Spring 4-9-2024

Optimizing NBA Roster Construction

Nick R. Riccardi
Syracuse University

Recommended Citation

Riccardi, N. (2023). Optimizing NBA Roster Construction. Academy of Economics and Finance Journal, 14, 28-38.

Optimizing NBA Roster Construction

Nick Riccardi, Syracuse University

Abstract

This study aims to quantify the effect that complementary player types have on team success in the National Basketball Association. Using cluster analysis, player-seasons are redefined from their traditional basketball positions to better encompass the roles that players play. For the 10 seasons of data, the best player for each of the 30 teams in the league is determined and teams are grouped based on the cluster of their best player. Ordinary Least Squares regressions are performed to test what player types fit together best. The results of this study show the importance of complementary workers to a firm's success.

JEL Codes: C1, J0
Keywords: team performance, NBA, clustering

Introduction

The success of a professional sports team is dependent on, among other factors, the level of talent that the team possesses and how the players on the team fit together. The National Basketball Association (NBA) is no exception to this and perhaps the most reliant on these two factors given that only five players for a team are on the court at once and a team's best players play most of the game. Therefore, it is vitally important for NBA general managers to acquire players who are good enough to bring the team success while also fitting together well on the court. Traditional basketball mindsets can make this task more daunting than it already is. With how the game is currently played in the era that emphasizes three-pointers[i] and the use of analytics, players are playing in less position-specific roles than they have before. This means that general managers need to assess the specific style in which a player plays when they evaluate them.

Not only do general managers need to be able to determine the playstyle of a player, but they also need to be able to figure out which other playstyles fit well with that style. Building an NBA team begins with the process of putting complementary players around the team's best player. In this study, player positions will be redefined based on the styles of players across the league. Using cluster analysis, player-seasons will be grouped together based on similar statistics. Linear regression models will be conducted with the purpose of figuring out which playstyles work well with others. The goal is to determine the types of players that complement specific styles of play in order to figure out how to build around the style of a team's best player.

Literature Review

Positions in basketball have always followed the traditional specification of having five players on the court, usually a point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C). McMahan (2018) states that this declaration of categorizing players into these five positions is the result of overall NBA strategy. Definitions of the roles for each of these positions are provided on the NBA's website ("Basketball Positions," nd). A point guard is a player who "runs the offense and usually is the team's best dribbler and passer. The point guard defends the opponent's point guard and tries to steal the ball." Shooting guards are described as "usually the team's best shooter. The shooting guard can make shots from long distance and also is a good dribbler." A small forward is perhaps the most versatile of the five positions, given that he "plays against small and large players. They roam all over on the court. Small forwards can score from long shots and close ones." The next position, a power forward, "does many of the things a center does, playing near the basket while rebounding and defending taller players. But power forwards also take longer shots than centers." The last of the five traditional positions is the center, which "is the tallest player on each team, playing near the basket. On offense, the center tries to score on close shots and rebound. But on defense, the center tries to block opponents' shots and rebound their misses."

The increase in overall athleticism and skill of players coupled with the use of analytics in today's modern NBA has changed the way these traditional positions are viewed. Now, with the increase in the importance of three-point shooting and being able to switch defensive assignments, basketball has become a positionless game. The current NBA commissioner, Adam Silver, shared this sentiment before Game 1 of the 2022 NBA Finals, saying, "We're a league that has moved increasingly to positionless basketball" (Aschburner, 2022). This is in part due to players like Dirk Nowitzki, a 7'0" power forward who could shoot from long distances with the same skill as guards. In today's NBA, power forwards and centers,

also known in combination as "bigs," are not only encouraged to shoot three-pointers, but without the skill, can become obsolete. Over the past two decades in the NBA, teams have been taking more three-pointers each year, with a major increase in threes attempted per game starting in 2013-2014 (Shea, nd). For this reason, the center position, generally assigned to the tallest players in basketball, is dwindling. Ziller (2017) notes that this change in playstyle has led to the typical duties of a center no longer being needed.

The NBA is in need of redefining positions by classifying players by their player type, not the traditional position they play. Many researchers have attempted to solve this problem through machine learning techniques, such as cluster analysis. Cluster analysis can be used to group together players based on similarities of statistics. Having the goal of finding complimentary player types in mind, only research considering team success in relation to clustering NBA players will be discussed. One such example comes from Kalman and Bosch (2020), who modeled lineup efficiency after clustering 3,608 NBA players from 2009-2018. The statistics chosen for clustering in this paper were based on skills, habits, and opportunity. Their clustering algorithm resulted in nine different clusters of players. From there, the authors modeled five-man combinations of clusters to predict Net Rating, which is the scoring differential per 100 possessions. A linear regression model was conducted to determine the effect that the number of players from each of the nine clusters within the combination of five players has on Net Rating. Lastly, a Random Forest Model was used to predict the Net Rating of all possible five-man combinations of clusters.

Zhang et al. (2018) clustered 354 players from the 2015-2016 season based on experience, weight, and height which resulted in five clusters described by these three factors. The authors then analyzed the distribution of clusters across different levels (based on performance) of teams. Patel (2017) also used one season of data to cluster 486 players from the 2016-2017 season based on per-100-possession stats. The clustering resulted in four clusters of players, which the author described as "The Paint Protectors," "The Supporters," "The Shooters," and "The Insiders." The author did not find significant relationships between team success and the cluster membership of a team's players. However, a significant result was found regarding the distance of players from their cluster centroid and team success.

Duman et al. (2021) clustered players within their traditional basketball position. That is, within each of the five traditional basketball positions, clustering was performed to distinguish the player types within a position. Four different clusters were created amongst players who were identified as point guards, shooting guards, and small forwards, respectively, while five clusters were created for power forwards and six for centers. The authors found the clusters of each of the five positions that were part of the most successful teams as well as pairs of clusters across two traditional positions. Osken and Onay (2022) used the clustering of NBA players to predict the outcome of NBA games. In this paper, players from the 2012-2013 to the 2017-2018 seasons were clustered using box score, advanced efficiency, and shot selection data. The authors predict the winners of NBA games using an artificial neural network that takes into account the minutes played by each cluster for each team as well as factors such as win percentages of teams, month of season, and days of rest for teams. The prediction accuracy was greater than 75%.

Furthermore, Tsai (2017) clustered all NBA players who averaged one shot per game or more from 2010 to 2016. First, the author clustered players using cumulative shot chart data. The shot charts were converted into heat maps which were then converted into a data matrix of players' shots. K-means clustering was then used on this matrix, and it was determined that seven clusters were the optimal amount. A second cluster analysis was based on player performance statistics, 16 in total, that were taken from the NBA's website. This clustering led to eight different groups. The author then took each of the matrices created for the two k-means clustering analyses and combined them into one. This matrix revealed an R-squared value of .71 in terms of its correlation to predicted winning percentages for teams throughout the league.

This study adds to the existing literature regarding clustering NBA players and team success by evaluating team success under the condition of the player type of the team's best player. This context is important because teams do not always have access to the types of players they desire and need to optimize their roster under the constraints of available players and assets. Additionally, the same makeup of players in a lineup will inherently be different depending on the role that the team's best player plays. For example, if a team's best player is a pass-first player, then team success will be highly dependent on the ability of the players around the best player to score off their passes. On the other hand, if the best player on a team is a score-first, high-shot attempt player, then the players around him might need to be good at setting screens and providing space on the court for the best player to get shots off. These two examples might have the same makeup of playerstyles on the team, but without the context of the playstyle of the best player, the distinction can't be made on whether the playstyles are a good fit. Furthermore, this study includes the 10 most recently completed regular seasons of the NBA. These 10 seasons begin with the start of the current three-point era that the NBA is in. Therefore, this study is a novel approach to finding the relationship between player types and team success in the NBA by considering the differences in complementary player types based on the player type of a team's best player as well as including the most recent data available.

Data and Methodology

The data gathered for this study are in the form of player-seasons gathered from basketballreference.com. That is, each row of data represents a given NBA player during a given season. This includes per-game, shooting, advanced, totals, and play-by-play statistics for every player from the 10 seasons between 2013-2014 and 2022-2023. The per-game statistics represent typical box score data such as points, assists, rebounds, blocks, steals, field goal percentage, and more. The totals statistics encompass the same statistics as per-game but are sums for each player for the entire season. To get an understanding of where and how players are taking their shots, the shooting statistics shine a light by providing the percentage of shots each player takes from five distance ranges, their field goal percentage from each range, the percentage of their shots that are assisted, and more. The distance ranges include zero to three feet away from the basket, three to 10 feet, 10 to 16 feet, 16 feet to the three-point line, and three-pointers. Advanced statistics include rate statistics, such as rebounding percentage, assist percentage, and steal percentage, as well as metrics derived to encompass a player's value on offense, defense, and in total. The author invites the reader to access the glossary provided by Basketball Reference which includes definitions for many of the statistics gathered ("Glossary," nd). An example of a definition for a rate statistic is that of TRB% (Total Rebound Percentage), which is "an estimate of the percentage of available rebounds a player grabbed while he was on the floor." Lastly, play-by-play statistics were recorded to get an estimate of where each player plays a majority of their minutes based on traditional player positions. Based on these estimates, each player will be assigned the position that they most frequently play.

Limitations were imposed on which player-seasons would be included in the analysis. Any player who did not play at least half of the season (41 games in a normal NBA season[ii]) as well as at least 10 minutes per game were taken out of the dataset. The number of games played minimum requirement was implemented because if a player has a small sample size, their effect on a team's winning percentage could be disproportionate. As far as the minutes per game minimum, the goal was for players in the analysis to have an established role on their teams. Playing at least 10 minutes per game warrants being part of a consistent rotation.

These restrictions limited the dataset to 3,145 players over 10 seasons. With these player-seasons, a k-means clustering algorithm was performed with the goal of redefining the traditional player positions into functional player types. K-means clustering "aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster" ("K-means clustering," nd). The algorithm works by minimizing within-cluster variances, or squared Euclidean distances. To start, centroids are randomly selected to serve as beginning points for clusters and then iterative calculations are performed for centroid positioning optimization (Medium, 2018). Since k-means clustering is an unsupervised algorithm, the results need to be interpreted by examining what similarities observations have within each cluster as well as what distinguishes the observations in each cluster from the other clusters.

Out of the performance statistics gathered for the player-seasons, the statistics used for analysis were total rebounding percentage (TRB%), assist percentage (AST%), steal percentage (STL%), block percentage (BLK%), turnover percentage (TOV%), standard deviation of the proportion of shots taken from each distance range (SDSH), shot percentage from each distance range, percentage of two-point makes that were assisted, percentage of three-point makes that were assisted, free throw rate, and personal fouls per game. Rate statistics were used rather than their per-game or totals counterparts to eliminate the opportunity that players get. More important than the counts that players rack up for these statistics is how often they occur when the player is on the floor.

Before the cluster analysis was conducted, each variable was standardized to not disproportionately influence the separation of observations based on differences in scales of variables. Given the large number of variables used for clustering, Principal Component Analysis (PCA) was used to reduce the number of dimensions. PCA is a technique for feature extraction, which creates new variables (the number of which is the same as the number of the original variables) that are combinations of the variables supplied (Brems, 2017). The created variables, called components, are ordered by the proportion of variance they explain between the observations. The first component explains the most amount of variance while the last explains the least amount. Dimensionality reduction is achieved by taking the number of dimensions that account for a certain threshold of cumulative variance explained. In this case, the threshold was set at 75% of the variance explained which was accomplished by the first seven principal components. With these seven components, k-means clustering was performed. The optimal number of clusters, six, was determined by the balance of the count of observations in each cluster and the between sum of squares accounting for 52.1% of the total sum of squares.

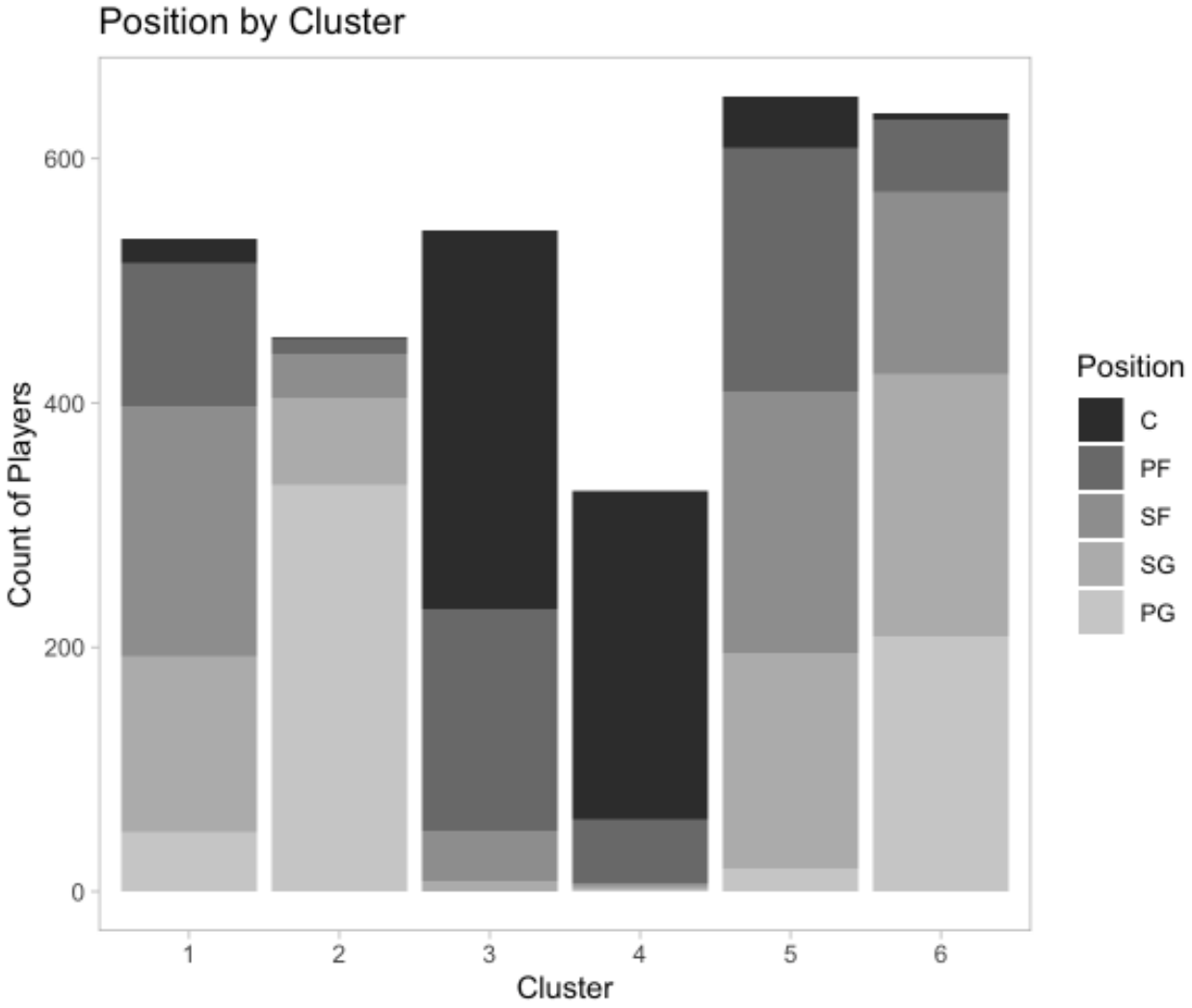
After the cluster analysis was conducted and each player-season was assigned to one of the clusters, Ordinary Least Squares regression models were run to figure out how each cluster impacted a team's success depending on which cluster the team's best player came from. Team success was determined by Pythagorean win percentage, which was the dependent variable in the models. Pythagorean win percentage is an estimate of a team's strength based on the number of points they score and allow over a season. A team's best player was based on which player had the greatest Real Plus-Minus (RPM). RPM is an advanced statistic created by ESPN that provides a "Player's estimated on-court impact on team performance, measured in net point differential per 100 offensive and defensive possessions" (ESPN, nd). Limitations were put into place for which players could be declared a team's best player. The best player for each team was selected from the players on that

team who played at least 25 minutes per game. This was implemented to prevent role players from being selected as the best player on a team. By playing limited minutes, a player's RPM may not realistically account for their contribution to their team's success. Teams were then grouped together based on which cluster their best player resided in. The independent variables in each of the models were the minutes played by a team's best player and the sum of the minutes played by each cluster for the rest of the players for a team.

Cluster Analysis

The first step in analyzing the results is getting an understanding of what separates the observations into their respective clusters. Figure 1 provides the first glance at the clusters by showing the breakdown of traditional positions by cluster. While the goal of the clustering was to move from traditional positions to player types, it is still informative to analyze the positional breakdown.

Figure 1: Traditional positions breakdown for each cluster



Clusters one and five have similar compositions, being mostly occupied by shooting guards, small forwards, and power forwards. Almost three-fourths of the second cluster is comprised of point guards, while only 14 of these 454 players are bigs (13 power forwards, one center). Clusters three and four, on the other hand, are made up of mostly bigs, with cluster three having 90.8% of players being bigs with zero point guards and cluster four having 97.7% bigs (82% centers). Cluster six has a distribution that's not as refined as clusters two, three, and four while not being as spread out as clusters one and five. Roughly 90% of this cluster is made up of point guards, shooting guards, and small forwards, with guards accounting for about two-thirds of the players within the cluster and almost split perfectly between the two guard positions.

Next, visualizing the shot profile of the clusters helps to understand a major aspect of the offensive side of the game. Figure 2 below shows the mean proportion of shots taken from each distance range for the six clusters. Like the positional breakdown, clusters one and five also have similar shot distributions. These two clusters take the highest proportion of shots from beyond the three-point line (0.466 for cluster one, 0.505 for cluster five), and the smallest proportion of shots between three feet and the three-point line (0.264 for cluster one, 0.296 for cluster five). Clusters two and six also have similar shot distributions, sporting the two lowest SDSH, showing that they take shots from all over the court. The greatest difference in proportion between these two clusters across any distance range is just 0.046 in the zero to three feet range. Lastly, clusters three and four take the shortest distance shots across the six clusters. While cluster three is relatively balanced (third lowest SDSH), over 60% of shots from these players come from within ten feet of the basket. Still, this number pales in comparison to the 83.7% of shots that come from within 10 feet for cluster four. Additionally, while cluster three still takes over one-fifth of their shots from beyond the three-point line, cluster four players only have a proportion of 0.006 in that range.

Figure 2: Proportion of shots taken from each distance range by cluster

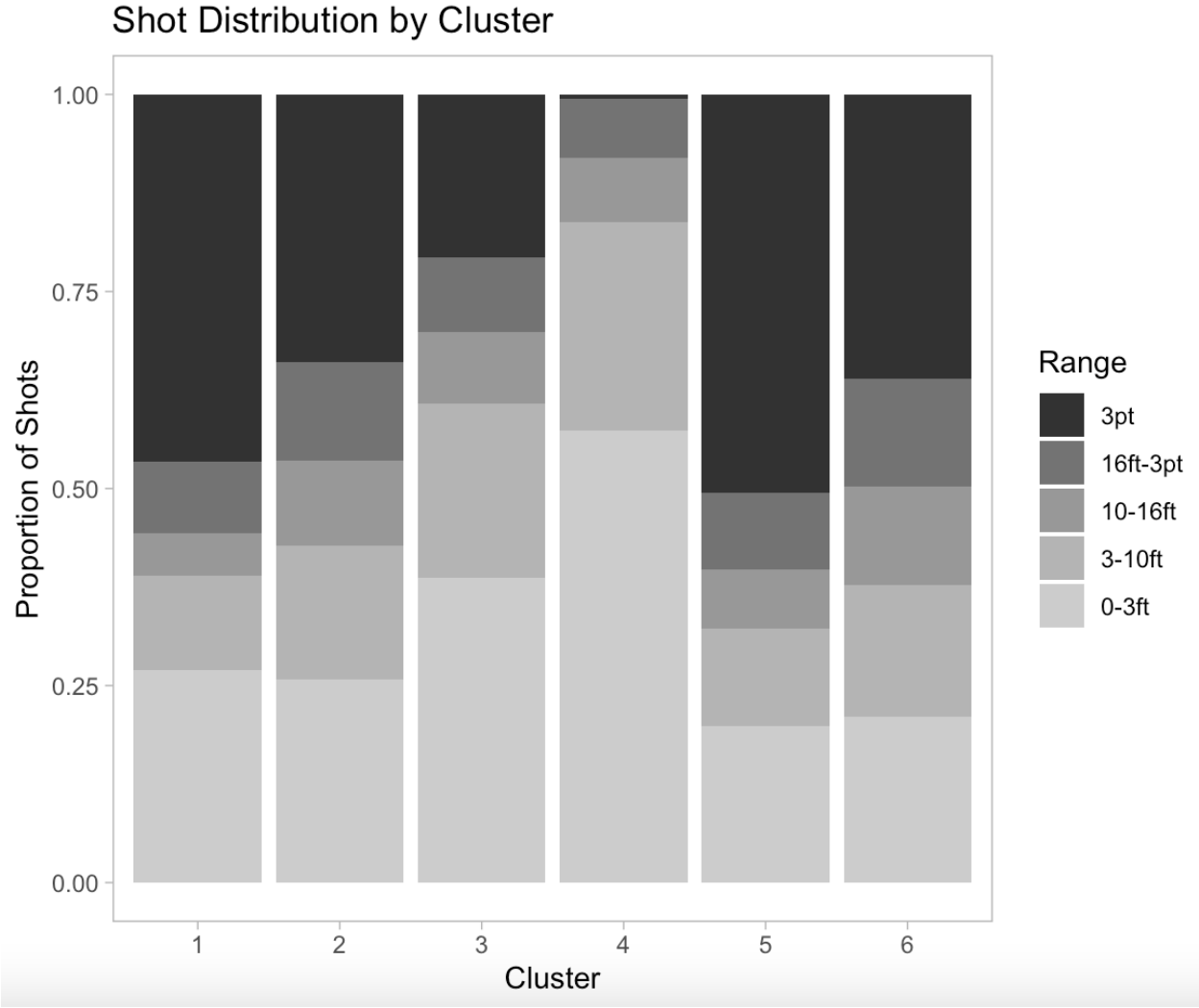


Table 1 provides the means of variables by each cluster. Bolded variables are those that were explicitly used in the dimensionality reduction and subsequent cluster analysis. The appendix provides definitions for each of the variables listed. Figure 1, Figure 2, and Table 1 provide the information necessary to describe the players in each cluster and give them labels based on their functional roles. For cluster one players, it is easier to focus on the negative aspects of their averages rather than the positives, considering how few there are. Cluster one players are the worst offensively out of the six clusters. These players rank last in overall field goal percentage, having the lowest field goal percentages in the three ranges between zero and 16 feet from the basket and second to last between 16 feet and the three-point line.

Table 1: Variable means by cluster

Variable	Cluster One	Cluster Two	Cluster Three	Cluster Four	Cluster Five	Cluster Six
TRB%	8.54	7.48	14.8	17.2	8.33	7.23
AST%	9.92	28.9	11.6	8.22	8.57	18.4
STL%	1.73	2.17	1.39	1.32	1.27	1.46
BLK%	1.37	0.949	3.25	3.70	1.17	0.853
TOV%	12.2	16.2	12.5	14.5	9.39	11.8
SDSH	0.200	0.136	0.166	0.246	0.201	0.123
FG% 0-3ft	59.8	61	70.1	67.5	67.2	62.7
FG% 3-10ft	29.7	38	43.4	40.5	41.5	41.6
FG% 10-16ft	29.9	39.9	39.5	36.2	41.5	43.1
FG% 16ft-3pt	31.6	38.3	37.4	26.5	38.3	40.9
3P%	34.5	33.3	32.4	0.96	36.5	35.5
% 2P Ast'd	57.2	27.9	65.5	68.7	66.1	37.2
% 3P Ast'd	92.4	69.1	95	2.5	94.9	78.9
FT Rate	0.219	0.276	0.316	0.404	0.176	0.237
PF/G	1.84	2.07	2.51	2.33	1.65	1.89
Min/G	21.8	27.7	25.3	21.7	22.2	27.4
PTS/G	7.71	13.6	12.5	8.29	9.03	14.1
TRB/G	3.37	3.85	6.79	6.91	3.31	3.65
AST/G	1.51	5.28	1.97	1.19	1.33	3.24
STL/G	0.772	1.21	0.719	0.578	0.582	0.81
BLK/G	0.346	0.317	0.945	0.96	0.3	0.28
FGA/G	6.63	11	9.43	6.06	7.39	11.5
FG%	41.5	43.5	51.7	56	44.5	44.4
3PA/G	3.12	3.71	1.97	0.025	3.66	4.09
FTA/G	1.45	3.26	3.01	2.4	1.33	2.84
USG%	16.3	22.2	20.2	16.5	16.9	22.3
Avg. Dist	15.1	13.8	9.91	4.68	16.5	14.7
% Shots 0-3ft	0.269	0.257	0.387	0.573	0.198	0.211
% Shots 3-10ft	0.12	0.17	0.221	0.264	0.124	0.167
% Shots 10-16ft	0.054	0.109	0.091	0.082	0.074	0.124
% Shots 16ft-3pt	0.090	0.124	0.095	0.074	0.098	0.137
% Shots 3P	0.466	0.34	0.207	0.006	0.505	0.361
% FG Ast'd	73.5	41.9	71.6	68.2	80.6	52.2

Their AST% is marginally better than the bottom two clusters in the statistic and their USG% ranks last, showing that teams generally do not turn to them on the offensive side of the game. However, cluster one players provide value on defense and from beyond the three-point line. The cluster averages the second highest STL%, third highest BLK% (highest among non-big-dominated clusters), and third highest three-point percentage. Players in cluster one can be considered defensive specialists. Common players in this cluster include Danny Green (6x), Kentavious Caldwell-Pope (5x), and P.J. Tucker (6x).

The most important statistic to describe cluster two players is AST% in which they average the highest percentage by a wide margin. Additionally, cluster two players average the highest STL%. Both findings are not surprising given that the cluster is dominated by point guards. These players average the lowest percentage of their field goals being assisted, the second highest USG%, the most minutes played per game. A couple downfalls to players in this cluster are that while they average the highest AST%, they also average the highest TOV%, and they are not particularly great at shooting from any specific distance range. Given their highest AST% and low assisted percentage, players in cluster two are playmakers, creating shots for their teammates and themselves. Some notable players from cluster two are Chris Paul (10x), James Harden (10x), and LeBron James (8x).

Cluster three players are skilled across many statistics. They average the second highest TRB%, third highest AST%, second highest BLK%, highest field goal percentages within 10 feet, and the second highest free throw rate. While they take almost 40% of their shots beyond 10 feet, they are not as effective as close to the basket, generally shooting about average from farther distances. Still, these players can stretch the floor with outside shooting while being highly effective close to the basket, are strong rebounders, can move the ball well, and have a strong defensive presence. Players in cluster three can

generally be thought of as versatile bigs. Examples of players from this cluster include Anthony Davis (9x), Giannis Antetokounmpo (9x), and Nikola Jokic (7x).

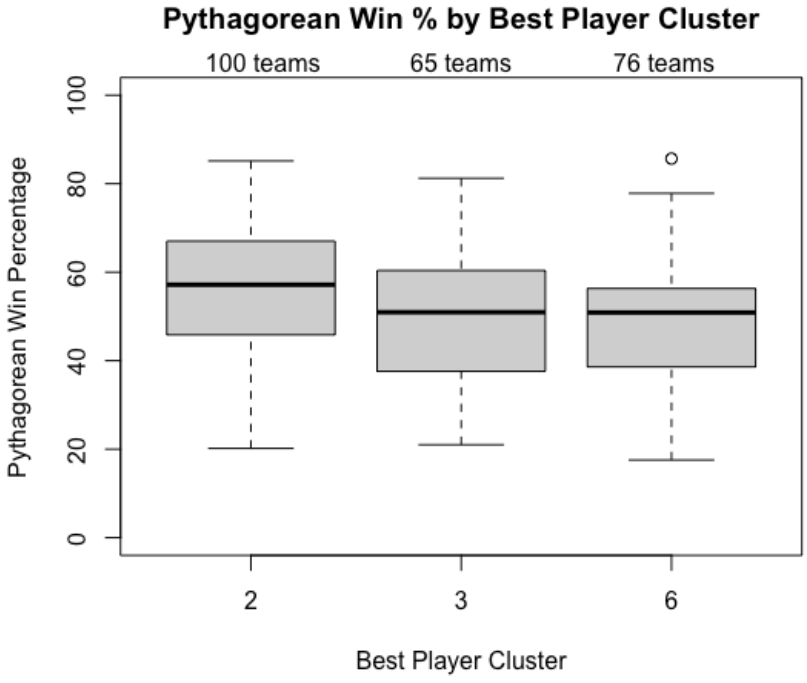
Players in cluster four can also generally be thought of as bigs, but in the traditional sense. These players fit closely to the definition of a center, highlighted by ranking first in TRB% and BLK% and rarely taking shots outside of 10 feet from the basket, hence the highest SDSH belonging to the cluster. Their average shot distance is less than five feet, they are second lowest in USG%, highest in field goal percentage, lowest in AST%, second lowest in STL%, and likely only take threes when they must heave the ball up at the end of the shot clock or quarter, evident by their extremely low three-point rate and three-point percentage. Once again, this cluster consists of traditional bigs. A few players that belong to this cluster often are Andre Drummond (9x), DeAndre Jordan (8x), and Rudy Gobert (9x).

There are a few defining characteristics of players in cluster five. Most evident of these are statistics relating to three-point shooting. These players take the greatest proportion of three-point shots while averaging the highest percentage from three as well. Players from this cluster are also effective between 10 feet and the three-point line and have the highest percentage of their field goals assisted. Other than their ability to catch and shoot the ball, they do not provide much value elsewhere. Cluster five players are low in AST% and TRB% and average the lowest STL% and third lowest BLK%. While their TOV% is the lowest of the six clusters, this is a microcosm of not having the ball in their hands for long, evident by their low USG%, low free throw rate, and high assisted percentage. Players from this cluster can be described as perimeter shot takers. Prime examples of cluster five players include JJ Redick (6x), Terrence Ross (8x), and Klay Thompson (7x).

Lastly, cluster six players have an affinity for scoring the ball. Their lowest SDSH amongst the clusters show that they are willing to take shots from anywhere on the court and they tend to be effective everywhere outside of three feet from the basket, ranking first or second in field goal percentage across the remaining distance ranges. Not surprisingly, these players take the most shots per game, average the most points, and have the highest USG%. Although they tend to take many shots, they also create for their teammates, evident by ranking second in AST%. Cluster six players tend to not be as effective on the defensive end, averaging the lowest TRB% and BLK%. Players in cluster six are well-rounded scorers. Some players commonly found in cluster six are Bradley Beal (8x), DeMar DeRozan (10x), and Jordan Clarkson (9x).

The next step in the analysis is to group teams together based on the cluster of the team's best player. As it was mentioned before, the best player on a team was determined by RPM. Given the playstyles of the six clusters, teams are inherently more likely to have their best player belong to a particular cluster. Therefore, it is unsurprising to see that 20 teams had their best player come from cluster one, 100 teams from cluster two, 65 teams from cluster three, 18 teams from cluster four, 21 teams from cluster five, and 76 teams from cluster six. Given these sample sizes, it is only appropriate to analyze teams whose best player is from cluster two, three, or six. These three clusters account for over 80% of all teams. Figure 3 presents boxplots of teams' Pythagorean win percentage grouped by best player cluster.

Figure 3: Distribution of Pythagorean win percentage grouped by best player cluster



Teams with a playmaker (cluster two) as their best player tend to fair the best out of the three groupings. Welch two-sample t-tests confirm this result. There was a significant difference in mean Pythagorean win percentage between cluster two teams (M = 55.7, SD = 14.6) and cluster three teams (M = 49.1, SD = 15.6) at the 1% level (p = 0.008) and a significant difference between cluster two teams and cluster six teams (M = 48.4, SD = 15.7) at the 1% level (p = 0.002). A Welch two-same t-test of mean Pythagorean win percentage between cluster three and cluster six teams did not yield a significant result.

Empirical Models

With an established understanding of the six clusters and teams grouped by the cluster of their best player, the last step in the analysis is to create linear regression models. Three models were specified, one each for cluster two teams (Model I), cluster three teams (Model II), and cluster six teams (Model III). The dependent variable for each model is Pythagorean win percentage and the independent variables are the total minutes played for the season by a team's best player (BPM) and the sum of minutes played for the season by each of the six clusters (denoted as C1M for cluster one minutes, C2M for cluster two, and so on). Although the dependent variable has natural lower (0) and upper (100) bounds, predictions supplied by the regressions were within the possible range of values[iii], so censoring was avoided. The following tables display summary statistics for the independent variables for the three subsections of teams. All variables are scaled in thousands of minutes.

Table 2: Cluster two teams variable summaries (in thousands of minutes)

Variable	Minimum	Median	Mean	Maximum	Std. Deviation
BPM	1.173	2.439	2.424	3.125	0.389
C1M	0	2.402	2.529	8.268	1.798
C2M	0	1.159	1.292	5.967	1.390
C3M	0	2.903	2.887	6.885	1.777
C4M	0	1.371	1.697	5.548	1.524
C5M	0	3.333	3.622	9.536	2.175
C6M	0	2.843	2.909	7.807	1.808

Table 3: Cluster three teams variable summaries (in thousands of minutes)

Variable	Minimum	Median	Mean	Maximum	Std. Deviation
BPM	1.510	2.284	2.253	3.030	0.360
C1M	0	2.396	2.502	7.757	1.773
C2M	0	2.266	2.317	5.215	1.242
C3M	0	2.041	2.028	5.946	1.495
C4M	0	0.549	0.989	4.203	1.187
C5M	0	2.780	3.022	7.340	1.740
C6M	0	4.022	3.973	10.271	2.316

Table 4: Cluster six teams variable summaries (in thousands of minutes)

Variable	Minimum	Median	Mean	Maximum	Std. Deviation
BPM	1.267	2.360	2.344	3.122	0.426
C1M	0	1.626	2.059	5.704	1.682
C2M	0	2.055	2.047	6.499	1.472
C3M	0	2.716	2.839	6.977	1.780
C4M	0	1.700	1.610	4.381	1.329
C5M	0	3.047	3.109	9.335	2.079
C6M	0	3.418	3.179	7.395	1.942

Ordinary Least Squares regressions were conducted by the specificities offered above. A Breusch-Pagan test revealed heteroskedasticity in Model III, so weighted least squares were applied to the equation as done by Yobero (2016). Table 5 presents the results of the three models.

The results of the models can best be analyzed by comparing the coefficients of significant variables. For example, in Model I, all six clusters have significant and positive coefficients, suggesting that all player types are beneficial to be put around playmakers. Individual coefficients represent the predicted percentage increase in Pythagorean win percentage by increasing minutes played by a particular cluster by 1000 minutes. However, what's important to note in terms of making

predictions is the significant negative constant. While the constant is meaningless in this context (a team cannot have a negative win percentage), it is important in making predictions.

Table 5: Regression results

	Model I (Cluster two teams)		Model II (Cluster three teams)		Model III (Cluster six teams)	
Variable	Coefficient	Prob.	Coefficient	Prob.	Coefficient	Prob.
Constant	-43.176	0.029**	-20.060	0.173	-35.781	0.026**
BPM	10.939	0.002***	16.676	0.002***	14.268	0.000***
C1M	5.011	0.000***	2.079	0.108	2.771	0.044**
C2M	4.015	0.004***	-0.739	0.602	1.911	0.159
C3M	5.776	0.000***	1.863	0.258	3.837	0.002***
C4M	4.549	0.002***	1.655	0.392	5.095	0.001***
C5M	6.154	0.000***	4.955	0.000***	3.973	0.000***
C6M	2.676	0.044**	1.940	0.042**	3.042	0.013**

*** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level

As an example, take C6M in Model I, as this is the lowest coefficient across all six clusters. In a normal NBA season, if a team does not play any overtime games, over 82 games they will total 19,680 minutes played[iv]. Using this number as well as the average minutes played by the best player on cluster two teams, if the remaining minutes on a cluster two team were given to only cluster six players, the predicted Pythagorean win percentage is just 29.5%. For context, the lowest prediction for cluster two teams was 36.2%. Contrast this with giving all the remaining minutes to cluster five players, which has the greatest coefficient of the six clusters, and the number jumps to 89.5%. While both cases are unrealistic as lineups need a balance of players, and there are natural upper limits of playing time by cluster outlined by the maximum number of minutes that each cluster received in their relative grouping, they show the predicted difference of the impact of different clusters on team success. Therefore, the results highlight the opportunity costs of giving certain clusters minutes over others. If a cluster two team were to take 1000 minutes that were played by cluster six players and gave them to cluster five players instead, their predicted Pythagorean win percentage would increase by about 3.5%, all else equal. Table 6 below shows the relative rank of clusters for each model and subsection of teams.

Table 6: Supporting clusters ranks by subsection of teams

	Cluster two teams		Cluster three teams		Cluster six teams	
Cluster	Coefficient	Rank	Coefficient	Rank	Coefficient	Rank
1	5.011	3	Insignificant	T-4	2.771	5
2	4.015	5	Insignificant	T-4	Insignificant	6
3	5.776	2	Insignificant	T-4	3.837	3
4	4.549	4	Insignificant	T-4	5.095	1
5	6.154	1	4.955	1	3.973	2
6	2.676	6	1.940	2	3.042	4

Discussion

Several takeaways related to complementary player types and predicted Pythagorean win percentage are provided by the regression models:

1. Unsurprisingly, all models suggest that the biggest impact comes from a team's best player. In all three models, BPM is significant with the largest positive coefficient.
2. The models suggest that cluster five players are universally the most impactful, ranking as the greatest coefficient of the six clusters for cluster two and three teams and the second highest for cluster six teams. This reflects the current state of the NBA and the need for role players to be able shoot three-pointers. Their catch and shoot ability and low USG% pairs nicely with any playstyle.
3. The relationship between cluster two and six players suggests that these two clusters are incompatible. Playing cluster six players on teams with a playmaker as the best player provides little value and cluster two players were the only ones not significant in Model III. This relationship makes sense as both types of players want the ball in their hands and having both on the court at once would likely lead to chemistry issues.
4. While cluster two and six players do not appear to be helpful to one another, cluster three players rank second for cluster two teams and third for cluster six teams, suggesting they are the most compatible as supporting players out of these three clusters.

5. Although cluster two teams tend to have greater success than cluster three and six teams, as supporting players, cluster two players do not seem to be helpful. Minutes played by this cluster are insignificant in Models II and III and rank second to last in Model I.
6. Model II suggests that cluster three teams are the least dependent on the makeup of the roster. Instead, a recipe for success for these teams is availability of their best player, given by the large, positive coefficient for BPM, and having players that can shoot the ball well, evident by clusters five and six being the only significant variables.
7. The cluster with the most varied impact across the three models is cluster four. The cluster ranks first for cluster six teams, fourth for cluster two teams, and is insignificant for cluster three teams. The insignificance in Model II is not surprising, as playing a traditional big with a versatile big would likely clog up the area near the basket for both players and not provide much defensive versatility. Cluster six teams, however, welcome the defensive prowess of cluster four players, while cluster two teams seem to benefit more by the defensive abilities of cluster three and one players rather than those from cluster four.
8. Cluster one players only seem to provide value to cluster two teams, ranking second to last for cluster six teams and not being significant in Model II.

Conclusion

This study aimed to quantify the effects that different playstyles of NBA players have on a team's success. By first clustering NBA players based on a variety of statistics that describe the way they play, traditional player positions were redefined to more accurately account for what players do when they are on the court. All 300 teams over the 10 years of data were put into groups based on the playstyle of their best player. Ordinary Least Squares regressions were conducted to show what playstyles are best to put around different styles of players. The models suggest that depending on the playstyle of a team's best player, different playstyles are impactful on team success. Teams whose best player can be described as a playmaker have higher predicted Pythagorean win percentages when surrounded by perimeter shot takers, versatile and traditional bigs, and defensive specialists than when minutes are given to other playmakers and score-first players. When the best player on a team is a versatile big, the model suggests that only perimeter shot takers and well-rounded scorers impact success compared to other playstyles. Lastly, teams that have a well-rounded scorer as their best player are predicted to benefit most from traditional bigs, perimeter shot takers, and versatile bigs.

Notes

i. The author invites the reader to look at Coach & A.D. (nd) or other sources for a glossary of basketball terminology.

ii. The 2019-2020 NBA season was shortened due to the COVID-19 pandemic. Games played by teams during this season ranged from 63 to 75. The games played minimum requirement was based on the number of games each individual team played. The 2020-2021 NBA season was shortened to 72 games. The games played minimum requirement was set at 36 for this season.

iii. Predictions for Model I ranged from 36.2%-76.5%, Model II ranged from 19.1%-70.2%, Model III ranged from 20.9%-65.8%.

iv. The 2022-2023 average minutes played per team was 19,828 due to a small number of games going to overtime.

References

Aschburner S (2022) Commissioner Adam Silver discusses league's 'positionless basketball' at Finals press conference. NBA https://www.nba.com/news/nba-commissioner-adam-silver-discusses-leagues-positionless-basketball-at-annual-finals-press-conference

Brems M (2017) A one-stop shop for Principal Component Analysis. Towards Data Science https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Coach & A.D. (nd) A glossary of basketball terms. https://coachad.com/glossary-of-basketball-terms/

Duman EA, Sennaroğlu B, Tuzkaya G (2021) A cluster analysis of basketball players for each of the five traditionally defined positions. Journal of Sports Engineering and Technology https://doi.org/10.1177/17543371211062064

ESPN (nd) NBA Real Plus-Minus - 2023-24. https://www.espn.com/nba/statistics/rpm/_/sort/DRPM

Kalman S, Bosch J (2020) NBA lineup analysis on clustered player tendencies: a new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates. In 14th annual MIT Sloan Sports Analytics Conference
McMahan I (2018) How (and why) position-less lineups have taken over the NBA playoffs. https://www.theguardian.com/sport/blog/2018/may/01/how-and-why-position-less-lineups-have-taken-over-the-nba-playoffs
Medium (2018) Understanding k-means clustering in machine learning. Towards Data Science https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
NBA (nd) Basketball positions. Jr NBA https://jr.nba.com/basketball-positions/
Osken C, Onay C (2022) Predicting the winning team in basketball: a novel approach. Heliyon https://www.cell.com/heliyon/pdf/S2405-8440(22)03477-6.pdf
Patel R (2017) Clustering professional basketball players by performance [Master's thesis, University of California Los Angeles]. https://www.proquest.com/docview/1989767866?fromopenview=true&pq-origsite=gscholar
Shea S (nd) The 3-point revolution. https://shottracker.com/articles/the-3-point-revolution
Sports Reference (nd) Glossary. Basketball Reference https://www.basketball-reference.com/about/glossary.html
Tsai P (2017) A metallurgical scientist's approach to predicting NBA team success. https://towardsdatascience.com/a-metallurgical-scientists-approach-to-predicting-nba-team-success-4bfa7b2bd6a7
Wikipedia (nd) K-means clustering. https://en.wikipedia.org/wiki/K-means_clustering
Yobero C (2016) Methods for detecting and resolving heteroskedasticity. RPubs https://rpubs.com/cyobero/187387
Zhang S, Lorenzo A, Gómez MA, Mateus N, Gonçalves B, Sampaio J (2018) Clustering performances in the NBA according to players' anthropometric attributes and playing experience. Journal of Sports Sciences https://doi.org/10.1080/02640414.2018.1466493
Ziller T (2017) The center position in the NBA is dying. https://www.sbnation.com/2017/7/17/15981266/nba-center-position-bad-free-agents

Appendix

Appendix: Variable descriptions, some quoted from basketballreference.com

Variable	
TRB%	An estimate of the percentage of available rebounds a player grabbed while they were on the floor
AST%	An estimate of the percentage of teammate field goals a player assisted while they were on the floor
STL%	An estimate of the percentage of opponent possessions that end with a steal by the player while they were on the floor
BLK%	An estimate of the percentage of opponent two-point field goal attempts blocked by the player while they were on the floor
TOV%	An estimate of turnovers committed per 100 plays.
SDSH	standard deviation of the proportion of shots taken from each distance range (derived)
FG% 0-3ft	Field goal percentage on shots between zero and three feet from the basket
FG% 3-10ft	Field goal percentage on shots between three and 10 feet from the basket
FG% 10-16ft	Field goal percentage on shots between 10 and 16 feet from the basket
FG% 16ft-3pt	Field goal percentage on shots between 16 feet from the basket and the three-point line
3P%	Three-point percentage
% 2P Ast'd	Percentage of two-point shots made that were assisted
% 3P Ast'd	Percentage of three-point shots made that were assisted
FT Rate	Number of free throw attempts per field goal attempt
PF/G	Personal fouls per game
Min/G	Minutes played per game
PTS/G	Points scored per game
TRB/G	Total rebounds per game
AST/G	Assists per game
STL/G	Steals per game
BLK/G	Blocks per game
FGA/G	Field goal attempts per game
FG%	Field goal percentage
3PA/G	Three-point attempts per game
FTA/G	Free throw attempts per game
USG%	An estimate of the percentage of team plays used by a player while they were on the floor.
Avg. Dist	Average shot distance (in feet)
% Shots 0-3ft	Proportion of shots taken between zero and three feet from the basket
% Shots 3-10ft	Proportion of shots taken between three and 10 feet from the basket

% Shots 10-16ft	Proportion of shots taken between 10 and 16 feet from the basket
% Shots 16ft-3pt	Proportion of shots taken between 16 feet from the basket and the three-point line
% Shots 3P	Proportion of shots taken from beyond the three-point line
% FG Ast'd	Percentage of shots made that were assisted (derived)