My Growth as a Data Scientist

When I think about the general roles and responsibility of someone who works as a data scientist, I feel a clear cut definition would be one gathers and cleans data from an input, and returns in in a meaningful way back to another. As a result, I will focus on this input and output, and how I have shown major progress in it through this last quarter in Stats 331.

I feel like one of the hardest things that stems from data science concepts is the act of importing and cleaning data. Where some things can be repeated on a fast scale, this is the only topic where there is no "guide" and the only way to figure out how to clean your data is to manually go into it and see what is wrong. I think as a result, when first ran into this topic I was often left with confusion cause there is essentially no right answer. The endless amount of errors such as different import functions from different packages, raw data where NA values are referred to as something else, or even multiple rows that contain no actual data stunned me into trying to avoid it at all costs. However, I feel like the first week I started to make a huge improvement was the practice activity in Week 4 where we imported government data about military spending for different data. The process of cleaning this excel file to me seemed to make any future data set we were given to be very easy in comparison. Not only did it require a new function as it was an excel file, but the NA values were encompassed of things such as "0" or "XXX". Even when importing it, there was descriptive paragraph in some rows and as a result we had to use skips or limits as to how much rows of data we actually needed. I feel like as a result this activity gave me the ability to be able to use R in a more comprehensive way. Before, I would have just assumed to go directly into the excel file and remove any unnecessary rows or manually change the NA values. However, this gave me external knowledge of how that process is not needed.

A second key role that a data scientist takes on is the ability to translate this cleaned data set into a meaningful output for others to understand. Initially, this was also a task that I struggled on greatly. I think this stemmed from the same reasoning as above in which there is no clear answer. However, the role of a data scientists needs to have the ability to think about explaining things in a unique way. I think for this, a key resource that helped me was the "From Data to Viz" website. For me, this planted a blueprint about what different types of variables I was using, and following that I had a set of graphs I could choose on my own. I think a clear example of this being shown is in my Lab 7. Given a numeric variable with time, I thought it was best to choose a line graph indicating the change in null values over time to

best illustrate where the incomplete data was taken from. I think overall I have shown growth as a data scientist student by being able to reasonably take in inputs through raw data and also provide output visualizations. In the future, I hope to continue to try and grow my skills in both these areas.