Capstone Project #1: The Milestone Report

The Proposal - Popping The Question?

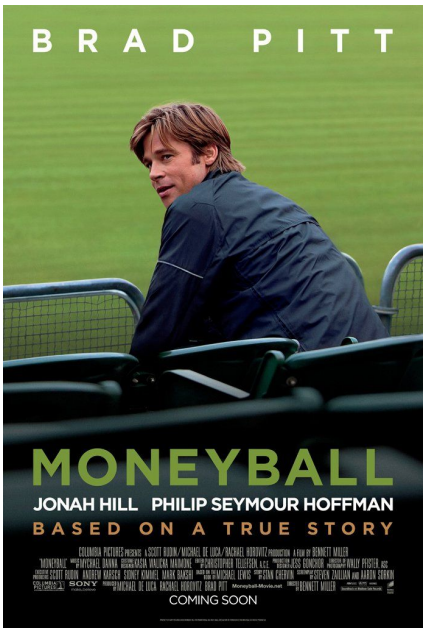


Image via Amazon

The original motivation for this project was trying to predict if the Sacramento Kings were going to make the playoffs this year. So far, I've found some very interesting data and insights with regards to both team-level and individual player-level data that could potentially help answer this question. However, I've decided to pivot in a new direction, namely to focus on predicting an individual player's win share.

There are a few reasons for this:

- Firstly, as of today (March 29, 2019), the Sacramento Kings are in ninth place in the NBA's Western Conference. This normally wouldn't be a bad position to be in late in the season, however, the San Antonio Spurs, who are currently in eighth place are six and a half games ahead of them with only seven games left in the regular season.

  What does this mean? That it is virtually impossible for the Kings to make the playoffs this year (FiveThirtyEight has the odds at <1%). While I could still theoretically continue in this direction, I think the value it would provide would be minimal at best.

- Secondly, I'm a big fan of the movie *Moneyball*, which is about how the Oakland A's used analytics to compete with (and beat) teams like the Red Sox and Yankees despite having one of the lowest payrolls in all of baseball. As I was doing my exploratory data analysis, particularly when it came to data on individual players, I kept asking myself: which players are diamonds in the rough, so to speak?

We all know players like LeBron James, Kevin Durant, and Anthony Davis, who are well-known superstars who stats back up the fact that they are for the most part the primary contributors to their teams' success. There are so many other players in the league though and I became particularly interested in trying discovering the

players that may not be particularly popular with the fans but are (or at least should be) very popular with the front office personnel of NBA teams. This may be of particular value considering that for teams that don't make the playoffs, they will look towards finding players in the offseason via free agency to bolster their roster with the hopes of making the playoffs in 2019-2020.

So in summary, I am pivoting the focus of my project. Initially, I was trying to answer the following question: 'will the Sacramento Kings make the playoffs this season?'. Due to a change in circumstances and perspective, I will now focus on trying to answer the following question: can I predict an individual player's win share? By answering this question, I hope to discover deeper insights into how much player's and their individual skill sets contribute to a team's win total.

The Data 'Rodeo' - Wrangling the data

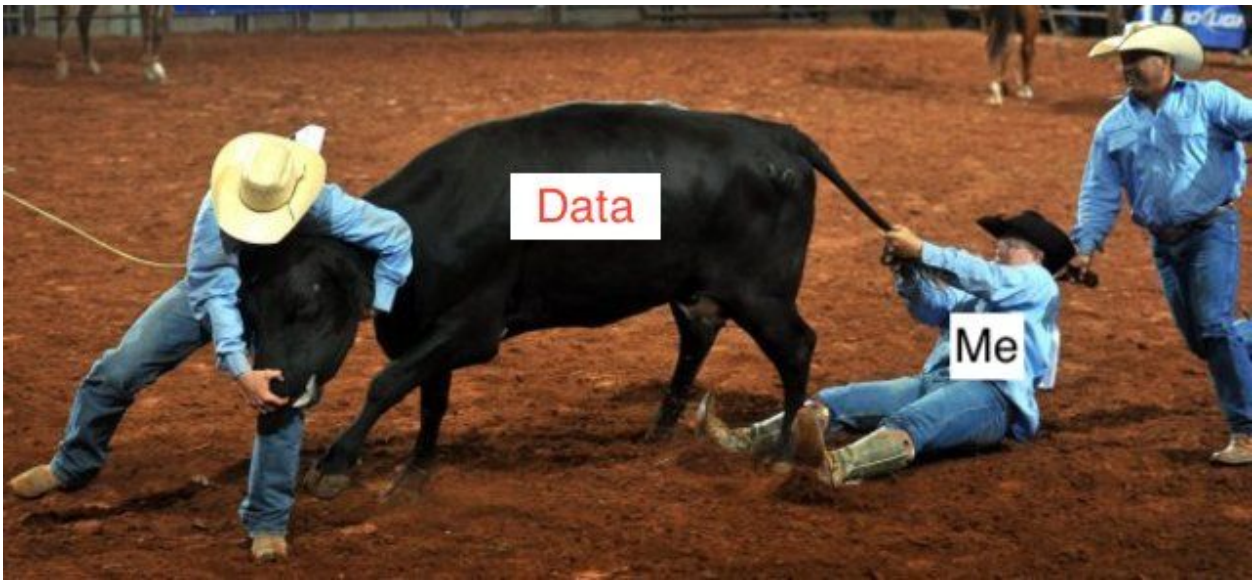


Image by Times Record News Wichita Falls

This project originally got started with one rather simple question: will the Sacramento Kings make the playoffs in 2019? However, since writing the project proposal, the Kings have gone from a record of 25 wins and 25 losses to a record of 40 wins and 42 losses. This record puts them in 9th place in the Western Conference, 6.5 games back of the 8th place San Antonio Spurs. With only two games left it is impossible for them to make the playoffs this year. But, hey there is always next year! While I was exploring the individual player data though, another question came to mind: can we predict how good a player is and how much he contributes to a team winning?

So we have a question we want to try and answer, what is the next step! Data wrangling! And in this case, no we won't be herding livestock. Instead, we're going to be wrangling NBA data from the wild, wild web! With certain fields, getting data that is useful can be the most difficult part of the project, if you can even find it at all. However, the NBA has seen a boon in the last few years when it comes to analytics. The Golden State Warriors have risen from an also-ran to perennial NBA title contender, and they credited much of this success to their use of analytics (check out this article). The rest of the NBA has followed suit, with nearly every team having some sort of data analytics department. This has not only helped teams on the court, but it's also been a boon to fans too particularly those who are analytically inclined.

Side note: I find it interesting that the two teams that started the 'analytics revolution' in their respective sports are both located in Oakland (i.e. the Oakland Athletics for MLB and the Golden State Warriors for the NBA).

After the excitement about finally picking a question I wanted to answer for this project, I was hit with the realization of how do I get the data to even start this? Luckily, this question was quickly answered. I follow Nate Silver's blog FiveThirtyEight and in addition to analysis on politics, culture, and economics, the blog has a section completely devoted to the NBA. Namely, it predicts a teams chance of making the playoffs which are updated after every game and depth chart revision. After reviewing the details of how FiveThirtyEight NBA predictions worked, I noticed one of their references was basketball-reference.com so I decided to check it out.

What I found amounted to the 'El Dorado' of basketball statistics. There was data on individual players and teams dating back to the late 1940's! For the most recent seasons, they had data sets on advanced statistics like Player Efficiency Rating (measures per-minute production), Win Shares (estimate of the number of wins contributed by a player) and Box +/- (estimate of the points per 100 possessions a player contributed above a league-average player). I'd found my treasure chest, now it was time to gather the data and get it ready for analysis!

With this in mind, I decided to use individual player data that combined statistics like points, assists, and rebounds (just to name a few) on a per 100 possession basis and advanced statistics like true shooting percentage and win shares (which I'll get to a little later). For clarification, why did I decide to use per 100 possession versus per game? In one word: pace. Teams play the game at different speeds, with some being more oriented towards the fast-break and others geared more towards slowing things down and running plays on the offensive end. As a result, teams that play faster will have more possessions per game on average than slower teams which has a direct impact on player stats. Per 100 possession statistics eliminate this discrepancy by showing on average how many points (or rebound, steals, etc.) he score per 100 possessions.

Luckily, the data wrangling process was not as hard as I was expecting. Using the urllib package with BeautifulSoup I was able to create a function that scraped data from basketball-reference.com and returned the raw data in a pandas DataFrame. Here is a step-by-step of how the function worked:

1. Create variable that stores URL we'll be scraping
2. Create variable -- html -- that opens the url
3. BeautifulSoup then passes through the website and returns it as an object
4. Uses list comprehension to find column headers and rows (i.e. player statistics), stores them in variables called 'header' and 'player_stats', respectively
5. Use pandas.DataFrame with the variable 'player_stats' as the data and 'headers' for the columns argument
6. Adds a column called 'Year' (mainly for organization purposes)
7. Returns the DataFrame

Using this function I was able to scrape per 100 possession and advanced statistics from basketball-reference.com for the past ten seasons (which did not include 2018-2019 as it is not complete yet). When I originally gathered data, I had to copy and paste into a text editor and then save it as a CSV file, which was quite tedious and was not particularly fun. This process sped up that process significantly, as gathering the data took seconds.

However, the raw data was not clean and there were a few steps I needed to take in order to make it usable for analysis. Luckily with a little experimentation, I was able to create two functions -- clean_per_poss &

clean_advanced -- that returned cleaned up pandas DataFrames ready for analysis. Here is the general overview of what they did:

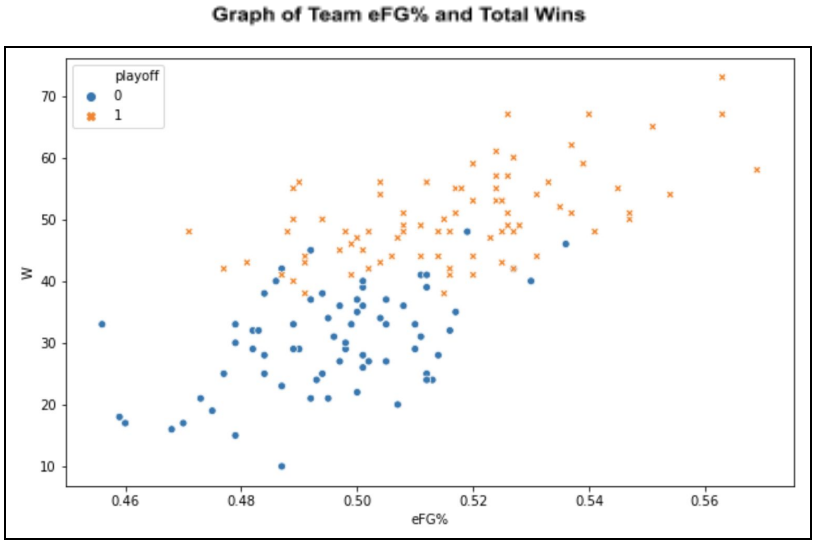
1. Dropped duplicate players
2. Replaced ‘None’ rows with ‘NaN’
3. Dropped rows with ‘NaN’
4. Dropped blank column(s) -- the advanced stats DataFrame returned two blank columns while the per-possession one only had one
5. Converted variables to numerics
6. Reset the index

While it took some time up front to tinker and create these functions, it helped cut down the clean-up time significantly in the long run. Once they were created, all that we needed to do was pass the raw DataFrame into the function and a second later a clean DataFrame was passed out! Then we merged the two data sets (per 100 possessions and advanced stats) together and voila, we had data for every player that played in the NBA for that particular season!

At the very end, we concatenated each seasons DataFrame together to return a data set that combined all ten seasons into one DataFrame. After checking and addressing null values -- which essentially was just replacing the null values with 0 because most of the observations had not taken a 3-point shot for example -- we had a clean and usable data set! Next step: exploratory data analysis!

Discovering a New World - Initial Exploratory Data Analysis of NBA Players

When I first began this project, I was initially focused on whether or not the Sacramento Kings would make the playoffs. I pulled team statistics from the past 5 seasons and individual player statistics from the past 10 seasons and began some initial exploration.



From the initial EDA of the team stats, there was a clear pattern amongst many of the variables indicating which teams made the playoffs from those that didn’t[1]. However, I was particularly fascinated by the individual player stats.

This was for a few reasons. Firstly, while basketball is a team sport, the talent level of each individual player directly contributes to that particular team's ability to win. Basically, the more talented your players, the higher the likelihood you are going to win games.

Secondly, looking at this from a team perspective is similar to looking at the economy from a macro perspective. You are able to assess a general direction but aren’t really able to capture the small nuances that

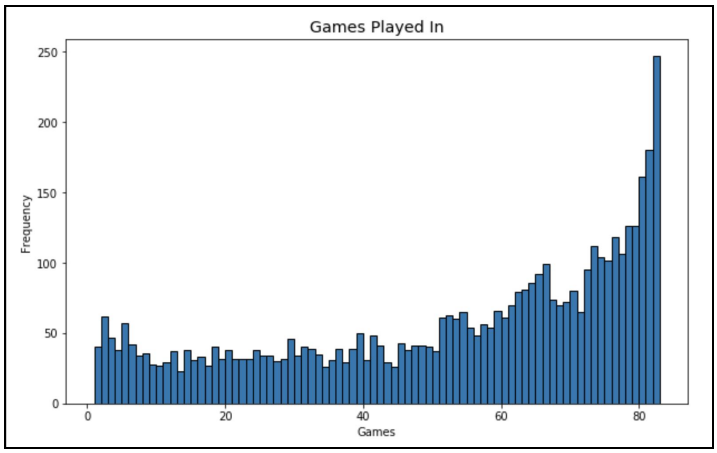
[1] See graph; this shows the relationship between a team’s effective field goal % and their win total. As we can see there is a fairly strong positive relationship between the two variables in addition to a clear ‘cut-off’ around 40 wins, with those below generally not making the playoffs while those above generally do.

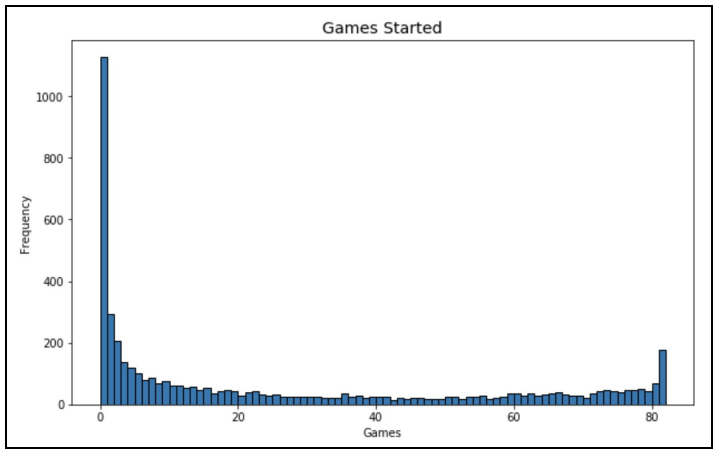
cause the economy to go that particular direction. The more I explored, the more I became interested in trying to assess an individual player's ability to contribute.

My initial hypothesis was that by examining each part of a player's game, we could highlight the skills that contribute to arguably the most important aspect of professional basketball: winning. Then I started asking myself questions like 'does a lockdown defender have a bigger impact on winning than a ball-handler?'. Or 'does being a better offensive player play a more significant impact than a defensive-oriented player?'.

Lastly, I had a lot more data to work with. For the team data, I could gather hundreds of observations while for individual players I could gather thousands, depending on how many seasons back I wanted to go. It seemed like a no-brainer and is why I decided to go the individual-player route.

Histograms of Games Played In and Total Games Started





When it comes to my exploratory data analysis, my focus was primarily on visualizing the forty or so potential explanatory variables and their relationship with our target variable, win shares per 48 minutes. Why did I choose this particular metric? Well for one thing 48 minutes is equal to 1 NBA game. We could've gone down the more macro route of trying to predict overall win shares (for the whole season) but I didn't think this would provide the same value as what a player did on a game-by-game basis. Additionally, I planned on comparing this number — i.e. win shares per 48 minutes — with the salary to generate a unique perspective on the value a player provided to a team compared to their cost (i.e. their salary).

One of the first interesting aspects of the data was in regards to the number of games played in. It was significantly left-skewed, with a sharp uptick around the 60 game mark[2]. This potentially signaled that while teams have between 13-15 players on their active roster on any given night, that only a portion of these players actually get playing time.

When we took a look at the number of games started[3] , there was further evidence supporting the previously stated 'portion of players' hypothesis. The overwhelming majority of players started exactly zero games! From there, the histogram quickly descends with a relatively small number of players starting between 20 and 70 games. At 82 games, there is a spike which gives us some evidence that teams could be using a core group of starters, leading to only a portion of players playing in any given game.

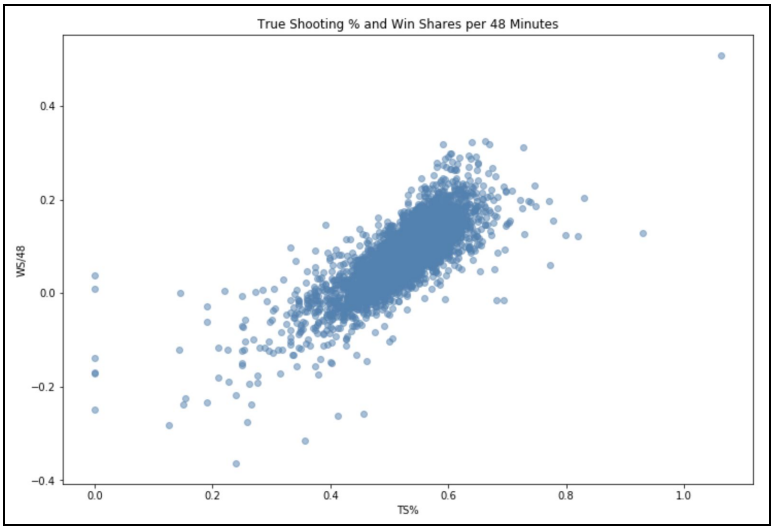
[2] Refer to histogram titled *Games Played In*

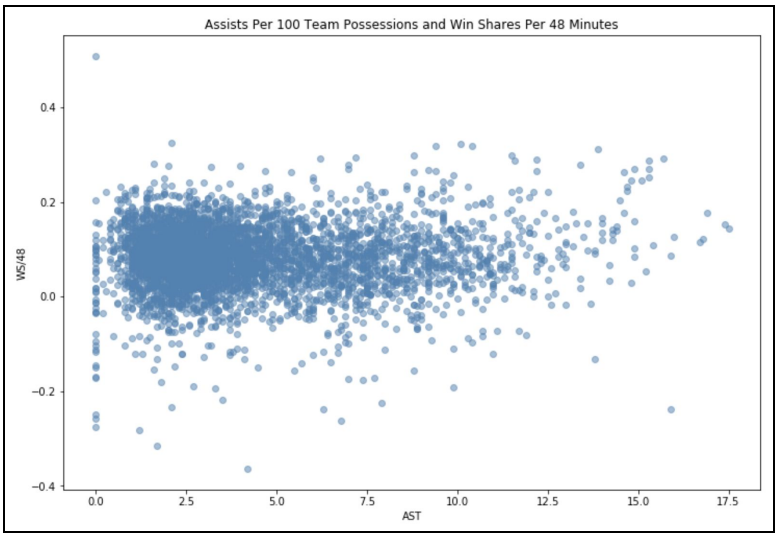
[3] Refer to histogram titled *Games Started*

Now when creating the first scatter plots — focusing on win shares per 48 minutes and minutes played, field goals, and 3-point field goals — there was a significant distortion of the graphs due to outliers. To address this, we used the PER statistic which stands for player efficiency rating.

```
There are 2526 observations that meet the minimum 58-game threshold.
This is in comparison to 4759 in the original data set.
Thats a difference of 2233 observations! Or, 46.92162218953562 % of observations.
```

Now why use PER for this? Firstly, because it is somewhat similar to win shares in that it uses both offensive and defensive statistic to give a holistic picture of a player's overall skill set. Secondly, it had a statistical cut-off in that players had to have played 6.09 MPG (minutes per game). After instituting the PER cut-off the distortion was for the most part eliminated. I want to note however that I didn't institute the second part of the cut-off which was that the player had to have played in more than 500 minutes total that season as well. The primary reason for this was due to the fact that it would have eliminated approximately 46% of my data set if I had and I did not want to eliminate so much data for what at best would be a marginal return[4].





That being said there were some interesting takeaways from the exploratory data analysis with some being expected and others being rather unexpected/disappointing.

The first, and perhaps the most obvious was the correlation between shooting and win share. Field goals made, field goal percentage and true shooting percentage[5] — which essentially combines field goal percentage, free throw percentage, and three-point percentage into one — all appeared to have a strong positive impact on win share. Basically, the more you shoot and the more shots you make, the higher your likely contribution is going to be towards your team winning.

The second takeaway was that assists[6] and rebounds don't appear to have any significant relationship with a player's win share. This was rather surprising since outside of scoring these are perhaps the two most common statistics used to describe how 'good' a player is. Additionally, it is generally assumed that better teamwork generally equals more wins but we're not seeing that to be the case at least with these particular graphs. With regards to rebounding, if you squint just right there may appear to be a slight positive

[4] See above Jupyter Notebook output which compares the number of observations from original data set to a data set where observations played in a minimum total of 58 games.

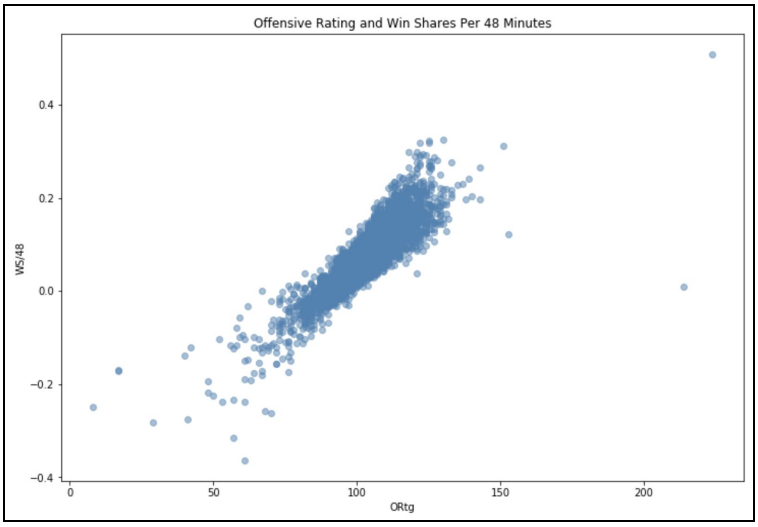
[5] Refer to scatter plot titled *True Shooting % and Win Shares per 48 Minutes*

[6] Refer to scatter plot titled *Assists Per 100 Team Possessions and Win Shares Per 48 Minutes*

correlation with win share but overall I'd say their relationship is ambiguous at best.

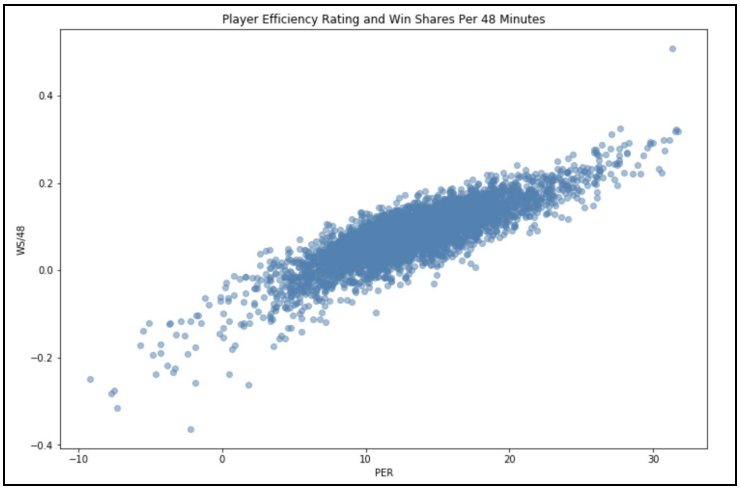
The third takeaway was that offense is king[7]. Most of the variables associated with defense — like steals and blocks — returned as indiscernible blobs with no relationship whatsoever with win share. After seeing the relationships with offensive variables and shooting in particular, being a good offensive player who can shoot the ball well looks to have more of an impact on winning than being a good rebounder who can block shots.

In a way this makes sense. A hypothesis I have is that while grabbing steals or blocking shots may lead to a team gaining additional possessions, the rarity of these respective statistics (averages usually hover around two per 100 possessions for both) cancels out any positive impact they may have.



The fourth and perhaps most important takeaway is the potential for multicollinearity between variables. While I didn't technically test for this, there are quite a few variables that interact with each other. A prime example of this would be the three variables associated with FG — field goals made, field goals attempted and field goal percentage. The three have a very intricate relationship, with making more field goals (or shooting more) having a direct impact on field goal percentage.

There are a few that may not be as obvious, especially to those unfamiliar with basketball statistics. One is PER — player efficiency rating — which is calculated by combining numerous offensive and defensive variables together. There are weights given to each but we'll have to be careful when it comes time to determine which features to use for our machine learning algorithms.



Don't just see, observe: despite a strong relationship with WS/48, we have to be careful going forward using variables like PER because they could induce multicollinearity.

[7] Refer to scatter plot titled *Offensive Rating and Win Shares Per 48 Minutes* on the following page