WHAT RELATES TO A SONG'S POPULARITY ON SPOTIFY?

Fitri Oktaviani

ABOUT DATASET

- Dataset : 30000 Spotify Songs
- Almost 30,000 Songs from the Spotify API. See the readme file for a formatted data dictionary table.

	track_id	track_name	track_artist	track_popularity	track_album_id	track_album_name	track_album_release_date	playlist_name	
0	6f807x0ima9a1j3VPbc7VN	I Don't Care (with Justin Bieber) - Loud Luxur...	Ed Sheeran	66	2oCs0DGTsRO98Gh5ZSl2Cx	I Don't Care (with Justin Bieber) [Loud Luxury...	2019-06-14	Pop Remix	37i
1	0r7CVbZTWZgbTCYdfa2P31	Memories - Dillon Francis Remix	Maroon 5	67	63rPSO264uRjW1X5E6cWv6	Memories (Dillon Francis Remix)	2019-12-13	Pop Remix	37i
2	1z1Hg7Vb0AhHDiEmnDE79l	All the Time - Don Diablo Remix	Zara Larsson	70	1HoSmj2eLcsrR0vE9gThr4	All the Time (Don Diablo Remix)	2019-07-05	Pop Remix	37i
3	75FpbthrwQmzHlBJLuGdC7	Call You Mine - Keanu Silva Remix	The Chainsmokers	60	1nqYsOef1yKKuGOVchbsk6	Call You Mine - The Remixes	2019-07-19	Pop Remix	37i
4	1e8PAfcKUYoKkxPhrHqw4x	Someone You Loved - Future Humans Remix	Lewis Capaldi	69	7m7vv9wlQ4i0LFuJiE2zsQ	Someone You Loved (Future Humans Remix)	2019-03-05	Pop Remix	37i
...	...	...	...	...	...	...	...	...	
32828	7bxnKAamR3snQ1VGLuVfC1	City Of Lights - Official Radio Edit	Lush & Simon	42	2azRoBBWEEEYhqV6sb7JrT	City Of Lights (Vocal Mix)	2014-04-28	♥ EDM LOVE 2020	6j
32829	5Aevni09Em4575077nkWHz	Closer - Sultan & Ned Shepard Remix	Tegan and Sara	20	6kD6KLxj7s8eCE3ABvAyf5	Closer Remixed	2013-03-08	♥ EDM LOVE 2020	6j

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   track_id                  32833 non-null  object
 1   track_name                32828 non-null  object
 2   track_artist              32828 non-null  object
 3   track_popularity          32833 non-null  int64
 4   track_album_id            32833 non-null  object
 5   track_album_name          32828 non-null  object
 6   track_album_release_date  32833 non-null  object
 7   playlist_name             32833 non-null  object
 8   playlist_id               32833 non-null  object
 9   playlist_genre            32833 non-null  object
 10  playlist_subgenre         32833 non-null  object
 11  danceability              32833 non-null  float64
 12  energy                    32833 non-null  float64
 13  key                       32833 non-null  int64
 14  loudness                  32833 non-null  float64
 15  mode                      32833 non-null  int64
 16  speechiness               32833 non-null  float64
 17  acousticness              32833 non-null  float64
 18  instrumentalness          32833 non-null  float64
 19  liveness                  32833 non-null  float64
 20  valence                   32833 non-null  float64
 21  tempo                     32833 non-null  float64
 22  duration_ms               32833 non-null  int64
dtypes: float64(9), int64(4), object(10)
memory usage: 24.1 MB
```

- The dataset has 23 columns with these dtypes: 9 float64, 4 int64, and 10 object (string/text) columns.

TOOLS

- Google Colab
- Replicate
- Looker Studio
- Github
- IBM

PROJECT OVERVIEW

- Objective: Identify patterns associated with track popularity using genre, audio features, and release year.
- Key Questions:
    1. Which genres are most popular on average?
    2. How do danceability and energy compare by genre?
    3. What does the valence (positivity) distribution look like in top genres?
    4. How has tempo changed over time?
    5. Which audio features correlate most with popularity?
- Unit of Analysis: One row = one track; popularity = Spotify score (0–100).
- Use IBM Granite 3.3-8b

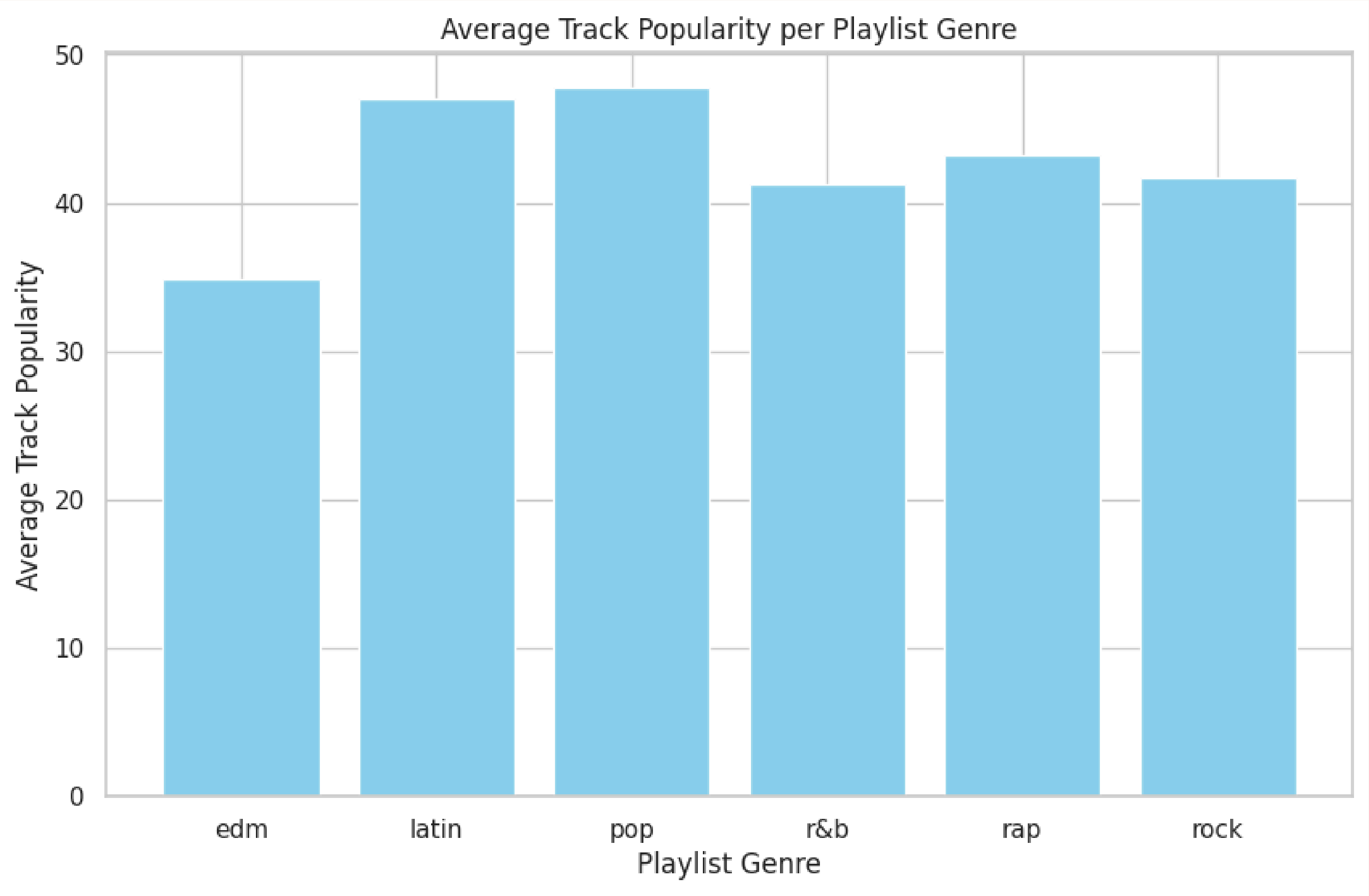
ANALYSIS PROCESS

- Columns used: track_popularity, danceability, energy, valence, tempo, loudness, acousticness, instrumentalness, liveness, speechiness, duration_ms, playlist_genre, playlist_subgenre, track_album_release_date.
- Type handling: Safe numeric coercion (errors='coerce'); per-chart NA drops.
- Aggregations:
    a. Mean popularity by playlist_genre.
    b. Mean danceability & energy by playlist_genre.
    c. Overlaid histograms of valence for top-5 genres by count.
    d. Mean tempo by release_year.
    e. Correlation matrix of audio features (+ popularity).
- Visualization: Matplotlib/Plotly; previews of aggregated tables for sanity checks.
- Quality controls: Top-N filtering for readability; optional 95% CI error bars.

INSIGHT & FINDINGS

GENRE POPULARITY

```
agent.invoke({"input": "Which genre has the highest average popularity? Visualization: Bar chart of average track popularity per playlist genre"}
```

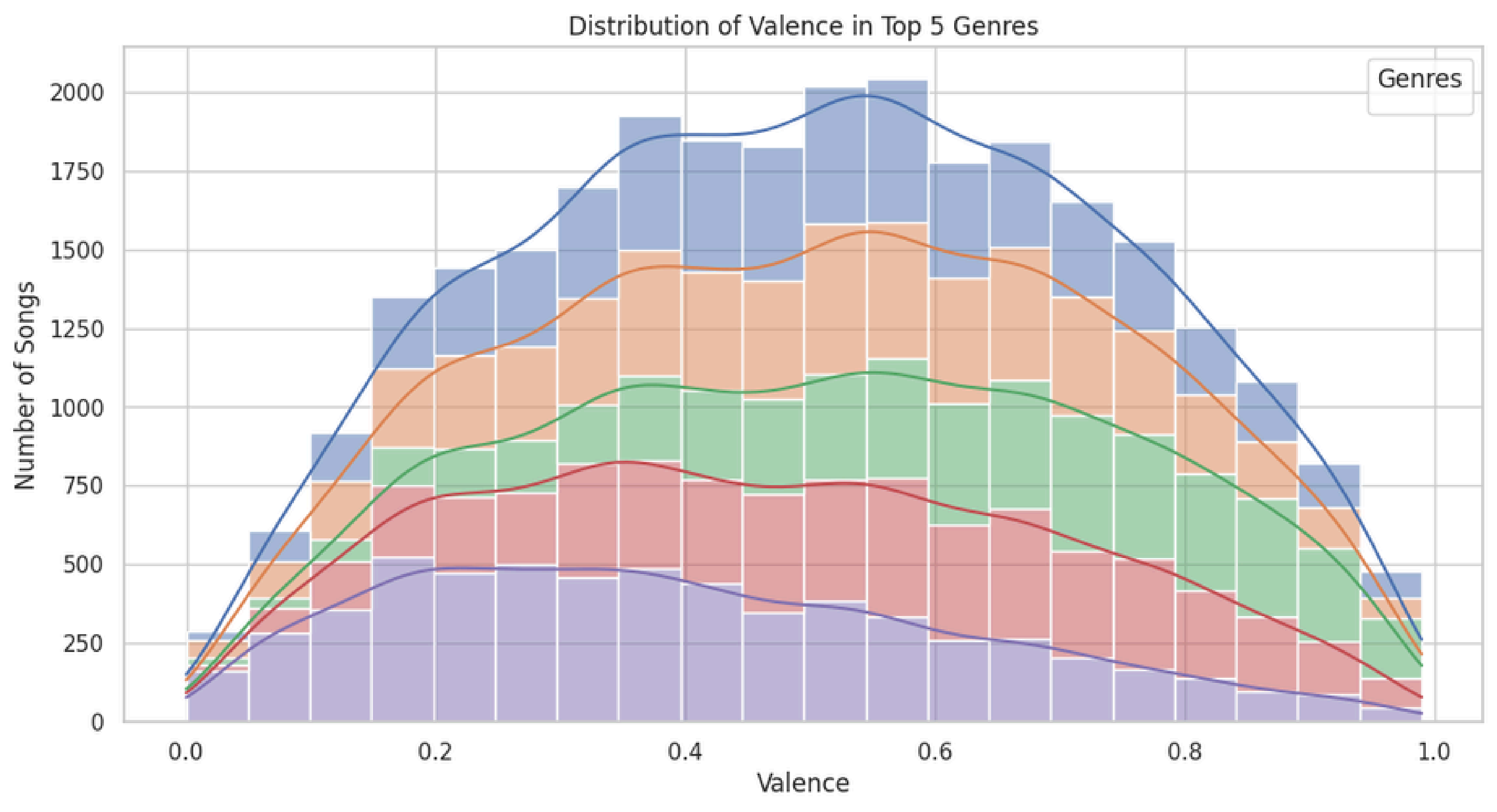


From the visualization, pop has the highest average popularity (~47.7), just above latin (~47.0)—a ~0.7-point gap, so they're essentially tied at first glance. R&B sits around 41.2, while EDM is lowest (~34.8). This suggests mainstream genres (pop/latin) tend to score higher in popularity than EDM in this sample.

INSIGHT & FINDINGS

DISTRIBUTION OF VALENCE (POSITIVITY) IN POPULAR GENRES

```
agent.invoke({"input": "How is the distribution of valence (positivity) in popular genres?"})
```

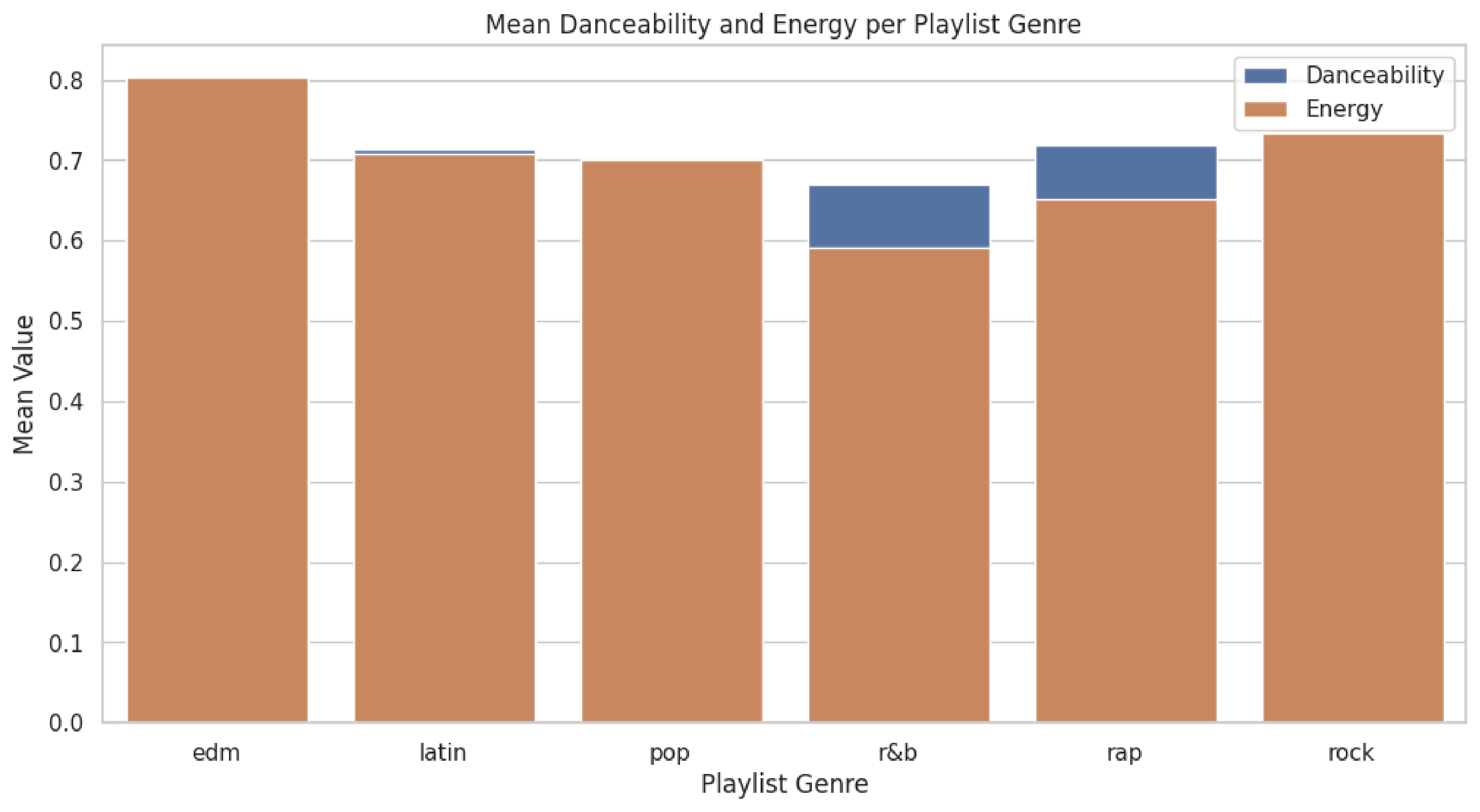


The histogram shows all five top genres (EDM, rap, pop, R&B, latin) clustering at valence 0.3–0.7 with a peak around 0.5–0.6—i.e., mostly neutral-to-cheerful moods. Pop and latin skew slightly higher (brighter), rap and R&B slightly lower (moodier), EDM has the widest spread but still centers around mid-positive, and extremes (<0.2 or >0.9) are rare—suggesting a mainstream preference for moderately cheerful tracks.

INSIGHT & FINDINGS

DANCEABILITY AND ENERGY COMPARE IN EACH GENRE

```
agent.invoke({"input": "How does danceability and energy compare in each genre)?"})
```

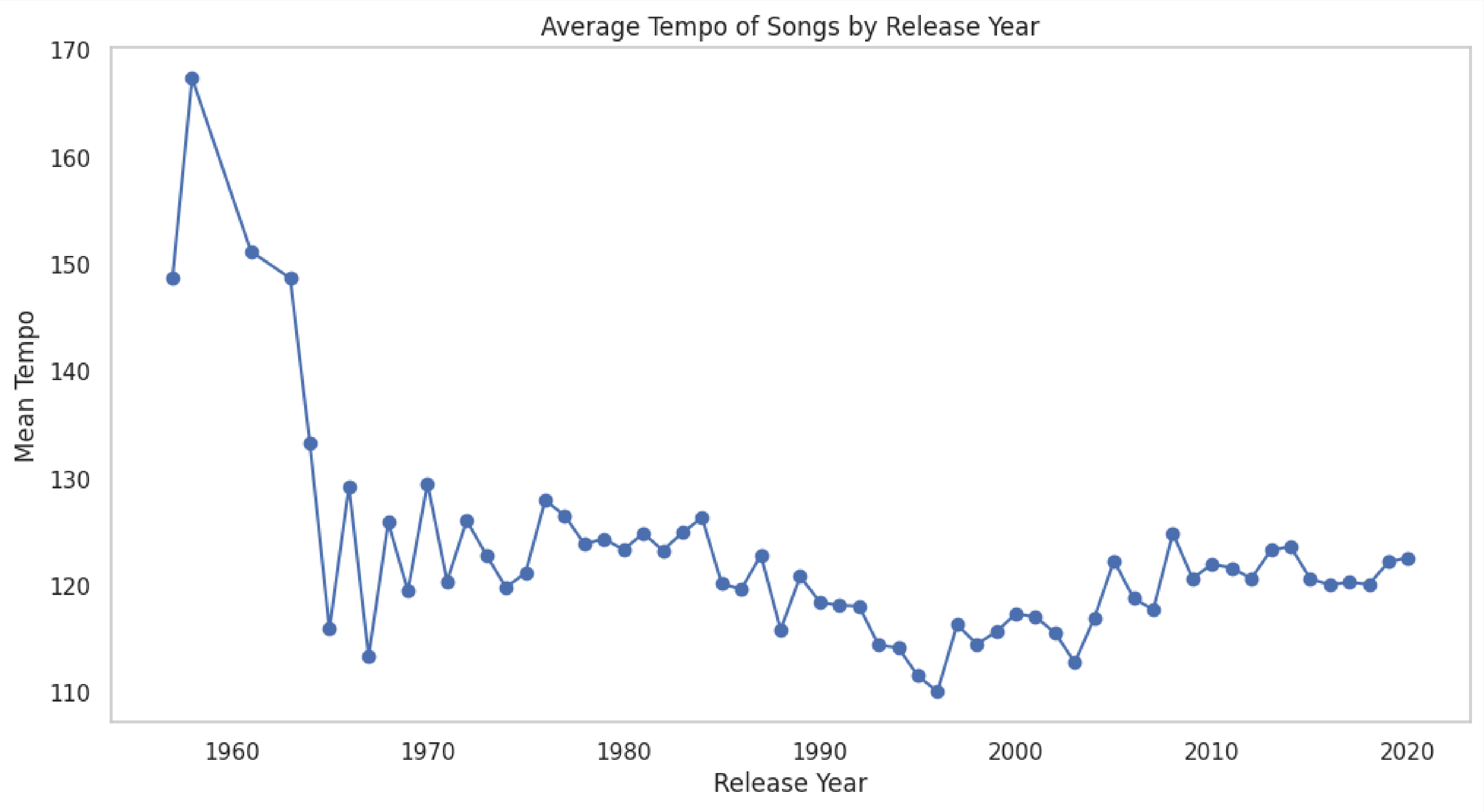


EDM leads in energy (~0.80) and is also highly danceable (~0.80); rock is fairly energetic (~0.73). Pop and latin sit around ~0.70 and are balanced (latin's danceability slightly above its energy). Rap has high danceability (~0.72) but lower energy (~0.65), making it groovier than aggressive. R&B shows the lowest energy (~0.59) with mid danceability (~0.67), matching a mellower profile. Overall, danceability generally tracks energy across genres—except rap, which over-indexes on danceability relative to energy.

INSIGHT & FINDINGS

AVERAGE TEMPO OF A SONG CHANGES OVER THE YEARS

```
agent.invoke({"input": "Create line chart release_year vs mean tempo, How the average tempo of a song changes over the years"})
```

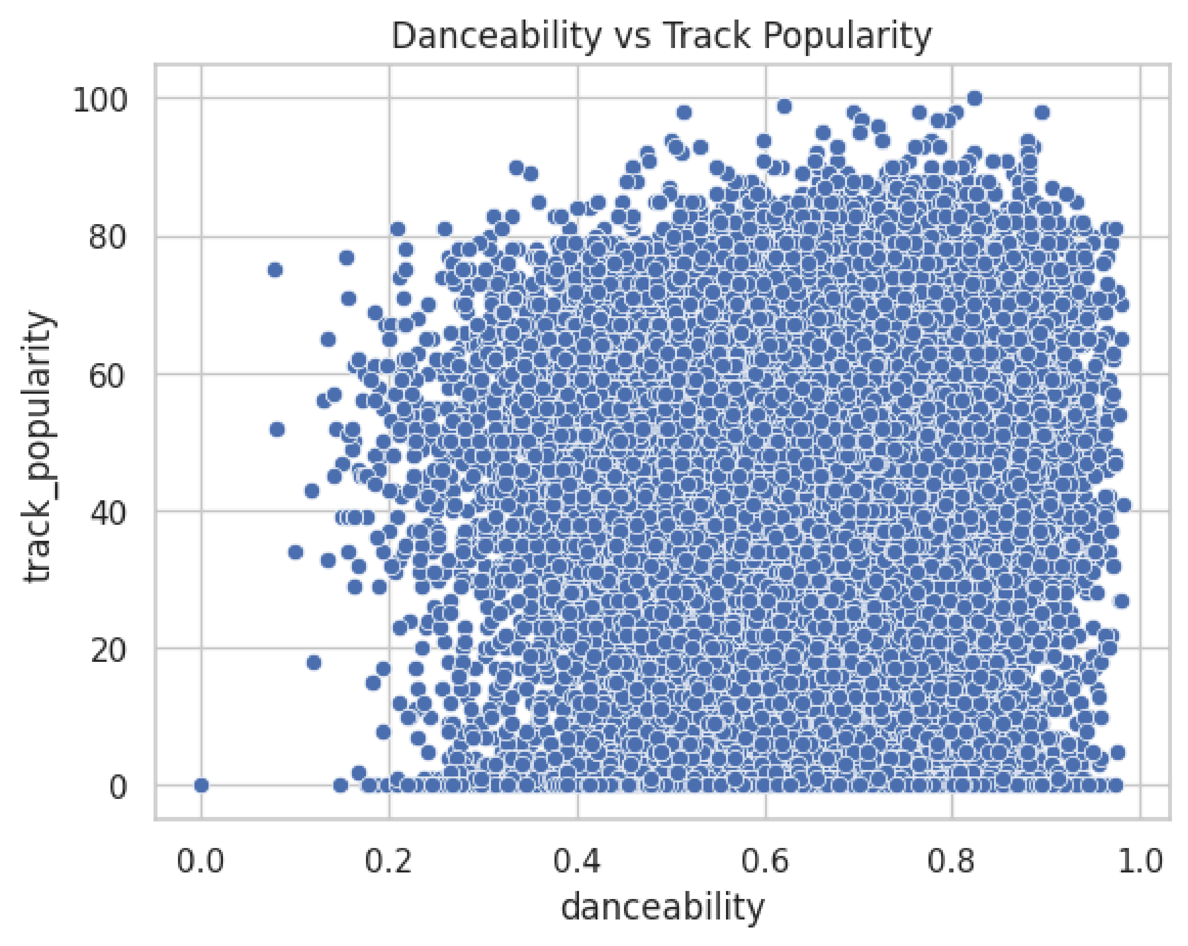


Average song tempo fell sharply from the late-1950s/early-1960s (~150–165 BPM) to the mid-1960s (~115–130 BPM), then stayed roughly stable through the 1970s–1990s near ~120 BPM with a mid-1990s dip (~110–115 BPM); from the 2000s to 2020 it inches up and flattens (~118–123 BPM), indicating that after the early drop tempo is broadly constant with small fluctuations likely driven by genre mix and playlist curation.

INSIGHT & FINDINGS

DANCEABLE SONGS TREND TO BE MORE POPULAR

```
agent.invoke({"input": "create catter plot,do more danceable songs tend to be more popular?"})
```

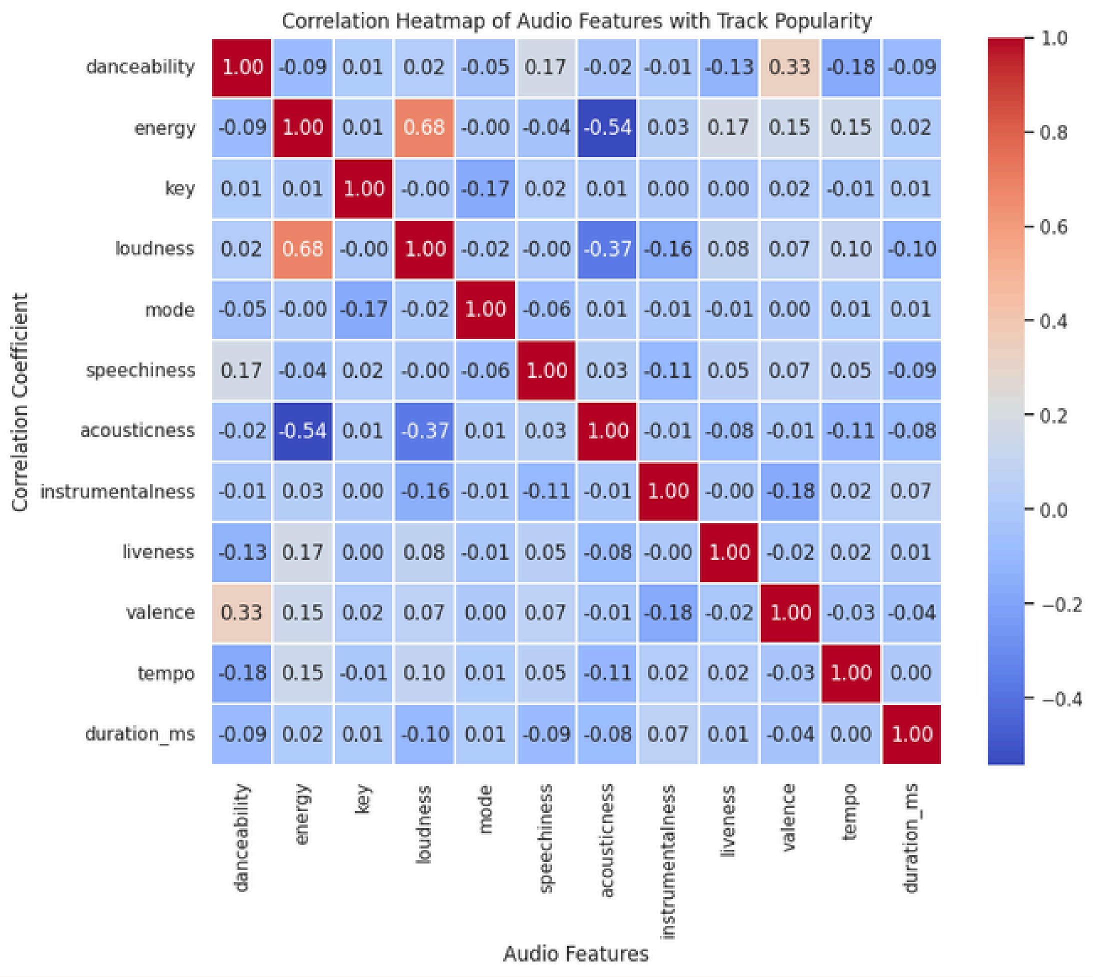


The scatter shows a weak positive correlation between danceability and track popularity: songs with medium–high danceability (~0.6–0.9) more often reach high popularity, whereas at low danceability (<0.3) highly popular points are rare.

INSIGHT & FINDINGS

AUDIO FEATURES ARE MOST CORRELATED WITH POPULARITY

```
agent.invoke({"input": "create a heatmap of which audio features are most correlated with popularity?"})
```



The heatmap shows that most audio features have only weak correlations with popularity (near 0), so no single metric “makes” a hit; danceability and valence are small positives, while acousticness, tempo, and duration_ms are small negatives. Independently of popularity, energy ↔ loudness is strong (~0.68) and danceability ↔ valence is moderate (~0.33); other pairs are generally weak.

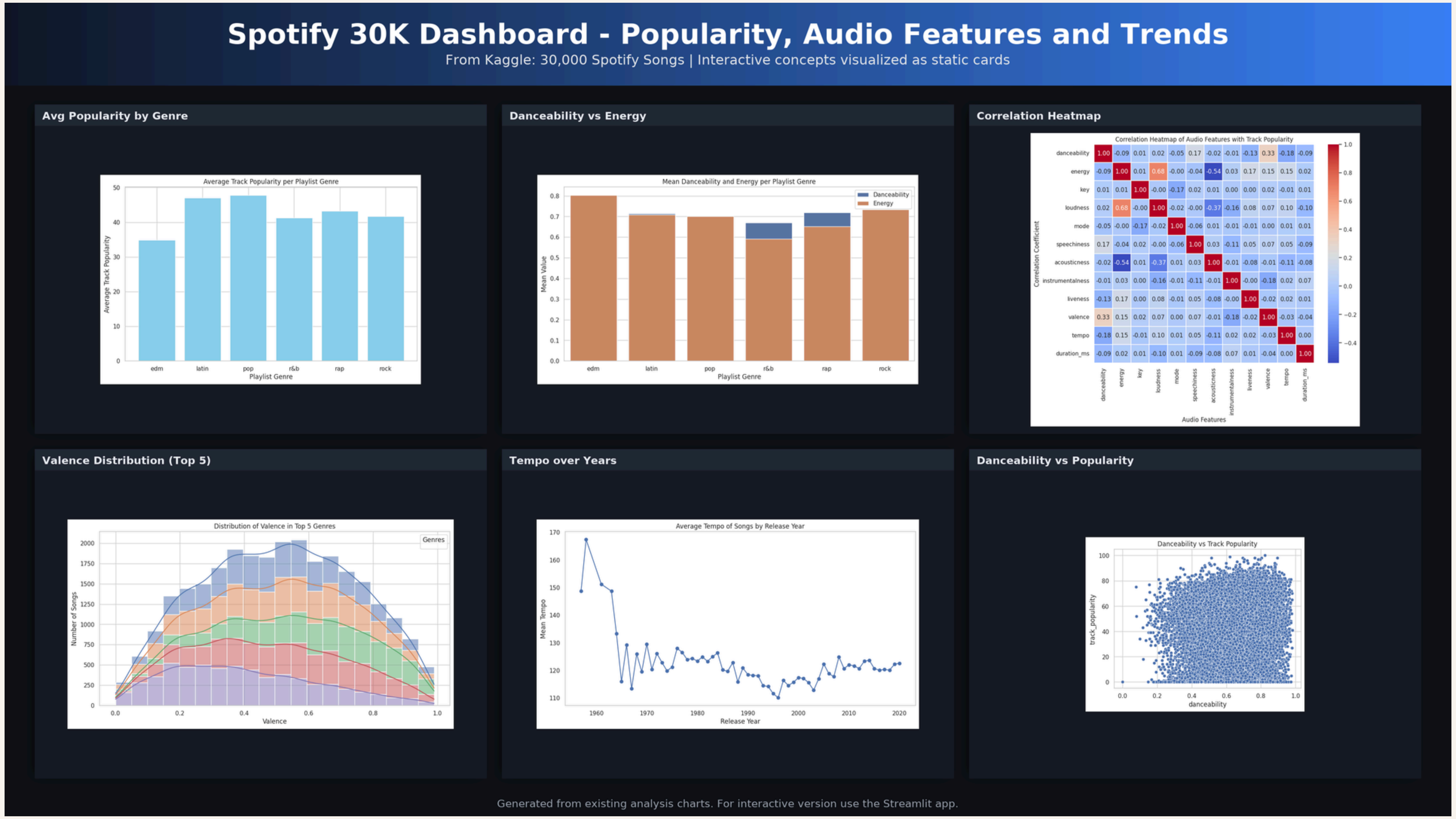
CONCLUSIONS

- Pop & Latin lead average popularity; EDM trails in this dataset.
- High energy + high danceability are common in favored genres, but no single feature determines hits.
- Moderate-to-high valence (brighter mood) dominates popular genres.
- Tempo has been broadly stable since the 1970s (~120 BPM).

RECOMMENDATIONS

- Artists/Producers: Prioritize memorable, danceable hooks, positive valence, and energetic mixes, while keeping artistic differentiation.
- Playlist Curators: Use Pop/Latin as anchor genres; blend Rap for groove and EDM for energy peaks.
- Experiments: A/B test high-valence + mid-tempo (118–123 BPM) vs moderate-valence + faster tempo for engagement.
- Further Analytics: Add 95% CI / ANOVA, break down by subgenre and release year, and build a multivariate prediction model.

DASHBOARD


Spotify 30K Dashboard - Popularity, Audio Features and Trends
From Kaggle: 30,000 Spotify Songs | Interactive concepts visualized as static cards
Avg Popularity by Genre
Average Track Popularity per Playlist Genre
Playlist Genre
Danceability vs Energy
Mean Danceability and Energy per Playlist Genre
Playlist Genre
Correlation Heatmap
Correlation Heatmap of Audio Features with Track Popularity
Audio Features
Valence Distribution (Top 5)
Distribution of Valence in Top 5 Genres
Valence
Tempo over Years
Average Tempo of Songs by Release Year
Release Year
Danceability vs Popularity
Danceability vs Track Popularity
danceability
Generated from existing analysis charts. For interactive version use the Streamlit app.

AI SUPPORT EXPLANATION

IBM Granite acted as an analytics co-pilot: shaping research questions, generating & refining Python code (pandas/Matplotlib/Plotly/Streamlit), fixing schema issues (e.g., track_genre → playlist_genre), parsing mixed dates into a robust release_year, and helping craft correlation-based insights and recommendations. Every AI output was reviewed, executed, and validated by me.

ROLES BY PHASE

- Scoping & Ideation: Proposed focused questions (genre vs average popularity; danceability vs energy; valence distributions; tempo over years; correlation heatmap).
- Data Preparation: Produced templates for schema auditing (df.info()), safe numeric coercion, targeted NA drops, and mixed-format date parsing to release_year.
- Rapid Debugging: Diagnosed errors (e.g., KeyError on genre columns) and suggested robust fallbacks and parsing strategies.
- EDA & Visualization: Generated reproducible prompts/snippets for bar, grouped bars, overlaid histograms, line charts, heatmaps, and boxplots (with optional 95% CIs).
- Narrative Building: Helped translate numbers into association-focused insights and practical (non-causal) recommendations.

AI SUPPORT EXPLANATION

EXAMPLE PROMPTS

- “Create a robust release_year from mixed track_album_release_date; coerce numerics; line-plot mean tempo by year.”
- “Bar chart: average track_popularity by playlist_genre; validate columns; show top-N and 95% CI.”
- “Grouped bars: compare danceability vs energy across playlist_genre, sorted by danceability.”
- “Build a correlation heatmap for audio features incl. track_popularity; print the top correlations.”

THANK YOU

Linkedin

Gmail

Bentome