**No, media coverage does not always mean an NBA player is statistically dominating.**

Jake Caldwell

11/15/2022

**Introduction**

Kyrie Irving currently sits atop the NBA media’s most vilified throne. After posting the link to a documentary Rolling Stone deemed as antisemitic, social media erupted in a frenzied discussion about Irving’s character. Although Irving is considered by many to be an excellent basketball player, he has recently been featured in the media due to his off-court actions (Stinar, 2022). Other NBA players have also suffered from similar flare ups as well. When Michael Jordan left the NBA after winning his third NBA title in 1993, the focus was on his possible suspension from the league from gambling (Smith, 2009). Furthermore, when LeBron James left the Cleveland Cavaliers he was labeled as an immature coward (Amico, 2020). Even though these men, more so James and Jordan, are thought of as great basketball players, their widespread media presence does not always focus on that. Their jobs are to win games for their team, but some of the coverage they get has nothing to do with that. On the contrary, when their teams are struggling and their actual play is getting talked about, they are often blamed for their team’s failure to produce (Reiter, 2022). This suggests an odd relationship between greatness and media coverage.

I’ve mentioned a few players who are thought of as great. But what does great mean? Over the course of the last decade, analytics surrounding the NBA have become increasingly popular (Steinberg, 2015). Analytics can tell us what players are statistically above average. But professional pundits and unknown twitter users alike also have a stake in the conversation about which NBA players are considered good (Scarano, 2009). From this, there are two views on how to consider a someone a good basketball player: Statistically and the Eye Test. The Eye Test is a term that describes watching a player to determine if they are “good” (Scarano, 2009). What is problematic is that the word good in the context of the Eye Test is extremely subjective. If professional basketball pundits favor use the Eye Test, then it is fair to ask if the professional talk show hosts and the average social media user engage in a deep statistical analysis of players to back up their claims? That begs the question “does media coverage correlate to the statistical prowess of an NBA player?” A question which seems to have no clear answer. In the early 2000s, ESPN was known for highlighting the statistical anomalies from the best players in the game. But now the ever-growing number of talk shows look to highlight only the most dramatic storylines (Posey, 2021). It seems as if dramatic value has transcended statistical value.

How does one determine statistical value? What statistics are the most important? Are there other factors that influence how well a player performs that aren’t tracked.There isn’t one agreeable answer to these questions. Instead, researchers concerned with the NBA have come up with “advanced statistics” that measure efficiency, in game production, and per minute effectiveness (Khan, 2013). Instead of focusing on the accumulation of points or rebounds, advanced statistics often examine the relationship between these countable stats and time on the court, which team won, how long a player holds the ball, and other similar in game situations. For example, the GameScore statistic is a linear combination of a player’s points, assists, rebounds, steals, blocks, turnovers, and personal fouls in a game (Page et al. 2013). This statistic attempts to accumulate a player’s positive contributions in a game (Points, rebounds, assists, blocks, and steals) and subtract their slip-ups (turnovers and personal fouls). The result is a score, typically around 8, that is interpreted as a player’s basic statistical contribution to a game. For reference, the highest GameScore ever record was a sixty-nine point, six assist, eighteen rebound, four steal, and one block effort by Michael Jordan against the Cleveland Cavaliers on March 28th, 1990: 64.60 (Sports Reference LLC, 2022). Another prevalent advanced stat is Player Efficiency Rating, PER for short. This, albeit harder to interpret, measures how efficient a player is. In essence, it is like the GameScore statistic, but it accounts for contributions on a per minute basis (Sports Reference LLC, 2022). One thing to notice in both GameScore and PER is that there are far more offensive statistics used to calculate the score. Steals and blocks are trackable, but they aren’t the end-all-be-all for measuring defensive effectiveness. There are other “un-statifiable” qualities about defense that matter. For instance, there are players that might not produce many blocks and steals but are still hard to score against. Defensive spacing is also an important aspect to basketball (Franks et al., 2015). Players that constantly swarm their opponent or make opposing players think twice before shooting are also great defensive players, although their blocks and steals might not be as evident (Franks et al. 2015). The importance of defense becomes harder to track when all we have for basic defensive stats are steals and blocks.

Because there are many advanced stats to interpret, it is hard to judge the completeness of each. At the end of each NBA season, sportswriters, coaches, players, and pundits are asked to vote on the players that will win awards. Last season, Nikola Jokic won MVP, Tyler Herro won Sixth Man of the Year, and Marcus Smart won Defensive Player of the Year. One may wonder if statistics went into deciding each of these winners. Most experts say that it is the best player on the best team (NBA.com Staff, 2017). But what does best player on the best team mean? One way to determine this is to look at which players produce the most wins for their team. One way to do this is to look at a player’s Win Share statistic. Win Shares are an attempt to divvy up credit for a teams win to players based on their offensive and contributions (Sports Reference LLC, 2022). However, the math behind this is hard to interpret. Because advanced statistics, like Win Shares, are hard to decipher, experts have tried to come up with their own ways to grade players. In 1999, when many thought Michael Jordan deserved MVP, Karl Malone took home the award. IBM’s production added to team production per minute model yielded results that matched this conclusion (Berri, 1999). David J. Berri created his own model that split a players offensive and defensive contributions to a team up and then subtracted their negative contributions on either end of the floor (Berri, 1999). The result of Berri’s test determined that neither Michael Jordan nor Karl Malone should’ve been MVP, but instead Dennis Rodman, a teammate of Michael Jordan’s, should’ve won (ibid). It is clear to see that the plethora of advanced measures of performance cloud discussions of value and production.

Although there are statistical ways to measure efficiency and production, non-statistical conjecture is still prevalent when discussing the importance of NBA players. When the three-point line was added in 1980, the game of basketball shifted emphasis from high percentage two-pointers to three-pointers. Players like Stephen Curry and Damian Lillard became increasingly valuable for their ability to shoot threes from distances that Jerry West, the NBA’s current logo, might’ve deemed as impossible. Furthermore, players come into the league with various physical measurements such as wingspan, vertical jump, and height. Players can also have a high basketball IQ and the will to play hard. The above are examples of intangibles (NBADraft.com, 2016). Intangibles are qualities that players may possess that cannot be taught. For example, someone with a long wingspan might make a better defender than someone with a shorter wingspan when they are taught the same techniques. Although this might not seem like it contributes to statistical production, a player’s wingspan can mean they can block more shots and a player’s movement without the ball means the draw more attention from defenders. These attributes make it even harder to statistically determine if a player is good or not. For example, Stephen Curry might be the greatest three-point shooter of all time, but it is his off the ball movement that has defenders struggling to cover him on defense. One can look at his impressive career three-point shooting percentage without ever considering the effort he puts in to get open. Adding intangibles to the puzzle of what makes a player good further complicates the picture.

This research, however, will be strictly focused on statistical production. When pundits and social medias users alike talk about what players are better than others, it is unclear if they are strictly using their own observations or statistical analysis. Thus, I wish to examine the relationship between the familiarity of NBA players and their statistical production. This begs the question: Does statistical production have a possible correlation to the recognizability of NBA players. In 2021, the NBA released their 75th anniversary team. According to the council of pundits and players that were asked to decide the team, these are the best players to ever play the game of basketball. I want to see if the people responsible for this team may have used a deep statistical analysis to determine if a player should make this team or not and answer the question: Does statistical production of an NBA player correlate with how recognizable they are? To answer this question, I looked at the basic in game statistics that make up two advanced statistics, Player Efficiency Rating and GameScore, using a Random Forest model to tell me the significance of each basic statistic. Next, I conducted a survey that asks 30 subjects to identify members of the NBA’s 75th Anniversary team to judge recognition. Then, using the information from the random forest, I created my own statistic that combines the significant parts of Player Efficiency Rating and GameScore and use that to determine the top 50 most statistically dominant NBA players over the course of their careers. Lastly, I will compare the list of the statistically dominant and recognizable players using a dissimilarity score.

**Methods**

**Data and Survey Design:**

For this research, I had two main sources of data: 1. BasketballReference.com provides the basic in game statistics of every NBA player averaged across each season and 2. A participant-based questionnaire that asks subjects to identify NBA players from a picture of their face.

The use of a partial order knowledge structure allowed me to account for subjects having varying degrees of expertise of statistical and historical NBA knowledge. I asked each subject 5 questions of increasing difficulty. The number of questions the subject got correct correlated with the number of players they correctly identified.

Because the survey data involves human subjects, I applied for IRB approval and the committee granted my survey to be exempt from IRB review due to minimal risks of participation and guaranteed anonymity. Indirect identification is impossible for this data because each response is de-identified and the data will be aggregated into a summary. This means no single observation can be pulled out from the summary statistics. I accounted for some potential experience bias in this data by using five questions at the beginning to determine the level of expertise each participant has regarding the NBA. I stored the response data in an encrypted file folder and have since disposed of the responses file to make sure that there can be no backwards identification.

The second data set I am using is per game statistics for each NBA player from 1981-2022 (Sports Reference LLC, 2022). There are roughly 19,000 observations in the dataset. The data included the following statistics:

|  |  |
| --- | --- |
| Abbreviation | Name and meaning |
| Season | Season – The year/season that the following observations took place in |
| MP | Minutes Played - total minutes a player was on the floor |
| AST | Assists - Number of passes a player made that immediately resulted in a made shot |
| FG | Field Goals - Number of made shots in a game |
| ThP | Three Pointers – Number of made three-pointers in a game |
| FT | Free Throws- Number of made free throws in a game |
| VOP | Value of possession - average points per possession |
| TOV | Turnovers - number of times the player lost the ball to the other team |
| FGA | Field Goals Attempted - total number of shots taken by a player (make or miss) |
| ThPA | Three-pointers Attempted – total number of three-pointers taken in a game |
| DRB | Defensive Rebounds- number of rebounds grabbed by a player while on defense |
| ORB | Offensive Rebounds - number of rebounds grabbed by a player while on offense |
| FTA | Free Throws Attempted - Number of free throws taken (made or missed) |
| STL | Steals - number of times the player stole the ball from the other team |
| BLK | Blocks - number of times the player blocked the shot of an opposing player |
| PF | Personal Fouls - Number of illegal plays the player committed |
| TRB | Total Rebounds - DRB + ORB |
| PTS | Points - Total points scored by a player during a game |

**Table 1:** These are the statistics that are contained in the data from Basketball Reference. The left column contains the name of the variable column in the dataset and the right column contains the full name of the stat and a short description of the variable.

The most recognizable stats from the data are PTS, AST, and TRB because those are the stats that are usually reported in an NBA player’s box score (their totals in one game) (Table 1). There is also a time component inherent in NBA data: the season. The data has every player that played at least one game in one season. However, as mentioned previously, I excluded a player’s observations if they played in less than 21 games. This data is open source and can be shared if a citation crediting Basketball Reference’s parent company Sports Reference LLC is provided (Sports Reference LLC).

**Statistical Review**

The two advanced statistics that I focused on in this research were Player Efficiency Rating and GameScore. They are both calculations involving the basic statistics found in the Basketball Reference dataset (Table 1). Player Efficiency Rating, PER for short, is a measure of the positive and negative contributions provided by a player to their team on a per-minute basis.

However, there are a few other statistics that go into the PER equation:

|  |  |
| --- | --- |
| Abbreviation | Name and meaning |
| VOP | Value of Possession – Average number of points scored on one possession, usually 1 (Sports Reference LLC, 2022) |
| DRB% | Defensive Rebound Percentage – The percentage of a player’s rebounds that are made when that player is on defense |

**Table 2:** Value of possession is used to compare how often the player surpasses the league average points per possession and Defensive Rebound Percentage is used to determine the value of a rebound (Offensive Rebounds are more valuable but occur less often).

PER also uses team average assists, field goals, as well as the league averages of those stats to determine the player’s effectiveness compared to a league average. Below is the full equation for PER:

2

PER is a complex measurement, but in essence it measures how much time on the court is spent producing positive contributions by a player.

GameScore, on the other hand, is easier to interpret. Instead of measuring contributions on a per-minute basis, GameScore takes the basic statistics, minus field goal percentage, and creates a linear combination of them to generate a total statistical output for one game. Like PER, GameScore adds the positive contributions, like points and assists, and subtracts the negative ones, like turnovers and fouls. The equation for GameScore is:

**Equation 2:** The scalars attached to each statistic are measures of importance. See Page et al. for more information.

GameScore can simply be thought of as an overall measure of a player’s in-game statistics (Equation 2).

However, there are limitations that come with both PER and GameScore. Since PER is a time sensitive nature, players who score two points and only play one minute might end up having a comparable efficiency rating to someone who scored forty points in twenty minutes (Equation 1). On the other hand, GameScore has arbitrary scalar weights assigned to each statistic. Looking at the equation for GameScore, points and steals have the same weight, but blocks are discounted at .7 (Equation 2). In summary, PER places a lot of emphasis on making the most of a player’s time on the court and GameScore emphasizes scoring and forcing (or loosing) turnovers. That is why I have decided to take an in-depth look at both measures and what statistics matter the most. Using this information, I will then create my own statistic that measures production.

Given that PER and GameScore are not included in the Basketball Reference data, I calculated them to create two more columns. Most of the variables in the data are used to calculate PER and GameScore, making them all variables of interest. However, I also want to highlight the importance of three-pointers. Because the NBA has become more of a three-point centric league, I included it in my statistic and therefore it is also a variable of interest.

**Feature Selection and Statistic Creation**

I performed feature selection on PER and GameScore so I can create a hybrid statistic that contains the most influential features of PER and GameScore. I used a random forest algorithm from the randomforest package version 4.7-1.1 (Liaw et al., 2002) using R version 4.2.2 (2022-10-31). Below is a diagram that represents how the feature selection process looks:

Consult previous research and assign weights to variables of importance

Run optimized random forest and plot variable importance

Run random forest models with varying # of trees to optimize for MSE

Add PER and GameScore to the Data Frame

**Figure 1:** This is the feature selection process that employs the random forest algorithm. More specifically, the importance of each variable will determine how said variable will be weighted in the new statistic that I create.

This will repeat for both PER and GameScore. Once both were run and optimized for MSE, I then created my new statistic, Seasonal Production. To validate that Seasonal Production was created according to both PER and GameScore, it had to be numerically like PER and GameScore. The best PERs are upwards of 30 and the best GameScores are upwards of 50. Given this, I checked that Seasonal Production fit in this range of scores.

I then calculated each player’s career average Seasonal Production and took the top 50 highest. This generated my list of the best statistical producers in the NBA from 1981-2022. I took the top 50 because that is how many players are in my survey. This allowed me to have two ranked lists, one on familiarity and the other on statistical performance, of equal length to compare.

Seasonal Production, as previously stated, is a hybrid of PER and GameScore. It combines the linear combination structure that GameScore uses with the per-minute production that PER employs. I was inspired by a well-known NBA youtuber, JxmyHighroller, to combine these two statistics because he used both PER and GameScore to perform some surface level analytics on players that are typically considered to be in the greatest of all time conversation (JxmyHighroller, 2021). However, the results of the random forest feature selection and the work previously done by David J. Berri are what lead to the actual structure of the equation.

**Dissimilarity Calculation**

After both ranked lists have been generated, the last step in this project is to compare the lists to see if they are similar enough to say that familiarity is possibly correlated with statistical production. To do this, I used the Bray-Curtis and Jaccard dissimilarity test found in the R-package vegan (Oksanen et al., 2022). Bray-Curtis is commonly used in ecology to measure the similarity between species abundance at different sites (Ricotta et al., 2017). The Bray-Curtis test checks both lists for their contents and the rank of their contents, the Jaccard test just checks the contents of the lists. I used these in my research by equating abundance to rank and containment: the higher ranked a player is, the more abundance they would have and if a player is in both lists, then the player is seen favorably by statistics and pundits. I took the ranks of each list and reversed them, making the highest ranked player a fifty and the lowest a one. I did this to keep the output of the Bray-Curtis test consistent with its normal interpretation. The output of the Bray-Curtis test is a score between zero and one, zero being that the lists are the same and one meaning that the lists are completely different. If the output of the Bray-Curtis test is closer to zero than one, then there is a higher likelihood that there is correlation between recognizability and statistical production of NBA players. The output of the Jaccard Index is also between zero and one, but this time, if the value is closer to one, then the lists have more observations in common (Bag et al., 2019). If the output of the Jaccard test is gJreater than 0.5, then there is more evidence of possible correlation between recognizability and statistical production of NBA players.

I also used the Bray-Curtis index to measure the ranks of the NBA 75 team players in terms of their recognizability and their statistical production. This sub-test yielded a deeper understanding of the relationship between familiarity and statistical production.

**Results**

**Summary of Data**

The question I am trying to answer with this research is, “is recognizability correlated with the statistical production of NBA players?” First, the table below summarizes the key variables from the Basketball Reference data:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PTS | Mean | Median | StdDev | Minimum | Q1 | Q3 | Max |
| PTS | 9.5 | 8.2 | 5.9 | 0.3 | 4.8 | 13.0 | 37.1 |
| AST | 2.1 | 1.5 | 1.9 | 0.0 | 0.8 | 2.9 | 14.5 |
| TRB | 3.9 | 3.4 | 2.5 | 0.1 | 2.1 | 5.2 | 18.7 |
| ORB | 1.1 | 0.9 | 0.8 | 0.0 | 0.5 | 1.6 | 6.9 |
| DRB | 2.8 | 2.4 | 1.8 | 0.1 | 1.5 | 3.7 | 12.3 |
| FG | 3.5 | 3.1 | 2.2 | 0.1 | 1.8 | 4.9 | 13.4 |
| FGA | 7.8 | 6.8 | 6.8 | 0.4 | 4.1 | 10.7 | 27.8 |
| ThP | 0.5 | 0.2 | 0.6 | 0.0 | 0.0 | 0.9 | 5.3 |
| ThPA | 1.5 | 0.7 | 1.8 | 0.0 | 0.1 | 2.6 | 13.2 |
| FT | 1.7 | 1.3 | 1.4 | 0.0 | 0.7 | 2.4 | 10.3 |
| FTA | 2.3 | 1.8 | 1.82 | 0.0 | 1.0 | 3.2 | 13.1 |
| STL | 0.7 | 0.7 | 0.4 | 0.0 | 0.4 | 1.0 | 3.7 |
| BLK | 0.4 | 0.3 | 0.5 | 0.0 | 0.1 | 0.6 | 5.6 |
| TOV | 1.3 | 1.2 | 0.8 | 0.0 | 0.8 | 1.9 | 5.7 |
| PF | 2.0 | 2.1 | 0.7 | 0.2 | 1.5 | 2.6 | 4.8 |
| MP | 22.6 | 22.5 | 9.2 | 2.3 | 15.1 | 30.5 | 43.7 |
| PER | 15.0 | 14.5 | 3.6 | 4.3 | 12.5 | 16.9 | 34.9 |
| GameScore | 7.0 | 6.0 | 4.5 | -0.2 | 3.4 | 9.7 | 27.8 |

**Table 3:** This table has the summary statistics for the variables of interest. Refer to table 1 for definitions of the variables in this table.

The average basic box score for an NBA player is 9.5 points, 2.1 assists, and 3.9 field goals on 3.5 made field goals (Table 3). The average NBA player also turns the ball over 1.3 times a game and commits 2.0 fouls. On the defensive side of the ball, the average NBA player records 0.7 steals and 0.4 blocks. These numbers result in an average PER of 15.0 and GameScore of 7.0.

There is a large discrepancy between the third quartile and the maximum PER and GameScore in the data set. For reference, random forest models were used to run regression on GameScore and PER because random forests don’t require assumptions.

Chart, histogram

Description automatically generated

**Figure 2:** This plot compares the distribution of career average PER and GameScore. The dotted line represents two standard deviations away from the mean. The PER distribution looks to be approximately normal, whereas the Gamescore distribution looks to be skew right.

The mean PER is slightly higher than the mean Gamescore distribution, but both distributions show limited observations beyond the two standard deviation mark (Figure 2). To more clearly visualize this, below is a plot that compares PER and Gamescore on the same axis.

Chart, scatter chart

Description automatically generated

**Figure 3:** This plot shows the relationship between GameScore and PER. The black lines are located at the 3rd quartile of PER and GameScore. The upper right area of the plot are the standout statistical producers in the NBA over the course of their careers. The blue dots represent players on the NBA 75th anniversary team.

Most of the NBA players sit outside of the upper right quadrant of figure three. The players that made it into the upper right quadrant are, considering PER and Gamescore, the most productive players of all time. Most of the NBA 75th Anniversary team also resides in the upper right corner as well (Figure 3). Before I ran the random forest model for feature selection, I found the top 50 players of all time based on career average PER and GameScore. I then compared them to the players in the NBA 75th Anniversary team to get a preliminary understanding of any connections between stats and supposed greatness. For reference, random forest models were used to run regression on GameScore and PER because random forests don’t require assumptions.

Diagram

Description automatically generated

**Figure 4:** This Venn diagram shows the number of items that the top 50 highest PER, GameScore, NBA 75th anniversary players that played after 1981 lists have in common. PER shares 14 with the NBA 75 team, while GameScore shares 29.

Looking at both PER and GameScore, the GameScore top 50 list has more players in common with the NBA 75 team than the PER list (Figure 4). Because PER is a much more advanced calculation than GameScore, PER targets some players that don’t stand out as much as the players on the 75th anniversary team (NBA.com, 2021).

**Statistic Creation**

Before I got into my random forest models, I checked to see what one regression tree for both PER and GameScore would look like.

Diagram, timeline

Description automatically generated

**Figure 5:** This is one singular regression tree for PER. It does use almost all the components from PER, but PTS is used to split 3 different nodes, which is something to consider.   
Timeline

Description automatically generated

**Figure 6:** This is one singular regression tree for GameScore. The glaring observation here is that all nodes are split using points.

Both trees, more so the GameScore tree, overwhelmingly use points to split the nodes for regression (Figures 5 and 6).

To create my own statistic, I used a random forest regression algorithm on both PER and GameScore to determine the importance of each component that is included in the Basketball Reference data.

Chart

Description automatically generated with medium confidence

**Figure 7:** The variable importance plot for the PER random forest model. Points is the most important variable, followed closely by Field Goals Made, but that makes sense because the number of field goals made contributes to points scored.

A picture containing text, sky, screenshot

Description automatically generated

**Figure 8:** Field Goals Made is the most important variable for GameScore random forest model, followed by Points. Again, same relationship between points and made field goals.

Both models have points and field goals as the most important variables (Figure 7 and 8). PER uses a combination of points, field goals, field goals attempted, free throws, and free throws attempted to build a scoring metric, which makes sense because the variables that as close to being as important as points in both random forests (Figure 7 and 8).

Given the importance of each variable from both models, I created the statistic called Seasonal Production:

Seasonal Production = (Table 1)

**Equation 3:** This is the stat that I created using the feature selection from the random forest models. It has weights like GameScore but emphasizes some of the least important statistics that are relevant. Refer to table 1 for variable definitions

The equation keeps the scoring metric provided by PER and has a similar structure to GameScore. However, it splits up offense, defense, and negative plays into three distinct per minute categories. The main takeaway from the random forests was that points and the related shooting stats were the most important, but some of the other statistics, like rebounds, assists, and steals, needed to be brought to greater attention. I also added three point makes to the stat because in today’s NBA the three pointer is of upmost importance and contributes heavily to scoring.

Chart, histogram

Description automatically generated

**Figure 9:** This is the distribution of Seasonal Production once it has been created. It has a more skewed distribution than GameScore does, but that is because it places a higher emphasis on underrepresented stats in the PER and GameScore random forest models. This punishes players who rely heavily on scoring.

For validation, the average Seasonal Production is around 10 and the max is just under 40, which fits alongside of PER and GameScore. Furthermore, the larger amount of skew in the distribution of Seasonal Production places more players below 2 standard deviations away from the average (mean) (Figure 9). I then used this to generate my top 50 most productive players over the course of their career list (Appendix A3).

**Survey Results**

After 30 respondents answered the survey questionnaire, I aggregated their responses into counts of their responses to the initial five question partial order knowledge structure and a ranked list of the players from the NBA 75th Anniversary team based on how often they were recognized (Appendix A4). Below is a summary of the results of the partial order knowledge structure.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Statistic | Mean | Median | Std. Dev. | Min | Q1 | Q3 | Max | N |
| POKS Score | 3.0 | 3.0 | 1.3 | 1 | 2 | 4 | 5 | 30 |

**Table 4:** This table contains the summary statistics of the survey partial order knowledge structure. Refer to the methods as to why this was included in the survey.

Most of the respondents answered at least two to three out of the five total questions. Those first three involved knowing what NBA and PPG stands for and who the all-time leading scorer is (Appendix A1). Most subjects had at least some basketball knowledge (Caldwell, 2022).

I also asked the subjects questions about each player they recognized. Only a few of the respondents were able to tell what a player’s career box score was. Fewer said they were a fan of each player, but it was a mixed bag when it came to knowing the last team each player played at least a 20-game season with. Furthermore, all respondents agreed that they recognized the players they saw because of media coverage and not because of watching basketball often.

The resulting list of the most recognizable players had arguably the most covered players by the media at the top: Stephen Curry, LeBron James, Kobe Bryant, Michael Jordan, Larry Bird, and Magic Johnson. The rest of the list varied, but there were some players, like Dave Cowens and Nate Archibald, who did not get recognized a single time (Appendix A4).

**Dissimilarity Metric**

The last step in my research is to use the Bray-Curtis dissimilarity metric to answer my research question: “Is there correlation between recognizability and statistical production of NBA players?” Before I computed the Bray-Curtis metric, I looked at the number of players that were included in both the statistical list and the familiarity list.

Diagram

Description automatically generated

**Figure 9:** This Venn diagram illustrates the number of players that the familiarity list and statistical production list share. As you can see, both lists have 50 entries and share 24 entries, which is less than that of GameScore from figure 5.

The lists share less than 50% of their players with each other (Figure 10). I then calculated the Bray-Curtis metric by hand as well as using the vegan package in R (Oksanen, 2022). From these calculations, both yielded a dissimilarity score of 0.493. This is an intermediate score (Ricotta, 2017). This means that I cannot answer my research question because an intermediate Bray-Curtis score means that there isn’t enough difference (or similarity) in either ranked list to rule out either no possible correlation or some possible correlation between familiarity and statistical production. The Jaccard metric between the familiarity and statistical production list is 0.661. This means that lists share roughly 66% of their contents.

Lastly, the Bray-Curtis metric between the ranked familiarity of the NBA 75 players and the NBA 75 players ranked Seasonal Production is 0.179. Based on the Bray-Curtis metric, this means that there might be some correlation between familiarity of familiarity and statistical production solely among the NBA 75th Anniversary team.

**Discussion**

As stated in the results, the dissimilarity between the 50 greatest statistical producers in the NBA and the most recognizable members of the NBA 75th Anniversary team was roughly 50%. This means that it is a coin flip to whether there might be correlation between familiarity and statistical production of NBA players. This could be because I only sampled 30 students at a Liberal Arts college or because of the statistic I created. However, I hypothesized that there would be no correlation between stats and recognition, and that is not the case here. However, when I looked at the similarity between the rankings of the NBA 75th Anniversary team based on my survey results and my new created statistic, I found that there was a high degree of similarity. This could be evidence that the players in the NBA are continually getting better and that the next generation of players will replace the old guard at the top . The NBA is full of players who help their teams win with their contribution’s night in and night out. For example, on November 13th, 2022, Joel Embiid had 59 points, 11 rebounds, 8 assists, and 7 blocks in a win over the Utah Jazz (Sports Reference LLC, 2022). This resulted in a GameScore of 54.4, only fifteen less than the highest GameScore ever recorded. Embiid did make my 50 best statistical producers list but was not a member of the NBA 75th Anniversary team (Appendix A3). In the past three years, 50-point scoring efforts are becoming more common (Quinn, 2020). The landscape of the NBA is changing, from the ever-increasing role of the three-point shot to the regular occurrence of big games, some of the great players of the past may becoming statistically forgotten.

One insight I found particularly interesting was that all participants in the survey said that they recognized the players because of media presence rather than watching basketball. Even the respondents who scored a five on the partial order knowledge structure agreed. If I made the question more general and asked if “they recognized the player because of media coverage or following basketball regularly”, those that scored a four or five might’ve said because they follow the NBA. However, the vagueness of following could mean that they watch game highlights or follow their favorite teams on social media. But both of those mediums might be considered media coverage by some people. There is a strong correlation between basketball knowledge and social media engagement. This would also result in those more knowledgeable seeing the more dramatic storylines about other NBA players as well. This begs another sub question to my research: Does non-basketball related media coverage make an NBA player more recognizable than basketball related coverage? In our current media climate, this question may produce an interesting answer.

I tried to create a stat that made some of the less important variables from PER and GameScore more significant. This stat produced a list that placed Magic Johnson and LeBron James as the two greatest statistical producers of all time (Appendix A3). While this might check out for some basketball fans, others might point out that neither Magic nor LeBron were prominent scorers. Between them, they have only led the league in scoring once (LeBron in 2008) (Sports Reference LLC, 2022). Looking at the variable importance for PER and GameScore, points scored per game was one of the most important variables for calculating both. One could argue that points scored could be an end-all-be-all in discussing who is the best basketball player ever, as points are what determines what team wins a game. But as Figure 4 shows, high PERs and GameScores, which also use other basic stats, are for a select group of players. Thus, statistics isn’t the only thing that determines greatness.

Furthermore, players, like Kyrie Irving, have a preceding reputation. One could argue that players have a basketball reputation and status that determines their salary and trade value (Ertug et al., 2013), but they may also have a non-basketball reputation that is more recognizable. Obviously, NBA players are still humans and have lives outside of basketball, but it is their job to go out and put their best effort out on the court, but should they be held to the lofty standards that the media holds them to? I cannot answer that question with my findings, but it is an interesting question to ask.

I reached an inconclusive result for the relationship between stats and familiarity, which could be due to the limitations of the study. For one, there are similar concepts to familiarity that I could’ve used. Jersey sales and TV ratings are a few measures of popularity of an NBA player that I could’ve examined more quantitatively than recognizability (Hausman et al., 1997). Similarly, my sample size and sample location present hidden biases that don’t allow the results of the survey to be generalized to a broader population. Furthermore, the statistic I created was through my interpretation of the random forest models I used, as well as the prior research in determining value that I read. My statistic has biases present because I created it. This is where value determination becomes murky. The stats I used come from observable moments in an NBA game, and I weighted them based on their importance values from the random forest models I ran, but at the end of the day, I am the one who split up offense, defense, and mistakes. I am also the one who chose to examine these splits on a per minute basis. I did base these findings in analysis and research, but that doesn’t mean I was completely bias free when creating the stat. It would be interesting to see many NBA analysts create a myriad of stats to test my research question on.

The data I used was only from the regular season, which is another limitation because of the exclusion of the playoffs. However, the playoffs have a very different competitive environment than the regular season. There are higher stakes, less games to play, and a higher degree of intensity (Horowitz, 2018). Some players elevate their games to will their teams to wins, while others shy away from the proverbial spotlight. Running the same study but strictly for the playoffs would produce different results than my study, but there would be fewer observations (players) and games played. It would still be another interesting perspective on my topic.

In the future, this research produces sub questions that should be answered. I previously mentioned that the creation of more stats by more statisticians could result in different dissimilarity scores. Similarly, examining different samples of people with a sample size of greater than 30 could also alter the results I found. On top of this, I would like to examine two sub questions: “Does the age of a survey respondent affect which players are more recognizable?” and “What advanced statistic most closely relates to the players selected to the NBA 75th Anniversary team?” Both questions call out my statistical selection process and my sample that I surveyed. I believe these to be my biggest limitations of my research. Examining the new questions that I suggested may give me better insight into why I got the results that I did.

**Appendix:**

**A1:** Survey Questionnaire Structure:

1. **Level of Expertise:**
   1. What does the NBA stand for?
   2. What does PPG stand for?
   3. Who is the all-time leading scorer in the NBA (Regular Season)?
   4. What does PER stand for?
   5. Who is the last player to win 6th man of the year?
2. **Player photos**
   1. If they recognize the player shown, participant will be asked the following:
      1. What is the player's career stat line (Points per game, Assists per game, Rebounds per game)?
      2. What team has the player most recently played for?
      3. Are you a fan of this player?
3. **Debriefing Question**
   1. Do you think that you recognized the players that you did because of media coverage or watching basketball regularly?

**A2:** Oral Informed Consent Language:

* Participants are being asked to answer questions about their knowledge of NBA statistics and general interest. Then, they are asked to look at ten pictures of NBA players from the NBA 75 team (Who played in the 1980 season and after) to see if they can identify them. If they recognize a player, participants will be asked a few questions about the player's stats, player’s most recent team, and whether the participant is a fan of said player. Finally, a quick debrief at the end of the conversation about whether the participant has seen any of the players they recognized in the media (social media, on TV, etc.).
* The purpose of these tasks is to generate a list of the most recognized players that can be compared to a ranked list generated by statistical testing on the researcher's end. The comparison will result in a correlation measure to determine if familiarity is correlated with statistical production.
* The potential benefits of this project are as follows: Possible increased interest in the NBA, a fun break from work (questionnaire is like a game), giving someone who might not be super interested in the NBA something to talk about if they are in a conversation about it.
* The one potential risk is embarrassment for not recognizing any of the players. However, this is not judged based on correctness. Think of it more as a game.
* Some possible alternatives to participating in this research is to compare the players that you see on social media or on TV with the players that win MVP or have a great stat line night in and out.
* Participation is completely voluntary, and you can stop participating at any time with no consequences.
* Please acknowledge that you are 18 years of age or older.

**A3:** Statistical Production Ranked List:

|  |  |
| --- | --- |
| Player | SProd |
| Magic Johnson | 30.64442 |
| LeBron James | 29.42305 |
| Trae Young | 28.62075 |
| Michael Jordan | 28.40423 |
| Luka Doncic | 28.297 |
| Chris Paul | 27.11406 |
| Larry Bird | 26.93427 |
| James Harden | 26.27836 |
| Damian Lillard | 25.9602 |
| Stephen Curry | 25.54558 |
| Russell Westbrook | 25.35943 |
| Kevin Durant | 25.02743 |
| Nikola Jokic | 24.243 |
| John Stockton | 24.20105 |
| Kevin Johnson | 23.561 |
| Ja Morant | 23.40933 |
| Allen Iverson | 23.2915 |
| John Wall | 23.2855 |
| Kyrie Irving | 22.9196 |
| Isiah Thomas | 22.72685 |
| Giannis Antetokounmpo | 22.15278 |
| Charles Barkley | 22.13027 |
| Tim Hardaway | 22.05369 |
| Tyrese Haliburton | 21.99133 |
| Anthony Davis | 21.9367 |
| Karl-Anthony Towns | 21.64543 |
| Joel Embiid | 21.57667 |
| Ben Simmons | 21.56325 |
| Clyde Drexler | 21.46444 |
| Kobe Bryant | 21.39053 |
| Jason Kidd | 21.35495 |
| Zion Williamson | 21.3425 |
| Karl Malone | 21.26363 |
| Alex English | 20.80955 |
| LaMelo Ball | 20.723 |
| Donovan Mitchell | 20.6882 |
| Steve Nash | 20.60171 |
| Stephon Marbury | 20.45464 |
| Dwyane Wade | 20.20676 |
| Gary Payton | 20.1675 |
| Adrian Dantley | 19.98364 |
| De'Aaron Fox | 19.9164 |
| Deron Williams | 19.84123 |
| Kemba Walker | 19.81055 |
| Jimmy Butler | 19.62018 |
| Gus Williams | 19.5895 |
| Devin Booker | 19.53757 |
| Kawhi Leonard | 19.43333 |
| Mark Price | 19.4113 |
| Norm Nixon | 19.41057 |

**A4:** Familiarity Ranked List:

|  |  |
| --- | --- |
| Player | Rank |
| Stephen Curry | 50 |
| LeBron James | 49 |
| Kobe Bryant | 48 |
| James Harden | 47 |
| Michael Jordan | 46 |
| Kevin Durant | 45 |
| Magic Johnson | 44 |
| Larry Bird | 43 |
| Kawhi Leonard | 42 |
| Jason Kidd | 41 |
| Allen Iverson | 40 |
| Dirk Nowitzki | 39 |
| Chris Paul | 38 |
| Shaquille O'Neal | 37 |
| Tim Duncan | 36 |
| Scottie Pippen | 35 |
| Russell Westbrook | 34 |
| Kevin Garnett | 33 |
| Carmelo Anthony | 32 |
| Steve Nash | 31 |
| Charles Barkley | 30 |
| Clyde Drexler | 29 |
| Anthony Davis | 28 |
| David Robinson | 27 |
| Damian Lillard | 26 |
| James Worthy | 25 |
| Reggie Miller | 24 |
| Hakeem Olajuwon | 23 |
| Giannis Antetokounmpo | 22 |
| Dwyane Wade | 21 |
| Ray Allen | 19 |
| Julius Erving | 20 |
| Dennis Rodman | 18 |
| Patrick Ewing | 17 |
| Gary Payton | 16 |
| Bill Walton | 15 |
| Robert Parish | 14 |
| Kareem Abdul-Jabbar | 13 |
| Moses Malone | 12 |
| Isiah Thomas | 11 |
| John Stockton | 10 |
| George Gervin | 9 |
| Karl Malone | 8 |
| Paul Pierce | 7 |
| Dominique Wilkins | 6 |
| Kevin McHale | 5 |
| Bob McAdoo | 4 |
| Elvin Hayes | 3 |
| Tiny Archibald | 2 |
| Dave Cowens | 1 |

# **Works Cited**

Amico, S. (2020, July 21). *Moon: LeBron leaving Cavaliers first time was 'like a funeral'*. Retrieved from SI: https://www.si.com/nba/cavaliers/nba/cavaliers/nba-amico/jamario-moon-lebron-james-miami-heat

Berri, D. J. (1999). Who Is "Most Valuable"? Measuring the Player's Production of Wins in the National Basketball Association. *Managerial and Decision Economics 20(8)*, 411-427.

Caldwell, J. (2022, November 9). NBA Familiarity Survey Response Data.

Desmarais, M. C., Maluf, A., & Liu, J. (1995). User-expertise modeling with empirically derived probabilistic implication networks. *User modeling and user-adapted interaction*, 285-315.

Ertug, G., & Castellucci, F. (2013). GETTING WHAT YOU NEED: HOW REPUTATION AND STATUS AFFECT TEAM PERFORMANCE, HIRING, AND SALARIES IN THE NBA. *The Academy of Management Journal, 56(2)*, 407-431.

Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015). CHARACTERIZING THE SPATIAL STRUCTURE OF DEFENSIVE SKILL IN PROFESSIONAL BASKETBALL. *The Annals of Applied Statistics, 9(1)*, 94-121.

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC bioinformatics, 14(1)*, 1-8.

Hausman, J. A., & Leonard, G. K. (1997). Superstars in the National Basketball Association: Economic Value and Policy. *Journal of Labor Economics, 15(4)*, 586-624.

Horowitz, I. (2018). Competitive Balance in the NBA Playoffs. *The American Economist, 63(2)*, 215-227.

Khan, E. (2013, October 18). *Advanced NBA Stats for Dummies: How to Understand the New Hoops Math*. Retrieved from Bleacher Report: https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math

Kinney, D. W. (2009). Internal Consistency of Performance Evaluations as a Function of Music Expertise and Excerpt Familiarity. *Journal of Research in Music Education, 56(4)*, 322-337.

Liaw, A., & Wiener, A. (2002). Classification and Regression by randomForest. *R News*, 18-22.

NBA.com. (2021, October 21). *NBA 75TH ANNIVERSARY TEAM ANNOUNCED*. Retrieved from NBA: https://www.nba.com/news/nba-75th-anniversary-team-announced

NBA.com Staff. (2017, April 12). *BLOGTABLE: WHAT CRITERIA MATTERS MOST IN MAKING YOUR MVP DECISION?* Retrieved from NBA: https://www.nba.com/news/blogtable-what-criteria-matters-most-making-mvp-decision

NBADraft.net. (2016, June). What does actually being intangible mean?

Oksanen, J., Simpson, G. L., & Blanchett, F. G. (2022). vegan: Community Ecology Package.

Page, G. L., Barney, B. J., & McGuire, A. T. (2013). Effect of position, usage rate, and per game minutes played on nba player production curves. *Journal of Quantitative Analysis in Sports, 9(4)*, 337-345.

Posey, J. (2021, October 29). *Many NBA players don't trust the media today, and I can't blame them*. Retrieved from Basketball News: https://www.basketballnews.com/stories/many-nba-players-dont-trust-the-media-today-and-i-cant-blame-them

Reiter, B. (2022, April 7). *Why LeBron James deserves full blame for Lakers' failures, and should no longer have say in solution*. Retrieved from CBS Sports: https://www.cbssports.com/nba/news/why-lebron-james-deserves-full-blame-for-lakers-failures-and-should-no-longer-have-say-in-solution/

Ricotta, C., & Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, 201-205.

Scarano, D. (2009, March 25). *The Eye Test: Watching Makes a Difference*. Retrieved from Bleacher Report: https://bleacherreport.com/articles/144924-the-eye-test

Schimek, M., Budinska, E., Ding, J., Kugler, K., Svendova, V., Lin, S., & Pfeifer, B. (2022). TopKLists: Inference, Aggregation and Visualization for Top-K Ranked Lists. *R package version 1.0.8*. https://CRAN.R-project.org/package=TopKLists. Retrieved from CRAN.

Sigler, K. J., & Sackley, W. H. (2000). NBA players: are they paid for performance? *Managerial finance, 26(7)*, 46-51.

Smith, M. (2009, February 28). *Michael Jordan's First Retirement: Was It a Secret Suspension?* Retrieved from Blearcher Report: https://bleacherreport.com/articles/131997-mjs-1st-retirement-was-it-a-secret-suspension

Sports Reference LLC. (2022, October 19). *2021-22 NBA Player Stats: Per Game*. Retrieved from Basketball Reference: https://www.basketball-reference.com/leagues/NBA\_2022\_per\_game.html

Steinberg, L. (2015, September 18). *CHANGING THE GAME: The Rise of Sports Analytics*. Retrieved from Forbes: https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/?sh=3d537d8d4c1f

Stinar, B. (2022, September 24). *Kyrie Irving's Shockingly Low Ranking In ESPN's Top-100 Players*. Retrieved from si.com: https://www.si.com/fannation/nba/fastbreak/news/kyrie-irving-shockingly-low-ranking-in-espns-top-100-players#:~:text=ESPN%20ranked%20Kyrie%20Irving%20as,for%20the%202022%2D23%20season.&text=The%20Brooklyn%20Nets%20did%20not,their%2082%20regular%20season%2