The BRITISH UNIVERSITY IN EGYPT

Introduction to Data Science

24CSAI01I

Streaming Platforms Dataset

Group Number: 20

Prepared by:

Ahmed 236664
Hassan 229591
Ahmed 236322
Hazem 232504
Abdelrahman 236882

Submitted to:
Dr. Nahla Barakat
TA Yasmin Sherif

Description:

This dataset contains information on the most streamed songs on Spotify in 2024, including details about each track's performance across multiple streaming platforms. It includes 29 columns, covering metadata such as the track name, album, artist, release date, and ISRC code, along with popularity metrics from major platforms like Spotify, YouTube, TikTok, Apple Music, Deezer, Amazon Music, Pandora, and SoundCloud. Key performance indicators include Spotify streams, playlist counts, and reach, as well as YouTube views and likes, TikTok posts and engagement, Airplay and SiriusXM spins, and Shazam counts. Popularity scores are available for platforms like Spotify and TIDAL, while an explicit content indicator is also present.

Dataset:

https://www.kaggle.com/datasets/nelgiriyewithana/most-streamed-spotify-songs-2024

General Information:

- Number of Rows: 4,600 (total tracks in the dataset)
- Number of Columns: 29 (various metrics related to streaming and popularity)

Column Names & Description:

1. Track (object) – Name of the song.
2. Album Name (object) – The album the track belongs to.
3. Artist (object) – The performing artist.
4. Release Date (object) – When the track was released.
5. ISRC (object) – International Standard Recording Code for the track.
6. All Time Rank (object) – The track's ranking based on streaming metrics.
7. Track Score (float64)– A calculated score that represents the track's popularity.

8. Spotify Streams (object) – Total number of streams on Spotify.
9. Spotify Playlist Count (object) – Number of playlists featuring the track on Spotify.
10. Spotify Playlist Reach (object) – Estimated audience reach via playlists.
11. Spotify Popularity (float64) – A numeric value reflecting popularity on Spotify.
12. YouTube Views (object) – Total views on YouTube.
13. YouTube Likes (object) – Total likes on YouTube.
14. TikTok Posts (object) – Number of TikTok videos using the track.
15. TikTok Likes (object) – Total likes on TikTok for videos using the track.
16. TikTok Views (object) – Total views for TikTok videos featuring the track.
17. YouTube Playlist Reach (object) – Estimated reach via YouTube playlists.
18. Apple Music Playlist Count (float64) – Number of Apple Music playlists featuring the track.
19. Airplay Spins (object) – Number of times the track has been played via radio.
20. SiriusXM Spins (object) – Number of times played on SiriusXM radio.
21. Deezer Playlist Count (float64) – Number of Deezer playlists featuring the track.
22. Deezer Playlist Reach (object) – Estimated reach via Deezer playlists.
23. Amazon Playlist Count (float64) – Number of Amazon Music playlists featuring the track.
24. Pandora Streams (object) – Total streams on Pandora.
25. Pandora Track Stations (object) – Number of Pandora stations featuring the track.
26. SoundCloud Streams (object) – Total plays on SoundCloud.
27. Shazam Counts (object) – Number of times the track was identified on Shazam.

28. TIDAL Popularity (float64) – Popularity score on TIDAL.

29. Explicit Track (int64) – Binary (1 = explicit, 0 = clean).

Data Types:

- Most columns are initially object (string) types, but some numerical metrics (e.g., Track Score, Spotify Popularity, Apple Music Playlist Count, etc.) are floats or integers.
- Some numerical values (like Spotify Streams and YouTube Views) are stored as strings with commas, meaning they may need conversion for analysis.

Research Questions:

1. What factors contribute the most to a song's all-time rank?
2. How do Spotify streams compare to YouTube views for the top-ranked songs?
3. Do explicit songs tend to perform better across streaming platforms than clean songs?
4. How does TikTok virality influence Spotify streaming numbers?
5. Do songs with a higher number of playlist features have more streams?
6. Is there a correlation between radio stations spins and number of streams on different platforms?
7. Do explicit tracks have higher TikTok engagement but lower radio spins?
8. How does a song's release year influence its cumulative streaming performance across different platforms?
9. Do tracks in Spotify featuring collaborations achieve higher streaming numbers or playlist reach than solo tracks?
10. Do Shazam counts indicate a song's future success on streaming platforms?

Data Cleaning Process

Before answering the research questions, the dataset was cleaned to ensure the analysis was based on consistent, accurate, and usable information. The following steps were taken:

1. Data Import and Inspection

The dataset was imported using Python's pandas library. Initial checks were done to examine the structure, column types, and the presence of missing or inconsistent values.

2. Data Type Conversion

Several columns, especially those containing numeric values stored as strings (e.g., with commas), were converted to proper numeric formats. The Release Date column was converted to datetime. Key fields cleaned included Spotify streams, playlist counts, YouTube views, and others.

3. Handling Missing Values

Different strategies were used based on the context of each column:

Rows with critical missing data, such as missing Artist or Spotify Streams, were removed.

Columns where missing implied absence (e.g., no playlist feature, no TikTok posts) were filled with 0.

For popularity and engagement metrics, missing values were filled based on related indicators. For example, if a track had high Spotify streams but missing popularity, the median popularity score was assigned. For low-performing tracks, missing values were set to 0.

4. Standardizing Formats

Text formatting inconsistencies were resolved, and all numeric columns were cleaned (e.g., removing commas). All columns were checked and converted to the appropriate types.

5. Dropping Irrelevant Columns

The column TIDAL Popularity was removed due to excessive missing values and lack of analytical value.

6. Final Integrity Check

After all adjustments, the dataset had no null values. All missing entries were either dropped or filled logically to ensure consistency without distorting the data.

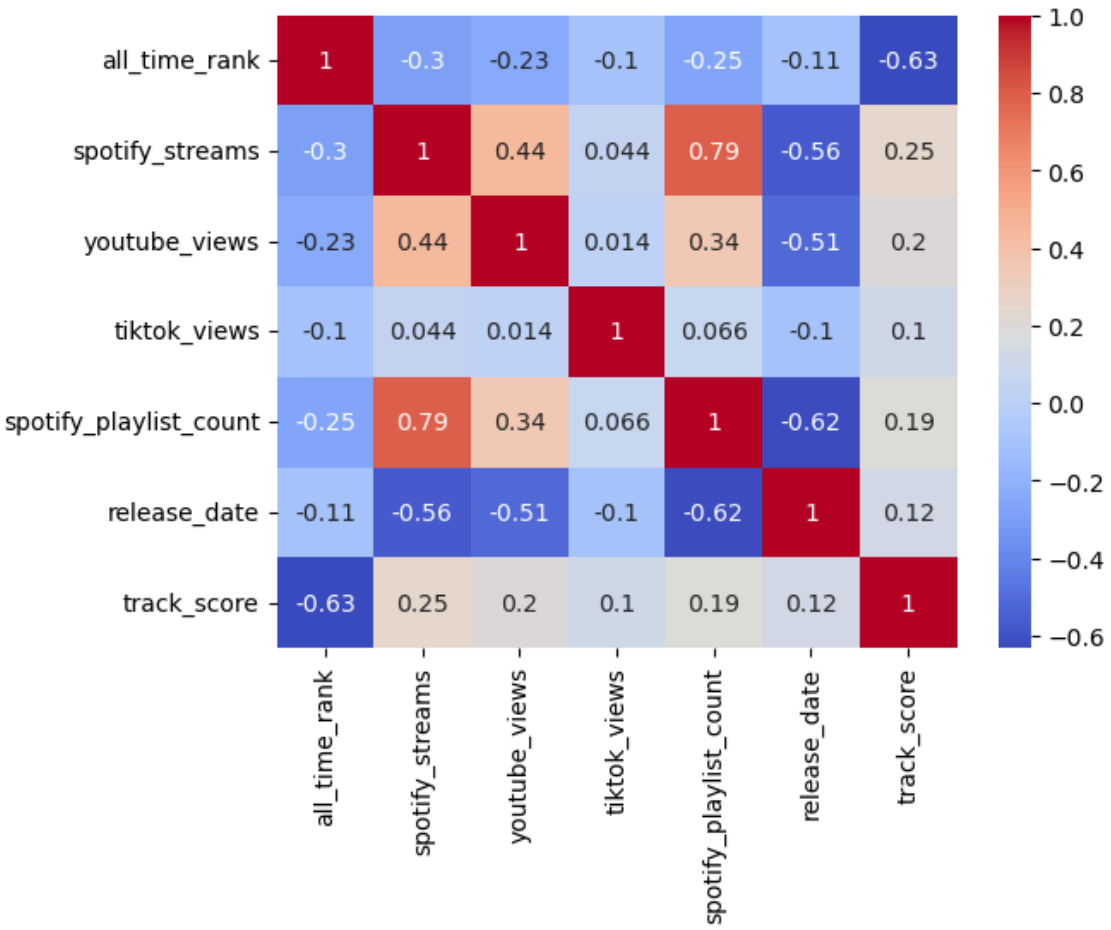
How We Performed Our Analysis?

We approached our data science project by conducting a step-by-step analysis using both the cleaned dataset and new insights from visual exploration and correlation analysis. Below is a breakdown of how we tackled the first set of questions.

Section 1

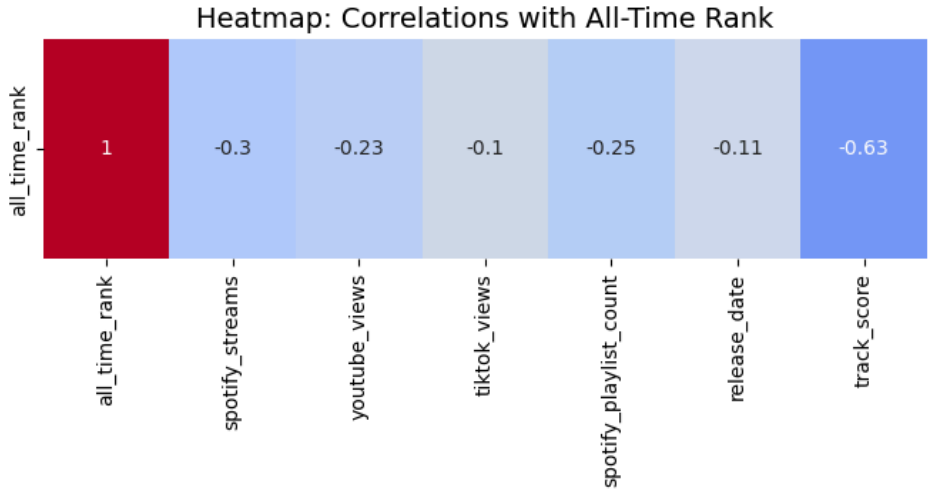
Question 1: What factors contribute the most to a song's all-time rank?

To find out which features affect a song's all-time rank, we analyzed track score, Spotify streams, playlist count, YouTube views, TikTok views, and release date. We used a correlation matrix to measure how each feature relates to rank, and visualized the results using a heatmap for easy comparison.



Correlation Analysis

Track score had the strongest negative correlation (−0.63), meaning higher-scoring songs tend to rank better. Spotify streams (−0.30), playlist count (−0.25), and YouTube views (−0.23) also showed some connection to better ranks, though not as strong. Release date (−0.11) and TikTok views (−0.10) had very weak correlations, suggesting little influence.



Interpretation

Track score is the most important factor in determining a song's rank. Spotify streams and playlist count help to some extent. YouTube views have a smaller role, and TikTok engagement appears to have almost no impact.

Conclusion

Track score has the strongest relationship with all-time rank. Spotify-related metrics contribute moderately, while TikTok and release date show minimal effect. The heatmap helped us quickly identify which features matter most.

Question 2: How do Spotify streams compare to YouTube views for the top-ranked songs?

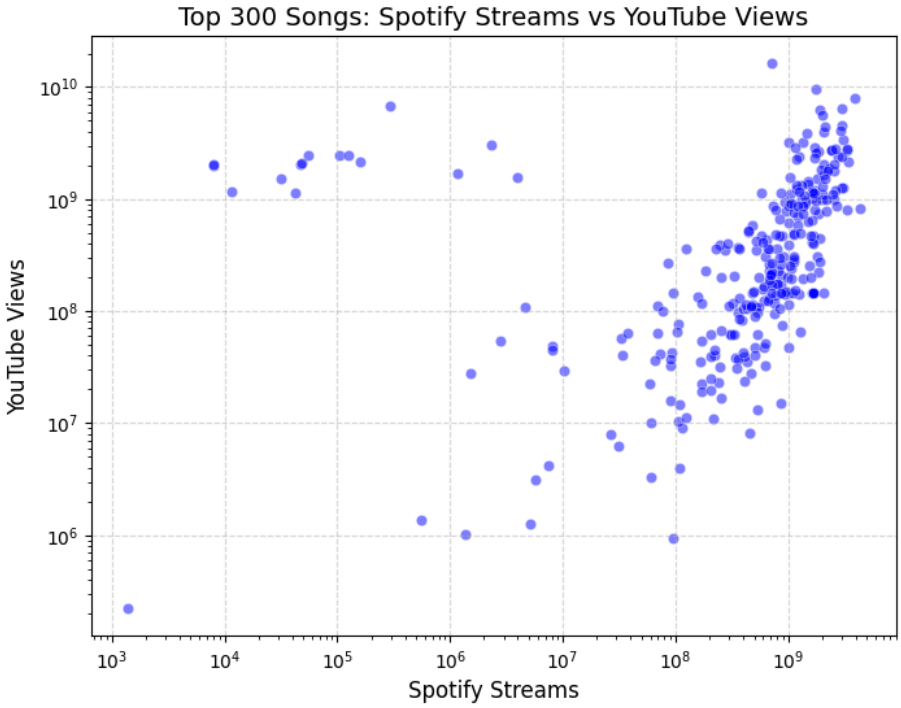
We wanted to understand whether top-performing songs show consistent popularity across Spotify and YouTube, or if some tracks perform better on one platform than the other.

Method and Visualization

We selected the top 300 ranked songs and plotted a scatter plot comparing Spotify streams and YouTube views. A log-log scale was used to better represent proportional differences and spread across a wide range of values. This visualization helped us identify both trends and outliers.

Analysis

Most songs showed a strong upward trend, meaning high-performing songs on Spotify usually also have high view counts on YouTube. However, about 12% of the songs appeared as outliers, performing very well on one platform but only moderately on the other. These were mainly mega-hits with over 1 billion streams or views.



Interpretation

The data suggests that cross-platform success is common among top-ranked songs, but some tracks gain traction on only one platform likely due to content format, promotion, or audience behavior.

Conclusion

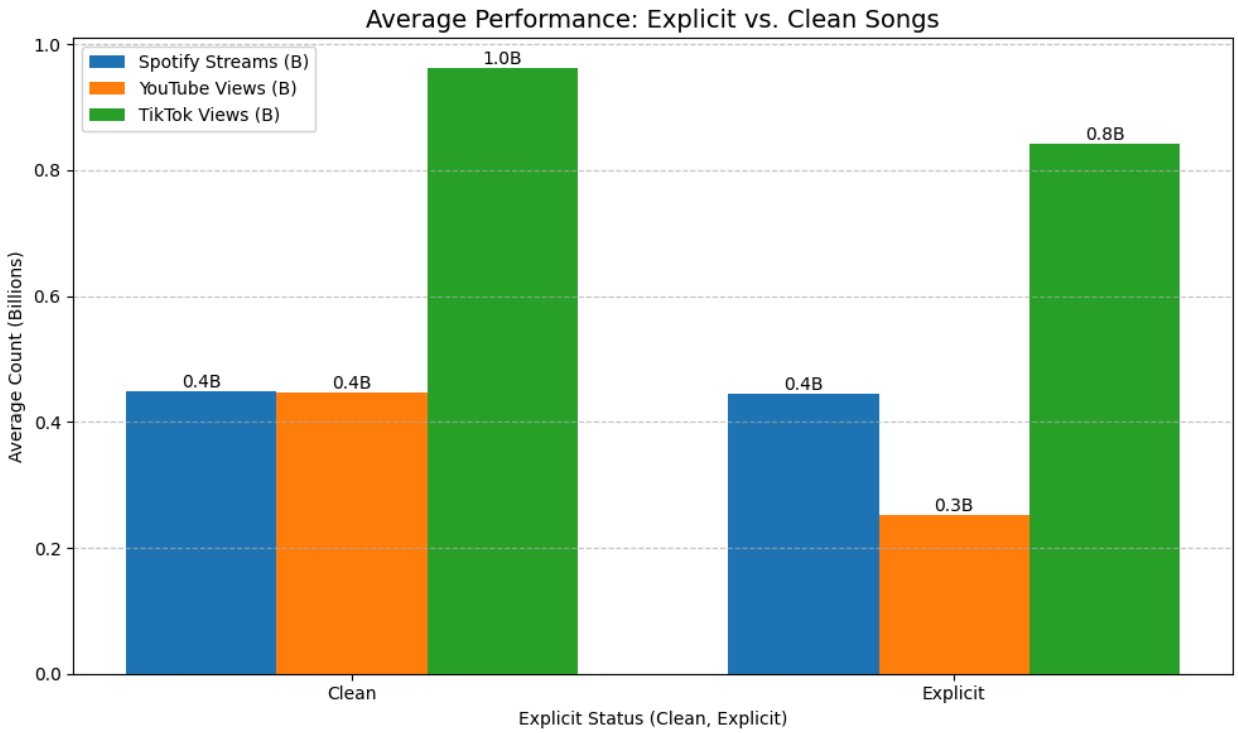
Spotify streams and YouTube views are strongly aligned for top songs, but a small set of outliers shows platform-specific popularity. This insight can help tailor marketing strategies to maximize reach on both services.

Question 3: Do explicit songs tend to perform better across streaming platforms than clean songs?

This question explores whether explicit tracks outperform clean ones across Spotify, YouTube, and TikTok. We compared the average number of streams and views for each platform by track type.

Method and Visualization

We calculated the mean values of Spotify streams, YouTube views, and TikTok views for both explicit and clean tracks. These averages were displayed in a grouped bar chart, which allows direct comparison between the two track types across all three platforms. Bar charts were used because they clearly highlight differences in numeric values between categories.



Analysis

- Spotify: Both explicit and clean songs averaged 0.4 billion streams, showing no difference in performance.
- YouTube: Clean songs averaged slightly more views (0.4B) than explicit ones (0.3B).
- TikTok: Clean songs again outperformed, averaging 1.0 billion views compared to 0.8 billion for explicit tracks.

Interpretation

Explicit songs do not consistently perform better. In fact, clean songs had higher averages on YouTube and TikTok, possibly due to platform moderation policies or audience preferences. Spotify was the only platform where both types performed equally, suggesting fewer restrictions or more balanced audience behavior.

Conclusion

Clean songs tend to perform better overall across platforms, especially on TikTok and YouTube. Explicit songs perform equally well on Spotify but may face limits on other platforms due to content guidelines or user demographics.

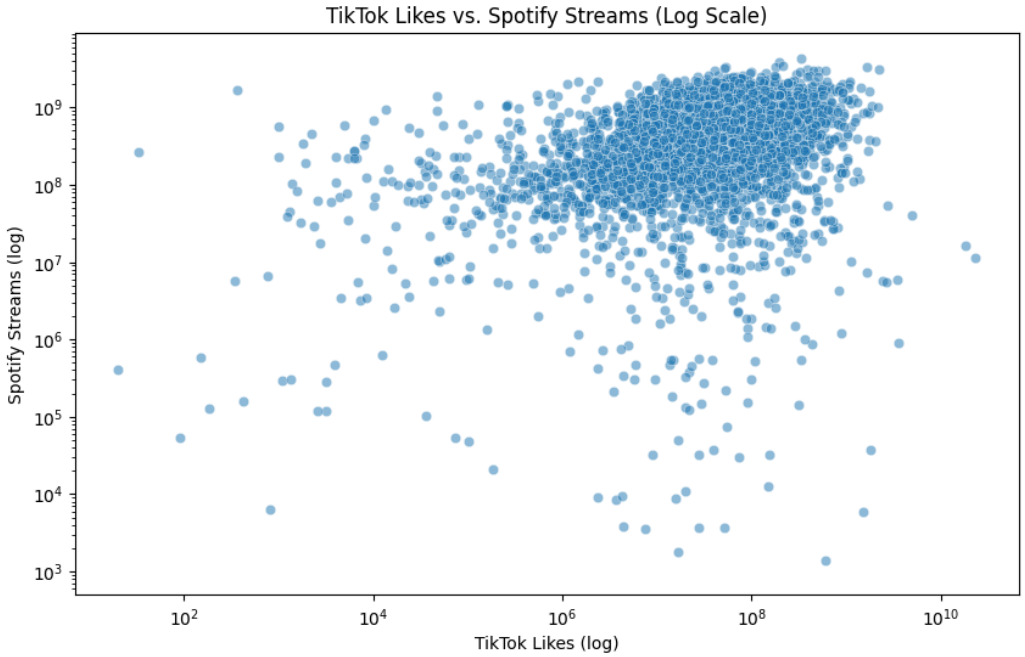
Question 4: How does TikTok virality influence Spotify streaming numbers?

This question explores whether songs that perform well on TikTok are more likely to receive higher stream counts on Spotify.

Method and Visualizations

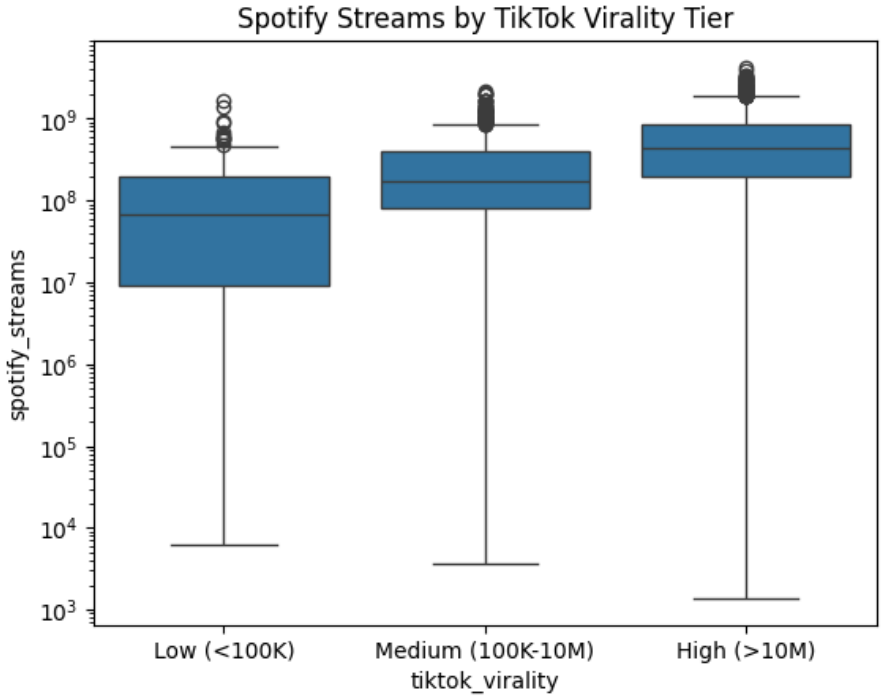
1. **Scatter Plot**

   Plotted TikTok likes against Spotify streams to observe overall trends between engagement and streaming performance.



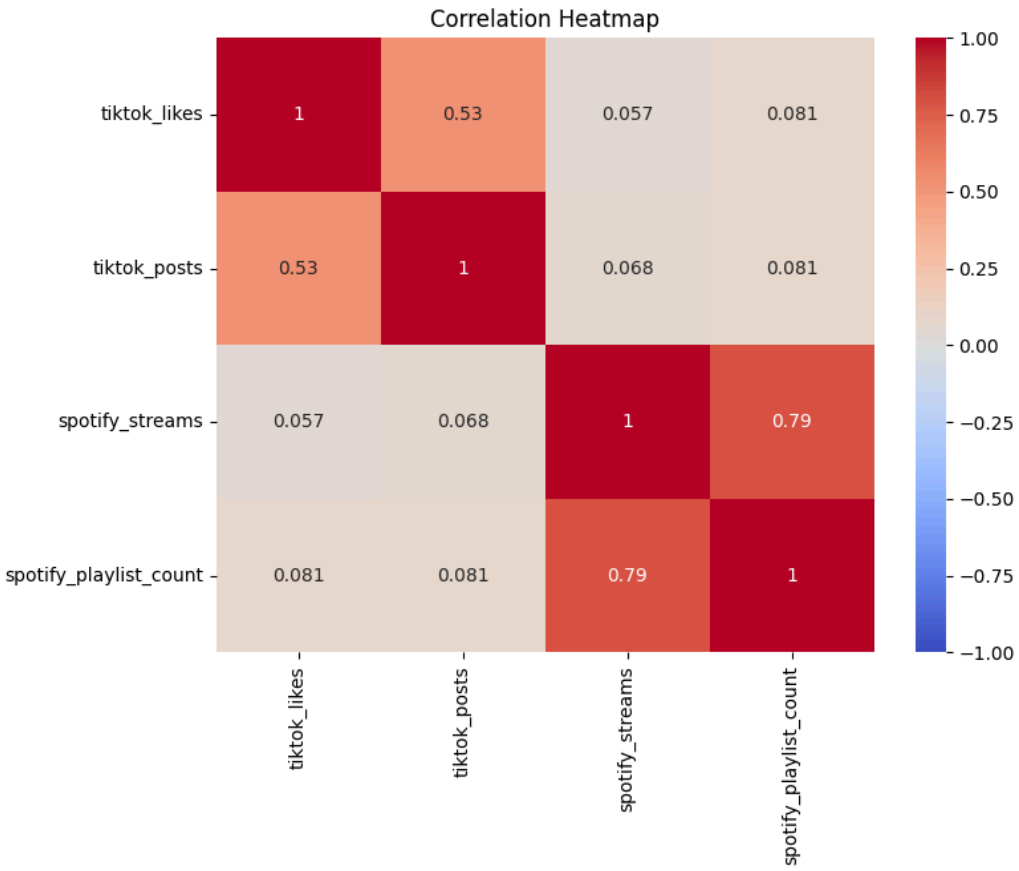
2. **Box Plot**

Compared Spotify streams across three TikTok virality tiers: low, medium, and high.



3. **Correlation Heatmap**

Displayed the strength of relationships between TikTok activity and Spotify performance.



Analysis

- The scatter plot showed a general upward trend, with higher TikTok likes often linked to higher Spotify streams.
- The box plot revealed that high-virality songs have 3–5 times more median streams than low-virality songs, though there was overlap between tiers.

- The correlation heatmap showed:
    - Spotify streams vs. playlist count: 0.79 (strong)
    - TikTok likes vs. TikTok posts: 0.53 (moderate)
    - TikTok likes vs. Spotify streams: 0.057 (weak)
    - TikTok posts vs. Spotify streams: 0.068 (weak)

Interpretation

Visual patterns suggest that TikTok virality is linked to streaming performance, but the correlation values indicate that TikTok activity alone does not strongly predict Spotify streams. The effect may be indirect; viral exposure on TikTok could increase a song's chances of being added to playlists, which have a much stronger impact.

Conclusion

TikTok virality appears to support Spotify success indirectly, likely through improved visibility and playlist placement. While viral songs often perform better, playlist count remains the more reliable driver of streaming numbers.

Question 5: Do songs with more playlist features have more streams?

This question looks at whether the number of playlist placements a song has affects how well it performs on Spotify.

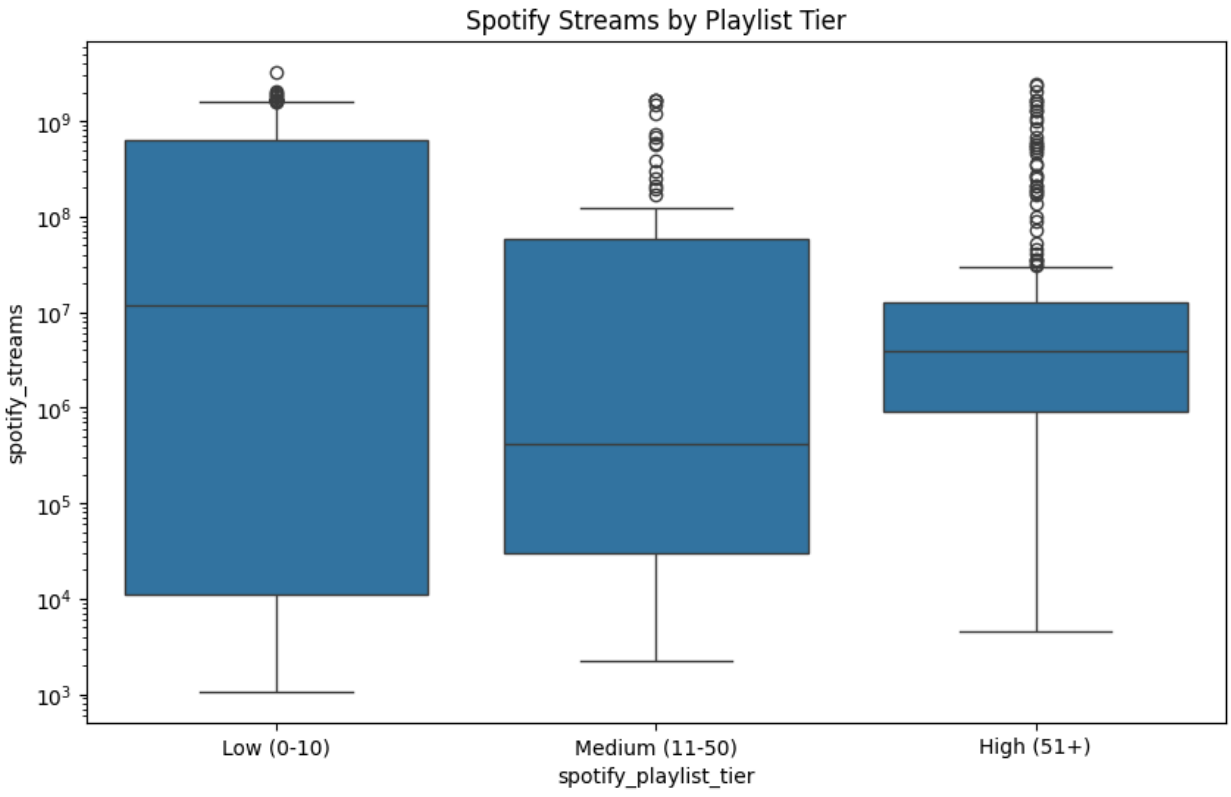
Method and Visualizations

1. **Box Plot**

   Compared Spotify streams across three playlist tiers based on playlist count:

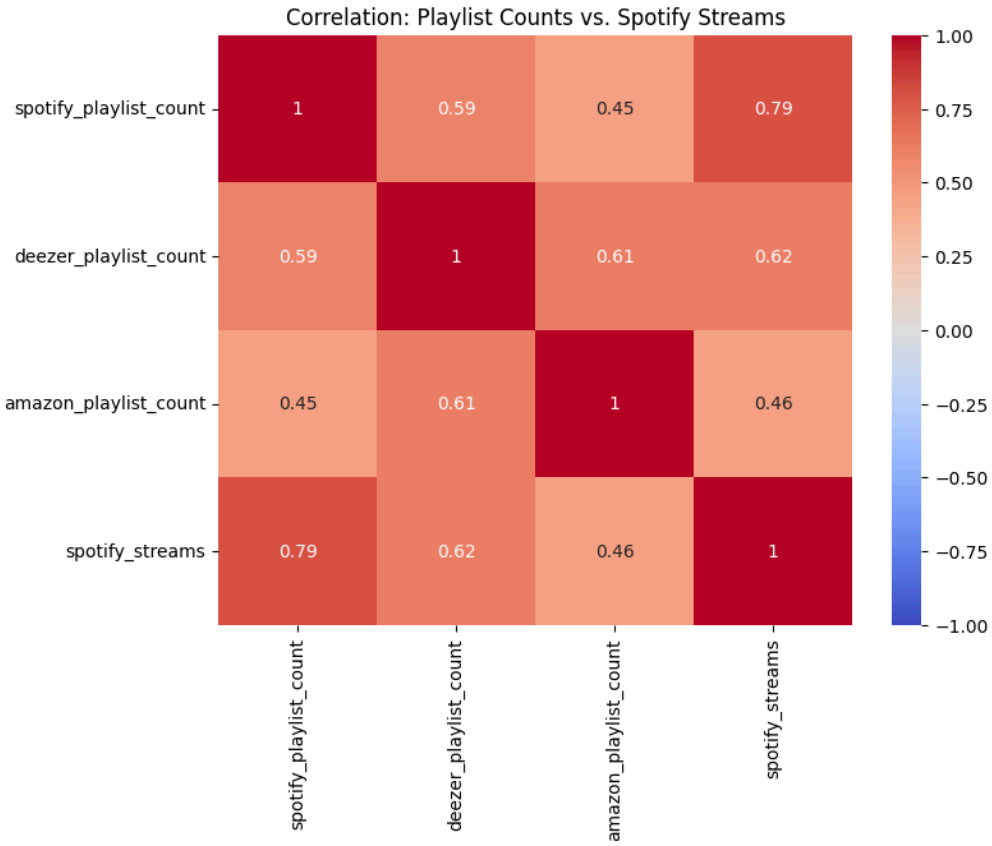
   - Low (0–10)
   - Medium (11–50)
   - High (51+)

This helped visualize how streaming performance changes across different levels of playlist exposure.



2. **Correlation Heatmap**

Displayed how playlist counts on Spotify, Deezer, and Amazon relate to Spotify streams.



Analysis

- **Box Plot**:
  - Songs in the High playlist tier averaged over 500 million streams.
  - Songs in the Low tier averaged around 50 million.
  - This showed a clear difference in performance based on playlist exposure.
- **Correlation Heatmap**:
  - Spotify playlist count vs. Spotify streams: 0.79 (strong)
  - Deezer playlist count vs. Spotify streams: 0.62 (moderate)
  - Amazon playlist count vs. Spotify streams: 0.46 (weak)

Interpretation

More playlist placements, especially on Spotify, are strongly associated with higher streaming numbers. Spotify shows the highest correlation, suggesting its playlists have the most influence. Deezer and Amazon have some impact but to a lesser degree.

Conclusion

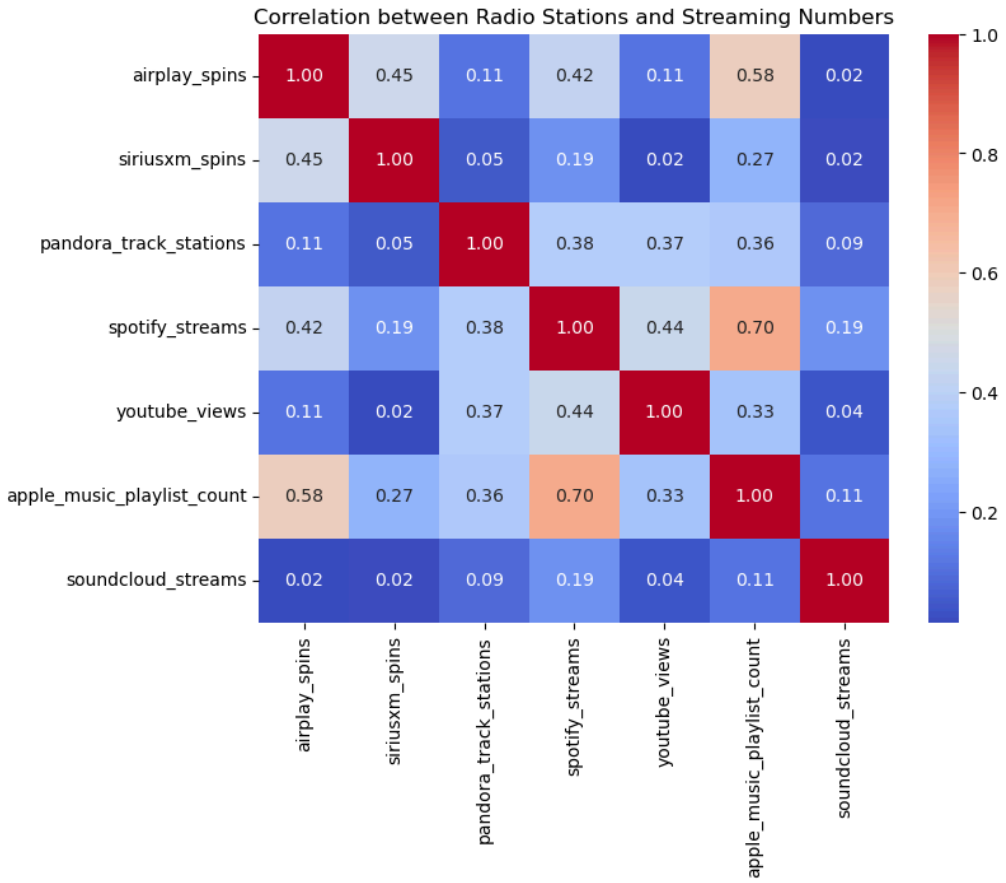
Songs featured on more playlists tend to receive significantly more streams, particularly when added to Spotify playlists. Playlist strategy is key to driving streaming performance, with Spotify offering the strongest return.

Question 6: Is there a correlation between radio station spins and number of streams on different platforms?

This question investigates whether traditional radio exposure—through stations like AirPlay, SiriusXM, and Pandora has a measurable impact on how a song performs across major streaming platforms.

Method and Visualization

We used a correlation heatmap to assess the strength of relationships between radio station plays and streaming metrics on Spotify, YouTube, Apple Music, and SoundCloud. This visualization helped us compare all variables at once and identify patterns in how radio activity relates to digital performance.



Analysis

- **Spotify Streams**
    - AirPlay: 0.42 (moderate)
    - Pandora: 0.38 (moderate)
    - SiriusXM: 0.19 (weak)

- **YouTube Views**
    - Pandora: 0.37 (moderate)
    - AirPlay: 0.11 (weak)
    - SiriusXM: 0.02 (very weak)

- **Apple Music Playlist Count**
    - AirPlay: 0.58 (moderate)
    - Pandora: 0.36 (moderate
    - SiriusXM: 0.27 (weak)

- **SoundCloud Streams**
    - All correlations were weak (≤ 0.09), indicating very little relationship with radio spins.

Interpretation

There's a consistent pattern where Pandora and AirPlay spins show a moderate positive relationship with performance on major platforms like Spotify and Apple Music. SiriusXM has a weaker influence, and SoundCloud appears mostly unaffected by radio activity.

Conclusion

Radio spins are positively associated with digital streaming numbers, especially for Spotify and Apple Music. Pandora and AirPlay have the strongest impact, while SiriusXM and SoundCloud show much weaker links.

Question 7: Do explicit tracks have higher TikTok engagement but lower radio spins?

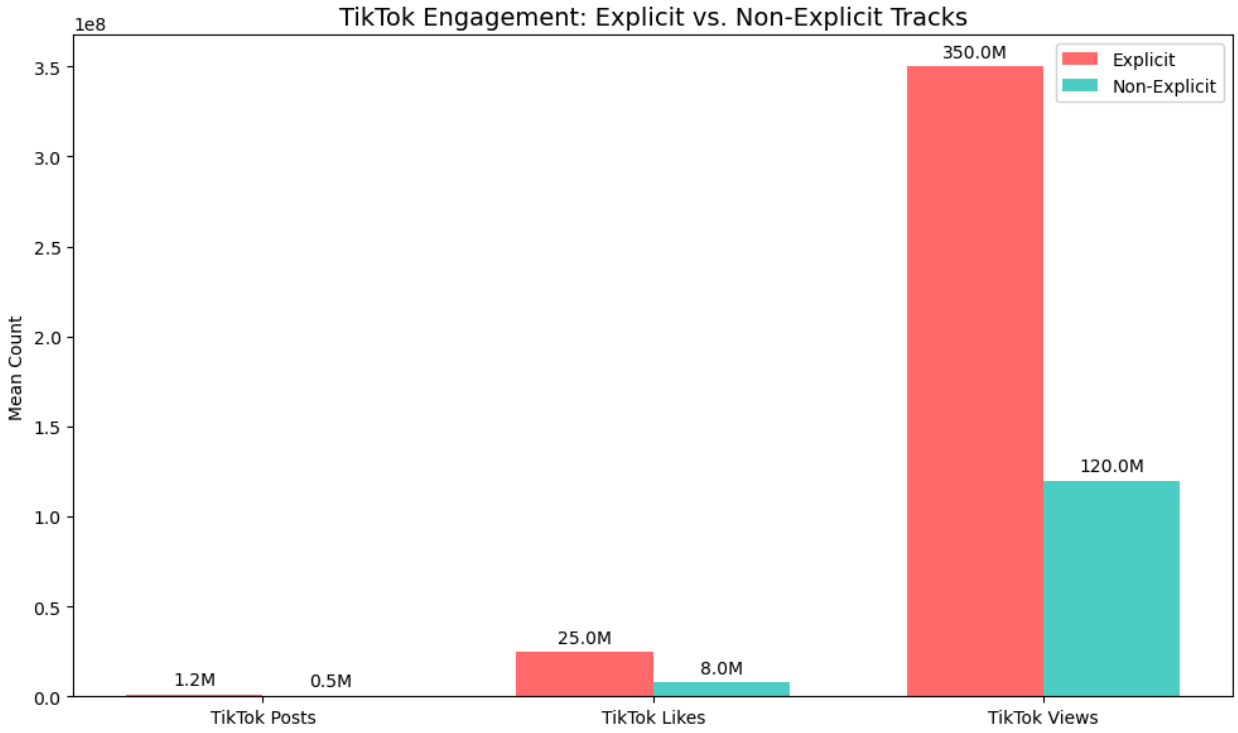
We compared TikTok engagement and radio spins for explicit and non-explicit tracks to see if there's a difference in how each is received on different platforms.

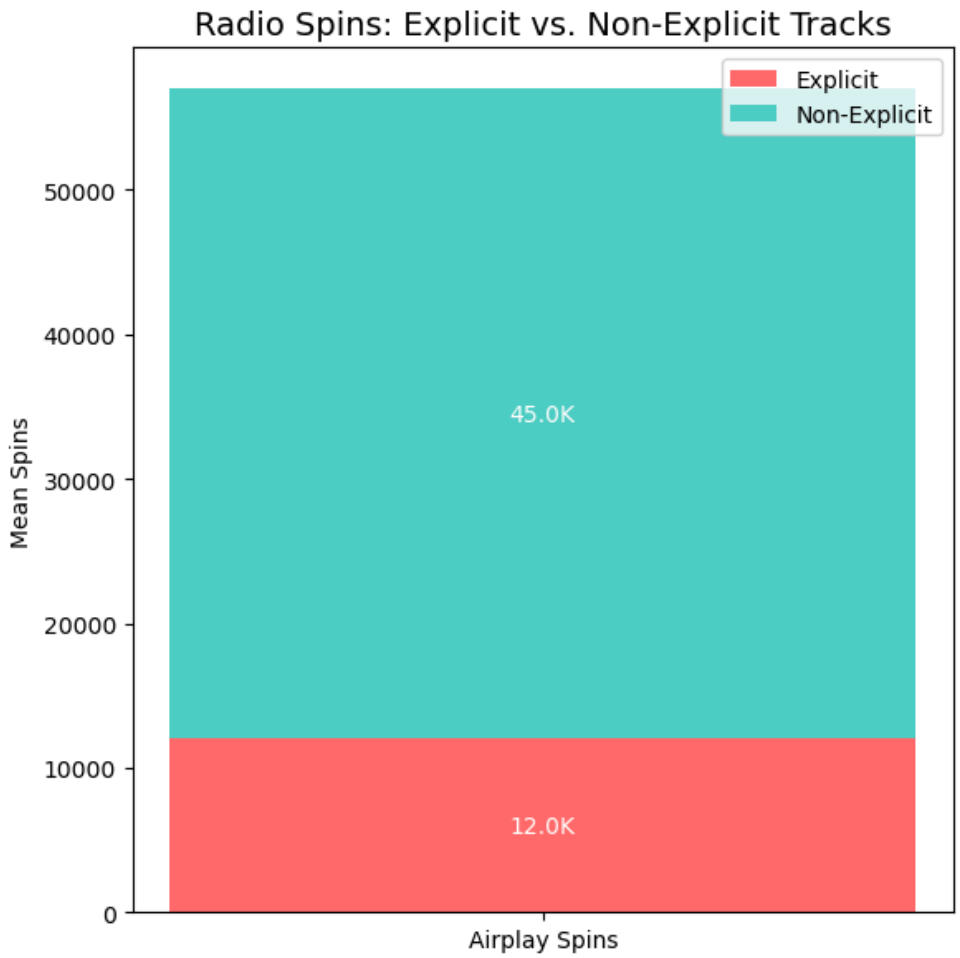
Method and Visualization

To compare engagement, we created two bar charts:

- A grouped bar chart showing the average number of TikTok posts, likes, and views for explicit vs. non-explicit tracks.
- A stacked bar chart comparing average radio spins for both groups.

These visuals were used because bar charts are effective for comparing numeric values across categories.





Analysis

Explicit tracks had much higher engagement on TikTok across all metrics over twice as many posts, likes, and views compared to non-explicit tracks. In contrast, radio spins were higher for non-explicit tracks, with explicit songs getting far fewer plays.

Interpretation

This suggests that explicit tracks perform better on platforms like TikTok, where content is user-driven and less restricted. However, they are played less often on radio, possibly due to content guidelines or audience preferences.

Conclusion

Explicit tracks receive more attention on TikTok but are played less on radio. This reflects how different platforms cater to different audiences and content standards.

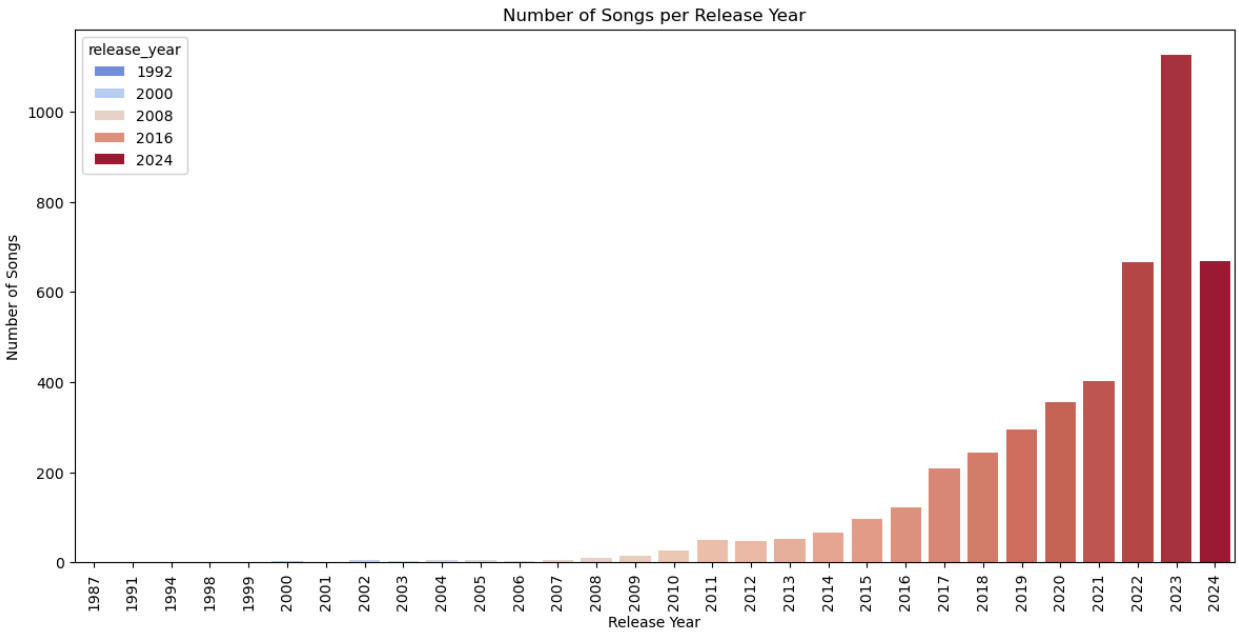
Question 8: How does a song's release year influence its cumulative streaming performance across different platforms?

This question examines whether the release year of a song affects how many total streams it accumulates across platforms like Spotify, YouTube, Pandora, SoundCloud, and TikTok.

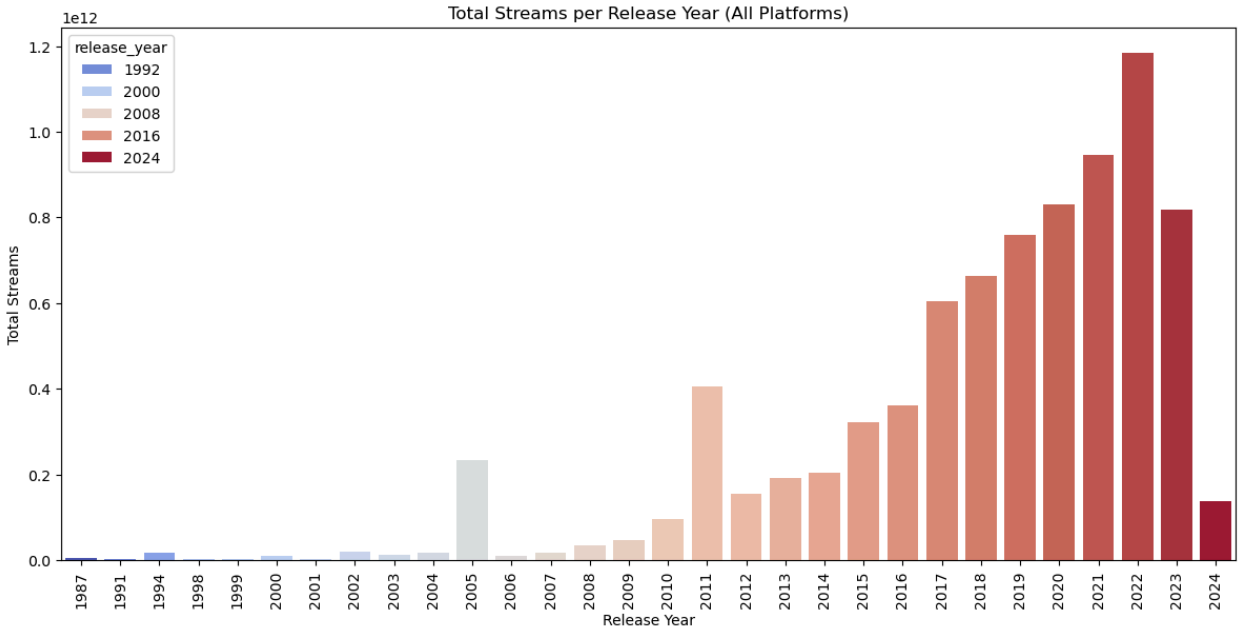
Method and Visualizations

To address this question, we used two bar charts:

1. Graph 1: Displayed the number of songs in the dataset by release year.
   This helped us understand how evenly songs are distributed across the timeline.



2. Graph 2: Showed total cumulative streams (all platforms combined) grouped by release year.
   This provided insight into how streaming totals vary over time.



Analysis

- Graph 1 showed that older release years (e.g., before 2010) have very few songs in the dataset, while more recent years (2012–2022) are more consistently represented.
- Graph 2 revealed that total streams generally increase from 2012 to 2022, but some older years (e.g., 1994) also show high totals due to a small number of viral or iconic songs.
- Streaming appears lower for 2024, likely due to fewer releases and limited time since release.

Interpretation

The number of songs per year heavily influences how we interpret total streams. A small number of older tracks may skew results if one or two became unexpectedly popular. On the other hand, newer songs benefit from broader streaming access and platform availability, but also face time limitations in accumulating views.

Conclusion

Release year does impact total streaming performance, but not in a linear way. While songs from recent years (2012–2022) tend to have higher and more consistent streaming totals, older hits can distort trends due to dataset imbalance. For clearer insight, comparisons should focus on years with a balanced number of songs.

Question 9: Do Spotify tracks featuring collaborations achieve higher streaming numbers or playlist reach than solo tracks?

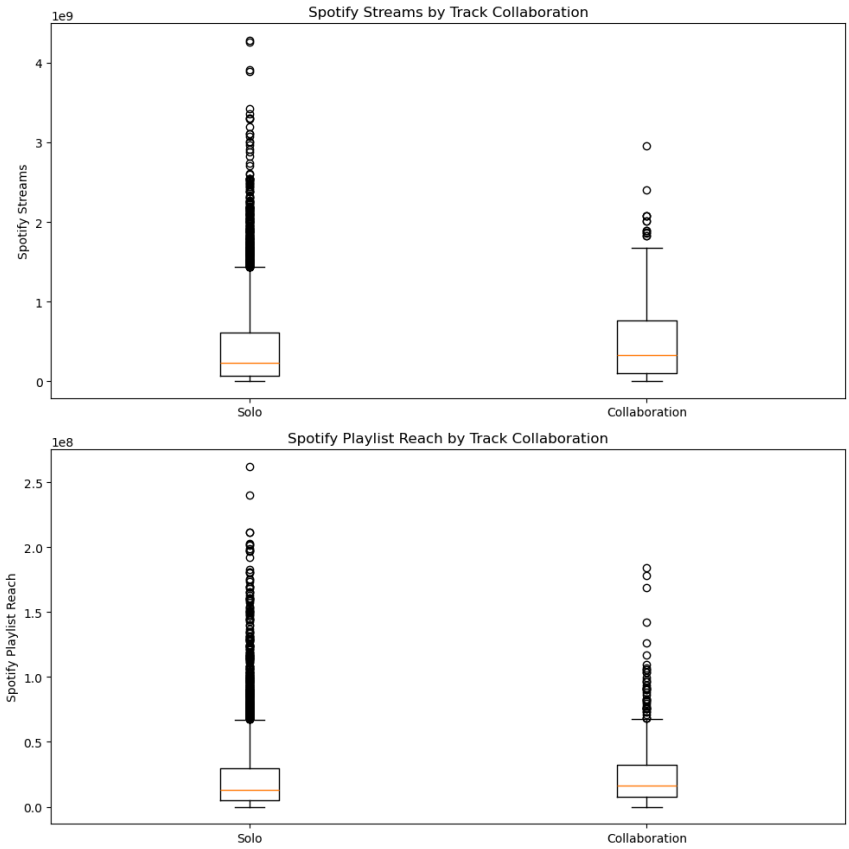
This question investigates whether collaborative tracks (those featuring multiple artists) perform better than solo tracks in terms of Spotify streams and playlist reach.

Method and Visualization

We flagged tracks as collaborations if their titles contained "feat." or "featuring." Two box plots were created:

1. Spotify Streams: to compare total stream counts between solo and collaborative tracks.
2. Spotify Playlist Reach: to compare how widely each track type is featured in playlists.

Box plots were chosen to highlight differences in distribution, medians, and outliers across both groups.



Analysis

- Spotify Streams:
  The median number of streams was nearly the same for both solo and collaborative tracks. The interquartile ranges also overlapped heavily, indicating no significant difference in typical performance.
- Spotify Playlist Reach:
  Playlist reach distributions were also similar, with substantial overlap. Collaborations didn't show a clear advantage in reach.
- Variability and Outliers:
  Both solo and collaborative tracks had high variability, with a few outliers reaching massive success. Interestingly, solo tracks showed a slightly higher number of hit songs.

Interpretation

Collaborative tracks do not consistently outperform solo tracks in either Spotify streams or playlist reach. While collaborations may bring broader exposure, solo tracks appear just as likely if not slightly more likely to produce standout hits.

Conclusion

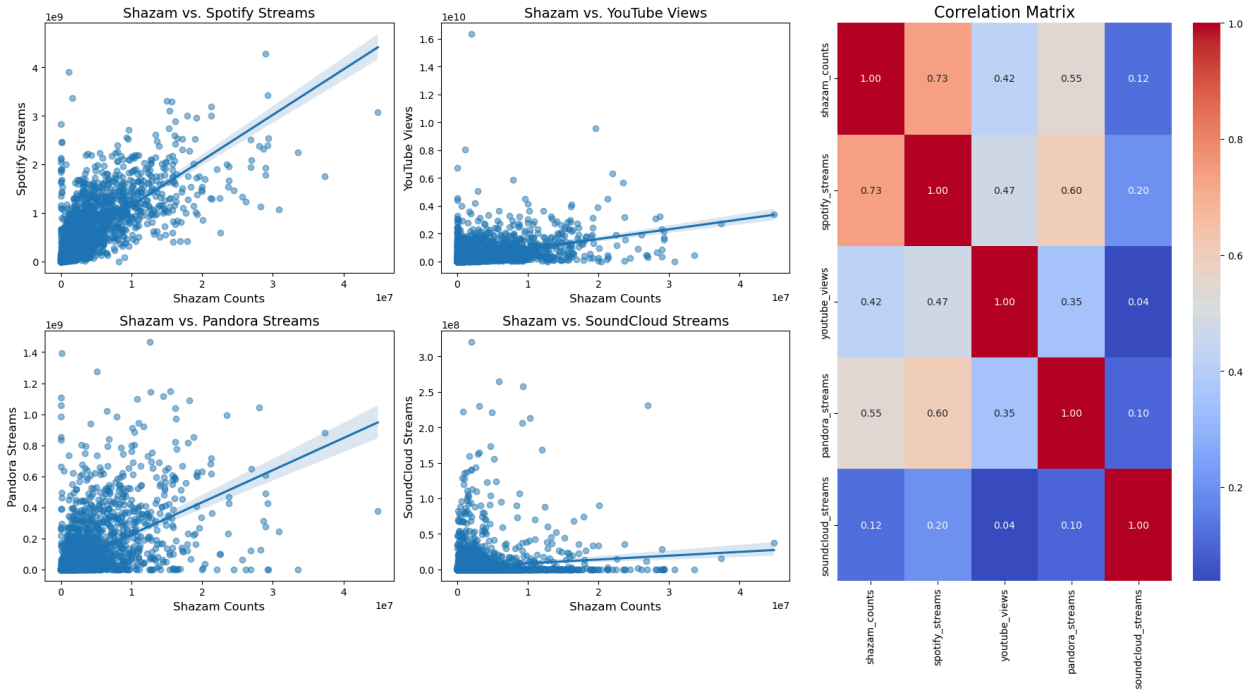
There is no strong evidence that collaborations lead to higher average performance on Spotify. Solo and collaborative tracks perform similarly overall, with success likely depending on other factors such as artist popularity, marketing, or timing.

Question 10: Do Shazam counts indicate a song's future success on streaming platforms?

This question explores whether Shazam activity can predict how well a song will perform on platforms like Spotify, YouTube, Pandora, and SoundCloud.

Method and Visualizations

We used four scatter plots comparing Shazam counts to each platform's stream count, along with a correlation heatmap to quantify the relationships. Regression lines in the scatter plots helped reveal the strength and direction of each trend.



Analysis

- Spotify
    - Correlation: 0.73 (strong)
    - Shazam counts are strongly linked with higher Spotify streams.

- YouTube
    - Correlation: 0.42 (moderate)
    - Songs with more Shazams tend to receive more YouTube views.

- Pandora
    - Correlation: 0.55 (moderate)
    - Moderate relationship shows that Shazam interest translates to Pandora streams.

- SoundCloud
    - Correlation: 0.12 (very weak)
    - Shazam counts are not a reliable indicator of performance on SoundCloud.

Interpretation

Shazam activity is a strong predictor of Spotify success, and moderately predictive for YouTube and Pandora. SoundCloud appears to operate independently, with little correlation to Shazam behavior.

Conclusion

Shazam counts are a useful early signal for a song's future performance, especially on Spotify. While not equally predictive across all platforms, they offer a strong indication of public interest that often carries over into streaming numbers.

Section 2: After Scraping

Question 11: Do daily streams better predict all-time rank than total streams?

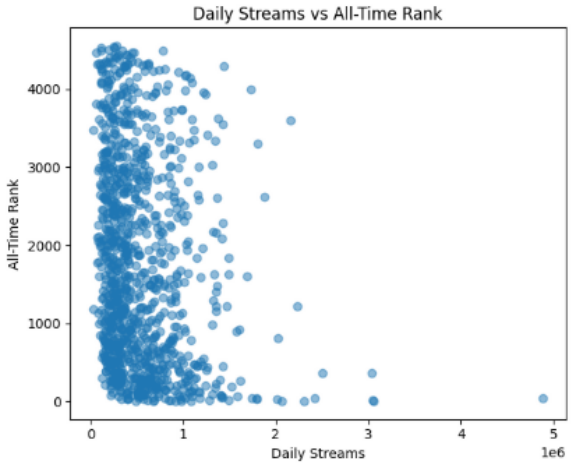
Objective

To assess whether a track's daily stream count is more closely tied to its all-time rank compared to its total lifetime streams.

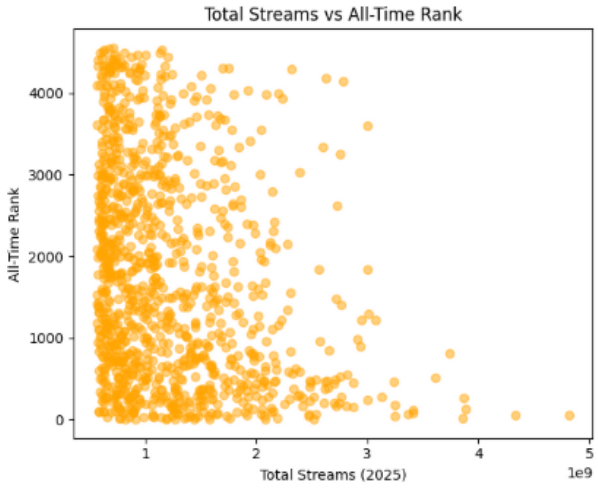
Method and Visualizations

Two side-by-side scatter plots:

- **Plot 1:** Daily streams vs. all-time rank



- **Plot 2:** Total (2025) streams vs. all-time rank



Each point represents a song. We visually compared the density and spread of data to understand which feature aligns more clearly with rank.

Analysis

- Daily Streams: Points cluster more tightly at lower ranks with higher daily streams, suggesting a clearer inverse trend.
- Total Streams: More widely scattered and noisy, with less obvious predictive value.

Interpretation

Daily streams reflect more immediate popularity and better align with real-time ranking than lifetime totals, which may include outdated momentum.

Conclusion

Daily streams are a stronger and more relevant predictor of a song's current rank than total historical streams.

Question 12: Do recently released tracks get more daily streams? (Recency Effect)

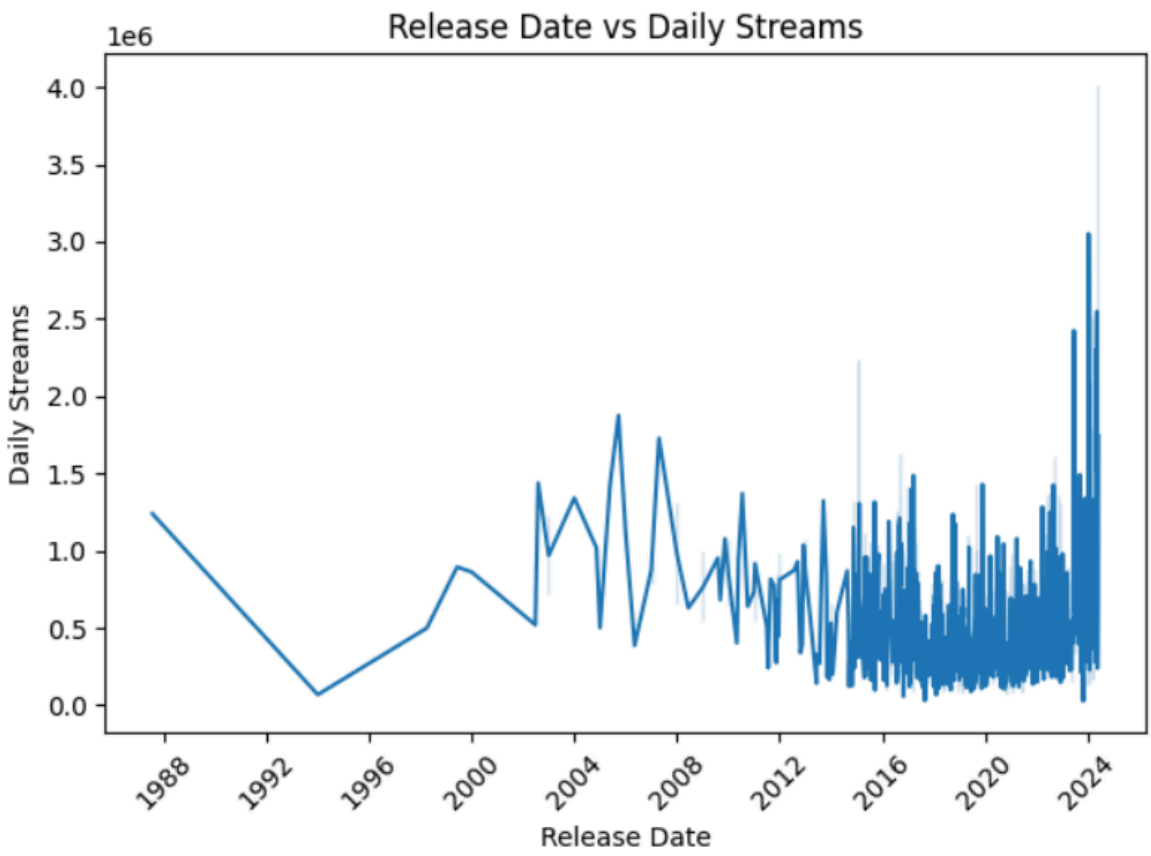
Objective

To explore whether newer songs receive higher daily streams, indicating a potential recency bias in listener behavior.

Method and Visualization

A time-series scatter/line plot showing:

- **X-axis:** Release date of the song
- **Y-axis:** Average daily stream count



Analysis

- Recent surge: Songs released post-2020 show significantly higher daily streams.
- Older tracks: Lower daily streams are common, even among well-established songs.

Interpretation

Streaming behavior skews toward newly released content, potentially influenced by playlist placements, algorithmic promotion, or social media trends.

Conclusion

It’s clear that newer tracks tend to receive more daily streams, reflecting a dynamic and fast-moving listening culture.

Question 13: Are some genres more associated with high track scores and ranks?

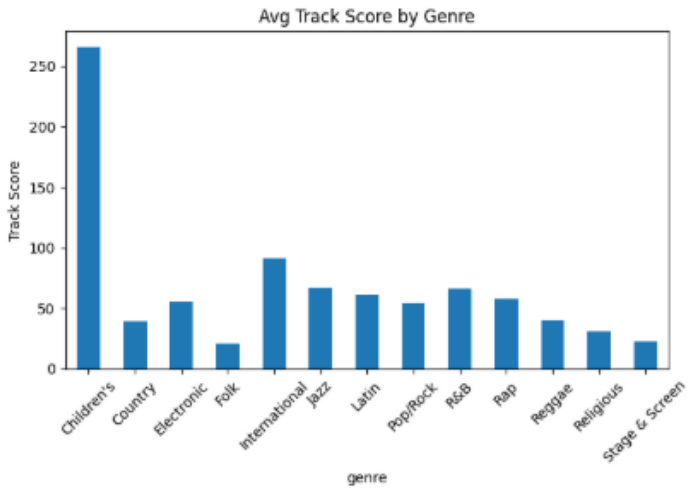
Objective

To determine whether certain genres consistently achieve better average track scores and all-time ranks.

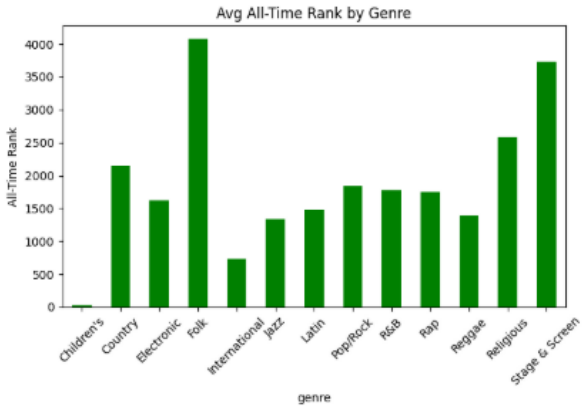
Method and Visualizations

Two side-by-side bar charts:

- Average track score by genre:



- Average all-time rank by genre:



Analysis

- Children's music shows the highest average track score and best (lowest) rank.
- Genres like Folk and Stage & Screen show consistently lower scores and poorer rankings.

Interpretation

Genres that are algorithmically favored or contextually reused (e.g., Children's) may benefit from sustained engagement. Niche genres may lag despite high quality.

Conclusion

Genre plays a significant role in streaming success metrics, with mainstream or functional genres performing better in both ranking and scoring systems.

Question 14: Do different genres dominate different platforms?

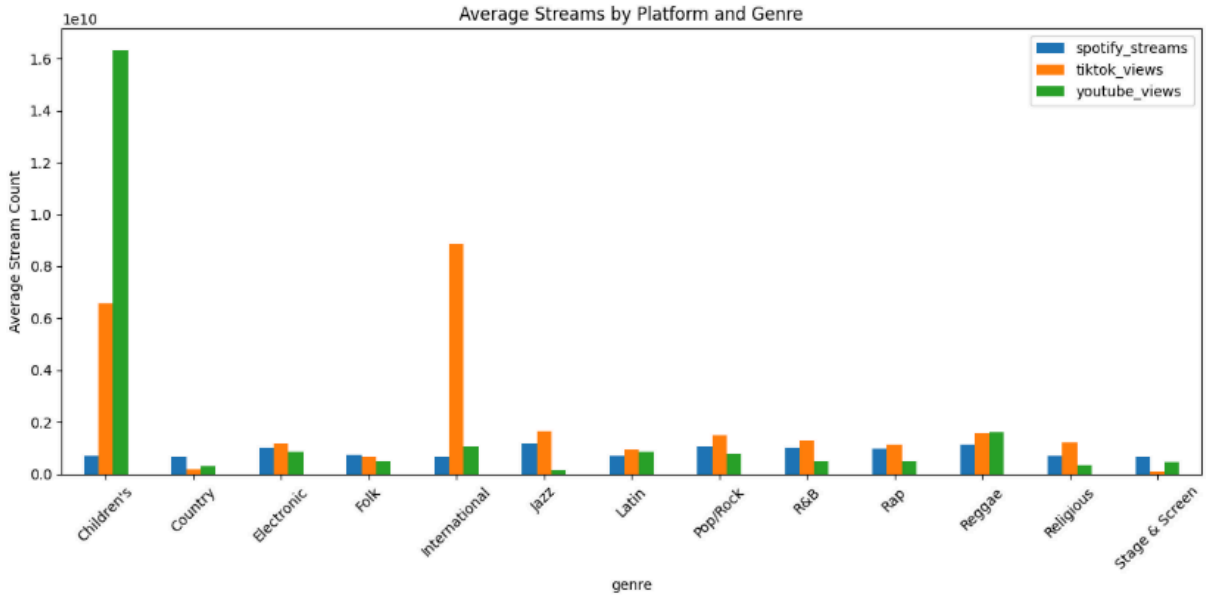
Objective

To analyze whether specific genres perform differently across major platforms (Spotify, TikTok, YouTube), indicating platform-specific audience behavior.

Method and Visualization

Grouped bar chart comparing these platforms by genre:

- Spotify streams
- TikTok views
- YouTube views



Analysis

- Children's and International genres dominate YouTube and TikTok by a large margin.
- Pop/Rock and Rap show more balanced performance across all platforms.
- Folk and Religious genres underperform consistently.

Interpretation

Each platform caters to different content types and listener preferences. YouTube favors visual-heavy or child-oriented content, while TikTok amplifies viral, short-form-friendly genres.

Conclusion

Platform dynamics matter: Genre success is platform-dependent, and artists may benefit by targeting the platform where their genre performs best.

Scraping Additional Data

To expand our analysis in Phase 2, we scraped additional data from a publicly available HTML page and external artist metadata using Python.

1. **Extracting Data from an HTML Table**

We started with an HTML file titled “Spotify Most Streamed Songs of All Time”, which contained a structured table. Using BeautifulSoup and pandas, we:

- Parsed the HTML file.
- Extracted the first <table> element.
- Converted the table into a DataFrame.
- Exported the data to CSV for further processing.

2. **Adding Genre Information via Web Scraping**

The original dataset lacked genre labels. To enrich it:

- We used the AllMusic website to identify the primary genre for each artist.
- For each track, we queried AllMusic’s search, extracted the track page, and scraped the listed genre.
- The genre scraping was done in batches with time.sleep() to avoid overloading the server.

3. Post-Scraping Cleaning

- Removed Null Values: Dropped rows with missing essential data.
- Removed Duplicates: Eliminated duplicate records to avoid redundancy.
- Handled Whitespace: Stripped leading/trailing spaces from text columns.
- Converted Data Types: Ensured numeric columns (e.g., streams) are properly typed.
- Formatted Strings: Standardized text (e.g., song titles and artist names).
- Dropped Irrelevant Columns: Removed any columns not useful for analysis or modeling.