A lot of the CSV file work and pandas data frame manipulation we used for this assignment drew on techniques that I had already learned in previous data science classes, such as DS2003 (Visualizing Data) and a course in the statistics department. I enjoyed refreshing those techniques and reminding myself how to work in CSV files, clean data, and explore data. However, those classes both worked primarily in R—before this assignment, I had never cleaned data in Python or using pandas. While many of the functions and rhythms were the same, the specific language and syntax used was different from R. It was a great opportunity to learn how to work with CSV files in a new language, which I'm sure will be extremely useful in the future for projects, whether academic or professional. Working with pandas was in many ways easier than "piping" databases as you do in R, and I'm glad to have this experience in Python.

From my previous data science classes, I think I have fairly okay data hygiene and tend to clean and organize my data as a first step, even if it's not required, just so I have to familiarize myself with the data and its structure. I did that on this assignment first, casting the columns into the proper data type so that I could run analyses on them later, since they were initially stored as objects. Knowing to do this first and foremost rather than waiting until it was necessary is a skill I learned in previous classes, and helped me avoid mistakes this time around. For example, when it came time to find the total points scored by all players my code worked immediately since I had already done all the background work in the earlier steps.

Analyzing ACC basketball statistics is not that different from analyzing any other type of data, in which data scientists and interested parties would like to look at the global sum of a variable, relevant averages, outliers, top performers, and other statistics that lend color to the world the data describes. However, it's important to let the questions come from the data, so that the analysis is specific to the dataset in question and actually results in interesting and important conclusions. The Python and pandas skills I have learned will be easily put to use in other contexts, but it's also up to me to make sure that I mold them to the context at hand. Despite requiring similar techniques, the analysis for a basketball league versus a healthcare company is driven by very different variables and dynamics that have to be kept in mind.