CapProj1 Mini-Proj: Data Wrangling

Source of the Data and Method of Acquirement:

The data was scraped and collected from Basketball-Reference and ESPN. The data includes statistics about all the NBA players from 2014-2018.

Links: https://www.basketball-reference.com/, http://www.espn.com/nba/statistics/

The goal of this project is to create an accurate predictor for the metric RPM. Real-Plus-Minus is an unbiased comparative metric to evaluate player value. Specifically, a player's estimated on-court impact on team performance, measured in net point differential per 100 offensive and defensive possessions. The data collected from these two sites, will help achieve the goal of creating this predictor.

What kind of cleaning steps did you perform?

Once the data was collected via python requests. The html format needed to be converted (BeautifulSoup), and then needed to be sorted through such that the contents of the tables were stored into some combination of lists and dictionaries. The data was then arranged into a DataFrame, where it would be easily analyzed. This process was done for both sites.

How did you deal with missing values, if any?

The two DataFrames would have to be merged together such that no data would be lost or changed. The data also had to be arranged such that all information for a player that year was on a single row, and then the two could be merged into one. Some players were missing from the data because the spelling of their names was different in each DataFrame. Those names had to be changed to match the others. Additionally, some positions were different, so the same had to be done for them. No outliers were removed, due to the high number of variables and proportionally lower number of datapoints. At this point, the data had been cleaned and sorted enough to be saved into a csv for EDA.