Descriptive Analysis

As shown by the following boxplot of Arrival Delays grouped by each Airline, there is a rather large presence of outliers present within our data that is consistent all throughout. These outliers will skew any average we collect since the values are so varied, so we opt to choose the median arrival delay instead.

A graph with black lines and numbers

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

Mean arrival delay vs Median arrival delay

As shown, the difference between the two is pretty significant, highlighting the importance of using the boxplot to visualize data of interest. In order to further understand how our variables relate to one another, we construct a heatmap to easily see the correlation between each one and ARR\_DELAY.

A screenshot of a computer

AI-generated content may be incorrect.

As shown by the following highlighted portion, DEP\_DELAY is the only variable present in our data that has a high correlation with ARR\_DELAY. The second highest correlation would be CARRIER\_DELAY at .57, which is not very correlated. Because of this information, we choose DEP\_DELAY to be our predictor variable for our Linear Regression Model in the following section. After running an OLS on 4 of the biggest airlines (American, United, Southwest, Delta) using the departure delays as our predictor variable, this is the result given:

A screenshot of a computer

AI-generated content may be incorrect.

This shows us that every 1 unit (minute in this case) increase in DEP\_DELAY will result in a .9155 unit (minute) increase in ARR\_DELAY, ceteris paribus. The P Value for this coefficient is .000, making it highly significant within our model. This regression also has an R^2 of .917, which means that 91.7% of the variance in ARR\_DELAY can be explained by DEP\_DELAY, making it a very strong predictor. This all makes sense in practice as well, since if a plane departs later than supposed to, it is going to arrive that same number of minutes late at the next airport. This is the data plotted on a scatterplot with the regression line:

A graph with a red line

AI-generated content may be incorrect.

Chat-GPT(4.0).(2025/04/06). “how do i create a basic scatterplot with the regression line included in python”

Generated using OpenAI Chat-GPT.