Visualizing Airline Flight Delay and Cancellation Data

Lebedev Georgiy Konstantinovich, Jiang Yi, Tsakalidou Ioanna

Data Visualization, Milestone 1

1. Dataset

We will explore the Airline Flight Delay and Cancellation Data from the US Department of Transportation's Bureau of Transportation Statistics. This dataset contains comprehensive information about domestic flights operated by major air carriers in the United States covering the period from Aug 2019 to Aug 2023. The dataset can be found in the US Department of Transportation, Bureau of Transportation Statistics.

For every flight, it contains the following fields:

- FlightDate
- Reporting_Airline
- Flight_Number_Reporting_Airline
- Origin & OriginCityName
- Dest & DestCityName
- CRSDepTime
- DepTime
- DepDelay (early departures show negative numbers)
- TaxiOut
- CRSArrTime
- ArrTime
- ArrDelay (early arrival show negative numbers)
- TaxiIn
- CRSElapsedTime
- ActualElapsedTime
- AirTime
- Cancelled
- CancellationCode
- Diverted
- CarrierDelay
- WeatherDelay
- NASDelay
- SecurityDelay
- LateAircraftDelay

The complete dataset consists of approximately 29 million rows, which is too large to visualize interactively.

Therefore, we will work with a reduced version containing about 3 million rows, which still provides a comprehensive view of flight patterns while being manageable for visualization purposes.

Some preprocessing will be required to address potential issues in the data:

- Handling missing values, especially in delay reason attributions
- Converting time formats for consistency
- Creating derived fields such as delay categories and seasonal indicators
- Merging with supplementary data such as airport geographic coordinates for mapping purposes

2. Problematic

Our objective is to develop a data visualization tool that enables users to analyze patterns in flight delays and cancellations in the United States. The tool will allow users to explore how various factors—such as time of year, day of the week, carrier, route, and weather conditions—affect flight reliability and performance.

Target Audience: Our tool is designed for anyone who is interested in exploring aviation related data. It can satisfy the needs of a large variety of users, from passengers who want to optimize their flight booking decisions based on historical reliability data to data analysts and aviation agencies for proposing better air traffic control and operation rules. For example, a frequent business traveler might use our tool to determine which day of the week or what time of a day has the lowest delay probability on their regular route. Similarly, airline operators could analyze which routes/airports experience the most seasonal variations in on time performance.

Specifically, our tool will include the following visualizations:

- **Interactive Map**: Our primary visualization will feature a map of the United States that displays flight routes as lines connecting origin and destination airports. The thickness of these lines will represent the volume of flights, while color coding will indicate average delay times. Users can filter by specific airlines, time periods, or delay thresholds, with the map updating dynamically to reflect these selections.

When the users adjust filters or select specific regions, our tool will automatically update several supplementary visualizations:

- Plot 1: Average Delay by Time of Day and Day of Week. This heatmap will display how delays vary throughout the day and across different days of the week.
- Plot 2: Delay Causes Distribution. This plot will break down the causes of delays (weather, carrier, national aviation system, security, late aircraft) across different carriers and routes.
- Plot 3: Seasonal Patterns in Flight Cancellations. This visualization will show how cancellation rates and reasons vary across different months and seasons.

3. Exploratory Data Analysis

Our exploratory analysis in the `flight_data_exploration` notebook revealed several interesting patterns:

- The distribution of delays follows a right-skewed pattern, with most flights experiencing either no delay or delays under 30 minutes, but with a significant "long tail" of extreme delays.
- Distinct seasonal patterns appear in both delay frequency and duration, with winter months (particularly December and January) showing higher delay rates.
- There are notable differences in performance metrics between major carriers, with some consistently outperforming others in on-time arrival.
- Certain airports and routes demonstrate significantly higher delay rates, suggesting local factors may play important roles in flight delays.
- The geographical distribution of delays shows regional patterns, with some areas of the country experiencing more consistent issues than others.
- Day of week effects are pronounced, with mid-week flights generally experiencing fewer delays than weekend flights.

4. Related Work

Airline delay data has been the subject of numerous studies and visualizations, including predictive modeling for delay forecasting [3], network analysis of the air transportation system [1], and impact assessments of weather events [2] on flight operations.

While these works provide valuable insights, they typically focus on specific aspects of the data or fixed time periods. Our approach differs by:

- Providing a comprehensive interactive visualization tool that allows users to dynamically explore multiple dimensions of the data simultaneously
- Including the most recent post-pandemic data (through August 2023), capturing the recovery and restructuring of the airline industry
- Focusing on the user experience by creating intuitive interfaces that make complex data accessible to non-technical audiences
- Incorporating geographic context and temporal patterns in an integrated visualization system

We draw inspiration from flight tracking visualizations such as FlightAware's Misery Map, the New York Times' interactive features on transportation systems, and Edward Tufte's principles of data visualization that emphasize clear presentation of complex multivariate data.

References

[1] Ajayi, Joseph, Yao Xu, Lixin Li, and Kai Wang. 2024. "Enhancing Flight Delay Predictions Using Network Centrality Measures" Information 15, no. 9: 559. https://doi.org/10.3390/info15090559

[2] Li, Qiang and Jing, Ranzhe and Dong, Zhijie Sasha, 2023. "Flight Delay Prediction With Priority Information of Weather and Non-Weather Features", Trans. Intell. Transport. Sys., no. 7: 7149, https://doi.org/10.1109/TITS.2023.3270743

[3] Aravinda, Jatavallabha, Jacob, Gerlach and Aadithya, Naresh, "Deciphering Air Travel Disruptions: A Machine Learning Approach", arXiv, 2024, https://arxiv.org/abs/2408.02802