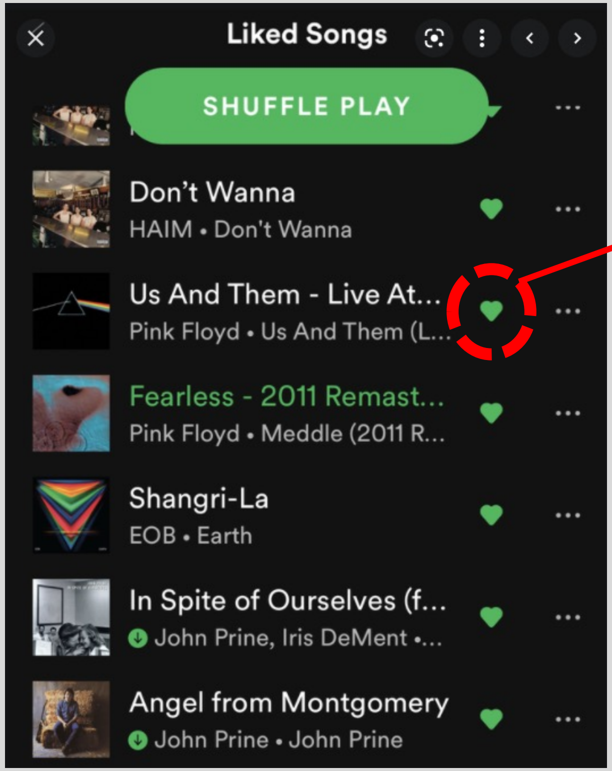
Spotify: A Deeper Dive

Analyzing my listening habits on Spotify through EDA & PCA.

Matt Reilly

The Objective?



Using my "Liked" songs on Spotify, quantify and visualize my listening habits.

Identify trends, correlations, and insights that can be made from my "liked" music history.

Data Extraction

- Use *SpotifyR* to query my profile's playlist
- Iterate through each track's ID to extract features
- Combine identification & features into single data frame

SpotifyR package

get_playlist( )

get_track_audio_features( )

Features

Identifier

Danceability	Energy	key	Loudness	mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	tempo	track_id	duration_ms	time_signature	track_name	artist_name	artist_id
0.806	0.913	6	-4.825	0	0.0467	0.00266	0.0932	0.0855	0.627	125.012	5jV0kgwhWafAdeJPVFcZpQ	263160	4	Want It	Black V Neck	2l0xOjnrmYsxNoQ0QI3G5a
0.806	0.638	8	-5.698	1	0.0424	0.149	0.0000137	0.119	0.858	104.976	2E7R8kXD7qZpvfW25F7gUW	195413	4	Prosecco	Patrik Jean	5QCf1Qb08Q4E3EPnyo8mw1
0.684	0.913	11	-4.203	0	0.0673	0.0395	0.39	0.0911	0.249	127.992	683i8O9hKQK1j9ai3mCo5M	355093	4	Funhouse - Digital Dog Remix	P!nk	1KCSPY1glIKqW2TotWuXOR
0.866	0.954	10	-7.38	1	0.0636	0.0953	0.429	0.123	0.9	124.965	7uIsUgH28J66pUa6nL2uSU	212000	4	Break Ya Neck	Sloth	5iD9mn2inFzm1u3jvr9egi
0.747	0.749	0	-6.337	1	0.0558	0.0798	0.00000179	0.318	0.914	121.964	2LAtELE0xGyMKcvNlxiyF3	138200	4	1 Day 2 Nights	HRVY	28y6CyJNkGNjJQKrlx4AmN
0.626	0.826	4	-4.474	0	0.0574	0.000482	0	0.335	0.461	122.976	1vZPWU4KC0Ao1XlIlJntL6	193840	4	Disco Love	The Saturdays	15qI5w4XJFLRMwOp2VrlD5
0.762	0.775	2	-4.348	1	0.034	0.0603	0	0.119	0.834	121.023	1HxcMzgxGXvteXSrTBbpu7	179507	4	Ego	The Saturdays	15qI5w4XJFLRMwOp2VrlD5

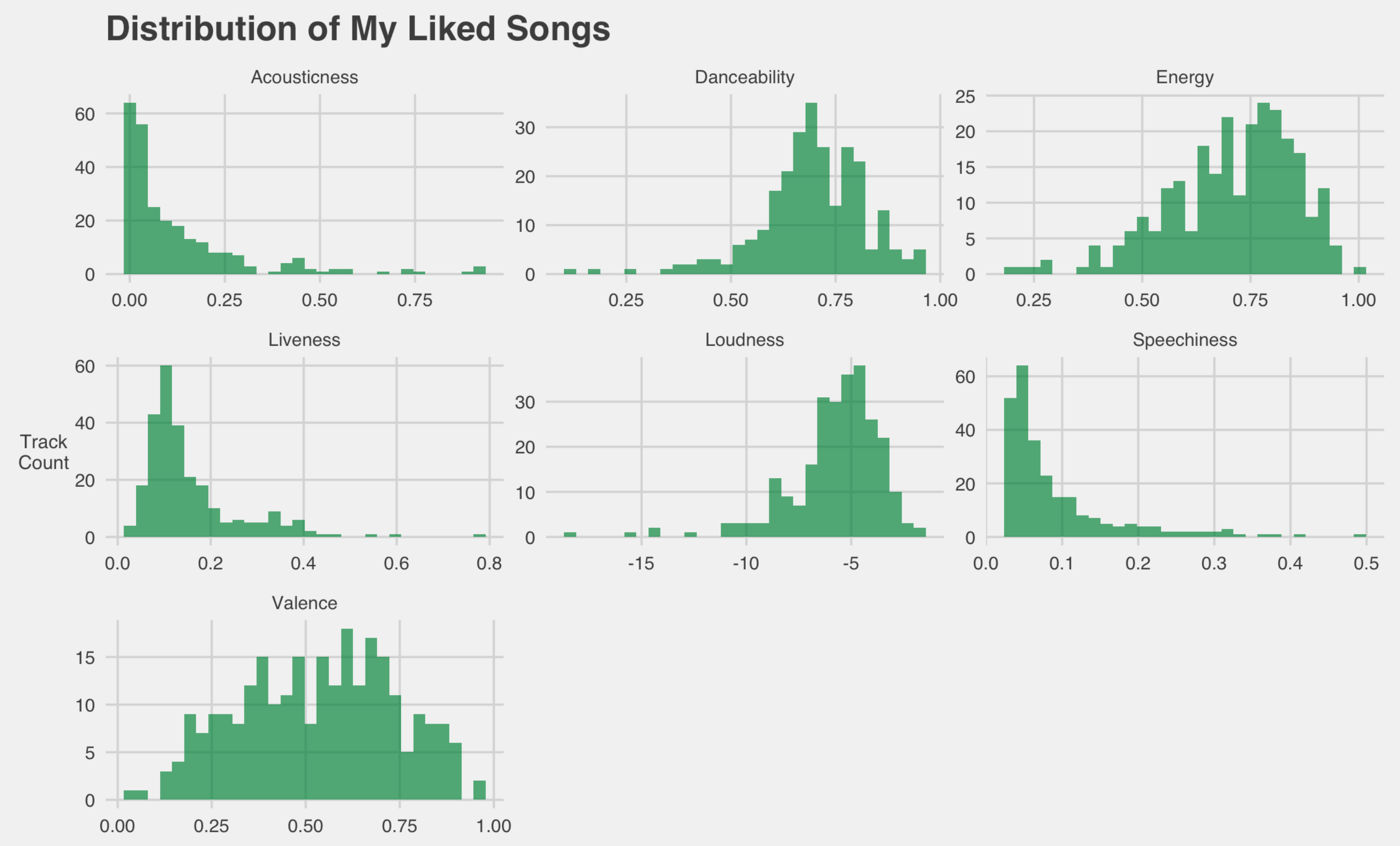
Exploratory Analysis

259 Tracks

221 Artists

Based on distributions, I tend to prefer more **danceable** and **energetic** tracks, rather than acoustic or “speechy” tracks.

I also like louder music.

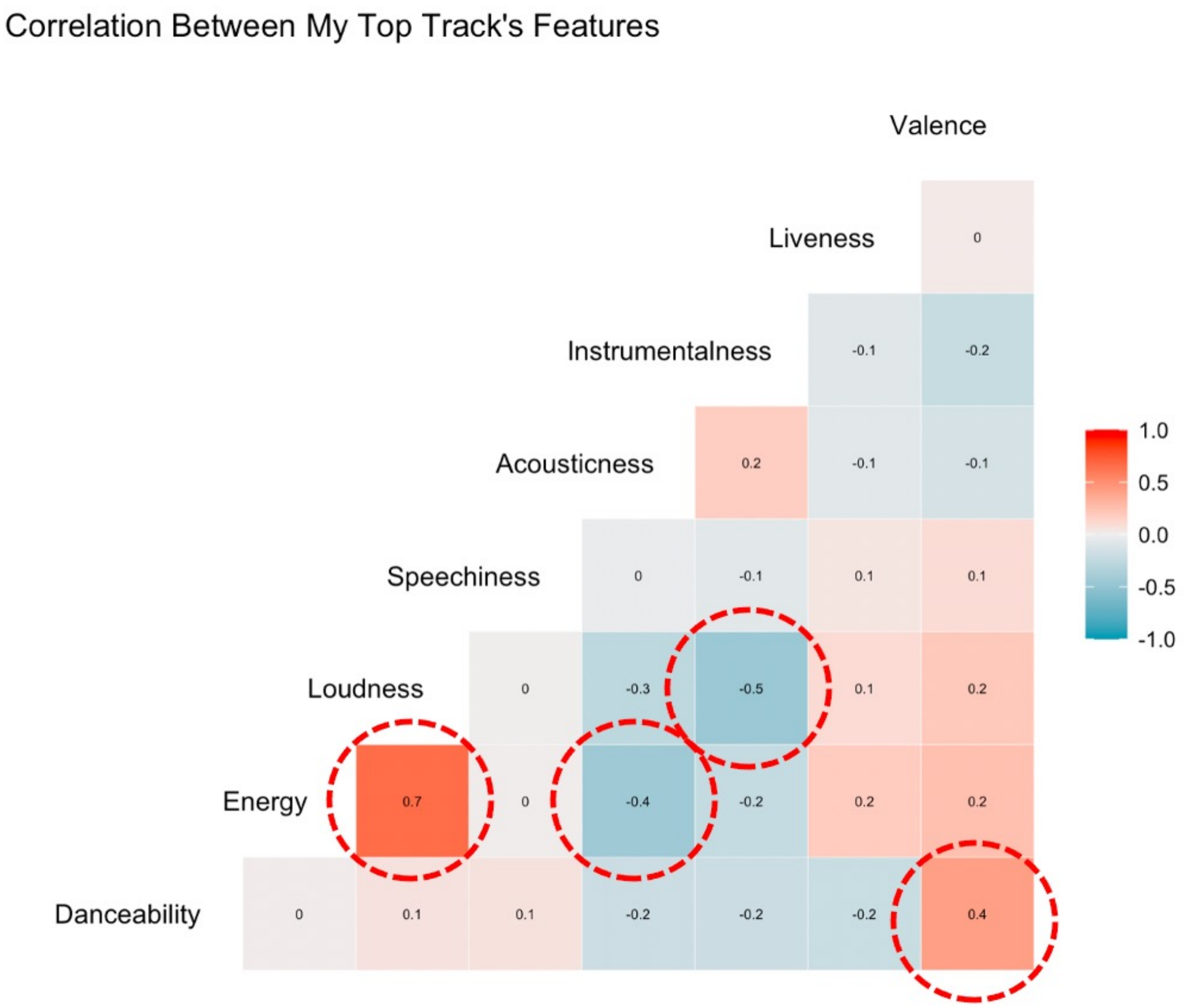


Feature Correlation

Correlation shows how close variables are to having a linear relationship with each other.

When my songs were **loud**, they tended to be more energetic and less instrumental.

The more danceable songs also tended to have a higher valence *(positivity conveyed by track).*



= "hotspot" relationships

PCA

INFO SLIDE:

Unsupervised method to reduce dimensions of data set

Collapse track features into **Principal Components** to more easily visualize

= PC1, PC2, PC3, PC4 …

PC1 Explains highest variation between features

Visual Example:

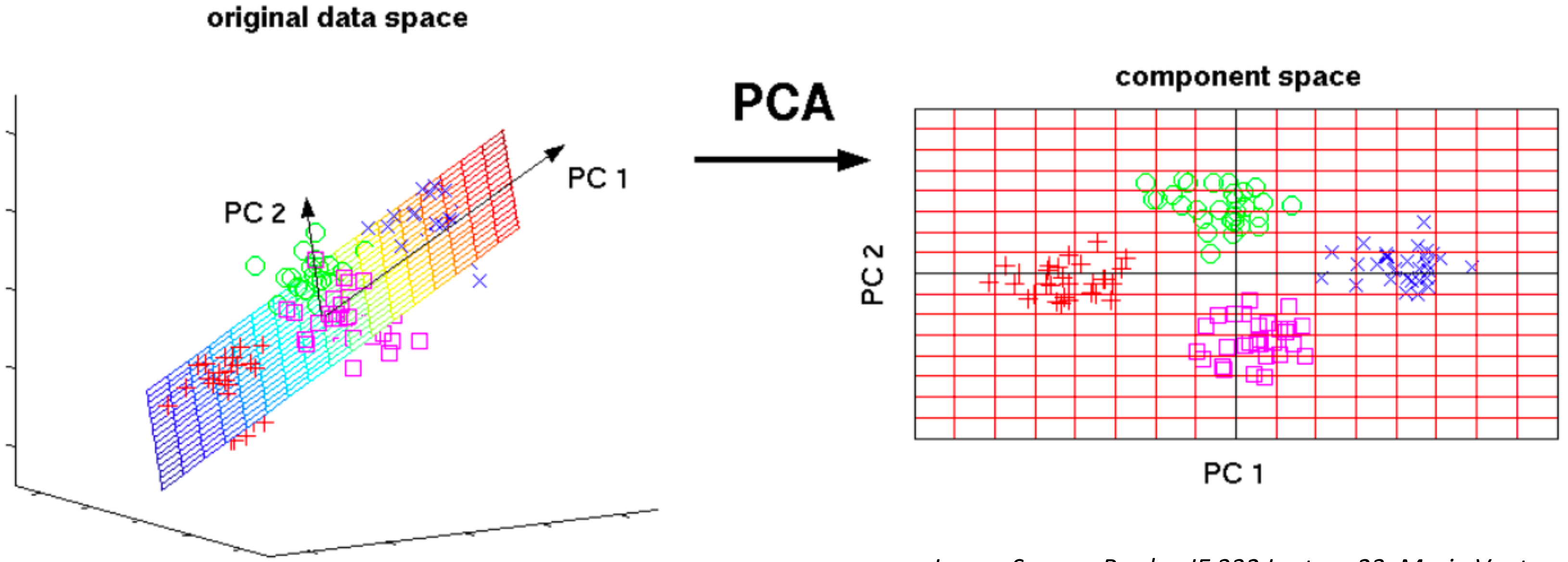
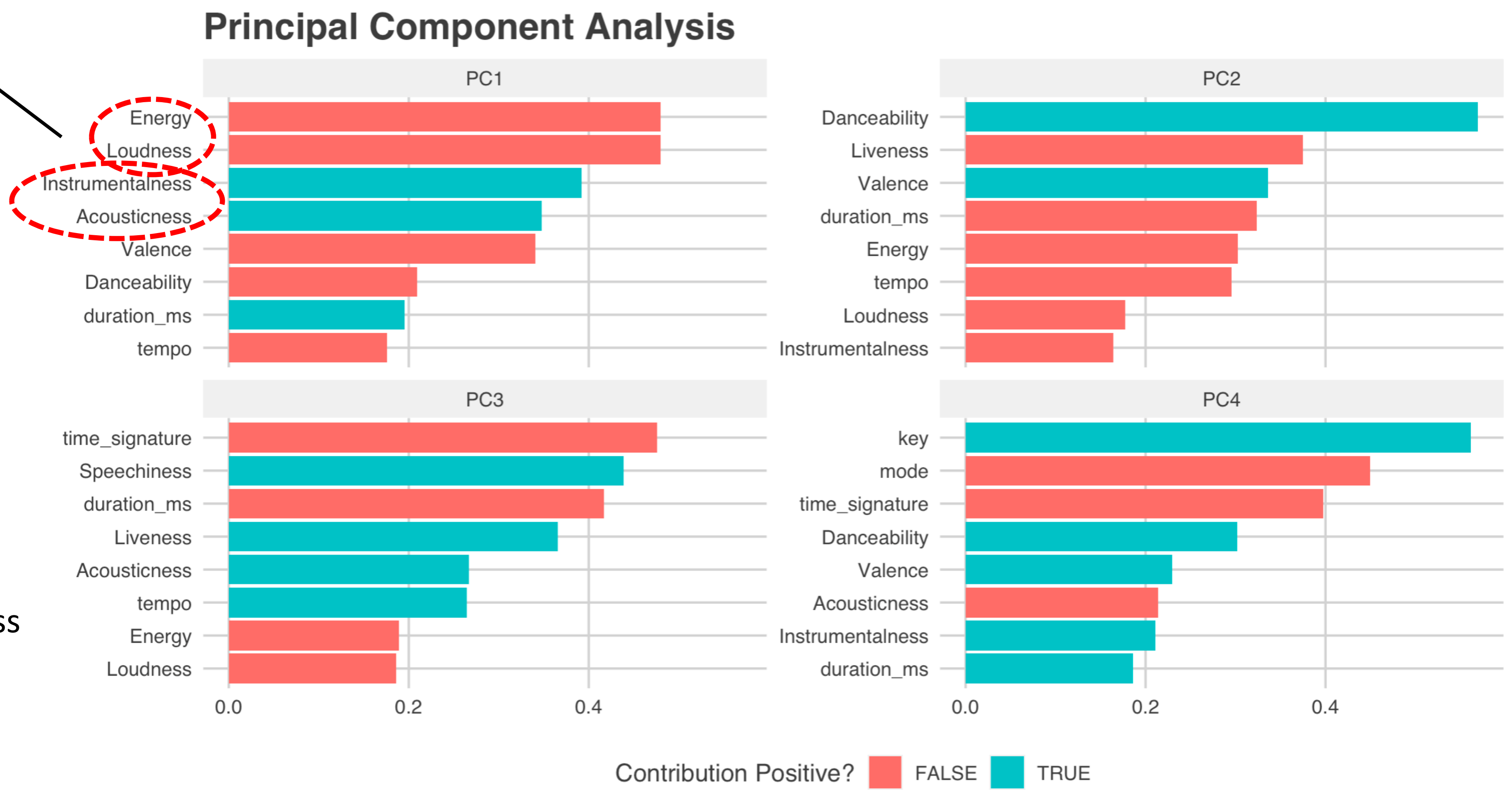


Image Source: Purdue IE 332 Lecture 23, Mario Ventresca

PCA

Correlated features explain variation in opposing directions



Features with high correlations tender to have a similar effect on variance explanation

Energy & Loudness

Instrumentalness & Acousticness

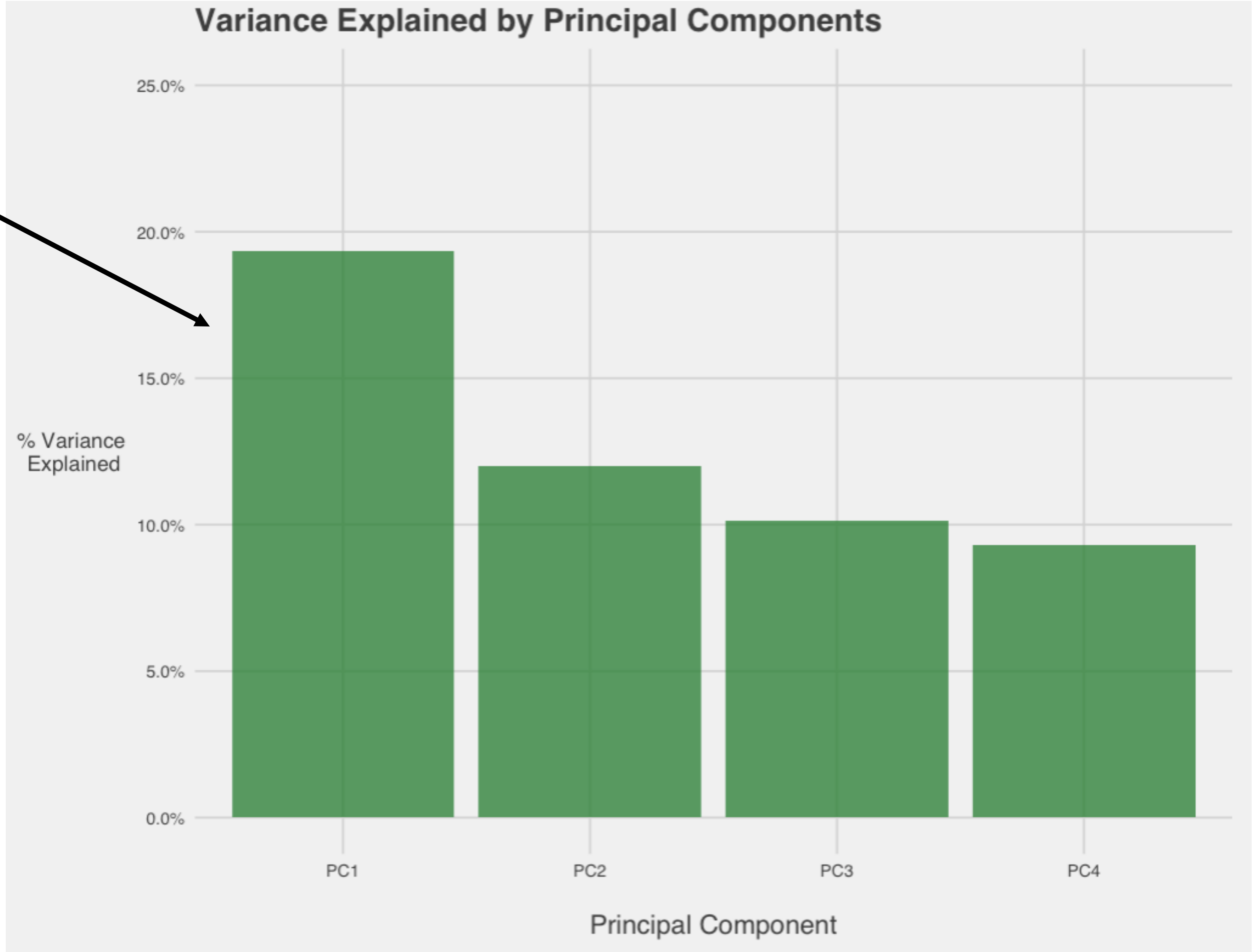
Valence & Danceability

PCA

PC1 only explains ~19% of variance, PC2 drops to ~12% *(31% total)*

Similarity of track features may cause minimal component explanation

My "Liked" songs are not that different?

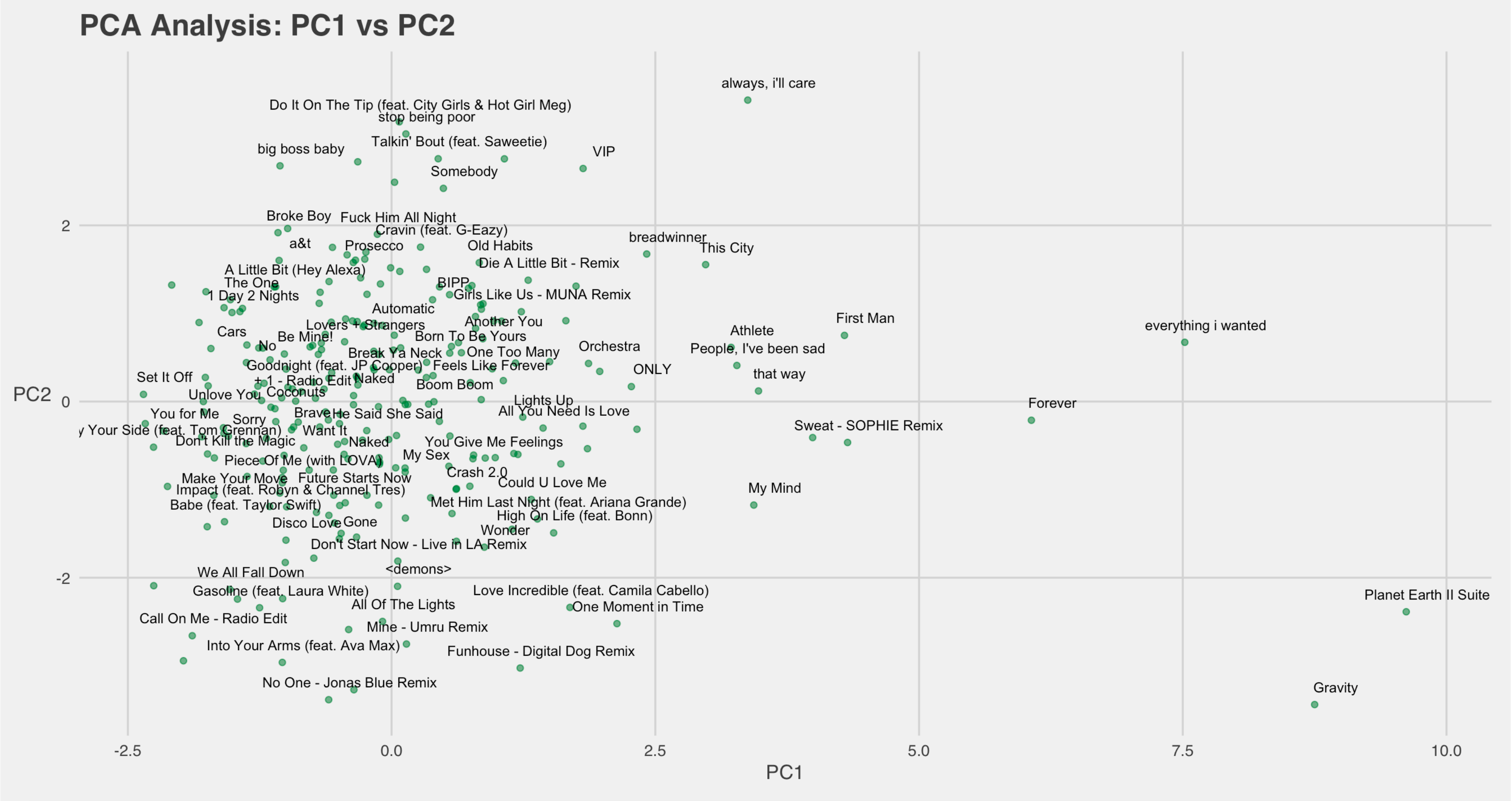


PCA

PC1 vs PC2 scatterplot (2D)

Tracks are not well defined by 2 PC's & are far from 2D subspace

For modelling, use a different dimensionality reduction method or include more features



PCA

Factor Mapping on PC1 vs PC2

Illustrates relationship between features & components

⇩

Most of my “liked” songs are energetics + loud