## Data Wrangling

This is the preliminary step before thinking about models and analysis. Essentially, I am just browsing and rummaging through data here. I first look at the structure of the data, and then the data types. The next step is seeing how many values are missing. Usually, if something has above 60% of data missing, I will consider dropping it. In this case, phone_type has all data missing, so it was an easy choice to drop. Create some visualizations to view the distributions of data. Other than scores, most of the data was categorical, so not much to see. Finally, output a cleaner version of data to work with.

## Exploratory Data Analysis

This is the step where you perform statistical analysis to see correlation between features. The goal is to identify several key ones to create a model with. Main visualization is a heat map of all the features. Lighter shade/color means more correlation. While most features, had no correlation, the scores were very correlation with each other. Next were scatterplots to visualize the relationship between the independent and dependent variables. Most of them were unclear, so I used the derived scores as the safest choices for features.

## Preprocessing and Training

This step begins the machine learning process. The model I chose was a K nearest neighbor classification model. I first removed any extra features, and dropped duplicate leads. The lead scores did not change, so I just took the first one. If they had, I would have chosen to keep the entry with the most recent matched_date. Then check for any remaining nulls or empty data and impute. Since the features were scores, I chose to use the median to impute. I split the data 70/30. Even though the scores seemed to be out of 100 already, I scaled them just to be safe. When picking the K neighbors for the model, I compared the mean error from K = 1 to K = 40. I found that K of 6-40 gave the least mean error, but using these values would probably lead to overfitting in future data, so values between 3-5 would be optimal. The metric I used to measure performance was a confusion matrix, and mean accuracies when put through cross validation. The KNN model had a 94% accuracy when K = 5. Created a pipeline and saved the model for bookkeeping purposes and to streamline the data inputting process.

## Modeling

I use my model to predict the application outcome for the uncontacted leads. Then using the sum of the scores as a secondary indicator, I sorted the leads from best to worst. The ones with a positive application prediction and highest score would be the ones to contact first.