Problem 1: Clustering

1.1- Dataset head

	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

Number of columns- 7
Number of rows- 210

#	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

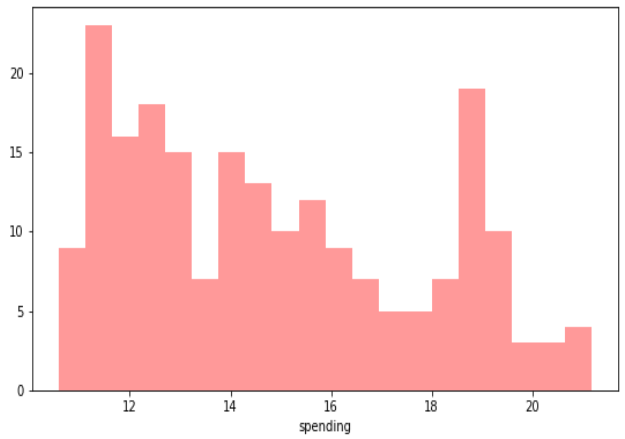
From info of data we can infer that there is no null value in the dataset and all variables in the dataset are floats.

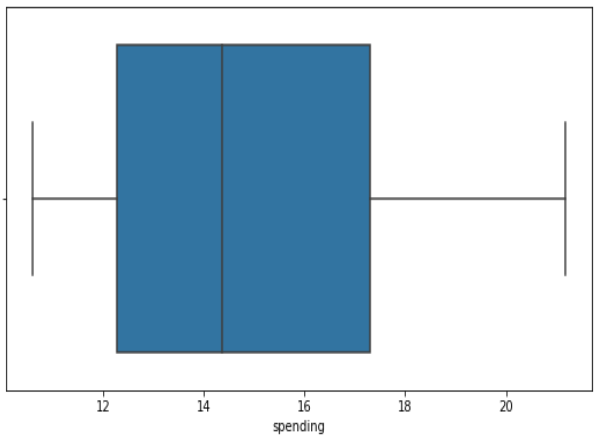
	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.85	2.91	10.59	12.27	14.36	17.30	21.18
advance_payments	210.0	14.56	1.31	12.41	13.45	14.32	15.72	17.25
probability_of_full_payment	210.0	0.87	0.02	0.81	0.86	0.87	0.89	0.92
current_balance	210.0	5.63	0.44	4.90	5.26	5.52	5.98	6.68
credit_limit	210.0	3.26	0.38	2.63	2.94	3.24	3.56	4.03
min_payment_amt	210.0	3.70	1.50	0.77	2.56	3.60	4.77	8.46
max_spent_in_single_shopping	210.0	5.41	0.49	4.52	5.04	5.22	5.88	6.55

Based on the summary we can see that the mean amount spent by customers is 14,850 and mean probability of full payment is 87%. Based on mean and median values we can make out that data is normally distributed.

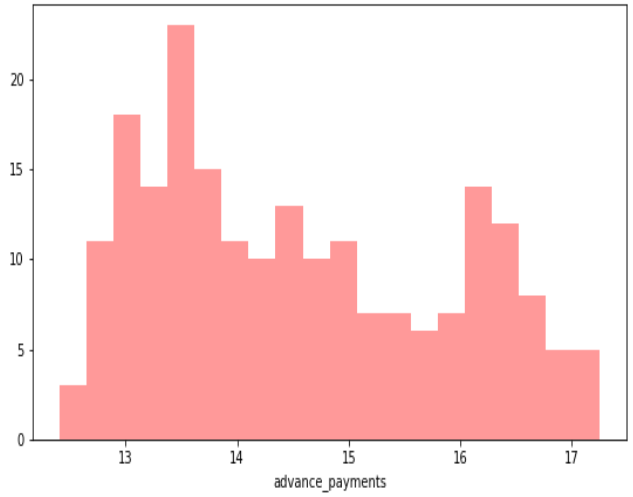
Univariate Analysis

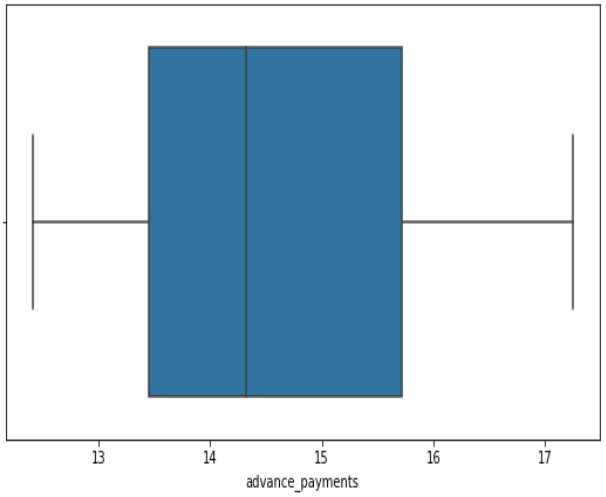
Spending
Skew: 0.4



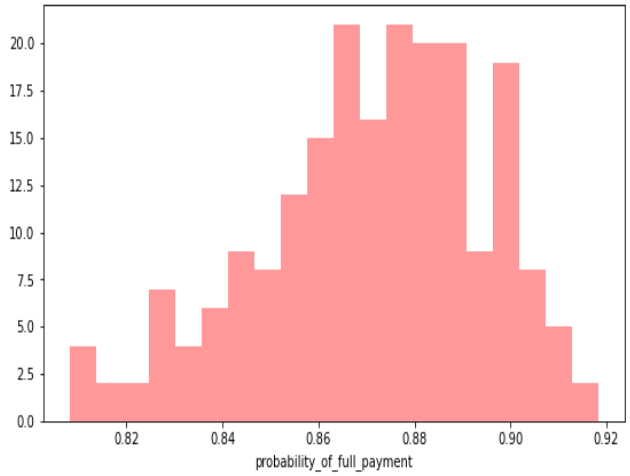


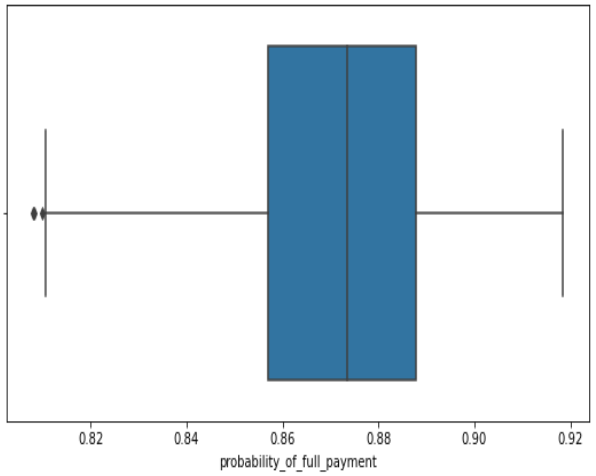
Advance_payments
Skew: 0.39



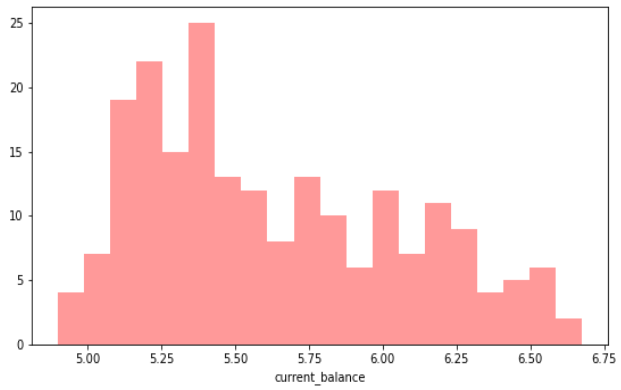


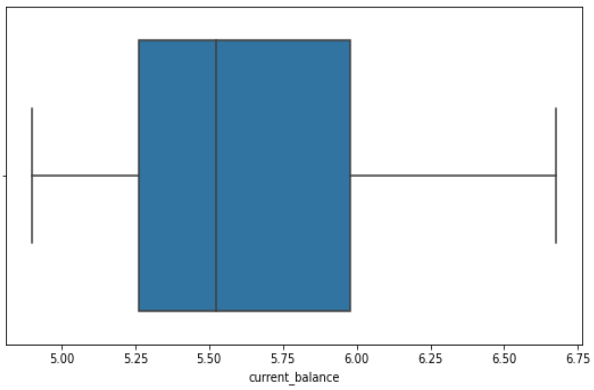
Probability_of_full_payment
skew : -0.54



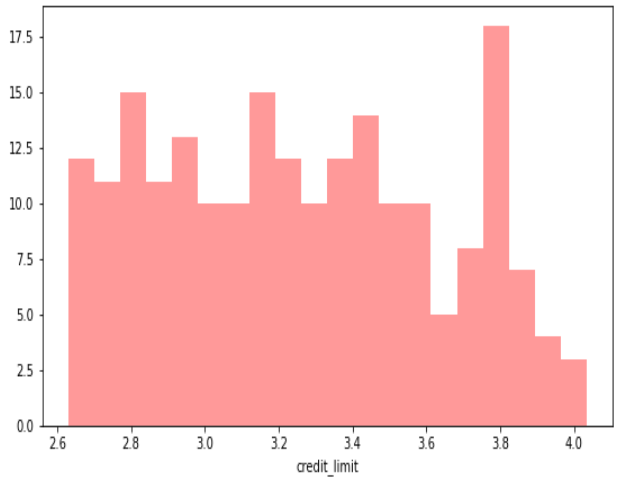


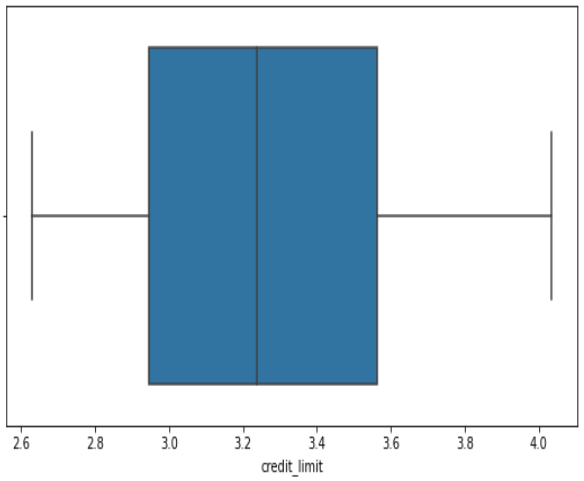
Current_balance
skew : 0.53



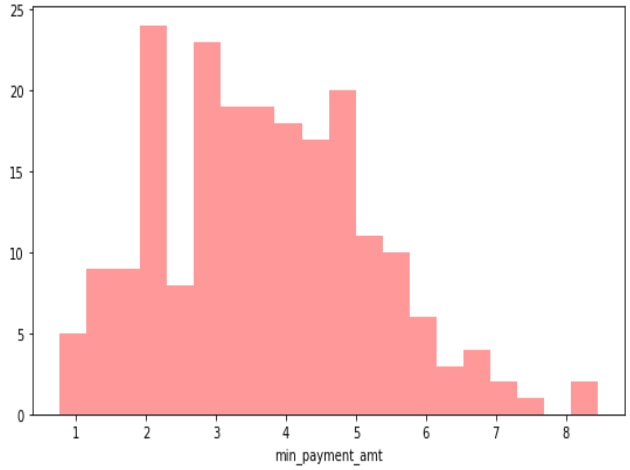


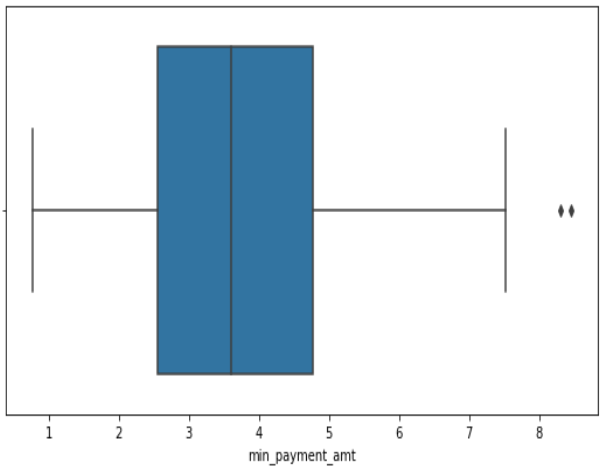
Credit_limit
skew : 0.13



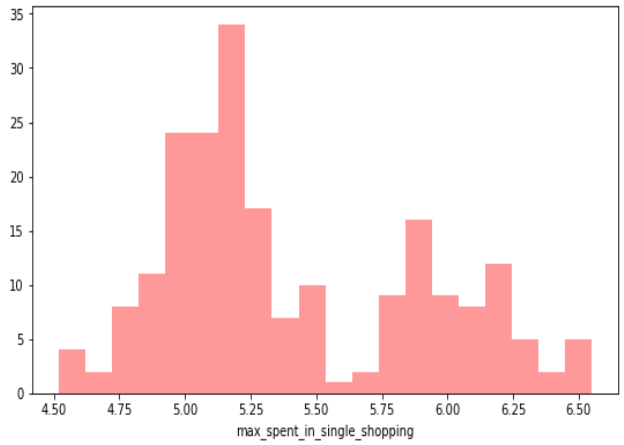


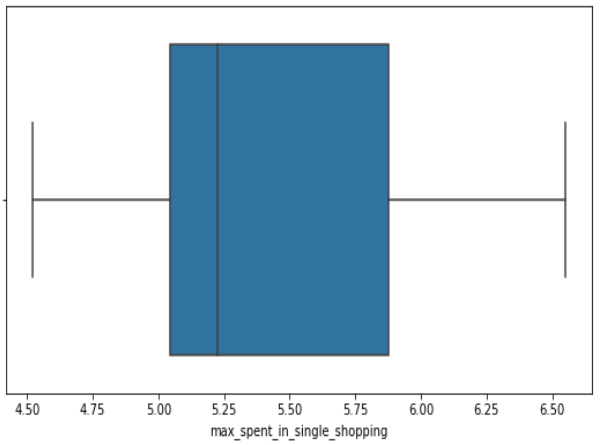
Min_payment_amt
skew : 0.4





Max_spent_in_single_shopping
skew : 0.56

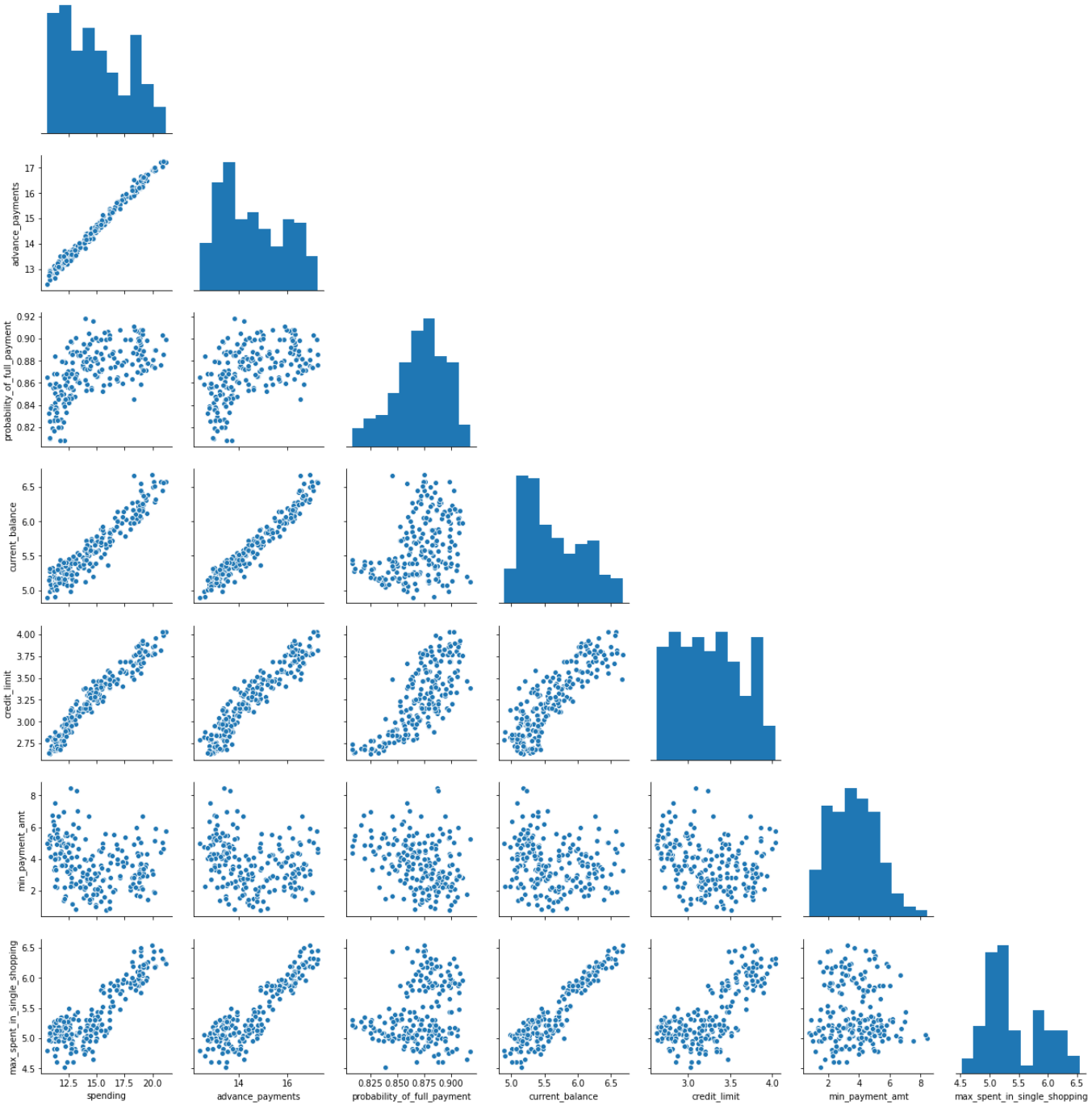




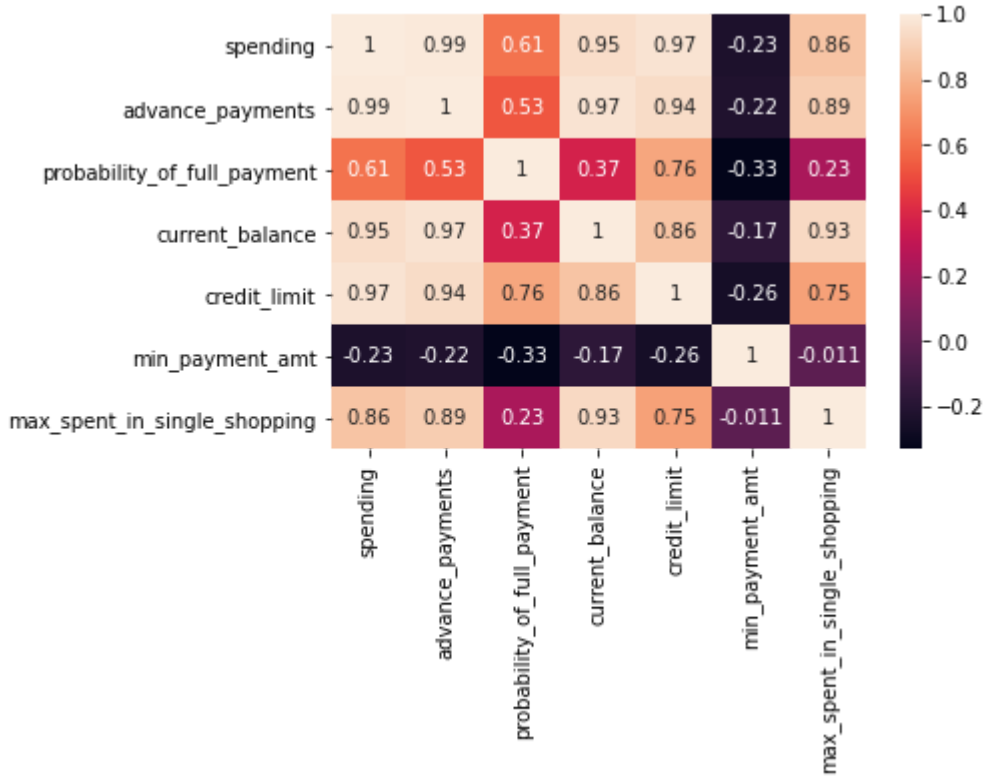
From the diagrames we can see that only Probability_of_full_payment variable is left skewed.
All other variables are right skewed.
Box plots show that Probability_of_full_payment has outliers in the lower side and
Min_payment_amt has outliers in the higher side.

Bivariate analysis

Pairplot:



Heatmap:



From pairplot and Heatmap we can see positive correlation between following variables:
-Advance_ payment and spending
-Current_balance and spending
-Current_balance and advance_payment
-Credit_limit and spending
-Credit_limit and advance_payments
-Credit_limit and current_balance
-Max_spent_in_single_shopping and spending
-Max_spent_in_single_shopping and advance_payments
-Max_spent_in_single_shopping and current_balance

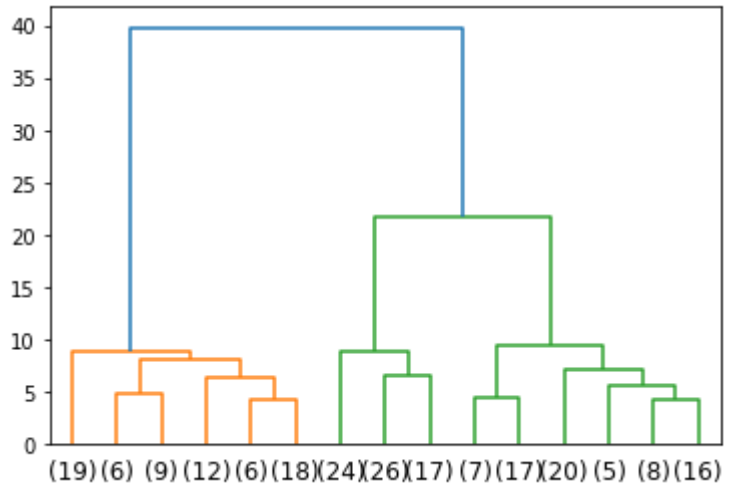
1.2 - Scaling

Clustering uses distance based algorithms for cluster formation. If variables in the data have large differences in their variances it will impact the algorithm .Variables with large variances will have more influence on the construction of clusters.Thus all variables need to be scaled before clustering.
We have used Standardscaler from sklearn to perform scaling. Standardscaler uses z score and after scaling mean=0 and standard deviation=1.

1.3- Hierarchical clustering

Dendrogram is imported from scipy.cluster and dendrogram is created by using scaled data and ‘wards’ linkage.Dendrogram is created by using Agglomerative clustering.

Dendrogram:



By looking at the dendrogram we can divide the data into 2 clusters:

Cluster1- yellow colour
-Contain 70 rows
Cluster2- Green colour
-Contain 140 rows

We can get the number of clusters from dendrogram by using the fcluster method.
Import fcluster from scipy.cluster.
In fcluster we are using ‘Wards link’ as linkage and ‘Distance’ as criterion. By giving 25 as the cutoff for distance on y axis we will get 2 clusters.

Cluster 1 - in cluster 1 are those customers who spend more on a monthly basis. Their probability of making full payment is 88%.

Cluster 2 - in cluster 2 are those customers who spend less than cluster 1 on a monthly basis. Their probability of making full payment is 86%.

1.4- K-Means Clustering

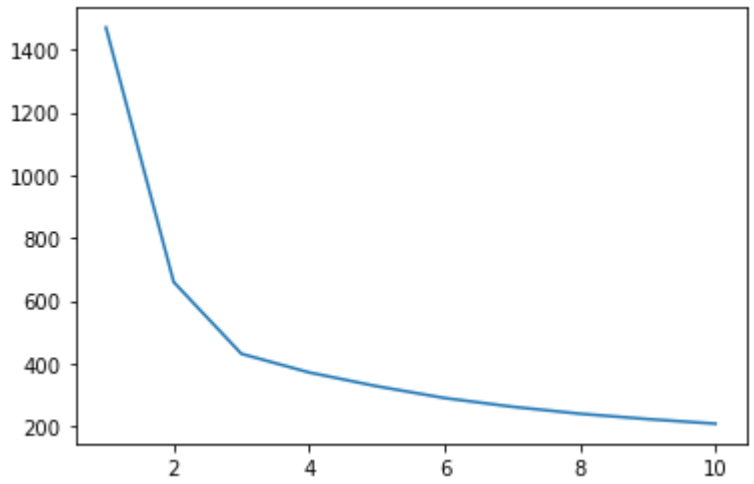
Imported k means from scipy.cluster and applied k means on scaled data.

i) wss calculation(inertia with different values of k)- on performing wss calculation with different values of k ranging from 1 to 10 we have seen a massive decrease in value from k=1 to k=2. From k=2 to k=3 there is no massive drop in value. Hence on the basis of inertia the optimal number of clusters is 2(k=2).

Values of inertia with k ranging from 1 to 10:

k=1, 1469.9999999999998,
k=2, 659.171754487041,
k=3, 430.6589731513006,
k=4, 371.30172127754196,
k=5, 327.3281094192775,
k=6, 289.7743349593589,
k=7, 261.99257202366164,
k=8, 239.74382812197695,
k=9, 222.65810758445298,
k=10, 208.06758801782854

ii)wss plot(Elbow plot)-



Plot shows massive decrease in values at the level of 2 on x axis.

Inference - Based on inertia and wss plot(elbow plot) optimal number of clusters for this data is 2.

Silhouette score is 0.47. It means all rows have been clustered correctly to there corresponding clusters.

1.5-

i) Hierarchical clustering-

Dataset head after adding cluster as a column to original dataset

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
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	2
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

Cluster 1- Description of cluster 1

	count	mean	std	min	25%	50%	75%	max
spending	70.0	18.37	1.38	15.38	17.33	18.72	19.14	21.18
advance_payments	70.0	16.15	0.60	14.86	15.74	16.21	16.56	17.25
probability_of_full_payment	70.0	0.88	0.01	0.85	0.87	0.88	0.90	0.91
current_balance	70.0	6.16	0.25	5.71	5.98	6.15	6.31	6.68
credit_limit	70.0	3.68	0.17	3.27	3.55	3.69	3.80	4.03
min_payment_amt	70.0	3.64	1.21	1.47	2.85	3.63	4.46	6.68

max_spent_in_single_shopping	70.0	6.02	0.25	5.44	5.88	5.98	6.19	6.55
clusters	70.0	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Cluster 2- Description of cluster 2

	count	mean	std	min	25%	50%	75%	max
spending	140.0	13.09	1.55	10.59	11.82	12.77	14.35	16.63
advance_payments	140.0	13.77	0.70	12.41	13.21	13.66	14.30	15.46
probability_of_full_payment	140.0	0.86	0.02	0.81	0.85	0.87	0.88	0.92
current_balance	140.0	5.36	0.23	4.90	5.18	5.35	5.52	6.05
credit_limit	140.0	3.05	0.25	2.63	2.84	3.04	3.23	3.58
min_payment_amt	140.0	3.73	1.63	0.77	2.46	3.60	4.88	8.46
max_spent_in_single_shopping	140.0	5.10	0.23	4.52	5.00	5.09	5.22	5.88
clusters	140.0	2.00	0.00	2.00	2.00	2.00	2.00	2.00

Promotional strategies for cluster 1: As the mean monthly expenditure of this cluster is more, we can give them discount vouchers on our partners websites, So that they can spend more using the credit card.

Promotional strategies for cluster 2: As the mean monthly expenditure of this cluster is less, we can encourage them to spend more by giving cash back using a rewards points based loyalty program .

ii) K means clustering-

Dataset head after adding k means cluster as a column to original dataset

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	k means cluster
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	0
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

Cluster 1(kmeans1)- Description of cluster 1

	count	mean	std	min	25%	50%	75%	max
spending	133.0	12.93	1.43	10.59	11.75	12.72	14.11	15.99
advance_payments	133.0	13.69	0.64	12.41	13.19	13.57	14.21	14.94
probability_of_full_payment	133.0	0.86	0.02	0.81	0.85	0.87	0.88	0.92
current_balance	133.0	5.34	0.21	4.90	5.18	5.33	5.48	5.79
credit_limit	133.0	3.03	0.24	2.63	2.82	3.03	3.20	3.58
min_payment_amt	133.0	3.83	1.61	0.86	2.59	3.64	4.92	8.46

max_spent_in_single_shoppin g	133.0	5.08	0.20	4.52	4.96	5.09	5.22	5.49
k means cluster	133.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Cluster 2 (kmeans 2) - Description of cluster 2

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
spending	77.0	18.16	1.48	15.38	16.84	18.55	19.11	21.18
advance_payments	77.0	16.05	0.64	14.86	15.55	16.18	16.50	17.25
probability_of_full_payment	77.0	0.88	0.02	0.85	0.87	0.88	0.90	0.91
current_balance	77.0	6.13	0.26	5.62	5.92	6.11	6.28	6.68
credit_limit	77.0	3.66	0.19	3.23	3.50	3.68	3.80	4.03
min_payment_amt	77.0	3.48	1.28	0.77	2.55	3.37	4.39	6.68
max_spent_in_single_shoppin g	77.0	5.97	0.29	5.09	5.84	5.96	6.18	6.55
k means cluster	77.0	1.00	0.00	1.00	1.00	1.00	1.00	1.00

Promotional strategies for cluster 1: As the mean monthly expenditure of this cluster is less, we can encourage them to spend more by giving cash back using a rewards points based loyalty program .

Promotional strategies for cluster 2:As the mean monthly expenditure of this cluster is more, we can give them discount vouchers on our partners websites, So that they can spend more using the credit card.

Problem 2- CART-RF-ANN

2.1-

Dataframe head-

	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

Number of columns = 10
Number of rows = 3000

#	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

Based on the info of the dataset there are no null values. There are 3 types of data :
i) Float data type - commission and sales.
ii) Object data type - Agency_code, Type, Claimed, Channel, Product name, Destination.
iii) int64- Age, Duration.

Summary of data:

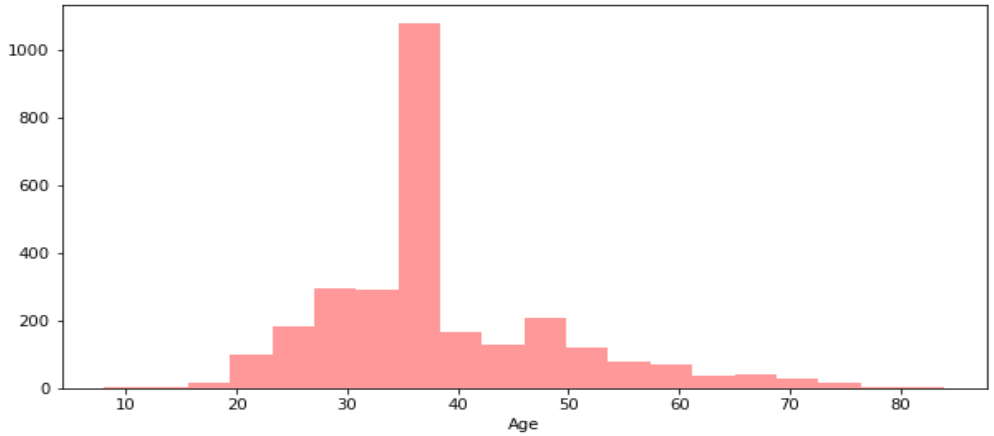
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000	NaN	NaN	NaN	38.091	10.4635	8	32	36	42	84
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000	NaN	NaN	NaN	14.5292	25.4815	0	0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000	NaN	NaN	NaN	70.0013	134.053	-1	11	26.5	63	4580
Sales	3000	NaN	NaN	NaN	60.2499	70.734	0	20	33	69	539
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Based on a summary of data we can see that the data is not normally distributed.

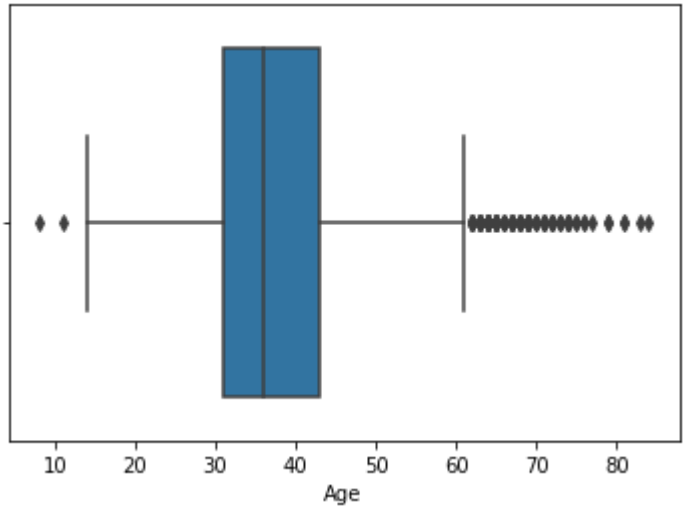
There are 139 duplicate rows. As duplicate rows will affect tree formation by increasing branches we are going to remove duplicate rows.After removing duplicate rows number of rows decreases to 2861.

Univariate analysis-

1.Variable- Age
plot-Distplot
skew:1.1



It is right skewed and mean age of travellers is 38 years

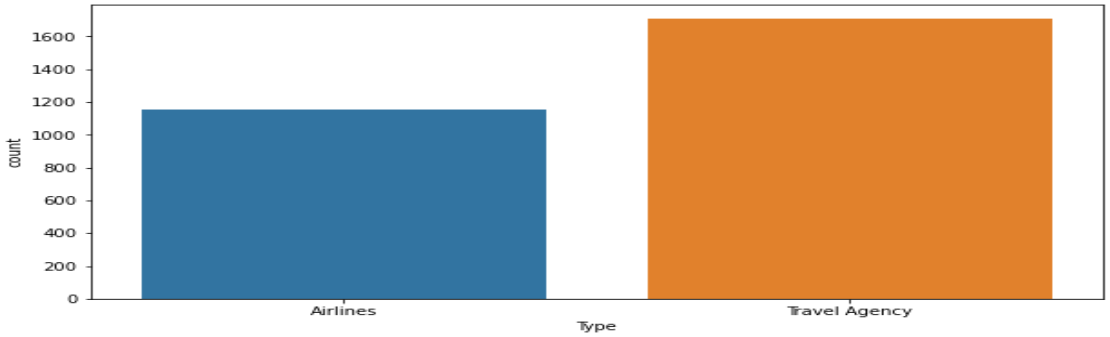


Box plot shows there is outliers in both sides

2.Variable- Agency_code
Plot- count plot

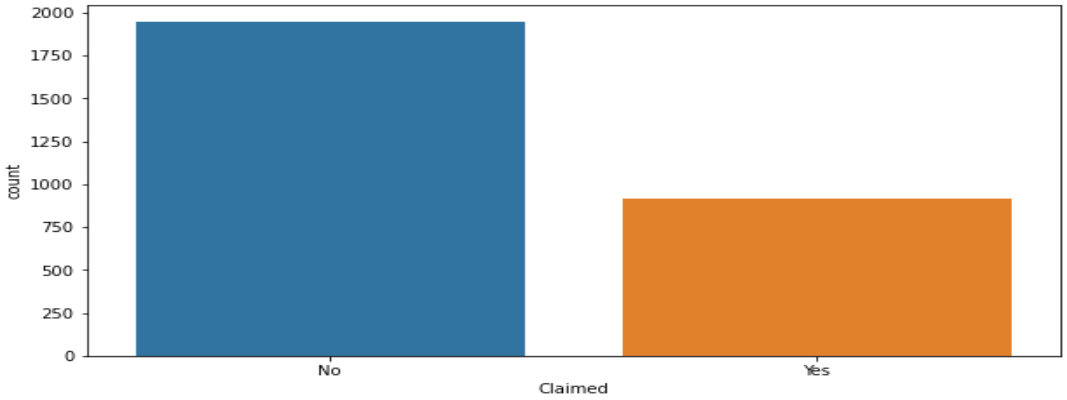
Maximum booking is through agency with code EPX and minimum booking is through agency with code JZI

3. Variable- Type
   Plot- Countplot



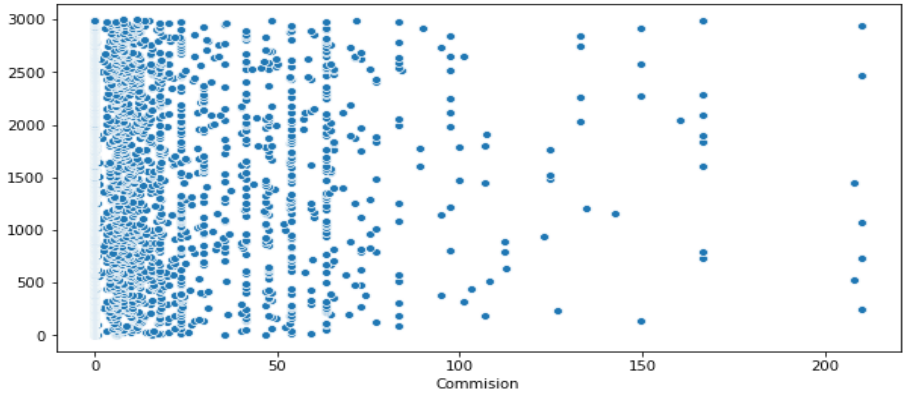
Maximum booking around 1600 bookings are done through a travel agency and around 1150 bookings are through airlines.

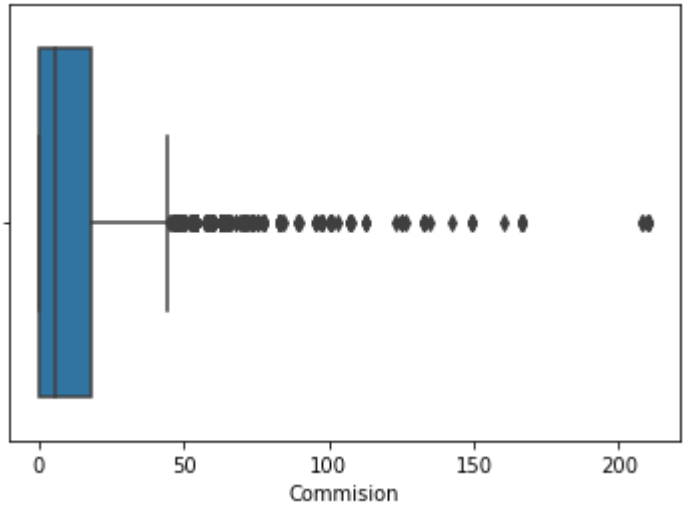
4. Variable- Claimed
   Plot- Countplot



Around 1900 people have claimed insurance and around 900 have not claimed insurance.

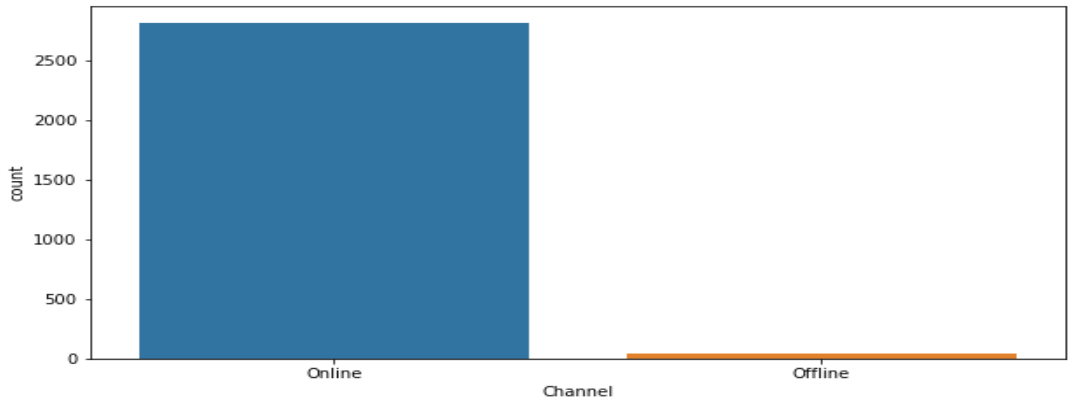
5. Variable- Commision
   Plot- Scatter plot





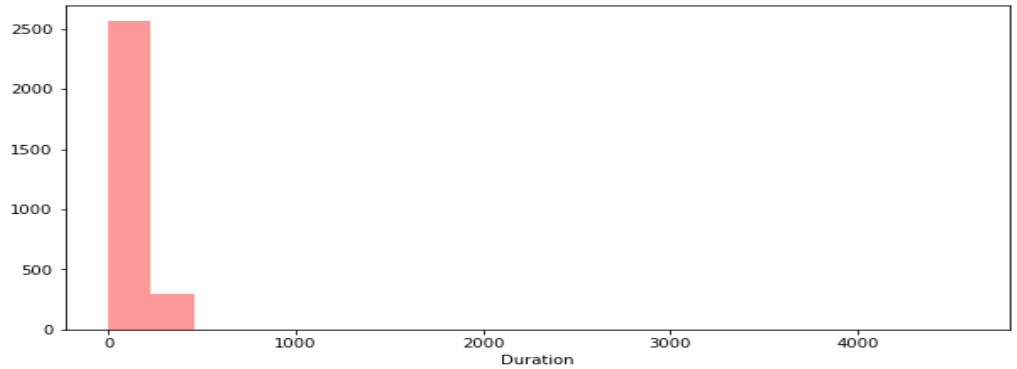
Boxplot shows commision has outliers in the higher side.

6. Variable- Channel
   Plot- Countplot

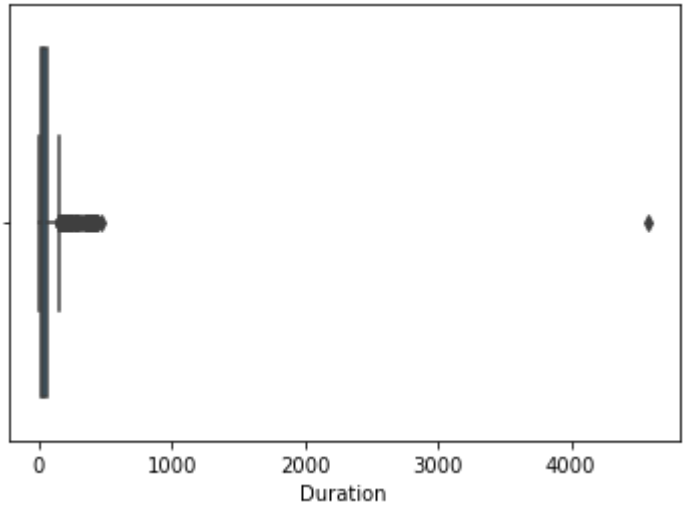


Maximum people >2500 have done their bookings online.

7. Variable- Duration
   Plot- Distplot
   Skew-13.79



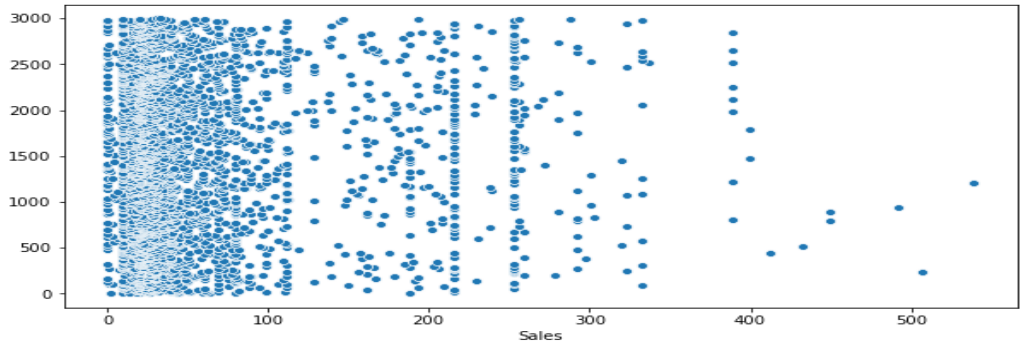
This is highly right skewed.

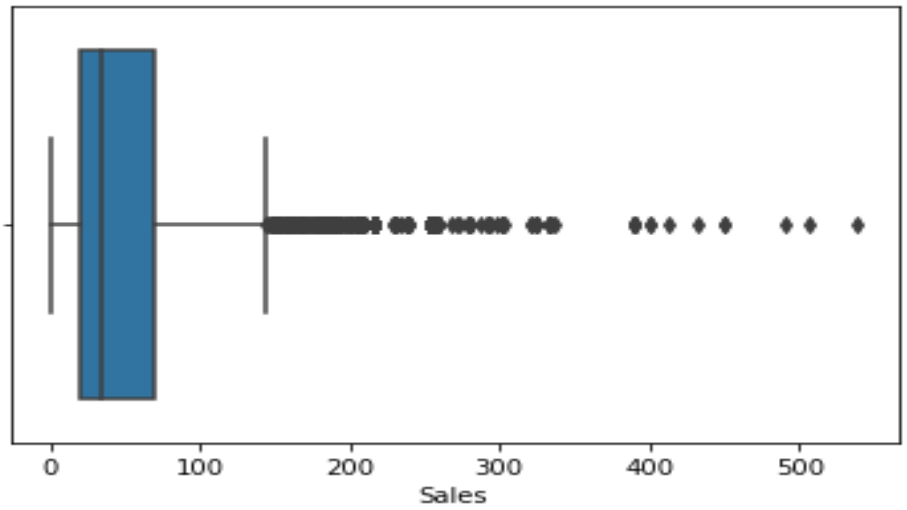


Boxplot shows duration has outliers in the higher side

8. Variable- Sales

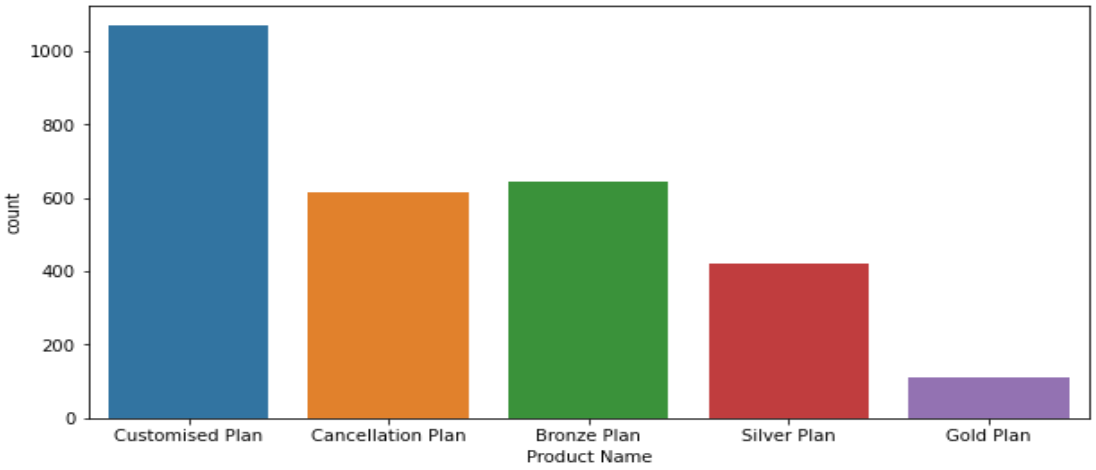
Plot- Scatterplot





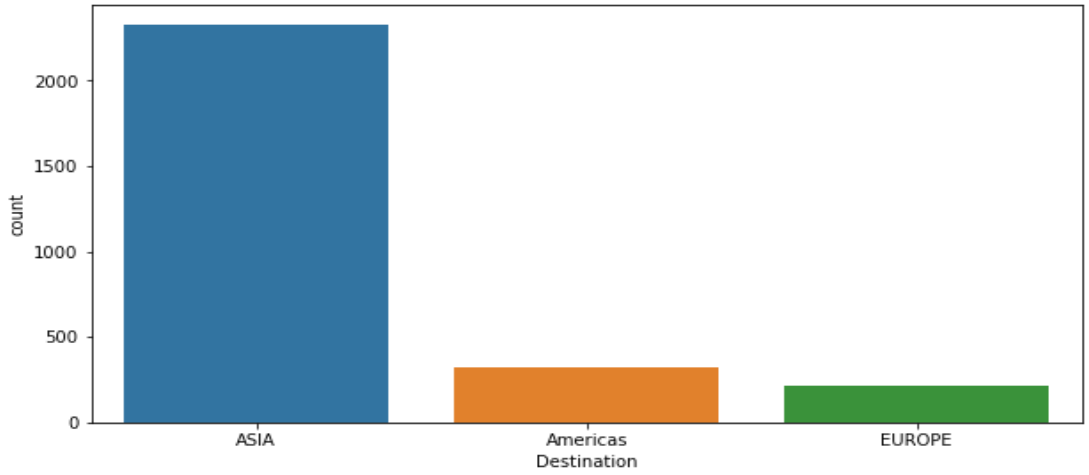
Boxplot shows Sales has outliers in the higher side

9. Variable- Product name
   Plot- countplot



Maximum people (around 1100) have taken customised plans and minimum people(around 100) have taken gold plans.

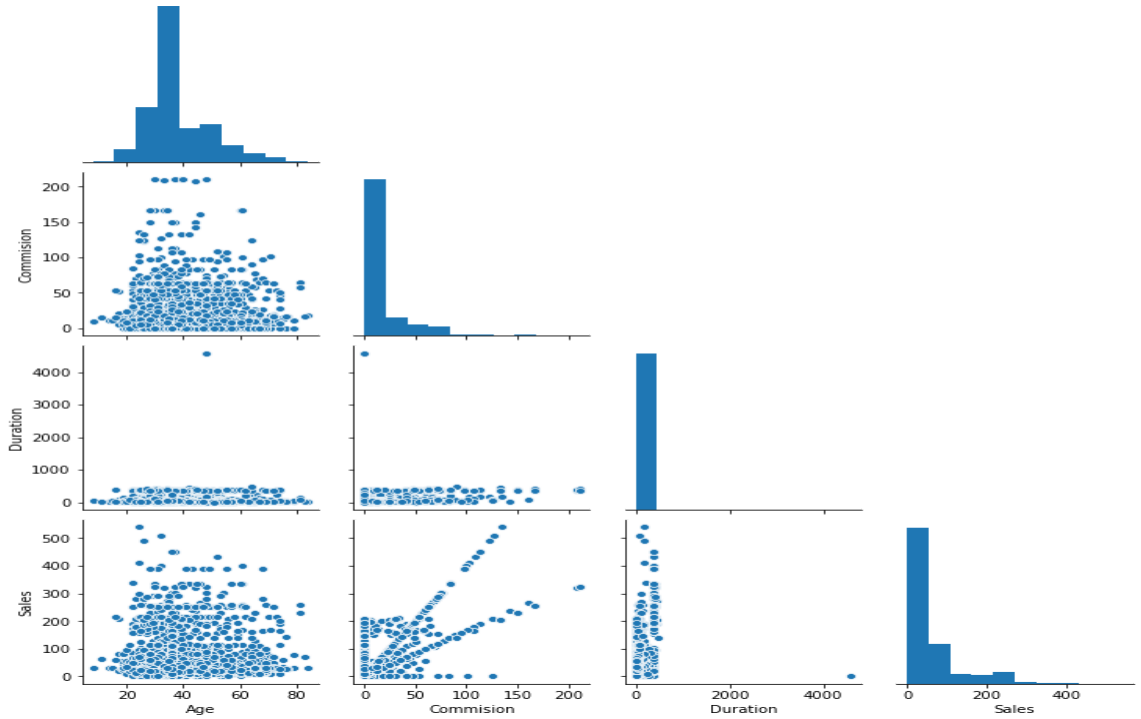
10. Variable- Destination
    Plot- Countplot



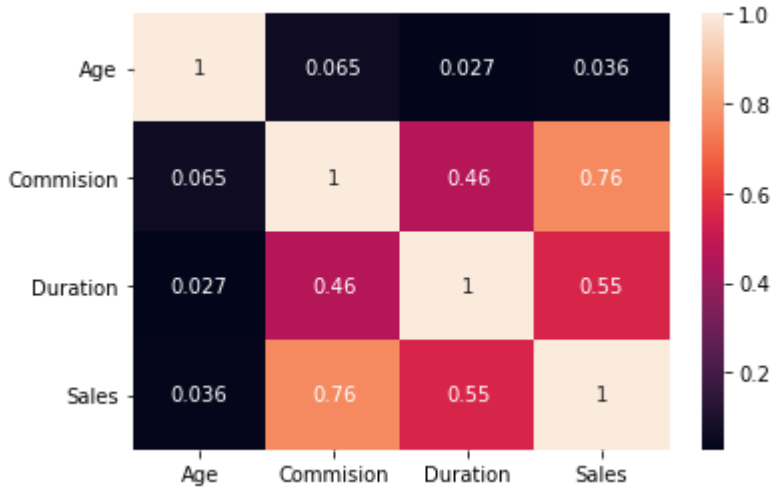
Maximum(>2000) people are travelling to Asia and minimum (<250) people are travelling to Europe.

Bivariate analysis-

Pairplot-



Heatmap-



Pairplot and heatmap do not show strong positive correlation between any variables. There is moderate positive correlation between commission and sales.

2.2-

-All the object type variables are converted into categorical type because cart,random forest and artificial neural networks take only numerical or categorical variables.
- Data set has been split into Dependent variable(Y) and independent variables(X).
- Both dependent and independent variables are then split into 70% train set and 30% test test by using train_test_split function from sklearn.

1) CART Model- cart model has been created by using DecisionTreeClassifier from sklearn by using 'Gini' as a criterion. Decision tree is created by using the Agency_code variable as a root node.
   - As the tree is overgrown pruning is required to prevent overfitting of the dataset.
   -pruning is done by using gridsearch.Prunning is done by using following best gridsearch parameters-
      Max_depth- 8
      Min_samples_leaf- 20
      Min_samples_split- 50

2) Random Forest Model- Random forest model is formed by using RandomForestClassifier from sklearn.
   -Best gridsearch parameters for random forest model are-
      Max_depth- 10
      Max_features- 4
      Min_samples_leaf- 20
      Min_samples_split- 60
      N_estimators- 101

3) Artificial Neural Network Model- neural network model is formed by using MLPclassifier from sklearn.
   -Best gridsearch parameters for neural network are-
      Hidden_layer_sizes- 100
      Max_iter- 10000
      Activation- 'relu'
      Solver- 'adam'
      Tolerance level- 0.01

2.3-

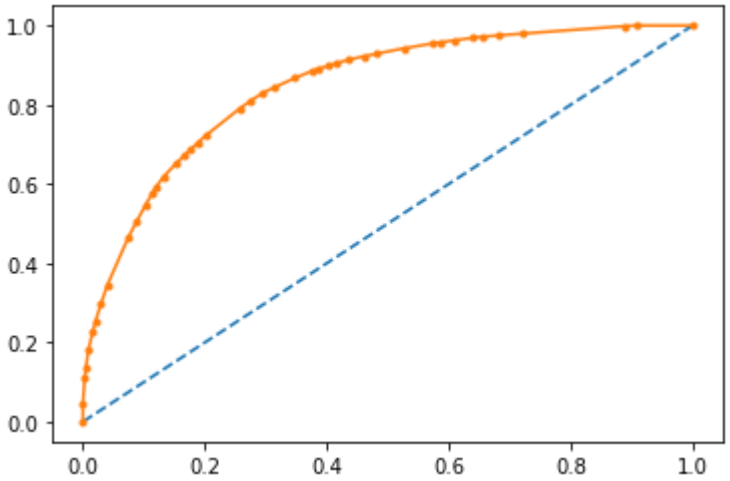
1) CART Model-

i) Train set- By fitting train set into cart model will get following information from classification report and confusion matrix-

Accuracy- 79%
Precision- 82%
Recall- 88%

True negative- 1183
True positive- 341
False negative- 302
False positive- 176

AUC Score-0.846
ROC Curve-



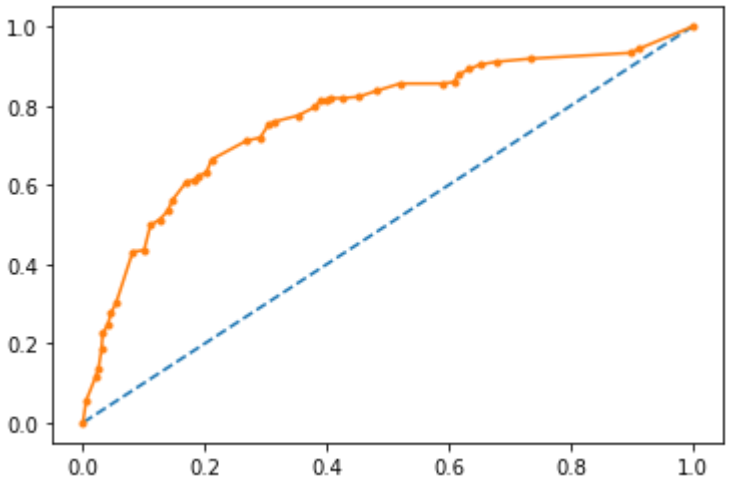
ii) Test set- By fitting test set into cart model will get following information from classification report and confusion matrix-

Accuracy- 76%
Precision- 80%
Recall- 86%

True negative- 506
True positive- 145
False negative- 126

False positive- 82

AUC Score- 0.769
ROC Curve-



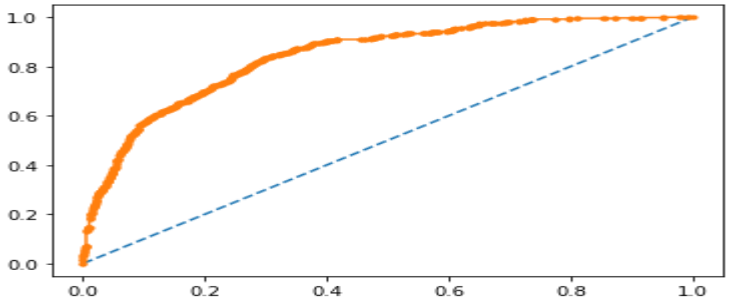
2) Random Forest Model-

_____i) Train set- By fitting train set into cart model will get following information from classification report and confusion matrix-

Accuracy- 79%
Precision- 82%
Recall- 90%

True negative- 1220
True positive- 369
False negative- 139
False positive- 176

AUC Score- 0.844
ROC Curve-

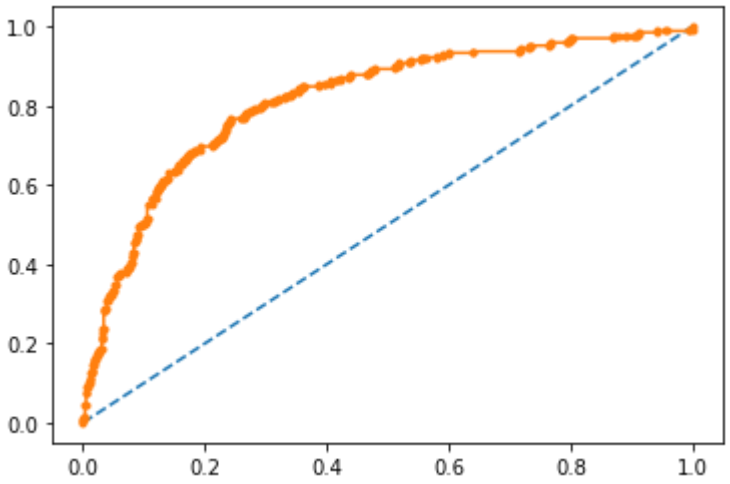


ii) Test set- By fitting test set into cart model will get following information from classification report and confusion matrix-

Accuracy- 78%
Precision- 81%
Recall- 89%

True negative- 525
True positive- 149
False negative- 122
False positive- 63

AUC Score- 0.815
ROC Curve-



3) Artificial Neural Network-

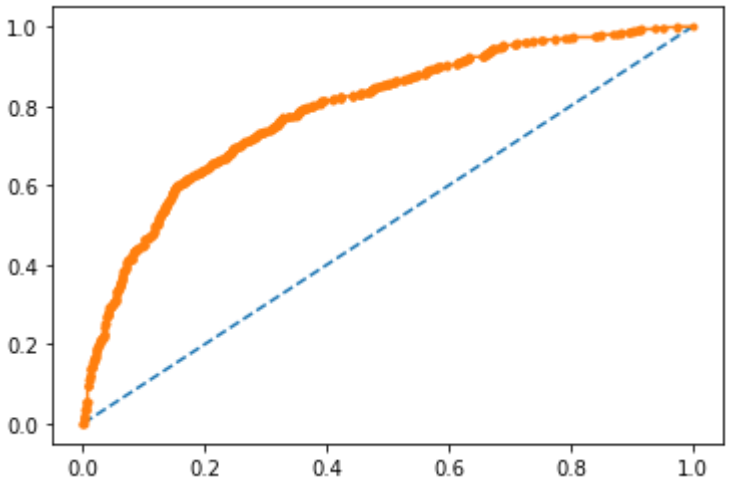
i) Train set- By fitting train set into cart model will get following information from classification report and confusion matrix-

Accuracy- 86%
Precision- 80%
Recall- 87%

True negative- 1183
True positive- 341
False negative- 302
False positive- 176

AUC Score-0.79

ROC Curve-

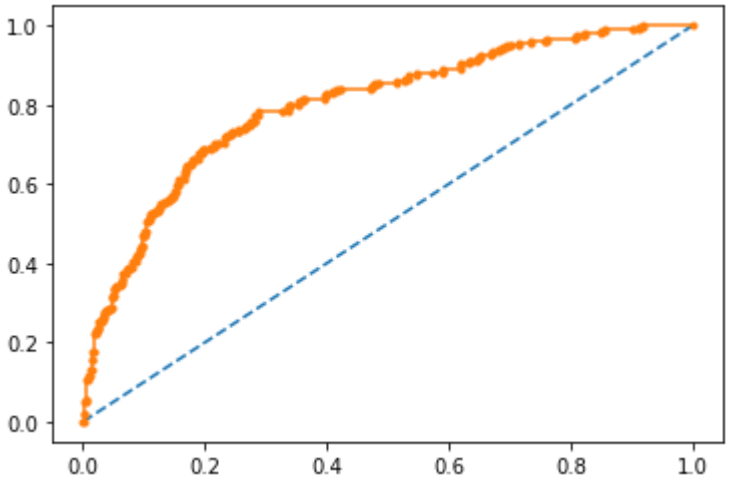


ii)_Test set- By fitting test set into cart model will get following information from classification report and confusion matrix-

Accuracy- 77%
Precision- 80%
Recall- 89%

True negative- 523
True positive- 142
False negative- 129
False positive- 65

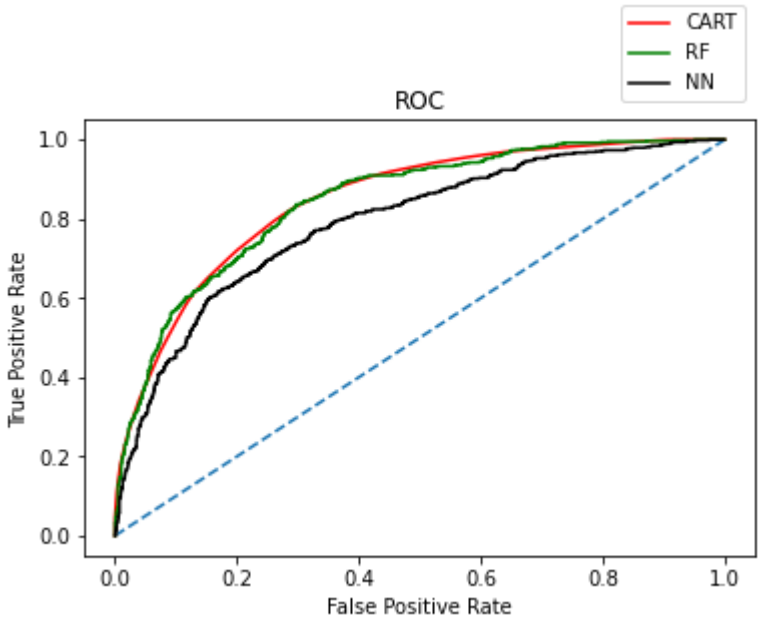
AUC Score- 0.80
ROC Curve-



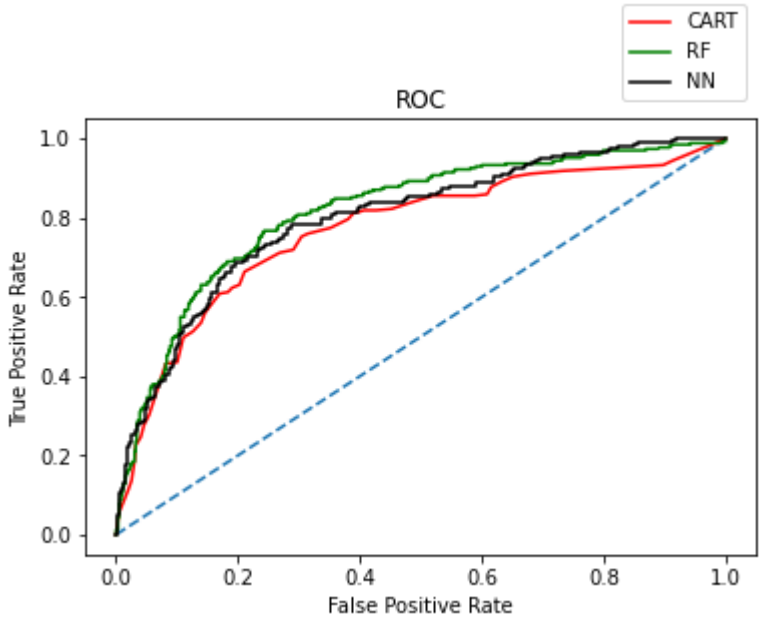
2.4-

- Based on accuracy random forest model is better
- Based on precision artificial neural network is best
- comparison of Roc curve-

Train set-



Test set-



Based on the graph we can see that the random forest model is performing best for both train and test data.

- Based on AUC score, the Random forest model is better than other 2 models.

2.5-

We will give them following recommendations-

i) To use the random forest model for future prediction of insurance claims.

ii) Model is 79% accurate.

iii)Model has 82% precision.