

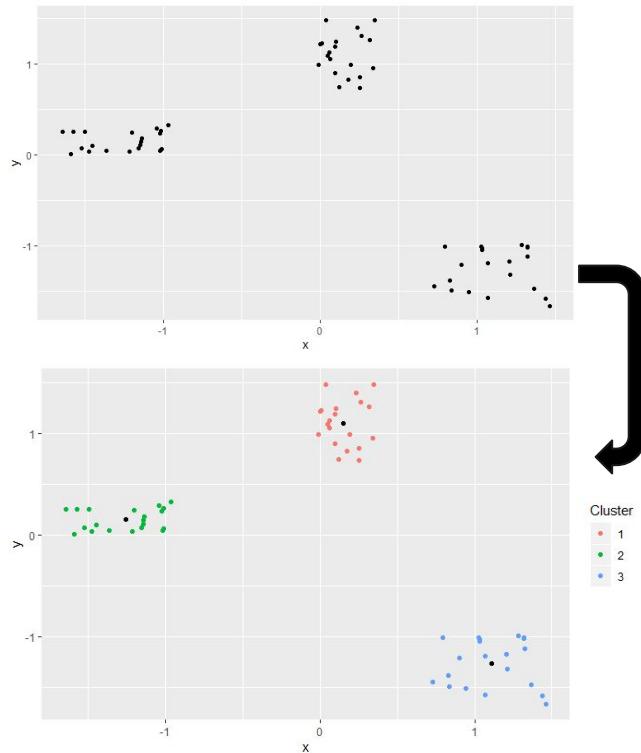


Clustering With PCA and K-means

Brandon Chan

What is clustering?

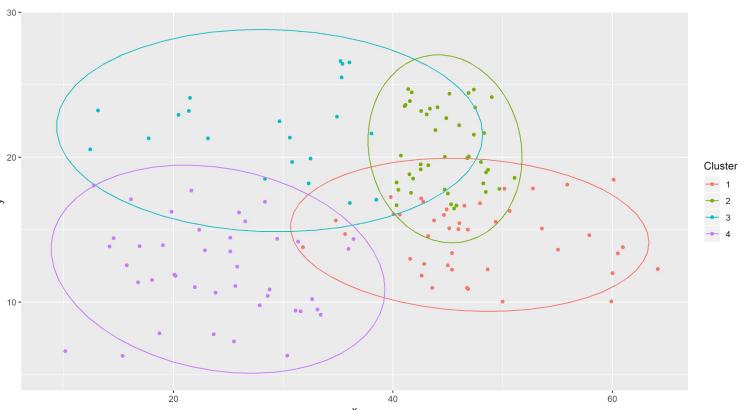
- Clustering is when you group similar data points together
- One group of similar data points makes up a cluster
- One key characteristic of clustering is that the groups/clusters are not predefined



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

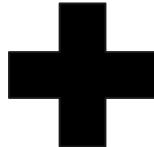
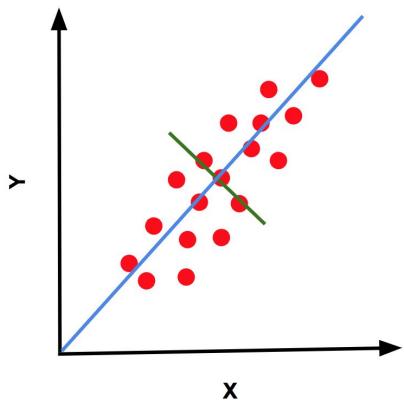
	nameShort	gr	pos	pcfFrequency	phi	tfs	fpm	pp	ppw	ppwTeam	ppwTeamS	ppwPerCent	phiPerGame	phiPerGame	ppwPerGame	ppwPerGame	ppwPerGameS	fpmTotal	ppwTotal	phiTotal	ppwTotal	phiTotal
ATL	Atlanta	81	207	0.32773209	170	179	81	0.02121209	104	301	2,3555156	2,30976423	0.73030423	1,45679102	1,45679102	1,45679102	116	0.44701012	0.4446			
SAC	Sacramento	82	298	0.16751702	345	274	118	1.165549	428	59	1,89701010	2,41751707	3,34145141	1,41616516	1,42659262	1,42659262	158	0.43335766	0.803			
ATL	Atlanta	83	148	0.09860540	175	152	83	1.182430	445	60	1,89300000	2,41750000	3,34145140	1,41616516	1,42659260	1,42659260	70	0.4495097	0.579			
TOR	Toronto	84	159	0.16751702	180	152	81	1.165549	428	66	2,02030535	2,30976423	0.73030423	1,45679102	1,45679102	1,45679102	99	0.4495097	0.579			
CLE	Cleveland	85	74	0.30989101	64	60	12	0.02121209	145	80	1,71795174	1,86959852	0.69695852	1,45679102	1,45679102	1,45679102	43	0.4495097	0.579			
CLE	Cleveland	86	148	0.12050103	21	19	7	1.121051	491	14	1,42827474	1,92000000	1,97142327	0.82000000	1,97142327	1,97142327	8	0.4495097	0.579			
TOR	Toronto	87	253	0.20968803	269	229	101	1.142320	424	71	1,89200000	2,41620000	3,24162000	1,42630000	1,45630000	1,45630000	125	0.4495097	0.590			
WIC	ChicagoCubs	88	192	0.09709802	163	178	83	0.047917	207	281	2,09889987	2,32777775	0.78700000	1,99773223	1,99773223	1,99773223	115	0.3935208	0.481			
CHI	ChicagoCubs	89	46	0.14715800	45	43	17	0.076705	244	263	2,09300009	2,40454545	1,93454545	0.77272727	1,81678168	1,81678168	26	0.3935208	0.523			
LAL	Lakers	90	123	0.21317203	68	118	33	0.076705	90	405	2,37671238	2,38916958	2,51069360	0.7021768	0,80361644	0,80361644	65	0.7199102	0.415			
F	GSW	91	143	0.20237765	169	162	59	0.076705	242	263	2,13891508	2,38916958	2,51069360	0.7021768	0,80361644	0,80361644	55	0.4495097	0.590			
BOS	Celtics	92	49	0.20237765	47	45	17	0.076705	244	250	2,13891508	2,38916958	2,51069360	0.7021768	0,80361644	0,80361644	24	0.4495097	0.568			
MIN	Timberwolves	93	77	0.16277302	273	216	80	1.080300	388	156	1,23770023	1,56545453	2,02191001	1,31310103	1,82278762	1,82278762	125	0.4120050	0.585			
TOR	Toronto	94	160	0.39711893	173	179	80	0.086173	155	340	2,89597101	2,55208597	0.95523929	1,75154216	1,75154216	1,75154216	102	0.3404009	0.471			
LAL	Lakers	95	168	0.30968806	162	158	58	0.086173	219	277	2,52380006	2,41750145	2,35820000	1,52238000	1,52238000	1,52238000	102	0.3404009	0.523			
WAS	Washington	96	266	0.15203400	309	241	110	1.118105	425	60	2,34903044	3,76526168	2,99024029	1,34164041	1,97763000	1,97763000	131	0.45617614	0.585			
NOR	NewOrleans	98	38	0.31462001	45	35	18	1.142320	447	45	2,11111111	2,60000000	1,64444444	1,20000000	0,64444444	0,64444444	17	0.14202571	0.585			
MEM	Memphis	99	41	0.20947605	58	52	19	0.090520	208	287	1,78174428	1,87174428	1,87174428	0.84444444	0,84444444	0,84444444	13	0.2419987	0.587			
IND	Indiana	100	24	0.20947605	240	24	13	0.090520	68	437	1,95714248	1,42827474	0.82000000	1,24282771	1,71423271	1,71423271	7	0.2419987	0.548			
CHI	Chicago	101	142	0.09650000	206	148	48	1.171051	478	18	2,00000000	2,54220968	1,8390917	0.8390917	1,00000000	0,81	0,4495097	0.547				
POR	Portland	102	214	0.14894603	233	202	64	1.080705	381	114	3,97140398	3,32687413	2,48671429	1,20000000	1,60774209	1,60774209	110	0,4495097	0.551			
ATL	Atlanta	103	101	0.34848101	94	69	30	0.059080	100	356	2,04666666	2,08088800	1,97777777	0,86666667	1,31111101	1,31111101	59	0,33707965	0,477			
TOR	Toronto	104	158	0.11095000	181	148	66	1.145971	427	66	2,03330303	3,01066687	2,66333303	1,10000000	1,32233323	1,32233323	80	0,43204749	0,571			
DAL	Dallas	28	88	0.42947717	101	79	36	1.147705	420	65	2,05464278	3,46279582	1,54719511	1,46279582	43	0,4595002	0,564					
LAL	Lakers	4	4	0.34905000	6	2	2	1.121051	491	2	2,05464278	3,46279582	1,54719511	1,46279582	43	0,4595002	0,547					
NYC	NewYork	73	204	0.35163300	241	191	76	0.02121209	311	184	2,74450055	3,90121966	1,64140456	1,32701704	1,68604010	1,68604010	118	0,3617014	0,528			
WAS	Washington	77	208	0.33000434	310	258	103	1.172051	385	130	3,75334075	3,02087403	1,32447132	1,37106214	1,96719349	1,96719349	153	0,3333475	0,539			
WAS	Washington	80	210	0.20504001	233	187	79	1.119051	399	95	2,83200000	2,93770000	0,98700000	1,32700000	1,35500000	1,35500000	108	0,4424998	0,564			
MIN	Minnesota	85	172	0.53700000	164	150	50	0.058480	212	283	2,04913085	2,52707862	2,52707862	1,79820777	1,79820777	1,79820777	100	0,33333303	0,483			
MDM	Memphis	86	162	0.33176872	135	141	48	0.08202303	163	260	3,47727273	3,36618162	3,24044545	1,90000009	2,11568598	2,11568598	92	0,4042553	0,485			



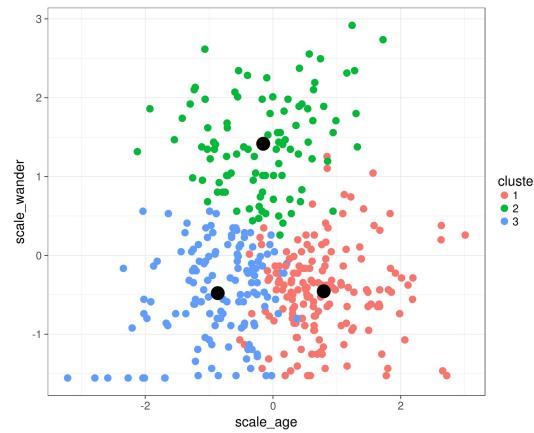


How?

Principal Component Analysis

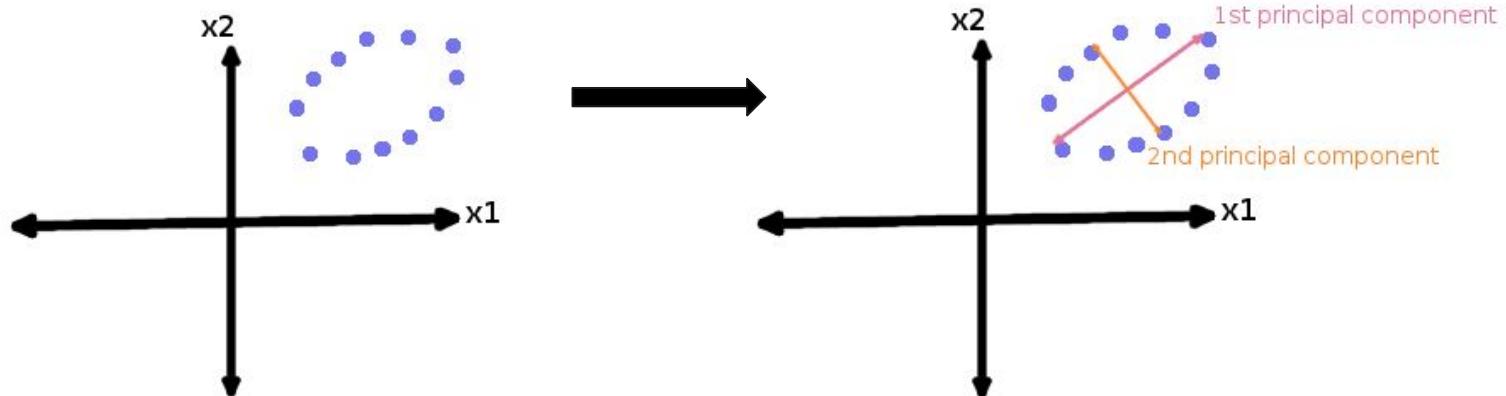


K-means clustering



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

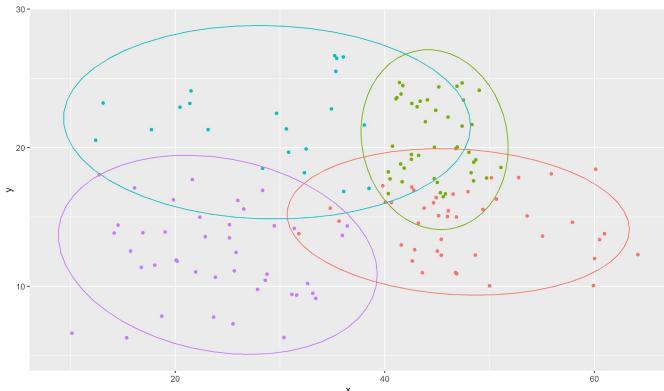
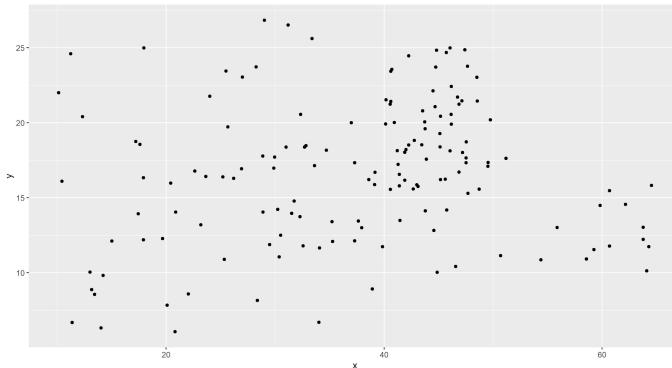
- Each axis in the new coordinate system is known as a Principal Component
- Each Principal Component represents one way that the data points can be transformed
- Each transformation results in a different amount of variation among the data points
- The Principal Components are ordered from the highest 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:
 - The user decides how many clusters they want to have
 - The k-means algorithm decides where the centers of those clusters should be
 - Every data point is assigned to the nearest cluster



The Data

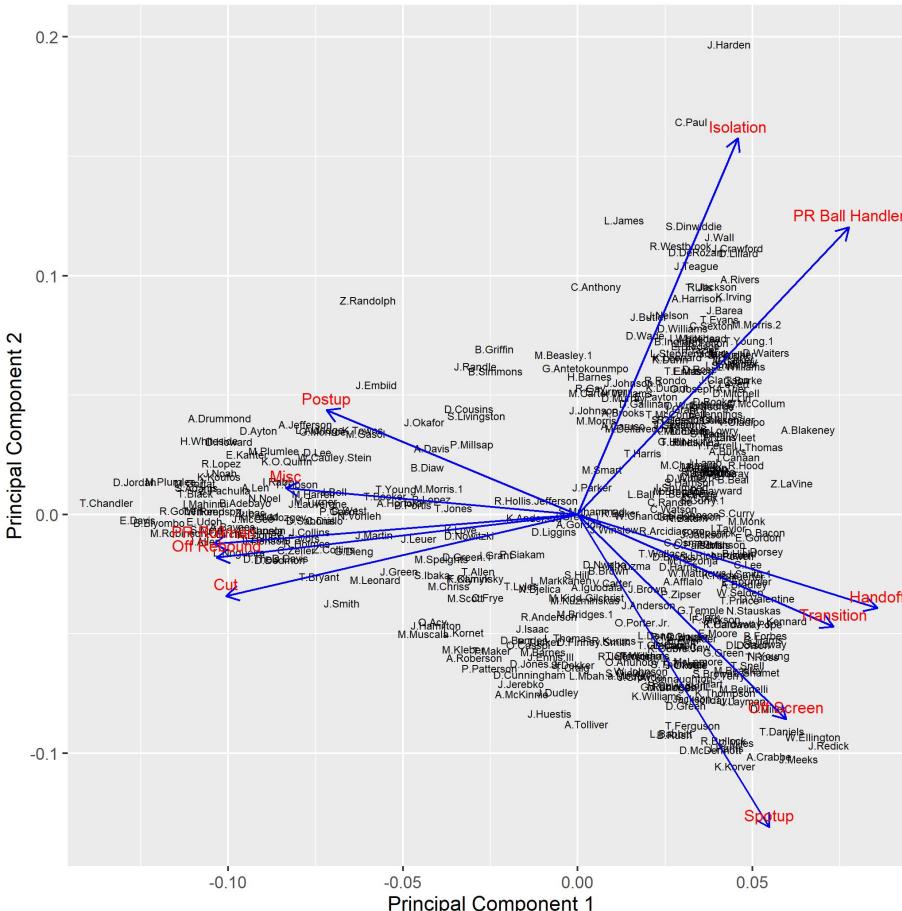
One data point - One NBA Player who played at least 1000 minutes in the past 3 seasons

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

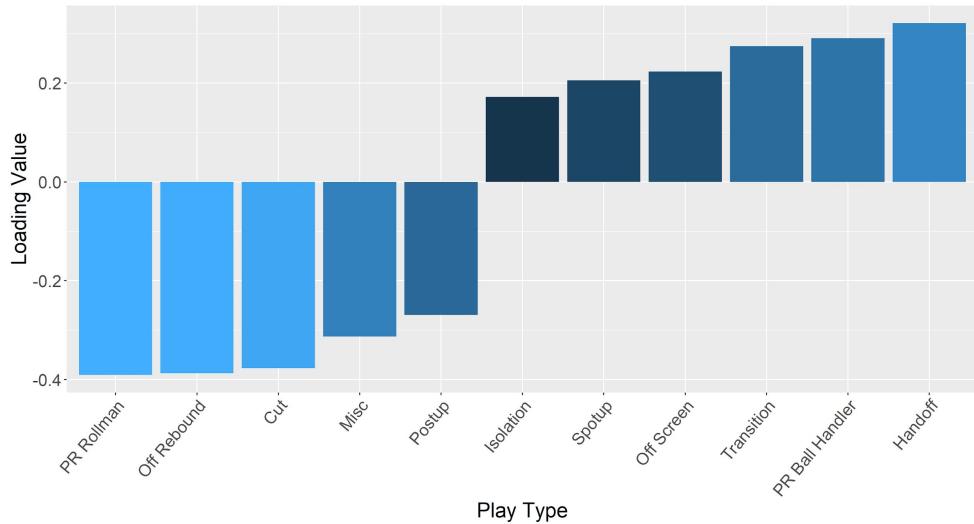


Biplot





PC1 Loadings

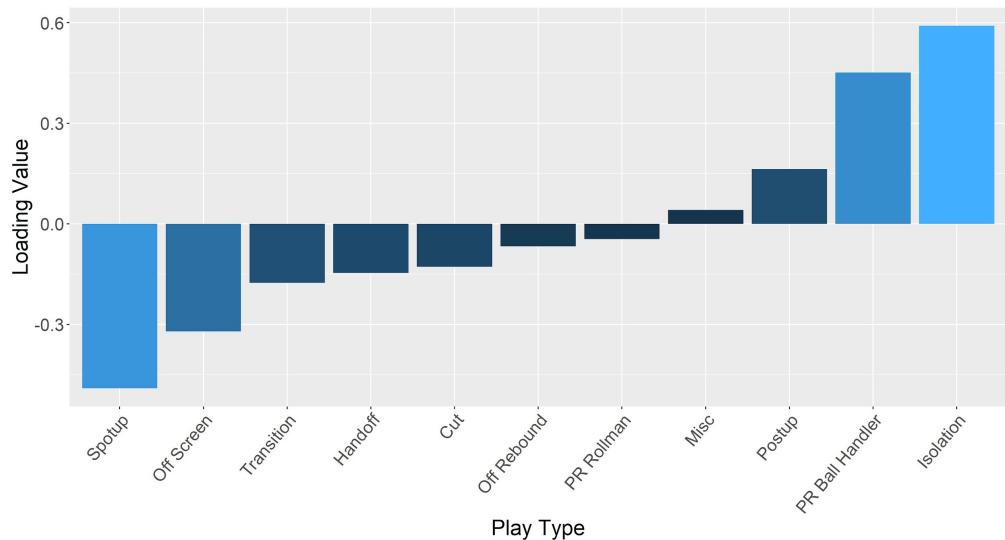


- High Handoff, PB Ball Handler, and Transition
- Low PR Rollman, Off Rebound, and Cut
- Separates big men from guards and forwards





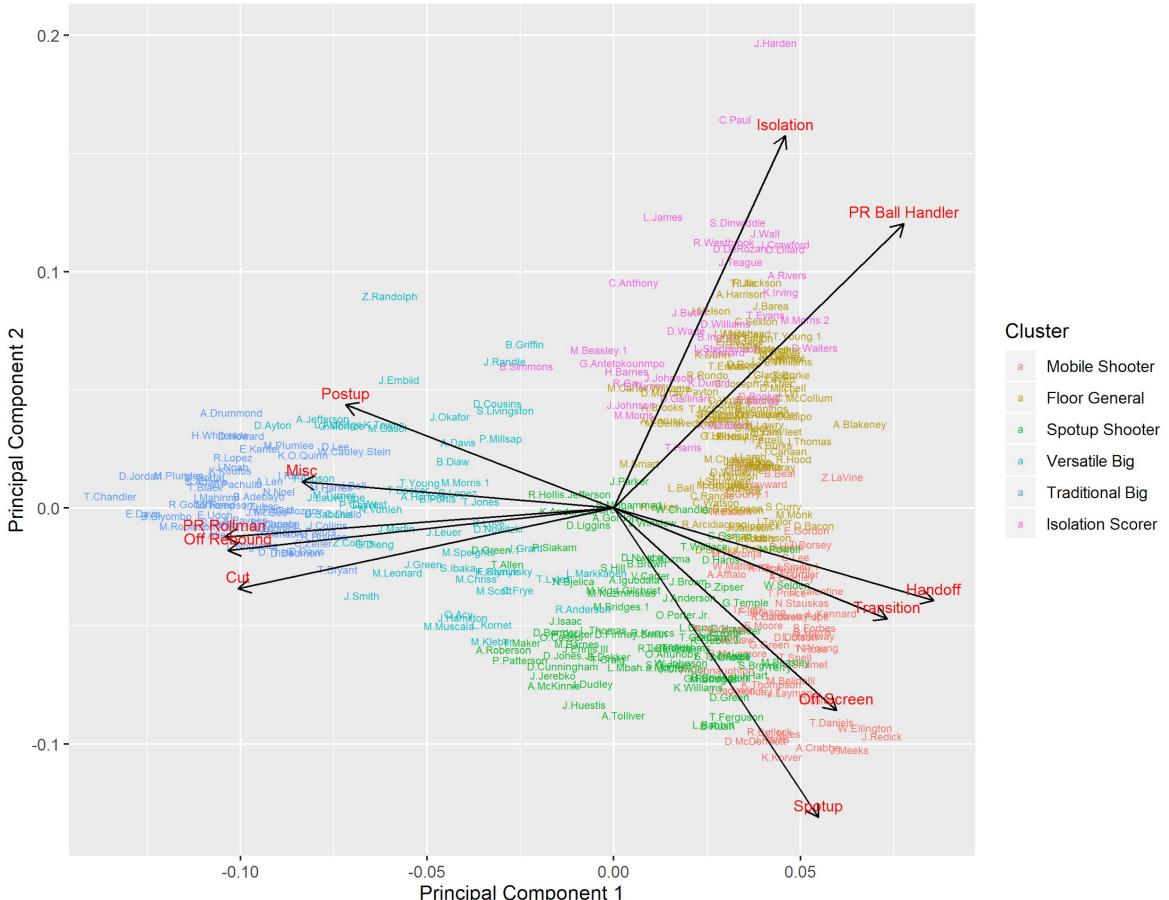
PC2 Loadings



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

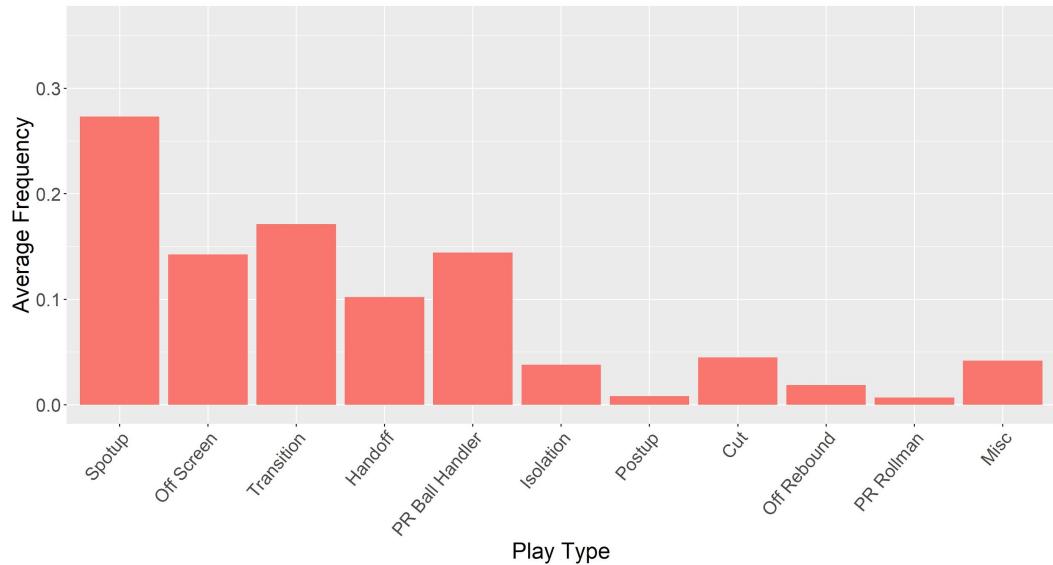


Clusters





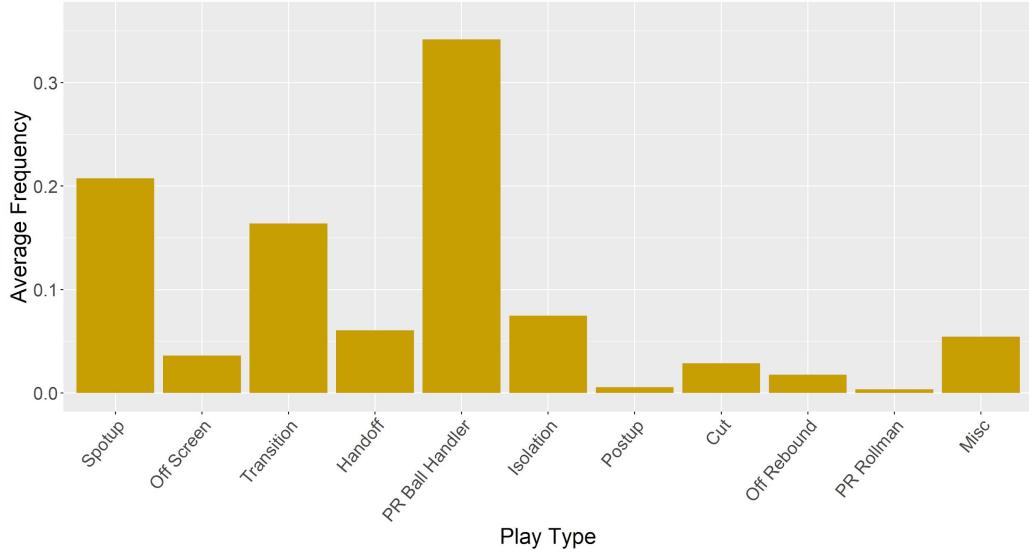
Mobile Shooters



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



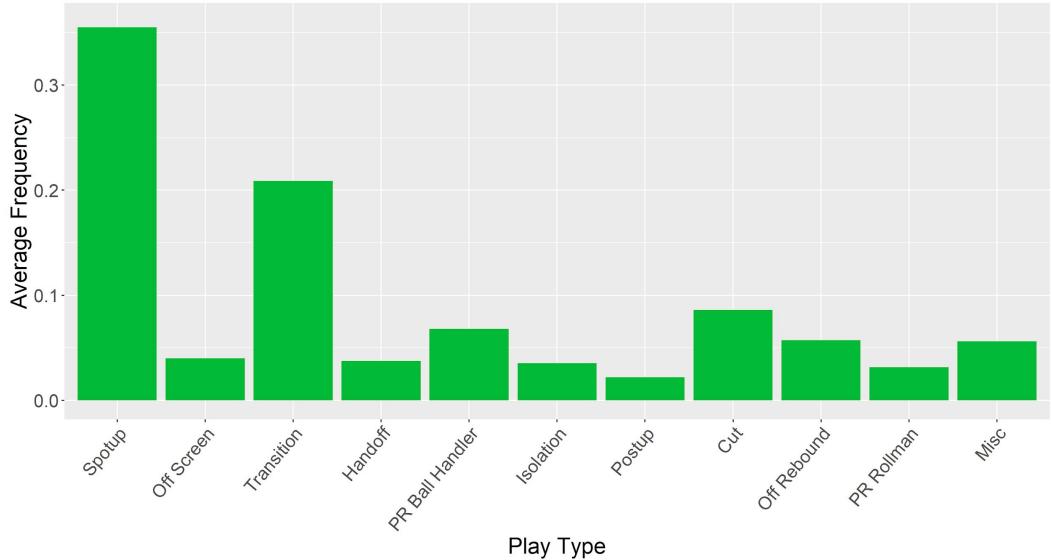
Floor Generals



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



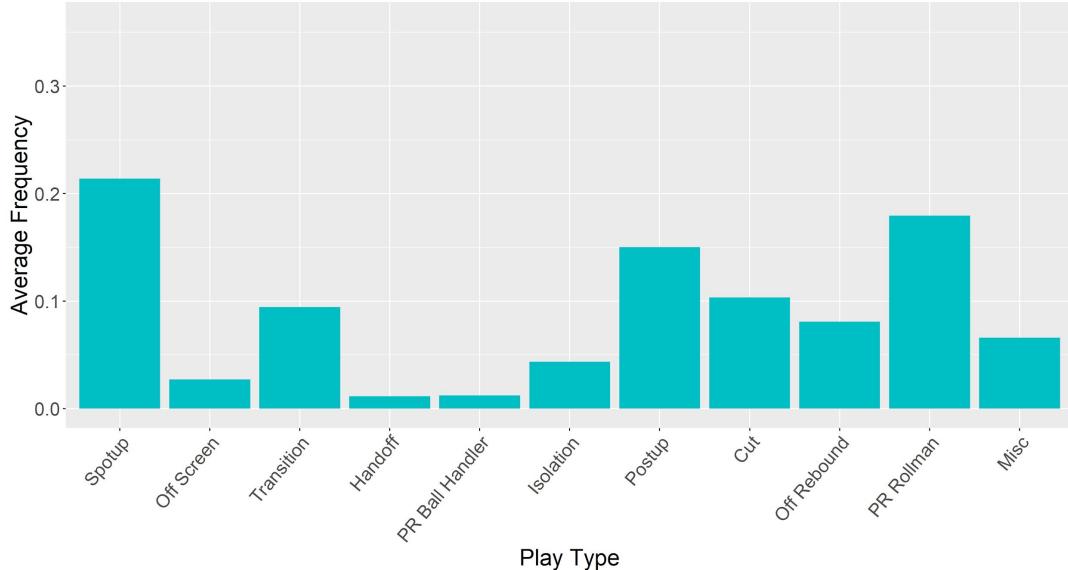
Spotup Shooters



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



Versatile Bigs

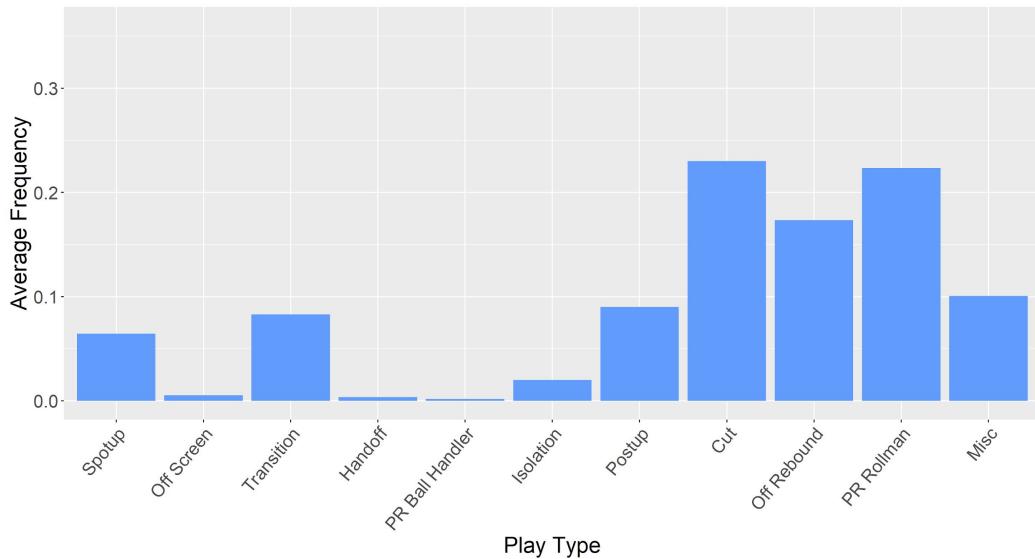


- High Spotup, PR Rollman, Postup
- Low Handoff, PR Ball Handler, Off Screen
- Key members: Demarcus Cousins, Anthony Davis, Blake Griffin





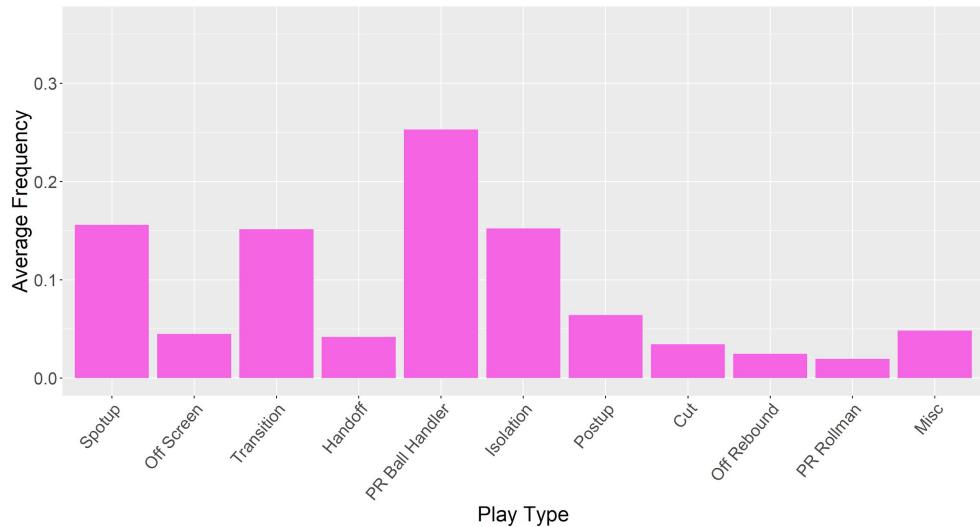
Traditional Bigs



- 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

Postups
Being the rollman
Grabbing rebounds
Cuts
Miscellaneous

Ball Handlers

Isolation
Running the pick and roll

Shooters

Spotting up
Running in Transition
Handoffs
Running off screens



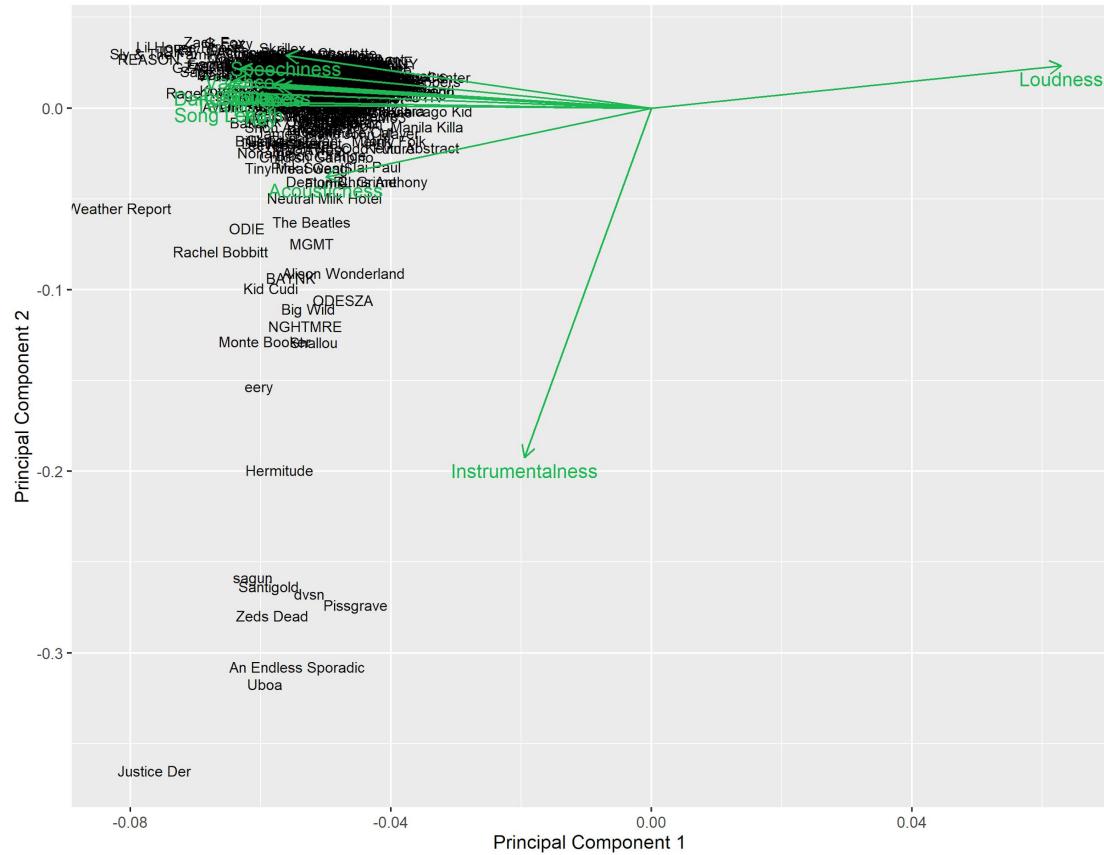
The Data

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:

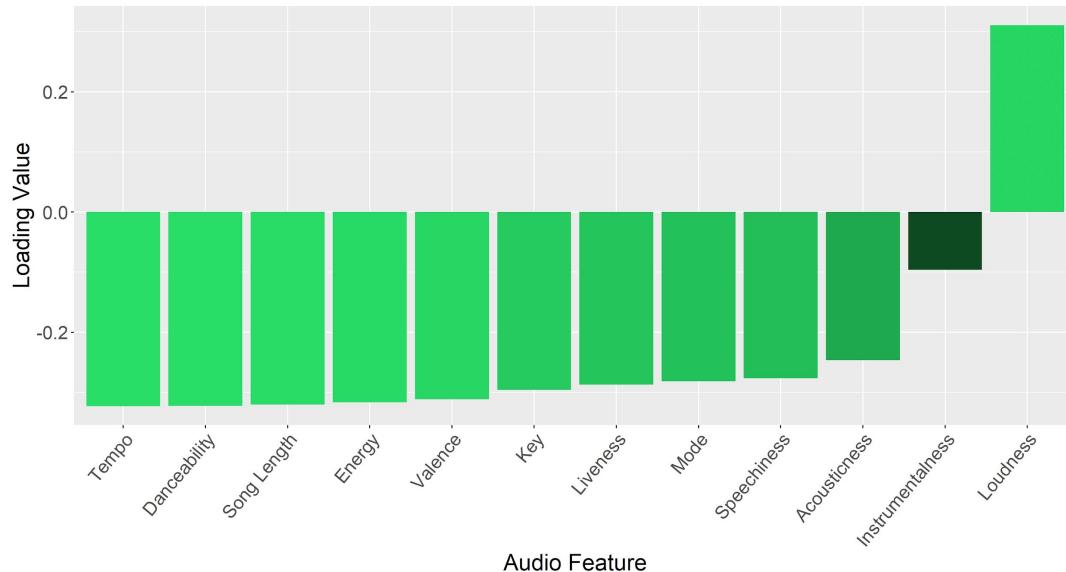
- Tempo
- Danceability
- Song Length
- Energy
- Valence
- Key
- Liveness
- Mode
- Speechiness
- Acousticness
- Instrumentalness
- Loudness







PC1 Loadings

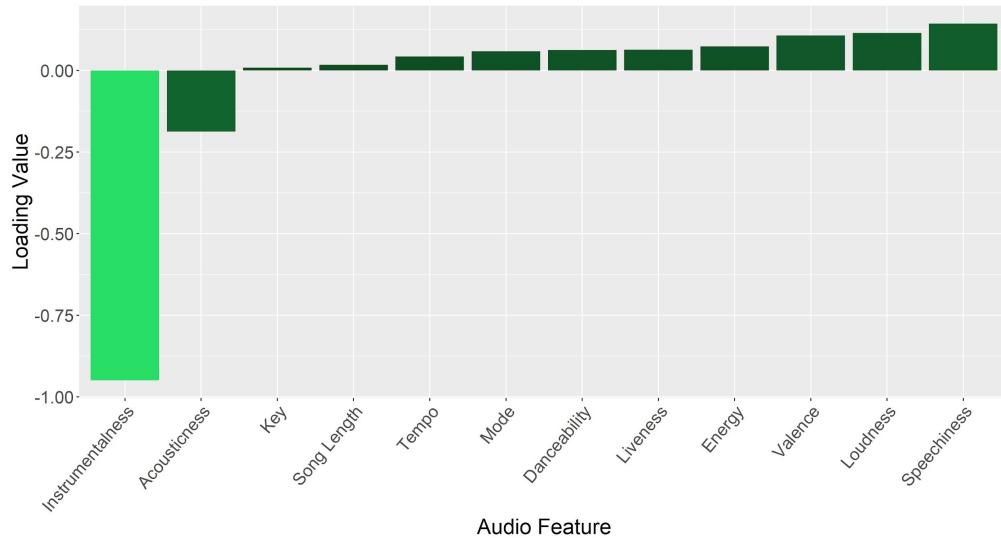


- High Loudness
- Low everything else
- Separates loud artists from not loud artists





PC2 Loadings

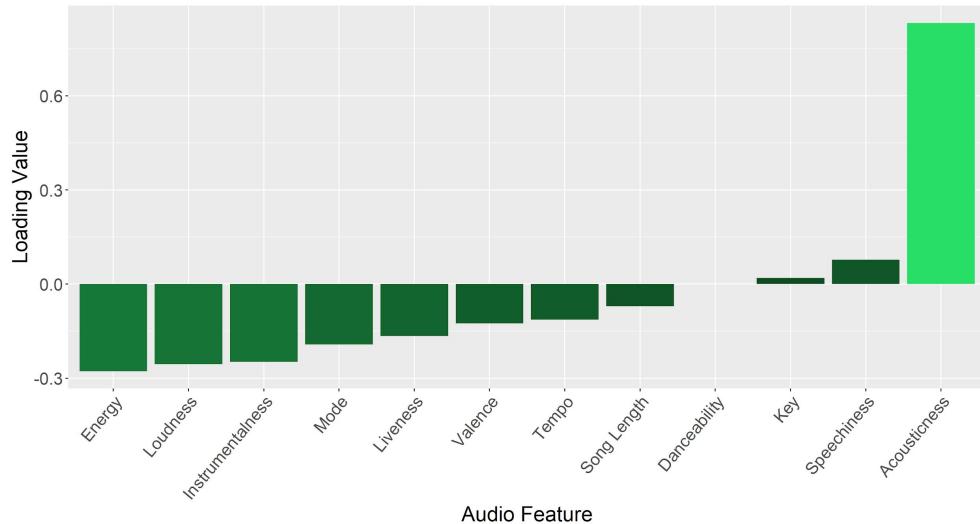


- High most variables
- Very low Instrumentalness
- Separates artists who play instrumentals and artists that use vocals





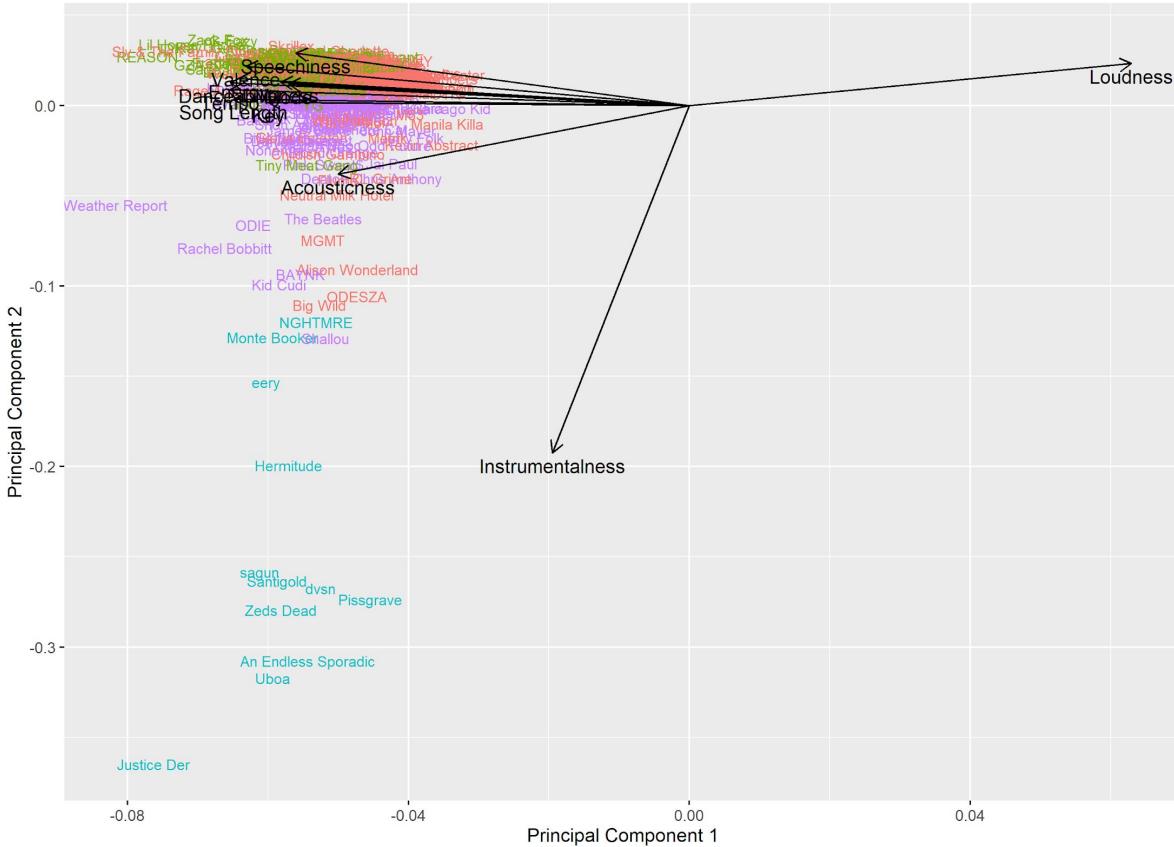
PC3 Loadings



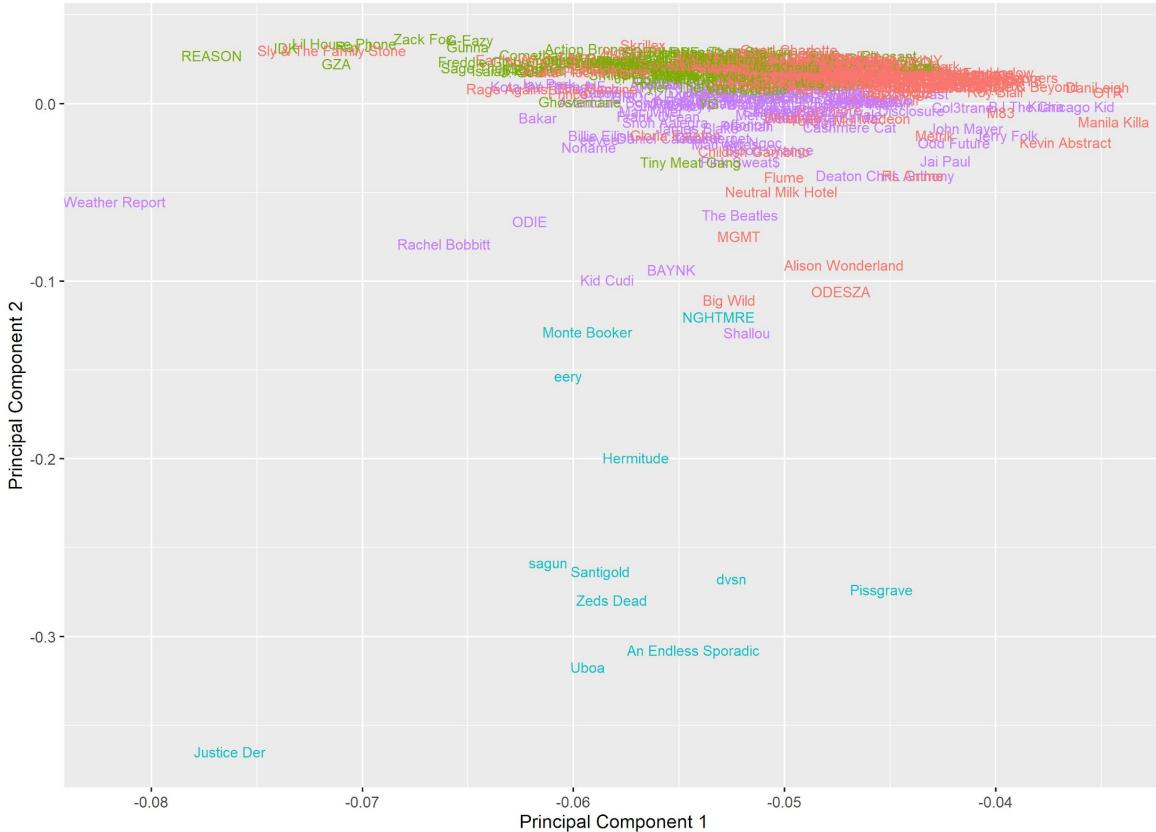
- High Acousticness
- Low everything else
- Separates acoustic instrumentation artists from artists that use other instruments



Clusters

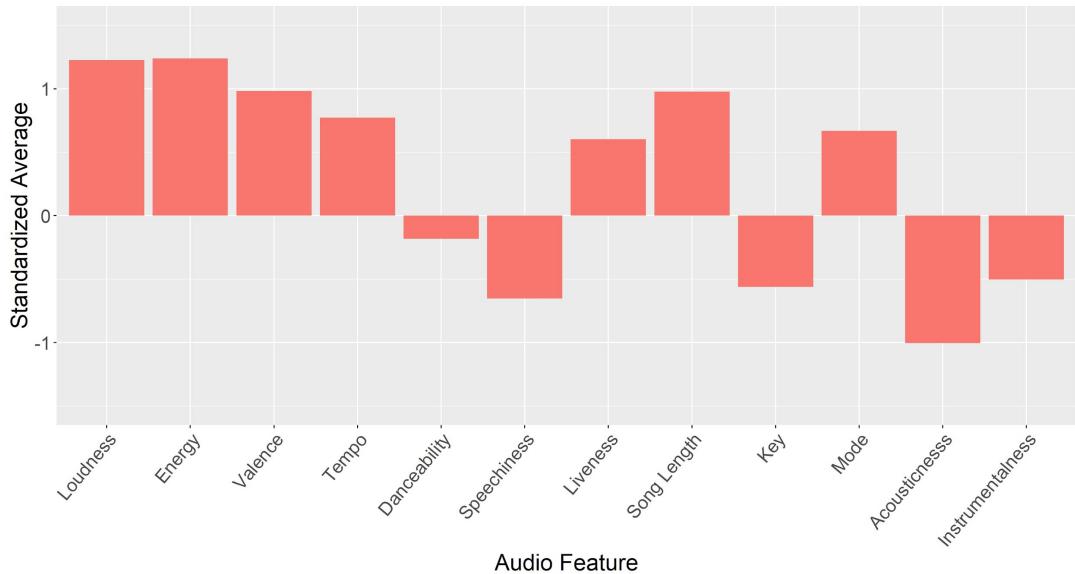


Clusters





Turn Up

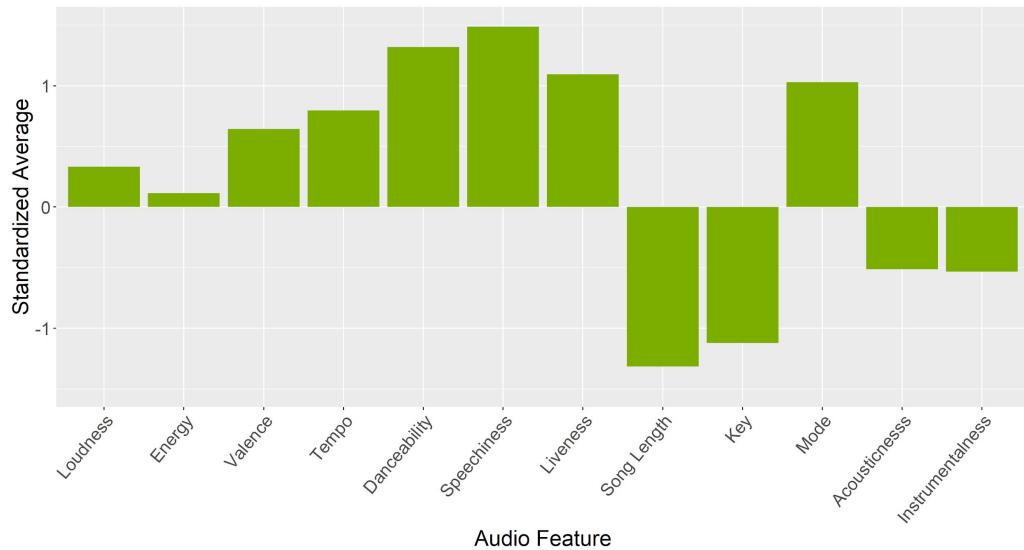


- High Loudness, Energy, Valence
- Low Acousticness, Speechiness, Instrumentalness
- Key members: Taylor Swift, Rage Against the Machine, Flume





Rap

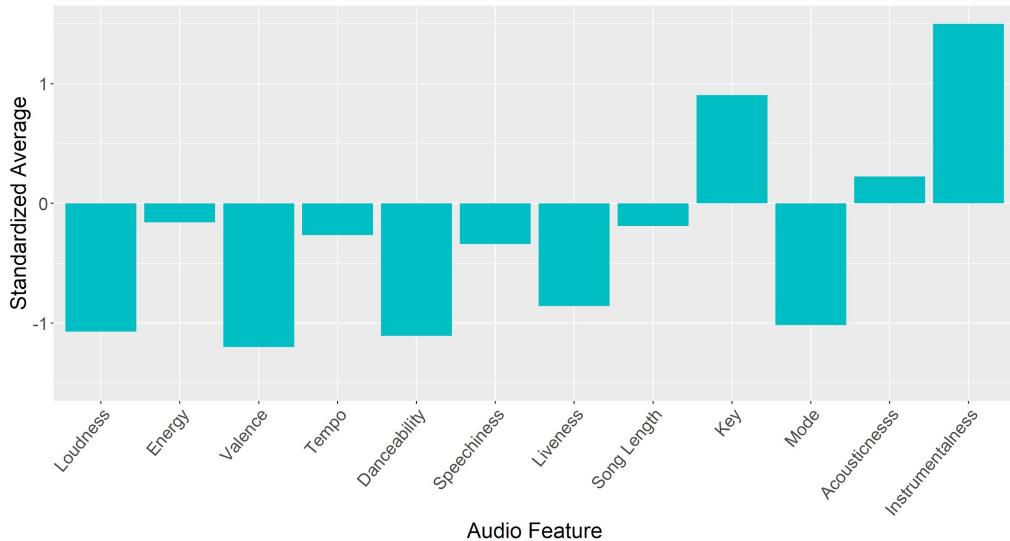


- High Speechiness, Danceability, Liveness
- Low Song Length and Key
- Key members: Kendrick Lamar, J. Cole, Wiz Khalifa





Instrumental

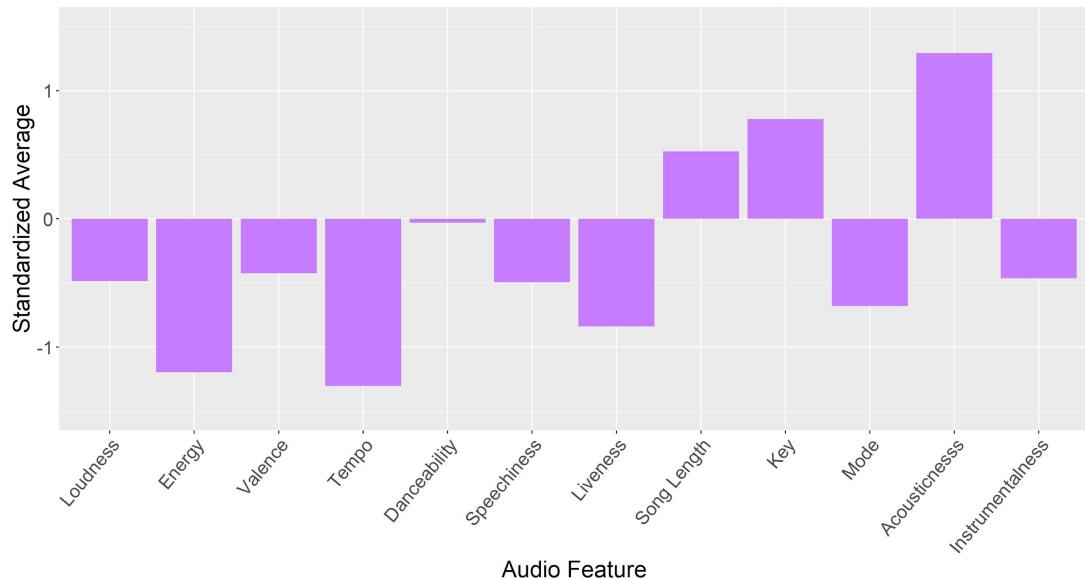


- Very high Instrumentalness and key
- Low everything else
- Key members: Justice, Derby, Hermitude





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

Loud
Energetic
Happy

Rap

Lyrical
Danceable

Instrumental

Instrumental

Turn down

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

