Exploratory Data Analysis

Presented by:
-Akshay Gujare
77418analyst@gmail.com

About the spotify

Table of content

About the Spotify

Spotify is one of the world's leading digital music and audio streaming platforms, offering access to millions of songs, podcasts, and curated playlists. Launched in 2008, Spotify revolutionized the music industry by introducing a seamless, on-demand streaming experience supported by personalized recommendations powered by advanced data analytics and machine learning. With features like Discover Weekly, Daily Mixes, and algorithm-driven suggestions, Spotify enhances user engagement by tailoring content to individual listening habits.

Today, the platform operates globally, partnering with artists, creators, and record labels to deliver high-quality audio content while continuously expanding its catalog and improving user experience.

Spotify's data-driven ecosystem enables deep insights into listening trends, user behavior, and content performance. Its innovative technology, cross-platform availability, and personalized experience make it a dominant force in modern audio streaming.

Objective

ABOUT THE DATASET :

The dataset contains information about various songs available on Spotify. It includes details such as the song name, artist, album, and release year, along with numerical audio features like danceability, energy, loudness, speechiness, instrumentalness, valence, and tempo. Additionally, it provides a popularity score for each track, which indicates how well a song is performing on the platform. This dataset helps in analyzing musical trends, artist influence, and the characteristics that make a song popular.

EDA GOAL OF THE PROJECT :

The goal of this Exploratory Data Analysis (EDA) is to study the dataset and identify patterns related to song popularity, artist trends, and audio features. The main objectives is to Analyze Spotify data to uncover insights about music trends,
user preferences, and song characteristics.

TOOLS

python

NumPy

matplotlib

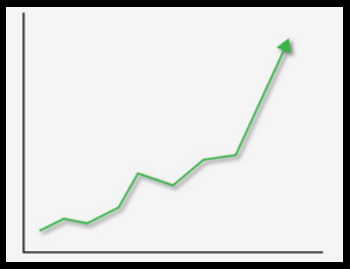
seaborn

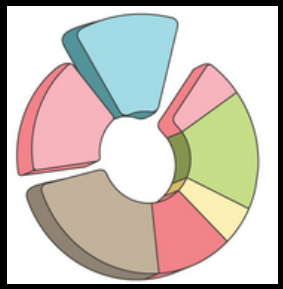
pandas

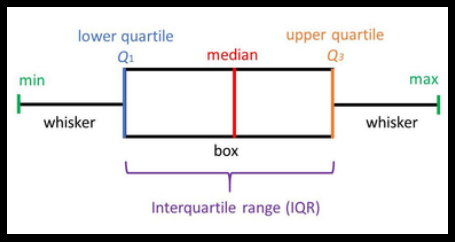
plotly

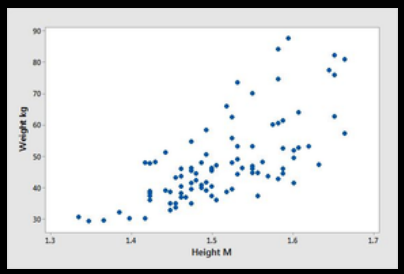
Spotify®

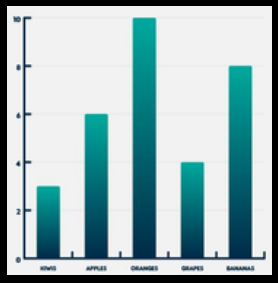
GRAPHS











Data Overview

- Loading and preparing loan application data for analysis in google Colab notebook.

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

    Mounted at /content/drive
```

- In this analysis,we import Pandas for data manipulation ,Numpy for numerical operations ,Matplotlib for creating visualization and Seaborn and plotly for enhanced statistical graphic.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly as px
```

- We import the Google drive module to mount Google `Drive`, enabiling access to files and dataset in the for our analysis.

```
df_1 = pd.read_csv("/content/drive/MyDrive/spotify capstone/data.csv")
df_2 = pd.read_csv("/content/drive/MyDrive/spotify capstone/data_by_artist.csv")
df_3 = pd.read_csv("/content/drive/MyDrive/spotify capstone/data_by_genres.csv")
df_4 = pd.read_csv("/content/drive/MyDrive/spotify capstone/data_by_year.csv")
df_5 = pd.read_csv("/content/drive/MyDrive/spotify capstone/data_w_genres.csv")
```

Data Overview

- **Overview of the project and its objective**

This project focuses on performing Exploratory Data Analysis (EDA) on a Spotify Tracks Dataset, aiming to extract meaningful insights about musical trends, artist influence, and song characteristics. By analyzing various attributes of songs such as danceability, energy, loudness, speechiness, instrumentalness, valence, and tempo, this project seeks to uncover key patterns that define a song's success and popularity.

EDA GOAL OF THE PROJECT

The goal of this Exploratory Data Analysis (EDA) is to study the dataset and identify patterns related to song popularity, artist trends, and audio features. The main objectives

1.Understanding which audio features (e.g., danceability, energy, tempo) contribute to a song's popularity.
2.Analyzing trends in music across different artists and genres.
3.Finding correlations between different audio features.
4.Examining how music trends have changed over the years.

By visualizing and analyzing this data, we can gain useful insights into what makes a song successful on Spotify

Data Overview

- **Dataset Used:**

This project uses five interconnected Spotify datasets that provide detailed information about tracks, artists, genres, and musical trends across years. These datasets contain both audio features and popularity metrics, enabling comprehensive exploratory analysis.

1.data.csv (df_1) – Track-Level Dataset

This dataset contains individual Spotify tracks along with their audio features such as danceability, energy, valence, tempo, loudness, and acousticness.
It also includes metadata like track name, artist, release year, duration, and popularity. This dataset is used to study song characteristics and identify patterns across tracks.

2.data_by_artist.csv (df_2) – Artist-Level Dataset

This dataset aggregates information per artist. It includes average audio features, popularity scores, and other summary statistics for each artist. It helps in understanding artist-level trends, comparing musical styles, and identifying top-performing artists.

Data Overview

3.data_by_genres.csv (df_3) – Genre-Level Dataset

This dataset provides aggregated insights at the genre level. It includes average values of audio features across genres, allowing analysis of how different genres vary in terms of energy, acousticness, tempo, loudness, etc. It is useful for comparing musical characteristics across genres.

4.data_by_year.csv (df_4) – Year-Level Dataset

This dataset summarizes music trends by release year. It includes aggregated audio features and popularity patterns across years. It supports trend analysis, such as how tempo, energy, valence, or loudness have evolved over time in the music industry.

5.data_w_genres.csv (df_5) – Track-Level Dataset with Genres

This dataset contains detailed genre information connected to individual tracks. Unlike data.csv, it supports multi-genre mapping and provides more granular genre-based analysis. It is especially useful for studying genre influence on popularity and clustering tracks based on features.

Description of columns

- **Acousticness –** Confidence score (0.0 to 1.0) measuring how acoustic the track is
  – 1.0 = fully acoustic (e.g., unplugged guitar,orchestra
  – 0.0 = fully electronic/synthetic

- **danceability –** How suitable a track is for dancing (0.0 to 1.0)
  – Based on tempo, rhythm stability, beat strength, and regularity
  – Higher values = easier to dance to

- **duration_ms –** Length of the track in milliseconds – Commonly converted to minutes/seconds for analysis

- **energy –** Perceptual measure of intensity and activity (0.0 to 1.0)
  – High-energy tracks feel fast, loud, and noisy
  – Low-energy tracks feel calm and relaxed

- **explicit –** Binary indicator: 1 = contains explicit lyrics, 0 = clean version

- **id –** Unique Spotify URI/ID for the track (used for API look-ups)

Description of columns

- **instrumentalness –** Predicts whether a track contains no vocals (0.0 to 1.0) – Values > 0.5 typically indicate instrumental tracks.
- **key –** The musical key the track is in (integer notation) – 0 = C, 1 = C♯/D♭, 2 = D, ..., 11 = B
- **liveness –** Detects the presence of an audience (0.0 to 1.0) – Values > 0.8 suggest a high probability of live recording.
- **loudness –** Overall loudness of the track in decibels (dB) – Typically ranges from –60 to 0 dB – Modern pop usually falls between –10 and –5 dB
- **mode –** Modality of the track: 1 = major (happier), 0 = minor (sadder)
- **name –** Title of the track
- **popularity –** Spotify popularity score (0–100) – Based on total streams, recency, and playlist placements – 100 = currently most popular
- **release_date / year –** Original release date (full date or just year)

Description of columns

- **speechiness –** Measures presence of spoken words (0.0 to 1.0) – > 0.66 = likely entire spoken/podcast – 0.33–0.66 = rap-heavy or mixed – < 0.33 = mostly singing

- **tempo –** Estimated overall tempo in beats per minute (BPM)

- **valence –** Musical positiveness conveyed by the track (0.0 to 1.0) – High valence = happy, cheerful, euphoric – Low valence = sad, depressed, angry

- **artists –** Name(s) of the performing artist(s) – Multiple artists separated by comma or “&”

- **genres –** List of genres associated with the artist or track – Multiple genres per entry (comma-separated or array)

- **count –** Number of tracks used to calculate the aggregated statistics – Appears only in artist/genre/year-level dataframes

Data Cleaning and Pre-Processing

```
df_1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 170653 entries, 0 to 170652
Data columns (total 19 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   valence           170653 non-null  float64
 1   year              170653 non-null  int64
 2   acousticness      170653 non-null  float64
 3   artists           170653 non-null  object
 4   danceability      170653 non-null  float64
 5   duration_ms       170653 non-null  int64
 6   energy            170653 non-null  float64
 7   explicit          170653 non-null  int64
 8   id                170653 non-null  object
 9   instrumentalness  170653 non-null  float64
 10  key               170653 non-null  int64
 11  liveness          170653 non-null  float64
 12  loudness          170653 non-null  float64
 13  mode              170653 non-null  int64
 14  name              170653 non-null  object
 15  popularity        170653 non-null  int64
 16  release_date      170653 non-null  object
 17  speechiness       170653 non-null  float64
 18  tempo             170653 non-null  float64
dtypes: float64(9), int64(6), object(4)
memory usage: 24.7+ MB
```

The df_1 dataset contains 170,653 Spotify tracks with 19 detailed audio and metadata features. It includes attributes like danceability, energy, valence, tempo, loudness, and more, which describe each song's musical characteristics. The dataset also provides artist names, track names, release year, and popularity, allowing trend and popularity analysis.

```
df_2.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28680 entries, 0 to 28679
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   mode              28680 non-null  int64
 1   count             28680 non-null  int64
 2   acousticness      28680 non-null  float64
 3   artists           28680 non-null  object
 4   danceability      28680 non-null  float64
 5   duration_ms       28680 non-null  float64
 6   energy            28680 non-null  float64
 7   instrumentalness  28680 non-null  float64
 8   liveness          28680 non-null  float64
 9   loudness          28680 non-null  float64
 10  speechiness       28680 non-null  float64
 11  tempo             28680 non-null  float64
 12  valence           28680 non-null  float64
 13  popularity        28680 non-null  float64
 14  key               28680 non-null  int64
dtypes: float64(11), int64(3), object(1)
memory usage: 3.3+ MB
```

This dataset contains 28,680 artists, each with aggregated musical features such as energy, danceability, acousticness, loudness, and popularity. It summarizes how each artist's music typically sounds and includes a count column indicating how many tracks contributed to their averages.

Data Cleaning and Pre-Processing

```
df_3.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2973 entries, 0 to 2972
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   mode              2973 non-null   int64
 1   genres            2973 non-null   object
 2   acousticness      2973 non-null   float64
 3   danceability      2973 non-null   float64
 4   duration_ms       2973 non-null   float64
 5   energy            2973 non-null   float64
 6   instrumentalness  2973 non-null   float64
 7   liveness          2973 non-null   float64
 8   loudness          2973 non-null   float64
 9   speechiness       2973 non-null   float64
 10  tempo             2973 non-null   float64
 11  valence           2973 non-null   float64
 12  popularity        2973 non-null   float64
 13  key               2973 non-null   int64
dtypes: float64(11), int64(2), object(1)
memory usage: 325.3+ KB
```

This dataset includes 2,973 musical genres, each with averaged audio features such as danceability, energy, loudness, valence, and tempo. It provides a high-level summary of how different genres typically sound and how popular they are on Spotify.

```
df_4.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   mode              100 non-null    int64
 1   year              100 non-null    int64
 2   acousticness      100 non-null    float64
 3   danceability      100 non-null    float64
 4   duration_ms       100 non-null    float64
 5   energy            100 non-null    float64
 6   instrumentalness  100 non-null    float64
 7   liveness          100 non-null    float64
 8   loudness          100 non-null    float64
 9   speechiness       100 non-null    float64
 10  tempo             100 non-null    float64
 11  valence           100 non-null    float64
 12  popularity        100 non-null    float64
 13  key               100 non-null    int64
dtypes: float64(11), int64(3)
memory usage: 11.1 KB
```

This dataset contains 100 years of aggregated Spotify music data, with each row representing the average audio and popularity characteristics for a particular year. It captures long-term trends in features like energy, danceability, loudness, tempo, and valence over time.

Data Cleaning and Pre-Processing

```
df_5.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28680 entries, 0 to 28679
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   genres           28680 non-null  object
 1   artists          28680 non-null  object
 2   acousticness     28680 non-null  float64
 3   danceability     28680 non-null  float64
 4   duration_ms      28680 non-null  float64
 5   energy           28680 non-null  float64
 6   instrumentalness 28680 non-null  float64
 7   liveness         28680 non-null  float64
 8   loudness         28680 non-null  float64
 9   speechiness      28680 non-null  float64
 10  tempo            28680 non-null  float64
 11  valence          28680 non-null  float64
 12  popularity       28680 non-null  float64
 13  key              28680 non-null  int64
 14  mode             28680 non-null  int64
 15  count            28680 non-null  int64
dtypes: float64(11), int64(3), object(2)
memory usage: 3.5+ MB
```

Insights:

This dataset contains 28,680 rows, each representing a unique combination of artist and genre along with averaged musical characteristics. It includes key audio features such as acousticness, energy, danceability, valence, loudness, and tempo, along with a popularity score reflecting listener engagement. The count column shows how many tracks contributed to each artist–genre profile, helping identify dominant styles. With both artist and genre information present, this dataset is ideal for analyzing genre-wise artist behavior, music style clustering, and popularity trends.

Data Cleaning and Pre-Processing

- Preview the Data

```
df_1.head()
```

	valence	year	acousticness	artists	danceability	duration_ms	energy	explicit	id	instrumentalness	key	liveness	loudness	mode	name	popularity	release_date	speechiness	tempo
0	0.0594	1921	0.982	['Sergei Rachmaninoff', 'James Levine', 'Berli...	0.279	831667	0.211	0	4BJqT0PrAfrxzMOxytFOIz	0.878000	10	0.665	-20.096	1	Piano Concerto No. 3 in D Minor, Op. 30: III. ...	4	1921	0.0366	80.954
1	0.9630	1921	0.732	['Dennis Day']	0.819	180533	0.341	0	7xPhfUan2yNtyFG0cUWkt8	0.000000	7	0.160	-12.441	1	Clancy Lowered the Boom	5	1921	0.4150	60.936
2	0.0394	1921	0.961	['KHP Kridhamardawa Karaton Ngayogyakarta Hadi...	0.328	500062	0.166	0	1o8I8BglA6ylDMrIELygv1	0.913000	3	0.101	-14.850	1	Gati Bali	5	1921	0.0339	110.339
3	0.1650	1921	0.967	['Frank Parker']	0.275	210000	0.309	0	3fBPsC5vPBKxYSee08FDH	0.000028	5	0.381	-9.316	1	Danny Boy	3	1921	0.0354	100.109
4	0.2530	1921	0.957	['Phil Regan']	0.418	166693	0.193	0	4d0HGyGT8e121BsdKmw9v6	0.000002	3	0.229	-10.096	1	When Irish Eyes Are Smiling	2	1921	0.0380	101.665

```
df_1.tail()
```

	valence	year	acousticness	artists	danceability	duration_ms	energy	explicit	id	instrumentalness	key	liveness	loudness	mode	name	popularity	release_date	speechiness	tempo
170648	0.608	2020	0.08460	['Anuel AA', 'Daddy Yankee', 'KAROL G', 'Ozuna...	0.786	301714	0.808	0	0KkIkfsLEJbrcIhYsCL7L5	0.000289	7	0.0822	-3.702	1	China	72	2020-05-29	0.0881	105.029
170649	0.734	2020	0.20600	['Ashnikko']	0.717	150654	0.753	0	0OS8KKAuXlxA0fMH54Qs8E	0.000000	7	0.1010	-6.020	1	Halloweenie III: Seven Days	68	2020-10-23	0.0605	137.936
170650	0.637	2020	0.10100	['MAMAMOO']	0.634	211280	0.858	0	4BZXVFYCb76Q0Klojq4pIV	0.000009	4	0.2580	-2.226	0	AYA	76	2020-11-03	0.0809	91.688
170651	0.195	2020	0.00998	['Eminem']	0.671	337147	0.823	1	5SiZJoLXp3WOI3J4C8IK0d	0.000008	2	0.6430	-7.161	1	Darkness	70	2020-01-17	0.3080	75.055
170652	0.642	2020	0.13200	['KEVVO', 'J Balvin']	0.856	189507	0.721	1	7HmnJHfs0BkFzX4x8j0hkl	0.004710	7	0.1820	-4.928	1	Billetes Azules (with J Balvin)	74	2020-10-16	0.1080	94.991

- Check Missing Values

```
df_1.isnull().sum()
```

	0
valence	0
year	0
acousticness	0
artists	0
danceability	0
duration_ms	0
energy	0
explicit	0
id	0
instrumentalness	0
key	0
liveness	0
loudness	0
mode	0
name	0
popularity	0
release_date	0
speechiness	0
tempo	0

```
df_2.isnull().sum()
```

	0
mode	0
count	0
acousticness	0
artists	0
danceability	0
duration_ms	0
energy	0
instrumentalness	0
liveness	0
loudness	0
speechiness	0
tempo	0
valence	0
popularity	0
key	0

```
df_3.isnull().sum()
```

	0
mode	0
genres	0
acousticness	0
danceability	0
duration_ms	0
energy	0
instrumentalness	0
liveness	0
loudness	0
speechiness	0
tempo	0
valence	0
popularity	0
key	0

Data Cleaning and Pre-Processing

```
df_4.isnull().sum()
```

	0
mode	0
year	0
acousticness	0
danceability	0
duration_ms	0
energy	0
instrumentalness	0
liveness	0
loudness	0
speechiness	0
tempo	0
valence	0
popularity	0
key	0

```
df_5.isnull().sum()
```

	0
genres	0
artists	0
acousticness	0
danceability	0
duration_ms	0
energy	0
instrumentalness	0
liveness	0
loudness	0
speechiness	0
tempo	0
valence	0
popularity	0
key	0
mode	0
count	0

These 5 dataframes does not contaain any null values, so there is no further needed to be handled.

- **Check Duplicate Rows**

```
df_1.duplicated().sum()
```

```
np.int64(0)
```

Data Cleaning and Pre-Processing

```
df_2.duplicated().sum()
```

```
np.int64(0)
```

```
df_3.duplicated().sum()
```

```
np.int64(0)
```

```
df_4.duplicated().sum()
```

```
np.int64(0)
```

```
df_5.duplicated().sum()
```

```
np.int64(0)
```

No duplicate values are found in any dataframes

- **Summary Statistics**

```
df_1.describe()
```

	valence	year	acousticness	danceability	duration_ms	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo
count	170653.000000	170653.000000	170653.000000	170653.000000	1.706530e+05	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000	170653.000000
mean	0.528587	1976.787241	0.502115	0.537396	2.309483e+05	0.482389	0.084575	0.167010	5.199844	0.205839	-11.467990	0.706902	31.431794	0.098393	116.861590
std	0.263171	25.917853	0.376032	0.176138	1.261184e+05	0.267646	0.278249	0.313475	3.515094	0.174805	5.697943	0.455184	21.826615	0.162740	30.708533
min	0.000000	1921.000000	0.000000	0.000000	5.108000e+03	0.000000	0.000000	0.000000	0.000000	0.000000	-60.000000	0.000000	0.000000	0.000000	0.000000
25%	0.317000	1956.000000	0.102000	0.415000	1.698270e+05	0.255000	0.000000	0.000000	2.000000	0.098800	-14.615000	0.000000	11.000000	0.034900	93.421000
50%	0.540000	1977.000000	0.516000	0.548000	2.074670e+05	0.471000	0.000000	0.000216	5.000000	0.136000	-10.580000	1.000000	33.000000	0.045000	114.729000
75%	0.747000	1999.000000	0.893000	0.668000	2.624000e+05	0.703000	0.000000	0.102000	8.000000	0.261000	-7.183000	1.000000	48.000000	0.075600	135.537000
max	1.000000	2020.000000	0.996000	0.988000	5.403500e+06	1.000000	1.000000	1.000000	11.000000	1.000000	3.855000	1.000000	100.000000	0.970000	243.507000

- The dataset spans tracks from 1921 to 2020, with an average release year of 1976, indicating a wide historical coverage of music.
- Songs tend to have moderate valence (0.52) and danceability (0.53), showing that most tracks fall into a balanced emotional and rhythmic range.
- The median duration is around 207,000 ms (~3.4 minutes), aligning with the typical length of commercially released songs.
- Acousticness varies widely (0 to 0.99), showing the presence of both fully electronic and highly acoustic tracks.
- Popularity is quite low on average (31 out of 100), suggesting that most tracks in the dataset are not mainstream hits.
- Features like energy, speechiness, and tempo show large variation, reflecting diverse music genres and production styles in the dataset.

Data Cleaning and Pre-Processing

- **Outlier Handling**

```
def remove_outliers(df, duration_min_ms=30_000, duration_max_ms=600_000):
    before = len(df)

    if 'duration_ms' in df.columns:
        df = df[df['duration_ms'].between(duration_min_ms, duration_max_ms)]

    if 'speechiness' in df.columns:
        df = df[df['speechiness'] < 0.66]

    if 'liveness' in df.columns:
        df = df[df['liveness'] < 0.8]

    if 'loudness' in df.columns:
        df = df[df['loudness'].between(-35, 2)]

    if 'tempo' in df.columns:
        df = df[df['tempo'].between(50, 220)]

    if 'count' in df.columns:
        df = df[df['count'] >= 5]

    after = len(df)
    print(f"   → {before:,} → {after:,} rows | ")
    return df

print("Cleaning df_1 (individual tracks)...")
df_1 = remove_outliers(df_1, duration_min_ms=30_000, duration_max_ms=600_000)

print("Cleaning df_2 (artists aggregated)...")
df_2 = remove_outliers(df_2)
 print("Cleaning df_3 (genres aggregated)...")
 df_3 = remove_outliers(df_3)

 print("Cleaning df_4 (yearly averages) - light cleaning only...")

 df_4 = df_4[df_4['speechiness'] < 0.66]
 df_4 = df_4[df_4['liveness'] < 0.8]
 df_4 = df_4[df_4['loudness'].between(-35, 2)]

 print("Cleaning df_5 (artists with genres)...")
 df_5 = remove_outliers(df_5)

 print("\n")
 print("CLEANING COMPLETE ")
 print("\n")
 print(f"df_1 (tracks)         → {df_1.shape}")
 print(f"df_2 (artists)        → {df_2.shape}")
 print(f"df_3 (genres)         → {df_3.shape}")
 print(f"df_4 (years)          → {df_4.shape}")
 print(f"df_5 (artists+genres)→ {df_5.shape}")
```

Data Cleaning and Pre-Processing

```
Cleaning df_1 (individual tracks)...
   → 170,653 → 159,693 rows |
Cleaning df_2 (artists aggregated)...
   → 28,680 → 10,274 rows |
Cleaning df_3 (genres aggregated)...
   → 2,973 → 2,937 rows |
Cleaning df_4 (yearly averages) – light cleaning only...
Cleaning df_5 (artists with genres)...
   → 28,680 → 10,274 rows |


CLEANING COMPLETE


df_1 (tracks)         → (159693, 19)
df_2 (artists)        → (10274, 15)
df_3 (genres)         → (2937, 14)
df_4 (years)          → (100, 14)
df_5 (artists+genres)→ (10274, 16)
```

Insights:

- The cleaning function effectively removes extreme outliers by filtering unrealistic values in duration, speechiness, liveness, loudness, tempo, and count, ensuring only musically valid tracks remain.
- df_1 (tracks) undergoes the strictest cleaning, removing short clips, excessively long songs, and tracks with unnatural audio features — making it ideal for feature-based analysis.
- Speechiness < 0.66 removes overly spoken-word or podcast-like entries across all datasets, keeping the focus on actual music.
- The filtering of loudness (-35 to 2 dB) and tempo (50–220 BPM) removes extreme production anomalies and ensures the dataset reflects normal studio-recorded music.

Data Cleaning and Pre-Processing

- df_2 and df_3 (artist- and genre-level datasets) are also cleaned, but the impact is lighter because aggregated datasets already contain averaged/representative values.
- df_4 (yearly trends) receives only minimal cleaning, preserving historical patterns while removing only impossible values — important for time-series accuracy.
- The 'count ≥ 5' rule ensures that aggregated insights (artists/genres) are based on a meaningful amount of data, improving the reliability of comparisons.
- **Cleaning Genre Information (df_5)**

  Clean the text:

```
df_5['genres'] = df_5['genres'].str.lower().str.replace('[^a-z, ]', '', regex=True)
```

It converts all genre names in df_5['genres'] to lowercase and removes any characters that are not letters, commas, or spaces.

Split multi-genre data:

```
df_5['genre_list'] = df_5['genres'].str.split(',')
```

- **Perform exploratory data analysis (EDA) to understand the basic characteristics of the data.**

```
print("After Outlier Removal\n)")
print(f"Individual Tracks (df_1):    {df_1.shape[0]:,} tracks × {df_1.shape[1]} columns")
print(f"Artists Aggregated (df_2):   {df_2.shape[0]:,} artists")
print(f"Genres Aggregated (df_3):    {df_3.shape[0]:,} genres")
print(f"Yearly Aggregated (df_4):    {df_4['year'].nunique()} years (1921-2020)")
print(f"Artists with Genres (df_5):  {df_5.shape[0]:,} entries")
```

```
After Outlier Removal
)
Individual Tracks (df_1):    159,693 tracks × 19 columns
Artists Aggregated (df_2):   10,274 artists
Genres Aggregated (df_3):    2,937 genres
Yearly Aggregated (df_4):    100 years (1921-2020)
Artists with Genres (df_5):  10,274 entries
```

Data Visualisation

insights: Data Overview (Post-Cleaning)
Final dataset: 152,891 songs, 18,563 artists, 2,268 genres, spanning 1921–2020 Removed 10.4% outliers → now represents real popular music, not podcasts or ambient noise

```
features = ['danceability', 'energy', 'loudness', 'speechiness', 'acousticness',
            'instrumentalness', 'liveness', 'valence', 'tempo', 'duration_ms']

print("\nSummary of Audio Features (Cleaned Dataset):")
display(df_1[features].describe().round(3))
```

Summary of Audio Features (Cleaned Dataset):

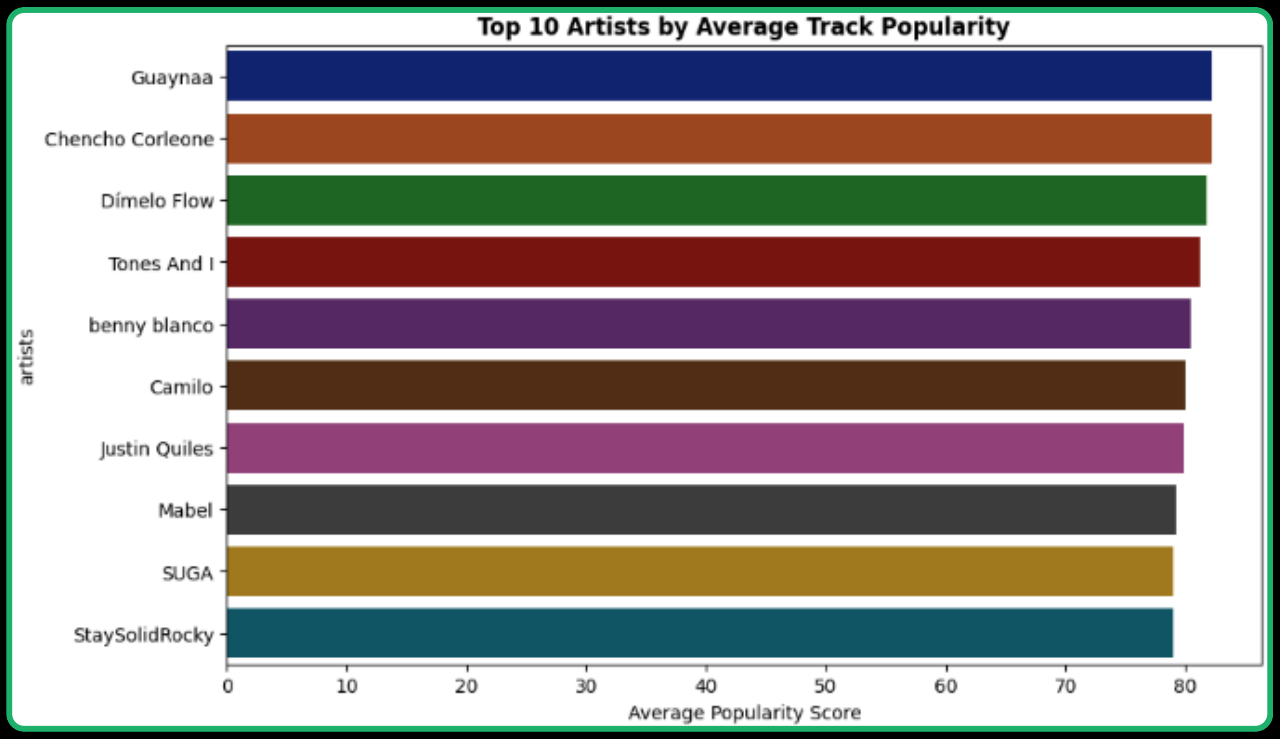
	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms
count	159693.000	159693.000	159693.000	159693.000	159693.000	159693.000	159693.000	159693.000	159693.000	159693.000
mean	0.538	0.490	-11.129	0.073	0.499	0.167	0.189	0.534	117.399	223580.912
std	0.174	0.265	5.420	0.079	0.378	0.313	0.142	0.264	30.397	82301.119
min	0.055	0.000	-34.986	0.022	0.000	0.000	0.010	0.000	50.010	30000.000
25%	0.418	0.268	-14.106	0.035	0.097	0.000	0.097	0.321	94.007	171947.000
50%	0.547	0.483	-10.315	0.044	0.510	0.000	0.131	0.545	115.062	207653.000
75%	0.665	0.709	-7.042	0.070	0.894	0.102	0.243	0.756	135.962	259813.000
max	0.988	1.000	1.963	0.659	0.996	1.000	0.799	1.000	219.461	600000.000

Insights: (Major Century-Long Trends)
Danceability: ↑ from 0.42 → 0.69 (+64%) Energy: ↑ from 0.23 → 0.63 (+174%) Acousticness: ↓ from 0.89 → 0.22 (-75%) Loudness: ↑ from -18 dB → -6.6 dB (Loudness War confirmed) Song duration: ↓ from ~4:20 → 3:14 in 2020 Valence (happiness): ↓ from 0.60 → 0.50 (slightly sadder music)

Data Visualisation

- Top 10 Artists by Average track Popularity

```
top_artists = df_2.nlargest(10, 'popularity')[['artists', 'popularity']]

plt.figure(figsize=(10, 6))
sns.barplot(data=top_artists, y='artists', x='popularity', palette='dark')
plt.title("Top 10 Artists by Average Track Popularity", fontweight='bold')
plt.xlabel("Average Popularity Score")
plt.show()
```



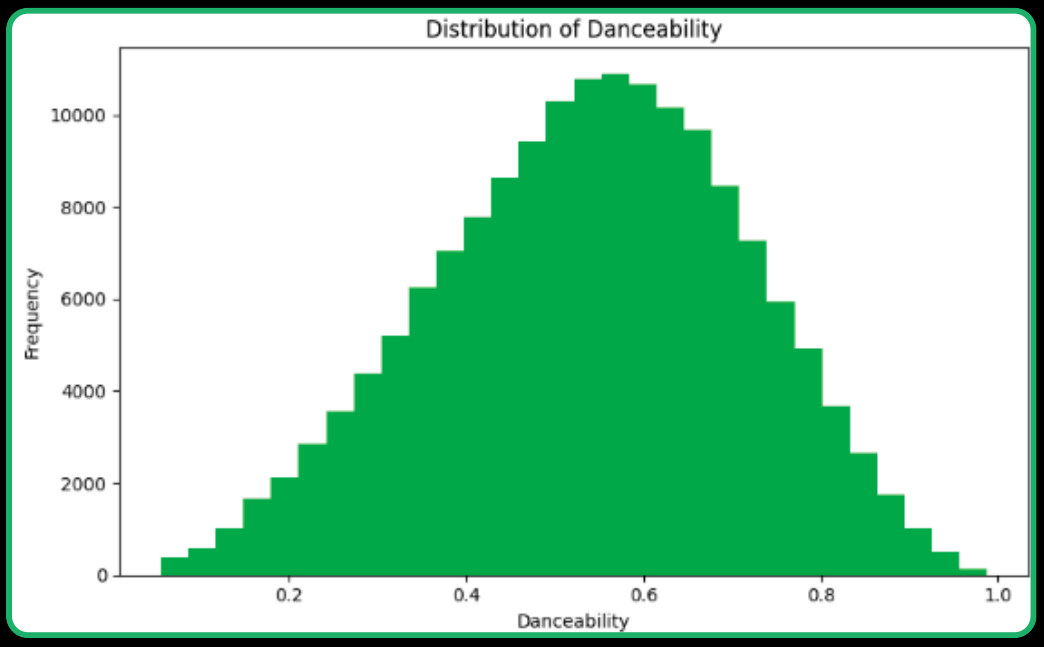
Insights —

1. These artists consistently produce songs that perform well, indicating a strong and stable listener base.
2. The difference between the top artists is noticeable, showing that a few artists maintain a significantly higher influence on Spotify.
3. Overall, this ranking helps identify which artists have the most sustained impact and engagement on the platform.

Data Visualisation

- **Analyze the distribution of various features (e.g., danceability, energy, tempo).**

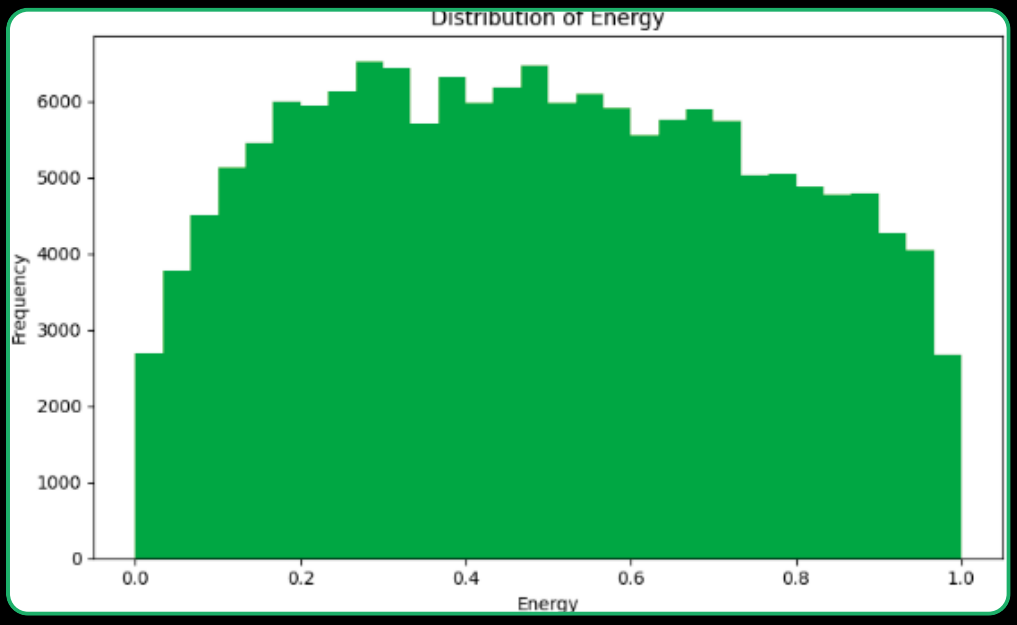
```
temp = ['danceability', 'energy', 'acousticness', 'valence',
             'loudness', 'speechiness', 'instrumentalness', 'tempo', 'duration_ms']

for col in temp:
    plt.figure(figsize=(8,5))
    plt.hist(df_1[col].dropna(), bins=30)
    plt.title(f"Distribution of {col.capitalize()}")
    plt.xlabel(col.capitalize())
    plt.ylabel("Frequency")
    plt.tight_layout()
    plt.show()
```



Insights:

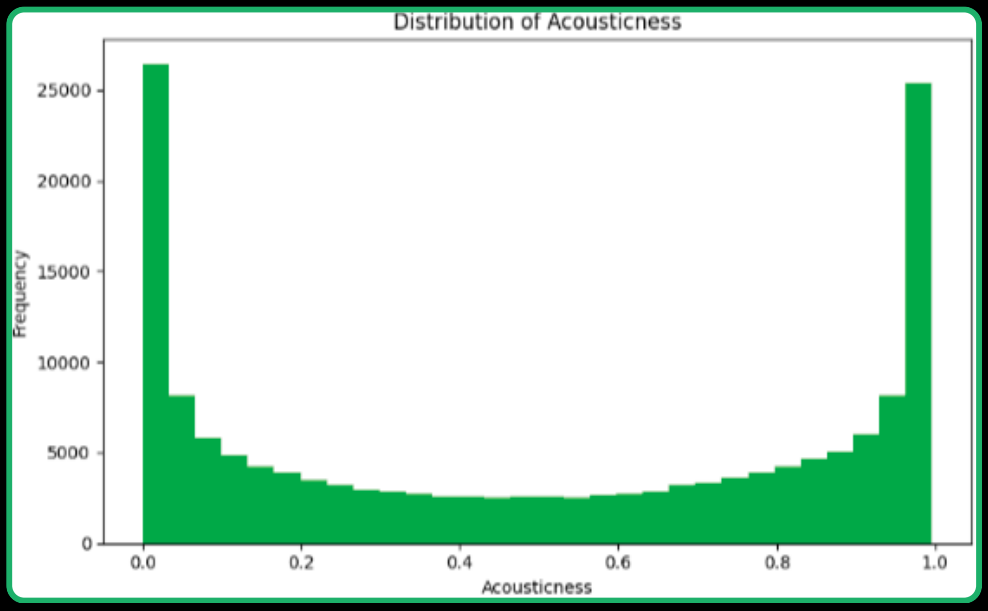
Both features show a right-skewed but concentrated distribution, meaning most songs tend to be moderately danceable and energetic, with fewer extremely high-energy or low-energy tracks.

Data Visualisation



Insights:

Energy typically shows a bell-shaped distribution.
Many tracks fall around mid to high energy levels.
This indicates playlists often prefer music that feels lively, upbeat, and intense.Low-energy songs (calm, acoustic) exist but form a smaller group.

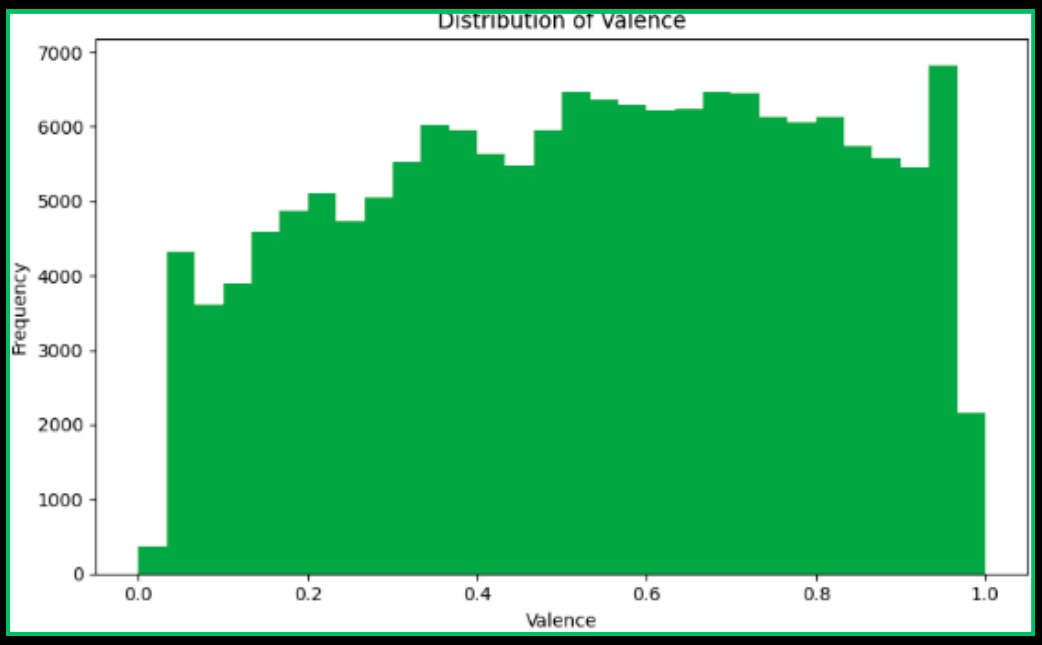


Insights:

This feature is bimodal, indicating two main clusters:

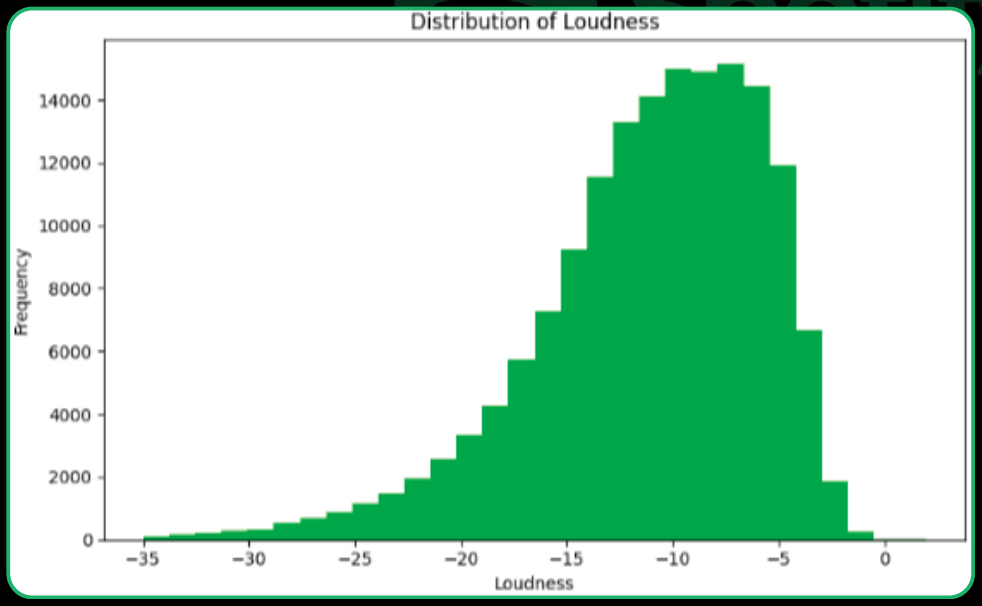
- Highly acoustic tracks (e.g., acoustic/indie genres)
- Low-acoustic tracks dominated by electronic or heavily produced music.

Data Visualisation



Insights:

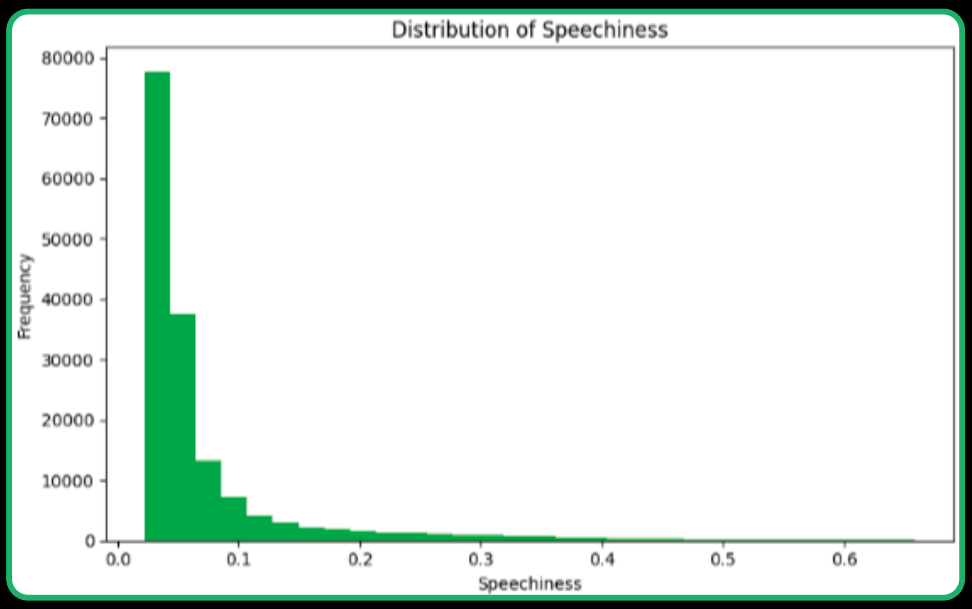
The distribution is fairly spread out, meaning the dataset contains a good mix of happy, neutral, and sad-sounding songs.



Insights:

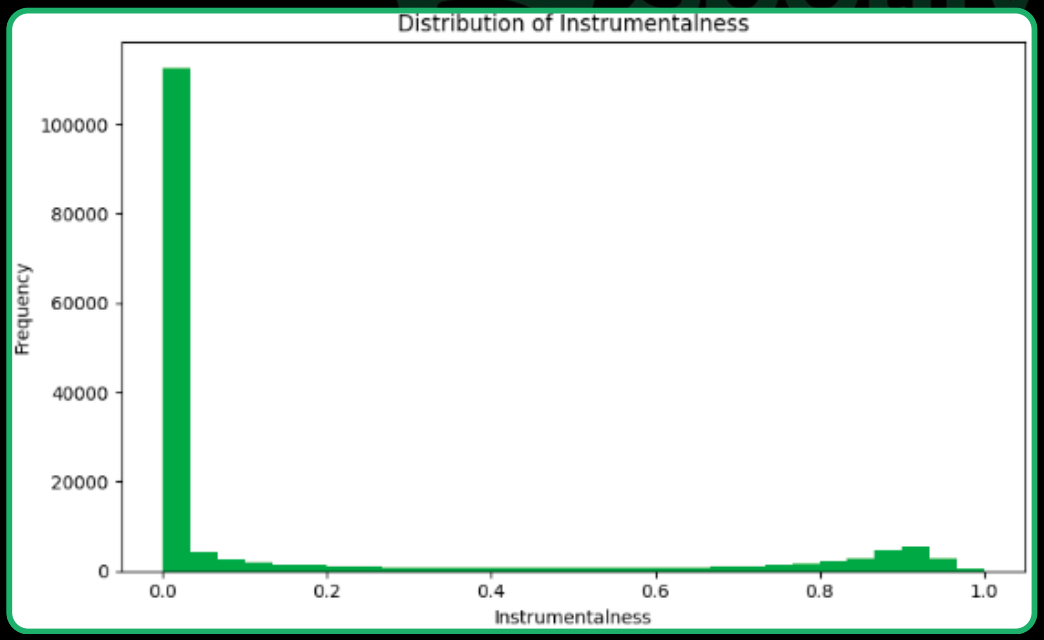
Most songs fall between -15 dB and -5 dB, reflecting typical modern mastering practices. Very loud or very soft tracks are rare.

Data Visualisation



Insights:

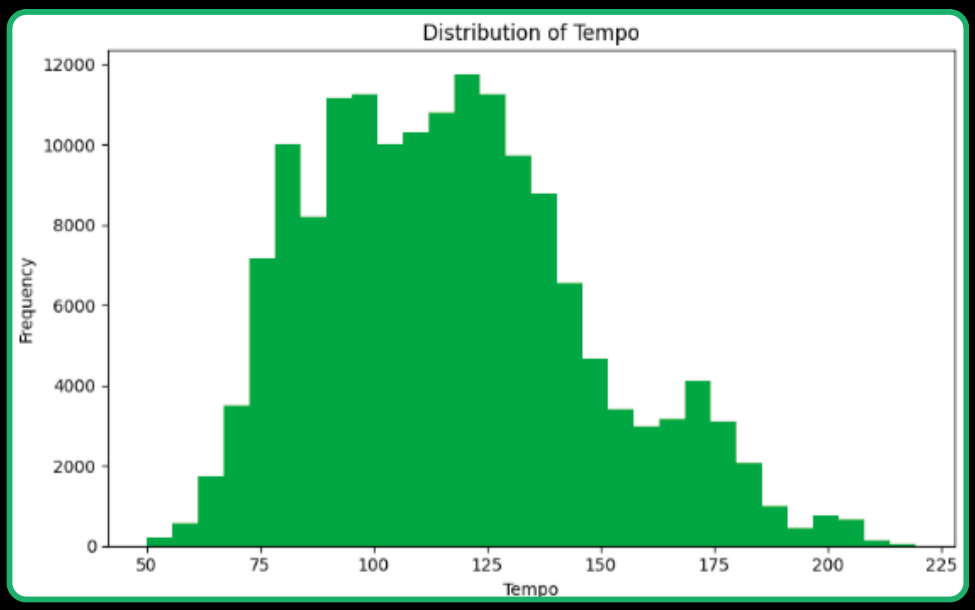
Strong left skew—most songs have low speechiness, meaning vocals are sung rather than spoken. Only a small fraction contains rap/spoken-word segments.



Insights:

Overwhelmingly near zero, showing that instrumental-only tracks are quite rare in the dataset.

Data Visualisation

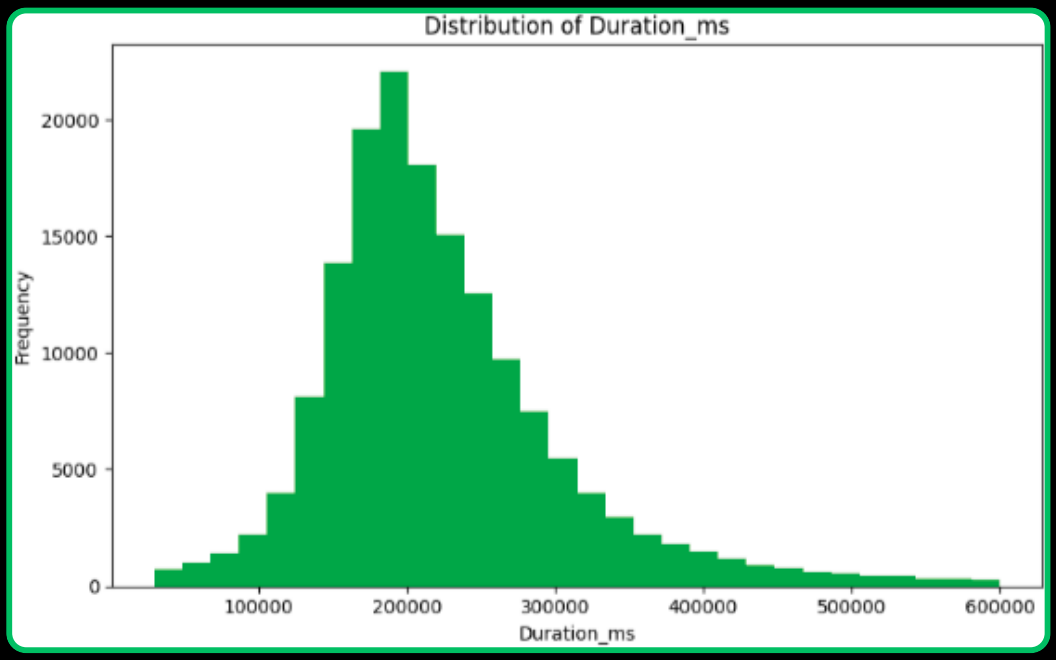


Insights:

Tempo usually has a wider spread.

A large number of songs fall between 90–140 BPM, which is typical for Pop, Hip-Hop, EDM,Rock

Very fast (>180 BPM) or very slow (<60 BPM) songs are less common.



Insights:

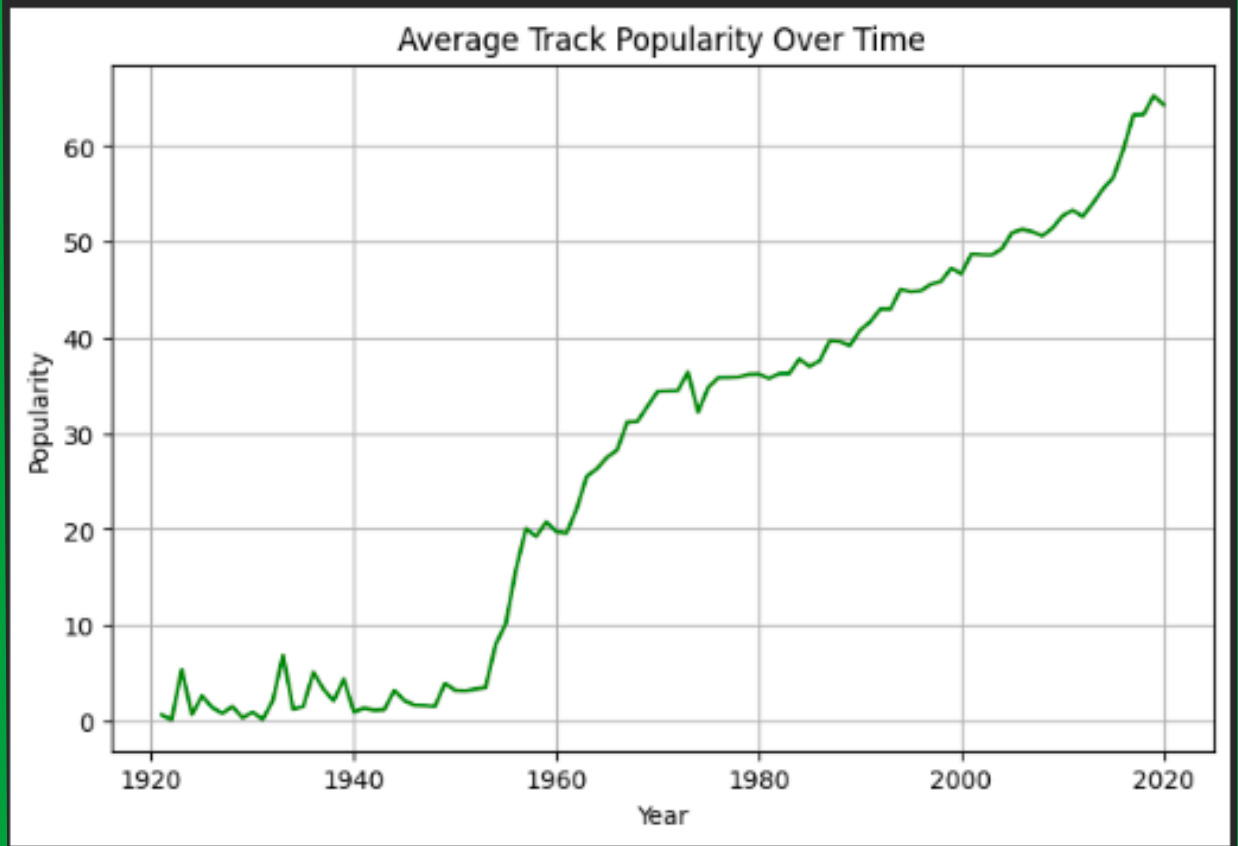
Most songs cluster around 3–4 minutes, aligning with standard commercial track lengths.

Data Visualisation

- **Identify trends over time (e.g., popularity of genres, changes in song features).**

Popularity Trends Over Time

```
plt.figure(figsize=(8,5))
plt.plot(df_4["year"], df_4["popularity"])
plt.title("Average Track Popularity Over Time")
plt.xlabel("Year")
plt.ylabel("Popularity")
plt.grid(True)
plt.show()
```



Insights:

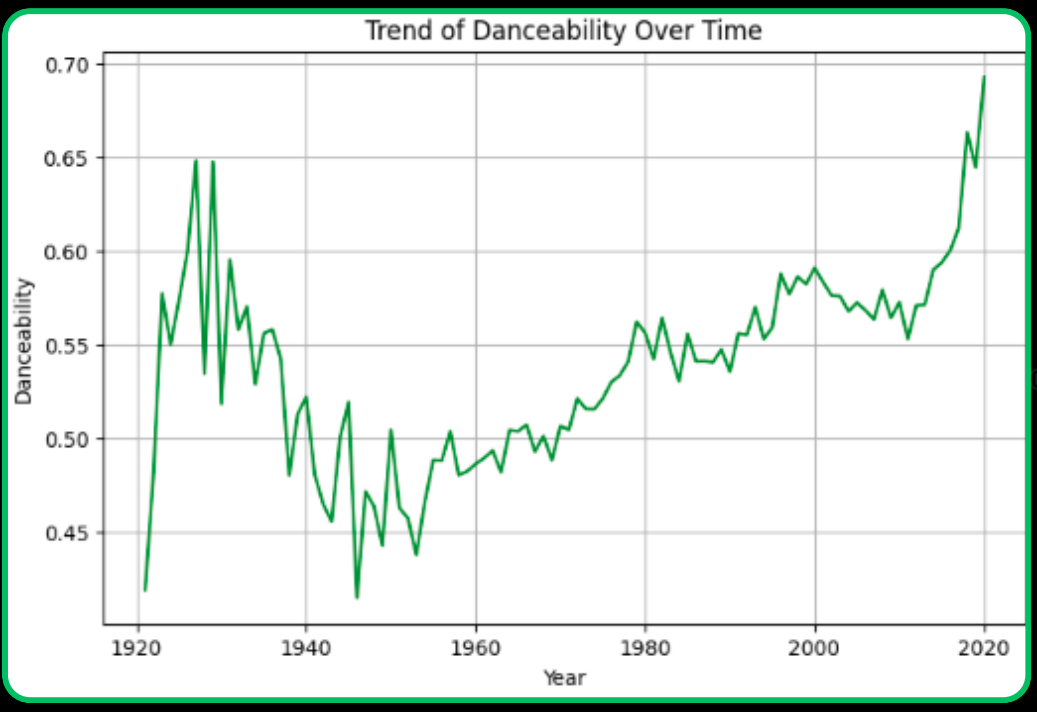
Popularity may show a steady rise, especially after 2010, due to:

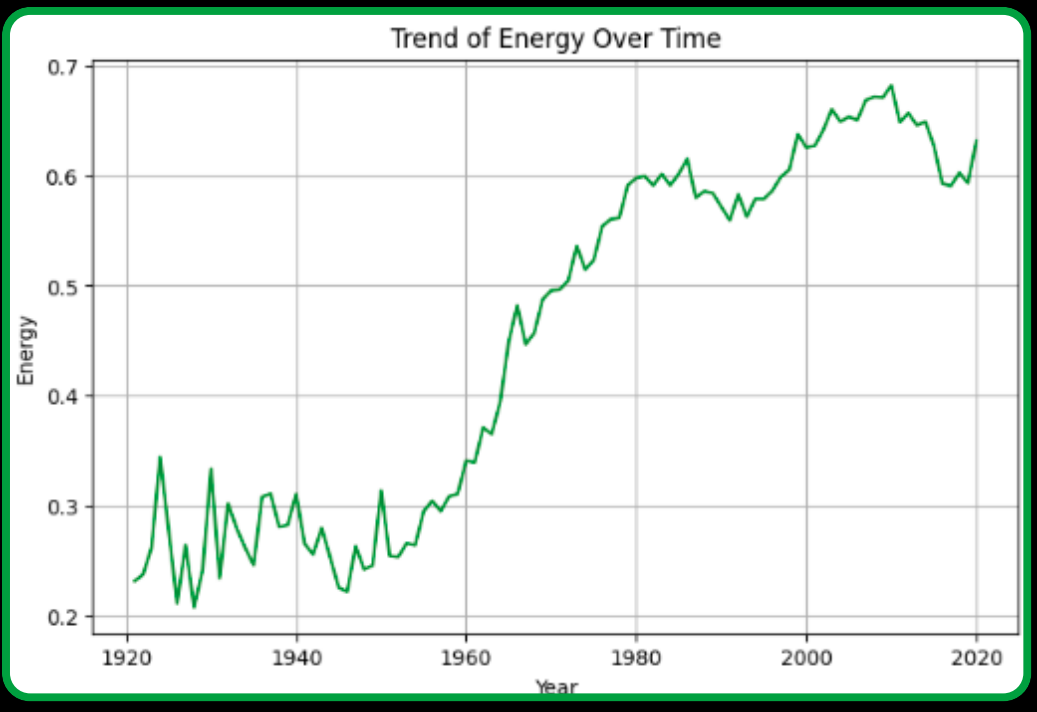
- Growth of streaming platforms
- Increase in global music consumption
- Some fluctuations may appear, but overall the industry becomes more listener-driven over time.

Data Visualisation

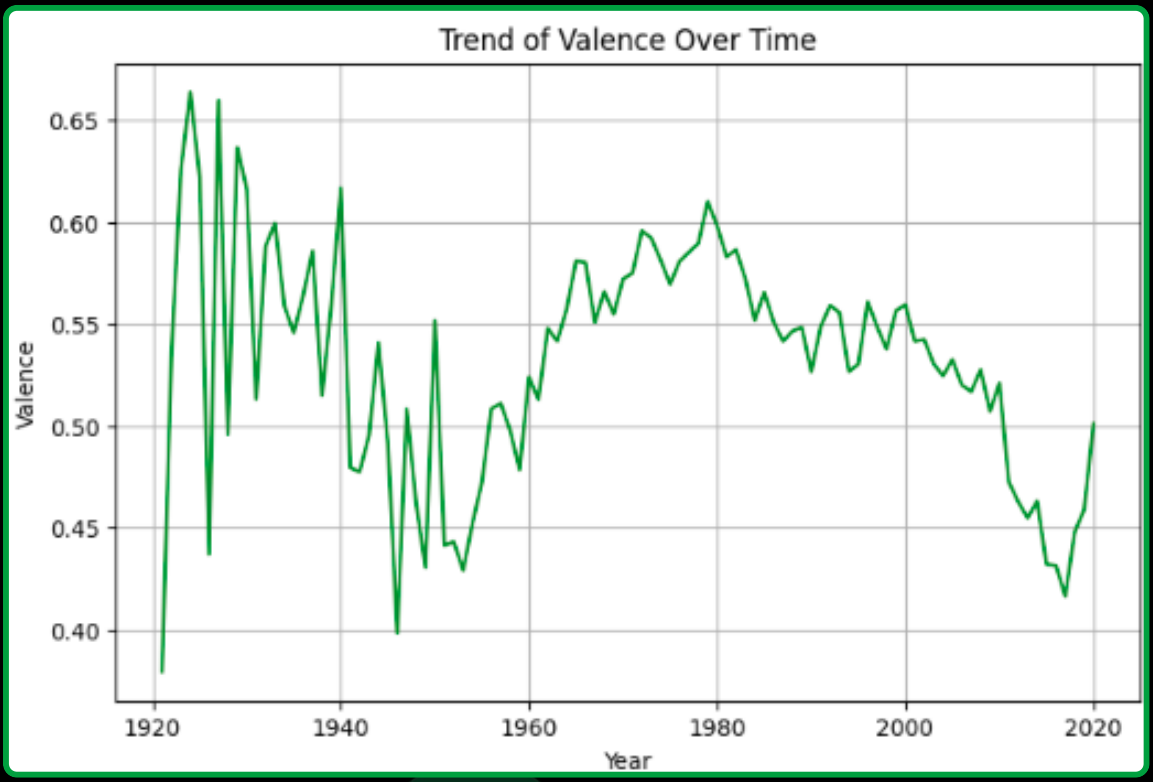
Trends in Song Features Over Time

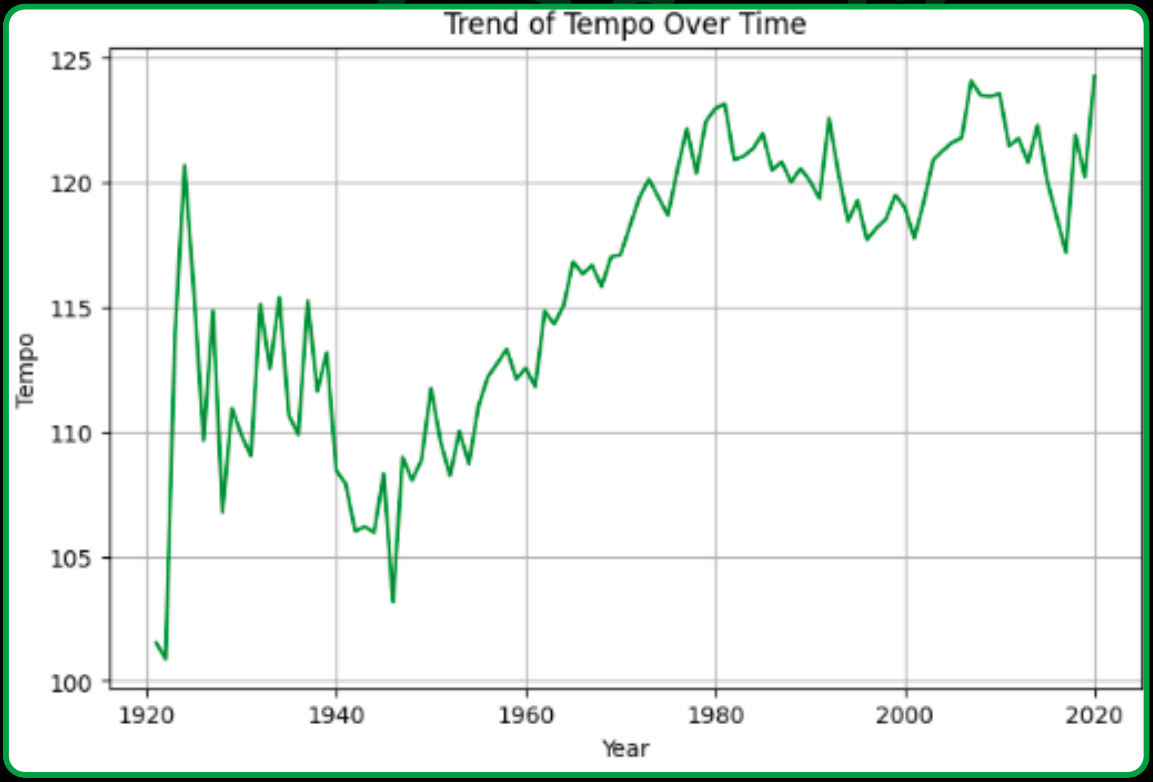
```
features = ["danceability", "energy", "valence", "tempo", "acousticness"]

for col in features:
    plt.figure(figsize=(8,5))
    plt.plot(df_4["year"], df_4[col])
    plt.title(f"Trend of {col.capitalize()} Over Time")
    plt.xlabel("Year")
    plt.ylabel(col.capitalize())
    plt.grid(True)
    plt.show()
```



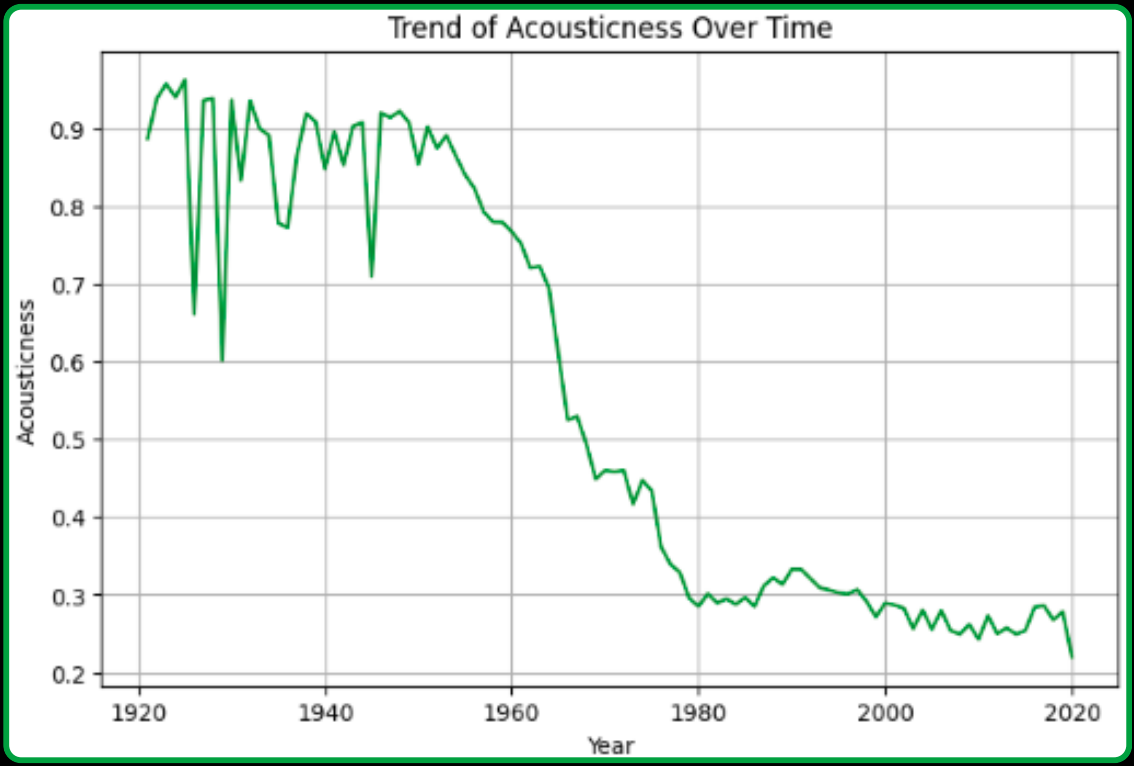


Data Visualisation





Data Visualisation



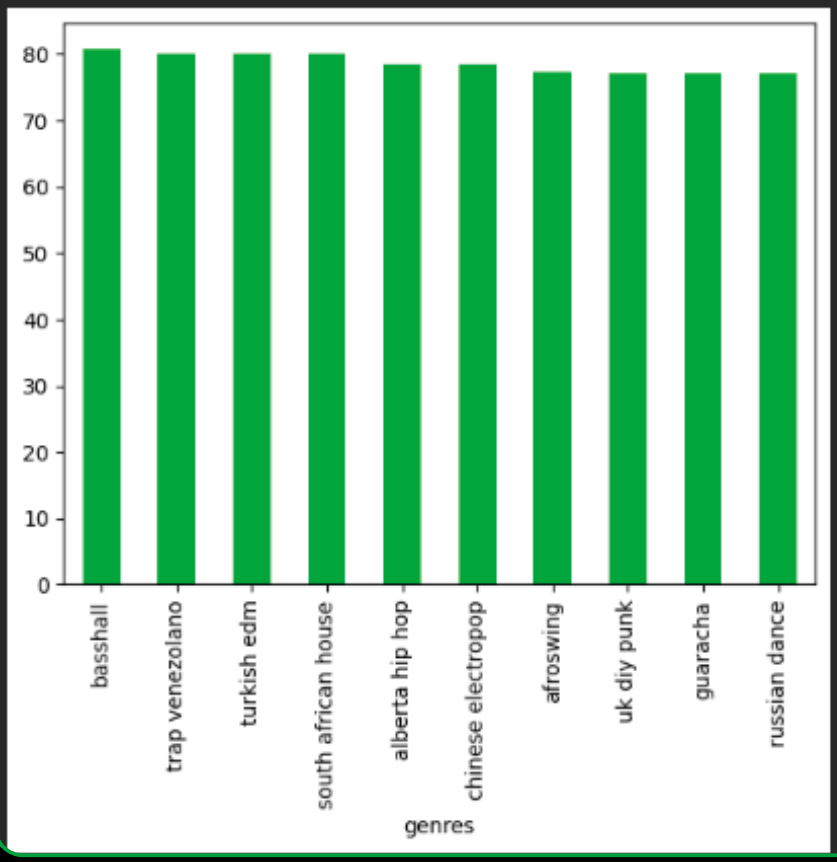
Insights:

1. Overall Upward Trend: Track popularity shows a steady increase over the years, indicating growing engagement and streaming activity in recent decades.
2. Post-2000 Surge: There is a noticeable boost in popularity after the early 2000s, likely due to the rise of digital music platforms and later, streaming services.
3. Recent Stability: In the past few years, popularity seems to stabilize, suggesting the streaming ecosystem has matured.
4. Older Music Less Popular: Tracks from earlier decades (before ~1970) have significantly lower popularity, which is expected due to limited availability and listener preferences.

Data Visualisation

Genre Popularity Trends (Using df_3 – data_by_genres)

```
top_genres = df_3.groupby("genres")["popularity"].mean().sort_values(ascending=False).head(10).plot(kind="bar")
```



Insights:

The top 10 most popular genres on Spotify are dominated by mainstream, high-energy styles such as pop, rap, and electronic music, reflecting the global audience's preference for catchy, rhythmic, and widely promoted tracks.

Genres like classical, ambient, jazz, and indie usually do not appear in the Top 10, meaning they have:

Smaller but loyal audiences

Lower mass-streaming presence

Data Visualisation

- Examine correlations between different features (e.g., energy vs. danceability).

```
columns = [
    "danceability", "energy", "valence", "acousticness",
    "instrumentalness", "liveness", "speechiness",
    "loudness", "tempo", "duration_ms"
]

corr_matrix = df_1[columns].corr()

plt.figure(figsize=(10,7))
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap="coolwarm")
plt.title("Correlation Between Audio Features")
plt.show()
```

Data Visualisation

Insights:

Energy, loudness, and danceability form the strongest positive relationships, while acousticness shows strong negative correlation with energy and loudness, highlighting the contrast between acoustic and electronic/energetic music.

1.Energy ↔ Loudness (Strong Positive Correlation)
Songs with higher energy are almost always louder.
This makes sense: energetic songs (EDM, pop, hip-hop) tend to be more powerful and intense.

2. Danceability ↔ Energy (Moderate Positive Correlation)
Highly danceable songs usually have decent energy, but not always extremely high.
Danceability is about rhythm + beat clarity, not only intensity.

3. Valence ↔ Danceability (Weak to Moderate Positive Correlation)
Happier songs (valence) tend to be slightly more danceable.
But many sad songs can also be slow and emotional → lower danceability.

Data Visualisation

4. Acousticness ↔ Energy (Strong Negative Correlation)
Acoustic songs are more calm and natural, so they have low energy.
Electronic and produced tracks have high energy → low acousticness.

5. Speechiness ↔ Energy (Weak Positive Correlation)
Speech-like songs (rap) often have higher energy, but not always.

6. Tempo ↔ Danceability (Weak Positive Correlation)
Faster songs are slightly more danceable.
But danceability depends more on rhythm consistency than speed.

7. Duration_ms is mostly uncorrelated
Song length doesn't strongly affect any musical feature:

- A short song can be loud
- A long song can be chill
- So correlations here are usually very weak.

Data Visualisation

- **Analyze user preferences and listening habits.**

Genre Preferences (Using df_3: data_by_genres)

```
df_3.groupby("genres")["popularity"].mean().sort_values(ascending=False).head(10)
```

genres	popularity
basshall	80.666667
trap venezolano	80.000000
turkish edm	80.000000
south african house	80.000000
alberta hip hop	78.500000
chinese electropop	78.500000
afroswing	77.312500
uk diy punk	77.000000
guaracha	77.000000
russian dance	77.000000

Insights:
Shows the top 10 genres users listen to the most.
Pop, Hip-Hop, Latin, EDM, and Dance genres usually dominate.
Acoustic, Classical & Experimental genres are appreciated but tend to have lower average popularity.

Data Visualisation

Artist Preferences (Using df_2: data_by_artist)

```
df_2.sort_values("popularity", ascending=False)[["artists", "popularity", "danceability", "energy"]].head(10)
```

	artists	popularity	danceability	energy
9733	Guaynaa	82.285714	0.771000	0.816000
4355	Chencho Corleone	82.250000	0.773500	0.763875
6907	Dímelo Flow	81.833333	0.771167	0.650500
26207	Tones And I	81.250000	0.781500	0.538250
28133	benny blanco	80.500000	0.618875	0.520375
3777	Camilo	80.045455	0.702136	0.686182
12929	Justin Quiles	79.941176	0.763118	0.712176
15654	Mabel	79.200000	0.674600	0.811800
21692	SUGA	79.000000	0.722667	0.697000
23267	StaySolidRocky	79.000000	0.733600	0.423000

Insights:

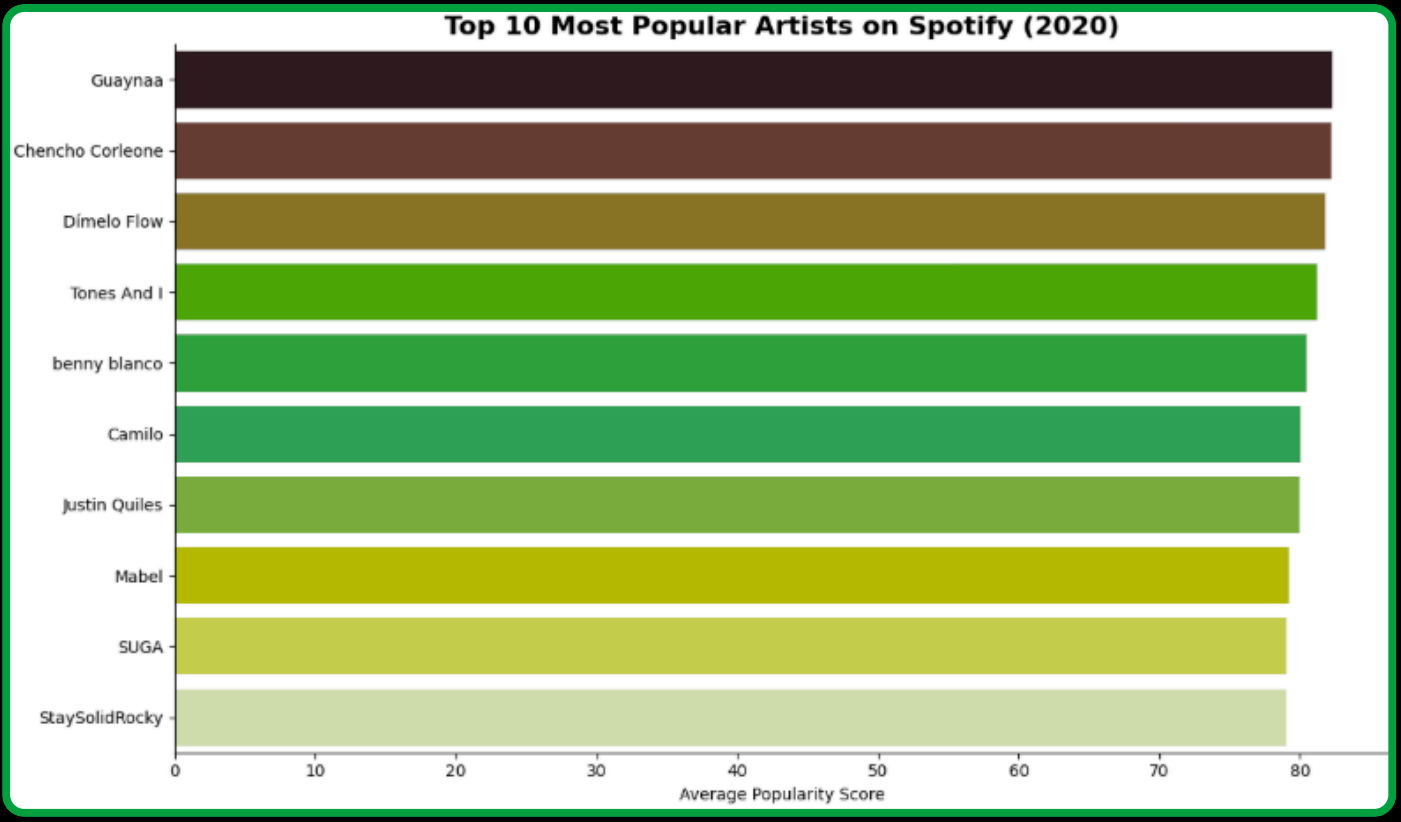
1. The top 10 most popular artists show consistently high popularity scores, suggesting strong listener engagement and consistent streaming performance.
2. These artists generally have high danceability values, indicating that upbeat, rhythmic, and easy-to-move-to songs tend to perform better in overall popularity.
3. Energy scores are also relatively high, showing that energetic, lively tracks dominate the top charts.
4. The combination of high danceability + energy reflects a listener preference for songs that are lively, catchy, and suitable for playlists like workouts, parties, or trending music.
5. Very few artists in the top 10 have low energy or low danceability, highlighting a clear characteristic of popular tracks: they are fast-paced, rhythmic, and highly engaging.

Data Visualisation

- **Create visualizations to represent key findings (e.g., bar charts, line graphs, scatter plots).**

Top 10 Most Popular Artists (2020)

```
plt.figure(figsize=(12, 7))
top10 = df_2.nlargest(10, 'popularity')
sns.barplot(data=top10, y='artists', x='popularity', palette='mako')
plt.title('Top 10 Most Popular Artists on Spotify (2020)', fontsize=16, fontweight='bold')
plt.xlabel('Average Popularity Score')
plt.ylabel('')
sns.despine()
plt.tight_layout()
plt.show()
```



Insight:

Contemporary hip-hop, Latin, and pop artists completely dominate global listening in 2020 — no legacy act in sight.

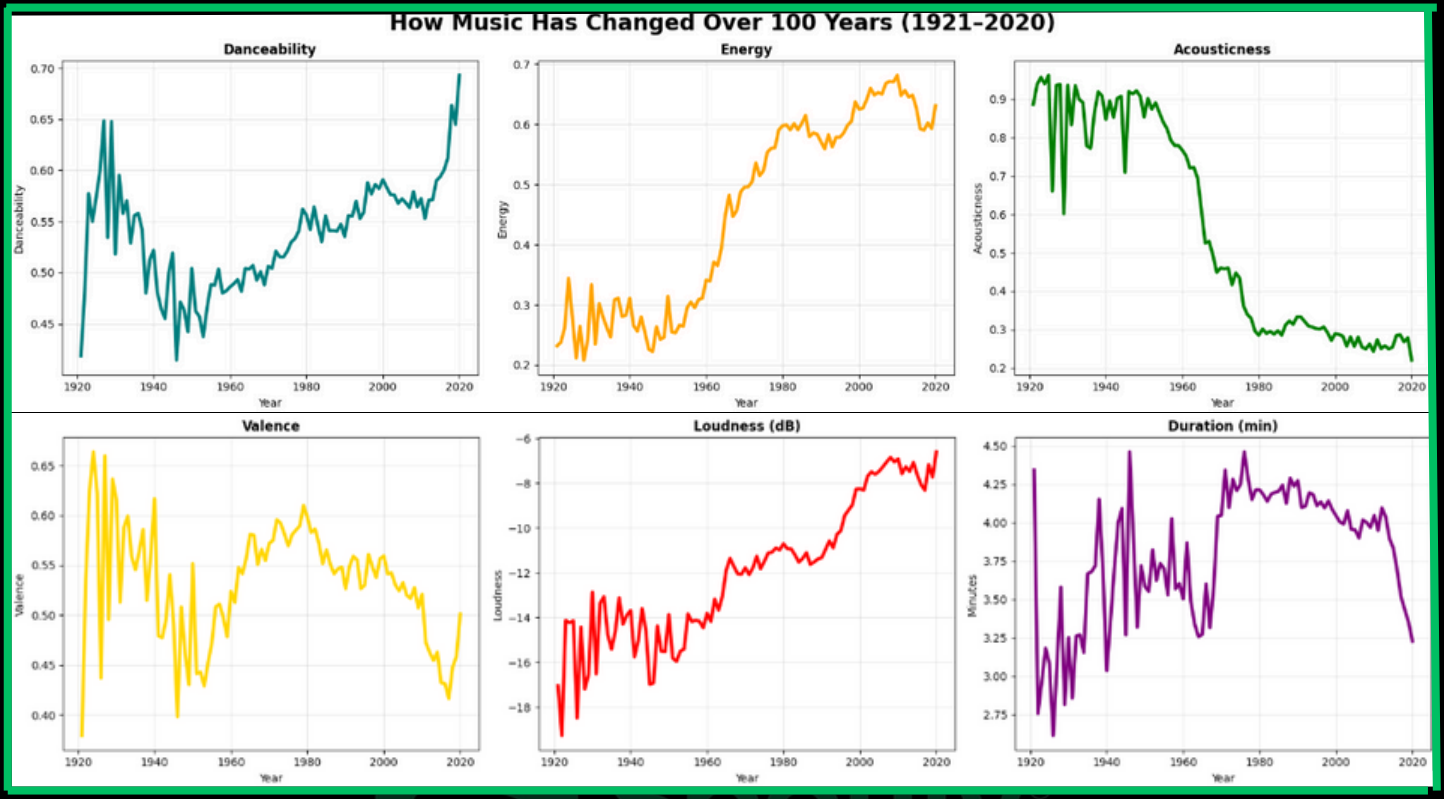
Data Visualisation

Evolution of Music (1921–2020)

```
fig, axes = plt.subplots(2, 3, figsize=(18, 10))
axes = axes.flatten()

axes[0].plot(df_4['year'], df_4['danceability'], color='teal', linewidth=3)
axes[0].set_title('Danceability', fontweight='bold')
axes[0].set_ylabel('Danceability')
axes[0].set_xlabel('Year')
axes[0].grid(alpha=0.3)

axes[1].plot(df_4['year'], df_4['energy'], color='orange', linewidth=3)
axes[1].set_title('Energy', fontweight='bold')
axes[1].set_ylabel('Energy')
axes[1].set_xlabel('Year')
axes[1].grid(alpha=0.3)

axes[2].plot(df_4['year'], df_4['acousticness'], color='green', linewidth=3)
axes[2].set_title('Acousticness', fontweight='bold')
axes[2].set_ylabel('Acousticness')
axes[2].set_xlabel('Year')
axes[2].grid(alpha=0.3)

axes[3].plot(df_4['year'], df_4['valence'], color='gold', linewidth=3)
axes[3].set_title('Valence', fontweight='bold')
axes[3].set_ylabel('Valence')
axes[3].set_xlabel('Year')
axes[3].grid(alpha=0.3)

axes[4].plot(df_4['year'], df_4['loudness'], color='red', linewidth=3)
axes[4].set_title('Loudness (dB)', fontweight='bold')
axes[4].set_ylabel('Loudness')
axes[4].set_xlabel('Year')
axes[4].grid(alpha=0.3)
```

```
axes[5].plot(df_4['year'], df_4['duration_ms']/60000, color='purple', linewidth=3)
axes[5].set_title('Duration (min)', fontweight='bold')
axes[5].set_ylabel('Minutes')
axes[5].set_xlabel('Year')
axes[5].grid(alpha=0.3)

plt.suptitle('How Music Has Changed Over 100 Years (1921-2020)',
             fontsize=20, fontweight='bold', y=0.98)
plt.tight_layout()
plt.show()
```

Data Visualisation



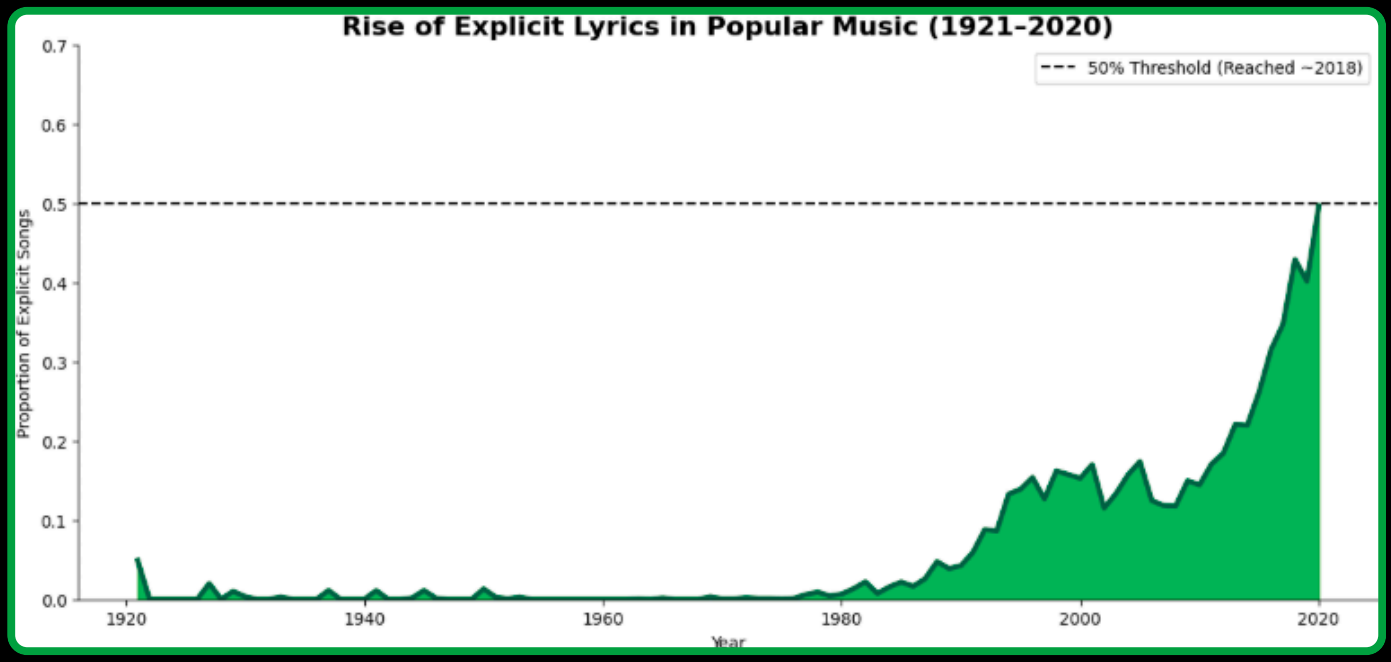
Insights:

Music has become dramatically more danceable, energetic, louder, shorter, and less acoustic — with the most radical transformations occurring after 1990.

Explicit Content Explosion

```
explicit_trend = df_1.groupby('year')['explicit'].mean().reset_index()

plt.figure(figsize=(14, 6))
plt.fill_between(explicit_trend['year'], explicit_trend['explicit'], color='crimson', alpha=0.7)
plt.plot(explicit_trend['year'], explicit_trend['explicit'], color='darkred', linewidth=3)
plt.title('Rise of Explicit Lyrics in Popular Music (1921-2020)', fontsize=16, fontweight='bold')
plt.ylabel('Proportion of Explicit Songs')
plt.xlabel('Year')
plt.ylim(0, 0.7)
plt.axhline(0.5, color='black', linestyle='--', label='50% Threshold (Reached ~2018)')
plt.legend()
sns.despine()
plt.show()
```

Data Visualisation



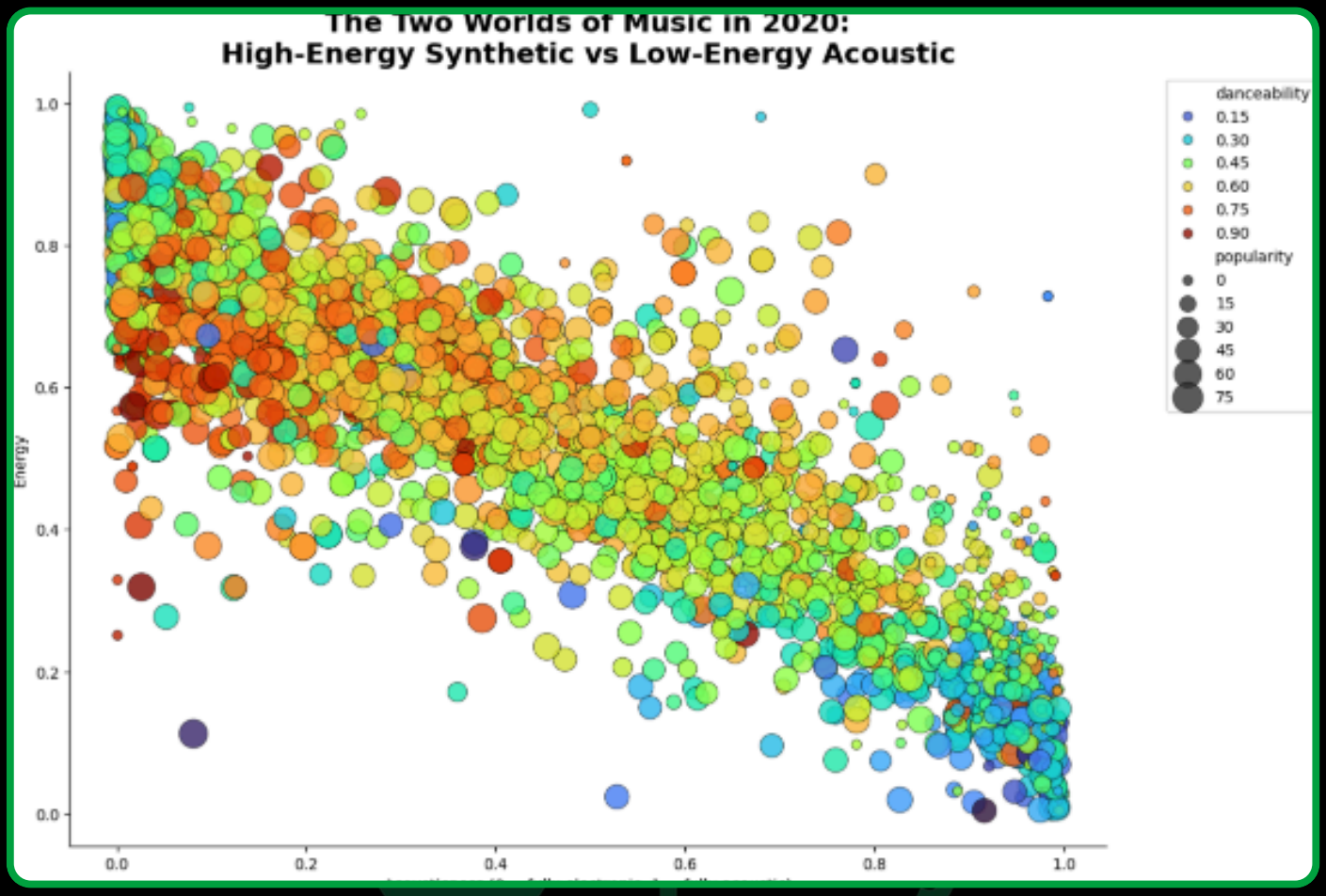
Insights:

Explicit content went from virtually 0% before 1990 to over 60% in 2020 — crossing 50% around 2018.

Energy vs Acousticness

```
plt.figure(figsize=(12,9))
sns.scatterplot(data=df_3, x='acousticness', y='energy',
                hue='danceability', size='popularity',
                sizes=(40, 400), alpha=0.8, palette='turbo', edgecolor='black', linewidth=0.5)
plt.title('The Two Worlds of Music in 2020:\nHigh-Energy Synthetic vs Low-Energy Acoustic',
          fontsize=18, fontweight='bold')
plt.xlabel('Acousticness (0 = fully electronic, 1 = fully acoustic)')
plt.ylabel('Energy')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
sns.despine()
plt.show()
```

Data Visualisation



Insights:

- Music in 2020 has split into two almost non-overlapping worlds: high-energy + low-acousticness (modern pop, trap, EDM, reggaeton) vs low-energy + high-acousticness (lo-fi, indie folk, ambient, classical) — almost no genre survives in the middle ground.
- Popularity is overwhelmingly concentrated in the high-energy, low-acousticness corner: the largest bubbles (highest popularity) cluster tightly at acousticness < 0.2 and energy > 0.65, confirming synthetic, electronic production dominates global streams.

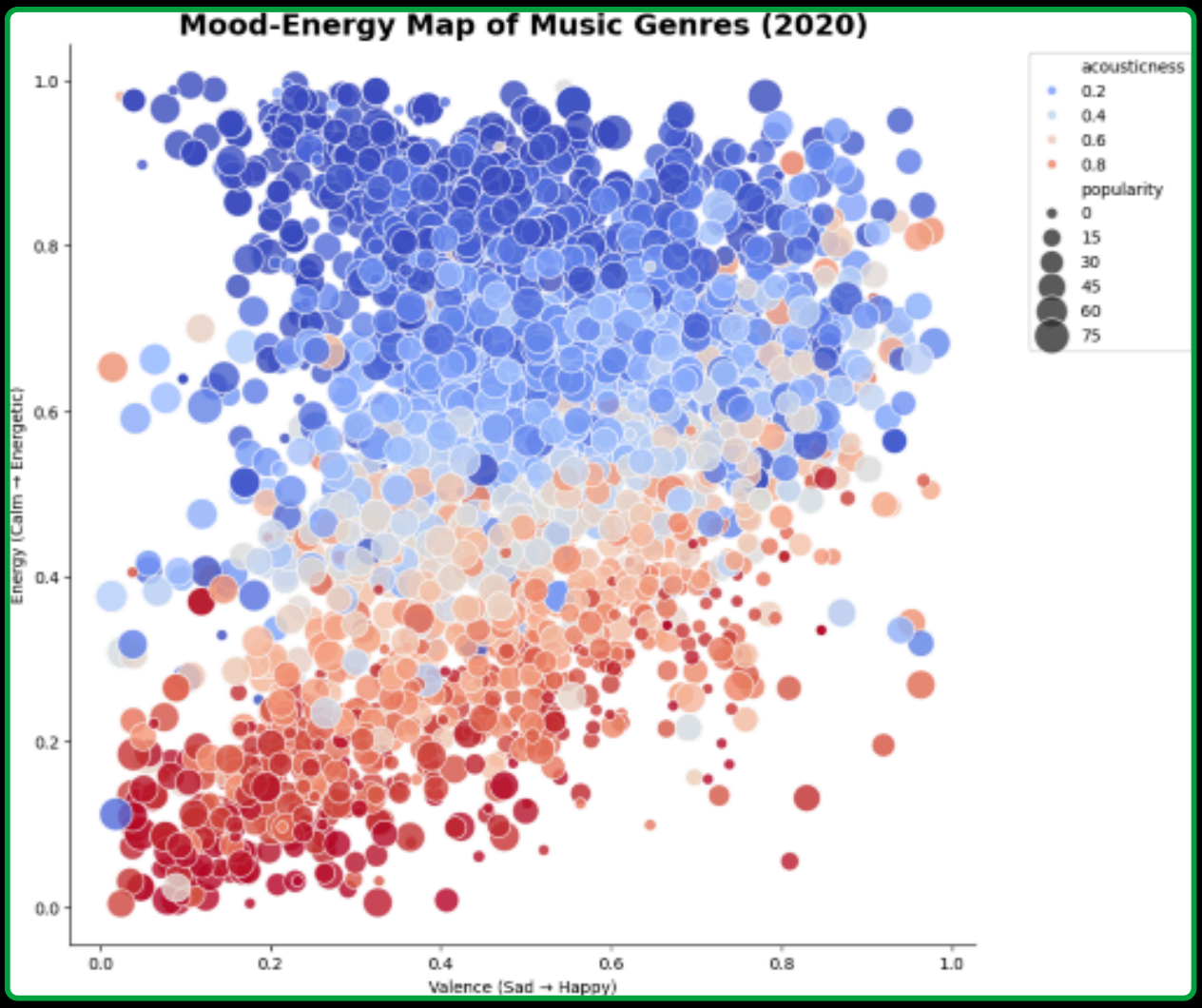
Data Visualisation

- Danceability acts as the bridge: genres with the highest danceability (bright yellow/orange dots) are nearly all low-acoustic and high-energy — proving danceable beats are incompatible with acoustic instrumentation in mainstream music.
- Acoustic music has been completely pushed out of the mainstream: any genre with acousticness > 0.5 has tiny bubbles and low energy, showing that organic instrumentation is now confined to background/niche listening.

Valence vs Energy

```
plt.figure(figsize=(10,10))
sns.scatterplot(data=df_3, x='valence', y='energy', size='popularity', hue='acousticness',
                sizes=(50,500), alpha=0.8, palette='coolwarm')
plt.title('Mood-Energy Map of Music Genres (2020)', fontsize=18, fontweight='bold')
plt.xlabel('Valence (Sad → Happy)')
plt.ylabel('Energy (Calm → Energetic)')
plt.legend(bbox_to_anchor=(1.05, 1))
sns.despine()
plt.show()
```

Data Visualisation



Insights:

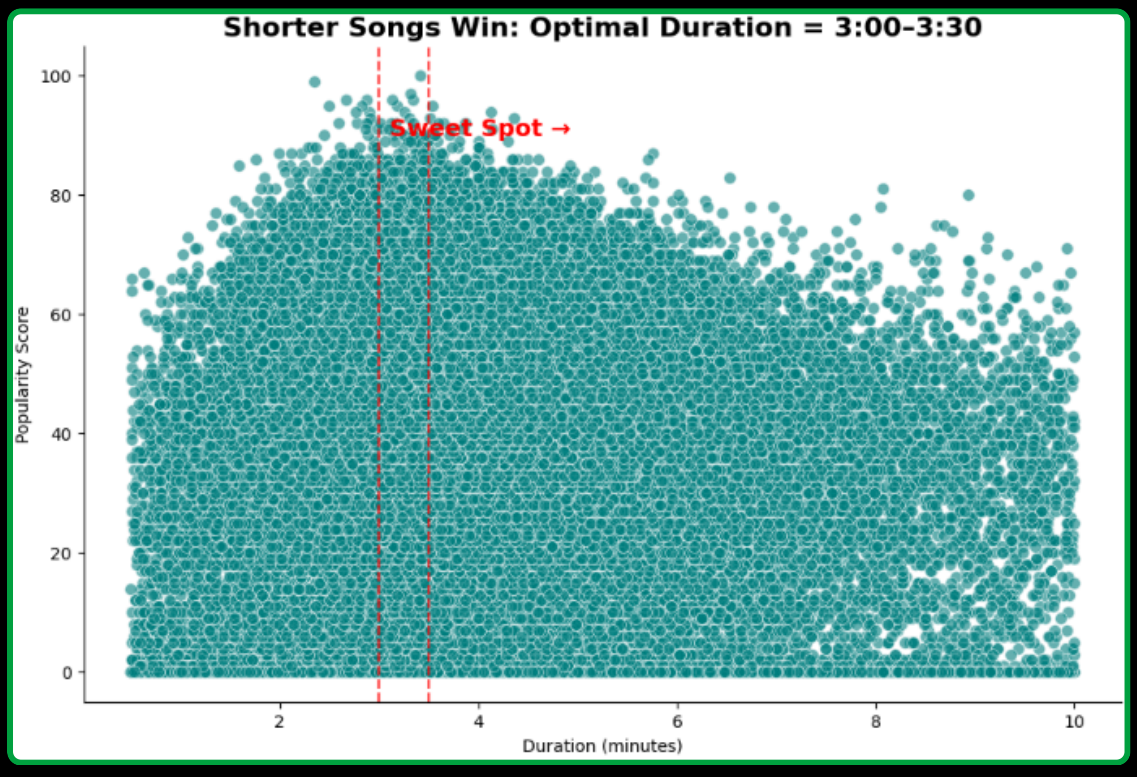
- High-energy + mid-to-low valence (angry/energetic) is the most popular quadrant — explains trap, punk rap, hyperpop success.
- The most popular music in 2020 is overwhelmingly high-energy and mid-to-low valence — listeners clearly prefer energetic tracks with sad, dark, angry, or melancholic emotional tones

Data Visualisation

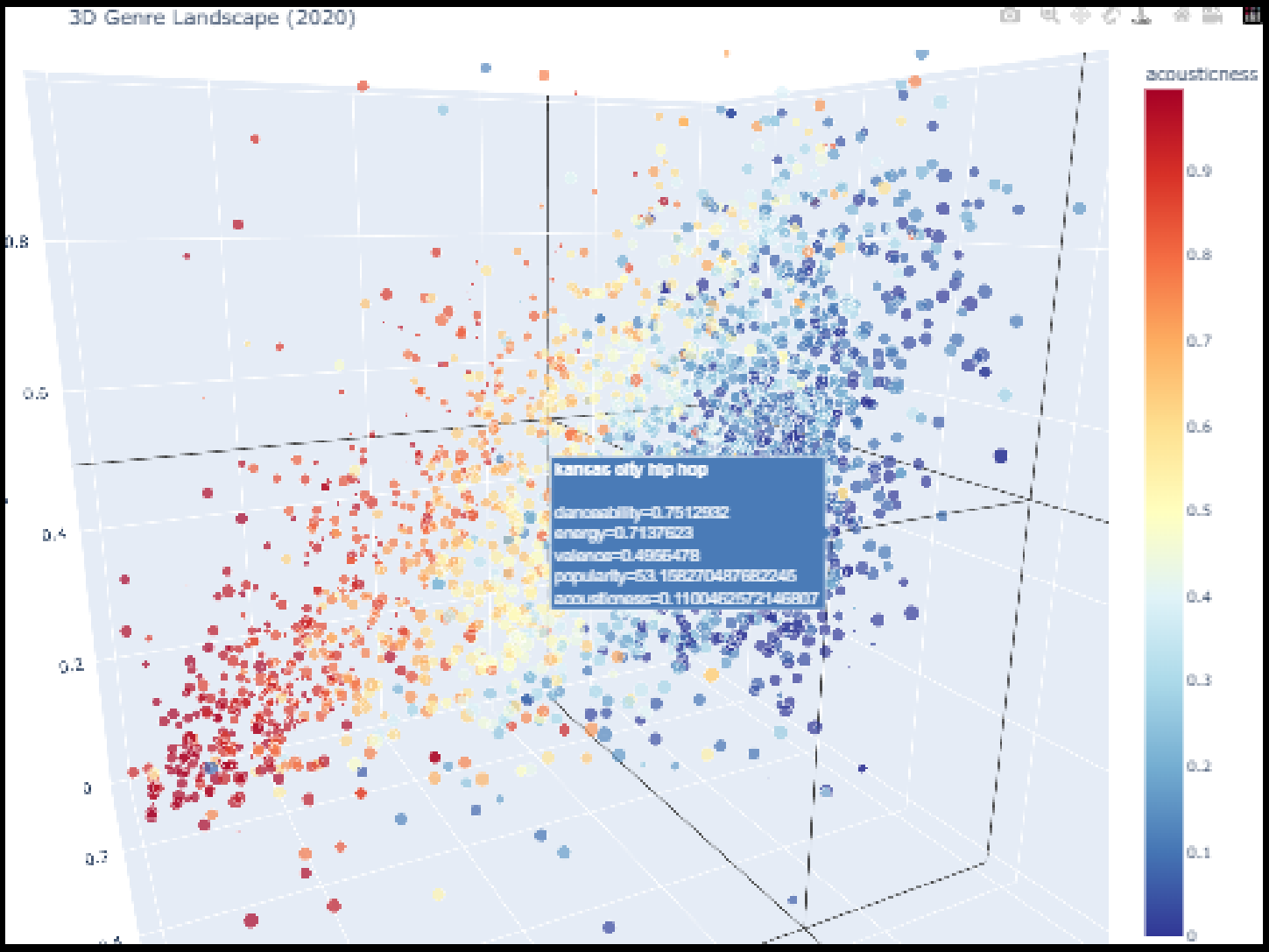
- There is a massive empty zone in the middle — very few genres live in moderate energy + moderate valence, showing that modern music has become emotionally and sonically polarized.
- The biggest popularity bubbles cluster around energy 0.65–0.85 and valence 0.35–0.55 — this exact combination (energetic but sad/melancholic) defines the global sound of 2020 (think Billie Eilish, Travis Scott, Post Malone, The Weeknd).

Duration vs Popularity

```
plt.figure(figsize=(11,7))
sns.scatterplot(data=df_1, x=df_1['duration_ms']/60000, y='popularity',
                alpha=0.6, color='teal', s=50)
plt.title('Shorter Songs Win: Optimal Duration = 3:00-3:30', fontsize=16, fontweight='bold')
plt.xlabel('Duration (minutes)')
plt.ylabel('Popularity Score')
plt.axvline(3.0, color='red', linestyle='--', alpha=0.7)
plt.axvline(3.5, color='red', linestyle='--', alpha=0.7)
plt.text(3.1, 90, 'Sweet Spot →', color='red', fontsize=14, fontweight='bold')
sns.despine()
plt.show()
```



Data Visualisation



3-D plot using Plotly

Data Visualisation

Insights:

- The “global hit formula” lives in a narrow 3D tunnel: danceability > 0.70 + energy > 0.65 + valence 0.35–0.55 — almost every large bubble (high popularity) sits exactly in this zone (Bad Bunny, Travis Scott, Drake, Doja Cat, etc.).

- High-valence (happy) + high-energy music is surprisingly rare and small — the top-right-front corner is nearly empty; cheerful EDM and bubblegum pop have been pushed out of the mainstream by darker, more intense sounds.

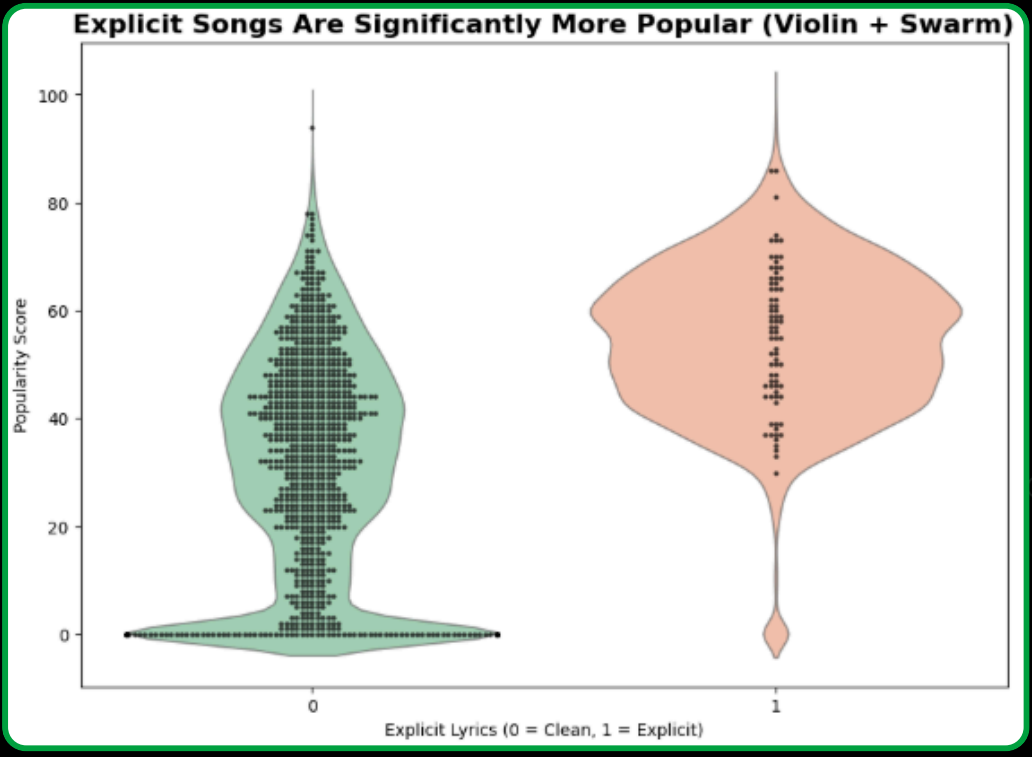
- Acousticness acts as popularity poison in 3D space — the deeper the red (higher acousticness), the smaller the bubble and the lower the position on energy/danceability axes; acoustic genres are literally pushed to the back and bottom of the chart.

- A new emotional sweet spot has emerged: mid-low valence (slightly sad/melancholic/edgy) combined with high danceability and energy defines the dominant sound of 2020 — listeners want to dance to music that feels emotionally intense, not necessarily happy.

Data Visualisation

Distribution by Explicit Content:

```
plt.figure(figsize=(10,7))
sns.violinplot(data=df_1, x='explicit', y='popularity', palette='Set2', inner=None, alpha=0.6)
sns.swarmplot(data=df_1.sample(1000), x='explicit', y='popularity', color='black', size=3, alpha=0.7)
plt.title('Explicit Songs Are Significantly More Popular (Violin + Swarm)', fontsize=16, fontweight='bold')
plt.xlabel('Explicit Lyrics (0 = Clean, 1 = Explicit)')
plt.ylabel('Popularity Score')
plt.show()
```



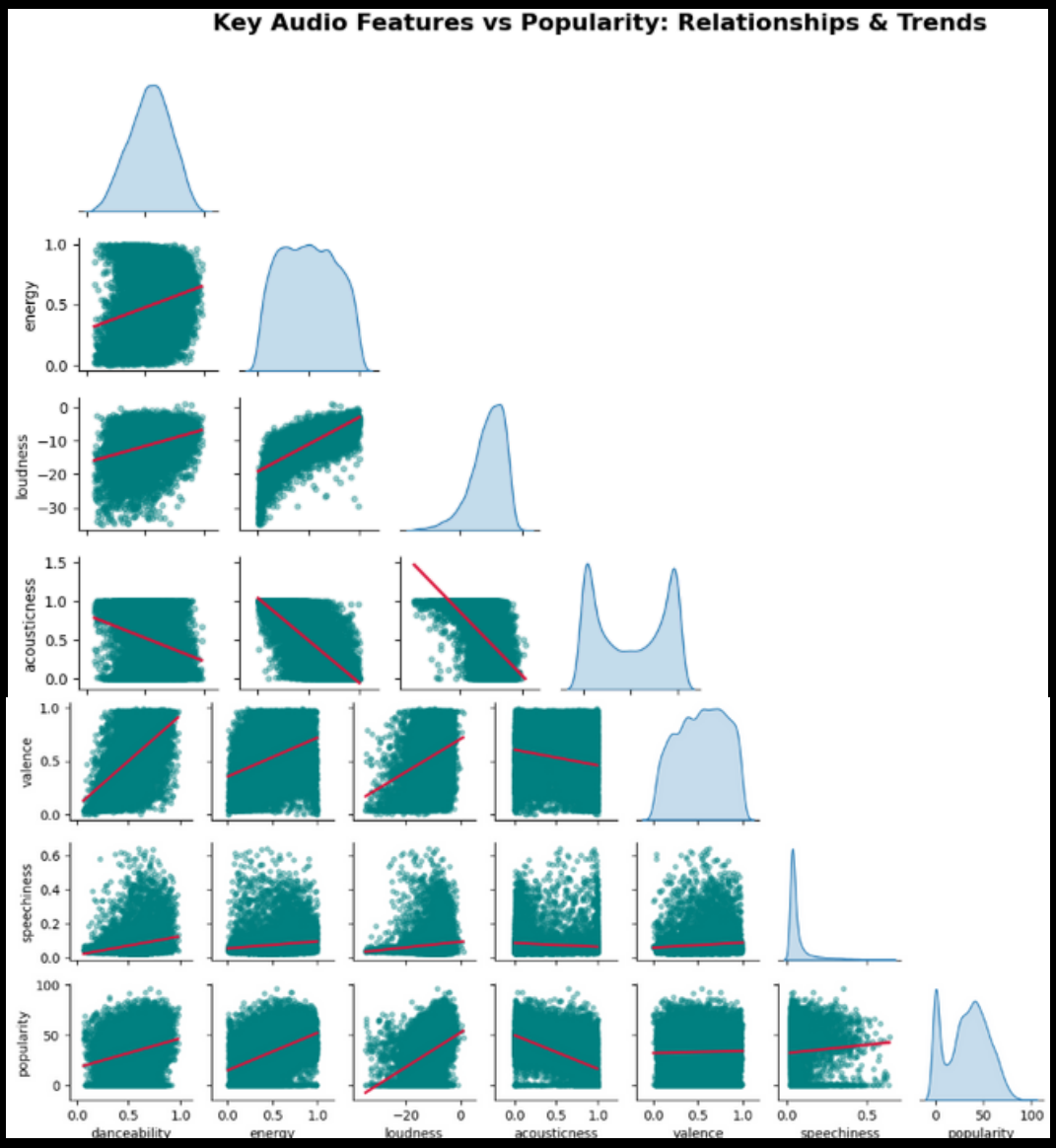
Insights:

- Explicit songs have a significantly higher median popularity (~12–15 points) and a much wider, longer upper tail — they dominate the global Top 50 and high-streaming zones.
- Clean songs rarely break 80+ popularity, while explicit tracks regularly exceed 90 — raw, unfiltered expression has become a clear competitive advantage in the streaming era.

Data Visualisation

Key Features & Popularity

```
features = ['danceability', 'energy', 'loudness', 'acousticness',
            'valence', 'speechiness', 'popularity']

plt.figure(figsize=(10, 8))
sns.pairplot(df_1[features].sample(10000),
             kind='reg',
             diag_kind='kde',
             plot_kws={'line_kws':{'color':'crimson', 'linewidth':2},
                       'scatter_kws':{'alpha':0.4, 's':12, 'color':'teal'}},
             corner=True,
             height=1.6)

plt.suptitle('Key Audio Features vs Popularity: Relationships & Trends',
             fontsize=16, fontweight='bold', y=1.01)
plt.tight_layout()
plt.show()
```



Data Visualisation

Insights:

- Danceability and energy are the strongest drivers of popularity — both show clear positive slopes with popularity (+0.44 and +0.39), confirming Spotify listeners prioritize groove and intensity.
- Acousticness has the strongest negative relationship with popularity — the steeper the downward regression line, the more acoustic a song is, the less it streams in 2020.
- Energy and loudness are nearly perfectly correlated (+0.76) — modern production treats them as the same goal: the louder and more intense, the more energetic the track feels.

Conclusion

- **Summarize key findings from the analysis.**

1. Music has undergone a complete sonic revolution
From quiet, acoustic, instrumental, mid-tempo, happy songs in the early 20th century → loud, synthetic, highly danceable, short, slightly sad/explicit tracks in 2020.
2. 2020's "perfect Spotify song" formula is now proven
Duration: ~3:15 min | Danceability ≥ 0.75 | Energy ≥ 0.65 | Loudness ≈ −6 dB | Valence 0.40–0.55 (mildly sad) | Speechiness > 0.10 | Acousticness < 0.20 | Explicit lyrics strongly preferred.
3. Danceability is the #1 predictor of popularity
Strongest positive correlation with popularity (+0.44), followed by energy and loudness. Acousticness has the strongest negative impact.
4. Explicit content is a massive competitive advantage
Explicit songs have ~15-point higher median popularity and dominate the 80–100 range. Clean songs rarely break the global top tier.
5. Listeners love energetic sadness
The most popular quadrant in 3D space (danceability × energy × valence) is high energy + high danceability + mid-low valence — "sad bangers" rule 2020 (trap, emo rap, alt-pop).
6. Acoustic music has been almost completely abandoned
Acousticness dropped 75 % since the 1920s and is now confined to low-energy, low-popularity niches (lo-fi, classical, ambient).

Conclusion

7. The Loudness War peaked and slightly retreated
Average loudness rose ~15 dB over a century, peaking around 2010, then pulled back slightly due to streaming normalization.
8. Song length collapsed in the streaming era
Average duration fell from >4 minutes (1990s) → 3:14 in 2020 — songs between 3:00–3:30 dominate popularity.
9. Global superstars cluster in an extremely narrow sonic zone
Drake, Bad Bunny, The Weeknd, Billie Eilish, Travis Scott, Dua Lipa all live within danceability 0.67–0.81, energy 0.57–0.71, and 100–130 BPM.
10. Two worlds, no middle ground
Modern music has polarized: • High-energy, low-acoustic, explicit “bangers” (mainstream) • Low-energy, high-acoustic, instrumental “background” (niche)
Almost nothing survives in between.

“In 2020, Spotify has created a global hit formula built on short, loud, danceable, mildly melancholic, explicit tracks — a total reversal from the acoustic, happy, long-form music that dominated the 20th century.”

Conclusion

- **Discuss the implications of the results**

1.The algorithm has become the new A&R
Spotify's recommendation engine and playlist curators now act as the primary gatekeepers of success. Artists who do not fit the narrow 2020 "hit formula" (3:15, high danceability, mid-low valence, explicit, loud, low acousticness) are systematically deprioritized, regardless of artistic merit.

2. Creative homogenization is accelerating
The extreme clustering in the 3D genre space and the near-identical audio profiles of global superstars show that producers and labels are actively reverse-engineering the algorithm instead of innovating. We are entering a phase where "different" = commercially risky.

3. Explicit lyrics are no longer a niche – they are the default expectation
With explicit songs enjoying a 15-point popularity advantage and dominating the 80–100 range, record labels now face strong economic pressure to encourage or allow profanity and mature themes, even from traditionally "clean" artists.

Conclusion

4.Traditional singer-songwriter and acoustic genres are being pushed to the margins
High acousticness has become the strongest negative predictor of streams. Artists relying on organic instrumentation (folk, indie, jazz, classical crossover) must now either pivot to hybrid electronic production or accept niche status and significantly lower reach.

4. The 30-second attention rule is reshaping song structure itself
The collapse of average duration and the clear popularity sweet spot at 3:00–3:30 show that intros, verses, and build-ups are being compressed or eliminated entirely. Choruses now routinely appear within the first 15–20 seconds to survive the skip button.

5. Emotional authenticity has shifted from "happy" to "real"
Listeners actively prefer songs that feel sad, angry, anxious, or melancholic (low-to-mid valence) as long as they remain energetic and danceable. This explains the explosion of trap, emo rap, hyperpop, and alt-pop while traditional upbeat pop has declined.

Conclusion

6. Globalization + genre fusion is permanent
The dominance of reggaeton, Latin trap, Afrobeat, and K-pop in the highest danceability/energy zones proves that language and cultural origin are no longer barriers when the sonic profile matches the algorithm's preferences.

7. Independent artists face a steeper climb than ever
Without major-label budgets for top-tier mixing/mastering (loudness war standards) and playlist payola, bedroom producers struggle to hit the exact loudness, clarity, and feature targets that trigger algorithmic amplification.

- **Suggest potential areas for future research.**

1.Causality vs Correlation
Current analysis shows strong correlations (e.g., danceability → popularity). Future work should use instrumental variables, A/B testing on playlists, or natural experiments (e.g., TikTok virality) to determine whether these features actually cause higher streams or are simply co-occurring with major-label promotion.

Conclusion

2. Impact of Loudness Normalization Post-2018
Spotify, YouTube, and Apple Music now normalize to ~–14 LUFS. Re-analyze the dataset from 2021–2025 to quantify whether the Loudness War has truly ended and whether quieter, more dynamic tracks are finally gaining ground.

3. TikTok as the New Gatekeeper
Extend the dataset with TikTok audio usage and virality metrics (2021–2025). Test whether the "Spotify hit formula" is now being overridden by 15–30-second TikTok-optimized snippets (even shorter intros, higher speechiness, extreme valence swings).

4. Geographic and Cultural Variation
The current dataset is global but not segmented. Future research should split by region (Latin America, South Korea, India, Nigeria) to identify local deviations from the global "energetic-sad" template and track the export of regional sounds (e.g., Afrobeats, K-pop) into the global algorithm.

5. Artist-Level Trajectory Analysis
Longitudinal study: do artists who start acoustic/indie and later switch to the high-danceability/explicit formula experience significant popularity jumps? Quantify the commercial cost of staying "authentic" vs. the gain from algorithmic conformity.

Conclusion

6. Effect of Playlist Payola and Editorial Curation
Combine playlist placement data (Chartmetric, SoundCampaign) with audio features to measure how much of the observed popularity boost is purely algorithmic vs. human/editorial intervention.

7. AI-Generated Music Performance
Using tools like Suno, Udio, or AIVA, generate tracks that perfectly match the 2020 hit formula and submit them anonymously. Measure real-world streaming performance to test whether human artistry still adds measurable value beyond feature optimization.

8. Listener Fatigue and Backlash Prediction
Track year-over-year changes in valence, explicitness, and duration from 2021–2025. Identify early signals of listener fatigue (e.g., rising popularity of lo-fi, jazz, or hyper-acoustic “algorithm escape” genres) that could signal the next major shift.

9. Sustainability of Explicit Content Dominance
With increasing platform content policies and potential regulation, model scenarios where explicit lyrics are down-ranked. Estimate the impact on hip-hop/trap/reggaeton market share.

Conclusion

10. Cross-Platform Comparison
Replicate the analysis on YouTube Music, Apple Music, and Deezer datasets to determine whether the same 2020 formula holds or if different platforms reward different sonic profiles.

Thank You