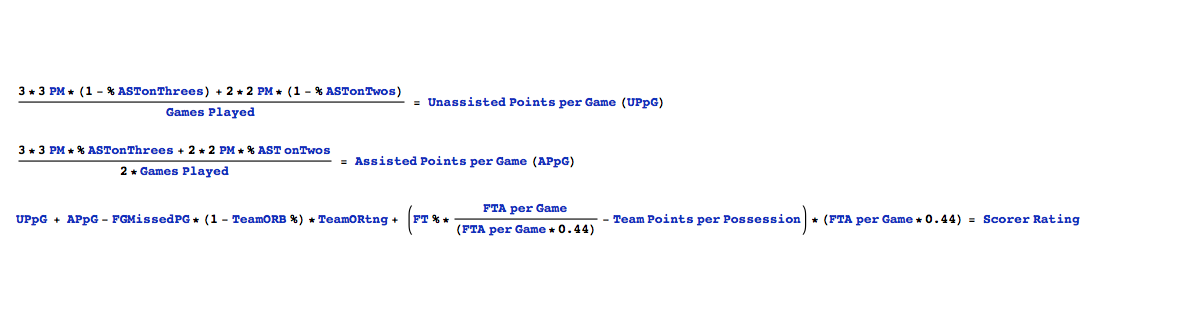
Danny Arenson

04/10/2017

Capstone 1: Data Wangling

For my first capstone project, I decided to work on NBA data. Specifically, predicting the “2016-2017 All- NBA Team Awards”, based on using a formula that was a calculated variable to assess a “Scorer Rating”, in effect providing a better weighted assessment of that player’s performance and not just on the basic in game statistics, ie) Points, Rebounds, Assists, etc… The formula is as follows:

At first glance this seemed like a daunting task, particularly with how/where I would be able to find this type of data, but after giving it some thought, and doing a bit of further research, I was able to realize that I could calculate most of these variables by using basic player statistics, because the complicated variables have a formula to calculate as well. Essentially, my plan was to take the basic player stats, use them to calculate the more complex variables, and take those variables that were derived and apply to the Scorer Rating formula.

I was able to find all of my data online with the various popular NBA stats websites, but they didn’t consist of data all together on one table. My next plan was to scrape the player data from each website by using Pythons package “Beautiful Soup”. I was able to find various helpful YouTube links which provided the proper steps for web-scraping. I seemed to be making some progress in web scraping, but for one reason or another, eventually hit a roadblock with being able to complete this task. I brought this up to my mentor in case he would be able to provide some insight and troubleshoot this issue. He then brought up an interesting idea, perhaps I was overthinking this and if I had the ability to copy and paste the data on to a spreadsheet, then I should just give that a try. So I did this, and fortunately it worked. It took making 5 separate tables from three websites to combine all player and team data together.

I took the combined initial player data and imported as a csv on python where I would then have my table and use the various python packages to evaluate the numbers and do any further data cleaning. There were no particular outliers within the data and because these websites are constantly being updated, there never seemed to be an issue with any missing values.

As mentioned earlier, I took many of the variables that were accessible online which made my initial table, but for those that I could not find online, I made a few new variables and added them as a new column to my table. After having all necessary variables, I then created the final variable for this table by combining all components.