Analysis of Flight Delays

1. Research Scenario and Questions

Scenario: Flight delays are a common issue for air travelers, causing inconvenience and disruptions to passengers, airlines, and airports. Understanding the factors that contribute to flight delays can help airlines and airports develop strategies to minimize these disruptions and improve overall service quality.

Research Questions:

1. What is the relationship between flight length and the likelihood of delay?
2. Are certain days of the week more prone to flight delays?
3. Are there specific airlines or airport routes with a higher probability of flight delays?

2. Data Set Description

The data set used for this project can be found here: https://raw.githubusercontent.com/datasets/openml-datasets/master/data/airlines/airlines.csv

The dataset includes the following variables:

1. **Airline:** The airline company operating the flight.
2. **Flight:** The flight number.
3. **AirportFrom:** The departure airport.
4. **AirportTo:** The arrival airport.
5. **DayOfWeek:** The day of the week the flight occurred (1 = Monday, 7 = Sunday).
6. **Time:** The departure time of the flight.
7. **Length:** The flight duration in minutes.
8. **Delay:** A binary variable indicating whether a flight was delayed (0 = true, 1 = false).

A random sample of 1000 lines will be used for the analysis. Data cleaning will involve checking for missing or duplicate values, as well as outliers.

3. Statistical Methods

The statistical methods that will be used for this analysis include:

1. **Descriptive statistics:**
   To provide an overview of the data and explore the distribution of variables.
2. **Chi-square test:**
   To investigate the association between categorical variables (e.g., DayOfWeek and Delay).
3. **Logistic regression:**
   To examine the relationship between multiple variables and the likelihood of a flight delay.

4. Results

The results of the analysis will be presented using tables and figures, which will include:

1. Summary statistics for each variable.
2. Bar charts and/or line graphs illustrating the relationship between DayOfWeek, Airline, AirportFrom, AirportTo, and the proportion of flights delayed.
3. Logistic regression output, including coefficients, odds ratios, and p-values, to determine the factors significantly associated with flight delays.

5. Conclusions and Limitations

The conclusions will summarize the key findings of the analysis in non-technical terms, highlighting the factors that are associated with flight delays. Additionally, any limitations of the analysis will be discussed, such as potential violations of statistical test assumptions or the limited generalizability of the results due to the small sample size. Furthermore, any recommendations for further research or strategies to reduce flight delays will be provided.