What Makes Songs Popular on Digital Streaming Platforms?

(COMP3125 Individual Project)

Colby Fournier
Data Science Fundamentals
(COMP 3125)

***Abstract*—This project studies what makes songs popular on streaming platforms like Spotify, and takes into consideration audio features such as energy, tempo, and more. Using a dataset that categorizes songs by their features, I was able to determine which features correlate to more popular songs on Spotify. The resulting data shows which features and genres perform better on streaming platforms.**

***Keywords*—music analysis, streaming platforms, audio features, popularity prediction, genre analysis**

I. Introduction (*Heading 1*)

Digital streaming platforms have revolutionized the way people listen to music in the modern era. Therefore, artists must understand what makes songs popular so that their music can reach a wider audience and produce more revenue. This project analyzes data taken from Spotify to find similarities between song features. Some of the questions I wanted to answer include: “Which sound features drive a song’s popularity?”, “How long are the most popular songs?”, “How do sound features differ between different music genres?”, and “Which genres perform better than others?” By finding answers to these questions, artists can use these results to create music that strikes a balance between being sonically unique and popular enough to capture more mainstream attention. This project is valuable because understanding what makes a song popular helps artists, record labels, and streaming platforms optimize new music production and marketing strategies, which generates more revenue.

II. Datasets

A. *Source of dataset*

The dataset used for this project was taken from Kaggle.com and is called “Spotify 1 Million Songs”. It contains very comprehensive information about songs spread across numerous different genres, also considering their features and popularity score. The data was collected from an application using Spotify’s Web API, which gave the author of the dataset access to the information needed about the songs collected.

B. *Characteristics of the datasets*

The dataset comprises over 1 million songs, providing opportunities to gather more diverse data. Each song has a list of associated characteristics that make every song unique, as listed in the dataset. The list includes:

- Danceability: how danceable the song is
- Energy: intensity of the song/how powerful it feels
- Acousticness: measures the amount of acoustics in a song
- Instrumentalness: ratio of instruments to vocals in a song
- Valence: How happy/upbeat a song feels
- Tempo: How fast a song is
- Loudness: How loud a song is
- Speechiness: How many spoken words are in a song
- Duration: How long the song is
- Popularity: Measured from 0-100, higher scores mean a song is more popular on Spotify
- Genre: musical style (pop, country, hip-hop, etc.)

This dataset was chosen for its large sample size and diverse genre selection, which makes it more representative and accurate to modern music trends. Steps taken for data cleaning include removing songs with missing features, deleting duplicate records, and converting song durations to minutes, as they were originally listed in milliseconds.

III. Methodology

To answer each question, many different Python and data visualization methods were used to obtain the best visualization that matches the data requested.

A. *Pearson Correlation Analysis*

- Assumptions: Implies there are no major outliers in the data, and that every entry has a linear relationship.
- Advantages: Easily highlights what sound features are the most important to finding the most popular songs
- Disadvantages: Does not handle more complex relationships between data points, only straight-line relationships.
- Why It’s Effective: A simple way to convey which sound features are the most important to making a song popular.
- Python Implementation: Uses pandas to find correlations between data entries and seaborn to create the data visualization using a heatmap, emphasizing which correlations are most important.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

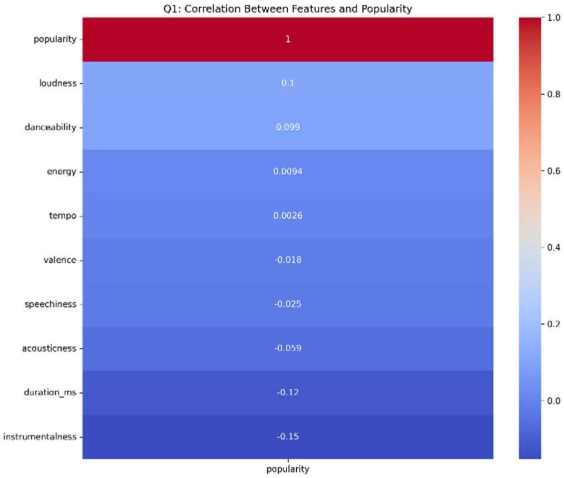
B. Linear Regression Analysis

Model is used to show the relationship between a song's popularity and multiple different sound features at the same time.

- Assumptions: Linear relationship between variables, independent observations, normal distribution, consistent prediction errors, low correlation between audio features
- Advantages: quantitative predictions and shows relative importance while controlling for other features. Also can identify which sound features have the highest statistical effect on a song's popularity.
- Disadvantages: Can be affected by outliers and audio features that are too similar, also only works well with straight-line relationships, like the Pearson Correlation Analysis
- Why It's Effective: Best method for showing how audio features work together in predicting popularity, also good for predictive modeling
- Python Implementation: SciKit linear regression module to create the model, fit() method for training and predict() method to test performance

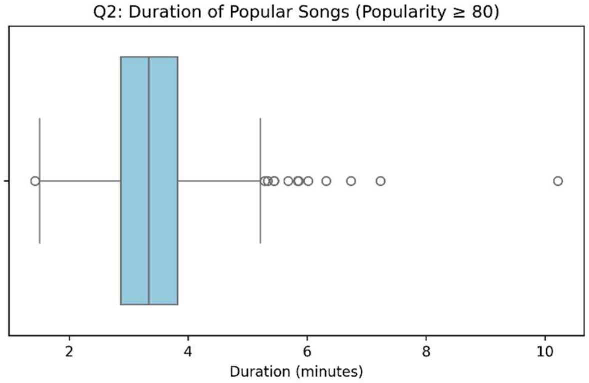
IV. Results

A. Question 1: Which sound features drive a song's popularity?



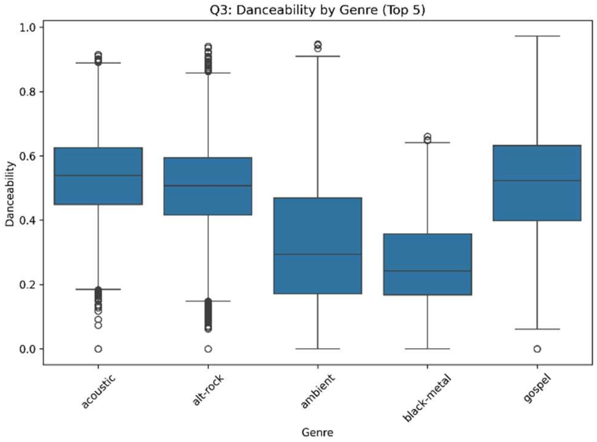
The heatmap shows which song features correlate with the highest song popularity scores. Songs with higher loudness, danceability, and energy scores all tend to produce more popular songs.

B. Question 2: How long are the most popular songs?



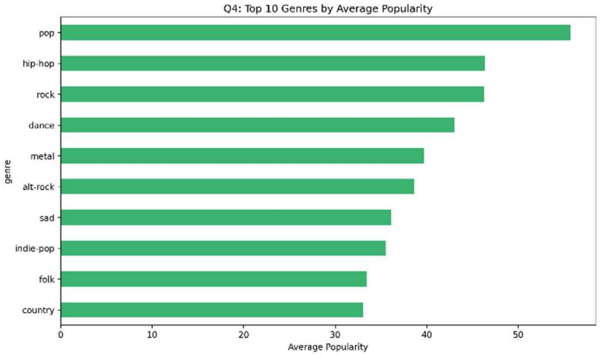
Given a song has a popularity score of at least 80, most of them are between 3 and 4 minutes long, with some outliers being less than 2 minutes long, and others being as long as 7 minutes. This suggests that typical radio-friendly remain in the 3-4 minute range, which likely remains consistent with past trends.

C. Question 3: How do sound features differ between different music genres?



One of the most important sound features collected, danceability, is compared across multiple different genres of music. EDM had the highest danceability score among the compared genres, while country had the lowest danceability score.

D. Question 4: Which genres perform better than others?



Pop music had the highest average popularity score among genres, while country music had the lowest score. This shows that pop music greatly outperforms other genres, likely because it is the most mainstream music genre.

V. Discussion

While working on this project, I felt as though I could not utilize all the data that I collected, which I believe is due to the dataset used being so large. Additionally, I believe some important factors were not taken into consideration to determine what makes a song popular. This might include things like the artist and their level of fame, the lyrical content of the song, marketing for the song/artist, and the number of times a song was played. Future studies should include all this information because I believe that they would be more influential in finding what makes songs popular on digital streaming platforms. These findings align with past studies showing that pop and EDM songs are generally more popular on digital streaming charts (Yee & Raheem, 2022). However, audio features alone cannot define what songs will be the most popular, as outside factors like an artist's fame and proper marketing play an equally important role.

VI. Conclusion

Features like loudness, danceability, and energy play a role in determining if a song becomes popular on digital streaming platforms like Spotify, which this project reflects exactly. Artists could use these features to attempt to gain more mainstream recognition. However, the weak correlations indicate that there are additional factors that go into determining mainstream success. For instance, a lesser-known artist who utilizes sound features that theoretically make up a popular song will likely not gain the mainstream attention they're seeking because of more popular artists doing the same thing. While the data suggests utilizing some of the features studied to make the "ideal" song, there are other, more subjective factors that arguably have a much larger effect on a song's popularity. Future work can combine this data along with social and marketing data for a wide number of artists to create more comprehensive prediction models.

References

[1] A. Joshi, "Spotify_1Million_Tracks," *Kaggle.com*, 2023. https://www.kaggle.com/datasets/amitanshjoshi/spotify-1million-tracks/data

[2] Y. K. Yee and M. Raheem, "Predicting Music Popularity Using Spotify and YouTube Features," *Indian Journal Of Science And Technology*, vol. 15, no. 36, pp. 1786–1799, Sep. 2022, doi: https://doi.org/10.17485/ijst/v15i36.2332.

[3] J. S. Gulmatico, J. A. B. Susa, M. A. F. Malbog, A. Acoba, M. D. Nipas, and J. N. Mindoro, "SpotiPred: A Machine Learning Approach Prediction of Spotify Music Popularity by Audio Features," *IEEE Xplore*, Mar. 01, 2022. https://ieeexplore.ieee.org/abstract/document/9776765 (accessed Jul. 24, 2025).

[4] P. Beesa et al., "Songs Popularity Analysis Using Spotify Data: An exploratory study," *Vidhyayana-An International Multidisciplinary Peer-Reviewed E-Journal*, vol. 8, no. si7, pp. 211-223, 2023. https://www.researchgate.net/publication/384286217_Songs_Popularity_Analysis_Using_Spotify_Data_An_exploratory_study