Spotify Song Popularity Analysis

Exploring the Relationship Between Danceability and Song Popularity

Project Information

Executive Summary

In today's music streaming ecosystem, over 100,000 songs are uploaded to Spotify daily. For artists, producers, and music labels, understanding what drives a song's success is crucial. This project investigates a fundamental question: **Does danceability influence song popularity on Spotify?**

Using a comprehensive dataset of 232,726 Spotify tracks and multiple linear regression analysis, Team C4 examines the relationship between danceability—a song's rhythmic suitability for dancing—and its popularity score. Through rigorous exploratory data analysis (EDA), statistical modeling, and genre-specific investigation, we provide data-driven insights into the audio features that shape listener engagement.

Our analysis reveals a **positive and statistically significant relationship** between danceability and popularity (coefficient = 15.90, $p < 0.001$). However, our findings also demonstrate that popularity is multi-dimensional, influenced by energy, valence, acousticness, loudness, and genre context. While danceability boosts a song's chances of success, it does not guarantee a hit—songs still need the right emotional and musical context to truly break through.

Business Challenge

Current State:

- **Market Saturation:** Tens of thousands of songs uploaded daily create intense competition for listener attention
- **Unknown Success Factors:** Artists and producers lack data-driven insights into which musical attributes drive popularity
- **Opaque Algorithms:** Spotify's recommendation systems and popularity scoring mechanisms are not transparent
- **Genre Complexity:** Different musical genres may respond differently to the same audio features
- **Intuition-Based Decisions:** Music production decisions often rely on subjective judgment rather than empirical evidence

Key Challenges:

- How can we quantify the relationship between audio features and popularity?
- Which musical characteristics matter most for streaming success?
- Do these patterns hold across different genres and musical styles?

Project Objectives

Our mission is to transform Spotify's audio feature data into actionable insights for the music industry:

- **Data Familiarization:** Thoroughly understand the Spotify Features dataset structure, variables, and business meaning
- **Exploratory Analysis:** Discover patterns, distributions, and relationships across audio features and popularity
- **Statistical Modeling:** Build multiple linear regression models to isolate danceability's effect while controlling for confounders
- **Genre Analysis:** Investigate how danceability-popularity relationships vary across musical genres
- **Causal Framework:** Design an ideal randomized experiment to establish causality (for future research)
- **Strategic Recommendations:** Develop actionable insights for artists, producers, and music platforms

Methodology

This project follows a comprehensive four-phase approach combining exploratory data analysis with rigorous statistical modeling:

Phase 1: Problem Definition & Hypothesis Development

Team: Chia-Chun Hung (Tony)

Activities:

- Define research question: What drives song popularity on Spotify?
- Formulate testable hypothesis: Songs with higher danceability have higher popularity
- Identify key variables: Danceability (IV), Popularity (DV)
- Justify business relevance and practical importance

Phase 2: Ideal Experiment Design

Team: Zhengyuan Pei (Sean), Hsin-Pan Chen (Blanca)

Activities:

- **Random Assignment:** Users randomly assigned to playlists with varying danceability levels (high/low/mixed)
- **Standardized Exposure:** Control for artist popularity, release dates, and algorithmic visibility
- **Confounder Control:** Hold energy, valence, tempo, and acousticness constant across conditions
- **Document Limitations:** Acknowledge observational data constraints and missing variables

Phase 3: Exploratory Data Analysis & Visualization

Team: Xiaoyu Ma, Yang-Hsuan Lin (Melody)

Activities:

- **Univariate Analysis:** Calculate summary statistics, create distributions (histograms, boxplots)
- **Bivariate Analysis:** Construct correlation matrix, create scatterplots
- **Genre Analysis:** Compare danceability-popularity patterns across musical categories
- **Pattern Identification:** Discover trends, outliers, and initial insights

Phase 4: Regression Modeling & Results

Team: Dat Nguyen, Victoria Nguyen (Vic)

Activities:

- Build multiple linear regression model with audio features
- Test variable significance and perform model diagnostics
- Interpret coefficients and quantify effect sizes
- Generate prediction intervals for example tracks
- Synthesize findings and develop recommendations

Dataset Overview

Source: Spotify Features.csv (Kaggle, published by Somu Mourya, 2023)

Size: 232,726 unique Spotify tracks across multiple genres and artists

Observation Unit: Each row represents one unique track at the song level (no aggregation)

Key Variables

Variable	Type	Description
popularity	Numeric (0-100)	Dependent Variable (DV): Track's popularity score
danceability	Numeric (0-1)	Main Independent Variable (IV): Suitability for dancing based on rhythm, beat strength, and regularity
energy	Numeric (0-1)	Control Variable: Intensity and activity level of the track
valence	Numeric (0-1)	Control Variable: Musical positiveness/cheerfulness conveyed by the track
tempo	Numeric (BPM)	Beats per minute, defining the track's speed and rhythm
acousticness	Numeric (0-1)	Control Variable: Measure of acoustic vs. electronic production quality
loudness	Numeric (dB)	Control Variable: Overall volume/loudness of the track
instrumentalness	Numeric (0-1)	Likelihood of the track being instrumental with no vocals
genre	Categorical	Musical genre classification (Pop, Rock, Jazz, Classical, etc.)

Key Deliverables

1. Technical Analysis Report

Comprehensive documentation including:

- Complete data dictionary with variable definitions and business meanings
- Descriptive statistics (mean, median, standard deviation, min, max)
- Correlation matrix and bivariate analysis
- Multiple linear regression model specification and results
- Model diagnostics (residual plots, VIF analysis, heteroscedasticity checks)
- Data visualizations (scatterplots, boxplots, distribution charts)

2. Python Analysis Scripts

Reproducible code for analysis:

- Jupyter notebook (spotify_analysis.ipynb) with complete workflow
- Data loading and cleaning procedures
- OLS regression implementation using StatsModels
- Prediction interval calculations and visualizations

3. Executive Presentation

Professional PowerPoint deck for stakeholders:

- Problem definition and business motivation (6-8 minutes)
- Ideal experiment design and causal framework
- Key findings and data visualizations
- Regression results and interpretation
- Strategic recommendations for artists and platforms

4. GitHub Repository

Open-source project documentation:

- Comprehensive README with project overview
- Code organization and repository structure
- Installation instructions and dependencies
- Links to dataset source and references

Expected Outcomes & Impact

Business Value

- **Data-Driven Music Production:** Artists and producers gain quantitative insights into audio features that influence popularity
- **Strategic Genre Positioning:** Understanding genre-specific patterns enables better targeting of audience preferences
- **Platform Optimization:** Streaming services can refine recommendation algorithms using validated audio feature relationships
- **Evidence-Based Marketing:** Music labels can make informed decisions about promotion and playlist placement

Academic Contributions

- **Regression Analysis Proficiency:** Practical application of multiple linear regression to real-world data
- **Causal Inference Framework:** Understanding limitations of observational studies and experimental design principles
- **Data Storytelling:** Translating statistical findings into compelling business narratives
- **Model Diagnostics:** Identifying and addressing heteroscedasticity, multicollinearity, and assumption violations

Tools & Technologies

Statistical Analysis

- **Python 3.12+:** Primary programming language for data analysis
- **StatsModels 0.14.6:** OLS regression modeling and statistical inference
- **Pandas:** Data manipulation and cleaning
- **NumPy:** Numerical computing and array operations

Visualization

- **Matplotlib:** Statistical plotting and charts
- **Seaborn:** Advanced statistical visualizations
- **Microsoft Excel:** Quick data exploration and summary statistics

Presentation & Documentation

- **Microsoft PowerPoint:** Executive presentation slides
- **Microsoft Word:** Technical reports and documentation
- **Jupyter Notebook:** Interactive code documentation
- **GitHub:** Version control and project sharing

Success Criteria

The project will be evaluated on:

- **Technical Excellence (30%):** Completeness of analysis, accuracy of regression modeling, quality of diagnostics
- **Statistical Rigor (25%):** Appropriate methodology, valid assumptions, proper hypothesis testing
- **Presentation Quality (20%):** Clear storytelling, professional slides, effective time management (6-8 minutes)
- **Insight Generation (15%):** Actionable recommendations, business relevance, strategic thinking
- **Team Collaboration (10%):** Contribution equity, professionalism, peer evaluation

Conclusion

This project represents a comprehensive analytical engagement that combines statistical rigor with practical business insights. By examining 232,726 Spotify tracks, we provide empirical evidence on the relationship between audio features and song popularity.

Our findings reveal that **danceability significantly influences popularity**, with a 0.10 increase in danceability associated with approximately 1.6 additional popularity points. However, we also demonstrate that popularity is **multi-dimensional**, shaped by energy, valence, acousticness, loudness, and genre context.

While our R^2 of 0.213 indicates that audio features explain only 21% of popularity variance, this is expected given that streaming success depends heavily on external factors like marketing budgets, playlist placement, artist recognition, and social media virality. Our analysis provides a **foundational understanding** of the audio-driven component of success.

Through this engagement, artists and music industry professionals gain **data-driven guidance** on audio optimization, while acknowledging that *danceability boosts a song's chances but does not guarantee a hit*. A song still needs the right emotional and musical context—and often the right promotional support—to truly break through.

Appendix

Team Roles & Responsibilities

- **Chia-Chun Hung (Tony):** Problem Definition & Motivation
- **Zhengyuan Pei (Sean):** Ideal Experiment Design
- **Hsin-Pan Chen (Blanca):** Ideal Experiment Design
- **Xiaoyu Ma:** Descriptive Statistics & Visualizations

- **Yang-Hsuan Lin (Melody):** Descriptive Statistics & Visualizations
- **Dat Nguyen:** Regression Model Development
- **Victoria Nguyen (Vic):** Regression Model Results & Interpretation

Project Resources

- **Dataset:** https://www.kaggle.com/datasets/somumourya/spotifyfeaturescsv-1
- **GitHub Repository:** https://ahsieh53632.github.io/music-attributes-and-popularity/
- **Course:** GSBA 545 - Regression Analysis, USC Marshall School of Business
- **Semester:** Fall 2025

—

Last Updated: January 2026
Project Status: Completed
Team C4 - USC Marshall School of Business