Predicting NBA Players' Average Points per Game Utilizing Player BMI

Authors:
Siju Abraham, Jose Alvarado, Andrew Bundy, Jeffrey Cook

6/17/2022

Introduction

Hypothesis

As a rather sports fanatical group, we wanted to focus our research to the athletic world. After scouring many datasets ranging from the MLB to NFL, we decided to focus on the NBA. It was a very relevant theme, as the 2022 NBA playoffs and finals were ongoing throughout our project. Once the appropriate dataset was selected, we analyzed its columns to propose a research question. Looking at the data, we were intrigued by the details regarding player stature, and this formed the basis for our research question. **Could a player's stature, specifically the player's body mass index (BMI), have any effect on the average amount of points scored per game by that player?**

Selection of Data

The NBA dataset our group had discovered was sourced from Kaggle*. It contains over 20 years worth of NBA statistics including data from the 2020 season. In total, the data ranges from 1996 to 2020. The dataset's 22 columns include, but are not limited to: a player's age, height, weight, and place of birth. It also includes details such as the player's team name, a player's draft year, and round pick. In addition to those stats, this dataset includes scoring details such as the average number of points scored per game, rebounds, assists, and total games played by any given player.

This dataset was originally pulled from Official NBA Stats | NBA.com and contained missing data that have since been filled. As the author of the dataset suggests, this dataset could be potentially used to analyze the geographical diversity of NBA players or how the body stature of a typical NBA player has changed over the years.

Kaggle NBA Dataset URL: https://www.kaggle.com/datasets/justinas/nba-players-data

However, for our research purposes, we needed to leverage the data to create a new column: the BMI, since this was not provided in the dataset. Interestingly, the columns included all the necessary information to calculate a player’s BMI. To calculate the BMI, we divided the player’s weight in kilograms by the player’s height in meters squared.

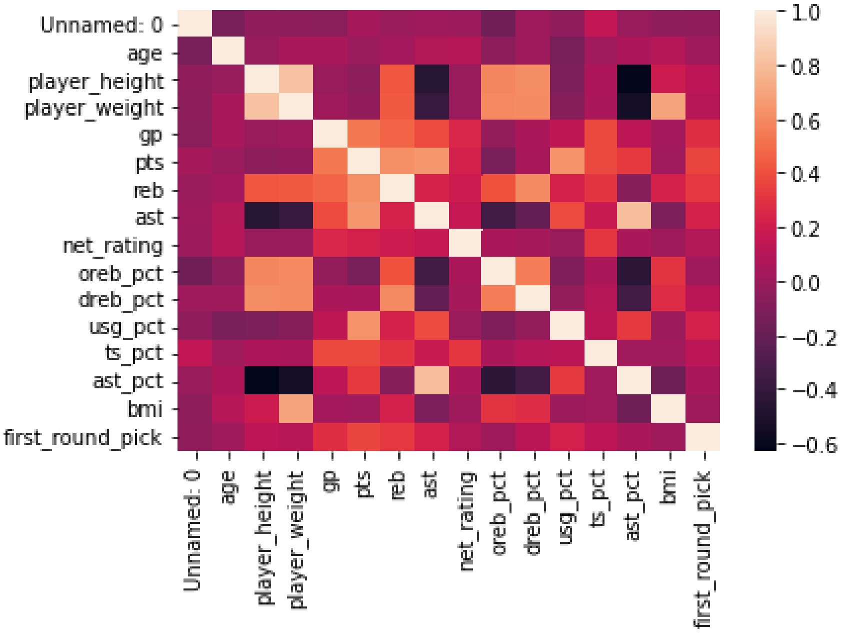
We used Scikit-learn’s StandardScaler function in order to scale the data. As was standard in our course, our train/test split we used was the following: 70% of the data was used to train the model, whilst 30% of the data was used to test the model. We used the kNN regression model with k set to the nearest 10 neighbors, as well as a linear regression model. We also employed the brute force method. Initially, this first run gave us a root mean squared error value of 6.07.

Initial Hypothesis

Methodology

We constantly referred to our dataset as it was our central source of information. To aid us in the inner logistics of the NBA draft, we referred to the draft's historical information from the NBA's official website and a Forbes article detailing NBA drafts. We added code to import our downloaded dataset from Github. This proved to be the simplest solution compared to importing it from Google Drive. Our code was written in Google Colab, allowing team members to collaborate freely, improving version control.

We used various visualization tools such as standard scatter plots. We also used a correlation heat map to detect any correlations between the various variables in the dataset, as well as tested the correlational scores between average points scored per game and other numeric variables. We also used the kNN Regressor and Linear Regression models for the BMI. Based on our initial hypothesis that BMI would impact a player's points per game, we created a model using KNN Regression while using z-score to transform the data.

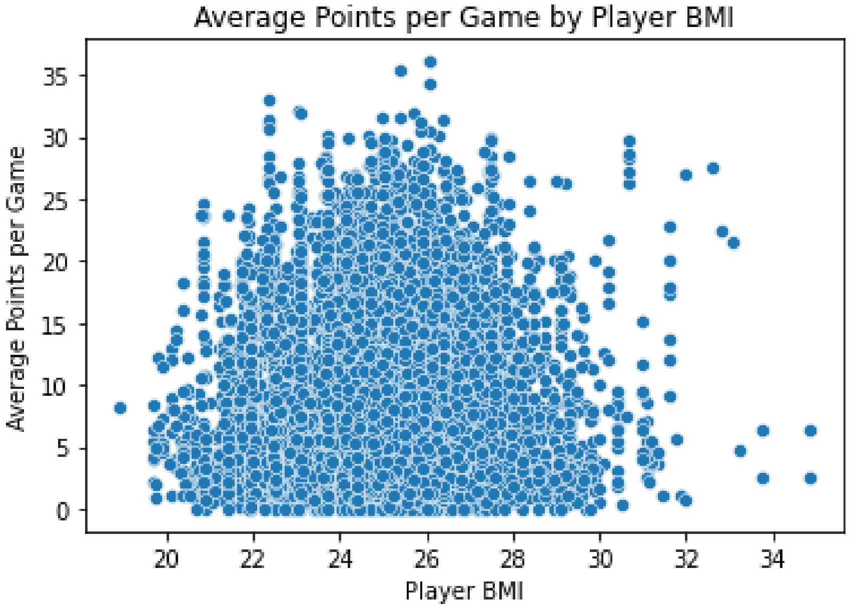


Heat map showing correlation scores between numeric columns

Additionally, we created a second model using linear regression for comparison purposes. We had to create a new model once we figured out that BMI was not a very good indicator of points scored per game. This decision was based on the correlation heatmap. We used different variables that were more highly correlated to points scored per game, in lieu of using BMI.

Results

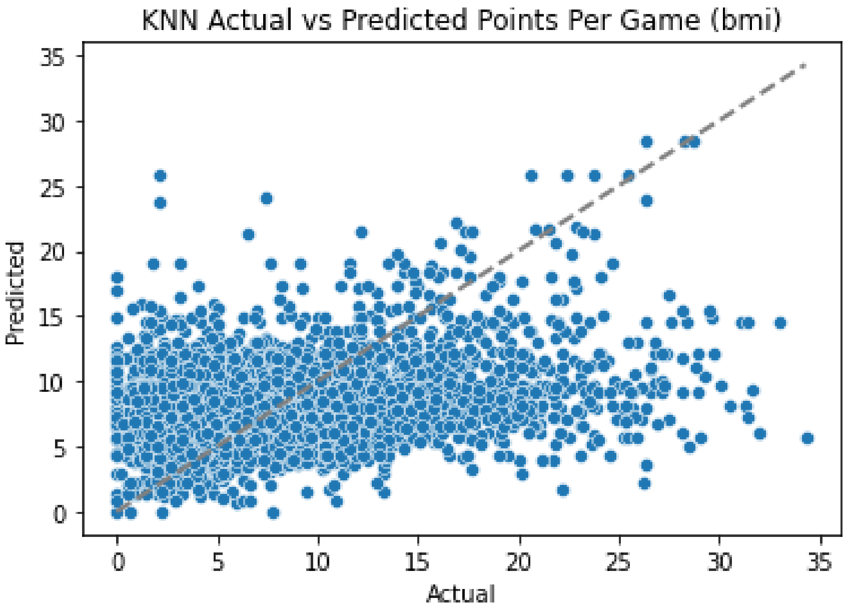
After careful analysis of the dataset, we ascertained the answer to our research question. We calculated BMI as previously mentioned and then plotted this data onto a scatter plot. For visualization, we created a scatter plot showing the Average Points per Game by Player BMI. On the x-axis, we graphed player BMI, while on the y-axis we graphed the average points per game.

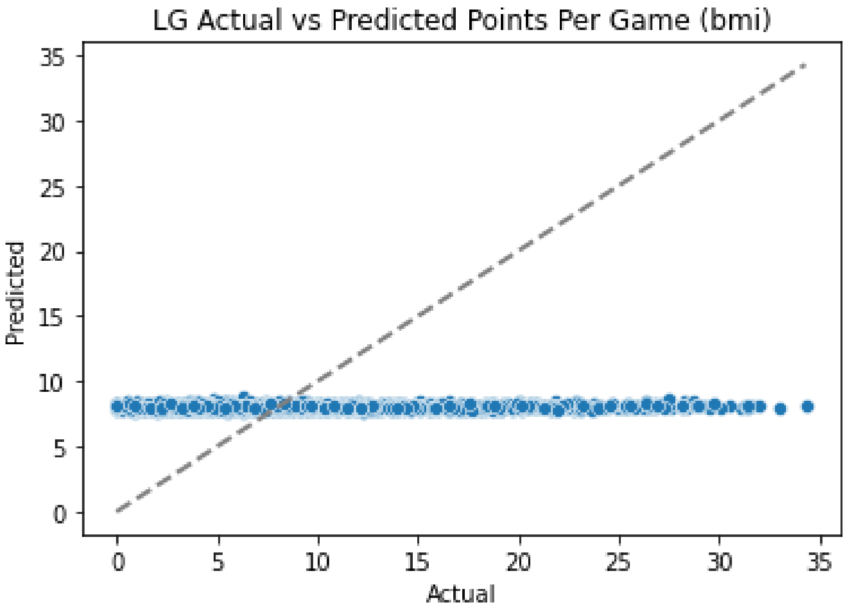


Initially the data seem to suggest a wide range of average points scored per game, but with deeper observation, there appeared to be a pattern indicating that the players with the highest scores fell within a BMI range of 24 - 28. While a BMI in this range did not necessarily lead to more points per game, it did yield the highest range of PPG compared to BMIs that fell below 24 or above 28. Additionally, most players fell within this range of BMI, which may account for the wider variety of scores observed.

However, when we used both kNN Regressor and Linear Regression models to predict average PPG using BMI, we found a very low correlation compared with other variables that we studied in our initial data exploration. The KNN Regressor model yielded an RMSE score of 6.01 after transforming our data using z-score. The Linear Regression model, by comparison, yielded a slightly higher RMSE of 6.15, and an R-squared score of 0.00.

These data led us to conclude that BMI played a very minor role, if any, in predicted average PPG. This supports our hypothesis that the reason for finding the highest average PPG in the 24 - 28 BMI range was a product of more players falling in that range, rather than BMI conferring some advantage to players that would lead them to score more points.


KNN Actual vs Predicted Points Per Game (bmi)
Predicted
Actual
0
5
10
15
20
25
30
35
0
5
10
15
20
25
30
35


LG Actual vs Predicted Points Per Game (bmi)
Predicted
Actual
0
5
10
15
20
25
30
35
0
5
10
15
20
25
30
35

Updated Hypothesis

Methodology

After finding next to no correlation between BMI and average PPG, we decided to look at the other columns that showed higher correlational scores with average PPG to see if they would be better predictors of average PPG. The top five scores were:

Assists	**0.656564**
Usage Percent	**0.638037**
Rebounds	**0.622818**
Games Played	**0.538367**
Shooting Efficiency	**0.378961**
First Round Pick Status	**0.368642**

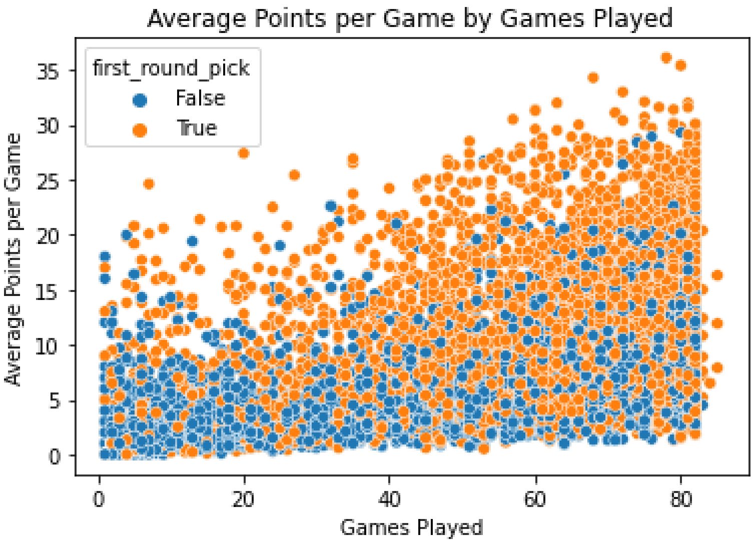
While assists and rebounds were two of the highest correlational predictors, we chose not to use them because they had a direct relationship with average points per game: namely, that scoring more assists and rebounds leads directly to scoring more points. Therefore, they are not necessarily predictors so much as factors that lead to higher scores.

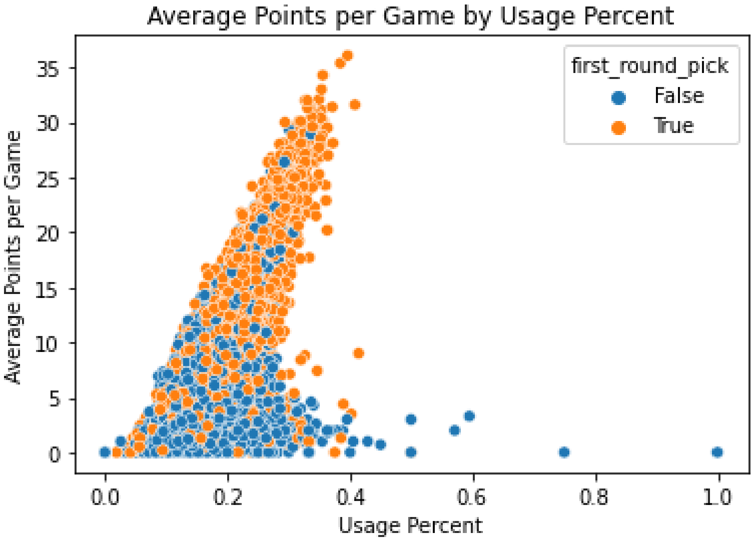
With that in mind, we chose four predictors that did not have a direct impact on PPG:

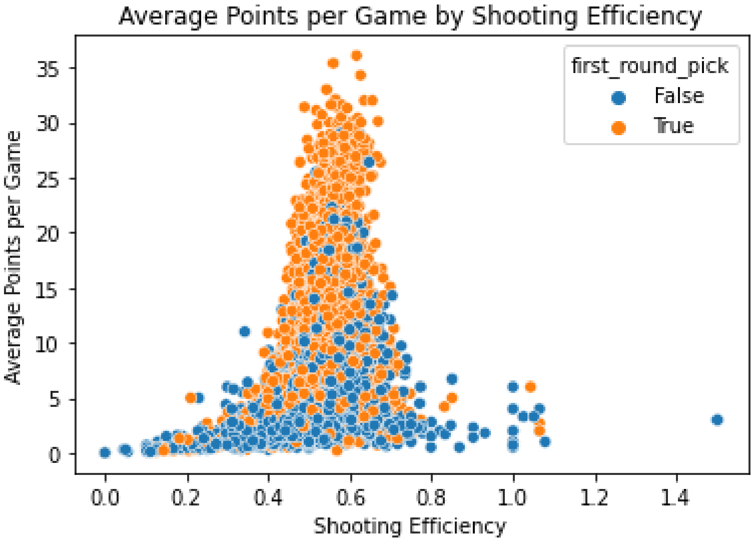
Usage Percent	**0.638037**
Games Played	**0.538367**
Shooting Efficiency	**0.378961**
First Round Pick Status	**0.368642**

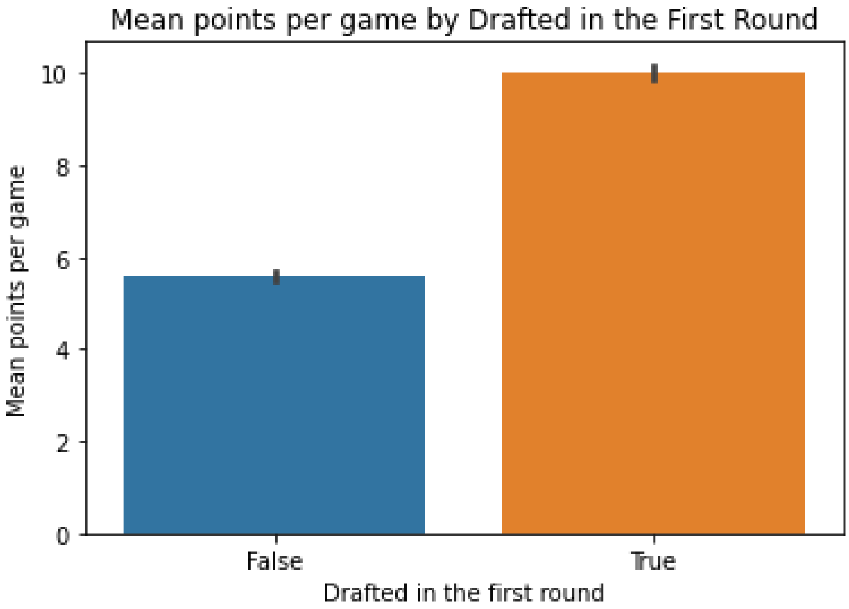
Usage percent, games played in a season, and shooting efficiency were all columns provided in the original dataset. Usage percent is defined as the percentage of team plays used by the player while he was on the floor. Games played is the number of games he played in a season. Shooting efficiency takes into account a player's 2- and 3-point shots, as well as free throws. First round pick status was a derived column indicating whether or not a player was drafted in the first round.

We created scatterplots to show the relationship between the first three predictors and average PPG, while also including a player's first round pick status. Since first round pick status is binary, we used a bar chart to show the mean of the PPG of players that were and were not selected in the first round.


Average Points per Game by Games Played
first_round_pick
False
True
Average Points per Game
0
5
10
15
20
25
30
35
0
20
40
60
80
Games Played


Average Points per Game by Usage Percent
first_round_pick
False
True
Average Points per Game
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8
1.0
Usage Percent


Average Points per Game by Shooting Efficiency
first_round_pick
False
True
Average Points per Game
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Shooting Efficiency

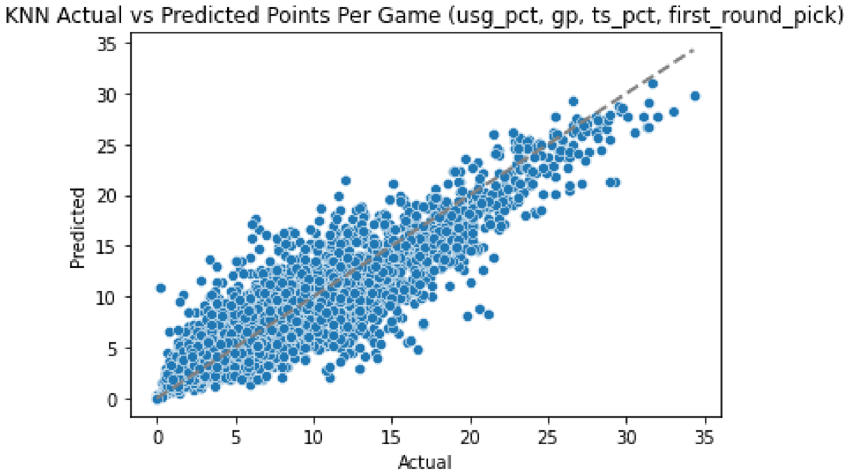


Results

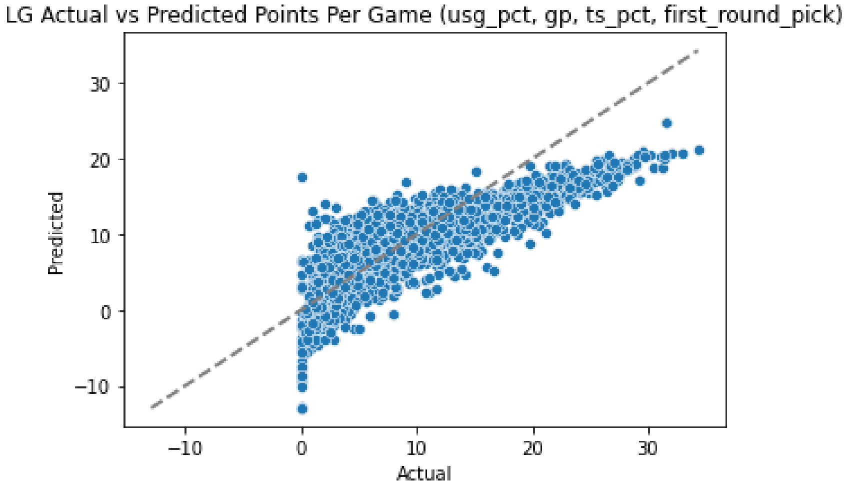
After examining the graphs, we felt more confident that we would be able to create models that would better predict average PPG. Rather than rely on a single predictor, we combined the four predictors to create a more complex, but hopefully more accurate model.

As we did with BMI, we started with a kNN Regressor with k = 10 nearest neighbors, utilizing a brute algorithm. We transformed the data using z-score and used a 70/30 split for our test/train split. We then scaled our data before creating our models.

Our new kNN Regressor model was far more accurate than our previous one that used only BMI to predict average PPG. We found our model had an RMSE of 2.71, less than half that of our original model. Additionally, plotting predicted vs. actual results showed a very clear picture of a model that demonstrated a clear pattern.



We also used a Linear Regression model so as to have one-to-one comparisons between our BMI models, and also to see how it differed from our four-predictor kNN Regressor model. We found the Linear Regression model's RMSE had grown to 3.54, with an R-squared value of 0.64. While these results proved better than both kNN Regressor and Linear Regression models utilizing solely BMI, our kNN Regressor model with four predictors was nearly 30% more accurate than our four-predictor Linear Regression model. Additionally, our Linear Regression model also predicted negative average PPG scores for some players, which are not valid scores and therefore would skew the overall accuracy of our predictions.



Conclusion

Through machine learning and data analysis, we conclude that our original hypothesis was incorrect. Our findings show that there is not a meaningful correlation between BMI and PPG. While we did find that the players with the highest PPG came from a BMI ranging from 24 and 28, this was likely due to the fact that most NBA players fall within that BMI range, rather than that BMI range specifically produces players with higher PPG.

As a result of the failure of our original hypothesis, we decided to create a new one and to test other factors with higher correlational scores to see if we could use these as a better predictor of players' PPG. Our analysis found that these other factors—the frequency with which a player is used, the player's shooting efficiency, and whether or not the player was drafted in the first round—were much better predictors of PPG than BMI. However, we cannot say that these factors necessarily cause players to score more points per game.

A more likely explanation for our findings is that players with higher skill are more likely to be drafted in the first round and more likely to be utilized by their coaches. Their shooting efficiency is likely an indication of their skill level, since more skilled players are probably scoring more frequently than less skilled players. The more a player is used, and the more accurate a player is, the more likely they are to score points. This would explain why such players have higher PPG than players who are utilized less often, as their skill level is higher, and they also gain more experience that could help improve their skill level.

Our findings could prove to be vital when coaches trade players or scout for potential NBA draft picks, whether similar college data and metrics could be used. Many aspects go into these decisions. The difference between a player's stats are minute, and every number matters. A single player could mean the difference between a multi-million dollar payout, or a complete and total loss for the team and franchise. While a player's stature does not appear to be as important as a player's skill level, BMI is used as a measure of physical fitness, and could potentially be a useful factor for coaches to consider, both when drafting players, and also when evaluating their current players.

Historical data like our NBA dataset prove vital to determining current trends, and could also be useful indicators of future trends. The more data there are, the better our predictions can be. It's these predictions that agents could use to accompany their observations of a player's skill as a way of finding the best of the best.

It should be noted that while average PPG is on the surface an important metric for analyzing player performance, there is more to basketball than simply scoring points. Defense, cooperation, player attitude, the amount of money available to a franchise, coaches, morale, and more all play important roles in success for both players and their teams. Some of these data are not readily available or difficult to quantify, and so for those analyzing performance there may be even more important factors than the ones easily accessible to researchers, coaches, staff, and agents.

Data alone cannot fully predict which players will become great and which will not. However, data are useful in many aspects of not just the NBA, but all sports. Continued analyses of available datasets can help people better understand which factors correlate with skill level—though it is important to remember that correlation does not imply causation.

Data analytics has become an important field in the world of sports, and with the rise in easily accessible data and even more quantifiable information, we can expect that trend to continue.

Resources

(n.d.) *NBA Advanced Stats*. NBA. https://www.nba.com/stats/

NBA.com Staff. (n.d.) *All-time NBA Draft History*. NBA. https://www.nba.com/news/history-draft

Steinberg, L. (2018, June 21). *The NBA Draft Process for Dummies*. Forbes. https://www.forbes.com/sites/leighsteinberg/2018/06/21/behind-the-scenes-the-nba-draft-process-for-dummies/?sh=534caa026095