Spotify: analysis of song popularity with reference to musical attributes and platform presence

Antonio Ciociola
Antonio Andrea Salvalaggio
Emanuele Rossi

Features

General information:
track_name, artist(s)_name, artist_count, released_year, released_month, released_day

Presence and popularity on various platforms:
in_spotify_playlists, in_spotify_charts, streams, in_apple_playlists, in_apple_charts in_deezer_playlists, in_deezer_charts, in_shazam_charts

Musical Attributes:
bpm, key, mode, danceability_%, valence_%, energy_%, acousticness_%, instrumentalness_%, liveness_%, speechiness_%

The Dataset

The dataset contains a comprehensive list of the most famous songs of 2023 as listed on Spotify. It provides insights into each song's musical attributes, popularity, and presence on various music platforms.

Our Goal

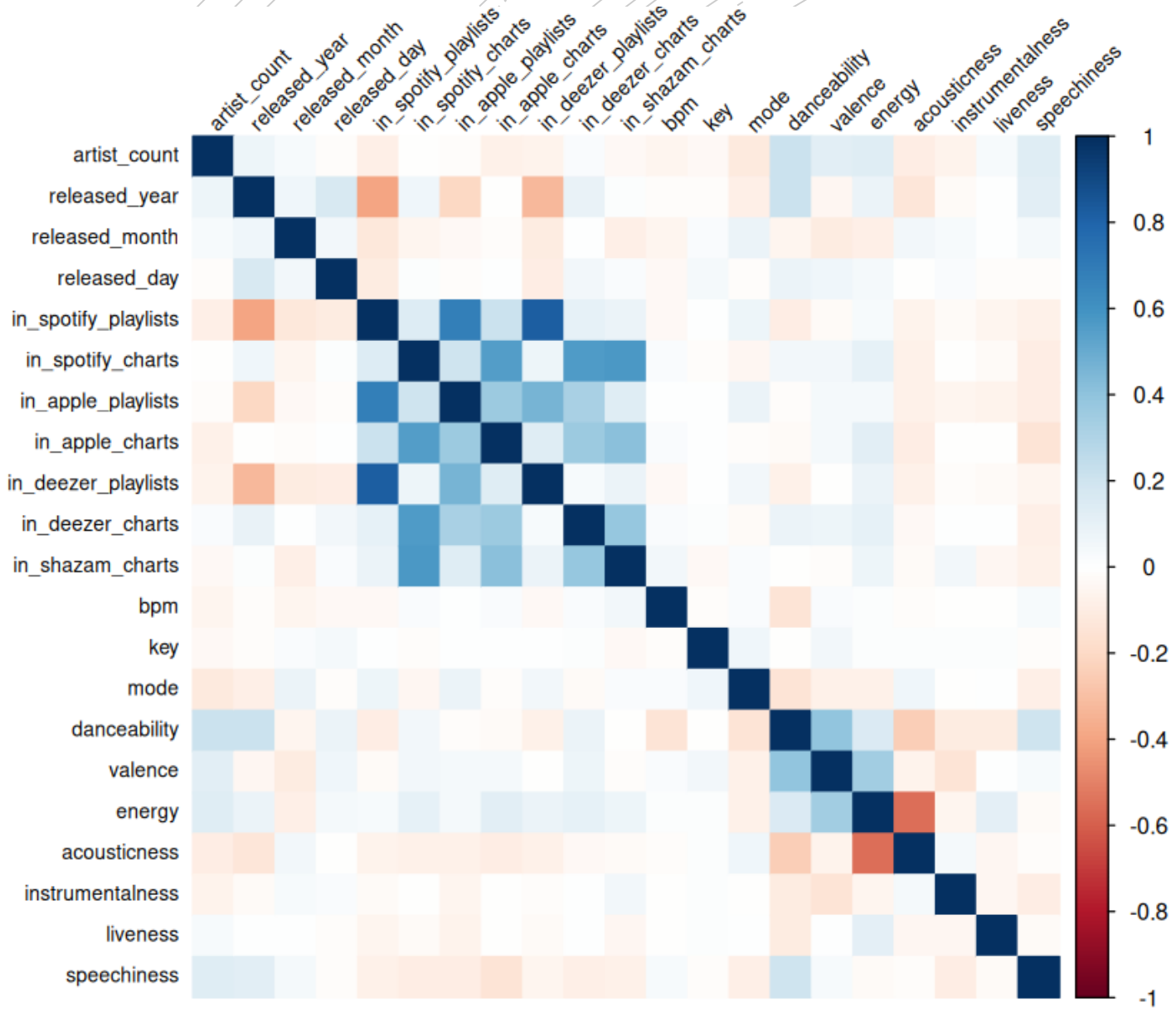
We aim at estimating a song's popularity (i.e. the number of streams) from its presence on various platform and musical characteristics. We expect to find high correlation between a song's diffusion and the number of times it's been listened (streamed), while we are unsure about the relevance of a song's attibutes for its popularity.

Preparing the Dataset

In order to use the dataset we had to make some corrections:

- We converted some text based features to numerical values: *mode* became either 0 (*minor*) or 1 (*major*), while *key* became the corresponding frequency in Hz
- Text based features that couldn't be converted, *track_name* and *artist(s)_name*, have been removed
- As the data on key and shazam charts was partially unavailable and both feature's contribution was either trivial (extremely low variance) or redundant (high correlation), we decided to remove these features

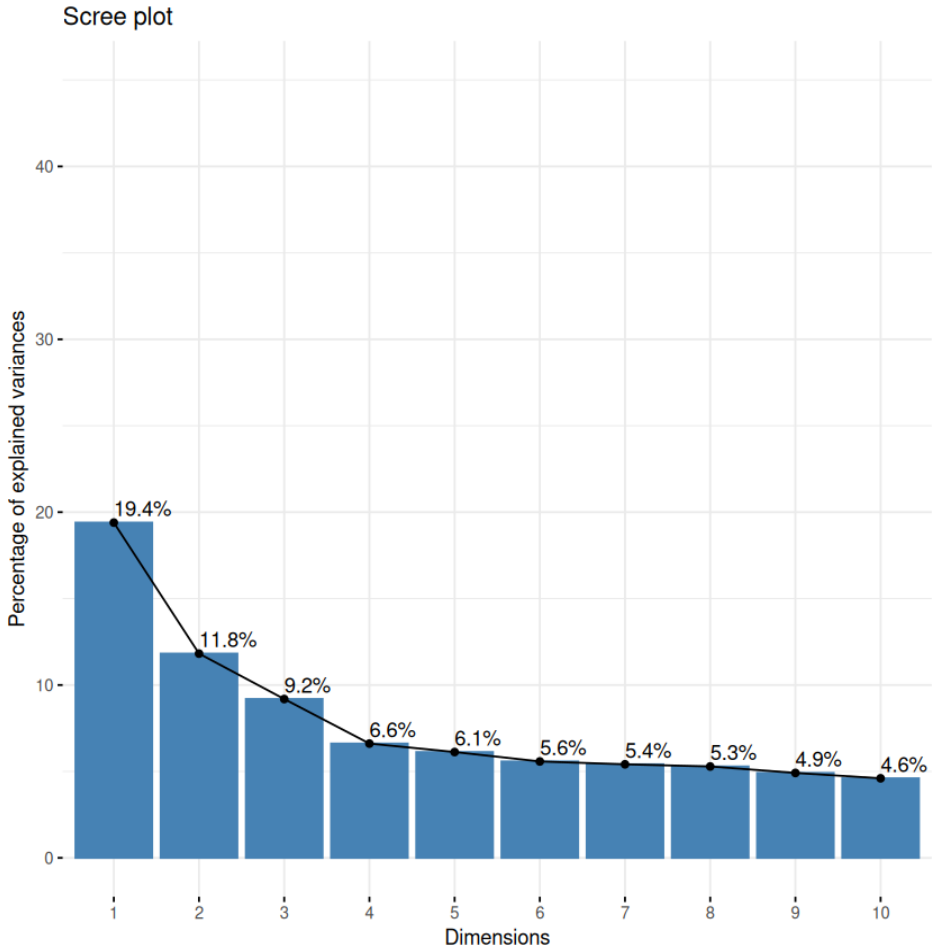
We also divided the dataset into training and test sets with the following proportions: 85% training, 15% test

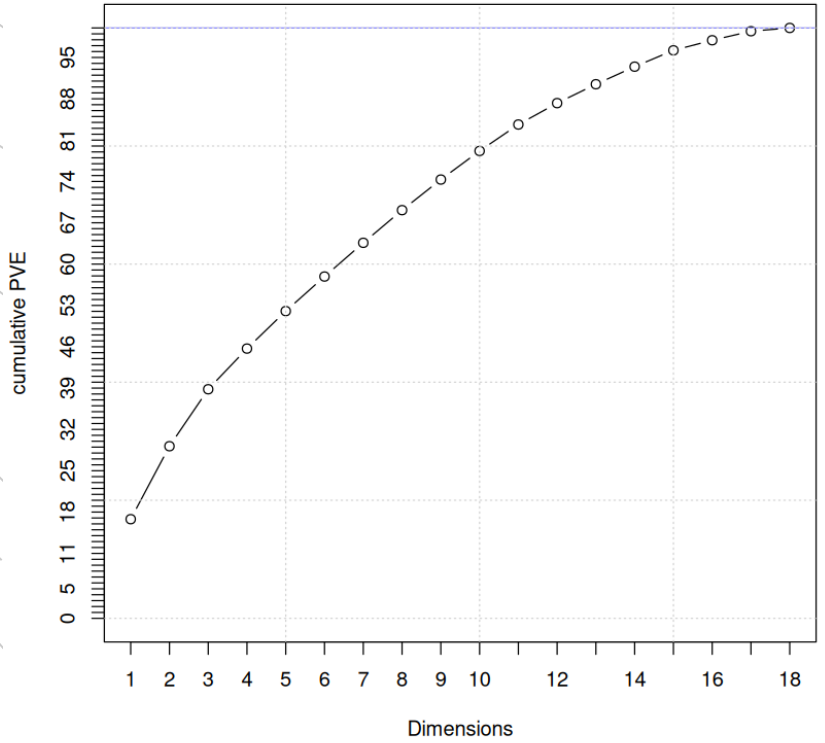


CORRELATION MATRIX

- The presence in a playlist is positively correlated to the presence in other playlists and, although not as strongly, in charts
- Acousticness is negatively correlated to energy, while valence is positively correlated to energy and danceability

PCA

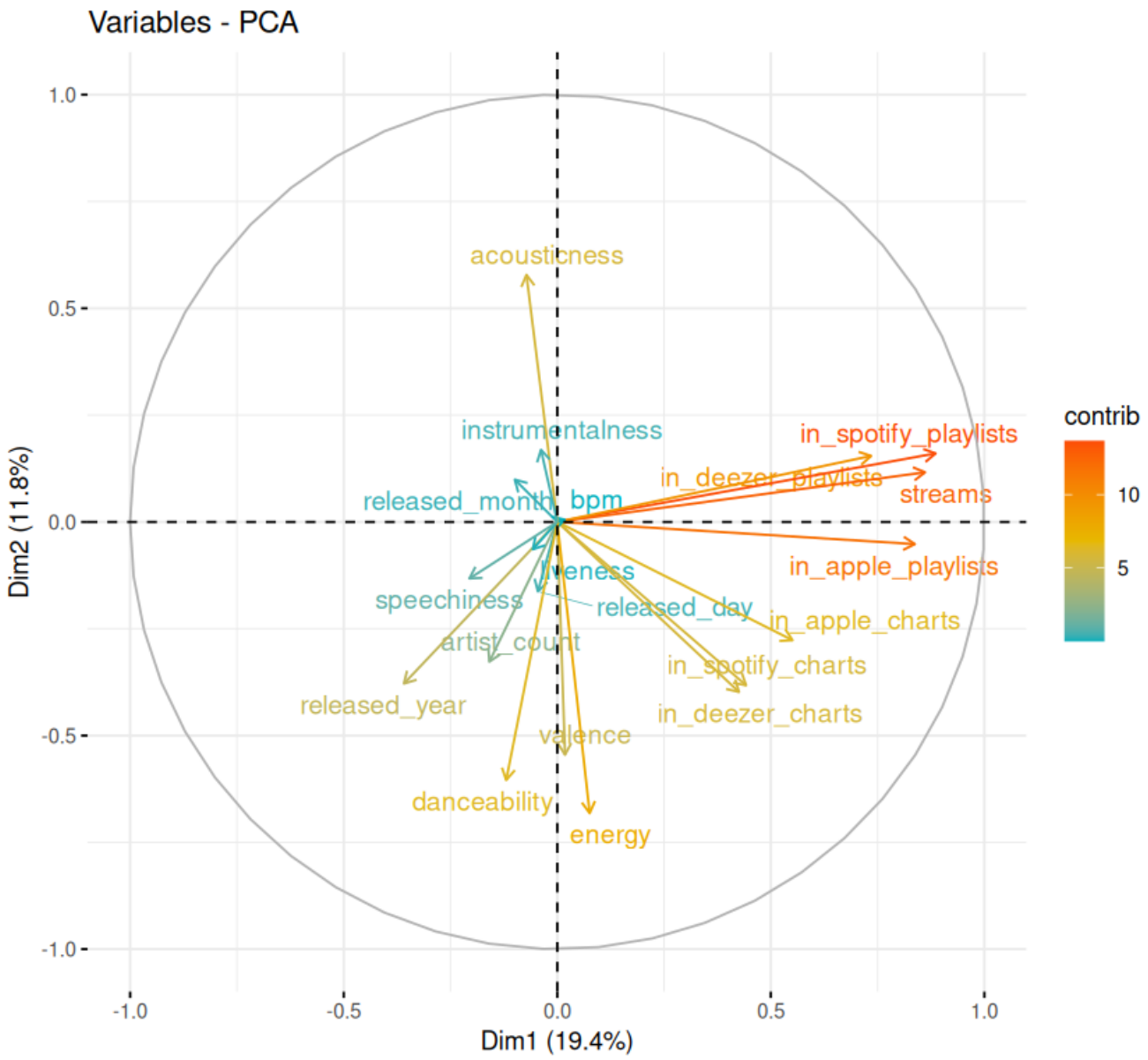




- We can clearly see there is no elbow in the cumulative PVE plot, with only few dimensions above the 90-95% line
- We chose not to remove any dimension because, as can be seen by the curve in both plots, the contribution of each dimension is non negligeable

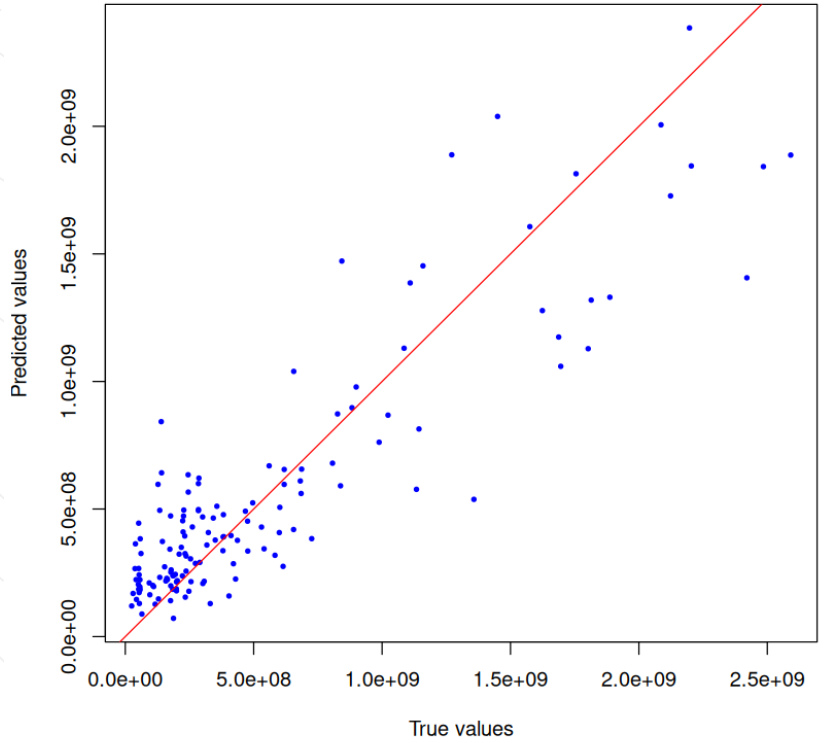
PCA

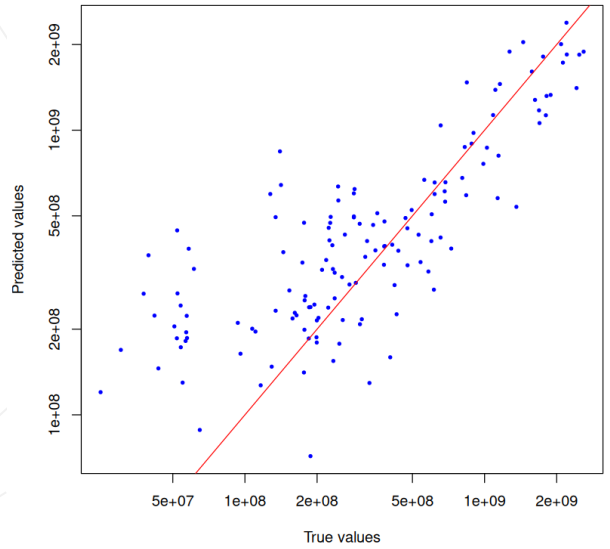
This plot shows the various feature's contribution to the first two dimensions, but as these amount to only 30% of the total variance, its significance is greatly reduced



Ridge

- As some of our feature have high correlation with each other we decided to use Ridge penalization
- These are the prediction results on the test set shown on both a linear and a logarithmic scale

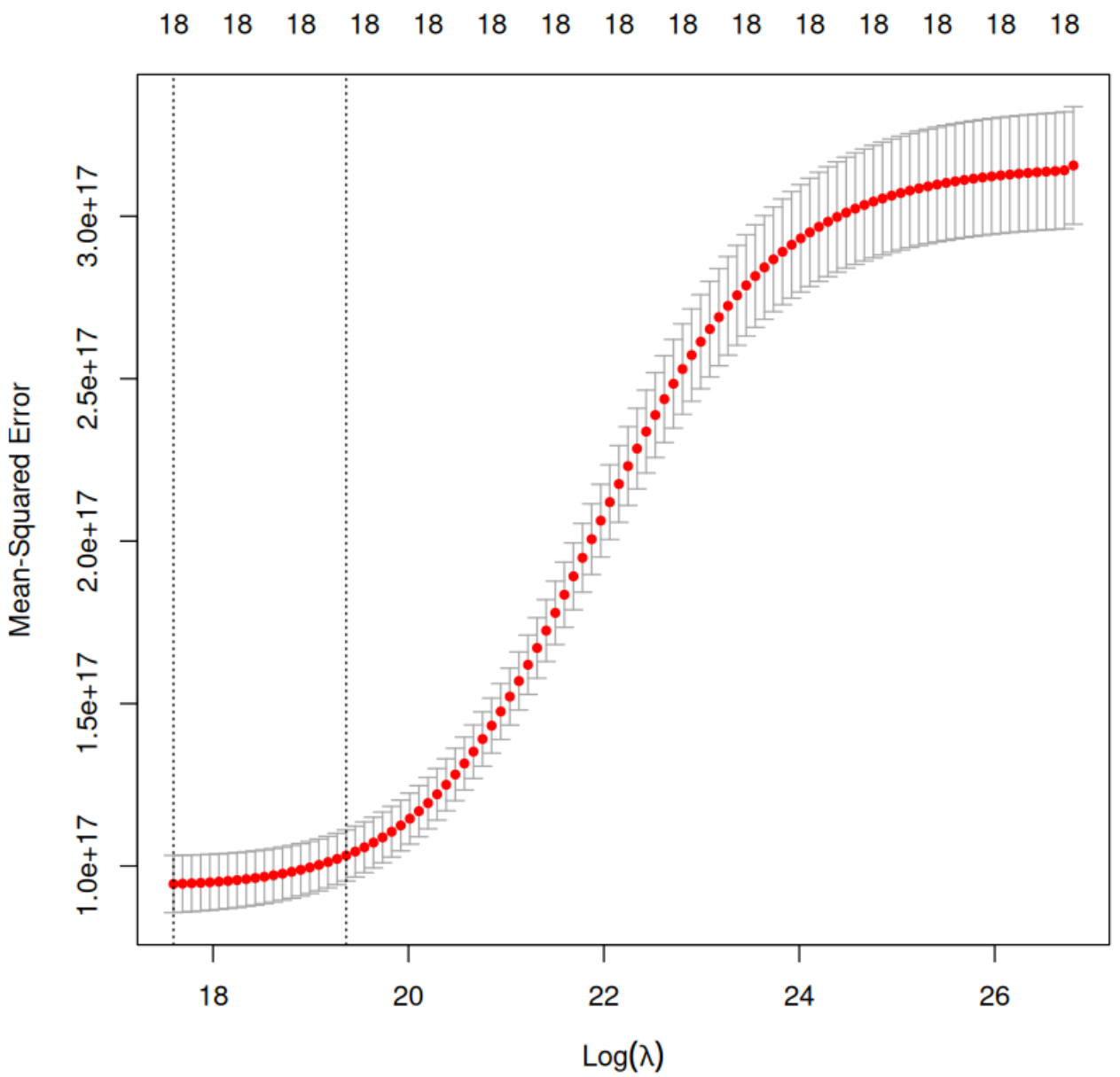




Ridge

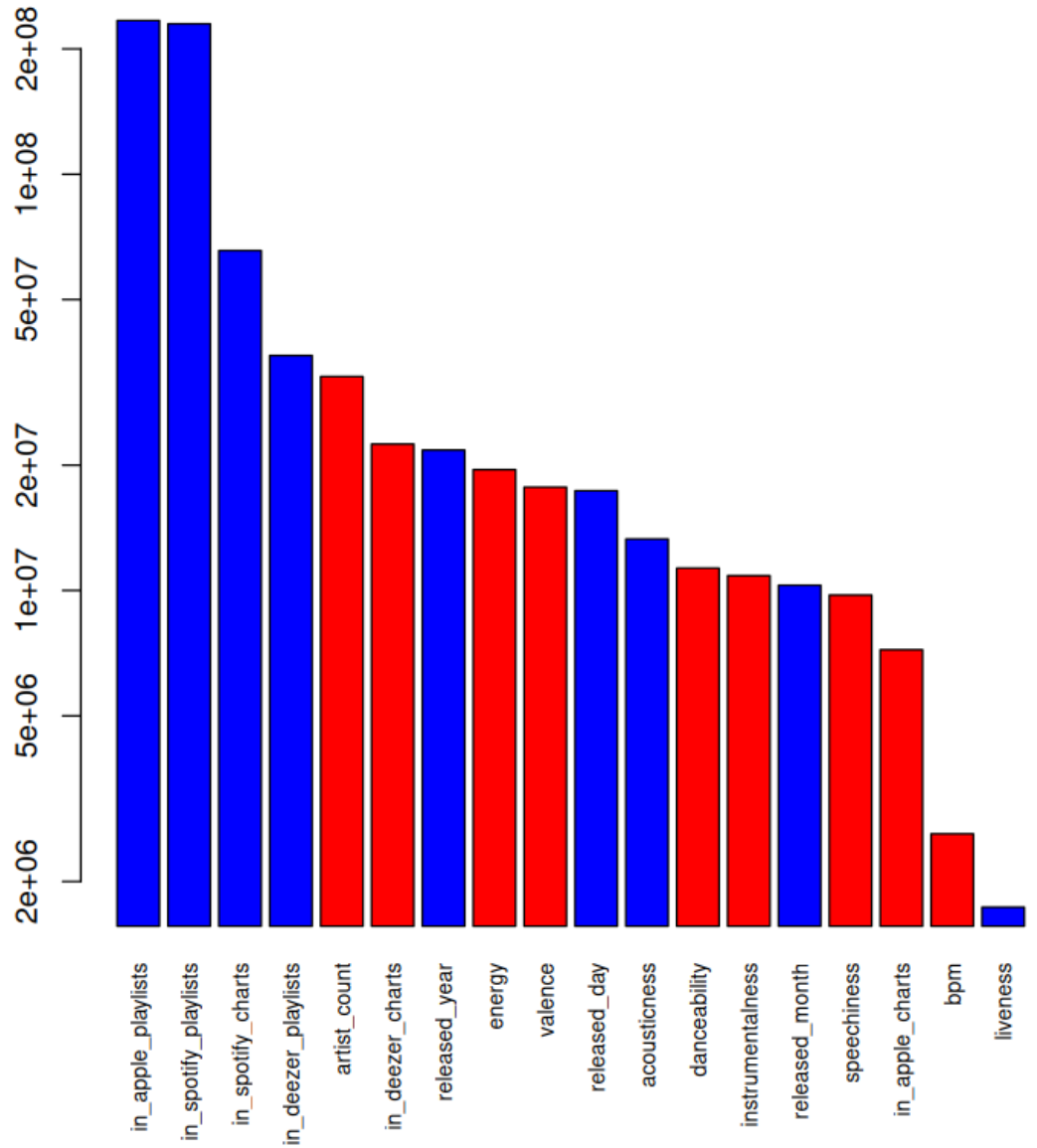
- We used cross-validation in order to find the best lambda parameter for ridge penalization
- As shown in the chart the best value of lambda to minimize the MSE is approximately 18

$$\hat{\beta}_{ridge} = argmin\{\|Y - X\beta\|^2 + \lambda\|\beta\|_{(2)}\}$$



Results

These are the betas resulting from our regression: as predicted a song popularity is mainly determined by its diffusion on various platforms (presence in playlists and charts). Another notable contribution is given by the release year: songs released more recently (in particular this year) are generally listened more.



References

- Dataset «Most Streamed Spotify Songs 2023» from kaggle: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023
- Material for the course «Statistical Learning & Large Data» by Prof. Chiaromonte available at: https://github.com/EMbeDS-education/ComputingDataAnalysisModeling20232024/wiki/SLLD
- R programming language:
  Project page: https://www.r-project.org/
  User Manual: https://cran.r-project.org/manuals.html

Thank you for your patience!