Does Statistical Excellence Influence an NBA Player's Popularity?

Jake Caldwell

10/28/2022

Introduction:

In current discussions about how well NBA players perform, the topic of statistical production is often overlooked by pundits. The goal of this research is to figure out if statistical production has some correlation with popularity. More specifically, do players that get featured in media often statistically perform better than lesser mentioned players? With the recent release of the NBA 75th anniversary team in 2021, those interested in basketball have rightfully questioned the inclusion criteria of certain players. Furthermore, many of those asked to vote on which players should be included in the 75th anniversary team are the same figures who, as far as I can tell, do not conduct in depth statistical analysis on player performance. Most of the time, the amount of accolades, like MVPs and championship rings, are used to judge a player's success. I, however, believe there is something to be said for a player's ability to play the game of basketball. I am pairing my own statistical analysis with a human subject based study that judges familiarity to see if the most recognizable players match the most statistically productive players. I think that the outcome of this study will most likely show that there might only be a weak correlation between familiarity and statistical production because there are plenty of players who are top 20 all time in points, assists, and rebounds that most people who regularly watch the NBA wouldn't be able to name.

Methods Summary:

Data:

There are two datasets being used in this research. The first comes from BasketballReference.com and contains season average stats for all NBA players from the 1981 to 2022 season (Sports Reference LLC, 10/19/2022). I chose this dataset because it contains the components of my variables of interest (Player Efficiency Rating and GameScore) and it also has more than 17000 observations after I filtered it with my inclusion criteria (Players who have played less than one season and less than 20 games in a season). The second dataset I am using is observations from survey responses. This data will come from a participant's answers to a short questionnaire (In appendix of this paper). I am accounting for bias in these surveys by using a partial order knowledge structure to grade participants' expertise (Desmarais, 1995). These observations will yield an ordered list of NBA players who were recognized by the participants the most. I will also be asking the participants some follow up questions about where they have seen the players they recognized and if they know any other information about a player they recognized. This data will be kept anonymous and confidential.

Methods:

To generate two ranked lists to compare, I will get one from the survey responses after I aggregate the responses into counts. The second ranked list will come from a random forest aided feature selection. I will then create a new stat based on the results on the random forest and rank players based on their career total. Once these two lists have been generated. I will be using Rank Based Ordering to compare the two lists (Webber, 2010). If the two lists have greater than a 0.5 RBO score, it can be said that there might be some correlation between familiarity and statistical production (Joshi, 2021).

Early Results:

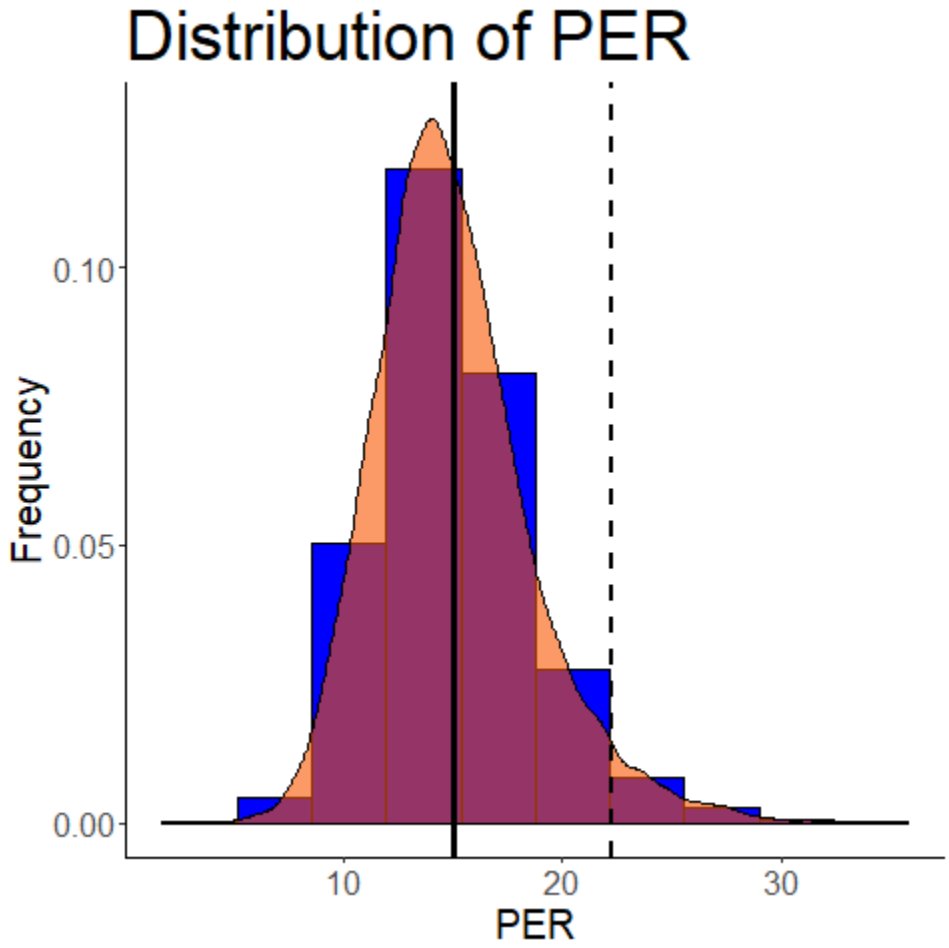
Now that I have laid out my aim and process of this project, I want to explore some of my early findings and areas of concern. First and foremost, I want to take a look at a summary of my variables of interest:

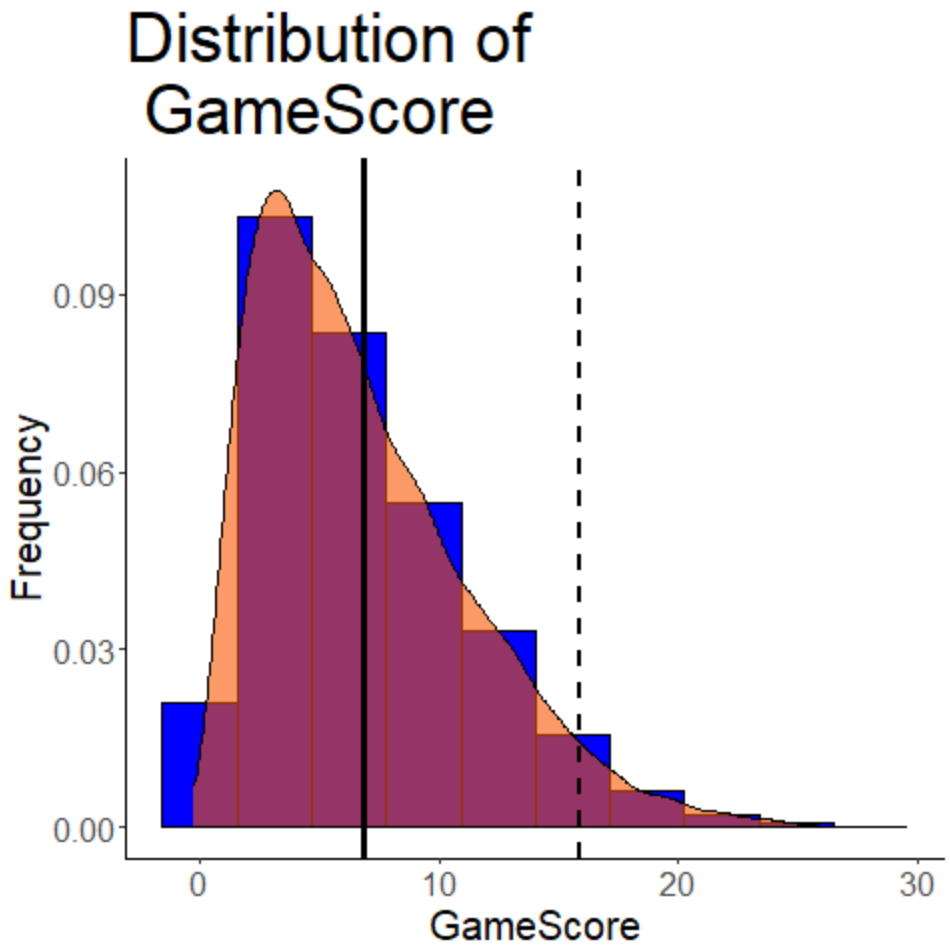
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	17051	26.828	4.058	18	24	30	43
PTS	17051	9.321	5.892	0.3	4.7	12.75	37.1
AST	17051	2.115	1.905	0	0.8	2.8	14.5
TRB	17051	3.932	2.509	0.1	2.1	5.2	18.7
FGpct	17051	0.453	0.061	0.152	0.414	0.489	0.778
TOV	17051	1.366	0.801	0	0.8	1.8	5.7
PF	17051	2.076	0.773	0.2	1.5	2.6	4.8
BLK	17051	0.459	0.516	0	0.1	0.6	5.6
STL	17051	0.736	0.459	0	0.4	1	3.7
FTpct	17051	0.735	0.117	0	0.676	0.815	1
GameScore	17051	6.864	4.505	-0.23	3.345	9.48	27.84
PER	17051	15	3.635	4.405	12.571	16.968	35.124

Variable definitions are provided in the appendix of this paper.

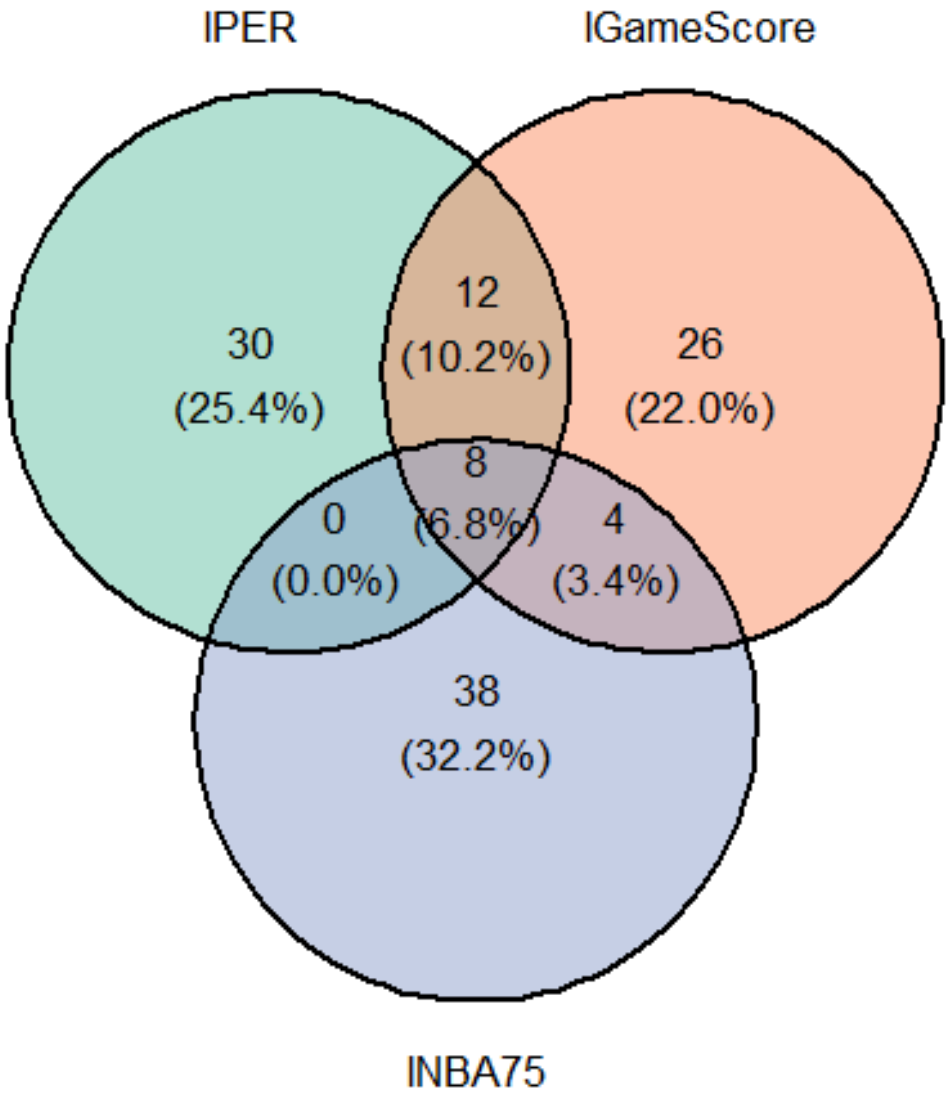
From this table, we can see that the average basic statline (Points, Assists, Rebounds, and Field Goal %) for an NBA player since 1981 is around 9.3 points, 2.1 assists, 4.0 rebounds, and 45.3% field goal percentage. The average NBA player also turns the ball over (TOV) around 1.3 times per game and accumulates about 2.0 personal fouls per game. On the defensive side of the ball, this average player records about 0.5 blocks and 0.7 steals per game. These stats in turn generate an average GameScore of 6.8 and Player Efficiency Rating of 15. However, there seems to be a quite large discrepancy between the average NBA players and the seasonal anomalies when it comes to GameScore and Player Efficiency Rating. Looking at the third quartile of both variables, we can see that 75% of all NBA players record a Player Efficiency Rating of 17.0 or

less and a GameScore of 9.5. These are not very far off from the mean PER and GameScore. On the contrary, the max PER in the dataset is 35.1 and the max GameScore is 27.48. This leads me to believe that the best statistical performers in the NBA make up less than 25 percent of the total players in my data. To show this more clearly, below is the distribution of PER and GameScore:





As you can see from the figures above, PER has a relatively normal distribution, whereas GameScore is slightly skewed to the right. The dotted lines in both plots are exactly two standard deviations from the mean. Which means that roughly 97.5 percent of players have a season average GameScore less than roughly 17 and a season average PER of less than roughly 23. The highest single season PER was recorded by Nikola Jokic in the 2022 season: 35.16. The highest single season GameScore was recorded by Michael Jordan in 1988. Given that the highest PER and GameScore are significantly larger than 2 standard deviations more than the average PER and GameScore, I decided to make two ranked lists using just PER and GameScore and compared those with the 50 NBA 75 team players I have selected for my survey. Below are the results:

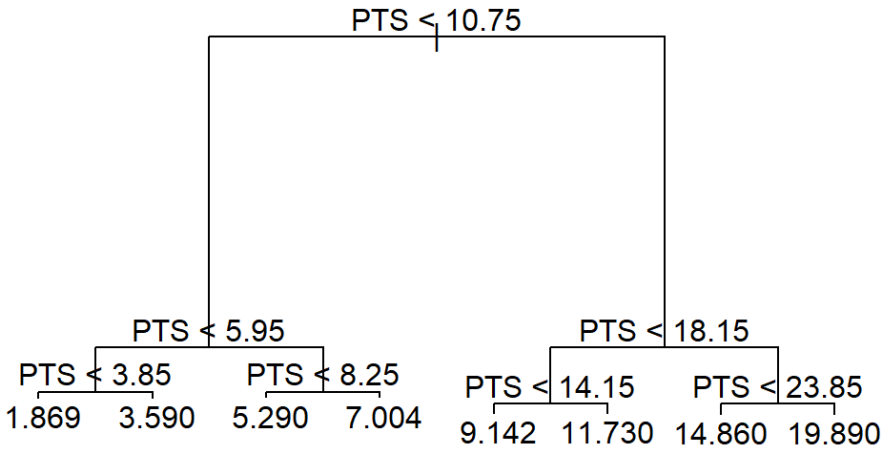


Instead of using single season PER and GameScore, I used career average. As it turns out, the top 50 career PER and GameScore list do not have much in common with the NBA 75th Anniversary team. Only 8% of the players are contained in all 3 lists. This leads me to believe that the sports writers and experts that selected the members of the anniversary team did not consider statistical production. It is important to know that PER and GameScore are just two supposed ways of measuring performance, but nonetheless, some of the players that are on the

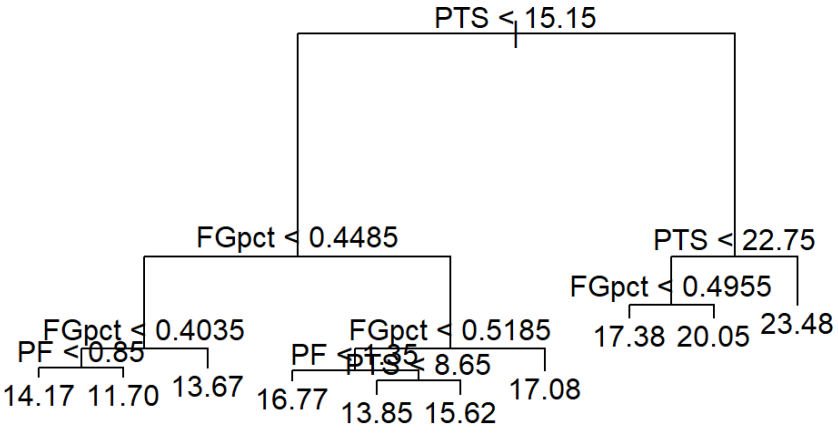
anniversary list are not statistical anomalies (Page, 2013). However, this makes sense because PER and GameScore alone do not define value; there is far more that goes into defining the statistical value of a player (Berri, 1999).

I then started on building my random forest model using the randomforest package to select the most integral features from PER and GameScore to build my own, more interpretable statistic (Garge, 2013). The first thing I did was run one single decision tree regression on PER and GameScore, below are the results:

GameScore Decision tree



PER Decision Tree



The game of basketball is not just about scoring, however, in the GameScore tree, only points per game was used to split. Furthermore, in the PER tree, points are used to split more so than any other variable. Even though I am selecting features from PER and GameScore through a random forest, these trees lead me to believe that I need to either standardize or deemphasize points in my new statistic.

Conclusion:

The main takeaways from these early results are that using the GameScore weighted format to build a stat is still easier to interpret, points are overemphasized in both PER and GameScore, the general perception of a player's performance doesn't match up to PER and GameScore, and the difference between an average player's statistical output and the top 50 highest PER and GameScore player's outputs is large. Going forward, I need to finish both random forest regressions for PER and GameScore, prune those average trees to deemphasize points, finish my collection of survey responses, and build the comparative lists. At this point in the research, it seems like there is a likely outcome of there being little to no possible correlation between familiarity and statistical production.

Appendix:

Link to github repository:

https://github.com/jake-caldwell/DA-401-Project/edit/main/README.md

Variable Table:

Abbreviation	**Name and Meaning**
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
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)
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 fouls the player committed
TRB	Total Rebounds - DRB + ORB
PTS	Points - Total points scored by a player during a game

Research Questionnaire:

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

Annotated Bibliography:

Basketball-Reference.com - Basketball Statistics and History. https://www.basketball-reference.com/. (09/21/2022)

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. http://www.jstor.org/stable/3108257

Desmarais, M. C., Maluf, A., & Liu, J. (1995). User-expertise modeling with empirically derived probabilistic implication networks. *User modeling and user-adapted interaction*, *5*(3), 283-315.

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.

Joshi, P. (2021, January 12). *RBO V/s Kendall Tau to compare ranked lists of items*. Medium. Retrieved September 29, 2022, from https://towardsdatascience.com/rbo-v-s-kendall-tau-to-compare-ranked-lists-of-items-8776c5182899

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.

Webber. (2010). *RBO: Rank biased overlap (Webber et al., 2010)*. RDocumentation. Retrieved September 29, 2022, from https://www.rdocumentation.org/packages/gespeR/versions/1.4.2/topics/rbo