Data Mining- Assignment

Problem 1: Clustering

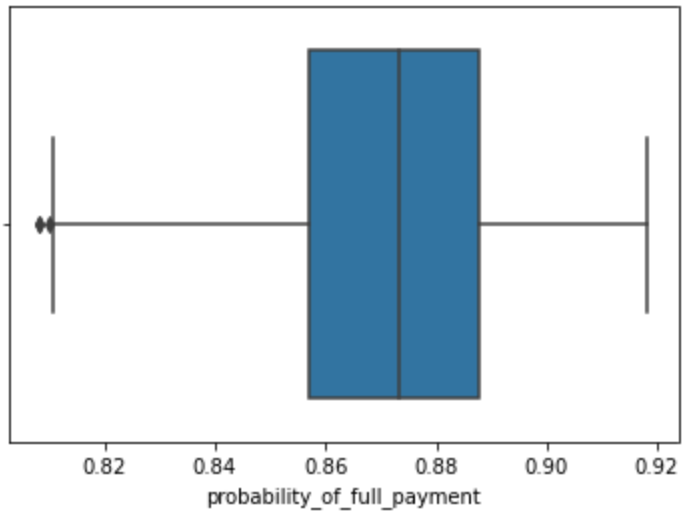
A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

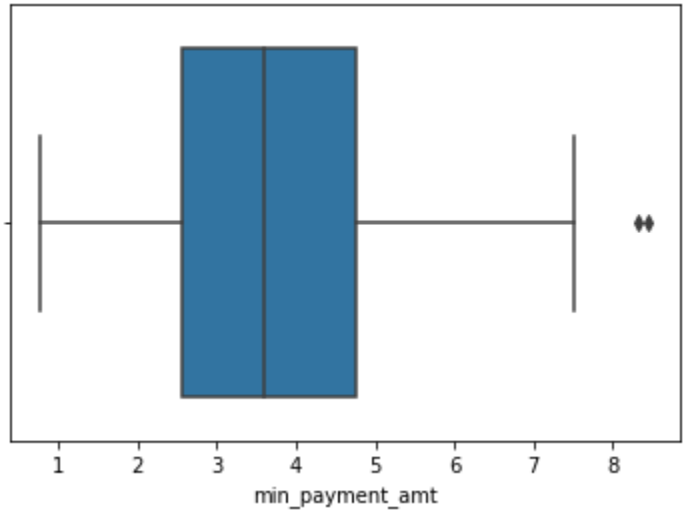
1.1 Read the data and do exploratory data analysis (3 pts). Describe the data briefly. Interpret the inferences for each (3 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

Solution:

The following are some observations after initial exploration of the data: (Details in Python file)

- The dataset consists of 210 rows and 7 columns.
- All the 7 attributes are numeric attributes and float data type.
- There are no missing values in the data.
- There are no duplicate records in the data set.
- Also, there are no bad data which is seen from the output of the 'info' command.
- There are a few outliers seen in the attributes, probability_of_full_payment and min_payment_amt as evident from the boxplots seen below (No outlier treatment is being done as per the instructions-FAQ in the question).
- Also, since probability ranges from 0 to 1, the outliers seen in the attribute, probability_of_full_payment should not be treated as they have been captured as per the real life scenario.
- There are no anomalies in the data as seen from the description of the data.





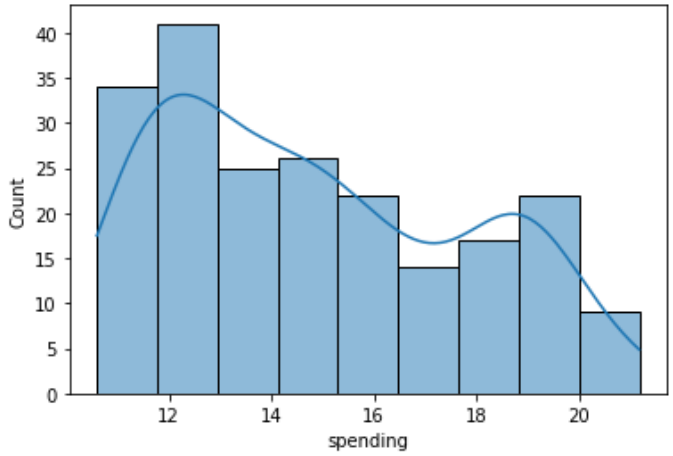
Data Visualization- Univariate Analysis

Insights from the Data

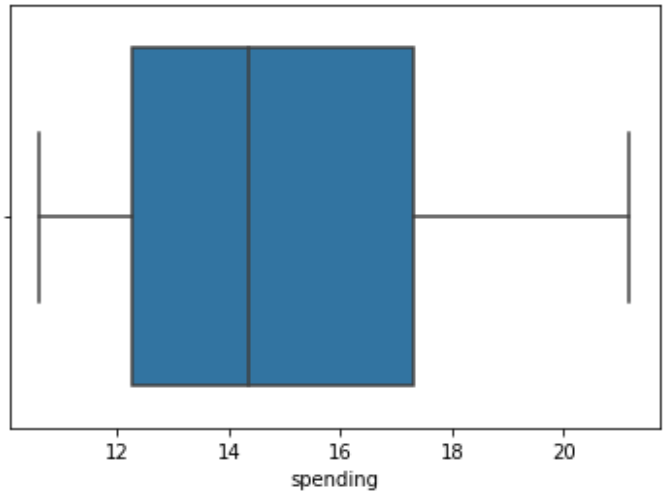
spending: Amount spent by the customer per month (in 1000s)

- The amount spent by the customer per month ranges from 10.59 to 21.18 (in 1000s)
- The mean amount spent is 14.85 and the median is 14.36 (in 1000s).
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the absence of outliers in the data.
- The Shapiro test indicates that the distribution is not normal and the skewness value = 0.40 indicating a right tailed distribution.

Distribution of spending



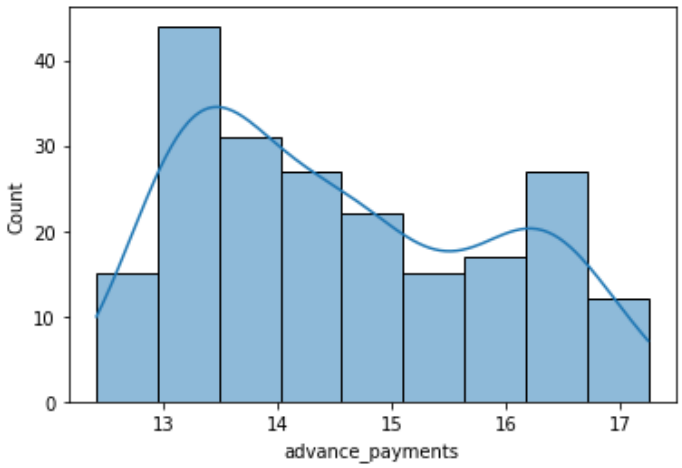
BoxPlot of spending



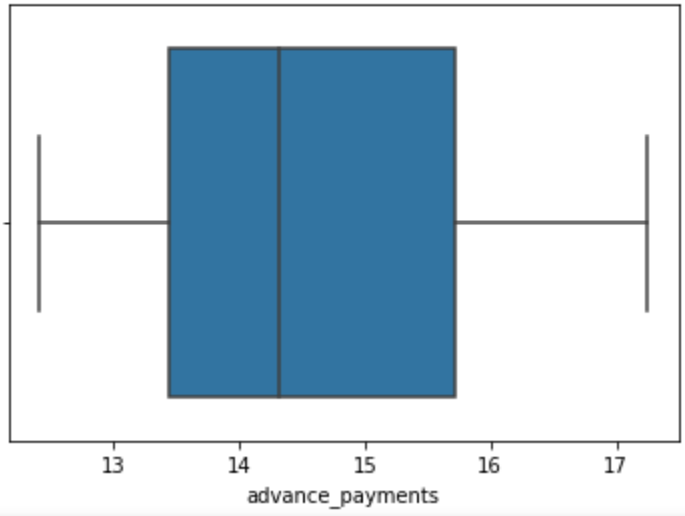
advance_payments: Amount paid by the customer in advance by cash (in 100s)

- The amount paid by the customer in advance by cash ranges from 12.41 to 17.25 (in 100s).
- The mean amount spent is 14.56 and the median is 14.32 (in 100s).
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the absence of outliers in the data.
- The Shapiro test indicates that the distribution is not normal and the skewness value = 0.38 indicating a right tailed distribution.

Distribution of advance_payments



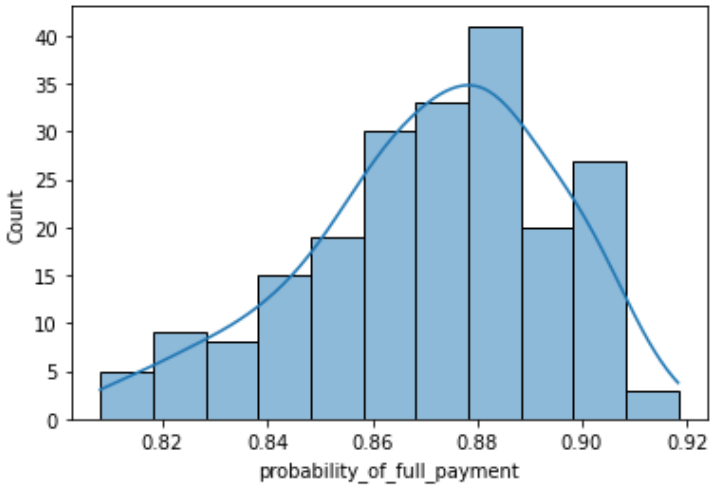
BoxPlot of advance_payments



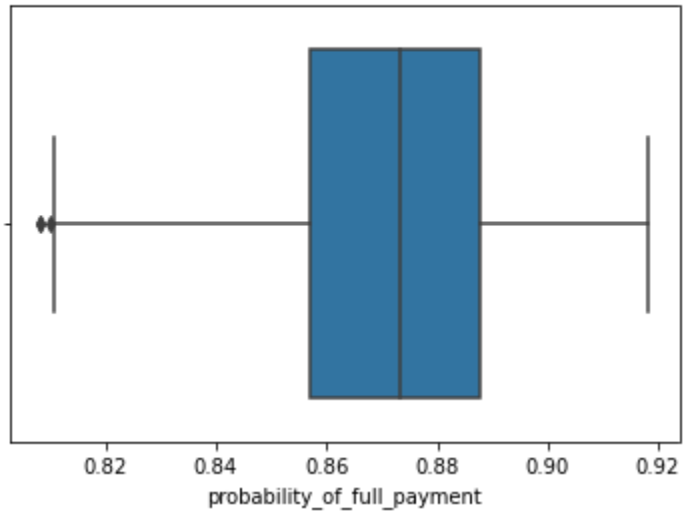
probability_of_full_payment: Probability of payment done in full by the customer to the bank

- The probability of payment done in full by the customer to the bank ranges from 0.81 to 0.92.
- The mean probability of payment done in full is 0.871 and the median is 0.873
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the presence of outliers in the data.
- The Shapiro test indicates that the distribution is not normal and the skewness value = -0.53 indicating a left tailed distribution.

Distribution of probability_of_full_payment



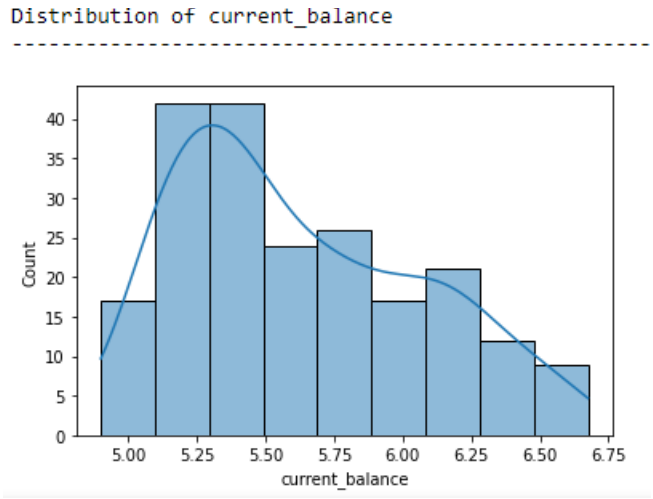
BoxPlot of probability_of_full_payment

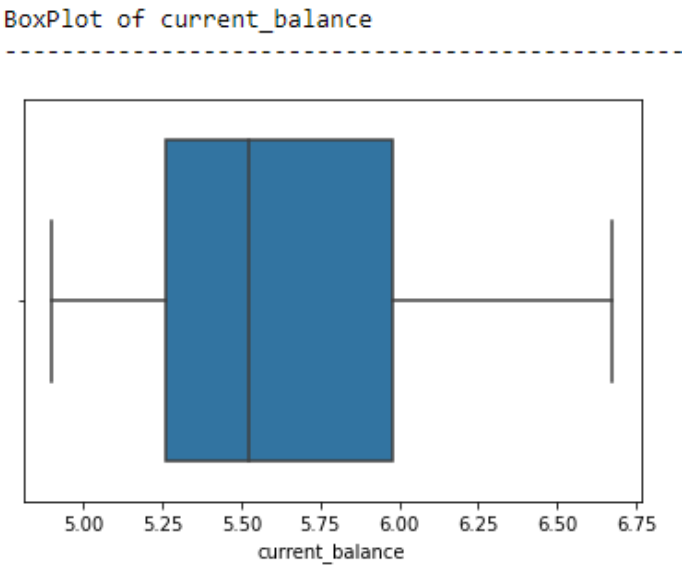


current_balance: Balance amount left in the account to make purchases (in 1000s)

- The balance amount left in the account to make purchases ranges from 4.90 to 6.68 (in 1000s).
- The mean balance amount left in the account to make purchases is 5.63 and the median is 5.52 (in 1000s)
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the absence of outliers in the data.

- The Shapiro test indicates that the distribution is not normal and the skewness value = 0.53 indicating a right tailed distribution.

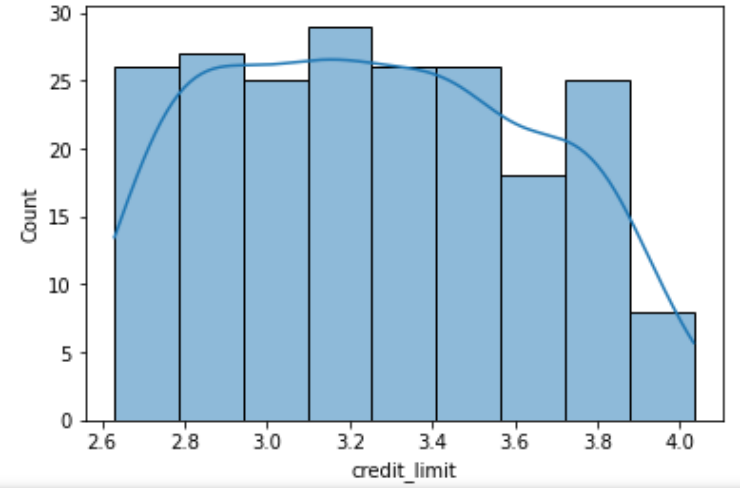




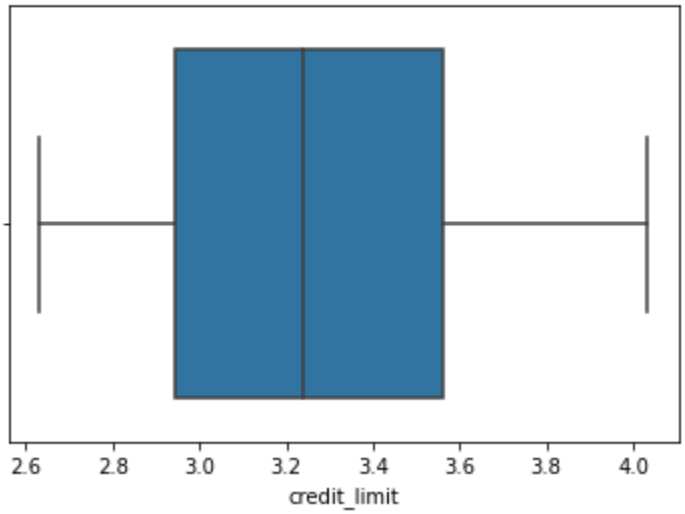
credit_limit: Limit of the amount in credit card (10000s)

- The credit limit ranges from 2.63 to 4.03 (in 10000s).
- The mean credit limit is 3.26 and the median is 3.24 (in 10000s)
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the absence of outliers in the data.
- The Shapiro test indicates that the distribution is not normal and the skewness value = 0.13 indicating a right tailed distribution.

Distribution of credit_limit



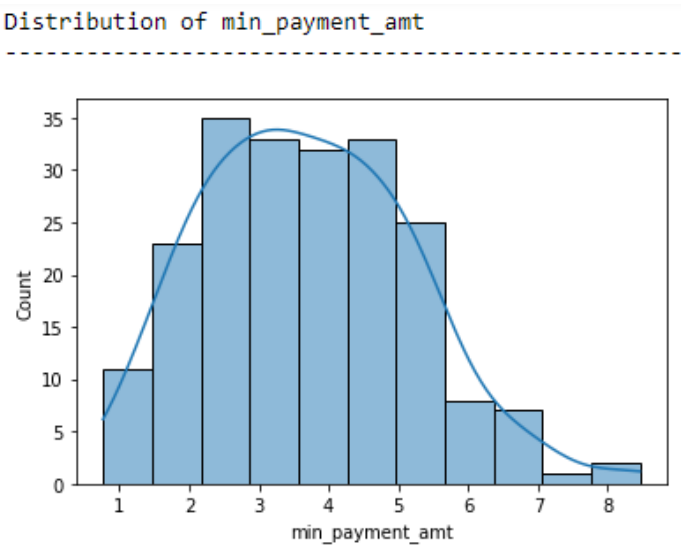
BoxPlot of credit_limit

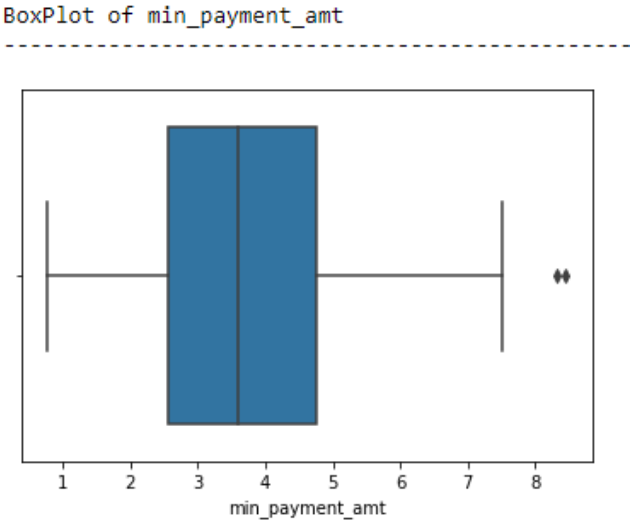


min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)

- The minimum paid by the customer while making payments for purchases mad monthly ranges from 0.77 to 8.46 (in 100s).
- The mean minimum payment amount is 3.7 and the median is 3.6 (in 100s)
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the presence of outliers in the data.

- The Shapiro test indicates that the distribution is not normal and the skewness value = 0.40 indicating a right tailed distribution.

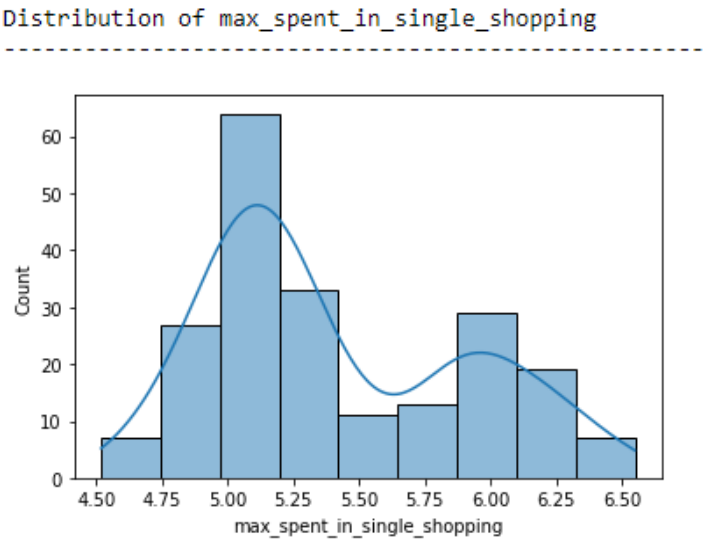


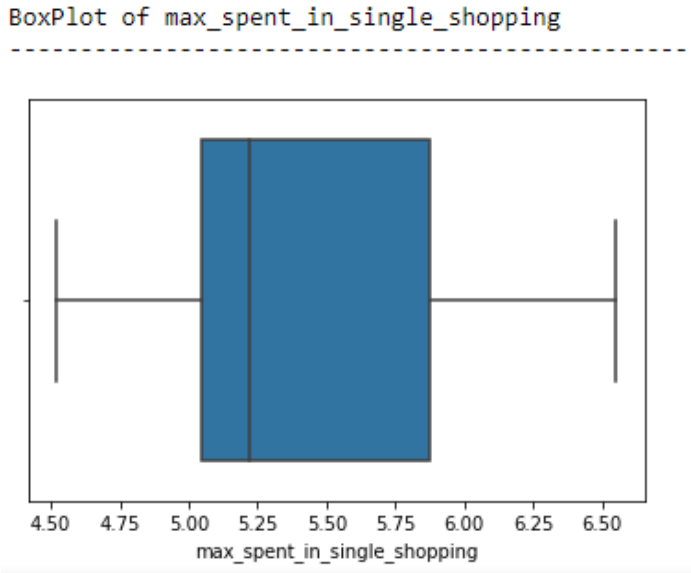


max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s)

- The maximum amount spent in one purchase ranges from 4.52 to 6.55 (in 1000s).
- The mean maximum amount spent in one purchase is 5.41 and the median is 5.22 (in 1000s)
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the absence of outliers in the data.

- The Shapiro test indicates that the distribution is not normal and the skewness value = 0.56 indicating a right tailed distribution.



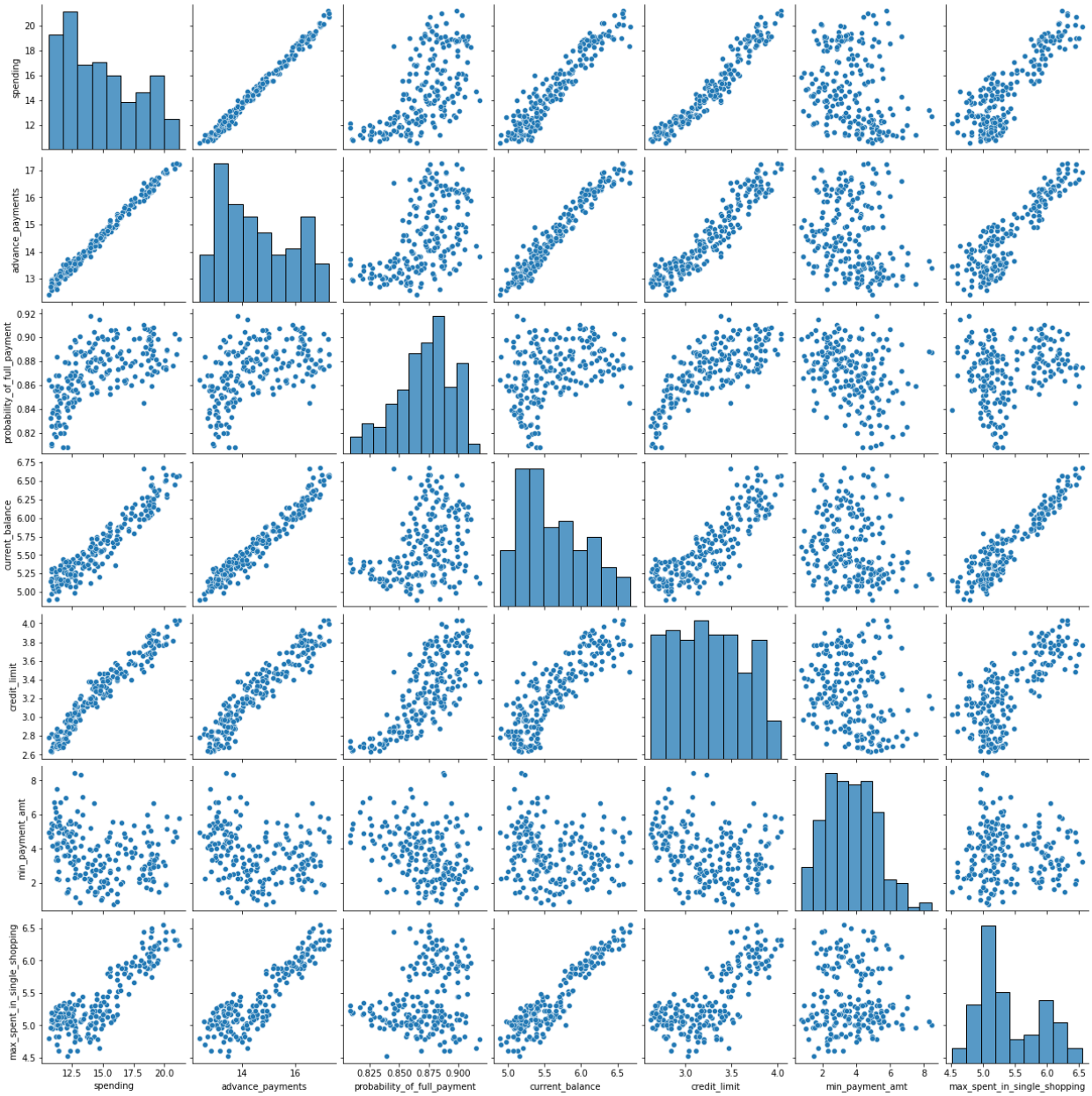


.**Multivariate Analysis- Bivariate Analysis**

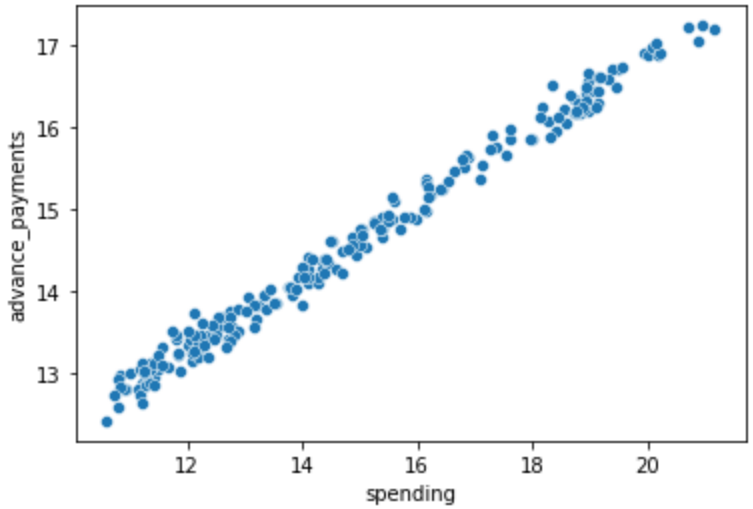
The below heat map gives the correlation between the 7 variables. The heat map indicates very high positive correlation between certain variables.


spending 1.00 0.99 0.61 0.95 0.97 -0.23 0.86
advance_payments 0.99 1.00 0.53 0.97 0.94 -0.22 0.89
probability_of_full_payment 0.61 0.53 1.00 0.37 0.76 -0.33 0.23
current_balance 0.95 0.97 0.37 1.00 0.86 -0.17 0.93
credit_limit 0.97 0.94 0.76 0.86 1.00 -0.26 0.75
min_payment_amt -0.23 -0.22 -0.33 -0.17 -0.26 1.00 -0.01
max_spent_in_single_shopping 0.86 0.89 0.23 0.93 0.75 -0.01 1.00
spending
advance_payments
probability_of_full_payment
current_balance
credit_limit
min_payment_amt
max_spent_in_single_shopping
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

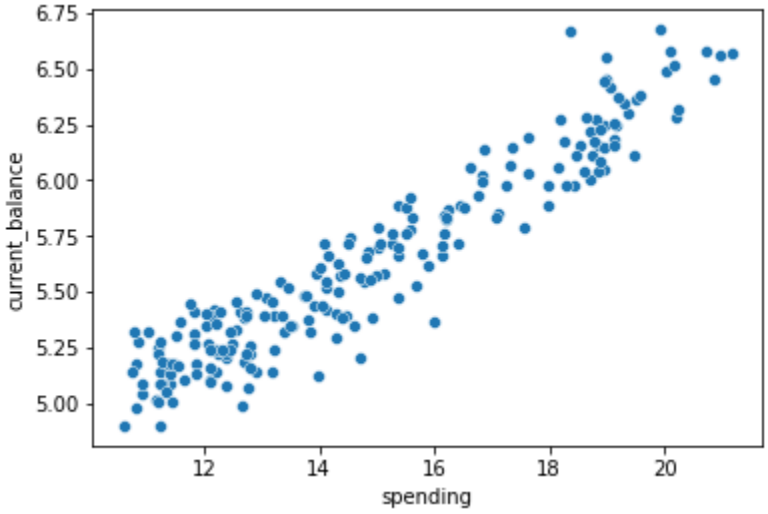
The below plot gives the pairwise scatterplot between the variables. The pairplot indicates a positive correlation between a number of variables.



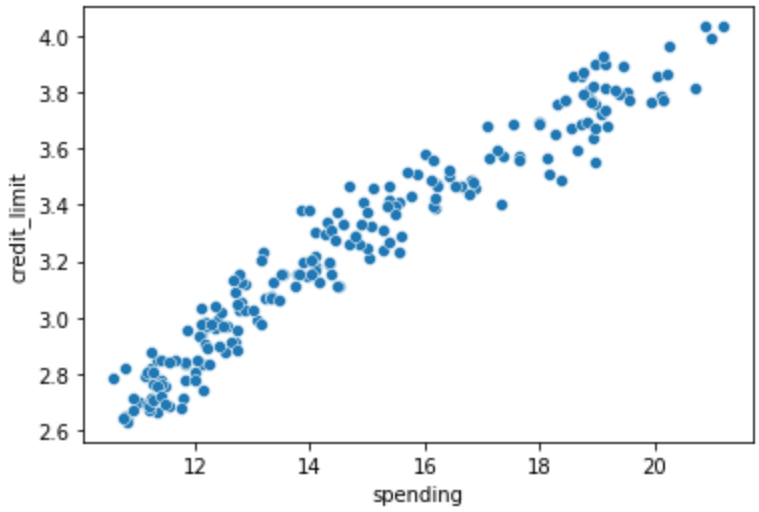
The plots of the variables with very high positive correlation rates (> 0.80) have been presented below**.**



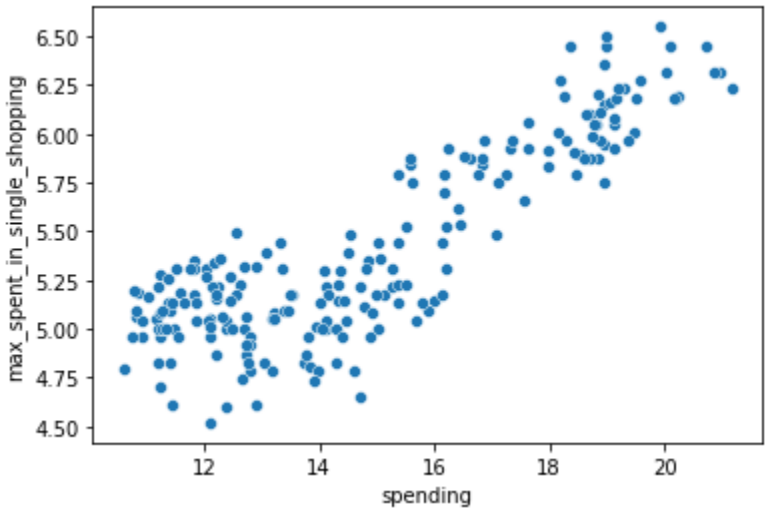
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.99) between spending and advance payments made by cash.



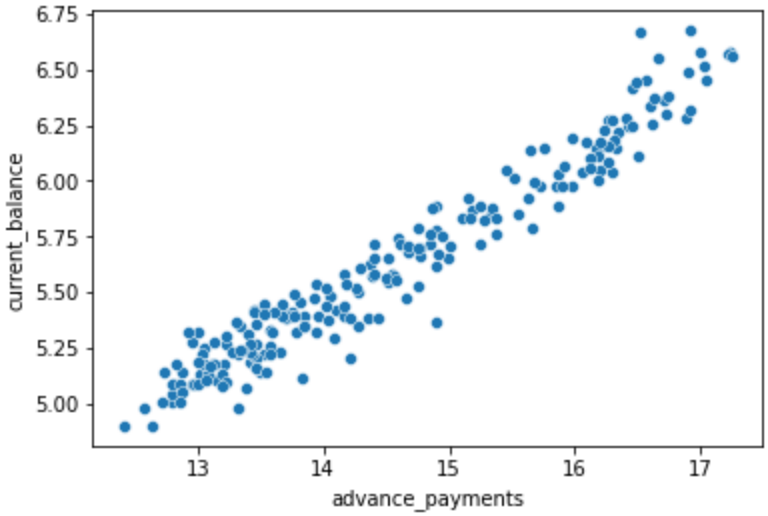
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.95) between spending and current balance.



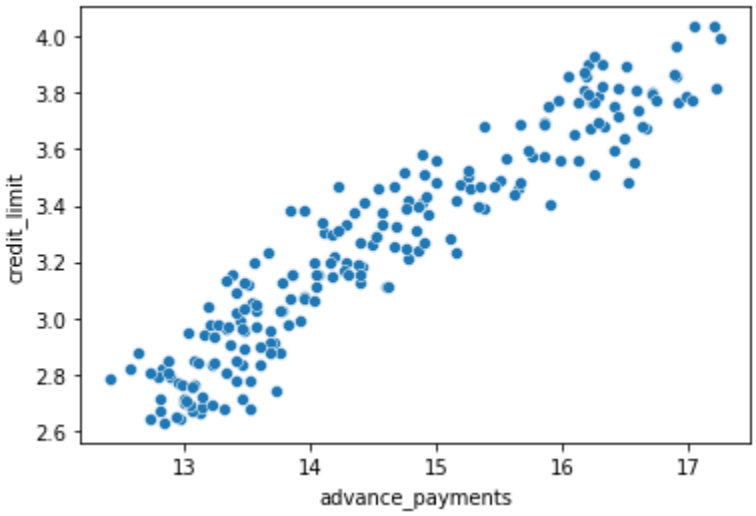
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.97) between spending and credit limit.



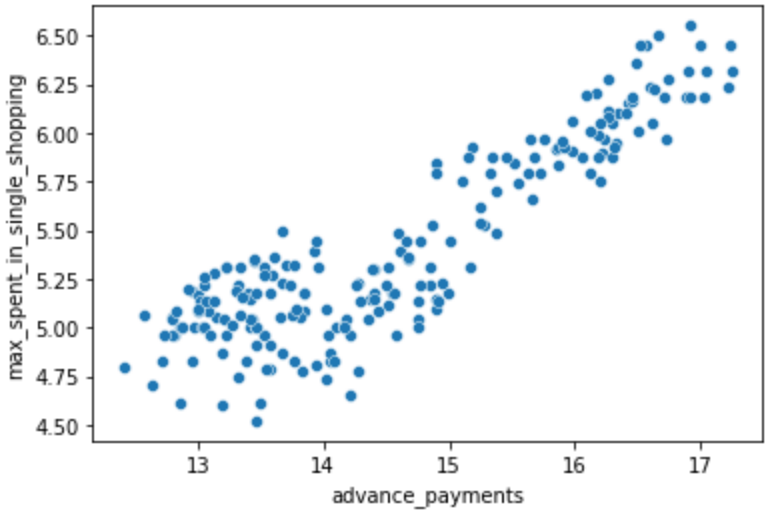
- From the heat map and the above scatter plot, we find a strong positive correlation (correlation coefficient = 0.86) between spending and maximum spent in single shopping.



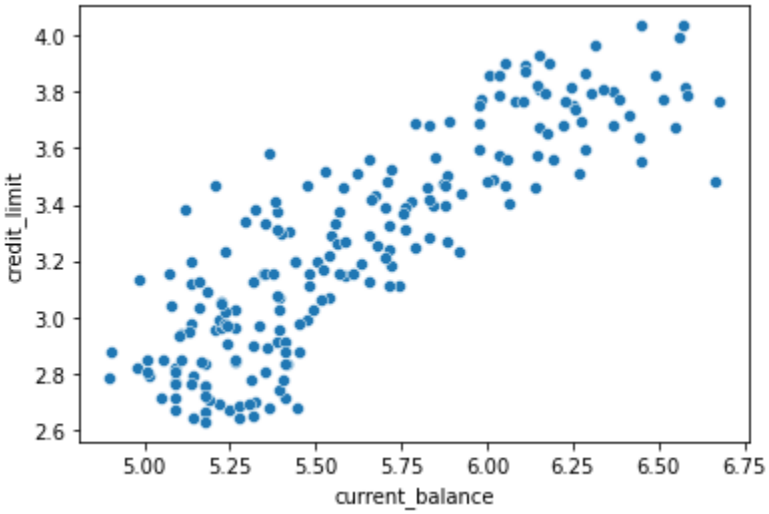
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.97) between advance payments and current balance.



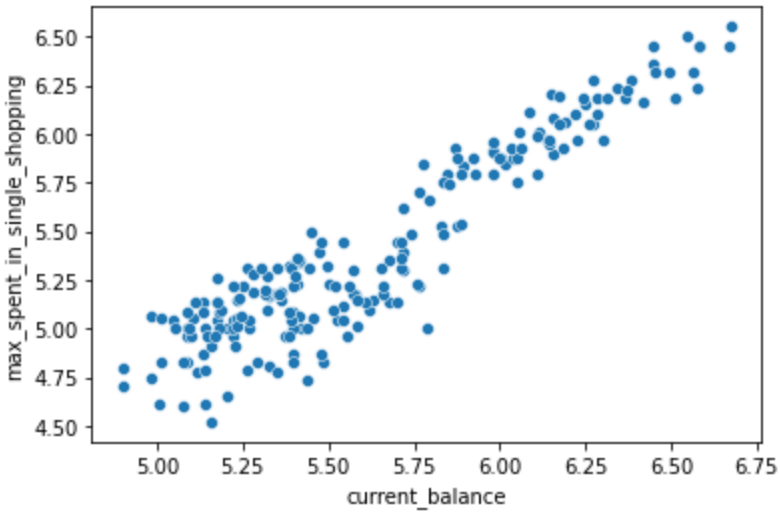
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.94) between advance payments and credit limit.



- From the heat map and the above scatter plot, we find a strong positive correlation (correlation coefficient = 0.89) between advance payments and maximum spending.



- From the heat map and the above scatter plot, we find a strong positive correlation (correlation coefficient = 0.86) between current balance and credit limit.



- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.93) between current balance and maximum spending.

The following are some observations inferred from the correlation values and the scatter plots.

- As the spending (amount spent by the customer per month) increases, the advance payments also increase which shows that high spending customers also seem to be paying advance payments in cash.
- As the spending (amount spent by the customer per month) increases, the current balance also increases. This tells us that high spending customers are people with high current balance.
- As the spending (amount spent by the customer per month) increases, the credit limit also increases. This shows that high spending customers have a high credit limit provided by the bank.
- Also, high spenders seem to be spending a large amount in a single shopping as seen from the spending versus max spent in single shopping plot.
- As advance payments increase, current balance also increases. This shows that people who pay a high advance payment amount also are customers with high current balance.
- As advance payments increase, credit limit also increases. This shows that people who pay a high advance payment amount also are customers with the advantage of a high credit limit.

- As advance payments increase, maximum amount spent in a single shopping also increases which shows that people who pay a high advance payment amount also are customers who spend a large amount in a single shopping.
- As current balance increases, credit limit also increase which shows that customers who maintain a high balance in their account are provided with a high credit limit by their bank.
- As current balance increases, maximum amount spent in a single shopping also increases which shows that customers who maintain a high balance in their account spend large amounts in a single shopping.

1.2 Do you think scaling is necessary for clustering in this case? Justify The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling and which method is he/she using to do the scaling. Can also comment on how that method works.

Solution:

Yes. Scaling is necessary for clustering in this case.

Scaling of the features is done to ensure that they are all almost the same magnitude so that each feature is equally important and which in turn helps the clustering algorithm. An important issue is that the range of the variables may differ by large magnitudes. If the original data is not scaled, this would result in putting more weights on the variables with a large range and thus resulting in poor model performance. Clustering algorithms use distance based metrics like Euclidean distance and hence, it becomes all the more significant to perform feature scaling.

In this dataset, it is seen that there is a significant difference in the range of the variables which can be seen from their standard deviation. The mean and standard deviation of the features are given below :

```
Mean of scaled data

spending                      14.847524
advance_payments              14.559286
probability_of_full_payment    0.870999
current_balance                5.628533
credit_limit                   3.258605
min_payment_amt                3.700201
max_spent_in_single_shopping   5.408071
dtype: float64
```

```
standard deviation of  data

spending                       2.909699
advance_payments               1.305959
probability_of_full_payment    0.023629
current_balance                0.443063
credit_limit                   0.377714
min_payment_amt                1.503557
max_spent_in_single_shopping   0.491480
dtype: float64
```

For example: the range of spending variable is between 10590 to 21180 whereas the range of advance payment is 1241 to 1725. When Euclidean distance is computed, the number $(21180-10590)^2$ is much larger than the number of $(1725-1241)^2$ which implies that the Euclidean distance metric will be dominated by the feature, 'spending' when compared to the variable, 'advance payment'.

We see that the variable, 'probability of full payment' ranges from 0.81 to 0.92 and hence while calculating the distance metric, the contribution of this metric will be even less significant. The current balance ranges from 4900 to 6680, the credit limit ranges from 26300 to 40300, the minimum paid amount ranges from 77 to 846, maximum amount spent in one purchase ranges from 4520 to 6550. It is clearly seen that the range of the features differ by large magnitudes which will result in poor clustering if scaling is not performed. Hence, scaling is a must for this dataset before we perform clustering.

There are two important scaling techniques- Standardization (Z-score Normalization) and Max-Min Normalization **(**Min-Max scaling)

The standardization method which is computed as x scaled = (x – mean)/ standard deviation is applied here which ensures that the attribute means are all 0 and standard deviation are all 1 which in turn assures that all the variables are all of the same magnitude as seen in the below tables. This method transforms the data into a normal distribution and works well for distance based algorithms like clustering.

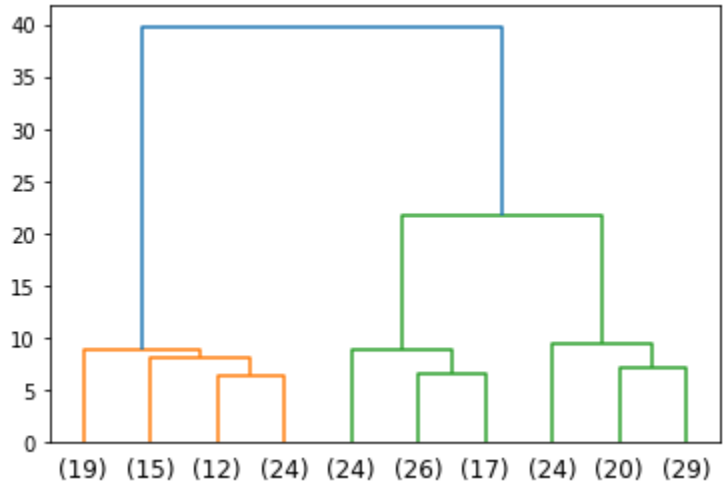
```
Mean of scaled data

spending                        9.148766e-16
advance_payments                1.097006e-16
probability_of_full_payment     1.260896e-15
current_balance                -1.358702e-16
credit_limit                   -2.790757e-16
min_payment_amt                 5.418946e-16
max_spent_in_single_shopping   -1.935489e-15
dtype: float64
```

```
standard deviation of scaled data

spending                          1.002389
advance_payments                  1.002389
probability_of_full_payment       1.002389
current_balance                   1.002389
credit_limit                      1.002389
min_payment_amt                   1.002389
max_spent_in_single_shopping      1.002389
dtype: float64
```

1.3 Apply hierarchical clustering to scaled data (3 pts). Identify the number of optimum clusters using Dendrogram and briefly describe them (4). Students are expected to apply hierarchical clustering. It can be obtained via Fclusters or Agglomerative Clustering. Report should talk about the used criterion, affinity and linkage. Report must contain a Dendrogram and a logical reason behind choosing the optimum number of clusters and Inferences on the dendrogram. Customer segmentation can be visualized using limited features or whole data but it should be clear, correct and logical. Use appropriate plots to visualize the clusters.

Solution:

Hierarchical clustering was applied to the scaled data using the ward linkage method. The following figure shows the dendrogram obtained from the clustering with the last 10 merges.

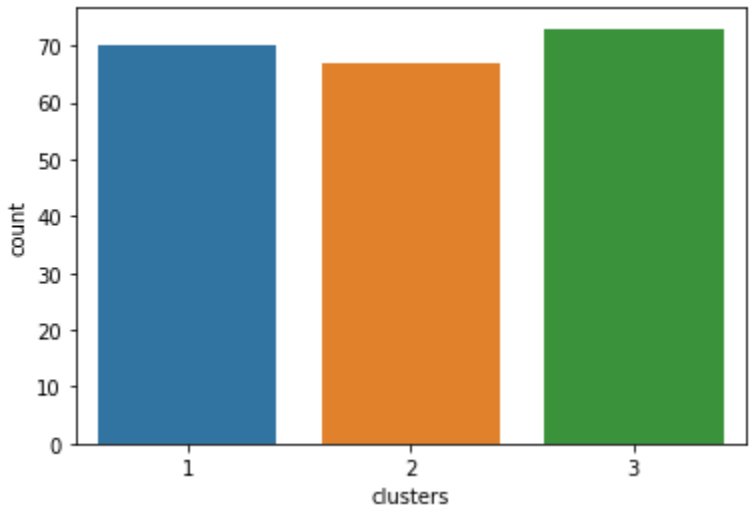


Although, the dendrogram shows two groups or clusters clearly, this will not help us with regard to the business problem and business impact. Fcluster is used with the distance criterion = 12 for forming the clusters. A parallel line to the x axis is drawn at distance 12, which intersects three vertical lines resulting in 3 clusters. In the last 10 merges, we are able to see 3 groups(2 green and 1 orange cluster)

The following shows the top 5 records of the datasets with the cluster details.

	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

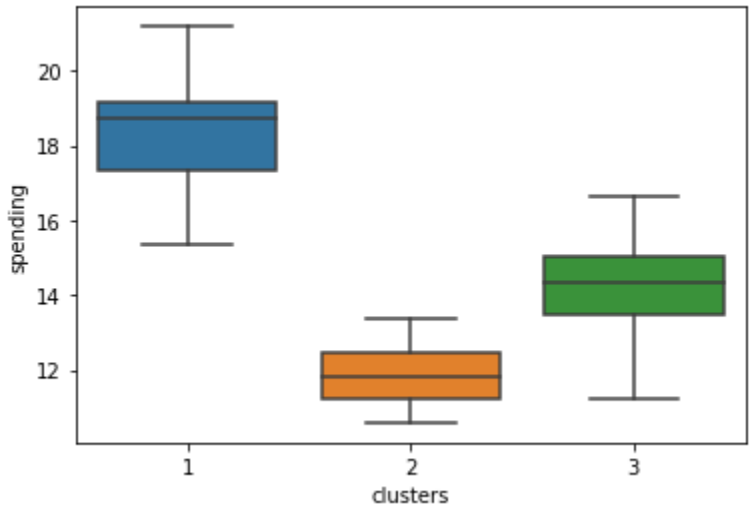
The following gives the count plot of the clusters.



There are 70 records belonging to cluster 1, 67 belonging to cluster 2 and 73 belonging to cluster 3.

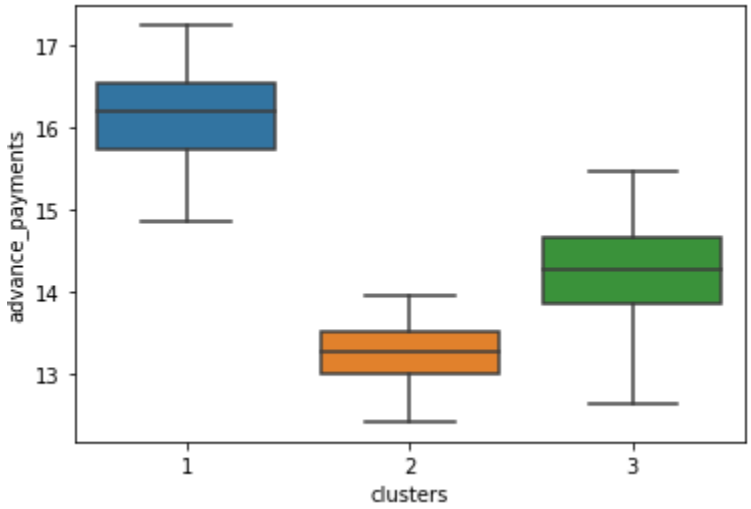
The following gives the average value of the variables and boxplots, cluster wise and shows the customer segmentation.

Spending- Amount spent by the customer per month (in 1000s)



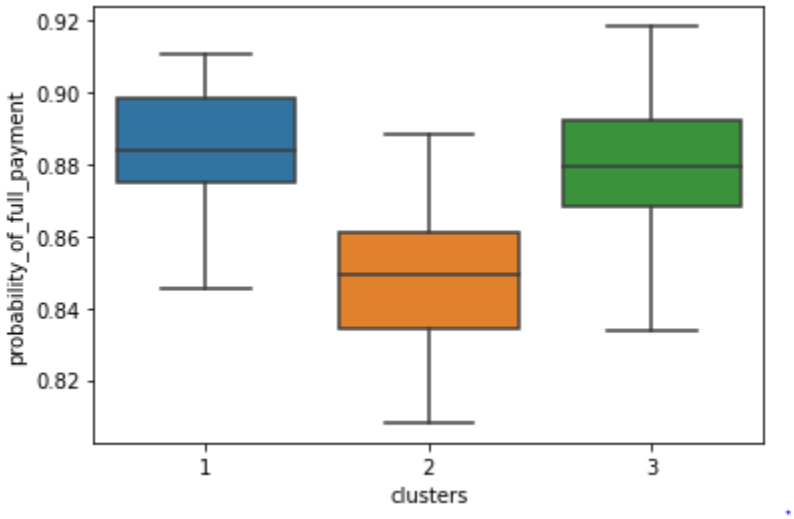

```
Average spending

clusters
1    18.371429
2    11.872388
3    14.199041
Name: spending, dtype: float64
```

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



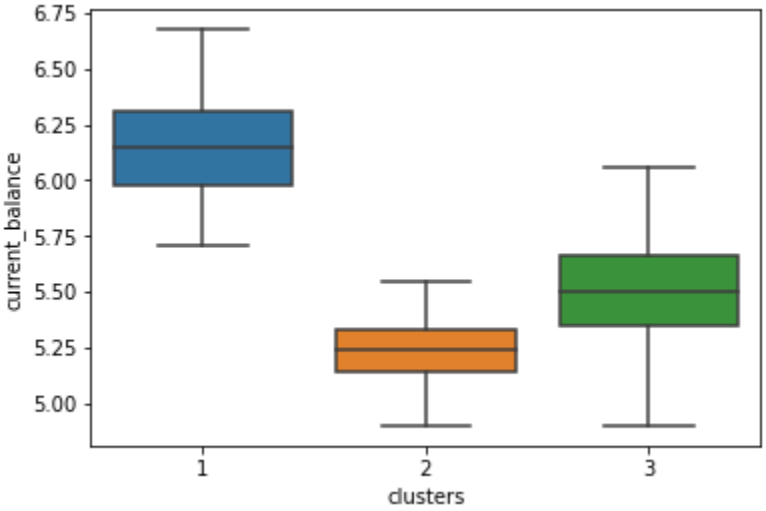
```
Average advance_Payments

clusters
1    16.145429
2    13.257015
3    14.233562
Name: advance_payments, dtype: float64
```

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



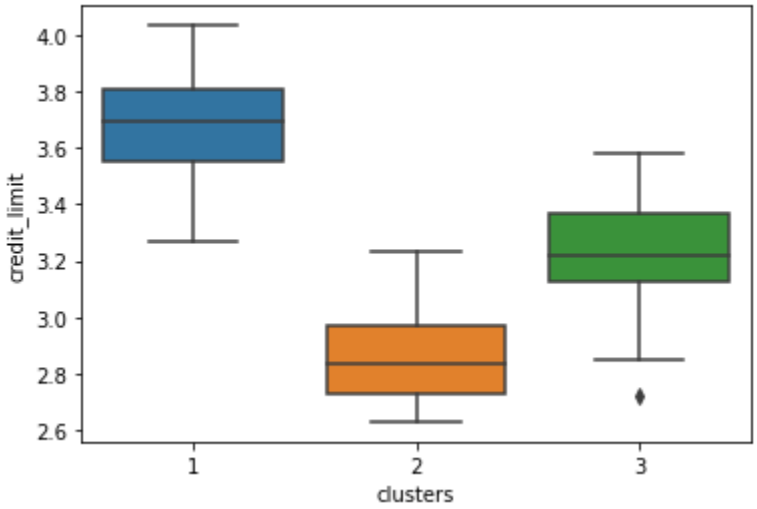

```
Average probability_of_full_payment

clusters
1    0.884400
2    0.848072
3    0.879190
Name: probability_of_full_payment, dtype: float64
```

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



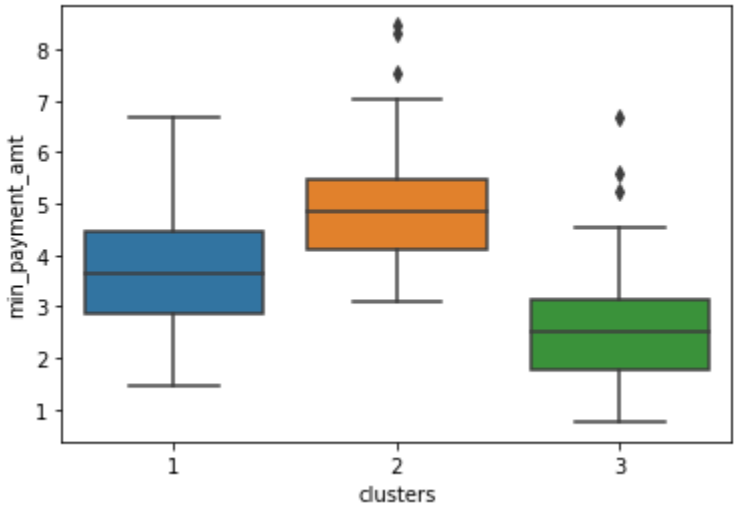
```
Average current_balance

clusters
1    6.158171
2    5.238940
3    5.478233
Name: current_balance, dtype: float64
```

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



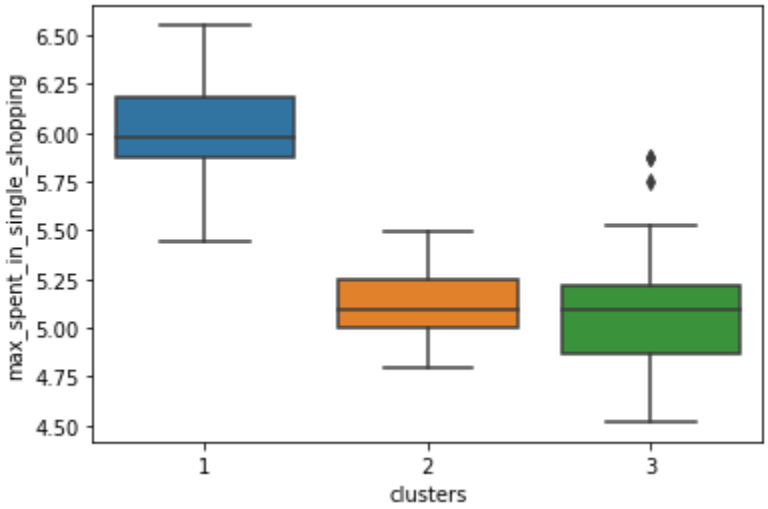

```
Average credit_limit

clusters
1    3.684629
2    2.848537
3    3.226452
Name: credit_limit, dtype: float64
```

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



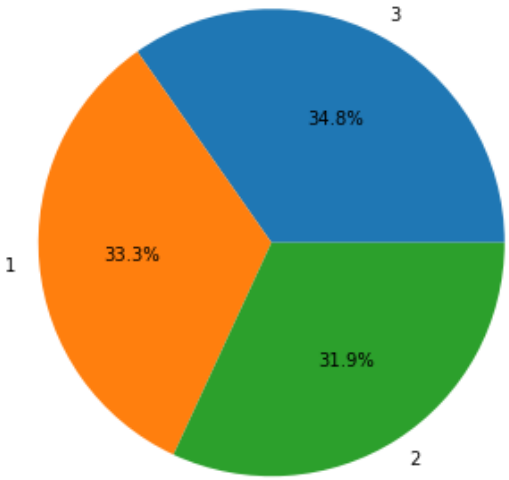
```
Average min_payment_amt

clusters
1    3.639157
2    4.949433
3    2.612181
Name: min_payment_amt, dtype: float64
```

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




```
Average max_spent_in_single_shopping

clusters
1    6.017371
2    5.122209
3    5.086178
Name: max_spent_in_single_shopping, dtype: float64
```

The following gives the pie chart of customer segmentation:



Following are a few observations on customer segmentation after analysis from the bivariate plots above:

- There are 70 customers belonging to cluster 1, 67 belonging to cluster 2 and 73 belonging to cluster 3.
- Cluster 1 consists of customers with highest spending, highest advance payments, highest current balance, highest credit limit, highest maximum spending in a single shopping and highest probability of full payments. With regard to minimum payment amount, they are second highest after cluster 2.
- Cluster 2 are customers with lowest spending, lowest advance payments, lowest current balance, lowest credit limit and lowest probability of full payments. They are the highest with respect to minimum payment amount and second highest with respect to maximum spent in a single shopping.
- Cluster 3 are customers who fall in between customers in cluster 1 and cluster 2 and are second highest with regard to the attributes spending, advance payments, probability of full payment, current balance, credit limit. They exhibit the lowest value with respect to minimum payment amount and maximum spent in single shopping.
- Looking at the observations, we can conclude that cluster 1 consists of wealthy/privileged customers, cluster 2 consists of not so wealthy/privileged customers and cluster 3 consists of customers who rank just below wealthy/privileged customers.
- For sake of better understanding, we could say cluster 1 are gold customers, cluster 3 are silver customers and cluster 2 are bronze customers.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters (2 pts). Apply elbow curve and silhouette score (3 pts). Interpret the inferences from the model (2.5 pts). K-means clustering code application with different number of clusters. Calculation of WSS(inertia for each value of k) Elbow Method must be applied and visualized with different values of K. Reasoning behind the selection of the optimal value of K must be explained properly. Silhouette Score must be calculated for the same values of K taken above and commented on. Report must contain logical and correct explanations for choosing the optimum clusters using both elbow method and silhouette scores. Append cluster labels obtained from K-means clustering into the original data frame. Customer Segmentation can be visualized using appropriate graphs.

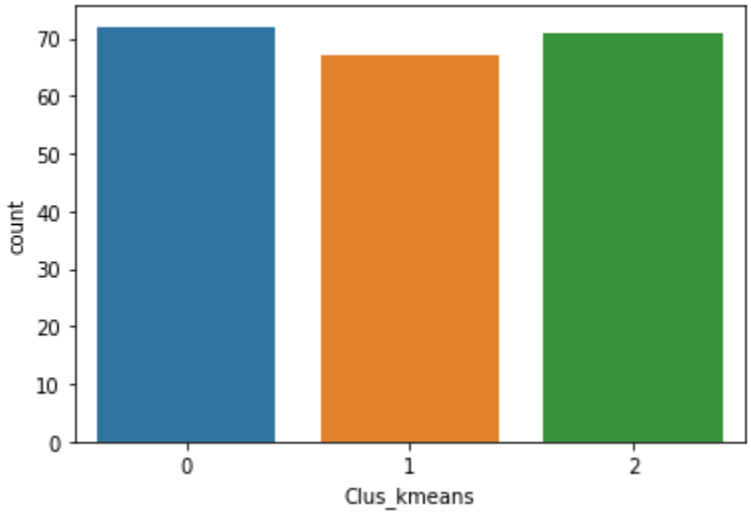
Solution:

K Means clustering was applied on scaled data and 3 clusters were found as the optimum number. The following gives the top 5 records with the kmeans cluster details which is given by the attribute Clus_kmeans.

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	Clus_kmeans
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	1
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0
17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1	1

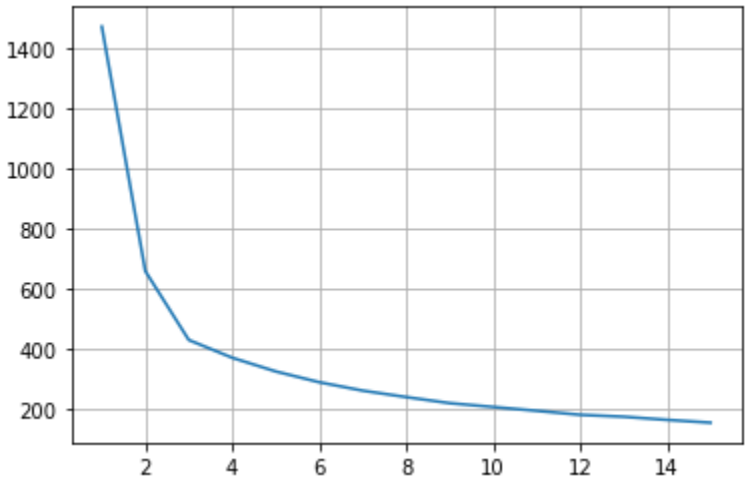
The clusters are cluster 0, cluster 1 and cluster 2.

The following gives the count plot of the number of clusters.



There are 71 records belonging to cluster 0, 72 belonging to cluster 1 and 67 belonging to cluster 2.

The following shows the elbow/WSS plot:



Elbow/WSS Plot

The within sum of squares (WSS-inertia) for clusters 2 to 15 are given below:

```
[1469.9999999999998,
 659.171754487041,
 430.6589731513006,
 371.65314399951626,
 326.311446829373,
 290.590030596822,
 262.64761255781275,
 241.26972395257218,
 221.16302211267288,
 208.47277143653196,
 195.77787225519222,
 182.52659249821758,
 176.14824535563054,
 165.76268242099326,
 156.43231840406148]
```

- The Kmeans algorithm was applied for number of clusters = 2 and number of clusters = 3.
- Quoting the definition, the silhouette coefficient is a measure of how well samples are clustered with samples that are similar to each other. Clustering models with high silhouette coefficient are said to be dense, where samples in the same cluster are similar to each other and well separated, and where samples in different clusters are dissimilar to each other.
- The silhouette score for no. of clusters = 2 was found to be 0.4658 and the silhouette score for no. of clusters = 3 was found to be 0.40.
- As per the WSS plot, the largest drop was found when no. of clusters = 2 (Inertia value drop from 1469.99 to 659.17) but the elbow joint as seen from the graph is when the number of clusters = 3.
- When the silhouette coefficient for each sample was computed for no. of clusters = 2, one negative value was found (-0.00617).
- Since, negative values are not acceptable, the silhouette coefficient for each sample for no. of clusters = 3 was computed and the minimum value was found to be 0.0027, a positive value.
- Hence, taking into account all the above factors, the optimum number of clusters was taken as 3.

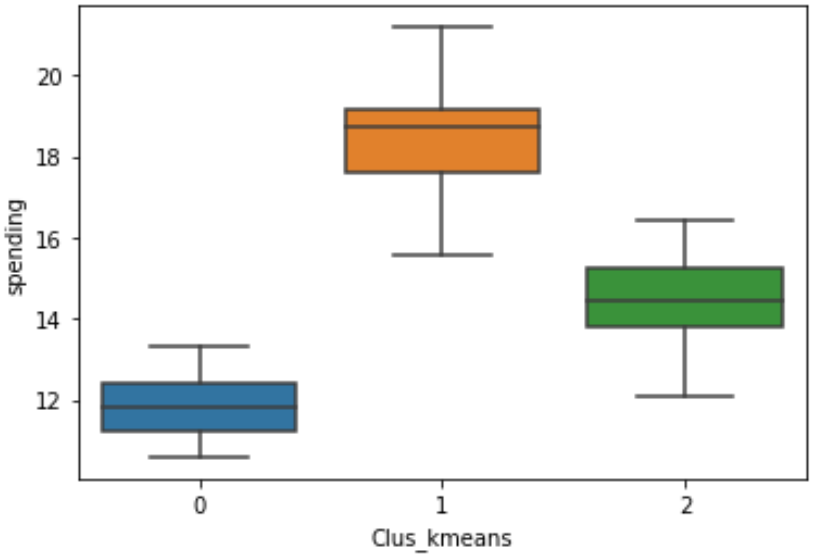
The following shows the cluster plot for 2 clusters.

The following gives the average value of the variables and boxplots, cluster wise and shows the customer segmentation.

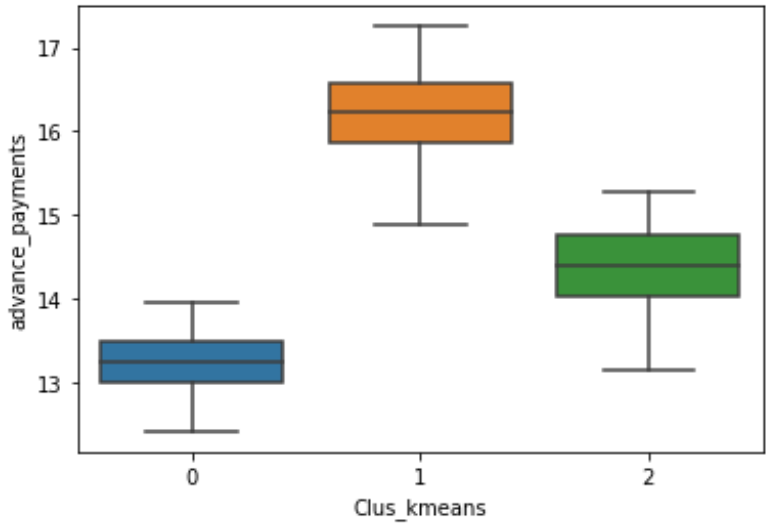
```
Average of spending
-----------------------------------------------------
Clus_kmeans
0    11.856944
1    18.495373
2    14.437887
Name: spending, dtype: float64

Average of advance_payments
-----------------------------------------------------
Clus_kmeans
0    13.247778
1    16.203433
2    14.337746
Name: advance_payments, dtype: float64

Average of probability_of_full_payment
-----------------------------------------------------
Clus_kmeans
0    0.848253
1    0.884210
2    0.881597
Name: probability_of_full_payment, dtype: float64

  Average of current_balance
  ----------------------------------------------
  Clus_kmeans
  0    5.231750
  1    6.175687
  2    5.514577
  Name: current_balance, dtype: float64

  Average of credit_limit
  ----------------------------------------------
  Clus_kmeans
  0    2.849542
  1    3.697537
  2    3.259225
  Name: credit_limit, dtype: float64

  Average of min_payment_amt
  ----------------------------------------------
  Clus_kmeans
  0    4.742389
  1    3.632373
  2    2.707341
  Name: min_payment_amt, dtype: float64
```

```
Average of max_spent_in_single_shopping
-------------------------------------------------
Clus_kmeans
0    5.101722
1    6.041701
2    5.120803
Name: max_spent_in_single_shopping, dtype: float64
```

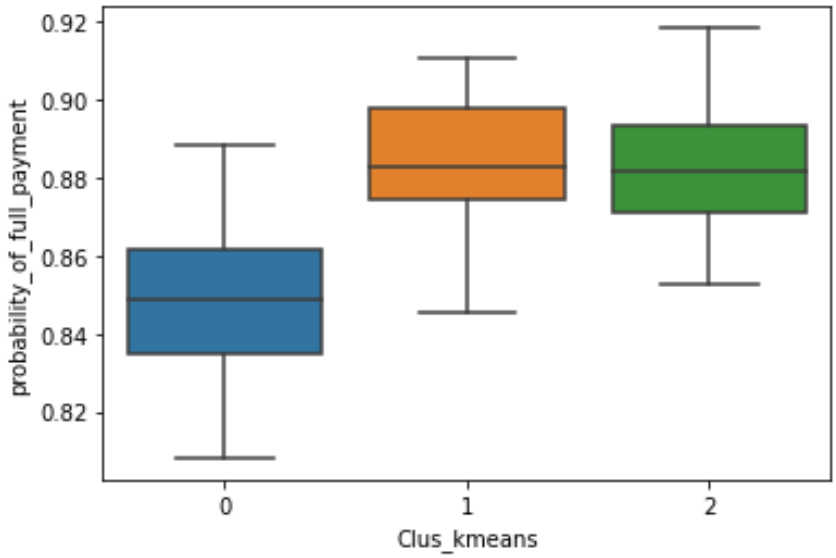
Spending- Amount spent by the customer per month (in 1000s)



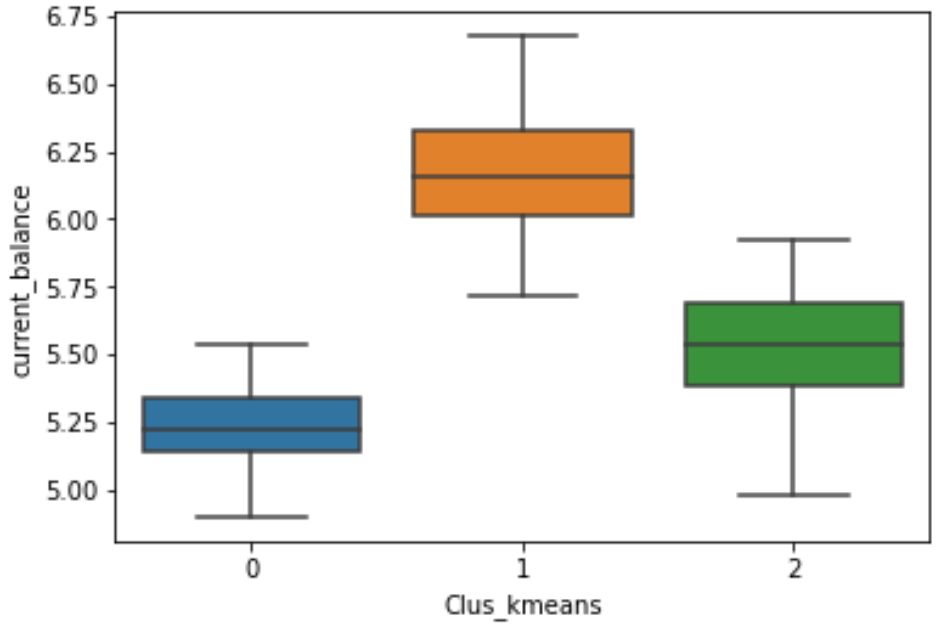
advance_payments: Amount paid by the customer in advance by cash (in 100s)



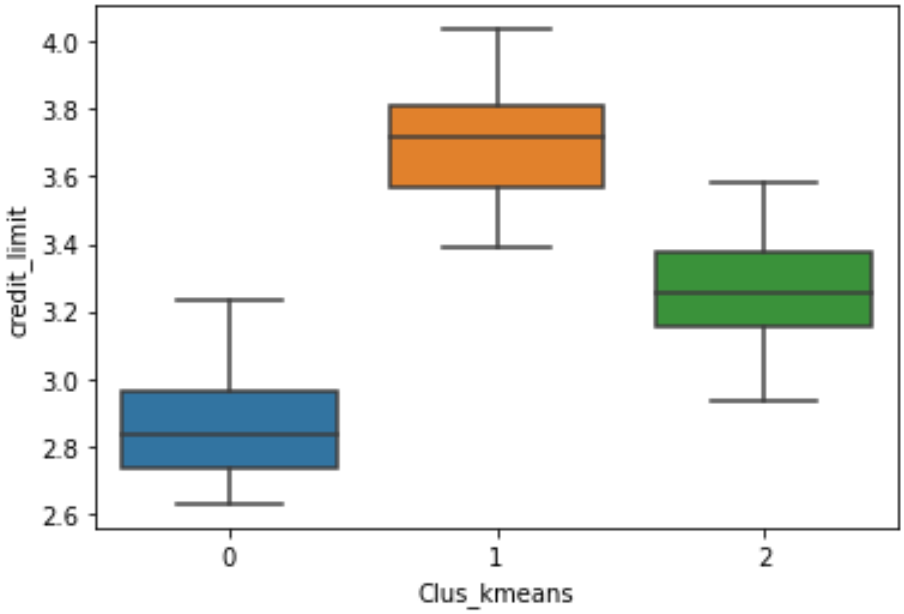
probability_of_full_payment: Probability of payment done in full by the customer to the bank



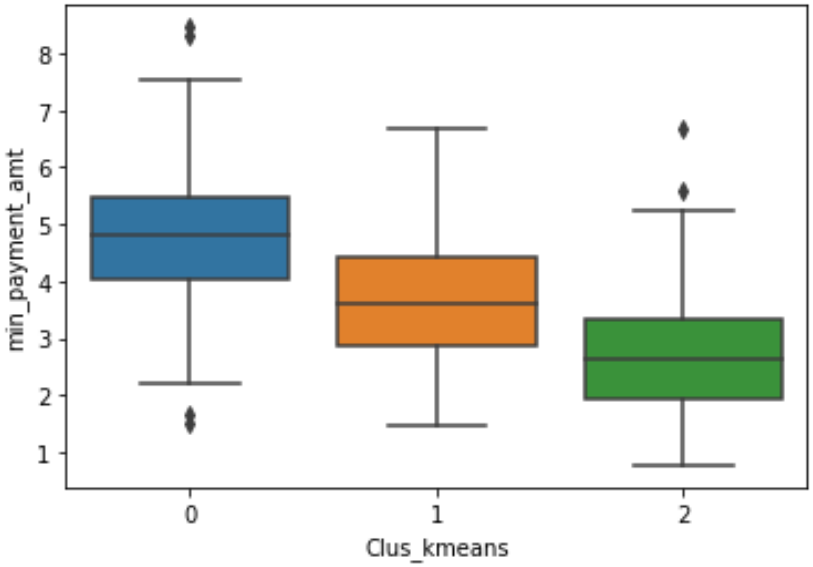
current_balance: Balance amount left in the account to make purchases (in 1000s)



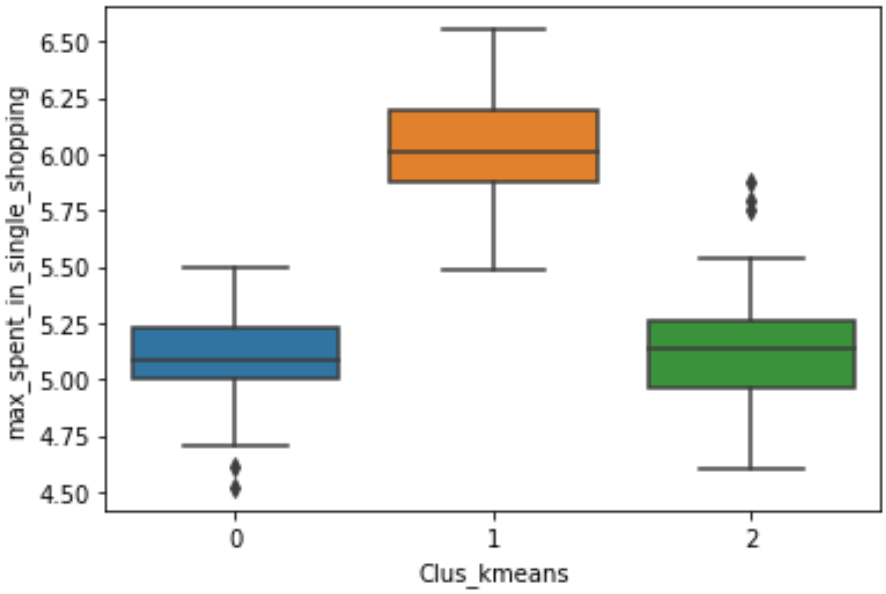
credit_limit: Limit of the amount in credit card (10000s)



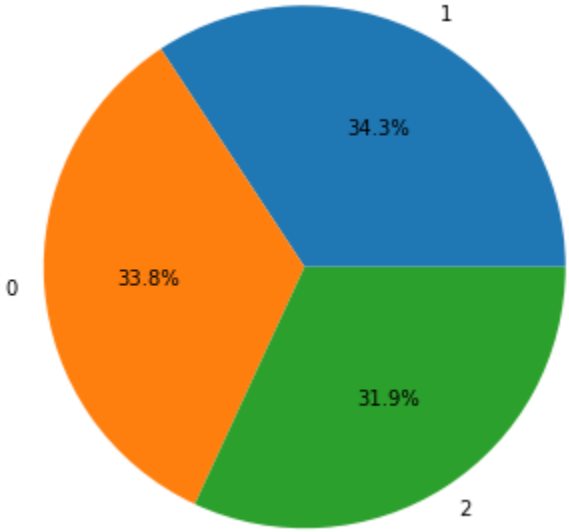
min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)



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



The following pie chart shows the customer segmentation for 3 clusters :



The following gives the clustering plot for 3 clusters.

Following are a few observations on customer segmentation after analysis from the bivariate plots above:

- There are 71 customers belonging to cluster 0, 72 belonging to cluster 1 and 67 belonging to cluster 2.
- Cluster 1 consists of customers with highest spending, highest advance payments, highest current balance, highest credit limit, highest maximum spending in a single shopping and highest probability of full payments. With regard to minimum payment amount, they are second highest after cluster 2.
- Cluster 0 are customers with lowest spending, lowest advance payments, lowest current balance, lowest credit limit, lowest probability of full payments and lowest maximum spent in a single shopping. They are the highest with respect to minimum payment amount.

- Cluster 2 are customers who fall in between customers in cluster 0 and cluster 1 and are second highest with regard to the attributes spending, advance payments, probability of full payment, current balance, credit limit and maximum spent in single shopping. They exhibit the lowest value with respect to minimum payment amount
- Looking at the observations, we can conclude that cluster 1 consists of wealthy/privileged customers, cluster 0 consists of not so wealthy/privileged customers and cluster 2 consists of customers who rank just below wealthy/privileged customers.
- For sake of better understanding, we could say cluster 1 are gold customers, cluster 2 are silver customers and cluster 0 are bronze customers.

1.5 Describe cluster profiles for the clusters defined (2.5 pts). Recommend different promotional strategies for different clusters in context to the business problem in-hand (2.5 pts). After adding the final clusters to the original dataframe, do the cluster profiling. Divide the data in the finalyzed groups and check their means. Explain each of the group briefly. There should be at least 3-4 Recommendations. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks will only be allotted if the recommendations are correct and business specific. variable means. Students to explain the profiles and suggest a mechanism to approach each cluster. Any logical explanation is acceptable.

Solution:

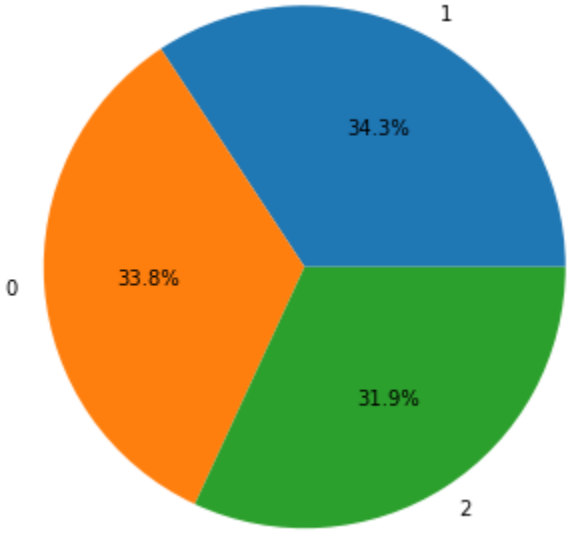
Hierarchical clustering and K Means clustering was applied to the data set. The following gives the top 5 records with the cluster details appended. The clusters attribute denotes hierarchical clustering and Clus_kmeans denote K Means clustering.

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	Clus_kmeans
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	1
10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0
17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1	1

The means of the clusters were computed as given below:

```
Average of spending
--------------------------------------------------
Clus_kmeans
0    11.856944
1    18.495373
2    14.437887
Name: spending, dtype: float64

Average of advance_payments
--------------------------------------------------
Clus_kmeans
0    13.247778
1    16.203433
2    14.337746
Name: advance_payments, dtype: float64

Average of probability_of_full_payment
--------------------------------------------------
Clus_kmeans
0    0.848253
1    0.884210
2    0.881597
Name: probability_of_full_payment, dtype: float64

  Average of current_balance
  ----------------------------------------
  Clus_kmeans
  0    5.231750
  1    6.175687
  2    5.514577
  Name: current_balance, dtype: float64

  Average of credit_limit
  ----------------------------------------
  Clus_kmeans
  0    2.849542
  1    3.697537
  2    3.259225
  Name: credit_limit, dtype: float64

  Average of min_payment_amt
  ----------------------------------------
  Clus_kmeans
  0    4.742389
  1    3.632373
  2    2.707341
  Name: min_payment_amt, dtype: float64

Average of max_spent_in_single_shopping
--------------------------------------------------
Clus_kmeans
0    5.101722
1    6.041701
2    5.120803
Name: max_spent_in_single_shopping, dtype: float64
```

The pie chart and clustering plot are seen from the below graphs.



After performing clustering, the following are some of the key observations.

- Customer segmentation has been arrived at with three groups of customers.
- Both hierarchical clustering and kmeans clustering has given us the almost exact grouping (Please refer solutions to questions 1.3 and 1.4).
- The three groups for the ease of explanation is being referred here as Gold, Silver and Bronze.
- The percentage of Gold customers is approx. 34.3%, Silver customers is 31.9% and Bronze customers are approx. 33.8%

Cluster Profiling:

Gold Customers:

- As the name suggests, this refers to the customers who bring a high amount of business to the bank.
- These customers spend the largest amount of money as can be seen from the spending attribute.
- They also pay the largest advance amount payments in cash.
- These customers hold the highest current balance.
- Since, these customers are high spenders, they also hold the highest credit limit.
- This group also spends the maximum in a single shopping.
- These customers have the highest probability of making full payments which shows that they are safe customers for the bank and the chances of them defaulting is almost nil.
- With regard to minimum payment amount, they are second highest.

Silver Customers:

- These customers fall in between the gold and the bronze customers.
- They are second highest or medium with regard to many of the attributes.
- They are medium spenders as seen from their spending attribute.
- With regard to the advance payments, they are only next to the gold customers.
- They have the second highest probability with respect to making full payment.
- They are the second highest with regard to their current balance.

- Since, their spending is not highest, the credit limit provided is also not the highest but second highest.
- As seen from their spending habits, they stand second highest with respect to maximum spending in a single shopping.
- However, they exhibit the lowest value with respect to minimum payment amount which is the minimum paid by the customer while making payments for purchases made monthly. This observation has to be taken into account as some of these customers might end up being defaulters.

Bronze Customers:

- Bronze customers exhibit the least values with respect to most of the attributes.
- They are people with the least spending.
- These are also customers who make the lowest advance payments.
- These customers maintain the least current balance.
- They have the lowest credit limit.
- These customers have the lowest probability of making full payments.
- They also spend the least in a single shopping.
- They are the highest with respect to minimum payment amount which is the minimum paid by the customer while making payments for purchases made monthly. This observation shows that there is very low chance that these customers will end up as defaulters.

Recommendations to the business: Promotional Offers

- As observed, Gold customers seem to be model customers who bring a lot of value to the business and who also make their payments on time, hence the bank can increase the credit limit for the Gold customers which will encourage more spending by the customers and hence more revenue to the bank.
- The bank may also contemplate the idea of offering the Gold customers add on credit cards which will appease the customers and which in turn will result in additional revenue for the bank looking at the spending habits of these customers.

- The Bronze customers are the highest with respect to minimum payment amount which is the minimum paid by the customer while making payments for purchases made monthly. This observation shows that there is very low chance that these customers will end up as defaulters. Hence, credit limit of these customers could be increased by the bank which will encourage their spending and thus improving the business of the bank and at the same time the limit for minimum payment amount for purchase made monthly could be decreased for these customers which will benefit both the customers as well as the bank.
- Since, Silver customers are medium spending customers, so as to encourage them, add on cards could be offered or their credit limit could be increased. It is better that only one of these options are offered to these customers since they seem to be the least with respect to minimum payment amount for purchase made monthly. This caution should be exercised by the bank so as to prevent defaulters.

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.

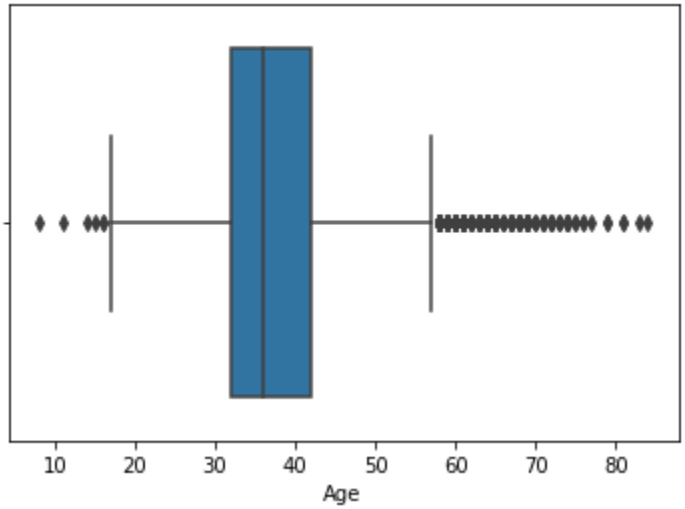
2.1 Read the data and do exploratory data analysis (4 pts). Describe the data briefly. Interpret the inferences for each (2 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

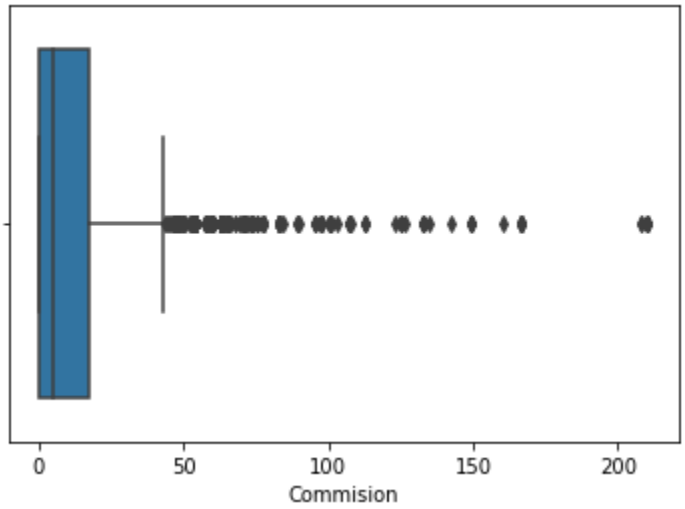
Solution:

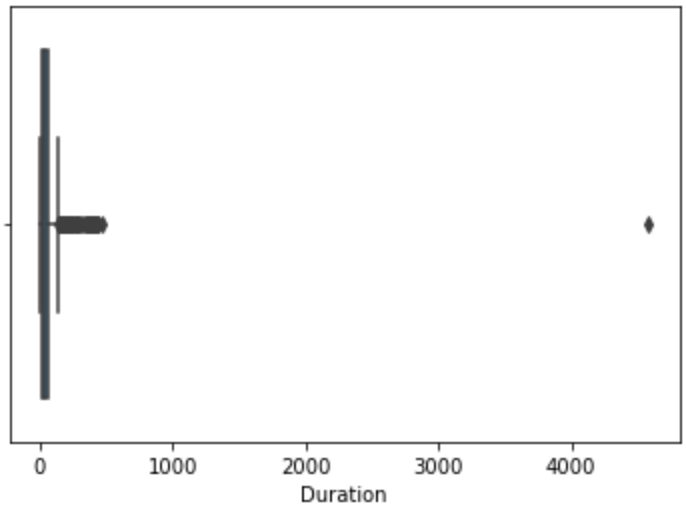
The following are some observations after initial exploration of the data: (Details in Python file)

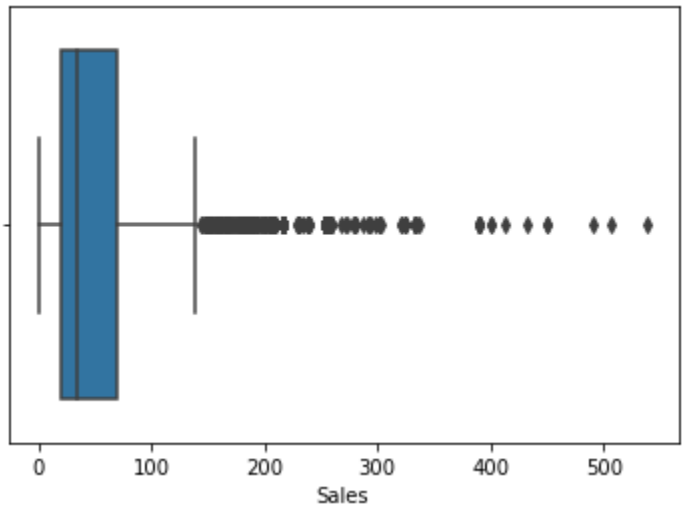
- The dataset consists of 3000 rows and 10 columns.
- 4 attributes are numeric attributes and 6 attributes are object data types.

- There are no missing values in the data.
- There are 139 duplicate records in the data set.
- Also, there are no bad data which is seen from the output of the ‘info’ command.
- There are outliers seen in all the 4 numeric attributes, as seen in the below boxplots (No outlier treatment is being done as per the instructions-FAQ in the question.)
- There is an anomaly seen in the duration attribute as seen from the description of the data.









Treating Anomalies:

The minimum value in the attribute, Duration is seen as -1 which is not possible as duration has to be a positive integer. This value has to be imputed. This value is imputed with the modal value which is 0. Since, this attribute though numeric is an integer value (discrete values) which denotes number of days, using the mode which is the most frequent value seemed to be suitable. Also, this being the minimum value seems to be better for this imputation.

Data Visualization- Univariate Analysis

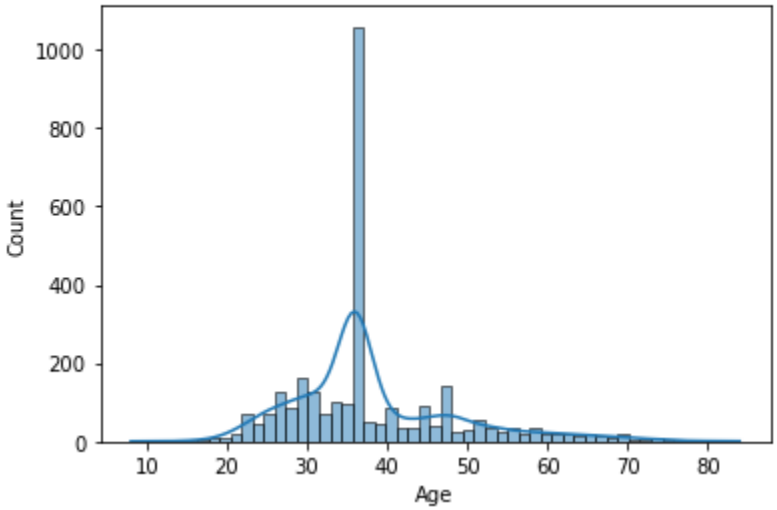
Insights from the Data

Age: Age of insured

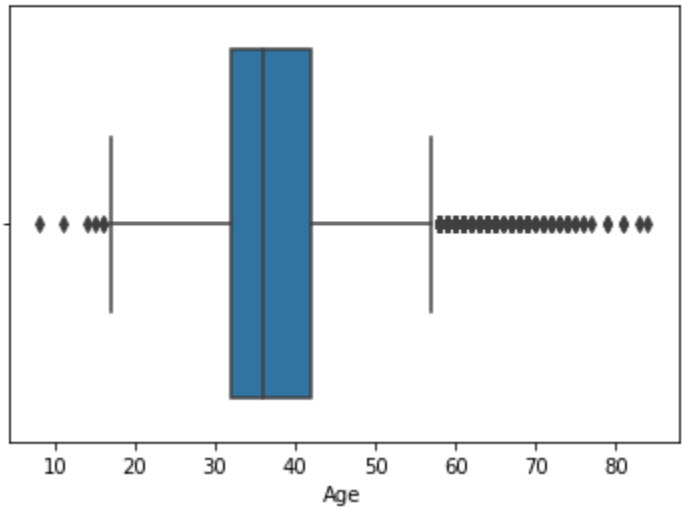
- The age of insured ranges from 8 to 84 years.
- The mean age is 38.09 and the median age is 36.

- Since the mean age is greater than the median age, this indicates that the distribution is right tailed.
- 75% of the claims have age of the insured less than or equal to 42 years.
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the presence of outliers in the data.

Distribution of Age



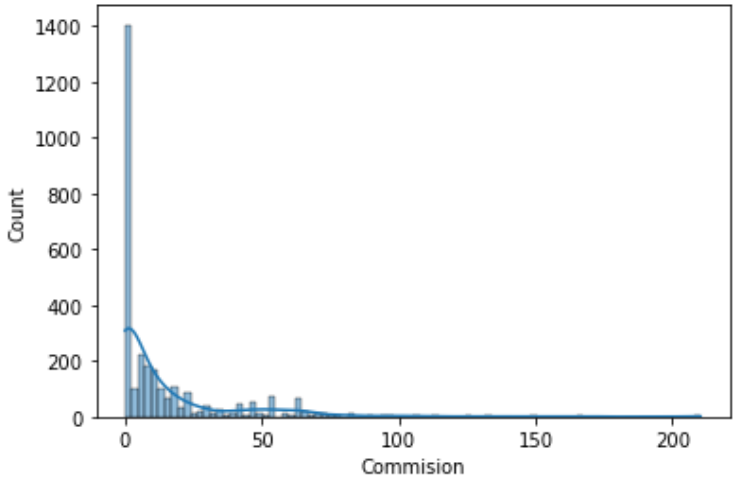
BoxPlot of Age



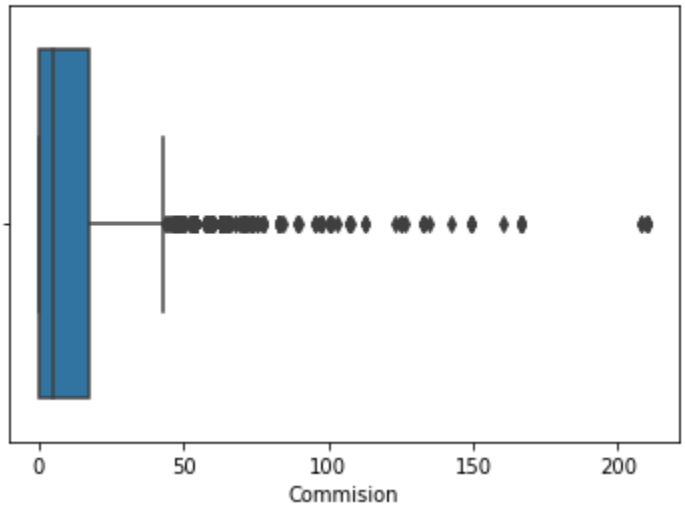
Commission : The commission received for tour insurance firm

- The commission received for tour insurance firm ranges from 0 to 210.21.
- The mean commission received is 14.53 and the median is 4.63.
- Since the mean is greater than the median, this indicates that the distribution is right tailed.
- 75% of the claims have commission received less than or equal to 17.24.
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the presence of outliers in the data.

Distribution of Commision

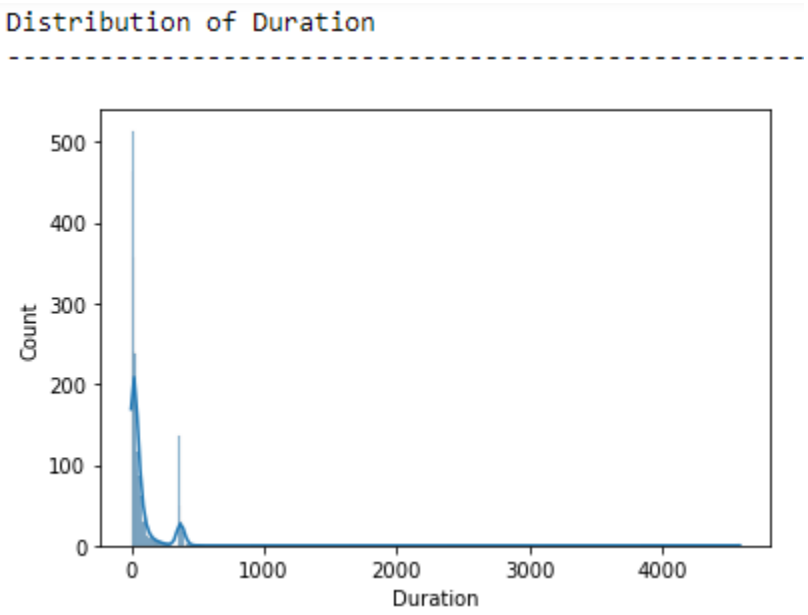


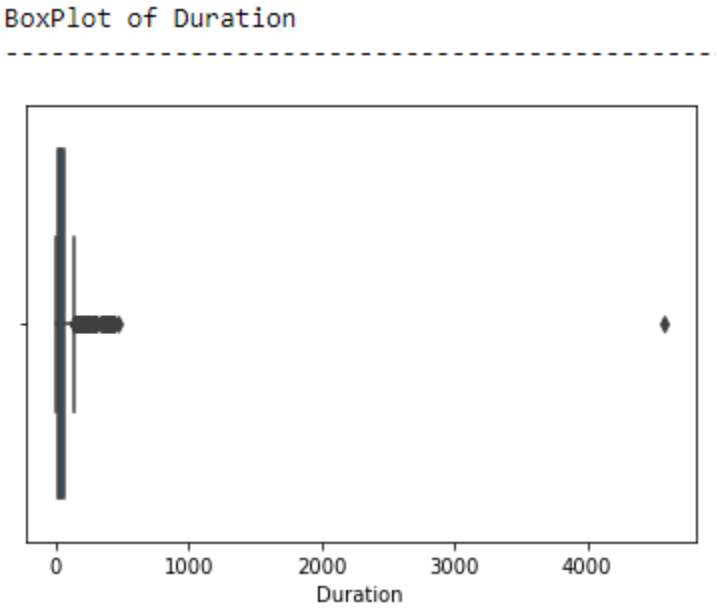
BoxPlot of Commision



Duration: Duration of the tour

- The duration of the tour ranges from 0 to 4580.
- The mean duration is 70 and the median is 26.5.
- Since the mean is greater than the median, this indicates that the distribution is right tailed.
- 75% of the duration is less than or equal to 63.
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the presence of outliers in the data.

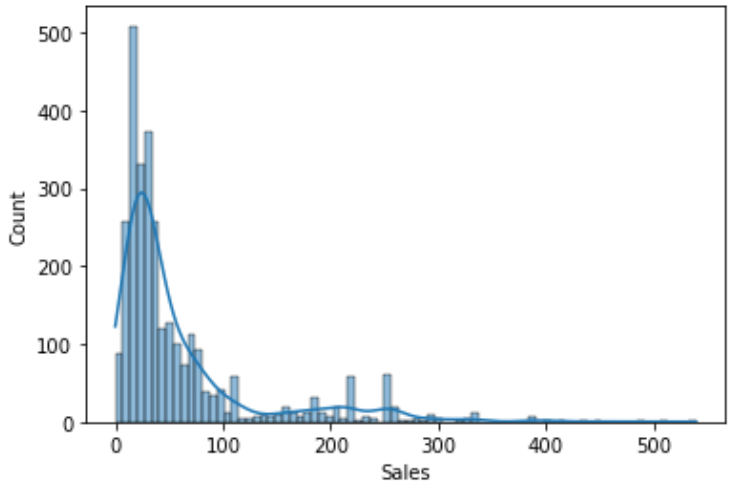




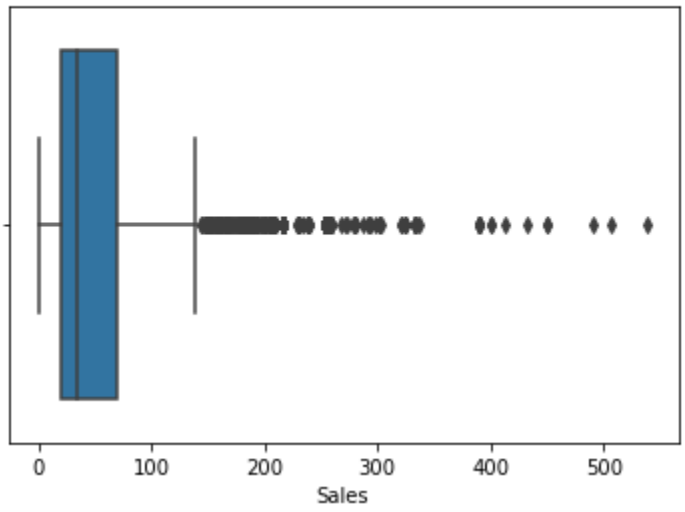
Sales: Amount of sales of tour insurance policies

- The amount of sales of tour insurance policies ranges from 0 to 539.
- The mean sales is 60.25 and the median is 33.
- Since the mean is greater than the median, this indicates that the distribution is right tailed.
- 75% of the sales is less than or equal to 69.
- The histogram and the boxplot of the distribution is given in the below figures.
- The boxplot indicates the presence of outliers in the data.

Distribution of Sales

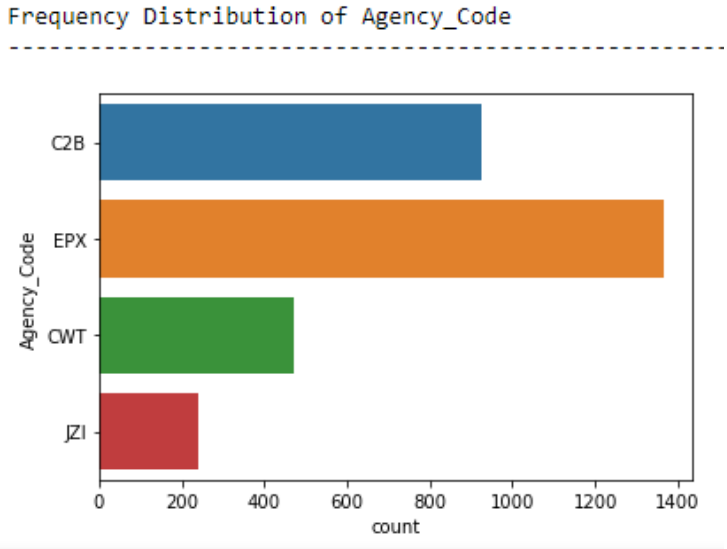


BoxPlot of Sales



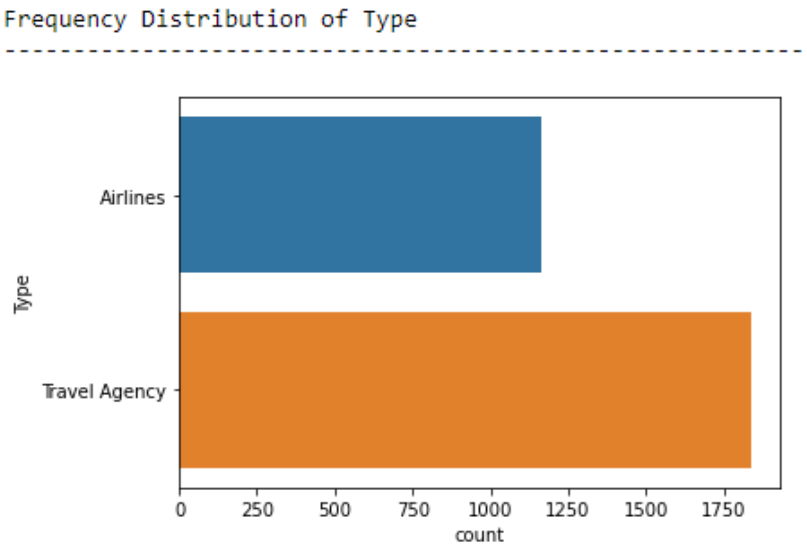
Agency_Code: Code of Tour firm

- There are 4 tour firms which operate.
- Maximum claims i.e. 45.5% is from agency with code EPX.
- Minimum claims i.e. 8% is from agency with code JZI.



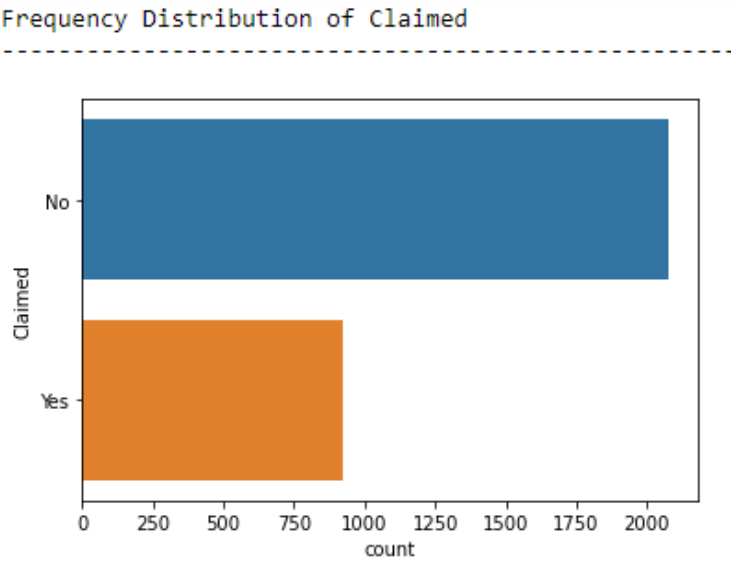
Type: Type of tour insurance firms

- There are 2 types of tour insurance firms.
- Maximum claims i.e. 61.2% is from the type Travel Agency.
- Minimum claims i.e. 38.8 % is from the type Airlines.



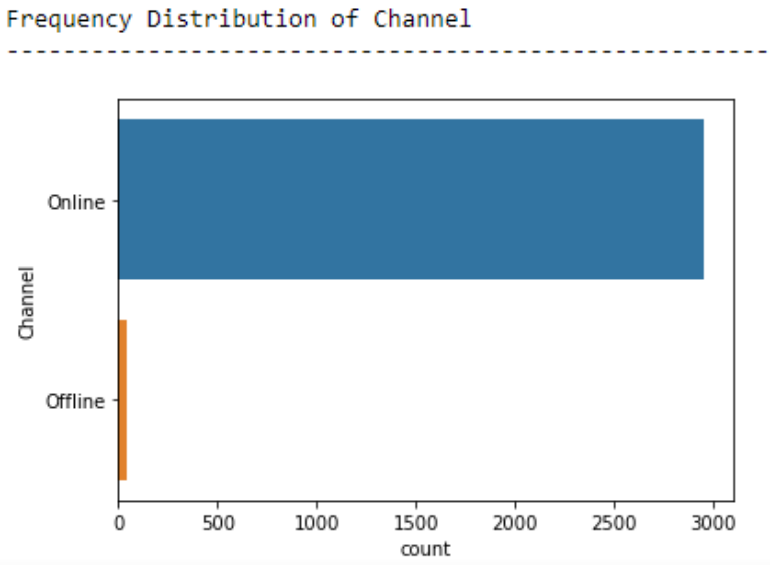
Claimed: Claim Status

- There are 2 types of claim status.
- Maximum claims i.e. 69.2% is of claim status ‘No’.
- Minimum claims i.e. 30.8 % is of claim status ‘Yes’.



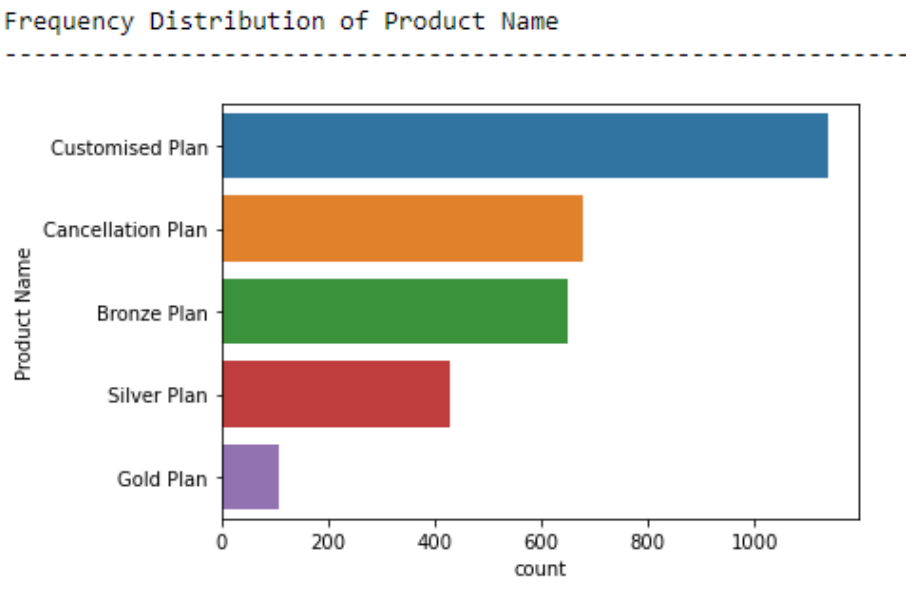
Channel: Distribution channel of tour insurance agencies

- There are 2 distribution channels of tour insurance agencies.
- Maximum claims i.e. 98.5% belongs to the distribution channel Online.
- Minimum claims i.e. 1.5 % belongs to the distribution channel Offline.



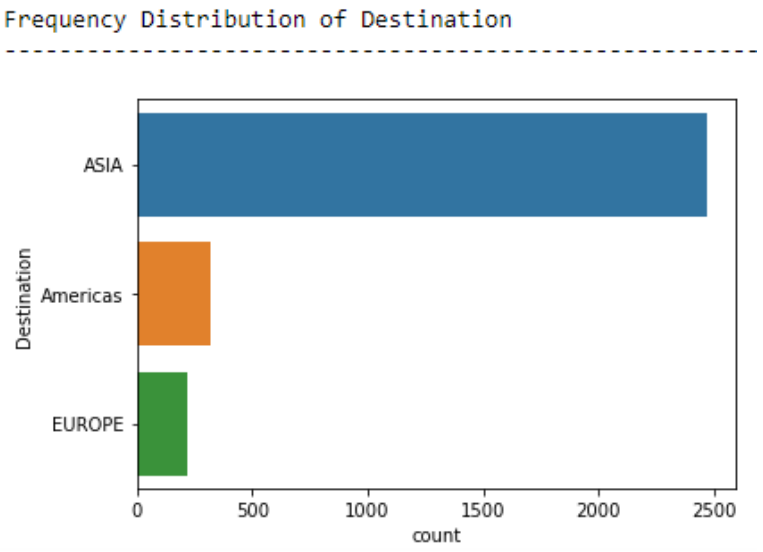
Product Name: Name of the tour insurance products

- There are 5 tour insurance products.
- Maximum claims i.e. 37.9% belongs to the product, Customised Plan.
- Minimum claims i.e. 3.6 % belongs to the product, Gold Plan.



Destination: Destination of the tour

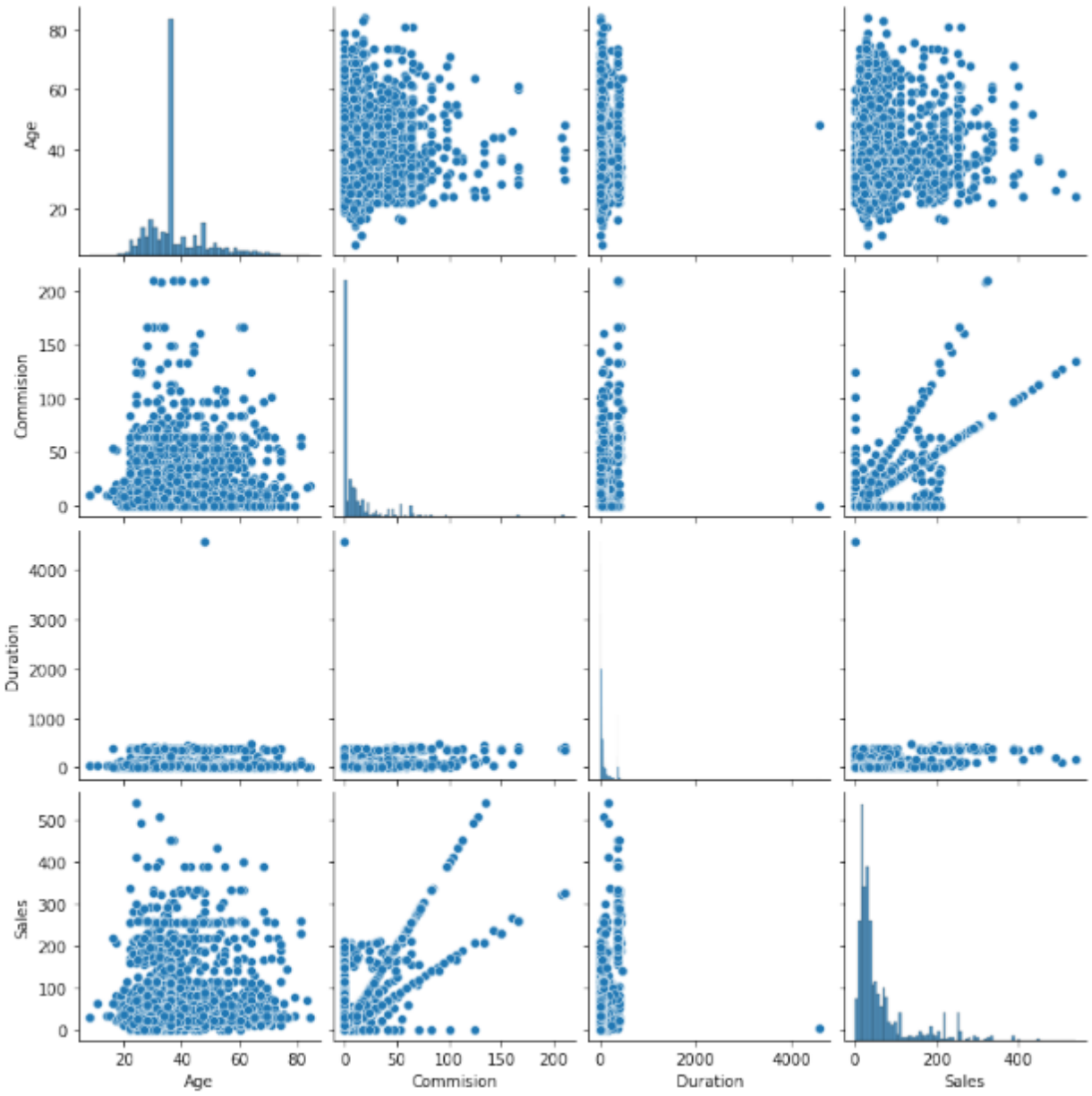
- There are 3 destinations.
- Maximum claims i.e. 82.2% is from destination, ASIA.
- Minimum claims i.e. 7.2 % is from destination, EUROPE.



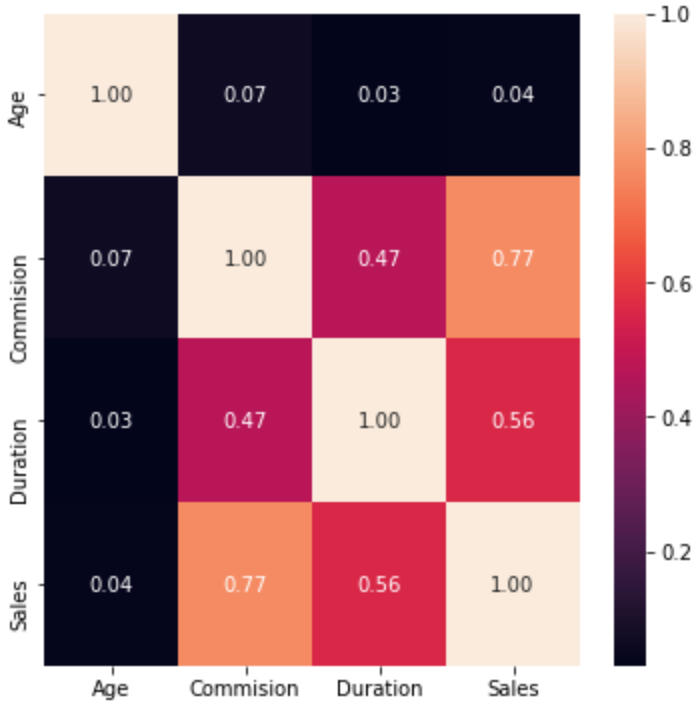
Data Visualization- Bivariate Analysis

Insights from the Data

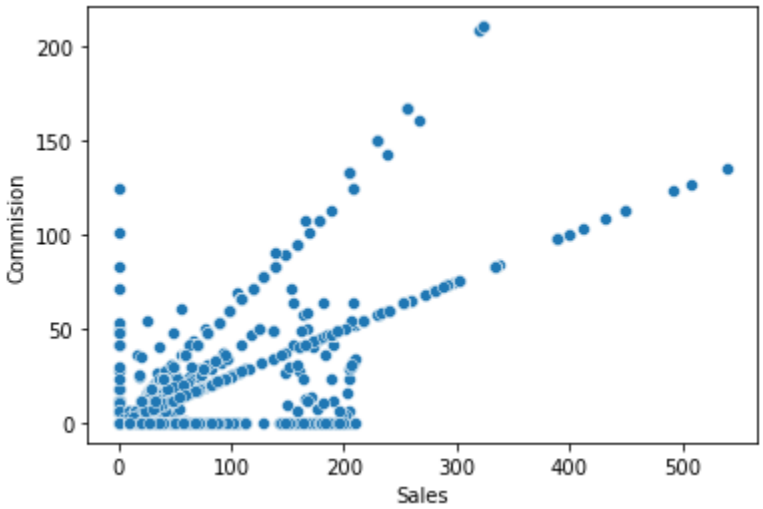
The following shows the pairplot of the numeric variables:



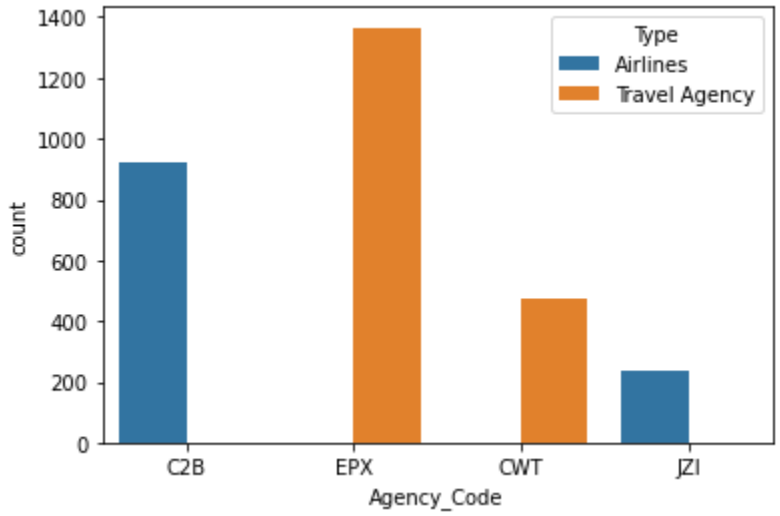
The following is the heat map and the correlation matrix:



- From the pair plot and the heat map, it can be concluded that no strong positive correlation lies between the variables.
- There seems to be a moderate positive correlation between the variables, Sales and Commission with a correlation coefficient of 0.77 as seen in the below scatterplot.
- It can be concluded that as Sales increases, the Commission also increases.



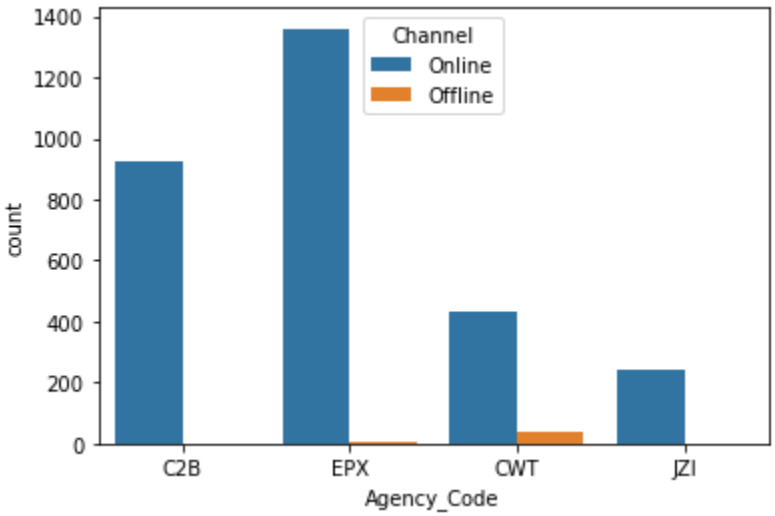
- The below count plot between the categorical variables Agency_Code and Type indicates that the agencies C2B and JZI are of Type Airlines whereas EPX and CWT are of Type Travel Agency.



- The following table gives the number of the agencies across the 2 types of tour firms.

Type	Airlines	Travel Agency	All
Agency_Code			
C2B	924	0	924
CWT	0	472	472
EPX	0	1365	1365
JZI	239	0	239
All	1163	1837	3000

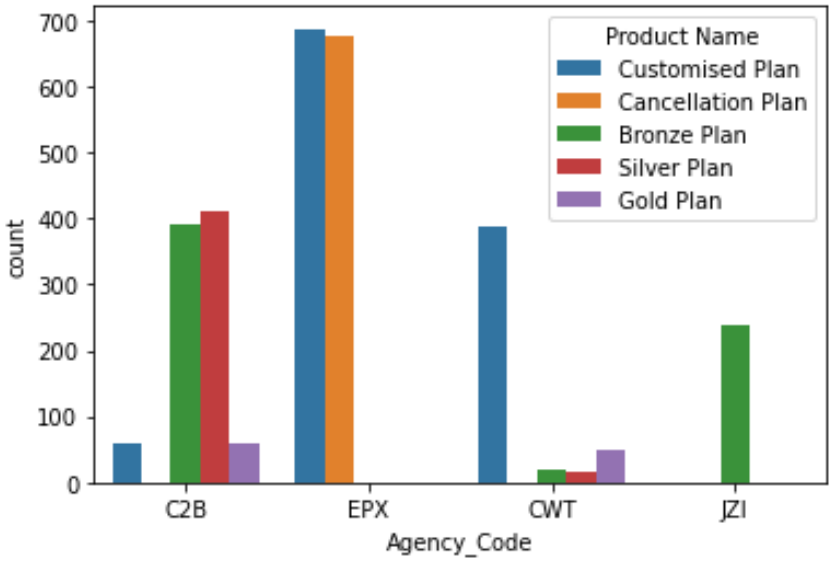
- The below count plot between the categorical variables Agency_Code and Channel indicates that the agencies C2B and JZI only operate online whereas EPX and CWT operate both online and offline.



- The following table gives the number of the agencies across the 2 Channels.
- The table clearly shows that in the case the agency EPX, the number of offline Channels is very low when compared to the Online Channel.
- Only 0.2 % of EPX agency operate offline as compared to 45.3% which operate online.

Channel	Offline	Online	All
Agency_Code			
C2B	0	924	924
CWT	40	432	472
EPX	6	1359	1365
JZI	0	239	239
All	46	2954	3000

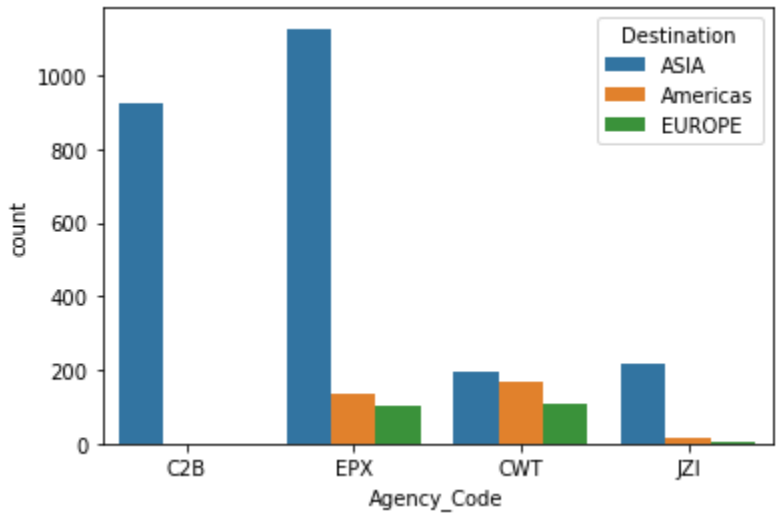
- The below is the count plot between the categorical variables Agency_Code and Product.
- The plot shows that the agencies C2B and CWT does not offer the product, Cancellation Plan.
- It is seen that the agency EPX only 2 products, Customised Plan and Cancellation Plan.
- The agency JZI offers only 1 product, Bronze Plan.



- The following table gives the distribution of different products offered by the agencies.

Product Name	Bronze Plan	Cancellation Plan	Customised Plan	Gold Plan	Silver Plan	All
Agency_Code						
C2B	392	0	60	60	412	924
CWT	19	0	389	49	15	472
EPX	0	678	687	0	0	1365
JZI	239	0	0	0	0	239
All	650	678	1136	109	427	3000

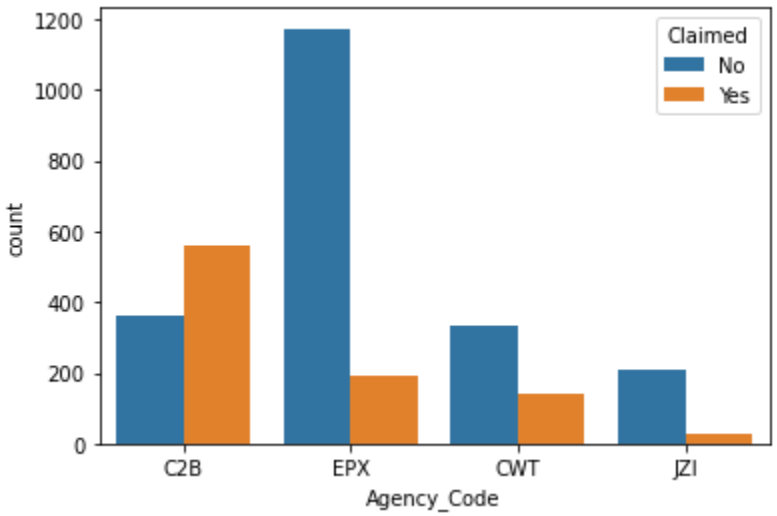
- The below is the count plot between the categorical variables Agency_Code and Destination.
- The plot shows that the agencies C2B offers tour in destination Asia only.
- It is seen that the agencies EPX, CWT and JZI offer tours in all destinations.



- The following table gives the distribution of the destinations offered by the agencies.

Destination	ASIA	Americas	EUROPE	All
Agency_Code				
C2B	924	0	0	924
CWT	194	170	108	472
EPX	1128	134	103	1365
JZI	219	16	4	239
All	2465	320	215	3000

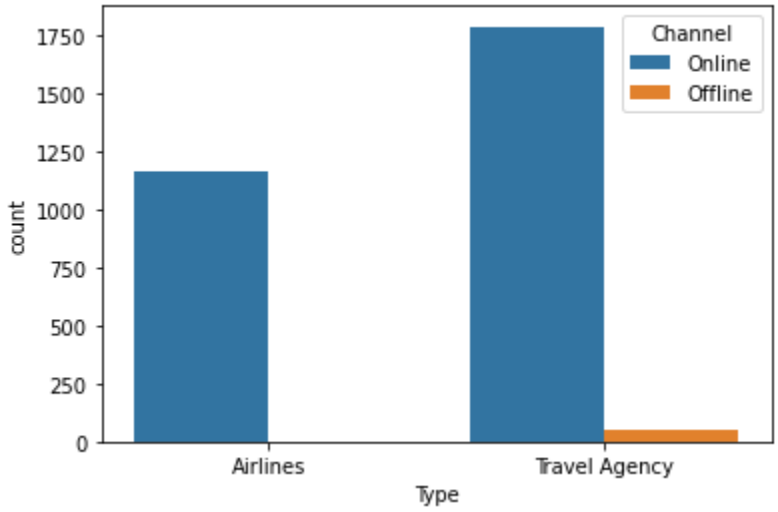
- The below is the count plot between the categorical variables Agency_Code and Claimed.
- The agency C2B has the maximum claims, 18.7%.
- The agency JZI has the minimum claims, 1%.



- The following table gives the distribution of the claim status by the agencies.

Claimed	No	Yes	All
Agency_Code			
C2B	364	560	924
CWT	331	141	472
EPX	1172	193	1365
JZI	209	30	239
All	2076	924	3000

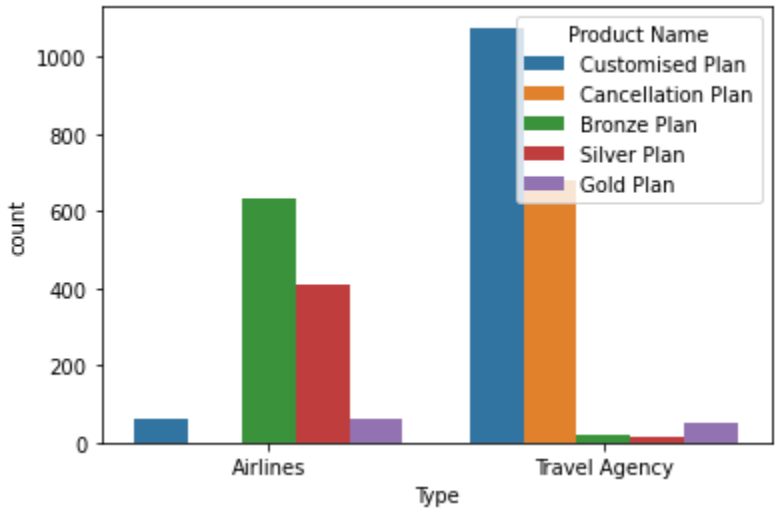
- The below is the count plot between the categorical variables Type and Channel.
- The tour type Airlines operates only the online channel.
- The tour type Travel Agency operates both online and offline. But the offline channel constitutes only 1.5%.



- The following table gives the distribution of Channel across tour type.

Channel	Offline	Online	All
Type			
Airlines	0	1163	1163
Travel Agency	46	1791	1837
All	46	2954	3000

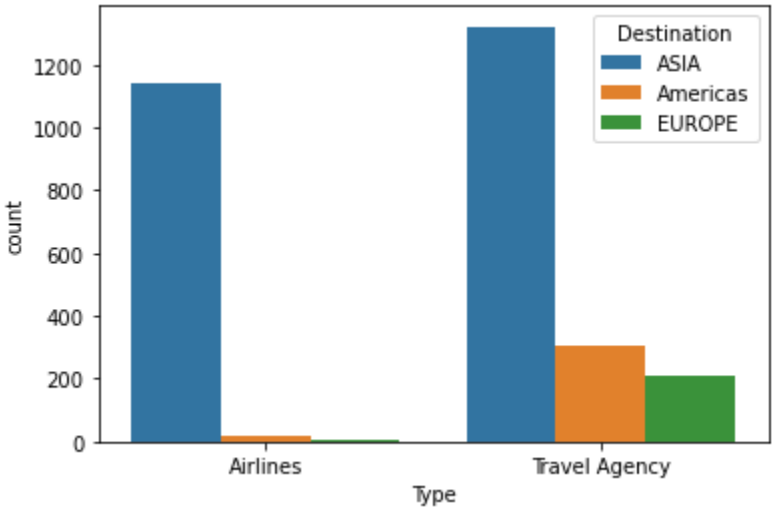
- The below is the count plot between the categorical variables Type and Product Name.
- The tour type Airlines does not offer the product, Cancellation Plan.
- Travel Agency's customized plan has a maximum value 35.9% and its silver plan is the minimum 0.5%.



- The following table gives the distribution of the various products across tour types.

Product Name	Bronze Plan	Cancellation Plan	Customised Plan	Gold Plan	Silver Plan	All
Type						
Airlines	631	0	60	60	412	1163
Travel Agency	19	678	1076	49	15	1837
All	650	678	1136	109	427	3000

- The below is the count plot between the categorical variables Type and Destination.
- The destination Asia seems to be the most popular among both the tour types.
- EUROPE is the least occurring destination among the tour types.

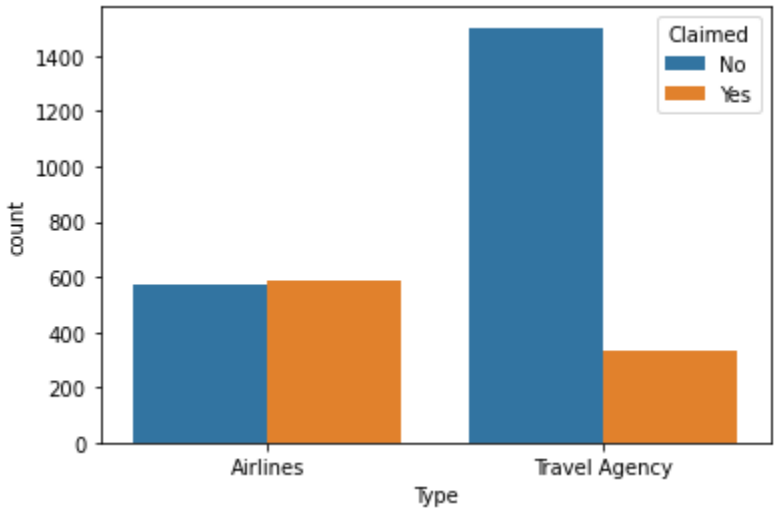


- The following table gives the distribution of the destinations across tour types.

Destination	ASIA	Americas	EUROPE	All
Type				
Airlines	1143	16	4	1163
Travel Agency	1322	304	211	1837
All	2465	320	215	3000

Destination	ASIA	Americas	EUROPE	All
Type				
Airlines	0.381000	0.005333	0.001333	0.387667
Travel Agency	0.440667	0.101333	0.070333	0.612333
All	0.821667	0.106667	0.071667	1.000000

- The below is the count plot between the categorical variables Type and Claimed status.
- Maximum claims is for tour type Airlines with 19.7%.

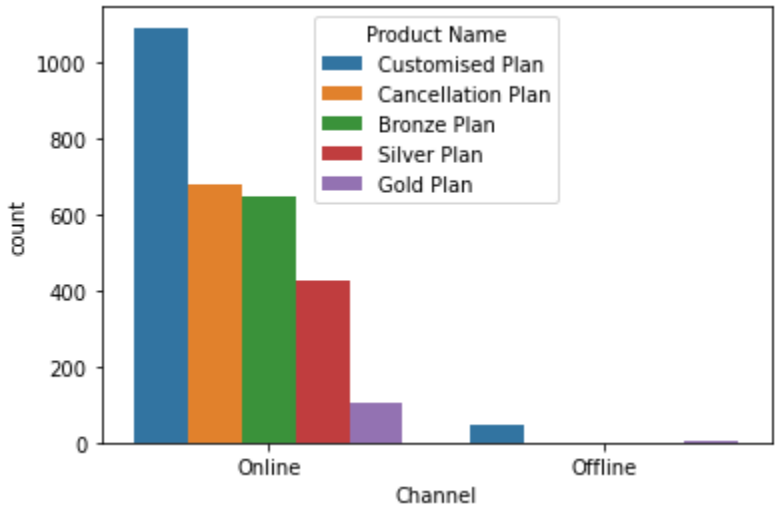


- The following tables gives the distribution of the claims across tour types.

Claimed	No	Yes	All
Type			
Airlines	573	590	1163
Travel Agency	1503	334	1837
All	2076	924	3000

Claimed	No	Yes	All
Type			
Airlines	0.191	0.196667	0.387667
Travel Agency	0.501	0.111333	0.612333
All	0.692	0.308000	1.000000

- The below is the count plot between the categorical variables Channel and Product Name.
- The Online channel offers all products.
- The Offline channel offers only Customised Plan and Gold Plan.
- From the distribution, it is clearly seen that Online Channel is the more preferred channel.

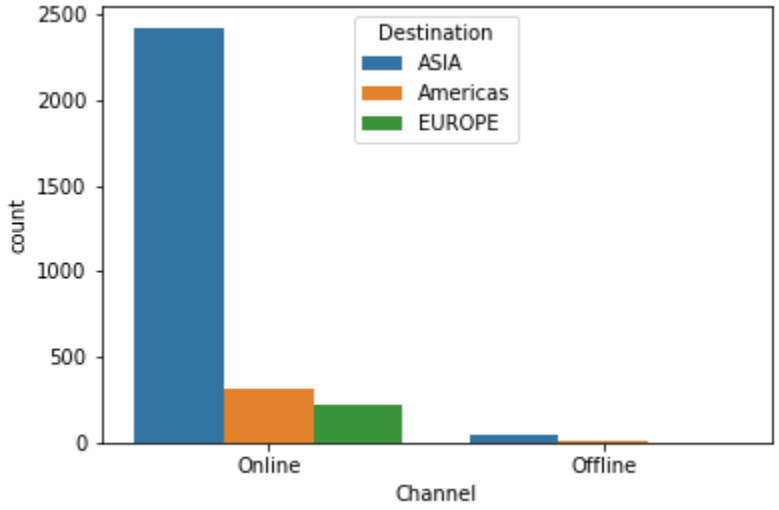


- The following tables gives the distribution of the products across channels.

Product Name	Bronze Plan	Cancellation Plan	Customised Plan	Gold Plan	Silver Plan	All
Channel						
Offline	0	0	44	2	0	46
Online	650	678	1092	107	427	2954
All	650	678	1136	109	427	3000

Product Name	Bronze Plan	Cancellation Plan	Customised Plan	Gold Plan	Silver Plan	All
Channel						
Offline	0.000000	0.000	0.014667	0.000667	0.000000	0.015333
Online	0.216667	0.226	0.364000	0.035667	0.142333	0.984667
All	0.216667	0.226	0.378667	0.036333	0.142333	1.000000

- The below is the count plot between the categorical variables Channel and Destination.
- Offline Channel has the destinations ASIA and Americas only.

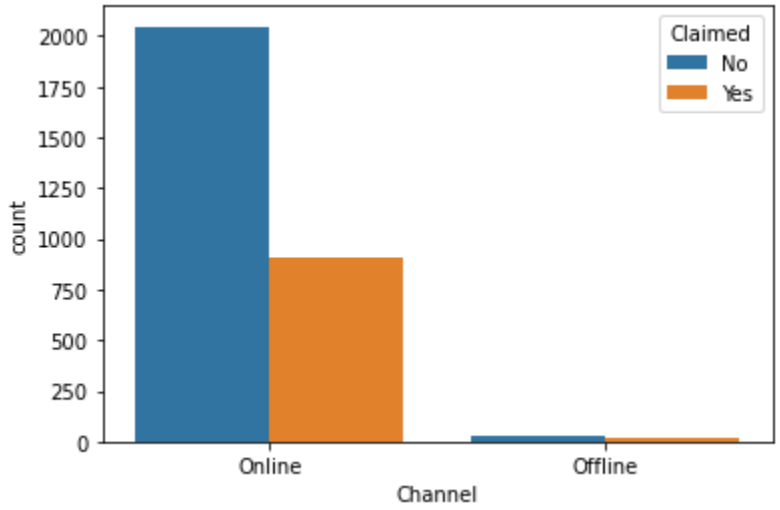


- The following tables gives the distribution of the destinations across channels.

Destination	ASIA	Americas	EUROPE	All
Channel				
Offline	42	4	0	46
Online	2423	316	215	2954
All	2465	320	215	3000

Destination	ASIA	Americas	EUROPE	All
Channel				
Offline	0.014000	0.001333	0.000000	0.015333
Online	0.807667	0.105333	0.071667	0.984667
All	0.821667	0.106667	0.071667	1.000000

- The below is the count plot between the categorical variables Channel and Claimed status.
- The maximum claims of 30.2% is for Online Channel.

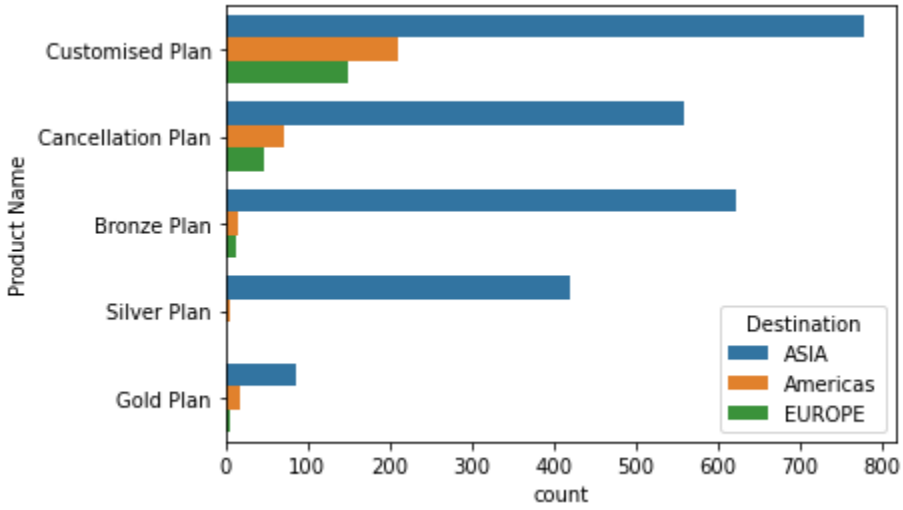


- The following tables gives the distribution of the claims across channels.

Claimed	No	Yes	All
Channel			
Offline	29	17	46
Online	2047	907	2954
All	2076	924	3000

Claimed	No	Yes	All
Channel			
Offline	0.009667	0.005667	0.015333
Online	0.682333	0.302333	0.984667
All	0.692000	0.308000	1.000000

- The below is the count plot between the categorical variables Product Name and Destination.
- The destination ASIA is the maximum preferred across all products.
- The destination EUROPE is the least preferred across the various products.

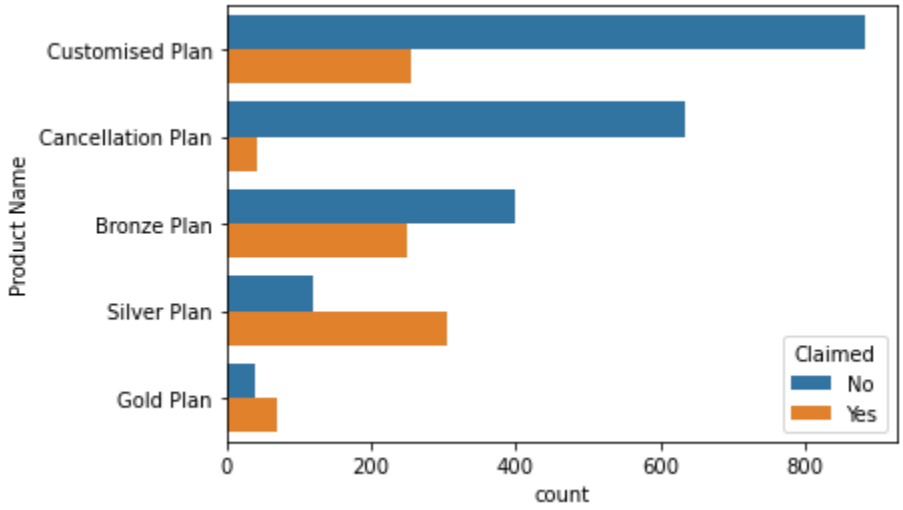


- The following tables gives the distribution of destinations across the different products.

Destination	ASIA	Americas	EUROPE	All
Product Name				
Bronze Plan	622	16	12	650
Cancellation Plan	558	72	48	678
Customised Plan	777	210	149	1136
Gold Plan	87	17	5	109
Silver Plan	421	5	1	427
All	2465	320	215	3000

Destination	ASIA	Americas	EUROPE	All
Product Name				
Bronze Plan	0.207333	0.005333	0.004000	0.216667
Cancellation Plan	0.186000	0.024000	0.016000	0.226000
Customised Plan	0.259000	0.070000	0.049667	0.378667
Gold Plan	0.029000	0.005667	0.001667	0.036333
Silver Plan	0.140333	0.001667	0.000333	0.142333
All	0.821667	0.106667	0.071667	1.000000

- The below is the count plot between the categorical variables Product Name and Claimed status.
- Maximum claims received, 10.2% is for the product Silver Plan.
- Minimum claims received, 1.4% is for the product Cancellation Plan.

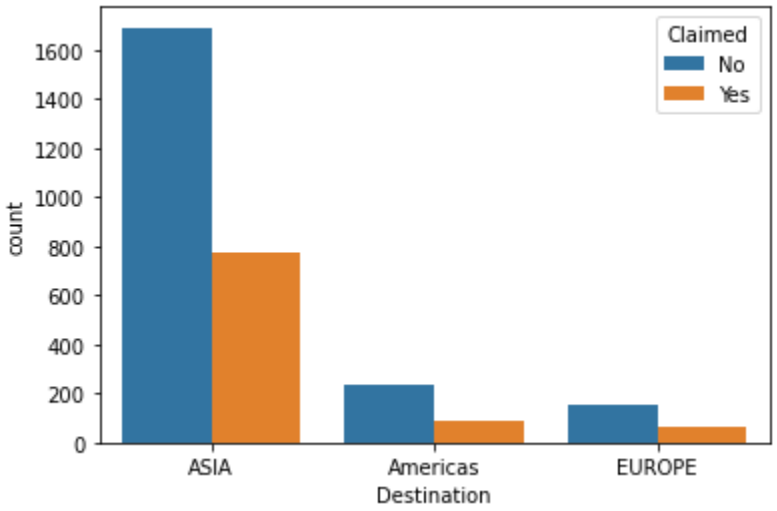


- The following tables gives the distribution of claims across the different products.

Claimed	No	Yes	All
Product Name			
Bronze Plan	399	251	650
Cancellation Plan	635	43	678
Customised Plan	882	254	1136
Gold Plan	39	70	109
Silver Plan	121	306	427
All	2076	924	3000

Claimed	No	Yes	All
Product Name			
Bronze Plan	0.133000	0.083667	0.216667
Cancellation Plan	0.211667	0.014333	0.226000
Customised Plan	0.294000	0.084667	0.378667
Gold Plan	0.013000	0.023333	0.036333
Silver Plan	0.040333	0.102000	0.142333
All	0.692000	0.308000	1.000000

- The below is the count plot between the categorical variables Destination and Claimed status.
- Maximum claims of 25.8% is for the destination ASIA.
- Minimum claims of 2.1% is for the destination EUROPE.



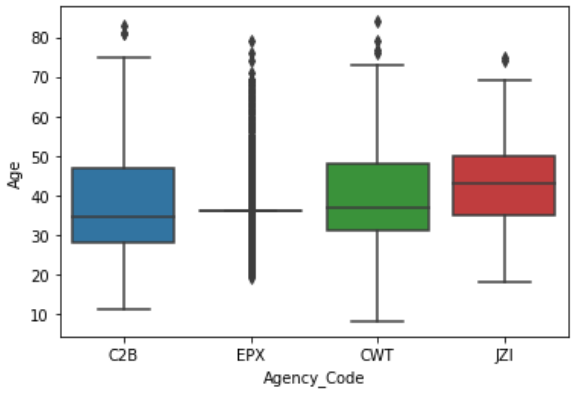
- The following tables gives the distribution of claims across the different destinations.

Claimed	No	Yes	All
Destination			
ASIA	1691	774	2465
Americas	232	88	320
EUROPE	153	62	215
All	2076	924	3000

Claimed	No	Yes	All
Destination			
ASIA	0.563667	0.258000	0.821667
Americas	0.077333	0.029333	0.106667
EUROPE	0.051000	0.020667	0.071667
All	0.692000	0.308000	1.000000

```
Mean of Age for Agency_Code
Agency_Code
C2B    37.765152
CWT    40.141949
EPX    36.832967
JZI    42.485356
Name: Age, dtype: float64
```

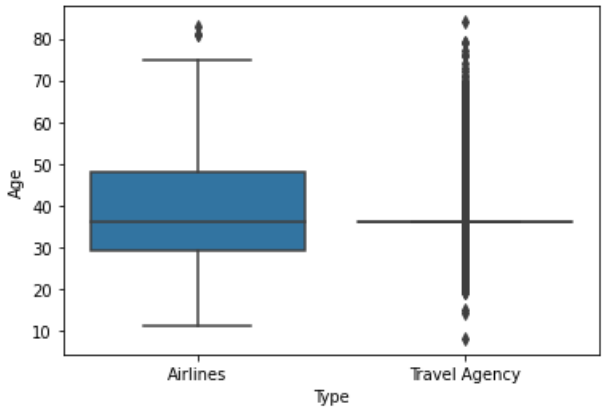
```
Plot of Age vs Agency_Code
--------------------------------------------------
```



- From the above plot, it is seen that the agency JZI has maximum age and the agency EPX has minimum age.

```
Mean of Age for Type
Type
Airlines         38.735168
Travel Agency    37.683179
Name: Age, dtype: float64
```

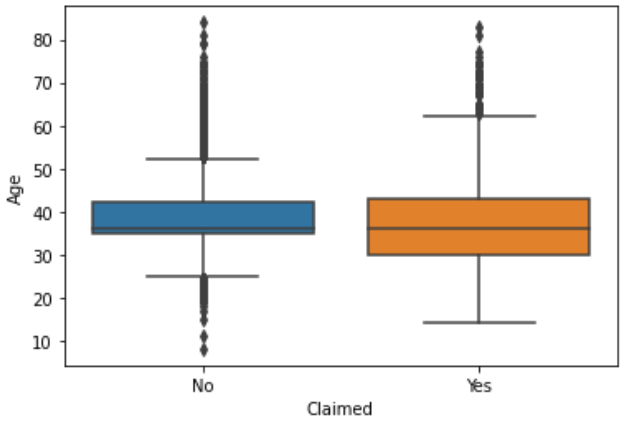
```
Plot of Age vs Type
--------------------------------------------------
```



- From the above plot, it is seen that tour type Airlines has a higher mean age.

```
Mean of Age for Claimed
Claimed
No     38.300578
Yes    37.620130
Name: Age, dtype: float64
```

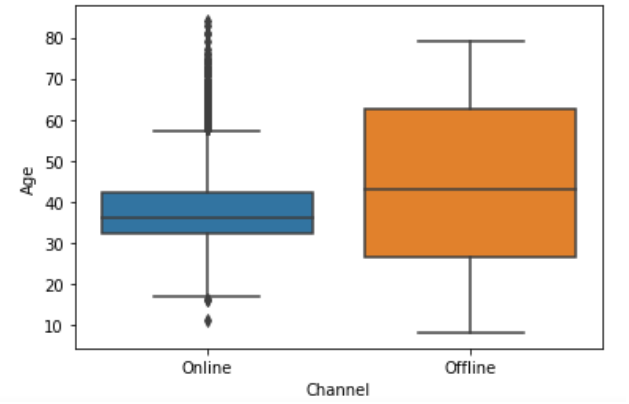
```
Plot of Age vs Claimed
--------------------------------------------------
```



- From the above plot, it is seen that the mean age for claims is slightly less than when compared to the non claims. The means are almost comparable.

```
Mean of Age for Channel
Channel
Offline    43.869565
Online     38.001016
Name: Age, dtype: float64
```

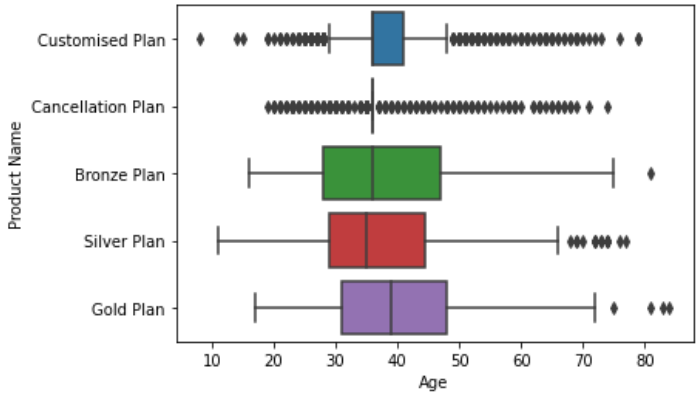
```
Plot of Age vs Channel
--------------------------------------------------
```



- From the above plot, it is seen that the mean age for Online Channel is lesser than the mean age for Offline Channel.

```
Mean of Age for Product Name
Product Name
Bronze Plan          38.412308
Cancellation Plan    36.497050
Customised Plan      38.608275
Gold Plan            41.908257
Silver Plan          37.782201
Name: Age, dtype: float64
```

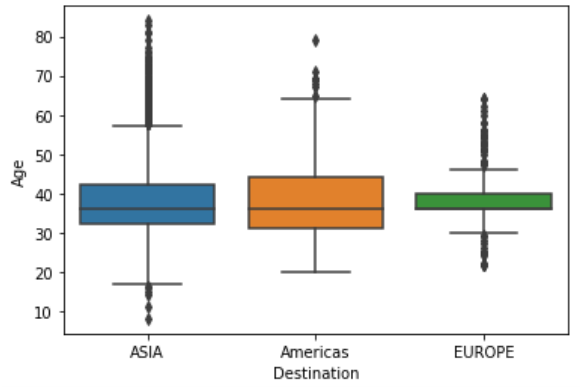
```
Plot of Age vs Product
----------------------------------------------------------------
```



- From the above plot, it is seen that the mean age for the product, Gold Plan is the highest and the mean age for the Cancellation plan is the least.

```
Mean of Age for Destination
Destination
ASIA         38.048276
Americas     38.481250
EUROPE       38.000000
Name: Age, dtype: float64
```

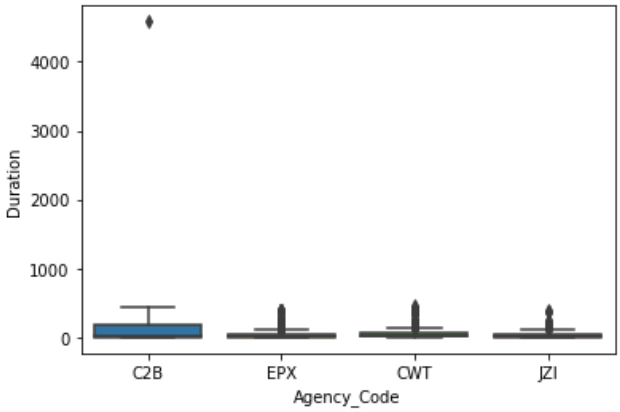
```
Plot of Age vs Destination
-------------------------------------------------
```



- From the above plot, it is seen that the mean ages for all the destinations are almost the same.

```
Mean of Duration for Agency_Code
Agency_Code
C2B    119.404762
CWT     64.733051
EPX     43.374359
JZI     41.485356
Name: Duration, dtype: float64
```

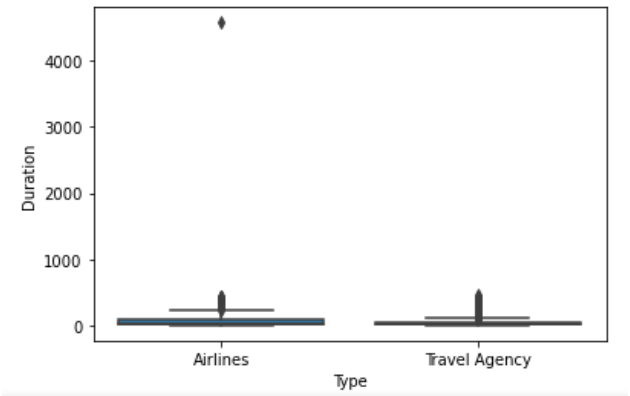
```
Plot of Duration vs Agency_Code
--------------------------------------------------
```



- From the above plot, it is seen that the mean duration of tour for the agency C2B is the highest and that for the agency JZI is the least.

```
Mean of Duration for Type
Type
Airlines         103.392089
Travel Agency     48.862275
Name: Duration, dtype: float64
```

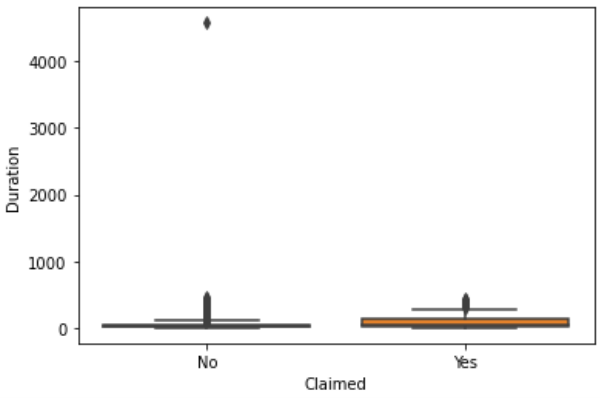
```
Plot of Duration vs Type
--------------------------------------------------
```



- From the above plot, it is seen that the mean duration of tour for tour type Airlines is the highest.

```
Mean of Duration for Claimed
Claimed
No      50.783719
Yes    113.179654
Name: Duration, dtype: float64
```

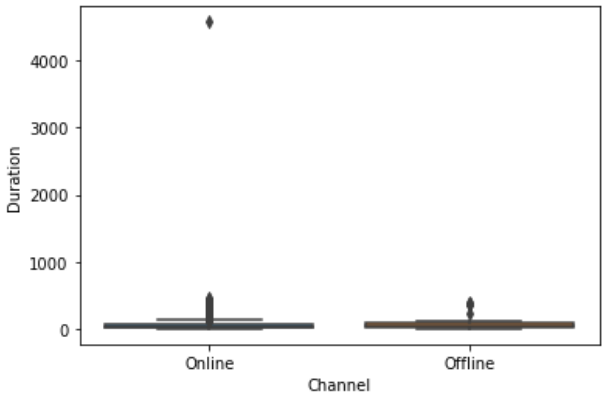
```
Plot of Duration vs Claimed
----------------------------------------------------
```



- From the above plot, it is seen that the mean duration of tour for the claims is the highest as compared to the non claims.

```
Mean of Duration for Channel
Channel
Offline    90.826087
Online     69.677387
Name: Duration, dtype: float64
```

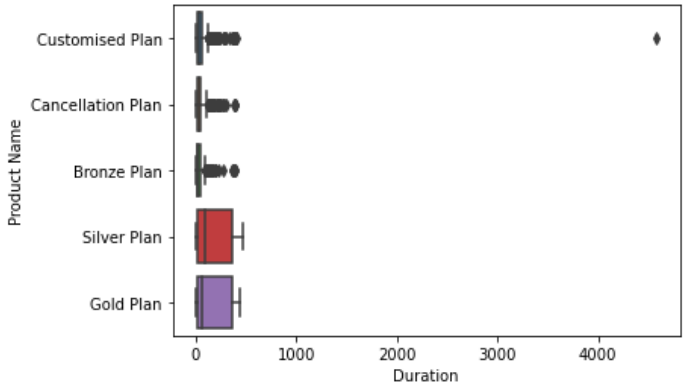
```
Plot of Duration vs Channel
----------------------------------------------------
```



- From the above plot, it is seen that the mean duration of tour for the Offline Channel is the higher than that of the Online Channel.

```
Mean of Duration for Product Name
Product Name
Bronze Plan           35.078462
Cancellation Plan     41.026549
Customised Plan       51.676937
Gold Plan            178.688073
Silver Plan          190.177986
Name: Duration, dtype: float64
```

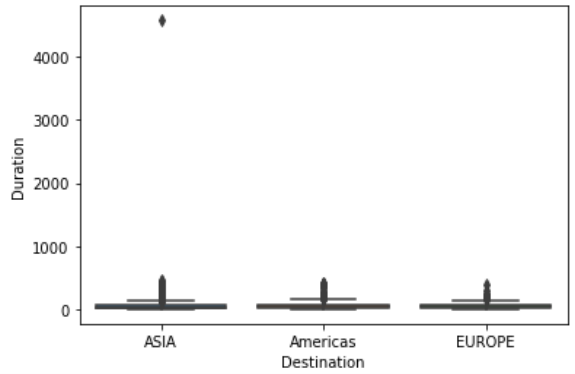
```
Plot of Duration vs Product
------------------------------------------------------------
```



- From the above plot, it is seen that the mean duration of tour for the product Silver Plan is the highest and the least is for the product Bronze Plan.

```
Mean of Duration for Destination
Destination
ASIA         70.443408
Americas     77.409375
EUROPE       53.911628
Name: Duration, dtype: float64
```

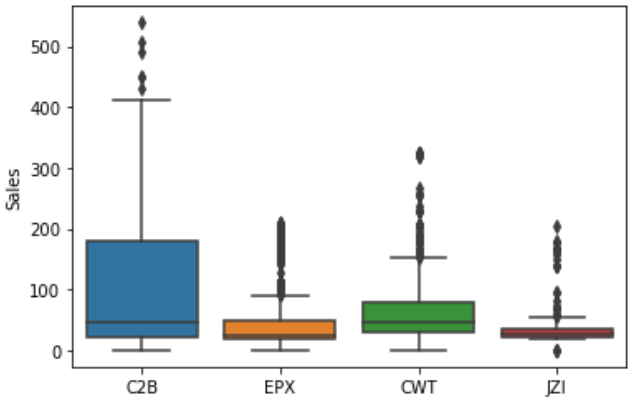
```
Plot of Duration vs Destination
------------------------------------------------------
```



- From the above plot, it is seen that the mean duration of tour for the destination Americas is the highest and the least is for EUROPE.

```
Mean of Sales for Agency_Code
Agency_Code
C2B    94.984632
CWT    66.834852
EPX    38.671810
JZI    36.196109
Name: Sales, dtype: float64
```

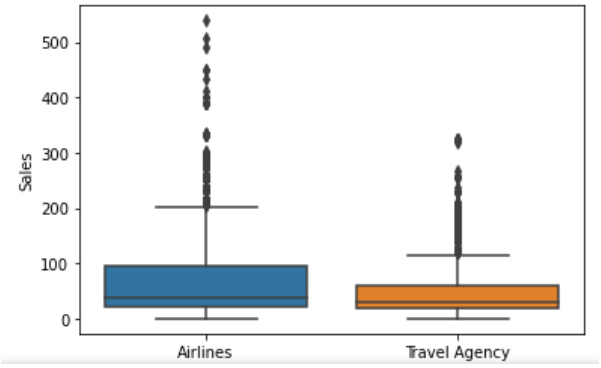
```
Plot of Sales vs Agency_Code
----------------------------------------------------
```



- From the above plot, it is seen that the mean sales of the agency C2B is the highest and that for agency JZI is the least.

```
Mean of Sales for Type
Type
Airlines         82.903414
Travel Agency    45.908040
Name: Sales, dtype: float64
```

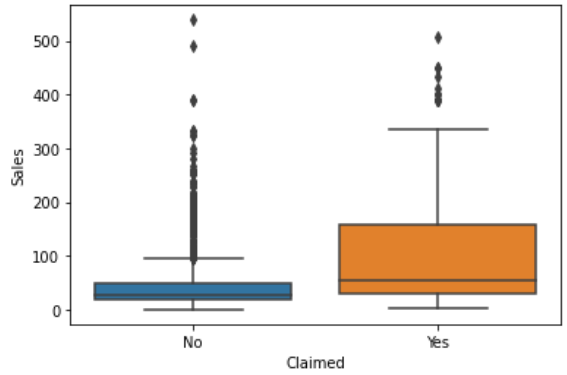
```
Plot of Sales vs Type
----------------------------------------------------
```



- From the above plot, it is seen that the mean sales of the tour type Airlines is greater than Travel Agency.

```
Mean of Sales for Claimed
Claimed
No     43.789133
Yes    97.233225
Name: Sales, dtype: float64
```

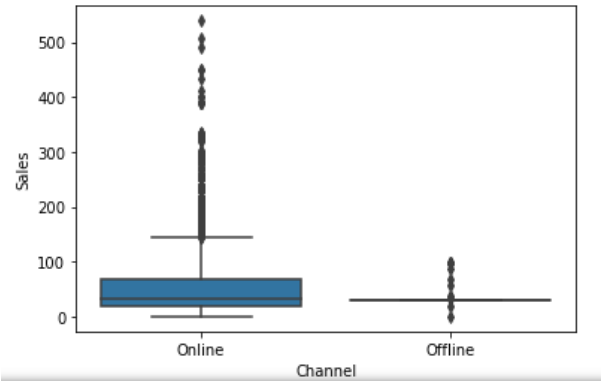
```
Plot of Sales vs Claimed
--------------------------------------------------
```



- From the above plot, it is seen that the mean sales for the claimed status. ‘yes’ is greater than the non claims.

```
Mean of Sales for Channel
Channel
Offline    39.043478
Online     60.580142
Name: Sales, dtype: float64
```

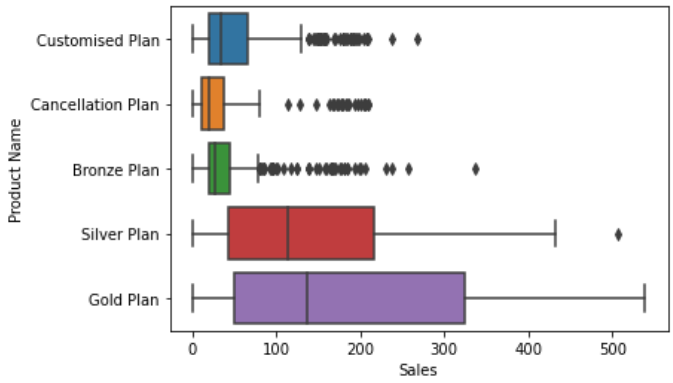
```
Plot of Sales vs Channel
--------------------------------------------------
```



- From the above plot, it is seen that the mean sales for Online channel is greater than the Offline channel.

```
Mean of Sales for Product Name
Product Name
Bronze Plan           39.446754
Cancellation Plan     31.965988
Customised Plan       47.863697
Gold Plan            179.743578
Silver Plan          139.276815
Name: Sales, dtype: float64
```

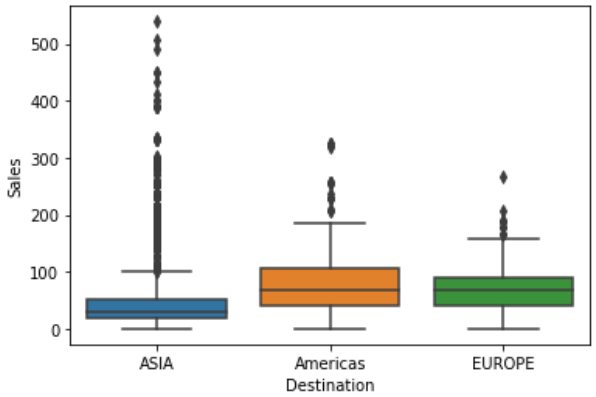
```
Plot of Sales vs Product
------------------------------------------------------------
```



- From the above plot, it is seen that the mean sales for the product Gold Plan is the highest and the mean sales for Cancellation Plan is the least.

```
Mean of Sales for Destination
Destination
ASIA         56.467513
Americas     82.573281
EUROPE       70.390093
Name: Sales, dtype: float64
```

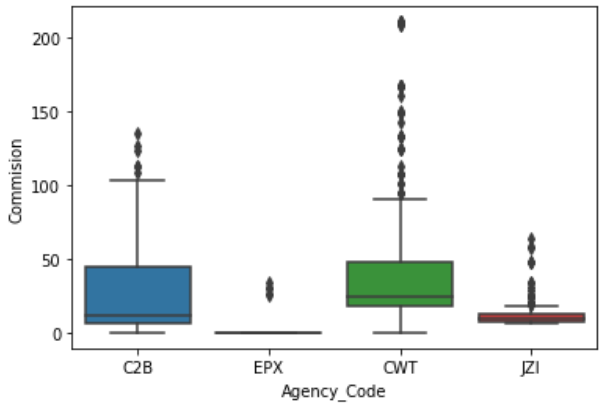
```
Plot of Sales vs Destination
-----------------------------------------------------
```



- From the above plot, it is seen that the mean sales for the destination EUROPE is the highest and the least for ASIA.

```
Mean of Commission for Agency_Code
Agency_Code
C2B    24.006169
CWT    39.144619
EPX     0.108425
JZI    11.638703
Name: Commision, dtype: float64
```

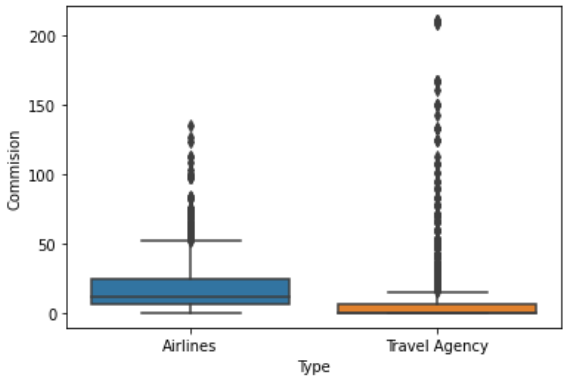
```
Plot of Commission vs Agency_Code
--------------------------------------------------
```



- From the above plot, it is seen that the mean commission for the agency CWT is the highest and the least for EPX.

```
Mean of Commission for Type
Type
Airlines         21.464617
Travel Agency    10.138410
Name: Commision, dtype: float64
```

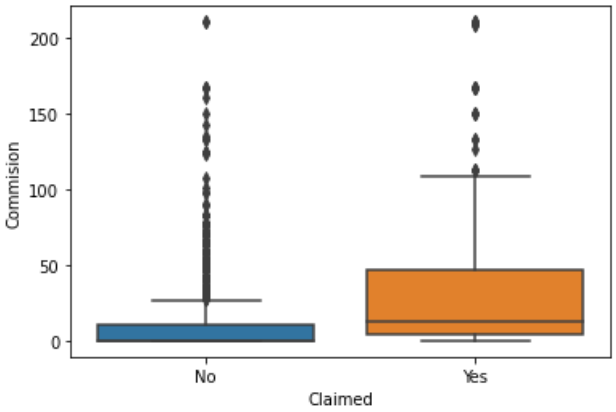
```
Plot of Commission vs Type
--------------------------------------------------
```



- From the above plot, it is seen that the mean commission for the tour type Airlines is the highest.

```
Mean of Commission for Claimed
Claimed
No      9.472606
Yes    25.890130
Name: Commision, dtype: float64
```

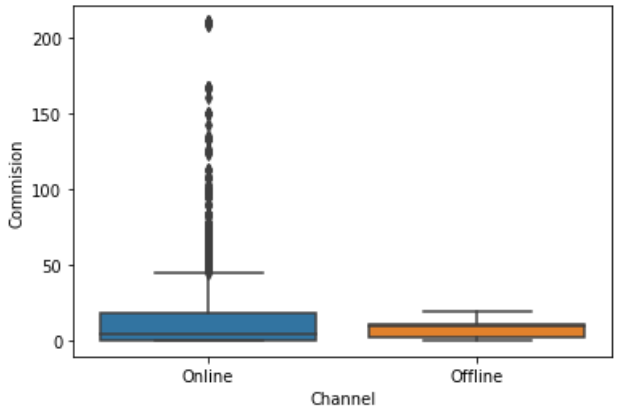
```
Plot of Commission vs Claimed
-----------------------------------------------------
```



- From the above plot, it is seen that the mean commission for claimed status ‘Yes’ is highest.

```
Mean of Commission for Channel
Channel
Offline     7.676957
Online     14.635907
Name: Commision, dtype: float64
```

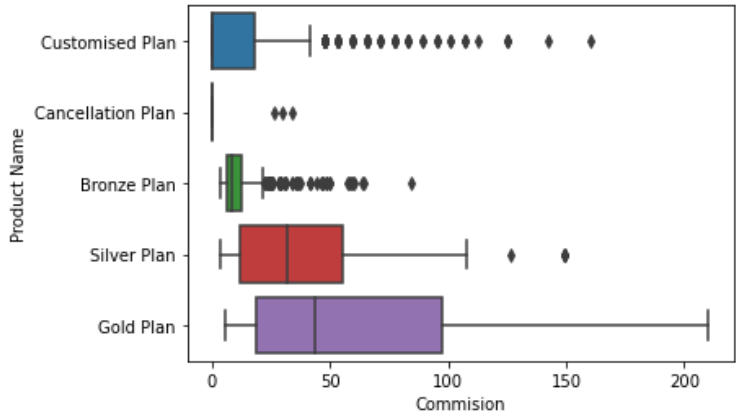
```
Plot of Commission vs Channel
-----------------------------------------------------
```



- From the above plot, it is seen that the mean commission for channel Online is highest.

```
Mean of Commission for Product Name
Product Name
Bronze Plan          11.322938
Cancellation Plan     0.132743
Customised Plan      11.654463
Gold Plan            67.195596
Silver Plan          36.472857
Name: Commision, dtype: float64
```

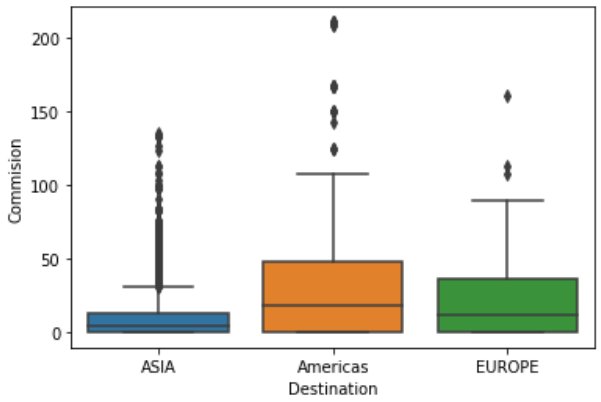
```
Plot of Commission vs Product
------------------------------------------------------------
```



- From the above plot, it is seen that the mean commission for the product Gold Plan is the highest and that for the Cancellation Plan is the lowest.

```
Mean of Commission for Destination
Destination
ASIA         11.732207
Americas     32.339906
EUROPE       20.088140
Name: Commision, dtype: float64
```

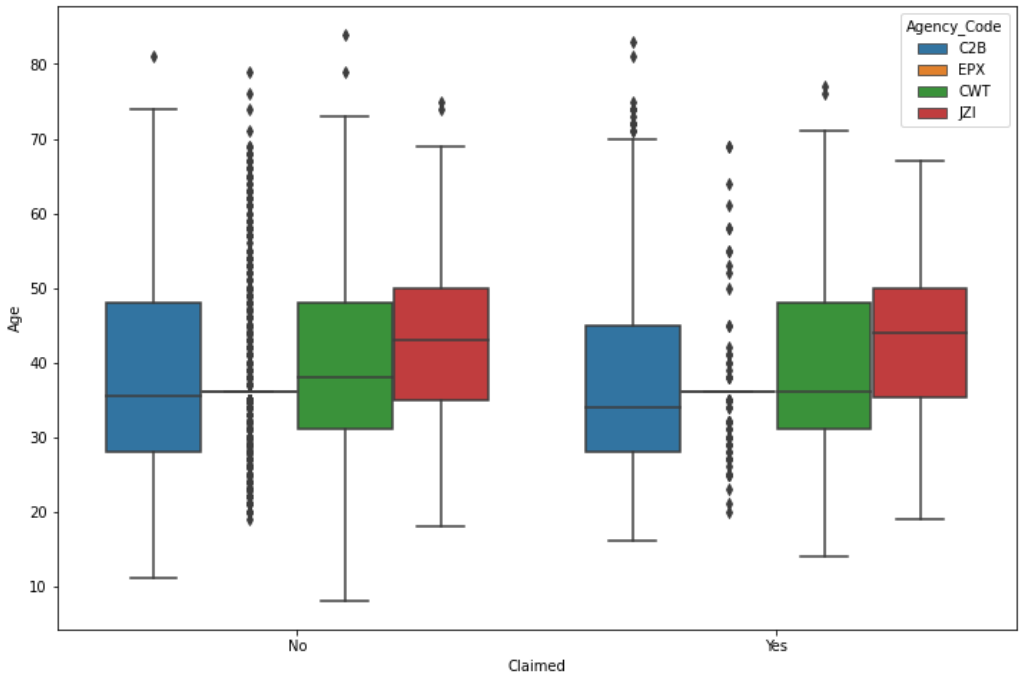
```
Plot of Commission vs Destination
------------------------------------------------------
```



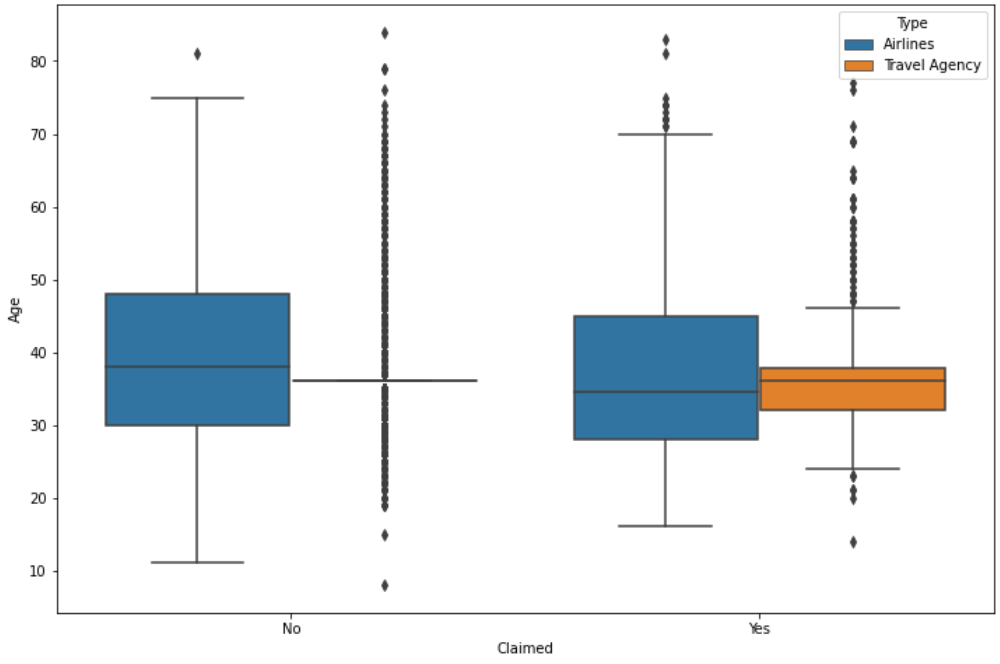
- From the above plot, it is seen that the mean commission for destination, Americas is the highest and that for the Asia is the lowest.

Data Visualization- Multivariate Analysis

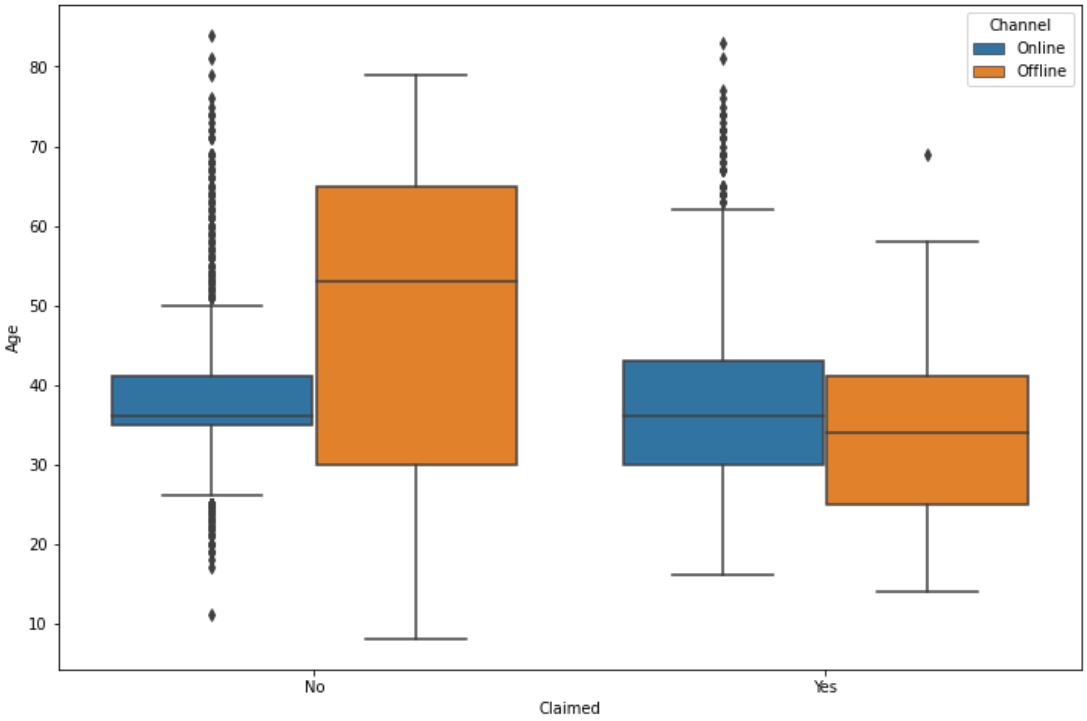
Insights from the Data



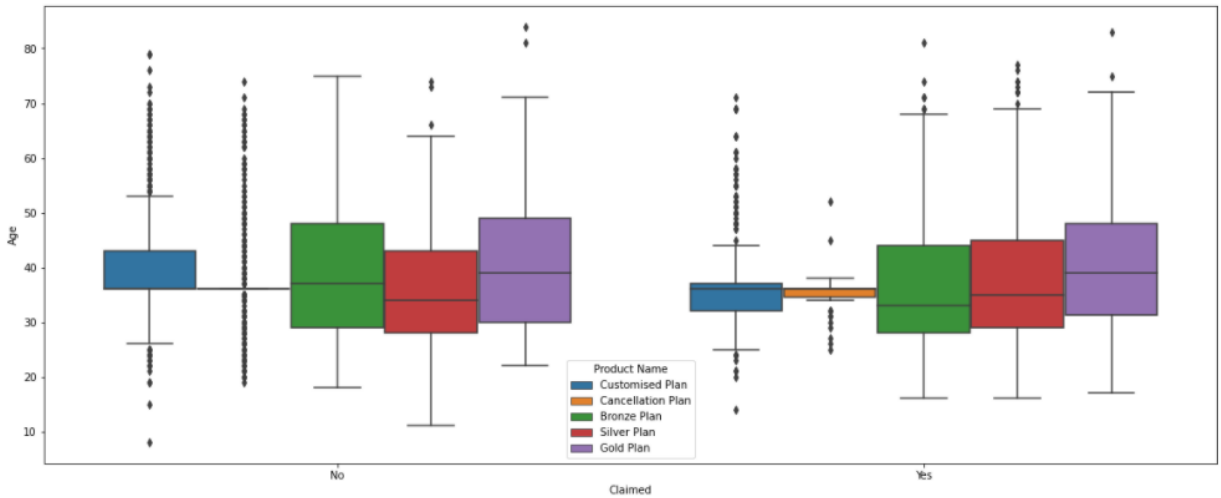
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median age for the agency JZI is the highest and that for C2B is the least.



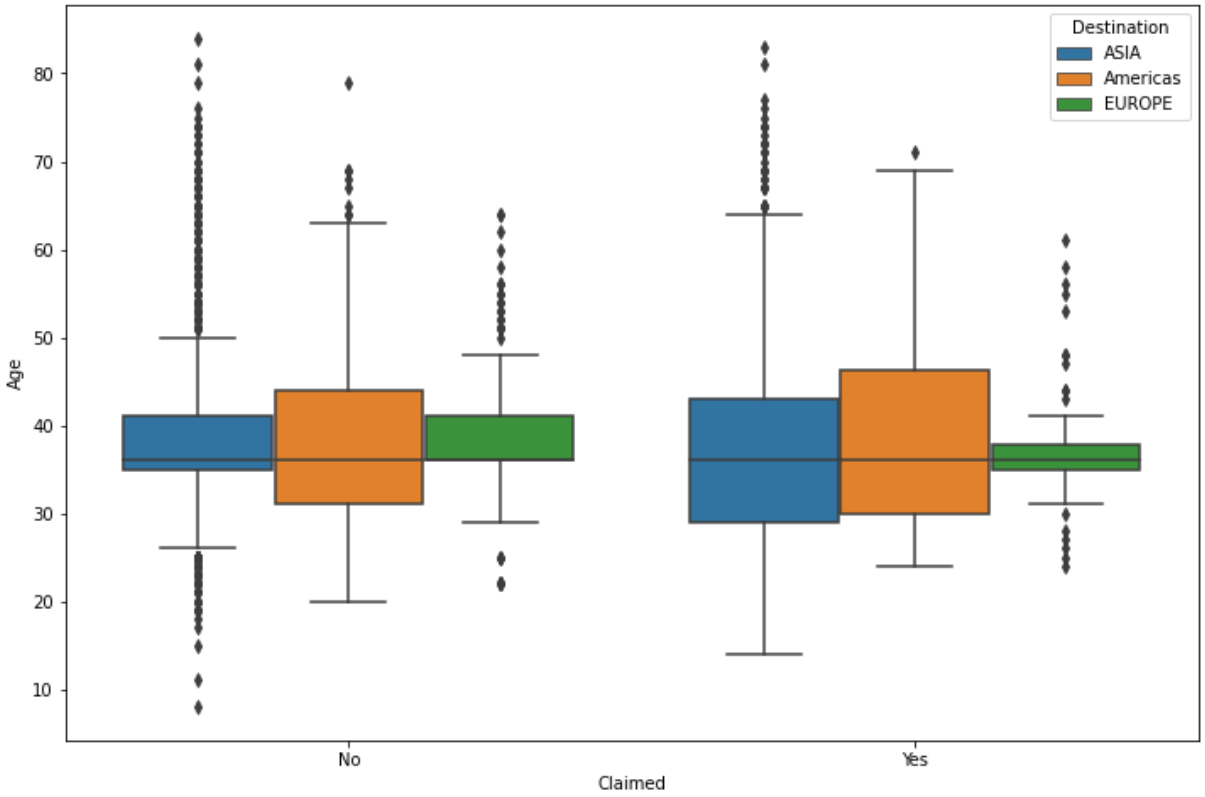
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median age for tour type, Travel Agency is maximum.



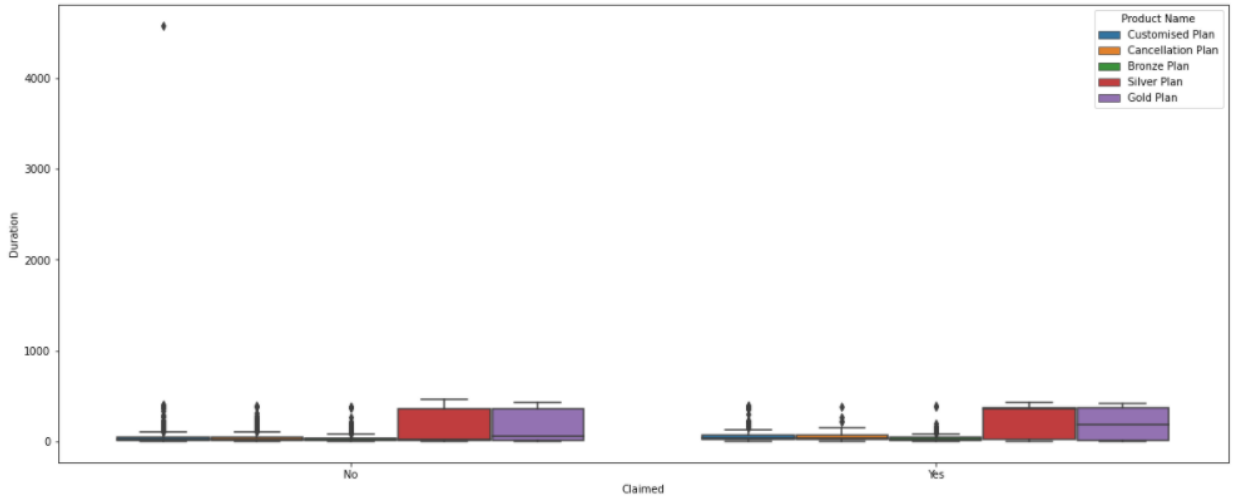
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median age for Online channel is maximum.



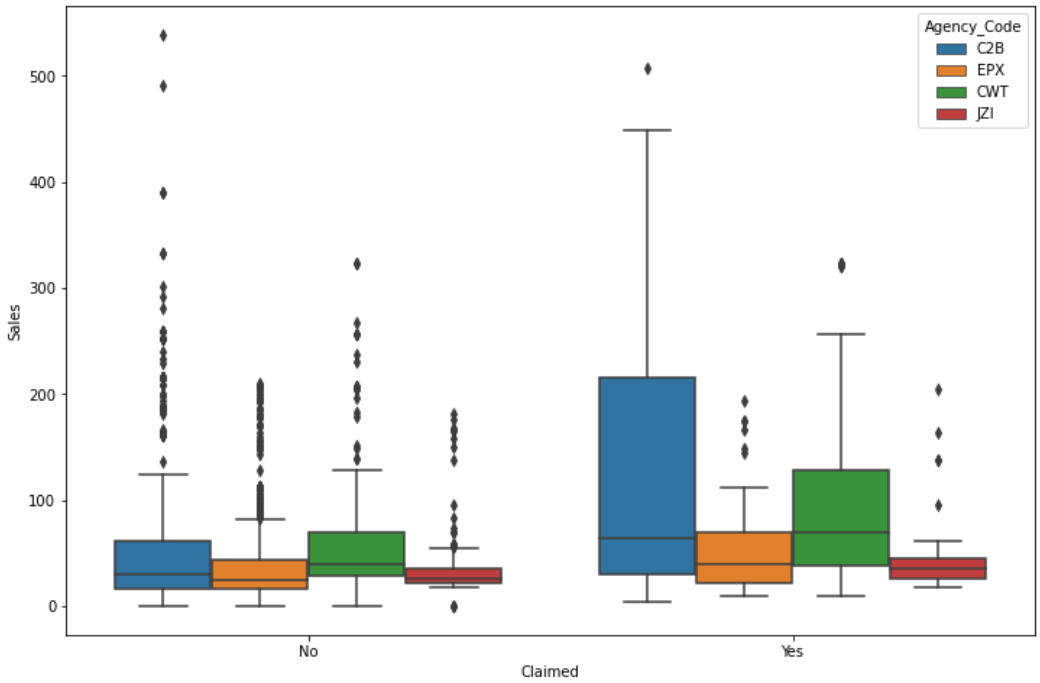
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median age for the product, Gold Plan is the maximum and that for Bronze plan is the minimum.



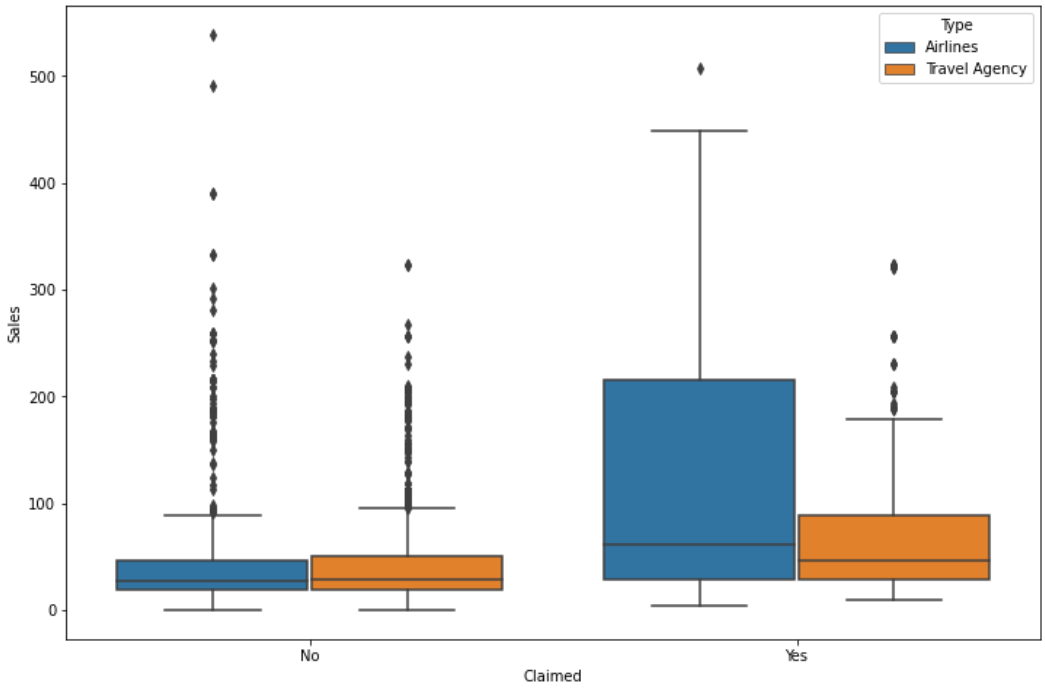
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median age for all the three destinations is almost the same.



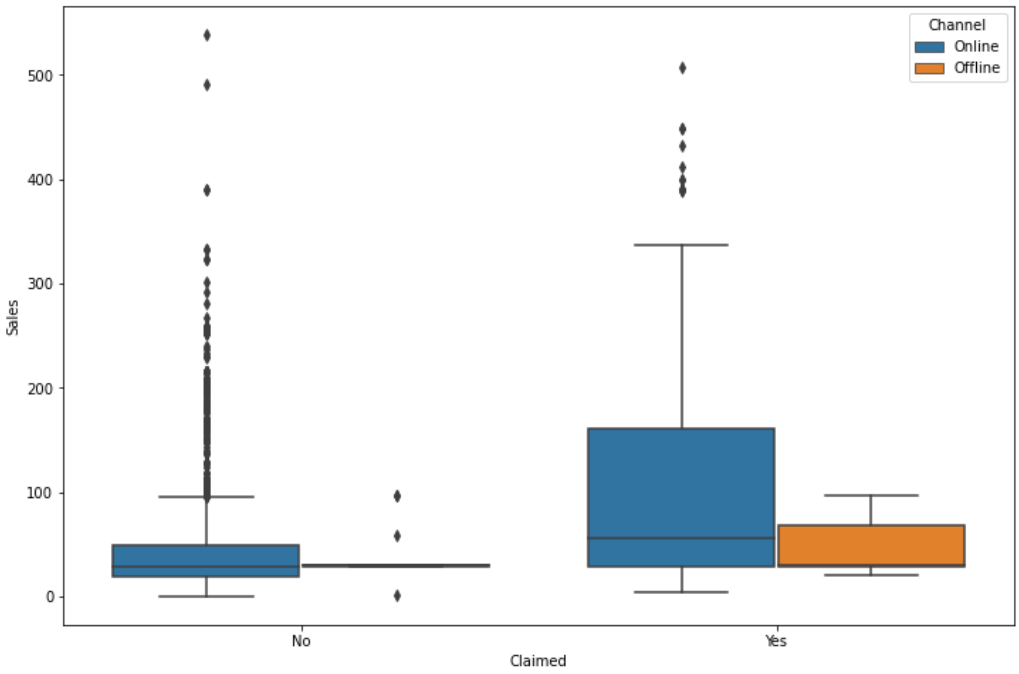
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median duration for the product, Silver Plan is the highest.



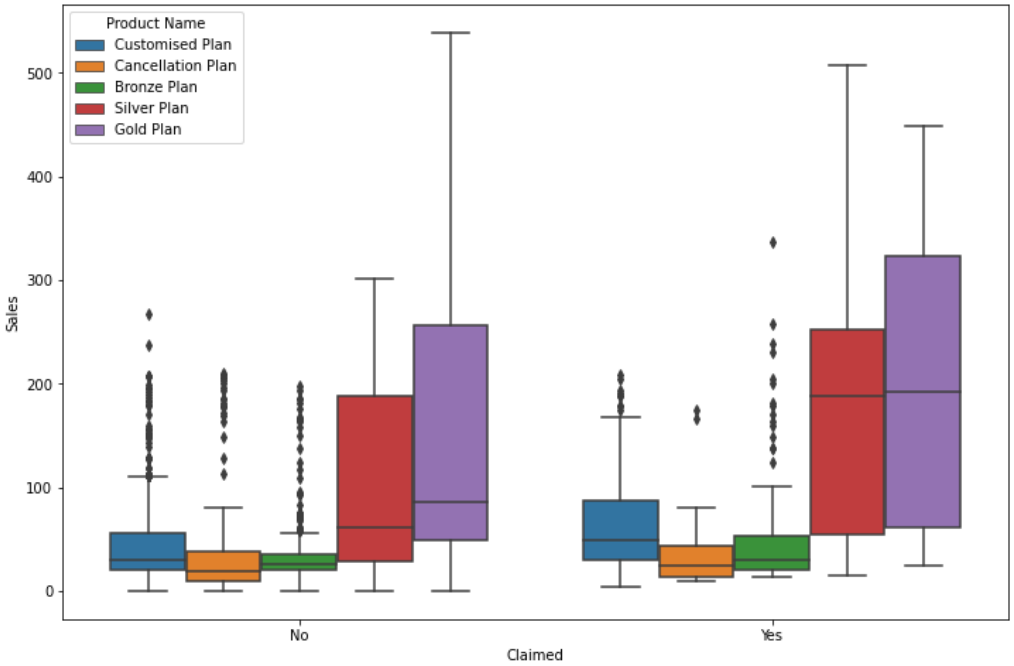
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Sales is the highest for agency CWT and it is the lowest for JZI.



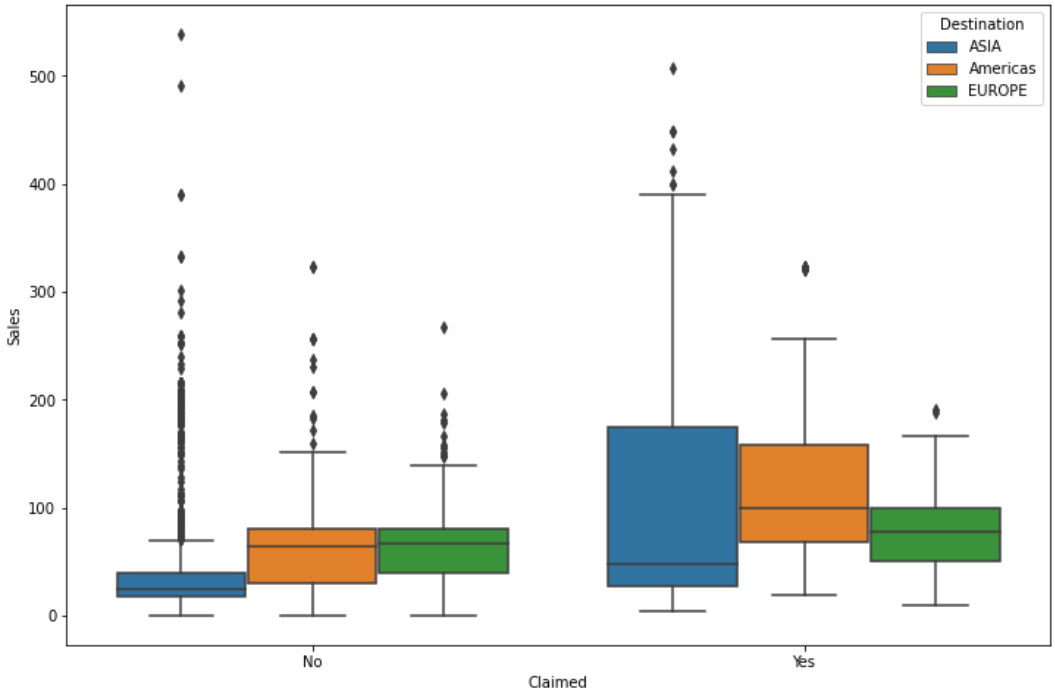
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Sales is the highest for the tour type, Airlines.



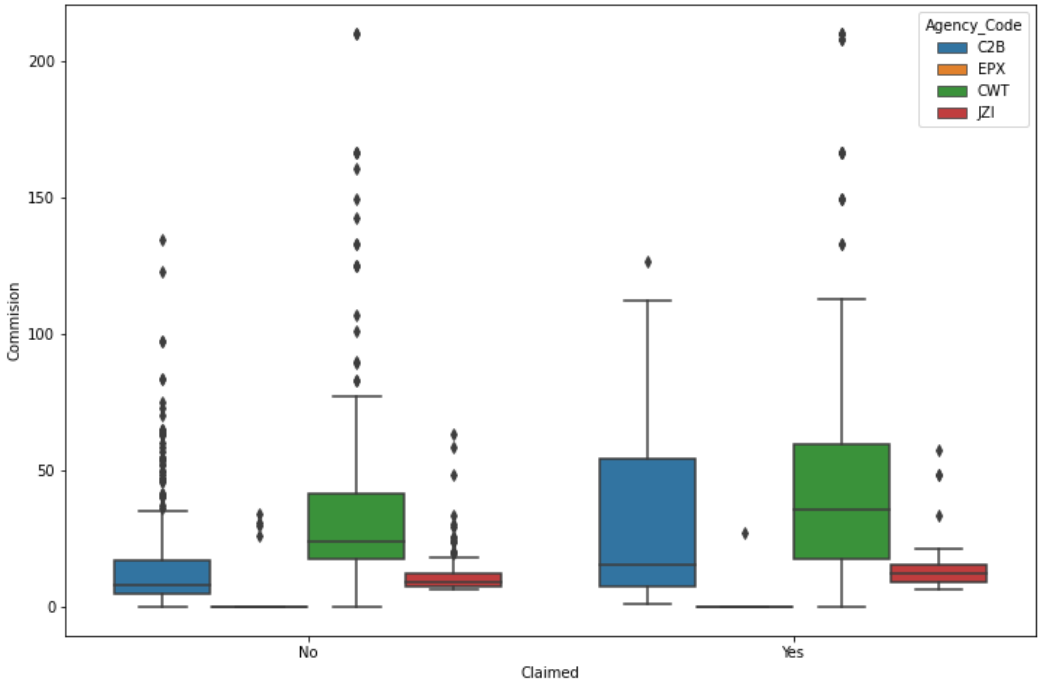
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Sales is the highest for Online Channel.



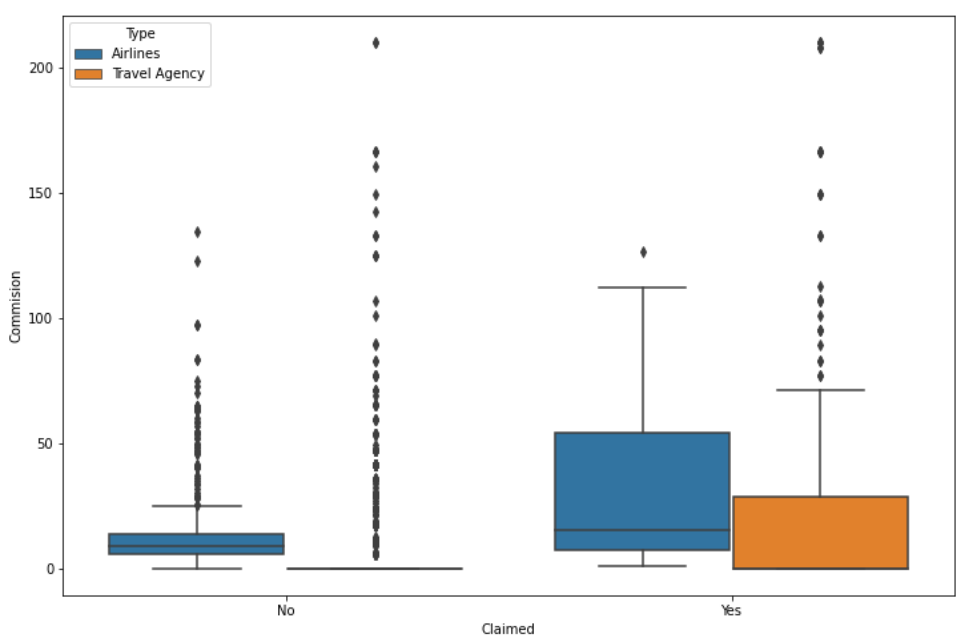
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Sales is the highest for the product Gold Plan and is the least for Cancellation plan.



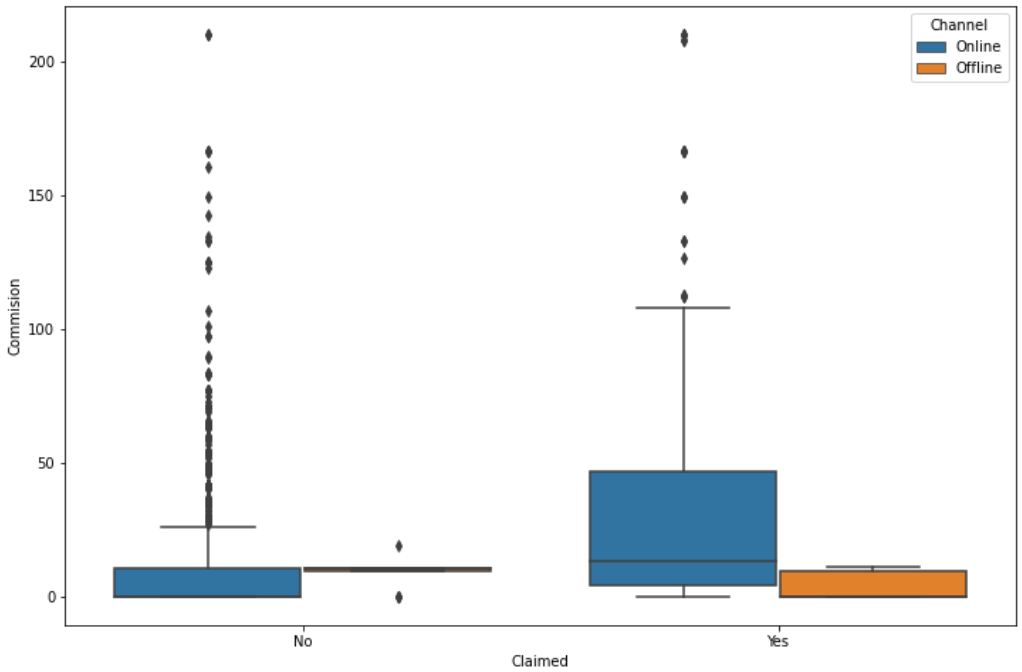
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Sales is the highest for destination Americas and the least for ASIA.



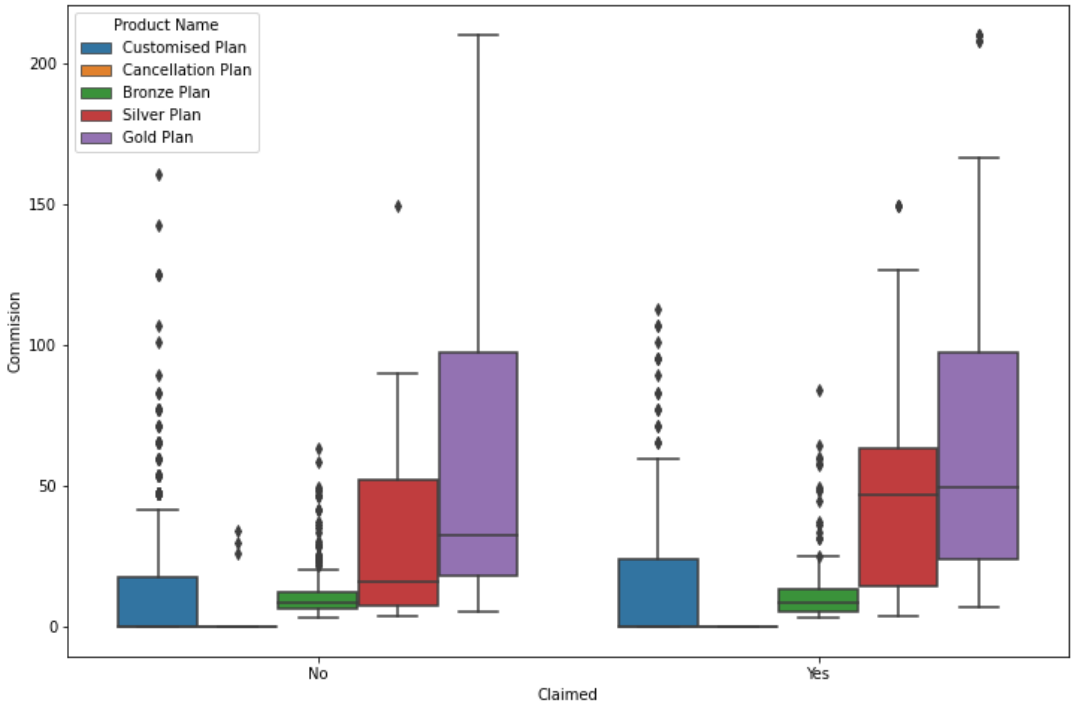
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Commission received is the highest for agency CWT and the least for EPX.



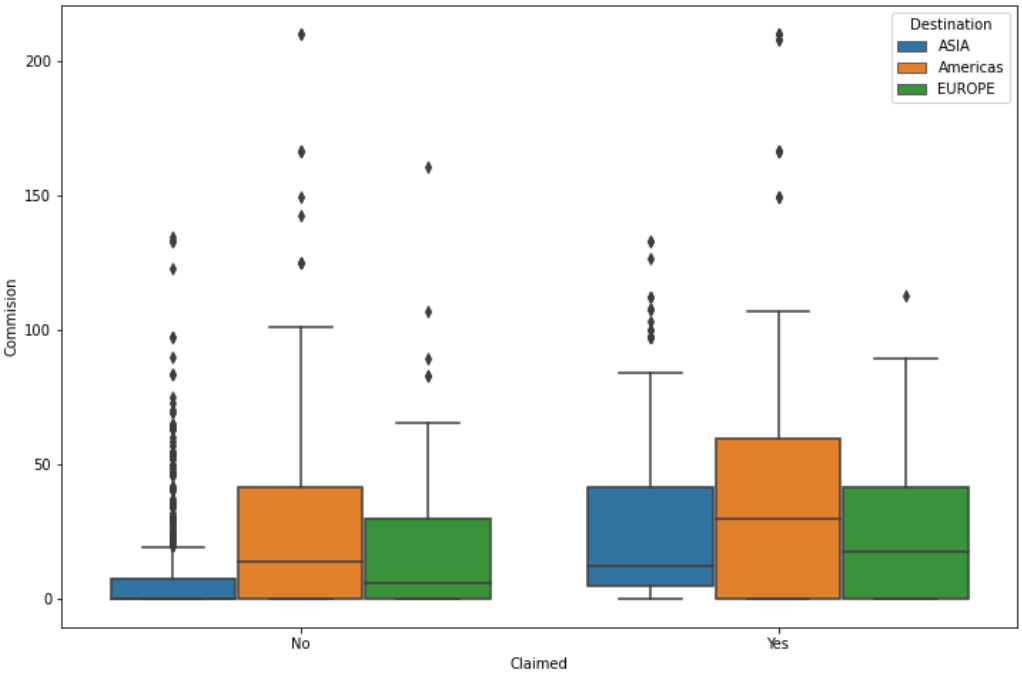
- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Commission received is the highest for tour type Airlines.



- From the above plot, it is seen that among the claims with claimed status ‘Yes’, median Commission received is the highest for the Online Channel.



- From the above plot, it is seen that among the claims with claimed status 'Yes', median Commission received is the highest for the product Gold Plan and the lower ones are the products Customised Plan and Cancellation Plan.



From the above plot, it is seen that among the claims with claimed status 'Yes', median Commission received is the highest for destination Americas and least for ASIA.

2.2 Data Split: Split the data into test and train(1 pts), build classification model CART (1.5 pts), Random Forest (1.5 pts), Artificial Neural Network(1.5 pts). Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed. Any reasonable split is acceptable. Use of random state is mandatory. Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Apply grid search for each model and make models on best_params. Feature importance for each model.

Solution:

CART Model

- Before applying the CART model, the object data has been converted to categorical data.
- The data set has been split into train set and test set in the ratio 70:30. This is the usual split which is assumed unless otherwise stated or required by the business and the client.
- The following gives the shape of the training data after the split.

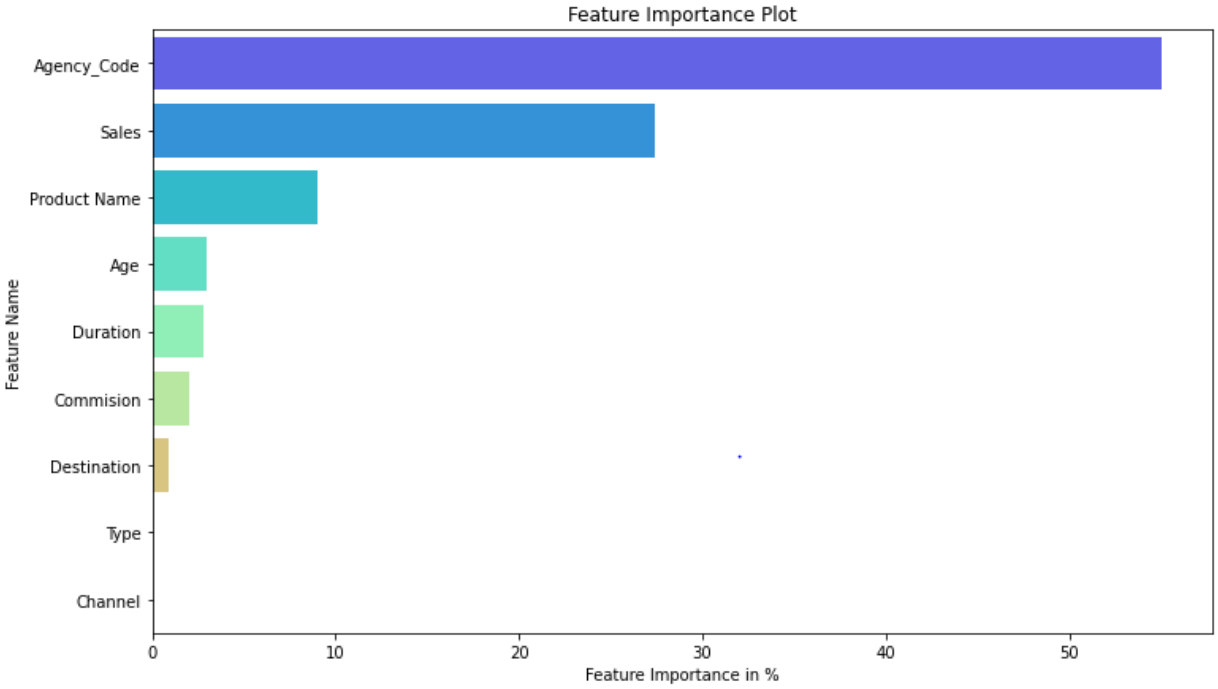
```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
Total observations is 3000
```

- The decision tree classifier was run with the criterion ‘gini’.
- Before applying the hyper parameter tuning and the grid search, the tree was fully grown and had approximately a depth > 20.
- Since pruning was required, the best parameters were found using the parameter grid.
- The following is the best parameters which were arrived at.

```
{'max_depth': 7, 'min_samples_leaf': 15, 'min_samples_split': 75}
```

- Random state has been included so as to ensure that the results are uniform.
- Feature importance for the model also has been arrived at as shown from below graph.

	Imp
Agency_Code	0.550452
Sales	0.274005
Product Name	0.089727
Age	0.029377
Duration	0.027699
Commision	0.019989
Destination	0.008751
Type	0.000000
Channel	0.000000



- From the above graph and table, the three important features for the model are Agency_Code, Sales and Product Name.

Random Forest Model

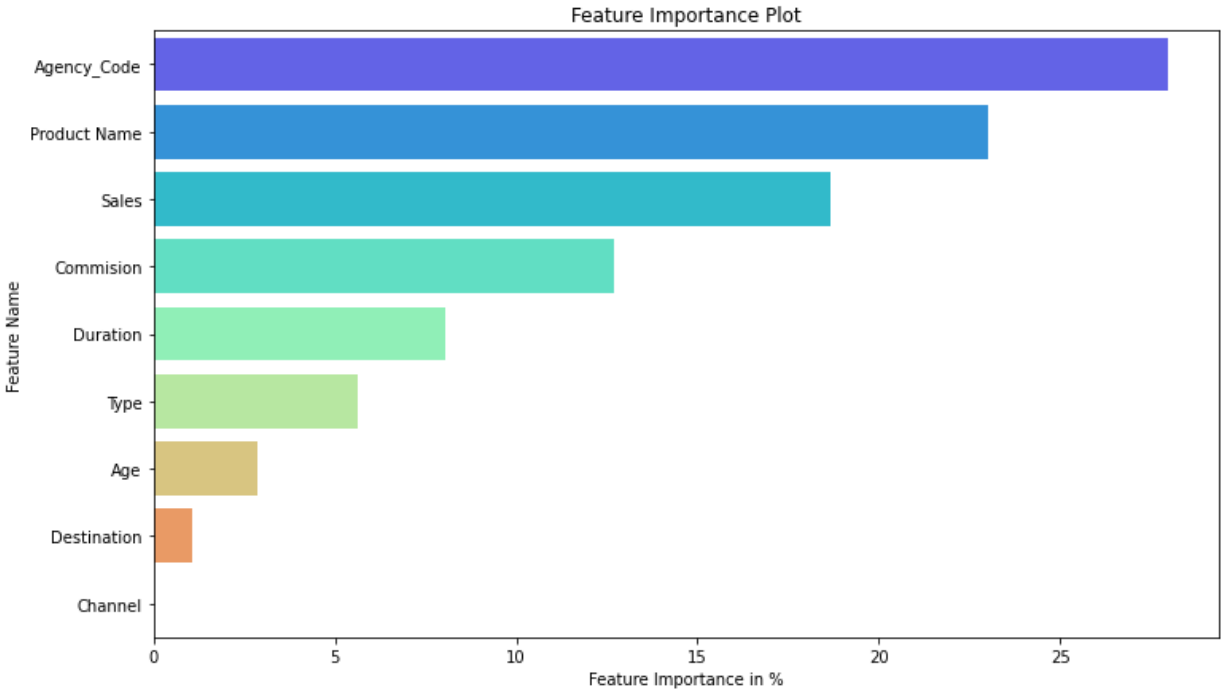
- The Random forest classifier algorithm was run using the parameter grid and grid search.
- The following was the best parameters which was arrived at.

```
{'max_depth': 8,
 'min_samples_leaf': 25,
 'min_samples_split': 75,
 'n_estimators': 301}
```

- Random state has been included so as to ensure that the results are uniform.

- Feature importance for the model also has been arrived at as shown from below graph.

	Imp
Agency_Code	0.279781
Product Name	0.230239
Sales	0.186664
Commision	0.127123
Duration	0.080287
Type	0.056357
Age	0.028688
Destination	0.010862
Channel	0.000000



- From the above graph and table, the three important features for the model are Agency_Code, Sales and Product Name.

Artificial Neural Network model

- The data was scaled using the standard scaler function.
- MLP classifier was applied to the data set.
- Random state has been used to ensure consistency in the results.
- After applying parameter grid and grid search, the following were the best parameters which were arrived at.

```
{'activation': 'relu',
 'hidden_layer_sizes': (100, 100),
 'max_iter': 500,
 'solver': 'sgd',
 'tol': 0.001}
```

- The activation function identified was the rectified linear unit with 2 hidden layers.
- The solver used was sgd- stochastic gradient descent with a tolerance of 0.001and the number of maximum iterations specified was 500.

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy (1 pts), Confusion Matrix (2 pts), Plot ROC curve and get ROC_AUC score for each model (2 pts), Make classification reports for each model. Write inferences on each model (2 pts). Calculate Train and Test Accuracies for each model. Comment on the validness of models (overfitting or underfitting) Build confusion matrix for each model. Comment on the positive class in hand. Must clearly show obs/pred in row/col Plot roc_curve for each model. Calculate roc_auc_score for each model. Comment on the above calculated scores and plots. Build classification reports for each model. Comment on f1 score, precision and recall, which one is important here.

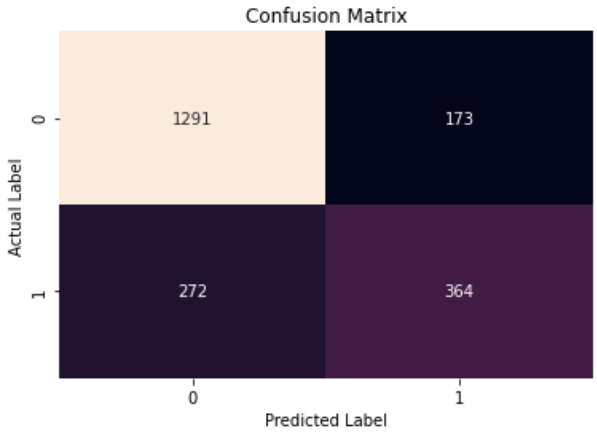
Solution:

Performance metrics for the models:

CART Model:

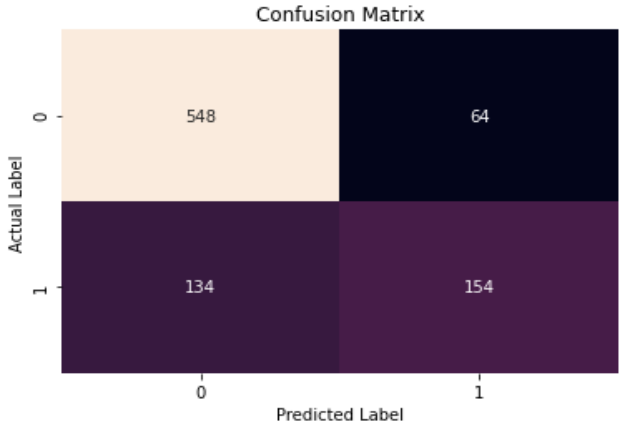
- The accuracy of the train set was found to be 0.79 and that of the test set was found to be 0.78.

- The following gives the confusion matrix of the train set:




```
True Negatives:  1291
False Positives:  173
False Negatives:  272
True Positives:  364
```

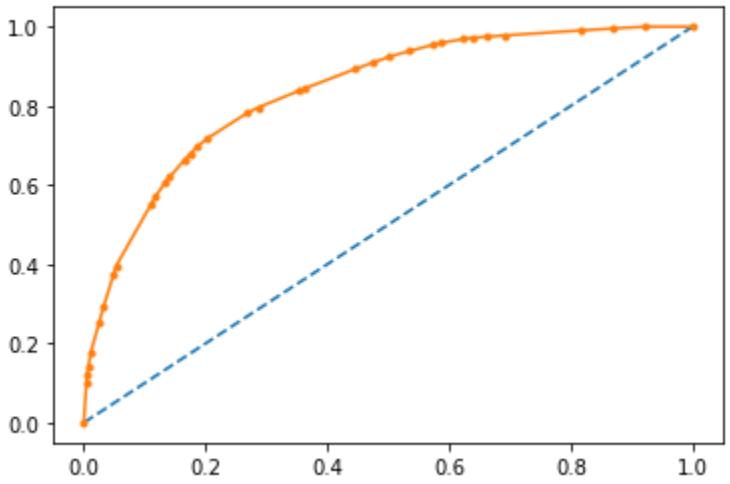
- The following gives the confusion matrix of the test set:




```
True Negatives:  548
False Positives:  64
False Negatives:  134
True Positives:  154
```

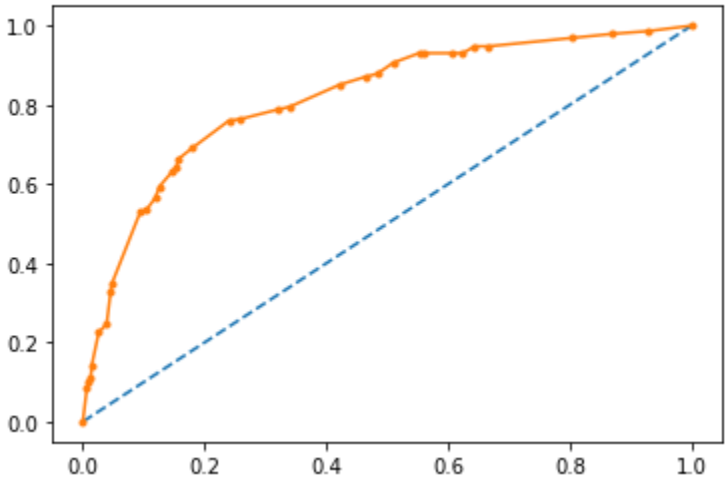
- The following is the ROC curve and ROC_AUC score of the train set:

```
AUC: 0.837
```



- The following is the ROC curve and ROC_AUC score of the test set:

```
AUC: 0.817
```



- The following is the classification report for the train set:

	precision	recall	f1-score	support
0	0.83	0.88	0.85	1464
1	0.68	0.57	0.62	636
accuracy			0.79	2100
macro avg	0.75	0.73	0.74	2100
weighted avg	0.78	0.79	0.78	2100

- The following is the classification report for the test set:

	precision	recall	f1-score	support
0	0.80	0.90	0.85	612
1	0.71	0.53	0.61	288
accuracy			0.78	900
macro avg	0.75	0.72	0.73	900
weighted avg	0.77	0.78	0.77	900

Inferences:

AUC on the training data is 83.7% and on test data is 81.7% indicating a relatively good performance model with respect to the claim prediction. The accuracy, precision and recall metrics are also almost similar between training and test set, which indicates no overfitting or underfitting has happened.

The precision of the test set is 0.71 which indicates that only 29 claims out of 100 claims are predicted wrongly as claimed status, ‘Yes’ and the remaining 71/100 are predicted correctly.

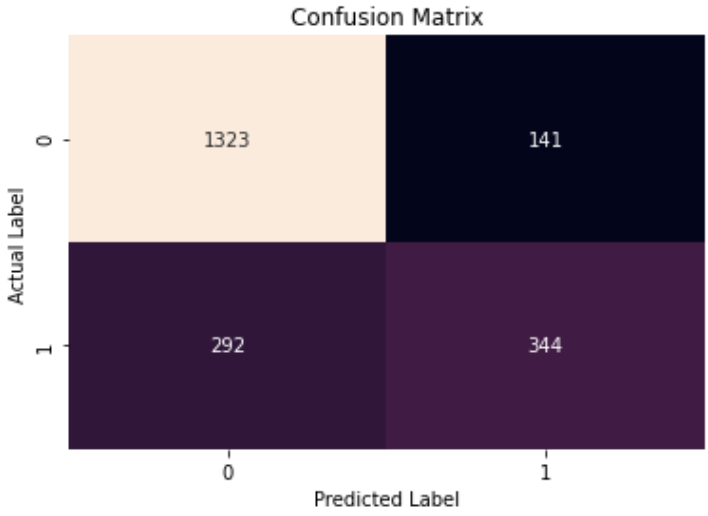
As for the business implication, the precision of the model may play a larger role in deciding the performance of the model since the insurance company should not be paying up for claims which have been wrongly classified as false positive. There are a total of 64 False Positives as can be seen from the confusion matrix.

Regarding the recall, it is 0.53 which indicates 47/100 claims with claimed status “Yes’ have been wrongly classified as No and their payments would be denied by the insurance company. This may cause a loss in reputation to the company. There are 134 false negatives as seen from the confusion matrix.

The Overall model performance is moderate and may be used to start predicting the claim status for any new claim submitted to the insurance company.

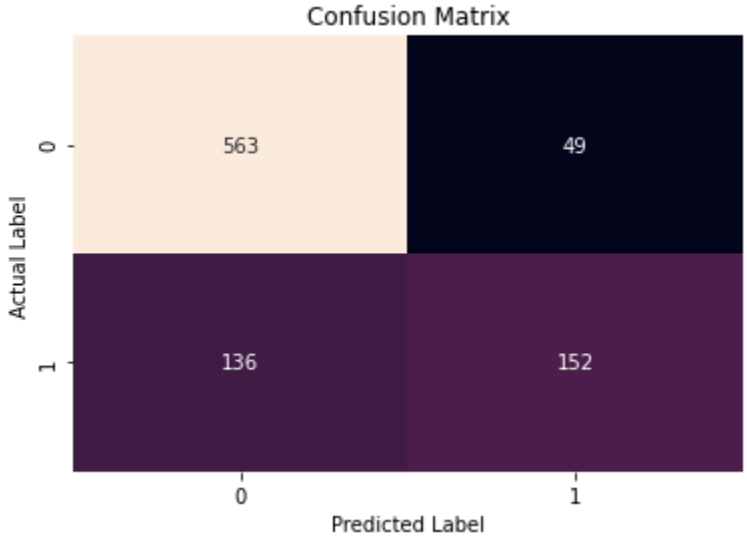
Random Forest Model:

- The accuracy of both the train set and the test set was found to be 0.79.
- The following gives the confusion matrix of the train set:




```
True Negatives:  1323
False Positives:  141
False Negatives:  292
True Positives:  344
```

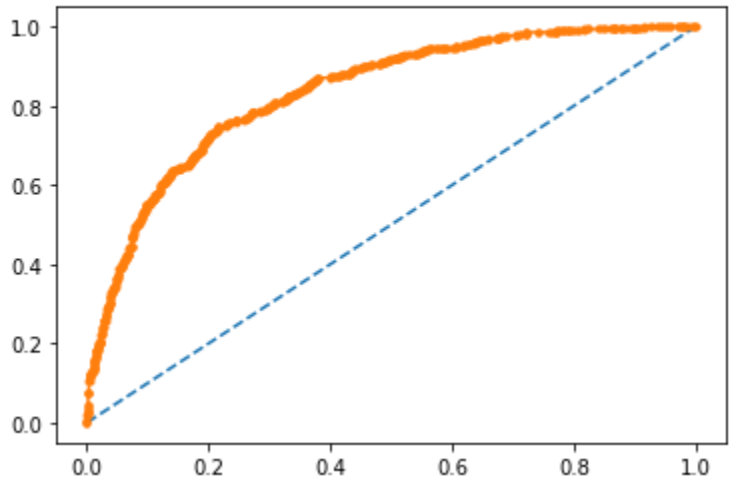
- The following gives the confusion matrix of the test set:



```
True Negatives:  563
False Positives:  49
False Negatives:  136
True Positives:  152
```

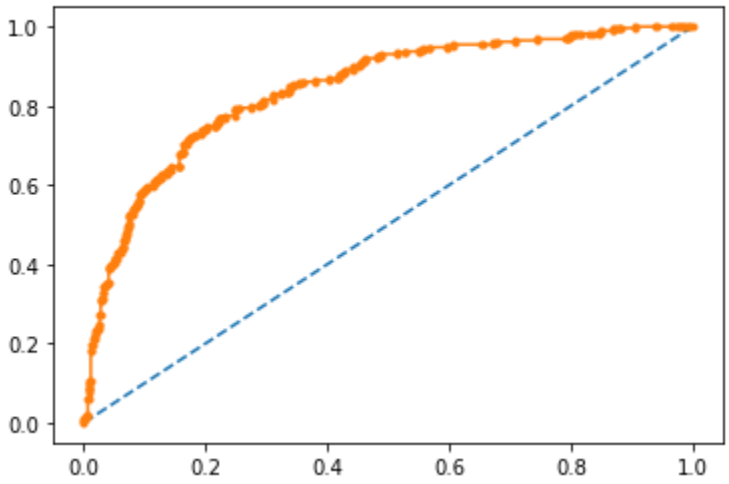
- The following is the ROC curve and ROC_AUC score of the train set:

```
AUC: 0.836
```



- The following is the ROC curve and ROC_AUC score of the test set:

```
AUC: 0.842
```



- The following is the classification report for the train set:

	precision	recall	f1-score	support
0	0.82	0.90	0.86	1464
1	0.71	0.54	0.61	636
accuracy			0.79	2100
macro avg	0.76	0.72	0.74	2100
weighted avg	0.79	0.79	0.78	2100

- The following is the classification report for the test set:

	precision	recall	f1-score	support
0	0.81	0.92	0.86	612
1	0.76	0.53	0.62	288
accuracy			0.79	900
macro avg	0.78	0.72	0.74	900
weighted avg	0.79	0.79	0.78	900

Inferences:

AUC on the training data is 83.6% and on test data is 84.2% indicating a relatively good performance model with respect to the claim prediction. The accuracy, precision and recall metrics are also almost similar between training and test set, which indicates no overfitting or underfitting has happened.

The precision of the test set is 0.76 (class 1 i.e. Claimed Status “Yes’) which indicates that only 24 claims out of 100 claims are predicted wrongly as claimed status, ‘Yes’ and the remaining 76/100 are predicted correctly.

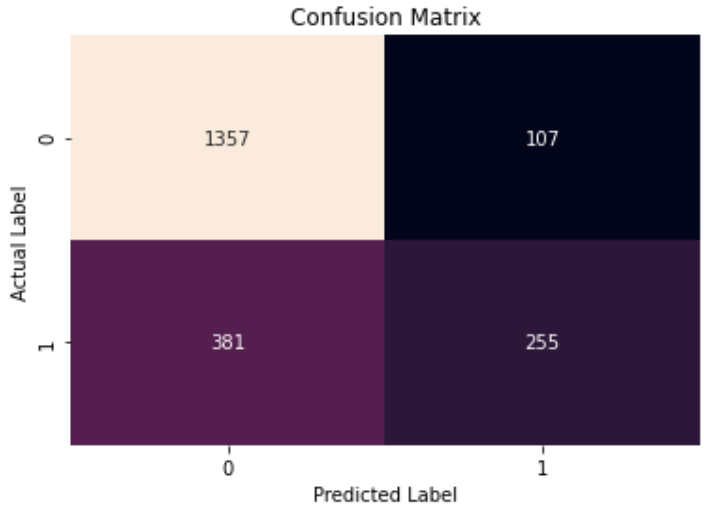
As for the business implication, the precision of the model may play a larger role in deciding the performance of the model since the insurance company should not be paying up for claims which have been wrongly classified as false positives. There are 49 false positives as seen from the confusion matrix.

Regarding the recall, it is 0.53 which indicates 47/100 claims with claimed status “Yes’ have been wrongly classified as No and their payments would be denied by the insurance company. This may cause a loss in reputation to the company. There are 136 false negatives as seen from the confusion matrix.

The Overall model performance is moderate and may be used to start predicting the claim status for any new claim submitted to the insurance company.

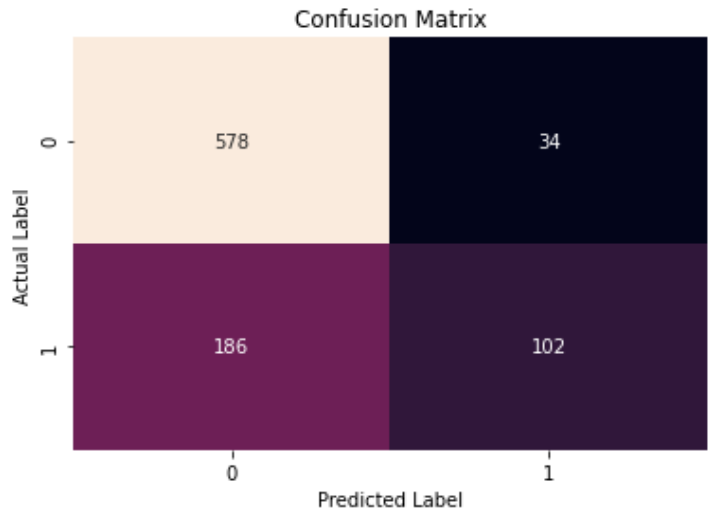
Artificial Neural Network Model:

- The accuracy of the train set was found to be 0.77 and that of the test set was found to be 0.76.
- The following gives the confusion matrix of the train set:




```
True Negatives:  1357
False Positives:  107
False Negatives:  381
True Positives:  255
```

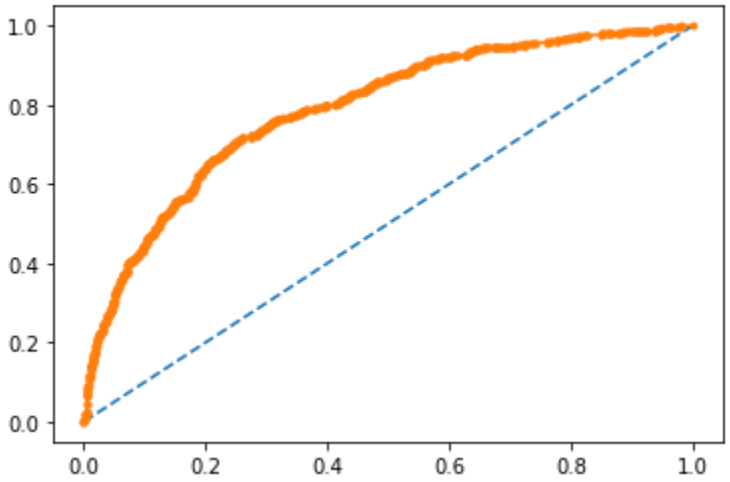
- The following gives the confusion matrix of the test set:



```
True Negatives:  578
False Positives:  34
False Negatives:  186
True Positives:  102
```

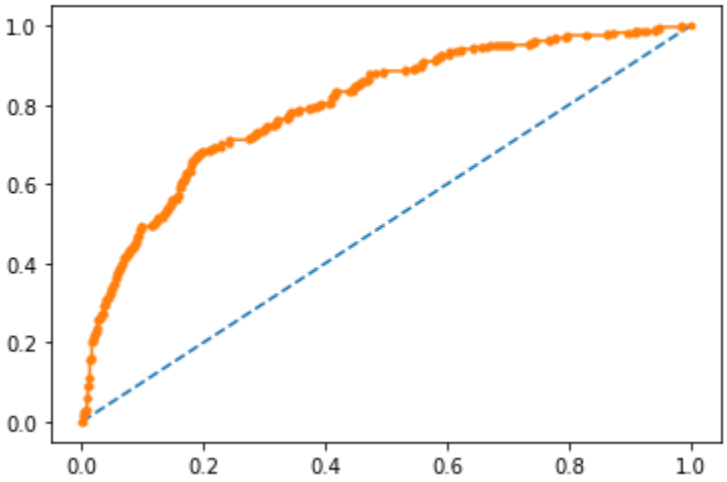
- The following is the ROC curve and ROC_AUC score of the train set:

```
AUC: 0.790
```



- The following is the ROC curve and ROC_AUC score of the test set:

```
AUC: 0.799
```



- The following is the classification report for the train set:

	precision	recall	f1-score	support
0	0.78	0.93	0.85	1464
1	0.70	0.40	0.51	636
accuracy			0.77	2100
macro avg	0.74	0.66	0.68	2100
weighted avg	0.76	0.77	0.75	2100

- The following is the classification report for the test set:

	precision	recall	f1-score	support
0	0.76	0.94	0.84	612
1	0.75	0.35	0.48	288
accuracy			0.76	900
macro avg	0.75	0.65	0.66	900
weighted avg	0.75	0.76	0.73	900

Inferences:

AUC on the training data is 79% and on test data is 79.9% indicating a relatively moderate performance model. The accuracy, precision and recall metrics are also almost similar between training and test set, which indicates no overfitting or underfitting has happened.

The precision of the test set is 0.75 (on class 1 i.e. Claimed Status “Yes’) which indicates that only 25 claims out of 100 claims are predicted wrongly as claimed status, ‘Yes’ and the remaining 75/100 are predicted correctly.

As for the business implication, the precision of the model may play a larger role in deciding the performance of the model since the insurance company should not be paying up for claims which have been wrongly classified as false positives. There are 34 false positives as seen from the confusion matrix.

Regarding the recall, it is quite low, 0.35 which indicates 65/100 claims with claimed status “Yes’ have been wrongly classified as No and their payments would be denied by the insurance company. This may cause a loss in reputation to the company. There are 186 false negatives seen from the confusion matrix.

The Overall model performance is moderate and should be used only after weighing the pros and cons to the business.

2.4 Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner (2.5 pts). Describe on which model is best/optimized (1.5 pts). A table containing all the values of accuracies, precision, recall, auc_roc_score, f1 score. Comparison between the different models(final) on the basis of above table values. After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.

Solution:

The following gives the table of performance metrics of the three models.

	Accuracy		**Precision**		**Recall**		**ROC_AUC Score**		**F1 Score**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
CART	0.79	0.78	0.68	0.71	0.57	0.53	0.837	0.817	0.62	0.61
Random Forest	0.79	0.79	0.71	0.76	0.54	0.53	0.836	0.842	0.61	0.62
Artificial Neural Network	0.77	0.76	0.70	0.75	0.40	0.35	0.79	0.799	0.51	0.48

- The accuracy scores of all the 3 models are quite similar.
- The precision of all the 3 models are also quite close in values.
- The recall of CART and Random forest models are the same but artificial neural network model has quite a lesser value of recall when compared to the other two models.
- The AUC scores are also quite close and almost comparable for all the 3 models.
- Due to the significant differences in the recall scores, the F1 score of artificial neural network model is quite less when compared to the other two models.

Final Model Selection:

- Since, the recall is significantly less than the other two models, we do not choose ANN model as we need to ensure both the precision and the recall are good.

- Among the CART model and the Random Forest model, the recall is the same for both the models but the precision is slightly better for the random forest model than that of the CART model.
- Also, the ROC_AUC score is again slightly better for the random forest model.

Conclusion:

- Taking into view the above points, the random forest model suits the best for the problem at hand.
- As for the business implication, the precision of the model plays an important role in choosing the model since, the insurance company should not be paying up for claims which have been wrongly classified (false positives).
- Recall also plays quite an important factor to the business since genuine claims (false negatives) would be denied payment by the insurance company. This may cause a loss in reputation to the company.

2.5 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

Solution:

Few insights from the data set:

- Mean sales for the product, 'Cancellation Plan' is the least.
- Mean sales at agency JZI is the least.
- The agency C2B has the maximum claims with claim status 'Yes', 18.7%.
- Maximum claims received, 10.2% is for the product, Silver Plan.

Business Recommendations:

- It is seen that that maximum claims with claimed status 'yes' has been received from the agency C2B. Hence, the insurance firm should look at auditing and double checking the claims from the tour agency C2B.

- Also, the product silver plan seems to be the most frequent in the claims. The business needs to revisit this product and identify any problems/issues associated with this product.
- The business can look at improving the sales through suitable strategies for the product, ‘Cancellation Plan’ which has the lowest sales.
- The business can also look at improving the sales at the agency JZI, which has the lowest sales, through appropriate business plans.
- It is recommended that the business choose the best model (Random Forest Model-Please refer Q 2.4) which ensures that the insurance company does not end up paying up for claims which have been wrongly predicted as genuine.
- At the same time, the firm should not reject genuine claims (which is due to an erroneous prediction) and deny the payments of such claims so as to protect its reputation and choose the appropriate model (Random Forest Model-Please refer Q 2.4).