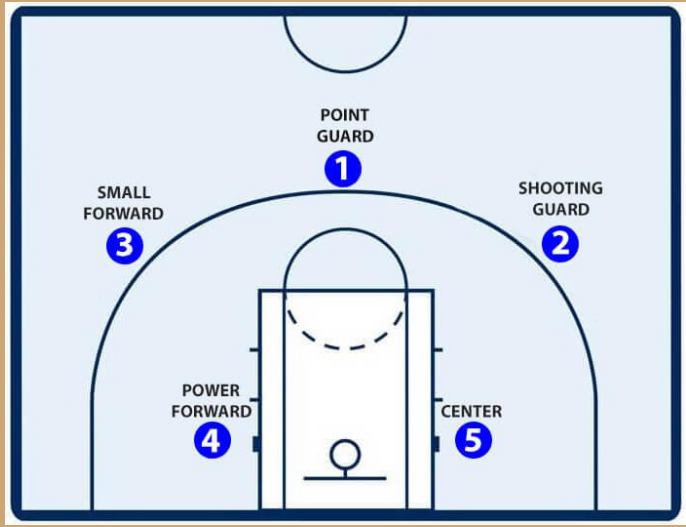
Clustering NBA Player Types

Identifying archetypes through unsupervised machine learning

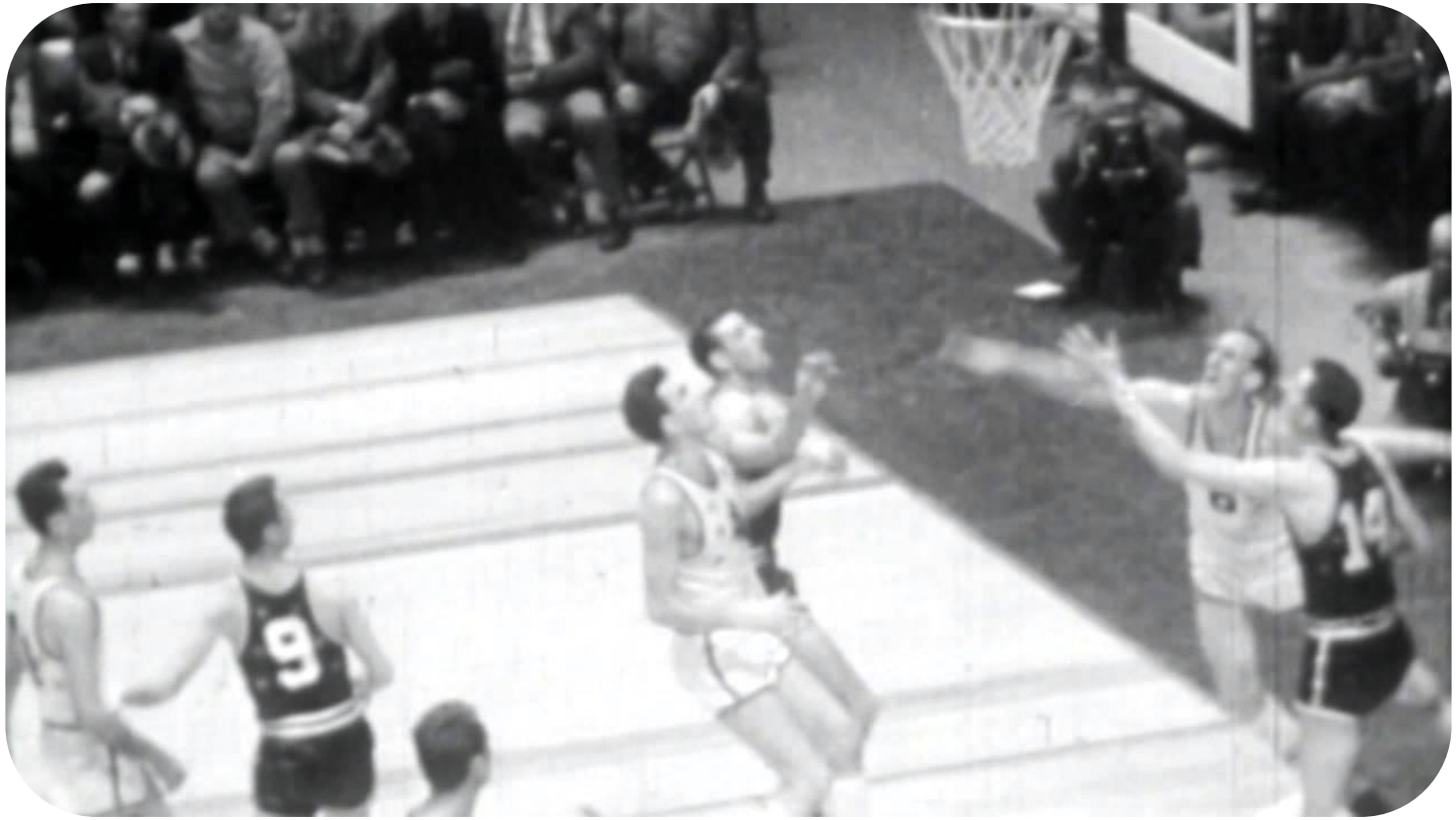
Remy Shea

Table of Contents



The five traditional roles of basketball have, for a long time, been useful tools to understanding the game.

James Naismith, Inventor of Basketball, 1861-1939



1954 NBA All-Star Game

$2.66 bn
2000-01

201%

$8.01 bn
2017-18

Increase in Revenue

Blake Griffin	Kevin Durant
NBA Superstar	
✔	✔
Around 6’11”	
✔	✔
Mostly Played PF in 2018/19	
✔	✔
Dated Kendall Jenner	
✔	✖

Table of Contents

Develop Metrics

Identify Archetypes

Inform Discussion

Develop metrics to measure play-style, identify player archetypes and inform discussions on roster and strategy

Table of Contents

HTML & BS4

Selection Criteria

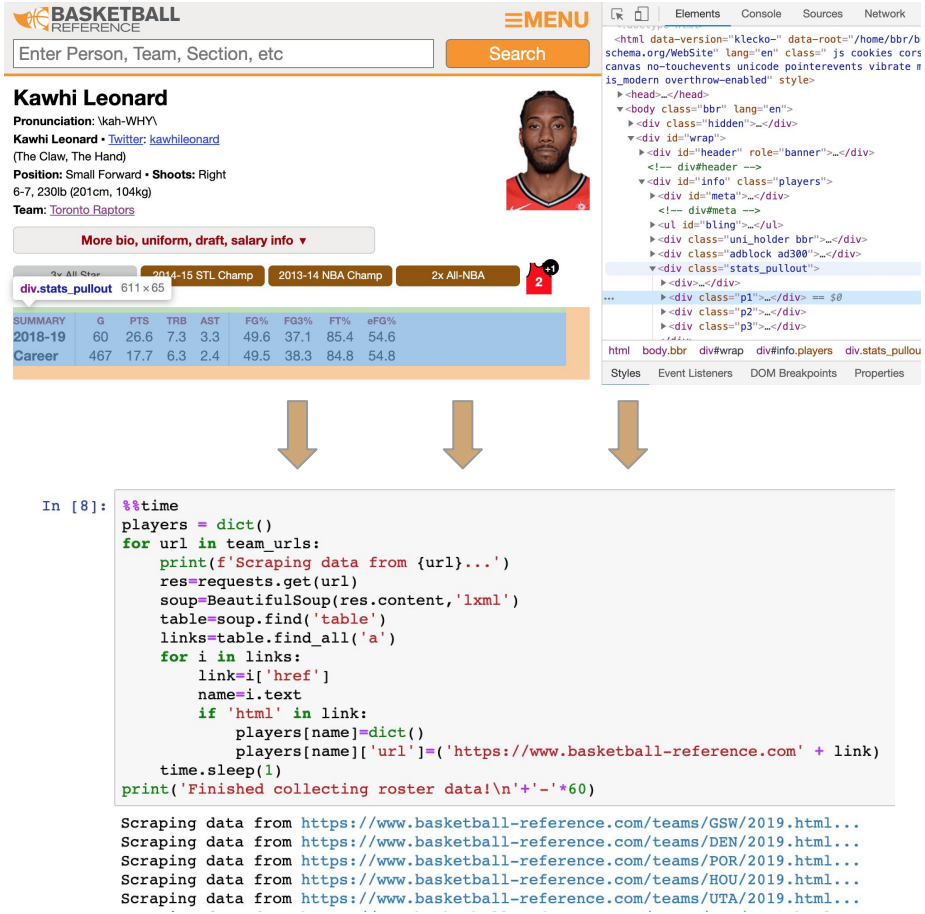
HTML & Web-Scraping

BeautifulSoup 4 Python library was used to make requests to:

http://www.basketball-reference.com/

Player URLs were extracted from team roster pages, which were then iteratively searched for Per-Game, Advanced, Shooting and Play-by-Play stats.

These data were saved to dictionaries indexed by player name.



Selection Criteria

497 VS 354

Players before selection criteria

Players after selection criteria

minimum 10 games played or 500 minutes played

Table of Contents

Small Data Sets

The Curse of Dimensionality

Multicollinearity

Small Datasets

The NBA has **30 teams**, 15 in each of the Eastern & Western conferences.

Each team is limited to **15 players** on a roster at any one time.

Most teams only give **8-10 players** consistent minutes every season.

Stricter selection criteria may cause the original 500 or so players to dwindle to maybe **300 players**.

Heuristic for maximum features in a dataset is the square root of the number of observations (354), or around **18 features**. We have **81 features**.

With enough features, even the mightiest datasets fall to sparsity

"The Curse of Dimensionality"

Multicollinearity

- Particularly a problem for clustering methods due to distance calculations
- Disproportionately accentuates certain areas of variance between players
- Adds complexity to model without capture additional variance

Table of Contents

Dimensionality Reduction

Explanatory Power

Variance Inflation Factor

Engineered Features

Positional Versatility

Root-mean-squared proportion of time spent at the traditional five positions.

The idea here is to capture the ability of a player to play many different styles of basketball, against different sizes of opponents.

Block-Foul Ratio

Blocked shots per 36 minutes divided by shooting fouls committed per 36 minutes.

An excellent defensive player will block shots whilst remaining in legal defensive position and not be called for fouls as often.

And-1 Completion Rate

And-1 plays per 36 minutes divided by shooting fouls drawn per 36 minutes.

Physical players are better able to finish through contact and convert more and-1 plays by scoring the basket despite being fouled.

Engineered Features

Draw Charge Rate

Offensive fouls drawn per 36 minutes divided by personal fouls minus shooting fouls minus own offensive fouls per 36 minutes.

Defenders can force turnovers from the opponents by establishing defensive position and drawing charges.

RMS Shot Distribution

Root-mean-squared proportion of shots from each distance from the basket.

Versatile players may attempt shots from every area of the floor, whereas specialists will have a higher RMS value.

RMS Shot Selection

Root-mean-squared element-wise product of proportion of shots from each area and field goal percentage from that area.

Smart players may be highly selective in only taking shots they know they will hit at a high rate.

Engineered Features

Assist-Turnover Ratio

Excellent passers will find their teammates for easy looks whilst limiting turnovers.

Efficient Shot Tendency

Smart players will disproportionately take corner 3s and dunks where possible, as they are the ‘best’ shots.

Isolation Effectiveness

Non-assisted field goal attempts divided by its sum with turnovers per 36 minutes.

Great 1-on-1 players will be able to score on their own without turning the ball over to the opponent.

Off-Def Rebound Ratio

Hustle players never give up on a play and will grab more rebounds off teammate misses.

Off-Def BPM Ratio

Offensively minded players will contribute more to team offensive productivity than defense.

Variance Inflation Factor

- Similar to Pairwise Correlation
- Quantifies multicollinearity in dataset
- “How well is one feature predicted by all the other features available?”

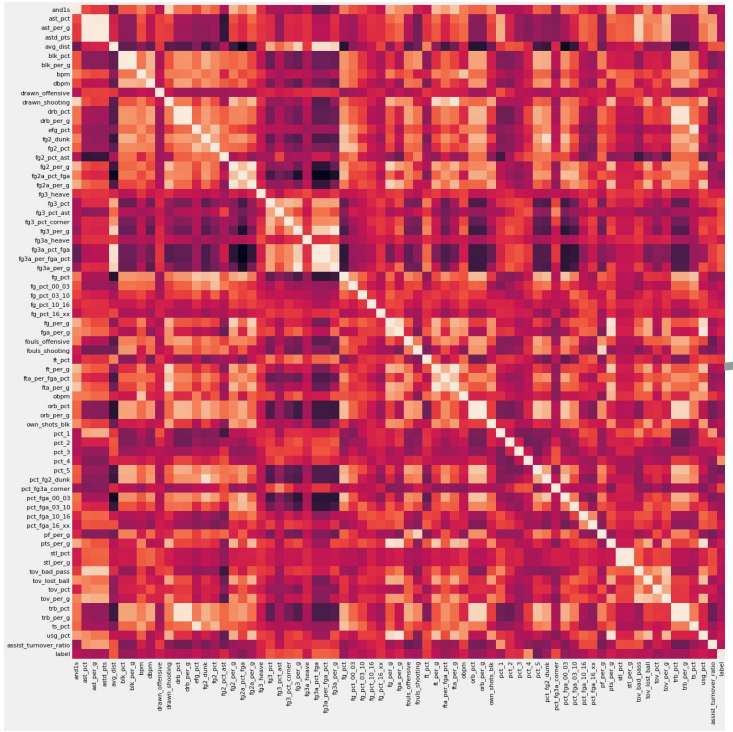
Before Feature Engineering

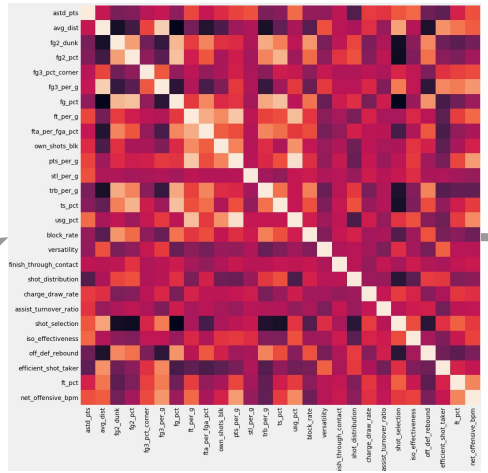
	VIF	features
27	inf	fg3a_per_fga_pct
26	inf	fg3a_pct_fga
18	1.228699e+06	fg2a_pct_fga
53	3.184417e+05	pct_fga_00_03
13	1.421711e+05	efg_pct
29	1.036312e+05	fg_pct
54	8.601442e+04	pct_fga_03_10
35	3.613340e+04	fga_per_g
67	3.249413e+04	ts_pct
55	2.887261e+04	pct_fga_10_16

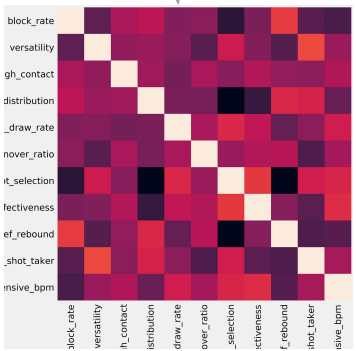
After Feature Engineering

	VIF	features
10	123.966097	pts_per_g
14	82.753929	usg_pct
6	68.338813	fg_pct
7	31.632734	ft_per_g
13	31.331972	ts_pct
1	30.331166	avg_dist
5	19.966155	fg3_per_g
21	12.361505	shot_selection
8	10.817946	fta_per_fga_pct
26	9.924090	net_offensive_bpm

Dimensionality Reduction Results







Num = number of features

mPWC = mean pair-wise correlation

sPWC = pair-wise correlation standard deviation

Num,	mPWC,	sPWC
81,	0.086,	0.088
27,	0.068,	0.083
11,	0.061,	0.043

Table of Contents

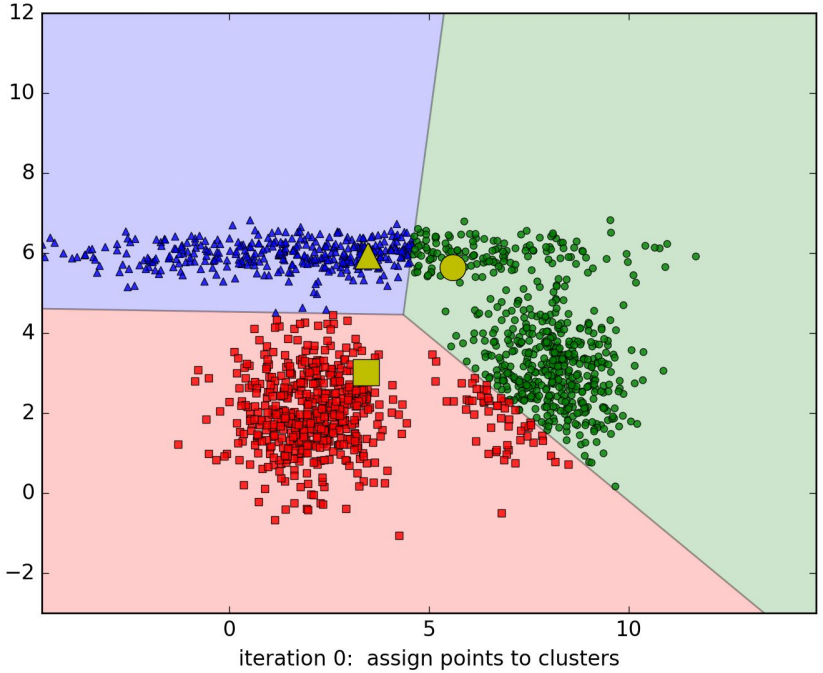
K-Means Clustering

DBSCAN

Silhouette Score

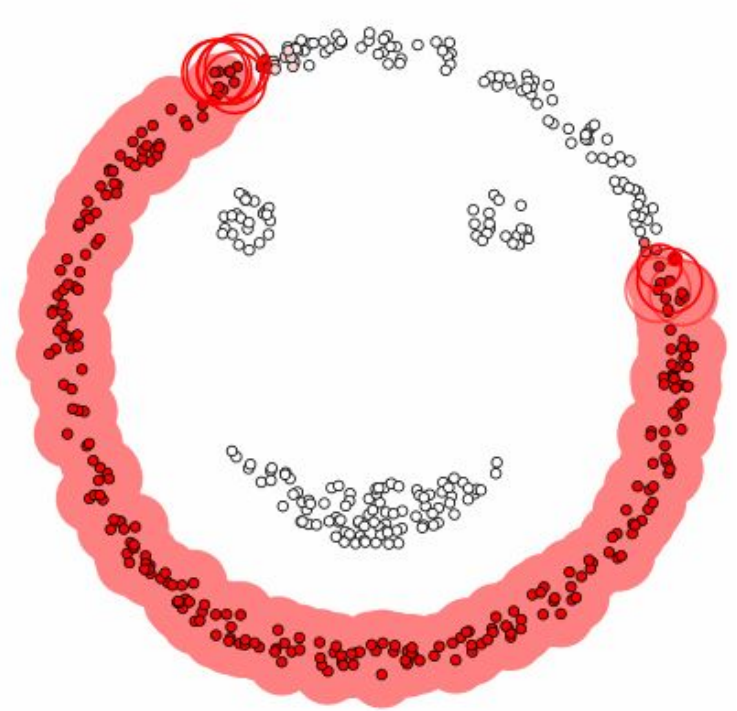
K-Means Clustering

1. Assign k centroids at random
2. Assign each datapoint to nearest cluster
3. Update centroid position with centroid mean
4. Repeat until convergence



Density-Based Spatial Clustering (DBSCAN)

1. Assign a cluster to a point.
2. Scan within a given radius (epsilon) of a data point for neighboring points
3. If there are enough (n) points in this vicinity, assign the point to the cluster.
4. Repeat this step for every datapoint in the vicinity until step 3 fails, then start a new cluster.



Silhouette Score

Homogeneity = average distance of a point to members of same cluster

Dissimilarity = average distance of a point to members of a neighboring cluster

Silhouette Score = Dissimilarity minus Homogeneity divided by the greater of the two.

-1 is terrible. +1 is excellent.

Table of Contents

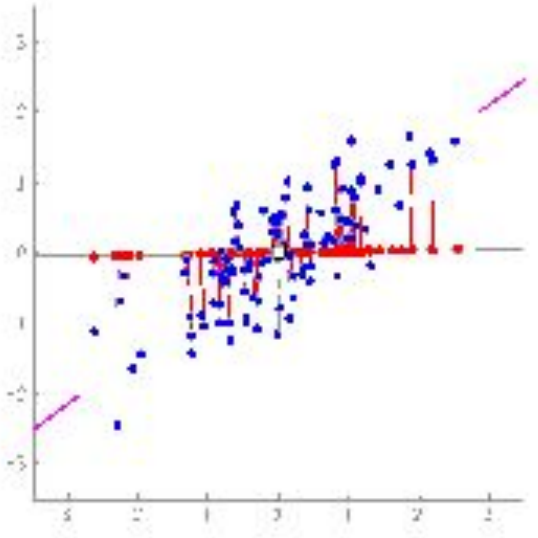
PCA

t-SNE

Interpretability

Principal Component Analysis (PCA)

1. Find the vector in the feature space to which each datapoint is the closest.
2. Repeat this process for all vectors perpendicular to all prior vectors.
3. Sort these eigenvectors by their eigenvalues.
4. Voila you've got a PCA.



t-distributed Stochastic Neighbor Embedding (t-SNE)

1. Scatter the data points into a low-dimensional space.
2. For each datapoint, sum the ‘pulls’ & ‘pushes’ from members of the same and other clusters respectively, according to their distance from the point, scaled by a t-distribution.
3. Move in the net positive direction.
4. Repeat for each datapoint.

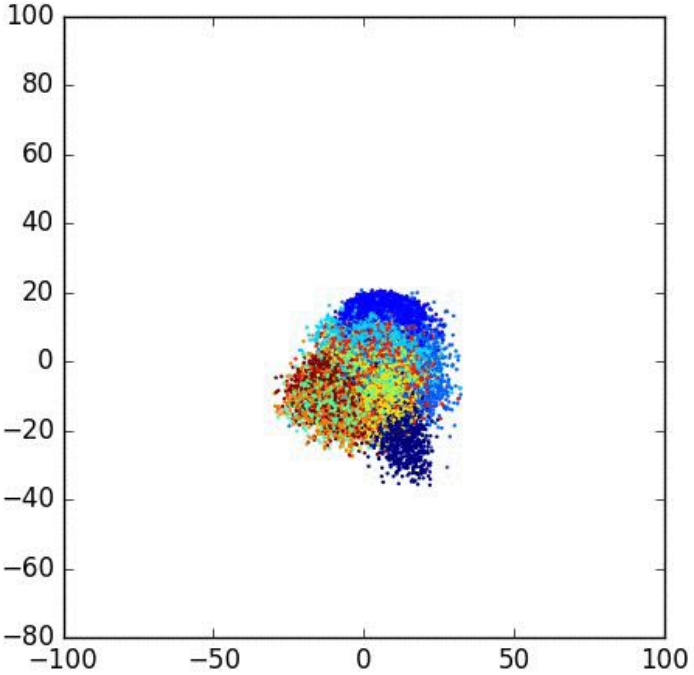


Table of Contents

Bringing it all together

What on earth happened?

Why did you even bother?

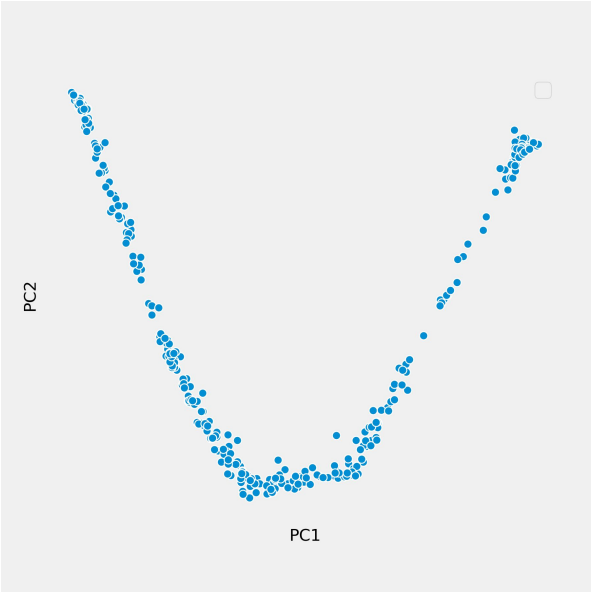
Original Features

A DBSCAN on a PCA of the original features with epsilon of 18, minimum 20 samples identified 6 clusters.

The first two principal components explained roughly 65% of the variance, and so were moderately suitable to visualize the data.

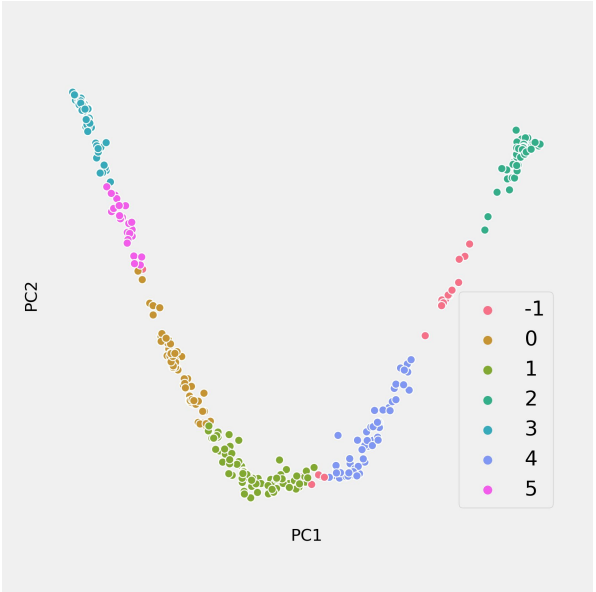
Although the silhouette score of **0.62** was very good, it is visually unconvincing.

Un-labeled PCA Visualization



DBSCAN Clustering on Original, PCA-Transformed Features

Labeled PCA Visualization



DBSCAN Clustering on Original, PCA-Transformed Features

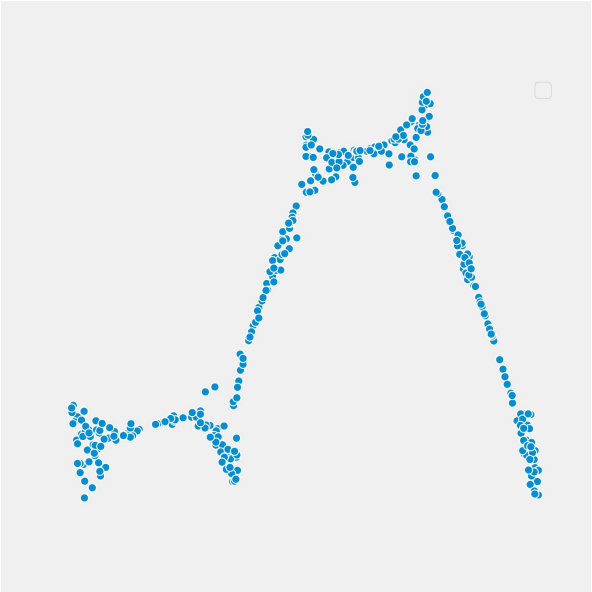
Original Features

The same DBSCAN on PCA of original features.

The t-SNE provides a more convincing view of the algorithms ability to separate players into clusters.

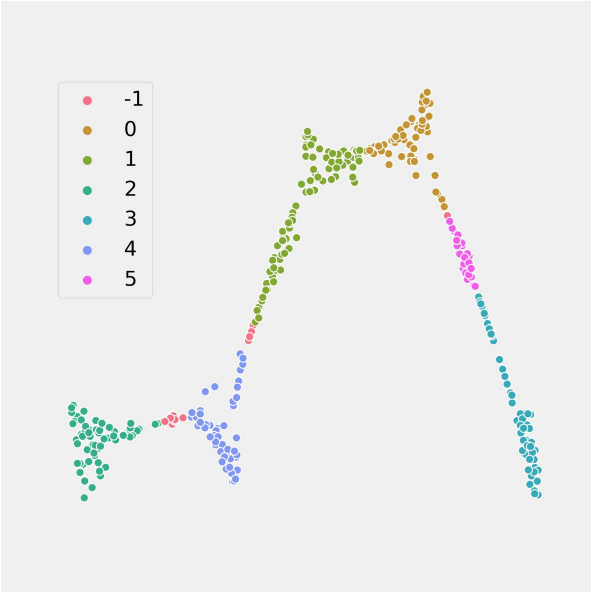
This highlights the limitations of PCA in terms of visualizing high-dimensional data.

Un-labeled t-SNE Visualization



DBSCAN Clustering on Original, PCA-Transformed Features

Labeled t-SNE Visualization



DBSCAN Clustering on Original, PCA-Transformed Features

Drawbacks with Using Original Dataset

- Multicollinearity causes clustering to be disproportionately aligned with differences along those duplicated features (shooting stats).

- If only concerned with play style, we want to limit player-talent, team-composition & team-success effects on clustering of players (PER, VORP, BPM).

- Potentially uninteresting clusters in attempting to answer player tendency (eye test), and clustering metrics may be misleading.

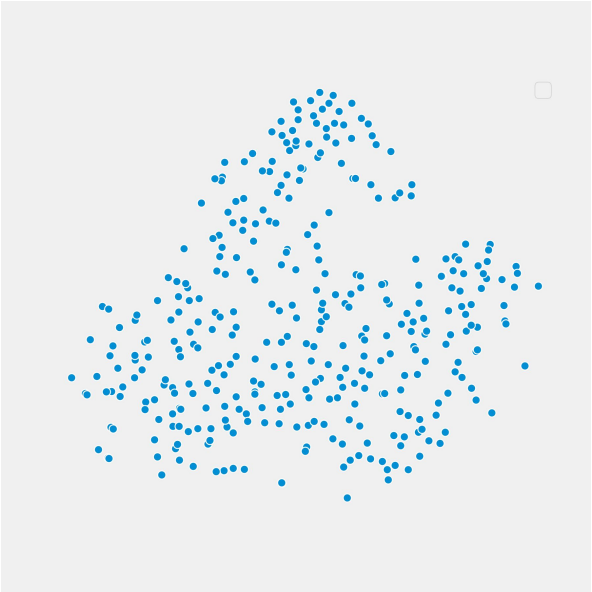
Selected Features

The 27 selected & engineered features produce a measly silhouette score of **0.145** in a K-means clustering with k = 4.

The unlabeled visualization shows that there may be one, at most two clear clusters in the data.

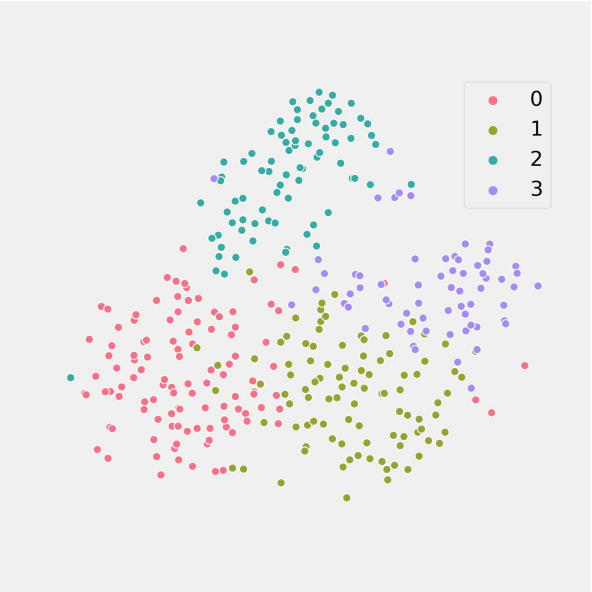
There exists a trade-off between collinearity and clustering

Un-labeled t-SNE Visualization



K-Means Clustering on Selected & Engineered Features

Labeled t-SNE Visualization



K-Means Clustering on Selected & Engineered Features

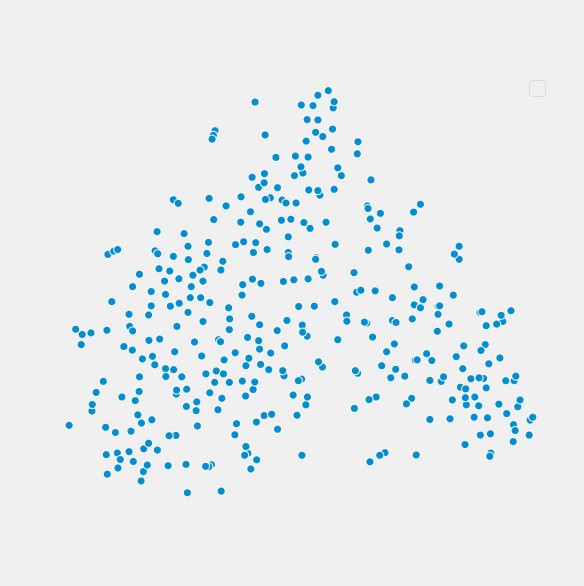
Engineered Features

The 11 engineered features produce an abysmal silhouette score of **0.123** in a k-means clustering with k = 7.

The t-SNE visualizations were chosen because the first two principal components of the engineered features did not explain enough variance.

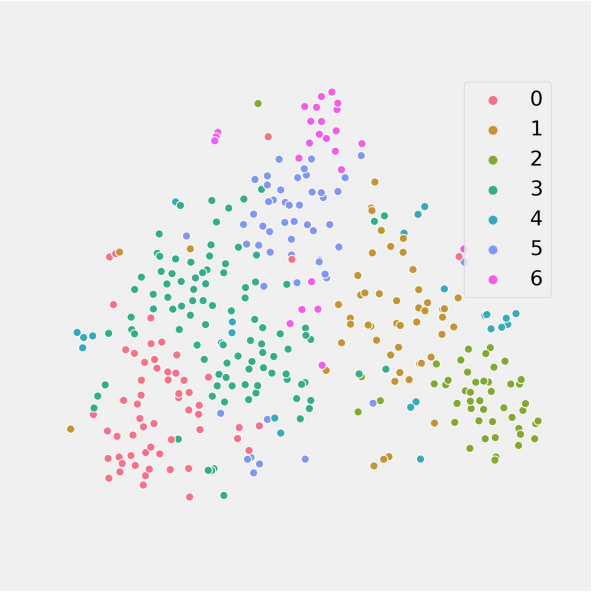
More feature engineering may be necessary.

Un-labeled t-SNE Visualization



K-Means Clustering on Engineered Features

Labeled t-SNE Visualization



K-Means Clustering on Engineered Features

Table of Contents

stats.nba.com & SeleniumWD

Tracking Player Improvement

League-Wide Trends In Time

NBA Advanced Stats

Players Hustle Leaders

PLAYERS TEAMS

Deflections Per Game

1.	Ben Simmons PHI	3.9
2.	Evan Fournier ORL	3.4
3.	Damian Lillard POR	3.3
4.	Paul George OKC	3.2

Deflections Per 36

1.	Khyri Thomas DET	7.2
2.	Jodie Meeks TOR	4.0
2.	Ben Simmons PHI	4.0
4.	Jared Dudley BKN	3.9

Deflections Totals

1.	Ben Simmons PHI	
2.	Damian Lillard POR	
3.	Kyle Lowry TOR	
3.	Klay Thompson GSW	

Table of Contents

Auto-Encoder Neural Nets

Additional Clustering Techniques

Additional Clustering Metrics

Advanced Methods

Advanced Clustering Methods:

- Agglomerative Clustering
- Mean-Shift Clustering
- Affinity Propagation
- Expectation Maximization & Gaussian Mixture Models

Advanced Clustering Metrics:

- Cluster Stability
- Homogeneity Score
- Davies-Bouldin Index
- Perplexity on Held-Out Data

Table of Contents

Transfer Learning : Rosters

Transfer Learning : Rookies

Unicorn/Dodo Metrics

Table of Contents

Thank You For Your Attention


STEPH CURRY
CURRY
30