Wendy Herrera

April 9, 2025

Flight Delays Analysis

For this code assignment, I was tasked to get the descriptive and predictive analytics on a data set about flight delays. I obtained important measures of data to get the descriptive analytics which then allowed me to create a predictive analytics regression model.

I began by getting the descriptive statistics of the data set column ARR\_DELAYS which contain the time of the flight arrival delays to better understand the data. I also obtained the statistics for flight delays based on airlines by grouping with the OP\_CARRIER\_NAME column and the statistics for destination airport by grouping with the DEST column. By reducing the data for these statistics, I was able to discover there are significant differences between the statistics of the entire data frame and the statistics of the data frame focused on key variables in the data.

A black screen with white text

AI-generated content may be incorrect.

This is the aggregated data of the flight delays of the original data set.

A screenshot of a computer screen

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

AI-generated content may be incorrect.

This is the descriptive data obtained from aggregating flight delays based on airlines & airports.

These results helped me reduce the scope of the data to focus on data in which flight delays were longer. I believed that by focusing on the top three airports and airlines that experienced longer delays, it would be easier to find better predictor variables for the regression model. The destination airports I chose were ECP, SUN, and SHD and the airlines were Frontier Airlines Inc., JetBlue Airways, and Spirit Air Lines. I used these airports and airlines to create two data sets and create correlation heatmaps. Based on these heatmaps, the variables with the strongest correlation to arrival delay are departure delay and carrier delay.

A colorful squares with white text

AI-generated content may be incorrect.

This heatmap contains the correlations between variables focused on airports of arrival.

A colorful squares with black and white text

AI-generated content may be incorrect.

This heatmap contains the correlations between variables focused on travel airline.

I continued my predictive analysis by creating a linear regression model and an ordinary least squares regression model. I created the regression model using the arrival delays as the dependent variable and testing the departure delays and the carrier delays as the independent variable. Both regressions had a p-value of 0, suggesting that there is a strongly significant relationship between flight arrival delays and departure delays and carrier delays. However, the scatter plot of the linear regression shows there are several outliers, and this relationship may not be exactly linear. Since departure delays had the highest correlation I chose it as final x variable.

A screenshot of a computer screen

AI-generated content may be incorrect.

This is the data obtained from the OLS regression with departure delays as the x variable.

A blue and black graph

AI-generated content may be incorrect.

This is the scatter plot of the predictive regression model with departure delays as the x variable.

In conclusion, I found that flight delay times could best be predicted with the equation y = 0.925x + 10.036. After about 10.036 minutes of flight delays, every minute of delay in departure causes the delay of arrival to increase by 0.925 minutes. Although this model is not perfect, it proves to have statistical significance and provides the most accurate results.