Spotify Weekly Top Songs Open-Source Data Project

Benjamin Ingram

August 2022

Executive Summary

Spotify is a Swedish audio streaming and media services provider founded in 2006. It is one of the largest music streaming service providers, with over 433 million monthly active users. Spotify offers digital copyright restricted music and podcasts, including more than 82 million songs, from record labels and media companies, and is available in 180+ countries. In this analysis, we will be exploring a dataset containing songs from Spotify's 'Weekly Top Songs' charts for each country for the weeks 2/4/21 through 7/14/22, in order to find insights into which artists, songs, and genres are most popular across geographical markets, and which song attributes are most prevalent in the most popular songs during the period.

Data Source

https://www.kaggle.com/datasets/yelexa/spotify200

The Spotify Weekly Top Songs is an open data source, collected by Spotify and made publicly available via Spotify's API. The data set was posted to Kaggle.com by a user in the community who obtained the datasets from Spotify and consolidated them into a single dataset. As the dataset was uploaded and shared with the Kaggle Community by a user in the community, the data collection process was reviewed and the data set briefly reviewed to determine its accuracy and authenticity. After reviewing the user's explanation on their GitHub and Kaggle, I believe the source is trustworthy.

Data Collection: This data set was collected from Spotify Charts and Spotify API and posted to Kaggle.com by a community user. The user first downloaded .csv files from the "Weekly Top Songs" charts on Spotify Charts for each country for the weeks 2/4/21 through 7/14/22. The files were then concatenated to create one combined file with each country, and additional data for each song was then obtained from the Spotify API and added to the dataset.

Data Contents: The dataset includes songs from the "Weekly Top Songs" Spotify Charts for each country for the weeks between 2/4/21 through 7/14/22. Variables include rank, artist, genre, track name, record label, peak rank, previous rank, weeks on chart, weekly number of streams, week, country, region, language, and various song measurements such as energy, loudness, speechiness, instrumentalness, etc.

Reason for Collection: I chose this data set because I am interested in exploring how music preferences differ among countries and regions in the world. As a frequent Spotify user, I am always interested in seeing which cities and countries my favorite artists are most popular. I can analyze the data set to obtain insights into music genre preferences by country and find what song attributes are most prevalent among top songs.

Data Cleaning & Consistency Checks

1. Dropped columns unnecessary/irrelevant to analysis including unnamed index column, uri, artist_id, artist_img, album cover, and pivot. Mostly unstructured data.
2. Renamed remaining columns for clarity including artists_num > number_of_artists, album_num_tracks > number_of_album_tracks, source > record_label, streams > weekly streams.

3. Removed 72 duplicate erroneous rows which only included column headers and no actual data to be analyzed.

4. Removed 395 rows with missing fields on song attributes (danceability, energy, key, mode, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration). Justified as the rows make up a very small percentage of total data set and the missing fields are important to analysis and wouldn't make sense to interpolate.

5. Updated data types to 20 variables. Entire dataset was originally string values, even though many of the variables were numeric. Variables were updated to integer and float data types as appropriate.

6. Updated -1 value in 'previous_rank' column to 'n/a' after review found that this likely indicated songs that were not charted in the previous week

7. Corrected formatting of 2 entries under 'release_date' field. (1967-09 > 09/01/1967, 1957-09 > 09/01/1957)

Data Profile

Number of rows and columns in the final dataset: 1,787,531 rows x 30 columns.

Column Details:

Variable	**Time Variance**	**Structure**	**Qualitative/Quantitative**	**Nominal/Ordinal Discrete/Continuous**
rank	Time-variant	Structured	Quantitative	Discrete
artist_names	Time-invariant	Structured	Qualitative	Nominal
number_of_artists	Time-invariant	Structured	Quantitative	Discrete
artist_individual	Time-invariant	Structured	Qualitative	Nominal
artist_genre	Time-invariant	Structured	Qualitative	Nominal
collab	Time-invariant	Structured	Qualitative	Nominal
track_name	Time-invariant	Structured	Qualitative	Nominal
release_date	Time-invariant	Structured	Quantitative	Discrete
number_of_album_tracks	Time-invariant	Structured	Quantitative	Discrete
record_label	Time-invariant	Structured	Qualitative	Nominal
peak_rank	Time-variant	Structured	Quantitative	Discrete
previous_rank	Time-variant	Structured	Quantitative	Discrete
weeks_on_chart	Time-variant	Structured	Quantitative	Discrete
weekly_streams	Time-variant	Structured	Quantitative	Discrete
week	Time-variant	Structured	Quantitative	Discrete
danceability	Time-invariant	Structured	Quantitative	Continuous
energy	Time-invariant	Structured	Quantitative	Continuous
key	Time-invariant	Structured	Quantitative	Discrete
mode	Time-invariant	Structured	Qualitative	Nominal
loudness	Time-invariant	Structured	Quantitative	Continuous
speechiness	Time-invariant	Structured	Quantitative	Continuous
acousticness	Time-invariant	Structured	Quantitative	Continuous
instrumentalness	Time-invariant	Structured	Quantitative	Continuous
liveness	Time-invariant	Structured	Quantitative	Continuous
valence	Time-invariant	Structured	Quantitative	Continuous
tempo	Time-invariant	Structured	Quantitative	Continuous
duration	Time-invariant	Structured	Quantitative	Continuous
country	Time-invariant	Structured	Qualitative	Nominal
region	Time-invariant	Structured	Qualitative	Nominal

language	Time-invariant	Structured	Qualitative	Nominal

Data Limitations and Ethics

1. Data only represents the period 2/4/21- 7/14/22, making it difficult to conduct time series analysis, identifying changing trends over a long period of time.
2. Songs with multiple artists are split into separate rows for each artist. This could result in inflated insights by counting these songs more than once during analysis. Songs with multiple artists may skew results.
3. Per the user who collected the data, “Since many artists had multiple genres, one of those genres was chosen for each row.” Without insight into how the genre was selected, I am placing trust that the user’s methodology is sound.
4. Analysis of popular music by country is limited to those countries that Spotify has made a “Weekly Top Songs” playlist for.

Questions to Explore

1. Which artist(s) have the most ranked songs across the world?
2. What genres are most popular, indicated by the genres that show up most frequently in song rankings?
3. Which genres are most popular in each country?
4. What songs have spent the most amount of time on charts during the time period?
5. What song attribute scores are most likely to result in a song reaching #1 rank on music charts? For example, are songs with high danceability scores more likely to reach #1?