Data and the Music Industry:
How Various Genres Encapsulate Spotify Trends

Shweta Dhar, Misha Gandhi, Mahnur Khalid, Kavya Moharana, Samidha Sampat

University of Illinois at Urbana Champaign: School of Information Sciences

IS 204: Research Design for Information Sciences

Professor Stephanie Besser

December 15, 2023

Abstract

The objective of our research is to discover how we can apply data analysis techniques to understand Spotify trends within a user's music and listening history to optimize promotional campaigns and music releases on Spotify. We will create predictive regression models to better understand the aspects that contribute to trends within user behavior on Spotify, specifically to understand which factors help shape promotional campaigns and music releases and also aspects that contribute to users' listening history. Furthermore, we will focus on how a user discovers their most listened to genre. This will allow us to understand how to structure future advancements for Spotify.

Our analysis uses the Spotify Top Hits Playlist (2010-2022) dataset containing data on various users from the music platform, and including information ranging from frequently-used playlists to most popular artists. We will be applying Python libraries, specifically: Pandas, Matplotlib, and Sci-KitLearn, along with varying statistical approaches, such as summary statistics. To fully understand the current trends, we will be visualizing the data through predictive modeling visualizations and regression analysis.

We were able to gauge that user algorithms and popularity in the media have a significant impact on Spotify user trends, which allows for the creation of user-personas. The data is also reliable because the users do not exhibit bias. The application of the dataset through Python and statistical methods have contributed to findings that align with our research question pertaining to user behavior, trends in media, and Spotify's generative technology.

Introduction and Literature Review

In a world where music is a significant part of life, we wanted to discover how data analysis techniques can be applied to understand Spotify trends within a user's music and listening history. Spotify is a music streaming service where users are able to make their own playlists, search for music, connect with friends, and make their own profiles. A wide variety of people ranging from pre-teens to middle-aged adults are familiar with the platform and have used it for many years. It is interesting to look at how music has developed over the years of using Spotify. Spotify also released Wrapped which collects data from a year and shows users which trend they have adapted. Specifically, looking at understanding which factors help shape user behavior is of particular interest. This will in turn allow understanding on how different factors help shape promotional campaigns, music releases, and what contributes to users' listening history. Not only that, but it will allow for some insight on future advancements for Spotify. By analyzing this topic, the application of understanding various user behaviors holds the potential to unlock insights into factors like effects on popularity, trends in popular songs, user preferences, and Spotify's evolving strategies and new technology to attract users to the platform.

Data analysis and machine learning algorithms refine recommendations and user driven results to understand the popular music in the industry. This allows members of the industry, singers and musical professionals to make informative decisions on new songs, lyrical content, emotional tones, popularity, and user personas. This is impactful as it gives music a new outlook on metrics to consider and optimizes their performances. All in all, it helps increase further understanding of popular music.

The research that we are focusing on relates to the topic of significance of user behavior within the Spotify platform, and its impact on different trends that arise within the platform. User behavior research is significant no matter the industry, because it is the basis of understanding consumer needs – an extremely integral part towards a company's success. Properly utilizing regression analysis and Python to understand which factors create users' music taste and trends within users' listening history is valuable in retaining Spotify's customer base. Furthermore, it integrates proper assets to improve the platform, optimizing Spotify's image. Marketing and emotions are often the most utilized approaches to increasing success of promotional business campaigns (Anderson, 2020, p. 2155). Applied computing is also a focus on retrieving the proper data to create the best recommendations for users on certain platforms in order to optimize their music taste. Specific Spotify behavior that has been studied have focused mainly on how its current algorithm focuses on diversifying users' music, alongside session-switching amongst users (Zhang, 2013, p.220). Additionally, understanding past research about global evolution of music preferences and its effects on listening trends (Terroso-Saenz, 2023) is essential to determining the answers to our current questions.

Methods

In finding data for this study, we searched for information related to user trends on music streaming platforms. We obtained our chosen dataset, "Spotify Top Hit 100 songs from 2010 to 2022", from Kaggle which contains 23 variables encompassing factors pertaining to a song's playlist, popularity, artist, album, and large range of audio features. For the data preprocessing and cleaning step we made sure to identify and remove any missing values and normalize any necessary numerical or categorical variables.

To begin the study, we conducted quantitative exploratory data analysis to gain important insights into the popular tracks, specifically with the audio feature elements such as danceability, energy, loudness, mode, tempo, and duration_ms. To better understand user behavior on Spotify, we created various visualizations such as scatter plots, bar charts, and line plots using Python's Matplotlib library. Summary statistics (mean, standard deviation, minimum, maximum) were then utilized for specific music trend analysis. These graphs were able to show us the commonalities between popular songs and provided an important overview of how the music landscape has evolved over the past decade while avoiding biases. Biases are often present when discussing genre and music tastes, but statistical analysis did not show biases. For user behavior analysis, we examined features like frequently played genres, artist diversity, and temporal listening patterns through variables like artist_genres, artist_popularity, 'year', and aggregating based on playlist_url. We examined the correlation between preferences for the six most popular genres on Spotify: Pop, Rap, Hip-Hop, Rock, Trap, and R&B. Python's Pandas library was used for these methods.

To understand how these trends can help shape promotional campaigns and music releases, we used predictive modeling with linear regression. Using Python's Sci-kitLearn library, we trained a LinearRegression model with artist_popularity as our independent variable and track_popularity as the dependent variable. To visualize the results, we created a scatter plot of the actual versus the predicted values. This model helped us examine the relationship between an artist's popularity and the popularity of a song.

Discussion and Results

An element that is reinforced in user trends on Spotify is how personalized the platform is. In our analysis, we were able to visualize the most popular artists and genres amongst users by changing conditions, such as geography, age, sex, etc. After performing our analysis, we see how users have the ability to create playlists of their liking, shuffle playlists that play similar songs as a song of their preference, play a radio that contains songs and artists within the same genre, etc. These aspects of the application allow for an abundance of personalization, which therefore heavily influence the behaviors and trends of users. These features were apparent in the results of our analysis.

First, we visualized the relationship between danceability and track popularity, by creating a scatterplot to see if the two had a correlation, to better understand why specific music trends on Spotify are the way they are. From the plot in Figure 1.1, we discovered that there is a weak positive correlation between danceability and track popularity, where in the plot, we see an increase in the data points. Specifically at the center of the plot, we see a cluster of data points between when danceability is 65 and 85, where the data points cluster and increment, showing a small correlation.

Beyond this, we aim to answer how these factors can help shape promotional campaigns and music releases. To dive into this analysis, we graphed the average duration of popular songs over the years to better understand how the average lengths of songs change, as seen in Figure 1.2. Throughout most of the years starting from 2000, the average duration changed slightly, until around 2006-2007 where the average duration jumped up drastically and fell back down to the following year. This pattern in the data appeared again around 2014, when the duration of songs

drastically increased and plummeted again the following year. After 2016, the average duration of songs continued to decrease until the current time period. Looking at this graph, we are not able to conclude that over the years, the average duration of popular songs have decreased drastically, despite a few sporadic increases.

Summary statistics were created to look into the descriptive analytics of how popular an artist is, how popular a track is, how danceable a song is, and the energy pertaining to a song. This analysis guided us towards answering what aspects contribute to trends within a user's music and listening history, which we strive to answer for more clarity on user behavior. It can be seen from the numbers in Table 1 that the popularity of the artist is similar to that of the track popularity in relation to how the data looks. This is because the IQR ranges are similar. In addition, as seen from the results, the mean for how danceable a track is and the energy associated with it is relatively in the middle. This means that there is no significant result. Thus, more information is needed to better understand how danceable a song is and how much energy it has.

This is also seen in the linear regression model in Figure 2. This is because it was used to understand the relationship between artist popularity and track popularity, which gives us insight on how a user most frequently discovers their most listened to genre. From this scatterplot, it can be seen that there is a strong relationship between the popularity of an artist and how famous a track is. Since there seems to be a strong correlation between the popularity of an artist in regards to how famous a track is. Since, the relationship seems strong, more research is needed to be done to understand how and why a track is popular. Moreover, what is needed in order to ensure that the track popularity is high when the artist is famous.

In answering whether interest in one genre leads to interest in another genre, we created a correlation heatmap which shows how strong the relationships are between the top six most

popular genres from the dataset: Pop, Rap, Hip-hop, Rock, Trap, and R&B. As seen in figure 3, the warmer, red colors on the heatmap indicate when two genres have a correlation closer to 1 which shows a stronger relationship. The colder, blue colors indicate less or negative correlation closer to -1, showing a weaker relationship. The correlation between rap and hip-hop is displayed as 0.67 which shows a clearer, moderate positive correlation with the hip-hop and trap genres. This implies that users who show interest in broader popular genres like hip-hop or pop may possibly be more interested in closely related subgenres such as rap or trap.

Limitations

It is evident that there are data limitations in relation to temporal scope, bias pertaining to geography, playlist-centric view, little information on a user profile, sample size, and evolving features of Spotify. Based off of these limitations, it becomes necessary to ask what data limitations are present when analyzing the Spotify dataset? How will these limitations influence the generalizability of research findings in relation to the broader user interface that Spotify has? By looking at the outcomes of the research, it is necessary to see that there are certain drawbacks that can help make future investigations fruitful. These limitations may not diminish the significance of what we found, but are still important to discuss. To begin with, even though the Spotify Top Hit Playlist dataset has many variables and a lot of data, it does not encapsulate the full range of user preferences and how they act in the app. Thus, it calls for future studies to investigate more and include more diverse datasets to have a wider variety of genres, regional preferences, and user demographics. Moreover, this analysis we have done only looks at data from 2010 to 2022. Thus, there is a temporal limitation in the data. It is not as current or relevant as it could be. In addition, although the dataset is good for capturing historical trends, having more nuanced examination of specific periods of time might pave way for better insights. This is

because this will allow for seeing change over time on the Spotify platform in regards to the influence of user trends. Another limitation with the research is that there may be potential biases and errors, this means that there needs to be more preprocessing techniques to enhance the accuracy. Our reliance on Python libraries and statistical methods can be reinforced by looking at machine learning models. This will allow us to capture how complex the data is in regards to user preferences.Moreover, the sample size is a limitation of our dataset. This is because if the sample size was larger, it would be closer to the population size. This would mean that it is more likely that the data and analysis would be more representative of the majority of individuals in the population. A larger sample size means that there is a high likelihood for better and more accurate results. All in all, to make our research better including a diversifying dataset, performing longitudinal analysis, incorporating analytical techniques, and mixing contextual variables will allow for a better understanding of music trends and behaviors on Spotify.

Some potential avenues for future work that could be conducted is relating to optimization, user segmentation and predictive analytics. Continuously improving and using real-time data will be so impactful towards Spotify because of the instant recommendations it could offer. This dataset has data from lots of detailed music groups but predicting using machine learning and artificial intelligence could take it to the next step. Tailoring promotional campaigns, advertisements, and techniques to identify the user groups could provide user segmentation and analysis. Music and applications also have an emotional element to it, so developing optimization algorithms for promotional content enhance the user's experience while making sure the ethics of data privacy and security are implemented. An application needs to make sure their users are being protected but how can they tailor it to the users privacy preferences and ensure trust; ethos, pathos and logos.

Conclusion

Society is becoming more data-driven than ever before. The more data that is presented and collected, the more we are able to examine user behaviors and trends to improve industries. The research that we focused on in regards to Spotify hit songs achieves a more solid understanding of user behavior, music trends and media-based algorithms. A more knowledgeable foundation of these specific aspects can help industry professionals make informative choices on new advancements onto their music. The research provides insights into the platform of Spotify through considerations for data to be improved via machine learning and data analytics to enhance current abilities. Optimization and personalization has always been the prime topic to cater to the user experience of Spotify, and with predictive analytics and real-time data, Spotify could reach higher to its potential.

References

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic Effects on the Diversity of Consumption on Spotify. International World Wide Web Conference, 2020 Issue, https://dl.acm.org/doi/abs/10.1145/3366423.3380281?casa_token=7gg97vV3rVAAAAAA:C006hv3LGoan72_di0IL915sAVhn5AZbL9_ro3JDdGKQG0niivI_bf_Awuh6yT6nJYI2FgDs22yIzfw

Josephine, J. (2022). Spotify Top Hit Playlist 2010-2022. Kaggle, https://www.kaggle.com/datasets/josephinelsy/spotify-top-hit-playlist-2010-2022/data

Terroso-Saenz, F., Soto, J., & Muñoz, A. (2023, January). Evolution of global music trends: An exploratory and predictive approach based on Spotify data. Entertainment Computing, Volume 44, https://www.sciencedirect.com/science/article/pii/S1875952122000593?casa_token=V59BMKBtVsgAAAAA:RKcmiTyimns2izWwZn7pV_5AT6sUISsnm4a74h9dULzyUAuhfarVVGqDdCLMRAQENwpZRqCmTIQ

Zhang, B., Kreitz, G., Isaksson, M., Ubillos, J., Urdaneta, G., Pouwelse, J. A., & Epema, D. (2013). Understanding user behavior in Spotify. *2013 Proceedings IEEE INFOCOM*, April 2013 Issue, https://ieeexplore.ieee.org/abstract/document/6566767

Tables and Figures

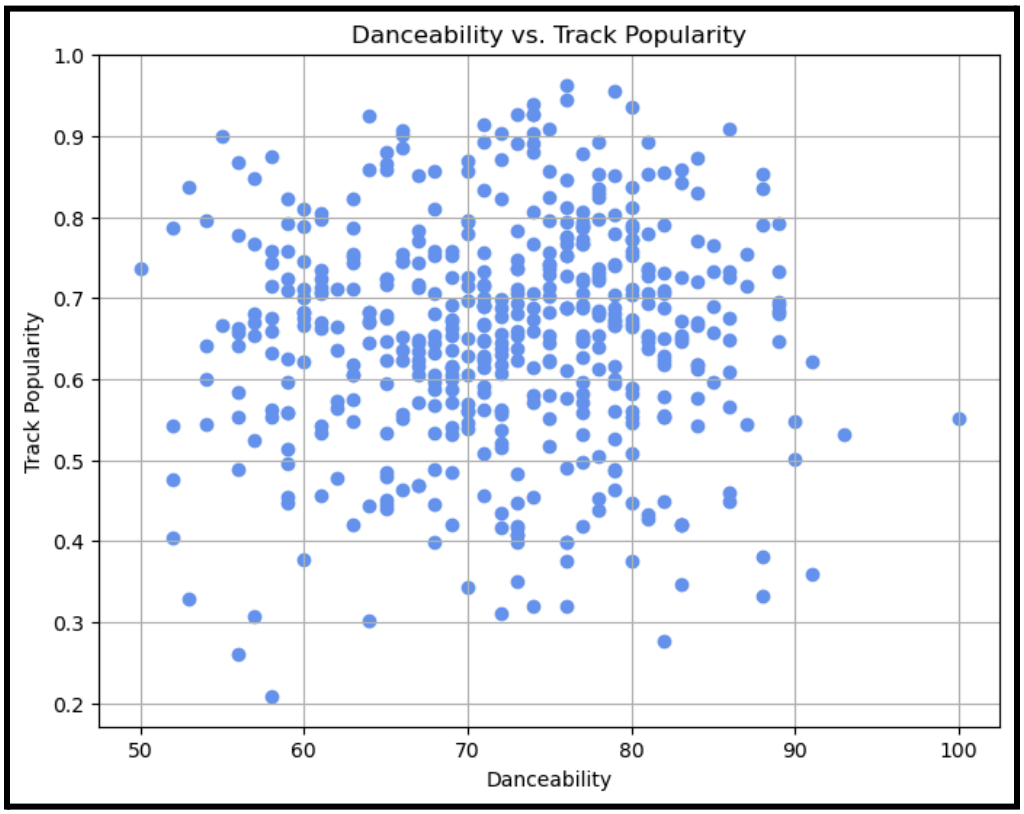

Figure 1.1 Exploratory Quantitative Analysis – Danceability vs. Popularity of Songs

	Track Popularity	Artist Popularity	Danceability	Energy
count	500.0	500.0	500.0	500.0
mean	72.2	72.59	0.66	0.7
std	8.87	12.51	0.14	0.16
min	50.0	35.0	0.21	0.06
25%	66.0	65.0	0.57	0.59
50%	73.0	74.0	0.67	0.72
75%	79.0	82.0	0.74	0.82
max	100.0	100.0	0.96	0.98

Table 1 Summary Statistics

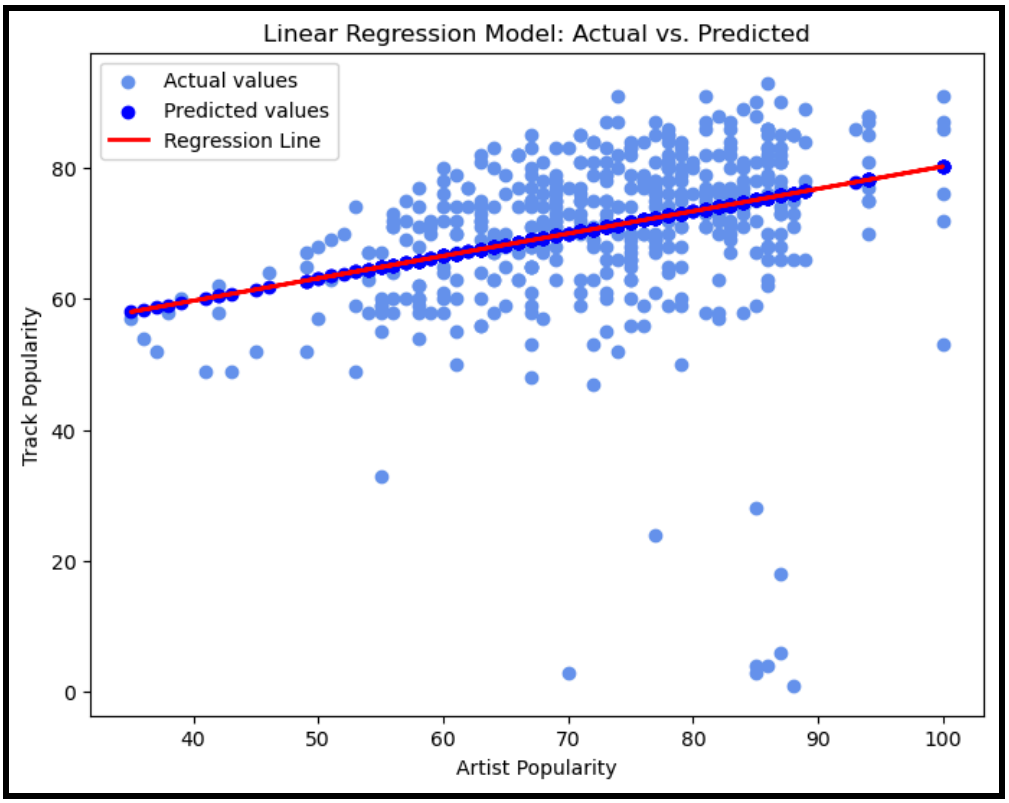

Figure 2 Linear Regression Analysis – Artist Popularity vs. Track Popularity

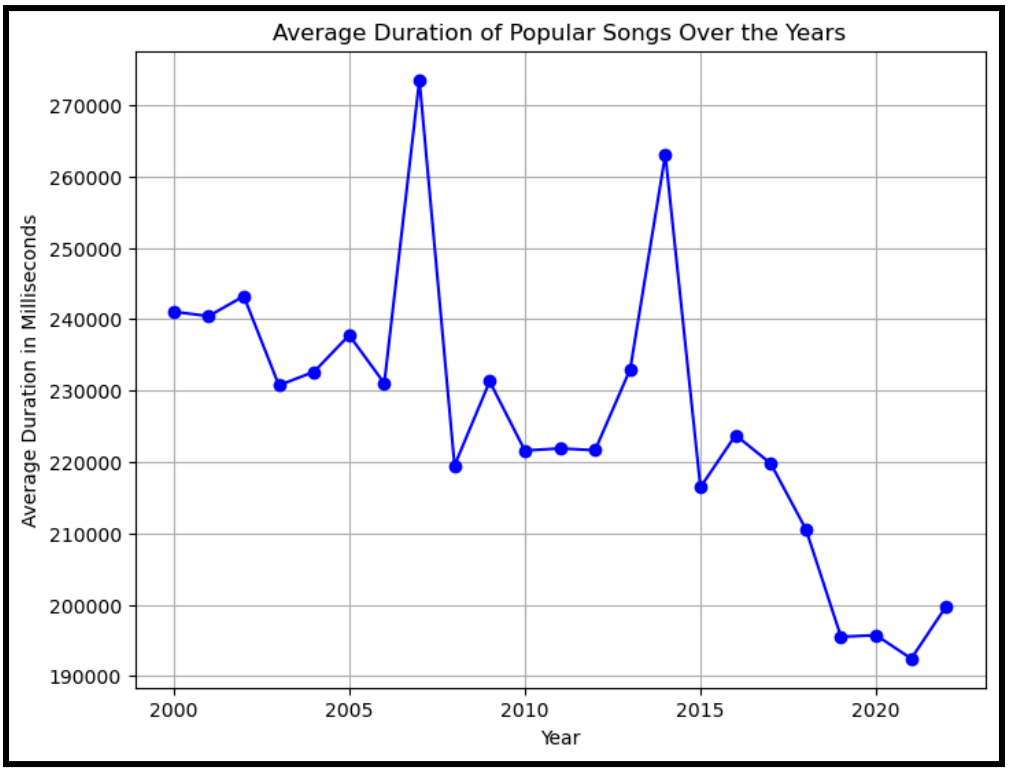

Figure 1.2 Exploratory Quantitative Analysis – Average Duration of Songs

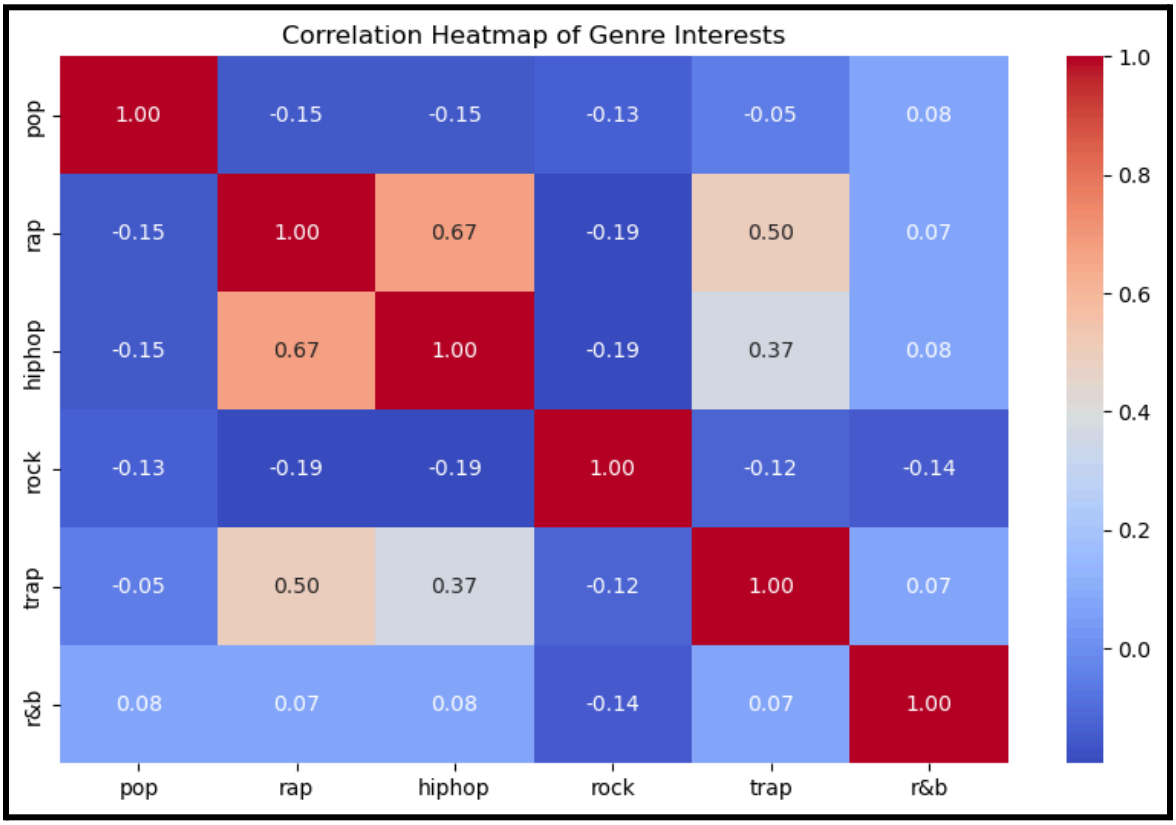

Figure 3 – Correlation between Popular Genres