Springboard DSC Program
Capstone Project One

Find the ideal player for a team

A study of changing style in NBA games and
a new approach of classifying player functions

Final Report

by Yiding Weng

February, 2019

Introduction

Every off-season the management of each NBA team has to find a way to improve its roster, usually through recruiting new players on the free market, or signing new contracts with their current players. But how to decide which new player to sign or which player on the old roster to keep? Will that player be able to provide the desirable contribution for the team? As each team tries to optimize the diversity and the integrity of the skill sets of the entire roster, there are problems that must be considered.

NBA teams are increasingly trotting out lineups with five players who can play and guard nearly any position. traditional positions don't accurately explain what a player's skillset truly is, they incorrectly oversimplify the skill sets of NBA players. Simply plugging players into one of five positions does not accurately define a player's specific skill set. Moreover, the misclassification of a player's position may lead teams to waste resources on developing draft picks that do not fit their systems.

Considering these changes, we need an effective way to designate positions in the NBA not based on basic physical traits such as height and weight, but in terms of function, such as shooting and defense. A framework for modern NBA positions is important towards our understanding for how players have evolved, and effective roster construction.

In this project I observed and verified the changing playing style in NBA games and players through data exploratory analysis, found a way to segment NBA players according to their performance statistics using clustering, and finally discussed the player profiles found through clustering.

Data

Proposed Data Collection:

aggregate NBA player's Statistics for 67 NBA seasons.

Data Location: https://www.kaggle.com/drgilermo/nba-players-stats/data

Historic NBA team records

Data Location: https://www.kaggle.com/druswick/nba-team-records-historic

Approach

The approach to the problem will be considered to have three parts:

1. Observe and verify the changing playing style in NBA games and players
2. Find a way to segment NBA players according to their performance statistics, using clustering
3. Discuss the profiles found through clustering

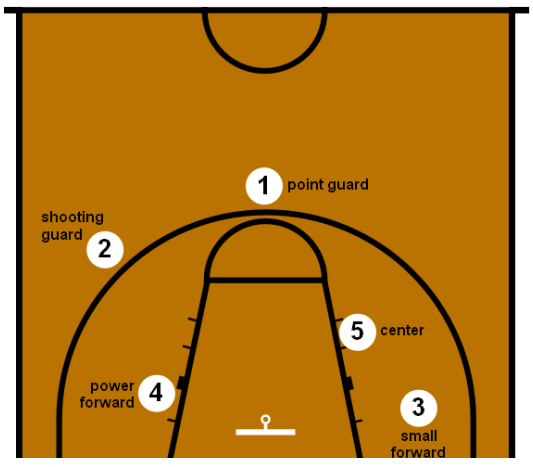
The first data collection includes the performance statistics of each individual player in each season. Using clustering methods to find out systematic ways of defining categories that specifies function or characteristics of players in the same category, compare that with the traditional way of defining player roles as PG, SG, SF, PF and C (see the definitions of these acronyms below).

For the second part understand whether each team has a balanced number of players in each new defining roles, and evaluate if the completion level is correlated to team regular season performance using the historic team records.

Data wrangling

Data cleaning

In dealing with player positions,
1. Point Guard (PG), typically the team's best ball handler and passer.
2. Shooting Guard (SG), usually able to shoot the ball from three-point range, they also tend to be the best defender on the team, as well as being able to move without the ball to create open looks:



Point Guard (PG), typically the team's best ball handler and passer.
3. Small forwards (SF), resembles a shooting guard more often than a power forward, which means they have quickness and strength at the same time.
4. Power forward (PF), similar role to C, able to score close to the basket while also being able to shoot mid-range jump shots.
5. Center (C), usually the tallest player of the team who plays near the baseline or close to the basket.

PG, SG, PG, PF, C are the 5 designated positions for professional players, but a very small portion of players played a role somewhere in-between two different positions, so that they were labeled with 2 positions such as
SF-SG, C-PF, SG-PG. Due to lack of information of which role these players were more inclined to play, I removed those players' information who played double roles.
For the stats that represent total values (others, as TS%, represent percentages), I transformed them into values per 36 minutes. The reason is to judge every player according to his characteristics, not the time they were on the floor.

Deal with missing values

Delete blank columns and rows.
There are significant missing values in 3P (3-Point Field Goals), 3PA (3-Point Field Goal Attempts), 3P% (3-Point Field Goal Percentage), 3Par (3-Point Attempt Rate). It's because 3-point line was not introduced to NBA until 1979, but for the sake of classifying modern players, I must assume that all players before 1979 have not attempted 3-pointers, so I will fill the missing values of these three columns with 0.
For the missing values in GS (Games Started), since number of Games started for each player does not affect their performance measuring, I will delete the GS column.
Missing values also appeared in ORB (Offensive rebounds), DRB (Defensive rebounds), STL(number of steals), BLK (number of blocks), TOV (number of turnovers) and their associated percentage rate(ORB%, DRB%, STL%, BLK%, TOV%), these measurements are not included in record before 1974, and TOV data are not present until 1978, These missing values cannot be filled with assumption, but at the same time these columns are too important to be removed from the construction of player classification, therefore all the data before 1978 have to be discarded.

Handle outliers

When using rate statistics (i.e. points per game) or cumulative statistics (i.e. total points) can be misleading when it comes to analysis because these statistics tend to inflate players with lengthier careers. To deal with outliers, I instituted a minimum

threshold of 40 games played. Also, I kept only players with more than 400 minutes for each season (with a 82 games regular season, that's around 5 minutes per game. Players with less than that will be only anecdotical and will distort the analysis).

Exploratory analysis

Data storytelling

Basketball players at different positions have very different performance in each category. For both management professionals and basketball audience, they both have certain traditional understanding or expectation on player functions. For example, player in the role of point guard (PG) are the ones with smaller statues, they tend to be the ball handlers most of the time, and often has higher number of assists in the game. And bigger players in the role of center (C) or power forward (Pf) tend to grab more rebounds and blocks since the rim is more accessible for them.

During data storytelling I examined how these common understandings of basketball player function stand firm in the light of real data, and we may find trend of changes in data attribute in relation to player position and time, which help us to see whether the function of each position has evolved into more specialized and distinctive in its own style: Does PG have more assist? Does center grab even more rebounds? Or whether the functions of different position have become more intermingled and interchange-able? Therefore, I divided the cleansed data according to position and according to time. The row data attributes I will look at are:

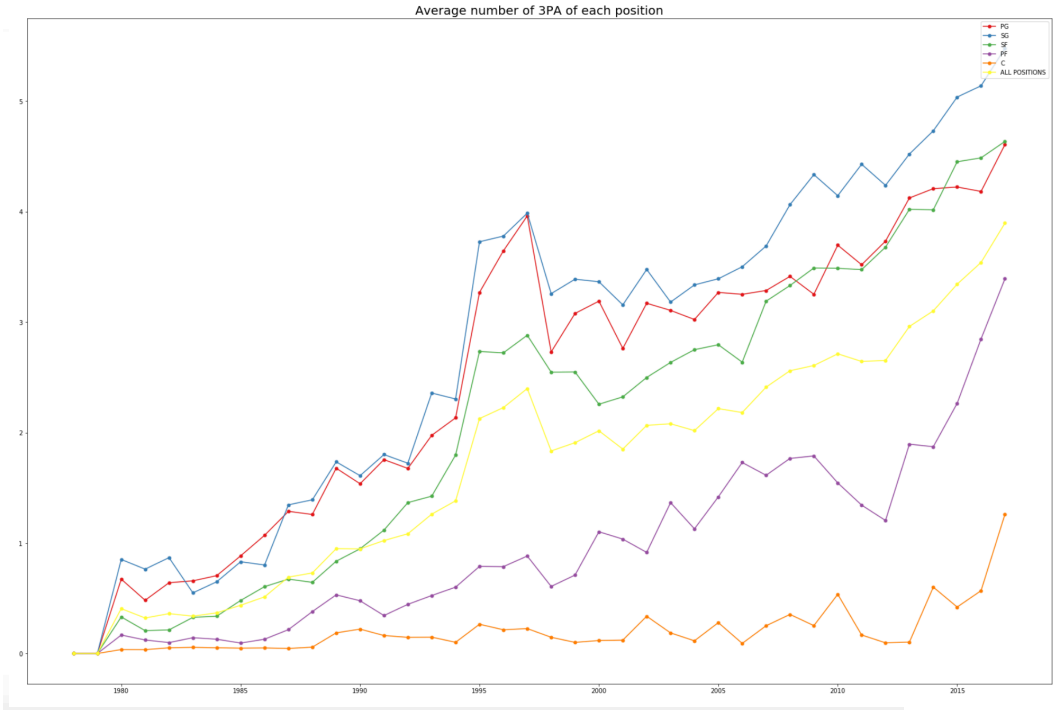
PTS - Number of points
3PA - number of 3-Pointer attempt
2PA - number of 2-Pointer attempt
AST - number of assists
TRB - number of total rebound
PF - number of personal fouls
FT - number of free throws
BLK - number of blocks

After visualizing the data with scatter plots of individual player information, for example:

3PA (number of 3-Pointer attempt) of individual player in each position from 1978-2017

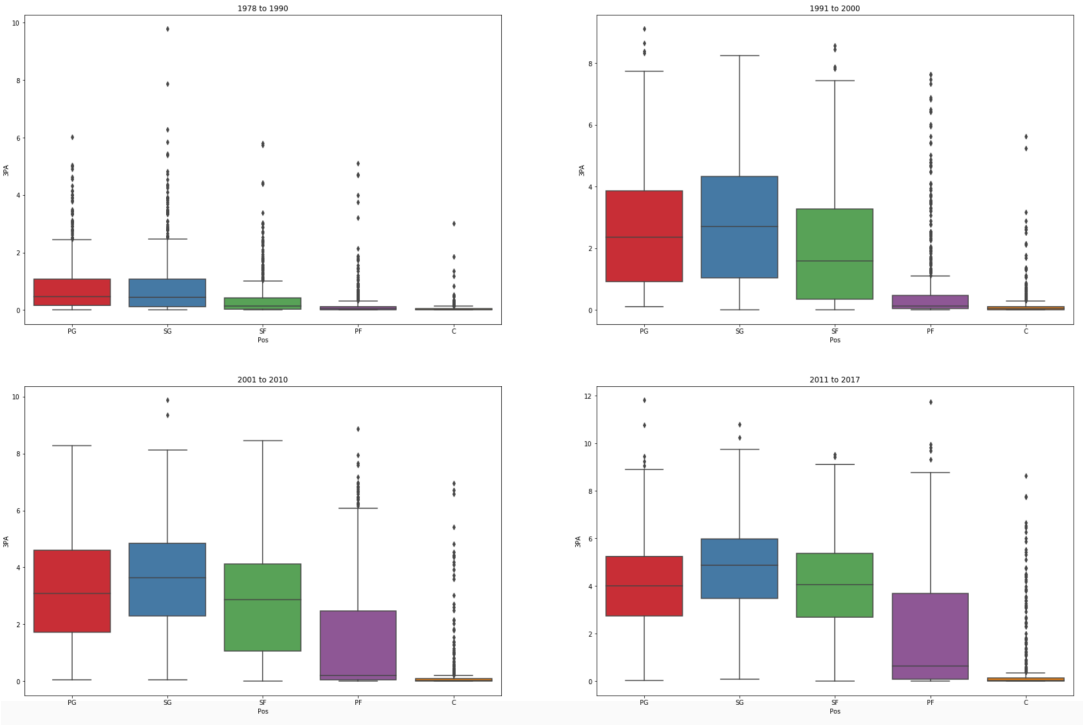
the line plots show average number for each position category for multiple attributes.

average number of 3-pointer attempts per game of each players in different positions



and box plot that presents player performance number distribution during different time period.

number of 3-pointer attempts for player distribution in each time period



we can draw a quick summary of the changing trend in each data category:

PTS - Center (C) and point guard (PG) took small forward (SF) and shooting guard (SG)'s places in leading score positions;
3PA - Overall increase, especially in point guard (PG), shooting guard (SG) and small forward (SF);
2PA - Overall decrease;
AST - Overall stable, point guard (PG) becomes less dominant in leading the number;
TRB - Overall stable, Center (C) becomes more dominant;
PF - Overal decrease;
FT - Overall decrease in FT, Center (C) and point guard (PG) took small forward (SF) and shooting guard (SG)'s places in leading number of free throws;
BLK - Overall stable, Center (C) becomes more dominant;

Therefore, according to certain attributes such as 3PA, TRB and BLK, we can say some functions for each positions have become more distinctive, where on the other end, attribute such as PTS, AST, and FT, suggest some function for certain positions has become less distinctive and can be interchangeable with other positions.

Statistical analysis

In this section we are going to use statistical analysis to explore if the observations from data storytelling can be validated.

Therefore, I used scoring points as a case study one of the previous observation. We have already noticed the change of role in leading the scoring data among the positions. Before 1990, players of SG had the highest number in scoring point, and according to the visualization we think SG players have been scoring less. I choose to use the data before 1990s and after 2010s since we want to compare the scoring data of SG from early and recent periods.

Preform hypothesis test

H0: $p = 0$ H1: $p \neq 0$
Null hypothesis:
There is no difference between the scoring points of SG before 1990 and after 2010 (average scoring points difference is 0, $p = 0$)
Alternate hypothesis:

There is difference between the scoring points before 1990 and after 2010 (average scoring points difference is not 0, $\rho \neq 0$)

We get a p-value of 0.0, which suggests that there is a statistically significant difference. We got a difference of 2.314 points between the means. You should combine this with the statistical significance. Changing by 2.314 points in 40 years is substantial.

This case study with one of the data attributes shows that all the validation is verifiable if we preform hypothesis test to each observation we made in multiple attributes.

Cluster modeling

Select and modify data for modeling

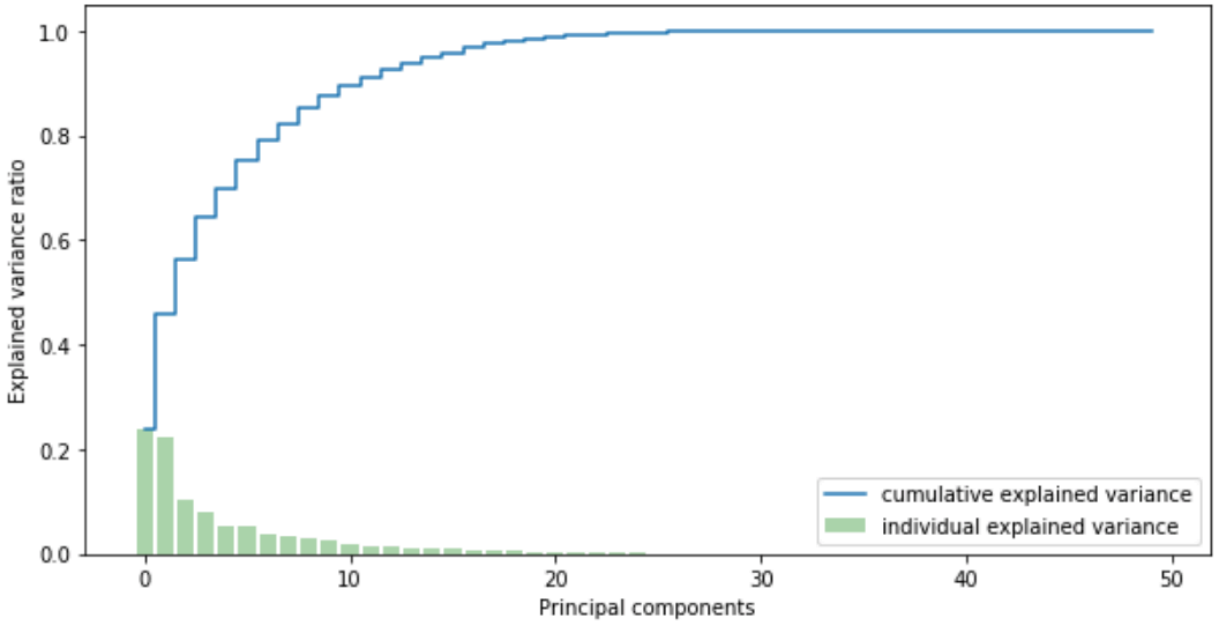
In this step, I used unsupervised learning methods to cluster the players in different time eras into categories that can better summarize their profiles. In order to observe the most obvious change in limited time, I chose two subsets have the largest time gap in between, which is pre-90 (players data before 1990) and 2000s (players data since 2010), and left aside the data in between.

Before sending the dataset into the clustering model, they must be modified through normalization, with the goal of changing the values of numeric columns in the dataset to a common scale, without distorting differences in the range of values. Since the column features of our dataset have different ranges, such as the range of scoring points is from 0 to 30+, and for height and weight is even larger, but for other features like number of blocks or steal, the range tends to be much smaller. So, for the sake of preventing the bias of the predicative model towards feature with higher number, we have to normalize the dataset.

Feature analysis with PCA

PCA stands for Principal Component Analysis, it is a method used to reduce number of variables in datasets by extracting important ones from a large pool. It reduces the dimension of data with the aim of retaining as much information as possible. In other words, this method combines highly correlated variables together to form a smaller number of an artificial set of variables which is called “principal components” that account for most variance in the data. PCA uses "orthogonal linear transformation" to project the features of a data set onto a new coordinate system where the feature explains the most variance is positioned at the first coordinate (thus becoming the first principal component).

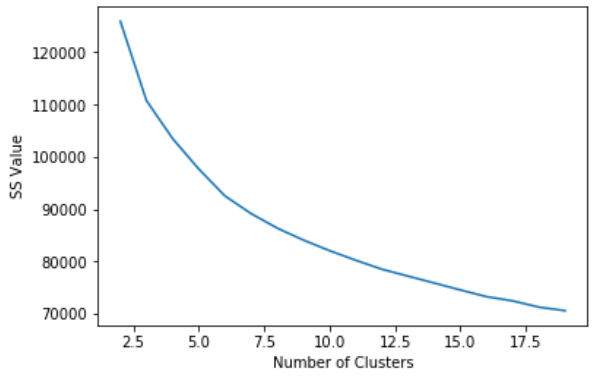
I plot out all the features in the order of the decreasing variance ratio (largest first) to find the features that explains the most variance.

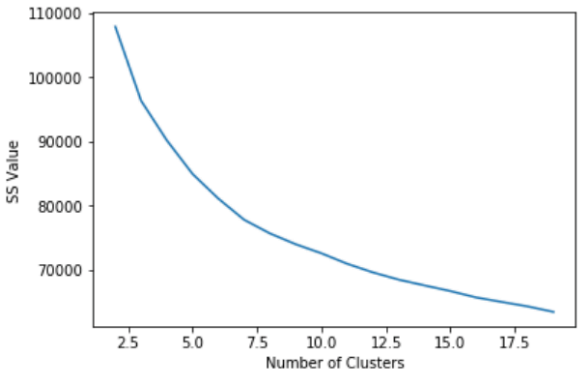


For datasets in both time periods, the first 6 features explained more than 75% of the variance. So I decided to move on with 6 principal components for clustering model.

Find the optimal number for clusters

With the two reduced dimension datasets, I explored the different choices of number for clustering, with the options from 2 to 20 clusters.





By using the elbow method, in both cases we can tell starting from 7, the SS value stops

decreasing as rapidly as before, so I will choose 7 as the number of clusters for both datasets.

Cluster analysis

By now all the data in two datasets are divided into 7 clusters, I observed the features of each of the cluster from normalized data in three different dimensions:

1. Physical feature assessment
2. Function assessment
3. Scoring assessment

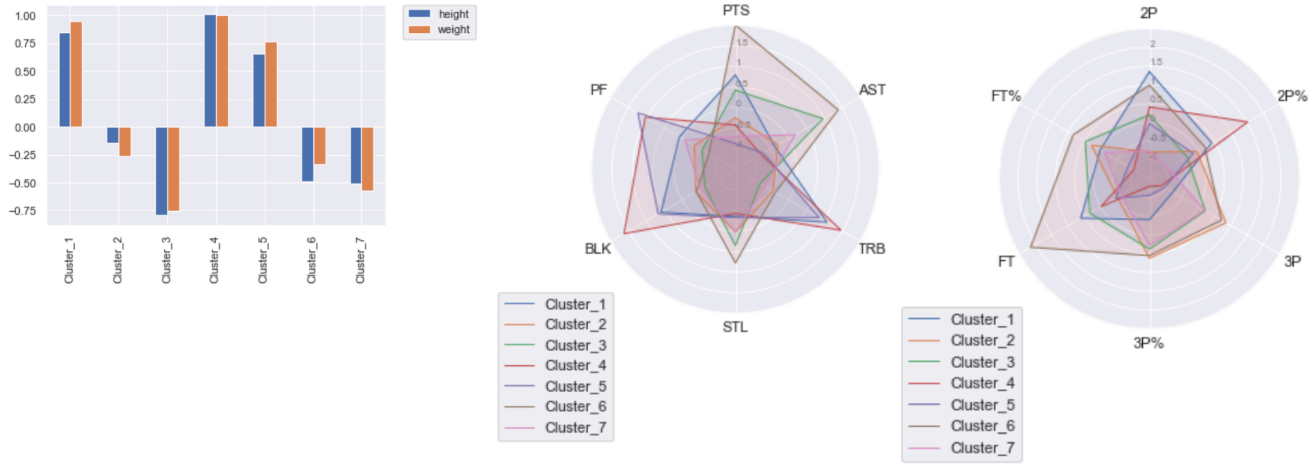
These three assessments, especially the 2nd and 3rd one are done through normalized data since we can equally compare the features at the same scale level.

And I also added the cluster label into the original (unnormalized) dataset, just to get additional insights from the raw data, such as PM (player minutes) and USG% (usage rate, stands for percentage of team plays taken by the player while on the floor.)

For pre-1900 data

Overall assessment of each cluster

Cluster_1: medium players particularly strong at mid-range and drawing free throws, take longest playing minutes, and most team plays
Cluster_2: small players who don't shoot much, have limited playing minutes
Cluster_3: big players who shoot very few, particularly strong in block and rebound, getting fouls, have limited playing time
Cluster_4: small players who particularly strong at distant-range, relatively strong in assist and steal
Cluster_5: big players who's relatively strong at mid-range and drawing free throws, particularly strong in getting rebounds and blocks, play long minutes
Cluster_6: big players have well rounded shooting options, average in all areas
Cluster_7: small players have well rounded shooting options, particularly strong in assist and steal



For post-2010 data

Overall assessment of each cluster

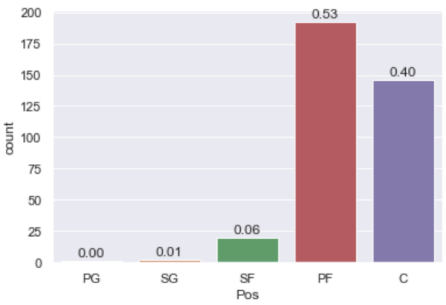
Cluster_1: big players who shoot mid-range and free throw often, well-rounded in all areas
Cluster_2: medium players who focus more on three points shooting, average in all areas
Cluster_3: small players have well rounded shooting options, relatively strong in assisting and stealing
Cluster_4: big players have good under basket shoots, bad at free throws, exceptionally strong in rebounding and blocking
Cluster_5: big players who don't shoot much, but particularly strong in rebounding and blocking, and also getting fouls
Cluster_6: small players who's exceptionally strong in all kinds of shooting, assisting and stealing, have highest usage
Cluster_7: small players focus more on three points shooting, average in all areas, less game time compare to other groups

Assign title to each cluster

The last step of cluster analysis is to look at the prominent players (the ones have the longest playing minutes) in each cluster to get a real feel of what are the players like, and see the distribution of traditional player position in the cluster.

Here is an example for the one of the clusters in post 2010 data

```
array(['Marc Gasol', 'Pau Gasol', 'Blake Griffin', 'Kevin Love',
       'Al Horford', 'Lamar Odom', 'Zach Randolph', 'Paul Millsap',
       'Tristan Thompson', 'LaMarcus Aldridge', 'Giannis Antetokounmpo',
       'Brook Lopez', 'David Lee', 'Al Jefferson', 'David West',
       'Serge Ibaka', 'Gerald Wallace', "Amar'e Stoudemire", 'Josh Smith',
       'Luis Scola', 'Andre Drummond', 'Nikola Vucevic', 'Elton Brand',
       'Chris Bosh', 'Joakim Noah', 'Greg Monroe', 'Carlos Boozer',
       'Rudy Gay', 'Marcin Gortat', 'Thaddeus Young', 'Anthony Davis',
       'Gorgui Dieng', 'Dirk Nowitzki', 'Andrea Bargnani',
       'Karl-Anthony Towns', 'Myles Turner', 'Luol Deng', 'Chris Kaman'], dtype=object)
```



Then I came up a title for each cluster to conclude the main function of that cluster, with the name of exemplary players and dominant positions as additional information.

For pre 1990 data

1. Middle man that carried the team - SG, SF
Isiah Thomas, George Gervin, Julius Erving, Magic Johnson.

2.Well-rounded small players with limited playing time - PG, SG, SF
Nick Weatherspoon, Jim Paxson, Charlie Scott, Armond Hill

3. Defensive big man who does the phyical work - C, PF
George Johnson, Clifford Ray, Paul Silas, Benoit Benjamin

4. Distant range shooter - PG, SG
Michael Adams, Reggie Miller, Johnny Newman, Jeff Hornacek

5. Elite big man with all traits - PF, C
Larry Bird, Moses Malone, Karl Malone, Hakeem Olajuwon

6. Well-rounded big man - SF, PF and C
Charles Oakley, Sam Perkins, Chuck Person, Alex English

7. Small players with all traits - PG, SG
John Stockton, John Havlicek, Bob Wilkerson, Kevin Porter

For post 2010 data

1. Big man with all traits - PF, C

Marc Gasol, Blake Griffin, Kevin Love, Dirk Nowitzki

2.Three-point shooter, outside defender - SG, SF, PF

Klay Thompson, J.J. Redick, Ray Allen, Kyle Korver.

3. Outside assistant - PG, SG

Ricky Rubio, Rajon Rondo, John Wall, Andre Miller

4. Under basket offender, ring protector - C, PF

Zaza Pachulia, Andrew Bogut, Andre Drummond, Rudy Gobert.

5. In between big men - PF, C

Lamar Odom, Tony Allen, Taj Gibson, Yi Jianlian

6. Elite player who carries the team - PG, SG and SF

LeBron James, Kobe Bryant, James Harden, Kevin Durant

7. Secondary ball handler, 3-point shooter - PG, SG and SF

Marco Belinelli, Corey Brewer, Matthew Dellavedova, Derek Fisher

Summary

There are obvious changes in player style between the two time periods, with at least 20 years of time gap in between them. We have noticed the optimal way to categorize players has become less relevant to the on-court positions, since before 1990 two of the seven clusters are mainly represented by 3 position:

2. Well-rounded small players with limited playing time - by PG, SG, SF
6. Well-rounded big man - by SF, PF, C

After 2010 three of the seven clusters are mainly represented by 3 positions:

2. Three-point shooter, outside defender - by SG, SF, PF
6. Elite player who carries the team - by PG, SG, SF
7. Secondary ball handler, 3-point shooter - by PG, SG, SF

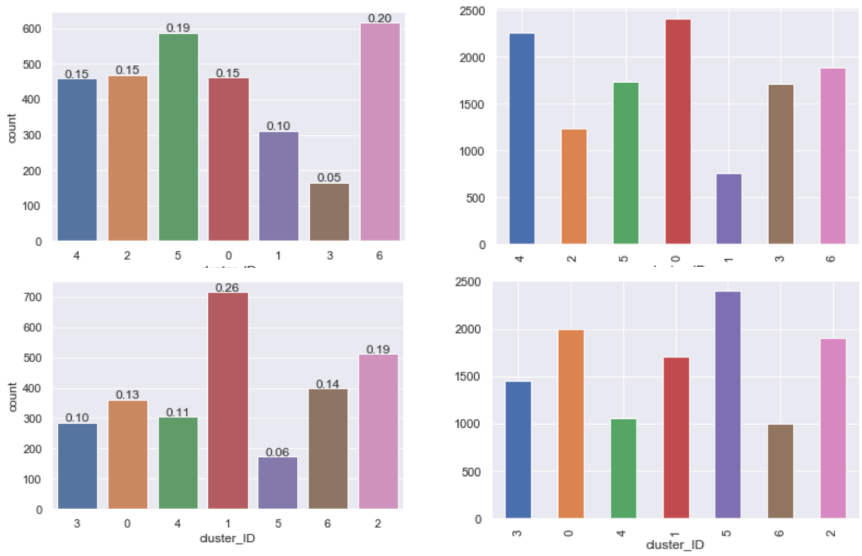
I wanted to see more of changing profiles and distribution of clusters throughout time, So I started by ordering all the clusters according to a fundamental physical feature-average player height, plot the number of players and the average playing minutes in each cluster.

Percentage of players in each cluster among total number

Average playing time of players in each cluster

Pre-1990 data

Post 2010 data



*The cluster_ID is the cluster number I have been using so

We perceive the proportion of big players (the first three clusters) have shrunk significantly, from 49% before 1990 to 34% after 2010. But we did not observe a drop in player's physical height during the data exploration, it means after 2010 more players tended to play 'small ball', which involves the game shifted towards more outside range, rather than near basket area dominated by big players. It's also explains the most populous cluster in post 2010 data is the 3D (three-pointers plus outside defending) player group, which takes 28% of all player number count. While the cluster in pre-1990 that resembles 3D the most is the (NO.3) Distant range shooter cluster, which only takes 5% of players number. Another observation is that the team carrier cluster (Middle man that carried the team) before 1990 takes 15% of all number, but the equivalent one (Elite player who carries the team) after 2010 takes only 6% of all player population.

From the playing minutes we can also observe the on-court minutes of smaller players (last three clusters) have increased significantly from pre1990 to post 2010, especially for the team carrying cluster 5. Elite player who carries the team) for the post 2010 data, though with the fewest members, but have the longest playing time among all groups.

Conclusions

The league players are increasingly more dominated by 'smaller' players, as their number takes the majority of the total population after 2010, it does not mean more players are becoming physically smaller, but means the game is moving away from near-basket towards outside range, which is also explained by the most populous cluster is the 3D (three-pointer plus outside defense) player group after 2010. The average game time for small players have also been getting longer. The increasing blurry of traditional player position within different cluster shows the increased freedom for players to switch between positions, this implies the speed and range of ball movement have increased, rather than simply move towards the big man inside or close to basket. The most elite players have shifted from the middle group to small group. Along with the insights we have found during the data exploration about the overall decreasing number of attempts of 2-pointer shoots and increasing number of 3-pointer-shots, plus fewer numbers of free throws and personal fouls, which seem to indicate that contemporary games are becoming less physical, fast-paced, and covering wider space. Therefore, we can state the NBA basketball playing style is moving towards 'small ball' after 1990s to 2010s.

Future work & Recommendations

Feature selection

I used PCA to reduce the dimension of original data, this method can improve computation speed and reduce the chance of modeling result being interfered by less relevant features. But it will be a better approach if have tried different methods for feature selection such as RFE (recursive feature elimination) and SelectKBest, and use silhouette score to compare the performance of all the clustering based on different feature selection options. More than that, I want to see if the cluster model will produce similar result without cutting away any features.

Cluster modeling for data from 1990 - 2010

In this project I have created clusters for data before 1990 and data after 2010 in order to observe the change of players from the largest time gap. But by doing so I assumed this change is gradual and increases as the time moves on. While it is possible the change is not linear and evolved towards one direction and merged back during certain time period, such phenomenon would not be noticed unless more study is done on these data.

Import Salary Data

A sustainable roster should not only include players of complementing skills sets, but should also maintain at an acceptable salary range. Even if the team manager picks players from different categories, there is certainly tendency to choose players with better statistics, but that usually also implies higher salary expectation. Therefore if the

salary is incorporated into the player information, it can provide extra perspective while comes to player trade consideration.

Team roaster analysis
Another idea to consider would be studying the combination of players from different clusters. Especially by looking at the team roaster in the past that have achieved significant success, we can find good insight about grouping various categories of players. Vice -versa, we can avoid the pitfall of combining wrong groups of players from past lessons.

Recommendation for clients

As I have suggested in the beginning, contemporary NBA players have evolved their functions beyond what their positions normally implies. Signing and trading players during off season involves careful consideration of the player functions. This project has provided general function to profile NBA players. A relatively successful team seeks to cut down their salary cap can consider trading their existing player with new player in the same function cluster with lower salary expectation. And for the team intends to improve their roaster performance should avoid trading player for the one within the similar function cluster, but for player from different ones to change the team chemistry.