Exploratory Data Analysis on Spotify Tracks

Subtitle: Data Science Project

Name: Arka Koley

Tools Used: Python (Pandas, Seaborn, Matplotlib)

Date: October 2025

Introduction

- This project explores Spotify track data using Exploratory Data Analysis (EDA).
- EDA is crucial for understanding patterns, trends, and relationships within a dataset before building any formal models.
- Objective: To identify the features that contribute to a song's popularity and to analyze how musical trends have evolved over time.

About the Dataset

The dataset contains a comprehensive collection of Spotify tracks, each detailed with over 20 distinct feature

Total Records: ~50,000 tracks

Key Columns:

- Identification: track_name, artist_name, year
- Performance Metric: popularity
-
- Audio Features: danceability, energy, acousticness, loudness, speechiness, valence, tempo, liveness, instrumentalness

EDA Objectives

Our analysis is structured around the following objectives:

1. Univariate Analysis: Explore the distribution and characteristics of individual features.
2. Bivariate Analysis: Study the relationships and interactions between pairs of features.
3. Multivariate & Correlation Analysis: Examine the correlations between multiple features simultaneously to uncover complex patterns.
4. Time Series Analysis: Track and analyze how musical features and popularity have changed across different years.
5. Generate Insights & Recommendations: Synthesize findings to provide actionable insights.

Data Cleaning & Preprocessing

To ensure the quality and reliability of our analysis, the following data cleaning steps were performed:

- Duplicate Removal: All duplicate rows were identified and removed.
- Missing Values: The dataset was checked for any missing values, which were handled accordingly.
- Data Type Conversion: Numeric columns were converted to their appropriate data types for accurate calculations.
- Data Verification: Column values were checked to ensure they fell within expected ranges and formats.

Univariate Analysis (Results)

Initial analysis of individual features revealed several key trends:

- Popularity: Most tracks cluster in the 40–70 range, indicating moderate popularity.
- Duration: The average song duration is approximately 3.5 minutes.
- Audio Features: A majority of the songs in the dataset are characterized by high danceability and energy.
- Musical Structure: The most common time signature is 4/4, and the majority of tracks are in a major key.

Bivariate & Correlation Analysis

Examining the relationships between features provided deeper insights:

- A strong positive correlation exists between danceability and energy.
- Popularity shows a slight positive correlation with both energy and danceability.
- Loudness and energy are highly correlated, which is an expected relationship.
- A negative trend was observed between acousticness and popularity, suggesting that acoustic songs tend to be less popular on average in this dataset.