**Airline Delays**

Descriptive statistics is useful to describe the data set by calculating central tendency. The data set “Flight\_Delays\_2018” provides insights on why flights were delayed in 2018. After reviewing the column information from Kaggle (Mu, 2019), The scope of the analysis is based on the date, which is broken down by month, carrier name, carrier delay, taxi in, taxi out, weather delay and late air craft delay .

**Step 1: Descriptive Analytics**

First, is to perform descriptive analysis on the dependent variable, ARR\_DELAY, to see what is happening.



A screen shot of a computer

AI-generated content may be incorrect. we can see that the average delay time is 63.60 minutes and most of the delays vary between 24 minutes to 75 minutes. The max states a 2,692-minute delay which might indicate an outlier or outliers.

Next, we perform descriptive analytics between an independent variable and the dependent variable to see if there might be some relationship, which then we can perform predictive analytics.

A screenshot of a computer program

AI-generated content may be incorrect.

**Month:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Carrier name:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Taxi in:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Taxi out:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Carrier delay:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Weather delay:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Late aircraft delay:**

A screenshot of a computer screen

AI-generated content may be incorrect.

Based on the information, we can see that there is no evidence of difference between the different months and the average arrival delay time and the different carriers and the average arrival delay time. The rest of the independent variables such as taxi in, taxi out, carrier delay, weather delay and late aircraft delay show that if the delay increases so does the average arrival delay.

Visualization tools help us understand the data and help us choose which independent variable can be used to predict arrival delay.

A computer code with text

AI-generated content may be incorrect.

**Box plots:**

A graph with lines and dots

AI-generated content may be incorrect. A graph with black lines and blue dots

AI-generated content may be incorrect.

Here, we see that there is not much deference between months or airlines and might not show if there is a relationship with the dependent variable ‘ARR\_DELAY’.

**Scatter plots:**

A graph of blue dots

AI-generated content may be incorrect. A blue dotted graph with white text

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect. A graph of a plane crash

AI-generated content may be incorrect.

Based on the scatter plots, ‘CARRIER\_DELAY’ shows stronger positive correlation than the other independent variables and will be used to perform predictive analysis to forecast arrival delays.

**Step 2: Predictive Statistics**

Now that we have chosen our independent variable, we can perform an OLS regression to find the line that best describes the relationship between both the dependent and independent variable. OLS will allow us to find the equation .

A computer screen shot of a program code

AI-generated content may be incorrect.

This returns the OLS regression results:

A screenshot of a computer

AI-generated content may be incorrect.

We see that from the ***coef*** column we can derive the equation, which is

. From this equation we can predict how long an arrival delay is if there is a carrier delay. We can state that holding everything else constant, for every minute there is a carrier delay, then there is an arrival delay by 0.83 minutes.

The **P>|t|**, lets us know if there is statistical significance between the independent variable and the dependent variable. The result is 0.000 which is below the significance level of 0.05 and we can infer that an arrival delay can be calculated by a carrier delay. **Prob (F-statistic)** also lets us know that our whole model is also significant. Lastly, **R-squared** lets us know that 32.2 percent of the dependent variable can be explained by the independent variable, Carrier delay. Therefore, 67.8 percent of arrival delays is explained by other variables.

Visualization of our OLS model:

A graph of a graph with dots

AI-generated content may be incorrect.

We can conclude that in 2018 the average arrival delay time was 63 to 64 minutes. From the descriptive analysis, we inferred that Carrier delay had a better relationship with Arrival delay and utilized visualization to help with the independent variable selection. From there, we used predictive statistics to see if in fact, there was a statistically significant relationship with the dependent variable and came up with a forecast equation. This can be viewed with the OLS model and the visualization.