Analyzing NBA Defensive Player of the Year: A Data Driven Approach

Peter D. DePaul III

05-25-2024

Contents

1 Introduction

In the NBA, the Defensive Player of the Year (DPOY) award can often be controversial. This is primarily because defense is difficult to quantify and interpret compared to offensive production. The goal of this report is to make the best effort to objectively determine who the most efficient and self-producing defender was this season. Keep in mind we are focusing on the individual's impact on their team.

2 Variable Table

When choosing the variables for this report, I will be focusing on the following individual player variables.

Table 1: Table of Variables

Variable	Description	Type
Deflections	Number of deflections	Integer
charges	Number of charges drawn	Integer
contested_shots	Number of contested shots	Integer
BLK	Number of blocks	Integer
STL	Number of steals	Integer
DBPM	Defensive Box Plus/Minus	Numeric
DRB	Defensive Rebounds	Integer
DFG_PCT	Defensive Field Goal Percentage	Numeric

I decided to use the variables from Table 1 because, of the available defensive statistics in basketball, these variables most directly explain a player's defensive capabilities. I included `contested_shots` because I believe it's unfair to compare player's defensive qualities, unless they contest and play a similar volume of shots. I will discuss further why I included certain variables, such as charges.

3 Exploratory Data Analysis

3.1 Data Collection Process

The advanced data statistics were scraped from NBA.com utilizing the data scraping tool "Data Miner". This includes the data files: `defense_2pt.csv`, `defense_3pt.csv`, `defense_dashboard.csv`, `hustle_stats.csv`, and `nba_combine_data`.

The `dpoy_voting.csv` along with the files beginning with `stathead_` were obtained from Basketball-Reference and StatHead respectively.

The data from `wingspans.csv` was collected by scraping this website using `Selenium` (SeleniumHQ 2024) in combination with `requests`. (Reitz and Contributors 2024)

Finally, the `rosters.csv` data was collected by utilizing the `nba_api` library in Python. (Patel and Contributors 2024)

3.2 Position's Impact on Receiving DPOY Votes

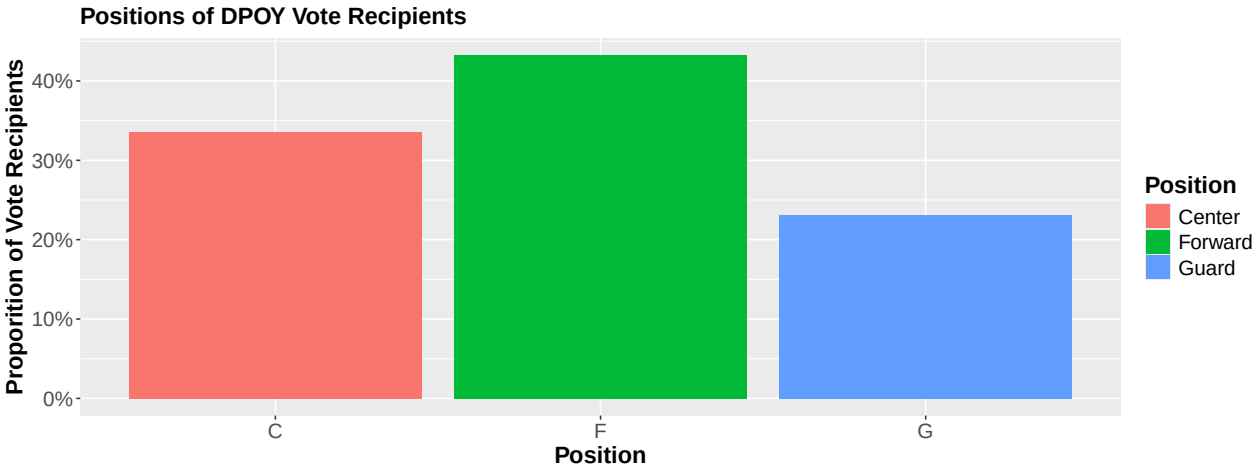

Figure 1: Positions of DPOY Vote Recipients

As we can see from Figure 1, a majority of Defensive Player of the Year vote recipients are those whose primary position is Forward, followed by Centers, and Guards. This is reasonable as Forwards often have more involved all-around roles on both offense and defense, like LeBron James. Defensive centers can be praised for their "anchoring" of the defense, like Ben Wallace.

3.3 Positional Impact on Winning Defensive Player of the Year

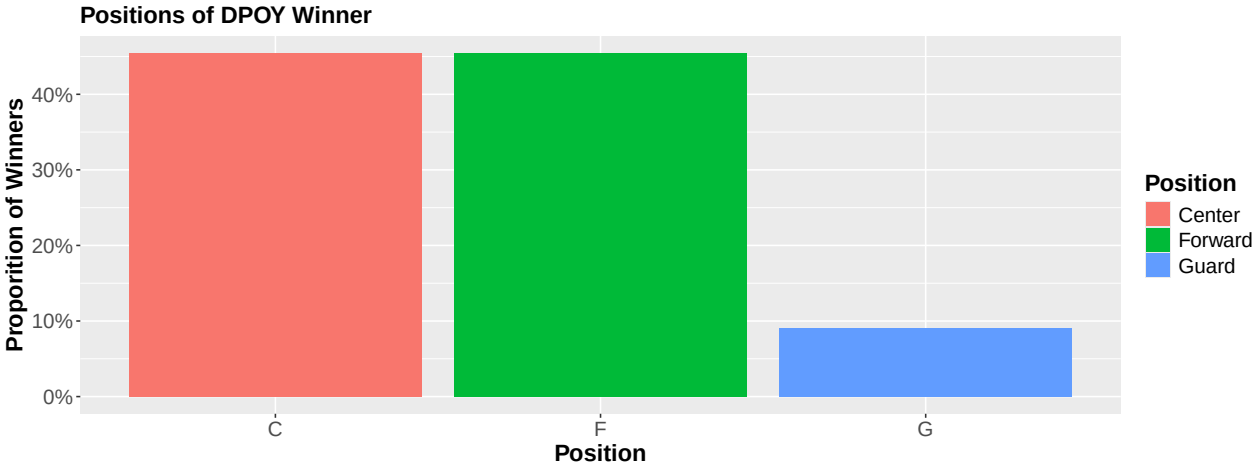

Figure 2: Positions of DPOY Winner

From Figure 2 we can see that in the last 11 seasons there has only been one Defensive Player of the Year who was a Guard, Marcus Smart in 2021-22. This suggests that voters in recent years are biased towards Centers and Forwards and may inherently devalue the credible defending capabilities of guards. This also suggests that the voters don't value defensive performance statistics, and there is some extraneous factor they value highly.

3.4 Taking a Look at Defensive Rankings

For the following defensive graphs and metrics concerning Defensive Rankings, I only utilized data for players in the top 80% of the NBA's DFGA season-by-season. I did this because I believe it's only fair for ranking purposes to compare players who see similar amounts of defensive volume.

3.4.1 Why does taking Charges Matter in Basketball?

One of the variables I expect some uproar about is the charges variable. Some might argue charges don't impact basketball that much. I disagree with this entirely. I believe that a charge, which results in a new offensive possession for your team strictly because of your ability to properly defend, is more meaningful than a lot of standard defensive plays.

I am interested in investigating how weight affects the ability to take charges. Another point that might be raised is the idea that the more you weigh, the more difficult it will be to take a charge. I don't necessarily find this to be true. It depends on the quality of the defender. A good defender who is large can still draw charges, but it might take a little more effort.

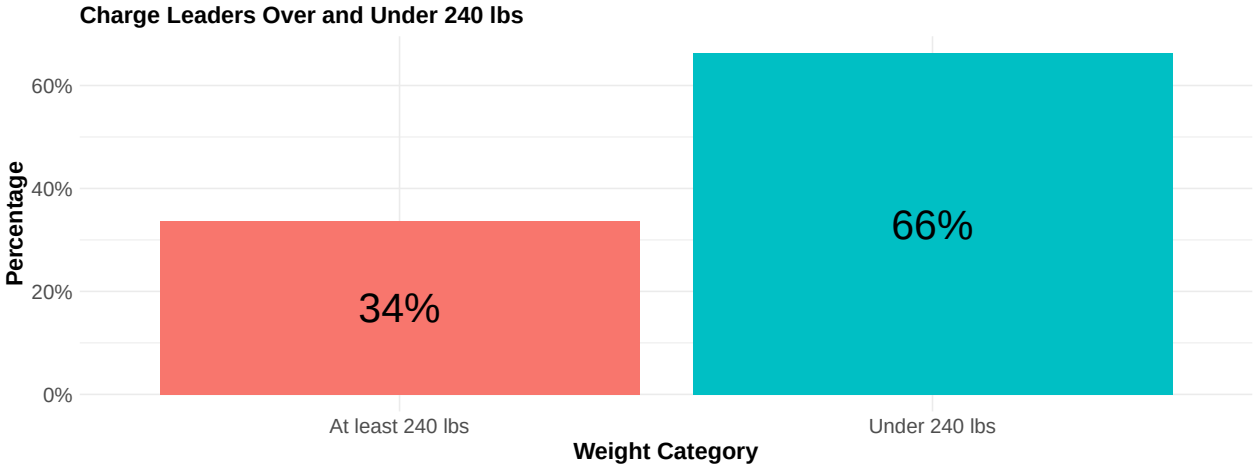

Figure 3: Charge Leaders by Weight

As we can see in Figure 3, my point stands. Since 2016-17 (when charges officially began being tracked), one-third of the charges taken leaders each season in the NBA have weighed at least 240 lbs. For reference, the mean and median weights within our data set are both around 220 lbs. This highlights the importance of taking charges on defense as a metric to establish individual success.

3.4.2 Why include Defensive Rebounding?

Rebounding is an important aspect of NBA success. Is rebounding more important on offense or defense? I'm not sure, but this article is about defense, so we'll focus on that. My main interests for including `DRB` as a variable are Rudy Gobert's elite defensive rebounding capabilities and to highlight the flaws of the `DWS` statistics, which account for `DRB` in its calculations.

The reason I'm addressing Defensive Rebounds is to give credit to Gobert's strengths. We all know he's great with blocks and Defensive Field Goal Percentage, but Defensive Rebounds are important for defense.

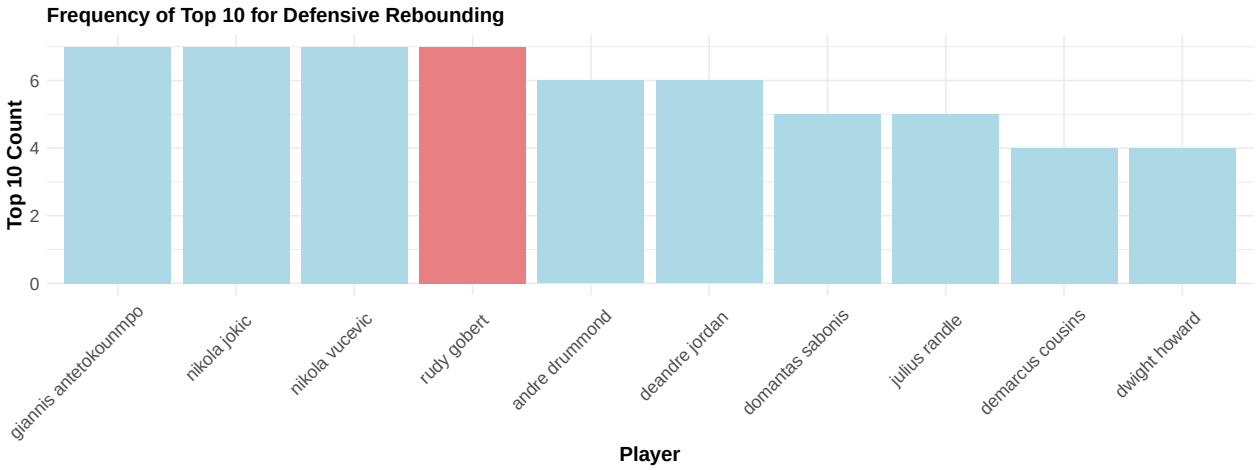

Figure 4: Top Defensive Rebounders since 2013-14

One of Gobert's best arguments for his DPOY candidacy most years is his incredible Defensive Rebounding skills. As we can see from Figure 4, Rudy Gobert is one of only four players who have the distinction of finishing top 10 in defensive rebounding seven out of the past 11 seasons. This is incredible and surpasses the number of elite rebounders such as Andre Drummond and DeAndre Jordan. However, this number is based on totals, whereas Drummond and Jordan do not log large minutes in recent years.

4 How does a player get more DPOY Votes?

For this process, I utilized the `xgboost` library along with the `tidymodels` interface to create a Gradient Boosted Decision Tree to analyze the important variables for predicting DPOY `vote_getter` status (either "Yes" or "No"). `vote_getter` establishes whether or not a player received at least 1 vote for DPOY.

4.1 Model Formula

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14}$$

Where:

Table 2: Explanation of Variables in the Regression Equation

variable	variable name	variable	variable name
$\hat{y}$	Predicted variable (vote_getter)	x_7	charges
β_0	Intercept term	x_8	def_LB
x_1	Season	x_9	Deflections
x_2	BLK	x_{10}	Height
x_3	DWS	x_{11}	Weight
x_4	DBPM	x_{12}	MP
x_5	DRtg	x_{13}	w_L_pct
x_6	DFG_PCT	x_{14}	DRB

4.2 Data Cleaning process

Most of my data cleaning process was performed using Python prior to the development of this report. Details for this can be found on the GitHub for this project. See file.

The most critical aspect of the data cleaning process for model development was data imputation. I utilized the Multiple Imputation by Chained Equations process (`mice` library in R). If you would like to read more about it (Buuren and Groothuis-Oudshoorn 2023). I found this to be a good use for this process because without imputation we still have more than 1,700 observations in our data set. This is a substantial amount of data for the process to base its decision on. The imputation used is important to keep in mind for the model and its interpretations, as there are imputations for the 2013-14 through the end of the 2015-16 seasons. Prior to 2016-17, there were no "hustle" defensive stats measured, which is why these values were missing.

4.3 Hyperparameter Tuning process

I utilized the `tune` interface of `tidymodels` (Kuhn and Wickham 2023) to hyperparameter tune the Gradient Boosted Tree to ensure the best performance given our training data. I used a 500-row random search matrix created using the `grid_latin_hypercube()` function, which searched for the ideal parameters of the boost_tree model, except for `sample_size` and `stop_iter`, which I restricted to 1 and 5, respectively.

4.3.1 Tuned Model Parameters

Table 3: Hyperparameter Values

hyperparameter	value
mtry	6
trees	138
min_n	3
tree_depth	8
learn_rate	$7.990\,051 \times 10^{-2}$
loss_reduction	$6.486\,839 \times 10^{-3}$
sample_size	1
stop_iter	5

4.4 Model Metrics and Performance

4.4.1 10 Fold Cross-validation Performance

Table 4: Cross-Validation Metrics

.metric	.estimator	mean	std_err
accuracy	binary	0.9568	0.0023
roc_auc	binary	0.9556	0.0089
specificity	binary	0.3825	0.0387

From the cross-validation metrics in Table 4, we are able to see the model has a high performance accuracy of 0.9568, which is outstanding performance. The `roc_auc` being 0.9556 also indicates that we are relatively close to a near perfect predictor for the data set. However the one thing of important note is the relatively poor performance of `specificity`. This is likely due to the limited occurrences of those receiving DPOY votes, with about 10-15 people per year receiving votes.

4.5 Performance on Test Data

Table 5: Table of Test Data Performance

Metric	Value
Accuracy	0.964
AccuracyPValue	0.011
Sensitivity	0.982
Specificity	0.659
Pos Pred Value	0.980
Neg Pred Value	0.675
Balanced Accuracy	0.820

As seen in Table 5, the `test` data set predictions achieved a high Accuracy of 96.4% along with a high Sensitivity of 98.2%. This points to the fact that the model excels at predicting when players will not receive DPOY votes. However I want to highlight a few lower-performing metrics. The Specificity achieved is 65.9%, which outperformed the cross-validation significantly. Additionally, the `AccuracyPValue` is less than 0.05. This indicates that at a 95% confidence level there is significant evidence to indicate that the model's Accuracy is better than a model achieved from randomly guessing.

Table 5 suggests that the model we have created using this imputed data performs at a high level overall, despite my desire for higher Specificity.

4.6 Variable Importance Plot

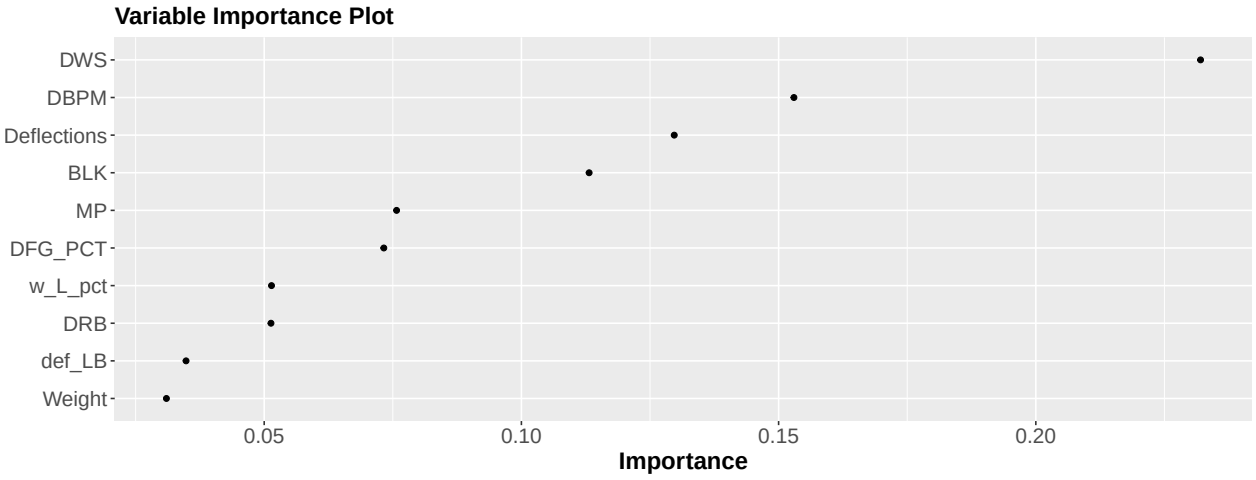

Figure 5: Variable Importance Plot

From the Variable Importance Plot of the model in Figure 5 we can highlight the 5 most important variables for determining who will receive a DPOY vote. These 5 variables are

- `DWS` - Defensive Win Shares
- `BLK` - Player's Total Blocks
- `MP` - Minutes Played
- `Deflections` - Pass Deflections
- `DBPM` - Defensive Box Plus-Minus

This goes along with what I would assume. To me these are 4 of the 5 best metrics we have for defining an NBA player's individual defensive impact. The only one of these variables I find problematic is `DWS` due to the method used for calculating it which I will discuss further.

4.6.1 Calculating Defensive Win Shares (DWS)

Defensive Win Shares are calculated using the formula:

$$\text{DWS} = \frac{\text{marginal defense}}{\text{marginal points per win}}$$

For a better understanding of the formula, I suggest visiting Basketball-Reference. However, to quickly address the flaws, it depends on both Defensive Rating and a team's defensive possessions per game for marginal defense. Marginal points per win also utilize PACE for part of the equation. The problems with Pace and Defensive Rating are that they don't really indicate defensive performance. They are both per 100 possession stats, and all they imply is how fast your team plays and, to some degree, how much your defense can slow down the other team. Anyone who watches basketball knows these stats are questionable in interpretation.

4.7 Interpretation of the Model

My goal with creating this model is to quantitatively understand what persuades an NBA DPOY voter to choose certain players. I believe I have mildly accomplished this goal. The model achieved moderately high Specificity while achieving extraordinary Sensitivity on the testing data. While it could perform poorly on future data, for what we have now, I would say it's a strong start.

5 Who Deserved the DPOY Awards?

5.1 The Highest Ranked Defenders from Each Season

Table 6: Table Highest Ranked Defenders

Player	Season	average_percentile	DPOY	diff_DPOY	diff_2nd
victor wembanyama	2023-24	87.35	0	24.49	13.06
draymond green	2022-23	83.04	0	25.89	10.42
nikola jokic	2021-22	76.83	0	13.65	6.35
draymond green	2020-21	81.57	0	3.69	3.69
anthony davis	2019-20	89.80	0	13.47	13.06
marc gasol	2018-19	77.51	0	1.82	1.22
anthony davis	2017-18	86.59	0	28.86	5.54
draymond green	2016-17	91.60	1	0.00	5.60

From Table 6 we are able to see the top ranked defenders by `average_percentile` variable since 2013-14 season.

I want to highlight the interesting part is the estimation suggests only Draymond Green's 2016-17 Season to be the only year in which the voters correctly determined the objectively best defender through these metrics. I find it important to highlight that often in these rankings too the highest rater defender is substantially higher ranking than the person who did win DPOY. This includes 2017-18, 2022-23, 2023-24 where all the highest ranked defenders did not win DPOY yet were almost 25 percentile points better on average than the DPOY winner for that year.

5.2 Let's talk about 2023-24

Table 7: Rudy and Wemby

	Percentiles							
Player	**Deflections**	**charges**	**BLK**	**STL**	**DBPM**	**DRB**	**DFG_PCT**	**average**
victor wembanyama	88.57	97.14	100.00	80.00	97.14	85.71	62.86	87.35
rudy gobert	37.14	34.29	88.57	17.14	74.29	91.43	97.14	62.86

From Table 7, these percentile variables represent their rankings among high-volume defensive players similar to their roles. This "high volume" refers to players in the top 20% DFGA by season. As we can see, Wembanyama was only worse at Defensive Rebounding and significantly worse at Defensive Field Goal Percentage. The defensive field goal percentage is problematic, but it likely has to do with his immaturity as a defender, which he'll grow out of with more experience.

Wemby's defensive game is significantly more versatile. He is in the top 20% of these high-volume defenders for deflections, charges, total blocks, total steals, defensive box plus-minus, and even defensive rebounding. His comparative weakness is his defensive field goal percentage.

Gobert, on the other hand, is primarily the king of the interior. He excels at getting blocks and preventing shots but doesn't contribute much outside of this.

Both players have a tremendous effect on their respective teams' defenses. Wemby is in the 97th percentile for Defensive Box Plus-Minus, and Gobert is in the 74th percentile. This is stellar for both. However, defensive box plus-minus is not a flawless statistic, as it neglects lineup combinations. For example, one reason Wemby is in such a high percentile for DBPM is that the Spurs were terrible at defense overall this year. When Wemby was off the court, they might as well not have played defense. Once Wemby stepped onto the court, his impact was felt immediately, leading to his high defensive box plus-minus.

6 Looking Forward and Future Considerations

Overall, there isn't much more to say regarding how voters choose the Defensive Player of the Year award. There's no perfect method for determining who gets votes, and often we're nowhere close to choosing correctly (based on my rankings).

Looking forward, I hope analyses like this can be used to build a higher-performing and more sound model. Perhaps it could even be used as a suggestion for future decisions concerning the Defensive Player of the Year. It may not be a perfect setup and ranking, but I believe it's better than whatever methods the voters are currently using. The so-called "eye test" often fails to meet the mark, and this is shown in the rankings. It shouldn't happen so often that players who are doing all they can on defense to better their team lose to players who are simply highly specialized in one aspect of defense.

7 Conclusion

From the data presented, we concluded that Rudy Gobert likely should not have been Defensive Player of the Year in the 2023-24 season; it should have been awarded to Rookie of the Year Victor Wembanyama. This is not the only time it has happened in recent memory. It is the sixth time in the last seven seasons that the award winner was not the highest-ranked defender. I built a model for predicting those who would receive Defensive Player of the Year votes, which performed moderately well. The model's importance is greater than one simply trying to predict the Defensive Player of the Year because it's up to the voters, and as I've shown, they're hard to predict.

Hopefully, further development can come along in this field of basketball analytics, as I believe defense needs the most attention from a statistical standpoint.

Bibliography

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2023. *Mice: Multivariate Imputation by Chained Equations*. https://CRAN.R-project.org/package=mice.

Kuhn, Max, and Hadley Wickham. 2023. *Tidymodels: A Collection of Packages for Modeling and Machine Learning*. https://CRAN.R-project.org/package=tidymodels.

Patel, Swar, and Contributors. 2024. "Nba_api: An API Client for NBA Statistics." https://github.com/swar/nba_api.

Reitz, Kenneth, and Contributors. 2024. *Requests: HTTP for Humans*. https://docs.python-requests.org/en/latest/.

SeleniumHQ. 2024. *SeleniumHQ Browser Automation*. https://www.selenium.dev/documentation/en/.