2021

GREAT LEARNING

PRAJOTH

[DATA MINING]

Table of Contents:

List Of Tables:

List OF Figures:

Problem 1: Clustering

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

Data Description:

1. spending : Amount spent by the customer per month (in 1000s)

2. **advance_payments :** Amount paid by the customer in advance by cash (in 100s)

3. **probability_of_full_payment :** Probability of payment done in full by the customer to the bank

4. **current_balance :** Balance amount left in the account to make purchases (in 1000s)

5. **credit_limit :** Limit of the amount in credit card (10000s)

6. **min_payment_amt :** minimum paid by the customer while making payments for purchases made monthly (in 100s)

7. **max_spent_in_single_shopping :** Maximum amount spent in one purchase (in 1000s)

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

Reading the dataset:

	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 1: Sample of Dataset

- Dataset contain 210 rows and 7 columns.

Checking info:

```
<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
```

Observation:

- Dataset contain seven float data types.
- There is no null value in data.
- There is no duplicate values in data.

Summary of Dataset:

	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

Table 2: Summary of Dataset

Univariate Analysis:

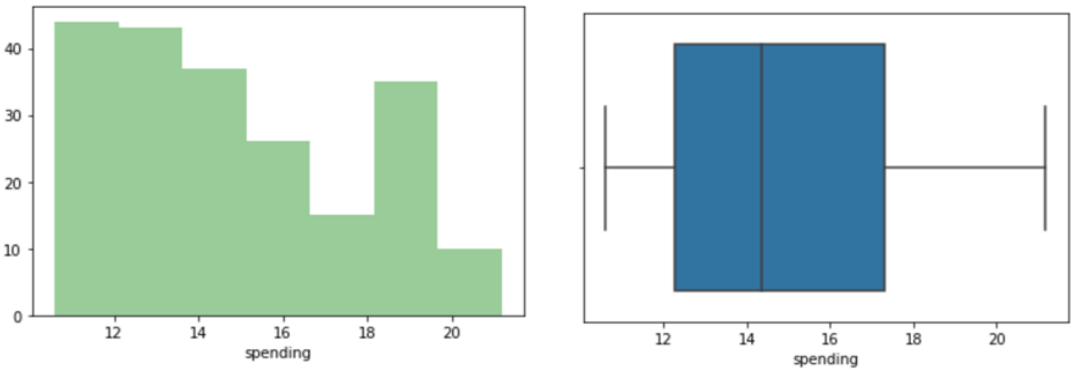

Figure 1: Spending

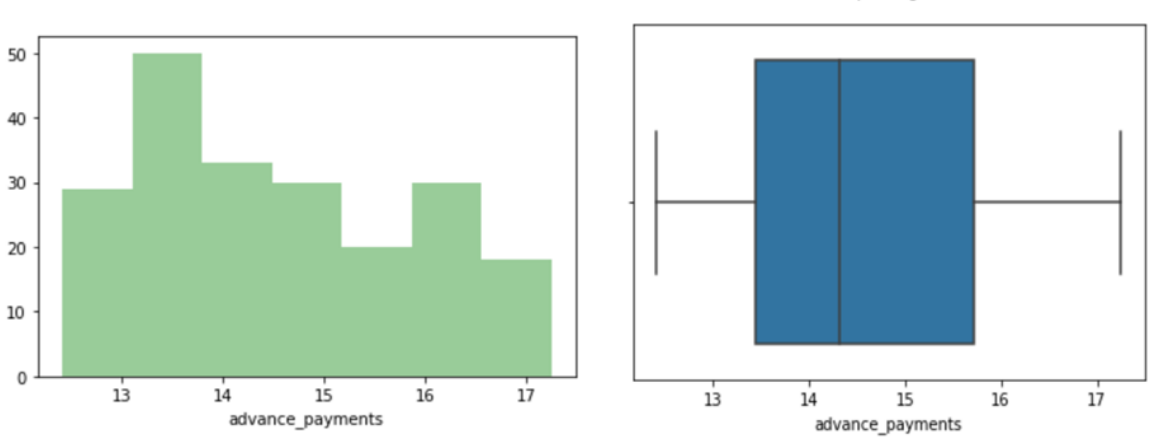

Figure 2: Advance Payments

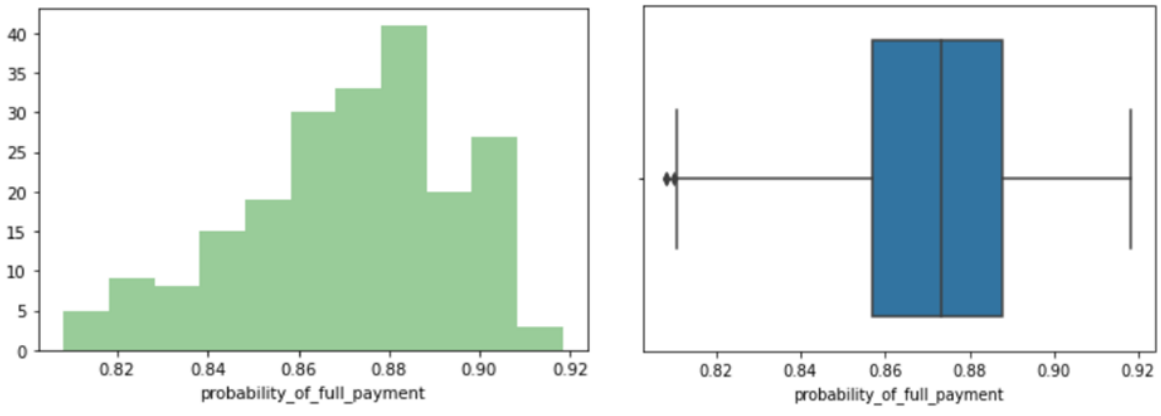

Figure 3: Probability of Full Payment

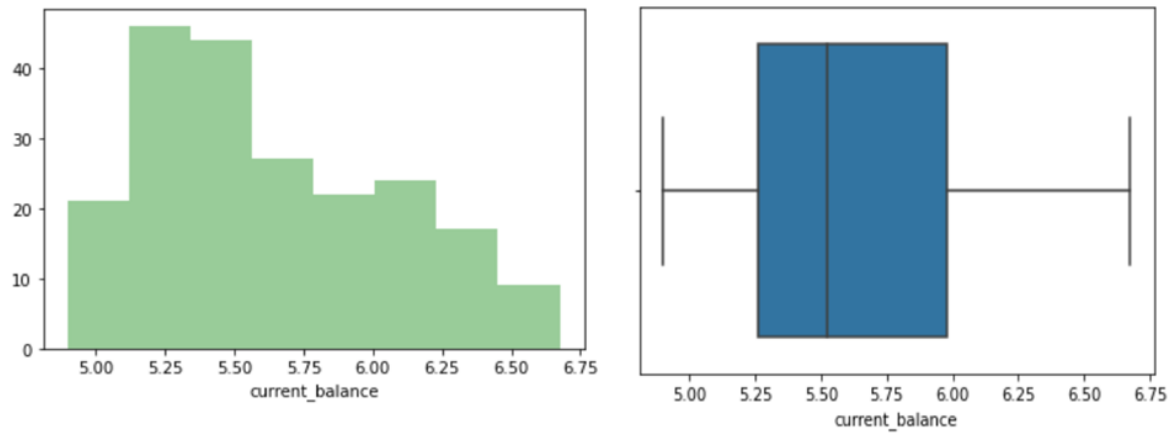

Figure 4: Current Balance

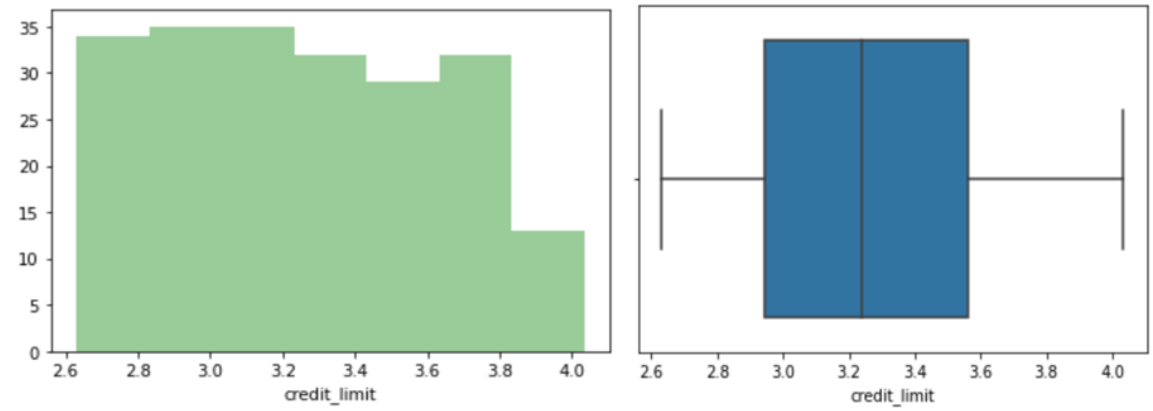

Figure 5: Credit Limit

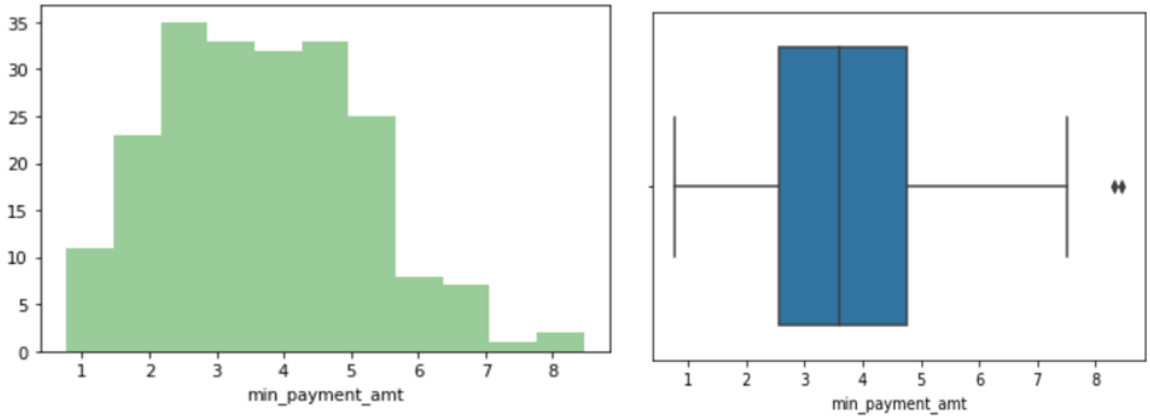

Figure 6: Minimum Payment Amount

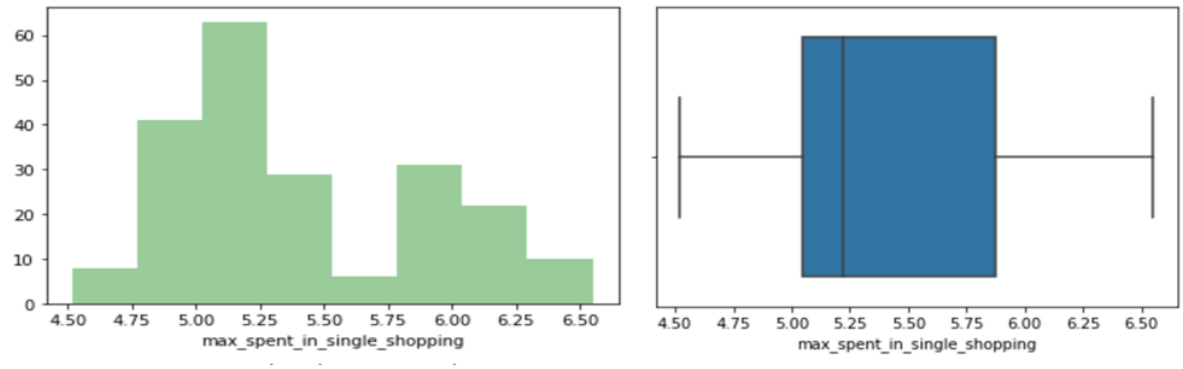

Figure 7: Maximum spend in Single Shopping

Observation:

- There are 7 features in dataset.
- Spending, Advance payment, current balance, credit limit, maximum spend in single shopping has no outliers.
- Probability of full payment, min amount payment has outliers.
- Means and Modes of all features seem to be equal.
- spending ranges from 10.590000 to 21.180000.
- Probability of full payment is left skewed.

Bivariate Analysis:

Pair Plot with Continuous Variable:

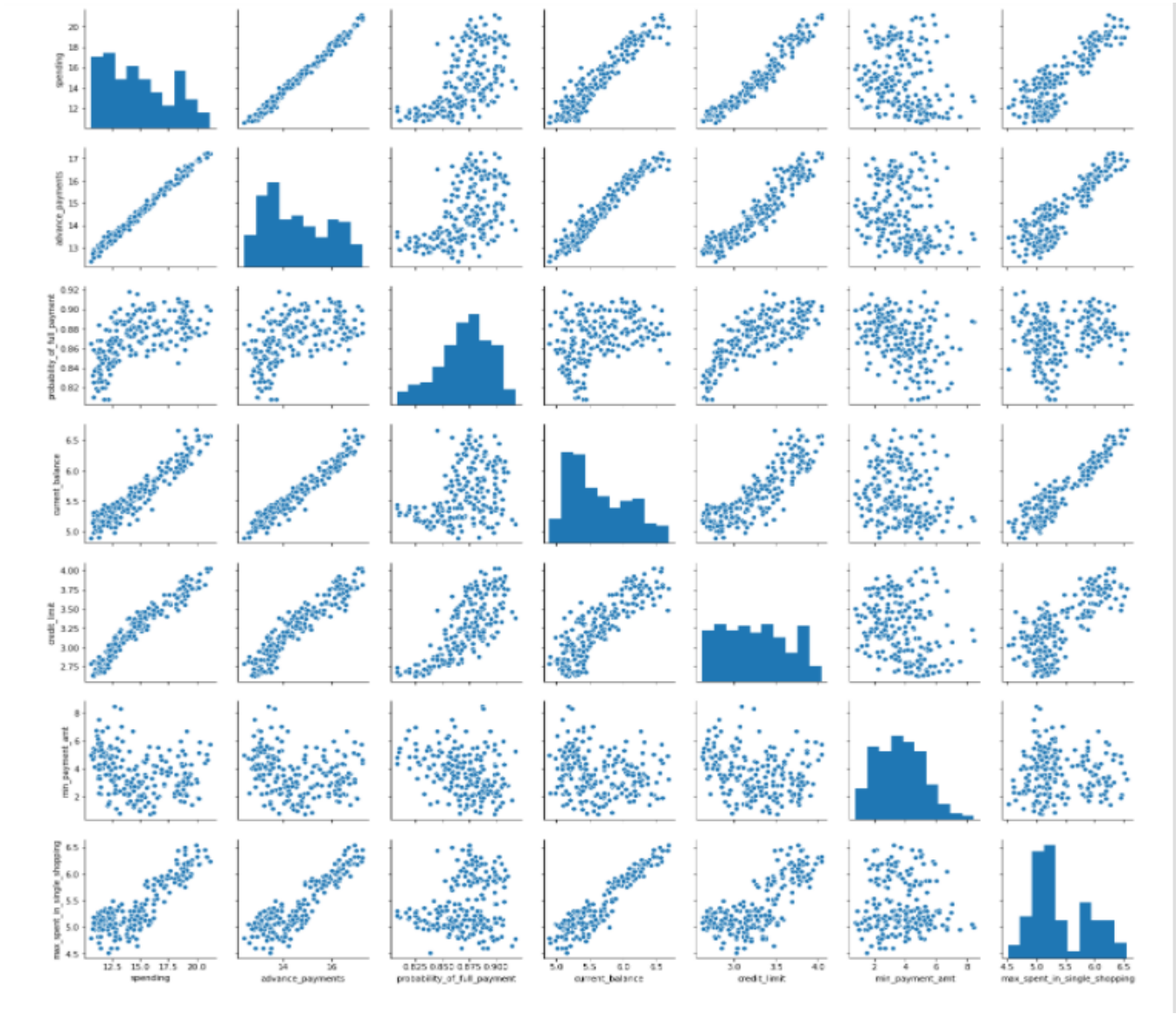

Figure 8: Pair Plot with Continuous Variable

Observation:

- High spending customers are good with advance payment, maintaining current balance, credit limit and max spent in single shopping.
- Customer having high current balance spending maximum amount in single shopping
- High credit limit customers doing more spending and advance payment.
- Probability of full payment and maximum payment amount are distributed abnormally with all features.

Heatmap:

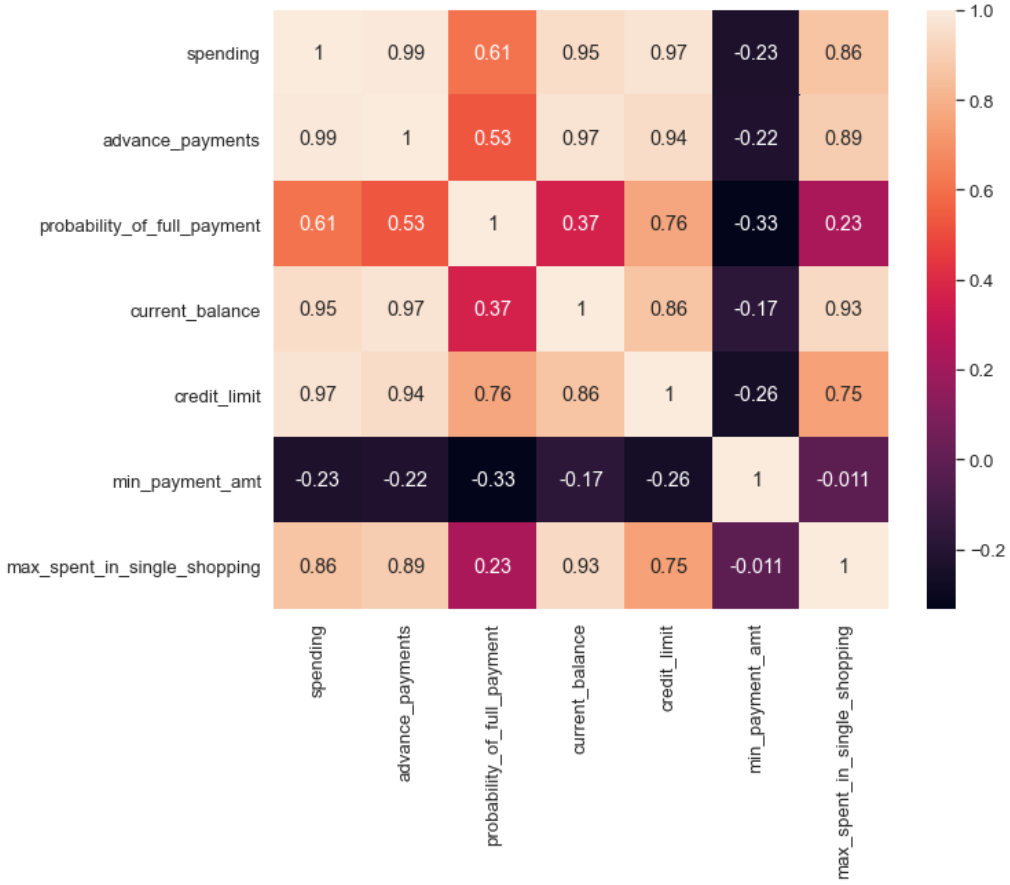

Figure 9: Heat Map

Observation:

- Current balance are highly correlated with spending, advance_payments,credit_limit and max spent in single shopping.
- Spending are highly correlated with advance payment, current balance, credit_limit and max spent in single shopping.
- Min payment amount is negatively correlated with all features.

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

- Scaling the data

- Yes. Clustering algorithms need feature scaling before they are fed to the algorithm. Since, clustering techniques use Euclidean Distance to form the cohorts, it will be wise.
- In clustering, you calculate the similarity between two examples by combining all the feature data for those examples into a numeric value. Combining feature data requires that the data have the same scale.

	0	1	2	3	4	5	6
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 3: Scaled Data

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

Hierarchical clustering:

- We use scaled data to implementing Wards linkage method for hierarchical clustering
- For Visualization we use Dentrogram.

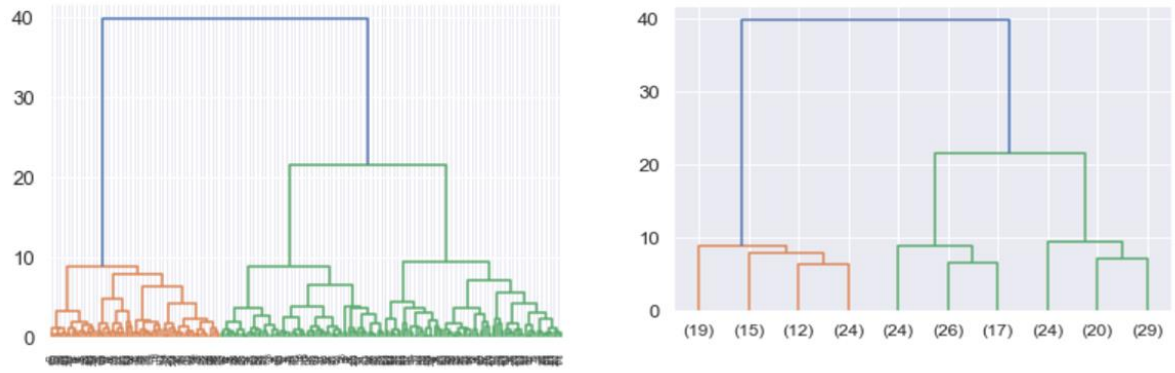

Figure 10: Dentrogram

Observation:

- The above dendrogram indicates all the data points have clustered to different clusters by wards method.
- To find the optimal number cluster through which we can solve our business objective we use truncate mode = lastp.
- Wherein we can give last p = 10 according to industry set base value. Data is formed in to 10 Clusters which shows in X-axis..

	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	3
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

Table 4: Hierarchical Clustering

Agglomerative Clustering:

- Agglomerative clustering uses **a bottom-up approach**, wherein each data point starts in its own cluster. These clusters are then joined greedily, by taking the two most similar clusters together and merging them.
- Set n_clusters=3, affinity='euclidean', linkage='ward' and store the result in another object 'Cluster agglo'.

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

Table 5 : Agglomerative Clustering

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.

Forming clusters with K = 1,2,3,4,5, and comparing the WSS:

cluster	Inertia for kmeans
n_clusters = 1	1469.99
n_clusters = 2	659.17
n_clusters = 3	430.65
n_clusters = 4	371.38
n_clusters = 5	327.21

Table 6: Inertia for K-means

Calculating WSS for other values of K - Elbow Method:

```
wss
```

```
[1469.9999999999998,
 659.171754487041,
 430.6589731513006,
 371.74655984791394,
 326.30618276116064,
 288.7694577022641,
 262.59786377312463,
 240.36349487462357,
 222.14573583325546,
 205.7991764666054,
 194.96722579215714,
 183.02445428635207,
 172.52016740220097]
```

```
a=[1,2,3,4,5,6,7,8,9,10,11,12,13]
```

```
sns.pointplot(a, wss)
```

Graph is drawn between inertia value for K-Means and Clusters(a), it is gradually decreasing between cluster 3 and cluster4. so we need to find **silhouette score between two clusters.**

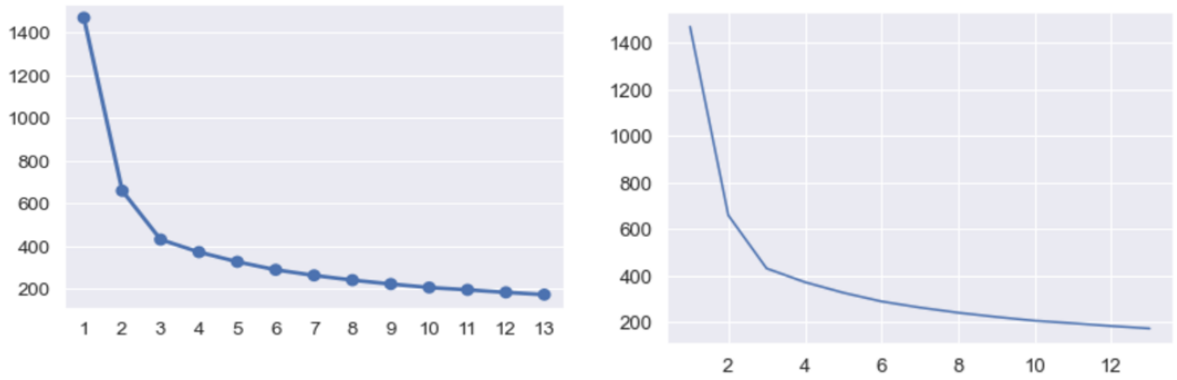

Figure 11: Graph

Cluster evaluation for 3 & 4 clusters: the silhouette score:

Clusters	Inertia for kmeans
n_clusters = 3	0.4007
n_clusters = 4	0.3276

Table 7 : silhouette score for Clusters

Silhouette score is better for 3 clusters when compared to cluster 4. So, final clusters will be 3.

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

Cluster Profiling for hierarchical clustering:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq	Group
clusters									
1	18.371429	16.145429	0.8844	6.158171	3.684629	3.639157	6.017371	70	High Spending
2	11.872388	13.257015	0.848072	5.23894	2.848537	4.949433	5.122209	67	Low Spending
3	14.199041	14.233562	0.87919	5.478233	3.226452	2.612181	5.086178	73	Medium Spending

Table 8 cluster profiling for hierarchical cluster

Cluster Profiling for K-Means Clustering:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	freq	Group
Clus_kmeans3									
0	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803	71	Medium spending
1	11.856944	13.247778	0.848253	5.23175	2.849542	4.742389	5.101722	72	Low spending
2	18.495373	16.203433	0.88421	6.175687	3.697537	3.632373	6.041701	67	High spending

Table 9 Cluster Profiling for K-Means Clustering

Cluster 0: Medium spending customers.

Cluster 1: Low spending customers.

Cluster 3: High spending customers.

Observation:

- Hierarchical clustering and K Means clustering is almost equal.
- In all clusters customers are good with max spent in single shopping and probability of full payment

Group 1: High Spending Customer

- Giving any reward points might increase their purchases.
- Bank need to tie up with top brands to provide best offers and cash back to customers
- Increase their credit limit and – Increase spending habits.
- Give loan against the credit card, as they are customers with good repayment record.
- Maximum max spent in single shopping is high for this group, so can be offered discount/offer on next transactions upon full payment.

Group 2: Medium Spending customer:

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

Group 3: Low Spending Customer:

- Customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate.
- Bank needs more concentration on promotions and advertisement cash backs and offers with more brands to increase the revenue from credit card sector.

Recommendation:

- Bank need to insist customer to use a credit card regularly and responsibly pay off as much as you can every month. ... If you keep your balance reasonably low and make on-time payments every month, you'll contribute to the positive growth of your credit history and scores.
- Bank need to encourage customers by explaining Some common advantages of credit card to use regurarly,

    ->Paying for purchases over time.

    ->Convenience.

    ->Credit card rewards.

    ->Fraud protection.

    ->Free credit scores.

Problem 2: CART-RF-ANN

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

Attribute Information:

1. Target: Claim Status

2. Agency Code: Age Code of tour firm

3. Type : Type of tour insurance firms

4. Channel: Distribution channel of tour insurance agencies

5. Product: Name of the tour insurance products

6. Duration: Duration of the tour

7. Destination: Destination of the tour

8. Sales: Amount of sales of tour insurance policies

9. Commission: The commission received for tour insurance firm

10 Age: Age of insured

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

Reading the dataset:

	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 10 Sample of Data

Observation:

- Dataset contain 3000 rows and 10 columns.
- Claimed is the target variable.

Info of dataset:

- Age and Duration are Integer type
- Commision and Sales are Float type.
- Agency code, Type, Claimed, Channel, Product name and Destination are Object type.
- So dataset contain two integer, two float and six object type variables.

Summary the dataset:

	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

Table 11: Summary of Data

Observation:

- There is no null values in dataset.
- There is 139 duplicate value in dataset. Since don't have unique identifier in dataset these duplicates may be a different customers so I am analyzing model with duplicates.

univariate Analysis:

Numerical Type:

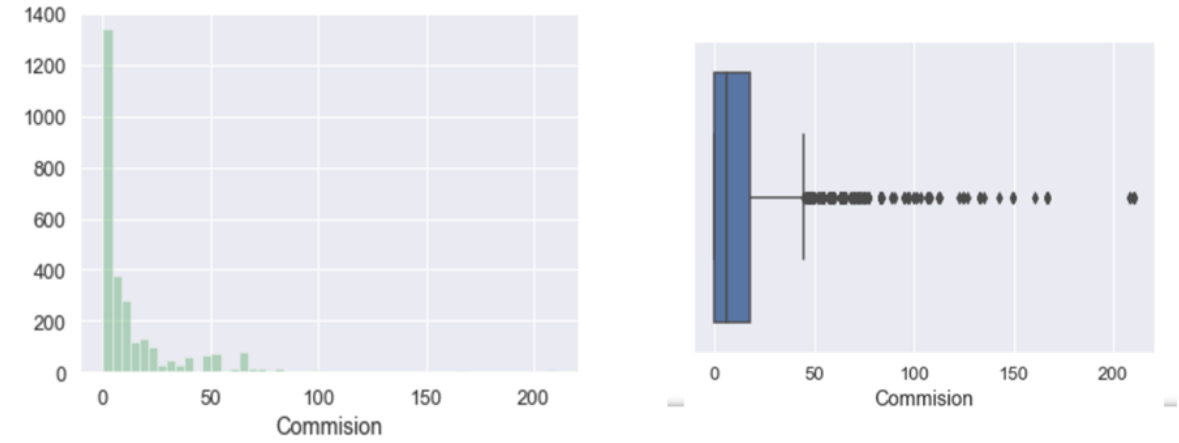

Figure 12: Commission

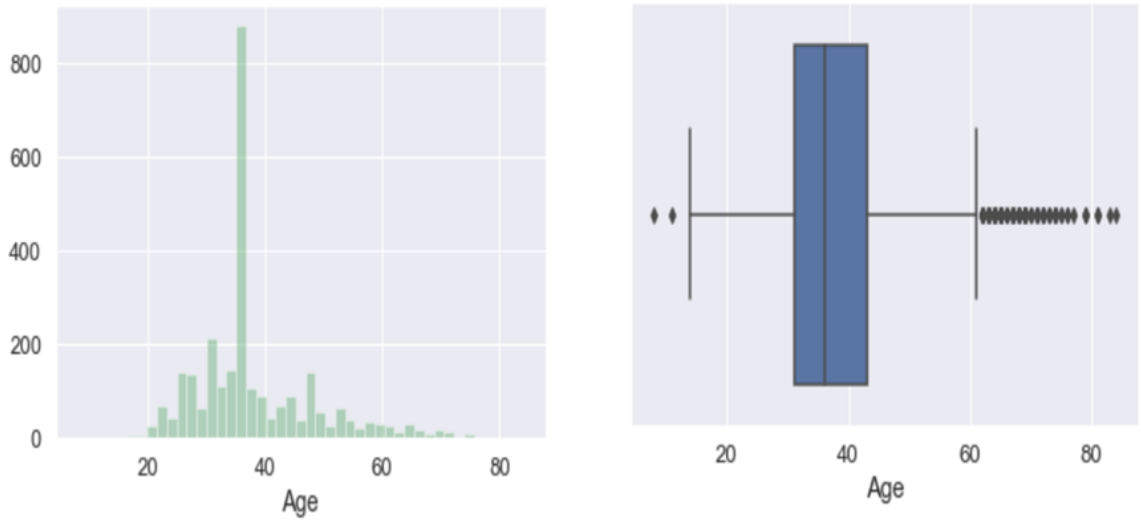

Figure 13: Age

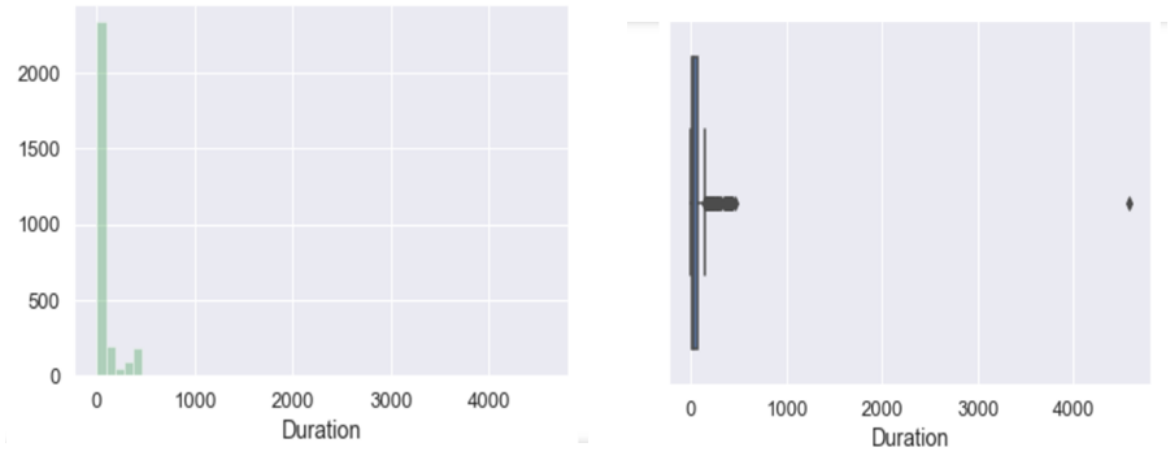

Figure 14: Duration

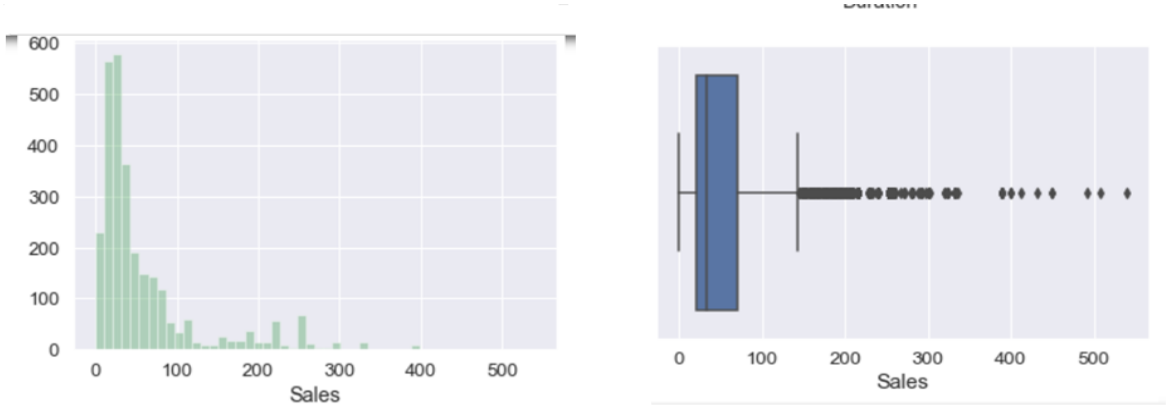

Figure 15: Sales

Observation:

- There are four numerical type in dataset.
- There are outliers in all the variables. Random Forest and ANN can handle the outliers.
- Minimum age is 8 and maximum age is 84.It is positively skewed. Age is distributed more between 30 to 50.
- Commision , sales and duration is right skewed.

Categorical Type:

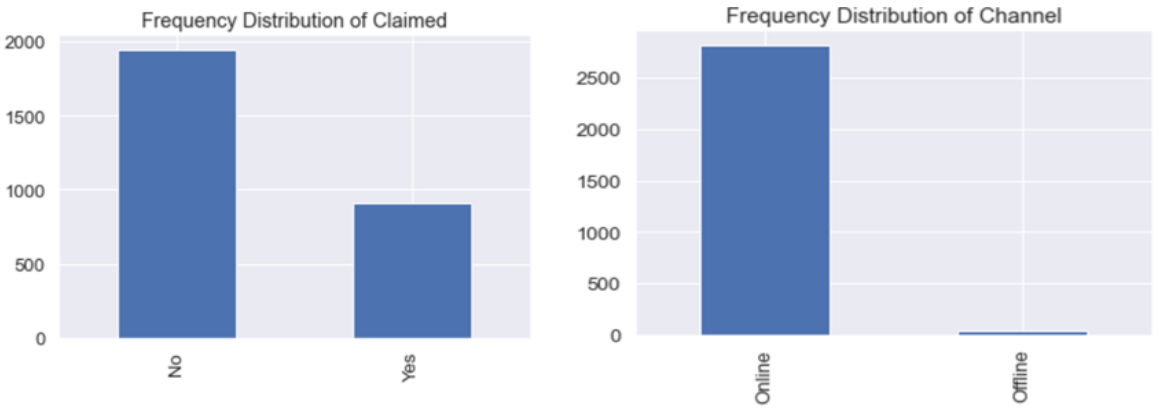

Figure 16: Bar chart for Claimed & Channel

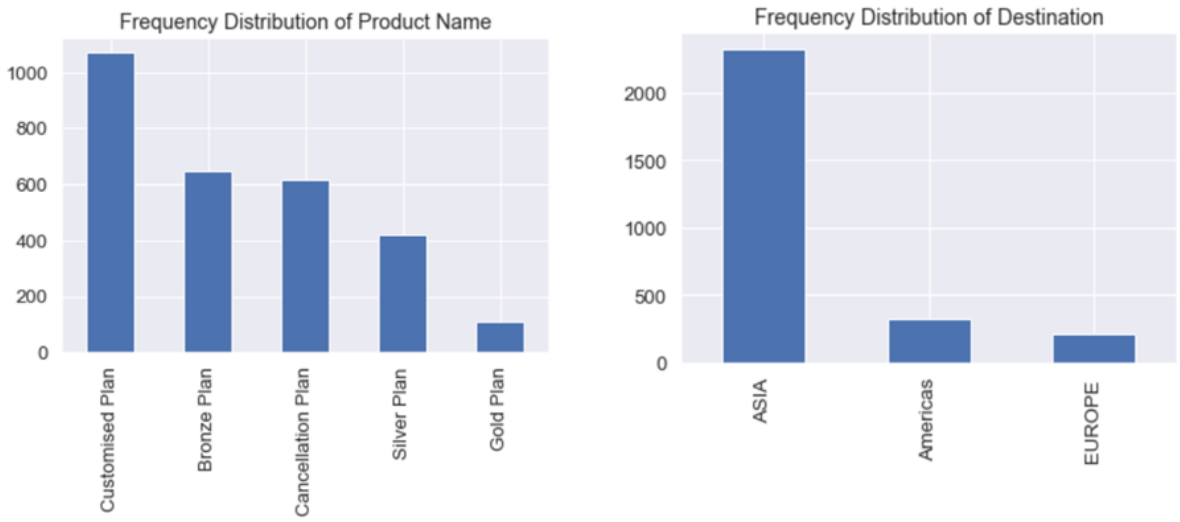

Figure 17: Bar Chart for Product & Destination

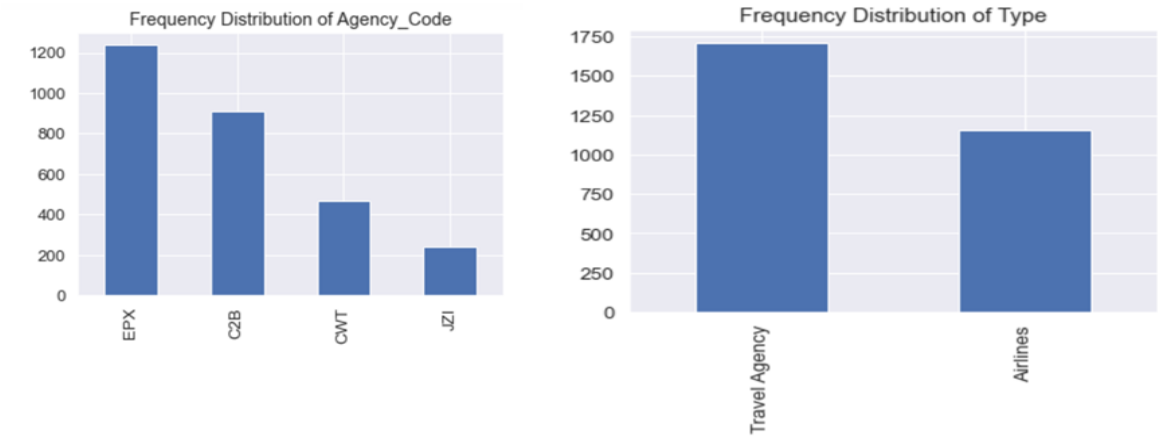

Figure 18: Bar Chart for Agency Code & Type

Observation:

- EPX has high number of customers which is 1238 and JZI has low number of customers which is 239.
- Customers are highly preferred travel agency than airlines.
- 914 customers has climed and 1947 customers are not climed
- Customers more preferred to use online platform than offline.
- More customers preferred customized plan than less customers choses gold plan.
- Asia has preferred more number of customers,second is America and Third is Europe.

Bivarient Analysis:

Agency code:

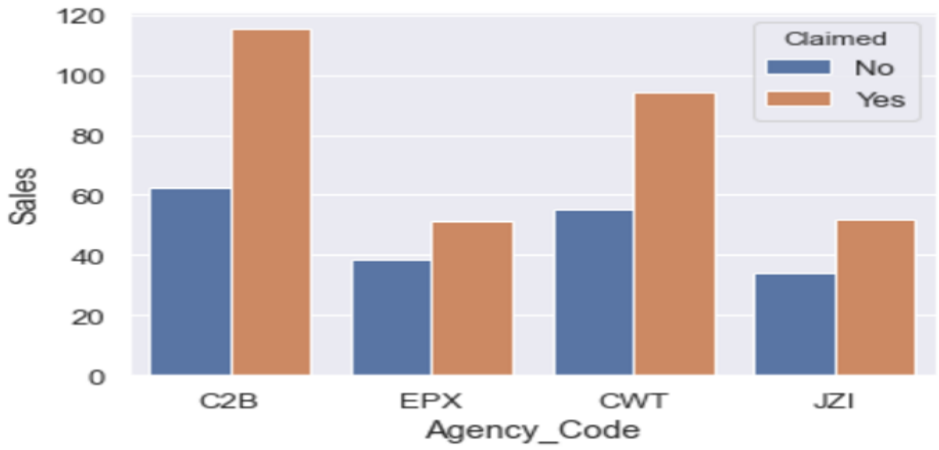

Figure 19: Bar Chart with Agency code & Sales with Claimed

Observation:

- The bar plot shows the split of sales with different agency code and also hue having claimed column.
- It seems that C2B have claimed more claims than other agency.

Type:

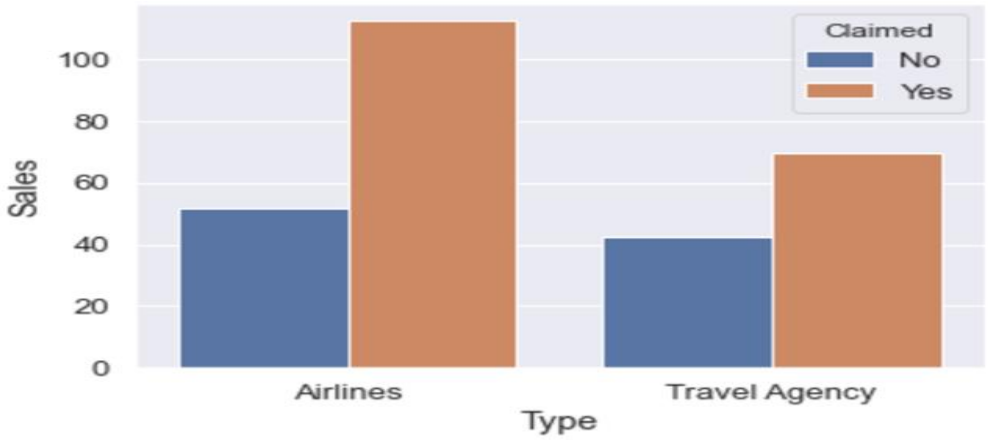

Figure 20 : Bar Chart for Type & Sales with Claimed

Observation:

- The bar plot shows the split of sales with different Type and also hue having claimed column.
- It seems that airlines have claimed more claims than travel agency.

Channel:

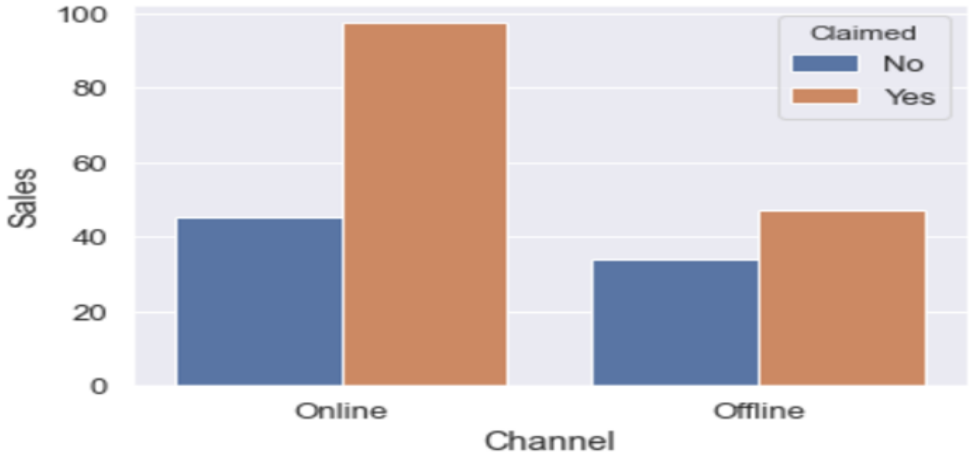

Figure 21:Bar chart for Channel & Sales with Claimed

Observation:

- The bar plot shows the split of sales with different channel and also hue having claimed column.
- It seems that online have claimed more claims than offline.

Product Name:

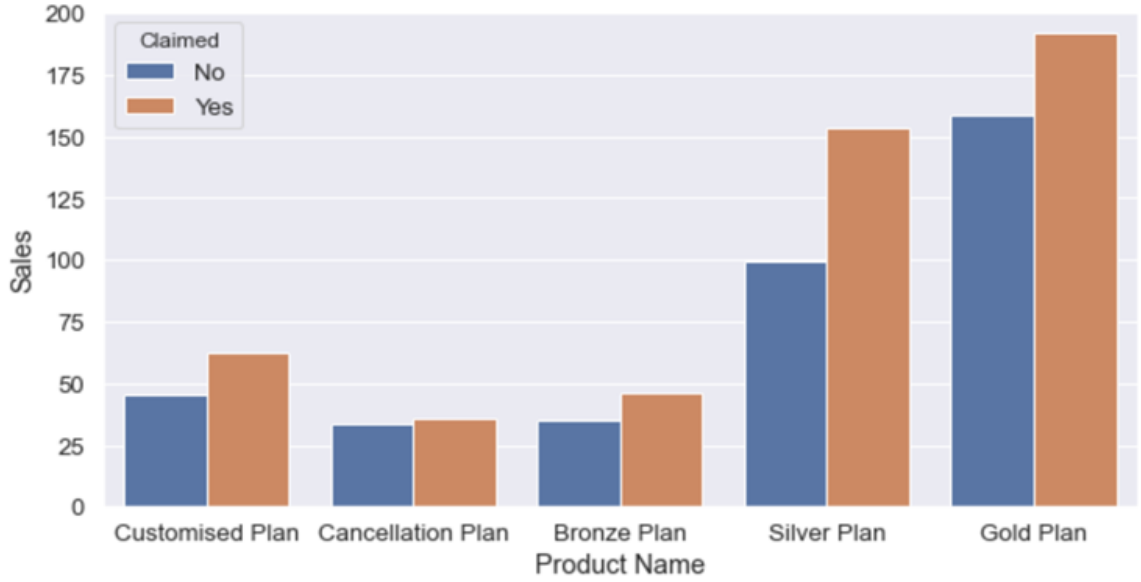

Figure 22: Bar Chart for Product & Sales with Claimed

Observation:

- The bar plot shows the split of sales with different Product name and also hue having claimed column.
- Gold plan has Claimed more than products.

- Gold plan has less customers but sales rate is high. so agency can get good profit.

Destination:

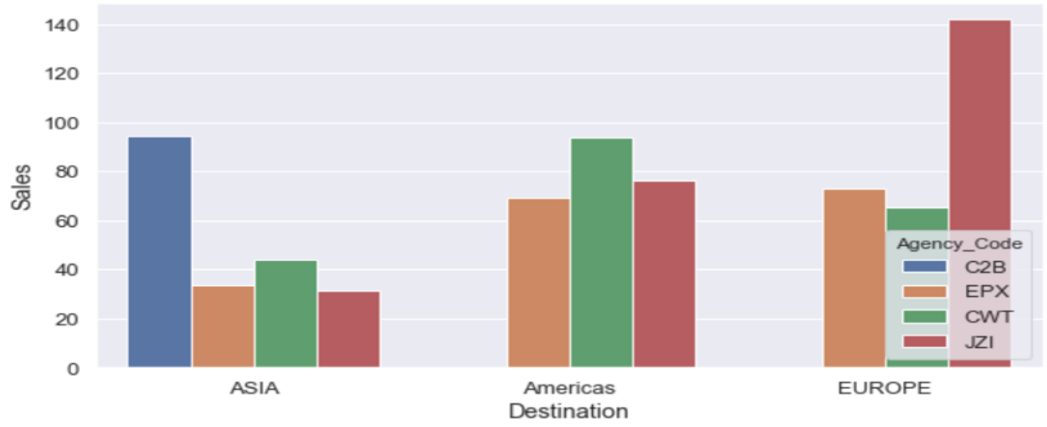

Figure 23: Bar Chart for Destination & Sales with agency Code

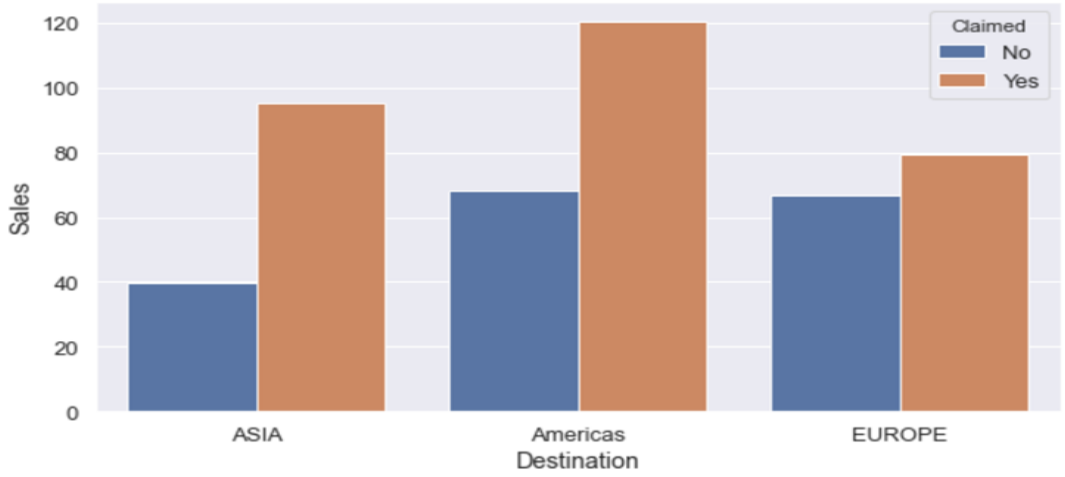

Figure 24: Bar Chart for Destination & Sales with Claimed

Observation:

- The bar plot shows the split of sales with different Destination and also hue having claimed column.
- customers who preferred America has Claimed more than other places
- Asia has highest frequency of customer but claiming rate is less than America.
- None of the customer travel to america and Europe fails to choose C2B agency
- Checking pairwise distribution of the continuous variables

Pair Plot with Continuous variable:

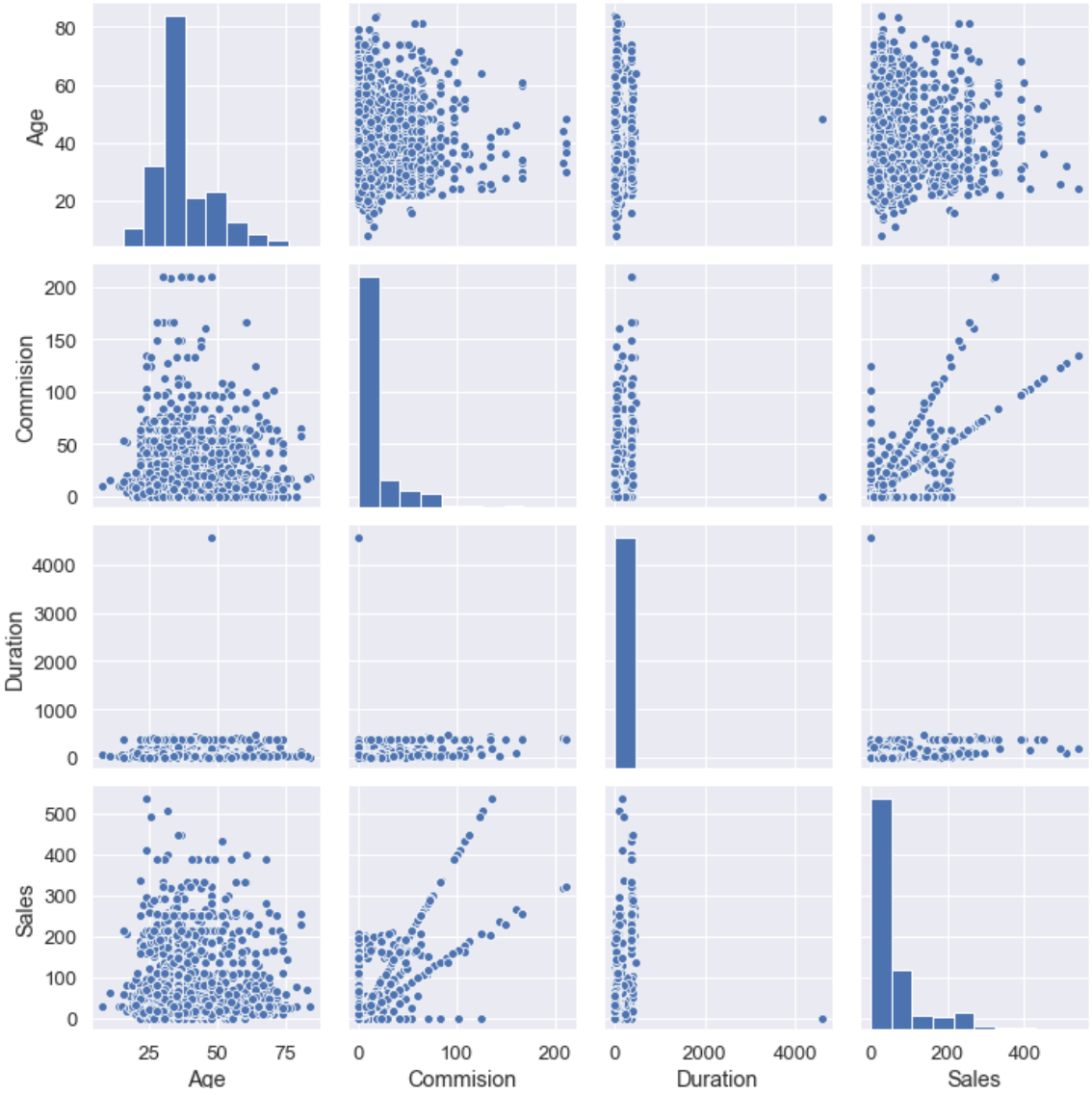

Figure 25: Pair Plot with Continuous Variable

Heat Map:

Figure 26: Heat Map

Observation:

- There are positive correlations between variables.Overall the magnitude of correlations between the variables are very less.
- There is no negative correlation

Converting Object data type into Categorical:

```
feature: Agency_Code
['C2B', 'EPX', 'CWT', 'JZI']
Categories (4, object): ['C2B', 'CWT', 'EPX', 'JZI']
[0 2 1 3]


feature: Type
['Airlines', 'Travel Agency']
Categories (2, object): ['Airlines', 'Travel Agency']
[0 1]


feature: Claimed
['No', 'Yes']
Categories (2, object): ['No', 'Yes']
[0 1]


feature: Channel
['Online', 'Offline']
Categories (2, object): ['Offline', 'Online']
[1 0]


feature: Product Name
['Customised Plan', 'Cancellation Plan', 'Bronze Plan', 'Silver Plan', 'Gold Plan']
Categories (5, object): ['Bronze Plan', 'Cancellation Plan', 'Customised Plan', 'Gold Plan', 'Silver Plan']
[2 1 0 4 3]
```

	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 12: Categorical Variable

Proportion of observations in Target classes:

- There is no issue of class imbalance here as we have reasonable proportions in both the classes

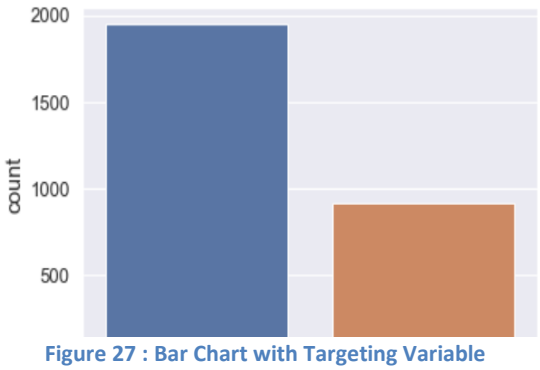

Figure 27 : Bar Chart with Targeting Variable

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

- Extracting the target column into separate vectors for training set and test set

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

Table 13: Extracting Target Column

Checking the dimensions of the training and test data:

```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
```

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

Cart conclution:	**Random Forest conclution:**	**Neural Network conclution:**
Train Data:	***Train Data:***	***Train Data:***
Accuracy: 79%	Accuracy: 81%	Accuracy: 78%
AUC: 84%	AUC: 86%	AUC: 82%
Recall:50%	Recall:58%	Recall:51%
Precision: 72%	Precision: 72%	Precision: 67%
f1-Score: 59%	f1-Score: 64%	f1-Score: 57%
Test Data:	***Test Data:***	***Test Data:***
Accuracy: 75%	Accuracy: 77%	Accuracy: 60%
AUC: 79%	AUC: 82%	AUC: 63%
Recall:38%	Recall:47%	Recall:63%
Precision: 73%	Precision: 73%	Precision: 42%
f1-Score: 50%	f1-Score: 57%	f1-Score: 51%

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

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.79	0.75	0.81	0.77	0.78	0.60
AUC	0.84	0.79	0.86	0.82	0.82	0.63
Recall	0.50	0.38	0.58	0.47	0.51	0.63
Precision	0.72	0.73	0.72	0.73	0.67	0.42
F1 Score	0.59	0.50	0.64	0.57	0.57	0.51

Table 14: Performance Matrices of all Models

Observation:

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

ROC Curve for the 3 models on the Training data:

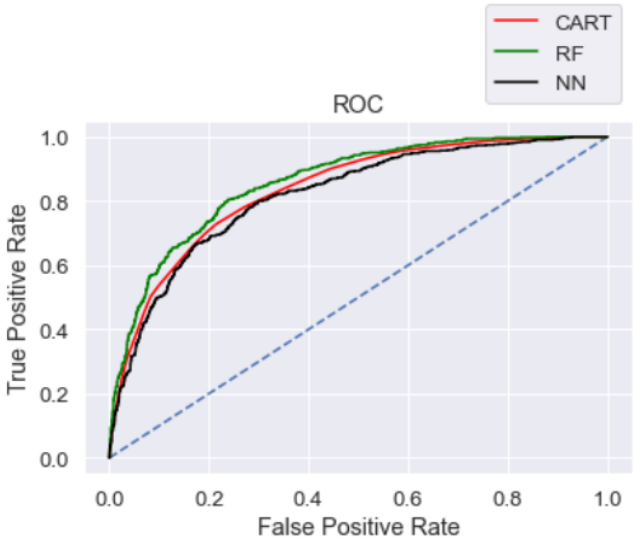

Figure 28: ROC Curve for 3 Models in Training Data

ROC Curve for the 3 models on the Test data:

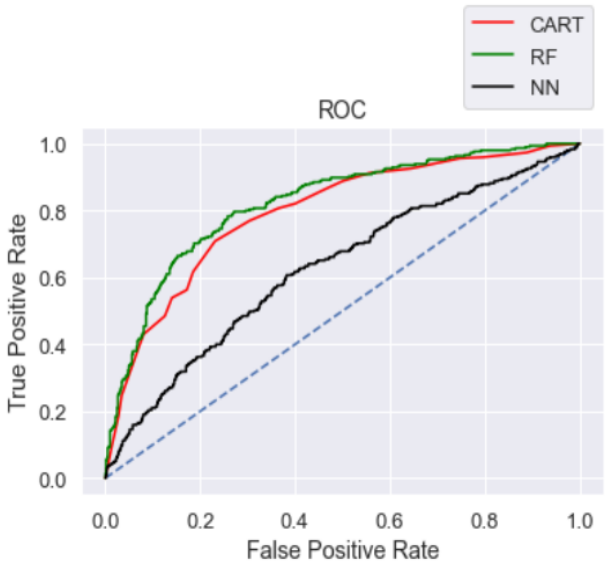

Figure 29 ROC Curve for 3 Models in Test Data

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

- Looking at the data, we need more data which will help us for better understand and predict the model in better way.
- JZI agency has low frequency of customers when compared to other agencies. So they should concentrate more on sales by giving advertisement and promotional campaign to increase the revenue.
- 98% of customer chooses online platform than offline, So online experience is better for customers we need to streamline in proper way which will be hassle free for customers. They can interact more customers in online, which will subsequently increase the revenue. Interesting fact is they providing insurance claim on offline business also...
- Travel Agency has high frequency of customers compared to airlines. Interesting fact is customer preferred airlines claiming rate is higher than customer preferred travel agency. So we need to evaluate why travel agency has low insurance claiming?..Need to overcome this situation to increase the revenue through travel agency.
- Customized Plan, Bronze Plan, Cancellation Plan has high customer frequencies but low claiming rate. so we need to work on marketing strategies to increase the sale in insurance claims .
- Silver Plan, Gold Plan has low customer frequencies but high claiming rate. we need to interact more customer in both plans so we can get more insurance claims so it will increase the revenue part.
- Interesting fact is none of the customer travel to America and Europe has not preferred C2B Agency but C2B agency are good with claiming rate. So need to focus on customers who travel to America and Europe to increase the revenue.

Key performance indicators (KPI) of insurance claims are

- Increase customer satisfaction which in fact will give more revenue.
- Regular feedback from customers and need to evaluate within tat.
- Need to reduce fraud transactions, deploy measures to avoid fraudulent transactions at earliest.
- Optimize claims recovery method.
- Reduce claim handling costs for customers.