Daiwik Swaminathan, Soma Tummala, Yashwant Sathish Kumar
CSC 466
Dr. Rwebangira

Reimagining NBA Player Categorizations Through Clustering

Background and Motivation

In this project, we set out to improve how NBA players are classified. Currently, players are classified using positions (Point Guard, Shooting Guard, Small Forward, Power Forward, Center), which has stood as the standard for years. However, with the evolution of basketball and play styles over time, these positions no longer accurately represent how players play, leading people to call the current game 'position-less'.

Therefore, we sought to cluster players into new, more representative roles using K-Means Clustering. Using data from the nbastatR library, we ran it through the K-Means clustering algorithm and were able to come up with 10 clusters, or roles, that better represented playing styles for players descriptively.

We believe that this information would be helpful for NBA fans and personnel alike to define players intuitively. For members of the coaching staff for NBA teams, defining players summatively by role allows them to quickly understand exactly what kind of production they'll be getting from players, which is helpful when scouting opponents or when trying to replace the value of players in the event of their long or short-term departure. This level of intuitiveness would also be experienced by fans, which could help new fans become close followers of the NBA given the decreased learning curve.

Previous Research

A previous study by Hassan Oukhouya utilized hierarchical clustering to classify NBA players into new positions. Oukhouya utilized principal component analysis to reduce dimensions of his data, scraped from NBA.com, then used hierarchical clustering to group players into related clusters based on the scraped statistics.

Oukhouya utilized different data from us, opting to scrape stats off of NBA.com, and these stats were different from what we used as well. Many of the stats were similar, pulled from the box score, but they were not "rate adjusted" meaning that they didn't account for time on the court. For example, as a measure of passing, Oukhouya used "assists" totals as one of his features while we used "assist percentage", which is a rate-adjusted metric.

Oukhouya explored grouping players in 3 or 4 clusters, providing insights into player groupings and performance metrics. He used the insights to come up with a measure of on-court value for players and ranked them based on importance.

Dataset

For our dataset, we wanted to leverage as much advanced/insightful data as possible. We knew that there were several websites that store immense amounts of data for every single player. In the end, we decided to go with Basketball-Reference.com so that we could make use of an R package (nbastatR) to easily scrape the data. Specifically, we were able to utilize the Bref_players_stats function within the nbastatR library to scrape player 'total' (box score stats) and 'advanced' stats (data that normally would not show up in a box score) from Basketball-Reference.com.

As for the time frame of data we wanted to take, we decided that it would be best to only select data from 2016 onwards. This is because 2016 was around the time in NBA history where play styles/schemes started to change significantly to involve the three pointer much more than before (Shea). This was one way which helped to standardize our data.

After performing the scraping, we ended up having nine seasons worth of data spanning 4,953 rows representing NBA players for each of their respective seasons from 2015-2016 to the current 2023-2024 season. With regards to what the data encompasses, here are the 14 attributes we had access to:

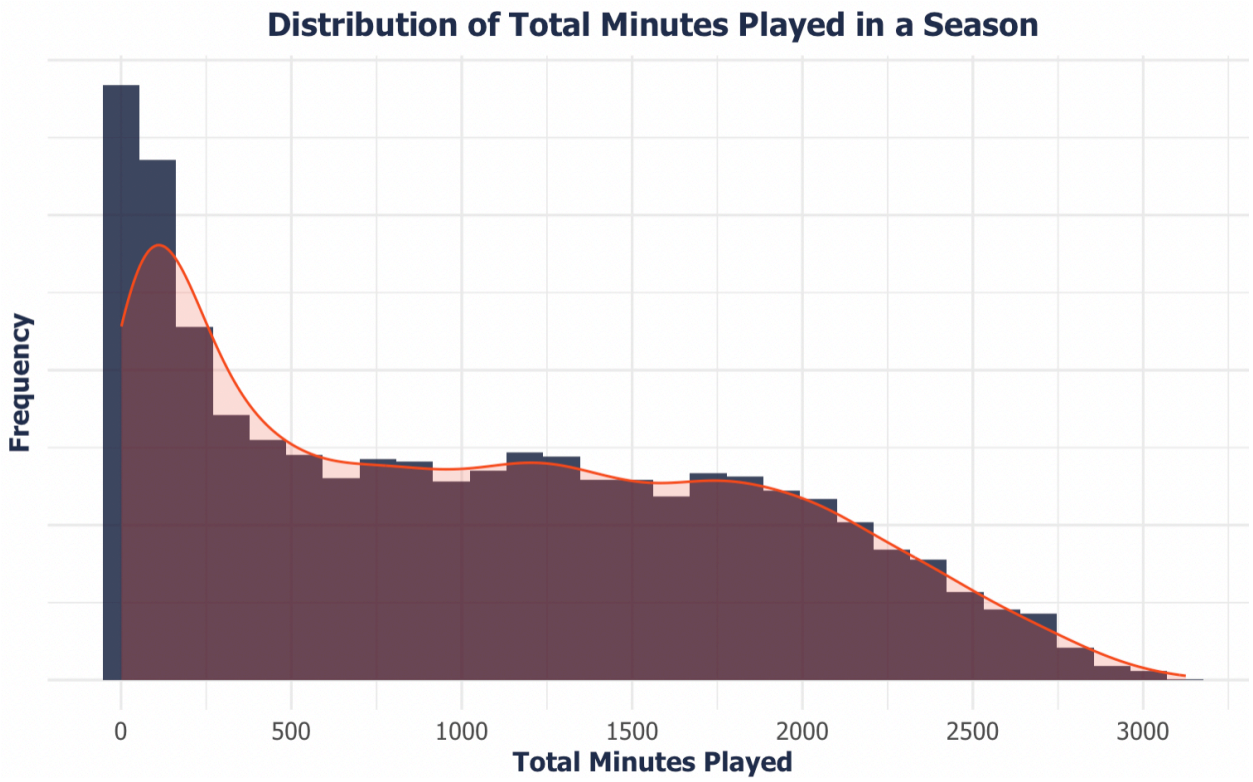
- pctFG3: percentage of 3-point shots made
- pctFG2: percentage of 2-point shots made
- pctFT: percentage of free-throw shots made
- fg3aPerMinute: 3-point shots attempted per minute
- fg2aPerMinute: 2-point shots attempted per minute
- pctFTRate: free throw shots / total number of field goal attempts
- pctORB: % of Total Team Offensive Rebounds a Player Collected while on the court
- pctDRB: % of Total Team Defensive Rebounds a Player Collected while on the court
- pctAST: % of Total Team Assists a Player Collected while on the court
- pctUSG: a % estimate of how many possessions a single player uses of their team
- pctSTL: % of Total Team Steals a Player Collected while on the court
- pctBLK: % of Total Team Blocks a Player Collected while on the court
- pctTOV: % of Total Team Turnovers a Player Collected while on the court

- ptsPerMinute: average of how many points per minute a player gets

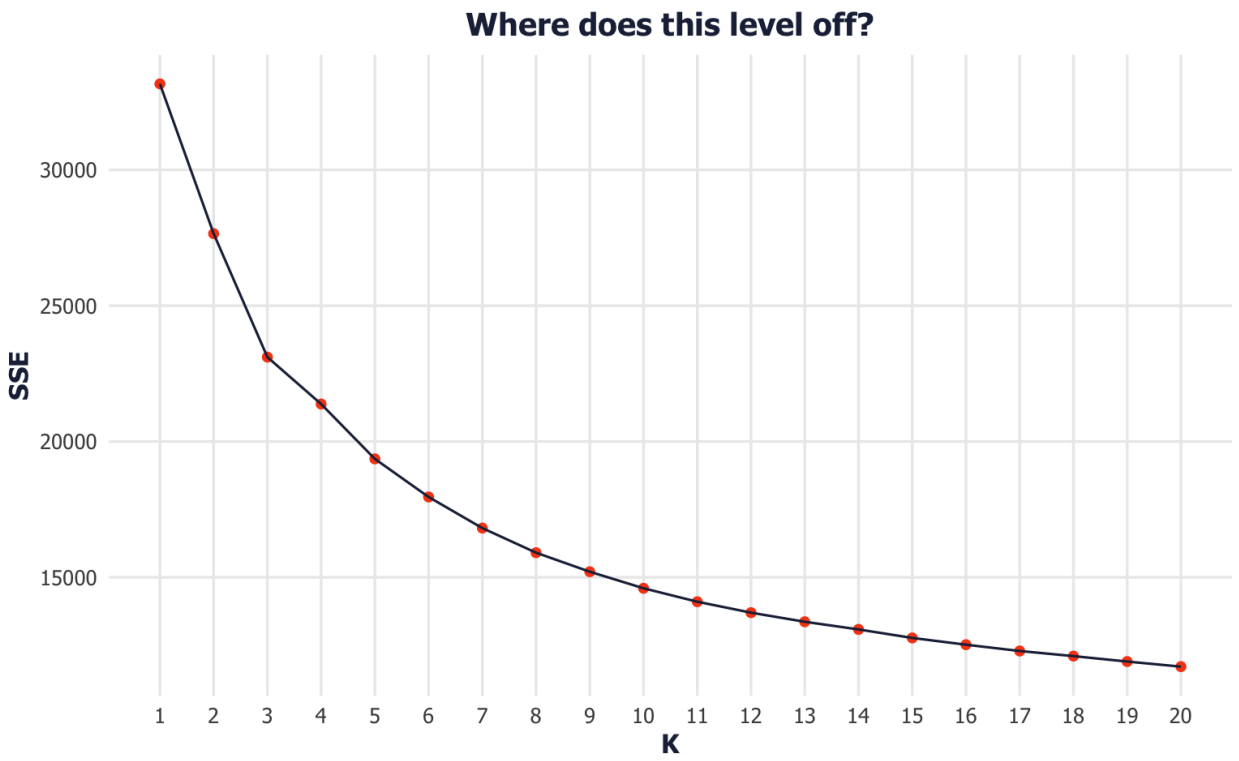
Methodology

We scraped basic and advanced stats based on various statistical attributes encompassing production on both the offensive and defensive side to the game. As a result, we ended up pulling data on 14 total attributes.

Then, we filtered for only the relevant players, those of whom saw a significant amount of court time. In order to do so, we plotted a frequency distribution of total minutes played for each player and noticed a dropoff at around 1000 minutes played. As a result, we filtered the data to keep rows for players that only played over 1000 minutes for each respective season.



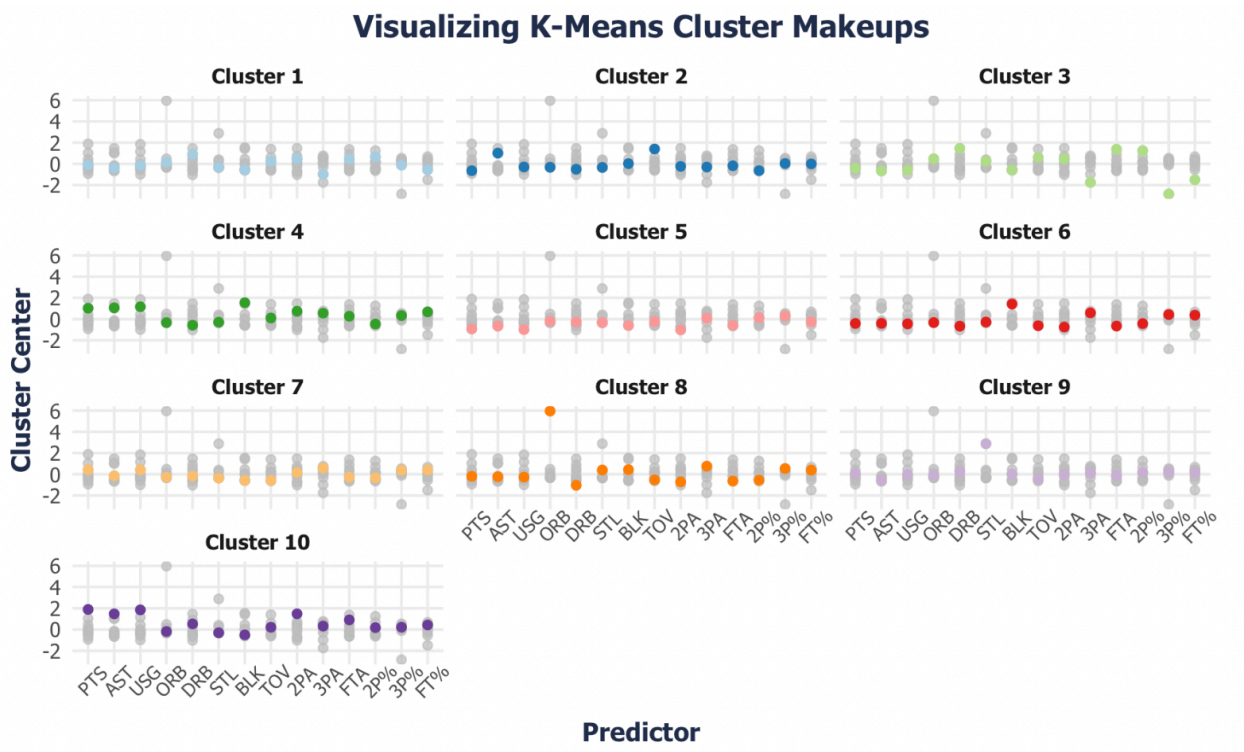
After cleaning the data, we were set up for our analysis. First, we standardized all the data using z-scores for interpretability. Each column value was now a standardized z-score relative to the rest of the column, which we thought would help with interpretability when visualizing our findings later. When moving forward with K-Means clustering, to decide which value of “k” to use, or the optimal number of clusters, we used the “Elbow Method”. As a part of this technique, we plotted the sum of squared differences for each value of k. Where the graph seemed to level off, forming the “elbow”, that was the optimal number of clusters. We ended up going with 10 clusters based on this.



Running K-Means after choosing k=10 clusters, we were able to come up with 10 different roles, which represented 10 groupings of players based on playstyle. In order to determine the playstyle represented by each cluster, we created visuals plotting the different predictor statistics against the standardized cluster centers. These centers stood as a representation for the value of each statistic for each group. For each role, we created a separate visual with its respective cluster centers highlighted. Based on the values shown for each statistical attribute, we were able to come up with names for each cluster. For example, a cluster whose center for PTS has a high value would indicate that players in that cluster were good at scoring.

Results & Visuals

Our K-Means model ended up producing 10 different clusters that grouped players into roles based on their playstyles. The visual below illustrates the difference between each cluster based on the values of the centers.



For each cluster, the (standardized) values of the centers for each statistic indicated the value of the statistic for the given cluster. Based on these center values for each cluster, we were able to come up with names for the clusters that appropriately summarized playstyles:

Cluster 1: Interior Slashers

- High frequency of 2-Point Shots
- High 2-Point Efficiency
- High frequency of free-throws drawn

Cluster 2: Pass-First Playmakers

- Rarely contribute to scoring
- High Assist (Passing) Rate
- High Turnover Rates

Cluster 3: Rebounding Big Men

- High Rebounding Rates
- High 2-Point %
- High 2-Point Attempts

Cluster 4: Elite Perimeter Scorers

- High scoring, assists, usage
- High 3-Point Percentage

Cluster 5: Pure Shooters

- High 3-Point Percentage
- High 3-Point Attempts
- Low rankings across other categories

Cluster 6: "3 & D" Players

- High 3-Point Percentage
- High 3-Point Attempts
- High Block Rate

Cluster 7: Average Perimeter Scorers

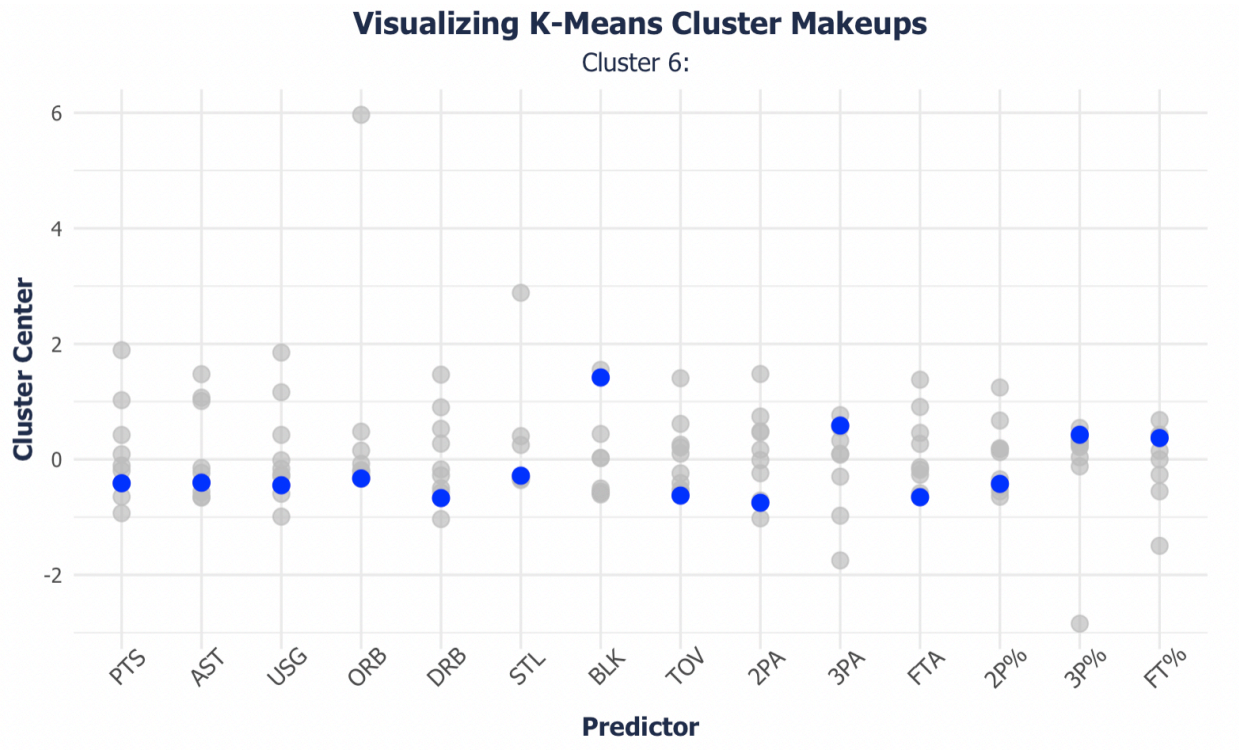
- High 3-Point Percentage
- High 3-Point Attempts
- Low Rankings Across Board for Others

Cluster 8: High-Energy Rebounders
- High Offensive Rebounding Rates

Cluster 9: Defensive Stoppers
- High Steal Rates
- High Rebounding Rates

Cluster 10: Offensive Generators
- High Scoring Rates
- High Assist Rates
- High Usage Rate

We also generated visuals for individual clusters such that they were easier to analyze:



Conclusions:

Breaking apart players into 10 roles as opposed to the current five positions used in the NBA provides benefits for coaches and players in the league.

As a coach, it is important to know who to substitute for injured players during games, and using the predefined NBA roles for this will not always be the best way to substitute a player. Taking LeBron James as an example, who is an offensive generator, it wouldn't make sense to replace him with the backup small forward, who may be a rookie without any experience on the team. It is better to look at both experience as well as overall playmaking ability to determine who replaces Lebron on the court, such as Austin Reaves because of their similar role as an offensive generator.

As a fan, getting into debates about the better player in the league is always an upcoming topic. Saying two players aren't comparable because they play different positions is not always the best way to compare play styles. Overall playing ability is important to consider, as well as how their role affects their team. A better statistic would be a player's over/under which is their impact on and off the court. This is a better indicator of a player's role as opposed to a player's assigned position.

Overall, breaking away from the norm of utilizing players' assigned positions when making decisions or talking about a player's ability is not completely representative of their skill. Using

roles gives a more well-rounded perspective on each player and provides better insights into their skills.

Works Cited

"Basketball Statistics and History | Basketball-Reference.com." *Basketball-Reference.com*, Sports Reference, www.basketball-reference.com/.

Bresler, Alex. "NbastatR Package - RDocumentation." *Www.rdocumentation.org*, RDocumentation, www.rdocumentation.org/packages/nbastatR/versions/0.1.10131.

Oukhouya, Hassan. "RPubs - NBA Player Performance Analysis: PCA, Hierarchical Clustering, and K-Means Clustering." *Rpubs.com*, RPubs, 2 Apr. 2023, rpubs.com/HassanOUKHOUYA/NBA_Player_Performance_Analysis_PCA_Hierarchical_Clustering_and_K-Means_Clustering.

Shea, Stephen. "The 3-Point Revolution." *Shottracker.com*, ShotTracker, shottracker.com/articles/the-3-point-revolution.