Flight Delays in the United States

Seara Chen
EECS
MIT
Cambridge, MA, USA
searac@mit.edu

Julia Fiksinski
EECS
MIT
Cambridge, MA, USA
fiksin@mit.edu

Margaret Wang
EECS
MIT
Cambridge, MA, USA
mxwang@mit.edu

***Abstract*—The onset of the COVID-19 pandemic rapidly brought the air transportation sector to a standstill in the spring of 2020. However, with vaccination rates rising, the number of flights nationwide for post-quarantine travel stands to grow as well. We sought to examine the air transportation sector's recent performance. We acquired data (dated from January 2016 to March 2021) regarding United States airline and airport on-time statistics and delay causes from the United States Department of Transportation's Bureau of Transportation Statistics website. The primary features we examined were geographic airport data, time-based flight delay data, and airline-based delay data. We created a scrollytelling presentation with numerous interactive visualizations and static charts to present these facets in a manner that encourages user exploration.**

I. Introduction

The air transportation sector is currently going through a period unlike any other. COVID-19 has rapidly brought the air transportation sector to a standstill with demand free-falling within a period of months. However, as vaccinations have begun to turn the tide on COVID-19 in the United States, this pattern will begin to turn. With post-quarantine travel, particularly air travel, reentering the collective minds of the United States, we revisit the air transportation sector's history of performance. We collected data regarding United States airline and airport on-time statistics and delay causes from the United States Department of Transportation's Bureau of Transportation Statistics website. Taking data spanning from January 2016 to March 2021, we examined several facets of the data and created a scrollytelling presentation to visualize them. First, geographic airport data, second, time-based flight delay data, and finally, airline-based delay data. The following sections will go into further detail on our work.

II. Related Work

The dataset used for this project comes from the Bureau of Transportation Statistics (BTS), which provides air carrier statistics for the U.S. Department of Transportation [1]. Alongside these statistics, BTS provides pie charts that visualize on-time arrival performance, categorize flight delays by cause, and elaborate on the specific types of delays within each delay category. We found the BTS visualizations useful for introducing viewers to the dataset, and we appreciated that the visualizations allow users to explore delay data for a particular air carrier, airport, or time period. We sought to expand upon this in our project not only by producing a wider variety of visualizations that focus on different aspects of the data, but also by allowing users to compare the delay statistics of multiple carriers.

When searching for examples of existing visualizations that allow for more user interactivity, we encountered FlightAware [2]. FlightAware manages a flight tracking and data platform, maintains a live heat map of flight delays and allows users to search by airport or flight number. Though we ultimately chose to use point size instead of color intensity as the encoding for the number of arrivals and delays in our map visualization, we found this webpage influential in both its visualization design and in its user-focused functionality.

As it did with most industries, the COVID-19 pandemic took a heavy toll on air travel. Although the project described in this paper focuses on air carrier delays in general and not on air travel as it relates to COVID-19, we examined visualizations depicting the impact of COVID-19 on the airline industry. One of the more influential articles we found was a scrollytelling visualization by The Washington Post that depicted the global air travel shutdown that took place between January and June of 2020 [3]. The visualization plots airports and their connecting flights on a rotating globe. Each slide in the narrative focuses on a different continent and displays an animation of the plummeting global flight count over the five-month period. We originally sought to create a visualization with connecting routes similar to this for the geographic component of our project. However, because the BTS dataset we used did not include any data regarding the origin locations of arriving flights, we used airport location to convey geographic information. Instead, we drew more inspiration from the narrative aspect of this article, as we thought that the scrollytelling approach could help engage more users with our project.

III. Methods

The development process involved data exploration, data wrangling, and web-based visualization development using D3.

Due to the complexity of our selected dataset, we needed to be particularly mindful of how to share the different facets of the data without overwhelming users. Instead of creating a comprehensive dashboard for users to explore and find points of interest themselves, we wanted to explore an approach where the dataset is broken down into individual aspects. This could allow users to explore freely within each aspect to build an understanding of the dataset section by section.

For our project, we chose to examine U.S. air travel data from January 2016 to March 2021. The dataset includes information on the numbers of total and delayed flights per airline, airport, and months. We found during our preliminary data exploration that certain airlines were missing data for some years. For example, Virgin America only has entries in our dataset from 2016 to early 2018 because the airline fully merged with Alaska Airlines in April 2018. BTS reports that it receives on-time data from "marketing airline networks... and operating carriers that have 0.5% of total domestic scheduled-service passenger revenue" [4]. We wanted to include air carrier comparisons over time for multiple visualizations in our project, so we felt it would still be meaningful to use data from airlines with missing years in our dataset instead of removing these airlines from the dataset.

During the data wrangling component of our project, we processed the data in three ways: (1) the percentage of carrier flights that are delayed, (2) the expected delay duration per carrier, and (3) the occurrence of each flight delay type scaled by carrier flights. For all three calculations, we grouped the dataset by air carrier and date (and therefore aggregated over all airports). The first calculation divides the number of delayed flights of by the total number of flights in a given period. The second is calculated by dividing the total duration of delays (minutes) by the total number of flights. These first two calculations serve separate purposes of communicating (1) the frequency at which flights are delayed and (2) the expected delayed time. We separated the two calculations to ensure that airlines that have frequent delayed flights are not misrepresented as the worst airlines if their delay duration is largely trivial. The third calculation (used in the visualization of delay types over time) divides the number of flights delayed due to each of five BTS-defined causes by the total number of flights in that time period and then scales it to show the number of delays per 10,000 arriving flights. Scaling this result makes comparing different delay causes easier, as certain delays only occur a small fraction of the time as others. These calculations are performed in a Python script and in the visualization-generating JavaScript files (all available in our GitHub repository).

We used the D3.js JavaScript library to create all of our visualizations, and we used JavaScript to create the scrollytelling interface [5]. For our narrative, we wanted to guide users through the different visualizations to explore airline delays by location, time, and delay type before revealing the top five airlines with the highest percentage of flights delayed and with the longest delays. Based on the peer critiques we received, we decided to add a set of static visualizations and more observations to augment our narrative.

IV. Results

Our visualization contains three main sections: (I) airports, (II) delay causes over time, and (III) air carrier comparison.

I. Airports

The first section of our scrollytelling visualization highlights patterns in geographic airport data. This consists of two interactive maps and and a donut chart.

We begin by displaying an interactive bubble map of all arrivals since 2016. This visualization contains several features, such as bubbles that scale in size with number of arrivals, on-hover tool-tips that display more detailed information about each airport, and zooming and panning for each state. The purpose of this visualization was to provide context towards the relative sizes of different airports (measured by number of flights) and which airports may be major nodes in the the air transportation industry. A map visualization was a natural extension of this goal and a cohesive and visually clear representation of the data. Each bubble on the map represents an airport. We chose bubble size to encode airport size as an intuitive and easily comparable indication of scale. The bubbles are semi-transparent so as to show any overlapping bubbles and the state lines underneath. Hovering on a bubble reveals more information about the selected airport. The additional data is hidden in a tool-tip to both keep the visualization uncluttered and support a moving point of focus. For further ease of comparison, the map also includes a bubble scale for three orders of magnitude of flights.

The Albers USA map projection was carefully selected to highlight the contiguous United States, as that is where the bulk of the data is located. Because US territories are not included in this projection, we included an additional rectangle containing data for territories and placed it on the map next to other non-contiguous land such as Alaska and Hawaii. Finally, because major metropolises tend to have more airports, certain areas of the visualization tend to be more crowded. In response, we were careful to order all bubbles such that none would be inaccessible due to being covered by a larger one. Furthermore, added support for zooming and panning exists so that even in crowded areas or areas with smaller airports, the viewer can easily focus on what they want to know.

The donut chart is a simple static visualization meant to introduce and provide context for the final visualization. This chart plots the frequency of each cause of flight delays. It shares the same color encoding of the rest of the visualizations depicting flight delay causes, but the chart also has its own labels so that the viewer does not need to go searching for a legend elsewhere. We are only plotting five delay types so the labels do not clutter the visualization. A donut chart was also chosen over a pie chart because there are only a few delay categories. Removing the center of the pie chart draws the viewer's focus to comparison of arc lengths, a more efficient method of comparison here.

The final visualization in this category is the buildup of the previous two. This visualization is another interactive bubble map. It contains all of the features in the first bubble map, with the addition of delay cause color-coding. Each cause is displayed in a list on the side as a combination of selection buttons and legend. When the viewer selects a button, the map

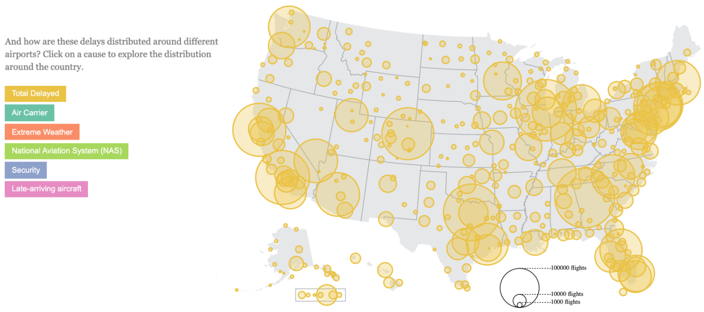

Fig. 1: Interactive bubble map of U.S. airports (section I). Size encodes number of delays and color encodes delay cause.

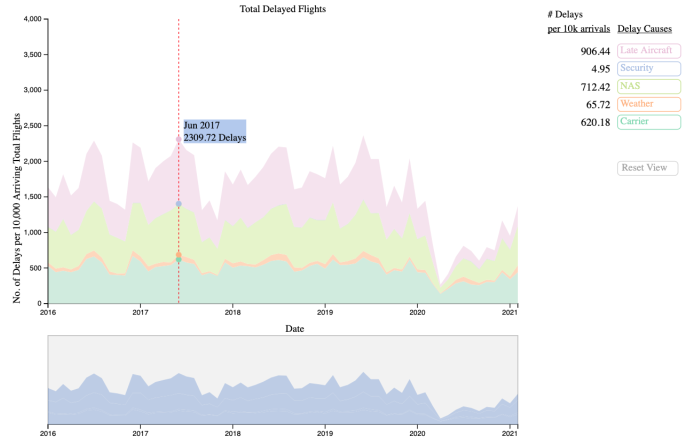

Fig. 2: Interactive stacked area chart displaying number of delays by cause from 2016-present (section II). Color encodes delay cause.

updates to show the corresponding delay data, and the bubbles change to match the color of the delay cause. In doing so, this visualization ties together the previous two in this section by displaying both the total number of delays and the number of flights delayed by each of the various causes.

II. Delay Causes over Time

The second section of our visualization explores change over time in the number of delays by cause. This consists of one interactive visualization, an animated clip, and a static chart. As mentioned in our methods section, we scaled the number of delays per 10,000 arriving flights for easier comparison of delay causes by frequency. As in the map and donut chart in section I, color encodes delay causes in this visualization. We considered presenting the number of delays by cause as lines instead of stacked areas, but we ultimately felt that a stacked area chart would more clearly convey the different proportions of delays by cause while still showing the total number of delays at each point in time. A drop-down element under the visualization caption on the left-hand side allows users to select any air carrier from the data, and the default view shows the total number of delays aggregated over all airlines. As with the individual airline calculations, the number of total delay flights is also scaled by the number of arrivals. The right-hand side of the visualization includes a legend for the delay causes; each entry in the legend serves as a selection button. Hovering over a particular cause's button highlights its related layer in the stacked area chart and dims the other layers, while clicking on a cause's button provides users with a new view: the individual area chart for that delay cause. This view hides other layers and moves the selected layer to start at the base of the y-axis so its height accurately conveys its number of delays. The individual view also changes the y-axis scale if necessary for delay causes that occur very infrequently (notably, security and weather delays). The legend also includes a reset button to let users return to the stacked view and view the full time period.

This chart also includes a vertical line that moves on hover and includes a tooltip that highlights the total number of delays at a moused-over point in time. While the tooltip shows the total number of delays, the counts of each individual cause appear to the left of their corresponding selection buttons in the legend. We added this feature in response to peer feedback, as it allows for a precise comparison of delay causes and counts at every point in time on the chart.

Users may also brush over part of the area chart to zoom into a particular time or double-click to reset the time scale. A "context map" view under the chart shows a miniature version of the area representing the total number of delays and includes a shaded rectangle that updates to highlight the time period in view as a reminder to users. If a user selects a different air carrier, the chart retains the current zoom (unless users double-click or click the reset button) to let users compare specific delay counts between airlines in a narrower time period. By offering multiple layers of interaction in this visualization, we enable users to more deeply explore not only how frequently delays occur due to different causes, but also how air carrier performance varies with respect to these causes.

The next component of the section is an animated clip of the interactive chart with a focus on January 2019 to December 2020. In this GIF, the vertical line and tooltip move across each month to highlight just how much the number of delayed flights (and relatedly, the number of total arrivals) plummeted at the start of the COVID-19 pandemic.

The final component in the section shows a static line graph of the total delayed flights per 10,000 arrivals each month for the years 2016 through 2020, with each line representing a different year (and color-coded as such). We aggregated the total number of delays over all airlines and airports for this chart. Interestingly, the graph also shows the uptick in flight delays (implying more flights) starting after September 2020. We added both the GIF and this static chart to give users a closer look at just how much air travel was affected by COVID-19. As mentioned earlier, while this was not the focus of our visualization, we felt it necessary to touch upon such an impactful situation.

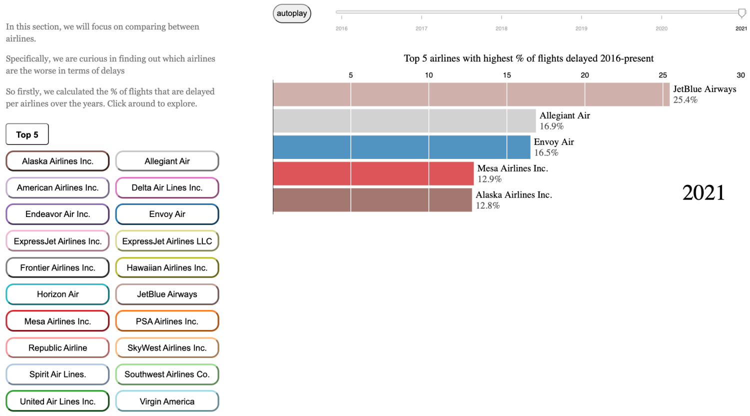

Fig. 3: Interactive bar chart depicting yearly airline rankings based on highest percentage of flights delayed (section III). Color encodes airline.

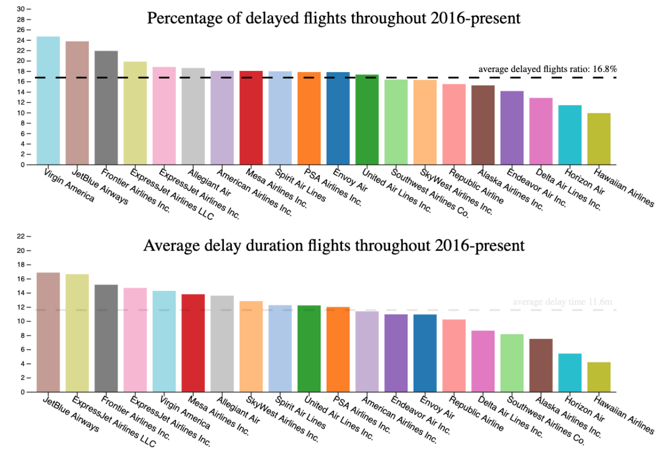

Fig. 4: Bar charts showing air carrier rankings throughout 2016-present by percentage of flights delayed and average delay duration (section III).

III. Air Carrier Comparison

The third section consists of three interactive bar charts, each of which depicts airline rankings based on their delay data. The purpose of this section is to address the question of which airlines are the worst with respect to flight delays. Given that the dataset contains information pertaining to 20 air carriers, we considered that this could be a lot of information for viewers to digest. We wanted to lighten the viewers' burden by using a simple mark that viewers would already be familiar with. We opted for a bar chart, as it is a classic mark for categorical comparison.

We chose color as the encoding channel to represent different air carriers because colors are classically used to differentiate categories. We explored the alternative of using airline logos instead to represent airlines, but some airlines' logos are so similar in color that they are hard to differentiate at a glance. Using logos also meant that users would sometimes need to read the text in the logo, which becomes complicated when an airline's logo is cut off due to the length of its corresponding bar.

The first two visualizations in this section show the changes in airline rankings between 2016 and 2021. We opted to put the bars in a horizontal fashion to emphasize the idea of rank and placed the airlines with the highest percentage of delays or longest delay durations on top. An interactive timeline allows users to select which year's rankings to view. An autoplay option also exists to automate the transition of the ranked bars. We hope that the animation of swapping bars' orders highlights the "rank" concept of the chart. It accentuates which airlines took over other airlines for the "worst" status in terms of delays.

On the left side of these ranking charts, we first offer the "Top 5" button to show the top 5 carriers with the worst delay ratios and durations. This is the default view, as it directly answers this section's primary question of which airlines are the worst when it comes to being on time. Each air carrier has a corresponding selection button; clicking an airline's button places its bar in the ranking view. This feature allows users to directly compare airlines that are important to them. Especially considering that the top 5 worst airlines are not necessarily well-known, we wanted a channel to directly introduce information that is relevant to the viewers. The need of this feature is also confirmed by the fact that in our peer review feedback, this feature was explicitly suggested.

In the last visualization of this section, we hope to give viewers a summary of all the airlines across the years, so we aggregated data over all years. The visualization contains two stacked charts: the upper chart ranks airlines by percentage of delayed flights, and the lower ranks airlines by average delay duration. We stacked the charts as opposed to putting them side-by-side to allow comparison between the results of the two calculations. Both charts use vertical bars sorted by value in order to display all 20 airlines together. We also introduce a reference line that represents the average value to further solidify the sense of comparison between airlines. This reference line also signals to viewers the above/below-average carriers. While creating this visualization, we discovered something interesting in this dataset: carriers that have higher percentage of flights delayed tend to have longer expected duration of delay as well.

V. Discussion

We presented the final project to a small group of people to receive feedback on our visualizations, interactions, and project design. These users stated that the interactions of the clickable areas (specifically, selection buttons of delay causes and air carriers) linked to the visualizations were satisfying to use. They also reported that they particularly enjoyed the amount of freedom they had with exploring different airports in the map visualizations (section I) and selecting various air carriers in the time and ranking visualizations (sections II and III). Additionally, users stated that they enjoyed the personal connection of viewing data for airports and airlines with which they were familiar. This positive feedback helps reinforce our design choices of creating multiple visualizations with numerous interactions that let users freely explore the various facets of dataset.

VI. Future Work

In many ways, the most interesting time period for this dataset is about to happen. Air travel has been heavily impacted by COVID-19, and as vaccinations lead to a lessening on social distancing precautions, air travel will become an interesting indicator. The airline industry will be much smaller for years to come, but there may also be other interesting changes. Will there be a sudden jump in flights for leisure, leading to more delays with the reduced number of airlines? Will the confined space of an airplane cabin remain a strong enough deterrent that flights stay infrequent for a longer period of time?

As another extension of our existing project and to provide even more usability to the user in this post-pandemic world, we could add a live section to our visualization where in addition to projecting historical data, we show live hotspots of delays and the most up-to-date data on which airlines provide the most timely travel. This would extend our project from just being one of reflection to one that more directly facilitates action.

We hope to continue to update our visualization to further examine the evolution of the airline sector.

Acknowledgment

We would like to thank the course staff of 6.859 for all their work in teaching this class this semester and for their support in the execution of this project.

References

[1] Bureau of Transportation Statistics. *Airline On-Time Statistics and Delay Causes*. 2020. URL: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp/.

[2] FlightAware. *Live Flight Tracking*. Apr. 18, 2020. URL: https://flightaware.com/live/airport/delays.

[3] Anthony Faiola, Lauren Tierney, and William Neff. “The virus that shut down the world”. In: *The Washington Post* (June 26, 2020). URL: https://www.washingtonpost.com/graphics/2020/world/coronavirus-pandemic-globalization/.

[4] Bureau of Transportation Statistics. *Understanding the Reporting of Causes of Flight Delays and Cancellations*. 2020. URL: https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations.

[5] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. “D^3 Data-Driven Documents”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2301–2309. ISSN: 1077-2626. DOI: 10.1109/TVCG.2011.185. URL: https://doi.org/10.1109/TVCG.2011.185.