Reflection 2 Richerson 1

The term “data cleaning” makes pure and raw data seem messy and disorganized. An example in class was with the survey from our professor. The survey was filled to the brim with over 500 responses, but the number of valid responses was only 257. This is because of other factors such as entering but never starting, starting but never finishing, and even when finished, some of the answers were the same. Still, they were worded differently as something as little as a capital letter vs a lowercase letter. This is all to say that instead of “data cleaning,” it is more like “data filtering” or “data reorganization”.

Usually, when prepping data, we can assume that all the responses we received on it were valid, or we could do the task of refining and combing through the data. Both could yield different results, like in the last example when asked, “How do you identify racially?” A lot of responses could have come from, say, Caucasian people, when in reality, most of those responses were unfinished, and the filtered average is a lot more vast.

We could filter it through looking through the most common responses. Such as a lot of people could have answered, “I like it,” but they are worded differently, like “I liked it, I like it, or like it”. By grouping similar responses, we can weed out the outliers and see what is the minority/majority of responses.

Looking at the responses, if they have to change or alter the data, like in a survey, I think they would try to keep the demographics and responses equal, like if there are responses like “Yes” and “I agree” those responses are the same and need to be put in the same category.