DS 740 Final Project Proposal

Jeff Watson

The data set I have chosen for the final project focuses on airline passenger satisfaction. The data set can be found [here](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction). The data is pre-split into a train and test set, but I will recombine these and use sample() to split out the groups. The data set contains 129,902 observations of 24 variables, 23 of which will be included in the final analysis. There are 4 categorical predictor variables and 18 quantitative predictor variables. Of those 18 quantitative predictors, 14 customer responses ranging from 0-5 rating various aspects of their travel. There are 393 NAs in one column (Arrival Delay in Minutes). As this is such a small subset of the observations (0.3%), I will omit these from the analysis.

The I will attempt to determine which factors are most indicative of a satisfied and dissatisfied customer. I will seek out which features are most and least valuable, according to airline customers. This information could be very valuable for people developing workflows and amenities for air travel by helping them focus time and money in the most effective/efficient areas.

I plan to use Random Forests and Artificial Neural Networks to do the predictions.