Exploratory Data Analysis on Flight Dataset

This is an EDA on the infamous Flights Dataset.

To Reproduce the result

- Run `downloadData.sh` to fetch original datasets and set up the working environment
- Put the datasets from `airlinesauxiliaryfiles.zip` provided with the assignment
- Run `wrangle.hive` to generate the datasets used for plots (saved to the folder `generated_datasets`)
- Run `main.R` to generate plots (generated plots will be stored in the folder `plots`)
- Run `main.py` to generate plots (generated plots will be stored in the folder `plots`)

Built With

- RStudio - The framework we used to create the analysis
- Jupyter Notebook - The environment we used to create the Python portion
- Hive - The framework we used to wrangling large datasets

Authors

- **Yoon Chang** - Portfolio
- **Yi Xu**
- **Yingyi Lai**
- **Junyao Chen**

Roadmap

1 Introduction

Imagine when we were in an airport ready for the vacation, but suddenly told by the broadcast that the flight was delayed, how frustrating and annoying we would be at that time. In order to have a better understanding of what lead to the delay and a better chance to avoid it, we decide to dive in to the airlines dataset we get from the US Department of Transportation's Bureau of Transportation Statistics (BTS). Focus on the airline data of 1998 and 2006, we would explore many possibilities that could cause the delay, such as wether condition, airport location, seasonal trend and so on.

There are four datasets we work on : *airlines, airports, carriers* and *plane-data*. The dataset *Airlines* contains 29 variables, recording flight's date, delay time and many other information. *Airport* include 7 variables, showing the specific location of each airports, such as city, state where it locates and an exact coordinate . The *carrier* dataset only has 2 variables, the carrier codes and their full name. As for the *plane-data*, it contains 9 variables recording comprehensive information of each plane. For example, it includes the responding manufacturer, model type and the year it produced. These are all valuable information for us to explore the causes for delay of planes, especially for the delay of departure and cancellation. In the following, we would join different dataset according to our need to analyze different aspect of airline delay situation. Since there is data in 1998

and 2006, we also want to see if there is any improvement made for the delay and cancellation situation through these years.

2 Wrangling

- byCarr (Hive)

  Grouping the dataset by the different carrier, I calculate the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time for each group. Selecting only 8 variables into this *byCarr* dataset, each record shows the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time for different carriers.

- byModel (Hive)

  Joining the *plane-data* dataset and *airlines* dataset, I group the new dataset by the year of data, manufacturer, model of the plane. Then I calculate the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time in each group. Selecting only 9 variables into this *byModel* dataset, each record shows the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time for different manufacturer and model of planes.

- byState(Hive)

  Joining the *airlines* dataset and *airport* dataset, I group the new dataset by the year of data and state where the airport locates. Then I calculate the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time in each group. Selecting only 8 variables into this *byState* dataset, each observation shows the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time for different year of data and state of the airport.

- byCord(Hive)

  Joining the *airlines* dataset and *airport* dataset, I group the new dataset by the coordination, lattitude and longitude of the airport. Then I calculate the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time in each group. Selecting only 8 variables into this *byCord* dataset, each observation shows the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time for different year of data, lattitude and longitude of the airport.

- bymyear(Hive)

  Joining the *airlines* dataset and *plane-data* dataset, I group the new dataset by the year of data and manufacture year of plane. Then I calculate the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time in each group. By selecting 8 variables into this *bymyear* dataset, each record shows the number of flight, number and rate of the delay, number and rate of the cancellation and average delay time for different year of data and different manufacture year of plane.

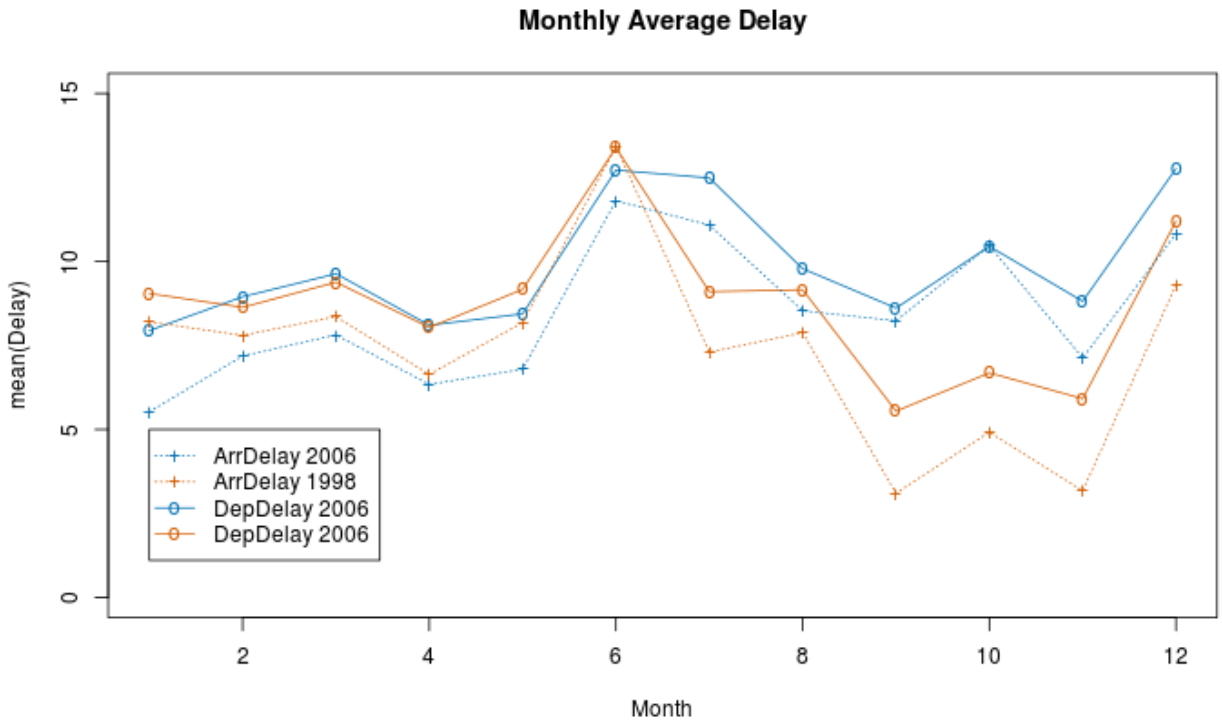
3 Visualization & Analysis

3.1 Feature Analysis

3.1.1 Delay

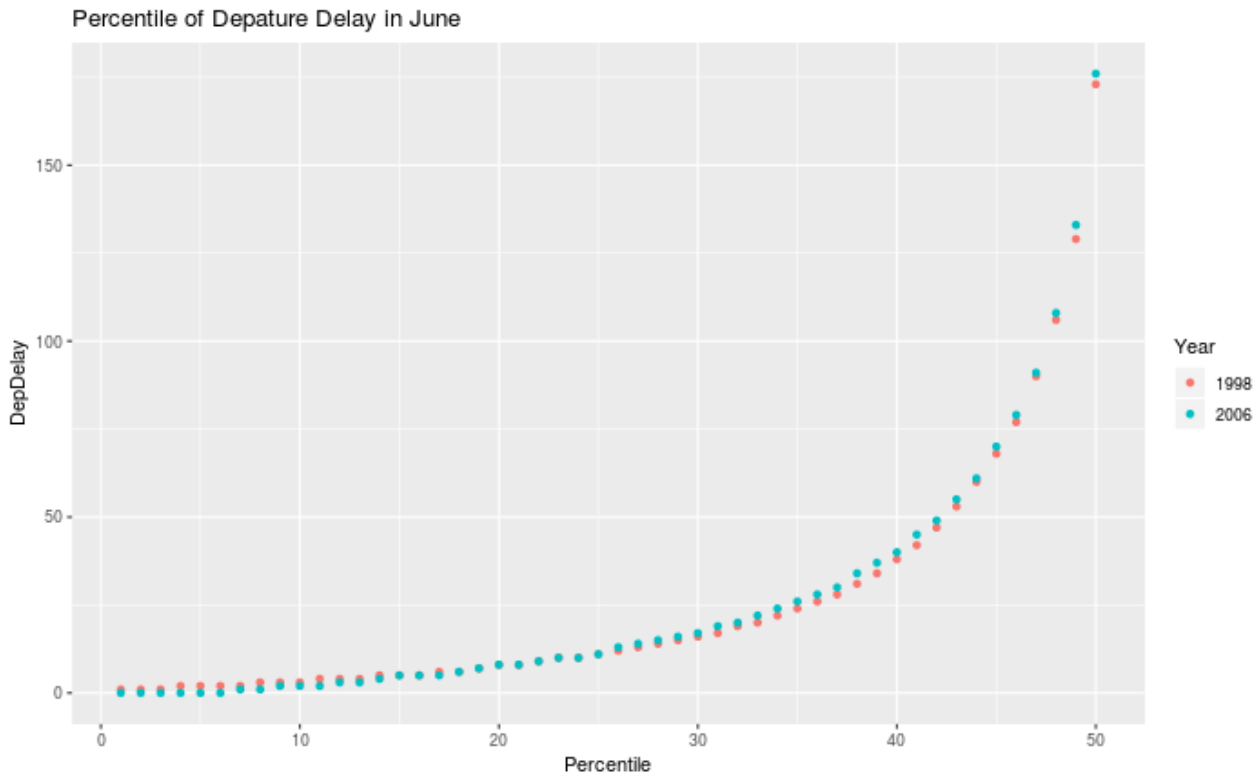
- Monthly / Day of Week trend of departure/arrival delay
  - Look through the year

  First we make the following plot to show how average arrival delay time and departure delay time of each month change through the year for 1998 and 2006.



As shown in the plot, all the lines generally follow the same trend for the first half of the year (January - June) and reach a peak at June. Then for June to December, 2006 generally have longer delay than 1998. For both of the two years, monthly average departure delays are longer than arrival delays. Base on this observation, we suspect there might be a relationship between the difference of expected time expansion minus actual time expansion and departure delay time. In another words, we wonder if departure delay can cause the flight spend less time to get the destination. We notice that there is a peak in June. Now let take another look into that.

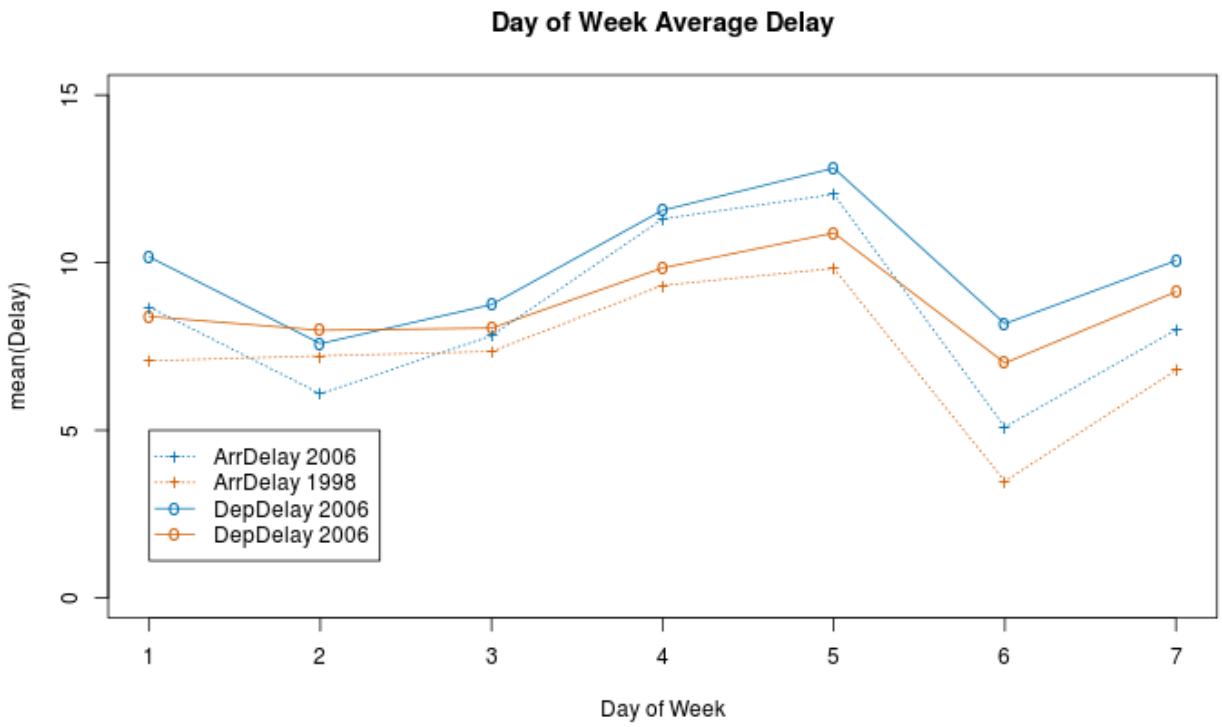
The following plot shows the percentiles of departure delay in June for the two years.



We can see from the plot that the departure delay for the two year roughly follow same distribution.

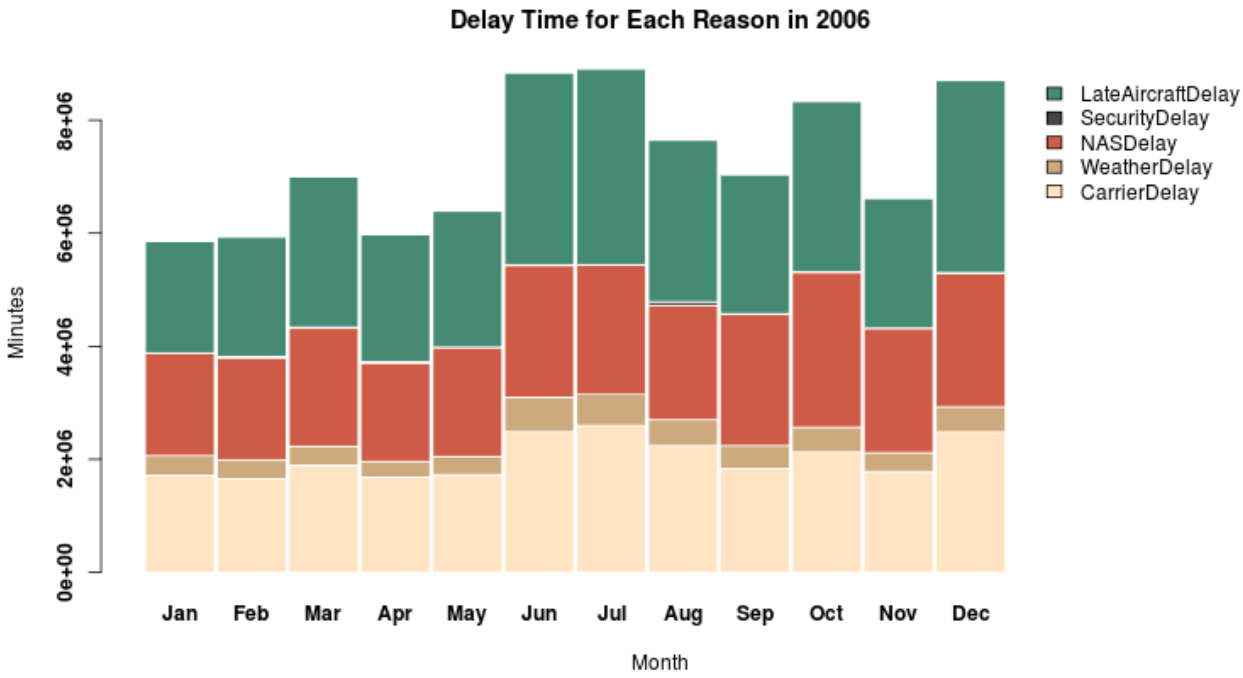
- Look through the week

Now we make the following plot to show how average arrival delay time and departure delay time of each day of week change through a week for 1998 and 2006.



According to the plot, for each of day in a week except for Tuesday, average delay in 2006 is higher than it in 1998 . Also we can see that Fridays have highest average delay time and Saturdays have lowest average delay time.

- Reason of delay in 2006

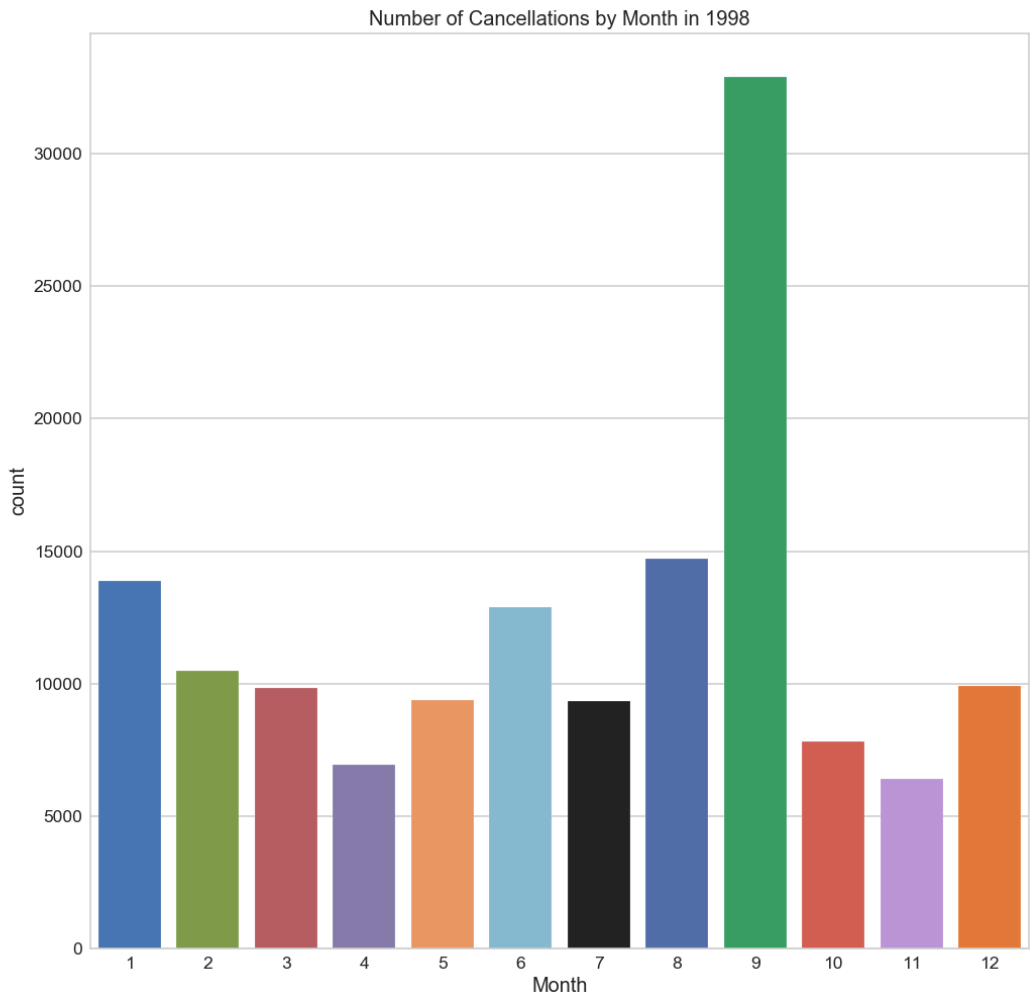


This graph shows the total time of delay by reason in each month of 2006. The y-axis indicates the sum of delay time in the unit of minute. For the sum, June, July and December had higher values, while January and February had relatively lower ones. For these five delay reasons, the distributions of them in each month were similar. Late aircraft, NAS, and carrier were three common reasons leading to most delay time, and the weather did not cause too much delay, which is unexpected. Delay caused by security was rare that we can hardly see the blocks with darker color from the graph.

Case Study: Hear Wave in 2006 We expected that there would be more delay time in winter, but actually June and July had the largest delay time caused by weather. Thus we might guess that some events or some extreme weather happened at that time. As we went back to the news, we found that there was a severe heat wave in 2006 summer that affected most of the United States and Canada, killing hundreds of people, and temperatures in many locations made the highest temperature records.[3]

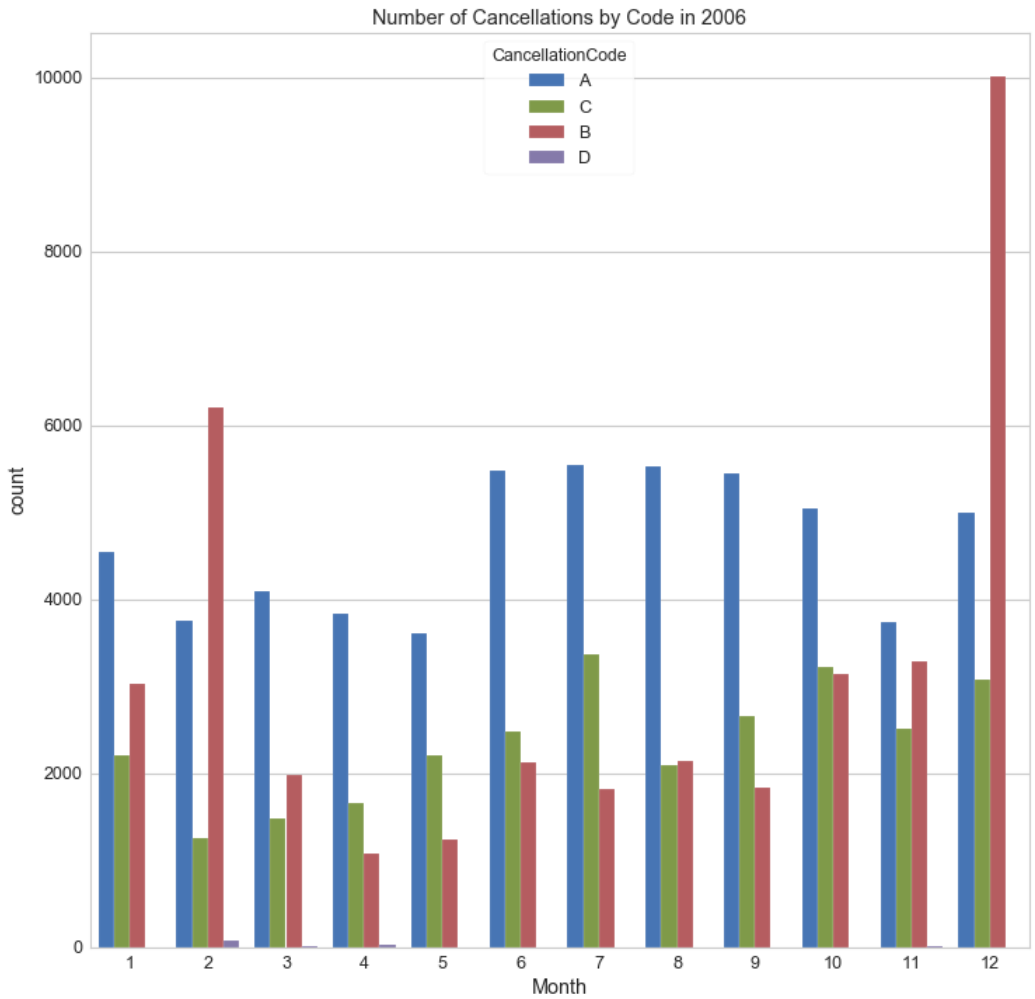
3.1.2 Cancellation

- Distribution Plots
    - Number of Cancellations by Month in 1998



In the above plot, there is a clear spike in September of 1998, and the rest of the months stay consistent in its cancellation numbers. Although the exact cause is not known in the data, there was Swissair Flight 111 that crashed near Peggys Cove, Nova Scotia, which killed 229 people on board.

- Number of Cancellations by Code in 2006




```
A = carrier, B = weather, C = NAS, D = security
```

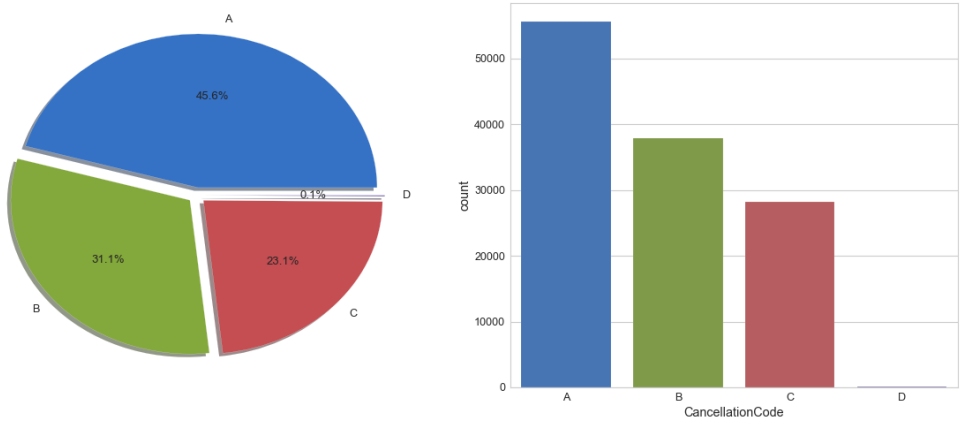
Generally, Code D (Secutiy) seems as it does not cancel flights much at all. Code A (Carrier) and Code C (NAS) stay consistent in their magnitudes among the causes of cancellation throughout the year. However, Code B (Weather) varies in its magnitude more dramatically depending on the month, spiking in December and February and calming down in summer months.

Case Study: Extreme Weather in 2006 February and December are the top 2 months that most flights got cancelled because of the weather. Thus we look back to the extreme weather in 2006.

In February 2006, there was a storm called the North America blizzard beginning on the evening of February 11th, and ending in Canada on February 13th. Some hardest hit areas included Connecticut, Delaware, New York, New Jersey, Virginia, Pennsylvania, Massachusetts, Maryland and District of Columbia. The major cities in the northeast received at least a foot of snow. Fatalities occured in several areas. [1]

Additionally, in December, there was a windstorm, Hanukkah Eve windstorm of 2006, happened, which was a powerful Pacific Northwest windstorm in the Pacific Northwest region of the United States and southern British Columbia, Canada between December 14, 2006 and December 15, 2006. It produced hurricane-forced wind and heavy rainfall, resulting in hunreds of millions of dollars in damage and serious casualties.[2]

- Distribution of Cancellation Codes



Based on the charts above, Code A (Carrier) cancellation takes about half of all the cancellations that occured in 2006. The second most common type is Code B (Weather), which is a bit more seasonal compared to the other two. Code D (Security) only takes 0.1% of all the cancellations.

3.2 Multivariate Analaysis & General Trends

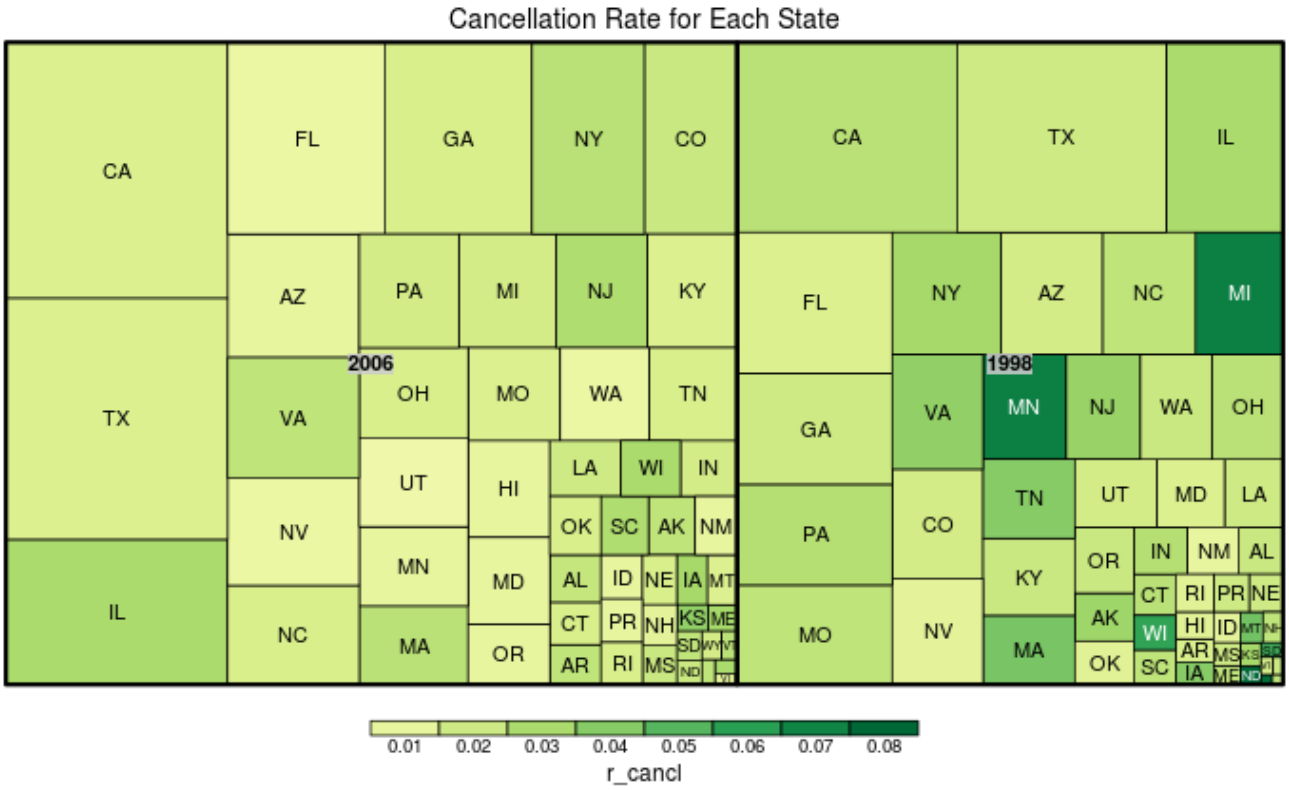
3.2.1 Treemap Analysis

We make the following treemaps to uncover span of the delays and cancellations over where the flight departure, which airline is the carrier and what kind of plane used for the flight.

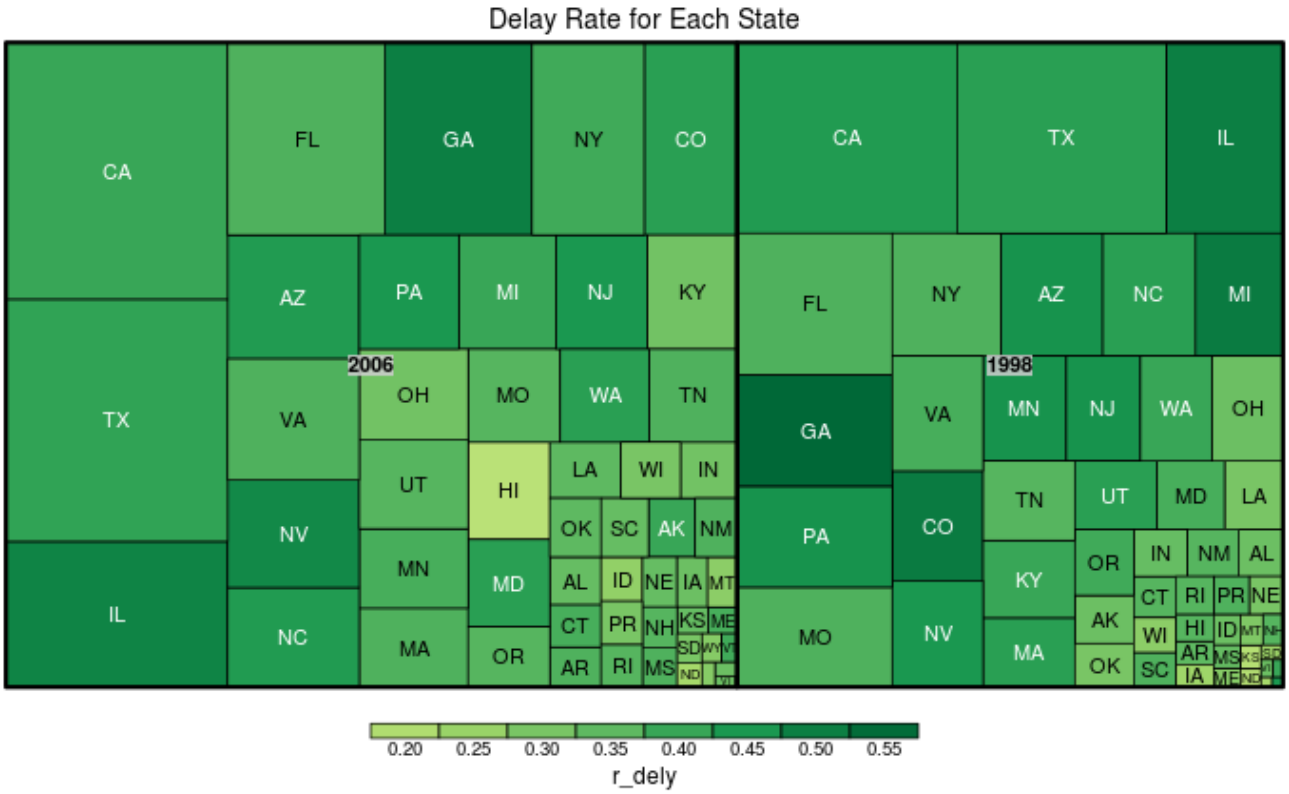
- Analysis for the states

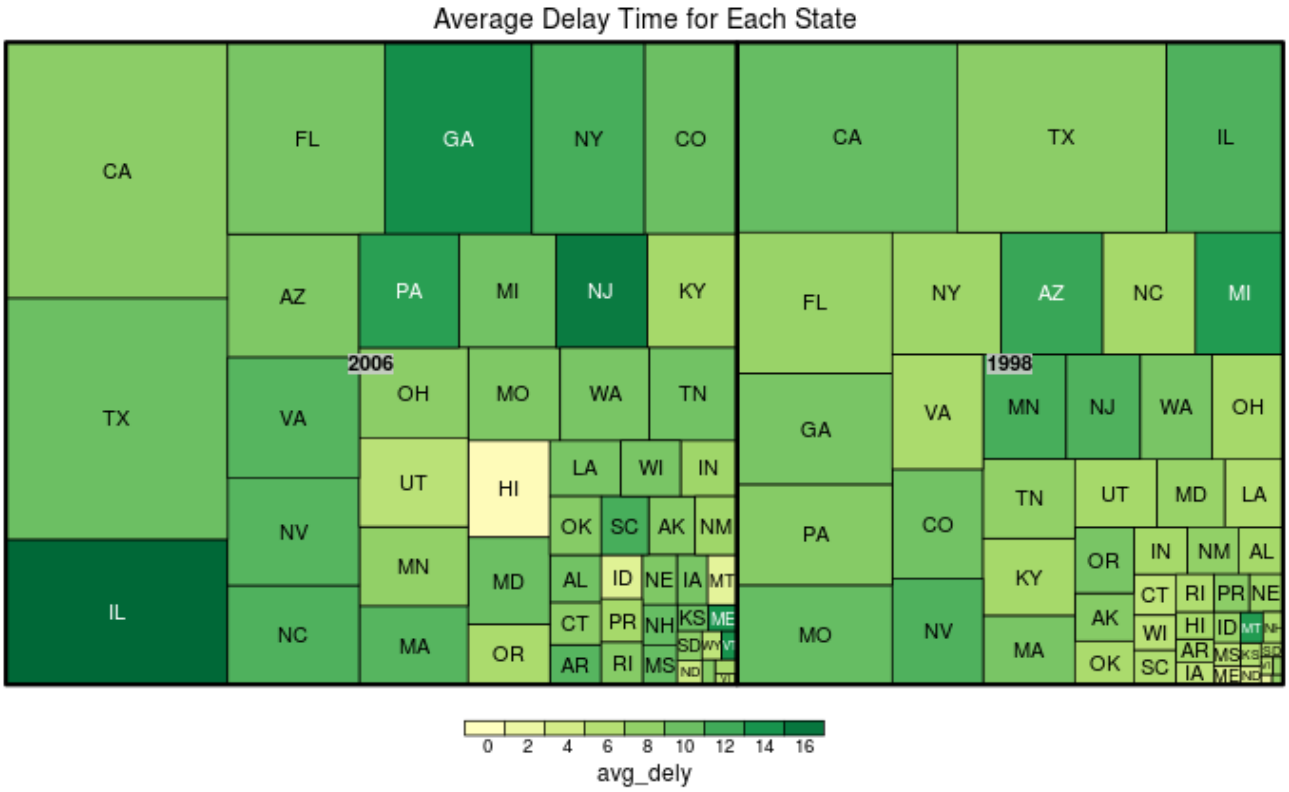
The following 4 treemaps show the cancellation rate, delay rate, and average delay time for each state. For all the 4 plots, every small rectangle represents a state, the sizes of the rectangles represent the number of flights recorded. And the color of them in each plot show the value of the 3 measures just mentioned respectively.

First we notice the total number expanded from 1998 to 2006. California had the largest number of flights in both 1998 and 2006.



In terms of cancellation rate, things were generally better in 2006 than it was in 1998. In year 1998, Mississippi and Minnesota had the worst cancellation rate.

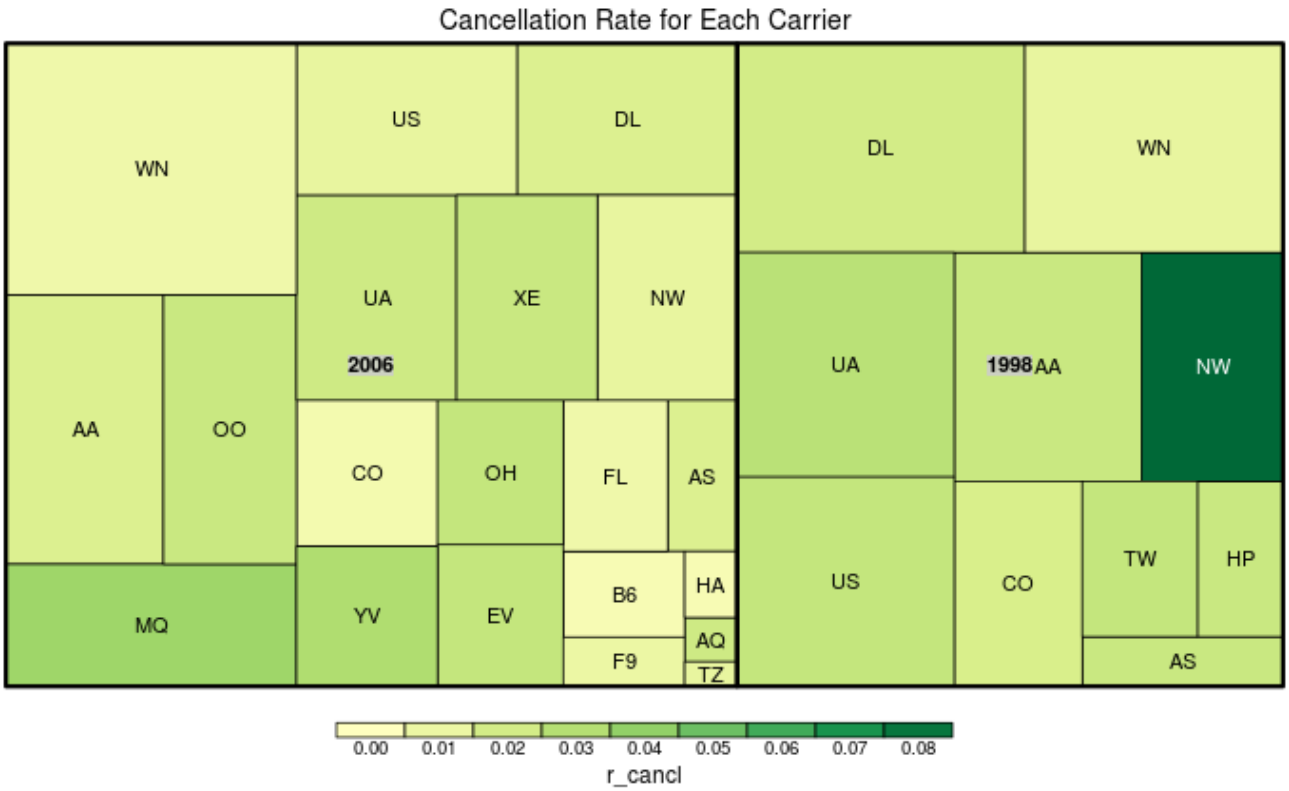




The 2 plots above tell us that people in Hawaii, Idaho and Montana are really since their flights almost never got delayed. On the other hand, our Illinois is one of the states where flight got delayed most.

- Analysis for the carriers

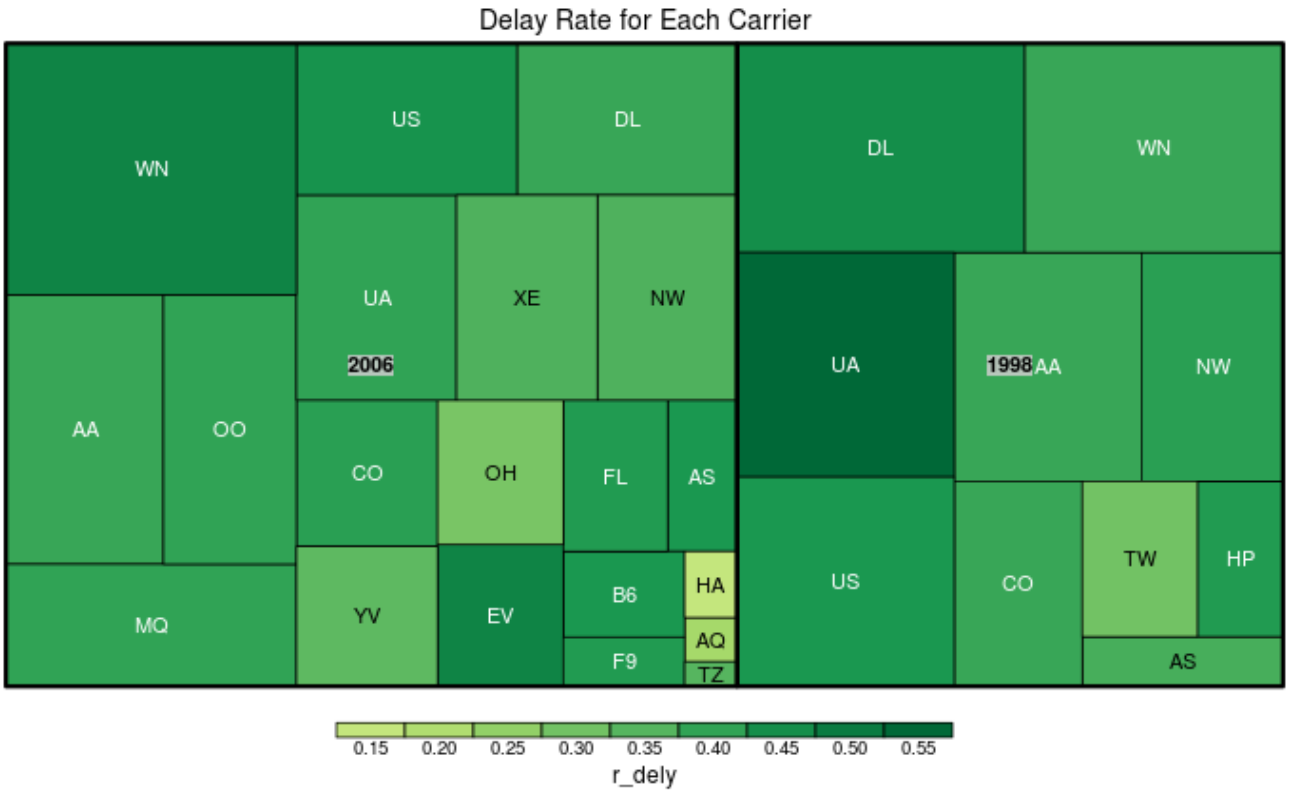
Now, lets find out flights of which carrier got cancelled or delayed most. For the 3 following treemaps, each rectangle stands for a carrier. First we notice that there are more carriers recorded in 2006 than in 1998.

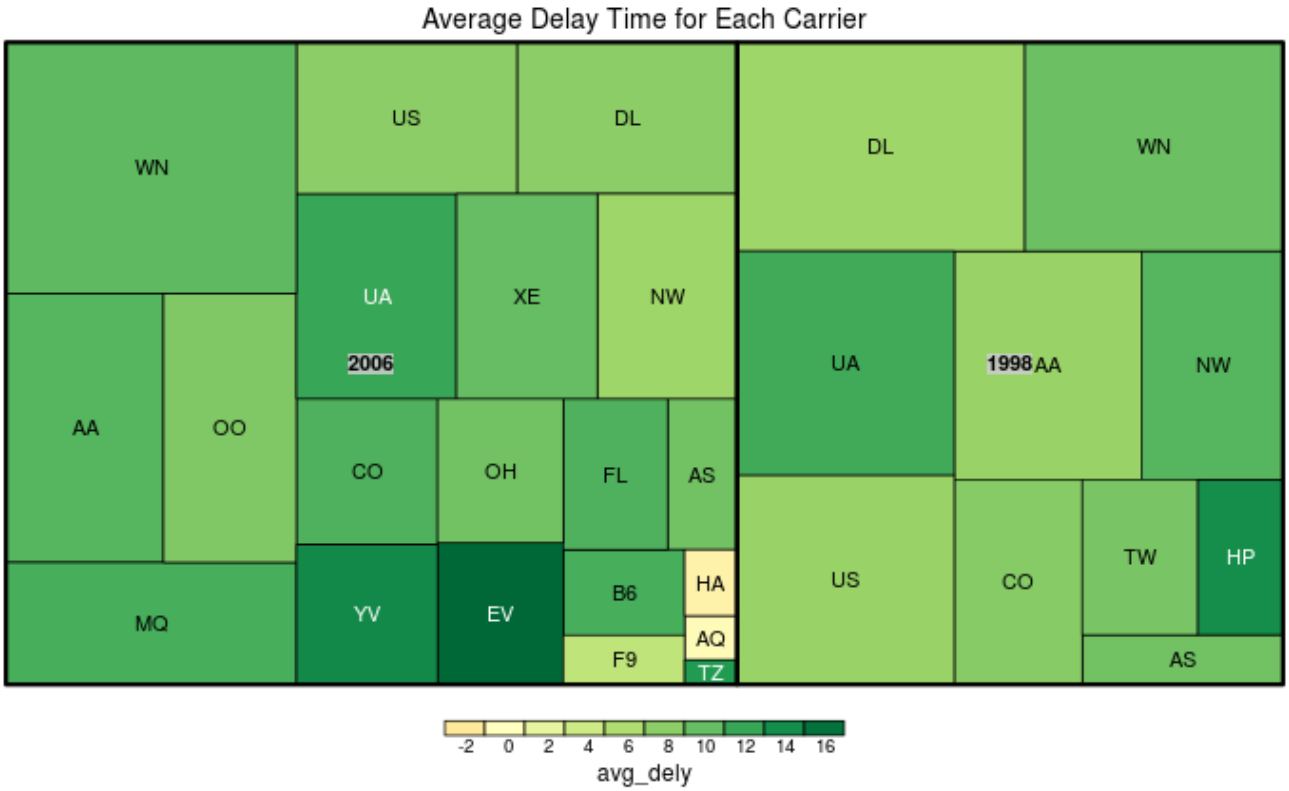


In 1998, NW (NorthWest Airlines) definitely stands out from the rest in terms of its magnitude of cancellation rate. In 2006, the top three carriers with highest cancellation rates are MQ (American Eagle Airlines), YV (Mesa Airlines), and EV (Atlantic Southeast Airlines).

Case Study: What happened to Northwest Airlines in 1998

Northwest Airlines fell into troubled labor relations in 1998. It walked away from the bargaining table, locked out its pilots and shut down the airline for more than two weeks. The airline sustained heavy losses as a result, and ended 1998 in the red, after being profitable since 1993.[4]



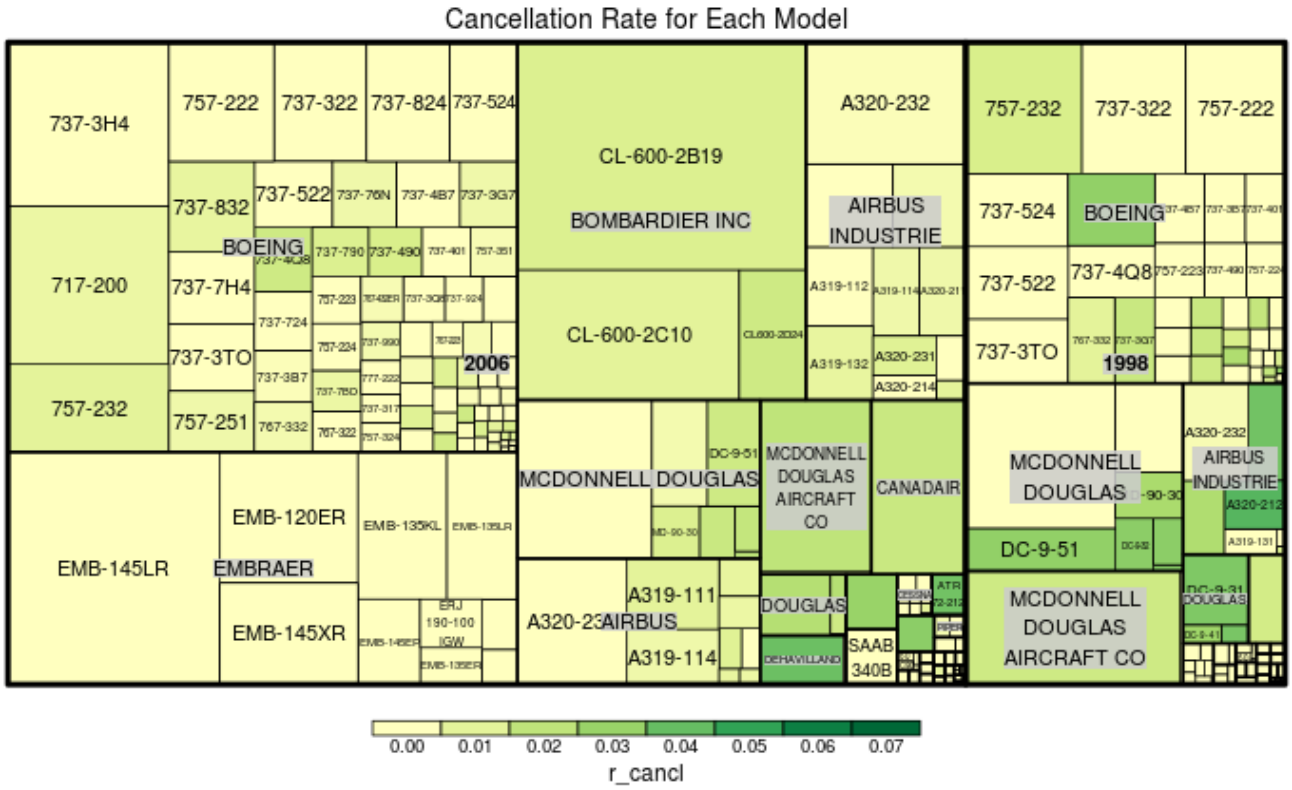


According to the 2 treemaps above, HA and AQ had really low delay time and rate.

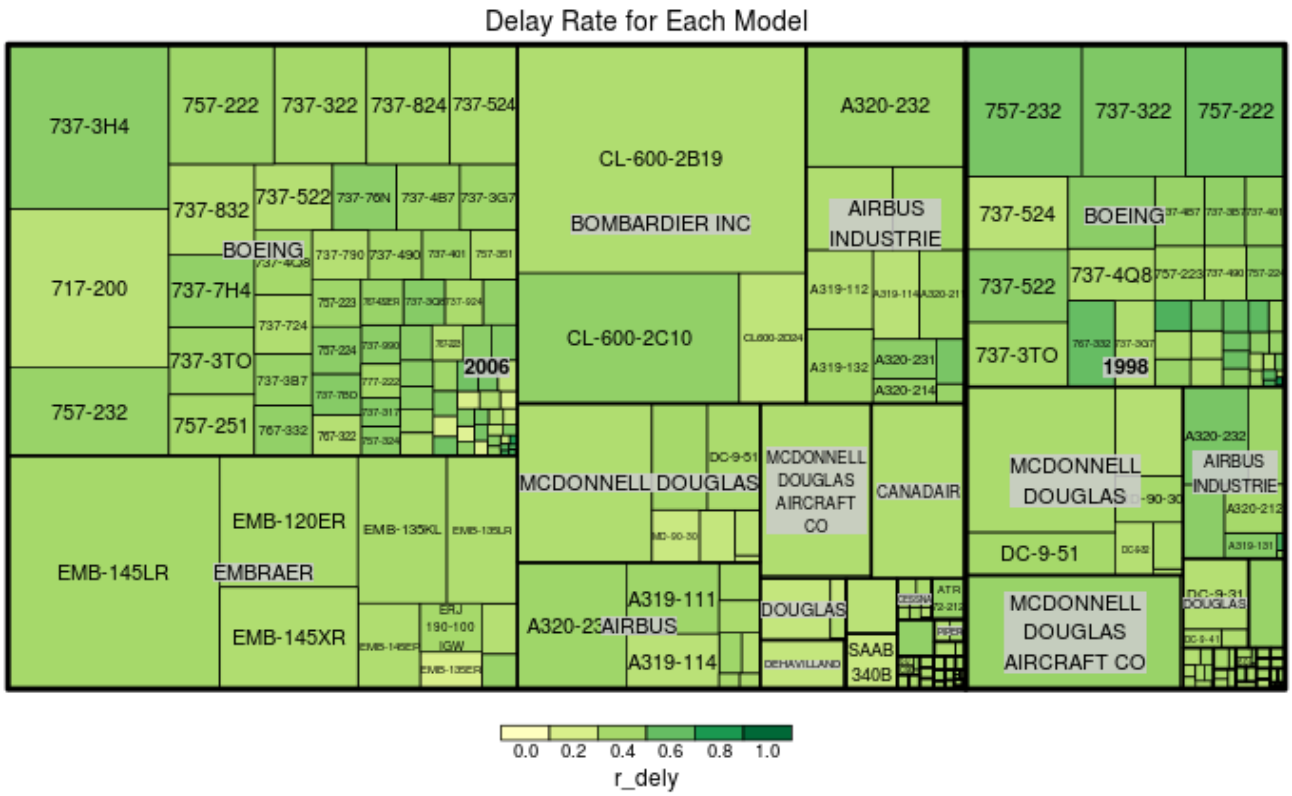
- Analysis for the models

The next two graphs show the relationships between cancellation rate and the plane model as well as the delay rate and the model type. The size of a square indicates the number of flights with that particular model. And a darker color indicates a larger ratio value.

For the two years, Boeing was the leading position of the number of flights. In 1998, McDonnell Douglas was in the second position while the ranking was lowered to the fifth in 2006. Embraer and Bombardier INC., which were not in the top 4 in 1998, became the second and third largest manufactures in 2006. The models in 1998 and 2006 differed a lot. Most old models were replaced by newer ones.

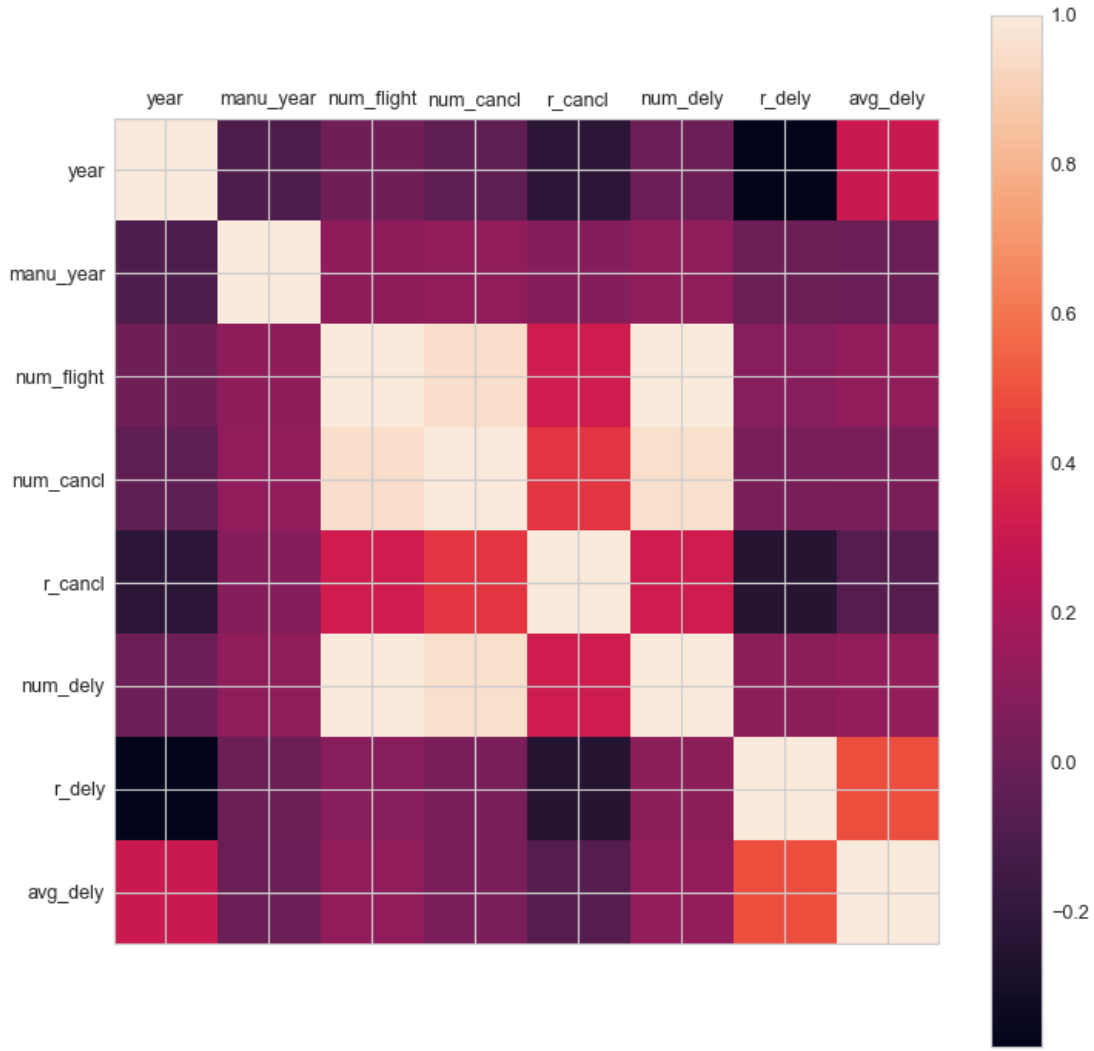


In 2006, most models in the top 8 manufactures had relatively low cancellation rates. On the contrary, large cancellation rates happened to flights with small manufactures. But this trend was not the same in 1998, the top several manufactures all had models with high cancellation rates.



This graph is not that interesting, because the colors are similar between models, manufactures and years. The delay rate was not affected significantly by models and manufactures.

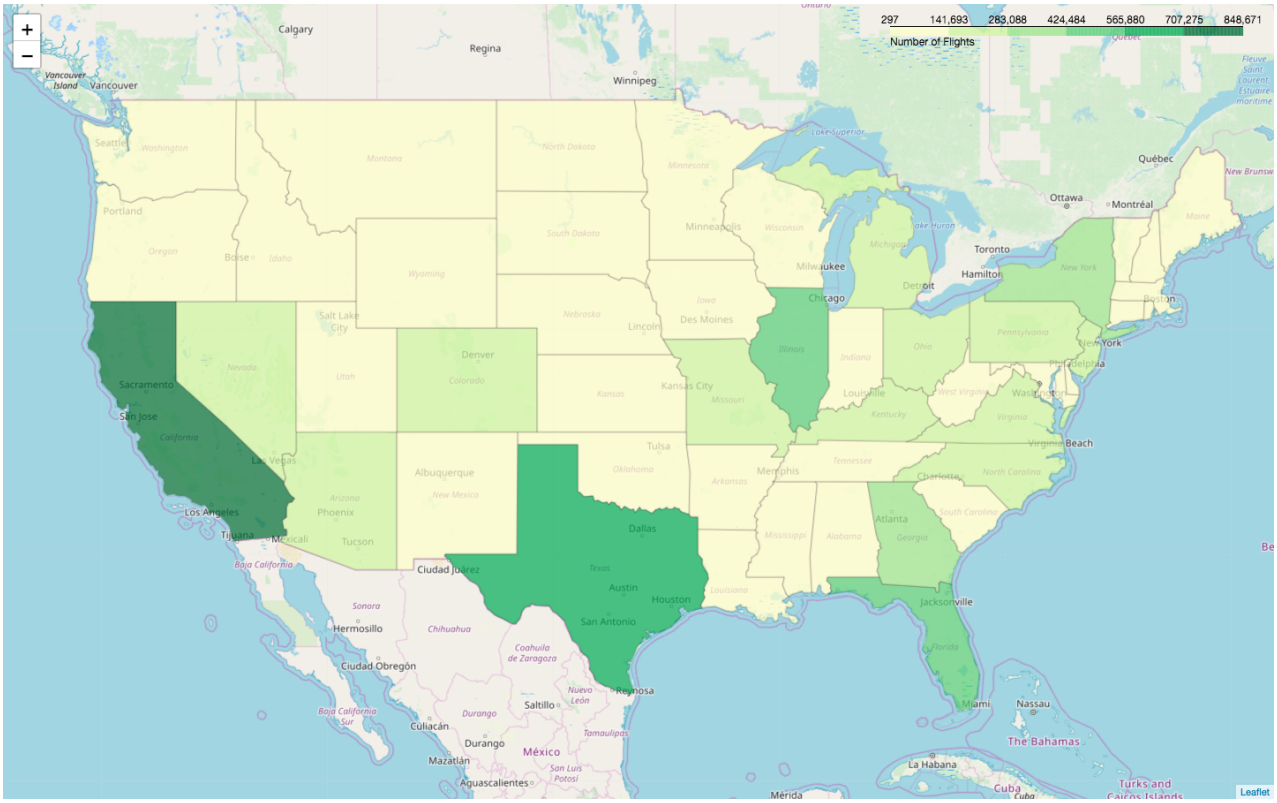
3.2.2 Correlation Heatmap Analysis



Some of the notable negative correlations include rate of delay & year and rate of cancellation & year.. As for the rate of delay and rate of cancellation versus delay, it makes sense that they would decrease over the years as airports and airlines stabilize and improve their procedures. Also, note the positive correlation between number of cancellation and number of delays versus number of flights. This is also natural since more flights could lead to more cancellations or delays.

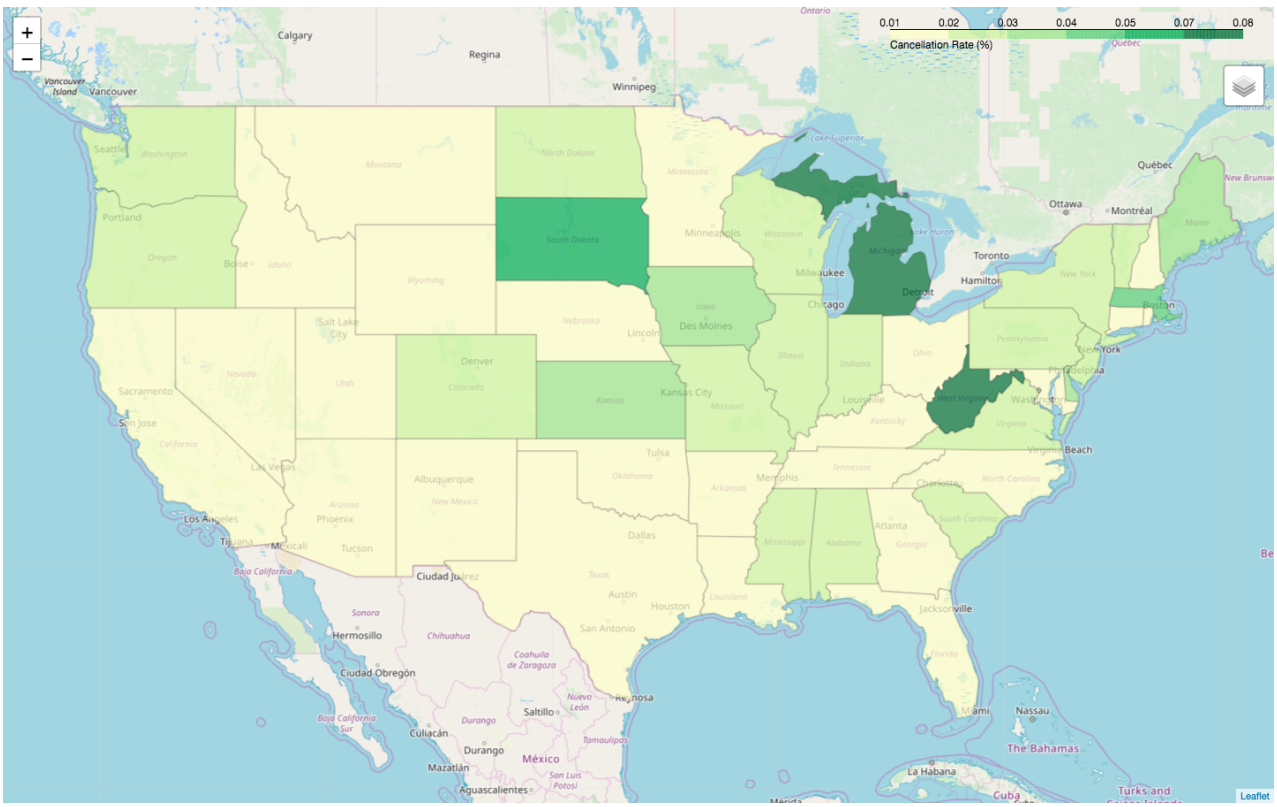
3.2.3 Geographical Heatmap Analysis

- By Number of Flights



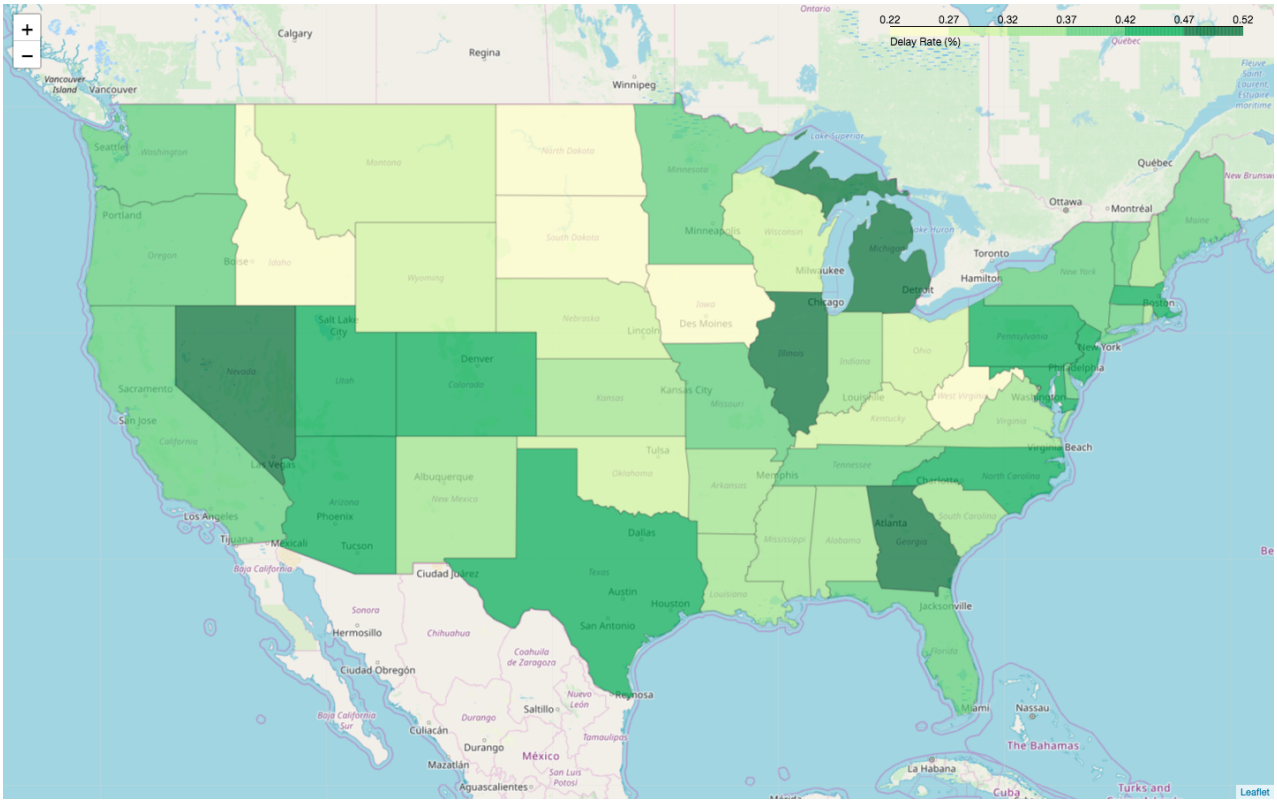
Based on the above map, the top five states with most number of flights are California, Texas, Illinois, Florida, and Georgia. Relatively, we could see that these five states have a substantial more amount of flights compared to the rest of the states.

- By Rate of Cancellation



Based on the above map, the top four states with the highest rate of cancelled flights are South Dakota, West Virginia, Michigan, and Maine. Again, we could see that these four states have substantially higher rate of cancellation compared to the rest of the states.

- By Rate of Delay

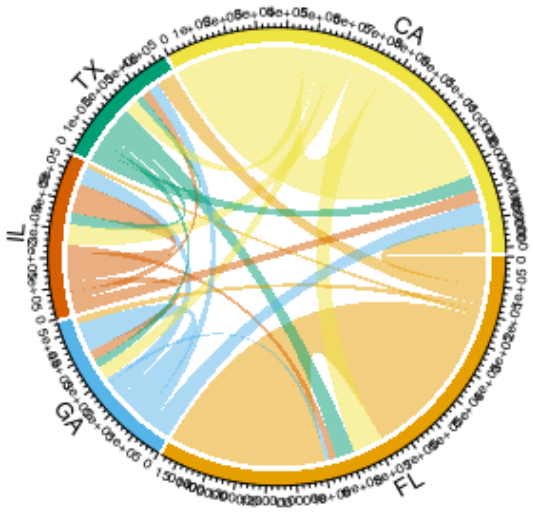


Based on the above plot, the top five states with the highest rate of cancelled flights are Illinois, Georgia, Michigan, Nevada, and Texas. It should be noted that the relative distribution of the above map follows a similar trend with the first geographical heatmap (By Number of Flights), which suggests that the rate of delay may be related to the number of flights in a certain state. However, the gap between the states are closer to each other, evident in much more even distribution of colors throughout the states.

3.2.4 Busiest states with most air routes

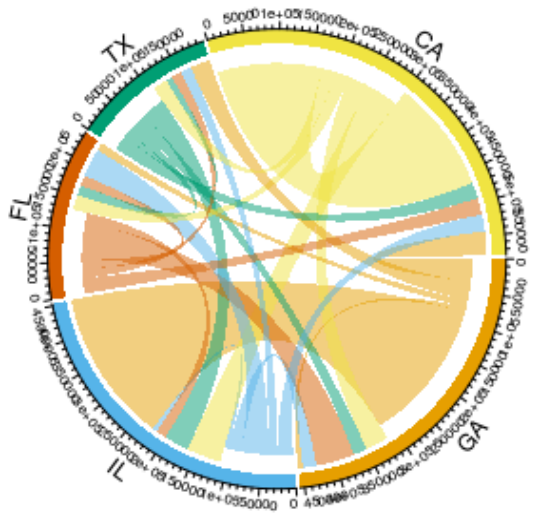
In order to have an overview that which place in America has the most air routes, we select the top five states base on the total number of flights for each state. The chord diagram below clearly illustrate the flights between the top five busiest states.

Chord Diagram of Flights Between States



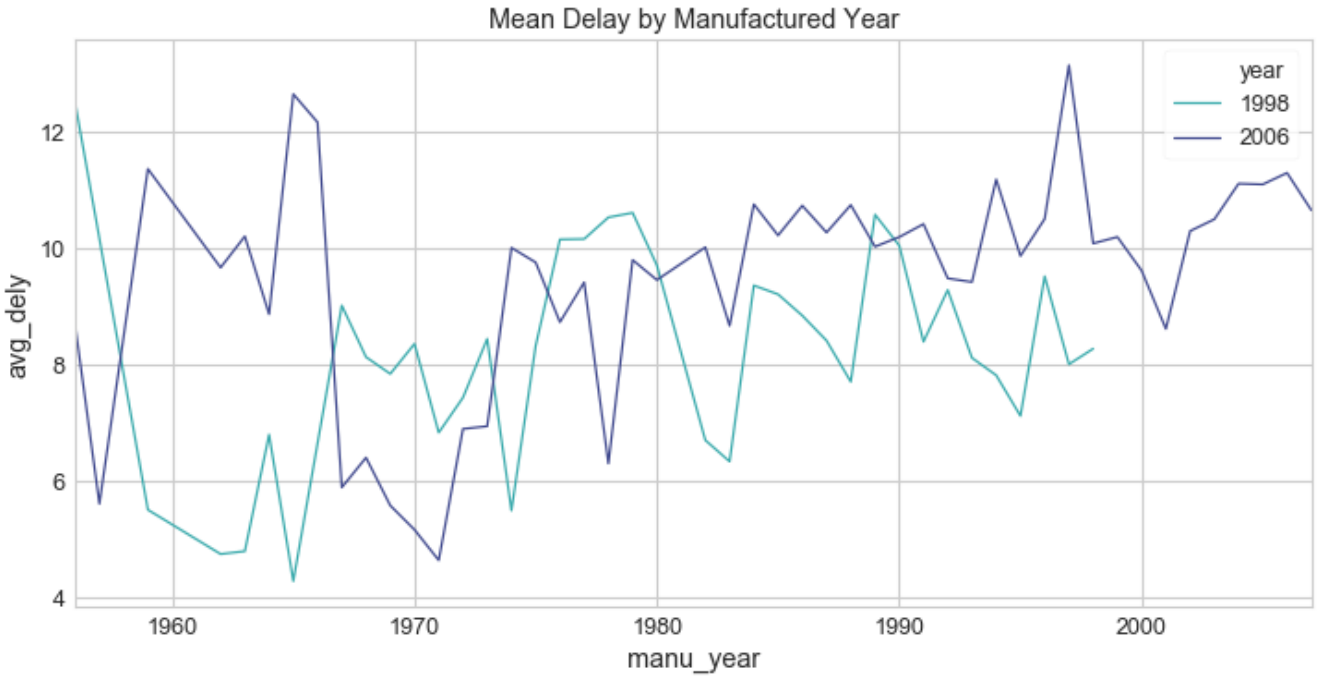
As we can see, the flights from Florida and California take up the majority of flights, following are Texas and Illinois, Georgia. Plus, I also subset the delayed flights seperately, to figure out the which air routes would delay badly.

Chord Diagram of Delayed Flights Between States



We really found something interesting here. Most delayed flights from Georgia are flying to Illinois, while 啊 great many flights delayed in CA are in-state flying, whose destination is also in California. As we noticed that the number of the flights in Florida almost take up one third of the total number of flights in these five states, but the number of delayed flight in Florida is relatively small compare to its total amount. We can infere that the weather in Florida is very good in most of the year and as a main tourist attraction, airports there have good control of plane in and out.

3.2.5 Does flight delay has anything to do with when the pane was manufactured



Based on the graph above, the average delay times have sharp declines and inclines throughout all of the years. This may be due to several possible causes. Certain planes stop flying after some age, leading to sudden decrease and increase during the cycles when old planes stop flying and new planes start flying. Also, this may be due to the fact that the overall quality of the planes in a certain year is much more influential in the delay times, leading to sharp differences in each year/cycle.

4 Contribution

Everyone equally contributed in this project.

Reference

[1] https://en.wikipedia.org/wiki/North_American_blizzard_of_2006

[2] https://en.wikipedia.org/wiki/Hanukkah_Eve_windstorm_of_2006>

[3] https://en.wikipedia.org/wiki/2006_North_American_heat_wave

[4] https://en.wikipedia.org/wiki/Northwest_Airlines

License

Acknowledgments

- Statistics 480: Data Science Foundations
- **Professor Darren Glosemeyer**