Data Visualization: Assignment 3

Siddharth Menon
IMT2022001
IIIT Bangalore
Bengaluru, India
Siddharth.Menon@iiitb.ac.in

Shreyank Bhat
IMT2022516
IIIT Bangalore
Bengaluru, India
Shreyank.Bhat@iiitb.ac.in

Vrajnandak Nangunoori
IMT2022527
IIIT Bangalore
Bengaluru, India
Vrajnandak.Nangunoori@iiitb.ac.in

***Abstract*—This document contains the visual analytics workflow in visualizing the kaggle dataset "Most streamed Spotify Songs 2024", incorporating additional datasets.**

I. Datasets Used:

1) Most Streamed Spotify Songs in 2024
2) Spotify Tracks Genre

Since the date cutoff for *Spotify Tracks Genre* dataset is 2022, even the *Most Streamed Spotify Songs in 2024* dataset had to be cut down to only have songs from 2022 or earlier, to facilitate a merge of the two datasets.

The following two workflows were aimed primarily at determining how much of a correlation the technical characteristics of a song has with its popularity, and how its popularity is affected by it. This also serves as the justification for the choice of secondary dataset.

A merge was performed between these datasets, using a string similarity function to compare names in the string columns.

II. Tasks

The objective for the analysis is to be able to observe and figure out the factors that contribute to the success or failure of songs and artists on streaming platforms. The factors considered include:

- Song information. The details include track information, various things like danceability, energy, loudness, liveness, tempo, genre etc.
- Artist information. The genres associated with the artists etc.

III. Visual Analytics Workflow

A. Workflow 1: Artist Characteristics and popularity

1) ***Iteration 1: Platform Relevance****:* For much of this analysis, we will be focusing only on the most relevant music streaming platforms. The 'relevance' of a platform is influenced mainly by two factors:

- What are the total number of streams/views received on a platform?
- Can the popularity of songs on one platform be directly attributed to its overall popularity which is reflected in other platforms too? A good example for this is, songs that do well on TikTok aren't necessarily popular music on other platform, and their popularity is attributed to their performance on one platform.

Identify applicable funding agency here. If none, delete this.

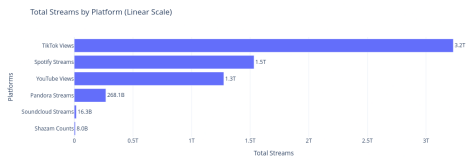


Fig. 1. Total Streams by Platform

Data

The data for loop 1 is a subset of the assignment 1 dataset alone. It includes:

- Track Name
- Spotify Streams
- YouTube Views
- TikTok Views
- SoundCloud Streams
- Pandora Streams
- Shazam Counts

Visualisation

Three visualisations were created with the above data. Among the two metrics listed earlier for gauging the 'relevance' of a platform, for the first metric we have a histogram with the total number of streams on each platform. We see that TikTok, Spotify and YouTube far outperform any other platforms in figure 1.

As for the second metric that we used, i.e. the relevance of the platform, we used a sunburst chart to check how the top ten songs by total streams perform on all of the platforms. If we find that certain platforms are key contributors to the overall streams of a song, we can take this as a sign of their relevance, as their streams on those platforms decide the popularity of the song.
Total streams were initially calculated by adding TikTok Streams as well. We quickly ran into a problem, as we found

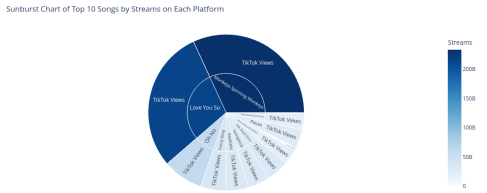


Fig. 2. Sunburst Chart: Top 10 songs by streams across all platforms

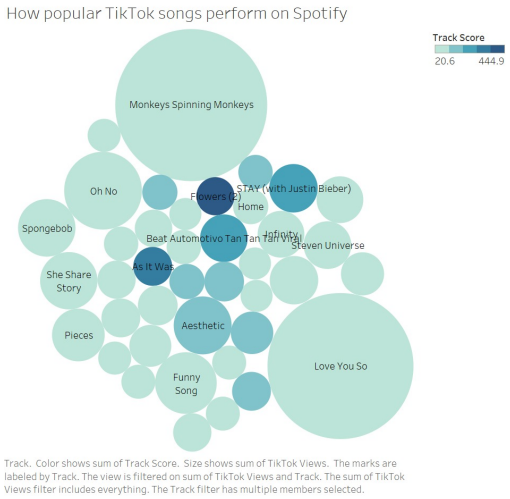


Fig. 3. Track Score vs TikTok popularity. Track Score (indicated by the colour channel), is a function of the Spotify Popularity of the song. The songs with the highest TikTok popularity aren't close to as popular on Spotify.

that TikTok seemed to far outweigh other platforms, and the top 10 songs were dominated by songs with many streams on TikTok but very few on the other platforms.

This is consistent with our observations from assignment 1, where we found that songs that do really well on TikTok do not do particularly well on Spotify and YouTube, possible because of the kind of music being used in TikTok, with their purpose often being only to complement the actual visual content. On the other hand, songs on Spotify and music videos on YouTube offer the music itself as their primary content. This can be seen in figure 3.

Therefore, for a more fair reflection, we also did the same by calculating total streams without TikTok views, to also study the class of songs which are popular, but due to TikTok's dominance, it gets drowned out in total streams.

Figures 4 and 5 show that after TikTok views are removed in the calculation of total streams, other platforms find better representation. It is worth noting that TikTok is included in the visualisation though, showing that songs that are popular on other platforms are popular on TikTok too, although it wasn't true for the converse.

Spotify and YouTube were the two best performing songs in the new visualisation, and this is seen in the analysis of a single artist, i.e. Ed Sheeran, when we drill down.

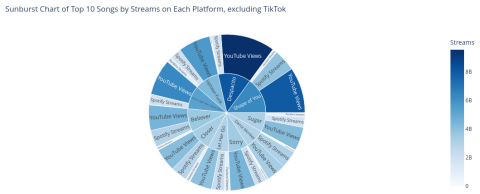


Fig. 4. Sunburst Chart: Top 10 songs calculated without TikTok Views

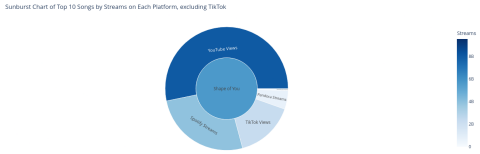


Fig. 5. Sunburst Chart: Platform relevance for a single song, Shape og You

Knowledge

TikTok, Spotify and YouTube are the most popular platforms, that is backed by sufficient statistical evidence.

*2) **Iteration 2: Artist popularity analysis**:* In this analysis we will be studying the popularity of artists across different streaming platforms. Our objective in this section are the following:

- How well have the top artists performed in the streaming platforms as compared to each other?
- Contribution of streams of the top 60 artists sorted on Total Streams
- Which of the platforms turn out to be most popular for the top 60 artists.
- Do some artists have a monopoly over a platform?

Data:

The data for iteration 2 is taken from the 'Most Streamed Spotify Songs 2024' which requires the following columns:

- Artist
- Youtube Views
- TikTok Views
- Spotify Streams
- And an additional column that takes the aggregate of the streams in each of these streaming platforms for each row 'Total Streams'
- Dataframe for the top 60 songs were taken sorted on total streams.

Visualisations:

We have plotted the following visualisations for this loop

- To begin the analysis, a general Tree Map was created to provide an overview of the streaming success of

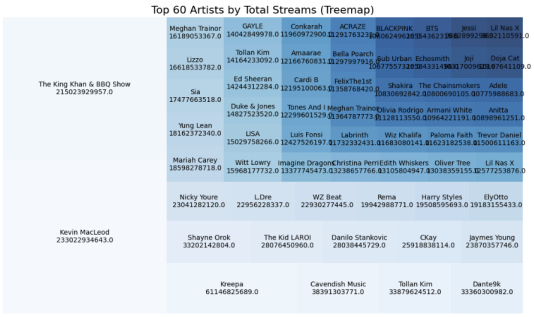


Fig. 6. Tree Map for top 60 artists

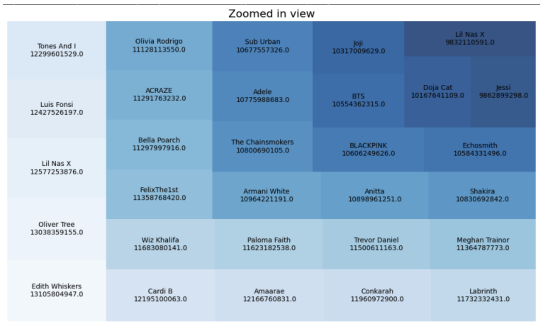


Fig. 7. Tree map for bottom half of top 60 artists(Zoomed in view)

various artists. This initial visualization, highlights the top 60 artists based on their total streams, depicted in Figure 6. Figure 7 shows the bottom half of the top 60 artists for a more clear view. By offering a clear comparison of streaming counts, this plot serves as an introductory perspective on how certain artists have achieved higher streaming numbers relative to others.

- We aimed to identify artists who demonstrated a monopoly over a particular platform and those who contributed significantly across multiple platforms. To analyze this, we aggregated the data for the top 60 artists and divided them into batches of 20. For each batch, we calculated the average number of streams across the platforms under consideration: YouTube Views, Spotify Streams, and TikTok Views. Using these averages, we plotted a pie chart for each batch, which visualizes the proportional contribution of each platform to the total streams for that group of artists.
  The pie charts provide an initial understanding of how streams are distributed across platforms. By analyzing these distributions, we can assess whether certain platforms dominate specific batches or whether there is a balanced contribution from multiple platforms. For instance, the presence of a dominant platform in the pie chart indicates that a particular group of artists relies heavily on that platform for their streams. On the other hand, a more balanced distribution suggests that the artists have a diverse audience spread across multiple platforms.

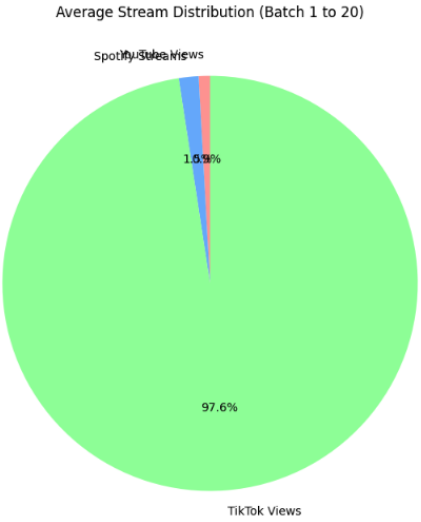


Fig. 8. Pie chart for artists ranked 1-20

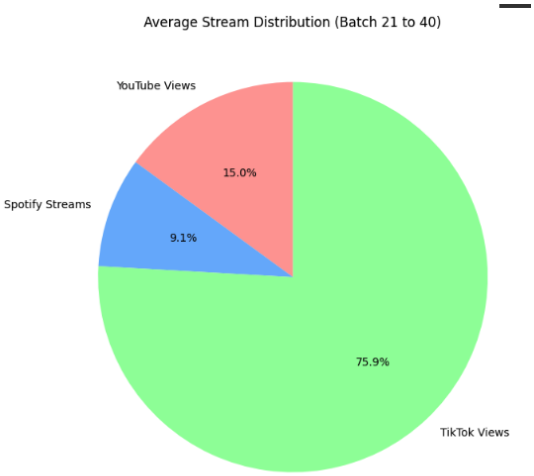


Fig. 9. Pie chart for artists ranked 21-40

This analysis serves as a foundational step to decide how to further investigate these patterns.Figure 8 - Figure 10 shows the visualizations.

- We plotted a bar chart at batches of twenty inferring from the pie chart. These bars were grouped for each artist showing the spotify, you tube and Tik Tok streams. This analysis helps in finding out those artists that had better distribution across platforms and those that focused on a single platform. Figure 11 - Figure 13 shows the bar charts.

Knowledge:

If we examine the three bar charts that were plotted, we

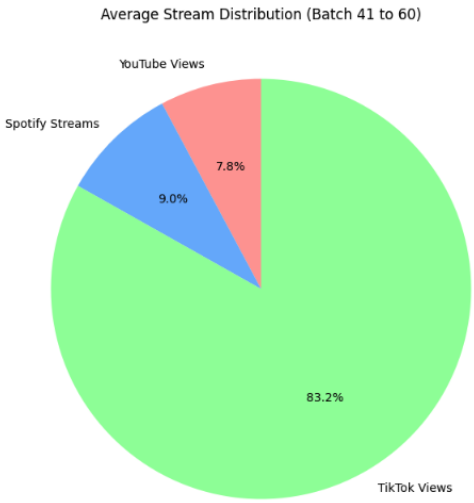


Fig. 10. Pie chart for artists ranked 41-60

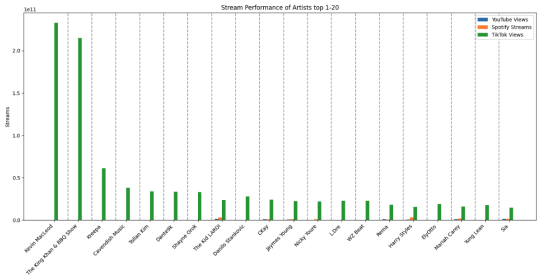


Fig. 11. Grouped bar charts for artists ranked 1-20

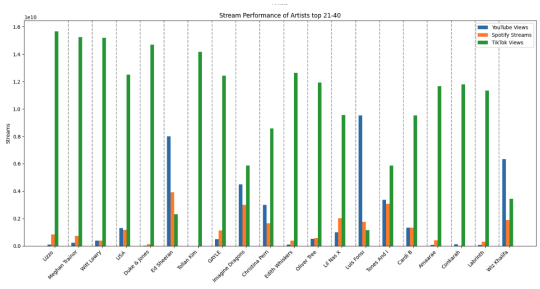


Fig. 12. Grouped bar charts for artists ranked 21-40

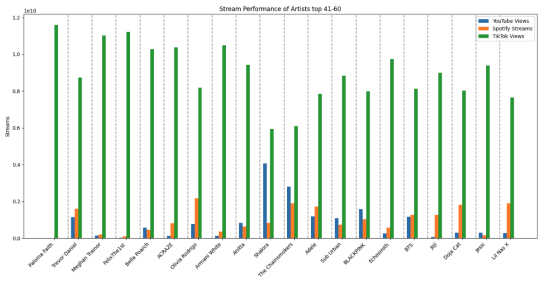


Fig. 13. Grouped bar charts for artists ranked 41-60

observe that the artists depicted in Figure 12 and Figure 13 exhibit a better distribution of streams across the platforms. In contrast, the artists in Figure 11 tend to focus their streams predominantly on a single platform. For example, Kevin MacLeod, The King Khan & BBQ Show, Kreepa, and Cavendish Music showed poor distribution of streams. In contrast to this, a little more even distribution of streams across platforms could be noted for Imagine Dragons, Wiz Khalifa, Ed Sheeran and The Chainsmokers. This difference portrays the diversity with respect to the level of platform dependency of the artists and their streaming strategies.

*3) **Iteration 3: Predicting Artist Popularity Distribution**:* This iteration of the workflow builds on the knowledge about the varying nature of artist dominance on different platforms - about how it can be very even sometimes, and also really skewed. Different machine learning models are experimented with to try and predict how an artist would fare on different platforms, by making predictions on their songs. The aim, very broadly, was to create a model that would tell which 'category' a song belonged to, with respect to its popularity on different platforms (performed much better in 1 platform vis-à-vis other platforms/performed well in 2 but not in a third, performed well in all 3 platforms almost equally) and then predicting the overall dominance of an artist as an aggregate/average of the performance of their songs. This would then be used to determine the strength of the correlation between the characteristics of song, in terms of the nature of the song, with its popularity on a platform, if at all there was any. The quality of the predictions made by our model would be an indicator of the presence and extent of said correlation.

Data:

A merge was performed on the 2 datasets being used in this assignment. The merge was on the (Track, Artist) tuple, where a string similarity check was performed to account for the cases where the same song appeared on two different datasets with slightly different track names or artist names. Filtering was performed based on the inferences from the first iteration, keeping only Spotify, YouTube and TikTok streams/views for further analyses.

Model

Two different approaches were tried for the creation of the model. One was to perform unsupervised clustering on the data, with the assumption that music that perform differently on different platforms must be fundamentally different from each other in terms of its characteristics (metrics such as tempo, energy, acousticness etc.).
Therefore, an unsupervised clustering was done using the K-Means clustering algorithm. Although the result (represented visually in 14 after reducing to 2 dimensions

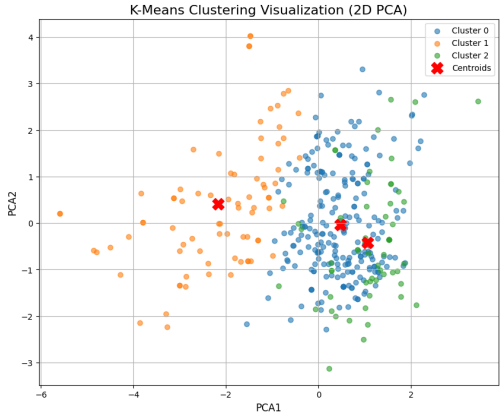


Fig. 14. Result of Unsupervised Clustering on Data

using PCA) does have 2-3 clusters which are vaguely separable, a count of the categories that points in each cluster belong to, it proved to be of no real significance, as all clusters were evenly composed.

A switch was made to then using a supervised model, the random forest classifier in this case. The model would output 3 True or False values [x y z] - with x corresponding to the prediction about whether or not the song performed well in Spotify, y about TikTok and z about YouTube.

Accuracy: The model yielded accuracies of 71%, 70% and 77% respectively for predicting performances on each of the platforms. This accuracy corresponds to predicting the dominance of each song on different platforms. However, when an aggregation of these results is used to predict the artist performance on different platforms, this accuracy was found to be higher, as was seen in manually tested cases, with data from outside the dataset being used to verify model predictions.

Knowledge:

The Random Forest Classifier model proves the existence of a weak, although existent correlation between song performance across platforms and the nature of the song itself. This correlation is stronger for songs and artists that do well on YouTube than Spotify or TikTok. The exact attributes that impact the performance of a song on a platform will be studied in the next iteration of this workflow.

4) ***Iteration 4****: Mapping artists and their song nature to the categories:* In this iteration, we build upon the insights gained from the third iteration to plot a Parallel Coordinates Plot (PCP). This visualization allows us to analyze the correlations between song characteristics and their nature. By examining these relationships, we aim to understand how various attributes influence a song's dominance across different platforms.

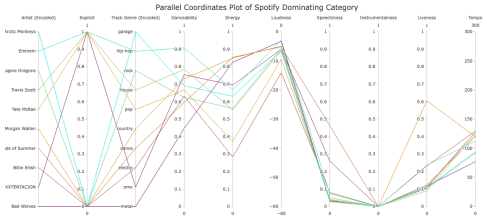


Fig. 15. PCP for songs in Spotify dominating category

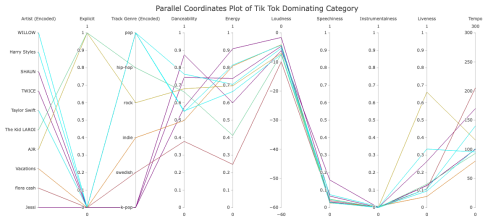


Fig. 16. PCP for songs in Tik Tok dominating category

Data:

We proceeded to make use of the merged data in iteration 3. Then we divided the data into 4 different csv files, where each csv file contained songs and its information belonging to a particular category, namely 'Spotify Dominant', 'Youtube Dominant' , 'Tik Tok dominant' , 'No dominance'. 10 artists from each category were taken to plot the PCP and to analyse the correlation. Due to the presence of greater variance in the 'Valence' and 'Acousticness', we decided to drop those columns from this pcp.

Visualizations:

Visualizations for each category has been plotted using the parallel coordinates plot which was built using the Plotly.js and D3.js api.

Figure 15 to Figure 18 shows the pcp for different categories.

For the nature of the songs like Danceability, Energy, Loudness, Speechiness, Liveness and Tempo we can compare to show how these influence the category that the songs belong to. Instrumentalness was close to 0 for all

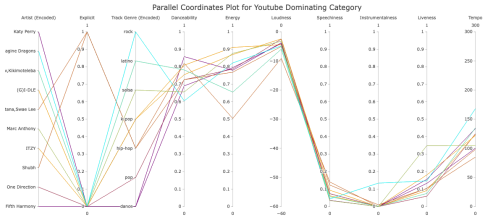


Fig. 17. PCP for songs in Youtube dominating category

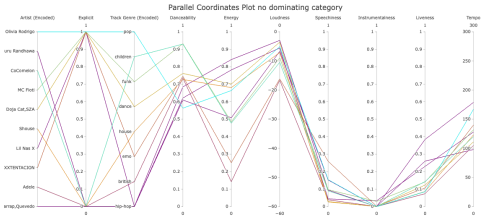


Fig. 18. PCP for songs in undefined dominating category

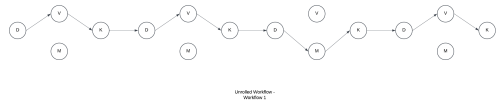


Fig. 19. Unrolled Workflow - Workflow 1

categories and The track Genres had no real pattern so we don't include it in the comparison below.

- For the **Spotify** dominant category(Figure 15), Danceability and Energy displayed no significant patterns due to high variance. However, Loudness consistently ranged between -15 and -5. Speechiness averaged around 0.1, while Liveness and Tempo were approximately 0.1 and 100 to 125, respectively.
- For the **YouTube** dominant category(Figure 17), there were fewer explicit songs, with non-explicit songs being more prevalent. Other features exhibited lower variance. Danceability averaged around 0.7, and Energy was between 0.8 and 0.9 for the majority of the songs. Loudness ranged from -10 to 0, Speechiness was between 0.05 and 0.15, Liveness averaged around 0.1, and Tempo averaged approximately 125.
- For the **TikTok** dominant category(Figure 16), explicit songs were again fewer in number. Danceability, Energy, Liveness, and Tempo showed high variance, making them less contributory to the classification of songs in this category. Loudness ranged from -10 to 0, while a significant portion of Speechiness fell between 0 and 0.1.
- For the **No dominant** category(Figure 18), Tempo primarily ranged between 100 and 150, and Liveness was predominantly close to 0.1. However, other features did not exhibit meaningful patterns due to high variance.

Knowledge

The classification of songs into distinct categories was clearly demonstrated in the PCP, highlighting a correlation between the nature of the songs and their respective categories.

B. Workflow 2

The insights from **Assignment - 1** were considered for songs released prior to the year 2024. However, since the additional dataset we are using contains songs released prior to 2022, we only consider the merge between the two datasets, taking into account songs released prior to the year 2022. In addition, the following analysis uses the knowledge from Workflow 1 that the TikTok, Spotify, YouTube are the most relevant platforms.

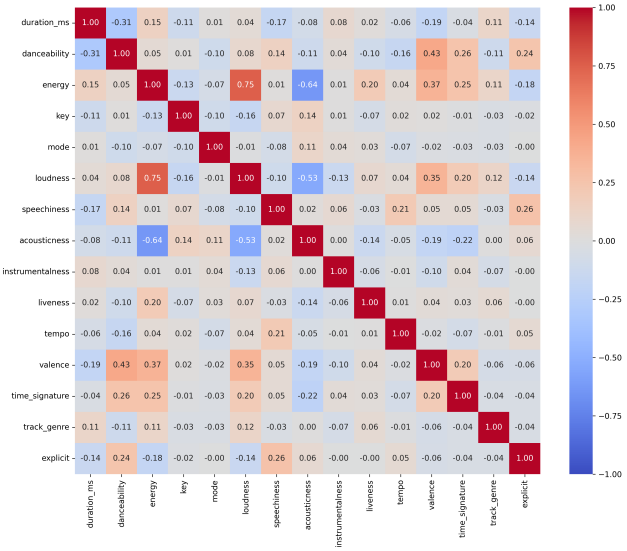


Fig. 20. Correlation Matrix for Dataset Features

Dataset Description: The merged dataset contains information regarding songs spanning approximately 45 genres. The following features will be used in the workflow:

- **track_name**: Name of the individual track.
- **duration**: Duration of the song in milliseconds.
- **danceability**: A score from 0 to 1 representing how suitable a track is for dancing based on various musical elements.
- **energy**: A measure of the intensity and activity of a track in the range 0–1.
- **key**: An estimated overall key of the track, represented by an integer where values map to pitches using standard pitch class notation (e.g., 0=C, 2=D, etc.).
- **mode**: Indicates the modality (major or minor) of a track, with major represented by 1 and minor by 0.
- **loudness**: The loudness of the track in decibels.
- **speechiness**: A score from 0 to 1 that represents the presence of spoken words in a track. Values closer to 1 indicate mostly spoken content.
- **acousticness**: A score from 0 to 1 that represents the extent to which a track possesses an acoustic quality.
- **instrumentalness**: A score from 0 to 1 representing the likelihood of a track being instrumental.
- **liveness**: A measure detecting the presence of an audience, where higher values indicate a higher probability of live performance.
- **tempo**: The overall estimated tempo of a track in beats per minute (BPM).

- **valence**: A score from 0 to 1 representing the positiveness conveyed by a track.
- **time_signature**: The number of beats in each bar of the track.
- **track_genre**: The genre of the track.

To understand the features in the additional dataset, we first analyze the correlation matrix of the numerical attributes to get a basic idea of the features themselves.

Insights from the Correlation Matrix

1) **Energy and Loudness**: Tracks with higher energy tend to be louder (**correlation: 0.75**).
2) **Acousticness and Energy**: Tracks with higher acousticness are less energetic (**correlation: -0.64**).
3) **Valence and Danceability**: Danceable tracks are often more positive in emotion (**correlation: 0.43**).
4) **Speechiness and Explicit**: Songs with spoken words are slightly more likely to be explicit (**correlation: 0.26**).
5) **Tempo and Other Features**: Tempo has no significant correlation with other features, suggesting it may vary independently.

From the insights of the previous workflow, we understand that TikTok, Spotify, and YouTube are the more relevant platforms among the various platforms available. This workflow aims to conclusively derive insights regarding what things contribute to a song's success or failure.

*1) **Iteration 1: Exploring the influence of features on platform-specific top performing songs**:* We first try to figure out how the core features, namely 'duration_ms', 'energy', 'loudness', 'speechiness', 'tempo' and 'instrumentalness' influence a song's popularity for the top-performing songs.

Inorder to figure out which factors are more prominent among the 3 platforms, we visualize the differences of means by using radar plots as shown in figures 21 and 22 (where figure 21 shows the scaled down values of the mean

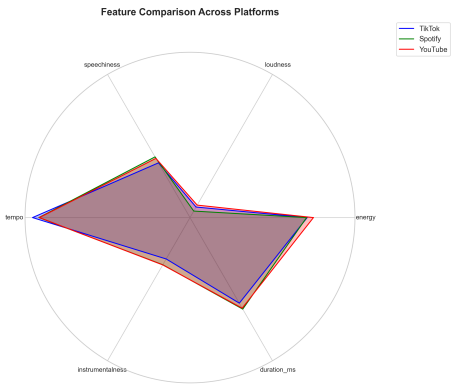


Fig. 21. Radar Chart: Min-Max scaled down means of the factors for the 3 platforms.

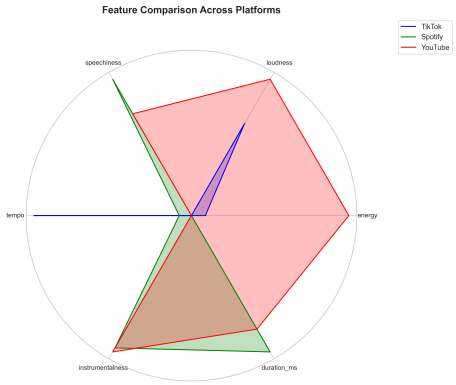


Fig. 22. Radar Chart: Min-Max normalized means of the factors for the 3 platforms.

values of the factors for a better idea of the actual values and figure 22 shows the min-max normalized values for the mean value of the features among the 3 platforms)

Insights

We draw our insights from Table 1, showing the values obtained from the plots.

- **Speechiness:** Songs with a relatively less speechiness score tend to perform better.
- **Tempo:** The 3 platforms have a high average, indicating that fast-paced songs tend to perform well, on spotify more than the other 2 platforms.
- **Instrumentalness:** We observe a very low average across the platforms, showing that songs with more vocals perform better.
- **Duration:** The better performing songs are close to 3 and a half minutes.
- **Energy:** A medium energy tends to influence the songs to getting more views, close to 0.65.
- **Loudness:** The 3 platforms have a loudness value close to -5 DB, indicating that louder songs are preferred.

	Ideal value	**Min value**	**Max value**	**The Influence (3 platforms)**
Speechiness	≈ 0.1	0.024	0.463	Spt > YT > TT
Tempo	≈ 130	48.718	205.561	TT > Spt > YT
Instrumentalness	≈ 0.003	0	0.703	YT > Spt > TT
Duration (sec)	≈ 203	94.2	436.7	Spt > TT > YT
Energy	≈ 0.651	0.142	0.974	YT > TT > Spt
Loudness (DB)	≈ -5.93	-16.6	-0.17	YT > TT > Spt

TABLE I
THE INFLUENCE OF THE 6 FACTORS ON THE TOP 3 PLATFORMS (TIKTOK - TT, SPOTIFY - SPT, YOUTUBE - YT)

We observe that a lower instrumentalness score is preferred and a low speechiness score is also preferred. However this would mean to have more vocals in the song and at the same have less spoken words indicating that songs with more

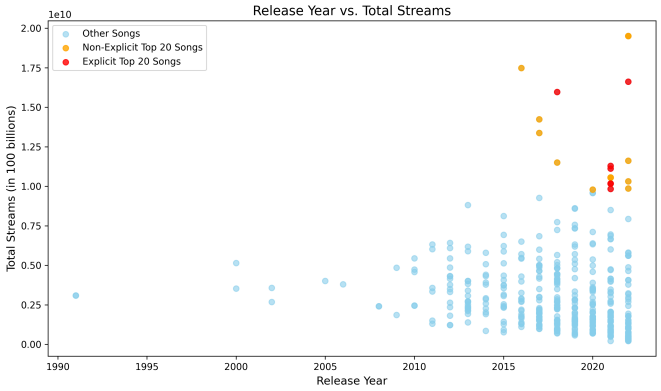


Fig. 23. Scatter Plot: Release Years of songs against combined streams.

humming noises could prove to influence a song's performance in a good way.

*2) **Iteration 2: An year-wise analysis of songs**:* We now try to analyse if the release years of the songs influence their reach by checking their combined views. We additionally, also check for the same using the 'popularity' score for performance on spotify.
Using a scatterplot, as shown in figure 23, we observe the release years of the songs and color the top performing songs.

Insights

- We observe from figure 23 that the songs released over the recent years (2018-2022) have more combined views compared to the songs that were released earlier. This could be due to many reasons, such as more content of songs due to advent in ease of sharing videos etc, or some other reasons.
- We also observe that there are some songs with explicit content (6 songs), mainly in the year 2021 when considering the top 20 songs. This could mean that a song would perform well being non-explicit rather than containing explicit content.

*3) **Iteration 3: Combined Influence of the factors and Release months of songs**:* We've understood from the previous loop that the songs released in the recent years performed the best. Now, we try to go a step further and check if there are any other reasons, such as festivals or any other holidays, that play a role in influencing a song's reach. For this, we try to analyse the trends in factors for top 20 songs, disregarding the influence of year for now as we consider the **years 2018-2022** only.
The figure 24 represents a grid a scatter plots, plotting the values of the 6 factors for the 3 platforms. The red colored scatter-points indicate that the points lie far away from the threshold values (the same as the ideal values in Table 1) and the green points indicate that the points lie close to the threshold values. The margins used for the 6 factors were:

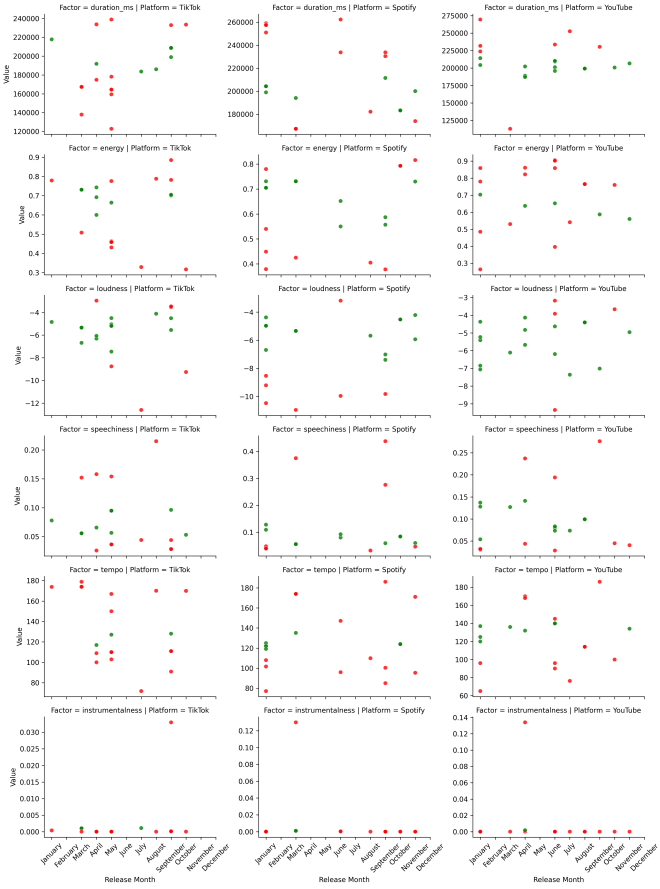


Fig. 24. Facet Grid of Scatter Plots: Release months of songs against the 6 factors's values.

speechiness - 0.05, tempo - 15, instrumentalness - 0.002, duration - 20 seconds, energy - 0.1, loudness - 2

	TikTok	**Spotify**	**YouTube**
Speechiness	0.05 - 0.15	0.05 - 0.15	0.05 - 0.15
Tempo	100-120,160-180	100-140	100-140
Instrumentalness	0.001	0.001	0.001
Duration (sec)	160-200, 230-240	180-220	180-220
Energy	0.55-0.75	0.55-0.75	0.55-0.75
Loudness (DB)	-4 to -8	-4 to -8	-4 to -8
Best Months	Mar-May, Sep	Jan, Sep-Nov	Jan, June (others too)

TABLE II
BEST VALUES FOR PLATFORM SPECIFIC SONGS

Insights

- **Seasonal Preferences:**
  - **TikTok:** The best seasons to release songs are **spring (March to May)** and **early autumn (September)**, aligning with increased engagement during these times.
  - **Spotify:** Songs tend to perform better during **winter (January)** and **autumn (September to November)**,

suggesting a preference for new music in colder, reflective months.
  - **YouTube:** While **January** and **June** stand out as optimal months, other months seem to work reasonably well, indicating a consistent performance across the year.
- **Threshold Comparisons:**
  - Across all platforms, songs with specific characteristics tend to align closely with the threshold values:
    - **Speechiness:** Ideal range is **0.05–0.15**.
    - **Tempo:** Optimal ranges vary:
      - **TikTok:** Prefers **100–120 bpm** or **160–180 bpm**, possibly for its short-form content and dance trends.
      - **Spotify and YouTube:** Favor a broader range of **100–140 bpm**.
    - **Instrumentalness:** Low instrumentalness (close to **0.001**) is common across platforms, suggesting vocals are key to popularity.
    - **Duration:**
      - **TikTok:** Prefers extremes, with durations clustering around **160–200 seconds** or **230–240 seconds**.
      - **Spotify and YouTube:** Stick to the mid-range, preferring **180–220 seconds**.
    - **Energy and Loudness:** Energy levels of **0.55–0.75** and loudness between **-4 to -8 dB** are universally favored, reflecting a balance of vibrant yet not overwhelming audio.
- **Platform-specific Trends:**
  - TikTok's song preferences exhibit greater variability, as seen by clusters of red points far from the thresholds. This may reflect the platform's trend-driven, dynamic nature.
  - Spotify and YouTube have more green points, indicating a stronger adherence to threshold ranges, possibly due to their more consistent, curated algorithms and user behavior.

*4) **Iteration 4: Genre analysis of songs**:* We now try to look at a different aspect of songs on each platform to figure out whether a certain genre dominates any platforms or if there is equal preference for different genre.
We do this by considering pie charts, one for each platform, and check the percentages of the different genres while considering top 20 songs and then top 50 songs to get a better understanding of how genre distribution varies with the number of songs considered. By analyzing the genre breakdowns for both smaller and larger sample sizes, we can assess whether platforms tend to favor a specific genre or whether there is a more balanced representation across multiple genres.
Thus, we observe from figures 25 and 26, that while considering top 20 songs, Pop, K-pop tend to dominate, with

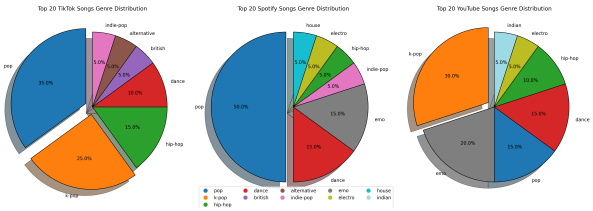


Fig. 25. Pie Chart: Genres of top 20 platform-specific songs.

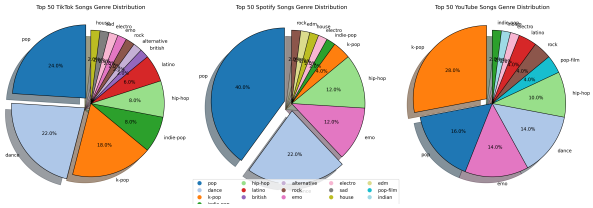


Fig. 26. Pie Chart: Genres of top 50 platform-specific songs.

TikTok favoring both, Spotify favoring Pop, and YouTube favoring K-pop in addition to having a better diversity of genres. When considering 50 songs, Pop continues to be dominant for the most part, with K-pop, dance, and emo also having a notable percentage. Other niche genres also tend to emerge such as 'rock', 'electro', 'indie-pop' with youtube showing a better percentage as compared to the other 2 platforms. Thus, considering everything, the genres 'pop', 'K-pop' and 'dance' tend to perform well, albeit different platforms favoring some of these.

We now try to confirm if the values of the 6 factors obtained from data exploration (in Table 2), still hold for the best genres to see their combined influence on a song's popularity. We do this by using a small multiples chart of violin plots to see the distribution of the values for the factors for the top 3 genres (considering all the songs) and check if they match with the values obtained in Table 2.

Thus, we observe from figure 27 that the 3 genres tend to exhibit values that are very close to the values obtained from table 2 for the 6 factors. However, we do notice an outlier in the 'dance' genre but can ignore it for the most part.

Thus, in conclusion, given a song with it's features (of genre, speechiness, loudness etc) the likelihood of it performing well can be maximised if it aligns with the following conditions:

- Considering the 6 factors, should have values close to those mentioned in Table 2.
- Considering when to release the song: Releasing the song in Autumn would likely result in better performance in all the 3 platforms. Releasing in months specified in Table 2 would give better results for a platform-specific popularity.
- Considering the genre of the song: The song is likely to perform well on TikTok if it is 'pop', 'dance' or 'k-pop', on Spotify if it is 'pop', 'dance' or 'emo' (but

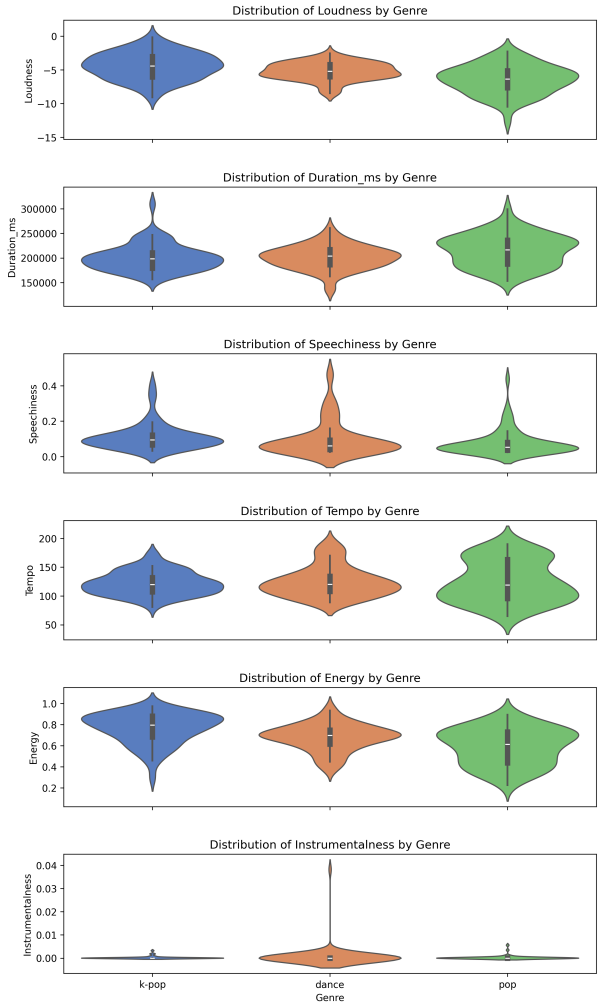


Fig. 27. Small Multiples Violin Chart: Distribution of the 6 factors for the 4 major performing genres.

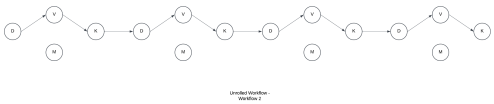


Fig. 28. Unrolled Workflow - Workflow 2

only to a certain degree), on YouTube if it is 'K-pop', 'emo', 'Pop' or even 'dance'.

IV. Work Distribution

- Siddharth Menon - IMT2022001
  - Preprocessing and Dataset Merging
  - Iterations 1 and 3 of Workflow 1
- Shreyank Bhat - IMT2022516
  - Workflow 1 - Iteration 2 - Artist Analysis
  - Workflow 1 - Iteration 4 - Mapping artists and their song nature to the categories
- Vrajnandak Nangunoori - IMT2022527
  - Workflow 2 - Iteration 1 - Influence of audio attributes of songs.
  - Workflow 2 - Iteration 2 - Influence of release year of songs.
  - Workflow 2 - Iteration 3 - Combined influence of the release months and the audio attributes.
  - Workflow 2 - Iteration 4 - Role of genre in determining song success.

Appendix

Assignment-1

Most Streamed Spotify Songs 2024

1st Siddharth Menon
(IMT2022001)
IIITBangalore
Electronic City, India
Siddharth.Menon@iiitb.ac.in

2nd Shreyank Bhat
(IMT2022516)
IIITBangalore
Electronic City, India
Shreyank.Bhat@iiitb.ac.in

3rd Vrajnandak Nangunoori
(IMT2022527)
IIITBangalore
Electronic City, India
Vrajnandak.Nangunoori@iiitb.ac.in

***Abstract*—This document presents a comprehensive analysis of the 'Most Streamed Spotify Songs 2024' Dataset available in Kaggle.**

I. Dataset Introduction

This dataset presents a comprehensive compilation of the most streamed songs on Spotify in 2024. It provides extensive insights into each track's attributes, popularity, and presence on various music platforms, offering a valuable resource for music analysts, enthusiasts, and industry professionals. The dataset includes information such as track name, artist, release date, ISRC, streaming statistics, and presence on platforms like YouTube, TikTok, and more. Key Features of the Dataset include:

1) **Track Name**: Name of the song.
2) **Album Name**: Name of the album the song belongs to.
3) **Artist**: Name of the artist(s) of the song.
4) **Release Date**: Date when the song was released.
5) **ISRC**: International Standard Recording Code for the song.
6) **All Time Rank**: Ranking of the song based on its all-time popularity.
7) **Track Score**: Score assigned to the track based on various factors.
8) **Spotify Streams**: Total number of streams on Spotify.
9) **Spotify Playlist Count**: Number of Spotify playlists the song is included in.
10) **Spotify Playlist Reach**: Reach of the song across Spotify playlists.
11) **Spotify Popularity**: Popularity score of the song on Spotify.
12) **YouTube Views**: Total views of the song's official video on YouTube.
13) **YouTube Likes**: Total likes on the song's official video on YouTube.
14) **TikTok Posts**: Number of TikTok posts featuring the song.
15) **TikTok Likes**: Total likes on TikTok posts featuring the song.
16) **TikTok Views**: Total views on TikTok posts featuring the song.
17) **YouTube Playlist Reach**: Reach of the song across YouTube playlists.
18) **Apple Music Playlist Count**: Number of Apple Music playlists the song is included in.
19) **AirPlay Spins**: Number of times the song has been played on radio stations.
20) **SiriusXM Spins**: Number of times the song has been played on SiriusXM.
21) **Deezer Playlist Count**: Number of Deezer playlists the song is included in.
22) **Deezer Playlist Reach**: Reach of the song across Deezer playlists.
23) **Amazon Playlist Count**: Number of Amazon Music playlists the song is included in.
24) **Pandora Streams**: Total number of streams on Pandora.
25) **Pandora Track Stations**: Number of Pandora stations featuring the song.
26) **Soundcloud Streams**: Total number of streams on Soundcloud.
27) **Shazam Counts**: Total number of times the song has been Shazamed.
28) **TIDAL Popularity**: Popularity score of the song on TIDAL.
29) **Explicit Track**: Indicates whether the song contains explicit content.

Identify applicable funding agency here. If none, delete this.

II. Analysis Goals

The objective of this analysis is to gain deeper insights into the dynamics of the music industry through visual exploratory analysis. Specifically, we aim to understand the following aspects:

1) **Trends Across Social Media Platforms:** Examine how different social media platforms have influenced the music industry, including their rise and decline in relevance.
2) **Artist Analysis and Listening Trends:** Investigate the trajectories of various artists, focusing on their emergence and potential resurgence in popularity.
3) **Impact of Playlist Curation:** Analyze how playlist placement affects the virality and longevity of songs, assessing the role of curated playlists in a track's success.

III. Data Pre-Processing

1) **Removal of Null Columns:** The column for TIDAL popularity was dropped as it contained only null values

across all entries, rendering it redundant for further analysis.

2) **Handling Missing Data:** Rows with over 50% null values were removed from the dataset. This step was crucial to ensure that the remaining data was sufficient for accurate analysis and modeling.
3) **Encoding Issues with Foreign Language Songs:** Songs in foreign languages that use non-Latin scripts (e.g., Japanese) were not correctly represented in the dataset due to encoding issues. Entries with song names rendered as gibberish characters (such as "ï¿½") were removed to maintain data integrity.
4) **Handling Duplicate Song Names:** For songs with identical names, a numerical suffix was added to distinguish between them. The second occurrence of a song was labeled as (2), the third as (3), and so on. This step ensures that each song entry is unique and identifiable.
5) **Addition of Rank by Year:** A new column, "Rank by Year," was introduced to capture the song's rank for each year. This addition helps in tracking the performance trends of songs over time.

IV. Data stories

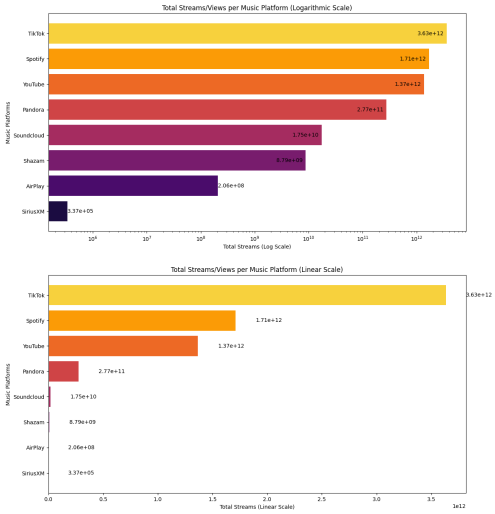
A. *Trends Across Social Media Platforms:*

The various platforms whose statistics are available in the Dataset are:

1) Spotify
2) Youtube
3) TikTok
4) Apple Music
5) AirPlay (radio stations)
6) SiriusXM
7) Deezer
8) Amazon Music
9) Pandora
10) Soundcloud
11) Shazam
12) TIDAL(Ignored in the report due to being an entirely NULL column)

1st Hypothesis:

Given its focus on music and its ease of accessibility, Spotify is expected to be the leading platform in terms of streams. It will likely be followed by YouTube and TikTok, which offer a wide range of features and attract more users. Shazam, which primarily serves to identify songs, is anticipated to rank lower than these platforms.



Total Streams/Views Accumulated by Music Platforms from most streamed Spotify songs in 2024

Fig 1: *Total Streams/Views Accumulated by Music Platforms from most Streamed spotify songs in 2024*

Observations: TikTok leads in terms of streams, followed by Spotify and YouTube. Shazam falls short in count despite being an app primarily used for identifying songs.

Possible reasons for the observed trends:

1) TikTok, released in 2016, rapidly accumulated users and became the most downloaded app in the United States by October 2018. Its appeal lies in its short, engaging content and widespread availability, which might contribute to TikTok's position as the most widely used platform.
2) Spotify, being a dedicated music streaming app, ranks second in terms of overall streams. YouTube, primarily used for longer videos, ranks third. Its high view count may be attributed to its long-standing presence and extensive user base.
3) Shazam's lower count could be due to the dominance of the top three platforms, which offer more features and better filtering options. As these platforms continue to evolve, users might prefer them over Shazam for streaming songs.

Considerations in the Graph:

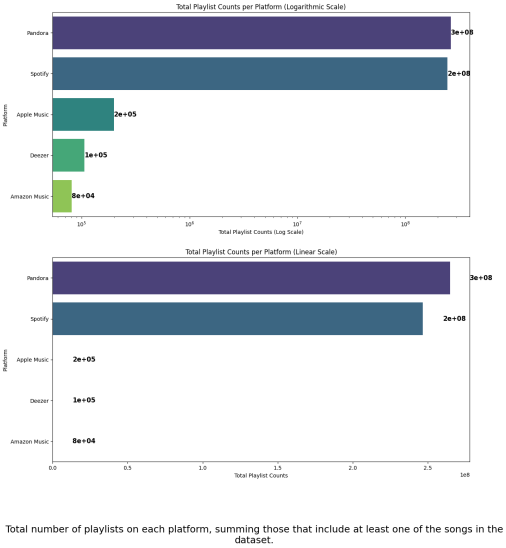
1) A horizontal bar plot was chosen for its aesthetic appeal, ease of readability.
2) The dataset contains columns for both streams and views, with some platforms reporting only one type. For consistency in plotting and analysis, streams have been used for platforms where they are available (e.g., Spotify) and views for others (e.g., YouTube), ensuring appropriate representation for each platform.
3) Text annotations have been used to display the exact values.
4) Marks used - lines
5) Channels used - Color(to represent categorical data i.e., platforms), Area(to represent quantitative data i.e., Total streams)

6) The first plot uses a logarithmic scale on the x-axis to better visualize the wide range of values. The second plot uses a linear scale on the x-axis to highlight differences in values that may not be as apparent in the logarithmic scale plot.
7) Platforms such as 'Apple Music', 'Deezer', and 'Amazon Music' are not included in the graph as the dataset does not contain columns representing streams/views for these platforms.

2nd Hypothesis:

Among the music streaming platforms 'Apple Music', 'Amazon Music', 'Deezer', 'Spotify', and 'Pandora', it is hypothesized that 'Apple Music' will likely lead in terms of total playlist count. This is due to the widespread popularity of iPhones, which come with the 'Apple Music' app pre-installed. Following 'Apple Music', 'Spotify' is expected to hold the second position due to its user-friendly interface, which is conducive to maintaining and managing playlists.

In third place, 'Amazon Music' is anticipated to perform strongly, benefiting from Amazon's global reach and extensive customer base. 'Pandora', a U.S.-based service owned by Sirius XM, is anticipated to place fourth, benefiting from its strong American user base. Lastly, 'Deezer', a French-based service, is expected to rank fifth, due to its smaller market size relative to Pandora's broader reach in the U.S.



Total number of playlists on each platform, summing those that include at least one of the songs in the dataset.

Fig 2: *Total Playlist Count on Each Platform for Playlists That Include At Least One of the Most Streamed Spotify Songs in 2024*

Key Observations:

1) Contrary to initial expectations, 'Pandora' has the highest total playlist count, likely driven by the large U.S. user base and the appeal of free streaming services. Figures 4 and 5 highlight that this high count is primarily due to a small number of songs contributing disproportionately to the platform's overall total.
2) As predicted, Spotify remains a strong contender in second place, which can be attributed to its user-friendly playlist creation features and ease of access.
3) The subscription-based nature of Apple Music and Amazon Music may explain their comparatively lower playlist counts, as many users gravitate towards free streaming options.
4) Despite offering free services, Deezer's lower popularity could be linked to France's smaller population, which limits its overall user base.
5) Figure 1 illustrates Spotify's leadership across most platforms in terms of streams and views, while Figure 2 reinforces its strong standing in playlist counts. Moreover, Figures 4 and 5 suggest that playlist counts on Spotify are more evenly distributed across songs, indicating a steady growth in playlists per track rather than a concentration of popularity around a few songs.

These findings underscore why both Pandora and Spotify outperform other platforms in playlist counts, reflecting higher levels of user engagement in curating and sharing playlists.

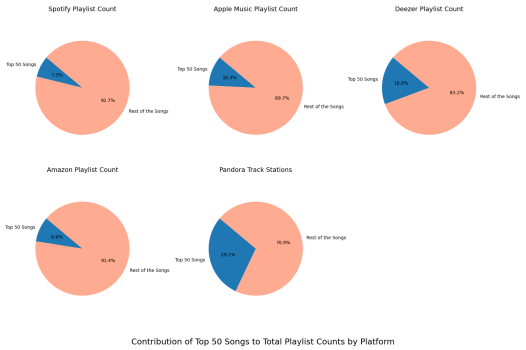


Fig 3: *Contribution of Top 50 songs to the Total Playlist Counts in Platforms*

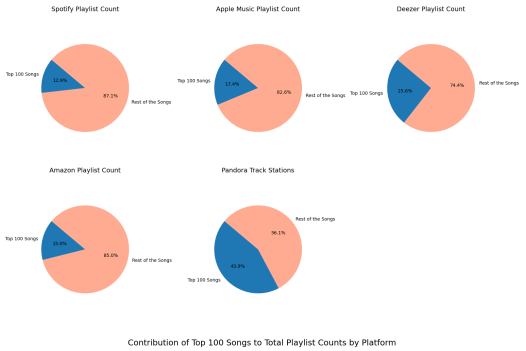


Fig 4: *Contribution of Top 100 songs to the Total Playlist Counts in Platforms*

Considerations in the Graph:

1) A horizontal bar plot was chosen for its aesthetic appeal, ease of readability. Pie charts were chosen as they best represented the contributions of different categories.

2) Only the platforms with a column representing the 'Playlist Count' have been plotted in Fig2, Fig3, Fig4
3) Text annotations have been used to display the exact values.
4) Marks used - Lines(in Fig2), Area(in Fig3, Fig4)
5) Channels used - Color(in Fig2, Fig4, Fig5 to represent categorical data i.e., platforms and top 100 songs), Area(in Fig2, Fig4, Fig5 to represent quantitative data i.e., Total Playlist Count, Percentages of playlist count)
6) The first plot uses a logarithmic scale on the x-axis to better visualize the wide range of values. The second plot uses a linear scale on the x-axis to highlight differences in values that may not be as apparent in the logarithmic scale plot.

3rd Hypothesis:

Songs released between 2019 and 2023 are likely to have contributed significantly to the total streams across music platforms. The COVID-19 pandemic, with its associated lockdowns and work-from-home mandates, led to a substantial increase in internet usage. This surge in online activity likely resulted in higher levels of music streaming and sharing, especially for songs released during this period, leading to their greater presence and circulation on the internet.

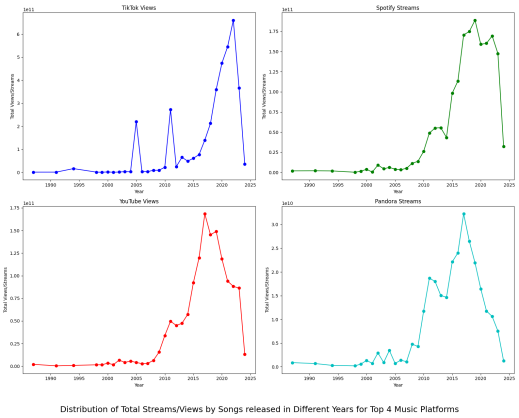


Fig 5: *Distribution of Total Streams/Views by Songs Released in Different Years*

Key Observations:

1) The top four platforms exhibit a notable surge in streams/views for songs released between 2019 and 2023, with a smaller increase observed from 2015 to 2019. This trend may be linked to the rise in internet users due to technological advancements or the increased ease of sharing and accessing content, leading to higher engagement during 2015-2023. The COVID-19 pandemic likely also played a role in driving peak streams/views during this period.
2) The slight spikes in views for songs from 2005 and 2011 on TikTok could be attributed to various factors. Cultural trends and nostalgia may have revived interest in these older tracks, while specific genres or styles from those years may be enjoying renewed popularity. Viral challenges or influencer-driven content may have also boosted their visibility. Additionally, significant marketing campaigns, collaborations, or changes in TikTok's algorithm or features could have elevated these songs' prominence. However, pinpointing the exact reasons for these trends remains challenging due to limited data.
3) Figure 6 reveals an increasing trend in the number of explicit songs released over the years, possibly reflecting a growing listener preference for explicit content. Figure 7 shows that among the top 100 performing songs on YouTube, Spotify, and TikTok, approximately 25-30% are explicit on YouTube, while Spotify and TikTok have higher proportions, with 45% and 50% of songs labeled explicit, respectively. This disparity may be due to YouTube's large child audience, necessitating stricter content filtering and monitoring. In contrast, Spotify and TikTok appear to have less rigorous content oversight.

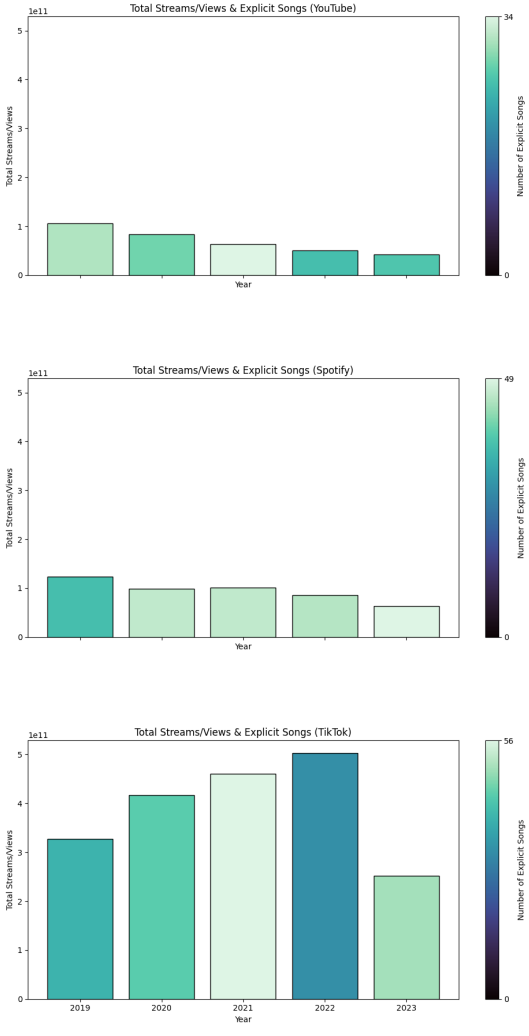


Fig 6: *Total Streams/Views for the Top 100 Songs Per Release Year (2019-2023) on the Top 3 Platforms (Based on Total*

Streams/Views)

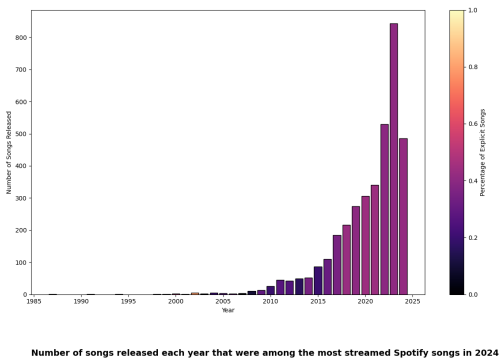


Fig 7: *Number of Songs Released in Different Years That Were Among the Most Streamed Spotify Songs in 2024*

Considerations in the Graph:

1) Bar plots have been used to visualize the different distributions in Fig6, Fig7.
2) Marks used - Lines(in Fig6, in Fig7)
3) Channels used -
   - Area: To represent the quantitative data, i.e., total streams/views of the top 100 songs and the total number of songs released each year, respectively, in Fig6 and Fig7.
   - Color: Represents the percentage of explicit songs and the number of explicit songs in the top 100 songs.
4) The linear color maps 'mako', 'magma' have been used for Fig6, Fig7 respectively.

4th Hypothesis:

As the number of posts on TikTok increases, the views and likes would also increase. On YouTube, as views increase, likes are expected to increase as well.

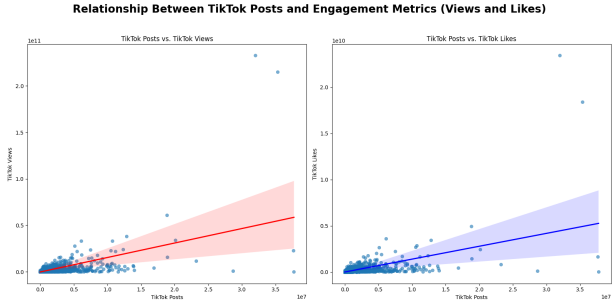


Fig 8: *Scatter Plot with a Regression Line Representing the Relationship Between TikTok Posts and Engagement Metrics (Views and Likes)*

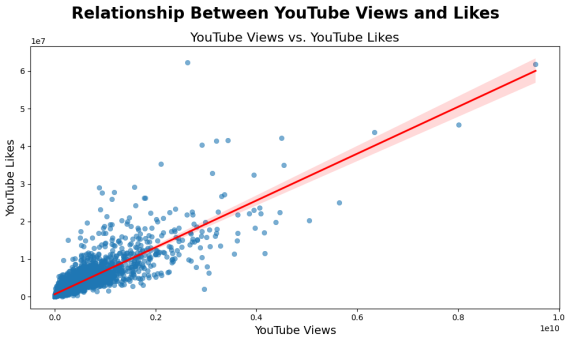


Fig 9: *Scatter Plot with a Regression Line Representing the Relationship Between YouTube Views and Likes*

Observations:

1) As the number of TikTok posts related to a song increases, both its views and likes tend to rise. This positive trend is evident from the graphs, suggesting that more posts generally lead to greater exposure, which drives higher engagement (views and likes). When a song is frequently featured in TikTok posts, it garners more attention, resulting in increased view counts and likes.
2) On YouTube, the data indicates that as the number of views increases, the number of likes also tends to rise. This correlation is apparent in the graphs, reinforcing the idea that higher view counts typically lead to more likes. This trend is likely due to the fact that as more people watch a video, it is more likely to receive positive feedback in the form of likes from viewers who enjoy the content.

Thus, the observed trends align with the hypotheses.

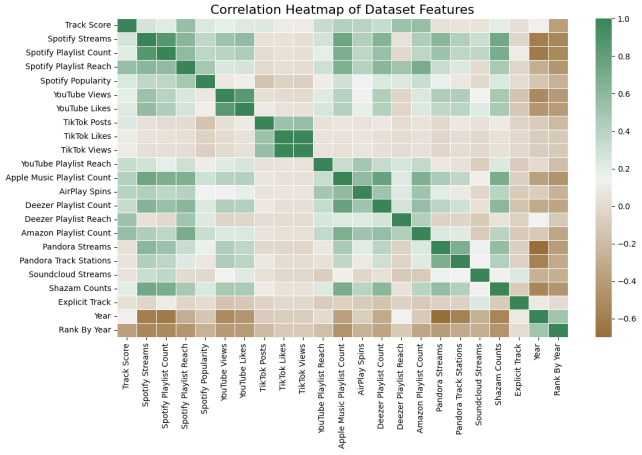


Fig 10: *Correlation heatmap showing the relationship between numeric features in the dataset.*

The correlation heatmap further reveals that both TikTok views and likes rise with an increasing number of TikTok posts, while YouTube likes also climb as YouTube views grow, supporting our hypothesis.

Considerations:

1) Scatter Plots Along With a Regression Line have been used to represent the relationships between the Views, Likes and Posts as shown in Fig8, Fig9 inorder to best visualize the general trend in the relationships. A correlation heatmap has been plotted to visualize the relationships between different numeric features in the dataset.
2) Marks used - Points(in Fig8, Fig9), cells(in Fig10)
3) Channels used -
   - X-position: To represent the number of TikTok Posts in Fig8, YouTube Views in Fig9
   - Y-position: To represent the number of TikTok Views, TikTok Likes, YouTube Likes in the individual graphs of Fig8, Fig9 respectively.
   - Color: To represent the numerical values of correlation in Fig10

5th Hypothesis:

The top-performing songs on these platforms are likely to contain a higher percentage of explicit content, with estimates around 50% for TikTok, 45% for Spotify, and 30% for YouTube. This trend is supported by the statistics showing an increase in explicit songs over the years, as illustrated in Fig. 6.

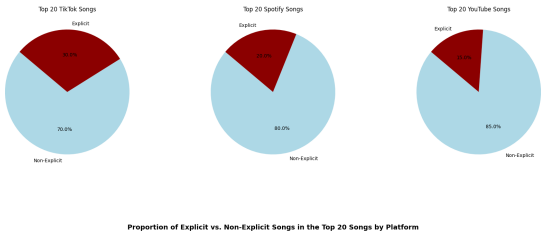


Fig 11: *Pie Charts Representing the Proportion of Explicit vs. Non-Explicit Songs in the Top 20 Songs by Platform*

We observe that TikTok has around 30% explicit content, while Spotify and YouTube have approximately 20% and 15%, respectively. This suggests that content filtering by these platforms plays a significant role in influencing the total streams/views. The relatively low percentage of explicit content on YouTube may indicate its efforts to filter content to appeal to a broader audience, including children.

Considerations:

1) Pie Chart has been used to represent the proportion of Explicit songs in the top 20 songs by platform as it best visualizes the same.
2) Marks used - Area(in Fig11)
3) Channels used - Area(to represent the percentages in Fig11), Color(to represent the categorical data in Fig11)

Year wise analysis of YouTube and Spotify streams

The number of Spotify streams and YouTube views has fluctuated over the years for songs released in different time periods. A combined bar plot showing the streams and views for songs from various years has been created to help analyze the popularity trends. Only the years after 2010 have been taken into consideration, as this was the time when both Spotify and YouTube had been around for a few years and had started gaining relevance.

6th Hypothesis:

YouTube is expected to lead for the better part of the 2010s decade, but Spotify will then exceed YouTube come the end of the same decade. Spotify is expected to build on this dominance for the remaining 5 years till 2024.

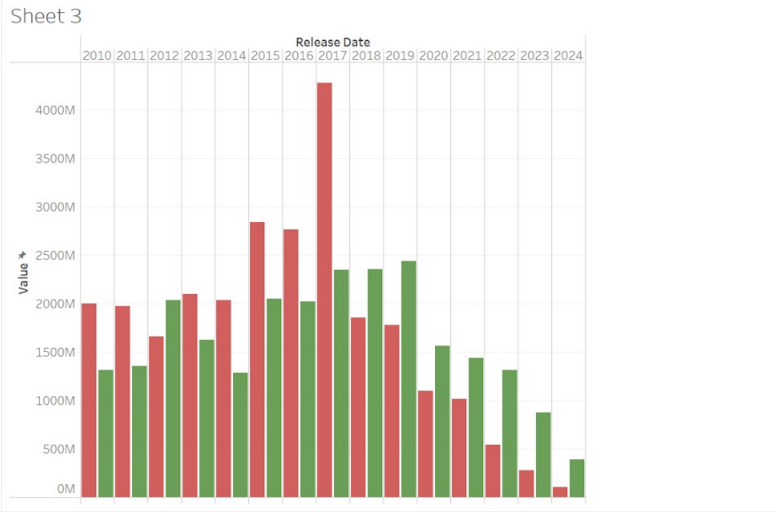


Fig 12: *Spotify vs YouTube - a yearly analysis (2010-2024)*

From the graph, we can infer the following insights:

1) There is a significant shift in YouTube views over the years. Views increase steadily until 2017, followed by a sharp decline up to 2024.
2) Spotify streams, on the other hand, show a more gradual change up to 2019, where they peak, then experience a rapid decline afterward.
3) Between 2010 and 2017, YouTube consistently had more views than Spotify streams. However, from 2018 to 2024, YouTube views fell below Spotify streams.
4) Several factors may contribute to these trends:
   - Spotify may have introduced more appealing offers for its premium subscription, attracting more listeners.
   - YouTube, being a convenient platform for music videos, may have seen a decline in views due to a decrease in the number of music videos released after 2017.

The hypothesis is found to be true for the most part. However, Spotify's views shrink continuously after 2020, where it was expected to increase. While it can be justified for the years 2024 and 2023 due to the lack of time these songs have had, for songs from 2020-2022, it appears that people do not listen to them much any longer.

Track Score based analysis for top TikTok songs

The below bubble chart visualizes the performance of popular TikTok songs on Spotify by showing two metrics: the sum of TikTok views and the track score on Spotify.

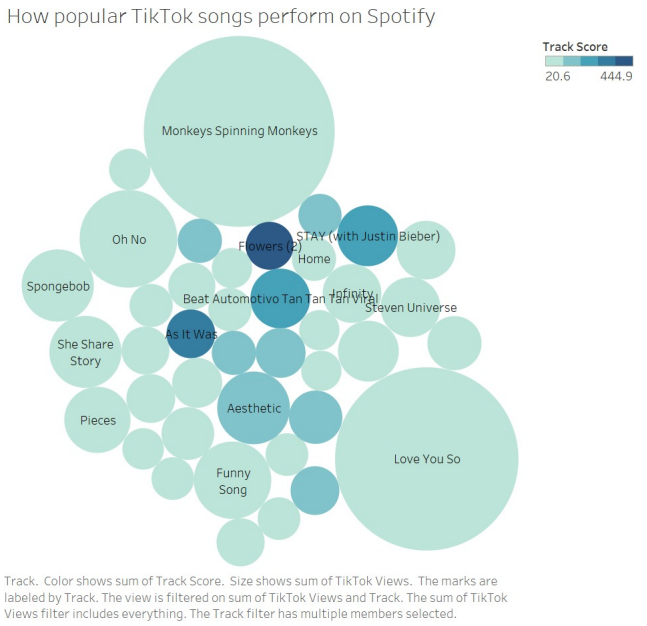


Fig 13: *Most listened to songs on TikTok and their Track Score*

Key Elements:

- **Bubble Size**: Represents the sum of TikTok views for each song. Larger bubbles indicate higher TikTok views.
- **Bubble Color**: Indicates the track score on Spotify, with darker shades representing higher track scores (ranging from 20.6 to 444.9).
- **Track Labels**: Each bubble is labeled with a song name, showing the corresponding song's popularity on both TikTok and Spotify.

Inferences:

1) **"Monkeys Spinning Monkeys"** has a large bubble, indicating it has a high number of TikTok views, but its lighter color suggests a relatively lower track score on Spotify.
2) **"Love You So"** has another large bubble, indicating significant TikTok views, but its color indicates a moderately lower Spotify track score.
3) **"STAY (with Justin Bieber)"** and **"Flowers"** are smaller bubbles, but their darker colors show they have a higher track score on Spotify, despite having fewer TikTok views compared to larger bubbles. Flowers also happens to be the sole representative of the highest Track Score Group, 356-444.
4) Songs like **"As It Was"** and **"Infinity"** show medium-sized bubbles, suggesting moderate TikTok views with varying track scores on Spotify.

This chart highlights how some songs that gain massive popularity on TikTok (large bubbles) don't always achieve equally high track scores on Spotify (lighter shades), while others with fewer views may perform better on the music platform.

To better understand the performance of Top TikTok songs on Spotify, here is a bar graph which groups Spotify songs by their "Track Score Group" (as split by the discrete colour map in the bubble diagram). Track Score Ranges on the X axis, and average views on the Y axis.

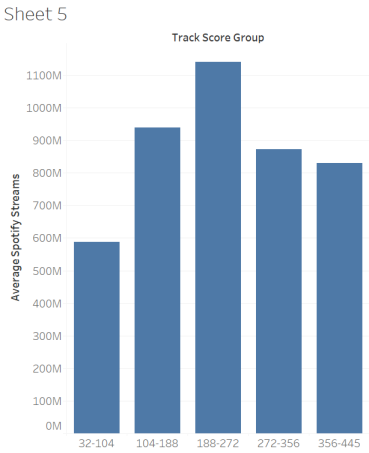


Fig 14: *How songs belong to every Track Score group (refer Fig 13) perform on Spotify*

Inferences:

1) The track score group "188-272" has the highest average Spotify streams, surpassing 1.1 billion streams.
2) Track scores between "104-188" and "272-356" also show significant average streams, nearing or exceeding 900 million.
3) The lowest average streams are associated with the "32-104" track score group, which accumulates less than 600 million streams. This also happens to be the group which dominates the top TikTok songs chart from earlier, indicating that the kind of songs people listen to on TikTok, probably as the background audio to short-format videos, are vastly different from what they like to hear on Spotify.
4) Tracks with mid-range scores tend to perform better in terms of average streams compared to tracks in the lowest or highest score ranges.

This graph helps visualize the relationship between track scores and their popularity on Spotify based on average stream counts.

B. Artist Analysis and Listening Trends

The tree map visualization below illustrates the popularity of the top artists based on their total Spotify streams. The color coding represents the average number of streams per song for each artist. While the colour of the boxes of each artist is not necessarily a representation of a balance between quality and quantity, it tells one about how many of an artist's songs top the charts - a darker colour would mean that an artist has few, but very big hits.

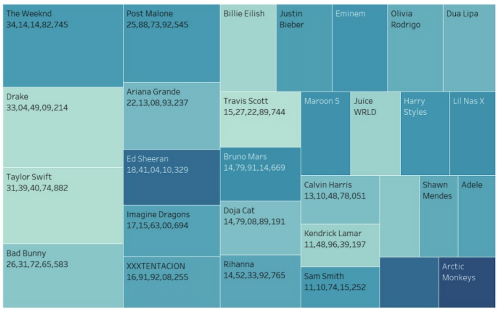


Fig 15: *Share of Spotify streams among the 30 most streamed artists*

We can see that the leading top 5 are *The Weekend*, *Drake*, *Taylor Swift*, *Bad Bunny*, and *Post Malone*

Explicit Songs Over the years

A collection of pie-charts to show the dominance and popularity of explicit songs over the years. The data clubs together 2 years, and takes the top 25 most popular songs from both years, and then looks for the percentage of explicit songs among the 50 songs put together.

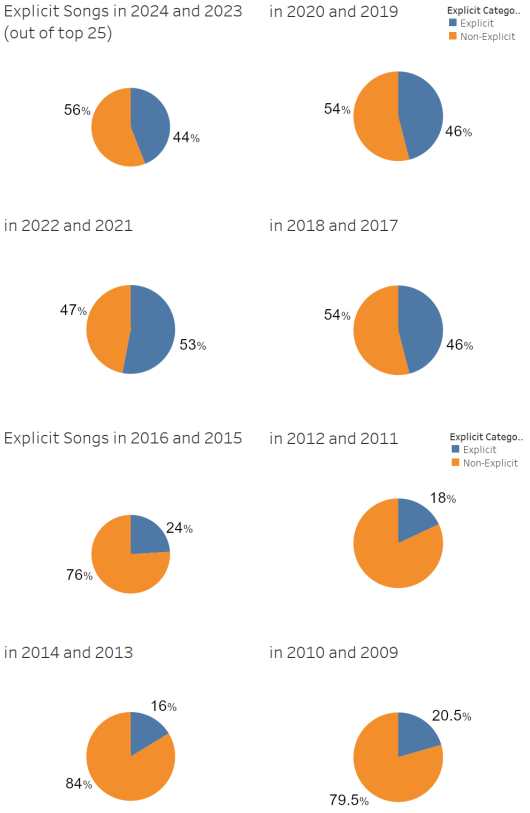


Fig 16 & 17: *Percentage of explicit songs among the top 50 songs, taken 2 years at a time for the last 16 years*

Inferences:

1) The year groups 2023-24, 2021-22, 2019-20 and 2017-18 all have an almost 50-50 split
2) The charts for the 8 years prior to that, however, show that there has been a big shift in trends. Moreover, this shift seems to be sudden and dramatic, as 2015-16 to 2017-18 is a change of 22%.

To find the reason for this trend, it was important to understand which kind of songs usually contain the most explicit content, and as per a research by Daniel Parris[1], 69% of explicit music usually belongs to the rap genre. Upon manual inspection of the dataset, one finds that majority of the rap music being listened to is from after 2018, possibly explaining the trend.

Artist Comebacks and Revivals: Below is a Gantt chart of the 5 artists with the longest interval between 2 successive songs that appear in the dataset.

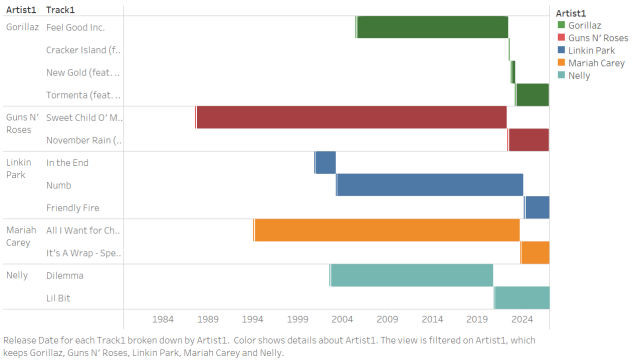


Fig 18: *Gantt chart showing longest gaps between successive appearances for artists*

Inferences:

1) Guns N' Roses leads the list with the highest gap between 2 successive songs, with a 35 year gap.
2) Mariah Carey follows, with a 29 year gap. Linkin Park is third, with a 21 year gap.

Upon reading from external sources, it was found that the reasons behind large gap can be due to a multitude of reasons. For *Guns N' Roses*, their 2022 release was a remaster and re-record of their 1991 song, November Rain[2] - it wasn't a new song altogether, but was one of the only releases by the band in an otherwise inactive era.

In the case of *Mariah Carey* and *Nelly*, the gap could only be explained by the limited success of their other releases during the gap - releases that failed to transcend through time and appeal to listeners today. It is likely that their new releases that make this dataset, too, might go forgotten soon and not find a lot of listeners few years from now.

The British band *Gorillaz* were on hiatus for much of the 2010s, justifying the release gap that coincides with that decade. They broke the hiatus in 2017 [3] and released their latest album in 2022.

For artists like *Linkin Park*, there are a combination of reasons. In their case, their releases from after 2003 fail to make the top charts in 2024, and they were inactive from 2017 till 2023.

C. *Impact of Playlist Curation*

Plot-1:

In analyzing the relationship between Spotify streams and playlist counts, we were curious to explore the connection between these two variables. Specifically, we hypothesized that for a given song, its number of streams could be directly influenced by the number of playlists it appears in. This idea stemmed from a natural assumption that as a song is added to more playlists, it gains exposure, leading to more listens. To test this hypothesis, we looked at data for the top 10 artists and their performance on Spotify, focusing on streams and playlist counts.

To better understand this relationship, we decided to plot a simple bar chart, where we ranked the top 10 artists based on their Spotify streams and Spotify playlist count. This allowed us to visually observe how these two metrics relate to each other.

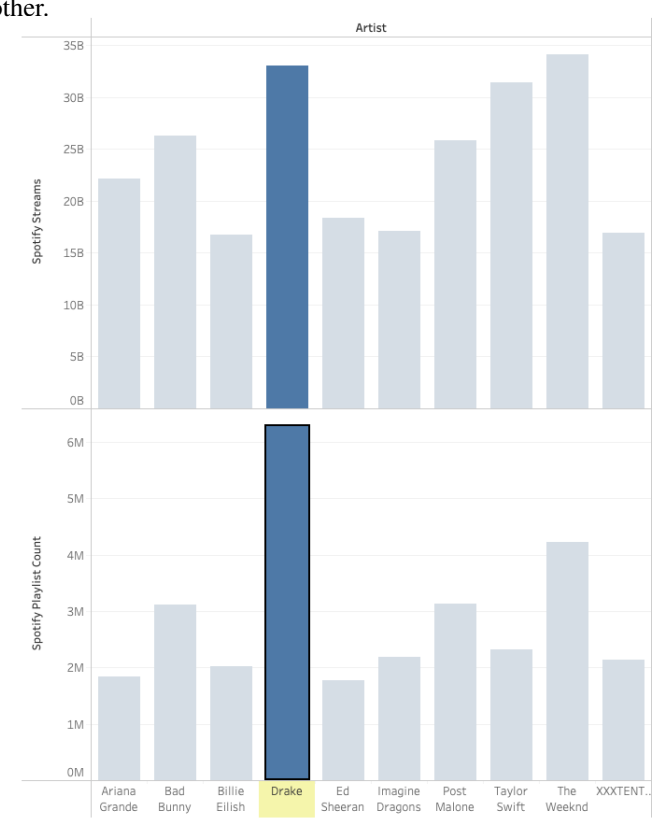


Fig 19

1) Artist vs Streams and Artist vs Spotify Playlist Count: The chart clearly shows how Drake, a standout artist, leads both categories. He has the highest number of streams and also dominates in terms of playlist count, which strongly supports our hypothesis. It implies that the more frequently an artist's music is included in playlists, the more streams it is likely to generate.

Looking at other artists, such as Bad Bunny, Ariana Grande, Billie Eilish, Imagine Dragons, and Post Malone, the trend continues to hold. For instance, Bad Bunny and Billie Eilish, though differing slightly in their number of playlists, still show a relationship between their streams and playlist placements. The patterns between Ariana Grande and Bad Bunny, as well as Imagine Dragons and Post Malone, further reinforce the idea that playlist count significantly impacts the stream count.

Thus, this visualization was essential in validating our hypothesis. This chart provided further clarity, making it clear that artists who are featured in more playlists tend to enjoy higher streams. Through this graphical representation, the connection became more tangible, with Drake emerging as the prime example of how playlist inclusion can boost an artist's overall performance on streaming platforms.

Plot-2:

We began with the hypothesis that a song's exposure across Spotify playlists could significantly influence its streaming numbers. To investigate this, we plotted two key metrics for the top ten artists by Spotify streams: the Spotify Playlist Reach, indicating the potential listener base through playlist inclusions, and the Spotify Streams themselves, representing actual listener engagement. The idea was to see if there's a direct correlation between the two, suggesting that being featured on more or high-reach playlists translates to higher streams.

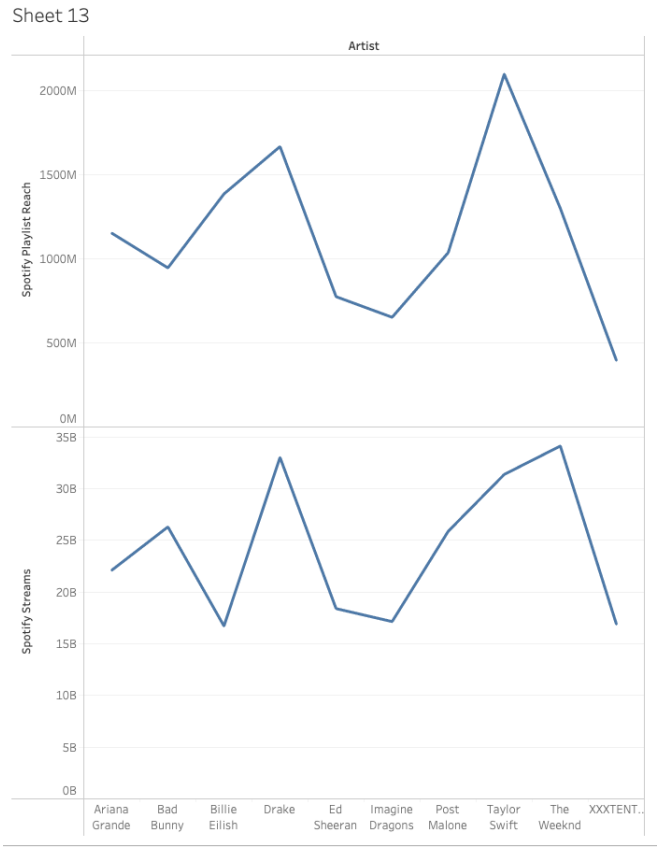


Fig 20

2) Artist vs Streams and Artist vs Spotify Playlist Reach: The resulting plot revealed a general trend where a higher Playlist Reach often coincides with increased streams, supporting our initial hypothesis. Most artists showed this pattern,

exemplified by sharp rises in both metrics simultaneously. However, interestingly, a few exceptions emerged, such as Ed Sheeran and Bad Bunny, who maintained high stream counts despite comparatively lower playlist reaches. This indicates that while playlist exposure boosts streaming numbers, certain artists can achieve high streams through strong fan bases or other means of engagement outside of playlists.

This analysis not only validated the significance of playlist reach in an artist's streaming success but also highlighted the complexities of music consumption, where factors beyond mere playlist presence—like artist popularity and fan loyalty—play crucial roles.

Plot-3:

Ideation:

The analysis of Spotify streams in relation to the explicitness of songs reveals an interesting trend. As shown in the top chart, which represents the number of explicit tracks by each artist, and the bottom chart, depicting the number of Spotify streams for the same artists, there is a notable correlation between these two variables. Artists with more explicit tracks tend to garner higher numbers of streams on Spotify, suggesting that explicit content plays a significant role in attracting listeners.

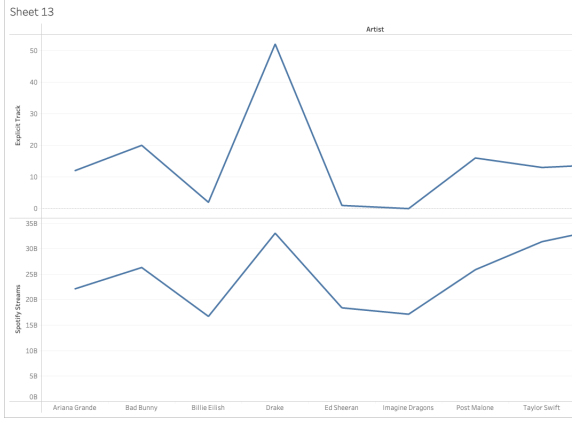


Fig 21

3) Artists vs Streams and artist vs sum of explicitness:

Inference

- Over the years, it appears that the demand for explicit songs on **Spotify** has risen more prominently compared to other streaming platforms. This could be attributed to changing audience preferences, where explicit content may resonate more with specific listener demographics on Spotify. The similarity in the trends of both charts implies that artists who produce more explicit content have been able to drive more engagement and attract larger listener bases.
- This correlation suggests that explicit content has become a strong factor in determining popularity on Spotify, which aligns with broader industry trends where explicit music often garners attention due to its rawness, relatability, or boldness. As Spotify continues to evolve as a dominant music platform, the importance of understanding listener preferences, particularly in relation to the explicitness of songs, becomes even more crucial for artists aiming to maximize their reach and popularity.

V. CONTRIBUTIONS:

The Tasks for visualizing the dataset were discussed in a meeting and all of us came up with our respective tasks.

1) Siddhart Menon: Performed Data Cleaning and Pre-Processing, added the column named 'Rank By Year', and visualized Task 2 using the following plots in Tableau in addition to the 6th Hypothesis
   - Figure 12 - Spotify vs YouTube
   - Figure 13, 14 - TikTok music analysis
   - Figure 15 - Artist Popularity Treemap
   - Figure 16, 17 - Song explicitness over the years
   - Figure 18 - artist comebacks and revivals
2) Shreyank Gopalakrishna Bhat: Helped add inferences in various plots and Visualized Task 3 using the following plots in Tableau
   - Fig 19, 20, 21 - Impact on artist curation and Stream analysis of Spotify based on factors such as explicit song count, spotify playlist count and spotify playlist reach.
3) Vrajnandak Nangunoori: Visualized Task 3 through the first 5 Hypothesis using the following plots in Python(code available in file 'Task1_Python_Code.py')
   - Figure 1 - Accumulated Streams/Views in Platforms
   - Figure 2 - Total Playlist Counts in Platforms
   - Figure 3, 4 - Contribution of top 50,100 songs to Playlist Counts in Platforms
   - Figure 5 - Distribution of Total Streams/Views by Release Years of songs.
   - Figure 6 - Total Streams/Views of Top 100 Songs by release year
   - Figure 7 - Number of Songs Released in Different Years
   - Figure 8, 9 - Relationships of Engagement Metrics in TikTok, YouTube
   - Figure 10 - Correlation Heatmap on Numerical Features
   - Figure 11 - Explicitness of Top 20 songs

VI. REFERENCES:

[1] The Rise of Explicit Music: A Statistical Analysis.

[2] Things That Are Different About New Version of Guns N' Roses' 'November Rain'

[3] Gorillaz is Back and It Feels Good