**CODE 8 - Airline Delays Analysis Report**

***Descriptive Analytics:***

*Part 1: Formulate a question*

My analysis is based off the question, “which airline has the most arrival delays and which of their independent variables have a linear relationship with arrival delays that could be used to predict future arrival delays?”

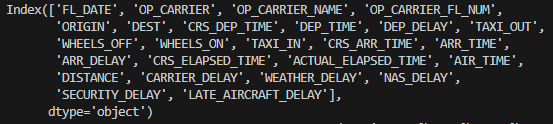
*Part 2: Identify useful columns & filter data*

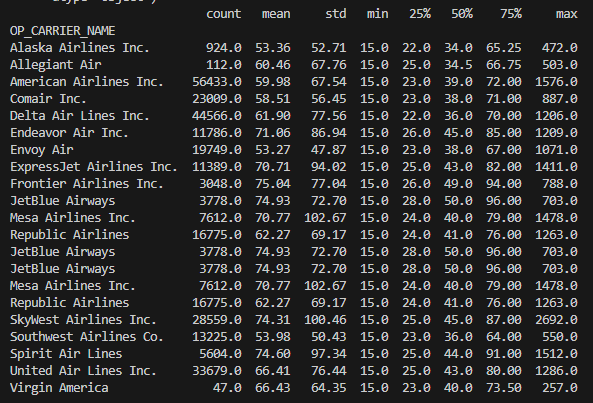
I began my analysis by importing applications that I will use, opening the airline delay’s file, and viewing the names of the columns in the dataset to identify which may possibly be useful. Furthermore, I chose to receive descriptive statistics on arrival delays based on the carrier’s name (OP\_CARRIER\_NAME) to find which airline had the most arrival delays.

The code and results are as follows:

A computer screen shot of a black screen

AI-generated content may be incorrect.





After viewing the statistics, I concluded that the airlines with the most arrival delays are American Airlines Inc. (AA), Delta Airlines Inc. (DL), and United Airlines Inc. (UA) based on having the highest “count” values. Due to this, I filtered my dataset to only view the values of those three airlines using the following code:

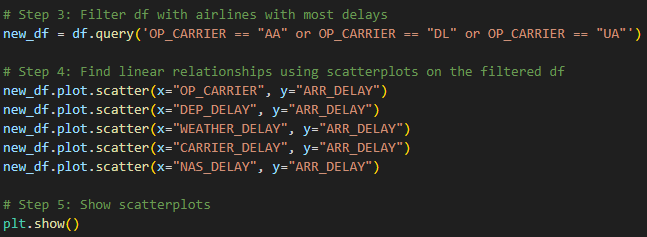
A screen shot of a computer program

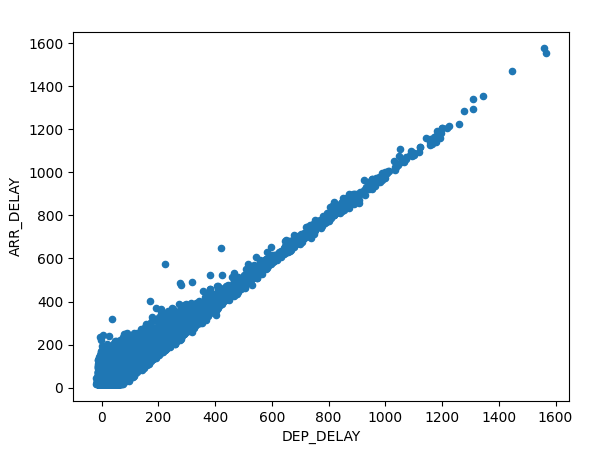
AI-generated content may be incorrect.

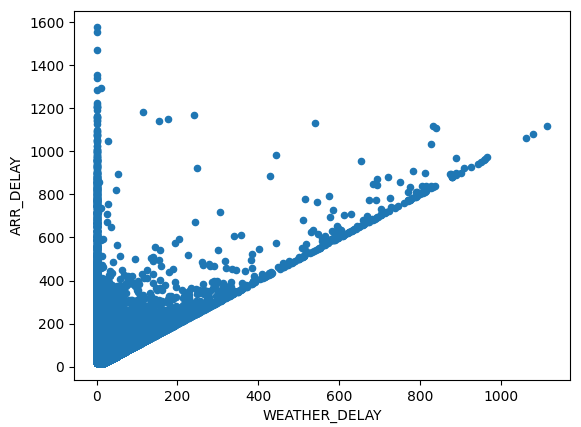
*Part 3: Create scatterplots*

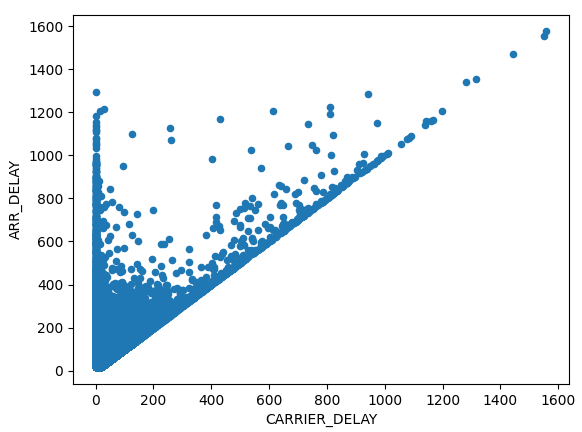
Next, I created scatterplots using the filtered data frame to identify any linear relationships between arrival delays and the other columns. By doing this, I found that the strongest correlations were formed using the following columns: DEP\_DELAY, WEATHER\_DELAY, CARRIER\_DELAY, and NAS\_DELAY. The reason for this is because once graphed, they formed scatterplots that showed straight lines demonstrating that an increase in these variables resulted in an increase in the number of arrival delays. However, DEP\_DELAY and CARRIER\_DELAY have the strongest correlation to the dependent variable as they form clearer lines and increase the quickest.

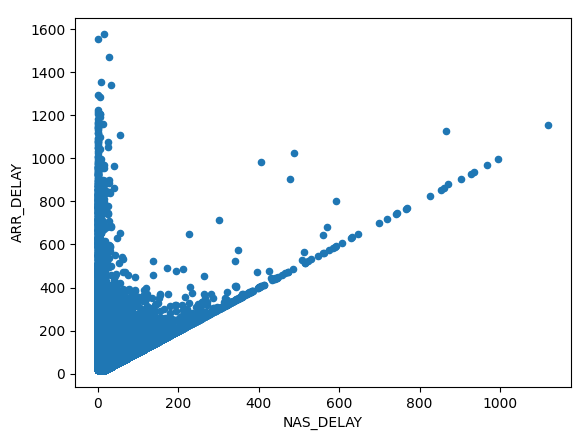
The code and results are as follows:











***Predictive Analytics:***

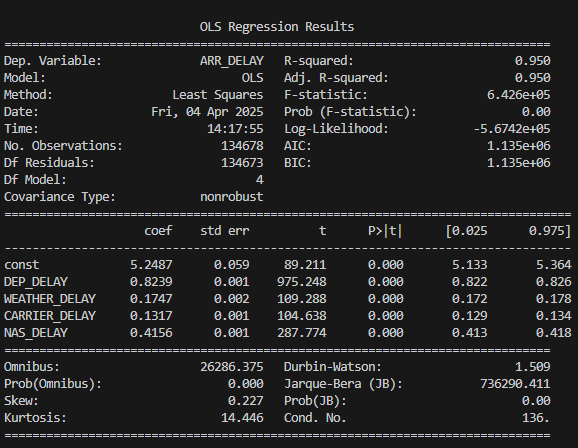
*Part 1: Run regression model*

For the predictive analytics portion of my analysis, I decided to create a regression model in order to validate if my independent variables were statistically significant to arrival delays. To begin, I specified the dependent variable (y) and independent variables (x) I would like to utilize from the filtered data frame. Next, I created an OLS using the x and y values then saved it to the variable “model”. Lastly, I printed the summary of the “model” variable to generate the regression model.

The code and results are as follows:

A screen shot of a computer

AI-generated content may be incorrect.



*Part 2: Analyzing findings*

By analyzing the regression model, there are several findings that are worth noting. First, the p-values (P>|t|) of each independent variable are less than .05; therefore, the predictor variables are indeed statistically significant when determining arrival delays. Moreover, both types of R-squared are 0.950 meaning this model is a very good choice when identifying future arrival delays. Last but certainly not least, to create the predictive linear regression equation, I used each “coef” value and put them in the proper format which resulted in:

**Arrival Delays = 5.2487 + 0.8239(DEP\_DELAY) + 0.1747(WEATHER\_DELAY) + 0.1317(CARRIER\_DELAY) + 0.4156(NAS\_DELAY)**

Equation elaborated:

* For every one-minute departure delay, the arrival will be delayed by 0.8239 minutes.
* For every one-minute delay due to the weather, the arrival will be delayed by 0.1747 minutes.
* For every one-minute delay caused by the carrier, the arrival will be delayed by 0.1317 minutes.
* For every one-minute delay caused by the National Airspace System (NAS), the arrival will be delayed by 0.4156 minutes.

Therefore, according to these results, departure and NAS delays have the most impact on arrival time in this predictive model.