OBA 410 Sports Analytics

Professor Shailesh Divey

14 March 2024

Sports Analytics Final Project

Logan Thornhill & Rahul Paudel

Introduction

The following project will be on basketball and the National Basketball Association. Basketball is a team sport in which two teams, each comprising five players, compete on a rectangular court. The primary objective is to shoot a basketball through the defenders' hoop while preventing the opposing team from shooting through their own hoop. The hoop is mounted 10 feet tall and players can score by making field goals worth 2 or 3 points depending on the distance from the hoop.

Basketball is governed by a set of rules that dictate gameplay to ensure fair competition such as dribbling, traveling, fouls, shot clock, inbounding, and many more. Basketball has grown into one of the most popular sports worldwide with millions of participants and fans across the globe. The National Basketball Association (NBA) is a professional basketball league in North America consisting of 30 teams divided into the Eastern and Western Conference. It is one of the major professional sports leagues in the United States and Canada and is considered the premier professional basketball league in the world. (Wiki). Teams compete in an 82-game regular season, in-season tournament, and playoffs to determine the league champion each year. Players can join teams through various methods including the draft, trades, and free agency. The NBA is widely popular through broadcasting, sponsorships, merchandise, and live games.

This project will delve into the intricate dynamics of the game of basketball. Through investigation into the fundamental principles of both basketball and the structure of the NBA, we are setting out to answer a central question: *How do true shooting percentage and a player's net rating affect*

a player's ability to generate points? We are aiming to uncover insights that highlight the relationship between player performance and team success, contributing to a deeper understanding of the game of basketball which is cherished by millions around the globe.

Data Set

Our data set comes from Kaggle **(**Kaggle Data**)**. This data set contains thousands of observations from decades of NBA games, so before we could begin our work, we had to clean the data to only display NBA players from 2018 to 2023.

Each row in this data set is for a unique player in the NBA, and each of them has a unique *player_name*. This data set includes a total of 21 variables, which are *player_name, team_abbreviation, college,* and *country* (**All of these are categorical variables, as they represent qualitative attributes).** The remaining variables are *age, player_height, player_weight, draft_year, draft_round, draft_number, gp, pts, reb, ast, net_rating, oreb_pct, dreb_pct, usg_pct, ts_pct,* and *ast_pct*. (**All of these are ratio variables, as they represent quantitative measurements).** *Age* represents the player's age, *player_height* represents a given player's height, *player_weight* represents a player's weight, *draft_year* represents the year a player was drafted, *draft_round* represents what round they were drafted in, *draft_number* represents the number that they were drafted, *gp* represents the number of games played in a given season, *pts* represents average points per game, *ast* represents average assists per game, *net_rating* represents a player's net rating, *oreb_pct* represents offensive rebound percentage, *dreb_pct* represents defensive rebound percentage, *usg_pct* represents a player's usage percentage, *ts_pct* represents a player's true shooting percentage, which is an adjusted shooting percentage that accounts for three-pointers and free throws, and *ast_pct* which represents a player's assist percentage.

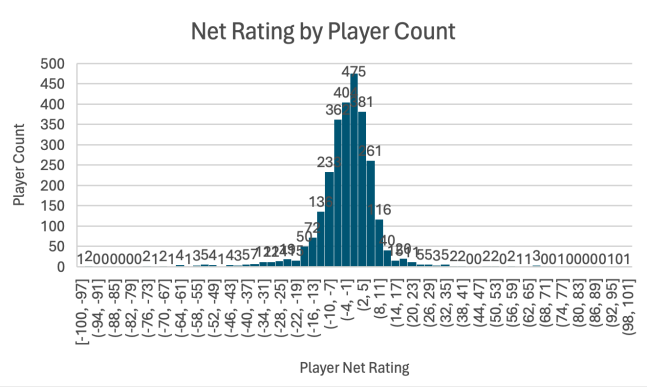
The original dataset contained data spanning back to 1996, however, the NBA has changed substantially throughout its history, and we only want to focus on the current age of the game. Therefore, we removed all player data before 2018, to get a better understanding of how player statistics affect a given team's success in our statistical analysis. To get a better understanding of both our dataset and the NBA in general, we made some frequency tables and histograms for both the categorical and ratio variables in our dataset. Below are the frequency tables we created:

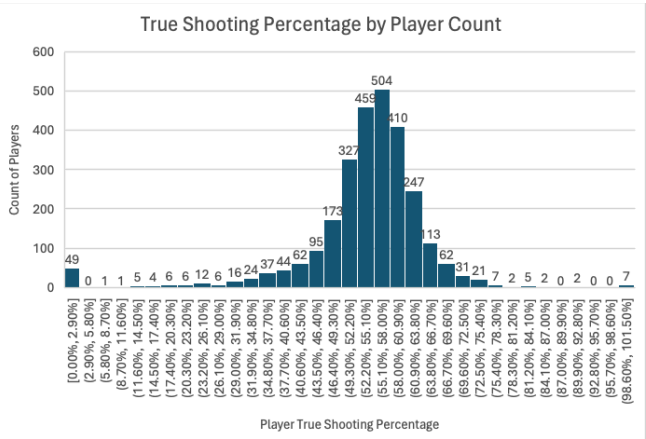
Row Labels	Count of player_name
USA	2117
Canada	96
France	57
Australia	47
Germany	34
Serbia	27
Spain	25
Croatia	23
Turkey	21
Nigeria	16
Brazil	15
Latvia	15
Lithuania	14
Slovenia	14
Cameroon	13
Greece	13
Bahamas	12
Japan	11
Ukraine	10
Italy	10
Senegal	9

Row Labels	Count of player_name
None	406
Kentucky	142
Duke	125
North Carolina	63
UCLA	63
Kansas	62
Arizona	59
Michigan	53
Virginia	42
Washington	42
Villanova	41
Indiana	40
Southern California	38
Gonzaga	36
Oregon	36
Michigan State	36
Florida	34
Florida State	33
Iowa State	32
Texas-Austin	31
Wake Forest	31

The goal of our frequency tables was to gain a better understanding of where NBA players come from. Because of this, we decided to make frequency tables for both countries of origin and college. According to the data, NBA players come from all corners of the globe, with the most coming from the United States, followed by Canada, France, Australia, and Germany between 2018 and 2023. As college goes, 14.8% of NBA players do not attend college, however, the colleges with the most representation in the NBA are Kentucky, Duke, North Carolina, UCLA, and Kansas.

For our histograms, we wanted to get a better understanding of the statistics that contribute most significantly to a player's ability to score points. We believe that the most significant statistics are net rating and true shooting percentage. Below are the histograms we created:





From the *Net Rating by Player Count* histogram, we found that 68.17% of NBA players have a net rating between -11 and 3.3 per game, while the most elite players average a net rating of more than 20, representing only 1.7% of the NBA player population. The highest-net rating players between 2018 and 2023 are Luka Doncic, Damian Lillard, Stephen Curry, Shai Gilgeous-Alexander, and Bradley Beal. A player's net rating measures a team's point differential per 100 possessions while a specific player is on the court (NBA).

For the *True Shooting Percentage by Player Count* histogram, we wanted to take a look at how true shooting percentage (TS%) might affect player points and net rating. TS% is a somewhat controversial NBA statistic, which is adjusted for three-pointers and free throws, and it essentially measures a player's efficiency at shooting the ball. It is calculated by equaling half of the points scored divided by the sum of the field goals attempted plus .475 times the free throws attempted (Wolfram). The vast majority (77.29%) of NBA players have a TS% between 46.4% and 63.8%, while the most elite NBA players have TS%s approaching the seventies. The NBA players with the highest TS% include Dereck Lively II, Daniel Gafford, Nick Richards, Grayson Allen, and Onyeka Okongwu.

$$TS\% = \frac{0.5\,PTS}{FGA + 0.475\,FTA}$$

We believe that the two statistics of Net Rating and TS% will help us the most in answering our research question. This is because Net Rating is a fundamental metric in the NBA that directly reflects a given player's offensive contribution to their team. Elite scorers play a crucial role in determining whether their team wins or loses, as they can put the team on their back offensively and create scoring swings. Along with this, TS% is a more comprehensive measure of point scoring that accounts for all forms of shooting. Unlike traditional shooting percentages, TS% more accurately assesses a player's scoring efficiency by considering the value of the shot that they are taking. A high TS% player indicates an elite scorer who can maximize their output and minimize wasted possessions. Understanding the characteristics of these variables helps analysts assess player performance, and make more informed decisions involving strategy and how the team should be managed.

Next, we conducted a correlation matrix to determine which variables are highly correlated with one another. We looked at the following variables from our dataset: age, player_height, player_weight, gp, pts, reb, ast, net_rating, oreb_pct, dreb_pct, usg_pct, ts_pct and ast_pct. The correlation matrix is below:

Correlation Matrix

	age	player_height	player_weight	gp	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct
age	1												
player_heigh	-0.013329	1											
player_weigh	0.125716	0.763548	1										
gp	0.121771	0.043115	0.065089	1									
pts	0.114942	-0.004472	0.045762	0.538275	1								
reb	0.104617	0.452176	0.471215	0.476961	0.64286	1							
ast	0.172757	-0.296169	-0.205479	0.392108	0.726671	0.391447	1						
net_rating	0.101036	0.00718	0.012527	0.200857	0.192343	0.163322	0.146122	1					
oreb_pct	-0.052459	0.5688	0.546805	-0.018785	-0.09716	0.413161	-0.228308	0.052594	1				
dreb_pct	0.040486	0.594869	0.584726	0.066482	0.108647	0.627544	-0.038695	0.053016	0.523371	1			
usg_pct	-0.007535	-0.077582	-0.024197	0.174893	0.696103	0.307203	0.509319	0.084518	-0.077887	0.094775	1		
ts_pct	0.108949	0.190957	0.206371	0.374006	0.339073	0.346235	0.144477	0.264617	0.149797	0.165076	0.119114	1	
ast_pct	0.116612	-0.425819	-0.312619	0.124762	0.411182	0.102894	0.795119	0.061912	-0.255579	-0.087417	0.46367	0.012467	1

The variable pairs with the highest positive correlations are the following:

1. ast_pct & ast (.795)
2. player_height & player_weight (.764)
3. ast & pts (.727)
4. usg_pct & pts (.696)
5. reb & pts (.643)

The variable pairs with the lowest correlations are the following:

1. ast_pct & player_height (-.426)
2. ast_pct & player_weight (-.313)
3. ast & player_height (-.296)
4. ast_pct & oreb_pct (-.256)
5. oreb_pct & ast (-.228)

The correlation matrix is a helpful source of analysis to see how variables relate to one another. A notable finding from the positive correlations is the high correlation between player height and player weight. The taller a player is, the more likely a player is to be heavier. Another notable finding is the positive correlation between assists and points. The more assists a player has, the more likely the player is to have more points per game. A notable finding from the negative correlation is between assists percentage and player height and weight. This finding could support the claim that players who are taller and heavier are more likely to have fewer assists, and this makes sense as the largest players on the court, centers, are rarely passing the basketball.

When looking at the table through the lens of our research question, we can gain a better understanding of how true shooting percentage, net rating, and points scored relate to each other. Average points scored has a relatively weak positive linear relationship with net rating (.192) and has a moderate positive linear relationship with true shooting percentage (.339). True shooting percentage has both a

moderate positive linear relationship with average points scored (.339) and with net rating (.265). This shows that both average points scored and net rating can potentially have a positive correlation with true shooting percentage. It was helpful to gain all of this insight before beginning our multiple regression analysis.

Analysis

Research Question: How do TS% and Net Rating affect a player's ability to generate points?

To answer this question, using a multiple regression model is most fitting for several reasons. Multiple regression allows us to analyze the relationship between more than one independent variable (TS% and Net Rating) with a single dependent variable (Points). We are trying to find out how both of these variables affect a given player's ability to generate points. A logistic model is used when dependent variables are either binary or categorical, and in our case, TS% and Net Rating are continuous. Even if we were able to force-fit a logistic model, interpreting the results and coefficients would be extremely challenging, and we might struggle to come to a logical conclusion.

A multiple regression model also allows us to integrate dummy variables, which lets us include categorical variables in our analysis. For our dummy variables, we wanted to learn more about how a player's age might affect Net Rating and TS%, therefore we included the dummy variable *IsOver25,* where players who are over the age of 25 are assigned a 1, and those who are younger are assigned a zero, to get a better idea of how age might affect Net Rating and TS%.

Below are the regression results from our model:

Multiple Regression Model

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.37247181
R Square	0.13873525
Adjusted R Square	0.13684651
Standard Error	5.97717752
Observations	2743

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	15745.57623	2624.262706	73.45392383	3.36801E-85
Residual	2736	97748.11731	35.72665106		
Total	2742	113493.6935			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**0.4234559**	1.261701347	0.335622924	**0.737180968**	-2.050527748	2.89743954	-2.0505277	2.89743954
net_rating	**0.03029629**	0.01004002	3.017552634	**0.002571733**	0.010609502	0.04998307	0.0106095	0.04998307
ts_pct	**16.0516671**	1.293402591	12.41041821	**1.90546E-34**	13.51552263	18.5878115	13.5155226	18.5878115
age	**-0.04166201**	0.046005969	-0.905578293	**0.365238841**	-0.131871955	0.04854794	-0.131872	0.04854794
isOver25 (1 if yes, 0 if No)	**1.10470554**	1.202714449	0.918510243	**0.358432763**	-1.25361474	3.46302582	-1.2536147	3.46302582
IsO25*Net-Rating	**0.0385983**	0.016527483	2.335400864	**0.019594413**	0.00619069	0.07100591	0.00619069	0.07100591
IsO25*TS%	**1.0149694**	2.09287316	0.4849646	**0.627740282**	-3.088802055	5.11874084	-3.0888021	5.11874084

As we can see from the model, the P-values for Net Rating, TS%, and isO25*Net Rating dummy variable are all highly statistically significant as their P-values are far less than .05. This means our hypothesis is true, and all of these independent variables significantly affect the points that a given NBA player will score. Due to the results of our model, we will not be including the variables that are not statistically significant. Below is the equation for our multiple regression analysis:

POINTS = .03(net_rating) + 16.05(ts_pct) +.04(isO25*net_rating)

We decided to include these three independent variables because they were the most statistically significant in our model. TS% has the P-value closest to zero, meaning that it is the most statistically significant variable in our dataset when observing point outcomes. Along with this, a high F-value (73.45) and a very low P-value indicate that our regression model as a whole is statistically significant.

Conclusion

In conclusion, our analysis aimed to understand how Net Rating and TS% affect a player's ability to generate points in the NBA. By closely examining 5 years of player data from 2018 to 2023, we found significant insights. In our multiple regression analysis, we found that both Net Rating and TS% have a substantial impact on a given player's ability to generate points for their team. More specifically, for every point increase in Net Rating and TS%, we found a corresponding increase in points scored. We also went deeper, investigating the influence of how age might impact point scoring, finding that players who are over 25 years old also positively impact point generation.

Moving forward, our analysis suggests that coaches and GMs can substantially benefit from focusing on improving players' shooting efficiency (TS%), and overall on-court impact (Net Rating), in order to reach the highest scoring and winning potential. By leveraging these findings, teams will be able to make more informed decisions in player evaluation, strategy, and roster construction, ultimately leading to improved player performance. Overall, this study provides valuable insights into the relationship between player performance metrics and team success in basketball, leading to a deeper understanding of the game and informing decision-making processes in the NBA and beyond.

Works Cited

Basketball True Shooting Percentage: Wolfram Formula Repository. (n.d.). Retrieved from https://resources.wolframcloud.com/FormulaRepository/resources/Basketball-True-Shooting-Percentage#:~:text=The%20true%20shooting%20percentage%20is,times%20the%20free%20throws%20attempted.

National Basketball Association. (2024). Retrieved from https://en.wikipedia.org/wiki/National_Basketball_Association

NBA. (n.d.). Retrieved from https://www.nba.com/stats/help/glossary

Players Advanced Leaders: Stats. (n.d.-a). Retrieved from https://www.nba.com/stats/players/advanced-leaders