Teresa Duong
Professor Williamson
DS 2002
23 September 2024

What did you learn about working with CSV files and pandas DataFrames in this assignment?

In this assignment I learned more about the different ways data can be stored such as in the CSV format and in dataframes. Specifically I learned more about reading CSV files into dataframes and converting them into a form where I can interact with them as fields and observations. In doing so, I learned more about data exploration, for instance, by displaying the first few rows of a dataframe and scanning its fields and the data types of values. I learned more about the importance of building a domain background knowledge for your data when I was trying to interpret unfamiliar basketball acronyms like G, MP, and TRB. I gained more familiarity with working with databases through building new dataframes to hold selected, or filtered data. I also learned more about data analysis by performing many groupings, sorts, and sums of the data. These skills can help me in the future to manage between different formats of data, and overall recognize the basic structures of dataframes that can lead to more complex operations.

What was the most challenging aspect of this assignment, and how did you overcome it?

The most challenging aspect of this assignment was performing specific analyses with an unfamiliar topic, especially in part 4, where we grouped the players by school and then found the total assists for the team. I think my difficulty came from not fully understanding the components of a dataframe as well as not fully understanding the subject matter. At first I interpreted the task as grouping the players by their school and additionally subgrouping the players by their team. I came to a major block doing this, as I tried to apply the groupby attribute twice to the dataframe and received an error about the dataframe not having the groupby attribute. When I realized that after applying the groupby attribute once to my dataframe, the dataframe is now a new object type (DataframeGroupBy), I then understood the reason it did not have a groupby attribute of its own, and I began to look for new approaches. Here I tried to create "for" loops to find the total number of assists for every team in every school. Doing this however, returned an error that the field "Team" did not exist, and upon closer inspection I realized that there was no other field that represented a player's "team". Only then did I research further and review the dataset to realize that a player's school was also the player's team, such as how UVA basketball players attend UVA as well as play for UVA's team. Realizing this led me to complete the question. I believe my strategies in working through this challenge came through making many attempts and learning by the error messages more about my inaccuracies in programming, as well as dataset understanding. With each new error, too, what I thought was important was going to reference texts like pandas documentation and class notebooks to understand more conceptually the programming that I am doing.

How do you think the insights gained from analyzing ACC basketball statistics could be applied to other real-world datasets?

I think the insights I gained from analyzing basketball statistics can certainly help with other real-world datasets because this basketball statistics dataset gave me exposure to large volume data and data with many fields that are more domain specific and descriptive. Datasets in previous assignments in this course have certainly built my knowledge on data science and data management, but they have also been smaller datasets with fewer rows and fields where I can view the entire dataset on one screen and check that my calculations make sense. In working with a large dataset, I can no longer rely on what I visually see to understand the dataset, but rather have to transfer many manual processes to data science processes. To me, this does make data exploration and manipulation more of a science. Certainly this will be useful in fields like healthcare where a hospital may admit thousands of patients that cannot all be accessed by a human scrolling on a computer screen. There will have to be data management systems in place for searching and organization to occur. As mentioned before, I also believe that studying this basketball statistics dataset would help me in approaching dataset projects in fields where terminology and commonly used metrics might be unfamiliar to me. Healthcare is also an example of a field with many acronyms and terminology that are not used casually such as cardiac information like SP (systolic pressure), DP (diastolic pressure), and SACT (sino-atrial conduction time). For data to be most accurate then, it is important for all members interacting with it to understand what is being measured, not only physicians and health team members, but also developers supporting healthcare technology.