NBA Player Type Clustering by Position Reflecting the Modern Basketball Trend

Namtaek Kwon

Finite Mixture Models

SungKyunKwan University

Nov 29, 2022

Outline

- Introduction
- Data Preprocessing
- GMM
- Conclusion

Introduction of Final Project

Traditional basketball has as few as three or as many as five positions.

- Guard (1.Point Guard, 2.Shooting Guard)
- Forward (3.Small Forward, 4.Power Forward)
- Center (5.Center)

Positional restrictions are gradually fading in modern basketball, and an increasing number of players are filling multiple roles.

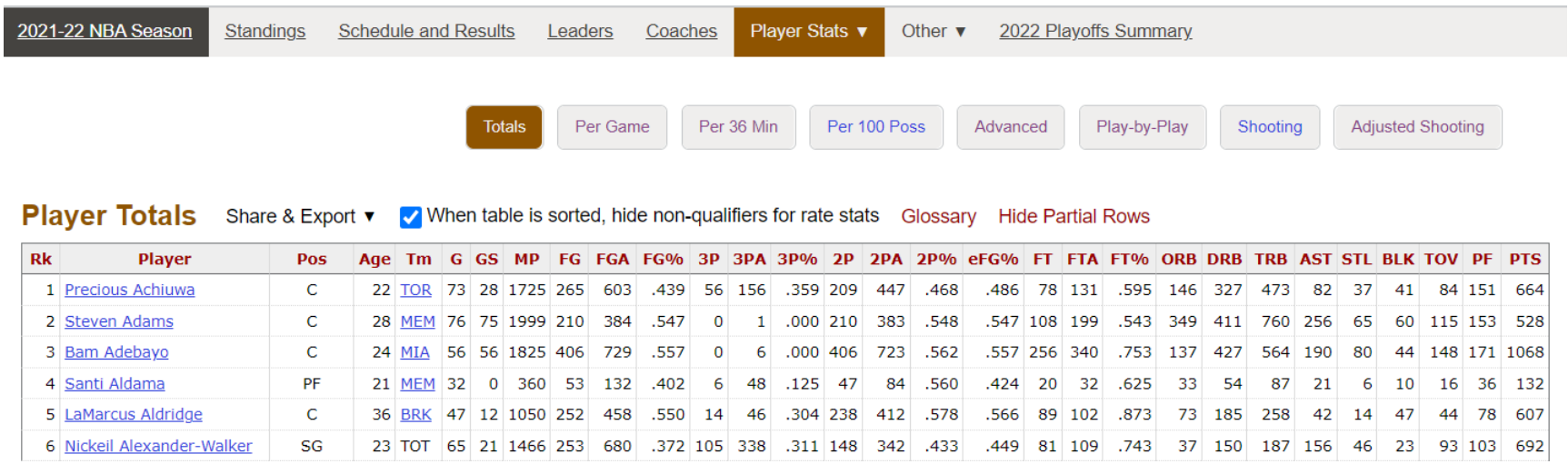
- Dual Guard (1+2), Swing Man (2+3), Stretch Forward (3+4), etc...

We try to subdivide player types for each position using clustering methods, rather than simply limiting to existing roles.

DATA

2021-2022 season NBA player stats

- Basketball-Reference.com
  - Player stats for per game, also advanced indices (ex. VORP).
  - There are so many indices, we should select the variables or reduce the dimensions.



2021-22 NBA Season | Standings | Schedule and Results | Leaders | Coaches | Player Stats ▾ | Other ▾ | 2022 Playoffs Summary

Totals | Per Game | Per 36 Min | Per 100 Poss | Advanced | Play-by-Play | Shooting | Adjusted Shooting

Player Totals Share & Export ▾ ☑ When table is sorted, hide non-qualifiers for rate stats Glossary Hide Partial Rows

Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Precious Achiuwa	C	22	TOR	73	28	1725	265	603	.439	56	156	.359	209	447	.468	.486	78	131	.595	146	327	473	82	37	41	84	151	664
2	Steven Adams	C	28	MEM	76	75	1999	210	384	.547	0	1	.000	210	383	.548	.547	108	199	.543	349	411	760	256	65	60	115	153	528
3	Bam Adebayo	C	24	MIA	56	56	1825	406	729	.557	0	6	.000	406	723	.562	.557	256	340	.753	137	427	564	190	80	44	148	171	1068
4	Santi Aldama	PF	21	MEM	32	0	360	53	132	.402	6	48	.125	47	84	.560	.424	20	32	.625	33	54	87	21	6	10	16	36	132
5	LaMarcus Aldridge	C	36	BRK	47	12	1050	252	458	.550	14	46	.304	238	412	.578	.566	89	102	.873	73	185	258	42	14	47	44	78	607
6	Nickeil Alexander-Walker	SG	23	TOT	65	21	1466	253	680	.372	105	338	.311	148	342	.433	.449	81	109	.743	37	150	187	156	46	23	93	103	692

DATA

Specific indices might be classified as follows.

- Offense : FG%, 3P%, 2P%, eFG%, AST, PTS, TS%, USG%
- Deffence : PF, TRB%, STL%, BLK%, TOV%
- Contributuon : PER, WS%48, BPM, VORP

Data

Naturally, the indicators do not reveal everything about the player.

(a) Stephen Curry -
All time 3-points shooter

(b) Russel Westbrook -
Triple-double machine

DATA

Naturally, the indicators do not reveal everything about the player.

- The indices is not overwhelming in Stephen Curry's case, but the attack unfolds through the space created by Curry leading the defense.
- Russel Westbrook was named MVP of the season and had a triple-double in the 2016-2017 season, but he later changed the method of counting VORP indicators.
- Kawhi Leonard is the league's best swingman and has twice been named Final MVP, but he also takes care of himself during the regular season.

VARIABLE FILTERING

The variables were pre-filtered, with a focus on ”ratio statistics”.

- Because all of the contribution indicators are ratio statistics, uniformity was achieved through filtering.
- Furthermore, the ratio is generally regarded as more significant.
- Of course, using more diverse variables is beneficial, but it was removed because the ratio and cumulative statistics move in similar directions when the dimension is reduced.

NA Imputation

NA was found in the data representing current ratio statistics and was replaced with 0.

- In the case of the center, there have been instances where players with greater influence under the goal.
- For example DeAndre Jordan, two-times rebound leaders, do not throw three points and thus cannot calculate the success rate.
  - It is reasonable to replace it with zero in this case.

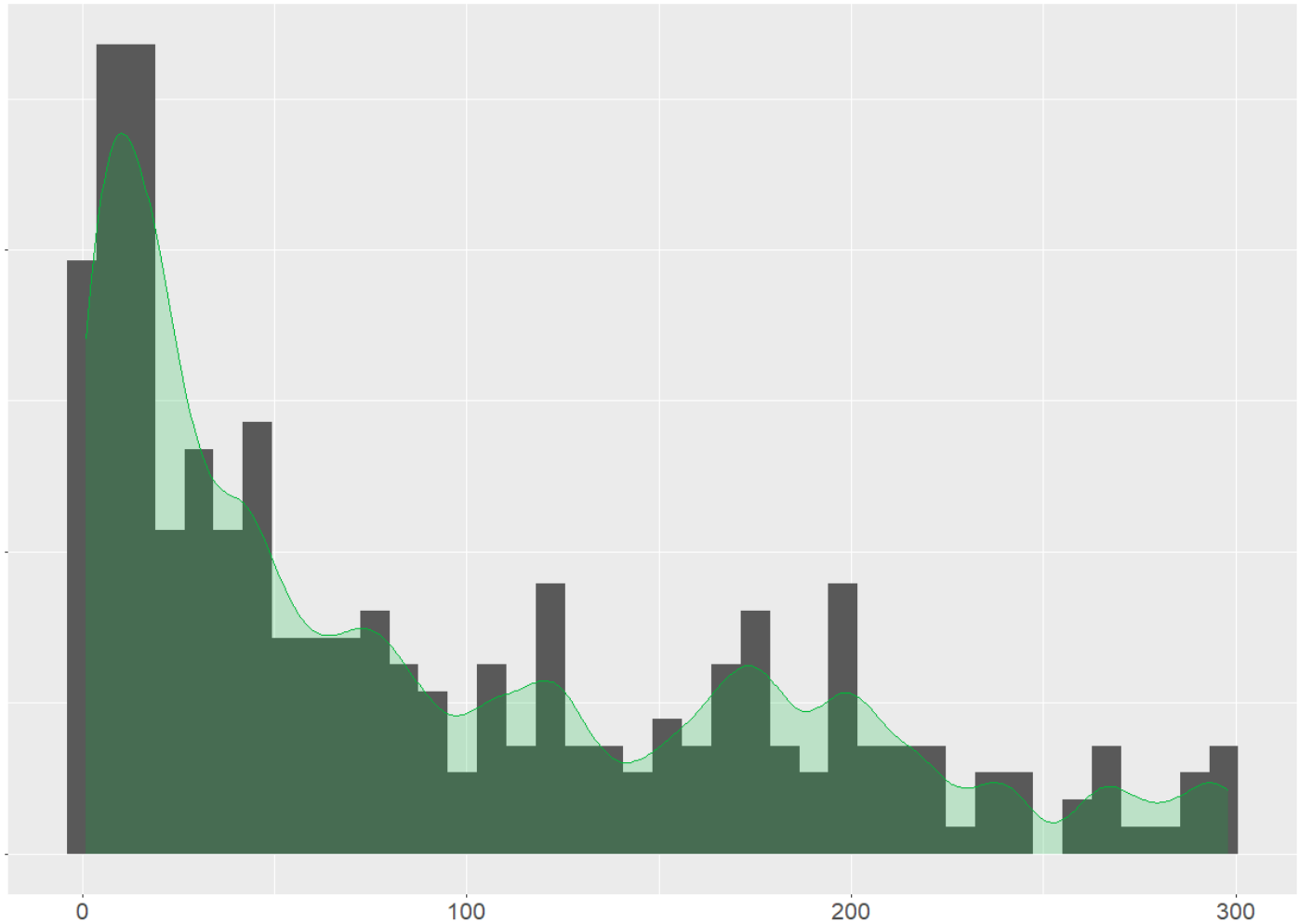
Treat Garbage Members

Furthermore, it was critical to limit the number of samples suitable for analysis.

- Players traded to other teams during the season were added together.
- Because garbage members are unimportant in the analysis, the analysis target was set at 200 minutes of playing time (less than 3 minutes on average in the season).
- Raising the playing time standard reduced the number of observations, while lowering the standard increased the influence of outliers.

Then, $n = 450,\ p = 17$.

Treat Garbage Members

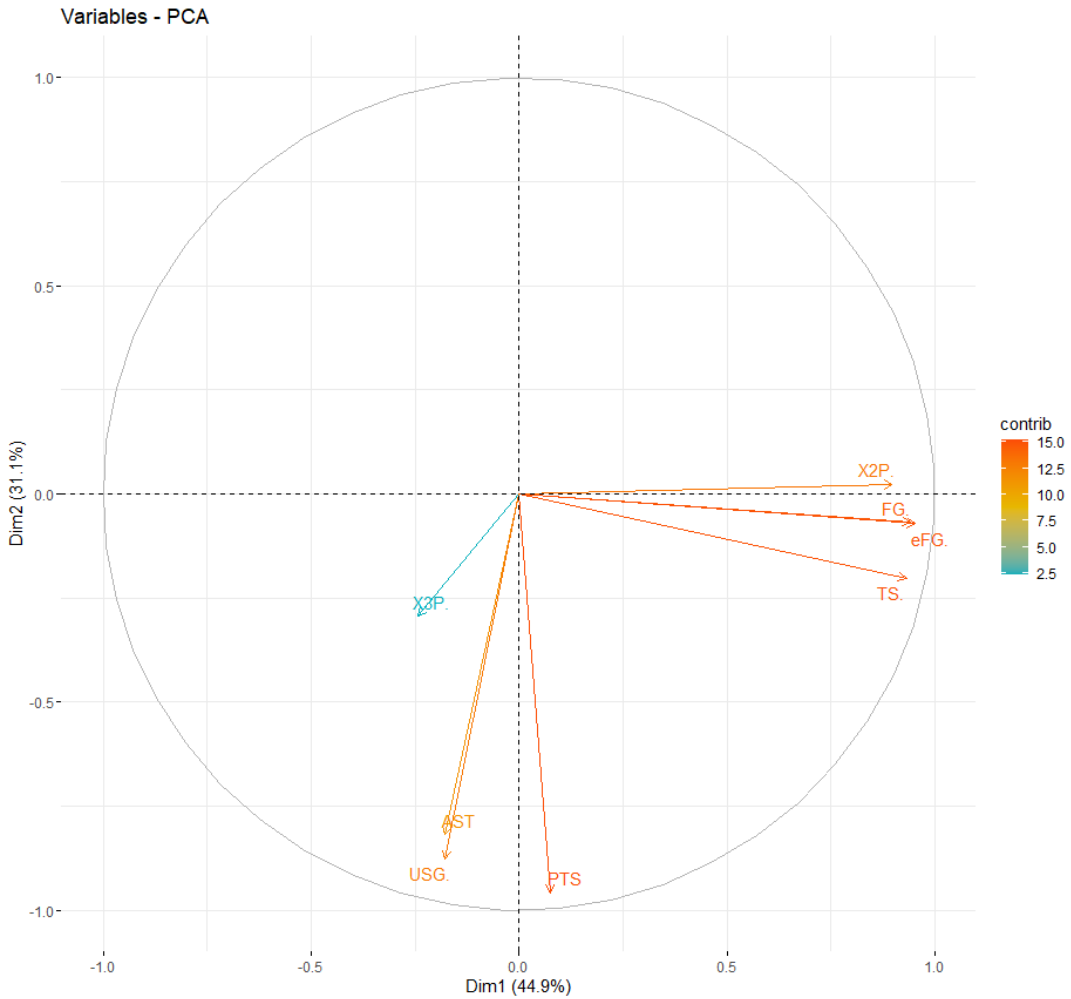


Dimension Reduction

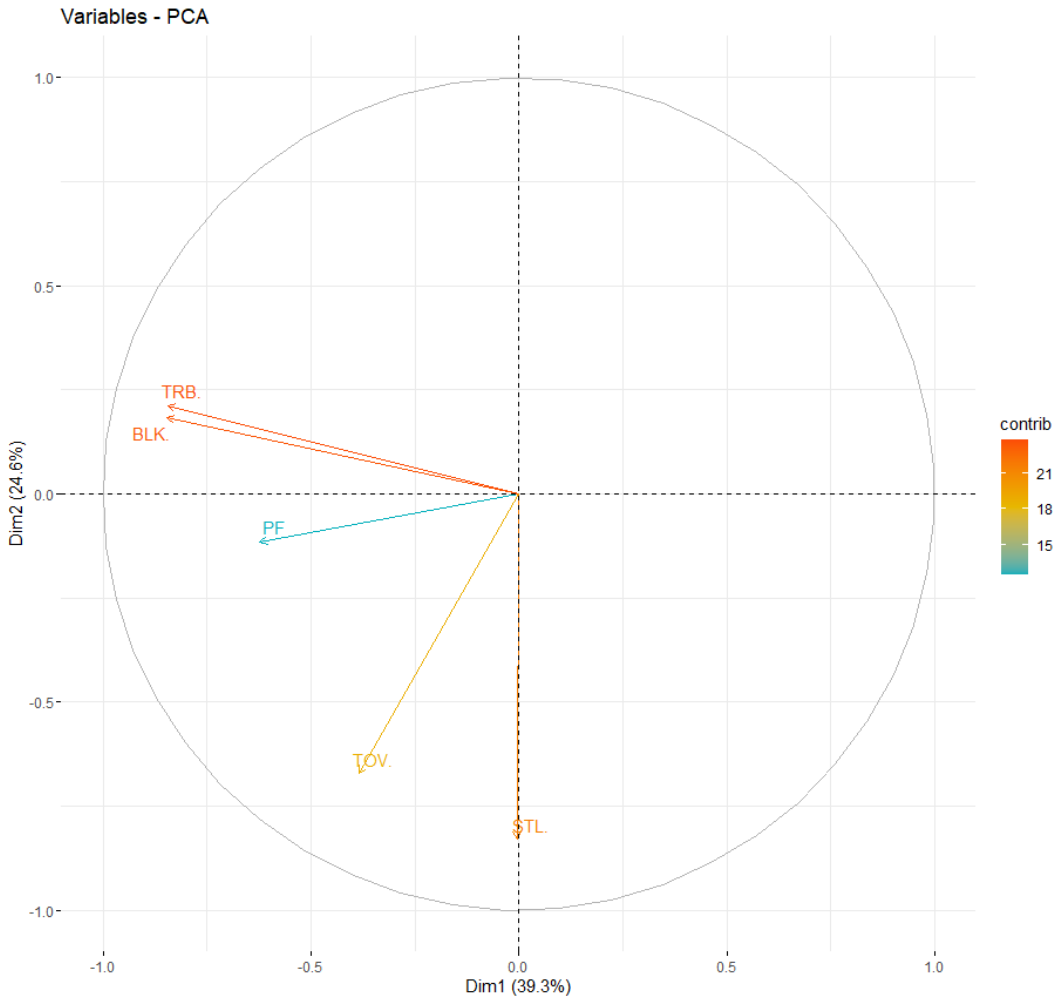
Because "$p = 17$" and "$n = 556$" are currently used, the number of variables must be reduced for proper clustering.

- Using prior knowledge, the indicator's information can be classified as attack, defense, or contribution.
- Dimensionality reduction is accomplished using PCA with two offense, defence indices and one contribution index, taking into account the number of each index.
  - Offense : FG%, 3P%, 2P%, eFG%, AST, PTS, TS%, USG.
  - Defence : PF, TRB%, STL, BLK, TOV.
  - Contribution : PER, WS%48, BPM, VORP.

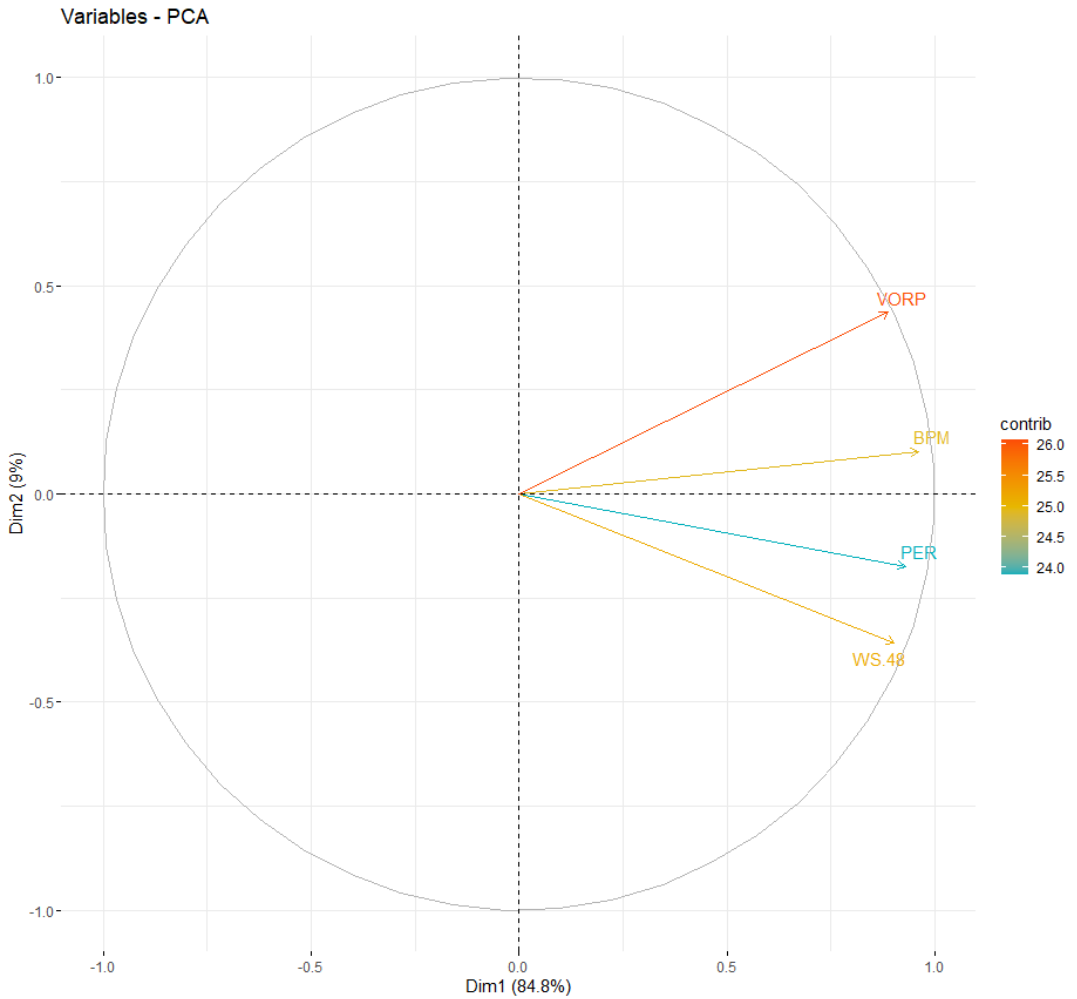
Dimension Reduction



Dimension Reduction



Dimension Reduction



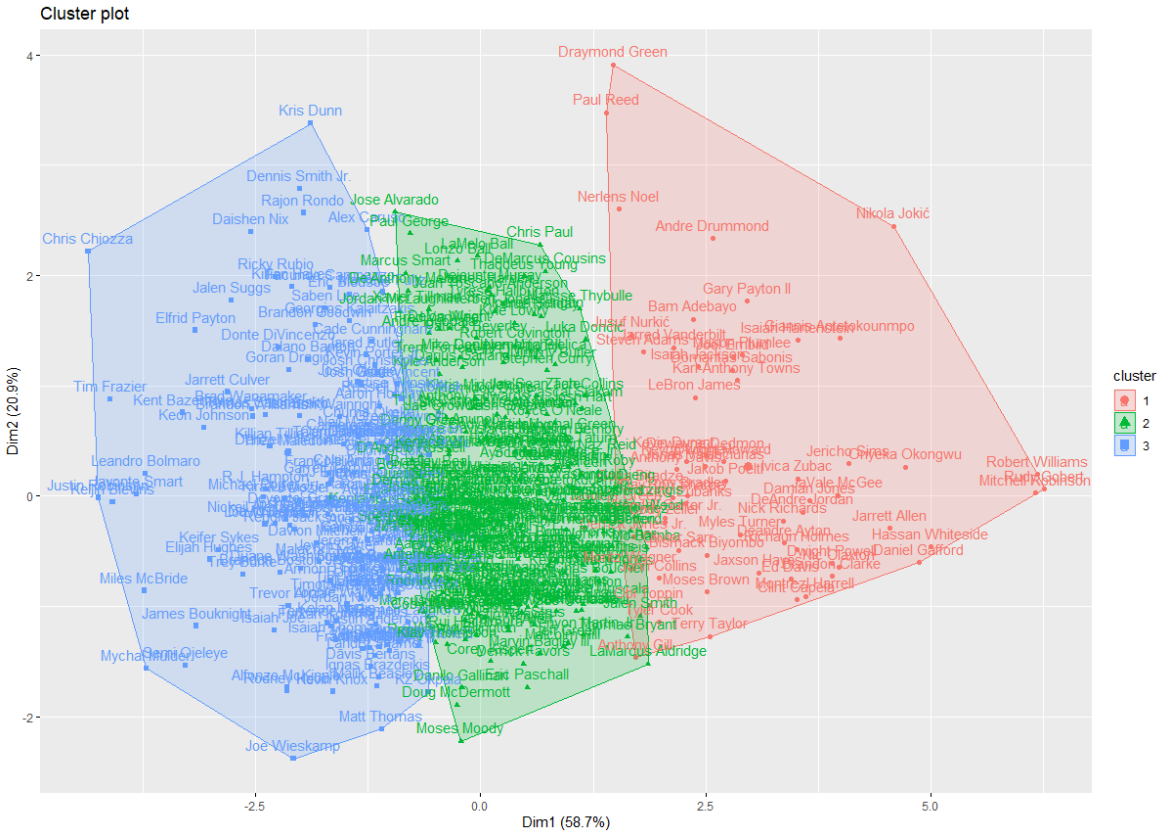
Dimension Reduction

The dimensionally reduced variables have the following meaning.

- ”Offense 1” index represent main ball handlers and scorers.
- ”Offense 2” index indicate the effectiveness of an attack.
- ”Defence 1” index indicates a height advantage.
- ”Defence 2” index represent defensive power while the ball is in play.
- ”Overall contribution” index show how much this player contributes to the victory.

EDA

This is the outcome of clustering all data, regardless of position.



We will proceed with clustering by dividing by position because we cannot get new information.

GMM

The analysis logic is summarized before performing clustering with GMM.

1. The analysis is started with each position for center, forward and guard.
2. Within the covariance structure of the variables used, a model is chosen.
3. The best model is chosen by comparing the silhouette values of the selected models for each variable.

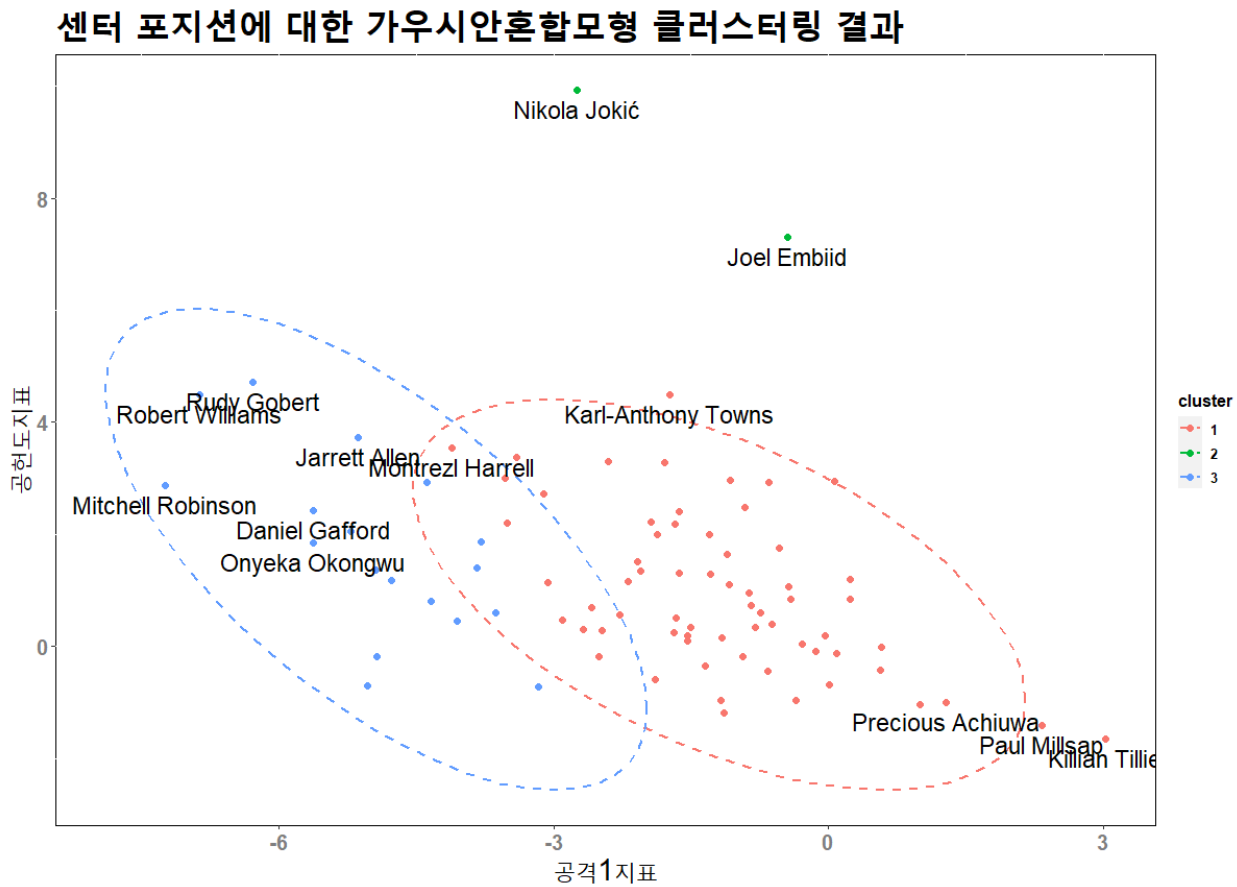
SILHOUETTE VALUES

The silhouette value is a quantitative measure of how well the clustering was done.

- Calculate the cohesiveness of each observation in a cluster and its distance from other cluster observations.
  - This is aggregated by taking the mean of all observations.
- The ”Rand Index” is appropriate in a simulation scenario where the correct answer is known, but it cannot be calculated if the correct answer is unknown.
- If the silhouette value is 0.5 or higher, it is considered successful; if it is 0.25 to 0.5, it is considered meaningful in some degree.

CENTER - GMM

For center, the silhouette value is 0.42 with 3 clusters in EEE covariance structure by using "Offense 1" & "Overall contribution".



CENTER - GMM

Center athletes are classified into three types.

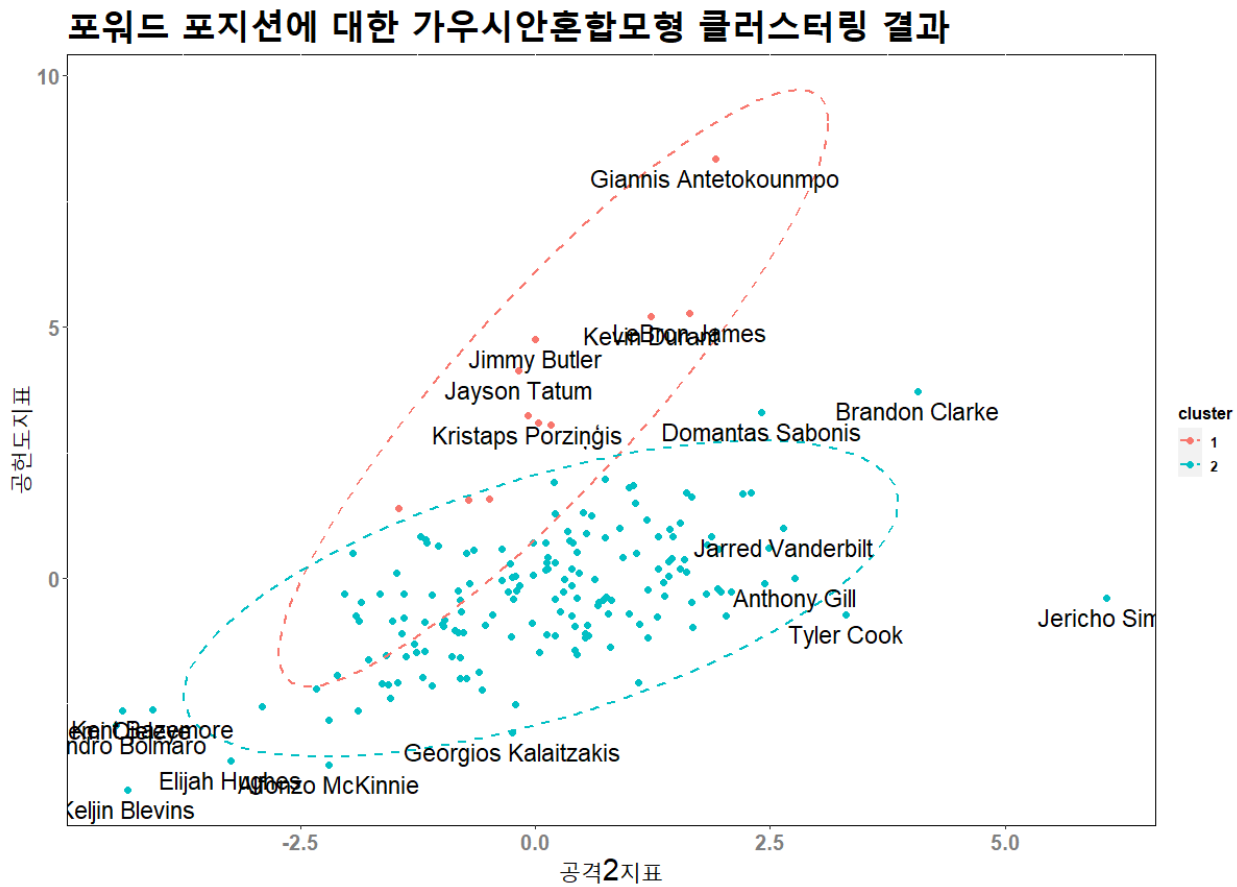
1. The offensive center, such as the Karl-Anthony Towns, is the first type.
   - Even if their contribution to victory is minor, it is a center type who can supplement the team's insufficient offense.

2. The MVP-class center is the second type.
   - They can not only play the center position, but also shoot and coordinate games.

3. A defensive center, such as Rudy Gobert, is the third type.
   - Rudy Gobert is the defender of the year, and players with excellent under-the-goal control, such as Gobert, can be described as such.

FORWARD - GMM

For Forward, the silhouette value is 0.46 with 2 clusters in EVV covariance structure by using "Offense 2" & "Overall contribution".



FORWARD - GMM

Forward athletes are classified into two types.

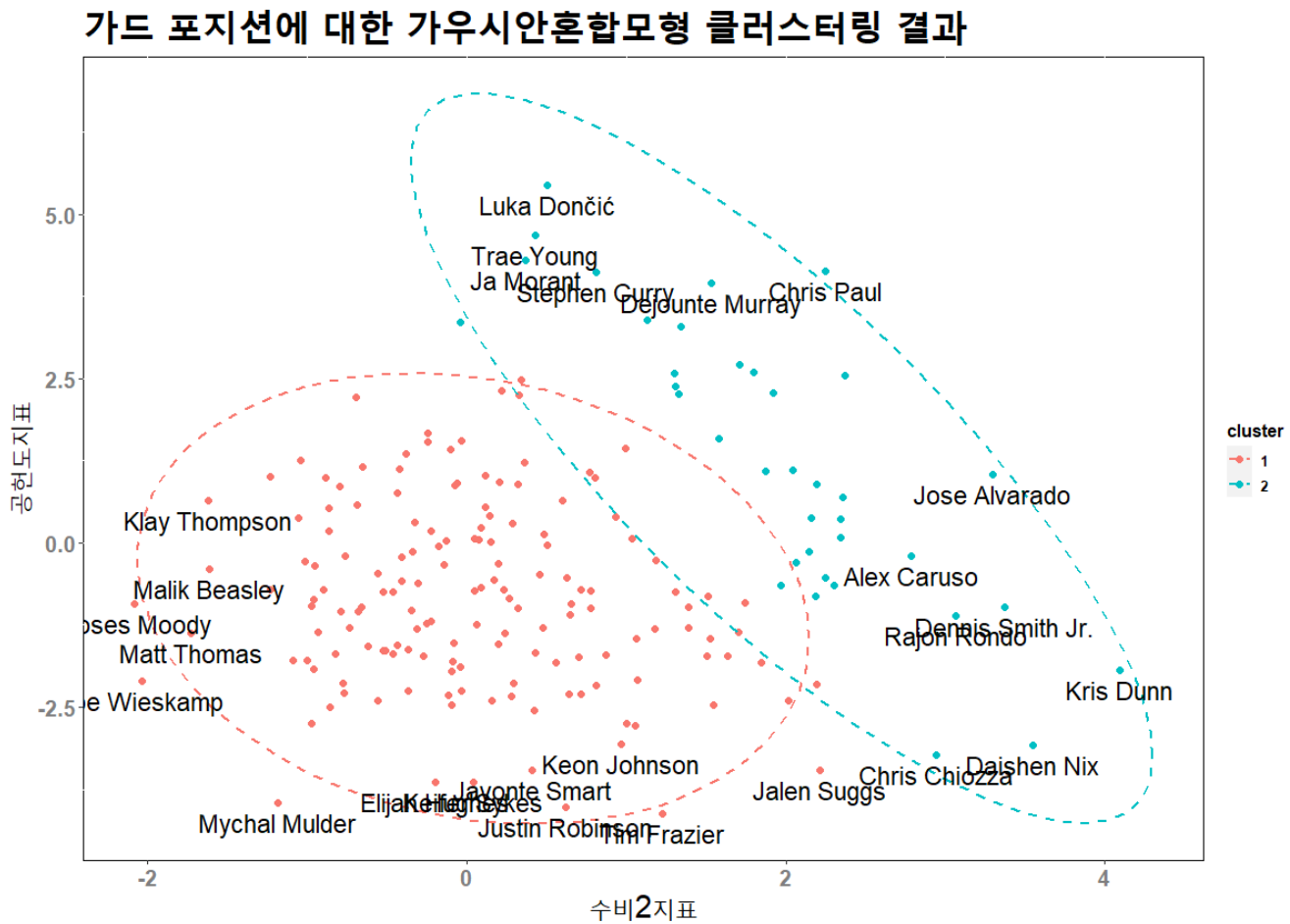
1. The first type is the ”superstar.”
   - These are players who can take high efficiency in attack and lead it to victory.
   - Small forward players such as LeBron James and Kevin Durant have dominated NBA Finals MVP over the last decade.

2. A ”non-superstar” player is the second type.
   - It has more ball possession time than superstar players in comparison.

Overall, superstar players chose the cluster, and these two aspects appeared even when other variables were used.

GUARD - GMM

For Guard, the silhouette value is 0.36 with 2 clusters in EVE covariance structure by using "Defence 1,2" & "Overall contribution".



GUARD - GMM

Guard athletes are classified into two types.

1. The first type of player is the ”Catch and Shooter.”
   - He’s the type of player who, like Klay Thompson, doesn’t take many possession but can play shooting guard for the team.
2. The second type is more difficult to understand.
   - To begin with, he is an ace guard, like Luka Doncic, Chris paul, and Stephen Curry, who lead the overall attack as ”main ball handlers.”
   - Then there’s Alex Caruso, a role player who contributes to the team through excellent defense rather than scoring.

In the case of guards, the analysis process reveals that the clustering results were not well produced.

CONCLUSION

I'll discuss the limitations and improvements I noticed during the analysis.

- First, it was difficult to effectively use various variables.
  - It was more natural to interpret the "PC" variables based on prior knowledge of the players rather than proceeding with the interpretation while looking at the reduced variables.
- Second, rather than the actual player types, some players influenced the outcome of clustering.
  - This phenomenon occurred because "Hero Ball" is possible depending on the superstar's personal competence in the case of basketball.
  - Because more clusters are desired, the Bayesian nonparametric method "DPMM" may be more appropriate in this case.