Executive Summary: Project Proposal

**Description of Dataset and Reason for Interest**

The data set that I will be using for the term project is on time performance, describing arrival delays for non-stop domestic flights by major air carriers for the year of 2008. This dataset was collected by the Bureau of Transportation Statistics (BTS) to specifically track the on-time performance, describing the arrival date for non-stop domestic flights by major air carriers for the year of 2008,which captured the merging and acquiring of certain airlines and carriers. I believe a lot can be learned from this year as there a great amount of detailed data that I might not be able to dive deep into over the span of various years with such a great deal of data. Next steps, after this project, could include taking assumptions from this semester’s data and comparing those findings across various years.

Nonetheless, the data set includes variables for departure and arrival delay, origin and destination airports, flight numbers and actual departure information, et cetera. I am particularly interested in this data set for two reasons: 1) my constant travels to and from Boston year round, as I am originally from Chicago, or have had opportunity to travel abroad and 2) my background in engineering psychology/human factors that has trained me to be curious about human behavior and how it contributes to certain processes/services.

From this data, I anticipate being able to find possible trends in flights being on time or delayed, and if delayed by how much, and whether that interval means there might be another trend emerging. More specifically, the variables month, day of the month, and day of the week have a higher likelihood of being delayed for certain or many airlines (identified by the variable Airline ID). For example, it would be interesting to see if and which airlines are more likely to have flights delayed in the last week of May, on a Friday than on a Wednesday, in the first week of October.

I believe that this data analysis replication across the various airlines will be useful information for algorithms used for Hopper and other flight booking agencies in search for the cheapest flight. Customers might want to know the cost or tradeoff of booking a cheap flight and whether delays for a certain airline is higher based on a set of trends. Initially how this data has been used was through the Bureau of Transportation Statistics to report airline performance and causes of flight delays. On their website, there are published, monthly, air travel consumer reports that are available to the general public. These findings, unfortunately, are not the easiest to read or understand what is meant by the data. In efforts to understand them myself, I believe this project would be a great opportunity to begin understanding so that I can learn how this information could better be presented.

**Possible Identified Data Issues**

Possible issues that I’ve identified is interpreting variables (i.e. total ground time away). For the missing data, I will need to figure out the relevance of those variables and what percentage of missing data there is for certain columns/rows. Regarding interpreting the variables, I will need to do more research on what these terms might mean and most significant variables that are normally attributed to flight delay. Fortunately, on the BTS website, they have provided detailed descriptions of the variables used. This will help me to determine which variables are most important to my analysis and which ones are less so.

**Provided Link to Dataset**

<https://public.enigma.com/datasets/on-time-performance-2008/0e506417-3449-4051-accb-83c743e6e18d>