Benjamin Levy Levy 1

Dr. Scotina

24 July 2022

Week One Reflection: Statistics in the Real World

While the first week of the Statistics in the Real World course was more so an introduction to R and data analytics/statistics, I was able to acquire some useful skills and knowledge. One major thing that I learned was data visualization, in which a statistician can represent data visually to be able to better identify trends and correlations. I found learning about the various types of graphs to be interesting. For example, I did not know that histograms were only used for representing fully numerical data, whereas bar plots/bar graphs are only used for categorical/qualitative data. Besides the concepts behind these plots, I learned how to implement them in the R programming language, which I hadn’t had much exposure to prior to engaging in this course. Another useful skill I was able to learn this week was data wrangling, which is the modification of data. A prime example of data wrangling is merging datasets (dataframes). Specifically, there are four main types of joins: inner join, left join, right join, full join. All these joins entail the use of a key, which essentially constrains what data remain. Inner join removes any data whose key’s value is not in the other dataset. For example, if merging a dataset with country codes as its key with a dataset of country names with country codes as its key, if one dataset has the code “CA” but the other does not, that datum will be removed. Left and right join work similarly, though the left or right dataframe, respectively, is persisted, while the right or left dataframe, respectively, has data removed based on the presence of keys in the opposite dataframe. On the other hand, full join persists both tables.

The knowledge I gained from class this past week will assist me in being successful to a larger degree in my study. For reference, I am conducting a study with Dr. Munmun De Choudhury, an associate professor at the Georgia Institute of Technology. Our study will involve identifying teens of color on social media at risk of suicide using both machine learning and deep learning techniques from a unique, private-facing dataset. Being able to effectively represent data from our data source will be an important and outstanding skill for our study, hence my interest in this course and its content.