CS573 Final Project

Investigation of Flight Delay of Major Carrier in the US

Team: Mei Yang, Huayan Sun and Hamid Mansoor

Table of Contents

Part 1

This part aims to answer some of the overview questions that were outlined in the process book section of the final project description. We have added our proposal and the changes we have made since to this part.

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.

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:

Initially we used simple bar charts to look at our data which has nine major carriers and thirty major airports. We checked the average delay time of flights. We noticed that different carriers operate different flight routes and for the same routes by which we mean flights with same origin and destination, different carriers have different delay performance. For a single carrier, the average delay time for the same flight can be varied in different months which is caused by weather or other reasons. Thus, we are interested in comparing the delay performance of different carriers and investigate on the causes of delay so that we can provide guidance to users when they book flights.

Design Evolution:

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 was going to 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 a 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 were 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 the that specific airline. That is, the lower the overall average delay time, the larger the font size.

Based on the feedback we got from Professor, we changed our initial design a little bit. We think it is better to make a more user-friendly and user-centric application rather than just cool visualizations that users may not know how to read or interpret it.Thus, we decided to remove the Chord Diagram as proposed in the proposal and add a table where users can input their trip plan and get a sense of the chance of delay they may get.

We implemented the table and made several iterations of it to get the most important metrics. We also changed it so that it now adds donut charts for the delay cause for the carriers listed in the table.

For the treemap, after the prototype demonstration, we liked Prof. Harrison's feedback and added a grid of treemaps rather than a single one for all the years and months. Right now, we have a grid of 12 treemaps, one for every month.

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.

The first visualization is a Parallel Coordinate plot (Fig. 1-2) showing the flight ontime arrival performance of nine major carriers from year 2010 to 2015. The latest data we have is till September 2016, but since the winter months can have high flight delay rates which may make it not accurate to compare 2016 with other years we only look show the results till year 2015. The intent of this plot is to show the general flight ontime percentage of different carriers and how they change over years. When you hover over a specific carrier, its data will be highlighted in red while other carriers are hidden. First, users can get a sense of the flight ontime percentage of carriers in any year from 2010 to 2015. Second, users can compare ontime performance among different carriers and see if they are improving over the years.

The second visualization is a Map (Fig. 3-4) showing the flight routes and their average delay time for a specific month of a specific carrier, which is the average of the past 6 years from 2011 to 2016. Nine carriers are listed at the top where you can click on each one to select it. Each dot on the map represents an airport and you can select any one as origin. When you hover over it, the selected origin become black and all its destination airports will show in blue connected with lines (orange means there is a delay and green means no delay). At the same time, the average delay time in minutes are shown in the bar chart at the right. You can hover over to any origin and the visualization will change accordingly. The month is set to be January by default, and you can click on the slider to select any month to see the delay performance from this origin in each month. You can also click to change to another carrier to see the flight routes from that different carriers provide from this origin.

The third visualization is a Table (Fig. 5)

The table requires three inputs: origin, destination and the month. The user specifies these using the three respective selectors. After clicking the show button, the table is populated with the flight delay data for the major carriers operating those routes. The user has the ability to sort the carriers in ascending or descending order by any of the measure in the table by clicking on the arrows in the column headers.

The proportions of flight delay causes are displayed in donut charts, each one representing a carrier. This segment gets updated every time the metrics change.

The last one is a 3 x 4 grid of Treemaps (Fig. 6-7)

For the final visualization we implemented a grid of treemaps. Each treemap represents one month of the year. The individual tiles represent the proportion of each cause of cancellation. Hovering over the tiles highlights that particular cause and carrier across all the other treemaps as well, giving the user an easy way to compare across all the 12 months.

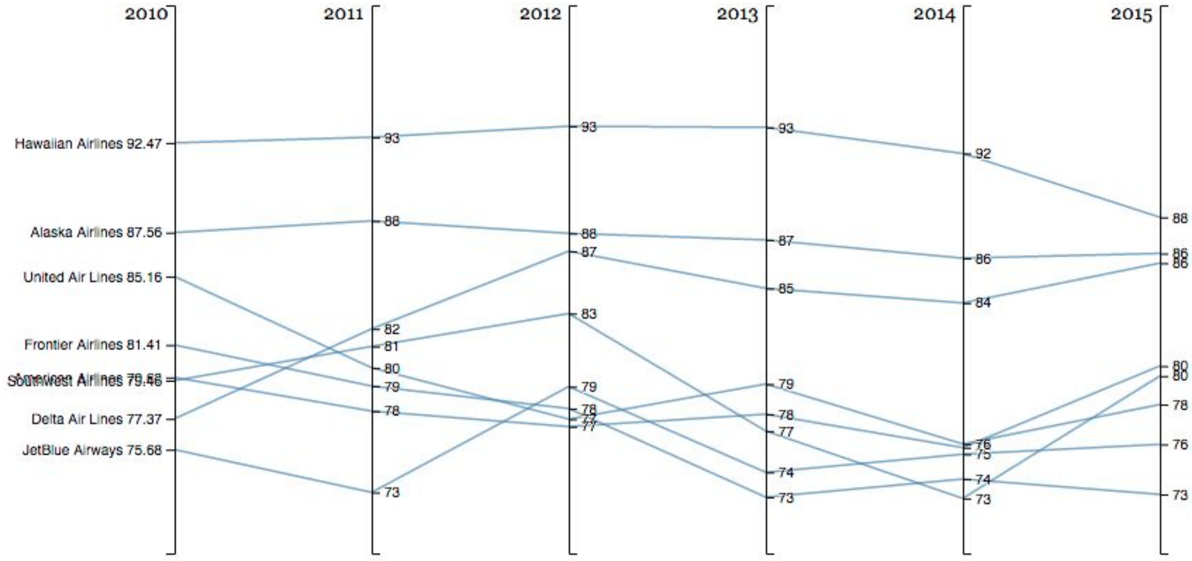


Fig. 1

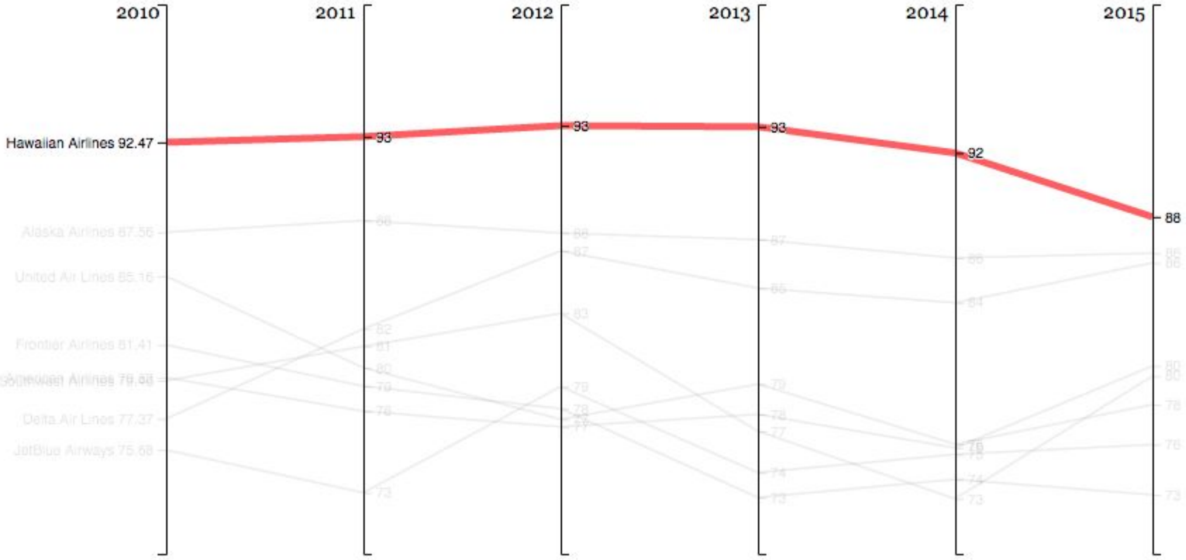

2010
2011
2012
2013
2014
2015
Hawaiian Airlines 92.47
93
93
93
92
88
Alaska Airlines 87.56
United Air Lines 85.16
Frontier Airlines 81.41
American Airlines 79.68
Southwest Airlines 79.68
Delta Air Lines 77.37
JetBlue Airways 75.68

Fig. 2

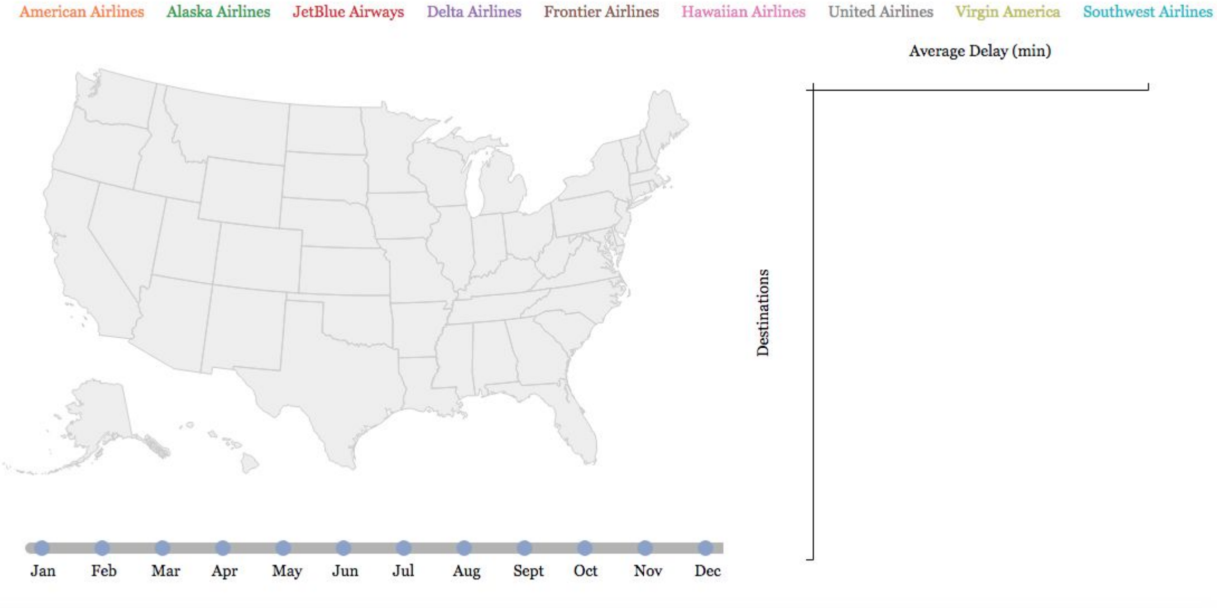

American Airlines
Alaska Airlines
JetBlue Airways
Delta Airlines
Frontier Airlines
Hawaiian Airlines
United Airlines
Virgin America
Southwest Airlines
Average Delay (min)
Destinations
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sept
Oct
Nov
Dec

Fig. 3

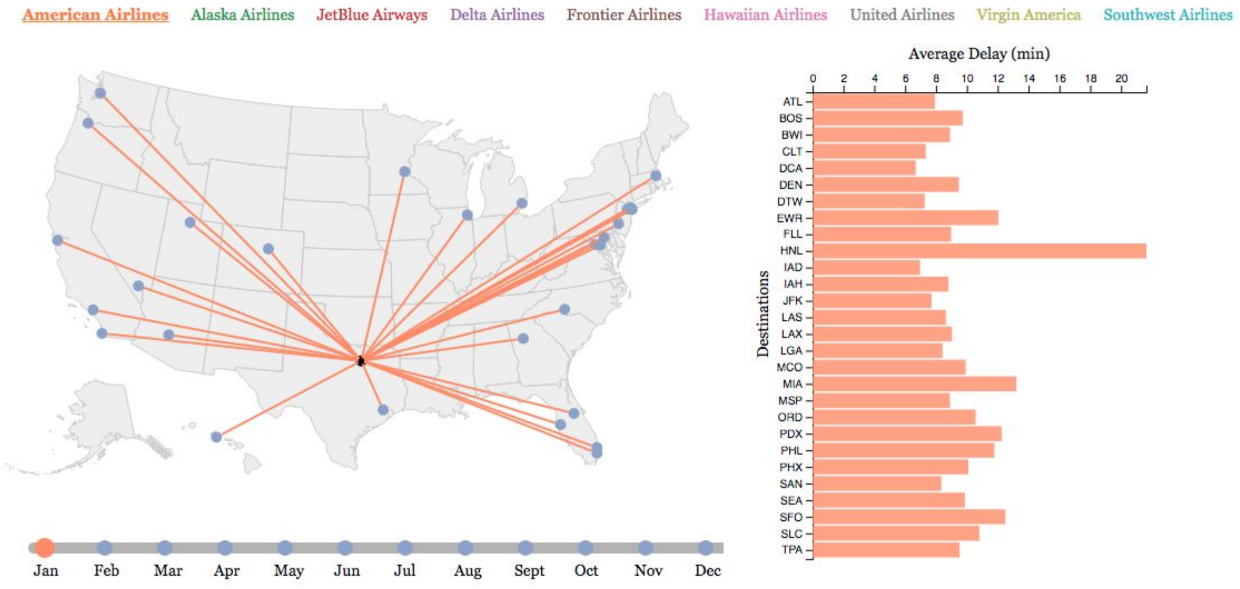
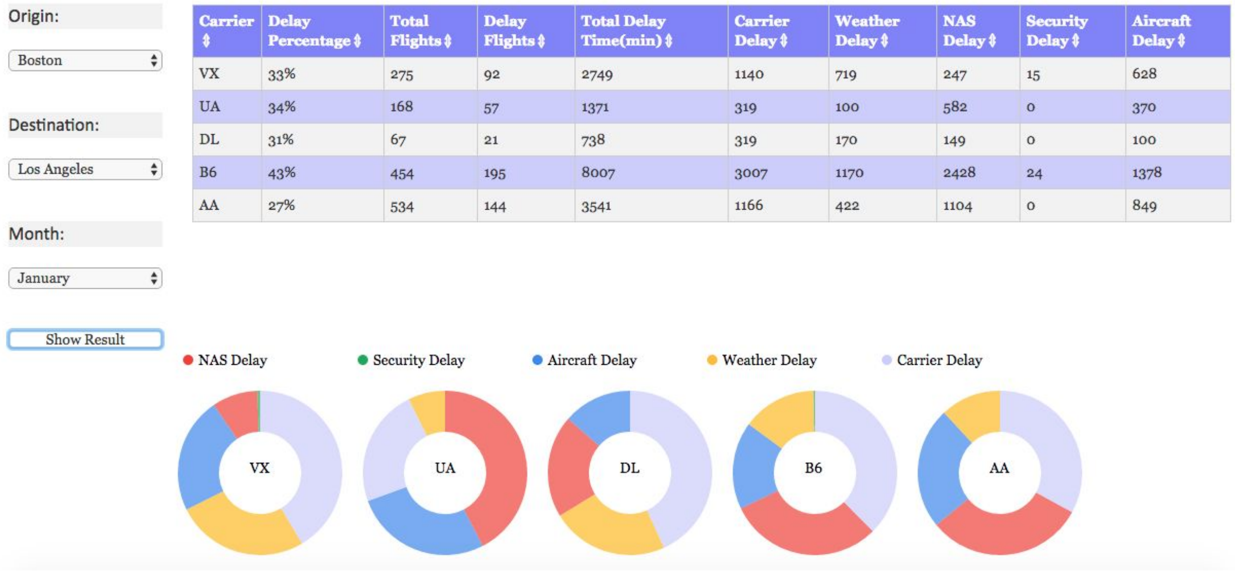


Fig. 4



Origin:

Boston

Destination:

Los Angeles

Month:

January

Show Result

Carrier ⇕	Delay Percentage ⇕	Total Flights ⇕	Delay Flights ⇕	Total Delay Time(min) ⇕	Carrier Delay ⇕	Weather Delay ⇕	NAS Delay ⇕	Security Delay ⇕	Aircraft Delay ⇕
VX	33%	275	92	2749	1140	719	247	15	628
UA	34%	168	57	1371	319	100	582	0	370
DL	31%	67	21	738	319	170	149	0	100
B6	43%	454	195	8007	3007	1170	2428	24	1378
AA	27%	534	144	3541	1166	422	1104	0	849

Fig. 5

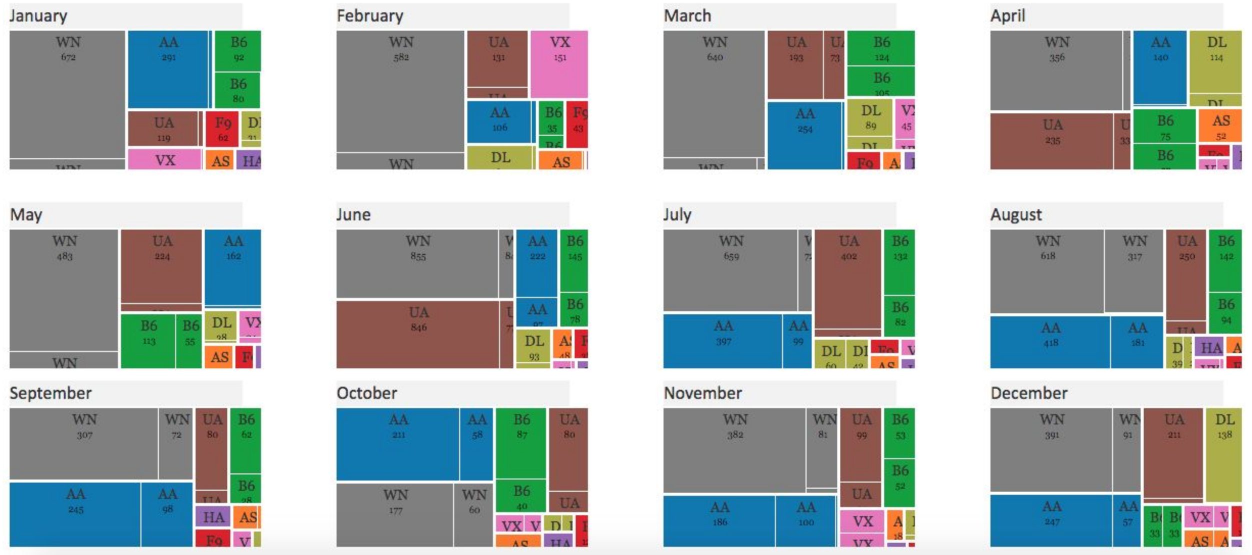


Fig. 6

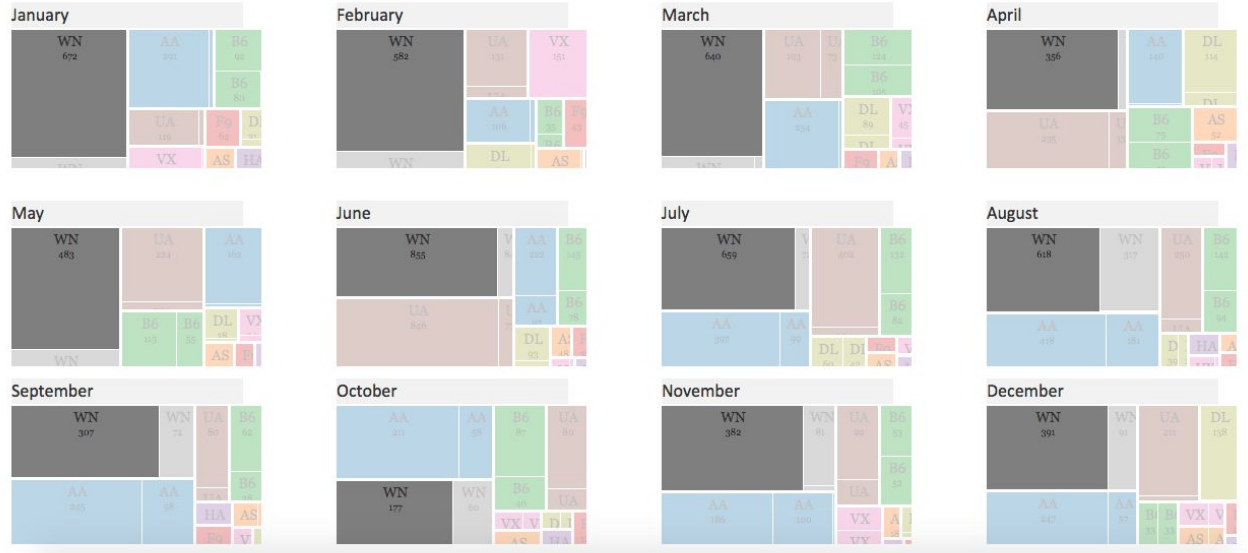


Fig. 7

<u>Evaluation</u>:

From the Parallel Coordinate plot, we learned the ontime performance of nine carriers over year from 2010 to 2015. In detail, Delta Airlines has improved, Hawaiian Airlines, United Airlines and Frontier Airlines have decreased, and other five carriers stay pretty much the same.

From the Map, we found that for the same carrier, for example, American Airlines, the average delay time tends to be high during summer time which is July to September. We think this is because during summer vacation, lots of people especially students travel a lot which can cause flight delay due to too many passengers. Surprising, the delay performance is not that bad during winter months as we expected due to weather cause. That is why we want to explore the causes of delay in detail in the Table next.

From the Table, we found out that delays are somewhat correlated with the busiest seasons namely December and January. This table is not really meant to provide the user with a way to make generalizations about the data. We expect the user to use this table when they have decided on the route and the month they are travelling in.

From the Treemap grid, it was obvious that there would be more cancellations for larger carrier and more cancellations during busy times of the year.

Part 2

This part contains the details of our in person meetings. We include things like sketches, design changes, data processing etc.

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:

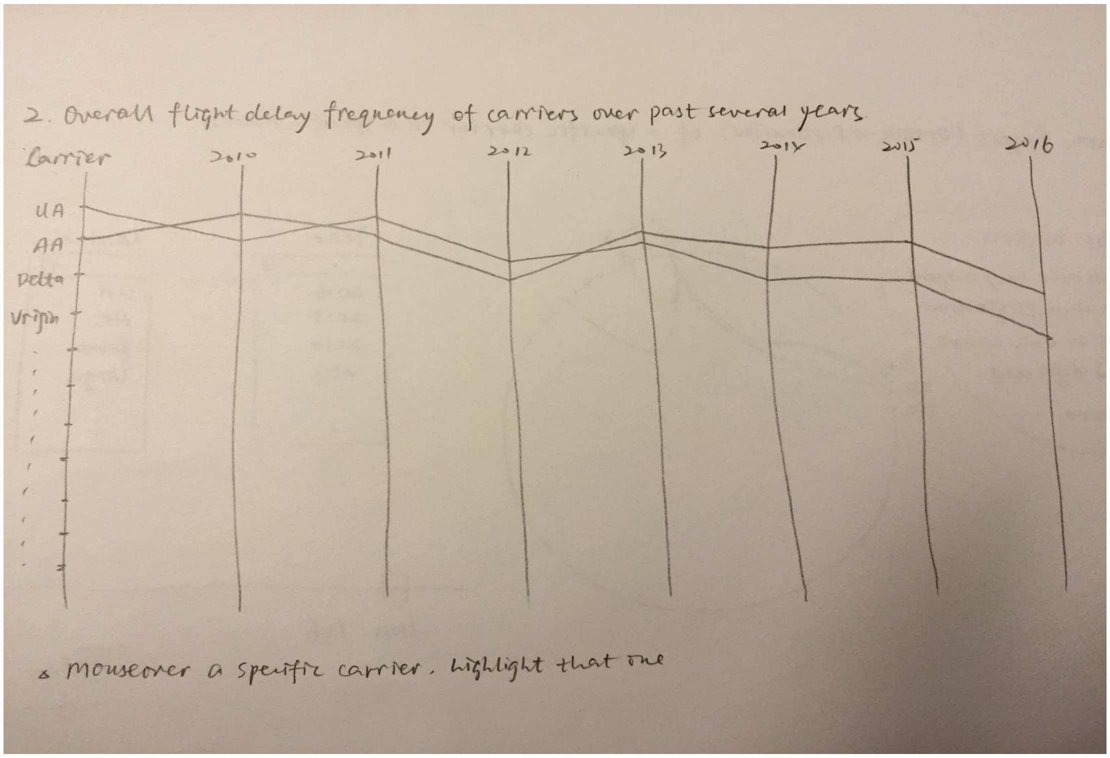
Parallel Coordinates

Chord Diagram

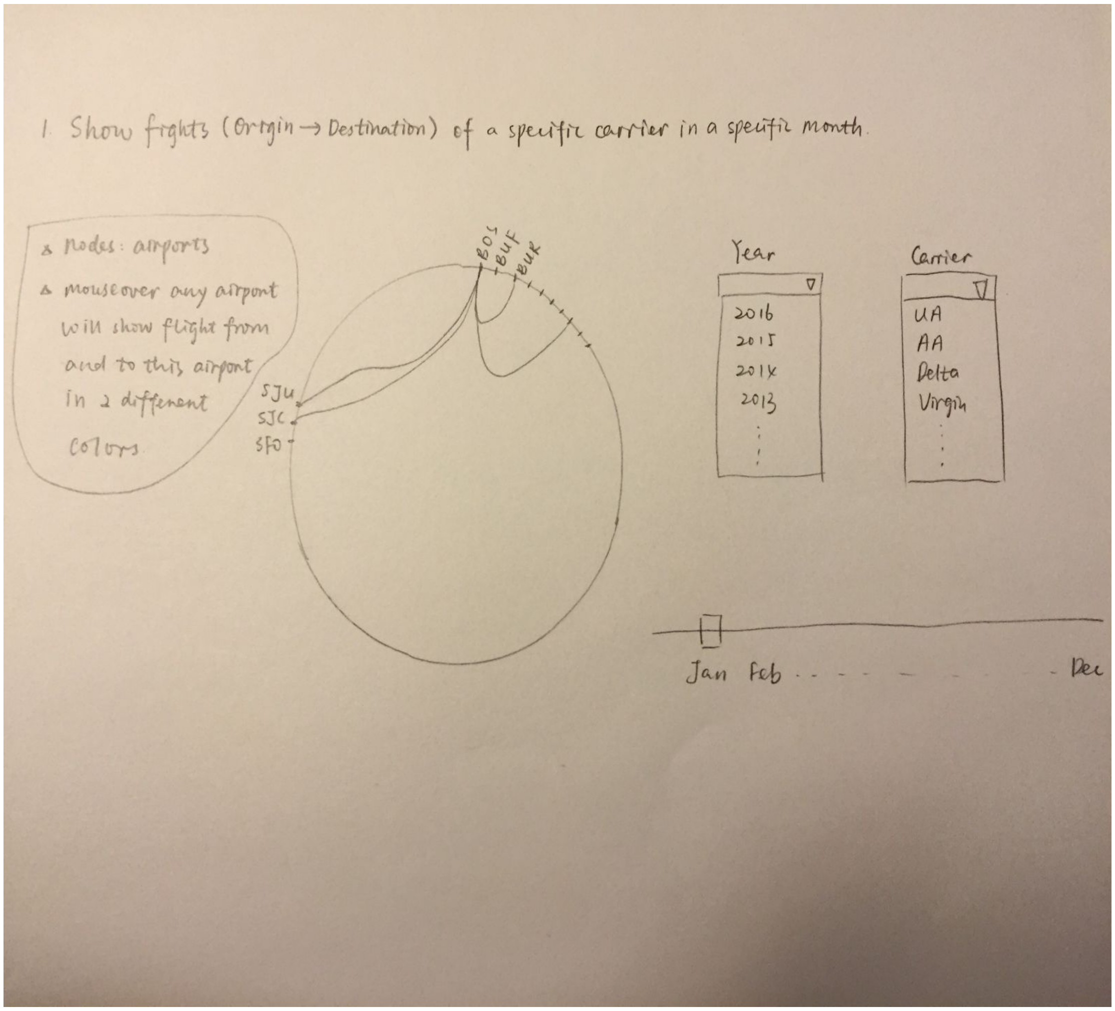
Map showing all the major airport in US

Treemap

We spent this meeting coming up with the project proposal. We discuss in detail, our rationale for these choices in the proposal. Here are some of our sketches.


2. Overall flight delay frequency of carriers over past several years.
Carrier
2010
2011
2012
2013
2014
2015
2016
UA
AA
Delta
Virgin
* Mouseover a specific carrier, highlight that one

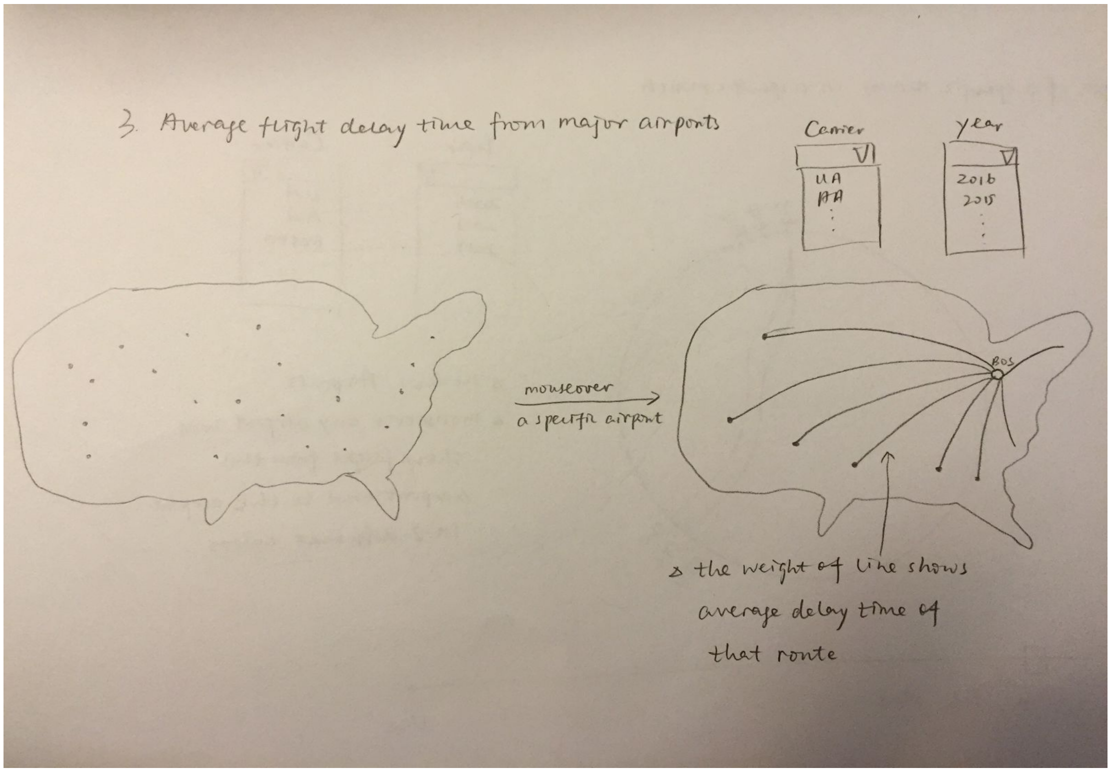
(Figure 1)



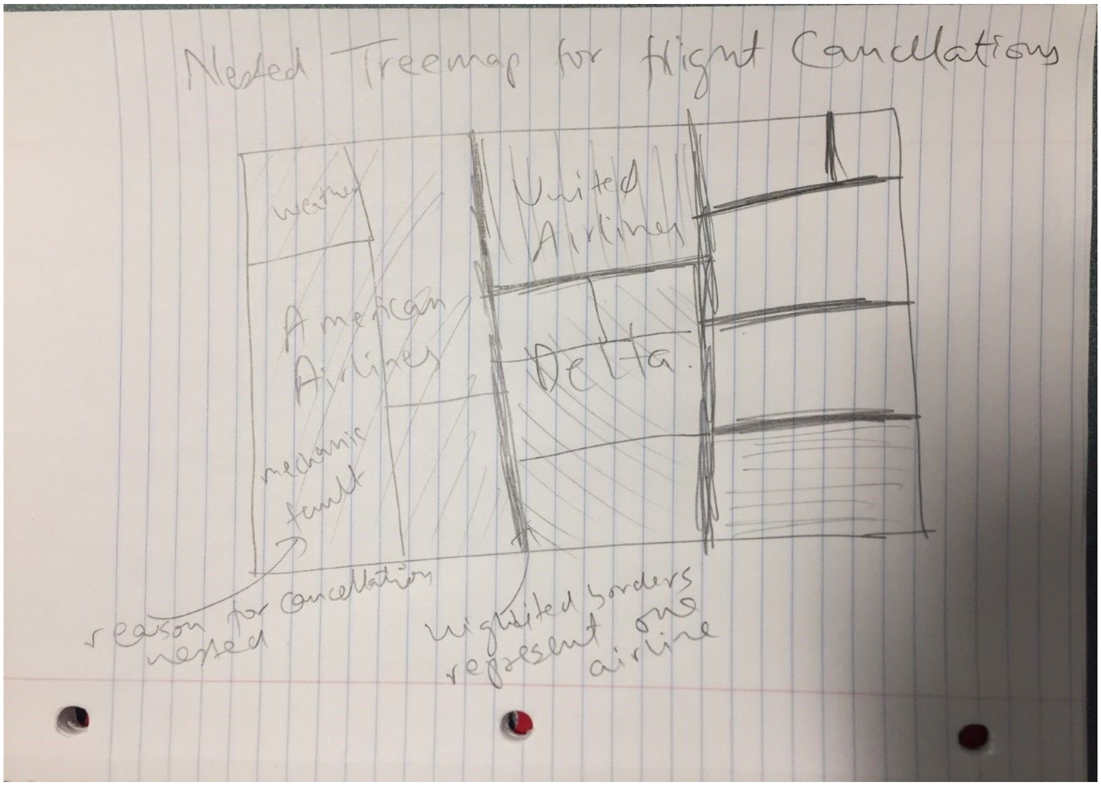
1. Show fights (Origin → Destination) of a specific carrier in a specific month.

- nodes: airports
- mouseover any airport will show flight from and to this airport in 2 different colors.

(Figure 2)



(Figure 3)


Nested Treemap for flight Cancellations
weather
American Airlines
mechanic fault
United Airlines
Delta
reason for cancellation nested
highlited borders represent one airline

(Figure 4)

Date: Nov 8th

We were all busy with the Assignment 5. The dispatching and brushing 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 in the data section above.

Date: Nov 15th

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:

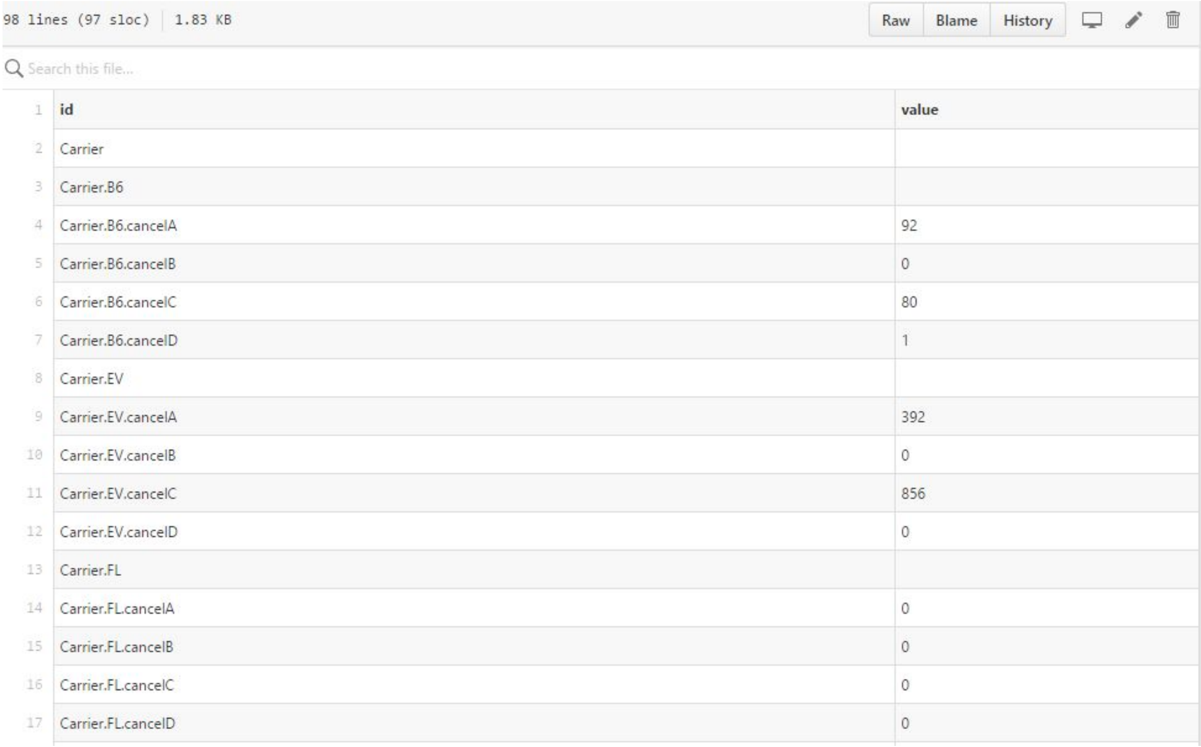
Hamid: Parallel Coordinates Plot and treemap for cancellation data

Huayan: Map of major airports

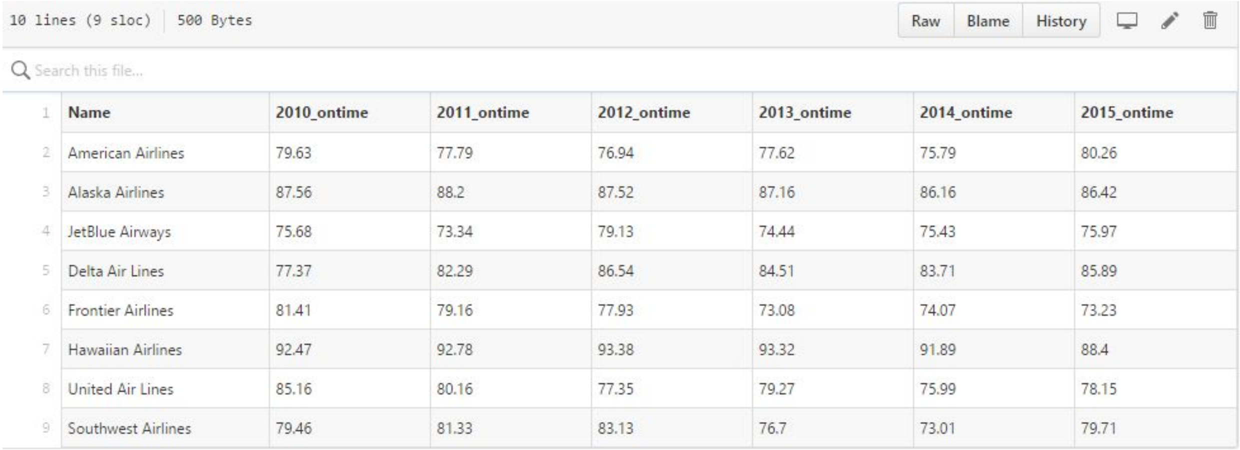
Emma: Table to allow the user to filter airlines

Date: Nov 18th

We met today to discuss our progress and especially one aspect of our project: data. As we have mentioned before the data we had was massive. It would be impossible to simply load all that up in d3.csv. We have been working on some data manipulation to first process the data on our end and then load it into our app. Huayan ran some javascript processes to filter out the useful data. Hamid wrote a python script to aggregate the cancellation data into something that would be easily useable with a treemap. Since we are covering only 30 major airports and 9 major airlines, it makes no sense to process the other data. We will be adding our JS and python code to our submission. Here are two examples of how our simplified data looks:



id	value
Carrier	
Carrier.B6	
Carrier.B6.cancelA	92
Carrier.B6.cancelB	0
Carrier.B6.cancelC	80
Carrier.B6.cancelD	1
Carrier.EV	
Carrier.EV.cancelA	392
Carrier.EV.cancelB	0
Carrier.EV.cancelC	856
Carrier.EV.cancelD	0
Carrier.FL	
Carrier.FL.cancelA	0
Carrier.FL.cancelB	0
Carrier.FL.cancelC	0
Carrier.FL.cancelD	0

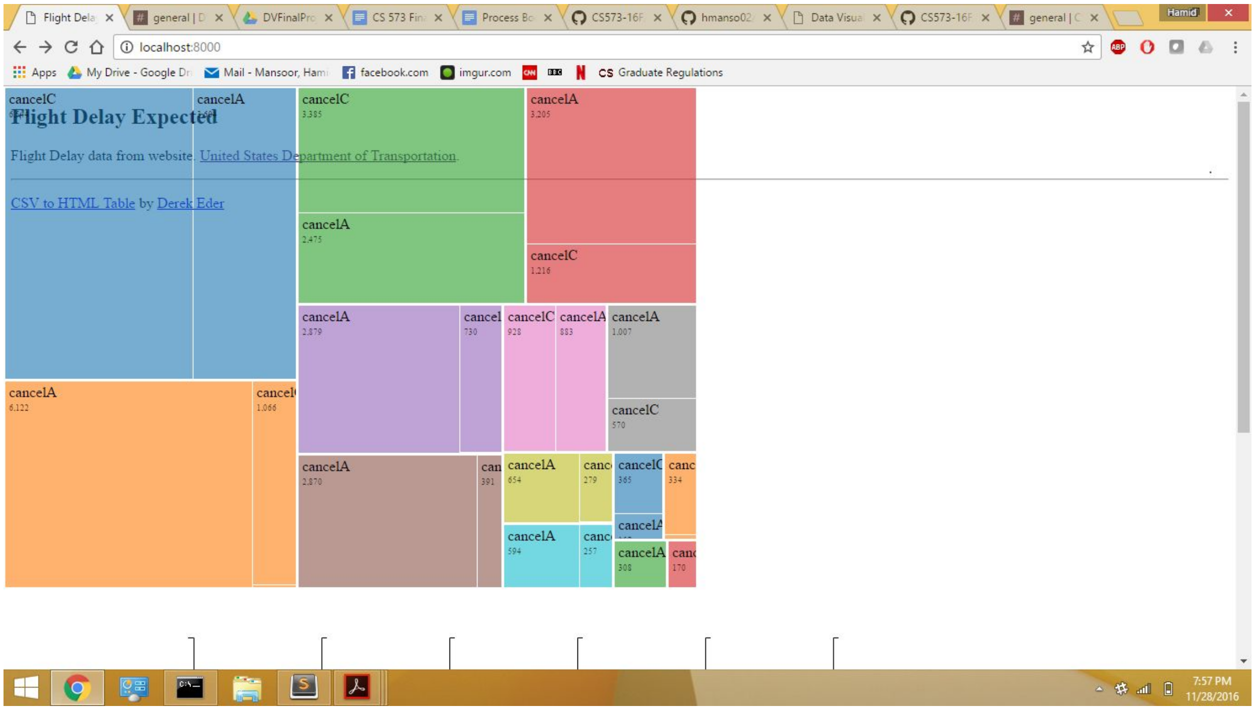


Name	2010_ontime	2011_ontime	2012_ontime	2013_ontime	2014_ontime	2015_ontime
American Airlines	79.63	77.79	76.94	77.62	75.79	80.26
Alaska Airlines	87.56	88.2	87.52	87.16	86.16	86.42
JetBlue Airways	75.68	73.34	79.13	74.44	75.43	75.97
Delta Air Lines	77.37	82.29	86.54	84.51	83.71	85.89
Frontier Airlines	81.41	79.16	77.93	73.08	74.07	73.23
Hawaiian Airlines	92.47	92.78	93.38	93.32	91.89	88.4
United Air Lines	85.16	80.16	77.35	79.27	75.99	78.15
Southwest Airlines	79.46	81.33	83.13	76.7	73.01	79.71

Date: Nov 22nd

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.

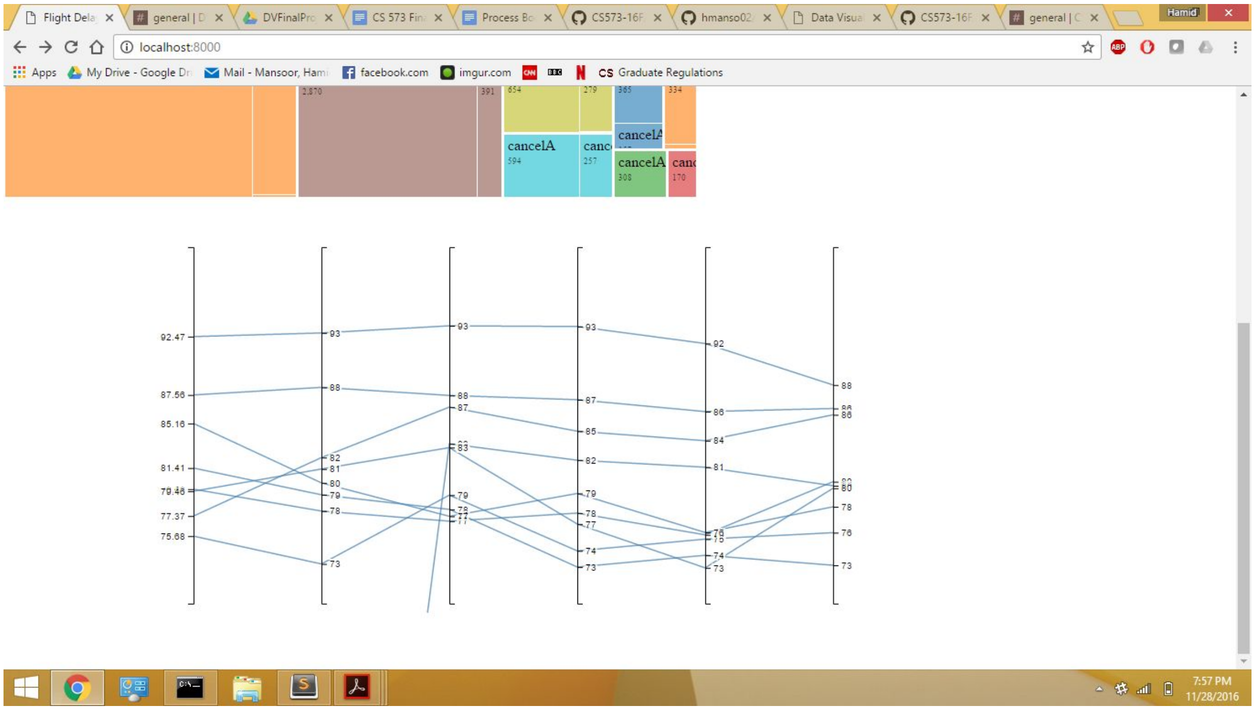
Here are screenshots of where we are at right now:

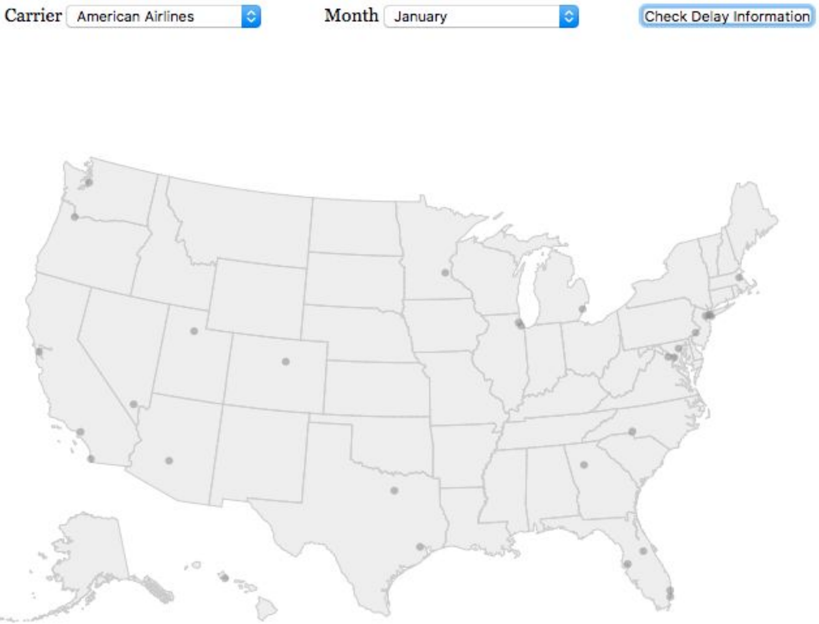


Flight Delay Expected

Flight Delay data from website. United States Department of Transportation.

CSV to HTML Table by Derek Eder

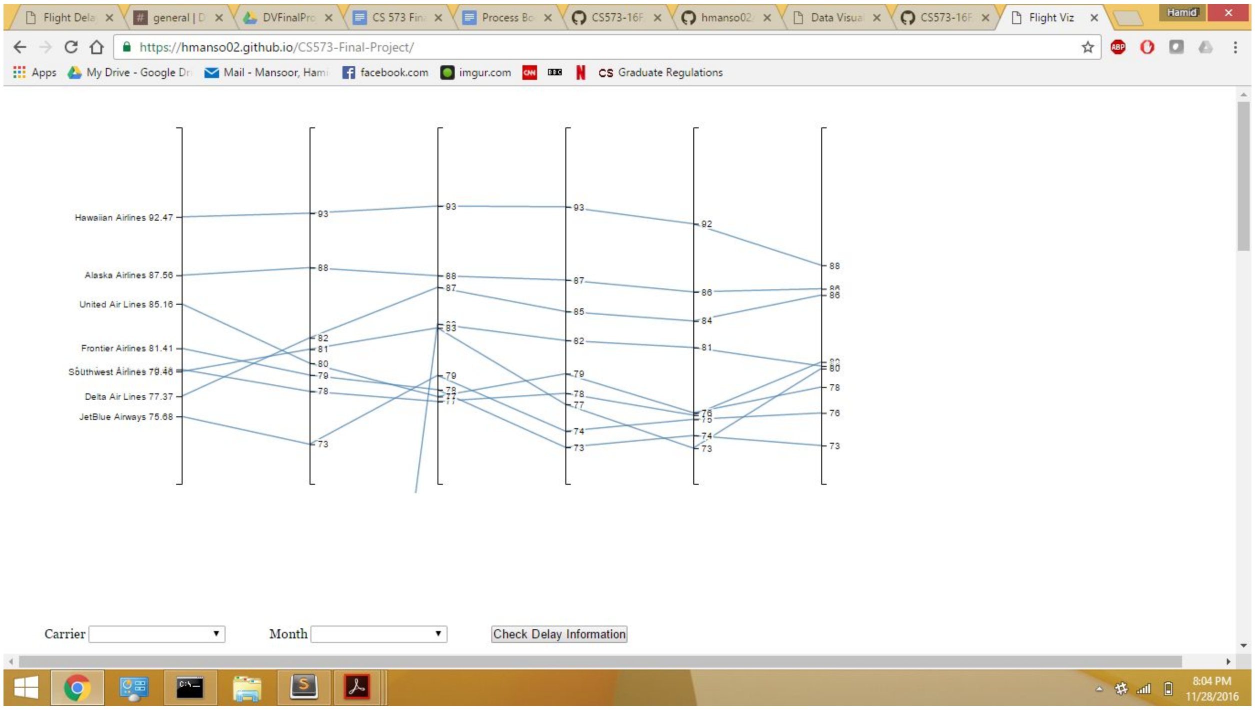


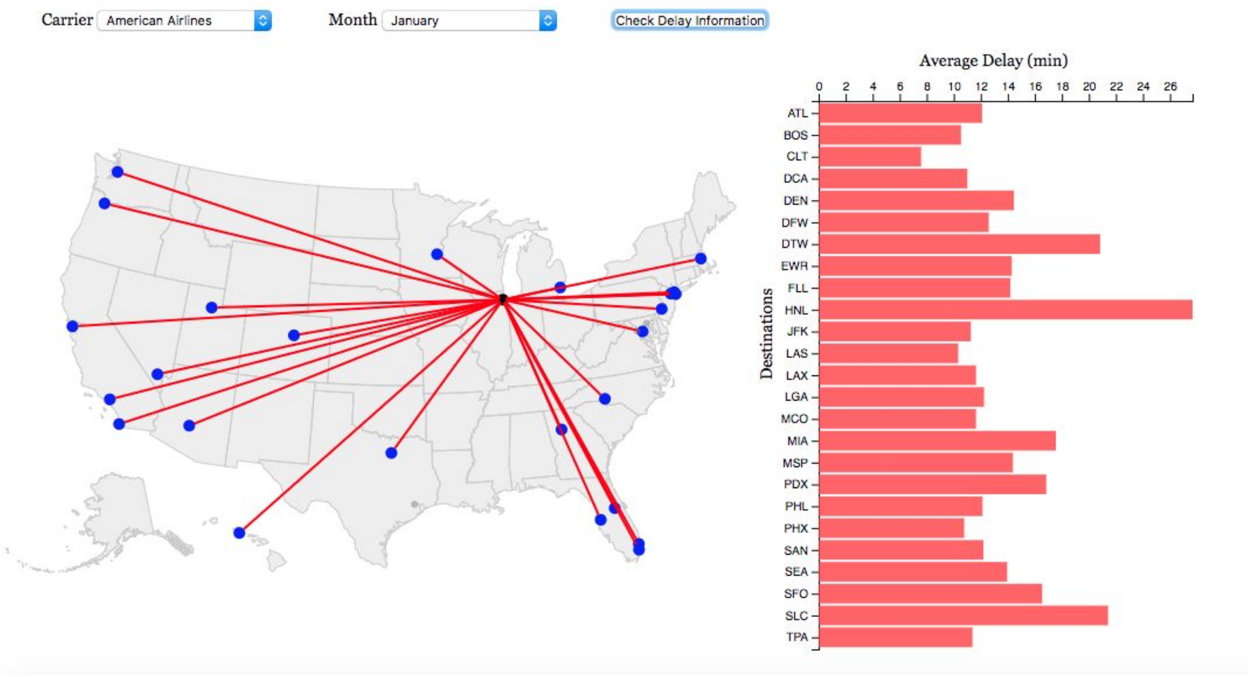


The styling is off in a bunch of places and the table is not functional. We will be working for the rest of the day to fix this.

Date: Nov 28th

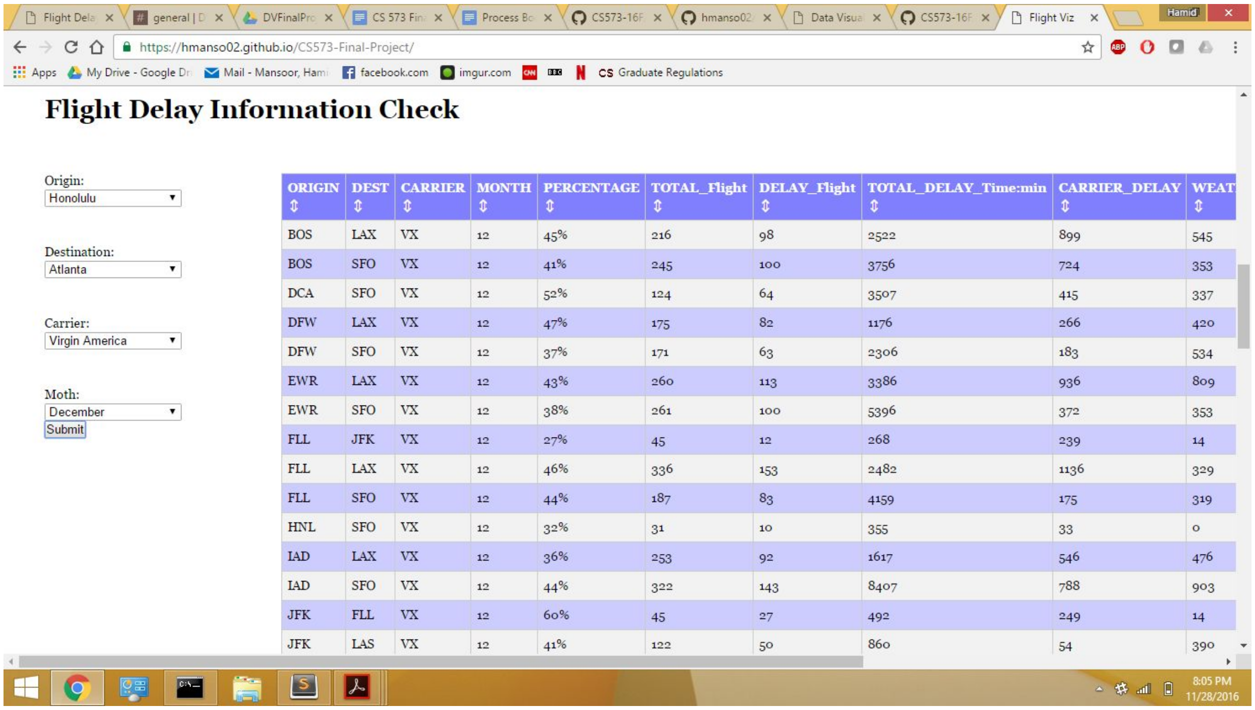
We continued work on the presentation of the prototype. We have added all the required visualizations to one page. We are attaching a few screenshots to show our progress:





Carrier: American Airlines

Month: January



Flight Delay Information Check

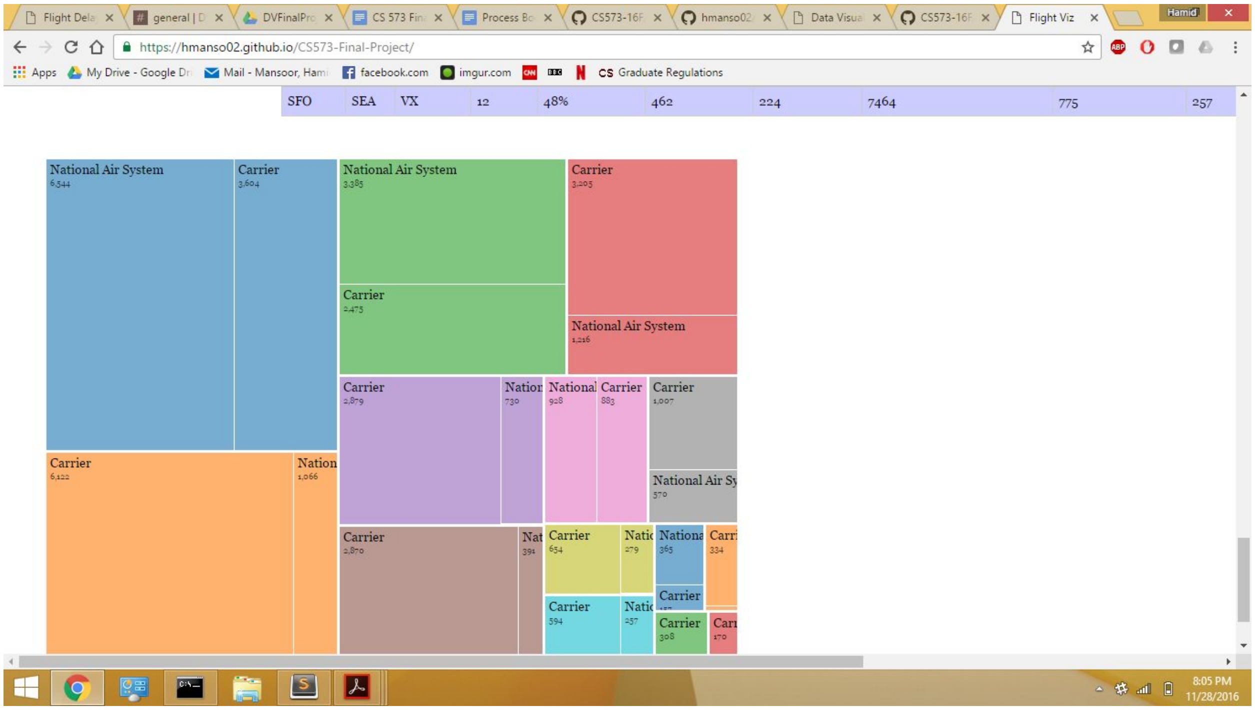
Origin: Honolulu

Destination: Atlanta

Carrier: Virgin America

Moth: December

ORIGIN ⇕	DEST ⇕	CARRIER ⇕	MONTH ⇕	PERCENTAGE ⇕	TOTAL_Flight ⇕	DELAY_Flight ⇕	TOTAL_DELAY_Time:min ⇕	CARRIER_DELAY ⇕	WEAT ⇕
BOS	LAX	VX	12	45%	216	98	2522	899	545
BOS	SFO	VX	12	41%	245	100	3756	724	353
DCA	SFO	VX	12	52%	124	64	3507	415	337
DFW	LAX	VX	12	47%	175	82	1176	266	420
DFW	SFO	VX	12	37%	171	63	2306	183	534
EWR	LAX	VX	12	43%	260	113	3386	936	809
EWR	SFO	VX	12	38%	261	100	5396	372	353
FLL	JFK	VX	12	27%	45	12	268	239	14
FLL	LAX	VX	12	46%	336	153	2482	1136	329
FLL	SFO	VX	12	44%	187	83	4159	175	319
HNL	SFO	VX	12	32%	31	10	355	33	0
IAD	LAX	VX	12	36%	253	92	1617	546	476
IAD	SFO	VX	12	44%	322	143	8407	788	903
JFK	FLL	VX	12	60%	45	27	492	249	14
JFK	LAS	VX	12	41%	122	50	860	54	390



Date: Dec 9th

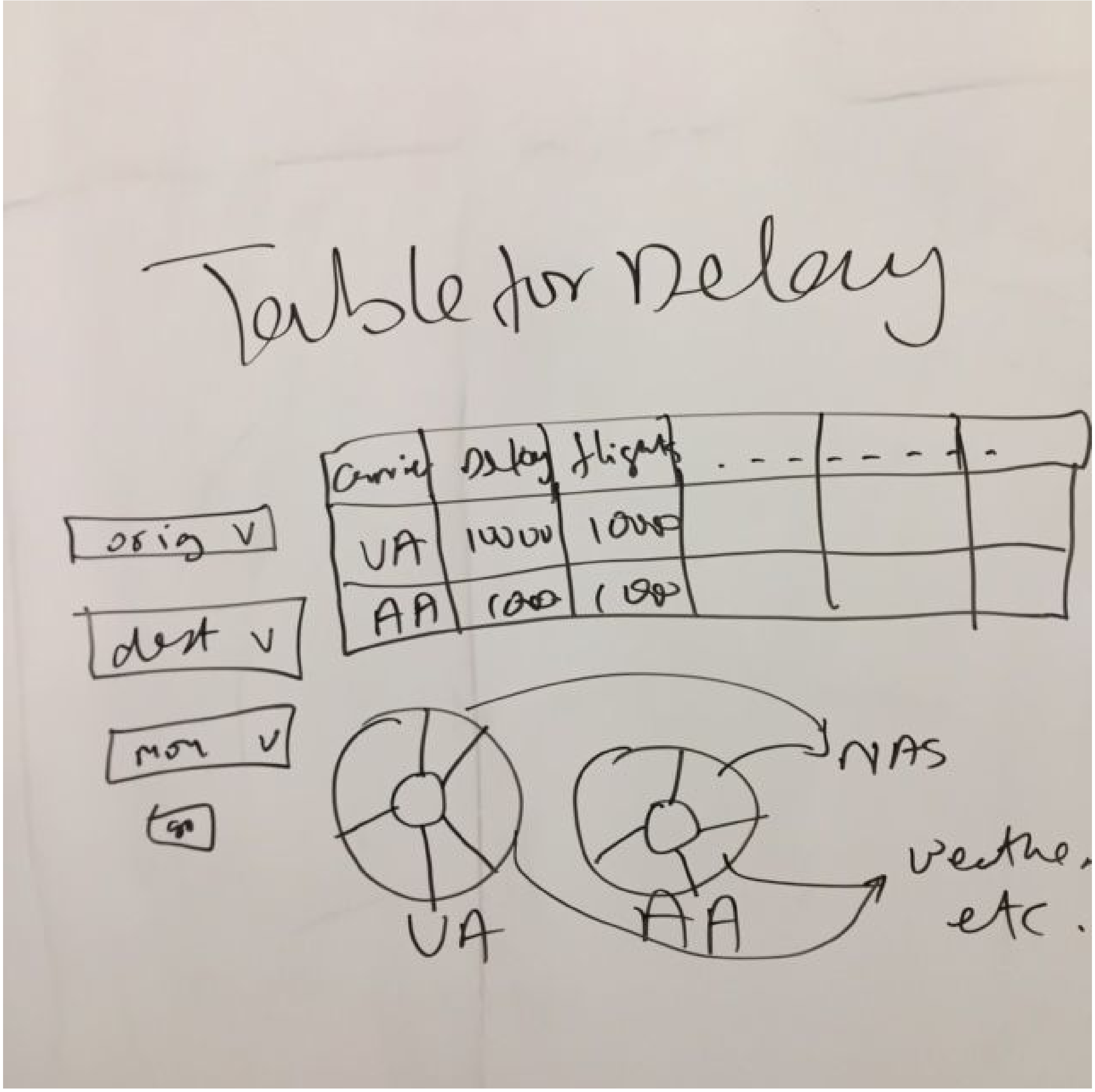
We got some feedback when doing demonstration on Dec 6th in class. And we made several changes on our design and visualizations.

For parallel coordinates, we added mouseover to highlight the carrier selected. We also decided to blur out the rest of the carrier data when a user mouses over a single carrier. In the example below, for the PCP, when the user hovered over Hawaiian Airlines, it highlighted that particular line and blurred out the rest of the data.

We also noticed and important bug with the PCP. There are some tick marks on the axes that get obscured because the intersection point at that particular axis is common between more than one line. We also noticed in the left most axis with the names of the airlines, that the text became illegible in some cases. We tackled that issue by blurring out the tick marks and the associated tick texts for all non selected lines on mouseover.

For the map, first, we added the axises for bar chart at the beginning instead of showing up after user input carrier and month. Second, we removed dropdown list for carrier and month and allow user to select them directly from the visualization so that they can easily compare delay performance among different carriers and different months for a specific carrier. Third, we changed the click to mouseover when selecting origins to make the interaction more smooth and easier for the user. Last, we changed the color palette from red/green referencing on colorbrew website.

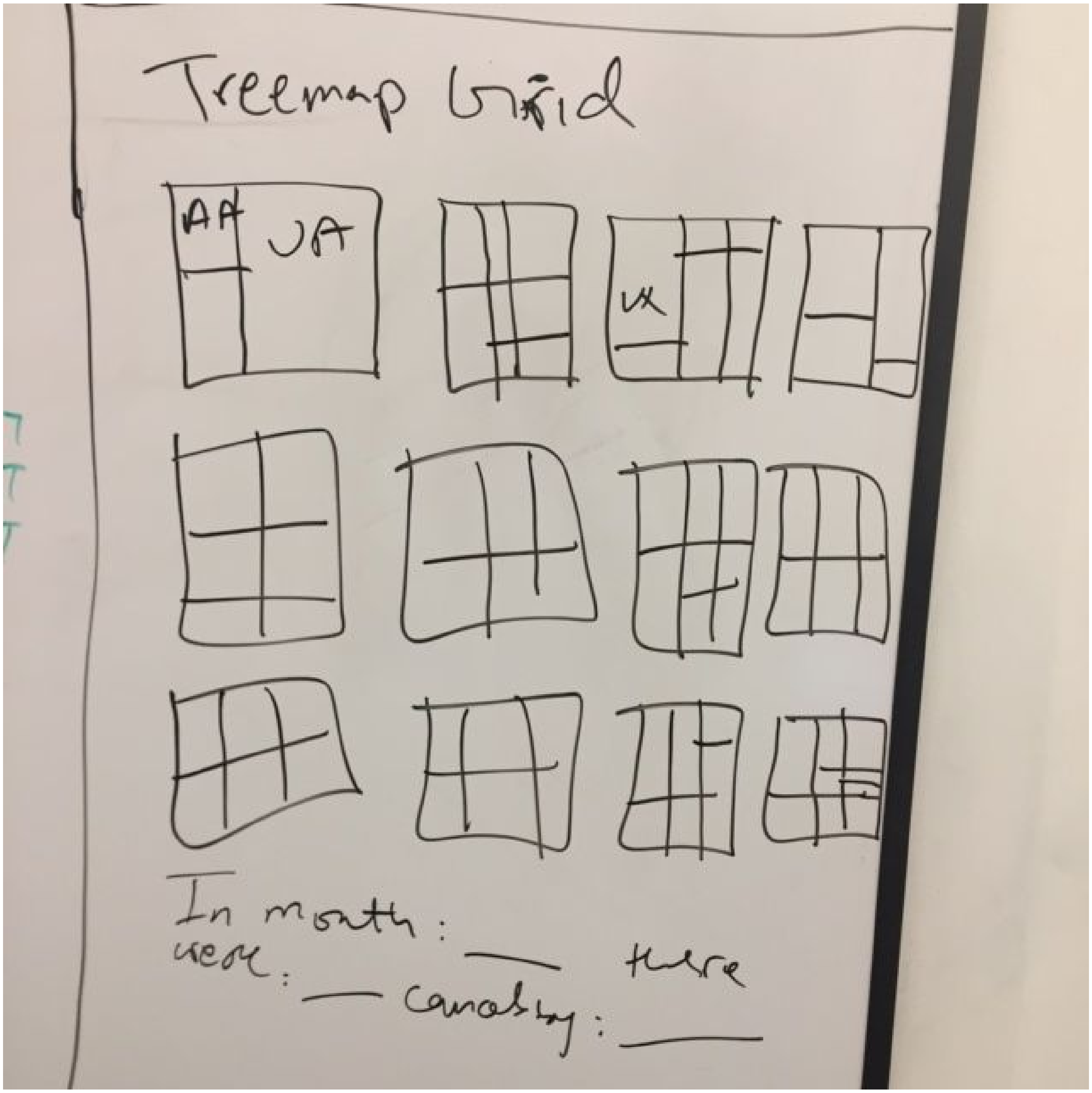
For the table, we deleted some columns that are not useful. We want the users to focus on the proportion of flights that are delayed across multiple carriers. We also thought about dynamically adding donut charts to show the proportion of each type of delay across the carriers. It would help the users figure out if some carriers are more prone to delays that are within their control to improve. Here is the sketch of what we came up with:



Carrie	Delay	flights	. . .	- - -	-
UA	10000	1000			
AA	100	100			

Before we the prototype presentation, we discussed the treemap because it did not seem like it providing very useful information. Users are not particularly concerned about the overall number of cancellations for a number of years. It made more intuitive sense to break up the cancellation data into seasons or months as some months would obviously have a higher number of cancellations in comparison with other months such as December and January because to snowstorms etc.

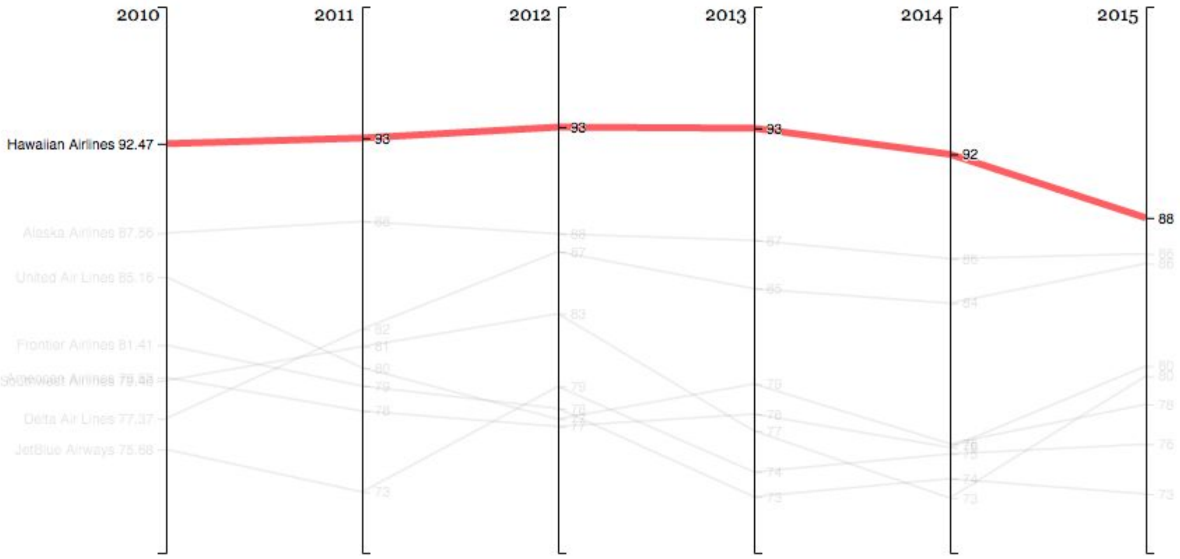
After our presentation, we discussed our idea of adding another layer on top of our current treemap that would allow the user to filter by the month. Prof. Harrison thought that it would be fine to do that. He also suggested a treemap grid for the 12 months. We liked that idea as it would provide a quick glimpse to the user of the performance of all the major carriers across all the months of the year. We drew the sketch of what this would look like:

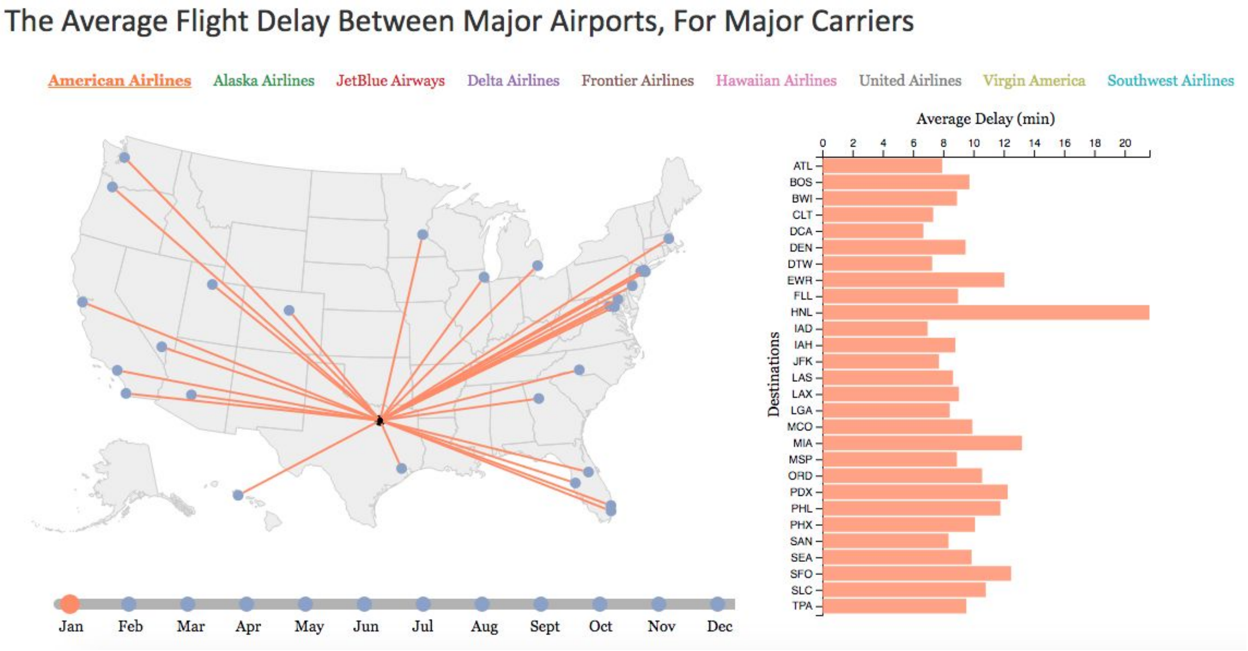


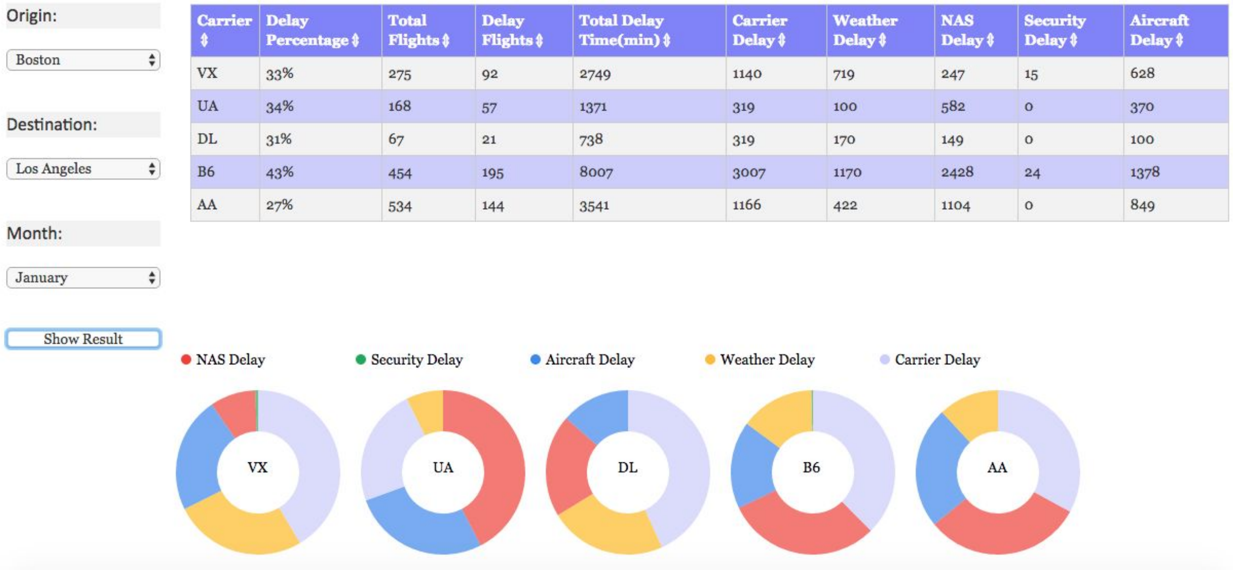
We ended our meeting after we decided on the changes to implement.

Date: Dec 11th

We met today to go over the changes we discussed and implemented after our last meeting about the feedback we received.







Origin:

Boston

Destination:

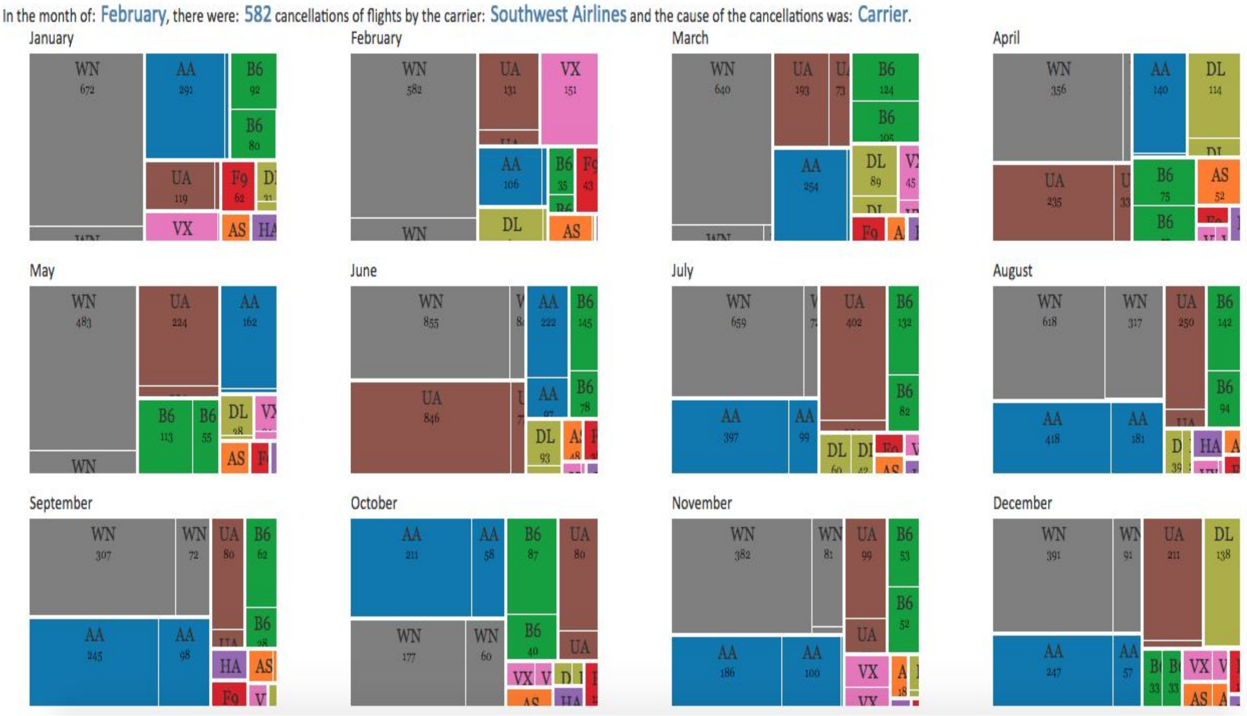
Los Angeles

Month:

January

Show Result

Carrier ⇕	Delay Percentage ⇕	Total Flights ⇕	Delay Flights ⇕	Total Delay Time(min) ⇕	Carrier Delay ⇕	Weather Delay ⇕	NAS Delay ⇕	Security Delay ⇕	Aircraft Delay ⇕
VX	33%	275	92	2749	1140	719	247	15	628
UA	34%	168	57	1371	319	100	582	0	370
DL	31%	67	21	738	319	170	149	0	100
B6	43%	454	195	8007	3007	1170	2428	24	1378
AA	27%	534	144	3541	1166	422	1104	0	849

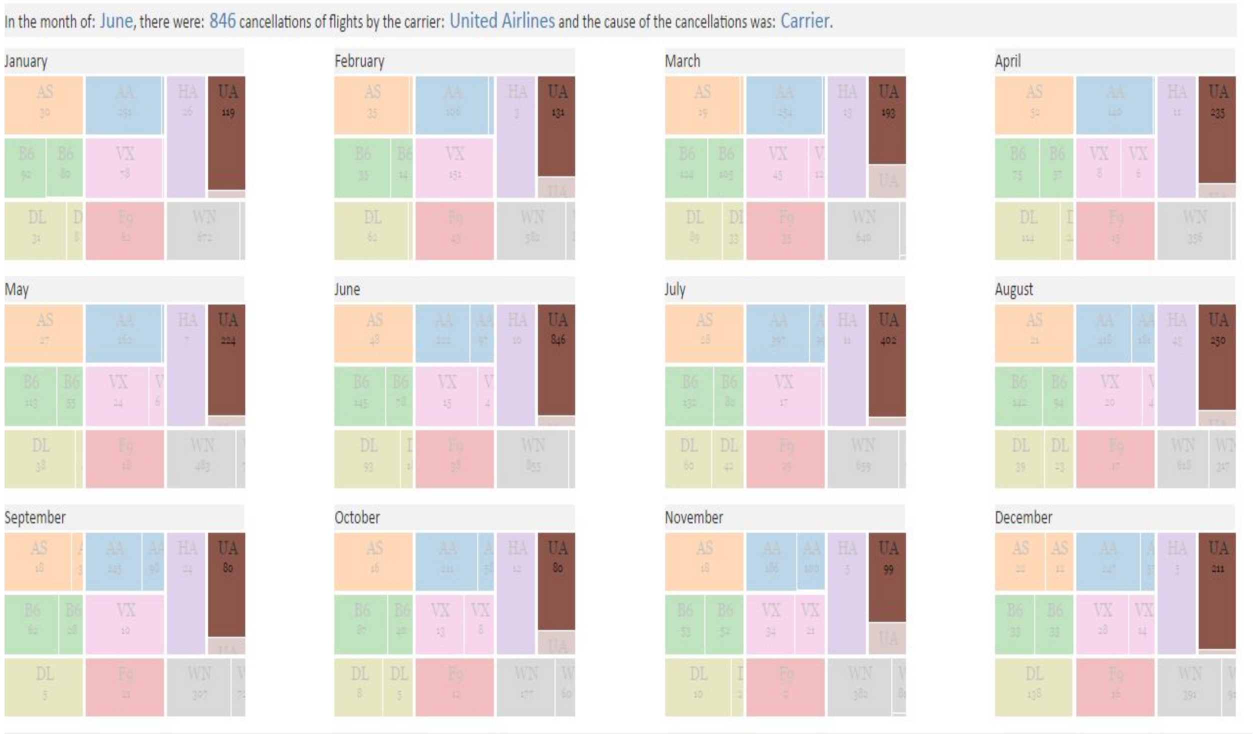


In the month of: February, there were: 582 cancellations of flights by the carrier: Southwest Airlines and the cause of the cancellations was: Carrier.

At last, we'll add some text around the visualizations to help users better understanding how to use our website to get the useful information.

Date: Dec 13th

Most of today as spent on finalizing the changes we had recently made. We added some storyline text to our visualizations to tie them all together and give the user some easy instructions. One major issue we found later after Dec 11th meeting was that the treemap grid was showing the tile sizes for the carriers based on the absolute number of flights. That is not very useful as larger carriers will undoubtedly have larger cancellation numbers. To fix this, we now show the proportion of each cause of cancellations within a carrier. This means that the overall carrier size in the treemap is consistent but the tiles within have varying sizes. Here is a screenshot below:



In the month of: June, there were: 846 cancellations of flights by the carrier: United Airlines and the cause of the cancellations was: Carrier.

We have also made some nice CSS changes to make the app look aesthetically pleasing.

Part 3

This part contains our concluding thoughts on the project and potential future work.

Conclusion

We built an app with multiple visualizations to let users make informed decisions about their travel plans based on the data they are able to explore about flight cancellations and delays across multiple major carriers and major airports. We were able to successfully implement multiple visualizations to facilitate the exploration of flight data. We hope this tool can be used by consumers to improve their travel plans. There are several important limitations and proposed features that we have discussed below.

Future work

One of the biggest limitations to our work was that we only included 6 years, while the data is available for about 31 years. Utilizing all this data could have given us a clearer and more reliable view of the overall nature of this industry. However, it is important to note that the sheer enormity of this data set makes this analysis very time consuming. We also think our visualizations could be more interconnected. This would allow the user to spend less time and gain more important information.