

CS573 Final Project

Investigation of Flight Delay of Major Carrier in the US

Team: Mei Yang, Huayan Sun and Hamid Mansoor

Date: Oct 27th

We met to go over the several ideas for our project. We combed through various online resources provided in the project description. Our general goal was to make a visualization story that will be accessible to the average person i.e. someone with a basic literacy in assessing visualizations.

We found an interesting data source on the Department of Transportation about flight data. The data was fairly well detailed and filtering features on the website were also very useful. We decided to go with that data source.

We discussed several visualizations and narrowed it down to four:

Parallel Coordinates

Chord Diagram

Map showing all the major airport in US

Treemap

We spent this meeting coming up with the project proposal. We discuss in detail, our rationale for these choices in the proposal. We have copied that section in the process book below:

Visualization Design

We want to provide the viewer with an overall view of how the rates at which flight delays and cancellations have changed over time for multiple carriers. We want to implement several visualizations that would help us achieve that.

To start, we will implement a Parallel Coordinate plot that will show the viewer the proportion/percentage of flights that were delayed or cancelled for every major carrier. We intend to show data for about 10 years. The users will be able to filter out carriers based on the percentage of flights that were delayed/cancelled. It makes sense to show this to the user first as it provides an overall generalized, cumulative view. Parallel coordinate plots are particularly effective to show temporal data as they allow the viewer to quickly see an overall trend. We have provided a sketch of our idea in Figure 1 below. We thought about some other ideas such as a bar chart matrix, to show the changes over the years.

The next visualization will be a Chord Diagram (Figure 2) to show the flights for different carriers in one month for the major airports. The airports will be arranged outside the ring and each of the connections will represent a flight. We allow the user to view only one month's data for an airline because anything more than that would really clutter up the diagram. For lines for the

flights that are delayed will be colored red. This visualization is meant to give viewers a way to see the seasonal changes in the delays for particular airlines. In addition, it will provide users a way to see which flight connections have the most or least likelihood of delays. A chord diagram is a particularly effective visualization for this problem as we have multiple destinations and origins and viewers are interested to see the links between them.

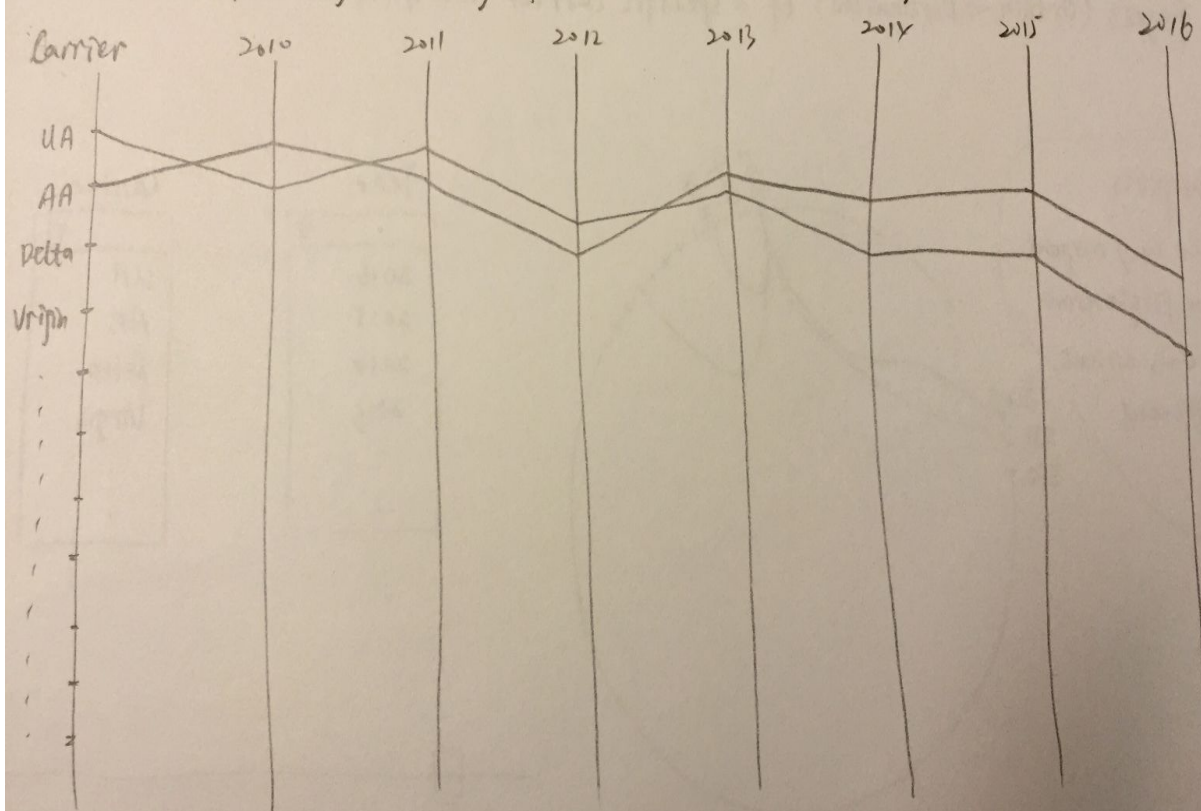
We want to allow the users to narrow down their search further by specifying a single airport. We will be using a map of the US and will have geographic markings for the major airports (Figure 3). Hovering over an origin airport will show connecting lines to other major destination airports. The width of the connecting lines will be proportional to the average delay time between the origin and destination airports. We thought about having a large interactive grid square with all the origins on the vertical scale and all the destinations on the horizontal scale. This did not seem like a great idea as the grid would get too complicated. A map seems more intuitive when discussing displaying information about cities/airports.

In addition to data for flight delays, we want the users to be able to see data about flight cancellations. We will make a Nested Treemap which will show the proportion of scheduled flights that were cancelled for each major carrier. Within each node of the treemap, representing one carrier, we will show the proportions of causes for the cancellation such as weather, aircraft faults, security related etc. We hand drew a prototype in Figure 4. We thought about using a grouped bar chart for this visualization, but it was also going to get too cluttered on the x axis to make much sense of the data.

One of the optional visualizations that we are considering is a word cloud at the end where the words will be the names of the airlines and the font size will reflect the overall average delay for that specific airline. That is, the lower the overall average delay time, the larger the font size.

We will be adding some sketches that we made to the proposal and to this process book.

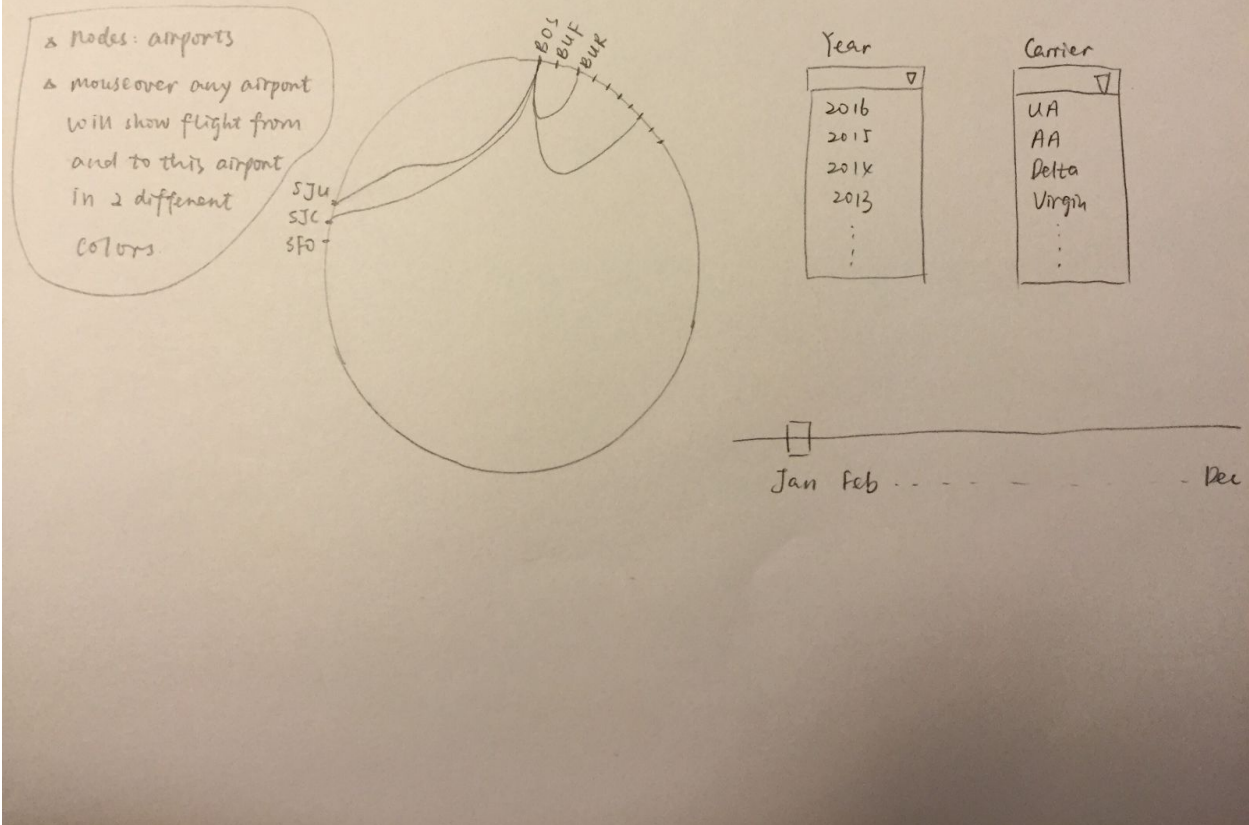
2. Overall flight delay frequency of carriers over past several years.



& Moreover a specific carrier, highlight that one

(Figure 1)

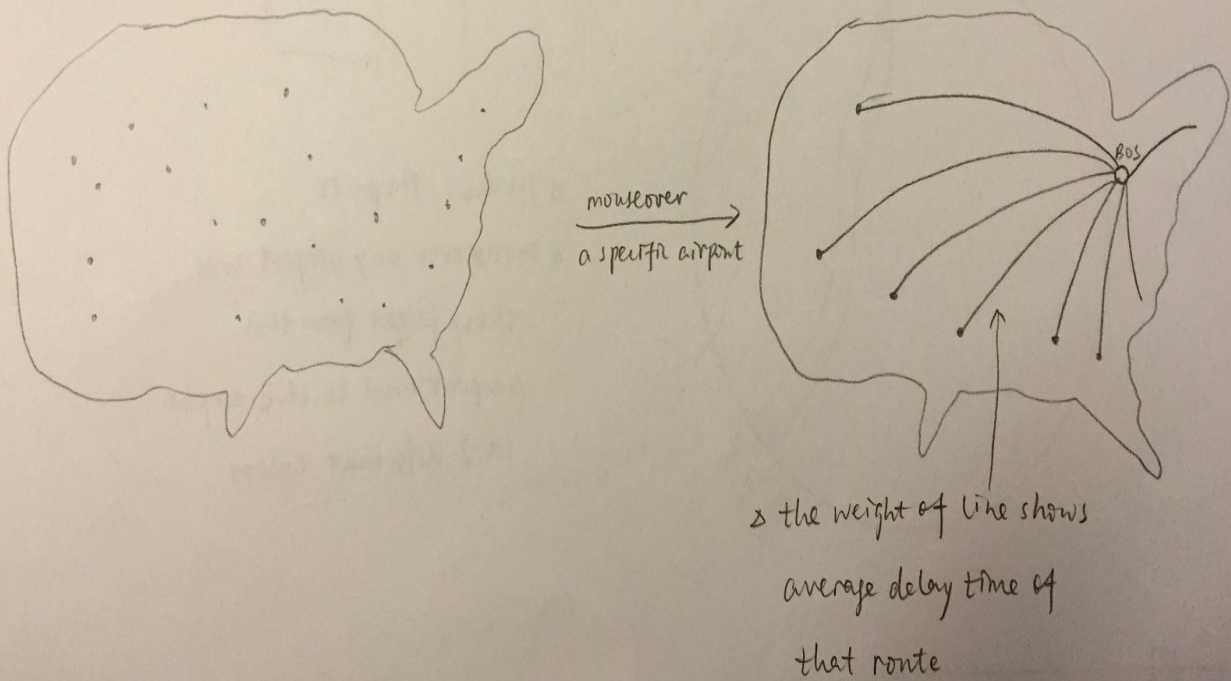
1. Show flights (Origin → Destination) of a specific carrier in a specific month.



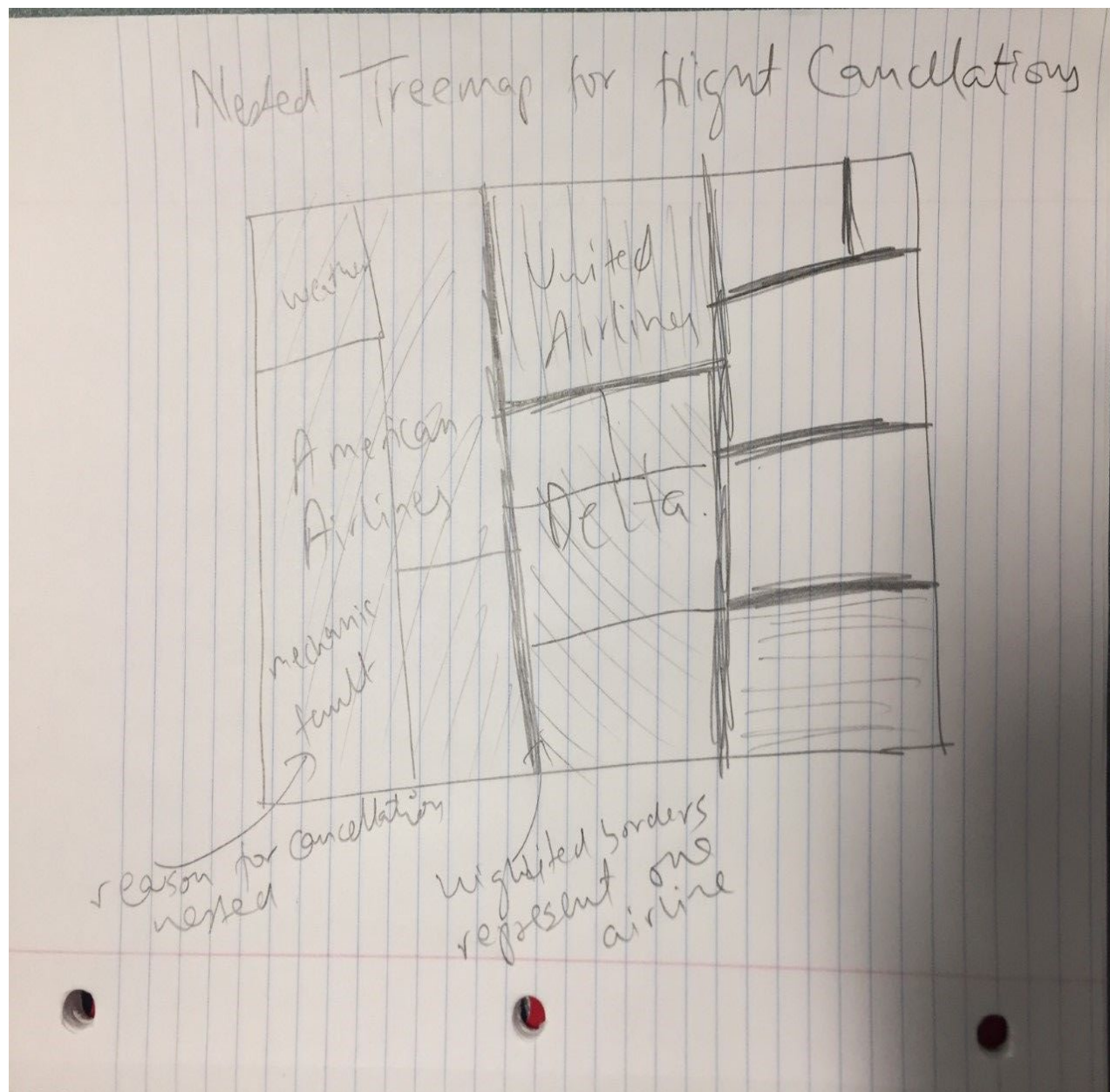
(Figure 2)

3. Average flight delay time from major airports

Carrier	year
VI	VI
UA	2016
AA	2015
...	...



(Figure 3)



We have written about some of the requirements of the process book below:

Overview and Motivation

Air travel has become widely available to travellers in the US. Delays and cancellations by major airlines causes a lot of problems for travellers. We want to give viewers a tool

to make an informed decision about the airline, airport, time of travel etc. that they choose, based on historical flight data.

Related Work

We saw some great examples in class, especially the MBTAVIZ project that inspired us to create a story to inform average users about travel.

Questions

One of the main questions we thought that viewers might have is how airlines fare against one another, given all other factors constant such as the origin, destination, time of year etc. We will allow the user to have several opportunities to compare airlines. We also thought that an important question that viewers may have is how airlines performed over time.

We had a pretty long meeting and we figured out some logistics such as setting up a github repo, setting up a google docs folder for process book, setting up a slack account and discussing our schedules. We have agreed to keep in constant contact over slack should any of us make any changes to the project.

Date: Nov 8th

We were all busy with the A5. The interaction techniques that we will learn from this assignment will be very useful in the final project.

We spent the previous week doing some data analysis. We want to get the data in proper shape before we move ahead with the actual implementation of the visualizations. We have described our methods below:

Data

We downloaded the dataset from United States Department of Transportation (http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time). It provides flight data for a specific month in a year that the user selects, and we downloaded the data from year 2011 to 2016 (the latest data at this moment is September 2016). We selected the fields we are interested in for each flight when we get the csv file, including carrier, origin airport, destination airport, arrival delay time

(min), delay time for five causes of delay, whether the flight is canceled and the cancellation cause, etc.

There are nine major carriers and more than 6000 airports total in the dataset. Considering it is hard to read showing all the airports in a single visualization and people rarely check flights of those small airports, we narrowed our scope to 30 major airports according to their rank in number of flights (https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States). We wrote a script in JavaScript to first filtered out the flights with origin or destination that we are not interested in. For example, there are total 454,879 flights in September 2016, and we narrowed that down to 189,212 flights within our scope of 30 major airports. We used the script to group the flights first by carrier then by month to integrate the flight information of a specific carrier in a specific month from year 2011 to 2015 into one csv file. When user inputs a carrier and month, we'll read flight data from that file and compute the average delay time for the visualizations.

For the on time data for the parallel coordinates plot and the cancellation data for the treemap, we used python to group together the data by airline and years. We used python because it is fast, which is very important given that we were manipulating millions of lines of csv. We will be attaching the python script to our submission. We simply used the "csv" library in python, which has a lot of cool features to features to allow for data exploration.

Exploratory Data Analysis: What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

Design Evolution: What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

Implementation: Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

Evaluation: What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

Date: Nov 17th

We met with Prof. Harrison today for feedback. One crucial piece of feedback he had was that we needed to make our story more interactive and simpler for the user since we are aiming for the average user. He suggested a simpler visualization to allow the users to filter out the major airlines and rearrange the data based on some parameters. We met after that and discussed some ways to do that. We decided to remove the chord diagram as it did not seem to present the information in a clear and straightforward way.

We looked up some table examples that we could leverage into our visualizations. Emma agreed to take on the table portion. We decided to continue working on our respective visualizations which were:

Hamid: Parallel Coordinates Plot and treemap for cancellation data

Huayan: Map of major airports

Emma: Table to allow the user to filter airlines

Date: Nov 27th

We met today to go over our progress and prepare for the prototype presentation. We have most of the visualizations up and running. We have some basic styling issues and functionality issues that we anticipate finishing before the presentation. The presentation will show our four visualizations and we expect to have a basic version up and running to show the class the direction in which we are headed.

Here are screenshots of where we are at right now:

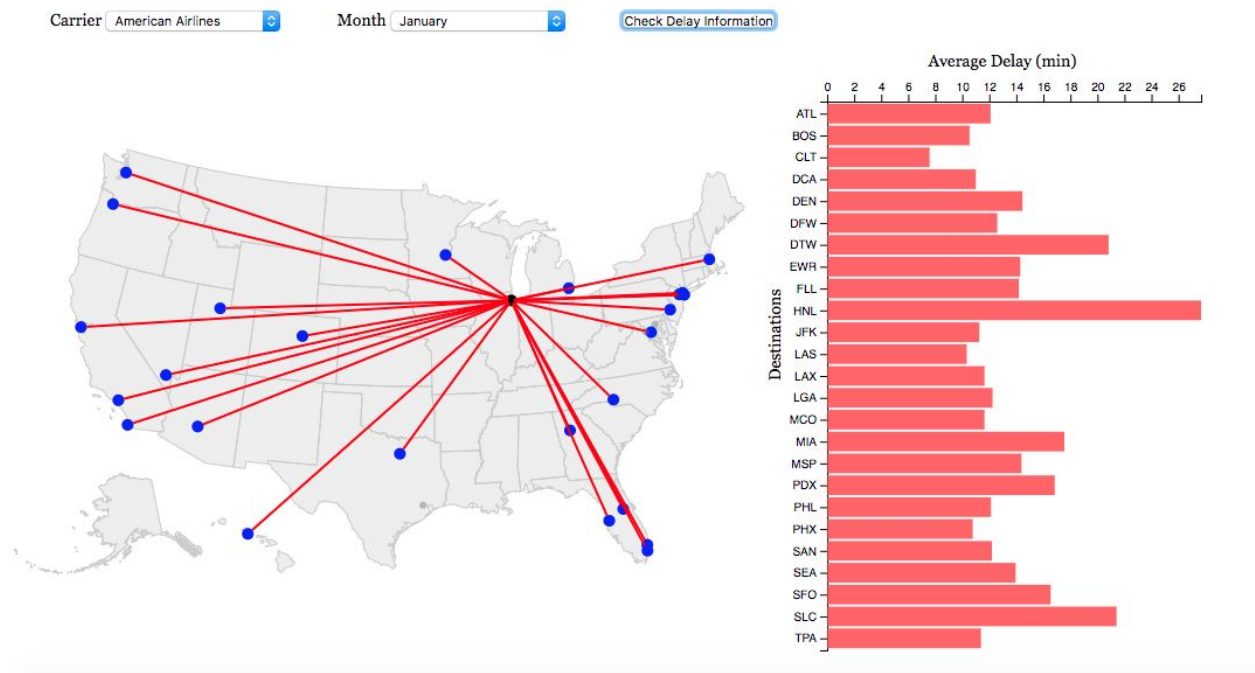
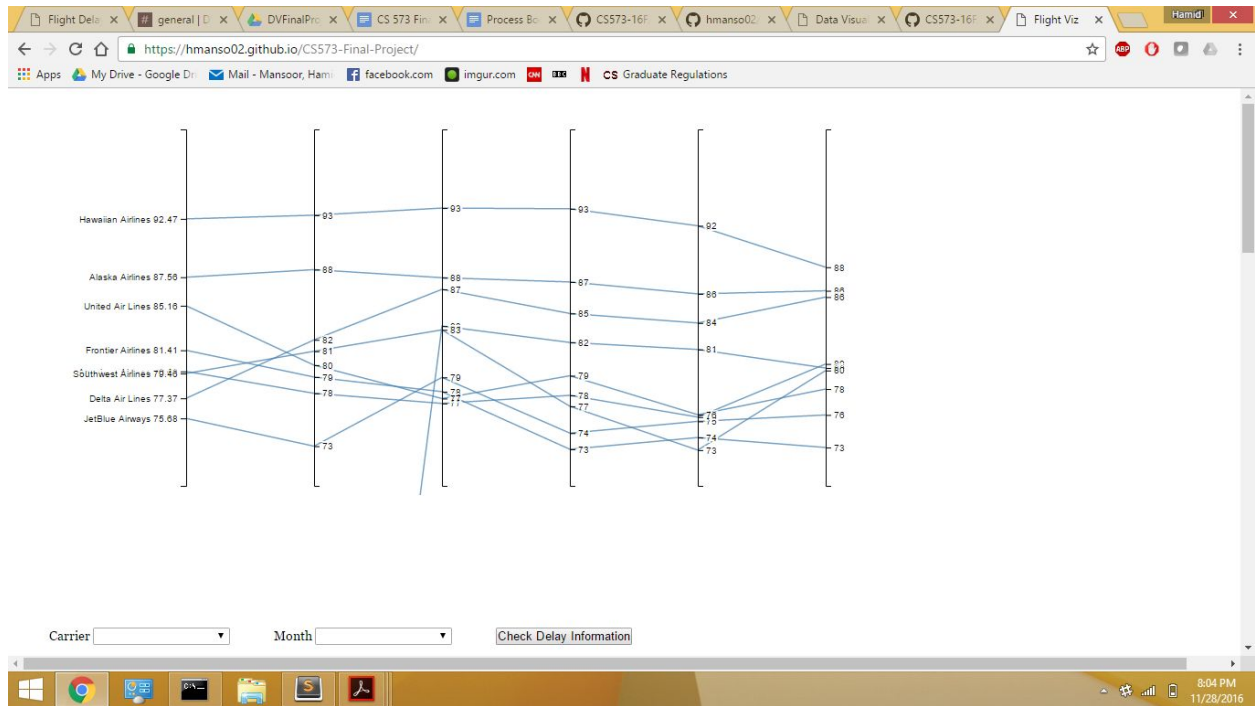
Carrier Month [Check Delay Information](#)

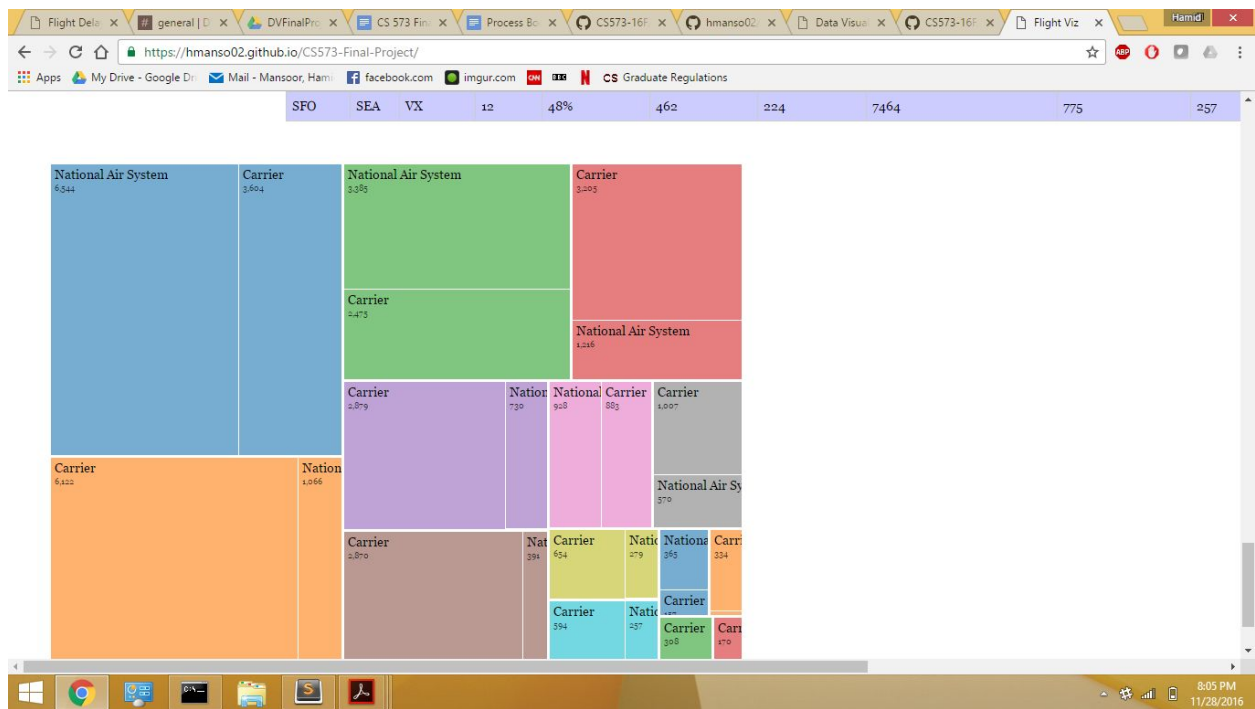
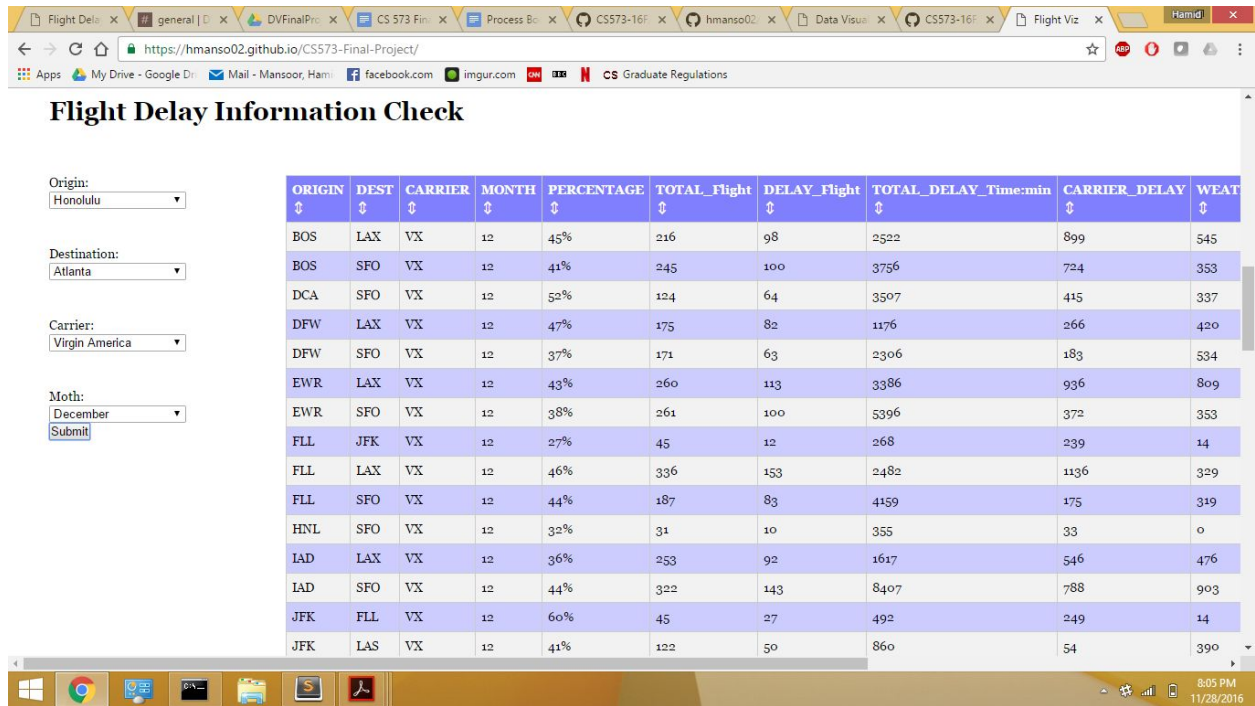


The styling is off in a bunch of places and the table is not functional. We will be working for the rest of the day to fix this.

Date: Nov 28th

We continued work on the presentation of the prototype. We have added all the required visualizations to one page. We are attaching a few screenshots to show our progress:





We think we are in a good place right now. All our visualizations are functional. We still need some work on the styling.