Chris Le

Final Report: Defining NBA Player Archetypes Using Clustering

1. Problem Statement

Each year, millions of individuals assemble teams to compete for the championship in fantasy sport leagues. According to the Fantasy Sports and Gaming Association, 62.5 million people played fantasy sports in 2022, with 22% of participants playing fantasy NBA basketball.[1] The fantasy basketball market is projected to grow from $3.8 billion in 2021 to $6.7 billion by 2030.[2] NBA fantasy leagues are won and loss by how much their players contribute to their team's statistical categories

Traditionally, NBA players are assigned into 5 positions: Point Guard, Shooting Guard, Small Forward, Power Forward, and Center based on the general strategic roles they play on the court. However, NBA players are more fluid than the traditional designations would suggest as players have begun to expand their basic skill set to assert dominance all over the court. These players no longer can be defined by the positions they play in. With that in mind, I hope to provide a new way to group players into certain classifications beyond positioning so that NBA teams to capture their statistical production on the floor. Accordingly, team managers can form rosters with players based on how they statistically impact the team.

The goal of this project is to use unsupervised learning models to define NBA player archetypes that represent NBA players by how they truly statistically contribute to their team on the floor better than traditional positions.

References:

1. https://thefsga.org/industry-demographics/
2. https://dataintelo.com/report/global-fantasy-basketball-market/

2. Dataset

The NBA player stats dataset was retrieved from Kaggle. This dataset contains 2021-2022 regular season NBA player stats per game.

Column Descriptions

- Rk : Rank
- Player : Player's name
- Pos : Position
- Age : Player's age
- Tm : Team
- G : Games played
- GS : Games started
- MP : Minutes played per game
- FG : Field goals 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
- 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

3. Data Wrangling

The raw NBA Player Game Statistic dataset contained 812 rows with 30 columns. The final shape of my NBA Player Game Statistic dataset was 500 rows with 26 columns.

Issue 1: Duplicate Player Entries

The first problem encountered when exploring the raw dataset was that it contained duplicate player names. This was because the dataset accounted for players who played on multiple teams during the NBA season by having an entry with a player's total statistics as well as separate entries with statistics based on each team a player was on. For example, Aaron Holiday played 63 total games (TOT) during the NBA regular season, but 41 games were with the Washington Wizards (WAS) and 22 games were with the Phoenix Suns (PHO). So, there would be three separate Aaron Holiday entries: Aaron Holiday (TOT), Aaron Holiday (WAS), and Aaron Holiday (PHO). Since we are only focusing on a player's comprehensive season statistics, the additional 207 rows containing multiple team player's statistics spit into different teams were dropped.

Issue 2: Columns Unrelated to Game Statistics

There were 30 features, but not all of them were relevant to our analysis of a player's individual game statistics. Thus, the rank (RK), age (Age), team (Tm), and games started (GS) columns were dropped.

Issue 3: Inconsequential NBA Players

Lastly, a baseline check of the distribution of feature values was done to spot any potential outliers that may skew player statistics. There is potential for a player to skew the data if their statistics were based only on a handful of games. A more in-depth look at the distribution of games played revealed that there was a consider portion of players skewing the data to the right. Since we want to analyze NBA players who have realistically opportunities to play and impact the game, we filtered the dataset to only contain players who played at least 10 games. This filtered out 105 players from the dataset.

4. Exploratory Data Analysis

Heatmap

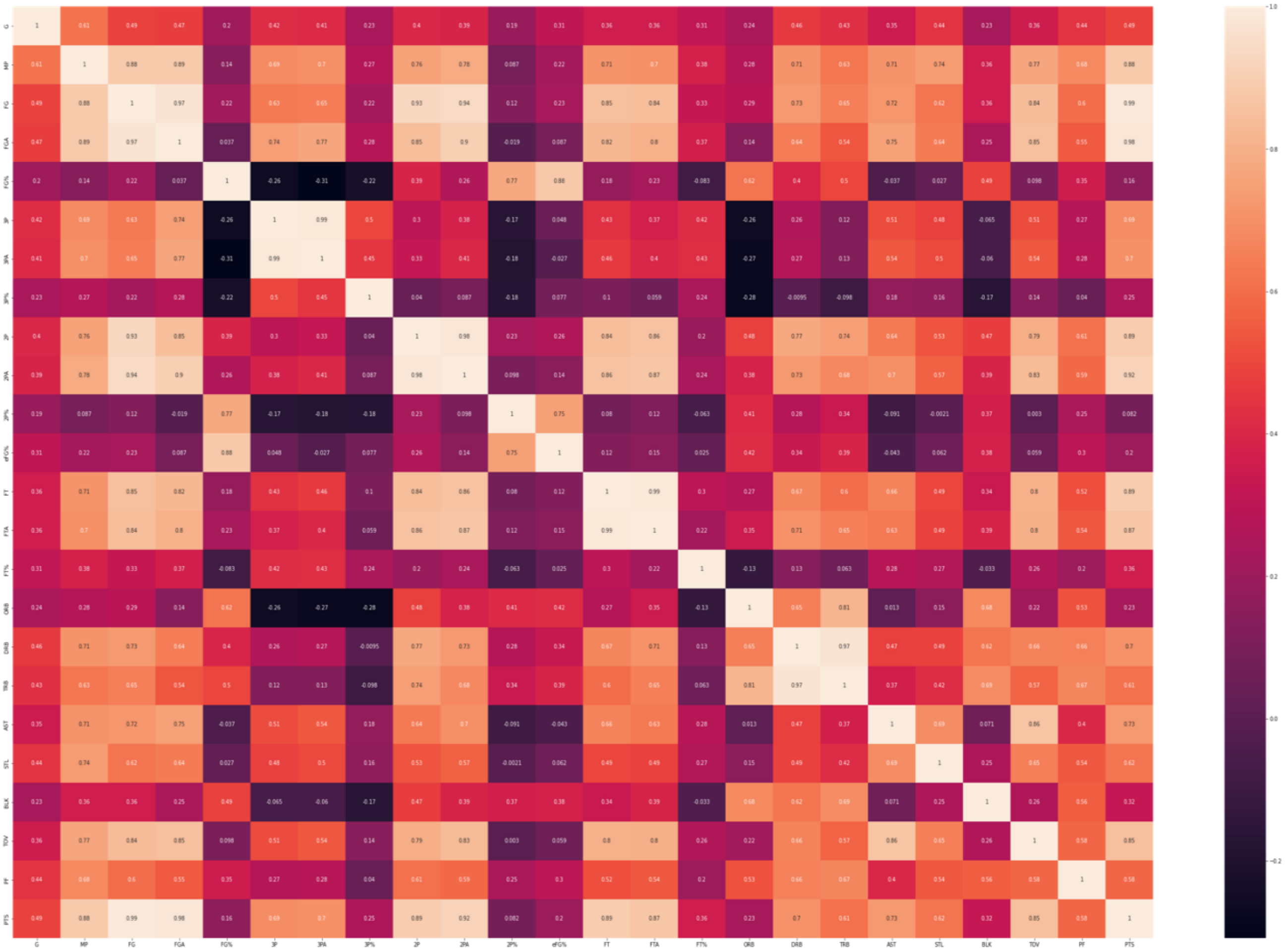

Figure 1: A heatmap representing the correlations between different features.

Besides player names (Player) and position (Pos) columns, the features in the NBA dataset were all numerical. Figure 1 shows a heatmap which revealed that there were many features that were highly correlated with each other: "FG"/"FGA", "3P"/"3PA", "2P"/"2PA", "FT"/"FTA", and "DRB"/"TRB". These correlations were expected because those stats are inherently dependent on each other. Some features that were surprisingly highly correlated were: "AST"/"TOV", "FT"/"2P", "FG"/"2P". Assists (AST) may be associated with more Turnovers (TOV) because players who pass the ball often have the ball in their hand more frequently and thus have more chances to turn the ball over.

Principal Component Analysis (PCA)

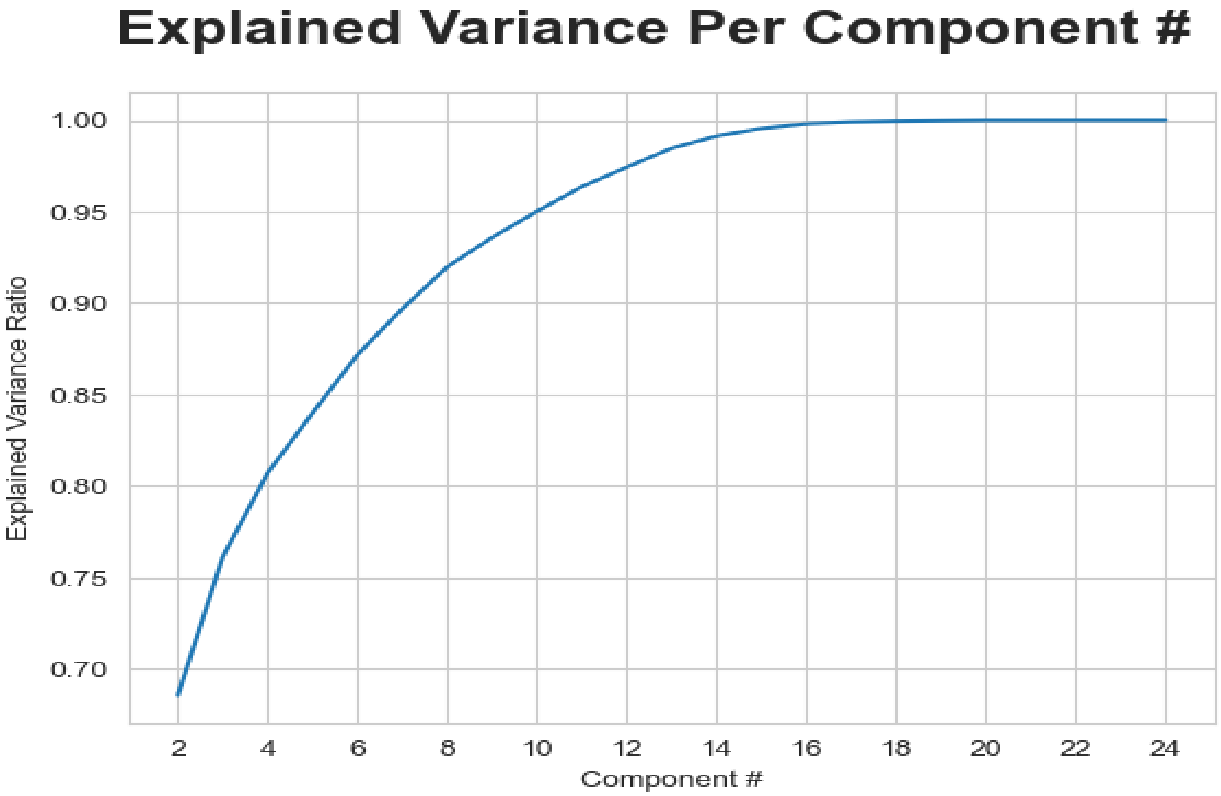

Figure 2: PCA showing how much each principal component contributed to the variance.

Figure 2 depicts that the first 5 components account for ~84% of the variance and the first 10 components for ~95% of the variance. I applied a Principal Component Analysis on the data as a means for dimensionality reduction for more efficient analysis. I narrowed it down to the first 10 PCA components because they account for about 95% of the variance in the data.

5. Clustering Models

K-Means Clustering

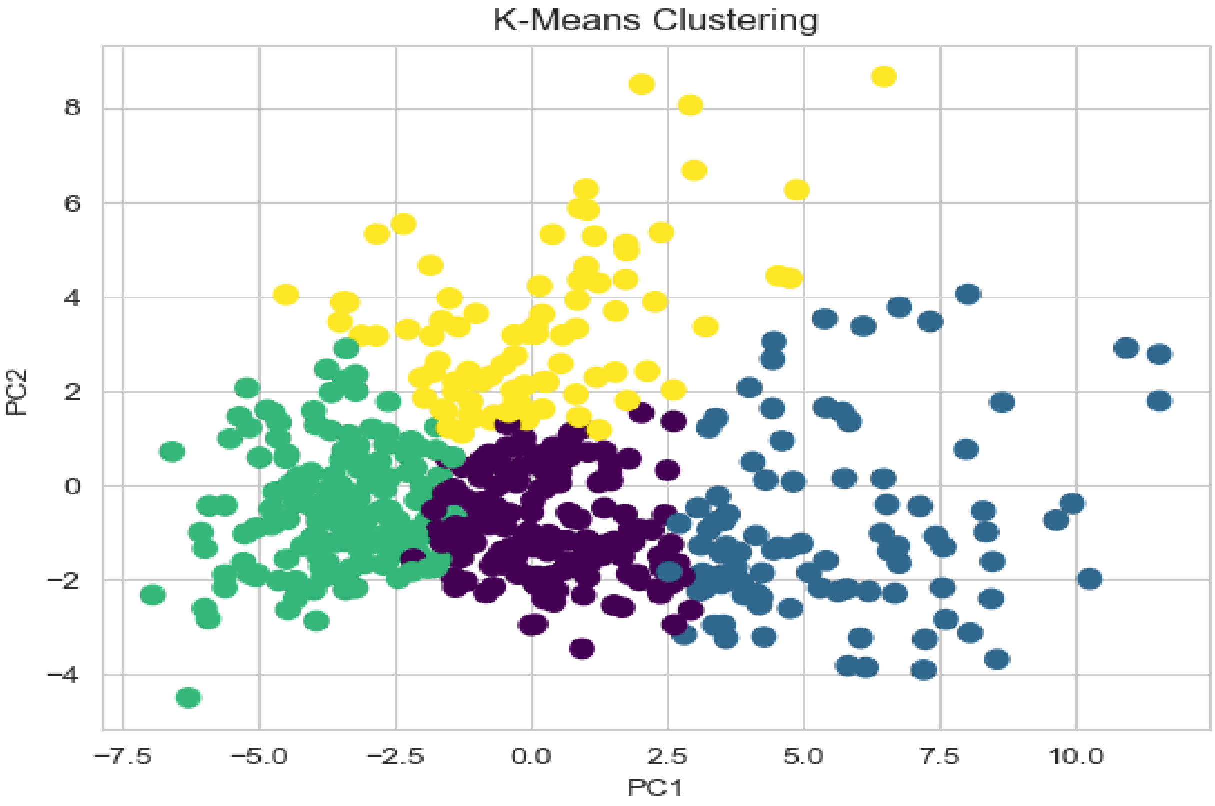

Figure 3: Scatterplot Visualizing K-Means Clusters

The first unsupervised learning model I applied was K-Means Clustering. To identify the optimal number of K to choose for my model, I used the Silhouette Method. Based on the Silhouette Method, the optimal number of K was K=4, with an average silhouette score of 0.222. Figure 4 displays a scatterplot of the K-Means Clustering Model with K=4.

Agglomerative Hierarchal Clustering

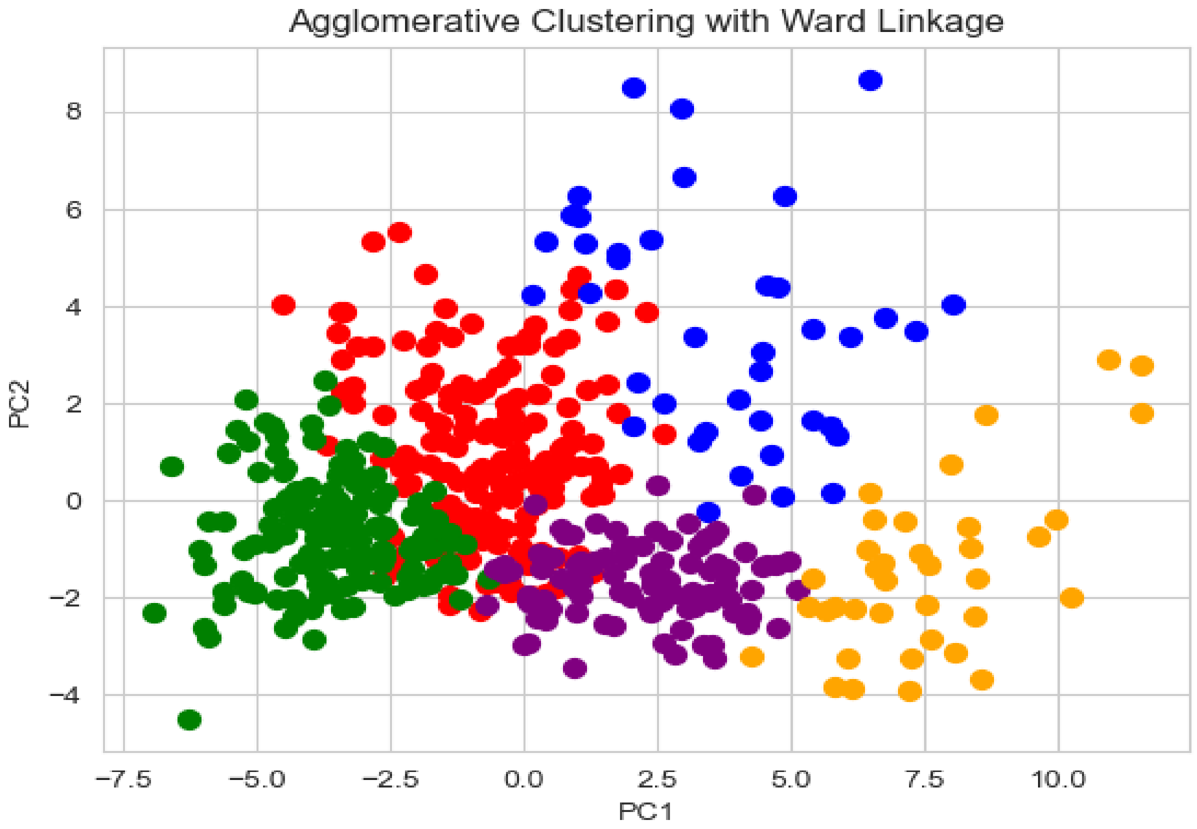

Figure 4: Scatterplot Visualizing Agglomerative Hierarchal Clusters

The second unsupervised learning model I applied was the Agglomerative Hierarchical Clustering with Ward linkage. The optimal number of clusters was 5, which I determined using a dendrogram which allowed me to identify the area on the chart with the highest vertical distance that does not intersect with any clusters. I then applied the Agglomerative Clustering method to the data and figure 4 demonstrates a visualization of the clusters on a scatterplot.

Final Model Selection

For the final model selection, I decided to use the Agglomerative Clustering with Ward Linkage. While the K-Means Clustering had a higher Silhouette Coefficient score of 0.222 (n_clusters = 4) compared to 0.184 for Agglomerative Clustering (n_clusters = 5), I chose to use the Agglomerative Clustering because it offered a clear selection for the optimal number of clusters at K=5 and I also felt that 4 clusters were not enough to differentiate a pool of 500 players.

6. Cluster Interpretation

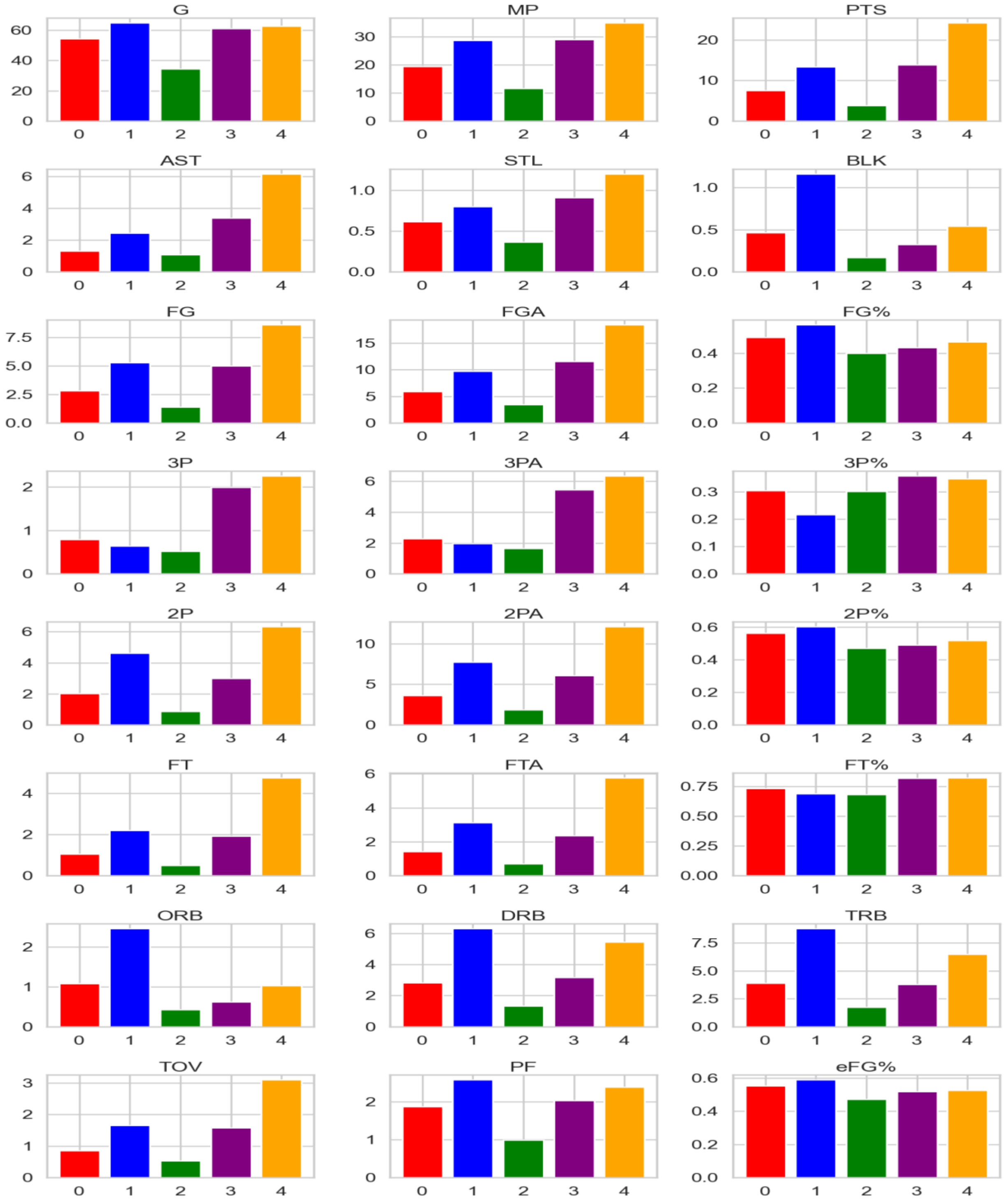

Figure 5: Bar Charts with Aggregated Statistics Based on Clusters

In order to analyze the clusters, I aggregated the clusters to calculate cluster-level statistics based on the mean to visualize how each cluster differs between each feature (Figure 5).

Cluster 0 (Red): Role Players

- Averaged > 54 games played at 19.4 minutes
- Low- to Mid- tiered in most categories: PTS, AST, STL, BLK, TRB
- Notable Players: LaMarcus Aldridge, Danny Green, Danilo Gallinari

Cluster 0 had the highest count of players in the pool with 182. I would interpret this group as the role players because while they do not excel in any particular area, but they play quality minutes and contribute to a wide range of statistical categories.

Cluster 1 (Blue): Traditional Big Men

- Highest-tiered in TRB, BLK, FG%, 2P%
- Low-tiered in 3PA, 3P%, FT%
- Notable Players: Jarret Allen, Rudy Gobert, Clint Capela

Cluster 1 had players that I would consider as traditional big men– meaning they play close to the basket and have high rebounding numbers, blocks, and field goal percentage. However, they are unable to stretch the floor and shoot, thus they have low 3-point and free throw percentage.

Cluster 2 (Green): Bench Players

- Averaged only 34 games at 11.5 minutes
- Lowest-tiered in every category besides 3P%
- Notable Players: Jose Alvarado, Kent Bazemore, Troy Brown Jr.

Cluster 2 was the cluster with the second highest number of players at 145. I consider them bench players due to their low number of games and minutes played– indicating their inability to consistently get on the floor. As a result, these players have the lowest averages across the board.

Cluster 3 (Purple): 3- and D- Players

- 2nd highest-tiered in 3P, FT%, PTS
- High in defensive statistics: STL, BLK
- Notable Players: Desmond Bane, Mikal Bridges, Malcom Brogdon

Cluster 3 had players I would categorize as 3-and-D players. These players specialize mainly in 3-point shooting and are key defensive assets to the team. Accordingly, this cluster had high steals, blocks as well as 3-point makes.

Cluster 4 (Orange): Versatile All-Around Players

- Highest-tiered in most categories including MP, PTS, AST, STL, FTA, TOV
- Notable Players: Giannis Antetokounmpo, Stephen Curry, Lebron James

Cluster 4 had versatile players that contributed across the board and had the highest averages in most categories. They are rare in the NBA, and so, this cluster had the lowest number of players with only 37.

Takeaway

Through Agglomerative Hierarchal Clustering, I was able to group players into five different archetypes that had a blend of players from the traditional positions. Using these redefined clusters, teams can now assemble rosters with NBA players based on their on-the-court statistical contributions rather than what positions they played. Accordingly, these clusters can provide an efficient way for NBA fantasy team managers to shore up their teams by picking players based on their statistical category needs.

7. Future Research

- The dataset used for these models only contained data for the 2021-2022 NBA season. The clustering models can be improved by incorporating data from multiple NBA seasons.
- This clustering model provide 5 clusters. In the future I want to group players into even more clusters to see if there is a way to further differentiate players.
- I would like to expand the dataset to incorporate advanced gameplay statistics that may illustrate playstyles such as spot-up shots, screen assists, isolated field goal attempts, transition shots made, deflections, charges drawn, or loose balls recovered. How do individual playstyle characteristics influence statistical contributions?