Clustering

* taken from Ron Yurko's slides

Q. Cluster NBA players into offensive Roles or Archetypes based on How they play.

* First, need variables that describe How they play.

* Todd Whitehead: try using Synergy Play Type data!

```
[1] "FGA_freq_Spotup"          "FGA_freq_PRRollMan"           "FGA_freq_Postup"
[4] "FGA_freq_Cut"             "FGA_freq_OffRebound"          "FGA_freq_PRBallHandler"
[7] "FGA_freq_Isolation"       "FGA_freq_OffScreenOrHandoff"
```

* Then, with this data, how to cluster players??

What is **unsupervised learning?**

We have p variables for n observations $x_1, \ldots, x_n$, and for observation i:

$$x_{i1}, x_{i2}, \ldots, x_{ip}$$

- *unsupervised*: none of the variables are **response** variables, i.e., there are no labeled data

Think of unsupervised learning as **an extension of EDA...**

- $\Rightarrow$ **there is no unique right answer!**

What is clustering (aka cluster analysis)?

ISLR 10.3:

> *very broad set of techniques for finding subgroups, or clusters, in a dataset*

- observations **within** clusters are **more similar** to each other,
- observations **in different** clusters are **more different** from each other

How do we define **distance / dissimilarity** between observations?

- e.g. **Euclidean distance** between observations i and j

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Units matter!

- one variable may *dominate* others when computing Euclidean distance because its range is much larger
- can standardize each variable / column of dataset to have mean 0 and standard divation 1 with `scale()`
- **but we may value the separation in that variable!** (so just be careful...)

What's the clustering objective?

- $C_1, \ldots, C_K$ are *sets* containing indices of observations in each of the K clusters
  - if observation i is in cluster k, then $i \in C_k$
- We want to minimize the **within-cluster variation** $W(C_k)$ for each cluster C_k and solve:

$$\underset{C_1,\ldots,C_K}{\text{minimize}}\Big\{\sum_{k=1}^{K} W(C_k)\Big\}$$

- Can define using the **squared Euclidean distance** ($|C_k| = n_k =$ # observations in cluster k)

$$W(C_k) = \frac{1}{|C_k|}\sum_{i,j\in C_k} d(x_i, x_j)^2$$

- Commonly referred to as the within-cluster sum of squares (WSS)

Lloyd's algorithm

1) Choose K random centers, aka **centroids**

2) Assign each observation closest center (using Euclidean distance)

3) Repeat until cluster assignment stop changing:

- Compute new centroids as the averages of the updated groups
- Reassign each observations to closest center

Converges to a local optimum, not the global

Results will change from run to run (set the seed!)

Takes K as an input!

things to check when clustering

- do the units of the variables make sense?
- Should you standardize the variables?
- is "nstart", the number of starting random configurations, large enough?

So, how do we choose the number of clusters?!

There is no universally accepted way to conclude that a particular choice of K is optimal!

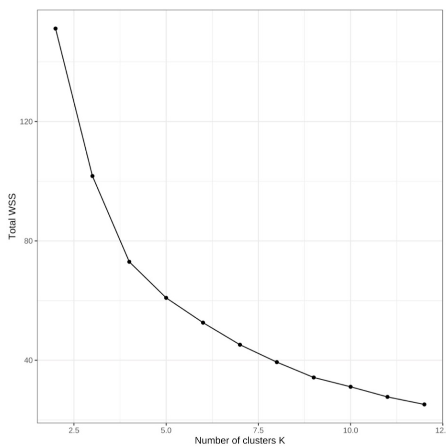
Popular heuristic: elbow plot (use with caution)

Choose K where marginal improvements is low at the bend (hence the elbow)

This is just a guideline and should not dictate your choice of K!

Gap statistic is a popular choice (see `clusGap` function in `cluster` package)

Next Tuesday: model-based approach to choosing the number of clusters!



Better alternative to `nstart`: K-means++

Pick a random observation to be the center c_1 of the first cluster C_1

- This initializes a set $Centers = \{c_1\}$

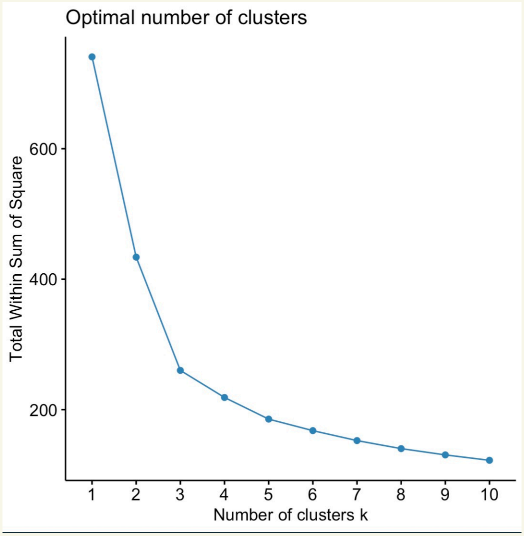
Then for each remaining cluster $c^* \in 2, \dots, K$:

- For each observation (that is not a center), compute $D(x_i) = \min_{c \in Centers} d(x_i, c)$
  - Distance between observation and its closest center $c \in Centers$
- Randomly pick a point x_i with probability: $p_i = \frac{D^2(x_i)}{\sum_{j=1}^n D^2(x_j)}$
- As distance to closest center increases $\Rightarrow$ probability of selection increases
- Call this randomly selected observation c^*, update $Centers = Centers \cup c^*$
  - Same as `centers = c(centers, c_new)`

Then run K-means using these $Centers$ as the starting points

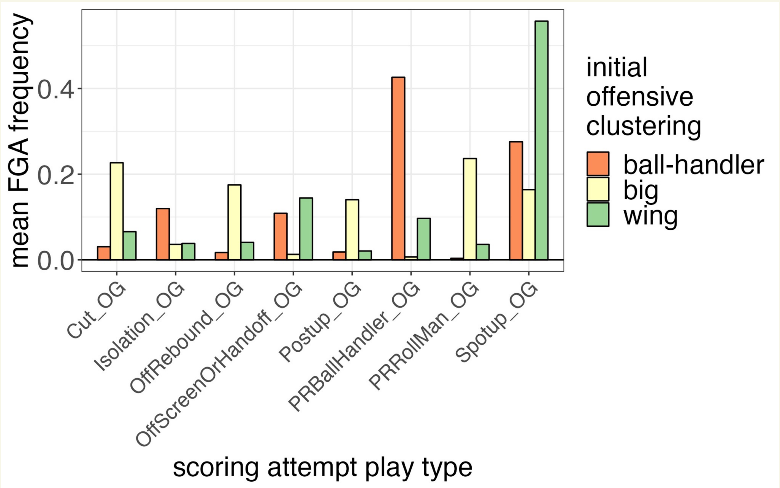
Time to Cluster NBA players into offensive Roles by Synergy FGA play type frequencies

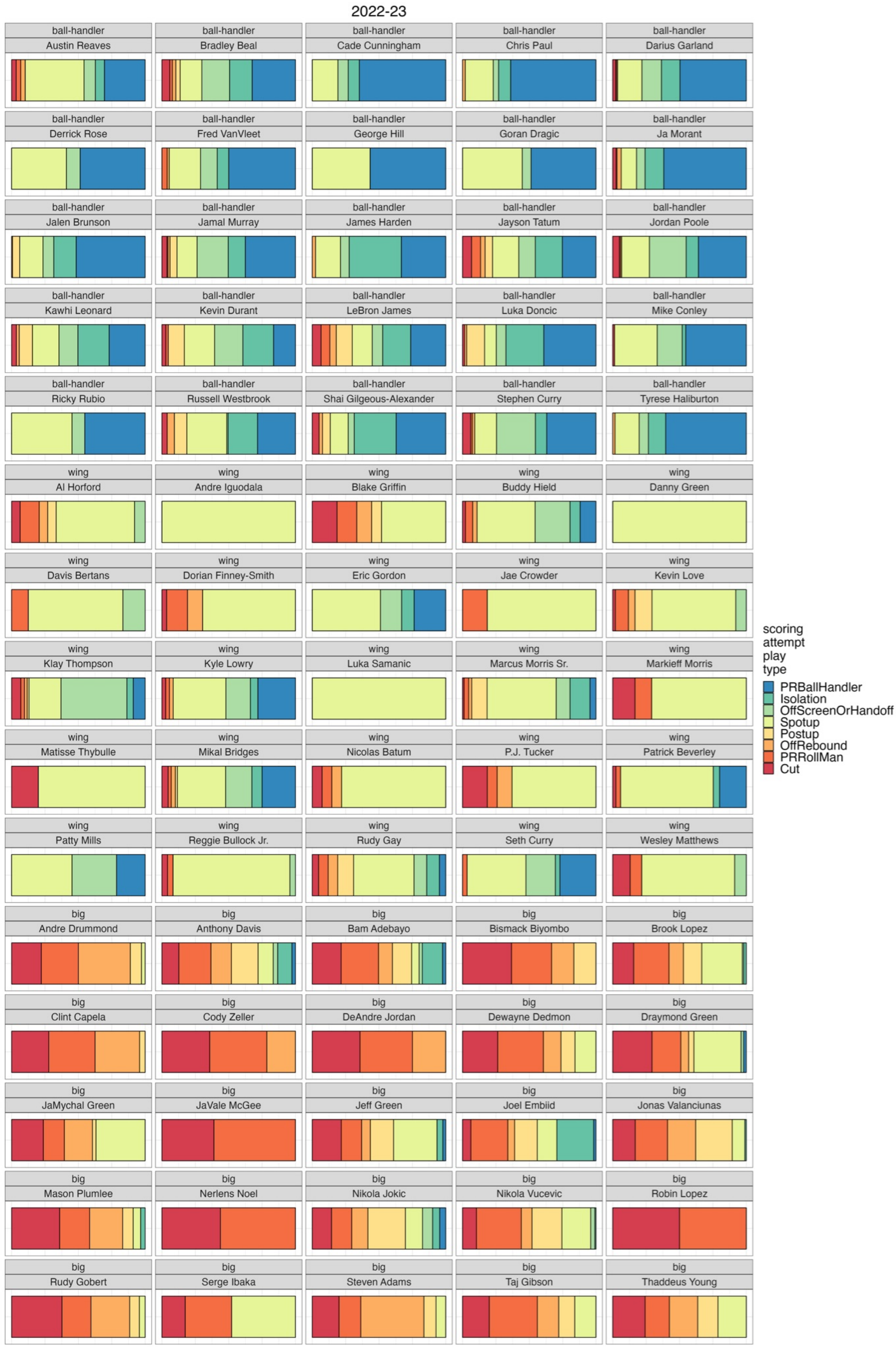
```
### Hard Clustering: K means
set.seed(387397)
km1 <- flexclust::kcca(mat_train, K1)
print(km1)
```



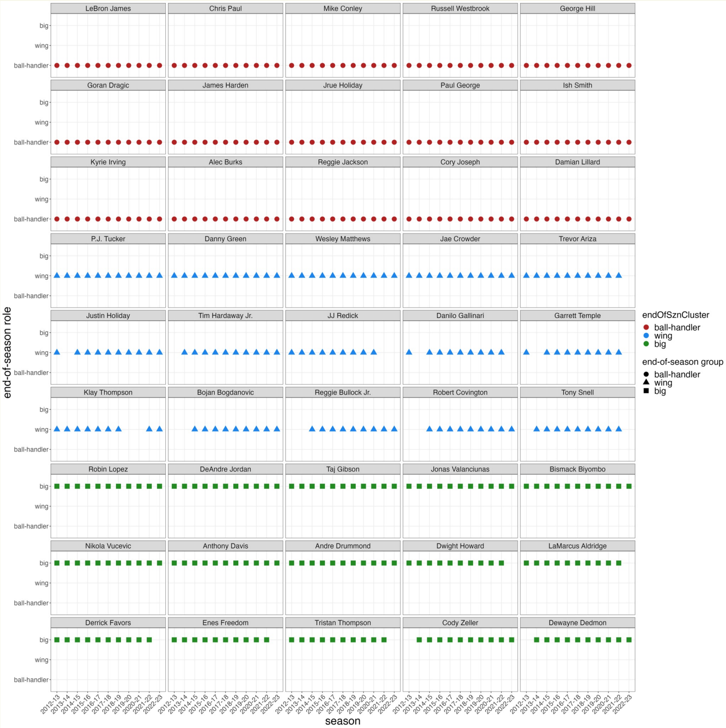
Elbow Plot

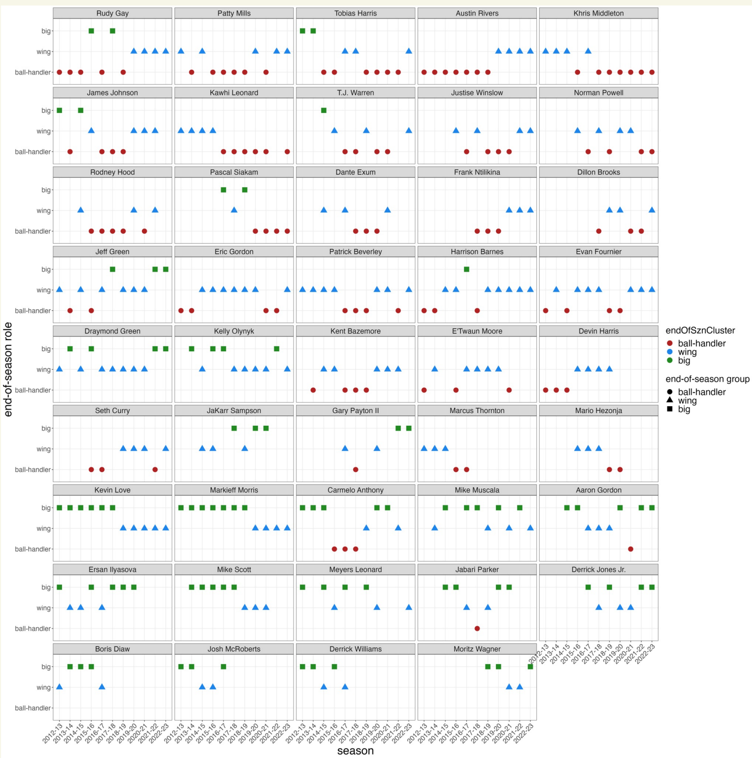
Began with K=3 clusters



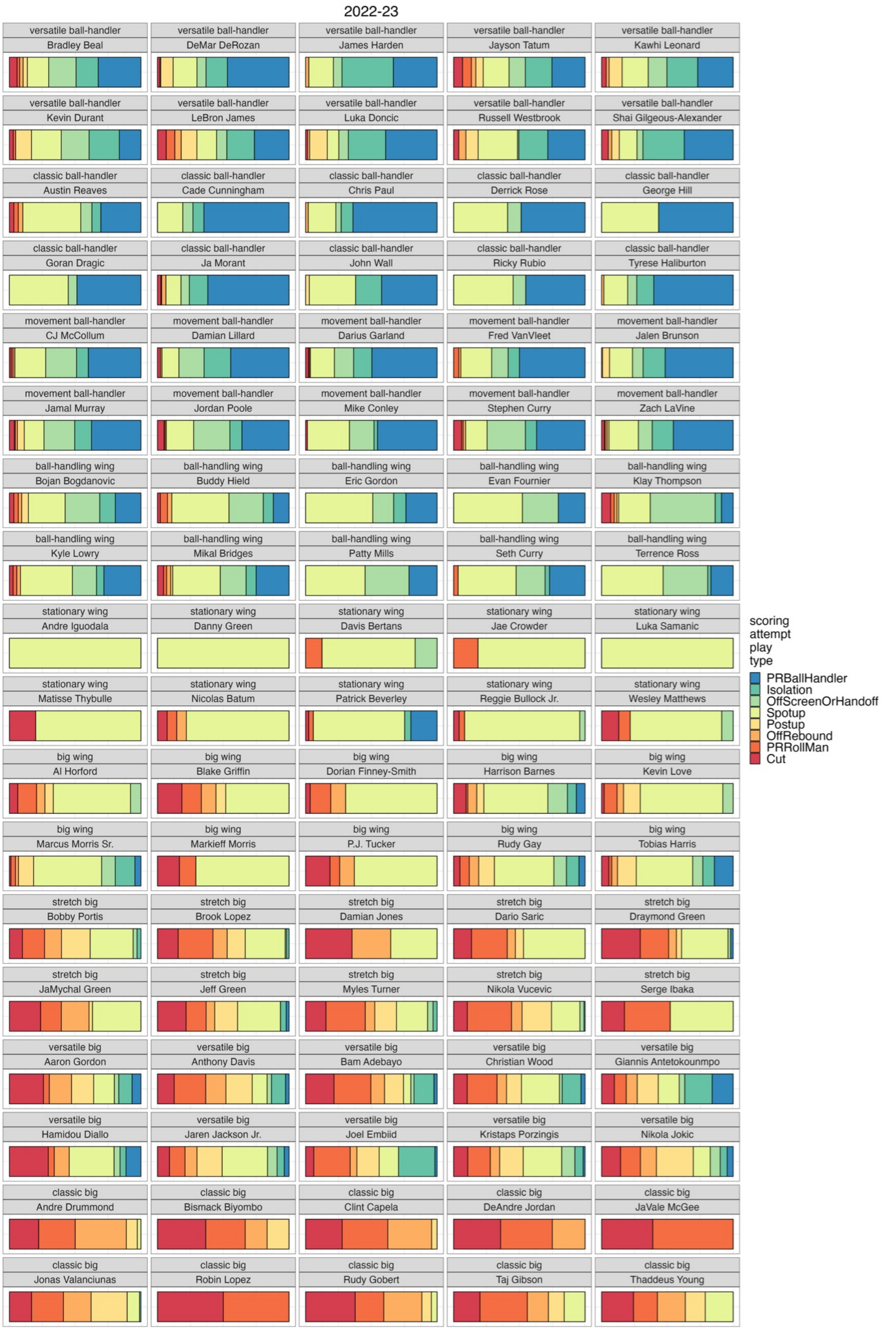


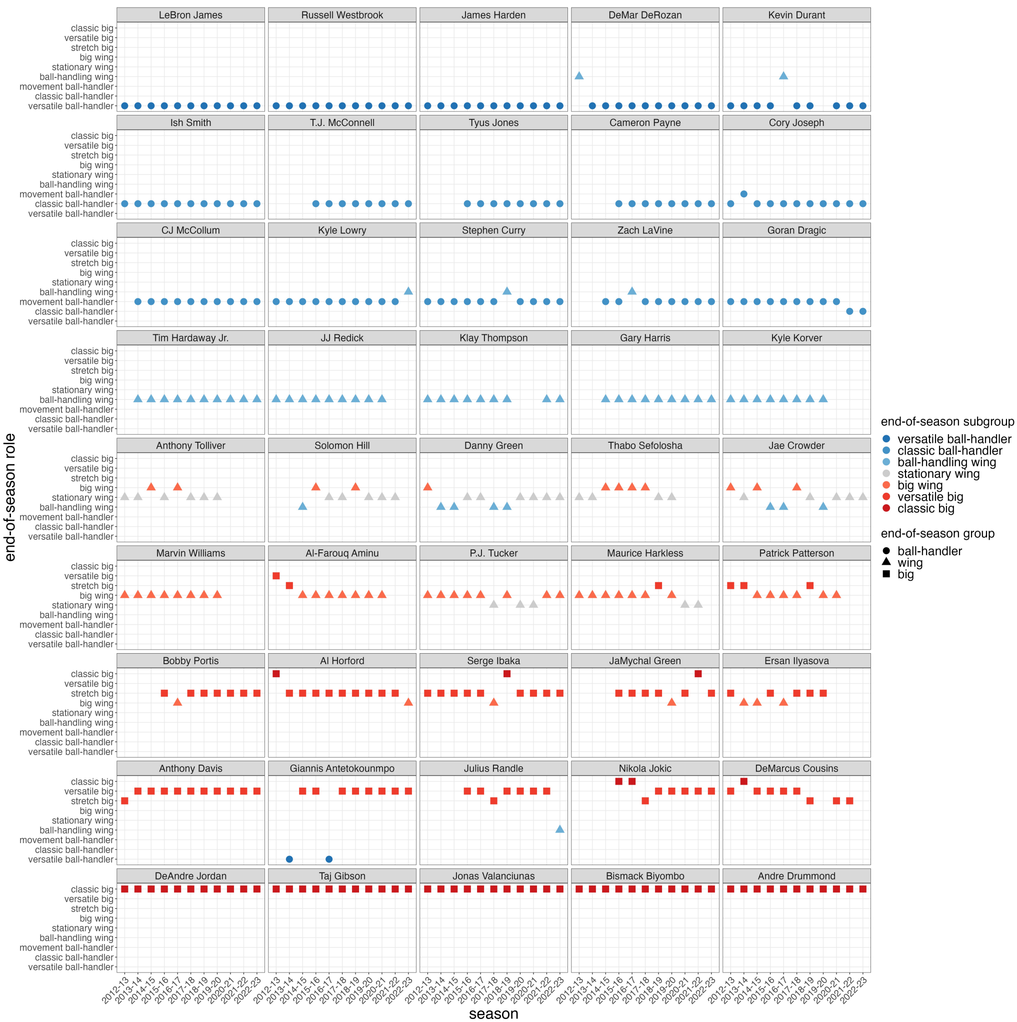
2022-23
ball-handler
Austin Reaves
Bradley Beal
Cade Cunningham
Chris Paul
Darius Garland
Derrick Rose
Fred VanVleet
George Hill
Goran Dragic
Ja Morant
Jalen Brunson
Jamal Murray
James Harden
Jayson Tatum
Jordan Poole
Kawhi Leonard
Kevin Durant
LeBron James
Luka Doncic
Mike Conley
Ricky Rubio
Russell Westbrook
Shai Gilgeous-Alexander
Stephen Curry
Tyrese Haliburton
wing
Al Horford
Andre Iguodala
Blake Griffin
Buddy Hield
Danny Green
Davis Bertans
Dorian Finney-Smith
Eric Gordon
Jae Crowder
Kevin Love
Klay Thompson
Kyle Lowry
Luka Samanic
Marcus Morris Sr.
Markieff Morris
Matisse Thybulle
Mikal Bridges
Nicolas Batum
P.J. Tucker
Patrick Beverley
Patty Mills
Reggie Bullock Jr.
Rudy Gay
Seth Curry
Wesley Matthews
big
Andre Drummond
Anthony Davis
Bam Adebayo
Bismack Biyombo
Brook Lopez
Clint Capela
Cody Zeller
DeAndre Jordan
Dewayne Dedmon
Draymond Green
JaMychal Green
JaVale McGee
Jeff Green
Joel Embiid
Jonas Valanciunas
Mason Plumlee
Nerlens Noel
Nikola Jokic
Nikola Vucevic
Robin Lopez
Rudy Gobert
Serge Ibaka
Steven Adams
Taj Gibson
Thaddeus Young
scoring attempt play type
PRBallHandler
Isolation
OffScreenOrHandoff
Spotup
Postup
OffRebound
PRRollMan
Cut


LeBron James
Chris Paul
Mike Conley
Russell Westbrook
George Hill
Goran Dragic
James Harden
Jrue Holiday
Paul George
Ish Smith
Kyrie Irving
Alec Burks
Reggie Jackson
Cory Joseph
Damian Lillard
P.J. Tucker
Danny Green
Wesley Matthews
Jae Crowder
Trevor Ariza
Justin Holiday
Tim Hardaway Jr.
JJ Redick
Danilo Gallinari
Garrett Temple
Klay Thompson
Bojan Bogdanovic
Reggie Bullock Jr.
Robert Covington
Tony Snell
Robin Lopez
DeAndre Jordan
Taj Gibson
Jonas Valanciunas
Bismack Biyombo
Nikola Vucevic
Anthony Davis
Andre Drummond
Dwight Howard
LaMarcus Aldridge
Derrick Favors
Enes Freedom
Tristan Thompson
Cody Zeller
Dewayne Dedmon
big
wing
ball-handler
end-of-season role
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22
2022-23
season
endOfSznCluster
ball-handler
wing
big
end-of-season group
ball-handler
wing
big

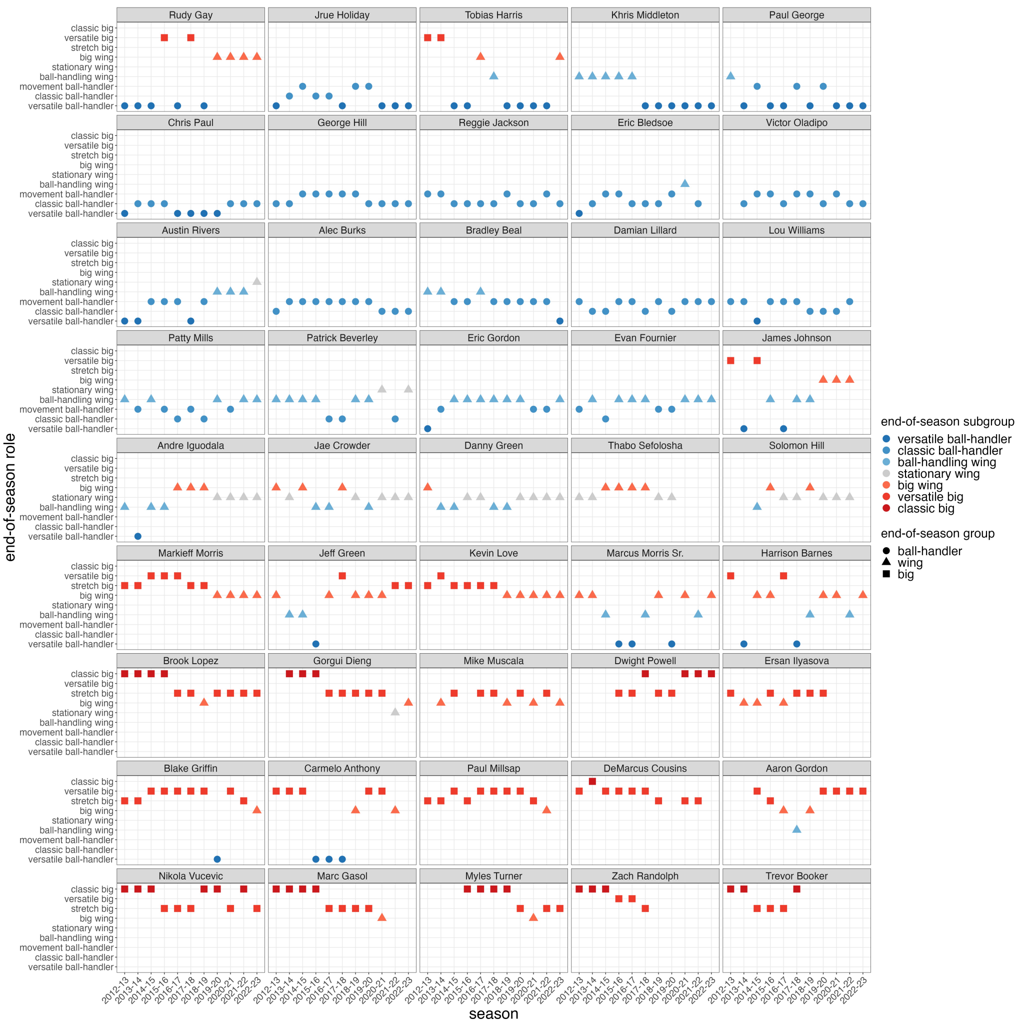

Rudy Gay
Patty Mills
Tobias Harris
Austin Rivers
Khris Middleton
James Johnson
Kawhi Leonard
T.J. Warren
Justise Winslow
Norman Powell
Rodney Hood
Pascal Siakam
Dante Exum
Frank Ntilikina
Dillon Brooks
Jeff Green
Eric Gordon
Patrick Beverley
Harrison Barnes
Evan Fournier
Draymond Green
Kelly Olynyk
Kent Bazemore
E'Twaun Moore
Devin Harris
Seth Curry
JaKarr Sampson
Gary Payton II
Marcus Thornton
Mario Hezonja
Kevin Love
Markieff Morris
Carmelo Anthony
Mike Muscala
Aaron Gordon
Ersan Ilyasova
Mike Scott
Meyers Leonard
Jabari Parker
Derrick Jones Jr.
Boris Diaw
Josh McRoberts
Derrick Williams
Moritz Wagner
big
wing
ball-handler
end-of-season role
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22
2022-23
season
endOfSznCluster
ball-handler
wing
big
end-of-season group
ball-handler
wing
big

I further clustered the 3 groups into more refined Roles:





LeBron James
Russell Westbrook
James Harden
DeMar DeRozan
Kevin Durant
Ish Smith
T.J. McConnell
Tyus Jones
Cameron Payne
Cory Joseph
CJ McCollum
Kyle Lowry
Stephen Curry
Zach LaVine
Goran Dragic
Tim Hardaway Jr.
JJ Redick
Klay Thompson
Gary Harris
Kyle Korver
Anthony Tolliver
Solomon Hill
Danny Green
Thabo Sefolosha
Jae Crowder
Marvin Williams
Al-Farouq Aminu
P.J. Tucker
Maurice Harkless
Patrick Patterson
Bobby Portis
Al Horford
Serge Ibaka
JaMychal Green
Ersan Ilyasova
Anthony Davis
Giannis Antetokounmpo
Julius Randle
Nikola Jokic
DeMarcus Cousins
DeAndre Jordan
Taj Gibson
Jonas Valanciunas
Bismack Biyombo
Andre Drummond
end-of-season role
classic big
versatile big
stretch big
big wing
stationary wing
ball-handling wing
movement ball-handler
classic ball-handler
versatile ball-handler
season
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22
2022-23
end-of-season subgroup
versatile ball-handler
classic ball-handler
ball-handling wing
stationary wing
big wing
versatile big
classic big
end-of-season group
ball-handler
wing
big


Rudy Gay
Jrue Holiday
Tobias Harris
Khris Middleton
Paul George
Chris Paul
George Hill
Reggie Jackson
Eric Bledsoe
Victor Oladipo
Austin Rivers
Alec Burks
Bradley Beal
Damian Lillard
Lou Williams
Patty Mills
Patrick Beverley
Eric Gordon
Evan Fournier
James Johnson
Andre Iguodala
Jae Crowder
Danny Green
Thabo Sefolosha
Solomon Hill
Markieff Morris
Jeff Green
Kevin Love
Marcus Morris Sr.
Harrison Barnes
Brook Lopez
Gorgui Dieng
Mike Muscala
Dwight Powell
Ersan Ilyasova
Blake Griffin
Carmelo Anthony
Paul Millsap
DeMarcus Cousins
Aaron Gordon
Nikola Vucevic
Marc Gasol
Myles Turner
Zach Randolph
Trevor Booker
end-of-season role
classic big
versatile big
stretch big
big wing
stationary wing
ball-handling wing
movement ball-handler
classic ball-handler
versatile ball-handler
season
2012-13
2013-14
2014-15
2015-16
2016-17
2017-18
2018-19
2019-20
2020-21
2021-22
2022-23
end-of-season subgroup
versatile ball-handler
classic ball-handler
ball-handling wing
stationary wing
big wing
versatile big
classic big
end-of-season group
ball-handler
wing
big