U.S. Air Travel Analysis

Air travel has become an important part of our society today. It brings people closer together and transports them to places both for business and for pleasure. Air travel is the fastest form of commercial transportation that is available.

For this class project, I performed exploratory data analysis on U.S. air travel data. In particular, I focused on the dataset presented here in Kaggle: <https://www.kaggle.com/usdot/flight-delays>. It is based on flight delay and cancellation data from the Department of Transportation (DOT) Bureau of Transportation Statistics for 2015.

The questions I looked to ask of the data include the following:

* Which airline was the best in terms of different metrics such as delays or cancellations?
* Which airport was the best?
* Is there a relationship between airline and airport for such metrics?
* Is there a time (ex. day of the week) where travelling is the best? I have often heard that Monday mornings are the worst times because of business travelers and that certain days around holidays are worse than others.
* Are there variables that can be represented by certain distributions? For example, do any variables follow the Pareto distribution or 80/20 rule? Are there certain flights or airports that are responsible for most of the delays?

The hypothesis:

* Airlines that are rated the best by travelers also perform the best.
* Airports that are rated the best by travelers also perform the best.
* Is there a relationship between airline and airport for such metrics?
* Mondays and Thursday are the worst days for travelling.
* Certain flights or airports are responsible for most of the delays.

The variables considered for the EDA:

* DAY\_OF\_WEEK - Integer 1 - 7 corresponding to the day of the week. 1 is Monday and 7 is Sunday.
* AIRLINE - Letter code corresponding to the airline for the flight.
* ORIGIN\_AIRPORT - Airport code corresponding to the flight's origin airport.
* DESTINATION\_AIRPORT - Airport code corresponding to the flight's destination airport.
* DEPARTURE\_DELAY - Integer value corresponding to the departure delay for the flight. Computed from SCHEDULED\_DEPARTURE and DEPARTURE\_TIME.
* ARRIVAL\_DELAY - Integer value corresponding to the arrival delay for the flight. Computed from SCHEDULED\_ARRIVAL and ARRIVAL\_TIME.

Outcome of the EDA:

* My initial hypothesis questions were very subjective in nature. I had to come up with an approach to use the data to show objectivity and to define the meaning of “best” or “worst”.
* Histograms for all the variables mentioned were created. The airline histogram indicated that SouthWest Airlines had the most flights by far, but I question that because I have not heard it mentioned as the biggest U.S. airline. The histogram for day shows that Thursday has the most flights, and Saturday has the fewest. The histograms for both departure and arrival delays show that most flights did not suffer long delays.
* Probability Mass Functions (PMFs) for airlines were skewed to left and showed that airline data for the most part had arrival delays that were closer to zero or negative. The PMF for Alaska Airlines which is top ranked according to J.D. Power versus other airlines had lower probability as the arrival delay time increased. PMFs for airports showed similar characteristics where arrival delays were closer to zero or negative. For J.D. Power’s top-rated large airport, Detroit Metropolitan Airport, had lower probability compared to other airports as the arrival delay time increased.
* The Cumulative Distribution Function (CDF) supports why travelers rate Alaska airlines higher as arrival delay of 10 minutes or less is expected over 80% of the time.
* For linear regression, it showed a direct relationship between departure delay and arrival delay. This makes absolute sense since a plane that takes off later will likely arrive delayed as well. I am not sure if variables like these can be considered independent from each other.

For this project, I did not do as much data cleanup as I would have liked. I just noticed that there was a problem with airport data where name reference used numbers instead of letters for a time period. It would have been nice to apply the fix for the data that I read online. Also, the variables in the list were mainly categorical, so I was at a lost on how to use them correctly to answer some of the questions as I do not recall many of the examples in our textbook using categorical variables.

There was some difficulty determining how to answer questions of “best.” It would likely have required thorough analysis of each candidate compared to every other candidate. I believe if I made further calculations from the existing data and used them as new variables, it would help to answer more of the questions. One assumption I had that was incorrect based on the data was that most flights were delayed. The data showed that flights were generally on time with some even arriving earlier. It makes me wonder if airlines have padded the time for their flights with additional time to account for variables beyond just the flight time.

I learned a lot from this course, but I feel I could use more practice in applying the concepts that were taught. I want to improve the analysis that I can perform with a data set. I would also like to be more familiar with the distributions and metrics for describing relationships.