Data Delinquents: Analyzing Spotify's 2023 Top Hits

Omar Al Achcha, Siena Theivagt, Trang Nguyen

School of Information; The University of Texas at Austin

I 310D: Introduction of Human-Centered Data Science

Dr. Abhijit Mishra

December 6, 2023

Data Delinquents: Analyzing Spotify's 2023 Top Hits

"Spotify is a pioneer in music streaming" (Pendlebury, 2023). To understand Spotify's significance in the streaming industry, it is pertinent to understand the roots and origins of said industry. With roots in peer-to-peer sharing platforms such as Napster, popularized in 1999, streaming was not an unfamiliar method of consuming music. However, due to the informal nature of these platforms, they would soon prove as unethical as these peer-to-peer sharing platforms sometimes promoted the illegal consumption of music, as the free downloading platform was riddled with copyright and ownership issues. Music streaming would soon take off again, as streaming giants such as Pandora and Spotify were launched in the mid 2000's, gaining traction at an unprecedented pace (Brewster, 2023). This brings us to the present day, where Spotify has maintained its reputation as a music streaming giant, harnessing over 515 million streamers worldwide as of 2023 (Shepherd, 2023). These music streams not only provide revenue for artists, but also provide publicity, exposure, and easier access to growing their fanbases.

Spotify's reputation and significance in the music streaming industry is a primary factor in our purpose and goal for this project; we aim to find correlations between different variables within Spotify's 2023 top hits, such as "danceability" and "energy", and determine if these variables can provide meaningful insights regarding music streams. In order to determine if these variables significantly impact streams, we will perform group comparisons of the means within the variables "danceability" and "energy", and perform a regression analysis for the two variables respectively. We will leverage the insights from our analysis to help artists who desire to be on the top hits list achieve their goal, focusing on three research questions that will shape our analysis along with our hypotheses:

1. Do songs with high danceability receive a significantly higher number of streams compared to songs with low danceability?

2. Are more energetic songs associated with a statistically higher number of streams compared to less energetic songs?
3. Is there a strong regression relationship between the variables danceability and energy? And what are the details of that relationship, if any?

Data Description

We obtained our dataset from Kaggle, which was titled “Most Streamed Songs Spotify 2023”, and published by author Nidula Elgiriyewithana. The dataset contained over 900 unique records of the top hits on Spotify from the year 2023, with descriptive variables such as song title, artist name, and stream count, as well as variables depicting audio features, such as danceability, valence, and beats per minute. For our project, we decided to focus on the relationship between the variables streams, which is used interchangeably with popularity, danceability, and energy. Both danceability and energy present themselves as percentages. According to Spotify’s Web API Referencer, danceability is calculated by “a combination of musical elements including tempo, rhythm stability, and overall regularity”, while energy is calculated by “a perceptual measure of intensity and activity” (Spotify, 2023).

Pre-processing: Cleaning and Sorting

In order to make the dataset more interpretable and fit for our analysis, we removed 15 rows that we thought would be unfit for our hypotheses, which included variables such as key, mode, and beats per minute. We also sorted the data by streams in descending order, as they were in a randomized order originally; this enhanced the readability and interpretability of the dataset. Finally, we removed all the columns that are not related to our research questions.

Methodology

In order to gather meaningful insights based off of our three research questions listed in the introduction, we will conduct statistical analyses to gather statistically significant results. Our first two research questions focus on whether the values of the variables “danceability” and

“energy” affect the number of streams, with both variables being separated into the two categories of “high” and “low” respectively; the comparison of the means of these two categories within each variable will be analyzed. The threshold marker for the separation of high and low is set at 50%, with values above the threshold being classified as “high” and the values below the threshold being classified as “low”. Due to the nature of our first two research questions involving group mean comparisons, we decided that comparing group means would be best suited for determining whether the higher danceability group and higher energy group received more streams compared to their lower counterparts.

In our third research question, we intend to explore the relationship between the two previously discussed variables of “danceability” and “energy”, determining whether there is a strong regression between the two variables. To achieve this, we will perform a regression analysis, gathering insights from the R-squared value that is to be calculated.

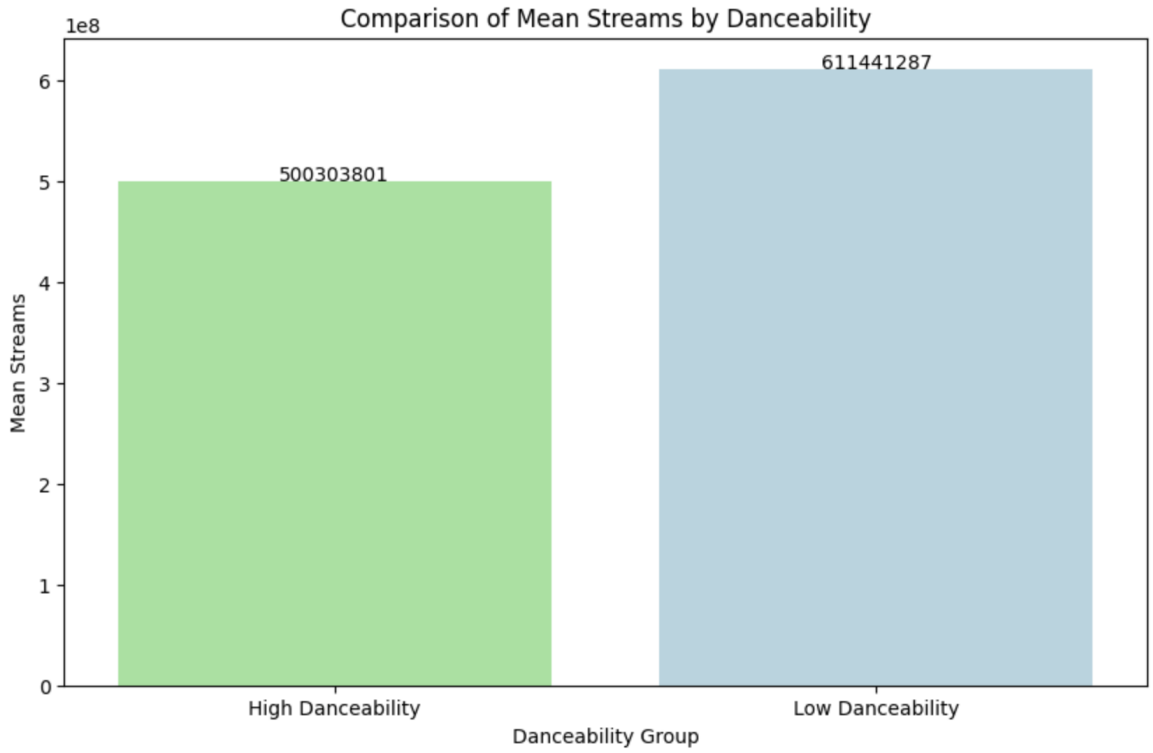
Analysis and Results: First Question

Our analysis will be guided by our three research questions, therefore being separated into three parts. Our first research question is as follows:

1. Do songs with high danceability receive a significantly higher number of streams compared to songs with low danceability?

As noted in our analysis plan, we will be conducting a comparison of group means, comparing the means of the groups “high danceability” and “low danceability” to determine whether a certain group receives a significantly higher amount of streams. Our first null hypothesis is H0 = Mean (High Danceability) group <= Mean (Low Danceability), with the alternative hypothesis being H1 = Mean (High Danceability) group > Mean (Low Danceability). We used the SciPy scientific computing library to run our statistical tests, calculating the means for each group respectively; the mean for the high group was calculated to be 500303801.36 streams, while the mean for the low group was calculated to be higher at 611441286.62 streams. Our results

provided sufficient evidence to fail to reject the null hypothesis. We used the SeaBorn library to create the visualization, a bar chart, that is attached below. It shows the group means of both the high and low categories for each group respectively, with the number of streams being labeled on the y-axis.



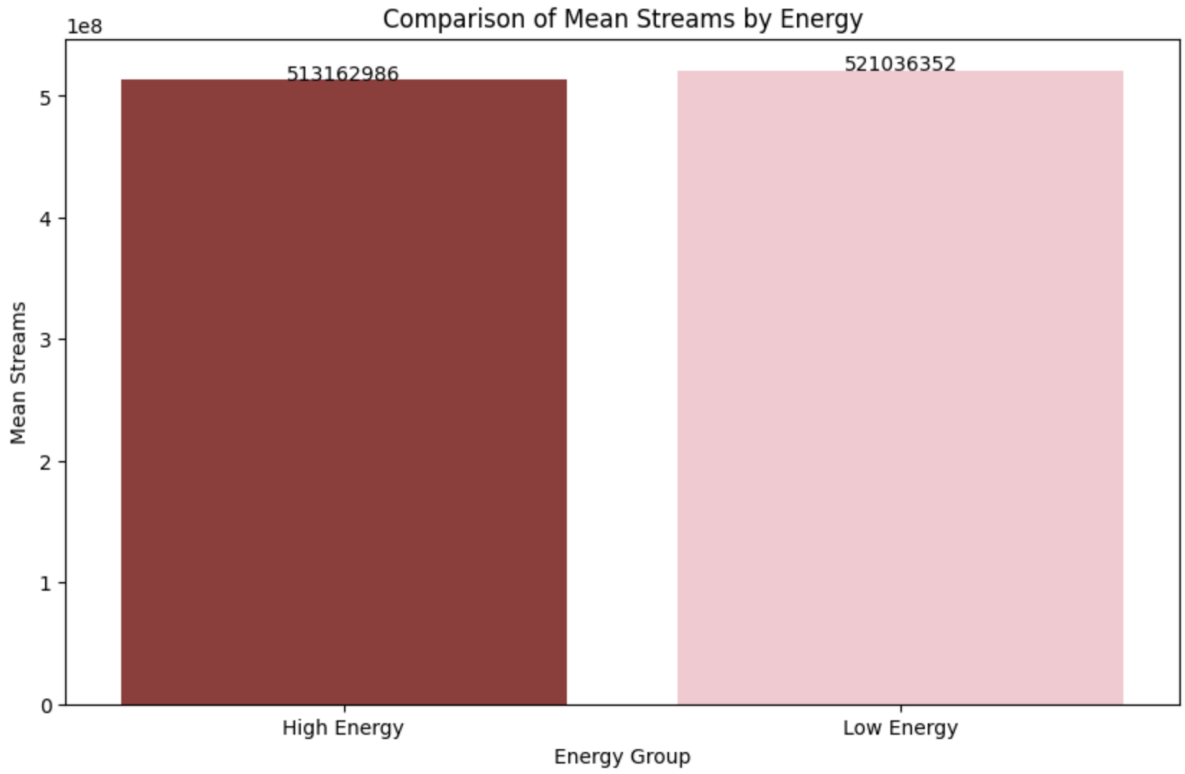
Analysis and Results: Second Question

Our second research question also uses a group comparison of means, with the research question being as follows:

2. Are more energetic songs associated with a statistically higher number of streams compared to less energetic songs?

We will be comparing the means of the groups “high energy” and “low energy” to determine whether a certain group receives a significantly higher number of streams. The null hypothesis is H0 = Mean (High Energy) group <= Mean (Low Energy), with the alternative hypothesis being H1 = Mean (High Energy) group > Mean (Low Energy). The mean for the high group was

calculated to be 513162986.05 streams, with the low group's mean being calculated to be higher 521036352.43 streams. Our results provided sufficient evidence to fail to reject the null hypothesis. Again, we used the SeaBorn library to create the visualization that is attached below. It shows the group means of both the high and low categories for each group respectively, with the number of streams being labeled on the y-axis.



Analysis and Results: Third Question

Our third and final research question is as follows:

3. Is there a strong regression relationship between the variables danceability and energy? And what are the details of that relationship, if any?

Our final statistical test will be seeing if there is a strong regression between the two variables discussed in our first two research questions. Our null hypothesis is H0 = There is no strong regression relationship between danceability and energy, with our alternative hypothesis being that H1 = There is a strong regression relationship between danceability and energy. For this

research question we used the SciKit-learn python library, using the linear regression function. The slope (coefficient) of the regression was calculated to be 0.22, while the R-squared value was calculated to be 0.039. The slope of 0.22 suggests that there is a positive relationship between danceability and energy; more specifically, it implies that for every 1% increase in danceability, the energy also increases by 0.22% on average. The R-squared value of 0.039 suggests a very low relationship or regression between the two variables; however, since it is a nonzero number, it also suggests some level of association albeit low, with the value of 1 being the highest possible correlation. The R-squared value does not give us sufficient evidence to reject the null hypothesis, therefore we fail to reject the null hypothesis. The regression analysis is visualized in the graph attached below, with “energy” being on the y-axis and “danceability” being on the x-axis.

Conclusion

After comparing the high and low category group means for our variables “danceability” and “energy”, we failed to reject both hypotheses, determining that the low categories for both variables had a higher mean of streams than their high counterparts. Based on these results, we canconclude that songs that have lower danceability and lower energy have higher streams, therefore positively influencing their chances of being on the Spotify Top Hits lists, or in this case, the Spotify Top Hits of 2023. We also determined that the relationship between the variables danceability and energy is weak, with the R-squared value calculated at 0.039. However, since the number is a nonzero value, there is suggested to be a very weak regression relationship.

Limitations: Scope and Usage

Our limitations largely deal with the nature of our dataset and its limited scope. Our dataset is limited to the year 2023, therefore our results obtained from analyzing this dataset cannot be reliably applied onto other years. Another large limitation is due to Spotify’s usage by country; Spotify is primarily used in North America and Europe, meaning that our data is heavily influenced by those regions, and our results would therefore not be reliably applicable to other countries (Iqbal, 2023). This is especially true for regions and countries who almost exclusively use other platforms outside of Spotify, such as South Korea with the Melon streaming service (Dredge, 2023).

Potential Applications and Future Work

In the future, additional variables should be analyzed in relation to streams in order to gain more meaningful insights on increasing streams. For our study, we analyzed the variables danceability and energy; in order to build and broaden the potential applications of this study, we believe that adding variables such as valence and liveness may be beneficial. We believe that as our study grows and more complexities are added, it will also be possible to create a predictive model that learns how to predict which songs will be a hit and make it to the top lists.

Another potential application is user-centered; the ability for users to curate playlists based on user preferences in musicality. This may look like Spotify improving their recommendation algorithm and accounting for similarities in variables such as danceability and energy to overall enhance user experience. Our final potential application ties into our project goal of helping artists achieve their goals of making it into Spotify's top lists; we believe that our results can provide artists with guidance in what songs garner more streams, therefore allowing artists to refine their releases accordingly.

References

Brewster, W. (2023, August 21). *The History of Music Streaming*. Mixdown Magazine. https://mixdownmag.com.au/features/the-history-of-music-streaming/

Dredge, S. (2023, May 11). *Melon still the most popular South Korean streaming service*. Music Ally. https://musically.com/2022/08/10/melon-still-the-most-popular-south-korean-streaming-service/

Iqbal, M. (2023, August 2). *Spotify revenue and Usage Statistics (2023)*. Business of Apps. https://www.businessofapps.com/data/spotify-statistics/

Pendlebury, T. (2023, October 24). *Best Music Streaming Service of 2023*. CNET. https://www.cnet.com/tech/services-and-software/best-music-streaming-service/

Shepherd, J. (2023, June 26). *23 essential spotify statistics you need to know in 2023*. The Social Shepherd. https://thesocialshepherd.com/blog/spotify-statistics

Spotify. *Get track's audio features*. Web API Reference | Spotify for Developers. https://developer.spotify.com/documentation/web-api/reference/get-audio-features#

Appendix

Dr. Mishra asked whether we were conducting a t-test and what kind of t-test was being used. We initially thought of using an independent sample t-test, but later determined that a group comparison of means would be better fit for our project goals and expectations.

Dr. Mishra's follow up question was concerning how our insights could be potentially used to improve or enhance user experience while using Spotify. As discussed previously in the conclusion, we believe that our findings and the potential applications of those findings can be used to help users curate playlists; this may look like Spotify improving their recommendation algorithm in order to account for the discussed variables. On a similar note, Spotify can also provide a page where users can potentially explore songs that fall into their preferred musicality variables, such as songs with high danceability and energy.

-