# Flights Analysis

*K Iwasaki*

*May 21, 2017*

## Contents



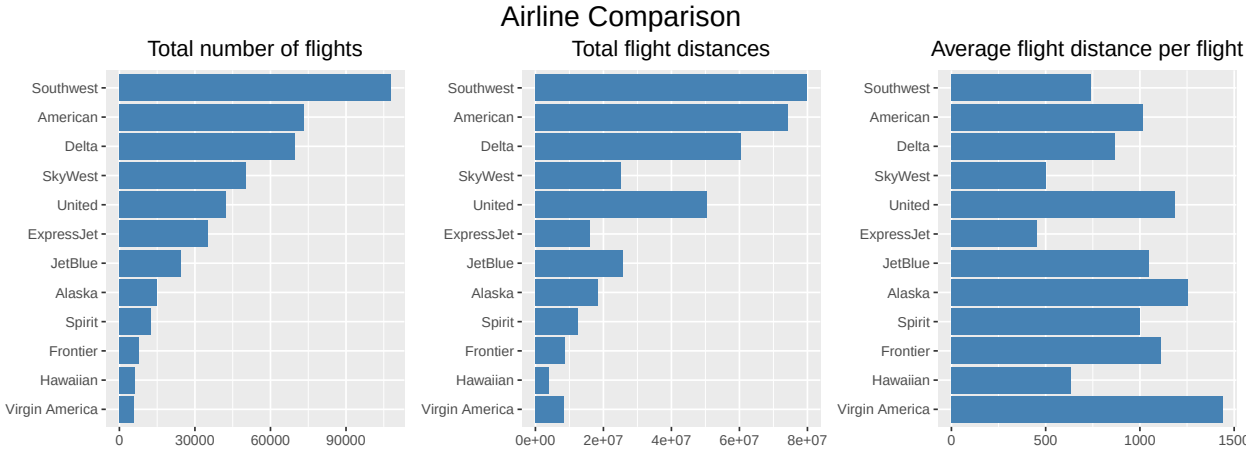
## Introduciton

If you travel frequently, you are likely to have experienced a flight delay. You might have wondered how frequently a flight delays. Airline on-time performance data is public. The U.S. government makes the data available here. I used the data for January 2017 for my investigation. The data contains 450,017 columns and variables including carrier, flight, departure delay, arrival delay, and airport. It's important to note that the data doesn't include international flights. Lastly, R is a stastical computing language I used for this analysis.

My focuses in this investigation are mainly around:

- Understand on-time performance by airlines at high level. This helps to set a stage for detail analysis.
- Investigate deplays in depth. Objective is to uncover useful insights to avoid flight delays for future travel.

## Which airline performs best?

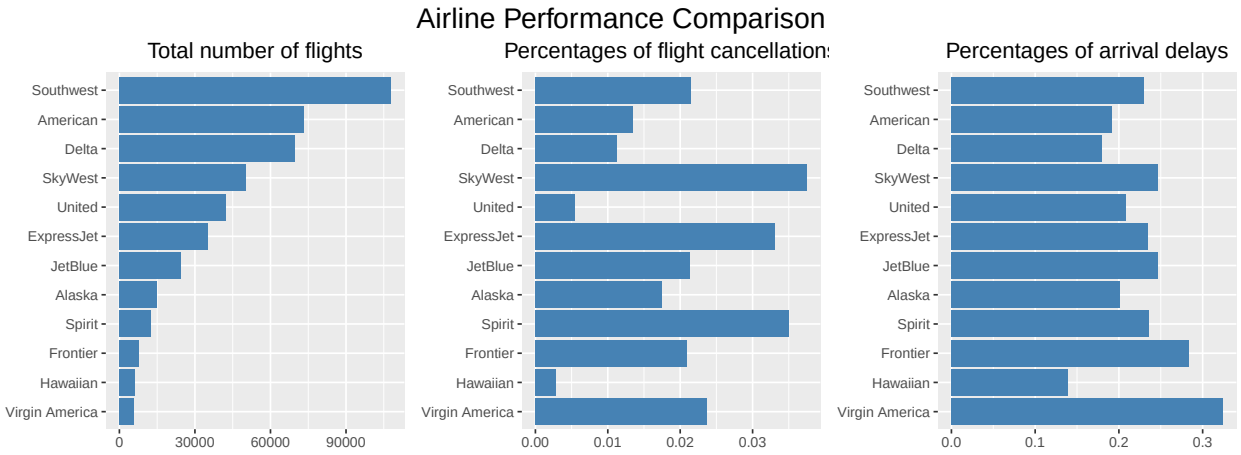


Before investigating on-time performance for airlines, I wanted to understand a big picture: which airline fly more frequently and more distance. It turns out Southwest operated the largest number of flights 107,785 in January 2017. In other words, Soutwest had average 3,477 flights in day. American and Delta followed the Southwest with 73,132 and 69,813 flights respectively for the month. Total flight distances for airlines show similar a ranking to the ranking for the total number of flights. I also added a chart for average flight distance per flight for airlines for reference.

With that high-level big picture in mind, let's dive into the on-time performance for airlines. I focused on two metrics: cancellations and arrival delays. I didn't include departure delay because I don't personally mind depature delay as long as my flight arrive on time.

In January 2017, overall 2% of flights were cancelled. Look at airlines, SkyWest is the worst, cancelling 3.7% of its flights. Sprint and ExpressJet followed Skywest with 3.5% and 3.3% cancellation ratio respectively. Investigating cancellations can be interesting. I might analyze the data to see if certain airports have higher cancellation rate, if at certain time of the day cancellation rate is high, and so forth.

It is no wonder frequent travellars (even non-frequent travellar) encounter flight delays. Unfortunately, flight delay is in fact frequent. 22% of flights arrived at their destination with more than 15 minuetes delay. Most of airlines have 20-30% of their flights delayed for arrival. Given arrival delays is almost ten times more frequent to cancallations, I focus on arrival delays for the rest of my analysis.




# Investigate the delays further

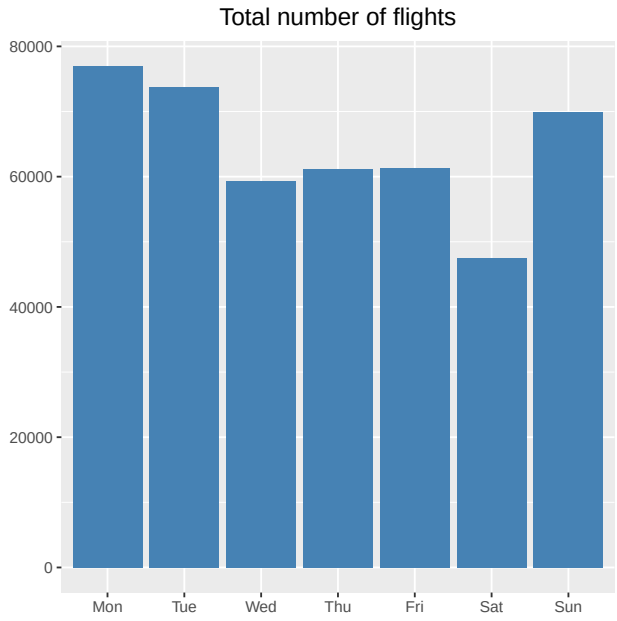
I recently read an article about tips for booking cheapest airline tickets. The article reads Tuesday, Wednesday, and Saturday are cheapest days to fly. Bigger airports(particularly hubs) often have cheaper airfares. These insights stimulates my thinking. Is there any particular days that have less or higher rate of arrival delays? Is there any particular time when arrival delays are more or less frequent? Does flight distance influence on arrival delays? Do bigger airports have more arrival delays or less arrival delays? Let's dive in! From this section, I drop five airlines with fewest number of flights from the analysis to simplify visualizations.

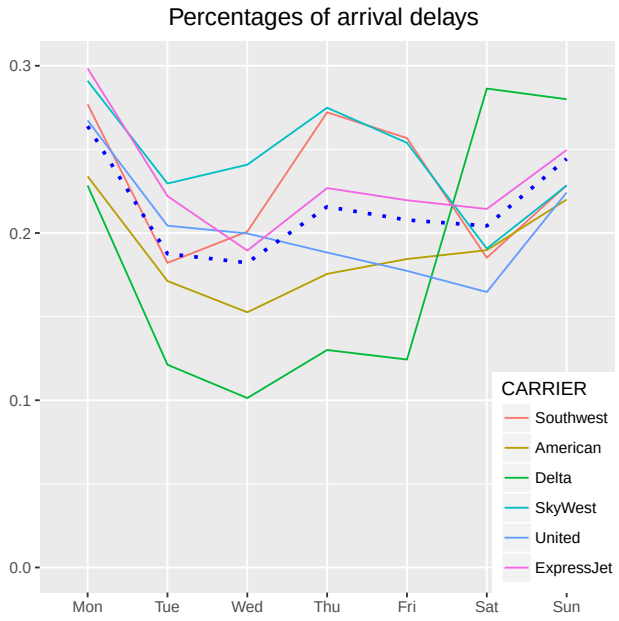
## Is there any particular days have less or higher rate of arrival delays?

Before investigating arrival delays for days of week, I looked at distribution of flights throughout a week. Monday, Tuesday, and Sunday have relatively higher number of flights compared to Wednesday, Thursday, Friday and Saturday. Saturday is clearly the lowest. One interpretation is that people fly out on Sunday, Monday, Tuesday for business then fly back later in the same week.

**- More flights more delays in general but some exceptions.** Look at a chart on the right below which includes blue dotted line for mean value. A pattern for percentages of arrival delays aligns with one for total number of flights throughout a week. Days with more flights such as Sunday and Monday have higher percentages of arrival delays. However what's interesting is that Saturdays has relatively high percentages of arrival delays despite its relatively low total number of flights aday. On Tuesdays, while total number of flights is relatively high, percentages of arrival delays is relatively low.

**- Delta and American perform better than average while Southwest, ExpressJet, SkyWest perform worse.** Arrival delays breakdowns for airlines provides insights. Delta shows best performance compared to its peers, running flights at lower percentages of arrival delays from Monday to Friday. American manages flights at lower than average arrival delays throughout a week. Southwest, Skywest, and ExpressJet performs worse than the average overall.

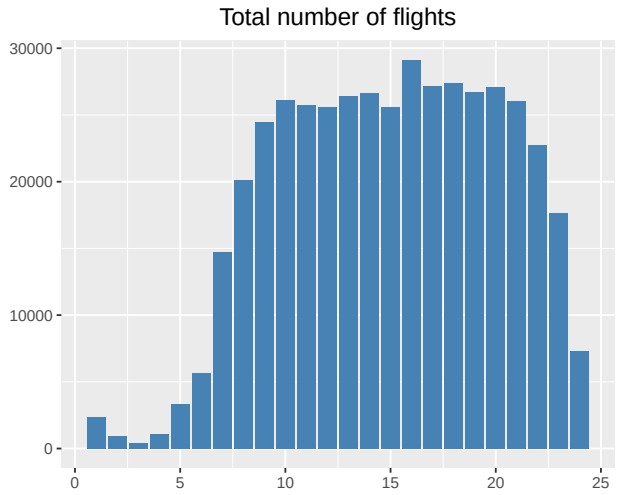


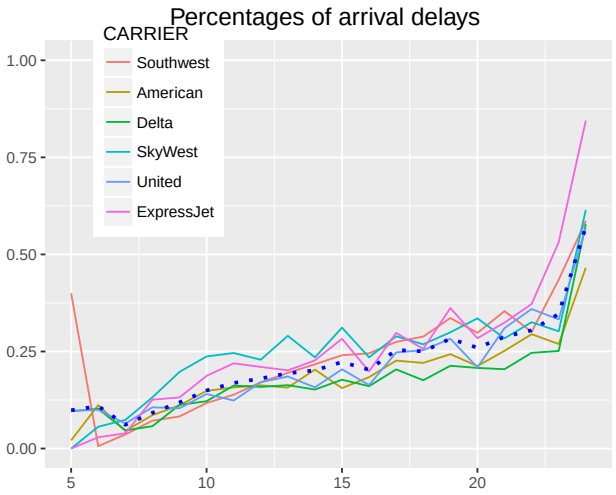




## Is there any particular time when arrival delays are more or less frequent?

Flight schedule is in line with other public transforpotation: there are more flights from 10am to 9pm and less in the early morning and night. Percentages of arrival delays follow clear upward trend from 5am to midnight.

**- The later you fly, the more likely your flight get delayed.** It is interesting to observe that while total number of flights is almost constant from from 10am to 9pm, percentages of arrival delays continue to rise in the same time frame. One interpretation is that one arrival delay affects other flights and the delay get accumulated throughout a day. It is also worth noting that all the airlines show the same trend.



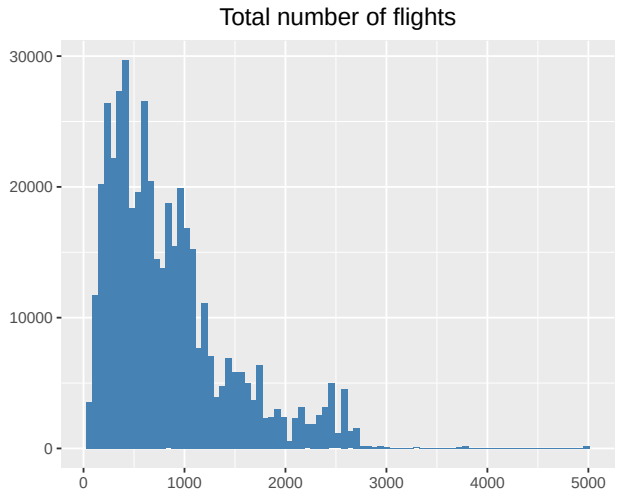



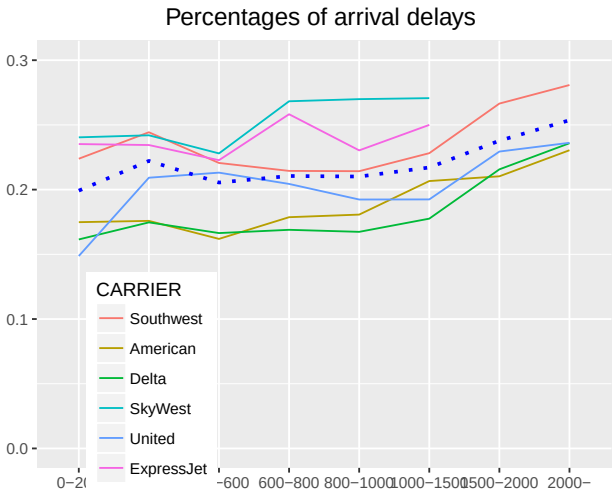

## Does flight distance influence on arrival delays?

Since this dataset includes only dometistic flight, flight distance is clustered below 1000 miles. Meidan vlaue for flight distance is 687 miles and mean is 852 means. Flights with longer distance pulls mean to the right.

**- The more distance the higher arrival delays.** Originally I was thinking, if it flies longer distance and longer duration, it has more chance to catch-up while flying. Thus, longer distance flights should have less arrival delays. It seems this is not a case in real world. Look at percentages of arrival delays. Mean values for percentages of arrival delays are somewhere between 20% and 25% for any flight distances.

**- Delta and American are a winner again.** Airline breakdown shows that better performers are Delta, American and United, managing arrival delays lower than the mean for almost all flight distances. Skywest, Southwest, and ExpressJet struggles to manage their flight on time.



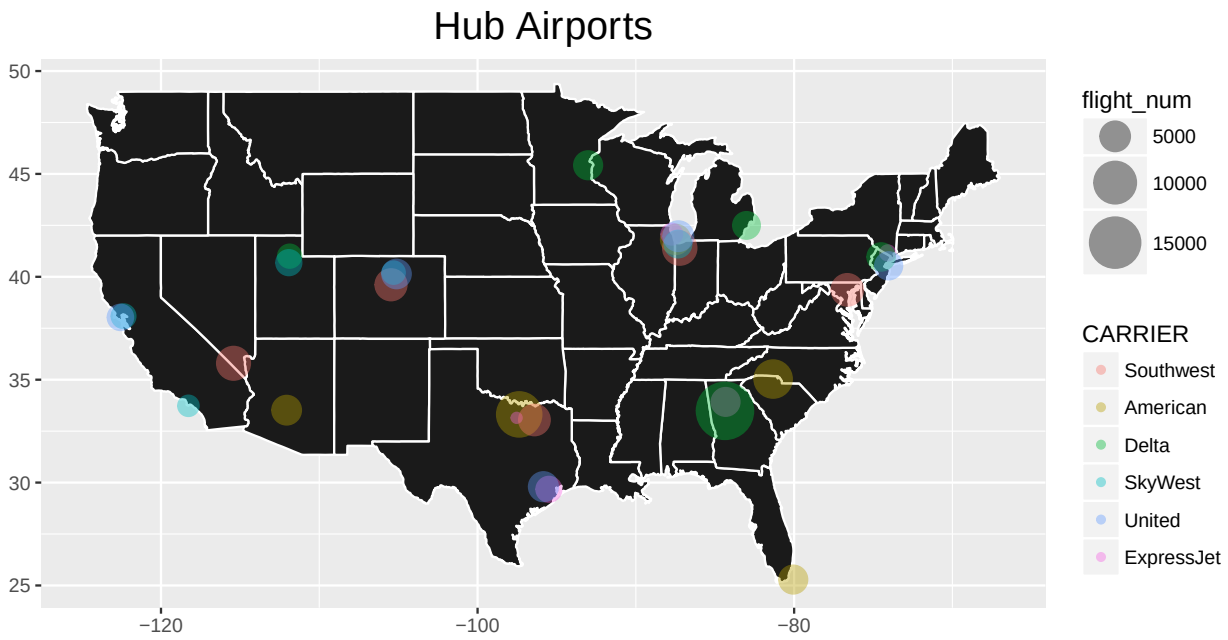




## Do hub airports have more arrival delays or less arrival delays?

In order to tackle this question, first I found hub airports for each airlines by getting total number of arrival flights per ariport for each airline and by filtering top five airports for each airlines. Excluding duplication, I got 18 airpots and visualized them on the map as below.

**Hub airports — visualize where they are and their flight volume**

In the map, each dot reflects number of flights for each airline. For example, if you are insterested in looking at hub airports for Delta, you find some green dots with various sizes. The largest green dots is at Atlanta, Georgia. In other words, Atlanta airport is the largest hub for Delta. You can also observe green dots at Minneapolis, and Detroit.




**Hub airports — uncover patterns for on-time performance**

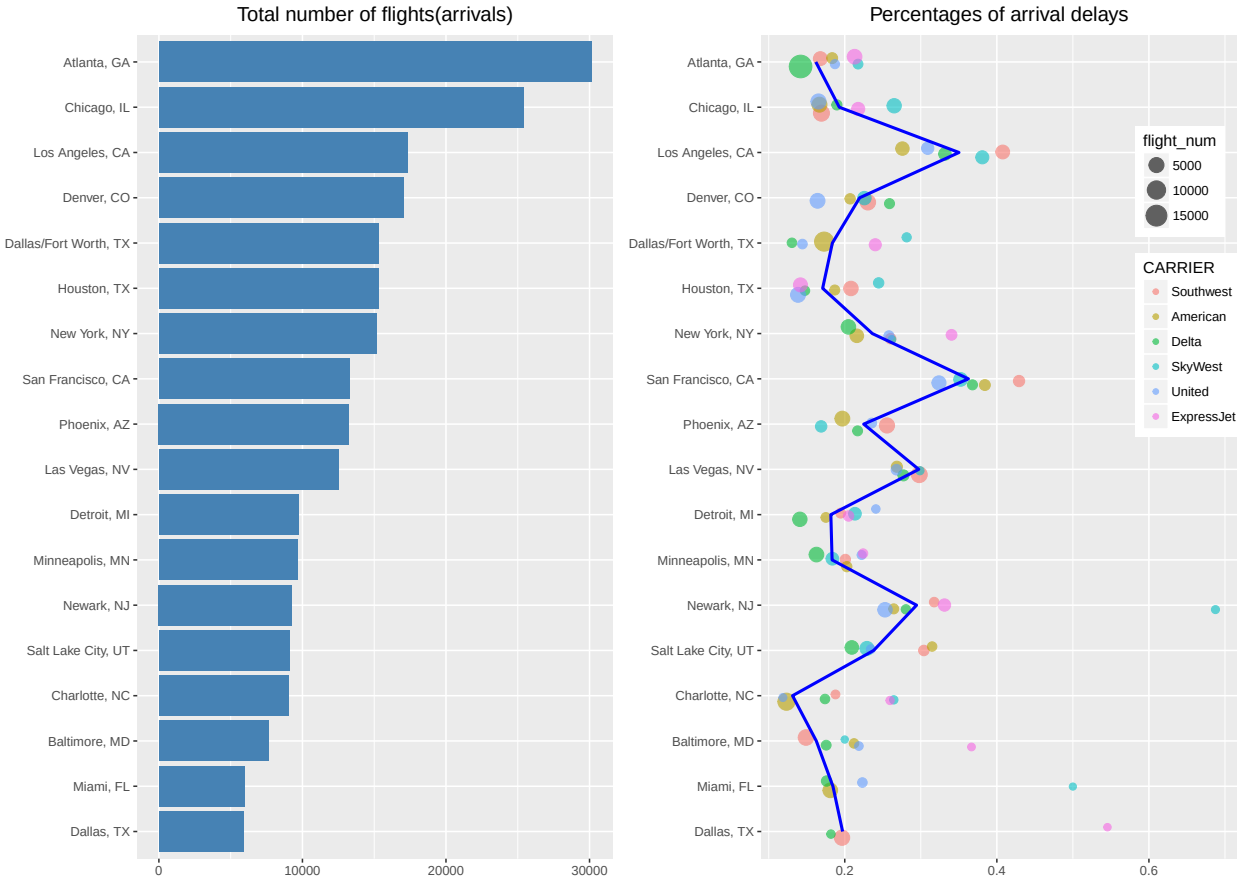
Second step in this analysis for hub airpots is look at on-time performance for each of hub airpors. I ordered airports by the total number of arrival flights in descending order. The chart on the right contains some valuable information to understand on-time performance in detail: the blue line representing a mean value for percentages of arrival delays for each station, and the dots representing a mean value for percentages of arrival delays for each station for each airline with size of a dot corresponding to flight volume. Key takeaways are as followings:

**- On-time performances vary across hub stations.** 16% of flights arrives more than 15 mins at Atalanta and 13% at Charlotte while 35% at Los Angles and 36% at San Francisco. A mergin between the best and worst score is about 20% which is too large to ignore.

**- On-time performances for airlines cluster to some extent at each hub station.** This is notable at Atlanta and Las Vegas where best performer and worst performer sit in a mergin of less than 10%. Most of stations show 10-15% mergin between best and worst performer.

Note: I simplified the analysis by combining airports in a same city into one. A larger airport handles much higher flight volumes than a smaller airport in a same city. Thus, my simplification ends up focusing on a

larger airport in each city.




# Conclusion

**1. Travel in a middle of a week. Avoid Monday and Sunday.** Tuesday and Wednesday are a good day to fly with average 19% and 18% arrival delays of more than 15 minuets while 26% on Monday and 24% on Sunday.

**2. Book a flight at earlier time in a day. Avoid flights arriving destination after 22pm.** From 7am to 8am, only 6% of flights arrive their destination with more than 15 minutes delay. The number goes up toward the end of a day. From 10pm to 11pm, it is as high as 30% for arrival delays.

**3. Flight distance doesn't matter much for arrival delays** In contrast to our intuition that longer distance flight catches up delay and thus doesn't arrive late, flight distance doesn't influence at least percentage of arrival delays.

**4. Choose better hub airports if possible. Avoid Los Angeles and San Francisco.** There are good and bad hub airports. If you make transit at a hub station to make flight cheaper, you should avoid some bad airports. Here is a black list: 35% arrival delays at Los Angeles, 36% at San Francisco, 30% at Las Vegas, and 29% at Newark.