Professional Basketball Shot Log Data Assessment
Summary Report

April 30, 2018

Ye-in Jeon
Sam Kupfer
Amy Mykityshyn
Krystin Sinclair
Vishnupriya Venkateswaran

Background

Sports data has long been collected and analyzed for strategy and player evaluation. For example, the outcome of every baseball players' at-bats are recorded, where the ball was hit, who fielded it, and the outcome of each play (out, walk, made to base, etc.). Recent advancements in technology have made data collection for faster-paced team sports, like basketball, easier and now video recorders and computers are used to log every detail of even these rapid performance metrics. Teams have successfully used this data for strategy development and player evaluation (Cohen, 2016). The media uses it for producing insights that can be presented by analysts and writers; and gamblers or fantasy league players use it to gain an advantage in predicting player and team performance.

Question of Interest

In basketball, winning is based on a team's relative success in scoring points against their opponent. The coaching staff develops plays and team strategies that improve their team's chance of success. A core aspect of these strategies is how to construct their plays and how to instruct their players to take the most effective shots. Our goal was to develop a model that demonstrates what factors are important to shot success and see if we could use that model to accurately predict whether a given shot would go in the basket.

This question meets the SMART criteria for research question development: it is **S**pecific, in that it seeks an answer that is narrowly focused; it is **M**easurable, the data we found are comprehensive measures of basketball shots for an entire season; it is **A**chievable, the dataset was able to be analyzed in R Studio with desktop computing resources; the question is **R**elevant to professional basketball players, recruiters, and fans who want to understand player performance; and it is **T**ime oriented, we believe that this question is straightforward enough to be answered in the given time and the data is already collected.

Data Overview

We followed the epicycles of analysis process for our research, so after developing our research question, we searched out data that could help us answer the question. The professional National Basketball Association (NBA) has records of shot logs on its web site (National Basketball Association). We were able to access, via Kaggle, the NBA shot log that contained every shot taken in the 2014-2015 season (128,069 shots) for our analysis.

The shot_log dataset consisted of 21 variables and 128,069 shot data points. The variables about the shot taken were descriptive: Game ID, matchup (teams playing), location (home or away), player name, and name of closest defender. We have outcome variables: which team won and the final point differential. We were most interested in data pertaining to the specific shot being recorded. Those variables were: game clock (time left in the quarter), shot number, period (quarter), shot clock, number of dribbles, touch time, shot distance, point type (2 or 3 points), shot result (made or missed), and closest defender distance. Some of the variables just represented the same information in a different manner (shot_result: made or missed; field goal made: 0 or 1).

Missing Data

Data for the variable “shot_clock” was missing in some instances. This makes sense, based on what we know about basketball. In professional basketball, the team on offense has 24 seconds to take a shot once they take possession of the basketball. The shot clock is a timer designed to increase the pace and scoring of the game. For shots when the game clock had less than 24 seconds in any quarter, there was no shot clock because the game clock would run out first. However, we were interested in including data when the shot clock was less than 24 seconds because teams still had a time limit in effect to score points at the end of each quarter. Therefore, we filled in the missing shot clock values with the time left in the period for those cells where there was no shot clock data listed.

This dataset also does not provide data concerning free-throw accuracy. Free throws are penalty shots that players get to take. They are worth one point and are taken from a set location and are not under time pressure. While we could determine how many points each team scored in a game from free throw (single point) shots from our data, the data set does not provide free throw accuracy for individual players.

Exploratory Data Analysis

Through the use of Str(), summary(), and sd() in R, we performed initial data exploration. This showed that shot distance, a continuous variable, at the third quartile is 22 feet. In basketball shots approximately 24 feet from the basket and closer are worth 2 points and are worth 3 points from further away. Therefore about three quarter of the shots are worth 2 points in this season. Point type, a categorical variable, showed that 74% of the shots were in fact 2 pointers. Field goal made is another categorical variable. This one represents the success of the shot, whether or not it went in. This is our variable of interest. The summary function in R showed that 45% of the shots were successful and the field goal data is summarized in Table 1.

Table1. Field Goals

Field Goals	Made %	Missed %	Shot Type %
2-pointers	49%	51%	74%
3-pointers	35%	65%	26%
All Field Goals	45%	55%	128,069

Among 128,069 shots taken, 45% of them were made and 55% of them were missed. Also, 74% of them were 2 pointers and 26% of them were 3 pointers. As you can see in Table 1, 49% of 2 pointers were made and 35% of 3 pointers were missed.

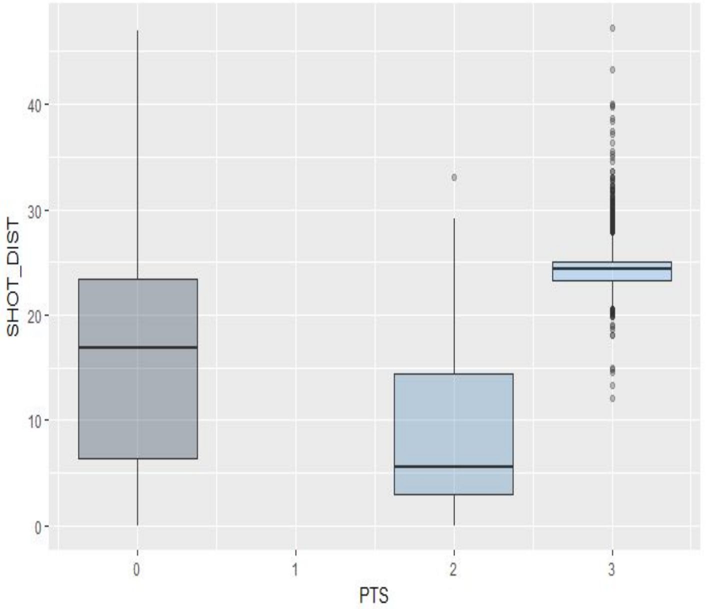


Figure 1. Points Scored by Shot Distance

The next step in data exploration was the use of table, histograms, box plots, and correlation, which were performed in R. Figure 1 shows the distribution of the shot distance by points. This showed that about half of all 2 point shots are successful whereas only 35% of all 3 point shots are successful. 3 point shots are typically considered to be more difficult because the ball has to travel farther.

The distributions shown in Figure 2 describe the field goals that were attempted by all players during the 2014-2015 professional basketball season. The first chart shows the number of shots taken by distance (in feet) from the basket. It shows that most shots that were 2 point shots, taken less than about 24 feet away. It also shows a steep drop off after 24 feet. The majority of 3 point shots are taken right around the 24 foot mark, which is about the average distance from the basket of the 3 point line, which is marked on the court. The second graph shows the probability of success of the shot at each distance. This has a downward trend, the further from the basket, the less likely the shot will be made. Between 0 and 2 feet, the shot accuracy is high, which is expected because these are shots taken very close to the basket. There is an increase at 46 feet, however this is an outlier. There are very few shots taken at this distance, making the probability more variable.

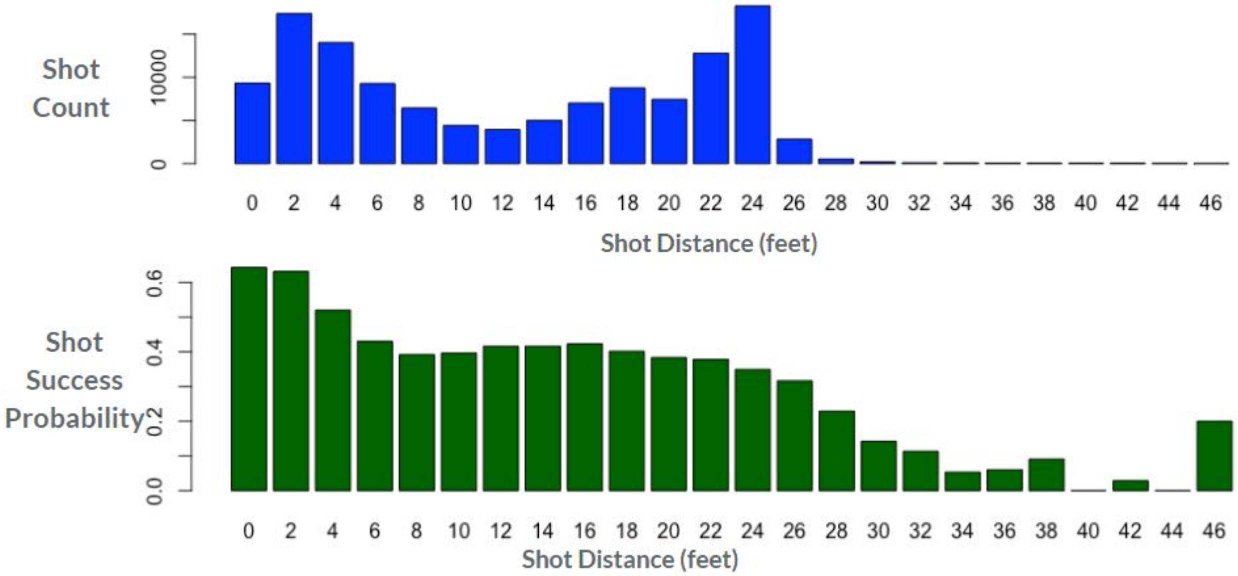


Figure 2. Number of Shots Taken and Shot Success by Shot Distance

The correlation matrix did not prove to be very insightful. The only highly correlated variables were those with redundancy. For example, touch time and dribble. While it is possible to hold the ball and therefore have a longer touch time without increasing dribbles, it would be quite difficult to do the opposite. If a player is dribbling many times, then the amount of seconds that the player has the ball would logically be greater. Points and field goals made also show high correlation. Field goal made is the success of the shot, a binary variable. Points is how many points the team received for each attempted shot, zero, (if the shot was missed), two, or three. Obviously, the points variable already includes information about whether a shot was made or missed. One of the higher correlations that may be interesting is closest defender distance and shot distance. The closer a player is to the basket the closer the opposing team is to the player. When watching a game this is usually quite apparent, and the high correlation is our numerical evidence. Unfortunately, field goal made (the variable of interest) did not show high correlation with many of the other variables. Shot distance and point type are the two variables with the strongest correlation to field goal made. Logically, this does make sense. If the ball has to travel further, it would be more difficult to make the shot. Point type is directly related to shot distance, and the correlations shows this.

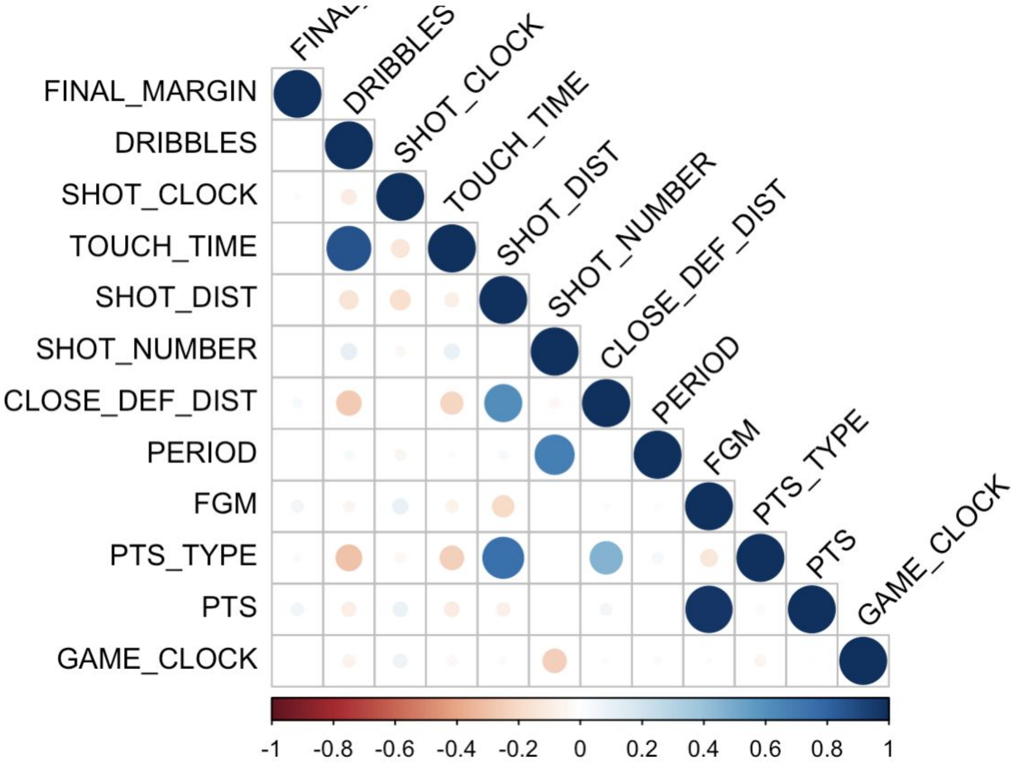


Figure 3. Correlation Values for Variables of Interest

Building a Model

Our question of interest is a traditional classification problem: “How accurately can we predict whether a particular shot will go in or not?” We have shot data that tells whether a shot was made (or not) as well as many other potentially contributing variables. We focused on the variables that pertain to the specific shot. We chose to exclude from the model variables that depict the outcome, for example whether or not the team won the game, because these outcome variables would have been unknown at the time at which the shot was taken. We also excluded descriptive variables, such as game ID, player ID and player name. Which player is taking the shot could have a strong relationship with shot success, but it is a nominal/categorical variable with too many categories to be useful in our model. After the outcome and descriptive variables were excluded, a Cp analysis was used to find which of the variables could potentially be used to make the best model. The analysis showed that only game clock should be dropped. All of the other variables together contributed to a model that had the lowest Cp.

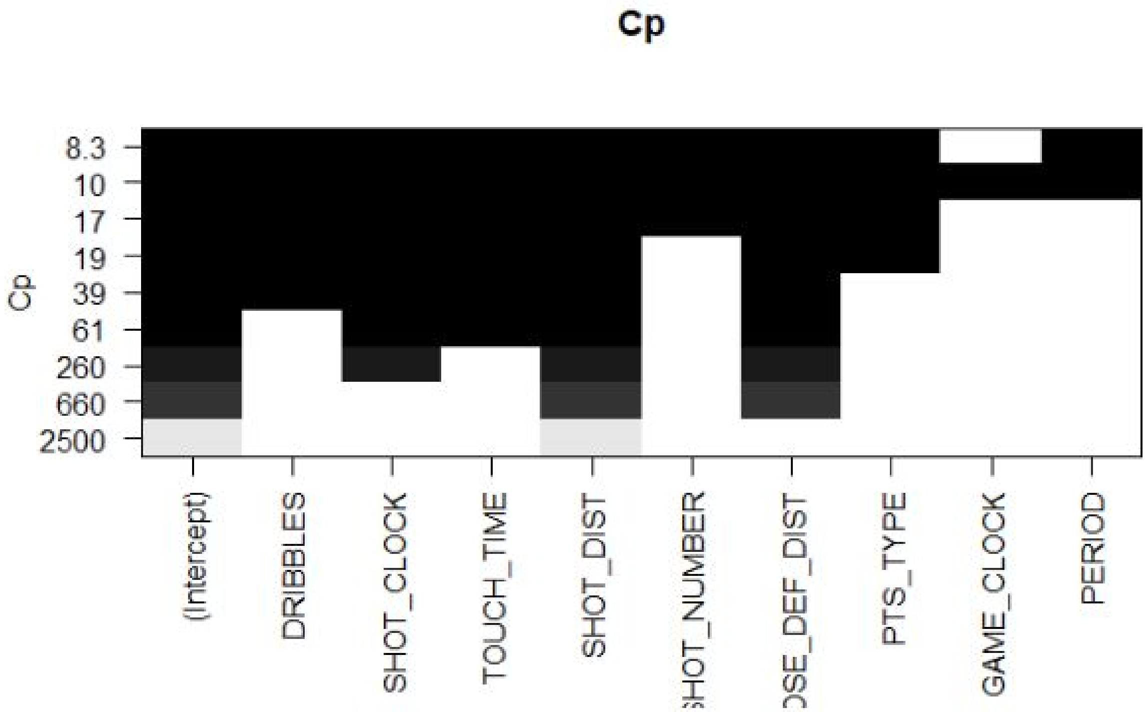
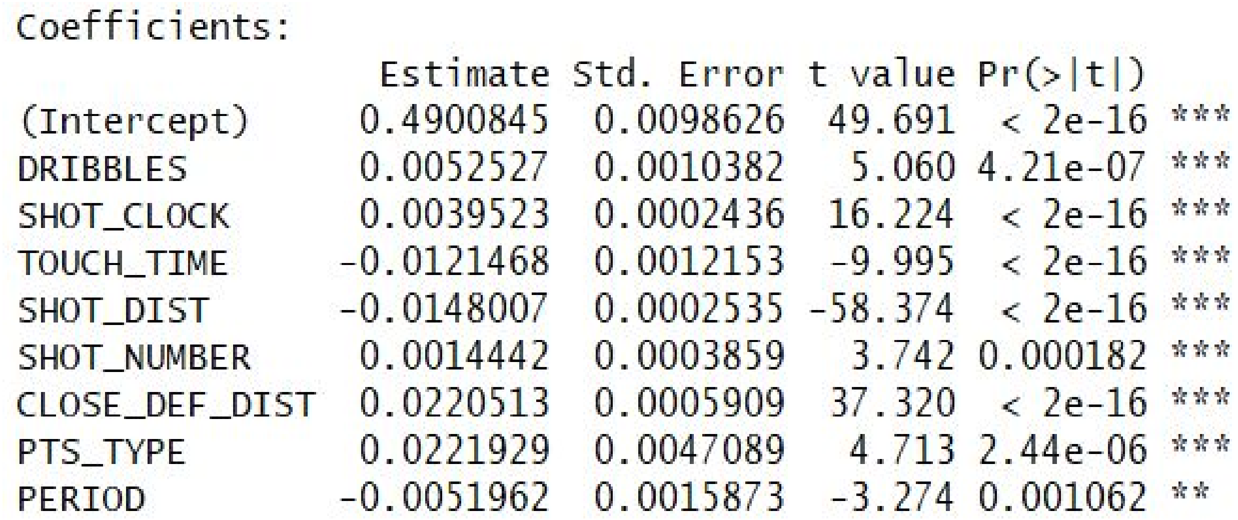


Figure 4. Model Building Exercise as a Function of Cp

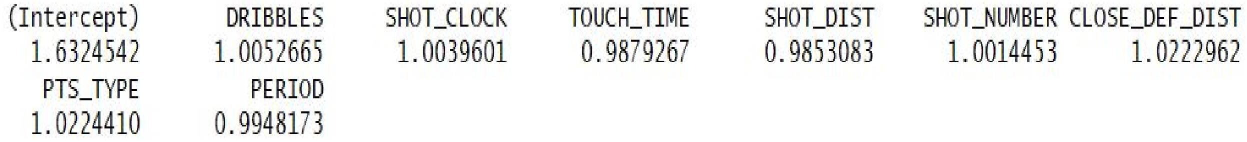
In order to answer the question, a predictive model showcasing the likelihood of shot success would be best practice. The dependent variable is binary, making a logistic regression the best choice. A logistic regression model was fit with “field goal made” as the dependent variable and the independent variables shown to make the best model from the Cp analysis. The output from the logistic regression model is shown below.



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.4900845  0.0098626  49.691  < 2e-16 ***
DRIBBLES        0.0052527  0.0010382   5.060 4.21e-07 ***
SHOT_CLOCK      0.0039523  0.0002436  16.224  < 2e-16 ***
TOUCH_TIME     -0.0121468  0.0012153  -9.995  < 2e-16 ***
SHOT_DIST      -0.0148007  0.0002535 -58.374  < 2e-16 ***
SHOT_NUMBER     0.0014442  0.0003859   3.742 0.000182 ***
CLOSE_DEF_DIST  0.0220513  0.0005909  37.320  < 2e-16 ***
PTS_TYPE        0.0221929  0.0047089   4.713 2.44e-06 ***
PERIOD         -0.0051962  0.0015873  -3.274 0.001062 ** 
```

Figure 5. Logistic Regression Model Coefficients

All of these coefficients are showing very low p-values. This means that the coefficients are significantly different from zero, which means that all of the variables are useful in this model. Intuitively, the signs of the coefficients make sense. For instance, shot distance has a negative coefficient, meaning that as a player moves closer to the basket, the likelihood of shot success would increase. Shot clock has a positive coefficient. The lower the number on the shot clock, the less time a player has. The player may feel pressure increasing in their situation and may rush their shot, which could make it harder to have a successful shot.



```
   (Intercept)       DRIBBLES     SHOT_CLOCK     TOUCH_TIME      SHOT_DIST    SHOT_NUMBER CLOSE_DEF_DIST 
     1.6324542      1.0052665      1.0039601      0.9879267      0.9853083      1.0014453      1.0222962 
      PTS_TYPE         PERIOD 
     1.0224410      0.9948173 
```

Figure 6. Odds Ratios for Model Predictor Variables

These exponential coefficients, shown in Figure 6 will help us to answer predictive questions about shot success. One question that we wanted to answer was how much does a one foot decrease in shot distance improve the likelihood of shot success. The odds ratio for shot distance is .985. A player is 1.5% more likely to make a shot for every foot closer to the basket they are, when taking the shot. Another question is how much does one foot increase in closest defender distance improve the likelihood of shot success. The odds ratio for closest defender is 1.022. A player is 2.2% more likely to make a shot for every foot farther away the closest defender is to the player taking the shot. A third question that can be answered with these odds ratios is: how much does one second increase in shot clock improve the likelihood of shot success? The odds ratio for shot clock is 1.00396. A player is .396% more likely to make shot for every extra second on the shot clock.

Model Reliability

The Hosmer and Lemeshow Goodness of Fit test and the area under the curve of receiver operating characteristic (ROC) were used to test the overall reliability of the model. Although we were able to build a model that had eight predictor variables that were significant contributors, we found that the model itself is not a very good predictor of shot success. The Hosmer and Lemeshow Test had a very small p-value, showing a significant difference between the model and the observed data. The area under the curve had a value of .6313, which indicates a poor model - .5 would mean that our model was no better than a "random" model, .8 is generally considered to be a good model. The sensitivity (true positives) and specificity (true negatives) plot is shown in Figure 7.

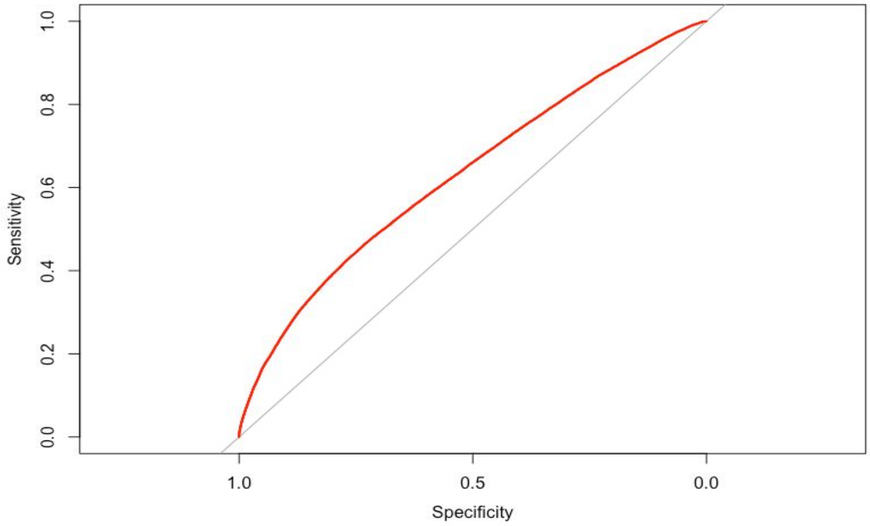


Figure 7. Model Specificity vs. Sensitivity

Conclusions and Future Work

The model we developed through the use of logistic regression did not prove to be a good predictor of field goal success for 2014-2015 professional basketball players. Despite having over 128,000 shot records, the variability in the data likely contributed to this result. It may be possible to build a better model with more data points, taken over more professional basketball seasons. There is too much variability in our data, such that a general model that includes most of the predictor variables for shot success does not really explain much of the variance in the outcome. Another way to improve the predictive performance of our model is to add better predictor variables. Perhaps more detailed player characteristics, like height or years of experience, would improve our model's predictive performance.

Or, it may be useful to reduce the between-player variability by looking instead at one individual's performance over several seasons. If a reliable model can be developed for a single player, models for other players could be developed and these models could be compared to see if there is convergence of the predictor variables that make the largest contributions to predicting probability of success for field goals made. If we are able to get to that point, a useful next step might be to test our composite model by comparing its results for shots taken by teams (groups of players) that eventually won the games versus teams that lost, i.e., did the teams win that took better shots, based on our predictive model?

One thing to keep in mind is that we cannot build a comprehensive model to assess an individual player's overall basketball skills or performance from this data because the data set doesn't include free throw accuracy, which is an important skill for a professional basketball player and also contributes to a team's overall points in a game.

References

National Basketball Association. NBA Advanced Stats. April 4, 2018 from https://stats.nba.com/player/201956/shots-dash/

Cohen, B. (2016). The Golden State Warriors Have Revolutionized Basketball. April 6, 2016. https://www.wsj.com/articles/the-golden-state-warriors-have-revolutionized-basketball-1459956975