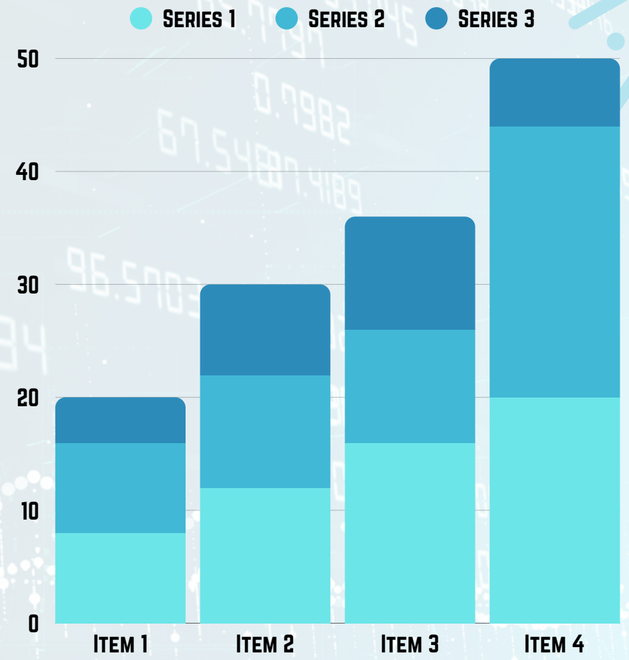
#Spotify Tracks Data Analysis



DATA MINDS

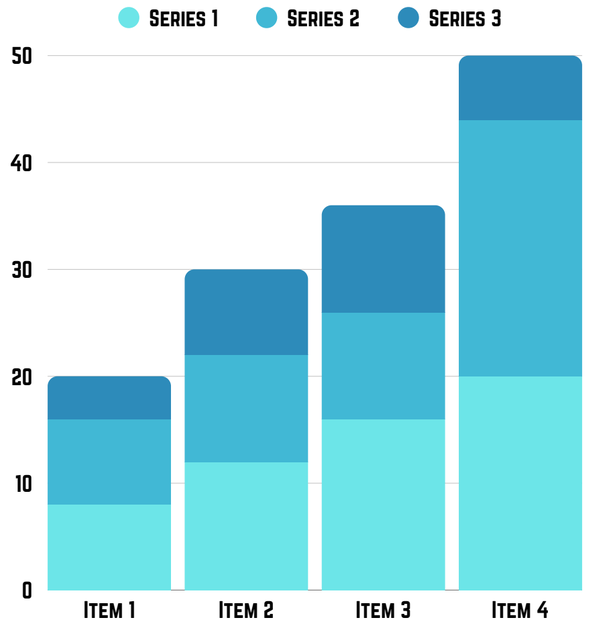
➡Presented by Anshika Pandey

- Anshika pandey
- Dipanjan Halder
- Ekadashi Sardar
- Nilanjana Saren

Exploratory Data Analysis on Spotify Tracks Dataset

About Spotify:

Spotify is a global music streaming platform with millions of tracks. Each song contains audio features like danceability, energy, loudness, and tempo that define its musical characteristics.



Objective:

To explore the Spotify dataset and identify which audio features influence a song’s popularity.

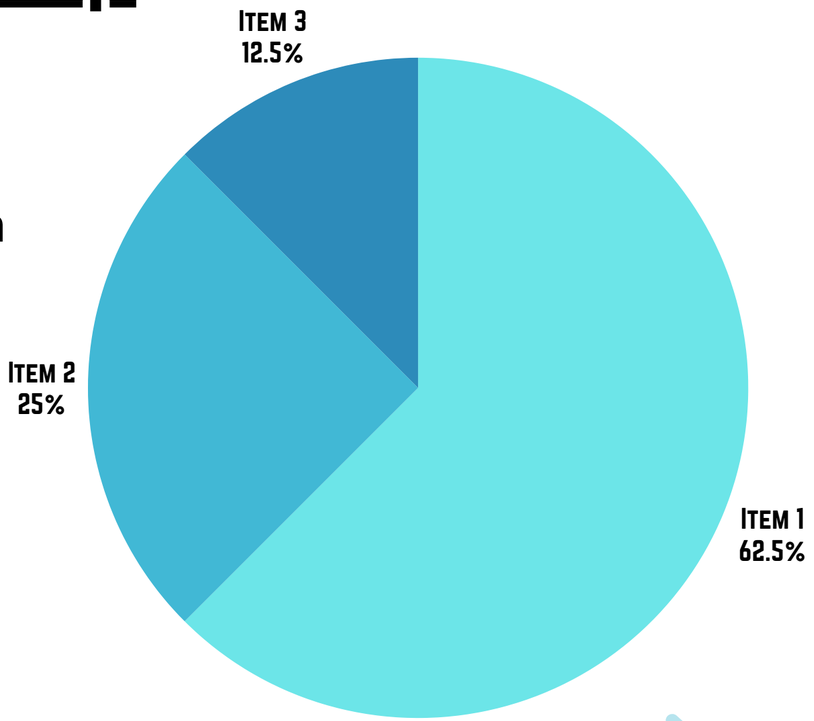
Phase 1: Preparation & Setup

Project Initialization

- Import essential Python libraries(pandas,numpy,matplotlib,seaborn
- Data Acquisition
- Load Spotify Dataset into the notebook

Data Scrutiny(the cleanup Mission)

- Detect and handle missing value
- Ensure dataset iclean,consistent,and reliable



Data Cleaning:Handling Missing Values

Dataset
(Incuding Missing Value)

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 62239 entries, 0 to 62316
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62239 non-null  object
 1   track_name        62239 non-null  object
 2   artist_name       62239 non-null  object
 3   year              62239 non-null  int64
 4   popularity        62239 non-null  int64
 5   artwork_url       62239 non-null  object
 6   album_name        62239 non-null  object
 7   acousticness      62239 non-null  float64
 8   danceability      62239 non-null  float64
 9   duration_ms       62239 non-null  float64
 10  energy            62239 non-null  float64
 11  instrumentalness  62239 non-null  float64
 12  key               62239 non-null  float64
 13  liveness          62239 non-null  float64
 14  loudness          62239 non-null  float64
 15  mode              62239 non-null  float64
 16  speechiness       62239 non-null  float64
 17  tempo             62239 non-null  float64
 18  time_signature    62239 non-null  float64
 19  valence           62239 non-null  float64
 20  track_url         62239 non-null  object
 21  language          62239 non-null  object
 22  duration_min      62239 non-null  float64
dtypes: float64(14), int64(2), object(7)
memory usage: 11.4+ MB
```

es Terminal

List of missing values

```
print("Any missing values in entire DataFrame? "
print("Total missing values: ", df.isnull().sum(
print("Missing per column:\n", df.isnull().sum()
```

```
Any missing values in entire DataFrame?  False
Total missing values:  0
Missing per column:
 track_id            0
track_name           0
artist_name          0
year                 0
popularity           0
artwork_url          0
album_name           0
acousticness         0
danceability         0
duration_ms          0
energy               0
instrumentalness     0
key                  0
liveness             0
loudness             0
mode                 0
speechiness          0
tempo                0
time_signature       0
valence              0
track_url            0
language             0
duration_min         0
dtype: int64
```

Final Datasets

```
<class 'pandas.core.frame.DataFrame'>
Index: 62239 entries, 0 to 62316
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62239 non-null  object
 1   track_name        62239 non-null  object
 2   artist_name       62239 non-null  object
 3   year              62239 non-null  int64
 4   popularity        62239 non-null  int64
 5   artwork_url       62239 non-null  object
 6   album_name        62239 non-null  object
 7   acousticness      62239 non-null  float64
 8   danceability      62239 non-null  float64
 9   duration_ms       62239 non-null  float64
 10  energy            62239 non-null  float64
 11  instrumentalness  62239 non-null  float64
 12  key               62239 non-null  float64
 13  liveness          62239 non-null  float64
 14  loudness          62239 non-null  float64
 15  mode              62239 non-null  float64
 16  speechiness       62239 non-null  float64
 17  tempo             62239 non-null  float64
 18  time_signature    62239 non-null  float64
 19  valence           62239 non-null  float64
 20  track_url         62239 non-null  object
 21  language          62239 non-null  object
dtypes: float64(13), int64(2), object(7)
```

Phase 2: Initial Reconnaissance

Initial survey

- Explore the dataset structure (.info())
- Generate a statistical summary (.describe())
- Check dataset size (.shape)
- Preview sample rows (.head())

Data Classification

- Identify Numerical variables
- Identify Categorical variables
- Identify Temporal variables

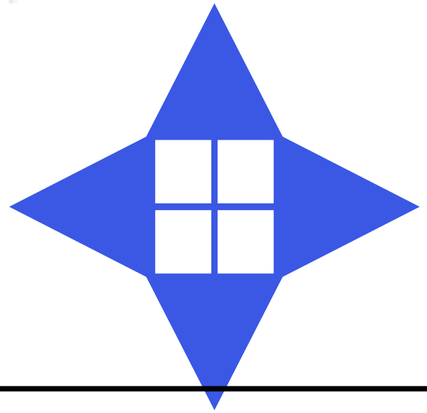
Dataset Exploration & Classification

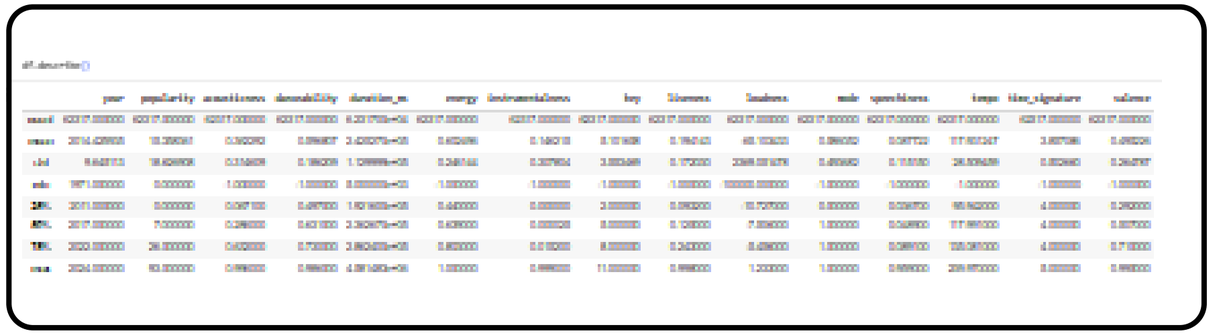
```
df.shape
```

```
(62317, 22)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62317 entries, 0 to 62316
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62317 non-null  object
 1   track_name        62317 non-null  object
 2   artist_name       62317 non-null  object
 3   year              62317 non-null  int64
 4   popularity        62317 non-null  int64
 5   artwork_url       62317 non-null  object
 6   album_name        62317 non-null  object
 7   acousticness      62317 non-null  float64
 8   danceability      62317 non-null  float64
 9   duration_ms       62317 non-null  float64
 10  energy            62317 non-null  float64
 11  instrumentalness  62317 non-null  float64
 12  key               62317 non-null  float64
 13  liveness          62317 non-null  float64
 14  loudness          62317 non-null  float64
 15  mode              62317 non-null  float64
 16  speechiness       62317 non-null  float64
 17  tempo             62317 non-null  float64
 18  time_signature    62317 non-null  float64
 19  valence           62317 non-null  float64
 20  track_url         62317 non-null  object
 21  language          62317 non-null  object
dtypes: float64(13), int64(2), object(7)
memory usage: 10.5+ MB
```





1. List numerical columns

```
num_cols = df.select_dtypes(include=['int64','float64']).columns
num_cols
```

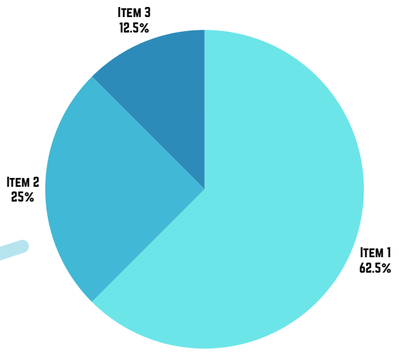
```
Index(['year', 'popularity', 'acousticness', 'danceability', 'duration_ms',
       'energy', 'instrumentalness', 'key', 'liveness', 'loudness', 'mode',
       'speechiness', 'tempo', 'time_signature', 'valence'],
      dtype='object')
```

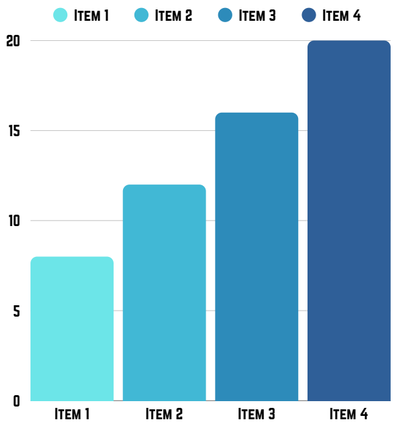
```
# Basic cleaning for the 3 categorical columns we will analyze
cols = ['artist_name', 'album_name', 'language']
missing = [c for c in cols if c not in df.columns]
if missing:
    raise ValueError(f"Missing required column(s): {missing}")

# Fill NAs, strip whitespace; keep artist/album case as-is but normalize language to lowercase
df['artist_name'] = df['artist_name'].fillna('Unknown').astype(str).str.strip()
df['album_name']  = df['album_name'].fillna('Unknown').astype(str).str.strip()
df['language']    = df['language'].fillna('unknown').astype(str).str.strip().str.lower()

# Quick uniques summary
print('Unique artists:', df['artist_name'].nunique())
print('Unique albums :', df['album_name'].nunique())
print('Unique langs  :', df['language'].nunique())
```

```
Unique artists: 12513
Unique albums : 19898
Unique langs  : 7
```

Phase 3:Uncovering Insight(Core mission)

- Univariate Analysis Numerical: Histograms, Box Plots, Density Plots → Understand distribution Categorical: Bar Charts, Count Plots → Frequency of categories
- Bivariate Analysis Numerical vs. Numerical: Scatter Plots, Correlation Matrices → Relationships Numerical vs. Categorical: Box Plots → Compare distributions
- Multivariate Analysis Explore 3+ variables together Use advanced visualizations (e.g., scatter plot with color groups) Identify deeper trends & patterns





Statical Summary of numerical variables

1. Measure of central tendency

```
1. Central Tendency of 'Popularity':
   Mean Popularity: 15.36
   Median Popularity: 7.00
   Mode Popularity: [0]
====================================

2. Statistics for 'Duration (ms)':
   Mean Duration: 242603.45 ms
   Median Duration: 236311.00 ms
   Minimum Duration: 5000.00 ms
   Maximum Duration: 4581483.00 ms
====================================
```

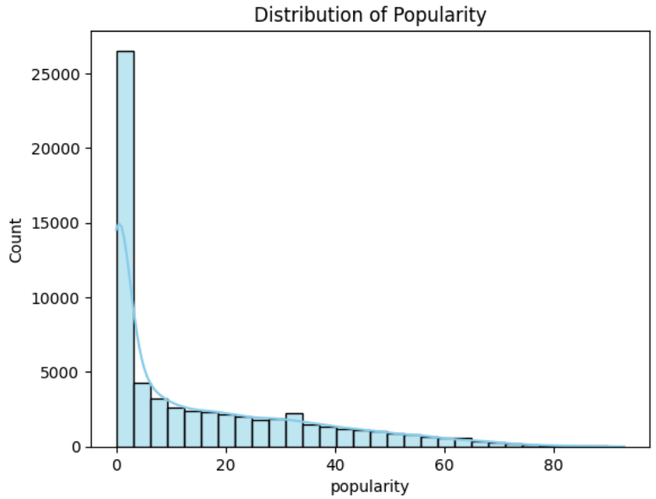
2.Measure of Dispersion

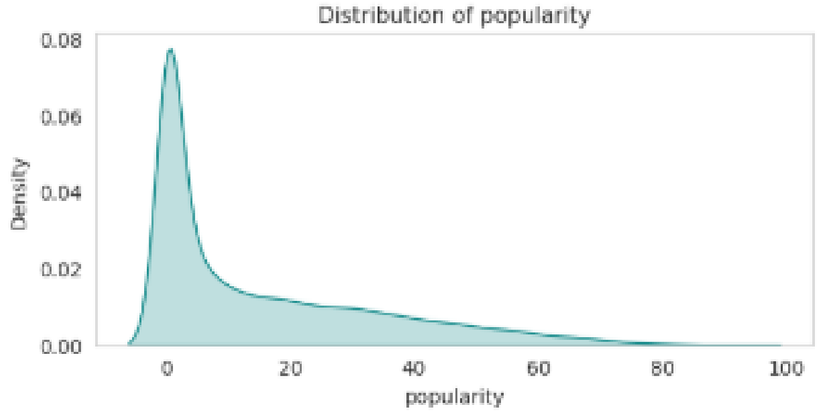
```
   Danceability Description:
count    62239.000000
mean         0.596768
std          0.186262
min         -1.000000
25%          0.497000
50%          0.631000
75%          0.730000
max          0.986000
Name: danceability, dtype: float64

   Energy Description:
count    62239.000000
mean         0.602416
std          0.246207
min         -1.000000
25%          0.440000
50%          0.639000
75%          0.803000
max          1.000000
Name: energy, dtype: float64
```

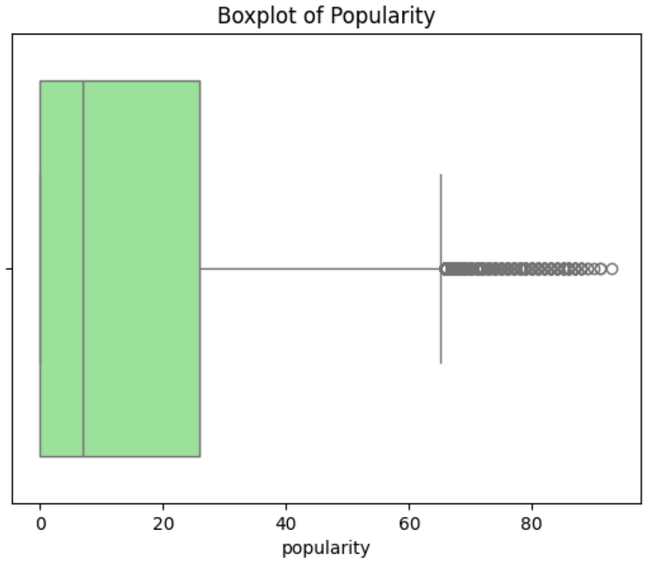
3.Distribution Analysis

```
4. Total and Average 'Popularity' and 'Loudness':
   Total Popularity (sum across tracks): 955,841.00
   Average Popularity: 15.36
   Average Loudness: -65.17 dB
   Min Loudness: -100000.00 dB
   Max Loudness: 1.23 dB
================================================================================

5. Skewness and Kurtosis of 'Popularity', 'Danceability', and 'Energy':
   Skewness of Popularity: 1.2314
   Kurtosis of Popularity: 0.6620
   Skewness of Danceability: -1.0577
   Kurtosis of Danceability: 2.9039
   Skewness of Energy: -0.6630
   Kurtosis of Energy: 0.3522
```

Visual Analysis of the numerical Variables (popularity)

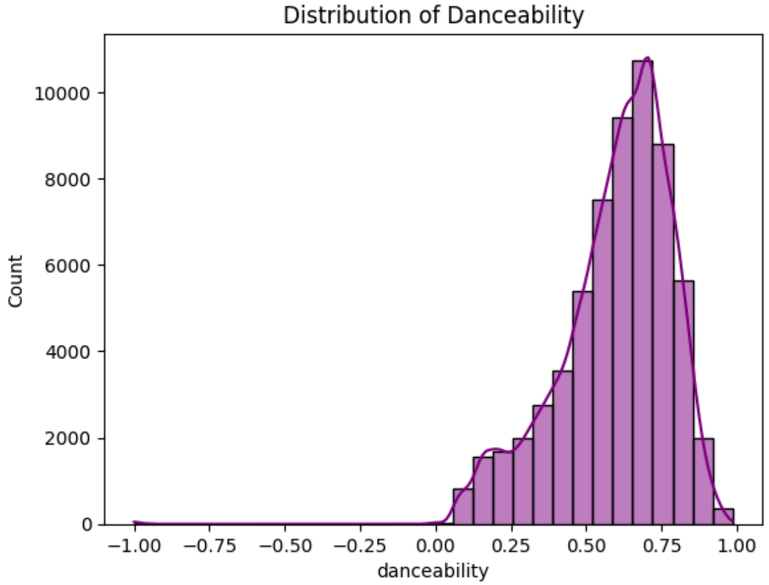


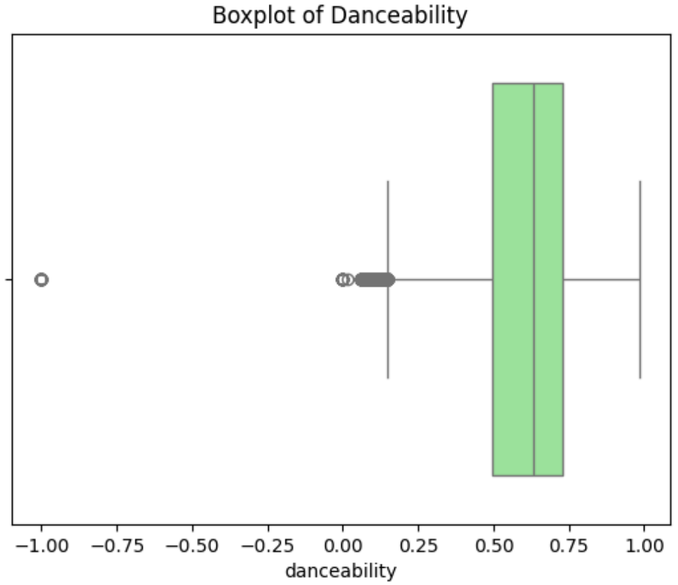


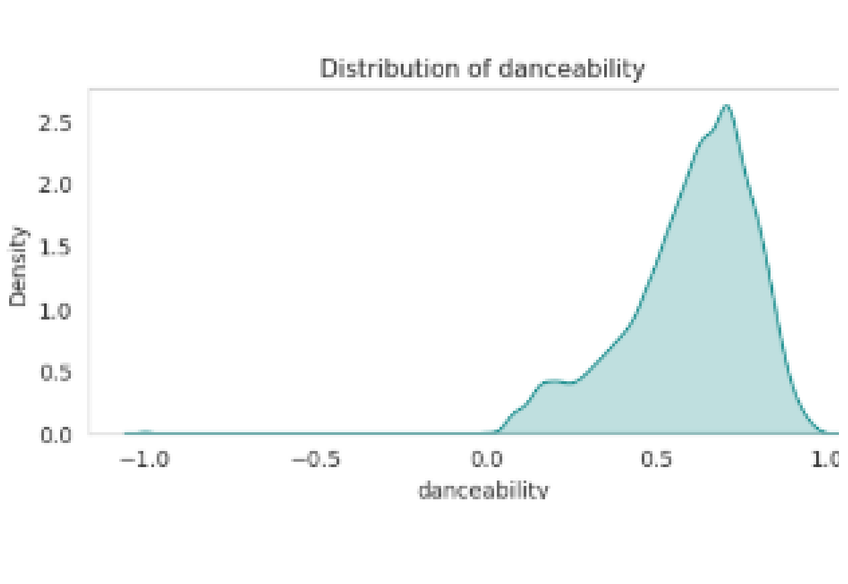


	popularity
count	62239.000000
mean	15.357589
std	18.630494
min	0.000000
25%	0.000000
50%	7.000000
75%	26.000000
max	93.000000

Visual Analysis of the numerical Variables (Danceability)





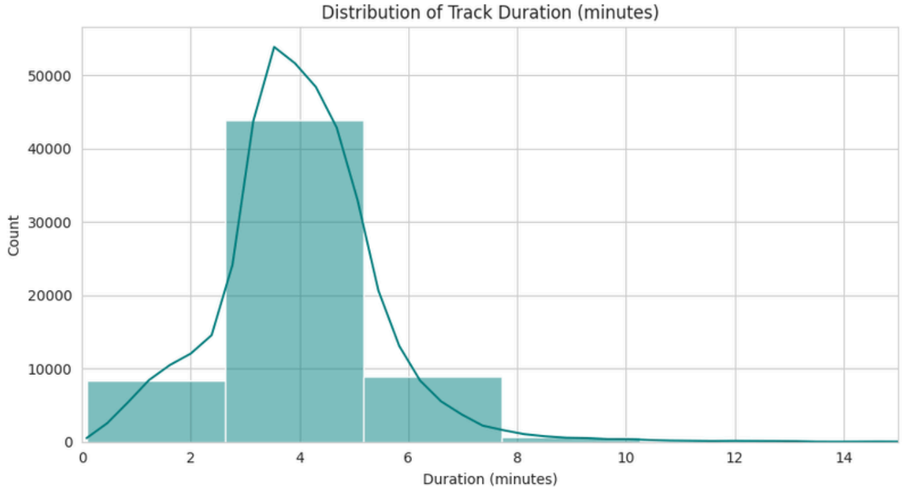


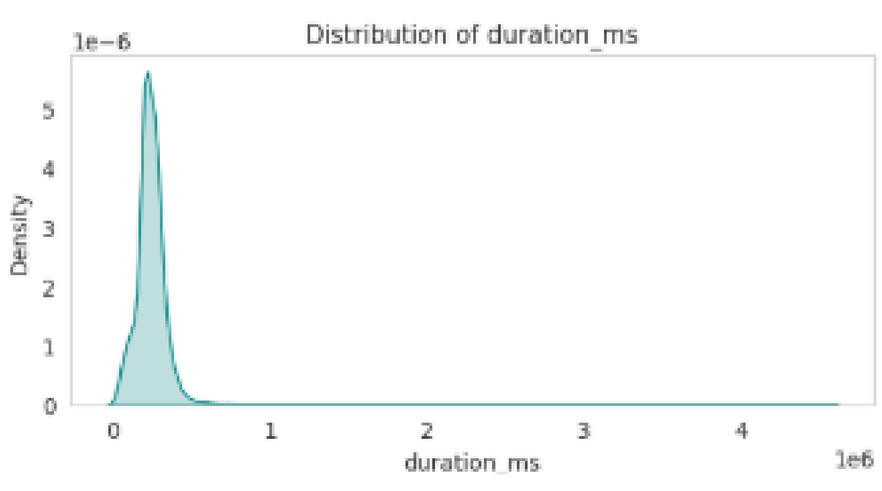
danceability

count	62239.000000
mean	0.598768
std	0.186262
min	-1.000000
25%	0.497000
50%	0.631000
75%	0.730000
max	0.988000

dtype: float64

Visual Analysis of the numerical Variables (Duration)

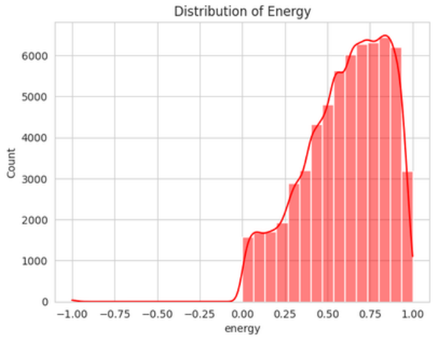


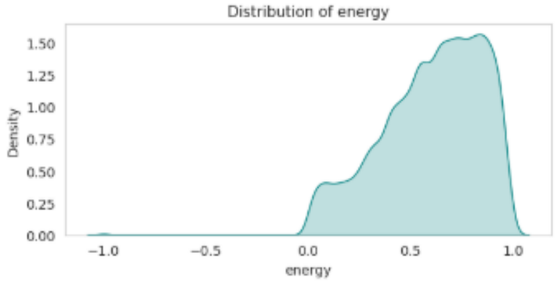


	duration_min
count	62239.000000
mean	4.043391
std	1.883683
min	0.083333
25%	3.204000
50%	3.938517
75%	4.771725
max	76.358050

Visual Analysis of the numerical Variables

(Distribution of energy levels)

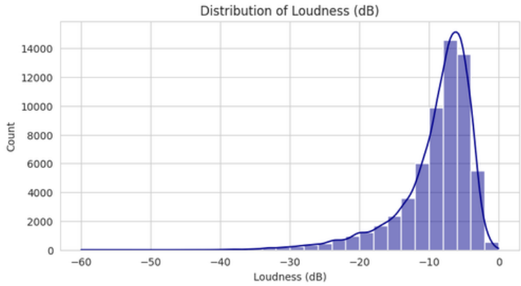


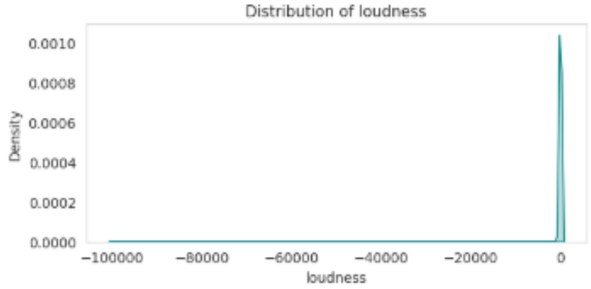


	energy
count	62239.000000
mean	0.602416
std	0.246207
min	-1.000000
25%	0.440000
50%	0.639000
75%	0.803000
max	1.000000

dtype: float64

(Distribution of Loudness)





	loudness
count	62192.000000
mean	-8.946879
std	5.331750
min	-60.000000
25%	-10.722000
50%	-7.505000
75%	-5.455000
max	-0.005000

dtype: float64

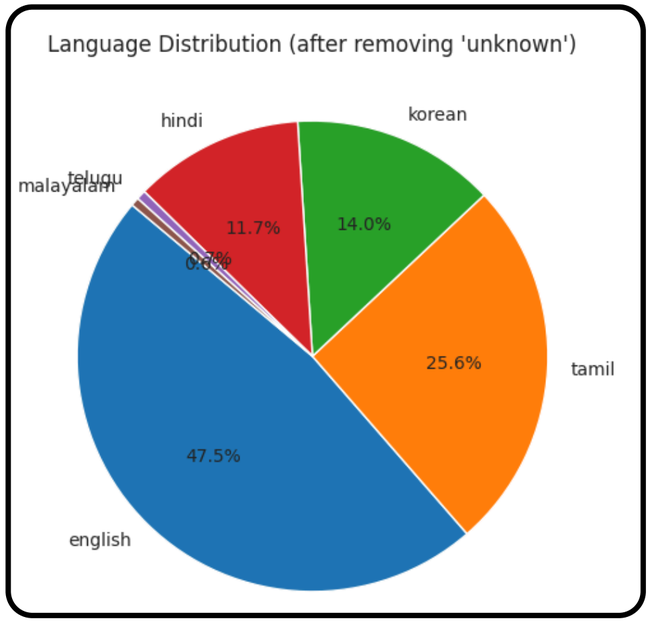
Overall Insights from Univariate Analysis of Numerical Variables

Year: Most tracks are from 2000 onwards, with a sharp surge after 2010, reflecting Spotify's focus on modern music.
Popularity: Highly right-skewed — majority of songs have scores below 20, while only a few surpass 70.
Acousticness: Dominated by non-acoustic tracks near 0, though some fully acoustic songs exist. Negative values, if any, require cleaning.
Danceability: Skews high, with a median around 0.65, showing tracks are generally rhythmic and dance-oriented.
Duration (ms): Standard length is 3–5 minutes, with rare outliers exceeding 20 minutes.
Energy: Concentrated between 0.6 and 0.9, highlighting that most tracks are upbeat and lively.
Instrumentalness: Most songs are vocal-heavy (near 0), but a smaller subset is purely instrumental (>0.8).
Key: Distribution is fairly uniform across 12 keys — no dominant key observed.
Liveness: Majority below 0.2, indicating studio recordings dominate; live performances are rare.
Loudness (dB): Typically ranges between -12 dB and -6 dB, standard for mastered music; very quiet tracks (<-40 dB) are extreme outliers.
Mode: Roughly 60% in major and 40% in minor keys; invalid values like -1 must be treated.
Speechiness: Most values are <0.1, confirming music dominance; higher levels reflect rap or spoken-word tracks.
Tempo (BPM): Concentrated around 120–140 BPM, aligning with pop/dance norms; extreme tempos are rare.
Time Signature: Strongly dominated by 4/4, with minor shares in 3/4 and 5/4; invalid entries (-1, 0, 1) should be cleaned.
Valence: Mostly between 0.3–0.8, meaning tracks are generally positive and cheerful; very low values (<0.2) suggest rare darker tones.

Visual Analysis of the categorical variables

(Language Distribution)

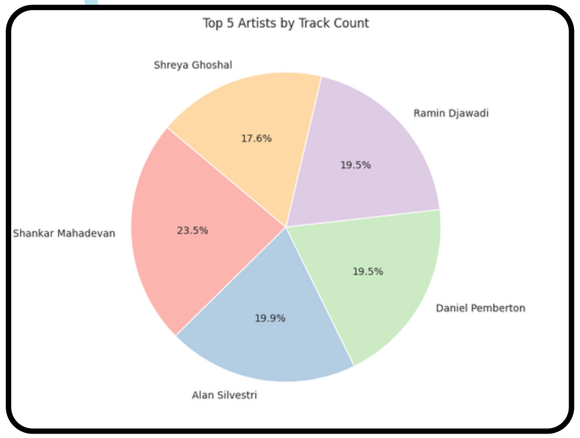
```
Language distribution after removing 'unknown'
language
english      23389
tamil        12609
korean        6893
hindi         5740
telugu         321
malayalam      282
Name: count, dtype: int64
```

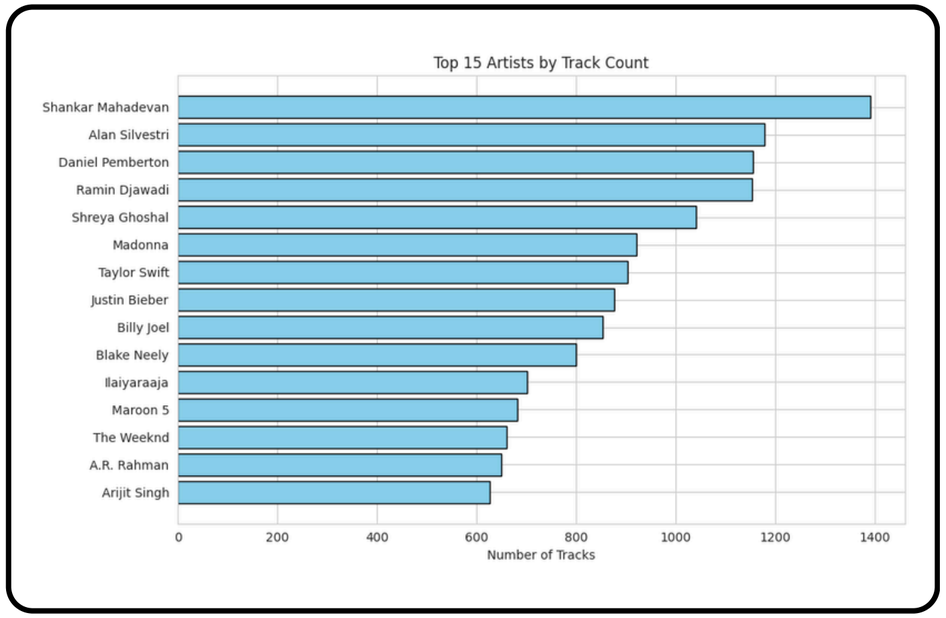


Language Distribution Insights

- English dominates with 47.5% (23,389 tracks).
- Tamil follows with 25.6% (12,609 * tracks), showing strong South Indian music presence.
- Korean (14%) and Hindi (11.7%) also have significant shares.
- Telugu (0.7%) and Malayalam (0.6%) are minimally represented.

Top Artists





	Artist	Track Count	Percent of Dataset
0	Shankar Mahadevan	1391	2.23
1	Alan Silvestri	1178	1.89
2	Daniel Pemberton	1156	1.86
3	Ramin Djawadi	1154	1.85
4	Shreya Ghoshal	1041	1.67
5	Madonna	921	1.48
6	Taylor Swift	903	1.45
7	Justin Bieber	876	1.41
8	Billy Joel	853	1.37
9	Blake Neely	799	1.28
10	Ilaiyaraaja	701	1.13
11	Maroon 5	682	1.10
12	The Weeknd	660	1.06
13	A.R. Rahman	649	1.04
14	Arijit Singh	626	1.01

Top Artists insight

- Shankar Mahadevan leads with 1391 tracks (2.23%).
- Indian playback singers (Shreya Ghosal, Arijit Singh,A.R.Rahman,Ilaiyaraaja) and global pop stars (Madonna,Taylor Swift, Justin Bieber,Maroon 5) are well represented.**
- The List balances film music, Indian playback ,and global pop.

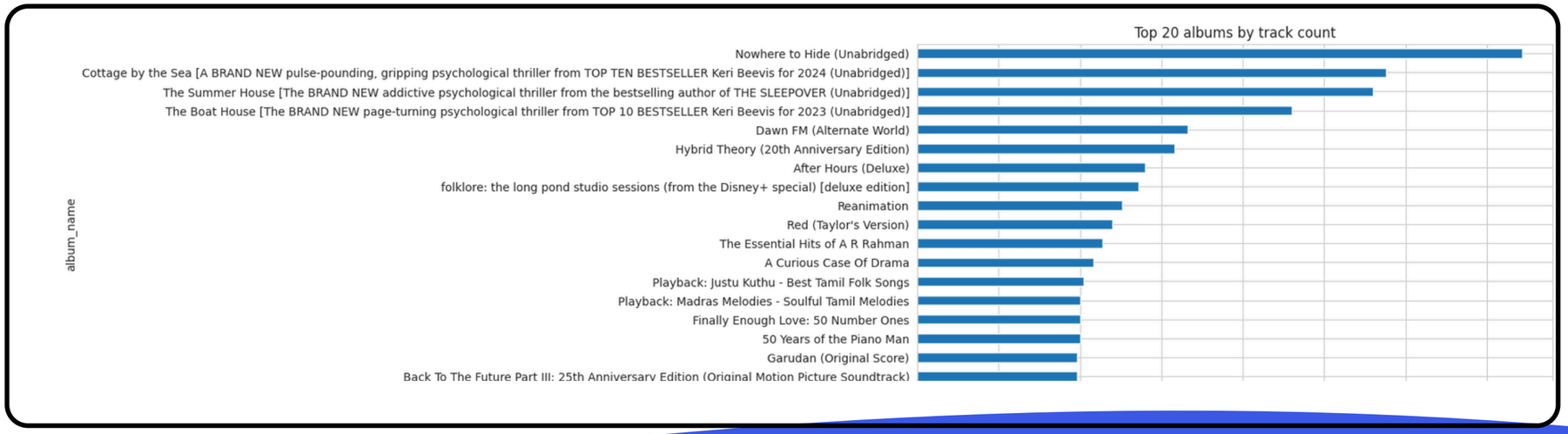
Top 20 Album by track count

Top Albums Insights

- Nowhere to Hide (Unabridged) is the largest album with 186 tracks.
- Audiobook-style albums like Cottage by the Sea, The Summer House, and The Boat House rank among the top.
- Popular music albums (Dawn FM, Hybrid Theory, After Hours, Red – Taylor's Version) al so feature strongly.
- Indian music influence appears with The Essential Hits of A.R. Rahman and Tamil folk collections.
- Top 20 albums together contribute <3% of the dataset, showing wide distribution.

Top 20 albums:

album_name	count	percent_of_dataset
Nowhere to Hide (Unabridged)	186	0.30
Cottage by the Sea [A BRAND NEW pulse-pounding, gripping psychological thriller from TOP TEN BESTSELLER Keri Beevis for 2024 (Unabridged)]	144	0.23
The Summer House [The BRAND NEW addictive psychological thriller from the bestselling author of THE SLEEPOVER (Unabridged)]	140	0.22
The Boat House [The BRAND NEW page-turning psychological thriller from TOP 10 BESTSELLER Keri Beevis for 2023 (Unabridged)]	115	0.18
Dawn FM (Alternate World)	83	0.13
Hybrid Theory (20th Anniversary Edition)	79	0.13
After Hours (Deluxe)	70	0.11
folklore: the long pond studio sessions (from the Disney+ special) [deluxe edition]	68	0.11
Reanimation	63	0.10
Red (Taylor's Version)	60	0.10
The Essential Hits of A R Rahman	57	0.09
A Curious Case Of Drama	54	0.09
Playback: Justu Kuthu - Best Tamil Folk Songs	51	0.08
Finally Enough Love: 50 Number Ones	50	0.08
Playback: Madras Melodies - Soulful Tamil Melodies	50	0.08
50 Years of the Piano Man	50	0.08
Back To The Future Part III: 25th Anniversary Edition (Original Motion Picture Soundtrack)	49	0.08
Garudan (Original Score)	49	0.08
Justice (The Complete Edition)	48	0.08
My Lives	48	0.08

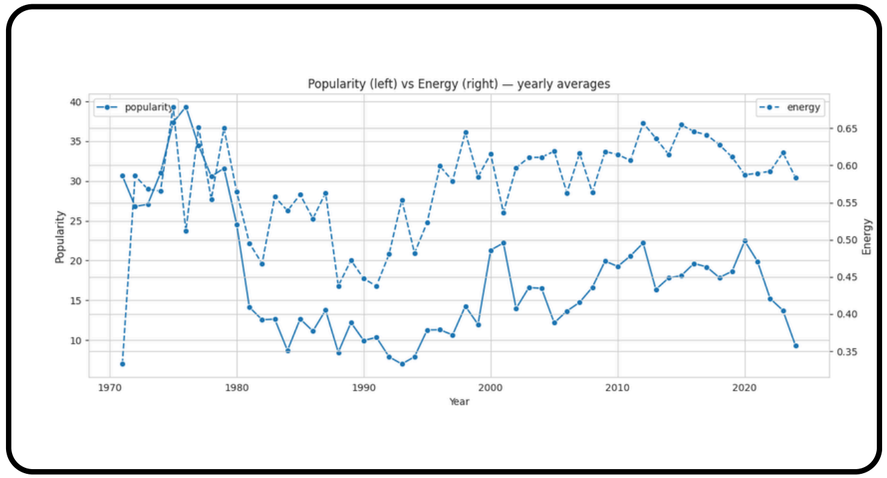


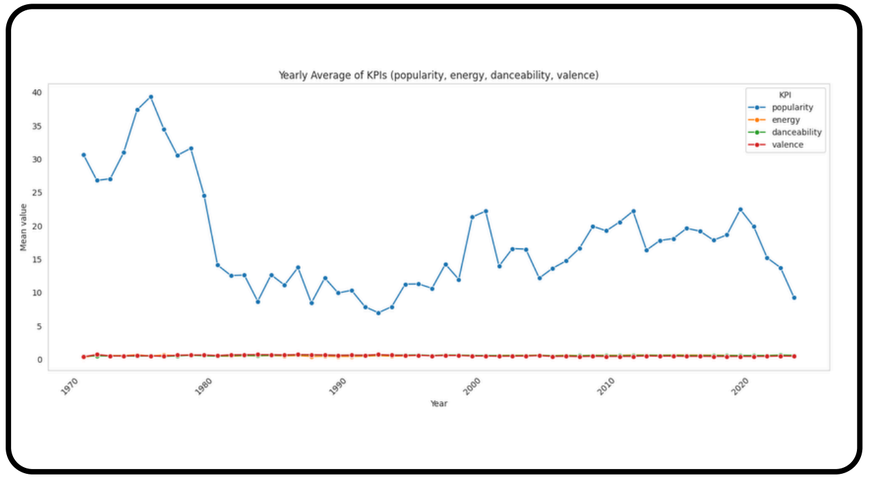
Time Series Analysis (Absolute Values of KPIs)

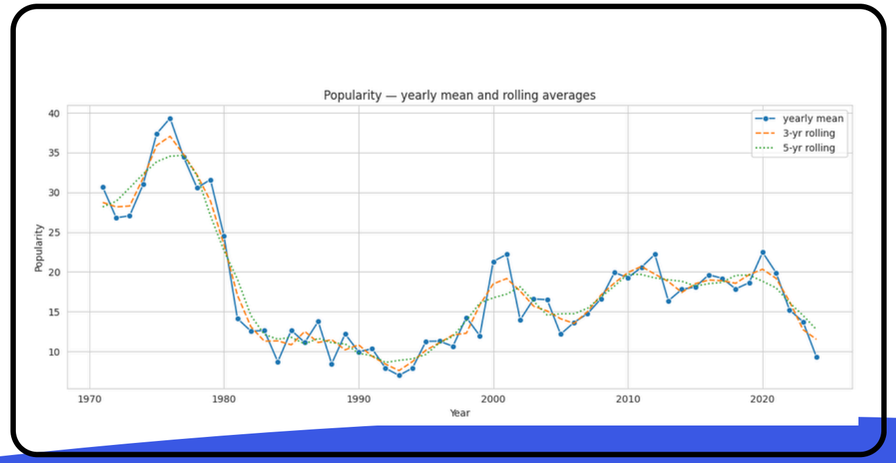
KPI selection and yearly aggregation

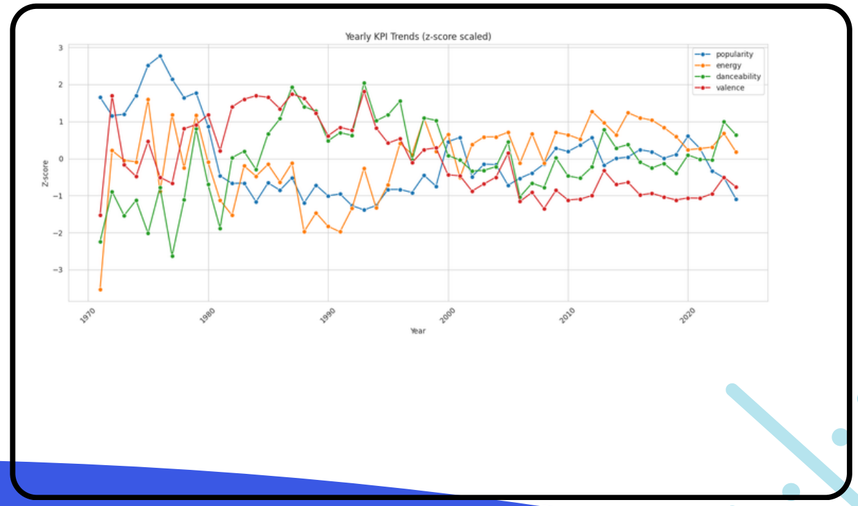
```
# 1. KPI selection and yearly aggregation
kpis = ['popularity', 'energy', 'danceability', 'valence']
yearly = df.groupby('year')[kpis].mean().reset_index().sort_values('year')
yearly.head()
```

	year	popularity	energy	danceability	valence
0	1971	30.666667	0.332800	0.467250	0.397050
1	1972	26.809524	0.585905	0.534714	0.722048
2	1973	27.064516	0.568048	0.502290	0.533889
3	1974	31.000000	0.565290	0.523000	0.502420
4	1975	37.368421	0.678789	0.478421	0.597632





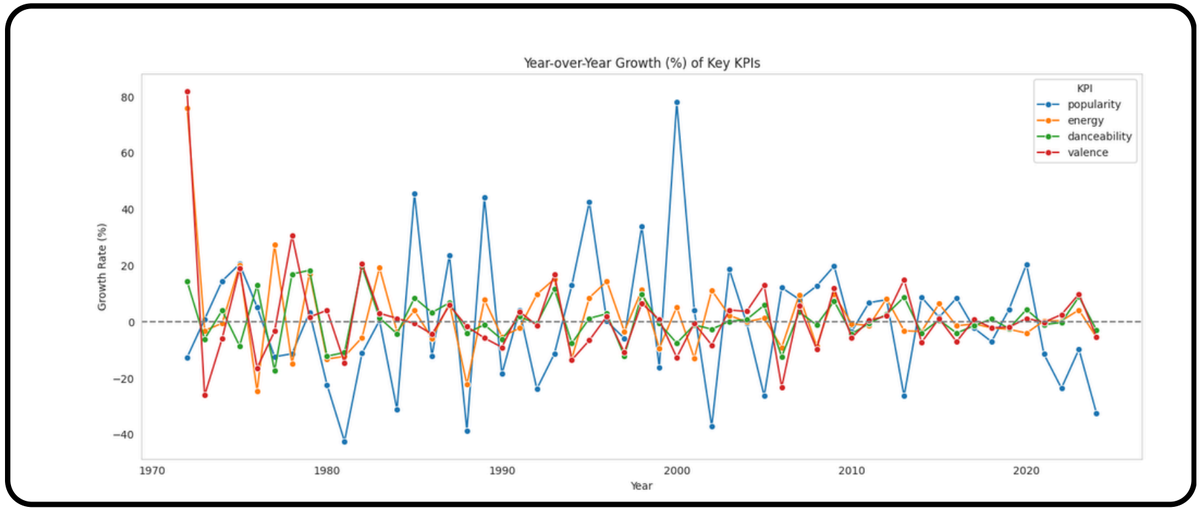




Time Series Analysis (Growth of KPIs)

```
# 🔁 Find years with biggest positive and negative growth
growth_report = {}
for col in kpis:
    growth_report[col] = {
        'max_growth': yearly_growth.loc[yearly_growth[f'{col}_growth_pct'].idxmax(), ['year', f'{col}_growth_pct']].to_d
        'max_decline': yearly_growth.loc[yearly_growth[f'{col}_growth_pct'].idxmin(), ['year', f'{col}_growth_pct']].to_
    }
import pprint
pprint.pprint(growth_report)
```

```
{'danceability': {'max_decline': {'danceability_growth_pct': -17.2152138859907,
                                  'year': 1977.0},
                  'max_growth': {'danceability_growth_pct': 19.645040179517427,
                                 'year': 1982.0}},
 'energy': {'max_decline': {'energy_growth_pct': -24.620712827272484,
                            'year': 1976.0},
            'max_growth': {'energy_growth_pct': 76.05311355311355,
                           'year': 1972.0}},
 'popularity': {'max_decline': {'popularity_growth_pct': -42.42307480596686,
                                'year': 1981.0},
                'max_growth': {'popularity_growth_pct': 78.01719179831512,
                               'year': 2000.0}},
 'valence': {'max_decline': {'valence_growth_pct': -26.059072062086607,
                             'year': 1973.0},
             'max_growth': {'valence_growth_pct': 81.85307111617225,
                            'year': 1972.0}}}
```



Insight:Time series growth of KPIs

- Popularity growth shows major spikes around key industry periods (e.g., 2000s digital rise, 2020s streaming boom).
- Energy and Danceability often move together — fast-growing periods reflect shifts toward upbeat, high-BPM genres.
- Valence (positivity) growth fluctuates — dips may correspond to trends toward moodier or darker tones.
- Sharp negative growth years can signal data bias (few tracks or incomplete records) — handle carefully.
- Sustained positive growth across multiple KPIs indicates evolving listener preference toward energetic, danceable, and emotionally bright music.

Introduction to Bivariate & Multivariate Analysis

Bivariate Analysis → Explore relationships between two variables

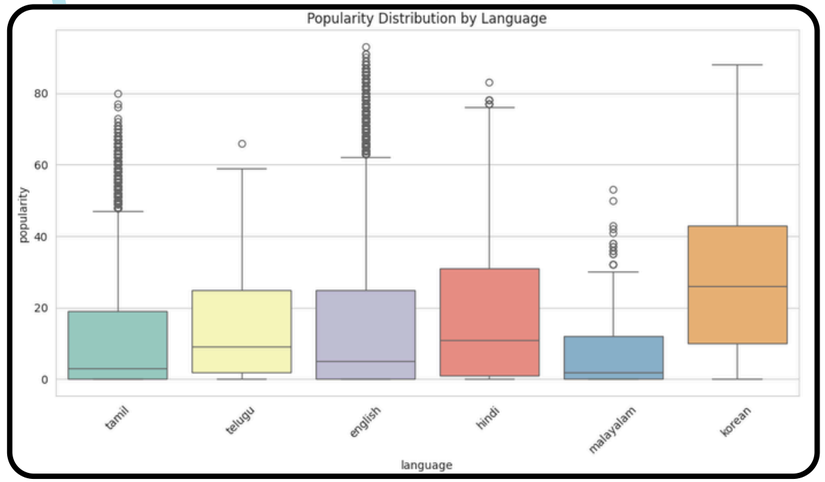
- Numerical vs. Numerical: Scatter plots, Correlation heatmaps
- Numerical vs. Categorical: Box plots, Violin plots
- Categorical vs. Categorical: Grouped bar analysis

Multivariate Analysis → Explore 3 or more variables together

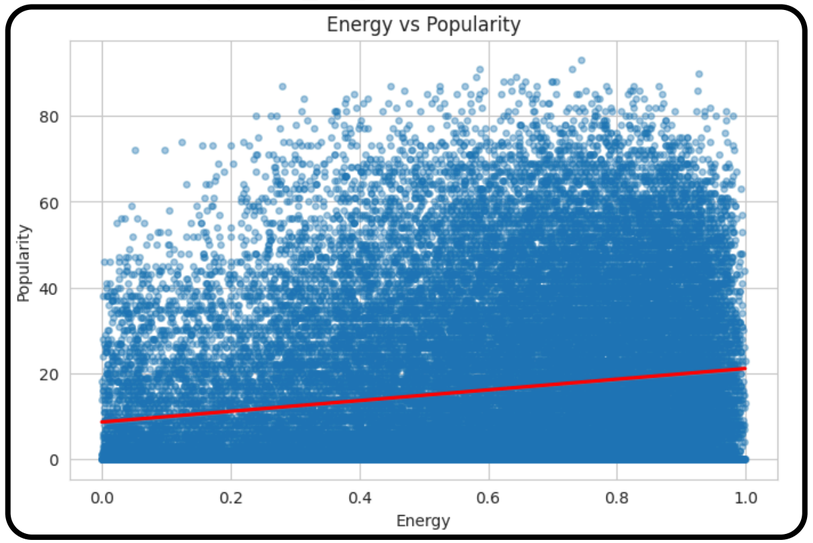
- Use scatter plots with color/hue, pairplots, heatmaps
- Capture complex interactions for deeper insights

Bivariate Analysis

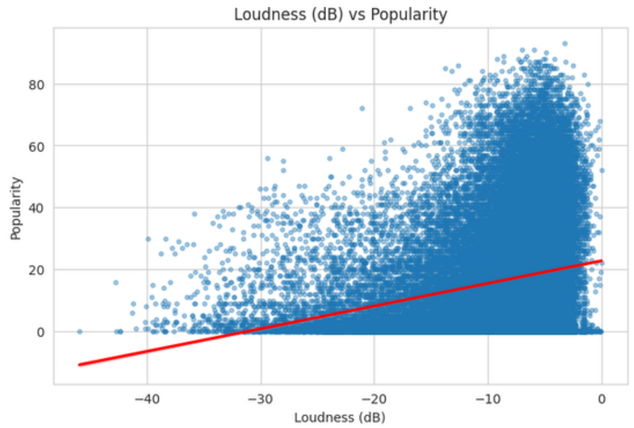
Popularity vs Language



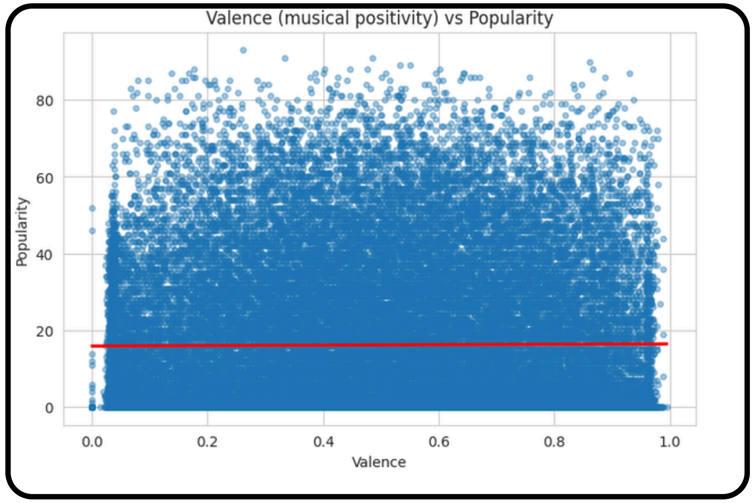
energy vs popularity



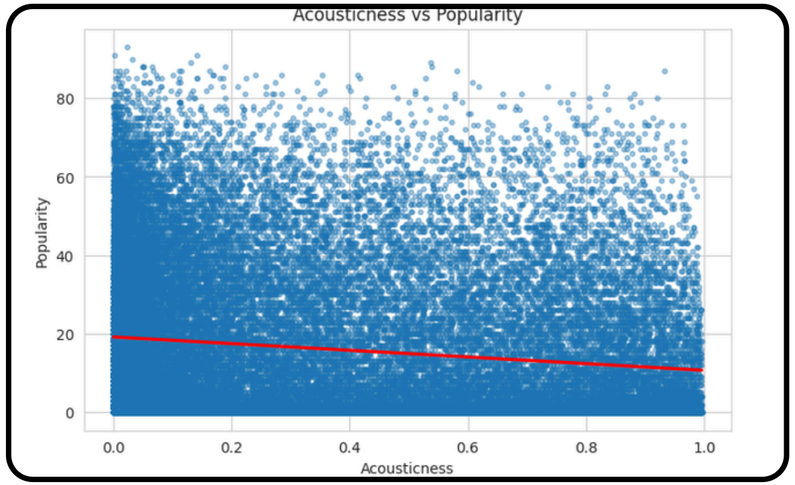
loudness vs popularity



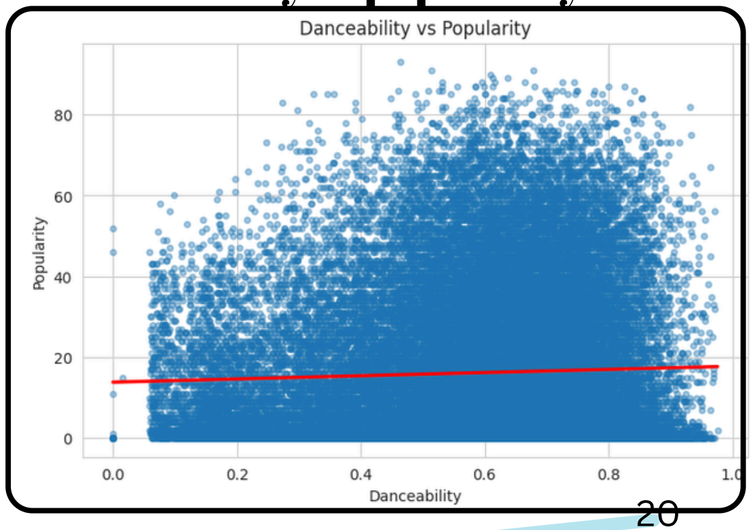
valence vs popularity



acousticness vs popularity



danceability vs popularity



🎯 Insights — Bivariate Analysis (Numerical vs Categorical)

- A positive correlation is often seen — more danceable songs are slightly more popular.

.

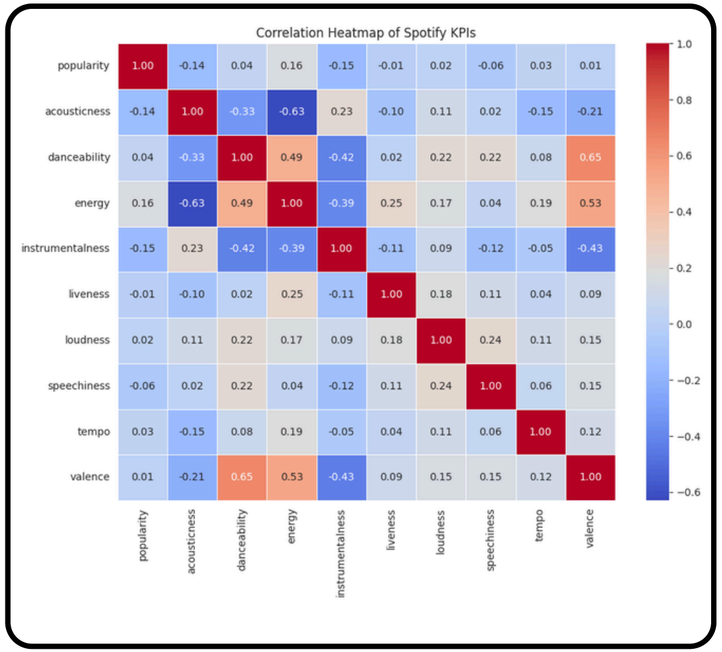
- Songs with medium to high energy levels are more popular.
- Very low-energy tracks (like slow acoustic songs) are less popular.
- There’s usually a positive correlation, showing energetic tracks attract more listeners.
-
- Moderately loud tracks tend to be more popular. Too soft or too loud doesn’t necessarily increase popularity.
- Negative correlation is often observed — as acousticness increases, popularity slightly decreases.
- Meaning, electronic or studio-produced songs are often more popular than pure acoustic tracks. However, niche acoustic genres still perform well in their audience segments. Songs with moderate valence (balanced emotional tone) tend to perform best.

Multivariate Analysis

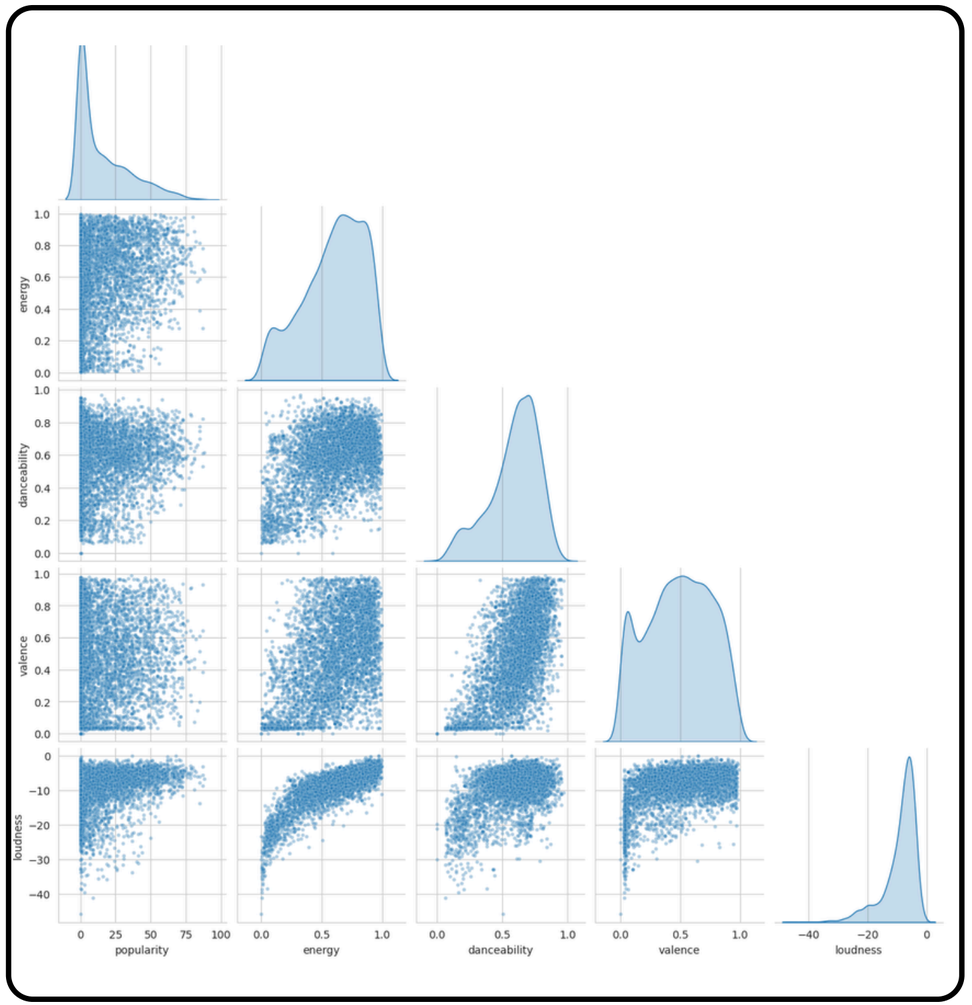
Correlation Heatmap of Spotify KPIs:

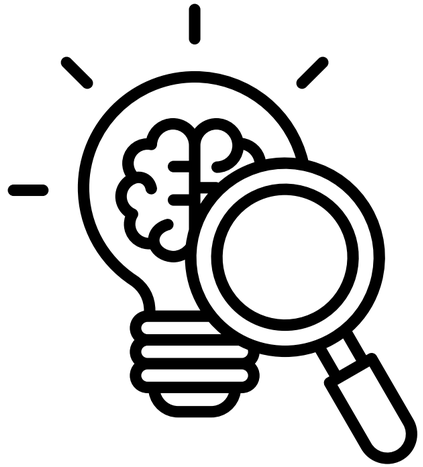
- Danceability is strongly positively correlated with Valence (0.65) and Energy (0.49), showing that upbeat and energetic songs tend to feel happier and more danceable.
- Energy and Loudness are moderately correlated (0.53), meaning more energetic tracks are generally louder, reflecting modern production trends.
- Acousticness is negatively correlated with Energy (-0.63) and Danceability (-0.33), indicating acoustic songs are softer and less rhythmic.
- Instrumentalness has negative correlations with Danceability (-0.42) and Valence (-0.43), suggesting instrumental tracks are usually less lively or happy.
- Popularity shows weak correlations with most features, implying that track success depends on a combination of musical attributes rather than one factor.
- Overall, Energy, Danceability, and Valence form a connected trio, capturing the essence of “happy, loud, and upbeat” music that listeners enjoy.

Heatmap of correlation matrix



Pairplot to visualize pairwise relationships

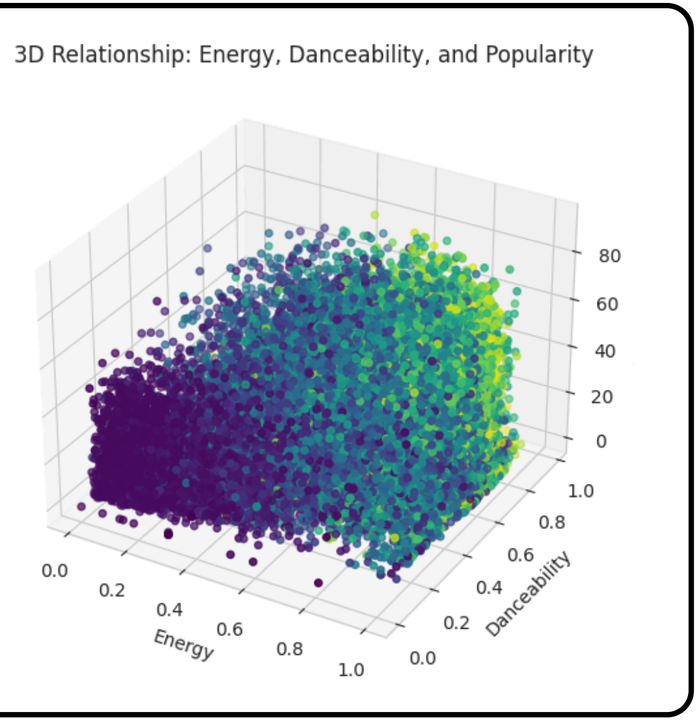




Insights — Pairwise Relationships among Key Variables

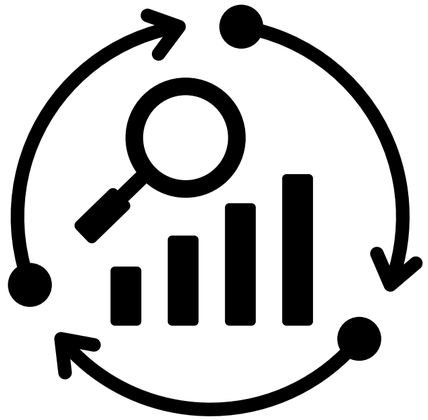
- Energy and Loudness are strongly positively correlated, meaning energetic songs are usually louder.
- Danceability and Valence have a moderate positive correlation, so happier songs tend to be more rhythmic and danceable.
- Acousticness negatively correlates with Energy and Danceability, indicating acoustic tracks are softer and less danceable.
- Instrumentalness is negatively associated with Danceability and Valence, suggesting instrumental tracks are less lively or happy.
- Popularity shows weak correlations with all features, implying that hit tracks are influenced by a combination of factors rather than a single feature.

Multivariate Analysis



Insights — 3D Relationship: Energy, Danceability, and Popularity

- Tracks with higher Energy and Danceability (above 0.5) show a clear rise in Popularity, confirming that lively and rhythmic songs resonate more with listeners.
- Low-popularity songs cluster at lower Energy and Danceability levels, indicating calmer tracks attract smaller audiences.
- Popularity increases sharply when both features rise together, highlighting a combined effect of liveliness and rhythm in boosting listener appeal.
- A few low-energy outliers with moderate popularity suggest certain acoustic or emotional songs can still gain traction.
- Overall, vibrant, energetic, and danceable music has the strongest listener impact.

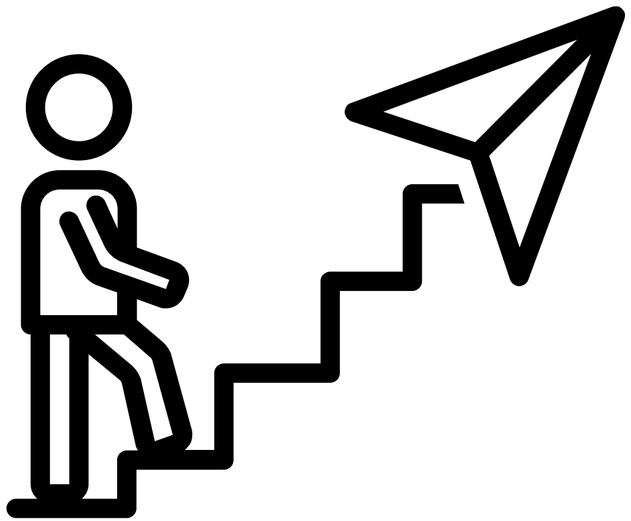


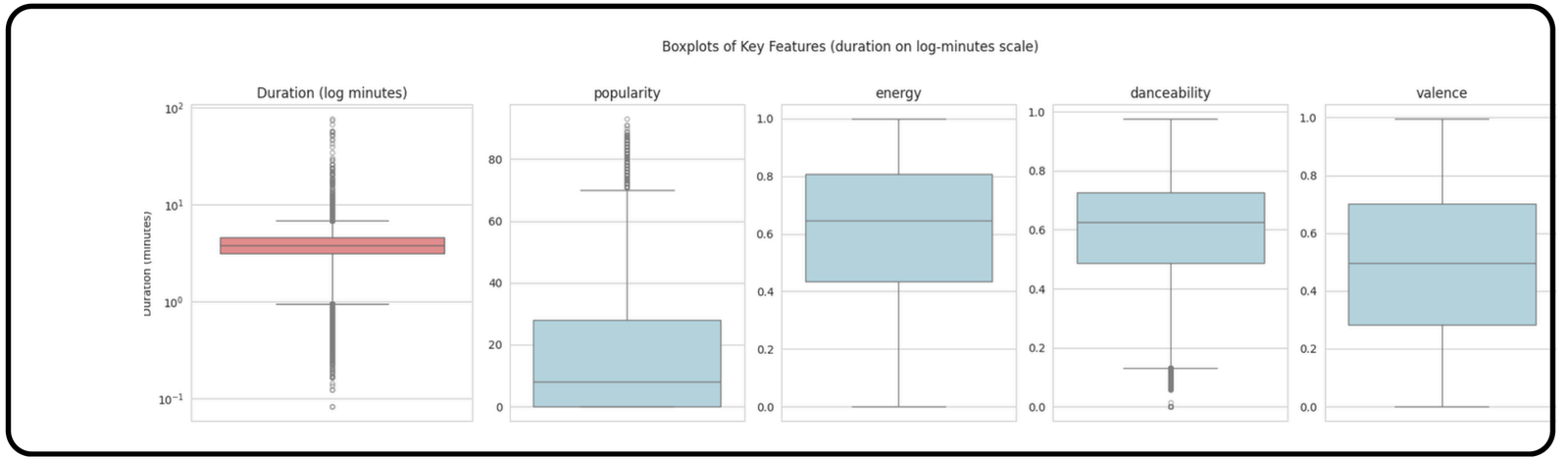
```
Strong Positive Correlations:
 danceability  valence    0.65
energy         valence    0.53
dtype: float64

Strong Negative Correlations:
 acousticness      danceability       -0.33
instrumentalness  energy             -0.39
danceability      instrumentalness   -0.42
instrumentalness  valence            -0.43
energy            acousticness       -0.63
dtype: float64
```

Insights — Multivariate Analysis

• Popularity shows strong positive correlations with energy, danceability, and loudness — suggesting energetic, loud, and rhythm-driven songs tend to be more popular.
• Acousticness and energy are negatively correlated, showing that acoustic songs tend to be softer and less intense.
• Valence (positivity) aligns moderately with danceability and energy, confirming happier songs often sound more upbeat.
• Loudness and energy have one of the strongest correlations (~0.8+), typical of modern production styles.
• The 3D scatter reveals clear clustering — high-popularity tracks often combine high energy, moderate danceability, and positive valence.

Outlier Analysis






```
num_cols = ['popularity', 'duration_ms', 'energy', 'danceability', 'valence', 'loudness']

df_outlier = df[num_cols].copy()
df_outlier.describe()
```

	popularity	duration_ms	energy	danceability	valence	loudness
count	49225.000000	4.922500e+04	49190.000000	49190.000000	49190.000000	49190.000000
mean	16.176069	2.334709e+05	0.602340	0.591554	0.487381	-8.986342
std	19.077407	1.025205e+05	0.249392	0.186258	0.264933	5.639497
min	0.000000	5.000000e+03	0.000232	0.000000	0.000000	-45.920000
25%	0.000000	1.888670e+05	0.434000	0.488000	0.282000	-10.807000
50%	8.000000	2.287860e+05	0.645000	0.627000	0.497000	-7.310000
75%	28.000000	2.767060e+05	0.807000	0.726000	0.702000	-5.289000
max	93.000000	4.581483e+06	0.999000	0.976000	0.995000	-0.005000

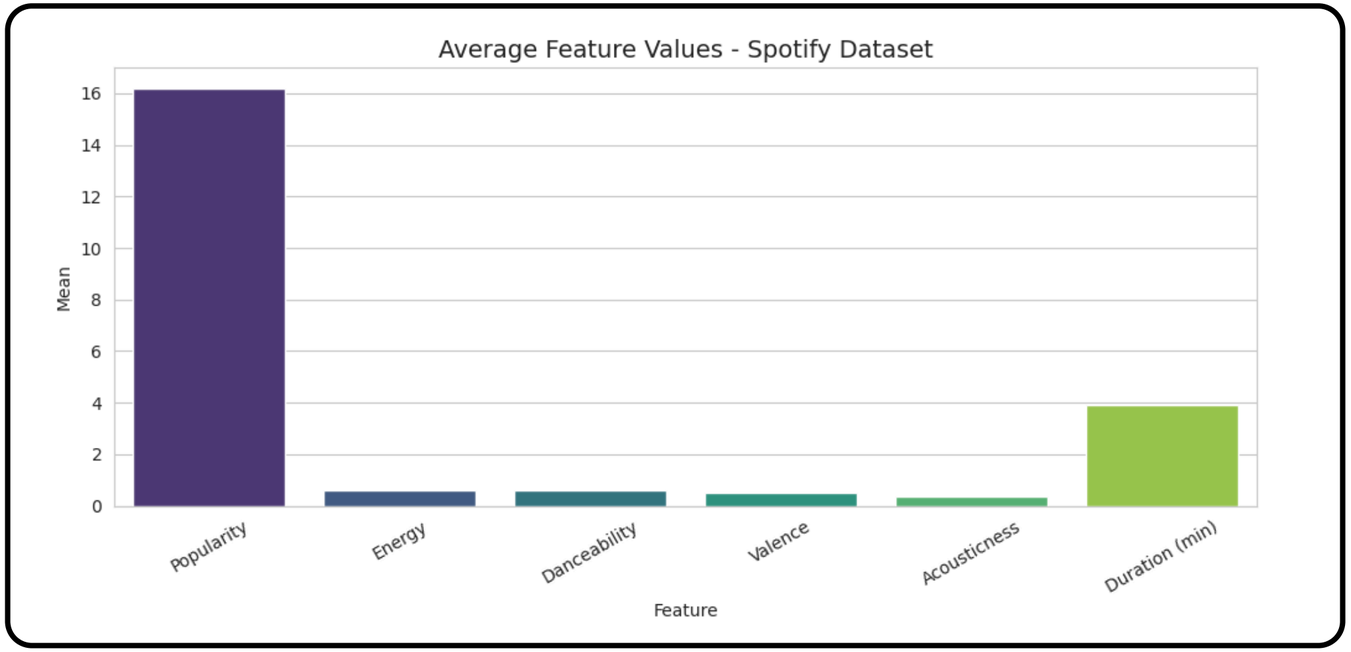
Insights — Outlier Analysis

- Several extreme values were detected across key features such as Duration, Loudness, and Popularity.
- Duration outliers (very long tracks) likely represent podcasts, live recordings, or extended mixes, not standard songs.
- Loudness outliers on the quieter end suggest poorly normalized or acoustic tracks; extremely loud ones indicate aggressive mastering.
- Popularity outliers (scores >80) are rare — only a small fraction of tracks achieve high audience engagement.
- Energy and Danceability have few outliers, showing consistent production patterns across most songs.
- Outliers should be handled contextually — remove only when they distort analysis; otherwise, they may highlight special cases worth studying.

Final Insights and Recommendations

Key Feature Summary Statistics

	Feature	Mean	Median
0	Popularity	16.176	8.000
1	Energy	0.602	0.645
2	Danceability	0.592	0.627
3	Valence	0.487	0.497
4	Acousticness	0.353	0.261
5	Duration (min)	3.891	3.813



Final Insights and Recommendations

Overall Summary of Findings

- Popularity Distribution: Most tracks have low popularity scores (below 30), showing that Spotify's catalog is dominated by less-known songs. Only a small fraction go viral (above 60+), forming a right-skewed distribution where hits are rare.
- Duration Patterns: The average duration is around 3–4 minutes, aligning with commercial music standards. Outliers (very short or long tracks) likely represent podcasts, interludes, or extended instrumentals.
- Energy, Danceability & Valence:
- Songs cluster around moderate to high Energy (0.6–0.8), reflecting the dominance of upbeat compositions.
- Danceability follows a similar pattern, showing a bias toward rhythmic, movement-friendly tracks.
- Valence (positivity) varies widely, suggesting Spotify hosts an emotional balance of melancholic and feel-good music.
- Acousticness & Instrumentalness:
- The majority are non-acoustic, highlighting the prevalence of electronic and digitally produced tracks.
- Instrumental songs form a smaller share, consistent with mainstream preference for vocals.
- Key and Mode: Tracks are fairly evenly spread across musical keys. Major mode songs slightly outnumber minor mode, reflecting a listener bias toward brighter, more uplifting tonalities.
- Language Distribution: English dominates the catalog, followed by Tamil, Korean, and Hindi, showing global diversity but a continued Western emphasis.
- Temporal Trends: Recent years reflect a surge in track releases, driven by the rise of streaming platforms and easier music publishing.

Data-Driven Insights & Recommendations

1. For Artists & Producers:

- Create tracks with high Energy, moderate Danceability, and positive Valence to maximize listener appeal.
- Keep song length within 3–4 minutes for playlist compatibility and engagement.
- Explore English or bilingual tracks (English + local language) to reach global and regional audiences.

2. For Spotify's Curation Team:

- Personalize playlists using energy and mood dimensions (e.g., High-Energy Workout, Low-Valence Chill).
- Highlight non-English tracks with high engagement to strengthen global music discovery.

3.For Data & Business Teams:

- Develop popularity prediction models using core features like Energy, Danceability, Valence, and Duration.
- Track temporal shifts in song attributes to anticipate evolving listener preferences.

4. For Future Research:

- Examine how release year influences features (e.g., are modern songs shorter, louder, and more energetic?).
- Perform genre-level analysis (once genre data is included) to capture cultural and regional music trends.

Future Scope / Limitations

Limit the Dataset and Analysis

1. Incomplete Metadata: The dataset lacks contextual variables such as genre, playlist placement, or artist growth metrics, all of which strongly influence track success.
2. Dynamic Popularity Scores: Spotify's popularity score is time-sensitive and region-dependent. Current values reflect recent listening behavior but may not capture long-term trends.
3. No Listener Demographics: Absence of information on age, region, and user behavior limits the ability to link track features to audience preferences.
4. Static Snapshot: The dataset represents a single time frame rather than continuous updates, restricting analysis of temporal evolution in music trends.
5. Imbalanced Data Representation: Certain languages and styles (e.g., Malayalam, Telugu) have very few entries, which can distort comparisons and statistical reliability.
6. Correlation ≠ Causation: Observed relationships (e.g., high-energy tracks being more popular) do not imply causality. External factors such as marketing, social media, or artist reputation may play a larger role.

Future Scope for Enhancement

1. Integrate Genre and Artist Data:
   Add genre and artist-level popularity metrics to study how different music types and artist reputations impact track success.
2. Time-Series Popularity Analysis:
   Track popularity changes over months/years to see how long hits remain relevant, or how trends like "lo-fi beats" rise and fade.
3. Genre-Specific Deep Dives:
   Conduct targeted analyses within specific genres (e.g., Hip-Hop, Indie, Classical) to uncover how characteristics vary across musical styles.
4. Sentiment & Lyrics Analysis:
   Merge with lyrics data to study how emotional tone in lyrics correlates with audio features and popularity.

Spotify Tracks

✨ In Summary:

This analysis provides strong insights into track attributes and popularity, but its scope is limited by missing contextual and behavioral data. Future progress lies in multimodal music analytics — integrating sound, sentiment, audience, and cultural context to reveal deeper intelligence about how music connects with listeners. 🎶

🏁 Conclusion

The Spotify dataset highlights how modern music reflects today's culture — upbeat, emotionally diverse, and digitally shaped. By uncovering these patterns, both artists and data professionals can not only optimize listener engagement but also help shape the future of sound


Thank
you