INF552 Data Visualization - Flight Delays Visualization Project
`https://remi.gr3z.eu/inf552.php`

Rémi Grzeczkowicz
École Polytechnique
remi.grzeczkowicz@polytechnique.edu

Ines Yaici
École Polytechnique
ines.yaici@polytechnique.edu

Abstract

This report introduces a visualization project focused on analyzing and understanding patterns and causes of flight delays in the aviation industry. Utilizing a dataset from the U.S. Department of Transportation's Bureau of Transportation Statistics, the project aims to provide valuable insights to stakeholders, including airlines, airports, and regulators, for developing strategies to reduce delays and enhance the passenger experience but also customers to help them select the best airline for their flight.

The technical setup involves a web server and MySQL Database, allowing for efficient data querying through SQL structures. Visualization tools such as Leaflet, D3, and Vega-Lite are employed to create intuitive and informative representations of flight delay data. The project offers a dual perspective through an Airport View and Flight View, providing users with interactive and dynamic insights into airport locations, delays, and flight connections.

Various graphical analysis techniques, including bar charts, stacked bar charts, bubble charts, and histograms, are employed to present complex datasets in a clear and informative manner. Comparative analyses reveal patterns and trends, enabling a deeper understanding of delay causes and their impacts on different airlines.

The implementation of interactive elements and user interface design facilitates seamless navigation between views, allowing users to explore airport locations, average delays, and flight connections. Challenges related to the dataset size are addressed through query optimization and indexing, ensuring efficient data retrieval and visualization.

Finally, this visualization project contributes to increased transparency in the aviation sector by providing stakeholders and the general public with accessible and understandable insights into flight delays. The innovative combination of data analysis and visualization techniques offers a valuable tool for decision-making, ultimately working towards the goal of reducing delays and improving the overall passenger experience in air travel.

1. Introduction

Overview of flight delays and their impact. Objectives of the visualization project.

This report presents a visualization project focused on the analysis of flight delays. Using a comprehensive dataset, our goal is to identify patterns and causes of delays in the aviation industry. This analysis aims to provide insights that can help reduce delays and improve the passenger experience. The project combines detailed data analysis with innovative visualization techniques to make complex flight delay information accessible and understandable.

2. Dataset and Motivations

The dataset for this visualization project comes from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics , which covers many aspects of flight delays. It includes variables such as flight dates, times, duration, delay lengths, reasons for delays, and geographic data of flight routes. The dataset provides a comprehensive overview of flight operations in the U.S. and their punctuality over a defined period of time, allowing a detailed analysis of the performance of the aviation industry.

The primary motivation for selecting this dataset is to understand and visualize the patterns and causes of flight delays, a major concern in the aviation industry that affects millions of passengers annually. By analyzing this data, we aim to identify trends and correlations that can inform stakeholders, including airlines, airports, and regulators, to develop strategies to reduce delays and improve the passenger experience. In addition, this visualization project aims to make complex flight delay data more accessible and understandable to the general public, thereby increasing transparency in the aviation sector. Also, this visualization can help to reintroduce the competition between airports, especially in areas where they are many airports, such as the San Francisco Bay, and then force airports to improve their performances.

3. Technical setup

To realize this project, we decided to use a wide range of technical tools. Our work is run on a web server, so it does not depend of a local setup. Then data are stored in an MySQL Database allowing us to query on it using the powerful SQL structure alongside the possibility to build indexes. For example, we built indexes on Airlines, States, FlightDate and Delay because that was the four main filters we use to query our data.

To build our visualization, we decided to use Leaflet for the map view. The map is filled using D3 and the SVG on the map is added thanks to a special package that allows us to link a layer over the map. Also, because the layer is over the map, and not on the map, we are able to create some interactions between the user and the D3 plot. Simpler graphs were built using Vega-Lite.

Also, queries on the data are fetched by d3.json() which calls a PhP page and gives parameters using the method GET (it means encoded in the URL). Then the PhP page does the request on our MySQL database and return the result as a json.

4. Visualization Techniques

This section outlines the various visualization techniques and libraries used in the project. The goal of these techniques is to present the flight delay data in an intuitive and informative manner, allowing for easy interpretation and analysis of complex datasets.

4.1. General view on the map

4.1.1 Airport View

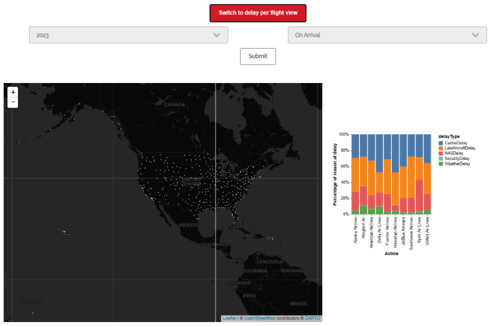


Figure 1: Airport view

The Airport view (Figure 1) initializes custom symbols to represent airports on the map. The map, which is centered and zoomed appropriately, uses OpenStreetMap tiles for a detailed geographic background. Airports are dynamically drawn using D3.js, with each airport marked by its geographic coordinates. The map's interactive features allow users to zoom in and out, providing a dynamic view of airports in different regions. The locations of the markers are updated in real time, ensuring accuracy as users interact with the map. On mouse over, the name of the airport appears, so it is easily accessible.

This visualization provides a clear geographic overview of airport locations, highlighting the spatial distribution and regional density of airports.

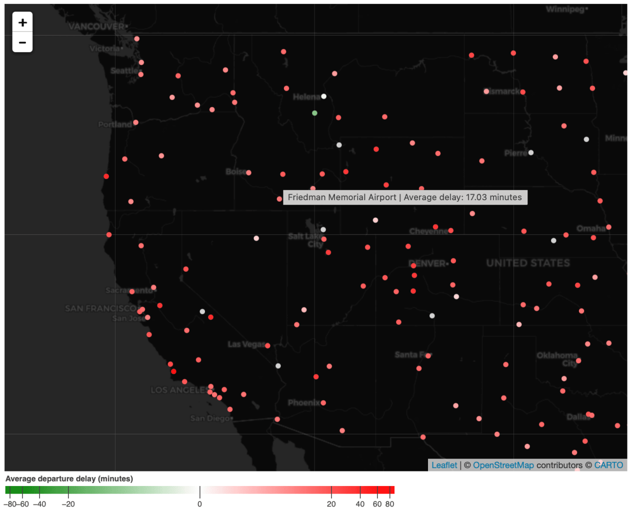


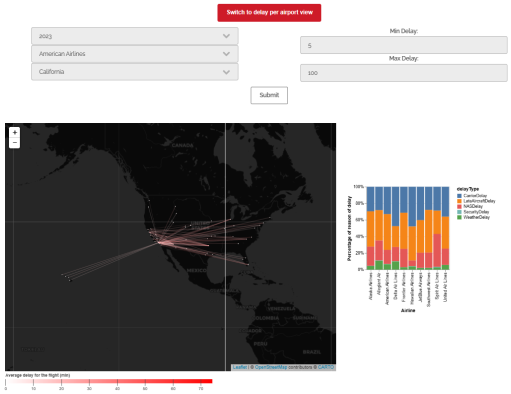
Figure 2: Airport view with query for 2023 on departure

Then, by selecting a year and a sens (On arrival or On departure), the airport plot now colors using a scale from green to red, showing on average for the year, the delay on arrival (resp. on departure) at the airport. On mouse over, the name of the airport and the associated delay appear. So, the user can have a global view of the delay at each airport, and then get the precis value for the airport they want. Figure 2 gives the result of a query for year 2023 on departure.

4.1.2 Flight View

The Flight View (Figure 3) integrates flight path data, including origin and destination coordinates, to plot flight paths on the map. Each flight is represented by lines or symbols that illustrate the routes taken. The map is equipped with event listeners for zooming and view resetting, ensuring that flight paths are accurately displayed and updated in response to user interactions.

This aspect of the visualization allows you to explore flight routes, shedding light on flight patterns, including frequency and geographic extent of routes. The interactive capabilities (understand : on mouse over, all the flights except the one selected are colored in light gray and opacity is reduced. Thus, a tool tip shows which airport are concerned by the link and what is the average on arrival delay for this



Switch to delay per airport view

2023

American Airlines

California

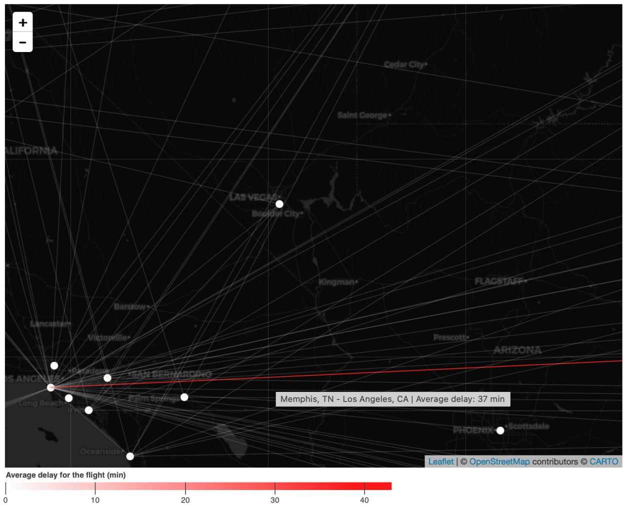
Min Delay: 5

Max Delay: 100

Submit

CANADA
delayType
CarrierDelay
LateAircraftDelay
NASDelay
SecurityDelay
WeatherDelay
Percentage of reason of delay
Airline
Average delay for the flight (min)
Leaflet | © OpenStreetMap contributors © CARTO

Figure 3: Flight view



Memphis, TN - Los Angeles, CA | Average delay: 37 min

Figure 4: Flight view for Delta Airline, California on 2023

connection) of this view provide deeper visibility into flight connections and the extensive network of air travel (see Figure 4). It allows users to analyze and understand the complexity of flight operations and provides valuable perspectives on how air travel connects different regions.

4.2. Graphical Analysis of Flight Data

Describe different graphs used and what they represent. Analyze the patterns and trends revealed by these graphs.

4.2.1 Bar chart Visualization : Causes of flight Delays by Airline

The bar chart visualization processes (Figure 5) delay data by categorizing the occurrence of different delay types across different airlines. This dynamic chart updates based on user-selected delay types, showing the distribution of delays such as ’WeatherDelay’, ’CarrierDelay’, etc. for each airline. The setup includes defining scales and axes, with the x-axis representing airlines and the y-axis indicating the percentage of delays in term of time.

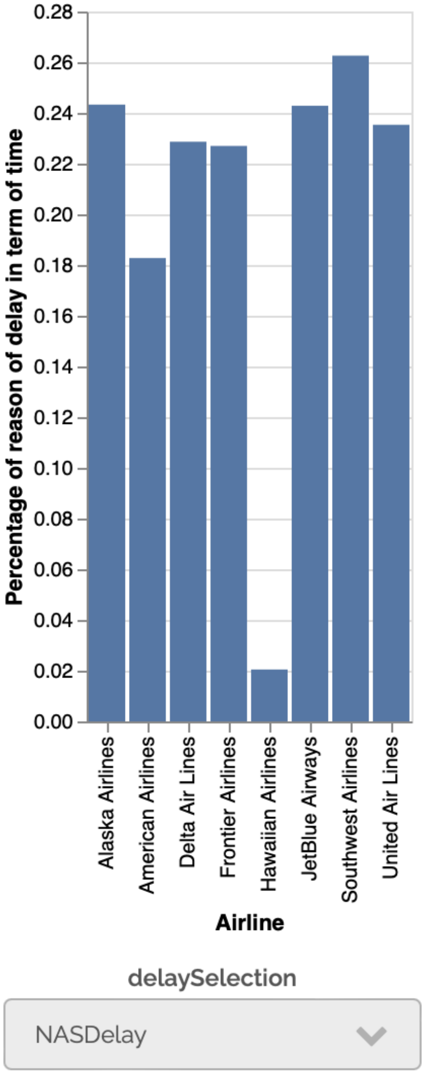


Figure 5: Flights delay at San Francisco Airport in 2023

This simple approach to implementation highlights the key aspects of building the bar chart, emphasizing the use of Vega-Lite for dynamic data representation and the considered design choices that enhance the chart’s effectiveness and user experience.

Comparative analysis: The bar chart format allows for easy comparison of delay causes between airlines. This can reveal patterns and outliers in delay events across the industry.

Interactive Insights: The interactive nature of the visualization allows users to explore different types of delays (Figure 6), providing a deeper understanding of the factors

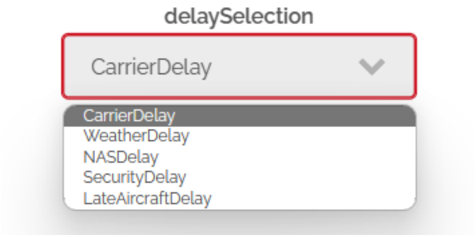


Figure 6: Delay types

that affect airline punctuality.

4.2.2 Stacked Bar Chart Visualization : Delays by cause and airline

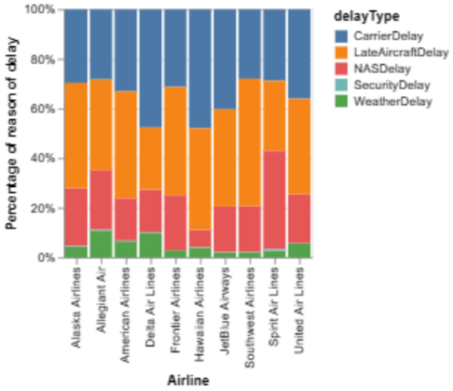


Figure 7: Delays by cause and airline in 2023

In this stacked bar chart visualization (see Figure 7), flight delay data is processed and grouped by airline, with each delay type categorized and color-coded for clarity. The implementation includes defining a color scale to differentiate delay causes and creating a legend for easy reference. The chart itself displays airlines on the x-axis and the percentage of delays on the y-axis, with each bar representing an airline and segmented to show the percentage of each delay type. These segments are stacked to show the total delay impact per airline.

Comprehensive overview: The stacked bar chart format provides a full analysis of the causes of delays for each airline, making it easy to see which types of delays are most prevalent.

Comparative analysis: By stacking the different types of delays, the chart allows for comparative analysis between airlines, highlighting which airlines are more affected by certain types of delays.

4.2.3 Bubble Chart of Delay Causes

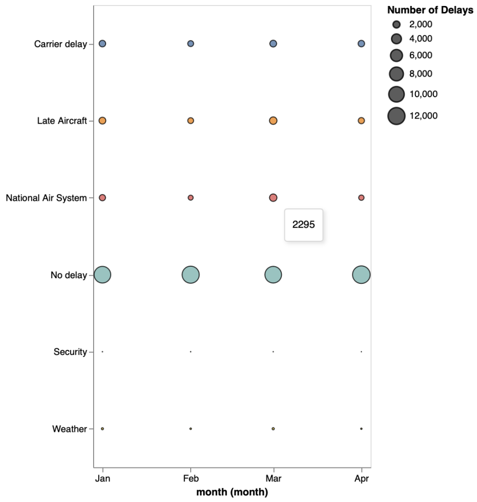


Figure 8: Number of flight per delay type per month at San Francisco airport in 2023

This graph (Figure 8) uses circles to represent delays, with the x-axis representing the month and the y-axis representing different causes of delays. The size of each circle correlates to the number of delays in terms of flights for that cause in a given month. This visual format allows for easy comparison of delay causes over time. Color coding is also used to differentiate between delay causes, making the graph easier to read. In addition, bubble chart are efficient to show general information, but not good for precise number. So, just by placing their cursor over a circle, users can get the precise number of flight affected by this delay for this month.

Cause Analysis Over Time: The Bubble Chart uniquely illustrates the distribution and frequency of different causes of delay over time, allowing for month-to-month comparison of delay factors.

Quantitative representation: The size of each bubble quantitatively represents the number of delays, making it easy to see which delay causes are most prevalent in any given month.

4.2.4 Comparative analysis of the flight delays of different airlines

Comparative Analysis of Flight Delays Across Airlines The visualization on Figure 9 was created using Vega-Lite. It processes the dataset in a sophisticated way with a special focus on the delay durations and the respective airline networks. A density calculation is applied to the *'ArrDelayMinutes'* field, it quantifies the probability distribution

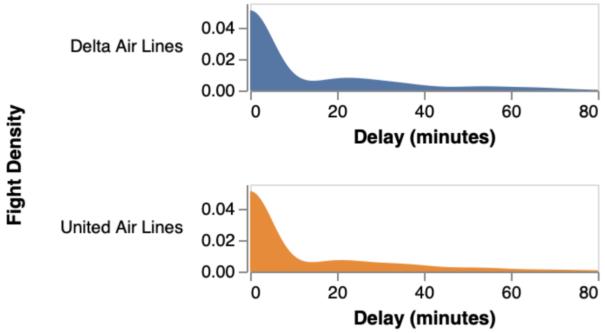


Figure 9: Comparative airline delay distribution analysis

of flight delays for each airline. The visualization is presented as a series of area plots, one for each airline. This method allows a direct and comprehensive comparison of the delay patterns. The implementation uses color coding to differentiate each airline, allowing for quick distinction and recognition.

Probability Density Estimation: density values are not direct counts or proportions, but density estimates. They represent the probability per unit on the x-axis.

Normalization: Density values are normalized so that the area under the density curve over the entire x-axis for each airline is equal to one.

4.2.5 Bar Chart of Flight Count by Airline

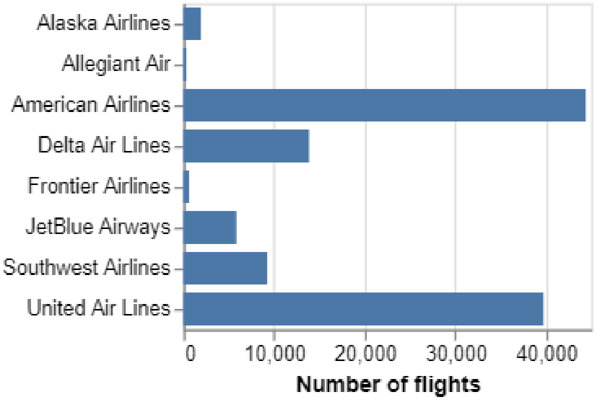


Figure 10: Flight Count by Airline

The Bar Chart of Flight Count by Airline (see Figure 10) provides a clear visual representation of the operational scale of each airline by showing the total number of flights. This is essential for understanding how flight delays are distributed across airlines, especially when comparing the operational complexity of larger airlines with higher flight volumes to smaller airlines. This analysis provides key insights into each airline's efficiency and reliability, which are critical to assessing their performance in managing schedules.

4.2.6 Delay per airline per destination

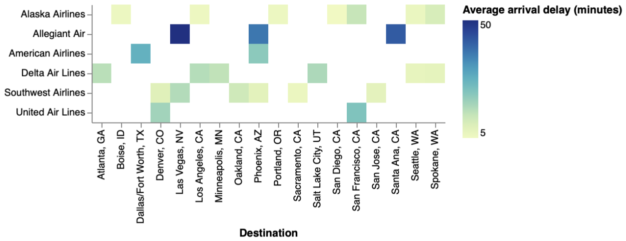


Figure 11: Delay per airline per destination departing from Spokane airport in 2023

This 2D Histogram Heatmap (see Figure 11) is a useful tool to reestablish competition between airline in an airport. Indeed, it shows all destination available at the selected airport and which airlines serves it while showing the average delay per airline per destination. So, If you customer, you can easily choose the airline which will be the most likely to get you to your destination on time.

4.3. Time Series Analysis of Flight Delays

The visualization (see Figure 12) provided is a time series graph representing flight delays throughout the day. It uses a two-layer approach to show both individual delay events and the overall trend of delays over time.

This two-tiered approach helps identify patterns and outlier events, providing both a macro and micro view of operational efficiency. Together, these elements provide a comprehensive view that is useful not only for immediate operational decisions, but also for strategic planning and analysis.

5. Interactive Elements and User interface

Putting thighs together, we designed the user navigation this way. First when the page is called, the user initially arrives on the Airport view (Section 4.1.1) showing nothing more than all the airports in our database on the map. Then they can select a year and a sens to show the average delay at each airport on the selected sens. They can also swap to the flight view (Section 4.1.2) using a button on the top of the visualization. Here they can ask for the average delay for flights departing from or arriving to the given state, using the select airline for a specific year. We would have prefer to be able to show all airlines at the same time, but the number of links is already big enough. This would have lead to a messy view, hardly understandable, especially for states containing big airport, such as Georgia with Atlanta.

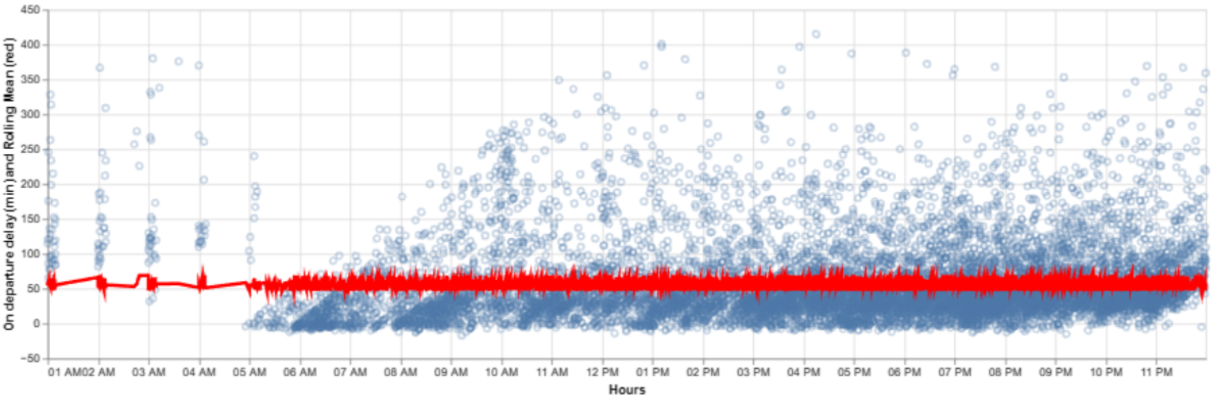


Figure 12: Day Rolling Average Plot

The switch between views does not require to reload data thanks to the property ”display:none” that allows us to switch between different view instantly.

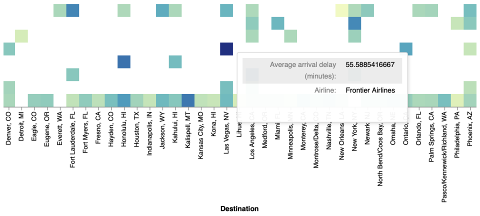
Thus, clicking on airport, on the map, no matter the view, will give access to the airport specific graphs.

To enhance user experience, we have also implemented a loading screen as illustrated below in Figure 13.

Figure 13: Loading screen

6. Challenges and Solutions

One of the main challenge is the size of our dataset. Ideed, it contains 29 637 504 entries, even after removing all entries linked to small airports (airports which have a IATA code starting by a number). To deal with this issue, we decided to limit every data queries to a selected year. Also, the Flight view is restricted by airline, because if not, they were to many data to plot on the map. In addition, the loading time was very long, depending of the query. To tackle this issue, we built indexes in our database management system which allows us to reduce the query to almost a minute, depending of the query (for instance, the flight query selecting American Airline, Alaska and 2023 in almost instantaneous, where the flight query selecting Delta Airline, Georgia, 2023 takes about 1 minute). Thus, we added a loading gif (Figure 13) so it shows to the user that something is happening while at the same time, it prevents the user from spamming a button because nothing seems to happen while they make multiple queries so it increases the query time.



Average arrival delay (minutes): 55.5885416667
Airline: Frontier Airlines

Figure 14: Delay per airline per destination departing from San Francisco airport in 2023

Another challenge was the adaptation of the graph showing delay per airline per destination (explained at Section 4.2.6). Indeed, on Figure 11, we chose a small airport on purpose. For big airport such as San Francisco (see Figure 14), because they are link to much more destination, the 2D Histogram can overflow the screen, leading to an unclean view of the data while at the same time, the user could be able to see some destination, but not airlines and the legend. To solve this, we decided to add a sub division in our HTML code, creating a horizontal scroll bar in case of overflow. In addition, we added a tooltip on mouse over, so if the legend and airline are not showed on screen, the user can still access the information.

We used the same trick to depict the distribution of delays for each airline (cf. Figure 9) when too many airlines are present.

7. Conclusion and Future Insights

In conclusion, this visualization project demonstrates the power of combining detailed data analysis with sophisticated visualization techniques to make complex information about flight delays accessible and understandable. It provides a comprehensive view of flight delays in the aviation industry, showing patterns, trends and causes. This project not only provides a valuable resource for aviation community to develop strategies for improvement, but also increases transparency and helps the general public make informed decisions. The use of interactive elements and a user-friendly interface enhances the experience.

We've concentrated our efforts on graphics that focus on airports details and analysis. Future improvements of our visualization may include deeper analysis of airline performance or a broader comparative view of different airports, considering aspects such as delay frequencies, causes, and duration.

Furthermore, with improved computational efficiency, we could enhance our visualization with statistical insights showing trends over years through interactive time series analysis. However, due to our already long loading time, we did not add such graphs.