Identifying Groups of Rebounding and Passing Forwards or Centers in the NBA

Kelvin Tiongson
kelvinjtiongson@lewisu.edu
DATA-51000-001, Summer 2020
Data Mining and Analytics
Lewis University

I. Introduction

The goal of this analysis was to identify different groups of forwards and centers in the NBA who are good rebounders and good passers. The analysis takes a look at NBA statistics from the shortened season of 2019-2020. The data that was used was scraped from a website that summarized players statistics in the NBA from the 2019-2020 season [1]. This analysis would help an executive working in the NBA or a person looking to build their unique fantasy basketball team by finding better passing rebounders. Rebounds and assists are critical statistical categories in real life and in the fantasy sports realm. In basketball, rebounds and assists are two categories that are always looked upon as a certain player's productivity. In fantasy basketball, those are two of the main categories in a competition.

The future sections of this report describe the dataset, methodology, results along with a discussion, and a conclusion. Section II contains a description of the dataset, frequency distribution, and descriptive statistics of rebounds per game and assists per game. The methodology for analysis is presented in section III. The two cluster methods chosen in this analysis were k-means clustering and hierarchical agglomerative clustering. In section IV, the reports and results of the analysis are discussed. Finally, section V provides conclusions about the tiers of these players in regard to their ability to rebound and pass the basketball.

II. Data Description

The dataset used in the analysis included twenty-nine columns. From the large number of columns, the attributes that were used were the ID of the player, full name, team that the player plays for, player's assists per game, player's rebounds per game, player's total rebound percentage, and a player's assist percentage. The goal of the analysis was to look at players who were not considered guards. Therefore, the attributes used for clustering were RPG, APG, total rebound percentage, and assist percentage. These attributes were normalized so that one feature was not outweighing another. Guards were labeled "G". There were positions that had "F-G" and "G-F". These positions were for players who could be considered as guards or forwards. These players were removed from the analysis. Therefore, the positions that were specifically looked at were centers or forwards. These players had a position labeled as "C", "F", "F-C", and "C-F".

TABLE I. 2019-2020 NBA Player Stats

Attribute	Type	Example Value	Description
ID	Nominal (primary key)	35	Record identifier
FULL NAME	Nominal (string)	"Lebron James"	Name of the player
TEAM	Nominal (string)	"Mia"	Team of the player
APG	Numeric (real)	2.4	Assists per game
RPG	Numeric (real)	9.4	Rebounds per game
TRB%	Numeric (real)	19.3	Total Rebound Percentage
AST%	Numeric (real)	13.5	Assist Percentage

From Figure 1, the frequency distribution of rebounds per game for forwards and centers varies but has a higher frequency between 2 and 6. The mode of rebounds per game is 2.5. In Figure 2, the frequency distribution of assists per game for forwards and centers has a higher frequency between 1 and 2. The mode of assists per game was 0.7.

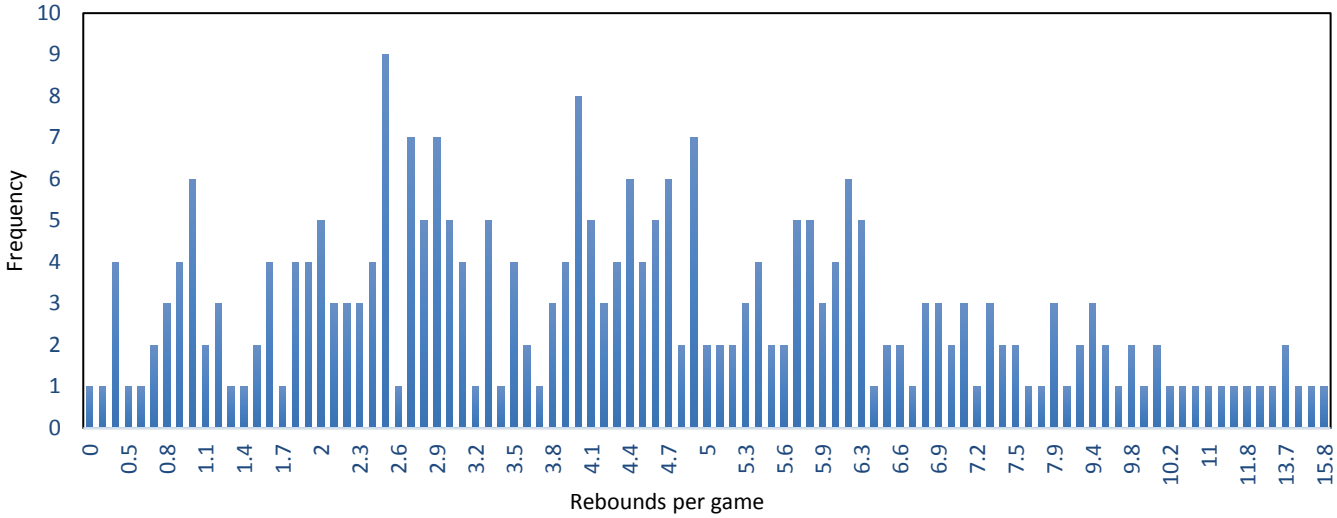

Fig 1. Frequency Distribution of Rebounds per game

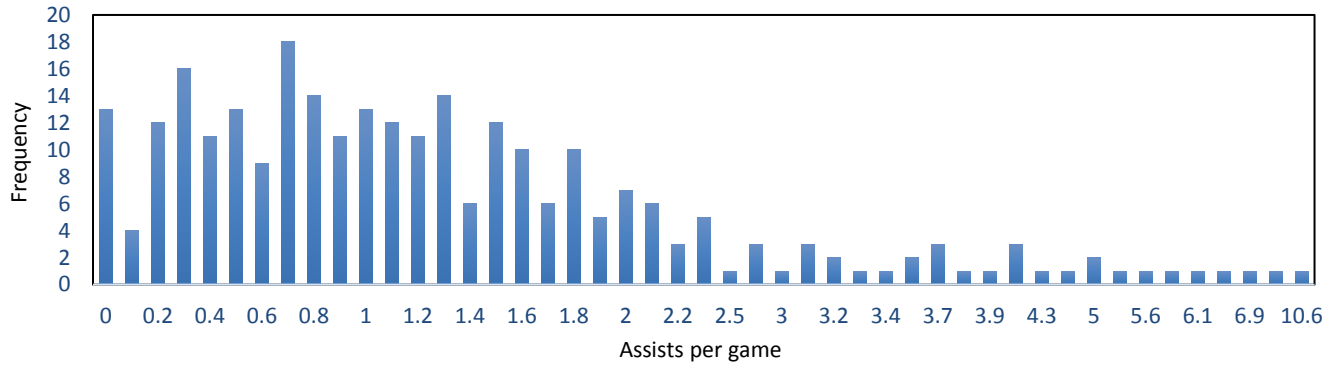

Fig 2. Frequency distribution of assists per game.

Wanting to take a deeper dive into the data, I decided to compute the descriptive statistics of each numeric category. From Table II, the two numeric attributes had a mean of 1.41 for APG and a mean of 4.61 for RPG. Their standard deviation was 1.38 for APG and 2.88 for RBG. The maximum value for each category was 10.6 for APG and 15.8 for RPG. Graphing the frequencies and calculating the mean of the two statistical categories, it is clear that forwards and centers rebound the basketball better than passing the basketball.

TABLE II. DESCRIPTIVE STATISTICS OF APG AND RPG

	APG	**RBG**
Count	275	275
Mean	1.412	4.610182
Standard Deviation	1.384073	2.882094
Minimum	0	0
25%	0.55	2.5
50%	1.1	4.2
75%	1.8	6.1
Maximum	10.6	15.8

III. Methodology

Looking at the data, I wanted to use both k-means and hierarchical agglomerative clustering methods in order to take a look and see if there were different groups of players that were forwards or centers that could pass and rebound the basketball. For the cluster analysis, the four attributes used were rebounds per game, assists per game, total rebounds percentage, and assists percentage. Rebounds per game and assists per game calculate the average a player rebounds and assists per game played. Total rebounds percentage is an estimated percentage of available rebounds grabbed by the player while the player is on the court [2]. Assists percentage is an estimated percentage of teammate field goals a player assisted while the player is on the court [3]. Beginning with k-means clustering, the silhouette scores calculated from Figure 3 shows that 2 clusters was the best choice for k.

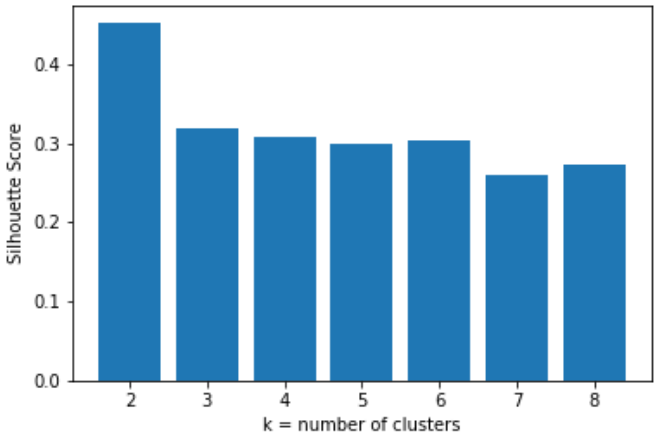

Figure 3. Silhouette score for k number of clusters

Even though each data point was put into its calculated cluster, there were a few outliers within them. So, I decided to look at hierarchical clustering to see if players were placed into different clusters. With hierarchical clustering, the distance metric that made the most sense to use between the statistical features was Euclidean distance. The linkage function that was chosen was the average linkage. Since I chose k = 2 number of clusters for k-means clustering, I felt that it was best to choose 2 clusters for hierarchical clustering.

IV. Results and Discussion

Figure 4 separates the data into two clusters. From Figure 4, cluster 1 represents players who are not as talented at passing but can rebound the basketball. Cluster 2 represents the better players that can rebound and pass the basketball. These players are better in one category. These players in cluster 2 can also be better than average at both statistical categories. However, in Figure 4 there are three points from cluster 2 that seem to be within cluster 1. Table III takes a look at these clusters. Table III displays the calculation of the Euclidean distance between those three points and the two centroids of the clusters. The calculation from Table III supports that each player belongs in cluster 2.

TABLE III. Players that seem to be in Cluster 1 for K-Means clustering

Player Name	**RPG**	**APG**	**Distance from C1 centroid**	**Distance from C2 centroid**
Mason Plumlee	0.3354	0.2264	0.2470	0.3145
Dragan Bender	0.3734	0.1981	0.2874	0.3451
Marquese Chriss	0.3924	0.1792	0.3091	0.3664

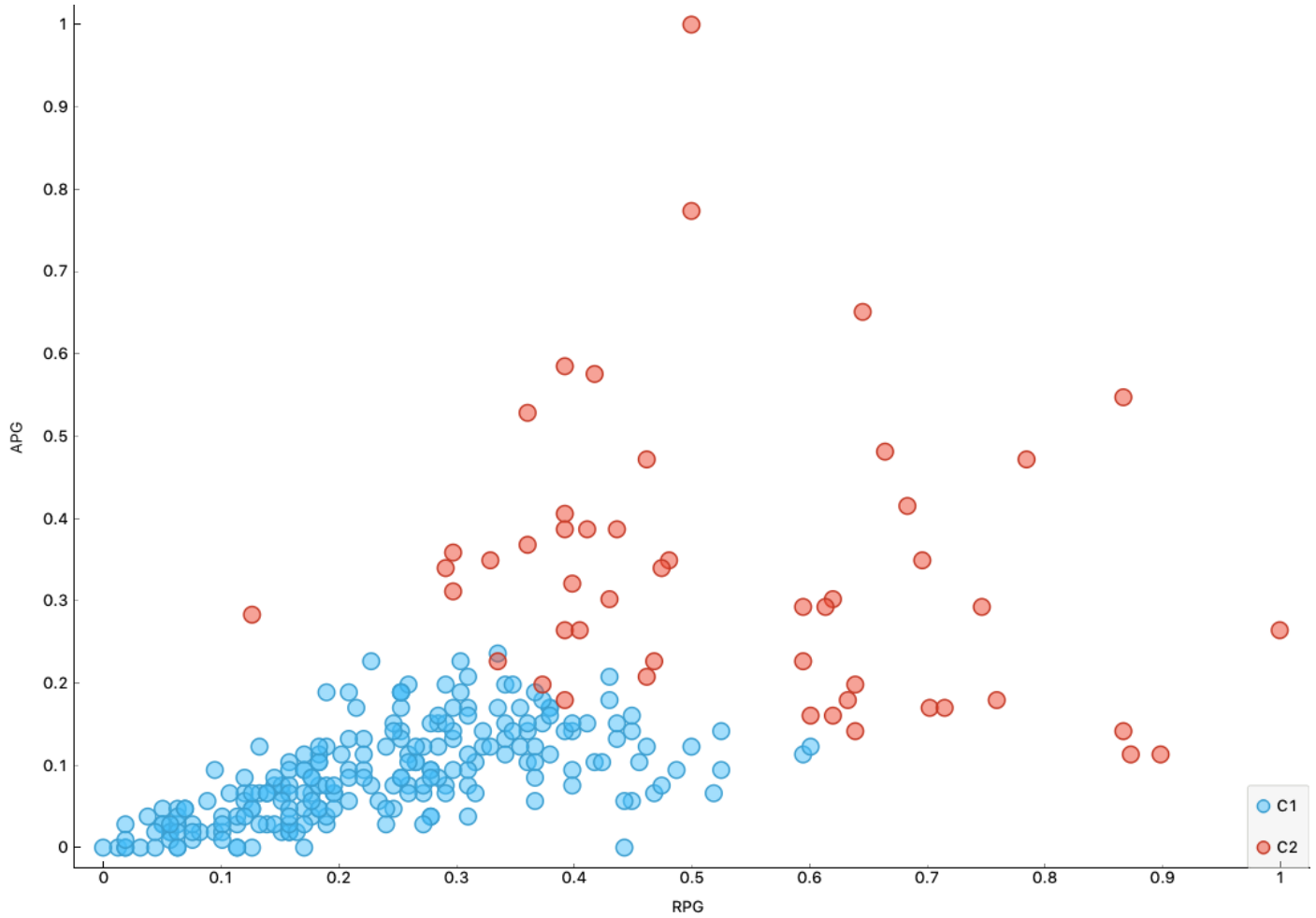

Figure 4. K-means cluster of players

TABLE IV. CLUSTER, LABEL, AND CENTROID OF K-MEANS CLUSTERS

Cluster	Label	Centroid
C1	Tier 2	(0.0886, 0.2364)
C2	Tier 1	(0.3340, 0.5409)

From Figure 5, I felt that I agreed with this clustering of the players using the hierarchical clustering method more than the clustering from the k-means clusters in Figure 4. Cluster 1 represents the elite passers and rebounders. Cluster 1 represents players who are exceptional at rebounding and passing. These players perform much better in these categories than the players in cluster 2. Cluster 2 represents the rest of the players who do not stand out like the players in cluster 1. The players in cluster 2 are very good in one category. These players are also greater than or equal to the average in both passing and rebounding. This cluster also represents players who are not very good in ether statistical category.

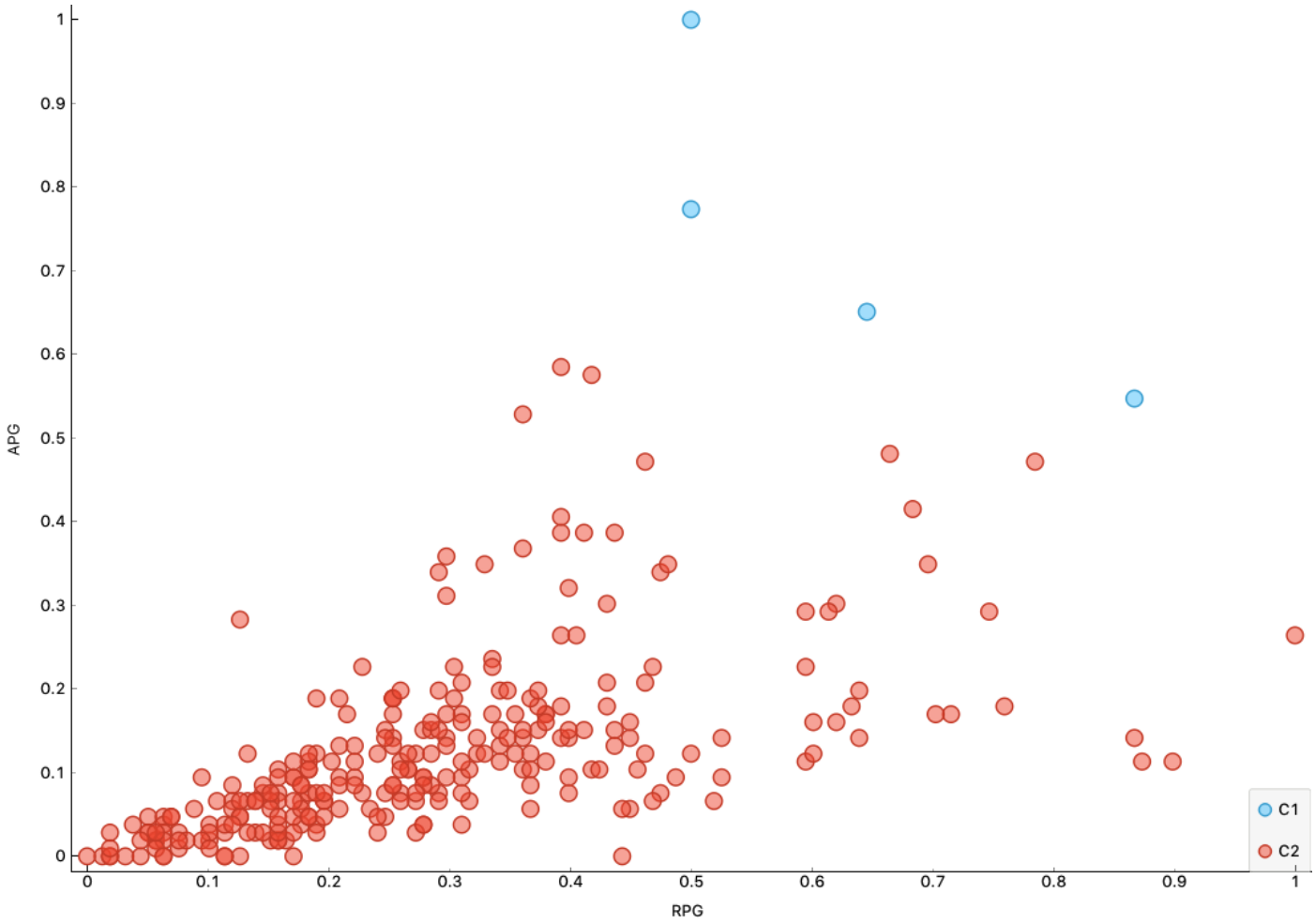

Figure 5. Hierarchical cluster of players

TABLE V. CLUSTER, LABEL, AND CENTROID OF HIERARCHICAL CLUSTERS

Cluster	Label	Centroid
C1	Elite Passers and Rebounders	(0.742925, 0.628165)
C2	Not Elite at Both Passing and Rebounding	(0.124208, 0.286819)

In basketball, two positions that would be considered a guard are a shooting guard and a point guard. A shooting guard is defined as the team's best shooter and is a good dribbler. A point guard is defined as the player that runs the offense and usually is the team's best dribbler and passer. These players are assumed to have the ball in their hands a majority of the time on the floor. There are two positions that would be labeled as a forward. One position is the small forward, who plays against small and large players. Small forwards roam all over on the court. The other forward position is the power forward, who does things such as playing near the basket while rebounding and defending taller players. Last, we have the center, which is the tallest player on each team, playing near the basket [4]. Centers and forwards have different roles than guards. They do not touch the basketball as much as guards do. My analysis takes a look at players who pass the ball more or have an ability to pass the ball. Better passing leads to better team play [5].

V. CONCLUSIONS

Within the report I have identified the rows that were removable with positions that were labeled as "G", "G-F", and "F-G". I removed the unnecessary columns shortening the dataset to each player's id, full name, team, assists per game, rebounds per game, total rebound percentage, and assist percentage. I graphed the frequency distribution and computed the descriptive statistics of the assists and rebounds per game, noting that a majority of these players are better at rebounding than passing. Before clustering the data, I normalized the features so that one feature does not outweigh the others. Then I clustered the players using k-means

clustering and hierarchical clustering. From the calculated silhouette score when using the k-means cluster, I clustered the players into two groups. I chose the same number of clusters when clustering the players with hierarchical clustering. Defining and labeling between the two clustering methods, I felt that hierarchical clustering method did a better job of clustering players in terms of identifying the players who NBA executives or fantasy players should target.

This report has shown that clustering of the forwards and centers can help NBA executives and NBA fantasy competitors target certain players in terms of their productivity for assists and rebounds in a basketball game. Focusing on the hierarchical clusters, there are four players that stood out above the rest. I would be targeting the four players that were in cluster 1 of the hierarchical clustering in Figure 5. This clustering of the forwards and centers helps show that there are different groups of players that vary within those statistical categories. This helps show that there are different players to target depending on the criteria that an executive or a competitor was looking for. The clustering of players with specific statistical categories can be used for all positions and for different categories depending on what is needed and desired for the NBA team or fantasy team.

REFERENCES

[1] https://www.nbastuffer.com/2019-2020-nba-player-stats/
[2] https://www.nbastuffer.com/analytics101/rebound-percentage/
[3] https://www.nbastuffer.com/analytics101/assist-percentage/
[4] https://jr.nba.com/basketball-positions/
[5] https://nba.nbcsports.com/2018/04/07/frustrated-wizards-coach-scott-brooks-were-a-selfish-basketball-team-right-now/