Traveler’s Board: Visualizing flight delays and causes with possible recommendations.

Krishna Vamsi Kurumaddali, Andrew Chan, Patrick Nguyen, Van-Phi Quy, Manjunath Gujjar

Abstract— In the aviation industry, our goal is to focus on specific airlines and their flight data. In particular, we are looking at average flight delays, flight routes, and number of flights. Our goal for this visualization is to aid the user in understanding which airline is best for them using a dashboard. Previous research shows that this topic has been explored including flight delay distributions, prediction models showing expected delays, and map representations of flight routes. It is important to explore this area of study because with new insights users will be able to better optimize their air travel by choosing better airlines that are more reliable or have less delays for their use case. The primary focus of this project revolves around visualizing statistics of airline and airport delays. The main finding for our data is that most airlines have an average delay time of around 24 minutes, the East Coast has more delays than the West, and that in the summer people tend to travel from north to south. Using our data users can discover flight patterns better which can help them plan their flights.

Index Terms— Flight Delays, Data Visualization, Map Visualization.

1 Introduction

Our visualization project is centered around the aviation industry. Airline data is a crucial part of providing insight into various aspects such as flight routes, flight delays, and flight amount. Users of the project will be general travelers who are interested in using an airplane as their method of travel. By using flight data provided by the Department of Transportation from 2015 [6],[7] we have created a tool filled with visualizations focused primarily on flight delays. In general, our visualization will help users understand more about the airlines they use and can aid in developing strategies to find the best times to fly. The tasks we have outlined for the user is that they should be able to identify which airlines are most busy and also discover each airline's average delay to other airlines.

Given the available visualizations regarding this subject, users aren’t able to view multiple aspects of flight data succinctly. We have found that with the abundance of flight data, there are too many existing visualizations that suffer from overloading users with cluttered and dense maps (See Figure 9), graphs, and other idioms. To combat this, we opted to create visualizations that have clear focus without extraneous information. Creating a set of visualizations that adhere to expressiveness and effectiveness principles can be difficult. We want to avoid misrepresenting data according to their semantics and avoid creating ineffective visualizations by overloading them with information. Our work provides different representations of flight delays which easily allow any user to identify and compare which airlines and airport delays are going to be most important and relevant to themselves.

2 Related Works

Wang’s paper [1] attempts to assess the elements that contribute to flight delays and assist corporations in better planning to mitigate the effects; nevertheless, we may use this paper to offer customers with the flight conditions of each airline and the causes for delays in order to filter the best airlines for their needs. The data used in this paper, includes a variety of elements, including weather, physical faults, delayed timings, and crew-related problems. In this study, airline company correlations to delay and delay rate components are given a strength rating and represented in a bar graph. In fact, all of the visualizations displayed in Wang’s paper are limited to bar graphs. In Figure 1, Wang’s paper presents different airline carriers according to their delayed flights ratios. Although we feel that it is important to represent the percent amount of delayed flights each airline has, it is not as effective as some of our visualizations. Our implementation, presented in Figure 10, does something similar but instead of displaying the percentage of delays via a bar graph, we present the average delay time for that flight. Figure 10 is more effective in representing data that is relevant to a typical user. In addition, our visualization is presented in ascending order from top to bottom, which lessens the cognitive load on our users. According to this study, airline companies, departure time, and airline delay are significantly associated, which can be a big interest for choosing flights.

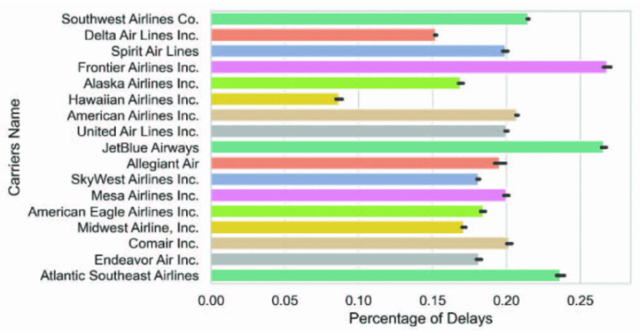
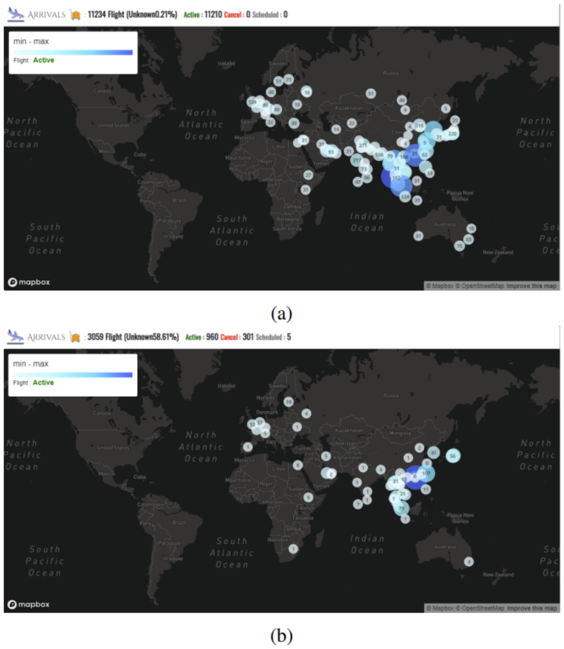


Fig. 1. Monthly airline delay counts [1].

Khamlae’s paper [2] describes the architecture and implementation of a web based interactive visualization of flight data over geographic space. In particular, the visualization allows users to see a summary of flight arrivals and departures at different points in time for different airports. This was done by using bubble maps to encode flight volume at airport locations. In addition, the application supports choropleth maps where flight arrival density is encoded according to different colored bins. We are going to follow a similar approach but encode two attributes for each airport. The two attributes we are encoding are the average delay and number of flights by airport. See Figure 13 for our implementation. We have found that encoding only one attribute and representing it via color does not do enough to achieve our user tasks. Encoding the average delay alongside the number of flights gives proper context to the expected average delay per airport. We extend and expand on the original idea that was discussed in Khamlae’s paper. Using a map will give geographical context to the popular airports and flights throughout the country which is necessary if we want to visualize data as a summary.


ARRIVALS
11234 Flight (Unknown0.21%) Active : 11210 Cancel : 0 Scheduled : 0
min - max
Flight : Active
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
mapbox
© Mapbox © OpenStreetMap Improve this map
3059 Flight (Unknown58.61%) Active : 960 Cancel : 301 Scheduled : 5

(a)

(b)

Fig. 2. Bubble map of arrival flight density at Bangkok Suvarnabhumi Airport, January vs May 2020 (a) Arrival flights in January (b) arrival flights in May [2]

Alternative approaches for analyzing and displaying flight data, such as machine learning, have been attempted in other works [3]. Liu presents an overview of applying machine learning on flight data for flight delay prediction. The data visualization presented in this paper ranges from showcasing flight delay distribution through a histogram graph, bar charts, scatter plots (complex). But the visualization which stands out is the treemap visualized for the flight delays and number of flights in different United States. Through this paper, we can observe some of the issues that arise in shoving data into an idiom and deeming it effective. A treemap is intended to make outliers in the data popout. Although it is doing its intended job by highlighting the states with the most average arrival delays, it suffers due to the lack of clarity. Treemaps suffer for our use case as we want users to be able to identify their specific state and compare against other states' flight delays. Treemaps are not effective at making individual comparisons. We opted instead to use a geographic heat map showing average flight delays by state (see Figure 7.1). This is much more effective as it not only provides geographic context for our users but also makes visual comparisons much easier and more organized. It is just as easy to see the outliers and visual popout without compromising visual clarity.

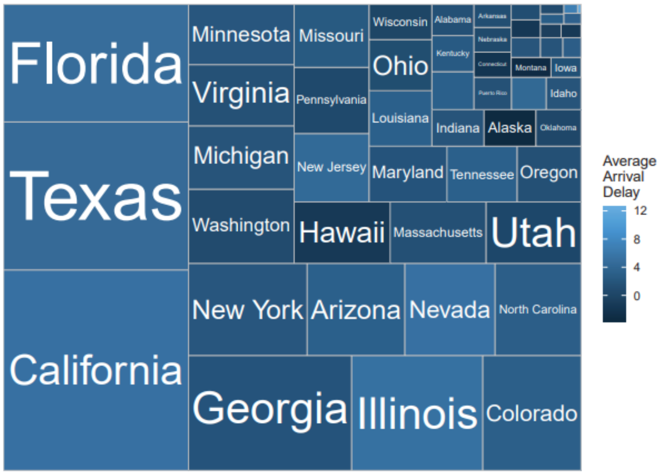


Fig. 3. Flight delay and number of flights trends by state [3].

Further previous studies [4] describes two types of flight data vis: Local, low-moderate data sizes used for decision making, and high level exploratory vis that aims to get as much data shown as possible. In terms of exploratory vis, the article explains multiple ways to show information in clearer and more useful ways. For example, instead of showing exact flight paths, they can be simplified into singular lines, with color or density representing how much traffic flows between airports. Flight ranges can also be emphasized by decreasing occlusion, which reduces hotspots for multiple flights. Figure 4 presents an approach to simplify and encode information wisely so that flight visualization channels are discriminable and separable. While we did not directly implement the connection of flights in our visualization tool, we instead took this philosophy and applied it to Figure 13. Although we used two channels with some interference (color and size) we understand the drawbacks of such an approach. However, compared to other combinations of encoding (e.g. width and height) we opted for our implementation instead.

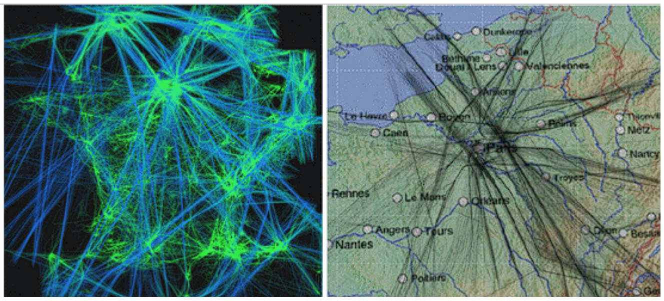


Fig .4. A map of flights over France on July 5th, 2006. The flights are color-coded by height, with higher altitudes represented by darker colors. (b) A zoomed-in view of the Paris area [4].

Our implementation of the visualization tool uses static data as its baseline. All the analysis and visual representation of data originates from that source. However, there are attempts to take real time streams of flight data and do analysis on that. Aljubairy presents his paper discussing the architecture of their system and some visualizations that they can present including boxplots, heatmaps, and matrices of flight delays present in China. Figure 5.1 shows a heatmap displaying the amount of delays present at airports in China. The visualization lacks clarity as it is missing any form of labeling or legend to identify the visual attributes on display. We approached our heatmap differently by observing average flight delays across different states as the airport density in the US is much more than that of China. In addition, figure 5.2 displays the delays and departure performance for airports using a box plot to display the number of delays per minute for different airlines. This visualization is not effective as the number of outliers contained within the data is considerably dense. When the quartiles are indiscriminable it is better to use a differing idiom. Although we do not display the same data, Figures 10 and 12 characterize flight delays by different airlines in a more comprehensive manner. The bins are discriminable and easy to make comparisons while still being able to observe outliers.

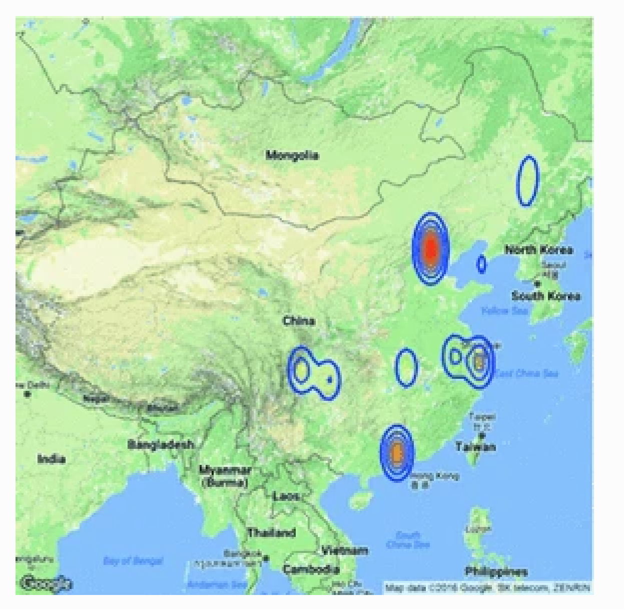


Fig. 5.1 Heat map visualizing the amount of delay at each airport [5].

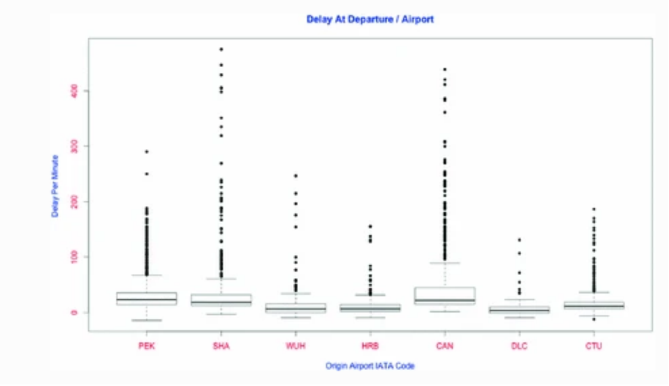


Fig. 5.2. Boxplot showing delay at departure performance for airports [5]

3 Implementation

Our project's main goal is to display visualizations made with tableau to convey aviation statistics. The main way our users will see the visualizations is through a dashboard that contains different visualizations. Some visualization examples can be seen in Figure 6,7,9,10,11,12,13 and 14. For the creation of our project we went through different steps to bring all the information together. These steps include curating data, creating visualizations, and creating the dashboard.

3.1 Linking Datasets

Our main data source for this project is a Kaggle Data set using Flights.csv [6] and three other CSVs [7]. The specific attributes that we used from the data for many charts are: Date of flight, Airline, Origin Airport, Destination Airport, Scheduled Departure, Departure Time, Departure Delay, Air Time, Scheduled arrival, arrival time, arrival delay, diverted, and canceled.

3.2 Creating the Visualizations

There are 3 data files for each flight data, airline company data and airport data with their location [6]. For the best use of these datasets to make better visualizations, we have connected all these datasets with specific attributes. Flight and airline datasets are connected with the airline codes given by the flight details within the airlines dataset. The landing and departure information is taken from the flight data set and linked with the airport dataset using the airport name and collected the location details to use in representing the flight in geographical maps.

3.3 Creating the Dashboard

During rapid prototyping, we used these datasets, and the attributes that we had already created and tagged to experiment with a whole bunch of possible visualizations using tableau. Then, from these early designs, we cut and removed any that were too cluttered, hard to read, or could be implemented in a completely different style.

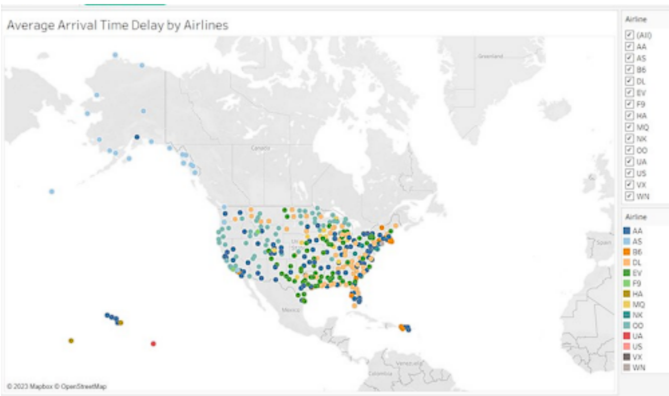


Fig. 6 An early alpha release prototype attempting to show both airport locations and which airlines used them. Was ultimately removed for being able to show one airline per airport at a time.

From the ones remaining, we started analyzing them to see which ways were best to improve them, while also making sure to remember the core task behind each vis.

3.4 Data Specification

The last part of our project should consist of making a skeleton dashboard that developers will use to add pages that contain our visualizations. The Dashboard will be made in tableau and have a menu bar on top that allows navigation through the pages. After our dashboard is created we can add our pages that can display visualizations. To start we want to add a homepage that describes what our data is and what purpose it serves. We will also include our sources here. This will be our page with the most complicated visualization that users will be able to interact with and learn all about airline delays throughout the U.S. The last page, the chart page allows the user to navigate through a lot of data contained in different visualizations. Ex: line chart to show airport delays over time, contour map/dot map to show most used airports, heatmap, and beeswarm chart.

3.5 Capabilities

The user can navigate the website with a mouse. The user interacts with the charts through the mouse. The user cannot upload or modify the data. Data is streaming to the front-end. The purpose of putting this data together is so the user can:

- Users can {identify, outliers} (Identify airlines/airports that are most busy, most on time).
- Users can {discover, distribution} of attributes (see at a glance average delays of airports).
- Users can {compare, features} of airports/airlines.

4 Results

Our project sought to inform a user about different aviation statistics and we have successfully done this by creating different visualizations. When looking at our dashboards a user is able to clearly see which U.S states and airlines have the longest delay times, and even when they occur. Each visual is unique and is able to be interacted with by the user. Below we have outlined the visualizations we have made along with a description.

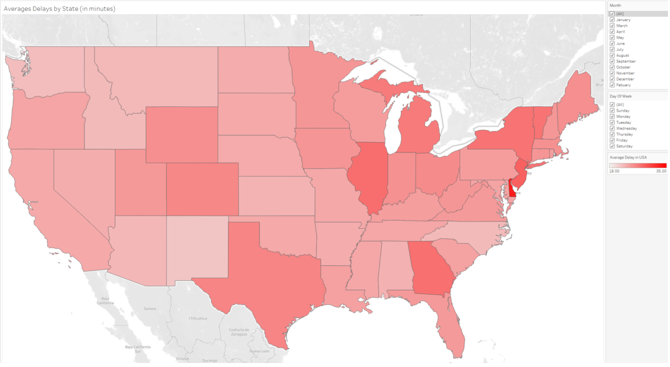


Fig. 7.1 Average Delays by State (in minutes)

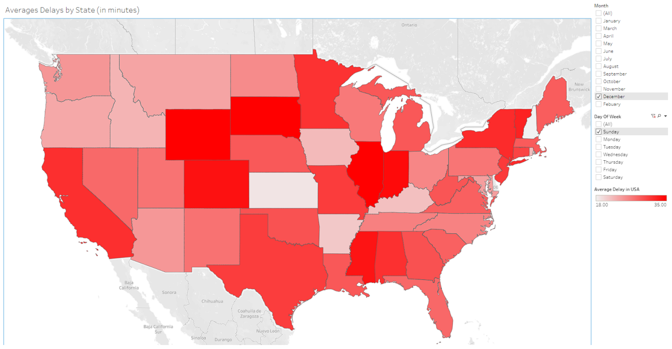


Fig. 7.2 Average Delays by State (in minutes), set to only Sundays in December.

Using color to effectively display data and trends was a major consideration on this specific idiom. One could ignore color,but that results in very bland displays such as shown in Figure 8. The user has to read every state's average delay, with no easy way to compare at a glance and see which states are better or worse. Thus, we can try to color the map, possibly with 2 colors to show better or worse off, such as in Figure 9. This is much better than the pure white map. However, green implies "good" while red implies "bad". In light of this, perhaps the whole map should be in shades of red. After all, even the best states have delays of around 18 minutes on average, which still should not be considered good. Thus we come to Figure 7.1 and 7.2, which we have chosen for our final version. It displays the states in shades of red to signify that, while every state has delays, there are worse states and better states. While it is a cultural concept that red means bad, it is a universal enough concept that we feel comfortable using it. Furthermore, using a monochrome color scale lets us show delays between states using luminance and saturation instead of hue. Hue is, physically, not a feature of color that lends itself to comparisons; is green "higher" or "lower" than red?

Furthermore, we have also added filters for month and day of the year. Using these, we can let the user explore seasons and compare those as well. The day of the week can be useful to determine when the user would prefer to fly in order to avoid delays. Figure 7.2 is a very good representation of this. In December, Sundays (and weekends in general) have much bigger delays than weekdays, suggesting to the user that it may be better to avoid Sunday as travel day. We have elected not to bin the colors into groupings, and instead let every state essentially have an individual color. Thus, when filtering, it becomes more obvious when a state changes color, while still allowing large changes in delays between states to be seen due to the large contrast.

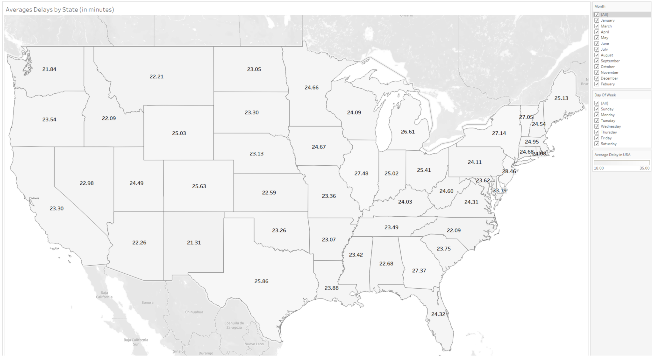


Fig. 8. Average Delays by State, with no color.

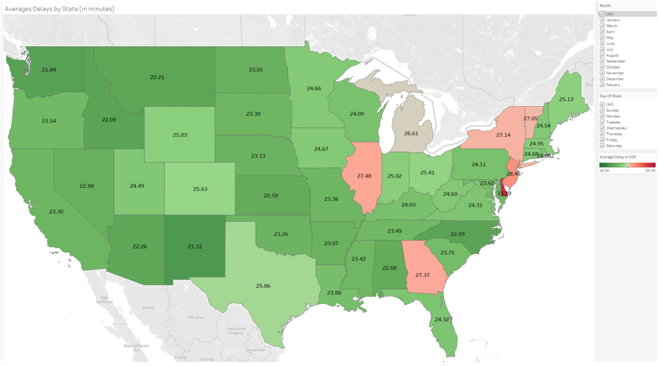


Fig. 9. Average flight Delays by State,with 2 color shades.

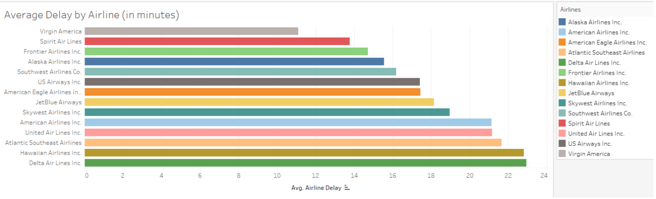


Fig. 10. Average delay by airline, calculated by summing the delays and dividing by the total number of flights per airline.

Figure 10 allows users to compare the average delay time against the other airlines, which in turn allows the user to know which airline is most likely to have a longer delay. From a quick view the user is able to distinguish between the least delayed flights and most frequent, however when a user selects a specific airline they are able to see the exact time each flight is delayed as well. The data itself is taken from 3 datasets, using flight keys to match airlines and other relevant data to eventually accumulate the delay times for each airline. Once we have the sum we can divide by the total number of flights for that airline to get the average delay time. Figure 10 is effective in allowing the user to quickly and easily see which airlines take the longest delays, some improvements could be adding a ranking.

Figure 10 uses color and horizontal length as its visual channels. The color assigned to each bar represents the specific airline associated with the delay. The length of the bars for each airline, once sorted for delay time, lets the user quickly compare in an efficient manner.. Overall, the graph is very simple and to the point, some trade offs of this is that the scope of our data representation is limited. If we wanted to include more data the graph could become more cluttered and we would lose the specific information we want to show.

Originally, to represent the average delays by airline we had used a bubble plot. One thing that was limiting in the bubble visualization was the size channel. The average delay times are not very distributed and more concentrated around a central number, resulting in it being hard to discern the difference between different bubbles. The juxtaposition of the bars in a bar chart is more intuitive than the bubbles. The cognitive load placed upon the user is much less due to the fact that spatially the bars are more organized. Although the bubbles might be a more aesthetically pleasing design it is an inferior choice for our visualizations. With all these drawbacks and trade-offs of a bubble plot specific to our data focus, we decided against the bubble chart and to run with a simple bar chart for this specific visualization in Figure 10.

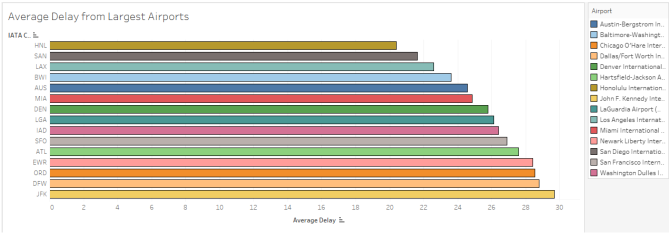


Fig. 11. Average flight delays by largest airport.

Figure 11 is similar to Figure 10 which allows users to observe the average delay time but between the largest airports in the US instead. For a user to observe the entire story regarding delays in flights it is important to not only consider the airline they might go with, but the airport that said airlines are departing from.

The linkage of the data and accumulation of delay time follows the same process stated in Figure 10. The difference being the attribute we are grouping by, instead of accumulating delay times between airlines, we are doing so by the airports. Using the raw data alone, it was extremely hard to see which airports were busiest and their delays. To determine which airports are considered the "largest", we observed the departure activity for each airport. The airports with the most departures were deemed as more active than the others, and a separate filter was created to filter our data to only the top 15. From there, we could calculate the average delay per airport, similar to Figure 10.

Calculating this new attribute to our airports creates an objective measure to rank airports. This improves the match between data semantics and the task we want the user to accomplish.

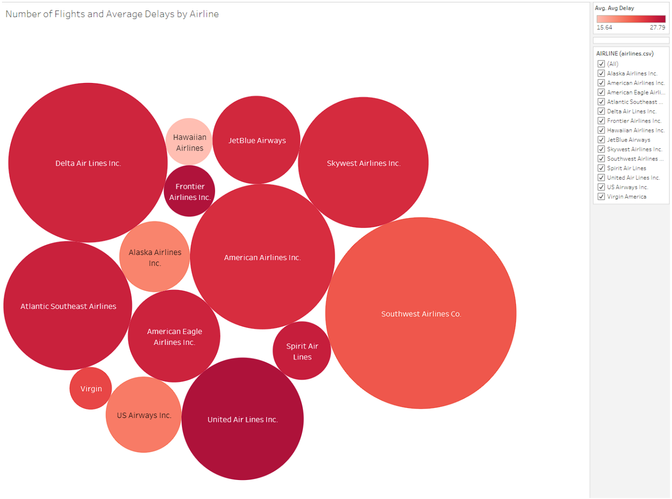


Fig. 12. Number of flights operated taken as size, Average Overall Delays taken as color component with a White to Red color scale. The vis also has interactive filters for Airlines with specific information based on number of flights and overall average delay.

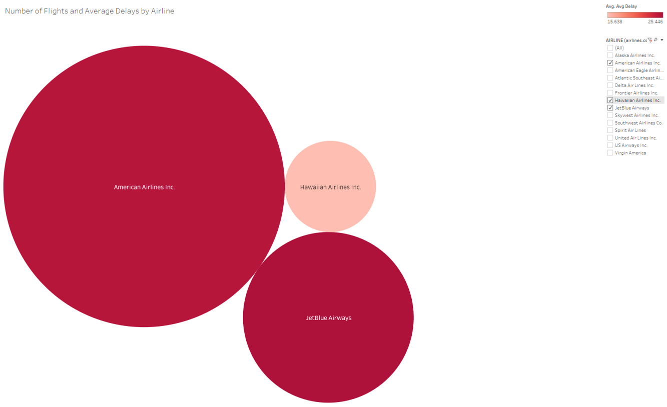


Fig. 13. Number of flights operated taken as size, Average Overall Delays taken as color component with a White to Red color scale. The vis also has interactive filters for Airlines with specific information based on number of flights and overall average delay.

Figure 12 displays the number of flights operated by each airline with overall delay encoded as color, with darker reds meaning longer average delays. We chose to use a bubble plot for our preferred vis with these specific attributes, after experimenting with bubble plots for Fig 10. The much larger variability in the number of flights from each airline makes size a much more readable factor when encoded as area, as some of these airlines run many times more flights than others, while average delays are in a much more narrow band of numbers. The users can conclude each airline's average overall delay by looking at the color and the amount of flights they operate per year with the size of the bubble. A major feature of this bubble plot is the fact the user can filter the data by the airlines of their choice, as seen in Figure. 13, to more directly compare as few as two airlines. A user can quickly identify that the smaller airline, Hawaiian, has a much smaller average delay than the larger carriers. This visualization is at its best at a glance; the user can quickly identify the highest or lowest delay airlines, as well as which carriers are large or small, while comparing the two channels. While this bubble plot is very interesting, there are some drawbacks to this idiom. The most glaring feature is the difficulty in displaying more detailed comparisons. Spirit Airlines is obviously much smaller than Delta Airlines, but is it bigger or smaller than Hawaiian in Figure. 12? By how much?

From the previous iterations of the same bubble plot, what makes Figure 12 stand out is that it is not based on one single attribute which was average airline delay but also lets the user generally compare and make conclusions based on two derivations of data. The previous display, only airline average delays, is much better shown as a quick bar graph like in Figure 10. In analyzing how effective this visualization is, we have to talk about clarity. Both channels are visually distinct; the size of a bubble and the color of a bubble have some interference. However, in our case, we feel that there are not too many marks at once, letting them still be effective channels. A small dark red bubble and a small pale bubble are easy to tell apart, and the same is true for large ones. No information is hard to read at a glance, making our choice of channels effective.

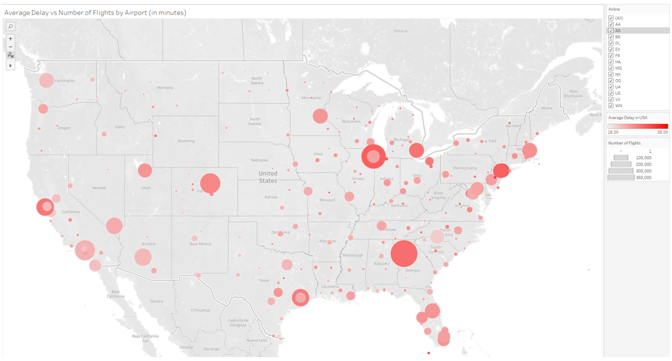


Fig. 14. Airport locations, with the color of the airport representing the average delay time and the size of the airport representing the number of flights in a given year.

As shown in Figure 14, we were trying to have marks that showed 3 different attributes; airport locations, number of flights, and average delays at the airport. Using spatial positioning to show locations was an easy choice, as they could be shown on a map of the US. Taking inspiration from Figure 7, we let color/saturation represent the average delay at the airport, while the size of the mark represents how busy the airport is. While color and size can interfere with each other, we found that a large, pale mark is visually distinct from a large, red mark and a small pale mark. Thus, our chosen encodings worked well for what we wanted to display. The biggest visual issue we have are airports that are extremely close together. Around LAX (Los Angeles International), there are 2 smaller airports that are inside the larger mark, occluded by LAX's mark and making them hard to see, especially because their delays aren't that much different from LAX. To help alleviate this, when zooming into the visualization, the marks stay the same size relative to the webpage, while the map zooms in. That way, a user can zoom in to get a better look at airports close to each other, so that airports that occlude each other in a larger view get visually separated. However, this still makes it hard to compare these smaller airports with those in another region of the U.S. entirely.

To let the user explore the data further, we added filters for specific airlines. Using this, the user can see which airports are most often used by major airlines, as well as let us see if different airlines using the same airport have significantly different delays.

Overall, this visualization is easy to read and lets the user quickly compare large and busy airports. However, there is no easy way to search for specific airports. Unless the user knows exactly where the airports they are looking for are, searching for a tiny airport among hundreds is not an easy feat. Furthermore, a detailed analysis in the differences between 2 or more airports is hard as well. The user would have to hover over every mark to get the exact average delay time and number of flights from the airport. The airline filter feature has a similar problem; There is no way to get 2 different maps of airlines side by side, instead forcing the user to switch between them constantly to get the data they want.

5 Future Directions

Currently, the visualizations presented in this paper (Figures 7-13) place a great focus on average airline delay times. We can explore the other measures of central tendency to give full context for the flight data we are given. No one measure of central tendency is the best so it might be helpful to juxtapose visualizations between the 3 measures in a pane view for a more well rounded visualization. Our primary stretch goal for this project was to upload our dataset into a database so that data queries and data mutations will be optimized and stored in one central location. We have kept everything within the confines of Tableau due to the ease of access in implementation. The processing of the data is quite slow. In the event that we were interested in implementing interaction with the data, such as filtering and highlighting regional data, the latency would be detrimental to the user experience. A possible improvement upon our data analysis could be to try out different methods. As mentioned in Jiang's paper [3] they had used machine learning of aviation data to predict flight delays. I believe that with the right methods, we could expand our visualization tool to incorporate data produced by predictive machine learning algorithms. In addition, we can include correlation analysis between the causes of delays. This is an example of an aspect of our flight data that we had yet to explore. Ideally, our tool should use the most recently released Department of Transportation data. A stretch goal would be to fetch the new yearly data and upload it to our public tableau and generate new and up to date visualizations.

6 Conclusion

In conclusion, this paper showcases how visualizations can aid travelers with recommendations by presenting them with flight delays, cancellations, filters based on time which is in days, weeks and months. This paper addresses some of the limitations which can be encountered in conventional dashboards by interactive dashboards which are informative. Visualizations are based on solving real world problems faced in airports like showing delays based on average, per airline, per state in the US. Therefore, this paper highlights the potential visualizations which improves the travel experience for passengers by providing necessary information to make informed personal decisions during their journey.

References

[1] H. Wang, "Big Data Visualization and Analysis of Various Factors Contributing to Airline Delay in the United States," *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, Sanya, China, 2022, pp. 177-181, doi: 10.1109/BDICN55575.2022.00042.

[2] P. Khamlae, C. Nimpattanavong, W. Choensawat and K. Sookhanaphibarn, "Visualization System for Air Traffic data," *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, Kobe, Japan, 2020, pp. 213-214, doi: 10.1109/GCCE50665.2020.9291971.

[3] Y. Jiang, Y. Liu, D. Liu and H. Song, "Applying Machine Learning to Aviation Big Data for Flight Delay Prediction," *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Calgary, AB, Canada, 2020, pp. 665-672, doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00114.

[4] T. Klein, M. van der Zwan and A. Telea, "Dynamic multiscale visualization of flight data," *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, Lisbon, Portugal, 2014, pp. 104-114.

[5] Aljubairy, A., Shemshadi, A., Sheng, Q.Z., "Real-Time Investigation of Flight Delays Based on the Internet of Things Data." *12th International Conference on Advanced Data Mining and Applications*, Gold Coast, QLD, 2016, pp. 788-800

[6] R. Bell, "flights.csv". figshare, 12-Sep-2019, doi: 10.6084/m9.figshare.9820139.v1.

[7] D. of T. Statistics, 2015, "2015 Flight Delays and Cancellations," Kaggle, 09-Feb-2017. Online. Available: https://www.kaggle.com/datasets/usdot/flight-delays.