Through this assignment, I learned several different skills or techniques necessary for wrangling data in python. These fundamentals are particularly important because data wrangling and manipulation is often the most difficult aspect of a data scientist's job – as we often must work with several different data sources, each with their own unique quirks and formatting. And it is nearly impossible to interpret find significance or value from large data sets if the data is not uniform. Fortunately, in this case, the data was very tidy to begin with. This allowed us to work on our skills with manipulation. Particularly, I think filtering and grouping are two of the most important manipulation skills a data scientist can know. Because although each row is related to a specific observation, we often want to analyze specific groupings or values of observations to find correlations. These skills are essential to future data analysis projects in that they allow us to essentially make entirely new data sets from existing ones. This allows infinitely more utility than just using the primary data set in analysis.

I think the most challenging aspect of this assignment was the coding. As someone who has prior experience in R, I am much more comfortable with the syntax of dplyr, which I feel is about as close to plain English as one can get with coding, and I often found myself using R syntax when trying to approach a problem. I overcame this, though, by writing what I though the solution would be in R syntax, and then going through step-by-step to find the python equivalent function and syntax. Another problem I had was with the groupby function. When coding, I prefer to solve the problem in steps. For instance, first I would write the code filter all players with more than 500 minutes, test it, then write the code to sort players in descending order, test it, and then write the code to only display the top 5 players. However, with the groupby function this is often not possible, because you can't groupby, see the grouping and then apply a function (such as sum) to it. It all has to be done at once, which I struggled with. I again overcame this by

writing out what I needed to do for each problem first, and then approaching the actual coding aspect step-by-step.

I think that sports data provides a particularly good avenue for practicing data analysis which can be applied to the real world. This is because the data is incredibly abundant, there's a vast plurality of different metrics, and the data can be grouped in several different ways. This mirrors the real world, where nowadays data is collected from nearly every facet of our lives. Business transactions and health records provide comparable avenues for data analysis, but having the skills to manage and analyze this data are crucial. Additionally, finding ways to analyze, manipulate, and compare different sports metrics provides a good introduction to other real-world applications as well. Again, with business, there are several indicators, both complex and simple, which we can use to analyze real-world data. Sports provides a simpler introduction to using these metrics to create potential models, relationships, and graphical outputs that can be used to find meaningful insights from the data.