

Top Spotify Songs in 73 Countries Vladimir Makarenkov

Introduction

Music is a huge constantly changing industry, where the success of a song can depend on several factor. Those factors may be related to current trends in the world of music, but also to listeners.

Goals

The goal was to find out which songs and genres are popular in different countries, analyze some changes in rankings of songs, and, most importantly, predict song popularity based on the information given and machine learning.

Data and methods

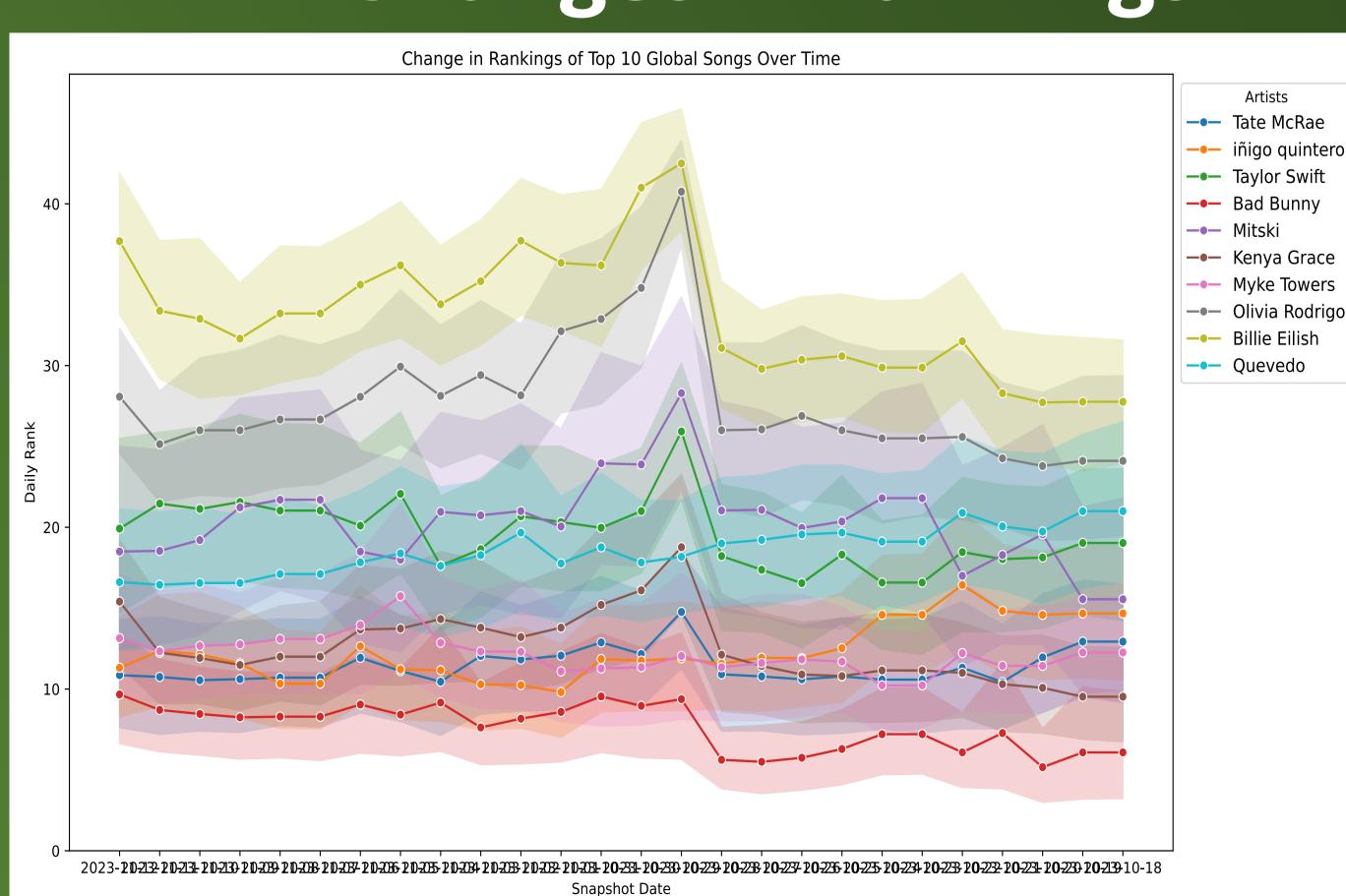
In this project, the Kaggle dataset "Top Spotify Songs in 73 Countries" was used. It contains information about the top 50 songs in 73 countries. The dataset didn't need much modification and was overall ready for use. However, for finding the most popular songs in each country, the 'daily_rank' column had to be modified to find the song with the minimum average daily rank.

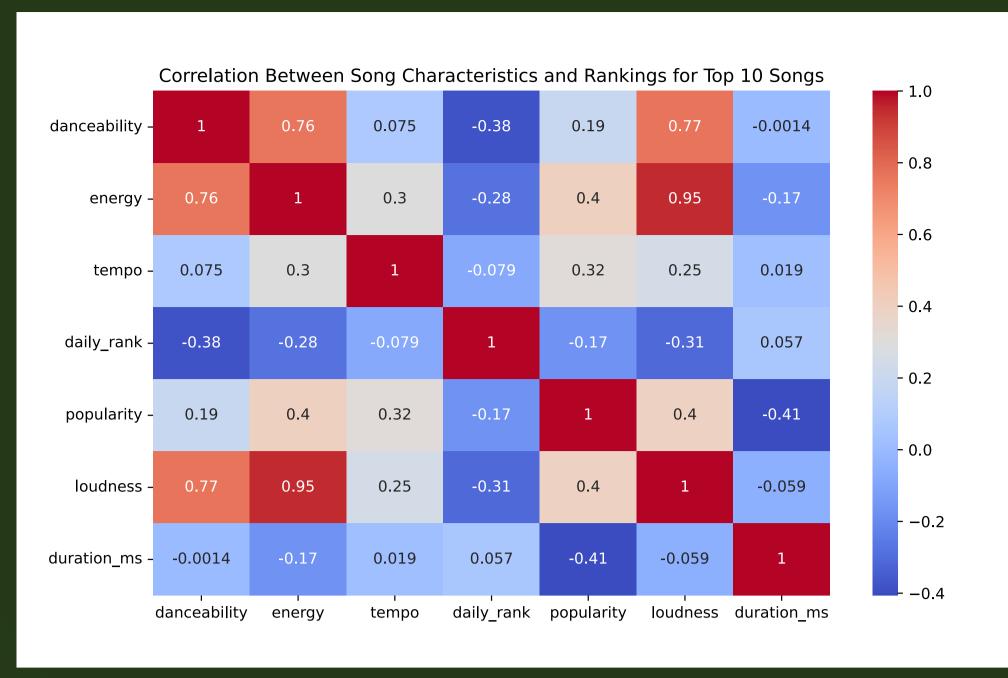
Top Songs and Genres



This table displays songs with the best daily ranking over time (snapshots were taken between 18.10.23 - 12.11.23). To summarize, for most countries, it's their local or similar-to-local song. In other cases, the best-ranked songs were globally viral songs. The top genres that appeared the most were HIP-HOP, POP, K-POP, and Reggaeton. In Latin-American countries Reggaeton dominated, while in Asia, there was more K-POP. In other countries, the results were mixed, but mostly the genres were the same.

Changes in rankings





A subset of relevant song characteristics, including danceability, energy, loudness, daily rank, tempo and popularity is selected. The correlation matrix of these characteristics is computed and visualized as a heatmap. Heatmap shows that there is no really correlation between daily rank and features like energy and danceability. Yet there is some correlation in popularity, loudness, energy and tempo.

Predicting future popularity

Predictive models are implemented using Random Forest, Linear Regression, and Gradient Boosting. The features include danceability, energy, key, valence, and tempo. The dataset is split into training and testing sets. The best-predicted popularity was 83 with Gradient Boosting, 77 with Linear Regression, and 75 with Random Forest. In this situation, Random Forest had the smallest predicted popularity, but its mean absolute error was low as well, so it should still be better than Linear Regression and almost the same as Gradient Boosting. For this popularity, the song had a tempo of 100 bpm, danceability and energy at the 0.8 level, valence at 0.9, and the second song key."

