

Andrew Bauer

DATA 6560 – Sports Analytics

17 Nov 2025

## Data Quality and Source Validation

### The Value of 3pt Shooting as Players Age

It is important to ensure that all data for this project is valid and sourced in a consistent and unbiased manner. Thus, I have chosen basketball-reference.com as my sole source for this data. Basketball reference provides raw, numerical data for quantitatively statistics in a consistently organized table. This data is updated daily and is sourced by industry leading company, Sportradar. The data encompasses all NBA players that played in the 2024-2025 season. There are no coverage gaps, and only missing values when a player did not play enough time to record statistics in said category. However, these players will be omitted from our calculations as their data can skew our relationships.

While there is little bias in the way the data is collected, there is some bias in the relationships themselves. Players who do not receive many minutes will not generate as much 3pt data or win shares, leading to an over-representation of poor 3pt performance leading to poor win share. To mitigate this bias, we will have to carefully filter out players who receive very little playing time and make “meaningless” contributions.

After compiling some of the data and exploring basic trends between 3pt shooting and win shares, it is evident that there are some extreme outliers. For example, when evaluating win shares vs 3 point makes per game, there is a very loose linear trend, with all players having between 0 and 12 win shares, except for 2 individuals (the two MVP candidates) with around 16. Anomalies such as this make it difficult to establish trends, however, because our data set is comprised of roughly 500 players, eliminating two data points will still give us an accurate picture, thus these points may be omitted in the future. Another anomaly that will have to be addressed is the Lebron factor. When evaluating average win shares by age, win shares increase with age until 29, and then taper off slowly. However, at age 40, there is Lebron, who completely skews the data because of his high win share and the small amount of players playing at age 40. This anomaly will be ignored in the data because the goal of the project is to focus on how to value players that produce win shares based on 3pt shooting, outside of their declining athleticism. Once this data is cleaned up, we will start to dive deeper into assessing the value of the 3pt shot by assessing which 3pt attributes correlate the most to win shares through the use of regression modeling.