Stephen Malacaria

Flight Data Project

Dataset

The dataset I chose can be found at https://www.kaggle.com/datasets/usdot/flight-delays?select=flights.csv . The dataset contains three tables: the *airlines* table contains 14 rows, relating a two-letter airline code to the airline's name; the *airports* table contains 322 rows and contains data about each airport, including IATA code, coordinates, state, and name; and then *flights* table contains nearly 6 million rows and 31 columns tracking domestic flights in the US in 2015. Each flight has data about the scheduled departure and arrival times, any delays, any particular reasons for the delays, and of course the origin and destination airports.

Motivations / Goals

I initially wanted to explore a map of flights by airline and the different connections between various cities for each airline, exploring what cities were hubs for different airlines. But then my focus shifted to airline delays, and my goal was to show the delays experienced by each airline at a user-selected set of airports. My final goal was to allow users to select a group of airports on a map, and then some charts would update with statistics about those airports. The statistics they would see would be the portion of flights at those airports operated by each airline, the average arrival delay of outbound flights by each airline, and the average arrival delay of inbound flights by each airline. Rather than showing these statistics for all 14 airlines, I would show the statistics for the big three US airlines (United, Delta, and American), and then a fourth category grouping all Other airlines together.

Manipulating the Data

With nearly 6 million rows and 31 columns, the flights table was actually too large to load into d3 all at once. To solve this issue, I first trimmed away all of the columns that I would not use, leaving me with just four: airline, origin airport, destination airport, and arrival delay. However, I realized that processing real-time calculations over 6 million rows while a user was operating the visualization may have caused issues. So instead, I pre-process calculated the statistics for each airport in advance. This led to an additional 13 rows in the *airports* data table, for example: total flights, United flights, Delta flights, United inbound delay, Delta outbound delay. This greatly reduced the scale of live processing I would have to do, with summations for each of the statistics having to be calculated over, at most, 322 rows rather than nearly 6 million.

Future Work

I think the visual layout of the webpage could be improved. It doesn’t really function like a website as I’ve seen in others’ projects; it’s just a standalone tool. I could also add more thought to the colors I chose. In terms of functionality, there are a number of original goals that I did not accomplish, such as additional filtering tools besides the brush.