Project 1 -- Group 4 Write Up

Matthew Adent, Khalil Locke, Chase Mueller, Marcanthony Solorzano

Data Boot Camp

Prof. Alexander Booth

Introduction

Whether we realize it or not, music plays a profound role in shaping our daily lives. It has the unique ability to influence our emotions, bringing comfort in times of sorrow and amplifying joy during moments of celebration. Music has the power to heal, unite, and inspire. With these ideas in mind, we knew music would be an engaging and meaningful focus for our project. After all, who doesn't love music? It's rare to find someone who doesn't enjoy it in some form

After exploring several options, we selected the Spotify and YouTube dataset from Kaggle, created by Salvatore Rastelli et al. This dataset stood out for its clarity, organization, and wealth of information. It provided an excellent foundation for us to ask interesting, data-driven questions while staying within the dataset's scope. This dataset is an excellent starting point for aspiring data analysts due to its accessibility and versatility.

Beyond our love for music, we were curious about the differences between YouTube and Spotify users. This dataset offered rich insights into each platform's emotional tones and popularity trends. By analyzing this information, we aimed to uncover how digital streaming shapes global music trends. Both Spotify and YouTube play significant roles in driving the music industry's revenue, making them essential to understanding consumer behavior and listening preferences. Insights from this analysis could inform strategies for artists and record labels, helping them tailor their content to better resonate with audiences on each platform.

Introduction to the Data Set, Data Cleaning, and Feature Engineering

The dataset contains 20,718 unique tracks. It also includes 28 columns for each track for us to analyze. The columns within the CSV are Unnamed, Track, Artist, Url_spotify, Album, Album_type, Uri, Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Duration_ms, Stream, Url_youtube, Title, Channel, Views, Likes, Comments, Description, Licensed, and official_video. After reviewing and discussing our questions, we decided that five columns were irrelevant to our work. Those columns were Unnamed, Url_spotify, Uri, Url_youtube and Description.

Once these columns were dropped, we were left with the following for our dataset.

```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Artist             20718 non-null  object
 1   Track              20718 non-null  object
 2   Album              20718 non-null  object
 3   Album_type         20718 non-null  object
 4   Danceability       20716 non-null  float64
 5   Energy             20716 non-null  float64
 6   Key                20716 non-null  float64
 7   Loudness           20716 non-null  float64
 8   Speechiness        20716 non-null  float64
 9   Acousticness       20716 non-null  float64
 10  Instrumentalness   20716 non-null  float64
 11  Liveness           20716 non-null  float64
 12  Valence            20716 non-null  float64
 13  Tempo              20716 non-null  float64
 14  Duration_ms        20716 non-null  float64
 15  Title              20248 non-null  object
 16  Channel            20248 non-null  object
 17  Views              20248 non-null  float64
 18  Likes              20177 non-null  float64
 19  Comments           20149 non-null  float64
 20  Licensed           20248 non-null  object
 21  official_video     20248 non-null  object
 22  Stream             20142 non-null  float64
```

From there, we broke the dataset down further to best suit our individual needs for the project and the questions we were working on.

Question 1 - How do the features differ between the top 15% of songs on YouTube and Spotify?

The first question dives into the differences between the most popular songs on YouTube and Spotify, focusing on their audio features and comparing acousticness between the platforms. Examining the top 15% of songs based on views and streams reveals interesting trends and insights about what makes a hit on each platform. We calculated the 85th percentile for views and streams to identify high-performing songs, the threshold for the top 15%. Songs above these thresholds were separated into two groups: one for YouTube and one for Spotify, with each dataset labeled by platform.

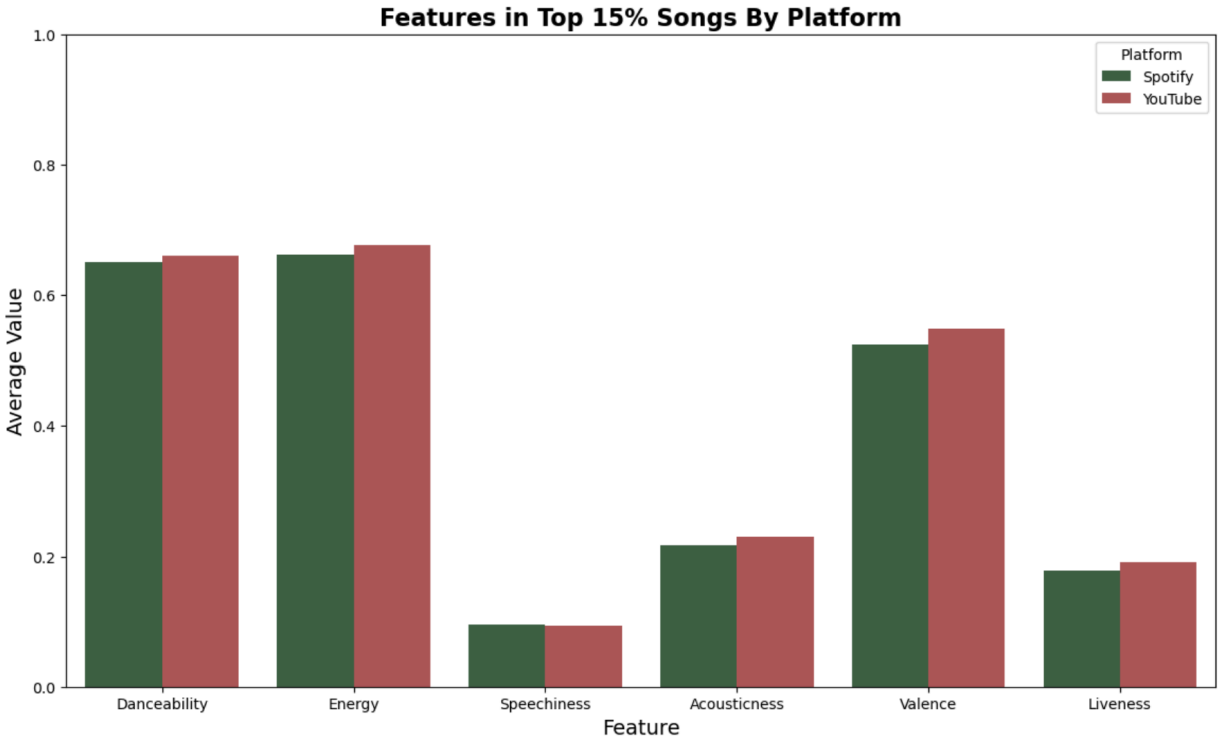
Feature Means by Platform:

	Platform	Danceability	Energy	Loudness	Speechiness	Acousticness	Valence	Liveness	Tempo
0	Spotify	0.649855	0.661293	-6.438411	0.095405	0.216407	0.524071	0.177659	121.120469
1	YouTube	0.659846	0.677289	-6.087433	0.094407	0.230838	0.548021	0.191912	121.907129

A statistical t-test was used to see if there was a significant difference in the "acousticness" (a measure of how acoustic a song sounds) between the top YouTube and Spotify songs. The test checked whether this difference could just be random noise. If the p-value from the test was less than 0.05, the difference was considered accurate and significant. If the p-value was 0.05 or higher, it suggested no meaningful difference between the platforms. Our results determined a statistical significance in the difference in the average acousticness.

```
P-Value: 0.018
The difference in average acousticness in the top 15% of Spotify vs YouTube songs is statistically significant.
```

Beyond acousticness, we calculated the average values for audio features like Danceability, Energy, Tempo, and more for both YouTube and Spotify. This helped paint a broader picture of the songs dominating each platform. To make the comparisons easier to understand, we used bar charts. Most features were shown on one plot, while Loudness and Tempo (which have very different scales) were displayed separately. These visuals highlight how the platforms differ in terms of song characteristics.



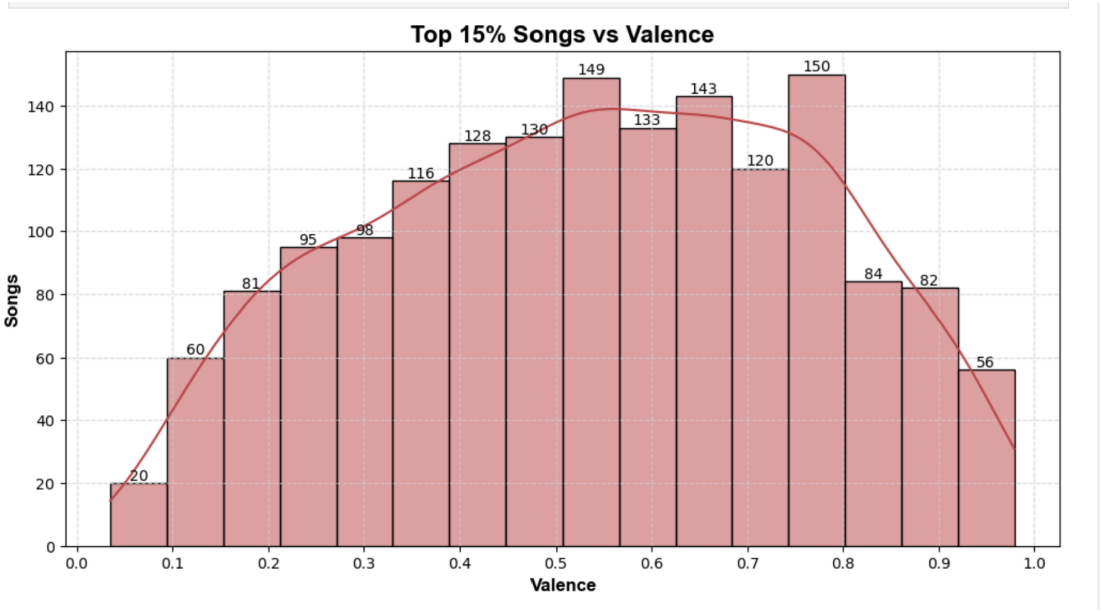
The t-test results, feature averages, and visualizations reveal apparent differences between popular songs on YouTube and Spotify. Popular songs on YouTube are more energetic, loud, and generally upbeat than Spotify's most popular songs. By understanding these patterns, musicians, content creators, and industry professionals can better tailor their work to match the unique audience preferences of each platform.

Question 2 - How are the emotional tones of top-streamed and most-viewed songs distributed?

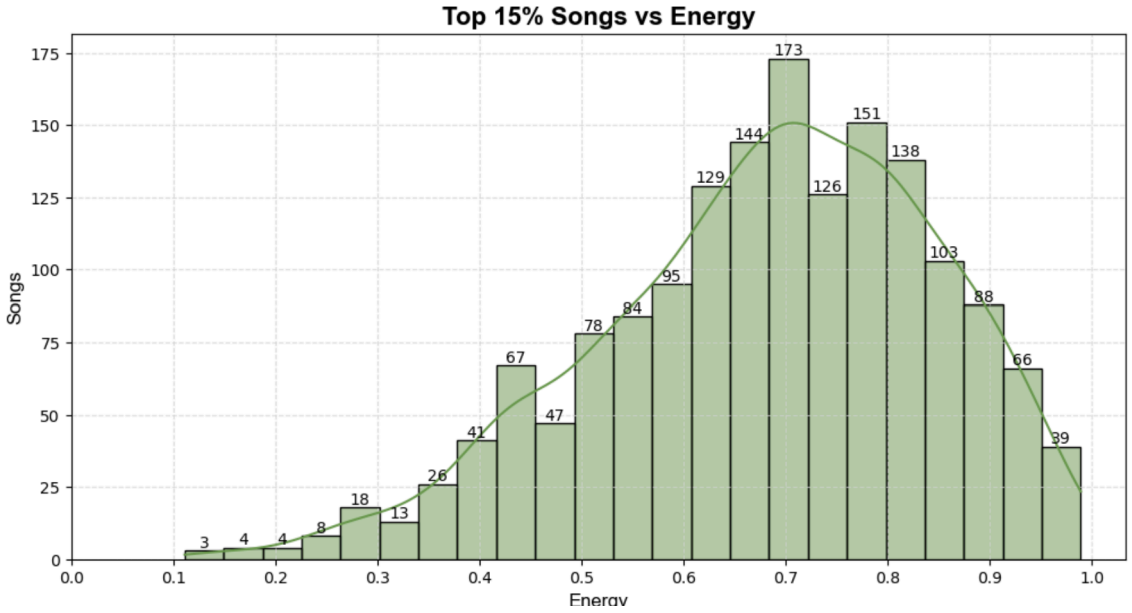
Our goal was to understand how these emotional tones are distributed among popular songs and to identify any trends or patterns in their appeal.

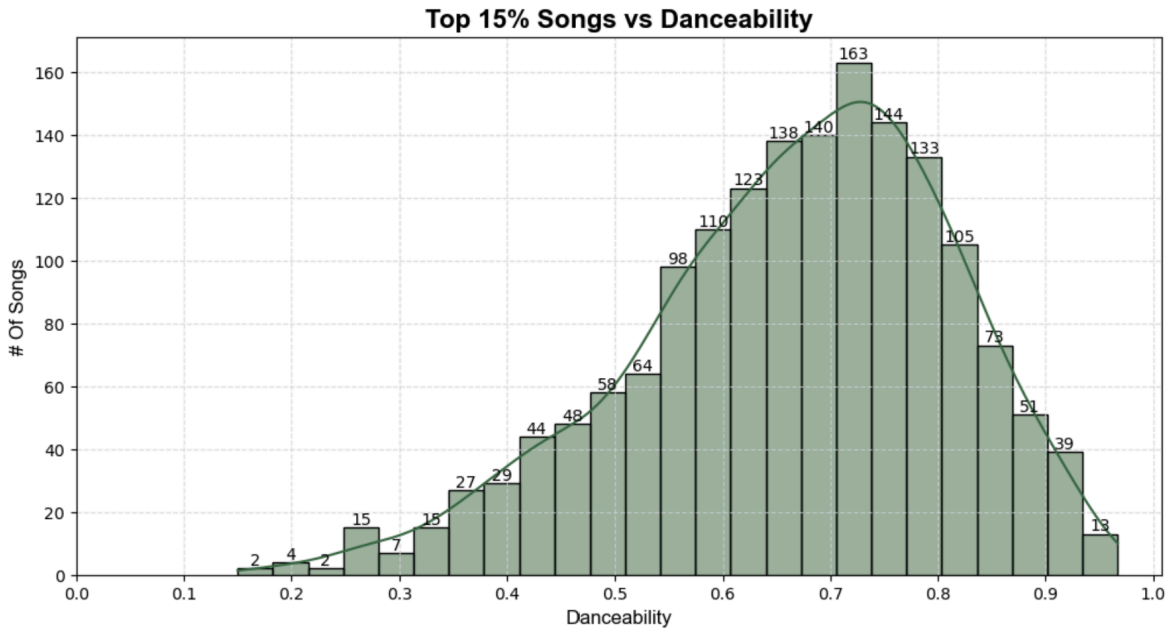
To focus our analysis, we chose the top 15% of most streamed and most viewed songs because these tracks represent the most popular and widely listened to music across platforms, giving us insight into the songs that resonate most with audiences. We specifically focused on valence, danceability and energy because these attributes capture a song's emotional tone, physical engagement and intensity, which are key elements that likely influence its popularity.

Starting with the valence histogram, valence measures how positive or negative a song sounds, with scores ranging from 0 to 1. A lower score represents a more sad, depressing, or negative tone, while a higher score reflects a happier, upbeat, or more positive tone. The first peak occurred between 0.5 and 0.6, and the second was 0.7 and 0.8. These peaks suggest that popular songs often have normal to high valence, meaning they generally sound positive or uplifting. However, the range also has diversity, showing that songs with slightly lower valance can still gain popularity.

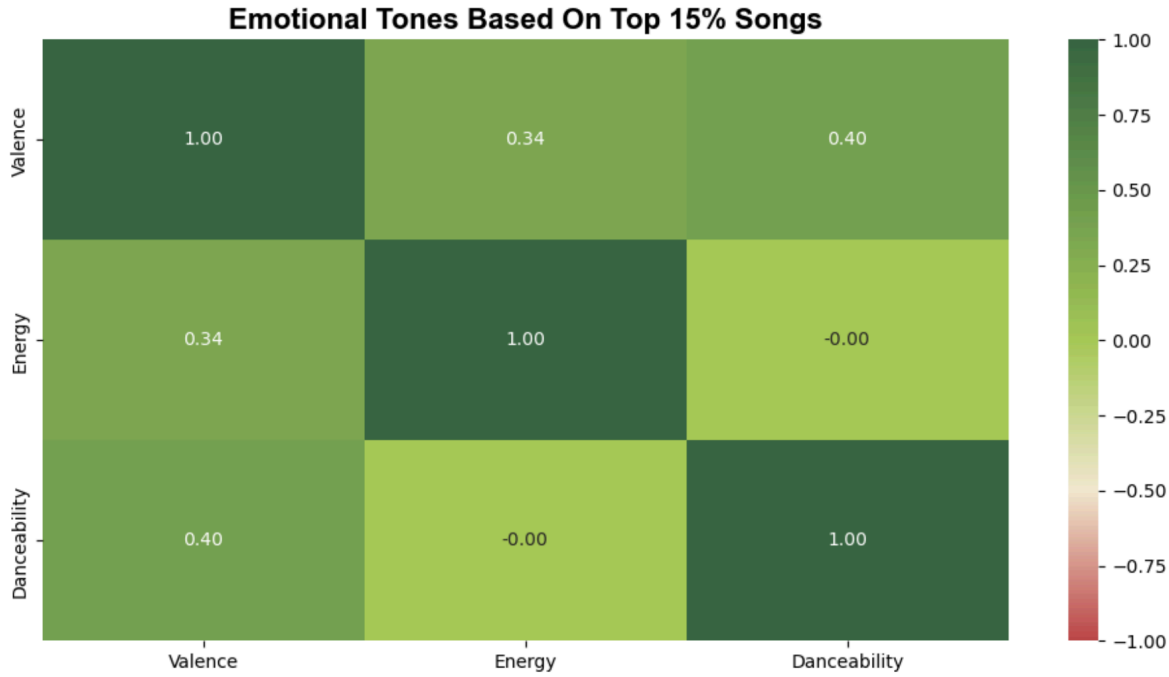


As we can see, we have a different story. Both of these features displayed a left-skewed distribution. While most popular songs maintain moderate to high energy and danceability, very low-energy or danceability songs are rare. For example, the energy histogram peaked around 0.7, with a decline as energy approached 1.0, indicating that while high-energy tracks are popular, excessively high energy levels might not be as appealing to the general audience. This is similar to danceability, which peaked between 0.7 and 0.8, suggesting that songs people can move tend to dominate the popular charts.





Lastly, we created a heat map. This aligns with what we observed in the histograms: songs with higher valence often have more energy and are easier to dance to. While the correlation isn't substantial, it suggests a general trend that helps explain the appeal of popular songs.



Altogether, these findings suggest that audiences gravitate toward songs that balance positive emotional tones with normal to high energy and danceability. Tracks that can energize, uplift, and engage listeners are more likely to dominate both Spotify and YouTube charts.

The distribution of emotional tones among the top 15% of songs across both platforms highlights the power of upbeat and high-energy music. But at the same time, there is enough diversity to show that songs with slightly more sad tones can still find their place in the spotlight.

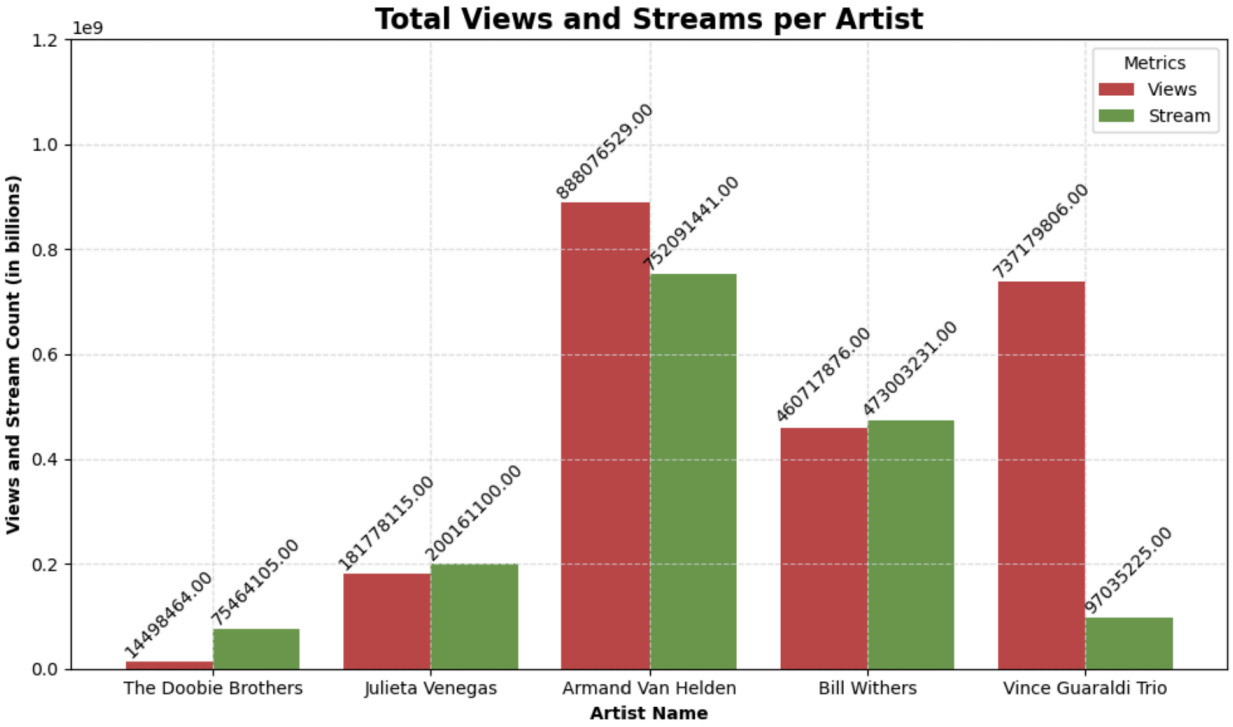
Question 3 - What kind of correlation do the features of a song have with each other?

The next step in our analysis was to explore the correlations between songs and their features. To do this, we examined two key aspects: the relationship between total views and streams for a random sample of artists and the broader correlations between song features across the dataset. We began by analyzing the total views and streams for a random sample of five artists in the dataset. This comparison aimed to show how different artists are received on YouTube versus Spotify. To visualize the data, we used a grouped bar chart, which displayed the aggregated views and streams side by side for each artist. The five randomly selected artists and their primary genres were as follows:

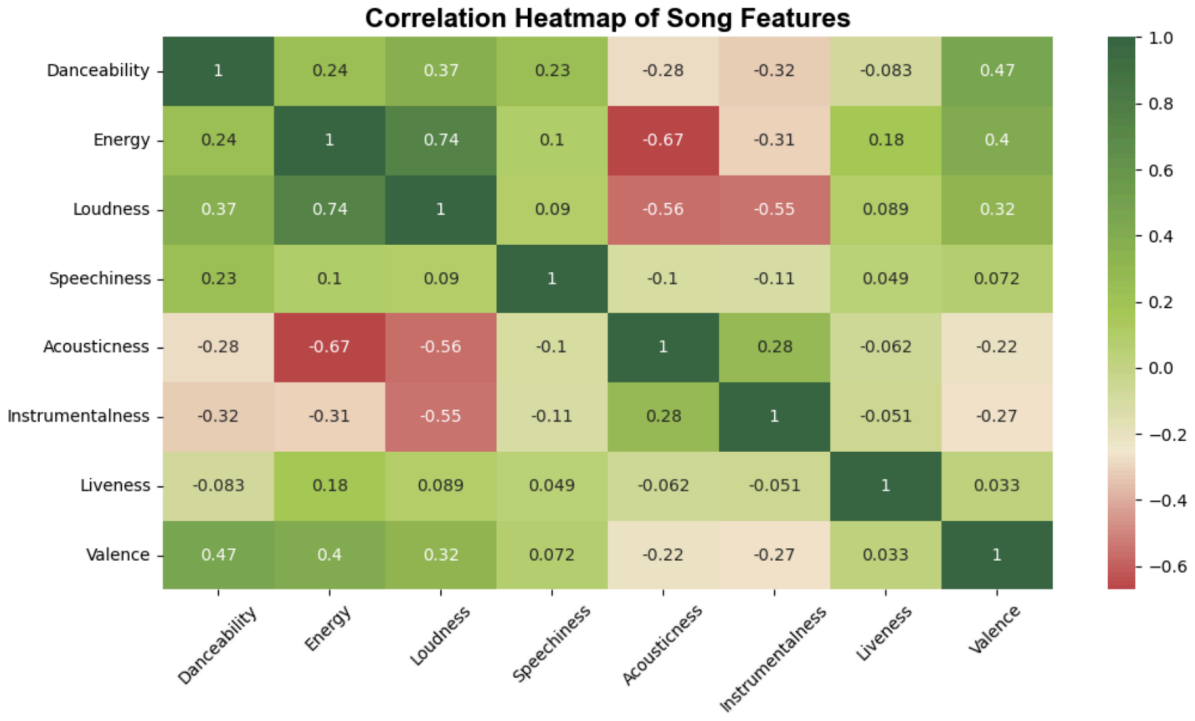
- Doobie Brothers - Soft Rock
- Julieta Venegas - Latin Pop
- Armand Van Helden - EDM
- Bill Withers - R&B/Soul
- Vince Guaraldi Trio - Jazz

Although the dataset didn't include genre information, a quick search allowed us to identify each artist's primary style. With this additional context, the variation in views and streams across the sample was unsurprising. However, one standout observation was the

significant disparity between YouTube views and Spotify streams for The Vince Guaraldi Trio. Known for their iconic *Charlie Brown Christmas* album, their popularity on YouTube surges during the holiday season. Songs like *Linus and Lucy* are staples of holiday playlists and are often played repeatedly, driving up YouTube views. This seasonal pattern doesn't translate to Spotify streams, which likely explains the difference in platform performance.



After examining the random sample of artists, we analyzed the overall relationships between song features. This broader analysis aimed to uncover patterns and correlations among attributes like Danceability, Energy, and Tempo across the entire dataset. To visualize these relationships, we used a heatmap. The heatmap allowed us to display all the features together in a single, intuitive view, making it easier to identify strong or weak correlations at a glance. This dual approach of comparing artist-specific platform performance and analyzing overall feature correlations provided a comprehensive understanding of the dynamics within the dataset.

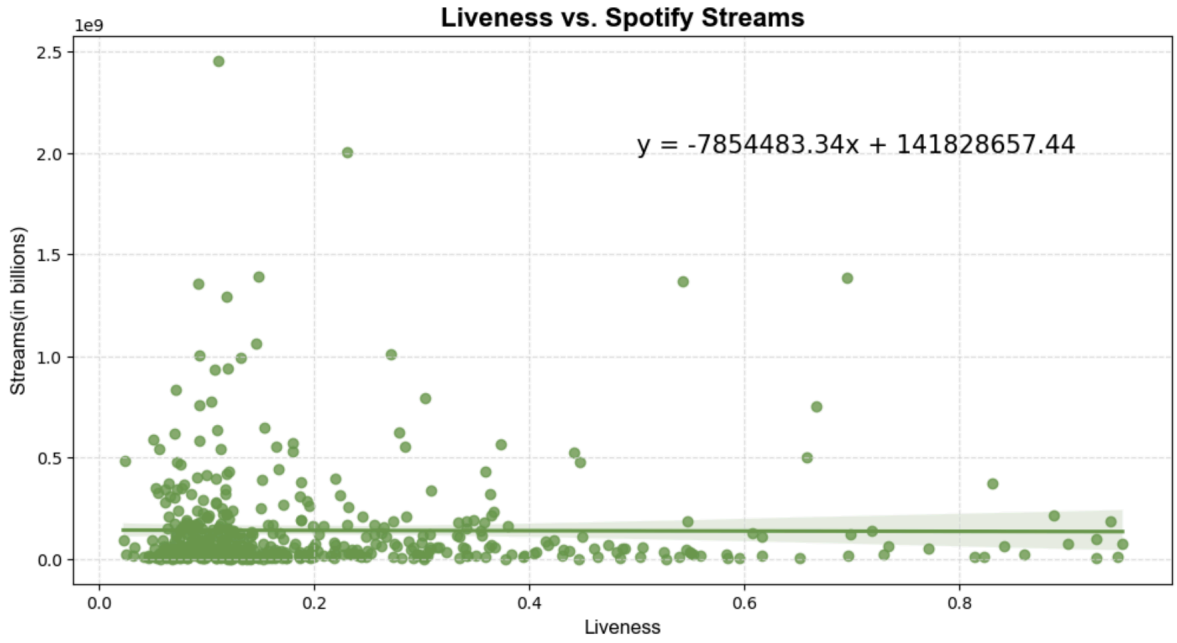


The heatmap above highlights positive and negative correlations between the features of songs in the dataset, providing valuable insights into the relationships between various attributes. One unsurprising finding is the negative correlation between a song's energy and acousticness. Acoustic tracks are typically not designed to energize or excite listeners; instead, they emphasize raw emotion and authenticity, encouraging listeners to focus on the lyrics and the song's more profound meaning. This contrast likely arises because high-energy tracks rely on electronic or instrumental elements to drive movement and excitement. In contrast, acoustic songs prioritize a more subdued, introspective experience. The negative correlation suggests that, generally, as a song becomes more acoustic, its energy decreases, reflecting the distinct purposes these two features serve in shaping the listening experience.

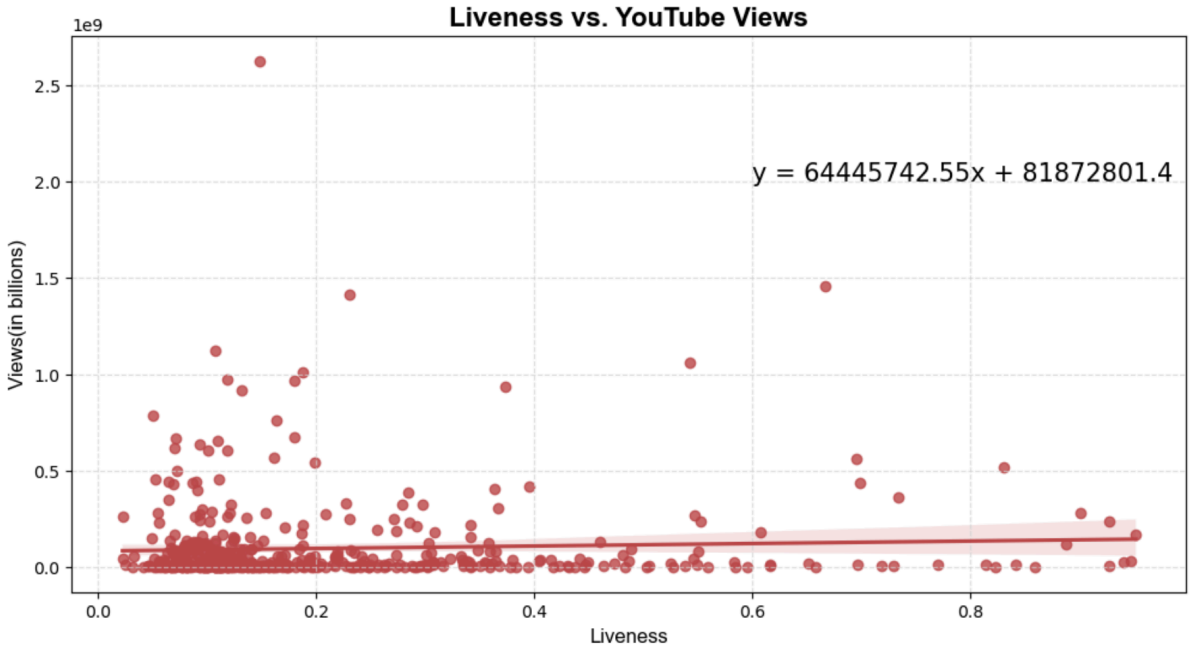
Additionally, this negative relationship emphasizes the differing emotional tones these features evoke. While high-energy songs are more likely to create an atmosphere of excitement or urgency, acoustic songs foster a sense of calmness or emotional depth. Understanding this relationship can help artists, producers, and platforms better tailor music to the emotional responses they wish to elicit from listeners, whether energizing them for a workout or providing a reflective mood for quiet moments.

Question 4 - Are high-liveness songs generally more popular on YouTube, and are low-liveness songs generally more popular on Spotify?

We performed a linear regression analysis to understand how "Liveness" relates to song performance on Spotify and YouTube. This approach allowed us to explore whether variations in "Liveness," which measures the extent to which a song has a live-performance feel, impact the number of streams on Spotify and views on YouTube. The relationship between "Liveness" and Spotify streams was analyzed by plotting "Liveness" (x-axis) against the total number of Spotify streams (y-axis, in billions) and fitting a linear regression model. The resulting equation for the regression line was y = −7,854,483.34x+141,828,657.44. The scatterplot, which included a regression line with a 95% confidence interval, revealed a clear trend: as "Liveness" increases, the number of Spotify streams decreases. This negative correlation suggests that songs with higher "Liveness" scores—often characterized by the ambiance of live performances—are less favored by Spotify users. This finding aligns with the platform's focus on polished, studio-recorded tracks, typically preferred for casual or focused listening experiences.



A similar analysis examined the relationship between "Liveness" and YouTube views. The regression model produced the following equation: y=64,445,742.55x+81,872,801.4 The scatterplot and regression line, again with a 95% confidence interval, showed a positive correlation. Songs with higher "Liveness" scores tended to have more views on YouTube. This trend may reflect YouTube's appeal as a platform for live performances and visually engaging content. Unlike Spotify, YouTube offers a more dynamic medium where the energy and authenticity of live music can enhance user engagement.



Call to Action

Several trends emerge from the dataset that artists can leverage to maximize engagement across platforms:

1. **High-Energy Music:** For music that is more live, energetic, or high-energy, YouTube is the ideal platform for engagement. The visual and dynamic nature of the platform aligns well with the appeal of such tracks.

2. **Relaxed Listening:** Spotify is likely the better platform for more comfortable and casual listening experiences. Its emphasis on curated playlists and continuous streaming makes it a preferred choice for users seeking laid-back audio experiences.

These insights suggest that emerging artists may want to tailor their music to the specific strengths of each platform, potentially optimizing their chances of creating a hit based on the platform's unique audience preferences.

Bias and Limitations

Several limitations of the dataset should be acknowledged to understand the scope and potential biases of the analysis:

1. **Multiple Platform Users:** The dataset does not account for users who engage with both Spotify and YouTube, meaning that views and streams are counted independently across platforms without considering any overlap. This could lead to overestimating engagement metrics for users who frequently switch between the two platforms.
2. **Absence of Genre Information:** The dataset lacks genre information for each song, restricting any analysis based on genre classification. As a result, we could not explore how different musical genres perform across platforms or how they may influence user behavior.
3. **No Timestamps for Views or Streams:** The absence of timestamps for when views or streams occurred makes it impossible to track trends over time. Without this temporal data, we could not analyze the seasonal or longitudinal patterns of song popularity on either platform.
4. **Geographic Data Missing:** The dataset must include information on the regions where the songs are most popular. This lack of regional context could skew the results, as higher user activity in certain areas (such as regions with large user bases on Spotify or YouTube) may disproportionately influence the findings.

Future Work

Future research could benefit from several enhancements to deepen the analysis:

1. **Incorporating Temporal Data:** Introducing dates into the dataset would allow for analyzing how emotional tones and streaming/viewing patterns evolve. This would help identify seasonal or event-driven trends, such as increased engagement during holidays or special events.
2. **Exploring Demographic Data:** By incorporating user-specific metadata, such as age, location, or gender, it would be possible to examine how different demographics engage with music. This could provide insights into whether emotional tones resonate differently across various user groups and offer a more nuanced understanding of platform usage.
3. **Developing a Predictive Model:** A predictive model incorporating features like energy, danceability, and valence could help forecast a song's likelihood of high engagement on YouTube and Spotify. This would enable more accurate predictions of a song's performance based on its acoustic characteristics and could inform content strategies for artists and record labels.

References

Pandas Development Team. Pandas Documentation.
https://pandas.pydata.org/docs/reference/frame.html.

"Python | Pandas DataFrame." GeeksforGeeks.
https://www.geeksforgeeks.org/python-pandas-dataframe/.

OpenAI. "ChatGPT." OpenAI, https://openai.com/chat/.

Module 5 & 6, Data Analytics Bootcamp.

Spotify and Youtube - Kaggle Dataset
https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube