# The Common Denominators of Spotify Hits

Taha Shakeel, Elwin Li, Roger Jin, Tait Duan

2025-12-07

## Contents



## Introduction

Music streaming platforms like Spotify have transformed how listeners discover and engage with songs, creating an environment where popularity can shift rapidly and trends can emerge overnight. With millions of tracks competing for attention, understanding what distinguishes highly streamed songs from the rest has become an intriguing question for both listeners and music analysts. Our goal is to understand how factors such as song title, release date, social media presence, and playlist presence impact a song's success on Spotify. This analysis will illustrate how even small component's of a song's presentation may contribute to success.

## Dataset

Our dataset is sourced from Kaggle and contains comprehensive streaming and playlist metrics for the most streamed songs on Spotify in 2024. The data captures performance across multiple music streaming

platforms and social media channels, this provide us a multidimensional view of how songs achieve success in the modern streaming era.

The quantitative variables include:

- `Spotify.Streams`: Total number of streams on Spotify
- `Spotify.Playlist.Count`: Number of Spotify playlists containing the song
- `Spotify.Playlist.Reach`: Total audience size across all Spotify playlists containing the song
- `Spotify.Popularity`: Spotify's proprietary popularity score for the song
- `Deezer.Playlist.Count`: Number of Deezer playlists containing the song
- `Deezer.Playlist.Reach`: Total audience size across all Deezer playlists containing the song
- `Amazon.Playlist.Count`: Number of Amazon Music playlists containing the song
- `Apple.Music.Playlist.Count`: Number of Apple Music playlists containing the song
- `YouTube.Views`: Total views on YouTube
- `YouTube.Likes`: Total likes on YouTube
- `TikTok.Posts`: Number of TikTok posts featuring the song
- `TikTok.Likes`: Total likes across TikTok posts featuring the song
- `TikTok.Views`: Total views across TikTok posts featuring the song
- `Track.Score`: Overall track performance score
- `TIDAL.Popularity`: TIDAL's popularity score for the song

The qualitative variables include:

- `Track`: Song title
- `Artist`: Artist or band name
- `Album.Name`: Name of the album containing the song
- `Release.Date`: Date the song was released
- `ISRC`: International Standard Recording Code (unique identifier for the recording)
- `Explicit.Track`: Binary indicator of whether the song contains explicit content (0 = clean, 1 = explicit)

Each row of the dataset represents one song with its associated streaming metrics across platforms. The dataset contains 4,600 observations after cleaning, with some missing values in certain platform-specific metrics where songs may not be available or tracked on all services.

# Research Questions

## Research Question 1: Do the most common words in song titles correspond to the songs with the highest Spotify streams?

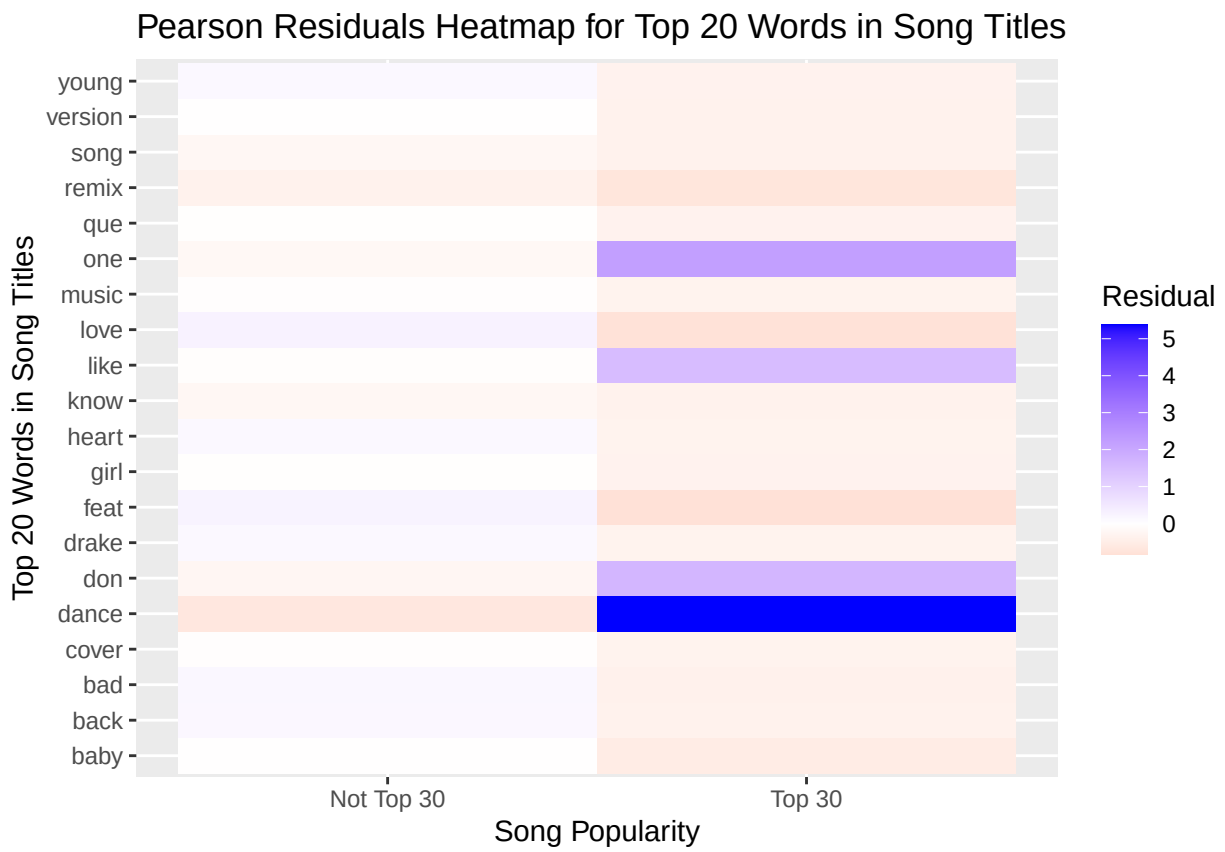
The title of a song plays a role in how listeners discover and remember music, and certain words may capture and retain attention more effectively than others. By examining whether the most common words in song title appear more frequently in the highest-streamed Spotify tracks, we can find out if linguistic patterns in titles are associated with commercial success.

**Graph 1: Word Cloud of 100 Most Frequent Words in Song Titles**


After cleaning the song titles to remove stop words and insuring all the words are formatted the same, we get the following word cloud. From the plot we can see that an overwhelming number of the most steamed Spotify songs contain the word 'feat', indicating a lot of the songs feature another artist. The words 'love', 'like', 'baby', and 'remix' also standout, indicating that these words recur across a broad range of tracks. There are also a couple of artists names that are common in song titles like 'travis', 'drake', 'cardi', and 'don'. This suggests that they were featured in several most streamed songs.

Now that we have identified the top words, we can take the top 20 words and create a contingency table comparing their presence across all songs versus the top 30 most stream songs on Spotify.

**Graph 2: Pearson Residuals Heatmap for Top 20 Words in Song Titles**




This heatmap visualizes how strongly each word is associated with Top 30 streamed songs versus non–Top 30 streamed songs on Spotify, after adjusting for chance. Cells shaded dark blue indicate much higher-than-expected counts, while lighter or neutral colors indicate weaker or no associations. One of the patterns we can see is that words like 'dance', 'like', and 'one' show strong positive Pearson residuals in the Top 30 column, indicating that that titles containing these words may be over represented among the top preforming tracks. However, most of the top 20 words showed weak or neutral associations. Indicating that a vast majority of words have no meaningful relationship with streaming success. Then there are words like 'feat', 'love', and 'remix' that have higher negative residuals in the Top 30 column. However, those are also the top 3 frequent words which could be the reason they have a negative residual. Those words are bound to appear outside the top 30 streamed songs on Spotify more than inside because of the high frequency.

```
chi
```

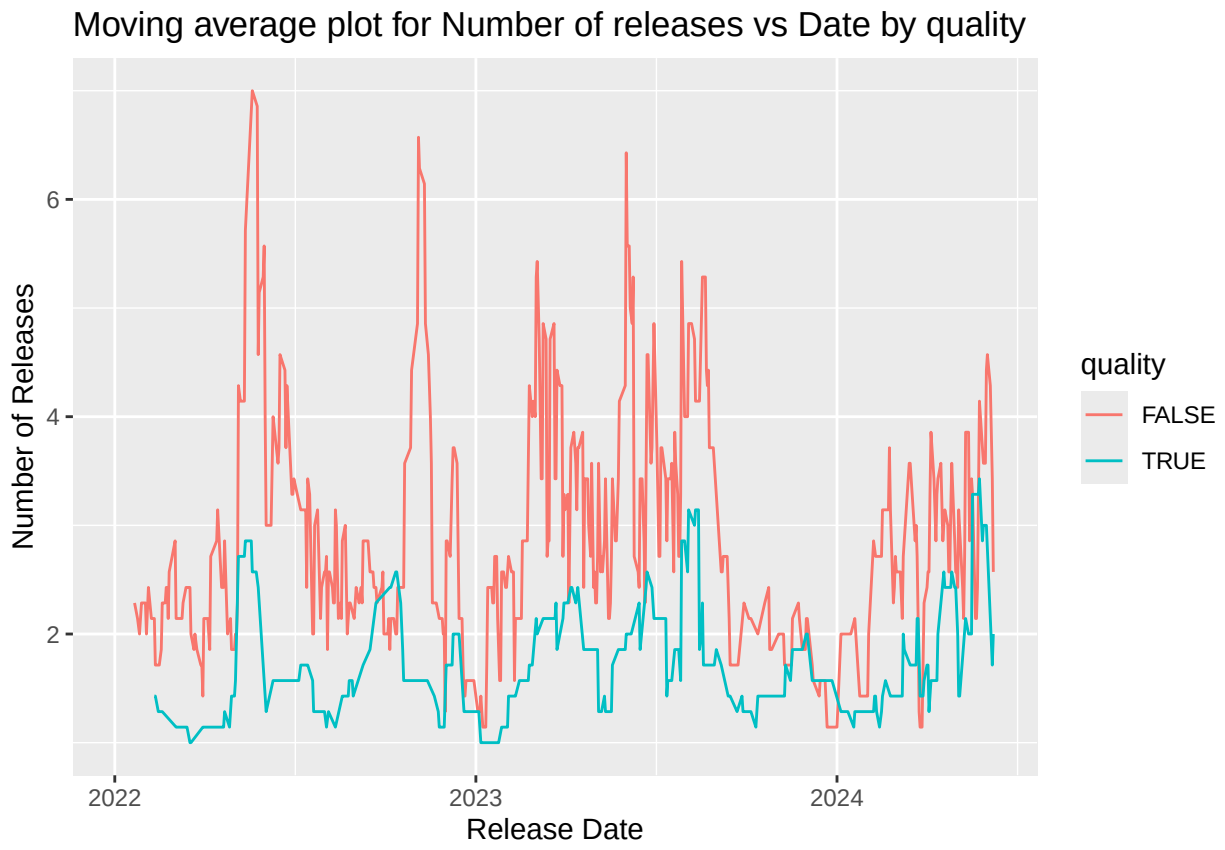
```
##
##  Pearson's Chi-squared test
##
## data:  cont_matrix
## X-squared = 65.307, df = 38, p-value = 0.003828
```

The chi-square results show a statistically significant relationship between word choice in song titles and whether a song appears in the Top 30 most-streamed tracks. The most frequently used words show little connection to popularity. But a small subset of words, 'dance', 'one', and 'like', appears disproportionately in top 30 streamed songs, as highlighted in the Pearson residuals heatmap. Overall, this suggests that word choice alone does not broadly predict streaming success, but certain title words may be associated with higher-performing songs.

## Research Question 2: Which times of release seems to be more common than others, and how does the time of release affect its overall quality and performance?

Next, we want to see if there are more common times where artists release songs, and if it affects the quality of the song. This is interesting because there could be seasonal factors that can affect whether or not a song is good or not. To see this, we make two graphs, a moving average time series plot, and a seasonal decomposition plot.

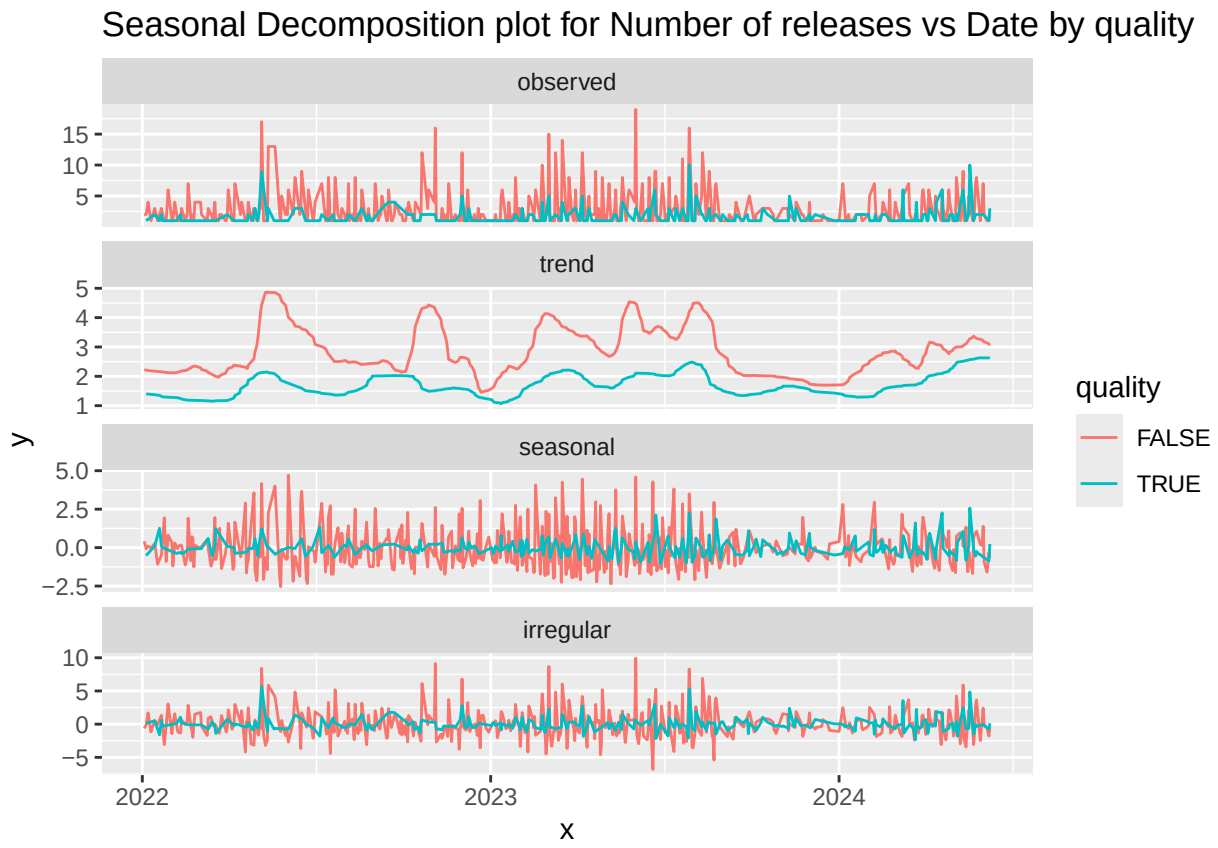
**Graph 1: Moving average plot for Number of releases vs Date by quality**




The moving average plot shown above shows the trend between the release date of songs from 2022 to 2024 and overall quality and performance of the songs. The so called "quality" variable is tracked using the `Track Score` variable in the dataset, which is calculated based on a variety of factors. We are using any `Track Score` of above or equal to 50 being a song that is considered quality, and a score of below 50 will not be considered to be a quality song. As we can see the red line, which is the low quality songs, have many spikes in number of releases throughout the time frame. The quality songs are less volatile. The most obvious high frequencies of releases seems to occur around the middle of the year. This is consistent throughout all the years. However, for 2022 it seems like there is also a surge of releases towards the end of the year, which is not consistent with 2023.

The trends of these two moving averages seems to be about the same, meaning that the release date probably does not effect the overall quality the songs end up being. Although, we can still see sometimes for example in late 2022, the low quality songs had a higher number of releases and high quality songs had lower number of releases than normal. The same thing on a smaller scale occurs around the same months in later 2023, which may suggest that high quality songs are less commonly released around the august-october months.

To see if there is actually a pattern in the months, we will make a seasonal decomposition plot. ### Graph 2: Seasonal Decomposition plot for Number of releases vs Date by quality



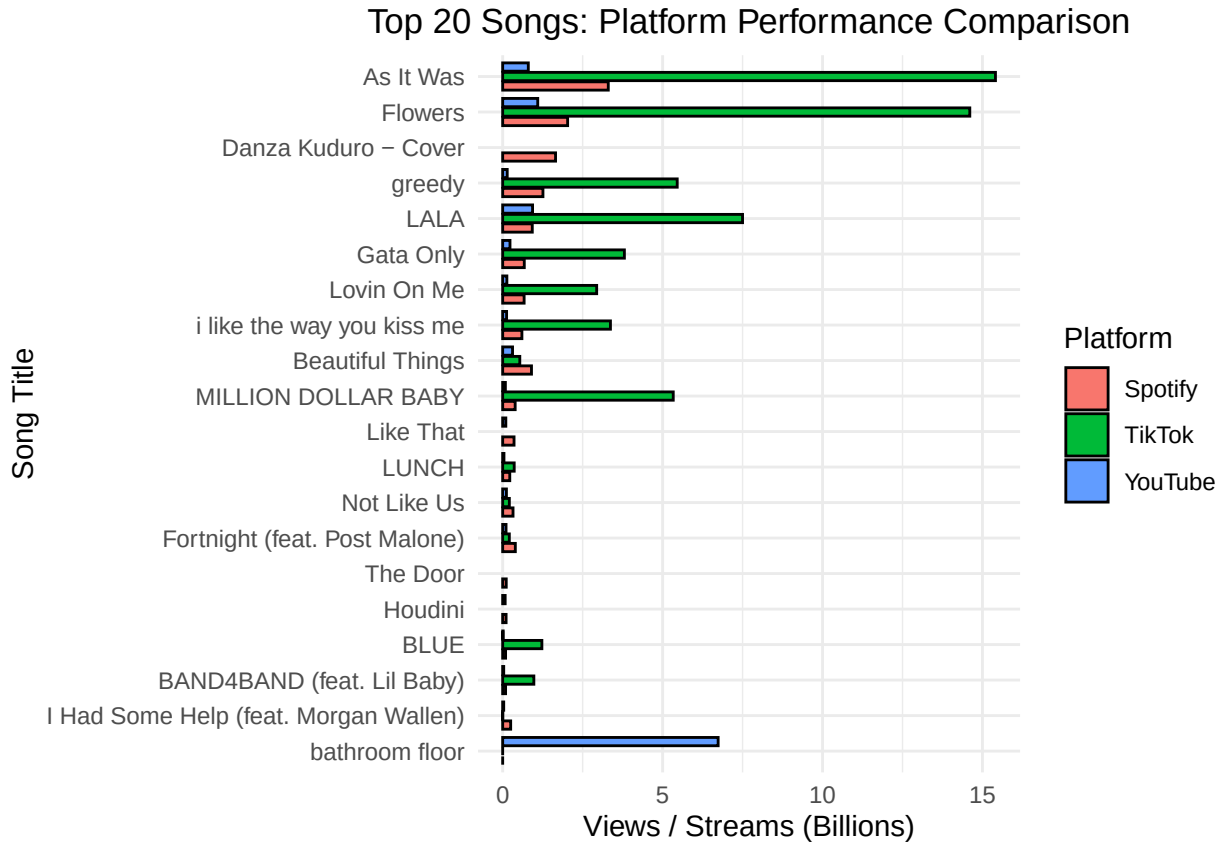

This seasonal decomposition plot is made with frequency 12, for the 12 months of the year, and we also made the plot on a modified version of the dataset, which is grouped by months. We can see that the low and high quality lines seems to mostly follow the same patterns and trends, and doesn't really deviate much against each other. This means that the effects of the months of the year on quality are very weak, if there is any at all.

## Research Question 3: Which factors most strongly predict a song's cross-platform success, and do songs achieve popularity through dominance in one area or through broad multi-platform engagement?

Understanding what drives song success in the streaming era is crucial for the music industry. With platforms like Spotify, TikTok, and YouTube shaping listening habits differently, a key question emerges: do songs achieve maximum popularity by dominating one platform, or do they require coordinated engagement across multiple channels? This analysis identifies which factors play the most influential roles and examines whether maximum success requires dominance in one area or coordinated engagement across multiple platforms.

Before exploring which factors predict cross-platform success, we first visualize how the most popular songs perform across the major platforms-Spotify, TikTok, and YouTube. This comparison provides a high-level view of the variability in multi-platform engagement and motivates the need to understand what drives these differences.

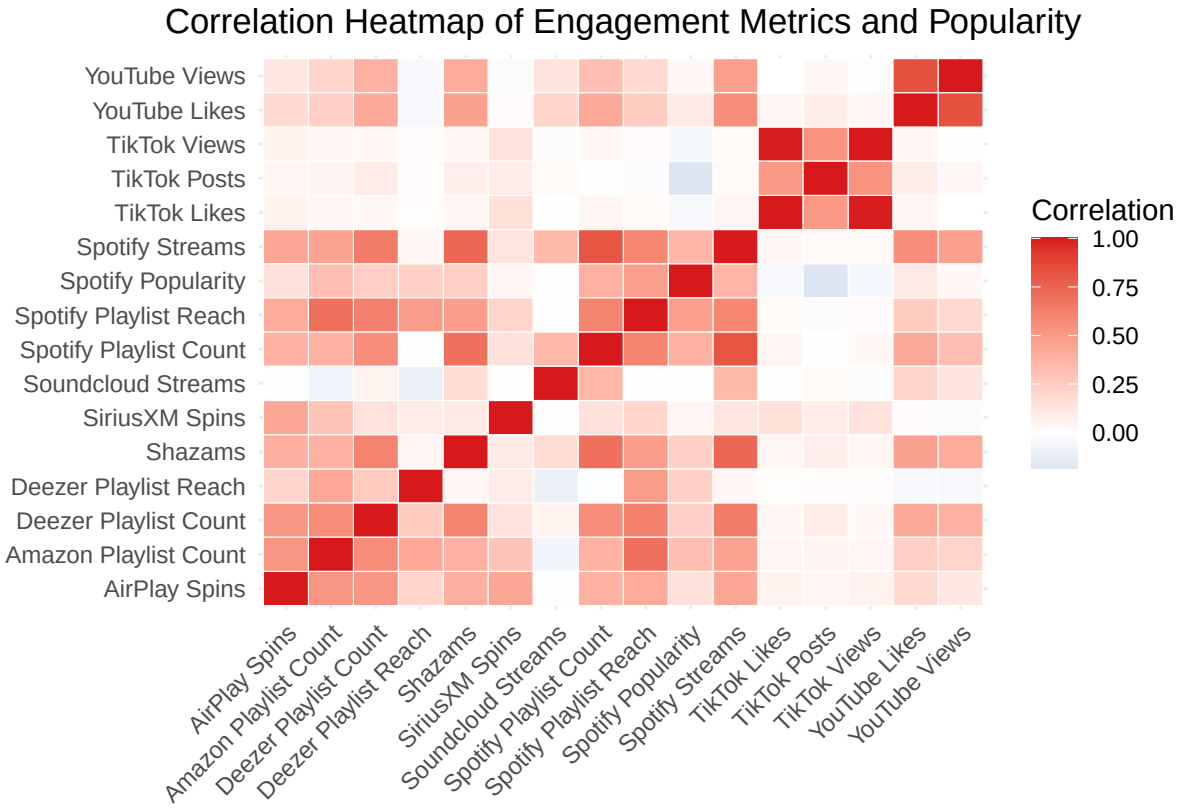
**Graph 1: Bar Plot: Platform Performance Comparison**



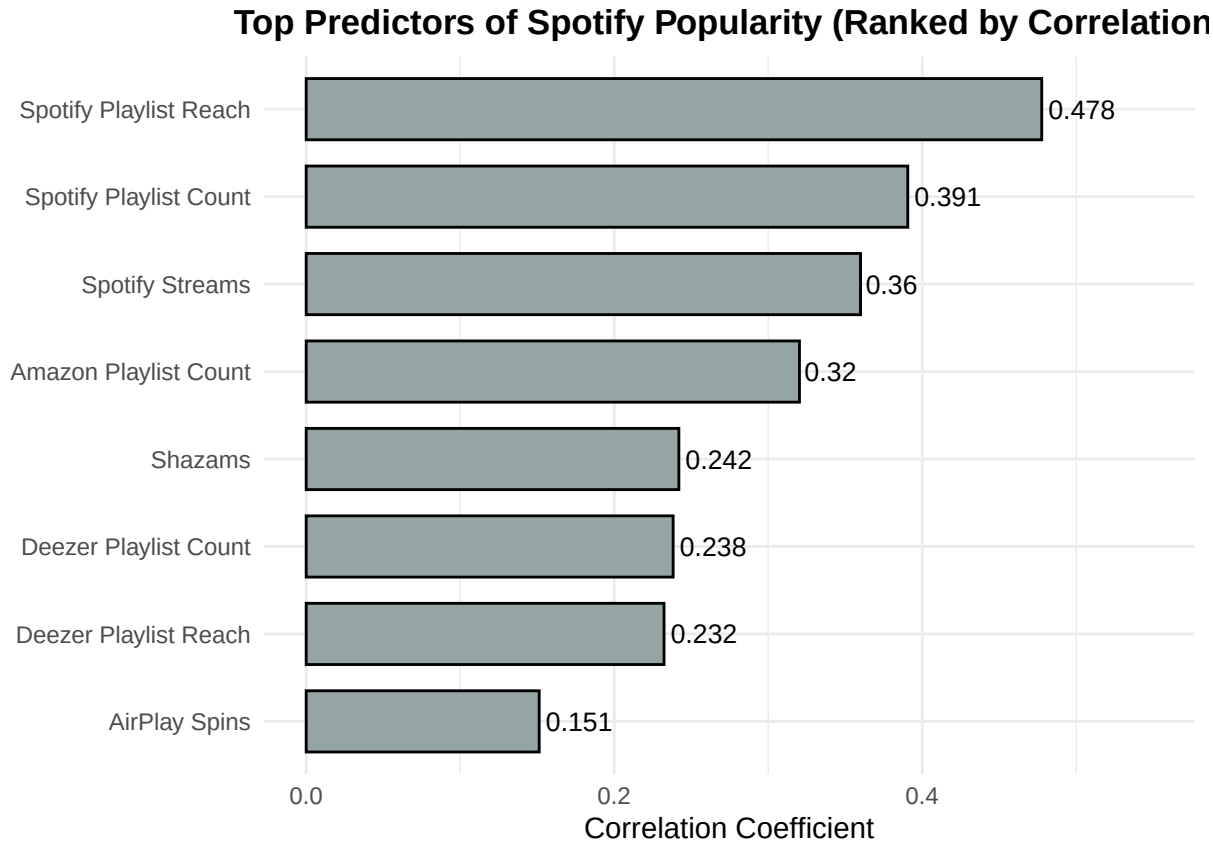

Graph 1 shows that the top 20 songs vary dramatically in their cross-platform performance. Some tracks, such as "As It Was" and "Flowers," have extremely high TikTok views relative to Spotify and YouTube, while "bathroom floor" is almost entirely YouTube-driven. Most songs show modest Spotify streams compared to their TikTok or YouTube presence. These differences demonstrate that songs do not become popular in the same way-some rely on TikTok virality, others on YouTube, and a few have more balanced performance. This variability directly motivates our research question: if songs can succeed through different platforms, does maximum popularity require dominance in one area, or does it require strong performance across all platforms? The answer to this question has important implications for understanding modern music promotion strategies.

Because Graph 1 shows that songs can perform very differently across platforms, we next examine whether specific engagement metrics are systematically related to overall popularity. To do this, Graph 2 computes a correlation matrix of key predictors-including TikTok activity, playlist exposure, YouTube engagement, and radio spins-to identify which factors are most strongly associated with popularity.

**Graph 2: Correlation Heatmap of Engagement Metrics and Popularity**



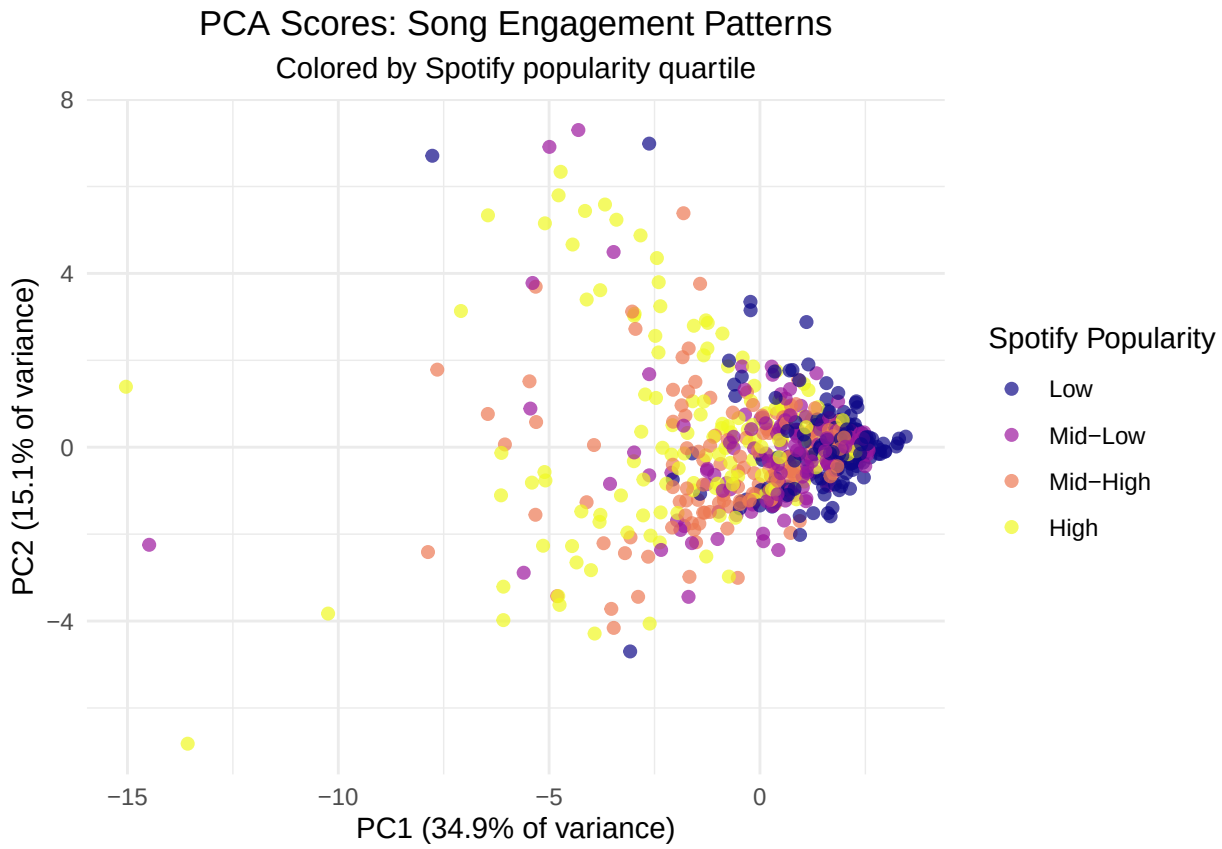
**Graph 3: Bar Plot: Top Predictors of Spotify Popularity (Ranked by Correlation)**



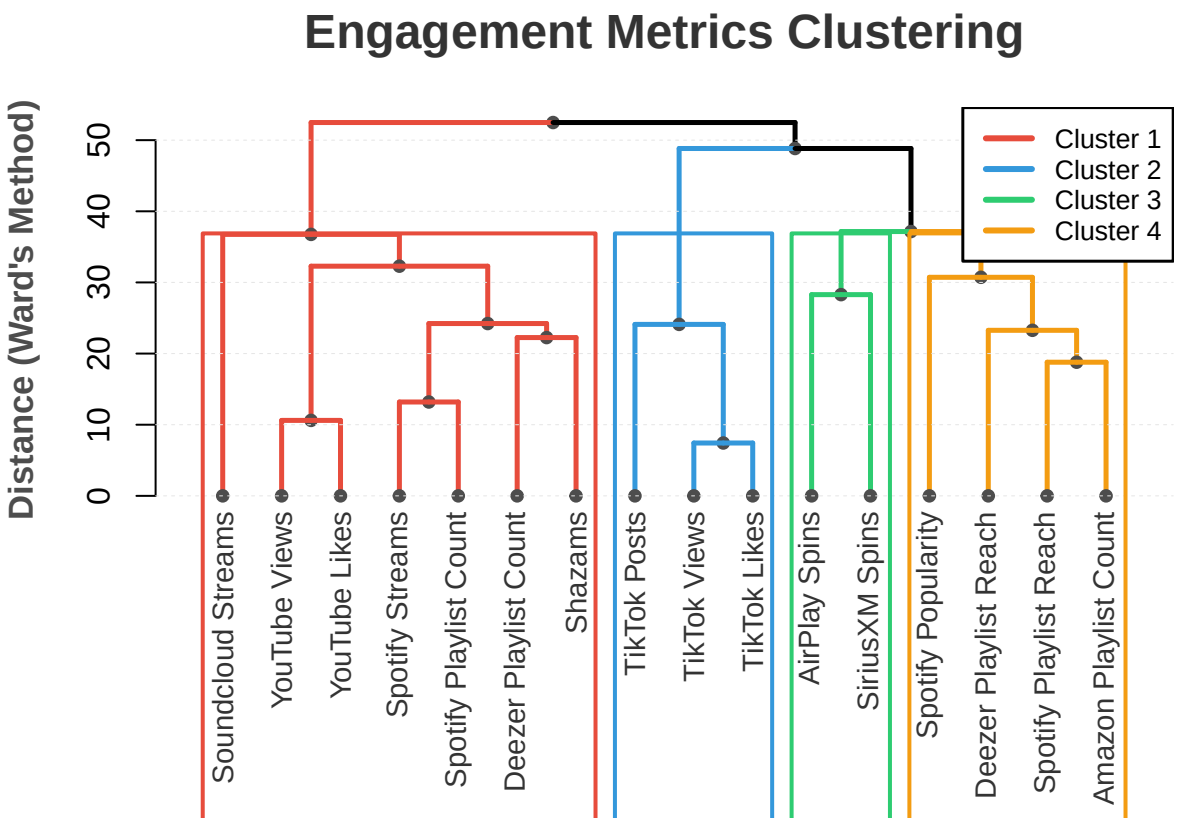

The correlation heatmap and ranking reveal which factors best predict Spotify popularity. The top five predictors are: Spotify Playlist Reach ($r = 0.478$) - strongest predictor, Spotify Playlist Count ($r = 0.391$) - strong predictor, Spotify Streams ($r = 0.360$) - moderate predictor, Amazon Playlist Count ($r = 0.320$) - moderate predictor, and Shazams ($r = 0.242$) - weak-to-moderate predictor. This ranking demonstrates that playlist exposure shows the strongest individual relationship with success, with the top two predictors both being playlist-related metrics. The heatmap also reveals clustering patterns: TikTok metrics (views, likes, posts) form a tight internal cluster with high inter-correlation ($r > 0.7$), while radio spins (AirPlay/SiriusXM) cluster separately, indicating these represent distinct promotional ecosystems. However, while these correlations identify which individual factors are strongest, they do not yet answer whether maximum success requires dominance in one area (e.g., extremely high playlist reach) or balanced strength across multiple platforms. To address this, we turn to dimensionality reduction techniques.

While correlation analysis identifies which individual factors are strongest, it cannot tell us whether top songs achieve success by excelling in one dimension or by achieving broad multi-platform engagement. To answer this question, we use Principal Component Analysis (PCA) to reduce the complexity of multiple overlapping metrics and examine whether high-popularity songs cluster in a specific region of the engagement space.

**Graph 4: PCA Scores: Song Engagement Patterns**




The PCA scores plot directly answers our second research question about dominance versus breadth. PC1-which explains 34.9% of total variance-captures a general "overall engagement intensity" axis: songs with high values on PC1 score high on playlist exposure, streaming volume, and cross-platform activity simultaneously. When songs are color-coded by Spotify popularity quartile, a clear gradient emerges: the highest-popularity songs (yellow points) cluster strongly toward the right side of PC1, indicating they achieve broad engagement across multiple dimensions rather than dominating in a single area. Lower-popularity tracks are more dispersed and scattered throughout the plot, suggesting more platform-specific or inconsistent engagement patterns. The tight clustering of high-popularity songs at high PC1 values demonstrates that maximum success requires coordinated multi-platform engagement, not dominance in one channel. While individual songs can be platform-specific (as Graph 1 showed), reaching the highest popularity tiers requires strong performance across playlists, social media, radio, and direct streaming simultaneously.




The dendrogram reveals that engagement metrics naturally cluster into four distinct ecosystems: TikTok virality (views, likes, posts), traditional radio (AirPlay and SiriusXM spins), playlist curation (Spotify/Deezer/Amazon playlist metrics), and direct consumption (YouTube views/likes, SoundCloud streams, Spotify streams). This clustering demonstrates that these promotional pathways operate semi-independently, providing context for understanding the multi-platform engagement patterns observed in the PCA. This clustering demonstrates that these promotional pathways operate semi-independently, providing context for understanding the multi-platform engagement patterns observed in the PCA.

**Which Factors Predict Success?**

Correlation analysis reveals that Spotify Playlist Reach ($r = 0.478$) is the strongest individual predictor of cross-platform success, followed by Playlist Count ($r = 0.391$), Spotify Streams ($r = 0.360$), Amazon Playlist Count ($r = 0.320$), and Shazams ($r = 0.242$). Playlist exposure emerges as the most powerful single factor, with the top two predictors both being playlist-related metrics. TikTok and YouTube metrics show weaker correlations, indicating that social media virality, while important, is less predictive than editorial playlist curation.

**Dominance vs. Broad Engagement?**

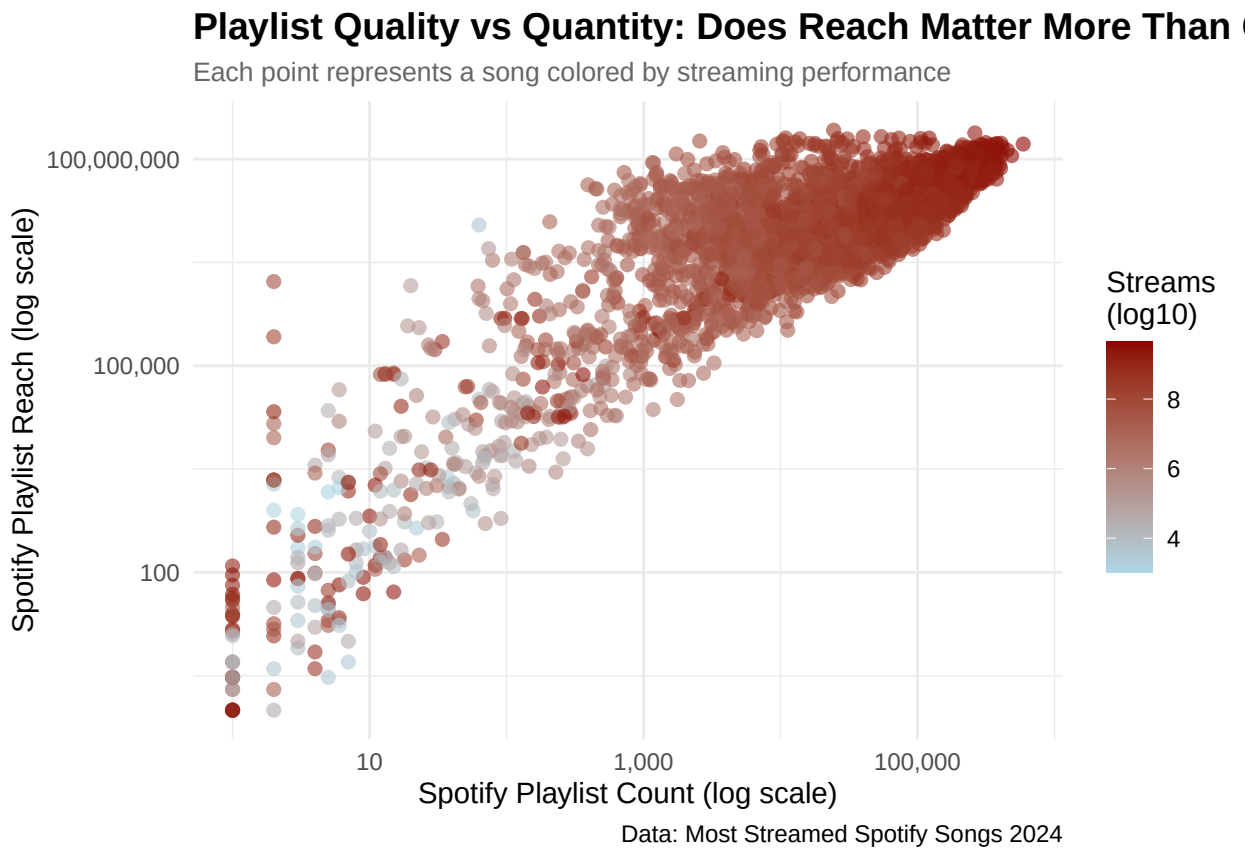
The PCA analysis directly answers whether songs achieve maximum popularity through dominance in one area or through broad multi-platform engagement. The results are clear: high-popularity songs cluster at high PC1 values, indicating they score strongly across multiple engagement dimensions simultaneously. This means that maximum success requires coordinated multi-platform engagement rather than dominance in a single channel. This means that maximum success requires coordinated multi-platform engagement rather than dominance in a single channel. While individual songs can be platform-specific, reaching the highest

popularity tiers demands strong performance across playlists, social media, radio, and direct streaming. Lower-popularity songs show more dispersed patterns, suggesting they lack this comprehensive engagement. The hierarchical clustering provides additional context by revealing that the metrics contributing to PC1 represent four distinct promotional ecosystems, reinforcing that maximum success requires engagement across multiple independent pathways rather than specialization in one.

## Research Question 4: Does playlist reach matter more than playlist count? Can songs be on fewer playlists but still crush it if they're the right ones?

Building on the previous question's findings, we now can look into the specific dynamics within playlist metrics. This analysis investigates whether being on more playlists automatically translates to greater streaming success, or if the quality of playlist placement (measured by audience reach) plays a more significant role. We employ an exploratory scatterplot to examine the relationship between playlist count and reach across thousands of songs, revealing substantial variation in how efficiently different playlists convert placement into audience exposure. Additionally, we use Principal Component Analysis (PCA) to identify underlying patterns in how songs achieve success across multiple platforms (Spotify, Deezer, Amazon Music, Apple Music), this allow us to determine whether distinct "playlist strategies" exist (some focused on volume, others on targeting high-reach playlists). Through this analysis, we aim to provide insights into optimal playlist placement strategies for artists and labels seeking to maximize streaming performance.

**Graph 1: Scatterplot: Playlist Reach vs. Playlist Count**




Based on the scatterplot above, it shows a clear positive trend: songs on more playlists reach more people. But what's really interesting is the huge variation among songs with similar playlist counts. For example,

around 1,000 playlists, some songs only hit a million listeners, while others reach tens of millions. This shows that not all playlists are equal, it indicates that some simply have much bigger audiences.

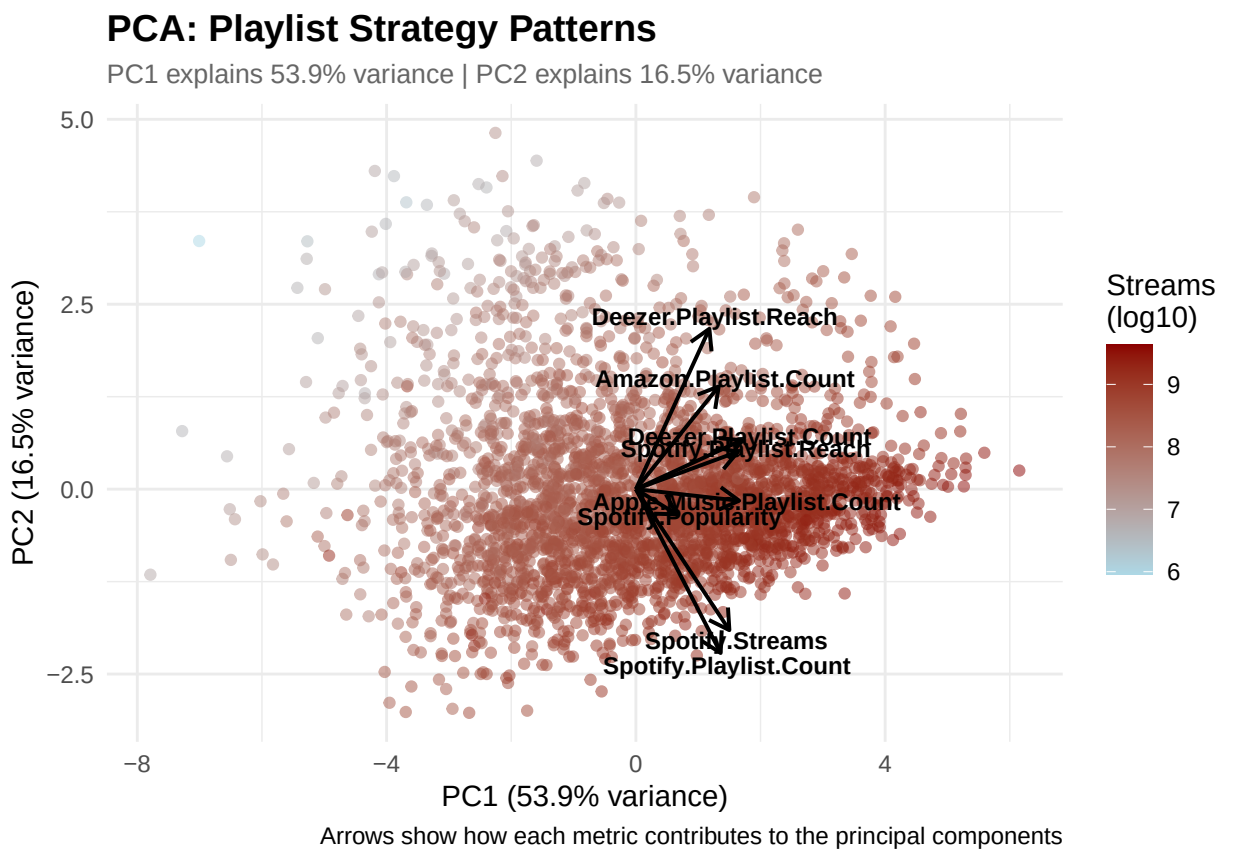
The darkest red points (the most-streamed songs) are mostly in the upper-right corner. This suggests that real hits don't just get added to a lot of playlists—many of those playlists also have large followings. So yes, playlist count matters, but which playlists you land on is just as important for reach.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.0775 1.1504 0.9700 0.7430 0.55575 0.52295 0.41437
## Proportion of Variance 0.5395 0.1654 0.1176 0.0690 0.03861 0.03418 0.02146
## Cumulative Proportion  0.5395 0.7049 0.8226 0.8916 0.93016 0.96435 0.98581
##                            PC8
## Standard deviation     0.33691
## Proportion of Variance 0.01419
## Cumulative Proportion  1.00000
```

```
##                                  PC1         PC2
## Spotify.Playlist.Count     0.3400202 -0.55372578
## Spotify.Playlist.Reach     0.4099105  0.12676225
## Deezer.Playlist.Count      0.4242590  0.16533960
## Deezer.Playlist.Reach      0.2948082  0.54195784
## Amazon.Playlist.Count      0.3322608  0.34696290
## Apple.Music.Playlist.Count 0.4141974 -0.03913098
## Spotify.Streams            0.3750476 -0.47654708
## Spotify.Popularity         0.1638127 -0.08515870
```

The PCA output tells us that just two components can summarize most of the story about 70% of all the variation in the data. PC1 alone covers over half(53.9%). Importantly, PC1 works like a general "hit song" score. If a song scores high here, it's doing well on every metric: more playlists, bigger reach, more streams, and higher popularity. PC2(16.5%) is where I think it gets interesting. It seems to separate songs that are big on Spotify (with lots of Spotify playlists and streams) from songs that gain traction through wider reach or other platforms like Deezer and Amazon. This last point feels important for the research. It looks like "reach" and "count" aren't measuring the same thing-where a song gets placed adds its own layer to success. beyond just the number of playlists.

**Graph 2: PCA: Playlist Strategy Patterns**




Looking at the PCA score plot, each song is a dot placed by its PC1 and PC2 scores, and the color shows how many streams it has (on a log scale). What really jumps out is a clear color shift from left to right. The darkest red dots (the most streamed songs) are clustered on the right side, where PC1 is high. This fits perfectly with PC1 being a general "success" score which we obbserved before: songs scoring high there are doing well across the board and also have the most streams.

Moving up and down along PC2, though, the colors don't change much. That tells us PC2 isn't really about how many streams a song gets, but more about how it gets them, whether it's mostly through Spotify or gains traction from playlists on other platforms.

This all ties back to the bigger question about playlists. You don't see any massively streamed songs sitting on the far left with low playlist counts, so from our perspective, quantity still obviously matters. But the full picture from the PCA suggests something more: the biggest hits live where both playlist numbers and playlist reach are strong. It's not just "get on as many playlists as possible." The data implies that getting onto playlists with bigger, wider audiences (the higher-reach ones) adds a real boost beyond just the count. In short, it looks like a "quality plus quantity" strategy is what separates the biggest hits from the others.

# Conclusion

Across our analyses, we explored how different factors relate to Spotify song popularity, focusing both on patterns in song titles and broader engagement across platforms. Our chi-square test and Pearson residual heatmap indicate that while most common words in song titles show no meaningful link to success, a few specific terms—such as 'dance', 'one', and 'like' appear disproportionately in the Top 30, suggesting that word choice has limited but not entirely negligible influence on streaming performance. When examining

seasonal patterns in release data, the moving averages and seasonal decomposition reveal that high and low-quality songs follow nearly identical trends throughout the year, which mean release month has little effect on eventual popularity aside from minor seasonal dips. The clustering and PCA results highlight that popularity is influenced more by cross-platform success than by title wording or release timing. Major hits happen when playlist reach, TikTok virality, radio play, and other forms of exposure work together, with playlist placement serving as the strongest driver. Taken together, these findings suggest that while certain words may slightly coincide with higher-performing tracks, streaming success is overwhelmingly tied to multi-platform amplification rather than the structure of a song title or the timing of its release.

## Future work

Our dataset lacks genre classifications and audio features like tempo, danceability, and energy. Future work could use Spotify's Audio Features API to investigate whether certain musical characteristics drive platform-specific success, whether high-energy tracks perform better on TikTok while melodic tracks gain YouTube traction. Understanding these patterns would help explain why certain songs achieve cross-platform success beyond just engagement metrics.

We examined release timing as a snapshot, but tracking how songs gain and maintain popularity over time would provide deeper insights. Questions like "How quickly does TikTok virality translate to sustained Spotify success?" or "Do songs released in certain months have longer chart lifespans?" would require time-series data showing weekly or monthly streams rather than cumulative totals.

Our analysis identifies correlations between metrics and success, but establishing causal relationships would require more sophisticated approaches. Additionally, building predictive models to forecast which songs might become hits would need training data that includes songs that failed to chart-our current dataset only includes the most streamed songs of 2024, introducing bias that limits our ability to identify what truly distinguishes successful songs from unsuccessful ones.