Airline Dataset satisfaction

1. We were asked to add additional features to the data set besides the ones you selected for approval in the following file: pairings and data sets - what we added is:

    - **Total In-flight Service Score**: Combining ratings for various in-flight services (such as inflight wifi service, seat comfort, inflight entertainment, etc.) into a single score.
    - **Total Ground Service Score**: Combining ratings for ground services (such as ease of online booking, check-in service, baggage handling, etc.) into a single score.
    - **Total Delay Score**: Combining departure delay and arrival delay into a single score to measure overall delay experience.
    - **Travel Distance Category**: Categorizing flight distance into short, medium, and long-haul flights.
    - **Overall Satisfaction Score**: Combining ratings for various aspects of the flight experience (in-flight service, ground service, etc.) into a single satisfaction score.
    - **Time of Day**: Extracting information from departure/arrival times to categorize flights into morning, afternoon, evening, and night flights.

2. Describe your dataset & features before processing it –
    - **Dimensions**: Determine the number of rows (instances) and columns (features) in the dataset using the shape attribute or the info() method.

    - **Data Types**: Check the data types of each feature using the dtypes attribute or the info() method. This helps identify whether features are numerical or categorical.

    - **Summary Statistics:** Calculate summary statistics for numerical features, such as mean, median, minimum, maximum, and standard deviation, using the describe() method.

    - Unique Values: Determine the number of unique values and frequency counts for categorical features using the nunique() and value_counts() methods.

3. Describe the process of cleansing, fixing, prepping the dataset –
   We load the dataset into a DataFrame df and split it into features (X) and target variable (y).
   We define preprocessing steps for numerical and categorical features using scikit-learn's Pipeline and ColumnTransformer.
   We define a pipeline that combines preprocessing steps with any modeling steps.
   We fit the pipeline to the training data and transform both the training and testing data.

Optionally, we convert the transformed data back to DataFrame format.
This code provides a structured approach to preprocess the dataset, handle missing values, scale numerical features, encode categorical features, and prepare the data for modeling.

4. Describe your dataset & features after processing it –
   - Dimensions: Ensure the number of rows and columns remains appropriate.
   - Data Types: Confirm if there are any changes in data types.
   - Summary Statistics: Reassess distributions and central tendencies of numerical features.
   - Unique Values: Verify encoding techniques and check for loss of information.
   - Missing Values: Ensure missing values have been handled appropriately.
   - Feature Engineering: Describe any new features created.
   - Initial Observations: Note any notable changes or insights gained from preprocessing.

5. Describe distribution of interesting features and what can be learned about them –
   Here's a concise summary of what we obtain from the provided code using your dataset:

   1. Visualizing Distributions: We generate histograms for selected features to understand their distributions, including features such as 'Age', 'Flight Distance', 'Inflight wifi service', and 'Seat comfort'.

   2. Summary Statistics: We calculate summary statistics (mean, median, etc.) for these features to quantify their central tendency and variability.

   3. Skewness and Outliers: We measure skewness and identify outliers in the data to understand its distributional properties and identify potential anomalies.

   4. Correlation Matrix: We analyze the correlation between selected features to understand their relationships and identify potential redundancies or correlations.

   By examining these aspects of the data, we gain insights into its characteristics, patterns, and potential issues, which can inform further analysis or preprocessing steps.

6. Assume multiple assumptions about the data, explain why you made those assumptions based on the work you've done so far:
   Here are some example assumptions we can make about the data and reasons for making them based on the provided code and dataset:

1. **Assumption**:
There may be a positive correlation between 'Inflight wifi service' and 'Online boarding' ratings.
Reasoning:
Passengers satisfied with inflight wifi may also have positive experiences with online boarding, as both relate to the convenience of technology use during the travel process.

2. **Assumption:**
Younger passengers may give higher ratings for 'Inflight entertainment'.
Reasoning:
Younger individuals often have higher expectations for entertainment options, and they may prioritize this feature more than older passengers.

3. **Assumption:**
Passengers traveling for business purposes ('Type of Travel' = 'Business travel') may be more likely to rate 'Inflight wifi service' higher than those traveling for personal reasons ('Type of Travel' = 'Personal Travel').
Reasoning:
Business travelers may have a greater need for inflight wifi to stay connected for work purposes, leading them to prioritize this feature more than leisure travelers.

4. **Assumption**:
Passengers in higher class categories ('Class' = 'Business' or 'First') may give higher ratings for 'Seat comfort' compared to those in the 'Eco' (Economy) class.
Reasoning:
Higher class passengers typically pay more for additional comfort and amenities, including seat comfort, which may lead to higher ratings in this category.

5. **Assumption**:
Longer 'Flight Distance' may correlate with higher ratings for 'Inflight entertainment'.
Reasoning:
Passengers on longer flights may value entertainment options more to alleviate boredom during extended travel times.

7. Prove and disprove the assumptions you have made :

   We load the dataset and proceed to analyze each assumption using statistical methods and visualizations.
   For each assumption, we calculate relevant statistics, such as correlation coefficients or mean ratings, and visualize the relationships using box plots.
   The results will help us determine whether the assumptions are supported by the data or if they need to be revised based on the evidence.

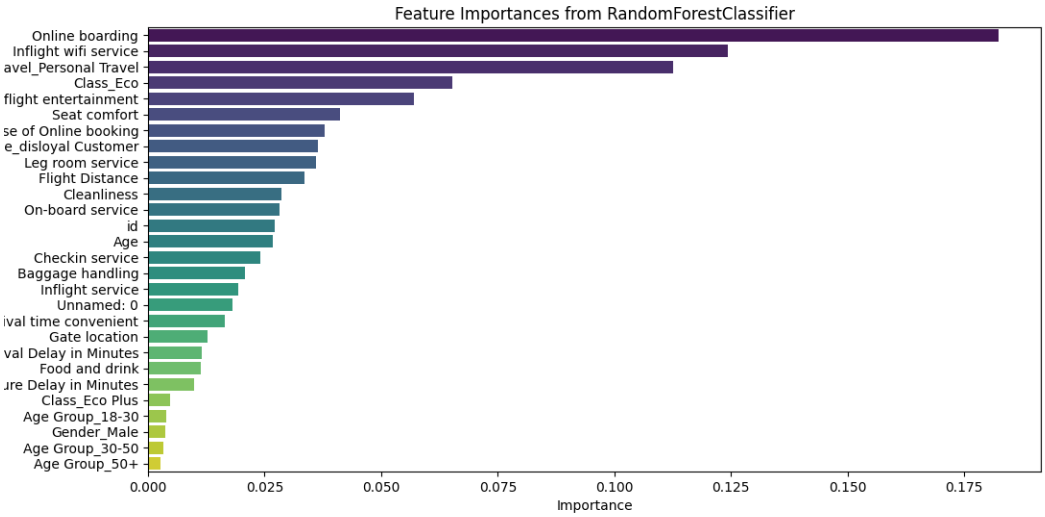
8. Find interesting correlations between features, explain these correlations and prove your claims using plots –

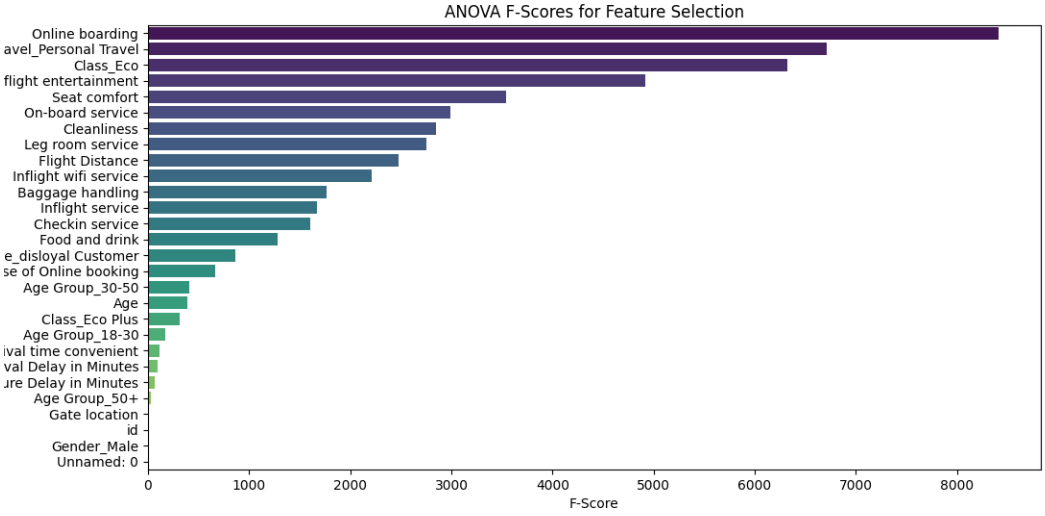
   To find interesting correlations between features, we can calculate the correlation matrix for numerical features in the dataset and identify significant correlations. Then, we can visualize these correlations using heatmaps to provide a clear understanding of the relationships between features.
   We load the dataset and select numerical features for correlation analysis.
   We calculate the correlation matrix using the .corr() method.
   We visualize the correlation matrix as a heatmap using seaborn's heatmap() function.
   Correlation values close to 1 indicate a strong positive correlation, while values close to -1 indicate a strong negative correlation. Values close to 0 suggest little to no correlation between features.
   This visualization allows us to identify interesting correlations between features and gain insights into how they relate to each other. We can then interpret these correlations based on domain knowledge and the context of the dataset

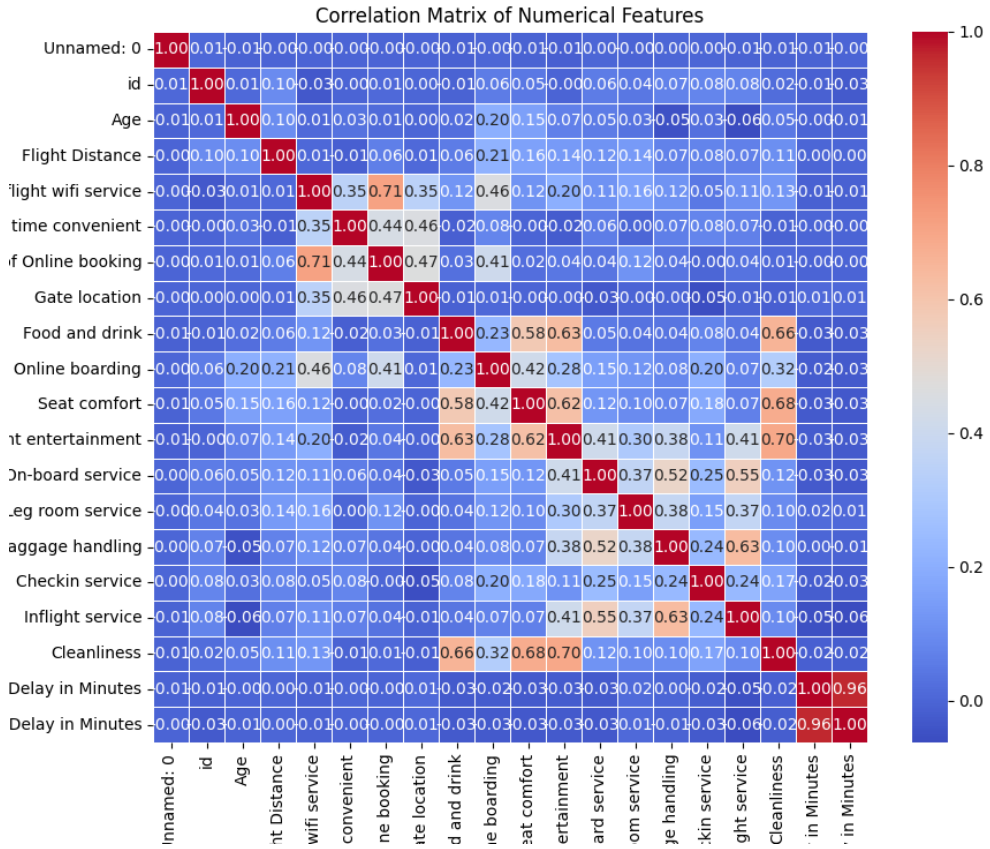
9. we employed ANOVA F-scores for feature selection and RandomForestClassifier for model-based analysis to identify the most influential features on customer satisfaction. Using ANOVA F-scores, we assessed each feature's importance statistically and visualized the results through a bar plot. Additionally, RandomForestClassifier helped quantify feature importances by training on the dataset and extracting relative importance values. Combining both techniques provided insights into key features driving customer satisfaction in the dataset, enhancing our understanding of its underlying factors.

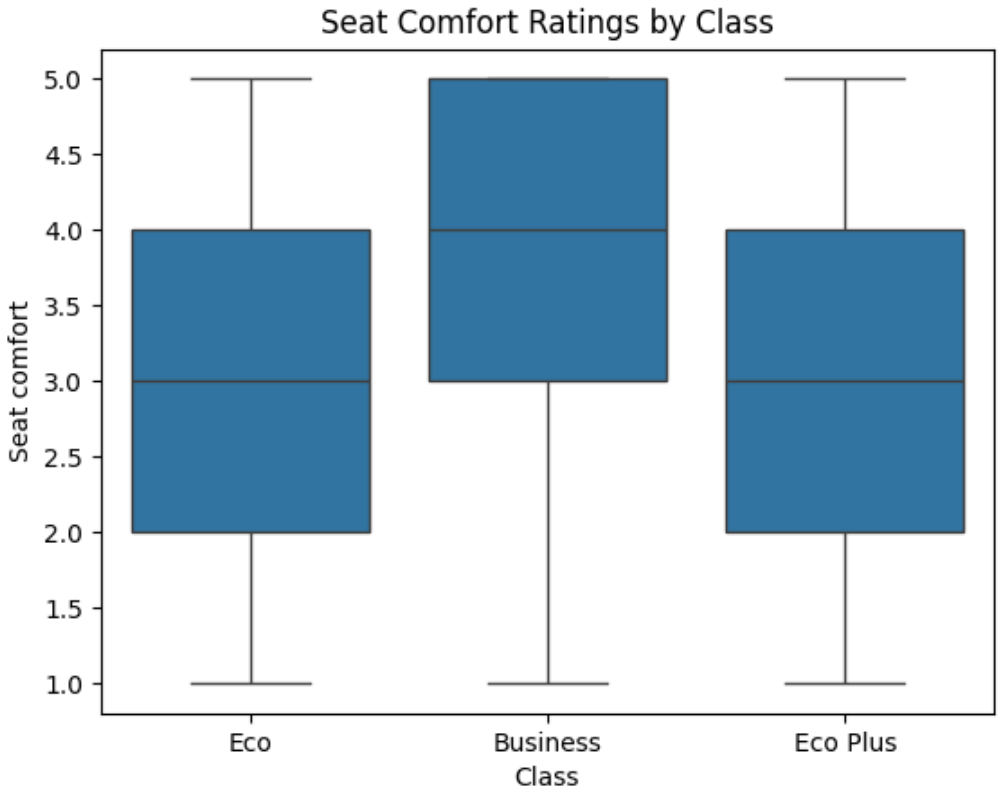
10. Exploring this dataset was a journey! We started by looking closely at how it's set up and what's inside. Then, we carefully cleaned up the data, making sure it was all good to go for analysis. This meant fixing any missing pieces and making sure categories were sorted out properly. As we dug deeper, we noticed some interesting connections between different parts of the data. It was cool to see how certain things affected passenger happiness the most. By testing out different ideas and assumptions, we got a better handle on how the data works. This whole experience taught us how important it is to get the data right before building models.

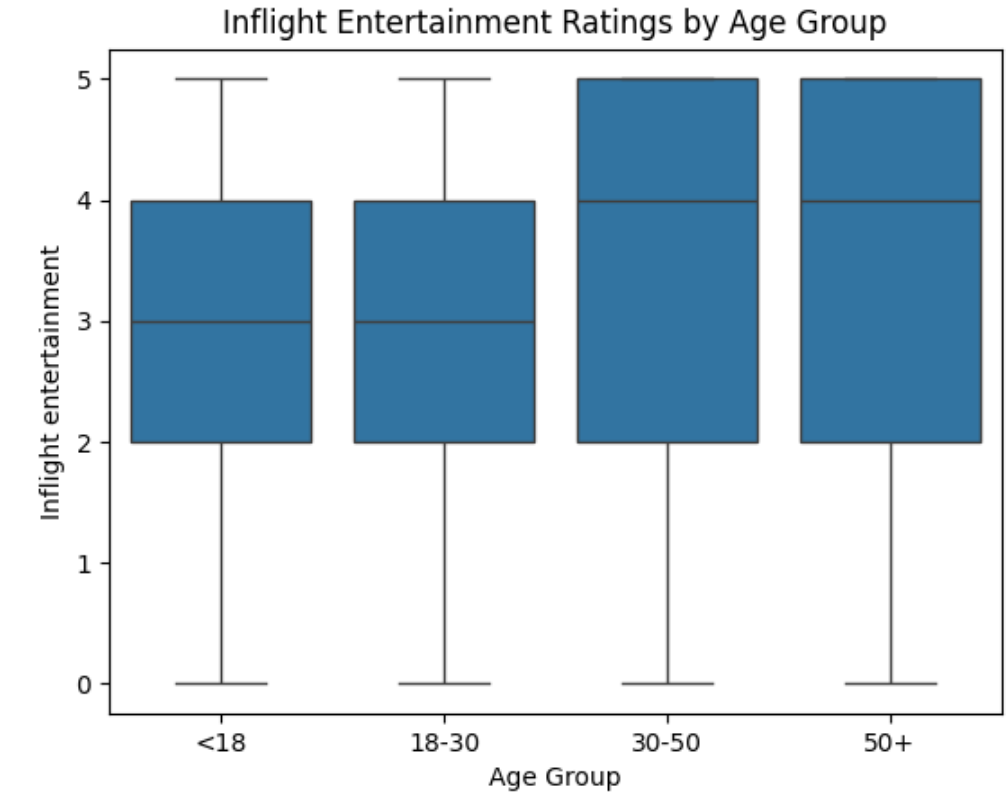
Graphs output:

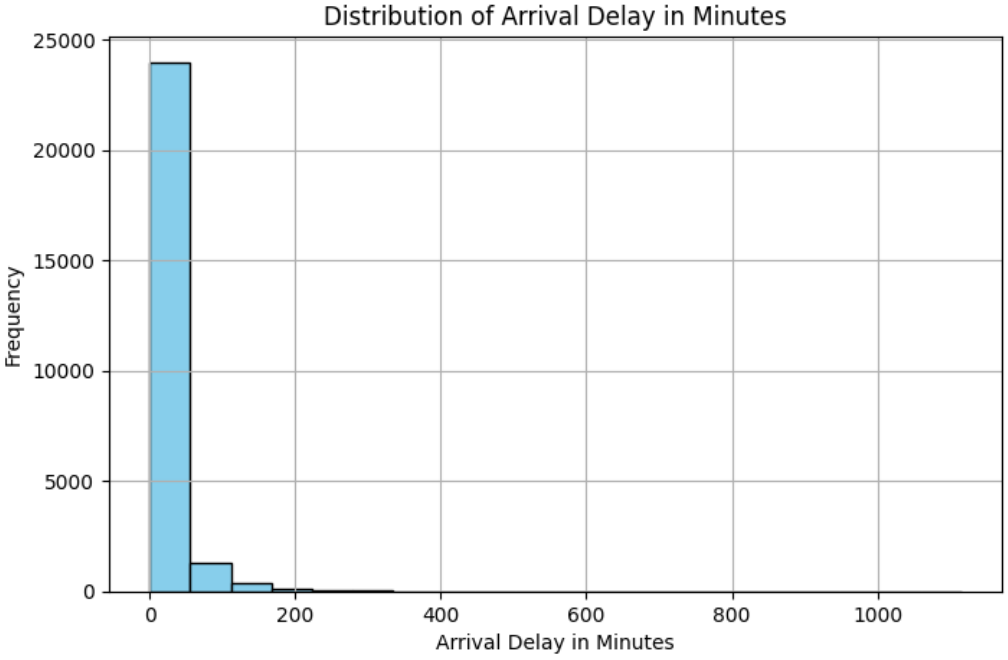


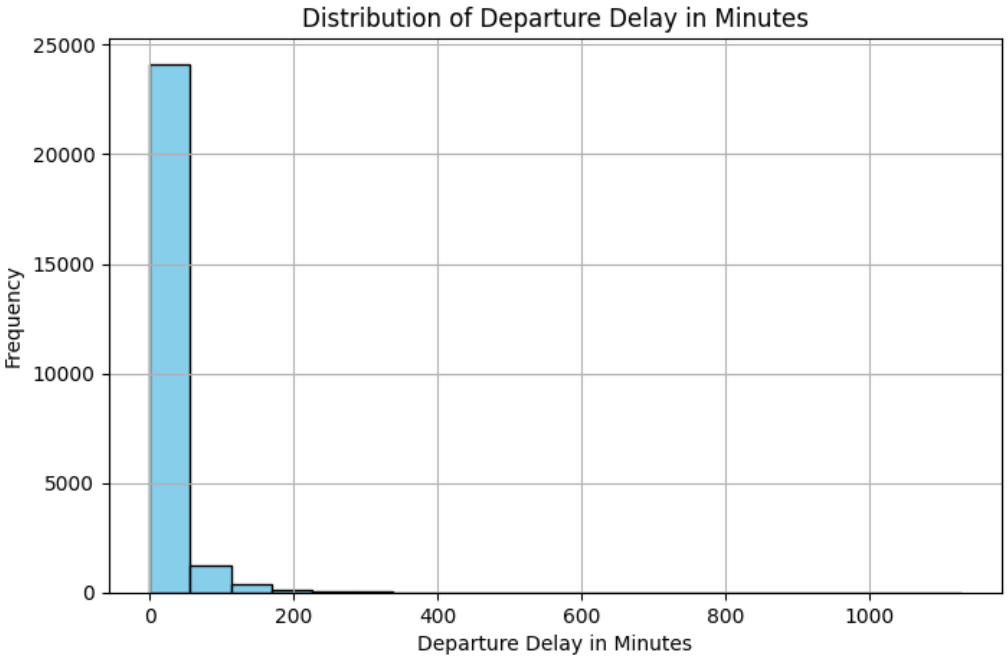


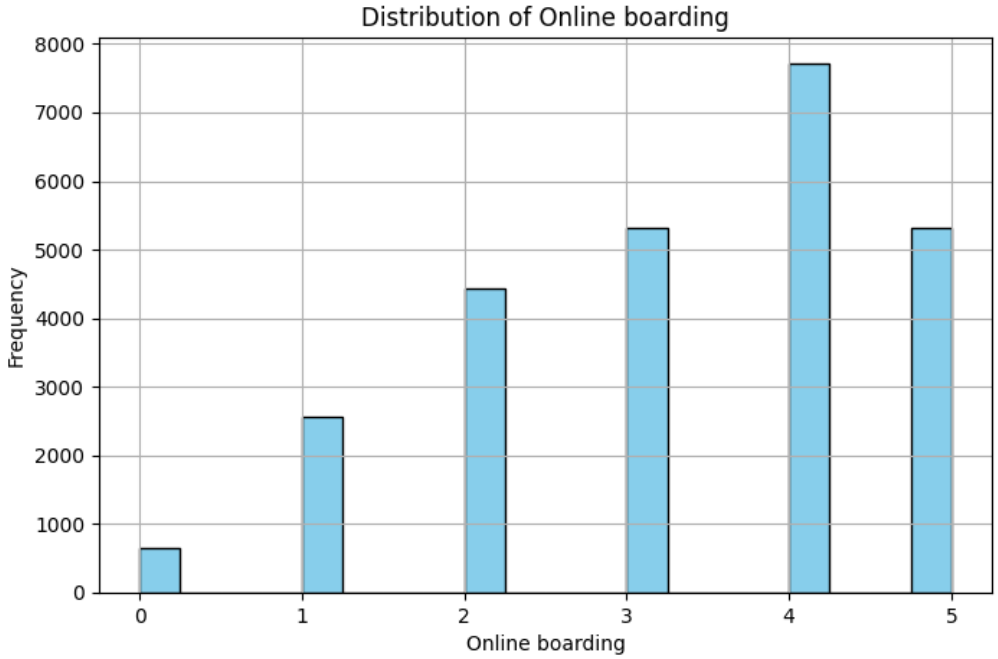

Correlation Matrix of Numerical Features
Unnamed: 0
id
Age
Flight Distance
flight wifi service
time convenient
f Online booking
Gate location
Food and drink
Online boarding
Seat comfort
nt entertainment
On-board service
eg room service
aggage handling
Checkin service
Inflight service
Cleanliness
Delay in Minutes
Delay in Minutes
1.0
0.8
0.6
0.4
0.2
0.0

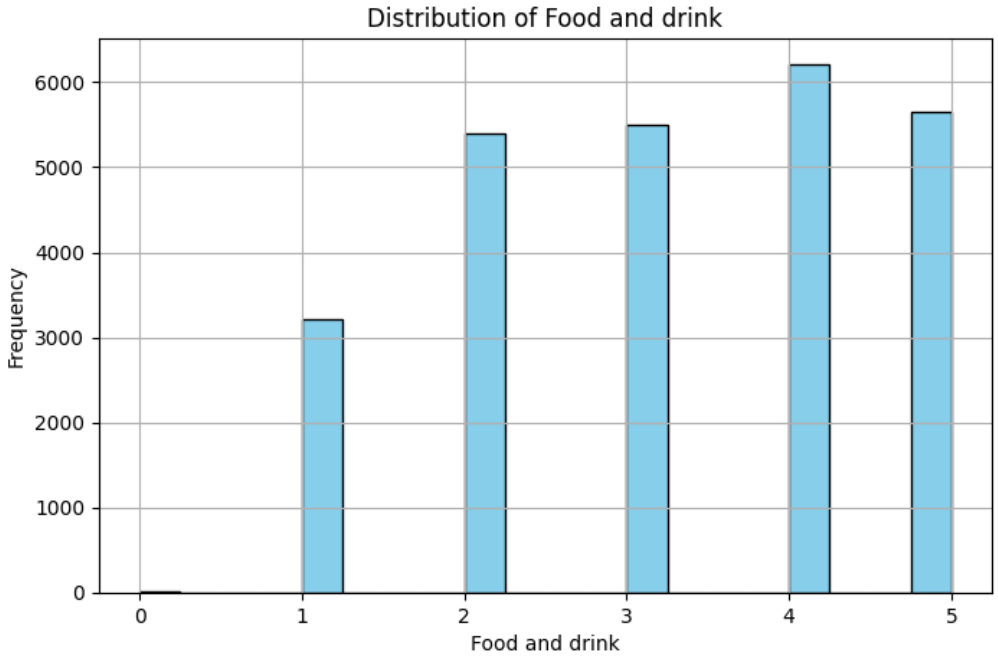

Seat Comfort Ratings by Class
Seat comfort
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Eco
Business
Eco Plus
Class


Inflight Entertainment Ratings by Age Group
Inflight entertainment
0
1
2
3
4
5
<18
18-30
30-50
50+
Age Group

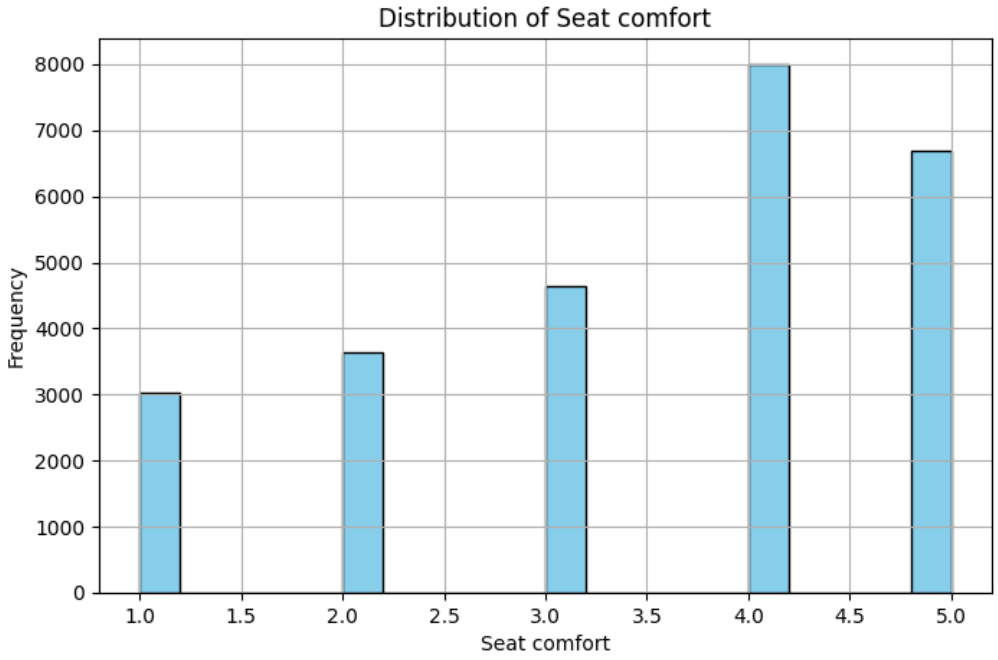

Distribution of Arrival Delay in Minutes
Frequency
0
5000
10000
15000
20000
25000
0
200
400
600
800
1000
Arrival Delay in Minutes

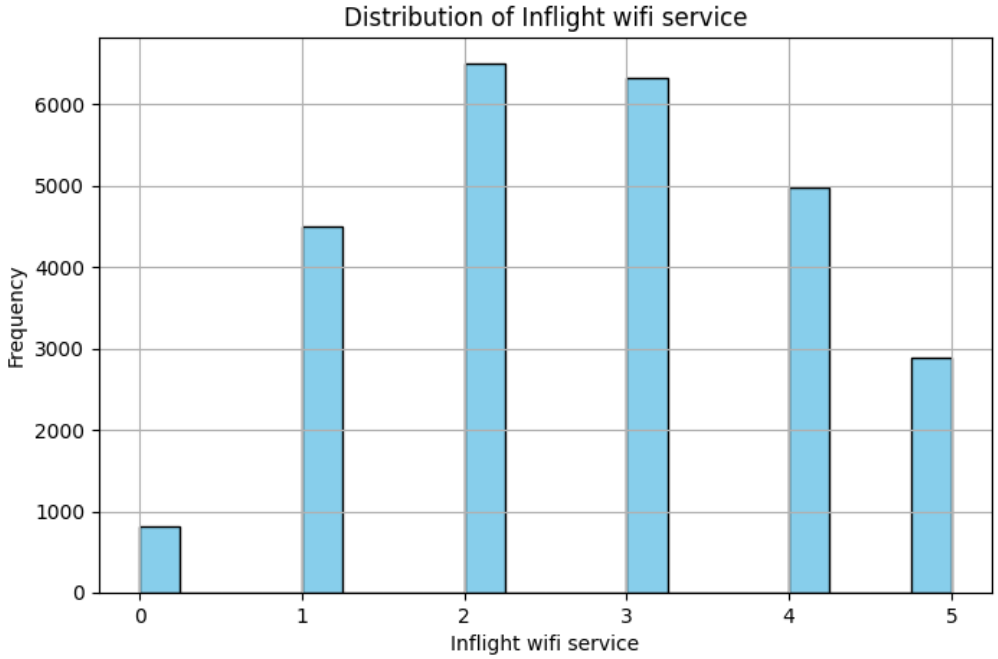

Distribution of Departure Delay in Minutes
Frequency
Departure Delay in Minutes
0
5000
10000
15000
20000
25000
0
200
400
600
800
1000

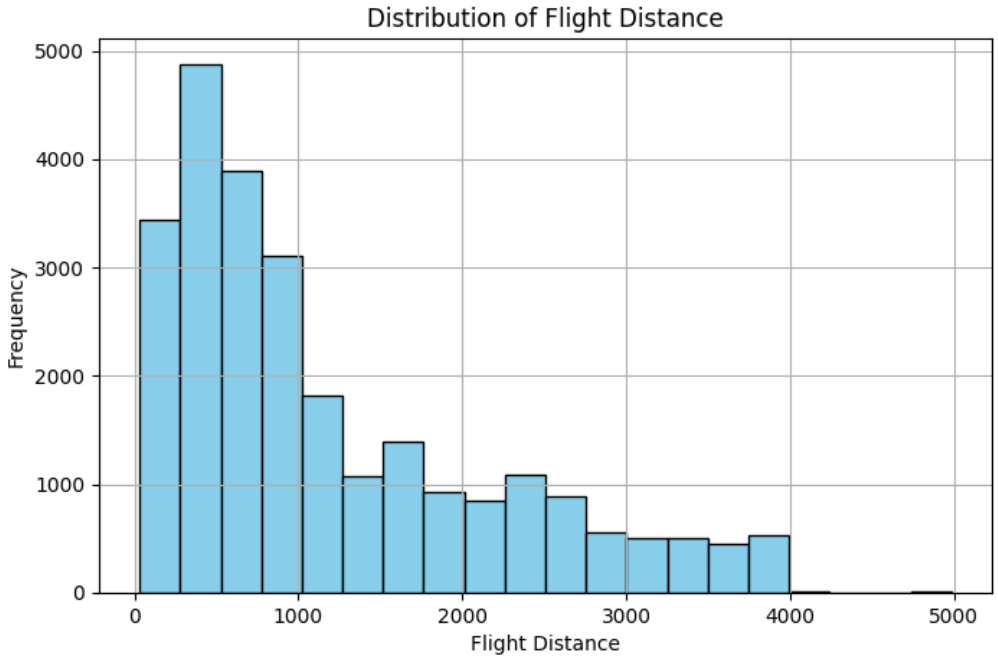

Distribution of Online boarding
Frequency
Online boarding
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
2
3
4
5

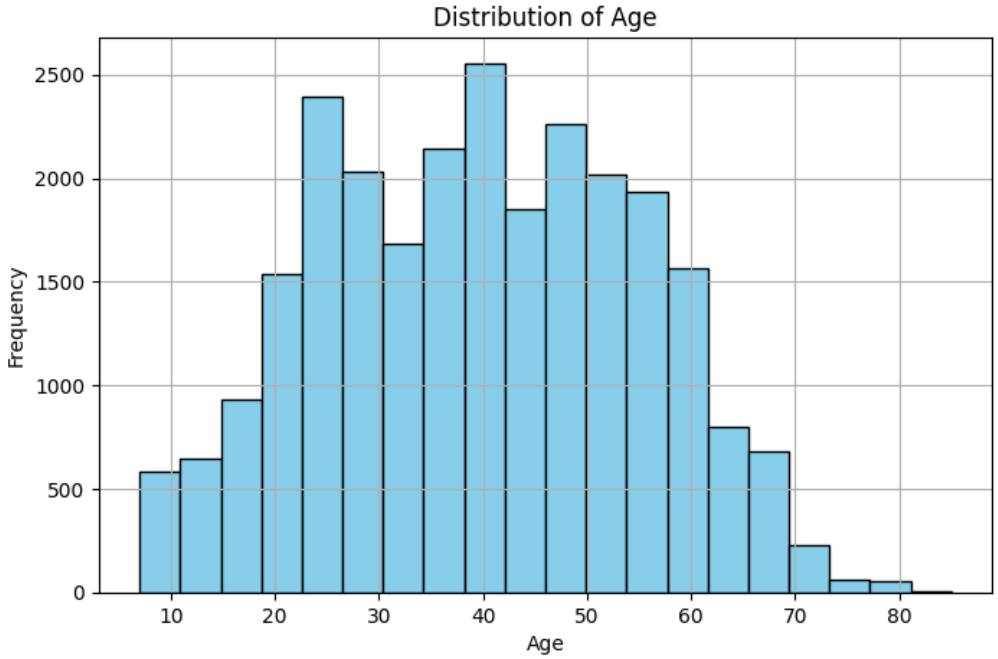

Distribution of Food and drink
Frequency
Food and drink
0
1000
2000
3000
4000
5000
6000
0
1
2
3
4
5


Distribution of Ease of Online booking
Frequency
Ease of Online booking
0
1000
2000
3000
4000
5000
6000
0
1
2
3
4
5


Distribution of Seat comfort
Frequency
Seat comfort
0
1000
2000
3000
4000
5000
6000
7000
8000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Distribution of Inflight wifi service
Frequency
Inflight wifi service
0
1000
2000
3000
4000
5000
6000
0
1
2
3
4
5


Distribution of Flight Distance
Frequency
Flight Distance
0
1000
2000
3000
4000
5000


Distribution of Age
Frequency
Age
0
500
1000
1500
2000
2500
10
20
30
40
50
60
70
80