Analysis of Spotify Top 50 Songs of 2022: US, Italy, Japan

Sergei Korotkikh

19-12-2022

Contents

Link to the repository: https://github.com/DataAccess2020/Capstone_Spotify

Number of commits: 43

Number of pull requests: 1

Number of issues: 3

1. Introduction

This project is devoted to analyzing the features of Spotify Top-50 songs of year 2022 in the United States, Italy and Japan. Using the Spotify API, I obtained the data for the features for each song. I was interested to see if there are any features that make a song popular, and if such features are the same around the world.

2. Literature

Since it is my first research project that involves statistics, I searched for similar analyses to use one as a guideline. I found a research paper called *"What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs"* in which Top 100 US Trending Songs of 2017 and 2018 were analyzed. According to this article, the most trendy songs in the US scored high in danceability and energy. In addition, their means for liveness and instrumentalness were low [1, p.79-82].

When choosing the countries for the analysis, I took into consideration the significance of the country's music market in the world, the region in which the country is located, and the number of "authentic" songs on the playlist of top songs on Spotify. USA and Japan were the two biggest music markets in the world in 2021 [2, p.10], so I decided to choose them. Other countries (Germany, South Korea, China) either have many American songs or are located in the same region (North America or Asia). So, as my third country I have chosen Italy.

3. Hypotheses

Since most people love 'happy' songs that they can easily dance to, I suppose that this year again the top songs should score high in danceability, energy and valence (the variable which shows how 'happy' the song is [3]). Furthermore, I think that these features should not be different in the USA, Japan and Italy. In other words, there should not be any statistically significant difference between any features.

Thus, my hypotheses are:

1. The most popular songs are high in **danceability** and **energy**.
2. The most popular songs' features **are not different** in the three countries analyzed.

4. Data

The tasks for my project are the following:

1. Extract the data for playlists with the most popular songs in the USA, Japan and Italy;
2. Extract the data (artist, title, position, features) for each song in the playlist;
3. Create the table with all the songs and three tables for each country's songs;
4. Perform tests (ANOVA, t-test) on the variables;
5. Visualize the data.

For my analysis, I needed to have the data frame for each country's top songs, as well as the features of the songs. First, I logged in with the credentials into the system. Then, in the Spotify app, I got the IDs of the three playlists (*Top Tracks of 2022 USA*, *Top Songs Italia 2022*, *Top Tracks of 2022 Japan*) and made the request using the API to get the information about the songs in the playlists. There are 50 songs in each playlist, which makes it total of 150 songs (n=150).

Here I encountered the difficulty: the API returned a large list with much unnecessary information about each song. To get the features for each song, I needed only the track ID. Then, using another API link, I created the data frame with the songs' features. It was weird for me to see that the values in the table were saved as lists, not as vectors. Luckily, the `unlist` function helped me.

Here is an example of getting the features for top 50 USA songs:

```
# Loading packages
library(tidyverse)
library(httr)
library(stringr)
library(cowplot)
library(knitr)


# Creating a table with USA song IDs ----

# the link to the playlist
songs_usa <- GET(url = url_usa,
                 config = add_headers(authorization = token))

# extracting the 'large list' to know the way to song ID
songs_usa <- httr::content(songs_usa)

# creating a vector which will contain songs' IDs
song_id_usa <- vector(length = 50)
```

```r
# loading the IDs into the vector
for (i in 1:50) {
  song_id_usa[[i]] = songs_usa$items[[i]]$track$id
}


# Creating a table with artist name and song title USA ----

# creating a vector for the URL for each song
vec_id_usa <- vector(length = 50)

# creating the URL for each song for API to use
for (i in 1:50) {
  vec_id_usa[i] <- str_c(track_url, song_id_usa[i])
}

# creating a table which will contain song ID, artist and title
names_usa <- data.frame(id = character(),
                    artist = character(),
                    title = character())

# adding the values to the table
for (i in 1:50) {
  names_usa[nrow(names_usa) + 1, ] = c(song_id_usa[i],
         songs_usa$items[[i]]$track$artists[[1]]$name,
         songs_usa$items[[i]]$track$name)
}


# Creating a table with songs features USA ----

# creating a list which will contain songs' features
features_usa <- vector(mode = 'list', length = 50)

# extracting the songs' features into the list
for (i in 1:50) {
features_usa[[i]] = GET(url = vec_id_usa[i],
                        config = add_headers(authorization = token))
features_usa[[i]] = httr::content(features_usa[[i]])
}

# creating a table with songs' features;
# deleting the unnecessary features;
# creating the table with ID, artist, title and features
songs_f_usa <- do.call('rbind', features_usa)
songs_f_usa <- songs_f_usa[, -c(12:16)]
usa <- cbind(names_usa, songs_f_usa)


# Changing tables' values from lists to vectors ----
for (i in 4:16) {
  usa[, i] <- unlist(usa[, i], use.names = F)
}
```

```r
# Adding a song position on the chart to the table ----
usa <- add_column(usa, position = 1:50, .after = 1)
```

I have done the same with the Italian and Japanese playlists.

5. Analysis and visualization

I was interested in the analysis of the means of features for each country. I grouped the songs by countries and counted the means for their features.

Having analyzed the means of the features, I have been able to obtain the following results.

The most popular songs of 2022 in all three countries score high on *danceability* (>0.6 on the scale from 0 to 1 [3]), *energy* (>0.63) and *loudness* (>-7 on the scale from -60 to 0 [3]). At the same time, the songs have low scores in *speechiness* (<0.12), *acousticness* (<0.23), *instrumentalness* (<0.02) and *liveness* (<0.22). It means that people from the USA, Japan and Italy like energetic songs that are easy to dance to. Also, the songs are not instrumental, they are recorded in the studio, and almost all the words in the songs are sung (not spoken).

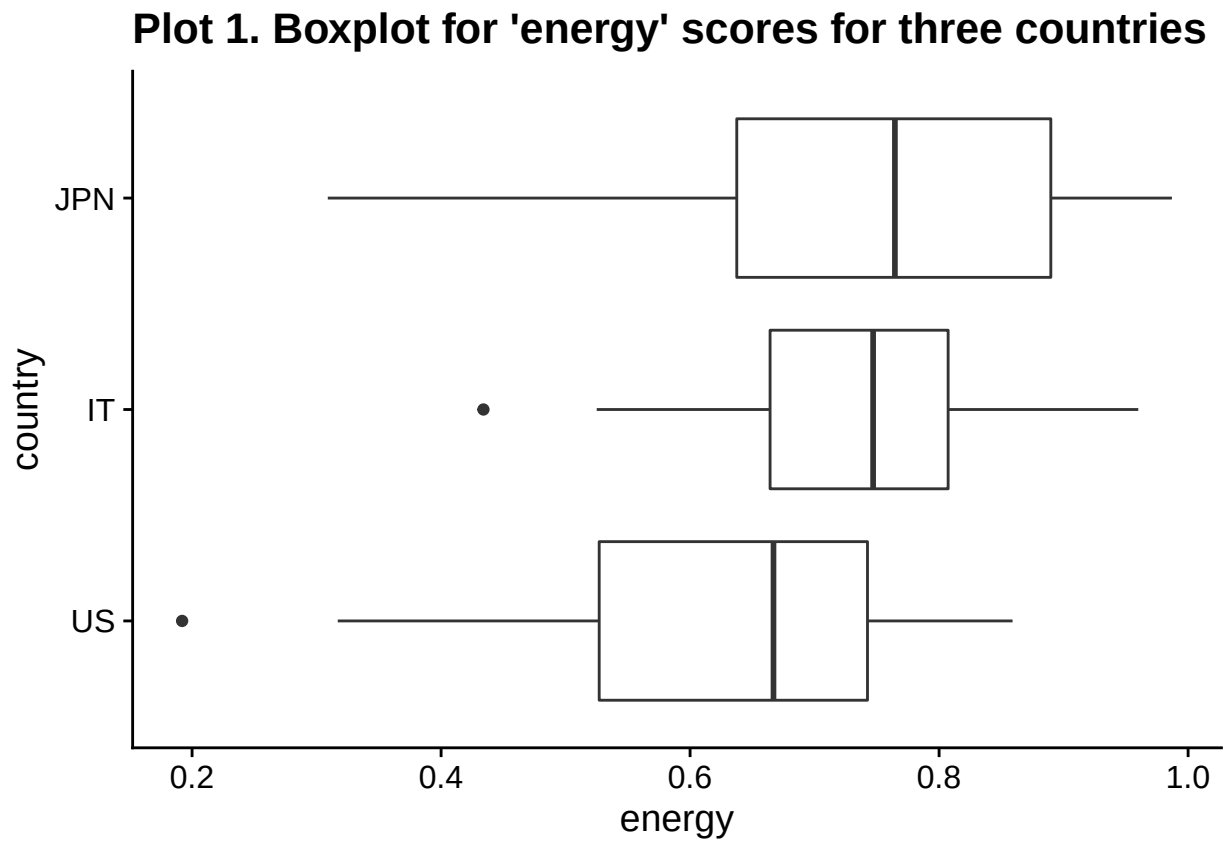
I have also paid attention to the differences in the means in particular features between the countries. For example, the *danceability* feature means are almost equal in the US and Italy (0.67790, 0.67388), but it is different in Japan (0.60980). t-test (0.05) demonstrates statistically significant difference between the *danceability* means in Italy and Japan:

Table 1: t-test on Italy - Japan danceability

P value	0.0211
t and df	2.344, 98

It means that Japanese people prefer songs that are slightly less danceable than Italians and Americans.

On the other hand, American people enjoy less energetic songs (0.63474) compared to people from Italy and Japan (0.73332, 0.74724):



ANOVA analysis has also been performed on other variables to see if there are any statistically significant differences between the means of the three countries. The tests on *loudness*, *speechiness*, *valence* and *duration* of the songs returned the statistically significant p-values.

Table 2: ANOVA: p-values for songs' features

variable	p
loudness	0.000
speechiness	0.008
instrumentalness	0.461
liveness	0.052
valence	0.025
tempo	0.614
duration_min	0.000

For example, this is the ANOVA test that pays attention to the songs' *valence*:

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## country       2  0.367 0.18366   3.765 0.0254 *
## Residuals   147  7.172 0.04879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although some tests returned statistically significant values, not all of them are practically significant. For instance, for ANOVA on *speechiness* p-level is less than 0.01. However, the *speechiness* values for Italy, Japan

and the USA do not vary much: 0.09, 0.07 and 0.11 accordingly (on the scale from 0 to 1). Still, practical significance, as well as the statistical one, has been found in *valence* and *duration* variables. Japanese people prefer happier songs (0.604) than Italians and Americans (0.509, 0.492). In addition, top songs in Japan are longer (~ 4 minutes) compared to the Italian and American ones (~ 3 and 3.5 minutes).

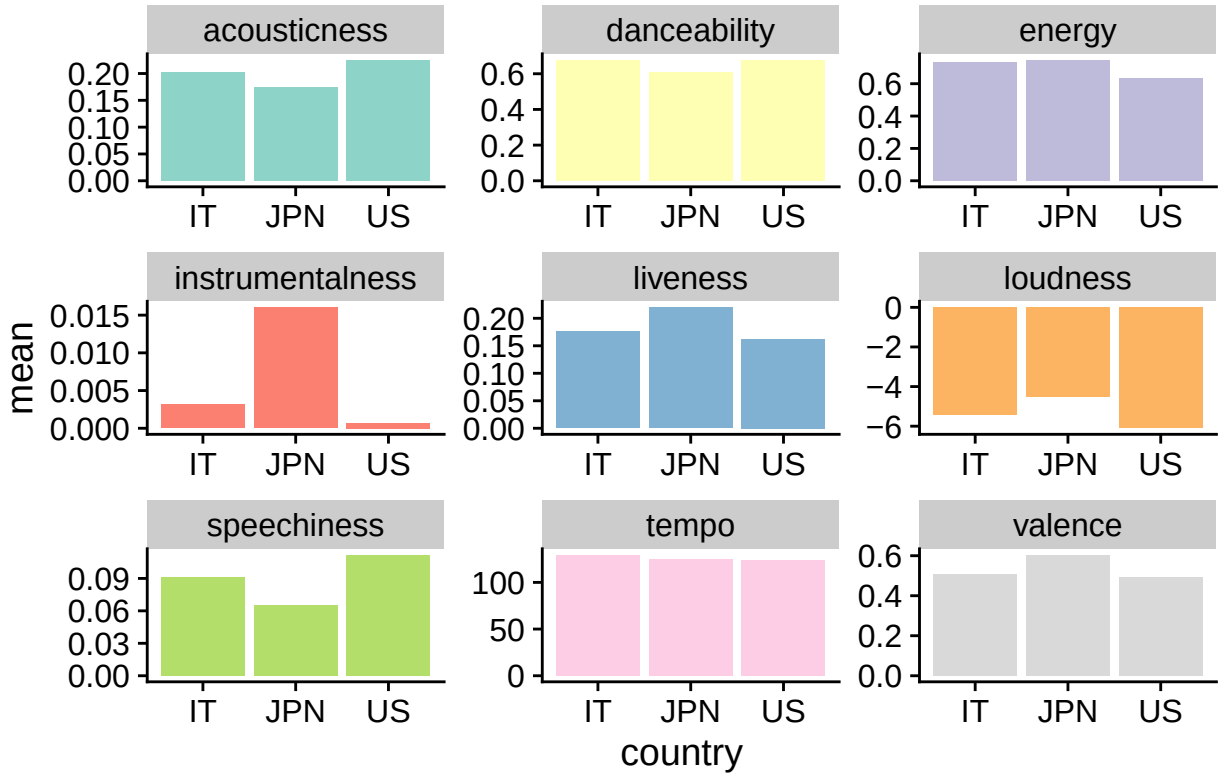
Then, I decided to create a plot which would show the differences in all nine features across the countries. I was thinking of using the `facet_wrap` function of `ggplot2` package. I needed to create a table which would have the country as the first column, the mean score for a specific variable as the second one, and the variable for this mean score as the third one. It took me some time to work on the code, but in the end I managed to develop one:

```
means_first <- cbind(all_means_main[1:3, 1],    # creating the table with
                                                # the column 'country',
                     all_means_main[1:3, 2],    # 'mean'
                     names(all_means_main[2])) # and 'variable', and
                                                # inserting the first value in there
# setting the names of the columns
names(means_first)[1] <- 'country'
names(means_first)[2] <- 'mean'
names(means_first)[3] <- 'variable'

# inserting the values of other variables
for (i in 3:10) {
  means <- cbind(all_means_main[1:3, 1], all_means_main[1:3, i], names(all_means_main[i]))
  names(means)[1] <- 'country'
  names(means)[2] <- 'mean'
  names(means)[3] <- 'variable'
  means_first <- rbind(means_first, means)  # joining the first table
                                           # and the table with a particular
                                           # variable
}
```

Here is a plot comparing nine features across the countries:

Plot 2. Differences in features across countries



6. Conclusion

After the analysis, I came to the following conclusion.

One of my hypothesis proved to be right, and the other one wrong. People in three different parts of the world *do* love energetic music which they can dance to. What is more, for each of top 50 songs (2022) in all the three countries such features as *acousticness*, *speechiness*, *liveness* and *instrumentalness* are low.

There are some differences in features' scores across the countries as well. For instance, songs in the Japanese playlist on average are longer, less danceable, and more 'happy' than those in American and Italian playlists. At the same time, American people prefer less energetic songs than Italians and the Japanese. Mean scores for some other features (*loudness*, *speechiness*) have statistically significant difference, but in practice the difference in values is negligible.

7. Reflection

It was extremely interesting to work on this project! I have known about Spotify songs' features for a long time, but I did not know how I could perform a proper analysis and reach relevant conclusions. Now that I know some basics of statistics, R, HTML and API, I can choose the topic for the research on my own, as well as extract the transform the data for my personal needs, analyze it and visualize it.

The most useful things for me during the project development were the `httr` package which helped me to extract the data, and the *for-loop*, thanks to which I was capable of transforming the data and filling the tables much faster. I also referred to the Spotify API documentation and *stackoverflow* very often: they both helped me when I was stuck.

I am very proud of what I have done and I am sure that the skills I practiced while working on this project will help me in my future studies and career.

Sergei Korotkikh

8. References

1. Al-Beitawi, Zayd, Mohammad Salehan, and Sonya Zhang. *"What makes a song trend? Cluster analysis of musical attributes for Spotify top trending songs."* Journal of Marketing Development and Competitiveness 14.3 (2020): 79-91.
2. IFPI. *Global Music Report.* URL: https://globalmusicreport.ifpi.org/
3. Spotify Web API | Spotify for Developers. Get Track's Audio Features. URL: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features