What makes a successful professional basketball player?*

How player performance influences offensive success in the NBA

Nikhil Bhambra (1004194702/bhambra3)

April 20th, 2022

Contents

*This analysis repository is available at https://github.com/bhambra3/final_project/. Word Count: 1661 < 1725

1 Introduction

Basketball is a high intensity, low stoppage-time team sport wherein five players attempt to control a bouncing ball and shoot it into the opponent teams net, which is a bucket hoisted about ten feet in the air. While simple in concept, success requires players to commit themselves both physically and mentally to out maneuver other players, co-ordinate with teammates, shoot with precision often from great distances, all while running short sprints for a ninety minute interval. At an elite level, the limited on-court presence incentives teams maintain large player rosters and prepare to play with multiple combinations of players, as well as sourcing and training elite athletes to be the best in their role. This divide and conquer strategy is employed by all thirty teams in the National Basketball Association (NBA), America's premiere professional basketball league, in their quest for the Larry O'Brien Championship Trophy.

Teams obsessively over-optimize in their quest for greatness, which includes rigorous training and exercise, developing their strategic thinking, and spending time on team cohesion drills. While development is certainly an important focus, these systems support already-elite athletes trying to become the best in the league, and as such player selection can be seen as a more important aspect of building a successful team. In the past, this has manifested as a preference for very tall players for shooting prowess, and in higher average weight for defensive power and shooting strength (Curcic, 2017) (see Appendix A for this trend over time). This over-optimization leads to a unique competitive environment, one wherein competition is so high and playing conditions are controlled so closely that each game can be thought of as representative of each players true skill in an ideal competitive environment. However, there are many non-physical aspects of basketball which contribute to team success, and it is unclear if the historical preference for heavier and taller players has a direct correlation on success within the league.

This report will seek to answer the title question, *What makes a successful professional basketball player?* by analyzing a dataset of NBA performance and physicality metrics compiled throughout the 2021-2022 regular season as of February 2021. The aim of exploration is to see if a player's offensive rating can be modelled using individual performance metrics, to hopefully model this computed statistic in the absence of data. It's important to note that *success* in professional play comes in distinct forms. For example, high scoring, possessions, and assists are considered metrics of a strong offensive player, where as a defensive player may prioritize stealing, blocking, and stopping oncoming offensive players. In fact, statistician and assistant coach for the Washington Warriors Dean Oliver developed two metrics which he believed to most faithfully represent these disparate types player skill: *Offensive Rating (O-RTG)* and *Defensive Rating (D-RTG)* (Basketball Reference, 2021) (see Appendix B for complete formulas). **I will use Offensive Rating (O-RTG) as a potential response to see if any player metrics have a significant influence on the rating.**

To pursue this goal, I will build a linear regression model to predict the success metric above within the context of an NBA player dataset, and will discuss any considerations or adjustments which were made to the data. I will detail the process of selecting and optimizing the set of predictors in the model, and will defend my choices with numerical and visual summaries. Finally I will discuss the usefulness, limitations, and interpret the model within the context of the NBA.

2 Methods

We will fit a linear regression model using the R Statistical Language (R Core Team, 2022) on a dataset of 551 NBA players active in the 2021-2022 season, collected online from the official NBA Statistics website (NBA Statistics Team, 2022) and NBAStuffer.com (NBAStuffer Data Team, 2022). The analysis dataset is comprised of 27 performance statistics such as minutes played or rebound percentage, and includes basic info like age, height, and weight. Please see Appendix C for the complete data dictionary. Offensive Rating (O-RTG) will be used to judge player success. By examining multiple potential models for this measure of success, we will accomodate for varying priorities and skills among offensive players, and will consider various areas of play in the context of our model.

A linear regression model simply estimates the variable of interest (in this case, the players performance) as a sum of potential predictors multiplied by a constant, known as a linear combination. If we define *n=551* as our sample size and p as the number of predictors, then the model can be written as:

$$\hat{Y} = \beta_0 + \sum_{i=1}^{551} \beta_i x_i + \sum_{i=1}^{551} \beta_i x_i x_k + \epsilon$$

- $\hat{Y}$ represents the response value (either Offensive Rating or Defensive Rating).
- $\hat{\beta_0}$ represents the intercept of the linear regression line.
- $\sum_{i=1}^{n} \beta_i x_i$ represents the number of predictors and their coefficients, $i \in \{1, ..., n\}$.
- $\sum_{i=1}^{n} \beta_i x_i x_k$ represents the interaction between predictors for some other predictor $k \in \{1, ..., n\}$.
- ϵ is the error term and accommodates variance not otherwise modeled with our linear relationship.

To select the best performing model, we will employ backwards model reduction using the AIC metric to optimize our potential predictor set in a step wise fashion. By continuously removing and verifying the model fit, we can arrive at an optimal predictor set. In the context of this analysis, I am seeking a simple model with a lower number of potential predictors as I value interpretability and applicability to future inference for my model over explicit predictive accuracy, and I believe using less features will avoid over fitting due to the increased variance.

2.1 Checking Assumptions and Conditions

To ensure our model is appropriate for inference, we must first verify that our data is appropriate to fit a linear regression model on. We will check for violations to the four assumptions below by examining the shape and trend of our proposed predictors residual plots. We will verify:

1. That all attributes are normally distributed across players (Normality)
2. That all player data is independent/uncorrelated with each other (Independence)
3. That each predictor x_i used in the final model must be linearly associated with the response $\hat{Y}$. This can be satisfied by a strong linear association in the plot of the predictor and the response, as well as a random scatter in the residuals of our predictors x_i, and a random scatter in the residuals versus fitted values plot (Linearity)
4. The residuals of the predictors should be random scatters, without any apparent fanning or parabolic patterns (Homoscedasticity)

Then, after verifying that a linear model is appropriate for the data, we need to further verify that the model successfully models the relationship being proposed, that being the set of predictors is a salient linear relationship with respect to the response. We will verify:

5. That the response and fitted model values should either be linear, or completely random with no obvious clustering patterns (Condition 1)
6. That the proposed predictor set has no multi-colinearity, which can be verified by checking the pairwise plots for the predictor set. If there are any distinct relationships between any of our predictors, then perhaps we need to further simplify our predictor set (Condition 2)

2.2 Diagnostics and Handling Violations

If there are violations to normality or linearity, we can attempt to correct these errors using a Box-Cox transformation. This adjustment allows us to use data which is not applicable for a linear regression analysis by reshaping the non-normal explanatory variables into a normal shape. It involves taking the log(response), or raising it to some power, and the specific values of λ vary for each x_i. Below is the example for taking a box-cox transformation for a generic predictor:

$$\psi(x_i, \lambda) = \begin{cases} x_i^{\lambda} & \lambda \neq 0 \\ log(x_i) & \lambda = 0 \end{cases}$$

However, power transformations do not fundamentally alter the differences between data points, but rather the centre and scaling of the whole distribution of data points, and as such, power transformations introduce an aspect of variance into the model. Further, because the data is no longer representative of true observations in the dataset, using power transformations limits the applicability of the final model for inference. As such, we will limit use of these transformations unless required.

If there is a violation to homoscedasticity, it can potentially be handled using variance stabilizing transformation (VST). Simply, a VST is a function applied to the potential response variable to help control for non-normal behavior and variance. This can take the form of setting the response Y to $sqrt(Y)$ or $log(Y)$.

Generally, we want to validate Conditions 1 and 2 before verifying the four assumptions are satisfied because a condition violation tells us our model is completely inappropriate for the dataset. After ensuring the model is sufficiently linear and does not exhibit multi-colinearity, then we should check for Assumptions 1-4 and apply any adjustments as necessary. At this stage, if any of the assumptions fail and can not be corrected, the model will fail to be useful. However if none of the assumptions are violated, then Condition 1 is likely satisfied. Any violations to Condition 2 will require us to simply the predictor set, or to discuss the influence of multi-colinearily on the final model.

It would be helpful to validate how useful our model is on unseen data, and to do this we could split our dataset into training and testing subsets. However given the small size of the dataset, as well as the highly correlated events which drive this data (professional players play against professional players), it was decided to omit this aspect of validation from this analysis to avoid overfitting and modelling relationships which do not truly exist.

Finally, to assess the usefulness of our model, we will use the Adjusted Coefficient of Determination r^2_{adj} across models on the same response to judge which is "more" successful at modelling the proposed relationship. This metric is useful as it allows us to compare models with different numbers of predictors, and it prioritizes models which have fewer parameters. It can be written as:

$$r^2_{adj} = 1 - \frac{\left(\frac{RSS}{n-p-1}\right)}{\left(\frac{SST}{n-1}\right)}$$

- *RSS* represents the Residual Sum of Squares, the total amount of unexplained variance in our model.
- *n-p-1* represents an adjustment ratio which considers the number of parameters p in the model.
- *SS* represents the Sum of Squares, the total variance present in our sample data.
- *n-1* represents an adjustment ratio needed to correct for an under-estimated variance.

3 Results

3.1 Examining Model Variables

To begin, each of the twenty-seven continuous numeric variables were examined for roughly normal properties when graphed with respect to our success criteria: ORTG. To see an example of the linear relationship, please see Appendix D. After testing, the following predictors were selected as potentially appropriate to build a model with. Any rows with an asterisk in the final column indicates that the a transformation may be required before modeling this as a linear relationship. Below the table is the potential first linear model before attempting to optimize using AIC and backwards selection.

Figure 1: The initial subset of potential predictors (n=551)

Column Name	Meaning	Requires transformation? (*=YES)
Full Name (uID)	First and Last Name	Neither
Height	Height of the player in inches	O
Weight	Weight of the player in pounds	O
TO	Turnover rate	O
TRB	Total Rebound Percentage	O
X2PA	2-point shot attempts	O*
BPG	Blocks per game	O*
TOPG	Turnovers per game	O*
MPG	Minutes per game	O*
FTA	Free-throws attempted	O*
VI	Versatility Rating	O*
USG	Usage percentage	O
X2P	2-point percentage	O
X3P	3-point percentage	O
eFG	Effective Shooting Percentage	O
TS	True Shooting Percentage	O
PPG	Points Per Game	O*
RPG	Rebounds Per Game	O

- **O-RTG** ~ Height + Weight + TO + TRB + X2PA + BPG + TOPG + MPG + FTA + VI + USG + X2P. + X3P. + eFG + TS + PPG + RPG (17 potential applicable attributes)

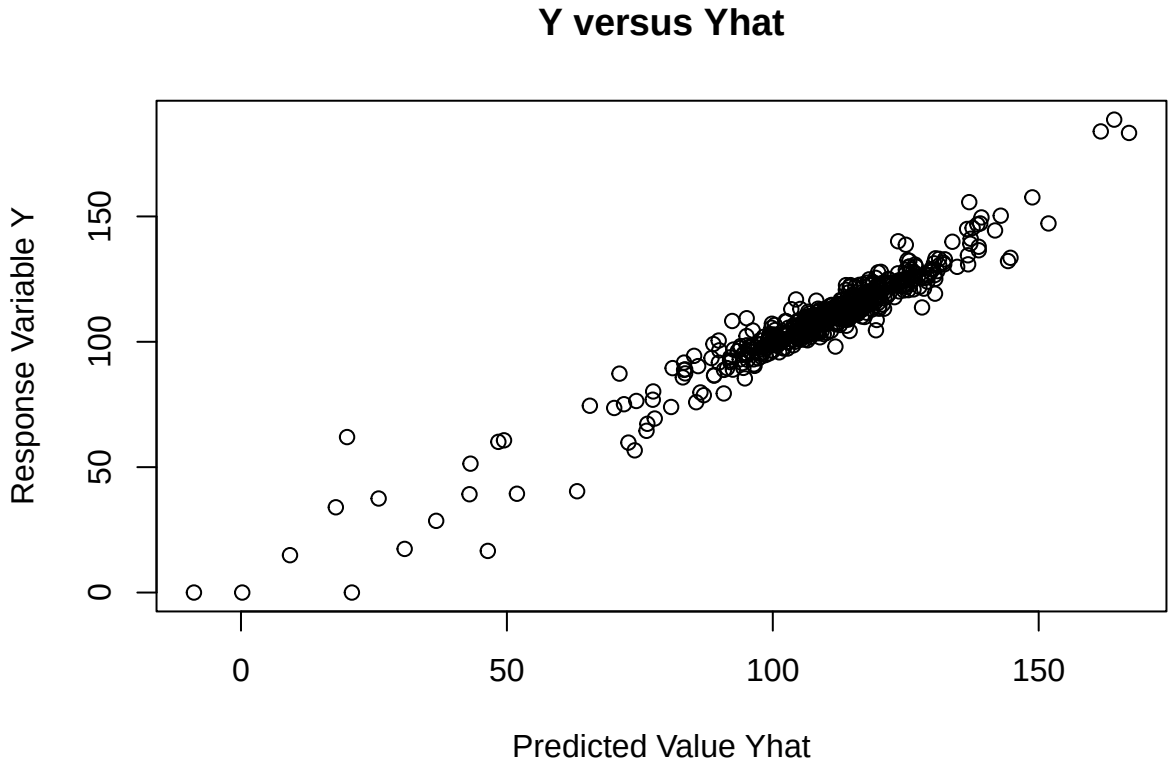
3.2 Model Selection Process

After running partial f-tests and comparing the adjusted r^2_{adj} value between reductions, we arrived at a final model which was characteristic of Offensive Rating. It contains 13 distinct predictors and has an adjusted $r^2_{adj} = 0.9219$. While this is good, we need to verify our assumptions still hold after defining our linear model. The final model is comprised of features from the list above, but we remove the following columns based on partial f-tests: X2PA, TOGP, X2P., and PPG.

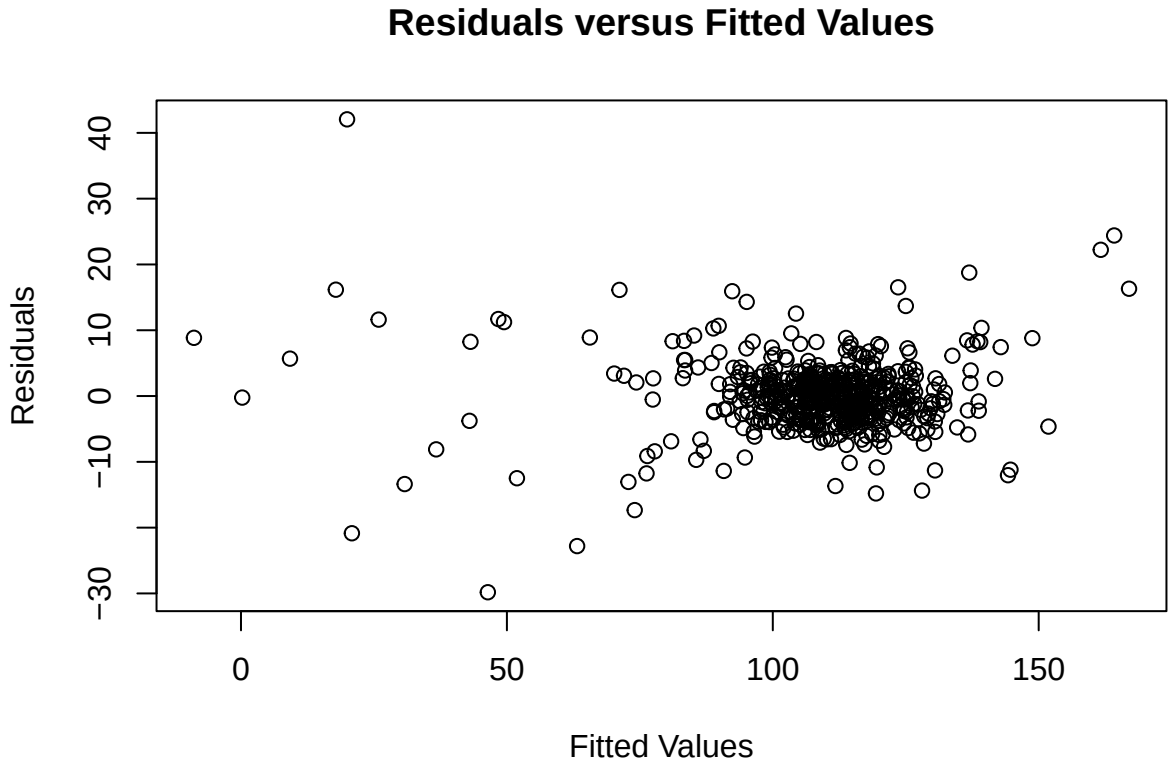
3.3 Goodness of Fit Considerations

To ensure our model is still valid after removing these terms, we must now verify our assumptions and conditions. Pairwise plots, applying transformations, and checking our conditions are not presented in this report.

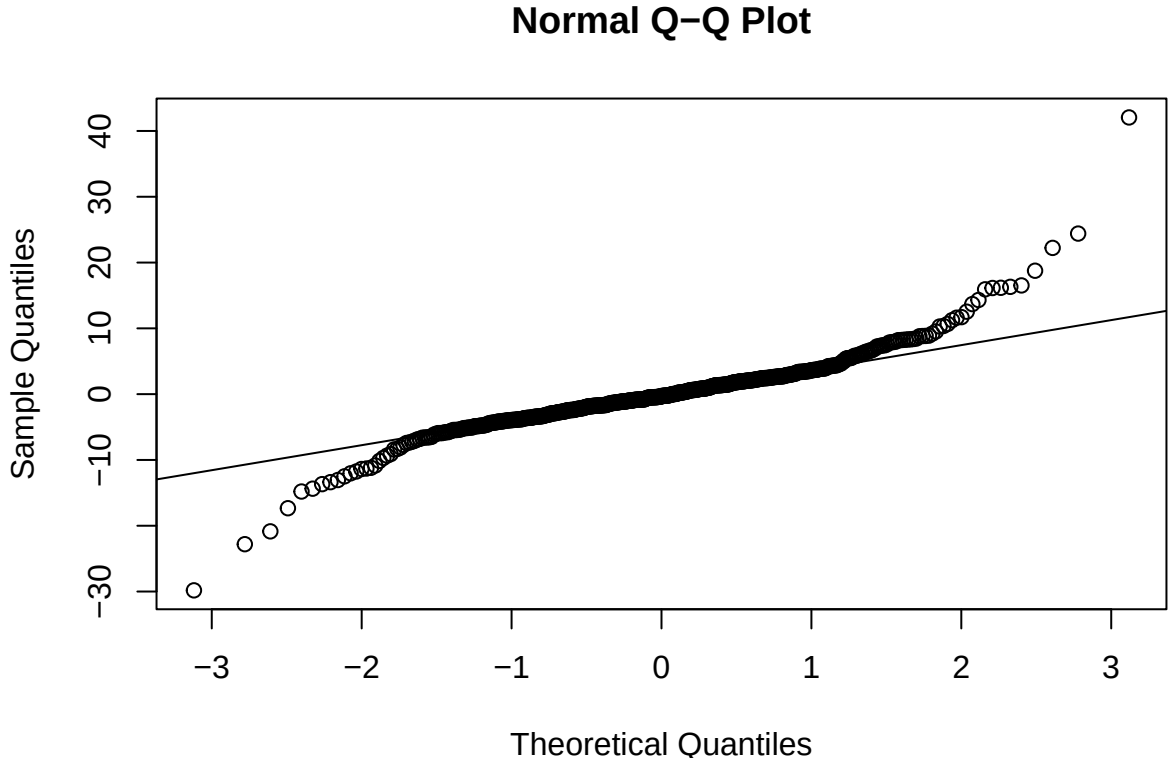
In the plot of our Response versus Model Values, we can see the characteristic linear pattern indicating that our model is indeed an appropriate way to describe Offensive Rating. There is high variance towards the extremes which limits usefulness of the model for the future.



In the Residuals versus Fitted Values plot, we see no discernible pattern which is ideal and should indicate uncorrelated errors, even though there is clustering around a centre and high variance towards the edges of the dataset. Given the context of the competitive landscape, it should not be surprising that extreme values are not well modeled.



To understand how our model compares to the distribution of our response, we can investigate the QQ-plot for the residuals of our ORTG model. As we can see from the above QQ plot, our model has a strong fit to the sample quantiles around the centre, but as value become more extreme they taper out in a chaotic fashion, which indicates very high variance for these extremes. This makes sense in the context of our model, as there are strong outliers in O-RTG which leads to this poor modelling.



4 Discussion

The adjusted r^2_{adj} for our Offensive model is 0.9219 on 13 attributes. While this is a close fit for middling values of O-RTG, this closeness is misleading for extreme values of O-RTG, which our model predicts poorly. This will cause future inference using this model to be noisy and perhaps unusable for values which fall outside the IQR.

4.1 Limitations of this Model

First and foremost, the metric O-RTG which we are attempting to model as a response variable is a computed statistic which was included in my initial dataset. As such, this metric could be overly idealized (rounded) because no information on data treatment was supplied. It is possible that O-RTG could be reported slightly differently because of the nature of computing statistics. Further, as this criteria is made up of seven other metrics (see Appendix B for full formula), it is trivial to see how these models can be modeled as linear combinations of other data. Not only does this mean the computed statistics are applicable and relevant to a linear investigation, but that using the metrics from the calculation as predictors should yield a theoretically strong linear relationship.

Next, the data used to create the analysis dataset(n=551,p=31) was manually created from two data sources, NBA.com (NBA Statistics Team, 2022) and NBAStuffer.com (NBAStuffer.com Data Team, 2022). The latter website is not an official data source for league play, and there was no published information about how the data was treated. As such the data itself could be viewed as a limitation because it is unclear how precise it is. Any players who did not have recorded statistics for at least one column were removed.

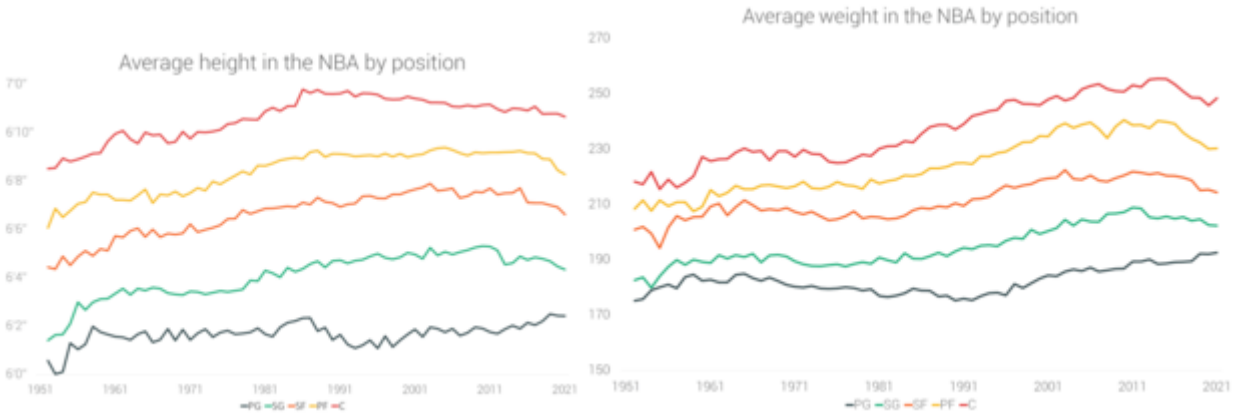
Next, using backwards AIC to simplify the model only explores one potential option for models of the response, and as such it may be useful to use randomization, brute-force, or investigation-driven selection to improve the model further. This analysis also fails to include interaction terms in the model and as such, it presents an over idealized model which will not handle variance caused by multicolinearity for future inference.

Because the model was not validated using training and testing data partitions, it is unclear how applicable this data will be for future observations. However not splitting the available dataset into these partitions allows our results to be more robust as they model relationships for the entire league, which is more useful in this context than a model derived from only a piece of league games. Particularly, these metrics are reflective of real-life games and if we did split our dataset it is not clear that the information would be reflective of real-world competitive restraints.

In the future, VIF should be calculated to determine if the variance of the model is more than what is acceptable, and if variance stabilizing transformations should be performed.

5 Appendix

Appendix A: Historical Trend for Taller and Heavier players (Curcic, 2017).



Appendix B: Formulas for success criteria (Sports Reference, 2022).

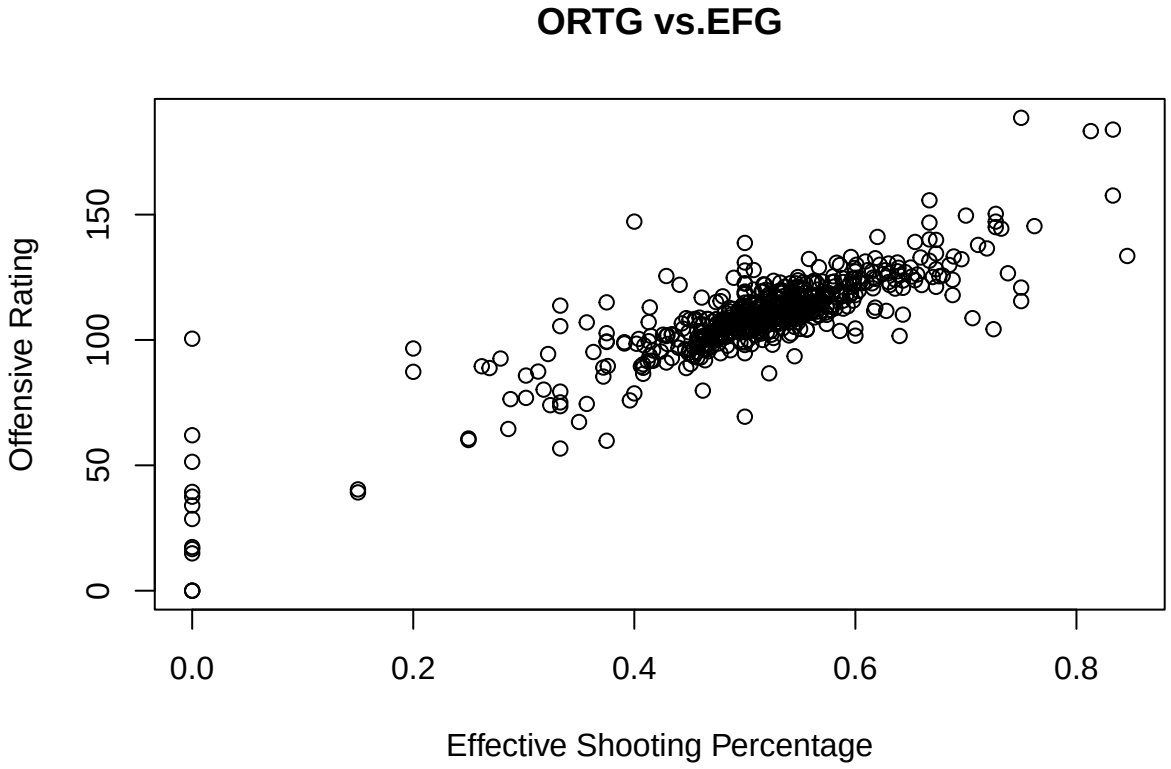
- **Offensive Rating** (for players) can be calculated as: Offensive Production Rating = (Points Produced / Individual Possessions) x OAPOW × PPG + FTM/FT * 3pt% + FG%

Appendix C: Performance metrics and their contextualized meanings

Column Name	Meaning
X	Unique Integer ID (1..551)
Full Name	First and Last Name
Height	Height of the player in inches
Weight	Weight of the player in pounds
ORTG	"Offensive Rating" Points produced per every 100 possessions
DRTG	"Defensive Rating Points allowed by the player per every 100 head-ons
Team	Current team
Position	Primary position
Age	Age in years
GP	Games played
MPG	Minutes per game
MIN	Minutes played percentage
USG	Usage percentage
TO	Turnover rate
FTA	Free-throws attempted
FT	Free-throws percentage
X2PA	2-point shot attempts
X2P	2-point percentage
X3PA	3-point shot attempts
X3P	3-point percentage
eFG	Effective Shooting Percentage
TS	True Shooting Percentage
PPG	Points Per Game
RPG	Rebounds Per Game
TRB	Total Rebound Percentage
APG	Assists per game
AST	Assist percentage

Column Name	Meaning
SPG	Steals per game
BPG	Blocks per game
TOP	Turnovers per game
VI	"Versatility Rating"

Appendix D: Predictor versus Response relationships for O-RTG



References

Curcic, Dimitrije. 2017. *67 Years of Height Evolution in the NBA - in-Depth Research.* Athletic Shoe Reviews, RunRepeat.com. runrepeat.com/height-evolution-in-the-nba.

NBA Statistics Team. Feb 18 2022. *Player Bios.* United States of America: National Basketball Association. www.nba.com/stats/players/bio/.

NBAStuffer.com Data Team. Oct 20 2021. *NBA Stats 2021/22: All Player Statistics in One Page.* NBAStuffer. www.nbastuffer.com/2021-2022-nba-player-stats/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reference, Basketball. n.d. "Calculating Individual Offensive and Defensive Ratings." *Calculating Individual Offensive and Defensive Ratings.* Sports Reference LLC. https://www.basketball-reference.com/about/ratings.html.