Joe Lollo
LIS 545
Dr. Nic Weber
9 March 2023

Curation Protocol Report:
Top Songs of 2022 on Spotify and TikTok, as Data

Project Abstract:

Since its release in 2016 and rise to popularity in the 2020s, TikTok has influenced cultural production in a variety of ways. Music in particular is prevalent on the platform in the form of short snippets accompanying dance moves or comedic performances, with the music being noted as "possessing a virality that often goes beyond the platform, profoundly influencing the music industry" (Montag et. al., 2021). While the influence of social platforms on what makes "popular" music has been highly documented in qualitative research and journalism, there is a fascinating potential for analytical work comparing TikTok's top global songs with global music streaming charts, to assess the influence of social media on users' music listening and streaming habits, and the lasting influences of trends and phenomena in popular culture and social media. This data repository aims to document the 100 most popular songs on TikTok and on global charts on Spotify over a 12-month period from January to December 2022, quantifying the popularity of each song and artist through the applications' own analytics. This includes each song's highest position, and the number of weeks they spent, on the Spotify charts, as well as each song and artist's popularity on TikTok.

The repository, titled **"Curation-Protocol,"** is hosted on my personal GitHub page: https://github.com/ChessPiece21/LIS-54X/tree/main/Curation-Protocol

About the Data:

The repository holds three datasets, all stored as comma-separated value (CSV) files. The main dataset, titled **"2022-Music-Charts.csv,"** is a combination of two datasets gathered from two separate application programming interfaces and joined into one dataset using the statistical programming language R. The first dataset, with the original version included as **"2022-Spotify-Charts.csv,"** was downloaded from the Spotify API. The second dataset, with the original version included as **"2022-TikTok-Charts.csv,"** features the top 100 songs on TikTok in 2022. This dataset was uploaded to the data science website Kaggle by user Sveta151, and was gathered using the TikTok API in a similar way to my own Spotify dataset. In R, I joined the separate datasets together using a "full join" method, meaning that "all records in two separate tables were joined, in a new, larger table, by a shared variable" (Wickham, 2016, 12.4.2). In my case, the shared variable I joined the datasets together with was song title, as both tables shared the column "track-name," and 37 songs were featured in both datasets.

Potential Audiences and Users:
This repository is meant to be an accessible starting point for computational and mixed-methods researchers in the fields of popular culture studies, communication/media studies, and digital humanities/cultural analytics to analyze the relationship between social media and music charts, and analyze what makes "popular" music. Data journalists or scientists outside of academia may also be interested in this dataset, which can be used for similar analyses and visualizations of 2022's music culture.

Modifications:
I made a few modifications to the data beyond the joining of the two datasets in R, however. I did them all using the spreadsheet application OpenRefine. Firstly, I added uniform resource indicators (URIs) to all songs on the TikTok charts that weren't also on the Spotify charts. I did this manually, by searching for each song on Spotify based on their titles on the TikTok dataset and then exporting the URI from Spotify and importing it into the combined dataset's "spotify_uri" column. This was mostly done to standardize every row on the dataset, but it also helps researchers attest that each song in the dataset's "track_name" column is what it purports to be.

I also added two Boolean ("TRUE"/"FALSE") variables to each song called **"spotify_chart"** and **"tiktok_chart"** that indicated whether each song was present on the Spotify or TikTok charts – at least one of these variables is "TRUE" for each song. Similarly to the above change, I made this modification to provide researchers with context where each of the top 100 songs in the datasets come from, normalizing the sources from both variables.

Lastly, I renamed a few variables, namely those about the charts, to tell which app each of these variables came from. This renaming and standardization is particularly important for re-use in data and cultural analytics, since without the additional context, researchers will be confused as to which chart positions were from which app. The renamed values are listed on the table below in bold, with their original source and original names in separate columns:

Original Dataset	Original Name	Renamed Variable
2022_Spotify_Charts	chart_pos	**spotify_chart_pos**
2022_Spotify_Charts	chart_weeks	**spotify_chart_weeks**
2022_Spotify_Charts	uri	**spotify_uri**
2022_TikTok_Charts	track_pop	**tiktok_track_pop**
2022_TikTok_Charts	artist_pop	**tiktok_artist_pop**

Documentation:
Metadata (using Dublin Core Schema):
This is specifically for 2022_Music_Charts, which is the refined dataset:

Attribute	**Value**
Title	Top 100 Songs of 2022 on Spotify and TikTok
Description	This dataset includes the top 100 songs on the charts across the social media application TikTok and the music streaming application Spotify, from January 2022 to December 2022. The intended audiences are researchers in the field of cultural analytics and media studies, as well as data journalists and scientists. This data was curated for a course at the University of Washington.
Source	Spotify; Kaggle; TikTok
Publisher	Joe Lollo
Contact Point	Joe Lollo, lollo21@uw.edu
Date Created	2023-02-24
Date Modified	2023-02-26
Subjects	popular music; social media; popular culture; streaming
Language	English
Temporal (*Coverage*)	2022-01-01 to 2022-12-31
Access URL	https://raw.githubusercontent.com/ChessPiece21/LIS-54X/main/Curation-Protocol/2022-Music-Charts.csv
Access Level	Public
Access Rights	This data is freely available to the public.
Type (*File Type/Format*)	CSV

This metadata is also included as an XML file on my repository titled "Music-Curation-Metadata.xml," also adhering to Dublin Core standards: https://github.com/ChessPiece21/LIS-54X/blob/main/Curation-Protocol/Music-Charts-Metadata.xml

Readme File:
The repository has a Readme file, meant as a guide for researchers. It contains an explanation of the data, a data dictionary (also stored as a CSV on the repository), and links to relevant files within the repository. It is stored as a Markdown file following GitHub's documentation standards:
https://github.com/ChessPiece21/LIS-54X/blob/main/Curation-Protocol/README.md

Explanation of Documentation:
I chose Dublin Core as my metadata schema because it is commonly used for describing the online preservation and dissemination of cultural artifacts, including pieces of music (Dini, 2007). Since the primary audiences are researchers and academics, they may already be familiar with this metadata schema and can further reuse this data. Spotify's style metadata style guide mentions that its "unique metadata syntax" takes inspiration "from numerous commonly-used metadata standards…[including] the Dublin Core Metadata Initiative," although their metadata files are in JSON files rather than XML (Spotify Metadata Style Guide, 2021). Dublin Core's documentation is very detailed, including examples of each field and comprehensive metadata (in multiple formats), which helped me implement best practices in this metadata schema.

I included most of the attributes that are required for metadata using Dublin Core's standards, in both the table and XML formats, but also included optional metadata attributes regarding the rights that users have to freely access this data on the web, including attributes for "accessRights" and "accessLevel" adhering to the Dublin Core standards. I also removed the "spatial" attribute, which indicates spatial coverage of the data, since this data is not tied to a specific location, and was not as important as the "temporal" attribute that was included and is rather important in explaining the dataset's coverage.

In the Readme file, written using Markdown and stored on the repository as **"README.md,"** I explained the data normalization process, the file naming conventions, and provided links to my metadata and data dictionary. The data dictionary provides explanations for each variable in 2022_Music_Charts, including their units of measurement (such as milliseconds for "duration_ms" and beats per minute for "tempo_bpm") when applicable. I was originally going to include the data dictionary as a table on the Readme file, but because it is very difficult to format tables in Markdown, it is included as a CSV in the repository:
https://github.com/ChessPiece21/LIS-54X/blob/main/Curation-Protocol/Data-Dictionary.csv

Reflection:
I enjoyed curating this data as it is on a topic that I am passionate about outside of school, and I believe that it has a strong potential for reuse in data-driven research. During my curation process, I re-read the *Library Carpentry* modules from class, specifically "Formatting Data Tables in Spreadsheet Applications." The modules' suggestion to "[fill] in missing data if you

have the resources to do so" led me to add the URIs for all the TikTok songs that were not in the Spotify dataset. Although locating and inputting URIs manually took a significant amount of time and labor, this process is very valuable with reusability of the data in mind: researchers can verify that each song is what it claims to be in the dataset. I fortunately did not encounter any field-name problems in the merged dataset, but I followed *Library Carpentry*'s suggestion to "avoid spaces, numbers, and special characters" when naming each column, knowing that problems would often arise if used in a statistical program for future analysis (*Library Carpentry*, 2016-2022).

After merging the data, I used OpenRefine to normalize the new dataset, taking steps such as adding extra variables and renaming a few variables to denote the application they came from. Fellow curators and data management professionals can easily update this data, and future researchers and analysts can directly use the merged data file without having to worry about data transformation. After merging and normalizing the data, I created a data dictionary and supporting metadata. This was done to provide complete documentation of the dataset and describe its main components, such as variable names, descriptions, and naming conventions. After reading this documentation, users and curators can learn more about the data and update the data or the curation process as needed. I provided the two raw datasets from TikTok and Spotify, with no modifications made, as well as my full merged dataset, which contains all the normalization efforts described earlier in this document. This decision was deliberate, as researchers could benefit from having the raw data to not only compare it to the combined data but to perform analysis on one specific application's charts only, having data available to them immediately rather than having to filter it out themselves.

I chose GitHub as my data repository since it is a stable repository and open to the public. It is entirely free and easy to use, meaning users can easily find and download the data. The GitHub interface in particular is easy to navigate, meaning users can find and download data quickly. The data's CSV file format helps with the accessibility and reusability as well, since it is also open and requires no proprietary software. Since I have a few years of experience using GitHub, it also felt natural to use it for this project, as the interface and processes were familiar to me.

My main concern about the curation of this data is the small picture this data provides users with about 2022's music charts, as it was only the top 100 songs over a year. A larger dataset, or individual datasets that could later be joined, of the top songs over each month of 2022, could have given researchers a broader picture of the music charts in 2022 across these apps. I chose not to do this for a few reasons: firstly, the TikTok dataset I gathered only had the top songs over a twelve-month period, meaning that I would have to learn how to work with a separate API to get monthly data alongside monthly Spotify data; second, given the scope of this assignment, I thought that refining and curating a dataset that large would require a significant amount of labor, almost as much as the API work. I decided that a larger, more unwieldy dataset would be more difficult to manage actively over time, and that this smaller sample of one year's worth of charts, rather than individual months, is an example of "carefully excavated" data, as discussed in Catherine D'Ignazio and Lauren Klein's *Data Feminism*. Larger datasets are usually less tractable, which makes it harder for potential re-use and analysis, and with this assertion in mind, D'Ignazio and Klein point out that smaller datasets are "easier to trace back to [their] material and cultural contexts," (D'Ignazio/Klein, 2018). Data that requires more labor to gather will consequentially require more labor to curate and manage over a long term, and eventually reuse, and to avoid an unfair amount of labor, I chose to value tractability and make this broader yet smaller sample of a year's worth of TikTok and Spotify data. While the items within my dataset are rather expressive and detailed, the dataset's relatively small size makes it more tractable and easier to trace back to its original context, which the inclusion of URIs helps with too. Since potential users should know the context behind the decision to only feature a yearly chart, I decided to include this potential "data quality issue" in my repository's Readme file.

Gathering and refining this data showed me the importance of well-curated and reusable research data. Since there is not much data journalism or analysis work that compares the presence of specific music on different platforms or performs a comparative analysis of trends in social media and streaming music charts, I believe that there is a high potential for reuse, as comparative analysis with other applications can help researchers discover the extent of TikTok's impact on cultural production.

References:

D'Ignazio, C., Klein, L. (201 8). *Data Feminism*. MIT Press. https://data-feminism.mitpress.mit.edu/

Dini, L. (2007). "Dublin Core Metadata Standards" (Translated). Conference on Digital Libraries and Metadata, University of Bolzano. https://slideplayer.com/slide/4908383/

Montag, C., et. al. (2021). "On the Psychology of TikTok." *Frontiers in Public Health* vol. 9, no. 1: 73-79. https://www.frontiersin.org/articles/10.3389/fpubh.2021.641673/full

Wickham, H. (2016). "12.4.2: Uniting Data." *R For Data Science*. https://r4ds.had.co.nz/

"Spotify Metadata Style Guide" (2021) Version 2. Accessed 1 March 2023. https://cdn.smehost.net/helpcentertheorchardcom-orchardprod/wp-content/uploads/2019/07/Spotify-Metadata-Style-Guide-v.2.pdf

"Tidy Data for Librarians" (2016-2022). *Library Carpentry*. Accessed 2 March 2023. https://librarycarpentry.org/lc-spreadsheets/