**Michael English**

**Data Wrangling Report and Statistics**

***Data Wrangling***

The Data Wrangling process for the capstone project, “Estimating Passenger Time and Activity at Hartsfield-Jackson International Airport, was a little hectic at first but slowly became more readily clear in mind. One of the main premises of the project is to calculate the timing and load factor of flights so as to form accurate approximations of peak times for population at the airport. There are three primary data sets being used:

1. **Flight Routes (Bureau of Transportation Statistics [BTS]):** This database lists the number of passengers for all routes leaving Atlanta in the month of March, 2019. This provides aggregate amounts and data for flights enplaned in Atlanta. Further, it gives an impression of which routes have the greatest traffic throughout the month which can give a better approximation from the statistical model.
2. **Maximum number of seats (from the Releasable Aircraft Database, Federal Aviation Administration)**: While some routes might be more popular, it doesn’t completely mean that those flights will have certain aircraft to be used. Need to get the *type, tail numbers,* and *number of seats* for each aircraft being used to give overestimate of population based on the flights analyzed in a given time frame.
3. **BTS Database of Flights:**  This will provide most pertinent data as it is giving the precise dates for flights departures, which will give a more accurate description of the population in smaller intervals. The dataset has been refined to focus on the top six cities in terms of traffic.

* From the **BTS Flight Routes** database, have parsed data from the collective data on carriers and enplaning on particular routes from all of America to Atlanta, GA. This database is also restricted to March 2019.
* From the Flight Routes subset, have removed all rows that contain data on passenger listings less than 100; in the timeframe of a month, routes with less than 100 passengers per month are mostly like indicative of freight vehicles/planes and the employees that work them, and are negligible in broader calculations.
* Added a column, *Average Passengers per Day;* this will provide a static context for the original data by determining the average count over an average of 30 days in each month. This, while not completely precise, gives an estimate of how many passengers were flying per day on each route from Atlanta.
* For the **BTS Database of Flights,** there is similar data, have removed redundant columns and leaving data that is pertinent to the particular timeframe of March 2019, developing a chronological consistency between all of the datasets.
* This dataset lists dates and tail numbers for all flights enplaning in Atlanta, GA in March 2019.
* So far, in comparison, the T-100D Dataset shows total passengers for the various *routes* departing from Atlanta, whereas the OnTime Reporting Database gives access to tail numbers that can be used to determine the load capacity of each flight enplaning in Atlanta for the entirety of March 2019.