We wanted to pull data from three different sources. Two of these sources were from basketball-reference.com and the other was from espn.com. Basketball-reference had one dataset that included common stats like points, three-pointers, two-pointers, etc. They also have a separate dataset with “advanced” stats that include less common stats like offensive win share, defensive win share, and player efficiency rating. The third dataset from espn contains the salaries of players.

We scraped data from each of the sites using Beautiful Soup. Each of the websites urls included the year the stats are from which made it easy to loop through several different years. The function we wrote takes in a list of years the user would like data from. It creates the urls with the first year on the list and scrapes the data from each of the three websites and creates three temporary dataframes. A column for players names exists on all three dataframes so we can loop through each player and see if their name exists on all three dataframes. If it doesn’t, that name is ignored because we will be missing critical information. If it does exist on all three, we index those names, create an array that includes data on that player from all three dataframes and that completed array gets added to the final dataframe. This process is completed for every player and then loops through every year that was requested. At the end, we do some cleaning to deal with missing values and players that play on different teams in the same year. We have worked with data as far back as 2010 with this scraper. We haven’t used data before this point to reduce the impact of inflation, changes to salary caps, and how the game has evolved over time.