Music Popularity Predictor

Final Project - Group 8

Matthew Yeon, Neha Gundavarapu, Jenny Belson

TOO MANY DEMO SUBMISSIONS FROM ARTISTS

The listening process is very long, tedious, and many demos never even get heard.

Project Goal

GOAL

Build a popularity predictor model for labels to streamline their process and efficiently select artists

HYPOTHESIS

If we develop a model that accurately predicts which artists or songs will become popular, Eagle Records will retain a competitive spot in the record label industry

HOW IT WORKS

[Hit Predictor Model] [Song Similarity Tool]

Enter Song
search criteria
(enter spotify URL)

Choose Song
by entering an
artist name and
song by that artist

Choose Decade
to see if the song
would be a hit or
not in that decade

Accuracy Test
to evaluate quality
of the logistic
regression model

Search Song
to return a playlist of
similar songs for a
better understanding
of the input song

Enter desired number
of **similar songs**

Receive a playlist
of similar songs

13 Audio Features

These features are used to determine the potential popularity of a song.

- **Danceability** [0.0 ~ 1.0]
- **Acousticness** [0.0 ~ 1.0]
- **Energy** [0.0 ~ 1.0]
- **Instrumentalness** [0.0 ~ 1.0]
- **Liveness** [0.0 ~ 1.0]
- **Loudness** [-60 ~ 0 db]
- **Speechiness** [0.0 ~ 1.0]
- **Tempo** [bpm]
- **Valence** [0.0 ~ 1.0]
- **Duration** [ms]
- **Key** [0 = C, 1 = C?/D?, 2 = D, …]
- **Mode** [major: 1 | minor: 0]
- **Time signature** [0 ~ 5]

[DATASET A]

The Spotify Hit Predictor Dataset (1960 - 2019)

Source : **kaggle.com**

Over 40,000+ tracks labeled hit (1) or flop (0), with features fetched via Spotify's Web API

[DATASET B]

Spotify API Data

Source : **developer.spotify.com**

Data directly accessed by connecting to Spotify's API

Data for Machine Learning

The Spotify Hit Predictor Dataset (1960 - 2019)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	track	artist	uri	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	chorus_hit	sections	target
2	Wild Things	Alessia Cara	spotify:track:2ZyuwVvV6Z3XJaXIFbspeE	0.741	0.626	1	-4.826	0	0.0886	0.02	0	0.0828	0.706	108.029	188493	4	41.18681	10	1
3	Surfboard	Esquivel!	spotify:track:61APOtq25SCMuK0V5w2K	0.447	0.247	5	-14.661	0	0.0346	0.871	0.814	0.0946	0.25	155.489	176880	3	33.18083	9	0
4	Love Someone	Lukas Graham	spotify:track:2JqnpexlO9dmvjUMCaLCLJ	0.55	0.415	9	-6.557	0	0.052	0.161	0	0.108	0.274	172.065	205463	4	44.89147	9	1
5	Music To My Ears (feat. Tory	Keys N Krates	spotify:track:0cjfLhk8WJ3etPTCseKXtk	0.502	0.648	0	-5.698	0	0.0527	0.00513	0	0.204	0.291	91.837	193043	4	29.52521	7	0
6	Juju On That Beat (TZ Anthen	Zay Hilfigerrr & Zayion M	spotify:track:1lItf5ZXJc1by9SbPeljFd	0.807	0.887	1	-3.892	1	0.275	0.00381	0	0.391	0.78	160.517	144244	4	24.99199	8	1
7	Here's To Never Growing Up	Avril Lavigne	spotify:track:0qwcGscxUHGZTgq0zcaqk	0.482	0.873	0	-3.145	1	0.0853	0.0111	0	0.409	0.737	165.084	214320	4	32.17301	12	1
8	Sex Metal Barbie	In This Moment	spotify:track:75BGF4LC7AOLFfxn6ukZDI	0.533	0.935	0	-3.704	1	0.128	0.0139	0	0.168	0.481	140.092	262493	4	21.0451	14	0
9	Helluva Night	Ludacris	spotify:track:0flKDWZq11997Fb2ptkQvu	0.736	0.522	2	-8.02	1	0.116	0.0299	0	0.108	0.369	97.547	200387	4	60.21027	10	1
10	Holiday With HH	No Bros	spotify:track:7LBa0KNFR8IY3g7LOfXqu8	0.166	0.985	7	-2.886	1	0.17	0.00183	0.0142	0.958	0.139	174.725	252787	4	31.23583	11	0
11	My Last	Big Sean Featuring Chris	spotify:track:70tFuqBcduJv15bEnOPRTh	0.387	0.773	8	-5.685	1	0.17	0.098	0	0.209	0.368	78.629	254120	4	23.30245	9	1
12	Break Up In The End	Cole Swindell	spotify:track:5Z19ylxppfnfdP4JH0u8oj	0.507	0.372	1	-8.433	1	0.0303	0.481	0	0.271	0.257	86.422	199693	4	36.66287	10	1
13	Cirrus	Bonobo	spotify:track:2lJ4d8MCT6ZlDRHKJ1br14	0.64	0.844	2	-8.412	0	0.0374	0.395	0.933	0.0827	0.364	119.042	352247	4	80.60317	13	0
14	Theme From "Bus Stop"	Jackie Gleason	spotify:track:5Jd78KUwqhZcY5msCpsDL	0.245	0.0935	5	-19.343	1	0.0373	0.748	0.254	0.0963	0.107	124.385	222787	4	20.536	12	0
15	Crawling Back To You	Daughtry	spotify:track:6BDtTzjbJ5kKKSWcJT8MlX	0.438	0.919	0	-2.91	0	0.0495	0.00674	0	0.158	0.195	151.026	225813	4	34.01444	8	1
16	Maze of Martyr (Official Dom	Dj Mad Dog	spotify:track:1hW21b9IQeETvRqMwnn2	0.32	0.99	1	-2.454	1	0.344	0.00902	0.00032	0.107	0.0424	178.107	232541	4	41.48721	14	0
17	Hotline Bling	Drake	spotify:track:0wwPcA6wtMf6HUMpIRd	0.891	0.625	2	-7.861	1	0.0558	0.00261	0.000176	0.0504	0.548	134.967	267067	4	69.38968	8	1
18	Cut Her Off	KCamp Featuring 2 Chain	spotify:track:2Vevs2eAQNNb7NTpKj5kq	0.769	0.611	8	-2.85	1	0.039	0.098	0	0.221	0.0901	144.037	243333	4	38.21223	12	1
19	Beautiful People	Chris Brown Featuring Be	spotify:track:0iSaO7CfL9NgXdM8Meu2u	0.417	0.806	5	-5.339	0	0.16	0.0703	0.00637	0.0841	0.545	127.887	226773	4	14.73978	10	1
20	Survival	Eminem	spotify:track:3stOygN0I7ClvkEB2LJGbv	0.459	0.899	2	-2.978	1	0.21	0.0038	0	0.126	0.437	176.384	272417	4	30.26082	10	1
21	Squidwards Nose (feat. Kg Pr	Joey Trap	spotify:track:2hLhiONy3FcndFJc2CiC2e	0.634	0.88	1	-1.714	1	0.293	0.169	0	0.474	0.548	150.959	119249	4	37.52226	5	0
22	Windshield	Greensky Bluegrass	spotify:track:7GI1Weh21oGJYeSbrtOyR	0.48	0.548	0	-9.119	1	0.0328	0.627	0.00502	0.205	0.322	90.109	224853	4	24.45777	9	0
23	Don't Speak (Instrumental)	Joseph Sullinger	spotify:track:1DV7nyw8OigFfEiJ3yEFj6	0.526	0.228	0	-12.975	0	0.0542	0.975	0.903	0.106	0.239	143.295	241496	4	40.78298	9	0
24	Faster	Matt Nathanson	spotify:track:6plKFdrBnKF0y3CRuceTDh	0.742	0.853	9	-4.147	1	0.0393	0.00743	4.79E-06	0.332	0.95	107.03	208280	4	43.42073	10	1
25	Sugar	Robin Schulz Featuring Fr	spotify:track:5tf1VVWniHgryyumXyJM7	0.636	0.815	5	-5.098	0	0.0581	0.0185	0	0.163	0.636	123.063	219043	4	31.44339	10	1
26	Badges	Yohuna	spotify:track:4gW4lIdFDob87TaoyREAH	0.481	0.199	7	-14.253	1	0.0326	0.949	0.618	0.12	0.147	127.996	229500	4	47.06457	8	0
27	Art House Director	Broken Social Scene	spotify:track:5feuZknKlJYGMPvWvKrpw	0.464	0.766	7	-5.298	1	0.041	0.000122	0.0275	0.544	0.335	130.458	212173	5	105.47915	6	0
28	10.000 Falls	Logical Terror	spotify:track:774fmDmDlvsKYdcdBnjp1Z	0.429	0.924	11	-6.456	0	0.135	0.00136	9.78E-06	0.345	0.209	180.033	241583	4	62.60494	12	0
29	The Healing Process	Koh Lantana	spotify:track:23puVz6Rhiq8Wax9KxnZtV	0.241	0.0553	3	-23.605	1	0.0336	0.926	0.93	0.108	0.0643	135.859	161442	1	22.19678	11	0
30	Love Don't Run	Steve Holy	spotify:track:1dXUWskP4zy7Inqpfy5hf6	0.512	0.532	0	-3.28	1	0.0301	0.475	0	0.0993	0.256	147.473	219267	4	33.96792	9	1
31	Lockjaw	French Montana Featurin	spotify:track:7iaw359G2XT14uTfV9feip	0.615	0.648	5	-3.792	0	0.22	0.0411	0	0.277	0.26	169.912	223147	4	59.76709	9	1
32	Sleep Hibernation	Moon Laika	spotify:track:0ek2PwrDkUWRqoaTq6WI	0.142	0.013	7	-33.299	1	0.0428	0.984	0.909	0.0994	0.0708	72.917	175132	4	55.05616	7	0
33	You Should See Me In A Crow	Billie Eilish	spotify:track:3XF5xLJHOQQRbWya6hBp	0.678	0.533	4	-10.485	1	0.186	0.462	0.219	0.139	0.323	150.455	180953	4	27.06075	9	1
34	Swish Swish	Katy Perry Featuring Nick	spotify:track:3OtMnyUaiipcAT23A8liyi	0.839	0.705	5	-5.194	0	0.0445	0.0184	1.77E-05	0.102	0.575	119.954	242520	4	25.02456	11	1
35	Hot Tottie	Usher Featuring Jay-Z	spotify:track:1Vot6YSxlnL52SGTN0XN9r	0.654	0.866	6	-4.332	0	0.286	0.00155	5.38E-06	0.0928	0.669	87.525	299333	4	22.25047	15	1
36	New Salem	Misery Index	spotify:track:4Hl8ohhuAt8AvjlGsyfuli	0.329	0.933	1	-4.336	1	0.0552	3.47E-06	4.09E-06	0.059	0.261	104.676	203960	4	29.48355	11	0
37	Raise Your Glass	P!nk	spotify:track:1gv4xPanlmH17bKZ9rOveF	0.7	0.709	7	-5.006	1	0.0839	0.0048	0	0.0289	0.625	122.019	202960	4	23.8094	8	1
38	Mr. Misunderstood	Eric Church	spotify:track:79dlOxvdsCDApoM8XChkp	0.385	0.808	7	-6.67	1	0.045	0.0629	0	0.33	0.546	129.22	319240	4	32.15759	8	1

Dataset of 60s / 70s / 80s / 90s / 00s / 10s

Step 1) Exploratory Logistic Regression & Data Analysis

<u>Data</u> : The Spotify Hit Predictor Dataset (1960 - 2019)

- Built logistic regression model to understand the coefficients of song attributes in a merged dataset of all decades from 1960 to 2019
- Explored how attributes of hits and non-hits varied over the decades and their distributions

	odds_ratio
Intercept	1.357260
danceability	26.015676
energy	0.141125
key	1.010091
loudness	1.116054
mode	1.490145
speechiness	0.041097
acousticness	0.254442
instrumentalness	0.033808
liveness	0.813380
valence	1.529481
tempo	1.001997
time_signature	1.148612
chorus_hit	0.997882

Logit Regression Results

Dep. Variable:	target	**No. Observations:**	41106
Model:	Logit	**Df Residuals:**	41092
Method:	MLE	**Df Model:**	13
Date:	Mon, 29 Nov 2021	**Pseudo R-squ.:**	0.2382
Time:	19:32:21	**Log-Likelihood:**	-21705.
converged:	True	**LL-Null:**	-28493.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3055	0.172	1.773	0.076	-0.032	0.643
danceability	3.2587	0.092	35.318	0.000	3.078	3.440
energy	-1.9581	0.100	-19.658	0.000	-2.153	-1.763
key	0.0100	0.003	3.010	0.003	0.004	0.017
loudness	0.1098	0.004	24.948	0.000	0.101	0.118
mode	0.3989	0.026	15.280	0.000	0.348	0.450
speechiness	-3.1918	0.156	-20.397	0.000	-3.499	-2.885
acousticness	-1.3687	0.053	-26.035	0.000	-1.472	-1.266
instrumentalness	-3.3871	0.067	-50.444	0.000	-3.519	-3.255
liveness	-0.2066	0.069	-2.984	0.003	-0.342	-0.071
valence	0.4249	0.059	7.214	0.000	0.309	0.540
tempo	0.0020	0.000	4.715	0.000	0.001	0.003
time_signature	0.1386	0.032	4.341	0.000	0.076	0.201
chorus_hit	-0.0021	0.001	-3.314	0.001	-0.003	-0.001

Step 2) Logistic Model for Prediction

1) Dropped **non-numerical data** from each decade song data

- **Track, Artist, URI, Chorus Hit, Sections**

2) Divided dataset into **<u>Train</u> (80%)**, **<u>Validation</u> (10%)**, and **<u>Test</u> (10%)** sets

- **Three steps to improve model accuracy**

3) Made predictions on test data for each decade after standardizing predictor sections of data

Step 3a) Evaluation of Logistic Regression Model

Built function to obtain decade-level LR model **accuracy scores** for...

Training Data

	LR Model Training Data Accuracy Scores
1960	0.661941
1970	0.666613
1980	0.688382
1990	0.744339
2000	0.819885
2010	0.801876

Test Data

	LR Model Test Data Accuracy Scores
1960	0.682081
1970	0.656812
1980	0.703757
1990	0.724638
2000	0.831633
2010	0.798752

Step 3b) Evaluation of Logistic Regression Model

Displayed **confusion matrix, precision score, recall score,** & **F1-score** for each decade

	1960s	1970s	1980s	1990s	2000s	2010s
Confusion Matrix	[[267, 161] [114, 323]]	[[215, 152] [115, 296]]	[[240, 97] [108, 247]]	[[200, 104] [48, 200]]	[[209, 71] [28, 280]]	[[238, 95] [34, 274]]
Precision Score	66.74%	66.07%	71.80%	65.79%	79.77%	74.25%
Recall Score	73.91%	72.02%	69.58%	80.65%	90.91%	88.96%
F1 Score	70.14%	68.92%	70.67%	72.46%	84.98%	80.95%

Step 4) Hit or Not Predictor

1) Used **spotipy library** to draw songs from Spotify data as a “**demo**”

2) Extracted **song features** from inputted song and ran through our logistic regression model, filtering on decade

3) Show user graph based on where their song falls on the “**danceability - loudness**’ scale

4) Converted to a **GUI format** using **Tkinter** library for ease and simpler user interface for non-technical record label employees (program is called “Music Moirai”)

Step 5) Song Similarities Generator

A) Included code defining a function to **identify nearest neighbors** through Euclidean distance calculations on numerical song features

B) User prompted to enter a **desired number of similar songs** to output ("neighbors")

C) Returns a **dataframe** of songs and corresponding artists

https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/

GUI : Using Tkinter

LIMITATIONS OF DATA AND PREDICTION MODEL

Defined features are **not the only indicators** that determine the popularity of the music

- Initial Popularity of an Artist
- Content of Lyrics
- Company Marketing
- Luck / Chance

Potential **Machine Bias**

- Historical data could have biases
- Popular trend may change over time

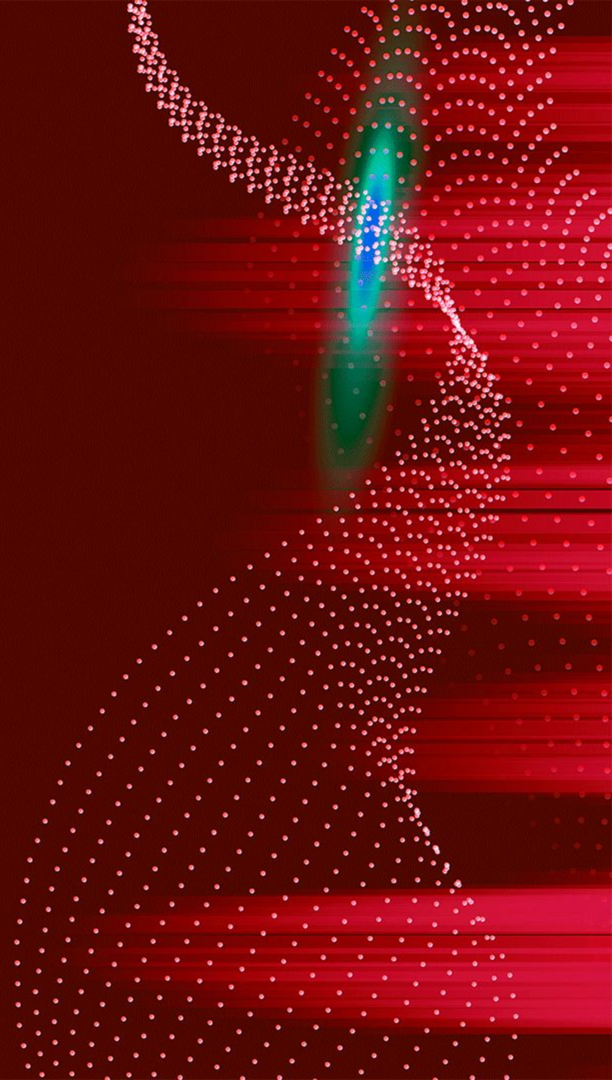
CHALLENGES

Applying **Logistic Regression Model** to predict an inputted songs

- ✓ **Solved by** building a function that trained, fit, and predicted with LR model in one place

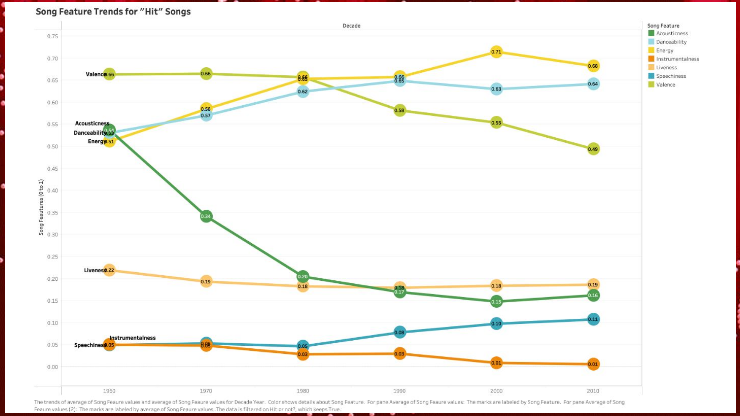
Embedding **error handling mechanisms** that occurs from user inputs

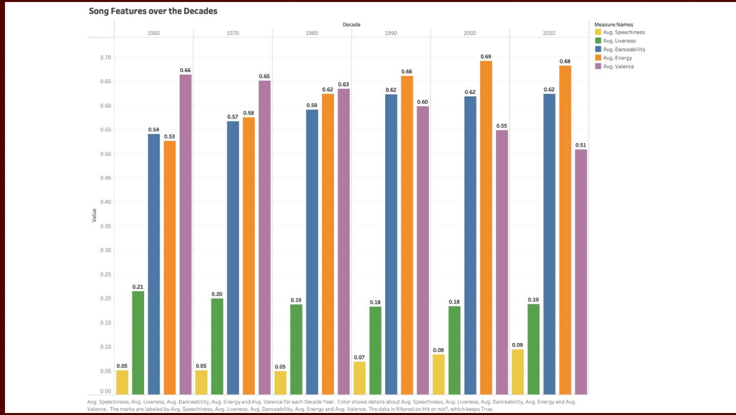
- ✓ **Solved by** using If-else blocks



Model-free Insights

- **Accousticness** & **Instrumentalness** is less indicative of a hit song
- High **Energy** & **Danceable** songs render to a more popular music over time

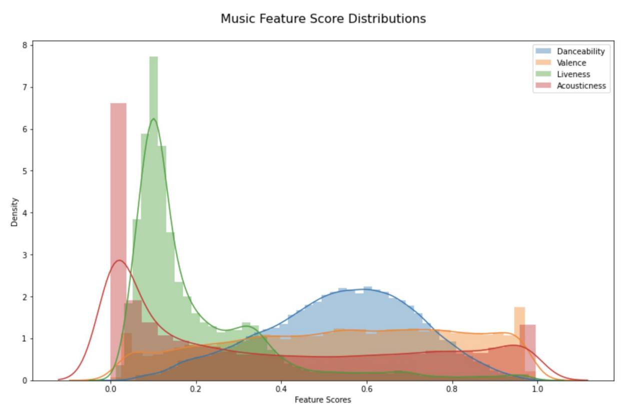




Model-driven Insights

- Logistic Regression Predictive Model based on the Hit Predictor dataset resulted in **60-80% accuracy** in predicting whether a Spotify song would be a hit or not in a specified decade

- Much higher distributions for **danceability** features compared to acousticness or liveness



	danceability	valence	liveness	acousticness
count	41106.000000	41106.000000	41106.000000	41106.000000
mean	0.539695	0.542440	0.201535	0.364197
std	0.177821	0.267329	0.172959	0.338913
min	0.000000	0.000000	0.013000	0.000000
25%	0.420000	0.330000	0.094000	0.039400
50%	0.552000	0.558000	0.132000	0.258000
75%	0.669000	0.768000	0.261000	0.676000
max	0.988000	0.996000	0.999000	0.996000

SWOT ANALYSIS

Strength

Highly accurate prediction model (60-80% accuracy score)

Weakness

Current model doesn’t include non-numeric indications of a song in the analysis (ex. lyrics)

Opportunity

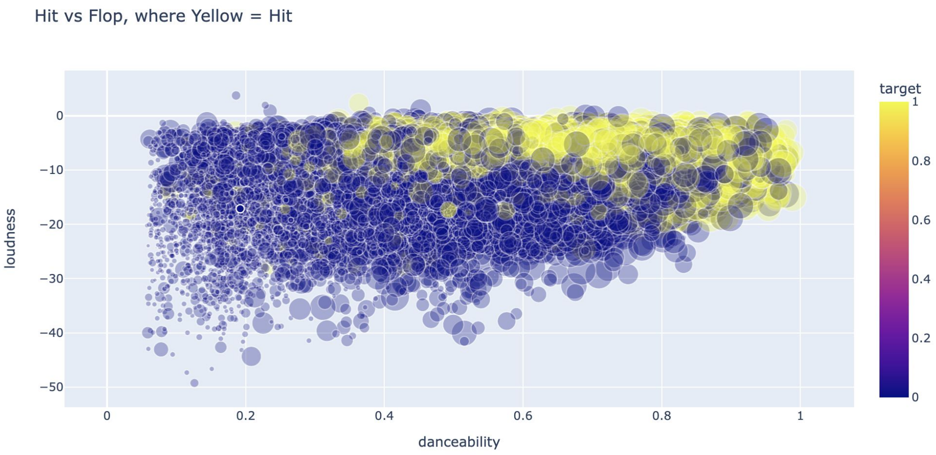
Improve the model to extract song attributes without relying on Spotify API
Develop algorithm that also analyze non-numeric features of the song to better predict hit songs

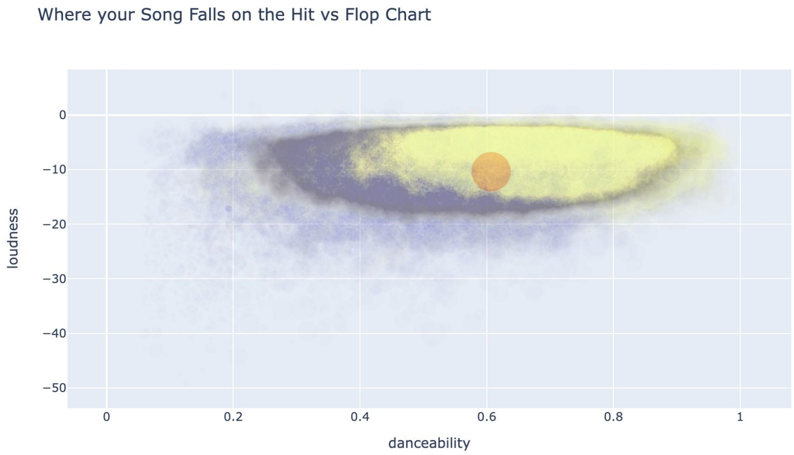
Threat

Tiktok, Youtube, or Spotify that also use advanced algorithm to predict hit songs

Current & Future Trend

Fast-paced and **Synthetically-sounding** or **Computer-generated-like songs**
(Ex. Dance Pop, EDM, Hip-Hop)





THANKS!

Any Questions?