Will They Claim It ??

Team – Runtime Terror

Problem Statement:

- As a Travel Insurance organisation, it is very important to analyse each and every customer in order to minimize the risk of unnecessary claim approval and also to see that genuine claims are not rejected so as to avoid law-suits which might the organisation need to face in future.

- Also, forecasting the upcoming claims helps to charge competitive premiums that are not too high and not too low. It also contributes to the improvement of the pricing models. This helps the company to be one step ahead of its competitors.

- Stake Holder – Chief Planning Officer

Table of Contents:

- EDA & Visualization
- Feature Engineering
- Feature Selection
- Model Building
- Conclusion

EDA & Visualization

What is the strength of the dataset?

```
In [12]: train.shape
Out[12]: (52310, 11)

In [14]: train.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52310 entries, 0 to 52309
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   ID                     52310 non-null  int64
 1   Agency                 52310 non-null  object
 2   Agency Type            52310 non-null  object
 3   Distribution Channel   52310 non-null  object
 4   Product Name           52310 non-null  object
 5   Duration               52310 non-null  int64
 6   Destination            52310 non-null  object
 7   Net Sales              52310 non-null  float64
 8   Commision (in value)   52310 non-null  float64
 9   Age                    52310 non-null  int64
 10  Claim                  52310 non-null  int64
dtypes: float64(2), int64(4), object(5)
memory usage: 4.4+ MB
```

Inference :

From a given dataset we could see there are total 52310 cases and 11 Features.

- Are there any data discrepancies? If they are what should be our approach to handle them?

```
df.isnull().sum()
```

Readings: We observed that there were no Null values.

- What kind of data Science problem is this?

Readings: We can see target variable "Claim" given in dataset hence this comes under Supervised Machine learning !!!!!!

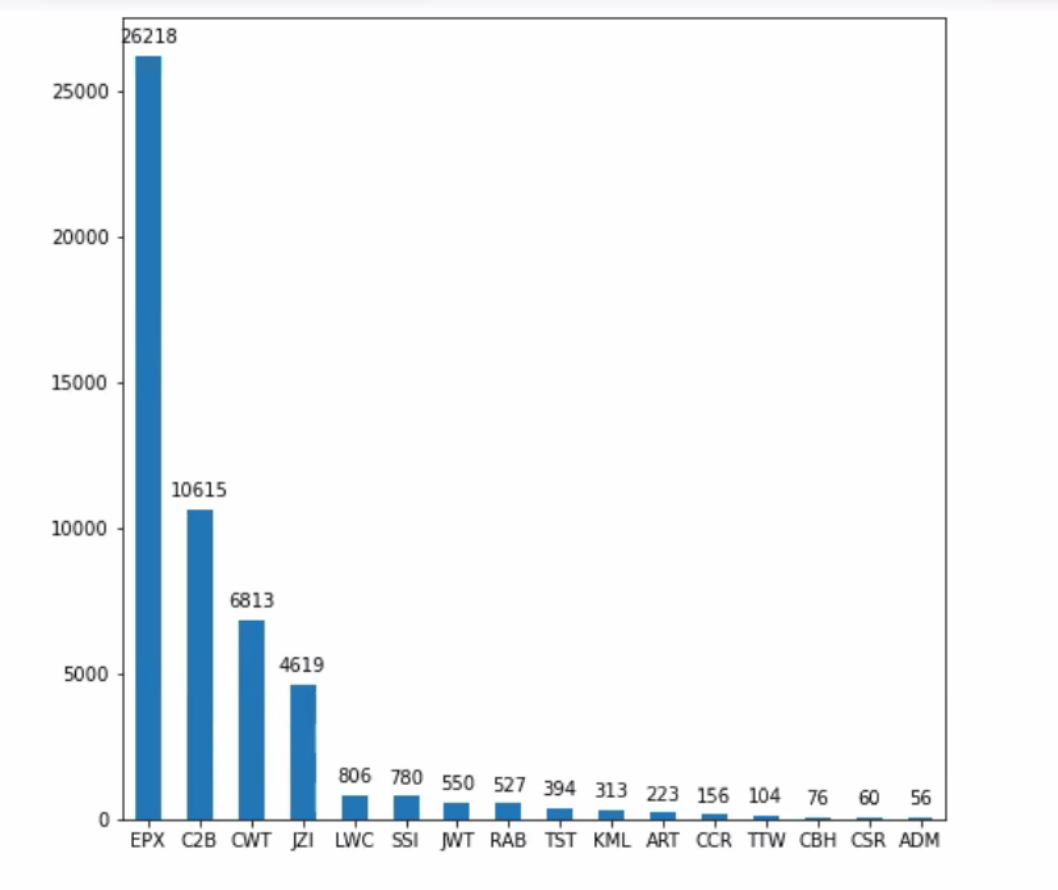
- How many object and numeric features are there?

Readings: We observe that except Duration, Net Sales, Commission (in value) and Age rest of them are of Categorical types. We have also observed that our target labels "Claim" is of Categorical type hence this problem statement comes under Classification Problem.

- Are there any feature which are not making sense and we can remove the same from dataset?

<u>Readings:</u> Post going through the entire data, we concluded that few of the feature are not going to help us in analysis and we can remove them.

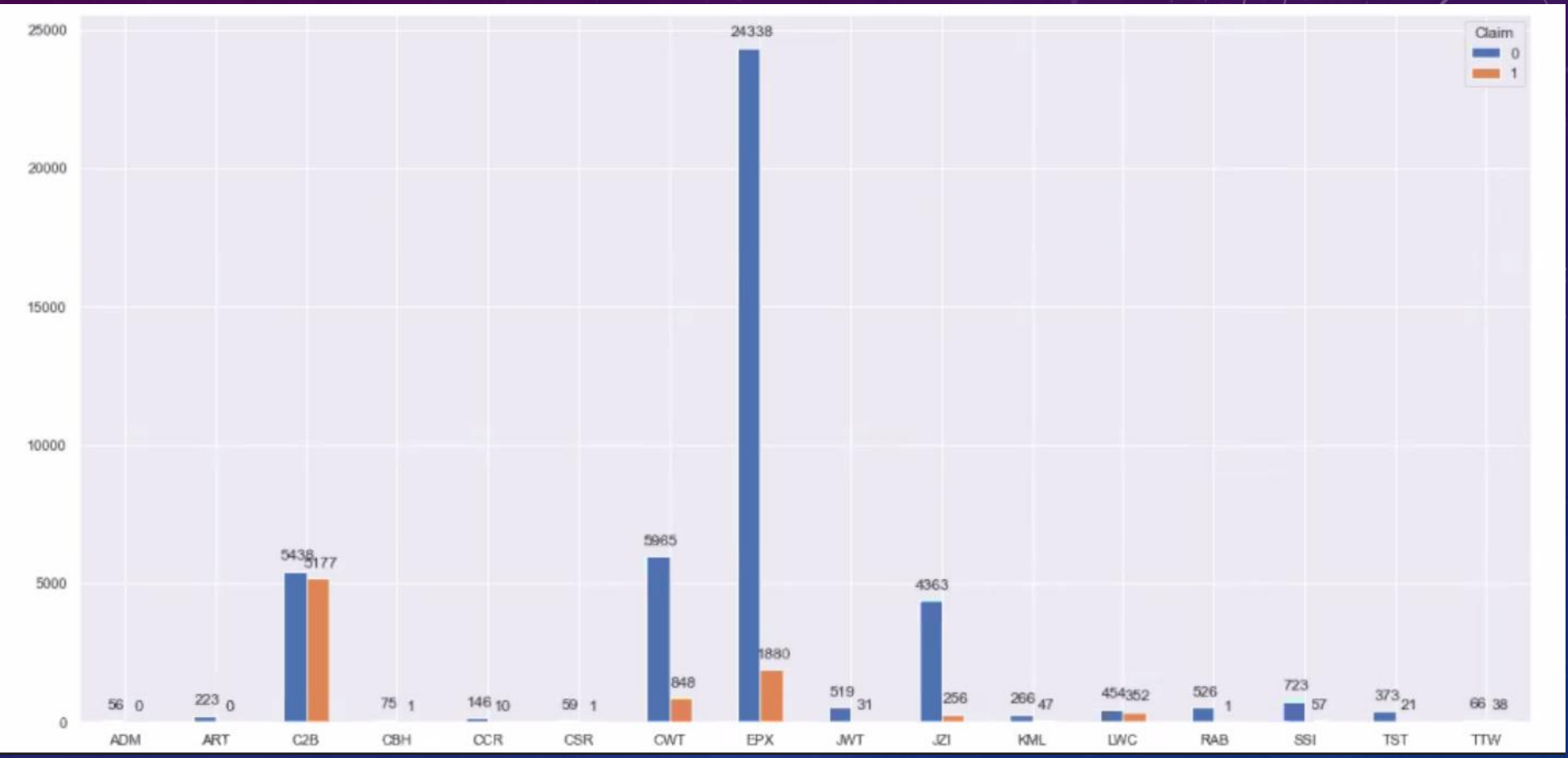
How many insurance policies were sold by each agency ?



Inference :

From above graph, we can conclude that EPX is selling most of the policies, while ADM is lowest in selling the policies

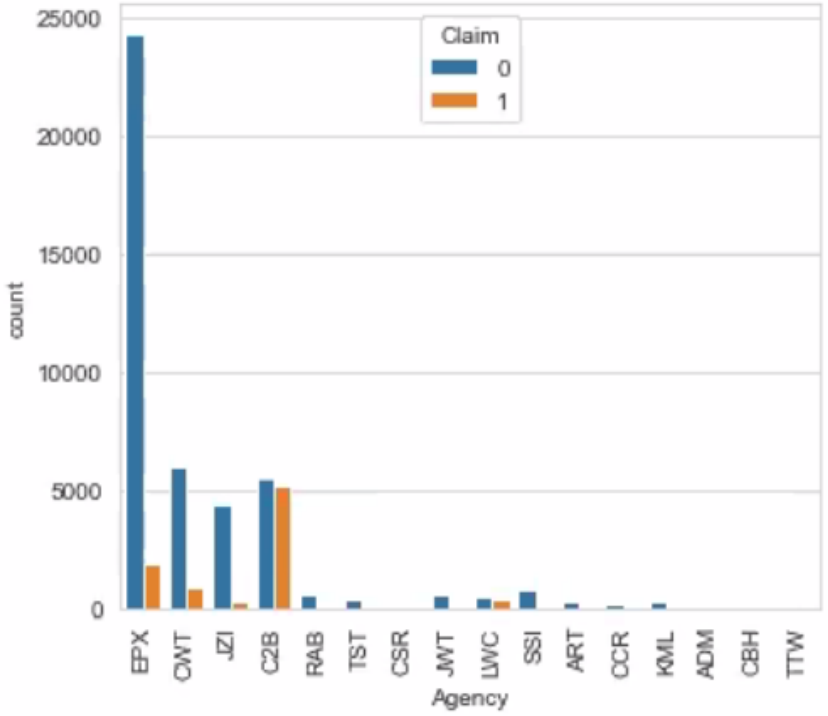
Claim analysis with respect to agency

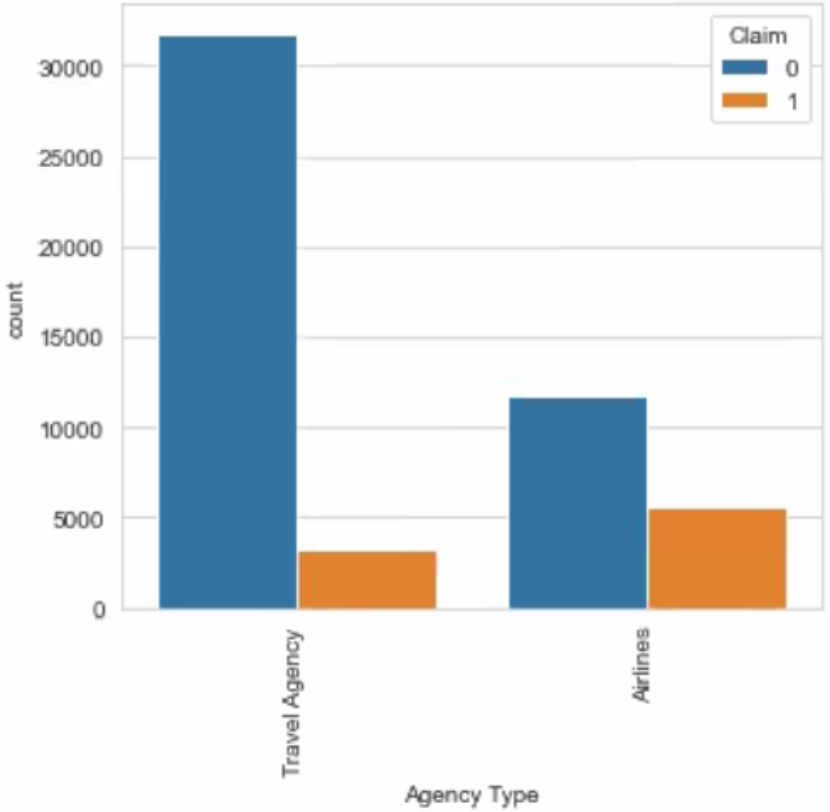


Inference :

From above graph, we observed the claim data with respect to each agency.

“Agency” & “Agency Type” wise “Claim Rate” & “Mean Commission”



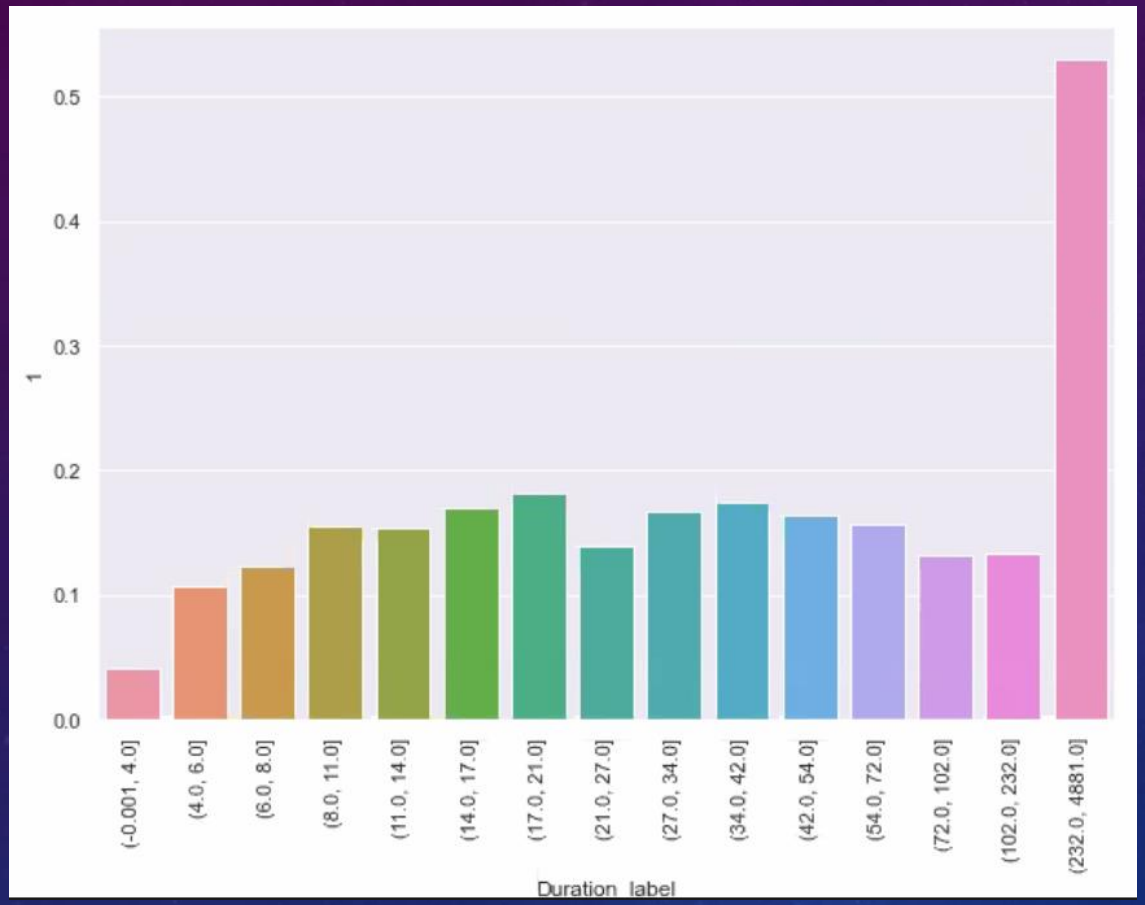


Inference :

Claim rate of particular agency and the line plot is for their average commission (Left graph)

Also , The Target variable is highly imbalanced. Not Claimed (Class 0) is dominating over Claimed(Class 1). Not Claimed is 83.33% and Claimed is 16.66%. (Right graph)

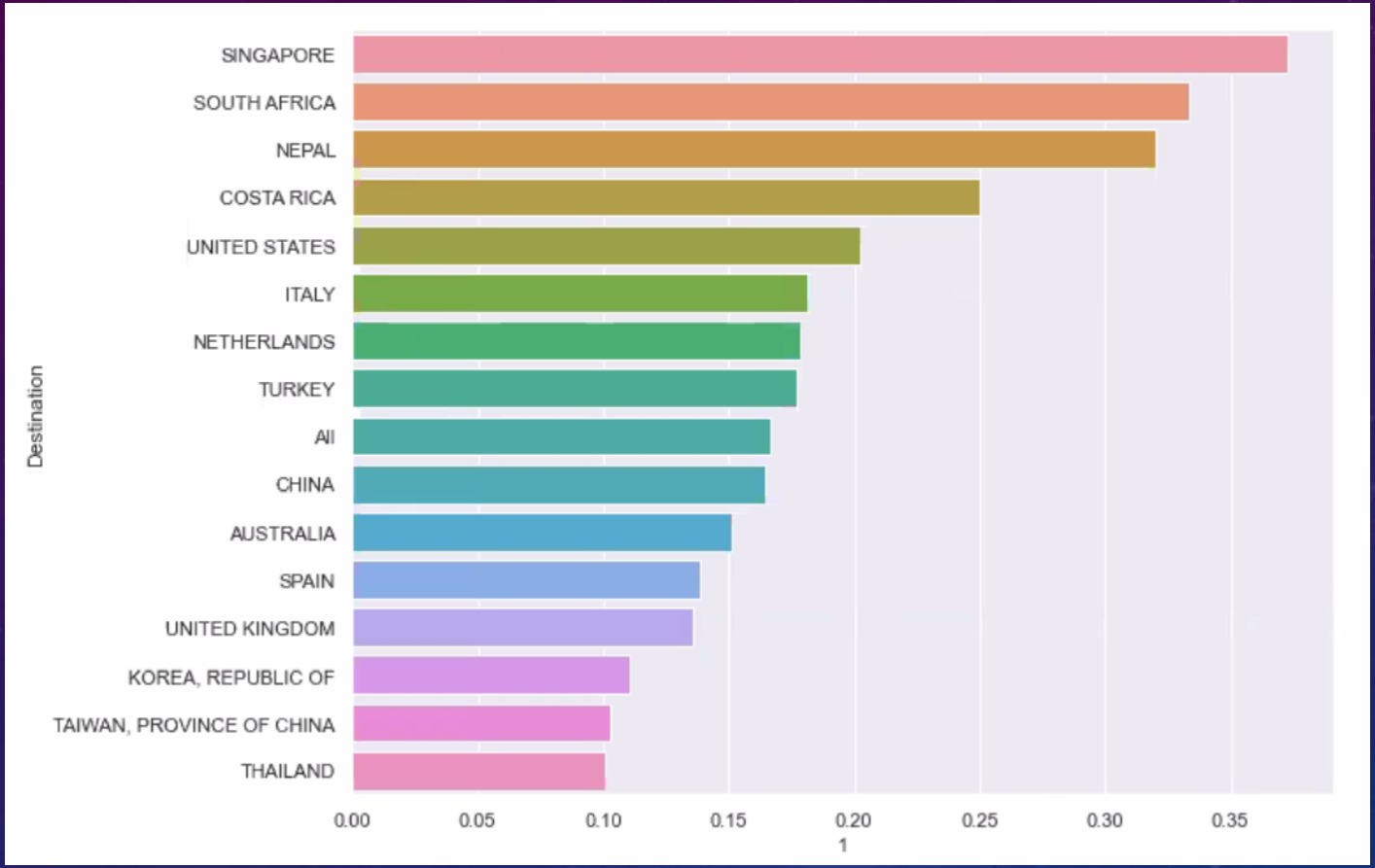
Duration impact on claim rate



Inference :

From above graph, we observed that which duration group is claimed maximum time

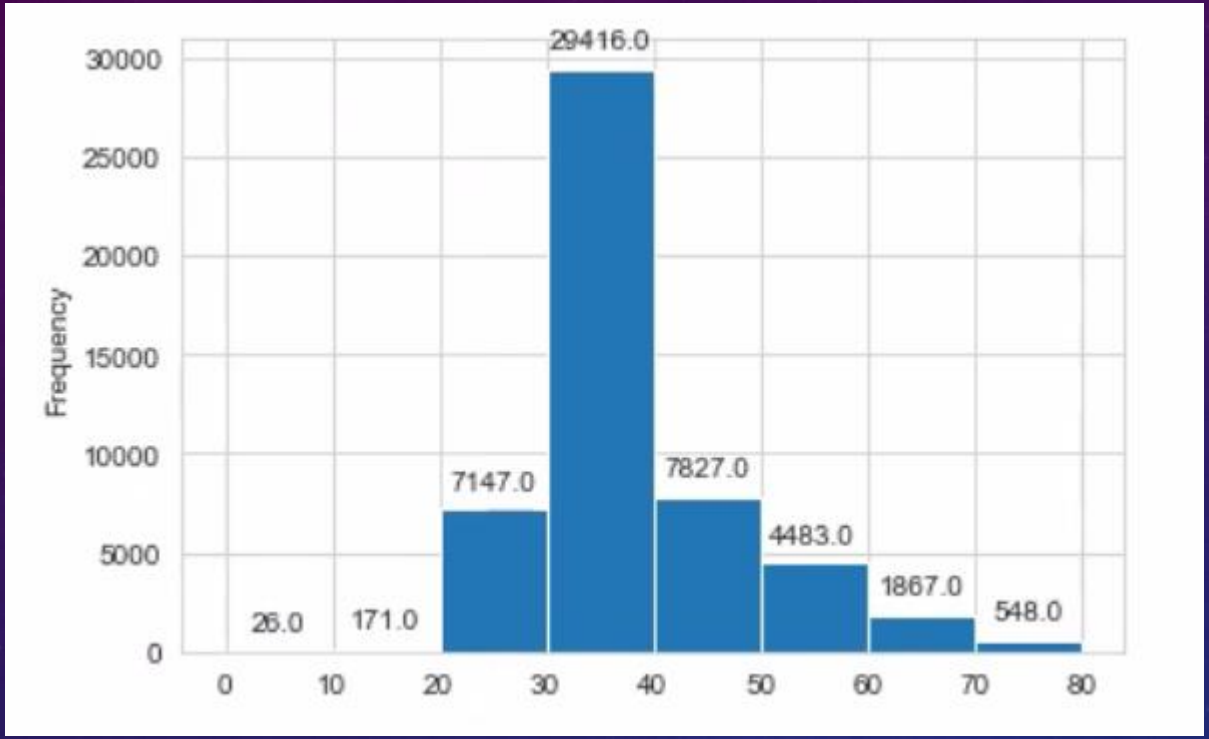
Impact of Destination on Claim rate



Inference :

From above graph, we observed that, Person visiting to Singapore is Claiming maximum time which is around (38%)

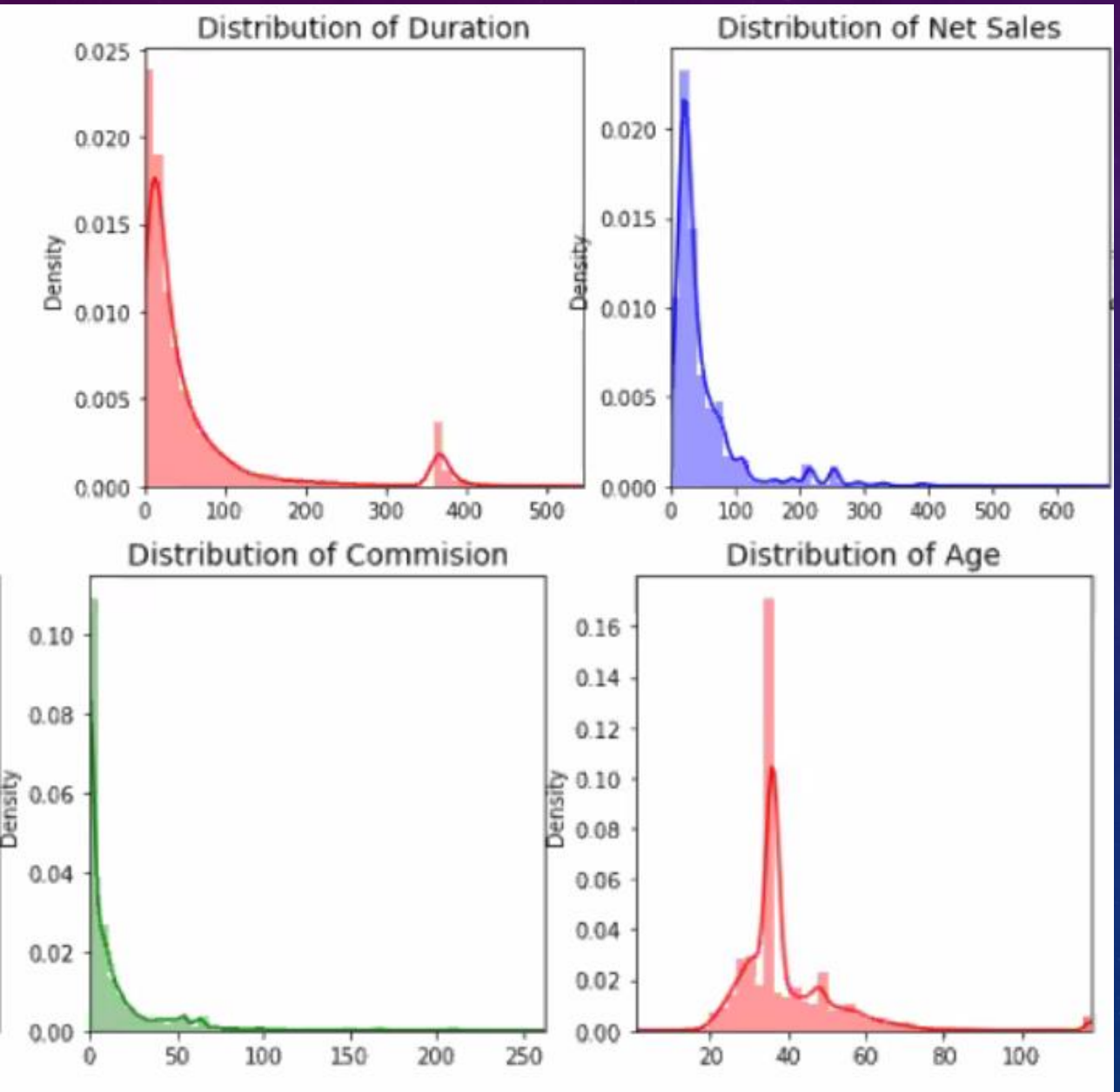
Age spread in sample



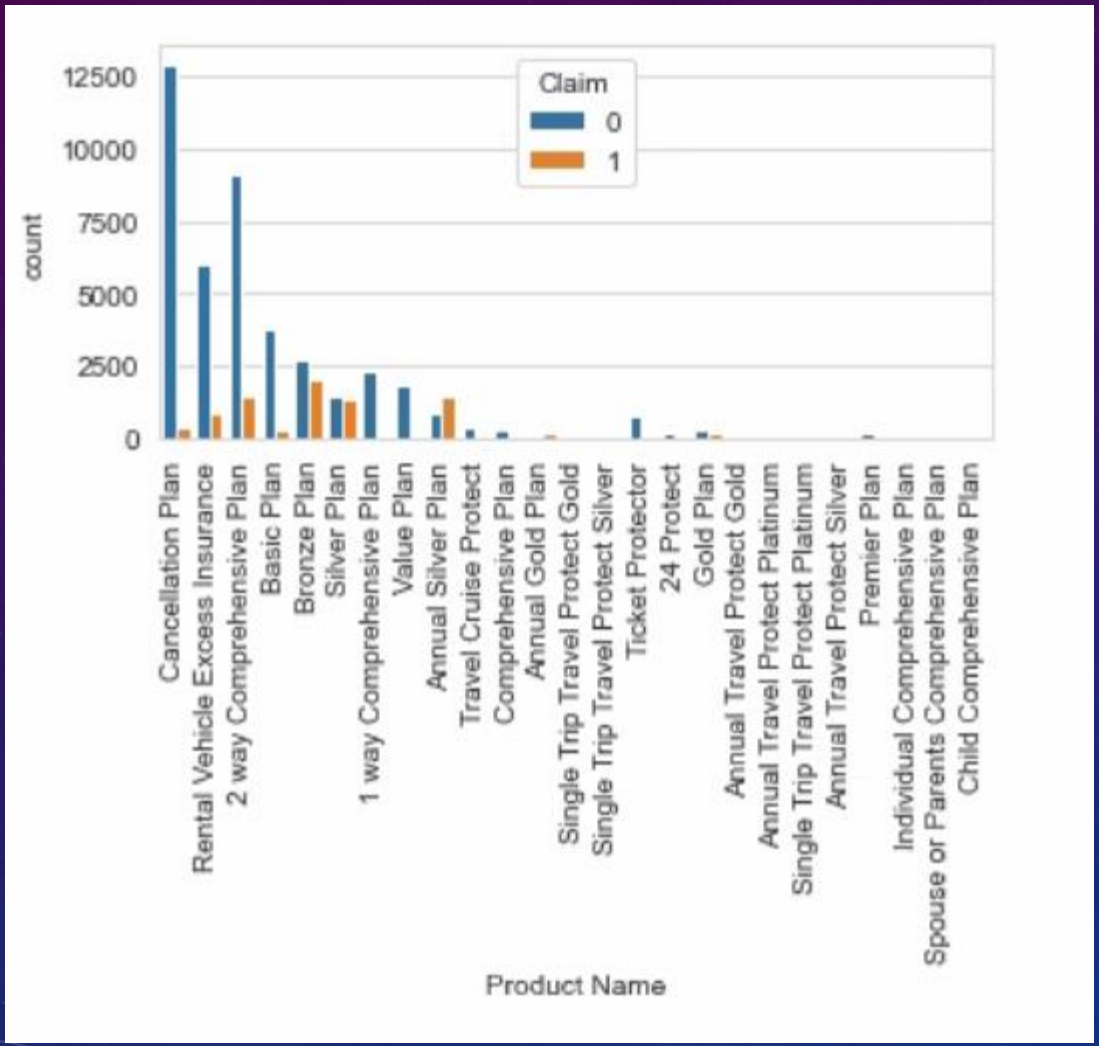
Inference

We can clearly observe that most of the customers lies in the Range of 25 to 55 years. One graph need to be added

Distribution of “Duration", "Net Sales”,”Commission”,”Age”



Product wise impact on Claim

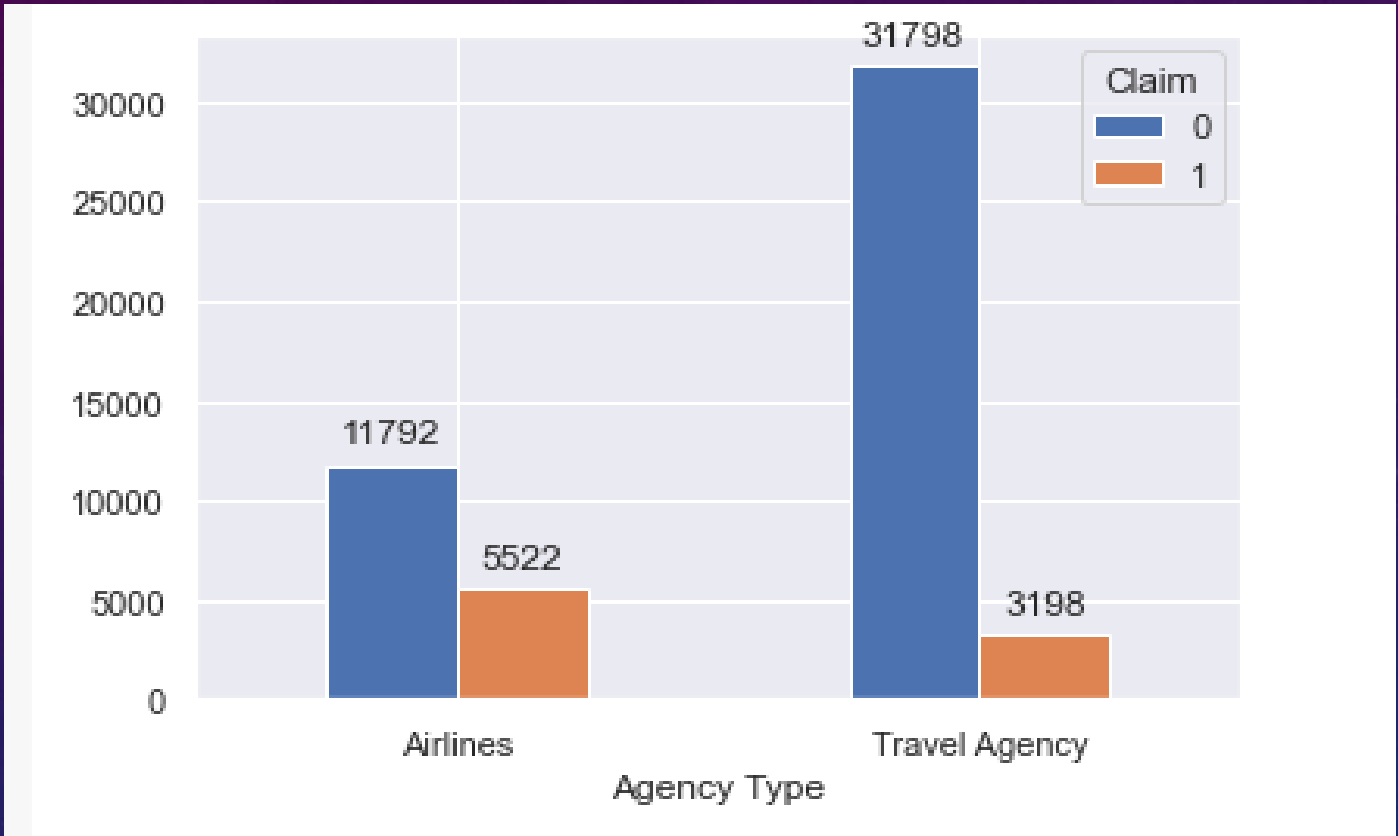


Inference :

- Here, from this graph we can observe that “Bronze Plan”, ”Annual Silver Plan", "Silver Plan” is Claimed most.

- Similarly, “Cancellation Plan”, ”2 Way Comprehensive Plan”, “Rental Vehicle Excess Insurance” is Most amongst Not Claimed.

Distribution of Claim across Agency Type?



Inference:

There are 43,590 observations (83.33%) for 'Not Claimed' insurances and only 8,720 (16.66%) 'Claimed' insurances. The data set is quite imbalanced. Most insurances are issued by Travel Agencies(66.90%) and the rest by Airlines.

Critical: 31.89% of insurances issued by Airlines are claimed whereas only 9.14% of insurances issued by Travel Agencies are claimed.
Insight: Claimed Insurances for Airlines is very high. The company should change the threshold of insurances sold by Airlines.

Insights for the 'ID' column

ID	Claim	Insurances sold	Net Sales
6031	1	13	1223.70
4299	5	9	1219.99
2200	5	11	1210.89
4972	2	11	1207.80
2566	0	15	1183.85
...	...	...	...
6130	2	7	-90.09

Inference:

From the plot above, these are top 5 customers who issued travel insurance from Safe Travel Inc multiple times and generated maximum sales for the company.

Insight:

Vouchers can be given to customers who give maximum sales to the company and not randomly. -marketing team, sales team

Observations for ID Column

ID	Claim	Insurances sold	Net Sales
5696	7	20	1143.86
8424	7	12	758.81

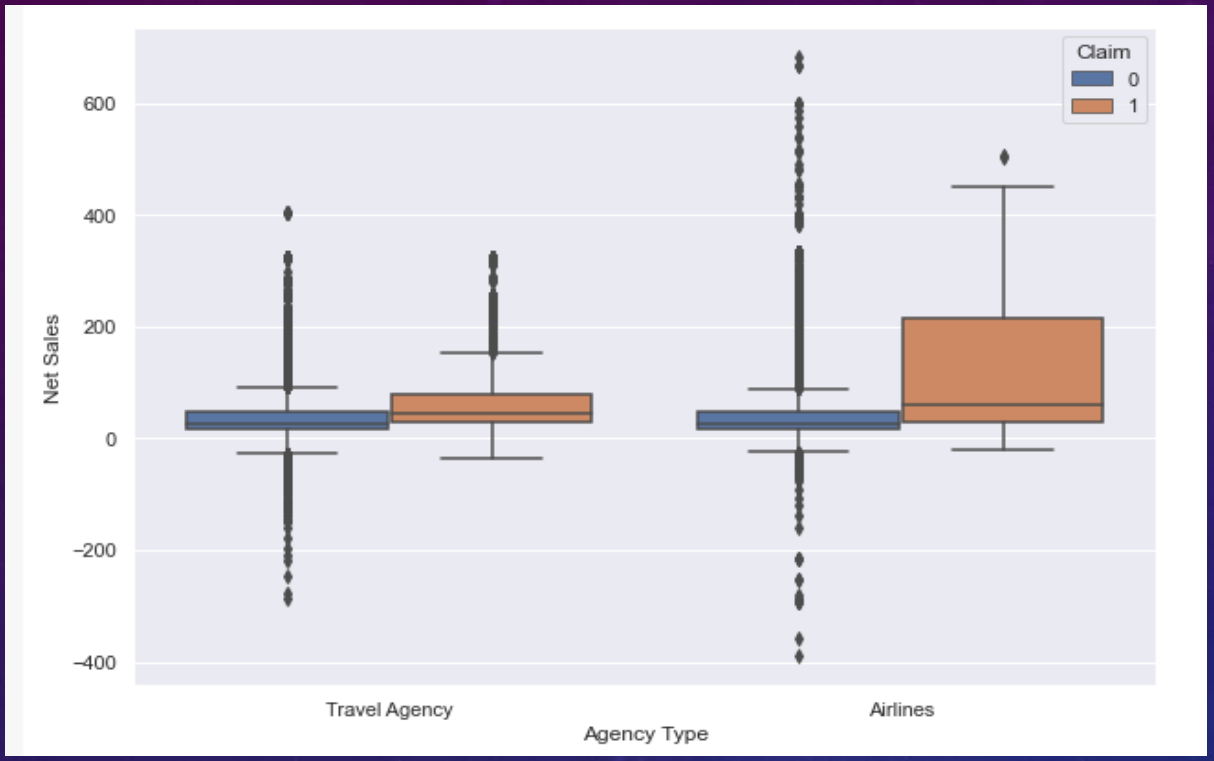
Inference:

- There are only 7992 unique customers who have bought multiple insurances. More than 50% customers in the data set have bought insurances between 5 to 8 times. Most customers have returned to buy travel insurances.
66.14%(5286) of customers have claimed their insurances at least once. 2706 customers (33.85%) were never claimed.

- 2 Customers (ID- 5696 & 8424) have claimed their insurance 7 times.

Insight: 83% of insurances issued are not claimed yet almost 66% customers get their claimed money. Simply from the data, most customers find it suitable to issue an insurance and hence come back several times but we don't know if the company is able to keep its profit margins.

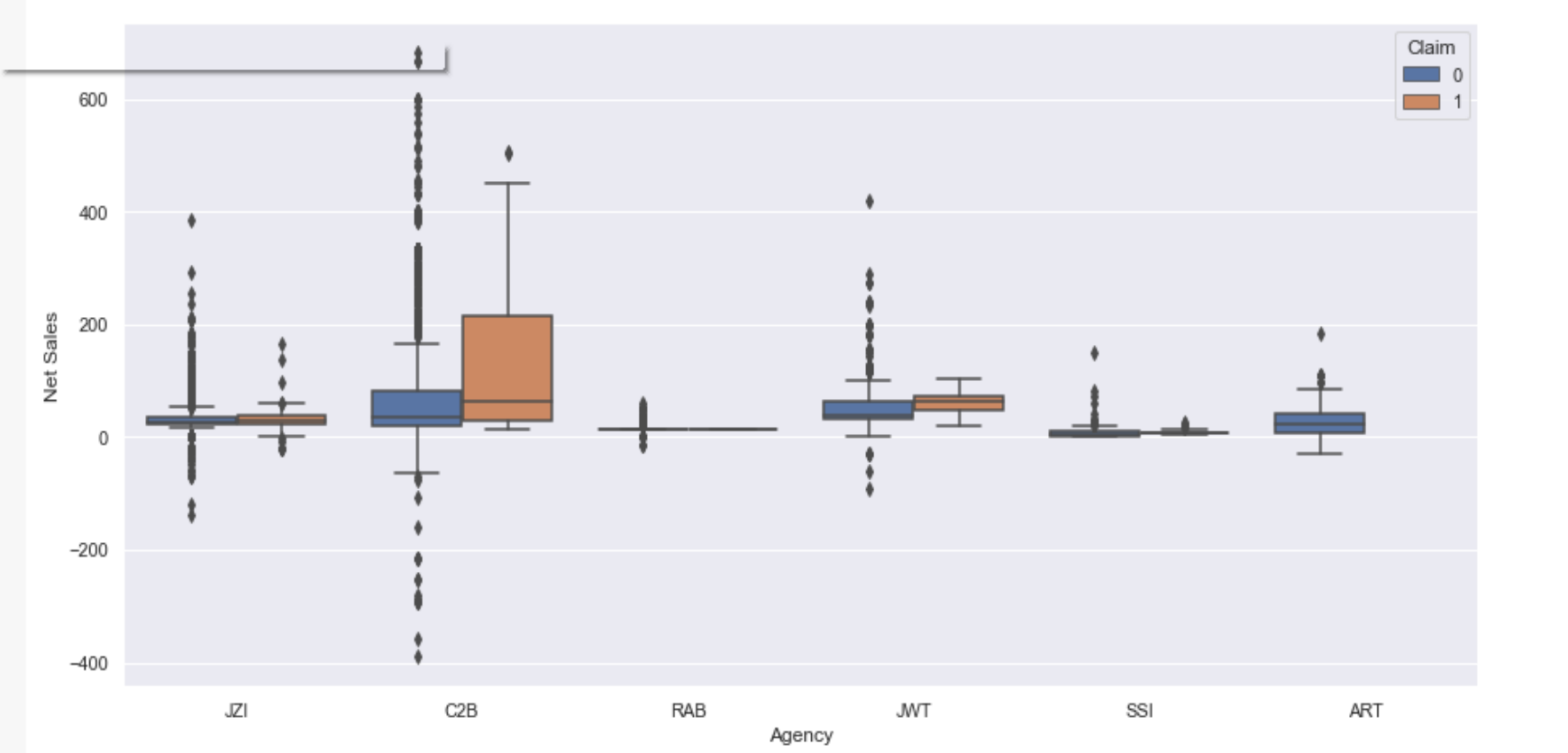
Columns- Agency, Agency Type, Product Name, Net Sales



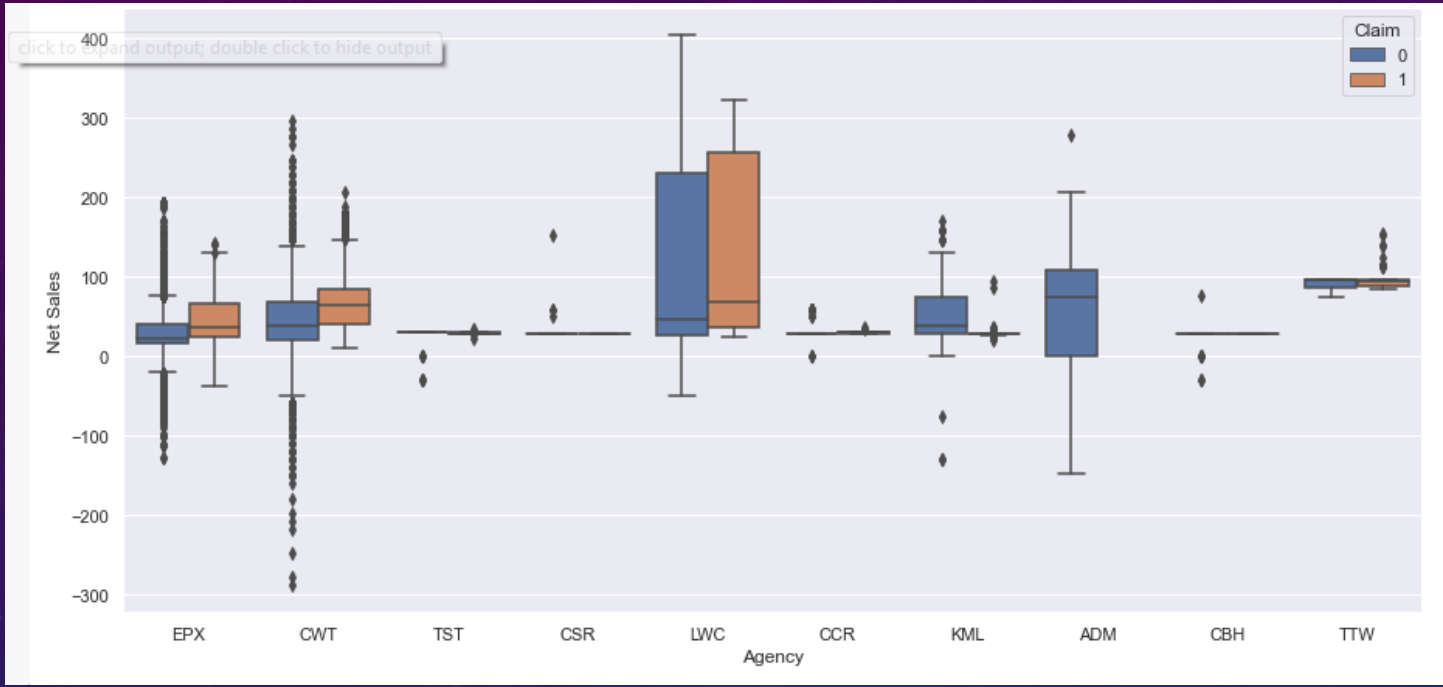
- From the data, 4,486 Insurances (31.24%) are claimed out of 14,358 insurances with Net Sales above 50 .
- Whereas, only 4,234 insurances (11.15%) are claimed out of 37,952 insurances with Net Sales below 50.

- Insurances with higher 'Net Sales' are claimed more often than with lower Net Sales. Though the total net sales from airlines & travel agencies are almost same because airlines are selling more expensive insurances, that is also the reason airlines has more insurances being claimed

Distribution of net sales across all agencies of Airlines.



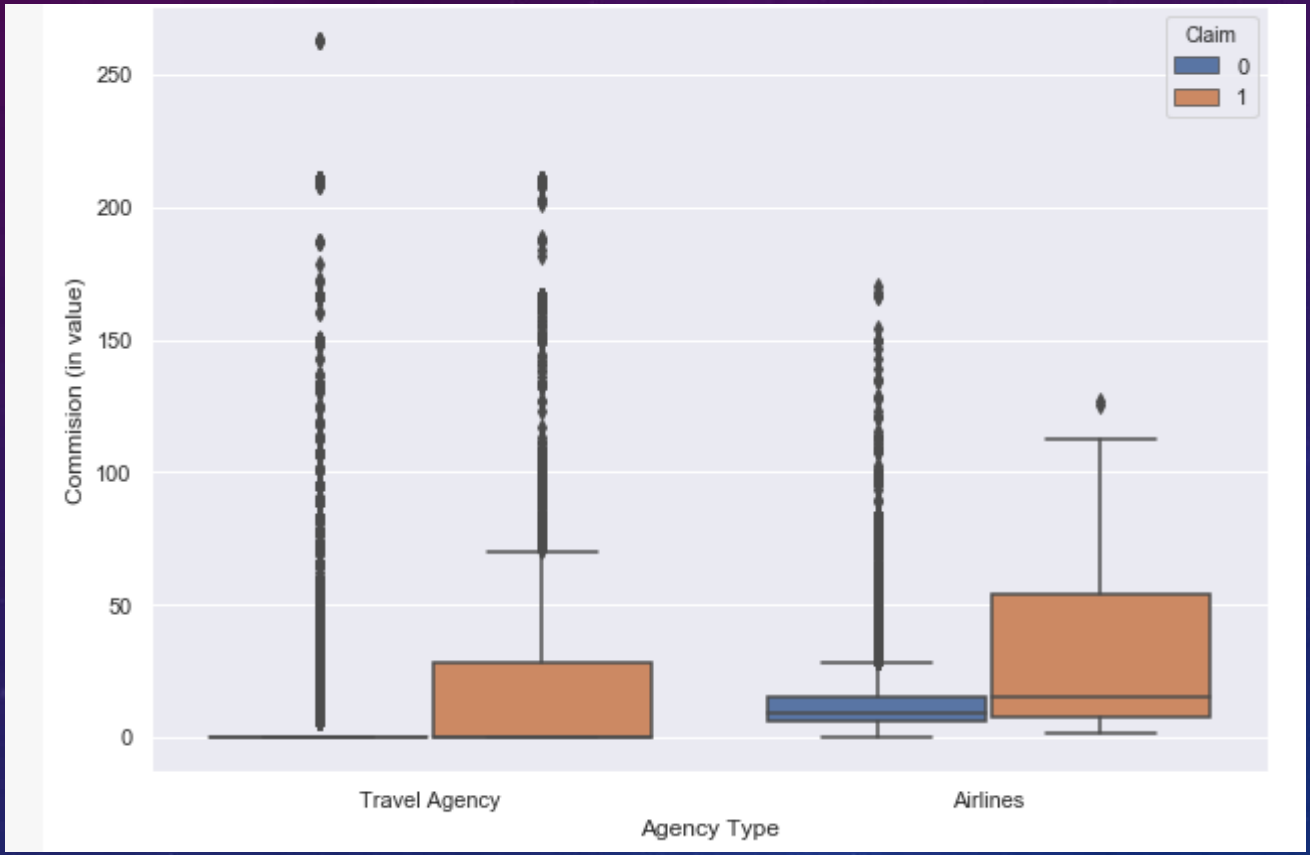
Distribution of net sales across all agencies of Travel Agencies.



Inference:

- The observation also fits well with how agencies sell insurances. Agencies such as C2B, LWC, EPX and CWT have more claimed insurances because they sell insurances for more Net sales.

How Commission effects the claim on The Basis of Agencies



Model Performances

Model	Accuracy	Precision	Recall
Random Forest	94	87	79
XG Boost	95	88	84
ADA Boost	94	81	84
Decision Tree	92	83	73

Best Model :

XG Boost was best , based on precision score

Conclusion

- Bronze Plan, Silver Plan and Annual Silver Plan **needs modifications in the product features** due to higher commission to sales roll outs and claim rates.

- Average age of personas travelling is 40 years.

- Destination wise, persons travelling to Singapore, South Africa and Nepal are claiming most.

- In Products, 'Annual Gold Plan‘ , 'Annual Travel Protect Gold‘ , 'Annual Silver Plan' and 'Single Trip Travel Protect Platinum' have Higher claim Percentage

Thank You

Team :
Prashant Yadav
Abhishek Bhoite
Manisha Kalita
Rohit gawande