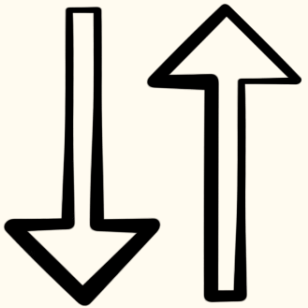
Spotify Charts Analysis

By: Liam McChesney

Project Description

Throughout this project I will be data mining a dataset full of Spotify charts data. I will be looking at the if there are meaningful differences between the two different types of charts in the data or between the different regions present in the data. Other questions include if there is something that causes a song to stay on the charts for longer or move up and down the charts at a specific pace. I also have access to different musical features of each song that could be analyzed for trends.



Prior Work

Angela Cottini did some work on the spotify Top 200 chart from 2020-2021. She focused on most frequent artists, song length, songs that stayed on the chart the longest, and release year. Her analysis does some work to find trends with the artist research. One interesting thing she found is that the number of times a song appeared on the Top 200 list was not correlated well with the number of followers the artist had.

https://rpubs.com/arcottini/821526

Reika Fujimura did analysis on what musical elements are different on a yearly basis between the songs on the Top 200 list.

https://www.reikafujimura.com/post/spotify-music-analysis

Dataset

The dataset I will be using is “Spotify Charts(All Audio Data)” pulled from Kaggle with just over one million rows. This dataset contains the global “Top 200” and “Viral 50” spotify charts from each world region from January 1st, 2017 to about two months ago. These charts are updated every two to three days and there is new data for each update. I have the dataset downloaded on my laptop and it can be reached at the link below.

Link: https://www.kaggle.com/datasets/sunnykakar/spotify-charts-all-audio-data

Proposed Work

Data Cleaning: There appears to be some error in some of the release dates as some of them are after the date they appeared on the chart. Some songs have null values in the popularity, available markets and release date columns.

Data Preprocessing: Breaking the data up between region would allow me to handle the data more efficiently. Converting the trend to a tertiary numerical value would allow some interesting calculations to be done on how the songs move up and down charts. Several columns such as URL are unnecessary and can be removed for ease of work.

Data Integration: This will not be necessary as the data is all in one dataset

List of Tools

- VSCode (Env)
- GitHub (Submissions)
- Jupyter Notebooks/ Python (Format/Language)
- Pandas(Data Manipulation Library)

Visual Studio Code

Evaluation

Results can be compared across multiple different axes. For example, any analysis can be compared between the different regions. I can also look and see if there are any correlations that appear that is apparent to the human eye but may not be obvious from a purely data point of view. Seeing if any artists have songs that stay on the charts longer or rise faster would allow people to try to emulate their song style for success.