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 project 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 18.18% drop in flights from September to October 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.) Fig. 2 is the same as Fig. 1 with the delayed and cancelled flights added to the figure. The most remarkable feature of this graph 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 911 attacks. With regards to the delays, these seem to follow the same trends as the total number of flights. Given more time, several

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

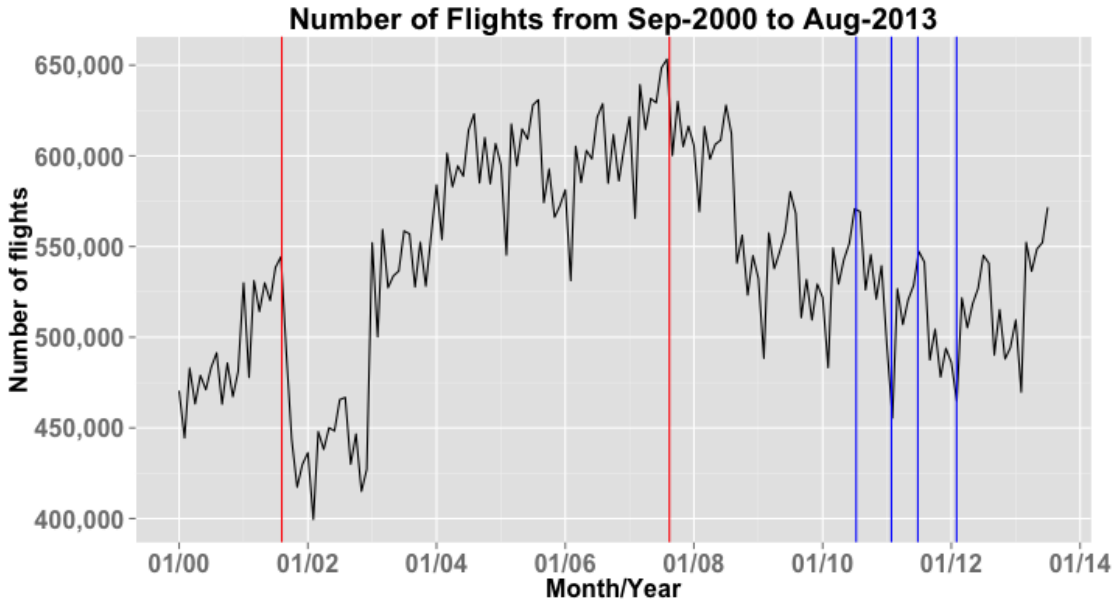


Figure 1: Number of US domestic flights from January 2000 to August 2013. The red lines indicate major events (9/11 terrorist attacks and beginning 2008 financial market crash after the subprime mortgage buble) and the blue lines indicate noticeable yearly trends.

more advanced methods would do a better job confirming this, 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).

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. 3) 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. 4. This bar graph seems to suggest that delays are rather steady with a slight increase around the summer period and Christmas holidays (again, common knowledge).

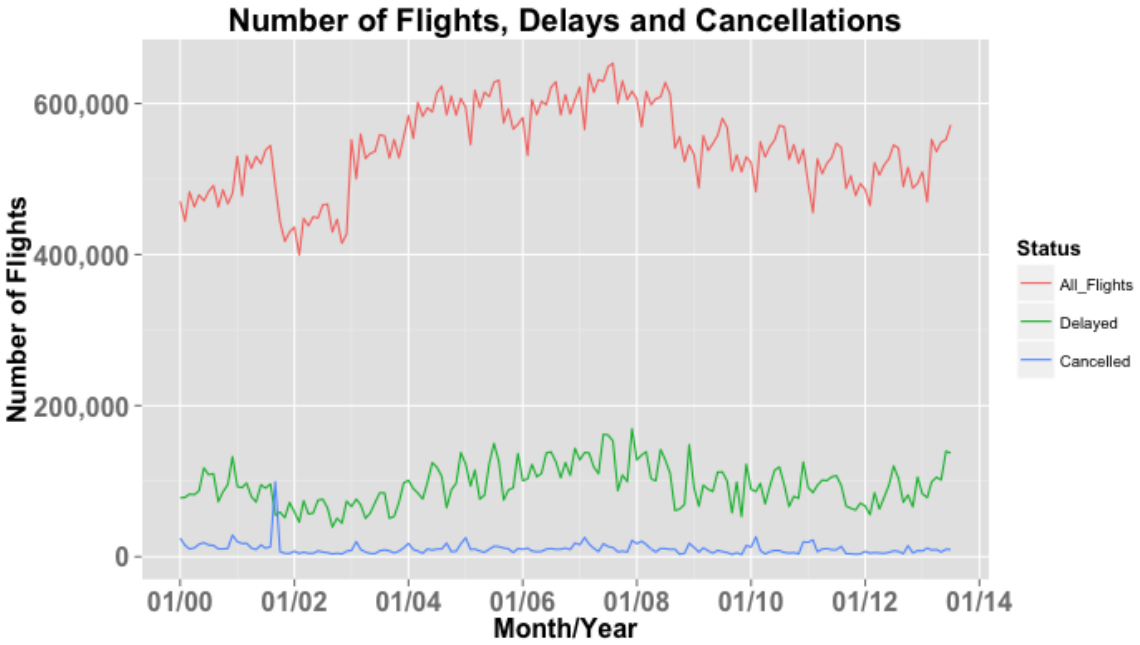


Figure 2: Example of Ixico image processing flowchart.

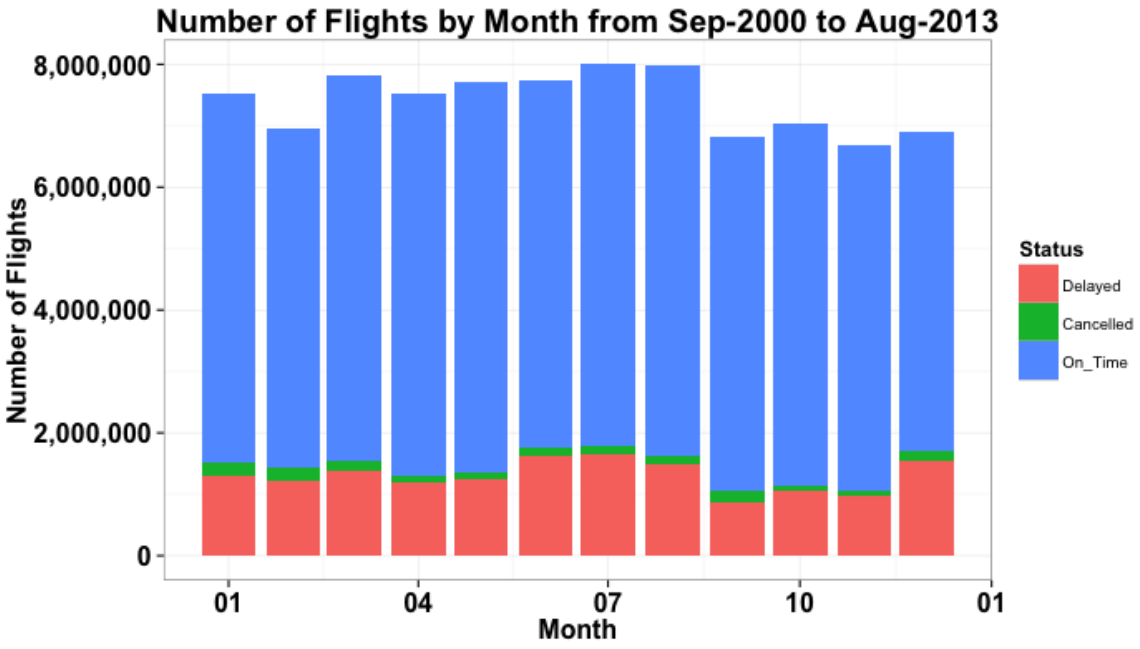


Figure 3: Example of Ixico image processing flowchart.

Now that we understand more about flight trends and delays/cancellations, it would be interesting to see what the major causes of delays are. Is there one airline that performs better than another? Is it better to depart from a state known for good weather? Fig. 5 breaks up the causes

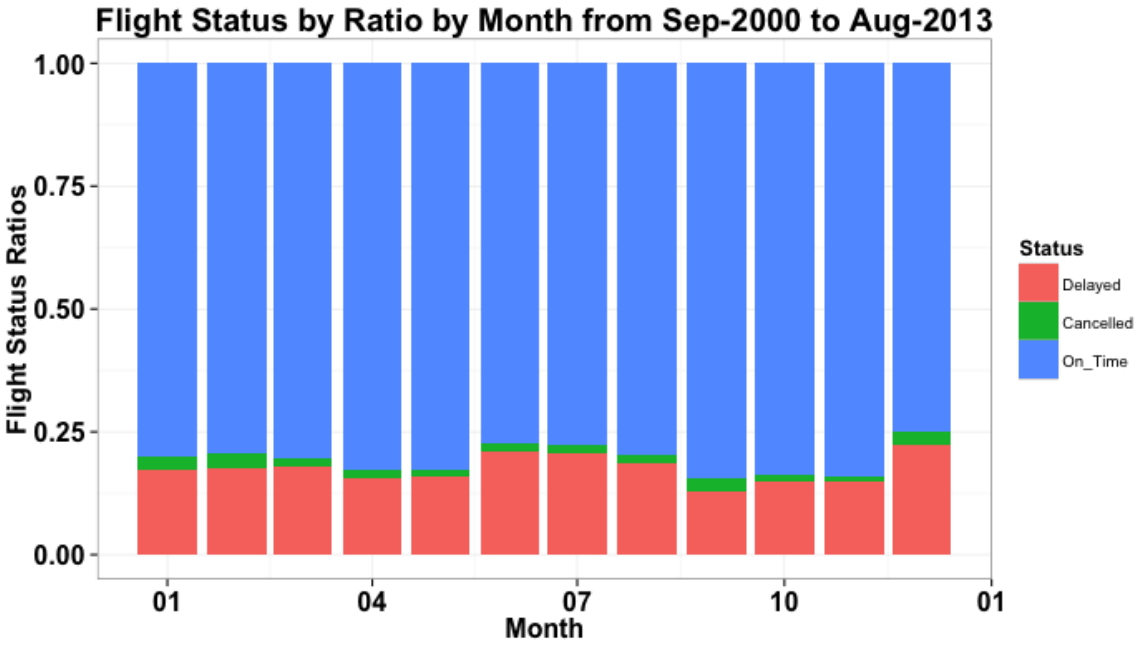


Figure 4: Example of Ixico image processing flowchart.

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.

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

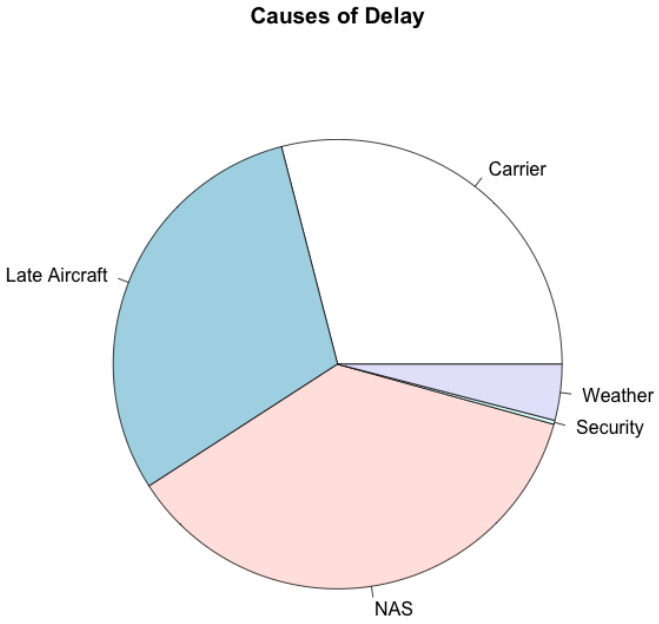


Figure 5: Example of Ixico image processing flowchart.

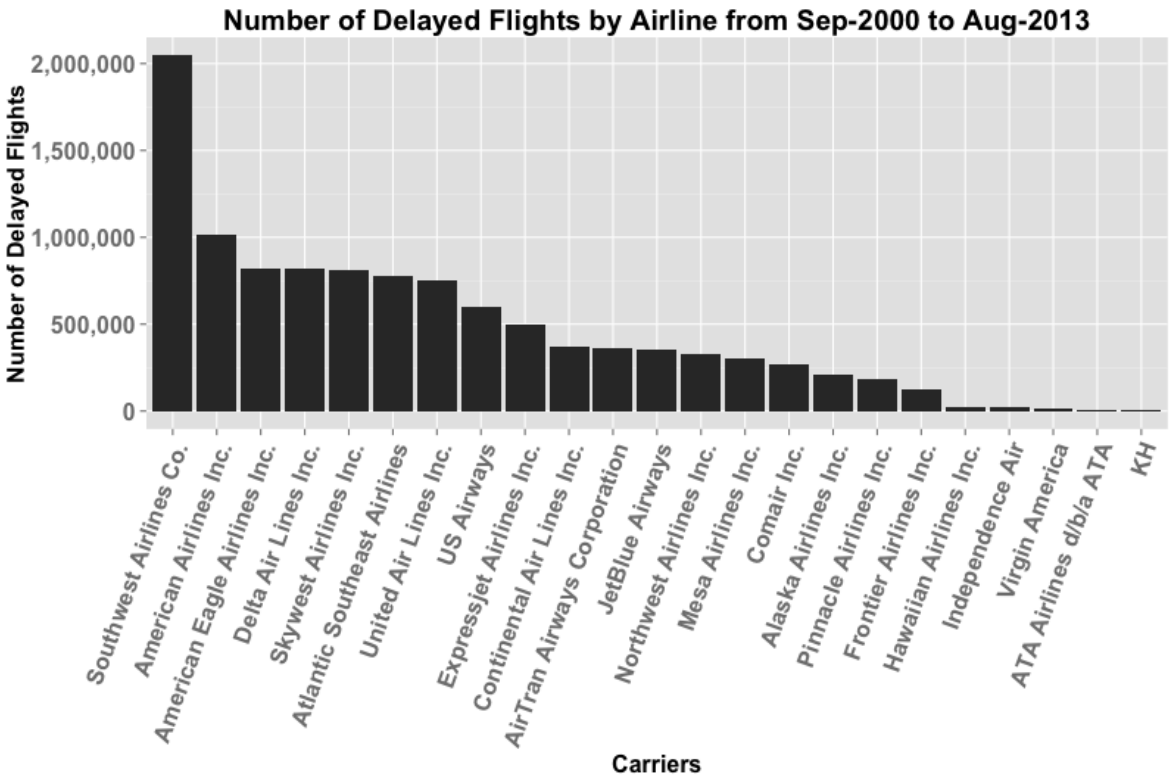


Figure 6: Example of Ixico image processing flowchart.

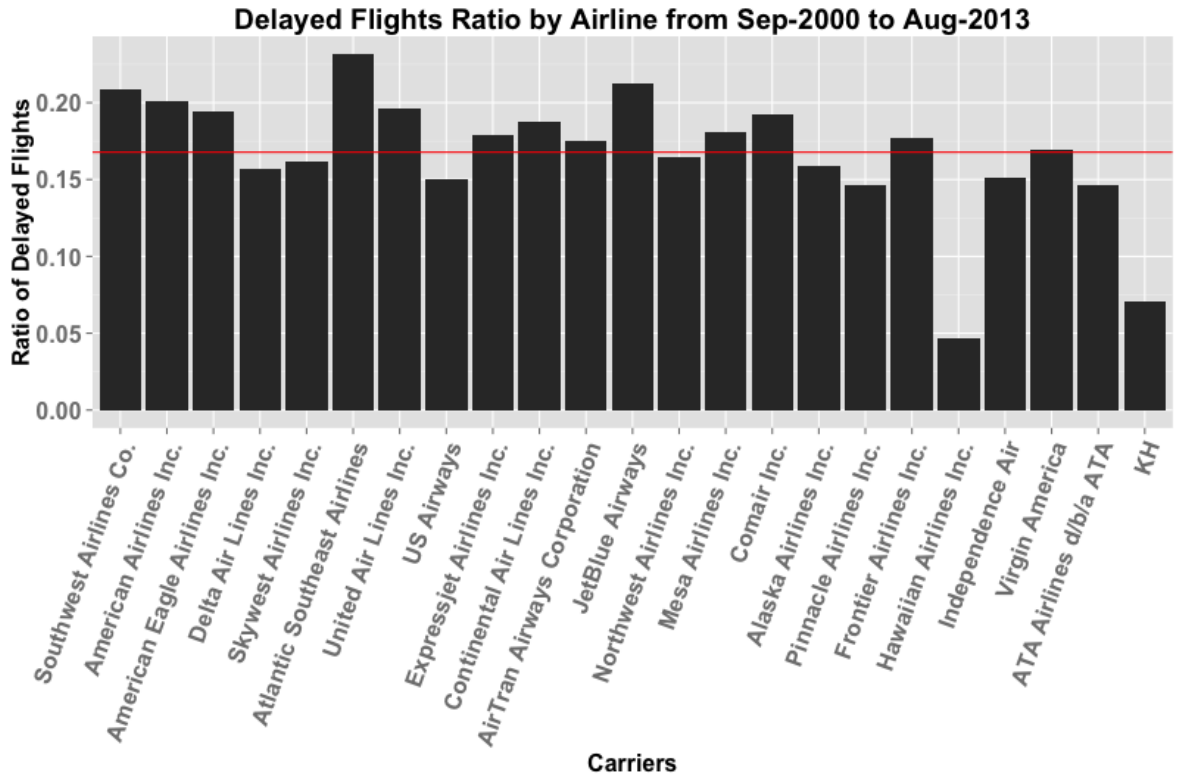


Figure 7: Example of Ixico image processing flowchart.

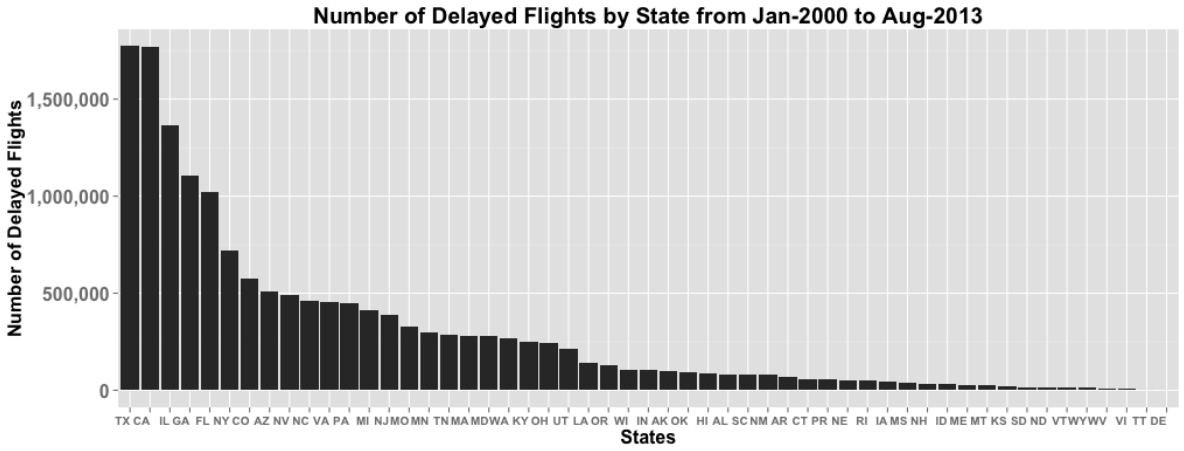


Figure 8: Example of Ixico image processing flowchart.

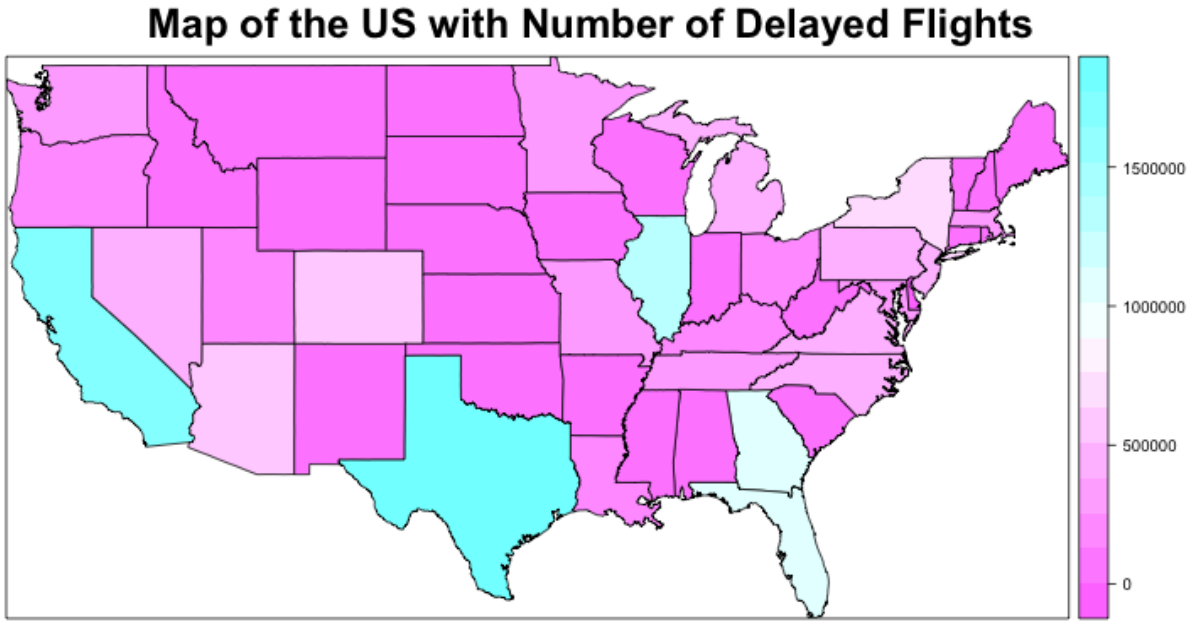


Figure 9: Example of Ixico image processing flowchart.

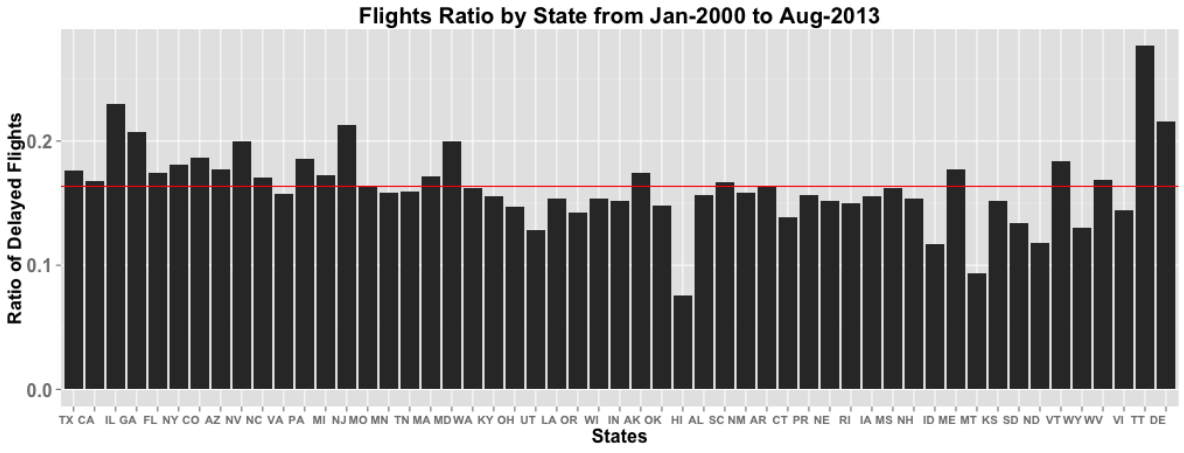


Figure 10: Example of Ixico image processing flowchart.

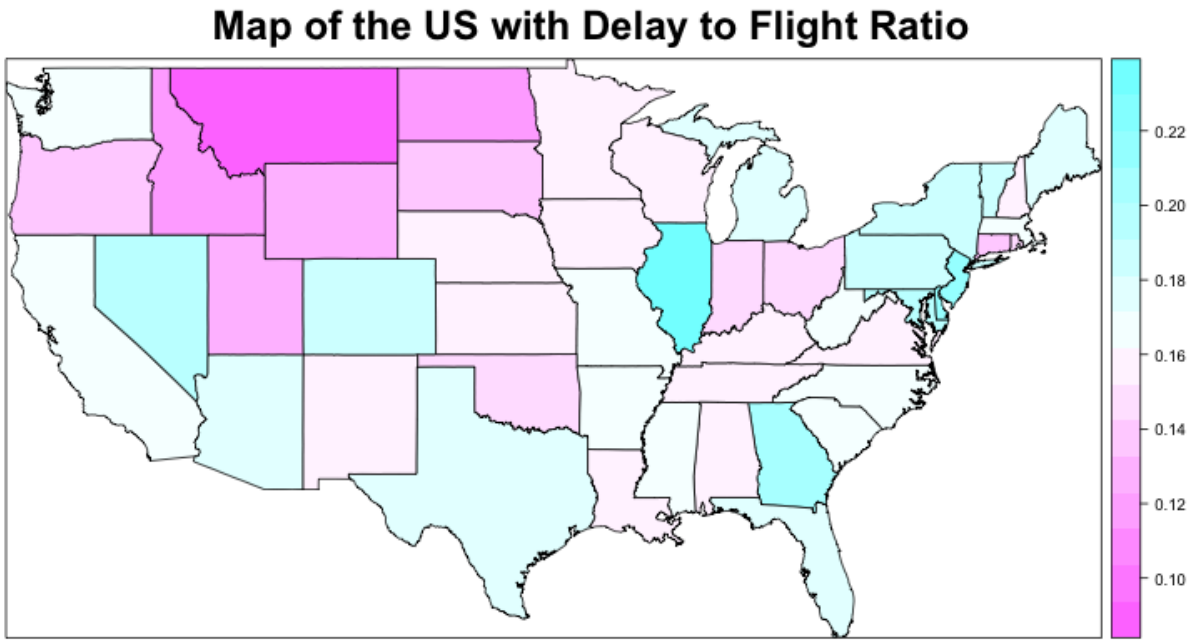


Figure 11: Example of Ixico image processing flowchart.