KJ Morgan

July 11, 2019

Statistics Research/ Data Wrangling

Why Are Flights Delayed?

**Remove Insignificant Variables-**

As I scan through the dataset, I saw that there was an abundance of variables that were not going to give me an accurate approach to my research question. Therefore, I had to use a simple logic approach to each variable. That approach was to simply ask myself do I see this variable impacting airlines. Going down the list, I saw that the delay times for Late Aircaft, Security, and Weather Delay were impactful enough to provide a statistical analysis. I went into RStudio to use a code to remove 26 variables of the starting 30, which leaves me with the 4 variables in the final dataset.

**Rename Variables-**

The variables in the dataset were given a name that came with the dataset. However, for my preference I would like for the variables to be named to where I understand them fully. Therefore, I used a code that renames all the variables in the dataset. I renamed the variables to Weather, Security, and Late Aircraft.

**Remove N/A-**

With respect to having a dataset with more than a million observations. The observations in the dataset were filled with N/A values. This is information that I do not need in the dataset. From this point, the N/A removal code was very beneficial because the actual values in the dataset will be the values being analyzed. The code deleted the N/A observations, which results me with my final dataset.

**Create a Subset of Data-**

The final subset of data comes from the final product of removing the unnecessary data in the dataset. The information that was not needed such as the N/A and extra observations were removed by the code. I found that to actually analyze my data to the standard was that I had to mine the data down to two major airports on different sides of the US. Those two airports are where I was going to analyze data from, so that there will be more of an accurate approach to the delay rate from major airports.