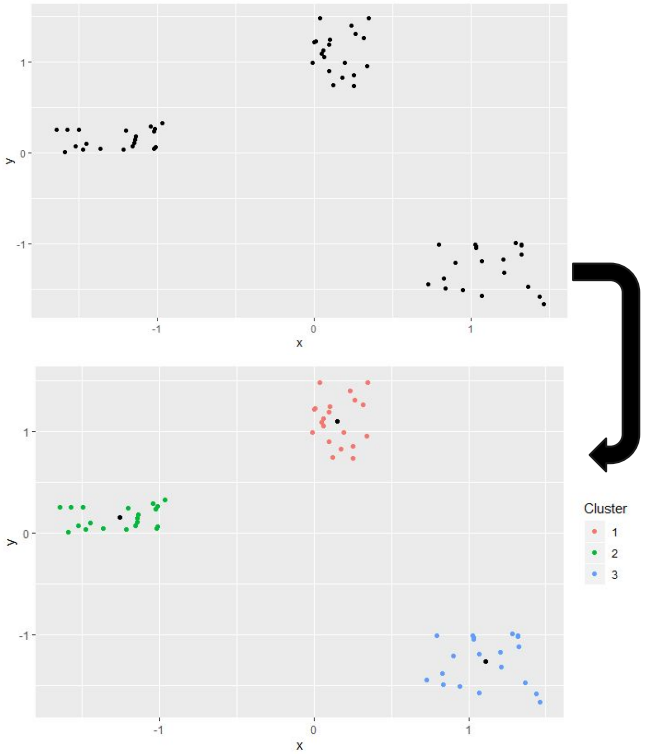
Clustering With PCA and K-means

Brandon Chan

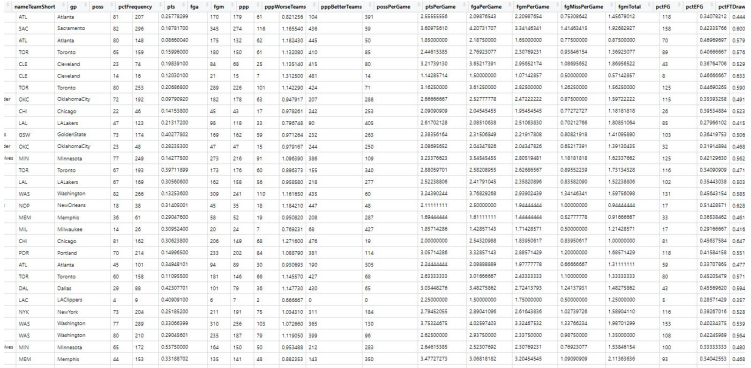
What is clustering?

- 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

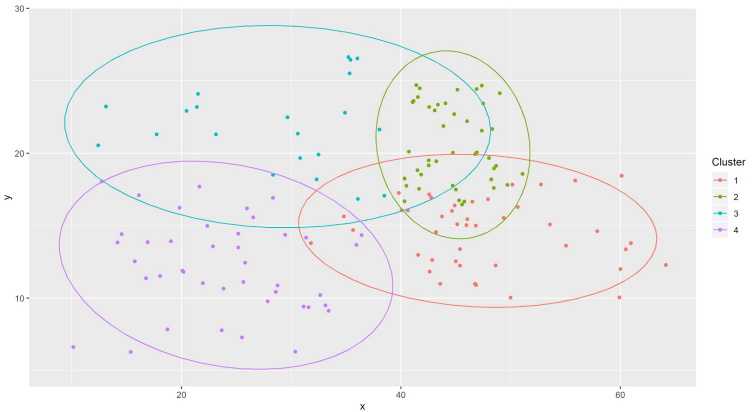


Why cluster?

- 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
- Also, by grouping together similar data points we can:
    - Find relationships among variables
    - Spot outliers
    - Uncover underlying patterns in the data

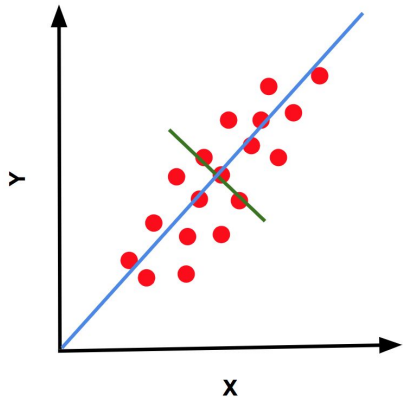


	nameTeamShort	gp	poss	pctFrequency	pts	fga	fgm	ppp	ppgWorseTeams	ppgBetterTeams	possPerGame	ptsPerGame	fgaPerGame	fgmPerGame	fgMissPerGame	fgmTotal	pctFG	pctEFG	pctFTDue
	ATL	Atlanta	81	207	0.25778299	170	179	61	0.621258	104	391	2.09976543	2.09976543	2.20987654	0.77530642	1.45679012	116	0.34079212	0.444
	SAC	Sacramento	82	206	0.16761700	345	274	118	1.165540	436	59	1.80975610	4.20731707	3.34146341	1.41455415	1.92682927	158	0.42333765	0.600
	ATL	Atlanta	80	148	0.08660040	175	132	62	1.182430	445	50	1.85000000	2.18750000	1.66000000	0.77500000	0.87500000	70	0.46996997	0.579
	TOR	Toronto	65	159	0.15096000	180	150	61	1.132080	410	95	2.44815385	2.76923077	2.30769231	0.93946154	1.36923077	89	0.40666667	0.576
	CLE	Cleveland	23	74	0.19939100	84	66	25	1.139140	415	60	3.21739130	3.65217391	2.86956522	1.08695652	1.56956522	43	0.37878706	0.529
	CLE	Cleveland	14	16	0.12032100	21	15	7	1.312500	481	14	1.14285714	1.50000000	1.07142857	0.92000000	0.57142857	6	0.46666667	0.833
	TOR	Toronto	80	253	0.20568000	289	226	101	1.142290	424	71	3.16250000	3.61250000	2.82500000	1.26250000	1.58250000	125	0.44692265	0.590
ter	OKC	OklahomaCity	72	192	0.09790920	162	178	63	0.947917	207	288	2.66666667	2.52777778	2.47222222	0.87500000	1.59722222	115	0.35393258	0.491
	CHI	Chicago	22	46	0.14153000	45	43	17	0.978261	242	253	2.09090909	2.04545455	1.95454545	0.77272727	1.18181818	26	0.39534884	0.523
	LAL	LALakers	47	123	0.21317200	98	110	33	0.796748	90	405	2.61702128	2.09510638	2.51063830	0.70212766	1.80851064	88	0.27999102	0.415
s	GSW	GoldenState	73	174	0.40277302	169	162	59	0.971264	232	260	2.38356164	2.31506849	2.21917808	0.80821918	1.41095890	103	0.36419753	0.509
ter	OKC	OklahomaCity	23	48	0.28233300	47	47	15	0.979167	244	250	2.08695652	2.04347826	2.04347826	0.65217391	1.39130435	32	0.31914894	0.466
ives	MIN	Minnesota	77	249	0.14277500	273	216	91	1.096390	386	109	3.23376623	3.54545455	2.80519481	1.16181818	1.62337662	125	0.42129630	0.562
	TOR	Toronto	67	193	0.39711399	173	176	60	0.896373	155	340	2.58209701	2.58208955	2.62686567	0.89552239	1.73134328	116	0.34090909	0.471
	LAL	LALakers	67	169	0.30560600	162	158	56	0.958580	218	277	2.52238806	2.41791045	2.35820896	0.83582090	1.52238806	102	0.35443038	0.503
	WAS	Washington	62	266	0.13253600	309	241	110	1.161650	435	60	2.24390244	3.76829268	2.93902439	1.34146341	1.59756098	131	0.45523154	0.595
	NOR	NewOrleans	18	36	0.31405001	45	35	18	1.164210	447	46	2.11111111	2.50000000	1.94444444	1.00000000	0.94444444	17	0.51428571	0.628
	MEM	Memphis	36	61	0.20567500	59	52	19	0.950820	208	287	1.66666667	1.61111111	1.44444444	0.82777778	0.91666667	33	0.35535452	0.481
	MIL	Milwaukee	14	26	0.30952400	20	24	7	0.769231	66	427	1.85714286	1.42857143	1.71428571	0.92000000	1.21428571	17	0.29166667	0.416
	CHI	Chicago	91	162	0.30623800	206	149	68	1.271600	478	19	2.00000000	2.54320988	1.63950617	0.83950617	1.00000000	91	0.45637584	0.647
	POR	Portland	70	214	0.14996500	233	202	84	1.088790	351	114	3.05714286	3.32857143	2.88571429	1.20000000	1.68571429	118	0.41584158	0.551
	ATL	Atlanta	45	101	0.24948101	94	89	30	0.930693	190	305	2.24444444	2.09888889	1.97777778	0.86666667	1.31111111	59	0.33707865	0.477
	TOR	Toronto	60	158	0.11090500	181	146	66	1.145570	427	68	2.83333333	3.01666667	2.43333333	1.10000000	1.33333333	80	0.45205479	0.571
	DAL	Dallas	29	66	0.42307701	101	79	36	1.196720	430	65	3.03448276	3.48275862	2.72413793	1.24137931	1.48275862	43	0.45599532	0.594
	LAC	LAClippers	4	9	0.40909100	6	7	2	0.666667	0	0	1.25000000	1.50000000	1.75000000	0.50000000	1.25000000	5	0.28571429	0.357
	NYK	NewYork	73	204	0.25185200	211	191	75	1.034310	311	164	2.79452055	2.89041096	2.61643836	1.02739726	1.58904110	116	0.39267016	0.528
	WAS	Washington	77	269	0.33066399	310	256	103	1.072860	385	130	3.75324675	4.02597403	3.32467532	1.33766234	1.99701299	153	0.40233475	0.559
	WAS	Washington	80	210	0.29045801	235	187	79	1.119050	399	96	2.92500000	2.93750000	2.33750000	0.96750000	1.39000000	108	0.42245989	0.564
ives	MIN	Minnesota	65	172	0.53750000	164	150	92	0.953488	212	283	2.64615385	2.52307692	2.30769231	0.79230077	1.53846154	100	0.53333333	0.480
	MEM	Memphis	44	163	0.33108702	135	141	48	0.882353	143	350	3.47727273	3.06818182	3.20454545	1.09090909	2.11363636	93	0.34042553	0.468

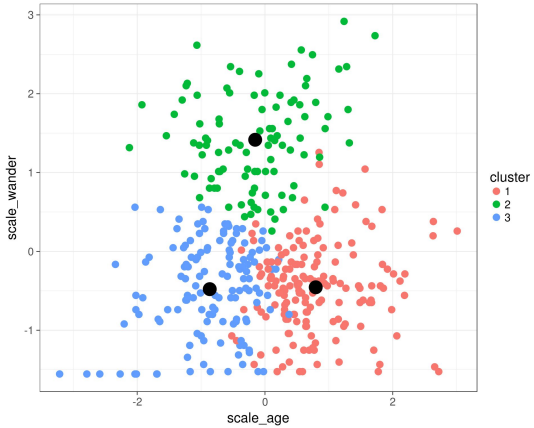


How?

Principal Component Analysis

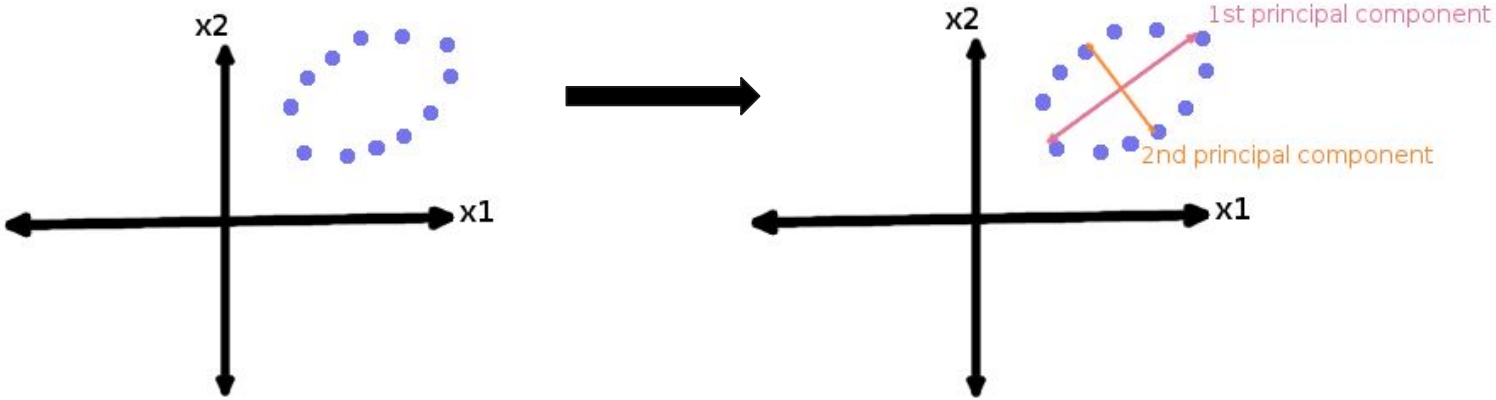


K-means clustering



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



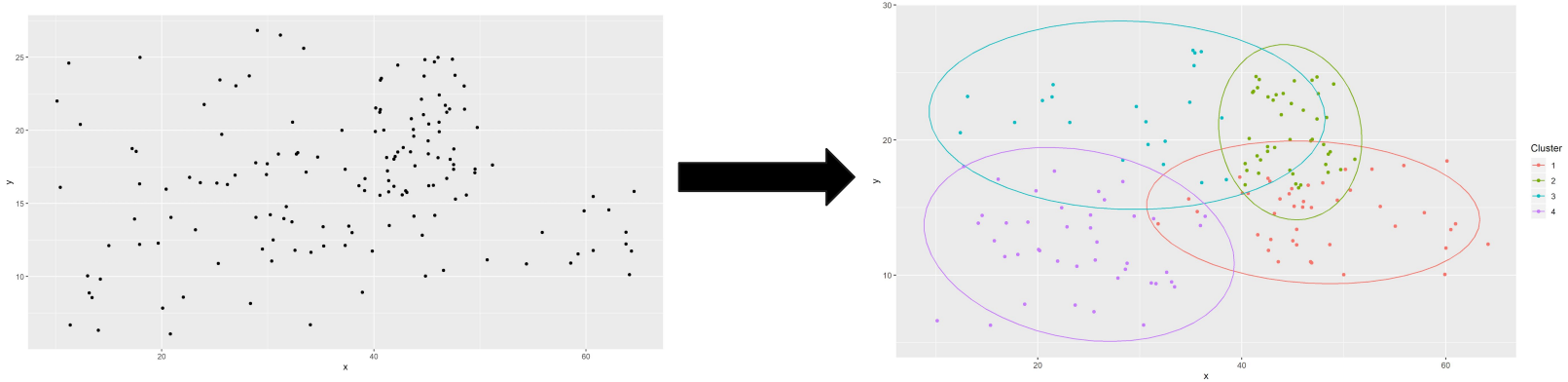
Some Things to Note

- 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



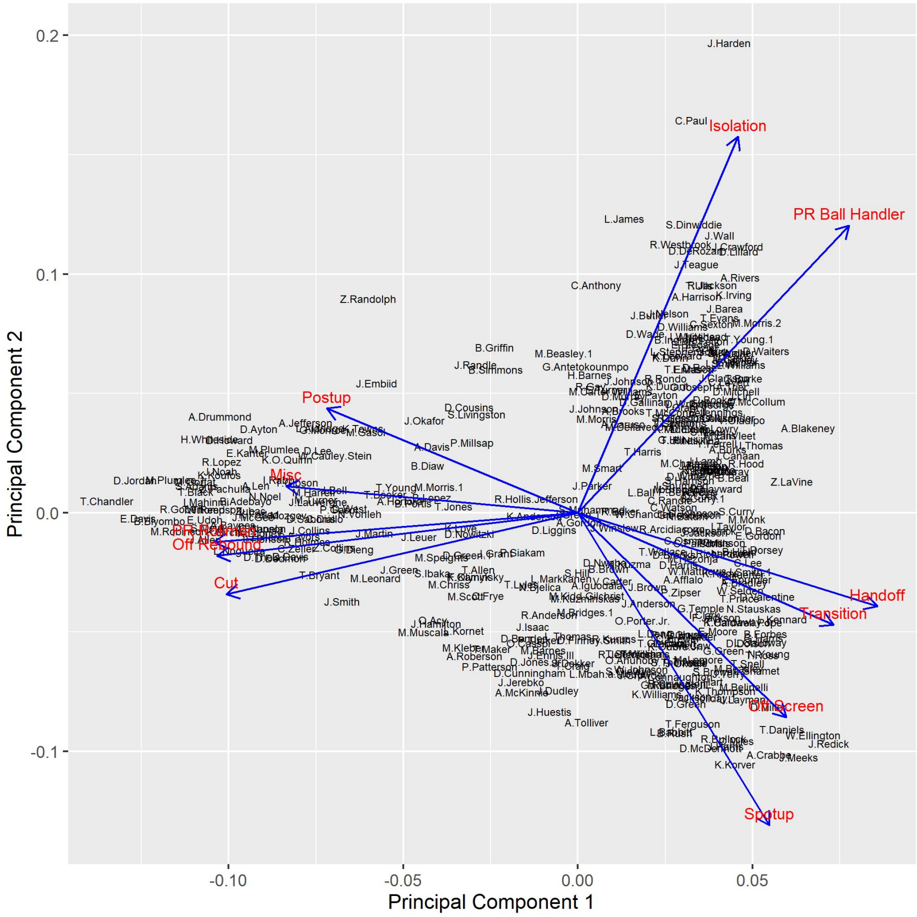
The Data

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:

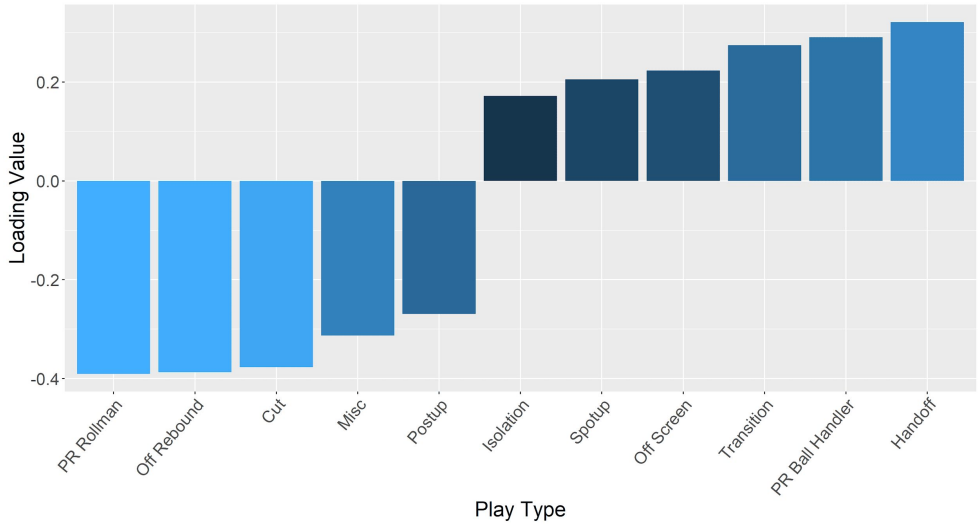
- Rollman in the pick and roll
- Off rebound
- Cut
- Postup
- Isolation
- Spotup
- Off screen
- Transition
- Ball handler in the pick and roll
- Handoff
- Misc

Biplot



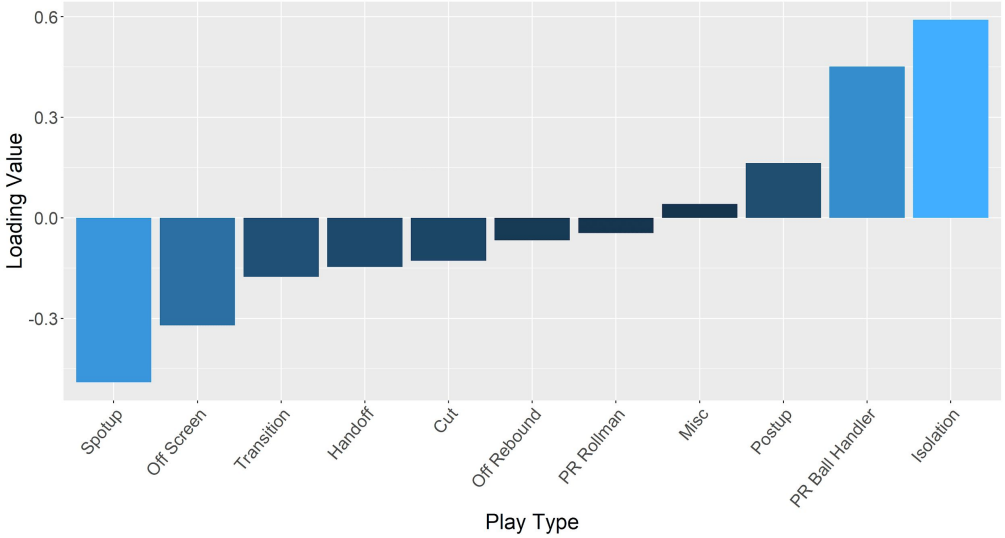
Principal Component 1
Principal Component 2
0.2
0.1
0.0
-0.1
-0.10
-0.05
0.00
0.05
Isolation
PR Ball Handler
Postup
Misc
Cut
Off Rebound
Handoff
Transition
Off Screen
Spotup
J.Harden
C.Paul
L.James
S.Dinwiddie
J.Wall
J.Teague
C.Anthony
A.Rivers
Z.Randolph
B.Griffin
M.Beasley.1
G.Antetokounmpo
J.Embiid
A.Drummond
A.Blakeney
D.Cousins
S.Livingston
J.Okafor
P.Millsap
B.Diaw
T.Chandler
Z.LaVine
R.Hollis.Jefferson
D.Liggins
J.Smith
T.Bryant
M.Leonard
J.Huestis
A.Tolliver
K.Korver
J.Redick
J.Meeks
T.Daniels
W.Ellington
A.Crabbe

PC1 Loadings



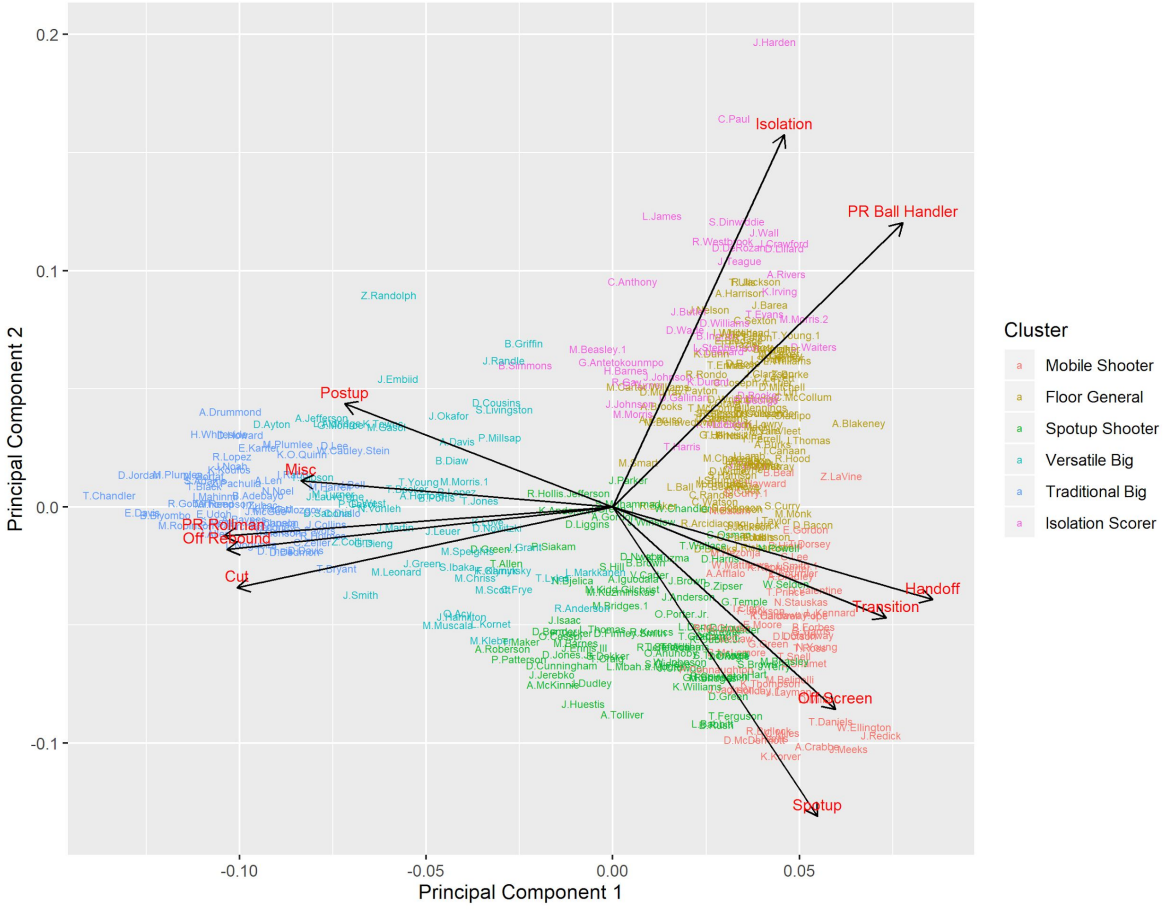
- High Handoff, PB Ball Handler, and Transition
- Low PR Rollman, Off Rebound, and Cut
- Separates big men from guards and forwards

PC2 Loadings

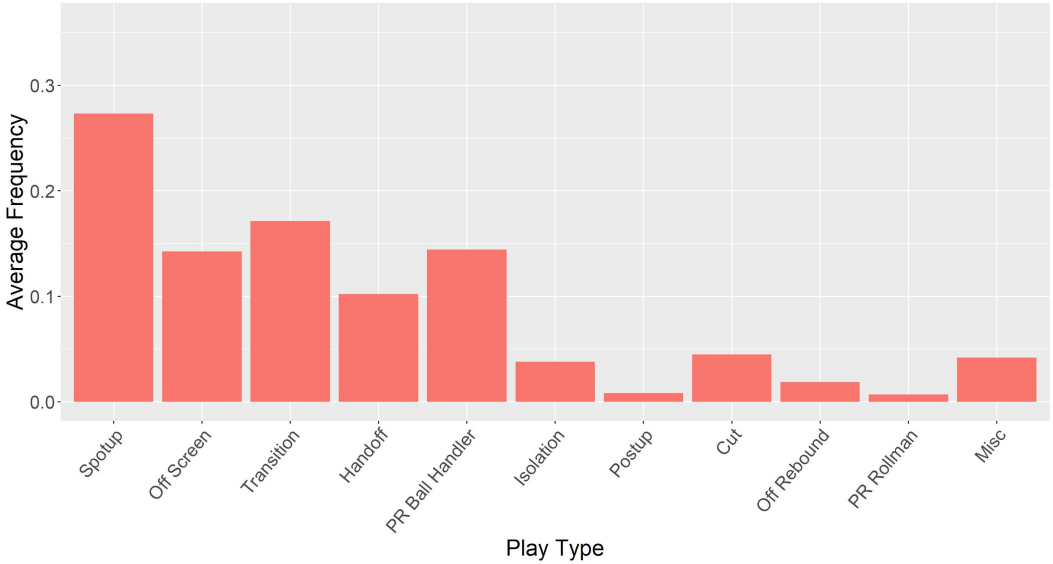


- High Isolation, and PR Ball Handler
- Low Spotup and Off Screen
- Separates ball handlers from shooters

Clusters

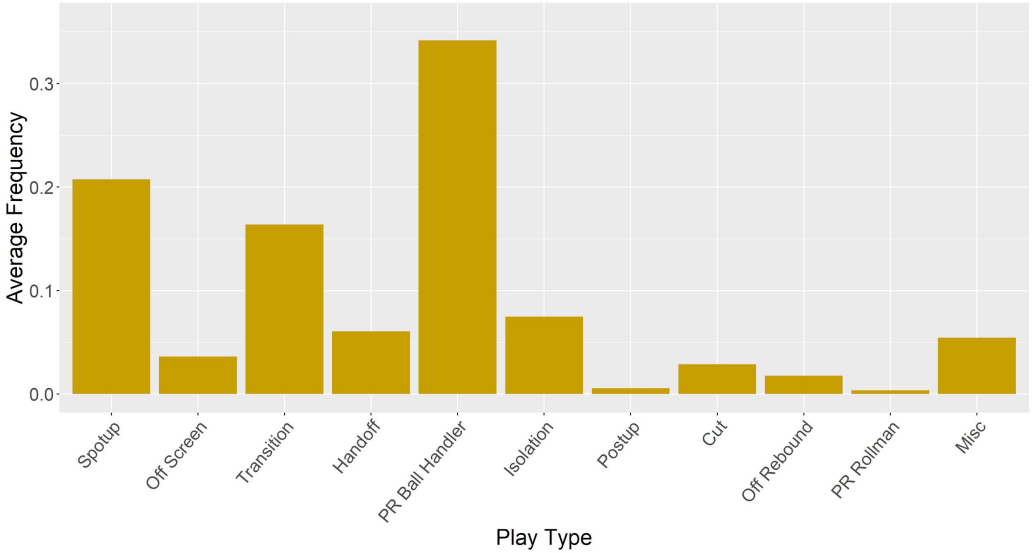

Principal Component 2
Principal Component 1
0.2
0.1
0.0
-0.1
-0.10
-0.05
0.00
0.05
Isolation
PR Ball Handler
Postup
Misc
PR Rollman
Off Rebound
Cut
Handoff
Transition
Off Screen
Spotup
Cluster
Mobile Shooter
Floor General
Spotup Shooter
Versatile Big
Traditional Big
Isolation Scorer

Mobile Shooters



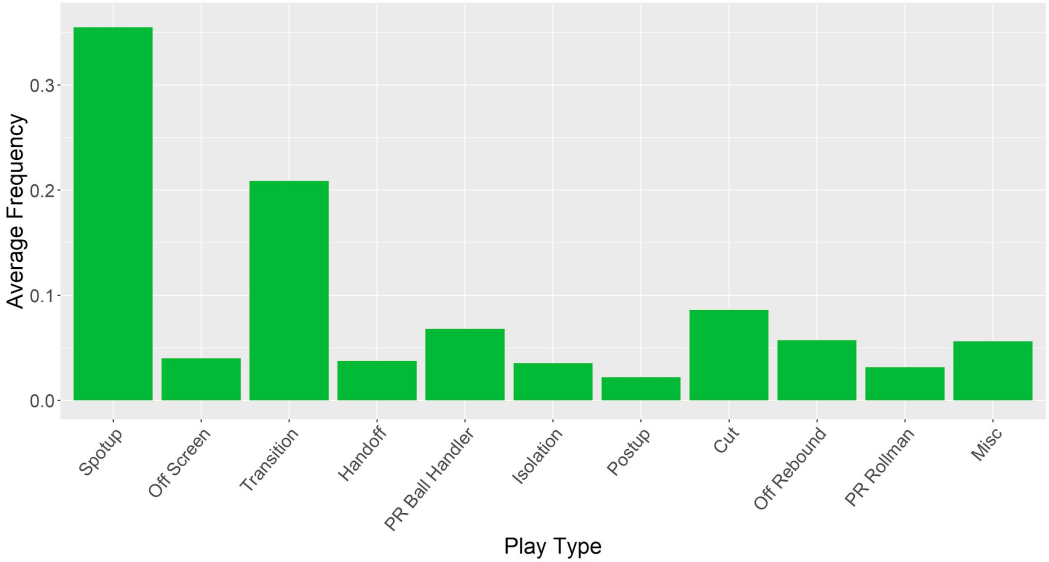
- High Spotup, Off screen, Transition
- Low PR Rollman, Postup, Off Rebound
- Key members: Buddy Hield, JJ Redick, Klay Thompson

Floor Generals



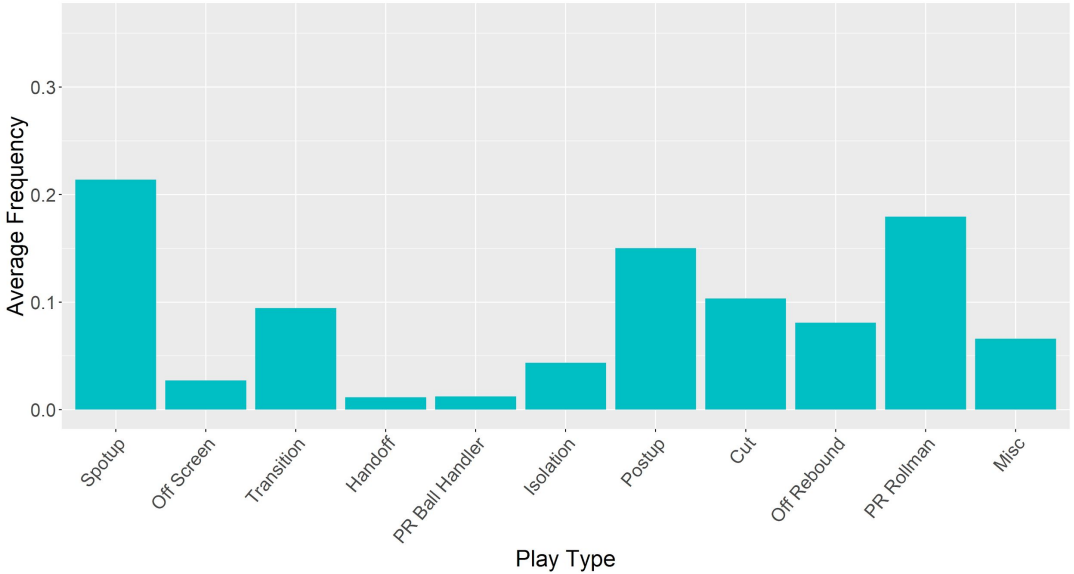
- High PR Ball Handler, Spotup, Transition
- Low PR Rollman, Postup, Off Rebound
- Key Members: Steph Curry, Lonzo Ball, Rajon Rondo

Spotup Shooters



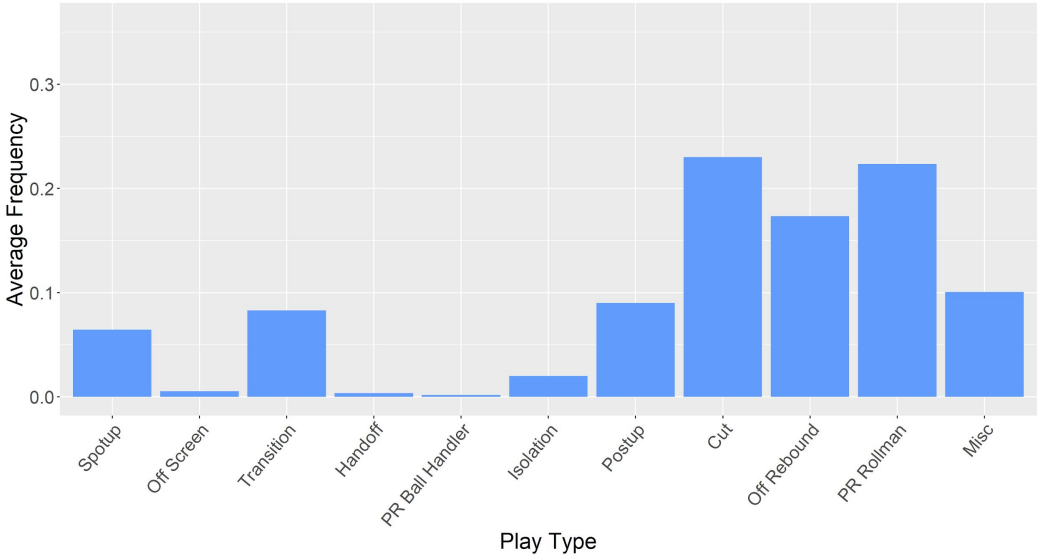
- High Spotup and Transition
- Low Off Screen and Postup
- Key members: Otto Porter Jr., Anthony Tolliver, Nemanja Bjelica

Versatile Bigs



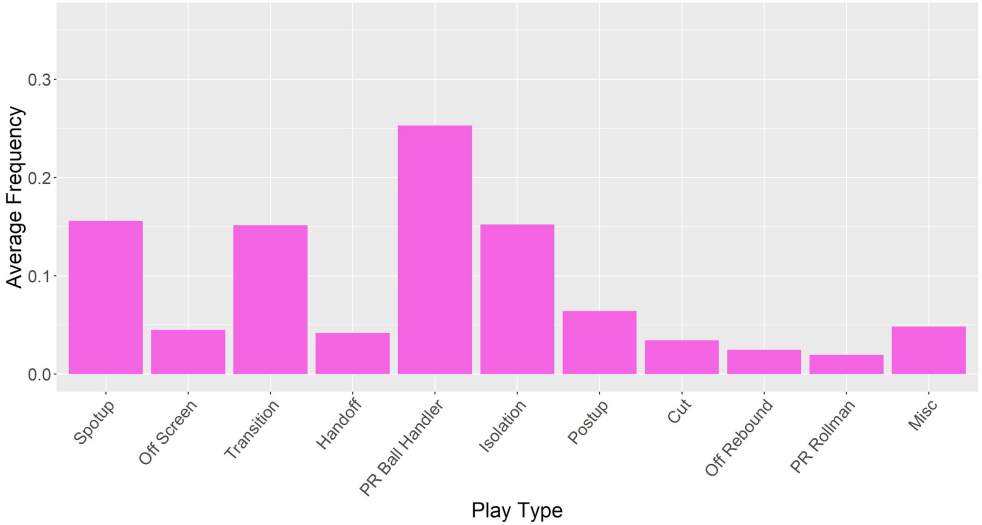
- High Spotup, PR Rollman, Postup
- Low Handoff, PR Ball Handler, Off Screen
- Key members: Demarcus Cousins, Anthony Davis, Blake Griffin

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		

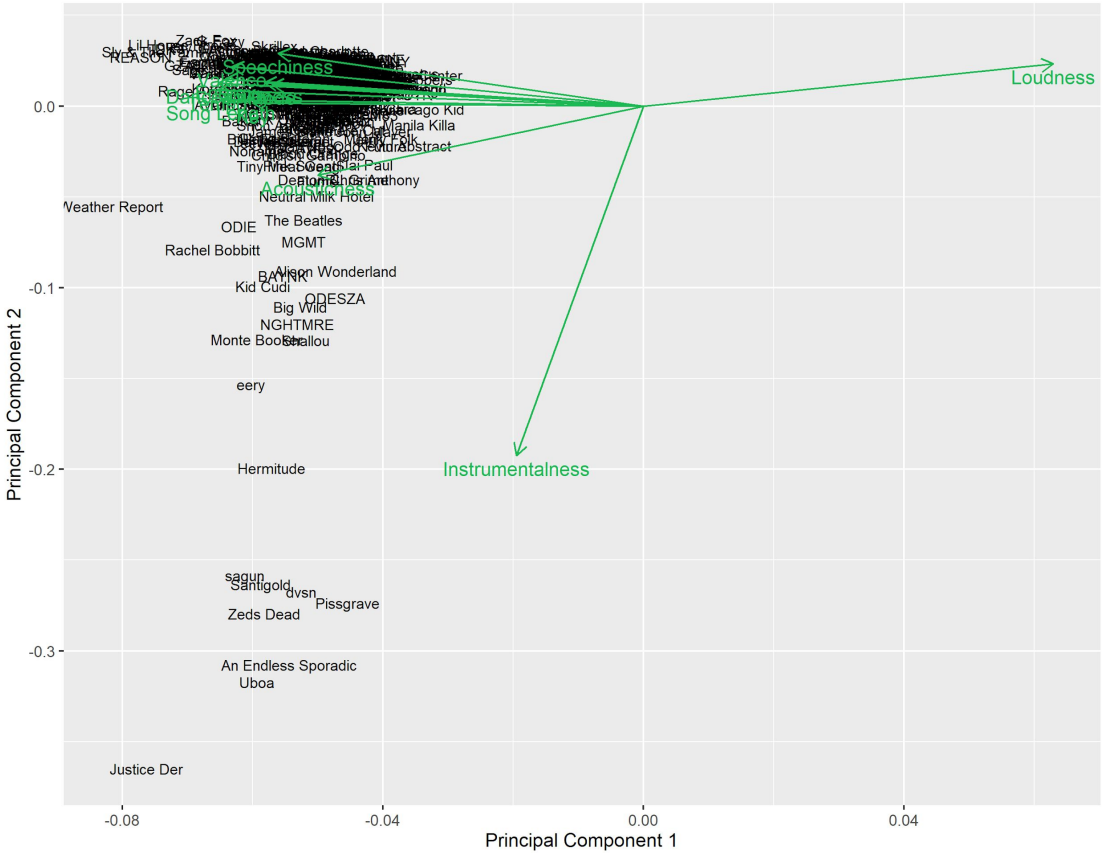
The Data

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

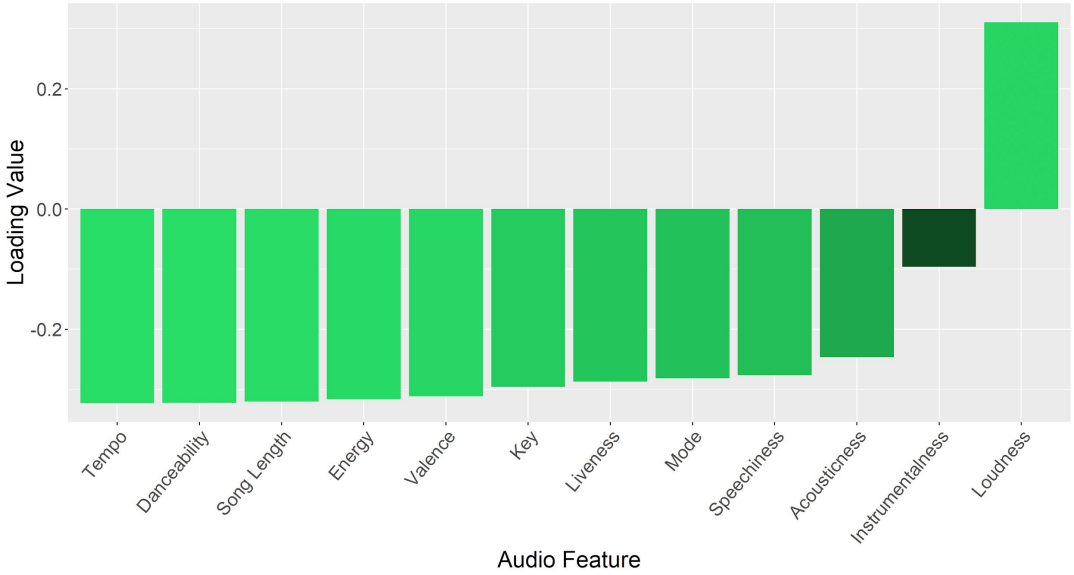
Variables - the average of these attributes from the songs I listened to from each artist:

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

Biplot

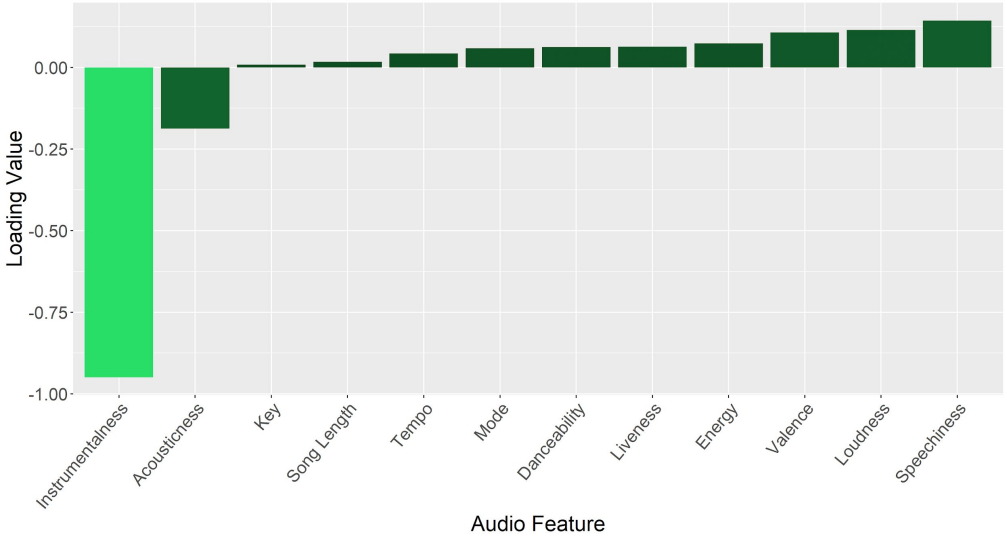

Speechiness
Loudness
Instrumentalness
Acousticness
Weather Report
ODIE
The Beatles
Rachel Bobbitt
MGMT
Kid Cudi
ODESZA
Big Wild
NGHTMRE
eery
Hermitude
Zeds Dead
Pissgrave
An Endless Sporadic
Uboa
Justice Der
Principal Component 2
Principal Component 1
0.0
-0.1
-0.2
-0.3
-0.08
-0.04
0.00
0.04

PC1 Loadings



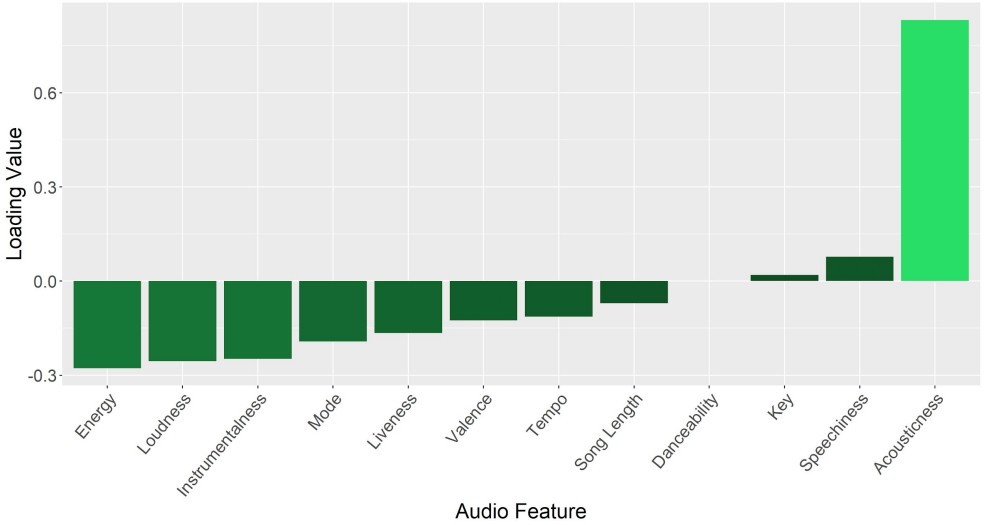
- High Loudness
- Low everything else
- Separates loud artists from not loud artists

PC2 Loadings



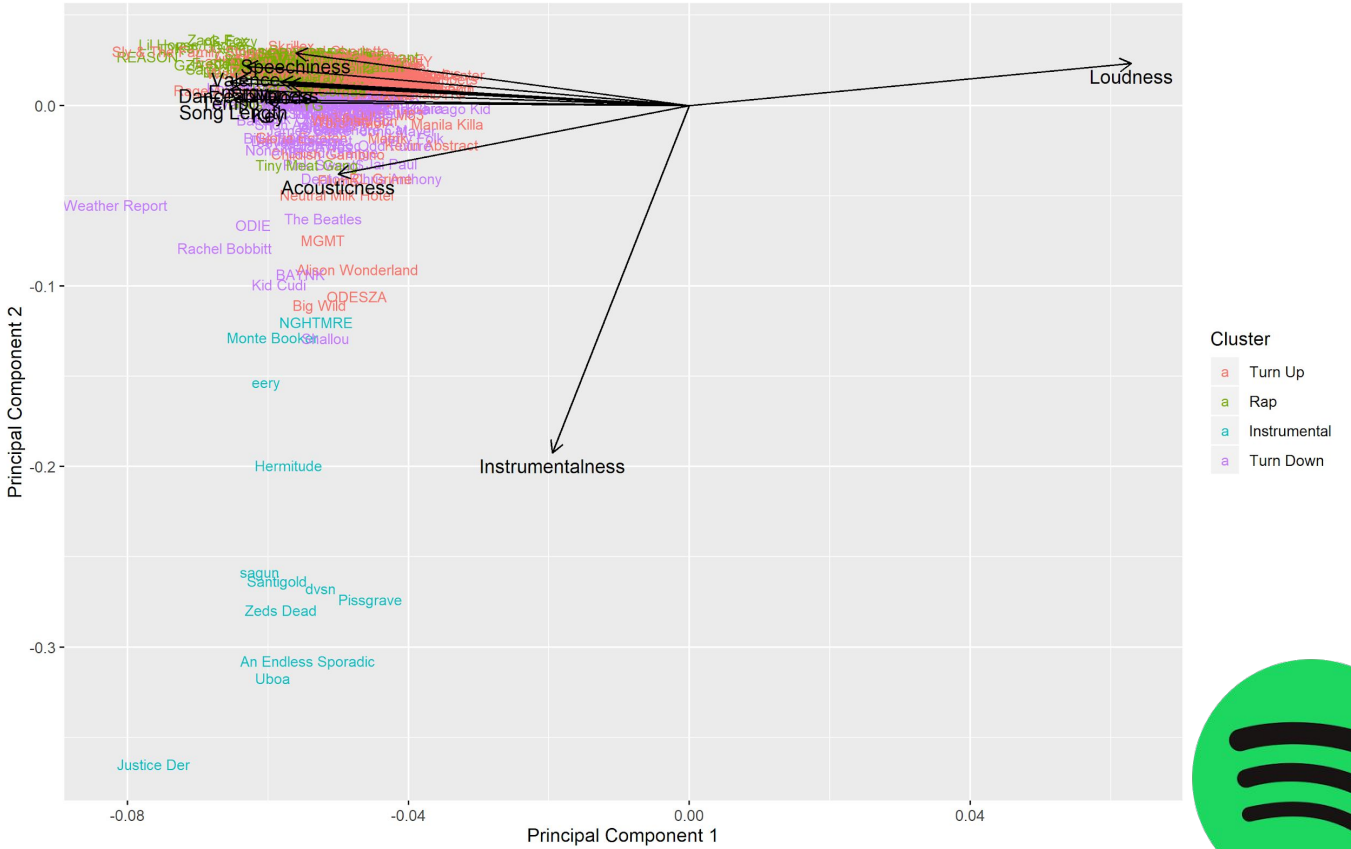
- High most variables
- Very low Instrumentalness
- Separates artists who play instrumentals and artists that use vocals

PC3 Loadings

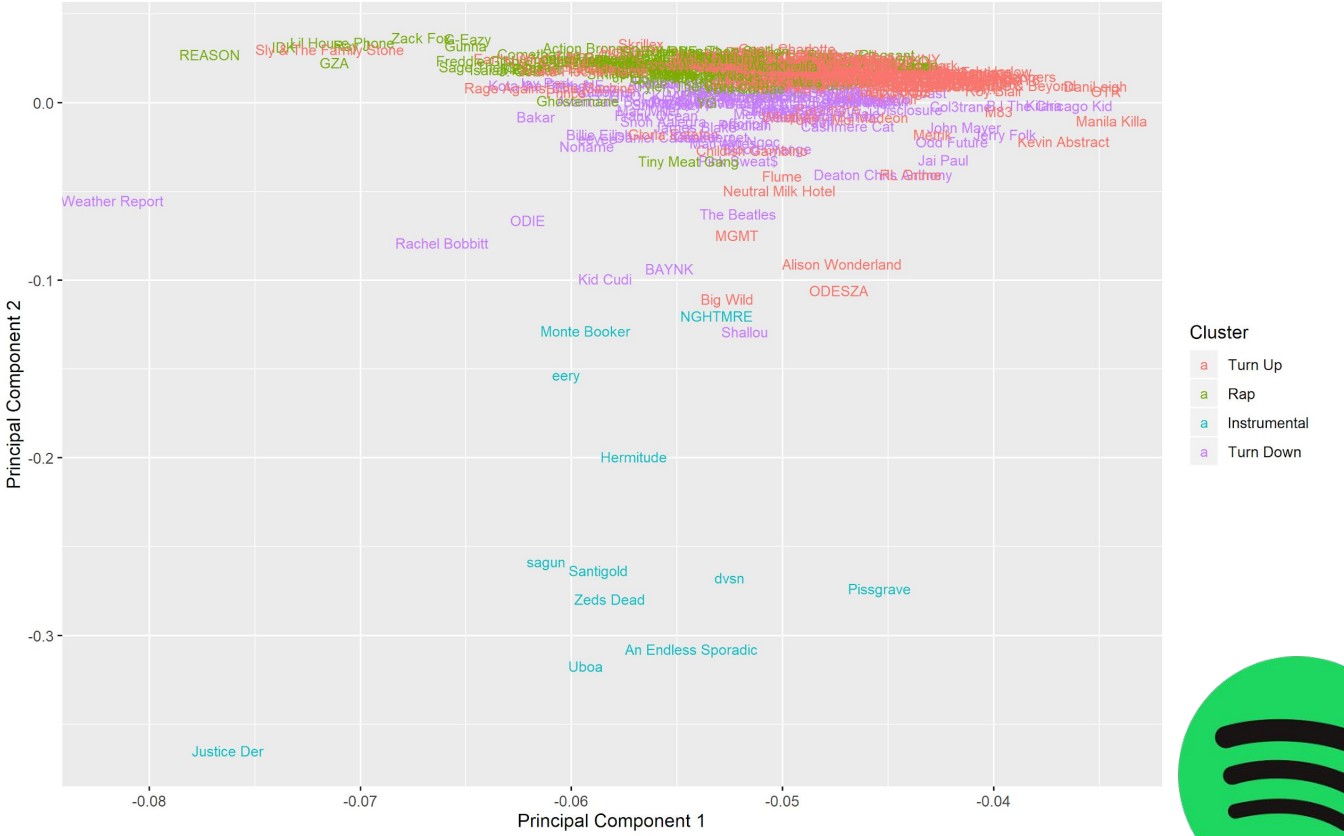


- High Acousticness
- Low everything else
- Separates acoustic instrumentation artists from artists that use other instruments

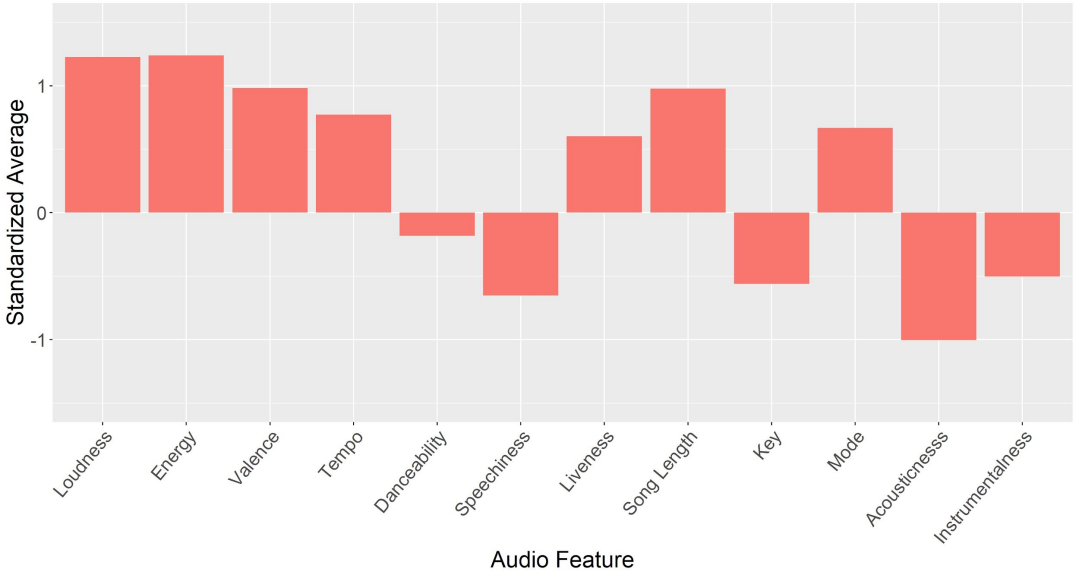
Clusters


Speechiness
Valence
Song Length
Acousticness
Loudness
Instrumentalness
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
Principal Component 2
Principal Component 1
0.0
-0.1
-0.2
-0.3
-0.08
-0.04
0.00
0.04
Cluster
Turn Up
Rap
Instrumental
Turn Down

Clusters

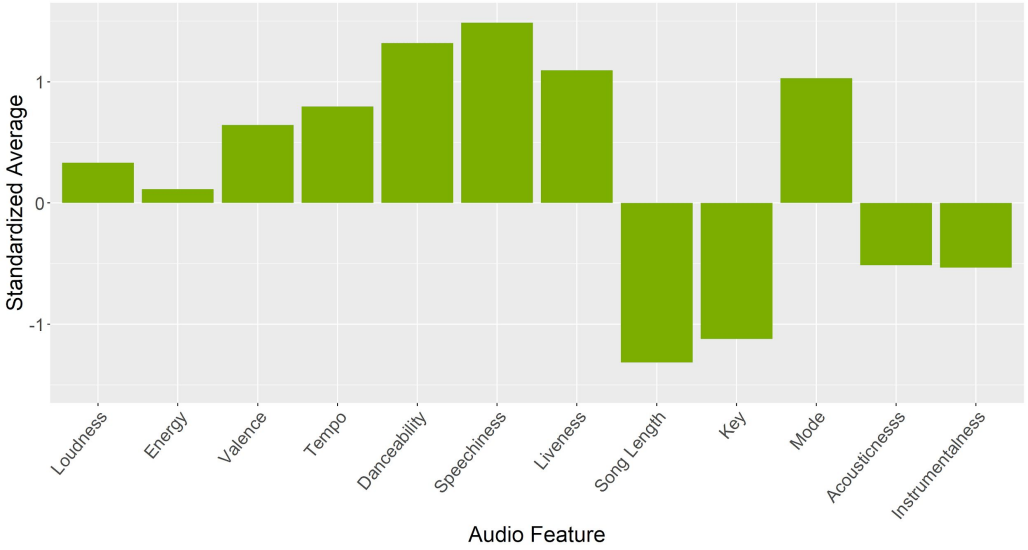

REASON
GZA
Zack Fox
Fazy
Gunna
Skrillex
Bakar
Ghostemane
Noname
Tiny Meat Gang
Flume
Neutral Milk Hotel
Weather Report
ODIE
Rachel Bobbitt
The Beatles
MGMT
BAYNK
Kid Cudi
Alison Wonderland
ODESZA
Big Wild
NGHTMRE
Shallou
Monte Booker
eery
Hermitude
sagun
Santigold
Zeds Dead
dvsn
Pissgrave
An Endless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

Turn Up



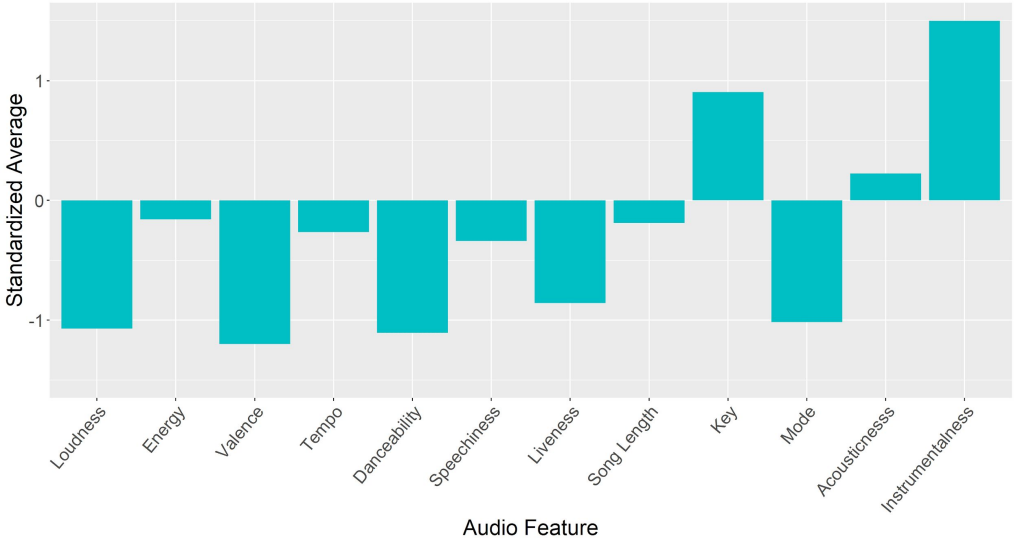
- High Loudness, Energy, Valence
- Low Acousticness, Speechiness, Instrumentalness
- Key members: Taylor Swift, Rage Against the Machine, Flume

Rap



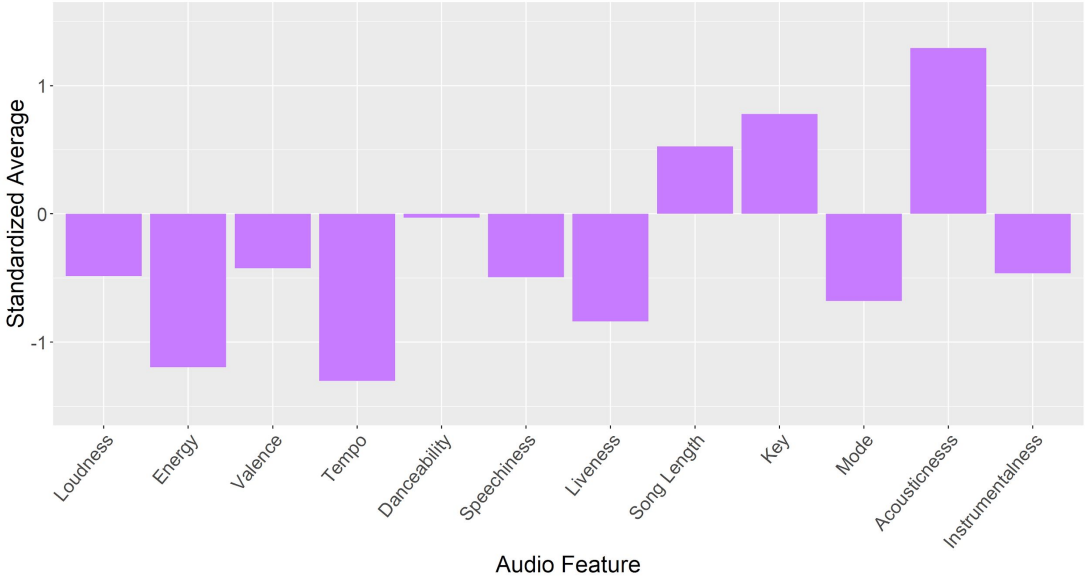
- High Speechiness, Danceability, Liveness
- Low Song Length and Key
- Key members: Kendrick Lamar, J. Cole, Wiz Khalifa

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

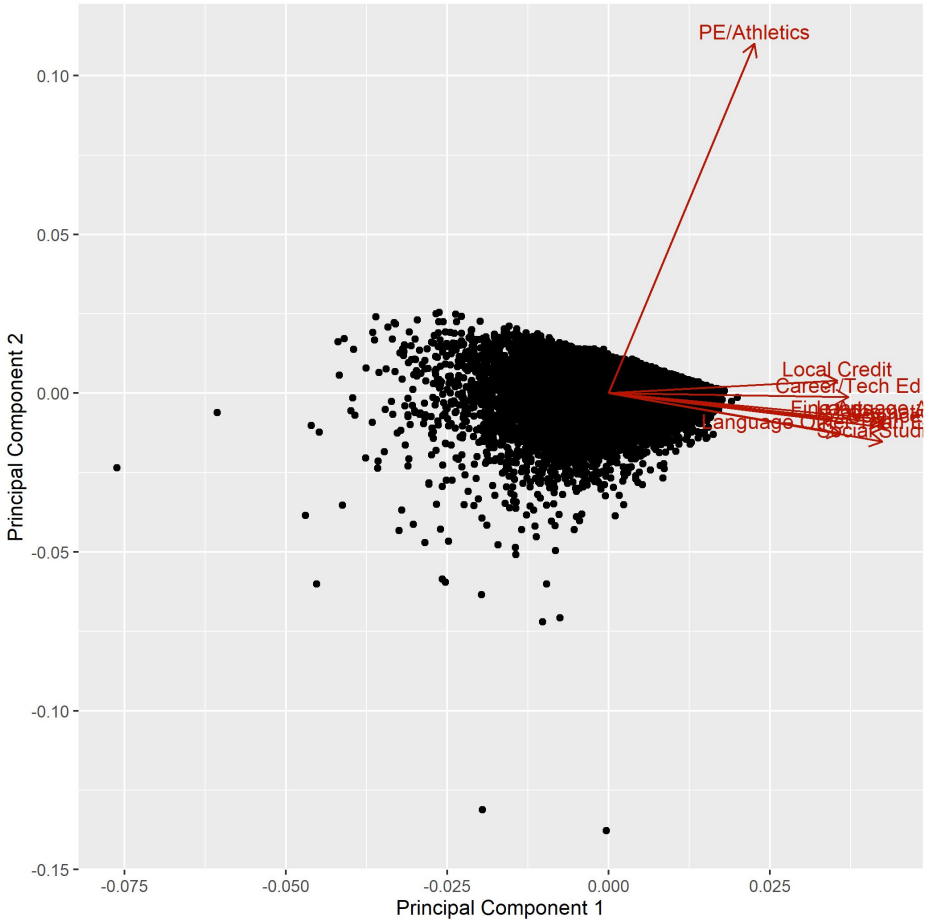
The Data

One data point - One student from the Aldine School District

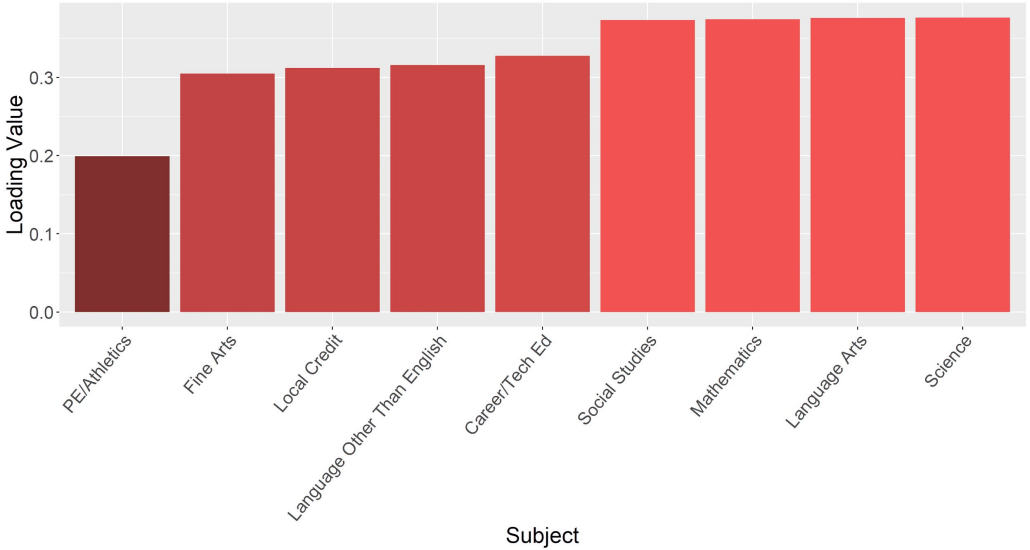
Variables - A student's average grade in these subjects:

- Science
- Language Arts
- Mathematics
- Social Studies
- Career/Tech Ed
- Language Other Than English
- Local Credit
- Fine Arts
- PE/Athletics

Biplot

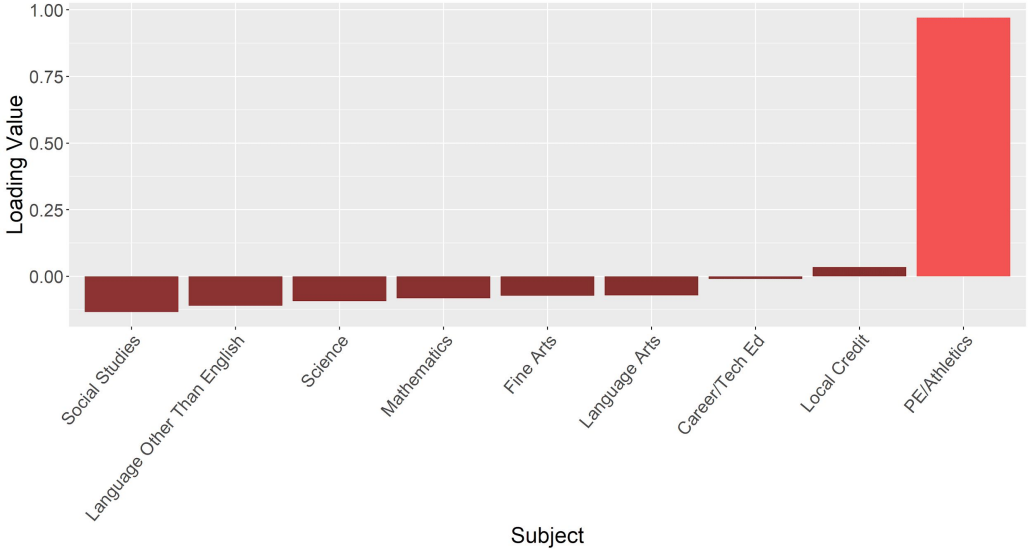

PE/Athletics
Local Credit
Career/Tech Ed
0.10
0.05
0.00
-0.05
-0.10
-0.15
Principal Component 2
-0.075
-0.050
-0.025
0.000
0.025
Principal Component 1

PC1 Loadings



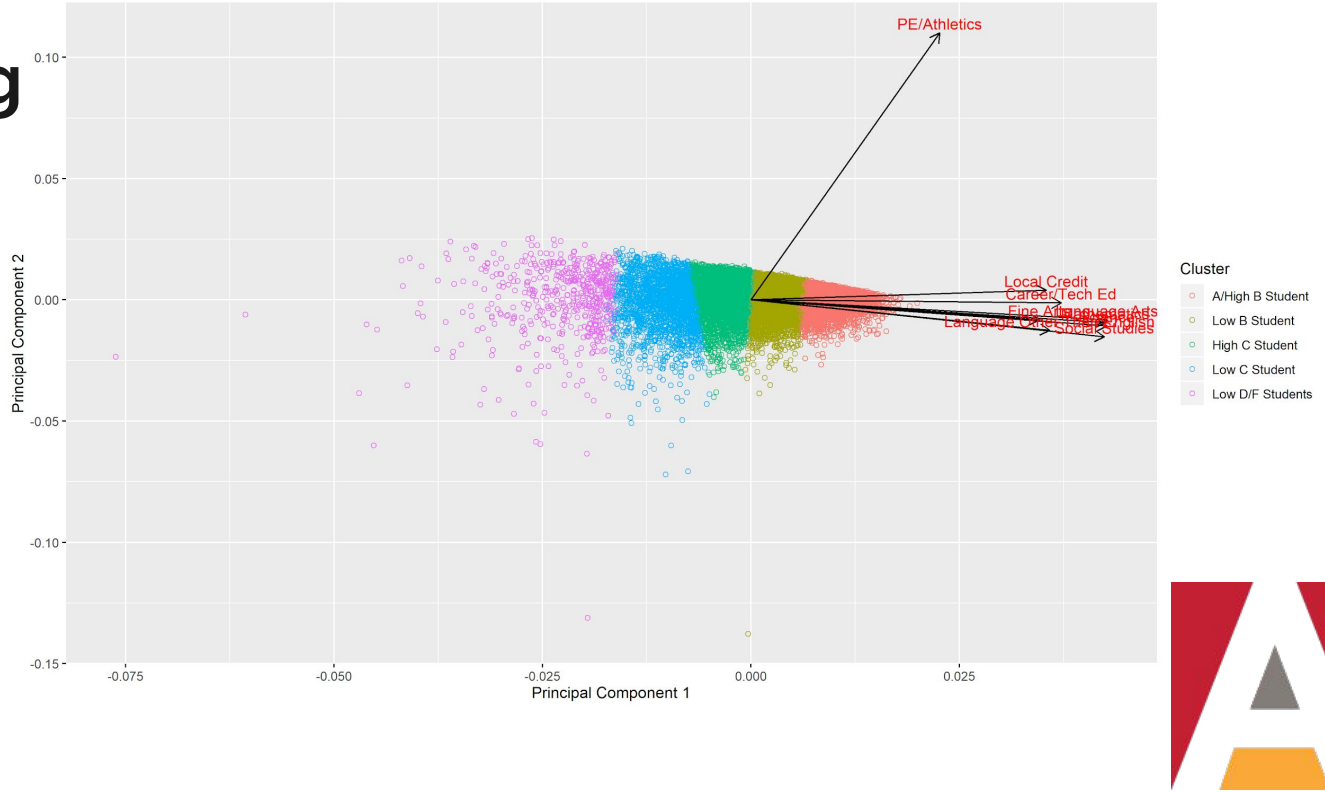
- High everything
- Separates students who get good grades from students who get bad grades

PC2 Loadings

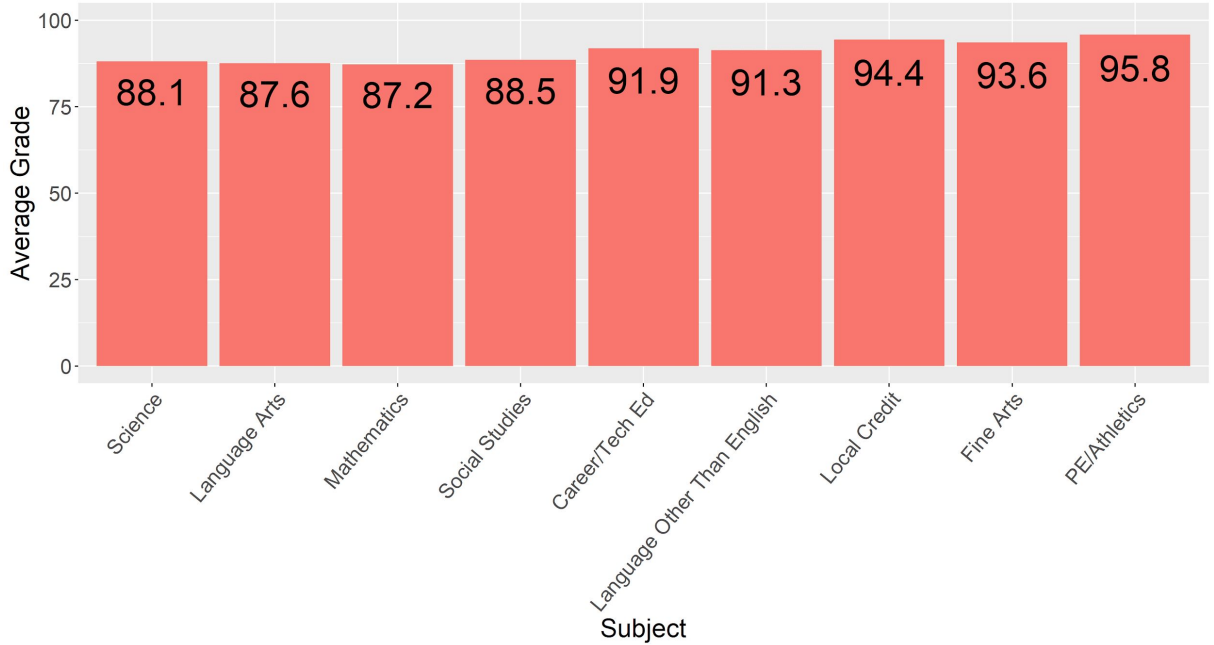


- High PE/Athletics
- Low everything else
- Separates students that do well in PE/Athletics from those that do not do well in PE/Athletics

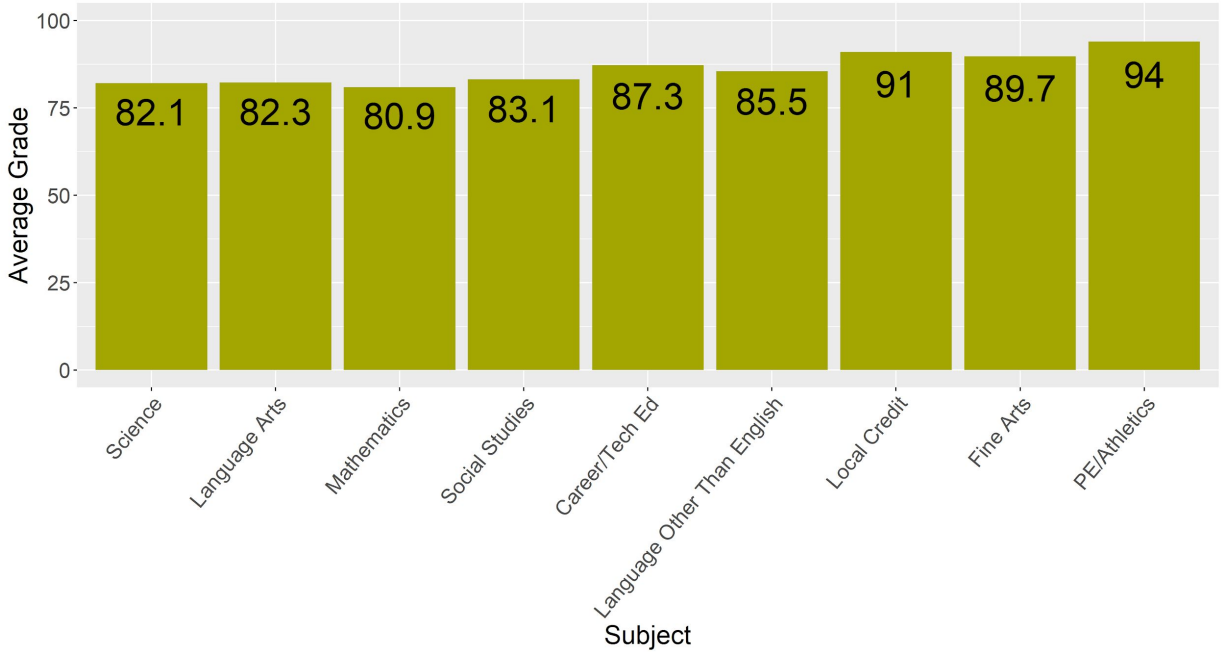
Clustering


PE/Athletics
Local Credit
Career/Tech Ed
Fine Arts
Language Arts
Social Studies
Principal Component 2
Principal Component 1
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.075
-0.050
-0.025
0.000
0.025
Cluster
A/High B Student
Low B Student
High C Student
Low C Student
Low D/F Students

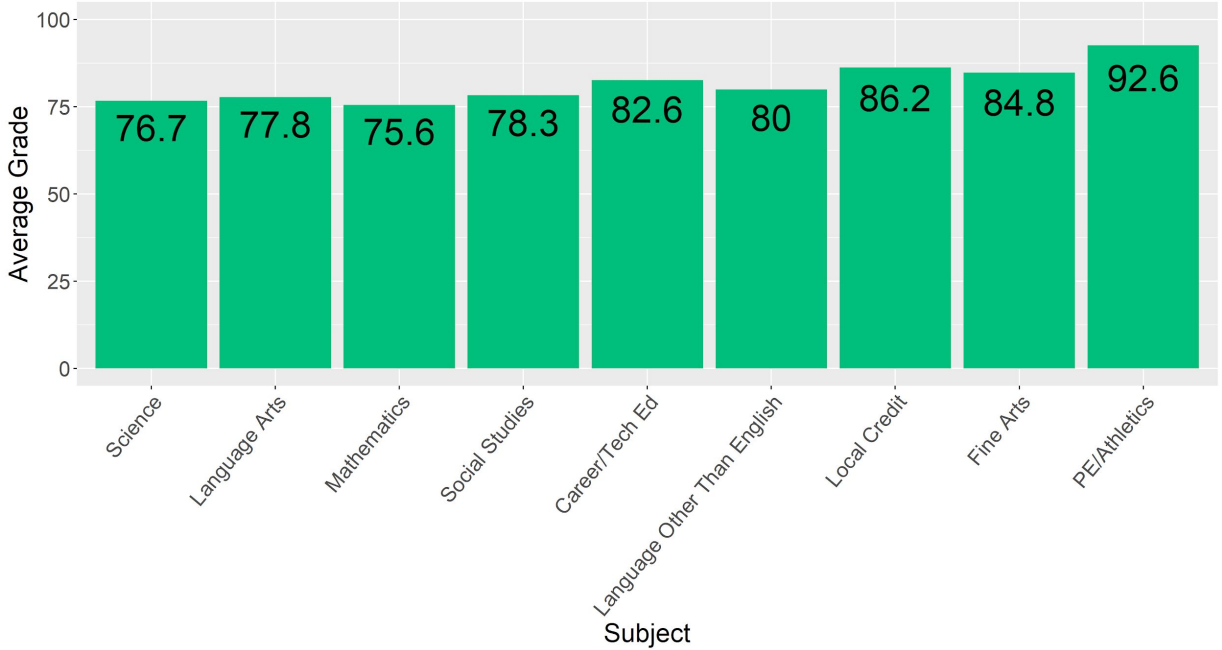
A/High B Students



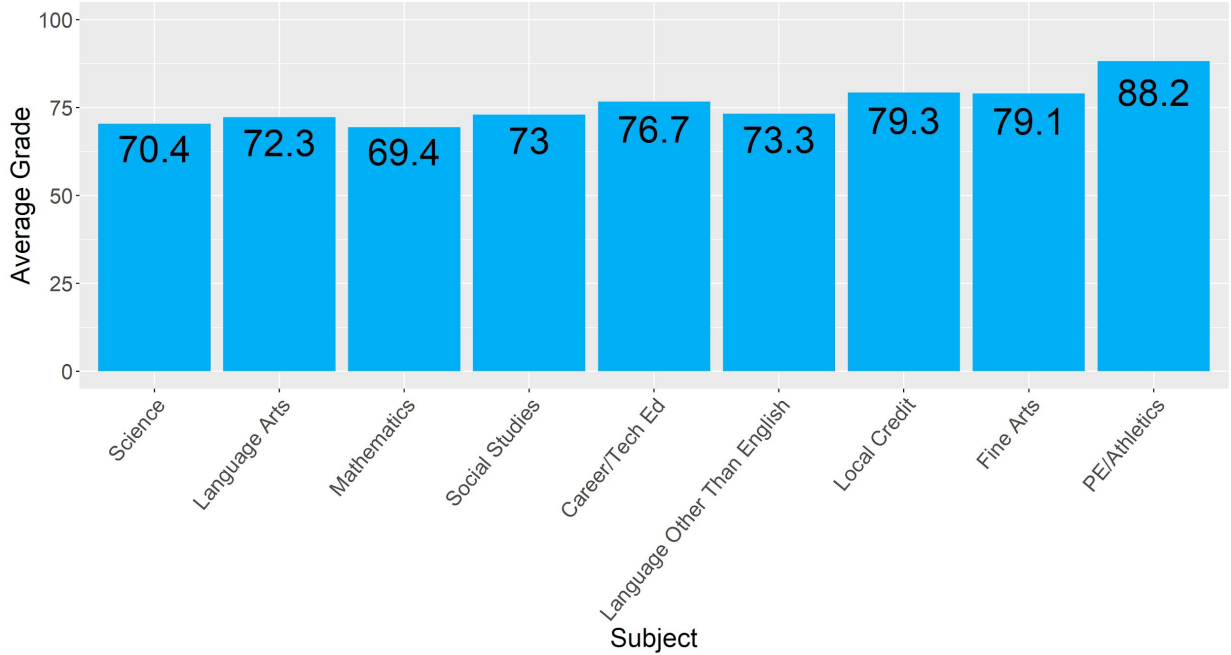
Low B Students



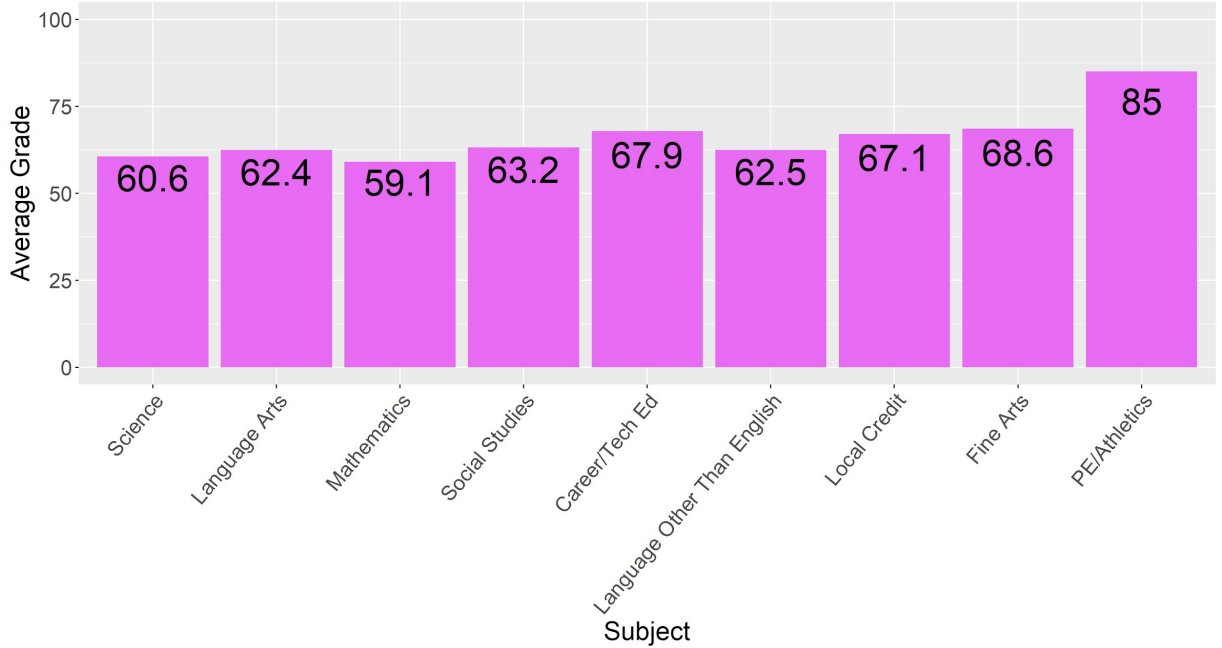
High C Students



Low C Students



Low D/F Students



Main Takeaways

- The best way to differentiate Aldine students is by separating students that get higher grades from students that get lower grades
- The second-best way to differentiate Aldine students is by separating students that do well in PE/Athletics and students that do not
- Student grades form a spectrum, rather than groups

Low Grades

High Grades

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