Visualizing Airline Flight Delay and Cancellation Data

Lebedev Georgiy Konstantinovich, Jiang Yi, Tsakalidou Ioanna

Data Visualization, Milestone 3

1. Introduction

In this project, we create a visualization tool to enable any US-based air traveler to look into flight delay and cancellation information.

We explore the Airline Flight Delay and Cancellation Data from the US Department of Transportation's Bureau of Transportation Statistics, which contains comprehensive information about domestic flights operated by major air carriers in the United States from Aug 2019 to Aug 2023. The complete dataset consists of approximately 29 million rows.

Our objective is to develop a data visualization tool that enables users to analyze patterns in flight delays and cancellations in the US. Our primary visualization will feature a map of the US that displays airports as colored weighted circles, with their size and color coding indicating the average delay times. Users can filter by delay thresholds and zoom in/out of the geographic region, with the map updating dynamically to reflect these selections. Based on this input, the tool shows delay statistics, such as the most and least delayed airports in a self-evident way, to aid users' flight booking decisions to avoid or prefer certain airports.

Our tool is designed for any US-based air traveler interested in minimizing the time spent at the airport. It caters to a wide range of users, from frequent business flyers needing to quickly assess potential delays for upcoming trips to occasional leisure travelers planning family vacations and seeking to avoid common travel headaches. By providing clear, intuitive visualizations of historical flight delay and cancellation data, our platform empowers users to make informed decisions, choose optimal travel times, and ultimately enjoy a smoother, more predictable journey.

We believe that our project and tool can be a helpful resource used both casually and in more professional circumstances.

2. Design Process

During the initial idea and planning phase, our team was motivated by our collective interest in air traffic data visualization. We recognized that while raw flight data is abundant, it often remains inaccessible and difficult to interpret for the average traveler. This presented a clear opportunity to transform complex datasets into an intuitive, visually engaging, and highly functional web application. Our goal was to empower users with the insights needed to navigate the often-turbulent skies of air travel, providing a tool that goes beyond mere flight tracking to offer a deeper understanding of delay patterns and cancellation risks geographically. For example, there are currently more than 5,000 airports in public use and an average of 45,000 domestic commercial flights in the US daily. This means that a vast and complex web of interconnected operations is constantly in motion, making it challenging for individual travelers to discern patterns or anticipate disruptions without specialized tools. The sheer volume of data generated by this system, from individual flight statuses to large-scale weather impacts, necessitates powerful visualization techniques to extract meaningful insights and provide actionable intelligence for travelers.

In determining what we wanted to visualize with the data, we also researched prior uses of the data, which include 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 goal is to provide a comprehensive interactive visualization tool that allows users to dynamically explore multiple dimensions of the data simultaneously. We believe that such a tool, incorporating geographic context and temporal patterns, will be valuable for all users.

Initial sketches: Our initial primary idea for visualization was a map of the US that displays average delay times per airport. Users can filter by specific airlines, time periods, or delay thresholds, with the map updating dynamically to reflect these selections.

Whenever the users adjust filters or select specific regions, our tool will automatically update several supplementary visualizations within the filtered parameters:

- **Plot 1: The most delayed arrival airports.** This bar chart will display the top 10 most delayed airports in the US within a period.
- **Plot 2: Average Delay by Time of Day and Day of Week.** This bar chart will display how delays vary throughout the day and across different days of the week.

Bring it all together, Figure 1 shows all the supplementary visualizations we wanted to present to the user, together with an interactive map. The first two figures depict the top 10 most delayed arrival and departure airports, while the latter two illustrate the average delay by time of day and day of week.

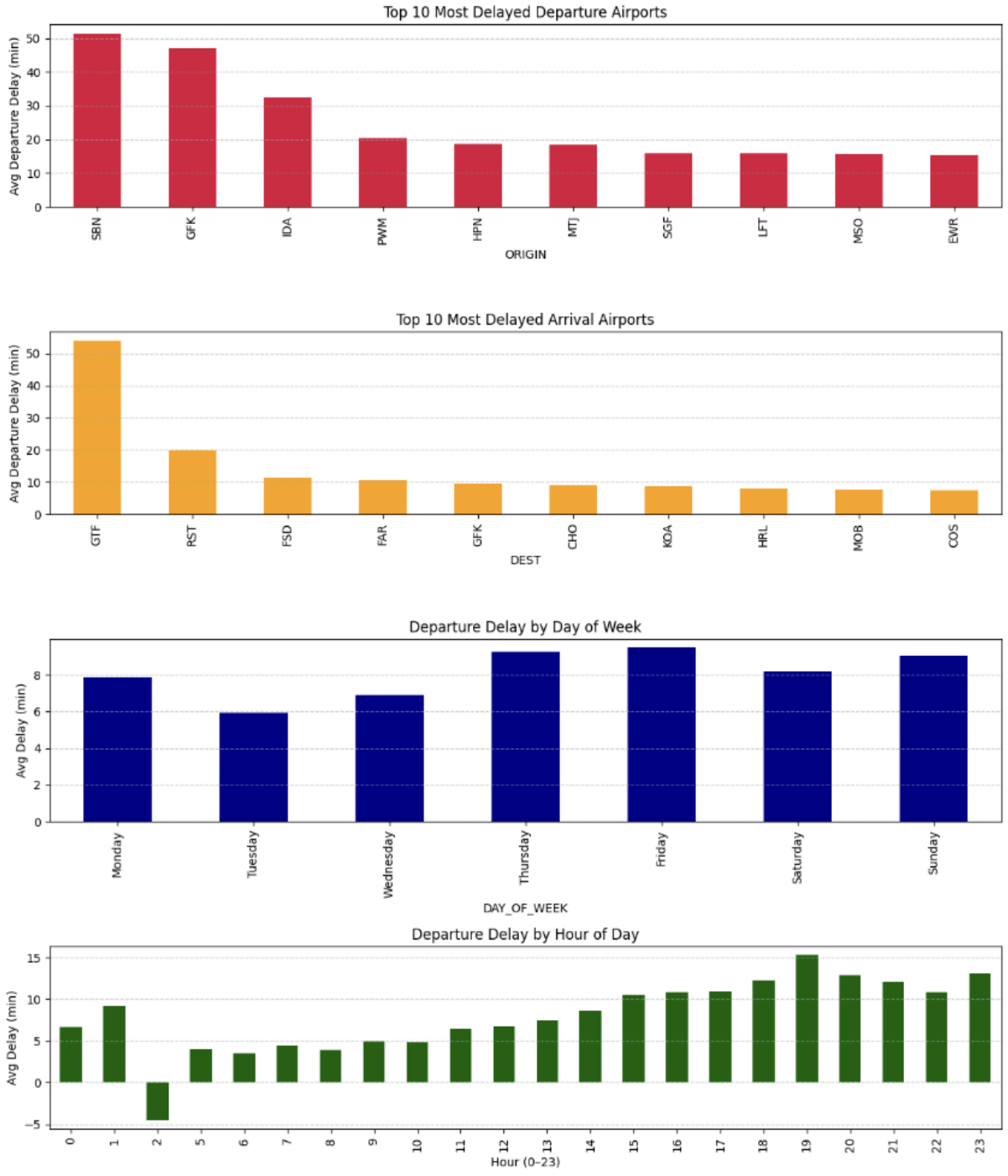

Figure 1: Sketch of Plot 1 and 2: Top 10 most delayed arrival airports; Top 10 most delayed departure airports; Departure delay by day of week; Departure delay by hour of day.

While working on the map component, we encountered the challenge of displaying the sheer volume of flight data without overwhelming the user or sacrificing performance. Given the tens of thousands of daily flights and over 5,000 public airports in the US, simply plotting every flight path or airport as a distinct point quickly led to severe clutter and unresponsiveness. The map would become a dense, unintelligible mess of overlapping lines and markers, making it impossible to discern any meaningful patterns or individual flight details. Furthermore, rendering so many individual elements dynamically resulted in significant performance bottlenecks, causing slow load times and sluggish interactions, particularly on less powerful devices. To overcome this, we shifted our approach to focus on plotting the airports only instead. This decision not only made the visualization more manageable, but also allowed for a more intuitive selection of data. Users can now choose their connection and destination selections geographically by hovering around the interactive map, facilitating a more natural and user-friendly experience.

Figure 2 shows our final iteration of the interactive map view of the project. Compared with our initial mockup and design, we removed some of the supplementary figures, given that the interactive map with color-coded and sized dots showing the delay range is self-evident. There is less need to further complicate users with more plots. We also include a delay filter in the upper-left region of the map to aid users in selecting the maximum delay that they would tolerate. Users can also zoom in/out geographically on the map to specified regions they are interested in for better visibility.

Figure 2: Final visualization of US-based flight delay

3. Technical Setup

We embed our visualizations in a website written using React. We use the Leaflet library for the interactive map and D3 for the accompanying graphs. For preprocessing and aggregating the raw flight data, we used Pandas, NumPy, GeoPandas, and scikit-learn. We use the preprocessed CSVs as the data backend for quick loading and lightweight interaction. For airport geolocation mapping, we used TopoJSON / GeoJSON. The website and database are deployed using GitHub Pages, which is accessible at https://com-480-data-visualization.github.io/com-480-team-kraska/.

The full dataset is too large (29 million rows) to visualize interactively, so we use a reduced version containing 10% of the flights on record (about 3 million rows sampled). The full dataset is available as a public dataset on US DoT statistics. The dataset contains flight

information between Aug 2019 and Aug 2023, covering pre- and post-pandemic periods. To accelerate the data backend serving, we preprocess the data to remove irrelevant columns, duplicates, and rows with missing values. Then, we aggregate the data by departure and arrival airports to accelerate the map view rendering and live adaptation. This significantly reduces the volume of data that needs to be queried and transferred, leading to faster response times for our visualizations and near-instantaneous front-end updates.

4. Distribution of Work

We met as a group several times and frequently communicated in a shared chat. In our meetings, we brainstormed the dataset we wanted to explore and the visualizations we wanted to make. After determining which visualizations we wanted to make, Ioanna and Yi sketched the mock-ups of the graphs, and Georgiy assembled them together to mock up the entire site. Georgiy, Ioanna, and Yi performed the initial data analysis. Georgiy built out the base of the website, focusing on the map and preprocessed integration, e.g., the query generation. Both Georgiy and Yi worked on the styling of the site. Ioanna generated the plots themselves. Ioanna and Yi did most of the write-ups for the milestones, but all members contributed.

5. Screencast

The screencast can be found in the following link: https://drive.google.com/file/d/1YFtQCGfewm6rL_OiIaplZ6HWOnbkdXg1/view?usp=sharing

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