PREDICTING SONG POPULARITY

Duygu Göksu
April 2, 2024
Capstone Project , Sprint 2

Spotify Popularity Scores

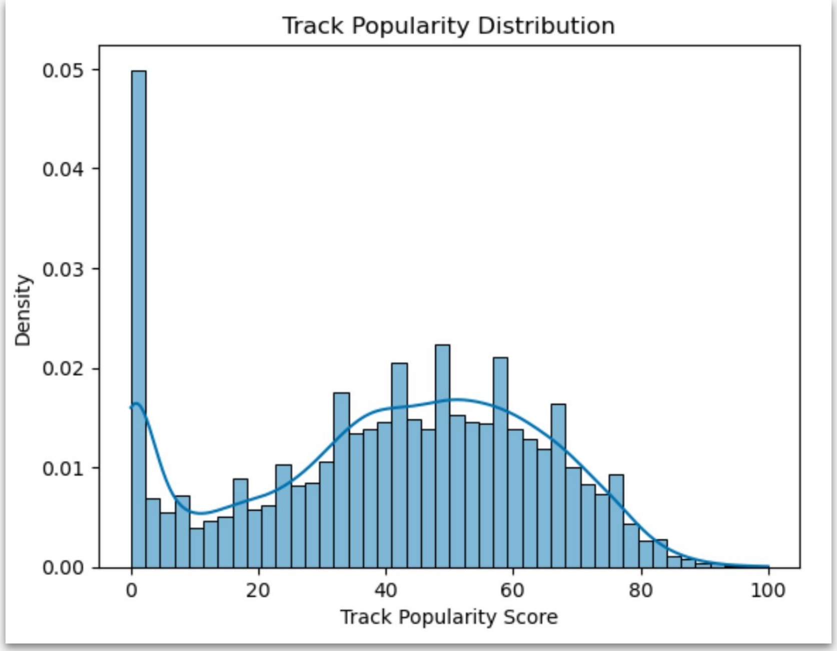
- Range from 0-100 tracking an artist's current popularity across Spotify
- Determined by recent stream count, save/skip rate, number of playlists*
- Artists with a popularity score closer to 50 and above are more likely to be in official Spotify playlists.**
- Very important for both artists and music label companies.

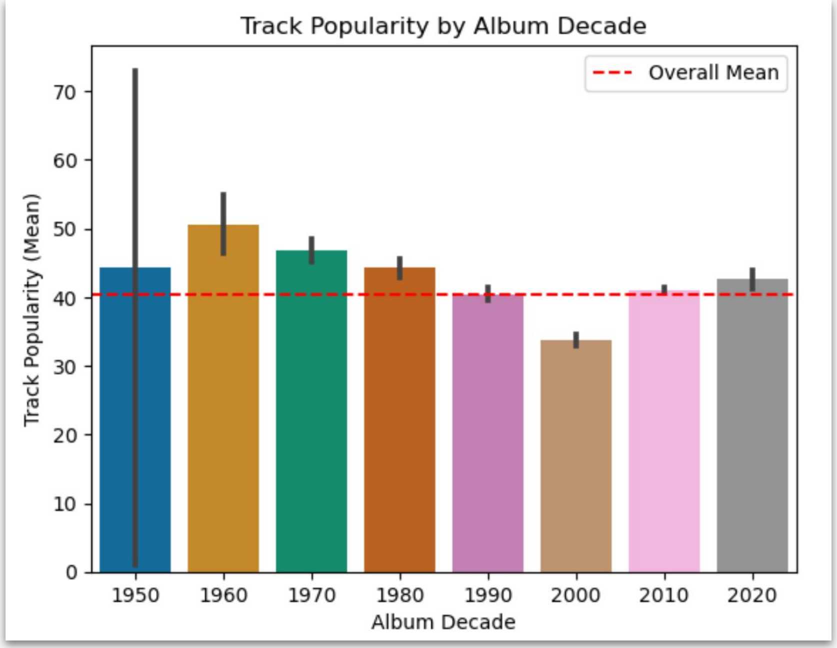
Impact: 11 million artists get paid about $0.003 - $0.005 per stream.***

Sources: *https://www.loudlab.org **https://medium.com ***https://www.searchlogistics.com

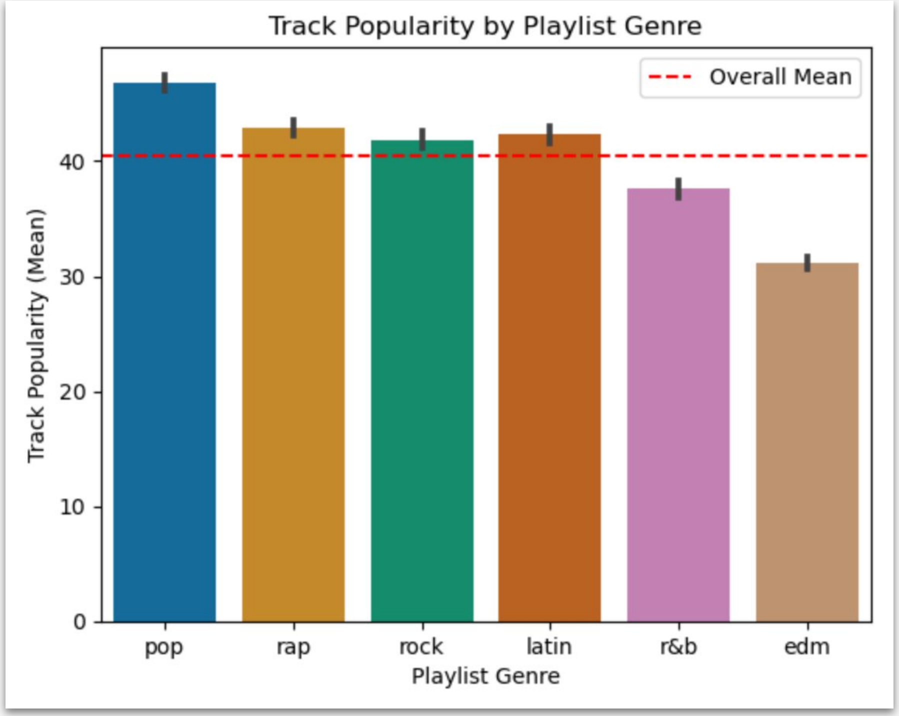
Data

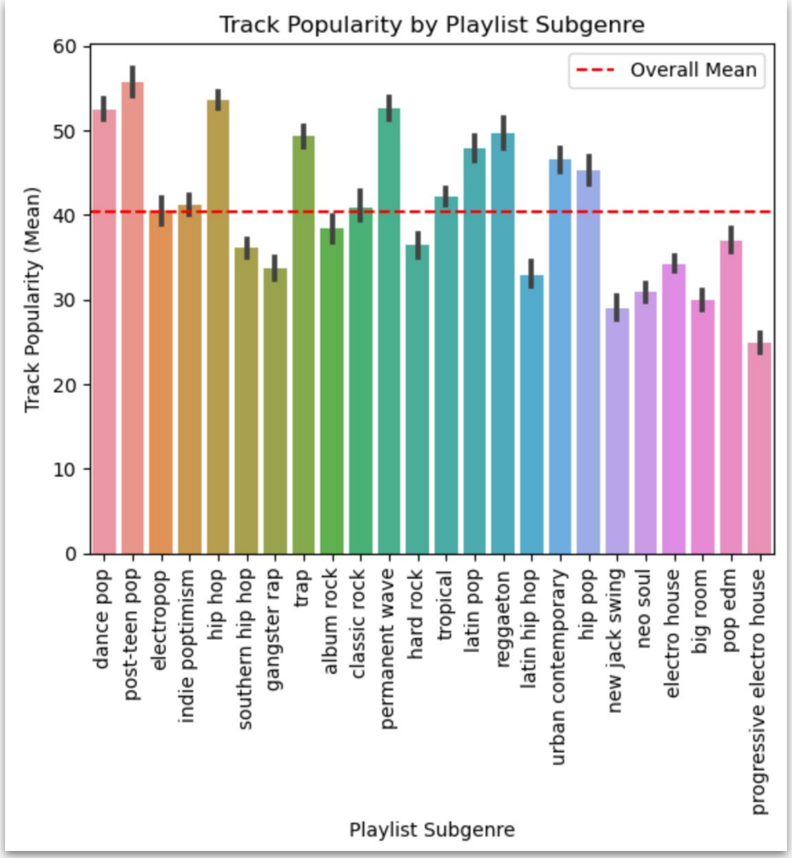
- 30000 Spotify Songs from Kaggle (**~26000** after cleaning)
- Target variable distribution is zero-inflated
- Most songs are from 2010s.



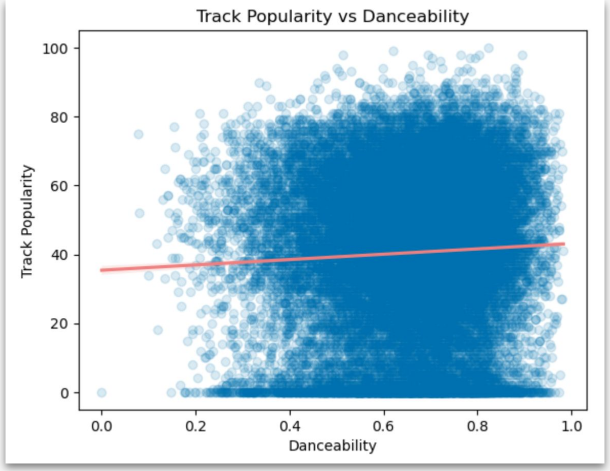


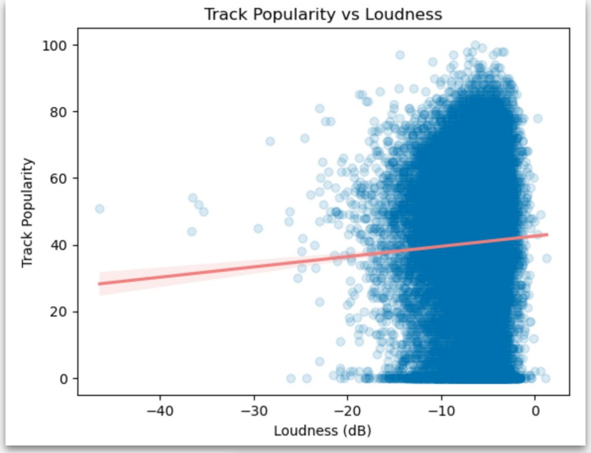
Song (sub)genres in the data

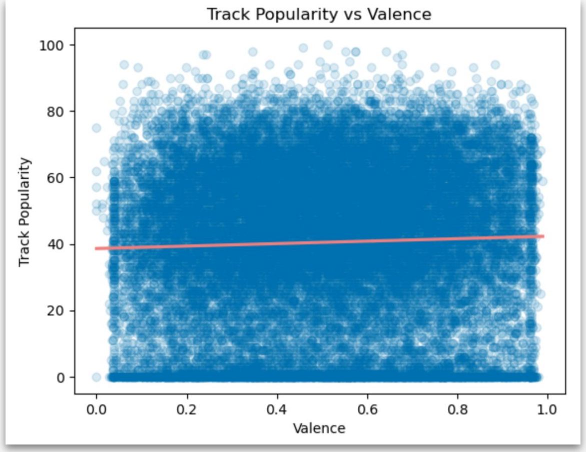


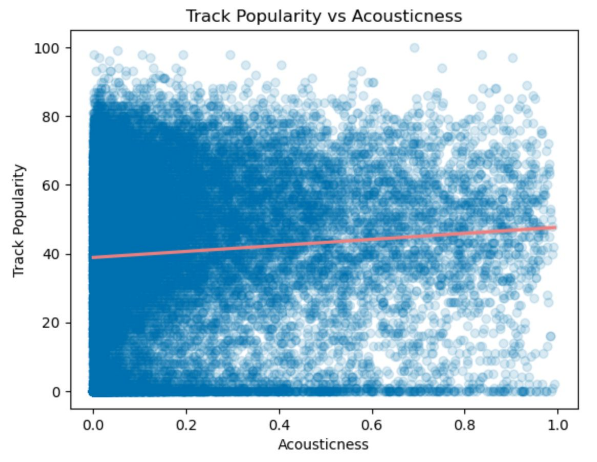


Numeric Features: with Positive Correlation

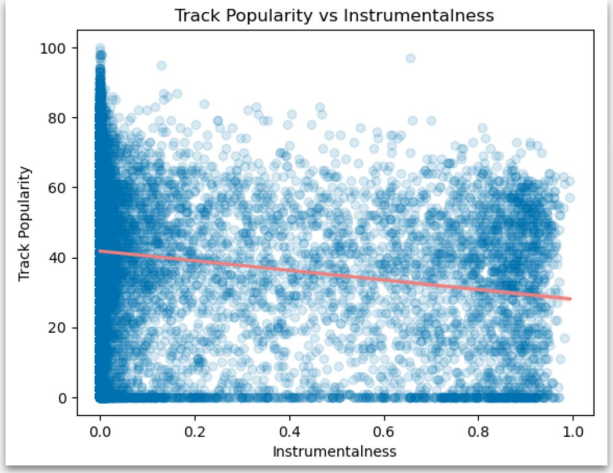


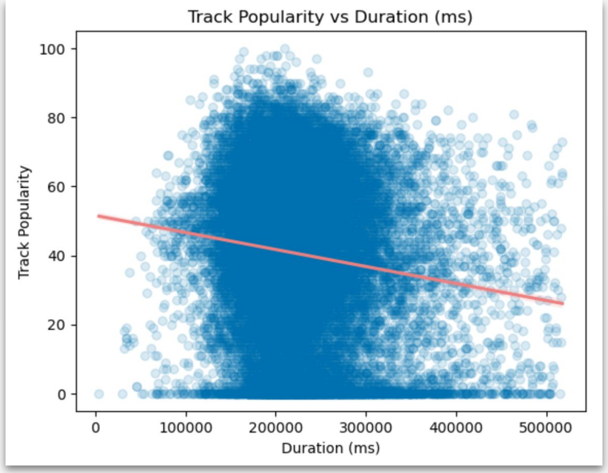


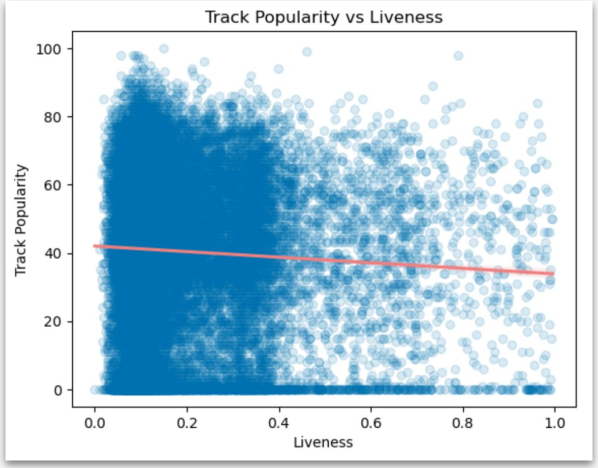




Numeric Features: with Negative Correlation







Modeling

Preprocessing:

- Log transformation on skewed features
- Standard Scaler on features with negative values
- MinMax Scaler on features with only positive values
- No Scaler on one-hot encoded columns
- Dropped 0 and 1 popularity scores in the second model

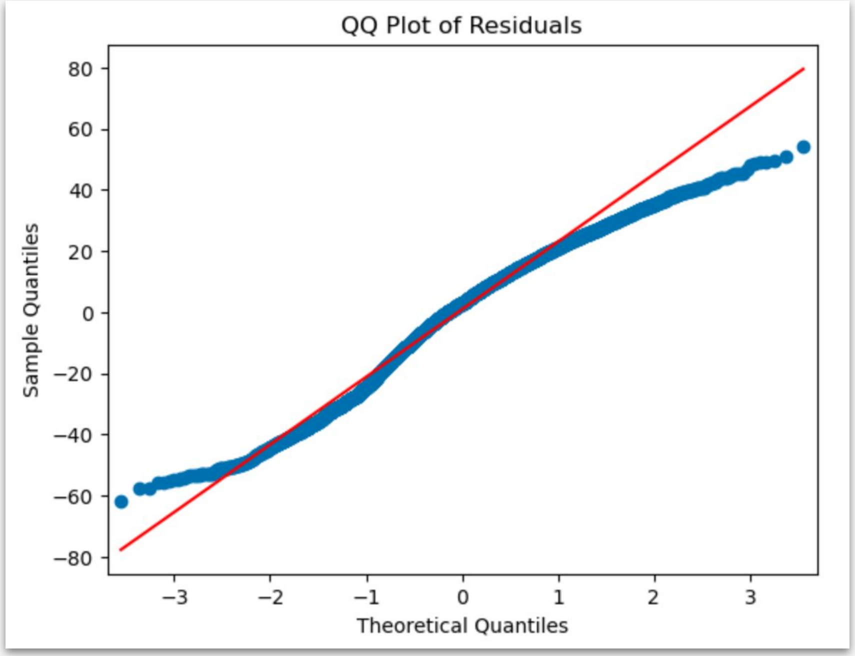
GridSearch (cv=5):

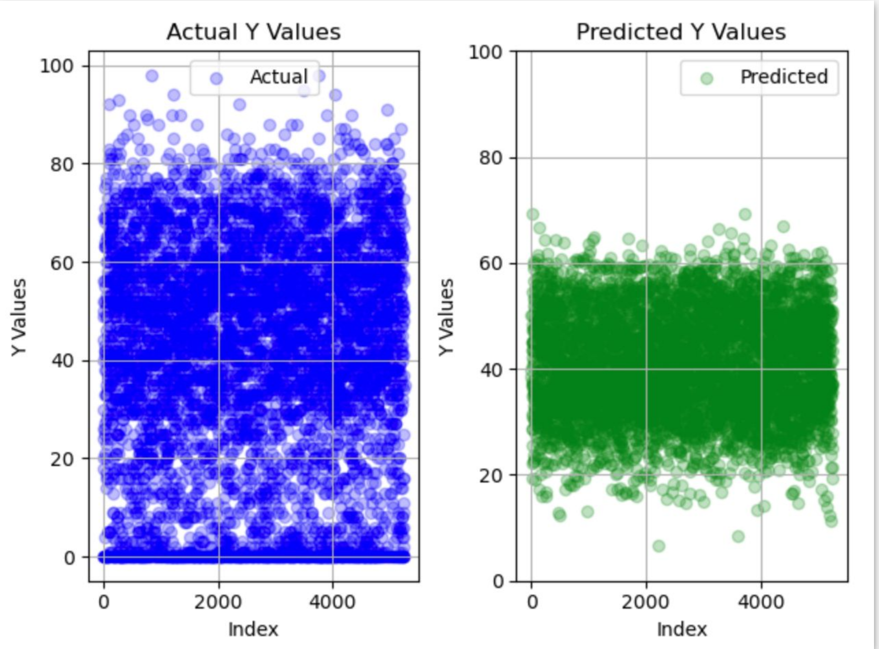
- Dimensionality reduction: PCA (0.9, None)
- Models: Lasso and Ridge Linear Regression
- Model alphas: [0.001, 0.01, 0.1, 1, 10, 100, 1000]
- Solvers for Ridge Linear Regression

Linear Regression Model 1

Lasso with alpha=0.001, and no dimensionality reduction.

R-squared: 0.179 (train), 0.178 (test) - Mean Absolute Error (test): 17.2

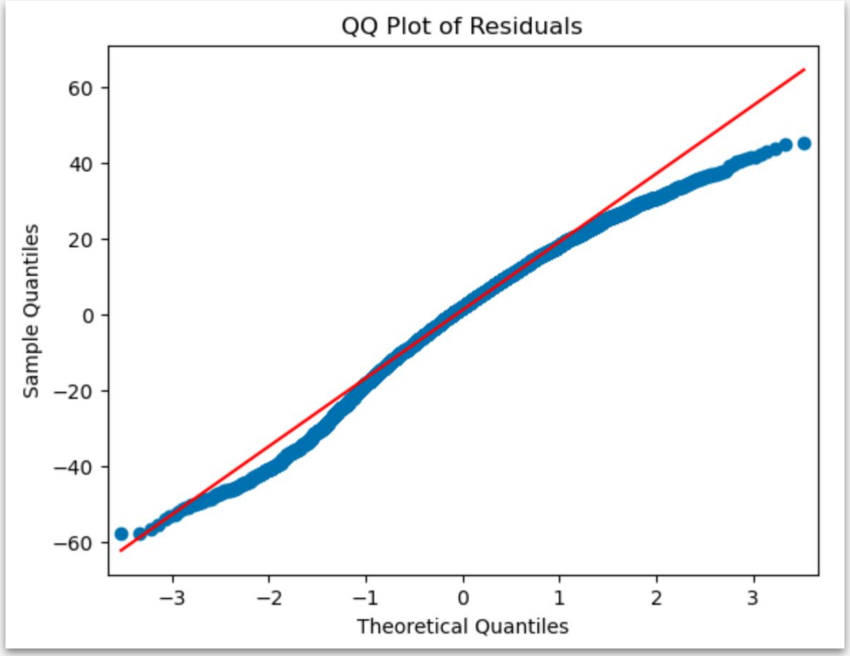


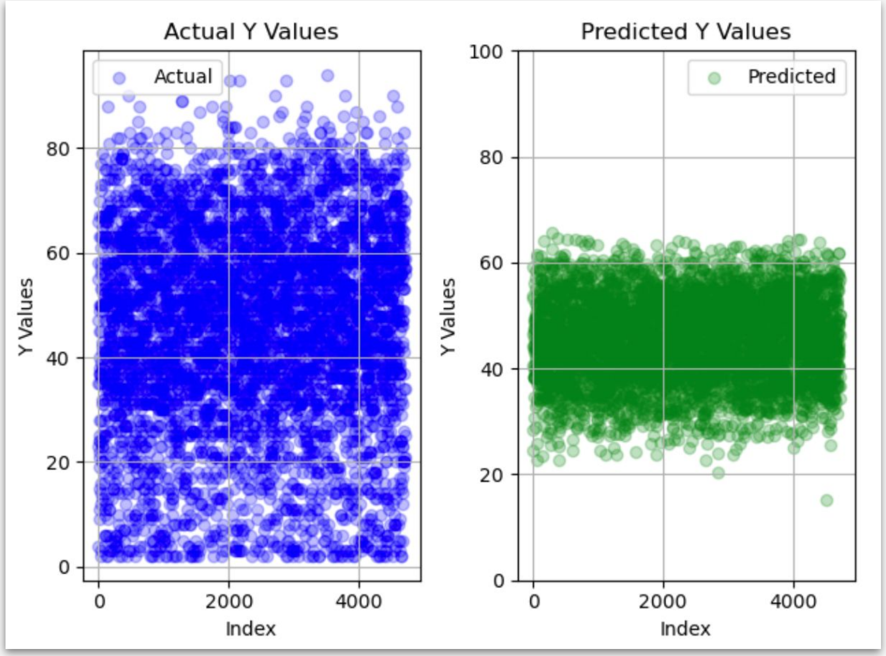


Linear Regression Model 2 (data=popularity>1)

Ridge with alpha=1, solver=lsqr, and no dimensionality reduction.

R-squared: 0.164 (train), 0.159 (test) - Mean Absolute Error (test): 14.5





Next Steps

- I’ve collected lyrics for the songs in the data, using Genius API and lyricsgenius library.
- I’ve cleaned these lyrics a lot. There were many playlists instead of lyrics, as wells random phrases inside lyrics text.
- This process has decreased the size of the data.
- The next step is feature engineering using these lyrics, and
- Building a more advanced model that is better with zero-inflated target values.
- I’ll keep it as a regression problem for now.

TEŞEKKÜRLER!