**Capstone Project, Data Wrangling**

This data set was already very clean and well-organized. Each observation was in its own row, and the variables were clear and concise. I changed the title of the “bmi” column to “bmi\_(weight/height^2)”, so that your average reader could have some idea of what the bmi metric means. This seemed like the one variable that an average reader may not have been able to understand. It’s less concise, but it seems like the additional explanation is meaningful, considering the high accessibility of the other variables.

I ran a few tests to double-check for “na” values or empty strings, and the data set didn’t contain any. Then, I ran a few tests to see if the data that was entered seemed reasonably correct. I ran “unique” tests on each column to examine the unique strings for each variable. This was very useful for “age”, “sex”, “children”, “smoker”, and “region”. Each string made sense within the given variable. For “charges” and “bmi”, there were many unique strings; with these being ratio and interval data, the categorization provided by “unique” didn’t summarize the data well. I then examined the highest 25 and lowest 25 values from each of those sets to check for outliers and found none. While I ultimately didn’t alter the data set very much, I ran tests to be sure that the data was as clean as I initially believed.