Project Report (Rename Appropriately)

Anonymous Author(s)

ABSTRACT

Over the past two decades, music has changed a lot. From rock to pop, then rap, to the music we hear today, we've seen different styles, ways of making music, and what people like to listen to. We used data sourced from Beatport to look at music from 2000 through 2023 for our analysis. With these two platform's data, we were interested in understanding how music has changed through several different factors, from human factors in the form of popular preferences to technological factors in the form of technological advances etc.

In this project we have used Google Cloud MySQL database to store Beatport data. After that we wrote python to connect to the database and used matplotlib to help us visualize and numpy to summarize the data. Not only programming, but we also used some papers to help us understand the project better. We concluded that in the early 2000's, rock and pop music may have been the dominant music, but as time went on, hip-hop and electronic dance music grew in popularity. Each genre of music may have its own specific keywords or phrases in its song titles. For example, in electronic dance music, words such as 'love', 'night' or 'dance' may be more common. In hip-hop, words related to street culture and freedom may be seen more often.

KEYWORDS

Music streaming; data analysis; music preferences; streaming services; Spotify; Beatport; listening habits; musical landscape; music patterns; 21st-century culture; genre analysis analysis; audio features; responsible data analysis; music trends; fair recommendation systems; music variety; API data collection; usage quotas

1 MOTIVATION AND GOALS

Music streaming sites like Spotify and Beatport have had a profound influence on 21st-century culture, with access to huge music libraries that have low barriers-to-entry that has never been seen before. As a result, these platforms have reshaped the music industry by empowering independent artists and transforming listening habits, contributing to a dynamic and ever-evolving musical landscape. Jack Webster[21] shows in his research that music tastes have evolved significantly with the emergence of music streaming, with people allowing to have more individual, unique tastes as Spotify and other music streamers recommend songs to individual users that allow them to easily discover music. Robert Prey[15] even argues that streaming services and the individuation that they allow for has had an effect on not just how music is consumed, but even how music sounds and is created.

This data set encompasses music tracks from 2000 - 2023, which will allow us to see the full impact of music streaming on the music itself by looking at the music features over time and how it has been affected from before streaming became widely available to today. Most streaming sites, such as Spotify and Beatport, were founded in the early 2000's, and gained popularity in the 2010's, as Van Dyke[20] stated in his article. Our goal with this data is to find patterns in the music features, such as music titles, how the music sounds, and how each genre has been changing, in the age of music streaming.

Our motivation to pick this data set was founded in a few areas of interest. Firstly, we believe the data set is quite unique, as some of the data, especially from the Beatport data, has attributes that are quantified and recorded that are usually seen as more or less subjective, such as how 'acoustic' a song is, how 'danceable' it is, etc. Being able to use data mining and analytic techniques to find patterns in these values is something that we would find really interesting. Another motivation to using this dataset and carrying out the analysis is that there are two independence data sources, Spotify and Beatport. This allows for us to explore data engineering techniques to remove duplicates and engineer the data best, especially with a large data set that was given.

This report is organized as follows. Section 2 discusses work done by others music streaming area, including how music tastes have changed during the past 20+ years. We will also look at previous work into data analysis in music, and what previous patterns have been shown. Section 3 presents a discussion of the design, architecture, and implementation of the data set, including cleaning of the data, and any normalization that was done. It will also go in detail of the transferring of data from the CSV files to a PostgreSQL database. Section 4 presents an analysis of the project, including the patterns that we found, and any clustering of the data that we find. This section will also go into detail of the lessons learned, and the difficulties found while during the analysis of the data. Section 5.2 discusses the legal considerations of the issues relevant to our project and section 6 discusses the ethical considerations of the issues relevant to our project. Section 7 presents the current state of the project, possible future work, and then concludes with a few final remarks.

2 RELATED WORK

Given the growing ubiquity of online streaming services, a sizable amount of research has been done in this area. Large data sets from streaming services like Spotify are relatively easy to obtain, allowing for researchers to freely conduct analysis. As the service subscription model replaces the previous ownership model, some researchers have explored how this shift has influenced music consumption patterns. Bronnenberg et. al.[8] found that the adoption of streaming services has increased new music discovery, and the variety of music that any individual user chooses to listen to. From the perspective of smaller artists, this makes it easier to get into the ears of listeners, but also makes it more difficult to stay in the spotlight.

Much research has also been done about the best predictors of popularity. For example, Sciandra and Spera[16] found associations between the audio features of tracks and Spotify's own popularity metric. While similar analyses have focused on linear and quadratic regression models, they used a Beta model with mixed effects. The audio features positively correlated with popularity were energy,

valence, and song duration. The audio features negatively correlated with popularity were speechness, instrumentalness, and liveness.

Based on the same Spotify data, Munoz et. al.[19] also explored predictors of song popularity in multiple countries around the world. In addition to the various metrics provided by Spotify, they also included many external factors in their analysis. For example, by taking into account daily temperature in various regions in the world, they found that Italians tend to listen to more danceable music in the spring and summer. With regard to the pandemic, they found that danceability of popular songs was negatively correlated with the number of Covid-19 cases in a particular country. Based on their analysis of historical patterns, they designed a model to predict the preferred type of music in the near future within a time frame of four months.

In a more focused exploration of the effects of the pandemic on music streaming trends, Timothy Yu-Cheong Yeung[23] looked into the popularity of nostalgic songs during the time of the pandemic. He found that Spotify users in the UK preferred older, more nostalgic, music when the pandemic began. Compared to the same time the previous year, users listened to more music older than five years.

3 DATA ENGINEERING

The data structure had to be engineered in a way to improve the data analysis and set up to allow efficient querying of the data. This required an initial setup of the database, cleaning and refactoring of the data, and improving data design and integrity. We also acknowledged further optimizations that could be done, and would allow more efficient data processing and querying, especially for larger data sets. But as this was out of scope of the project, we deemed it unnecessary to implement, but still outline what some of those changes could be.

3.1 Database Selection and Setup

McFarland's data set[12], after a quick data exploration, has been significantly normalized and set into a relational database format, so engineering the CSV files containing the data tuples to a relational database was relatively straightforward.

First, we had to select a relational database management system, which we selected MySQL. We selected MySQL rather than PostgreSQL, Microsoft SQL Server, or Oracle because of the findings by Poljak et al. [14], which showed several characteristics of MySQL that we found ideal for our project. Firstly, MySQL is an open-source project, which although it has a enterprise-level that requires payment, it's base-tier is free. MySQL was has also direct support with Python, which was the coding language we will use in both data engineering and analytic tasks. Thirdly, there is well-supported and documented, as it has been an industry standards for years and has been maintained by Oracle. Realistically, many of the current RDMS would have been acceptable for our needs.

After selecting our RDMS, we needed to create a database server to host our data. We decided against creating a local server, as there was several disadvantages in doing so, including cost, availability, and scalability. Instead, we decided to use a cloud-computing service, as it has a free tier available for research and development, it has high availability, as it's maintained by a third-party, and it's able to scale dynamically based off the needs of the RDMS. Amazon Web Services, Google Cloud, Microsoft Azure, and other cloud providers provide very similar services to host relational data, and Google Cloud had the most available documentation and was the easiest to set up.

We used IntelliJ DataGrip to import the data from the CSV files to the MySQL database hosted on Google Cloud. IntelliJ DataGrip is an IDE that specializes in relational database management [7], and is available with the educational subscription for free under as an RIT student. DataGrip has built-in import features for CSV files, which allowed for simple initial importing of the data.

3.2 Data Design and Integrity

Some engineering is required to help with the data quality and performance. To improve data quality and performance, we set up primary and foreign keys to create and enforce relationships between tables. These relationships were laid out in McFarland's data set [12]. To create the primary keys, simply manually removing the duplicate rows that may have occurred due to import errors, and setting the constraint. This was done manually, as there was few import errors that led to duplicate values. Setting up the foreign keys was a bit more tedious, as importing would lead to missing values that would lead to missing references in other table's foreign keys. To check for missing values, we looked for distinct values in the foreign keys that were "missed" in the referenced table. Using those values, we would search and "re-add" them from the original CSV file. Once there are no distinct values in the table that can't be referenced, we can then create the foreign key without errors. Once all of the relationships have been created, you are presented with the interconnected diagram shown in Figure 1.

To further improve data quality, we added constraints on some of the attributes to reflect real-world standards. Some date-time attributes such as updated_on, duration, release_date, etc. were initially TEXT data types, and were changed to TIMESTAMP, HOUR, DATE, etc. respectively. This will allow for efficient and ease in comparisons during the analysis of the data in Section 4.

ISRC, or International Standard Recording Code [11], is a set code standard that references a sound or music video recording, and allows artists to collect royalties on their music. This code is referenced in multiple tables, and allows us to create a relationship between the two independent data sets. The code requires a 12-character length of text; the first two characters being "US", "QM", or "QZ", the next 3 characters being alphanumeric characters, and last 7 characters being numeric values. In order maintain this code standard, we set the ISRC attributes to a char(12), and created a check to enforce the constraints laid above. This will ensure that any new rows, and any changes made to existing rows, will adhere to the code requirements.

We also analyzed the data for replicated data within the database, and found multiple attributes that were already directly or indirectly contained in other attributes. An example of this was the *duration* and *duration_ms* columns found in multiple tables. Both columns symbolized the same attribute of a tuple, the duration of a specific track or song, but the *duration* column was a text value that varied in format, and *duration_ms* was an integer value that was more

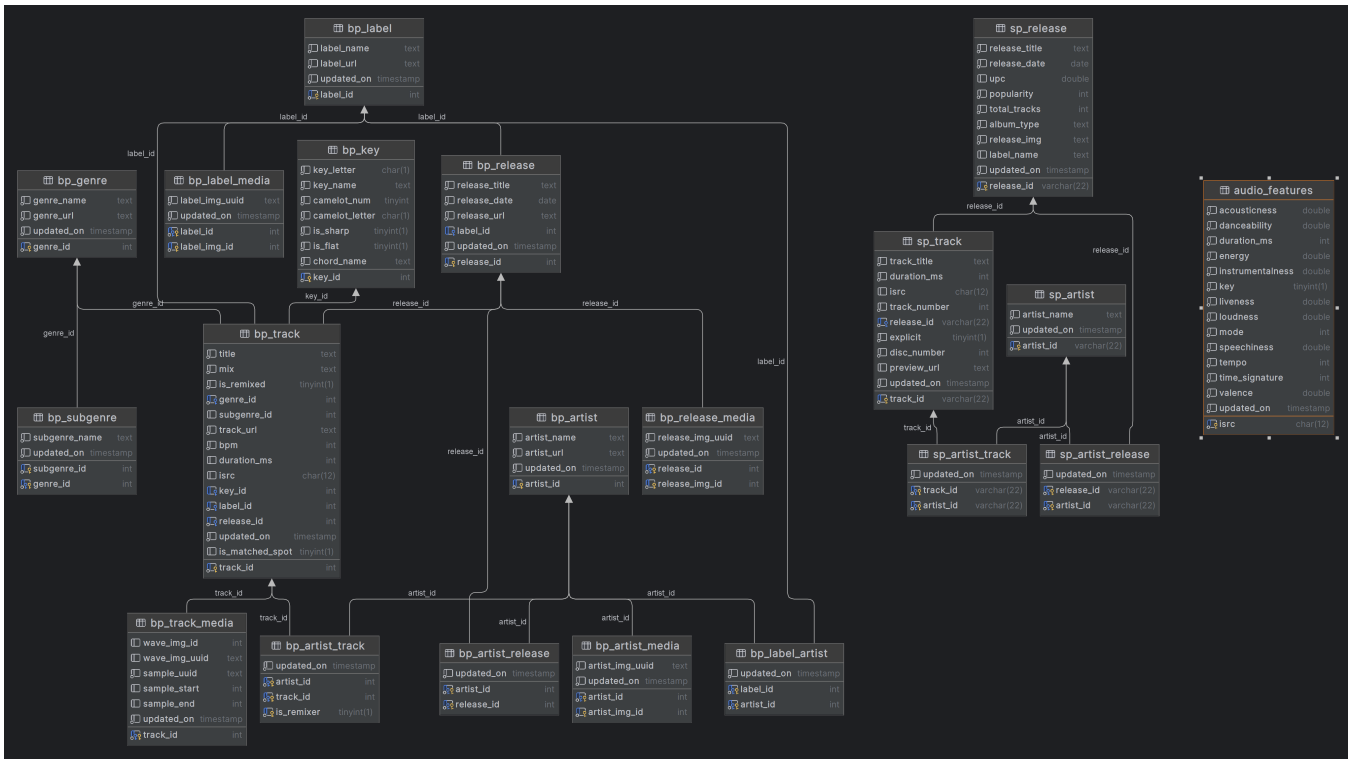

Figure 1: Entity-Relationship diagram of the music dataset

specific, we decided to remove the *duration* column to remove the duplicate data.

3.3 Future Optimizations

Future data engineering will go hand-in-hand with future development with the data analysis. There are attributes that can be further broken apart and normalized, but may lead to less optimal and efficient data analysis due to more complex and resource intensive queries from increased joining of tables together. For example, the chord related attributes in the *bp-key* table can be separated into their own table, but that may not be necessary or desirable based off the needs of the data analysis.

There are also examples of NULL values within attributes, which is an often clear sign of possible improvement in database systems[6]. A clear example of this is the *genre_id* and *subgenre_id* attributes in the *bp_track* table. All songs have a genre, but some songs have also a subgenre classification, and some songs don't. This could be normalized through a separate bridge table containing the *track_id* and *subgenre_id*, which would remove the use of NULL values. But this would require an increase of joining the tables together through the bridge table as described above. So although NULL values do exist within the current system, we believe that their existence allows for improved efficiency and processing of the data, but recognize that it isn't perfectly clean.

Overall, McFarland cleaned and engineered the data very well, and much of the data engineering was related to the transferring of the data and SQL semantics. Further improvement can be made, but are not entirely necessary, as there is not large amounts of duplicated data and further normalization may slow down queries or make them unnecessarily complicated.

4 DATA ANALYTICS

Data analysis was done using Python on data queried from a MySQL database.

The dataset contains music beginning from the year 1900 to the present, however it is significantly weighted toward the present with the vast majority of tracks being from the past three years. The distribution of tracks over time is shown in figure 2. Although it is possible that more music is being created overall, this is likely a result of the Beatport dataset being explored. Users may have a strong preference for recent music which strongly influences the distribution of tracks over time.

Aside from outliers appearing at the beginning of the 20th century, the duration of tracks increases slightly over time. The average duration of tracks in seconds over time is shown in figure 3.

Artists tend to produce music only within one or two different genres. However some artists produce music in over twenty different genres. A histogram of the number of genres produced by a single artist is shown in figure 4. This confirms the intuition that artists tend to produce music of a specific genre and rarely branch out to a wide variety of genres.

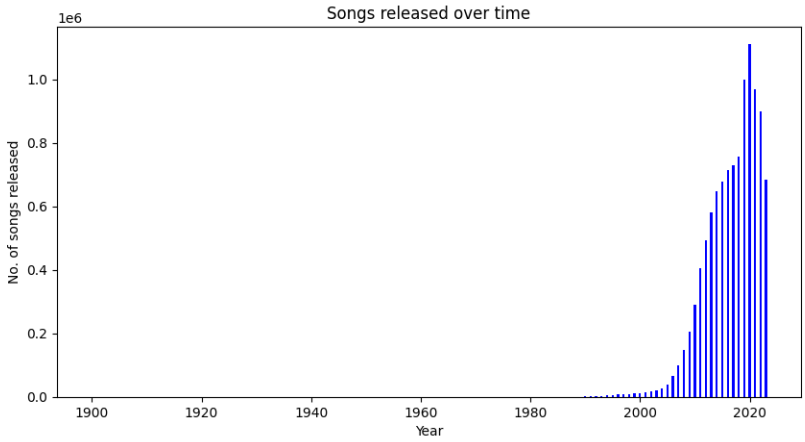

Figure 2: The number of tracks created over time

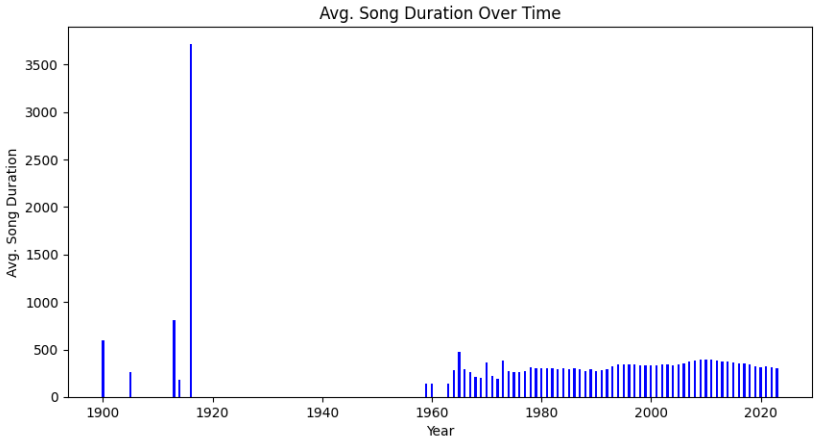

Figure 3: The average song duration over time

5 LEGAL CONSIDERATIONS

Working with third-party data involving music streaming platforms requires legal considerations of how the data is collected, as well as legal considerations relevant to music streaming and music streaming data. We will investigate into how our data was collected, and considerations that were or were not made, as well as general considerations when using third-party APIs. We will then also consider current legal issues involved within the music streaming industry, include AI-generated music, music streaming royalties, and transparency of data processing.

5.1 Data Collection

Scott McFarland[12] specifically laid out information about the dataset, including how the dataset was created. He specifically mentions that he created 6 separate accounts to bypass Spotify usage limits, and provides instructions of how to bypass these limits. This is in clear violation of Spotify's developer terms[18], specifically Section VI. Section VI pertains to access, usage, and quotas for developers. Under this section, they state that if a Spotify developer account reaches a quota limit, they can apply for a quota extension. They also specifically state that the use of separate developer accounts for the same application is strictly prohibited.

The collection of data for this specific data-set should not be deemed as a crime, as Scott McFarland's actions didn't have a significant impact on Spotify or any of it's functionalities, and it's violations of the legal terms was harmless. But this should be noted for future applications where collection of data from private third-parties is necessary. Legal terms and possible violations, such as usage quotas, should be noted and followed. Violations of these terms, however unlikely especially for personal development projects, can have lasting impacts and could lead to further API usage quotas and limitations, such as those seen by the Reddit Developer API[22].

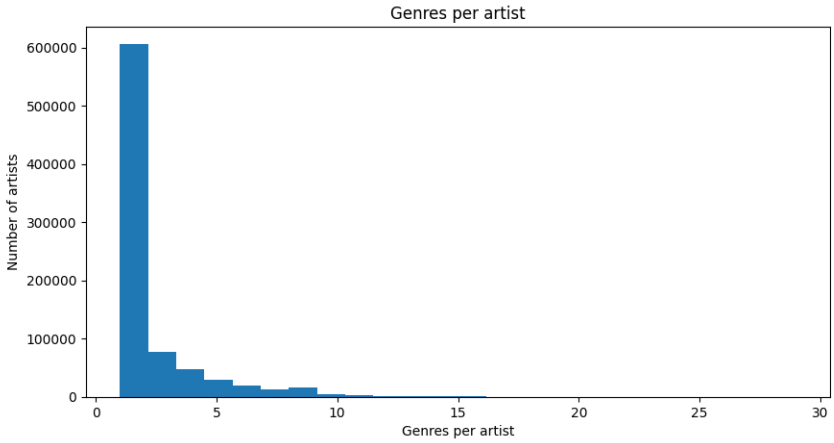

Figure 4: The number of genres per artist

5.2 Music Streaming

Music streaming and the generated data from these platforms have ran into various legal issues and scrutiny. First, music professionals are now concerned with the rise of AI-generated music. An article out of CNN[24] have shown that music companies like Universal Music Group have urged streaming services like Spotify and Apple Music to take on the problem of AI songs impersonating their artists. It is extremely difficult for these streaming services to detect and regulate AI generated content. Currently, there are few legal protections within this domain. The U.S. copyright office[13] has only given loose guidelines around AI-generated media. They state that they will analyze any given work based on the degree to which AI was used, and in what capacity. If in the future, streaming services like Spotify and Apple Music become required by law to filter out illegitimate AI-generated music from their libraries, this could create a major technical problem for them. As AI technology improves, it will become increasingly difficult to distinguish between genuine human-made songs and artificially generated ones.

Second, as music streaming services have taken over the music industry from traditional forms of distribution, songwriters are concerned they are not being compensated fairly, according to a recent article from Variety[3]. Part of the problem stems from a system set up for a record and CD-dependent industry of years past. Under this system, the majority of the royalty went to the performing artist and record label and the rest to the songwriters and publishers. This division made sense given the significant cost the label took on for the manufacturing and distribution of records and CDs. The Copyright Royalty Board published a final determination in 2018[5], taking a step in favor of songwriters by increasing the publisher's share from 11.4 to 15.1 percent. Currently, neither streaming services nor music labels are interested in sacrificing some of their share towards songwriters and publishers citing thin margins, and competition from other forms of media.

Lastly, as music streaming services deal with the personal information of their users, they are subject to data privacy regulations like the General Data Protection Regulation. Included in the GDPR[10] is the requirement for data controllers to communicate to users what data is being taken and in what way it is being used. If found to not be in compliance with these regulations, fines or other consequences may take effect. For example, in June of 2023, the Swedish Authority of Privacy Protection[9] (IMY) fined Spotify around 5 million euros for inadequate transparency with users about their data. This fine was levied because the IMY found that Spotify was not clear in its description of data use.

6 ETHICAL CONSIDERATIONS

Ethical considerations relevant to music streaming and music streaming data include data security, preventing fraudulent activity, proper communication with the public, and avoiding discrimination. These considerations will be discussed in relation to the ACM Code of Ethics.

Article 2.9 of the ACM Code of Ethics[2] relates to the importance of security in computing systems. As music streaming services currently process immense amounts of user data, they must take strong measures to protect it. In addition to protecting their data systems from outside attacks and breaches, other threats exist that are more difficult to identify. Streaming fraud is another example of fraudulent activity in music streaming services. Streaming fraud[1] occurs when individuals or groups create meaningless tracks and use bots to stream them at scale. Additionally, certain groups advertised as marketing agencies may sell their services to artists and use bots to increase the streams on those artists' songs. Apple music has recently taken measures[1] to reduce streaming fraud on their platform Apple Music. They now conduct reviews and deem certain streams fraudulent. Content providers are given monthly reports which include details on fraudulent streams. Offenders will also face financial repercussions or possible account termination. Despite these measures streaming fraud remains a large problem on their platform and further steps must be taken.

Article 2.7 of the ACM Code of Ethics[2] requires that computing professionals take measures to communicate their work and its impacts to the public. Whenever an individual or company is in charge

of a large-scale application like a music streaming service, they have an ethical obligation to offer some transparency to their users. The public should know how their data is being used, and of any important decisions that might affect them. Spotify[17] does this in part by publishing internal research on their R and D website. For example, Spotify researchers discussed the difficulties in designing fair and unbiased recommendation systems. They cited numerous challenges including competing fairness metrics, scarce implementation guidance from critics, and properly identifying appropriate areas of the system to implement changes. In this way, Spotify has made an effort to uphold ACM 2.7 by publicly communicating their efforts towards the fairness of their application.

In addition, this same Spotify research [4] relates to Article 1.4 of the ACM Code of Ethics [2] which requires computing professionals to prioritize fairness and avoid discrimination. Although preliminary, it represents a proactive step taken by Spotify to improve fairness on their platform.

7 CONCLUSIONS

Data analysis was done to understand and find patterns in the data. Trends were explored to discover how music has changed over time. Top genres were analyzed individually to find audio features associated with each. Genres were also analyzed by what words tended to occur in the titles of their tracks.

Future work may continue to find patterns and meaningful insights in the data of music streaming services. Interesting associations may be found between tracks made by the same artist. Artists may tend to prefer certain audio characteristics and share them across all their work. In addition, given the trends in the current data, qualities of future music may also be predicted.

This work was an investigation into the data of music streaming services. Such services and their data will only become more prevalent and impactful in the world. It is important for data analysts to ethically and responsibly use data to improve the way users consume music.

REFERENCES

[1] Apple 2023. *Apple: Apple Music streaming fraud reduced by 30%: How the company did it.* Retrieved November 20, 2023 from https://www.gadgetsnow.com/apps/apple-music-streaming-fraud-reduced-by-30-how-the-company-did-it/articleshow/105362527.cms

[2] Association for Computing Machinery. 2018. *ACM Code of Ethics and Professional Conduct.* https://www.acm.org/code-of-ethics.

[3] Jem Aswad. 2022. *Inside the multi-billion dollar battle royale over music-streaming royalties.* Retrieved November 20, 2023 from https://variety.com/2022/music/news/streaming-royalties-music-biz-dsps-spotify-1235327760/

[4] Lex Beattie, Dan Taber, and Henriette Cramer. 2022. Challenges in Translating Research to Practice for Evaluating Fairness and Bias in Recommendation Systems. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 528–530. https://doi.org/10.1145/3523227.3547403

[5] United States Copyright Royalty Board. [n. d.]. Rates and Terms for use of Nondramatic Musical Works in the Making and Distributing of Physical and Digital Phonorecords. https://www.crb.gov/rate/16-CRB-0003-PR/attachment-a-part-385-regs.pdf

[6] Ching-Hsue Cheng, Liang-Ying Wei, and Tzu-Cheng Lin. 2007. Improving Relational Database Quality Based on Adaptive Learning Method for Estimating Null Value. In *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*. 81–81. https://doi.org/10.1109/ICICIC.2007.350

[7] DataGrip Core Team. 2023. *DataGrip: The Cross-Platform IDE for Databases.* JetBrains, Prague, Czech Republic. https://www.jetbrains.com/datagrip/

[8] Hannes Datta, George Knox, and Bart J Bronnenberg. 2017. Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery. *Marketing Science* 37, 1 (April 2017), 5–21. https://doi.org/10.2139/ssrn.2782911

[9] EDPB 2023. *IMY issues an administrative fine against Spotify for shortcomings regarding transparency.* Retrieved November 20, 2023 from https://edpb.europa.eu/news/national-news/2023/imy-issues-administrative-fine-against-spotify-shortcomings-regarding_en

[10] European Parliament and Council of the European Union. [n. d.]. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj

[11] ISRC 2023. *International Standard Recording Code.* Retrieved November 21, 2023 from https://usisrc.org/about/index.html

[12] Scott McFarland. 2023. *10+ M. Beatport Tracks / Spotify Audio Features.* Retrieved October 19, 2023 from https://www.kaggle.com/datasets/mcfurland/10-m-beatport-tracks-spotify-audio-features

[13] U.S. Copyright Office. 2023. *Copyright and Artificial Intelligence.* Retrieved November 21, 2023 from https://www.copyright.gov/ai/

[14] R. Poljak, P. Poščić, and D. Jakšić. 2017. Comparative analysis of the selected relational database management systems. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 1496–1500. https://doi.org/10.23919/MIPRO.2017.7973658

[15] Robert Prey. 2018. Nothing personal: algorithmic individuation on music streaming platforms. *Media, Culture & Society* 40, 7 (2018), 1086–1100. https://doi.org/10.1177/0163443717745147

[16] Mariangela Sciandra and Irene Carola Spera. 2020. A model-based approach to Spotify data analysis: a Beta GLMM. *Journal of Applied Statistics* 49, 1 (Aug. 2020), 214–229. https://doi.org/10.1080/02664763.2020.1803810

[17] Spotify RD Research 2022. *Research Areas: Algorithmic Responsibility.* Retrieved November 20, 2023 from https://research.atspotify.com/algorithmic-responsibility/#:~:text=Research%20in%20algorithmic%20responsibility%20at,that%20teams%20can%20actually%20use.

[18] Spotify Terms 2023. *Spotify Developer Terms.* Retrieved November 21, 2023 from https://developer.spotify.com/terms

[19] Fernando Terroso-Saenz, Jesus Soto, and Andres Muñoz. 2023. Evolution of global music trends: An exploratory and predictive approach based on Spotify data. *Entertainment Computing* 44 (2023), 100536. https://doi.org/10.1016/j.entcom.2022.100536

[20] Elise VanDyke. 2021. *The Rise of Music Streaming Services.* Retrieved September 24, 2023 from https://globaledge.msu.edu/blog/post/57046/the-rise-of-music-streaming-services

[21] Jack Webster. 2019. Music on-demand: A commentary on the changing relationship between music taste, consumption and class in the streaming age. *Big Data & Society* 6, 2 (July 2019), 2053951719888770. https://doi.org/10.1177/2053951719888770

[22] Kyle Wiggers. 2023. *Reddit will begin charging for access to its API.* Retrieved November 21, 2023 from https://techcrunch.com/2023/04/18/reddit-will-begin-charging-for-access-to-its-api/

[23] Timothy Yu-Cheong Yeung. 2023. Revival of positive nostalgic music during the first Covid-19 lockdown in the UK: evidence from Spotify streaming data. *Humanities and Social Sciences Communications* 10, 1 (March 2023), 1–12. https://doi.org/10.1057/s41599-023-01614-0

[24] Vanessa Yurkevich. 2023. *Universal Music Group calls AI music a "fraud," wants it banned from streaming platforms. experts say it's not that easy.* Retrieved November 20, 2023 from https://www.cnn.com/2023/04/18/tech/universal-music-group-artificial-intelligence/index.html