Sahil Rajapkar
CIS3120 - Avinash Jairam
12/21/21

Homework 2 (Part 1)

For this project, I particularly enjoyed working on it simply because I got to choose a dataset of my liking. This gave me the freedom to continue my progress with web scraping and programming with the addition of my interests involved. So, I went ahead and chose data regarding the NBA, specifically that of which pertains to players in NBA history that have had the highest scoring records in terms of points and many other factors. I chose this dataset from Wikipedia because I simply thought it would be interesting to find some statistics through the use of technology. Nevertheless, I'd like to dive into my process for the homework assignment.

I began with scraping data from Wikipedia through the usual web scraping ideologies such as looking at the classes and tags, importing different packages, and many more concepts that we have learned in class and that I have mentioned before. After doing so, I created a program using a for loop and if statement to create a CSV file that matches the data of the table on Wikipedia within a personal folder of mine. This task was a bit challenging at first but I was able to manage through with the help of YouTube. Next, I'd like to speak about how to incorporate using a DataFrame.

I used a Data Frame to replicate that data that was placed in the CSV file. So, an easy way to think about this is I copied a table from Wikipedia to a CSV file then I copied that table into a DataFrame in Anaconda-Spyder. This was a nice transition I had seen of data which would, later on, help me in finding the stats for this table via Python instead of having to use some commands through Excel (which could also have been done but would have defeated the purpose of this class and the topic of programming especially for analytical purposes).

So, this was one of the easier sections. I mainly used the .describe() function to give me the stats of some designated columns (in my case, Rank, Points per Game, and the rest). I also learned one important thing from this which was that the type of the column mattered from an object to a float64. They each will give a different set of results, though still similar. But, lastly, how can my results be used to solve a problem to my dataset…Well, this is a great question, indeed. Let's say one wanted to know what the average (mean) points per game resonated for the top 50 NBA career-scoring leaders. Let's say you didn't want to go through the hassle of using a calculator, you can turn to a program! Not only would you be able to solve this problem of trying to get a distinct piece of information, you'd be able to apply this to all things ranging from the maximum, count, minimum, standard deviation, and so much more. The keyword is assistance!

For example, if I chose to scrape data about the NBA players (like this), then I could assist coaches in producing better yields. I can help them in trying to get an advantage over another team. I could help them get better insights in what the team and franchise are doing well in and not so well in. The opportunity is endless. As the saying goes, "All information is good, even when it's bad", so for this case, I would assume a coach would want as much information as possible to see the team, franchise, players, community, and all stakeholders alike win. This is just one of many factors that play into differentiating great teams from good teams, hence, basketball needs data at its best!