To Fly or Not to Fly?
(Or Analysis of Flights in the United States with a Focus on Delay Statistics)

Amirpasha Shirazinia
email: amirpasha.shirazinia@signal.uu.se

June 24, 2016

Abstract

In this report, we investigate flight data and statistics provided by United States Department of Transportation, the Bureau of Transportation Statistics. We basically analyze the volume of flights as well as the volume of delayed and canceled flights over the interval 2000-01-01 till 2015-12-31. We also analyze delay statistics during this interval with respect to airline companies and destinations. Finally, we will try to answer these questions: What is the trend of flight volume and delays during this interval? When is it more likely for a flight to get delayed or canceled? How is the performance of an airline company? How busy a destination?

1 Introduction

The US Bureau of Transportation Statistics has provided data on flights operated by passenger airlines between October 1987 and April 2016[1]. These comprehensive statistics represent general data, e.g., name of airline carriers, origin/destination cities and states, flight dates, etc., as well as performance statistics, such as delay in minutes, canceled or diverted flights, etc.

1.1 Motivation

As *flight delays* highly determine the quality of trips, we focus to a grater extent on the performance of flights with respect to this criterion. We also quantify the flight performances with respect to volume of flights operated by airlines. In general, delays are comprised of arrival and departure delays, which, according to the Bureau of Transportation Statistics, can be caused by carrier delay, weather delay, national air system delay, security delay and late aircraft delay. Throughout our analysis, we only consider arrival delay, since a typical traveler might be affected by this type of delay more frequently which can be crucial. For example, it is important if she/he shows up in a meeting at a destination in-time, or if she/he wants to enjoy her/his planned holiday travel, etc. Delays and volume of flights can have a great impact on the rating and ranking of airlines as well, which potentially leads to growth or downsize and/or profitability or loss.

1.2 Objectives and General Statistics

Following these motivations, we summarize the objectives of our report as follows:

- Analysis of volume of total/delayed/canceled/diverted flights.
- Analysis of delay statistics with respect to airline companies and destinations.

We look at the statistics in the interval of 16 years from 2000-01-01 till 2015-12-31. The amount of data that we investigate adds up to 46GB of storage comprised of 192 (16×12) csv files. Table 1.2 illustrates some general statistics extracted from the data under consideration.

[1]These statistics can be downloaded on monthly basis via `http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time`.

Table 1: General flight statistics

Period	Total flights	Average number of flights/year	Average number of flights/month
2000–2015	$\sim$ 102 M	$\sim$ 6.39 M	$\sim$ 533 K

2 Analysis of Volume of Flights

In order to offer some insights into evolving flight statistics, in Figure 1, we demonstrate the number of total flights as well as delayed[2], canceled and diverted flights per month during the studied time period.

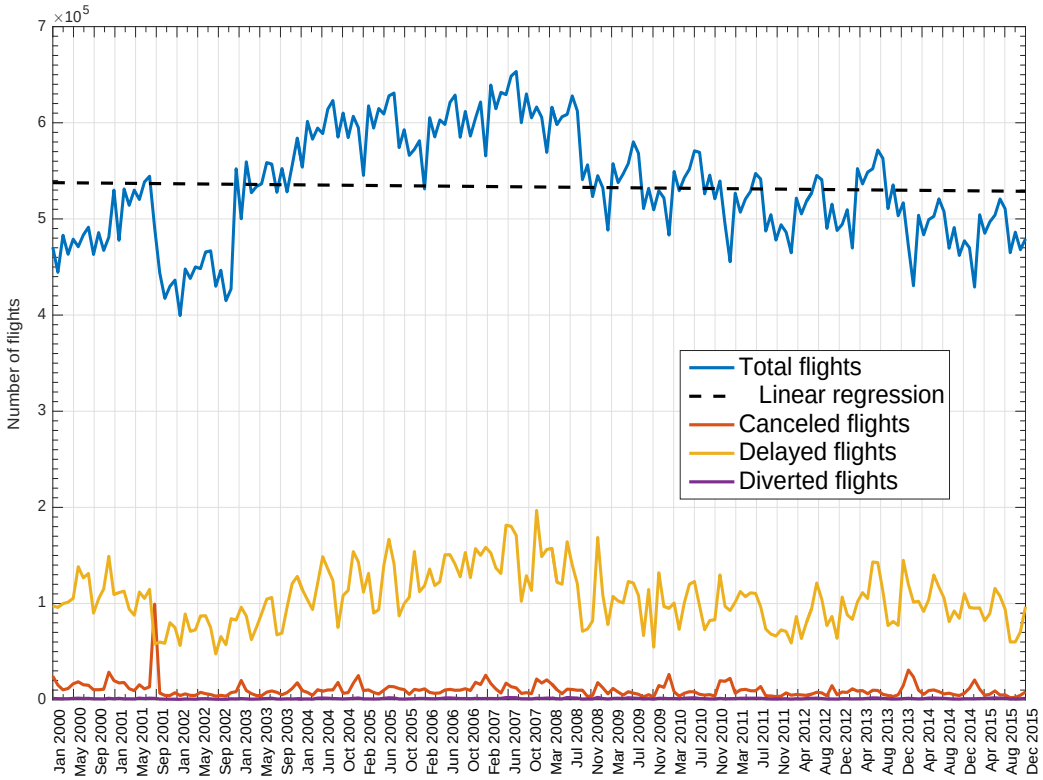

Figure 1: Volume of total flights, delayed, canceled and diverted flights from January 2000 to December 2015.

Observing Figure 1, some important events in contemporary US history can be noticed. On September 2001, the minimum number of flights, amount to 399 500, is operated by airlines which is related to the 9/11 terrorist attacks. Compared to its previous month, August 2001, the number of flights dropped off by 26.6%. These incidents impact the number of flights until early 2003. The number of flights rises to the maximum point, i.e., 653 300, on August 2007, the summer before the US financial crisis, which drastically affected US airline industry. For example, the number of flights from July 2008 dropped off by 7.6% compared to the same month in 2009. According to [1], between May 2008 and 2009, jobs were cut at several major airlines across the US, such as, United Airlines (6 600), American Airlines (4 900), Northwest Airlines (4 300). We also perform linear regression on the number of flights, which shows that it drops by 1.65% from January 2000 to December 2015. There are several reasons to this: expensive airline tickets, comfort in traveling by car, customer service rating, heavy security control, development of audio/video calling technology, just to name a few. In Figure 1, it can be also observed that in December 2007, maximum delayed flights, amounts to 196 800, have happened. According to CNN [2], 2007 set to be worst year for delays due to severe weather conditions. We also observe

[2]An arrival delay of more than 15 minutes is considered a delayed flight.

that the 9/11 incidents caused 99 320 canceled flights.

It is also interesting to track the percentage of delayed/canceled/diverted-to-total flights. In Figure 2, we demonstrate these ratios. Here, we note that although the low percentage of delayed flights around September 2001 is due to the low volume of flights and high percentage of flight cancelations. A linear regression method estimates that delayed flights drop off almost 1% from 2000 to 2015. Although this is good news (partly because of the drop in total flights), considering the whole duration from 2000 to 2015, on average, almost 20% of total flights have delay above 15 minutes, and also, on average, almost 2% of total flights have been canceled. In both figures, the number of diverted flights are much less than delayed or canceled ones.

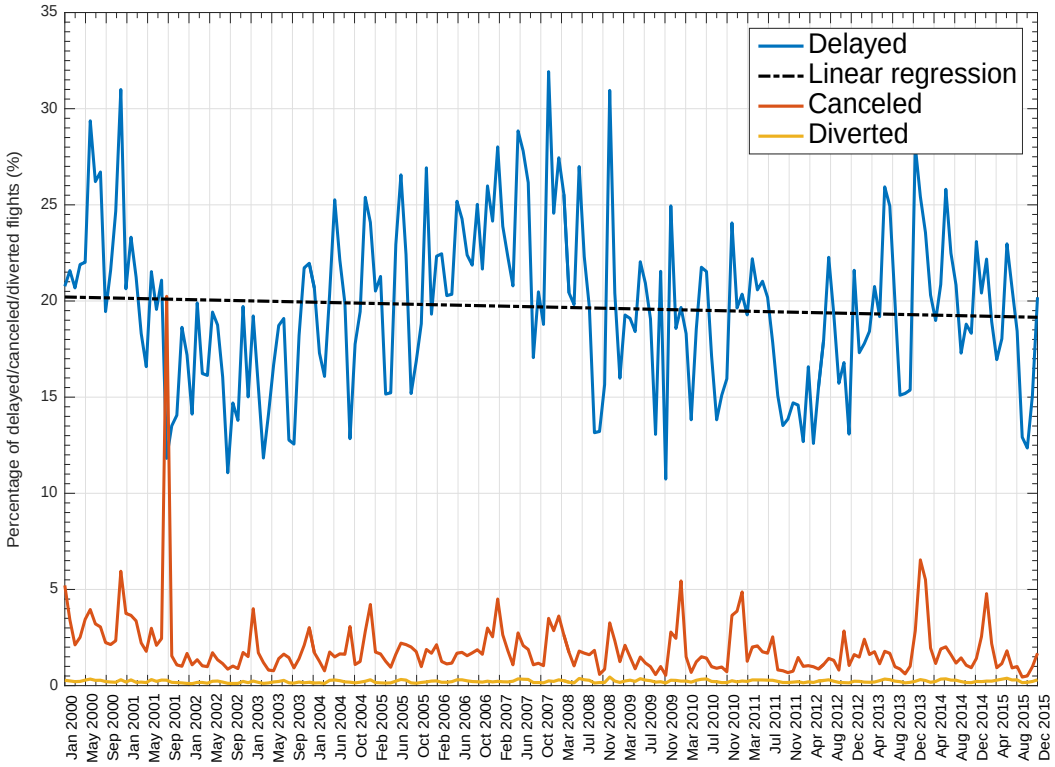

Figure 2: Percentage of delayed/canceled/diverted-to-total flights from January 2000 to December 2015.

From Figure 1 and Figure 2, we see that commonly in some months, there are peaks and in some months there as valleys in the volume of total, delayed and canceled flights. This behavior, as shown in Figure 3, can be distinguished in a more precise way by averaging the studied data over months during the period of 2000-2015.

There are interesting results that can be drawn from Figure 3. For example, as US locates in the northern hemisphere, people normally go on vacations in July and August. That is why there are large volume of flights operated by airlines and consequently more delays. As weather condition is stable in these month, the number of canceled flights is low, compared to those of e.g., December, January and February. More specifically, in February, on average, minimum amount of flights are operated by US airlines in the past years. Note also that the statistics in September is to some extent are influenced by the 9/11 incidents.

Next, we demonstrate a figure which determines the reliability of flights in the US in terms of robustness against delays and cancelations. For this purpose, empirical cumulative distribution functions (CDFs) of the number of delayed/canceled/diverted flights are plotted in Figure 4. Noting the median on the $y-$axis, we observe that with probability 50% almost 10 000 flights per month are subject to delay with more than 15 minutes. With the same probability, less than 8 800 flights are subject to cancelation. In addition, although the CDF of canceled flights is steep, rising to 1, the CDF of delayed flights does not have the same behavior

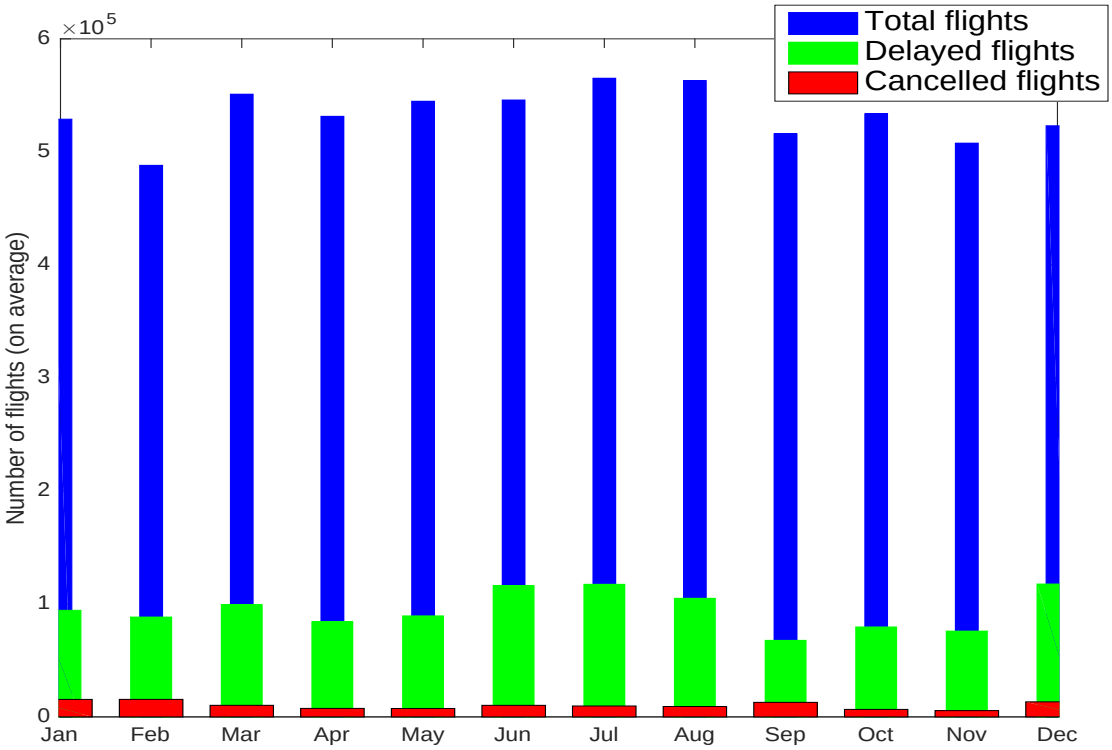

Figure 3: Average number of total/delayed/canceled flights per month

which shows the unreliability of flights to delays.

To emphasize the importance of delay in our analysis, we show another important figure which reveals the probability that a flight has a delay more than a given time amount. In Figure 5, we demonstrate the complementary CDF of length of delays during June, July and August 2015. This recent period is chosen as it is a holiday period for travelers in US. The intersecting line shows that more than 7% of flights have delays more than 60 minutes in these months.

These statistics show that travelers' vacations in summer could be annoyingly interrupted with lengthy flight delays. Hence, it seems critical to analyze the performance of airlines and destinations in terms of flight delays.

3 Analysis of Delay by Airlines and Destinations

In the previous section, we saw that the number of delayed flights as well as delay lengths are critical. Therefore, in this section, we focus on the volume of flights and the number of delayed flights with respect to airline carriers and destination states in order to find out which airlines and destination contribute more into the number of delayed flights. Moreover, it also reveals the performance of airlines.

For the sake of comparison, we grouped the whole 16-year period into two intervals, from 2000 to 2007, and from 2008 to 2015. From now on, we only consider all arrival delays more than 60 minutes, as it can be devastating for travelers, and it is a threshold where airlines are starting to get low-quality ratings.

In Figure 6 and Figure 7, we first show the number of flights operated by various airlines[3] in a sorted manner. It can be seen that Southwest Airlines (WN), the world's largest low-cost carrier, in both periods has the first rank in carrying out large volume of flights. The airlines has largely increased the volume of flights from the period of 2000-2007 to the period of 2008-2015, by more than 13%. The next rank, with a huge gap in flight volume, belongs to American Airlines (AA) and Delta Airlines (DL). It is also interesting to observe

[3]Airline abbreviations are selected based on IATA codes. Please search the airline code in `http://www.iata.org/publications/Pages/code-search.aspx`.

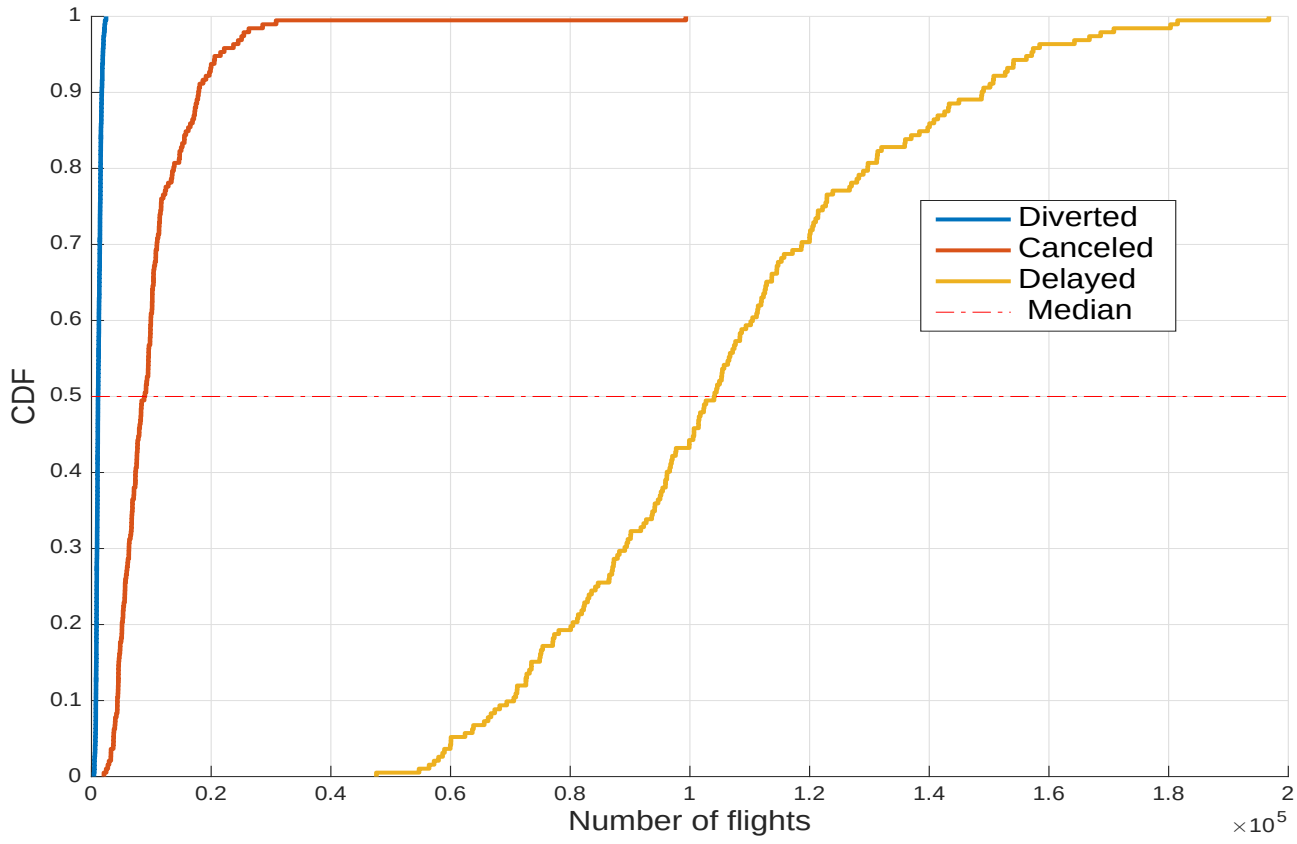

Figure 4: CDF of average delayed/canceled/diverted flights per month in the period 2000–2015.

that SkyWest Airlines (OO) has a huge rise in ranking in carrying out flights. Finally, the lowest rank belongs to Aeko Kula (KH) based in Hawaii.

To observe the performance of airline carriers, inFigure 8 and Figure 9, we illustrate the ratio between the number of delayed flights (+60 minutes) and number of flights for each airline. According to this criterion, the lower the value gets, the better the airline is. The best ranks belong to Hawaii-based airlines, Aeko Kula (KH) and Hawaiian Airlines (HW). The situation of Southwest Airlines (WN) is also very good knowing its high volume of flights. Delta Airlines (DL) and American Airlines (AA) have also stable situations. In the period of 2000-2007, Mesa Airlines (YV) and in the period of 2008-2015, Spirit Airlines (NK) got the lowest ranks. The latter is also one of ultra-low-cost carriers.

In the next figures, Figure 10–Figure 13, we look into the status of flight destination states, and compare them in terms of number of flights and delays in a sorted manner. The figures reveal that California (CA), Illinois (IL) and Texas (TX) are the busiest states in terms of number of destination flights and delays which was expected due to their sizes and populations. It can be also revealed that the total number of flights to these states dropped (see, e.g., the linear regression in Figure 1 for a reason) from the period of 2000-2007 to the period of 2008-2015, however the number of delayed flights have been increased. As a final note, Delaware (DE) is the least-busy state, again possibly due to its size and population.

4 Some Notes on Implementation

The implementation of this analysis is done in MATLAB and Java platforms since my recent projects have been carried out using these languages. I used MATLAB toolboxes to import and analyze statistics and to plot the results, and borrowed some data structures, such as hash tables, from Java libraries. It should be mentioned that the analysis can be also performed using Python and R.

In the spirit of reproducible results, I provide the codes for this study online, available on the following GitHub link: `https://github.com/amirpasha82` [3].

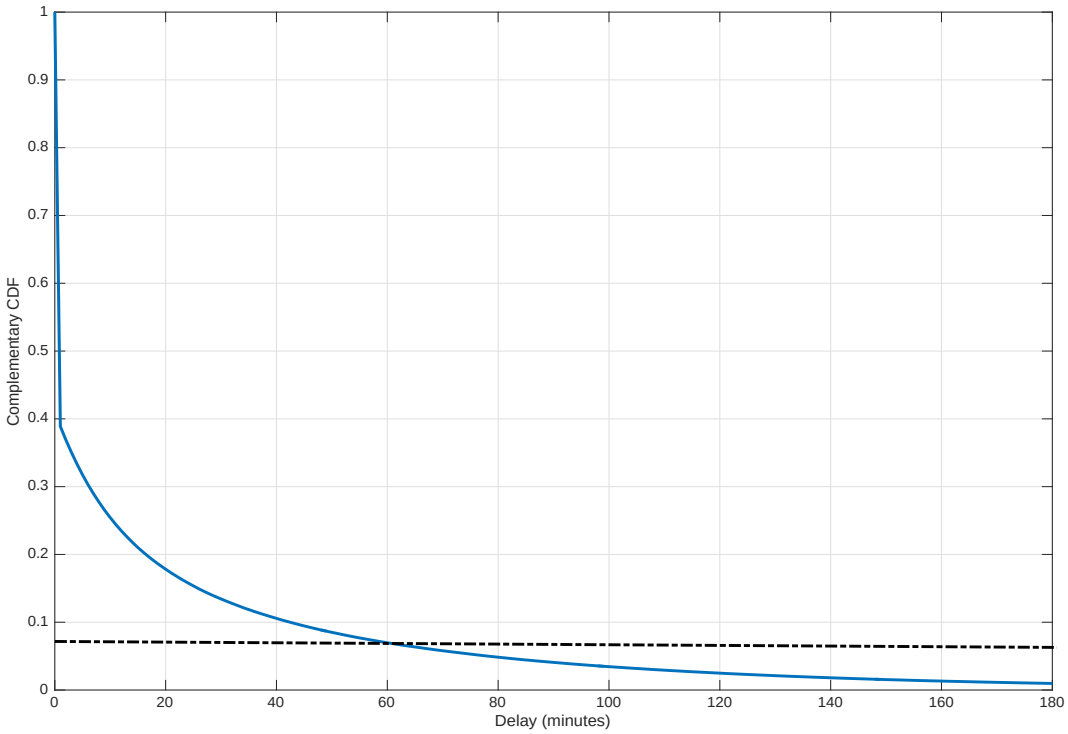

Figure 5: Complementary CDF of flight delays (in minutes) during June 2015 to August 2015. Almost 7% of flights have delays more than 60 minutes in these months.

5 Discussion and Extensions

In this report, we analyzed US flight statistics, from 2000 to 2015, provided by the Bureau of Transportation Statistics. We focused to a greater extent on delay statistics. For example, we showed, e.g., traveling in which months of the year leads to a more lengthy delays, and we showed the performance of airline carriers with respect to delayed flights over number of operated flights.

We note that the provided data is extensive, and can be looked at from many other angles. For example, we can analyze the data to discuss the following questions: Which airports are busier? Which days of the week are safer to travel in order not to encounter delay? What is the relation between the flight delay and flight distance?

So, fly or not to fly? Well, that is the question!

References

[1] R. Goyal and D. Negi, "Impact of global economic crisis on airline industry," *International Journal of Commerce, Business and Management*, vol. 3, no. 2, pp. 2319–2828, 2014.

[2] CNN. (2007) Fliers beware: 2007 set to be worst year for delays, report says. [Online]. Available: http://edition.cnn.com/2007/US/09/25/flight.delays/

[3] A. Shirazinia. (2016) Data science – US flights study. [Online]. Available: https://github.com/amirpasha82

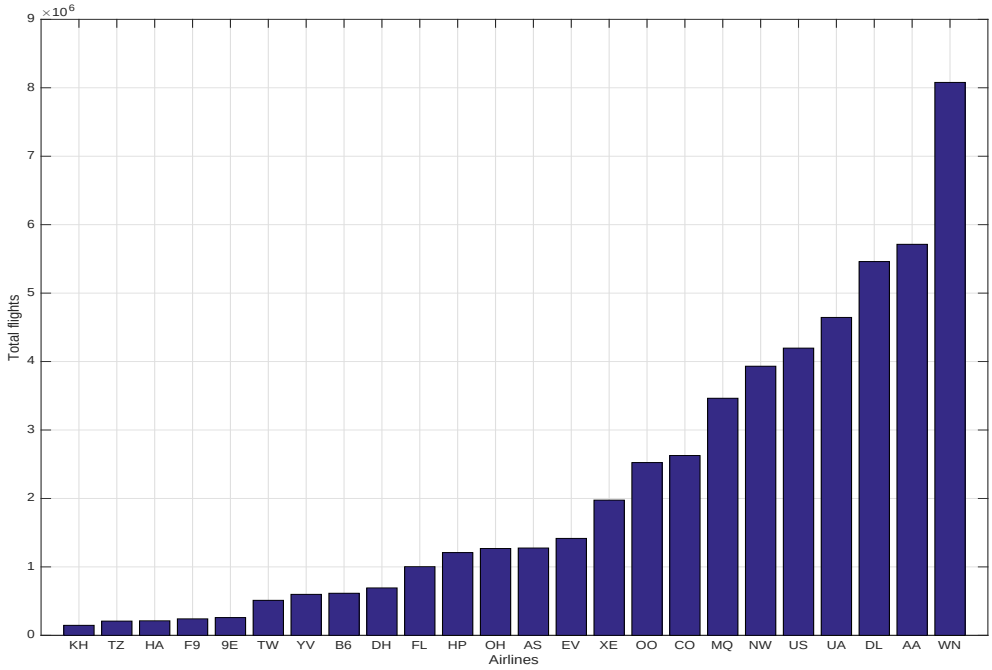

Figure 6: Number of flights by airlines from 2000 to 2007. Southwest Airlines (WN), the world's largest low-cost carrier, has the first rank in carrying out large volume of flights.

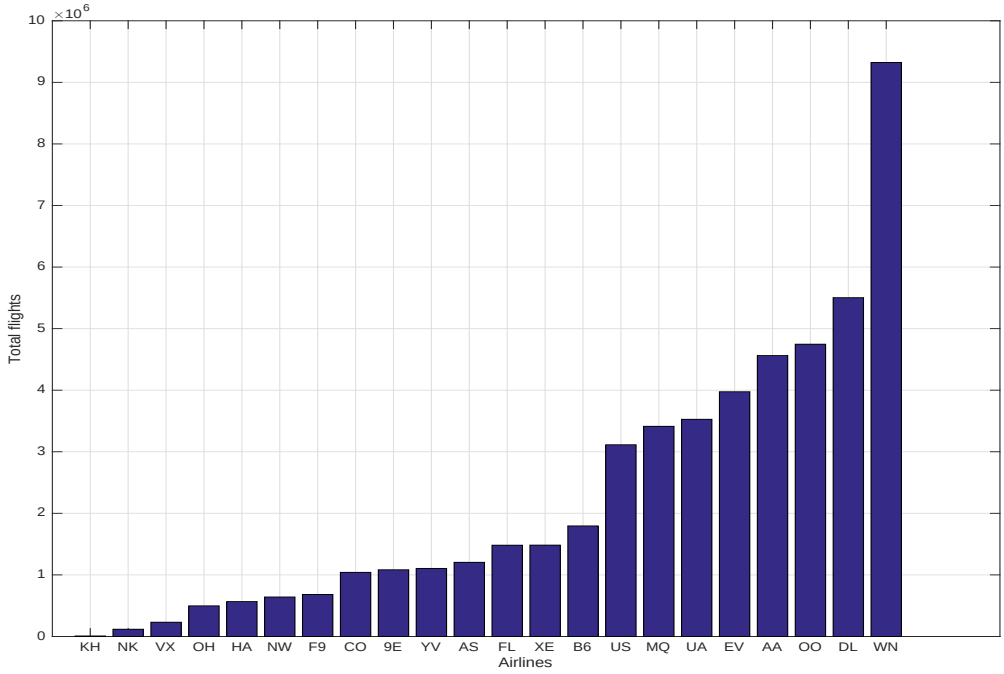

Figure 7: Number of flights by airlines from 2008 to 2015. SkyWest Airlines (OO) has a huge rise in ranking in carrying out flights.

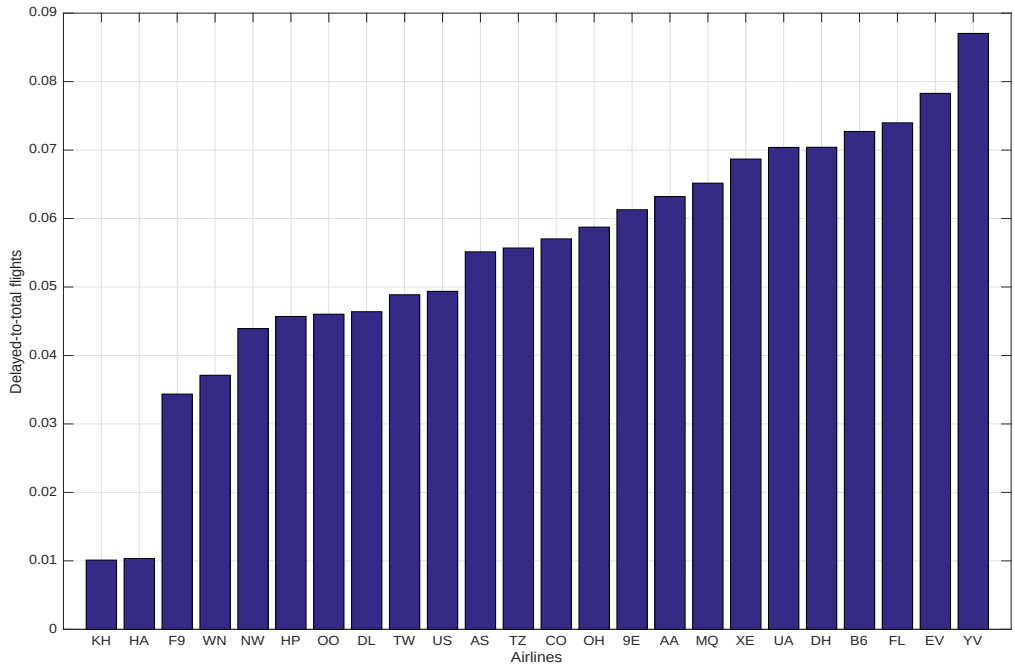

Figure 8: Delayed-to-total flights (operated by each airline) by airlines from 2000 to 2007

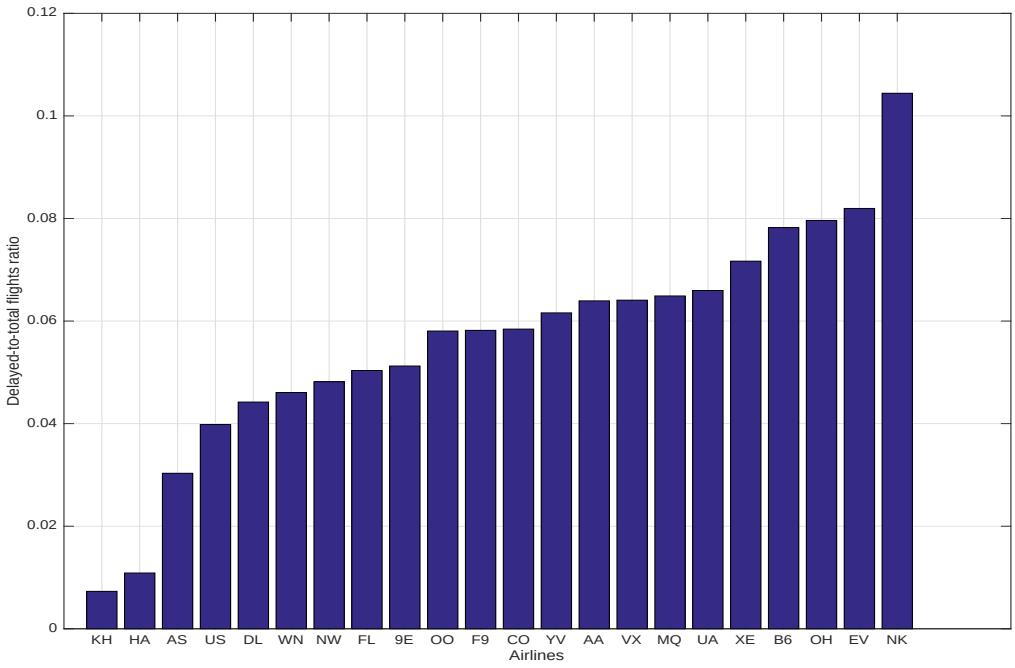

Figure 9: Delayed-to-total flights (operated by each airline) by airlines from 2008 to 2015. Spirit Airlines (NK), a popular ultra-low-cost carrier, gets the lowest rank.

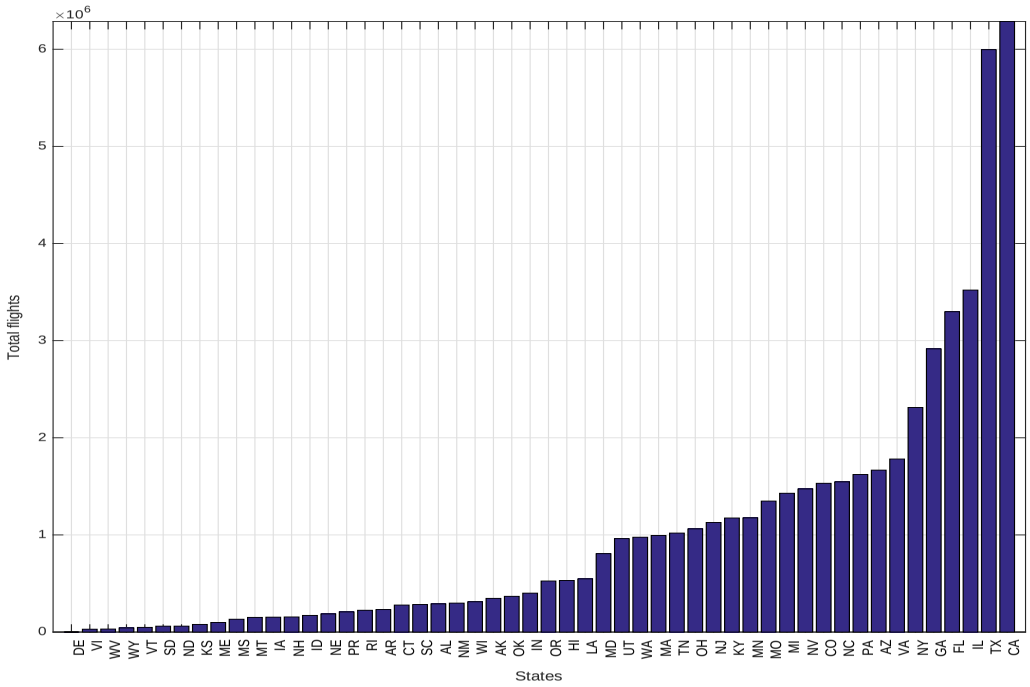

Figure 10: Number of flights vs. destination states from 2000 to 2007

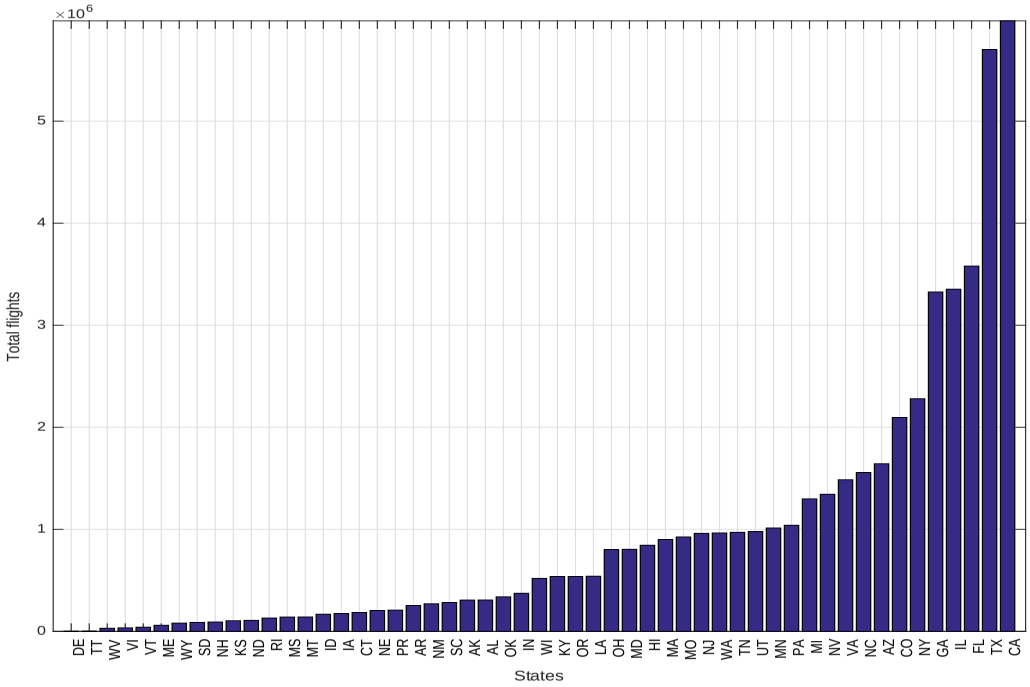

Figure 11: Number of flights vs. destination states from 2008 to 2015

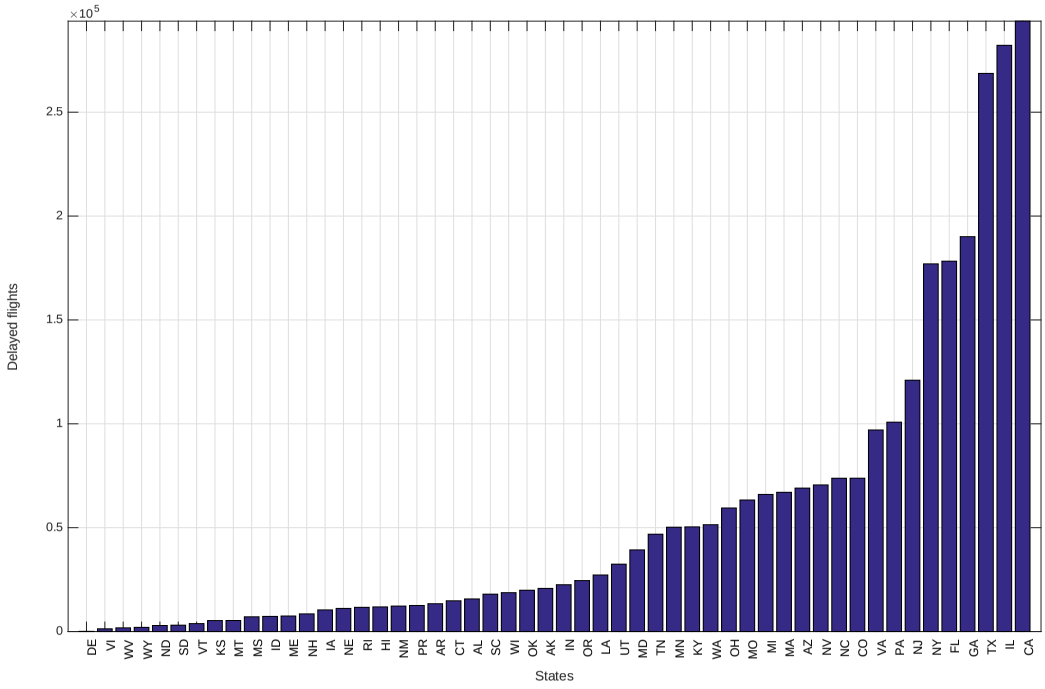

Figure 12: Number of delayed ($\geq$ 60 minutes) flights by destination states from 2000 to 2007

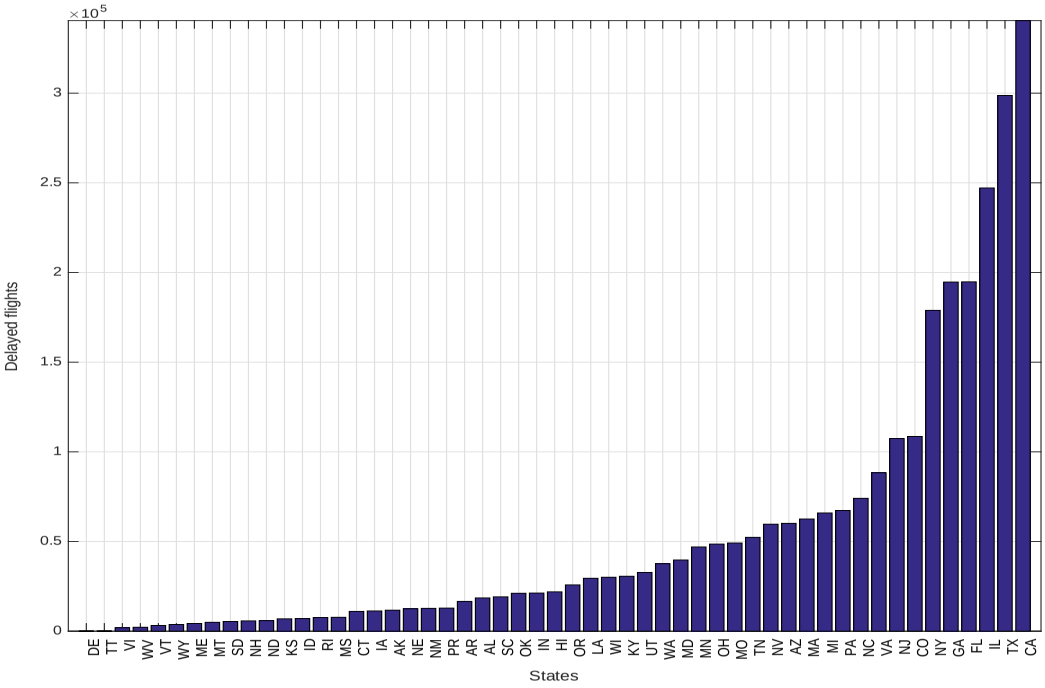

Figure 13: Number of delayed flights ($\geq$ 60 minutes) by destination states from 2008 to 2015