Pre Analysis Plan

This research project aims to develop a machine learning model that can distinguish between fabricated and real numeric datasets. We plan to analyze established fake and real datasets to identify patterns, then train the model to independently find these patterns and determine a classification of fake or real. This means that our data will look a bit different, since we are using multiple data sets essentially as observations, as opposed to a single data set with a uniform set of columns and rows.

For the learning type, we plan to use a mix of supervised and unsupervised learning. We will use supervised learning to look for the trends we identified in the beginning and unsupervised learning to look for trends we may have missed that the computer can independently identify. Some elements of feature engineering we have chosen to include are min-max normalization of each variable, one-hot encoding for categorical variables, and feature transformations as needed (such as log for skewness). Along those lines, we will also evaluate conformity with Benford's law by looking through the first digits as well as evaluating the proportion of missing values in the data set as a whole. We will be using the random forests classification model since we want to predict either fabricated or real as an outcome of each data set. To test the accuracy of our model and its predictions, we will use a confusion matrix. We will use this tool for the train/test set, as well as new data.

There are a few potential issues we may face with answering this research question. One issue is if the model wrongly classifies human error in the data set as an indicator of fake data. This may look like a certain degree of false data entry, which the model could interpret as fake because it would violate certain properties and patterns listed above. To prevent this, we will try to include a way that accounts for a certain degree (possibly 5%) of human error allowed in the data set. More than that threshold that would be classified as fabricated data. Another potential issue is unique forms of data cleaning needed to put the data set through the model. If the data is not already clean, this may lead to some other inaccuracies in prediction or even errors in computing by preventing the model from running. Since it is not possible to clean every data set in addition to fitting it to the model, we may focus on a specific kind of data and data sets that are already clean.