Alexandra Gates
Homework 2 Write Up

The pipeline is made for a k-nearest neighbors model. After getting the data, I made 3 different dataframes with it: one that had all the data, one that dropped columns that shouldn't be part of the correlation (in this case, personID, zip code, and delinquency in 2 years) and one that dropped only personID and zip code (mainly for the boxplots so I could split the data by delinquency).

I then did a correlation, boxplots, and scatterplots. Scatterplots are commented out in the submitted version because they take SO much time to run, but I used them to understand the data and how they changed for each dataframe or change made to the dataframe.

Z-scores were calculated for each data point in the set for all columns except zip code and person ID. I used the z-scores to determine outliers, defining an outlier as having a z-score greater than or equal to 1.96, however a user can redefine an outliers as they'd like. From this, I made a column that counted how many outliers there were in each row. I determined that if a row had more than 4 outliers (again, user can choose another threshold) in it, the individual may have been lying on parts of the survey or there was bad data collection. To see how it impacted correlations and relationships between variables, I kept the original dataframe, but also made one without rows with more than 4 outliers and another without any row that had outliers. Then looked at boxplots and scatterplots again. From here on out, I used the dataframe that didn't have rows with more than 4 outliers.

Next I looked at NAs. I created a function that found columns with NAs, another that put columns with nas in a list, and then a third to fill nas with a number of the user's choice (mean or median). I chose to use median because it is less likely to be affected by outliers. Before creating this function, I checked the median and mean of columns with missing data on the full dataset, the dataset with less outliers, and the dataset with no outliers (outliers as defined above). I chose median and the dataset with less outliers because the median wasn't as affected by outliers, and in case outliers wasn't someone lying, I wanted the algorithm to account for that. Those sound like opposites, but mainly I wanted to choose the stable one between median and mean, while still accounting for all data/not overfitting the data. I filled nas for all three dataframes, though, so user can do what they'd like.

Next I changed continuous variables to discrete variables. In this function, I wanted the user to have the choice of either cut or qcut. I chose cut because I didn't want each bin to have the same number of entries, and I chose age because I'm curious how the analysis changes by age range. For k-nearest neighbors, I did not use this.

Next I dummified categories. Rather than having zip code as a column, I changed it to multiple columns where each zip code is a column and the data is either a 1 or 0 depending on if it was in that zip code or not.

Lastly, I initialized the KNN model, fit it to the data, predicted the probability of each row having or not having had a delinquent episode in the last 2 years. I then determined how accurate the classifier is.

My accuracy was 83.7%.