DATA MINING PROJECT

DSBM

Done by: Hariharan Manickam

Problem 1: Clustering

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

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

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

We can observe that we we do not need to change the data type as all are float.

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

We can observe the basic statistics of the dataset in the table above.

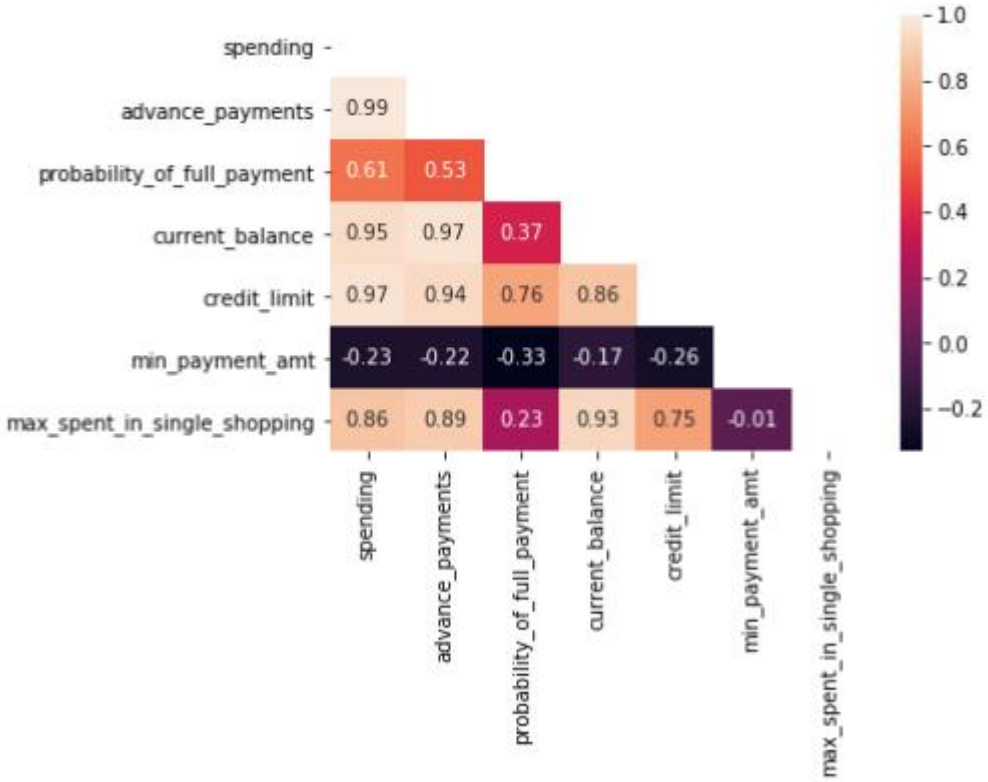
Conclusions from Univariate analysis

- Spending : Does not follow normal distribution. Average amount spent by customer is around 14000/month.
- Advance payment : Does not follow normal distribution. Average amount paid by customer is around 1400/month.
- Probability of full payment : Follows a normal distribution with a slight left skew. 87% of the customers give full payment on time.
- Current balance : Follows a normal distribution with a slight right skew.
- Credit limit : Does not follow normal distribution.
- Minimum payment amount : Follows a normal distribution with a very slight right skew.
- Maximum spent in single shopping : Does not follow a normal distribution.

We can observe from the dataset that the units are different for all the variables hence it will be difficult to compare. We should follow the approach of standardization/Normalization so that further analysis can be easily done.

Conclusions from bivariate analysis

1. Spending is directly proportional to advance payment, probability of full payment, current balance, credit limit and max spent in single shopping.
2. If there is increase in spending then the advance payment that the customer pays also increases.
3. The probability of full payment is around 90% when the spending is 20000 per month.
4. Expenditure is directly proportional to credit limit and current balance
5. It follows a similar pattern for max_spent_in_single_shopping, The only aberration is the No pattern found in minimum payment amount.
6. Advance payment is directly proportional to current balance, credit limit, max_spent_in_single_shopping.
7. Credit limit is directly proportional to the spending. advance payment, probability of full payment, min payment amount and max spent in shopping increases.
8. Current balance directly proportional to spending and advance payment made by the customer.



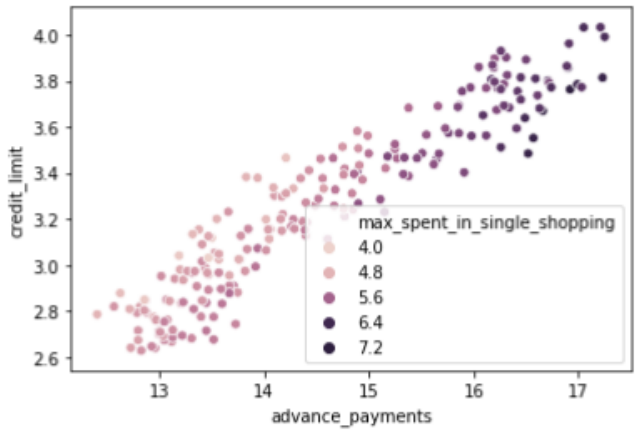
Highly positively correleated variables

1. Advance_payments and Spending: 0.99
2. Credit_Limit and Spending:0.97
3. Current balance and advance payments: 0.97
4. Current balance and Spending: 0.95
5. Credit_Limit and Advance payment: 0.94
6. Max_spent_in_single_shopping and Current_balance: 0.93
7. Max_spent_in_single_shopping and advance_payment: 0.89
8. Max_spent_in_single_shopping and Spending: 0.86
9. Credit_Limit and Current balance: 0.86

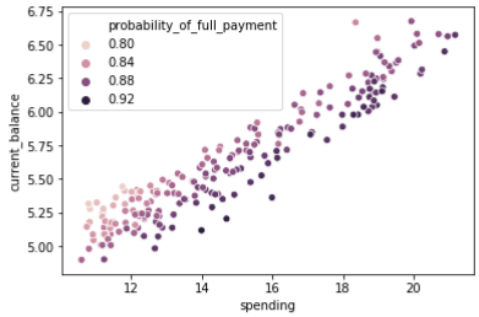
Slighltly negatively correleated variables

1. Min_payment_amt and probability of full payment: 0.33
2. Min_payment_amt and credit_limit: 0.26
3. Min_payment_amt and spending: 0.23
4. Min_payment_amt and advance_payments: 0.22
5. Min_payment_amt and current_balance: 0.17

Conclusions from multivariate analysis



We observe increase in current balance and spending of the customer increases the probability of full payment.



We observe increase in credit limit increases the advance payment made by the customer through cash also increases the spending in a single shopping.

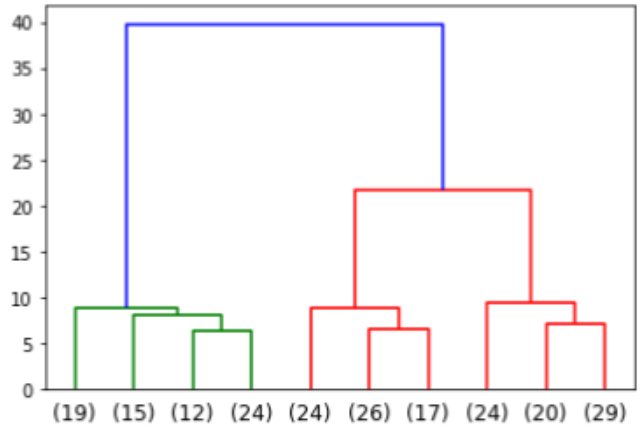
1.2 Do you think scaling is necessary for clustering in this case? Justify

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

From the table above, it is clear that we should scale the data as they are not of a uniform range. For example, The standard deviation for spending is is 2.9 whereas it is 0.02 for probability of full payment.

We have to scale the data because using the data without scaling might create some sort of bias in the outcome of the analysis. It is always important to standardize the data we use to prevent bias in the outcome.

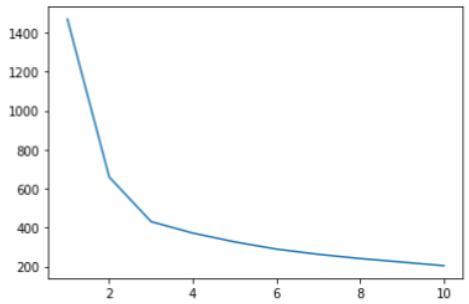
1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them



We can observe that it is best to split the customer profiles into 2 categories from the dendogram above.

We use the wardlinkage method and fclusters.

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.



From the plot, we can observe that the optimum number of clusters is 2 as there is a major dip from 1 cluster to 2 clusters. From 2 to 3 clusters, there is a slight dip. Post 3 cliusters, the plot seems to get flatter and flatter.

We run the kmeans with 2 clusters and add the labels with our dataframe

We find that all silhouette widths are positive. Hence the mapping is done correctly.

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

Observations

1. Cluster 0 (Conservative credit card users)

- 63% of these customers have an average spending of 13000 per month
- The probability of full payment is around 86%.
- These customers have a lower credit limit, current balance and pay lesser advance payments.
- This can be observed from the max_spent in single shopping with an average spending of 5000.
- The minimum payment amount for these customers is about 3800 and slightly better than the liberal group.

2. Cluster 1 (Liberal credit card users)

- 37% of these customers have an average spending of 18000 per month
- The probability of full payment is around 89%.
- These customers have a high credit limit, current balance and pay more advance payments.
- This can be observed from the max_spent in single shopping with an average spending of 6000.
- The minimum payment amount for these customers is about 3500.

1. Cluster 0 (Conservative credit card users)

- Probability of full payment is high
- Maintain higher current balance
- It is important to not lose these customers

Promotional Strategy

- Main goal is to retain customers
- Can do so by providing them special benefits, privilages, special memberships, etc

2. Cluster 1 (Liberal credit card users)

- Have lower spending limit
- Have a higher minimum payment amount
- Majority of customer base
- Need to get them to maintain a higher current balance

Promotional Strategy

- Need to get them to use more credit to make purchases
- Can do so by giving them more incentives to use their credit cards
- Such as cashbacks, discounts, points to avail, and spreading awareness about the perks of using credit

Problem 2: CART-RF-ANN

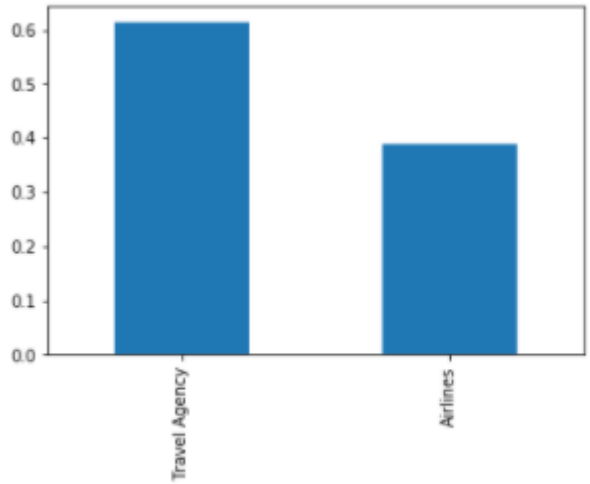
An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models'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

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

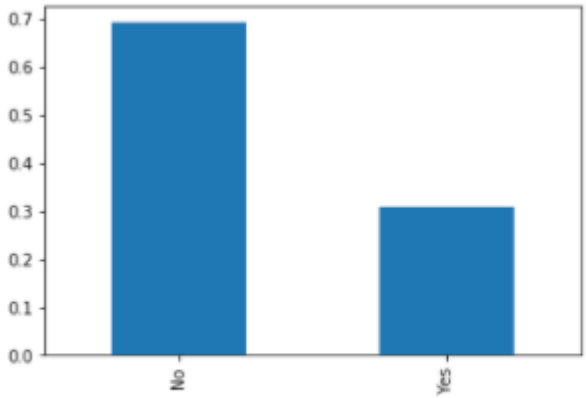
Conclusions from univariate analysis

1. Age: Does not follow a normal distribution.
2. Commision: Follows a heavily right skewed normal distribution.
3. Duration: Does not follow a normal distribution
4. Sales: Seems like a right skewed normal distribution but it does not follow a normal distribution.

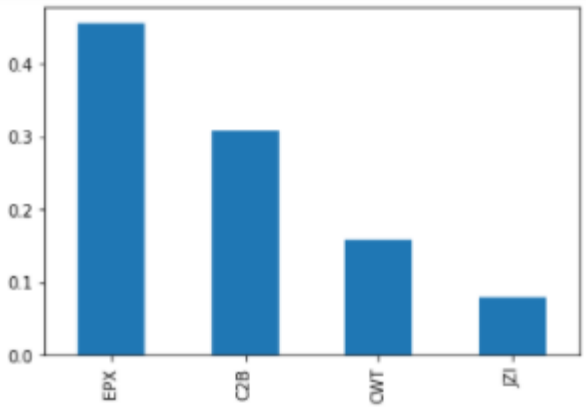
We do observe outliers but we do not need to treat them as we perform classification using a tree based model.



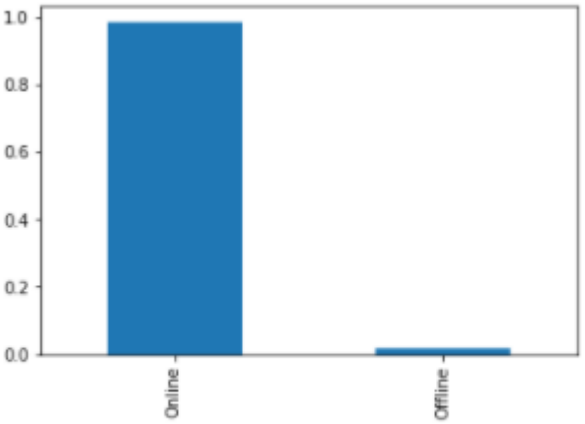
We observe 60% insurances taken are for travel agencies and the remaining 40% for airlines



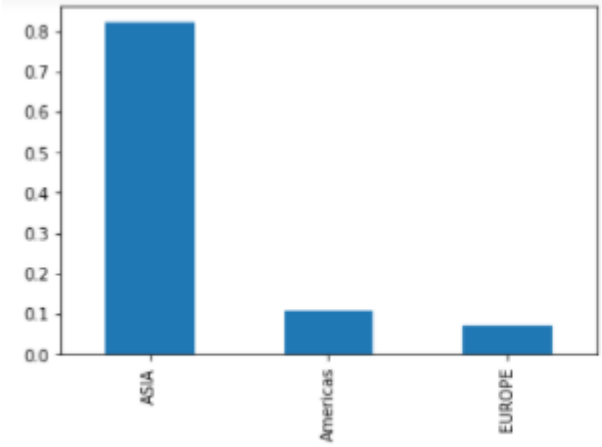
We observe that almost 70% do not claim the insurance and only 30% claim the insurance.



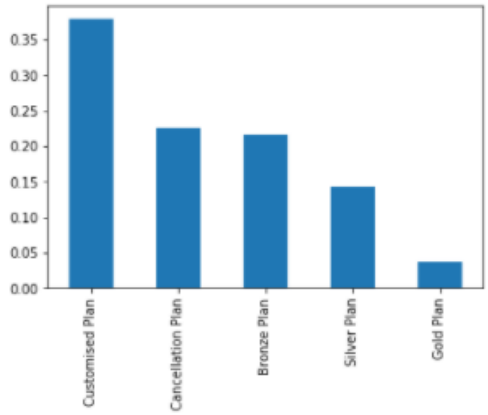
We can observe the agencies ranked from highest to lowest insurance policies given by the tour insurance provider.



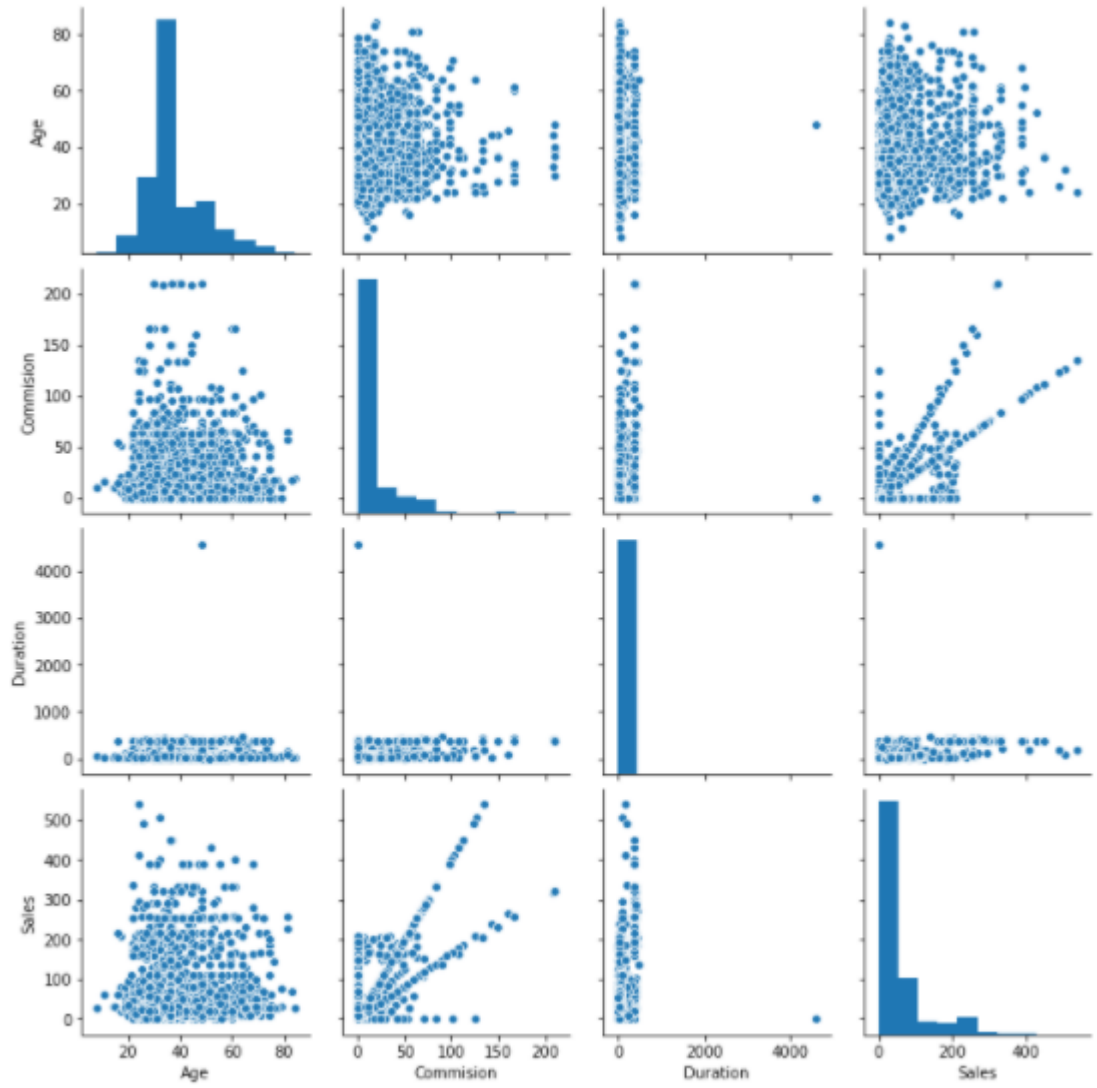
We observe that almost all insurance policies are applied online.



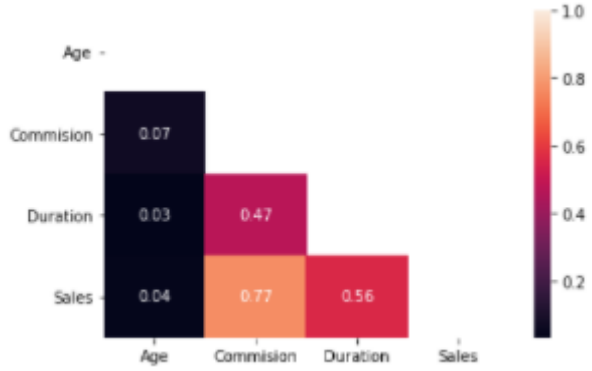
We observe that almost 80% insurance policies are for Asia and approximately 10% for AMericas and Europe each.



We can observe that most agencies prefer a customised plan. We can also observe that many do not prefer the gold plan offered by the company.



From the pairplot we can observe that commision and the volume of sales are directly proportional.



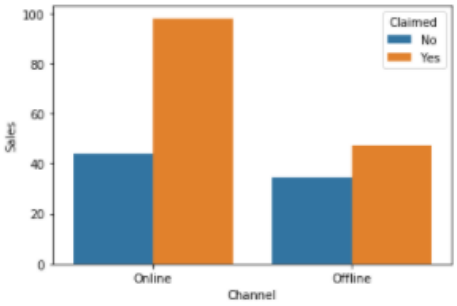
Positively correlated variables

1. Sales and Commision: 0.77
2. Sales and Duration: 0.56
3. Duration and Commision: 0.47

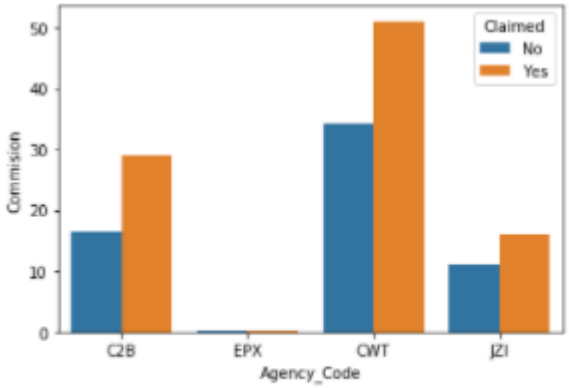
Non correlated variables

1. Age and Duration: 0.03
2. Age and Sales: 0.04
3. Age and Commision: 0.07

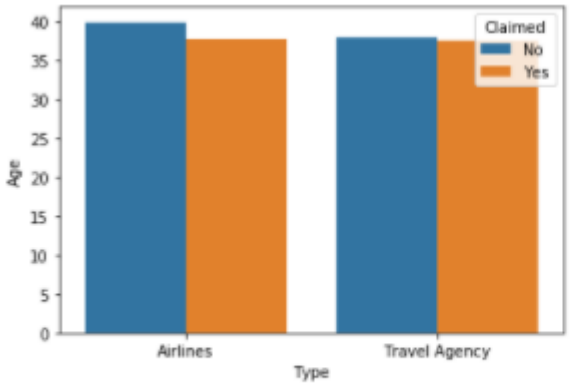
The dataset does not contain negatively correlated variables.



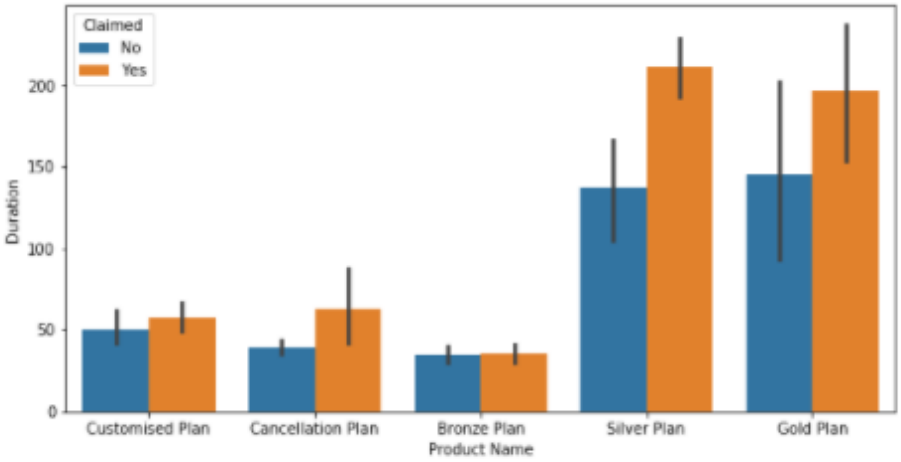
Sales are highest in the online channel compared to offline. The number of insurance policy holders who applied online who have claimed is almost 60% more than the number of insurance policy holders who applied offline.



The agency ranked by claims are as follows. CWT > C2B > JZI > EPX



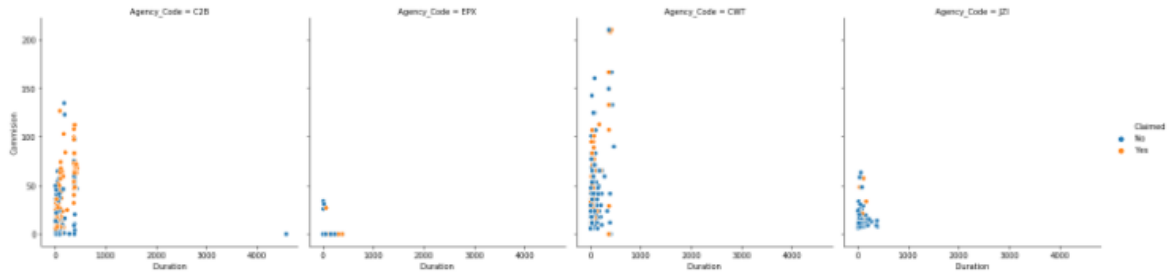
We can observe the average ages of the policy holders who have claimed and not claimed in the plot above.



The highest insurance plans claimed and not claimed are silver and gold and the duration of the policy is more than 100 days.

We observe that:

- 80% of the sales are from Asia
- Most of the sales of the tour insurance are between the age of 20 & 50 for the all the destinations



- The commision amount does notimpactthe duration of the tour insurance or the outcome of the claim status.
- We can observe that C2B and CWT agency code have the most number of claims

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.

	precision	recall	f1-score	support
0	0.79	0.88	0.83	605
1	0.68	0.52	0.59	295
accuracy			0.76	900
macro avg	0.73	0.70	0.71	900
weighted avg	0.75	0.76	0.75	900

- Higher precision score for claim status "No" compared to the claim status "Yes".
- High recall value of 89% which means that if the claim status is 'No' which is 70% of the claims are originally detected correctly.
- Good F1 score with 83% for the testing data.
- Recall value is 52% if the claim status is "Yes" which is 30% of claims received from the customer.
- Poor f1 score at 0.59
- Accuracy is 0.76
- Important to make sure that the customer who is not eligible for a claim is correctly identified because wrong identification of a claim since expenses will exceed the premiums which will result in a loss.

	precision	recall	f1-score	support
0	0.78	0.92	0.85	605
1	0.74	0.47	0.58	295
accuracy			0.77	900
macro avg	0.76	0.70	0.71	900
weighted avg	0.77	0.77	0.76	900

Observations (RF Model)

- Higher precision score for claim status "No" compared to the claim status "Yes".
- High recall value of 92% which means that if the claim status is 'No' which is 70% of the claims are originally detected correctly.
- Good F1 score with 85% for the testing data
- Recall value is 47% if the claim status is "Yes" which is 30% of claims received from the customer.
- Recall value lower than 0.5. Hence it is not good.
- F1 score is also poor at 0.58
- Accuracy is 0.77

This model performs 1% better than CART model.

	precision	recall	f1-score	support
0	0.79	0.89	0.84	605
1	0.70	0.53	0.60	295
accuracy			0.77	900
macro avg	0.74	0.71	0.72	900
weighted avg	0.76	0.77	0.76	900

- Higher precision score for claim status "No" compared to the claim status "Yes".
- High recall value of 89% which means that if the claim status is 'No' which is 70% of the claims are originally detected correctly.
- Good F1 score with 84% for the training data
- Recall value is 53% if the claim status is "Yes" which is 30% of claims received from the customer.
- Recall value is above 0.5 which is acceptable.
- F1 score is very average at 0.6
- Accuracy is 0.77

This model performs exactly as the CART model.

Observations

- Model performs very poorly and we do not use ANN model

	Accuracy	Precision	Recall	AUC ROC Score	F1 score
CART	0.76	0.75	0.76	0.787	0.75
RF	0.77	0.77	0.77	0.819	0.76
ANN	0.77	0.76	0.77	0.501	0.76

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

Observations (CART Model)

- Model gives most importance to Duration, Agency Code, Age, and Sales
- Duration and Sales are important parameters for the management because more sales mean more policies sold and longer duration of the policies mean more premiums
- Risk is higher as the the age of the customers increases.

Observations (RF Model)

- Higher precision score for claim status "No" compared to the claim status "Yes".
- High recall value of 92% which means that if the claim status is 'No' which is 70% of the claims are originally detected correctly.
- Good F1 score with 85% for the testing data
- Recall value is 47% if the claim status is "Yes" which is 30% of claims received from the customer.
- Recall value lower than 0.5. Hence it is not good.
- F1 score is also poor at 0.58
- Accuracy is 0.77

This model performs 1% better than CART model.

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

Conclusion & Recomendations

- We use the CART model as it takes the most relevant parameters for its model.
- Models can be used to identify who will not claim more accurately than who will claim.
- Business should use Duration, Age and Agency code to predict future claims