Michael English

Springboard Data Sciences Certification Program

Data Story

***Introduction***

Air travel consistently becoming a common part of the American experience. With ticket prices dropping lower, it has perpetuated the influx of passengers that are utilizing air travel, and over time has ultimately produced less efficiency in the airport systems due to the need for more personnel to accommodate needs, as well as the usage of space and facilities, increasing the rate of wear and tear. At Hartsfield-Jackson International Airport (HJ\_ATL), aptly named the busiest airport in the world, overcrowding is somewhat of an expectation as their enplaning numbers always seem to increase. In this research, aircraft load capacity and flight data will be used to ascertain populations at airports to determine and address rates of overcrowding.

Exact data could be produced having from direct access to passenger ticket and flight information, but such information is generally described sensitive by the U.S. federal government. However, the Federal Aviation Administration (FAA) has crafted an open aircraft database with content that indicates maximum passenger count list per aircraft, per manufacturer. Using the maximum number of seats on a flight as an overestimate, it will be feasible to create a predictive model that accurately depicts what a population size at the airport could be at given moments in time. As a bonus, the analysis of this data can potentially show its usage in application of scheduling personnel based on the number of people enplaning as they all have to go through security; it can also be used in approximating airport profits, making a basis from the airports cost figures for tickets and the number of passengers (tickets) that are enplaning in a given time frame.

***Data Wrangling***

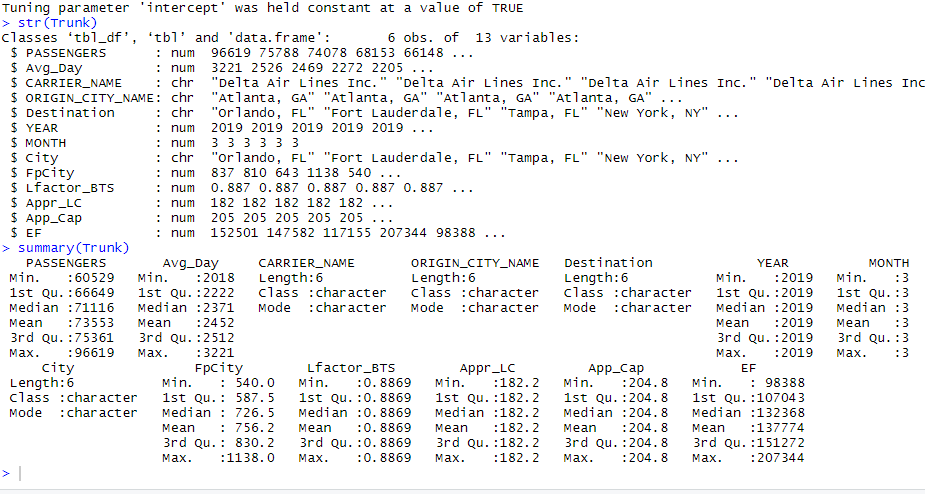


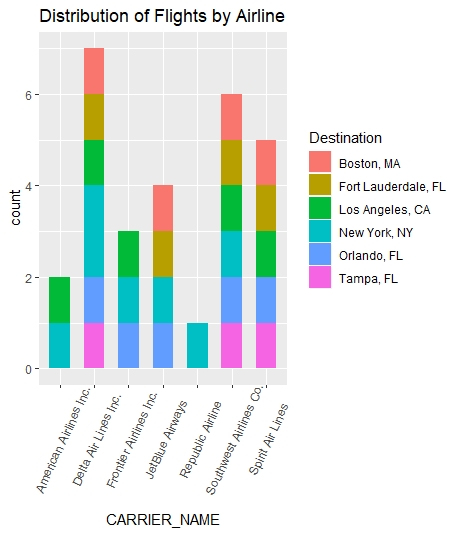
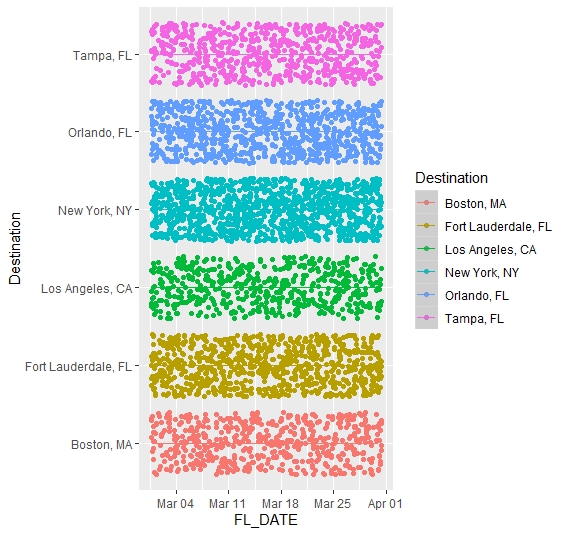
Figure 1. Structure and Summary of the Trunk Dataset.

The data used for the project results from two data sets that were merged into one. The first dataset, *Atlanta Routes*, was comprised of data on flight routes from Atlanta in March 2019, and the aggregate number of passengers for each route. The second database, BTS, was gifted from the Federal Bureau of Transportation Statistics. This database contained a list of ALL commercial flights departing from Atlanta in March 2019, as well as tail numbers which would allow for a more approximate estimate of load factor and thus the population size. Both databases were wrangled, and consolidated into one dataset, Trunk, which shows the aggregate passengers as well as all aircraft and their tail numbers that are for flight Atlanta. A limiting factor of this database and the two it was created from is the refinement of focus – the data isn’t as accurate as if we had data on *daily* *operations* and particular times of flights, which would have made calculating peak times for maximum population much easier.

***Statistical Analysis***

The statistical analysis of the datasets compared multiple data points in a scatter plot and histogram and analyzing the results to draw a conclusion on the population size based on the flights for the top six flight routes from Atlanta: Boston, MA; Fort Lauderdale, FL; Los Angeles, CA; New York, NY; Orlando, FL; and Tampa, FL.

In the figure below, each dot on the scatter plot representing a flight from the BTS database. Based on the density of the plots for each respective, city, we can determine which routes will house the most passengers and thus contribute mostly to the passenger population at the airport:

**

***Figure 2.***

***Figure 3.***

Figure 3 represents the breakdown of those same flights, but by airlines that queue at Hartsfield-Jackson International Airport. As the data shows, the airlines that have the most flights enplaning from Atlanta are Delta Air Lines, and Southwest Airlines, both whom have home hubs at the airport. These two airlines provide a balanced baseline as well as to having the most flights to these major cities and thus being used for analysis.

The next stage is to apply machine learning techniques such as linear regression or clustering in order to organize the data into a mathematical model. The basis of this module is to take the cleaned data and use statistics to analyze the data, and use machine learning to develop a model that approximate data points and other points in the future. We are looking forward to determining what regression models will produce the most accurate predictive results for the dataset.