Project Metadata

Title: **Flight Delay Analysis**
Group Member Names:

- Aditi Ravindra
  - aravindra@umass.edu
  - 33490623
- Yash Kamoji
  - ykamoji@umass.edu
  - 34032599
- Sreevidya Bollineni
  - sreevidyabol@umass.edu
  - 34738775

Github Repository: https://github.com/ykamoji/airport-delay-analysis

Background & Motivation

Our group chose to do this project on Flight Delay Analysis because all three of us frequently take flights as a form of travel, as we are not from Massachusetts. Flight delays are a very frustrating and unfortunately, common experience, which made this topic relevant to us personally and thus, more meaningful. Taking a deeper look into the reason and patterns behind delays with flights felt like an interesting and practical way to apply data analysis to a real-world problem we all regularly face. This project combines both our personal experiences as well as our drive for data-driven exploration; by examining these patterns in delays, we are eager to gain insights that can hopefully help frequent fliers, such as us, better anticipate delays and plan itineraries that go hand in hand with that.

Project Objectives

The goal of our project is to analyze flight delay data to identify common causes of delays based on factors such as location, time, and season. By visualizing this data, we want to answer key questions such as the most frequent reasons for delay, the percentage of delayed flights, any time or seasonal patterns that affect the punctuality of flight arrivals, and most dependable airline. Understanding these factors can help not only help passengers anticipate delays, but also allow airlines to be prepared and mitigate them more effectively. Overall, the insights we can gain through analyzing this flight data can help improve flight scheduling within airports, allow for better resource allocation, and enhance customer satisfaction among airlines, all of which contribute to a more efficient travel process.

Data

The Bureau of Transportation Statistics (BTS) is a part of the Department of Transportation (DOT) and is the primary source of transportation data for the U.S. government. They collect, analyze, and disseminate data on all modes of transportation. This dataset, known as the “Airline On-Time Performance Data”, encompasses detailed records of scheduled and actual departure and arrival times for non-stop domestic flights reported by certified U.S. air carriers.

Since June 2003, the airlines that report on-time data also report the causes of delays to BTS. This is available in a publicly accessible interface where users can explore various aviation databases, including the "Airline On-Time Performance Data," to retrieve specific datasets. The raw data can be downloaded in excel sheets, with the option to obtain one month of data at a time.

Data source link: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

Data Processing

For this analysis, we need to compile on-time flight performance data from the past year. However, since the public interface only provides access to data one month at a time, we must retrieve each month's dataset individually. Once all 12 months of data are collected, we will aggregate them into a single dataset for comprehensive analysis.

To maintain the integrity of our findings, we conducted a preliminary data-cleaning process to address key inconsistencies in the dataset. This included handling missing values for critical flight details such as departure time, origin, destination, and air time. Additionally, the raw data is provided in excel sheets that contain formatting discrepancies, requiring manual adjustments before the data can be processed for analysis.

Beyond basic cleaning, we also perform a final step to have read data efficiency. This involved removing redundant attributes that can be derived from existing columns, such as quarter (which can be inferred from the date) and airport name (which can be obtained from the airport ID). These optimizations were necessary to reduce the overall dataset size, ensuring faster data processing.

By analyzing the data over time, we can identify patterns in flight delays associated with different seasons, months, or specific time periods. This allows us to uncover trends such as seasonal fluctuations in delays, peak congestion periods, and variations in airline performance.

Additionally, we can compute key performance metrics, including the proportion of flights that were delayed, canceled, or diverted. We will explore insights such as the average delay per airport, rankings of the most and least delayed airports, and congestion levels at various airports.

For data preprocessing, we utilize a Python script to automate the entire workflow, eliminating the need for manual handling of individual excel files. Instead of manually opening, reading, and scrubbing each file, our script systematically processes all available excel files, extracting relevant data and merging them into a single consolidated dataset.

During this automated process, we also perform the necessary data-cleaning steps outlined earlier. This includes handling missing values, resolving formatting inconsistencies, and removing redundant attributes to optimize dataset size and efficiency. This approach ensures consistency, accuracy and significantly reduces the time and effort required for data preparation.

Visualization Design

Ideas

Aditi Ravindra
Yash Kamasi
Sreevidya Bollineni

Factors

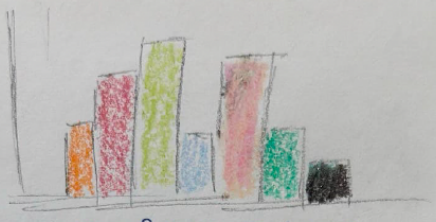
airports
- all/some? which airports to choose
- busiest/least busy airports
- most/least delays

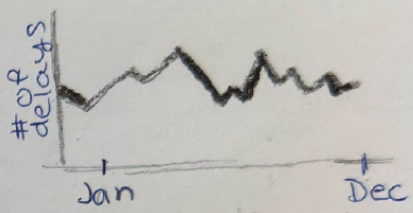
airlines
- all/some? which airlines to consider?
- most/least popular
- most/least departing/arriving flights
- "best" / "worst"

delays/performaree
- arrival/departure delays
- comparisons (which delay is most/least common)
- relevance
- analyze flights that are delayed only/discard rest. as it doesn't answer our question

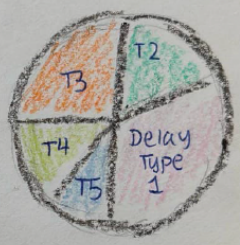
Dataset:

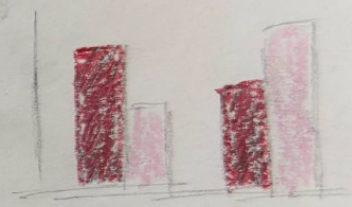
- years - timeline of data → monthly (csv)
- airport details (name, state, ID)
- airline details (name, ID, ...)
- diff types of delays (5)
- diversions



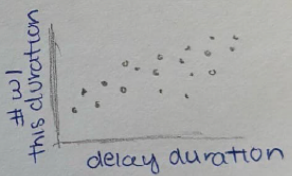


map, organized by state/color





- no delay
- delay



FILTER

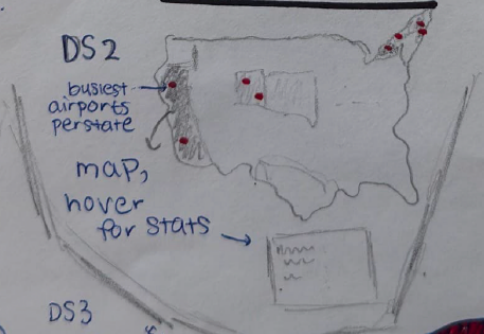
- which states have the most delays w/ arriving/departing flights?
- what is the most common cause of delay?
- what are the most/least efficient airports?

Visualization Ideas:

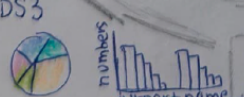
- map for initial view
- stats/% of diff types of delays (pie chart)
- best/worst airports in terms of delays (bar graph)
- show overall data by timeline (month, season) using line graph

CATEGORIZE

DS2

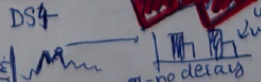


DS3



- shows % of delay by cause e.g. weather, etc. → to help answer q.
- best 5 and worst 5 airports (less delay = best ✓)

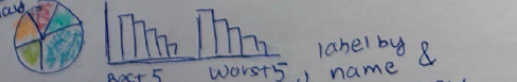
DS4



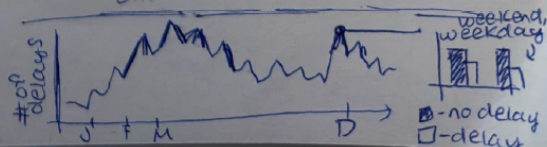
- line graph x-axis: month y-axis: # delays
- darker for more delays to make it easier to read

To show time/seasonal info

Combine & refine



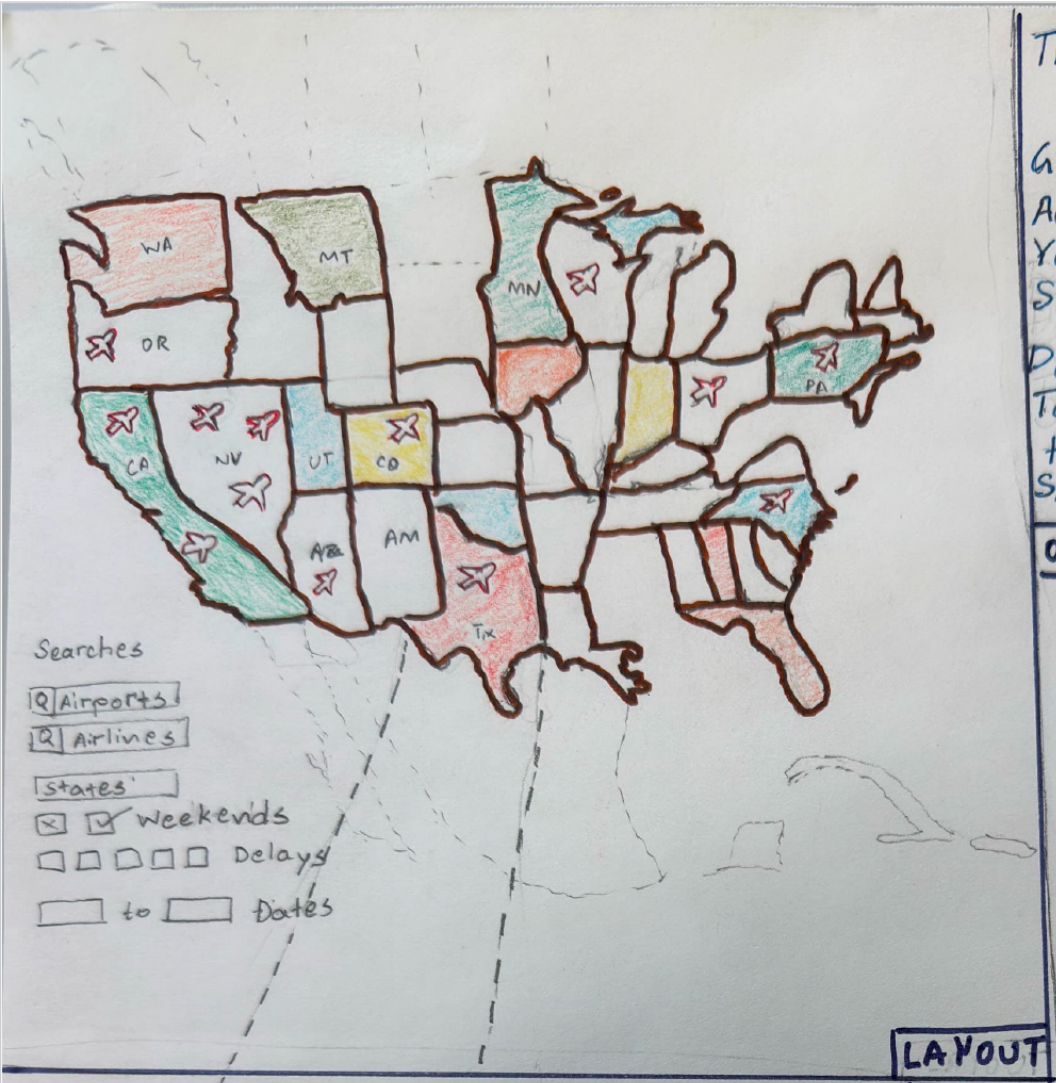
label by name & show best/worst 5



Title: Flight delay analysis

Group Members:
Aditi Ravindra
Yash Kamoji
Sreevidya Bollineni

Date: 2/26/2025
Task: Geographical Map of total delays in a state.
Sheet: DS 2

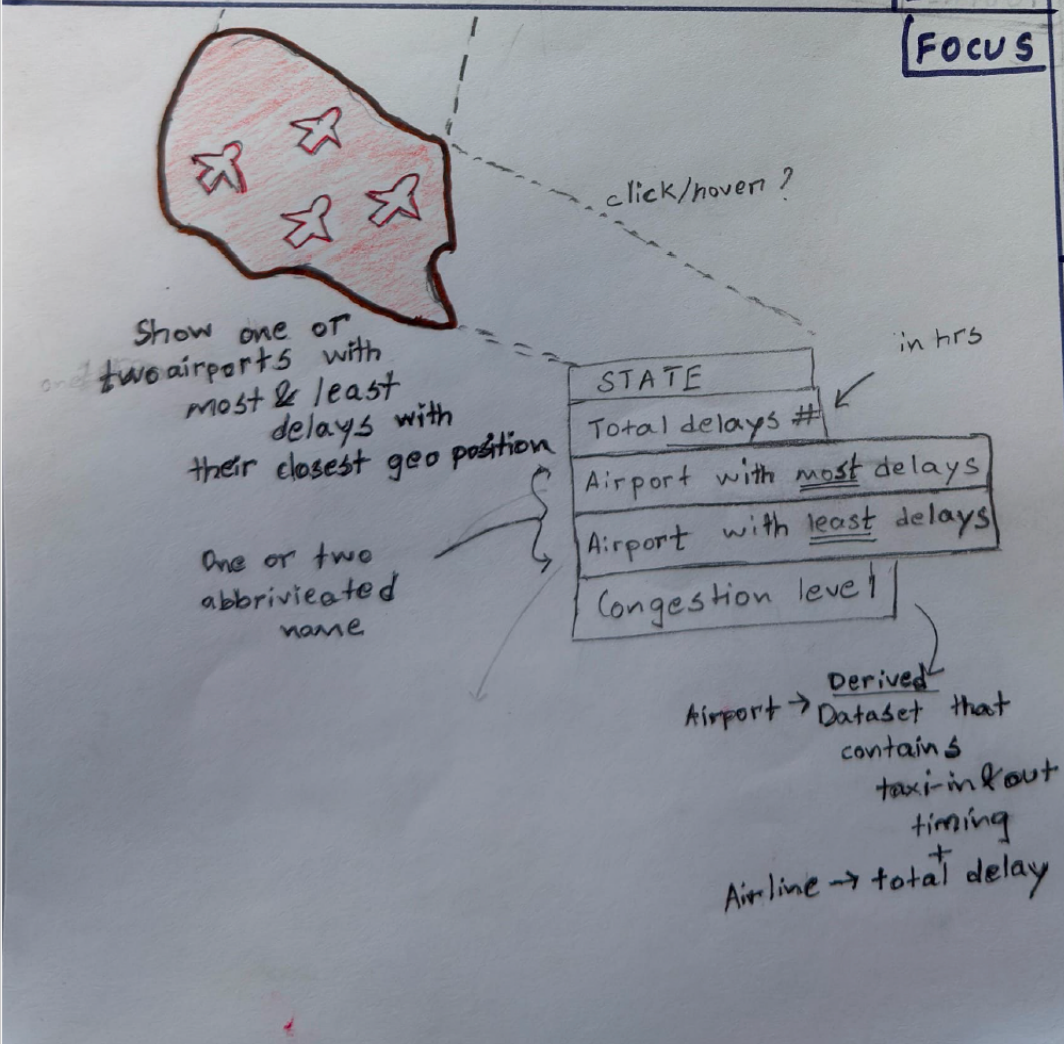


OPERATIONS

- User can hover on a state on the map.
- Popup with stats of that state.
- Filters / Selections on airports, airlines, states, weekday / weekend, date duration delay types.
- Map will have a gradient of shades on the states that correlates with the total number of delays

LAYOUT

FOCUS



DISCUSSION

MAP shows:
↳ Regions having darker shade have higher delay than lighter regions.
↳ Popup placement should not collide with map elements.
↳ More stats to be shown on hover?

Selection or filtering should grey out region which are irrelevant.
Color gradient scheme should be clear
↳ Considerations for visual impairments

Title : Flight delay analysis

Author : Aditi Ravindra
Yash Kamoji
Sreevidya Bollineni

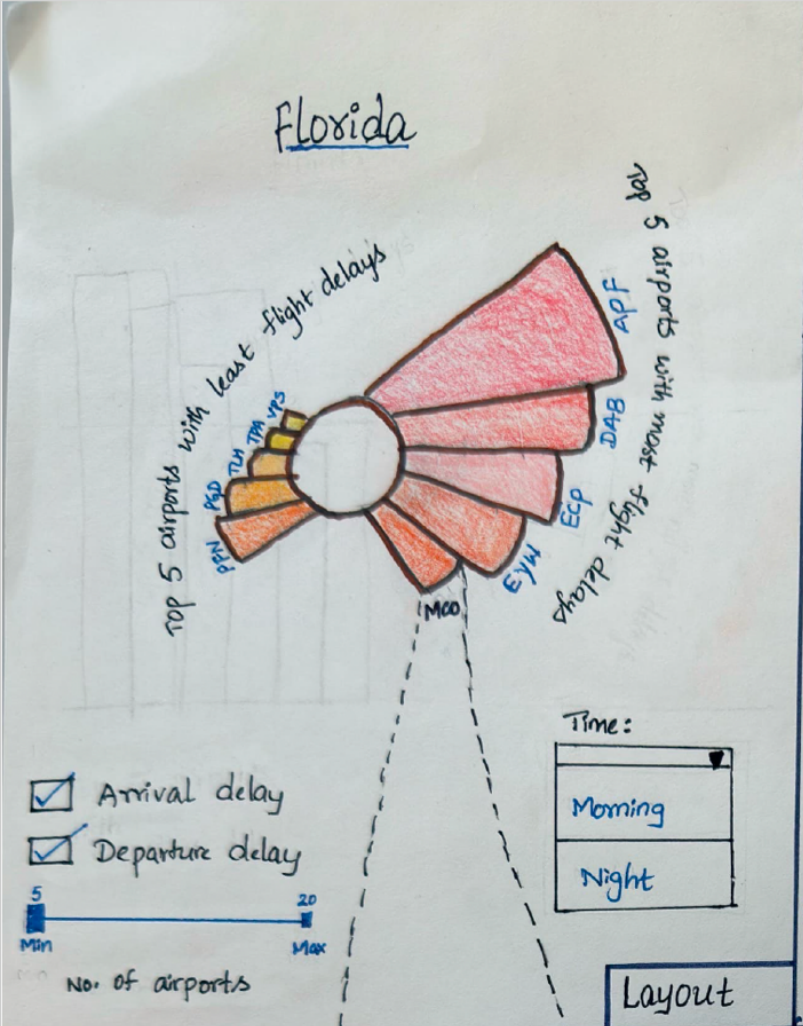
Date : 02/26/2025
Sheet : DS 3

Task : Representation of state-wise airports with the most and least flight delays

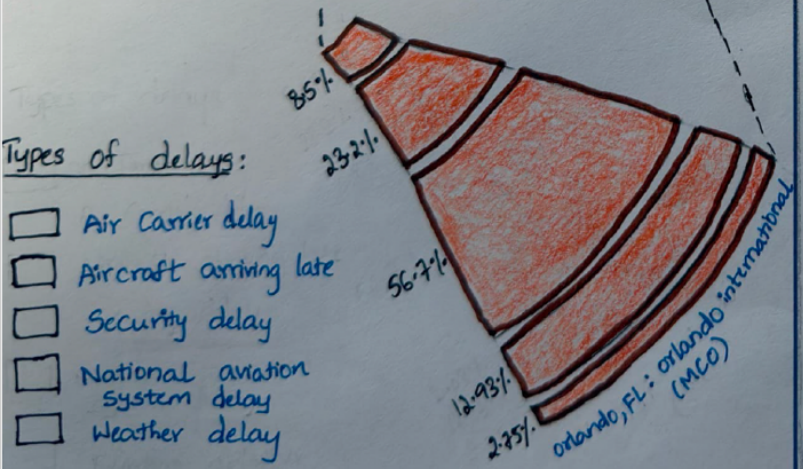
Operations :

1) User Selections :
   * Select the number of airports to display for most flight delays and least flight delays
   * Choose the type of delay to analyze :
     - Arrival delay
     - Departure delay
     - Both arrival & departure delays
   * Filter delays based on time of day :
     - Morning delays
     - Night delays
2) Hover Interaction :
   * When hovering over an airport, display :
     - The airline with the most delays at that airport
     - The airline with the least delays at that airport
3) Click Interaction :
   * Clicking on an airport shows a detailed breakdown of different types of delays for that airport

Layout

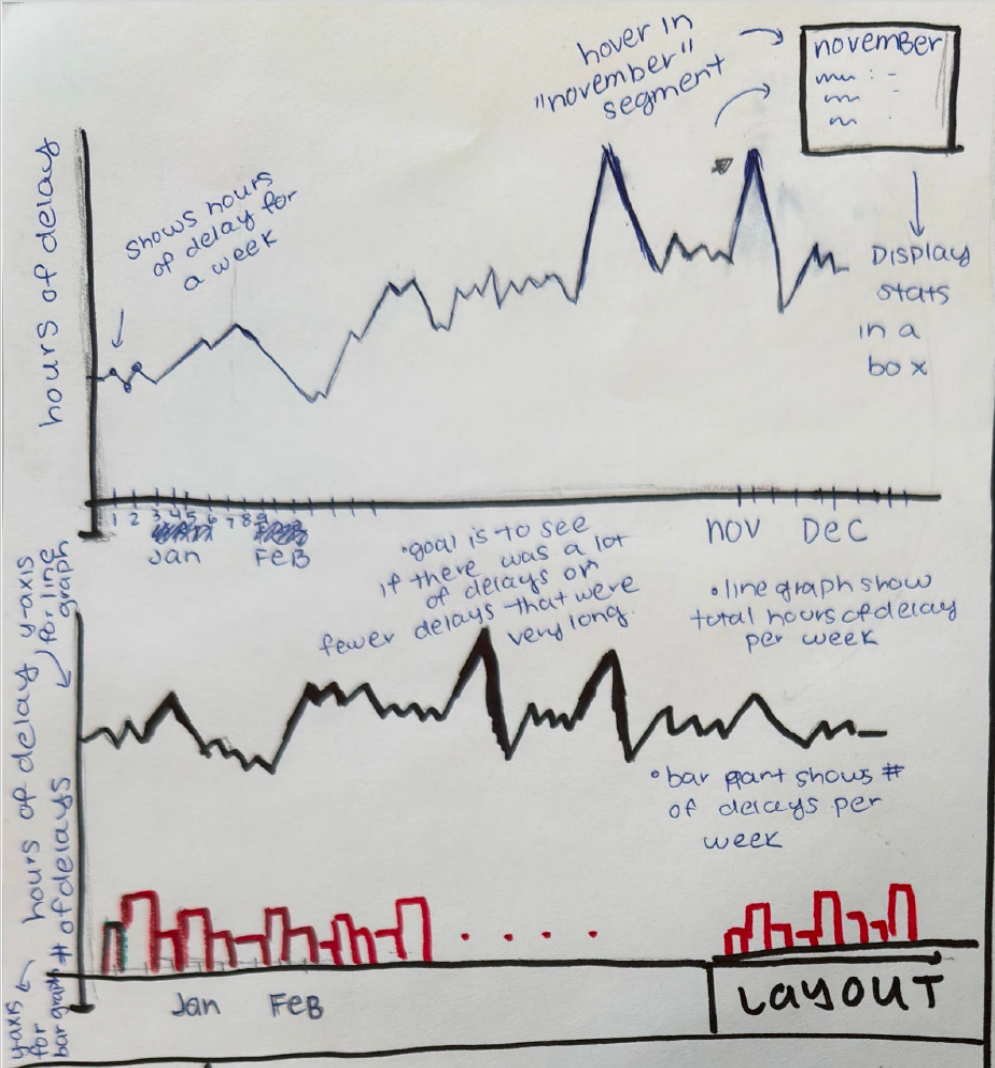


Focus



Discussion :

* Each airport, whether having the most flight delays or least flight delays, is represented as a segment around the circle.
* The radius of each segment corresponds to the total number of delays at that airport
* Airports with the highest number of flight delays are represented in
* Airports with the lowest number of flight delays are represented in
* When a user selects a segment (airport), a breakdown of different types of delays is displayed as a percentage.



LAYOUT

Title | Flight Delay Analysis

Authors | Aditi Ravindra, Yash Kamoji, Sreevidya Bollineni

Date | 2/26/2025

Tasks | ~~[illegible]~~ line graph, bar graph, stats, operations, details

Sheet | DS4

METADATA

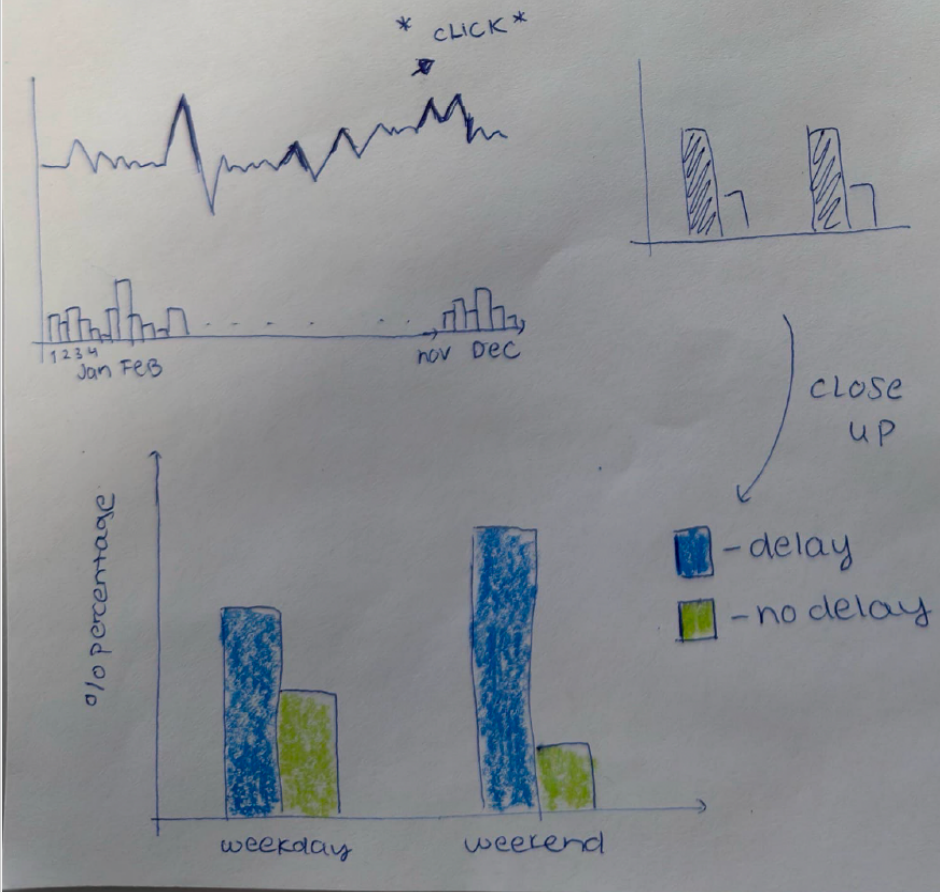
- The graph is segmented by month (Jan-Dec).
- When you hover over a particular segment, it will display statistics in a box (numbers, not a visualization) (% of each delay contribution)
- clicking on ~~the~~ a segment will load and visualize a bar graph on the side to the right w/ info on weekend/weekday delays.
- Possible? Load a visualization to the right and have the line graph (originally centered) slide to the left

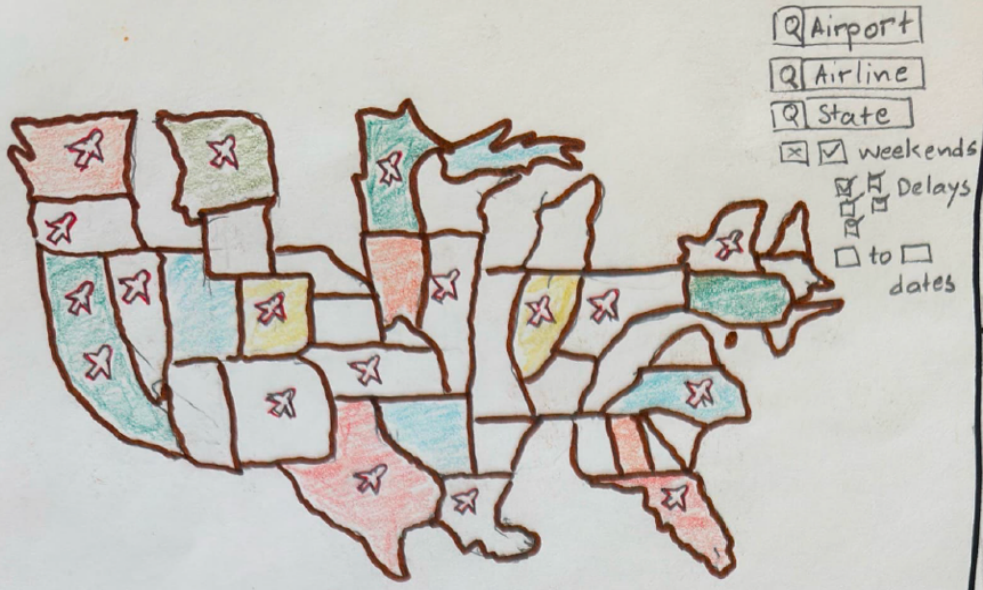
OPERATIONS

DISCUSSION

- 52 datapts for each week
- The visualization graphs a subset (yet to be chosen) of the datapoints, segmented by month. (more explanation in layout)
- The line gets darker / thicker as the # of delays (y-axis) gets higher.
- The x-axis is labeled by month, and week
- The y-axis by #s (by 10s, 100s, ...)
- When you click on a segment (segment "lines" - where it separates to diff month - not visible, but relatively clear), it shows a bar graph.
- The bar graph has 2 groups "weekends" and "weekdays" and shows # of delay/no delays — meant to help user visualize if weekends/days have more delays in a certain month

FOCUS





Title: Flight Analysis Data

Author: Aditi Ravindra, Yash Kamoji, Sreevidya Bollineni

Date: 2/26/2025

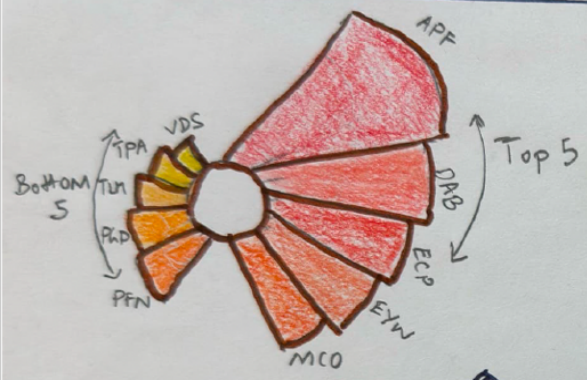
Sheet: DS5

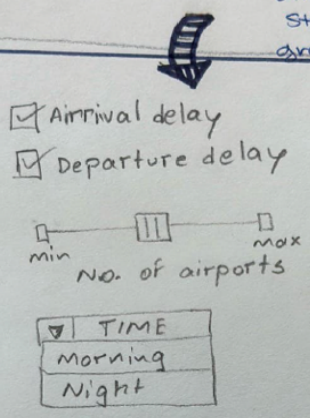
Tasks: Final decision, overview, operations, detail

- all data loaded
- filter allows user to look at a specific year, airport, airline, state, weekdays/weekends, dates.
- Hovering over a state displays stats, and little plane icons (2) per state rep most efficient airports.
- Clicking on a state on the map leads to state 2, which displays

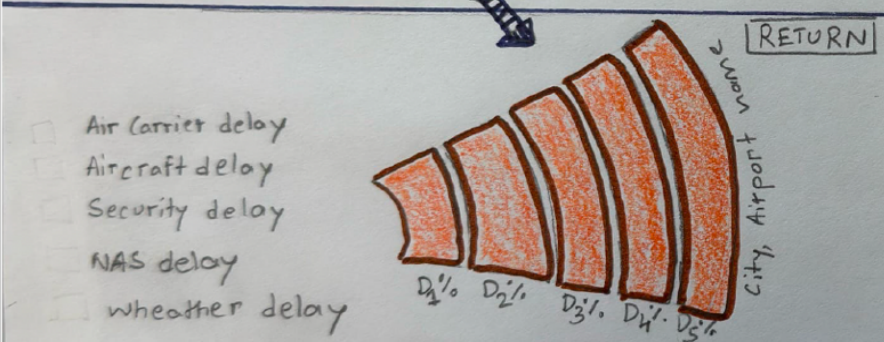
state 1 → map

other states gray out





- State 3, clicking ~~reset~~ show all takes you back to the og map (state 1) (all states colored/ none grayed out)
- #/hours of delays by week displayed in line + bar graph on same graph separately.
- Hovering displays info abt a month and clicking on a month's segment shows a bar graph to display c/o of delays on weekends/days.



OPERATIONS

DETAIL

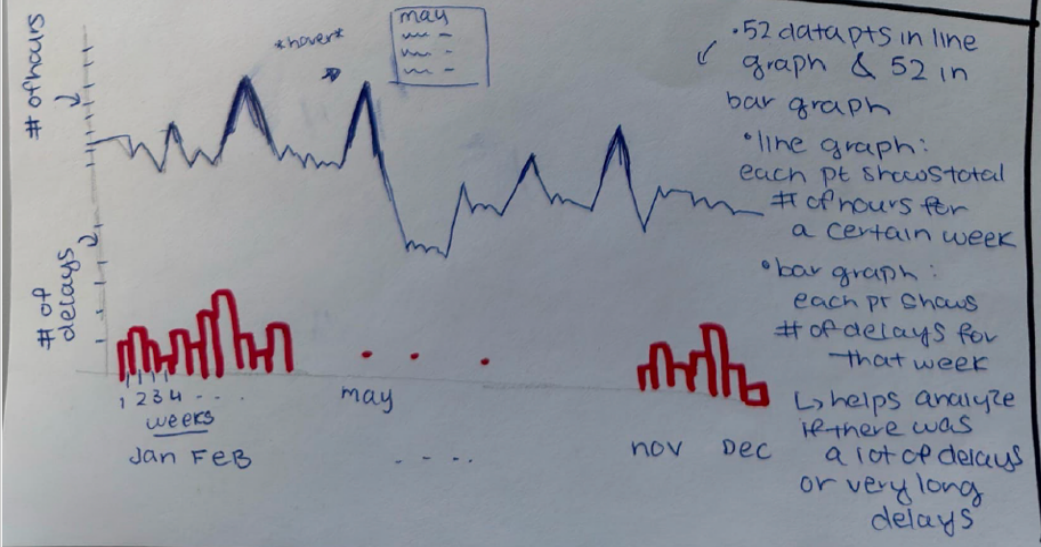
Public dataset: Needs to be cleaned & preproccessed using python (pandas).

Bootstrap support for UI controls & sizing.

Website and its viz will be built on vanilla HTLMS, CSS, Javascript.

All the stats and visualization calculations will be processed in python & then feed to website.

Color scheme selection is based on the views & variation of it data.



- 52 datapts in line graph & 52 in bar graph
- line graph: each pt shows total # of hours for a certain week
- bar graph: each pt shows # of delays for that week
  - ↳ helps analyze if there was a lot of delays or very long delays

Estimate:

5 data preparation
10 integrating data & its vizualization
10 Sheet #1 design
10 Sheet #2 design
10 Sheet #3 design

45 days ← total

Must-Have Features

Below are the essential features required:

1. **Interactive Filter & Selection Options:**

   Users can filter flight delay data by:

   - **Time**:
     - Year, Month, specific dates, weekday/weekend.
     - Flight delay time: morning or Night
   - **Location**: State, Airport
   - **Airline**: Select specific carriers.
   - **Delay Type**:
     - Filter by type of delay cause.
     - Toggle between arrival, departure, or combined delays for a particular state selected.

2. **Dynamic Geographic Delay Heatmap with Drill-Down:**

   - **U.S. Heatmap**:
     - Color-coded by average delay time per state (e.g., red for most delays, green for least delays).
   - **Click Interaction:** Zoom into a state-level view to display:
     - Top airports most and least delayed (ranked by number of delays), based on the selection of number of airports.
     - Heatmap intensity reflects severity.

3. **Radial Bar Chart (Nightingale Chart):**

   - Visualizes top airports with the highest and lowest delay counts.
   - **Radius length**: Represents total delay count per airport.
   - **Color coding**: Red (highest delays) to yellow (lowest delays).
   - **Customization**: Adjust the number of airports displayed (e.g., top 10, top 20).

4. **Hover Interaction for Detailed Insights:**

   - Hovering over a state displays:
     - Total number of delays in that state a
     - One or two airports with the most and least number of delays.
   - Hovering over an airport in a particular state reveals:
     - **Most delayed airline**: Airline with the highest delay count at that airport.
     - **Least delayed airline**: Airline with the lowest delay count at that airport.

5. **Clickable Elements for Delay Cause Breakdown**

   - Clicking on an airport displays a pop-out of the selected airport segment displaying breakdown of delay causes and their proportions:

        - Weather delays
        - Air traffic control delays
        - Carrier-related delays
        - Security-related delays
        - Late aircraft delays
    - Clicking on a Matrix Segment displays non-adjacent matrix comparisons (e.g., Jan vs. Dec).

6. Interactive Segmented Line Graph:

- **Segmented by Month (Jan-Dec):**
    - Analyzes delays across different months to identify patterns.
- **Hover Interaction:**
    - Displays total delays and delay causes for the selected month.
- **Click Interaction:**
    - Loads a bar graph comparing weekend vs. weekday delays for the selected month.
- **Dynamic Layout Adjustment:** The line thickness increases as the number of delays increases (thicker = more delays).

7. **Comparative Statistics & Ranking Dashboard:**

- **Ranking table**: Sorts airports by total delay count.
- **Highlighting**: Based on the number of airports selected to display:
    - Top airports with the most flight delays (red).
    - Top airports with the least flight delays (yellow).
- **Weekend vs. Weekday Comparison:**
    - Bar graph visualizing weekday vs. weekend delays to identify trends.

8. Axis Clarity & Scaling:

- **X-Axis:** Clearly labeled by month (e.g., Jan-Dec).
- **Y-Axis:** Uses an appropriate numerical scale (e.g., 10s, 100s) for delay counts.

9. Dynamic Visualization Layout & Real-Time Feedback:

- **Adjustable Layout:**
    - Visualizations resize and rearrange dynamically when new graphs are loaded.
- **Real-Time Feedback:**
    - Line graph thickness adjusts dynamically based on delay severity (thicker = more delays).

These features will provide actionable insights into flight delays, helping to identify key factors influencing flight disruptions and optimizing airport operations.

Optional Features

Below are the optional features that are nice-to-have:

1. **Animated Transitions Between Views:**
    - ❖ Smooth animations when switching filters or visualizations.
2. **Multi-Airport Comparison Tool:**
    - ❖ Users can select multiple airports and compare:
        - ➢ Delay patterns.
        - ➢ Airline performance.
        - ➢ Weather impact on flight punctuality.
3. **Alternative Travel Recommendations:**
    - ❖ Suggest alternative airports or airlines with fewer delays based on past data.
    - ❖ Provide best time to book flights based on historical delay trends.

These features would enhance user engagement and provide deeper insights but are not essential.

Project Schedule

Date	Task	Team Member
03/07	Preprocess data	Yash
03/21	Complete stats and visualization calculations	Yash & Aditi & Sreevidya
04/11	Implement design #1	Yash
04/18	Implement design #2	Sreevidya
04/18	Implement design #3	Aditi
04/30	Integrate all designs and complete website	Yash & Aditi & Sreevidya