NBA THEN AND NOW: A RE-CATEGORIZATION

Avinash Daniel, Justin Peter, Jaryt Salvo

Abstract: Basketball has changed drastically in the past four decades. Where, in the 80's and 90's, the most commonly taken shots were within the 3-point line, to today where a solid 3-point game is a must. Using data from "2020-21 NBA Player Stats: Per Game," we built multiple model clustering players of two different datasets using three clustering methodologies to determine a final model that best answers the question: ***"How can we operationalize players based on their statistics?"*** Our final model included three functionalities:

> ***In-the-Paints***: High rebounds and blocks; ***Generals***: High points, assists, and steals; and ***Versatiles***: quasi if-else category or catch-all for basketball's everyman.

Introduction

Have you ever been a fan of something for a long time and wondered, “How is [this thing] now related to what it was [X] years ago?” For us, [this thing] is the National Basketball Association (NBA). If you are into basketball, you know there are five players on the court per team. These five players have position names that often indicate their function on the court.

Our contention is that these positions meant something concrete in terms of describing their on-court function/output as expressed by their respective statistics. To reify, using positions we discussed above: centers would have few 3-point shot attempts and point guards would have many assists, both, as compared to other positions. We believe, what was meaningful in the categorization of players, e.g. “point guard” or “center,” has been lost.

Data Description

Data from “2020-21 NBA Player Stats: Per Game” statistics include 492 rows and 29 columns. The observations are made by sports analysts and updating after each game. The 29 columns include, written in smaller font as to take up less space:

Rk — Rank, alphabetically by surname.

Pos — Position: C, Center; PF, Power Forward; PG, Point Guard; SF, Strong Forward; SG, Shooting Guard.

Age — Player's age on 2/1 of the season.

Tm — Team: 30 categorical variable of 3-letter team name, e.g. CLE, Cleveland Cavaliers; DET, Detroit Pistons; MIA, Miami Heat.

G — Games played by player.

GS — Games Started.

MP — Minutes Played Per Game.

FG — Field Goals Per Game, total shoots made per game.

FGA — Field Goal Attempts Per Game.

FG% — Field Goal Percentage.

3P — 3-Point Field Goals Per Game.

3PA — 3-Point Field Goal Attempts Per Game.

3P% — 3-Point Field Goal Percentage.

2P — 2-Point Field Goals Per Game.

2PA — 2-Point Field Goal Attempts Per Game.

2P% — 2-Point Field Goal Percentage.

eFG% — Effective Field Goal Percentage. (This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.)

FT — Free Throws Per Game

FTA — Free Throw Attempts Per Game

FT% — Free Throw Percentage

ORB — Offensive Rebounds Per Game

DRB — Defensive Rebounds Per Game

TRB — Total Rebounds Per Game

AST — Assists Per Game

STL — Steals Per Game

BLK — Blocks Per Game

TOV — Turnovers Per Game

PF — Personal Fouls Per Game

PTS — Points Per Game

Research Question

New dynamics, techniques, strategies, coaching, and big data analytics have transformed the game of basketball from individuals performing five discrete position-functions—Center, Power Forward, Point Guard, Strong Forward, and Shooting Guard—to players wearing multiple hats, performing significantly different from players of the past in their same position.

> ***How can we operationalize players based on their statistics?***

As such, our research question is one of unsupervised learning. Unsupervised, unlike commonly called forth supervised methods, requires no response variables. We will perform a cluster analysis to determine the optimal number of clusters that describes groupings of our data.

Data Cleaning and Setup

From the onset, we removed columns not useful from a "functionality" point-of-view, these columns included: rank, age, team, and games started. Another variable we didn't find useful was minutes played. The main reason for this was because we filtered our final datasets to include only players who have played *at least* 70% of total games. Essentially, with that 70% threshold, we will have already filtered out players with insignificant number of minutes played.

3-Point Percentages NA's

After removing these columns above and players who played less than 70% of the total games played, we observed a summary of our data. First thing we noticed was that there were a few "NA's" under the 3-point percent column. We discovered this was because these players had never attempted a 3-point shot. Therefore, 3-point shots made divided by 3-point shots attempt (or our 3-point percentage column) is undefined because we can't divide by 0. To resolve these NA's, we used a function to fill all 3-point percentages with value of "NA" as 0[%].

Clarification of Player Positions

Next, we looked at a table of positions in our data. We found some dual-positions occurring in six instances. Although positions are not part of our functionality analysis, we will plot charts of the distributions between positions and clusters, so we want clearly defined positions for the position-cluster analysis.

These dual-positions included: C-PF, PG-SG, SF-SG, SG-PG, and SG-SF. The reason that these dual-positions appeared is because these players played for multiple teams and on each team, they played different positions. To determine which of the two positions these players fit better into, we used a K-Nearest Neighbor (KNN) algorithm.

KNN Process. Because our final analysis will be cluster-based, we decided KNN process is a good method for sorting our six players, because, after-all, nearest neighbors are likely to be part of the same cluster. The steps of our KNN process were:

1) Create a 70-30 training to testing split in our sub-setted data (including only the two positions of the player in question),
2) Find optimal number of nearest neighbors by performing caret library's knn() function on training and testing data and, using parameter kappa, to test the goodness of fitting of 1 through 60 nearest neighbors, and
3) Use optimal number of nearest neighbors to determine respective player's position based on their statistics and the positions of the optimal number of nearest neighbors.

Methodology

We constructed two sets of cluster models based on subsets from our web-scrapped data. First was a dataset with “Percentages” column removed. Second was a dataset with “Attempts” and “Made” columns removed. We did this because these three columns—attempts, made, and percentages—are a function of each other, viz. percentages equals made divided by attempts.

It was initially thought that the first dataset will produce a better clustering because we believe there is important information within the two-valued fraction compared to the single-valued percentage. Specifically, we believe there is a significant difference between 100% 3-point percentage when that comes from a player that shot and made 10 3-point shots versus the player who shot and made only 1 3-point shot.

Optimal Number of Clusters

Clustering methodologies we used were: Model Based, Kmeans, and Hierarchical. In terms of the R functions we used to cluster our two datasets, we had differing degrees of subjectivity in choosing the optimal number.

Kmeans. Starting with the most subjective—Kmeans—the optimal number of clusters is based on the parameter “within-cluster sum of squares” (wss). Essentially, wss is the sum distance within the centroids. As such, the relationship between wss and the number of clusters is exclusively decreasing and approaching 0. Because the wss graph looks like a reciprocal function, we had to pick an optimal number of clusters somewhat arbitrarily—nothing clear-cut.

Hierarchical. More definitive, Hierarchical clustering produced a possible small range of optimal clustering choices. Whereas you can always find lower wss by adding one more cluster, the Hierarchical method usually suggests 1 or 2 optimal numbers of clusters.

One factor contributing to this degree of subjectivity comes from how you believe the distribution should look. If you believe one cluster should contain more observations, you may choose one number; however, if you believe clusters should be more equally distributed, for example, you may choose another.

A final note on Hierarchical clustering: the method for which clusters were determined in our project was the Ward.D2 method. Compared to other Hierarchical methods, Ward.D2 produced the most similar clustering as Kmeans and Model Based clusters.

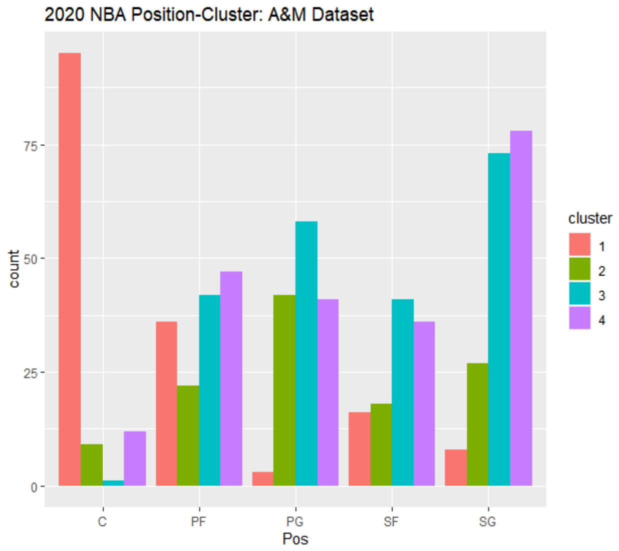
Model Based. Last and least subjective method for determining optimal number of clusters was that of Model Based. In the R function to construct Model Based clusters, we used parameter Bayesian information criterion (BIC). BIC attempts to find the best goodness of fit by neither overfitting nor underfitting. When comparing BIC, we look for the smallest value.

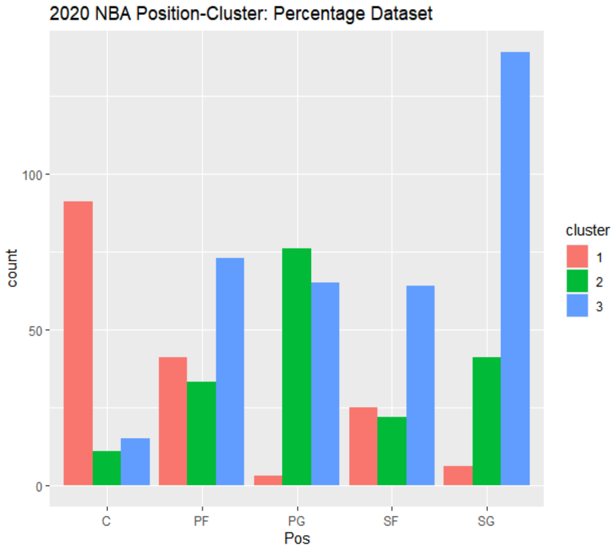
Because we were given a specific value of mathematical rigor and precision, we judge the subjectivity of this method was minimized and that its suggestion for optimal number of clusters to be most informative. That is, whatever number of clusters suggested by the Model Based clustering, we cross-checked whether that same number was appropriate and justifiable for the above two methods. As it turns out, in both data sets and both clustering methods; the Model Based optimal number of clusters were appropriate for Kmeans and Hierarchical.

Model Comparisons

In our two datasets, we found the above methodologies suggested 4 and 3 clusters, respectively. As such, we used 4 and 3 clusters for each method—Kmeans, Hierarchical, and Model Based—for their respective datasets. Below, we will compare:

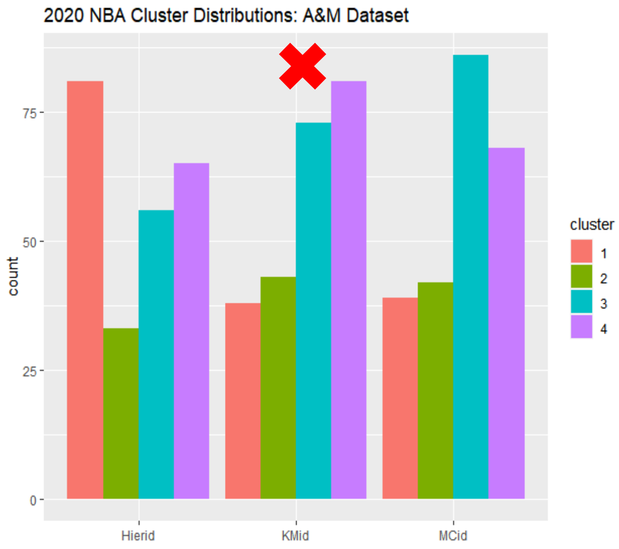
Clusters & Positions

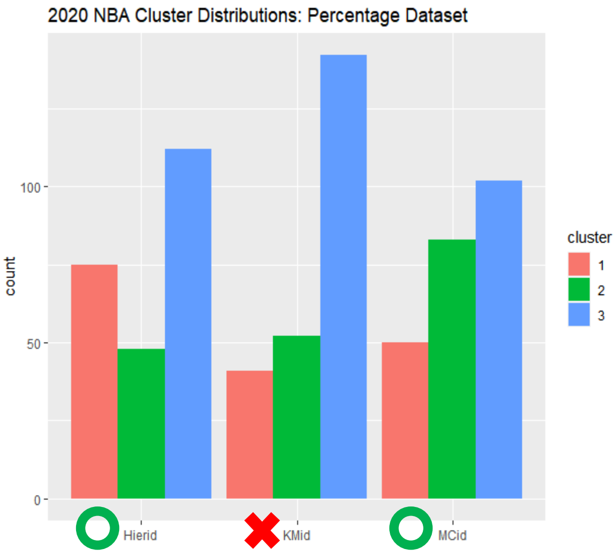

2020 NBA Position-Cluster: A&M Dataset


2020 NBA Position-Cluster: Percentage Dataset

Above, to the left, we have our attempts and made dataset split into its 4 clusters; and to the right, we have the percentage dataset split into its 3 clusters. First, we'd like to note how there is confirmation that our suspicion that positions don't foretell function. That is to say, we see a distribution of clusters, i.e. functionality, in both dataset over every position. Second, we note how certain positions have more distinct functionality than others. For example, in both datasets, we see Centers are predominately cluster 1 and Shoot Guards are mainly cluster 3(/4).

Cluster Distributions


2020 NBA Cluster Distributions: A&M Dataset


2020 NBA Cluster Distributions: Percentage Dataset

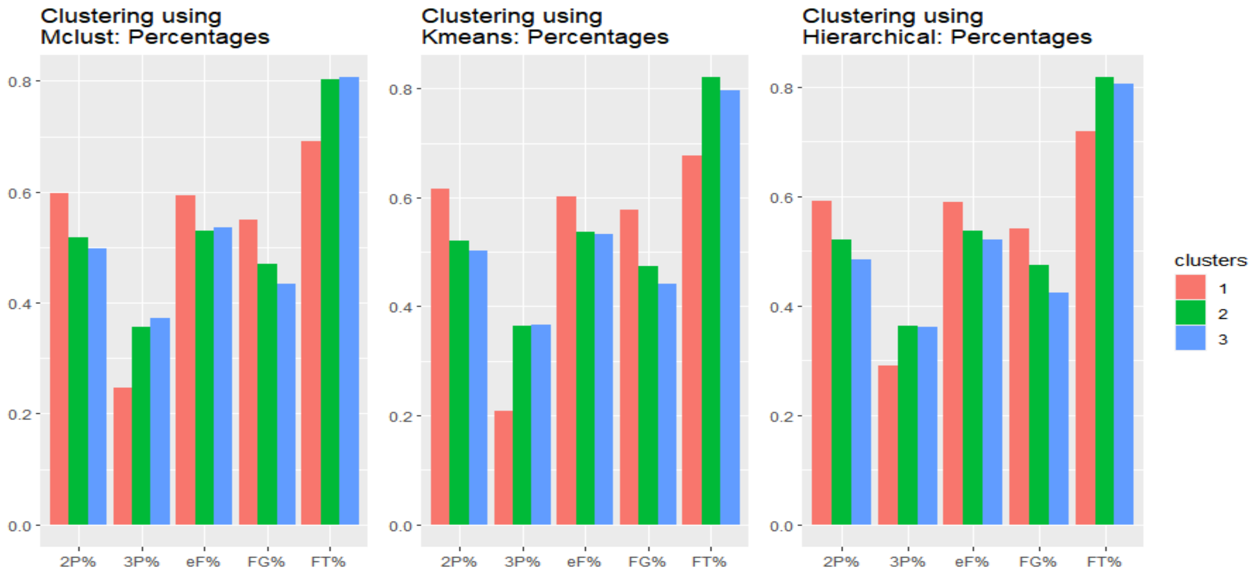
In these comparative bar charts, we look at the two dataset's cluster distribution based on their respective clustering method. Notice the "x" on the attempts and made chart. We put the "x" there because this dataset with its clustering were determined worse in terms of describing functionality. We will discuss this further below, but, in short, clusters 3 and 4 in the made and attempted dataset is more-or-less a high and low grouping of cluster 3 in our percentages dataset.

Now, we can focus on the percentage cluster distribution chart to determine the clustering methods we prefer most. Let's start by saying which method we found worst—Kmeans. We did not like Kmeans' clusters because 1) least equally distributed clusters and 2) highest proportion of cluster 3. Hierarchical and Model Based models are both promising. Best case for Hierarchical: least amount of cluster 2's, as we will see, this cluster may be expected to be smaller. Best case for Model Based: least amount of cluster 3's, as we will see, this cluster is a bit of an "else" statement of "if not clusters 1 and 2, then 3".

Final Model and Functionalities

Recall from above, we consider our ***percentage dataset with 3 clusters as our final model***. We will consider all clustering methodologies in our analysis because we feel the final decision on the method is necessary only whence we want to categorize our players. That is, generally, and as we will see, the patterns and features of each cluster in all three methodologies are *very* similar. Although, players sorted into each cluster are somewhat different, each cluster's functioning is almost identical. Let's see plots on each cluster's functionality:

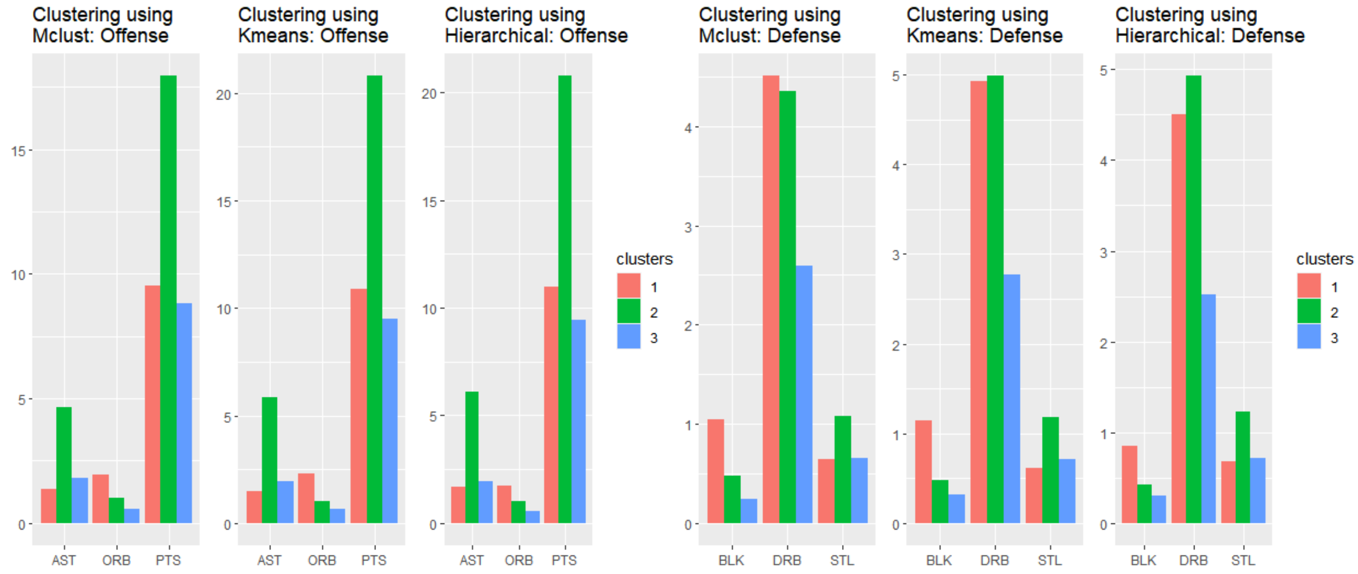
Shooting Percentages



Cluster 1. Cluster 1 has highest 2-point percent, effective field goal percentage, and field goal percentage. They also have the lowest 3-point percentage and free throw percentage. In terms of their shooting percentages, cluster 1's are the barbell group, never being in the middle, only highs or lows of each category.

Cluster 2 and Cluster 3. These two clusters don't really stand out against each other in statistics for shooting percentages. Every shot type, clusters 2 and 3 are neck-and-neck.

Offensive and Defensive Statistics

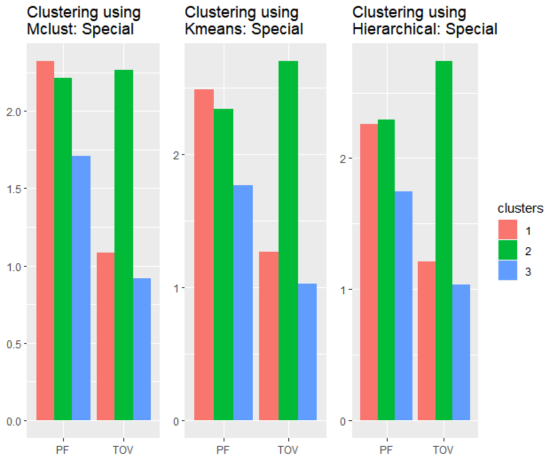


Cluster 1. Cluster 1 has highest rebounds—particularly offensive, but also total—and blocks. They tend to be low in the assists category.

Cluster 2. Here is where we see the key trait of cluster 2—a huge number of points and assists as compared to the other two clusters. They also steal more and have similar number of defensive rebounds as cluster 1's.

Cluster 3. In terms of offensive and defensive statistics, cluster 3 again seems uninteresting. They don't stick out as highs in any statistic, yet they are decisively low in blocks and rebounds—both offensive and defensive.

Personal Fouls and Turnovers



Cluster 1. Cluster 1's tend to have similar number of personal fouls as cluster 2's.

Cluster 2. Another key trait of cluster 2 is their high number of turnovers compared to other two clusters.

Cluster 3. Cluster 3's commit the least number of personal fouls and turn the ball over about as often as cluster 1's.

Naming Clusters

Based on the descriptions above, we will now name the clusters with the intent to evoke a network of associations that will paint a mental picture of their functionalities. The three names cluster names are; In-the-Paints, Generals, and Versatiles.

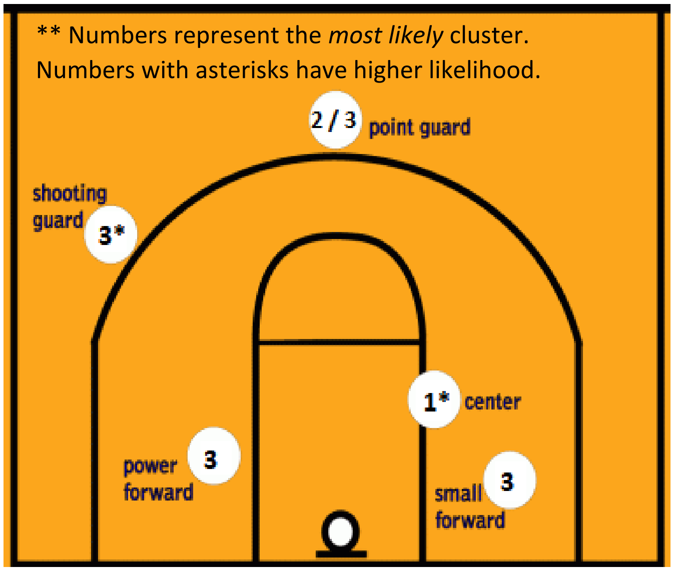
In-the-Paints. We call cluster 1 "*In-the-Paints*". Recall that cluster 1's have the most rebounds and blocks. This suggests these players are closer to the basket; blocks and rebounds overwhelmingly occur closer to the basket. This nearness to the basket also helps explain why these players have the highest shooting percentages—it is no wonder their shots, often very close to the basket, go in more often. Similar with their pathetic 3-point percentages, they are not used to being so far away from the basket, they can't help but to miss from that far away.

Generals. We call cluster 2 "*Generals*". Generals' key characteristics are high points, assists, steals, and turnovers. This is the playmaker, the quarterback of basketball, if you will. On offense, he controls the tempo of the game by controlling the ball the most (suggestive by the high number of turnovers) and facilitate scoring opportunities for himself (high points) and his teammates (high assists). Conversely to the General controlling the ball the most, on defense, he is guarding the opposing player whom is also controlling the ball the most. As such, we expect (and see) Generals have the most steals, as well.

Versatiles. We call cluster 3 "*Versatiles*". Versatiles tend to be very vanilla, if you will, not standing out really in any statistic. We mentioned before how we can crudely call this cluster an "if-else" category, where if a player is not an In-the-Paint or General, then they are thrown into this catch-all Versatiles grouping. They score the least points, they get the least rebounds, they foul the least, and turn the ball over least. The last two points, inversely compared to our Generals, suggests Versatiles 1) don't possess the ball on offense all too often, and 2) don't defend players whom possess the ball so often. It would appear they're decent at one-timers, that is, receiving a pass and directly shooting (low turnovers, yet high 3-point percentage).

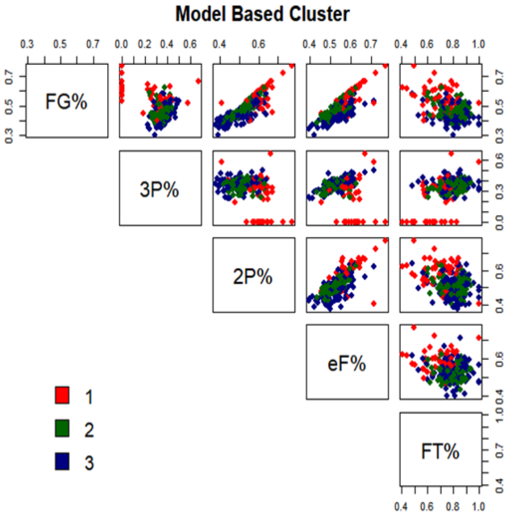
Visualizations

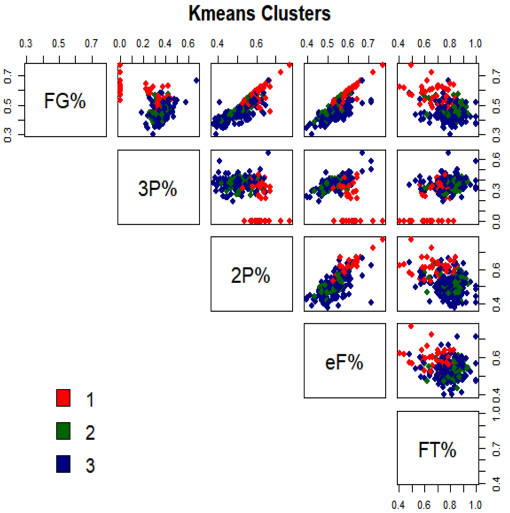
Now that we’ve seen statistical comparisons, let’s turn to area-based visualizations. First, let’s take a look at the most likely cluster of each position:

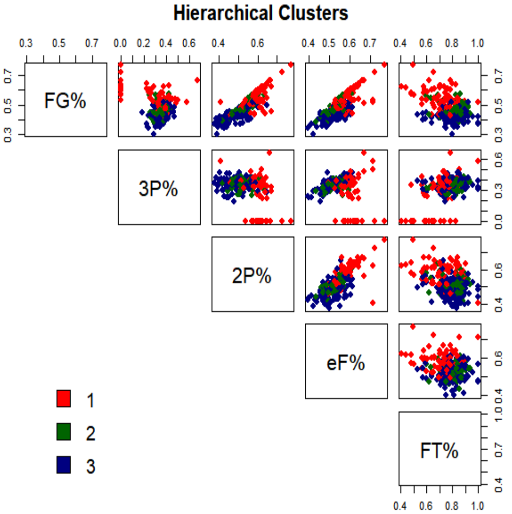

** Numbers represent the *most likely* cluster. Numbers with asterisks have higher likelihood.

In this picture, we used an above graph from the Model Comparison section under the “Clusters & Positions” heading. Keep in mind, each position has a distribution of potential clusters, however, we indicated a position’s *most likely* cluster based on the distributions. Also notice how Versatiles can play in 4 of the 5 positions. This is another reason for their being called Versatiles.

Pairs Plots

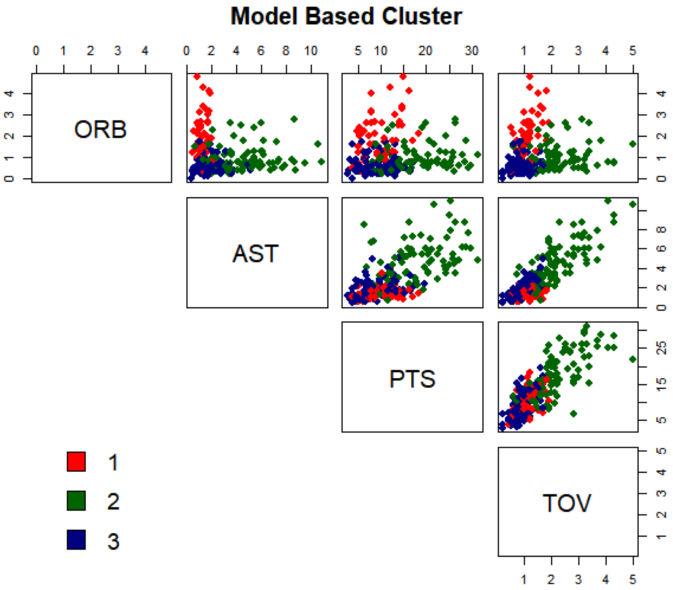

Model Based Cluster

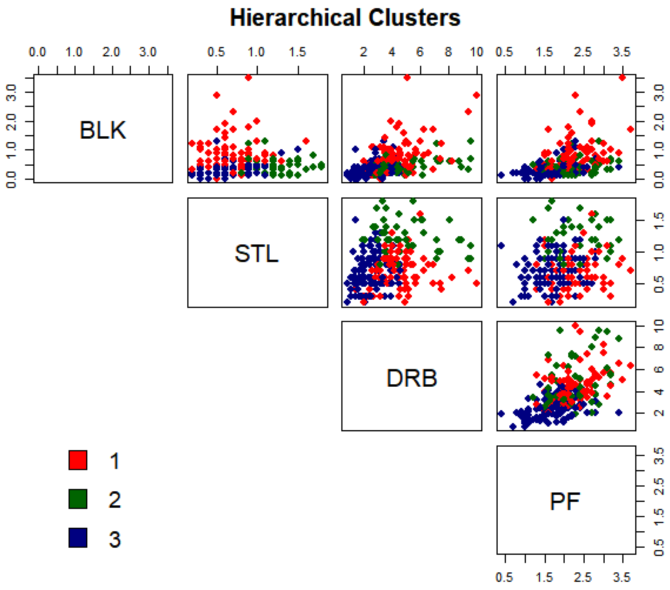

Kmeans Clusters


Hierarchical Clusters

We included the pairs plots of all three methodologies to see the similarities of general *areas* for each cluster. Although, individual players often change from Generals to Versatiles, for example, when sorted by Kmeans vs Model Based, the functions of each cluster remain approximately the same, save some hazy perimeter boundaries.

Having seen the similarities in the areas of clusters of the three methodologies, we will turn to offensive and defensive statistics of Model Based and Hierarchical models, respectively.





These pairs plots sing the same song as we explained in the previous section. In short:

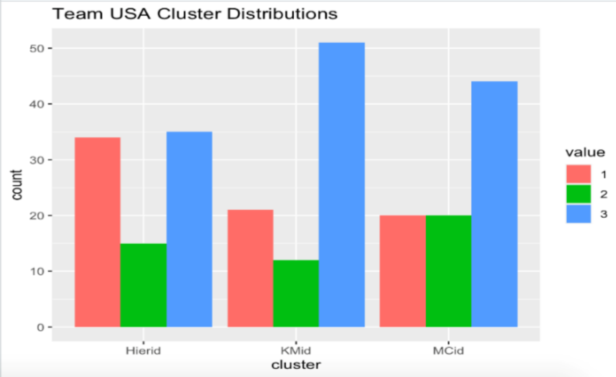
> ***In-the-Paints***: High rebounds and blocks; ***Generals***: High points, assists, and steals; and ***Versatiles***: quasi if-else category or catch-all for basketball's everyman.

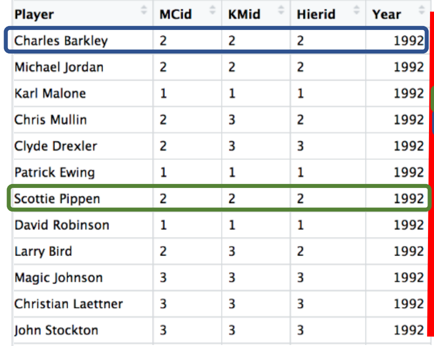
Extensions and Implications of Final Model

Now that we have this new framework for analyzing the NBA, we'd like to use this knowledge and apply it in two interesting scenarios: 1) Using our final model, what is the make-up of Team USA's Olympic team as compared to the NBA of 2020, and 2) In championship teams since 1980, how have functionalities changed over these years?

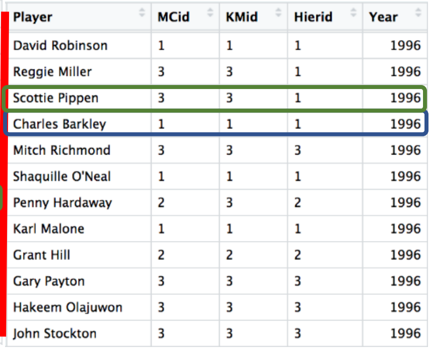
Olympics: Team USA

The Winter Olympics, when basketball is played, occurs once every four years. It is a huge event because, as we know, the Olympics is played by the top athletes in the world. Our hypothesis is that because the Olympics draws the best players from each country to compete once every four years, the vast majority of these players will be Generals. Below is a distribution of clusters from Team USA since 1992 as well as two sample tables to help explain our findings.





Player	MCid	KMid	Hierid	Year
Charles Barkley	2	2	2	1992
Michael Jordan	2	2	2	1992
Karl Malone	1	1	1	1992
Chris Mullin	2	3	2	1992
Clyde Drexler	2	3	3	1992
Patrick Ewing	1	1	1	1992
Scottie Pippen	2	2	2	1992
David Robinson	1	1	1	1992
Larry Bird	2	3	2	1992
Magic Johnson	3	3	3	1992
Christian Laettner	3	3	3	1992
John Stockton	3	3	3	1992



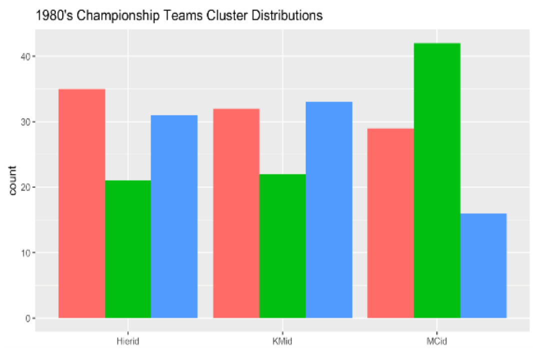
Player	MCid	KMid	Hierid	Year
David Robinson	1	1	1	1996
Reggie Miller	3	3	1	1996
Scottie Pippen	3	3	1	1996
Charles Barkley	1	1	1	1996
Mitch Richmond	3	3	3	1996
Shaquille O'Neal	1	1	1	1996
Penny Hardaway	2	3	2	1996
Karl Malone	1	1	1	1996
Grant Hill	2	2	2	1996
Gary Payton	3	3	3	1996
Hakeem Olajuwon	3	3	3	1996
John Stockton	3	3	3	1996

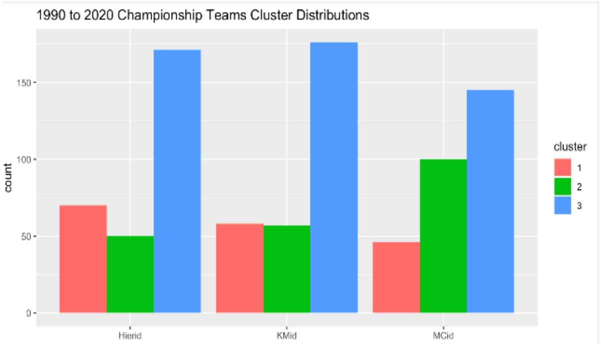
The bar graph displays the frequency of players that were categorized into one of three clusters. From the Hierarchical method, we see that Team USA has a near-equal distribution of In-the-Paints as Versatiles. From the Model Based method, we can see a major difference in that, similar to the distribution of today’s players, Versatiles are the highest frequency players.

Most interestingly we found two players that played in both 1992 and 1996 Team USA we identified very differently. Charles Barkley and Scottie Pippen were both classified as solid Generals on their 1992 teams, however, in 1996, Pippen performed mostly as a Versatile and Barkley as an In-the-Paint. This change led us to question the objectivity of our model. We realized something hugely important: Unlike positions that are player-dependent (player [X] *is* a Center or player [Y] *is* a Point Guard), our functionalities are *team*-dependent.

In short, we believe that by changing our focus from discussing positions (player-dependent) to functionalities (team-dependent) we will have “bumped-up” conversations from the individualistic, superstar level to team dynamics and relations between players—a whole new, interesting conversation to be had.

NBA Championship Teams since 1980s





The above plot has the data from our NBA championship team from 1980 to 2020. In this plot we could see that in the 80s, there was a high distribution of In-the-Paint players compared to Generals and Versatiles, however, starting from 1990, we see the categorizations similar to that of today's general distribution. It is revelatory to note how 1979-1980 season was the first season with the introduction of a 3-point shot. Previous seasons there was no reason to take a long shot, no added benefit of an extra point and all the negative of a harder shot. In effect, the 80's was the decade of learning a new game—the 3-point long game.

Conclusion

In reference to our research question, "How can we operationalize players based on their statistics?", we constructed a final model of operationalization with three categories:

> ***In-the-Paints***: High rebounds and blocks; ***Generals***: High points, assists, and steals; and ***Versatiles***: quasi if-else category or catch-all for basketball's everyman.

We (believe we have) discovered by further analysis that these functionalities describe basketball at a different level—instead of by an area-specific, individualistic position; we can talk about dynamics and interrelations within a team. Truly, the game has changed since the 80's if only that was a decade learning how to incorporate a 3-point shot. It took a full decade to really adjust to the new game and another two decades to fine tune into an approximate cluster distribution we see today. It’s been a great journey for players and fans alike and we look forward to the next decades and rule changes to come!

Limitations

First, we must address the question of model accuracy. We have no metrics. In fact, I suppose, in any new-clustering projects, there wouldn’t be any accuracy measures as there is no underlying *true* model to compare the constructed model against.

Second, these functionalities are time specific. That is, had we used data from the 80’s to construct our final model and had to name them; 1) demarcations of what is considered one cluster opposed to another would be different and 2) names that which evoke the functionality would also therefore be different. The game changes, we just wonder how fast and at what scale.

Third, methods of which to cluster data. For our project, we opted for an unsupervised method we learned in class. During our research, testing, experimentation, and implementation phases, we tried all even supervised clustering methods using PCAs, LDAs, stepwise, etc. These methods, in theory, can produce fine clusters of different numbers and functions, but it wasn’t, we decided, what we were ultimately aiming to achieve.

Future Questions

Already we outlined and began some preliminary extension questions to be explored such as changes of functionalities throughout the years. We think a proper timeseries analysis of the changes in functionalities would produce insightful results. Also, considering how when we refer to player functionalities, we are talking about it *within the context of a particular team*: a further question may involve the composition of teams that work best and/or worst against other team composition. There are still so many fun and interesting questions to be asked and explored.

References

"2020-21 NBA Player Stats: Per Game." *Basketball-Reference.com*, www.basketball-reference.com/leagues/NBA_2021_per_game.html.

Goldsberry, Kirk. "Seven Ways the NBA Has Changed since Michael Jordan's Bulls." *ESPN*, 30 Apr. 2020, www.espn.com/nba/story/_/id/29113310/seven-ways-nba-changed-michael-jordan-bulls.

Newport, Kyle. "Study Shows How NBA's Style of Play Has Changed since 1980." *Bleacher Report*, bleacherreport.com/articles/1955250-study-shows-how-nbas-style-of-play-has-changed-since-1980.

Notes and Slides from Dr.Sarkar

Skallas, Paul. "Refinement Culture (PT 1)." *Paulskallas.substack.com*, paulskallas.substack.com/p/refinement-culture.