Capstone Project - Predicting Flight Delays

I want to try and understand what factors are causing flight delays and see if I can predict whether a delay will occur or not in the future. People don’t enjoy flight delays and would like to avoid them. I think people would be interested in knowing what airline, day of the week, city, etc. would be most likely to cause a delay. If a customer knew the solution then they might book on certain days, with certain airlines and may consider going to a different airport nearby. A customer who is on a strict time constraint would be much more likely to do this. This could also be helpful from an airlines perspective as they may need to focus more attention on reducing delays under certain conditions. An airline may want to adjust their flight schedule during certain months to ensure that they are arriving on time. In order to solve this problem, I am going to use a 2015 flight delays and cancellations dataset from the Department of Transportation. This data was provided on Kaggle. I also think weather would be very useful during this analysis and will add historical weather data from NOAA's National Centers for Environmental Information (NCEI). The first step I will take in solving this problem is to filter out a smaller list of airports as the dataset is very large. I will include airports that will represent several different climates in the US. I will create many visualizations to understand the data. I plan on at least building a logistic regression model and random forest to predict whether a delay will or will not occur. My deliverables will include the code for the machine learning models, a report and a presentation on my findings.