Seed Scientific Candidate Assignment 1

U.S.A. Passenger Airline Study

Marc Dupuis
marc.f.dupuis@gmail.com
(980) 319-4717

November 12, 2013

1 Project Specifications and Objectives

The objective of this study is to investigate US passenger airline data using two datasets provided by Seed Scientific and the online Research and Innovative Technology Administration (RITA).

As stated in the prompt, the goal is to: *demonstrate to Seed Scientific my ability to manage a real-world data set, to communicate and contextualize summary statistics, and, most importantly, my creativity and analytical rigor in generating and testing hypotheses.* Furthermore, it is explicitly stated that the assignment is deliberately unspecific and exploratory by nature.

2 Introductory Note

Seed Scientific provided two datasets as CSV files which held information on the carrier identification tags and airport locations. In addition to this, I was directed to the RITA website[a], which holds data on US airline passenger flights. The available information is extremely diverse and rich, and contains data on departure dates, cities and airline as well as delays, cause of delays and cancellations.

The data from the RITA set is remarkably well organized and complete. There is no unexpected missing data and very little modifications had to be done in order to process the data.

3 Results

As an occasional passenger, the first questions that came to mind when I was given the data were: What do passenger flight trends look like? And what are the delay/cancellation trends?

The number of US domestic flights is illustrated in Figure 1 for every month from January 2000 to August 2013. Two traits immediately jump out: a significant drop in flights around September 2001 and a sustained but progressive drop beginning in September 2008 (as marked by the red lines). These two dates match the 9/11 terrorist attacks and the US financial market crash of 2008 respectively. The 9/11 attacks caused an 22.72% drop in flights from September to October 2001 and did not fully recover until a sudden surge around January 2003.

The blue lines mark other noticeable, yearly trends. There seems to be a surge in flights in summer and a drop around December (Christmas and New Year's eve). These trends were expected and are common knowledge.

Now that we've seen how the number of flights has evolved over the past 13 years, the next question is whether or not delays and cancellations have followed the same trends or if they have been increasing relative to the number of flights (I think we can all agree that it seems like every flight is delayed nowadays.) The most remarkable feature of cancellation and delays graphs is the "insignificance" of the number of cancelled flights relative to the total number of flights, as well as the sudden surge in cancellations around September 2001. Again, this surge marks the time of the 9/11 attacks. With regards to the delays, these seem to follow the same trends as the total number of flights. Given more time, several more advanced methods would do a better job confirming this,

[a] http://www.transtats.bts.gov

but the correlation between delayed flights and non-delayed flights is approximately 0.68 versus 0.04 for cancelled and on-time flights. This shows a rather strong correlation between number of flights and delays and a surprisingly low correlation between cancelled flights and other flights (including delayed flights).

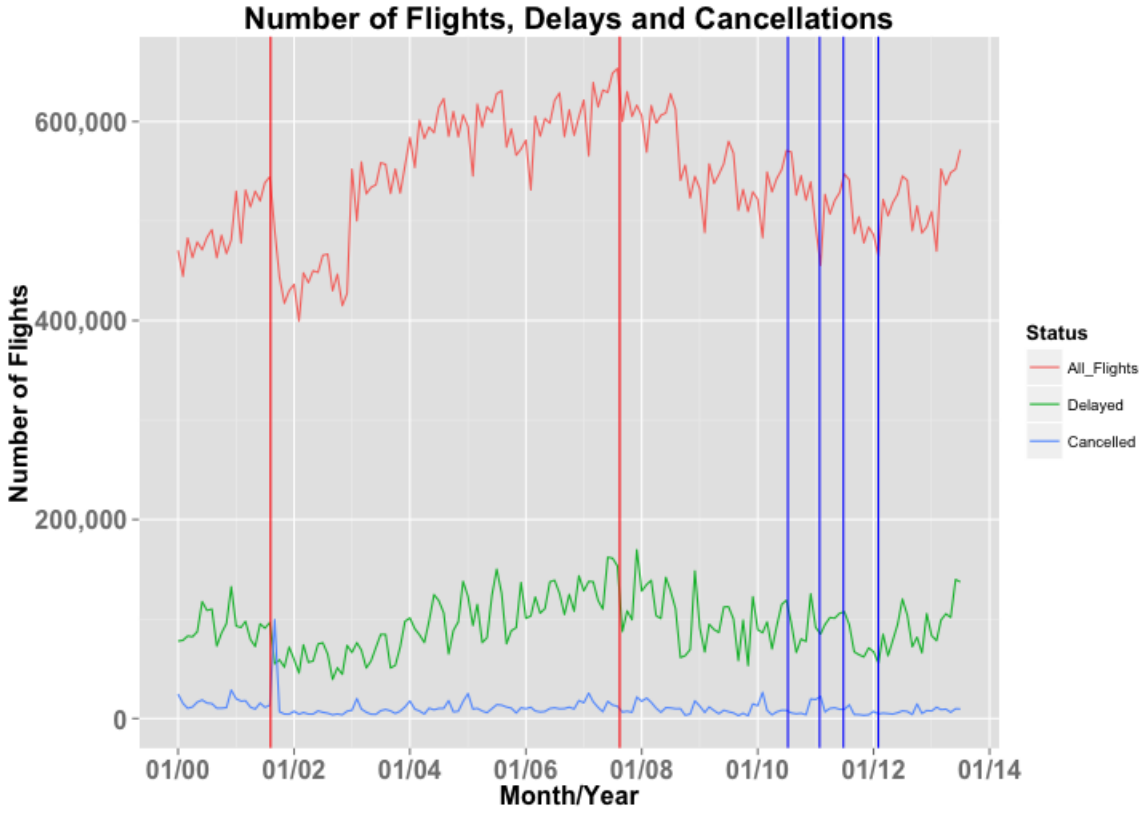

Figure 1: Flights, delays and cancellations from 2000 to 2013. The red lines indicate major events affecting the number of flights while the blue lines highly yearly trends.

Following the observations marked by the blue lines in Fig. 1, it would be interesting to look at the number of flights and delays by month (Fig. 2) The bar graph seems to support the idea of an increase of flights in the summer and a drop at the end of each year. However, the delays by month would be better expressed as a percentage of total flights. This is done in Fig. 3. This bar graph seems to suggest that delays are rather steady with a slight increase around the summer period and Christmas holidays.

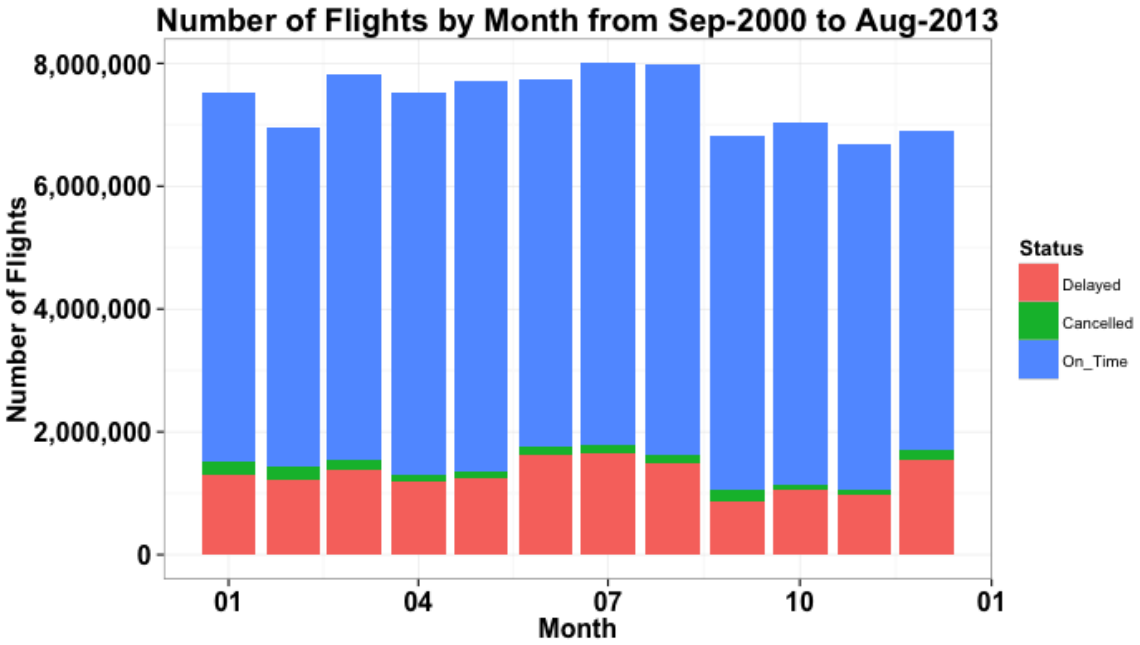

Figure 2: Delays by month from January 2000 to August 2013.

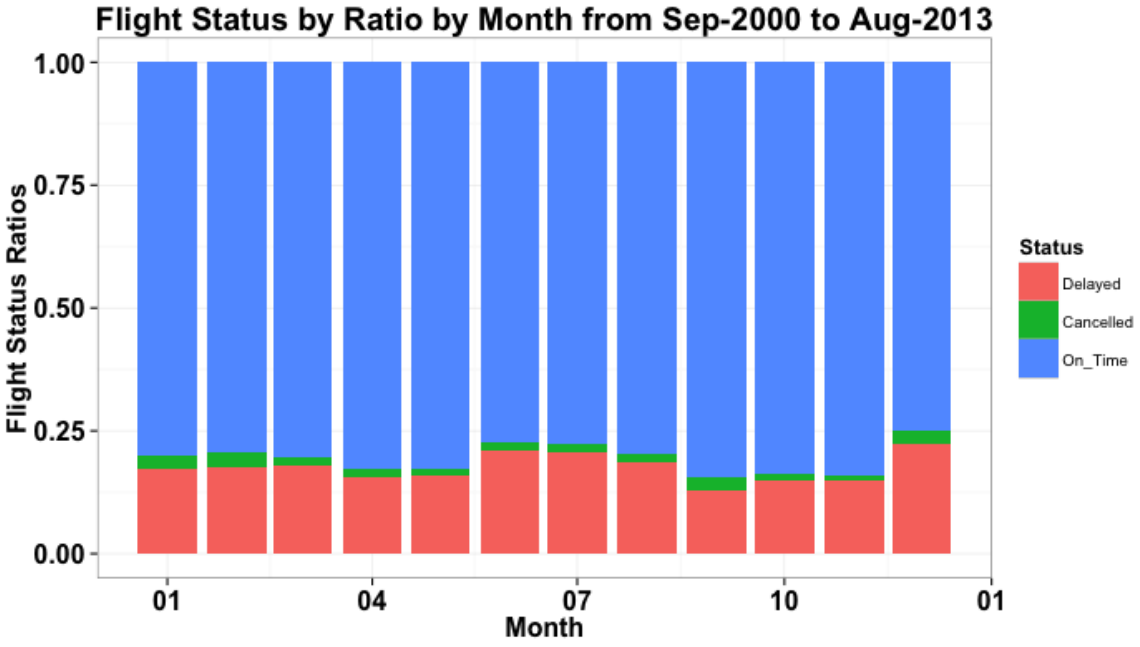

Figure 3: Delays by month expressed as a percentage of total flights each month.

Now that we understand more about flight trends and delays/cancellations, it would be interesting to look at the causes of delays. Is there one airline that performs better than the others? Is it better to depart from a state known for good weather? Fig. 4 breaks up the causes of delays into

five categories:

1. Late aircraft
2. National Airspace System (NAS) delay
3. Carrier delay
4. Weather
5. Security

With this figure however, it seems as if weather is not a frequent cause for delay, even though we would intuitively think that it is one of the main causes. This is because the NAS only categorizes delays as due to weather if the weather is severe[b]. Minor weather delays fall under NAS orders and thus contribute to the NAS section of the pie chart in addition to heavy air traffic and airport operations.

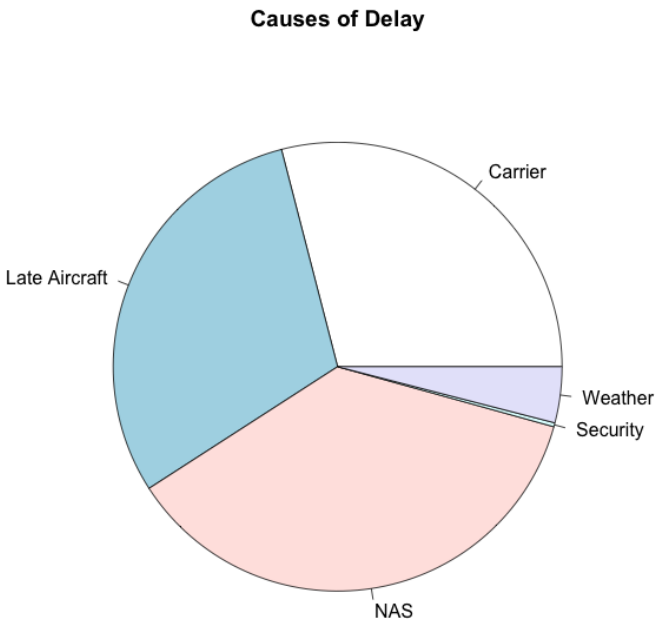

Figure 4: Pie chart illustrating main causes of US flight delays.

Now lets look at the delays by carrier relative to the total number of flights as illustrated in Fig. 5. Atlantic Southeast Airlines, JetBlue and Southwest Airlines are the three worst performing airlines, while Hawaiian Airlines and Aloha Air Cargo are the best performing and perform exceptionally well.

[b]http://www.rita.dot.gov/bts/help/aviation/html/understanding.html

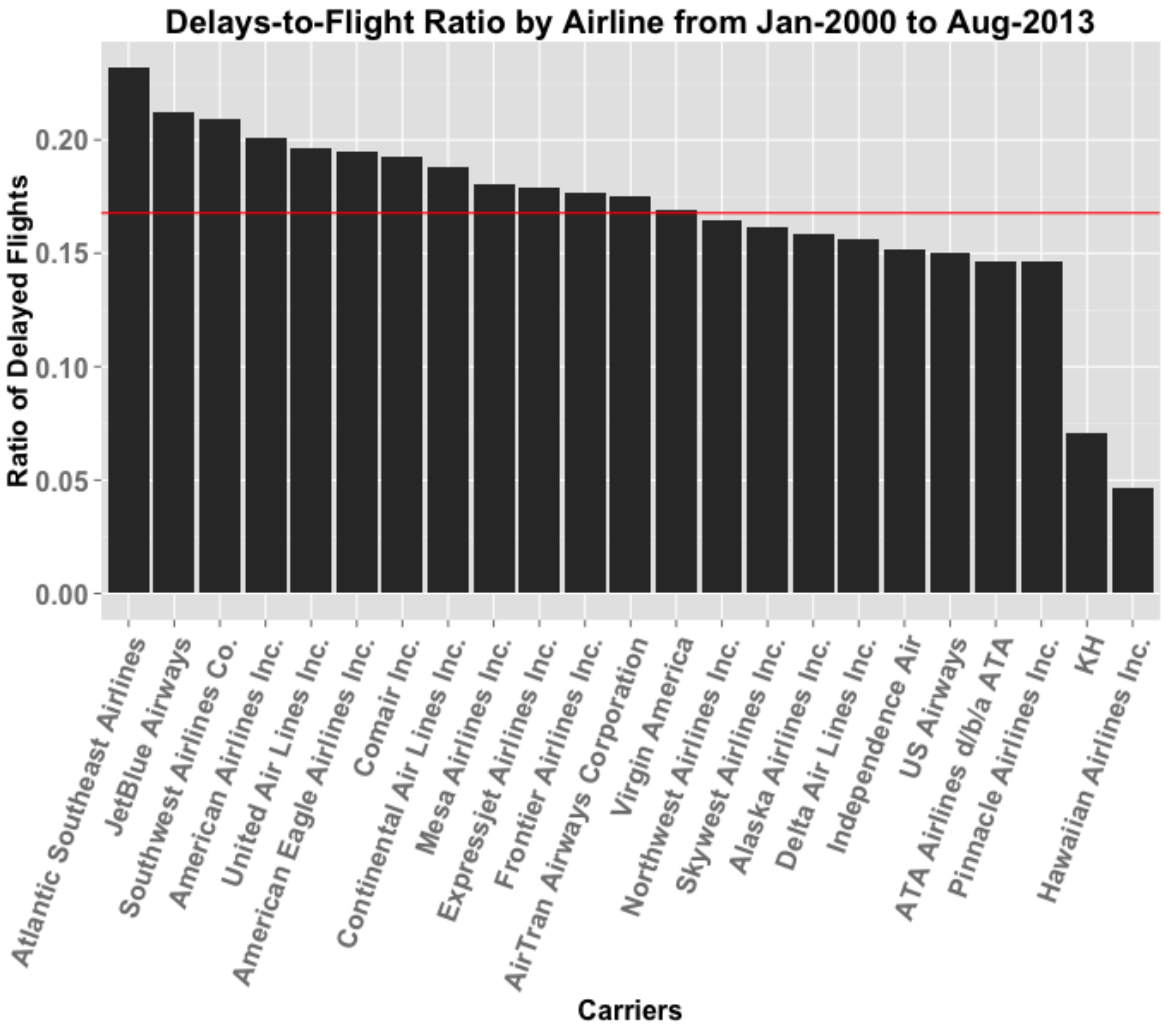

Figure 5: Delay-to-flight ratio by carrier.

So we now know which airlines to avoid in terms of departure reliability and which to recommend. Lets now look at the relationship between delays and departure state. Fig. 6 shows a fairly spread out distribution amongst all fifty states, Puerto Rico (PR), the Trusted Territories (TT) and the U.S. Virgin Islands (VI). As with the delays by carrier, we should look at this data relative to the number of flights departing from each of these states (Fig. 6). The best performers turn out to be Hawaii, Montana and Idaho. On the other hand, the worst performing states are the Truste Territories, Illinois, New Jersey and Delaware (as illustrated in Fig. 7.)

Finally, given the correlations that seems to exist between flight delay and carrier, month and departure state, if there is a significant connection between these features, a classifier should outperform random chance. Indeed, by applying a random forest, we can successfully determine whether or not a flight will be delayed with an 80.16% success rate[c]. However, the Receiver Operator Char-

[c]Additional features used to the ones mentioned are arrival state and distance between departure and arrival states,

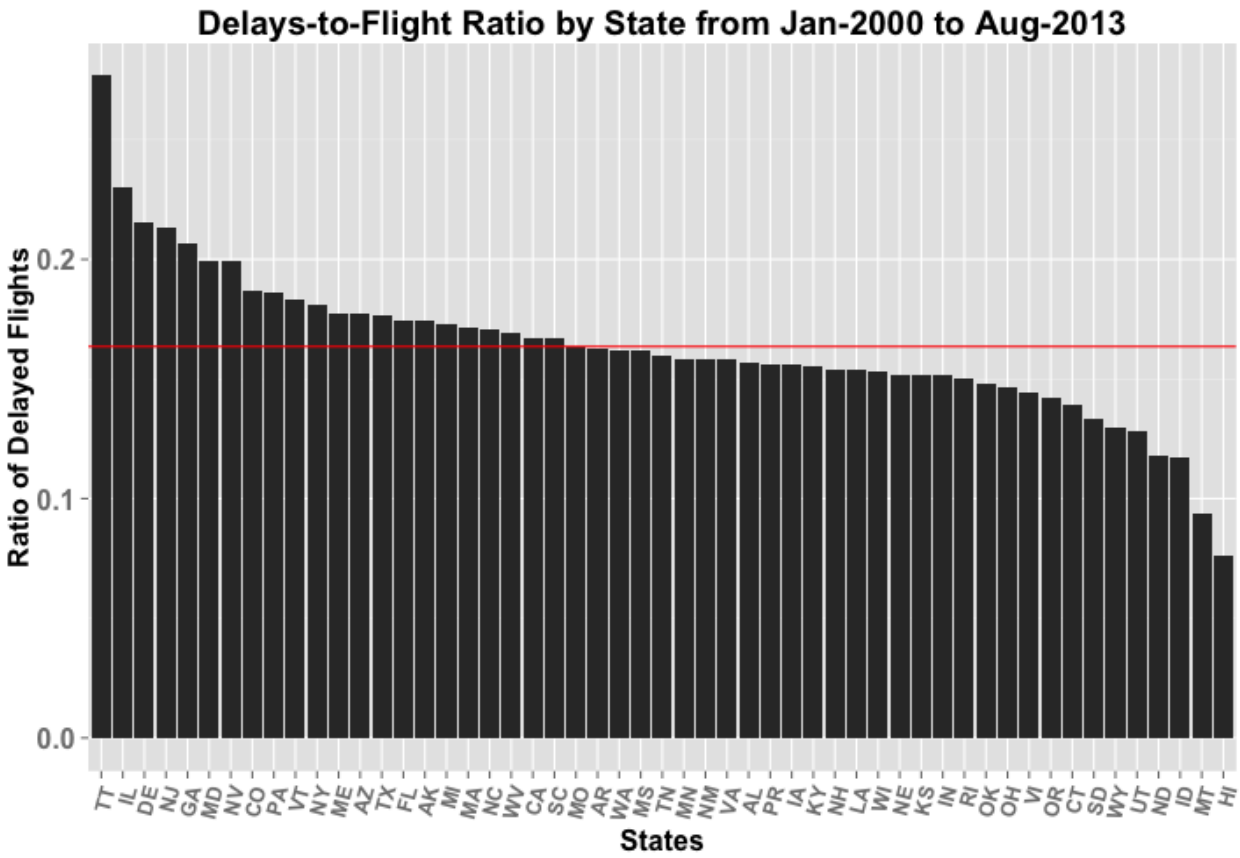

Figure 6: Delay-to-flight ratio by state.

acteristic (ROC) curve has an area of 0.607, which indicates that the classifier performs better than a random guess, but is not highly reliable.

4 Conclusion

There are several conclusions to draw from this investigation. First, as expected, customers flight patterns are affected by major disasters and the state of the economy. Preliminary indications seem to suggest that delays have not been increasing relative to the number of flights over the years and flight cancellations are relatively constant across time (aside from major disasters).

Second, delays are to be expected around the high traffic months during the summer as well as during the holiday season in December (we know how hard it can be to make it home for Christmas). However, the cause for delays is equally due to late aircrafts, carriers and NAS orders. It is difficult to truly assess the cause of a delay in the system. For instance, a late aircraft could be caused by weather in the state it came from or the carrier.

and many more can and should be looked into. The classifier is also based on a significantly reduced dataset composed of 6.25% of the available data.

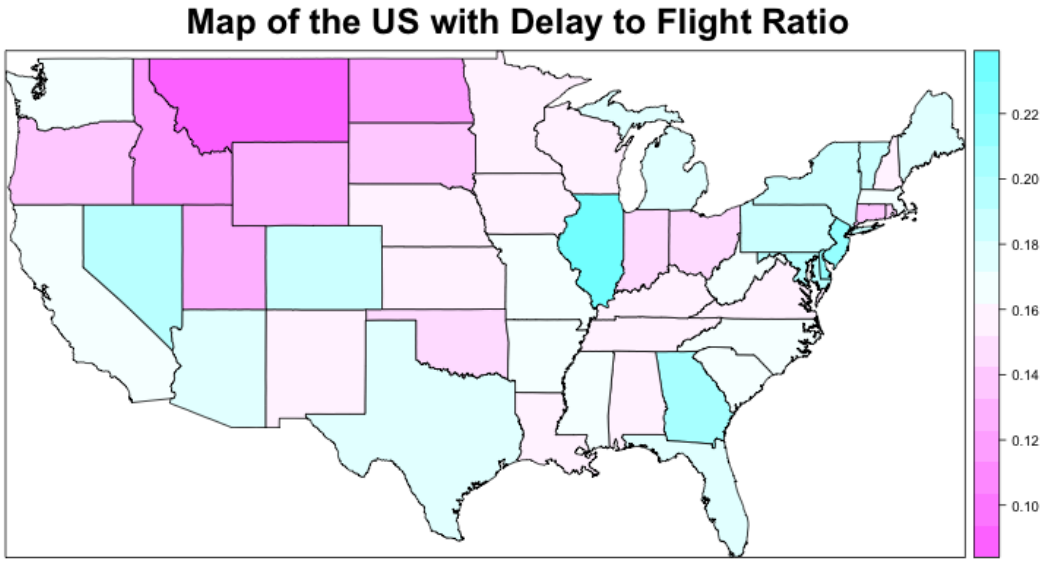

Figure 7: Heatmap illustrating the delay-to-flight ratio for each state on the continent with Montana and Illinois on opposite ends of the spectrum.

Finally, the scenario in which a passenger is least likely to experience a flight delay is by travelling from Hawaii on Hawaiian Airlines in September[d]. Similarly, a passenger is most likely to experience a delay if they travel at Christmas and they depart from Chicago with Southwest Airlines.

This study is very limited and is based on a number of assumptions which would have to be carefully considered in future investigations. The intent is to offer an initial perspective on flight patterns in the US and flight delay trends.

[d]It's obviously not this straight forward as there are other important factors such as weather, human and natural disasters, day of the weak etc.

Appendix

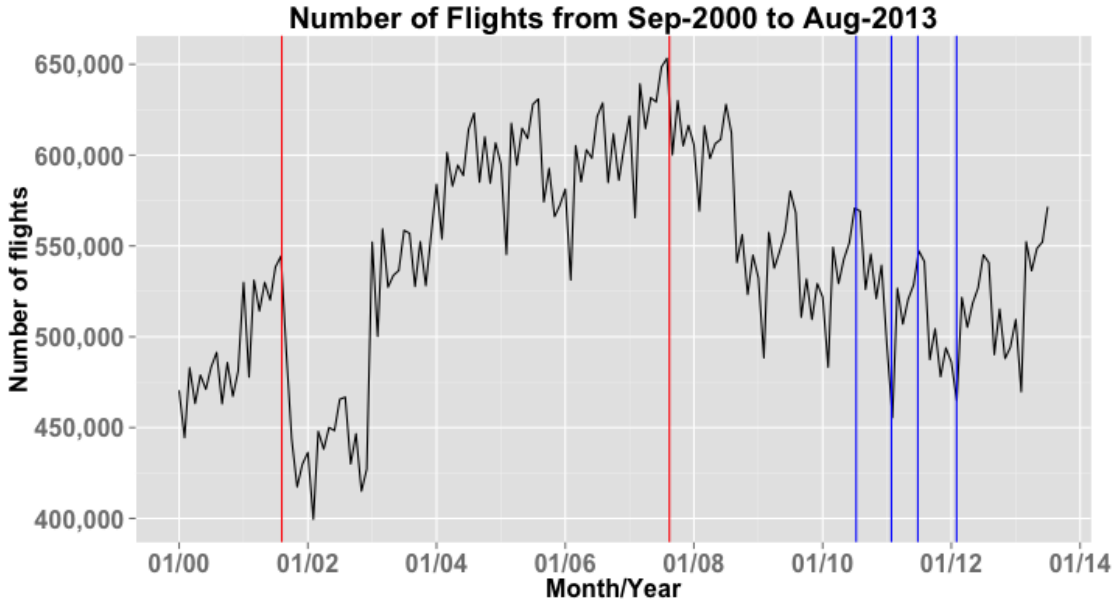

Figure 8: Number of US domestic flights from January 2000 to August 2013. The red lines indicate major events (9/11 terrorist attacks and beginning 2008 financial market crash after the subprime mortgage buble) and the blue lines indicate noticeable yearly trends.

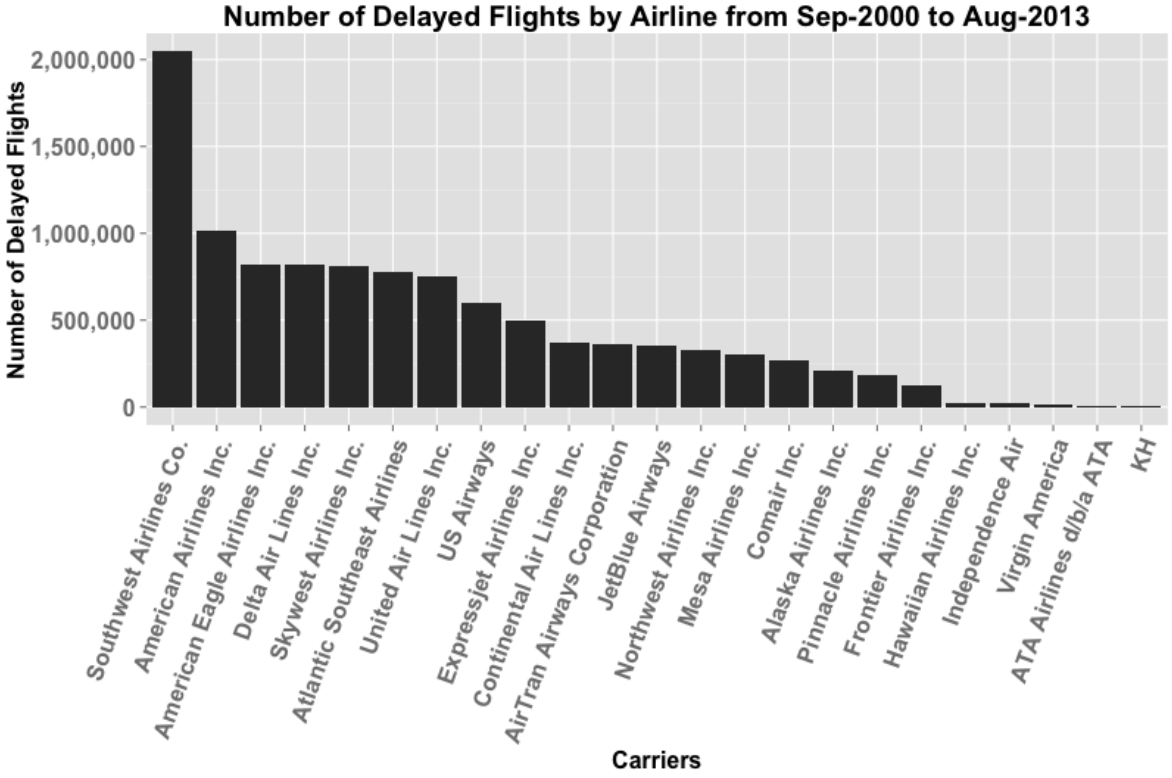

Figure 9: Total delays by carrier since January 2000. Southwest has over twice the amount as the carrier ranked by number of delays.

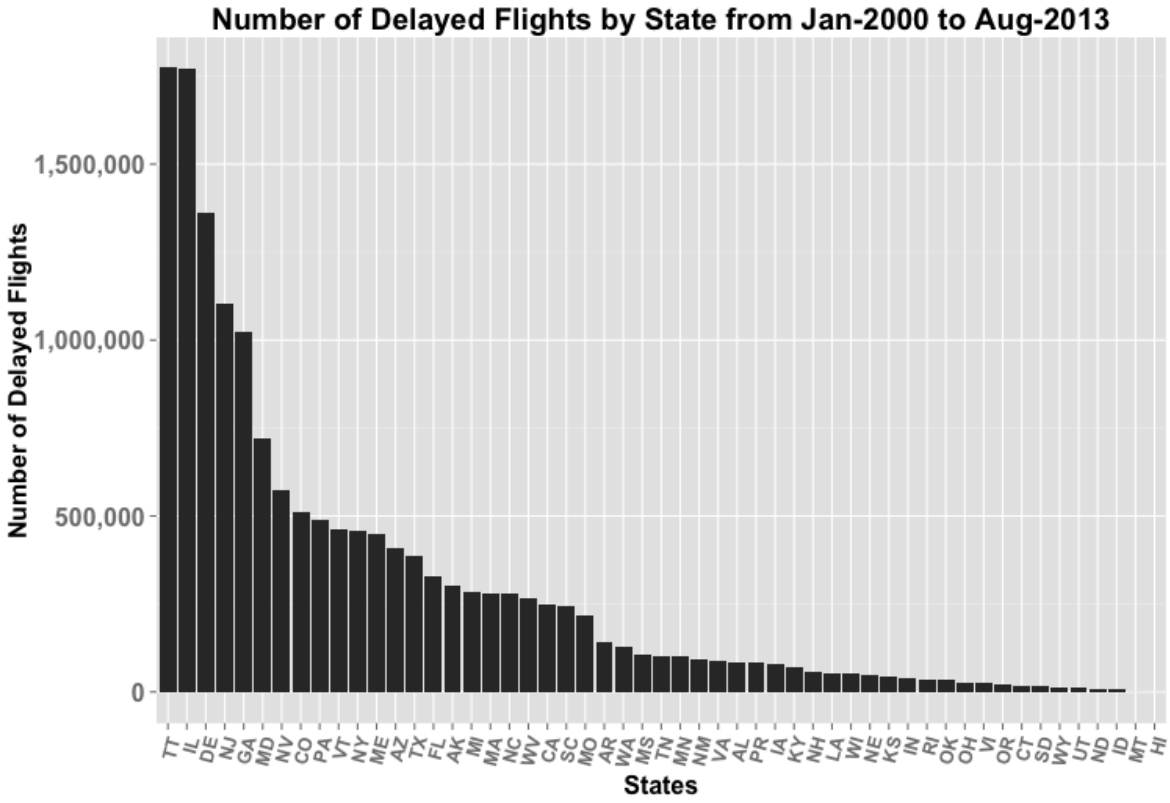

Figure 10: Delays by departure state. The top six are somewhat expected: Texas, California, Illinois, Georgia, Florida and New York. These six states host six of the major airport hubs: Houston, Los Angeles (LAX), O'Hare, Atlanta, Miami and JFK.

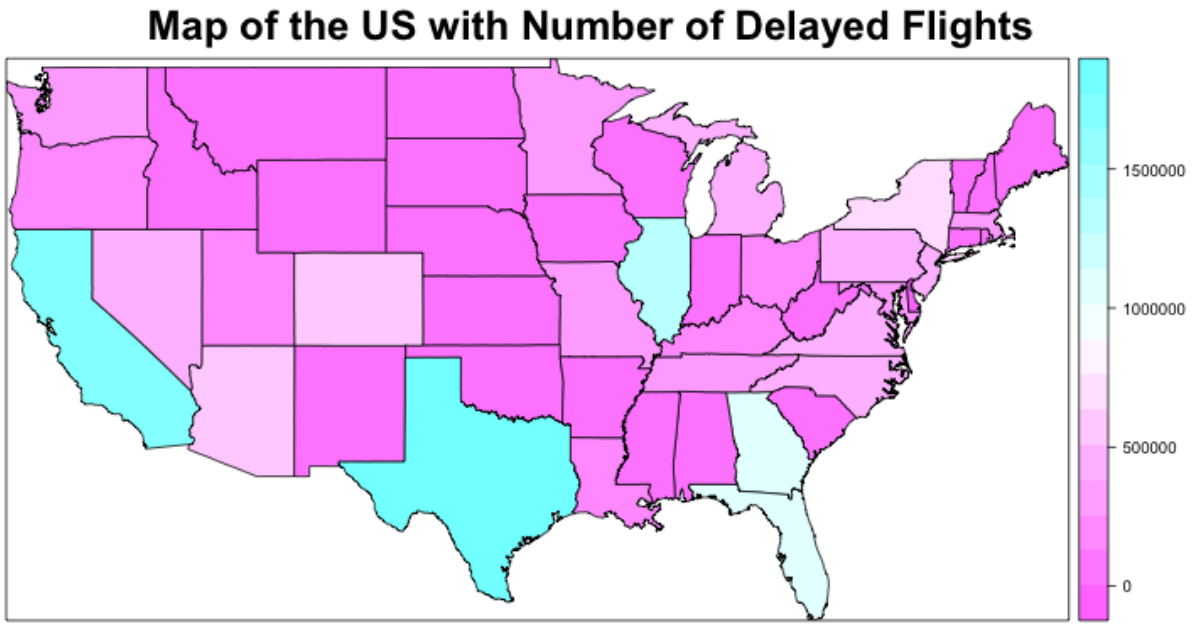

Figure 11: Heatmap illustrating the number of delays by state since January 2000.