Spotify Top Charts Analysis Progress Report

An analysis of musical trends across region, time and charts

Liam McChesney
Computer Science
University of Colorado - Boulder
Boulder, CO USA
limc8676@colorado.edu

PROBLEM STATEMENT

Throughout this project I will be data mining a dataset of Spotify Charts data. I will be looking at if there are meaningful differences between the two different charts in the data (Top 200 and Viral 50). I will also be looking at the differences between the different regions present in the data (regions include various countries from across the world). Other possible questions include analysis of what causes songs to stay on the lists for longer or to move up or down the chart at a specific pace. The dataset also contains musical features for each song. Using these I am able to analyze what is different between songs in each of the above categories.

LITERARY SURVEY

Angela Cottini did some work on the Spotify Top 200 chart from 2020-2021. She focused on the most frequent artists, song length, songs that stayed on the chart the longest, and release year. 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. Another interesting find was that in a two-year period over half of the songs on the list are released in the first year of the period as opposed to before the period or in the second year. This helps us to see a little bit about how long it typically takes songs to become popular and how long they stay popular. (Cottini 2021)

Reika Fujimura did analysis on what musical elements are different on a yearly basis between the songs on the Top 200 list from 2018 to 2022. They found significant trends around Christmas time in the US region. Consistently the acousticness and valence (positivity of a song) would increase up till Christmas and then drop off immediately after. Other than that, valence seems like the trait the changed most from year to year where all other traits follow a similar pattern each year. Comparison of musical traits between regions was also done. The United States and Canada have relatively similar data whereas the United States and Japan have large differences in categories such as energy and duration. (Fujimura 2022)

Both of the papers give a good basis for my project. I can take the type of work that Reika Fujimura did and apply it to different axes. For example, I could look at the different musical features across region as opposed to across time. I could also look at if there are specific artists that occupy the charts at a specific time of year (Mariah Carey at Christmas time perhaps). As for Angela Cottini's work I like the work she did to look at trends among artists. Much of the work I have planned to do with the musical features could also be done with artists to see if specific artists rise or fall at a certain rate or stay on a chart for longer. In this vein we could learn if perhaps well-established artists rise faster than up and coming ones.

DATASET

The dataset I will be using is the "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 various

countries from January 2017 till March 2024. 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. Rows include title, rank, date, artist, URL, region, chart, trend, streams, track_id, album, popularity, duration_ms, explicit, release_date, available_regions, af_dancability, af_energy, af_key, af_loudness, af_mode, af_speechiness, af_acousticness, af_instrumentalness, af_liveness, af_valence, af_tempo, and af_time_signature.

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

Edit: I was incorrect. This data set has about 26 million rows not just over one million. Originally, I had opened the file in excel to check how many rows it had but excel only showed me the first 1 million rows. As I have been doing preprocessing and cleaning, I discovered there are over 26 million rows.

PROPOSED WORK

Data Cleaning: There appear to be some errors in some of the release dates as some of them are after the date they appeared on the chart. I don't currently plan on using the release dates in my analysis but if I do decide to, I will run a check to flag each song with an improper or null release date and check them manually. Some songs have null values in the popularity and available markets columns. Again, I do not plan on using the available markets column but if I do, I will check and see if there is another row with the same song without a null value. If there isn't I will drop the data points with null values. As for songs with null popularity, I will again check to see if there are any other rows in a similar timeframe (about 1 month) that are not null. If there are not, these data points will be counted in their own group and reported with the data, so we are able to see how much data is lost to null values.

Data Preprocessing: Breaking the data up between regions would allow me to handle the data more efficiently (aggregation based on continents). Converting the trend to a tertiary numerical value (-1 for down, 0 for same and 1 for up) 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.

Edit: I realized that popularity doesn't have very many null values. Some artists have a 0 popularity because they aren't well known enough yet to have a popularity rating. Streams on the other hand has a lot of null values. About 5 million rows are missing this data which is about a fifth of the data. For this reason I will not be using it in my analysis. I didn't have a plan to use it anyways. I also realized that about 1% of my data is missing the af traits that much of my analysis will hinge on. Because of this I will split this data out and not use it. There are also 11 songs without a title and 18 without an artist. Since each of these small groups represents only one song each and I am unable to fill in the missing value I will pull them out of the data as well.

Data cleaning/preprocessing was effective and reduced file size by 22 GB down to under 4 GB. Some preprocessing was completed including the transformation of string fields into integer fields (trend and ch). The only thing that hasn't been completed is the aggregation of data based on continent. I am unsure if this will be useful as the original purpose was to break up the data into smaller chunks. Since file size was reduced drastically by cleaning and preprocessing this step may not be necessary. The following columns being dropped contributed to the drop in file size: url, track_id, album, streams, duration_ms, explicit, available_markets, and release_date.

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

EVALUATION METHODS

Results can be compared across multiple different axes. For example, analysis can be done between the different regions, across time, and between the two

charts. Seeing if any type of song stays on the charts longer or rise faster would allow people to try to emulate those songs for success. Analyzing when different types of songs are popular would enable artists to determine an optimal release date for their songs. One specific approach I could take if I have the time is to apply a k-means clustering to the data based on the musical qualities. This would allow us to split the songs into groups and then to analyze how the songs move up and down the charts. A way of analyzing a song or group of songs (using an average) movements on the charts would be a sort of velocity function. I can find the peak of each song/group of songs and then measure the average amount of time it spent at each position until it reaches its peak. This would be the ascending velocity. Doing the same for/songs groups of songs on the way off the chart would be the descending velocity. An ideal song would have a high ascending velocity and a low descending velocity. A last characteristic in this group would be what I will call "stickiness" or how long a song stays at its peak. A high stickiness means a song stays at its peak for a long time. All of these characteristics can be completed for different regions and then compared.

TOOLS

The following tools will be used for the project:

- VSCode: Environment
- GitHub: Submissions
- Jupyter Notebooks: Format
- Python: Language
- Pandas: DML Library

MILESTONES

1. **July 19th:** Data cleaning and data preprocessing completed **(Complete)**
2. **July 22nd:** Progress report completed **(Complete)**
3. **August 5th:** Evaluation completed **(Incomplete)**
4. **August 9th:** Final report completed **(Incomplete)**

RESULTS

Thus far I have not done any analysis. The only results I have are that about 1% of the data was lost to null data. Ultimately this seems to be a very good ratio for how unpredictable data can be. I am now ready to begin analysis as my data is cleaned and separated. I was very surprised at the large drop in file size during cleaning. Dropping unnecessary columns, converting some string fields to basic integer fields, and removing rows with nulls resulted in a drop of over 22 GB down to under 4 GB.

REFERENCES

[1] Angela Cottini, 2021. Analysis of Spotify's Top 200 Chart (2020-2021) DOI: https://rpubs.com/arcottini/821526

[2] Reika Fujimura. 2022. Spotify Music Chart Trend Analysis. DOI: https://www.reikafujimura.com/post/spotify-music-analysis