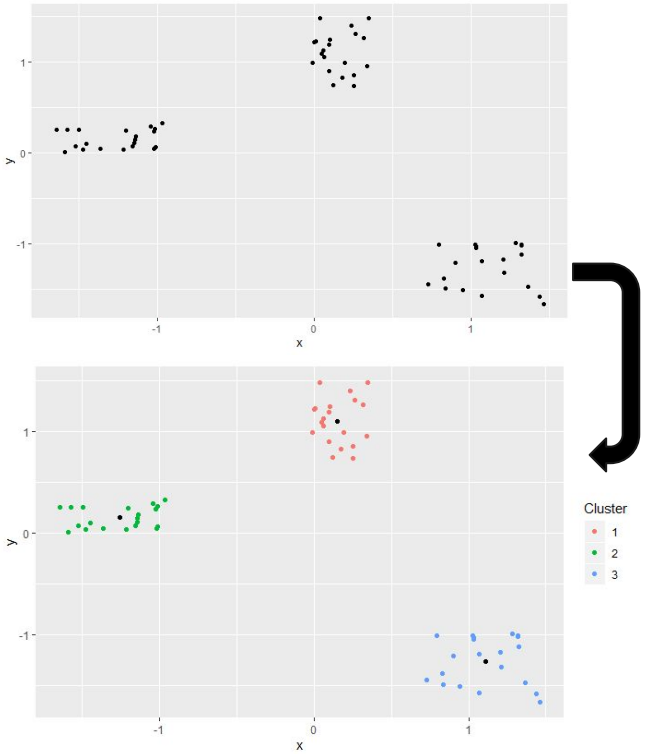
Clustering With PCA and K-means

Brandon Chan

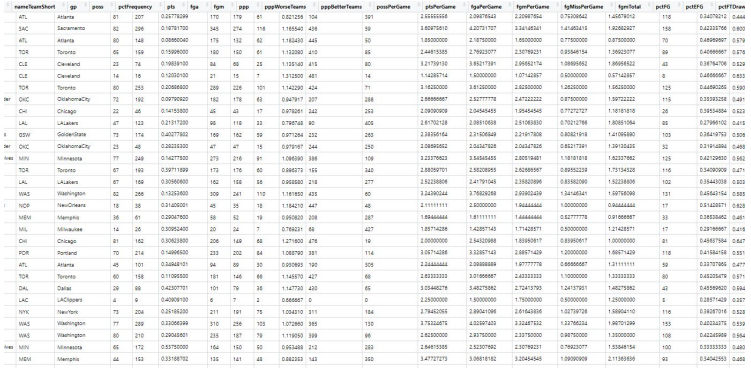
What is clustering?

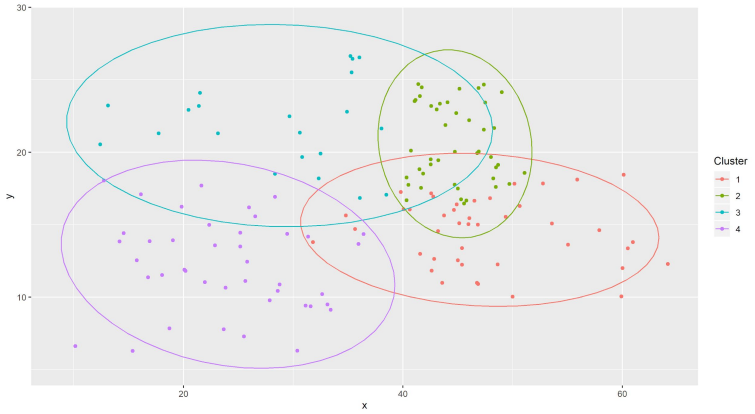
- Clustering is when you group similar data points together
- One group of similar data points makes up a cluster
- One key characteristic of clustering is that the groups/clusters are not predefined



Why cluster?

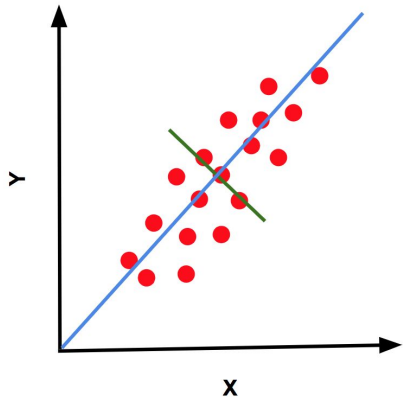
- Clustering is a great way to get the bigger picture of a data set
- Clustering is especially useful for understanding big data sets with lots of variables
- By grouping together similar data points we can:
    - Find relationships among variables
    - Spot outliers
    - Uncover underlying patterns in the data



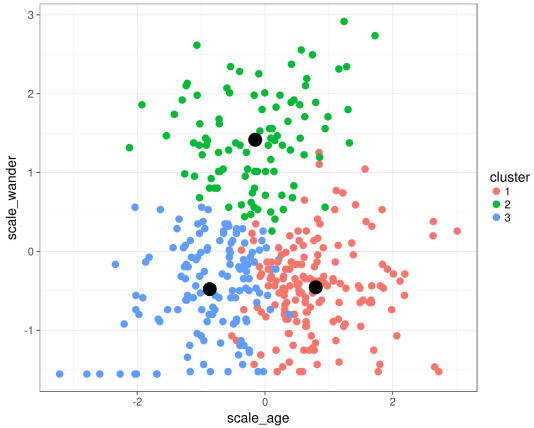


How?

Principal Component Analysis

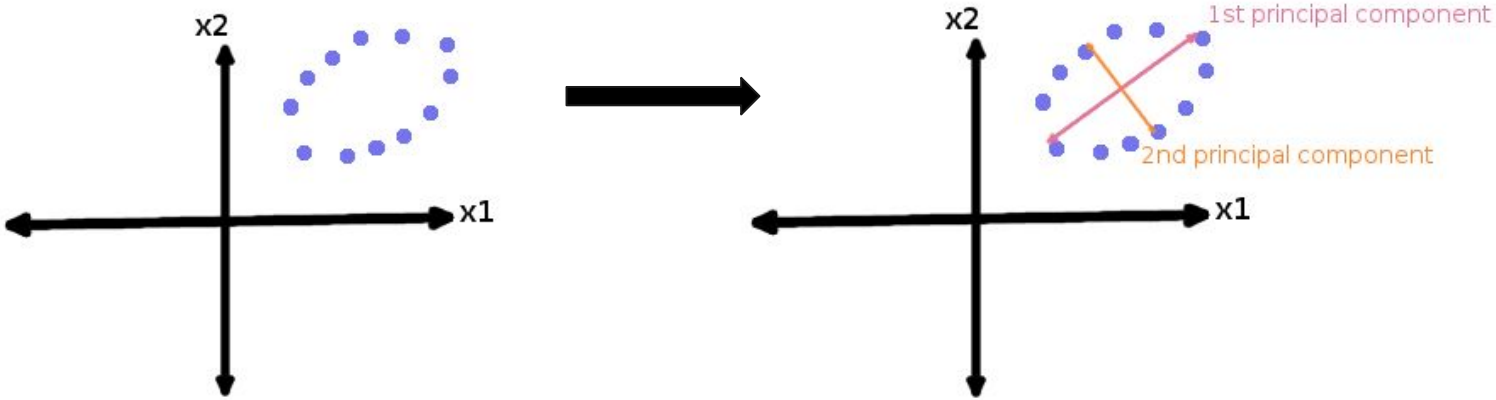


K-means clustering



Principal Component Analysis

- Principal Component Analysis (PCA) is a transformation of the original data
- PCA plots the original data points on a new coordinate system such that the variation among the data points is maximized



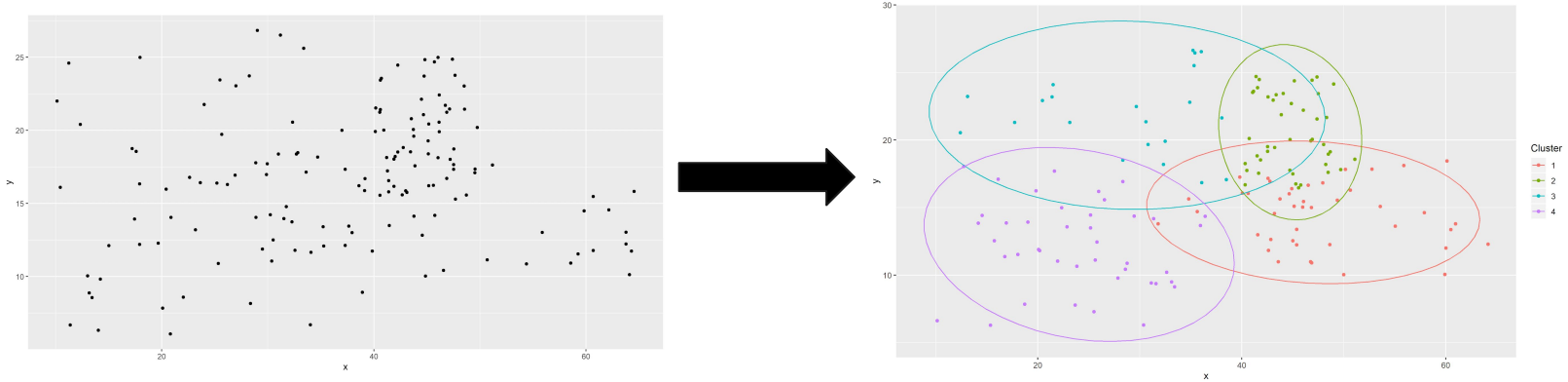
Some Things to Note

- Each axis in the new coordinate system is known as a Principal Component
- Each Principal Component represents one way that the data points can be transformed
- Each transformation results in a different amount of variation among the data points
- The Principal Components are ordered from the highest amount of variation to the least amount of variation

In summary, PCA transforms the points so that there is the MOST variation in the LEAST number of axes.

K-means Clustering

- Simply put, here's how k-means clustering works:
  a. The user decides how many clusters they want to have
  b. The k-means algorithm decides where the centers of those clusters should be
  c. Every data point is assigned to the nearest cluster



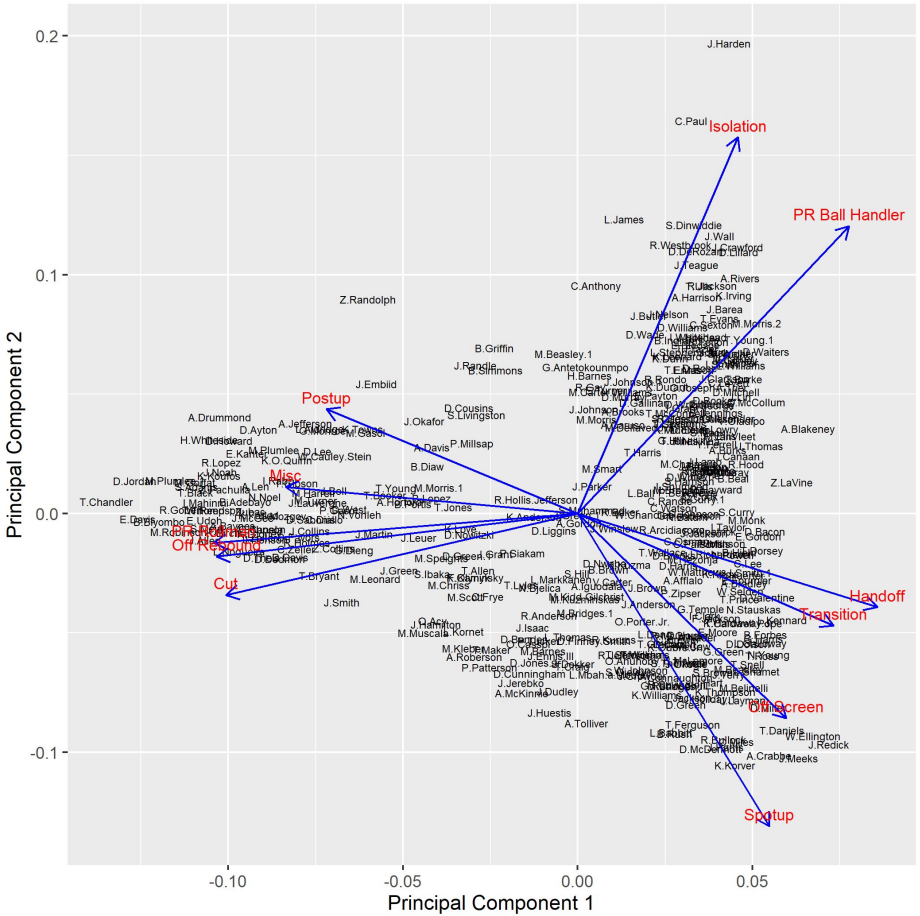
The Data

One data point - One NBA Player who played at least 1000 minutes in the past 3 seasons

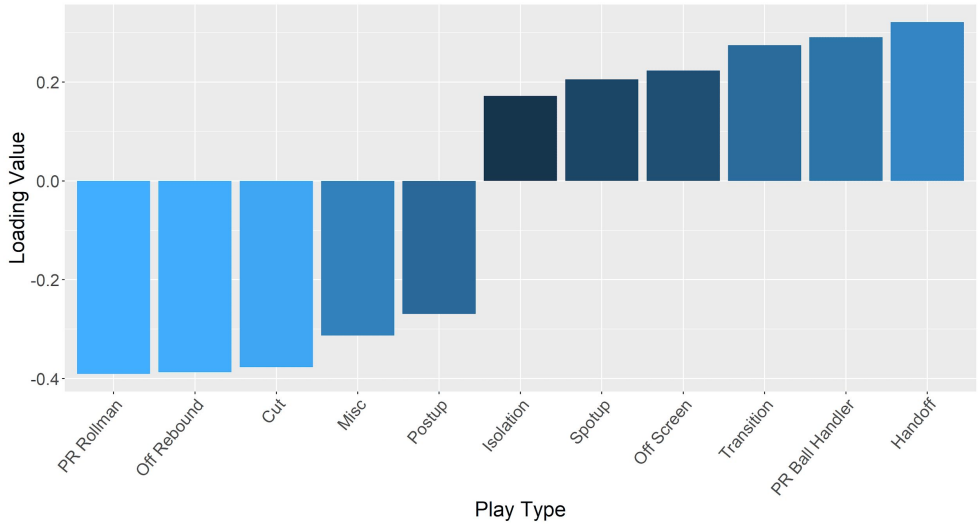
Variables - The frequencies at which they took part in the following plays:

- Rollman in the pick and roll
- Off rebound
- Cut
- Postup
- Isolation
- Spotup
- Off screen
- Transition
- Ball handler in the pick and roll
- Handoff
- Misc

Biplot

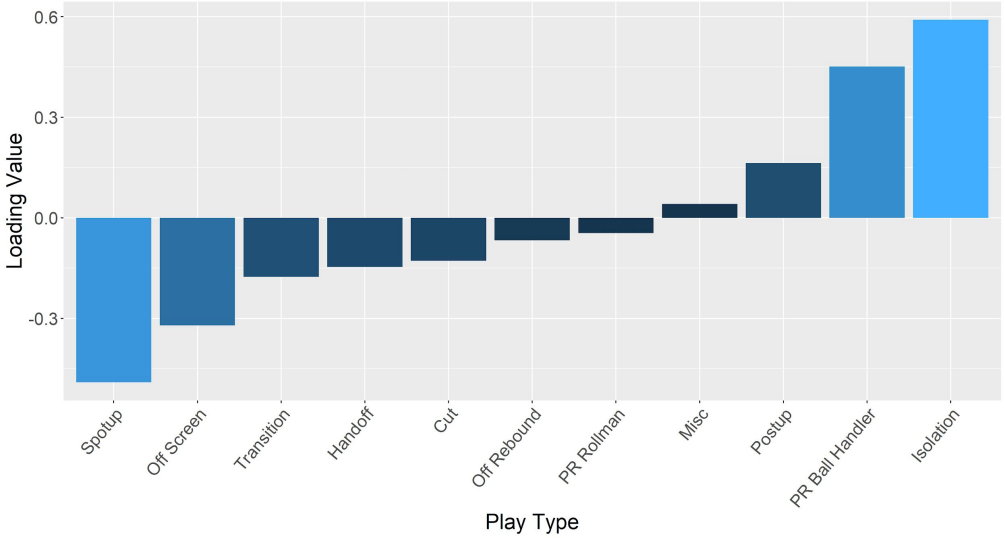

Isolation
PR Ball Handler
Postup
Misc
Off Reb/Putback
Cut
Handoff
Transition
OffScreen
Spotup
J.Harden
C.Paul
L.James
S.Dinwiddie
C.Anthony
Z.Randolph
B.Griffin
M.Beasley.1
G.Antetokounmpo
J.Embiid
D.Cousins
J.Okafor
P.Millsap
A.Drummond
A.Blakeney
Z.LaVine
T.Chandler
J.Parker
R.Hollis.Jefferson
D.Liggins
J.Leuer
J.Martin
T.Bryant
M.Leonard
J.Smith
M.Kidd.Gilchrist
R.Anderson
M.Bridges.1
O.Porter.Jr.
J.Isaac
M.Kornet
J.Huestis
A.Tolliver
A.McKinnie
D.Cunningham
J.Jerebko
P.Patterson
M.Kleber
A.Roberson
M.Muscala
J.Hamilton
T.Ferguson
K.Korver
D.McDermott
J.Redick
W.Ellington
T.Daniels
A.Crabbe
J.Meeks
Principal Component 2
Principal Component 1
0.2
0.1
0.0
-0.1
-0.10
-0.05
0.00
0.05

PC1 Loadings



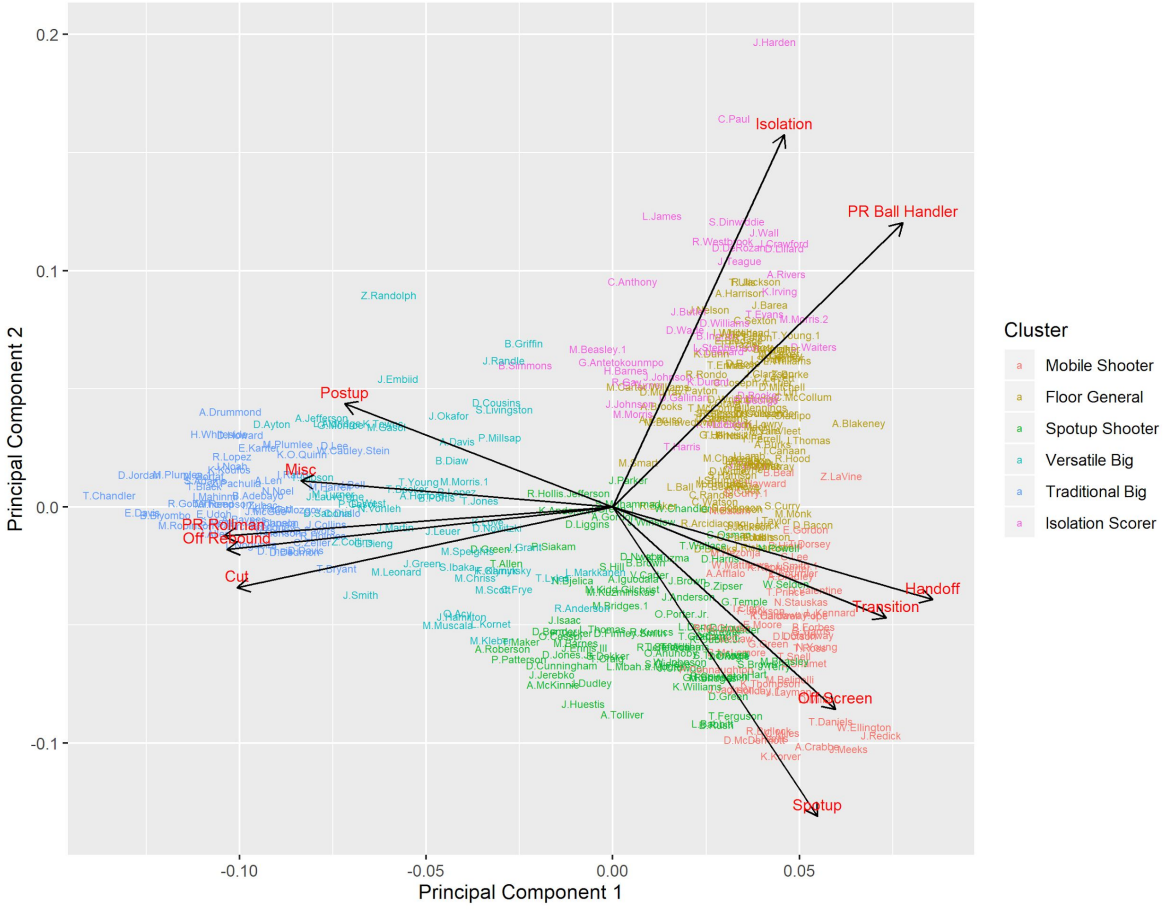
- High Handoff, PB Ball Handler, and Transition
- Low PR Rollman, Off Rebound, and Cut
- Separates big men from guards and forwards

PC2 Loadings

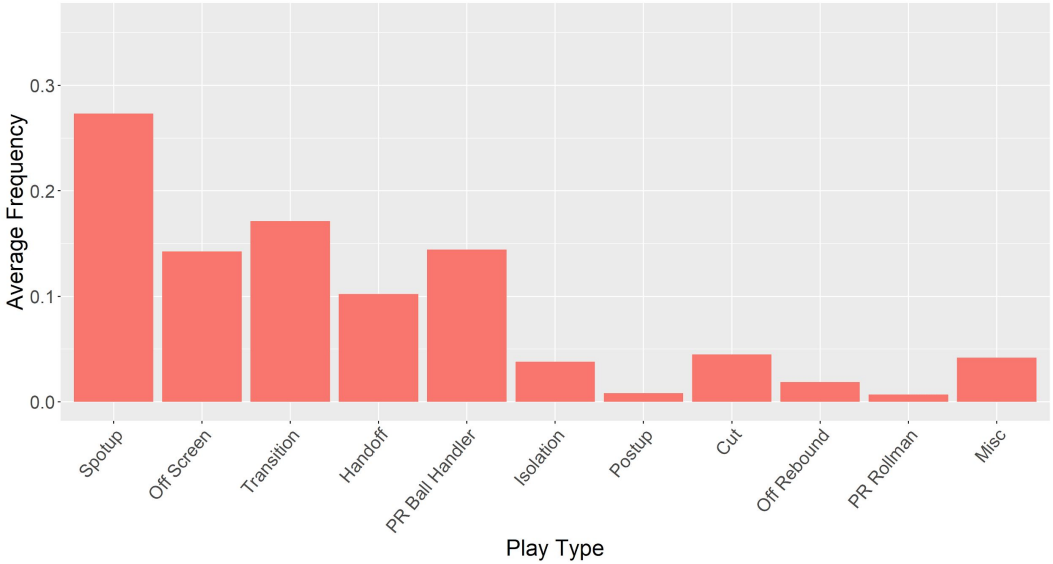


- High Isolation, and PR Ball Handler
- Low Spotup and Off Screen
- Separates ball handlers from shooters

Clusters

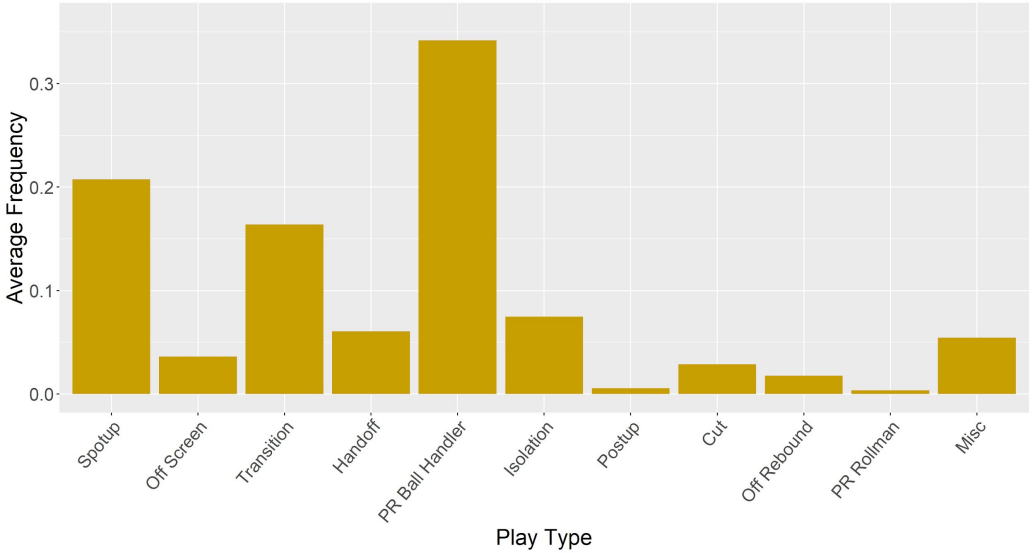

Principal Component 2
Principal Component 1
0.2
0.1
0.0
-0.1
-0.10
-0.05
0.00
0.05
Isolation
PR Ball Handler
Postup
Misc
PR Rollman
Off Rebound
Cut
Handoff
Transition
Off Screen
Spotup
Cluster
Mobile Shooter
Floor General
Spotup Shooter
Versatile Big
Traditional Big
Isolation Scorer

Mobile Shooters



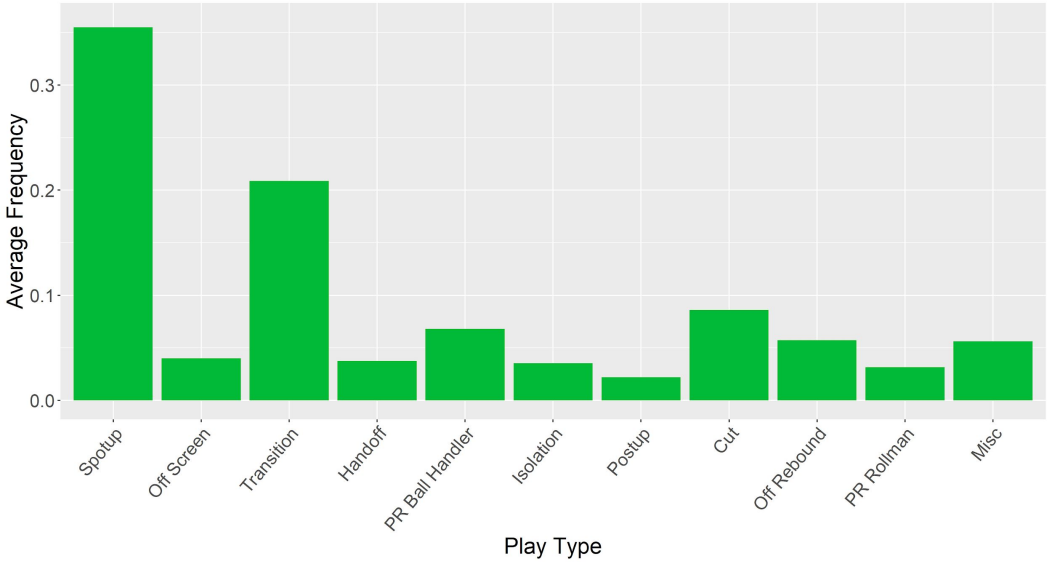
- High Spotup, Off screen, Transition
- Low PR Rollman, Postup, Off Rebound
- Key members: Buddy Hield, JJ Redick, Klay Thompson

Floor Generals



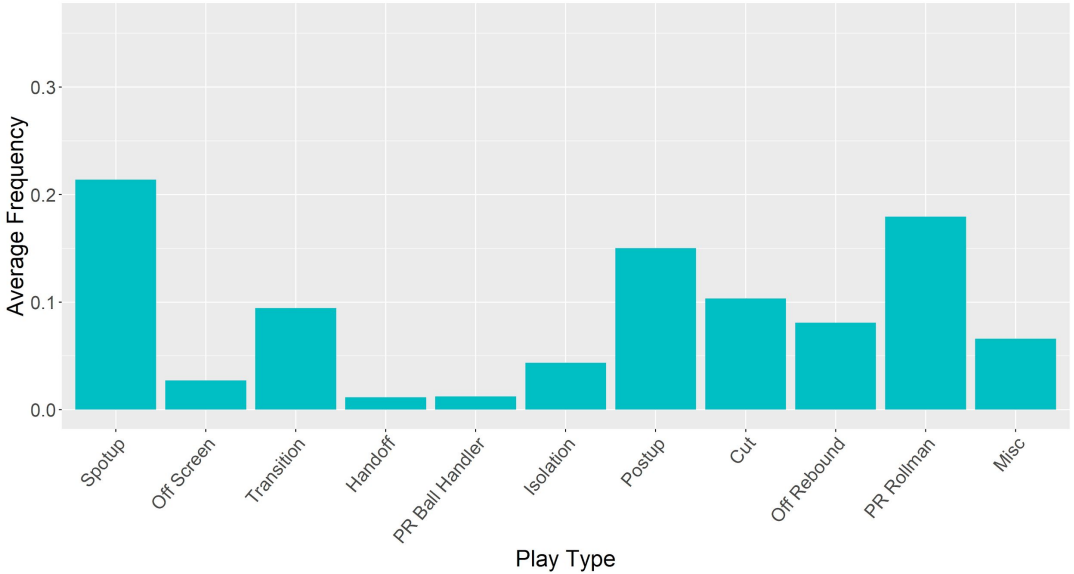
- High PR Ball Handler, Spotup, Transition
- Low PR Rollman, Postup, Off Rebound
- Key Members: Steph Curry, Lonzo Ball, Rajon Rondo

Spotup Shooters



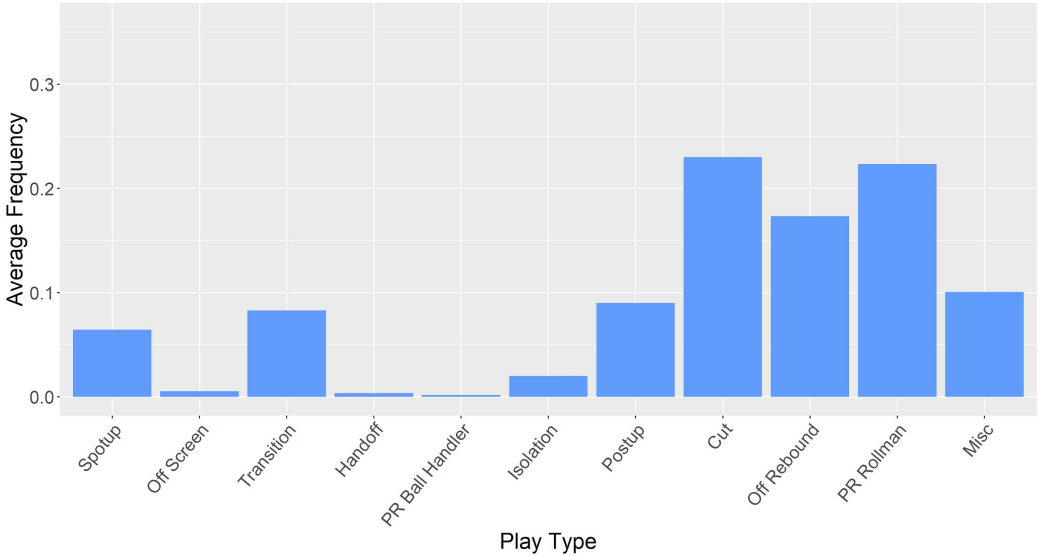
- High Spotup and Transition
- Low Off Screen and Postup
- Key members: Otto Porter Jr., Anthony Tolliver, Nemanja Bjelica

Versatile Bigs



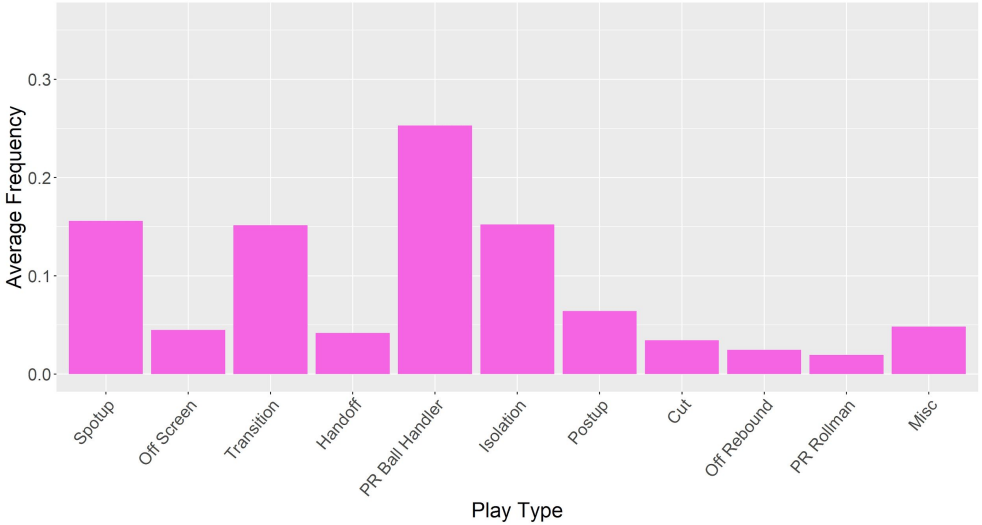
- High Spotup, PR Rollman, Postup
- Low Handoff, PR Ball Handler, Off Screen
- Key members: Demarcus Cousins, Anthony Davis, Blake Griffin

Traditional Bigs



- High Cut, PR Rollman, Off Rebound
- Low PR Ball Handler, Handoff, Off Screen
- Key members: Rudy Gobert, DeAndre Jordan, Andre Drummond

Isolation Scorers



- High PR Ball Handler, Isolation, Transition
- Low PR Rollman, Off Rebound, Cut
- Key members: James Harden, Lebron James, Russell Westbrook

Main Takeaways

- The best way to differentiate NBA players is by separating big men from non-big men
- The second-best way to differentiate NBA players is by separating ball handlers from non-ball handlers.

Big Men	Ball Handlers	Shooters
Postups	Isolation	Spotting up
Being the rollman	Running the pick and roll	Running in Transition
Grabbing rebounds		Handoffs
Cuts		Running off screens
Miscellaneous		

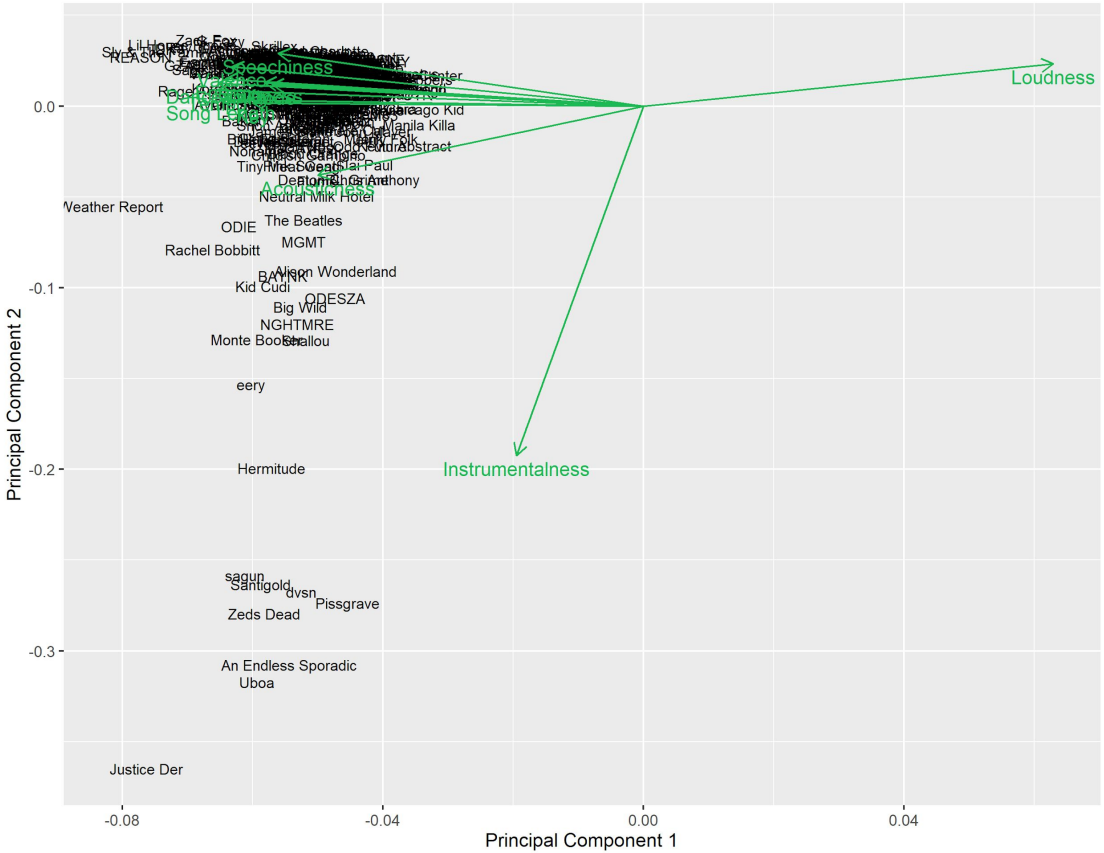
The Data

One data point - One artist that I have listened to in the past 4 months

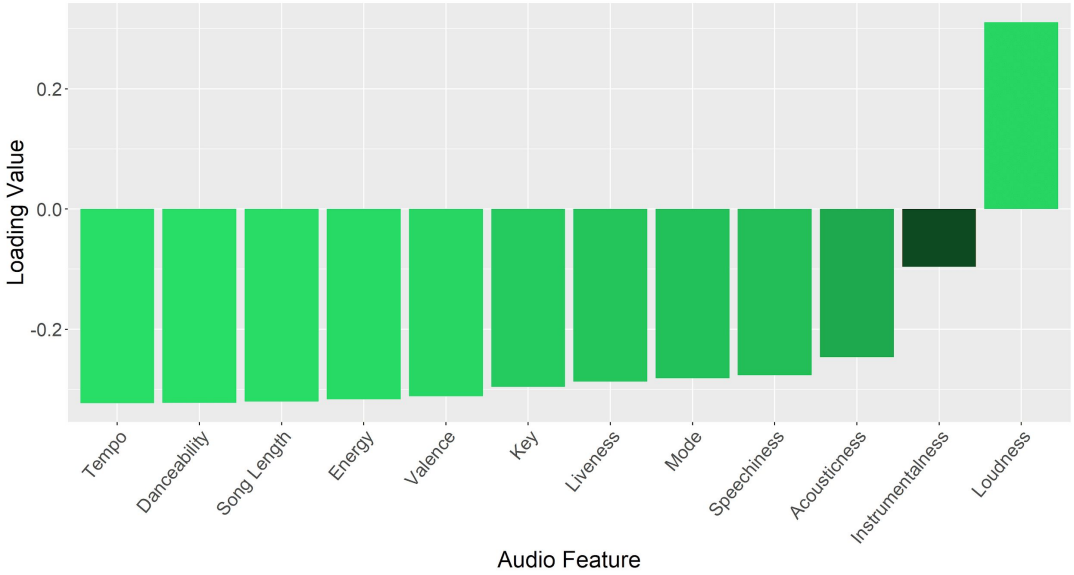
Variables - the average of these attributes from the songs I listened to from each artist:

- Tempo
- Danceability
- Song Length
- Energy
- Valence
- Key
- Liveness
- Mode
- Speechiness
- Acousticness
- Instrumentalness
- Loudness

Biplot

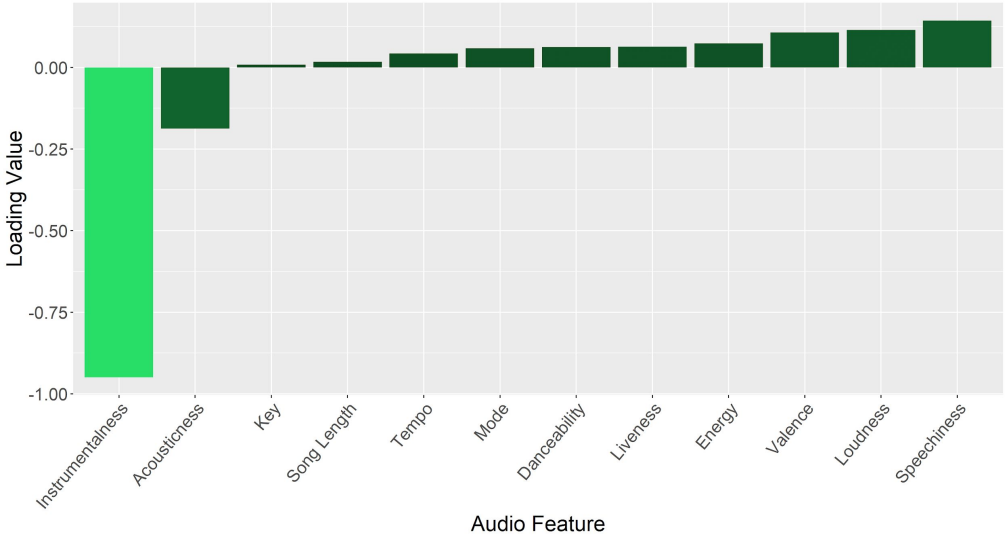

Loudness
Speechiness
Instrumentalness
Principal Component 2
Principal Component 1
0.0
-0.1
-0.2
-0.3
-0.08
-0.04
0.00
0.04
Weather Report
ODIE
The Beatles
Rachel Bobbitt
MGMT
Kid Cudi
ODESZA
Big Wild
NGHTMRE
eery
Hermitude
Zeds Dead
Pissgrave
An Endless Sporadic
Uboa
Justice Der

PC1 Loadings



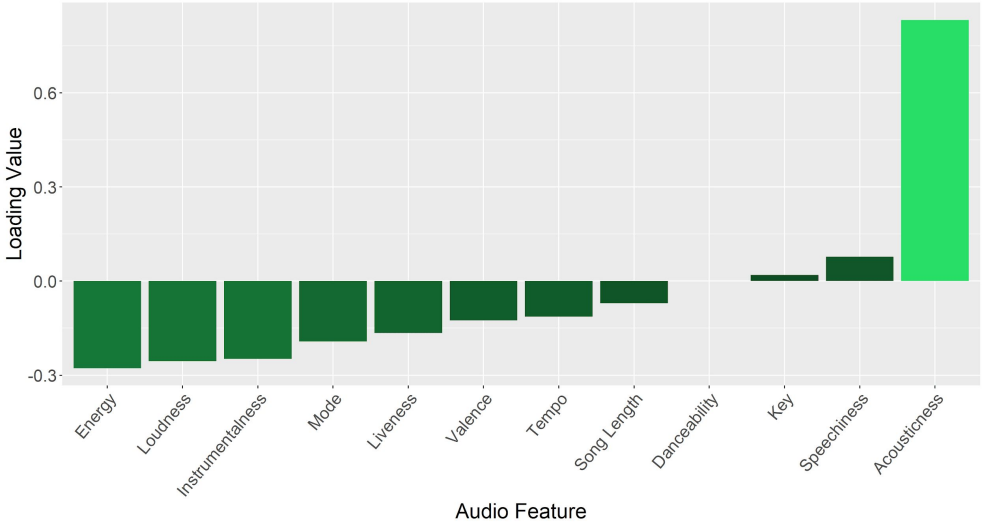
- High Loudness
- Low everything else
- Separates loud artists from not loud artists

PC2 Loadings



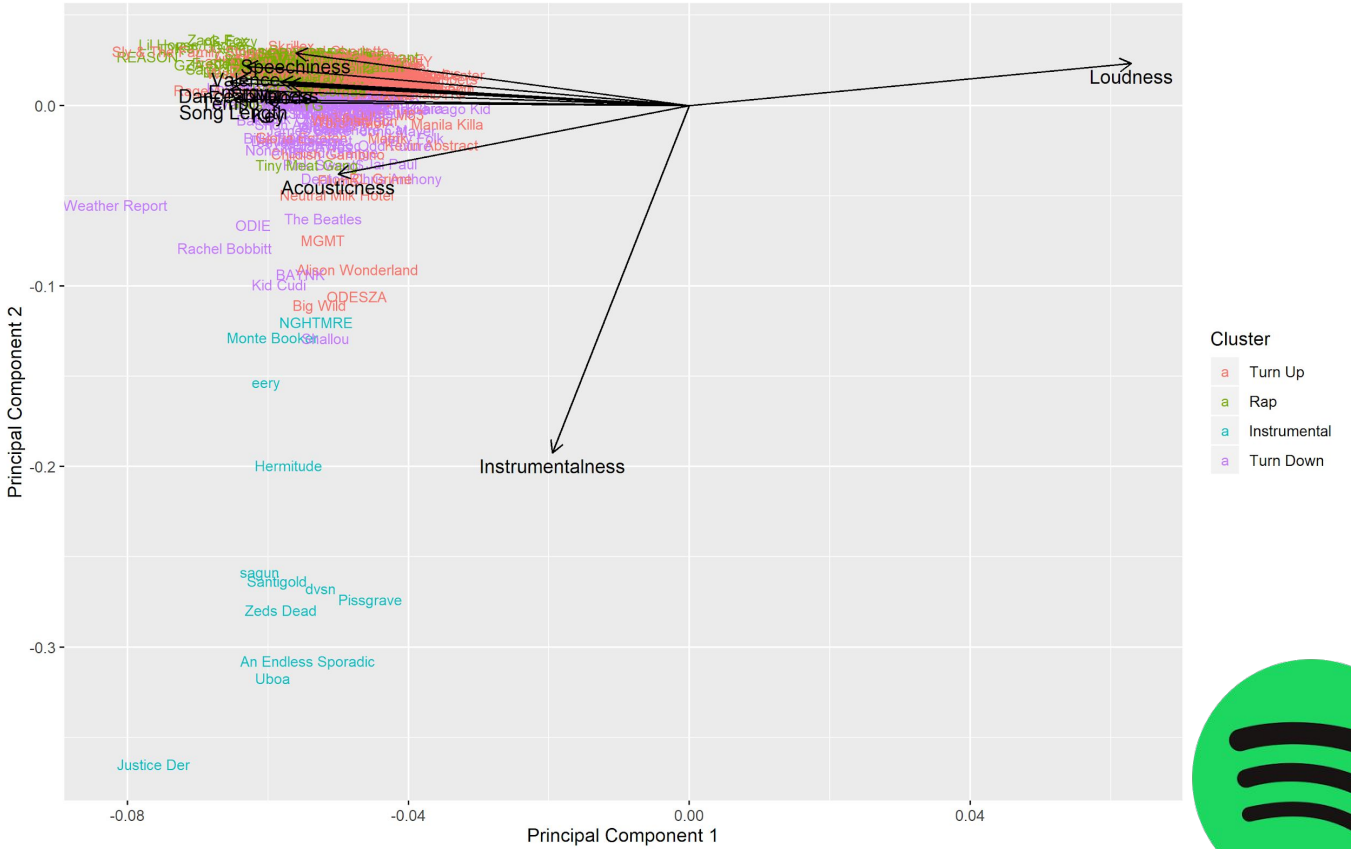
- High most variables
- Very low Instrumentalness
- Separates artists who play instrumentals and artists that use vocals

PC3 Loadings

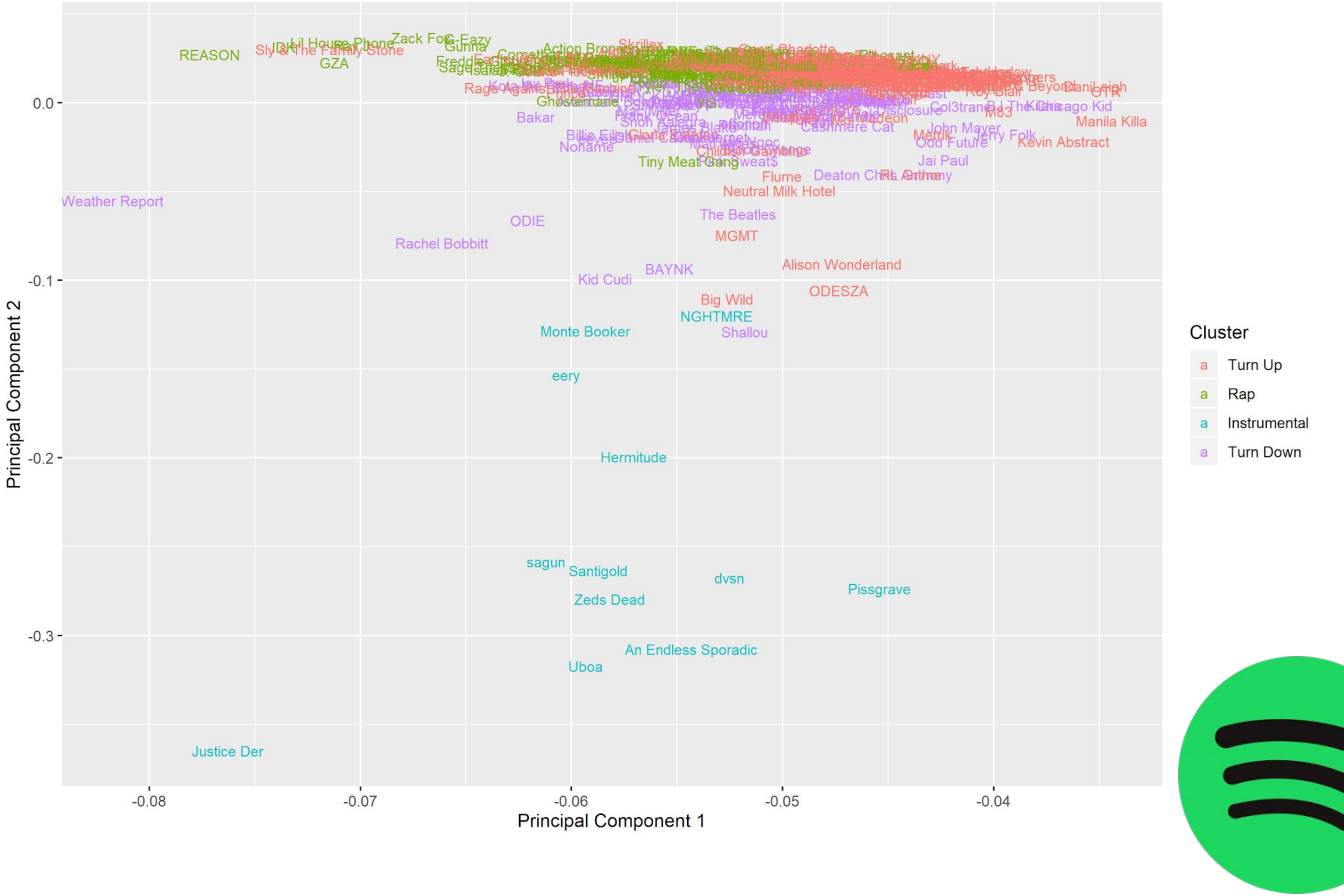


- High Acousticness
- Low everything else
- Separates acoustic instrumentation artists from artists that use other instruments

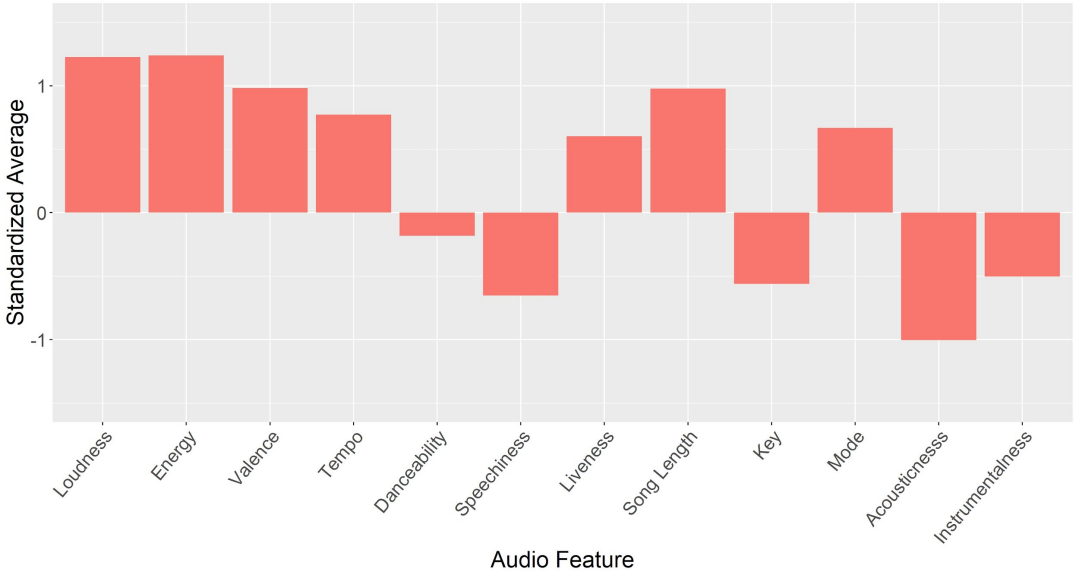
Clusters


Speechiness
Valence
Song Length
Acousticness
Loudness
Instrumentalness
Principal Component 2
Principal Component 1
0.0
-0.1
-0.2
-0.3
-0.08
-0.04
0.00
0.04
Weather Report
ODIE
The Beatles
MGMT
Rachel Bobbitt
Alison Wonderland
Kid Cudi
ODESZA
Big Wild
NGHTMRE
Monte Booker
eery
Hermitude
sagun
Santigold
dvsn
Pissgrave
Zeds Dead
An Endless Sporadic
Uboa
Justice Der
Manila Killa
Cluster
Turn Up
Rap
Instrumental
Turn Down

Clusters

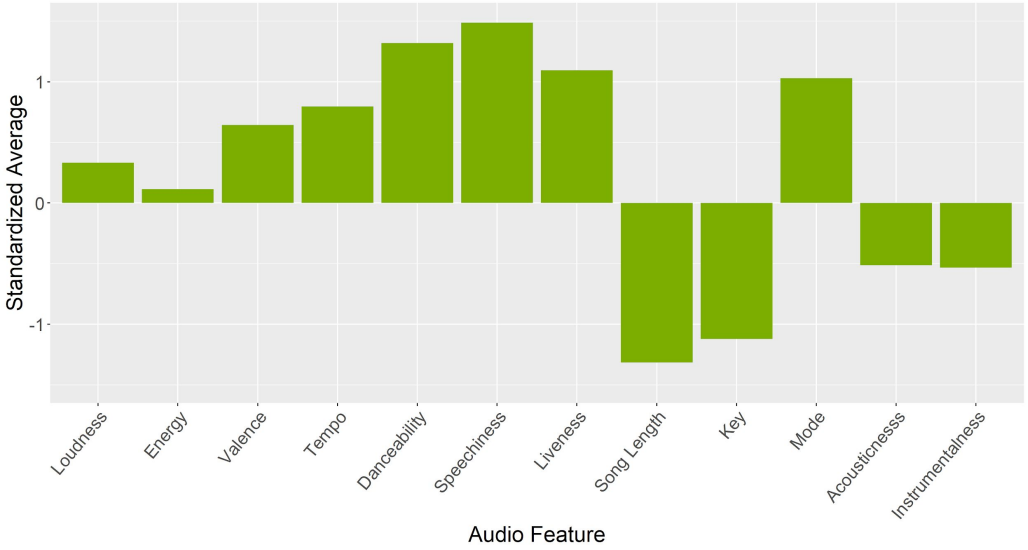

Principal Component 2
0.0
-0.1
-0.2
-0.3
-0.08
-0.07
-0.06
-0.05
-0.04
Principal Component 1
REASON
GZA
Zack Fox
Fazy
Gunna
Bakar
Manila Killa
Kevin Abstract
Jai Paul
Weather Report
ODIE
Rachel Bobbitt
The Beatles
MGMT
Neutral Milk Hotel
Flume
Alison Wonderland
ODESZA
BAYNK
Kid Cudi
Big Wild
NGHTMRE
Shallou
Monte Booker
eery
Hermitude
sagun
Santigold
Zeds Dead
dvsn
Pissgrave
An Endless Sporadic
Uboa
Justice Der
Cluster
Turn Up
Rap
Instrumental
Turn Down

Turn Up



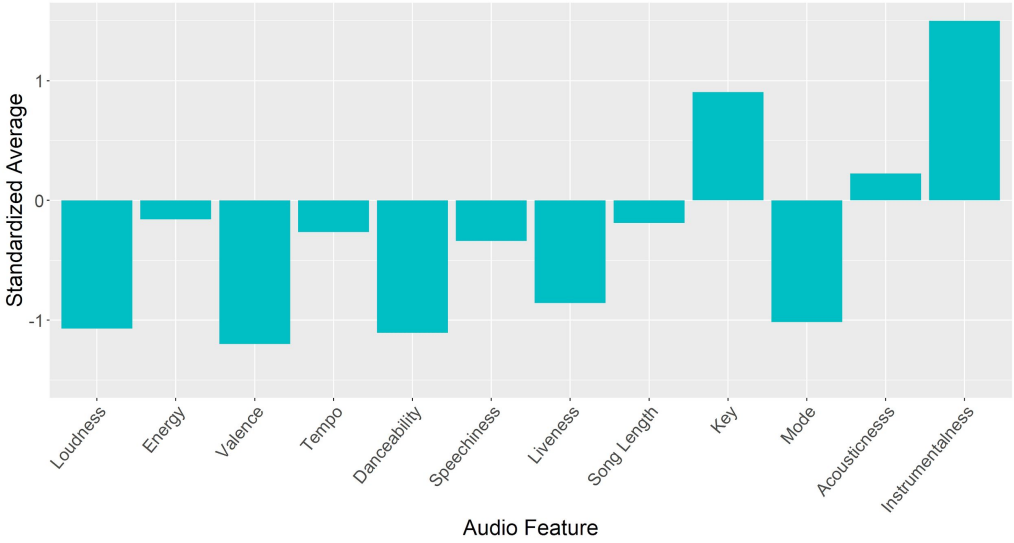
- High Loudness, Energy, Valence
- Low Acousticness, Speechiness, Instrumentalness
- Key members: Taylor Swift, Rage Against the Machine, Flume

Rap



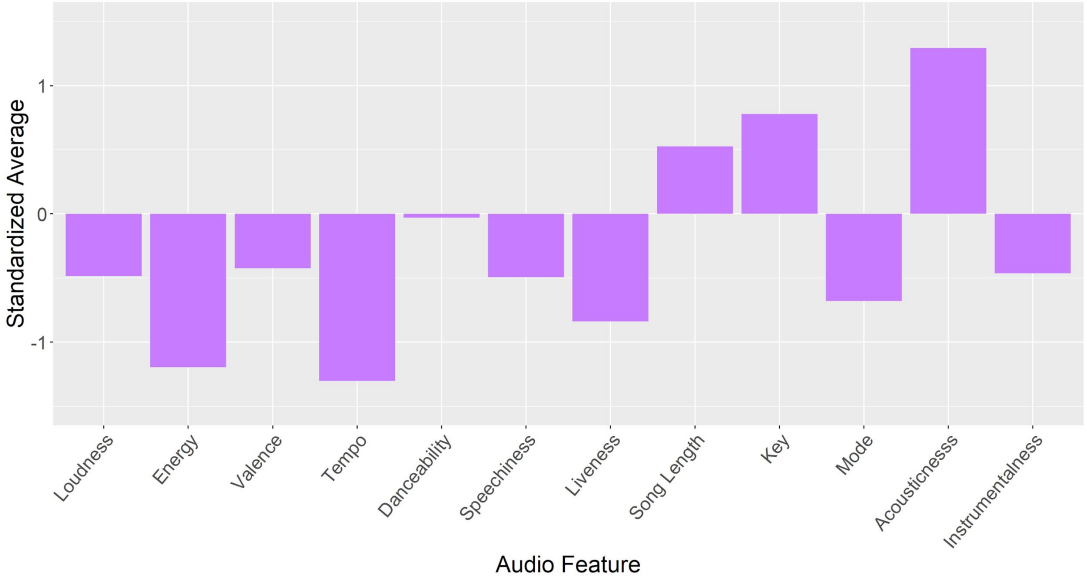
- High Speechiness, Danceability, Liveness
- Low Song Length and Key
- Key members: Kendrick Lamar, J. Cole, Wiz Khalifa

Instrumental



- Very high Instrumentalness and key
- Low everything else
- Key members: Justice Der, eery, Hermitude

Turn Down



- High Acousticness, Key, Song Length
- Low everything else
- Key members: Frank Ocean, Daniel Caesar, SZA

Main Takeaways

- The best way to differentiate my favorite artists is by separating the loud artists from the quiet artists
- The second-best way to differentiate my favorite artists is by separating the instrumental artists from the artists that use vocals

Turn up	Rap	Instrumental	Turn down
Loud Energetic Happy	Lyrical Danceable	Instrumental	Acoustic Slow Soft

Conclusion

- Clustering is a great way to understand and visualize the structure of a big set of data
- One way to cluster data points is by using Principal Components Analysis combined with K-means clustering
    - PCA emphasizes variation in the data, which makes natural groupings more apparent
    - K-means clustering decides what the groups are
- With a better understanding of our data set, our future analysis and decision-making is smarter and more informed