Spotify 2023
Exploratory Analysis

Name, Samia Mahmood (smahmood@bellarmine.edu)
Name, Joseph Guthrie (jguthrie@bellarmine.edu)

I. INTRODUCTION

This dataset is titled Spotify 2023. We found this dataset from Kaggle. The reason why we decided to pick this dataset was because it had a lot of upvotes which helped ensure that this was a trusted dataset for analytical purposes. Additionally, the dataset had a variety of variables which could be used and manipulated for data analysis. Given this variety within the variables, it provided more opportunities for data analysis as the data could be analyzed in a variety of ways.

II. DATA SET DESCRIPTION

There are 953 entries (rows) in the dataset, with 24 columns. The columns position, name, their datatypes, number of, and percentage of missing values are as follows in **Table 1.**

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
track_name	*Nominal, object*	*0%*
artist(s)_name	*Nominal, object*	*0%*
artist_count	*Interval, int64*	*0%*
released_year	*Interval, int64*	*0%*
released_month	*Interval, int64*	*0%*
released_day	*Interval, int64*	*0%*
in_spotify_playlists	*Ordinal, int64*	*0%*
in_spotify_charts	*Interval, int64*	*0%*
streams	*Ordinal, object*	*0%*
in_apple_playlists	*Ordinal, int64*	*0%*
in_apple_charts	*Interval, int64*	*0%*
in_deezer_playlists	*Ordinal, object*	*0%*
in_deezer_charts	*Interval, int64*	*0%*
in_shazam_charts	*Ordinal, object*	*5.25%*
bpm	*Ordinal, int64*	*0%*
key	*Nominal, object*	*10%*
mode	*Nominal, object*	*0%*
danceability_%	*Ratio, int64*	*0%*
valence_%	*Ratio, int64*	*0%*
energy_%	*Ratio, int64*	*0%*
acousticness_%	*Ratio, int64*	*0%*
instrumentalness_%	*Ratio, int64*	*0%*
liveness_%	*Ratio, int64*	*0%*
speechiness_%	*Ratio, int64*	*0%*

III. Data Set Summary Statistics

Since there are several columns in our data that do not contain data useful (or possible) to graph, these will be excluded, giving us 13 variables to examine the statistics for. The various statistics can be referenced in **Table 2**.

Table 2: Summary Statistics for Spotify 2023

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
in_spotify_playlists	*953*	*5200.125*	*7897.6*	*31*	*875*	*2224*	*5542*	*52898*
in_spotify_charts	*953*	*12.00*	*19.58*	*0*	*0*	*3*	*16*	*147*
in_apple_playlists	*953*	*67.81*	*86.44*	*0*	*13*	*34*	*88*	*672*
in_apple_charts	*953*	*51.90*	*50.63*	*0*	*7*	*38*	*87*	*275*
in_deezer_charts	*953*	*2.70*	*6.04*	*0*	*0*	*0*	*2*	*58*
bpm	*953*	*122.54*	*28.06*	*65*	*100*	*121*	*140*	*206*
danceability_%	*953*	*66.97*	*14.63*	*23*	*57*	*69*	*78*	*96*
valence_%	*953*	*51.43*	*23.48*	*4*	*32*	*51*	*70*	*97*
energy_%	*953*	*64.28*	*16.55*	*9*	*53*	*66*	*77*	*97*
acousticness_%	*953*	*27.06*	*25.99*	*0*	*6*	*18*	*43*	*97*
instrumentalness_%	*953*	*1.58*	*8.40*	*0*	*0*	*0*	*0*	*91*
liveness_%	*953*	*18.21*	*13.71*	*3*	*10*	*12*	*24*	*97*
speechiness_%	*953*	*10.13*	*9.91*	*2*	*4*	*6*	*11*	*64*

The following tables **Table 3** and **Table 4**, show the categorical variables within the dataset.

Table 3: Proportions for Mode Type

Mode	*Frequency*	*Proportion (%)*
Major	*550*	*57.71%*
Minor	*403*	*42.28%*

Table 4: Proportions for Type of Key

Key	*Frequency*	*Proportion (%)*
C#	*215*	*22.5%*
G	*96*	*10.07%*
G#	*91*	*9.5%*
F	*89*	*9.33%*
B	*81*	*8.49%*
D	*81*	*8.49%*
A	*75*	*7.86%*
F#	*73*	*7.66%*
E	*62*	*6.50%*
A#	*57*	*5.98%*
D#	*33*	*3.46%*

The following figure is a correlation matrix for all the continuous variables within the dataset.

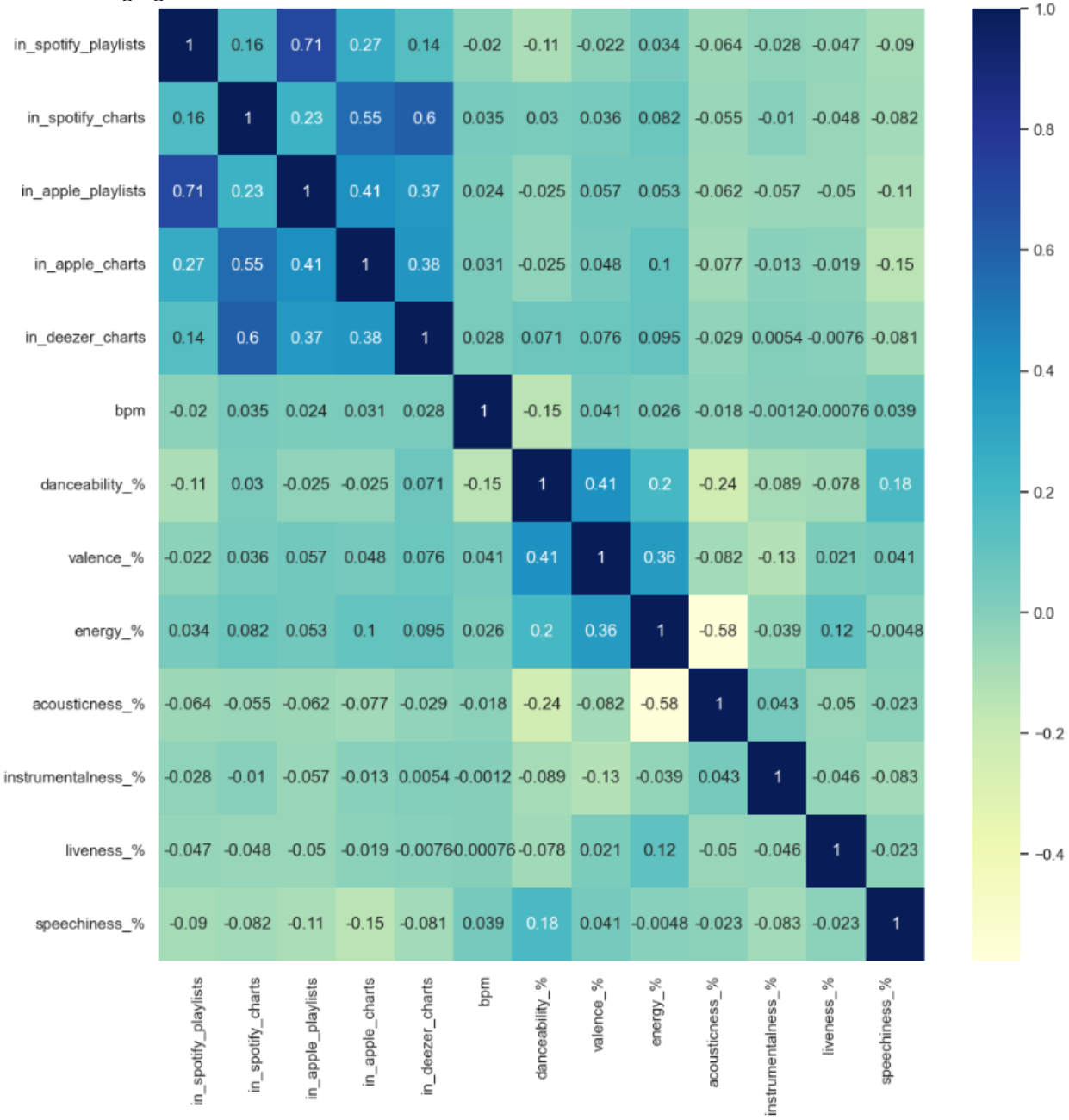

Figure 1: Correlation matrix pf continuous variables excluding categorical variables.

Table 5: Correlation Table/Tables

The following table displays the songs within the dataset that had the highest percentage of energy.

Track Name	***Artist Name***	***Energy %***
I'm Good (Blue)	*Bebe Rexha, David Guetta*	*97*
Murder In My Mind	*Kordhell*	*97*
That That (prod. & feat. SUGA of BTS)	*PSY, Suga*	*96*
Tá OK	*dennis, MC Kevin o Chris*	*96*

Bombonzinho - Ao Vivo	*Israel & Rodolffo, Ana Castela*	*95*
Merry Christmas	*Ed Sheeran, Elton John*	*94*
Every Angel is Terrifying	*The Weeknd*	*94*
Idol (「アイドル」)	*YOASOBI*	*94*
KICK BACK	*Kenshi Yonezu*	*94*
Freaks	*Surf Curse*	*94*

Table 6: Correlation Table/Tables

The following table displays the songs within the dataset that had the lowest percentage of energy.

Track Name	***Artist Name***	***Energy %***
Happier Than Ever	*Billie Eilish*	*24*
It's Beginning To Look A Lot Like Christmas	*Michael Buble*	*23*
Boyfriends	*Harry Styles*	*20*
Special	*SZA*	*20*
I'm Tired - From "Euphoria" An Original HBO Se...	*Labrinth*	*20*
Something In The Way - Remastered 2021	*Nirvana*	*20*
Sweet Nothing	*Taylor Swift*	*16*
The Christmas Song (Merry Christmas To You)	*Nat King Cole*	*15*
Heart To Heart	*Mac DeMarco*	*14*
What Was I Made For? [From The Motion Picture Barbie]	*Billie Eilish*	*9*

When looking at which tracks have the most energy and which tracks have the least energy, the song "I'm Good Blue" by Bebe Rexha and David Guetta had the highest percentage of energy while the song "What Was I Made For? from the Motion Picture Barbie" had by Billie Eilish the least percentage of energy.

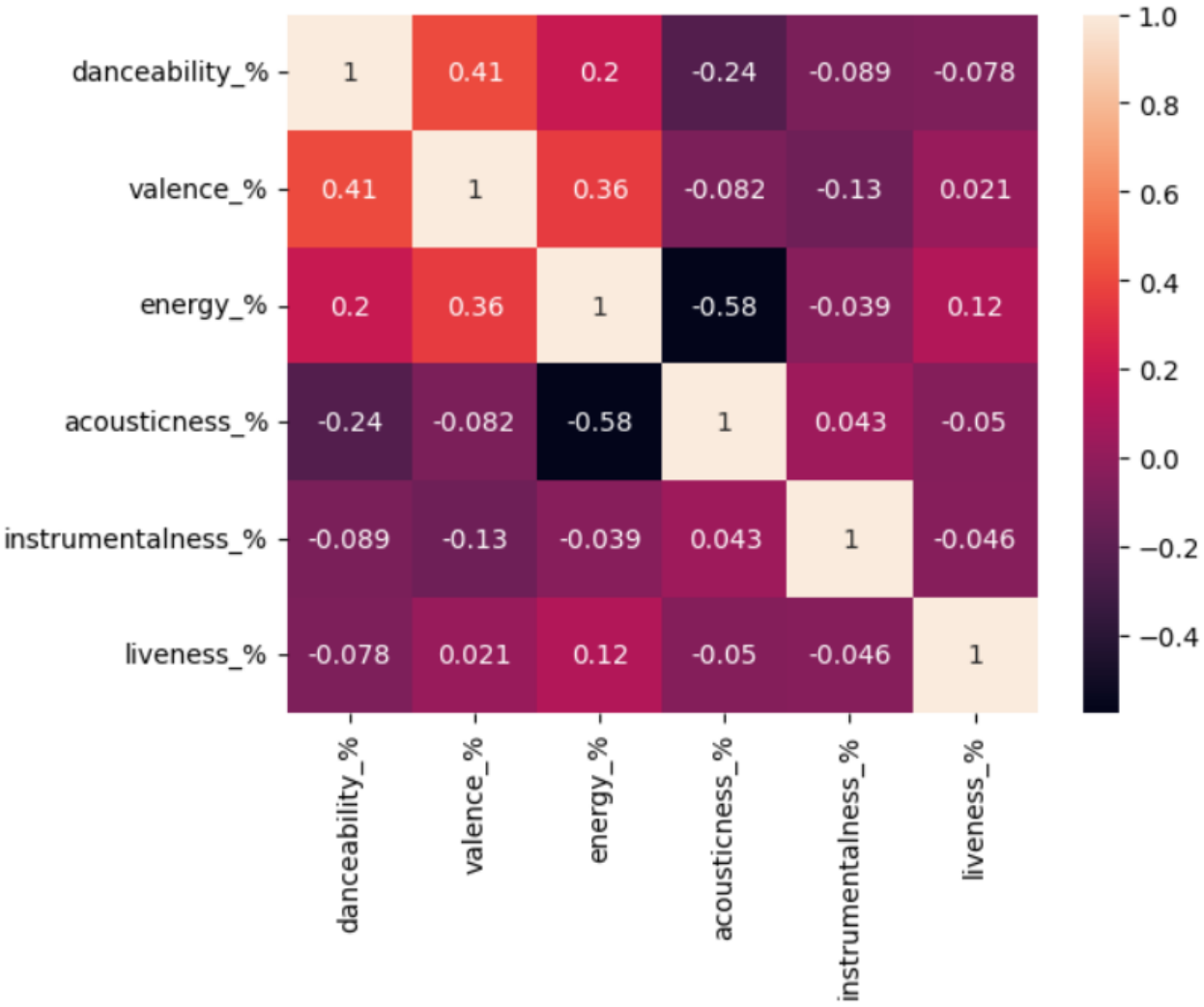

Figure 2. Correlation between danceability, valence, energy, acousticness, instrumentalness, and liveness

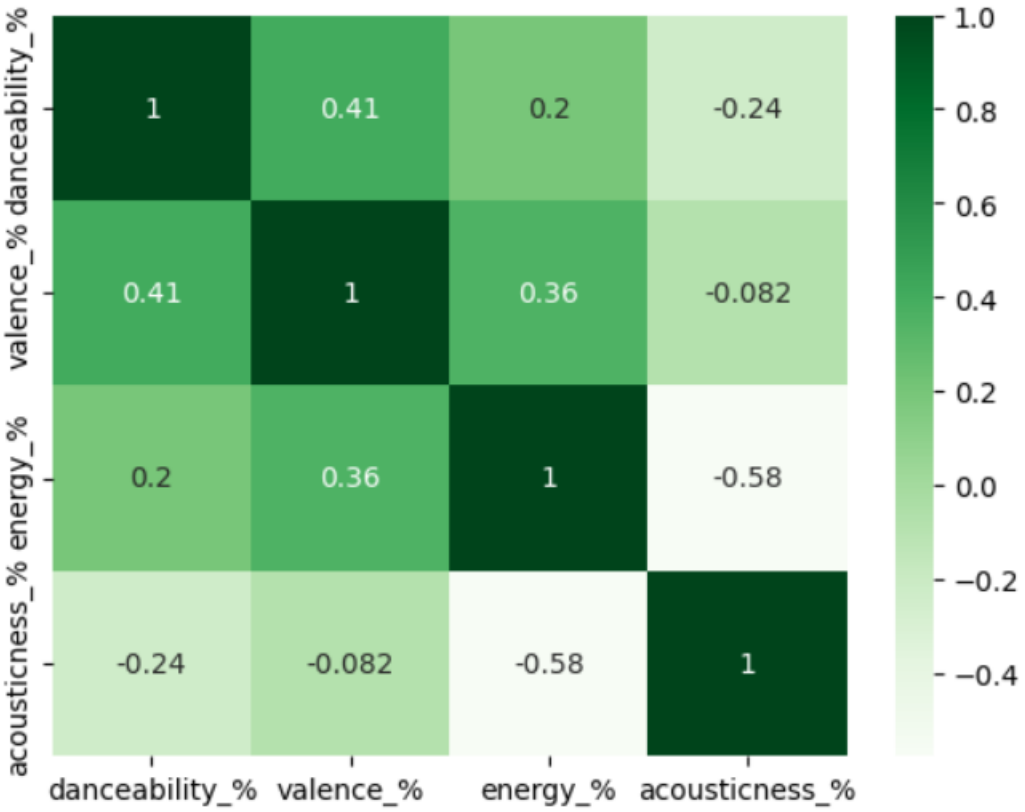

Figure 3. Correlation between danceability, valence, energy, and acousticness.

Based on the heat maps, there is a somewhat weak positive correlation between the percentage of valance and danceability. With a correlation value of 0.41. There is a somewhat strong negative correlation between the percentage of acousticness and percentage of energy. With a correlation value of -0.58. There is a weak positive correlation between the percentage of energy and percentage of valence. There is a very weak positive correlation between percentage of energy and percentage of danceability.

IV. DATA SET GRAPHICAL EXPLORATION

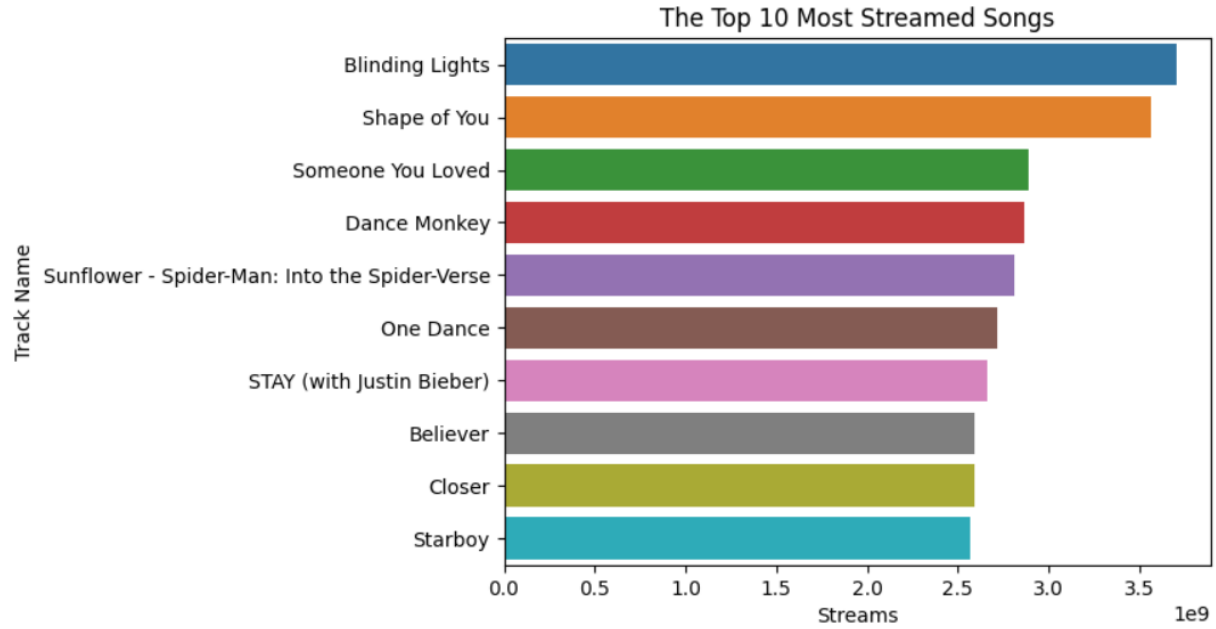

Figure 4. Comparing the Number of Streams for the Top 10 Songs and Track Name of each Song

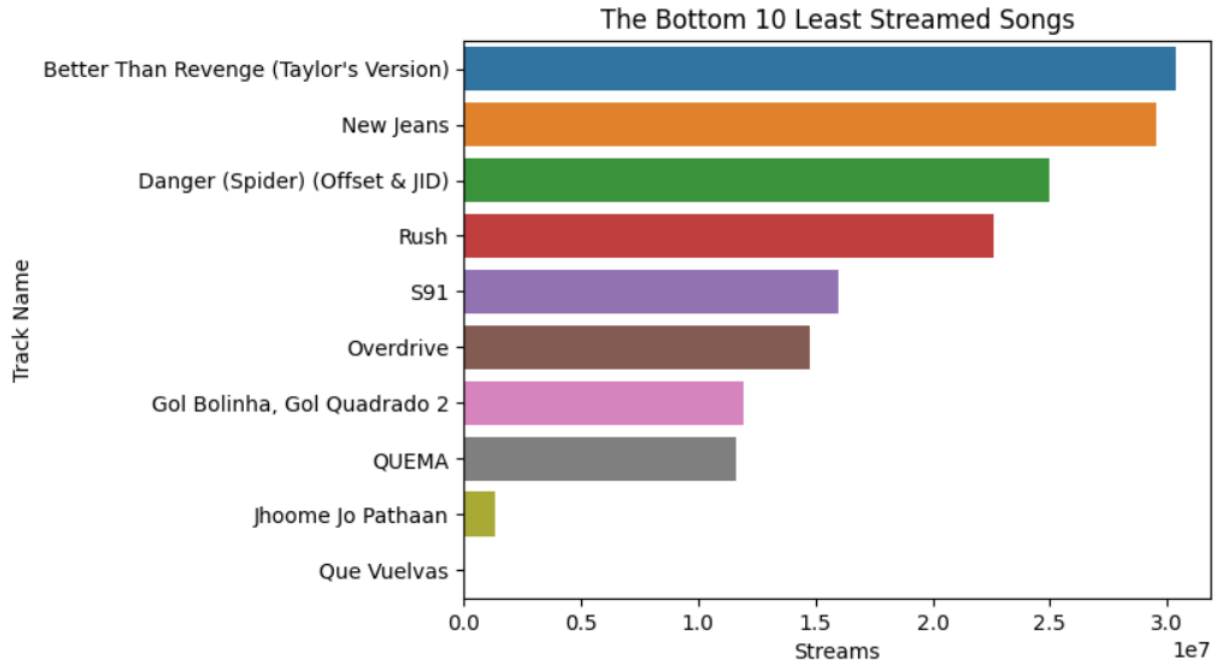

Figure 5. Comparing the Number of Streams for the Bottom 10 Songs and Track Name of each Song

Based on these bar graphs, the most streamed song was Blinding Lights by The Weeknd, the second most streamed song was Shape of you by Ed Sheeran, and the third most streamed song was Someone you Loved by Lewis Capaldi. The least streamed song was Que Vuelvas by Carin Leon and Grupo Frontera. The second least streamed song was Jhoome Jo Pathaan by Arijit Singh, Vishal Dadlani, Shekhar Ravjani, and Sukriti Kakar. The third least streamed song was QUEMA by Sog, Ryan Castro, and Peso Pluma.

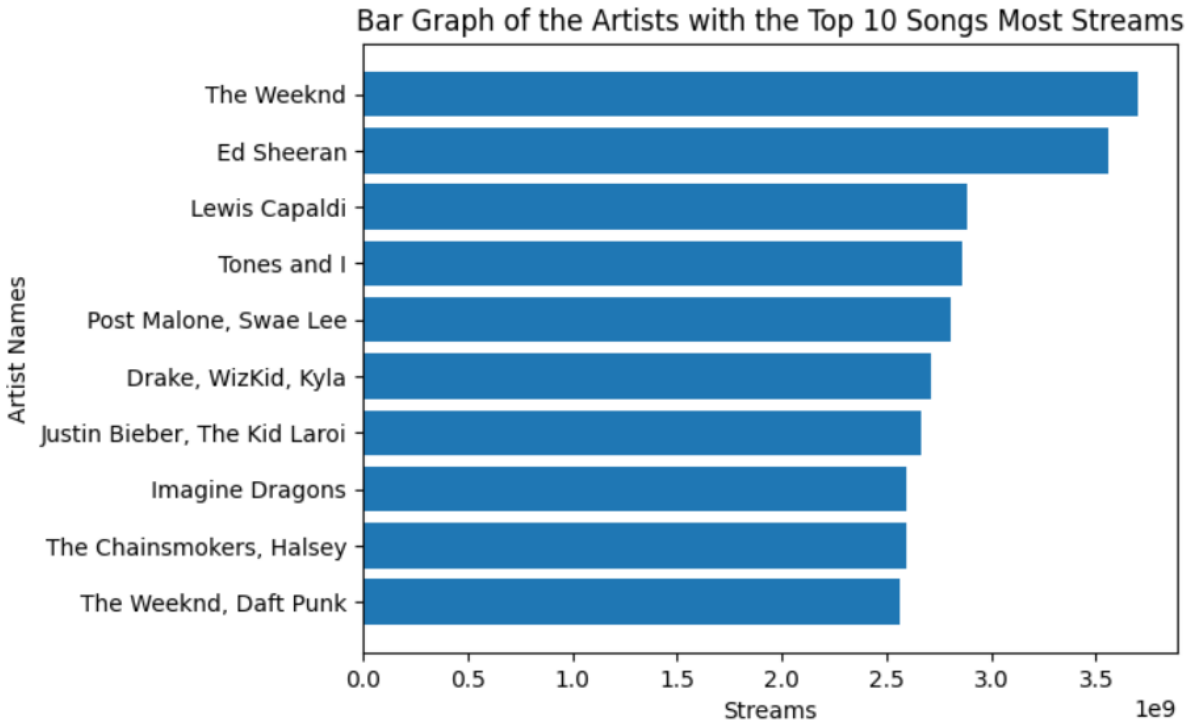

Figure 6: Comparing the Number of Streams for the top 10 songs and the artist's name.

Based on this bar graph, in 2023, The artist with the most streams on Spotify were the Weeknd with the artist with the second most streams being Ed Sheeran and the artists with the third most streams being Lewis Capaldi.

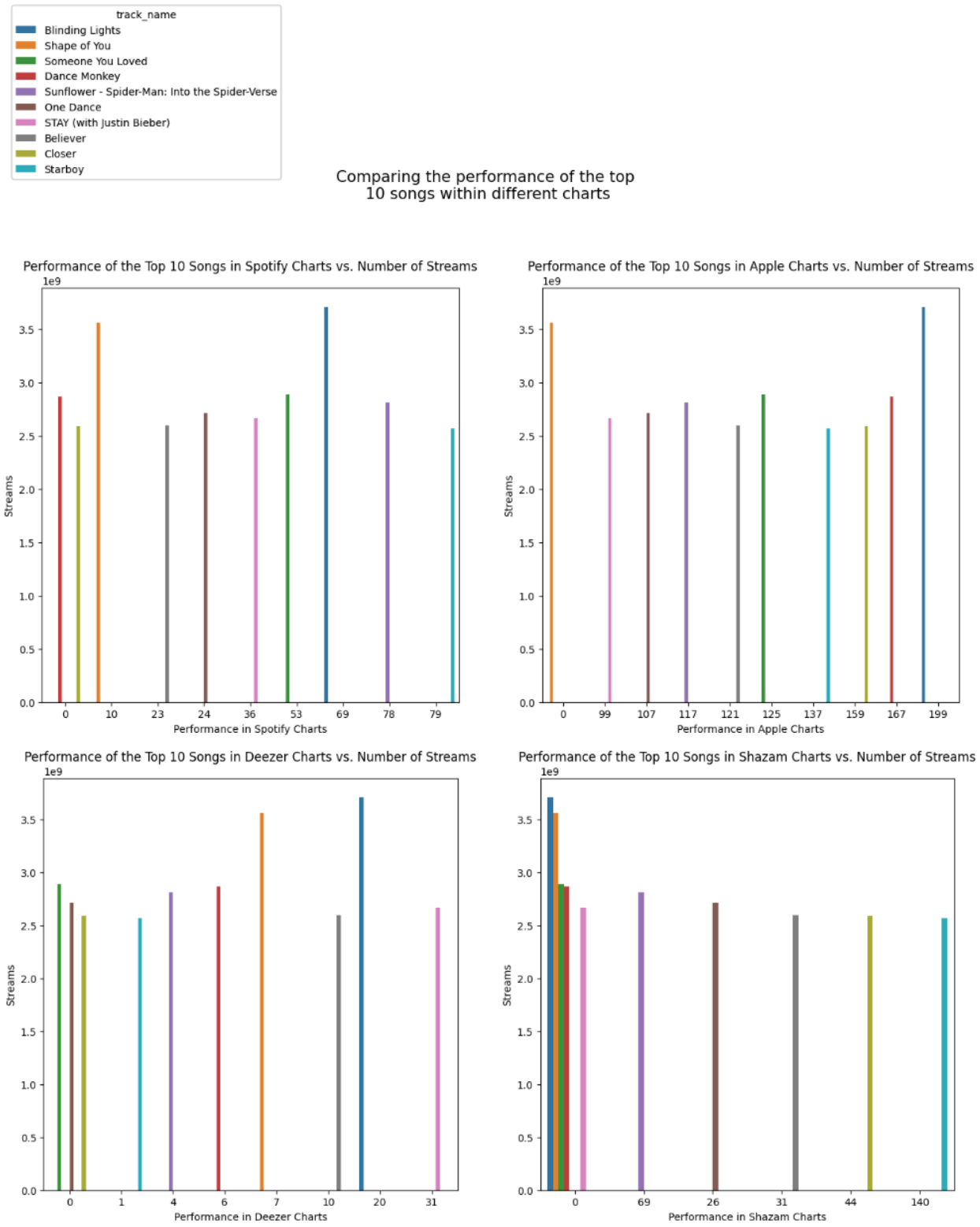

Figure 7: Comparing the performance of the top 10 songs within different charts.

Based on the bar graphs, the top streamed song "Blinding Lights" by The Weeknd seems to perform relatively well on Spotify charts, apple charts, and deezer charts relative to the other top 10 most streamed songs. However, its

performance on the shazam charts is very low relative to the other top 10 most streamed songs. Given the inconsistency between the number of streams each top 10 songs got and their corresponding performance across each chart, we can assume that there is no correlation between the number of streams and the performance of the song.

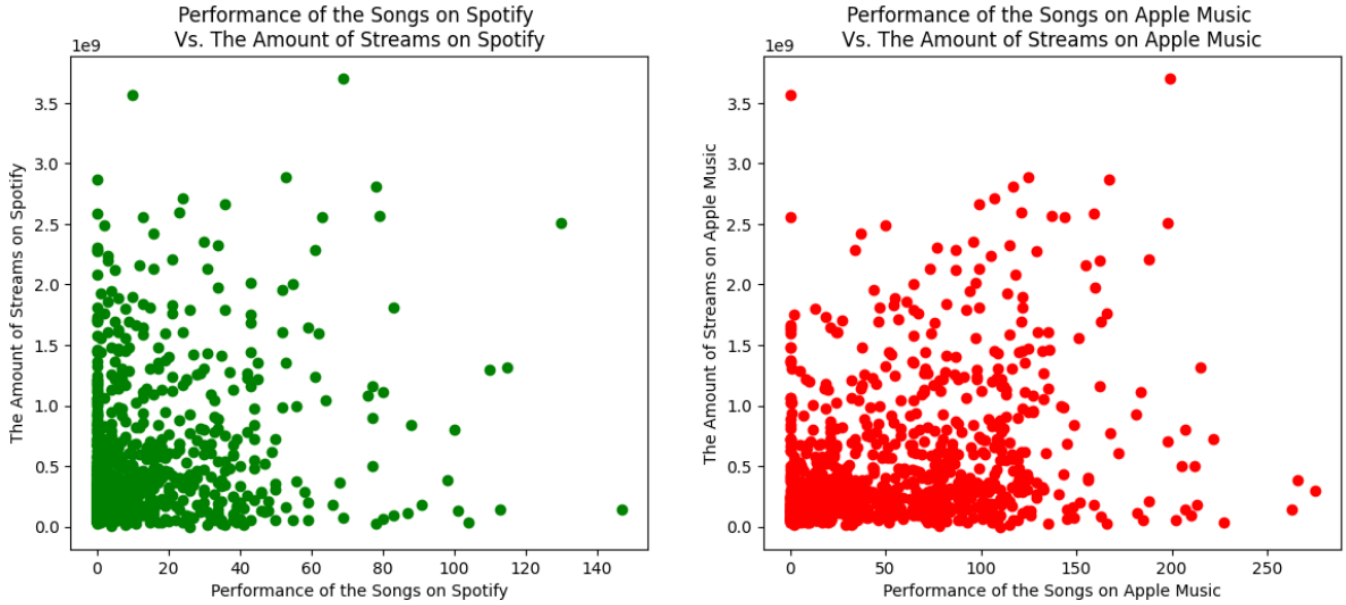

Figure 8: Correlation between the performance of the song and its streams for Spotify and apple

Based solely on the scatterplots, there seems to be a very weak correlation between the performance of the song and the number of streams it has within Spotify or apple charts. However, given how non-linear the graph is, this implies that there most likely is no correlation. The correlation value between the performance of songs on Spotify and the number of streams for each song is 0.24 making it a very weak positive correlation. The correlation value between the performance of songs on apple music and the number of streams for each song is slightly higher compared to Spotify with a value of 0.32. This is also a very weak positive correlation.

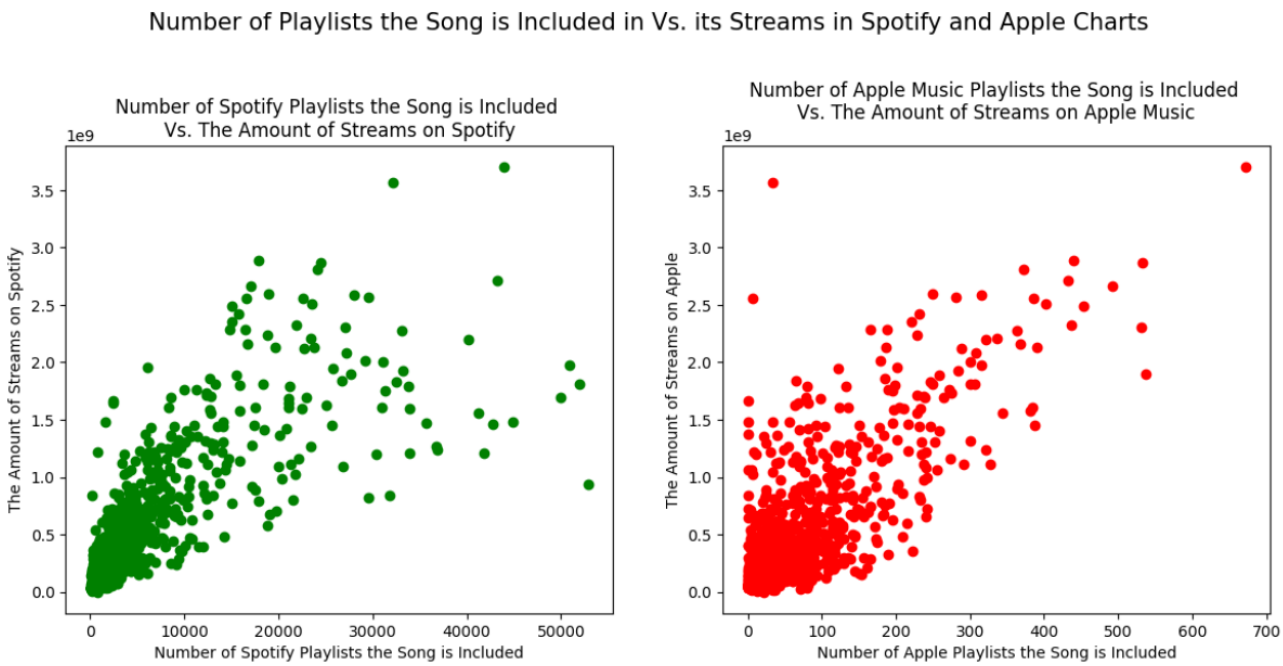

Figure 9: Correlation between the number of playlists the song is included in and its streams for Spotify and apple

Based on the scatterplots, there seems to be a relatively strong positive correlation between the number of playlists the song is included in and the number of streams it has in both Spotify and apple charts.

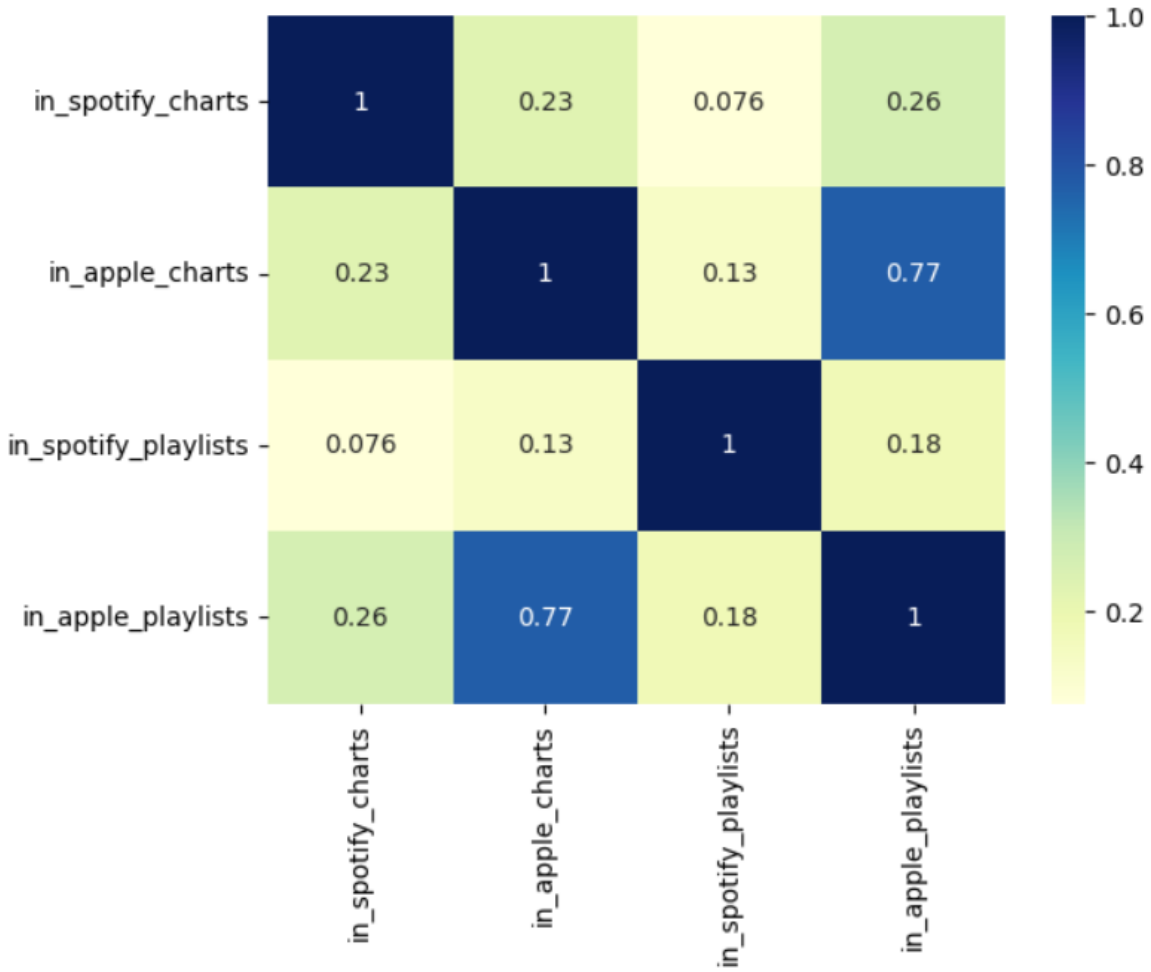

Figure 10: Correlation between the number of playlists the top 10 songs are included in and their performance in the Spotify and apple charts

When looking at the top 10 songs from the data set, interestingly, there is a strong positive correlation between the performance of the top 10 songs in apple charts and the number of playlists they are included in. However, there is no correlation between the performance of the top 10 songs in Spotify charts and the number of playlists they are included in.

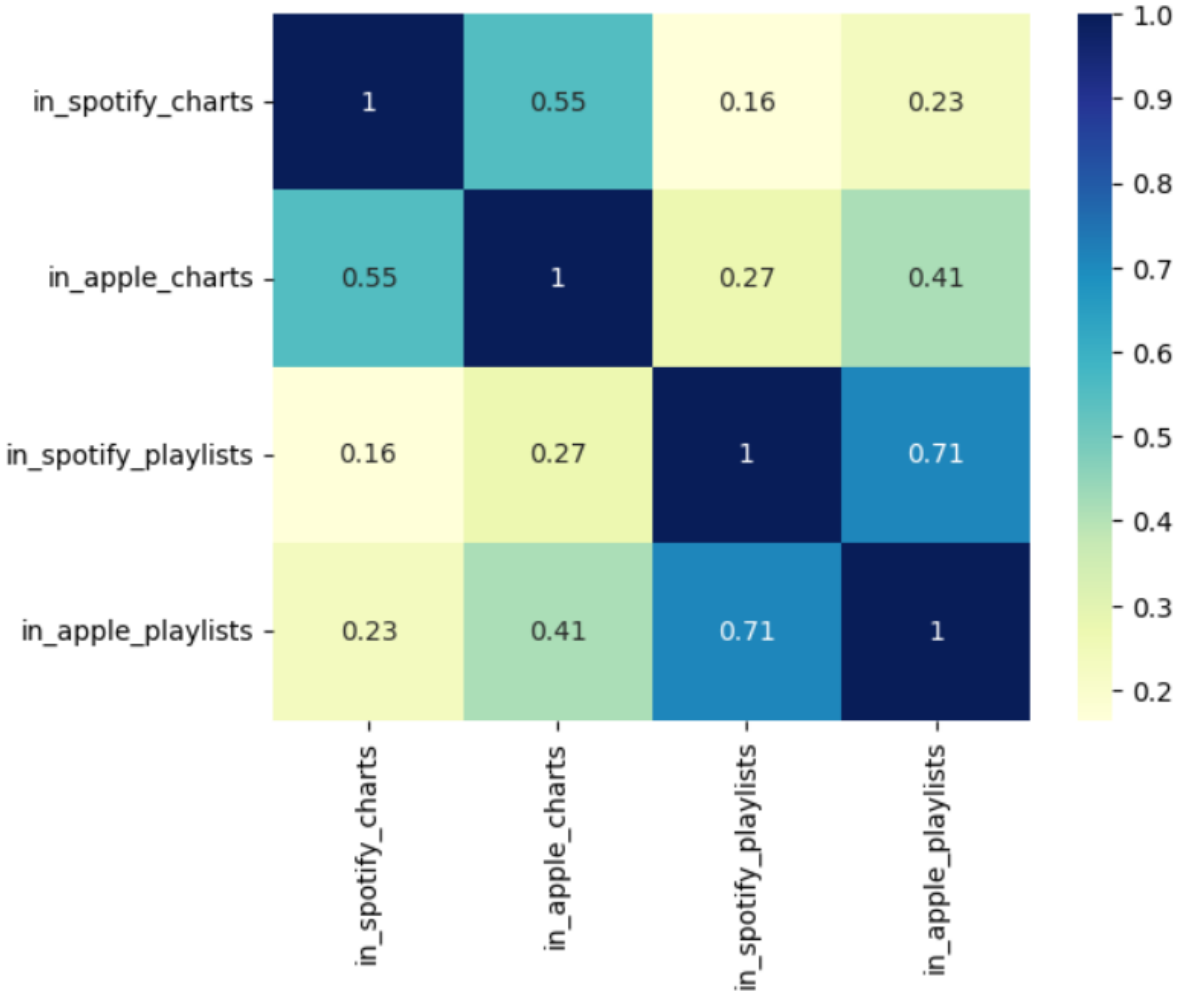

Figure 11: Correlation between the number of playlists the songs are included in and their performance in the Spotify and apple charts.

When looking at the songs from the whole dataset, there is a strong positive correlation between the number of Spotify playlists the songs within the dataset are included in and the number of apple playlists the songs within the dataset are included. Additionally, there seems to be a relatively strong positive correlation between the performance of the songs within Spotify charts and the performance of songs within apple charts. Interestingly, there is also a relatively weak positive correlation between the performance of the songs in apple charts and the number of playlists they are included in. However, there is a very weak positive correlation between the performance of the songs in Spotify charts and the number of playlists they are included in.

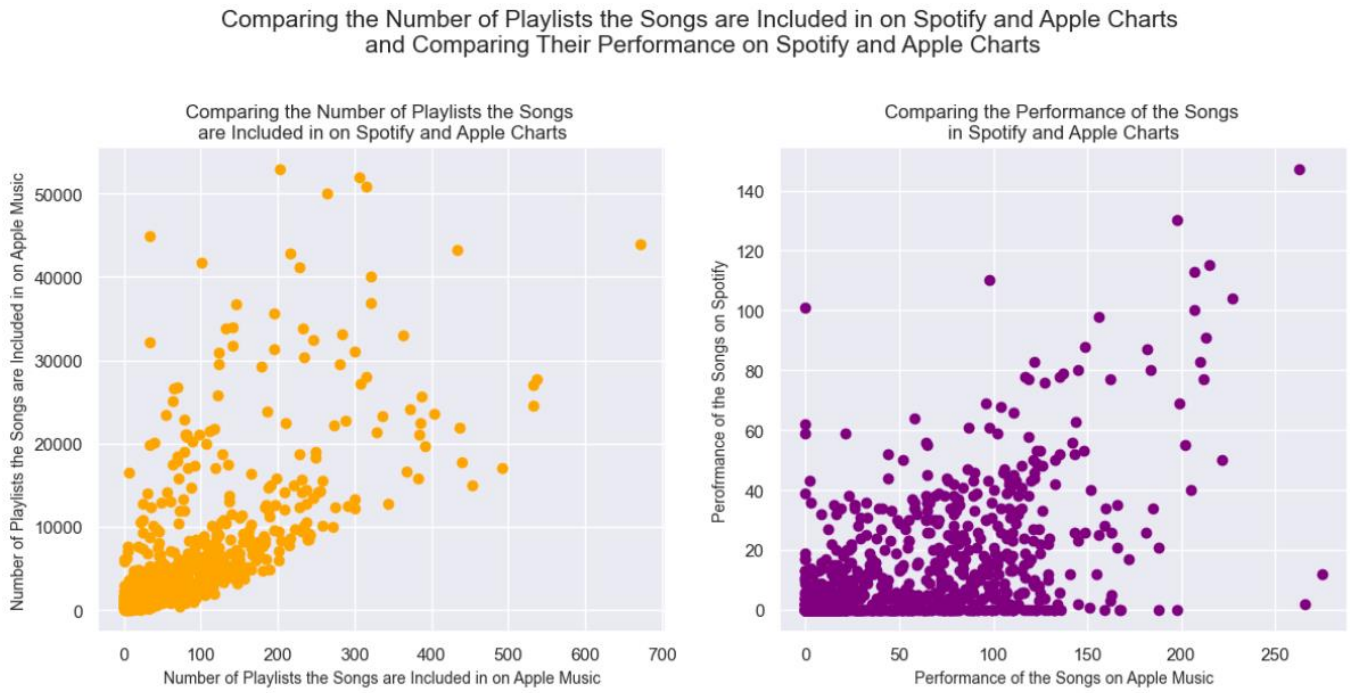

Figure 12: Comparing the Number of Playlists the Songs are Included in on Spotify and Apple Charts and Comparing Their Performance on Spotify and Apple Charts

Based on the scatterplot, there is a very strong positive correlation between the number of apple playlists the songs of this dataset are included in and the number of Spotify playlists the songs of this dataset are included in. However, by comparison, there is a slightly weaker positive correlation between the performance of songs within this dataset in apple charts and the performance of songs within this dataset in Spotify.

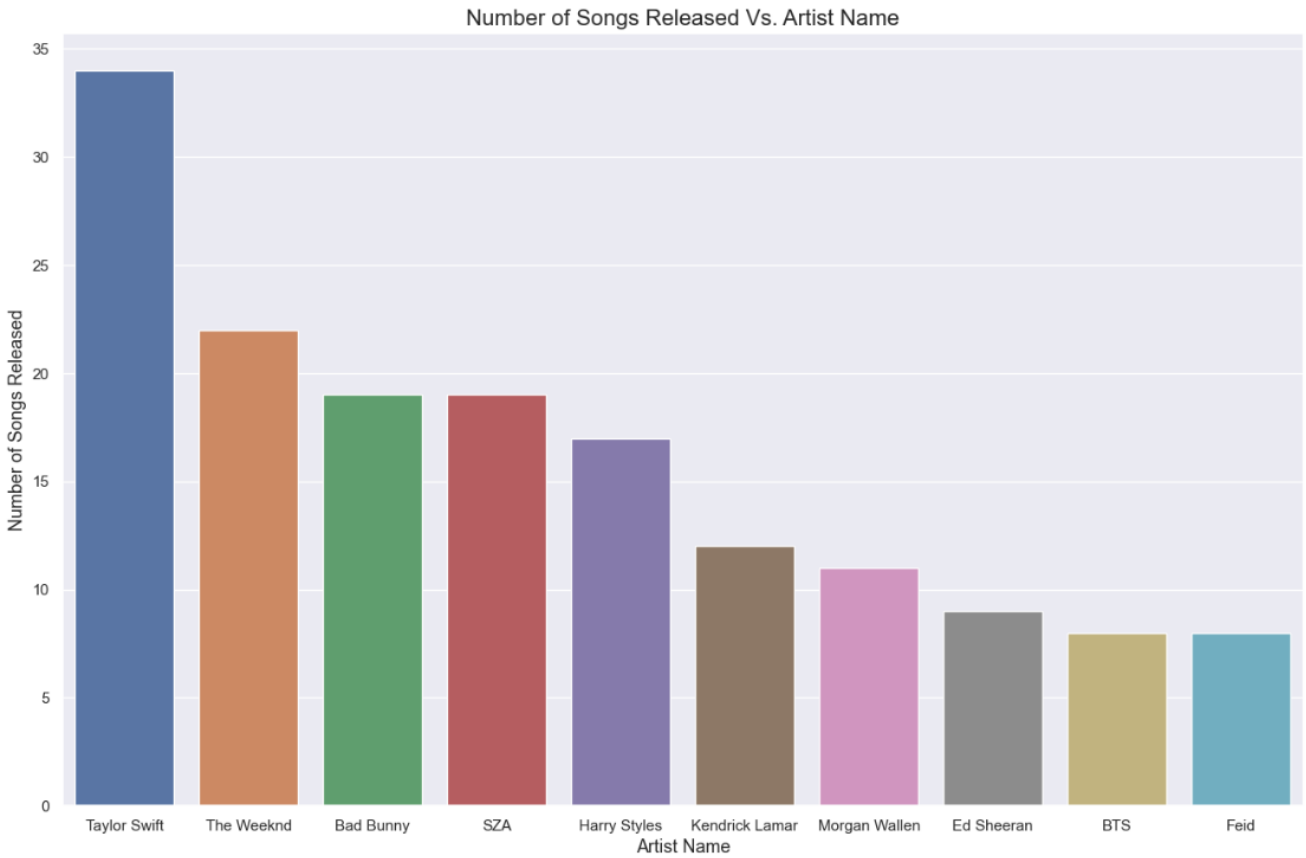

Figure 13: Visualization of which artist released the most number of songs.

The artist who released the greatest number of songs was Taylor Swift with a total of 34 songs.

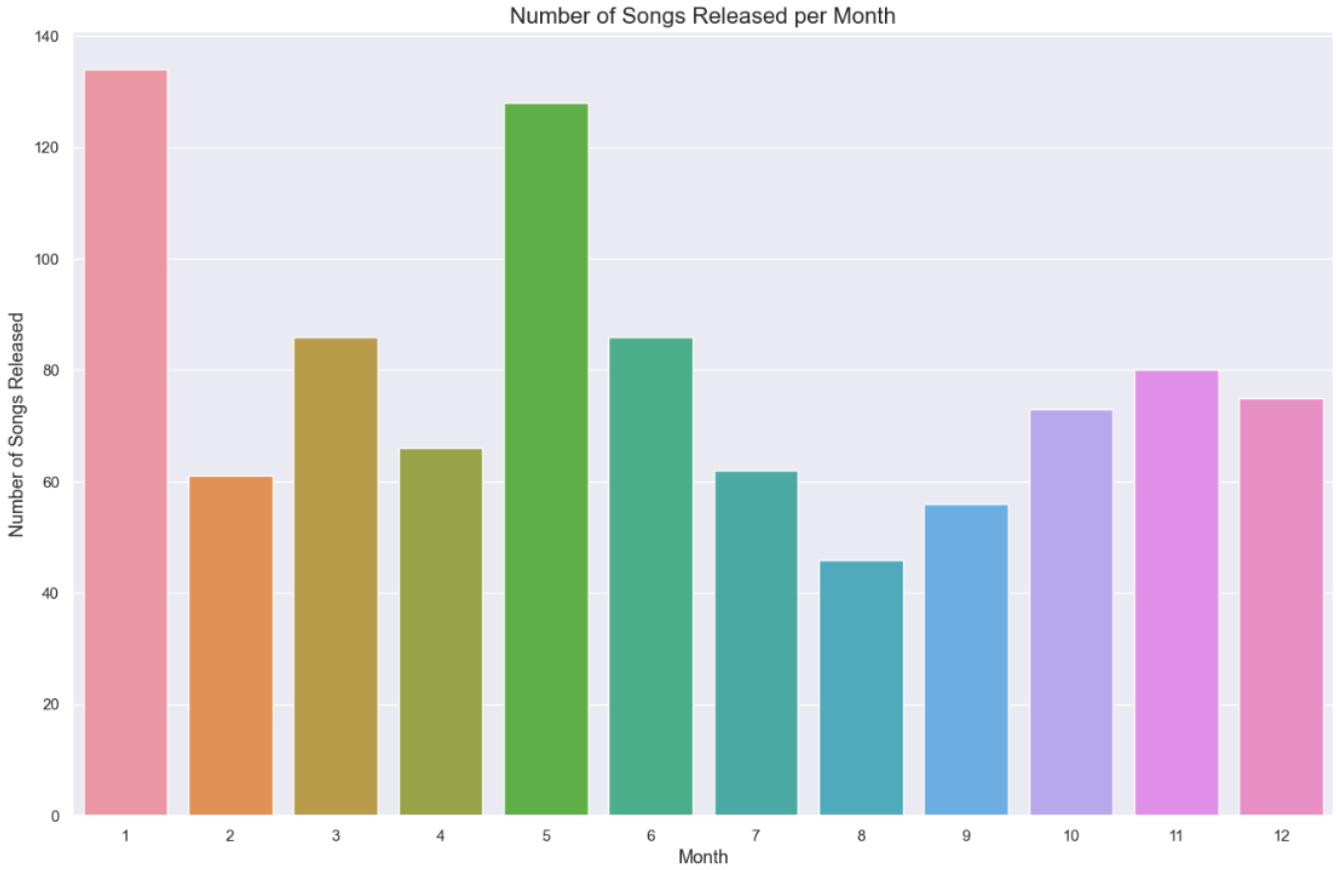

Figure 14: Visualization of which month had the greatest number of songs released.

Within this dataset, throughout the years, the most amount of songs released were in January with a total of 134 songs. The second month with the most amount of songs released was May with a total of 128.

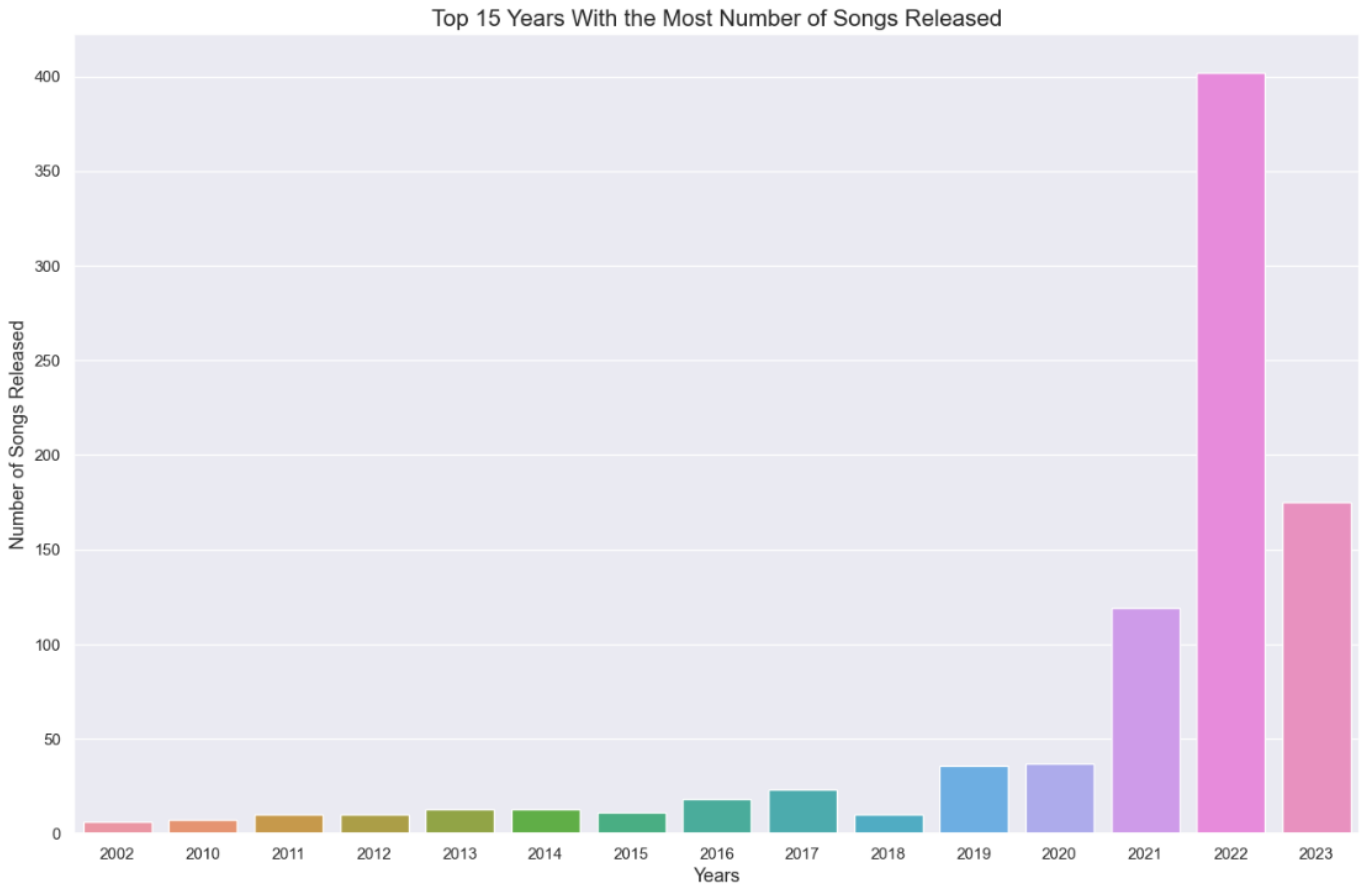

Figure 16: Visualization of which of the top 15 years had the greatest number of songs released.

Within this dataset, 2022 was the year with the greatest number of songs released with a total of 402. 2023 was the year with the next highest number of songs released with a total of 175. 2021 was the year with the third highest number of songs released with a total of 119.

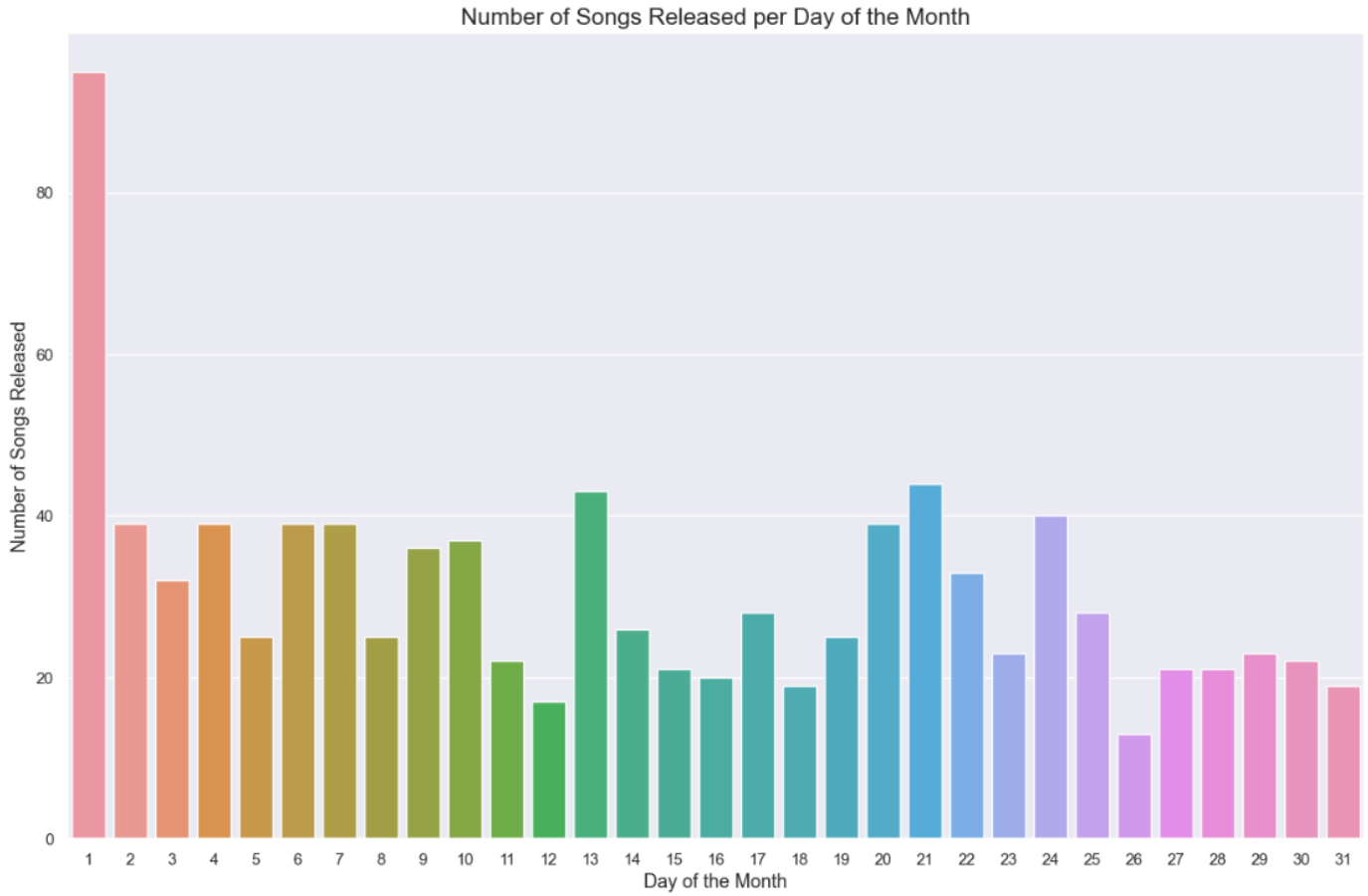

Figure 17: Visualization of which day of the month had the greatest number of songs released.

Within this dataset, throughout the years, the first day of the month is when the greatest number of songs are released with a total of 95 songs.

V. SUMMARY OF FINDINGS

Based on our findings, the most streamed song was Blinding Lights by The Weeknd while the least streamed song was Que Vuelvas by Carin Leon and Grupo Frontera. The artist who released the greatest number of songs was Taylor Swift with a total of 34 songs. Throughout the years, most of the songs were released in January and May with the first day of the month having the greatest number of releases. 2022 saw the highest number of songs released compared to any other year within the dataset.

The data within apple music was very interesting. When focusing on the top 10 songs, there is a strong positive correlation between the performance of the top 10 songs in apple charts and the number of playlists they are included in. When focusing on songs from the dataset, there is also a relatively weak positive correlation between the performance of the songs in apple charts and the number of playlists they are included in. By comparison, the correlation from the data within Spotify charts was very weak.

In terms of the correlations between performance, streams, and number of playlists for the songs within this dataset, there is a relatively strong positive correlation between the number of playlists the song is included in and the number of streams it has in both Spotify and apple charts. However, there is a very weak correlation between their performance of songs and the number of streams. Additionally, there is a strong positive correlation between the number of Spotify playlists the songs within the dataset are included in and the number of apple playlists the songs within the dataset are included.

When analyzing the various components of the songs from this entire dataset, there is a somewhat weak positive correlation between the percentage of valance and danceability. There is a somewhat strong negative correlation between the percentage of acousticness and percentage of energy. When focusing on the energy level within the songs, the track that has the most energy is "I'm Good Blue" by Bebe Rexha and David Guetta. The track with the least energy is "What Was I Made For? from the Motion Picture Barbie" by Billie Eilish.