

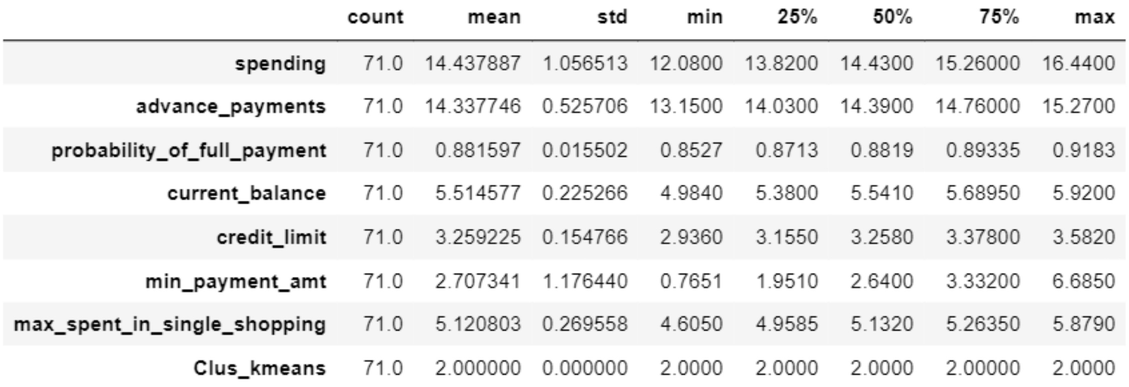
Cluster-2 customers:



	count	mean	std	min	25%	50%	75%	max
spending	67.0	18.495373	1.277122	15.5600	17.59000	18.7500	19.14500	21.1800
advance_payments	67.0	16.203433	0.546439	14.8900	15.85500	16.2300	16.58000	17.2500
probability_of_full_payment	67.0	0.884210	0.014917	0.8452	0.87465	0.8829	0.89805	0.9108
current_balance	67.0	6.175687	0.237807	5.7180	6.01150	6.1530	6.32800	6.6750
credit_limit	67.0	3.697537	0.166014	3.3870	3.56450	3.7190	3.80800	4.0330
min_payment_amt	67.0	3.632373	1.211052	1.4720	2.84800	3.6190	4.42100	6.6820
max_spent_in_single_shopping	67.0	6.041701	0.229566	5.4840	5.87900	6.0090	6.19250	6.5500
Clus_kmeans	67.0	1.000000	0.000000	1.0000	1.00000	1.0000	1.00000	1.0000

Table. 11

Cluster-3 customers:



	count	mean	std	min	25%	50%	75%	max
spending	71.0	14.437887	1.056513	12.0800	13.8200	14.4300	15.26000	16.4400
advance_payments	71.0	14.337746	0.525706	13.1500	14.0300	14.3900	14.76000	15.2700
probability_of_full_payment	71.0	0.881597	0.015502	0.8527	0.8713	0.8819	0.89335	0.9183
current_balance	71.0	5.514577	0.225266	4.9840	5.3800	5.5410	5.68950	5.9200
credit_limit	71.0	3.259225	0.154766	2.9360	3.1550	3.2580	3.37800	3.5820
min_payment_amt	71.0	2.707341	1.176440	0.7651	1.9510	2.6400	3.33200	6.6850
max_spent_in_single_shopping	71.0	5.120803	0.269558	4.6050	4.9585	5.1320	5.26350	5.8790
Clus_kmeans	71.0	2.000000	0.000000	2.0000	2.0000	2.0000	2.00000	2.0000

Table. 12

Promotional strategies:

- Based on the clusters obtained from k-means clustering strategy for different segment of customers is as follows:

For Type 1 Customers: (Represented by purple dots in the k-means cluster plot wherein both spending and current balance are low):

- If we compare the mean ‘current_balance’ of Type 1 customers (5.23) and Type 3 customers (5.51) (discussed below) are almost similar. But spendings vary for both. So, there are customers in this segment who are having the balance, but less interested in spendings. Banks should attract this particular customer segment with some kind of discounts, vouchers etc.

For Type 2 Customers: (Represented by Green Dots in the k-means cluster plot wherein both spending and current balance are high)

- This is probably a high value customer and hence special discounted pricing based promotional campaigns for this group to increase their spending and use current balance.

For Type 3 Customers: (Represented by Red Dots in the k-means cluster plot wherein both spending and current balance are moderate)

- Further analysis could be performed to understand what would make this segment to move into green area.

Problem 2 (CART-RF-ANN)

Problem statement:

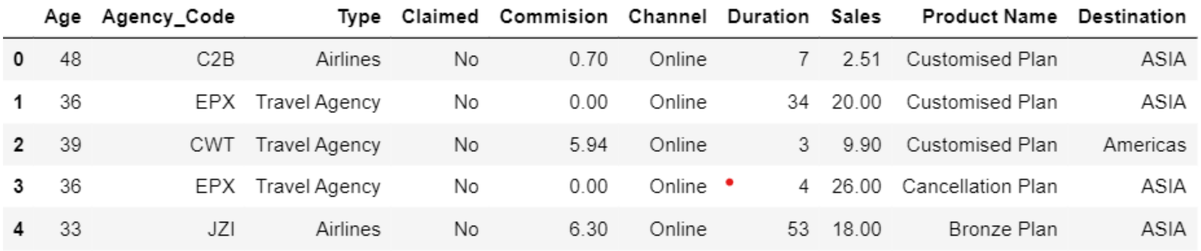
An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models'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

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

Exploratory Data Analysis:

- **Data description:**

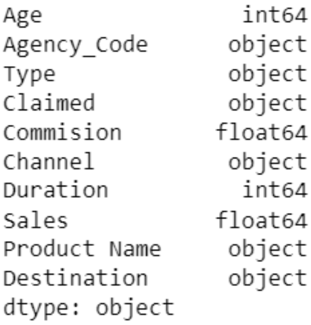
Reading the data file and loading first five records:



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

Table. 13

Dataset data types:

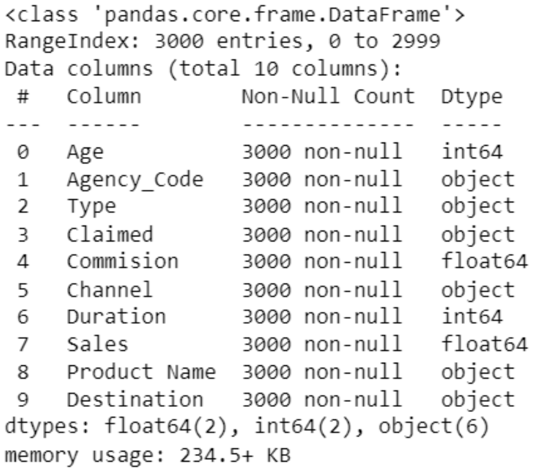


```
Age               int64
Agency_Code      object
Type             object
Claimed          object
Commision       float64
Channel          object
Duration          int64
Sales           float64
Product Name     object
Destination      object
dtype: object
```

Table. 14

- There are both numeric and object type features.

Dataset information:

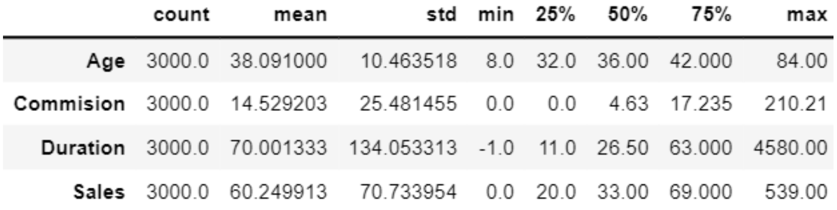


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Age           3000 non-null   int64  
 1   Agency_Code   3000 non-null   object 
 2   Type        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

Table.15

- There are no null values.
- There is total 4 numeric and 6 object type variables
- There is total 3000 records and 10 features in the dataset.
- There are 139 duplicated records.

Dataset description:



	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

Table.16

- **Insights:**

Let's describe each feature below:

1) Age: Mean age is ~38 years. There is no variation in this feature.

2) Commision:
   - Commision is in percentage on sales.
   - Minimum commision is 0% and maximum commission is ~210%
   - Out of two types of tour insurance firms available, Majority of the 'Travel Agencies' tou r insurance firm is taking zero percent commision on sales.\
   - Some of the 'Online' channels are taking more than 100% commision.

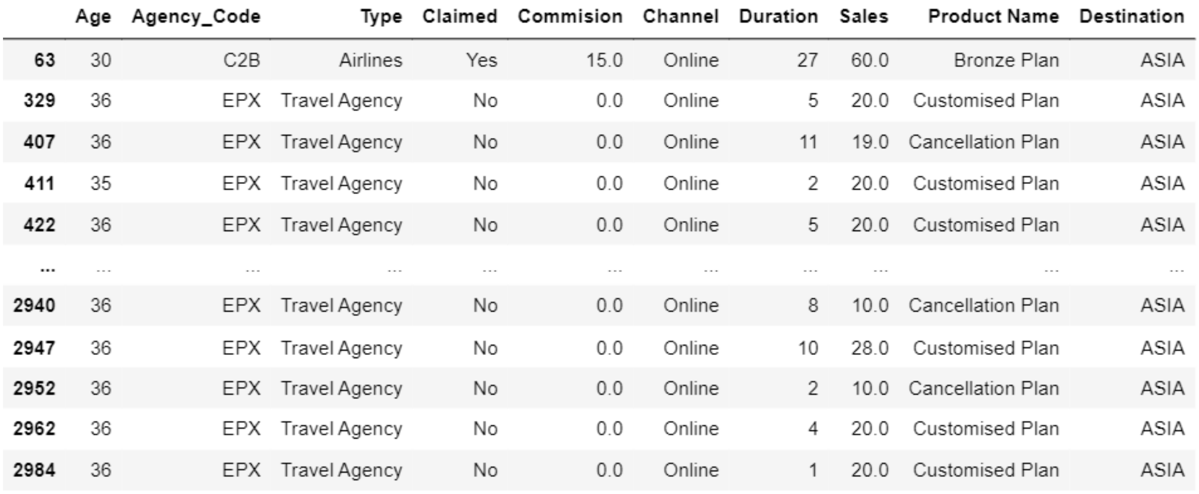
3) Duration:
   - On an average, customers have travelled 70 days using several term insurances avai lable.

- There is huge variation in the data for this feature.

4) Sales:
   - On an average, customers spent ~60000 on tour policies.
   - There is variation in the dataset for this feature.

➢ Data pre-processing:

- There are no null values in the dataset.
- There is total 4 numeric and 6 object type variables.
- There is total 3000 records and 10 features in the dataset.
- There are no anomalies present.
- Total 139 duplicated records are there in the dataset. Sample dataset of duplicated records is shown below:



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
63	30	C2B	Airlines	Yes	15.0	Online	27	60.0	Bronze Plan	ASIA
329	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
407	36	EPX	Travel Agency	No	0.0	Online	11	19.0	Cancellation Plan	ASIA
411	35	EPX	Travel Agency	No	0.0	Online	2	20.0	Customised Plan	ASIA
422	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
...	...	...	...	...	...	...	...	...	...	...
2940	36	EPX	Travel Agency	No	0.0	Online	8	10.0	Cancellation Plan	ASIA
2947	36	EPX	Travel Agency	No	0.0	Online	10	28.0	Customised Plan	ASIA
2952	36	EPX	Travel Agency	No	0.0	Online	2	10.0	Cancellation Plan	ASIA
2962	36	EPX	Travel Agency	No	0.0	Online	4	20.0	Customised Plan	ASIA
2984	36	EPX	Travel Agency	No	0.0	Online	1	20.0	Customised Plan	ASIA

Table.17

- There are 127 records are from zero percent 'Commision' records, 138 records are from 'ASIA' destination, 128 records are from 'Travel Agency' type insurance firm etc.
- But it can be of different customers, there is no customer ID or any unique identifier, so we are not dropping them off.

➢ Data visualization:

Univariate analysis:

- Let’s visualize all the numeric columns using hist plot and check the distribution nature of the features.

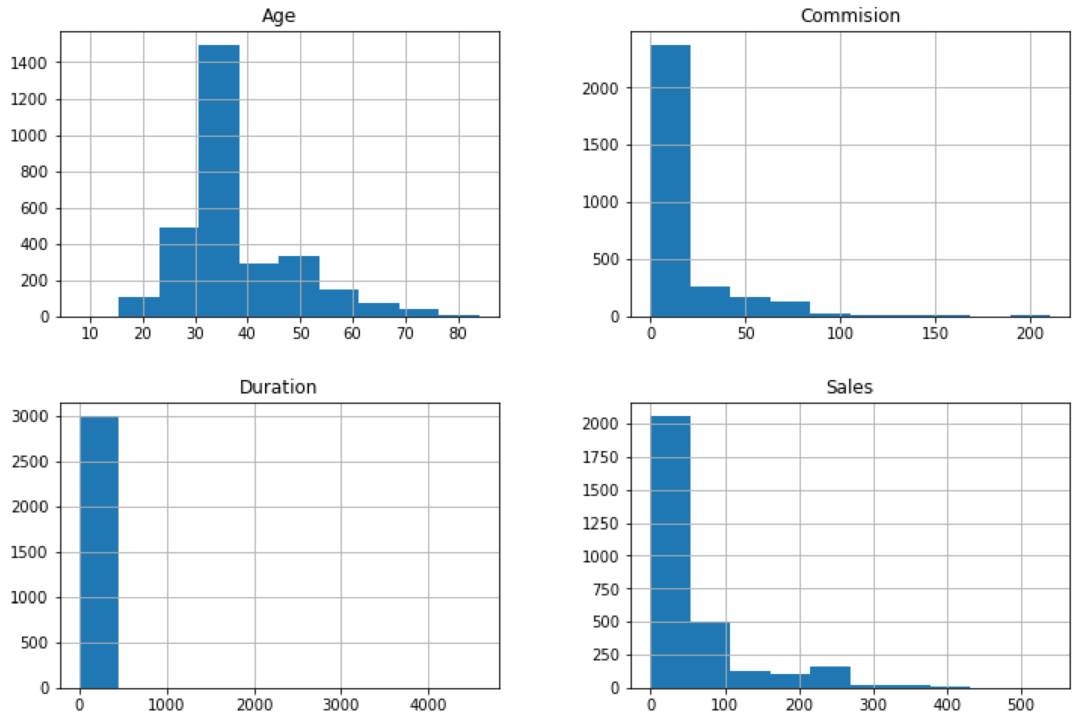
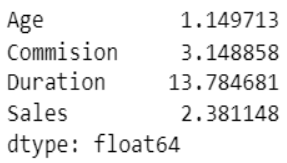


Fig.15

Checking skewness:



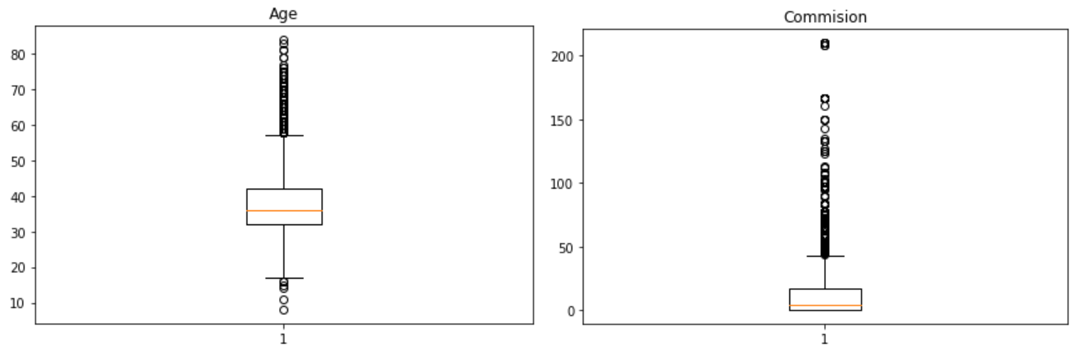
```
Age           1.149713
Commision     3.148858
Duration     13.784681
Sales         2.381148
dtype: float64
```

Table.18

Interpretations:

- None of the feature is normally distributed
- All numeric features are highly right skewed distributions and 'Duration' is ver y highly right skewed

Checking for outliers:



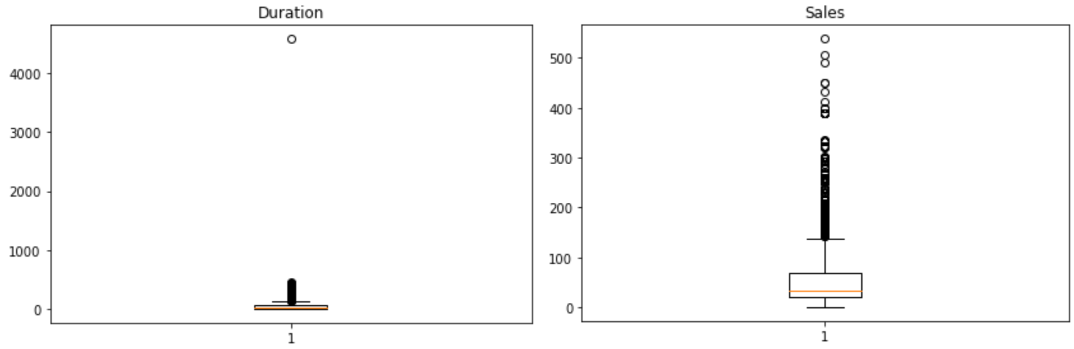


Fig. 16

- Outliers are present in every feature

Outlier treatment:

Let’s treat the outliers by IQR method.

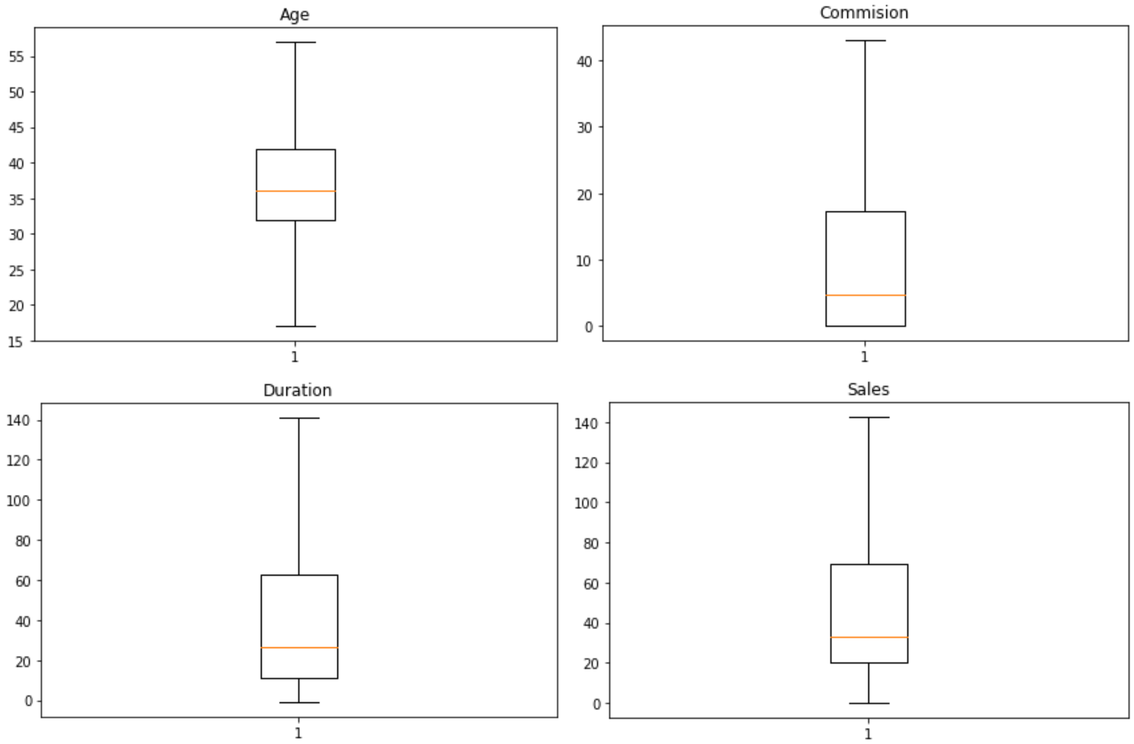


Fig. 17

- Outliers are treated by IQR method before making the models.

Bivariate analysis:

- Let’s plot the pair plot and heatmap to check correlation b/w the data features

Pair plot:

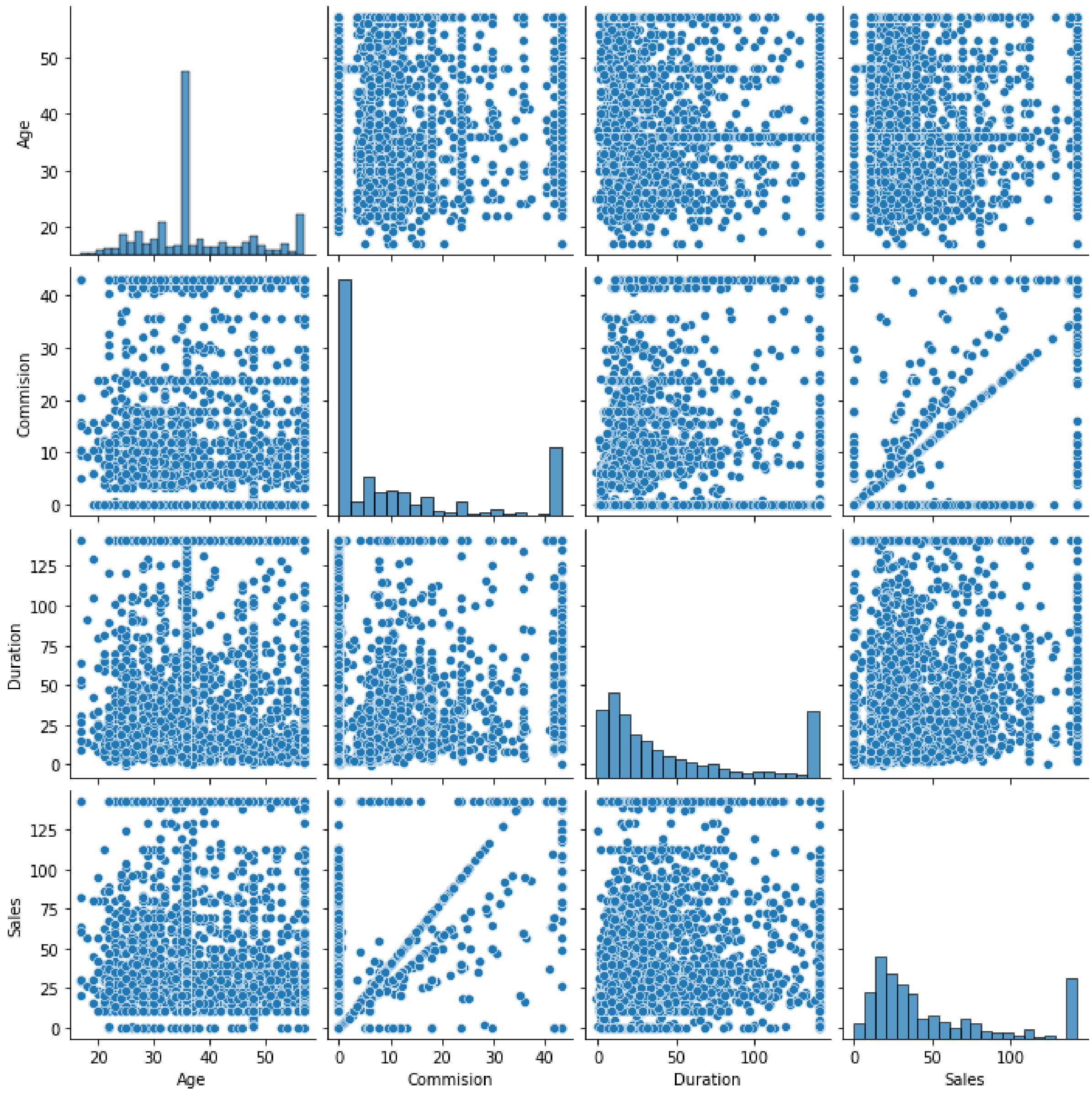


Fig.18

Heatmap:

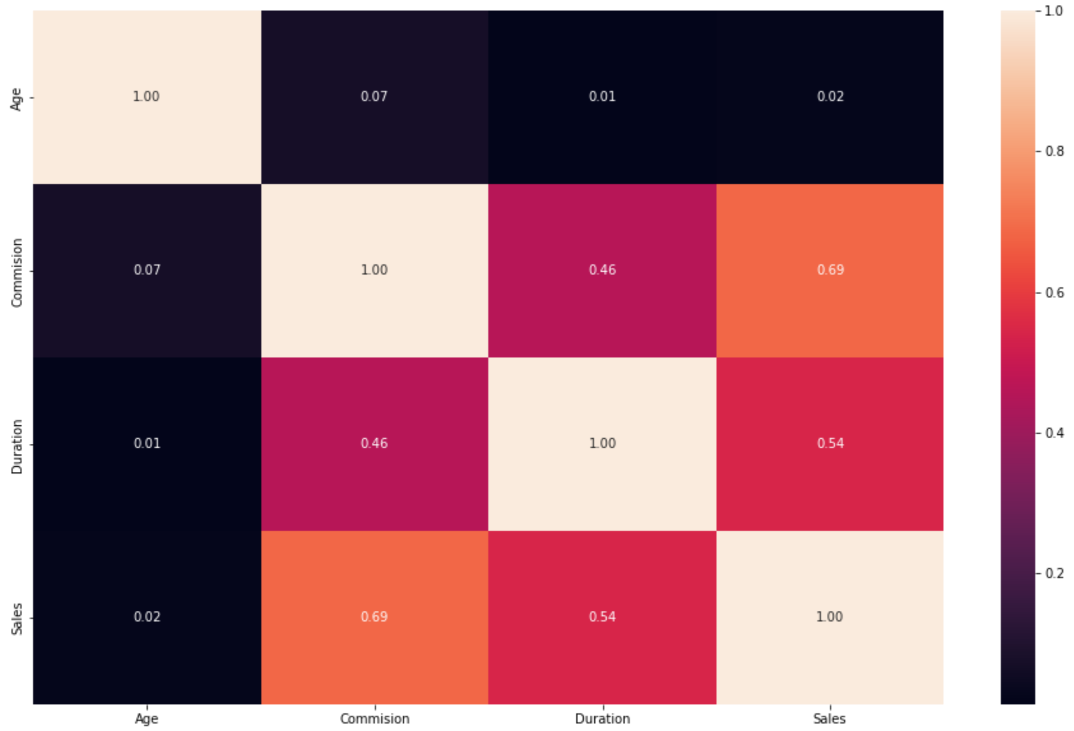


Fig.19

Insignts (From both pairplot and heatmap):

- 'Commision' and 'Sales' have moderate correlation among all the pairs.
- Remaining other pairs have weak to very weak correlation.

Multivariate analysis:

- Let’s analyse the sales for different features

Agency code:

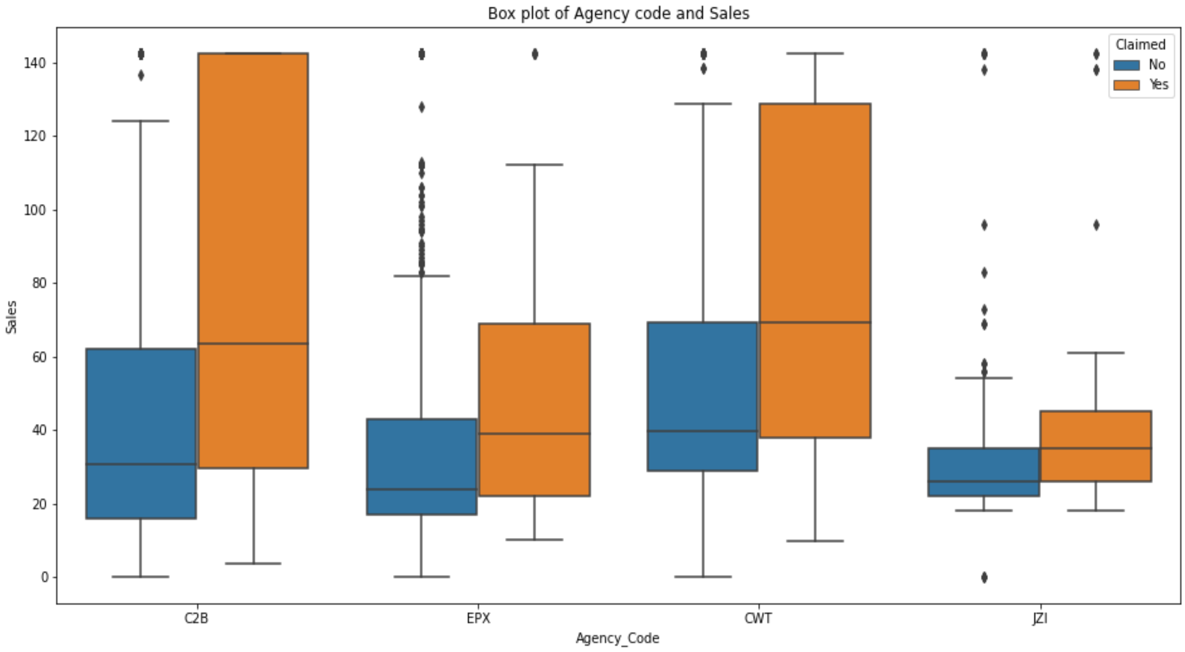


Fig. 20

Interpretations:

- C2B Agency has more sales for claimed customers.
- CWT Agency has more sales for non-claimed customers.
- JZI Agency has less sales for both claimed and non-claimed customers.

Type:

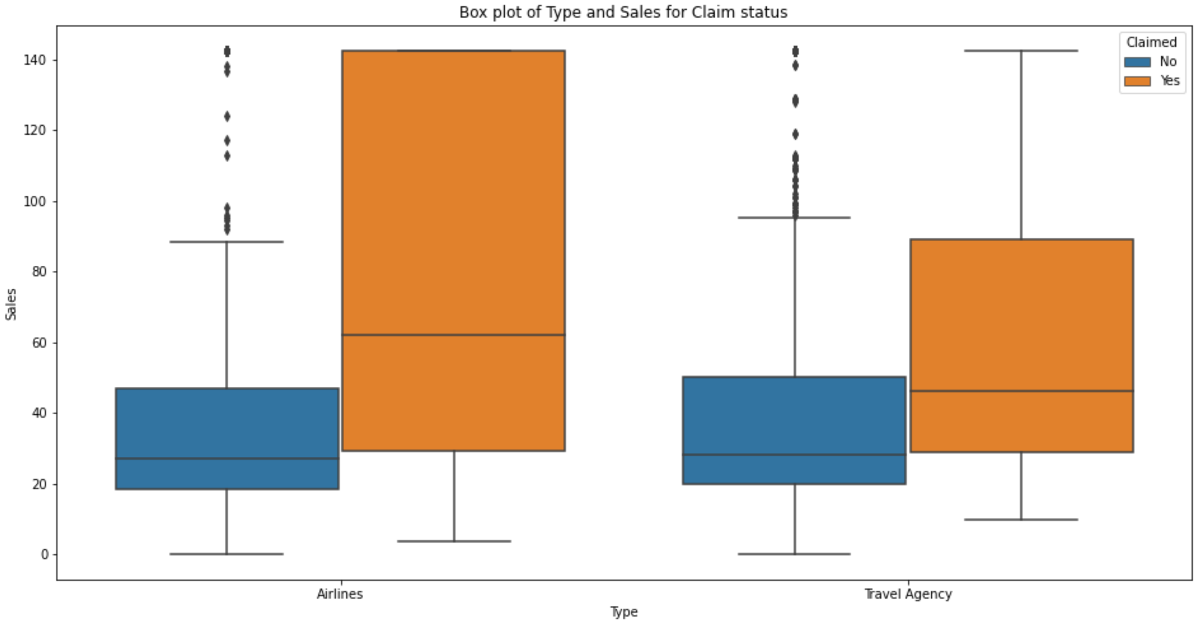


Fig. 21

Interpretation:

- 'Airlines' have more sales compared to 'Travel Agency' for claimed customers and a slight vice versa case for non-claimed customers.

Channel:

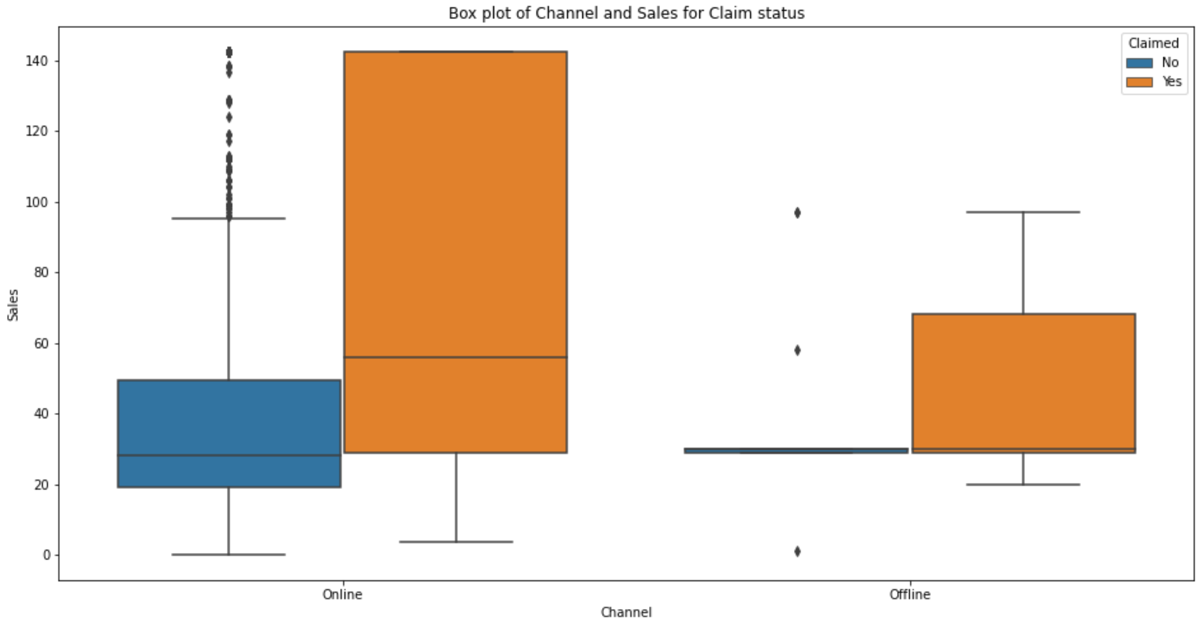


Fig.22

Interpretation:

- 'Online' services are more preferred for tour planning.

Destination:

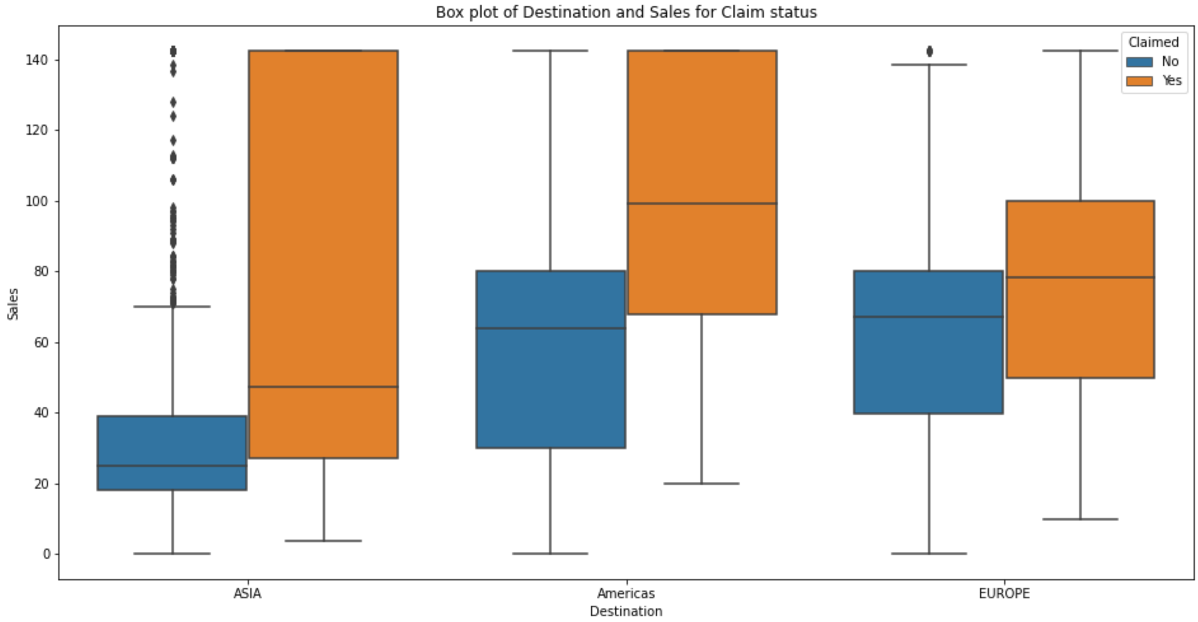


Fig. 23

Interpretation:

- Customers spending more across 'Asia' and 'America' compared to 'Europe'.
- On an average, 'America' has more sales for both claimed and non-claimed customers (almost similar with 'Europe' for non-claimed customers)

Destination vs Duration:

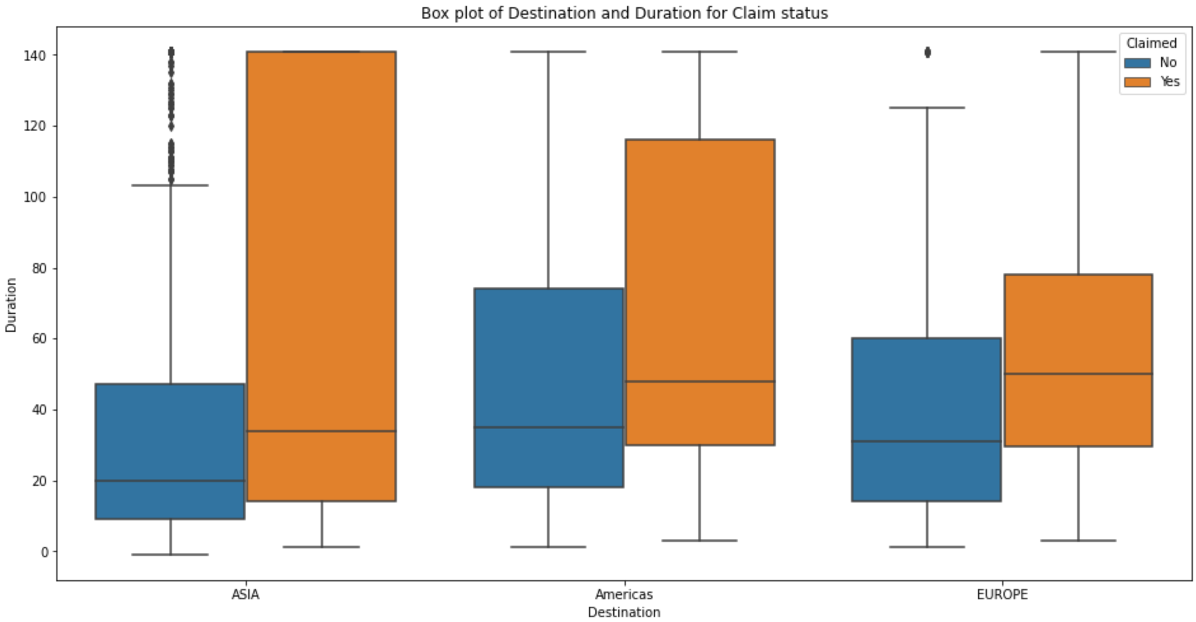


Fig. 24

- On an average, customers are spending more time in 'America', and then in 'Asia'

Let’s visualize sales vs commission on different products:

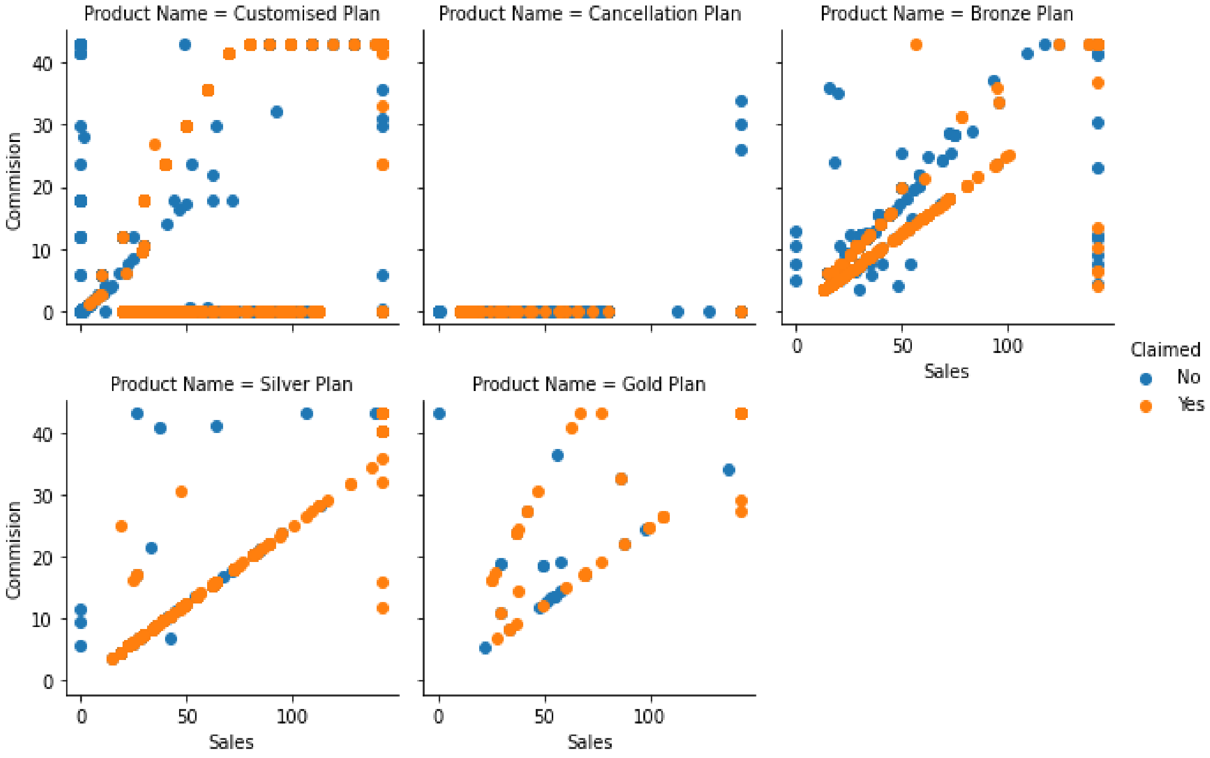


Fig.25

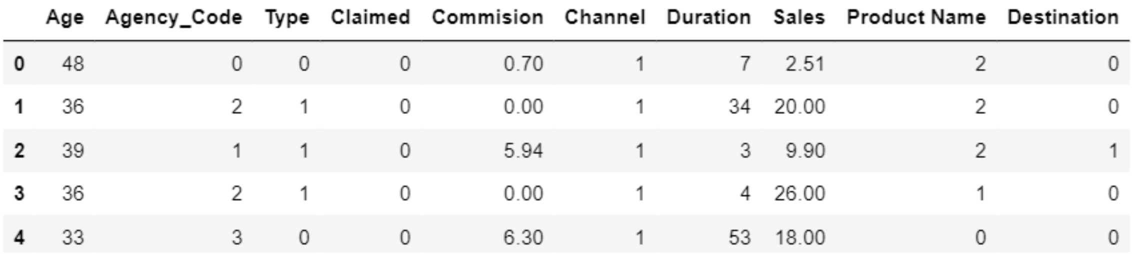
Interpretations:

- For 'Silver plan' and 'Gold plan', there is very good correlation b/w Sales and Commis ion.
- For 'Customised plan' and 'Bronze plan', there is very moderate correlation b/w Sales and Commision.
- Poor correlation exists for 'Cancellation plan' b/w Sales and Commision.

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

- **Data encoding:**

Converting object data into categorical/numercial data:



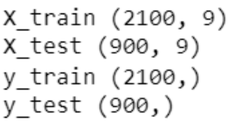
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0	0.70	1	7	2.51	2	0
1	36	2	1	0	0.00	1	34	20.00	2	0
2	39	1	1	0	5.94	1	3	9.90	2	1
3	36	2	1	0	0.00	1	4	26.00	1	0
4	33	3	0	0	6.30	1	53	18.00	0	0

Table. 19

- Dataset encoded successfully
- Let's copy the original dataset and use it for ANN model with scaling it.
- Let's build the CART and RF models without scaling it.
  Note: Splitting will be done by using random_state

Splitting the data into Train and Test set:

- Target variable is 'Claimed'
- Let's drop 'Claimed' variable for train dataset and pop it for test dataset
- Dataset splitting is done with 30% test dataset and 70% train dataset.
- Let's check the shapes of splitted dataset



```
X_train (2100, 9)
X_test (900, 9)
y_train (2100,)
y_test (900,)
```

Fig. 26

1) CART model:

- Let's build decision tree by using 'gini' criterion
- Decision tree is built and root node is shown below:

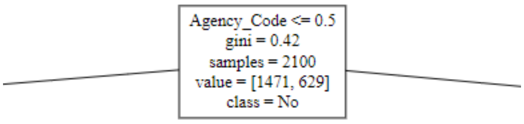
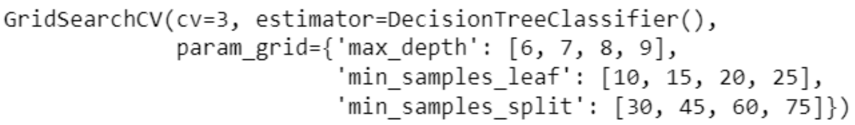


Fig. 27

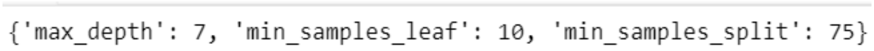
- As per tree diagram, best feature taken for decision tree is 'Agency_code' and its gini gain value is 0.42.
- From tree diagram, we can choose optimum depth as 10
- Now, let's get best parameters by using 'GridSearch' by setting the parameters as shown below:



```
GridSearchCV(cv=3, estimator=DecisionTreeClassifier(),
             param_grid={'max_depth': [6, 7, 8, 9],
                         'min_samples_leaf': [10, 15, 20, 25],
                         'min_samples_split': [30, 45, 60, 75]})
```

Fig. 28

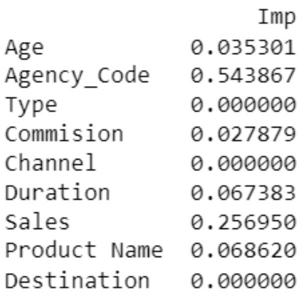
- Best parameters are as below:



```
{'max_depth': 7, 'min_samples_leaf': 10, 'min_samples_split': 75}
```

Fig. 29

- Feature importance's are below:



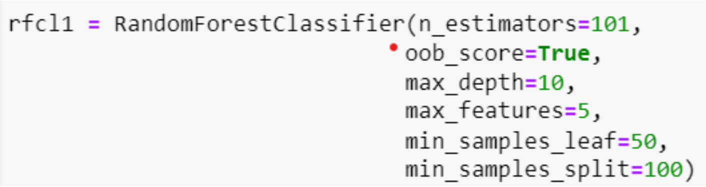
```
                   Imp
Age           0.035301
Agency_Code   0.543867
Type          0.000000
Commision     0.027879
Channel       0.000000
Duration      0.067383
Sales         0.256950
Product Name  0.068620
Destination   0.000000
```

Table. 20

- From 'GridSearch' also, we can see 'Agency_Code' is the best feature for the classification.

2) Random Forest model:

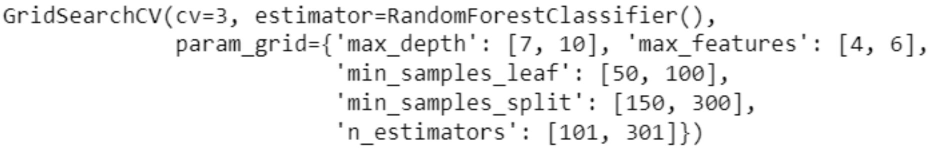
- Let's build RF model by defining some rough parameters initially as shown below:



```
rfcl1 = RandomForestClassifier(n_estimators=101,
                               oob_score=True,
                               max_depth=10,
                               max_features=5,
                               min_samples_leaf=50,
                               min_samples_split=100)
```

Fig. 30

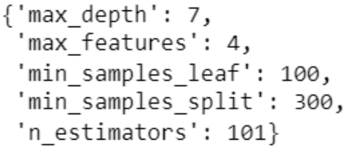
- Training data is fit into the model with the above defined parameters.
- Out of bag score for this defined model is 78.66%
- Now, let's get best parameters by using 'GridSearch' by setting the parameters as shown below:



```
GridSearchCV(cv=3, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [7, 10], 'max_features': [4, 6],
                         'min_samples_leaf': [50, 100],
                         'min_samples_split': [150, 300],
                         'n_estimators': [101, 301]})
```

Fig. 31

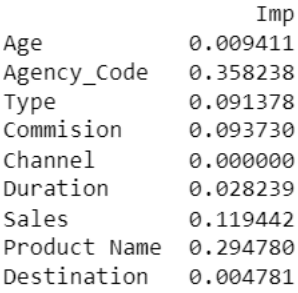
- Best parameters are as below:



```
{'max_depth': 7,
 'max_features': 4,
 'min_samples_leaf': 100,
 'min_samples_split': 300,
 'n_estimators': 101}
```

Fig. 32

- Let's build a by using the best parameters and use this for predictions further
- Feature importance's are below:



```
                   Imp
Age           0.009411
Agency_Code   0.358238
Type          0.091378
Commision     0.093730
Channel       0.000000
Duration      0.028239
Sales         0.119442
Product Name  0.294780
Destination   0.004781
```

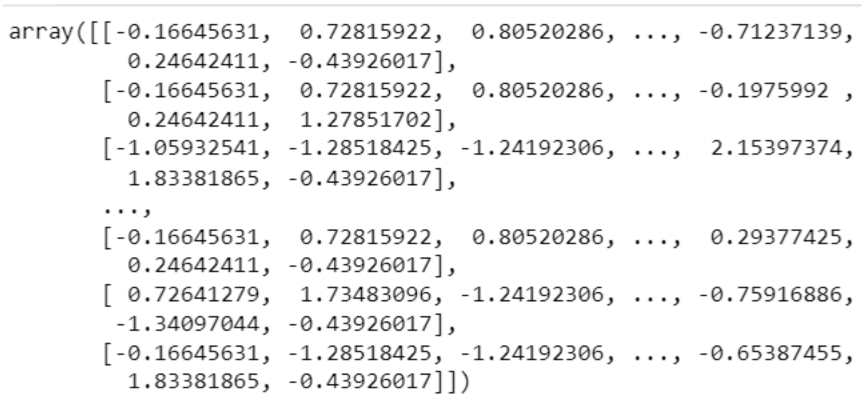
Table. 21

- From RF model also, it can be seen that 'Agency_Code' is the best feature for the classification.

3) ANN model:

- ANN requires scaling of the dataset
- Feature scaling: Scaled train dataset (fit.transform) and test dataset (transform)
- Sample scaled data of train and test dataset are as shown below:

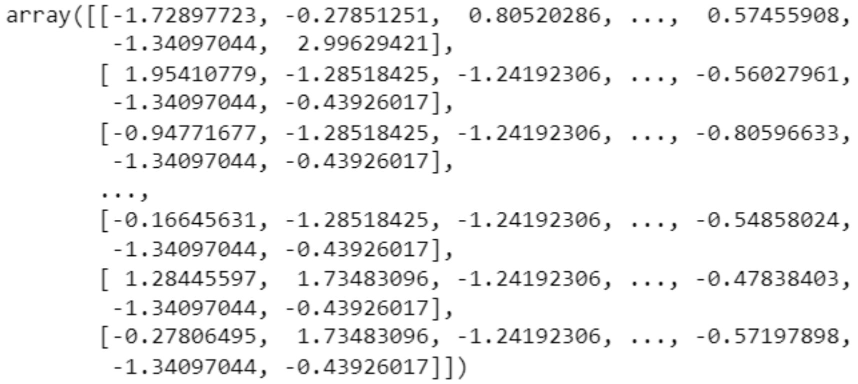
Train scaled data:



```
array([[-0.16645631,  0.72815922,  0.80520286, ..., -0.71237139,
         0.24642411, -0.43926017],
       [-0.16645631,  0.72815922,  0.80520286, ..., -0.1975992 ,
         0.24642411,  1.27851702],
       [-1.05932541, -1.28518425, -1.24192306, ...,  2.15397374,
         1.83381865, -0.43926017],
       ...,
       [-0.16645631,  0.72815922,  0.80520286, ...,  0.29377425,
         0.24642411, -0.43926017],
       [ 0.72641279,  1.73483096, -1.24192306, ..., -0.75916886,
        -1.34097044, -0.43926017],
       [-0.16645631, -1.28518425, -1.24192306, ..., -0.65387455,
         1.83381865, -0.43926017]])
```

Fig. 33

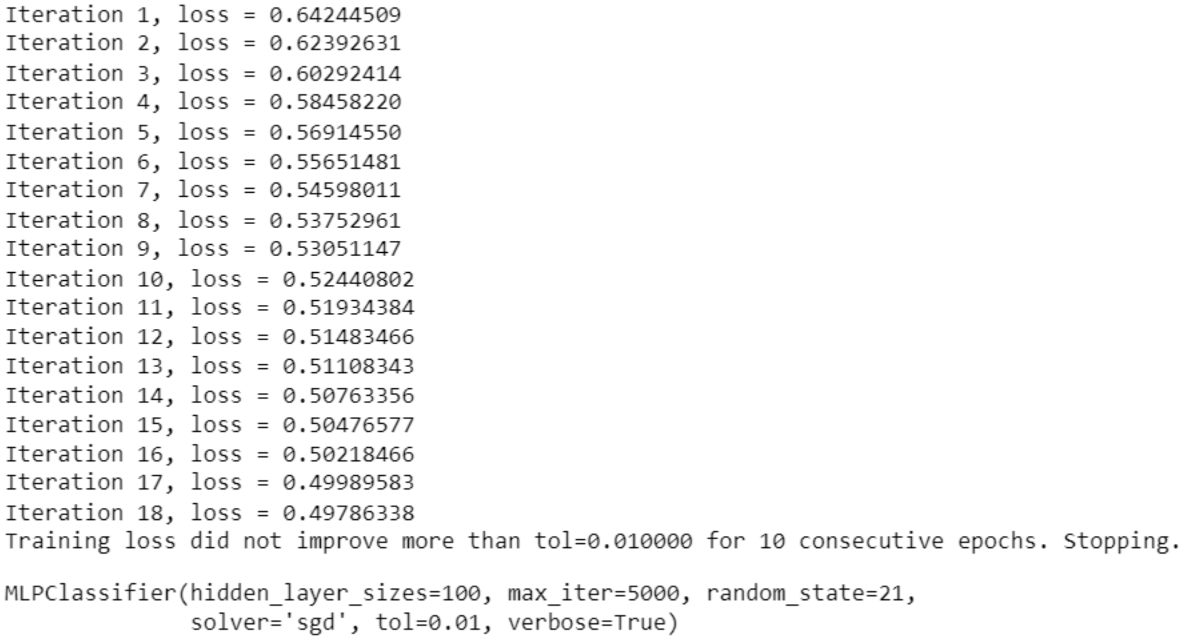
Test dataset:



```
array([[-1.72897723, -0.27851251,  0.80520286, ...,  0.57455908,
        -1.34097044,  2.99629421],
       [ 1.95410779, -1.28518425, -1.24192306, ..., -0.56027961,
        -1.34097044, -0.43926017],
       [-0.94771677, -1.28518425, -1.24192306, ..., -0.80596633,
        -1.34097044, -0.43926017],
       ...,
       [-0.16645631, -1.28518425, -1.24192306, ..., -0.54858024,
        -1.34097044, -0.43926017],
       [ 1.28445597,  1.73483096, -1.24192306, ..., -0.47838403,
        -1.34097044, -0.43926017],
       [-0.27806495,  1.73483096, -1.24192306, ..., -0.57197898,
        -1.34097044, -0.43926017]])
```

Fig. 34

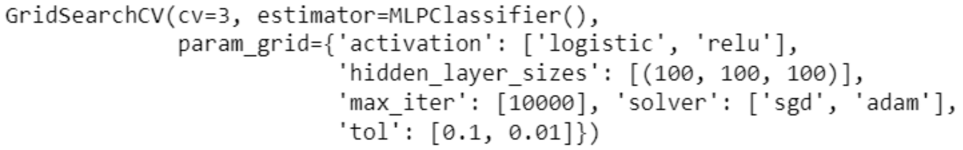
Let’s fit the train data using some rough parameters:



```
Iteration 1, loss = 0.64244509
Iteration 2, loss = 0.62392631
Iteration 3, loss = 0.60292414
Iteration 4, loss = 0.58458220
Iteration 5, loss = 0.56914550
Iteration 6, loss = 0.55651481
Iteration 7, loss = 0.54598011
Iteration 8, loss = 0.53752961
Iteration 9, loss = 0.53051147
Iteration 10, loss = 0.52440802
Iteration 11, loss = 0.51934384
Iteration 12, loss = 0.51483466
Iteration 13, loss = 0.51108343
Iteration 14, loss = 0.50763356
Iteration 15, loss = 0.50476577
Iteration 16, loss = 0.50218466
Iteration 17, loss = 0.49989583
Iteration 18, loss = 0.49786338
Training loss did not improve more than tol=0.010000 for 10 consecutive epochs. Stopping.

MLPClassifier(hidden_layer_sizes=100, max_iter=5000, random_state=21,
              solver='sgd', tol=0.01, verbose=True)
```

Fig. 35

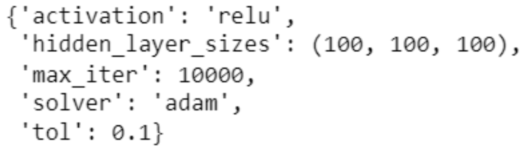
- Let’s do GridSearch to get best ANN model. Parameter given are as below:



```
GridSearchCV(cv=3, estimator=MLPClassifier(),
             param_grid={'activation': ['logistic', 'relu'],
                         'hidden_layer_sizes': [(100, 100, 100)],
                         'max_iter': [10000], 'solver': ['sgd', 'adam'],
                         'tol': [0.1, 0.01]})
```

Fig. 36

- Best parameters are as shown below:



```
{'activation': 'relu',
 'hidden_layer_sizes': (100, 100, 100),
 'max_iter': 10000,
 'solver': 'adam',
 'tol': 0.1}
```

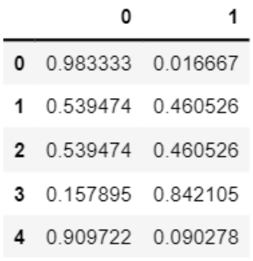
Fig. 37

- Let’s build the with these best parameters and use for the test predictions

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.

1) CART model:

- Accuracy score for train set 80.28%
- Accuracy score for test set 76.00%
- Let’s predict the test dataset using best model obtained above
- Probabilities obtained are as below:

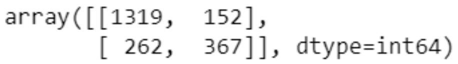


	0	1
0	0.983333	0.016667
1	0.539474	0.460526
2	0.539474	0.460526
3	0.157895	0.842105
4	0.909722	0.090278

Table. 22

Train dataset:

- Confusion matrix:



```
array([[1319,  152],
       [ 262,  367]], dtype=int64)
```

Fig. 38

- ROC_AUC score is 0.85
- ROC curve:

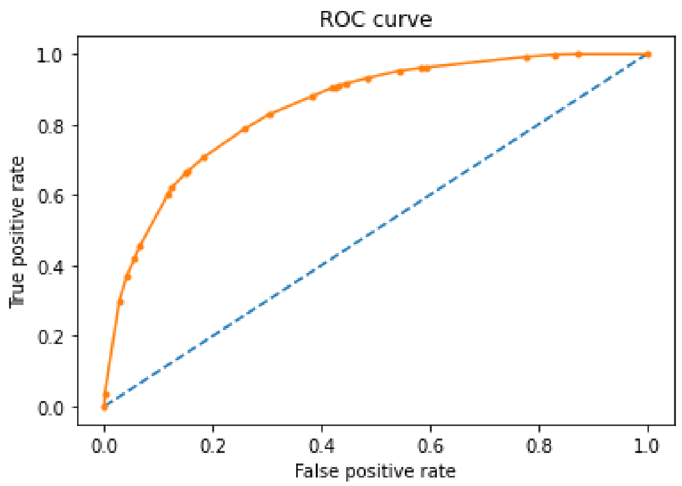
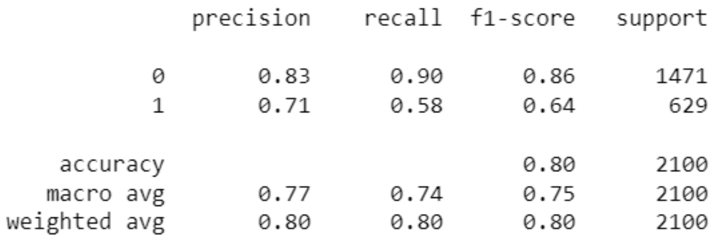


Fig. 39

- Classification report:

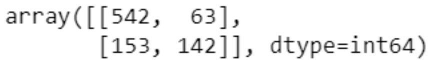


```
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1471
           1       0.71      0.58      0.64       629

    accuracy                           0.80      2100
   macro avg       0.77      0.74      0.75      2100
weighted avg       0.80      0.80      0.80      2100
```

Fig. 40

Test dataset:

- Confusion matrix:



```
array([[542,  63],
       [153, 142]], dtype=int64)
```

Fig. 41

- ROC_AUC score is 0.80
- ROC curve:

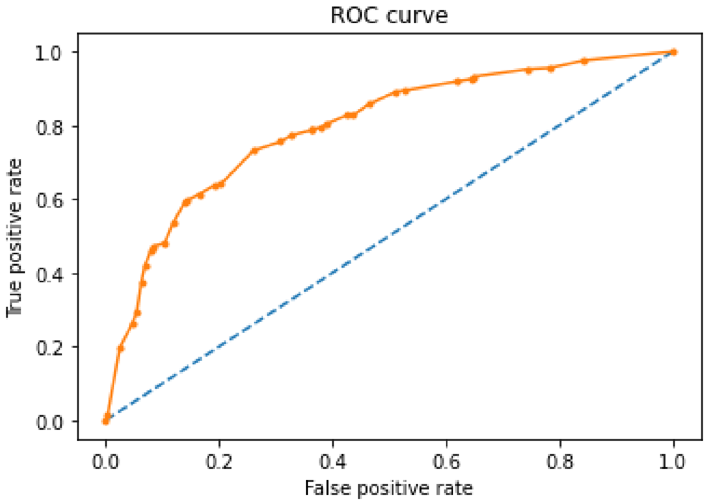
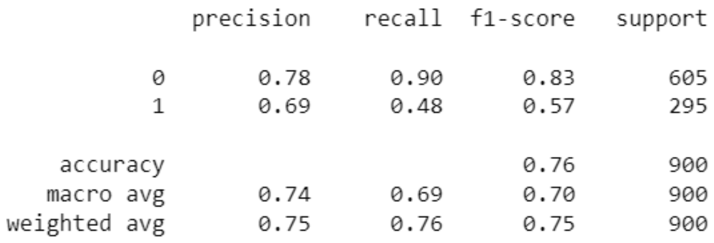


Fig. 42

- Classification report:

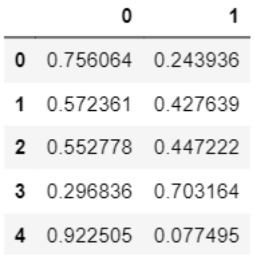


```
              precision    recall  f1-score   support

           0       0.78      0.90      0.83       605
           1       0.69      0.48      0.57       295

    accuracy                           0.76       900
   macro avg       0.74      0.69      0.70       900
weighted avg       0.75      0.76      0.75       900
```

Fig. 43

2) Random Forest model:

- Accuracy score for train set 78.00%
- Accuracy score for test set 74.66%
- Let's predict the test dataset using best model obtained above
- Probabilities obtained are as below:

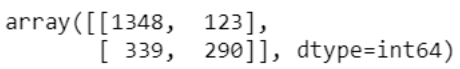


	0	1
0	0.756064	0.243936
1	0.572361	0.427639
2	0.552778	0.447222
3	0.296836	0.703164
4	0.922505	0.077495

Table. 23

Train dataset:

- Confusion matrix:



```
array([[1348,  123],
       [ 339,  290]], dtype=int64)
```

Fig. 44

- ROC_AUC score is 0.82
- ROC curve:

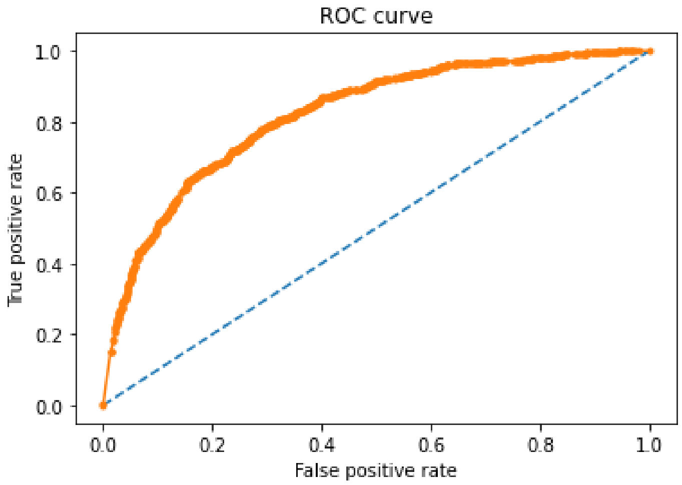
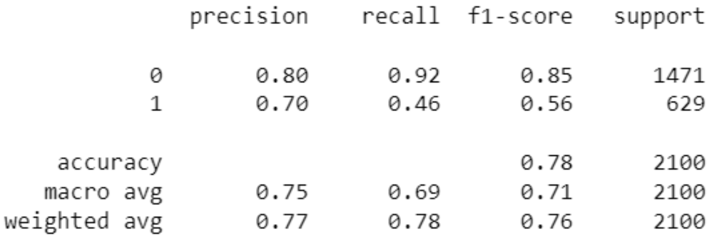


Fig. 45

- Classification report:

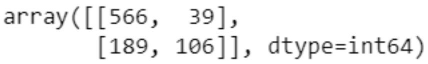


```
              precision    recall  f1-score   support

           0       0.80      0.92      0.85      1471
           1       0.70      0.46      0.56       629

    accuracy                           0.78      2100
   macro avg       0.75      0.69      0.71      2100
weighted avg       0.77      0.78      0.76      2100
```

Fig. 46

Test dataset:

- Confusion matrix:



```
array([[566,  39],
       [189, 106]], dtype=int64)
```

Fig. 47

- ROC_AUC score is 0.81
- ROC curve:

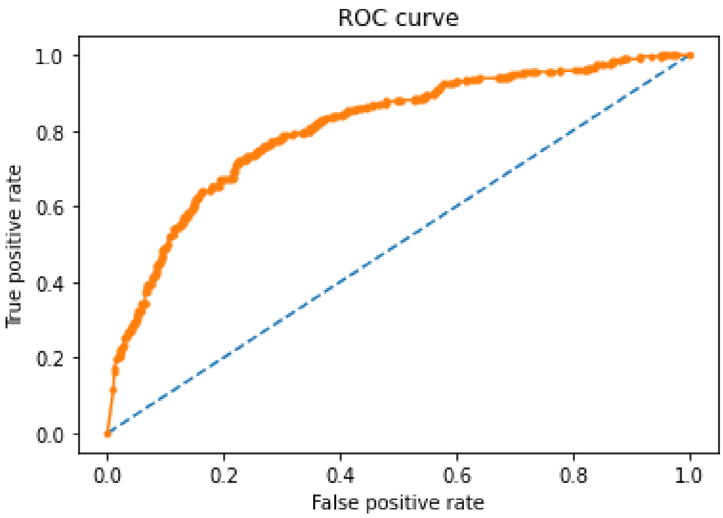
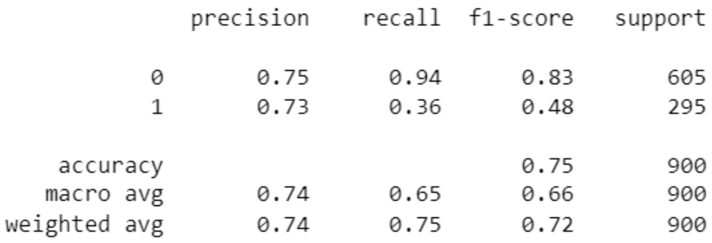


Fig. 48

- Classification report:

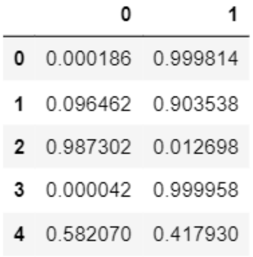


```
              precision    recall  f1-score   support

           0       0.75      0.94      0.83       605
           1       0.73      0.36      0.48       295

    accuracy                           0.75       900
   macro avg       0.74      0.65      0.66       900
weighted avg       0.74      0.75      0.72       900
```

Fig. 49

3) ANN model:

- Accuracy score for train set 78.47%
- Accuracy score for test set 76.00%
- Let's predict the test dataset using best model obtained above
- Probabilities obtained are as below:

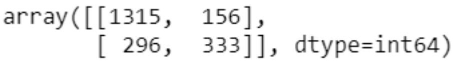


	0	1
0	0.000186	0.999814
1	0.096462	0.903538
2	0.987302	0.012698
3	0.000042	0.999958
4	0.582070	0.417930

Table. 24

Train dataset:

- Confusion matrix:



```
array([[1315,  156],
       [ 296,  333]], dtype=int64)
```

Fig. 50

- ROC_AUC score is 0.84
- ROC curve:

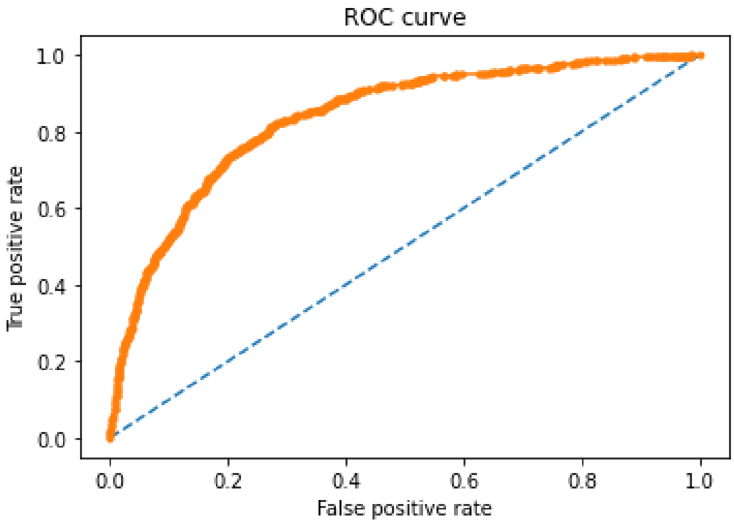
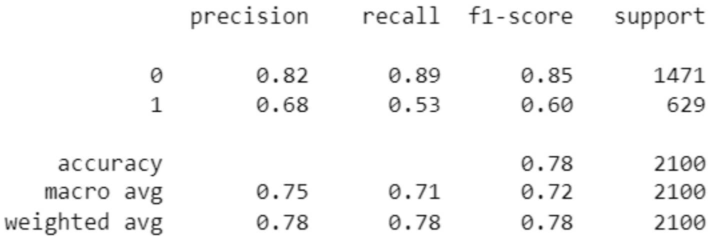


Fig. 51

- Classification report:

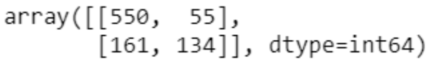


```
              precision    recall  f1-score   support

           0       0.82      0.89      0.85      1471
           1       0.68      0.53      0.60       629

    accuracy                           0.78      2100
   macro avg       0.75      0.71      0.72      2100
weighted avg       0.78      0.78      0.78      2100
```

Fig. 52

Test dataset:

- Confusion matrix:



```
array([[550,  55],
       [161, 134]], dtype=int64)
```

Fig. 53

- ROC_AUC score is 0.81
- ROC curve:

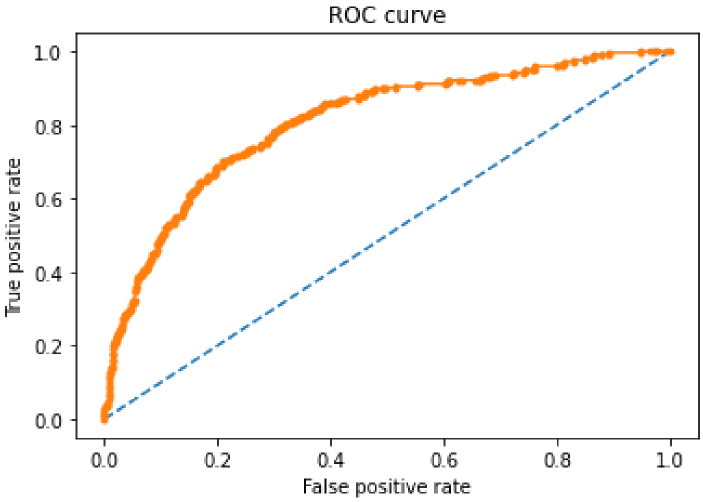
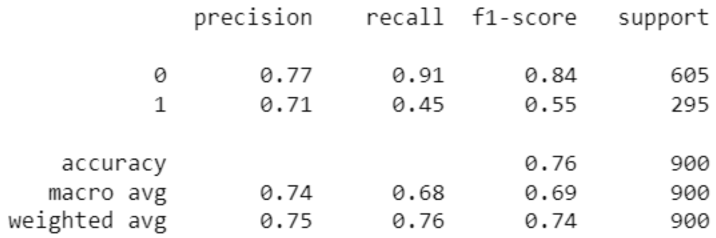


Fig. 54

- Classification report:

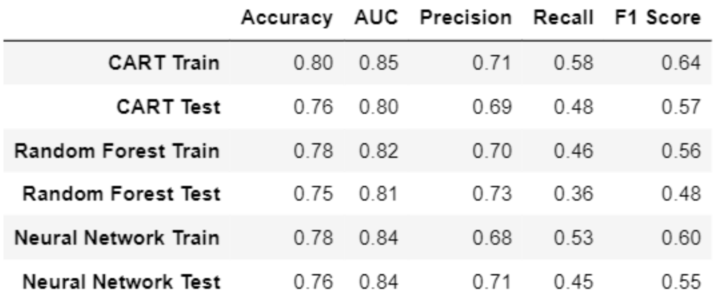


```
              precision    recall  f1-score   support

           0       0.77      0.91      0.84       605
           1       0.71      0.45      0.55       295

    accuracy                           0.76       900
   macro avg       0.74      0.68      0.69       900
weighted avg       0.75      0.76      0.74       900
```

Fig. 55

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

Table having Accuracy, AUC score, Precision, Recall, F1 score are mentioned below for both train and test dataset:



	Accuracy	AUC	Precision	Recall	F1 Score
CART Train	0.80	0.85	0.71	0.58	0.64
CART Test	0.76	0.80	0.69	0.48	0.57
Random Forest Train	0.78	0.82	0.70	0.46	0.56
Random Forest Test	0.75	0.81	0.73	0.36	0.48
Neural Network Train	0.78	0.84	0.68	0.53	0.60
Neural Network Test	0.76	0.84	0.71	0.45	0.55

Table. 25

CONCLUSION:

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

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

- I strongly recommended we collect more real time unstructured data and past data if possible.
- This is understood by looking at the insurance data by drawing relations between different variables such as day of the incident, time, age group, and associating it with other external information such as location, 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:

  i) Reduce claims cycle time,

  ii) Increase customer satisfaction,

  iii) Combat fraud,

  iv) Optimize claims recovery,

  v) 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.

THE END