**Analyzing the Factors Driving Song Popularity on Spotify:**

**Insights from Streaming Data**

**Team Members:** Mark Daza

Aarushi Kotwani

Surya Kailash Ramesh

Zahra Ahmadi Angali

**ABSTRACT**

This report analyzes Spotify’s Most Streamed Songs Dataset, sourced from Kaggle, to explore the factors influencing song popularity in the streaming era. The dataset includes information about musical features, streaming metrics, and platform presence across Spotify, Apple Music, and Deezer. Key questions include how musical attributes like danceability and energy correlate with streams, the role of playlist placements and release dates, and the impact of the artist collaborations. We also hope to reveal if the release dates of songs influence their popularity.

Findings reveal that most songs have fewer than 500 million streams, with only 10% of the songs we evaluated surpassing 1 billion, highlighting the dominance of global hits. Playlist placements significantly drive streaming success, while musical features like energy and danceability show weak correlations. Songs released after 2010 attained more streams, reflecting the impact of current popular culture. Cross-platform appeal and collaborations also influence success, though variability exists. These insights provide a better understanding of what drives streaming success, offering helpful guidance for artists, music curators, and marketers.

**INTRODUCTION**

Music streaming platforms like Spotify have transformed the way we experience music, making it easier for songs to reach audiences worldwide. This report uses data from the Spotify Most Streamed Songs Dataset, sourced from Kaggle, which provides comprehensive information about Spotify's most streamed songs. The dataset is enriched with additional insights from platforms like Apple Music, Deezer, and Shazam, offering a broader perspective on song popularity across multiple platforms.

We aim to answer key questions such as:

* What attributes make a song highly streamable across platforms like Spotify, Apple Music, and Deezer?
* Are there correlations between musical features (e.g., danceability, energy, and valence) and streaming success?
* How do platform-specific features, like playlist placements and chart rankings, influence the number of streams?
* Can we detect patterns related to the release date (year, month, or day) influencing a song's popularity?
* What are the most-streamed songs on Spotify?
* Do certain times of the year (month or day of the week) show higher release activity for popular songs?
* How do individual artists or collaborations impact streaming numbers?

The dataset includes details about song features (like tempo, energy, and valence), streaming metrics, release dates, and artist counts. By analyzing this data, we aim to uncover patterns and trends that explain what makes certain songs achieve massive success in today’s competitive streaming world.

**METHODOLOGY**

The dataset contained 953 songs with 25 variables, including details about tracks, musical attributes, streaming metrics, and platform presence. During the preprocessing stage, the streams column, which had only one missing value, was handled by removing the corresponding row, as it was essential for the analysis. Missing values in the in\_deezer\_playlists and in\_shazam\_charts columns were replaced with 0, under the assumption that these missing values indicated the song was not featured on these platforms. This approach ensured that no bias was introduced while retaining as much data as possible.

To address the 95 missing values in the key column, represented as empty strings, a Random Forest model was implemented. The dataset was prepared by replacing empty strings with NA and separating rows with missing and non-missing values. The non-missing data was split into a training set (70%) and a test set (30%) using stratified sampling. Relevant features describing musical characteristics, such as rhythm, mood, energy, acoustic qualities, and the number of streams, were selected for the model. An initial Random Forest model identified the most important features, but the model's accuracy was found to be too low, indicating that these variables were not efficient for predicting the key. As a result, it was decided to proceed with the analysis of this attribute without the 95 missing values. Visualizations were created to explore relationships between musical keys, positivity, and stream counts, providing insights into song popularity.

Non-essential columns, such as cover\_url (which contained image links), were excluded as they were irrelevant for statistical or correlation analysis. A correlation matrix was then computed using relevant variables, including streams, danceability, energy, and valence, to identify relationships between features. Furthermore, numerical transformations ensured that the data was uniformly scaled and suitable for visualization.

Visualization techniques such as histograms, scatterplots, and heatmaps were employed to explore patterns, distributions, and correlations. Boxplots were used to analyze the distribution of streams across playlist count ranges, providing insight into how playlist exposure impacts song popularity. These steps ensured that the dataset was optimized for accurate, clear, and insightful analysis. The methodology provided the foundation for understanding the dynamics of song popularity in the context of streaming platforms.

**COMPUTATIONAL RESULT**

A graph of a distribution of streams

Description automatically generated

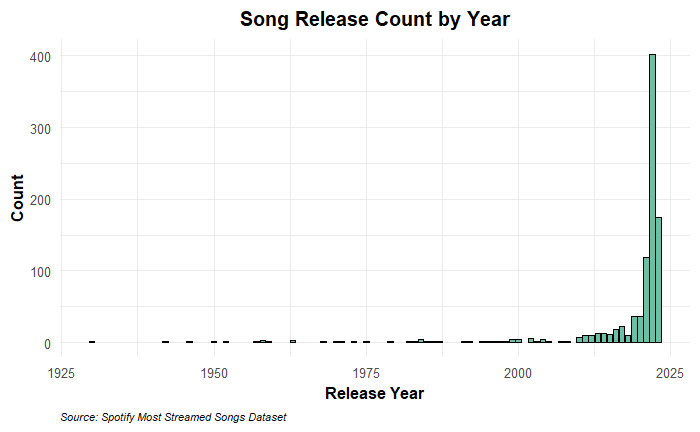
The analysis of the Spotify Most Streamed Songs Dataset revealed key insights into the relationships between song attributes, streaming metrics, and popularity.

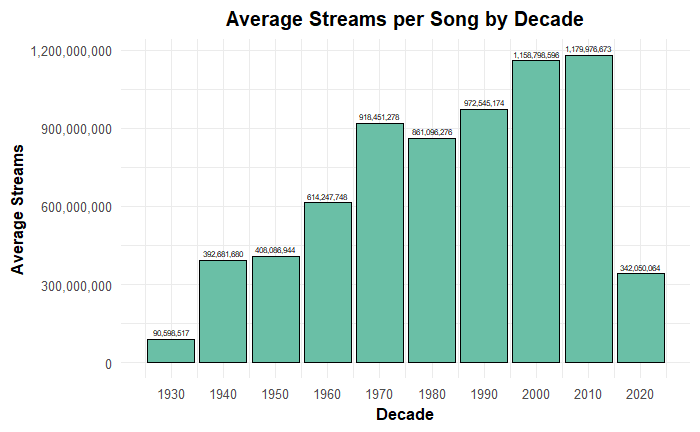
The histogram shows that 67.65% have fewer than 500 million streams, with the highest number of songs falling between 0 and 100 million streams. As the number of streams increases, the number of songs decreases. Only a few songs have over 1 billion streams, and even fewer surpass 2 billion. This shows that a few very popular songs get most of the streams, while most songs have much lower numbers.

A graph of a comparison of streams across playlist count

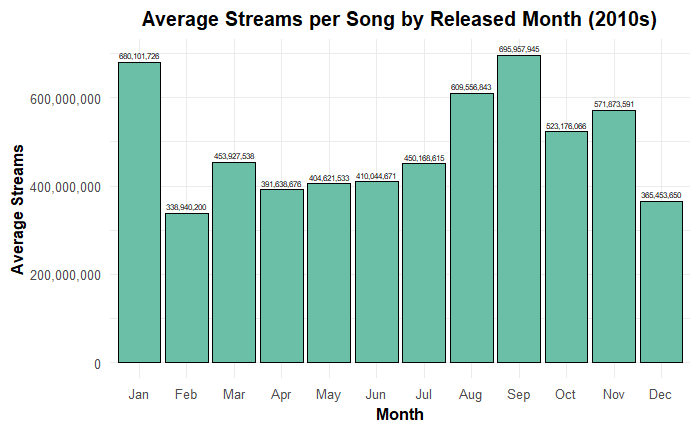
Description automatically generated

This graph shows how the number of playlists a song is added to relates to its total streams. Songs included in more playlists, especially in the 10,001+ range, tend to have much higher streams than songs in fewer playlists. For example, songs in the 0-500 range have very low streams on average, while songs in 10,001+ playlists have much higher streams. The red dots show outliers, meaning some songs perform unusually well, even with fewer playlists. Overall, songs in more playlists are streamed more, making playlist inclusion a key factor in the song’s popularity.

 This graph illustrates the release years of Spotify’s most-streamed songs and their respective counts. As you can see, streaming trends favor newer music, which will likely dominate the chart. Very few songs from before the year 2000 are represented in this dataset. Older songs may not appear on playlists as frequently, which impacts their streaming numbers. Most of the songs in this dataset were released after 2010, suggesting that Spotify’s most-streamed songs are skewed toward more contemporary releases.



This graph shows the average number of Spotify streams for songs categorized by decade. The plot indicates a consistent rise in average streams beginning in the 1930s, reaching a peak in the 2000s and 2010s. The lasting popularity of songs from earlier decades is still apparent, likely due to their continued presence in popular culture. Notable examples of these classics include "White Christmas" and "Rockin' Around the Christmas Tree," along with significant tracks from the 1960s, 1970s, and 1980s, such as "Have You Ever Seen the Rain," "Bohemian Rhapsody," and "Every Breath You Take." Songs released after 2000 show the highest average number of streams, suggesting that these tracks might better resonate with Spotify's listener base. In contrast, the lower average streams for songs from earlier decades indicate a prevailing preference among Spotify users for modern music, indicating that older tracks are less likely to achieve more streaming success.



The graph shows the average streams per song by month for songs released after 2010. By excluding older songs released before Spotify’s availability, the analysis intends to focus on trends specific to Spotify's platform, hoping to provide a clear view of how release timing may impact streaming success. Songs released in January and September show significantly higher average streams compared to other months, and one can interpret that these months may align with strategic release windows. January's high average could be attributed to listeners engaging with new music during the start of the year, while September’s peak may coincide with back-to-school energy. Interestingly, in December, a traditionally significant month in the holiday season, songs released during this month do not exhibit as high average streaming; we attribute this potentially to competition from well-established holiday classics tracks. This plot may suggest that release month could play a critical role in a song's streaming performance. Artists and record labels might benefit from prioritizing releases in January and September to the exposure and inclusions of their tracks on the radar of listeners during peak streaming periods.

A graph with different colored lines

Description automatically generated with medium confidence

This box plot compares the streaming counts of solo artist songs versus group collaborations. While the medians are similar, the spread of streaming counts for group collaborations is slightly larger, suggesting that collaborations have more variability in success. However, both solo and group songs can achieve high streaming numbers, indicating that collaboration alone is not a definitive factor for streaming success.

A graph of a number of years

Description automatically generated with medium confidence

The following visualization shows the relationship between songs' average BPM (beats per minute) and their release years, with bubble sizes representing total streaming counts. The x-axis represents the song's release year, while the y-axis measures the average BPM, capturing the tempo trends over time. From this chart, the average BPM of songs has obviously fluctuated across different decades. The tempo was highly variable in earlier years (pre-1950), with noticeable drops below 100 BPM and subsequent spikes exceeding 140 BPM. As the timeline progresses toward the modern era, post-2000, fluctuations are more moderate, and average BPM stabilizes within a narrower range between 100–140 BPM. The fact that larger bubbles clustered in recent years means that the songs released more recently have significantly higher streaming counts compared to older tracks. The size of the bubbles gives some idea about streaming popularity. Larger bubbles indicate that the tracks in the last twenty years (2000-2025) tend to have higher total streams, probably because of digital streaming platforms and changes in how people listen to music. Though sometimes high in BPM, older songs show smaller bubble sizes, reflecting lower overall counts of streams.

This visualization underlines historical trends in song tempos and the rising dominance of recent music in terms of streaming popularity. Moreover, it indicates that modern music favors a consistent BPM range, considering listener preferences and evolving music production methods.

A graph of a number of bars

Description automatically generated with medium confidence

This plot illustrates the distribution of streaming count depending on the number of artists working on a track, discriminating between solo artists (red) and collaborations/collaboration groups (green). The y-axis represents streaming counts on a logarithmic scale in order to emphasize the differences in streaming performance, while the x-axis represents artist count. From the boxplot, tracks performed by single artists-artist count equals 1-receive high streaming counts with a median close to the billion-stream mark. There is, however, considerable variation, as captured by the wide IQR and outliers below 10 million streams. In the case of collaborations - artist counts above 1-the median streaming counts are consistent for different sizes of collaborating artists, from 2 to 5 artists (although a bit lower than single-artist tracks). This means that even though collaborations obtain outstanding streaming performance, they do not consistently outperform single-artist tracks. Remarkably, when the count of artists increases beyond 5, collaborations become less common; still, the streaming counts remain quite stable, indicating that there are few but successful tracks in this range.

Interesting observations are those of 6-artist collaborations, where the streaming count distribution is increased significantly, reflecting the greater variability of performances. Meanwhile, tracks featuring 7 or 8 artists show narrow distributions, indicating consistently moderate streaming counts with no extreme outliers. While single-artist tracks are likely to be leading in terms of median streaming performance, collaborations still managed to be competitive; indeed, with increasing artist counts, the streaming success remained stable. This power analysis suggests that both single and collaborative tracks find their place in driving significant streams based on factors like artist popularity and song promotion.

A graph with red and purple squares

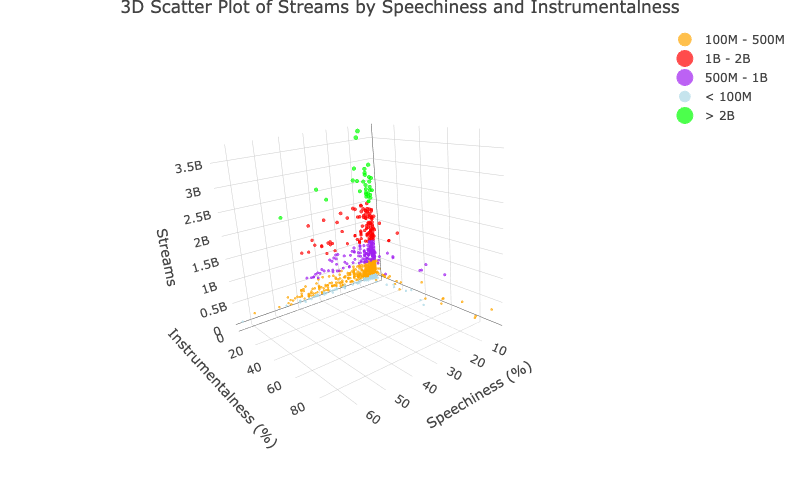
Description automatically generated

The heatmap reveals that musical features like danceability, energy, and valence (positivity) are closely connected, with high-energy songs often being both positive and danceable, creating an appealing combination for uplifting music. However, the weak correlation between streams and these attributes suggests that streaming success depends more on external factors such as playlist inclusion, artist popularity, and marketing efforts rather than just the song’s intrinsic qualities. Acoustic tracks, which show a negative relationship with energy and danceability, cater to a different audience, likely favoring simplicity and calmness over energetic beats. Overall, while musical attributes shape a song's character, external influences play a larger role in determining its popularity.

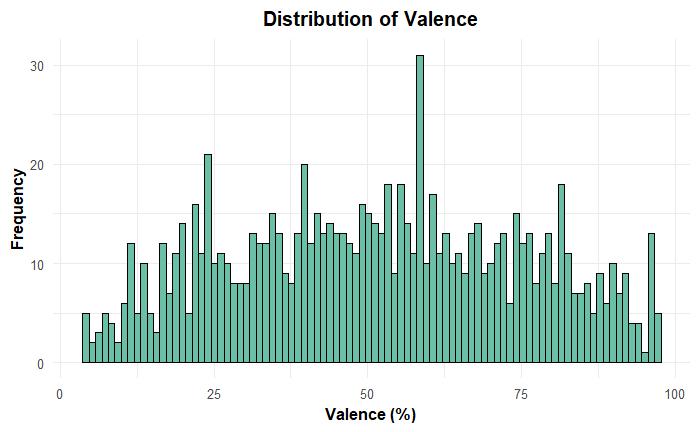
A diagram of a heatmap

Description automatically generated

The heatmap shows that musical features like positivity (valence), energy, and danceability are interconnected, with energetic songs often being both positive and danceable. However, the weak correlation between these features and streams indicates that a song’s musical qualities alone do not significantly drive its popularity.

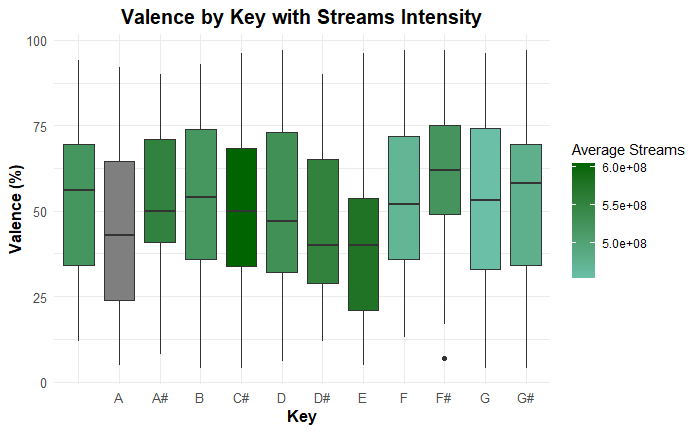


The 3D scatterplot offers an insightful exploration of the relationship between instrumentalness and speechiness and a track's streaming performance. Speechiness has a more varied and noticeable impact on streaming results than instrumentalness, which seems to have a concentrated pattern. The high streaming categories (>1 billion streams) are dominated by tracks with higher instrumentality and minimal speechiness (<10%), which corresponds to the preference for instrumental or barely vocalized genres. .   
  
Tracks with greater speechiness (>50%), on the other hand, typically stay in lower streaming categories (<100 million streams), indicating limited appeal to the general public. There are exceptions, too, when songs with intermediate speechiness (20%–40%) sometimes reach 1 billion streams, suggesting that particular vocal and instrumental combinations are popular with listeners. This analysis underscores that while instrumentalness is present, it is speechiness that plays a more dynamic role in influencing streaming success, particularly through its variability across high and low-stream tracks.



The analysis investigates the relationship between valence (a measure of a track's positivity) and streams (total plays). The scatterplot reveals that tracks with valence values between 40% and 70% dominate the higher streaming range (>1B streams), indicating that moderately positive tracks resonate well with audiences. This range often represents uplifting or feel-good tracks that are not overly sentimental or melancholic. In contrast, tracks with extremely low valence (<20%) or extremely high valence (>80%) predominantly fall in the lower streaming ranges (<500M streams), indicating that highly negative or excessively cheerful tracks have limited mass appeal.

Interestingly, there is considerable variability within the mid-range valence values (40%-70%), suggesting that while positivity contributes to a track's success, other factors like genre, artist popularity, and marketing also play a role. Tracks in this valence range frequently achieve over 1B streams, showcasing their broad appeal. On the other hand, highly negative or melancholic tracks, despite their niche appeal, rarely exceed 1 billion streams, demonstrating that extreme emotional tones may limit mainstream success. These results emphasize the importance of maintaining an emotional balance in tracks to attract a diverse audience and achieve higher streaming performance.



The "Valence by Key with Streams Intensity" plot, created without the 95 missing values in the key column, reveals how emotional tone and streaming popularity vary across musical keys. Keys like A# and D are associated with more positive and upbeat tracks, while keys such as C# and F lean toward a more subdued tone. Interestingly, keys like C# and F, despite their lower positivity, show higher streaming numbers, as indicated by the darker green shading. The variation in emotional tone is also notable, with some keys, such as D# and G#, displaying a wide range of valence, while others, like B, are more consistent. Additionally, a few standout tracks with extreme emotional tones are captured as outliers, particularly in the key F#. These results underscore the diversity in emotional tone and popularity across keys, with no single key dominating both metrics, reflecting varied listener preferences and track characteristics.

A diagram of a data visualization

Description automatically generated with medium confidence

The UMAP displays Spotify, where songs are clustered into six distinct groups based on their musical characteristics. The variables that were used for clustering include bpm (beats per minute), Danceability, valence\_% (positivity of a track), energy\_%, acousticness\_% (acoustic nature), instrumentalness\_% (amount of instrumentation), liveness\_% (live performance indicator), and speechiness\_% (spoken word presence). These features are representative of various sonic and compositional features that might help in the clustering of similar tracks together. UMAP is a very popular modern technique for dimensionality reduction, used for simplifying complex high-dimensional data sets. It projects high-dimensional data with many features (like the Spotify variables discussed above) into a lower-dimensional space-such as the two dimensions in this plot-while preserving as much of the structure in the data as possible. UMAP is a general method for dimensionality reduction that is particularly well-suited for exploratory data analysis; it is computationally efficient and often performs well in revealing cluster or other global structures in data that are not apparent in the high-dimensional representation. Here, it allows us to visualize the relationships between songs and the formation of distinct musical clusters.

In the UMAP visualization, clusters are color-coded (1–6), representing groups of songs with similar musical traits. For instance, Cluster 3 (green) and Cluster 4 (cyan) are well-separated, while Clusters 1 (red) and 6 (pink) are overlapping more, hence more similar. Cluster 2 (yellow) is the smallest group, which brings into focus a niche set of tracks with unique characteristics. UMAP reduces dimensionality by visualizing these clusters that help unfold hidden patterns in Spotify data, such as identifying energetic, acoustic, or highly danceable tracks, which makes it very useful for music analytics and recommendation systems.

**Discussion**

The findings from this study shed light on the factors contributing to a song’s popularity on Spotify. A song’s inclusion in playlists and charts emerged as a significant influence streaming success, with songs featured in more playlists achieving higher streams. These findings highlight the importance of playlist curation, which increases the visibility of songs and their chances of reaching wider audiences. While some musical attributes—such as energy, danceability, and valence—are interconnected and contribute to a song's popularity, they seem to have limited impact on total streams. We believe that external factors, such as marketing efforts, an artist's popularity, and other social media platforms, may play a more significant role in driving streams.

Our analysis of this dataset shows some interesting trends in streaming numbers. A handful of songs have incredibly high streams, while most others fall into a cluster of similar streaming numbers. It’s no surprise that tracks from the 2010s dominate—this era coincides with the boom of digital music platforms, which have completely transformed how we consume music. We also noticed that songs with moderate energy and valence tend to do well, therefore suggesting listeners tend to seek balanced, feel-good music. At the same time, we found there are some big outliers tracks with massive streaming numbers, which we attribute to cultural moments or viral trends.

Overall, this study highlights just how many factors come into play when it comes to a song’s streaming success. Musical traits matter, but so do external factors like marketing and cultural relevance. In the future, we suggest driving deeper into how things like social media, marketing strategies, and regional trends shape these patterns. As well as continuing exploring the effect of collaborations, genre preferences, and regional trends to understand better what drives success in the streaming era. This report offers a useful guide for artists, curators, and industry professionals seeking to optimize their strategies in the digital music ecosystem.

**References**:

[Kaggle](https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs)

[Stack Overflow](https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs)

[Musical Keys](https://www.classicfm.com/discover-music/music-theory/what-are-musical-keys/)

[ResearchGate – Emotional Music Classification](https://www.researchgate.net/figure/Emotional-classification-of-music-by-energy-and-valence-parameters-with-exemplary-songs_fig2_334084112)