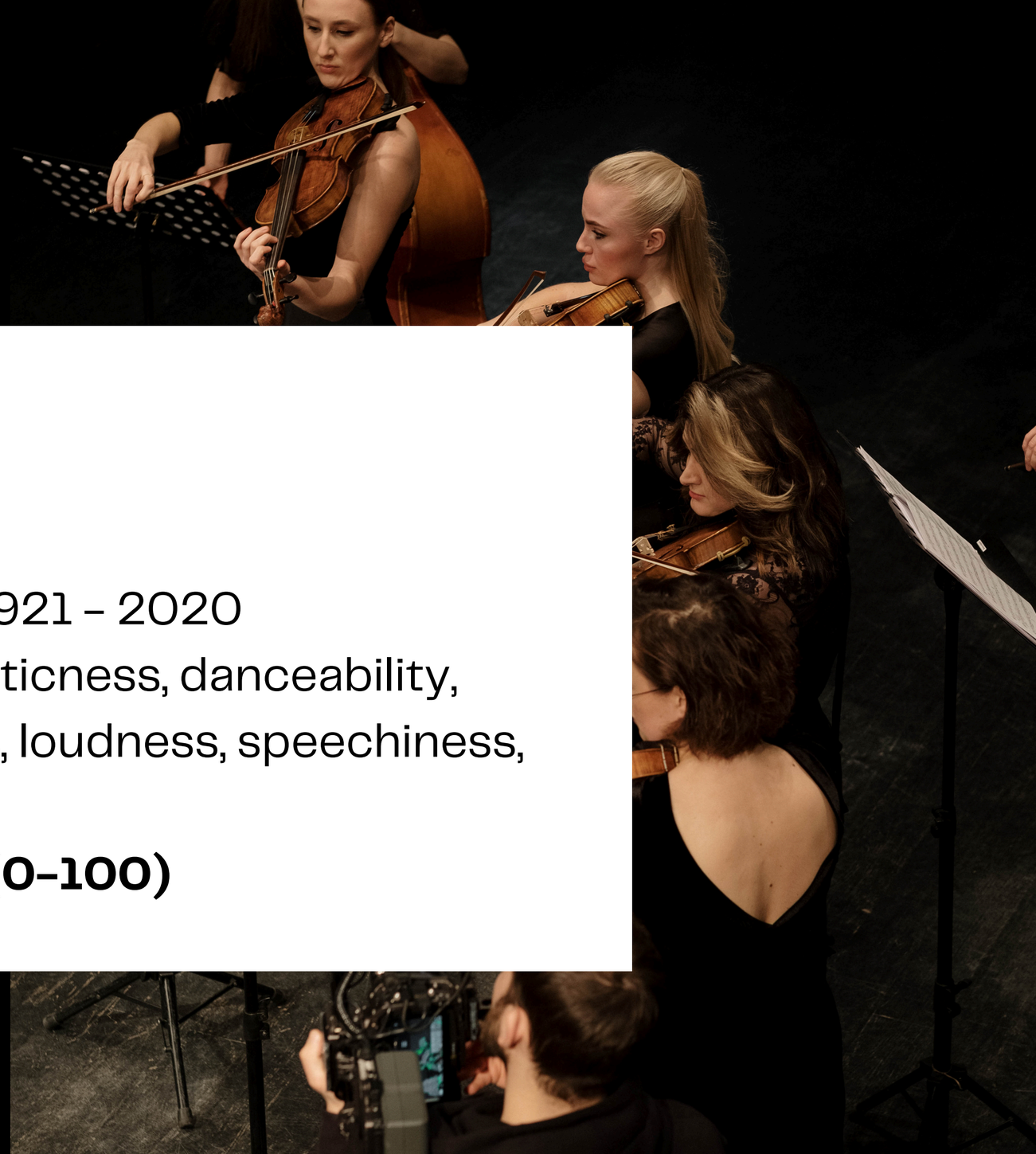

HOW YOU CAN MAKE A HIT ON:
SPOTIFY
Project 2, Group 7: Jennie, Luz, Austen
EXPLORE NOW



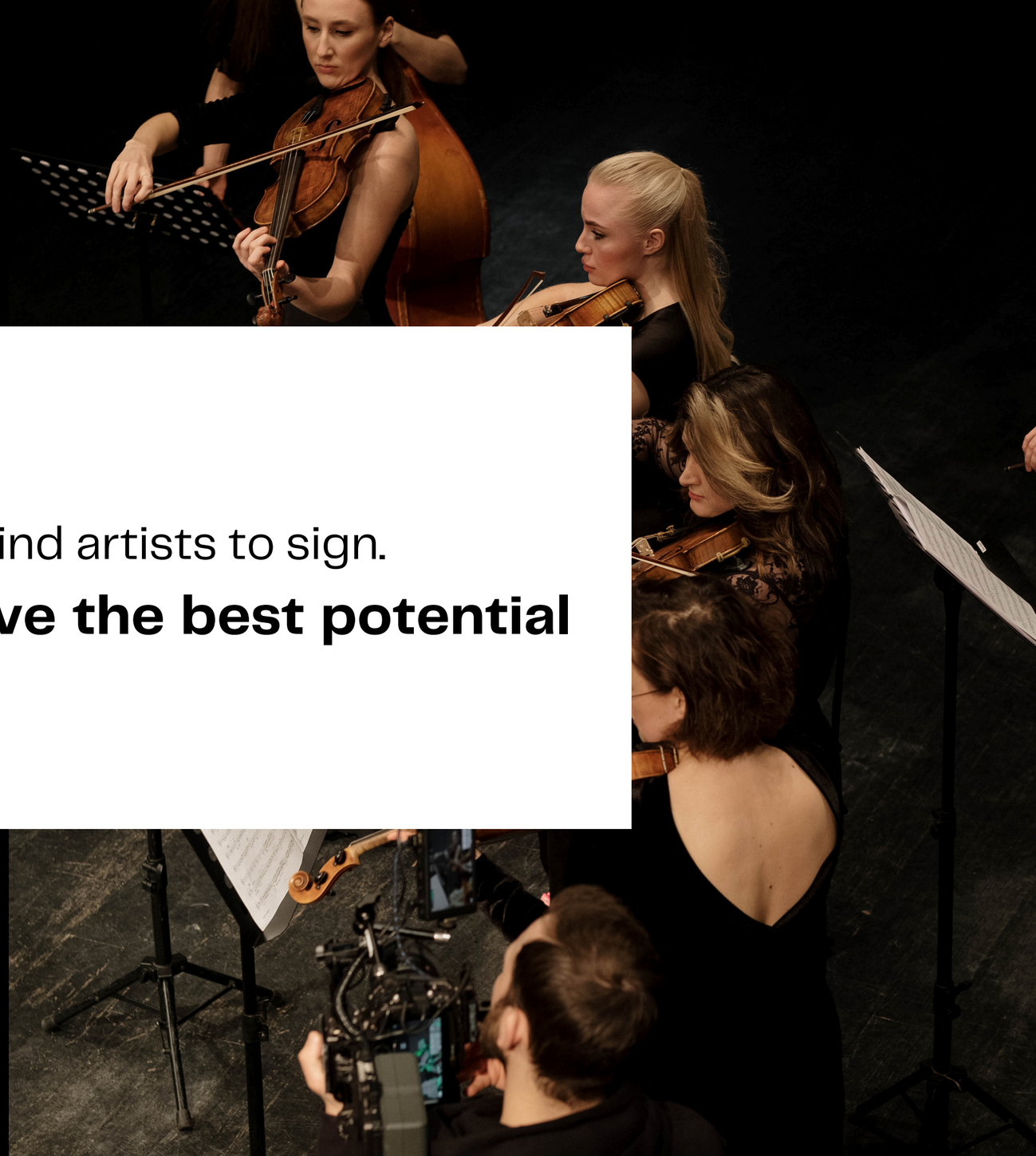
OVERVIEW

- **170k** rows
- **Year** (when the songs released): 1921 – 2020
- **10 song features:** duration, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, valence, tempo (beat per minute)
- Quality metric: **Popularity Index (0–100)**

AUDIENCE

Record executives who are trying to find artists to sign.

–> Which artists and songs have the best potential for success?



SPOTIFY POPULARITY INDEX

TOP 25% of Popularity (75% percentile)

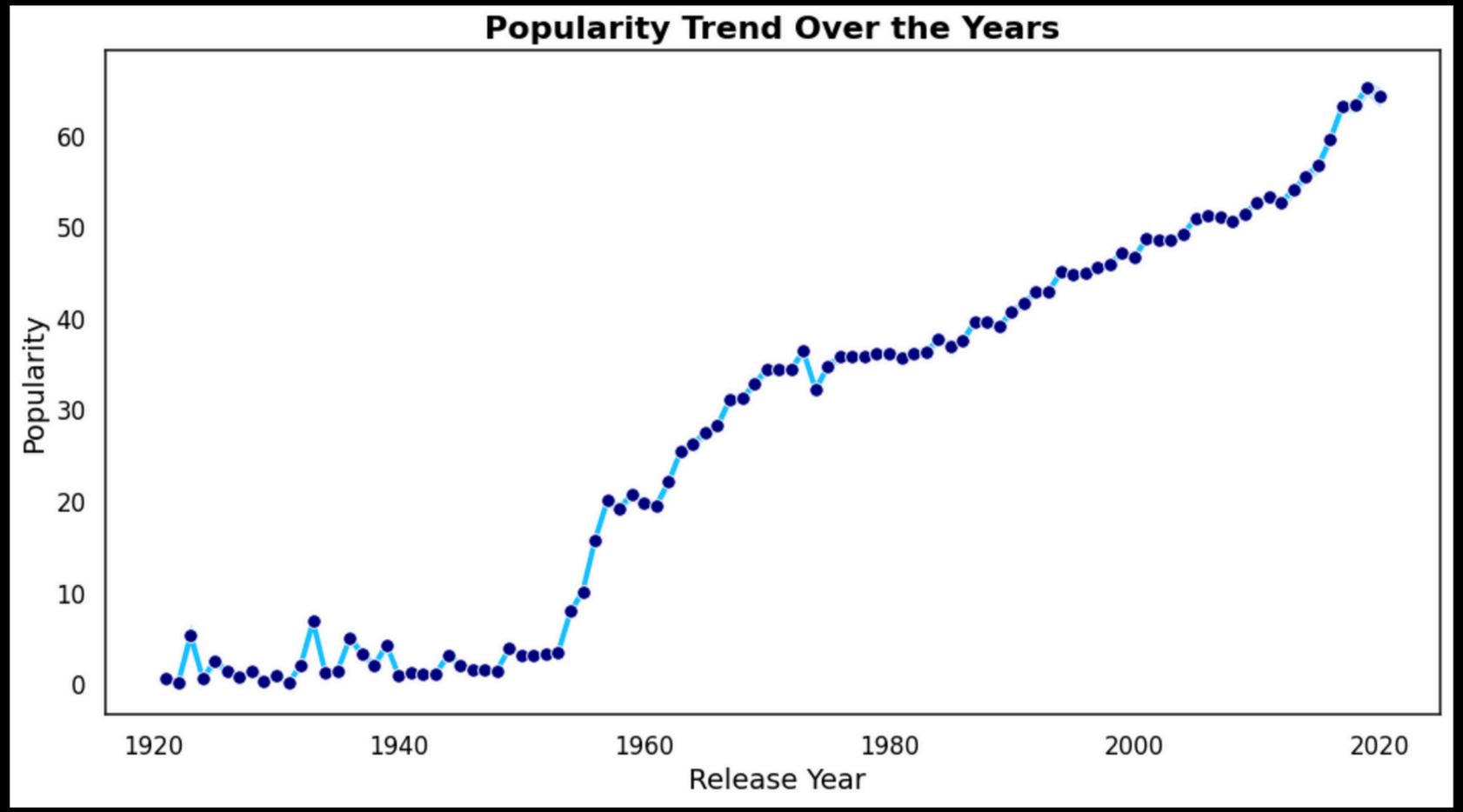
- 38k songs
- Popularity range from **48 - 100**

Calculated by different factors:

- Streams
- Engagement
- Recency
- Playlist Inclusion

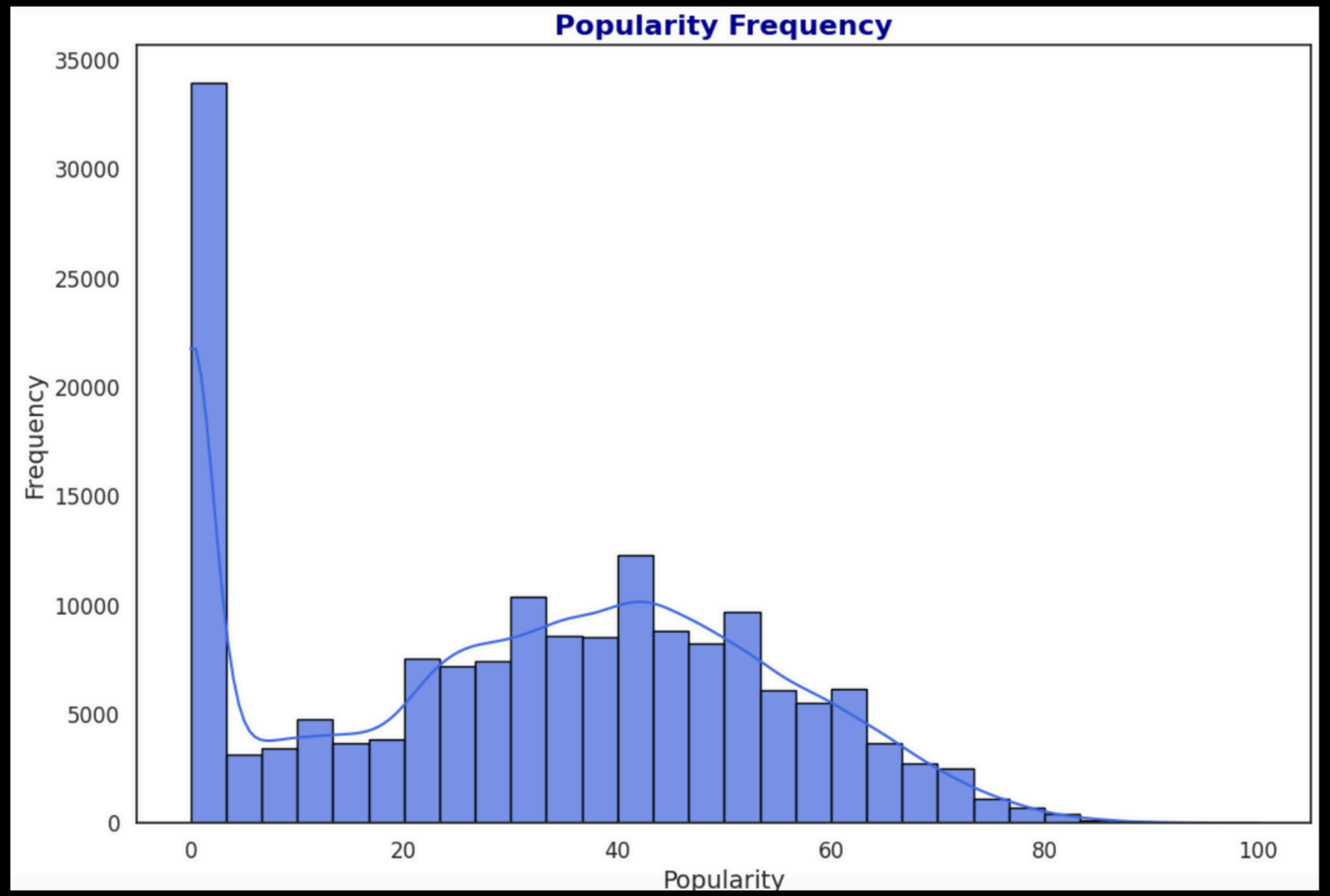
Success = Popularity >= 50

POPULARITY OF RELEASE YEARS



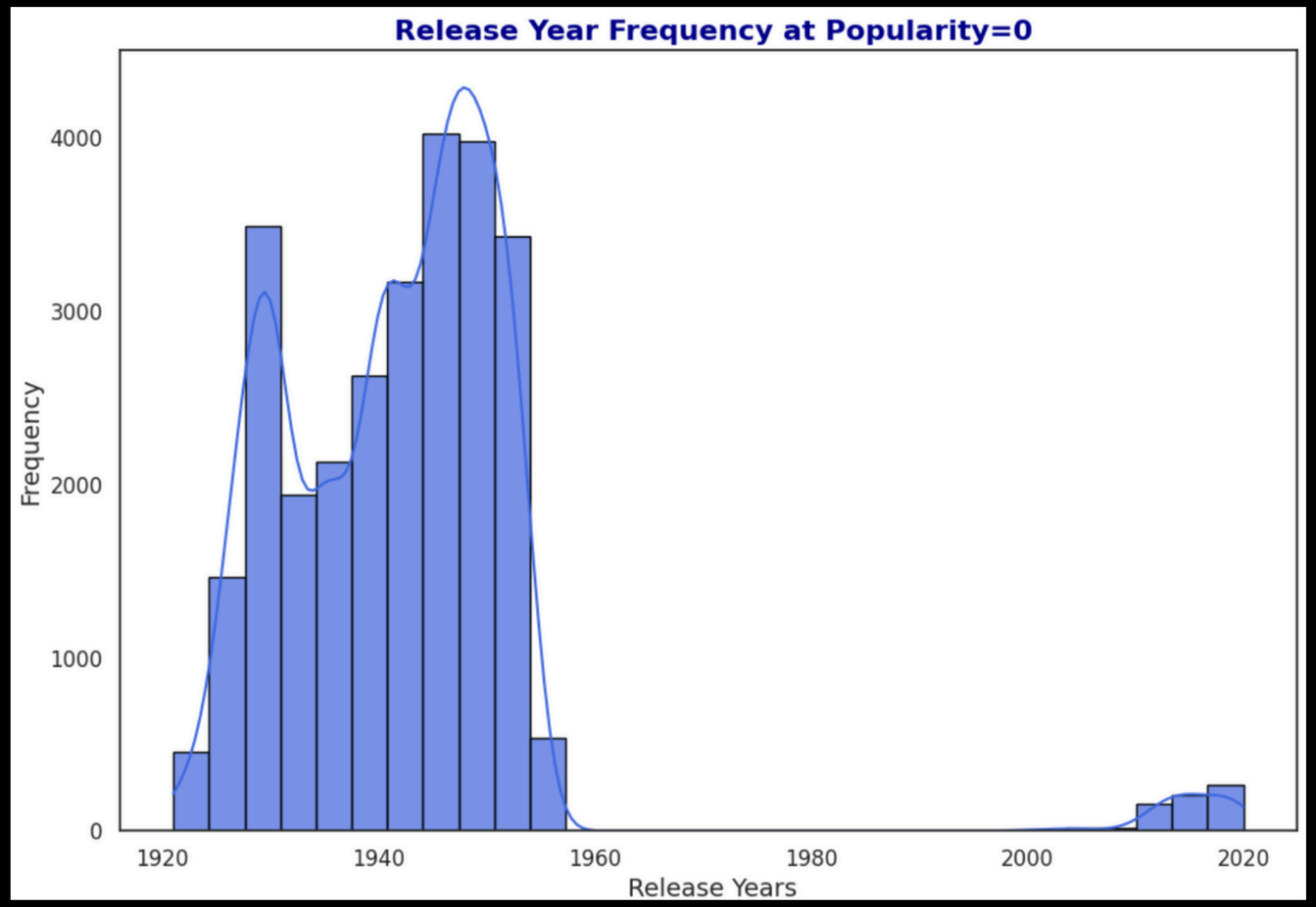
Popularity has increased in the most recent years, why?
Spotify was lauched in 2011 in the US

SPOTIFY POPULARITY FREQUENCY



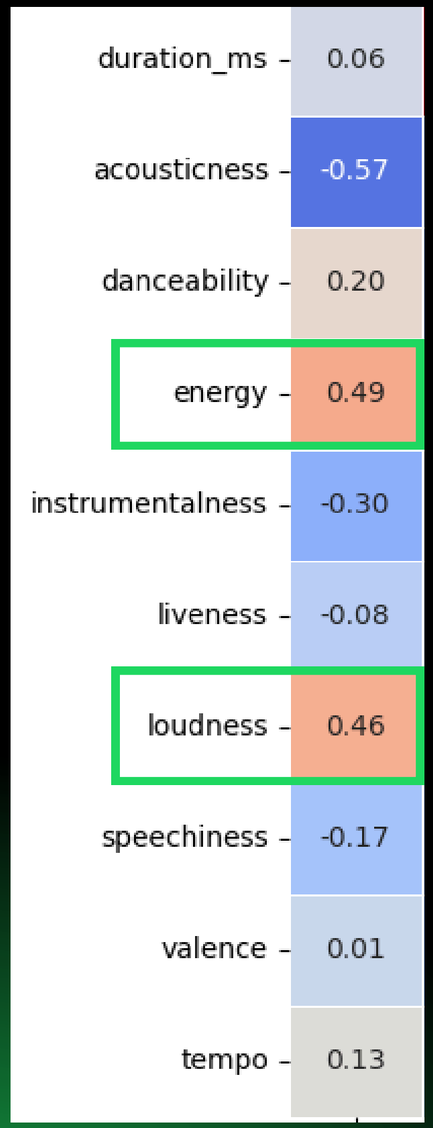
Most Frequent Popularity index is zero, why?

SPOTIFY POPULARITY FREQUENCY



Confirmation that songs with older release dates have lower popularity

WHICH FACTORS ARE MOST IMPORTANT FOR SUCCESS?



(Popularity >= 50)

39k songs

23% of the dataset

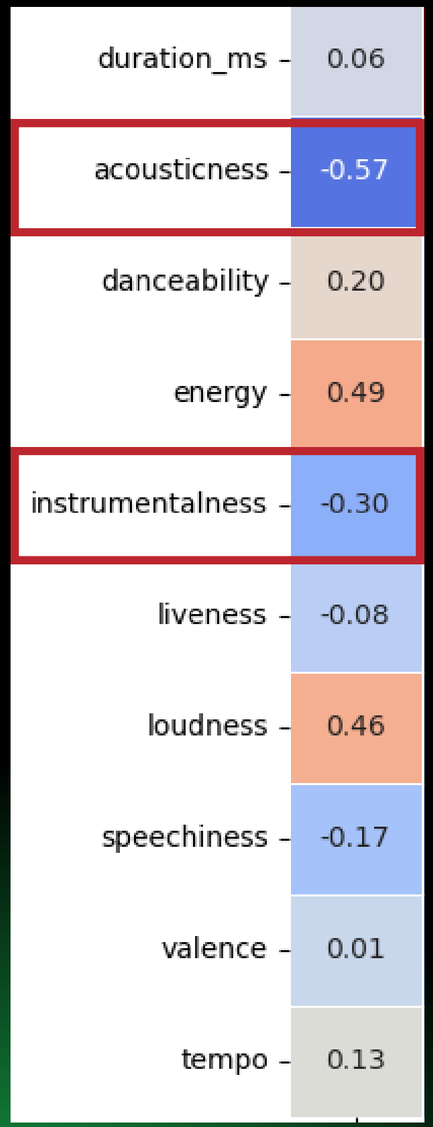
Energy (0.49): represents a perceptual measure of intensity and activity.

Energy vs **Loudness (0.46)** Correlation: **0.78**

Energy vs **Dancebility (0.20)** Correlation: **0.22**

Higher-energy songs tend to be louder, but not neccessary danceable

WHICH FACTORS ARE MOST IMPORTANT FOR SUCCESS?



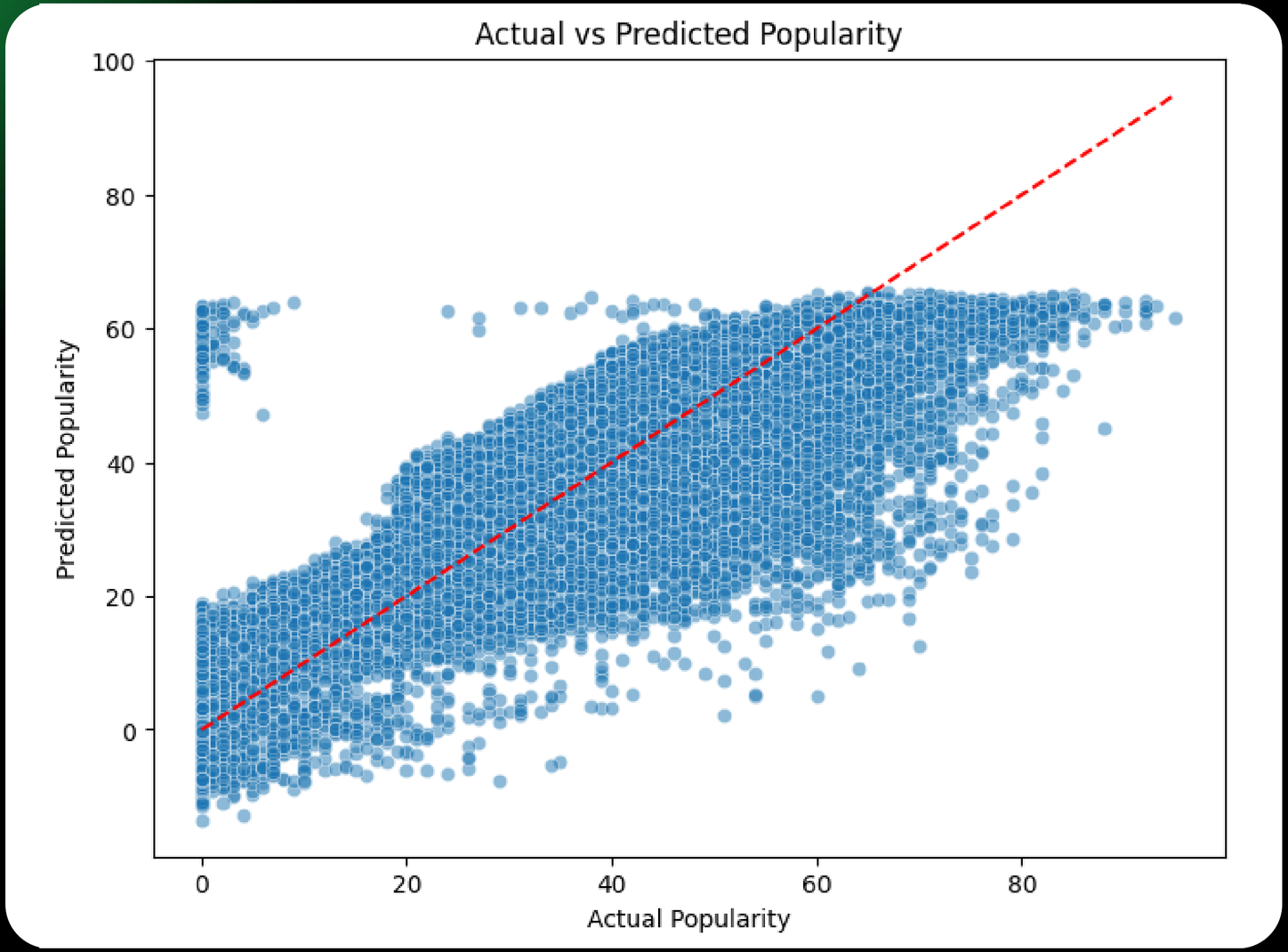
Acousticness vs Instrumentalness Correlation: 0.33

Popular songs are more likely to have **higher energy** and **less acoustic elements.**

HOW CAN WE PREDICT POPULARITY AND OTHER FEATURES?
WHICH MODELS ARE THE MOST ACCURATE?

MODELING

Linear Regression: Song Features

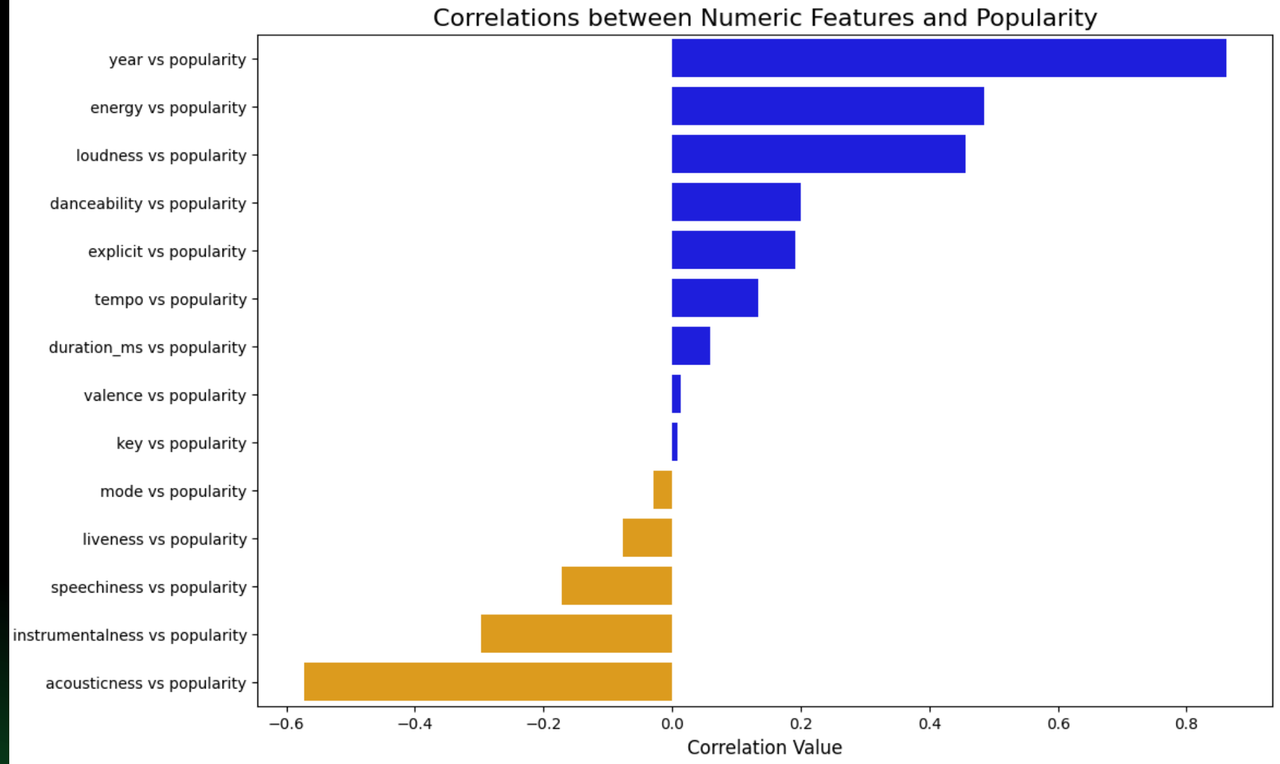


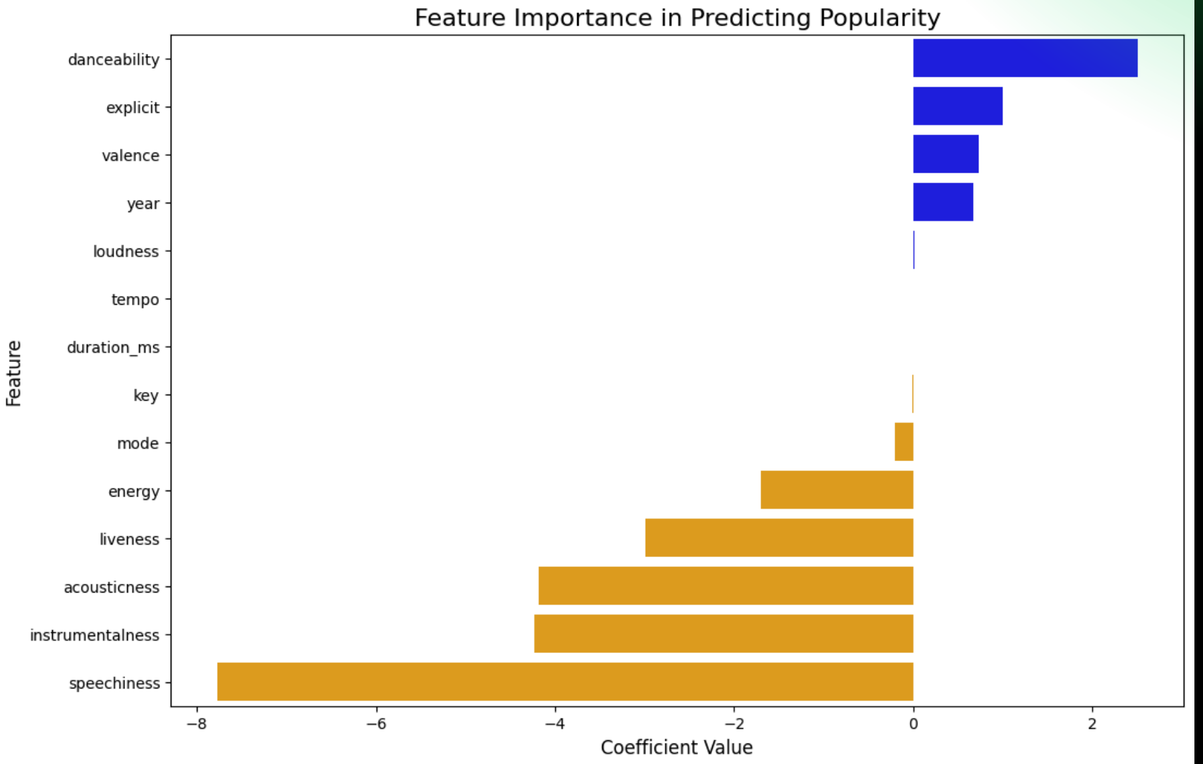
Mean Absolute Error:
7.984

R-squared:
0.759

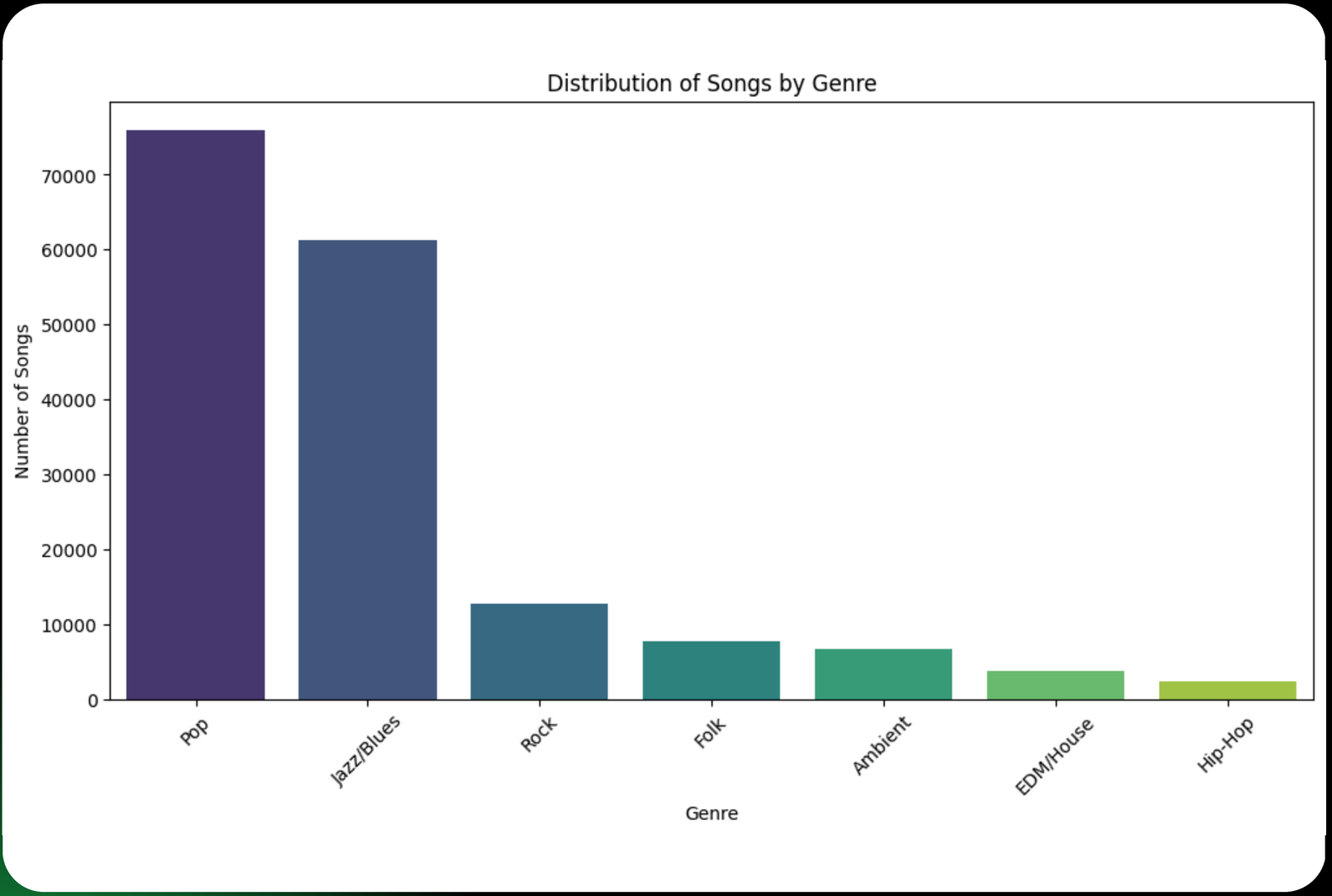
Root Mean Squared Error:
115.112

Linear Regression Features





GENRE ESTIMATE!



Purpose:

Assigns a genre to each song based on its audio features and release year.

Features Used:

Tempo, Energy, Danceability, Acousticness, Speechiness, Instrumentalness, Loudness, Year

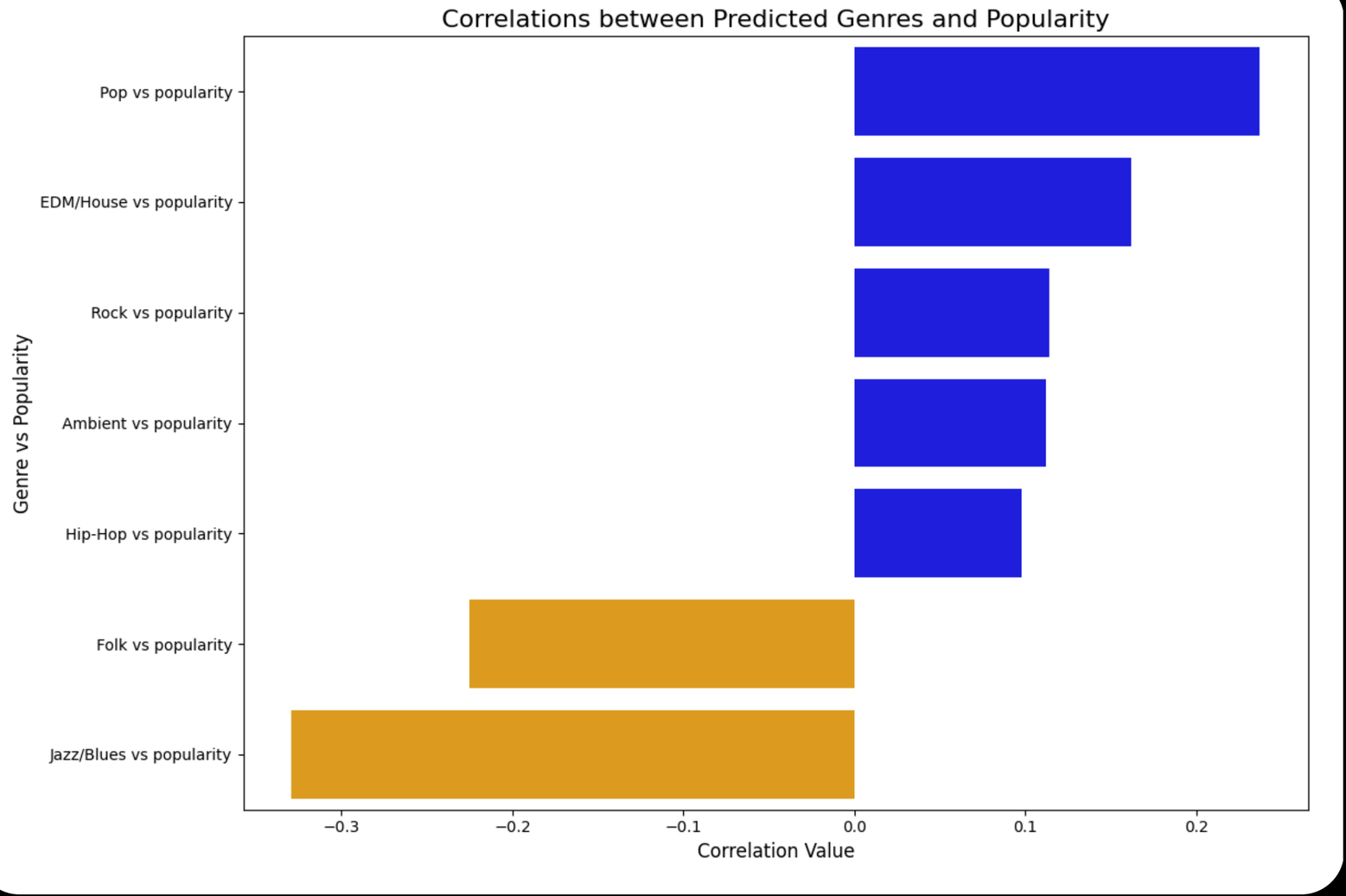
Example – EDM/House:

Year > 1980

Tempo (BPM) is in (120 – 160)

Dancability > 0.65 / 65%

GENRE CORRELATIONS



Pop:

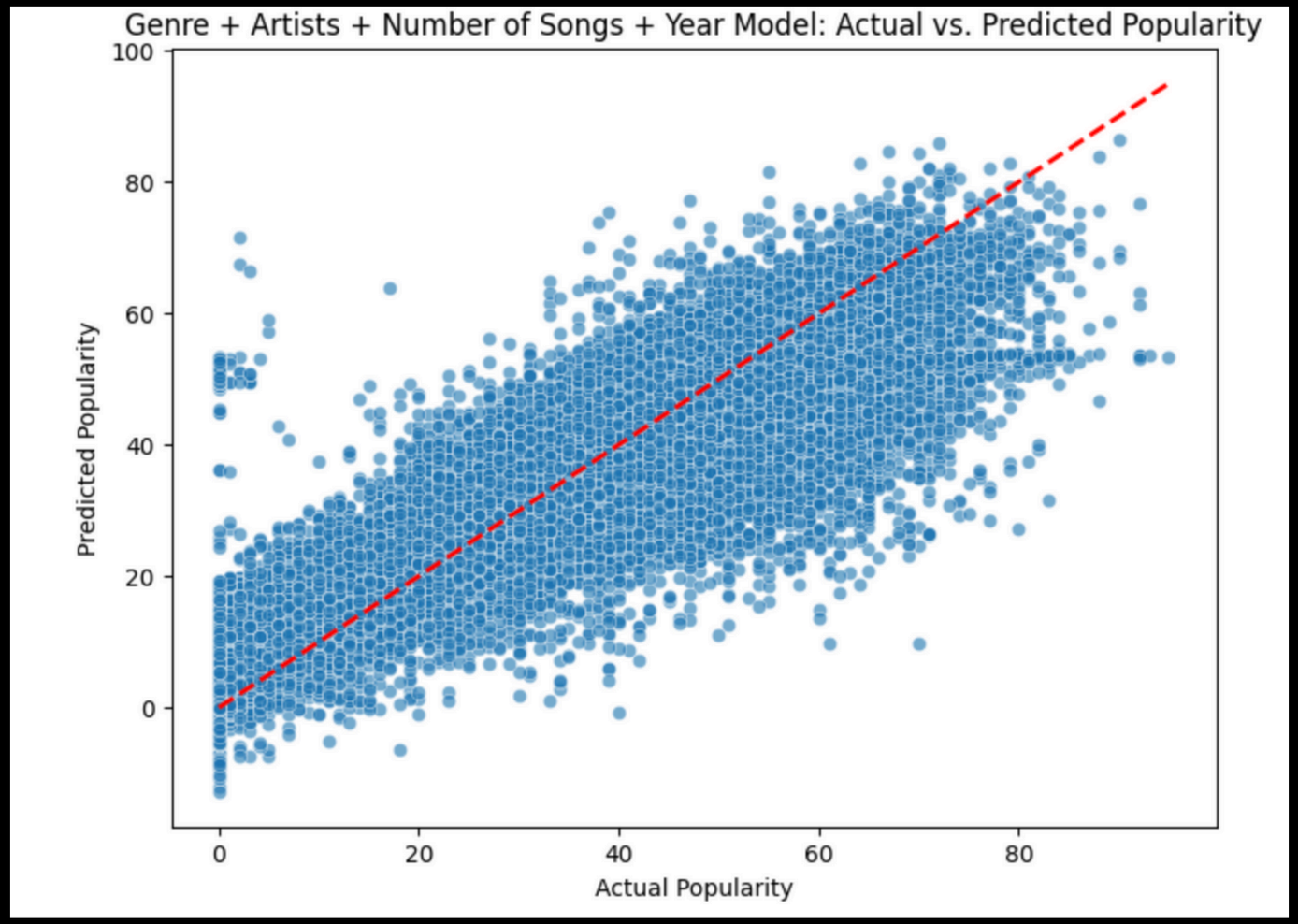
Most Popular Genre

Less Popular Genres:

Older Music

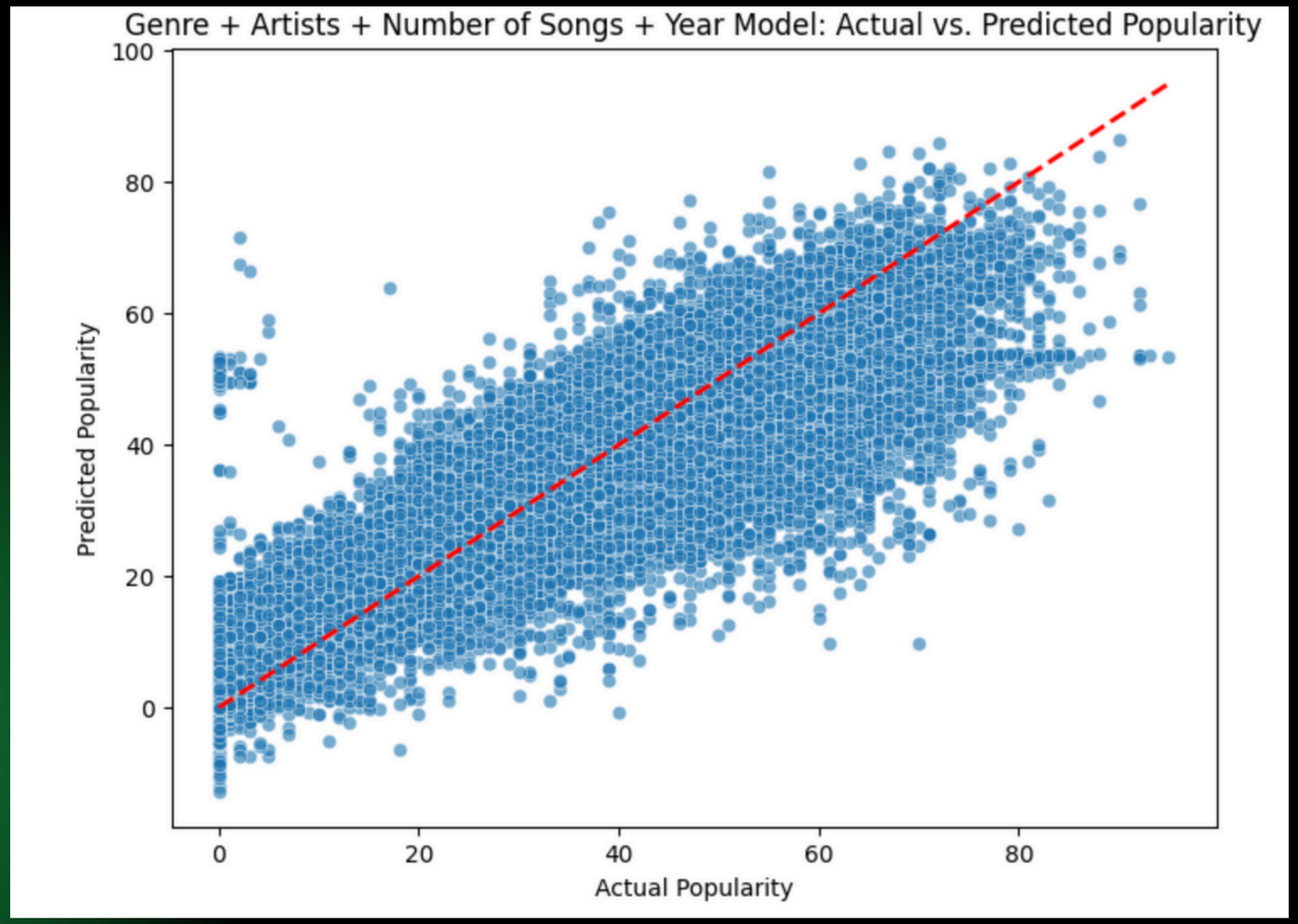
Not All Songs Here are Bad

Linear Regression:
Genre, Artists, Number of Songs, Year



Improves the model's performance compared to using only song-related features.

Linear Regression:
Genre, Artists, Number of Songs, Year



Mean Absolute Error:
6.84

R-squared:
0.805

Root Mean Squared Error:
96.51

Random Forest Regressor

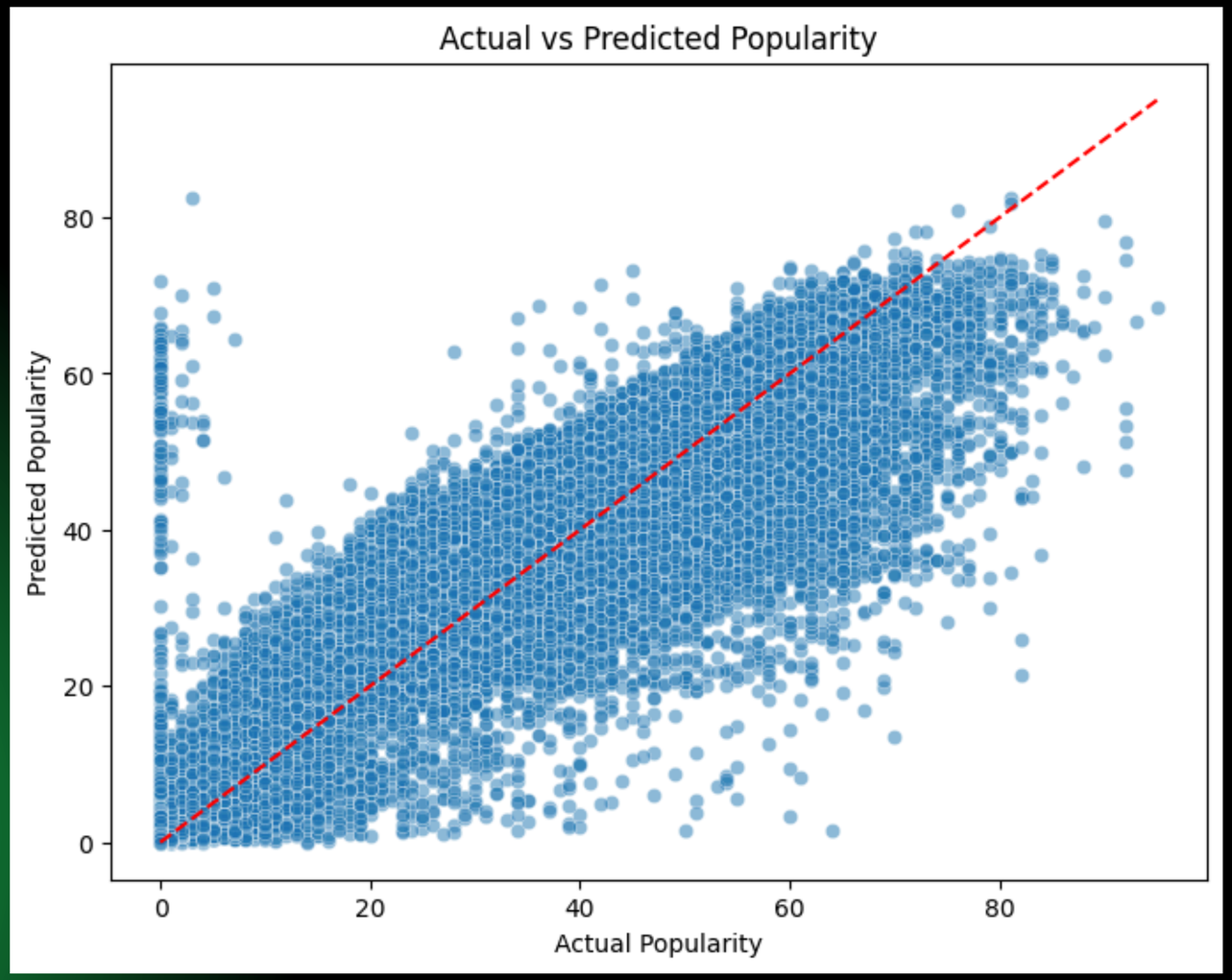
How It Works:

Builds many trees using random subsets of song features (loudness, acousticness, ...)

Each tree "votes" on a prediction; the average of these votes gives the final popularity score

Can capture more complex patterns in data, reduces overfitting compared to a single tree

Random Forest Regressor

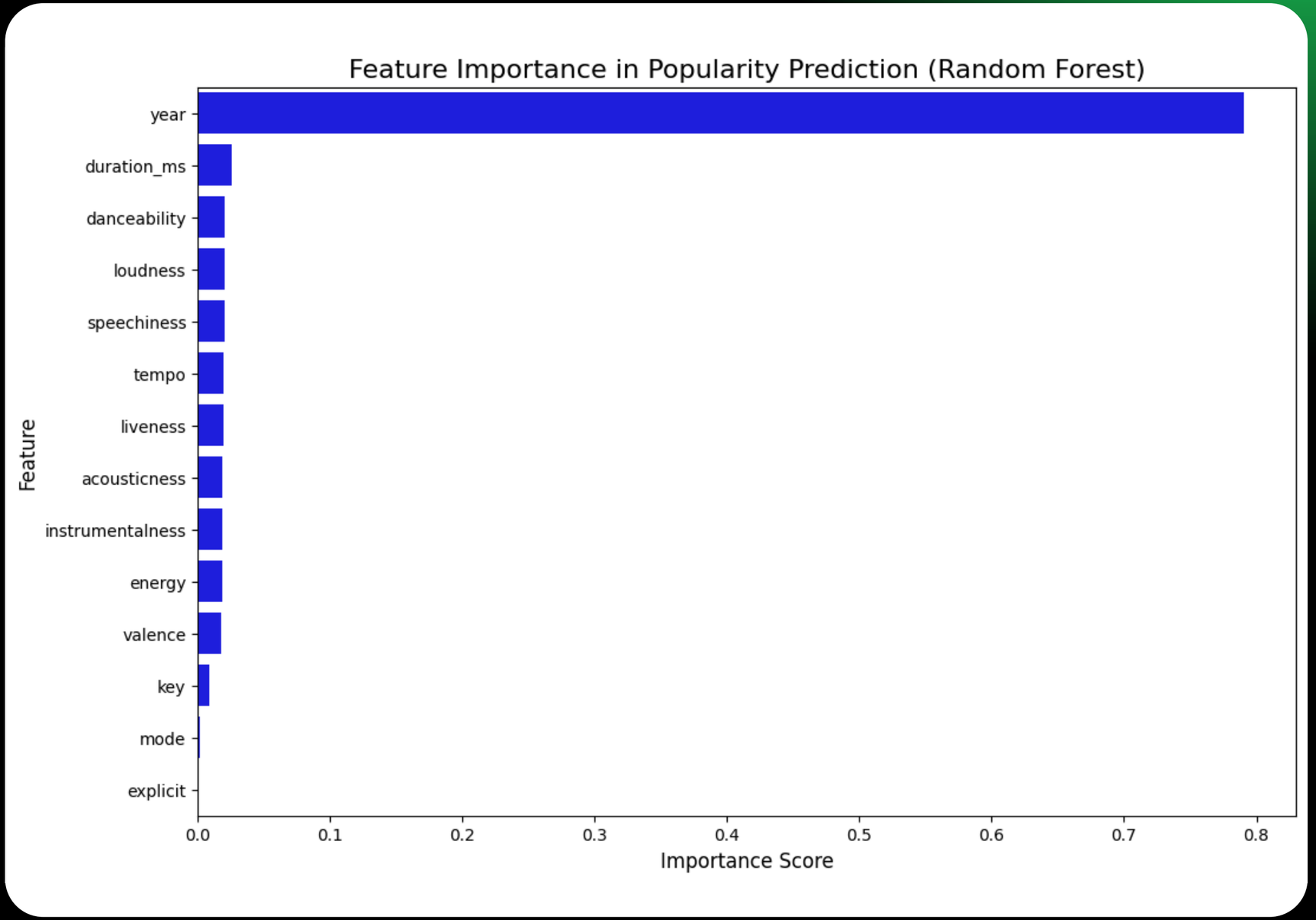


Mean Absolute Error:
6.749

R-squared:
0.809

Root Mean Squared Error:
91.220

Random Forest Regressor Features



CONCLUSION

- **Linear Regression:**
  - **Genre, artist identity, and year** significantly improves the model's ability to predict popularity

- **Random Forest Regressor** (~0.8 R-squared, 6.74 Mean Absolute Value)
  - Importance On **Acoustics, Speechiness, Loudness, and Length**

- **Logistic Regression** (~78% Accuracy)
  - Importance On **Loudness , Valence, and Danceability**

- **Trends** Seen Across Models:
  - Release date year increases likelihood of accurate prediction
- **Song Recommendation** works best when playlist is of the same taste