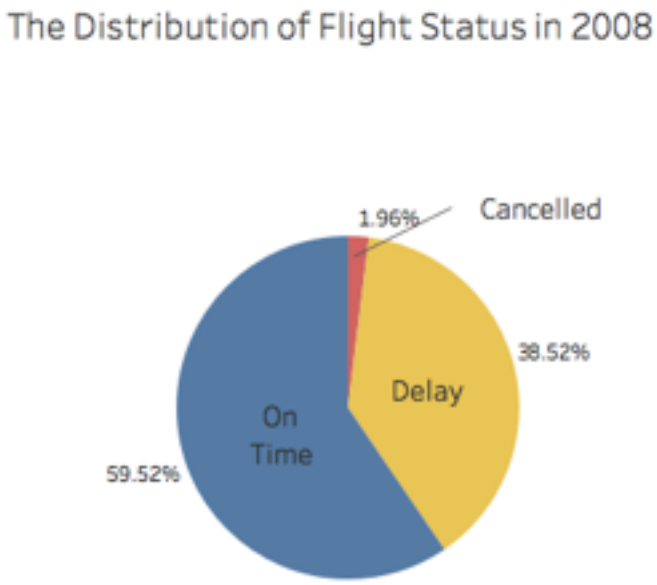
The Performance of Flight Delays In 2008

by Esther Xu
Sep 2017

1. Summary

The data are provided by Research and Innovative Technology Administration (RITA) and Federal Aviation Administration (FAA). They consists of flight arrival and departure details for all commercial flights within the USA in 2008.

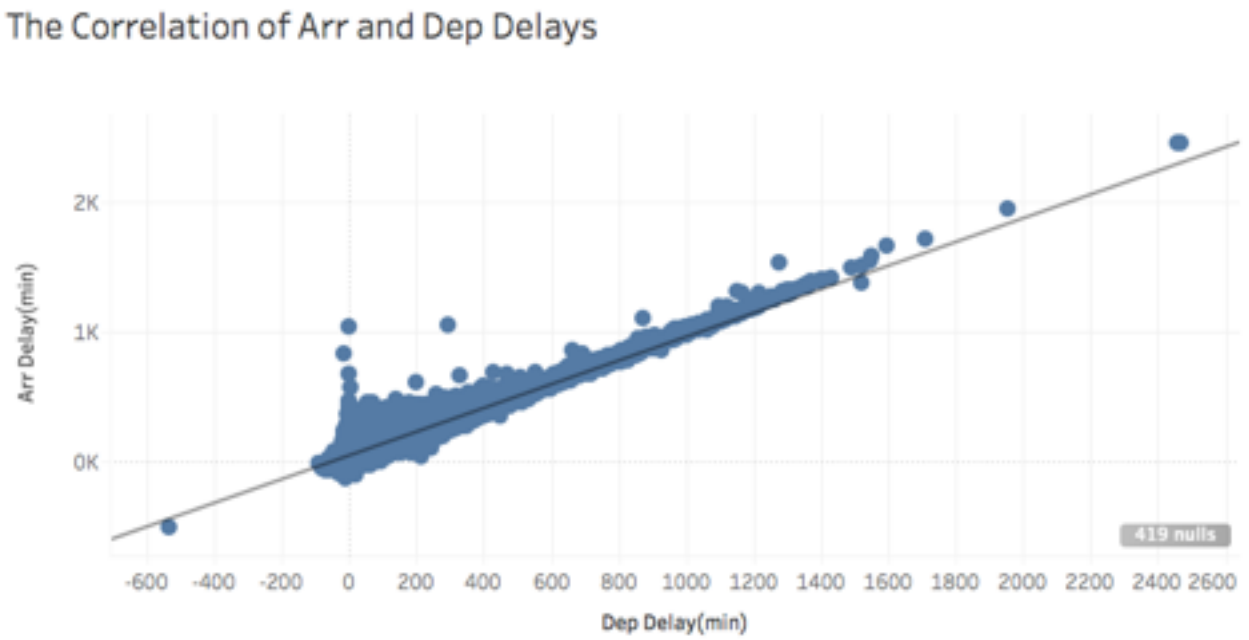
In this year, representing 20 commercial and 7,009,728 flights.The cancelled flights account for 1.96%, and the delayed's account for 38.52%.



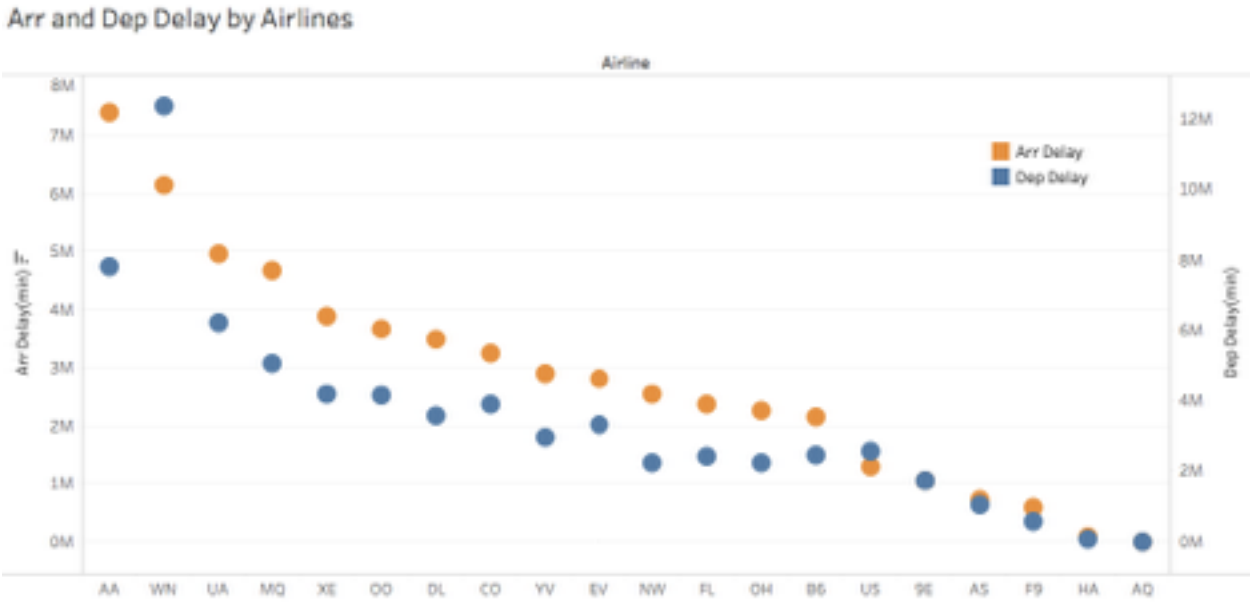
2. Design

Flight delays mainly includes departure and arrival delay. The arrival delay consists of carrier, weather, NAS, security, lateAircraft and other delay.

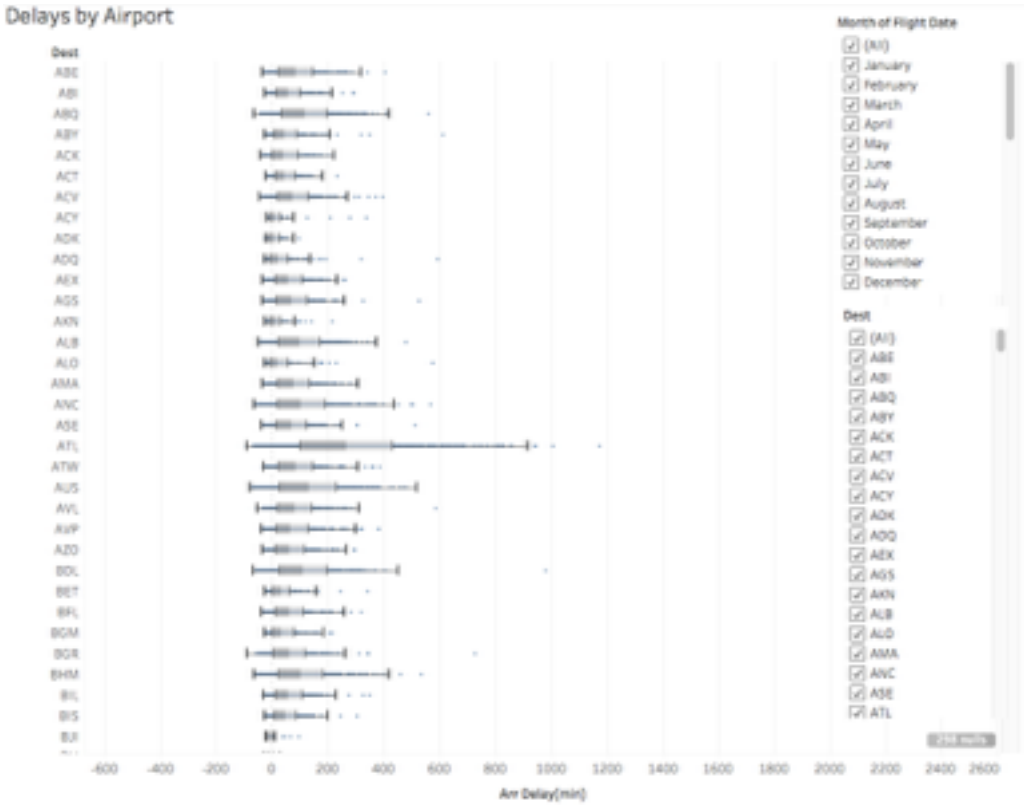
In the following visualization, using the scatter plot to show two variables with a given correlation which the shape of the data is strong.



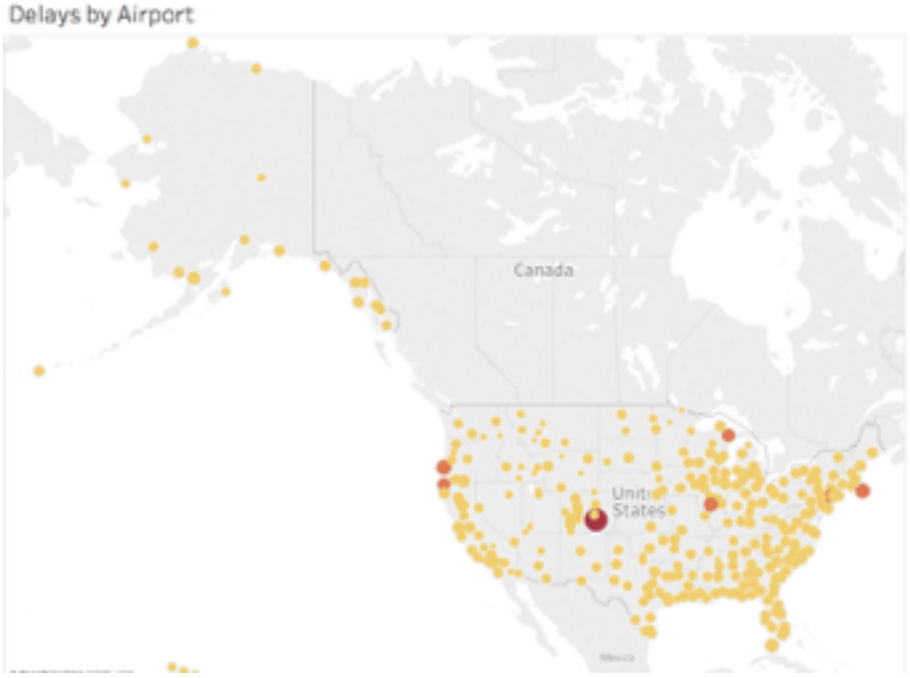
According to the following diagram,arrival delays were always greater than departure in the most airlines. Therefore, we will focus on arrival delay in the rest exploring.



2.1 Initial design

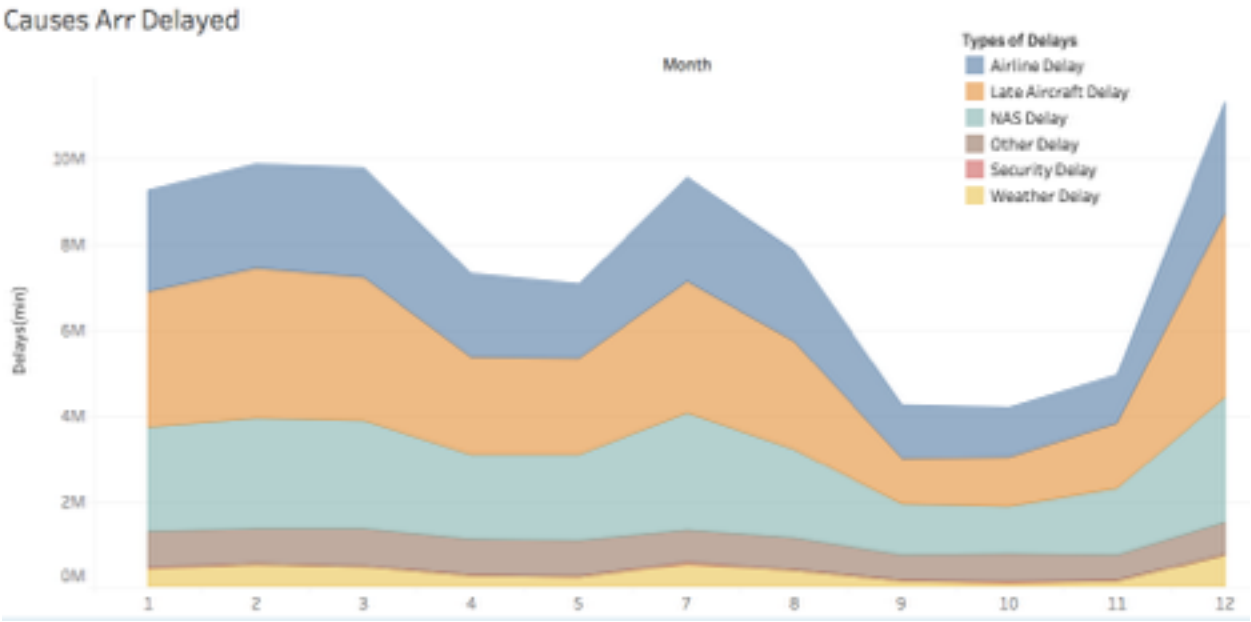


Above of my initial run to display the diagram of arrival delays by airport. Apparently, lots of airports are in there, and it can't be visually presented which airport has the highest or lowest delays.
The below map chart is clear to show. I used size and color represent how much arrival delays are and how the scale of arrival delays is in this airport.

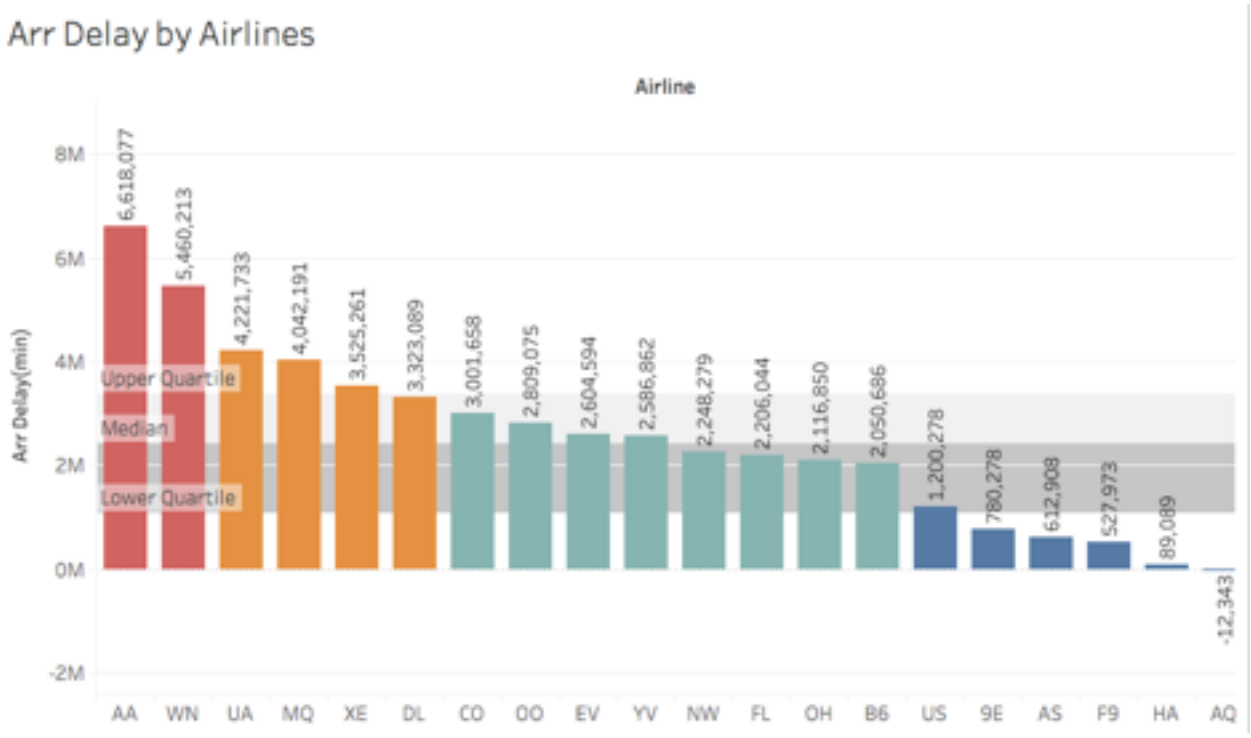


2.2 The Causes of the Arrival Delay By Month

Following diagram we can see the type of each delay varied from month to month. For example, the effect of weather delays was almost zero between September and November.It seems that September to November were a good time to travel.

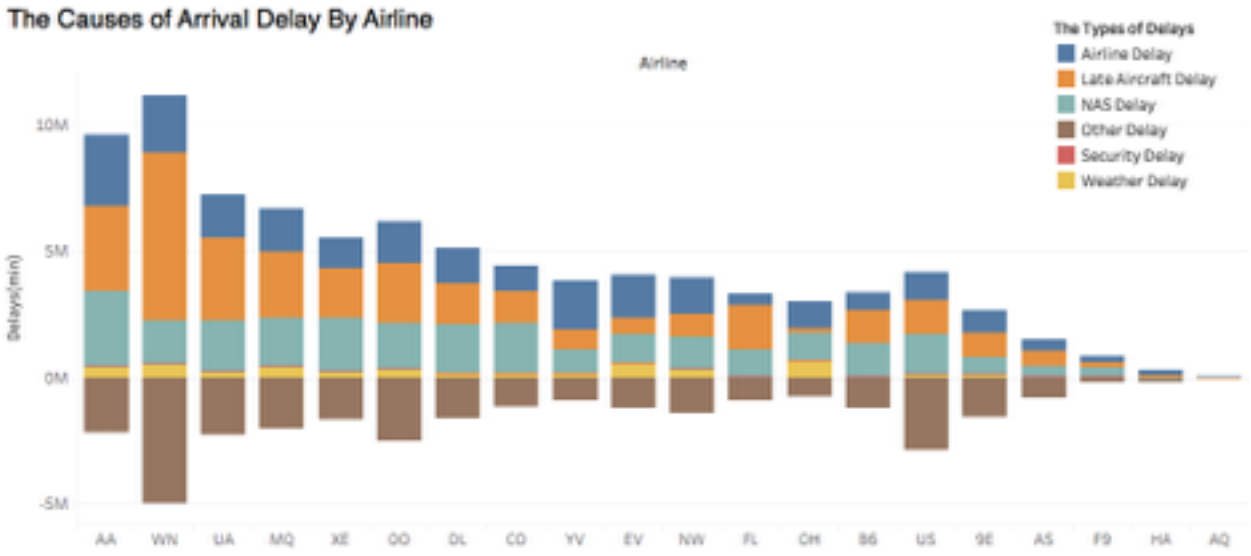


2.3 The Rank of the Total Arrival Delays By Airline



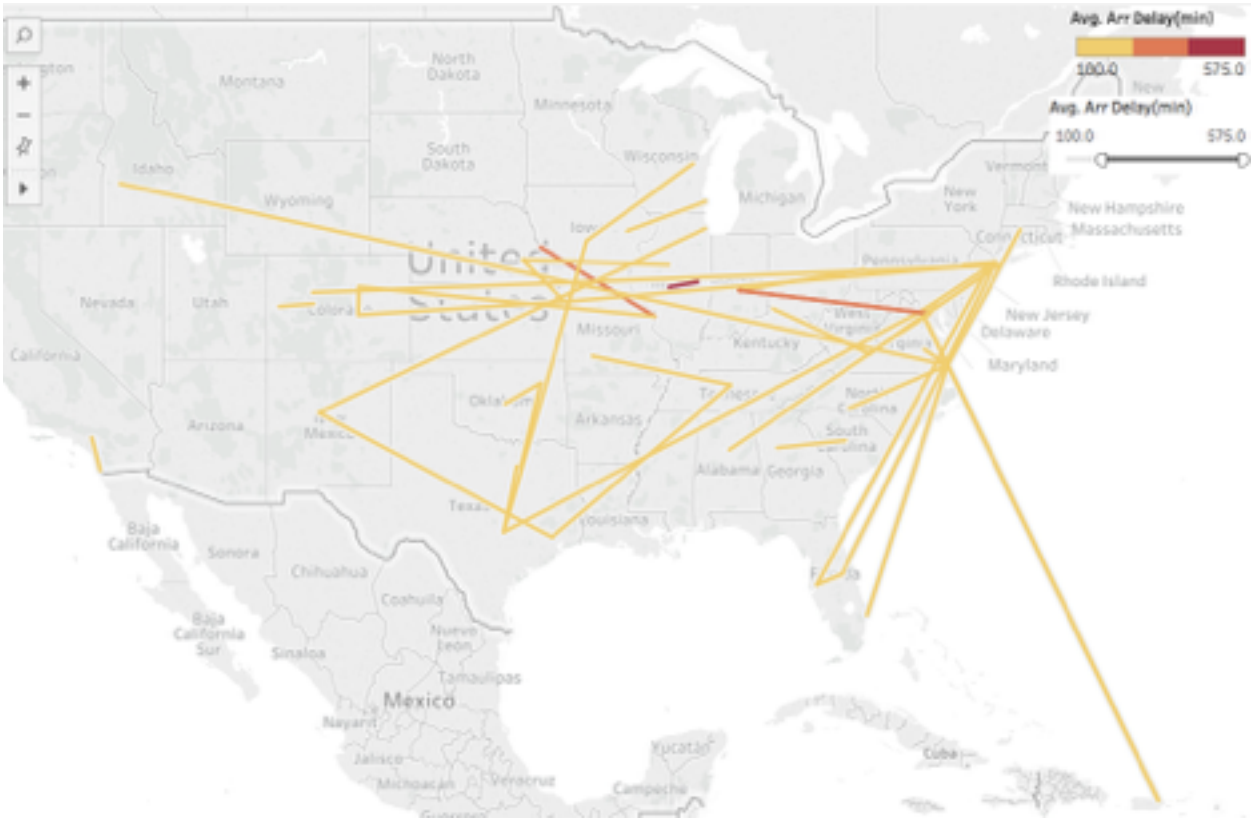
As you can see, AQ was the best airline which had the lowest arrival delays. Nevertheless, AA had 7,381,323 minutes arrival delays which indicate it was the worst airline in 2008.

2.4 The Causes of the Arrival Delays By Airline



Above diagram, AA airline were mainly influenced by airline, late aircraft and NAS delay.

2.5 The Relationship between Arrival and Distance



I plotted some air lines which the average of arrival delays are greater than 100 minutes.The three of highest arrival delays of distance are not the longest. There is no direct relationship between arrival delay and distance.

3. Feedback

Here some feedback that I got from my friends and applied to improve my diagrams:

Natalie - 'I notice a map and spots that provides information on where airport delays are occurring in, a pie chart and descriptions of the problem. My question would be why has airport delays been increasing; what is the cause of it. The relationship I notice is that flight delay's percentage is not far from flights that arrive on time. The main idea of the visualization is that flight delays are increasing. And I understand the graphic.'

Since she used the iPhone to watch the visualization which showed it too small,she noticed why has airport delays been increasing and what is the cause of it after I reminded. But I add some text on my diagrams to highlight what topic is.

Chris - ' I noticed that this data is only US. My question would be what is the total number of flights; what is WN,AA etc mean? According on the data distribution, the worst delay is in the central of US.I want to avoid flights to the central of US.'

First of all,I've already put the total number of flights in my caption,so I don't think I should put it into my diagrams. Second I put the name of airline into the details of makes in tableau , so that can explicit to what airline does the air symbol represent.

My mentor - 'Red-Green colors are not a good choice for using in comparisons (used in the first slide) since they are not color-blind friendly colors. The people with this disorder are not able to distinguish this two colors.'

I changed the Red-Green colors in the some diagrams.

4. Resources

Data Visualization in Tableau Course

https://www.tableau.com/learn/training

5. Data Source

2008.csv is taken from the following website in 2008:
http://stat-computing.org/dataexpo/2009/the-data.html

airports and carriers data are taken from the following website:
http://stat-computing.org/dataexpo/2009/supplemental-data.html

I recombine the data for showing the spider map,the following code:

In [39]:
```
"""
recombine the data to getting long and latitude data with carriers
"""
df = pd.read_csv('2008.csv')
airports_df = pd.read_csv('airports.csv')

# # print(airports_df.info())

origin_merge = pd.merge(df, airports_df, how = 'left', left_on=['Origin'], right_on=['iata'])

dest_merge = pd.merge(df, airports_df, how = 'left', left_on=['Dest'], right_on=['iata'])

concat = pd.concat([origin_merge,dest_merge])
concat['PathID'] = concat['Origin']+'-'+concat['Dest']

# pprint.pprint(df.shape)
# pprint.pprint(concat.shape)
df_new = concat[['airport','PathID','ArrDelay','UniqueCarrier',
                 'long','lat','Month','DayofMonth','Year']]

# dump data
df_new.to_csv('2008_geo.csv')
```

6. Visualization in Tableau Public

The following of initial version:
https://public.tableau.com/profile/esther2958#!/vizhome/Theperformanceofflightdelay/Story1

The following of final version:
https://public.tableau.com/views/Theperformanceofflightdelay-final/Story1?:embed=y&:display_count=yes&publish=yes