SPOTIFY DATASET ANALYSIS

AGENDA

DATASET DESCRIPTION

DETAILED ANALYSIS

INSIGHTS AND OBSERVATIONS

CONCLUSION

OBJECTIVE:

- To explore a large **Spotify track** dataset using data analysis.
- To understand how **audio features** (danceability, energy, tempo, etc.) are distributed.
- To study how these features related to **track popularity**.
- To identify patterns and insights useful for **music recommendation / understanding trends**.

DATASET DESCRIPTION

- Total tracks (rows): **62,317**
- Total attributes (columns): **22**

- Contains:
  - Track & artist information
  - Audio feature scores (0–1 scale mostly)
  - Popularity score (0–100)
  - Year & language

BASIC UNDERSTANDING OF THE DATASET

🎼 Categorical Columns:

- **track_id** – Unique identifier for each track in Spotify.
- **track_name** – Name/title of the song.
- **artist_name** – Name of the performing artist(s).
- **album_name** – Album to which the track belongs.
- **track_url** – Direct URL link to play the track on Spotify.
- **artwork_url** – URL for the album or track cover art.
- **year** – Release year of the track.
- **language** – Primary language of the track's lyrics

🎧 Audio & Numeric features:

- **popularity** – Integer score (0–100) representing track audience engagement & streams.
- **danceability** – How suitable a track is for dancing (-1 to 1) based on rhythm & beat stability.
- **energy** – Overall intensity and activity of the track (-1 to 1).
- **acousticness** – Probability that a track is acoustic, i.e., not electronic or heavily produced.
- **instrumentalness** – Likelihood that a track has no vocals (-1 to 1).
- **loudness** – Decibel measure of track loudness (range ~ −60 to 0).
- **speechiness** – Presence of spoken words in the track (speech-like content).
- **liveness** – Detects the presence of audience in the recording. (-1 to 1)
- **valence** – Positivity of the musical mood (happy vs sad sounding).
- **tempo** – Beats per minute (BPM), rhythm pace of the song.
- **duration_ms** – Length of the track in milliseconds.
- **key** – The key the track is in . .
- **mode** – Musical mode: major (1) or minor (0).
- **time_signature** – Estimated beats per bar.

DEEPER LOOK AT THE DATASET

Initial shape of the dataset (rows x columns)

```
Shape of dataset: (62317, 22)
```

Removing the duplicates based on track_id

```
Before removing duplicates: 62317
After removing duplicates: 62239
Duplicates removed: 78
```

Columns and their datatypes

```
Data types:
track_id             object
track_name           object
artist_name          object
year                  int64
popularity            int64
artwork_url          object
album_name           object
acousticness        float64
danceability        float64
duration_ms         float64
energy              float64
instrumentalness    float64
key                 float64
liveness            float64
loudness            float64
mode                float64
speechiness         float64
tempo               float64
time_signature      float64
valence             float64
track_url            object
language             object
dtype: object
```

22 columns:
13 floats
2 integers
7 objects

First 5 rows of the dataset (head)

First 5 rows:

	track_id	track_name	artist_name	year	popularity	artwork_url	album_name	acousticness	danceability	duration_ms	...	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence	track_url	language
0	2r0ROhr7pRN4MXDMT1fEmd	Leo Das Entry (From "Leo")	Anirudh Ravichander	2024	59	https://i.scdn.co/image/ab67616d0000b273ce9c65...	Leo Das Entry (From "Leo")	0.0241	0.753	97297.0	...	8.0	0.1000	-5.994	0.0	0.1030	110.997	4.0	0.459	https://open.spotify.com/track/2r0ROhr7pRN4MXD...	Tamil
1	4I38e6Dg52a2o2a8i5Q5PW	AAO KILLELLE	Anirudh Ravichander, Pravin Mani, Vaishali Sri...	2024	47	https://i.scdn.co/image/ab67616d0000b273be1b03...	AAO KILLELLE	0.0851	0.780	207369.0	...	10.0	0.0951	-5.674	0.0	0.0952	164.995	3.0	0.821	https://open.spotify.com/track/4I38e6Dg52a2o2a...	Tamil
2	59NoiRhnom3lTeRFaBzOev	Mayakiriye Sirikiriye - Orchestral EDM	Anirudh Ravichander, Anivee, Alvin Bruno	2024	35	https://i.scdn.co/image/ab67616d0000b27334a1dd...	Mayakiriye Sirikiriye (Orchestral EDM)	0.0311	0.457	82551.0	...	2.0	0.0831	-8.937	0.0	0.1530	169.996	4.0	0.598	https://open.spotify.com/track/59NoiRhnom3lTeR...	Tamil
3	5uUqRQd385pvLxC8JX3tXn	Scene Ah Scene Ah - Experimental EDM Mix	Anirudh Ravichander, Bharath Sankar, Kabilan, ...	2024	24	https://i.scdn.co/image/ab67616d0000b27332e623...	Scene Ah Scene Ah (Experimental EDM Mix)	0.2270	0.718	115831.0	...	7.0	0.1240	-11.104	1.0	0.4450	169.996	4.0	0.362	https://open.spotify.com/track/5uUqRQd385pvLxC...	Tamil
4	1KaBRg2xgNeCljmyxBH1mo	Gundellonaa X I Am A Disco Dancer - Mashup	Anirudh Ravichander, Benny Dayal, Leon James, ...	2024	22	https://i.scdn.co/image/ab67616d0000b2735a59b6...	Gundellonaa X I Am a Disco Dancer (Mashup)	0.0153	0.689	129621.0	...	7.0	0.3450	-9.637	1.0	0.1580	128.961	4.0	0.593	https://open.spotify.com/track/1KaBRg2xgNeCljm...	Tamil

5 rows × 22 columns

Univariate Analysis

What is Univariate Analysis?

Univariate analysis examines a single variable at a time to understand its individual distribution, range, and statistical behavior.

In the upcoming slides we will:

- ❑ study how each variable behaves independently using descriptive statistics.
- ❑ visualize numerical features using histograms, boxplots, and distribution graphs.
- ❑ interpret each feature’s spread using metrics like mean, median, quartiles, and range.
- ❑ identify skewness, outliers, and concentration ranges for each

Statistical Data…

```
Statistical Summary : popularity

count    62317.000000
mean        15.358361
std         18.626908
min          0.000000
25%          0.000000
50%          7.000000
75%         26.000000
max         93.000000
Name: popularity, dtype: float64
```

```
Statistical Summary : valence

count    62317.000000
mean         0.495226
std          0.264787
min         -1.000000
25%          0.292000
50%          0.507000
75%          0.710000
max          0.995000
Name: valence, dtype: float64
```

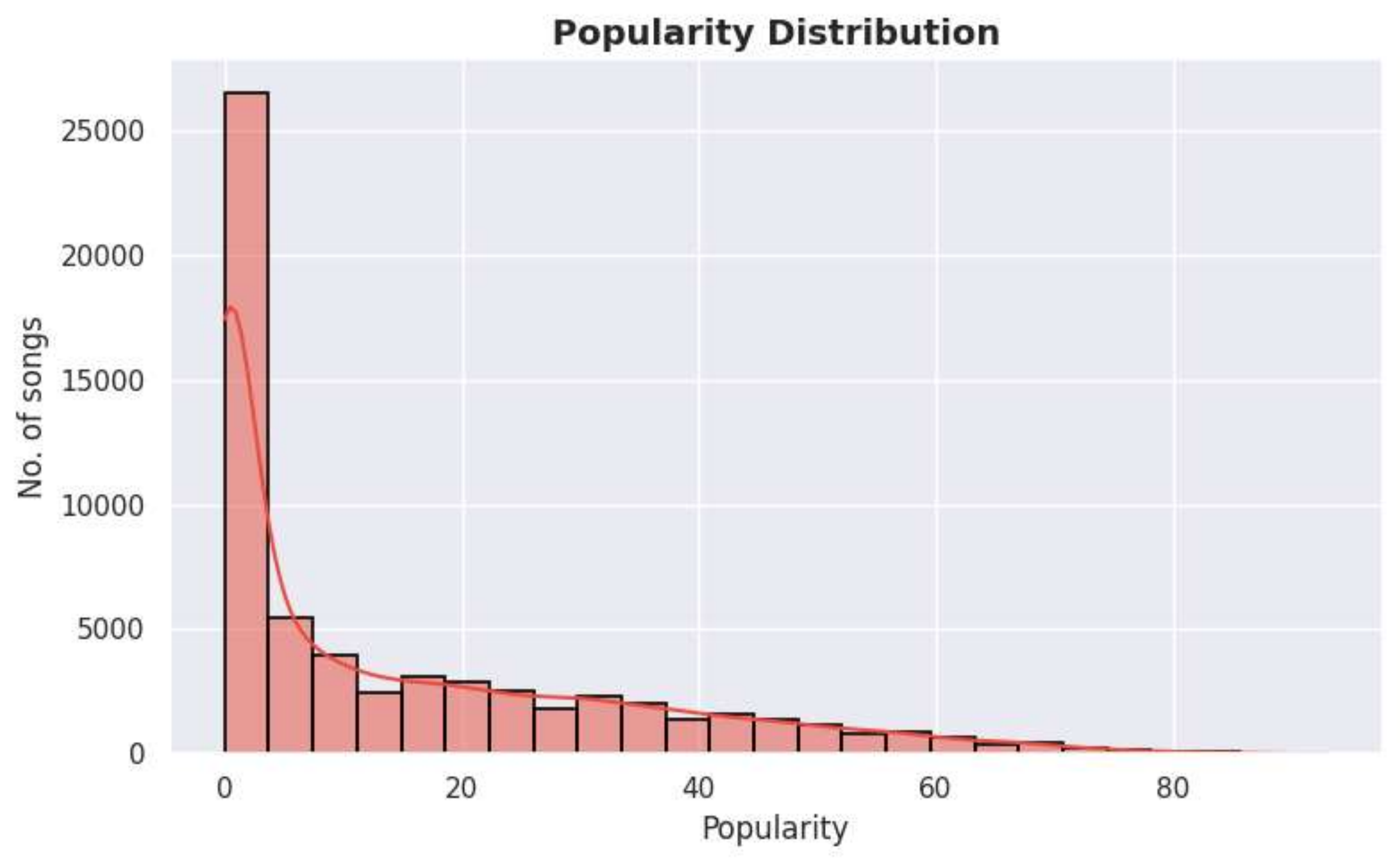
```
Statistical Summary : tempo

count    62317.000000
mean       117.931247
std         28.509459
min         -1.000000
25%         95.942000
50%        117.991000
75%        135.081000
max        239.970000
Name: tempo, dtype: float64
```

```
Statistical Summary : energy

count    62317.000000
mean         0.602496
std          0.246144
min         -1.000000
25%          0.440000
50%          0.639000
75%          0.803000
max          1.000000
Name: energy, dtype: float64
```

```
Statistical Summary : danceability

count    62317.000000
mean         0.596807
std          0.186209
min         -1.000000
25%          0.497000
50%          0.631000
75%          0.730000
max          0.986000
Name: danceability, dtype: float64
```

Lets try to understand the data more deeply by visualisation →___

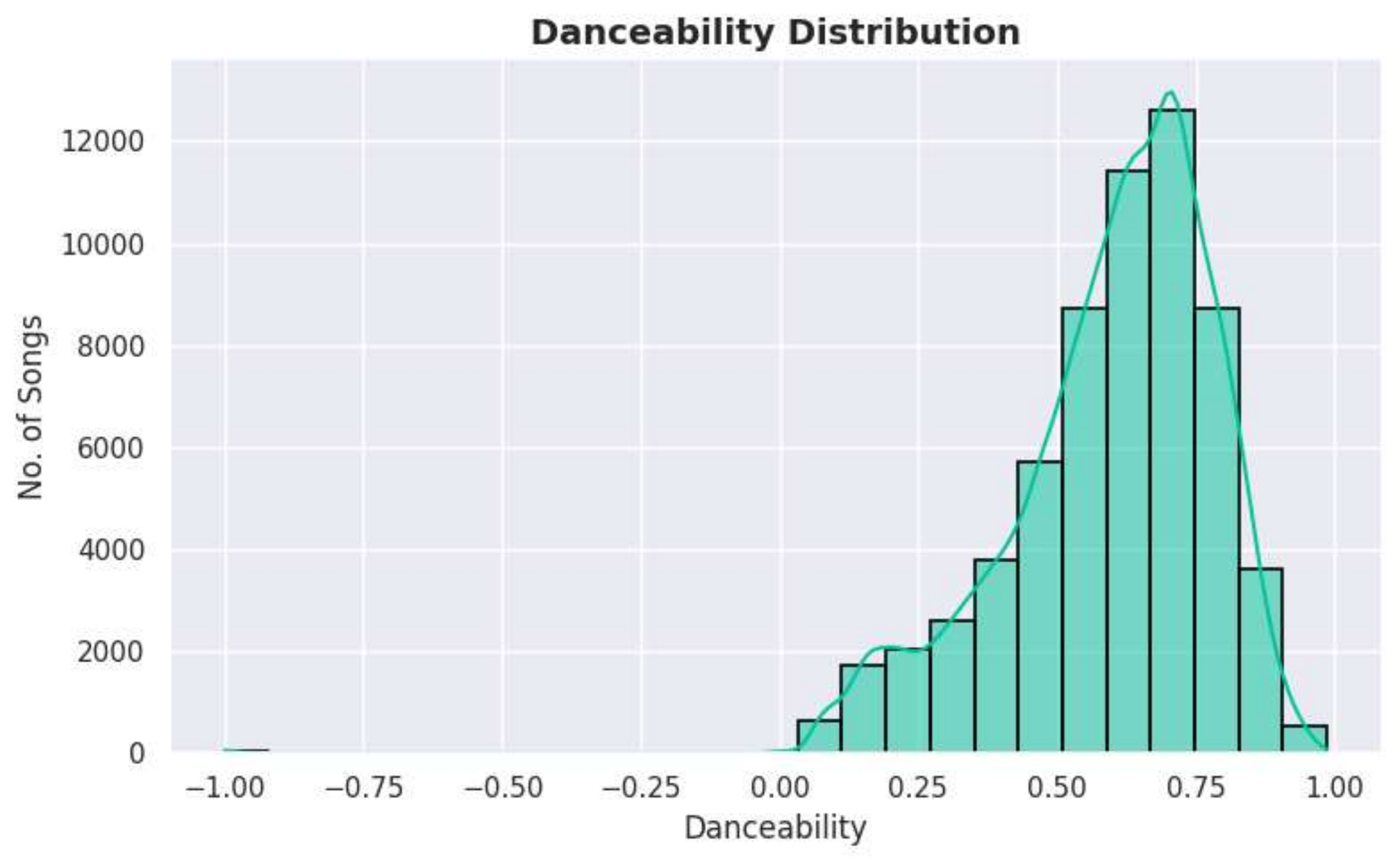
Popularity Scores



- The distribution is heavily skewed, meaning only a small portion of songs achieve very high popularity.
- Median popularity is low, which suggests that the majority of tracks receive limited engagement.
- This shows that on Spotify, a few viral songs dominate listener attention.

Danceability

Danceability reflects how suitable a track is for dancing — based on rhythm stability & beat strength



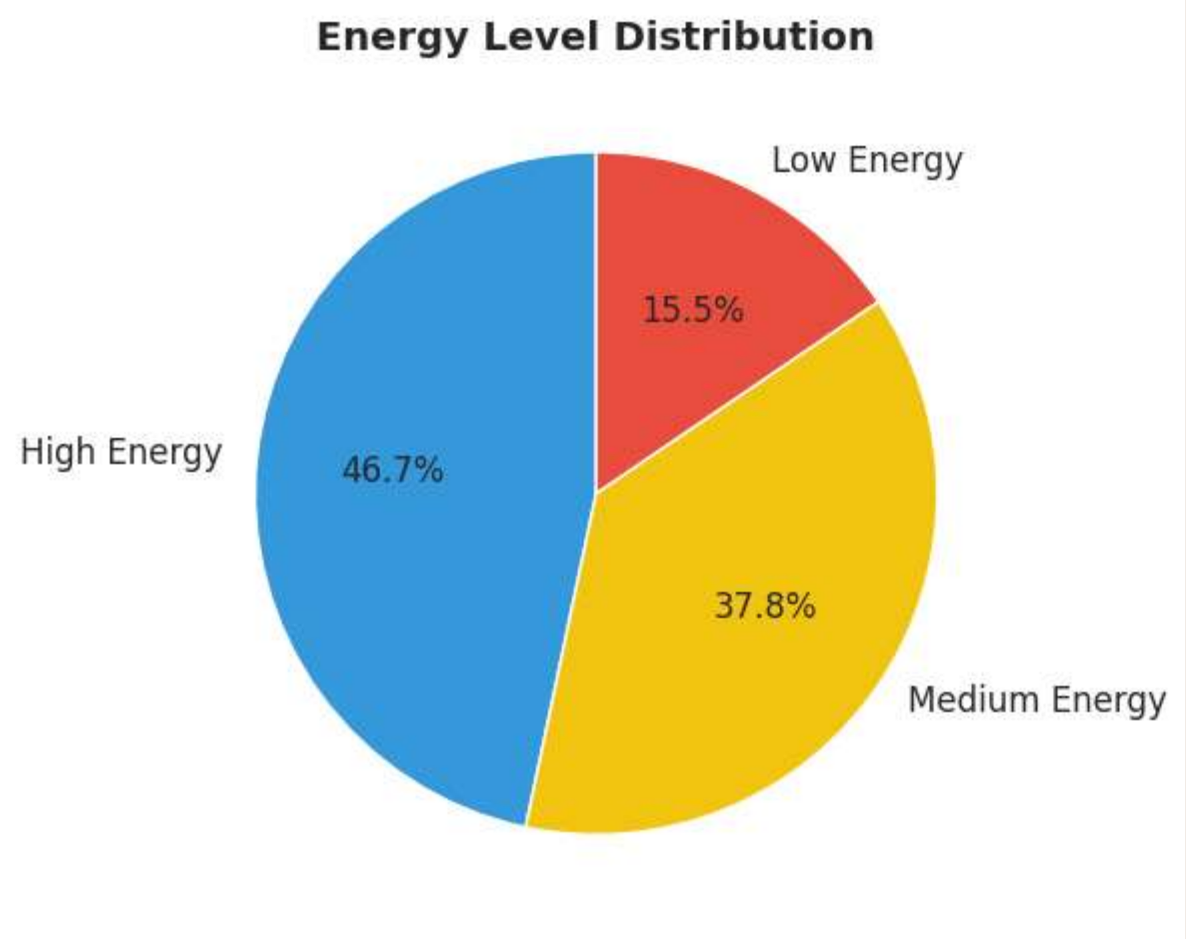
Most songs have **moderate-to-high** danceability **(0.5–0.8)**, showing Spotify's library favors rhythm-driven and upbeat tracks

Very **few tracks** have extremely low danceability **(< 0.3),** meaning most songs possess some rhythmic flow or beat alignment.

The peak density around **0.6–0.7** suggests Spotify's catalog leans toward tracks suitable **for casual or upbeat listening** rather than purely acoustic or spoken content.

Energy Levels

Energy measures intensity and activity — higher values sound stronger and more powerful



The energy level ranges are as follows :
Low Energy : 0 – 0.33
Medium Energy : 0.34 – 0.66
High Energy : 0.67 – 1.00

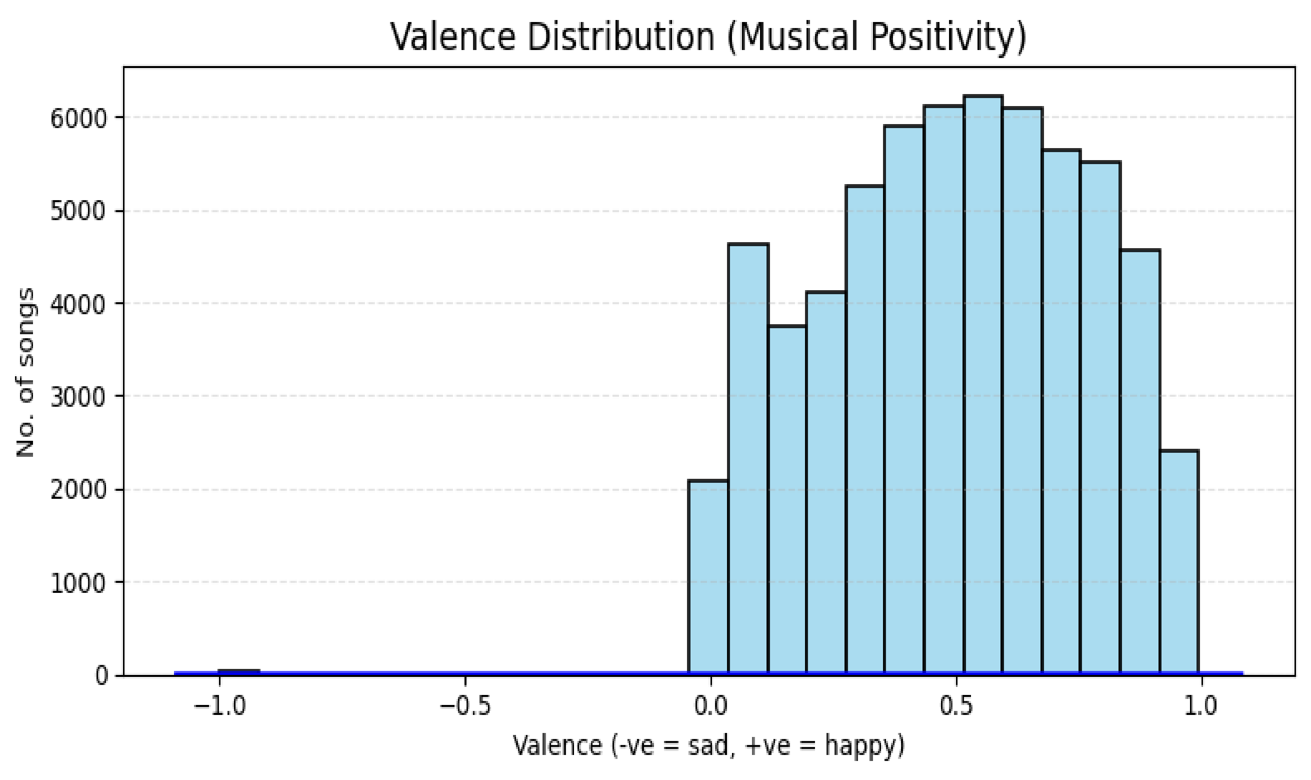
High Energy (0.67–1.00) dominates, comprising the majority of tracks, suggesting Spotify has a strong focus on balanced, moderately intense songs suitable for casual listening.

Low Energy (0.00–0.33) is the smallest slice, indicating that very few tracks are calm, acoustic, or mellow.

- Most Spotify tracks have **high energy**, with fewer high-energy songs and very few low-energy tracks, reflecting a bias toward **engaging and upbeat music**.

Valence

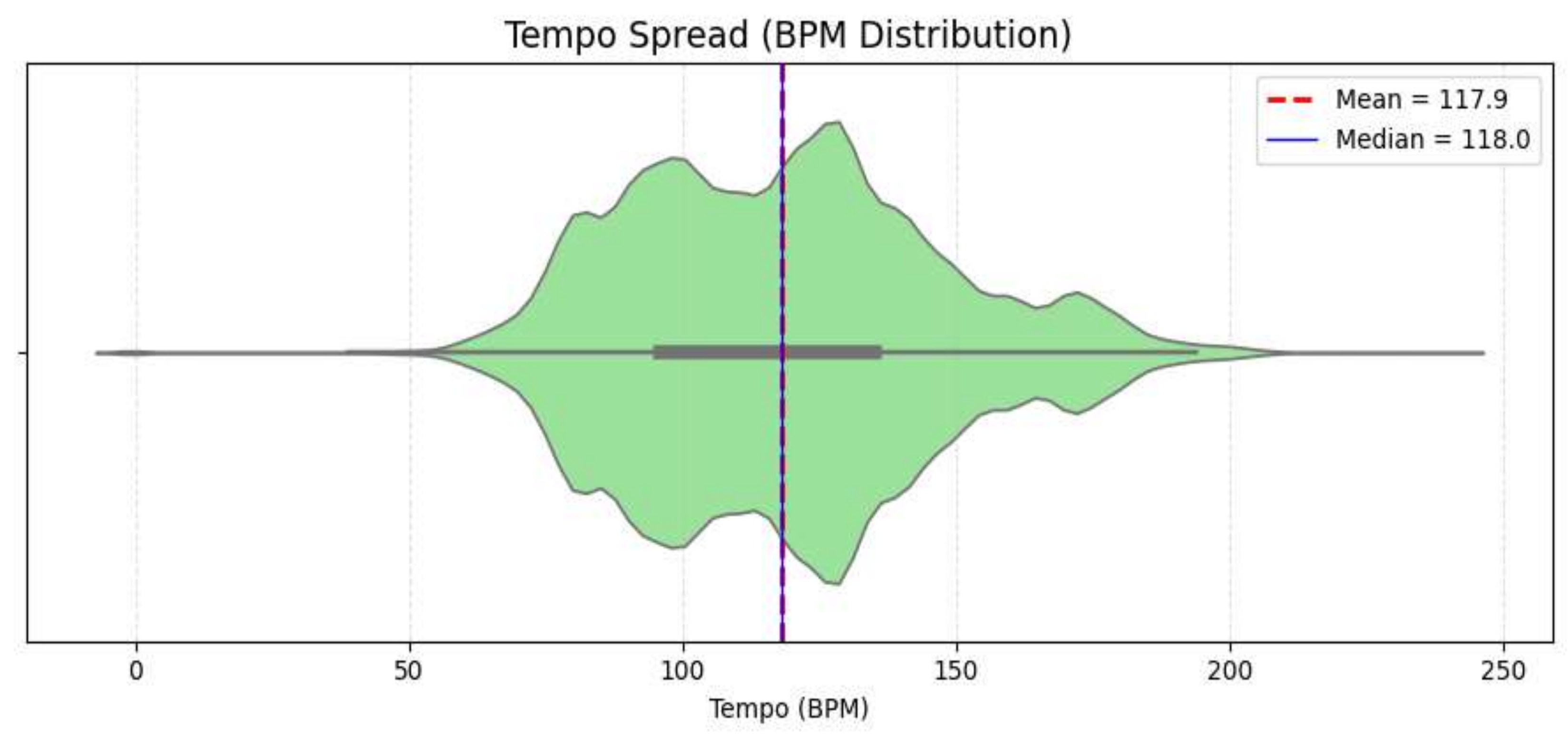
Valence measures how happy or positive the song sounds



- We observe songs spread mainly across 0 to 1
- Very negligible number of songs have valence point around –1
- This variety suggests that the listeners prefer to listen to jolly type of songs rather than sad ones

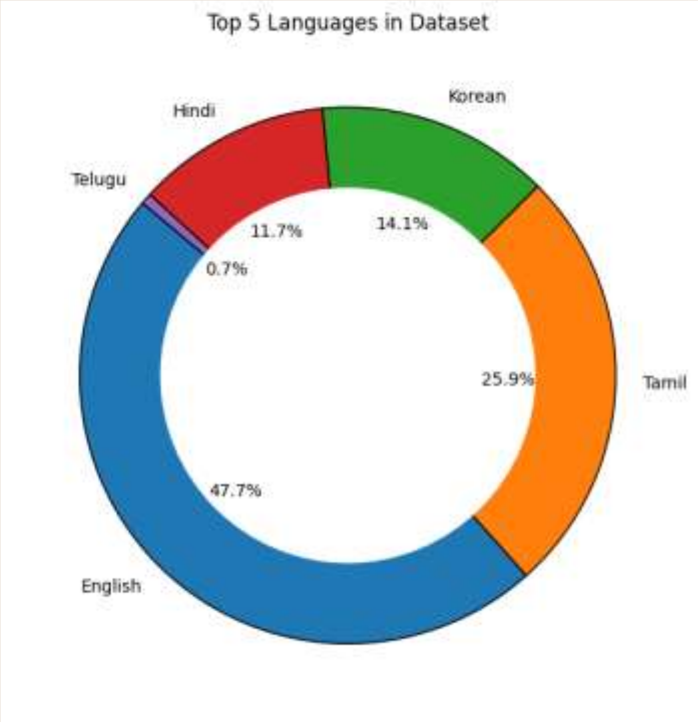
Tempo Distribution

Tempo determines the speed of the song — in beats per minute

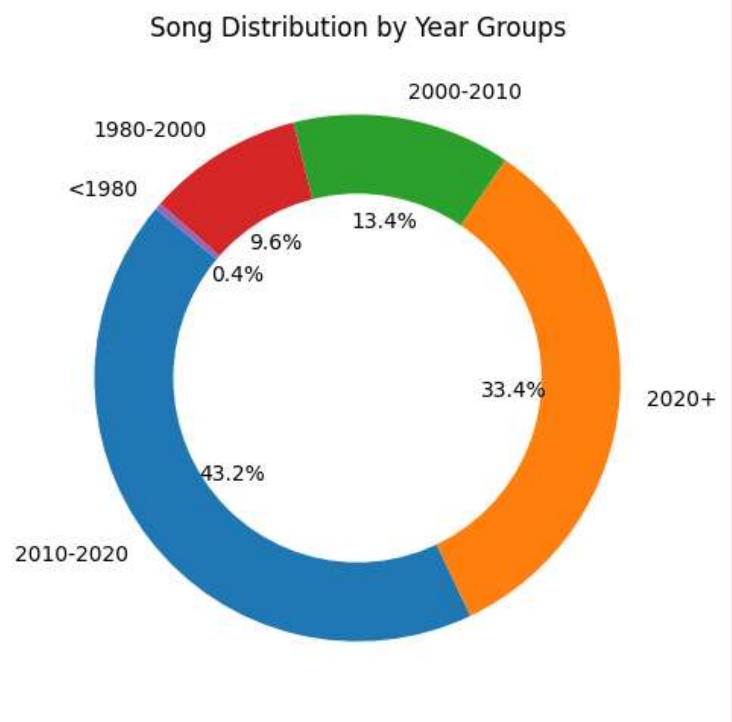


- Most songs cluster between ~100–130 BPM
- This tempo range is known to be naturally engaging to humans
- This aligns with typical pop music structure

Univariate Analysis for Categorical Data



This suggests that most songs available in this dataset are targeted toward a broad mainstream audience of a particular language



This suggests that popularity and audio features observed will be influenced by **modern music styles**, not older eras.

Key Insights

- ✓ **Popularity is heavily right–skewed**, meaning only a small percentage of songs achieve high listener engagement.
- ✓ **Danceability is moderately high for most tracks**, indicating that Spotify songs generally have a rhythmic, movement-friendly feel.
- ✓ **Energy levels are high in majority of tracks**, showing preference toward powerful, intense, lively music.
- ✓ **Valence (emotional positivity) mainly lies between 0 to 1**, meaning listeners prefer happy songs rather than sad songs.
- ✓ **Tempo clusters around ~100–130 BPM**, suggesting most tracks follow typical human-preferred rhythm ranges.
- ✓ **Language distribution shows dominance of one main language**, with others forming minor segments — indicating asymmetric linguistic representation.
- ✓ **Year-group distribution shows majority of tracks from modern eras (2010–2020)**, suggesting dataset bias toward recent streaming-era content

Bivariate Analysis

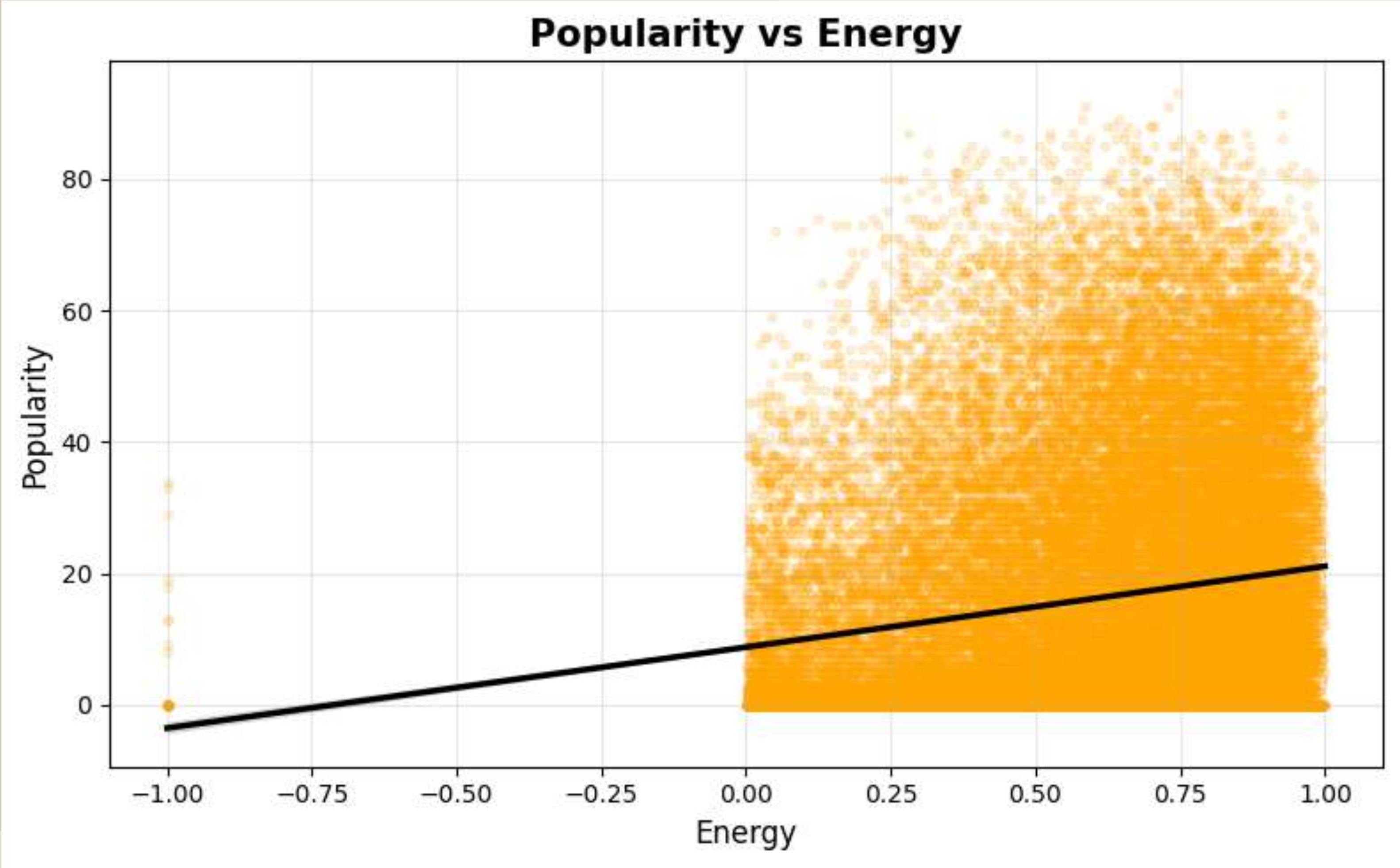
What is Bivariate Analysis?

Bivariate analysis examines the relationship between two variables to understand how changes in one affect the other.

In the upcoming slides we will:

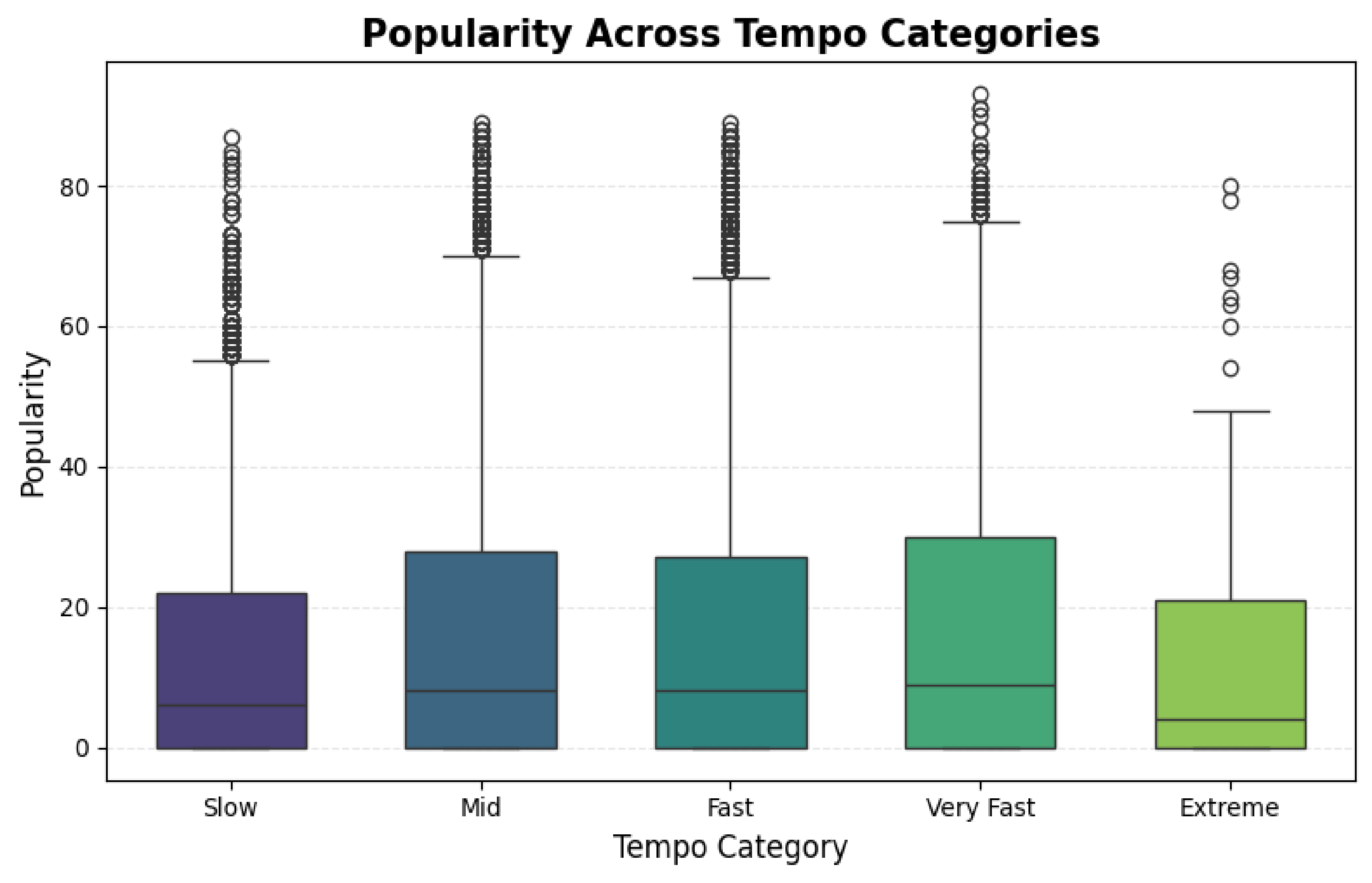
- analyze pairwise relationships between musical features.
- use scatter plots to observe directional trends between variables.
- apply boxplots and grouped comparisons to evaluate category-level patterns.
- see how features like tempo, danceability, and duration interact with popularity.
- interpret how two-variable combinations reveal stronger insights than individual features.

Popularity VS Danceability



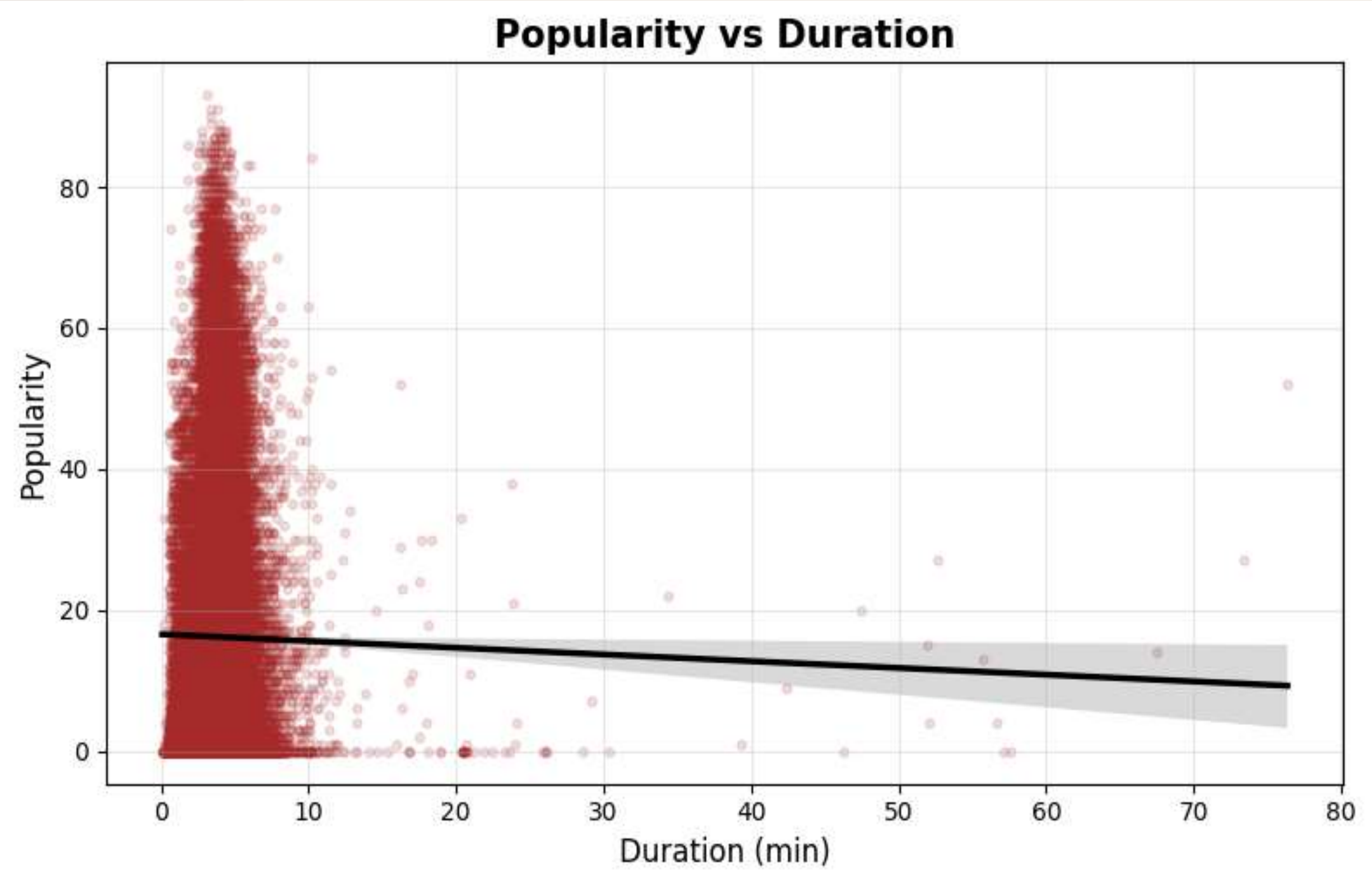
- Higher energy songs generally tend to be more popular
- Low-energy songs are less likely to gain listener attention
- The positive trend line shows a direct relationship

Popularity VS Tempo



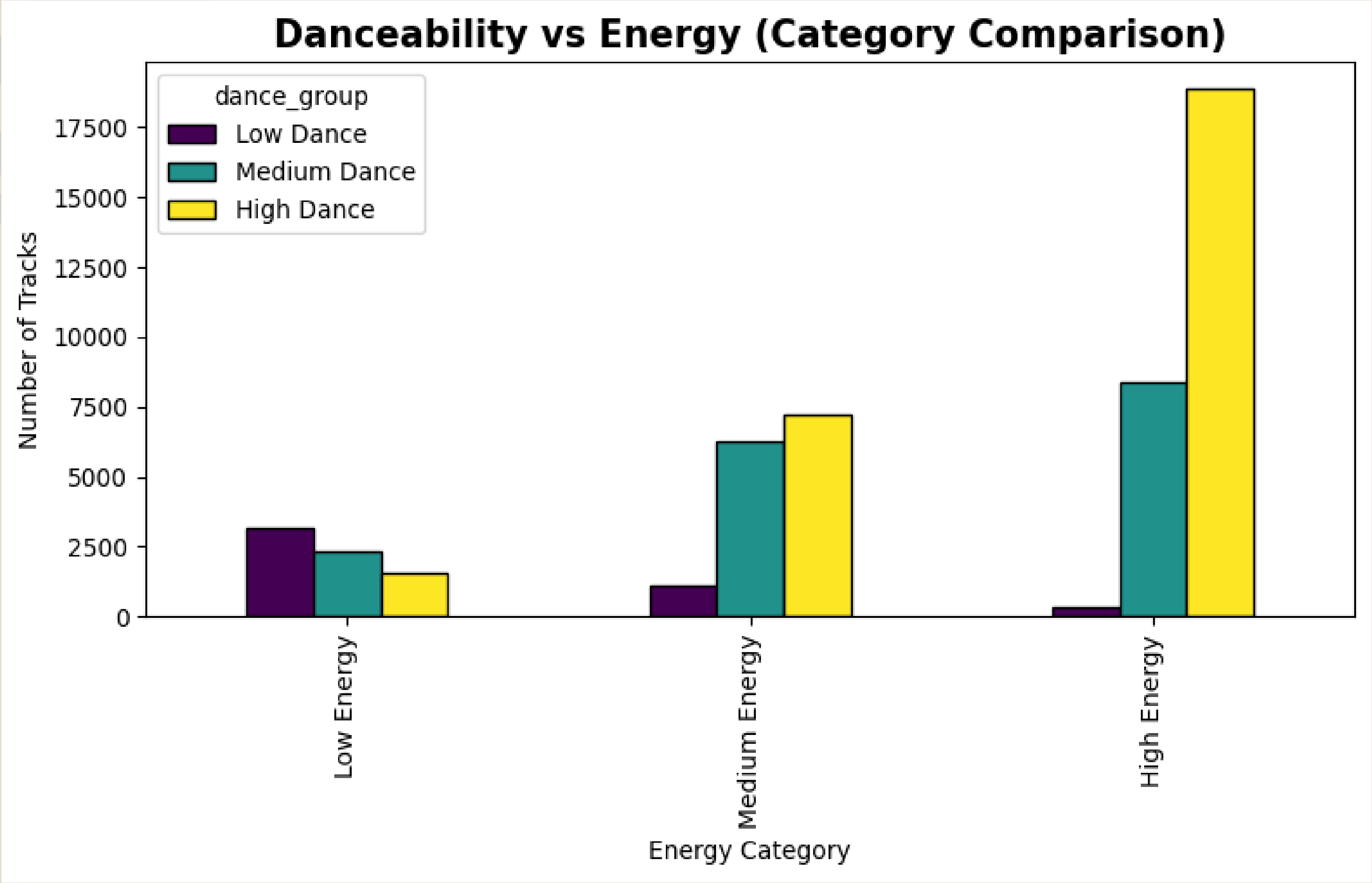
- Popular songs exist across all tempo ranges
- No single tempo category guarantees popularity
- Slightly higher variation in medium–fast BPM

Popularity VS Duration



- Songs with duration between **(~2–8 min)**, especially around **3–4 minutes**, tend to be more popular.
- shorter songs → higher replay value
- playlist-friendly
- streaming algorithm boosts replay count

Danceability vs Energy



- Clear positive trend between energy and danceability, showing that both tend to increase together.
- More energetic songs generally show higher danceability, meaning they have stronger rhythm suitable for movement.
- Low-energy songs typically have lower danceability, indicating they are calmer and less rhythm-driven.

Key Insights

- ✓ There is a mild positive trend between popularity and danceability, indicating that moderately danceable songs have slightly higher odds of becoming popular.
- ✓ Popularity does not depend strongly on tempo, with successful songs existing across slow, medium, and fast BPM categories.
- ✓ Popular songs are mostly within a typical duration range (about 2–4 minutes), while extremely long or short tracks are less likely to be hits.
- ✓ Energy and danceability show a positive relationship, meaning that songs with strong rhythmic energy tend to also be more dance-friendly

Multivariate Analysis

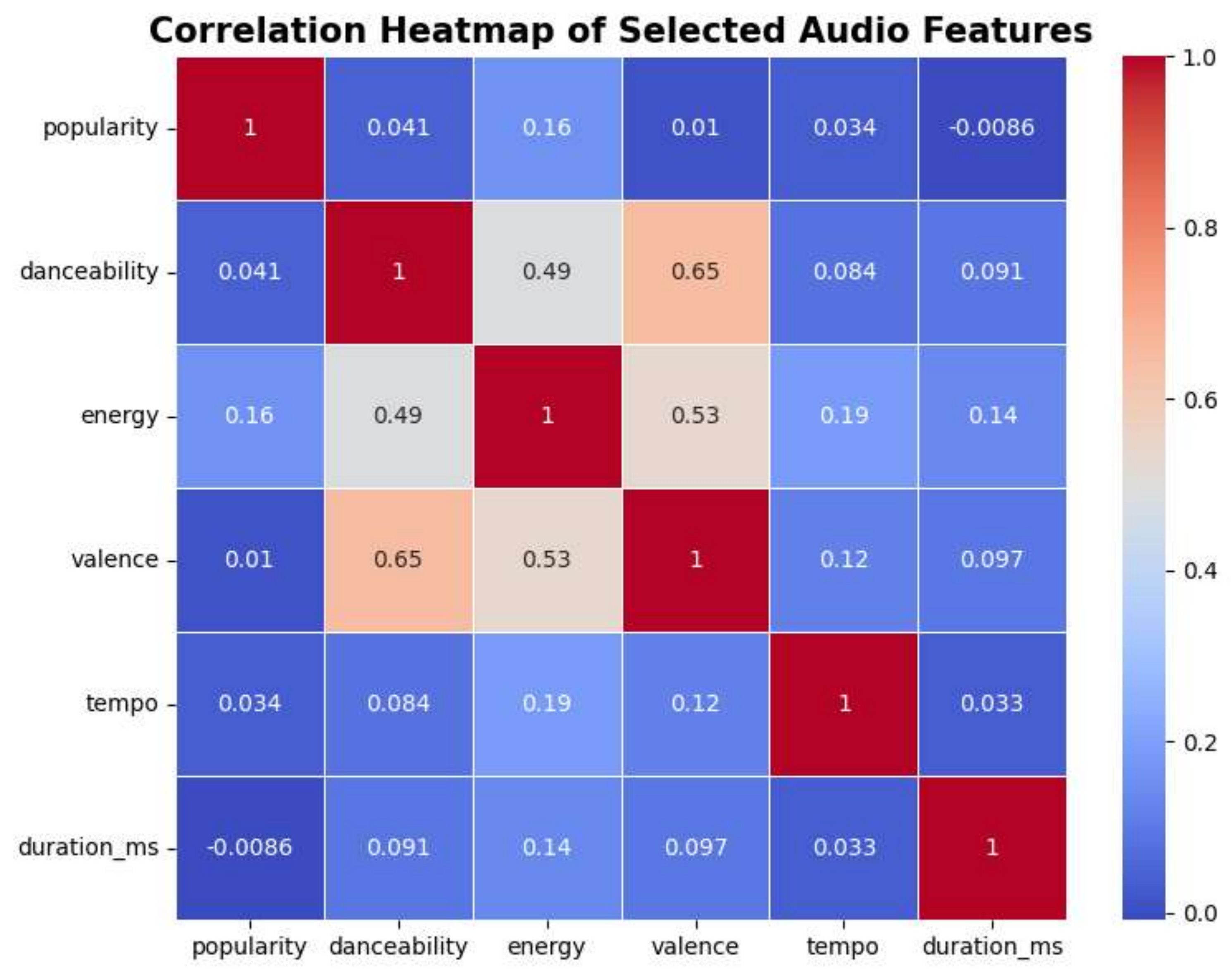
What is Multivariate Analysis?

Multivariate analysis examines relationships among three or more variables simultaneously to understand how multiple features together influence outcomes.

In the upcoming slides we will:

- ❑ analyze combined relationships among multiple audio features.
- ❑ observe how different attributes collectively influence song popularity.
- ❑ will visualize correlations between features using a heatmap.
- ❑ analyze how energy and danceability interact simultaneously with popularity.
- ❑ explore how song characteristics evolve over time using year-based trends.

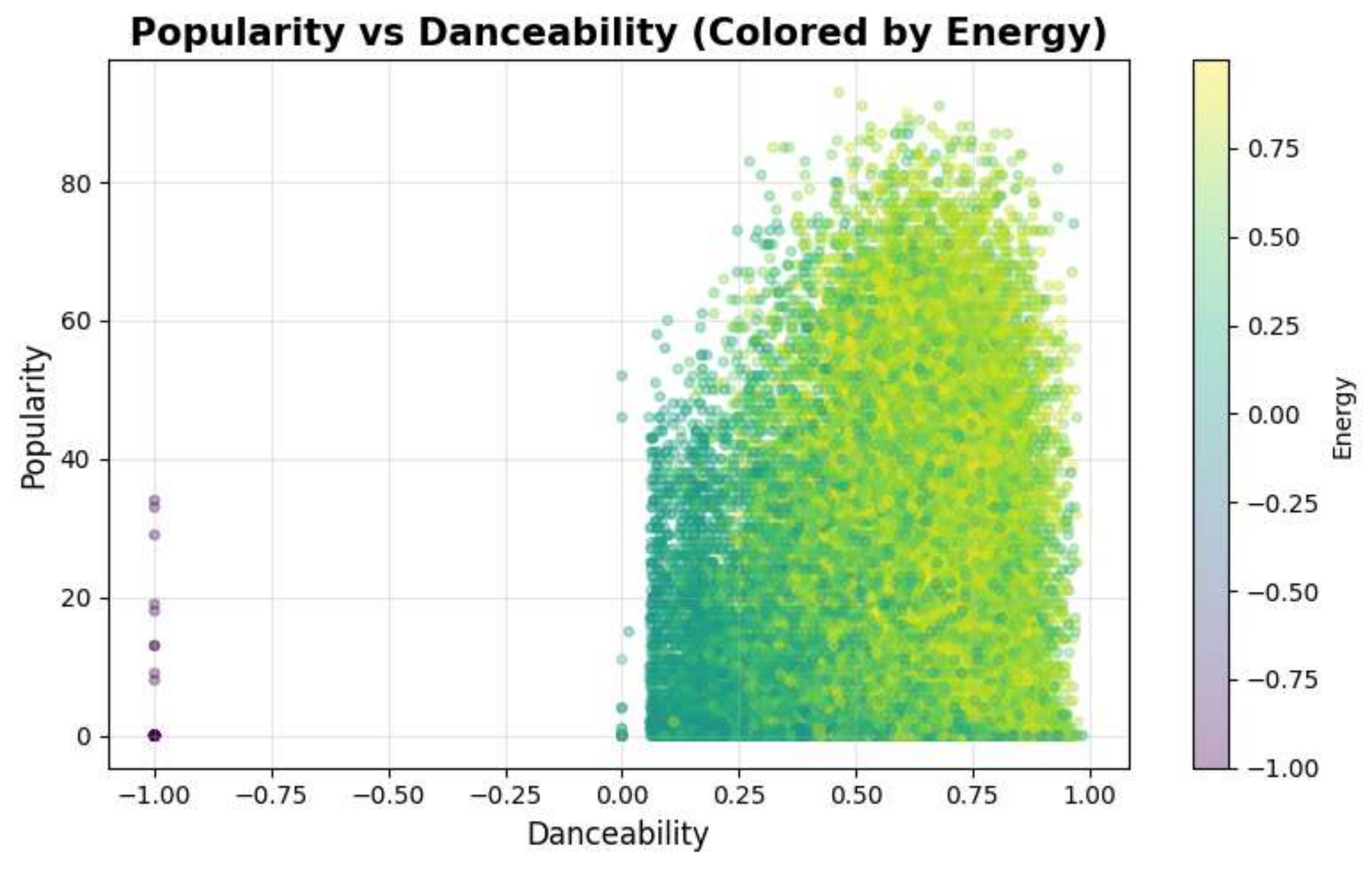
Correlation Heatmap



What this heatmap shows:
how strongly features correlate from –1 (negative) to +1 (positive)

- Energy and danceability show moderate positive correlation
- Popularity does not strongly correlate with any single feature

Popularity - Danceability - Energy

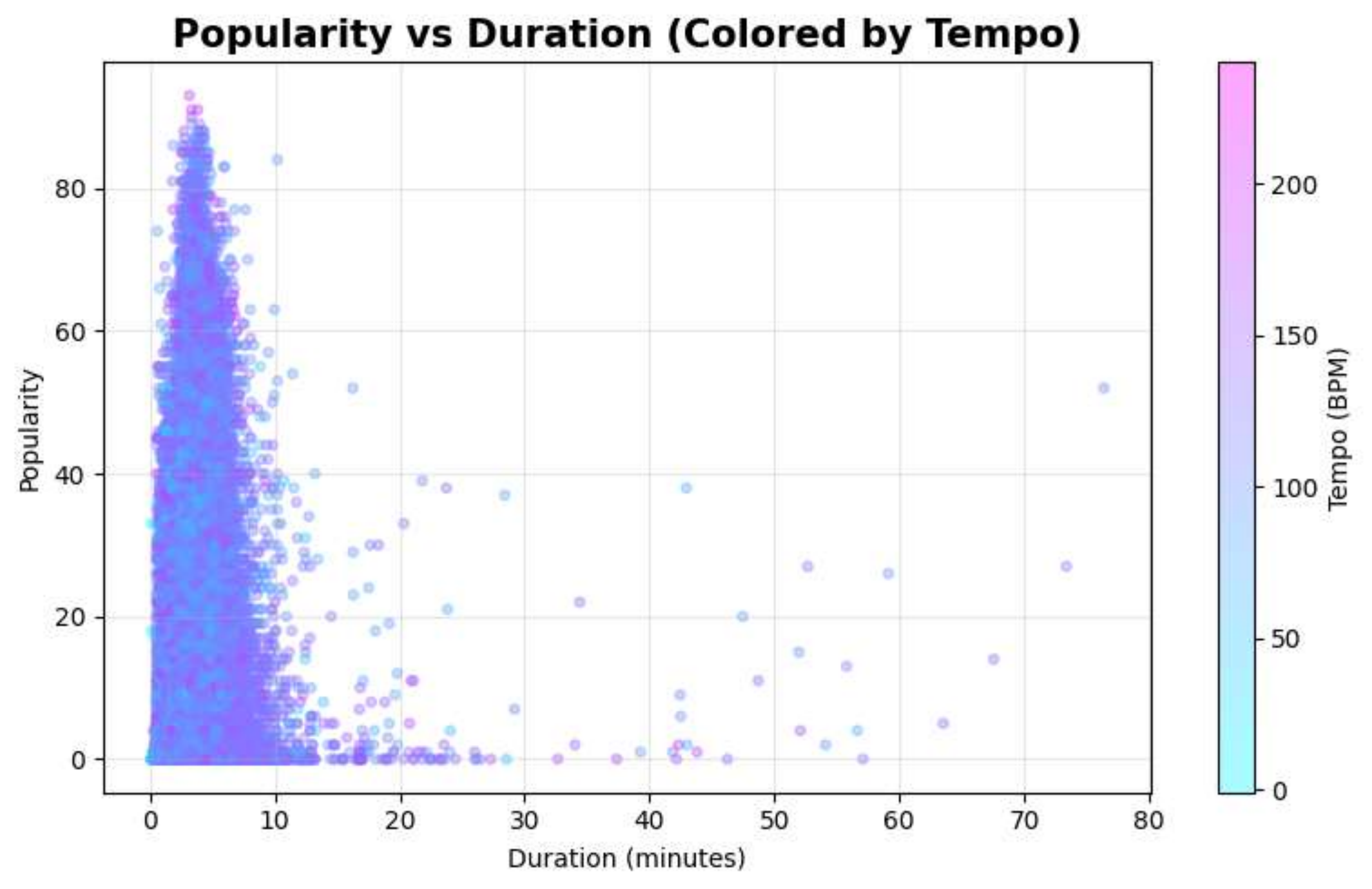


- Higher danceability songs often also exhibit higher energy.

- However, energy does not consistently result in higher popularity, since both energetic and calm tracks appear across popularity levels.

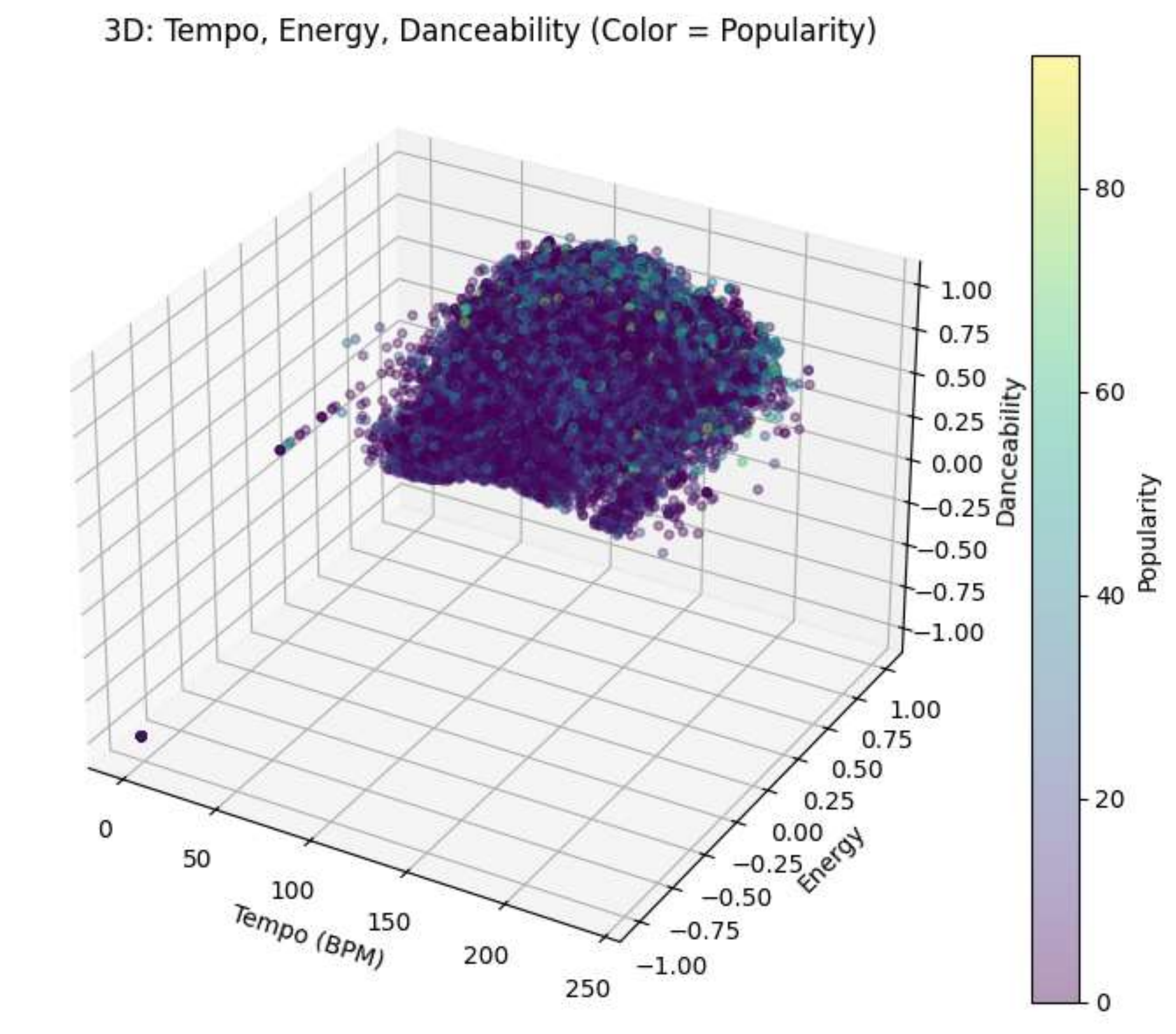
- Therefore, while energy strongly influences danceability, it does not strongly determine popularity.

Popularity – Duration - Tempo



- Popular songs typically fall in the standard duration range of around 2–4 minutes.
- Tempo variation across these durations shows that both slower and faster tempo tracks can be popular.
- This suggests that tempo and duration jointly influence engagement rather than individually dictating popularity.

Tempo – Danceability – Energy – Popularity



- Faster tempo songs tend to have higher energy, showing a natural link between speed and musical intensity.

- Higher danceability appears more consistently in tracks with higher energy, reinforcing their rhythmic drive.

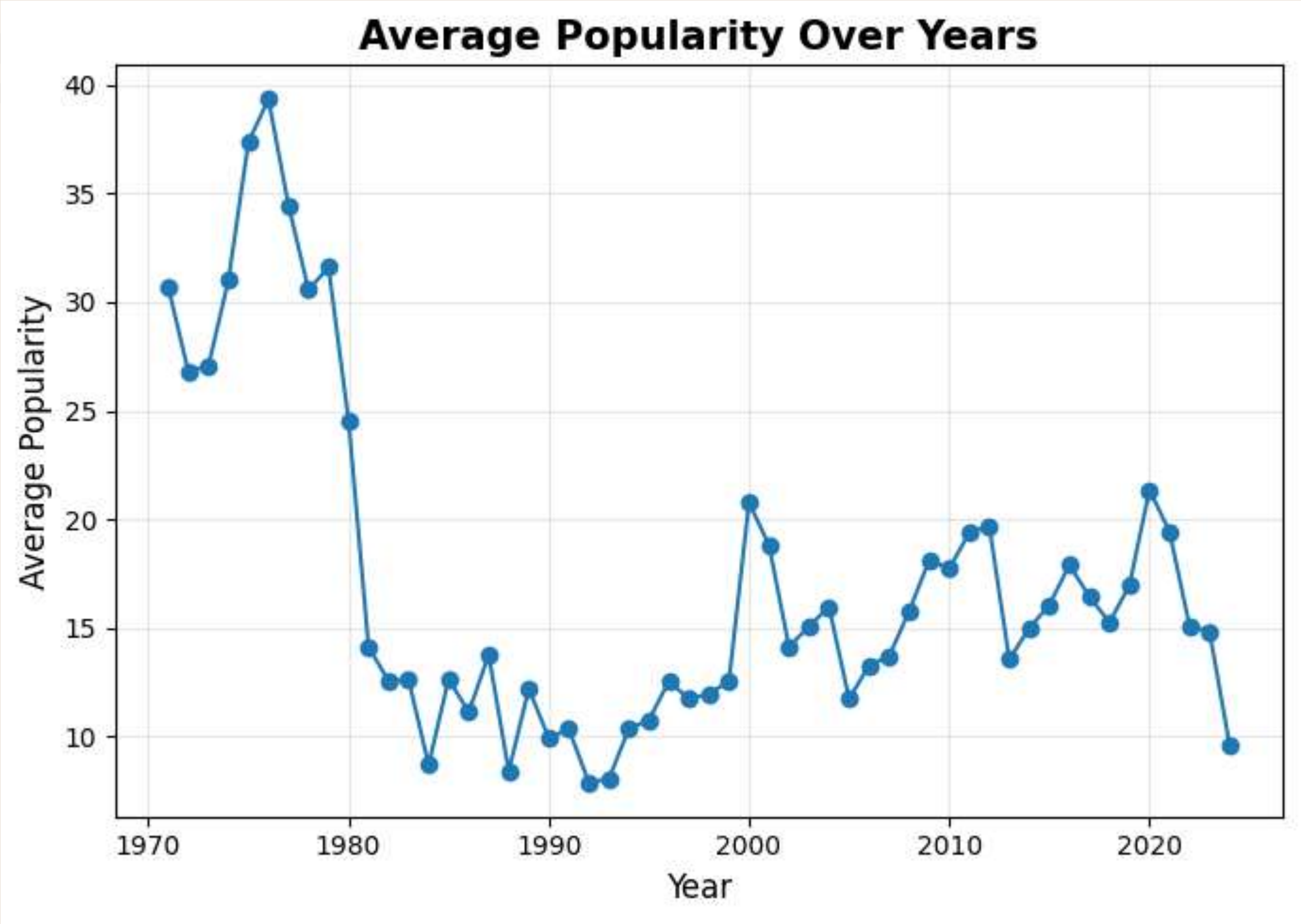
- Popularity (shown by color intensity) is spread across different combinations of tempo and energy, indicating that song success is not tied strictly to one specific musical pattern.

Key Insights

- ✓ Correlation analysis shows that danceability and energy move together, but popularity weakly correlates with any single feature, confirming it is multi-factor driven.
- ✓ Multivariate plots reveal that popularity emerges in balanced musical zones rather than at extreme values of tempo, energy, or danceability.
- ✓ Emotional tone (valence) and song pace show varied combinations, meaning hits can be upbeat, slow, energetic, or melancholic — diversity prevails.
- ✓ The interplay of tempo, energy, and rhythmic drive highlights structural patterns of modern music, but success still depends on broader audience and cultural context.
- ✓ Overall, popularity behaves as a combined outcome of multiple musical characteristics interacting rather than a simple linear dependency.

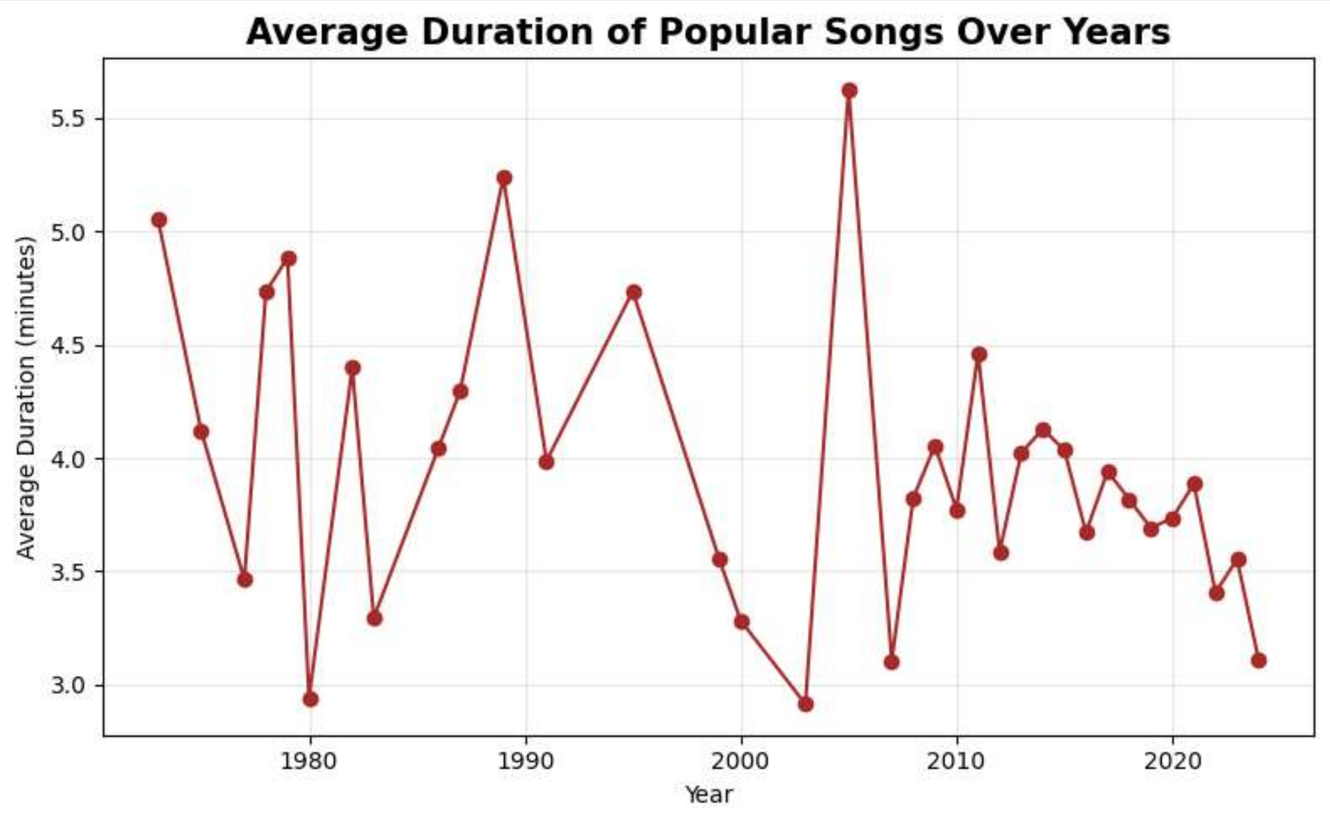
Time Series Analysis

Popularity – Time



- Older songs show higher average popularity because they have accumulated plays over longer periods.
- Classic or evergreen tracks maintain cultural value and continue to be streamed across generations.
- Newer songs may experience short-term spikes but have not had time to build sustained popularity.
- Popularity over time is influenced heavily by long-term listener behavior rather than just release date.

Duration – Time



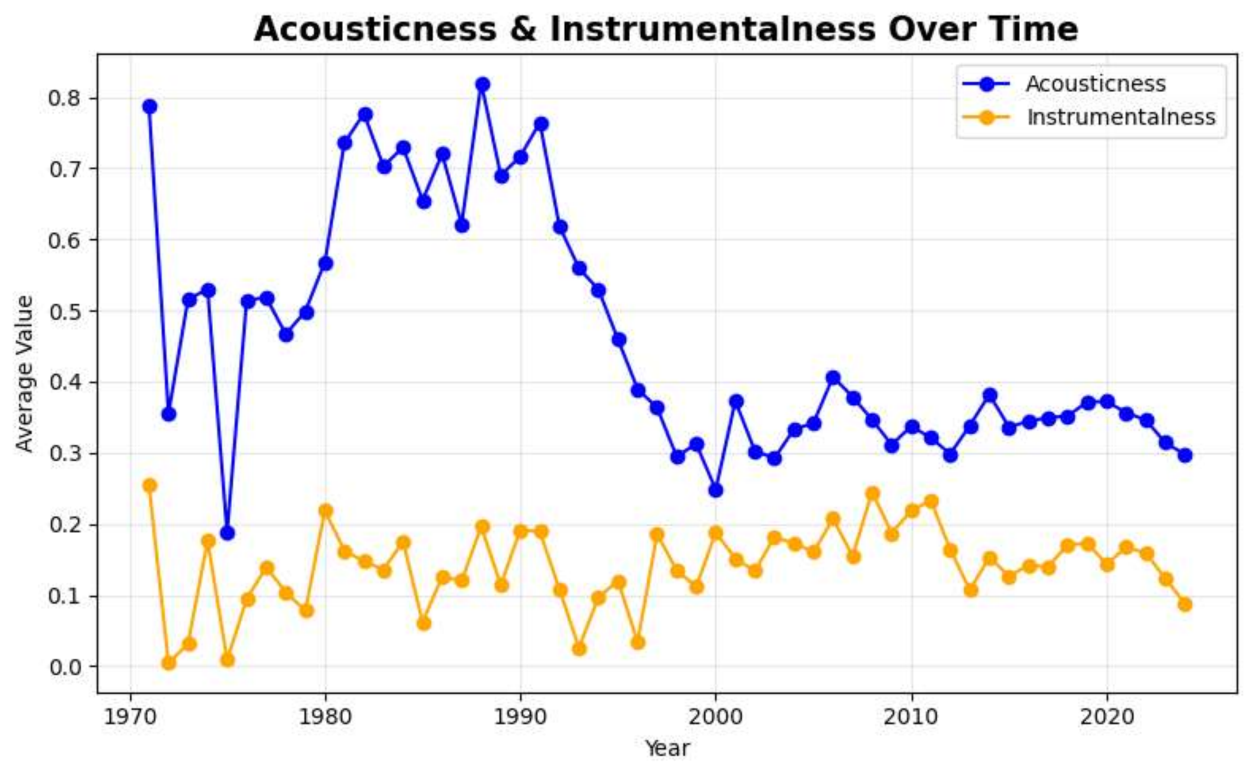
•Average song duration has decreased over time, showing a shift toward shorter tracks.

•Older music traditionally had longer compositions, often exceeding 4–5 minutes.

•Modern music is typically around 2–3 minutes long, likely to encourage replays and playlist sequencing.

•This trend aligns with shorter attention spans and streaming-optimized music structure.

Acousticness - Instrumentalness



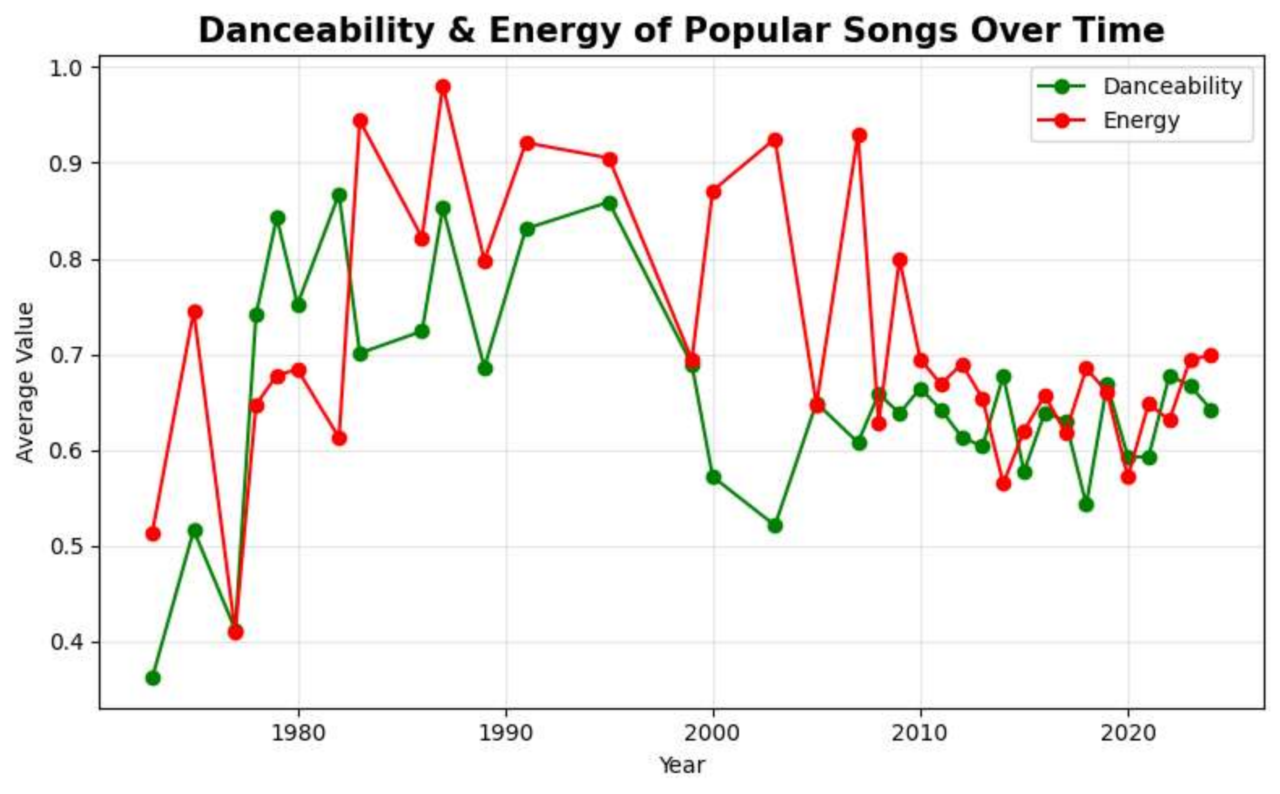
•Acousticness decreases across decades, indicating movement away from natural acoustic sounds.

•Instrumentalness remains consistently low, as most popular modern tracks are vocal-focused.

•The industry shift favors studio-produced electronic sound textures over organic instrumentation.

•Listener preferences appear to lean toward vocally expressive & digitally crafted tracks.

Danceability – Energy



- Popular songs consistently remain energetic over time, showing enduring preference for dynamic tracks.

- Danceability rises in recent years, reflecting a modern emphasis on rhythm & beat-driven music.

- Increasing danceability aligns with trends shaped by clubs, TikTok, reels, and playlist culture.

- Energy+Danceability together reveal a strong bias toward movement-friendly, upbeat compositions.

Key Insights

- ✓ Older songs show higher average popularity, as they have had more time to accumulate streams and achieve sustained cultural relevance
- ✓ Average song duration has decreased over time, with modern hit songs becoming shorter to encourage repeat streaming and faster engagement.
- ✓ Acousticness has steadily declined while instrumentalness remains low, reflecting a shift from natural acoustic sounds to digitally produced, vocal-centric music.
- ✓ Danceability has increased over the years, showing that modern music is more rhythm-focused and suited for movement, clubs, reels, and viral trends.
- ✓ Energy levels remain relatively high across decades, demonstrating that listeners consistently prefer strong, lively, dynamic tracks.
- ✓ Older songs show high accumulated popularity, indicating long-term listener loyalty and sustained cultural relevance over time.

Conclusion

- Popularity is not driven by any single musical attribute but rather by a combination of energy, danceability, tempo, and song duration.
- Most popular tracks fall into moderate ranges for multiple features, showing a preference toward balanced musical profiles.
- Emotional tone (valence) does not strongly determine popularity, meaning listeners appreciate both upbeat and melancholic music.
- Long-term streaming patterns suggest that older songs accumulate greater popularity over time, indicating sustained cultural value.
- Overall, music success appears to be multi-dimensional, reflecting complex listener behavior, marketing influence, and evolving musical trends.

Future Scope

- ❖ Conduct deeper genre-level analysis to understand how different genres evolve over time.

- ❖ Incorporate lyrics-based sentiment analysis to assess emotional meaning beyond valence score.
- ❖ Include external metrics such as social media trends, YouTube views, and radio play to broaden popularity interpretation.

- ❖ Use machine learning models to predict popularity using multiple musical features.

- ❖ Study demographic listening patterns (age groups, regions) to identify targeted audience behavior.

- ❖ Expand dataset beyond Spotify to compare platform-specific popularity across Apple Music, YouTube Music, etc.

THANK YOU!

PRESENTED BY:

TANVIR