SIS 1 : Data Collection & Preparation

Team Khan Kensey and Shakhzod Abdullaev

Top 100 Songs — Spotify vs Billboard
Shakhzod Abdullaev — 22B031601

Khan Kensey - 22B030608

Objective

Build a full data pipeline combining Spotify API and Billboard website scraping, clean and merge the two datasets, perform basic EDA and create visualizations to analyze the relationship between streaming popularity (Spotify) and chart ranking (Billboard Hot 100).

Data Sources

- Web scraping: *Billboard Hot 100* — https:// billboardtop100of.com/
    - Scraped song title, artist, and current chart rank for the top 100 songs.
- API: *Spotify Web API (via Spotipy)* — https://developer.spotify.com/documentation/web-api/
    - Collected track metadata: popularity score (0–100)
- API: *ReccobeatsAPI*— https://reccobeats.com/docs/apis/
    - Collected track metadata: danceability, energy, tempo (BPM), and duration (ms).

- Politeness:
    - Used browser-style headers (User-Agent) for scraping.
    - Added short delays between Spotify API calls to avoid rate limits.
    - Saved API results to a CSV to make the notebook reproducible.

Methodology

1. Scraping Billboard Hot 100:
    - Used requests + BeautifulSoup to parse the Billboard chart HTML.
    - Extracted Rank, Title, and Artist from the main chart list.

2. Spotify API Queries:
    - For each (Title, Artist) pair from Billboard data, queried Spotify Search API.
    - Extracted track popularity, danceability, energy, tempo, and duration_ms.
    - Saved results to spotify_data.csv for reproducibility.

Data Preparation and Join

- Normalization: lowercased all song and artist names,
- Join: left join on (title, artist) between Billboard and Spotify data.
- Post-cleaning: dropped nulls, removed duplicates
- Final columns: [Index, Billboard year, Rank, Song, Artist, Popularity, Danceability, Energy, Tempo, Valence]

EDA (Exploratory Data Analysis)

- Spotify Popularity:
    - Mean = 52.7, Min = 0, Max = 91.0.
    - Top 10 songs all above 85 popularity.
- Danceability & Energy:
    - Average danceability = 0.67, energy = 0.67.
    - Low positive correlation (r ≈ 0.05) between energy and tempo.
- Tempo (BPM):
    - Range 51–208 BPM, median around 122 BPM.
- Billboard Rank vs Spotify Popularity:
    - Weak negative correlation (r ≈ -0.01): almost zero – billboard standings doesn't represent spotifys popularity
    - Song can rank high on billboard, but be low in popularity on spotify

---

Key Findings

- Spotify Popularity and Billboard Rank are almost not correlated: high rank on billboard doesn't mean hight popularity on spotify

- Most Top 100 tracks fall in 100–130 BPM, aligning with mainstream pop rhythm.
- Energy and valence are moderately linked — upbeat and energetic tracks tend to sound happier.
- Energy is negatively correlated to accousticness – energetic songs are usually more electronic.