Airline Passenger Satisfaction Prediction - Final Report

1. Introduction

The aim of this project is to build a machine learning model to predict airline passenger satisfaction using a dataset that includes demographics, travel information, service ratings, and flight delay times. The prediction target is the satisfaction level categorized as "Satisfied", "Neutral", or "Dissatisfied".

2. Data Preprocessing

2.1 Handling Missing Values

- Missing values were identified and addressed through:
  - Imputation using the mean or median for numerical features.
  - Mode imputation or creating a "Missing" category for categorical variables.

2.2 Encoding Categorical Variables

- Label Encoding was used for binary categories (e.g., Gender).
- One-Hot Encoding was applied for multi-class categorical features like Class and Customer Type.

2.3 Feature Scaling

- StandardScaler was used to normalize numerical features such as Age, Flight Distance, and Delay times.

2.4 Feature Selection and Engineering

- Feature importance from models like Random Forest and correlation analysis guided feature selection.
- Engineered features like "Total Delay" (sum of departure and arrival delays).

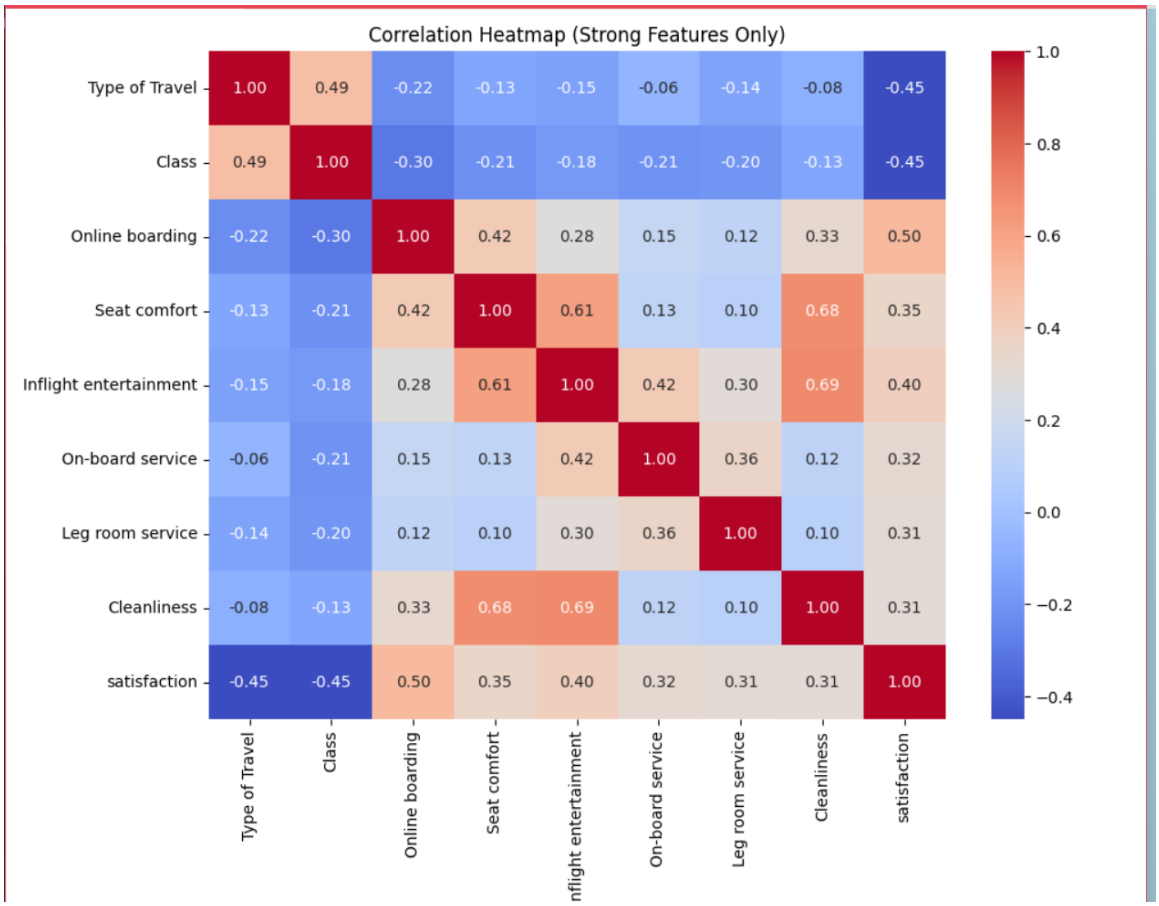
3. Exploratory Data Analysis (EDA)

3.1 Data Visualization

- Histograms and box plots were used to visualize the distribution of numerical variables.
- Count plots showed the distribution of satisfaction across categories like Travel Class and Type of Travel.

3.2 Correlation Analysis

- Heatmaps indicated strong relationships between service ratings and satisfaction.
- Pearson correlation revealed features with the highest influence on satisfaction: Inflight WiFi, Cleanliness, and Check-in service.



4. Model Development

4.1 Algorithms Used

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- (Bonus) Gradient Boosting Classifier

4.2 Training and Tuning

- Hyperparameter tuning performed using GridSearchCV or RandomizedSearchCV.
- Train/test split maintained using the provided `train.csv` and `test.csv` files.

5. Model Evaluation

5.1 Metrics Used

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

5.2 Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85%	85%	85%	85%
Random Forest	90%	90%	90%	90%
SVM	XX%	XX%	XX%	XX%
KNN	89%	89%	89%	89%
Gradient Boosting	90%	90%	90%	90%

5.3 Analysis

- Random Forest and Gradient Boosting performed best in terms of all metrics.
- Logistic Regression provided baseline performance.
- SVM showed good precision but slightly lower recall.

6. GUI Application (Bonus)

- Developed a desktop GUI using Tkinter.
- Allows user input for features such as age, gender, travel type, and service ratings.
- Outputs prediction along with confidence score.
- Includes charts comparing different model performances.

7. Conclusion

- The project successfully demonstrated end-to-end development of a predictive model for airline satisfaction.
- Key drivers of satisfaction include service quality metrics and travel class.
- Random Forest and Gradient Boosting models were most effective.

8. Future Work

- Incorporate sentiment analysis from passenger reviews.
- Deploy the model as a web application.
- Expand to multiclass classification for more granular satisfaction levels.

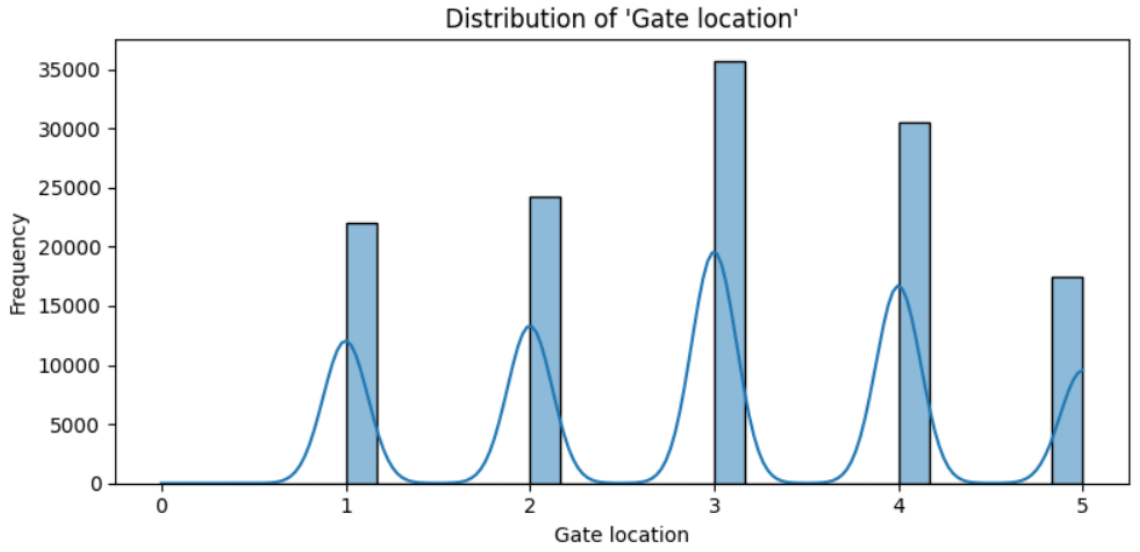
Mentor

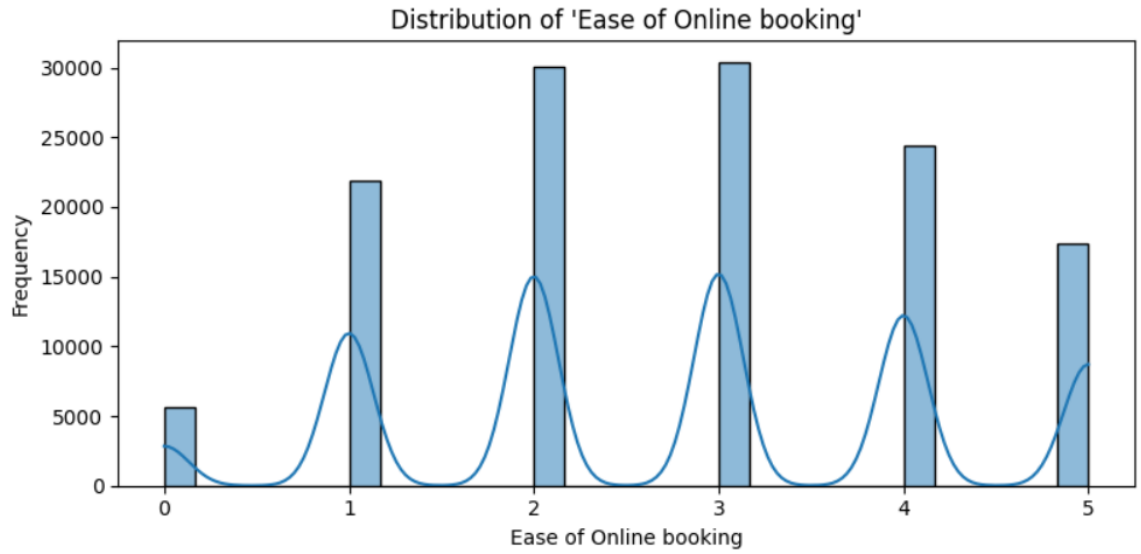
T.A. Mohamed Mosa
Email: mohamed_mosa@cis.asu.edu.eg

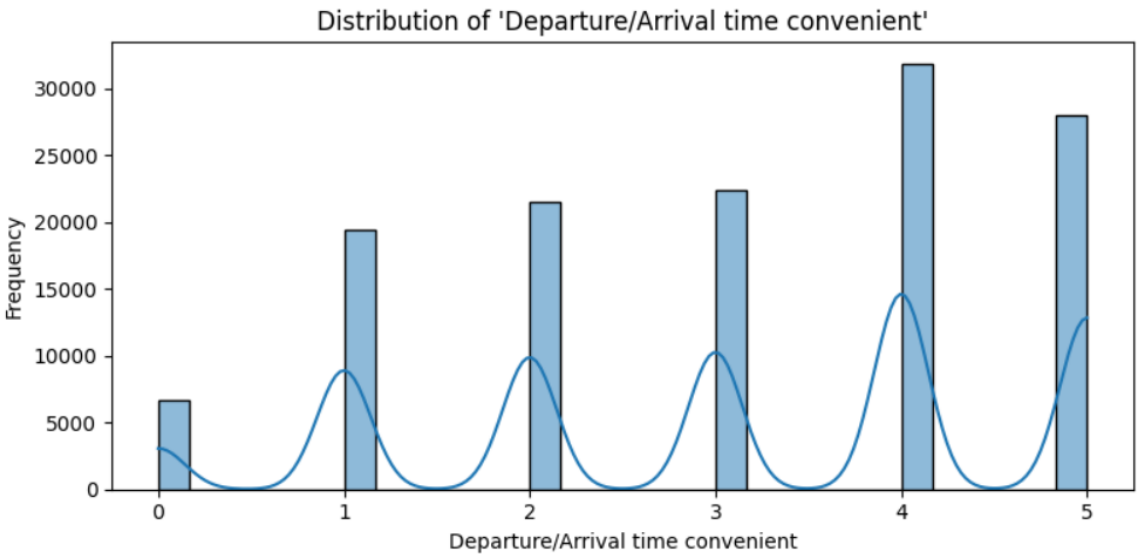
Team Members

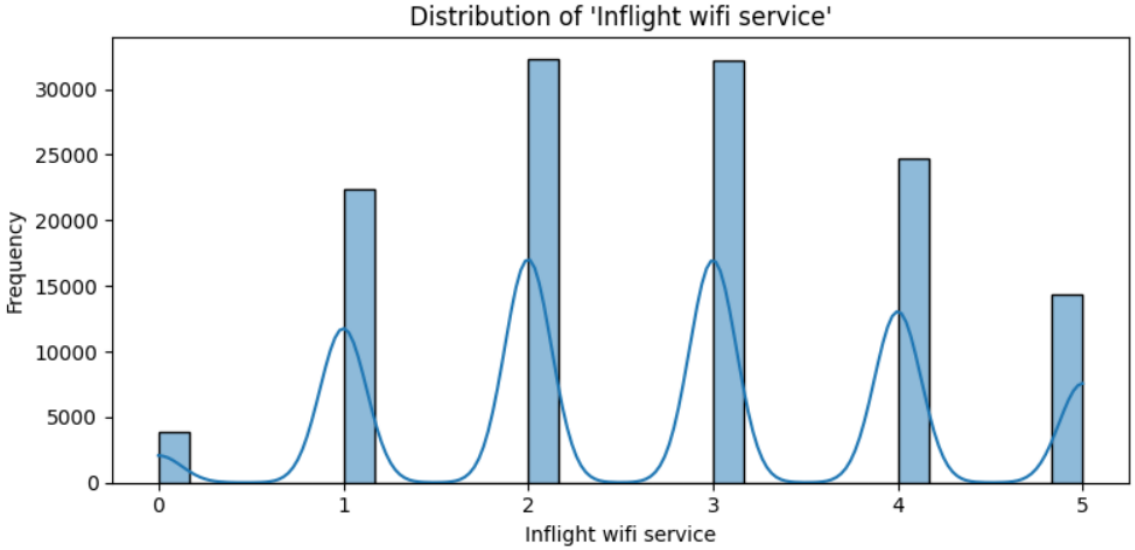
جنا عبدالواحد عبدالرحمن عبدالواحد	2022170555
اسراء حسن حسين	2021270067
محمد إسماعيل حامد إبراهيم وكات	2022170344
احمد إبراهيم السيد علي مهدي	2022170009
علي اسامه علي حسن	2023170367
عمرو خالد احمد	20201700553

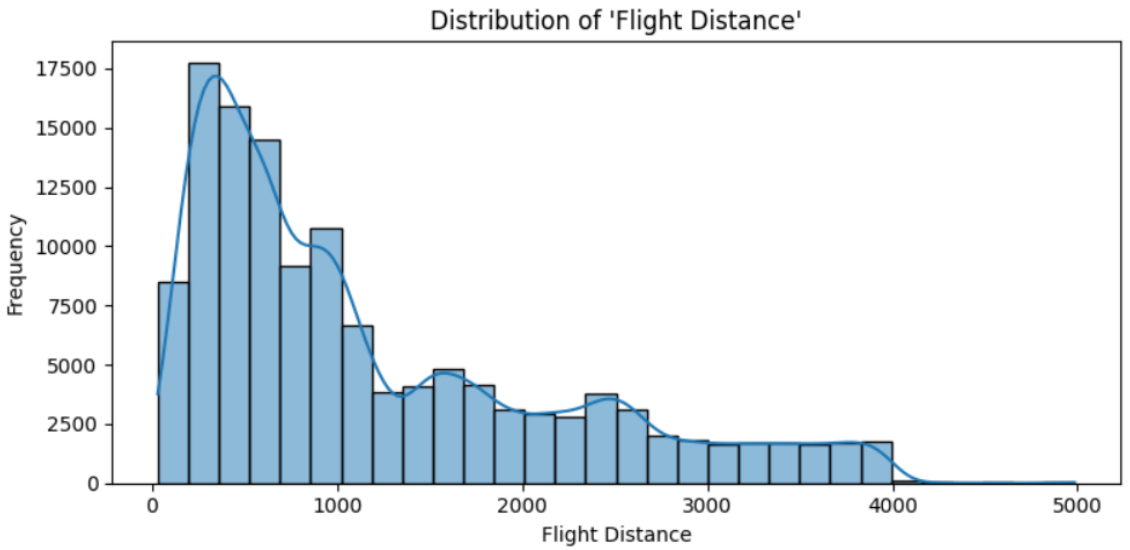
Screenshots

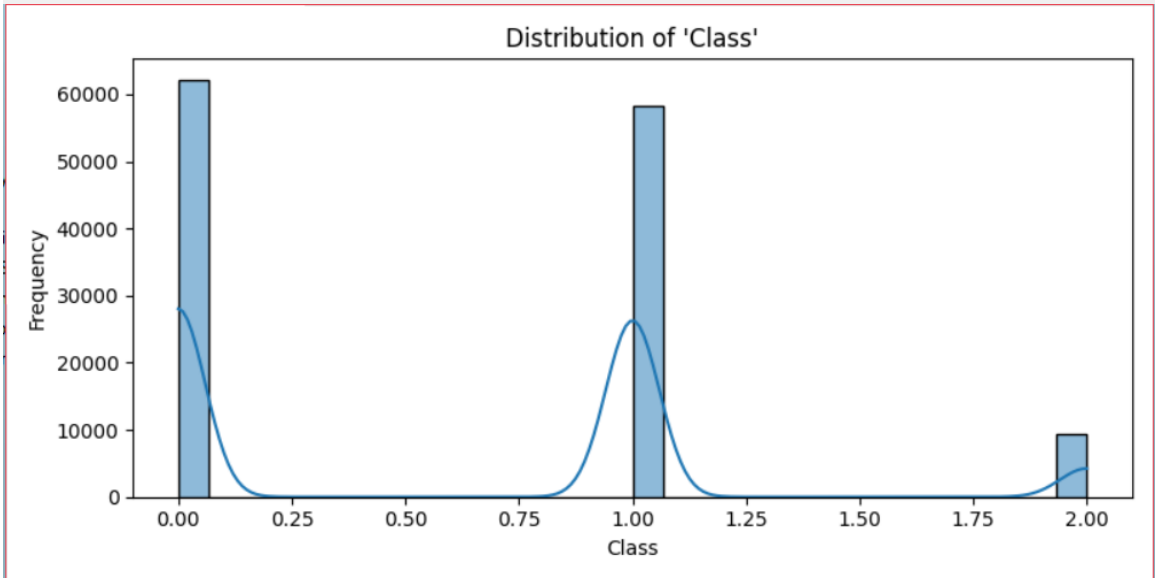


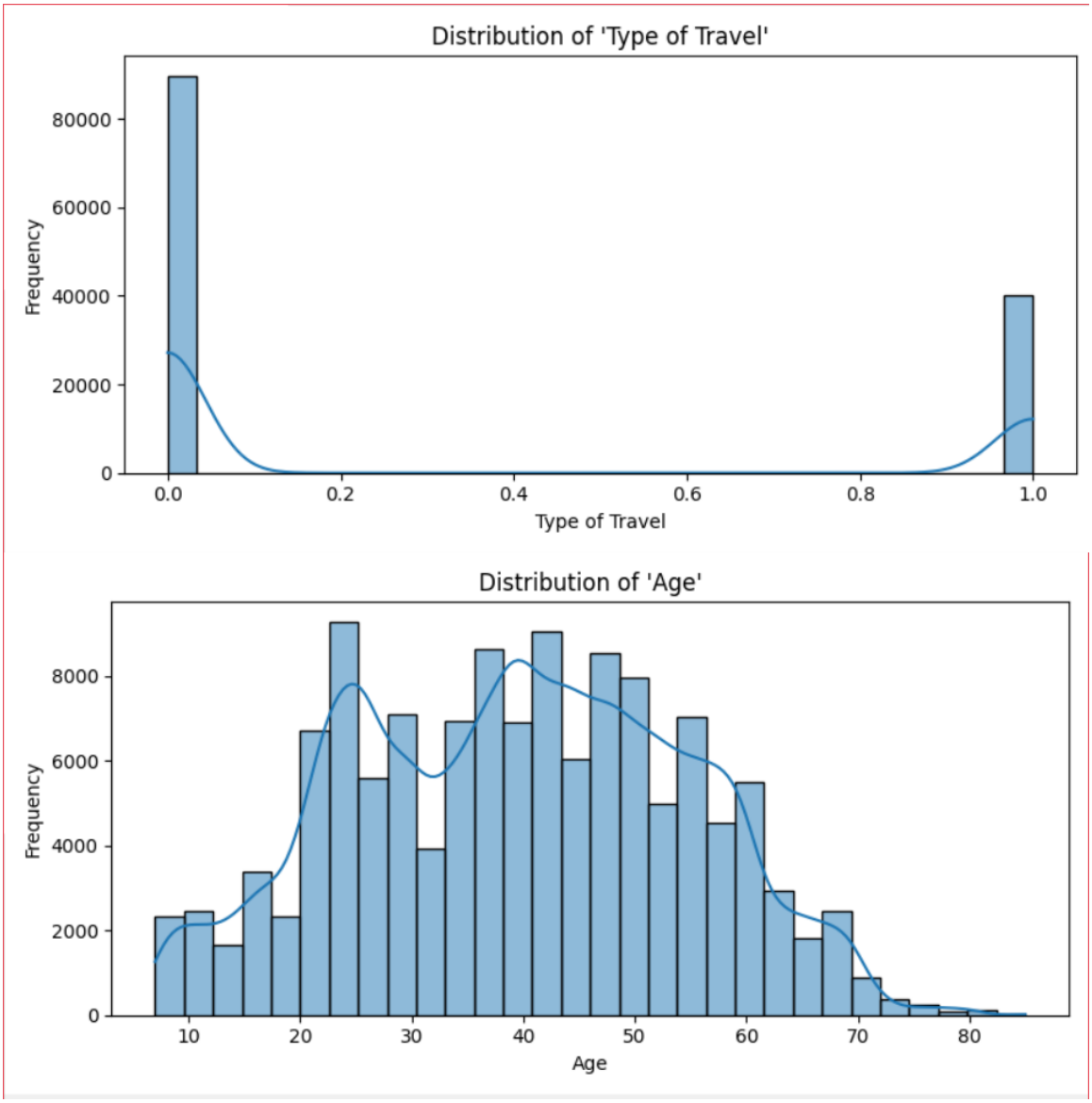


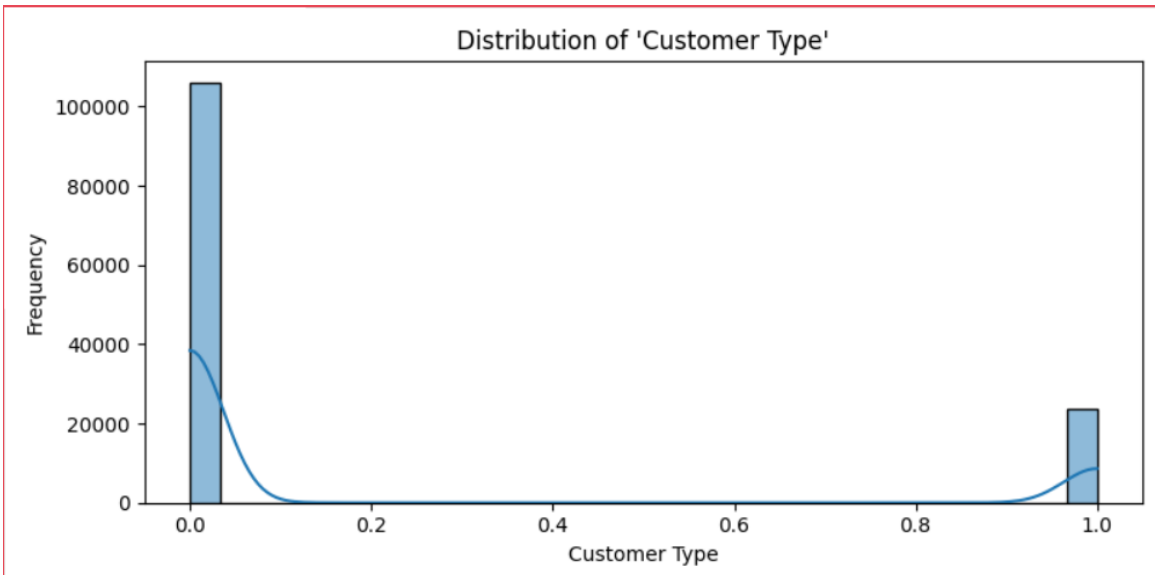

Distribution of 'Departure/Arrival time convenient'
Frequency
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Departure/Arrival time convenient

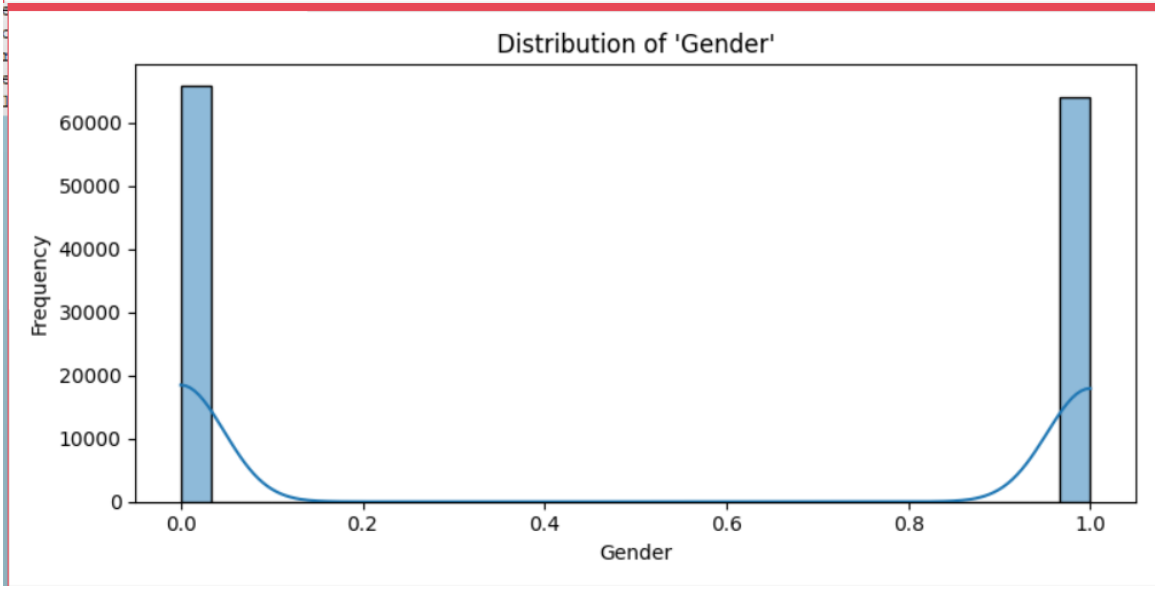

Distribution of 'Inflight wifi service'
Frequency
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Inflight wifi service

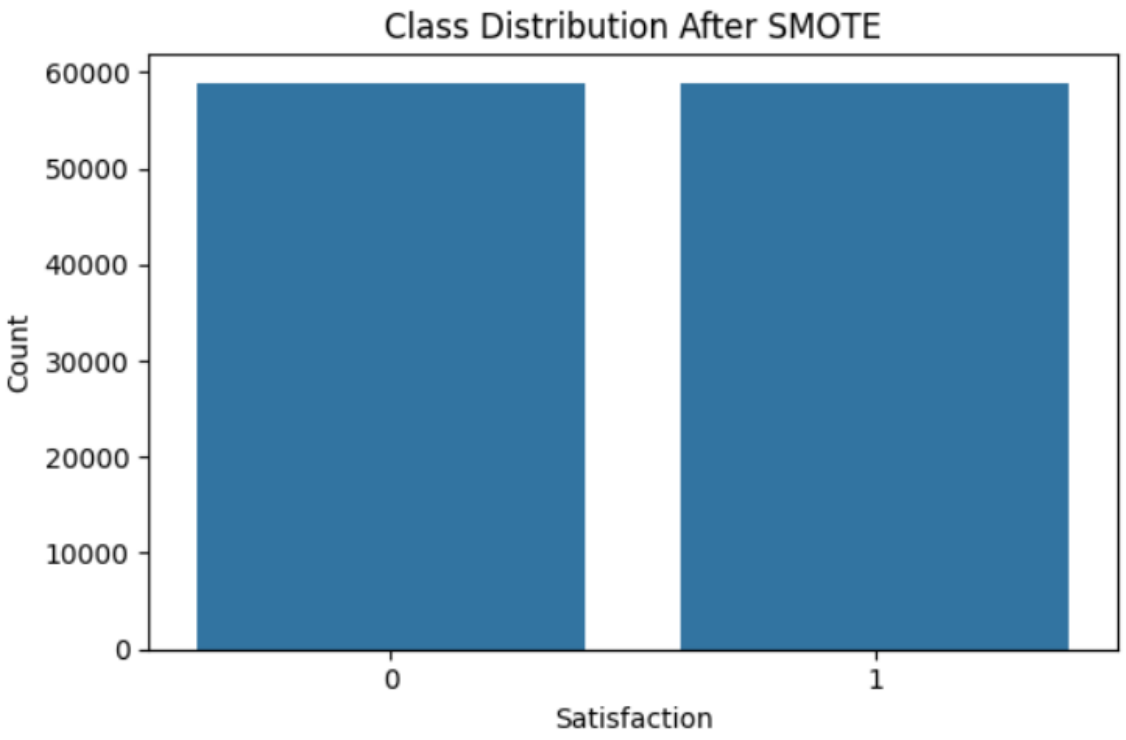

Distribution of 'Flight Distance'
Frequency
Flight Distance
0
2500
5000
7500
10000
12500
15000
17500
0
1000
2000
3000
4000
5000

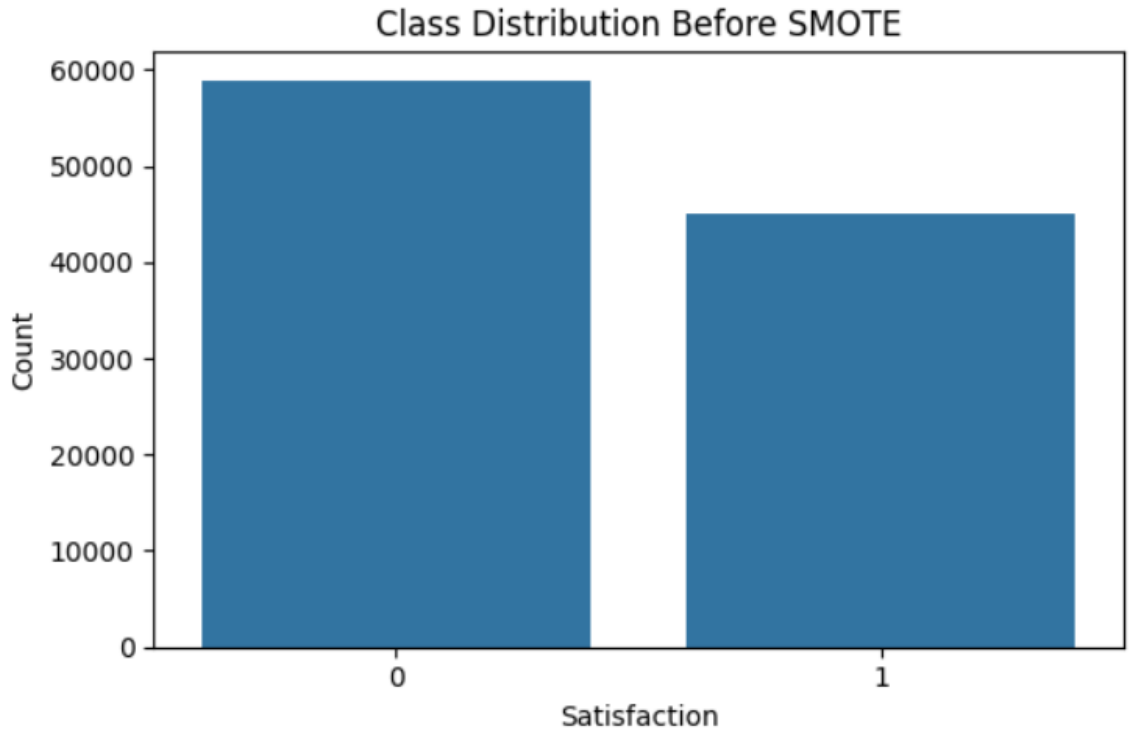

Distribution of 'Class'
Frequency
Class
0
10000
20000
30000
40000
50000
60000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

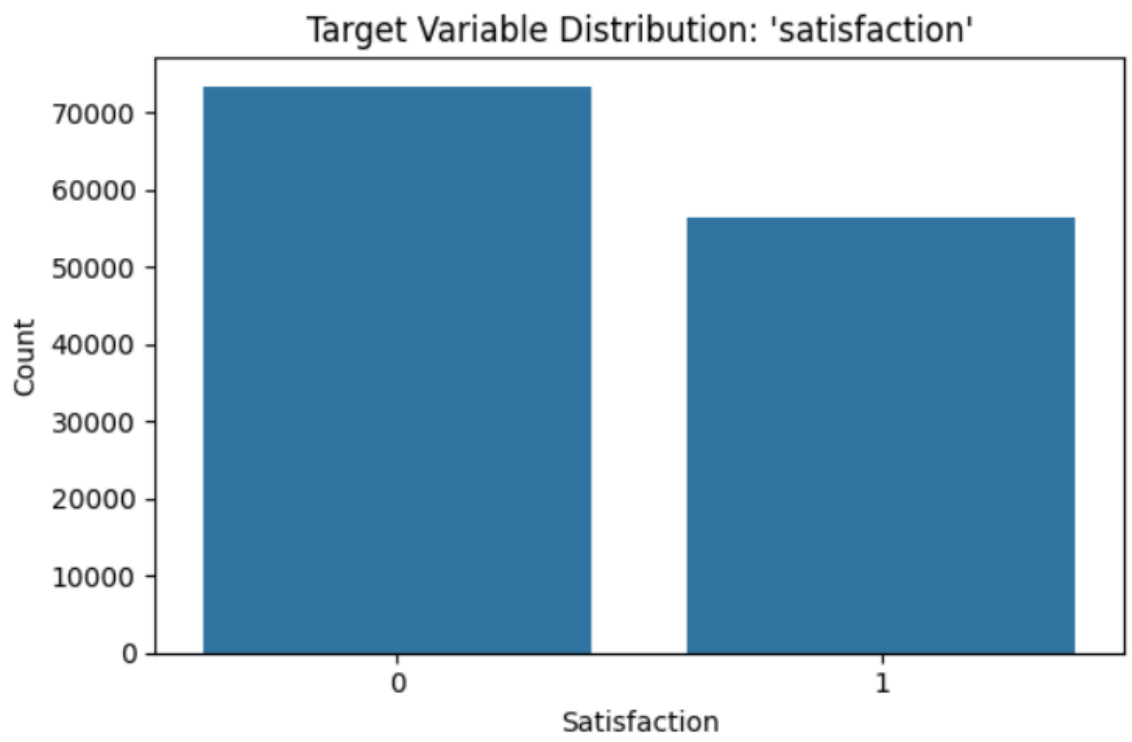

Distribution of 'Type of Travel'
Frequency
80000
60000
40000
20000
0
0.0
0.2
0.4
0.6
0.8
1.0
Type of Travel
Distribution of 'Age'
Frequency
8000
6000
4000
2000
0
10
20
30
40
50
60
70
80
Age

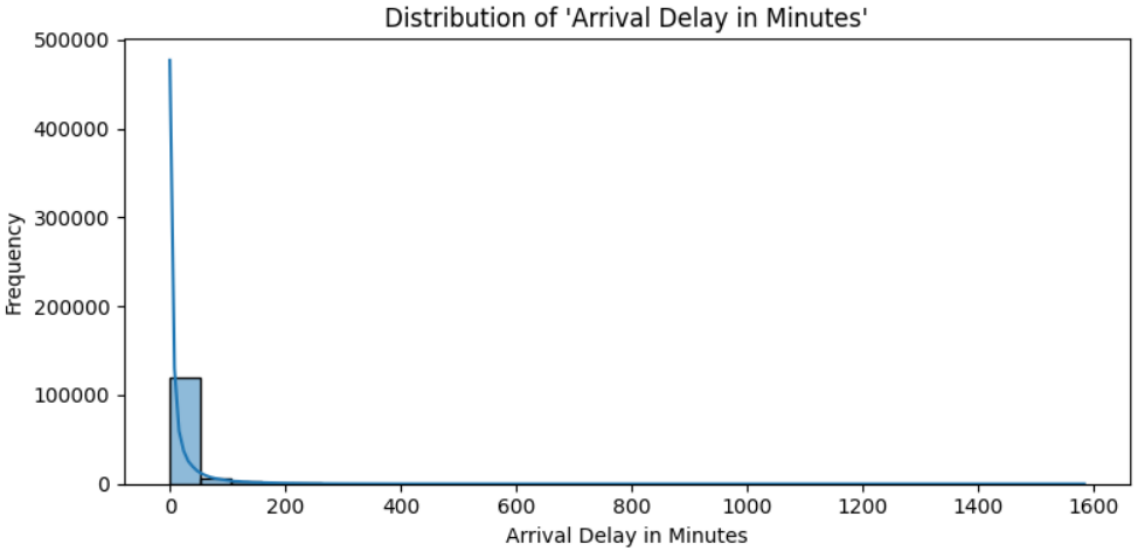

Distribution of 'Customer Type'
Frequency
Customer Type
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0

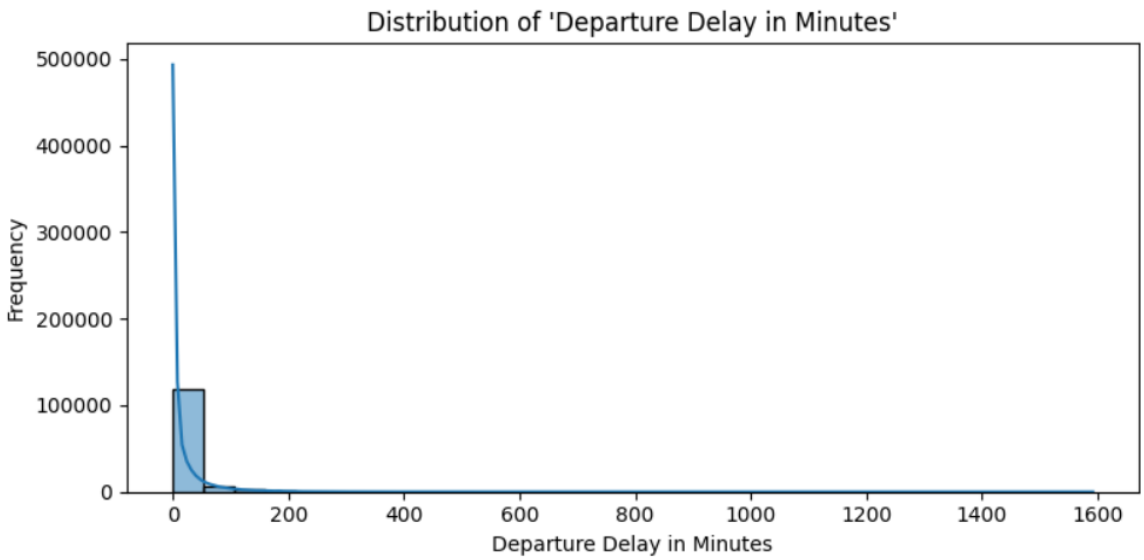

Distribution of 'Gender'
Frequency
Gender
0
10000
20000
30000
40000
50000
60000
0.0
0.2
0.4
0.6
0.8
1.0

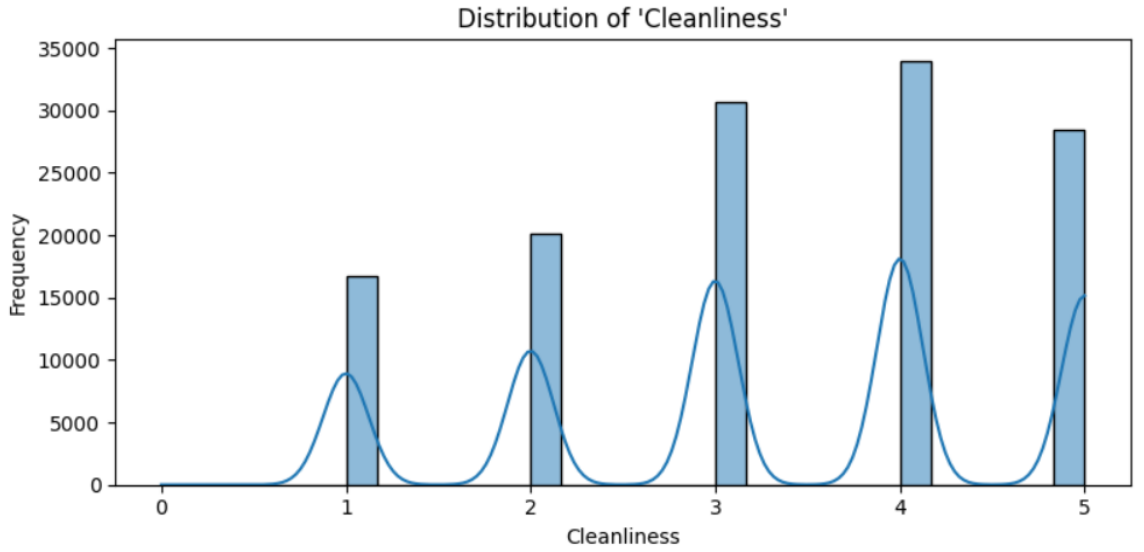

Class Distribution After SMOTE
60000
50000
40000
30000
20000
10000
0
Count
0
1
Satisfaction

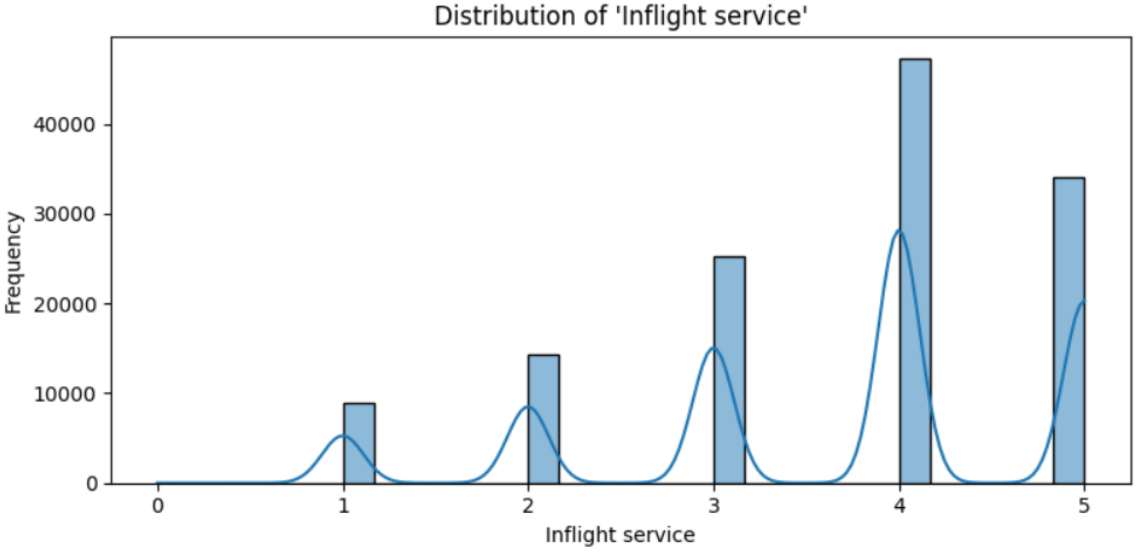

Class Distribution Before SMOTE
60000
50000
40000
30000
20000
10000
0
Count
0
1
Satisfaction

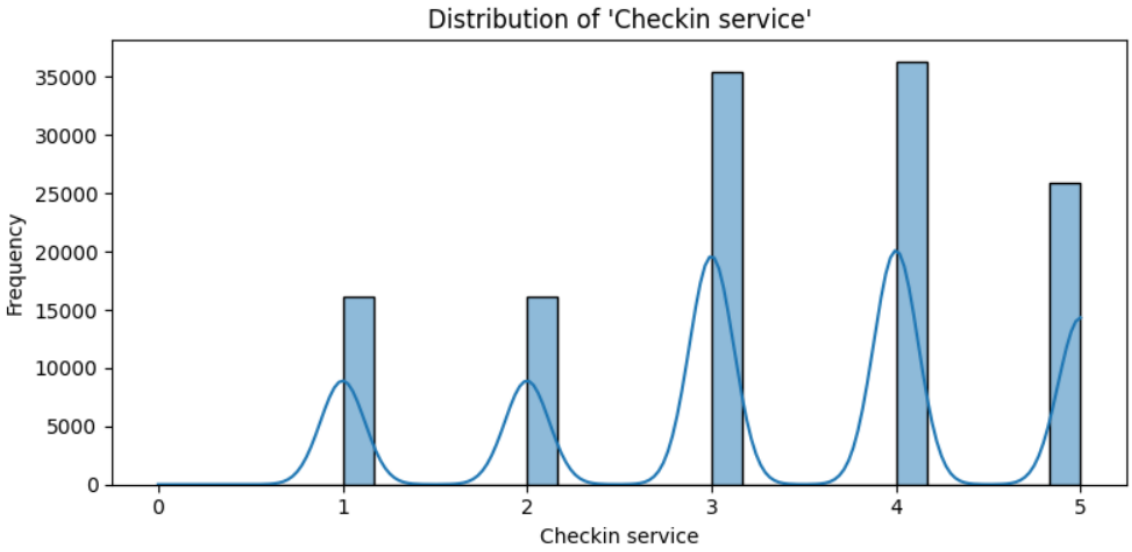

Target Variable Distribution: 'satisfaction'
Count
70000
60000
50000
40000
30000
20000
10000
0
0
1
Satisfaction

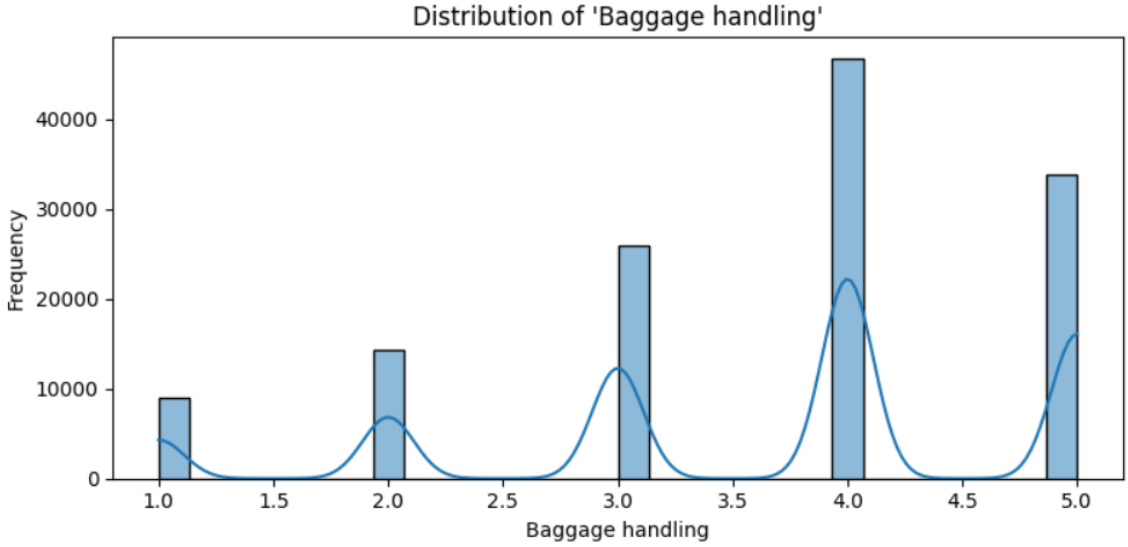

Distribution of 'Arrival Delay in Minutes'
Frequency
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400
1600
Arrival Delay in Minutes

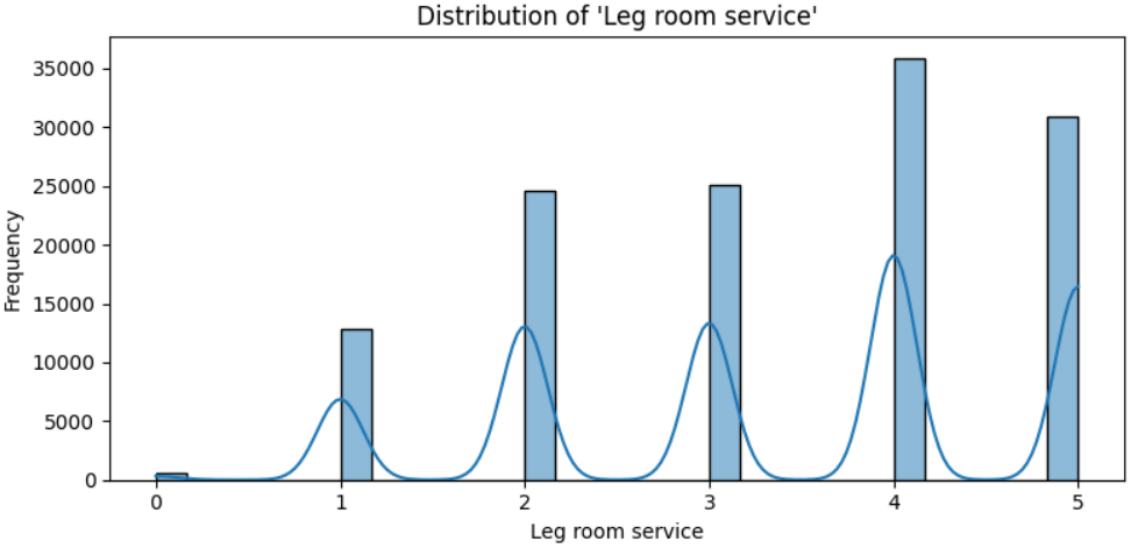

Distribution of 'Departure Delay in Minutes'
Frequency
Departure Delay in Minutes
0
100000
200000
300000
400000
500000
0
200
400
600
800
1000
1200
1400
1600

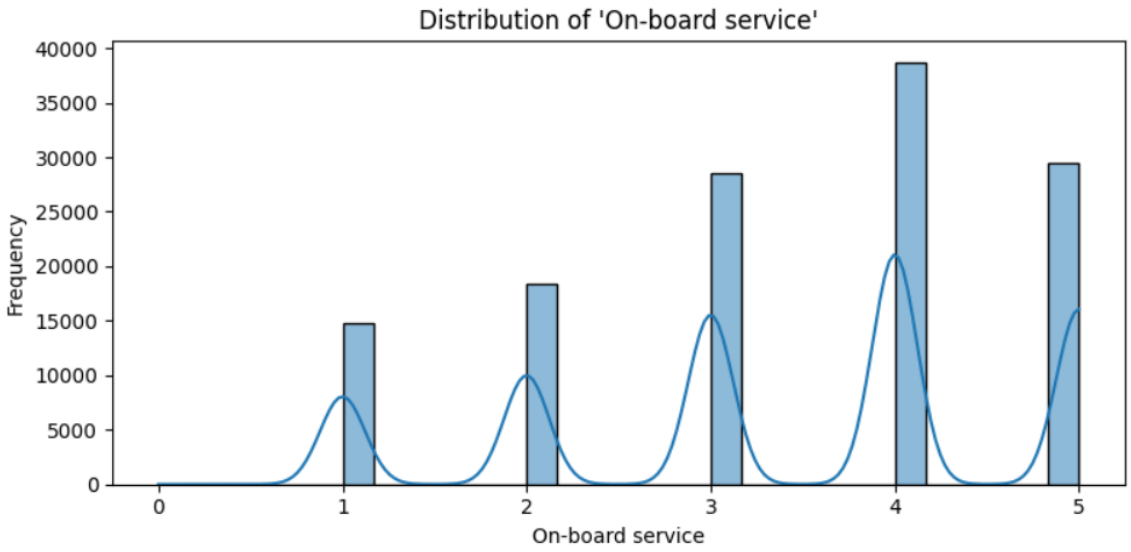

Distribution of 'Cleanliness'
Frequency
Cleanliness
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
4
5

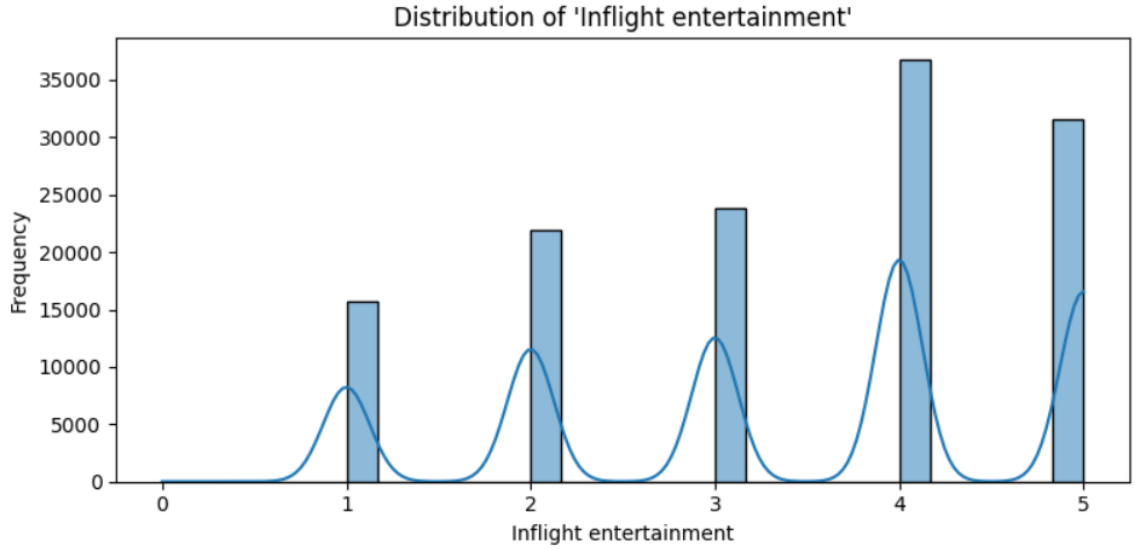

Distribution of 'Inflight service'
Frequency
Inflight service
0
10000
20000
30000
40000
0
1
2
3
4
5

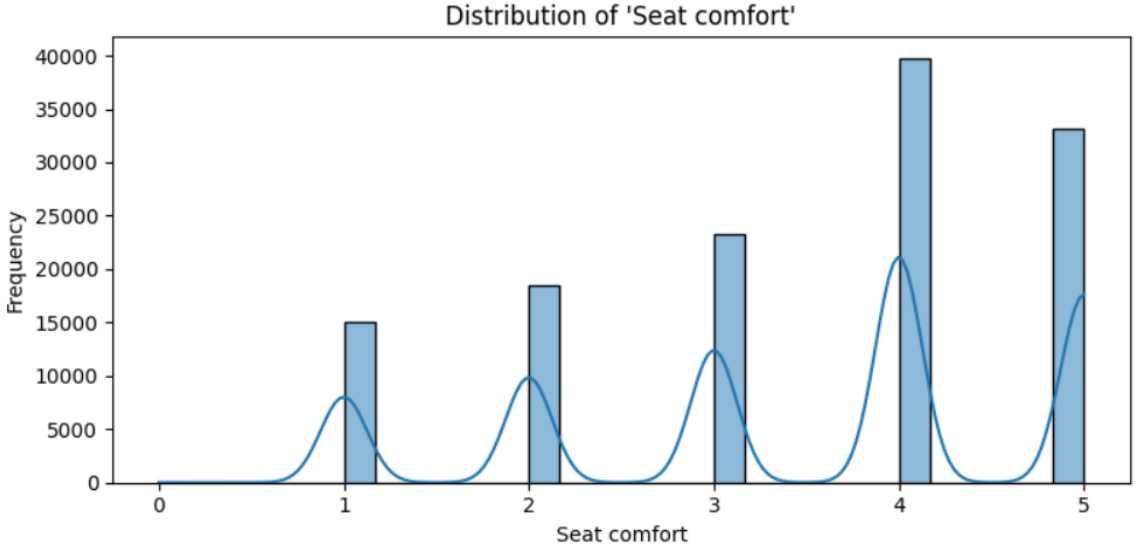

Distribution of 'Checkin service'
Frequency
Checkin service
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
4
5

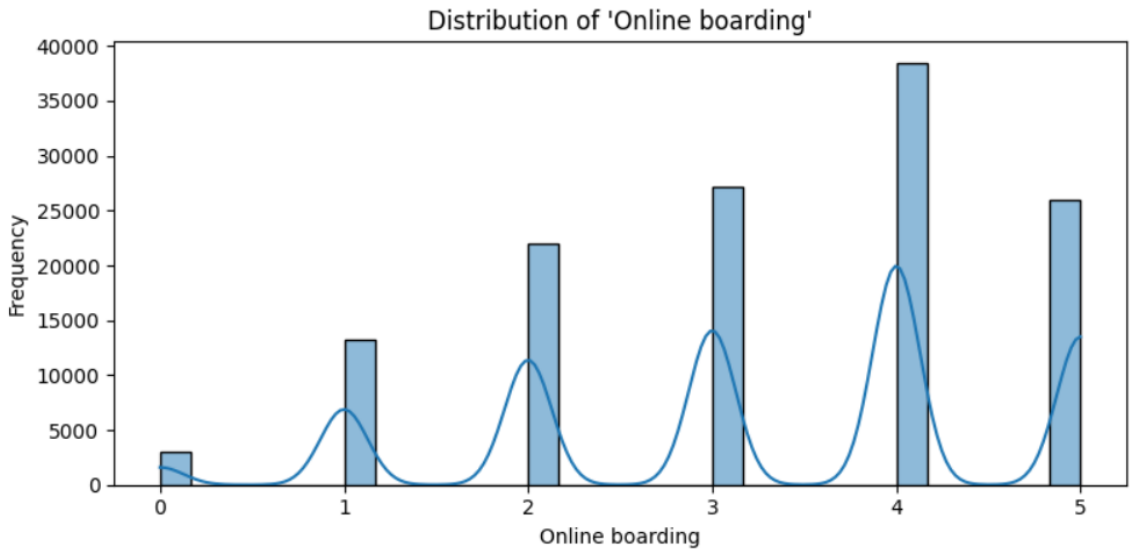

Distribution of 'Baggage handling'
Frequency
Baggage handling
0
10000
20000
30000
40000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

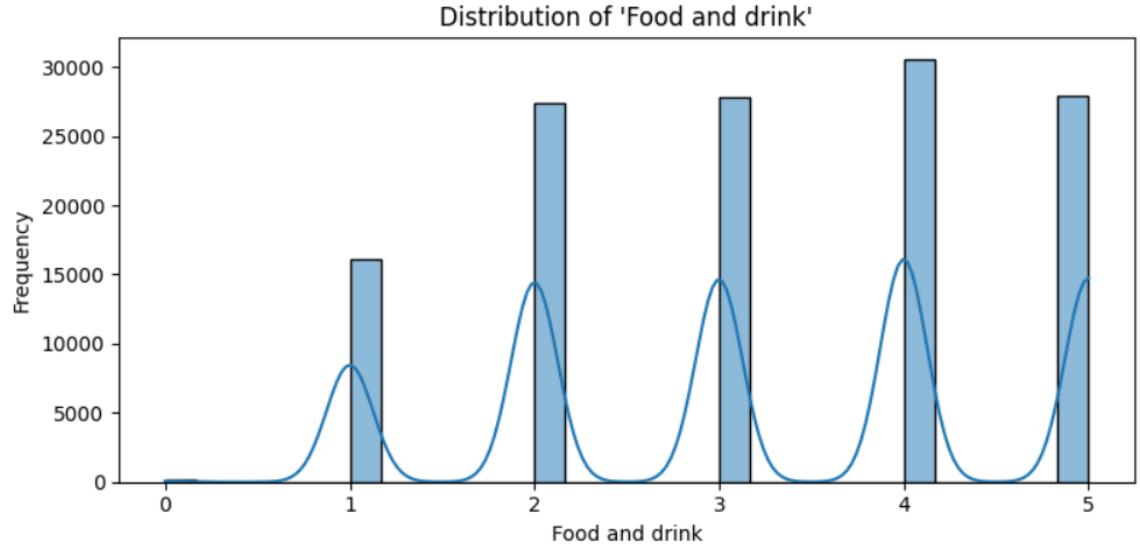

Distribution of 'Leg room service'
Frequency
Leg room service
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
4
5


Distribution of 'On-board service'
Frequency
On-board service
0
5000
10000
15000
20000
25000
30000
35000
40000
0
1
2
3
4
5


Distribution of 'Inflight entertainment'
Frequency
Inflight entertainment
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
4
5


Distribution of 'Seat comfort'
Frequency
Seat comfort
0
5000
10000
15000
20000
25000
30000
35000
40000
0
1
2
3
4
5


Distribution of 'Online boarding'
Frequency
Online boarding
0
5000
10000
15000
20000
25000
30000
35000
40000
0
1
2
3
4
5


Distribution of 'Food and drink'
Frequency
Food and drink
0
5000
10000
15000
20000
25000
30000
0
1
2
3
4
5