SpotifAI

Mid-Program Presentation

Presented by:
Aishwarya Kandam Vitta
Rishie Nandhan Babu
Rachel Lindor
Aadit Grover
Shashank Patoju

Group 4
Date: 11/1/2024

What is SpotifAI

- This project aims to build a recommendation system for spotify users based on their listening habits.
- Insights from the spotify dataset taken along with user specific data to recommend songs that the user might like!

Dataset Exploration

- Dataset appears mostly complete, with no missing values in any columns.
- Song duration was converted from milliseconds to seconds to enhance readability and ease of interpretation. Since most listeners think in terms of seconds and minutes, this conversion could make the data more intuitive. It also simplifies calculations and comparisons without affecting data integrity.

Current Progress

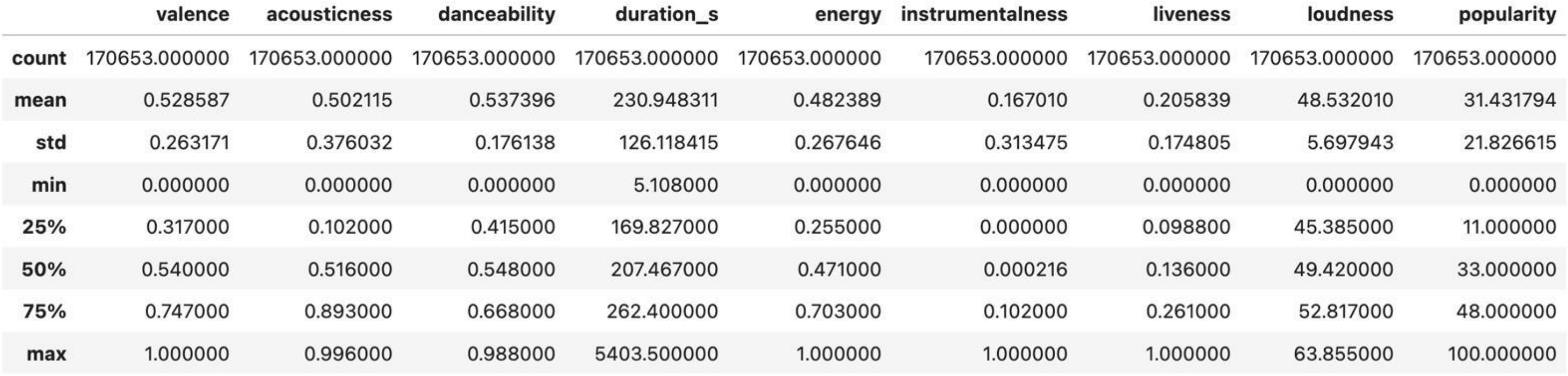
Progress till date:

- Data clean up and deriving parameters from already existing data.
- Performing EDA to find out how parameters in the dataset are related to each other.
- Deriving common statistics on the Spotify dataset.
- Correlating all the parameters to pick features that we may use moving forward to build upon our recommendation system.

General Dataset Statistics

- The following parameters were calculated on the dataset using pandas lib

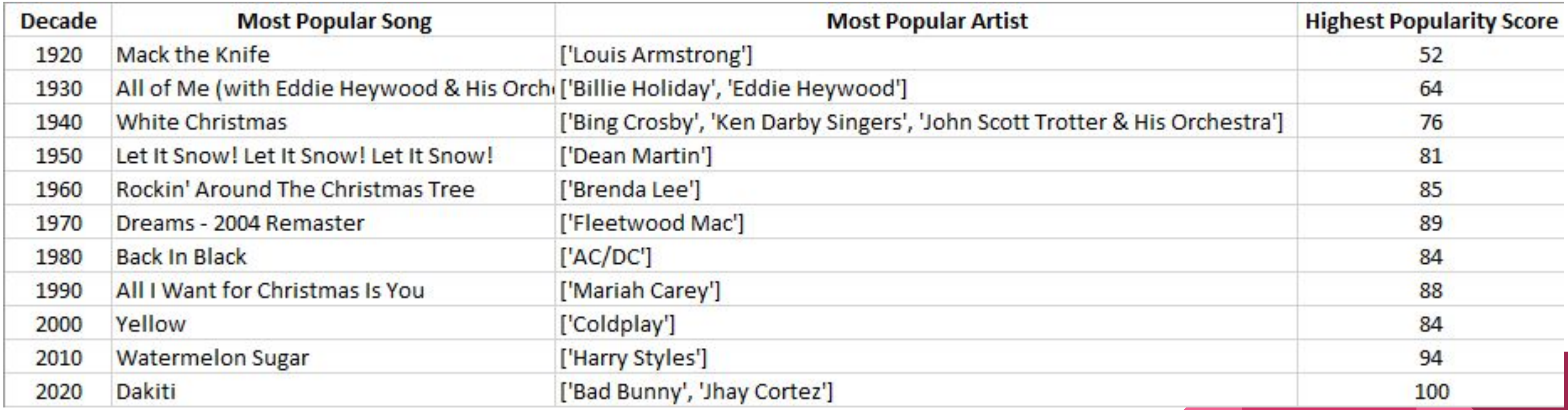
  Count, mean, std, min, 25%, 50%, 75% and max



	valence	acousticness	danceability	duration_s	energy	instrumentalness	liveness	loudness	popularity
count	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000
mean	0.528587	0.502115	0.537396	230.948311	0.482389	0.167010	0.205839	48.532010	31.431794
std	0.263171	0.376032	0.176138	126.118415	0.267646	0.313475	0.174805	5.697943	21.826615
min	0.000000	0.000000	0.000000	5.108000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.317000	0.102000	0.415000	169.827000	0.255000	0.000000	0.098800	45.385000	11.000000
50%	0.540000	0.516000	0.548000	207.467000	0.471000	0.000216	0.136000	49.420000	33.000000
75%	0.747000	0.893000	0.668000	262.400000	0.703000	0.102000	0.261000	52.817000	48.000000
max	1.000000	0.996000	0.988000	5403.500000	1.000000	1.000000	1.000000	63.855000	100.000000

Some Top Statistics from Dataset

Most Popular Song by Decade

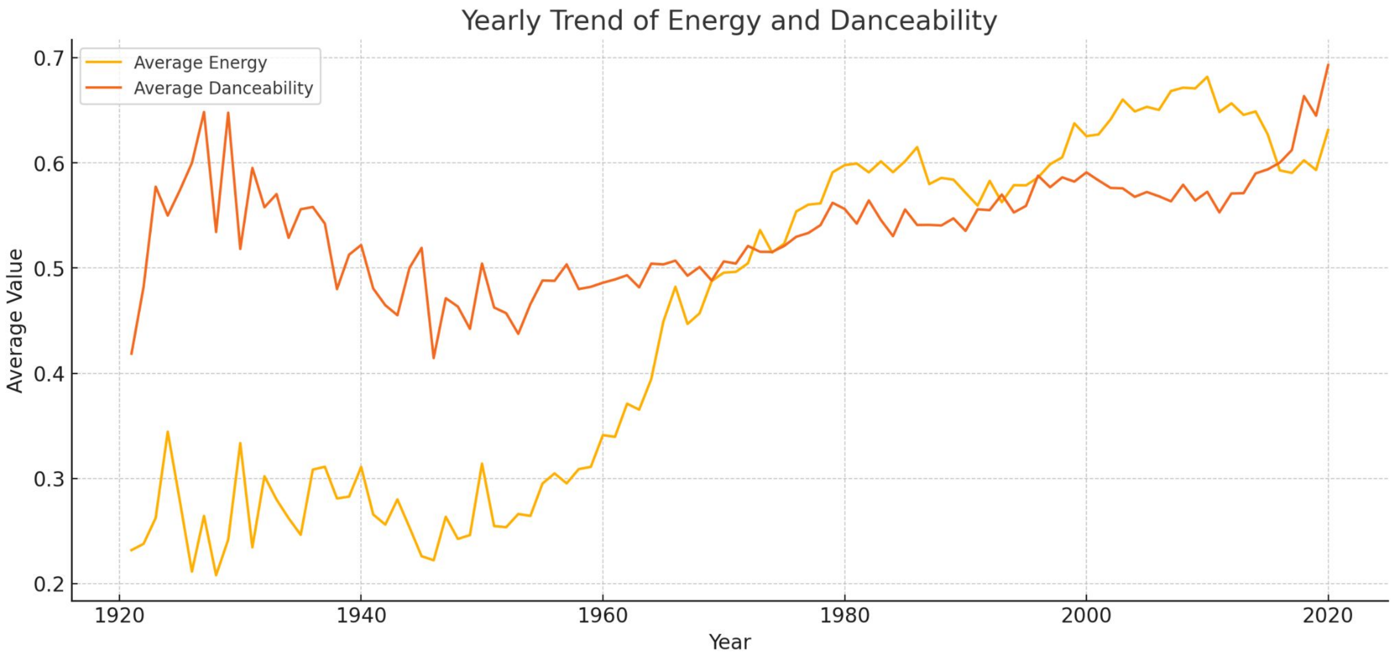


Decade	Most Popular Song	Most Popular Artist	Highest Popularity Score
1920	Mack the Knife	['Louis Armstrong']	52
1930	All of Me (with Eddie Heywood & His Orch	['Billie Holiday', 'Eddie Heywood']	64
1940	White Christmas	['Bing Crosby', 'Ken Darby Singers', 'John Scott Trotter & His Orchestra']	76
1950	Let It Snow! Let It Snow! Let It Snow!	['Dean Martin']	81
1960	Rockin' Around The Christmas Tree	['Brenda Lee']	85
1970	Dreams - 2004 Remaster	['Fleetwood Mac']	89
1980	Back In Black	['AC/DC']	84
1990	All I Want for Christmas Is You	['Mariah Carey']	88
2000	Yellow	['Coldplay']	84
2010	Watermelon Sugar	['Harry Styles']	94
2020	Dakiti	['Bad Bunny', 'Jhay Cortez']	100

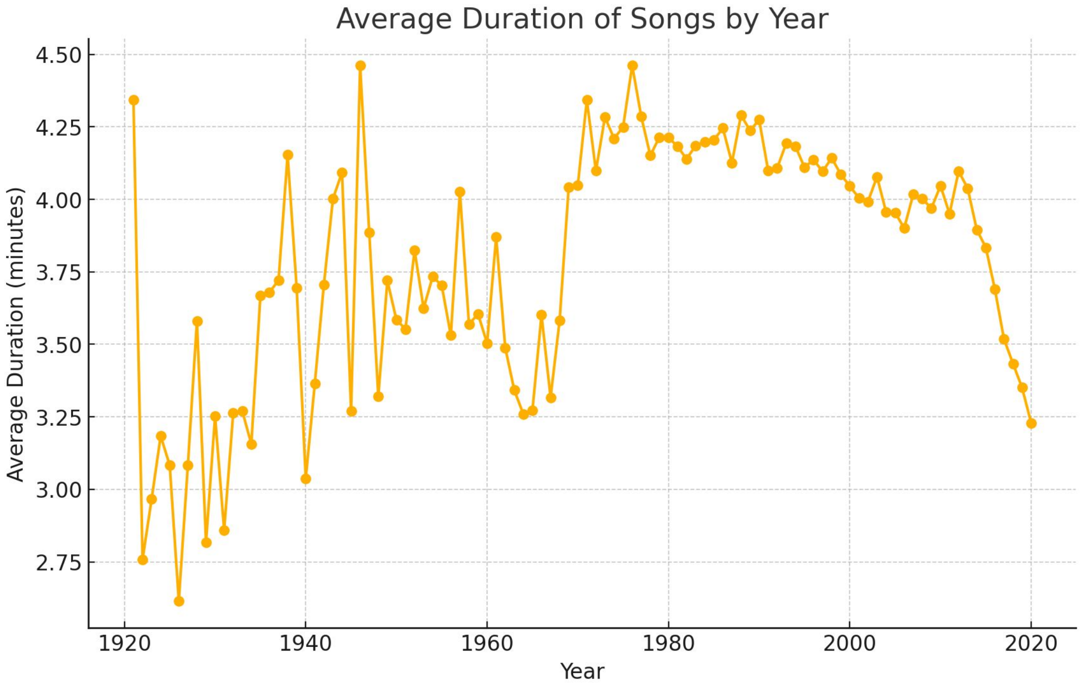
***Please note the some decades had multiple popular songs. For examples, the 1980s decade has 3 songs with an 84 popularity score and the 2000s has 4 songs with an 84 pop**

Data Visualization

Yearly Trend of Energy and Danceability



Average duration of songs by year



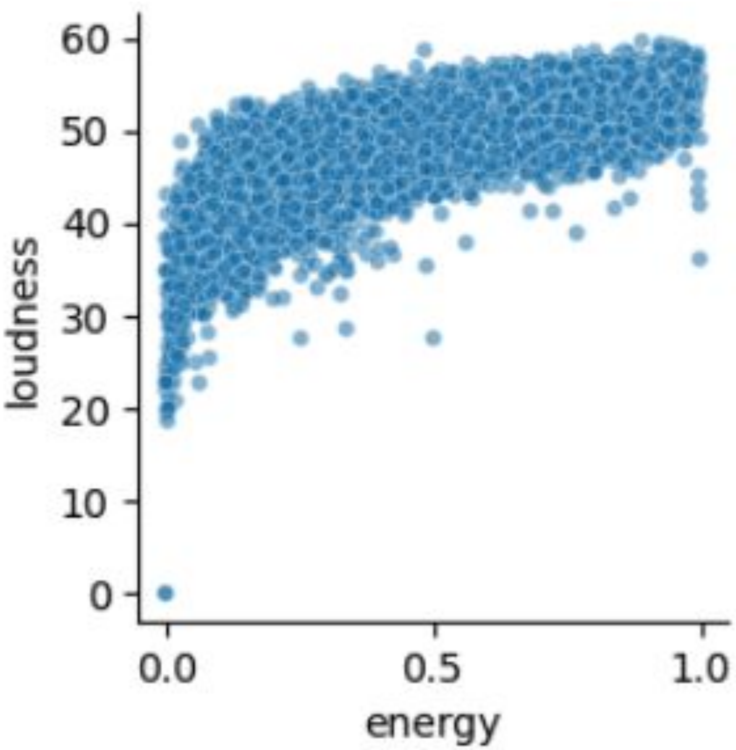
Shows us how on an average song duration has changed over the years!

Notice a general fall in duration from 2000s up until 2020.

Correlations

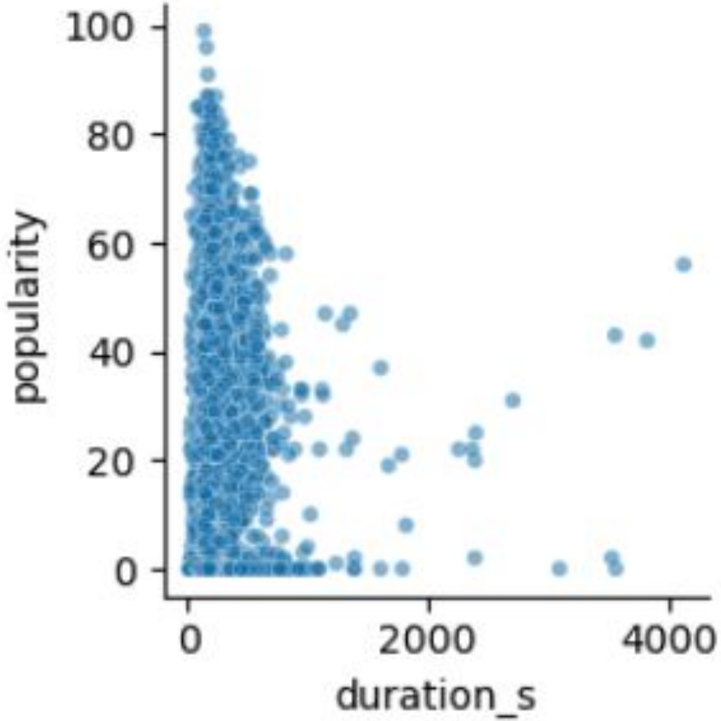
- Correlations between the different parameters in the dataset give an understanding of how closely the parameters are related.
- Positive correlation means the parameters move together
- Negative correlation means the parameters move opposite each other
- Vital for recommendation systems such as ours
- Helps making informed decisions and predictions.

Loudness and Energy Correlation



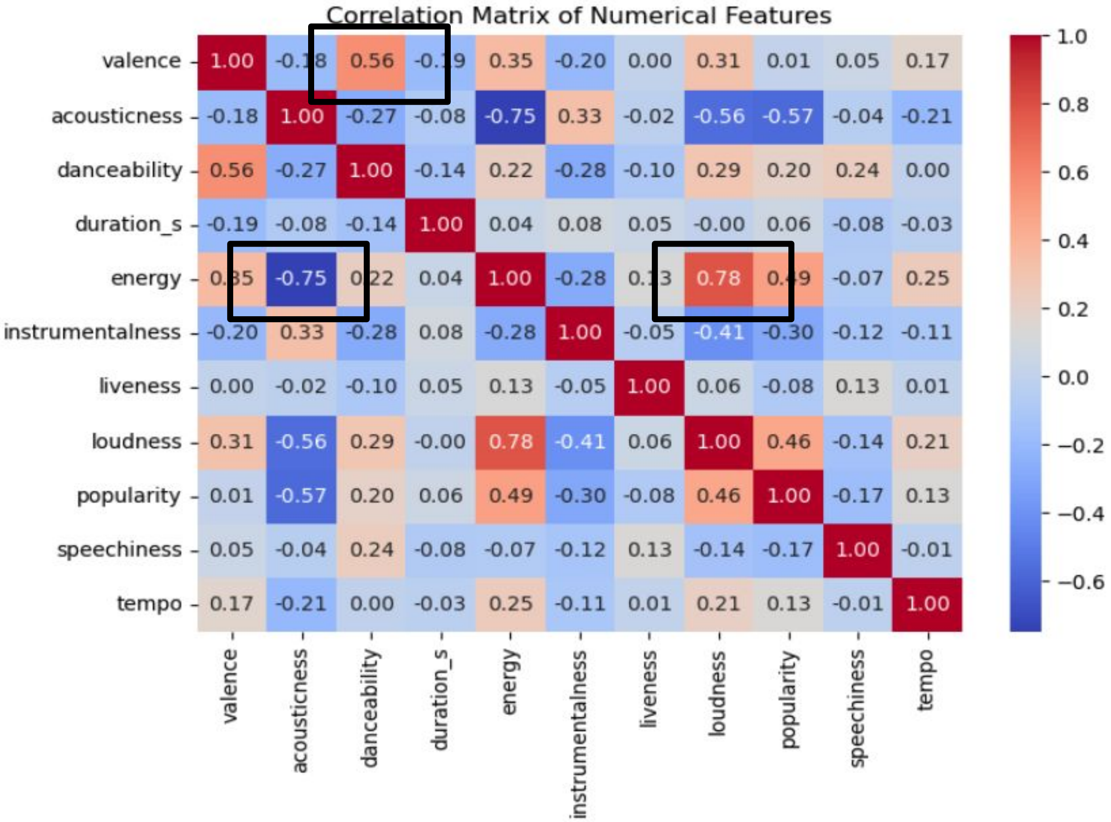
- Loudness, energy seem to be positively correlated.
- Note: Sampled 10,000 data points (from 170k) for clearer visual

Popularity and Duration(s) Correlation



- Shorter songs seem to be more popular
- Note: Sampled 10,000 data points (from 170k) for clearer visual

Correlating all parameters



Correlation Matrix of Numerical Features

	valence	acousticness	danceability	duration_s	energy	instrumentalness	liveness	loudness	popularity	speechiness	tempo
valence	1.00	-0.18	0.56	-0.19	0.35	-0.20	0.00	0.31	0.01	0.05	0.17
acousticness	-0.18	1.00	-0.27	-0.08	-0.75	0.33	-0.02	-0.56	-0.57	-0.04	-0.21
danceability	0.56	-0.27	1.00	-0.14	0.22	-0.28	-0.10	0.29	0.20	0.24	0.00
duration_s	-0.19	-0.08	-0.14	1.00	0.04	0.08	0.05	-0.00	0.06	-0.08	-0.03
energy	0.35	-0.75	0.22	0.04	1.00	-0.28	0.13	0.78	0.49	-0.07	0.25
instrumentalness	-0.20	0.33	-0.28	0.08	-0.28	1.00	-0.05	-0.41	-0.30	-0.12	-0.11
liveness	0.00	-0.02	-0.10	0.05	0.13	-0.05	1.00	0.06	-0.08	0.13	0.01
loudness	0.31	-0.56	0.29	-0.00	0.78	-0.41	0.06	1.00	0.46	-0.14	0.21
popularity	0.01	-0.57	0.20	0.06	0.49	-0.30	-0.08	0.46	1.00	-0.17	0.13
speechiness	0.05	-0.04	0.24	-0.08	-0.07	-0.12	0.13	-0.14	-0.17	1.00	-0.01
tempo	0.17	-0.21	0.00	-0.03	0.25	-0.11	0.01	0.21	0.13	-0.01	1.00

- Preprocessing of ‘loudness’: Added a constant to each value to shift the range into the positive domain.
- Gives us an accurate idea of which features are very closely related to each other.
- We plan to pick features based on this data moving forward.

Thank you!