Michael Surdek

4-3 Journal: Data Validation & Discovery Walkthrough

After reading the data into RStudio using the read.csv function and assigning each data set to a variable name, the next step was to perform a summary on each of the data sets. In this case, both data sets contained 10 columns and the summary function returns information and descriptive statistics for each column. Running the summary is a fast and simple way to get a better look at your entire data set. In the initial data assessment, we calculated the minimum, maximum, and mean of a couple variables of our choice, based on our initial questions and intuitions. This is a common process, but it can be misleading if your priors are incorrect or if the data is invalid. It is important to run the summary function on your data in order to make sure it is valid by looking at some important information for each variable. More often than not, you will notice aspects of your data of which you were not aware from your initial assessment. Additionally, not only does the summary function provide a five number summary of the numeric variables, but it also tells you the lengths and classes of your variables, which is important to know as you continue with your analysis.

The min, max, and mean that is provided by the summary function for each variable helps the data analyst get a better sense of the distribution of the data. It is important to know if your variables are distributed normally, skewed left/right, uniform, or in another common way. In general, if the mean of a variable is less than halfway between the min and max, you might think that the data is skewed right, and if the mean of a variable is more than halfway between the min and max, you might think the data is skewed left. These descriptive statistics do not tell you everything, but they are a good starting point. It is also important to dive further and find out the frequency of observations around the min, mean, and max.

Data assessment is the process of reviewing unfamiliar data, with the intent of determining if that data can be used to provide insights towards a particular question, problem, or situation. Data assessment consists of the initial steps that a data analyst will typically undergo before allocating further time, effort, capital, or other resources towards a potential project.