Cool Kids Playing With SQL

Raman Zatsarenko | Austin Couch | Srikar Sundaram | Ricky Gupta

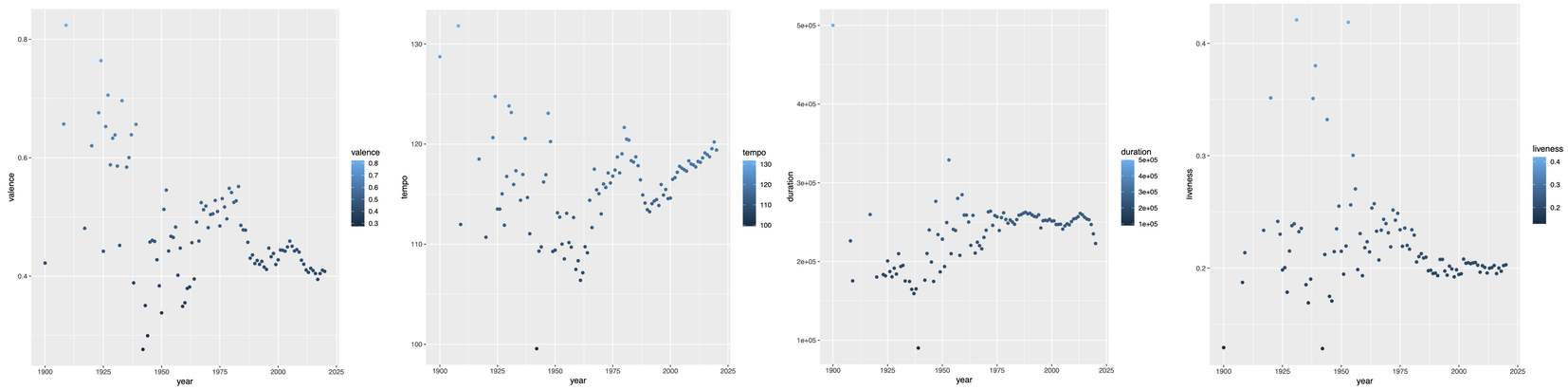
Introduction

Based on the music data we have extracted for our data base, we performed a comprehensive data analysis and discovered some interesting patterns in the data. We were also able to construct a regression model that could predict the year a song was produced in based on certain parameters. R and Minitab were used as the main analysis tools.

Findings

Notice a strong converging pattern in all plots from figure 1. It seems that as the year of production increases, the music produced becomes more similar, since the scattered values converge into a single value. This observation can be explained if we consider how the process of music production has changed over time. Before music used to be very experimental with all kinds of tempos, harmonics, etc. being used. Now music is primarily made by large producing labels, which follow a similar 'hit formula' to make any song as commercially successful as possible.

Data Patterns



(Fig. 1) Scatterplots Showing the Chaneges in Song Metrics Over the Years

Multiple Linear Regression Model

Using the correlation plot to investigate the relationships between parameters, we constructed a number of regression models trying to predict the year of production based on those parameters. Using selection criteria such as the R^2 adjusted and Mallow's Cp, we were able to construct a model that fits well and can be used to predict the year of production. Refer to the equation below.

$$year = 2384.99 + 4.58x_{i1} - 123.06x_{i2} + 70.74x_{i3} - 1.86x_{i4} - 7.18x_{i5} - 96.46x_{i6}$$

(Fig. 2) Model equation

Figure 2 describes the model we came up with using R and the lm() function. Here i = [1, ..., n], where n is the sample size. The Mallow's Cp criterion produced a value of 8.12, which is close to the number of parameters we are approximating and is considered a good result. The R^2 adjusted criterion was found to be 0.898, which is also considere a good result and makes the model usable. A residuals analysis was conducted to verify that a linear model indeed fits the data. The result of the analysis showed that a linear model is plausible.
Note that in Fig.2 the following parameters (in order) were used: loudness, acousticness, instrumentalness, tempo, duration, valence. If we specify a value of x for each of those, it is possible to approximate the release year.

Technologies

Communications & Information Sharing

Discord
Slack
Google Drive
Diagrams.net

Programming Tools & Database Technology

DataGrip
Git & Github
IntelliJ IDEA
Visual Studio Code
PostgreSQL

Analysis

R
Minitab

R Packages

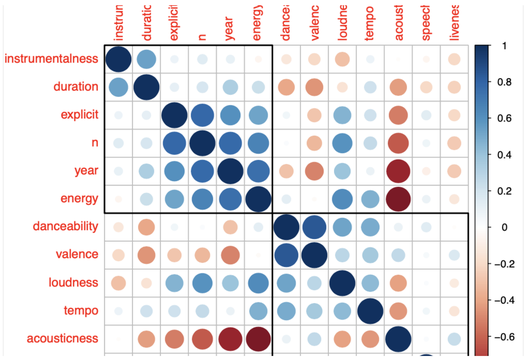
tidyverse
patchwork
factoextra
cluster
corrplot

Defining Metrics

All metrics are described as they are defined in the official Spotify API documentation.

Valence - describes musical positiveness conveyed by a track (0-1). Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
Tempo - the speed or pace of a given song (in beats/min).
Duration - how long is a given song (in miliseconds).
Liveness - the probability that the song was recorded with a live audience (0-1). A value above 0.8 provides a strong likelihood that the track is live.
Loudness - measurement of the average decibel level of the song.
Acousticness - describes how acoustic a song is (0-1). A score of 1.0 means the song is most likely to be an acoustic one.
Instrumentalness - represents the amount of vocals in the song (0-1). The closer this value is to 1.0, the more instrumental the song is.

Correlations



(Fig. 3) Correlation Matrix Showing Data The Relatedness Between Data When Summarized By Year.