Pinal Mistry, Giancarlo Giudice, Evan Duffy

Carleton Bootcamp

Project 1

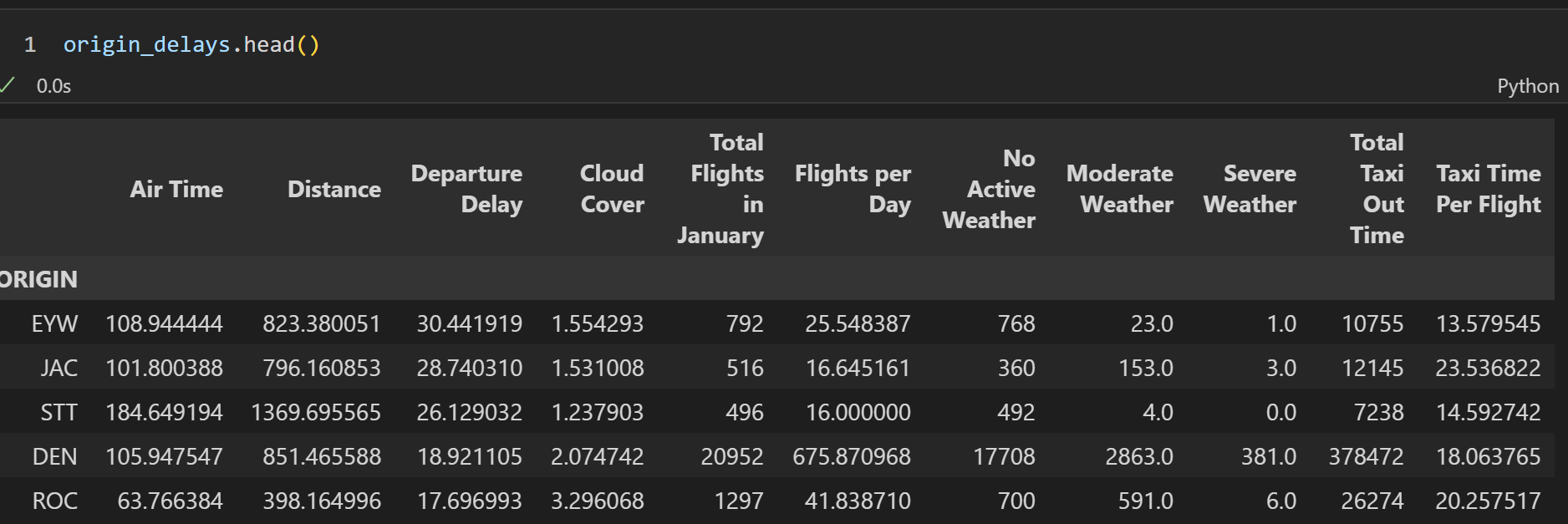
August 31, 2023

Project Writeup

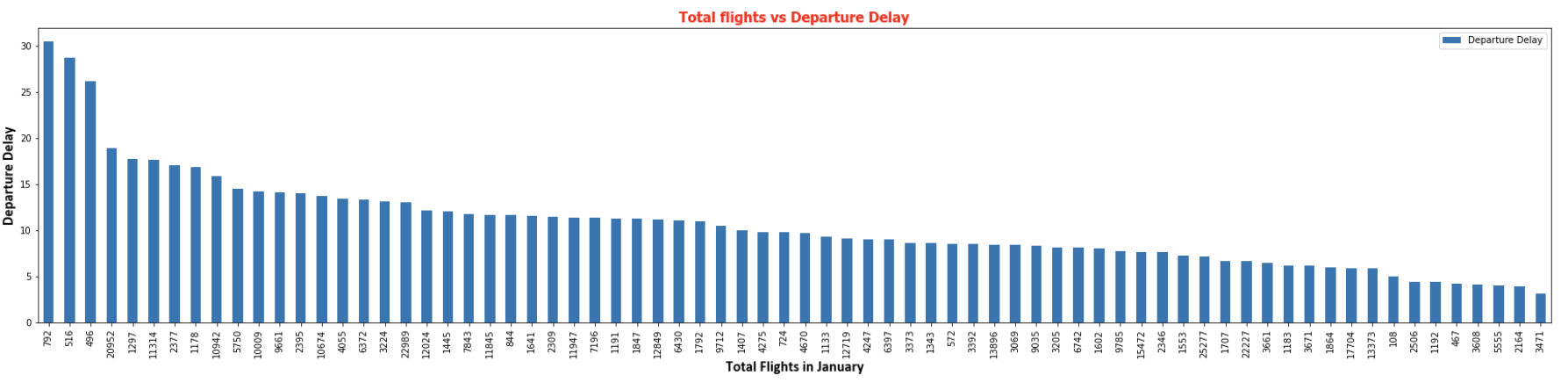
Our project sought to answer the following three research questions:

1. What are the biggest factors in determining flight delays?
2. Is there a subset of the data (Aircraft age, aircraft type, time in air, departure/arrival airport, etc.) that provides a true correlation to flight delays?
3. Does the number of delays early on in the year cause airlines to receive less business in the latter part of the year?

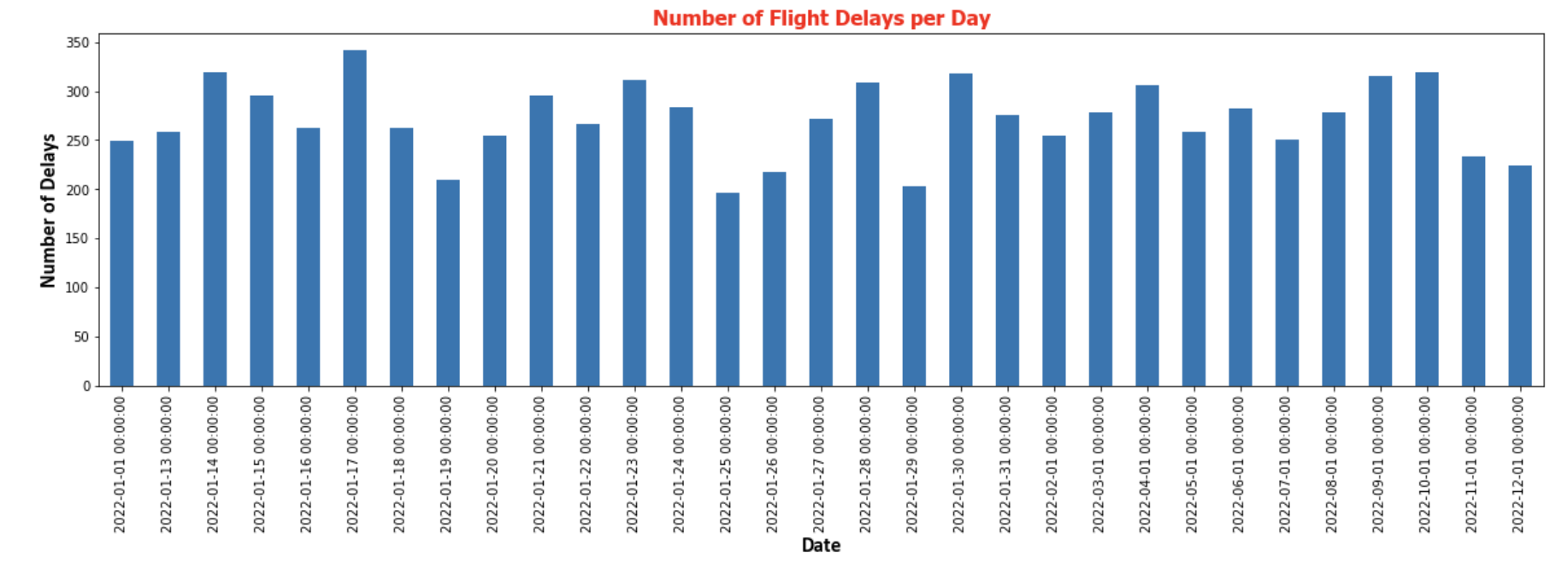
From our research, there does not appear to a set condition in determining factors that cause flight delays. We found that the airport with the highest average flight delay was EYW, Key West International airport, and that it was not the biggest airport, nor the busiest, and not even the airport with the worst weather. So based on these findings, we could not say for certain which factors were the main causes of flight delays. Taking the Head of our dataframe, you’ll notice that there is no specific pattern.



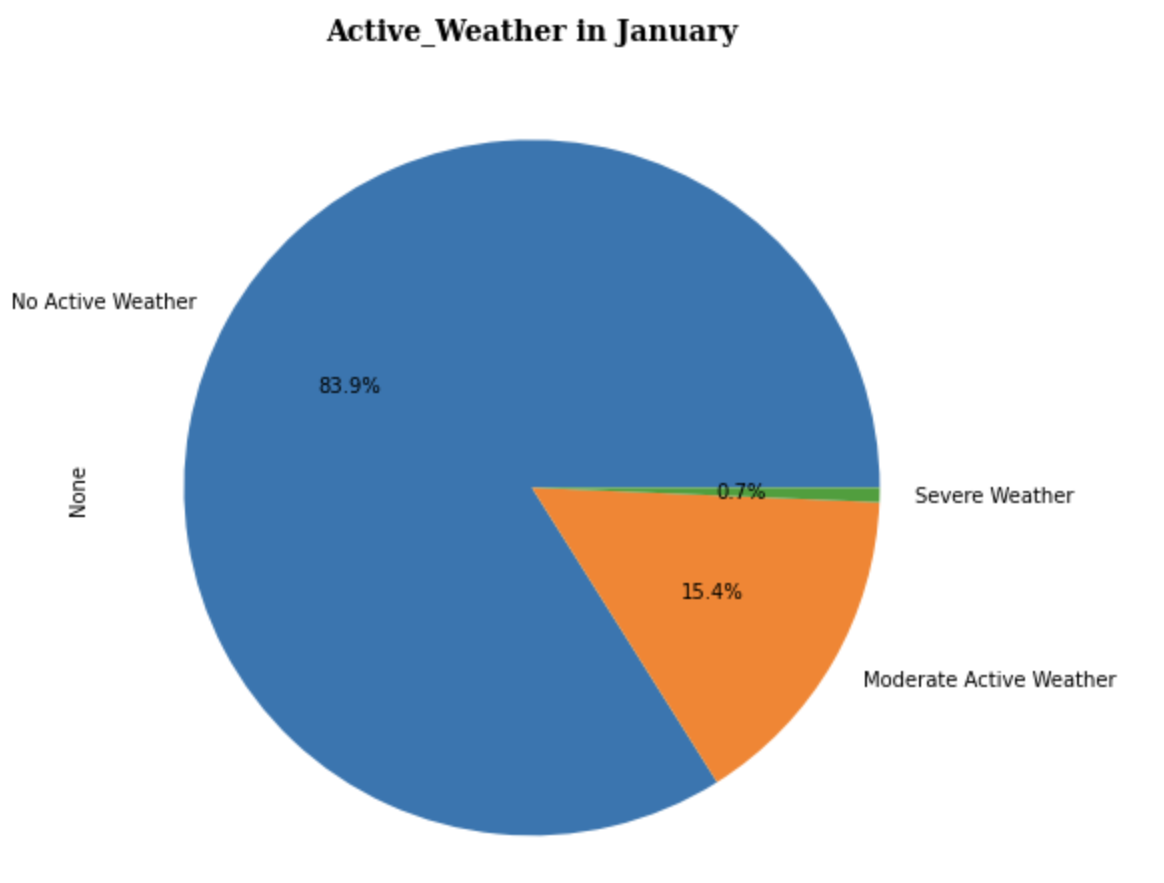
As you can see, there is no discernable pattern in any of the factors for determining flight delays.



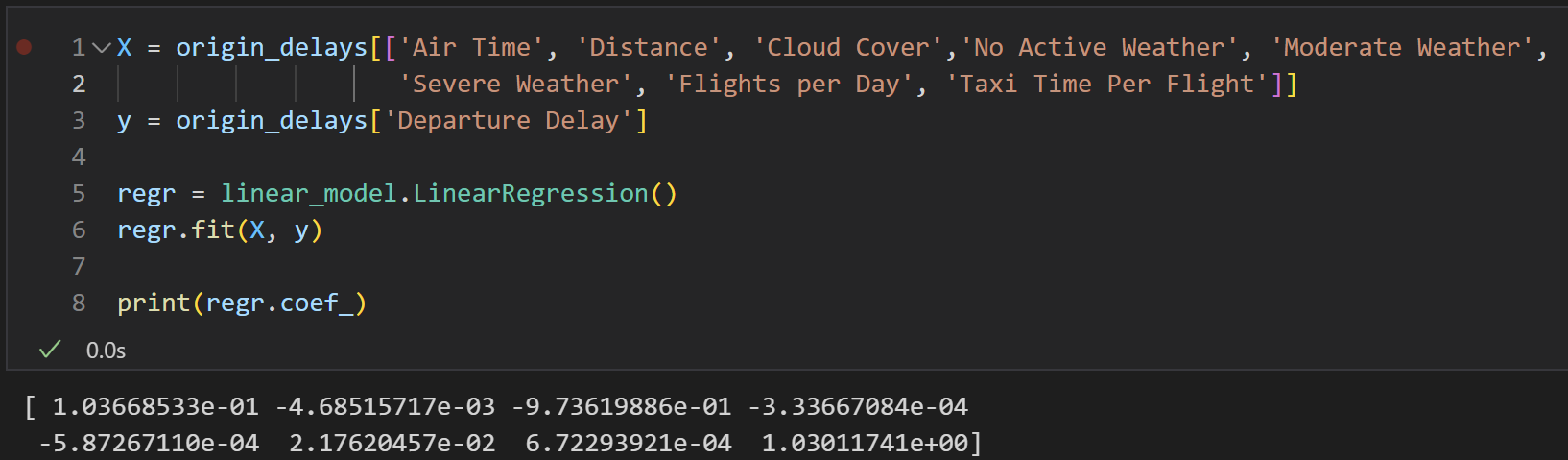
Amongst our Origin airports, represented in the x axis, we can see that there is no correlation between the total number of flights and the total delay time.



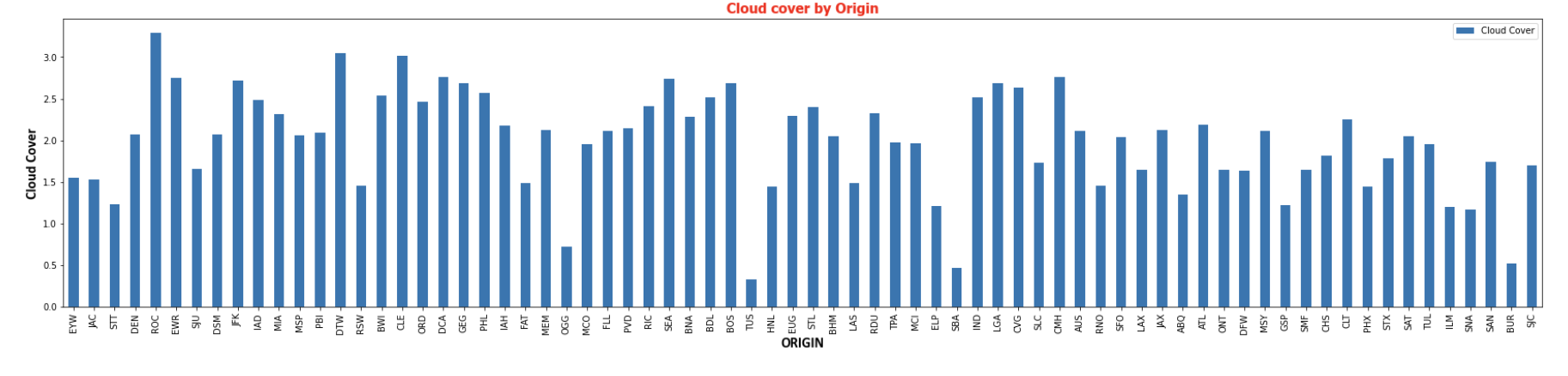
The same goes for the number of delays per day of the month. All of this goes to establishing a lack of trend amongst our data in relation to delays.



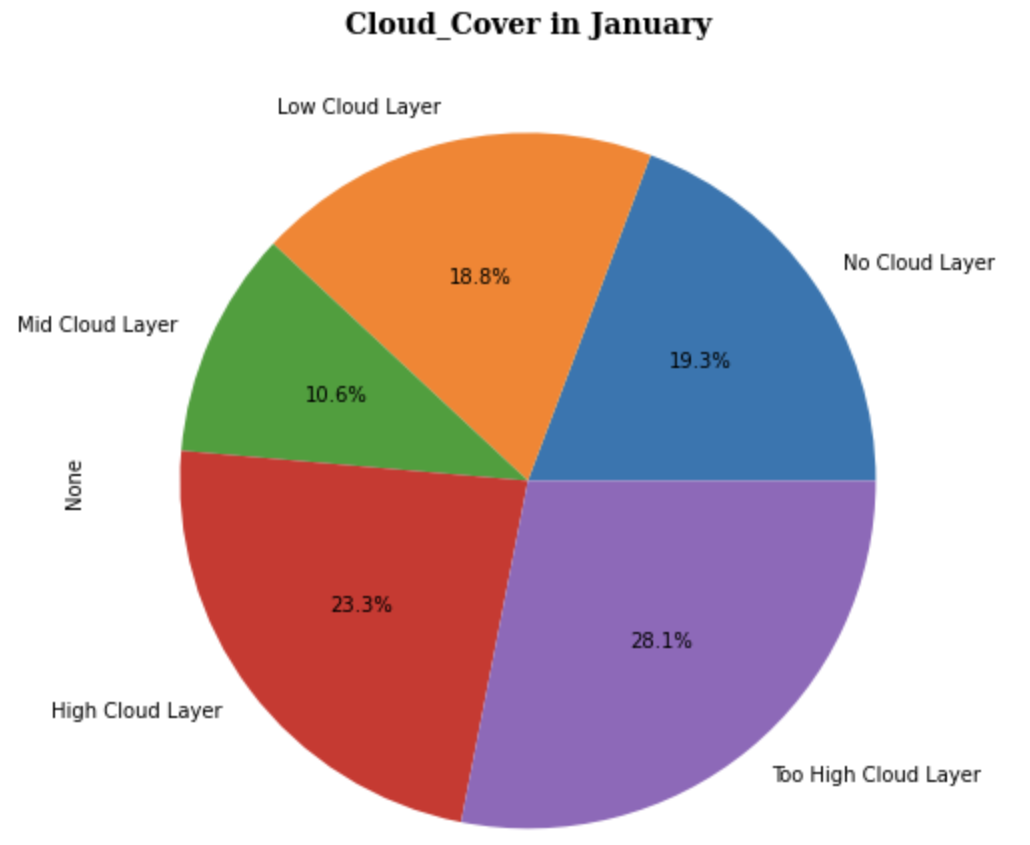
And with active weather being as small of a total of the total weather patterns, only 16.1% of all weather taken being moderate to severe, it makes sense that the amount that it impacts delays is minimal as there is perhaps not enough active weather for it to severely hamper flight times

As for determining a subset of the data, we decided to group our data by the origins of each flight to have a more even and compact dataset. From there, we determined that -amongst the data used- “Taxi Time per Flight”, which is the time it takes for the airplane to travel the tarmac in order to have a clear route for liftoff, was the most impactful factor in determining pre-flight delays via a multilinear regression. A multilinear regression creates a line of best fit through the data, and has a slope for each independent variable. This allows us to measure each variables’ impact on the dependent variable in order to determine the variable that has the largest effect on the dependent variable, in this case departure delays. Now as to why “Taxi Time per Flight” was the most impactful factor could be due to any number of reasons: Perhaps a flight took too long to takeoff and caused a domino affect on flights following it, or maybe pilots take longer than what the airports think they need to takeoff. 

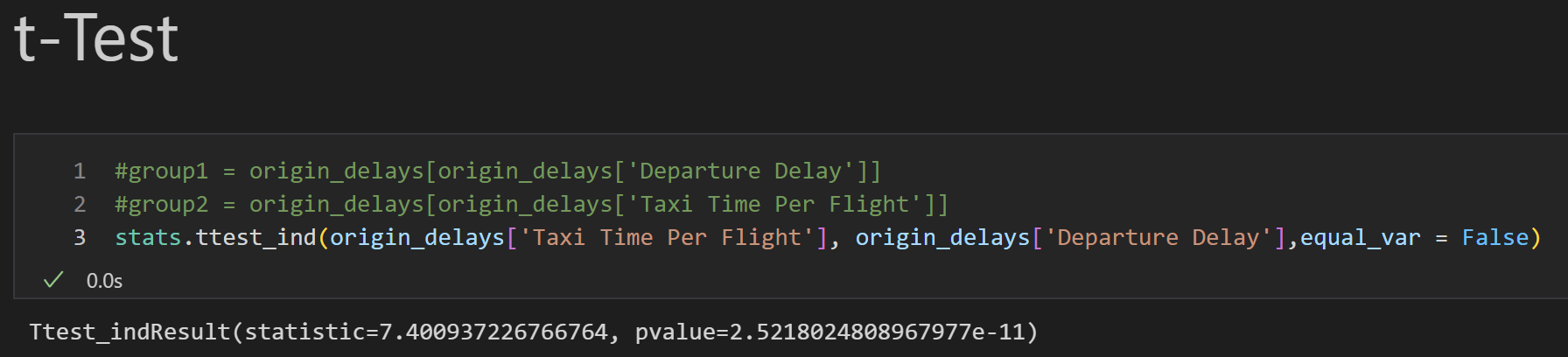
Above is the multilinear regression and a printout of the coefficients of each of independent variables. As we can see, the last coefficient is the most impactful and that corresponds to the last variable input, Taxi Time per Flight. From there, we ran a t test on Taxi Time per flight to ensure that it did have an affect on flight delays. Our Null Hypothesis was that Taxi Time per Flight did not have an effect on departure delays.



The data once again speaks for itself in a lack of correlation between our independent variables and the delays. EYW has middling cloud coverage but viewing the other graphs, has the highest average delays.

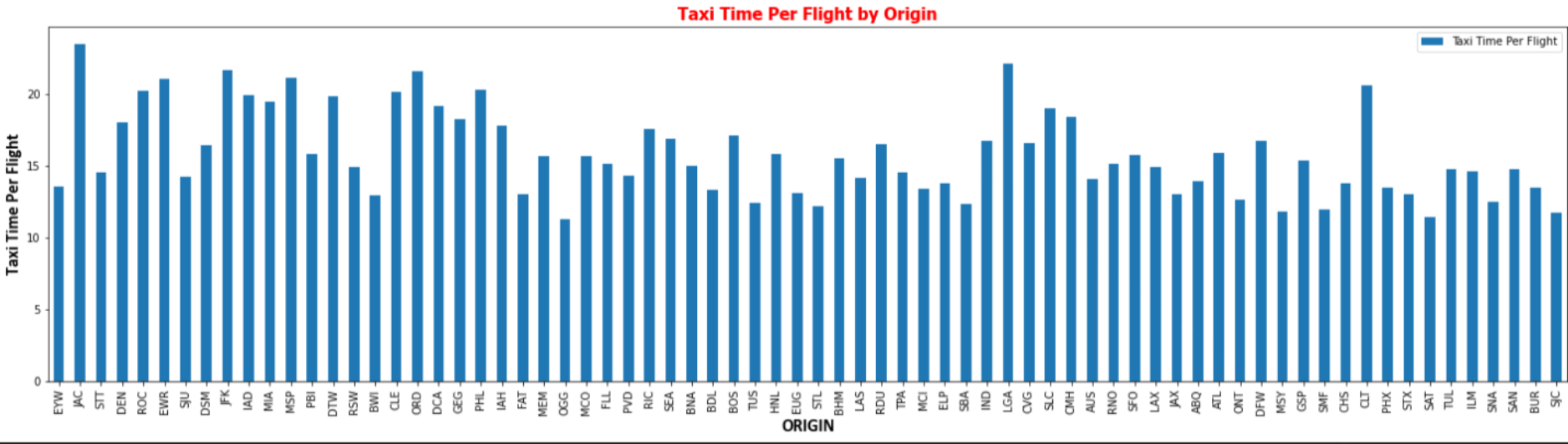


Interestingly, we can see that multiple layers of cloud coverage, which would impede visuals for both pilot and air traffic control, encompassed the majority of the month and would explain it having a strong positive influence on the regression, yet its coefficient is negative which would imply that the more cloud cover is present the less delays you would face.



As we can see from the test, we reject the Null Hypothesis as our p-value is roughly 2.522x10-11, or 0.0000000000252.

Our last research question was not answerable as the data set was too large initially to be manipulated and we decided that in order to be able to use the data that we would only use one month’s worth of data as that still contained over 500000 rows of flights. If we move forward with this project, parsing through different months would be an interesting direction for us to take.



As we have discussed, even with our most influential factor, there is no discernable pattern in the results. Our x axis is ordered from highest to lowest average delays. The left most is, EYW, has the highest average delays, but as we can see it taxi time is no where near the highest and there is pattern to rest of the data.

In conclusion, of out of all our possible factors, it appears as though Taxi Time is the best indicator of departure delays. We determined this through use of a regression and t test to verify our results. We are also unable to answer our last research question due to the compression of our dataset from a full year to just the month of January for the sake of being able to work with the data at a reasonable size. In the future, we would like to perform this analysis on the remaining 11 months of the year. Another interesting avenue to explore would be to look at the individual airlines and to see if there is a more discernable pattern leading to delays. Below shows delays per airlines and it might prove a rich avenue to explore.

