Datasheet

Dataset: Spotify Charts

Description:

Is a dataset of all daily hit charts curated by Spotify. Source: https://www.kaggle.com/datasets/dhruvildave/spotify-charts The dataset holds all the Top 200 and Viral 50 charts published by Spotify. A new chart is published every 2-3 days. This collection includes charts from the beginning of January 1, 2017 up until December 31, 2021.

Detailed description of the dataset's columns:

- title as string – the title of the song
- rank as integer – number in the ranks from 1-200
- date as date in format dd/mm/yyyy – the date on which the charts are published
- artist as string – the artist's name of the song
- url as string – the reference link of the song for the Spotify API
- region as category – the country where this song was in the charts
- chart as category – song is in the top 50 ranks, the chart is viral50 otherwise is top200
- trend as category – describes if a song gained or lost ranks, either MOVE_UP, MOVE_DOWN or SAME_POSITION
- streams as integer – the number of streams of the song over Spotify

The intended use of this dataset is either apply an exploratory data analysis of the Spotify charts (e.g., find correlation between music features and chart placement), to build a music recommender (for instance for playlists, albums or artists) or even a model to forecast charts for upcoming months.

Motivation:

The initial motivation for the gathering process of this dataset is unknown because the data was provided from a contributor on Kaggle with no further explanation. Also, how the data was collected was not described. Therefore, no detailed information on the motivation for this dataset can be given. For answering the research question, this dataset will be used for an exploratory data analysis to find a correlation of charts between countries.

Biases:

Since the dataset was collected from another user on Kaggle, the data collecting biases cannot be investigated.

- general problems
    - behavioural bias: Music listening behaviour is different between people. Not everyone will listen on the same streaming platform. This can cause problems when comparing this dataset to others, which publish charts.
- issues at the data source
    - functional bias: It is not very clear which characteristics are used to calculate the top charts in Spotify. This can be also a problem when this dataset is compared to a different one, which also

tracks charts.
    - external biases: Other factors can play a role in how people listen to music. For example, other video content heavy social media platforms, like TikTok or YouTube, can have a big influence on music consumption.

Billboard Top 100

Description:

Is a collection of the Top 100 charts on Billboard. Source of the dataset: https://www.kaggle.com/datasets/dhruvildave/billboard-the-hot-100-songs Source Billboard: https://www.billboard.com/ The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine. Chart rankings are based on sales, radio play, and online streaming in the United States. Every week, Billboard releases "The Hot 100" chart of songs that were trending on sales and airplay for that week. This dataset is a collection of all "The Hot 100" charts released since its inception in 1958. Detailed description of the dataset's columns:

- date as date in format dd/mm/yyyy – date of the chart
- rank as integer – the placement of the song in the ranking, ranging from 1-100
- song as string – the song title
- artist as string – the tracks artist
- last-week as integer – rank of the song in the previous week's charts
- peak-rank as integer – the top rank a track achieved in the charts
- weeks-on-board – the number of weeks a song was on the board

This dataset can be used for different applications. Its suites work on simple data analysis (e.g., find the best tracks of on certain artist over the years) and on machine learning to forecast the charts for different artists in the future. This dataset can also be used in combination with different datasets to find correlations between different charts datasets or to forecast which music features could influence charts in the future.

Motivation:

The data gathering process was made by a user on Kaggle and therefore, a description of the motivation cannot be made. There is not much information on how this dataset is generated or how the information from the Billboard was transferred to this dataset. To answer the proposed research question, the dataset will be used to model a forecast to predict future charts, with the exploratory data analysis from the previous dataset.

Biases:

Since the dataset was collected from another user on Kaggle, the data collecting biases cannot be investigated.

- general problems
    - population bias: Maybe older generations consume music in a different way. Probably, they will tend to listen to more radio
    - behavioural bias: Music listening behaviour is different between people. Not everyone will listen on the same streaming platform. This can cause problems when comparing this dataset to others, which publish charts.
- issues at the data source

- functional biases: the ranking system is loosely defined; it does not say which groups has the most weight. This could cause a track to be much higher in the ranking because of a higher stream count compared to a track with higher counts of radio play.
- external biases: data could be influenced by social media or other platforms because the rank is calculated per streams, radio play and sales. Also, which track is played in the radio