Clustering NBA Players

BY MITCHELL HERSEY

What are the standard positions in basketball?

- Point Guard
  - Runs the offense
  - Smallest player on the court
  - High assists, low rebounds
  - John Stockton, Chris Paul
- Shooting Guard
  - Scores the ball
  - Guards the best opposing guard
  - Michel Jordan, James Harden
- Small Forward
  - Versatility
  - Many different styles of player
  - Lebron James, Kevin Durant
- Power Forward
  - Strong and durable
  - Often can provide shooting range despite their heights
  - Tim Duncan, Kevin Garnett
- Center
  - Tallest player on the court
  - Rebounds, Dunks, and Alley Oops
  - Shaquille O'Neal

Is it really true though?

- Does a player's statistics actually support the existence of these five distinct positions?
- Can you determine a player's position strictly off of their box score?

	Totals								
WS	G	GS	MP	FG	FGA	2P	2PA	3P	3PA
4.6	73	59	31.1	4.2	7.8	4.0	7.2	0.2	0.6

		Shooting					
PF	PTS	FG%	2P%	3P%	eFG%	FT%	TS%
2.4	9.7	.538	.555	.333	.551	.683	.570

…sometimes

Clustering

- There have been 24,690 individual player seasons from 1950 to 2017
- Can this data create groups that resemble the traditional basketball positions?

- What attributes should be used?
- How many groups should there be?

What data is good data?

- Threshold of 400+ minutes (~5 minutes per game)
  - Allows for a reasonable sample of the player's on-court ability
  - Minutes first recorded in 1952
- Counting stats normalized to per 36
  - Project a player's contribution to a starting player's workload
  - Standard practice in the statistics community
  - Rebounds, Steals, Blocks not recording until 1974
  - Turnovers recorded in 1978
- Normalize stats again between 0 and 1

Attributes

1952 – 2017

- Points
- Assists
- True Shooting %
- Field Goal Attempts
- Field Goal %
- Three Point Attempts
- Three Point %
- Free Throw Attempts
- Free Throw %

1978 – 2017

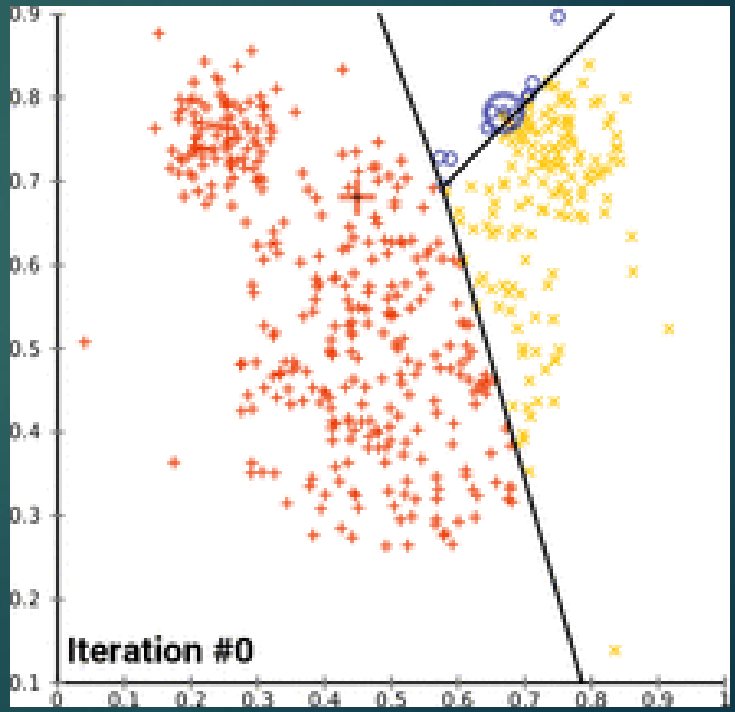
- All of the attributes from the other clustering
- Rebounds
- Assists
- Turnovers
- Two Point Attempts
- Two Point %
- Usage %

K-Means Clustering (Lloyd's Algorithm)

- Designate K random samples as the initial centers of K clusters
- Determine the closest cluster centroid for every season through Euclidean distance

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

- Re-adjust the centroid to be the mean of the seasons supplied to it
- Repeat until the centroids converge towards (approximately) a centroidal Voronoi tessellation

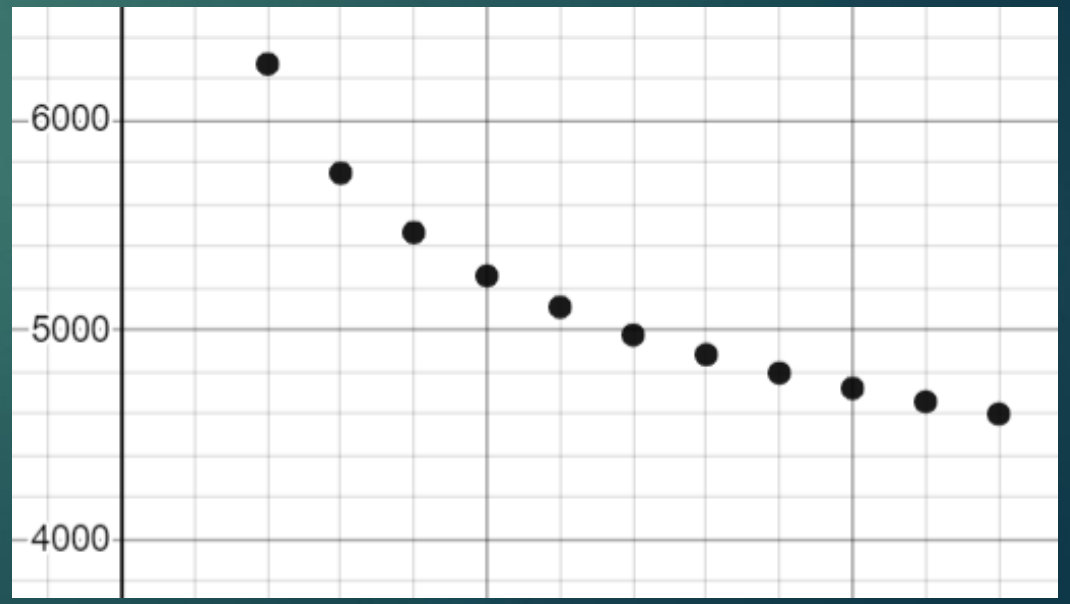


Random Sampling

- Final clusters are ultimately most impacted by the initial seed for that cluster. A bad seed can mean lopsided clusters and thus unusable data.
- Complete K-means multiple times and use the clustering with the lowest cumulative error between cluster centroid and assigned data.
- Sum of Squares Error $$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Choosing K

- Luckily we already have a K to compare to but not everyone is that lucky
- Is there a better K, can groupings be explained more accurately with more or fewer groups?
- Find the middle ground between usable clusters (1) and clusters without error (n)
- Turns out this middle ground is also about 5



Results (1978 – 2017)

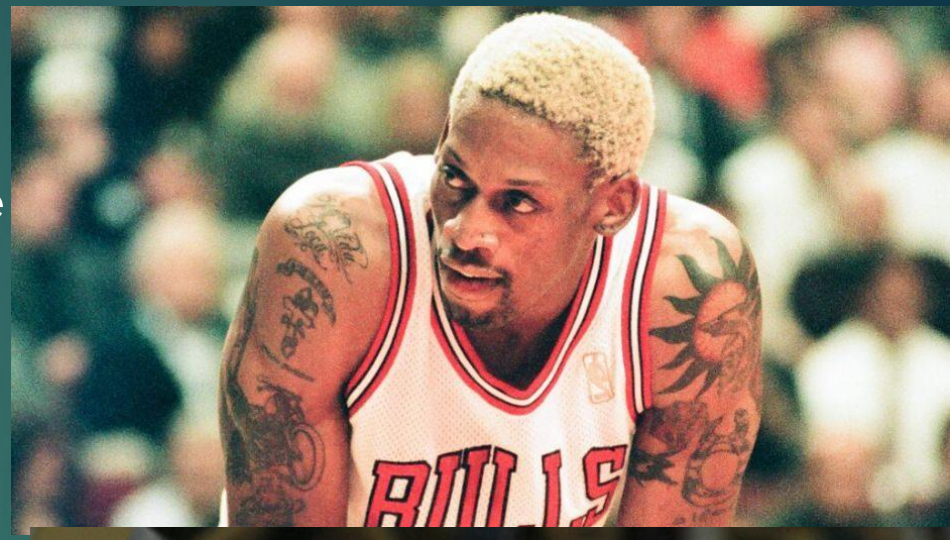
- Players cluster well into five groups, but now how you would think.
- 14,012 Eligible Seasons
- **Cluster 0 – 'All-Star Big-men'**
  - 2,738 Seasons
  - ~75% PF or C
- **Cluster 1 – '3 and D Wings (3 or D not guaranteed)'**
  - 3,534 Seasons
  - ~63% SF or SG
- **Cluster 2 – 'Traditional Point Guards'**
  - 2,414 Seasons
  - ~66% PG
- **Cluster 3 – 'GOATs'**
  - 2,206 Seasons
- **Cluster 4 – 'Early 2000s European Big-man signed to take 6 fouls from Shaq'**
  - 3,120 Seasons
  - ~86% PF or C

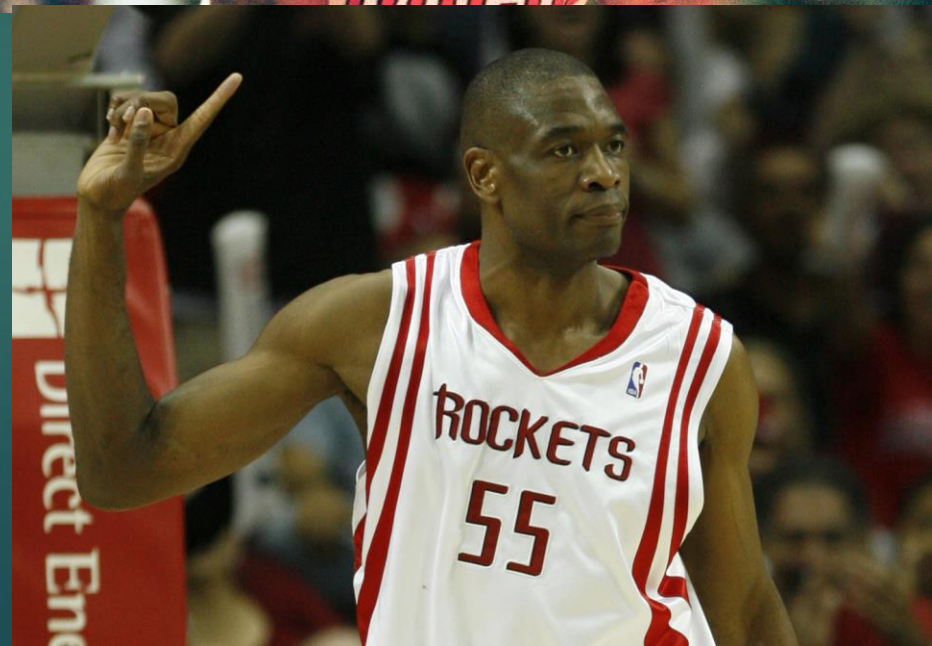
The Greatness of Clusters 0 and 3

- 666 Seasons by a Hall of Fame player
  - **Cluster 0** has 301
  - **Cluster 3** has 235
  - 80% of all Hall of Fame seasons (1978 – 2017)
  - 35% of total seasons
- Every player of the 1992 Dream Team spent the majority of their careers in Clusters 0 or 3
  - John Stockton a notable exception (Cluster 2)
- **Cluster 3** also contains the majority of the careers of current superstars like Lebron James, Kevin Durant, Stephen Curry, James Harden, etc.

The Mediocrity of Cluster 4

- Despite being the second largest group, there are only 44 seasons by a Hall of Famer in Cluster 4
  - 1 by Phil Jackson who is in the hall as a coach
- ~6% of Hall of Fame seasons from 22% of seasons
- Only 2 Hall of Famers spent the majority of their careers here (29 Seasons)
  - Dennis Rodman
  - Dikembe Mutombo





Extensions

- Implementation of more clustering algorithms
  - Fuzzy C-means
  - Compare groupings between the algorithms
- Analysis on this data set
- Data Visualization
  - Ggobi sort of bricks my computer
- Another clustering with even more attributes
  - Shot location (eg 0-3 ft, corner 3s)

Thank You!