Data Wrangling

The data set I am using for my first capstone project concerning financial well being has already been cleaned and wrangled in the sense that outliers, null values, nonsense responses, problematic values, etc. have been dealt with, as described in Section 4.2 of this document: https://s3.amazonaws.com/files.consumerfinance.gov/f/documents/cfpb_nfwbs-puf-user-guide.pdf. Instead, I used a data set from data.world that was meant to be used for data wrangling. It was accompanied with some documentation describing a hypothetical undergraduate applicant and what factors would influence which college they decide to attend.

The data set that I used contains data for colleges in the United States, such as their location, sizes, prices, admission rate, and so on. This was accompanied by a data set giving crime rates for cities also in the United States. First, I removed the irrelevant columns from the college data set, keeping only the information that our hypothetical applicant would want to know about. I then had to create a new column containing both the city and state data to avoid any ambiguity in the cases where different cities had the same names, but in different states. I could then merge the college and crime data sets using this new column. I removed the rows where city or state data wasn't available by performing an inner merge. Using describe() to take a quick look at the data makes it seem as there are no extreme outliers for any of our variables.

I then checked for null values. To deal with the null values in our crime related data, I simply created a summary 'OverallCrime' variable which would assign equal importance to each type of crime and ignore null values. This was then stored as percentile data. To deal with null values in the non-crime variables, such as admission rate and price, I simply dropped the rows where these data weren't available, since such information is of utmost importance to our hypothetical applicant, and thus shouldn't be filled in using averages or back-filling.