Sierra Jorgensen

LIS 545 - Winter 2026

Final Project Draft

Dataset: https://www.kaggle.com/datasets/syedaeman2212/top-100-songs-on-spotify-in-2025

Data and Metadata Profile

Introduction

This data set, which ranks the "top 100 songs on Spotify in 2025," is limited in scope. There is an assigned metric of 100 songs that narrows the allowable data, and the data set itself consists of only one CSV file that can be opened in Excel or Google Sheets. It contains only nine column headers and excludes other metadata elements such as *Instruments* or *label.* Metadata as the "establishment of knowledge, evidence, and truth" (Mayernik, 2020) is richly developed in the study and organization of music. One example of a music-focused metadata standard is published by MusicBrainz, an online music encyclopedia that also hosts a schema of term relationships created in SQL. The data and metadata expansion possibilities of this data set are wide.

Data

The data in this data set are both numeric, such as *1, 2, 3, 265.47* and *98,* as well as qualitative, like *Indie, Justin Bieber,* and *No.* An initial review flagged some possible errors in the presentation of the data in the CSV file: artist *Taylor Swift* is assigned the genre *Latin,* for example, which is an entity relationship that lacks real world support. Inconsistencies like this may represent problems in the collection stage of the Data Life Cycle (DataONE, n.d.), likely due to the fact that this is a synthetic dataset, "mimic[king] the patterns found in real-world data" (Shahid, "Provenance," n.d.).

Synthetic data has been explored as a viable method of protecting personal information, with one University of Washington research paper calling it 'a fundamental component of "people analytics," where sensitive, private data must be used to make high risk decisions'

(Howe, 2017). However, without sufficient metadata describing the sourcing and cultivation of the dataset, it is difficult to understand where in the Data Life Cycle problems truly arose.

Noting issues in the reliability of the data, it is important to separate interest in the topic at hand from interest in this specific curation. There are several key stakeholders to this data, generally speaking. Artists and their management teams, Spotify as a streaming service, advertisers, and listeners may approach this data from different perspectives, but it is meaningful to each. However, this specific data set as posted to kaggle.com has fewer stakeholders due to its small scale, limited metadata, and the popularity of publications sharing data of the same concept. The "About" section of this data set on Kaggle calls out its usefulness for data analysis projects, so the specific stakeholders invested in this data set are likely students researching data and metadata as opposed to musicians or streaming services.

Another detail showing support for student use of this dataset is its license. The dataset is licensed under Apache 2.0, which is an open source license that allows the user to "produce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form" (Apache, 2004).

Metadata

The metadata of the actual data set as represented by the column headers of the CSV file are *Rank, Song_Title, Artist, Genre, Release_Date, Spotify_Streams_Millions, Popularity_Score, Duration_Seconds*, and *Explicit*. The metadata of the data as hosted on kaggle.com are *Collaborators, Sources, Collection Methodology, License*, and *Expected Update Frequency.* The metadata is presented on the kaggle.com "Data Card," as well as in the CSV file of the data itself.

This is a small set of data with limited metadata. The metadata provided is sufficient for the purpose of the dataset, which is to show the most popular songs on Spotify in the year 2025, but may not be comprehensive enough to serve any other purpose. For example, *Spotify_Streams_Millions* and *Popularity_Score* describe numerical data answering the very question posed by this dataset. However, metadata for the sourcing and curation of the data is largely missing from the Data Card, limiting its trustworthiness and overall usability =.

This may additionally be in part because this appears to be ad hoc metadata and does not follow the Dublin Core or other well known metadata standards. Key to the issues described above is the absence of information regarding *Source,* one of the thirteen Dublin Core Elements cited as "[o]bjects, either print or electronic, from which this object is derived" (Weibel, 1995). The metadata of this data set seem compiled "as lubricants in disjointed, imprecise scientific communication" (Edwards et al. 2011, 684).

Generally, the data set is discoverable due to the descriptiveness of the ad hoc metadata elements used, like *Duration_in_Seconds.* However, it is unclear what some of the other metadata truly represent once removed from their place in the original CSV file and due to their unknown origin. *Popularity_Score* is an element meaningful in context but completely useless without further analysis. What is the score based off of? What other data inform the score? Who attributed the score? A user would likely be unable to use this data in any of their own writing or research without first figuring out how the popularity scores are calculated; at that point, one might have found a better dataset for their purposes.

No publications are listed or provided with the data set. This specific data set on kaggle.com comes up on the first page of results when Googling "top 100 songs Spotify 2025." Other Google keyword searches, like "Eman Shahid" (the named contributor of the dataset) and

“Eman Shahid Spotify” don’t bring up any additional sources. Likewise, neither this dataset nor its contributor appear in a UW Libraries search. The purpose of this data, to show the most popular songs on Spotify in 2025, is published in various lists by Spotify itself as well as many popular blogs and websites, like Pitchfork.

While a small data set consisting of only one CSV file, this data set is an interesting display of common, popular data rich with metadata and design expansion opportunities due to its current limitations.

Repository Profile: Internet Documents in Economics Access Service (IDEAS)

As discussed in the metadata profile of dataset "Top 100 songs on Spotify in 2025," there are issues in the presentation of the data as well as a dearth of provided metadata. Additionally, the timeliness of the dataset excludes many of the repository options available via the Registry of Research Data Repositories. Medieval music, scores of Bach, or Dutch songs through the ages are all represented, but recent Spotify stream counts are not. One promising repository called "One Million Songs" was produced from data collected by The Echo Nest, a Spotify-owned machine learning music intelligence program that appears, upon cursory research, to have been superseded by some sort of casino-focused blog (The Echo Nest, n.d.) and is not maintained nor current enough to be useful. However, further thinking regarding streaming services and the consumption of content realized roots in economic principles. Thus, "Internet Documents in Economics Access Service," or IDEAS, stood out as a reasonable fit for the dataset despite not being music-focused.

IDEAS is self-described as "the largest bibliographic database dedicated to Economics and available freely on the internet" (RePEc, n.d., para. 1). IDEAS draws upon data organized by Research Papers in Economics (RePEc), a service that collects machine readable metadata from original research in economics and organizes the primary data to enhance dissemination. Specifically, RePEc finds and mirrors metadata from archives into their own repository following the Guildford Protocol, a "set of rules for the publication and exchange of documents on the Internet" (Krichel, 2016). Per the Guildford Protocol, RePEc captures bibliographic data from institutions who archive their publications following the criteria outlined on RePEc's tutorial webpage: create or identify a usable server for data hosting, establish a web archive that

RePEc can access, and define metadata that is readable by RePEc's program. IDEAS encourages researchers to register with the site to claim authorship of any data mirrored from the web archives.

Institutions who host research in economics list their publications with RePEc, or individual authors may submit their work to the Munich Personal RePEc Archive if unaffiliated with a contributing institution. In 2020, a post entitled "Who are the authors registered with RePEc?" was published to The RePEc blog. The post states that there were over 60,000 authors with publications included in RePEc while there are only 25,000 members of the "three largest [economics] associations in the profession: The American Economic Association, the European Economics Association, and the Econometric Society" (The RePEc Blog, 2020). IDEAS organizes their registered users into two separate lists: those who designate themselves published authors while registering their account with RePEc, and those who do not.

RePEc provides detailed step-by-step instructions for contributors on exactly how to establish an archive with a machine readable layout and how metadata must be described and saved following the Research Documents Information Format, or ReDIF. ReDIF prioritizes simplicity and academic self-description, cataloging three classes of items: resources, tangibles, and collections (Krichel, 2015, 2.1). While the dataset "Top 100 Songs on Spotify in 2025" is not an academic publication, the dataset may utilize the template "ReDIF-paper" as a research report formally unpublished (Krichel, 2015, 2.1).

RePEc includes a "Step by Step" tutorial that prescribes the technological standards of the ReDIF templates in a few short sentences: "[a]ll these templates are held in simple text files. This means: no binary formatting as it is provided by word processors. No HTML markup. Always save files as text only. Also, the files holding templates need to have the extension *.rdf* or

.*redif*." (IDEAS, n.d., para. 2). IDEAS promotes access to economics research by both utilizing the simple ReDIF standard and translating the Guildford Protocol and their processes into a simple tutorial for researchers and readers alike.

Generally, there is not a personal exchange of information because RePEc pulls from archives using an automated process, not a direct submission. Volunteers who work on IDEAS may have contact information available in the "Personal Details" section of their IDEAS profile, such as email address, website, or social media handles. The repository suggests visitors contact Christian Zimmerman, described as being "in charge of" IDEAS (IDEAS, n.d., "Credits"), "only" if questions aren't answered in the FAQ.

IDEAS is an impressive repository for the dissemination of scholarly research. Simple for institutions to contribute, it is also easy for users to find and access data. No log in or registered credentials are required for users of IDEAS' website, making initial access easy and immediate. Results are plentiful as well: 573 hits were returned from a keyword search of *music* published in year *2025.* Clicking on one of the results brings users to a landing page with the authors, abstract, and citation provided. IDEAS provides the DOI and their own archive handle for this data, as well as the article's references, exportable in a variety of formats. References are extracted by CitEc, or Citations in Economics, an automated tool pulling references from documents listed by RePEc (CitEc, n.d.). Another subpage of this data shares article keywords and a link to some article statistics, including a graph and chart mapping abstract views and file downloads. These statistics are produced by LogEc, another tool of RePEc datasets that "collects accesses statistics" and "provides a convenient way of tracking trends in the profession" (LogEc, n.d., para. 1).

IDEAS does not host publications or articles, but indexes the bibliographic details of research that may be downloaded from external locations. While the metadata provided in “Top 100 Songs on Spotify in 2025” need translating to follow ReDIF standards, the simplicity afforded by the models RePEc uses and the opportunity to authors unaffiliated with academic institutions allows for the inclusion of this dataset.

Suggested Data Citation

Shahid, E. (n.d.). *Top 100 Songs on Spotify in 2025* [Dataset]. Kaggle.

https://www.kaggle.com/datasets/syedaeman2212/top-100-songs-on-spotify-in-2025

This citation follows the required principles outlined in the Joint Declaration of Data Citation Principles. The URL provided above is "globally unique," directs to a landing page unique to the dataset, and is machine readable (Fenner, et al, 2019).

Long-term Preservation Statement

This data set consists of one CSV file downloadable onto any device without downloading additional proprietary software. CSV files, or comma-separated values, are an open data format supported by Rule 4 of Hart, et al's "Ten simple rules for digital data storage" (2016) due to their non-prohibitive accessibility.

Copyright Statement

Apache 2.0 is an appropriate copyright license sufficient for this data set that includes non-confidential and non-sensitive data. The data contain no private or sensitive information, and the stated purpose of the dataset as displayed in the "About" section of its data card is for use in data science projects, so open access is reasonable.

Statement on Human Subject Considerations

Artist names are the only personal data in this data set. Spotify account data are not shared, and neither is any data that identifies private or sensitive information about the artists or producers of the songs listed in the data set. The data were not altered or adjusted for privacy because there are no real ethical or privacy concerns in the data in the first place, as in, there was no need for anonymization or randomization. Spotify as a service houses account holder information, including private financial data, however the synthetic data populated in this data set do not contain any user data whatsoever.

Works Cited

Apache Software Foundation. (2004). *Apache License, Version 2.0*. https://www.apache.org/licenses/LICENSE-2.0

Cargill, C. F. (2011). *Why standardization efforts fail*. *Journal of Electronic Publishing, 14*(1). https://doi.org/10.3998/3336451.0014.103

CitEc. (n.d.). *Frequently asked questions* (FAQ). https://citec.repec.org/faq.html

DataONE Best Practices Working Group. (n.d.). *Data Management Skillbuilding Hub*. DataONE. https://dataoneorg.github.io/Education/

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, 6(1). https://doi.org/10.1038/s41597-019-0031-8

Hart E, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, Poisot T, Woo KH, Zimmerman N, Hollister JW. 2016. Ten simple rules for digital data storage. *PeerJ Preprints* 4:e1448v2 https://doi.org/10.7287/peerj.preprints.1448v2

Howe, B., Stoyanovich, J., Ping, H., Herman, B., & Gee, M. (2017). *Synthetic data for social good*. *Bloomberg Data for Good Exchange*. https://faculty.washington.edu/billhowe/publications/pdfs/howe17bloombergdatasynthesizer.pdf

IDEAS/RePEc – Economics and Finance Research. (n.d.). https://ideas.repec.org/

Krichel, T. (2016). *Guildford Protocol*. RePEc and ReDIF documentation.

https://openlib.org/acmes/root/docu/guilp.html

LogEc. (n.d.). *About LogEc*. https://logec.repec.org/about.htm

Mayernik, M.S. (2020). Metadata. In B. Hjørland & C. Gnoli (Eds), *Encyclopedia of Knowledge Organization*. International Society for Knowledge Organization. https://www.isko.org/cyclo/metadata

MetaBrainz Foundation. (n.d.). *MusicBrainz Database*. MusicBrainz. https://musicbrainz.org/doc/MusicBrainz_Database

Munich Personal RePEc Archive. (n.d.). *Munich Personal RePEc Archive*. https://mpra.ub.uni-muenchen.de/

Shahid, E. (n.d.). *Top 100 Songs on Spotify in 2025* [Dataset]. Kaggle. https://www.kaggle.com/datasets/syedaeman2212/top-100-songs-on-spotify-in-2025

The Echo Nest. (n.d.). *The Echo Nest*. https://the.echonest.com/

The RePEc Blog. (2020, September 24). *Who are the authors registered with RePEc?* https://blog.repec.org/2020/09/24/who-are-the-authors-registered-with-repec/

Weibel, Stuart. 1995. "Metadata: The Foundations of Resource Description". *D-Lib Magazine* 1, no. 1. http://www.dlib.org/dlib/July95/07weibel.html