# Final Project Proposal

Joseph Kerkhof

DS 740

Spring 2020

# Overview

For my final project, I’d like to perform analysis on the [January Flight Delay Dataset](https://www.kaggle.com/divyansh22/flight-delay-prediction). This dataset is sourced from the U.S. Government and provides many columns of data like Airline, Date, Time, Flight Number, and a binary value if the flight was delayed by at least 15 minutes.

I’d either like to predict the likelihood of a flight being delayed (given as a confidence percentage) by at least 15 minutes, or a binary prediction (whether the flight was delayed or not). I’m not sure which path is most likely for success.

Audiences for my analysis could be the Federal Aviation Association (FAA) and the private airline companies. These audience members could use my analysis to reduce the risk of a flight being delayed and improve the number of on-time departures.

Determining the importance of predictor variables is going to be key in the analysis, so I’ll choose techniques where I can peer into the decisions of the model and create recommendations for the audience. I’ll start with using Linear Discriminant Analysis (LDA) and Logistic Regression to predict binary outputs, if this process doesn’t work out very well, it might be easier to predict a probability. In this case I will try Linear Regression. For each method, I will be sure to cross validate to make sure I am obtaining an honest assessment for prediction.