Abstract

By: Gabe Compton, Henok Gelan, Abdur Islam, Osaze Ogieriakhi, Justin Moos, Aden Athar, Jaeger Nelson

1. Introduction

In the digital age, a song's success is no longer confined to traditional metrics like radio airplay or album sales. Instead, platforms like Spotify, TikTok, and YouTube play a crucial role in determining which songs become hits. This project explores how external streaming and social media platforms influence Spotify's Track Score, which is a measure of a song's popularity based on streams and playlist presence. By using Principal Component Analysis (PCA) and correlation analysis, we aim to identify the platforms that most significantly impact Spotify popularity. Additionally, we investigate whether older songs regain popularity due to trends on platforms like TikTok, and whether recency plays a role in a song's success. Our work involves data aggregation, cleaning, and transformation to ensure accurate and meaningful insights.

With the rise of music streaming, songs can gain popularity through multiple channels, often independently of the artist's promotional efforts. A track may go viral on TikTok, leading to millions of additional Spotify streams, or a music video may gain traction on YouTube, reviving an older song. While some artists deliberately optimize their music for viral potential, the true impact of external platforms on Spotify success remains unclear. The primary objective of this project is to determine whether trends from external platforms significantly influence Spotify performance, particularly Track Score, Playlist Count, and Playlist Reach.

We set out to investigate key data mining questions, such as: Which external platform has the highest correlation with Track Score? Are older songs more likely to regain popularity if they trend on TikTok or YouTube? Does a spike in YouTube views predict an increase in Spotify Streams? Through this analysis, we aim to quantify the influence of different platforms on Track Score, helping artists and industry professionals understand whether virality leads to long-term success or merely short-lived attention.

2. Data Mining Task

Our approach involves analyzing Spotify's most streamed songs in 2024, a dataset containing fields such as Track Name, Artist, Album, Release Date, Track Score, Spotify Streams, Playlist Count, and Playlist Reach. To measure the influence of external platforms, we also consider view counts from YouTube, TikTok trends, and historical streaming data. The expected output includes PCA visualizations and correlation matrices that illustrate which platforms contribute most significantly to Spotify's Track Score. These insights will help determine whether recency, external trends, or playlist exposure play the most critical role in a song's popularity.

One of the key challenges we encountered was aggregating data from multiple sources, as Spotify, TikTok, and YouTube operate on different metrics. While Spotify directly reports streams and playlist placement, external platforms provide less structured engagement data. Another challenge was handling the time lag between virality and its impact on streaming numbers—a song may trend on TikTok weeks or months before seeing an increase in Spotify Streams. Standardizing and normalizing this data was crucial for meaningful comparisons.

3. Technical Approach

To address these challenges, we applied several data preprocessing techniques. First, duplicate entries and missing values were removed to ensure consistency. We then standardized numeric values such as stream counts, playlist reach, and YouTube views to enable proper PCA and correlation analysis. Since different platforms have different engagement models, we also experimented with lag analysis, comparing Spotify growth rates before and after a song trended on external platforms. Our final approach involved generating visualizations that highlight correlations between Spotify Track Score and external platform performance, revealing patterns in cross-platform influence on music popularity.

As part of our methodology, we faced additional challenges in data visualization and interpretation. Spotify's Track Score is inherently influenced by multiple factors, and separating organic growth from trend-driven spikes was complex. Additionally, older songs that re-enter popularity due to social media trends required careful analysis to ensure their streaming numbers weren't inflated by pre-existing popularity rather than external influence. By leveraging Python libraries for data processing, visualization, and machine learning, we successfully transformed our dataset into readable insights.

4. Evaluation Methodology

The dataset used in this project was sourced from Kaggle, containing streaming statistics for Spotify's most played songs of 2024. While generally accurate, it required significant cleaning and normalization due to duplicate, missing, and inconsistent entries. Once preprocessed, the data was analyzed using PCA, correlation coefficients, and time-series analysis to evaluate the relationship between Spotify Track Score and external platform influence.

To evaluate the effectiveness of our findings, we relied on several key metrics, including total views per platform, release date, song popularity rankings, and playlist reach. These metrics allowed us to compare trending vs. non-trending songs, providing insights into how external exposure translates into long-term streaming success. Our results, once finalized, will provide a comprehensive understanding of whether external platform trends directly impact Spotify Track Score, offering valuable takeaways for artists, labels, and streaming services.

By: Gabe Compton, Henok Gelan, Abdur Islam, Osaze Ogieriakhi

#1 Introduction

- **Motivation from real world applications for the task chosen**
    - Ideas
    - Does the recency of a song relate to its popularity?
    - How do the views and streams on different platforms affect track score? Which platform is the most influential. How does the release date
- **Examples of data mining questions we set out to investigate in this project**

Some data mining questions could be:

- Which platform has the highest correlation with Track score?
- Do older songs be more likely to regain popularity if they trend on platforms like TikTok or YouTube?
- Do songs that trend on TikTok experience a measurable increase in Spotify streams?
- Is there a correlation between TikTok or YouTube trends and increased playlist reach on Spotify?

- **State our personal motivation for selecting this particular project**

We were curious about how much each platform affects the overall popularity of a song. Does an old song have more popularity than other songs because it was popular on TikTok or another platform? How do trends influence how popular a song is?

- **Describe the challenges and approach**

Spotify and other platforms are not always extremely transparent with their data. However with API's and other means, the challenge of aggregating all of the data into one place was overcome. With this aggregated data, our next issue was filtering out incorrect or irrelevant data, so that we are only working with clean data. Once this data was cleaned, we were able to more filter the data using python libraries into readable and comparable data.

- **Summarize results(this is something we will have later for the final submission)**

#2 Data mining task

- **Clearly describe all the details of the task. What is the input data? What is the output of the data mining approach? Give examples to illustrate them.**

Input: Spotify's most streamed songs in 2024 with various fields like, Spotify streams, track score, artist, album, release date, and many more.

Output: PCA and correlation values based on various fields, along with graphs, diagrams and other visualization. We can have a PCA graph of all the different streaming platforms and track score to see what is the most influential to the score. (Other than spotify listens)

- **List all the data mining questions that you set out to investigate in this project.**
    - Which platforms have the most impact to track score
    - Relation of spotify streams compared to older platforms

- **List the key challenges to solving this task**

Our team will face quite a few challenges such as cleaning data, aggregating data into a usable format, comparing the data using python libraries, normalize and standardizing data, displaying the data, and more will appear as the project progresses.

#3 Technical Approach (#2 challenges is similar to #3 approach to solving)

- **Describe all the details of your algorithmic approach to solving this data mining task and/or answering the data mining questions.**

    Standardization and normalization of data

    Removing duplicates/combining

    Compare data and display

- **How are you addressing the challenges mentioned above**

What we will likely be using for addressing the challenges are likely going to be consulting large language models along with several coding sites like geeksforgeeks and looking into code used in kaggle related to our challenges

- An algorithmic pseudo-code and/or a figure (block diagram) to illustrate the approach will be good. (don't have yet and may not need this section)

#4 Evaluation Methodology

- **Explain the dataset and its source that you employed to study this task. Any specific challenges to using this data for your task?**

    The dataset we used was a messy dataset from Kaggle. This dataset was accurate, but had many duplicate and blank entries, so the data needed to be cleaned, normalized, and standardized before it could be used.

- **List the metrics you employed to evaluate the output of the data mining task and/or questions investigated. Justify their choice from a real-world applications perspective.**

Metrics we reviewed included total number of views per platform, the release date of each song, name of each song, artist of each song, and the album that each song was a part of. Each of these metrics were required to create a comparable dataset that we were able to use to generate our visualizations.

5. Results and discussion (ignore this is something for the final part of the project)

- Present and explain results in a step-by-step manner to tell us a story about what you have discovered by doing this project (all graphs and tables should be properly labeled with legends and captions. They should be self-sufficient to understand the results)

- What worked and why?

- What didn't work and why not?

6. For Reproduce Project(probably not going to be included as this is after completion)

- Clearly describe all the details of the reproduced algorithms. Why do you choose these three algorithms? Describe the simulation results and observations.

- Will you be able to reproduce the same results as those reported in the original paper? If not, what do you think are the reasons?

- What you have learned from these different papers?