NBA-Cluster-Caleb

Caleb Skinner

2023-06-09

Contents

Note: This analysis is founded on Tony Munoz's project "KMeansProject/NBAKMeans.Rmd" and Dr. Sturdivant's Outline "NBA ClusterModule.Rmd."

Overview

Cluster analysis is a statistical analysis tool that partitions observations into sub-populations of similar characteristics within the data set. This process can be useful, because similar observations often behave and respond to stimuli in similar ways. Identifying clusters can allow researchers to predict and draw conclusions on the behavior of certain groups. There are many popular topics that use cluster analysis: risk analysis, marketing, real estate, insurance, medical research, and earthquakes.

In this module, we'll use the clustering of NBA players as an example. Suppose you were an NBA General Manager interested in constructing a high-quality team. The best teams use lots of different kinds of players to achieve their goals. Golden State Warriors Guard Stephen Curry is an incredible shooter and ball-handler, but the Warriors need other kinds of players, too. A team comprised completely of Stephen Curry and his clones would struggle to defend or rebound the ball. The team would also struggle to give each Stephen Curry the playing time and shots that he has come to expect. Instead, General Managers can separate potential players into groups, because it helps them to identify their team needs. This is where cluster analysis proves useful.

For this exercise, imagine that you are the General Manager of the Dallas Mavericks. You are tasked with creating a strong, balanced team. Later in the module, you will have an opportunity to create hypothetical trade scenarios that could benefit the team.

Data

Our data for this exercise comes from the 2021-2022 NBA Season. This season, the Mavericks finished 4th in the Western Conference with 52 wins and 30 losses under coach Jason Kidd. They exceeded expectations and made the Western Conference Finals.

Our data includes 374 players. Each of these 374 players fulfilled our requirements of appearing in at least 25 games and playing an average of at least 12 minutes (a complete game is 48) in those games. Because of midseason trades or acquisitions, some of the players will appear in our data twice. That's because they fulfilled our playing time requirements for two different teams in the same season. The second iteration of the player will be marked with a 1 following his name (i.e. Smith becomes Smith1). We've divided the variables into two data sets.

The first set of variables are focused on determining the influence a player has on the game. Some of these variables are the players' minutes per game, total games played and started, points and rebounds per game, and field goal attempts per game. This will be helpful in clustering the players into groups of stars, average starters, and reserves. We've termed this data set "usage". Below is a data dictionary for the first set of variables.

Variable	Explanation	Example
Name	nba player's first and last name	Trae Young or Trae Young1
POS	playing position	PG (point guard), SG (shooting guard), SF (small forward), PF (power forward), C (center)
Team	abbreviation of city of player's team	atl (Atlanta), bos (Boston), etc.
GP	total games played	46, 70, etc.
GS	total games started	7, 56, etc.
MIN	minutes per game	18.2, 30.2, etc.
PTS	points per game	6.8, 14.9, etc.
AST	assists per game	1.1, 3.5, etc.
TO	turnovers per game	0.8, 1.7, etc.
STL	steals per game	0.5, 1.1, etc.
OR	offensive rebounds per game	0.5, 1.4, etc.
DR	defensive rebounds per game	2.3, 4.1, etc.
BLK	blocks per game	0.2, 0.6, etc.
PF	personal fouls per game	1.5, 2.4, etc.
FGM	field goals made per game	2.6, 5.5, etc.
FGA	field goals attempted per game	5.4, 12.2, etc.
3PM	3-point field goals¬ made per game	0.6, 1.9, etc.
3PA	3-point field goals attempted per game	1.9, 5.2, etc.
FTM	free throws made per game	0.8, 2.2, etc.
FTA	free throws attempted per game	1.1, 2.8, etc.
PER	player efficiency rating metric	11.74, 17.27, etc.

Variable	Explanation	Example
SC-EFF	scoring efficiency	1.162, 1.332, etc.
SH-EFF	shooting efficiency	0.48, 0.56, etc.

And here is a small slice of the usage data set.

Name	POS	Team	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK	P
Trae Young	PG	atl	76	76	34.9	28.4	9.7	4.0	0.9	0.7	3.1	0.1	1
John Collins	PF	atl	54	53	30.8	16.2	1.8	1.1	0.6	1.7	6.1	1.0	3
Bogdan Bogdanovic	SG	atl	63	27	29.3	15.1	3.1	1.1	1.1	0.5	3.5	0.2	2
De'Andre Hunter	SF	atl	53	52	29.8	13.4	1.3	1.3	0.7	0.5	2.8	0.4	2
Kevin Huerter	SG	atl	74	60	29.6	12.1	2.7	1.2	0.7	0.4	3.0	0.4	2

The second set of variables are helpful in determining a player's role or function in the game. Some of these variables are Field Goal Percentage, Height, and Weight. Lots of the common variables have been converted into per minute values in order to isolate their frequency. These players will be divided into sub-groups like scorers, big men, and wings. We've termed this data set "role". Below is a data dictionary for the second set of variables.

Variable	Explanation	Example
Name	nba player's first and last name	Trae Young or Trae Young1
POS	playing position	PG (point guard), SG (shooting guard), SF (small forward), PF (power forward), C (center)
Team	abbreviation of city of player's team	atl (Atlanta), bos (Boston), etc.
Height	height in inches	76, 81, etc.
Weight	weight in pounds	200, 234, etc.
PTSPerMin	points per minute	0.356, 0.515, etc.
ASTPerMin	assists per minute	0.055, 0.133, etc.
TOPerMin	turnovers per minute	0.036, 0.065, etc.
STLPerMin	steals per minute	0.023, 0.038, etc.
ORPerMin	offensive rebounds per minute	0.022, 0.066, etc.
DRPerMin	defensive rebounds per minute	0.101, 0.175, etc.
BLKPerMin	blocks per minute	0.009, 0.027, etc.
PFPerMin	fouls per minute	0.064, 0.099, etc.
FGP	field goal percentage	0.417, 0.496, etc.
FGMPerMin	field goals made per minute	0.131, 0.192, etc.
FGAPerMin	field goals attempted per minute	0.284, 0.419, etc.
3PP	3 point percentage	0.306, 0.379, etc.
3PMPerMin	3 point field goals made per minute	0.029, 0.072, etc.

Variable	Explanation	Example
3PAPerMin	3 point field goals attempted per minute	0.094, 0.192, etc.
FTP	free throw percentage	0.709, 0.842, etc.
FTMPerMin	free throws made per minute	0.039, 0.087, etc.
FTAPerMin	free throws attempted per minute	0.053, 0.112, etc.

And here is a small slice of the role data set.

Name	POS	Team	Height	Weight	PTSPerMin	ASTPerMin	TOPerMin	STLPerMin
Trae Young	PG	atl	73	180	0.814	0.278	0.115	0.026
John Collins	PF	atl	81	235	0.526	0.058	0.036	0.019
Bogdan Bogdanovic	SG	atl	78	220	0.515	0.106	0.038	0.038
De'Andre Hunter	SF	atl	80	225	0.450	0.044	0.044	0.023
Kevin Huerter	SG	atl	79	190	0.409	0.091	0.041	0.024

Part 1: Idea of similarity/distance - Interactive

Below is a data set of the ten Dallas Maverick Players from 2021-2022 that met our playing-time restrictions. Kristaps Porzingis was traded in the middle of the season, but he still met our playing-time qualifications for the Dallas Mavericks. For this example, we've combined a few of the variables from both the usage and role data sets. Pick any four of these players.

Name	Height	Weight	MIN	PTS	OR	DR	AST	STL	BLK	TO	2PA	2PP
Luka Doncic	79	230	35.4	28.4	0.9	8.3	8.7	1.2	0.6	4.5	12.8	0.528
Kristaps Porzingis	87	240	29.5	19.2	1.9	5.8	2.0	0.7	1.7	1.6	9.9	0.537
Jalen Brunson	73	190	31.9	16.3	0.5	3.4	4.8	0.8	0.0	1.6	9.6	0.545
Tim Hardaway Jr.	77	205	29.6	14.2	0.3	3.4	2.2	0.9	0.1	0.8	5.4	0.473
Dorian Finney-Smith	79	220	33.1	11.0	1.5	3.2	1.9	1.1	0.5	1.0	3.2	0.599
Dwight Powell	82	240	21.9	8.7	2.1	2.8	1.2	0.5	0.5	0.8	4.4	0.703
Reggie Bullock	78	205	28.0	8.6	0.5	3.1	1.2	0.6	0.2	0.6	1.6	0.550
Maxi Kleber	82	240	24.6	7.0	1.2	4.7	1.2	0.5	1.0	0.8	1.7	0.586
Josh Green	77	200	15.5	4.8	0.8	1.6	1.2	0.7	0.2	0.7	2.7	0.573
Sterling Brown	77	219	12.8	3.3	0.5	2.5	0.7	0.3	0.1	0.5	1.3	0.492

Student picks four players

Name	Height	Weight	MIN	PTS	OR	DR	AST	STL	BLK	TO	2PA	2PP
Dwight Powell	82	240	21.9	8.7	2.1	2.8	1.2	0.5	0.5	0.8	4.4	0.703
Maxi Kleber	82	240	24.6	7.0	1.2	4.7	1.2	0.5	1.0	0.8	1.7	0.586
Josh Green	77	200	15.5	4.8	0.8	1.6	1.2	0.7	0.2	0.7	2.7	0.573
Sterling Brown	77	219	12.8	3.3	0.5	2.5	0.7	0.3	0.1	0.5	1.3	0.492

Look at the four players that you selected. Compare their available statistics. Which of the four players are most similar kinds of players? Which variables make them similar? Which variables do they most differ? Which of the four players are the most "different"? Which variables differentiate them the most? Are they similar in any of the categories?

One common and effective way to compare the similarity of two points (or in this case, players) is the **euclidean distance formula**. The distance formula is found by the following formula:

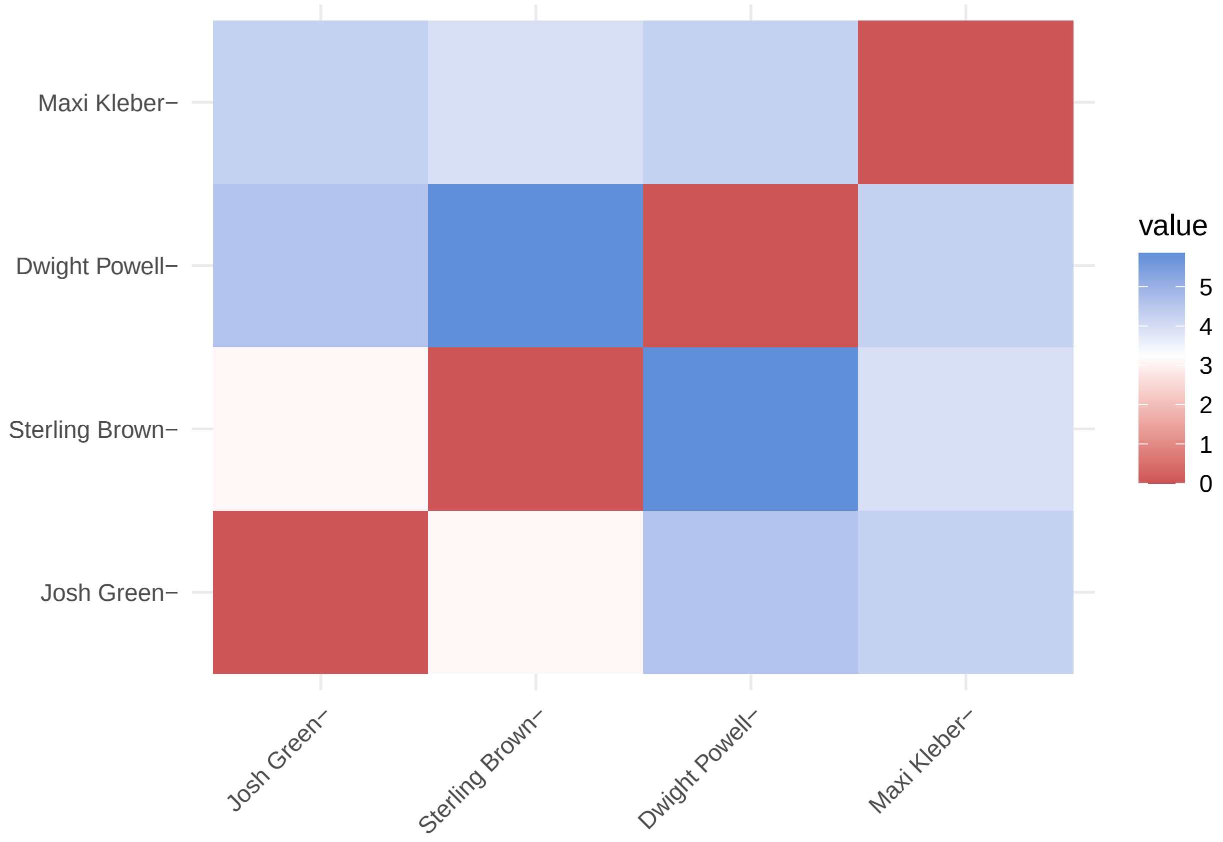
- d = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

You can visualize this as drawing the shortest line possible between two points and then measuring it. Right now, our variables are in different units (inches, pounds, points, percentage, etc.), so we'll standardize (more on this later) each of the variables, so the units are equal. This helps each variable to have equal importance in our distance formula.

Below is a table of the distances between each of the players. Match up the player in the column with the player in the row and you'll find the distance between them. The smaller the value, the more similar the players are.

```
##                  Dwight Powell Maxi Kleber Josh Green
## Maxi Kleber           4.269475                       
## Josh Green            4.554980    4.270914           
## Sterling Brown        5.846063    3.940473   3.102775
```

Below is a visualization of the distances. As the distances increase, the color changes from red to blue. Players matched with themselves will be dark red, because their distance is 0.



Do the tabulated results agree with your previous assessment? Which is more accurate: your original assessment or the similarity metric?

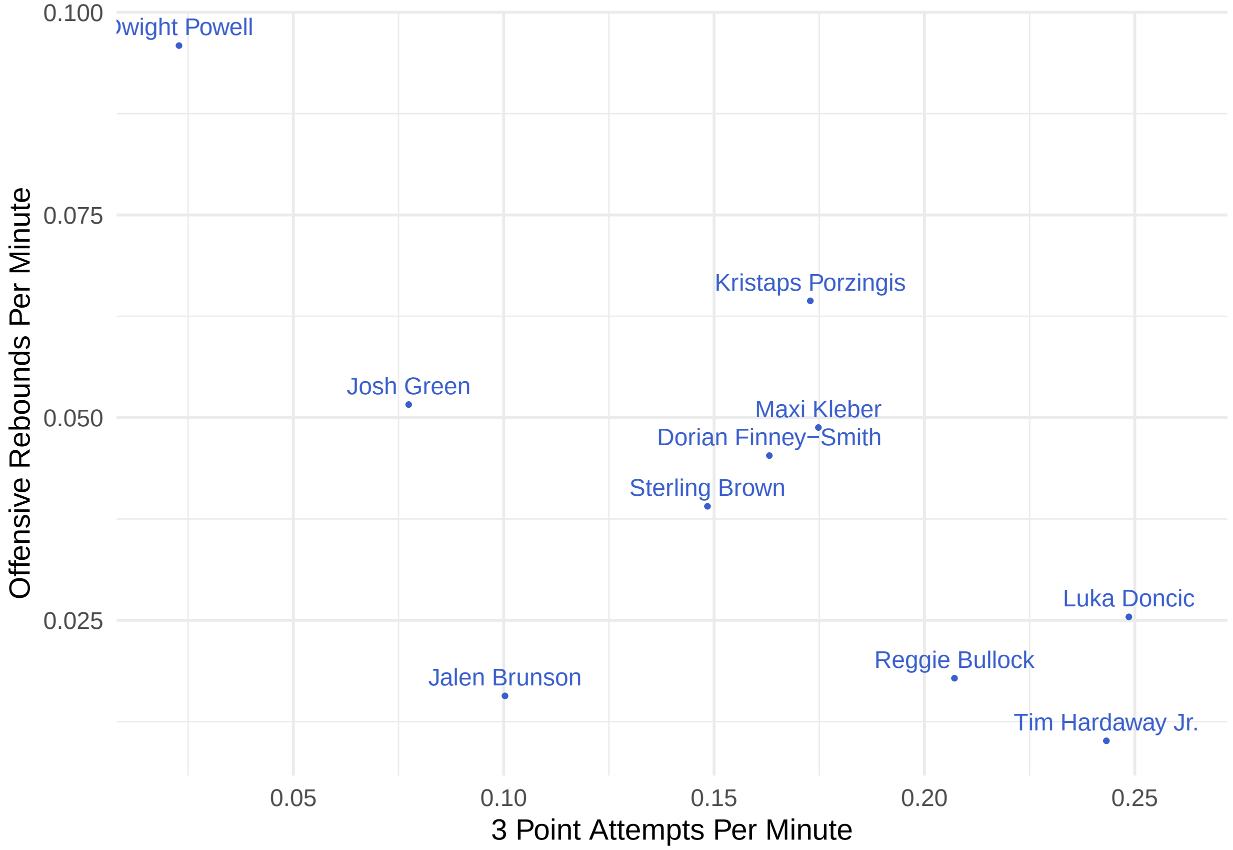
How “good” is the similarity metric? Explore the data for the players that are “closest/most similar” - what in the data makes them close? Do the same for the dissimilar players.

Part 2: Perform the clustering

Dallas Mavericks - Interactive

Calculating the distance between points is the first step in cluster analysis. The players with the smallest distance (or with the most similarity) between them are naturally placed in a cluster together.

How does the clustering actually work? As an illustration, we'll use a basic plot of the Offensive Rebounds and 3-Point Shooting of our Dallas Mavericks players. We've standardized the results by adjusting them to per-minute values.



What do you notice about the data? There are some obvious clusters, but some other lone points. How would you group the players? How would you describe these groupings? In cluster analysis, every point needs to belong to a cluster. Do any points not seem to have a cluster?

K-Means

Cluster analysis is the process of partitioning the data into sub-populations or clusters. This is done so that observations in the same cluster are more similar to each other than observations in a different group. These clusters then can be analyzed.

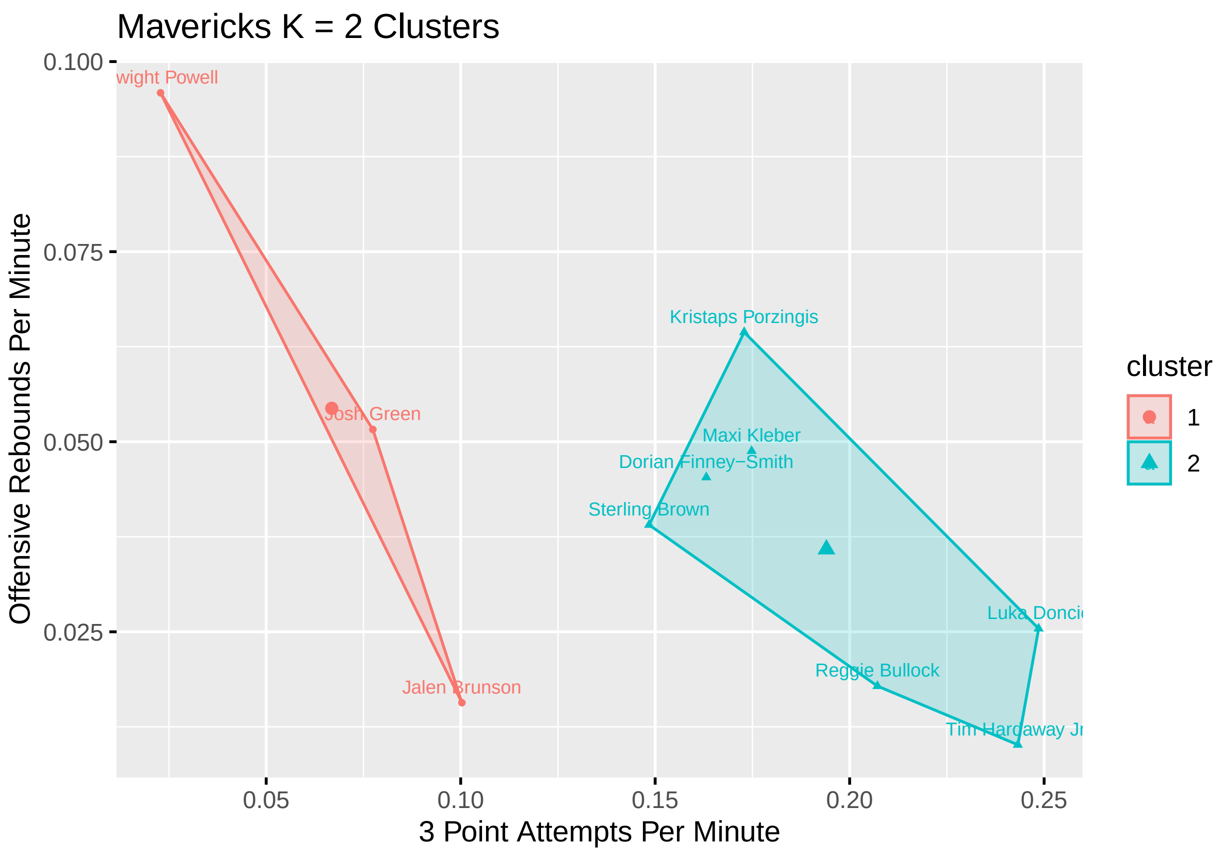
One common method to divide the data into these clusters is distance based and uses the **K-Means Algorithm**. The k-means algorithm partitions the data into clusters which can then be analyzed. Furthermore, this is performed in an unsupervized fashion. This means that the clusters are found by the algorithm and not predetermined by the researcher. In the NBA example, we cannot determine our clusters beforehand. The algorithm may confirm our original intuition, but this is not guaranteed.

The K-Means Algorithm assigns the data into clusters so that the sum squared distance between the center (or mean) of the clusters and each observation is minimized. At the end, the variance of the all the points within each cluster is as small as possible. One downside of the K-Means Algorithm is that users must predetermine the number of clusters they'd like to create. This is entered as the parameter, K. Let's say we want to separate our data into K = 2 clusters. The K-Means algorithm will go through four basic steps:

1. Randomly select two initial cluster centers.
2. Assign each observation to the closest center.
3. Calculate the mean of all the observations within each cluster. These cluster means become the new center of each cluster.
4. Repeat steps 2-3 until no further changes are made.

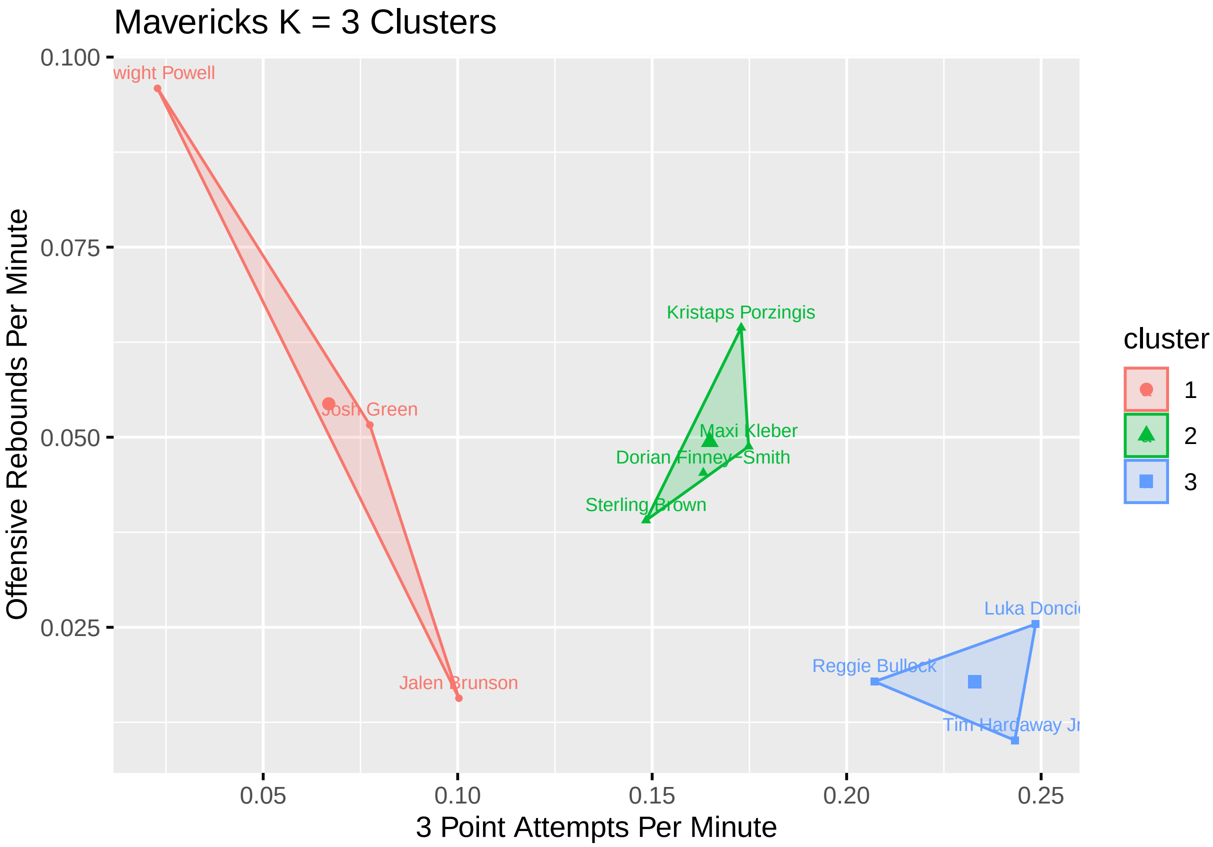
As these steps are followed, the clusters will move closer and closer to their final positions. Since the first step is to randomly assign cluster centers, the K-Means approach can occasionally yield different results. It's worth trying it a few different times with different starting points.

Before you look below, provide your estimation of the two clusters of our Dallas Mavericks players. Where would you anticipate the cluster centers to be located?

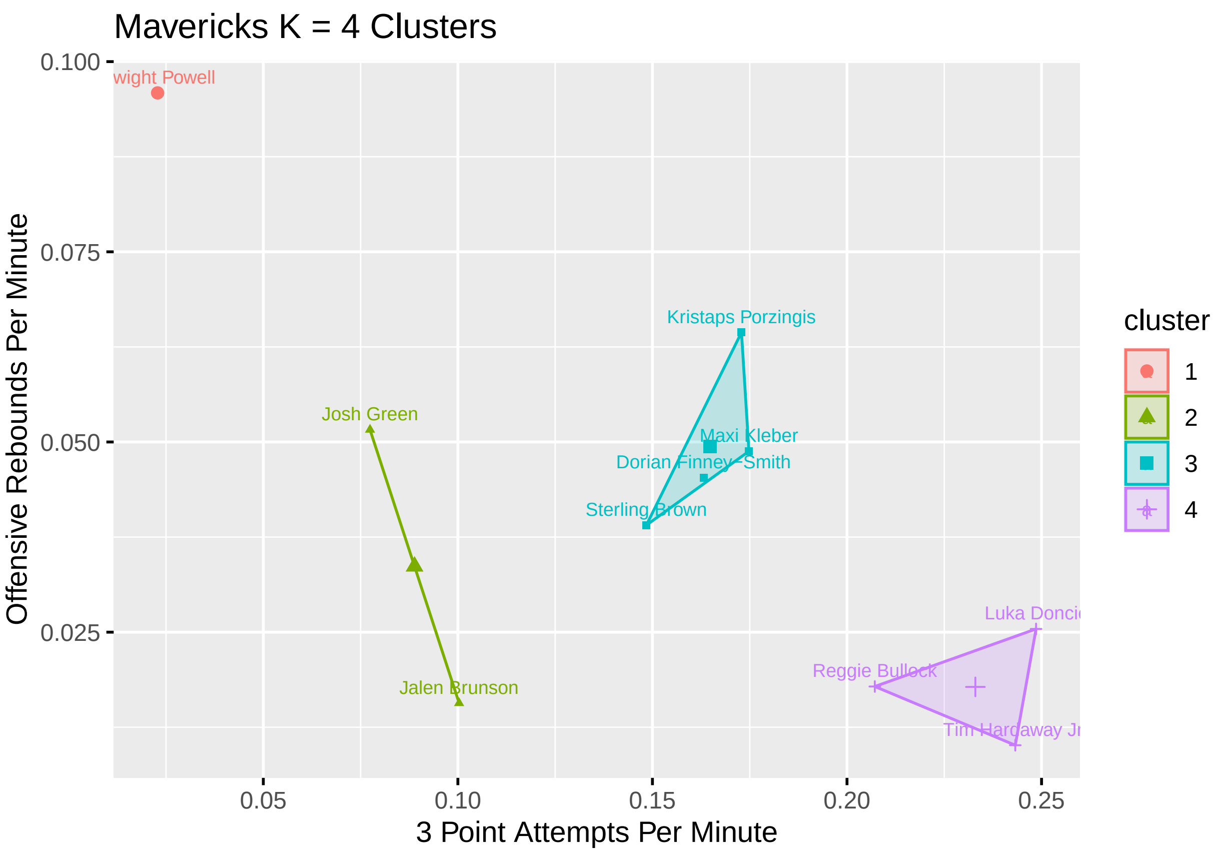


Is this how you would have grouped the players? Notice the large points in the middle of each cluster. These are the cluster centers. Are they where you expected?

How do you think the groupings will change with three clusters? We can easily tell K-Means to randomly assign three centers, and the process of assigning points to cluster means will continue exactly as before.



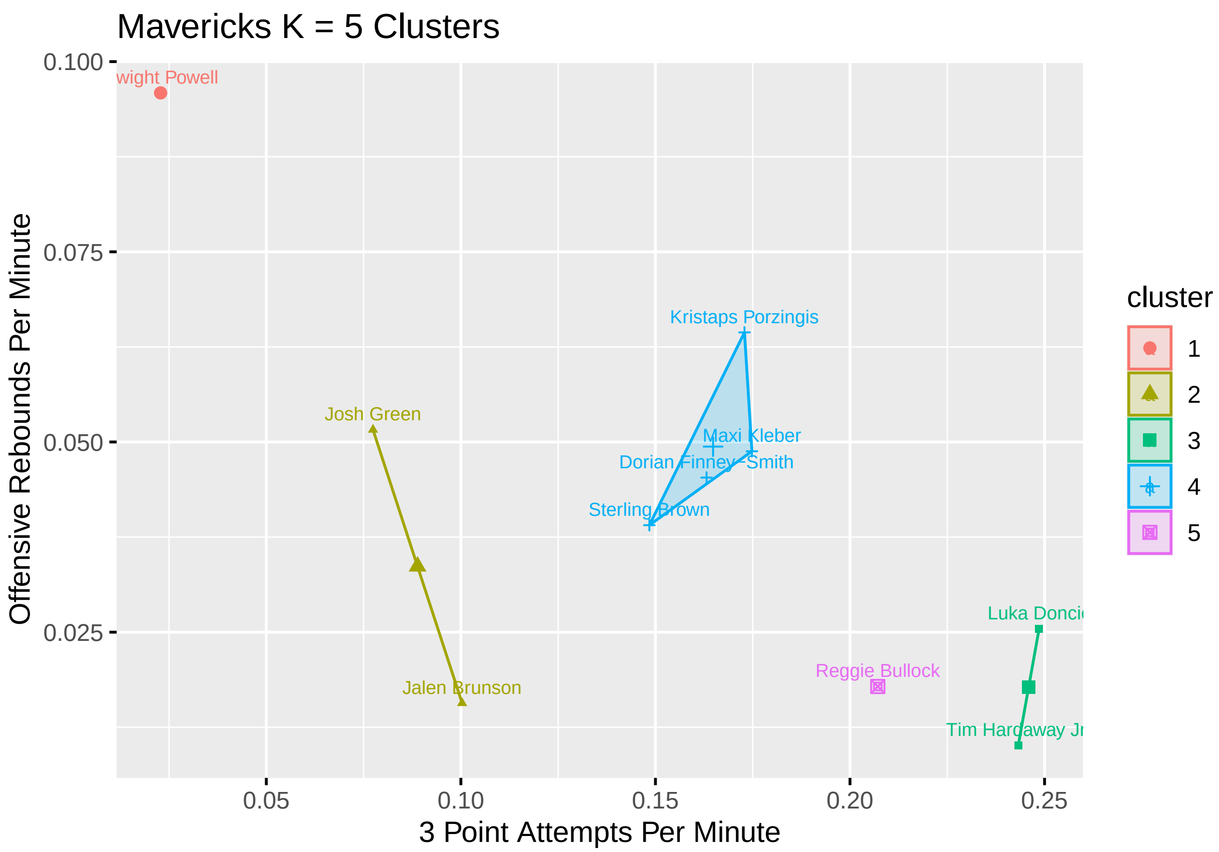
Or four clusters?



Notice Dwight Powell sits in a cluster of his own when K = 4. Is this helpful?

He is truly an outlier among the players. He attempts very few 3 Point Attempts per minute, but he gets more Offensive Rebounds than any other player by far. Clustering him by himself is helpful to identify him as an outlier, but it is unhelpful for comparing him with other players.

Take a look at K = 5 clusters.



At some point, the power of clustering the points begins to fade. Does Dwight Powell deserve to be in a cluster of his own? Possibly. Does Reggie Bullock? Definitely not.

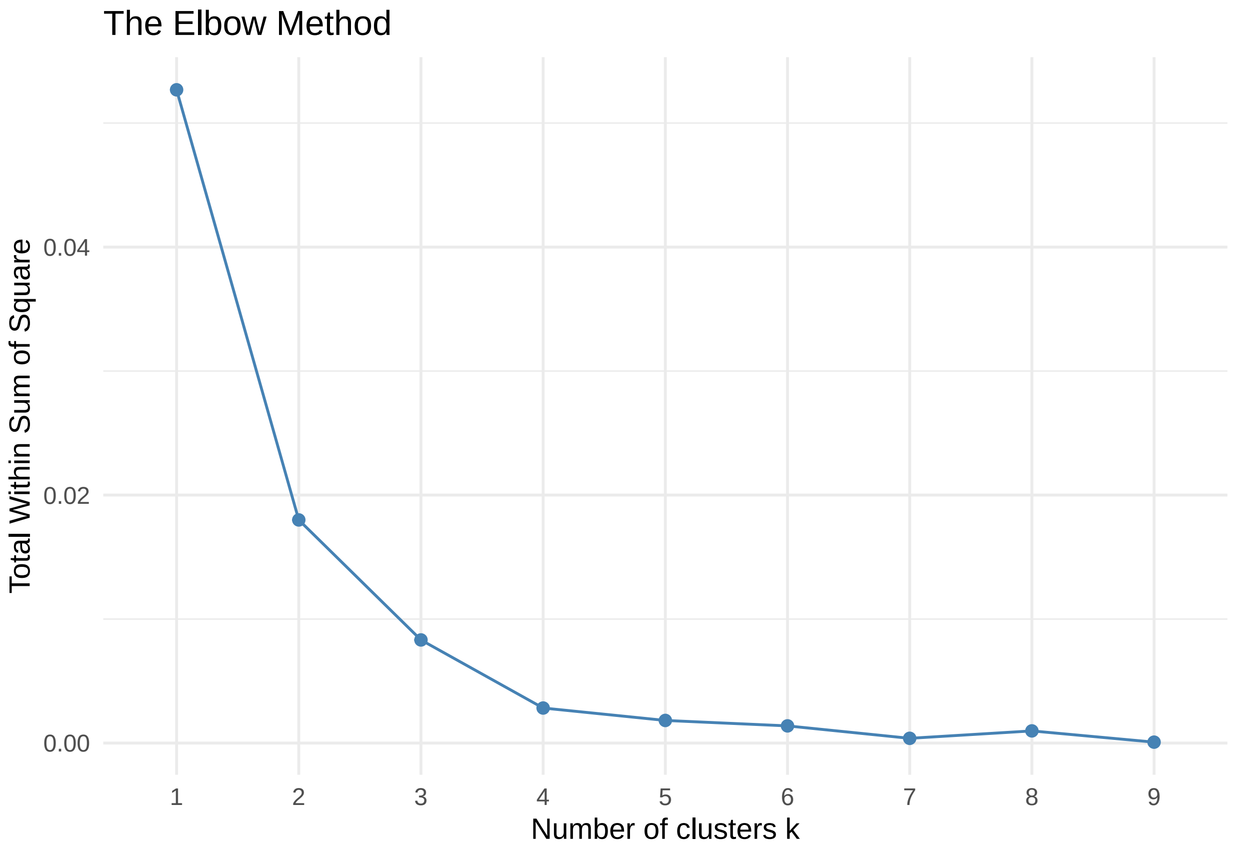
Which of the four values of K did you find most useful or accurate? Were there ever too few or too many clusters to analyze helpfully?

If you're interested, there are lots of other methods to perform cluster analysis. The hierarchical method, for example, is slightly more complex than the partitioning method. It begins with all points in their own clusters, and merges similar points together in a sequential manner. Model-based clustering, constraint-based, density-based, and grid-based are all other useful clustering methods. *Add a link? Include more information on these? Less?*

Optimal Number of Clusters

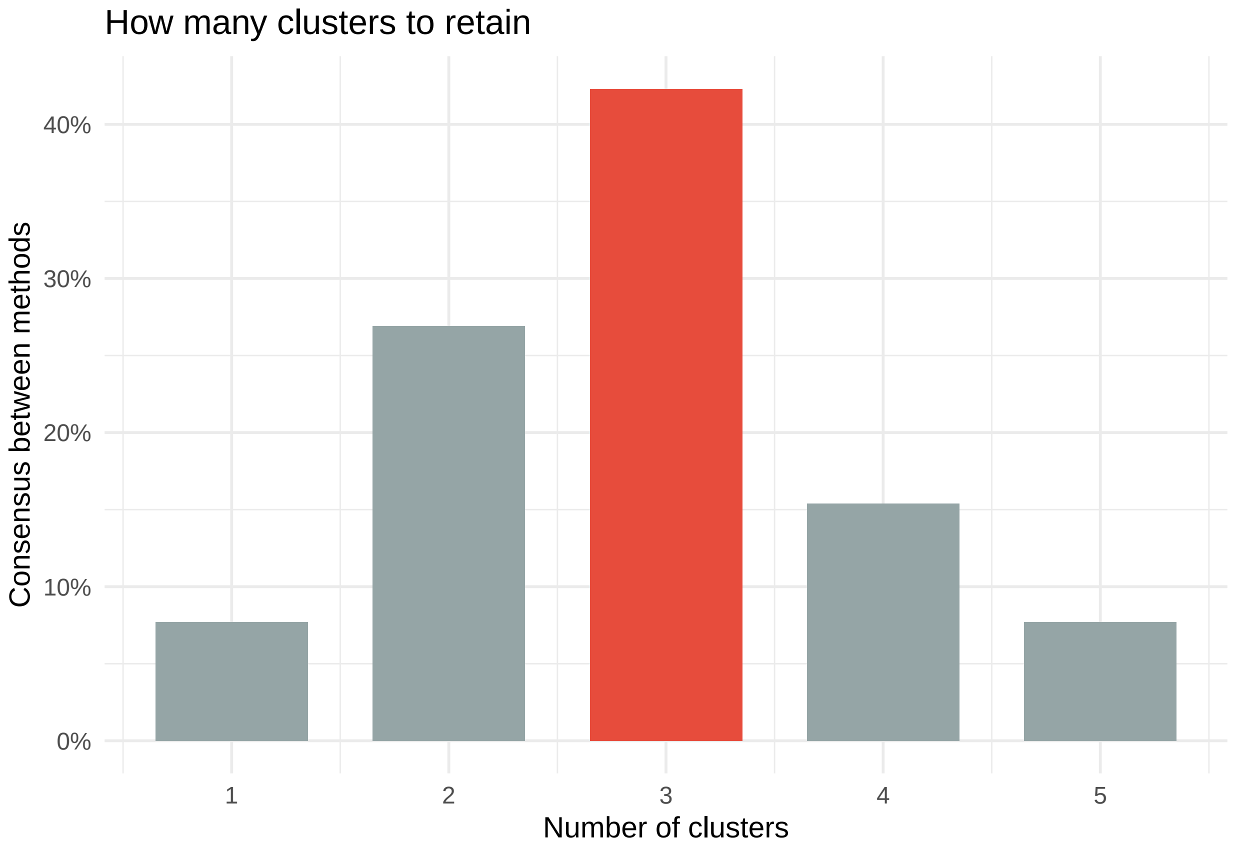
So, how can we choose the optimal number of clusters?

It's helpful to evaluate the effectiveness of the clusters for each value K. There are plenty of ways to test this effectiveness, but we'll walk through a common example called the **Elbow Method**. The Elbow Method totals up the distance between the centers of each cluster and their observations. This is called the **Total Within Summed Squares (TWSS)**. As K increases and more clusters are added to the model, the sum squared distance will decrease. Eventually, the value of each additional cluster diminishes. The Elbow Method plots the results, and the user can look for a point when increasing the number of clusters no longer proves useful. Often, this point looks like an Elbow.



The graph helpfully demonstrates that the value of each additional cluster decreases as more clusters are added. The bends in the graph indicate that clusters beyond four have little value. Despite being common, the Elbow Method is often ambiguous and difficult to interpret. Look for the bend in the Elbow Plot. K = 2, K = 3, and K = 4 all seem like reasonable conclusions from the visualization.

The Elbow plot is just one test to determine the optimal number of clusters. Two other popular methods are the Average Silhouette Method and the Gap Statistic Method. *LINKS to example on these* In all, there are dozens of methods to determine the ideal number of clusters and they often disagree. We'll take a consensus of 27 methods and proceed from there.



The tests give varied estimates for the optimal clusters, but it is up to the user to decide how many clusters you will include in your K-Mean Algorithm. It's common practice to choose several and compare the results of each.

From there, we would conduct our analysis of each cluster and examine the results.

Clustering Strength

After the clustering is completed, how can we analyze our clustering solution?

We want to reduce the Total Within Summed Squares (TWSS) or distance from each observation to its cluster mean, but we also want to minimize the total number of clusters used.

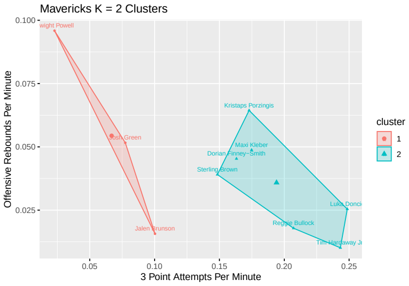
Two helpful measurements to summarize these preferences for our clusters are **intra-class similarity** and **inter-class similarity**.

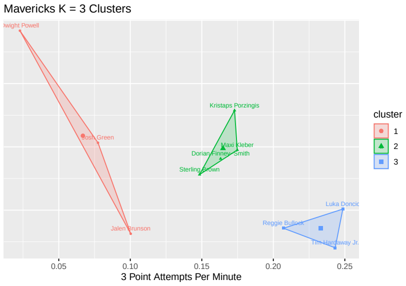
Intra-class similarity tests the relationship between observations of the same cluster. We want this similarity to be high. We want all the observations in a cluster to exhibit similar features.

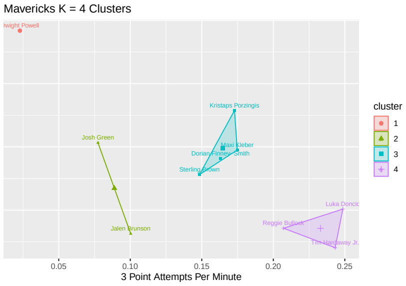
Inter-class similarity tests the relationship between different clusters. We want this relationship to be low. Ideally, each cluster is distinct and the observations within can be clearly assigned to a cluster.

As we increase the number of clusters, K. The intra-class similarity will increase, because observations will be assigned to smaller clusters that a more representative. However, the inter-class similarity will also increase, because the cluster centers are now closer together. This is why it is impractical to choose a large value for K.

Recall our clustering for the Dallas Mavericks players. Which K value has the highest intra-class similarity? Which cluster specifically? Which K value has the highest inter-class similarity?







Usage Data Set

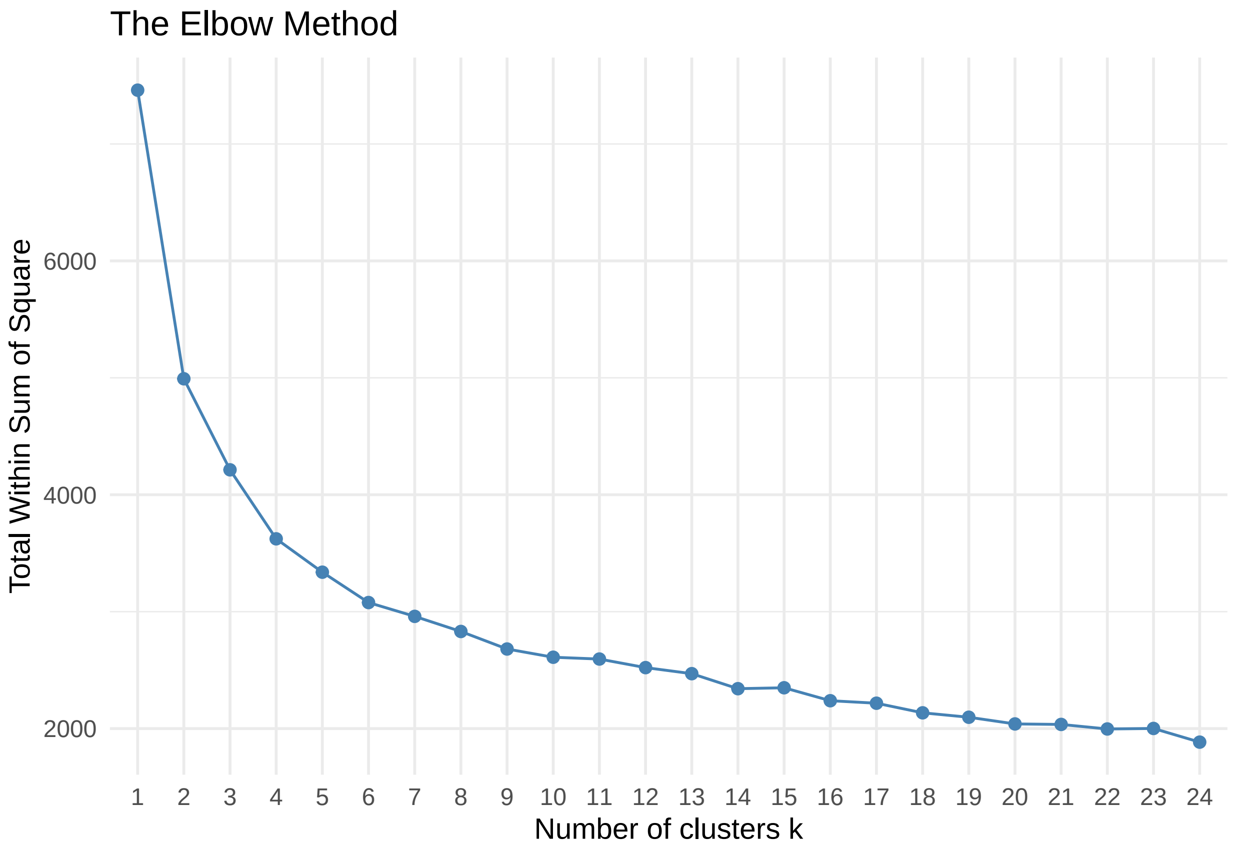
Let's focus now on our larger data sets with many more variables and observations. It seems like it'd be more complicated, but the process is almost exactly the same. One important distinction to remember is that the large number of dimensions make the data difficult to visualize. There are different methods that aid in this visualization. We'll walk you through the usage data set and demonstrate appropriate analysis, and then allow you to work through the role data set.

Remember the usage data set? It contains variables aimed at categorizing the workload and skill of the players. We hope to divide players into sub-groups like stars and bench players.

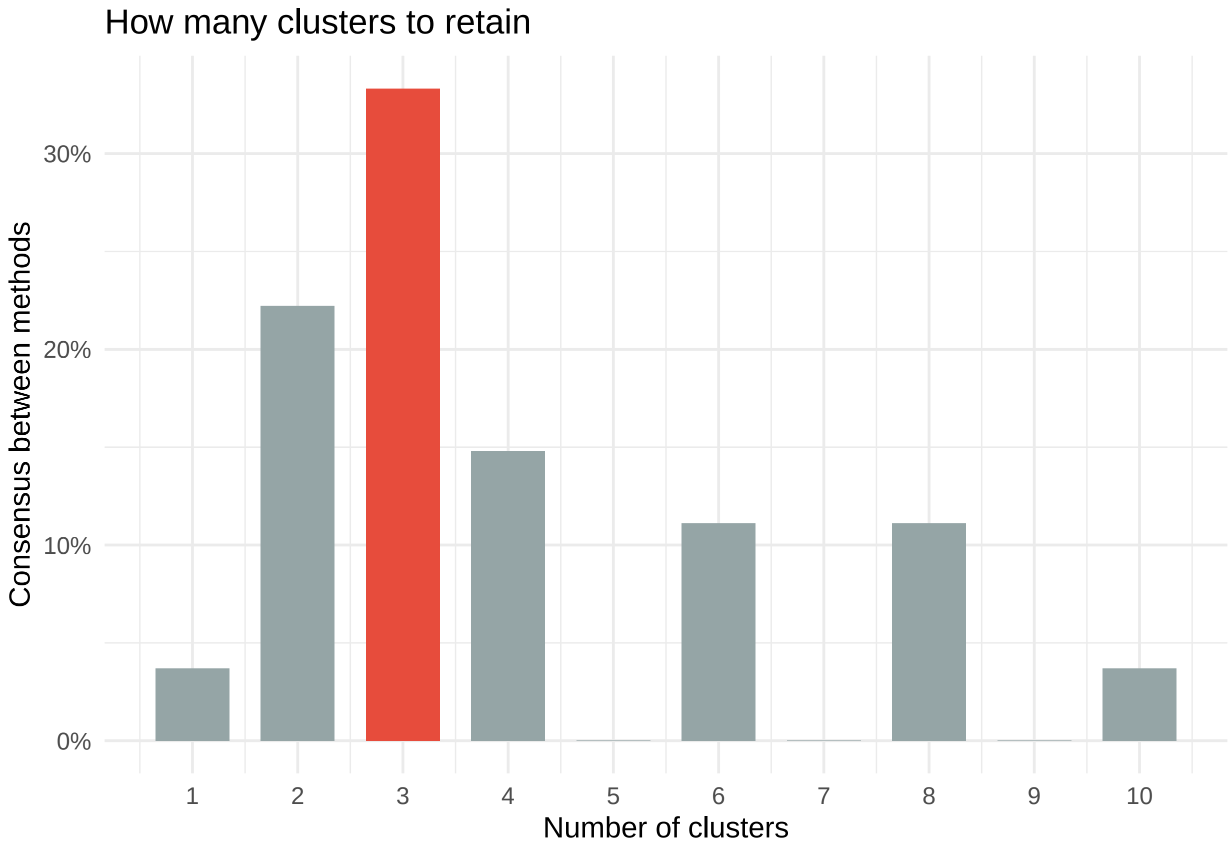
It is very important that we standardize the data first. Lots of our variables have different units. Games played and Blocks per game are hard to compare without scaling. Without standardizing, the large values- like Games Started or Games Played- will exert too much influence on the data. Now, each value is described in relation to the other observations. After standardizing, Trae Young's assist total is 3.656, so we know that he has a lot more assists than the average player in our data set. Often, the standardized data is difficult to contextualize, so we'll want to convert the data back for analysis. Below is a small glimpse into what our standardized data looks like.

Name	POS	Team	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK	F
Trae Young	PG	atl	1.212	1.690	1.478	2.819	3.656	3.186	0.346	-0.422	-0.191	-0.961	-0.
John Collins	PF	atl	-0.221	0.820	0.903	0.802	-0.386	-0.291	-0.479	0.857	1.514	1.296	1.
Bogdan Bogdanovic	SG	atl	0.365	-0.163	0.692	0.620	0.279	-0.291	0.896	-0.678	0.036	-0.710	0.
De'Andre Hunter	SF	atl	-0.286	0.782	0.762	0.339	-0.642	-0.052	-0.204	-0.678	-0.362	-0.209	1.
Kevin Huerter	SG	atl	1.082	1.085	0.734	0.124	0.074	-0.172	-0.204	-0.806	-0.248	-0.209	0.

Let's begin by taking a look at the Elbow plot of the usage dataset.



The Elbow plot shows that the algorithm experiences diminishing returns after K = 2 and K = 3. From the Elbow Plot, we would expect that the consensus lies somewhere between 2 and 5 clusters.



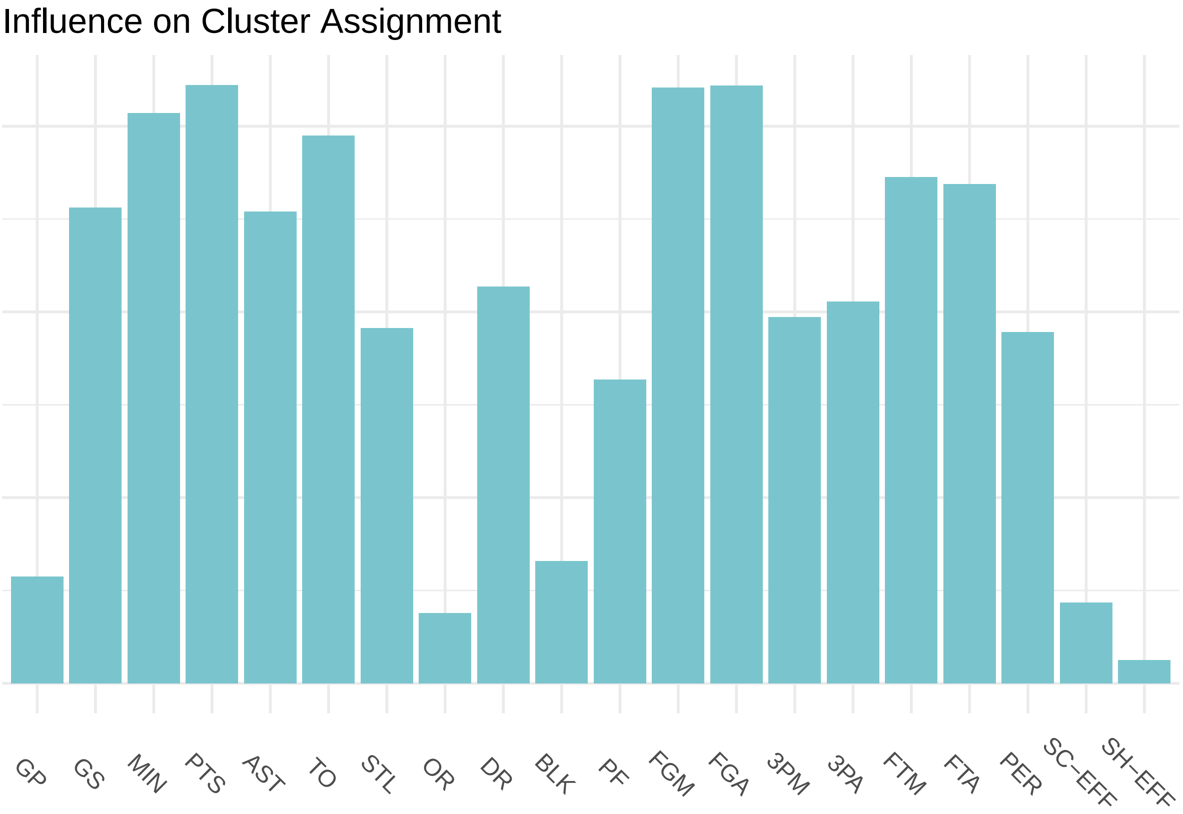
The tests favor three clusters. Some tests also prefer two and four clusters, so those models are worth a look.

K = 2 Clusters - Example

Let's start simple and begin with K = 2 clusters.

But before we begin, let's first look through the variables in our analysis and see which ones have the most influence on the clustering. If some have little or no influence, we can simplify our analysis by removing them.

The visualization below demonstrates the differences between our two clusters. The variables that have large differences are important in the clustering assignment. They greatly influence the assignment of an observation.



This type of exercise is essential for clustering analysis, because it allows one to see which variables are important to consider when classifying an observation.

This visualization scales the centers of the variables for each cluster and contrasts them. Variables with large positive or negative values have a large influence on the clustering. These variables help differentiate the cluster. Variables with an influence close to 0 have less importance.

We see a great diversity in the variables that possess significant influence on the clustering. Field Goals Made and Attempted and Points all seem to carry the largest amount of effect. Interestingly, both of the efficiency metrics: Scoring Efficiency and Shooting Efficiency both lack influence. Games Played, Offensive Rebounds, and Blocks all also don't contribute much to our clustering. We chose to remove Shooting Efficiency and keep the other four, but we easily could have removed them from our analysis.

Note for Reviewer. Removing the five variables causes a slight shift in the cluster assignment. This changes some of the analysis and points I was making on the outliers, and it makes comparison between K = 2 and K = 3 more difficult. We don't remove any of the variables when K = 3. Still, it could make things confusing to not remove variables with very little influence. I'm open to suggestions on what to do here.

Now that we've removed some variables. Let's see how many observations are within each cluster.

Cluster	Size
1	119
2	255

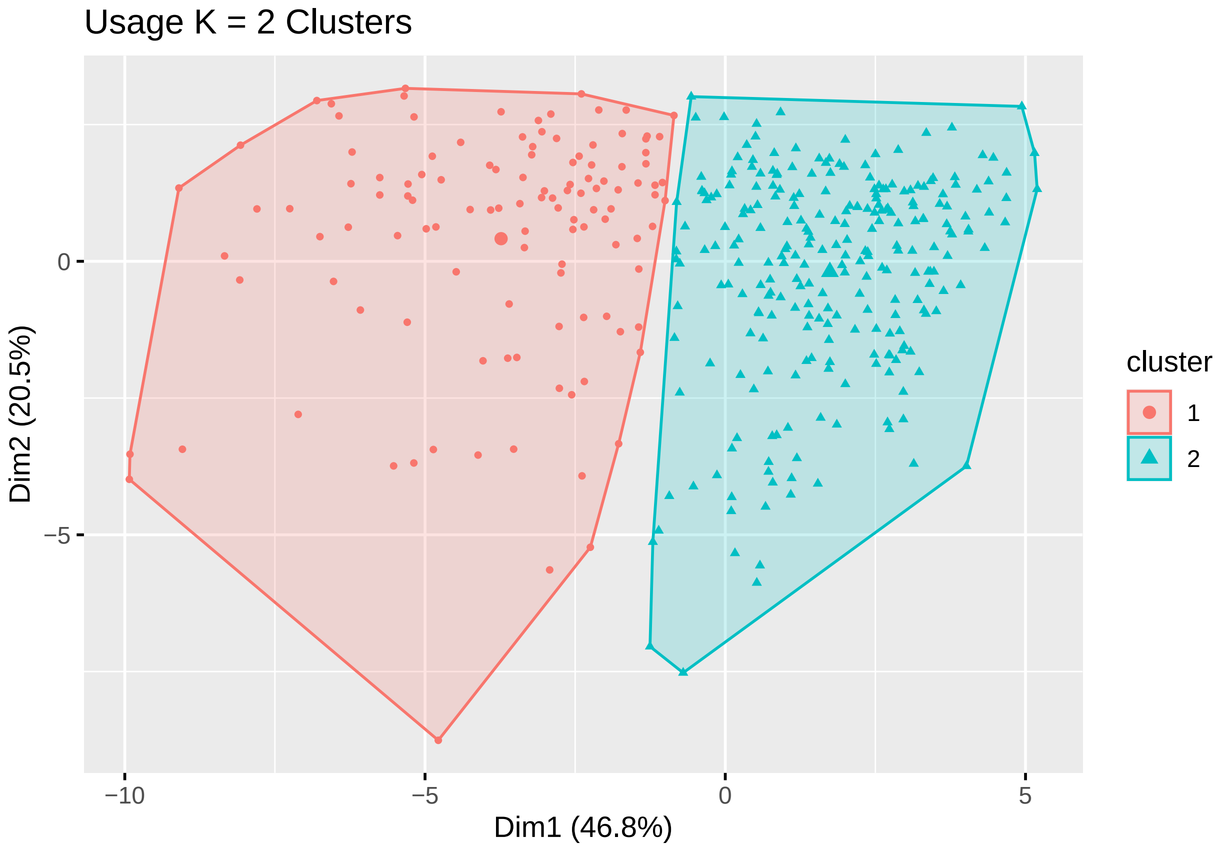
The clusters are not identical in size, and it's different enough that we should keep an eye on it. It's important to verify that each of the clusters contain a significant number of observations. Like we saw with Dwight Powell earlier, sometimes small clusters can tell us valuable information about the observations they contain.

The K-Means Algorithm will assign each observation a cluster and print out descriptive statistics of each cluster. This can give us a good idea of what makes up each cluster. We went back and unstandardized the data.

Clusters	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK	PF	FGM	FGA	3P
1	60.689	56.706	32.572	18.656	4.417	2.266	1.035	1.141	4.845	0.582	2.337	6.773	14.592	1.9
2	55.847	19.455	20.540	7.937	1.685	0.913	0.652	0.978	2.779	0.437	1.789	2.958	6.429	0.9

Generally, it looks like cluster 1 contains starter caliber players and cluster 2 includes the bench players. This helps to explain why cluster 1 is a bit smaller than cluster 2.

Now, let's look at the clusters graphically. This can help us to see how different the clusters really are from each other. The graph is created by combining the values of all the variables in a visually understandable way. This is through a process called Principle Component Analysis (PCA). *Link to more defined explanation of PCA.*



I can't figure out a way to increase the size of the centers. They are only slightly larger than the other points

Many of the observations in both clusters lie close to the border. This indicates that the division between the clusters was close and there may be some observations that could have been placed in either cluster. The centers are fairly close and located at about (-3,0) and (2,0).

There are several large outliers in both clusters, but especially in the lower portion of the visualization in both clusters and the left portion cluster 1.

Prototypes

To help us understand the clusters better, let's look at some players that fall very close to the cluster center. We'll call the players that represent the cluster well **prototype players**.

Name	Cluster	distance
Khris Middleton	1	1.8408
Miles Bridges	1	1.9950
Gordon Hayward	1	2.2011

Khris Middleton is our prototype player for cluster 1. Overall, his characteristics are most similar to the mean or center of the variables within cluster 1. Miles Bridges and Gordon Hayward also lie close to the center. These are the players right next to the center in the visualization. Let's look at their season statistics, so we can get a better understanding of cluster 1 as a whole.

This would be a good opportunity to play highlights of one of the players or show a picture or something to keep people engaged.

Name	POS	Team	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK	F
Khris Middleton	SF	mil	66	66	32.4	20.1	5.4	2.9	1.2	0.6	4.8	0.3	2
Miles Bridges	SF	cha	80	80	35.5	20.2	3.8	1.9	0.9	1.1	5.9	0.8	2
Gordon Hayward	SF	cha	49	48	31.9	15.9	3.6	1.7	1.0	0.8	3.8	0.4	1

The three players all play a similar position; one that allows them to contribute in all areas of the game. There was significant variety in the number of Games Played, but they Started in each game and received a lot of playing time. They all played over 30 Minutes per game and scored about 20 Points a game. Their Rebound, Assist, Block, and Turnover totals vary a little bit, but they are all fairly high. They all took and made roughly the same number of shots per game (15.2-15.9 FGA) and (6.8-7.5 FGM).

Let's move on to cluster 2. First, notice how much smaller the distances are from the cluster 2 center. More observations lie close to cluster 2's center than cluster 1. This is not entirely surprising, as there are almost 100 more players in cluster 2 than 1.

Name	Cluster	distance
Blake Griffin	2	1.1953
Torrey Craig	2	1.2661
Rudy Gay	2	1.2779

Blake Griffin is our prototype player of cluster 2. Torrey Craig and Rudy Gay are also strong representative of cluster 2 as well.

Name	POS	Team	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK
Blake Griffin	PF	bkn	56	24	17.1	6.4	1.9	0.6	0.5	1.1	3.0	0.3
Torrey Craig	SF	ind	51	14	20.3	6.5	1.1	0.8	0.5	1.2	2.7	0.4
Rudy Gay	SF	utah	55	1	18.9	8.1	1.0	0.9	0.5	1.0	3.4	0.3

Once again, the prototypes look like an average NBA player. They each played around 55 Games and Started in very few of them. They played about 17.1-20.3 Minutes a game and scored from 6.4-8.1 Points a game. Their Rebound, Assist, Steal, Block, Turnover, and Foul values are fairly low and generally close together. They also don't take as many shots as cluster 1 - only about 6 Field Goal Attempts per game.

Outliers

Now, let's look through some of the players that fall farthest from the center of their cluster. These players are **cluster outliers**. In these cases, the clustering least represents the observation. These players are very different from the center. It can be helpful to identify and explain outliers by comparing them to our prototype players. How do they differ? What attributes led to their classification?

Is there a way to only label a few of the points in the visualization

Name	Cluster	distance
Rudy Gobert	1	9.1597
Joel Embiid	1	8.8545
Myles Turner	1	6.6678

Name	POS	Team	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK
Rudy Gobert	C	utah	66	66	32.1	15.6	1.1	1.8	0.7	3.7	11.0	2.1
Joel Embiid	C	phi	68	68	33.8	30.6	4.2	3.1	1.1	2.1	9.6	1.5
Myles Turner	C	ind	42	42	29.4	12.9	1.0	1.3	0.7	1.5	5.5	2.8
Khris Middleton	SF	mil	66	66	32.4	20.1	5.4	2.9	1.2	0.6	4.8	0.3
Blake Griffin	PF	bkn	56	24	17.1	6.4	1.9	0.6	0.5	1.1	3.0	0.3

Sometimes, you'll need to do some digging on the outliers. We chose to show you Khris Middleton and Blake Griffin's characteristics again for comparison. Joel Embiid, Giannis Antetokounmpo, and Myles Turner represent two very different kinds of outliers. Embiid and Giannis are superstars. They finished second and third in the MVP voting in the 2021-2022 season. They are very far from the prototype of cluster 1, but they are even further from the prototype of cluster 2. These are the points near (-10, -5) in the visualization.

Myles Turner, however, possesses some attributes that could be classified as cluster 1 and cluster 2. He played lots of Minutes, Started most games, and had strong Rebounding values. However, his shooting numbers fall right between the clusters, and he doesn't tally very many Points, Assists, Steals, or Turnovers. This point is likely the (-5, -9) outlier in the visualization. He is a borderline case. *Is there a more statistical word for this?*

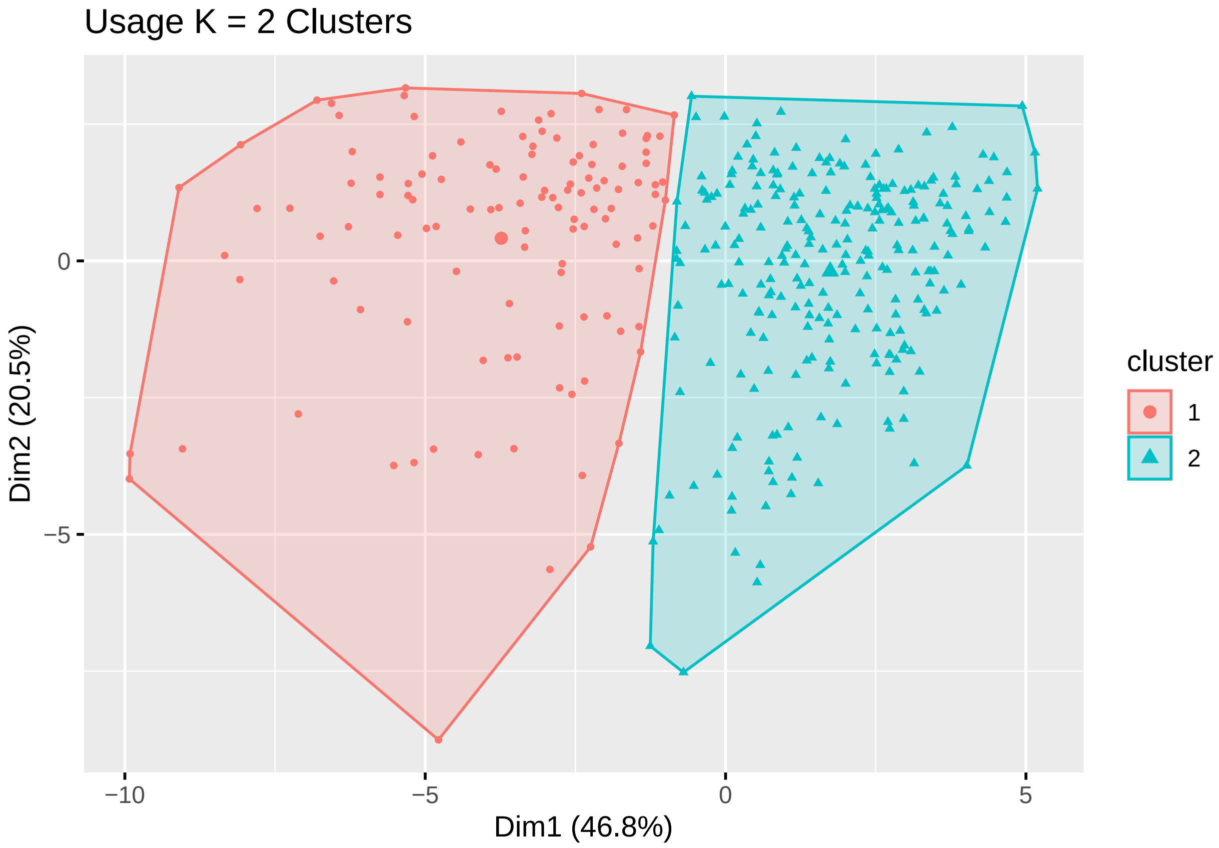
Name	Cluster	distance
Robert Williams III	2	7.4501
Mitchell Robinson	2	7.3269

Name	Cluster	distance
Clint Capela	2	6.4898

Name	POS	Team	GP	GS	MIN	PTS	AST	TO	STL	OR	DR	BLK	P
Robert Williams III	C	bos	61	61	29.6	10.0	2.0	1.0	0.9	3.9	5.7	2.2	2
Mitchell Robinson	C	ny	72	62	25.7	8.5	0.5	0.8	0.8	4.1	4.5	1.8	2
Clint Capela	C	atl	74	73	27.6	11.1	1.2	0.6	0.7	3.8	8.1	1.3	2

These cluster 2 outliers are all similar players. Robert Williams III, Mitchell Robinson, and Clint Capela are all big men. Like Myles Turner, they are players that play a lot of Games and Minutes, get lots of Rebounds and Blocks, but don't shoot very much. Our data emphasizes shooting a lot and perhaps this leaves players like these without an appropriate cluster. They are borderline candidates that perhaps would benefit from another cluster.

Now, let's analyze the strength of K = 2 clusters. For reference, we've repeated the visualization below.



The two clusters possess strong inter-class differences. For only two clusters, cluster 1 and cluster 2 are fairly distinct. The centers are far apart and demonstrate two different classifications of players. Cluster 1 is clearly a sub-population of starting, high-volume players and cluster 2 is a sub-population of bench players. Still, we've analyzed the outliers and found some players that could fall in either cluster. There could be some confusion for players like Robert Williams and Myles Turner. These players seem more similar to each other than most of the players in their own cluster. These outliers fall around (-2, -7). Check the visualizations again to see the cluster of players near there.

The intra-class similarity is fairly low. The clusters are large and have many outliers in each of the directions. Players like Giannis Antetokounmpo, Khris Middleton, and Myles Turner have little in common, but they are all grouped into cluster 1. Yet, most of cluster 1 produce larger values and most of cluster 2 have smaller numbers.

K = 3 Clusters - Interactive

Now, let's look at the consensus tests' most popular number of clusters: K = 3. Here, we'd like you to produce your own analysis of the results. If you need help, look back at the K = 2 example.

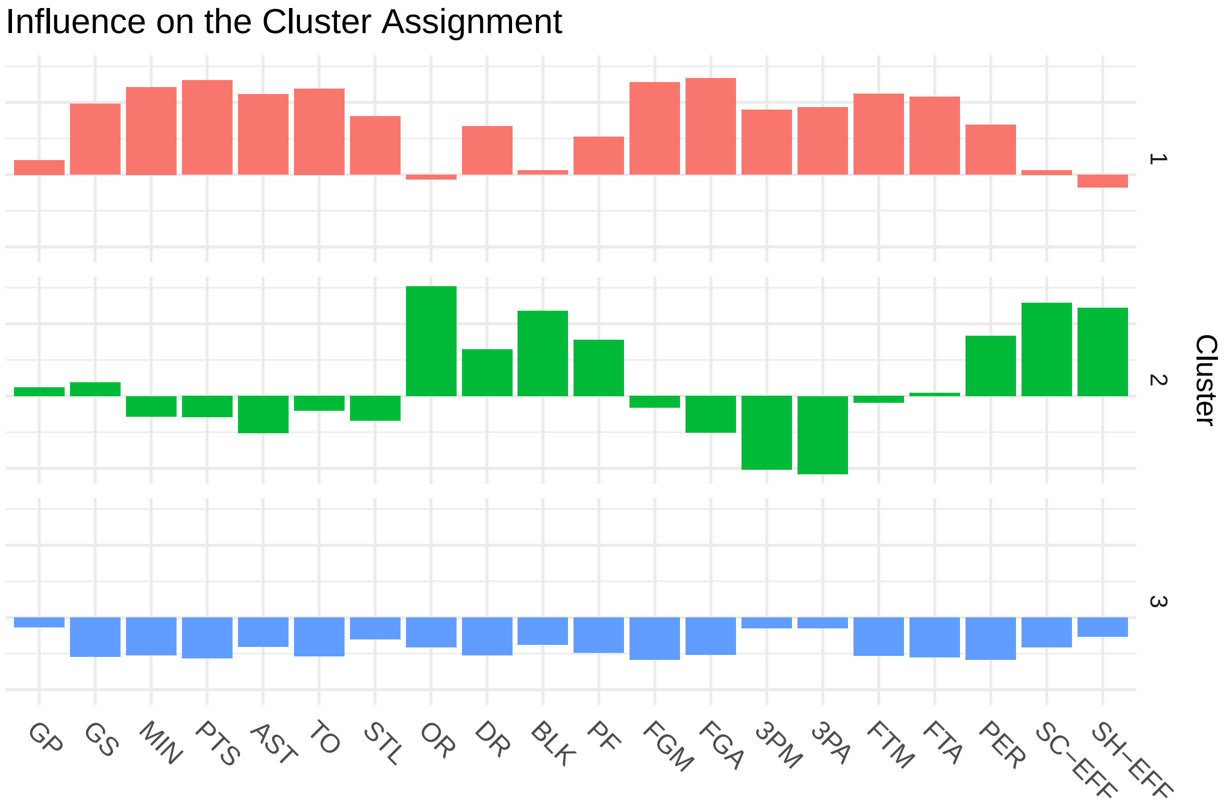
As you progress, fill out this table with descriptors of the three clusters. This will be helpful for you as you try to identify their distinctions.

Cluster	Description
1	
2	
3	

Once again, let's first look through the variables in our analysis and see which ones have the most influence on the clustering.

This visualization plots the centers for each variable in a cluster. At a glance, this helps us to understand the characteristics of each cluster. We can see that cluster 2, for example, has high offensive rebounds and blocks per game, but low 3 point attempts and 3 point makes.

It can also tell us what variables are unimportant. If a variable has the similar mean throughout all three clusters, then the variable does not help us to distinguish between the clusters. If a variable has a large positive value in one cluster and a large negative value in another, then that variable is very useful for classifying our data.



Before you analyze, remember that variables with a strong negative value still have large influence. It's just a negative association with a variable instead of a positive association.

What do you notice about the variables? Which kinds of variables possess significant influence? Some variables have a strong influence in one cluster, but a weak influence in another cluster. Why is this?

After analyzing, would you choose to remove any variables from the data?

Is there a better way to look at the variables and remove the less influential ones?

We chose to remove the Games Played variable, because its influence was close to 0 in all three clusters. All of the other variables had a large effect in some category.

Now that we've removed some variables. Let's see how many observations are within each cluster.

Cluster	**Size**
1	102
2	61
3	211

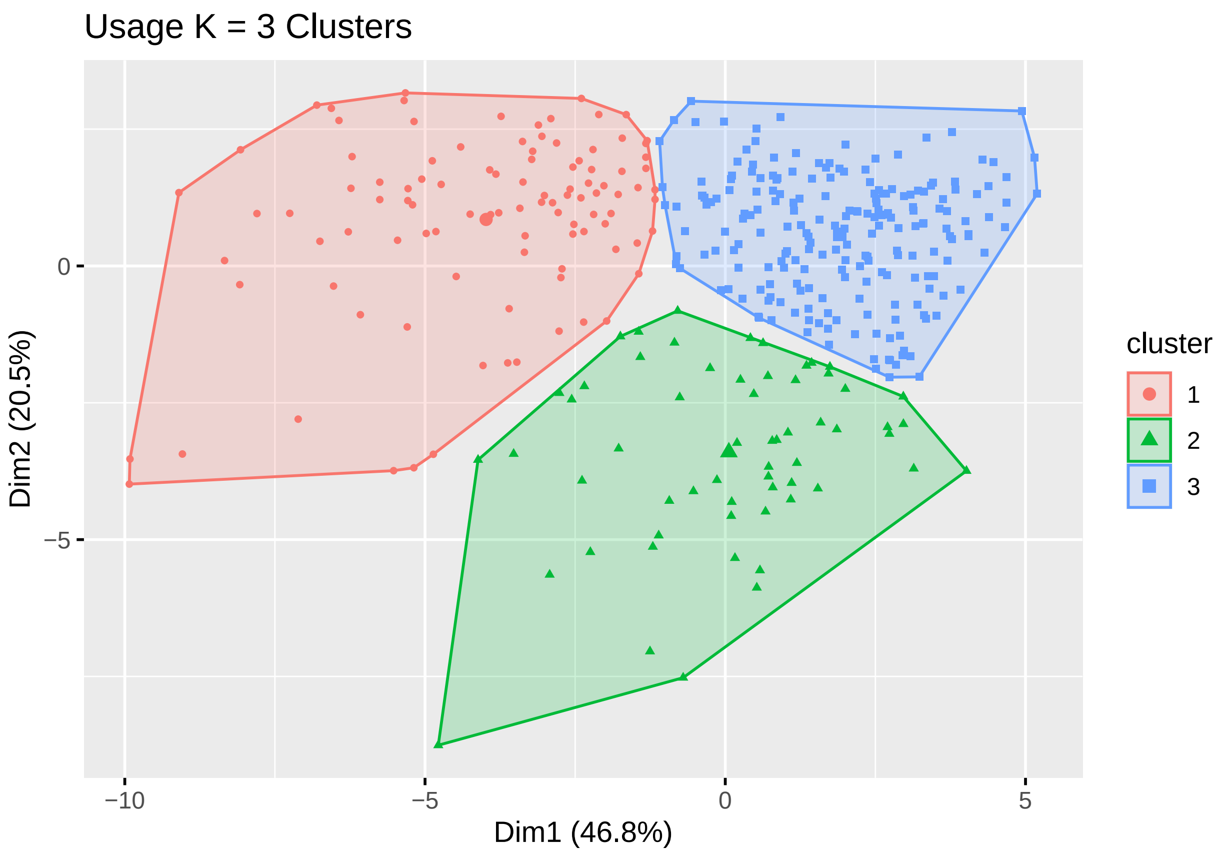
What do you notice about the cluster size? What could this tell us about the clusters?

The clusters are not identical in size, but the clusters are each large enough that there is no reason to be concerned.

Clusters	**GS**	**MIN**	**PTS**	**AST**	**TO**	**STL**	**OR**	**DR**	**BLK**	**PF**	**FGM**	**FGA**	**3PM**	**3P**
1	57.539	33.089	19.346	4.773	2.355	1.071	0.983	4.652	0.513	2.289	6.966	15.224	2.066	5.7
2	36.295	22.354	9.580	1.551	1.175	0.649	2.220	4.575	0.954	2.438	3.818	6.684	0.351	1.0
3	17.185	20.735	7.992	1.773	0.902	0.667	0.709	2.519	0.333	1.669	2.924	6.708	1.139	3.2

What do you notice about the cluster means? Without looking any further, how would you describe the three clusters? Jot down some notes in your table.

Now, let's look at the clusters graphically.



What do you notice about the visualization? Are there a lot of observations that reside on the border? Where are the centers and outliers of each cluster?

Compare the new visualization with the K = 2 visualization. Where did the third cluster come from? What kinds of players?

If you were to create a fourth cluster, what points would you group together?

Let's look at our prototype and outlier players. We've compiled them all into a table for you to compare and contrast.

Name	Cluster	Category	POS	Team	GS	MIN	PTS	AST	TO	STL	OR
Miles Bridges	1	Prototype	SF	cha	80	35.5	20.2	3.8	1.9	0.9	1.1
Malcolm Brogdon	1	Prototype	PG	ind	36	33.5	19.1	5.9	2.1	0.8	0.9
Khris Middleton	1	Prototype	SF	mil	66	32.4	20.1	5.4	2.9	1.2	0.6
Nikola Jokic	1	Outlier	C	den	74	33.5	27.1	7.9	3.8	1.5	2.8
Giannis Antetokounmpo	1	Outlier	PF	mil	67	32.9	29.9	5.8	3.3	1.1	2.0
Joel Embiid	1	Outlier	C	phi	68	33.8	30.6	4.2	3.1	1.1	2.1
Nic Claxton	2	Prototype	PF	bkn	19	20.7	8.7	0.9	0.8	0.5	1.9
Isaiah Roby	2	Prototype	PF	okc	28	21.1	10.1	1.6	1.0	0.8	1.7
Richaun Holmes	2	Prototype	C	sac	37	23.9	10.4	1.1	1.2	0.4	2.1
Robert Williams III	2	Outlier	C	bos	61	29.6	10.0	2.0	1.0	0.9	3.9

Name	Cluster	Category	POS	Team	GS	MIN	PTS	AST	TO	STL	OR
Myles Turner	2	Outlier	C	ind	42	29.4	12.9	1.0	1.3	0.7	1.5
Rudy Gobert	2	Outlier	C	utah	66	32.1	15.6	1.1	1.8	0.7	3.7
Damion Lee	3	Prototype	SG	gs	5	20.0	7.4	1.0	0.6	0.6	0.4
Ziaire Williams	3	Prototype	SG	mem	31	21.7	8.1	1.0	0.7	0.6	0.4
Rudy Gay	3	Prototype	SF	utah	1	18.9	8.1	1.0	0.9	0.5	1.0
Tomas Satoransky	3	Outlier	SG	no	3	15.0	2.8	2.4	0.7	0.4	0.6
Robert Covington	3	Outlier	PF	por	40	29.8	7.6	1.4	1.2	1.5	0.9
Buddy Hield1	3	Outlier	SG	sac	6	28.6	14.4	1.9	1.6	0.9	0.8

Here is a smaller table that may help you compare the players more easily.

Name	Cluster	Category	POS	Team	GS	MIN	PTS	AST	TO	STL	OR
Khris Middleton	1	Prototype	SF	mil	66	32.4	20.1	5.4	2.9	1.2	0.6
Isaiah Roby	2	Prototype	PF	okc	28	21.1	10.1	1.6	1.0	0.8	1.7
Damion Lee	3	Prototype	SG	gs	5	20.0	7.4	1.0	0.6	0.6	0.4
Joel Embiid	1	Outlier	C	phi	68	33.8	30.6	4.2	3.1	1.1	2.1
Rudy Gobert	2	Outlier	C	utah	66	32.1	15.6	1.1	1.8	0.7	3.7
Tomas Satoransky	3	Outlier	SG	no	3	15.0	2.8	2.4	0.7	0.4	0.6

Use the above tables to summarize each of the 6 categories. What kind of players belong in each category? Is there a lot of variation within the prototypes? Is there a lot of variation within the outliers? Which of the outliers are closest to a different cluster? Would you reclassify any of the outliers?

After looking through the clusters, why do you think cluster 2 is so much smaller?

Let's analyze the overall strength of K = 3 clusters. How does the intra-class similarity compare with K = 2? The inter-class similarity?

Comparing K = 2 to K = 3 - Mix

Often, it is interesting to compare the cluster results. Here, we tabulated the cluster assignments between K = 2 and K = 3. This can help us to see how the clustering with K = 2 overlaps with K = 3.

Cluster	1	2	3
1	102	13	4
2	0	48	207

What do you notice about the clustering distribution?

We can see that most players in cluster 1 from K = 2 stayed in cluster 1 when K = 3. We identified both of these clusters as the "starters," so this makes a lot of intuitive sense. Most of cluster 2 from K = 2 moved into cluster 3 when K = 3. The interesting transition comes with the middle cluster of K = 3. This cluster is full of big men that don't score a lot. They came from both cluster 1 and cluster 2 of K = 2. We saw this in our outlier analysis earlier.

What are the benefits and costs of both K = 2 and K = 3? Which would you choose?

Role Data Set

Now we move on to a second data set and we want to give you a lot more autonomy to test different clusters or outliers yourself. The data set is different, but the process is almost exactly the same. If you have questions, we'll give you hints or you can look back to the usage data set for a clear example.

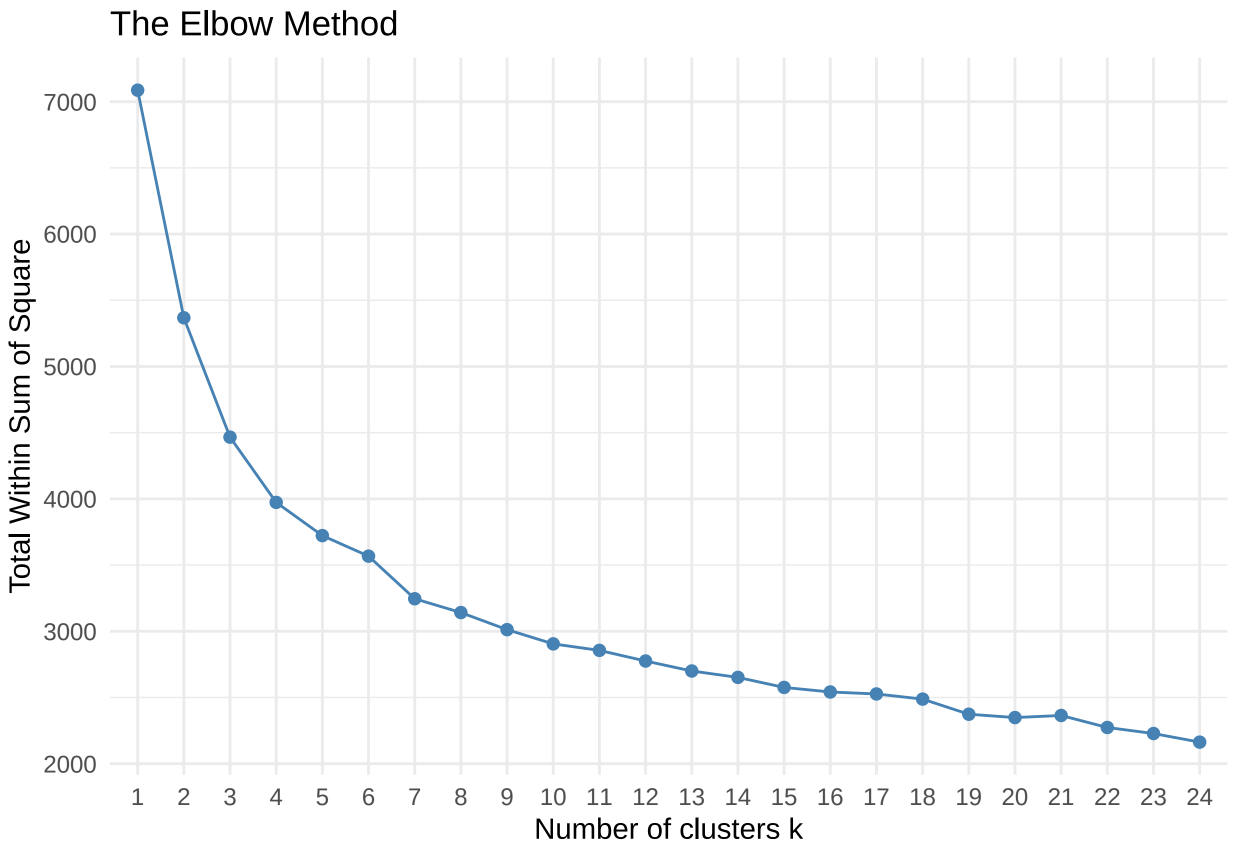
Remember the role data set? It contains variables aimed at categorizing the function and specific characteristics of the players. We hope to divide players into sub-groups like scorers, 3-point shooters, and rebounders.

Even though most of our data has been set to adjusted "per minute" quantities. It is still very important that we standardize the data first. Otherwise common values like points per minute will outweigh the effect of less common characteristics like blocks per minute. Now each variable is on the same scale. Often, the standardized data is difficult to contextualize, so we'll want to convert the data back for analysis. Below is a small glimpse into what our standardized data looks like.

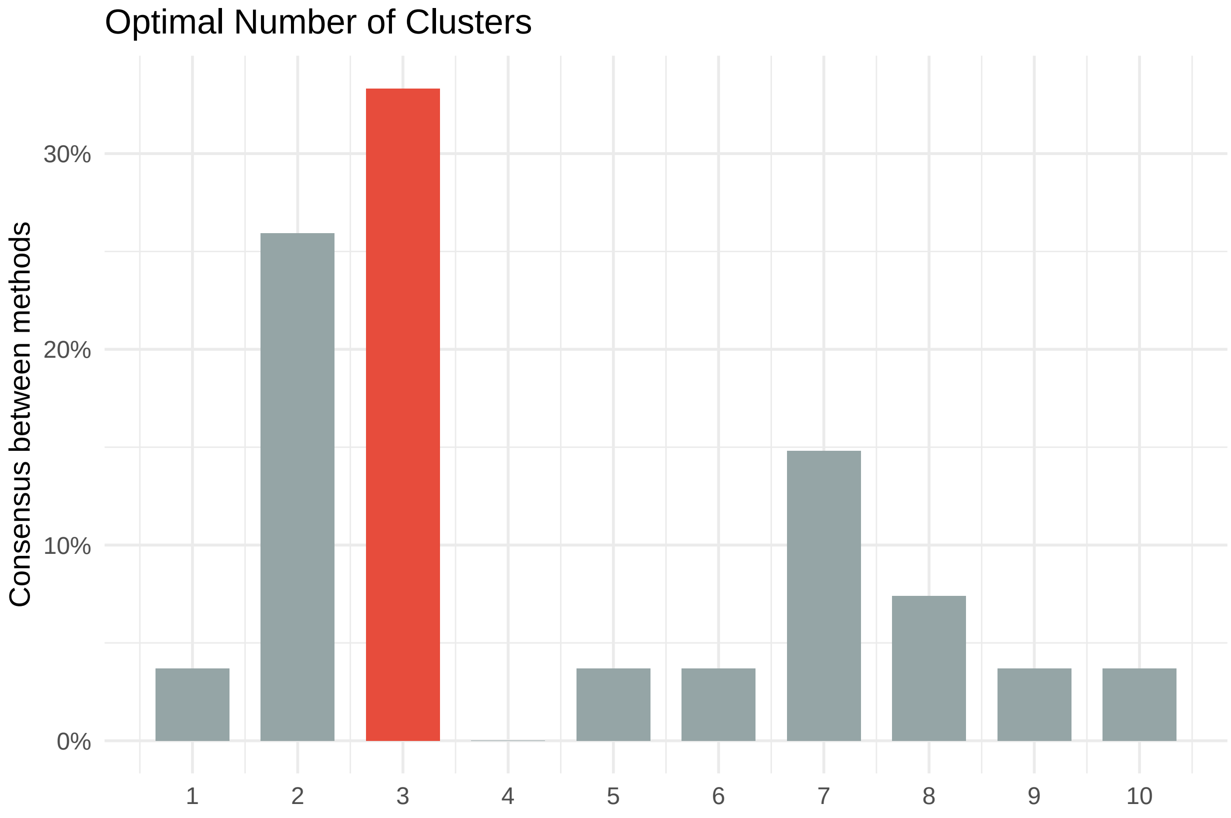
I could also give a short mini lesson on the importance of standardizing using games started and blocks or something like that.

Name	POS	Team	Height	Weight	PTSPerMin	ASTPerMin	TOPerMin	STLPerMin
Trae Young	PG	atl	-1.655	-1.496	2.824	3.171	2.751	-0.499
John Collins	PF	atl	0.842	0.774	0.620	-0.708	-0.773	-1.019
Bogdan Bogdanovic	SG	atl	-0.094	0.155	0.539	0.129	-0.692	0.469
De'Andre Hunter	SF	atl	0.530	0.362	0.036	-0.970	-0.420	-0.689
Kevin Huerter	SG	atl	0.218	-1.083	-0.277	-0.129	-0.558	-0.676

Let's check our Elbow plot to get an idea of the clustering.



What do you see from the Elbow plot? At what point do the returns diminish? How many clusters does the Elbow plot suggest?



There's a lot of variation in the preferred number of clusters. How many clusters would you choose to analyze? How many values of K would you like to analyze? This is totally up to you. Feel free to move back and forth through this section to analyze the data as much as you like.

We will be using K = 7 for the trade scenario portion, so we recommend you review through K = 7.

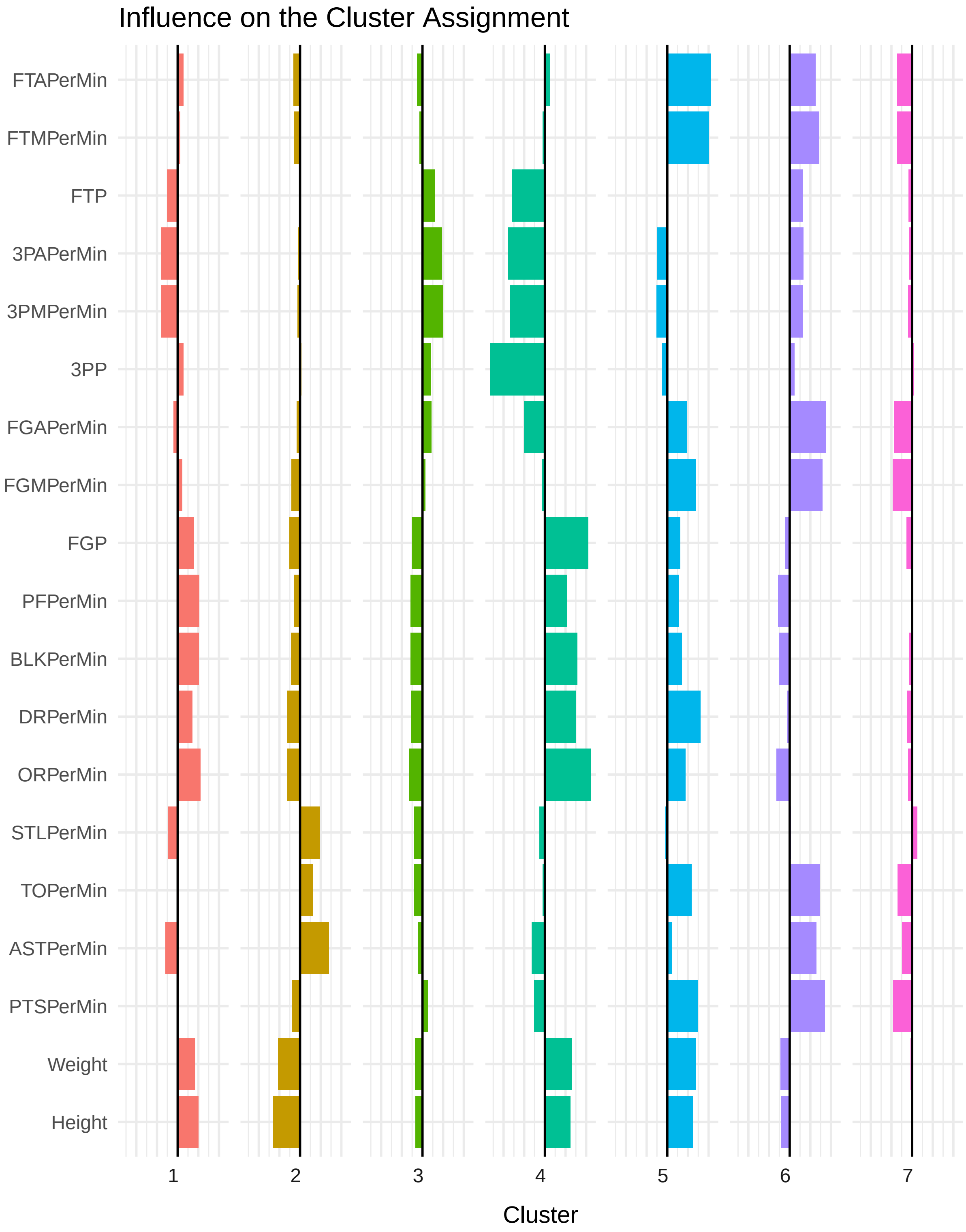
give them space to choose

Ok, you've chosen K = 7. Here is an empty table for you to describe each of the clusters. As you grow in understanding of each of the clusters, fill it out with a few distinguishing words. Make sure you can glance at the table and understand what separates one cluster from another.

Cluster	Description
1	
2	
3	
4	
5	
6	
7	

We'll begin by looking at the mean for each variable of a cluster. Remember, this can help us identify variables that are not useful and get a general understanding of the characteristics of each cluster.

There may be a lot of variables, so we flipped the coordinates of the plot to make it easier to read. A bar to the right indicates a positive association and a bar to the left indicates a negative association.



Sift through the variables to see if any are unused throughout the clusters. If so, this indicates that the variable does not help differentiate the data into clusters. You can remove it here:

If you chose a large number of clusters, it may be difficult to use this visualization to remove unimportant variables. Instead, you should be able to see some of the important attributes of each of the clusters. Be thinking of identifiers for each cluster. Which variables are important throughout?

Let's begin to analyze the numeric values of the centers. Look through each cluster's characteristics. What sticks out to you?

Which clusters are scorers? Which are rebounders? Which have higher assist numbers? Higher 3-point shooting? Are any two clusters similar? What differentiates them?

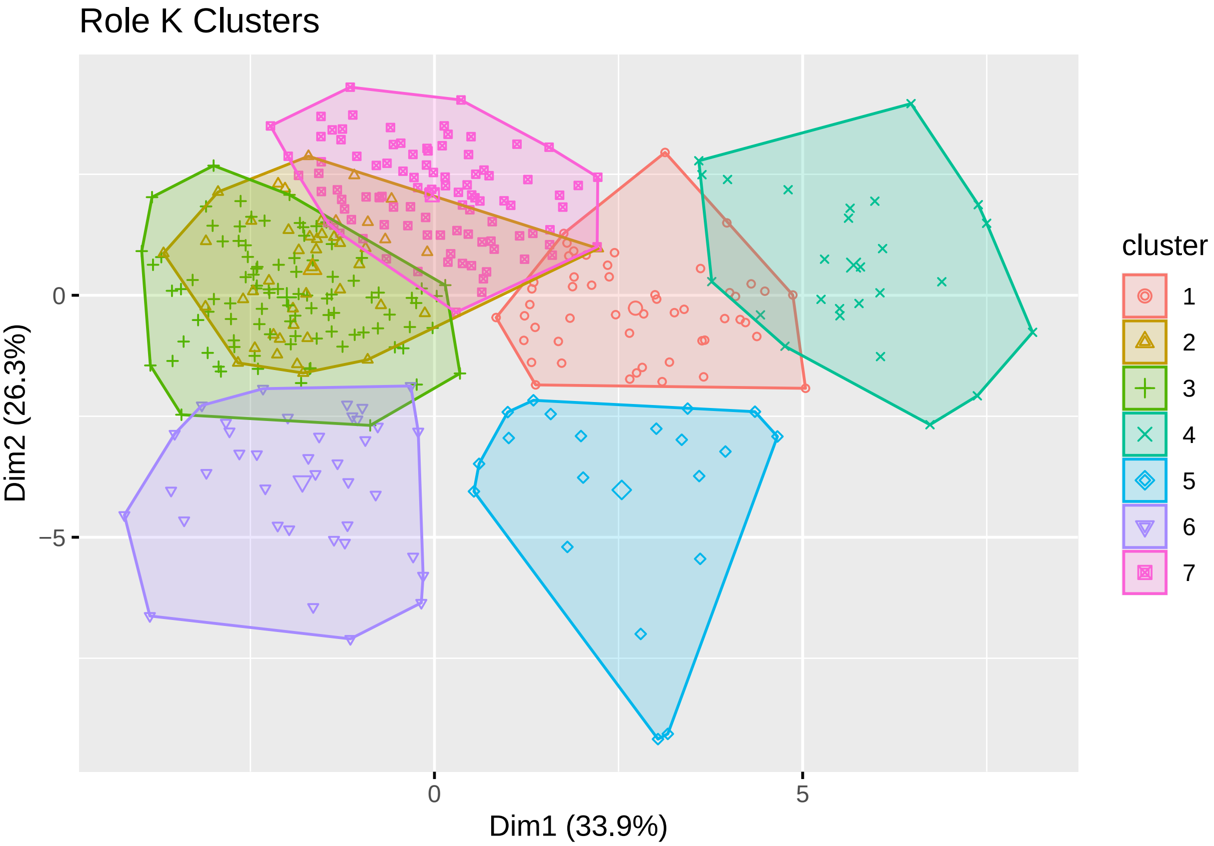
At this point, give a short descriptor of each cluster. Each cluster should be uniquely described.

Let's look at the size of each cluster.

Cluster	**Size**
1	51
2	45
3	96
4	26
5	20
6	38
7	98

Does this surprise you? Which clusters are large and small? Does this fit with your perception of the makeup of NBA teams?

Let's look at the distribution of the players.



What do you notice from the visualization? Remember, the dimensions cannot represent all the data, so we may have clusters that overlap. Imagine that there is a third dimension "Z" that explains another 30%-40% of the data.

Where are the cluster centers and outliers? Which clusters seem to be the closest together? Furthest away? Are any clusters more isolated than others? Is this supported by your previous analysis?

If you had to add another cluster where would it be? If you had to remove a cluster, where would it be?

Let's look at our prototype and outlier analysis.

First, we need to verify that our prototypes and outliers are prototypes and outliers. Now that we can change the number of clusters, its possible that you have some pretty small clusters. With a smaller sample size, we want to ensure that all our prototypes are indeed close to the cluster center and that all our outliers are indeed far away. In our K = 2 usage analysis, our prototypes were about 1-2.3 units away from the center. Our outliers were about 6-8.5. However, as K increases, the outlier distances should fall. Let's look at the distances from the center of our top 3 prototypes and outliers from each cluster to see how they compare.

Cluster	Category	Name	distance
1	Prototype	Trendon Watford	1.8523
1	Prototype	Isaiah Roby	1.8794
1	Prototype	John Collins	2.0773
1	Outlier	Isaiah Jackson	5.3644
1	Outlier	Tristan Thompson	6.8757
1	Outlier	Jakob Poeltl	7.1035
2	Prototype	Eric Bledsoe	1.6470
2	Prototype	Marcus Smart	1.6596
2	Prototype	Raul Neto	1.7083
2	Outlier	Josh Giddey	3.7481
2	Outlier	Jose Alvarado	3.8303
2	Outlier	Draymond Green	5.0848
3	Prototype	Coby White	1.3340
3	Prototype	Saddiq Bey	1.4581
3	Prototype	Lonnie Walker IV	1.4612
3	Outlier	Mike Muscala	4.0470
3	Outlier	Klay Thompson	4.1396
3	Outlier	Kevin Love	4.4379
4	Prototype	Ivica Zubac	1.9336
4	Prototype	Bismack Biyombo	1.9389
4	Prototype	Nic Claxton	2.3272
4	Outlier	Rudy Gobert	4.6189
4	Outlier	JaVale McGee	4.6530
4	Outlier	Thaddeus Young	5.0444
5	Prototype	Karl-Anthony Towns	2.0831
5	Prototype	Pascal Siakam	2.4719
5	Prototype	Jonas Valanciunas	2.5886
5	Outlier	Giannis Antetokounmpo	5.5937
5	Outlier	Joel Embiid	5.9766

Cluster	Category	Name	distance
5	Outlier	DeMarcus Cousins	6.0388
6	Prototype	Khris Middleton	1.5916
6	Prototype	Bradley Beal	1.6040
6	Prototype	Jaylen Brown	1.8622
6	Outlier	James Harden	4.0727
6	Outlier	Luka Doncic	4.0863
6	Outlier	Trae Young	4.1980
7	Prototype	Torrey Craig	1.3072
7	Prototype	Torrey Craig1	1.6834
7	Prototype	CJ Elleby	1.7221
7	Outlier	Xavier Tillman	4.2227
7	Outlier	Thaddeus Young1	4.2333
7	Outlier	Gary Payton II	5.0684

Which prototypes are the strongest prototypes? Which prototypes do you trust the most? Which are the strongest outliers? Would you disqualify any outliers or prototypes from the analysis (i.e. a supposed outlier is not far enough from the center or a labeled prototype is too far from the center).

Is this too long? I could remove the two long outliers table and only use the shorter one?

If you wish to disqualify a player from analysis, do it here:

Provide a space for the student to remove player's from the analysis. Assume student disqualifies Nic Claxton. Just for the heck of it.

Cluster	Size
1	51
2	45
3	96
4	26
5	20
6	38
7	98

Look again at the size of each cluster. Does this help explain any of your findings?

These outliers can be very different from each other. We'll need to look into them to see what kind of players they are. Once again, we'll show you the top 3 of each category first, and afterward a smaller table with only the top player.

Name	Cluster	Category	POS	Team	Height	Weight	PTSPerMin	ASTPerMin	TOP
John Collins	1	Prototype	PF	atl	81	235	0.526	0.058	0.

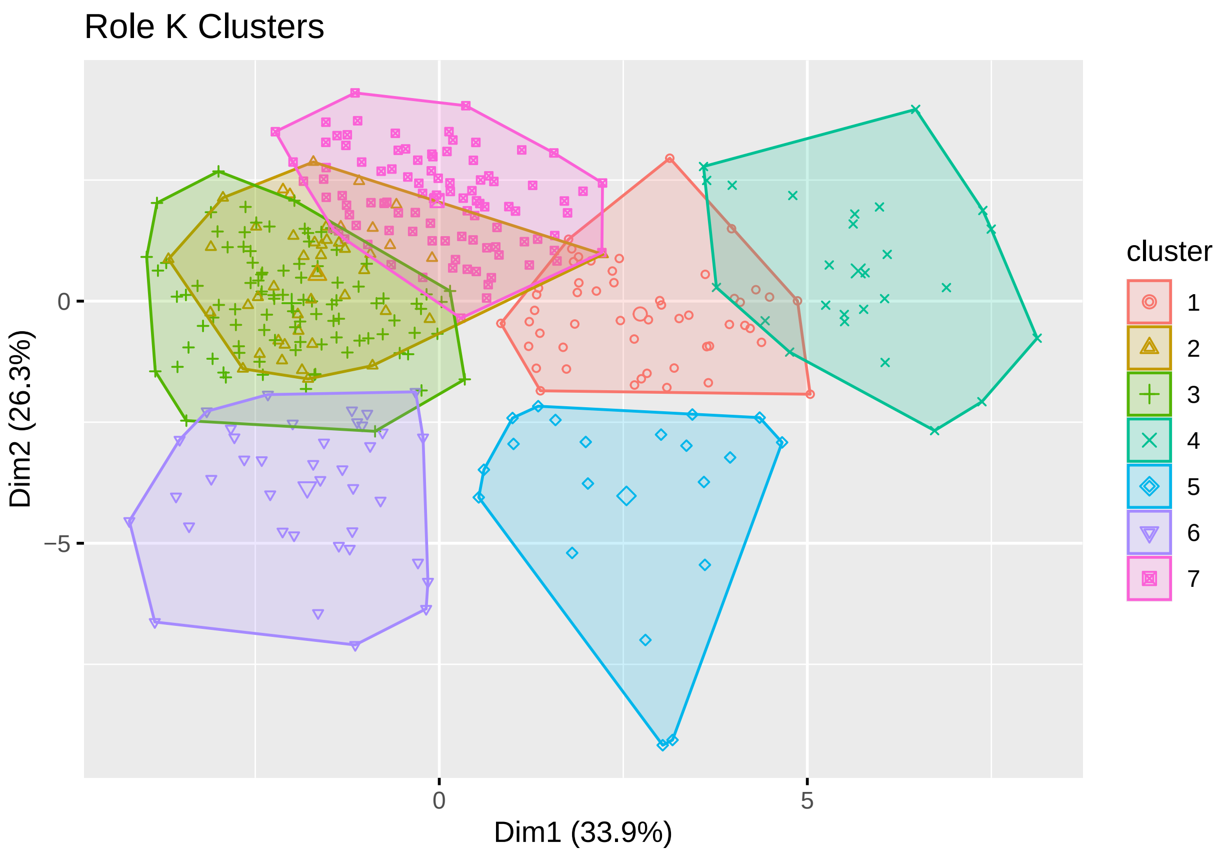
Name	Cluster	Category	POS	Team	Height	Weight	PTSPerMin	ASTPerMin	TOP
Isaiah Roby	1	Prototype	PF	okc	80	230	0.479	0.076	0.
Trendon Watford	1	Prototype	PF	por	81	240	0.420	0.094	0.
Isaiah Jackson	1	Outlier	F	ind	82	205	0.553	0.020	0.
Tristan Thompson	1	Outlier	C	sac	81	254	0.408	0.039	0.
Jakob Poeltl	1	Outlier	C	sa	85	245	0.466	0.097	0.
Marcus Smart	2	Prototype	PG	bos	75	220	0.375	0.183	0.
Eric Bledsoe	2	Prototype	SG	lac	73	214	0.393	0.167	0.
Raul Neto	2	Prototype	PG	wsh	73	180	0.383	0.158	0.
Draymond Green	2	Outlier	PF	gs	78	230	0.260	0.242	0.
Jose Alvarado	2	Outlier	PG	no	72	179	0.396	0.182	0.
Josh Giddey	2	Outlier	SG	okc	80	205	0.397	0.203	0.
Coby White	3	Prototype	PG	chi	77	195	0.462	0.105	0.
Saddiq Bey	3	Prototype	SF	det	79	215	0.488	0.085	0.
Lonnie Walker IV	3	Prototype	G	sa	76	204	0.526	0.096	0.
Kevin Love	3	Outlier	PF	cle	80	251	0.604	0.098	0.
Klay Thompson	3	Outlier	SG	gs	78	215	0.694	0.095	0.
Mike Muscala	3	Outlier	C	okc	82	240	0.580	0.036	0.
Ivica Zubac	4	Prototype	C	lac	84	240	0.422	0.066	0.
Bismack Biyombo	4	Prototype	C	phx	80	255	0.411	0.043	0.
JaVale McGee	4	Outlier	C	phx	84	270	0.582	0.038	0.
Thaddeus Young	4	Outlier	PF	sa	80	235	0.430	0.162	0.
Rudy Gobert	4	Outlier	C	utah	85	258	0.486	0.034	0.
Karl-Anthony Towns	5	Prototype	C	min	83	248	0.737	0.108	0.
Jonas Valanciunas	5	Prototype	C	no	83	265	0.587	0.086	0.
Pascal Siakam	5	Prototype	PF	tor	81	230	0.602	0.140	0.
DeMarcus Cousins	5	Outlier	C	den	82	270	0.640	0.122	0.
Giannis Antetokounmpo	5	Outlier	PF	mil	83	242	0.909	0.176	0.
Joel Embiid	5	Outlier	C	phi	84	280	0.905	0.124	0.
Jaylen Brown	6	Prototype	SG	bos	78	223	0.702	0.104	0.
Khris Middleton	6	Prototype	SF	mil	79	222	0.620	0.167	0.
Bradley Beal	6	Prototype	SG	wsh	75	207	0.644	0.183	0.
Trae Young	6	Outlier	PG	atl	73	180	0.814	0.278	0.
James Harden	6	Outlier	SG	bkn	77	220	0.608	0.276	0.
Luka Doncic	6	Outlier	PG	dal	79	230	0.802	0.246	0.

Name	Cluster	Category	POS	Team	Height	Weight	PTSPerMin	ASTPerMin	TOP
Torrey Craig	7	Prototype	SF	ind	79	221	0.320	0.054	0.
Torrey Craig1	7	Prototype	SF	phx	79	221	0.332	0.058	0.
CJ Elleby	7	Prototype	SG	por	78	200	0.287	0.074	0.
Gary Payton II	7	Outlier	SG	gs	75	195	0.403	0.051	0.
Xavier Tillman	7	Outlier	C	mem	80	245	0.364	0.091	0.
Thaddeus Young1	7	Outlier	PF	tor	80	235	0.344	0.093	0.

Below is the smaller table.

Name	Cluster	Category	POS	Team	Height	Weight	PTSPerMin	ASTPerMin	TOP
Trendon Watford	1	Prototype	PF	por	81	240	0.420	0.094	0.
Eric Bledsoe	2	Prototype	SG	lac	73	214	0.393	0.167	0.
Coby White	3	Prototype	PG	chi	77	195	0.462	0.105	0.
Ivica Zubac	4	Prototype	C	lac	84	240	0.422	0.066	0.
Karl-Anthony Towns	5	Prototype	C	min	83	248	0.737	0.108	0.
Khris Middleton	6	Prototype	SF	mil	79	222	0.620	0.167	0.
Torrey Craig	7	Prototype	SF	ind	79	221	0.320	0.054	0.
Jakob Poeltl	1	Outlier	C	sa	85	245	0.466	0.097	0.
Draymond Green	2	Outlier	PF	gs	78	230	0.260	0.242	0.
Kevin Love	3	Outlier	PF	cle	80	251	0.604	0.098	0.
Thaddeus Young	4	Outlier	PF	sa	80	235	0.430	0.162	0.
DeMarcus Cousins	5	Outlier	C	den	82	270	0.640	0.122	0.
Trae Young	6	Outlier	PG	atl	73	180	0.814	0.278	0.
Gary Payton II	7	Outlier	SG	gs	75	195	0.403	0.051	0.

Look through the prototypes and outliers. Compare their results with your previous findings. Do the prototypes of each cluster match up with your summary of the cluster? How do the outliers fit in? Two outliers can be very different. Pick a few outliers and determine their closest two clusters.



Analyze the K = 7 clusters as a whole. Are the clusters good? Do they have high intra-class similarity? What about a low intra-class similarity? If you were to do the analysis again, would you choose the same amount of clusters?

Compare lots of Ks

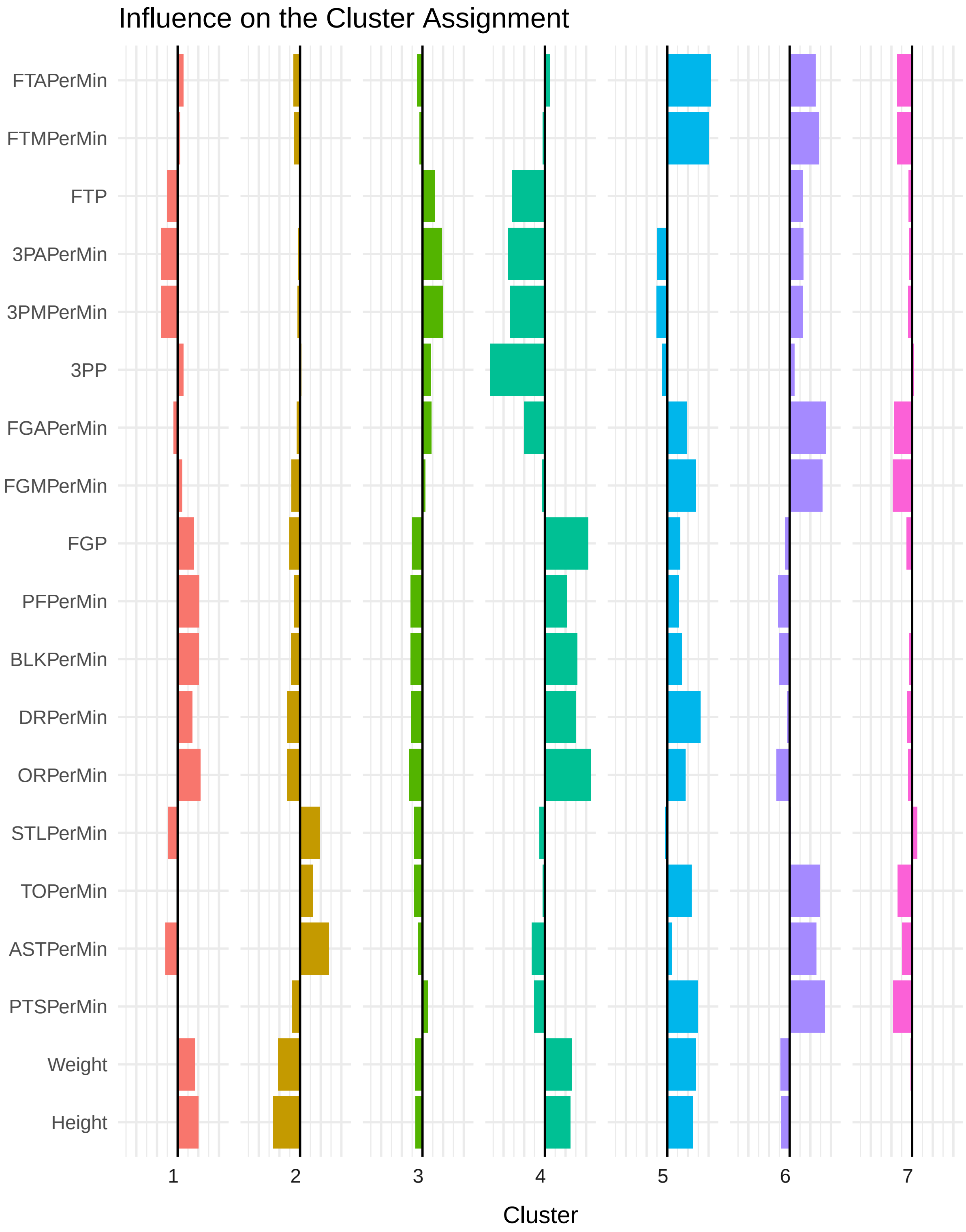
Select two values of K (between 2 and 10) to compare. This table can become very complex. Remember, the rows are the cluster assignment with the first value of K and the columns are the cluster assignment with the second value. Isolate and analyze one row or column at a time.

Cluster	1	2	3	4	5	6	7
1	0	9	17	0	8	38	0
2	1	35	79	0	0	0	90
3	50	1	0	26	12	0	8

Part 3: GM of Dallas Mavericks

Returning back to the Dallas Mavericks. Let's take a look at how the Mavericks players were clustered in our role dataset. Let's use K = 7. If you did not analyze K = 7 earlier, it is worth a look.

Below are a few visual reminders of each cluster's characteristics.



Before moving on, fill out this table to describe each cluster. Write a few descriptive words that distinguish each cluster. This will help you to organize your thoughts on each cluster. If you already completed this for K = 7 in the role dataset, then you are free to proceed.

Cluster	**Description**
1	
2	
3	
4	
5	
6	
7	

Caleb's estimation of 7 clusters. I'd like to provide them a blank table to fill out somehow. Like a text file table with two columns.

Cluster	**Description**
1	big men, mediocre scorers, kinda shoot deep
2	small point guards, facilitaters
3	meh players, 3 point shooters
4	big men, can't shoot deep at all
5	high-volume players, generally tall
6	high-volume players, average height
7	low production, very mediocre, likely corner 3 players

Mavericks Offseason Analysis

Now, let's look at the cluster assignments of our ten Dallas Mavericks players.

Name	Cluster	POS	MIN	Height	Weight	FGP	3PP	FTP	PTSPer	ORPerMi
Dwight Powell	1	C	21.9	82	240	0.671	0.351	0.783	0.3973	0.0959
Jalen Brunson	2	PG	31.9	73	190	0.502	0.373	0.840	0.5110	0.0157
Tim Hardaway Jr.	3	SF	29.6	77	205	0.394	0.336	0.757	0.4797	0.0101
Kristaps Porzingis	5	C	29.5	87	240	0.451	0.283	0.865	0.6508	0.0644
Luka Doncic	6	PG	35.4	79	230	0.457	0.353	0.744	0.8023	0.0254
Dorian Finney-Smith	7	PF	33.1	79	220	0.471	0.395	0.675	0.3323	0.0453
Reggie Bullock	7	SF	28.0	78	205	0.401	0.360	0.833	0.3071	0.0179
Maxi Kleber	7	PF	24.6	82	240	0.398	0.325	0.708	0.2846	0.0488
Josh Green	7	SG	15.5	77	200	0.508	0.359	0.689	0.3097	0.0516
Sterling Brown	7	SF	12.8	77	219	0.381	0.304	0.933	0.2578	0.0391

What do you notice about the player assignments? How many clusters do the Mavericks have represented? Which cluster is the most common on the Mavericks team?

Why is cluster 7 the most common? What kind of player is in cluster 7?

The Mavericks experienced a bit of turnover in the 2022 offseason. They'd already traded away C Kristaps Porzingis for SG Spencer Dinwiddie at the end of the 2022 season, and they lost productive SG Jalen Brunson to free agency. They traded away SF Sterling Brown and other assets for C Christian Wood during the 2022 Summer.

Let's assess the offseason moves of the Dallas Mavericks by looking at the opening day roster for 2023 and its cluster distribution. Below are the eleven players on the Dallas Mavericks roster at Game 1 of the 2023 season, a loss against the Phoenix Suns.

Name	Cluster	POS	MIN	Height	Weight	FGP	3PP	FTP	PTSPer	ORPerMi
Dwight Powell	1	C	21.9	82	240	0.671	0.351	0.783	0.3973	0.0959
Tim Hardaway Jr.	3	SF	29.6	77	205	0.394	0.336	0.757	0.4797	0.0101
Spencer Dinwiddie	3	PG	30.2	77	215	0.376	0.310	0.811	0.4172	0.0265
Davis Bertans	3	SF	14.7	82	225	0.351	0.319	0.933	0.3878	0.0136
JaVale McGee	4	C	15.8	84	270	0.629	0.222	0.699	0.5823	0.1392
Christian Wood	5	C	30.8	82	214	0.501	0.390	0.623	0.5812	0.0519
Luka Doncic	6	PG	35.4	79	230	0.457	0.353	0.744	0.8023	0.0254
Dorian Finney-Smith	7	PF	33.1	79	220	0.471	0.395	0.675	0.3323	0.0453
Reggie Bullock	7	SF	28.0	78	205	0.401	0.360	0.833	0.3071	0.0179
Maxi Kleber	7	PF	24.6	82	240	0.398	0.325	0.708	0.2846	0.0488
Josh Green	7	SG	15.5	77	200	0.508	0.359	0.689	0.3097	0.0516

The roster looks somewhat similar, but what classification of player did the Mavericks lose in the 2022 season and not return in the 2023 season? What classification of player did the Mavericks gain in the 2023 season?

Answer: They lost a cluster 2 player, lost a cluster 7 player, gained two cluster 3 players, and a cluster 4 player.

What kind of player is in cluster 2? What would losing this kind of player do to a team?

Dallas Mavericks Trade

Let's say you're the GM of the Dallas Mavericks after game 1 of the 2022-2023 season. Which players would you consider trading and what cluster of player would you hope to acquire? Which players are you willing to give up?

Answer: I think the correct answer here is give up any of cluster 3 or 7 for a cluster 2. Maxi Kleber is the most expendable because he has some features of 1,4,5 and some of 7. And they have excess of these players.

Select four players you are willing to trade and one cluster that you are looking for.

Name	Cluster	POS	MIN	Team	Height	Weight	FGP	3PP
Lou Williams	2	SG	14.3	atl	73	175	0.391	0.363
Dennis Schroder	2	PG	29.2	bos	75	172	0.440	0.349
Marcus Smart	2	PG	32.3	bos	75	220	0.418	0.331
Ish Smith	2	PG	13.8	cha	72	175	0.395	0.400
Lonzo Ball	2	PG	34.6	chi	78	190	0.423	0.423
Alex Caruso	2	SG	28.0	chi	76	186	0.398	0.333
Ricky Rubio	2	PG	28.5	cle	75	190	0.363	0.339
Brandon Goodwin	2	G	13.9	cle	72	180	0.416	0.345
Jalen Brunson	2	PG	31.9	dal	73	190	0.502	0.373
Facundo Campazzo	2	PG	18.2	den	70	195	0.361	0.301
Cory Joseph	2	PG	24.6	det	75	200	0.445	0.414
Killian Hayes	2	PG	25.0	det	77	195	0.383	0.263
Saben Lee	2	PG	16.3	det	74	183	0.390	0.233
Draymond Green	2	PF	28.9	gs	78	230	0.525	0.296
Kevin Porter Jr.	2	SG	31.3	hou	76	203	0.415	0.375
Josh Christopher	2	SG	18.0	hou	77	215	0.448	0.296
D.J. Augustin	2	G	15.0	hou	71	183	0.404	0.406
Tyrese Haliburton	2	PG	36.1	ind	77	185	0.502	0.416
T.J. McConnell	2	PG	24.1	ind	73	190	0.481	0.303
Keifer Sykes	2	G	17.7	ind	71	167	0.363	0.300
Eric Bledsoe	2	SG	25.2	lac	73	214	0.421	0.313
De'Anthony Melton	2	SG	22.7	mem	74	200	0.404	0.374
Tyus Jones	2	PG	21.2	mem	72	196	0.451	0.390
Kyle Lowry	2	PG	33.9	mia	72	196	0.440	0.377
Gabe Vincent	2	PG	23.4	mia	75	200	0.417	0.368
Jrue Holiday	2	PG	32.9	mil	75	205	0.501	0.411
Patrick Beverley	2	PG	25.4	min	73	180	0.406	0.343
Jordan McLaughlin	2	PG	14.5	min	71	185	0.440	0.318

Name	Cluster	POS	MIN	Team	Height	Weight	FGP	3PP
Jose Alvarado	2	PG	15.4	no	72	179	0.446	0.291
Josh Giddey	2	SG	31.5	okc	80	205	0.419	0.263
Theo Maledon	2	PG	17.8	okc	76	175	0.375	0.293
Jalen Suggs	2	SG	27.2	orl	76	205	0.361	0.214
R.J. Hampton	2	PG	21.9	orl	76	175	0.383	0.350
Chris Paul	2	PG	32.9	phx	72	175	0.493	0.317
Cameron Payne	2	PG	22.0	phx	73	183	0.409	0.336
Dennis Smith Jr.	2	PG	17.3	por	74	205	0.418	0.222
Tyrese Haliburton1	2	PG	34.5	sac	77	185	0.457	0.413
Davion Mitchell	2	PG	27.7	sac	74	205	0.418	0.316
Derrick White1	2	PG	30.3	sa	76	190	0.426	0.314
Tre Jones	2	PG	16.6	sa	73	185	0.490	0.196
Malachi Flynn	2	PG	12.2	tor	73	175	0.393	0.333
Mike Conley	2	PG	28.6	utah	73	175	0.435	0.408
Ish Smith1	2	PG	22.0	wsh	72	175	0.457	0.357
Raul Neto	2	PG	19.6	wsh	73	180	0.463	0.292
Aaron Holiday	2	G	16.2	wsh	72	185	0.467	0.343

From the list, choose a player you like from a team that has several of these types of players. They'd be more likely to part ways. Assess the strengths of the pertinent players and propose a trade! How does it look?

Feel free to make the trades as complex as you wish, but try to choose something that the opposing team would agree to.

Defend your proposed trade using the cluster information. You may add in some basketball knowledge if you like.

What do you think of this process? What are the strengths and weaknesses of evaluating a team based on cluster membership?