

Data Literacy Lab Final Project

By: Charles McCain , Arman Hassan, Delphine Liu, Jennifer Lee, Samuel Leggett, Ubence Lazo,
William Lorentz

Introduction:

The data set we've selected is the "Most Streamed Spotify Songs 2023." This dataset provides a complete overview of popular songs on Spotify, a leading music streaming service, during the year 2023. It includes key attributes such as the song's name (`track_name`), artist(s) name (`artist(s)name`), the number of artists involved (`artist_count`), release date (`released_year`, `released_month`, `released_day`), presence and rank on Spotify charts (`in_spotify_charts`), the total number of streams on Spotify (`streams`), and several musical features such as beats per minute (`bpm`), key, mode, danceability percentage (`danceability%`), valence percentage (`valence_%`), energy percentage (`energy_%`), acousticness percentage (`acousticness_%`), instrumentalness percentage (`instrumentalness_%`), liveness percentage (`liveness_%`), and speechiness percentage (`speechiness_%`). This dataset also presents the song's presence and rank on other platforms, including Apple Music, Deezer, and Shazam playlists and charts.

Research Question: To what extent, if any, does the percent of danceability of a song correlate with its number of streams for solo artists?

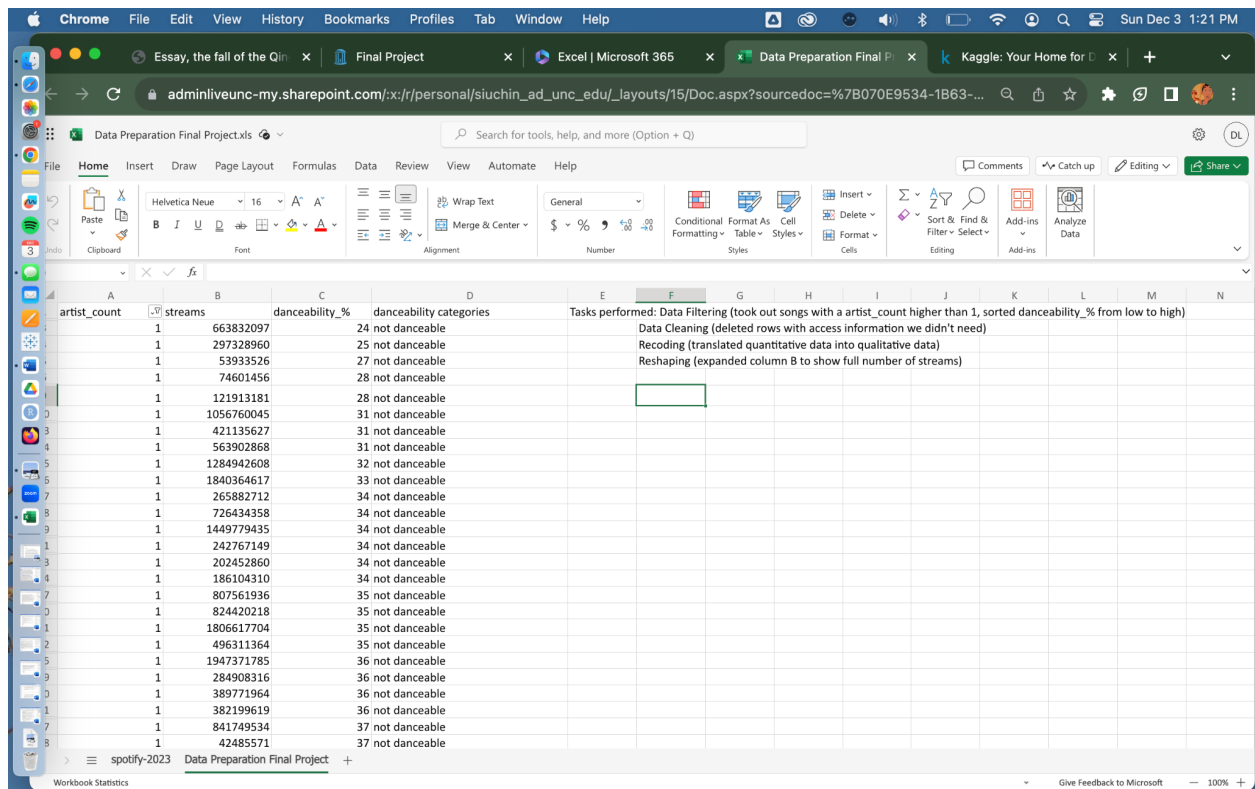
The first ethical consideration that comes to mind is the privacy of musical artists and their work. This dataset contains information about artists' songs that must be handled responsibly in order to maintain respect for both artists' individual and intellectual property rights. Another ethical issue that could arise is the analysis of an artists' popularity and success as it relates to their identity. It is important to be sensitive, avoiding any stereotypes or unfair judgments based on characteristics such as gender, ethnicity, or other personal attributes. Additionally, though our analysis is numerical, the number of streams a song may have does not represent the quality of a song and is not completely representative of an artist's talent or ability. Thus, an artist and their work should not be viewed solely on the basis of the number of streams they may have. Another potential consideration is, given the dataset includes information about artists who may not have consented to being part of the analysis nor this study, the ethical implications of using their data must be considered. Finally, it's crucial to acknowledge the ethics in assessing danceability as it is a fairly subjective concept, underscoring the importance that its analysis may not be completely representative of empirical data.

Methods:

We acquired our dataset from Kaggle. As a group, we were interested in finding data about the most streamed songs on Spotify. We searched on Kaggle and quickly found this dataset.

We worked with the “danceability_%”, “streams”, and “artist_count” variables. These were all numerical variables, giving us quantitative data that we translated into qualitative data. “Danceability_%” was measured in a percentage indicating how suitable the song is for dancing. “Streams” were measured by “total number of streams on Spotify, and “Artist_count” was defined as the “number of artists contributing to the song”, according to the Kaggle website.

Data Preparation:



artist_count	streams	danceability_%	danceability categories	Tasks performed:
1	663832097	24	not danceable	Data Filtering (took out songs with a artist_count higher than 1, sorted danceability_% from low to high)
1	297328960	25	not danceable	Data Cleaning (deleted rows with access information we didn't need)
1	53933526	27	not danceable	Recoding (translated quantitative data into qualitative data)
1	74601456	28	not danceable	Reshaping (expanded column B to show full number of streams)
1	121913181	28	not danceable	
1	1056760045	31	not danceable	
1	421135627	31	not danceable	
1	563902868	31	not danceable	
1	1284942608	32	not danceable	
1	1840364617	33	not danceable	
1	265882712	34	not danceable	
1	726434358	34	not danceable	
1	1449779435	34	not danceable	
1	242767149	34	not danceable	
1	202452860	34	not danceable	
1	186104310	34	not danceable	
1	807561936	35	not danceable	
1	824420218	35	not danceable	
1	1806617704	35	not danceable	
1	496311364	35	not danceable	
1	1947371785	36	not danceable	
1	284908316	36	not danceable	
1	389771964	36	not danceable	
1	382199619	36	not danceable	
1	841749534	37	not danceable	
1	42485571	37	not danceable	

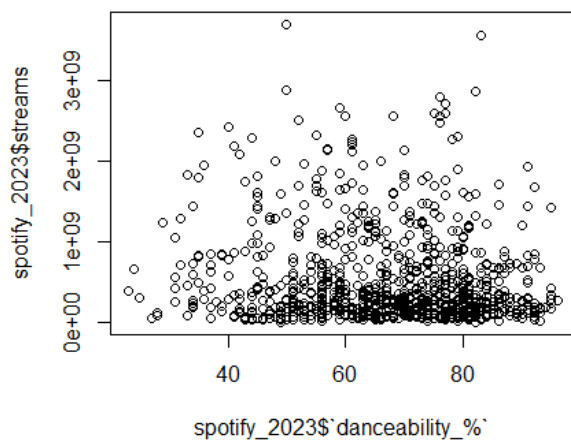
Data Preparation Final Project.xls.xlsx

Above, we have attached a screenshot of our data preparation excel sheet. In order to access the entire sheet, you must click on the attached link. This sheet uses the “danceability_%”, “streams”, and “artist_count” numerical variables. We thus used data cleaning to delete rows of data we did not need, reshaping by expanding column B to show the full number of streams, and recoding to translate our quantitative “danceability_%” data into qualitative data. We translated the quantitative “danceability_%” data into three categories: not danceable, danceable, and very

According to the statistics of average streams of each category of “danceability_ %” being “not danceable”, “danceable”, and “very danceable”, it appears that amount of streams and danceability do not have a strong correlation. In fact, based on the calculations it shows that the “not danceable” category of “danceability_ %” has the highest average streams. Following the “not danceable” category on highest average streams is the “danceable” category with 2nd highest streams out of the 3 categories. Then the category with the least amount of streams is the “very danceable” category. This data shows that having high streams does not correlate to having a high “danceability_ %.”

In our data visualization, we used R for statistical analysis, focusing on the relationship between “danceability_ %” and the number of streams. We calculated a correlation coefficient of -0.01054569, indicating a low correlation. However the p-value of 0.001119 implies a high statistical significance (using a significance level of 0.05), suggesting that the observed weak correlation is not just a chance occurrence.

Scatter Plot Data Visualization of 2023 Spotify Tracks’ percent of Danceability and Total Streams



Visualization Description: We created a scatter plot data visualization using R software by creating a spotify_2023 data frame and running a code that would correlate the “danceability_ %” column and “streams” column: `plot(spotify_2023$danceability_ %`, spotify_2023$streams)`. The scatter plot displays songs from spotify in 2023, with their danceability percentage on the horizontal line and the number of times they’ve been streamed on the vertical line. Looking at the spread of dots, we can see that songs of all types of danceability (low, medium and high) have a wide range of stream counts. There’s no obvious trend where more danceable songs have

higher streams, which suggests that a song's danceability doesn't necessarily mean it will be streamed more often.

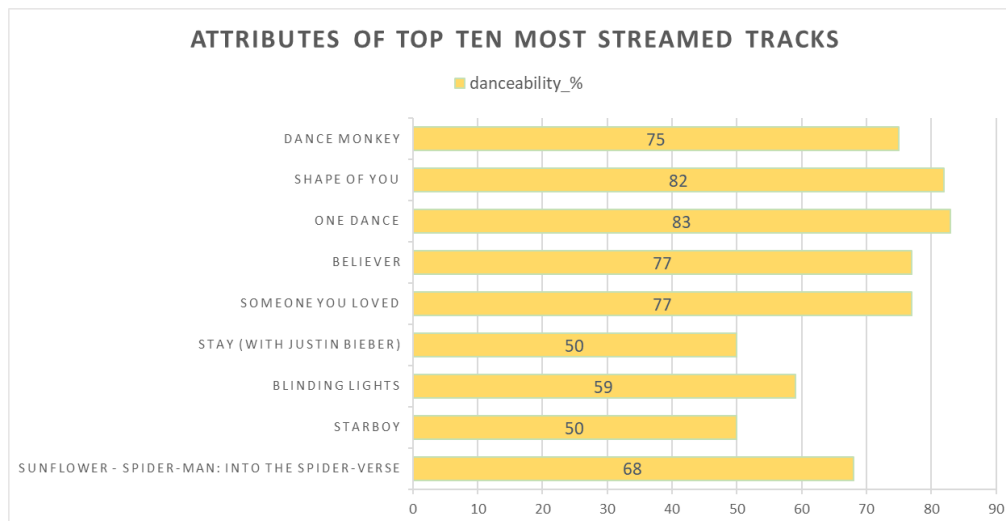
Calculating a Correlation Coefficient Inferential Statistics between danceability and streams using RStudio

```
> cor.test(spotify_2023$danceability_%,spotify_2023$streams, method= "pearson")  
  
Pearson's product-moment correlation  
  
data: spotify_2023$danceability_% and spotify_2023$streams  
t = -3.2686, df = 950, p-value = 0.001119  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.16786950 -0.04220224  
sample estimates:  
cor  
-0.1054569  
  
>
```

The correlation coefficient is -0.1054569, which tells us that there is a low correlation between danceability and streams

The p-value is .001119, which is low, close to 0, and tells us that there is a high statistical significance between danceability and streams.

Stacked Bar Chart of the Percent of Dancibility of the Top Ten Most Streamed Tracks on Spotify 2023



Visualization Description: We created a 2D bar chart data visualization using Microsoft Excel. We selected the data by filtering the streams column by “Top 10,” then we created a 2D bar chart based on the “track_name” column and “danceability_ %” column. This bar chart shows the danceability percentages of the top 10 most streamed songs on Spotify. Each bar represents a different song, and the length of the bar indicates how danceable each song is, with a higher percentage meaning more danceable. From the chart, we can see a mix of danceability scores among the most popular songs. Some have high danceability, while others are lower, which suggests that a song can be very popular on Spotify regardless of how danceable it is. This addresses our research question by showing that high streams don't always match with high danceability.

Conclusion:

In conclusion, our research question aimed to explore the relationship between the percent of danceability in a song and its number of Spotify streams for solo artists in 2023. Through our analysis, we observed a lack of correlation between these two factors, as represented by the relative random distribution in our scatter plot. Our correlation coefficient of -0.01054569 further affirmed the absence of a strong correlation between our two variables.

On the other hand, our data also revealed an intriguing statistic of a p-value of 0.001119 , signaling high statistical significance of our data on the relationship between danceability and the number of streams. The correlation coefficient is -0.01054569 , showing a weak correlation. In addition, our p-value being low indicates this weak correlation is unlikely to be due to chance.

Avenues for further research could be diving deeper into the nuances of danceability and researching potential confounding variables that influence the number of streams a song has. For example, one could explore whether there are specific genres or cultural factors that impact the danceability of a song and how this may be represented in its stream count. Additionally, our study focused on the year 2023; researching different time periods and ages in musical history could reveal more on what variables impact the number of streams a song has and how this may correlate with its danceability.

Finally, an avenue of further research could be investigating the impact of external factors like social media trends, marketing, or cultural events on the danceability and stream metrics for a song. This data and understanding could lead to a more comprehensive understanding of what propels audience engagement. As music streaming and technology continues to evolve in future years, this data can provide a foundation and basis for ongoing research that can help us further understand the intricate connections and correlations between musical attributes and stream counts.

Links:

<https://www.kaggle.com/datasets/nelgiriwithana/top-spotify-songs-2023/>

[Data Preparation Final Project.xls.xlsx](#)