Project: DATA MINING

Name- Manisha Rout
PGP-DSBA Online
Mar’22

Date: 24.04.2022

greatlearning
Power Ahead

Table of Content

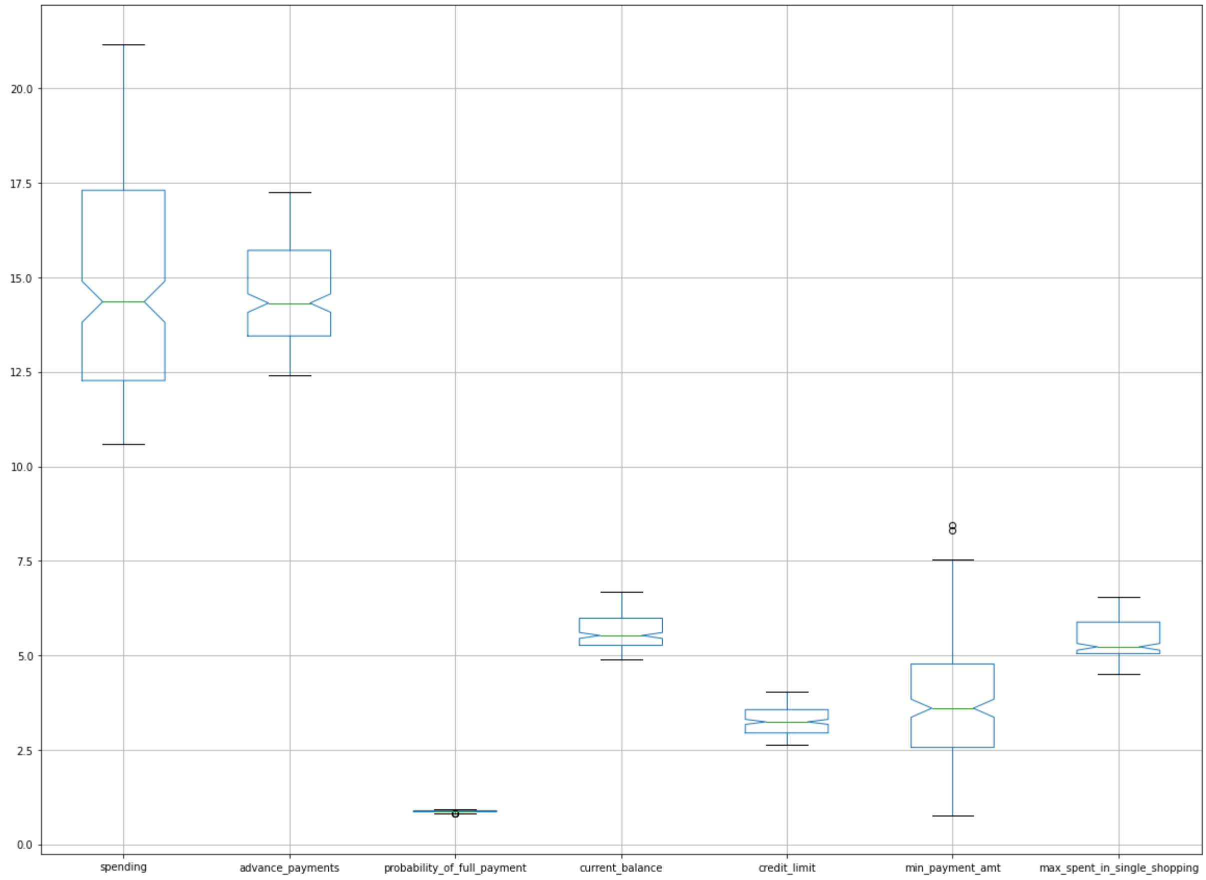
List of Figures

List of Tables

Problem 1.1

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

1. The given dataset has 210 rows and 7 columns. All the 7 features are numeric and float data type.
2. The dataset has no duplicate records and no missing values.
3. Outliers are present in two of the features: probability_of_full_payment and min_payment_amt which can be seen from the boxplot.
4. No anomalies present in the dataset
5. There is no bad data present in the dataset
6. Outlier treatment has not been done on the dataset since:
   - the outliers carry little weightage, they will not have any effective influence on clustering
   - Outliers of probability_of_full_payment features should not be treated; since, it is based on real life scenario and the probability always ranges from 0 to 1.



Univariate analysis

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

7. Amount spent by the customers per month ranges from 10.59 to 21.18
8. Average amount spent by the customer is 14.85

9. The mean is nearly equal to median however, the distribution is not normal which is evident from the boxplot and probability plot
10. Skewness of the spending attribution is 0.40 indicates a right tailed distribution
11. Outliers are not present for this attribution which is evident from the box plot

```
Description of spending
------------------------
count     210.000000
mean       14.847524
std         2.909699
min        10.590000
25%        12.270000
50%        14.355000
75%        17.305000
max        21.180000
```

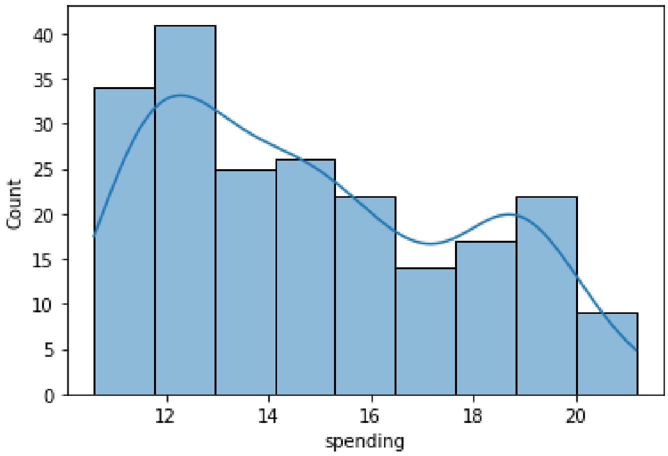


Figure 1: Histogram of Amount spent by the customer per month (in 1000s)

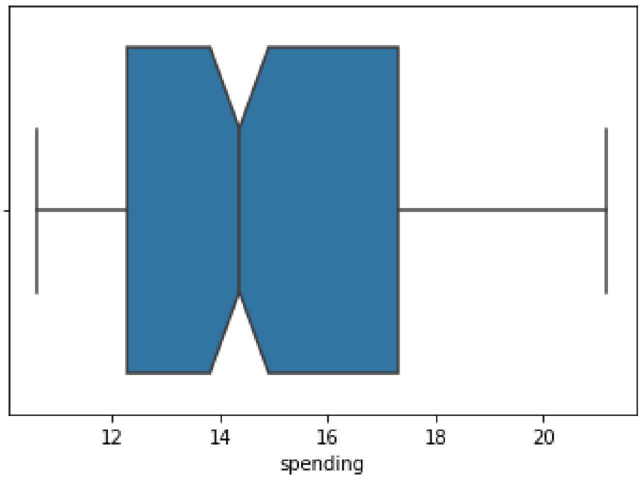


Figure 2 Boxplot of Amount spent by the customer per month (in 1000s)

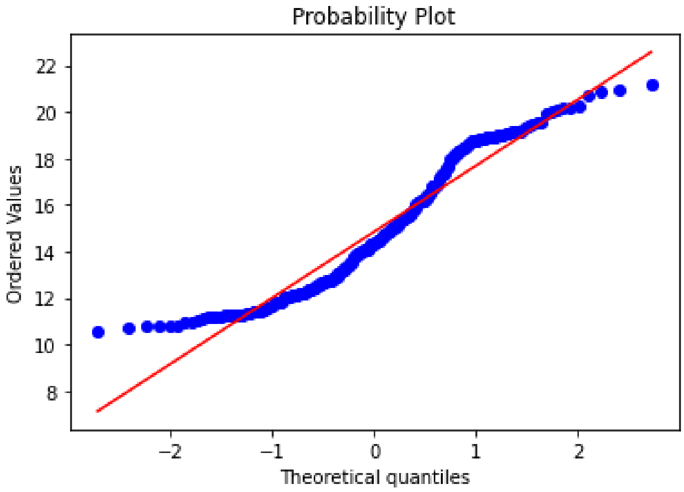


Figure 3 Probability plot Amount spent by the customer per month (in 1000s)

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

12. Amount paid by the customer in advance by case ranges from 12.41 to 17.25
13. Average amount paid by the customer in advance by cash is 14.56
14. The mean is not equal to median, the distribution is not normal which is evident from the boxplot and probability plot
15. Skewness of the Advance_payments attribution is 0.39 indicates a right tailed distribution
16. Outliers are not present for this attribution which is evident from the box plot

```
Description of advance_payments
------------------------------
count     210.000000
mean       14.559286
std         1.305959
min        12.410000
25%        13.450000
50%        14.320000
75%        15.715000
max        17.250000
```

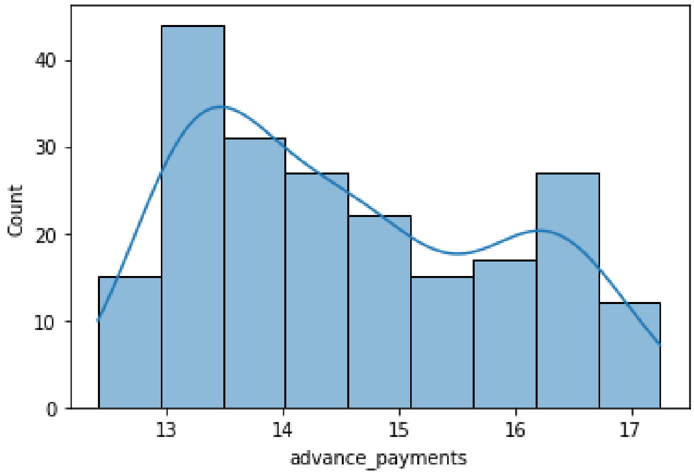


Figure 4: Histogram of Amount paid by the customer in advance by cash (in 100s)

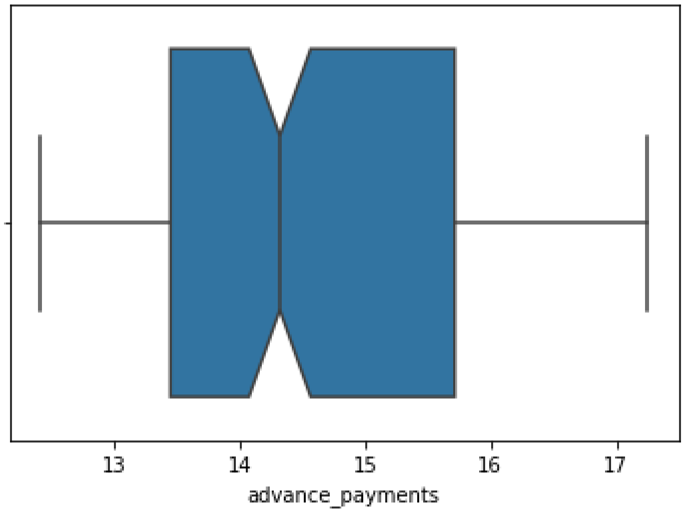


Figure 5 : Boxplot of Amount paid by the customer in advance by cash (in 100s)

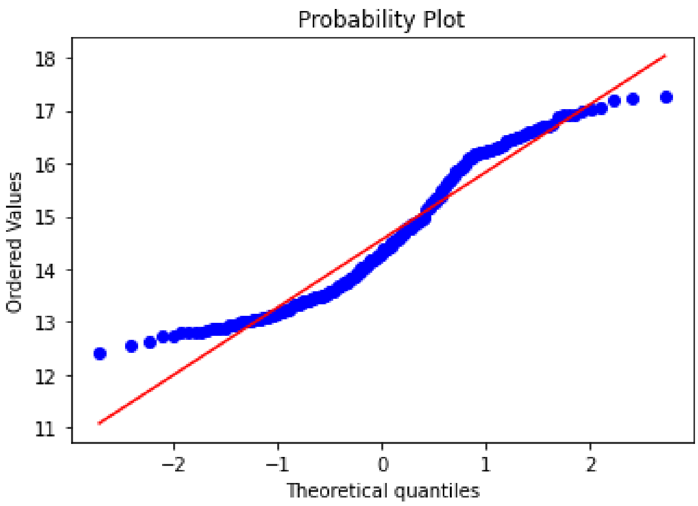


Figure 6 : Probability Plot of Amount paid by the customer in advance by cash (in 100s)

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

- Probability of payment done in full by the customer to the bank ranges from 0.80 to 0.92
- Average Probability of payment done in full by the customer to the bank is 0.87
- The mean is not equal to median, the distribution is not normal which is evident from the boxplot and probability plot
- Skewness of the probability_of_full_payment attribution is -0.53 indicates a left tailed distribution
- Outliers are present for this attribution which is evident from the box plot

```
Description of probability_of_full_payment
-------------------------------------------
count    210.000000
mean       0.870999
std        0.023629
min        0.808100
25%        0.856900
50%        0.873450
75%        0.887775
max        0.918300
```

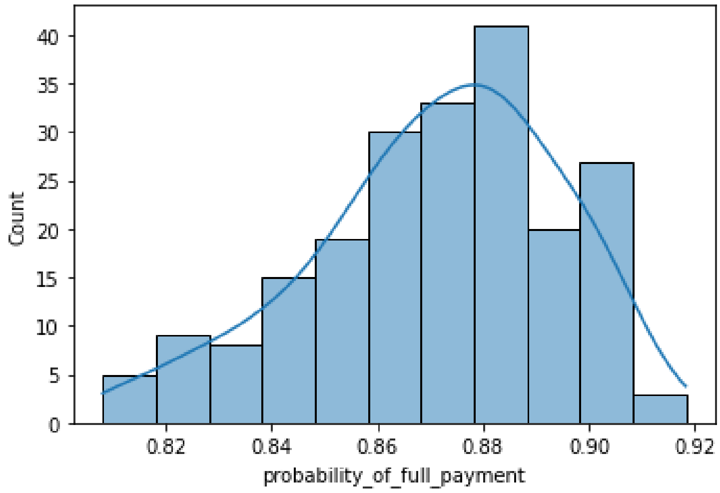


Figure 7: Histogram of probability_of_full_payment

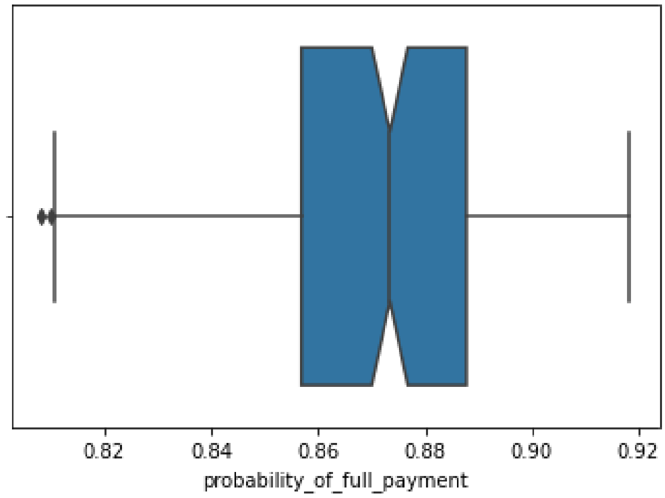


Figure 8: Boxplot of probability_of_full_payment

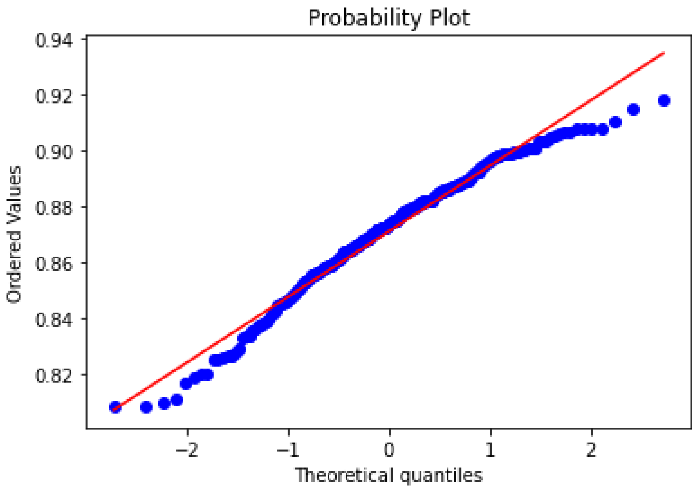


Figure 9: Probability Plot of Probability_of_full_payment

4. **current_balance: Balance amount left in the account to make purchases (in 1000s)**
   - Balance amount left in the account to make purchases ranges from 4.9 to 6.67
   - Average balance amount left in the account to make purchases is 5.62
   - The mean is nearly equal to median however, the distribution is not normal which is evident from the boxplot and probability plot
   - Skewness of the current_balance attribution is 0.53 indicates a right tailed distribution
   - Outliers are not present for this attribution which is evident from the box plot

```
Description ofcurrent_balance
------------------------
count    210.000000
mean       5.628533
std        0.443063
min        4.899000
25%        5.262250
50%        5.523500
75%        5.979750
max        6.675000
```

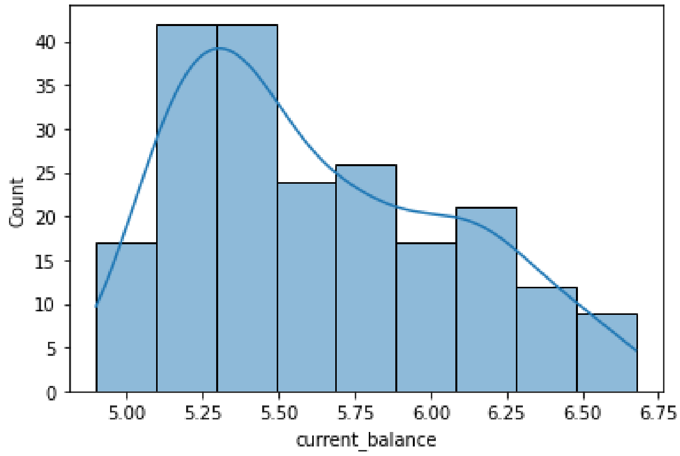


Figure 10: Histogram of Current_balance

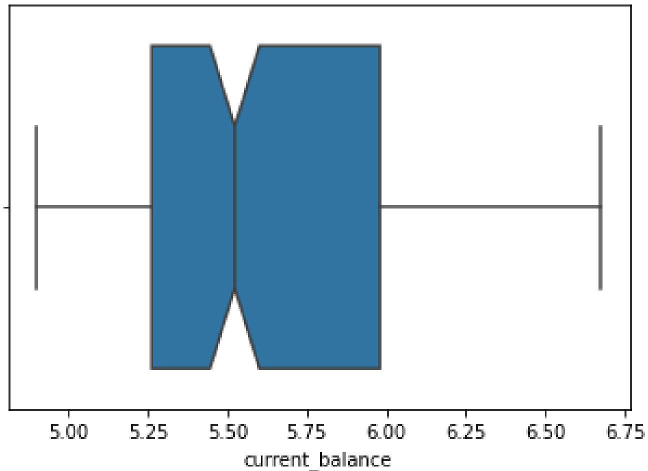


Figure 11: Boxplot of Current_balance

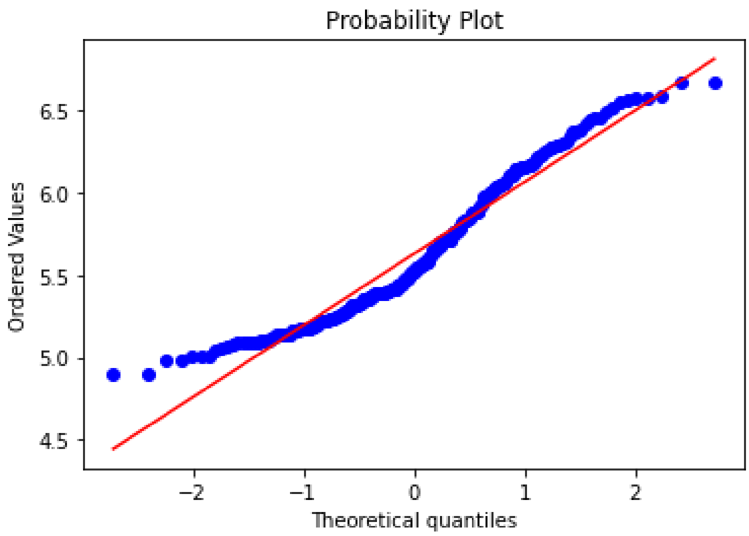


Figure 12: Probability Plot of Current_Balance

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

- Limit of the amount in credit card ranges from 2.63 to 4.03
- Average limit of the amount in credit card is 3.2
- The mean is not equal to median, the distribution is not normal which is evident from the boxplot and probability plot
- Skewness of the credit_limit attribution is 0.13 indicates a right tailed distribution
- Outliers are not present for this attribution which is evident from the box plot

```
Description of credit_limit
------------------------
count    210.000000
mean       3.258605
std        0.377714
min        2.630000
25%        2.944000
50%        3.237000
75%        3.561750
max        4.033000
```

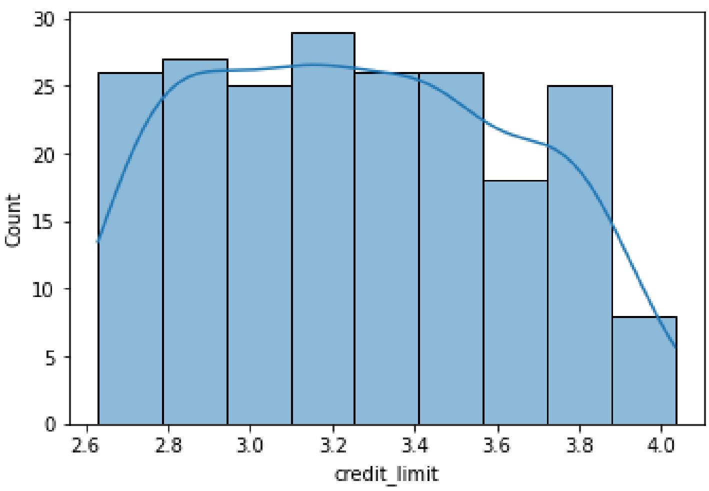


Figure 13: Histogram of credit_limit

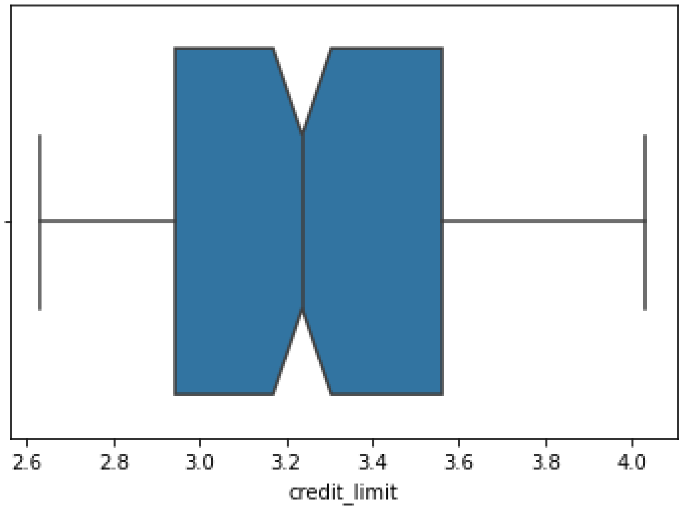


Figure 14: Boxplot of credit_limit

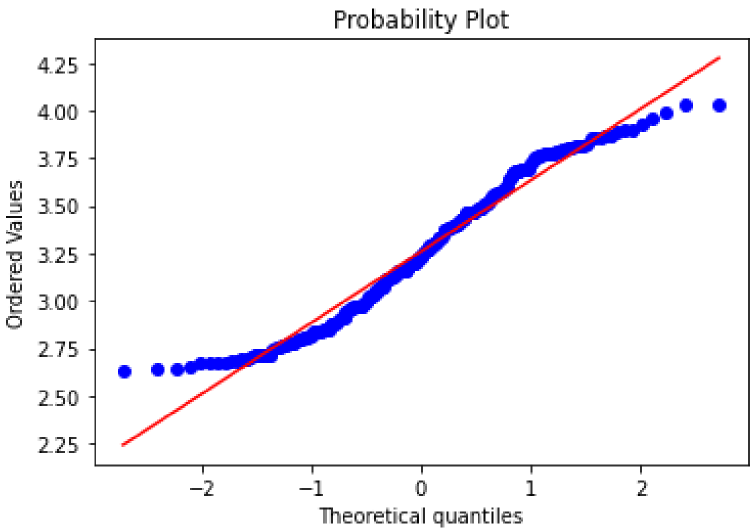


Figure 15: Probability Plot of Credit_limit

6. **min_payment_amt: minimum amount paid by the customer while making payments for purchases made monthly (in 100s)**
   - minimum amount paid by the customer while making payments for purchases made monthly ranges from 0.7 to 8.45
   - Average of minimum amount paid by the customer while making payments for purchases made monthly is 3.7
   - The mean is not equal to median, the distribution is not normal which is evident from the boxplot and probability plot
   - Skewness of the min_payment_amt attribution is 0.4 indicates a right tailed distribution
   - Outliers are present for this attribution which is evident from the box plot

```
Description ofmin_payment_amt
------------------------
count     210.000000
mean        3.700201
std         1.503557
min         0.765100
25%         2.561500
50%         3.599000
75%         4.768750
```

```
max          8.456000
```

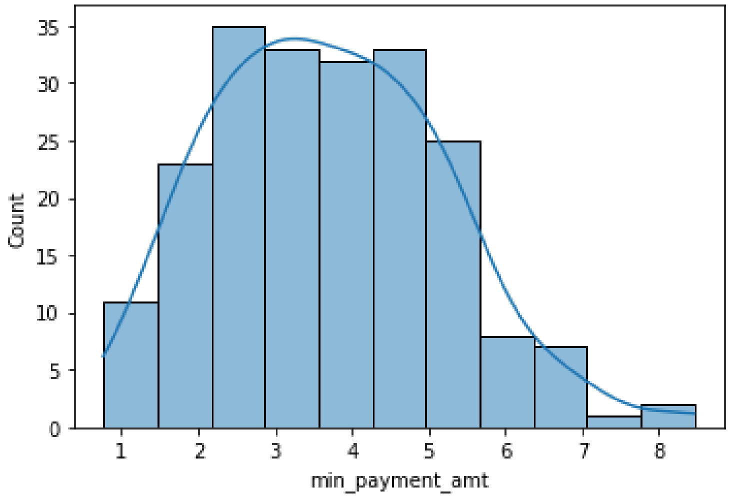


Figure 16: Histogram of min_payment_amt

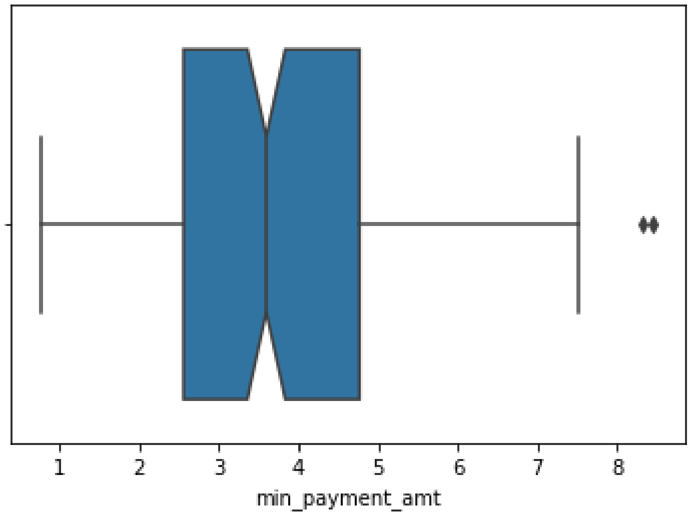


Figure 17: Boxplot of min_payment_amt

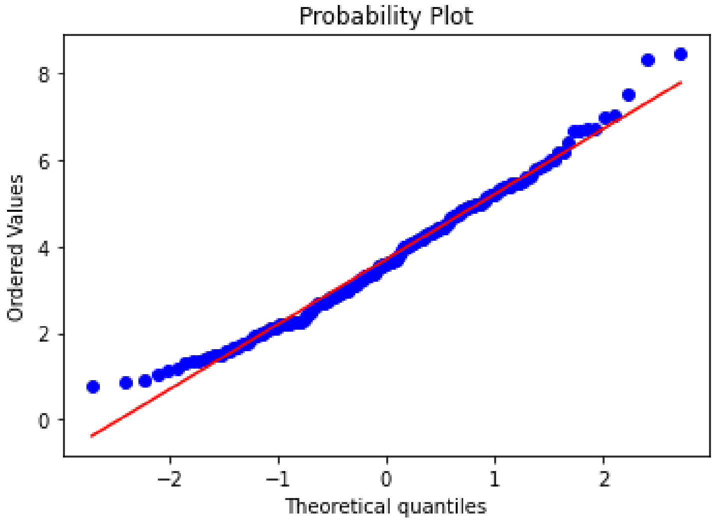


Figure 18: Probability plot of min_payment_amt

7. max_spent_in_single_shopping: (in 1000s) Maximum amount spent in one purchase
   - Maximum amount spent in one purchase in on ranges from 6.5 to 4.5
   - Average of Maximum amount spent in one purchase made monthly is 5.4
   - The mean is not equal to median, the distribution is not normal which is evident from the boxplot and probability plot
   - Skewness of the max_spent_in_single_shopping attribution is 0.56 indicates a right tailed distribution
   - Outliers are not present for this attribution which is evident from the box plot

```
Description max_spent_in_single_shopping
------------------------
count     210.000000
mean        5.408071
std         0.491480
min         4.519000
25%         5.045000
50%         5.223000
75%         5.877000
max         6.550000
```

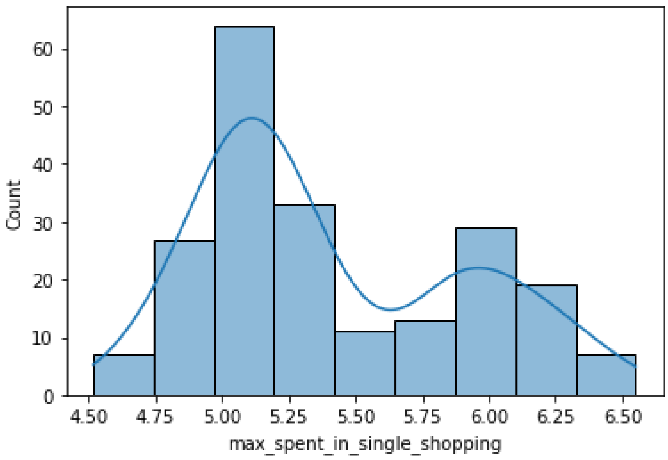


Figure 19: Histogram of max_spent_in_single_shopping

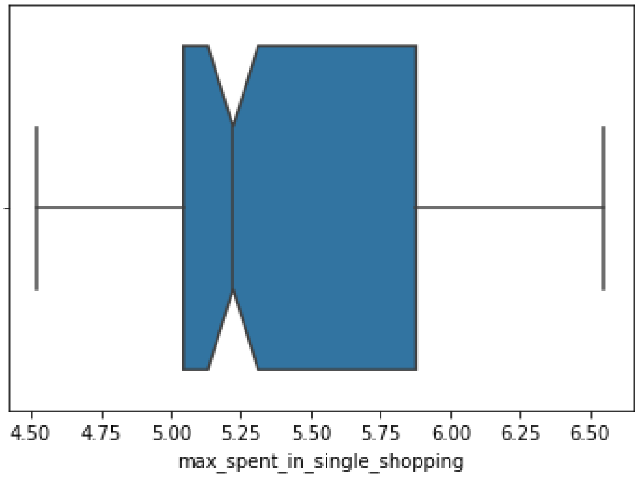


Figure 20: Boxplot of max_spent_in_single_shopping

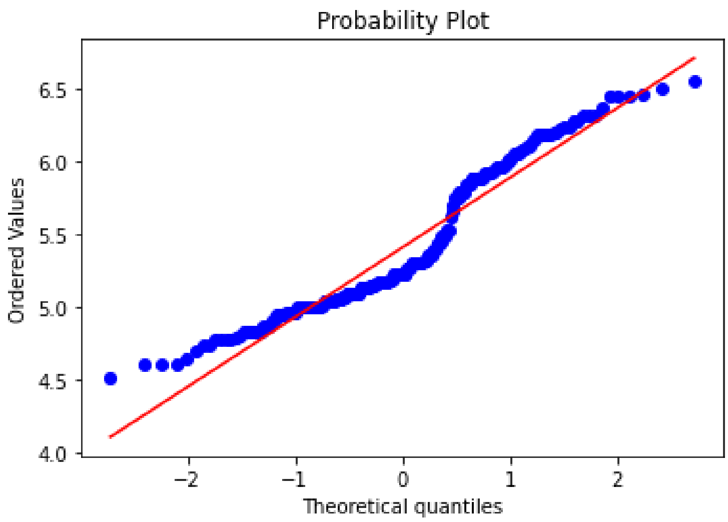


Figure 21: Probability Plot of max_spent_in_single_shopping

Multivariate-Bivariate analysis

greatlearning
Power Ahead

- Heat map shows the correlation between different attributes by assigning numbers as well as colours and Pairplot gives a graphical representation of correlation between different attributes.

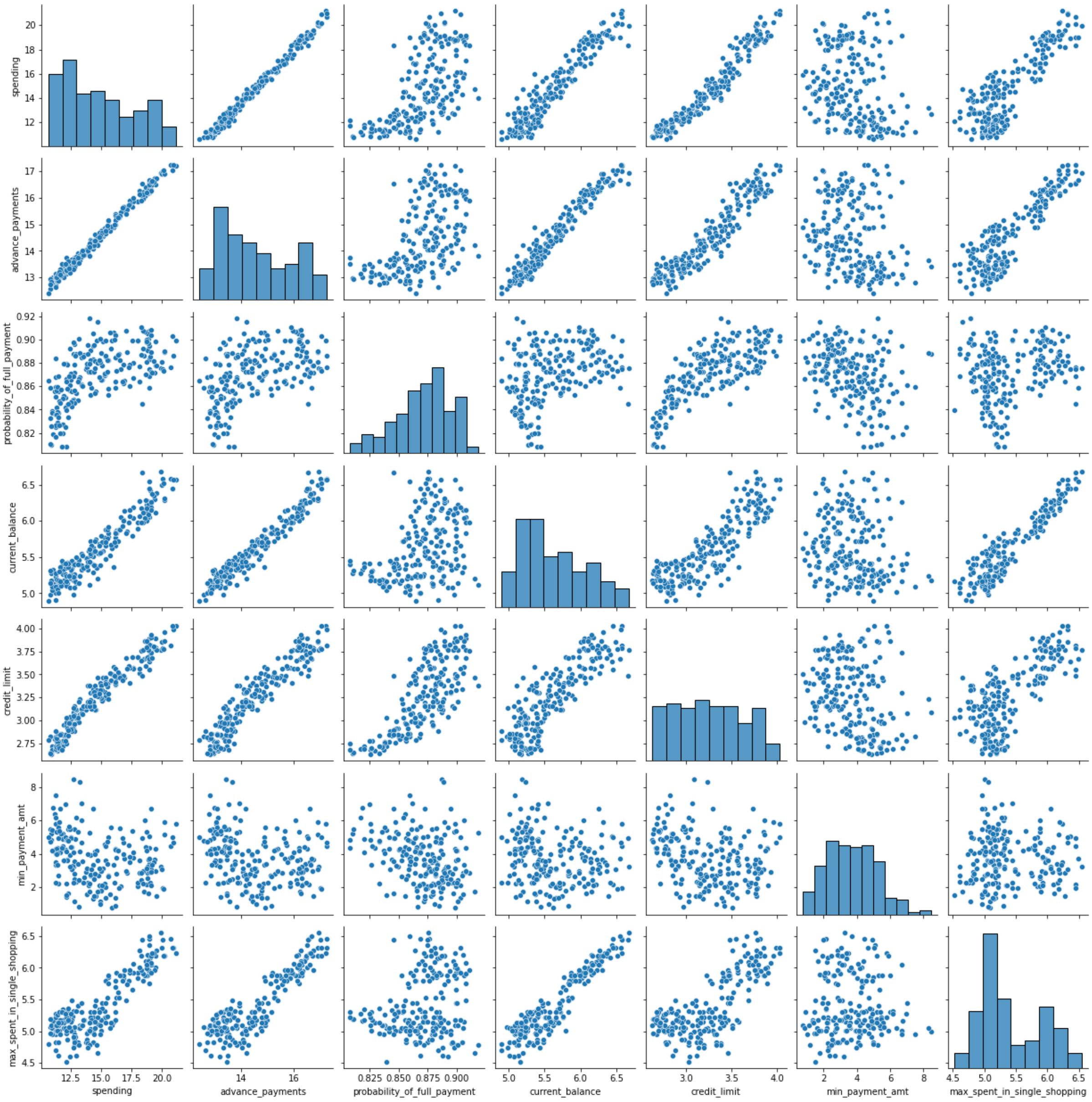


Figure 22: Pairplot for multivariate-bivariate analysis

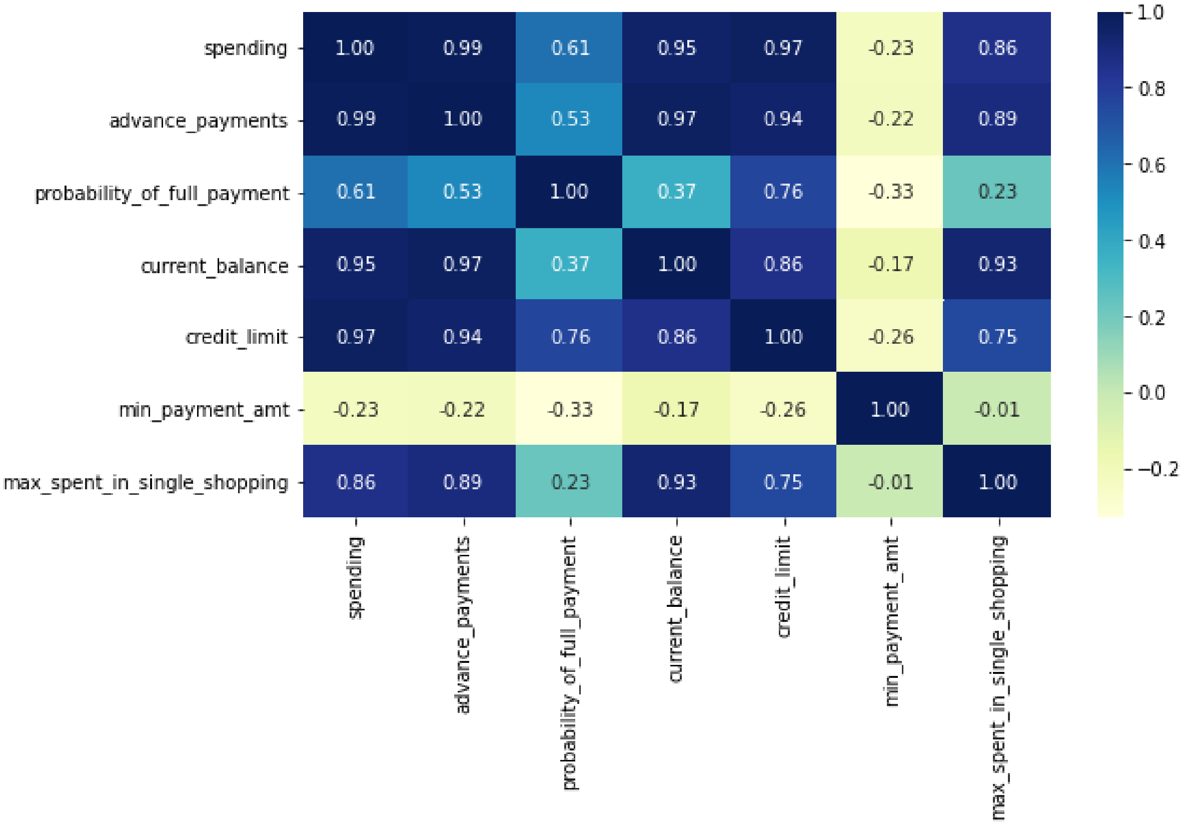
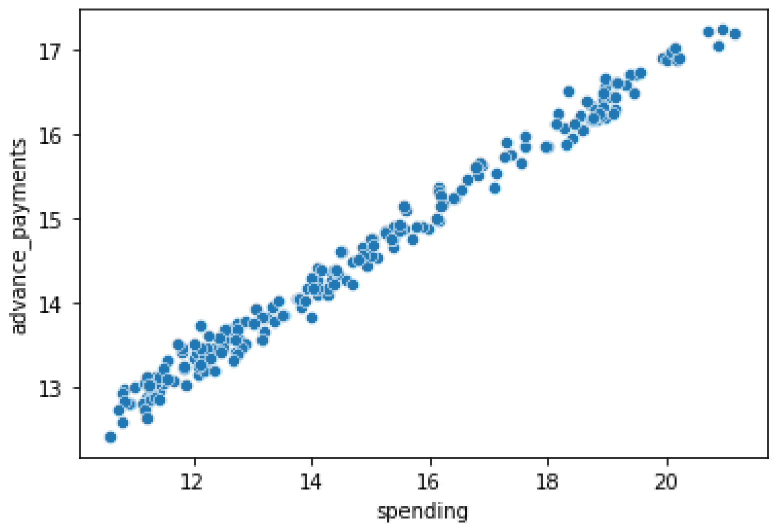


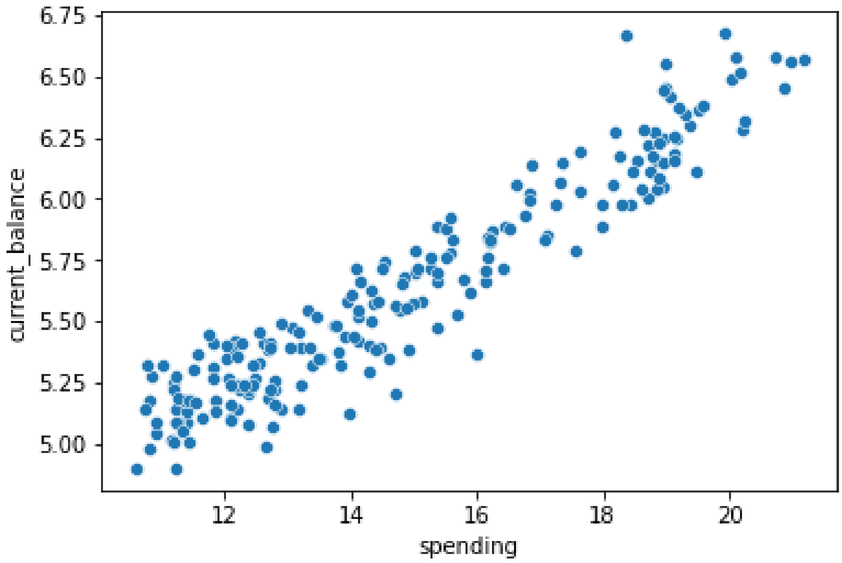
Figure 23: Heatmap for Multivariate-Bivariate Correlation

- The observations are as follows:

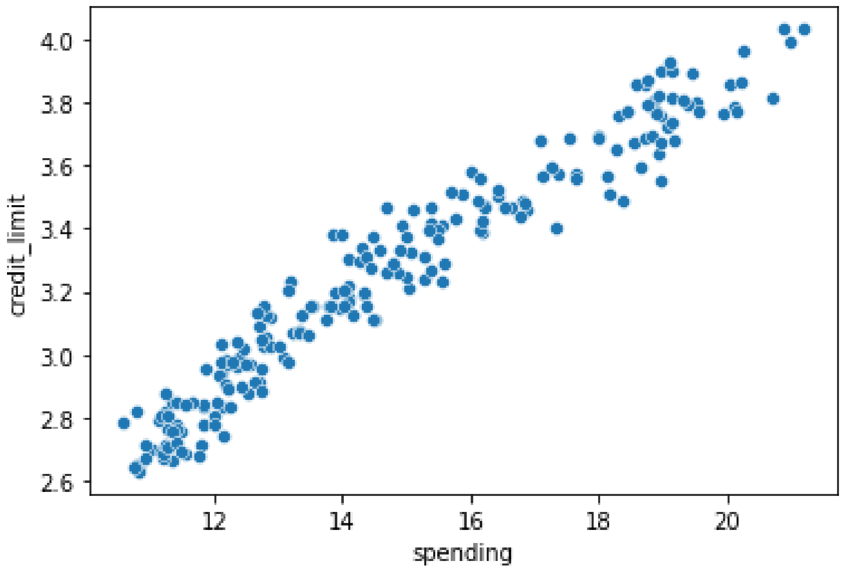
  I. There is a very strong positive correlation (0.99) between the spending and advance_payments; which infers that as the amount spent by the customer per month increases, the amount paid in advance by cash also increases. In short, Customer who spent maximum amount per month also paid maximum amount in advance by cash.



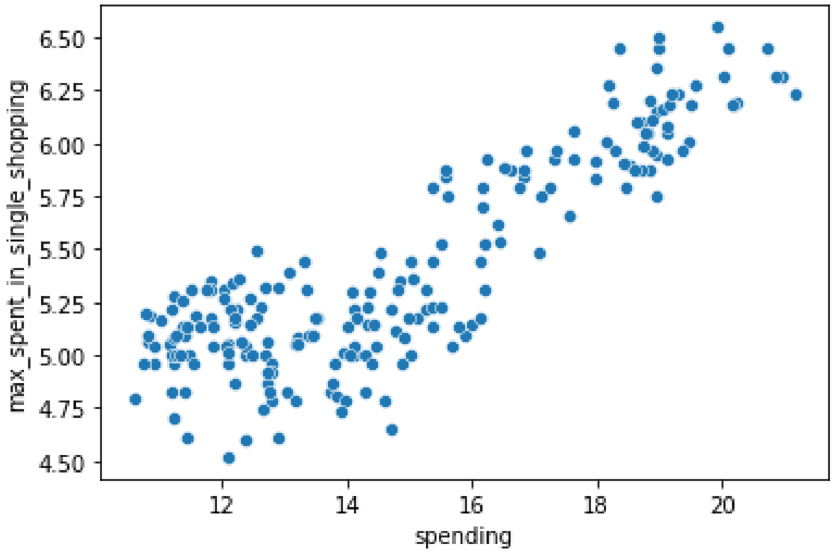
  II. There is a very strong positive correlation (0.95) between the spending and current_balance; which infers that as the amount spent by customers per month increases, their balance amount left in their account to make purchases also increases.



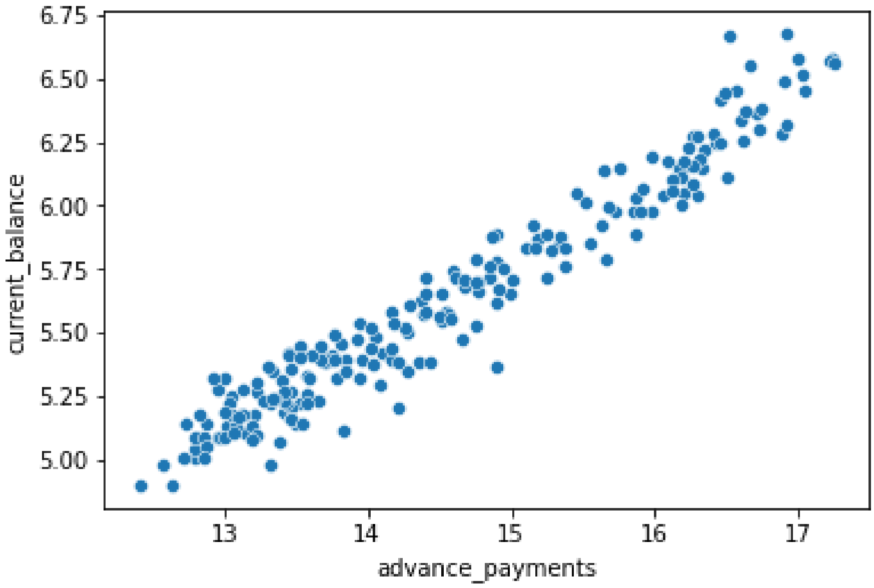
III. There is a very strong positive correlation (0.97) between the spending and credit_limit; which infers that as the amount spent by customers per month increases, Limit of the amount in their credit card also increases.



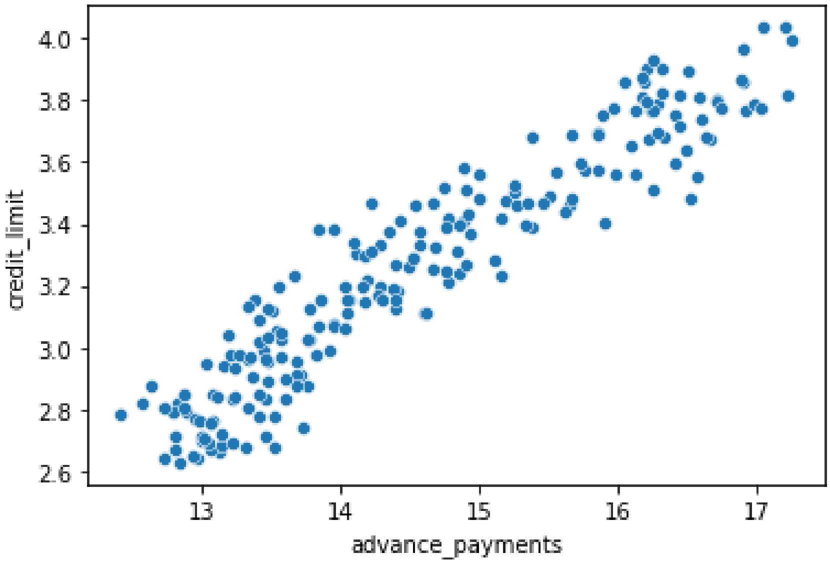
IV. There is a strong positive correlation (0.86) between the spending and max_spent_in_single_shopping; which infers that as the amount spent by customers per month increases, their maximum amount spent in one purchase also increases.



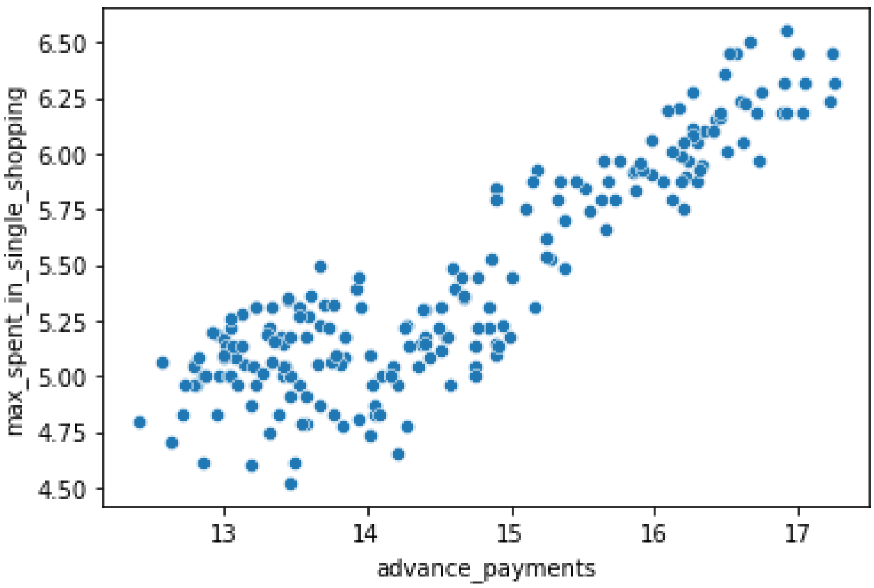
V. There is a very strong positive (0.97) correlation between the advance_payments and current_balance; which infers that as the amount paid by customers in advance by cash increases, their balance amount left in their account to make purchases also increases.



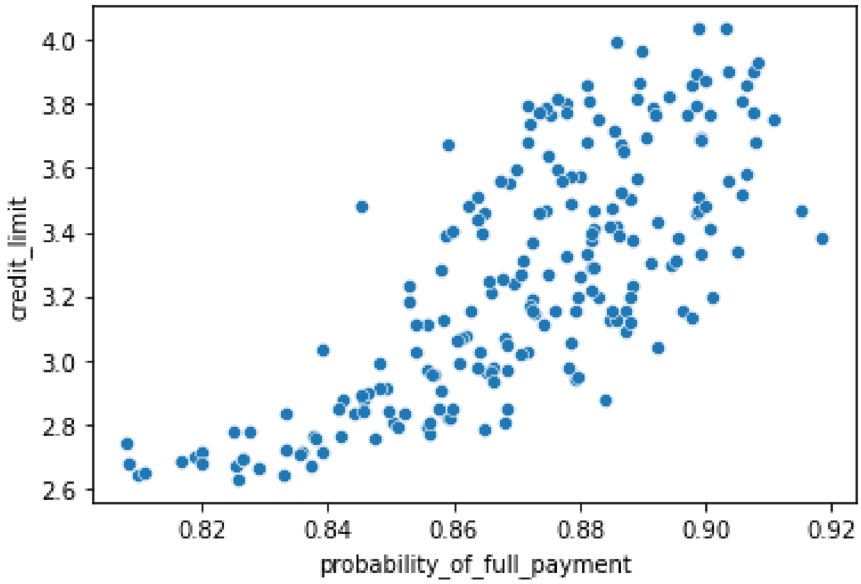
VI. There is a very strong positive (0.94) correlation between the advance_payments and credit_limit; which infers that as the amount paid by customers in advance by cash increases, Limit of the amount in their credit card also increases.



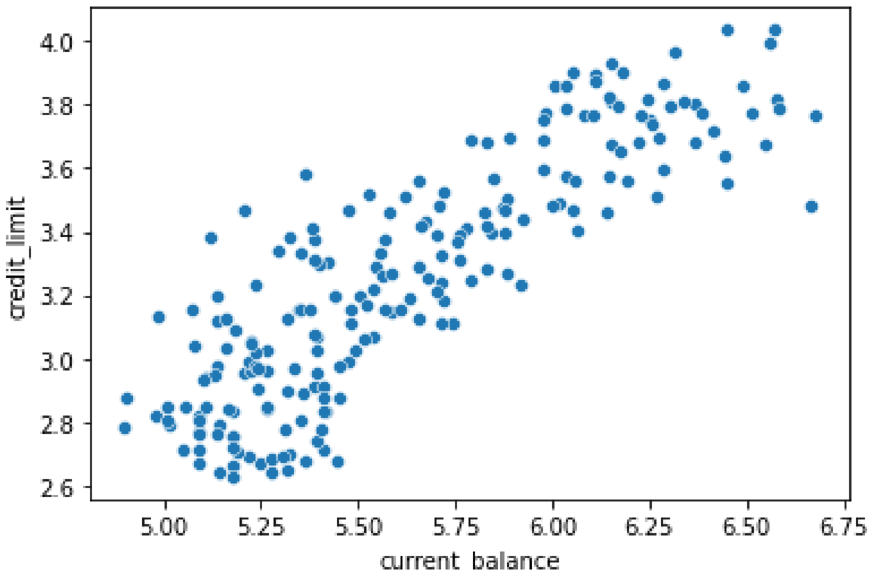
VII. There is a strong positive (0.89) correlation between the advance_payments and max_spent_in_single_shopping; which infers that as the amount paid by customers in advance by cash increases, their maximum amount spent in one purchase also increases.



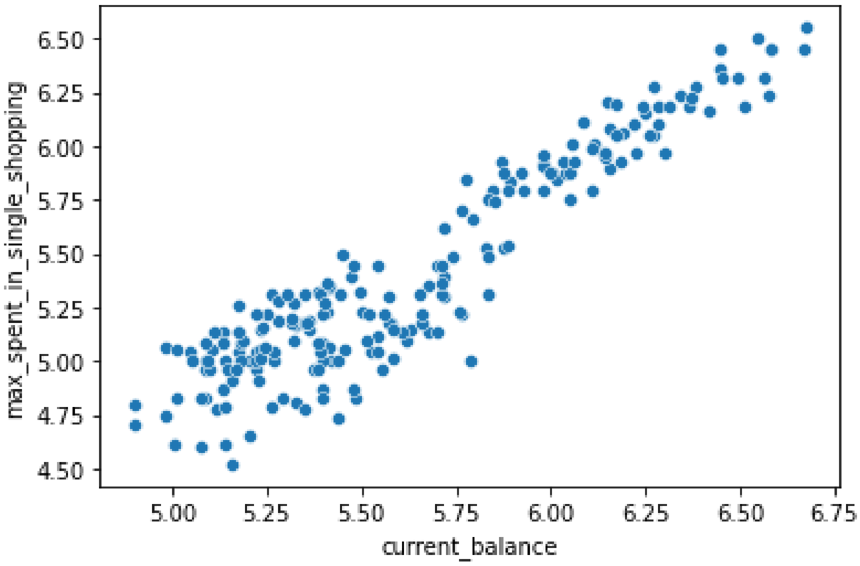
VIII. There is a positive (0.76) correlation between probability_of_full_payment and credit_limit; which infers that as the probability of payment done in full by the customer to the bank increases, their limit of the amount in the credit card also increases.



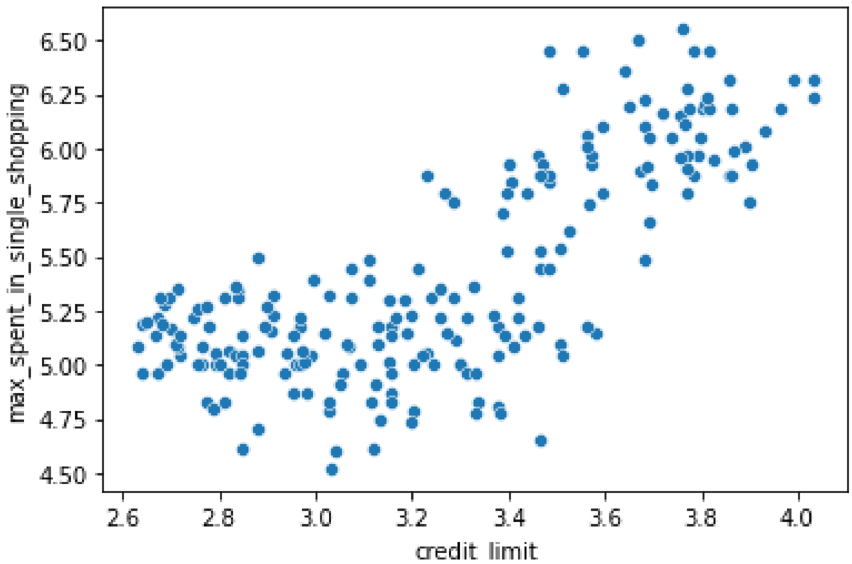
IX. There is a very strong positive correlation (0.95) between the current_balance and credit_limit; which infers as the balance amount left in customers account to make purchases increases, their limit of the amount in the credit card also increases.



X. There is a very strong positive correlation (0.93) between the current_balance and max_spent_in_single_shopping; which infers as the balance amount left in customers account to make purchases increases, their maximum amount spent in one purchase also increases.



XI. There is a positive correlation (0.75) between the credit_limit and max_spent_in_single_shopping; which infers as the limit of the amount in the credit card of customers increases, their maximum amount spent in one purchase also increases.



Problem 1.2

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

Yes, Scaling is necessary for clustering in this case

- Clustering is a fundamental unsupervised learning algorithm which uses distance-based methods for cluster formation. Hence, it gets highly influenced by the ranges of attributes
- For the given dataset it is evident that some attributes carry significantly more weightage compare to others and there is a significant difference in the range of attributes; Attributes with much larger range of values can influence the clustering output.

```
mean of unscaled 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
clusters                          2.014286
kmeans_clusters                   1.004762

standard deviation of unscaled 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
clusters                         0.827046
kmeans_clusters                  0.827156
```

- for example: the range of spending variable is between 10590 to 21180 whereas the range of min_payment_amt is 76 to 845. When any distance method is computed, for instance, Chebyshev Distance method the value of, max (|21180- 10590|) is much larger than the value of, max (|845- 76|); which implies that the clustering will be dominated by the feature, 'spending' when compared to the 'min_payment_amt'.
- Data scaling ensures that each attribute in the data is being weighted equally by the clustering algorithm. There are different types of standardization/ scaling method such as Min-Max Scaling, Z-Score Scaling, Scaling by StandardScaler.
- Formula for Min-Max Scaling= Xsc=X−Xmin/Xmax−Xmin. The data is scaled to a fixed range - usually 0 to 1
- Formula for Z-score Normalization= (x – mean)/ standard deviation, which ensures that mean is tending to 0 and standard deviation is tending to 1
- For the given dataset normalization done by StandardScaler
- It is evident that all the attributes are being weighted almost equally in the scaled data, which will give a better result for clustering. Hence, data scaling is required before clustering in this case.

```
Mean of Scaled Data
---------------------------------
0    9.148766e-16
1    1.097006e-16
2    1.243978e-15
3   -1.089076e-16
4   -2.994298e-16
5    5.302637e-16
6   -1.935489e-15
Standard Deviation of Scaled Data
---------------------------------
```

```
0     1.002389
1     1.002389
2     1.002389
3     1.002389
4     1.002389
5     1.002389
6     1.002389
```

Problem 1.3

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

- Hierarchical clustering was imported from the SciPy packages and applied to the scaled dataset
- 'Ward' linkage method is used, truncate mode was set at 10, which gives an output of dendrogram with last 10 merges. When linkage is "ward", only "Euclidean" distance is accepted as an affinity.

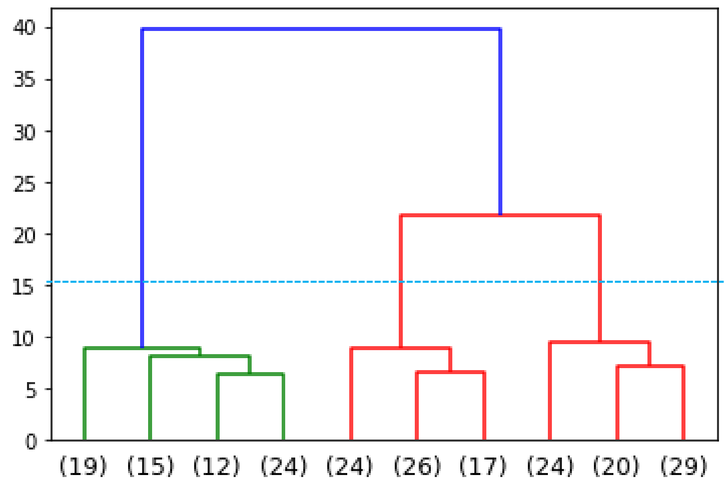


Figure 24: Dendrogram for Hierarchical clustering

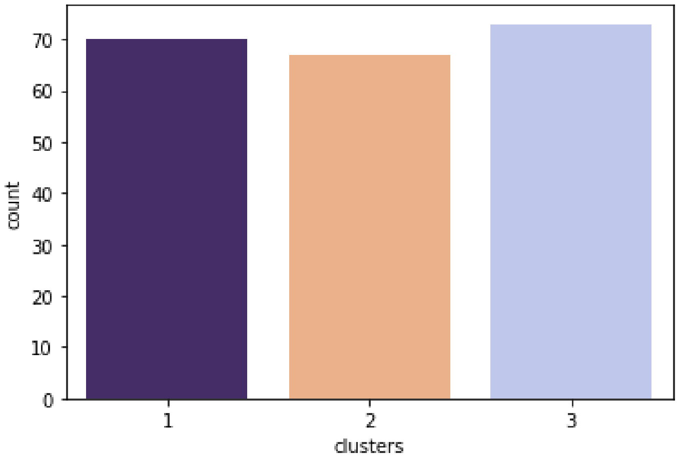
- 2 groups were seen clearly from the dendrogram. However, for the given business problem it will not have any significant impact.
- For further analysis Fluster is used with the distance criterion =15 for cluster formation, here 15 refers to the point on y-axis where a horizontal line is drawn to the x-axis, which derives the number of clusters. The number of clusters is equal to number of intersection points between the horizontal line and the dendrogram. 3 clusters are formed.

Figure 25: 3 clusters can be seen from the dendrogram

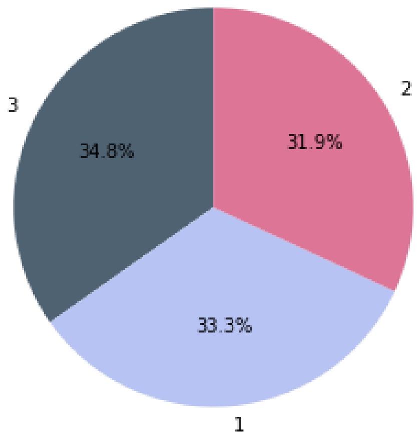
index	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.55	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1

3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1

- 70 records belong to cluster 1, 67 belongs to cluster 2, and 73 belongs to cluster 3 which is calculated by the value_counts function and can be seen from Count plot.



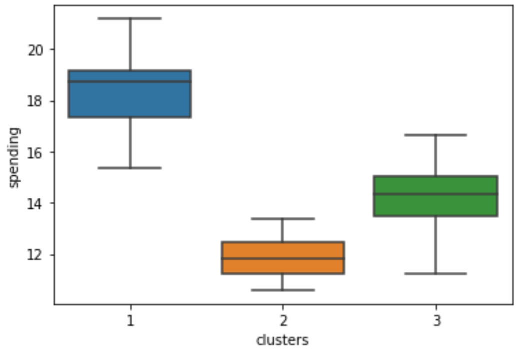
- The pie chart gives an idea about proportion of Customers in each cluster.



Bivariate analysis is done for stratification of Customers.

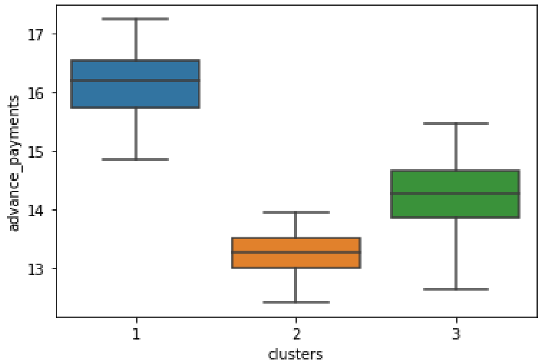
- **Spending- Amount spent by the customer per month (in 1000s)**

```
Mean of clusters in spending
-----------------------
clusters
1    18.371429
2    11.872388
3    14.199041
```



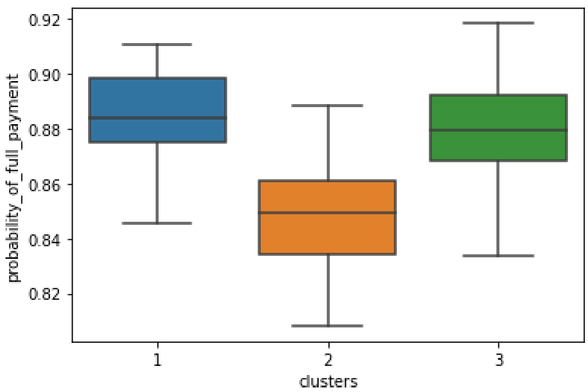
- **advance_payments: Amount paid by the customer in advance by cash (in 100s)**

```
Mean of clusters in advance_payments
-----------------------
clusters
1    16.145429
2    13.257015
3    14.233562
```



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

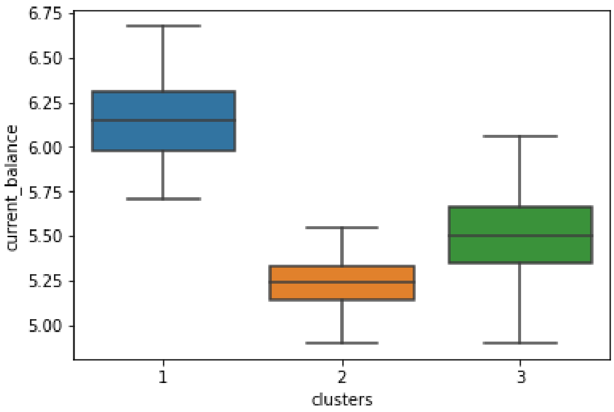
```
Mean of clusters in probability_of_full_payment
-----------------------
clusters
1    0.884400
2    0.848072
3    0.879190
```



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

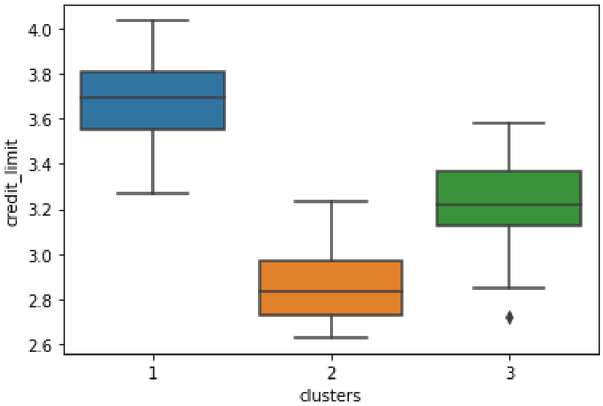
```
Mean of clusters in current_balance
-----------------------
```

```
clusters
1     6.158171
2     5.238940
3     5.478233
```



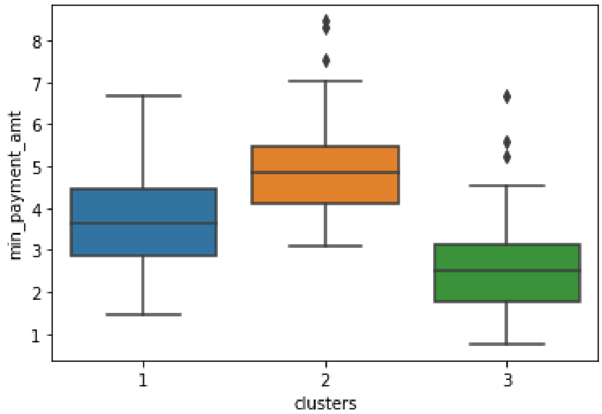
- **credit_limit: Limit of the amount in credit card (10000s)**

```
Mean of clusters in credit_limit
-----------------------
clusters
1     3.684629
2     2.848537
3     3.226452
```



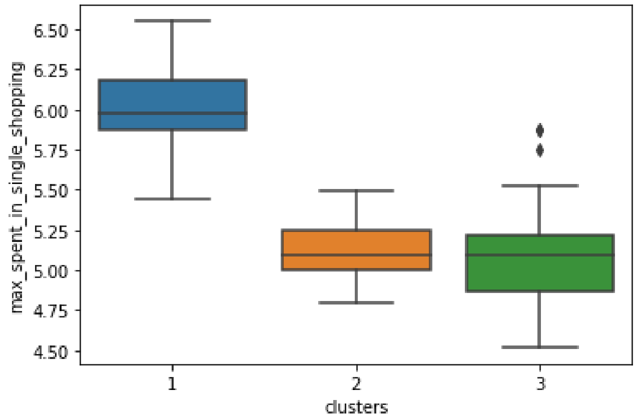
- **min_payment_amt: minimum paid by the customer while making payments for purchases made monthly (in 100s)**

```
Mean of clusters in min_payment_amt
-----------------------
clusters
1     3.639157
2     4.949433
3     2.612181
```



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

```
Mean of clusters in max_spent_in_single_shopping
-----------------------
clusters
1     6.017371
2     5.122209
3     5.086178
```



I. **Cluster 1**
   - It consists of 70 customers with highest spending, highest advance payments, highest probability of full 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 exhibit second highest after cluster 2. We can infer those customers belong to cluster 1 are wealthy.

II. **Cluster 2**
   - It consists of 67 customers with lowest spending, lowest advance payments, lowest current balance, lowest credit limit and lowest probability of full payments. They exhibit the highest with respect to minimum payment amount and second highest with respect to maximum spent in a single shopping. We can infer those customers belong to cluster 2 are not wealthy.

III. **Cluster 3**
   - It consists of 73 customers who fall in between customers in cluster 1 and cluster 2 and are second highest with regard to 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. We can infer those customers belong to cluster 3 lies just below the wealthy people.

Problem 1.4

Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

- K-means clustering was imported from the Sklearn packages and applied to the scaled dataset. Within sum of squares were checked for clusters (1 to 10).

```
1.1469.9999999999998,
2.659.171754487041,
3.430.6589731513007,
4.371.30172127754213,
5.326.36254154106985,
6.290.02485649252253,
7.262.16062941548705,
8.240.84566436921847,
9.220.0243198630256,
10.206.25580243471035
```

- There is a significant drop in WSS values from 1 to 2 (1469 to 659) and from 2 to 3 (659 to 430); However, from 3 onwards the drop becomes gradual. So, we have taken 3 clusters as optimum number of clusters.

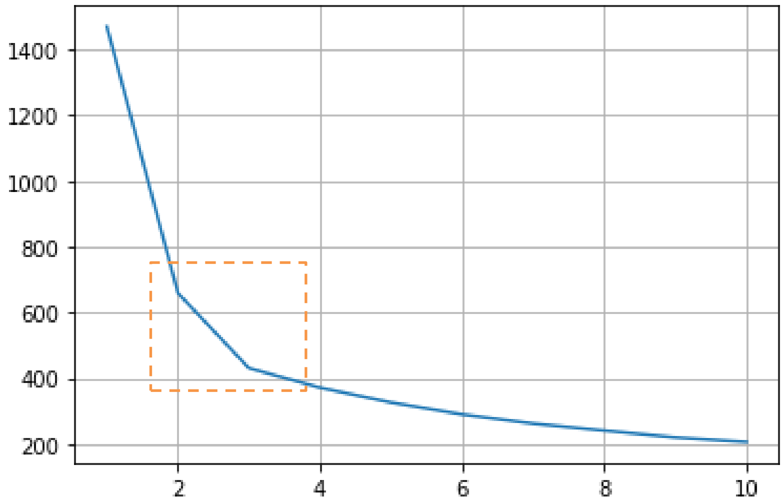


Figure 26: Elbow Plot

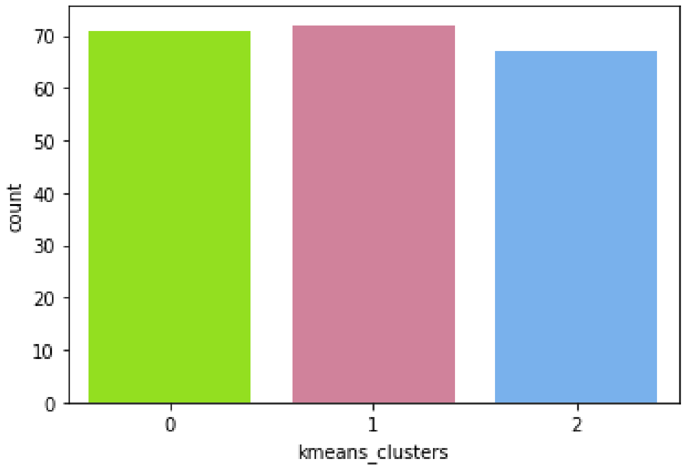
- As per the WSS plot there is a significant drop can be seen from 1 to 2; However, the elbow joint can be seen from the graph, when the number of clusters=3

- Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. The silhouette score for cluster 2 is 0.465, for cluster 3 is 0.40, for cluster is 0.32

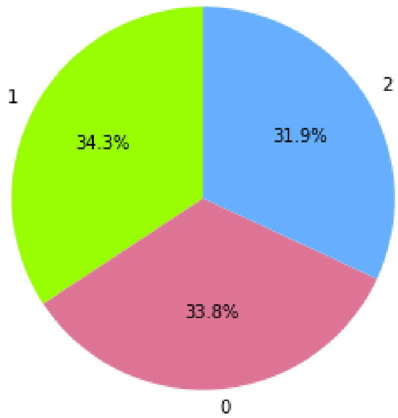
- the smallest value of silhouette sample for cluster 2 is -0.006 and cluster 4 is -0.05. A negative value indicates, observations might have got assigned to the wrong cluster which is not acceptable.

- the smallest value of silhouette sample for cluster 3 is 0.002. A positive value indicates, there is no observation that is incorrectly map within clusters. Hence, we have taken 3 clusters as optimum number of clusters.

- 71 records belong to cluster 0, 72 belongs to cluster 1, and 67 belongs to cluster 2 which is calculated by the value_counts function and can be seen from Count plot.



- The pie chart gives an idea about proportion of Customers in each cluster.

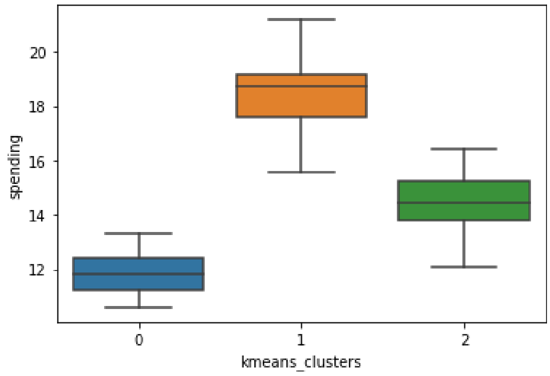


Index	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	kmeans_clusters
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.55	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	0
4	17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1

Bivariate analysis is done for stratification of Customers.

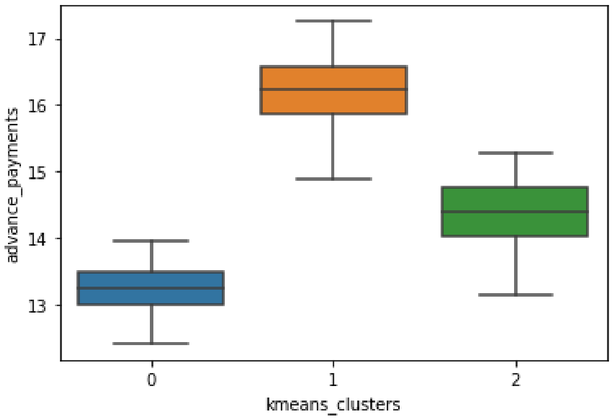
- **Spending- Amount spent by the customer per month (in 1000s)**

```
Mean of Kmeans_clusters in spending
-----------------------------
kmeans_clusters
0     11.856944
1     18.495373
2     14.437887
```



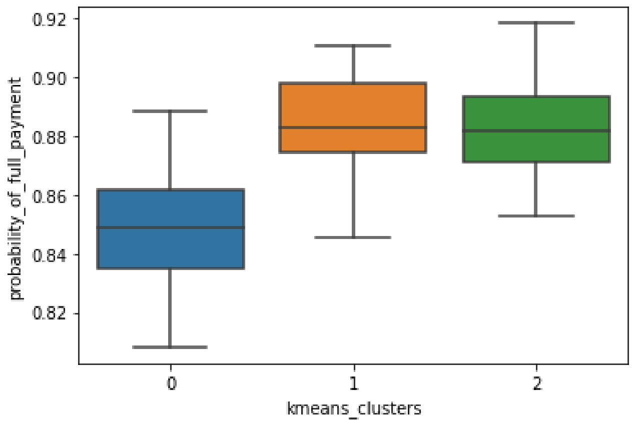
- **advance_payments: Amount paid by the customer in advance by cash (in 100s)**

```
Mean of Kmeans_clusters in advance_payments
-----------------------------
kmeans_clusters
0     13.247778
1     16.203433
2     14.337746
```



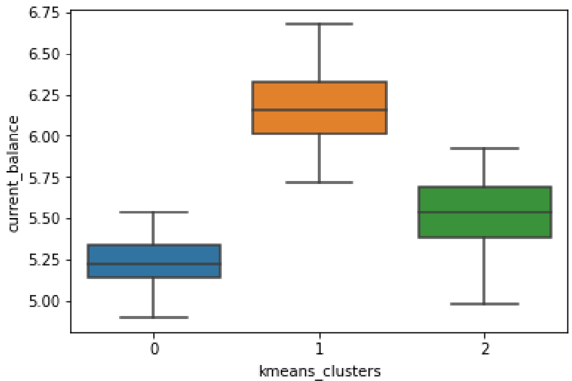
- **probability_of_full_payment: Probability of payment done in full by the customer to the bank**

```
Mean of Kmeans_clusters in probability_of_full_payment
-----------------------------
kmeans_clusters
0     0.848253
1     0.884210
2     0.881597
```



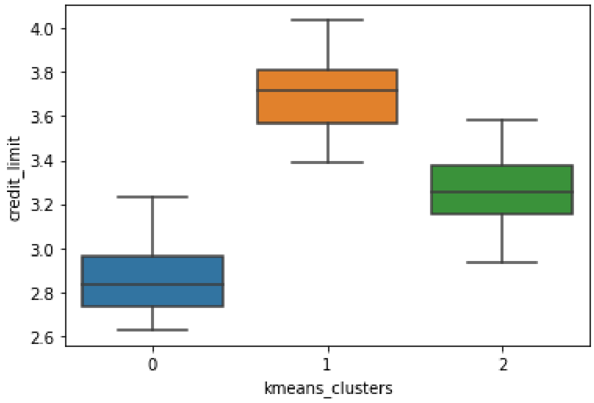
- **current_balance: Balance amount left in the account to make purchases (in 1000s)**

```
Mean of Kmeans_clusters in current_balance
-----------------------------
kmeans_clusters
0    5.231750
1    6.175687
2    5.514577
```



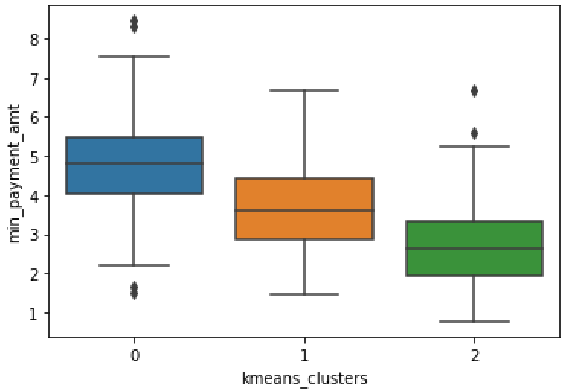
- **credit_limit: Limit of the amount in credit card (10000s)**

```
Mean of Kmeans_clusters in credit_limit
-----------------------------
kmeans_clusters
0    2.849542
1    3.697537
2    3.259225
```



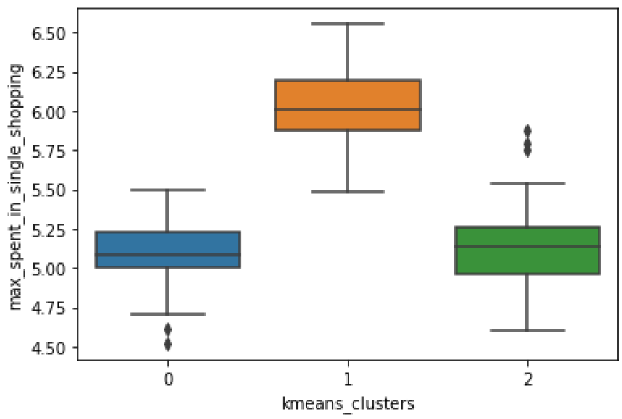
- **min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)**

```
Mean of Kmeans_clusters in min_payment_amt
-----------------------------
kmeans_clusters
0     4.742389
1     3.632373
2     2.707341
```



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

```
Mean of Kmeans_clusters in max_spent_in_single_shopping
-----------------------------
kmeans_clusters
0     5.101722
1     6.041701
2     5.120803
```



IV. **Cluster 0**

- It consists of 72 customers with lowest spending, lowest advance payments, lowest current balance, lowest credit limit and lowest probability of full payments. They exhibit the highest with respect to minimum payment amount. We can infer those customers belong to cluster 0 are not wealthy

V. **Cluster 1**

- It consists of 67 customers with highest spending, highest advance payments, highest probability of full 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 exhibit second highest after cluster 0. We can infer those customers belong to cluster 1 are wealthy

VI. **Cluster 2**

- It consists of 72 customers who fall in between customers in cluster 0 and cluster 1 and are second highest with regard to spending, advance payments, probability of full payment, current balance, credit limit. They exhibit the lowest value with respect to minimum payment amount and second highest in maximum spent in single shopping. We can infer those customers belong to cluster 2 lies just below the wealthy people.

Problem 1.5

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

- Both hierarchical clustering and K-means clustering has given us the almost similar clustering
- Both from Hierarchal clustering and K-means clustering we have obtained 3 optimum clusters suitable for the given business problem

Index	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters	kmeans_clusters
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.55	1	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3	2
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1	1
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0
4	17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1	1

- Below Data gives the Average of Clusters in K-Mean Clustering

```
Mean of Kmeans_clusters in spending
-----------------------------
kmeans_clusters
0    11.856944
1    18.495373
2    14.437887
Name: spending, dtype: float64
```

```
Mean of Kmeans_clusters in advance_payments
-----------------------------
kmeans_clusters
0    13.247778
1    16.203433
2    14.337746
Name: advance_payments, dtype: float64

Mean of Kmeans_clusters in probability_of_full_payment
-----------------------------
kmeans_clusters
0    0.848253
1    0.884210
2    0.881597
Name: probability_of_full_payment, dtype: float64

Mean of Kmeans_clusters in current_balance
-----------------------------
kmeans_clusters
0    5.231750
1    6.175687
2    5.514577
Name: current_balance, dtype: float64

Mean of Kmeans_clusters in credit_limit
-----------------------------
kmeans_clusters
0    2.849542
1    3.697537
2    3.259225
Name: credit_limit, dtype: float64

Mean of Kmeans_clusters in min_payment_amt
-----------------------------
kmeans_clusters
0    4.742389
1    3.632373
2    2.707341
Name: min_payment_amt, dtype: float64

Mean of Kmeans_clusters in max_spent_in_single_shopping
-----------------------------
kmeans_clusters
0    5.101722
1    6.041701
2    5.120803
Name: max_spent_in_single_shopping, dtype: float64
```

- Below data gives the average of clusters in Hierarchical clustering

```
Mean of clusters in spending
-----------------------
clusters
1    18.371429
2    11.872388
```

```
3    14.199041
Name: spending, dtype: float64

Mean of clusters in advance_payments
-----------------------
clusters
1    16.145429
2    13.257015
3    14.233562
Name: advance_payments, dtype: float64

Mean of clusters in probability_of_full_payment
-----------------------
clusters
1    0.884400
2    0.848072
3    0.879190
Name: probability_of_full_payment, dtype: float64

Mean of clusters in current_balance
-----------------------
clusters
1    6.158171
2    5.238940
3    5.478233
Name: current_balance, dtype: float64

Mean of clusters in credit_limit
-----------------------
clusters
1    3.684629
2    2.848537
3    3.226452
Name: credit_limit, dtype: float64

Mean of clusters in min_payment_amt
-----------------------
clusters
1    3.639157
2    4.949433
3    2.612181
Name: min_payment_amt, dtype: float64

Mean of clusters in max_spent_in_single_shopping
-----------------------
clusters
1    6.017371
2    5.122209
3    5.086178
Name: max_spent_in_single_shopping, dtype: float64
```

- From Hierarchical clustering we got 3 types of customers: cluster 1- wealthy, cluster 2- Not wealthy, cluster 3 – Below wealthy but above not wealthy customers. From K-means clustering we got 3 types of customer cluster 1- wealthy, cluster 0- Not wealthy, cluster 2– Below wealthy but above not wealthy customers

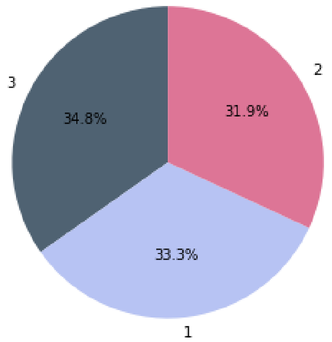


Figure 27: Pie chart of Hierarchial clustering

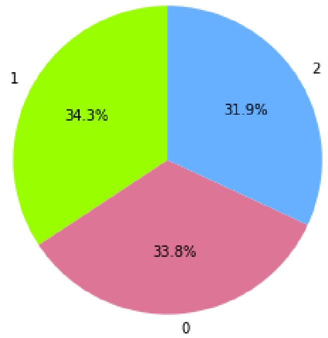


Figure 28: Pie chart of KMeans Clustering

- Let's do a customer segmentation for better understanding
  Wealthy - Tier 1 customers approx.34.3%
  Below Wealthy- Tier 2 customers approx.31.9%
  Not Wealthy- Tier 3 customers approx.33.8%

Profiling Of Clusters

- **Tier 1 Customers:**
  - These Customers obtained highest in all attributes among all three clusters except for the minimum payment amount
  - Since these customers spent highest maximum amount in one purchase, they obtained second highest in the minimum amount paid by the customer while making payments for purchase
  - They spend highest amount of money per month
  - They pay largest sum of money in advance by cash
  - Their limit of the amount in credit card is highest because these customers are high spenders.
  - They have highest Balance amount left in their account to make purchases
  - Since, these customers have the highest probability of making full payments, the probability of them falling in defaulter list is almost nil.
  - Overall, these are Customers who brings highest amount of Profit to the banks and are safe customers for the bank

- **Tier 2 Customers:**
  - These Customers obtained second highest in all attributes and lowest for the minimum payment amount
  - They spend second highest amount of money per month
  - They pay second largest sum of money in advance by cash

- Their limit of the amount in credit card is second highest.
- They have second highest Balance amount left in their account to make purchases
- They spent second highest maximum amount in one purchase
- These customers obtained the second highest probability of making full payments, However, the probability of them falling in defaulter list cannot be ignored; because these customers paid minimum amount while making payments for purchase which raise the concern that some of these customers might end up being defaulters

- **Tier 3 Customer:**

  - These Customers obtained lowest in all attributes among all three clusters and exhibit highest for the minimum payment amount
  - They spend least amount of money per month
  - They pay least sum of money in advance by cash
  - Their limit of the amount in credit card is lowest because these customers are least spenders.
  - They have least Balance amount left in their account to make purchases
  - These customers have the least probability of making full payments, however, these customers obtained highest amount while making payment for purchases made monthly, which makes them safe customer and there is a very little chance of them might end up being defaulters.

- **Business Recommendation for Promotional strategies**

- **For Tier 3 Customers:**

  - These customers obtained the highest rank with respect to minimum payment amount which is the minimum paid by the customer while making payments for purchases, which makes them safe customer and there is very little chance of them might end up being defaulters.
  - To attract these customers, Bank should reduce the amount of the minimum payment that will allow them to spend more. credit limit of these customers could be increased with rewards or discounts by the bank which will encourage them to spend more.
  - The above strategies will be beneficial for both the customers and bank and could improve the business of bank.

- **For Tier 1 Customers:**

  - These are ideal customers to the bank; they bring highest amount of revenues to the bank.
  - Since these customers are highest spenders, an increase in their credit limit will allow them to spend more, resulting more profit to the bank. Also, the bank should give them premium offers on each purchase which will encourage them to spend more.
  - Since, these customers have the highest probability of making full payments, the probability of them falling in defaulter list is almost nil. Hence, Bank should give them add on card or family card, which will please these customers, resulting in more spending by them and more benefit to the customer

- **For Tier 2 Customer:**

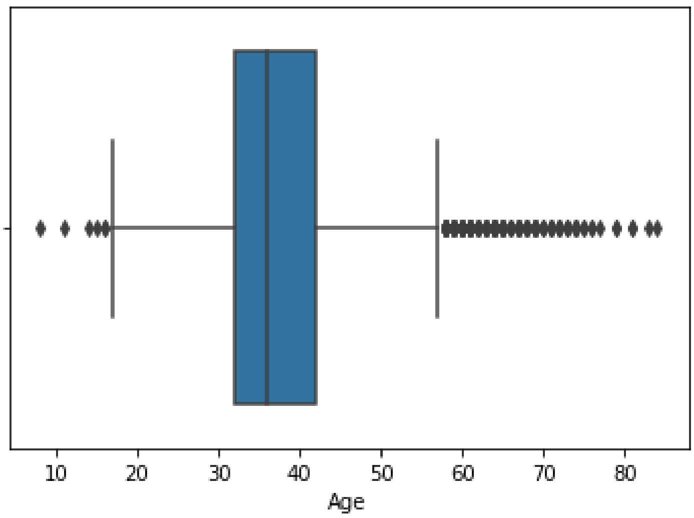
  - For these customers, it is better to increase the credit limit (along with rewards) or reduction in the amount of the minimum payment or an add on credit card will encourage them to spend more.
  - The bank should follow any one of the above strategies since these customers might end up being defaulters which can’t be ignored.

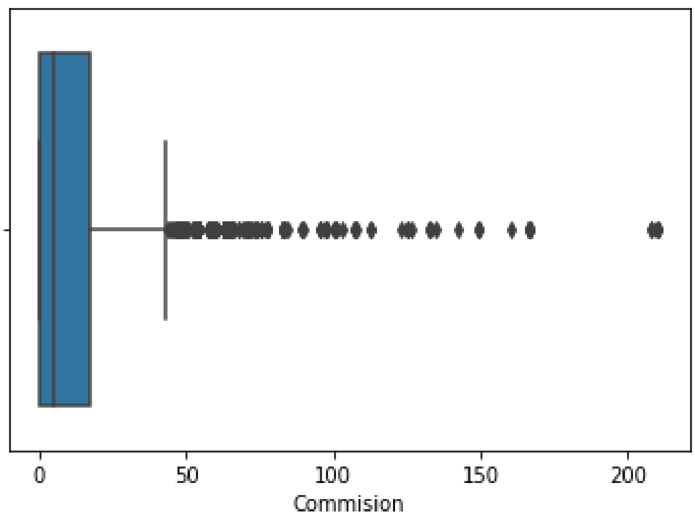
greatlearning
Power Ahead

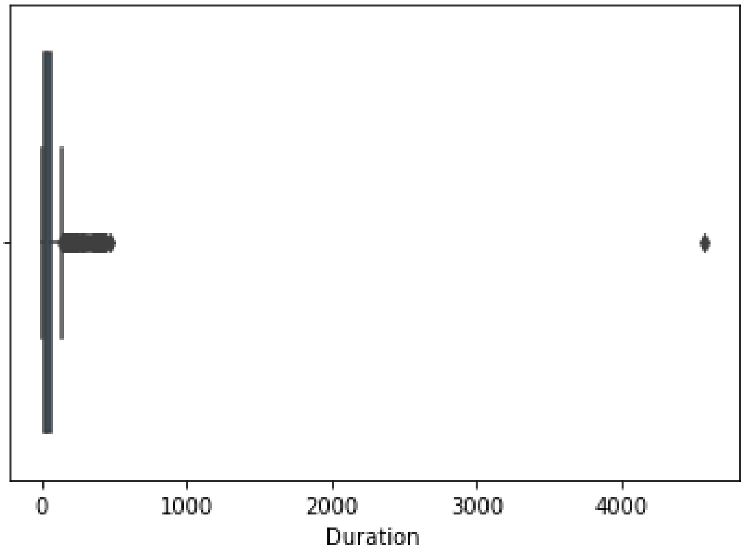
Problem 2.1

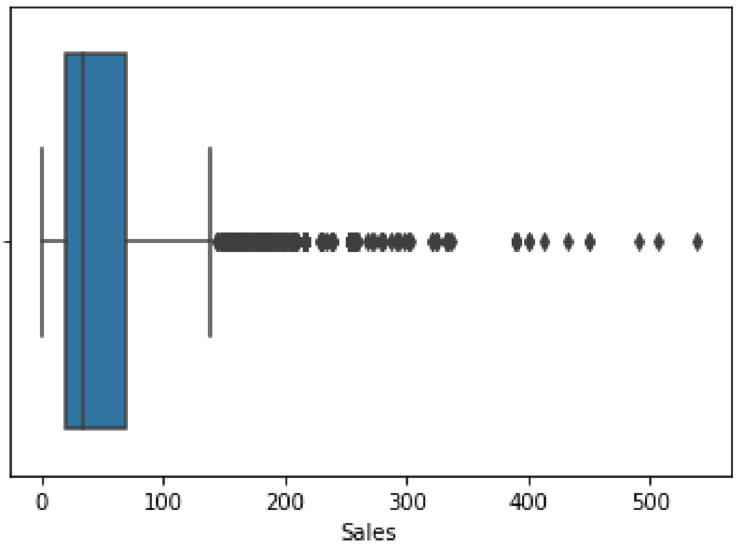
Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

- The given dataset has 3000 rows and 10 columns. Out of the 10 attributes 4 features are numeric (int, float) and 6 features are object data type.
- No bad data present in the dataset
- The dataset has no missing values.
- The dataset has 139 duplicate records, which contributes less than 5% of the dataset; Hence, it was not treated/removed.
- Outliers are present in all the 4 numerical features: Age, Commission, Duration, Sales. Since, Decision tree is susceptible to outliers therefore no outliers treatment is being done.
- Scaling has not been done on the entire dataset and later performed on the respective algorithm based on the requirement.
- There is an anomaly present in the duration attribution (-1)









Anomalies Treatment

- The describe function as well as the unique function shows that the minimum value for duration is seen as -1, which is not possible in real life scenario. This attribution is based on number of days; therefore, it is imputed with the mode (most frequent value) which is 0.

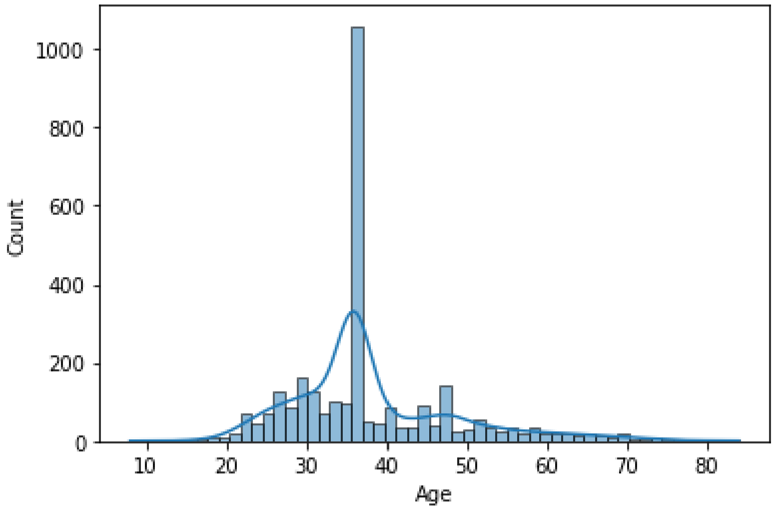
Univariate Analysis for Numeric Attributes

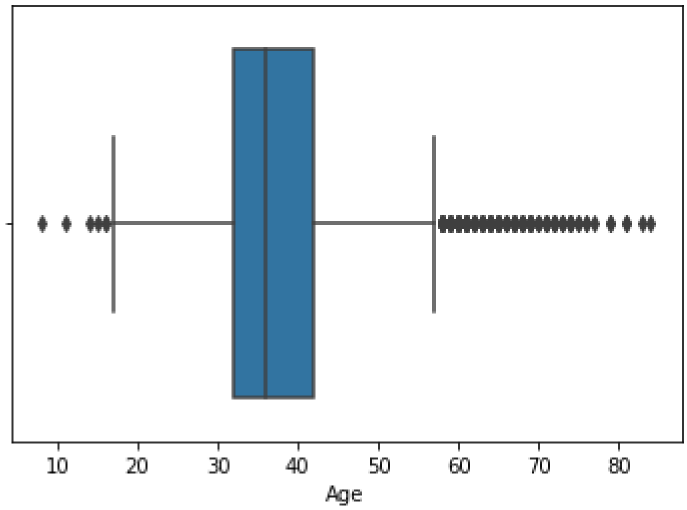
Age: Age of insured

- Age of insured ranges from 8 to 84 yrs
- Mean age is 38 yrs and median age is 36. Skewness is 1.1
- Mean is greater than median, indicates that the distribution is right tailed
- Outliers present in this attribute

```
Description of Age
---------------
count     3000.000000
mean        38.091000
std         10.463518
```

```
min          8.000000
25%         32.000000
50%         36.000000
75%         42.000000
max         84.000000
```



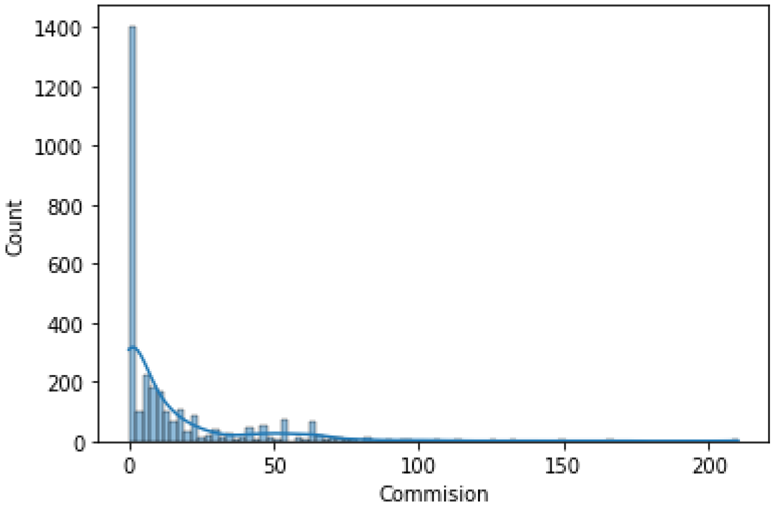


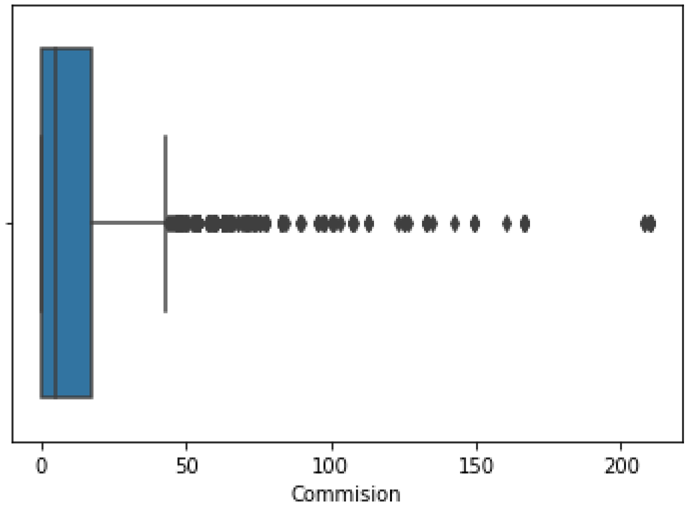
Commission: The commission received for tour insurance firm

- The commission received for tour insurance firm ranges from 0 to 210
- Mean of the commission received for tour insurance firm is 14.5 and median is 4. Skewness is 3.14.
- Mean is greater than median, indicates that the distribution is right tailed
- Outliers present in this attribute

```
Description of Commission
---------------
count     3000.000000
mean        14.529203
std         25.481455
min          0.000000
25%          0.000000
```

```
50%          4.630000
75%         17.235000
max        210.210000
```



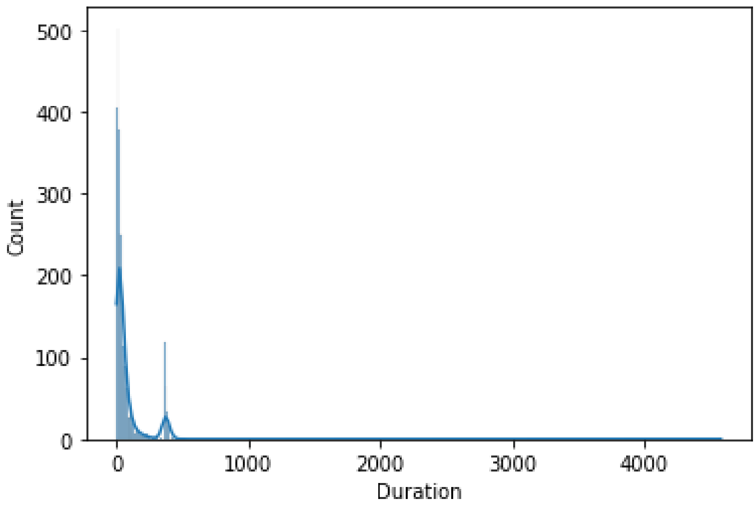


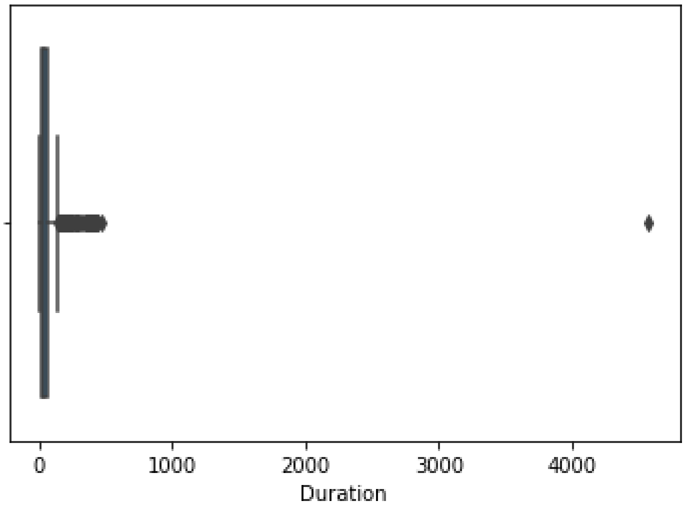
Duration: Duration of the tour

- The duration of tour ranges from 0 to 4580
- Mean of the duration of tour is 70 and median is 26.5. Skewness is 13.8.
- Mean is greater than median, indicates that the distribution is right tailed
- Outliers present in this attribute

```
Description of Duration
---------------
count     3000.000000
mean        70.001333
std        134.053313
min         -1.000000
25%         11.000000
50%         26.500000
```

```
75%         63.000000
max       4580.000000
```

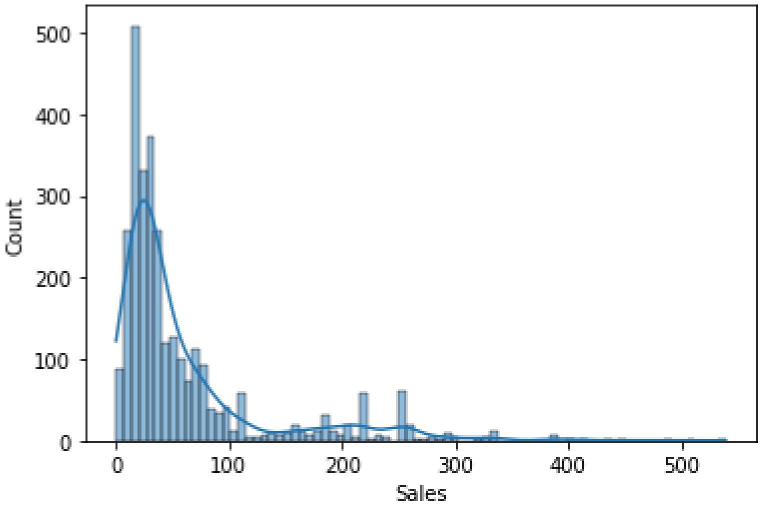


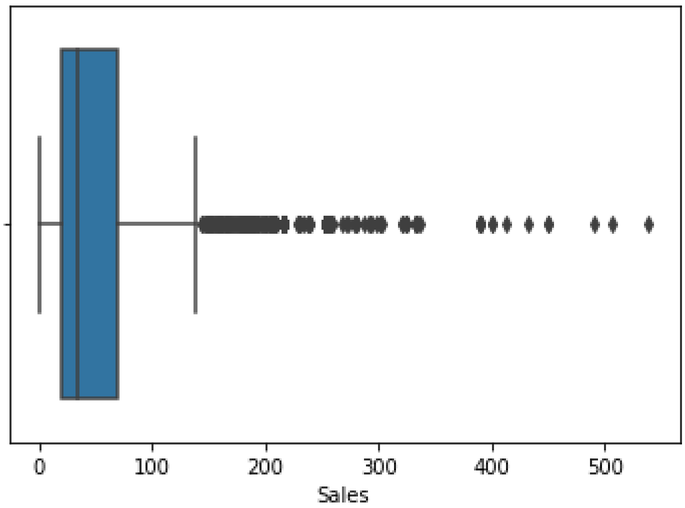


Sales: Amount of sales of tour insurance policies

- The amount of sales of tour insurance policies ranges from 0 to 539
- Mean of amount of sales of tour insurance policies is 60.25 and median is 33. Skewness is 2.4.
- Mean is greater than median, indicates that the distribution is right tailed
- Outliers present in this attribute

```
Description of Sales
---------------
count     3000.000000
mean        60.249913
std         70.733954
min          0.000000
25%         20.000000
50%         33.000000
75%         69.000000
max        539.000000
```



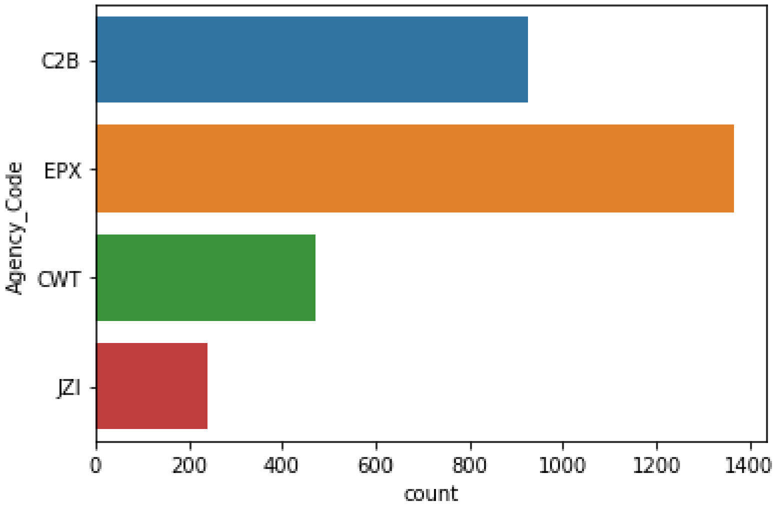


Univariate Analysis for Categorical Attributes

Agency_Code: Code of Tour firm

- There are 4 types of tour agency firm operates
- Agency code with EPX claims maximum 45.5% and code with JZI claims minimum

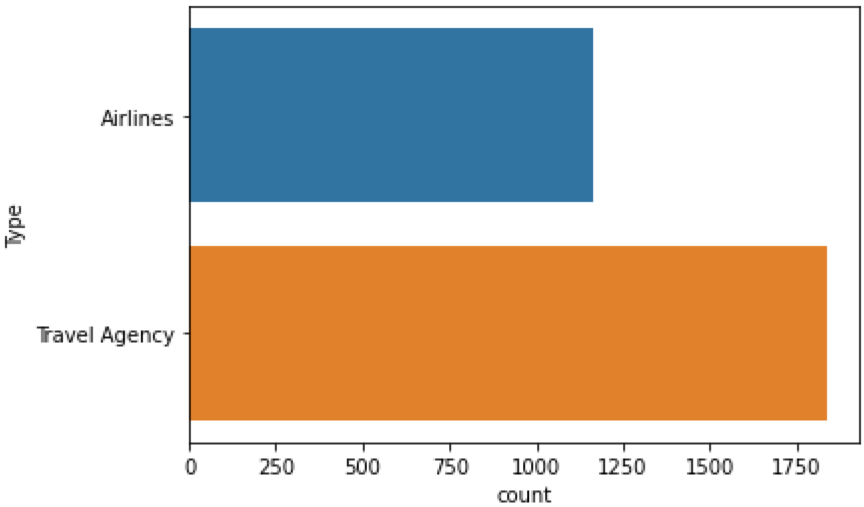
```
Percentage Value counts of Agency_Code
------------------------
EPX     0.455000
C2B     0.308000
CWT     0.157333
JZI     0.079667
```



Type: Type of tour insurance firms

- There are 2 types of tour insurance firm operates
- Travel agency types claims maximum 61.2% and Airlines types claims minimum 38.8%

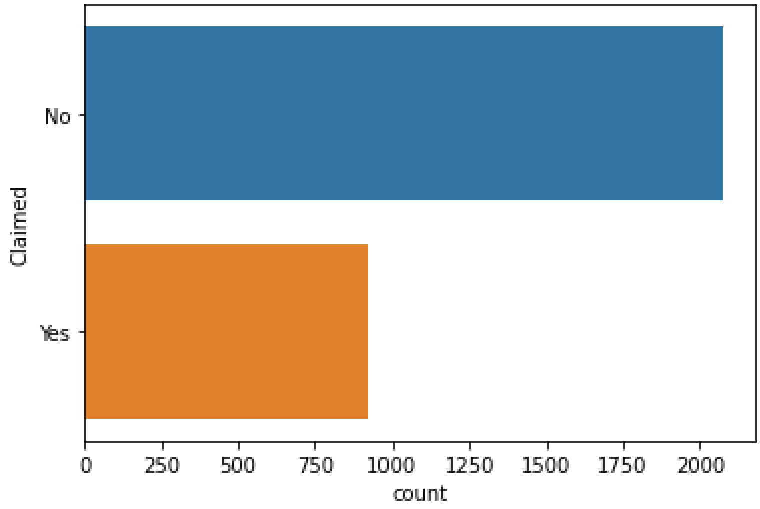
```
Percentage Value counts of Type
------------------------
Travel Agency    0.612333
Airlines         0.387667
```



Claimed: Claim Status

- There are 2 types of claim status 'Yes' and 'No'
- Claim status 'No' claims maximum 69.2%
- Claim status 'Yes' claims minimum 30.8%

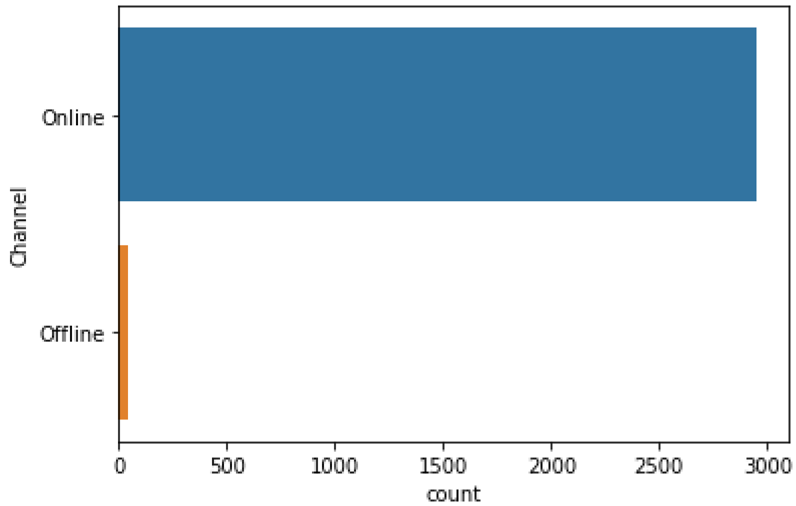
```
Percentage Value counts of Claimed
------------------------
No     0.692
Yes    0.308
```



Channel: Distribution channel of tour insurance agencies

- There are 2 types of Distribution channel of tour insurance agencies operates
- Online distribution channel of tour insurance agencies claims maximum 98.5%
- Offline distribution channel of tour insurance agencies claims minimum 1.5%

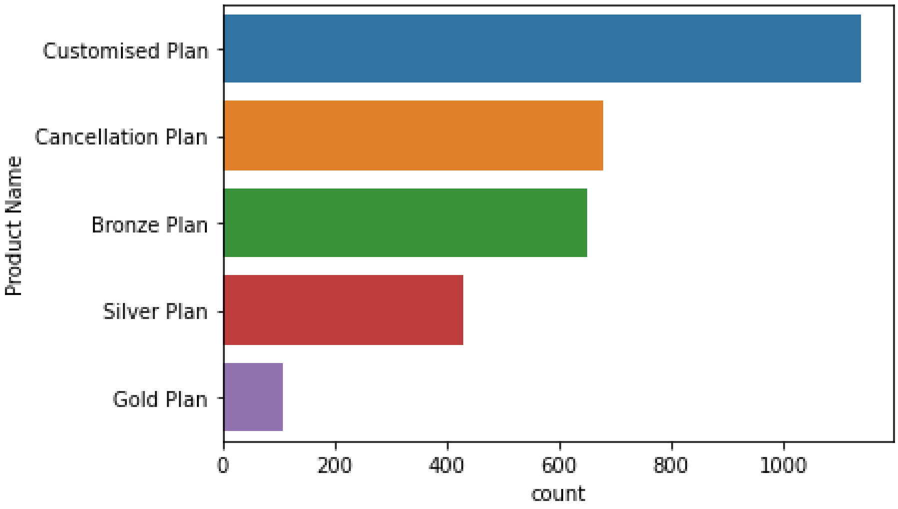
```
Percentage Value counts of Channel
------------------------
Online     0.984667
Offline    0.015333
```



Product Name: Name of the tour insurance products

- There are 5 types of tour insurance products exists
- Customized Plan tour insurance products claims maximum 37.9%
- Gold Plan tour insurance products claims minimum 3.6%

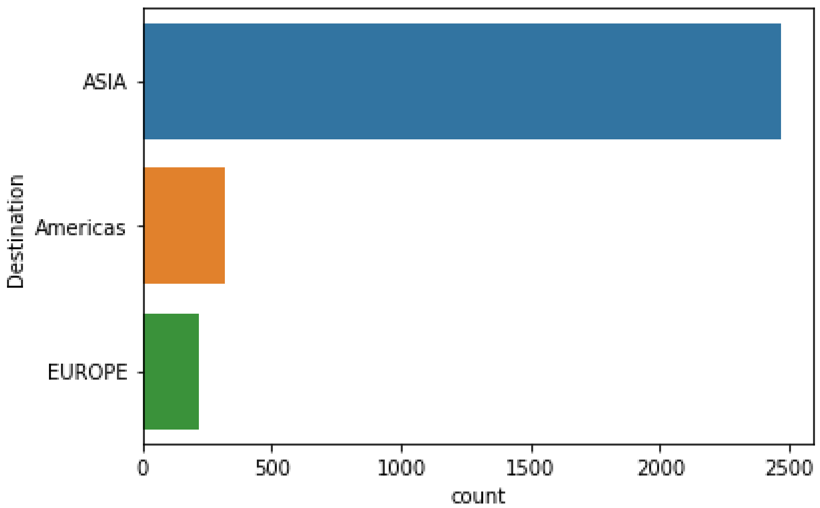
```
Percentage Value counts of Product Name
------------------------
Customised Plan      0.378667
Cancellation Plan    0.226000
Bronze Plan          0.216667
Silver Plan          0.142333
Gold Plan            0.036333
```



Destination: Destination of the tour

- There are 3 types of destination of the tour exists
- Asia for Destination of the tour claims maximum 82.2%
- Europe for Destination of the tour claims minimum 7.2%

```
Percentage Value counts of Destination
------------------------
ASIA        0.821667
Americas    0.106667
EUROPE      0.071667
```



Multivariate-Bivariate analysis

Heat map shows the correlation between different numeric attributes by assigning numbers as well as colors and Pair plot gives a graphical representation of correlation between different numeric attributes.

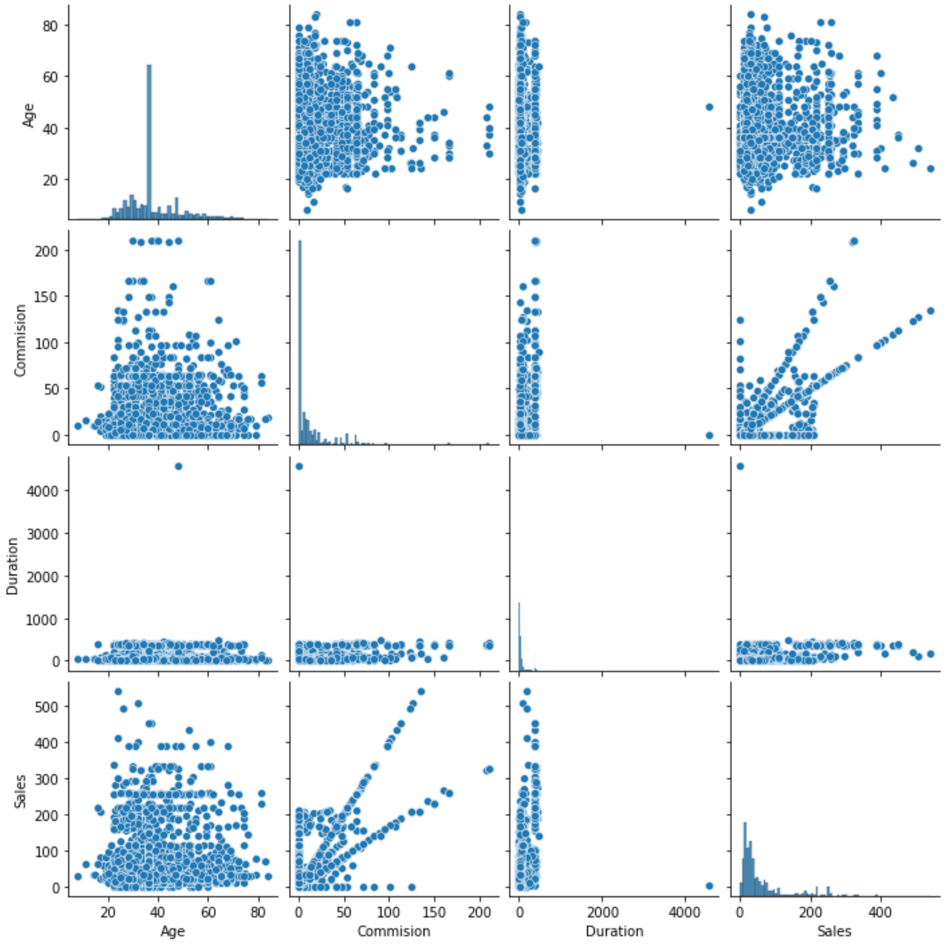


Figure 29: Pair Plot of numeric attributes

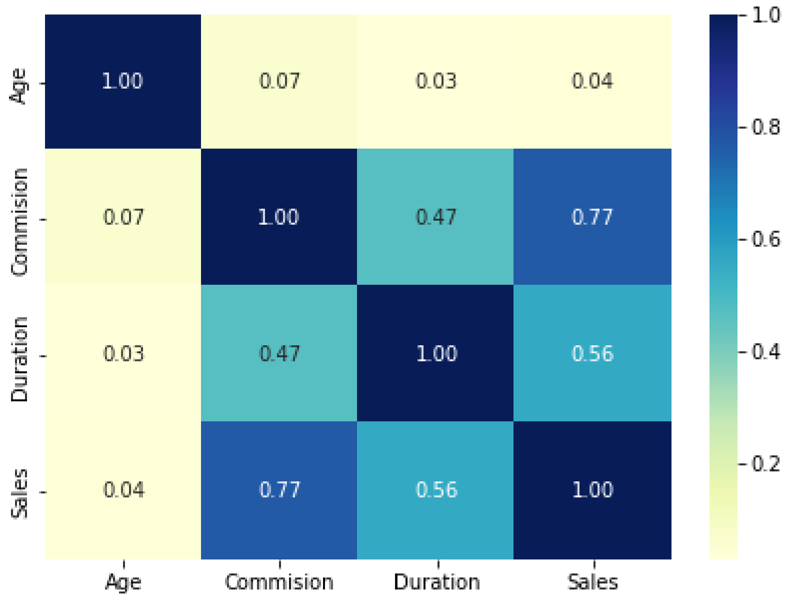


Figure 30: Heatmap of numeric attributes

- A positive correlation (0.77) can be seen between the commission and Sales, which infers as the amount of tour insurance increase, the commission received for tour insurance firms also increases

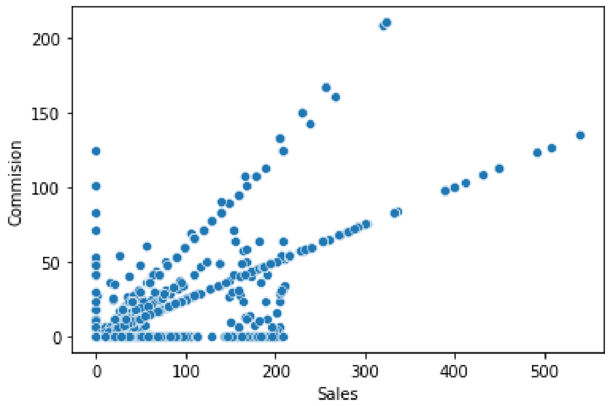


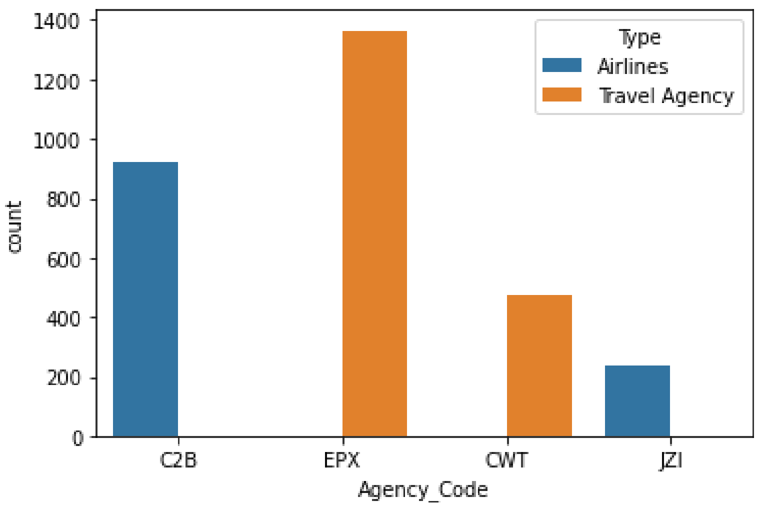
Figure 31: Pair plot between Sales and Commission

Coutplot and Crosstab between categorical variables executed and below are the observations

Bivariate analysis of Agency_Code with all other categorical variables

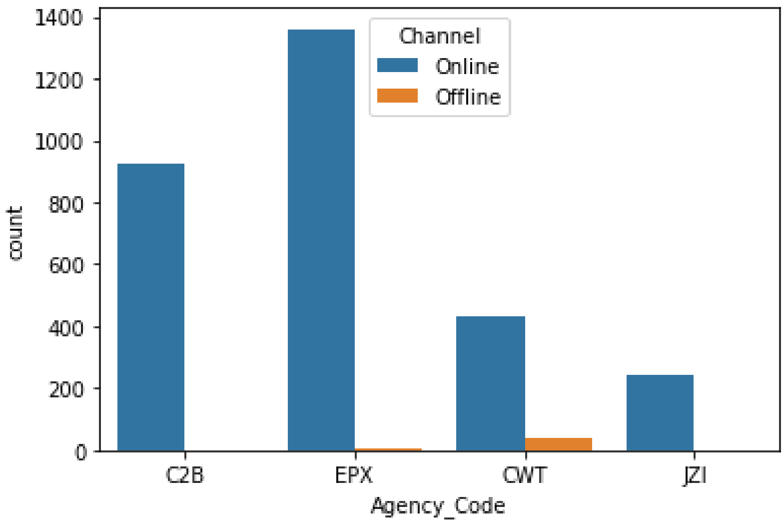
I. C2B and JZI tour firms are of Airlines type (1163) whereas EPX and CWT tour firms are Travel Agency type (1837)

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



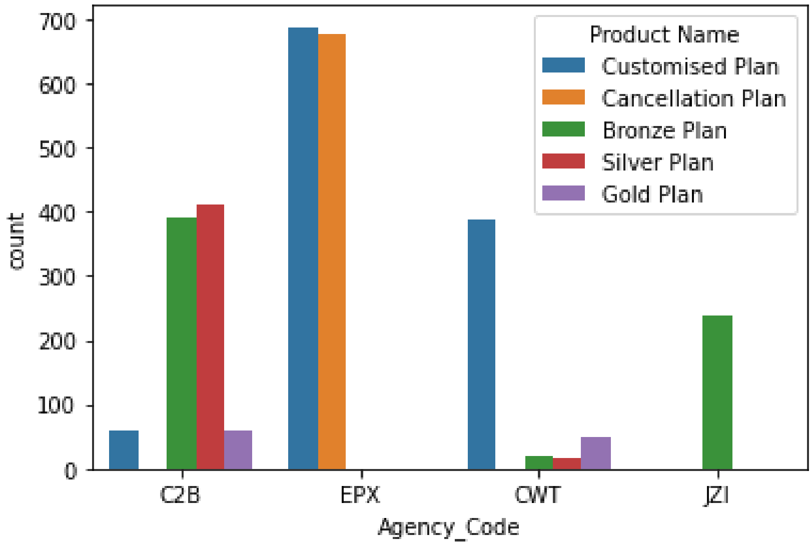
II. C2B and JZI tour firms only operates online distribution channel of tour insurance agencies (2954) and EPX and CWT tour firms operates both online and offline distribution channel of tour insurance agencies (46).

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



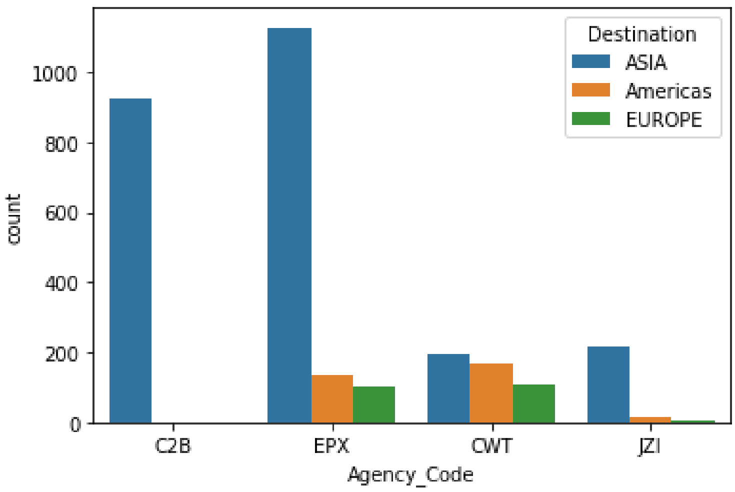
III. Bronze plan tour insurance products offers by C2B, CWT, JZI tour firms. Cancellation Plan product is only offers by EPX among all the 4 tour firms. Customized Plan Product is offers by all the 3 tour firms except JZI. Gold Plan and Silver Plan products offer only by 2 tour firms, C2B and CWT.

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



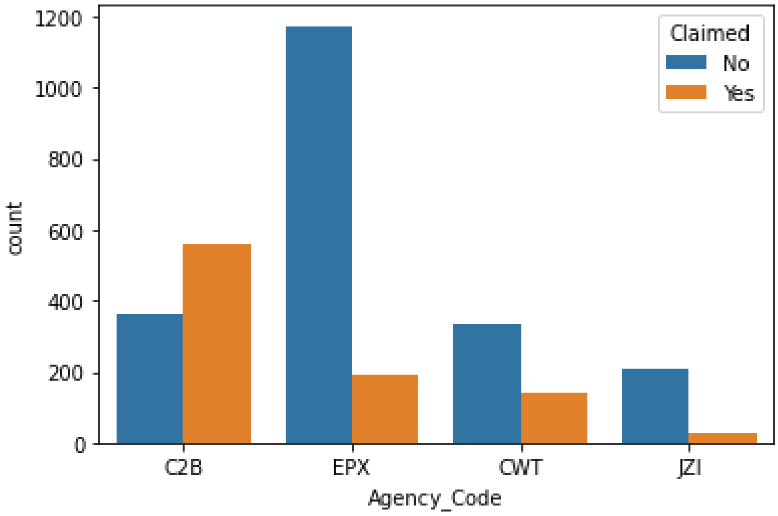
IV. All the 4 tour firms offer tours in all destination except C2B. C2B offers tour in destination to Asia only

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



V. All the 4 tour firms claims both Yes and No. C2B claims maximum Yes and EPX claims maximum no whereas JZI claims minimum claim status in both Yes and No

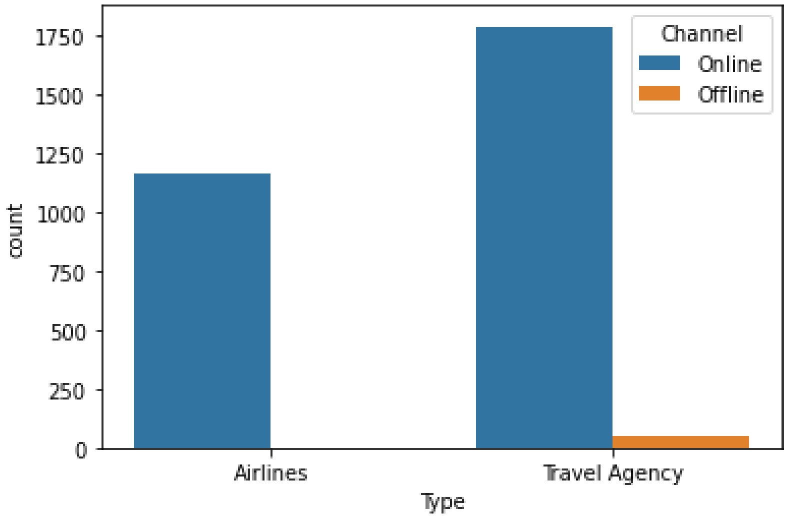
Claimed	No	Yes	All
Agency_Code			
C2B	364	560	924
CWT	331	141	472
EPX	1172	193	1365
JZI	209	30	239
All	2076	924	3000



Bivariate analysis of Type with other categorical variables

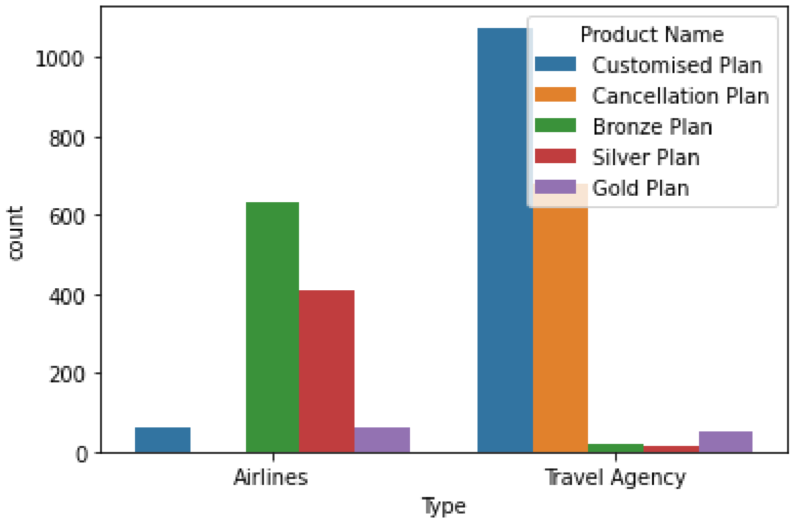
I. Airlines tour type only operates online distribution channel of tour insurance agencies whereas, Travel agency tour type operates both offline and online distribution channel of tour insurance agencies.

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

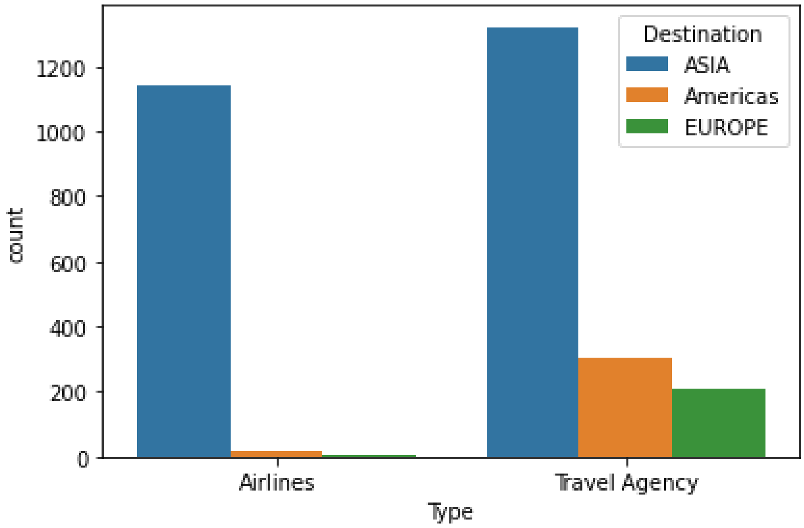


II. Bronze plan, Customized Plan, Gold plan and silver plan tour insurance products offer by both Airlines and Travel agency tour types. Airlines tour type doesn't offer Cancellation Plan.

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



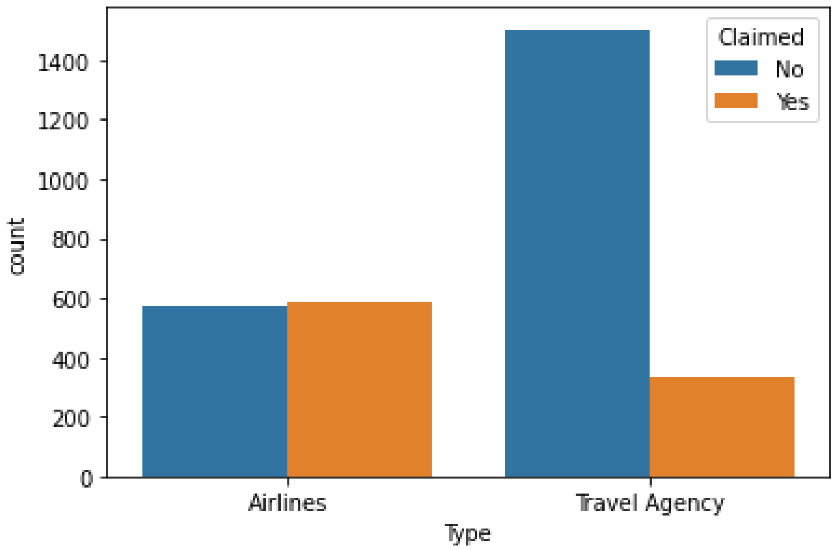
III. Both the Tour type offer all three tour destinations Asia, America and Europe. Tour Destination Asia seems more popular and Europe seems least popular in both the tour type.



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

IV. Both the tour type claims both Yes and No. Airlines tour type claims maximum Yes and Travel Agency tour type claims maximum No.

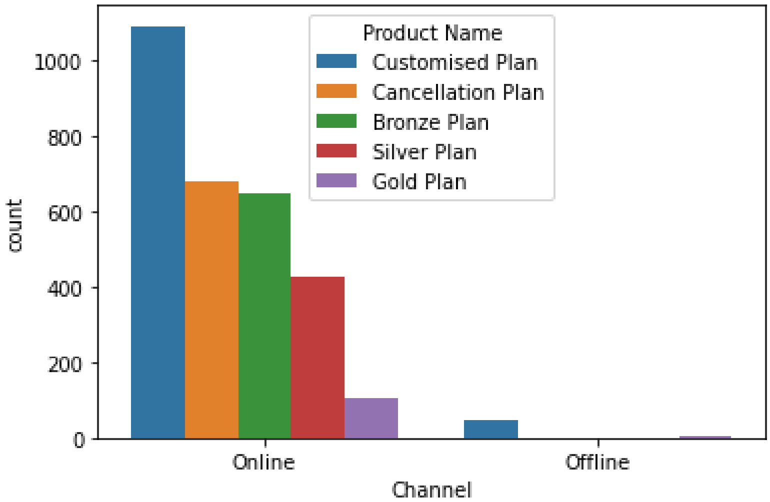
Claimed Type	No	Yes	All
Airlines	573	590	1163
Travel Agency	1503	334	1837
All	2076	924	3000



Bivariate analysis of Channel with other categorical variables

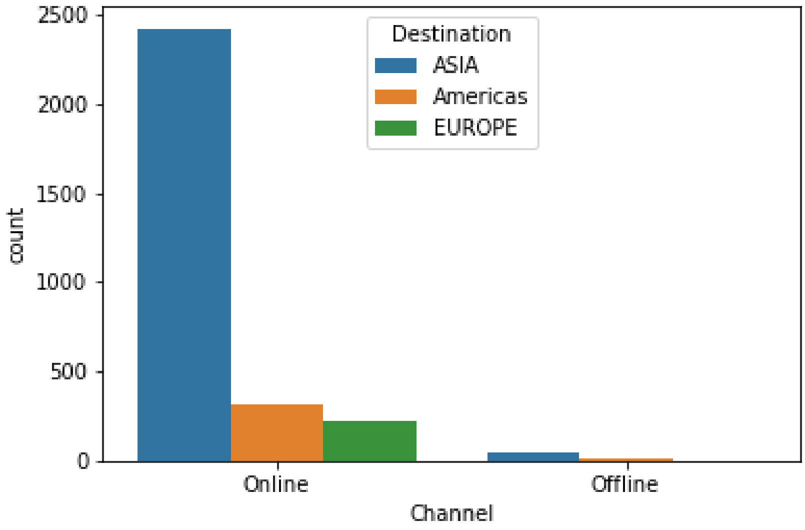
I. Online distribution channel of tour insurance agencies offers all the tour insurance products and offline distribution channel of tour insurance agencies offers only customized plan and gold plan

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



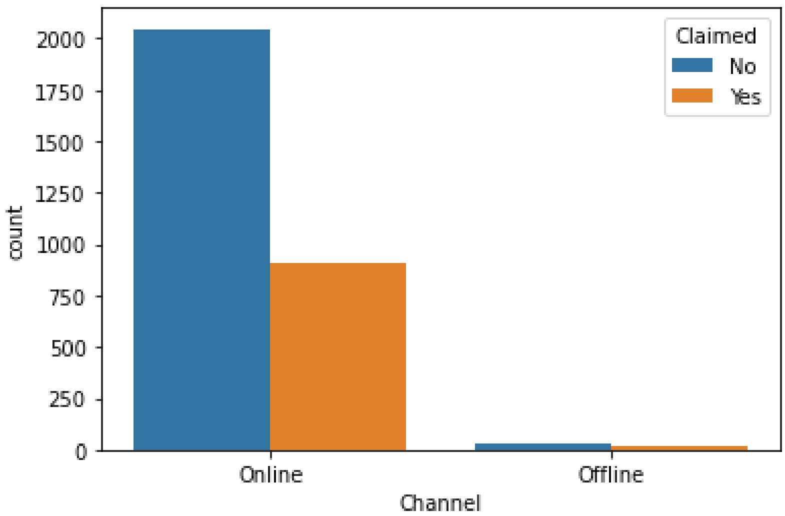
II. Online distribution channel of tour insurance agencies offers all the three-tour destinations and offline distribution channel of tour insurance agencies offers only two tour destination Asia and America. Asia seems more popular destination in both Online and Offline distribution channel of tour agencies

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



III. Both the distribution channel of tour insurance agencies claims both Yes and No. Online Channel claims both maximum Yes and No

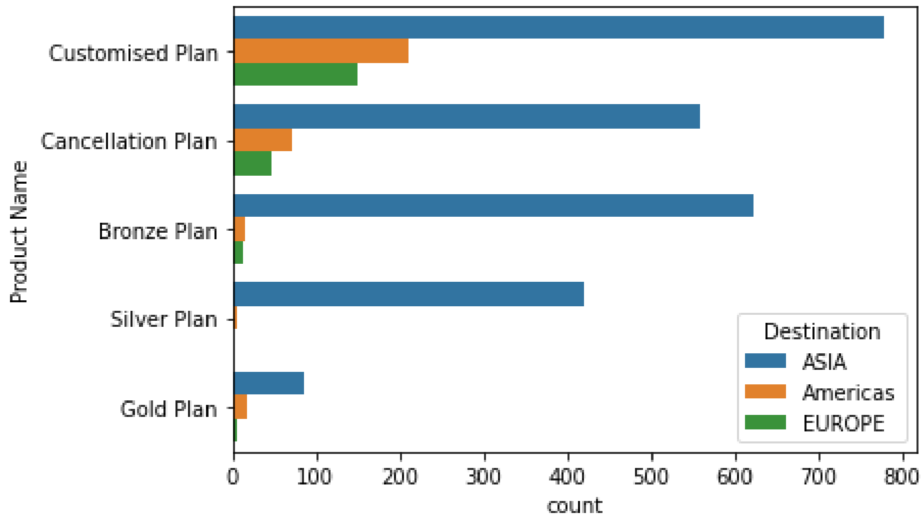
Claimed	No	Yes	All
Channel			
Offline	29	17	46
Online	2047	907	2954
All	2076	924	3000



Bivariate analysis of Product Name with other categorical variables

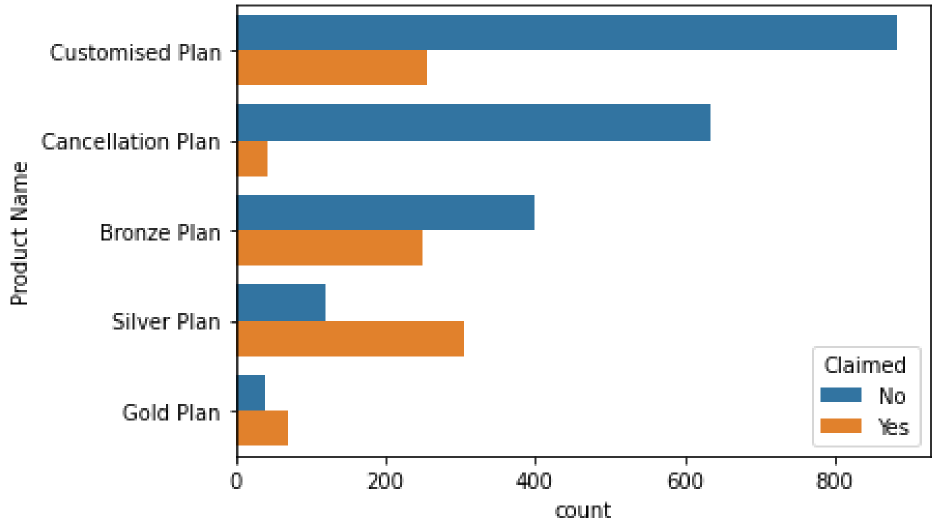
I. Tour Destination Asia is the maximum preferred across all the tour insurance products

Destination Product Name	ASIA	Americas	EUROPE	All
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



II. Tour insurance product- silver plan claims maximum Yes whereas, customized plan claims maximum no. Cancellation Plan claims minimum Yes and Gold plan claims minimum No.

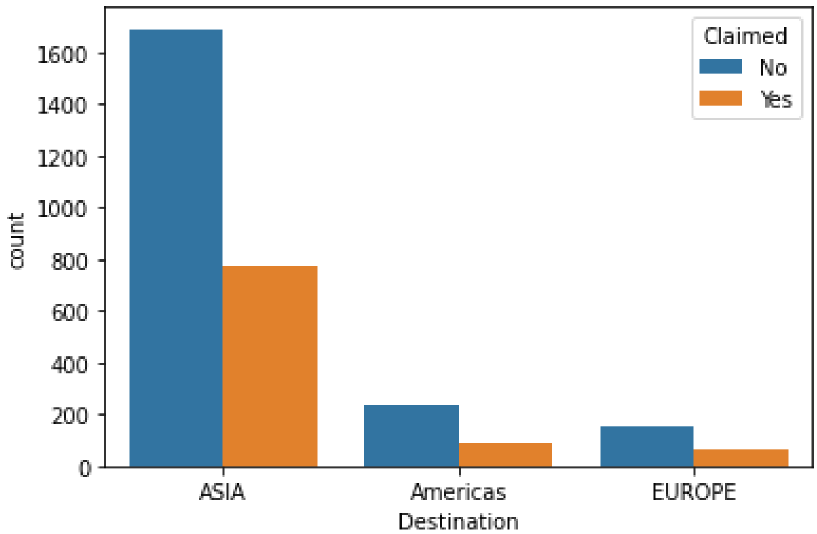
Claimed Product Name	No	Yes	All
Bronze Plan	399	251	650
Cancellation Plan	635	43	678
Customised Plan	882	254	1136
Gold Plan	39	70	109
Silver Plan	121	306	427
All	2076	924	3000



Bivariate analysis of Destination with Claimed variable

I. All three destination claims both the status Yes and No. Destination Asia claims both maximum Yes and No and Europe claims both minimum Yes and No

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

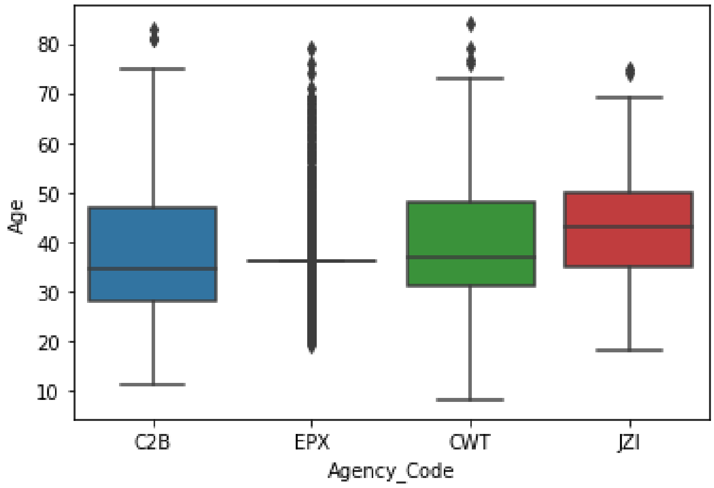


Bivariate analysis of num-cat attributes

Age with other Categorical attributes

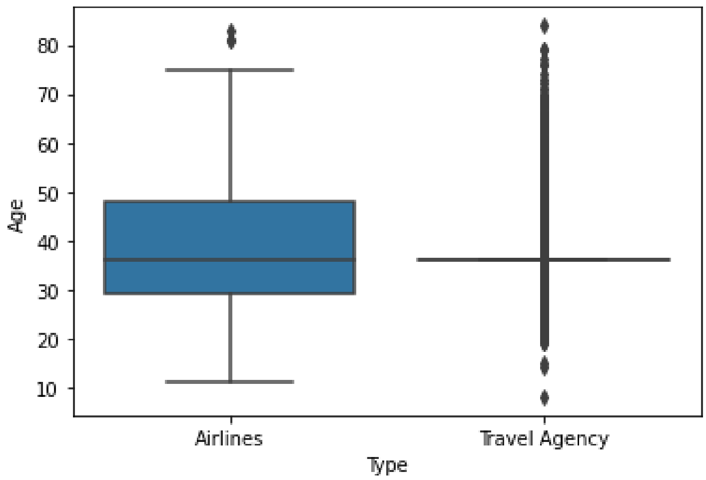
I. JZI tour firms has maximum mean age of Insured and EPX has minimum mean age of insured

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



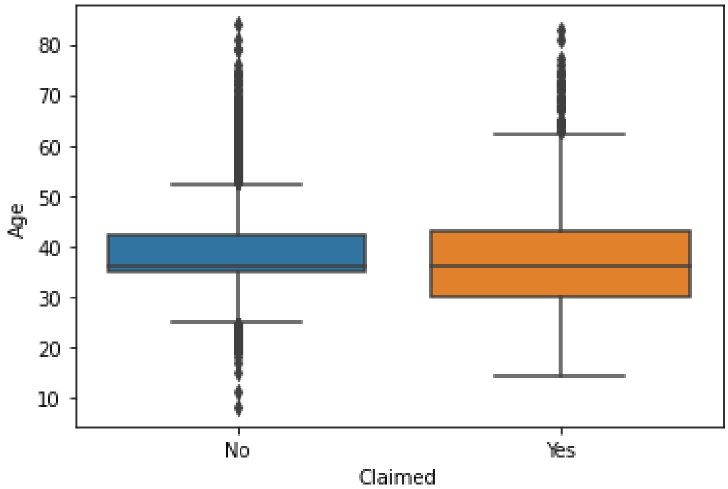
II. Airlines tour type has greater mean age of insured compare to Travel Agency

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



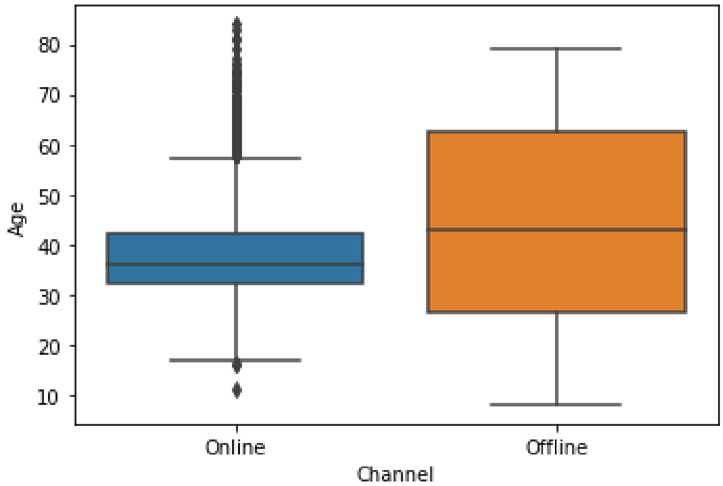
III. No claim status claims by greater mean age of insured than yes claim status. There average age of insured is almost comparable

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



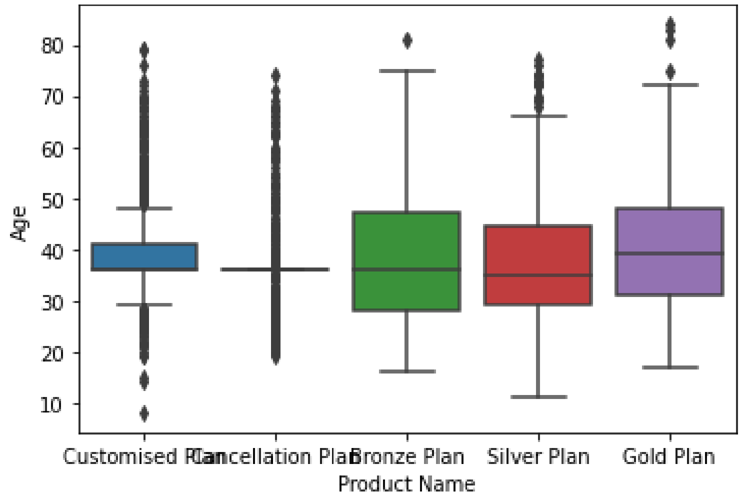
IV. Mean age of insured is maximum in Offline distribution channel of tour insurance agencies than online distribution channel

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



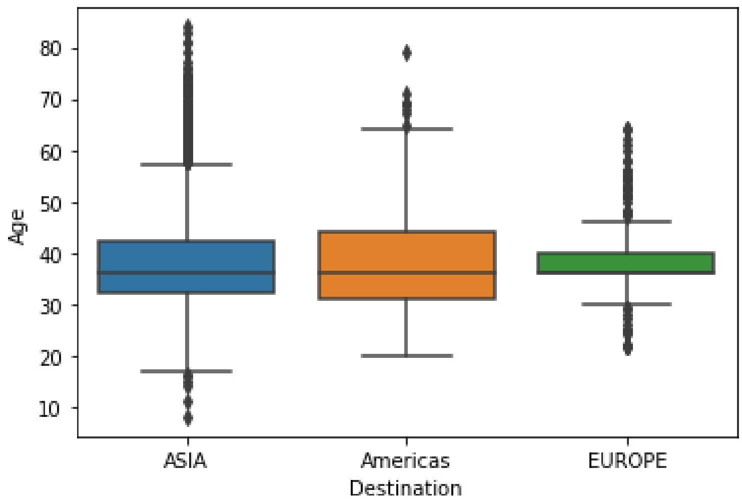
V. Tour insurance product- gold plan has maximum mean age of Insured, and cancellation plan has minimum age of Insure. Mean age of Insured is nearly same for Bronze Plan and Customized Plan

```
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
```



VI. All three-tour destination have almost same mean age of Insured

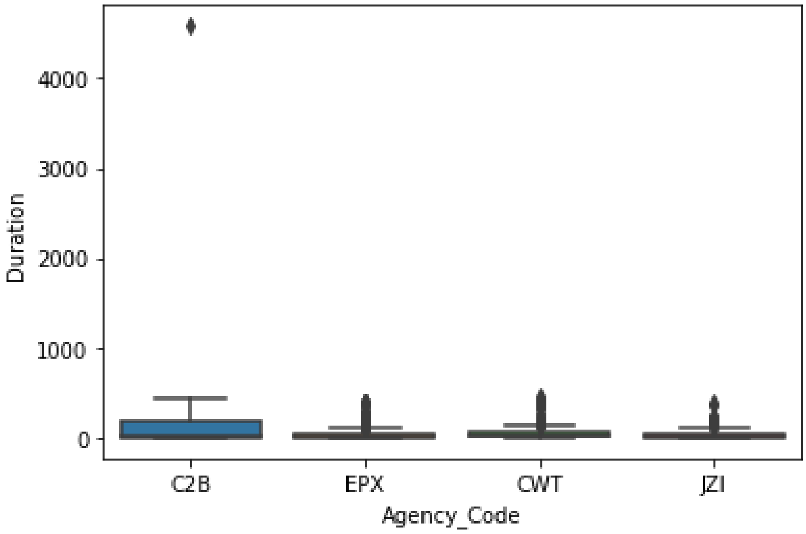
```
Mean of Age for Destination
Destination
ASIA        38.048276
Americas    38.481250
EUROPE      38.000000
Name: Age, dtype: float64
```



Duration with other Categorical attributes

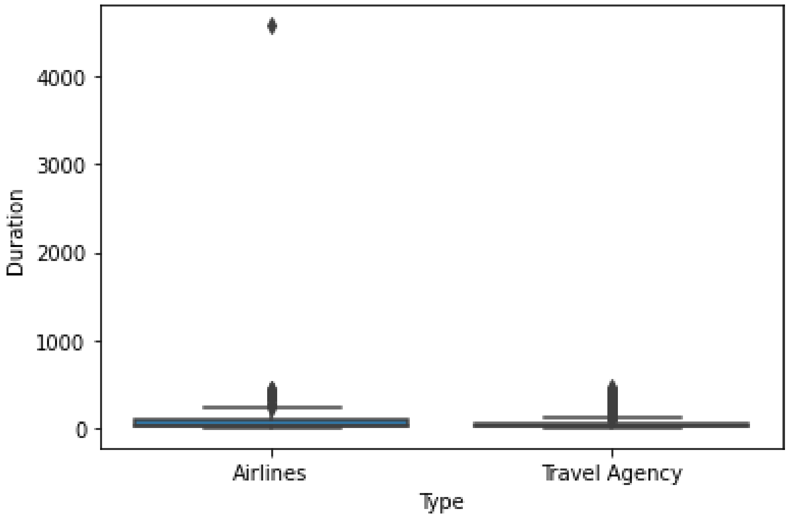
I. C2B tour firms has maximum mean duration tour and JZI has minimum mean duration tour

```
Mean of Duration for Agency_Code
Agency_Code
C2B     119.404762
CWT      64.733051
EPX      43.374359
JZI      41.485356
```



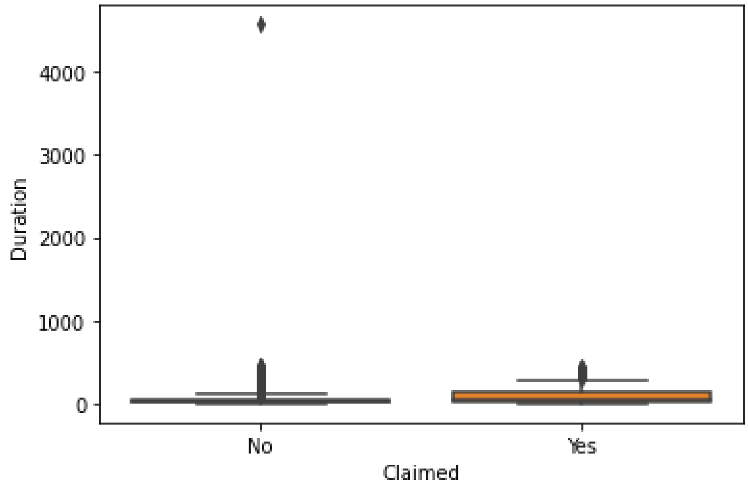
II. Airlines tour type has greater mean duration tour and Travel Agency has minimum duration tour

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



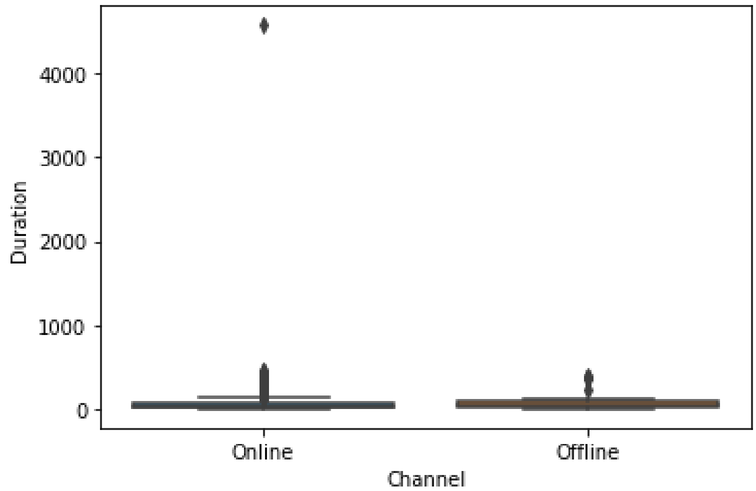
III. Mean duration tour for claim status Yes is greater than No

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



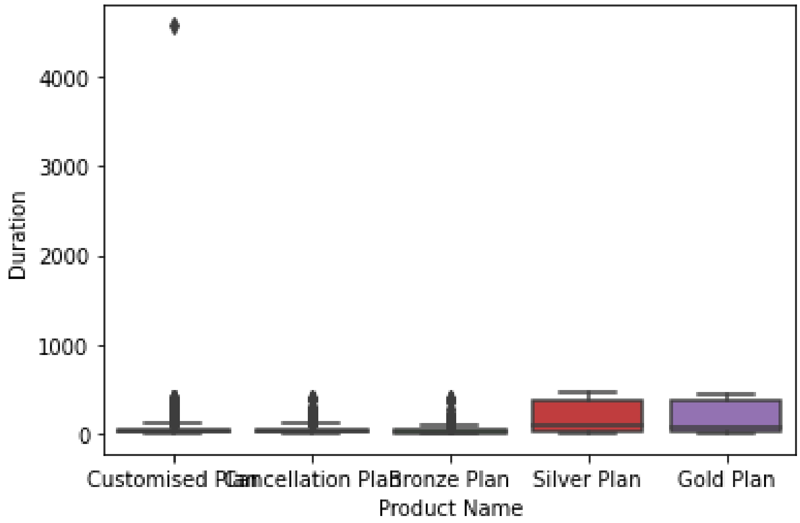
IV. Mean duration tour for Offline distribution channel is greater than online distribution channel

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



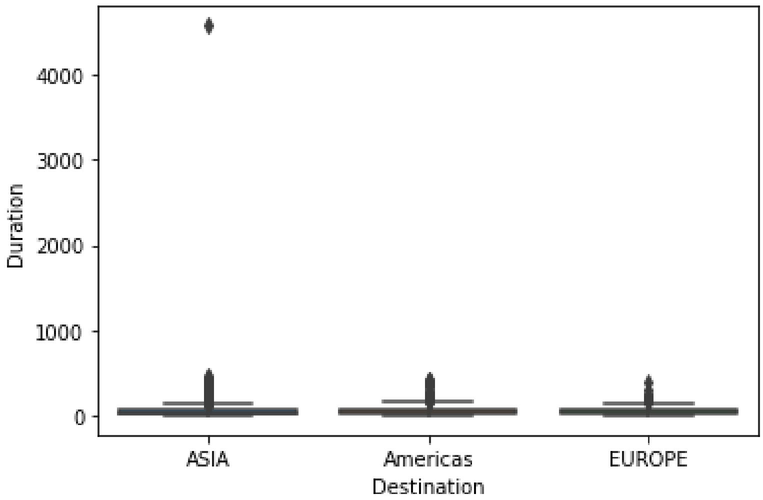
V. Mean duration tour is highest for silver-plan tour insurance product and least for bronze plan

```
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
```



VI. Mean duration tour is highest for America and lowest for Europe

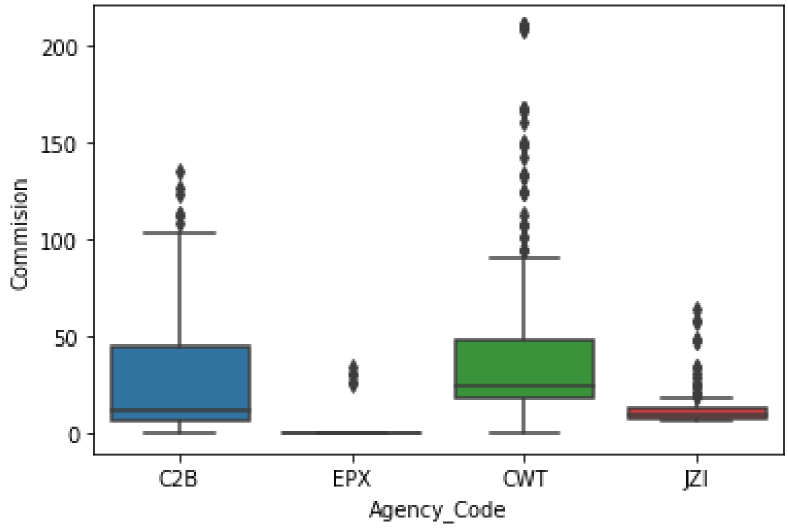
```
Mean of Duration for Destination
Destination
ASIA        70.443408
Americas    77.409375
EUROPE      53.911628
Name: Duration, dtype: float64
```



Commission with other Categorical attributes

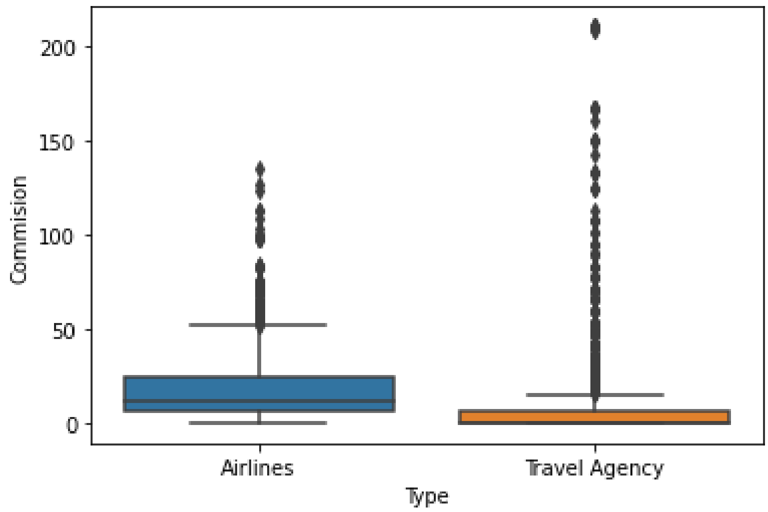
I. The commission received for tour insurance firm is highest for CWT and lowest for EPX tour firm

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



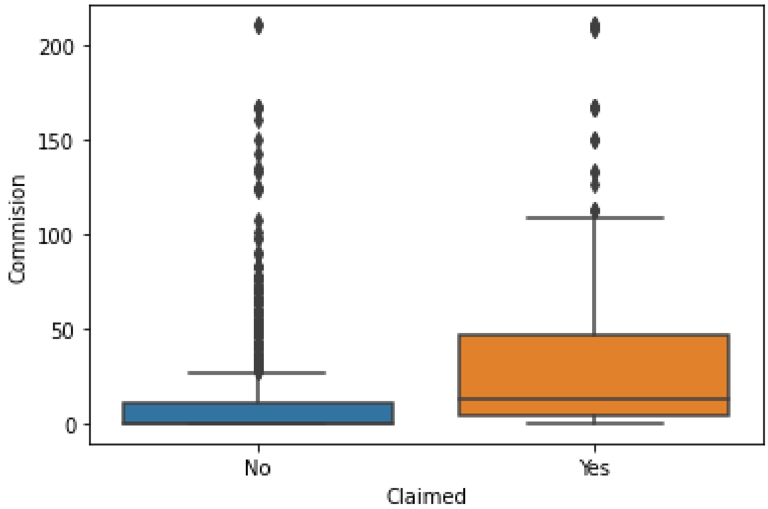
II. The commission received for tour insurance firm is greater for Airlines tour type compare to Travel Agency tour type

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



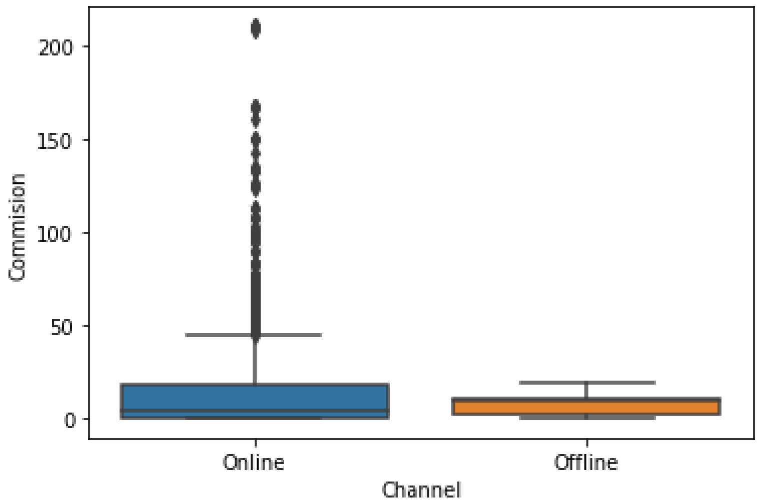
III. The commission received for tour insurance firm is greater for claim status Yes compare to No

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



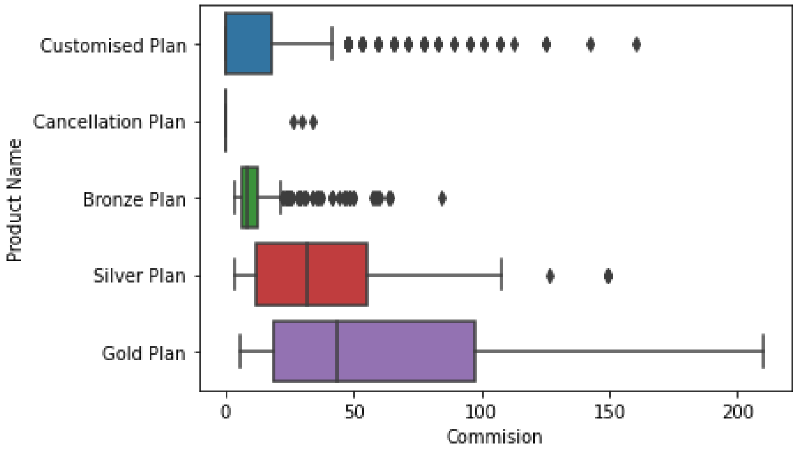
IV. The commission received for tour insurance firm for Online distribution channel is greater than offline distribution channel

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



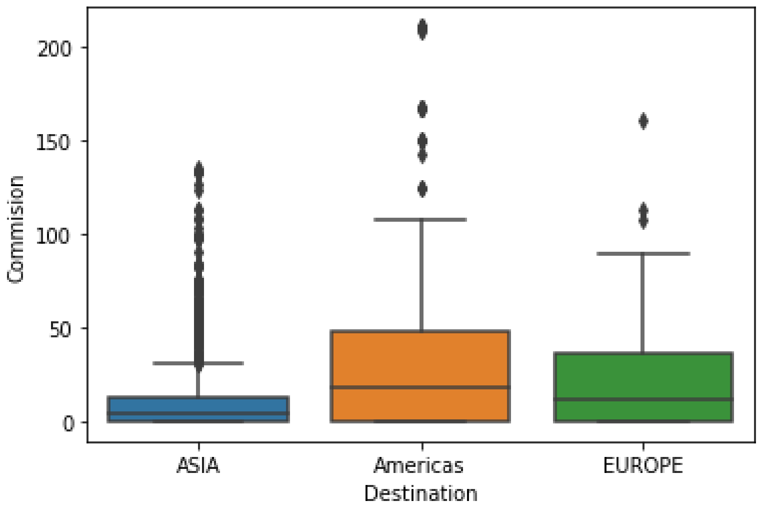
V. The commission received for tour insurance firm is highest for gold-plan tour insurance product and least for cancellation plan

```
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
```



VI. The commission received for tour insurance firm is highest for America and lowest for Asia

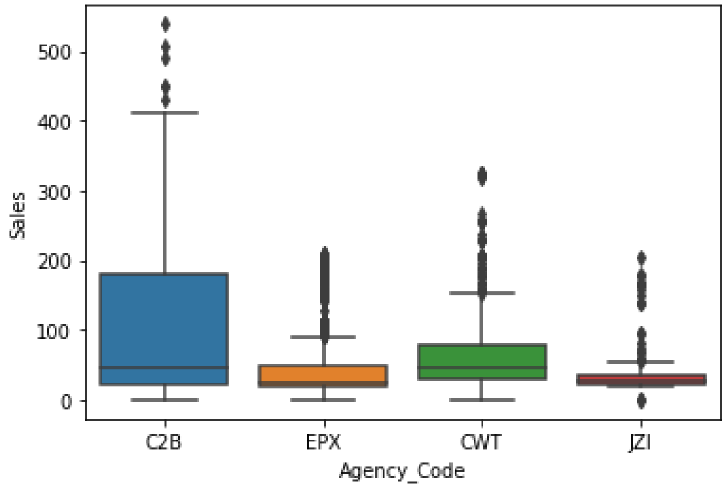
```
Mean of Commission for Destination
Destination
ASIA        11.732207
Americas    32.339906
EUROPE      20.088140
Name: Commision, dtype: float64
```



Sales with other Categorical attributes

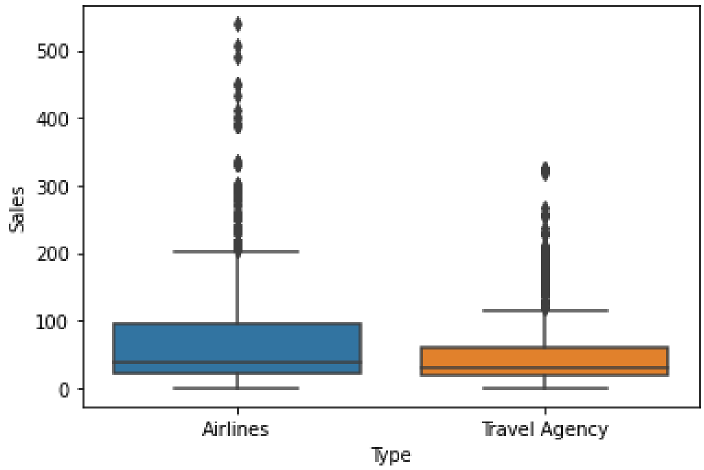
I. Amount worth of sales per customer in procuring tour insurance policies is highest for C2B and lowest for JZI tour firm

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



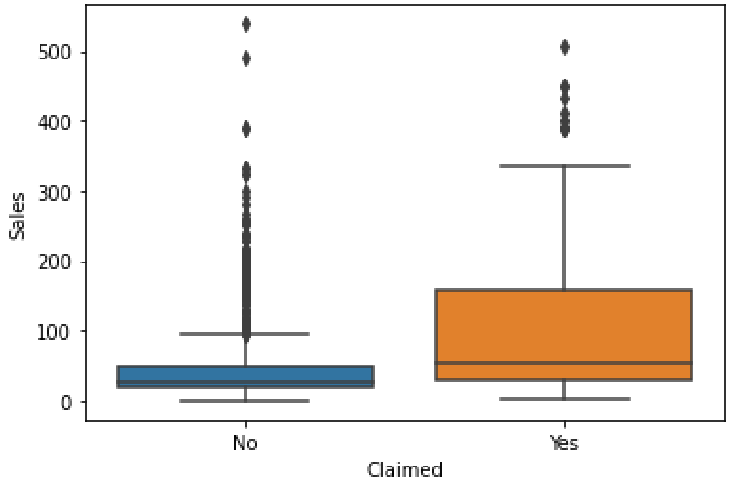
II. Amount worth of sales per customer in procuring tour insurance policies is greater for Airlines tour type compare to Travel Agency tour type

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



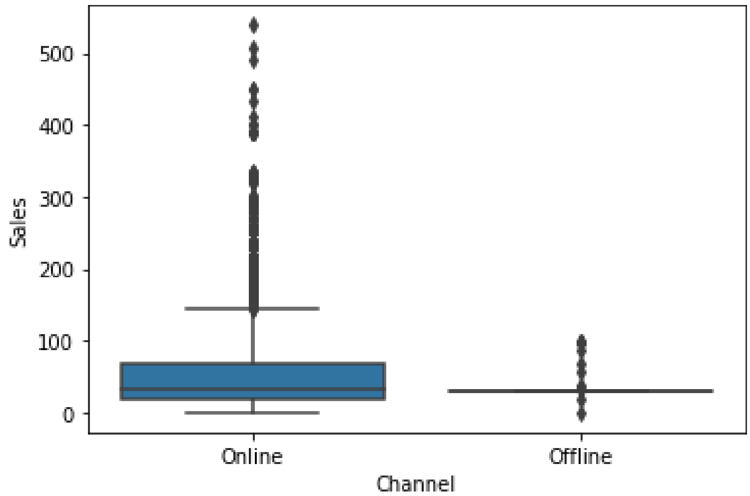
III. Amount worth of sales per customer in procuring tour insurance policies greater for claim status Yes compare to No

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



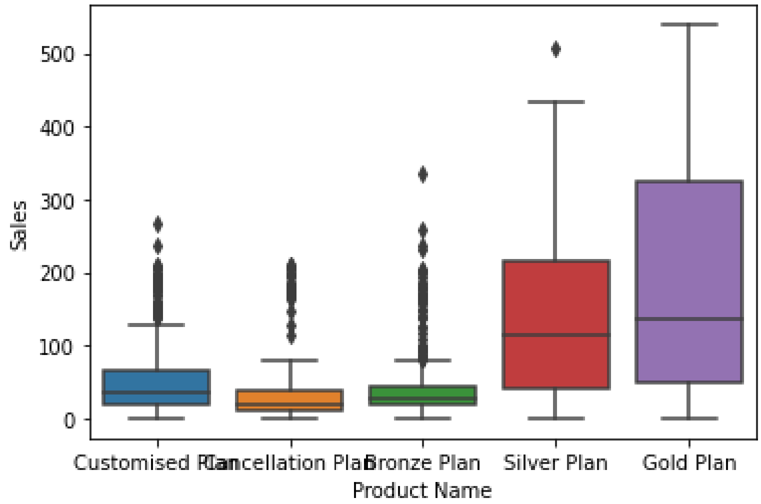
IV. Amount worth of sales per customer in procuring tour insurance policies for Online distribution channel is greater than offline distribution channel

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



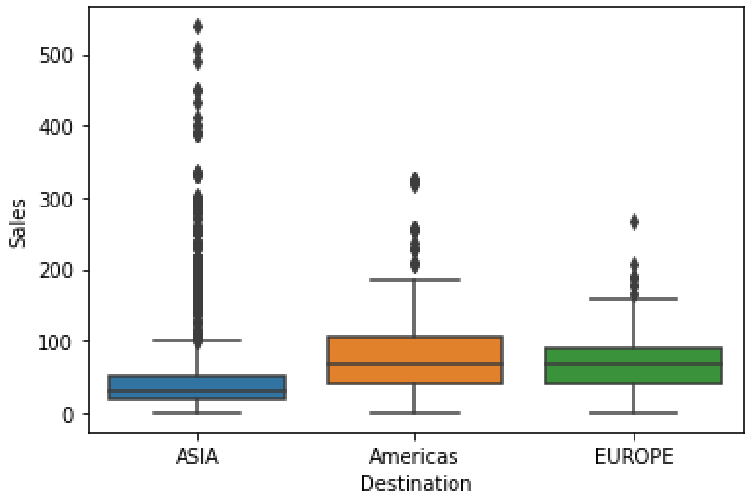
V. Amount worth of sales per customer in procuring tour insurance policies is highest for gold-plan tour insurance product and least for cancellation plan

```
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
```



VI. Amount worth of sales per customer in procuring tour insurance policies is highest for America and lowest for Asia

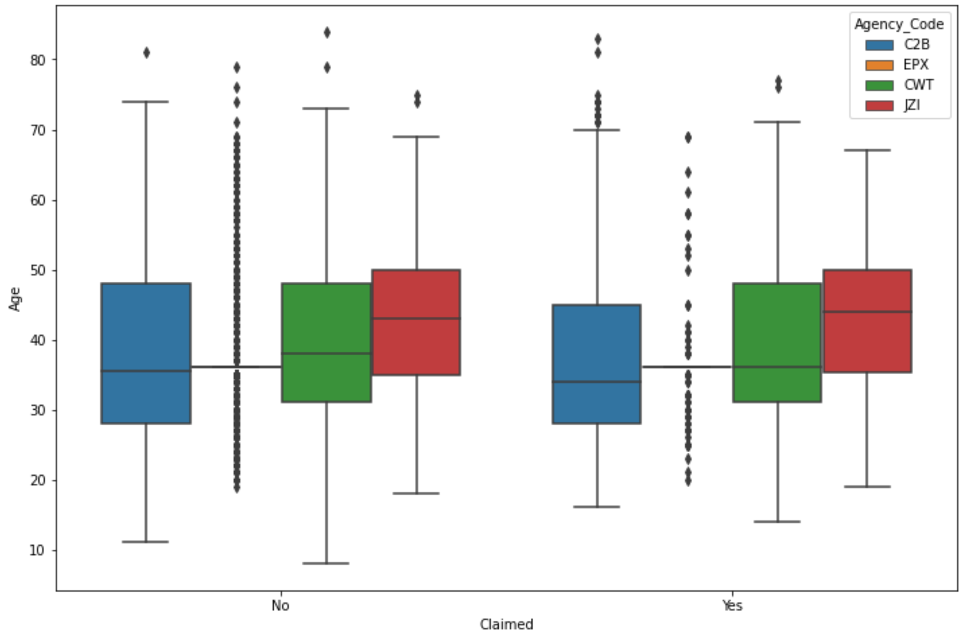
```
Mean of Sales for Destination
Destination
ASIA        56.467513
Americas    82.573281
EUROPE      70.390093
Name: Sales, dtype: float64
```



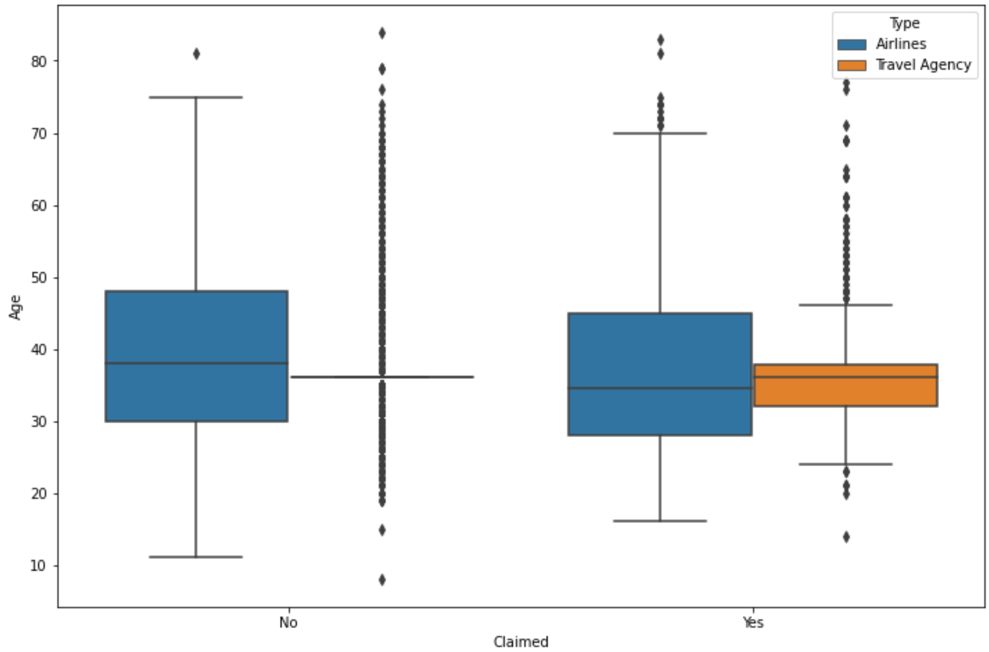
Multivariate analysis of num-cat attributes

Age-Claim- other categorical attributes

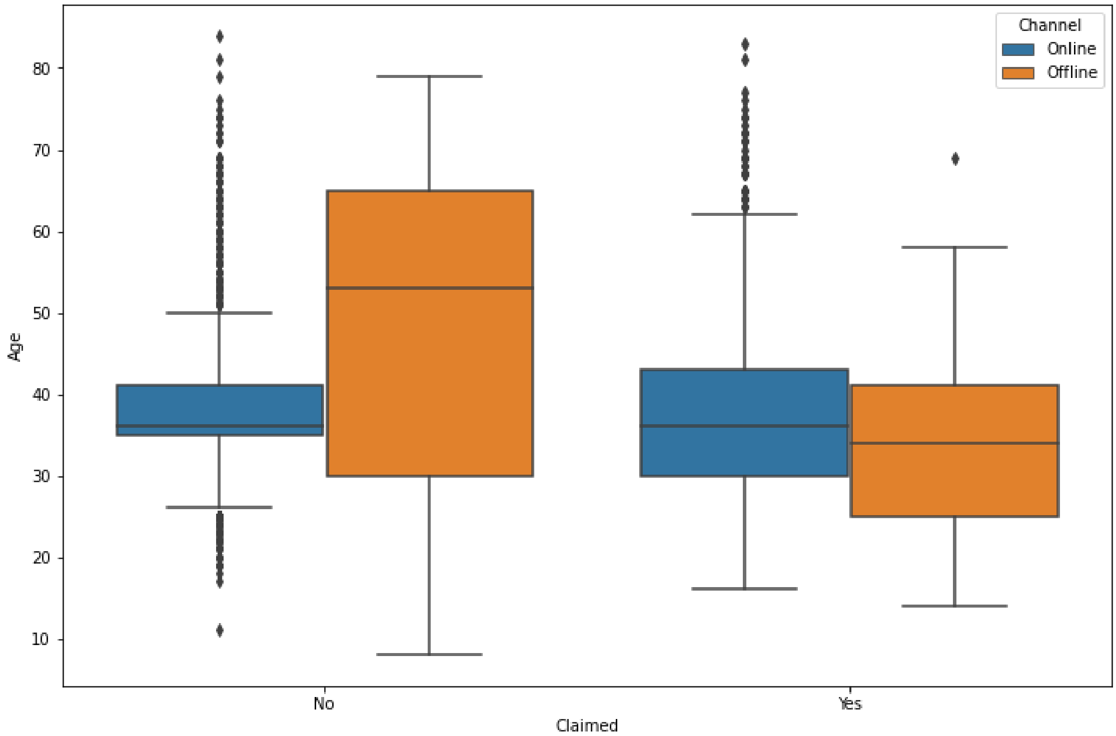
I. For both the claimed status 'Yes' and 'No' the median age of insured is highest in JZI tour firm and lowest in C2B tour firm



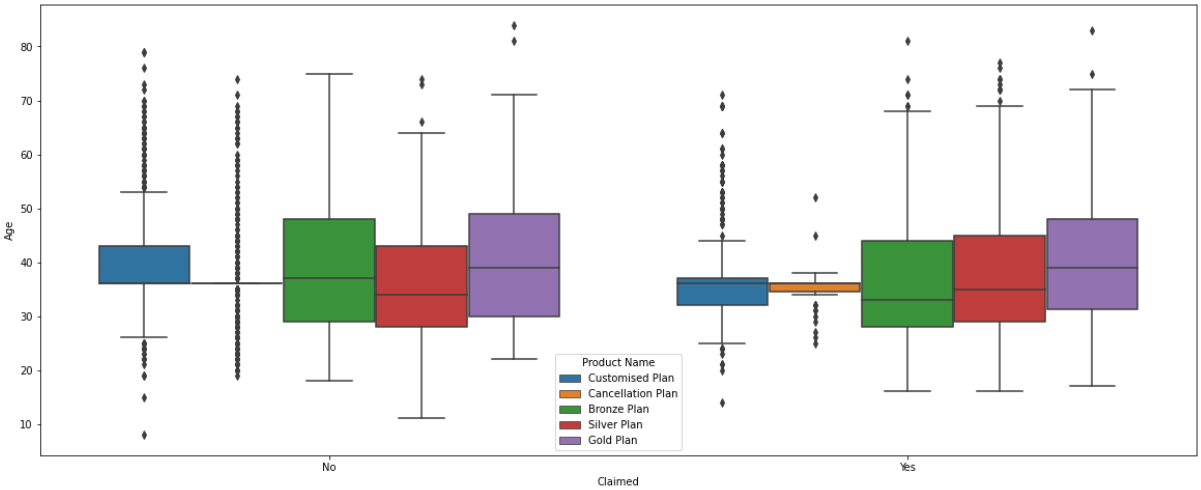
II. For claimed status 'Yes' the median age of insured is greater for travel agency tour type and for claimed status 'No' the median age of insured is lower for Airlines



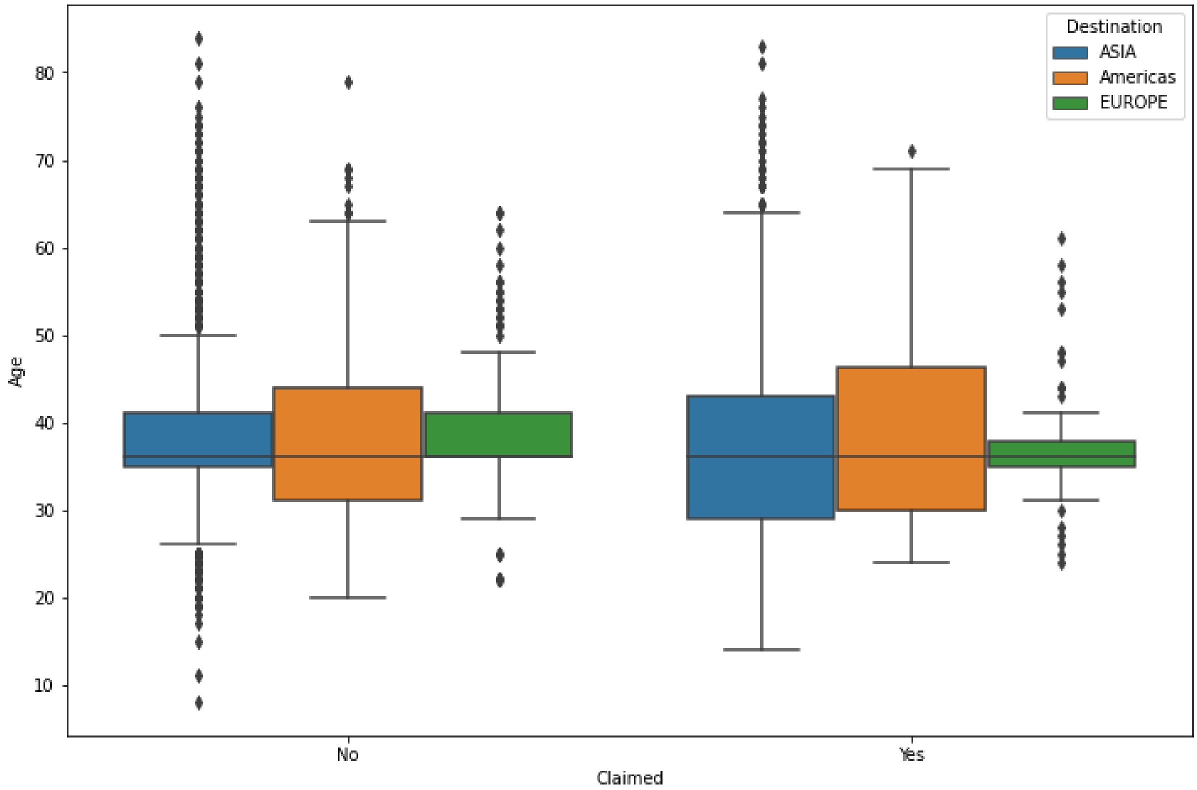
III. For claimed status 'Yes', the median age of insured is greater for Online distribution channel compared to offline. For claimed status 'No' the median age of insured is greater for Offline distribution channel compared to online



IV. For both the claimed status ‘Yes’ and ‘No’ the median age of insured is highest for Gold-Plan tour insurance product. For claimed status ‘Yes’, the median age of insured is lowest for bronze plan

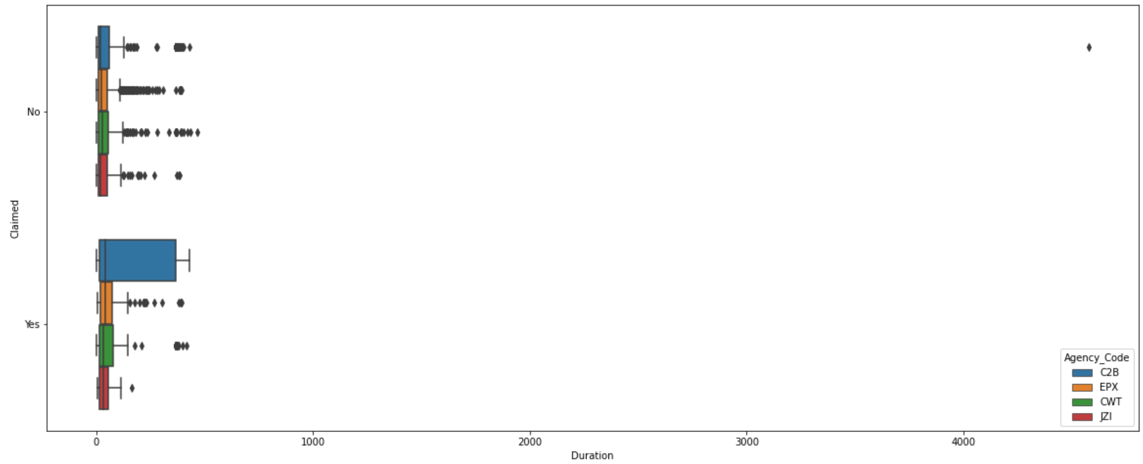


V. For both the claimed status ‘Yes’ and ‘No’ the median age of insured is almost same for all the three destinations.

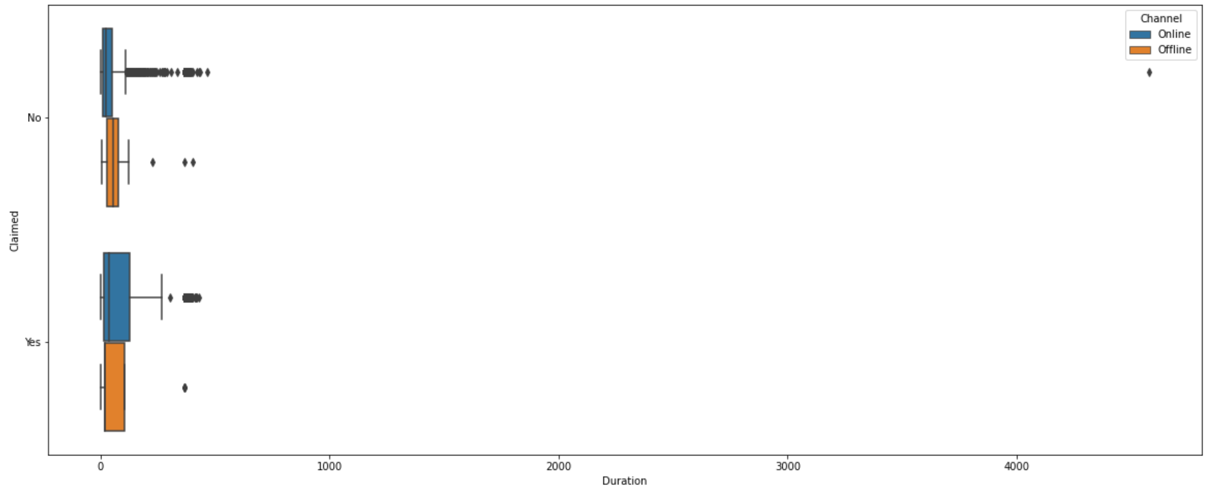


Duration-Claim- other categorical attributes

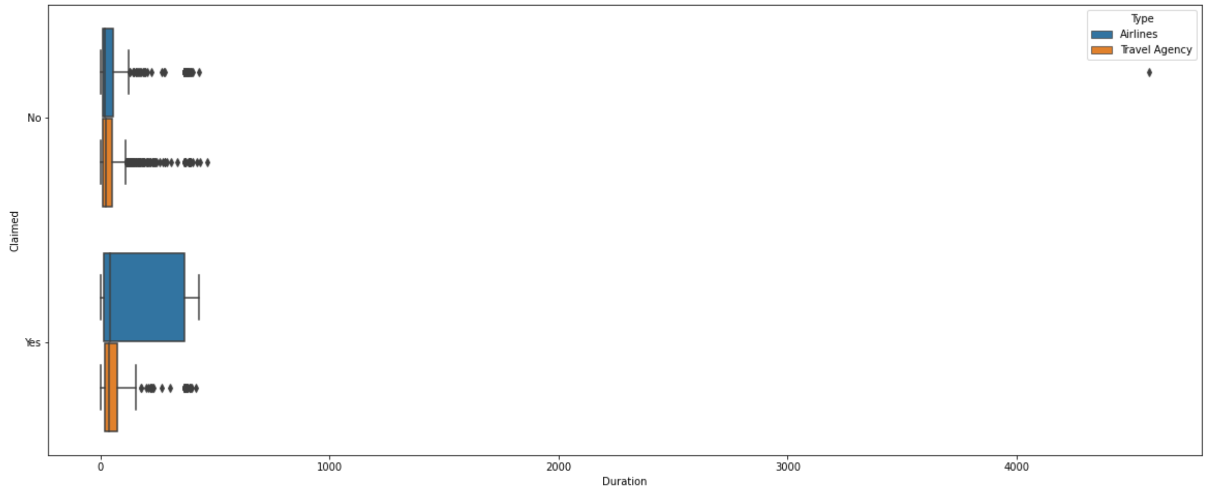
I. For both the claimed status ‘Yes’ and ‘No’ the median duration of tour is all most same for all the four tour firms



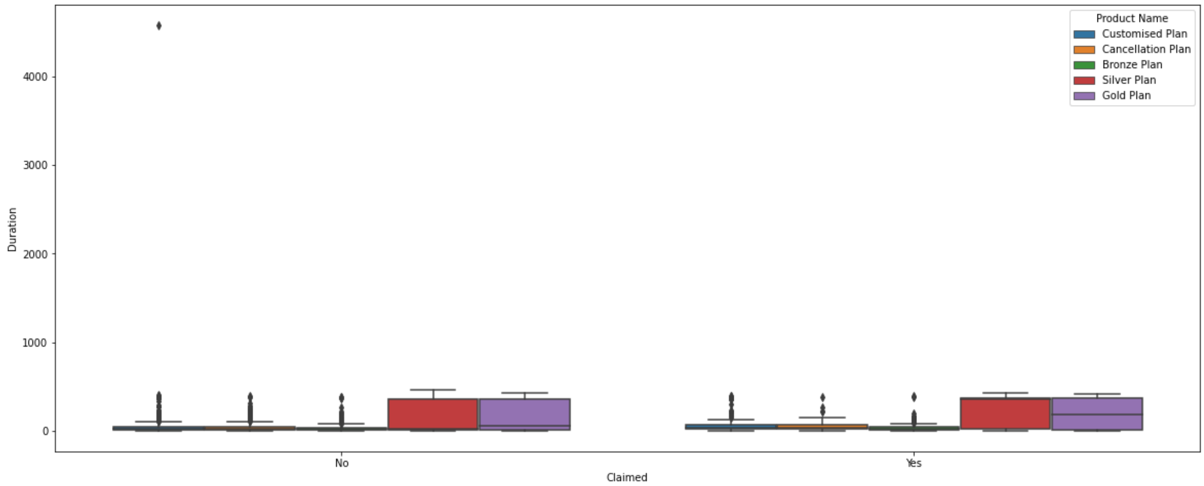
II. For claimed status ‘Yes’, the median duration of tour is greater for Online distribution channel compared to offline. For claimed status ‘No’, the median duration of tour is greater for Offline distribution channel compared to online



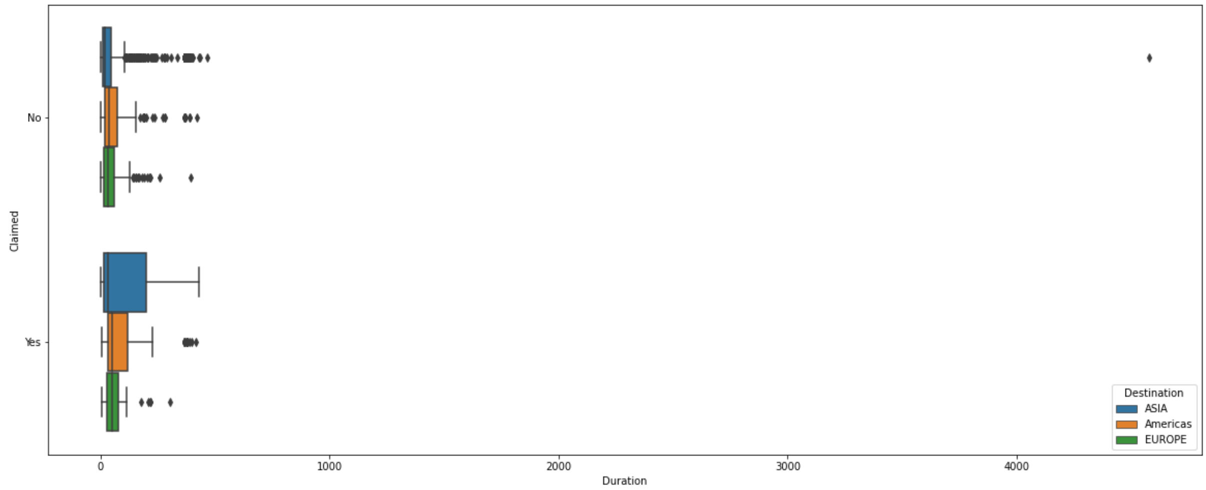
III. For Claimed status 'Yes' Both the tour type has same median



IV. For claimed status 'Yes', the median duration of tour is highest for Gold-Plan tour insurance product. For claimed status 'No' the median duration of tour is all most same for all 5 tour insurance products

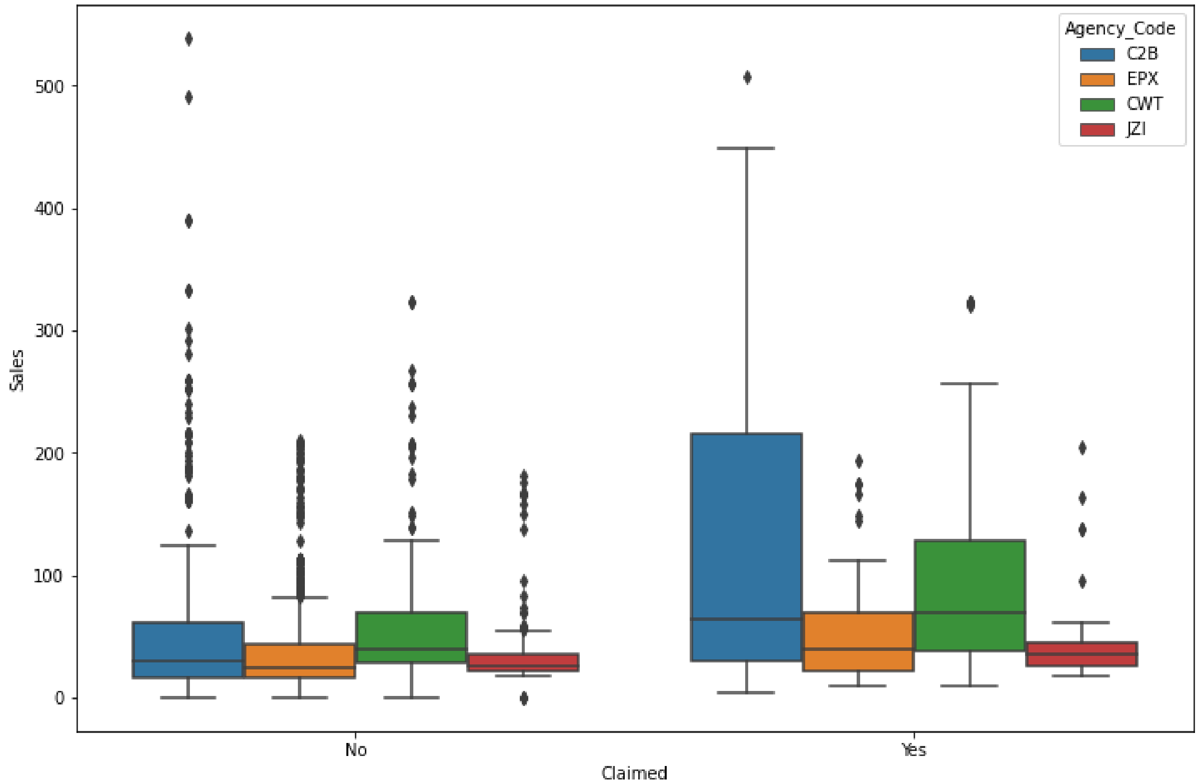


V. For both the claimed status 'Yes' and 'No' the median duration of tour is highest for tour destination America and least for Asia

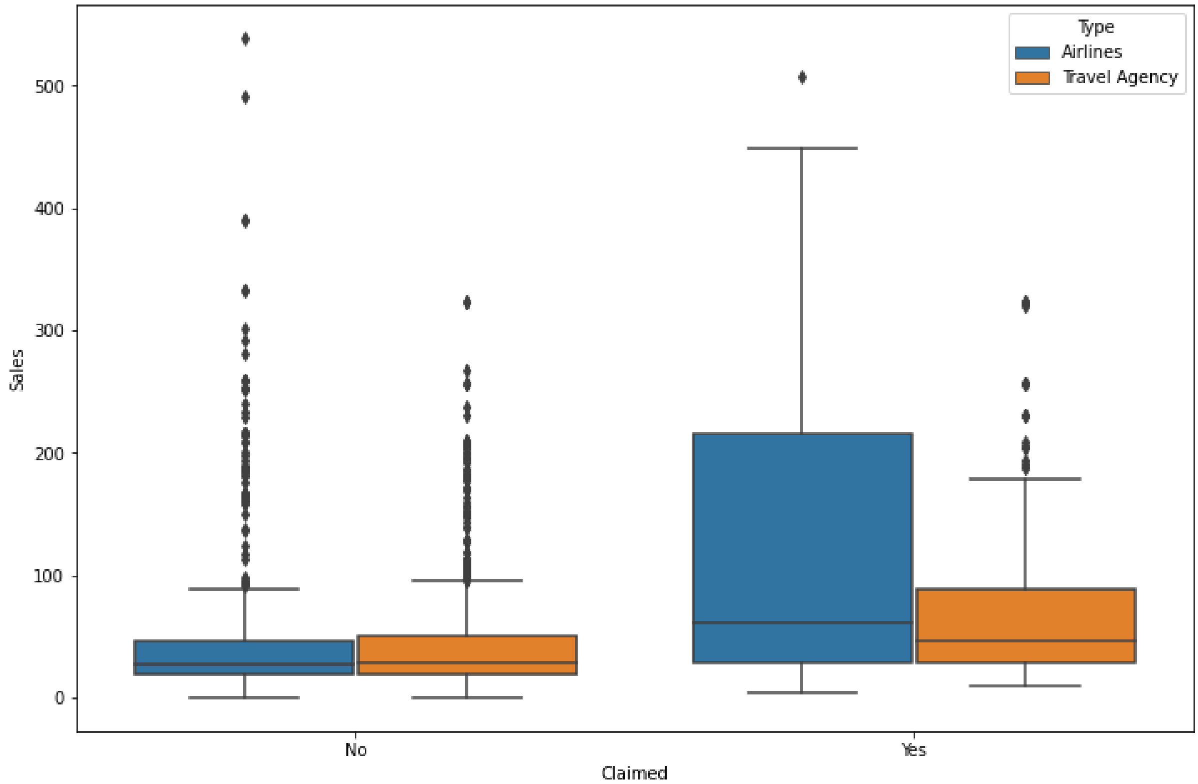


Sales-Claim- other categorical attributes

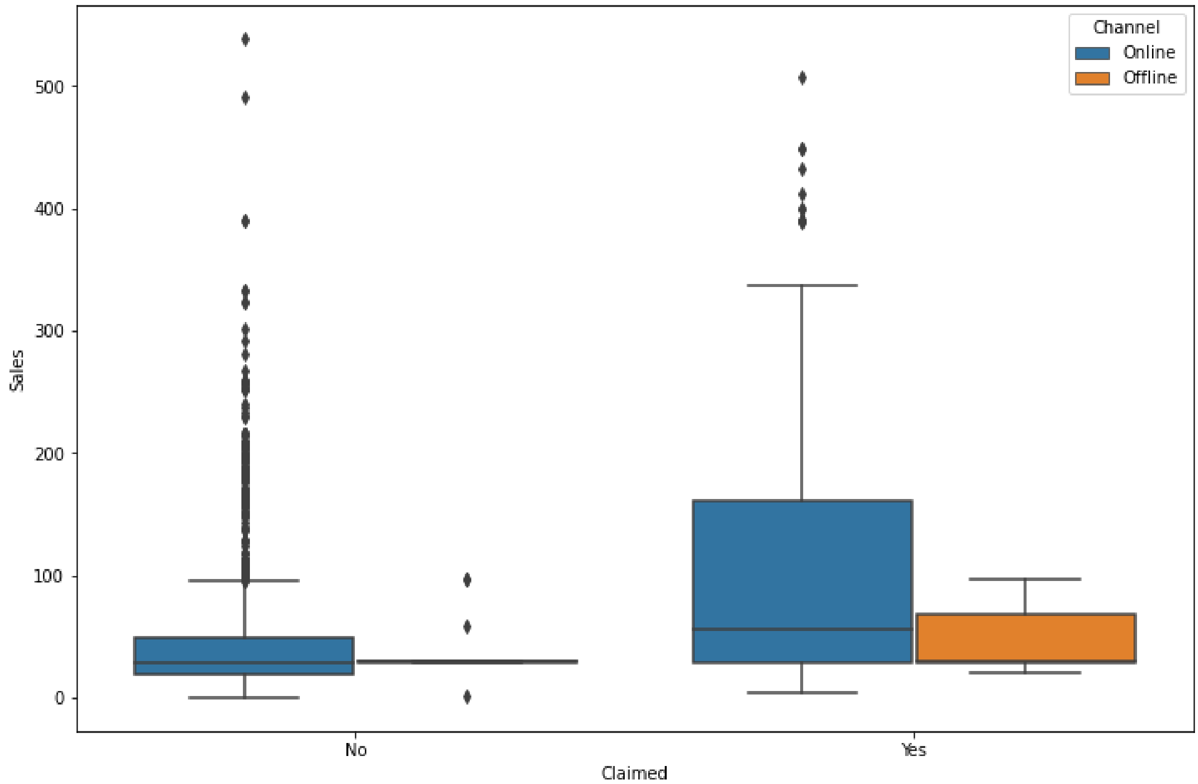
I. For both the claimed status 'Yes' and 'No' the median amount worth of sales per customer in procuring tour insurance policies is highest for both C2B and CWT tour firm and lowest for both EPX and JZI



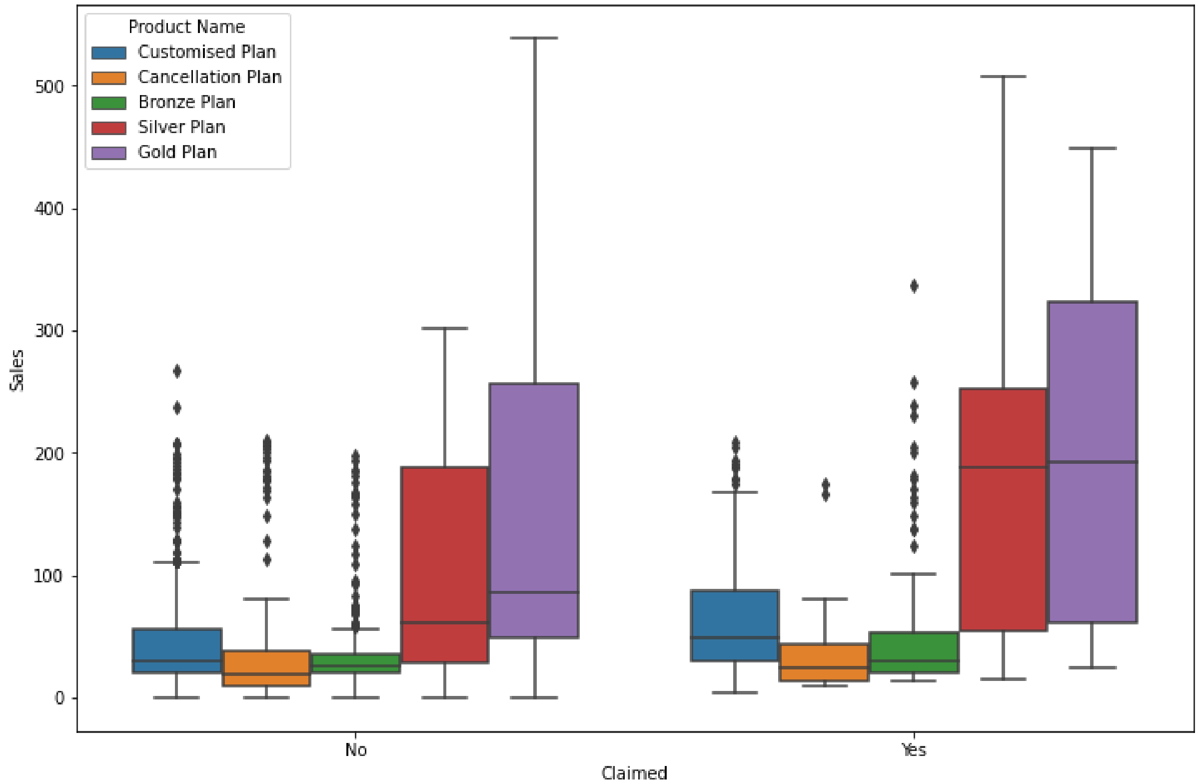
II. For claimed status 'Yes', the median amount worth of sales per customer in procuring tour insurance policies is greater for Airlines tour type. For claimed status 'No' the median amount worth of sales per customer in procuring tour insurance policies is same for both the tour type



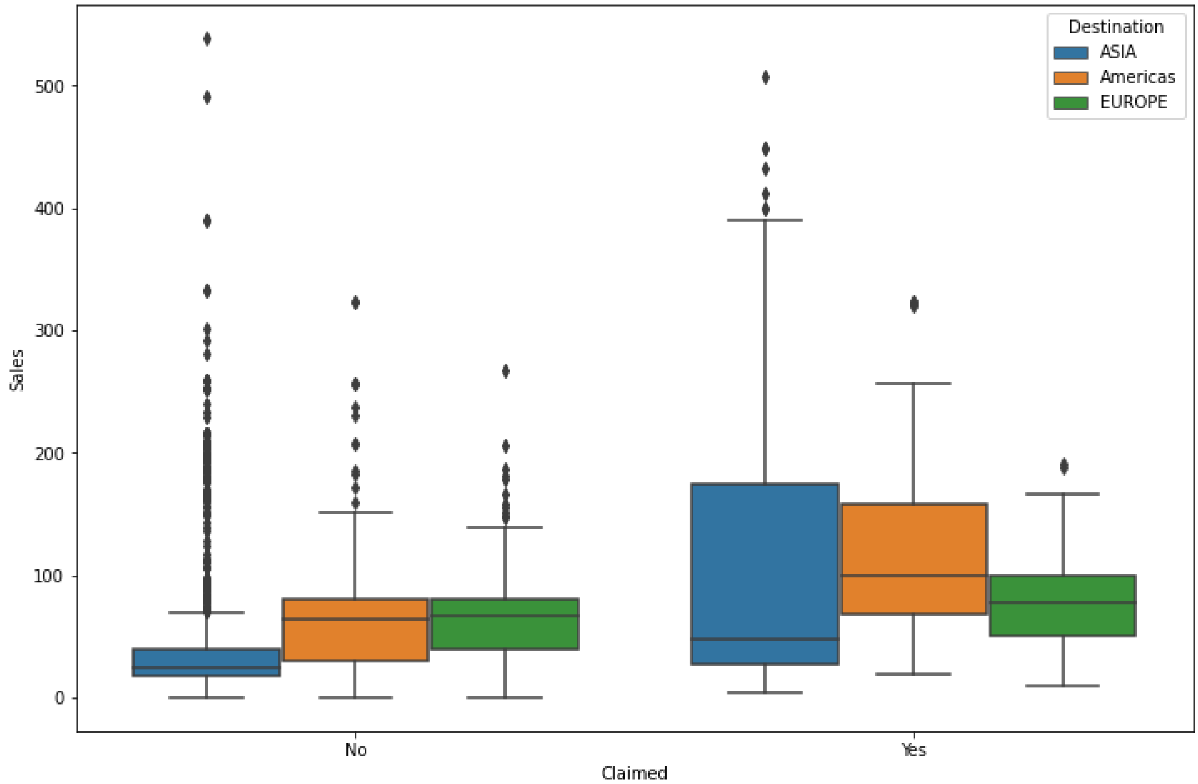
III. For claimed status ‘Yes’, the median amount worth of sales per customer in procuring tour insurance policies greater for Online distribution channel compared to offline



IV. For both the claimed status ‘Yes’ and ‘No’, the median amount worth of sales per customer in procuring tour insurance policies is highest for Gold-Plan tour insurance product and lowest is Cancellation-plan

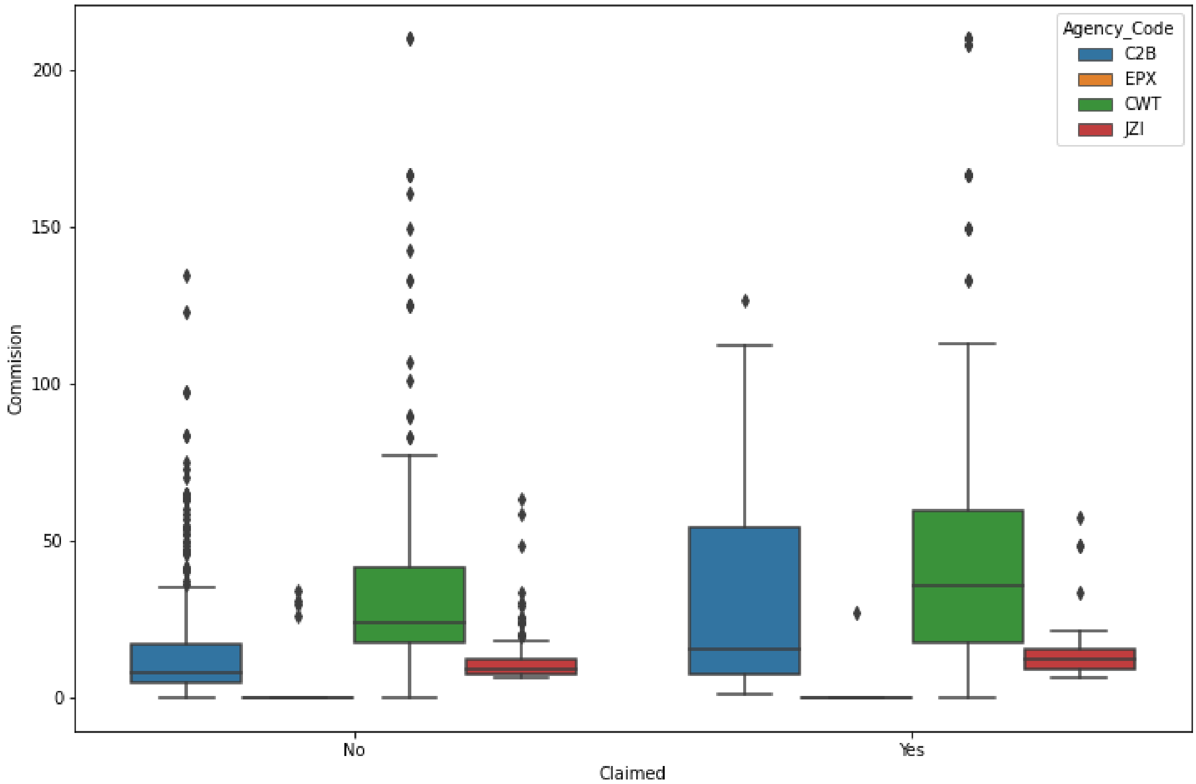


V. For claimed status 'Yes', the median amount worth of sales per customer in procuring tour insurance policies highest for destination America and lowest for Asia. For claimed status 'No', the median amount worth of sales per customer in procuring tour insurance policies highest for Europe and lowest for Asia

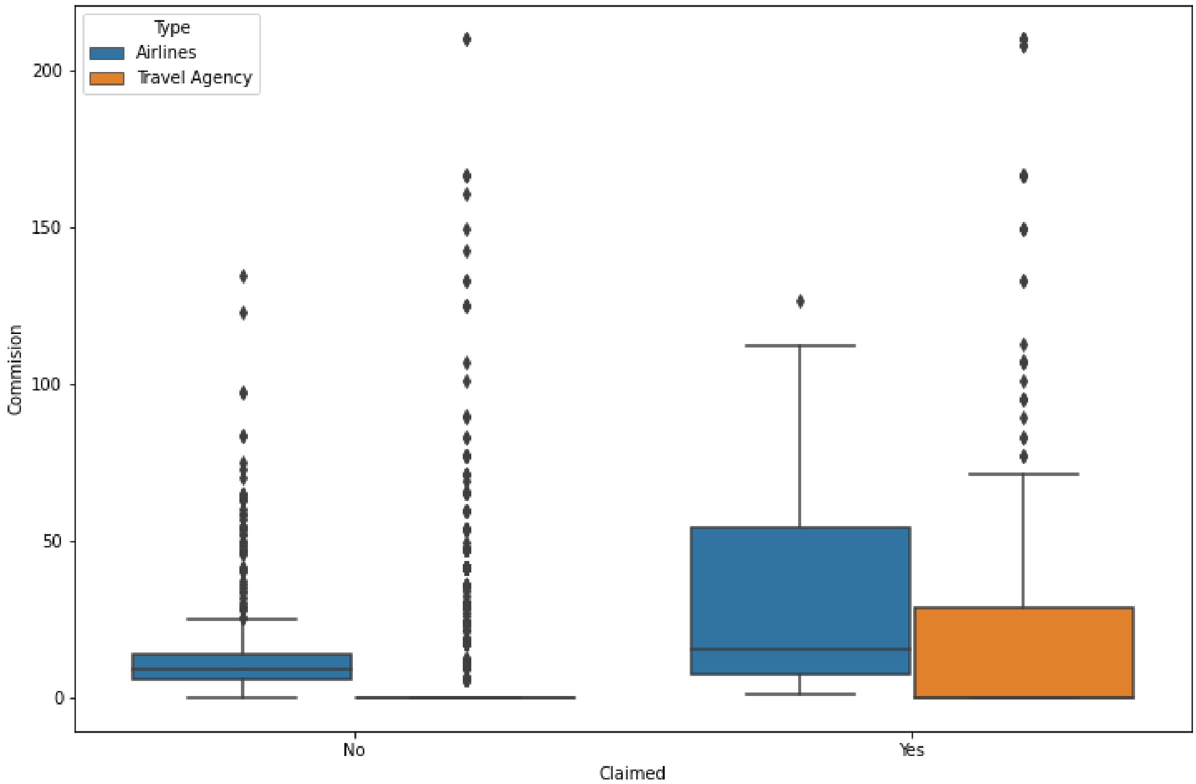


Commission-Claim- other categorical attributes

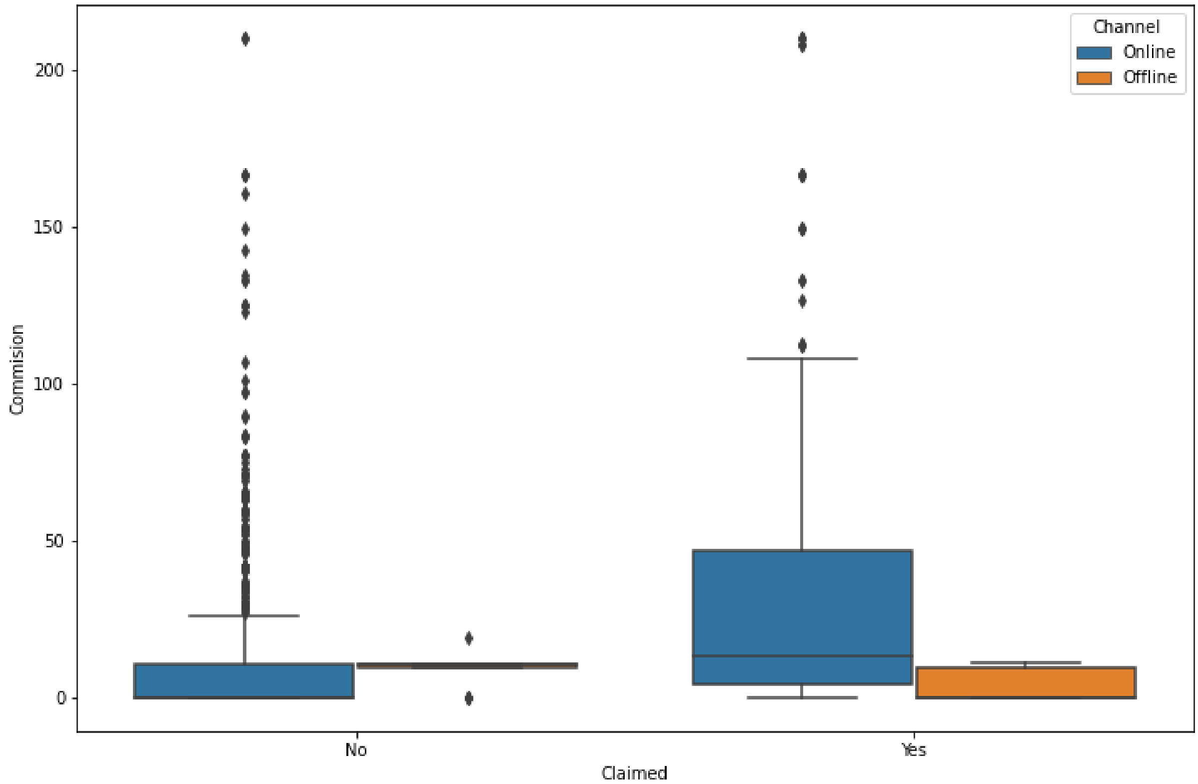
I. For both the claimed status 'Yes' and 'No' the median of the commission received for tour insurance firm highest in CWT tour firm and lowest in EPX tour firm



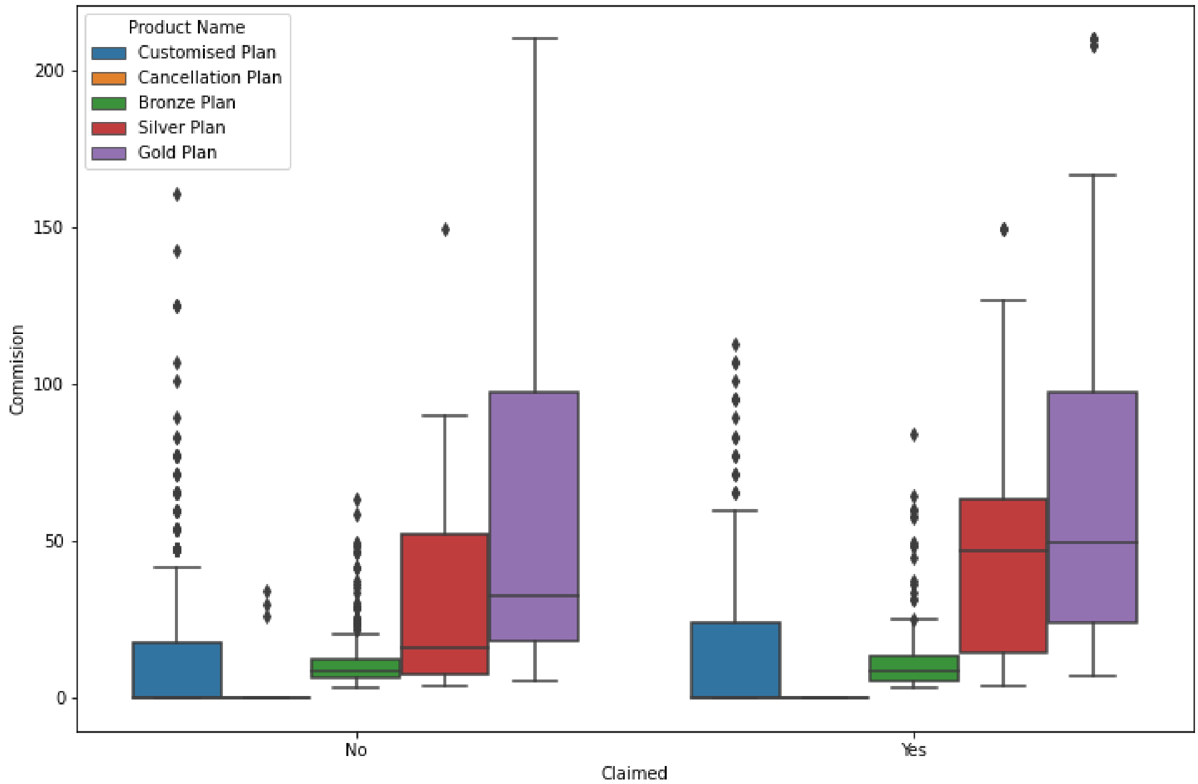
II. For both the claimed status 'Yes' and 'No' the median of the commission received for tour insurance firm is greater in Airlines tour type and least in Travel agency tour type



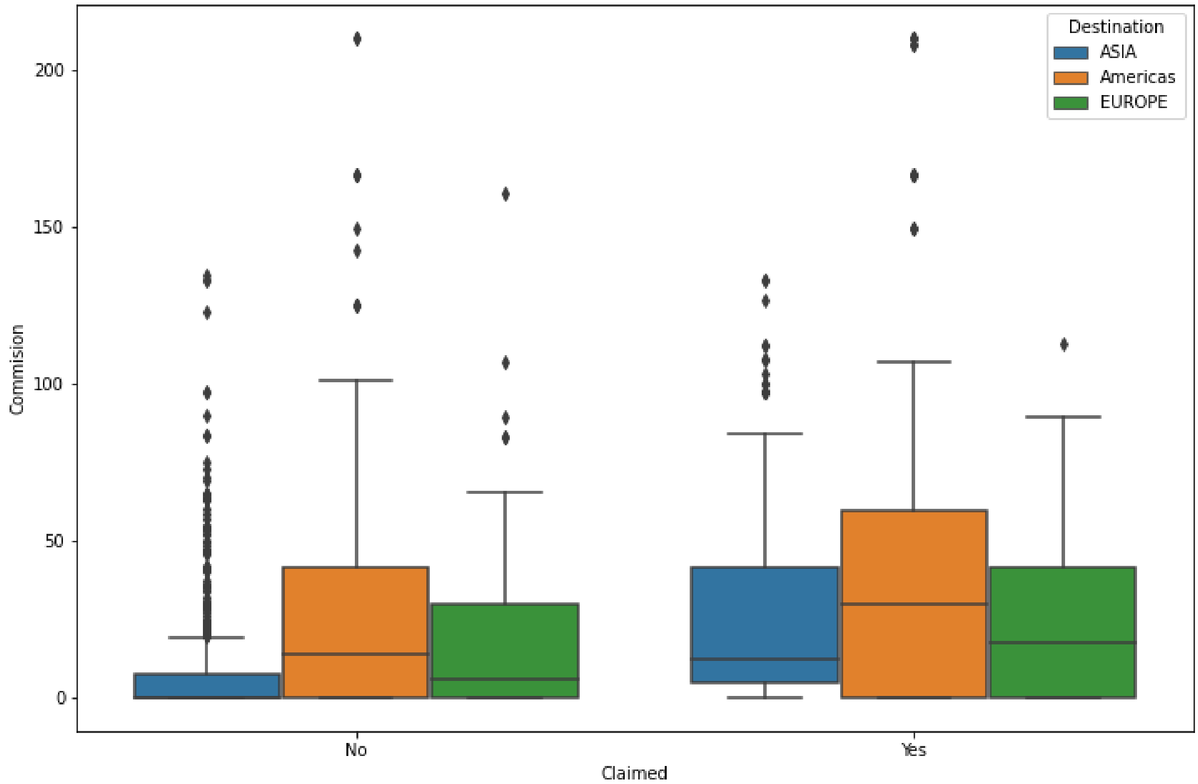
III. For claimed status 'Yes', the median of the commission received for tour insurance firm is greater for Online distribution channel compared to offline. For claimed status 'No', the median of the commission received for tour insurance firm is greater for Offline distribution channel compared to online



IV. For both the claimed status 'Yes' and 'No', the median of the commission received for tour insurance firm is greater for Gold-Plan tour insurance product and lowest is Cancellation-plan



V. For both the claimed status 'Yes' and 'No', the median of the commission received for tour insurance firm is highest for destination America and lowest for Asia



Problem 2.2

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

CART Model

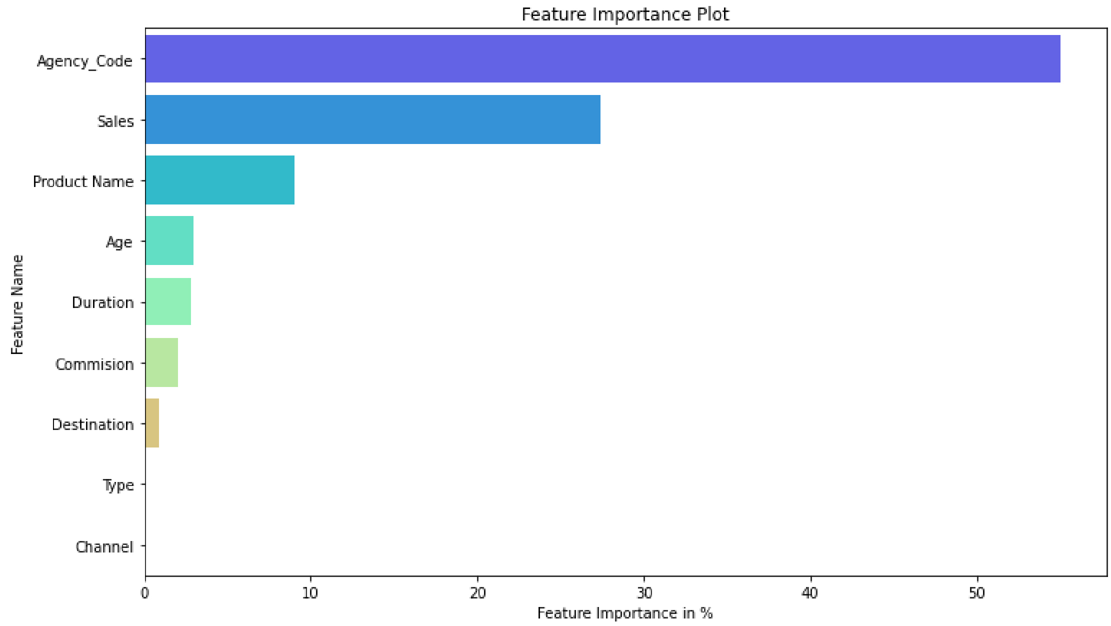
- For Decision tree building all the data should be in the form of numerical data type. In the given dataset there are 6 attributes are of object data type presents in the data; therefore, they are converted into categorical type with codes.
- From Sklearn packages train_test_split was imported and splitting of data is being done in 70:30 ratio (70% for train and 30% for test). Random state has been set to 0
- The shape of the data is as follows

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

- Decision tree classifies imported from Sklearn. tree package.
- For Decision Tree Classifier the criterion used is 'gini'-
  - $Gini = 1 - \Sigma {p_i}^2$
- Before pruning (hyper parameter/ grid search) the tree was allowed to grow fully and it had approximately depth of 20
- Pruning of decision tree helps to prevent overfitting the training data so that our model generalizes well to unseen data. Pruning a decision tree means to remove a subtree that is redundant which is not an useful split and replace it with a leaf node.

- Cross validation used as 3
- For better model we have performed grid search with different set of values for- max_depth': [7, 8, 9, 10], 'min_samples_leaf': [15, 20, 25], 'min_samples_split': [45, 60, 75]
- Best parameters have been found [max_depth': 7, 'min_samples_leaf': 15, 'min_samples_split': 75], which implies the decision tree has depth of 7, minimum sample leaf- 15 ensures that every leaf node/ terminal node has at least 15 observations in them, minimum sample split- 75 ensures that every node before splitting should have at least 75 observations
- Classification report for the particular model also generated. Accuracy obtained for train sample is 0.79 and for test sample is 0.78.
- Feature importance for the model also generated, which shows that Agency Code feature carries greater importance compared to all other attributes. Channel and Type attributes have no importance for the model

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



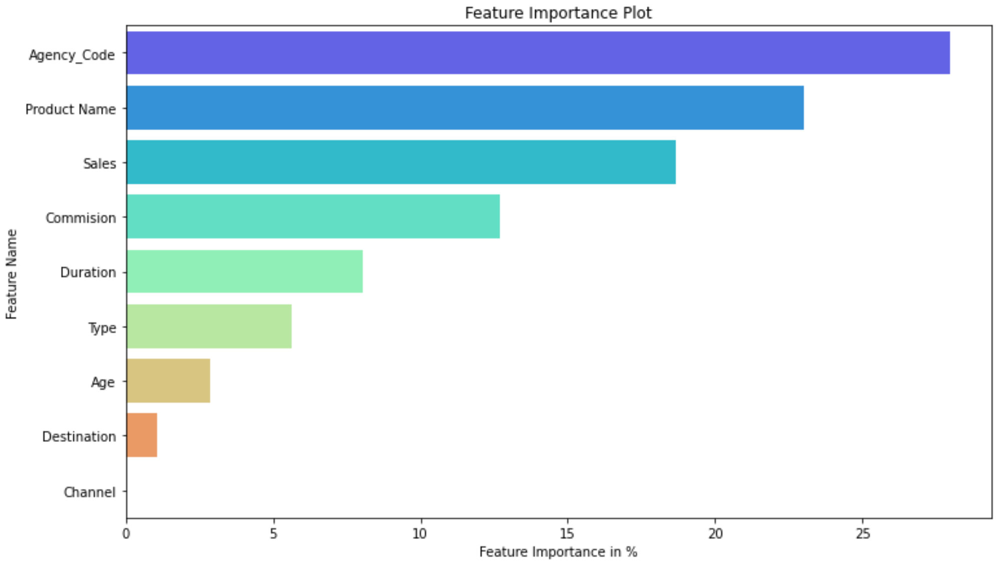
Random Forest Model

- For Decision tree building all the data should be in form of numerical data type. There are 6 attributes are of object data type presents in the data; therefore, they are converted into categorical type with codes.
- From Sklearn packages train_test_split was imported and splitting of data is being done in 70:30 ration (70% for train and 30% for test). Random state has been set to 0
- The shape of the data is as follows

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

- For Random Forest model the criterion used is 'gini'-
- Cross Validation used is 3
- For better model we have performed grid search with different set of values for- max_depth': [7, 8,] 'min_samples_leaf': [20, 25], 'min_samples_split': [60, 75], 'n_estimators: [101,301]
- Best parameters has been found [`max_depth': 8, 'min_samples_leaf': 25, 'min_samples_split': 75, 'n_estimator: 301'`], which implies the decision trees must have depth of 8, minimum sample leaf- 25 ensures that every leaf node/ terminal node has at least 25 observations in them, minimum sample split- 75 ensures that every node before splitting should have at least 75 observations, n estimator- 301 number of decision trees build within the random forest.
- Classification report for the random forest model also generated. Accuracy obtained for train sample is 0.79 and for test sample is also 0.79
- Feature importance for the model also generated, which shows that Agency Code feature carries greater importance compared to all other attributes and Channel has no importance for model.

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



Artificial Neural Network Model

- From Sklearn packages train_test_split was imported and splitting of data is being done in 70:30 ration (70% for train and 30% for test).
- Random state has been set to 0
- Artificial neural network models learn a mapping from input variables to an output variable. as such, the scale and distribution of the data drawn from the domain may be different for each variable (for example age and duration). Input variables may have different units (e.g. Years, Currency in the data set) that, in turn, may mean the variables have different scales. Differences in the scales across input variables may increase the difficulty of the problem being modeled. Hence, the given set scaled using Standard Scaler
- For scaling fit transform used on train data set and only transform used on test split
- Cross validation used is 3 and random state set to 0
- Best parameter has been found from the grid serch : `activation': 'relu','hidden_layer_sizes': (100, 100),'max_iter': 500,'solver': 'sgd, 'tol':0.01.` Number of hidden layer= 2, iteration is 500, activation function is ReLu- it is the function through which we pass our weighed sum, in order to have a significant output, namely as a vector of probability or a 0–1 output. Solver also known as optimization algorithm is sgd with tolerance 0.01- Stochastic Gradient Descent (it minimizes the loss according to the gradient descent optimization, and for each iteration it randomly selects a training sample)
- Classification report for the Ann model also generated. Accuracy obtained for train sample is 0.77 and for test sample is 0.76

Problem 2.3

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

CART Model-

- Below figures shows the confusion matrix for training and test set

Confusion Matrix

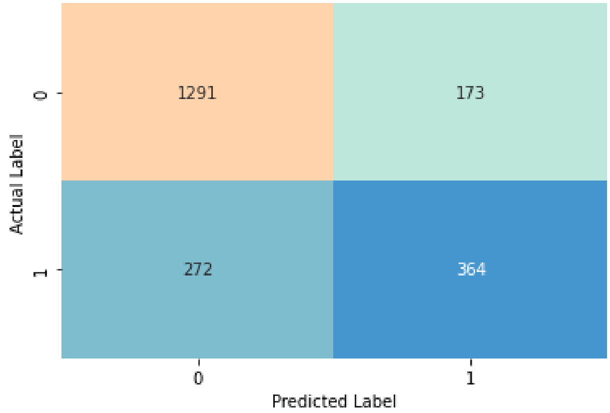


Figure 32: Confusion Matrix for train set

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

Confusion Matrix

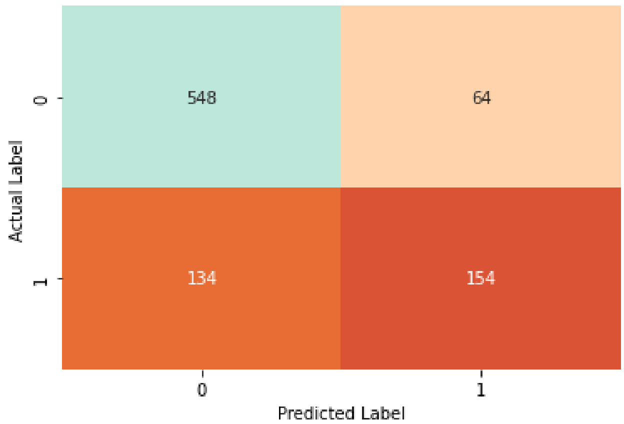


Figure 33: Confusion Matrix for test Set

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

- Below Reports shows the classification reports of both training and test 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

Figure 34: Classification report for train 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

Figure 35: Classification report for test set

- Below graphs shows the ROC curve for both training and test set

AUC: 0.837

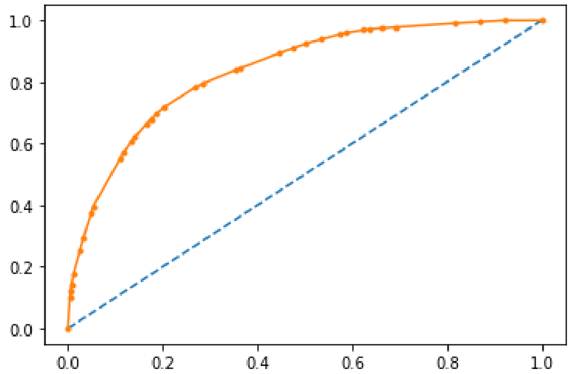


Figure 36: ROC curve for Train set

AUC: 0.817

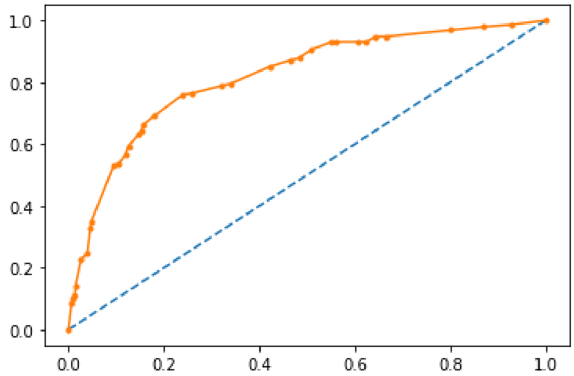


Figure 37: ROC curve for test set

<u>Inferences</u>

- Accuracy for the train set was found to be 0.79 and for test set 0.78
- Precision for claim status 'Yes' in the train set was found to be 0.68 and for test 0.71, In the test set, it implies that 0.29 were wrongly claimed as 'Yes'. From the confusion matrix of test set we can see that 64 observations are false positives
- Recall for claim status 'Yes' in the train set was found to be 0.57 and for test 0.53. This implies 0.47 were wrongly claimed as 'No'. From the confusion matrix of test set we can see that 134 observations are false negatives.
- The accuracy, Precision and recall values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model
- Precision metrics plays a very important role for this particular business problem. Since there are 64 false positives present, it could lead to a negative implication to the Insurance company.
- Recall metrics also have an implication to the business. Since, there are 134 false negatives present in, it could lead to have negative impression on the Insurance company which may lead to loss of customers.
- Area under the curve o training data is 83.7% and on test data is 81.7% which seems good. AUC graph foe both the test and train dataset are not flat which implies a good performance model
- Overall, it is a moderate model can be used for prediction

Random Forest Model

- Below figures shows the confusion matrix for training and test set

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

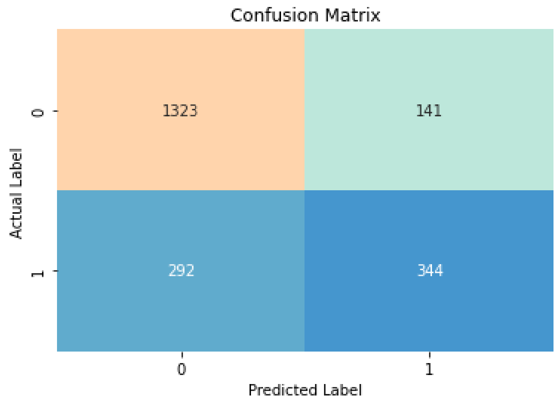


Figure 38:Confusion Matrix for Train set

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

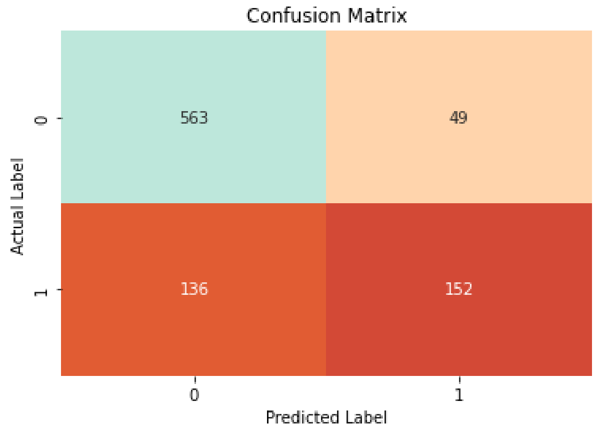


Figure 39: Confusion matrix for test set

- Below Reports shows the classification reports of both training and test 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

Figure 40: Classification report for train 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

Figure 41: Classification report for test set

- Below graphs shows the ROC curve for both training and test set

AUC: 0.836

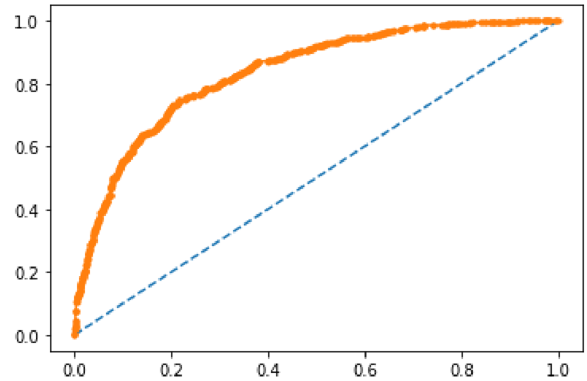


Figure 42:ROC curve for train set

AUC: 0.842

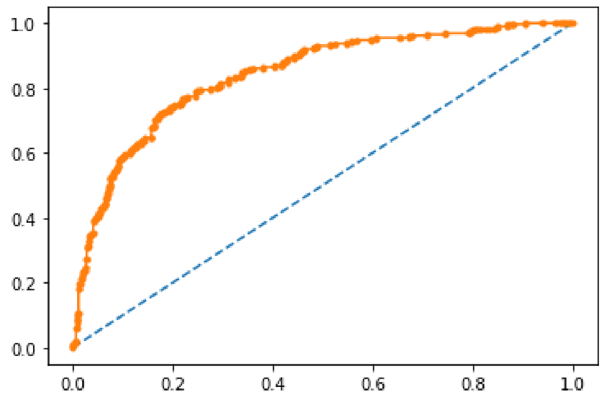


Figure 43: ROC curve for test set

<u>Inference</u>

- Accuracy for the train set was found to be 0.79 and for test set also 0.79
- Precision for claim status ‘Yes’ in the train set was found to be 0.71 and for test 0.76, In the test set, it implies that 0.24 wrongly claimed as ‘Yes’. From the confusion matrix of test set we can see that 49 observations are false positives
- Recall for claim status ‘Yes’ in the train set was found to be 0.54 and for test 0.53. This implies 0.47 wrongly claimed as ‘No’. From the confusion matrix of test set we can see that 136 observations are false negatives.
- The accuracy, Precision and recall values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model

- Precision metrics plays a very important role for this particular business problem. Since there are 49 false positives present, it could lead to a negative implication to the Insurance company.
- Recall metrics also have an implication to the business. Since, there are 136 false negatives present in, it could lead to have a negative impression on the Insurance company which may lead to loss of customers.
- Area under the curve o training data is 83.6% and on test data is 84.7% which seems good. AUC graph for both the test and train dataset are not flat which implies a good performance model
- Overall, it is a moderate model can be used for prediction

Artificial Neural Network Model

- Below figures shows the confusion matrix for training and test set

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

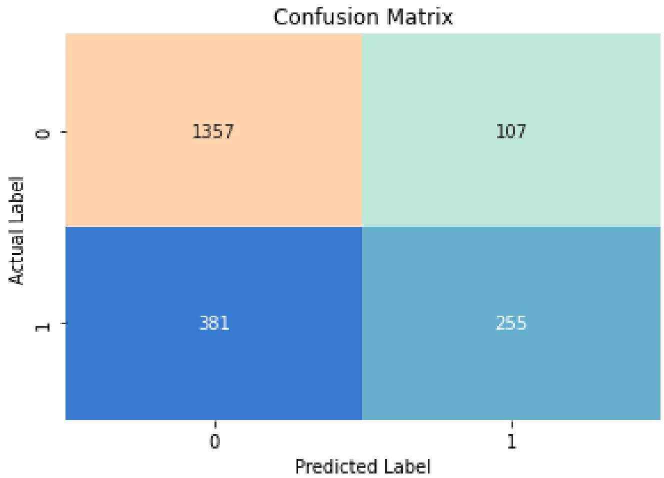


Figure 44: Confusion Matrix for Train set

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

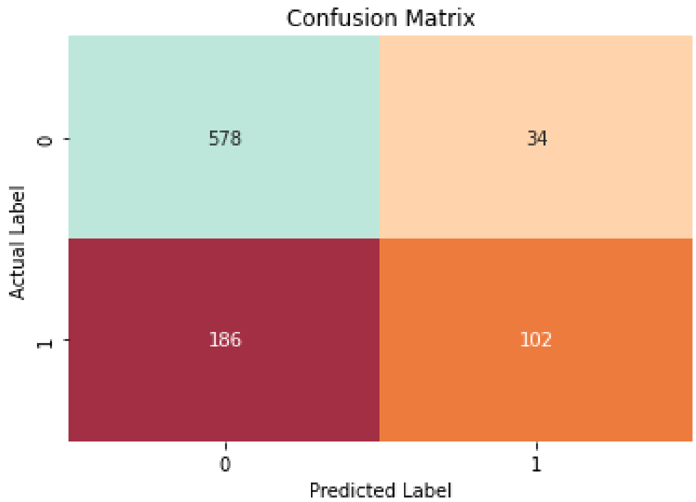


Figure 45: Confusion Matrix for Trest set

* Below Reports shows the classification reports of both training and test 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

Figure 46: Classification report for train 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

Figure 47:Classification report for test set

* Below graphs shows the ROC curve for both training and test set

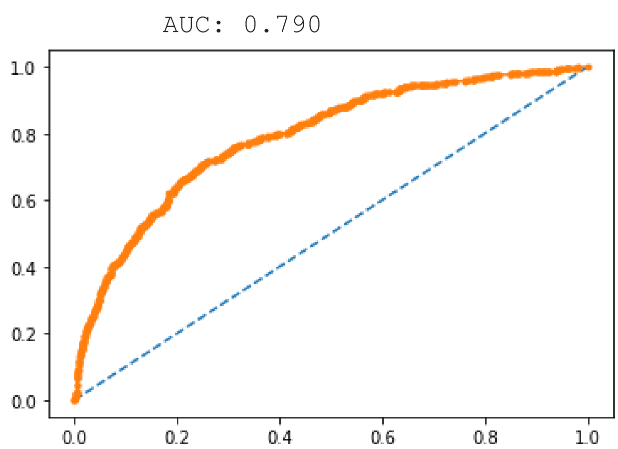


Figure 48:ROC curve for train set

AUC: 0.799

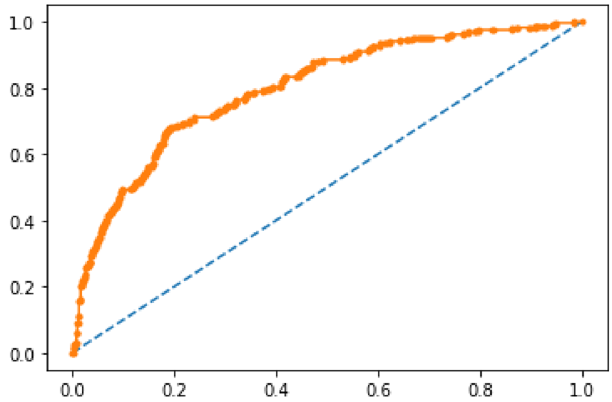


Figure 49:ROC curve for test set

Inference:

- Accuracy for the train set was found to be 0.77 and for test set 0.76
- Precision for claim status ‘Yes’ in the train set was found to be 0.70 and for test 0.75, In the test set, it implies that 0.25 of the total observation were wrongly claimed as ‘Yes’. From the confusion matrix of test set we can see that 34 observations are false positives
- Recall for claim status ‘Yes’ in the train set was found to be 0.40 and for test 0.35. This implies 0.65 were wrongly claimed as ‘No’. From the confusion matrix of test set we can see that 186 observations are false negatives.
- The accuracy and precision values are almost similar for both the training and test data set which implies no overfitting and underfitting happened in the model
- The recall values are less
- Precision metrics plays a very important role for this particular business problem. Since there 34 are false positives present, it could lead to a negative implication to the Insurance company.
- Recall metrics also have an implication to the business. Since, there are 186 false negatives present in, it could lead to have a negative impression on the Insurance company which may lead to loss of customers.
- Area under the curve o training data is 79% and on test data is 79.9% which seems good. AUC graph for both the test and train dataset are not flat which implies a good performance model
- Overall, it is a moderate model can be used for prediction.

Problem 2.4

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

	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	0.84	0.82	0.62	0.61
RF	0.79	0.79	0.71	0.76	0.54	0.53	0.84	0.84	0.61	0.62
ANN	0.77	0.76	0.75	0.75	0.40	0.35	0.79	0.80	0.51	0.48

Table:1 Comparison table of All the Models

Ref: RF- Random Forest, ANN- Artificial Neural Network

Observations

- It is evident from the table that accuracy metrics are similar for all the 3 models
- Precision metrics are similar for all the 3 models
- Recall metrics are almost similar for CART and RF; however, it is evidently low in ANN
- AUC scores are almost similar for all the 3 models
- F1 scores are almost similar in CART and RF; however, it is evidently low in ANN

Selection of Model

- Higher the F1 score better is the model, in case of ANN F1 score is quite less compared to the other 2 models. F1 score in ANN is less because of lower recall value. Hence, among the 3 models we are omitting ANN for final model.
- Values are almost same across all the metrics for both CART and RF. However, Precision and AUC scores are slightly better in RF model
- Hence, we are selecting Random Forest model as suitable model for claim staus prediction

Final Model- Random Forest (conclusion)

- Precision metrics plays a very important role for this particular business problem. Since there are false positives present, it could lead to a negative implication to the Insurance company. Therefore, the bank should stop paying for false positive claims
- Recall metrics also have an implication to the business. Since, there are false negatives present in, it could lead to have a negative impression on the Insurance company which may lead to loss of customers.

Problem 2.5

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

- Since, we have seen from the above analysis that their significant numbers of false positive and false negative present in our suitable, which could lead to less profit and negative reputation for the insurance company
- From the multivariate-bivariate analysis we have found few insights such as:
    - Tour agency firm C2B has the maximum claim status 'Yes', which can raise the concern
    - Tour agency firm EPX has the maximum claim status 'No', which can raise
    - Silver plan tour product receives the maximum claim among all other tour products
    - Mean sales at agency JZI are the least
    - For claimed status 'Yes', the median of the commission received for tour insurance firm is greater for Online distribution channel compared to offline

- Business Recommendation
    - The insurance company should inspect and review the claims coming from C2B agency firm, since maximum claim status ‘Yes’ comes from this firm. This may lead to identify the reason behind false positive and necessary actions should be implemented to reduce the false positive claims.
    - The insurance company should inspect and review the claims coming from Silver Plan, since maximum claim status ‘Yes’ comes from this firm. This may lead to identify the reason behind false positive and necessary actions should be implemented to reduce the false positive claims
    - The insurance company should inspect and review the claims coming from EPX agency firm, since maximum claim status ‘No’ comes from this firm. This may lead to identify the reason behind false negative and necessary actions should be implemented to reduce the false negative claims
    - Suitable strategies such as amount refund or some other strategies should implement by the insurance firm to improve the Cancellation Plan tour product.
    - Suitable strategies should implement to improve the ‘Yes’ claim status for JZI agency firm

greatlearning
Power Ahead