**NBA Player Roles Over Time - Clustering Across Eras**

Nicholas Sienicki and Erik Euler

**Problem Statement**

Change within any industry is inevitable, and professional sport is no different. The rise of analytics in sports has shifted traditional ways of valuing success, which has resulted in new approaches to pregame and in-game decision making based on metric-driven insights. This is especially present in the National Basketball Association (NBA) where team actions and strategies have greatly changed since the rise of analytics.

In professional basketball, player roles have evolved significantly in recent years, reflecting changes in strategies, rule adaptations, and even the skill sets players develop as they enter the professional ranks.

Focusing on one of the more significant shifts in recent years, perhaps the most transformative period in recent NBA history was the rise of the Golden State Warriors' dynasty, starting in 2014 with the hiring of Steve Kerr and emergence of Stephen Curry as one of the greatest offensive players of all time. Their success, driven by an emphasis on three-point shooting, versatile defense, and positionless play, has had a profound impact on how players are utilized in the league.

The objective of this paper is to use clustering methodologies to identify distinct player roles and compare player roles across two NBA seasons: the pre-Warriors dynasty era (2012-13) and today’s game (2023-24). To do so, a key piece of our approach will be evaluating feature usefulness within an unsupervised learning environment, a challenge with no “accuracy” metric from which to evaluate clustering effectiveness.

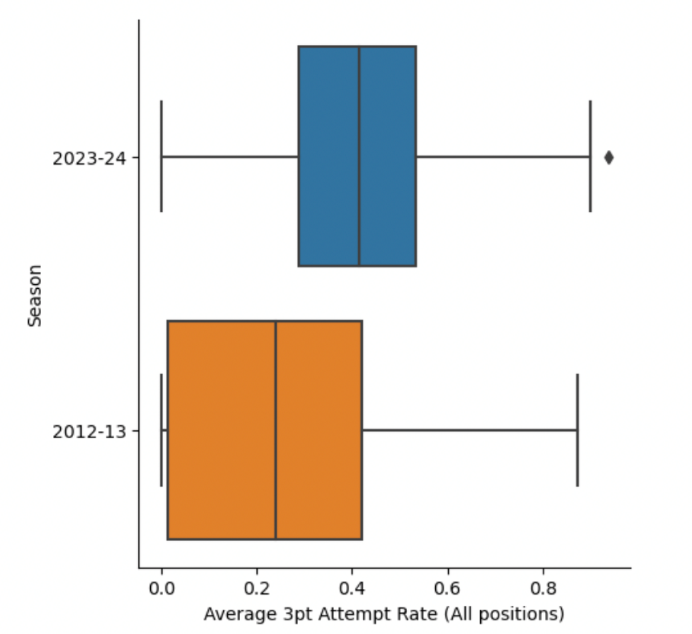
We will seek to provide the reader with an overview of the current landscape of player positions within basketball from our perspective and survey existing literature around player clustering and clustering in general. Then, we’ll use the Calinski-Harabasz index to select and tune a KMeans approach across a couple feature evaluation strategies to come up with the best KMeans clustering for our dataset. Finally, we’ll examine and compare our clusters to see what kind of insight our analysis provides to the evolution of player roles across the periods measured.

**Basketball Review**

By analyzing player roles before and after the Warriors’ emergence using clustering, we aim to demonstrate the new positionless nature of the game from a modeling lens, potentially revealing shifts in the importance of certain skills, the emergence of new player archetypes, or the decline of traditional roles. In addition, layering in a comparison to traditional positions (point guard, center, etc.) gives us a good sense of whether these clusters’ positional purity is consistent across the years investigated. The table below provides a brief description of the 5 traditional basketball positions.

|  |  |
| --- | --- |
| **Position** | **Description** |
| Point Guard (PG) | - Teams’ best passer/dribbler - Protect the ball, limit turnovers - Typically one of the shortest players in the lineup (this has shifted) |
| Shooting Guard (SG) | - Teams’ best shooter, especially from 3pt range  - Many offensive plays are designed to get SG an open shot  - Shorter in height as well, similar to PGs |
| Small Forward (SF) | - More versatile players  - Often the team's best defender on the court  - Can score both inside (closer to the basket) and outside - Typically bigger/taller than the Guards, but shorter than Power forwards and Centers |
| Power Forward (PF) | - More skilled taller players  - Majority of scoring comes closer to the basket, but this has shifted and PFs that can shoot are very valuable  - Good rebounders  -Usually teams 2nd tallest players but sometimes taller than Centers |
| Center (C) | - Often the team's biggest player in the lineup (both height and weight)  - Not usually as skilled offensively as other positions  - Known as good defenders, making interior scoring difficult for other team  - Teams best rebounder and shot blocker |

Below you will find box plots comparing various features from the 2012-13 season and the 2023-24 seasons, to illustrate how the game has changed in the past 10+ seasons.

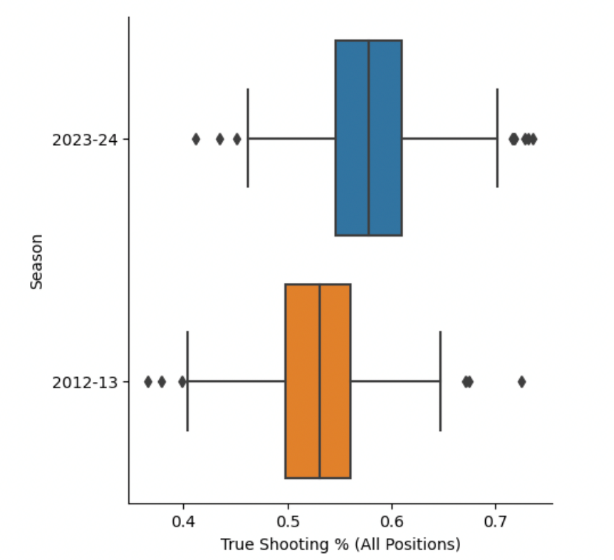
A diagram of a number of numbers

Description automatically generated with medium confidence

Figure : 3pt Attempt Rate (all positions) Figure : 3pt Attempt Rate (PFs and Cs)

First, it’s important to point out the shift in 3-point shooting that has occurred over the past 10 years. In Figure 1, a box plot is displayed tracking the rate of 3-point shot attempts from 2012-13 and 2023-24. The average rate in which all players attempt 3-point shots has increased by ~20% between these two seasons, from around 25% to 45%. This 45% 3-point shot attempt rate was seen as the 75th percentile (3rd Quartile) back in the 2012-13 season and more recently considered the league wide average.

What underpins this shift in 3-point shot attempts? Obviously, 3 points is more valuable on a possession-by-possession basis than 2 points. This has always been the case, but traditionally the distance that players need to shoot from has offset the point differential. Recent crops of players, however, are demonstrably better at shooting the ball than their predecessors. This is illustrated by the below plot that compares “True Shooting”, which is a measure of shooting efficiency that takes into account, 2-point shots, 3-point shots and free throws.

A graph of a number of players

Description automatically generated with medium confidence

Figure : True Shooting Percentage Figure : Player Height

True shooting % average has increased by about 5% and the various quartiles and upper bounds have increased as well, indicating an overall increase in league wide shooting skill.

As for players’ physical characteristics, when observing Figure 4 above, you’ll notice that the average height is about the same for both seasons observed, 200 centimeters or about 6 foot 6 inches in height. The interesting aspect of this plot is seen in the lower tails. The 12-13 season had many more players of a shorter height, 185 centimeters and below, compared to the 23-24 season. This illustrates the prioritization of height over the past 10 years, and how skillsets historically seen in shorter players are now seen in taller players as well.

**Literature Review**

Several evaluations of individual player roles exist attempting to go beyond traditional designations of player position. We see quite a few analyses that use player roles as a component of evaluation. In a 2017 evaluation of player technical and physical performance profiles, player position was used to compare performance between strong and weak teams (Zhang et al., 2017). An excellent survey of the impact of positionless players on overall team defensive results helped to provide our understanding of the potential to take our classification approach and fold it into a wider evaluation of team success (South, 2024).

The closest evaluation to our own we found, and a concept we hope to build upon in a slightly different direction, is a 2019 classification focused on player “versatility.” A Brazilian research team built a definition of versatility based on filling multiple roles on the court- defending, rebounding, and shooting characteristics (Rangel, Ugrinowitsch, and Lamas, 2019). Their evaluation used data from nine seasons in the Brazilian league to segment these versatile players from specialists. One of the points of analysis we will perform on top of our classification is to examine the profiles of the players in our clusters, so an examination of clusters for versatility could provide a cross-period and cross-country comparison point with existing research.

In developing our feature evaluation approach, we relied upon the work done in a 2013 paper evaluating functional magnetic resonance imaging of the brain. This work touches on the difficulty of tuning models in unsupervised learning contexts in general, and feature selection and extraction in particular (Wang et al 2013). The concept introduced, principal feature analysis (PFA), is based in principal component analysis (PCA) and extends PCA by using the methodology to identify the original features that explain the most variance in the dataset. This is distinct from PCA’s approach of using the linear combination of features, which result in new features that the authors deem “feature extraction” as opposed to feature selection.

To determine the effectiveness of feature selection, as well as in the selection of the clusters within the KMeans algorithm, we needed to find a way to compare clustering across approaches. There are many metrics that go about measuring cluster separation, but the one used in our approach is the Calinski-Harabasz (CH) index. Mathematically, the index measures the ratio of between-cluster variance and within-cluster variance, with the goal of creating tight clusters with minimal overlap with other clusters (Łukasik et al, 2016). Figure 5 shows the formulation, with ui representing each cluster’s centroid, U representing the overall “center of gravity” of the dataset, C representing the number of clusters, CL representing each cluster individually, and N representing the number of features:

Inserting image...

Figure : Calinski-Harabasz Index Formula

Well defined clusters have high values of this index, while smaller values indicate overlaps between clusters or wide dispersions within a cluster. This index will be the evaluation metric of the quality of various cluster permutations.

**Methodology**

We have gathered player data from NBA seasons 2012-13 and 2023-24 using data available from [Basketball Reference](https://www.basketball-reference.com/leagues/NBA_2024_per_game.html). For both periods, we collected basic statistics (e.g., points, rebounds, assists, shooting percentages) cleanly available out of game box scores, and advanced metrics (e.g., Player Efficiency Rating, True Shooting Percentage) that rely on other models of understanding of how player performance can be best related to team success.

After collecting the data, we needed to clean and normalize it. Clustering methodologies are dependent on normalized data without missing values, therefore establishing standards of usable data is important to our approach. Generally, we encountered few issues with the cleaning of our data, with the main adjustment being the interpolation of statistics related to three-point makes and attempts, which some players did not have. We normalized each feature to have values from 0 to 1 to ensure variance for each feature was weighted equally within the selected algorithms. Additionally, we limited the players clustered to those who had played more than 40 games.

In total, we found 65 features supporting the below feature categories, with examples of each below:

* **Scoring Role**: Points per game, shooting percentages, and usage rate
* **Playmaking Role**: Assists per game, usage percentage, turnovers
* **Defensive Role**: Blocks, steal percentage, and defensive rating statistics (win shares, defensive plus/minus)
* **Physical Attributes**: Height and weight
* **Overall Effectiveness**: Win Shares, Value over Replacement Player (VORP), Plus/Minus (BPM)

One of our primary concerns in this evaluation is one of covariance among features. We have a skew of the data toward shooting statistics, which that make up 27/65 features in our dataset. The fear is that we may be segmenting players into shooting roles that do not fully reflect their overall contributions on the court. Broader than that, our “advanced” statistics are often built upon the more basic features. Both are present. Our main approach to assuaging that concern is one of feature selection.

We’ll compare the performance of three views of the data for both periods we have available- the original full dataset, a hand-selected dataset limiting the shooting features to only the advanced shooting statistics, and features selected through the PFA method outlined above. The implementation of the PFA method is implemented as laid out in Wang et al’s 2013 PFA paper:

Matrix Y represents the player statistics with dimension *n x L*:

**

Where n is the number of features and L is the number of observations. **Σ** represents the covariance matrix of Y, and **Λ**is the diagonal matrix whose diagonal elements are the eigenvalues of **Σ. Σ** can then be expressed as**:**



where**A**is the eigenvector of the covariance matrix **Σ**, written as **A** = (**e**1, **e**2, …, **e***n*), **e***i* ∈ **R***n*, and**e***i*is an eigenvector corresponding to the eigenvalue*λi*.**A***q*represents the first*q*eigenvectors and**a***i*be the row vectors of**A***q*:



where each vector**a***i*represents the projection of the *i*th feature of**Y**to the *q*-dimension space. Features that are highly correlated will have similar weight vectors**a***i*, which can be used to remove features with redundant information (Wang et al, 2013).

Based on this, the PFA approach finds highly correlated features, which we can then use to remove those with redundant information. Better selection performance for PFA than the hand-picked model would go a long way to addressing the above problem of feature selection and covariance within our data. Our implementation uses the first eight principal components for **a***i* weight comparison to generate the eight features that capture the most variance within the dataset.

For each of the data views, we’ll run the KMeans algorithm for a selection of k values, and measure performance using the CH index. The clustering we will proceed with for each period will be the highest index value with the best-performing number of clusters. Our KMeans approach relies on limiting within-cluster sum of squares, which may or may not be the best way of clustering player roles.

After clustering has completed, we will take the sets of clusters for each period and compare them. We’ll link the centroids between eras by taking the minimum Euclidean distance, allowing for cross-period comparison of how roles have shifted. We’ll also be able to evaluate each cluster’s traditional position designation by computing a purity metric using the players’ listed traditional position, as well as empirically evaluate the clusters and links from a basketball perspective. Through these comparison points, we aim to draw a conclusion of how player roles have shifted in the periods measured.

**Results and Discussion**

After the creation of the datasets as outlined above, we were able to run a comparison of the effectiveness of the feature selection approaches for each period applying differing k values from 2 to 15 to the KMeans algorithm using the CH Index. As shown in Figures 6, 7, and in the table detailing the evaluation in the appendix, we see that the index peaks at 3 for the PFA selection methodology across our evaluation field and the PFA approach outperforms the hand-selection and the full model approaches. Features selected are also available in the appendix.

A graph with a line

Description automatically generatedA graph with a line and a dotted line

Description automatically generated

Figure : Calinski-Harabasz Index by k, PFA 2012-13 Figure : Calinski-Harabasz Index by k, PFA 2023-24

We also see that the hand-selected model outperformed the full model as well. As the PFA approach selected 8 variables, the hand-selected model included 40, and the full model included 65. The effectiveness of each feature selection technique corresponded to the number of features, although not linearly.

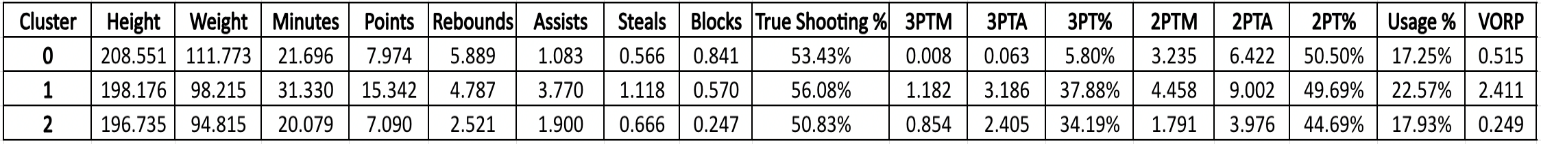
In general, we see ~15% increase in the CH index from 2012-13 to 2023-24 across the views created. This indicates better defined clusters in today’s player roles than in 2012-13. This is a somewhat surprising result given the expectation of blurrier roles as time and skillsets progress. The CH-optimal number of separable clusters also surprised us. The expectation was that we would see more specialization than the result above shows in both eras, but especially in 2012-13.

Next, we were able to begin examination of the clusters. As noted above, we linked clusters by the Euclidean distance between centroids. The results can be found in Figure 8:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Cluster Year** | **Closest Opposite Cluster** | **Euclidean Distance** | **Most Frequent Position** | **Positional Purity** | **Count** |
| 0 | 12\_13 | 2 | 2.565888 | C | 58.947368 | 95 |
| 1 | 12\_13 | 2 | 1.370890 | PG | 26.373626 | 91 |
| 2 | 12\_13 | 0 | 0.557599 | SG | 31.818182 | 154 |
| 0 | 23\_24 | 2 | 0.557599 | SF | 27.461140 | 204 |
| 1 | 23\_24 | 0 | 2.977291 | C | 81.08108108 | 37 |
| 2 | 23\_24 | 1 | 1.370890 | C | 21.186441 | 107 |

Figure : Cluster Evaluation

Below we will look at selected per game statistics as well as some physical characteristics that will paint a picture about the average or typical player within our created clusters based on our KMeans model and PFA approach to feature extraction. First, we will look at the 2012-13 cluster assignments.

Figure : 12-13 Custer Per Game Statistics

*Cluster 0 (2012 – 2013)*

Cluster 0 of the 2012-13 cluster assignments appears to represent the typical big man in the 2012-13 season. The average height is 208.5 centimeters, or slightly taller than 6’10. Additionally, this cluster outputs the most rebounds and blocks on average, two counting statistics typically dominated by centers and power forwards, as touched on earlier. This is perpetuated further in Figure 8 where we see that nearly 59% of players within this cluster identify as centers. This cluster takes and makes very few 3-point shots, attempting about 0.06 per game, and only making 5.8% of them on average. However, they do make the highest percentage of 2-point attempts, at 51.5% which intuitively makes sense as typically these players' shot attempts will occur close to the basket.

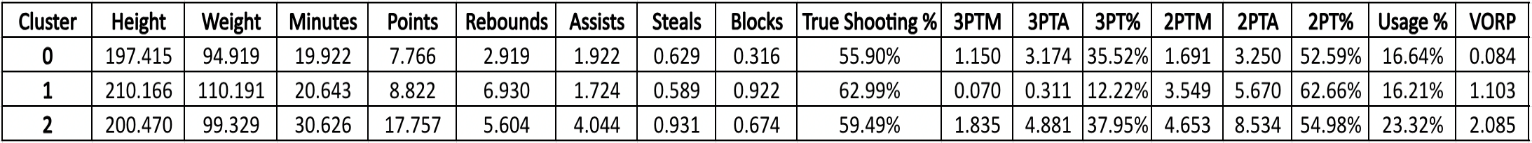
*Cluster 1 (2012 – 2013)*

Cluster 1 of the 2012-13 cluster assignments appear to be the main ball handlers on a team and the player who is the biggest offensive threat. This cluster of players averages the most points (15.3), assists (3.7), minutes (31.3), and usage rate (22.57%) which is a measure of the percent of the possessions “used” by an individual player (ending a possession in a made/missed shot attempt, turnover, or free throw). This cluster also takes the most shots per game at just over 12 shot attempts per game, and shoots at the highest percentage, with a True Shooting % of 56.08%. The typical makeup of this player is about 6’6 and the plurality of this cluster are listed as point guards; albeit at a low 26.4%.

*Cluster 2 (2012-13)*

Cluster 2 of the 2012-13 cluster assignments represents the role and bench players within a team. This cluster averages the lowest number of minutes per game at just over 20 minutes and scores the fewest points. On average these players take the fewest shots (just behind cluster 0) and have the lowest True Shooting % at 50.83%. These players do have higher usage than cluster 0 players, which could possibly have to do with their competent 3 points shooting, making about 34.2% of them. For this cluster, it is worth touching on the VORP statistic, which stands for Value Over Replacement Player. VORP hinges on the idea that a replacement level player historically costs a team 2 points over a 100-possession game, and VORP measures how much higher or lower a given player is performing relative to that hypothetical replacement player, based on on-court scoring margin (also known as Box Plus/Minus). A negative VORP indicates a player is performing worse than a replacement level player, while a positive VORP indicates a player is outperforming a replacement level player. Cluster 2 has a VORP of 0.249, meaning this group of players, on average, are performing marginally better than a replacement level player – further distinguishing this cluster as the role/bench players. The players in the cluster on average are about 6’5 and 31.8% of them are listed as shooting guards.

Next let's investigate the cluster assignments from the 2023-24 player data.

Figure : 23-24 Custer Per Game Statistics

*Cluster 0 (2023-24)*

Cluster 0 from 2023-24 appears to be the cluster representative of the bench and role players. These players on average are just below 6’6 and play just under 20 minutes per game. They have the lowest True Shooting % of the three clusters while having the second highest usage rate, just edging out cluster 1. These players take nearly as many 3-point shots as they do 2-point shots and make these 3-point attempts at a solid rate of 35.5%. It is worth noting that this cluster averages 0.629 steals per game, which if considered on a per minute basis exceeds the rate of the other clusters. Looking at VORP again, this cluster is barely outperforming a replacement level player. Lastly, it is important to note that 55% of all 23-24 eligible players were categorized into this cluster, and within the cluster, the highest frequency of position was small forward, but only at a low rate of 27.6%, indicating this cluster captured a wide breadth of players.

*Cluster 1 (2023-24)*

Cluster 1 seems to capture the 2023-24 big men (centers and power forwards). On average these players are just under 6’11, averaging 6.9 rebounds and 0.92 blocks a game. The players within this cluster yield the highest True Shooting % at nearly 63% and take nearly all their shots inside the 3-point line, only taking 0.3 3-point shots, a game as opposed to 5.6 2-point attempts. Despite this high True Shooting %, these players have the lowest usage rate at 16.21%, further illustrating the perceived value of being able to take and make 3-points shots consistently. It is also interesting to note that this cluster only averages 20.6 minutes per game, which is not even half of a 48-minute NBA game. Only 37 players were categorized into this cluster, and 81.08% of them are listed as centers.

*Cluster 2 2023-24*

Cluster 2 appears to represent the main ball handlers and offensive threats on a given team. These players average just over 30 minutes per game, 17.7 points, 4.04 assists, 5.6 rebounds and attempt 13.3 shots a game. These players possess the highest usage rate of 23.32%, and have a True Shooting % of 59.5%, which is quite high given they take almost 5 3-point shots per game. Regarding 3-point shooting, these players are the best 3-point shooters as well, making these shots at nearly 38%. This is the second largest cluster, with 118 players categorized into this cluster, and surprisingly, the most common position in this cluster is center, although the frequency is only 21.18% indicating many positions could be represented by this cluster. Players within this cluster are on average marginally shorter than 6’7, further exhibiting the potential positional versatility within this cluster.

*Comparing Clusters*

When observing the 2023-24 cluster groups, the ‘closest opposite cluster’ mapped as expected back to the 2012-13 cluster assignments. We ended up with three distinct groups from both seasons when examining these cluster assignments: role players, big men, and main offensive threats. Below we will compare these groups a bit further, examining how these roles changed over the past ten years.

First, it is hard to ignore the shift in 3-point shooting within each of the three distinct groups. Even when looking at the big men cluster, while the numbers are still quite low, they take more than four times as many 3-point shots and have increased the rate in which they go in by 6.42%. Both attempts and rate of making 3-pointers has also increased for the other two clusters, albeit not as drastic as seen in the big men. Further, as pointed out above, 21.18% of the offensive threat cluster, that attempts nearly five 3-point shots a game, were listed as centers, displaying how the traditional expectations of the center position are shifting.

Next, let's look at usage for our clusters across these seasons. Both the usage rates for big men and role players have decreased, while the usage for the offensive threats has increased. The most skilled scorers on a given team are relied upon more in the current era of basketball than the previous period measured. It is also worth looking at another advanced statistic in VORP, where the role player clusters’ VORP decreased from 0.249 to 0.084 showing a reduction of role player value closer to that of a replacement level player. Only the big men cluster saw an increase in VORP, more than doubling their perceived value over a replacement level player between our seasons of observation, while the offensive threats VORP decreased by about 0.33.

Visually, we can see the similarity in the clusters between the two selected seasons. Both create distinct differences when projecting the players’ full statistics into two principal components. This shows the success of the CH index in creating separable clusters:

A graph with colored dots

Description automatically generatedA graph with colored dots

Description automatically generated

Figure 1: PCA of Player Clusters, 2023-24 Figure : PCA of Player Clusters, 2012-13

The ‘closest opposite cluster’ ended up yielding some interesting results when considering the 2012-13 season. The cluster identified as big men in 2012-13 mapped closest to the offensive threats cluster from the 2023-24 season based on Euclidean distance. While it was the largest Euclidean distance for any of the three opposite clusters measures for these 2012-13 clusters, it is interesting that the linking approach yielded this result. We hypothesize this may have to do with the increased size of the best offensive players in more recent years, seeing players 6’9 and above exhibiting elite levels of offensive skill combined with rebounding and defensive statistics that are traditionally associated with power forwards and centers. We will touch on a few of these types of players in the section below.

**Top Players by Cluster**

Another way to further contextualize these clusters is to look at some of the best players to fall into each cluster. Notably, it is challenging to define “best” players in the NBA as there are so many star players with various skillsets, but an oft-cited statistic is Player Efficiency Rating (PER). PER was created by well renowned basketball analyst John Hollinger, and is made up of a complex formula that considers player positive contributions (made shots, free throws, assist, rebounds, blocks, steals), player negative outcomes (missed shots, turnovers, fouls), adjusts for the teams pace (measures a team's offensive possessions), and finally, adjusts this number on a per minute basis such that simply more playing time doesn't yield higher PER (Hollinger, 2007). Ultimately, a PER value of 28 or above will place a given player in the Most Valuable Player (MVP) conversation.

Given this metric is somewhat of a catch-all statistic, we excluded it from our model variables but have brought it back in to determine the top players from each cluster during our two seasons observed. Below are tables for each of the two seasons observed and the top three players by PER for each cluster.

*2012-13 Cluster 0 (Big men)*

|  |  |
| --- | --- |
| **Player** | **Player Efficiency Rating** |
| Brook Lopez | 24.7 |
| Andray Blatche | 21.9 |
| Anthony Davis | 21.7 |

Figure 11: 2012-13 Cluster 0 PER Leaders

*2012-13 Cluster 1 (Main offensive threats)*

|  |  |
| --- | --- |
| **Player** | **Player Efficiency Rating** |
| LeBron James | 31.6 |
| Kevin Durant | 28.3 |
| Chris Paul | 26.4 |

Figure 12: 2012-13 Cluster 1 PER Leaders

*2012-13 Cluster 2 (Bench/role players)*

|  |  |
| --- | --- |
| **Player** | **Player Efficiency Rating** |
| Eric Bledsoe | 17.5 |
| Chris Copeland | 16.8 |
| Will Bynum | 16.6 |

Figure 13: 2012-13 Cluster 2 PER Leaders

*2023-24 Cluster 0 (Bench/role players)*

|  |  |
| --- | --- |
| **Player** | **Player Efficiency Rating** |
| T.J. McConnell | 20.9 |
| Bol Bol | 18.7 |
| Kevin Love | 18.5 |

Figure 14: 2023-24 Cluster 0 PER Leaders

*2023-24 Cluster 1 (Big men)*

|  |  |
| --- | --- |
| **Player** | **Player Efficiency Rating** |
| Giannis Antetokounmpo | 29.9 |
| Domantas Sabonis | 23.2 |
| Andre Drummond | 23 |

Figure 14: 2023-24 Cluster 1 PER Leaders

*2023-24 Cluster 2 (Main offensive threats)*

|  |  |
| --- | --- |
| **Player** | **Player Efficiency Rating** |
| Nikola Jokic | 31 |
| Shai Gilgeous-Alexander | 29.3 |
| Luka Doncic | 28.1 |

Figure 14: 2023-24 Cluster 2 PER Leaders

There are a few intriguing aspects worth touching on in the above tables. First, the identification and classification of league MVPs into differing categories. Lebron James’ 2012-13 season is widely seen as one of the greatest seasons of the last 25 years, averaging 26.8 points, 8 rebounds, 7.3 assists, shooting 56% and almost unanimously winning league MVP. More recently, Nikola Jokic, has won three MVPs in four years. In 2023-24, he averaged 26.4 points, 12.4 rebounds, 9 assists, shot 58.3%, all while taking three 3-point shots a game and making about 36% of them. Both James and Jokic are classified as offensive threats.

As for the big men, it is interesting to compare these top three PER players from this cluster in these two seasons. In 2012-13, this cluster conformed to traditional expectations of their size- taking few 3-point shots, scoring close to the hoop, rebounding and blocking shots. The top two PER players in 2023-24, Giannis Antetokounmpo and Domantas Sabonis, aren't very different, however they have notably higher PERs and attain a much higher usage rate. Giannis Antetokounmpo, standing at 6’11 and possessing a 7’3 wingspan, won the MVP award in back-to-back years in 2019 and 2020. In 2023-24 he averaged 30.4 points, 11.5 rebounds, 6.5 assists, shooting 61%. However, he does not shoot 3-pointers well, only making 21.4% of them on less than 1 attempt per game. Domantas Sabonis was on the outskirts of the MVP conversation in 2023-24, averaging 19.4 points, 13.7 rebounds, 8.2 assists, shooting 59.4% and only taking 1.1 3-pointers a game.

Contrasting Sabonis and Antetokounmpo with James and Jokic, all MVP level players, shows the degree to which the model separates players based on shooting, especially in 2023-24.

Brook Lopez is an interesting case study individually. He entered the league in 2008-09 and up until the 2015-16 season, he had taken a *total* of 31 3-point attempts, occupying a very traditional Center role on his team, and playing that role well. Starting in 2016-17 and into today, he averages more than five 3-point attempts per game and makes these shots at a 36% rate. This career transformation exemplifies the shift in how today's NBA is played, and he subsequently was categorized in the offensive threat cluster in 2023-24, moving from the big men cluster he was previously categorized in in 2012-13. A 7’ foot player who can take and make 3’s presents extreme challenges for a defending team.

The role/bench player cluster PERs above yield the lowest values across the clusters, which tracks with the statistical observations above. These players are not widely depended on offensively and in both seasons observed, only average about 20 minutes per game. However, this does not mean there are not valuable players in these clusters. T.J. McConnell is currently one of the best backup point guards across the league. At only 6’1, a 20+ PER rating is impressive, averaging 10.2 points, 5.5 assists, 2.7 rebounds, 1 steal while only playing about 18 minutes per game. Eric Bledsoe, recording the highest role/bench player cluster PER in 2012-13 was also a very solid player. While he was a bench/role player in 2012-13, he went on to be a very effective starter and have a successful career, averaging 20+ points in back-to-back years from 2015-17. That said, this cluster could be used to identify players that could potentially break out if given a larger role. Role/bench players are instrumental to a team's success. Players in this cluster are by no means bad players and teams with reliable bench and role players often see themselves playing deep into the playoffs.

**Conclusion**

This project leveraged novel variable selection within unsupervised clustering to identify and analyze player roles in the NBA, comparing two pivotal periods: pre-Warriors era (2012-13) and modern era (2023-24). We sought to understand the impact of strategic innovations, such as the increased emphasis on spacing and versatility, on the evolution of player roles. Additionally, we also evaluated the effectiveness of traditional position designations.

On these fronts, we had mixed results. We showed that there are three distinct clusters of players in both eras: playmaking/scoring focused players, rebounding/defensive contributors, and role players. We saw that position mattered less to classification than player role. This shows that traditional positional designation does not necessarily correspond to on-court outcomes. Our linking exercise also showed broader shifts in usage across eras, especially among player roles.

We also found that there were limitations in our approach. We could further tune the effectiveness of the PFA approach by evaluating permutations of the number of variables selected. Including a pure PCA feature extraction approach could be a useful point of comparison as well. Limiting the hand-selected model to fewer variables may rival the PFA approach for separability, or further prove out the power of PFA.

Using the CH index as the only metric of success may limit us as well. Sacrificing cluster separability for more clusters with greater insight may drive better analysis on player roles by allowing for specialist clusters. In general, the difficulty in evaluating the success of unsupervised clustering proved a significant challenge.

The findings and approach provide valuable insights and a platform for further research for teams, analysts, and fans, offering a data-driven perspective on how the game of basketball and its player archetypes have evolved over the past decade.

**Bibliography**

Cirtautas, J. (2023, October). NBA Players, Version 6. Retrieved October 13, 2024 from <https://www.kaggle.com/datasets/justinas/nba-players-data>.

Hollinger, J. (2007, April). Hollinger: What is PER? UPDATED. ESPN.com. <https://www.espn.com/nba/columns/story?columnist=hollinger_john&id=2850240>

NBA.com (2016). Basketball Positions - Jr. NBA. *NBA.Com*. jr.nba.com/basketball-positions/.

Łukasik, S., Kowalski, P. A., Charytanowicz, M. and Kulczycki, P., "Clustering using flower pollination algorithm and Calinski-Harabasz index," *2016 IEEE Congress on Evolutionary Computation (CEC)*, Vancouver, BC, Canada, 2016, pp. 2724-2728, doi: 10.1109/CEC.2016.7744132.

Rangel, W., Ugrinowitsch, C., & Lamas, L. (2019). Basketball players’ versatility: Assessing the diversity of tactical roles. International Journal of Sports Science & Coaching, 14(4), 552-561. <https://doi.org/10.1177/1747954119859683>

South, C. (2024). A basketball paradox: exploring NBA team defensive efficiency in a positionless game. *Journal of Quantitative Analysis in Sports*. <https://doi.org/10.1515/jqas-2024-0010>

Sports Reference LLC "NBA Player Stats" Basketball-Reference.com – NBA Player Stats. Retrieved November 12, 2024 from <https://www.basketball-reference.com/leagues/NBA_2013_per_game.html>, <https://www.basketball-reference.com/leagues/NBA_2013_advanced.html>, <https://www.basketball-reference.com/leagues/NBA_2013_adj_shooting.html>, <https://www.basketball-reference.com/leagues/NBA_2024_per_game.html>, [https://www.basketball-reference.com/leagues/NBA\_2024\_advanced.html](https://www.basketball-reference.com/leagues/NBA_2013_advanced.html), <https://www.basketball-reference.com/leagues/NBA_2024_adj_shooting.html>

Wang L., Lei Y., Zeng Y., Tong L., and Yan B., Principal feature analysis: a novel voxel selection method for fMRI data, *Computational and Mathematical Methods in Medicine*. (2013) **2013**, 7, 645921, [**https://doi.org/10.1155/2013/645921**](https://doi.org/10.1155/2013/645921), 2-s2.0-84885579715.

Zhang, S., Lorenzo, A., Gómez, M. A., Liu, H., Gonçalves, B., & Sampaio, J. (2017). Players’ technical and physical performance profiles and game-to-game variation in NBA. *International Journal of Performance Analysis in Sport*, *17*(4), 466–483. <https://doi.org/10.1080/24748668.2017.1352432>

**Appendix**

**CH Index by Data View by K**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **K** | **CH Index** | **Dataset** | **K** | **CH Index** |
| pfa\_players\_12\_13 | 2 | 116.9966 | hand\_players\_23\_24 | 2 | 108.2655 |
| pfa\_players\_12\_13 | 3 | 127.0461 | hand\_players\_23\_24 | 3 | 94.89062 |
| pfa\_players\_12\_13 | 4 | 100.9865 | hand\_players\_23\_24 | 4 | 79.6478 |
| pfa\_players\_12\_13 | 5 | 94.5453 | hand\_players\_23\_24 | 5 | 68.67112 |
| pfa\_players\_12\_13 | 6 | 87.67832 | hand\_players\_23\_24 | 6 | 62.1788 |
| pfa\_players\_12\_13 | 7 | 78.39225 | hand\_players\_23\_24 | 7 | 56.0323 |
| pfa\_players\_12\_13 | 8 | 76.05792 | hand\_players\_23\_24 | 8 | 52.8402 |
| pfa\_players\_12\_13 | 9 | 71.61572 | hand\_players\_23\_24 | 9 | 46.28035 |
| pfa\_players\_12\_13 | 10 | 70.6243 | hand\_players\_23\_24 | 10 | 44.76498 |
| pfa\_players\_12\_13 | 11 | 66.47124 | hand\_players\_23\_24 | 11 | 43.03601 |
| pfa\_players\_12\_13 | 12 | 63.21994 | hand\_players\_23\_24 | 12 | 41.78022 |
| pfa\_players\_12\_13 | 13 | 64.93878 | hand\_players\_23\_24 | 13 | 39.59164 |
| pfa\_players\_12\_13 | 14 | 62.34078 | hand\_players\_23\_24 | 14 | 37.17511 |
| pfa\_players\_23\_24 | 2 | 96.90883 | players\_12\_13\_model | 2 | 87.61702 |
| pfa\_players\_23\_24 | 3 | 106.8017 | players\_12\_13\_model | 3 | 84.49286 |
| pfa\_players\_23\_24 | 4 | 95.98819 | players\_12\_13\_model | 4 | 71.06815 |
| pfa\_players\_23\_24 | 5 | 84.58555 | players\_12\_13\_model | 5 | 66.77439 |
| pfa\_players\_23\_24 | 6 | 77.42268 | players\_12\_13\_model | 6 | 61.27205 |
| pfa\_players\_23\_24 | 7 | 74.68216 | players\_12\_13\_model | 7 | 57.2198 |
| pfa\_players\_23\_24 | 8 | 68.69018 | players\_12\_13\_model | 8 | 48.71662 |
| pfa\_players\_23\_24 | 9 | 62.87426 | players\_12\_13\_model | 9 | 48.62581 |
| pfa\_players\_23\_24 | 10 | 61.82485 | players\_12\_13\_model | 10 | 46.15945 |
| pfa\_players\_23\_24 | 11 | 57.98884 | players\_12\_13\_model | 11 | 43.54489 |
| pfa\_players\_23\_24 | 12 | 55.82205 | players\_12\_13\_model | 12 | 40.37911 |
| pfa\_players\_23\_24 | 13 | 54.32342 | players\_12\_13\_model | 13 | 38.89905 |
| pfa\_players\_23\_24 | 14 | 52.78195 | players\_12\_13\_model | 14 | 37.26077 |
| hand\_players\_12\_13 | 2 | 92.74756 | players\_23\_24\_model | 2 | 97.27762 |
| hand\_players\_12\_13 | 3 | 89.81339 | players\_23\_24\_model | 3 | 92.8255 |
| hand\_players\_12\_13 | 4 | 74.3667 | players\_23\_24\_model | 4 | 78.21146 |
| hand\_players\_12\_13 | 5 | 71.07643 | players\_23\_24\_model | 5 | 70.68228 |
| hand\_players\_12\_13 | 6 | 63.81312 | players\_23\_24\_model | 6 | 60.74915 |
| hand\_players\_12\_13 | 7 | 58.84096 | players\_23\_24\_model | 7 | 55.60201 |
| hand\_players\_12\_13 | 8 | 52.36883 | players\_23\_24\_model | 8 | 50.64254 |
| hand\_players\_12\_13 | 9 | 49.15256 | players\_23\_24\_model | 9 | 47.78328 |
| hand\_players\_12\_13 | 10 | 46.06126 | players\_23\_24\_model | 10 | 46.1161 |
| hand\_players\_12\_13 | 11 | 44.03681 | players\_23\_24\_model | 11 | 42.76428 |
| hand\_players\_12\_13 | 12 | 40.92049 | players\_23\_24\_model | 12 | 41.42223 |
| hand\_players\_12\_13 | 13 | 38.70745 | players\_23\_24\_model | 13 | 39.26894 |
| hand\_players\_12\_13 | 14 | 36.85751 | players\_23\_24\_model | 14 | 37.21754 |

**Variable Selection**

|  |  |
| --- | --- |
| **Full Model : (players\_23\_24\_model/ players\_12\_13\_model)** | **['Age\_adv', 'G\_adv', 'TS%\_adv', '3PAr\_adv', 'FTr\_adv', 'ORB%\_adv', 'DRB%\_adv', 'TRB%\_adv', 'AST%\_adv', 'STL%\_adv', 'BLK%\_adv', 'TOV%\_adv', 'USG%\_adv', 'OWS\_adv', 'DWS\_adv', 'WS\_adv', 'WS/48\_adv', 'OBPM\_adv', 'DBPM\_adv', 'BPM\_adv', 'VORP\_adv', 'MP\_per\_gm', 'FG\_per\_gm', 'FGA\_per\_gm', 'FG%\_per\_gm', '3P\_per\_gm', '3PA\_per\_gm', '3P%\_per\_gm', '2P\_per\_gm', '2PA\_per\_gm', '2P%\_per\_gm', 'eFG%\_per\_gm', 'FT\_per\_gm', 'FTA\_per\_gm', 'FT%\_per\_gm', 'ORB\_per\_gm', 'DRB\_per\_gm', 'TRB\_per\_gm', 'AST\_per\_gm', 'STL\_per\_gm', 'BLK\_per\_gm', 'TOV\_per\_gm', 'PF\_per\_gm', 'PTS\_per\_gm', 'FG\_shoot', '2P\_shoot', '3P\_shoot', 'eFG\_shoot', 'FT\_shoot', 'TS\_shoot', 'FTr\_shoot', '3PAr\_shoot', 'FG+\_shoot', '2P+\_shoot', '3P+\_shoot', 'eFG+\_shoot', 'FT+\_shoot', 'TS+\_shoot', 'FTr+\_shoot', '3PAr+\_shoot', 'FG Add\_shoot', 'TS Add\_shoot', 'player\_height', 'player\_weight', 'ind\_usa']** |
| **Hand-Selected: (hand\_players\_23\_24/ hand\_players\_12\_13)** | **['Age\_adv', 'G\_adv', 'ORB%\_adv', 'DRB%\_adv', 'TRB%\_adv', 'AST%\_adv', 'STL%\_adv', 'BLK%\_adv', 'TOV%\_adv', 'USG%\_adv', 'OWS\_adv', 'DWS\_adv', 'WS\_adv', 'WS/48\_adv', 'OBPM\_adv', 'DBPM\_adv', 'BPM\_adv', 'VORP\_adv', 'MP\_per\_gm', 'ORB\_per\_gm', 'DRB\_per\_gm', 'TRB\_per\_gm', 'AST\_per\_gm', 'STL\_per\_gm', 'BLK\_per\_gm', 'TOV\_per\_gm', 'PF\_per\_gm', 'PTS\_per\_gm', 'FG\_shoot', 'FTr\_shoot', '3PAr\_shoot', 'FG+\_shoot', '2P+\_shoot', 'eFG+\_shoot', 'FT+\_shoot', 'TS+\_shoot', 'FTr+\_shoot', '3PAr+\_shoot', 'player\_weight', 'player\_height']** |
| **Principal Factor Analysis: (pfa\_players\_23\_24, pfa\_players\_12\_13)** | **['TRB%\_adv', 'DWS\_adv', 'eFG%\_per\_gm', '3PAr\_adv', 'FT\_per\_gm', 'AST\_per\_gm', '3P\_shoot', 'FT%\_per\_gm']** |