Travel Insurance Claim Prediction

Asyraf Al Rasyid

Table of Content

Business Problem

Problem Statement

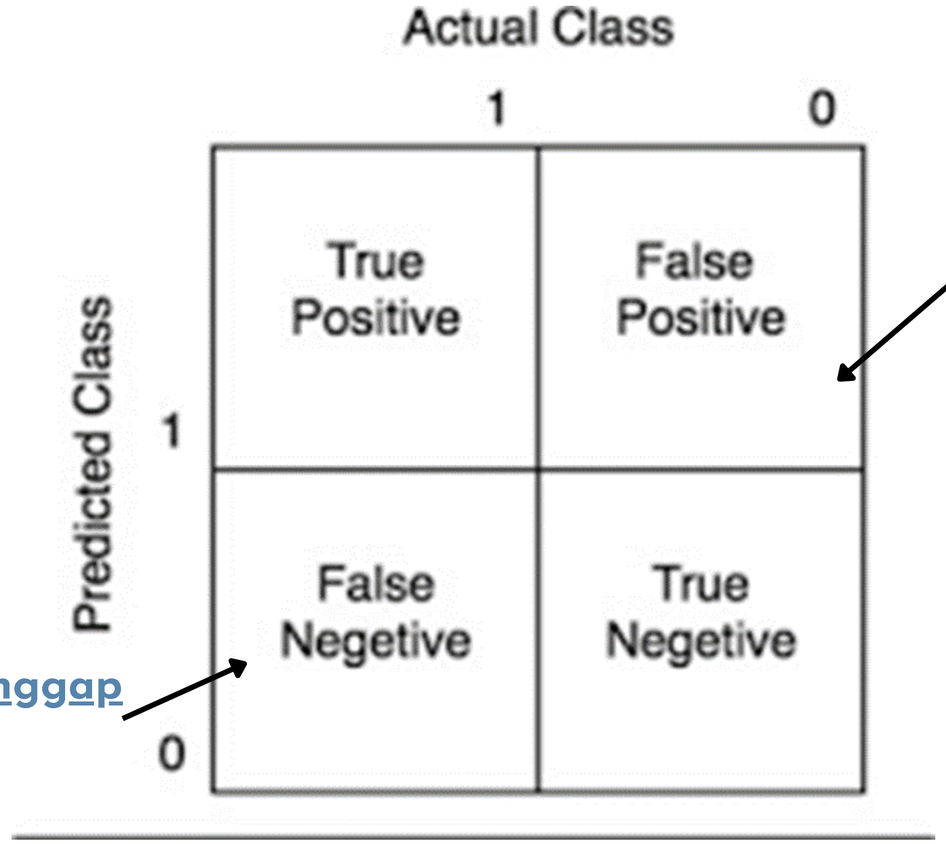
pengajuan klaim yang tidak terduga dapat memberikan dampak terhadap perusahaan. Model ini diharapkan dapat membantu perusahaan dalam membuat keputusan yang lebih tepat dalam menentukan apakah seseorang akan **melakukan klaim atau tidak pada asuransi perjalanannya**

Goals

Mengembangkan model prediktif yang dapat dengan akurat memprediksi kemungkinan klaim berdasarkan data yang tersedia.

Analytic Approach

menemukan suatu pola yang dapat membedakan pemegang polis yang akan melakukan klaim asuransi dan yang tidak



Seseorang yang dianggap tidak klaim tapi klaim

Seseorang yang dianggap klaim tapi todak klaim

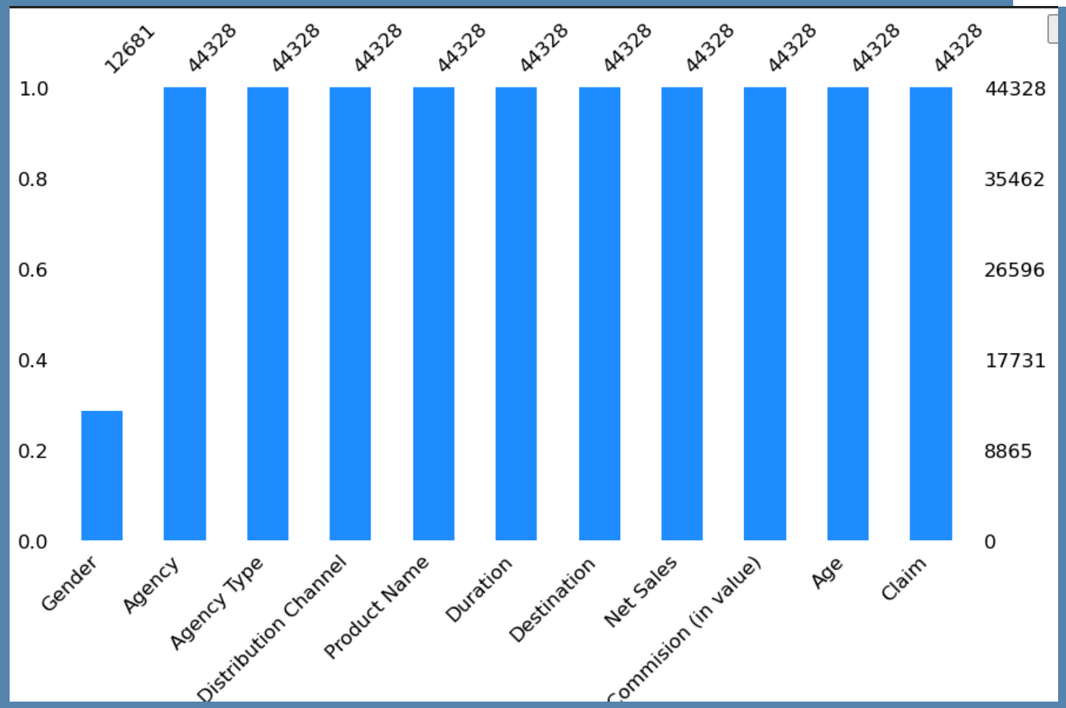
Data Understanding

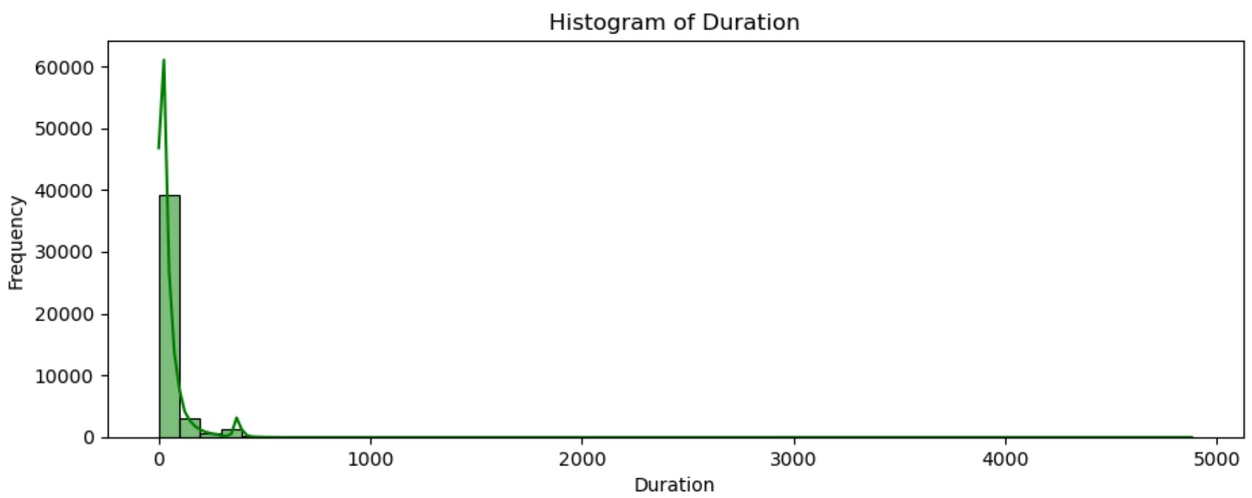
Jumlah baris dan kolom: (44328, 11)

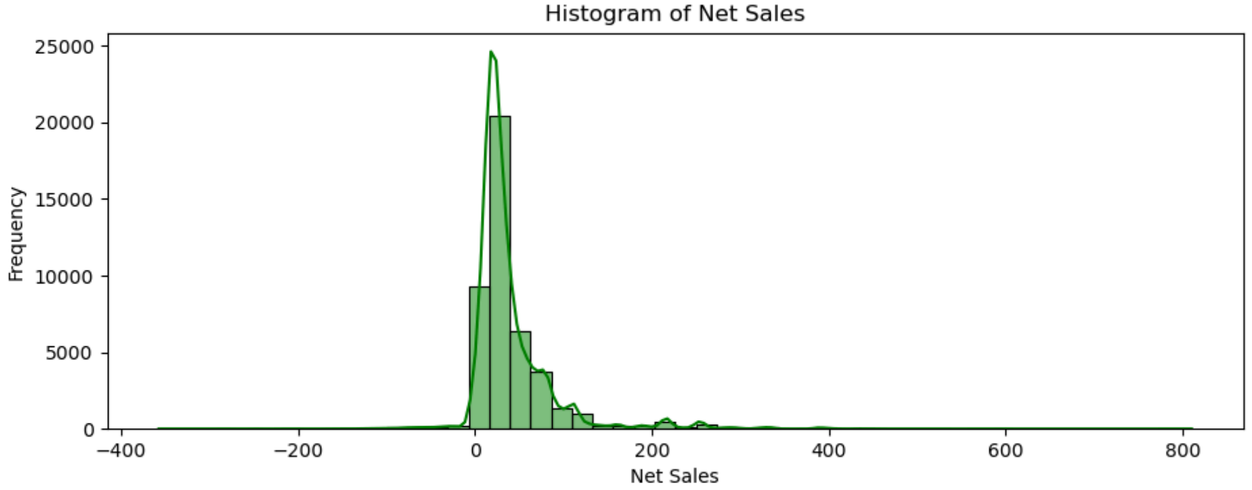
- Agency: Nama agensi asuransi perjalanan
- Agency Type: Tipe asuransi perjalanan
- Distribution Channel: Penyaluran produk kepada konsumen
- Product Name: Jenis produk asuransi yang digunakan turis
- Gender: Jenis kelamin
- Duration: Durasi perjalanan dalam hari
- Destination: Tujuan perjalanan
- Net Sales: Jumlah harga penjualan asuransi perjalanan (dalam dollar singapore)
- Commission (in value): Komisi yang didapatkan agensi asuransi perjalanan
- Age: Usia turis
- Claim: Claim status.

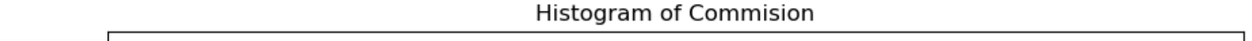
	Duration	Net Sales	Commision (in value)	Age
count	44328.000000	44328.000000	44328.000000	44328.000000
mean	49.424292	40.550948	9.707692	39.925600
std	109.153961	48.661970	19.625637	13.954926
min	-1.000000	-357.500000	0.000000	0.000000
25%	9.000000	18.000000	0.000000	35.000000
50%	22.000000	26.500000	0.000000	36.000000
75%	53.000000	48.000000	11.550000	43.000000
max	4881.000000	810.000000	283.500000	118.000000

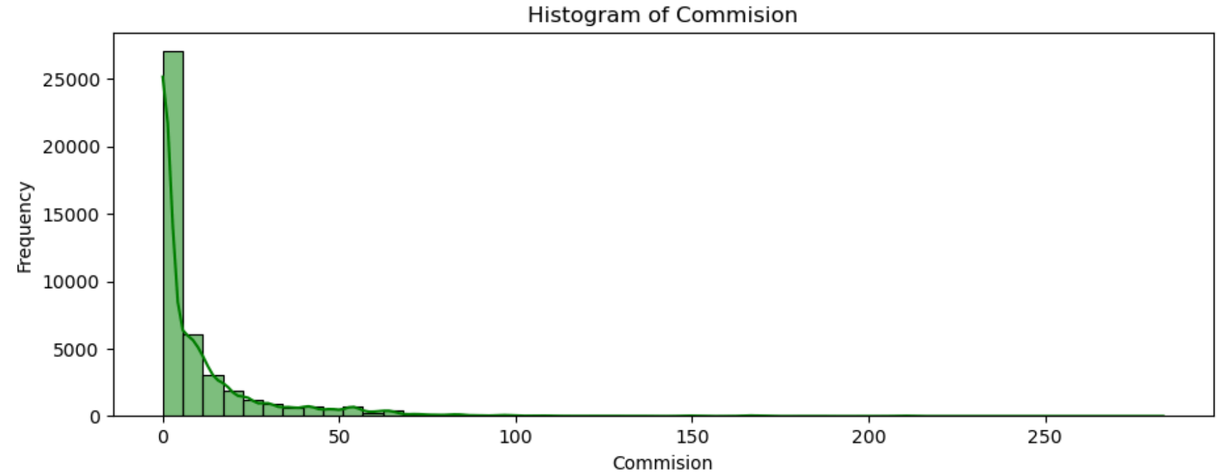
	Agency	Agency Type	Distribution Channel	Product Name	Gender	Destination	Claim
count	44328	44328	44328	44328	12681	44328	44328
unique	16	2	2	26	2	138	2
top	EPX	Travel Agency	Online	Cancellation Plan	M	SINGAPORE	No
freq	24656	32113	43572	12979	6504	9267	43651

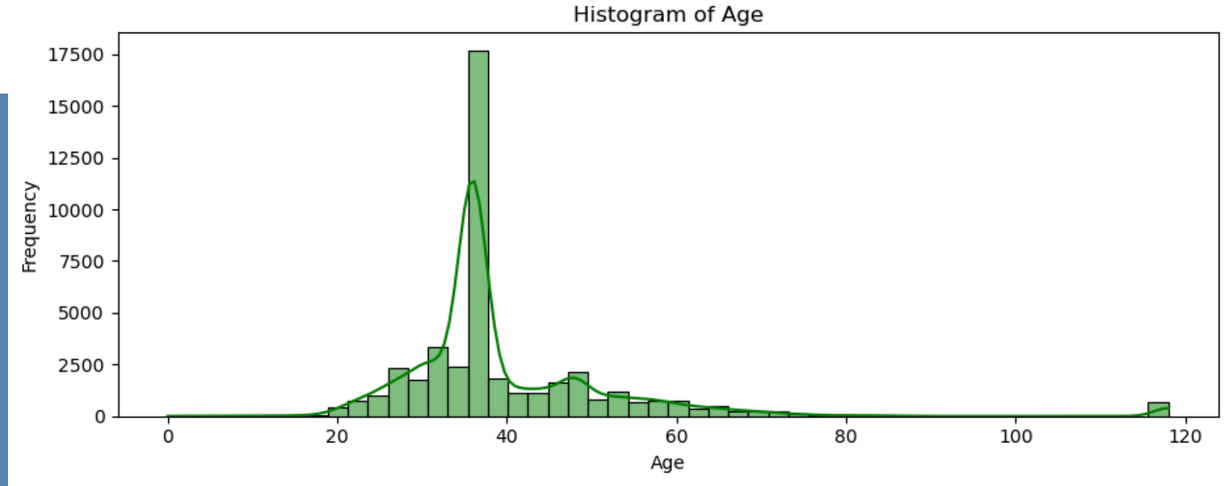


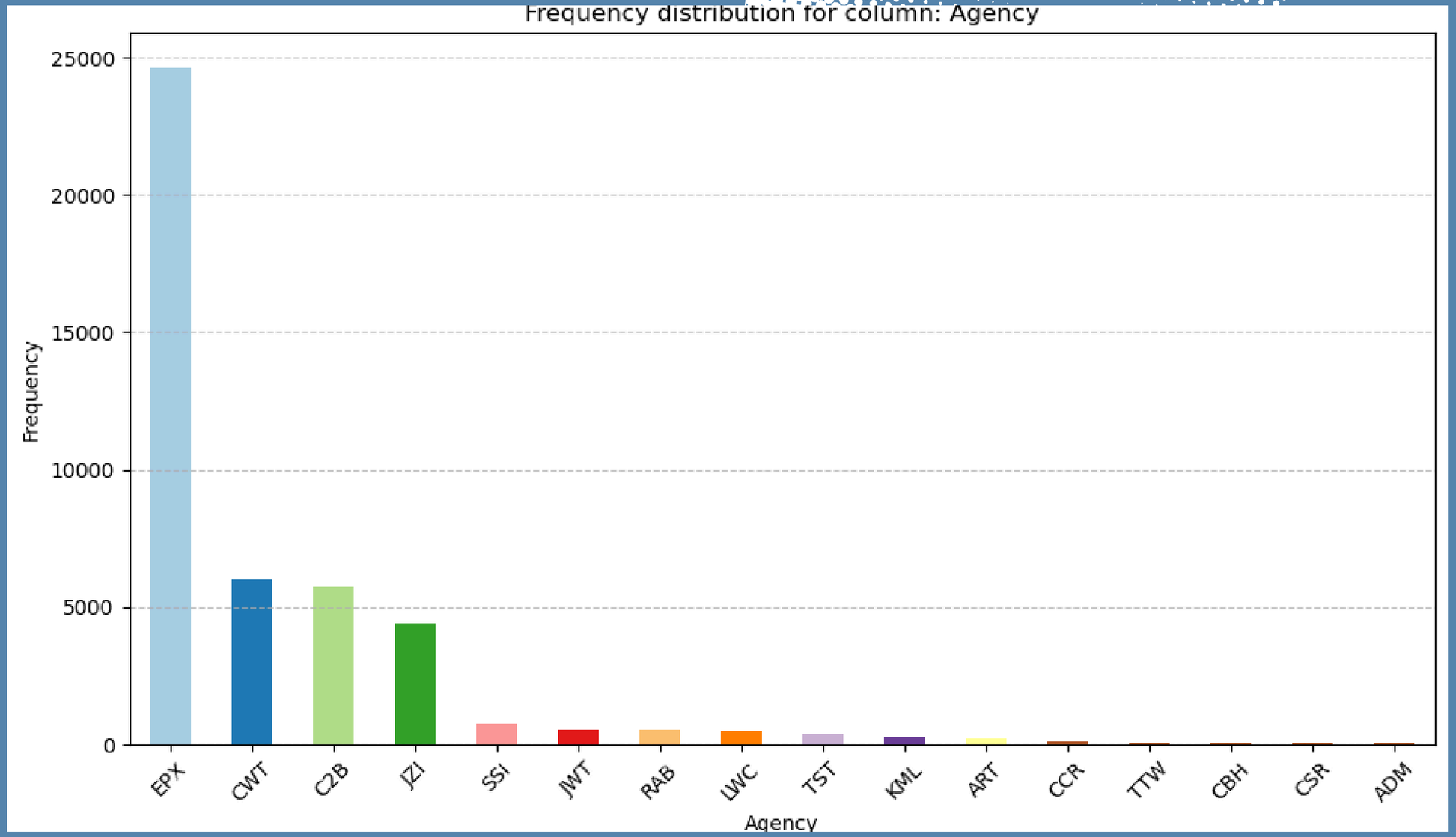

Histogram of Duration
Frequency
Duration
0
10000
20000
30000
40000
50000
60000
0
1000
2000
3000
4000
5000

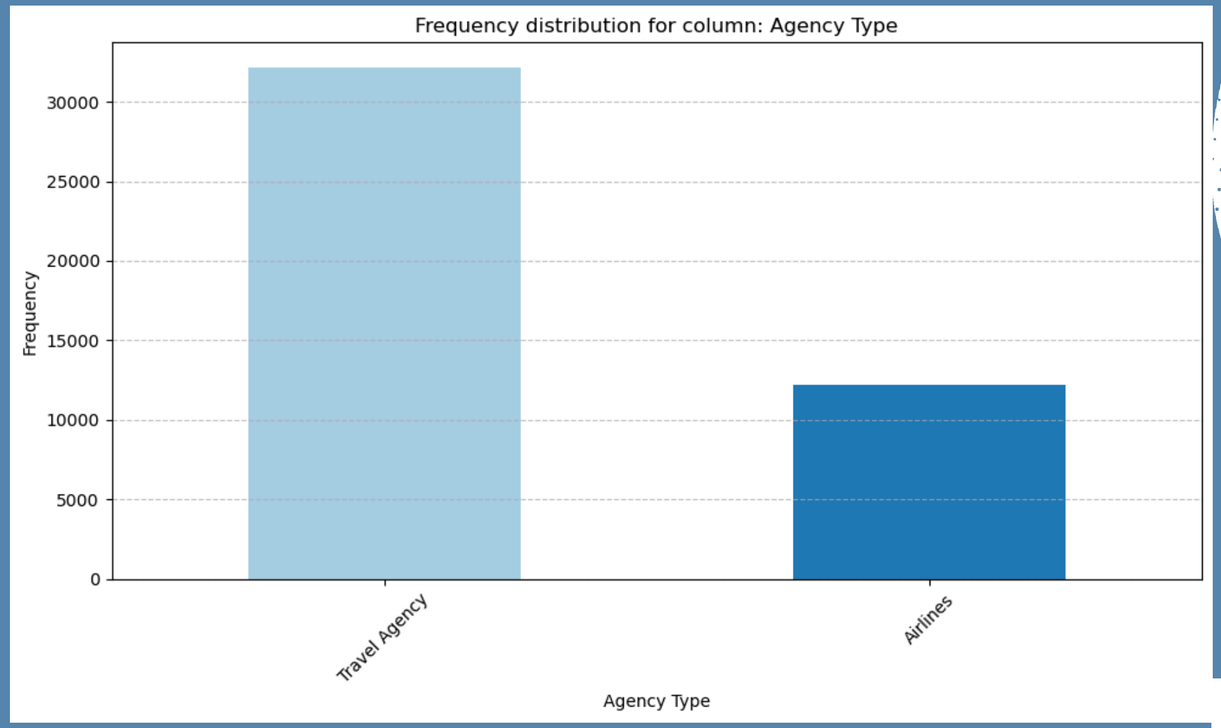

Histogram of Net Sales
Frequency
Net Sales
0
5000
10000
15000
20000
25000
−400
−200
0
200
400
600
800

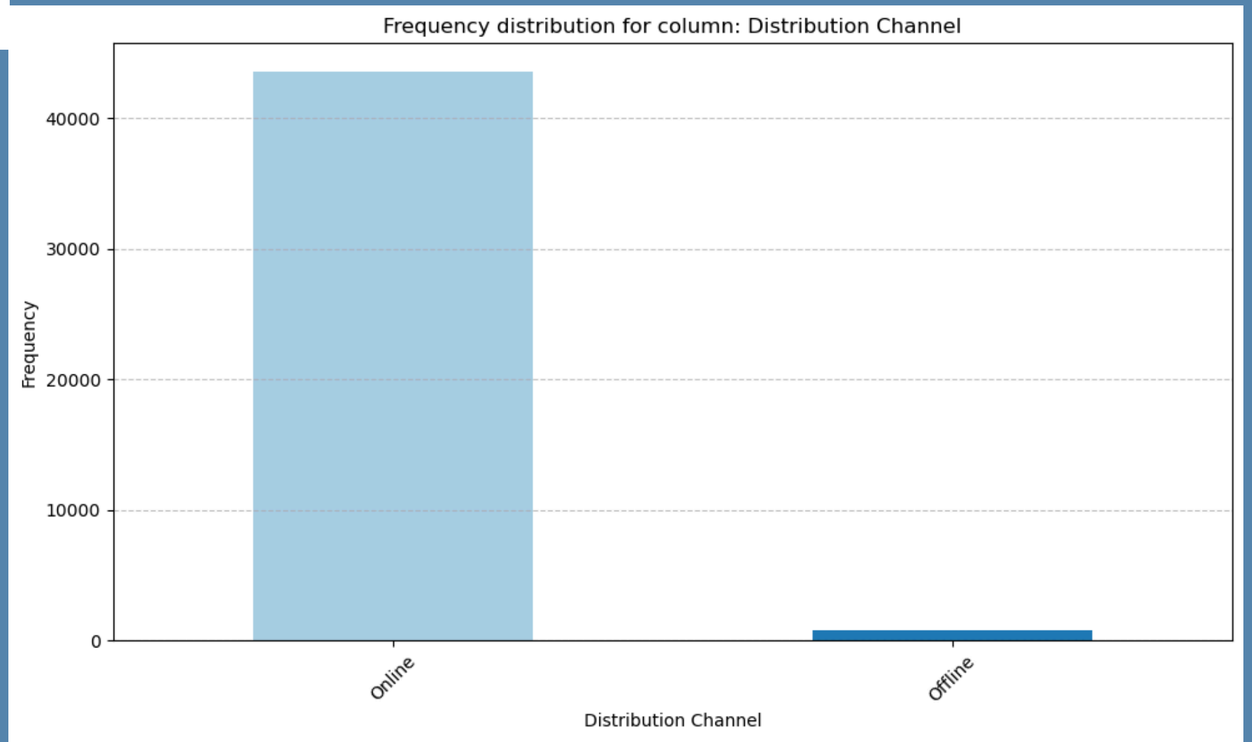

Histogram of Commision

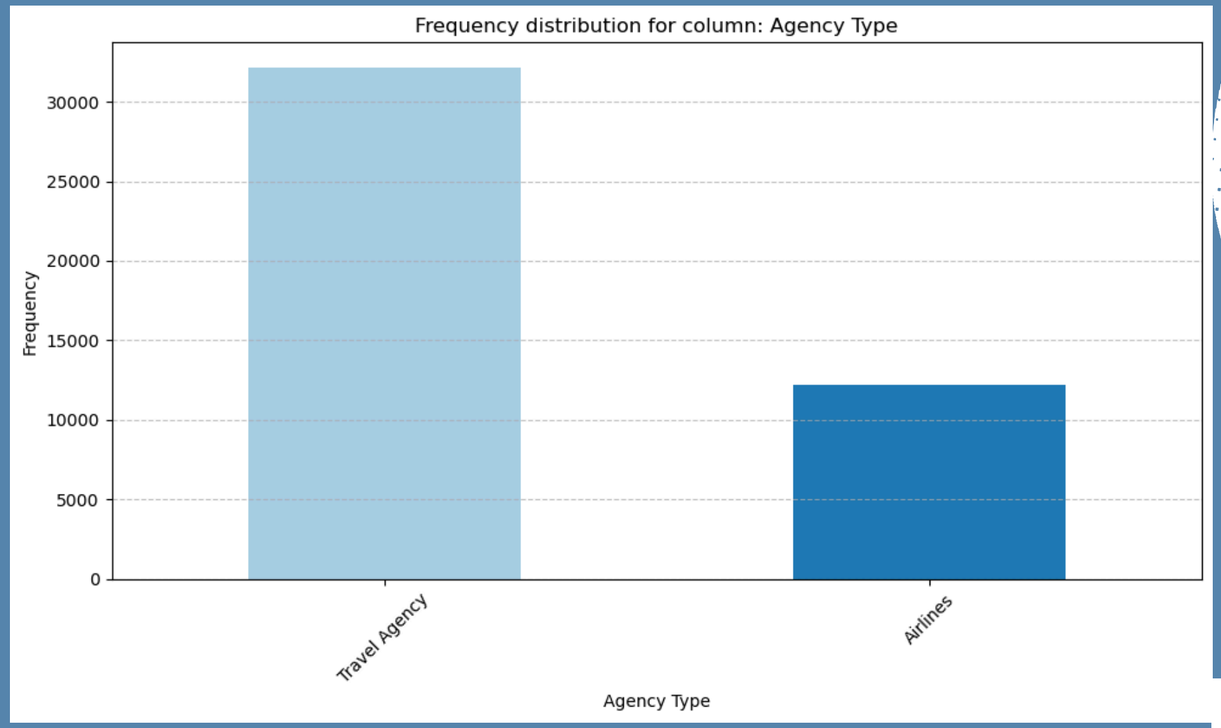

Histogram of Commision
Frequency
Commision
0
5000
10000
15000
20000
25000
0
50
100
150
200
250

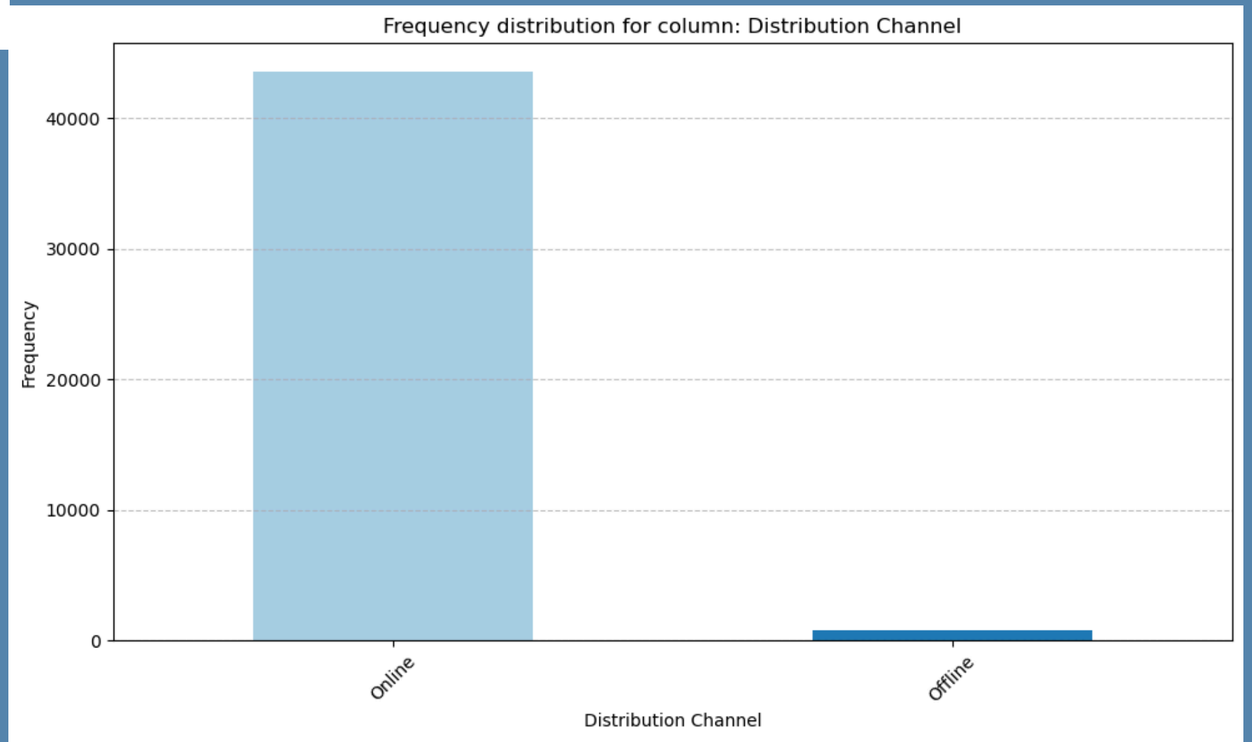

Histogram of Age
Frequency
Age
0
2500
5000
7500
10000
12500
15000
17500
0
20
40
60
80
100
120

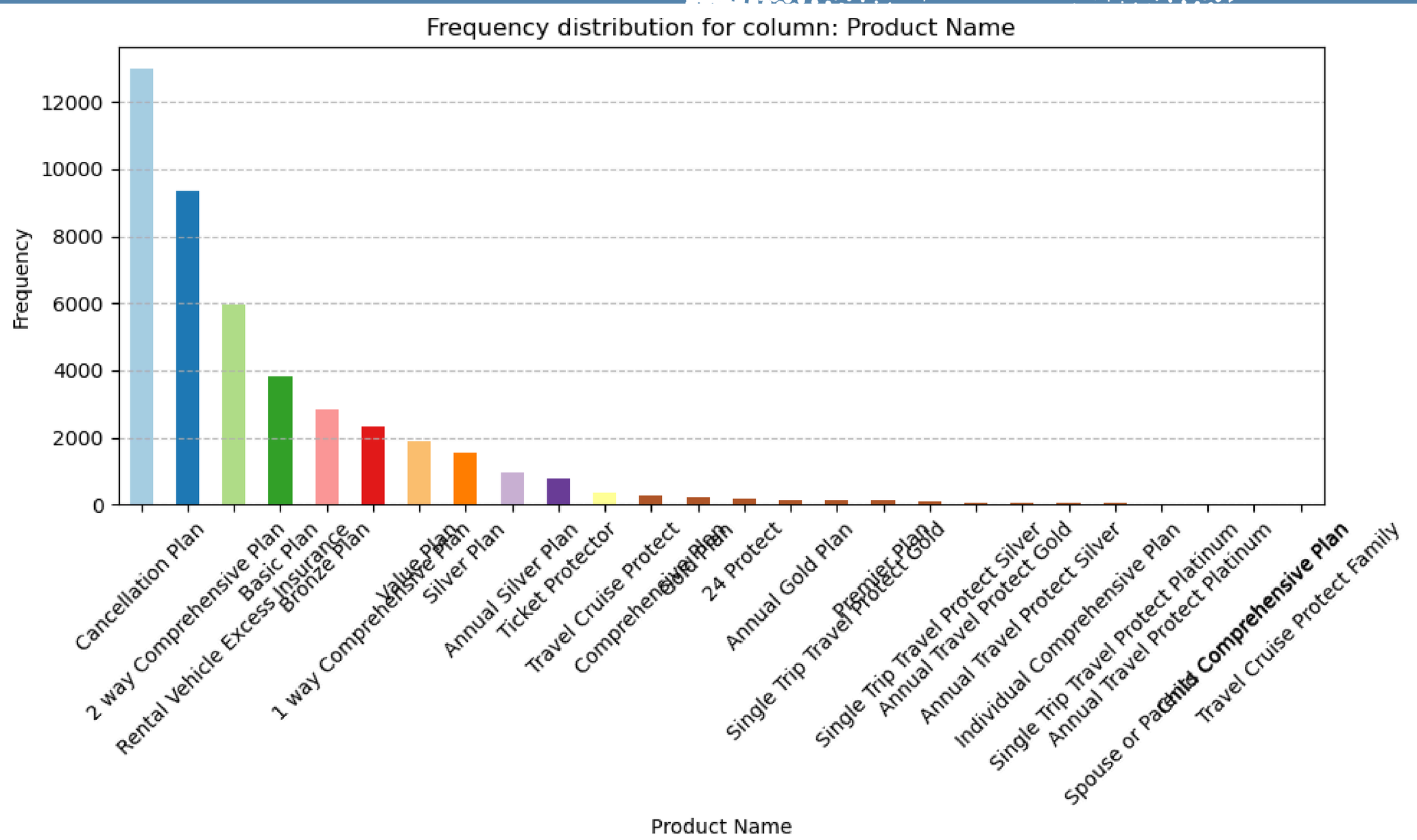

Frequency distribution for column: Agency
Frequency
25000
20000
15000
10000
5000
0
EPX
CWT
C2B
JZI
SSI
JWT
RAB
LWC
TST
KML
ART
CCR
TTW
CBH
CSR
ADM
Agency

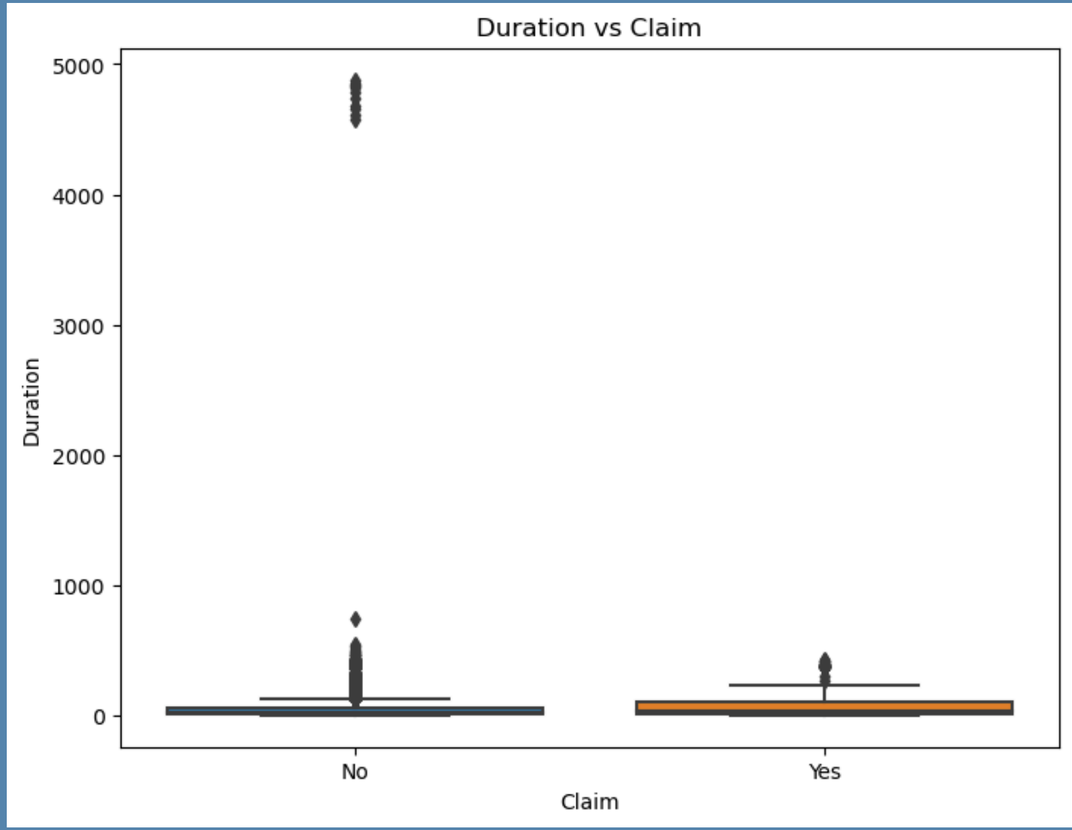

Frequency distribution for column: Agency Type
Frequency
30000
25000
20000
15000
10000
5000
0
Travel Agency
Airlines
Agency Type

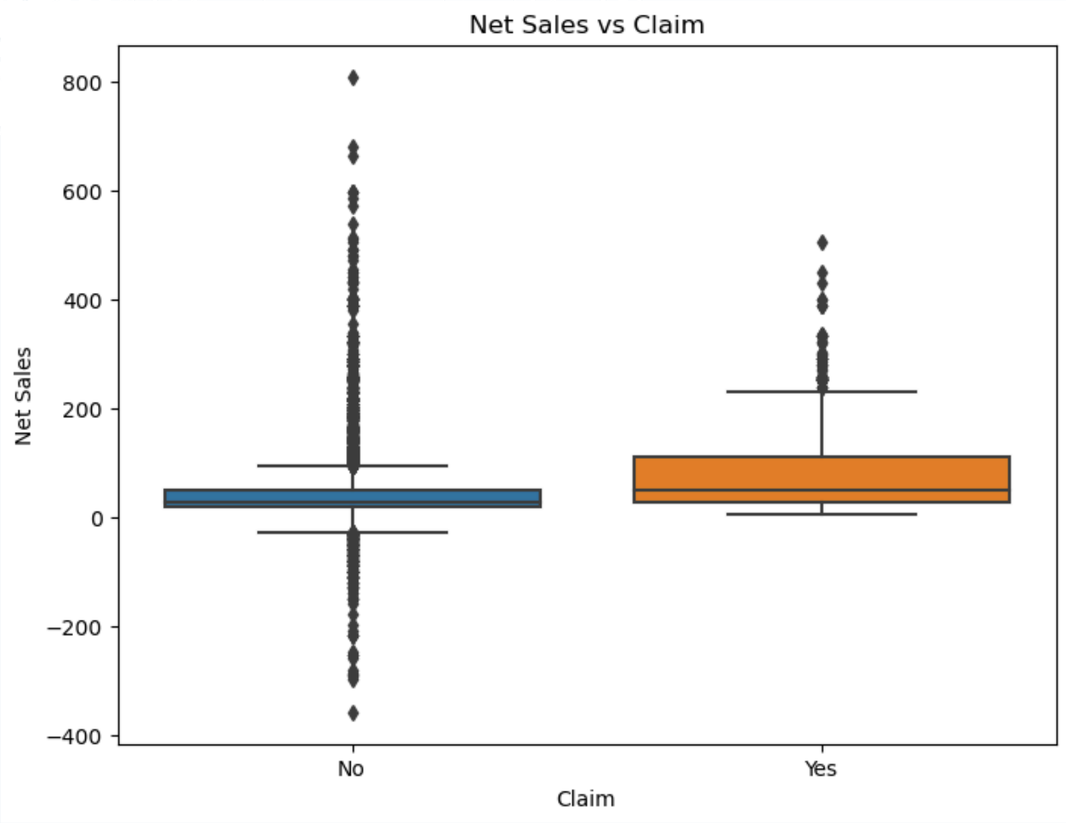

Frequency distribution for column: Distribution Channel
Frequency
40000
30000
20000
10000
0
Online
Offline
Distribution Channel

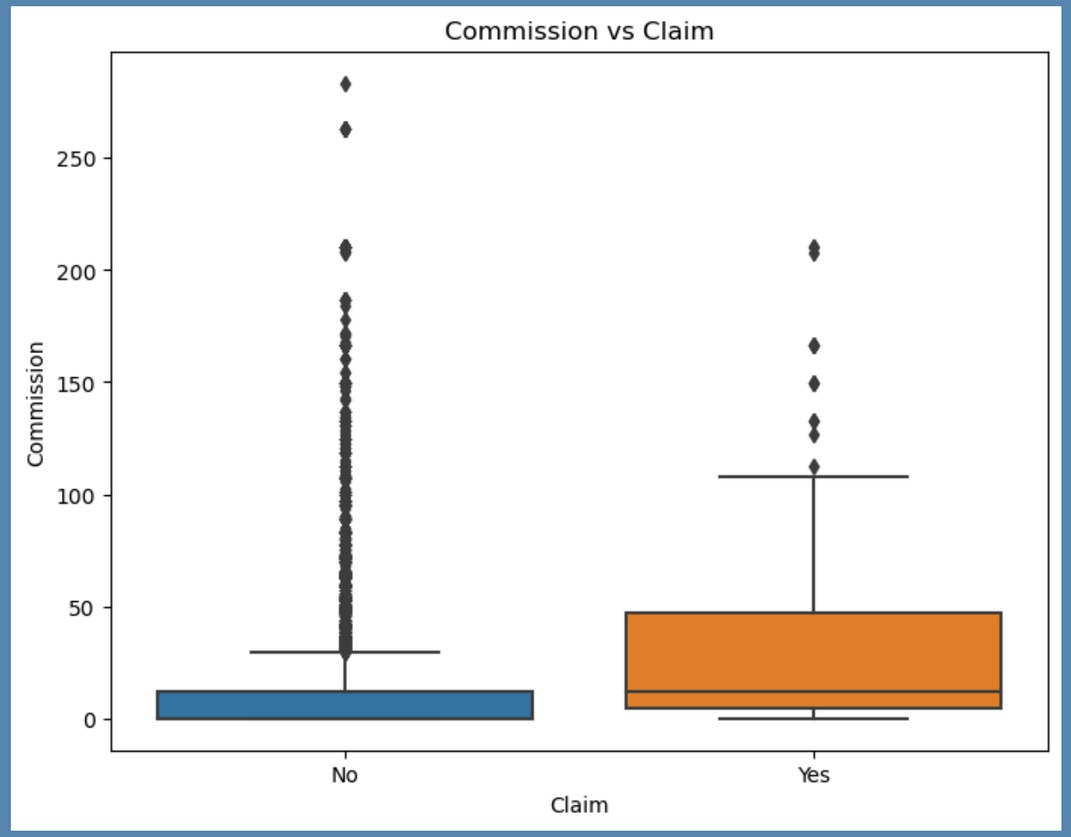

Frequency distribution for column: Agency Type
Frequency
30000
25000
20000
15000
10000
5000
0
Travel Agency
Airlines
Agency Type

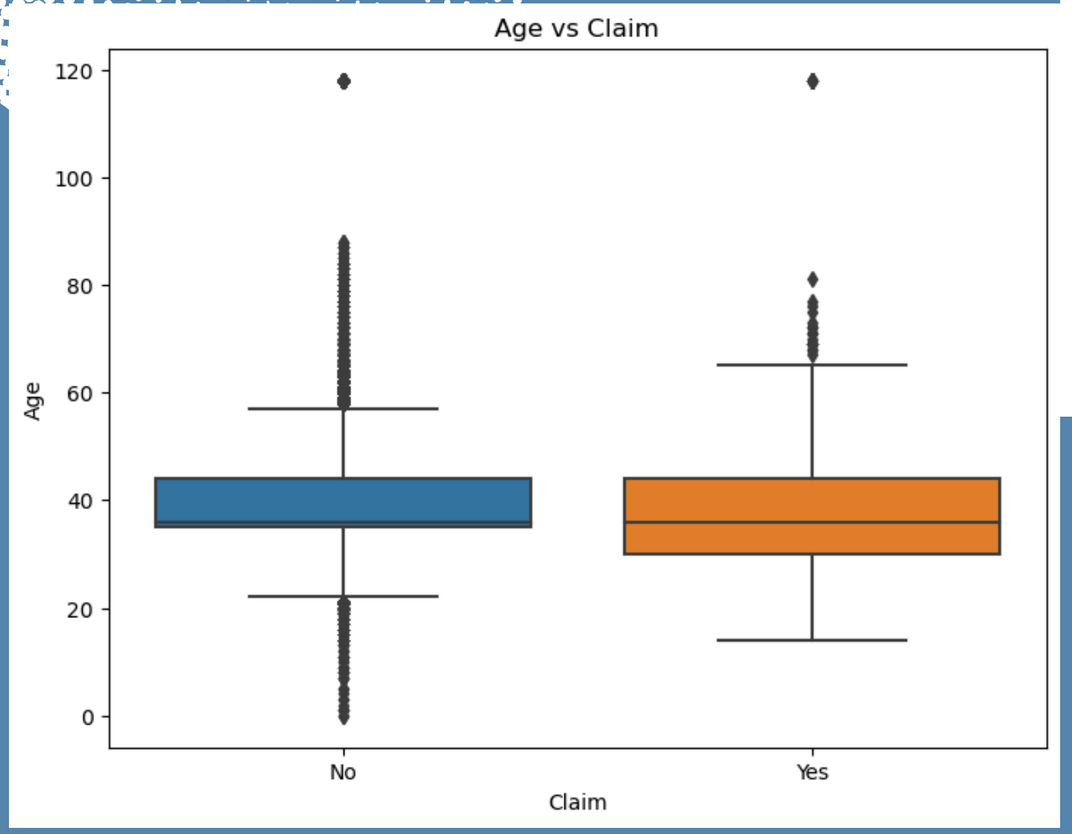

Frequency distribution for column: Distribution Channel
Frequency
40000
30000
20000
10000
0
Online
Offline
Distribution Channel

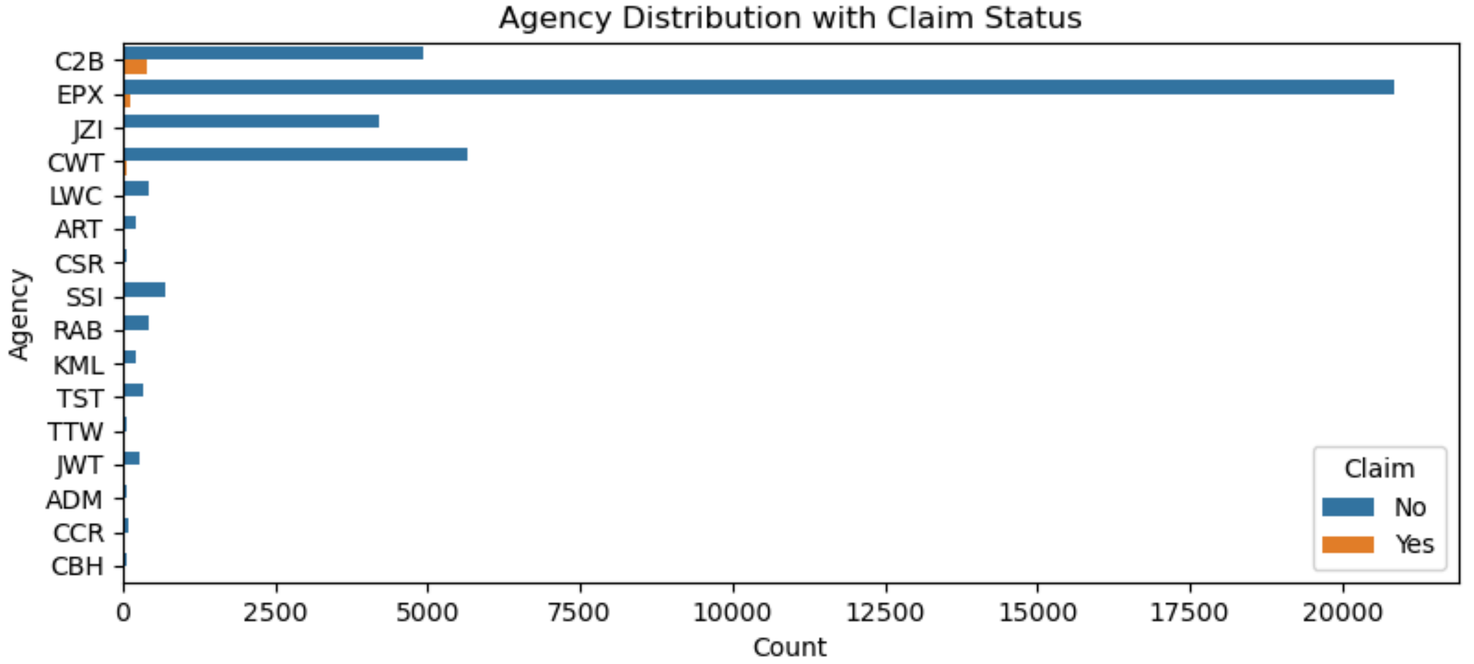

Frequency distribution for column: Product Name
Frequency
0
2000
4000
6000
8000
10000
12000
Cancellation Plan
2 way Comprehensive Plan
Basic Plan
Rental Vehicle Excess Insurance
Bronze Plan
1 way Comprehensive Plan
Value Plan
Silver Plan
Annual Silver Plan
Ticket Protector
Travel Cruise Protect
Comprehensive Plan
Gold Plan
24 Protect
Annual Gold Plan
Premier Plan
Single Trip Travel Protect Silver
Single Trip Travel Protect Gold
Annual Travel Protect Gold
Annual Travel Protect Silver
Individual Comprehensive Plan
Single Trip Travel Protect Platinum
Annual Travel Protect Platinum
Spouse or Partner Plan
Child Comprehensive Plan
Travel Cruise Protect Family
Product Name

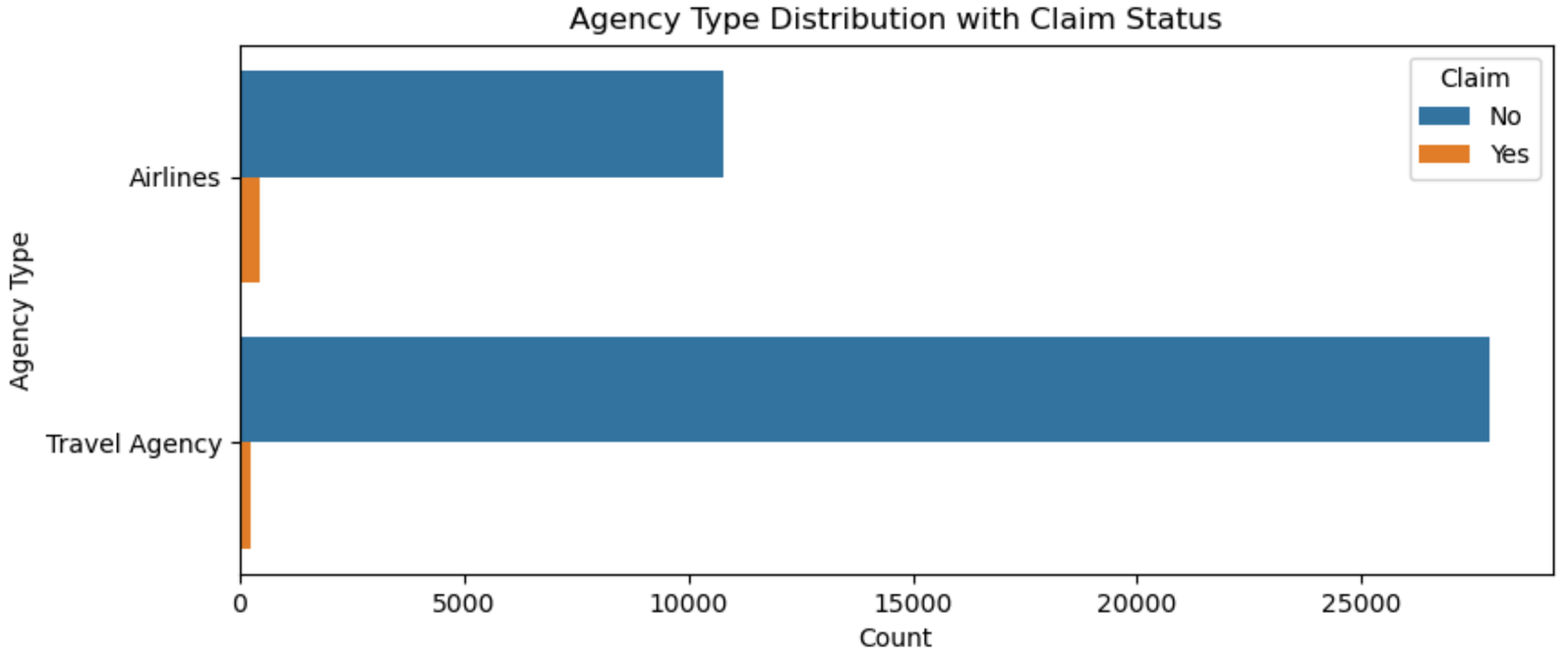

Duration vs Claim
Duration
5000
4000
3000
2000
1000
0
No
Yes
Claim

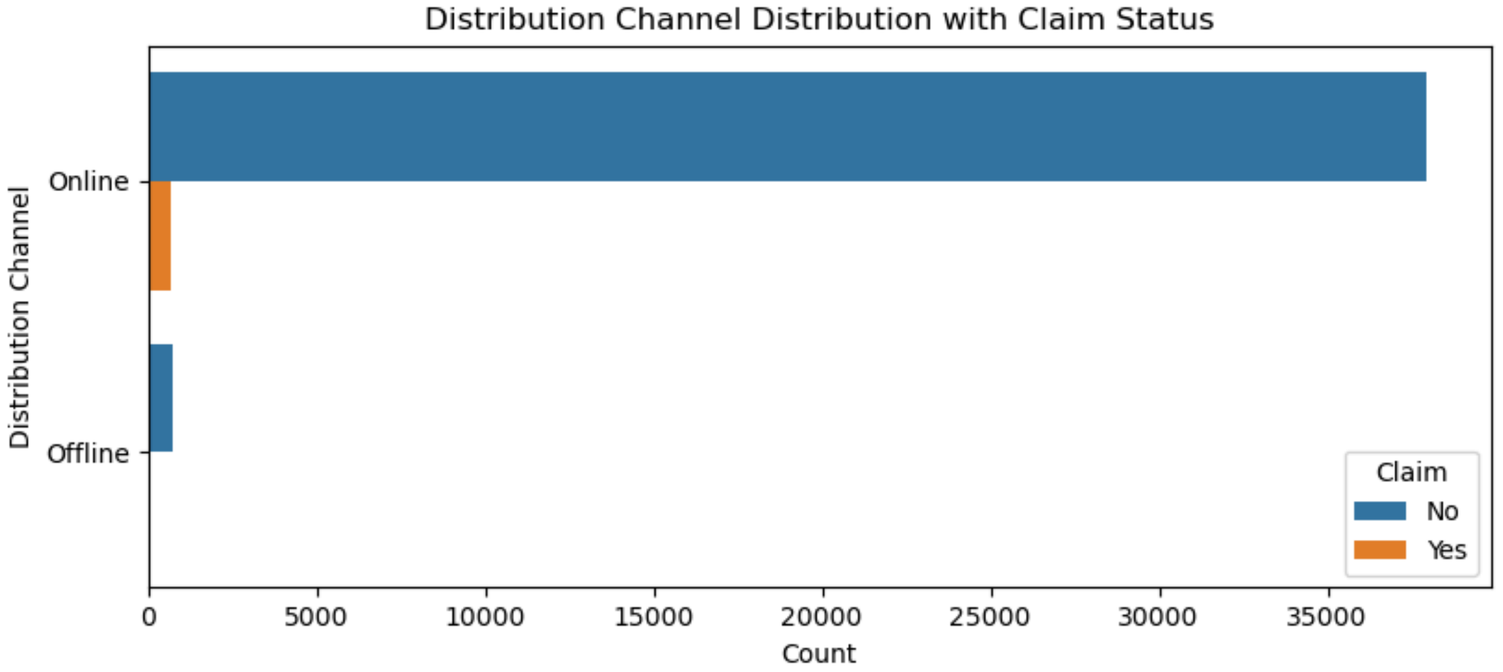

Net Sales vs Claim
Net Sales
800
600
400
200
0
−200
−400
No
Yes
Claim

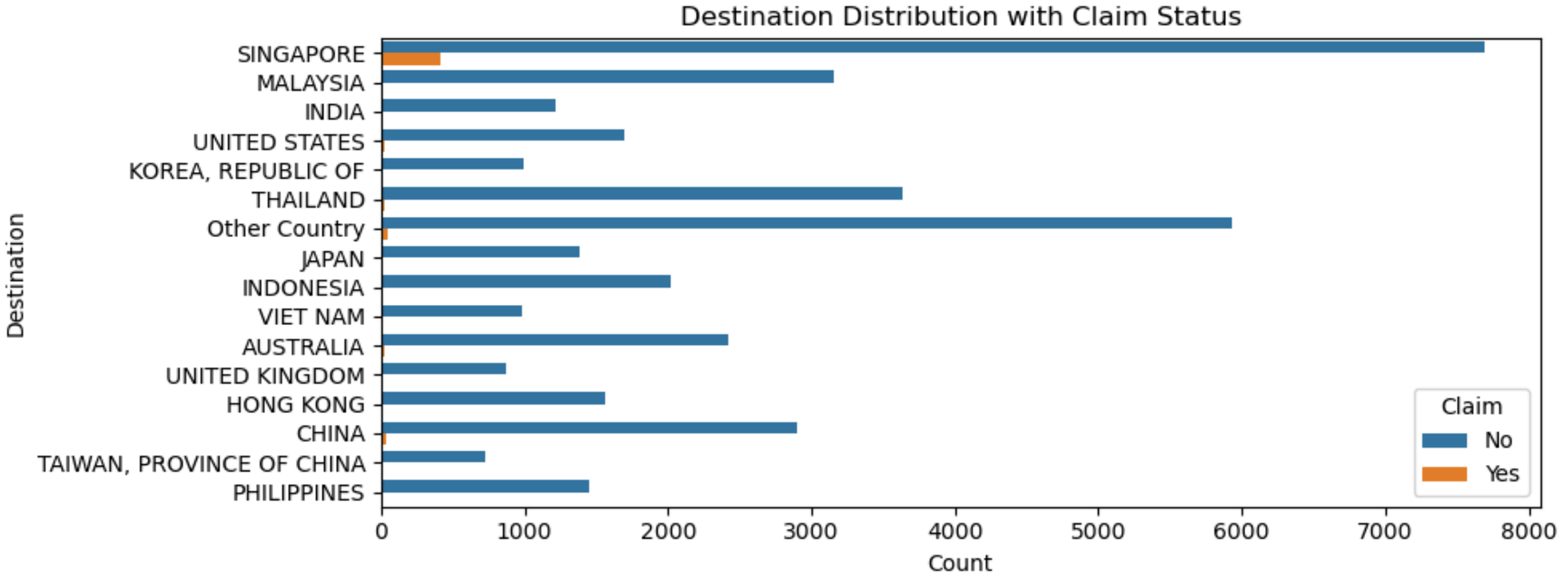

Commission vs Claim
Commission
250
200
150
100
50
0
No
Yes
Claim


Age vs Claim
Age
120
100
80
60
40
20
0
No
Yes
Claim

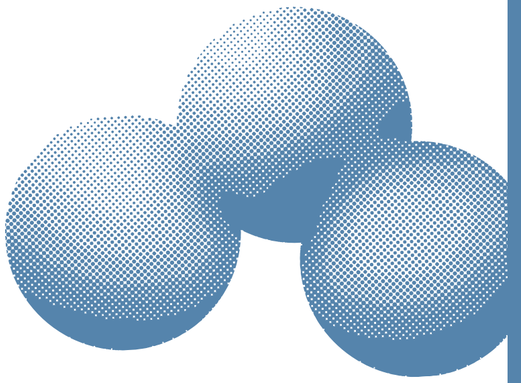

Agency Distribution with Claim Status
C2B
EPX
JZI
CWT
LWC
ART
CSR
SSI
RAB
KML
TST
TTW
JWT
ADM
CCR
CBH
Agency
0
2500
5000
7500
10000
12500
15000
17500
20000
Count
Claim
No
Yes


Agency Type Distribution with Claim Status
Airlines
Travel Agency
Agency Type
0
5000
10000
15000
20000
25000
Count
Claim
No
Yes


Distribution Channel Distribution with Claim Status
Distribution Channel
Online
Offline
Claim
No
Yes
0
5000
10000
15000
20000
25000
30000
35000
Count


Destination Distribution with Claim Status
Destination
SINGAPORE
MALAYSIA
INDIA
UNITED STATES
KOREA, REPUBLIC OF
THAILAND
Other Country
JAPAN
INDONESIA
VIET NAM
AUSTRALIA
UNITED KINGDOM
HONG KONG
CHINA
TAIWAN, PROVINCE OF CHINA
PHILIPPINES
Claim
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
8000
Count

Feature Engineering

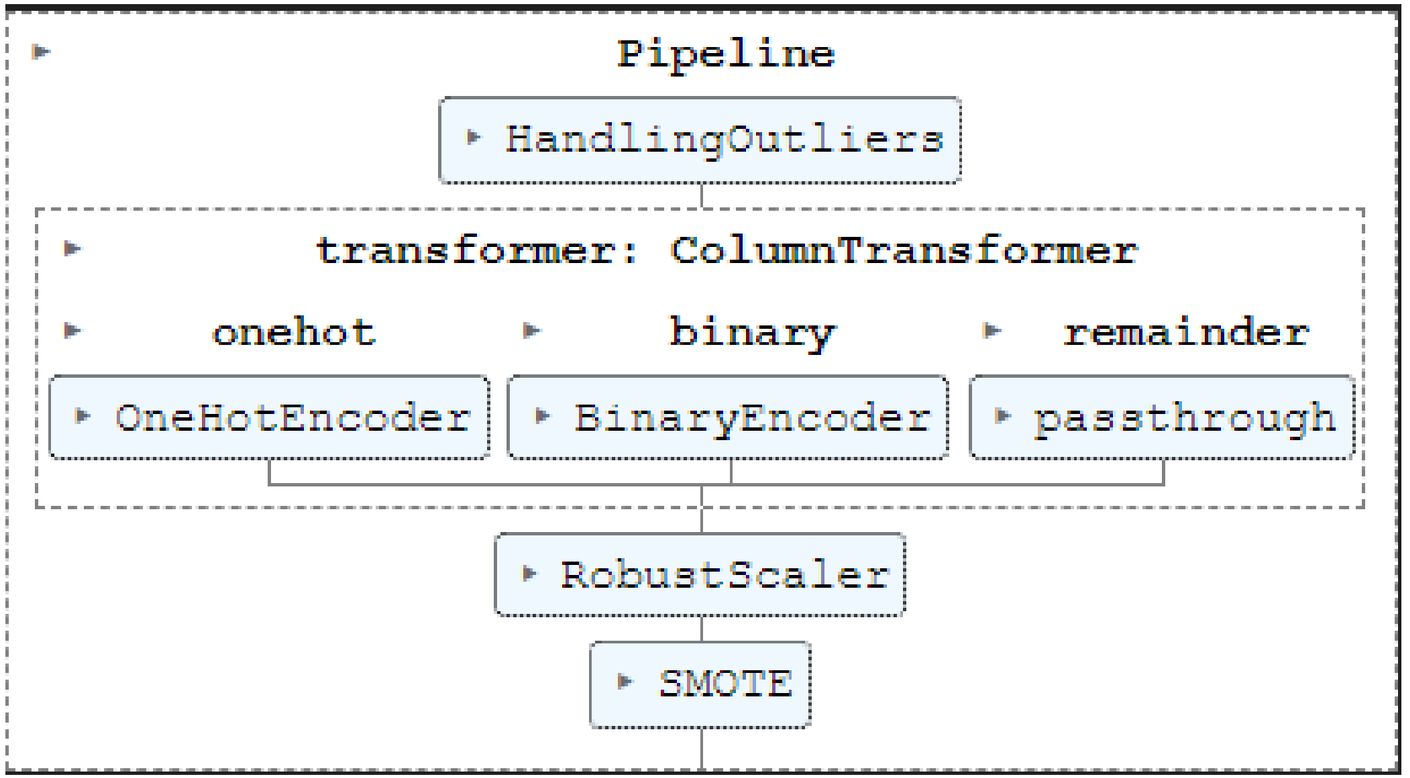


- Kolom Gender di Drop karena Missing Value nya 71%
- Data Duplicate di drop sebanyak 5000 Data
- Negative Value Duration di drop
- Destination menggunakan 15 data terbanyak, sisanya dikelompokkan
- Cardinality pada Agency, Destination, Product Name
- Imbalance Data pada tingkat mild pada target

Data Preprocessing

- Encoding pada feature kategorikal dengan One Hot Encoding dan Binary Encoding
- Membuat Feature dan Target
- Handling Outlier dengan Windsorizing
- Scaling dengan Robust Scaler
- Resampling Menggunakan SMOTE
- Dimasukkan kedalam Pipeline

Data Preprocessing



Modelling

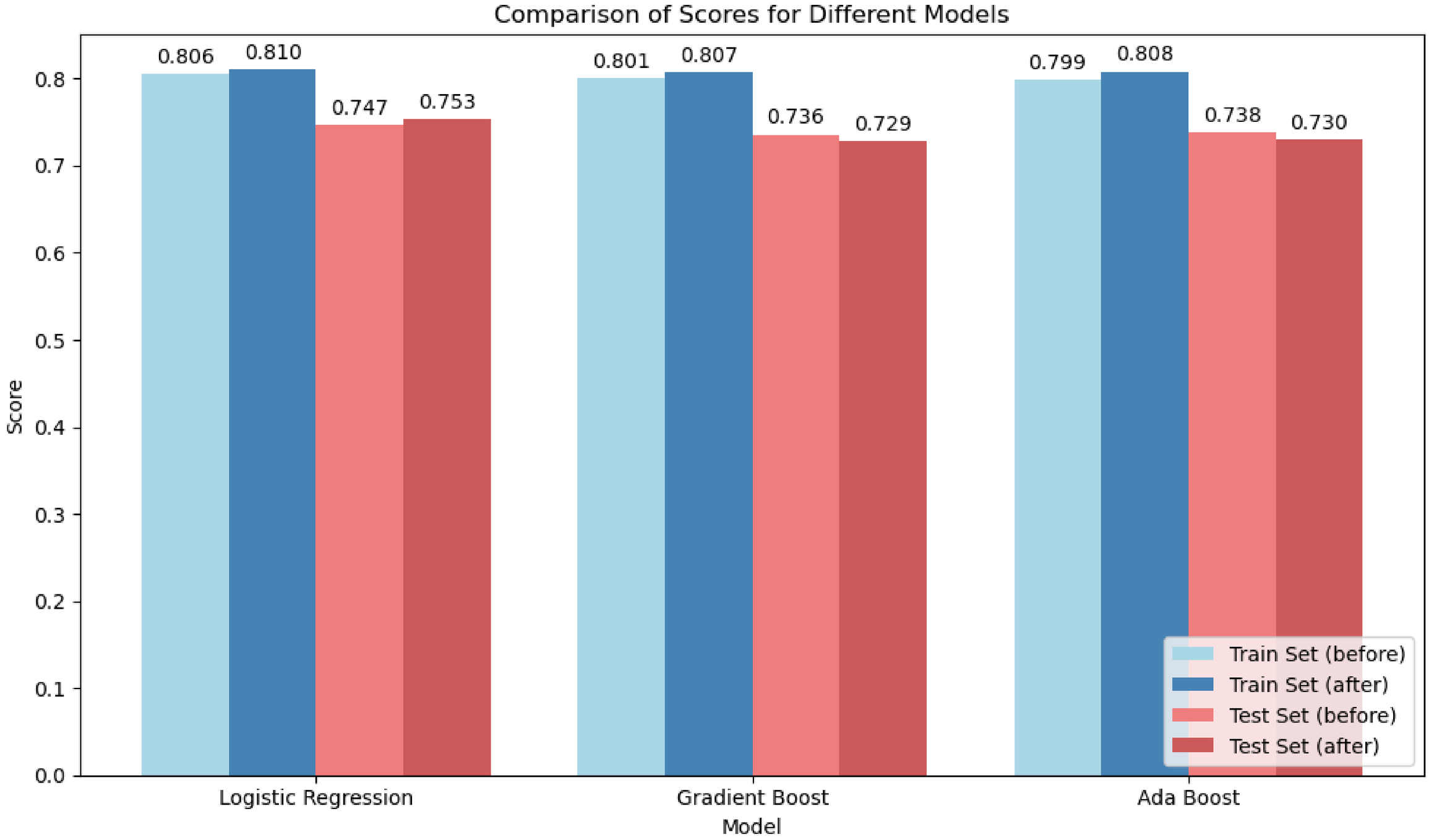
Logistic Regression, KNN Classifier, Decision Tree Classifier, Random Forest Classifier, Adaboost Classifier, Gradient Boost Classifier, dan XGBoost Classifier

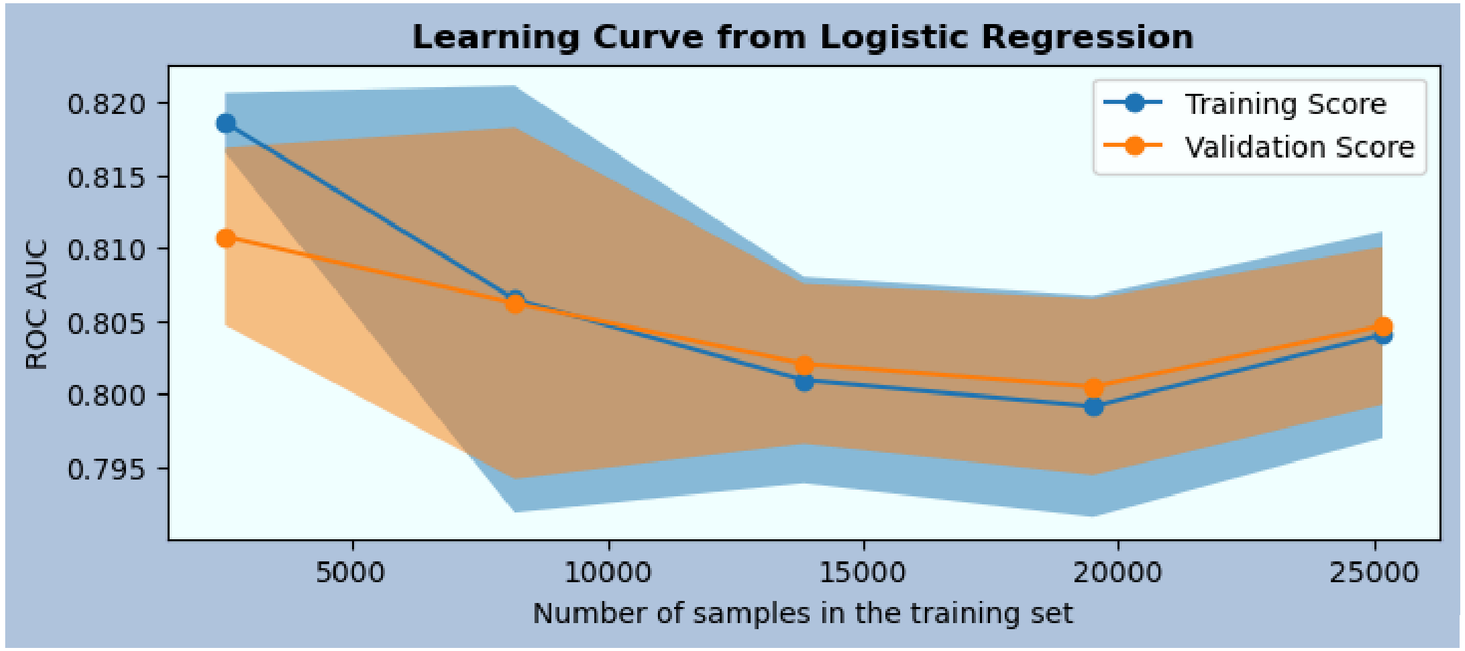
Cross Validation

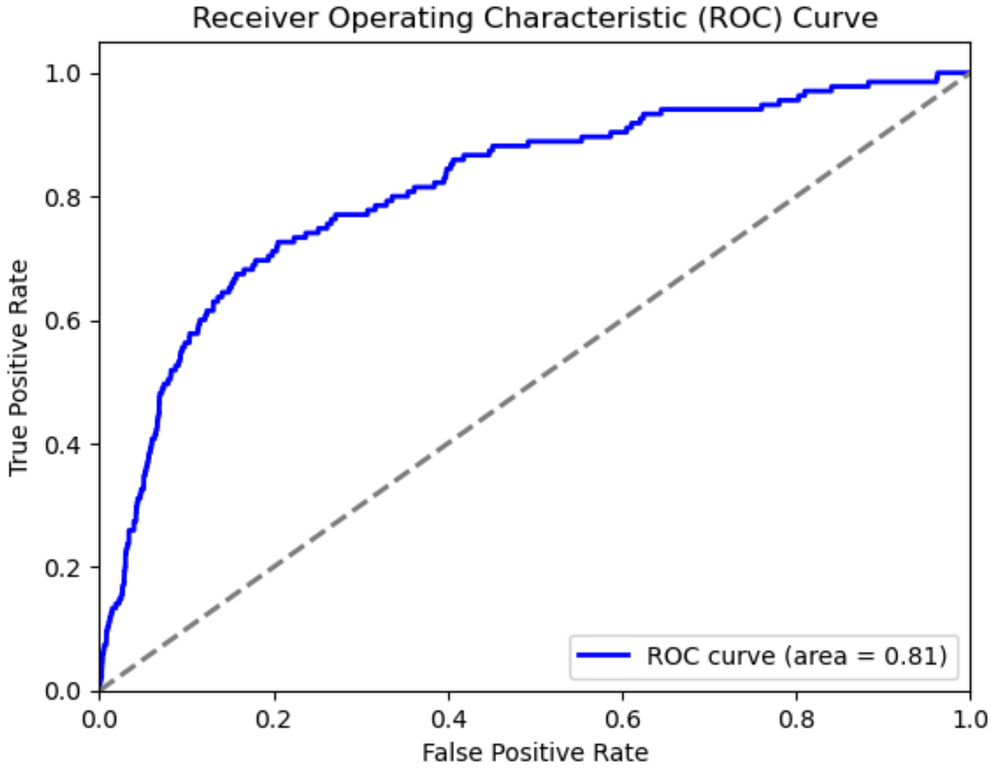
```
for algoritma in models:

        model_pipe = Pipeline([
            ('outlier', HandlingOutliers()),
            ('preprocessing', transformer),
            ('scaler', scaler),
            ('resampler', smote),
            ('model', algoritma)
        ])

        skfold = StratifiedKFold(n_splits = 5)

        model_cv = cross_val_score(
            model_pipe,                 # model yang sudah dimasukkan ke dalam pipeline
            X_train,                    # data sebelum di preprocessing
            y_train,
            cv = skfold,
            scoring = 'roc_auc',
            error_score='raise'
            )
```

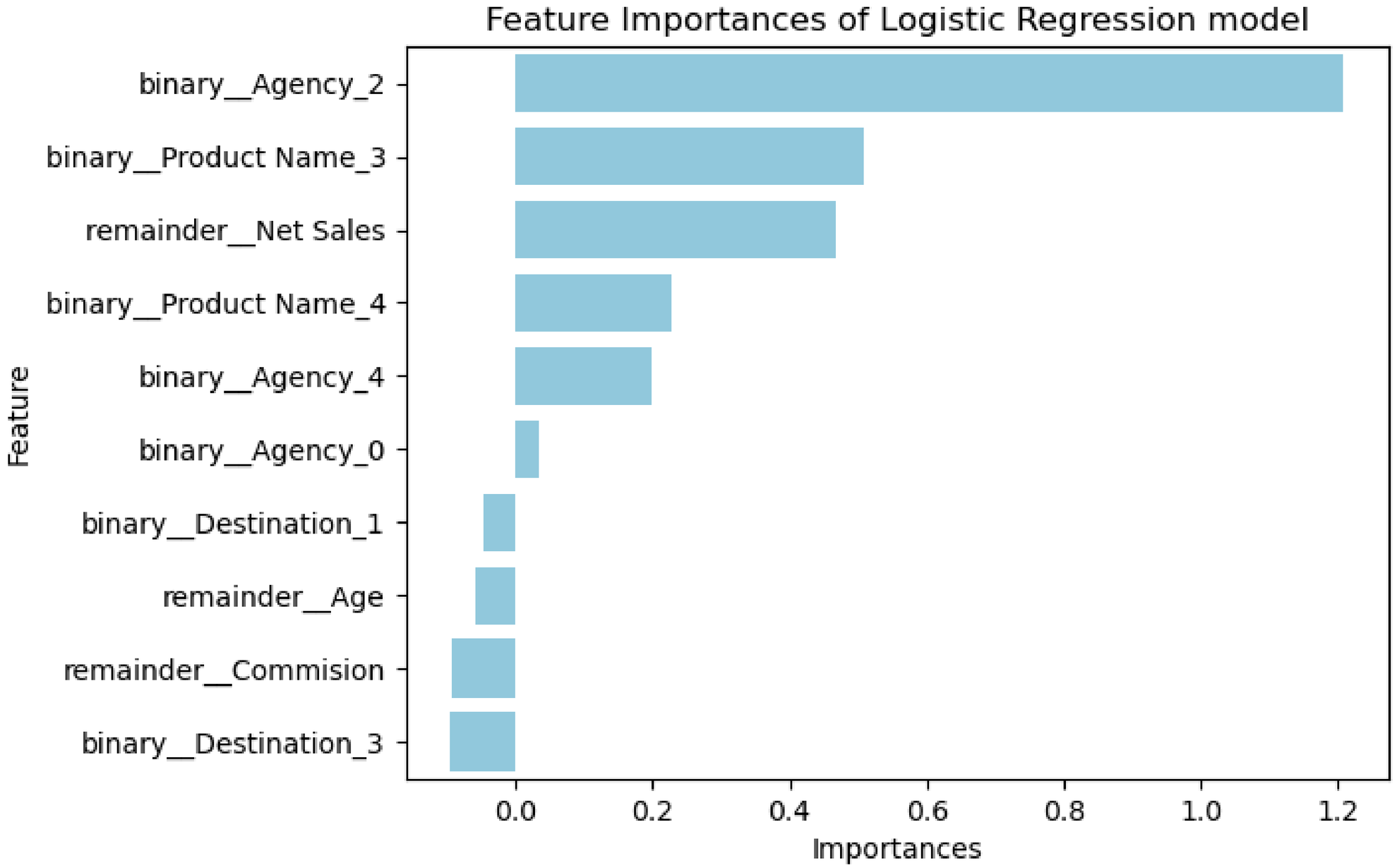
	model	mean	std	all_score
0	LogisticRegression	0.806732	0.023814	[0.8074, 0.8196, 0.7732, 0.8426, 0.7907]
5	GradientBoostingClassifier	0.801139	0.013131	[0.7953, 0.8109, 0.7856, 0.8215, 0.7924]
4	AdaBoostClassifier	0.799833	0.025325	[0.7995, 0.8211, 0.7607, 0.8319, 0.7859]
6	XGBClassifier	0.759585	0.011800	[0.755, 0.763, 0.7631, 0.7764, 0.7404]
3	RandomForestClassifier	0.735154	0.020823	[0.7663, 0.7296, 0.7474, 0.704, 0.7284]
1	KNeighborsClassifier	0.683993	0.011265	[0.7029, 0.6825, 0.6695, 0.6881, 0.677]
2	DecisionTreeClassifier	0.546512	0.006774	[0.5496, 0.5412, 0.5581, 0.5446, 0.5391]

	model	accuracy (test_set)
0	LogisticRegression	0.747032
4	AdaBoostClassifier	0.738925
5	GradientBoostingClassifier	0.736076
6	XGBClassifier	0.618830
1	KNeighborsClassifier	0.595237
3	RandomForestClassifier	0.579673
2	DecisionTreeClassifier	0.527045


Comparison of Scores for Different Models
0.806
0.810
0.747
0.753
0.801
0.807
0.736
0.729
0.799
0.808
0.738
0.730
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Score
Logistic Regression
Gradient Boost
Ada Boost
Model
Train Set (before)
Train Set (after)
Test Set (before)
Test Set (after)

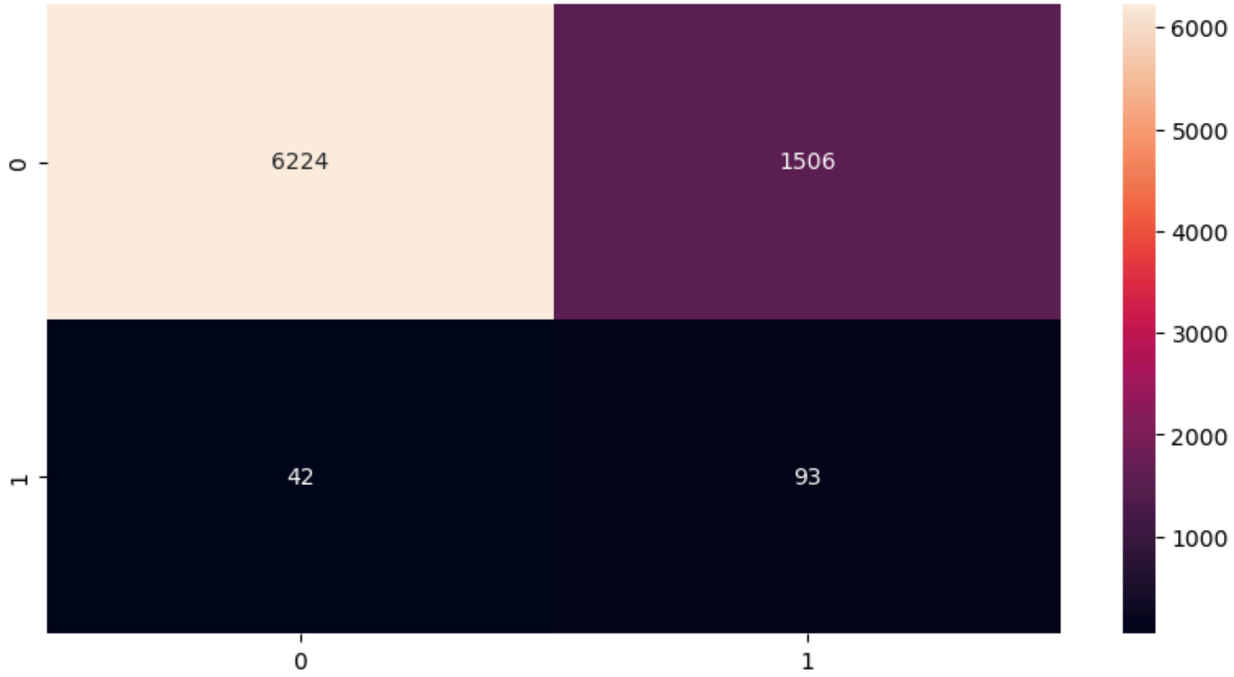

Learning Curve from Logistic Regression
Training Score
Validation Score
ROC AUC
0.820
0.815
0.810
0.805
0.800
0.795
5000
10000
15000
20000
25000
Number of samples in the training set


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.81)

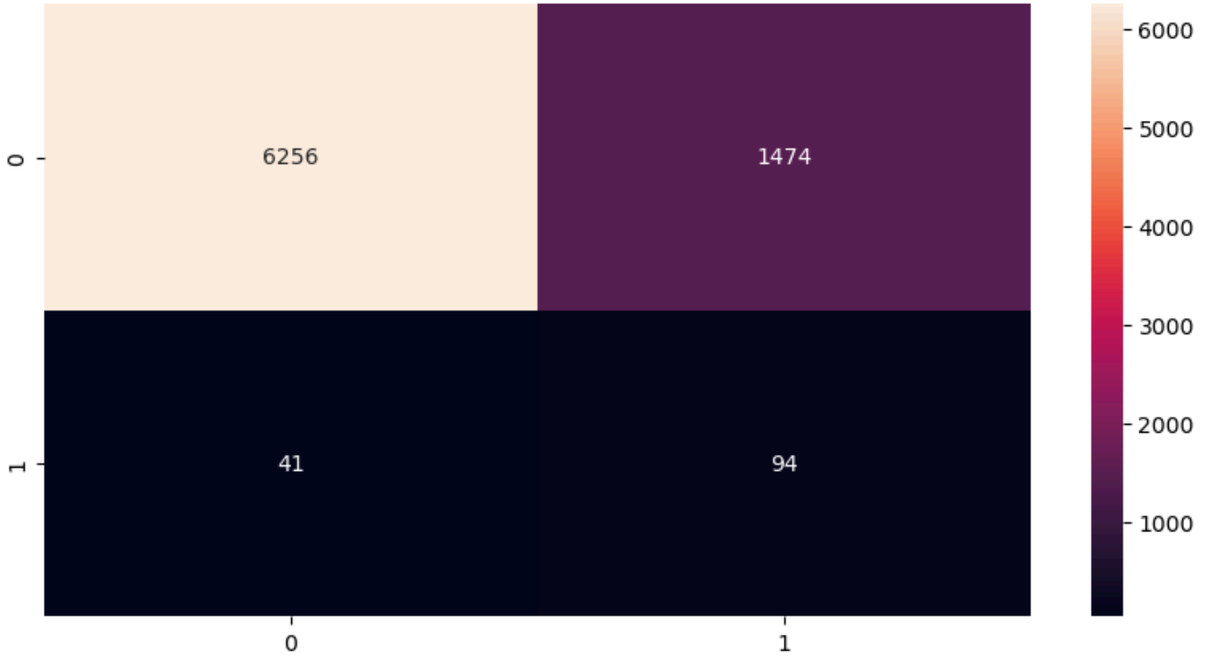

Feature Importances of Logistic Regression model
binary__Agency_2
binary__Product Name_3
remainder__Net Sales
binary__Product Name_4
binary__Agency_4
binary__Agency_0
binary__Destination_1
remainder__Age
remainder__Commision
binary__Destination_3
Feature
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Importances

Confusion Matrix

Sebelum



Sesudah



Cost Benefit Analysis

Faktor	Tanpa ML	Dengan ML
Biaya Awal	Rendah (terutama gaji)	Tinggi (infrastruktur, gaji, pelatihan)
Biaya Berkelanjutan	Tinggi (gaji, biaya kesalahan, pelatihan)	Sedang (pemeliharaan, pembaruan)
Akurasi	Lebih rendah (kesalahan manusia)	Lebih tinggi (presisi berbasis data)
Kecepatan Pemrosesan	Lebih lambat (proses manual)	Lebih cepat (proses otomatis)
Skalabilitas	Terbatas oleh sumber daya manusia	Skalabilitas tinggi
Kepuasan Pelanggan	Potensial lebih rendah (waktu pemrosesan lama)	Lebih tinggi (klaim cepat dan akurat)
Deteksi Penipuan	Kurang efektif	Lebih efektif

Conclusion

1. **Logistic Regression adalah model terbaik** untuk dataset ini karena performa yang cukup stabil dan baik pada dataset yang besar.

2. Hasil ROC AUC
   - **Train score sebelum tuning**: 75.66%
   -**Test score sebelum tuning**: 74.70%
   - **Train score setelah tuning**: 75.91%
   - **Test score setelah tuning**: 75.28%

3. Setelah tuning, model menunjukkan **peningkatan dalam recall, precision, dan accuracy.** Ini menunjukkan bahwa ***tuning membantu model menjadi lebih baik*** dalam mengidentifikasi klaim yang benar sambil mengurangi kesalahan dalam prediksi klaim.

4. Berdasarkan **feature importance** didapati bahwa seseorang akan klaim atau tidak klaim dapat dipengaruhi dimana pemegang polis tersebut membeli asuransi perjalanan dari suatu **agen asuransi.**

Recommendation Model

1. penggunaan hyperparameter tuning yang lebih luas akan membantu
2. Feature Engineering yang lebih dalam
3. Penggunaan metriks lain dengan penyesuaian problem statement
4. Mencoba menggunakan resampler lain seperti RandomOverSampling, SMOTENC, SMOTE-ENN

Recommendation Business

1. Akurasi yang Lebih Tinggi
2. Kecepatan Pemrosesan yang Lebih Tinggi
3. Penghematan Biaya
4. Keputusan yang Lebih Tepat
5. Peningkatan Kepuasan Pelanggan

Terima Kasih