Traveling is something that I like to do so I thought an analysis on flights would be interesting. I found a dataset on Kaggle that tracked January 2020 flights to make predictions for January 2021. Due to the virus, I did not want to make those assumptions as not as much travel is happening and foreign destinations are closed to the US right now. I decided to want to see if there is a trend between certain days traveled and the length of the flight. I also wanted to see what airlines tend to fly further.

Based on the EDA, my initial question and curiosity if flights tend to fly longer on certain days seemed to be false. From the testing, there does not seem to be any correlation or relation between the two. It seems that there are some flights that go out every day and that there is no set pattern to it. I thought there would have been a correlation between day and distance because longer flights may not happen every day and could happen every other day.

I think another option that I could have looked at are when the flights happen. Is the distance of farther travel happen more in the morning or more in the evening, could have been examined. For this I could have created an additional column that is an indicator if the flight happened in the morning or the evening based on departure time. I feel that maybe some other variables could have helped in the analysis, such as if the flight was international or domestic. A further study could have gone into this and a relationship between distance and flight type would have been related as international flights tend to be longer.

Overall, this study has helped that sometimes the data you look at does not go as planned. A big challenge for me was in the dataset I chose itself. A lot of the variables were more descriptive and were of the type string. Some of the integer data was more of a descriptive analysis than a statistical one, such as flight number. As far as the analysis goes, there were some areas that I needed to refresh my memory and used the assignments as examples. I do not come from a statistics background, so refreshers were needed, especially when examining PMF and CDF. This project though has taught the importance of EDA and what you can do in this step for data analysis.