Final Term Project Proposal – Flight Delays and Cancellations Dataset

DATS 6401

Dhatrija Sukasi

Final Term Project Proposal – Flight Delays and Cancellations Dataset

Dataset Selection and Justification

For my project, I have selected the U.S. Flight Delays and Cancellations Dataset from Kaggle. This dataset contains over 5 million flight records across multiple years, which comfortably exceeds the minimum requirement of 50,000 observations. Air travel plays a critical role in the global economy, and delays or cancellations have significant impacts on passengers, airlines, and operations. Therefore, analyzing flight delay patterns provides valuable insights into transportation efficiency, service quality, and operational planning. The dataset includes both categorical and numerical features, making it highly suitable for visualization and dashboard development.

Features and Classification

- Numerical Features:
    1. Departure Delay (minutes)
    2. Arrival Delay (minutes)
    3. Flight Distance (miles)
    4. Air Time (minutes)
    5. Taxi-Out Time (minutes)
    6. Taxi-In Time (minutes)
- Categorical Features:
    1. Airline Carrier
    2. Origin Airport
    3. Destination Airport
    4. Flight Status (on-time, delayed, cancelled)
    5. Day of Week / Month (derived from date)
- Potential Feature Engineering:
    - Create a binary variable for "Severely Delayed" (e.g., >30 minutes).
    - Categorize delays by cause (weather, carrier, late aircraft, etc.).
    - Extract time-based features (hour of departure, seasonality).

Proposed Static Plots

- Numerical Features:
    - Histogram of Departure and Arrival Delays.
    - Boxplots of Delays grouped by Airline.
    - Line charts showing trends in delays across months/years.
    - Scatter plot of Flight Distance vs. Air Time.
- Categorical Features:
    - Bar chart of delays per Airline.
    - Pie chart of delay causes (weather, carrier, etc.).
    - Bar chart of flights per Airport (top 10 busiest origins/destinations).
    - Heatmap of flights by Day of Week vs. Hour of Day.

Conclusion

The U.S. Flight Delays and Cancellations dataset offers a robust foundation for data visualization, enabling exploration of both numerical and categorical dimensions. It provides opportunities for feature engineering, statistical testing, and dimensionality reduction. The dashboard will communicate insights through static and interactive visualizations, making it relevant for stakeholders such as travellers, airlines, and policy makers. This dataset satisfies all project requirements and offers rich storytelling potential through data.