Airline Delay Writeup

Owen Queen

The Process

In this case, we were asked to explore weather delays to flights in Chicago O'Hare International Airport (ORD). This was a very open-ended problem, and I decided to take a simple approach to the data, analyzing and gathering weather and flight data to see which factors may have correlated. ……

If you are interested, I have included all the code I wrote for this project in public Github repository available at this address:

https://github.com/owencqueen/Melton_Scholars_Application

If you go to this page and open the README.md, I have notes about how to run my code and what each file contains.

Gathering Data

I started with the dataset provided in the writeup (transtats.bts.gov). This data was very helpful as it provided a comprehensive collection of flight data from around the United States, including variables such as time of weather delay, date of flight, flights that were cancelled, and other delays. I decided to analyze data for the entire year of 2019. This data is the most recently available, and by having an entire year, it would allow us to look at ORD in all seasons, not just the winter. This data set would be the basis of my analysis for the status of flights.

I was also able to find several other datasets that concerned weather and major events in the Chicago area. The first was local climatological data from NOAA (National Centers for Environmental Information), named "2019_weather_data.csv" in the github repo. This provided variables such as precipitation, snowfall, temperatures, and more weather data in Cook County, IL (where ORD is located). These variables were important in assessing the causation of climate factors on delay times. The second was storm data gathered from the Storm Events Database maintained by NOAA, named "2019_storm_data.csv" in the github repo. This data was interesting because it identified major weather events within Cook County. I used the identification of major weather events to select dates which might be of interest for our analysis.

Data Engineering

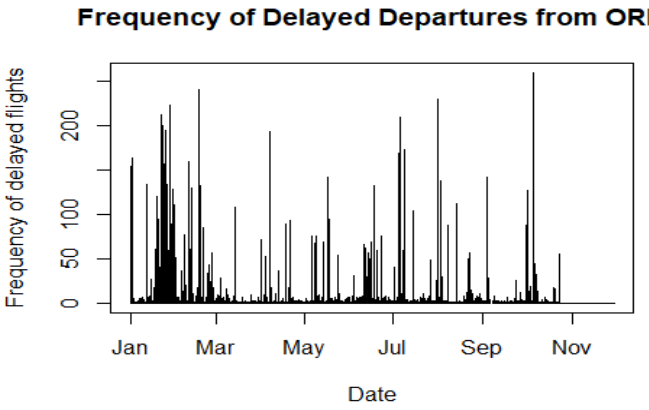
The majority of this project consisted of cleaning the data. These data sets were very well maintained, but to use the sets in general, they required construction of new data frames and modifications of old ones. To do this, I used R's powerful built-in features for data frames. Much of the project consisted of fixing tiny differences in dataframes to have our functions work in general cases.

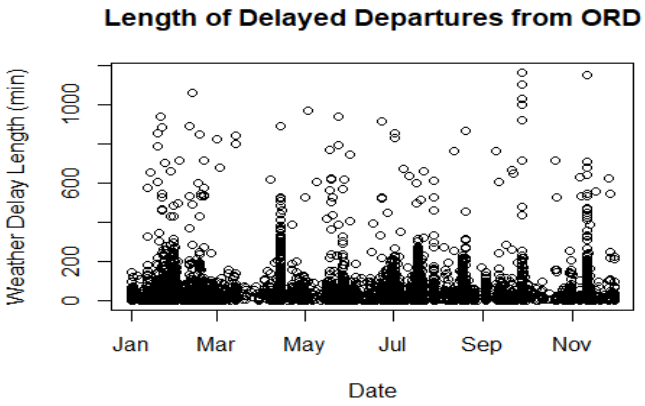
For more information on the programs I wrote to clean the data, see the Github repo link above and find the README.md file for more explanation of how everything works.

Results

Note: The bulk of my work was done in *year_analysis.R*, so you can see my code in the Github repo. I am not going to include any code in this report, only the resulting graphics and statistics.

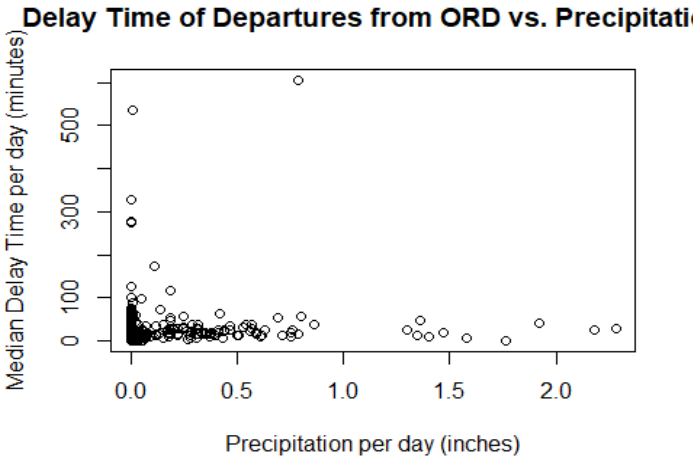
First, I wanted to simply understand when delays happen at ORD. The writeup seemed to suggest that this happens in the winter months, and I wanted to validate this statement.

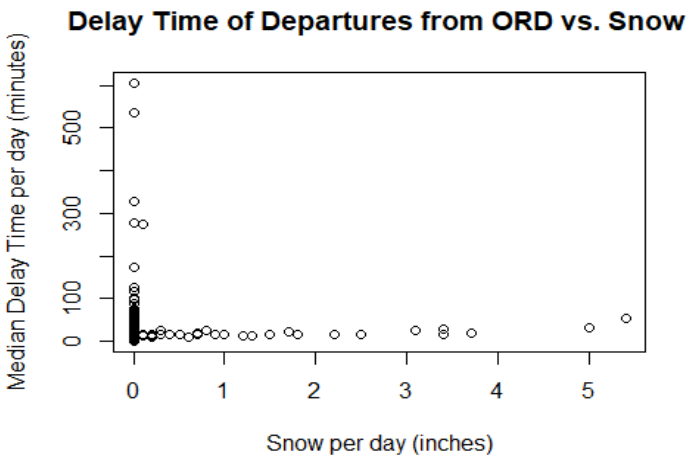




From this data, we see some clusters of more frequent delays around February, but generally, the spread for flight delays is relatively even. Thus, we cannot necessarily just analyze any specific time period - we must look at the entire year to understand the cause and effect of flight delays.

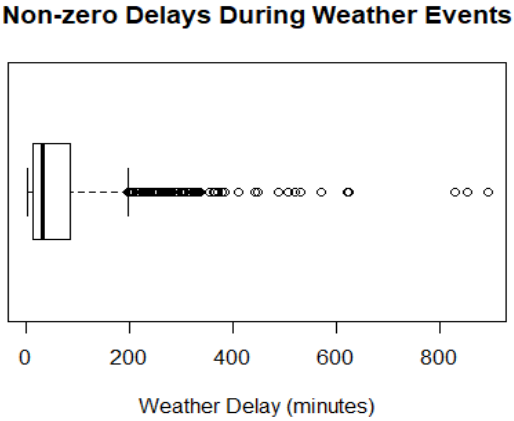
I then wanted to look at how precipitation affected delayed flights, because it seems reasonable that precipitation (snowfall or rainfall) would affect how a flight is delayed.



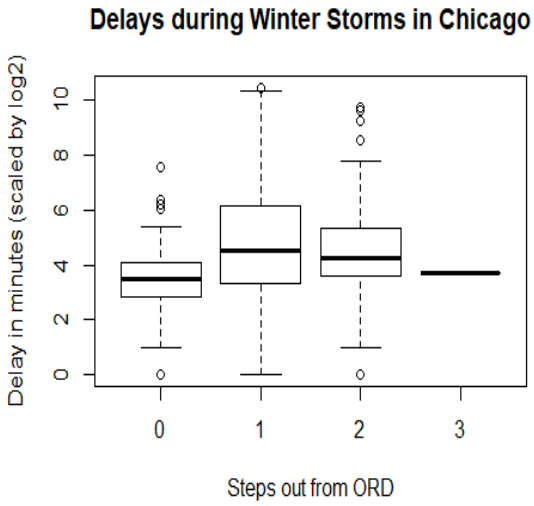


The Pearson's correlation coefficient (*R*) value for a linear regression model of delay time vs. precipitation is 0.0369; for delay time vs. snow, the *R* value is 0.0173. From these values and the plots, we can see that there is no distinct relationship between delay times and precipitation/snow. Therefore, I concluded that maybe instead of looking at just weather data, we should look at our data that identifies the dates that major storms occurred and see if this had a major effect on delays.

Upon analysis of the storm data (*2019_storm_data.csv*), you can quickly see that there are multiple storm descriptions, but not all of these would logically be significant on a flight being delayed or not. For example, one event description is a "Flood", but this may not be worthwhile to consider because a flood on the ground would not affect planes being able to land. Thus, I decided that for the sake of time, I would only choose a few major weather events to analyze. The weather events I chose were "Winter Storm", "Thunderstorm Wind", and "Heavy Rain" because these seemed to be the most available events that I assumed, from personal experience, may affect a flight being delayed or not.

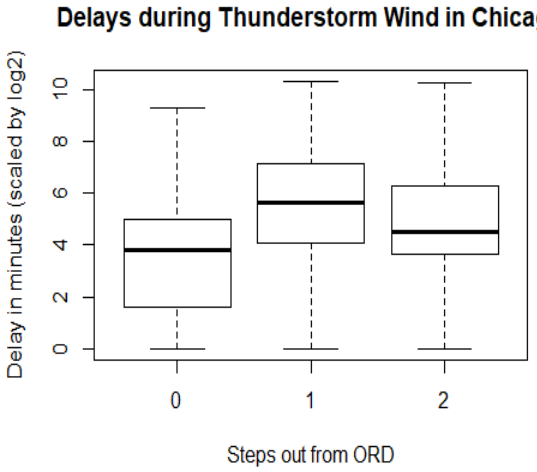


Upon analysis, we can see that of the flights that departed from Chicago in 2019, 2.77% of them were delayed. When looking only at flights that occured on days that Thunderstorms, Thunderstorm Wind, or Heavy Rain occurred, 69.8% of these such flights were delayed. In addition, I included a boxplot of delayed flights on days when these major weather events occurred, and we can clearly see that the presence of these major weather events typically corresponds to long delay times in departing flights. Armed with this knowledge, I decided to see how these delays percolate out from ORD on days when there is a major weather event in the Chicago area. After some preparation, I prepared the following boxplots:



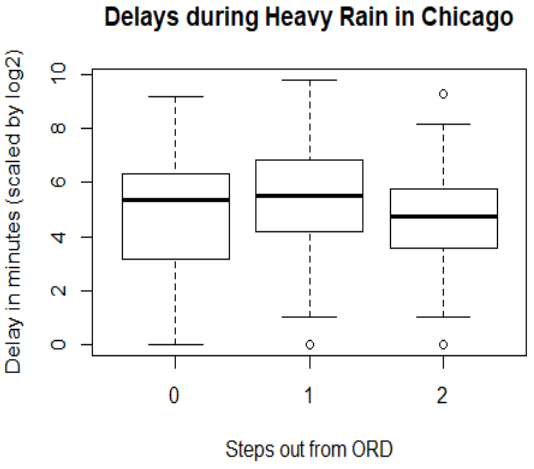
The y-axis of all of these plots (Delay in minutes) is scaled by log_2 because of large outliers that made the graphs difficult to view when plotted normally. The "Steps out from ORD" values correspond to how many steps away a flight is from ORD (built using the breadth-first search in the *five_func.R* file). For example, if the Steps out from ORD value is 1, that means that this is departing flights from aiports that received flights from ORD.

These boxplots show us that typically, delays at secondary airports are usually larger than the delays at ORD. Thus, the impact of these events is more severe for departures from secondary airports than it is at ORD itself. We cannot make any reasonable conclusions that this increase of delay time is solely caused by the weather events in Chicago because there are many factors to consider in this case. For example, this increase in delay could be because typically, flights originating from Chicago may be going to airports in the same geographic region, meaning those secondary airports would also be experiencing weather similar to Chicago. Thus, more analysis is required to make any conclusions about these findings.



Also, we can see that the size of the delay most goes down the further out you get from ORD. In some cases, such as with "Thunderstorm Wind", the delay at secondary airports is even larger than ORD, which may be due to the compounding delays that each airport will experience to account for the delay from ORD. Again, no reasonable conclusions can be made about this before further analysis is done on other factors involved.

I did some analysis on the arrival flights during these storms, but I did not include them in this report to save space. I found that arrivals were typically only delayed coming into ORD; thus, delays on flights coming from secondary airports were generally not affected.



In final, delays of departing flights from ORD typically correspond to major weather events, and these delays percolate down to other airports that are in the line of flights coming from ORD. Departures from secondary airports experienced greater delays than ORD when the Chicago area experience major weather events, and sometimes, these major weather events even corresponded to large delays in tertiary airports. Snow and precipitation amount in day as well as time of year are both not adequate predictors of delay time in ORD. Thus, ORD experiences delays all year, and it experiences these delays in all sorts of precipitation types and amounts.