Saad Alam

My Slide Rule

Data Science Intensive

Data Story, Take 1: Delays Out of O’Hare International Airport

1. Choosing a Dataset

The first step in attempting to create a Data Story was choosing a data set which I could utilize to gain some insights. I currently reside in Cincinnati, OH and travel by plane at least three to four times a year. Many of my flights connect to my final destination via O’Hare international and I cannot remember ever having a flight that departed on time out of O’Hare Airport in all my travels. I decided to get a sample of data that would allow me to gain deeper insight into the flight delay patterns at O’Hare Airport, and then see if I could use that information to rework my travel plans accordingly in the future.

To get a glimpse of the delay pattern I extracted a set of data from the Bureau of Transportation Statistics. A few rows of the specific data points I extracted are shown below in Table 1:

Table 1: O'Hare Delayed Flight Data

|  | **DAY\_OF\_WEEK** | **FL\_DATE** | **UNIQUE\_CARRIER** | **ORIGIN** | **ORIGIN\_CITY\_NAME** | **DEST** | **DEST\_CITY\_NAME** | **CRS\_DEP\_TIME** | **DEP\_TIME** | **DEP\_DELAY** | **CANCELLED** | **DISTANCE** | **CARRIER\_DELAY** | **WEATHER\_DELAY** | **NAS\_DELAY** | **SECURITY\_DELAY** | **LATE\_AIRCRAFT\_DELAY** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 2015-07-06 | AA | BOS | Boston, MA | ORD | Chicago, IL | 1930 | 2131 | 121 | 0 | 867 | 0 | 0 | 95 | 0 | 18 |
| **1** | 1 | 2015-07-13 | AA | BOS | Boston, MA | ORD | Chicago, IL | 1930 | 2003 | 33 | 0 | 867 | 0 | 0 | 29 | 0 | 0 |
| **2** | 1 | 2015-07-20 | AA | BOS | Boston, MA | ORD | Chicago, IL | 1930 | 1925 | -5 | 0 | 867 | 0 | 0 | 0 | 0 | 0 |
| **3** | 1 | 2015-07-27 | AA | BOS | Boston, MA | ORD | Chicago, IL | 1930 | 1938 | 8 | 0 | 867 | 0 | 0 | 0 | 0 | 0 |
| **4** | 1 | 2015-07-06 | AA | ORD | Chicago, IL | SEA | Seattle, WA | 1706 | 1703 | -3 | 0 | 1721 | 0 | 0 | 17 | 0 | 0 |

It basically consists of delay time information for flights coming in and out of Chicago. I filtered this dataset down to get only departing flights out of O’Hare International Airport.

1. Initial Findings and Further Questioning

The first thing I did was just to count the number of delayed flights in the month of July and also plot the number of delayed flights per day on a time series plot. There were a total of 13004 delayed flights out of O’Hare in the month of July. Figure 1 below shows the number of delayed departures per day:

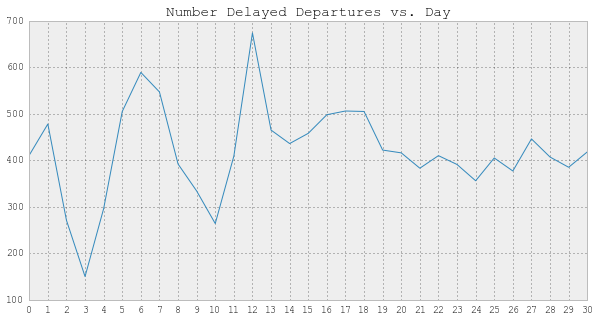


Figure 1: Number of Delayed Flights per Day

The numbers looked pretty high for the number of delayed flights out of one airport each day. But in order to be able to tell how likely it would be to be on one of those delayed flights I had to look at the percentage of flights out of O’hare that were delayed each day of the month. The result is shown in Figure 2 below:

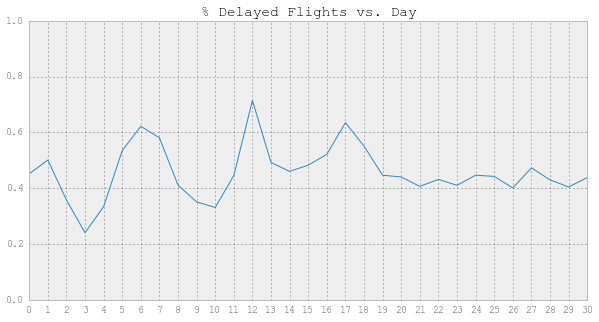


Figure 2: Pct Delayed Flights per Day ORD

I was quite flabbergasted by the result. Looking at the results from just one randomly selected month of data, it was amazing to see that on almost every day of the month, 40% or more of the flights were delayed. My first question looking at the time data was, is there a best time to travel during the month and/or week to minimize the chance of finding myself twiddling my thumbs in an O’Hare terminal gate?

Also, I personally have been flying United Airlines the majority of times that I have been flying through O’Hare. I wanted to look further into the data to see how Untied Airlines performed in terms of delayed flights during this month.

1. Trying to Answer Some Questions

To look further into the question of when the best time(s) are to travel, I looked into the distribution of delays based on day of the week. Figure 3 below shows the number of delayed flights and percentage of delayed flights as a function of day of the week:

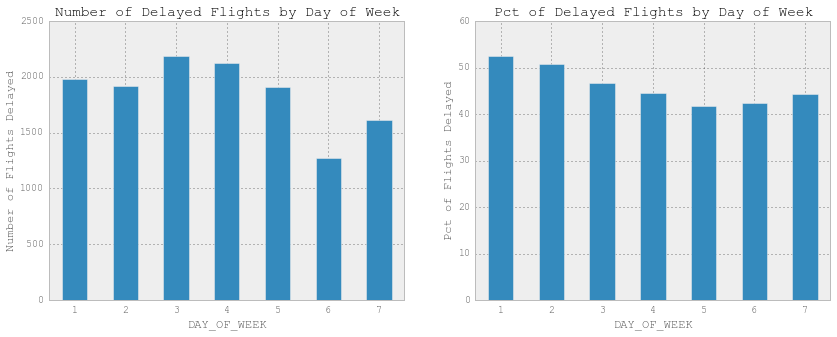


Figure : Delayed Flights vs. Day of Week

From the plots above it seems difficult to definitively say that any single day or 2 days were the best times to travel. However, there seems to be a slightly downward trend in percentage of delayed flights as one gets later in to the latter days of the week. I believe over a longer study of multiple sets of data I would look at this trend using several more weeks’ worth of data.

Next, I looked a little further into the performance of individual airlines. Figure 4 shows the numbers and percentages of delayed flights for each airline that flew out of O’Hare during July 2015:

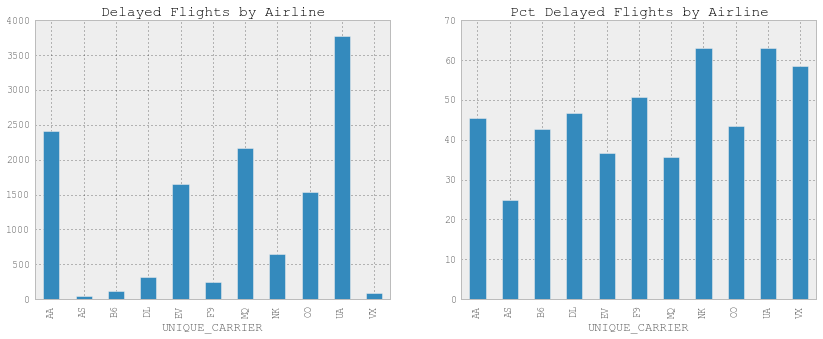


Figure : Delayed Flights by Airline

Sure enough, my experiences were confirmed by the data from July. United Airlines had the greatest number of delayed flights and was tied for the highest percentage of delayed flights during the month of July. Figure 5 below shows the United Airlines delayed flights data for each day of the month:

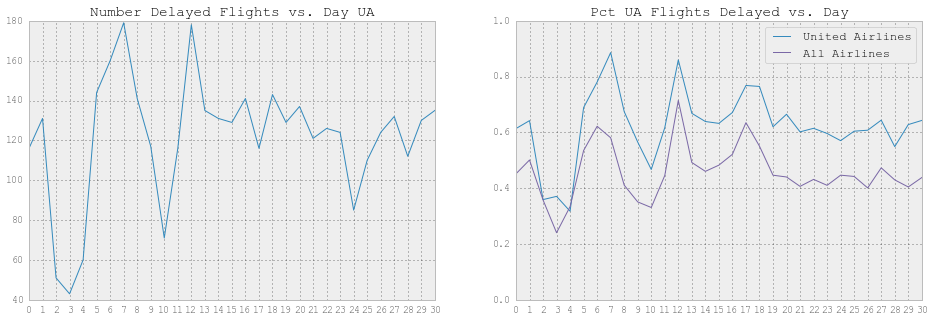


Figure : Delayed Flights per Day United Airlines

United Airlines had a higher percentage of delayed flights than field of airlines every day except for one during July of 2015. Looking at the results above, it certainly gives me reason to be skeptical of United Airlines’ efficiency in their fleet operation. Of course, one month of data is not sufficient to make this ultimate conclusion. However, it provides enough of an indication to further analyze the performance of the airline.

Assuming this result held consistent for several other months’ worth of data, I would have to closely consider my booking of flights via United Airlines through O’Hare. I would also like to perform this analysis for other airports to determine if United Airlines has a particularly difficult time at O’Hare or if this is the overall performance of the airline.

Lastly, I wanted to explore whether there was any sort of correlation between delay time and the flight time. I only took a very cursory jab at this by scatter plotting departure time vs. delay time using the O’Hare July data set. The result is shown in Figure 6 below:

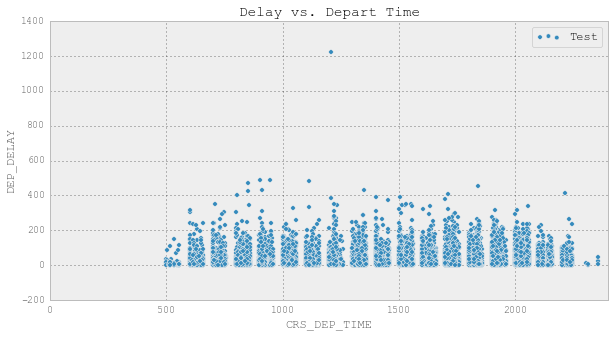


Figure : Delay Length vs. Scheduled Departure

It is difficult to tell if there is any correlation or trend from the scatter plot shown in Figure 6. I need to first look further into the likelihood of a delay as a function of scheduled departure time to see if there is a trend in that regard. This would provide me with a strategy for planning flight times as well.

1. Conclusions