CSE512 | Data Visualization

Assignment 2
Notebook

Aniket Handa
aniket@uw.edu

Exploratory Journey / Analysis

From the start I wanted to work on some question to which the answer is contradictory to the common belief or at least the visualization should be rich and useful. So, I started asking question with a hope to find something that the results into unexpected. I first explored the flight dataset, then went to movie database and then again came back to flight data. Here I only show the exploration of flight dataset just it make it less complicated.

Flight Dataset

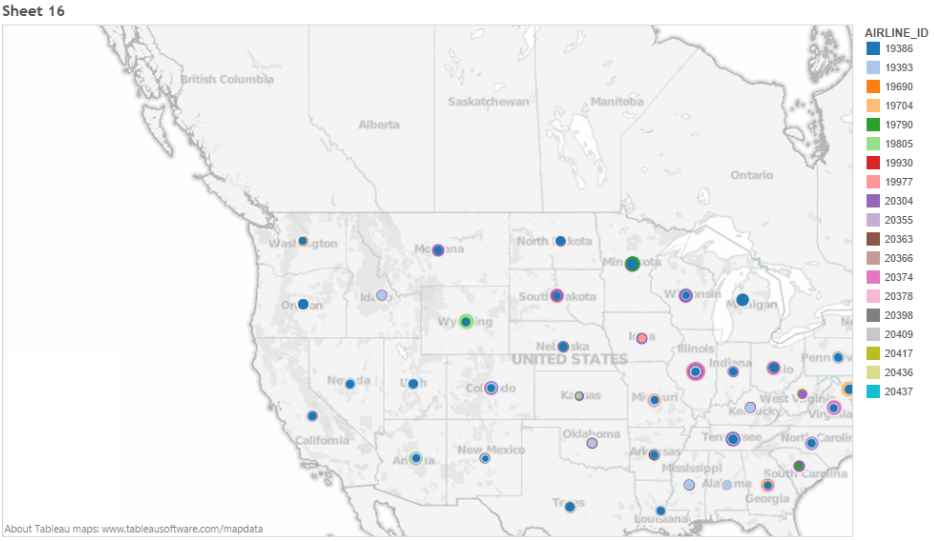
I mostly used Tableau for exploratory analysis. The first question that popped up when I saw the raw Flight data was that:

1) What's the relation between all these Measures? Is Arrival Delay correlated to Departure delay?

   In particular, first I wanted to know if Flight Arrival Delay is correlated to Flight Departure Delay. Though obvious, I just wanted to confirm. And yes indeed Departure Delay is depended upon the Arrival delay. Then I looked at various other delay measures such as due to security, etc. Many of them were not richly populated and had null values. For finding patterns I plotted scatters plot matrix of Measures that made sense together and also analysed the trend line. The trend line gave an overview of the relationship.

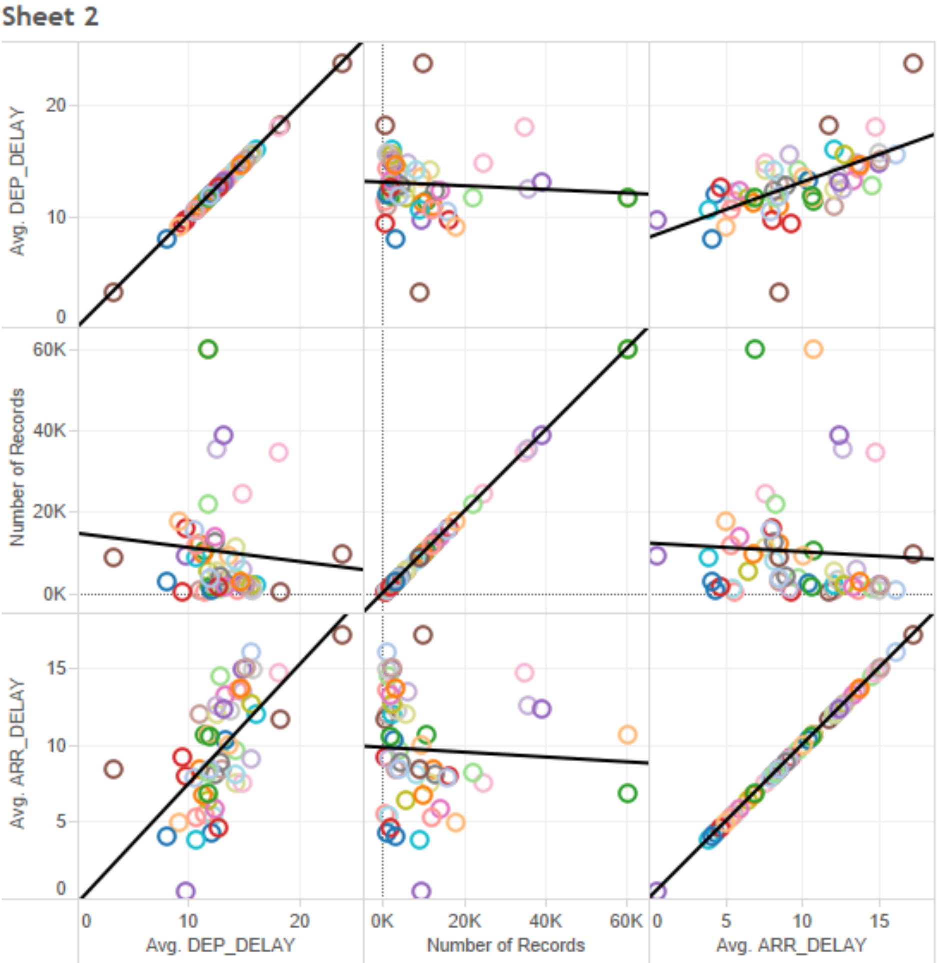
2) Which State has the highest number of flights?

   As I was analysing all these measures I realized the trend changes sharply if categorized state wise. This led me to the above question, which was trivial to answer. But was interesting to see California, Texas, Georgia, Florida to have more flight records than on the east side.



3) Do States with high number of records suffer from more delays?

I combined the above two questions. This resulted in predictable conclusion that most states with high flight traffic data do incur more delays than the ones with light flight traffic.



4) Are some Dimensions redundant?

Then I hopped on to discover the Dimensions. While exploring them I couldn't figure out the difference between few columns. So, I just plotted them against each other to see the points where they differ. I found that some of the columns are redundant, as in Unique Carrier, Carrier and Airline ID are nominal values, which remain same between two distinct data points. A quick background search gave the answer that Airline ID is actually the DOT ID issued to the Airline carrier for the plane. As the plane carrier doesn't change very often these fields remained constant in the month of December 2009.

5) Is Airtime and delays related?

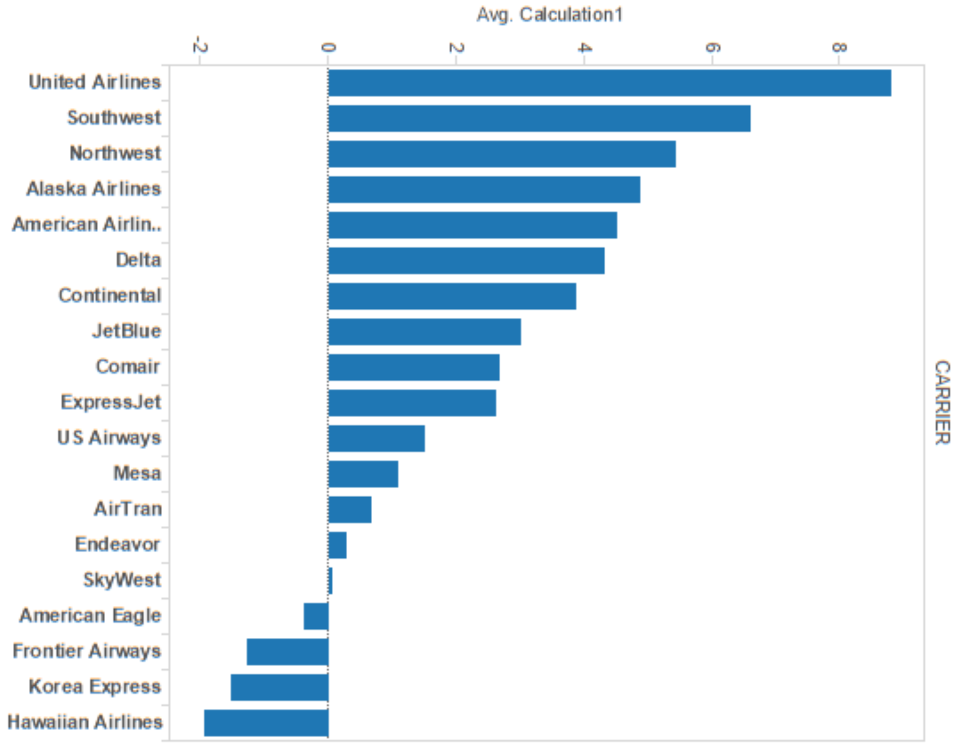
No, I couldn't find a compelling correspondence between any measure of delay and airtime of a flight.

6) Can we directly relate Carrier column with actual airline company?

This column was pretty comprehensive and didn't suffer from any ambiguities. There I could directly Alias all Carriers with actual airline name. I got this data by querying IATA.org.

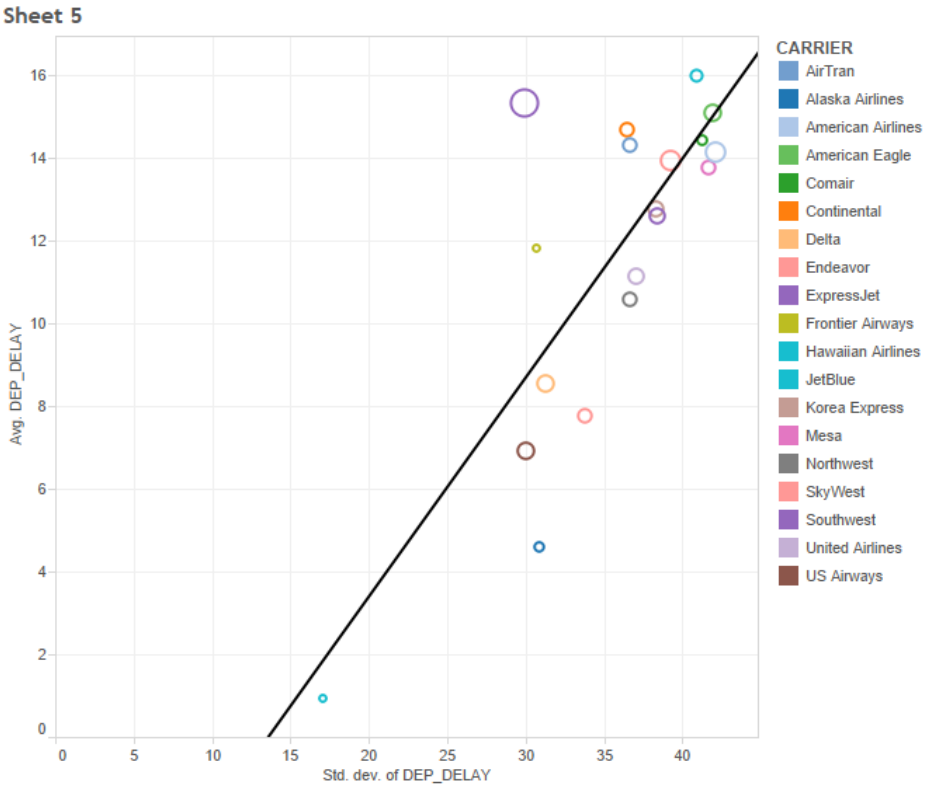
7) Which all flights have large difference in Arrival delay and Departure delay?

After getting flight names I was curious to see, which all airline companies introduce more delays during flight rather than covering up the time. For this I used a calculated field. The results showed that United Airlines added the most delay while Hawaiian Airlines was the best at covering up the lost time in case of a delayed flight.



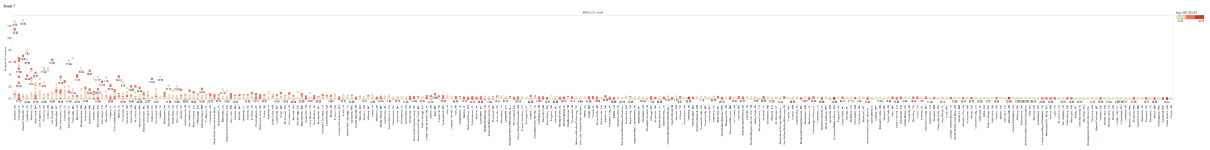
8) What is the std. deviation of departure delay? Does it relate to average departure delay?

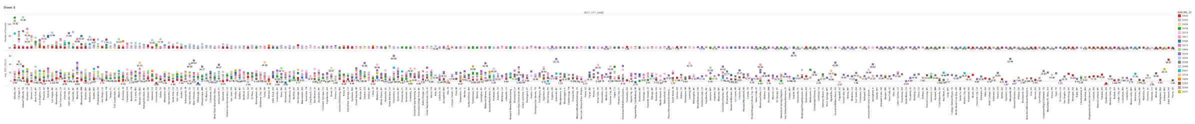
It was interesting to find that as the average departure time increase there deviation of data also increases with respect a to particular airline.



9) How are delays at the city level?

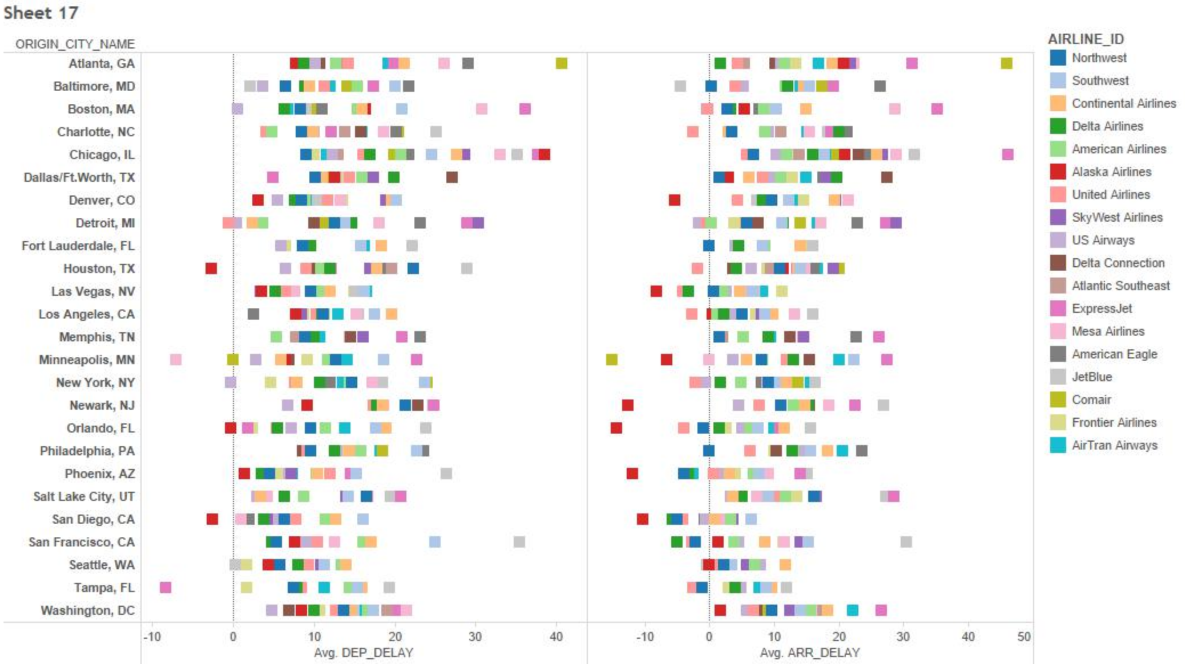
After having a good overview of delay information at the state level I went to city level. The list was long but interesting, as information about average delay can be useful to many.





10) But, How can I fit so many cities in a readable manner?

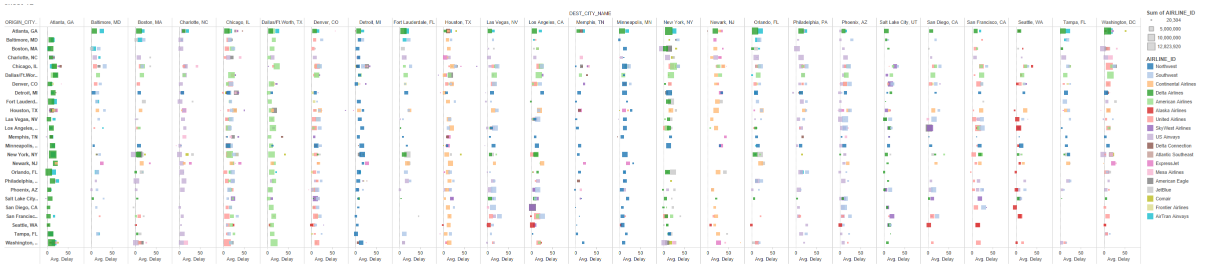
I tired really hard to come up with various techniques to fit the information in small space in a human interpretable manner, but couldn't find it exactly. What I did was I took top 25 busiest cities out of all for further analysis. For this I use filters by using 'Sets' feature in Tableau to select the top five cities with most number of flight records.



Average of DEP_DELAY and average of ARR_DELAY for each ORIGIN_CITY_NAME. Color shows details about AIRLINE_ID. The data is filtered on In / Out of Set 1 and In / Out of Set 2. The In / Out of Set 1 filter keeps In. The In / Out of Set 2 filter excludes Out.

11) Does it vary according to arrival destination?

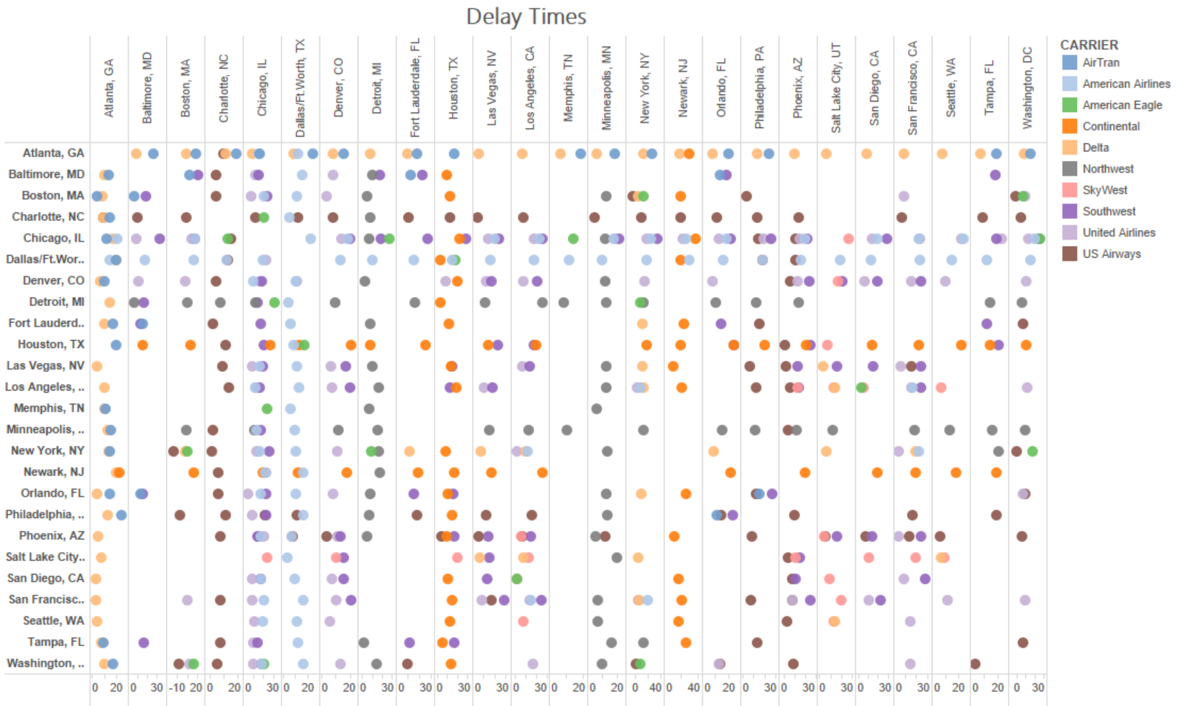
There isn't much point in knowing the delay only w.r.t origin city and not according to destination city, as the values varies a lot. So I made a 2D matrix of Origin and Destination city with delay times. Note, we use departure delay as departure and arrival were found to be quite correlated. The visualization wouldn't change much if we use arrival delay instead.

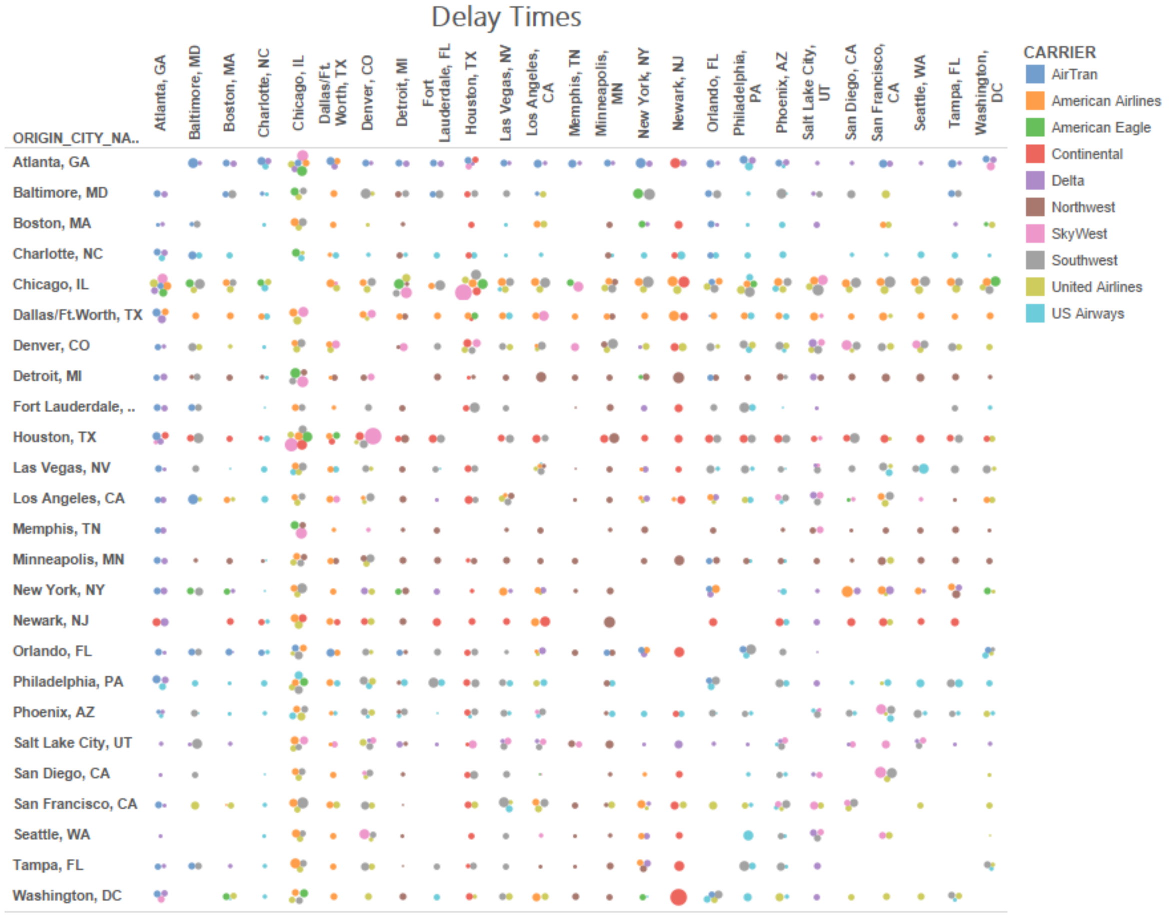


Average of DEP_DELAY for each ORIGIN_CITY_NAME broken down by DEST_CITY_NAME. Color shows details about AIRLINE_ID. Size shows sum of AIRLINE_ID. The data is filtered on In / Out of Set 1 and In / Out of Set 2. The In / Out of Set 1 filter keeps In. The In / Out of Set 2 filter excludes Out.

12) How to fit and make it expressive?

The visualization was again too big even after filtering with top 25 busiest cities. I tried various forms but wasn't happy with the results. I wanted the end result to be easily understood.

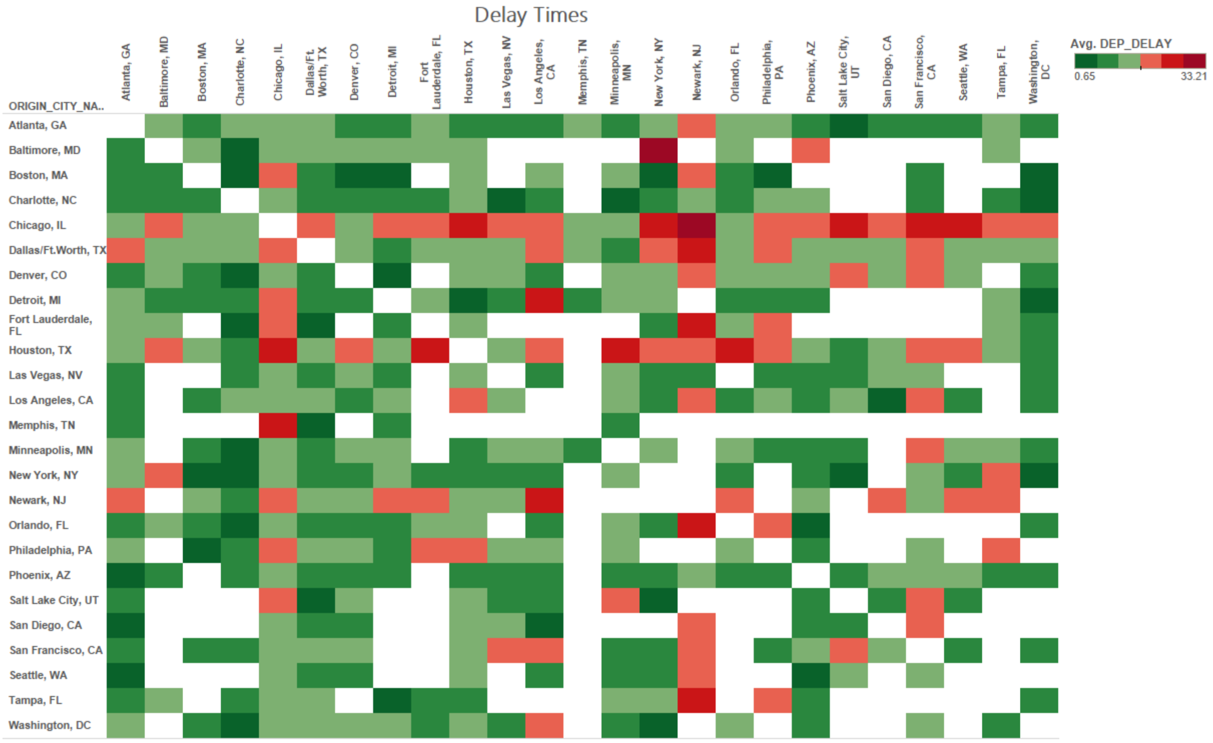




CARRIER (color) and average of DEP_DELAY (size) broken down by DEST_CITY_NAME vs. ORIGIN_CITY_NAME. The data is filtered on In / Out of Set 1, In / Out of Set 2, count of CARRIER and In / Out of Set 3. The In / Out of Set 1 filter keeps In. The In / Out of Set 2 filter excludes Out. The count of CARRIER filter includes values greater than or equal to 0. The In / Out of Set 3 filter excludes Out.

13) How about just average delay between these cites and not carrier information?

This got me to this heat map.



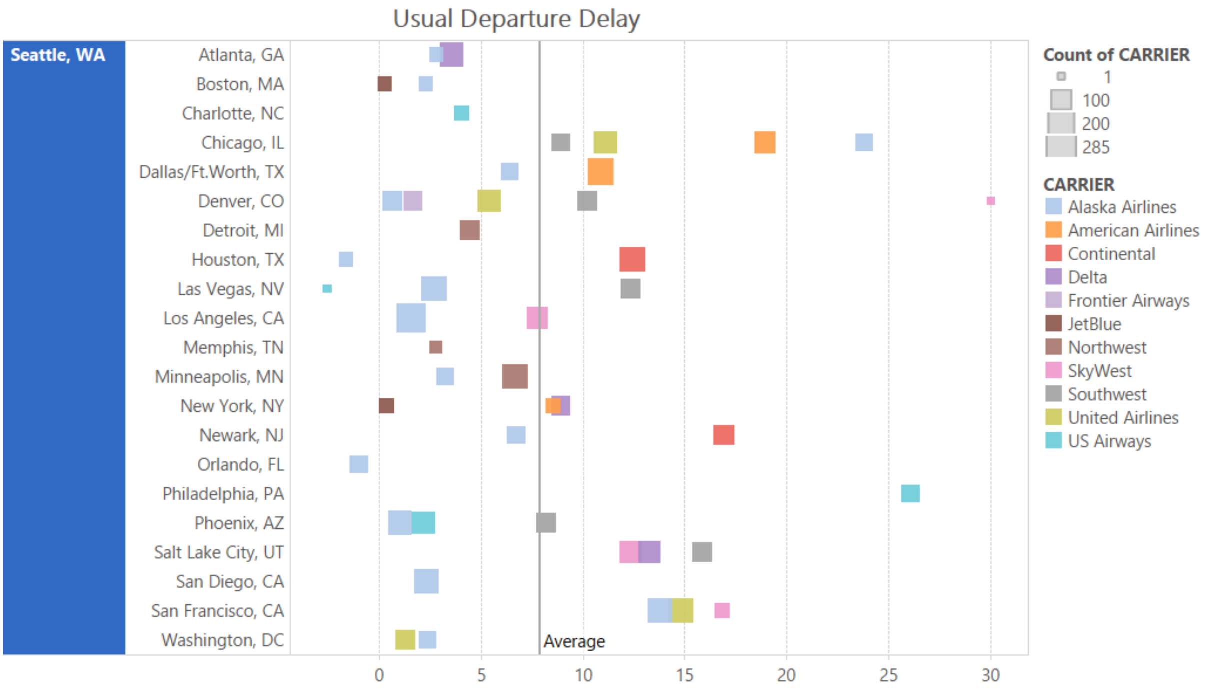
Average of DEP_DELAY (color) broken down by DEST_CITY_NAME vs. ORIGIN_CITY_NAME. The data is filtered on In / Out of Set 1, In / Out of Set 2, count of CARRIER and In / Out of Set 3. The In / Out of Set 1 filter keeps In. The In / Out of Set 2 filter excludes Out. The count of CARRIER filter includes values greater than or equal to 100. The In / Out of Set 3 filter excludes Out.

One can clearly see in the above visualization that Newark, NJ receives quite many delayed flights where as Chicago, IL departs many delayed flights. No wonder at the intersection of two, where the flight is from Chicago to Newark its deep red indicating a long delay from usual schedule.

This map was interesting, but I was asking the question from a perspective of frequent flyer who wishes to know which airline to board which results in least delayed time.

14) Which flight should I board for the odds of incurring least delay at a particular Airport?

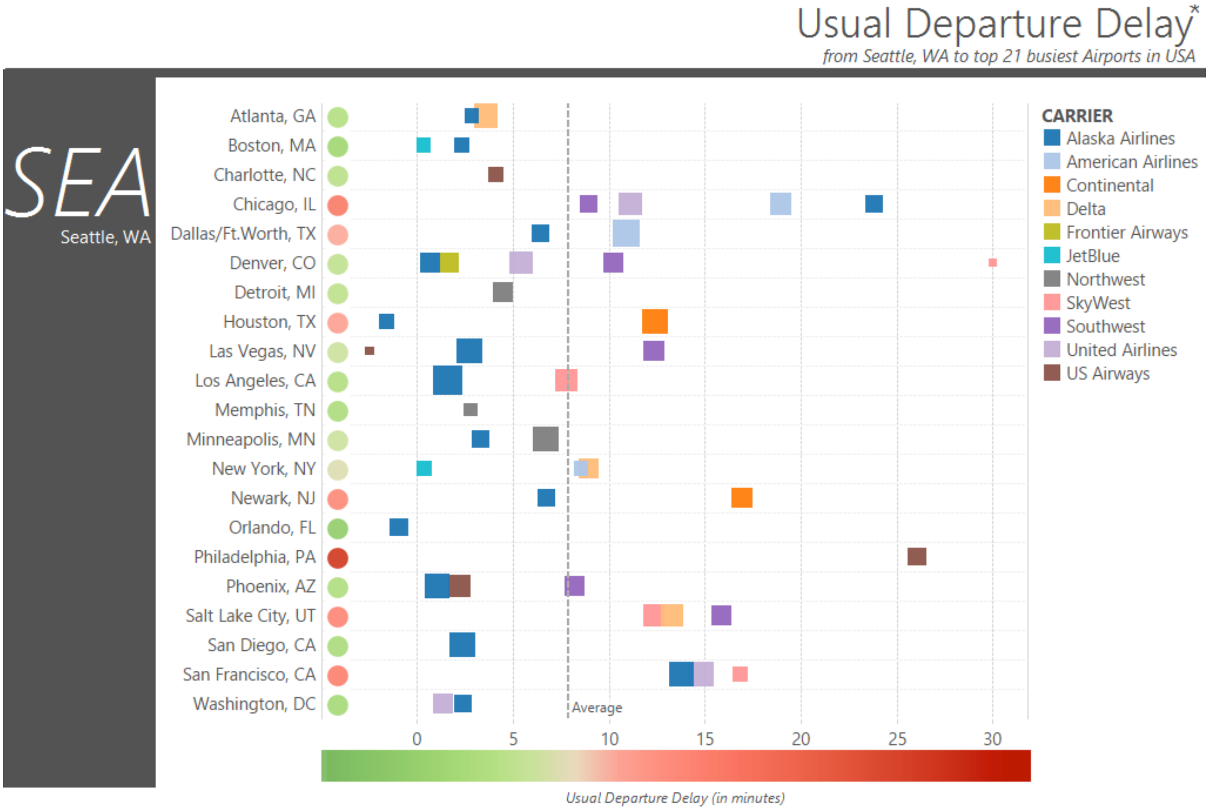
This question was intriguing as it would be extremely helpful for people deciding upon which airline to choose. This question because my final question. I choose Seattle to answer this question and came up with the following rough visualization, which I later refined.



This visualization could answer given the destination city that which airline would be the best bet if I want to incur least delay. I liked the fact that earlier heat map could give an overview of all flights departing. Therefore I added the features of the heat map to come up the final visualization.

Final Visualization

Q. Which flight should I board for the odds of incurring least delay at a particular Airport?



The visualization is designed for frequent travellers who wish to quickly pick a carrier given an origin city, and where departure delay matters. Here the display is shown with respect to the origin city as Seattle but it could be easily be made for any other city. This could be used either at the origin city airport or at travel booking websites, facilitating better time schedules for travellers. For example, if I want to go to Chicago I might not rely on Alaska Airline whereas if I want to go to Houston the same Airline might be a reliable choice. A total number of 52 cities were reached from Seattle via flights (Dec, 2009). But, for not making the visualization too complicated, it only shows top 21 busiest airports from the city. The list of top 21 busiest airport is determined by the total number of flight records each city holds. The cities are listed in alphabetical order for a quick lookup. Average Departure delay is shown below, with a bar running from green to red. The central colour of the bar is at the average (which is 7.88 minutes for Seattle). The coloured bar serves the dual purpose of a colour redundant scale for departure delay and also serves as a key for circles in front of every city. The circles represent the overall departure delay for the particular city, calculated by averaging out all the flight data to that city. One can quickly discern that flights to Chicago, IL and Philadelphia, PA usually run late from Seattle. Also flights from Denver are not that delayed except for an outliner by SkyWest. One can also find interesting patterns by overviewing the data; like the pattern formed by blue squares gives a feeling that Alaska Airlines depart usually on time. The size of the squares represents the number of flights by that airline. It can also be seen by size of squares that Alaska Airlines is big in Seattle. No wonder Seattle is primary hub of Alaska Airline.