ALY 6015 - Intermediate Analytics

Project Report

Analysis of Songs and its presence on YouTube and Spotify

Submitted By -

Devi Somalinga Bhuvanesh

Gurrup Kaur Gurbir Singh Soi

Vy Hoang Mai Duong

Northeastern University

Seattle

Professor Ghazal Tariri

Table of Content

Analysis of Songs and its presence on YouTube and Spotify

1. INTRODUCTION

Analysis of songs can provide valuable insights into the music industry and consumer preferences, making it a crucial field of study. Understanding the characteristics that make a song popular, such as the tempo, valence, and energy, can help music producers and marketers create music that appeals to consumers. Nowadays, Spotify and YouTube are considered among the most known platforms for distributing and consuming music and video content globally.

By analyzing the mainstream content on these two platforms, music industry professionals can also gain insights into the music consumption habits of consumers. The project analyses the characteristics of various songs from different artists worldwide, including several statistics of the audio version on Spotify and the official music video of the song on YouTube. There are 28 variables provided for each of 20,718 songs, including the name of the song, artist, album, danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration, number of streams, YouTube video information such as title, channel, views, likes, comments, description, licensed, and official video.

2. AIM

The aim of this study is to analyze the characteristics of songs and factors that contribute to their popularity on Spotify and YouTube, two major platforms for music distribution and consumption. By conducting an in-depth analysis, the study aims to provide valuable insights into the music industry, consumer preferences, and the impact of different factors on the reception of songs in the streaming era.

3. OBJECTIVES

- Explore the characteristics of songs on Spotify and YouTube, including danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and duration.
- Investigate the relationship between song characteristics and their popularity, using **linear regression analysis**. Identify the factors that significantly influence the popularity of songs and propose a suitable model for predicting song popularity.
- Analyze the effect of licensing on views by employing **logistic regression analysis**. Examine whether there is a significant difference in views between licensed and unlicensed songs on YouTube.
- Compare the reception of artists on YouTube and Spotify using **hypothesis testing of paired t-tests.** Determine if there is a meaningful difference in how artists are received on these two platforms.
- Provide valuable insights and recommendations for music industry professionals, producers, and marketers based on the findings of the analysis.

4. METHODOLOGY

The dataset utilized in this analysis was obtained from the Kaggle platform on April 19th, 2023. It is important to note that the information regarding each song in the dataset was collected on February 7th, 2023, indicating that the dataset is time sensitive. Based on the available dataset, the data is cleaned followed by a detailed data analysis process in the Python platform. An initial exploratory data analysis is carried out. Further, the analysis will employ three key Analytical methods; including:

- **Linear Regression Analysis** to uncover the potential relationship between the song characteristics and their popularity, then propose a suitable model to predict the song's popularity based on different factors.
- **Logistic Regression** is used to analyze the effect of licensing on views, using the number of views as the predictor variable and whether a song is licensed or not as the response variable.
- **T-tests are** employed to test whether there is a meaningful difference in how artists are received on YouTube compared to Spotify.

5. DATA CLEANING

The given dataset consists of information with **20,718 rows** of songs, its features, and their presence on Spotify and YouTube with **28 different attributes.**

5.1. Deletion of columns

- The first column in the original database is just a row index which is unnamed and not required for the analysis. In Python, an automatic row index value starting from "0" is generated. Hence, to avoid duplication and confusion, it was removed using the 'drop' function.
- The URL links of songs are not required for analysis. Therefore, two columns consisting of URL, i.e., Url_Spotify, and Url_Youtube are removed from the database.

Overall, 3 columns are removed from the database.

5.2. Duplication of values

Two instances were observed while examining the dataset -

1. Multiple artists have performed the same song/ track resulting in duplication of rows with the same song and its features.
2. The same song/track was also performed by different artists but had different features as well on YouTube and Spotify.

To identify the uniqueness of the song/track and its feature of the song, ‘Uri’ column helped to identify the unique rows. Therefore, unique songs and features were merged with "Uri" as string delimited by "," and saved in merged_track_data.

5.3. Missing Values

The database contains information on 18,862 songs, which includes 11 features such as Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, and Duration. However, two of the songs in the database do not have data recorded for these features.

Furthermore, the music video for each track is streamed on both YouTube and Spotify, and various attributes are recorded to measure their reach. When a video is streamed on YouTube, eight attributes are recorded, while Spotify records one feature. However, 404 tracks out of 18,862 were not streamed on YouTube, as indicated by missing data in fields such as "Title," "Channel," "Views," "Licensed," and "Official Video." Out of the 18,458 videos available on YouTube, 69 do not have information on likes, 96 do not have comments, and 384 do not have a description. Additionally, 550 out of the 18,862 songs are not present on Spotify.

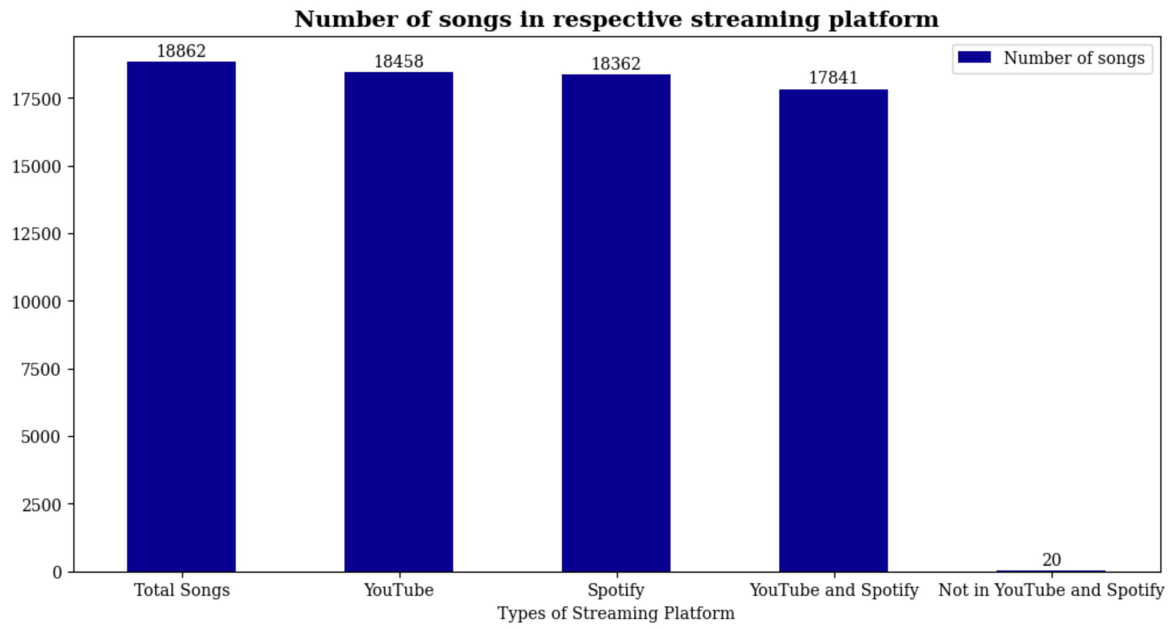


Figure 1. Number of songs in respective streaming platforms

Overall, of the 18,862 songs in the database, 95% (17,841) songs are available on both platforms. However, there are 20 songs (0.10%) that cannot be found on either YouTube or Spotify.

5.4. Outliers

To determine whether there is any significant variation in the features of the songs on YouTube and on Spotify, an outlier analysis was carried out. Outliers refer to values that are significantly different from the rest of the data points in a dataset. In the overall database, it was observed that YouTube comments of video songs have the farthest outlier from its mean, followed by duration of the video. There is no huge variation or outlier observed in three features of the songs, i.e., Key, Acousticness, and Valence.

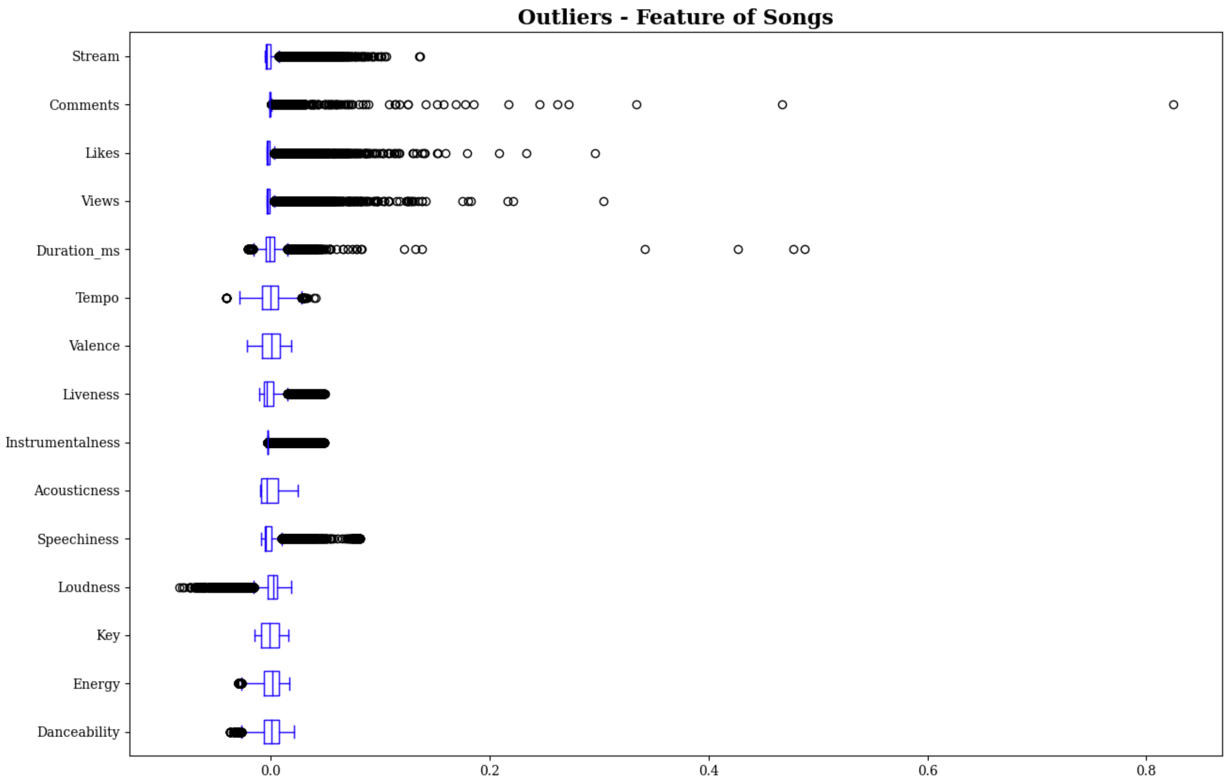


Figure 2. Outliers on Features of Songs

Table 1: Outlier Analysis for Features of Songs

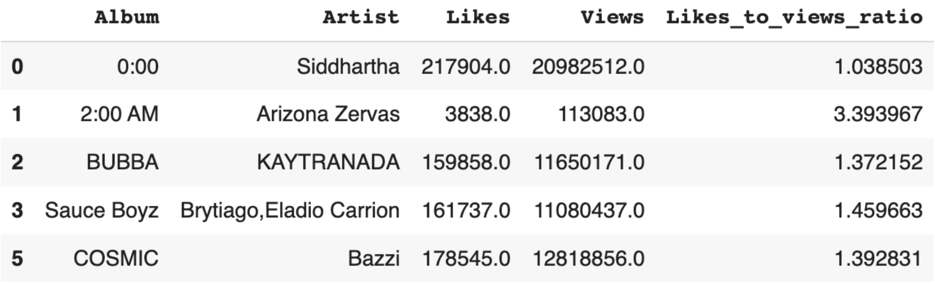
Features of Songs	Number of Outliers
Instrumentalness	4022
Speechiness	2405
Views	2378
Comments	2355
Likes	2336
Stream	2011
Liveness	1331
Loudness	1131
Duration_ms	692
Energy	281
Danceability	213
Tempo	57

The list above shows the number of outliers for various features of songs. The feature with the highest number of outliers is Instrumentalness, with 4022 songs having values that deviate significantly from the rest. Speechiness (2405), Views (2378), Comments (2355), Likes (2336), and Stream (2011) follow as the next five features with a high number of outliers. The least number of outliers were found in Tempo (57) features of the songs, followed by Danceability (213) and Energy (281).

5.5. New Data Column Added

A new column was added, that is, 'Likes_to_views_ratio', provides valuable information about the engagement and reception of songs on the respective platforms. It measures the relative proportion of likes received compared to the number of views a video has garnered.

Table 2: Inclusion of new column "Likes_to_views_ratio"



	Album	Artist	Likes	Views	Likes_to_views_ratio
0	0:00	Siddhartha	217904.0	20982512.0	1.038503
1	2:00 AM	Arizona Zervas	3838.0	113083.0	3.393967
2	BUBBA	KAYTRANADA	159858.0	11650171.0	1.372152
3	Sauce Boyz	Brytiago,Eladio Carrion	161737.0	11080437.0	1.459663
5	COSMIC	Bazzi	178545.0	12818856.0	1.392831

The 'Likes_to_views_ratio' provides insight into the level of audience engagement, appreciation, and reception of songs on the respective platforms, helping to identify songs that have resonated strongly with viewers and generated a higher proportion of likes compared to their view count.

6. DATA ANALYSIS

After removing irrelevant columns, duplicate values, and adding a new column, the dataset now consists of 18,862 records with 26 variables.

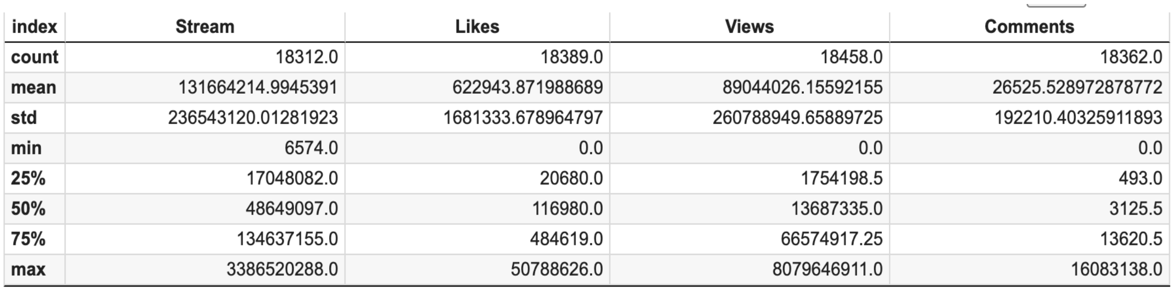
6.1 EXPLORATORY ANALYSIS

Exploratory analysis plays a crucial role in the study analysis by providing initial insights and understanding of the data.

6.1.1. Descriptive Key Analysis

Statistical characteristics on key indicators such as Stream, likes, views, and comments are chosen to understand their features which will help in further analysis.

Table 3: Descriptive Statistics on Stream, Likes, Views, Comments



index	Stream	Likes	Views	Comments
count	18312.0	18389.0	18458.0	18362.0
mean	131664214.9945391	622943.871988689	89044026.15592155	26525.528972878772
std	236543120.01281923	1681333.678964797	260788949.65889725	192210.40325911893
min	6574.0	0.0	0.0	0.0
25%	17048082.0	20680.0	1754198.5	493.0
50%	48649097.0	116980.0	13687335.0	3125.5
75%	134637155.0	484619.0	66574917.25	13620.5
max	3386520288.0	50788626.0	8079646911.0	16083138.0

Some of key statistics regard songs and videos performance include:

- Spotify streams peak at approximately 3.4 billion counts, average stream counts of 18,312 tracks is 131,664,214.

- YouTube views have the highest maximum value in all indicators, at more than 8 billion counts, but averaging at 89,044,026 which is lower than Spotify stream counts.
- All the indicators appear to be heavily right skewed with mean values much higher than median values.

6.1.2. Type of Songs

The presented pie chart illustrates the distribution of song types within the analyzed dataset, based on the 'Album_type' feature. This information can be useful in gaining insights into the overall structure and composition of the dataset, as well as in identifying which types of songs are most common in the music industry.

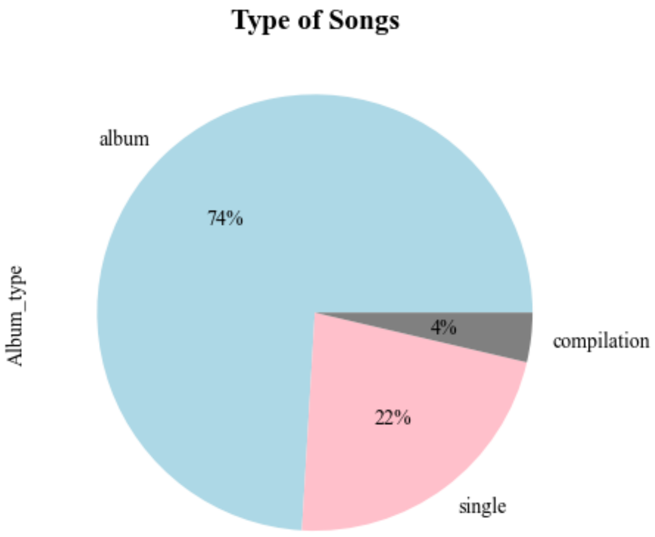


Figure 3. Type of Songs

It can be observed that a significant proportion of the analyzed dataset, approximately 74%, consists of songs categorized as 'Album'. In contrast, only 22% of the dataset comprises 'Single' songs, while 'Compilation' songs account for a mere 4%. These findings indicate that the dataset predominantly consists of songs that are part of a larger album, as opposed to standalone singles or collections of songs from diverse sources.

6.1.3. Type of Songs - Views on YouTube vs Stream on Spotify

The bar chart below illustrates the comparison between the total number of views on YouTube and streams on Spotify for different types of songs.

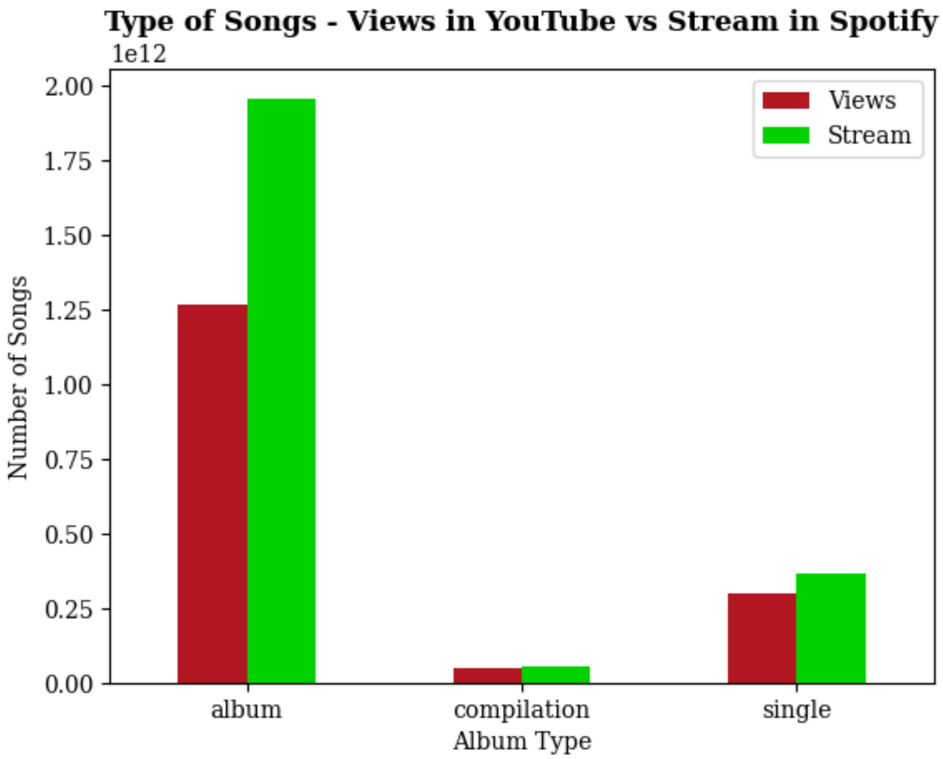


Figure 4. Type of Songs – Views in YouTube vs Stream in Spotify

Overall, it can be observed that there are more streams on Spotify than views on YouTube for each album type. Additionally, the album type with the highest number of streams and views is 'Album', followed by 'Single' and 'Compilation'. Moreover, it can be seen that the variation in the number of streams and views is greater for 'Album' compared to the other two album types.

Finally, for the 'Compilation' album type, there is almost an equal number of streams and views, indicating a balanced distribution of audience reach across the two platforms.

6.1.4. Official vs Unofficial video songs

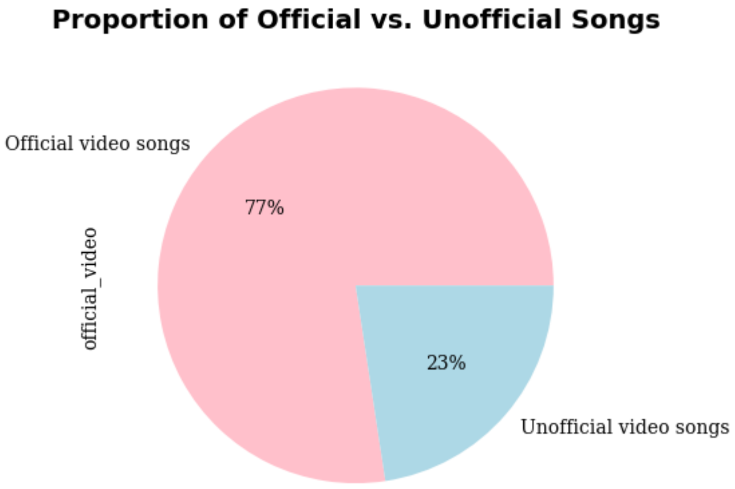


Figure 5. Proportion of Official vs. Unofficial Songs

The pie chart shows that a considerable proportion (77%) of songs on YouTube have official videos, while only 23% are without official videos. This suggests that the music industry places great importance on creating official videos for songs, as they have the potential to greatly influence the popularity and success of a song. They also serve as a means to promote the artist and generate revenue through advertising and partnerships. The high percentage of official videos may be indicative of the industry's focus on branding and marketing, as well as the significance of visual media in music consumption.

6.1.5. Features of Songs

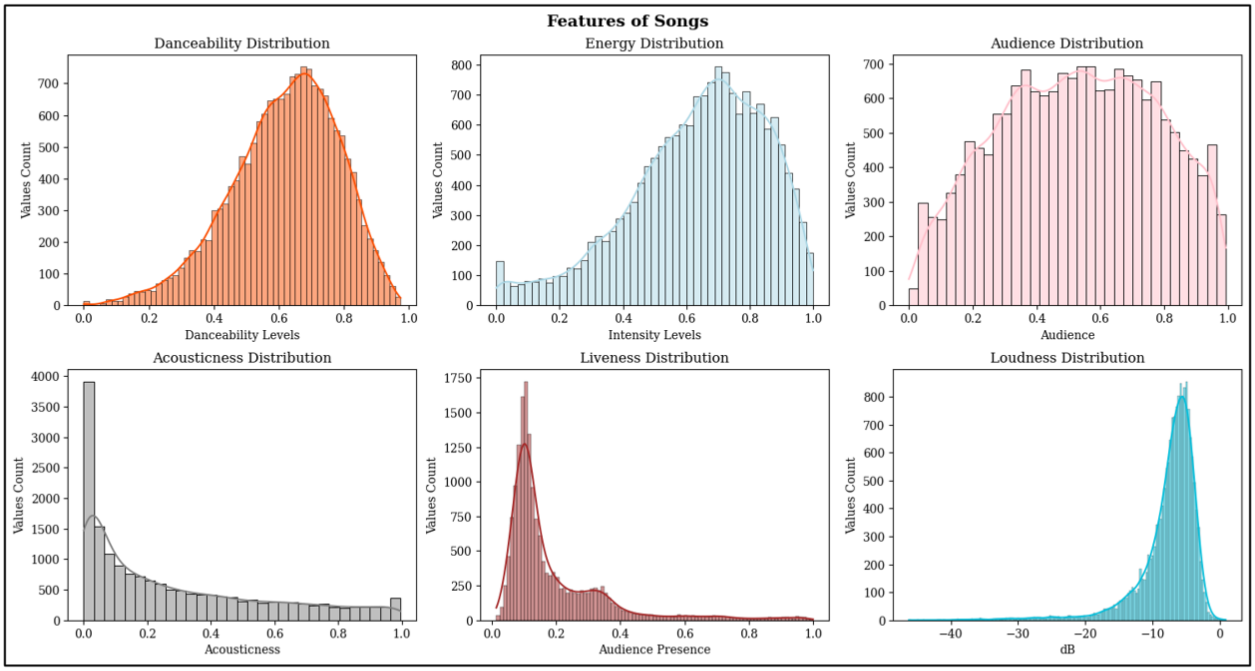


Figure 6. Features of Songs

The average danceability score of the songs in the dataset is 0.619, with a standard deviation of 0.165. This suggests that the songs have a moderate level of danceability. Similarly, the average energy score is 0.635, with a standard deviation of 0.215, indicating that the songs are moderately energetic. The mean valence score is 0.528, suggesting that the songs have a moderate level of positivity.

In terms of acoustic elements, the average acousticness score is 0.286, indicating a moderate presence of acoustic elements in the songs. The average liveness score is 0.191, implying that most songs were recorded in a studio rather than live performances. The mean loudness score is -7.648, suggesting that the songs have a moderate loudness level.

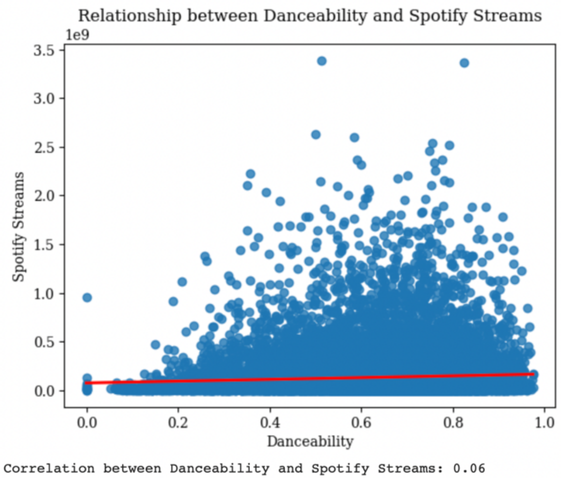
In addition, when it comes to popularity indicators, the average number of views is 92.19 million, and the average number of likes is 641,150. This indicates that the songs in the dataset are generally popular. Additionally, the average number of comments is 27,427, indicating that the songs generate engagement and discussion. The mean likes-to-views ratio is 1.211, meaning that, on average, there is slightly more than one like per view. This suggests that most viewers enjoy the songs.

Skewness and kurtosis are statistical measures used to describe the shape of a distribution. Skewness measures the degree of asymmetry in the distribution, while kurtosis measures the degree of peakness or flatness in the distribution.

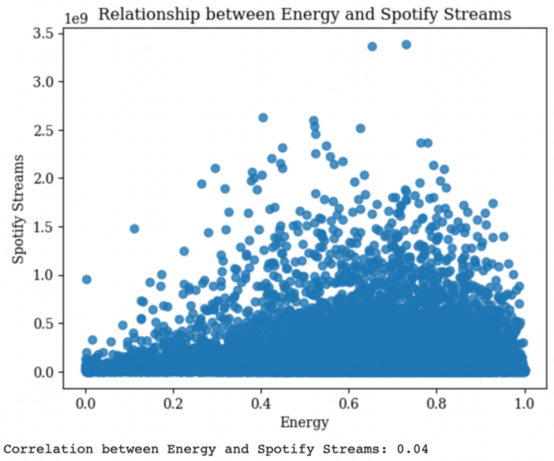
- For the features of Danceability, Energy, Liveness, Valence, and Likes_to_views_ratio, their means are greater than their medians, indicating positive skewness. This means that the distribution is skewed towards higher values, with a tail on the right side.
- On the other hand, Loudness and Acousticness have means that are less than their medians, indicating negative skewness. This suggests that the distribution is skewed towards lower values, with a tail on the left side.
- In terms of kurtosis, Views, Likes, Comments, and Stream exhibit a large difference between their maximum values and the third quartile (75th percentile). This indicates high positive kurtosis, meaning that the distribution has a sharp peak and heavy tails, indicating the presence of a few songs with very high values while most songs have relatively lower values.

Overall, the distribution of the given features is skewed and heavy-tailed, indicating that there are a few songs with very high values, and most songs have relatively low values.

6.1.6. Linear Relationship between song characteristics in Spotify and YouTube



```
Correlation between Danceability and Spotify Streams: 0.06
```

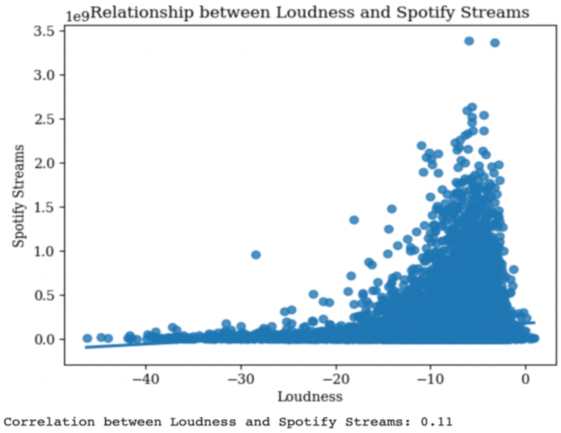


```
Correlation between Energy and Spotify Streams: 0.04
```

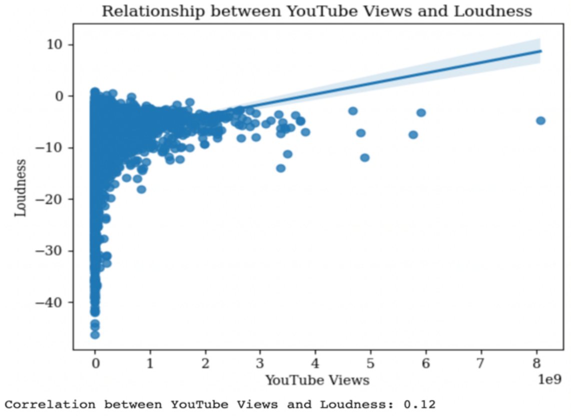
Figure 7. Danceability vs Spotify Stream

Figure 8. Energy vs Spotify Stream

- Danceability and Spotify streams have a 0.06 correlation coefficient, which suggests a marginally favorable association. This shows that there is little to no correlation between a song's degree of danceability and the quantity of Spotify streams it receives.
- A very slight positive link between energy and Spotify streams—with a correlation coefficient of 0.04—can be seen. This shows that there is little to no correlation between a song's energy level and the amount of Spotify streams it gets.



```
Correlation between Loudness and Spotify Streams: 0.11
```



```
Correlation between YouTube Views and Loudness: 0.12
```

Figure 9. Loudness vs Spotify Stream

Figure 10. Loudness vs YouTube

- Spotify streams and loudness have a weakly positive correlation (0.11), which suggests that the two variables are related. This shows that songs with greater volume levels may have a modest propensity to receive more Spotify streams, but that there are probably other factors at work that affect how many streams a song gets.
- The correlation coefficient between YouTube views and loudness, which is 0.12, also shows a weakly positive association between the two variables. According to this, there might be a tiny trend for songs with louder volume levels to gain more YouTube views, but it's probable that there are other factors at work that affect how many views a video gets as well.

6.1.7. Top 10 best performed Artists on YouTube and Spotify

When talking about the entertainment video and music industry, based on the number of views and stream counts, the insight into the current trend of content mainstream and consumption habits can be obtained.

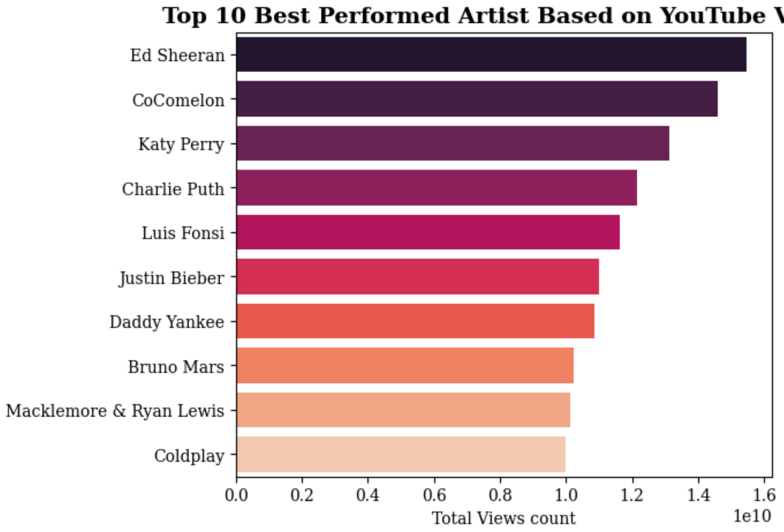


Figure 11. Top 10 best performed Artists on YouTube based on Views

By viewing the top viewed songs, the Best Performed Artist on YouTube is Ed Sheeran with an accumulated 15.5 billion views count for all his songs. Interestingly, coming closely in second place is CoComelon, a channel dedicated to children's content, with 14,6 billion views count.

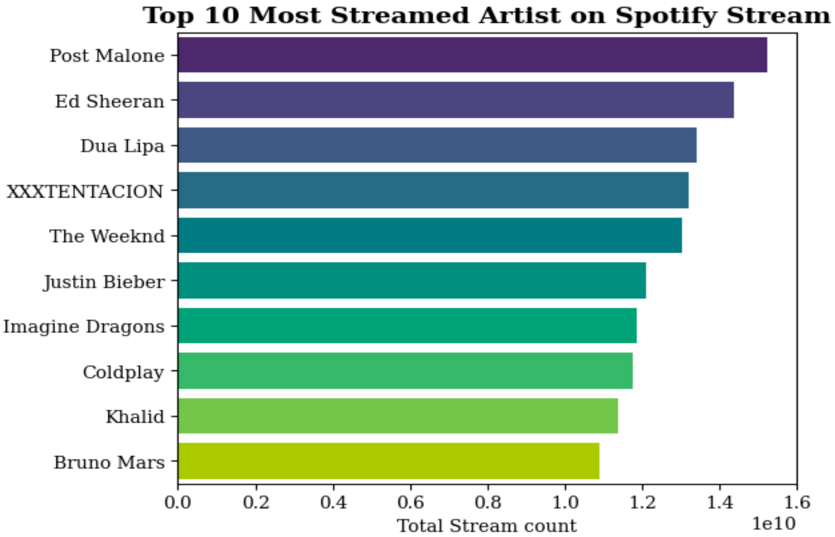
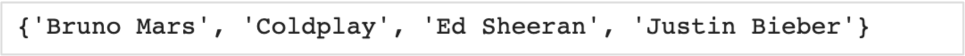


Figure 12. Top 10 best performed Artists on Spotify based on Streams

On the other hand, Post Malone is the most successful artist on Spotify with approximately 15.3 billion streams in total. Come in second, is Ed Sheeran with 14.4 billion streams.

From the figure 11 and 12, it is worth taking note that there is a difference between the Best performed Artist's list of 2 platforms. Spotify seems to favor Artists from The US and UK only. Meanwhile YouTube introduces a few more Artists from different countries and languages, and even a kid channel representative. Moreover, there are only 4 artists who simultaneously make their names on both platforms Top 10.



```
{'Bruno Mars', 'Coldplay', 'Ed Sheeran', 'Justin Bieber'}
```

Figure 13. Artists who made Top 10 Most viewed on both Spotify and Youtube

6.1.8. Top 10 best-performed Songs on YouTube and Spotify

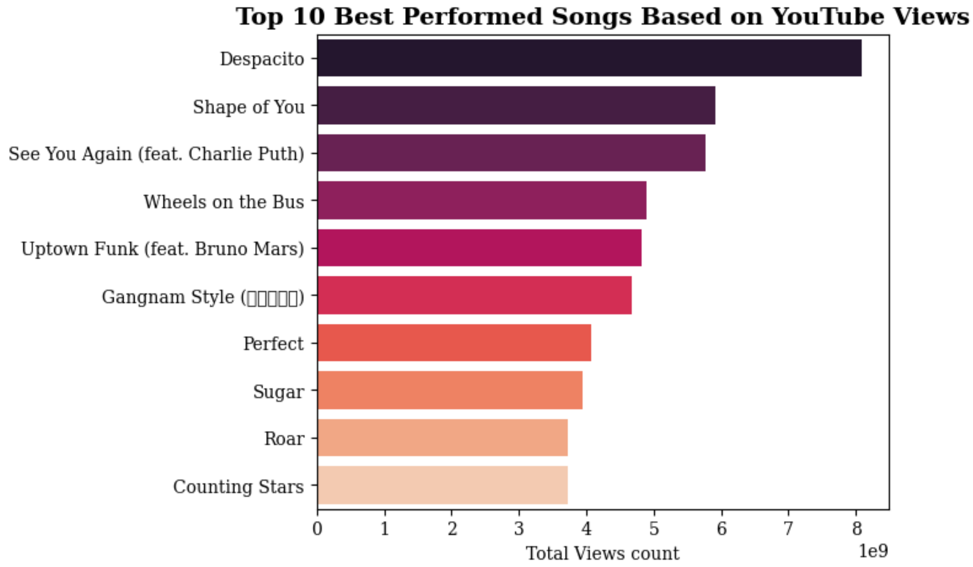


Figure 14. Top 10 best performed Songs on YouTube based on Views

When looking into individual songs, Despacito, a world-famous Spanish song is holding the golden star at more than 8 billion views count, more than 1 billion higher than the second most viewed song, Shape of You. It can also be observed that YouTube Top 10 views consists of a variety of English, Spanish and Korean songs.

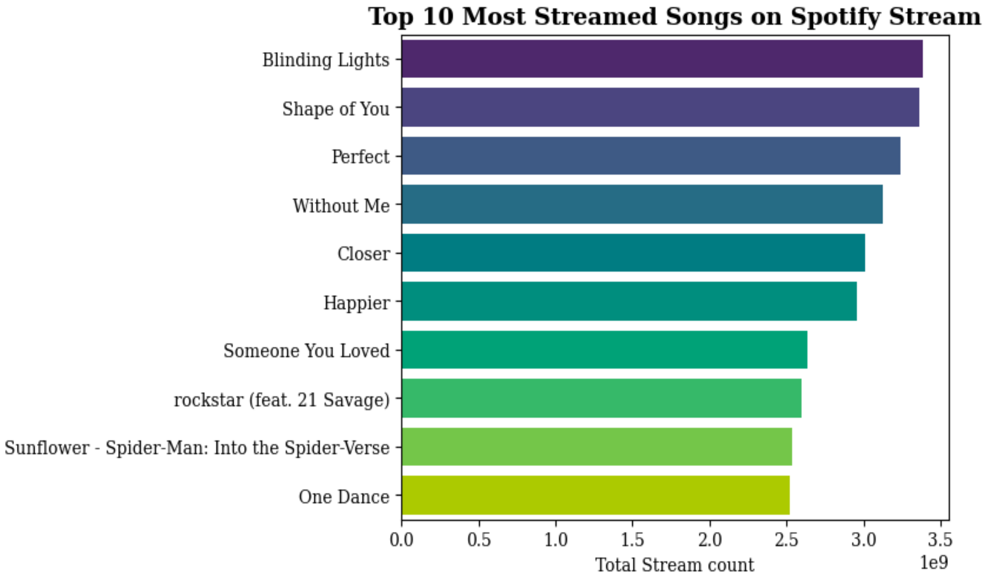
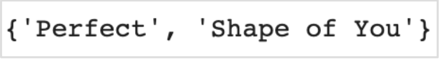


Figure 15. Top 10 best performed Songs on Spotify based on Stream

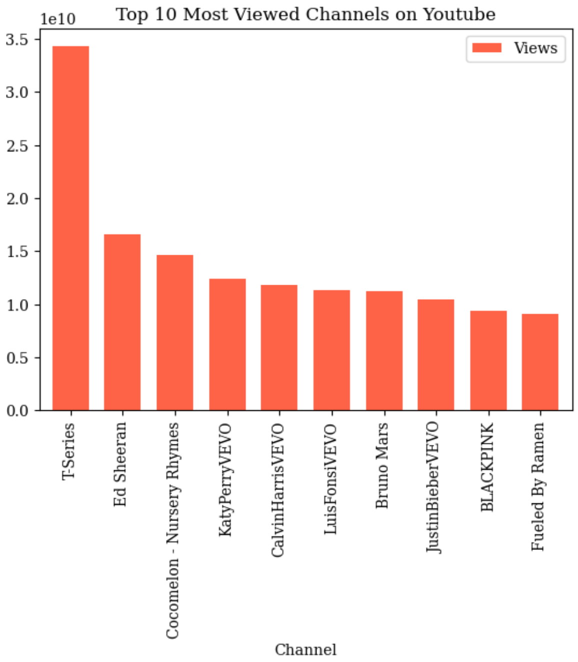
On Spotify, it is a tight race between the top 4 songs as the difference in views is very close. The top 10 viewed songs ranging from 2.52 billion to nearly 3.39 billion. There are only 2 songs that make it to the Top 10 played in 2 platforms, which are:



```
{'Perfect', 'Shape of You'}
```

Figure 16. Songs which made Top 10 most viewed on both Spotify and Youtube

6.1.9. Top Streamed Spotify Albums and Top Viewed YouTube Channels



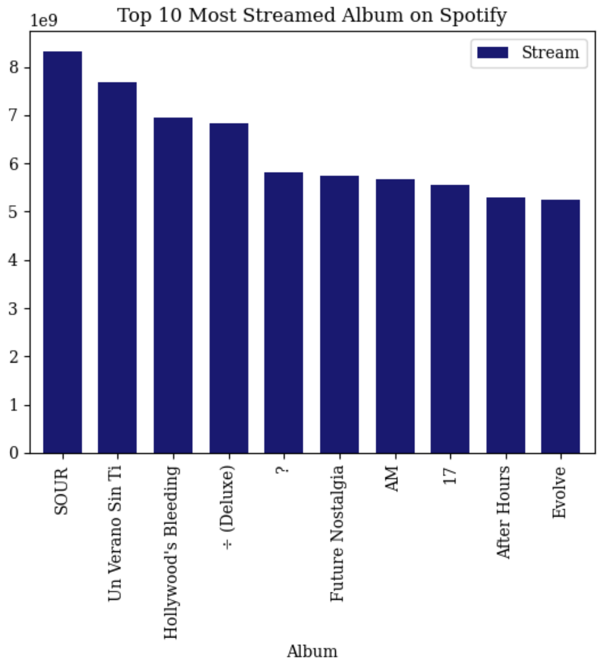
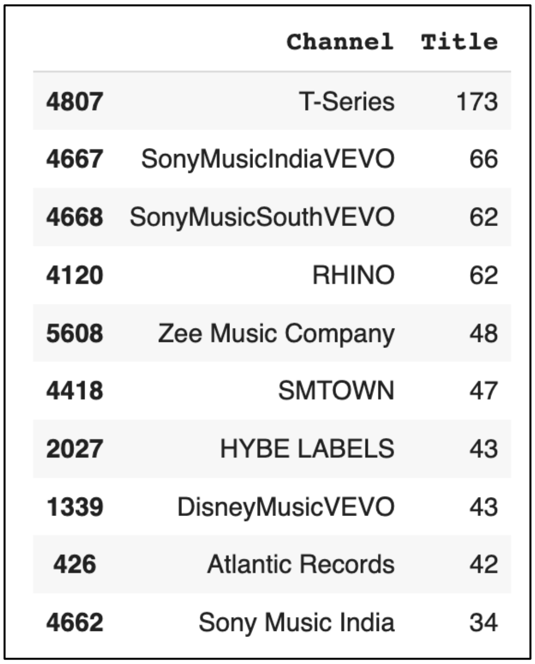


Figure 17. Top 10 viewed channels on YouTube

Figure 18. Top 10 streamed album on Spotify

From above, the top viewed channels and top streamed albums are compared. YouTube seems to gather more consumption with a higher count of visits. Artists also seem to be able to reach a larger audience from YouTube based on the counts of channel visits.

Entertainment companies are also taking advantages of this platform to reach more audience by looking at the YouTube channels with the most uploads are all Entertainment Music companies.



	Channel	Title
4807	T-Series	173
4667	SonyMusicIndiaVEVO	66
4668	SonyMusicSouthVEVO	62
4120	RHINO	62
5608	Zee Music Company	48
4418	SMTOWN	47
2027	HYBE LABELS	43
1339	DisneyMusicVEVO	43
426	Atlantic Records	42
4662	Sony Music India	34

Figure 19. Top 10 YouTube Channels

6.1.10. Top 10 YouTube Videos with Highest Likes to View Ratio

A YouTube indicator called "likes to view ratio" contrasts the quantity of likes a video has earned with the total number of views. This ratio might serve as a helpful gauge of how much an audience is watching and appreciating a video.

With 237,761 likes and 954,081 views, the video "j-hope 'Intro' Visualizer" has a likes to views ratio of 24.92%, as can be seen. This shows how interested and appreciative the viewers are of the video's content. Similar to that, the video "j-hope 'Safety Zone' Visualizer" has acquired 1,952,637 views and 453,910 likes, for a likes-to-views ratio of 23.25%.

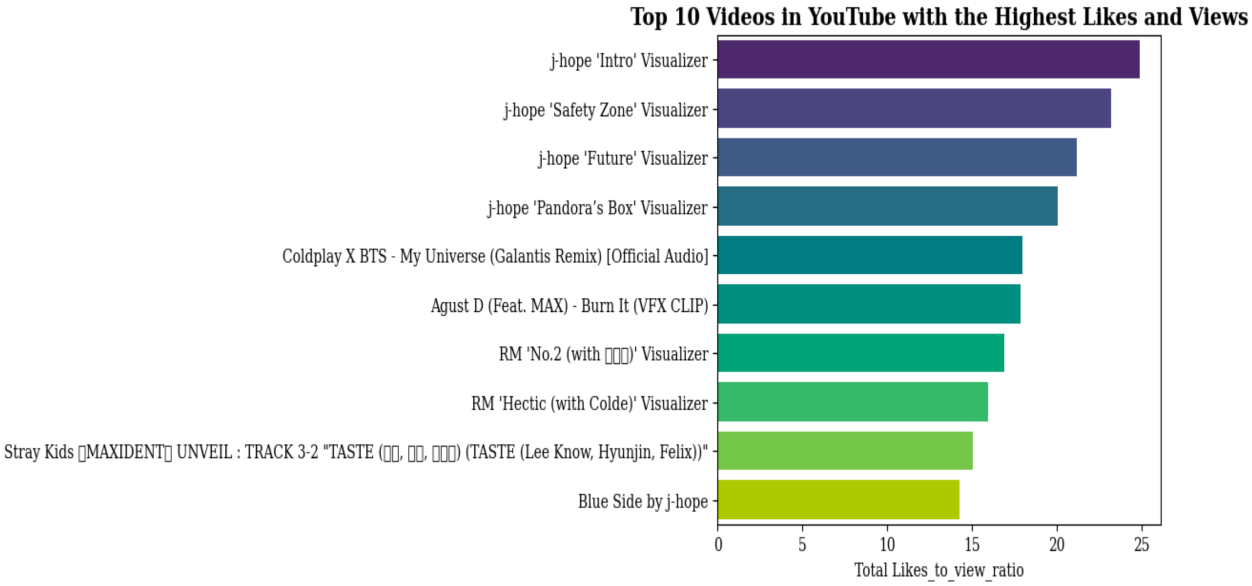


Figure 20. Top 10 YouTube Vides with highest likes and views

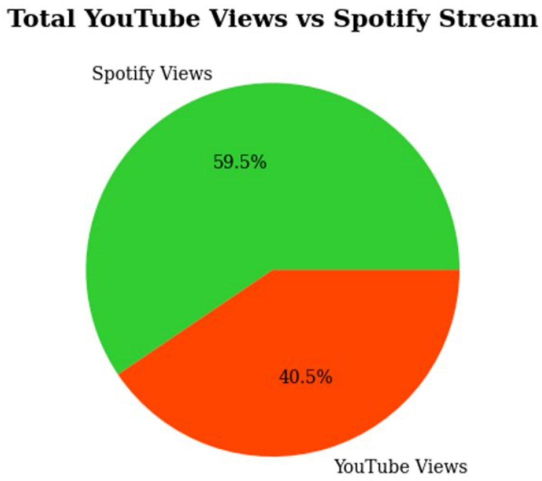
The video "Blue Side by j-hope" has received the most likes (2,357,216), but it has also received the most views (16,504,576), yielding a lower ratio of likes to views (14.28%). This shows that even though the video has a lot of likes, viewer engagement may not be as high as other videos.

7. PREDICTING MODELLING AND ANALYSIS

The following statistical techniques are employed in the study:

- Paired t-test of dependence samples
- Logistic Regression and Confusion Matrix
- Linear Regression and Correlation Matrix

7.1 Paired t-test of dependence samples



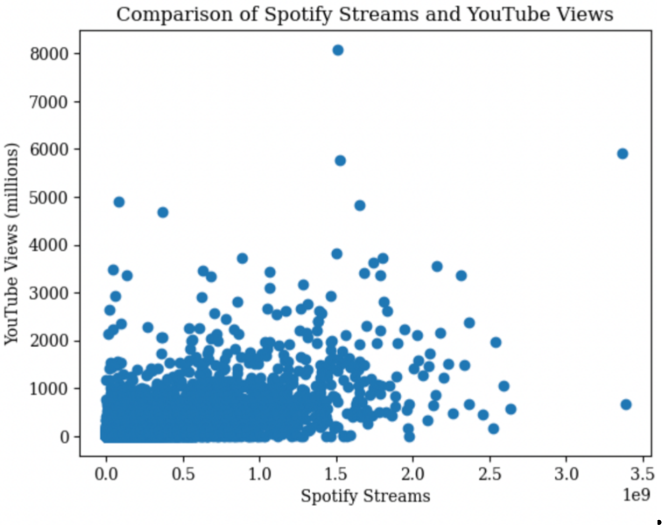


Figure 21. YouTube vs Spotify Views ***Figure 22. Scatterplot on YouTube vs Spotify Views***

Spotify streams contribute to 59.5% of the total views, while YouTube views contribute to 40.5% of the total. Spotify streams and YouTube views have a moderately positive correlation (0.60), which suggests that the two variables are related. This implies a correlation between an increase in YouTube views and a rise in Spotify streaming, and vice versa, although the association is not as strong as it would be if the correlation coefficient were closer to 1.

Beside the linear relationship between Spotify Streams and YouTube Views, another aspect that Artists and Entertainment companies might also be interested in is that would there be a significant difference between Spotify Streams and YouTube Views. For this question, the paired test or dependent t-test was employed to conclude an answer.

Table 4: Application of Paired t-test

Paired t-test of dependence samples	
Null Hypothesis	There is no significant difference between the reception Artists receive on YouTube and Spotify
Alternate Hypothesis	There is a significant difference between the reception Artists receive on YouTube and Spotify
Alpha	0.05
Type of Tail	Two-tail
Critical Value	1.96
t-statistic	-15.363285756565345
P-value	1.7806596212483008e-51
Degree of Freedom	3251
Result	There is a significant difference between the reception Artists receive on YouTube and Spotify

At 95% confidence, the test statistic is -15.363285756565345 which is a very large t-statistic. Based on the t-statistic, the Null hypothesis is rejected and conclude that there is a significant difference between the reception Artists receive on YouTube and Spotify

7.2. Logistic Regression and Confusion Matrix

The logistic regression model was used to examine how licensing affects the number of views on YouTube.

Most of the songs in the dataset are licensed, accounting for 70% of the total. This implies that most of the songs in the dataset are legally authorized and have the required permissions to be distributed and monetized on various platforms. On the other hand, the remaining 30% of the songs are unlicensed, which means they may not have the necessary legal permissions for distribution and monetization. This could be due to various reasons such as copyright issues, lack of authorization, or not meeting the necessary legal requirements.

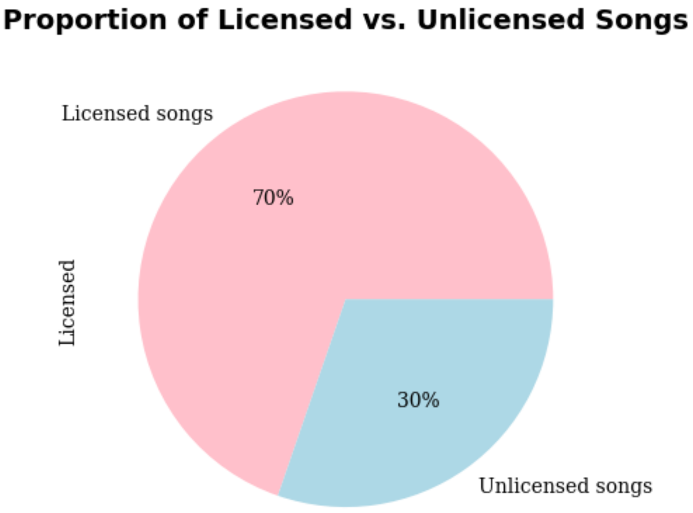


Figure 23. Proportion of licensed vs unlicensed songs

In this analysis, the number of views served as the predictor variable, while the response variable indicated whether a song was licensed or not. The first plot shows the actual data, while the second plot displays the predicted data based on the logistic regression model.

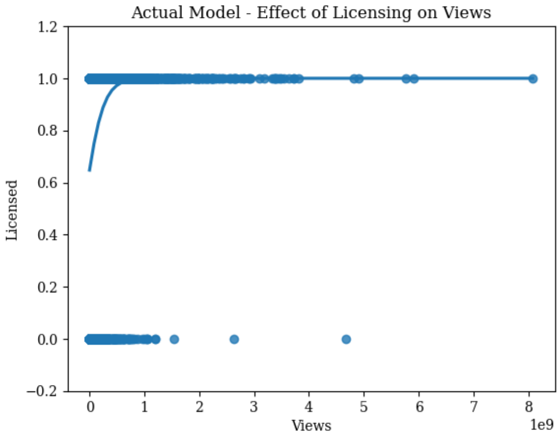


Figure 24. Actual Model on licensing vs views

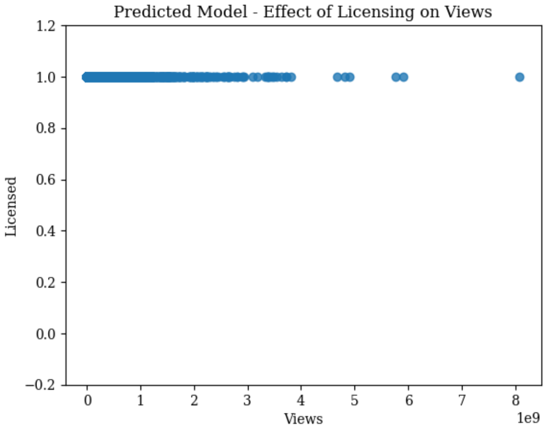


Figure 25. Predicted Model on licensing vs views

Based on the actual model, licensed YouTube videos generally tend to receive more views compared to unlicensed ones. However, the logistic regression model predicts that all videos with views are licensed, which is likely due to an imbalance in the dataset where there are more licensed songs than unlicensed songs. This dataset imbalance makes it challenging for the model to accurately predict the minority class (unlicensed songs) since it has less training data to learn from.

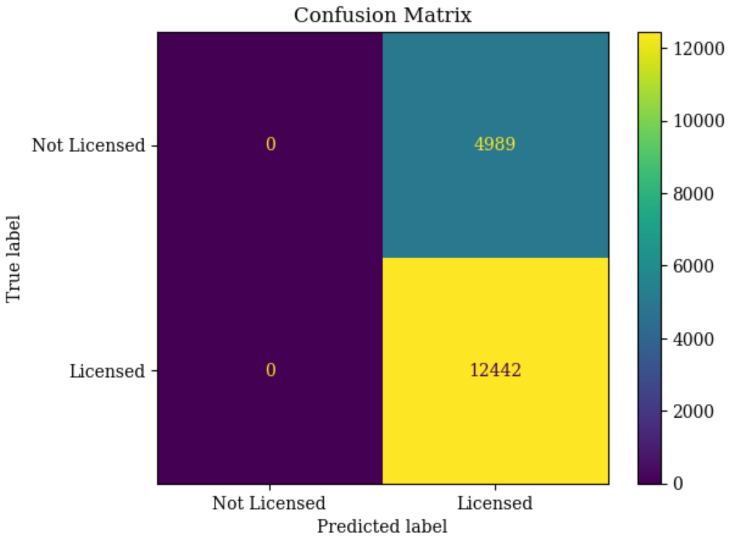
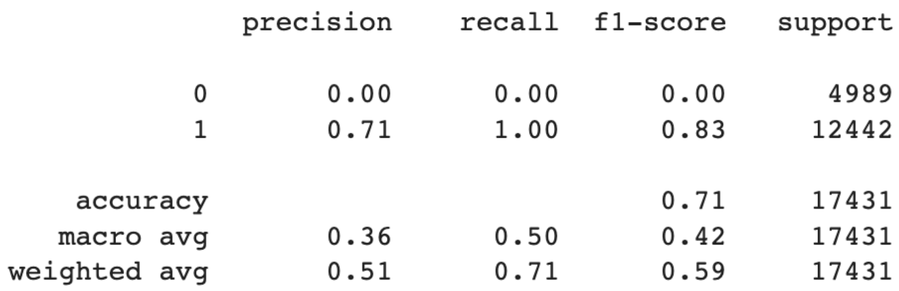


Figure 26. Confusion Matrix

The confusion matrix reveals that the model did not correctly identify any of the licensed songs as positive (true positives) and did not falsely identify any unlicensed songs as positive (false positives). However, it did wrongly identify 4,989 licensed songs as unlicensed (false negatives), while correctly identifying 12,442 unlicensed songs as negative (true negatives).



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      4989
           1       0.71      1.00      0.83     12442

    accuracy                           0.71     17431
   macro avg       0.36      0.50      0.42     17431
weighted avg       0.51      0.71      0.59     17431
```

Figure 27. Classification Report

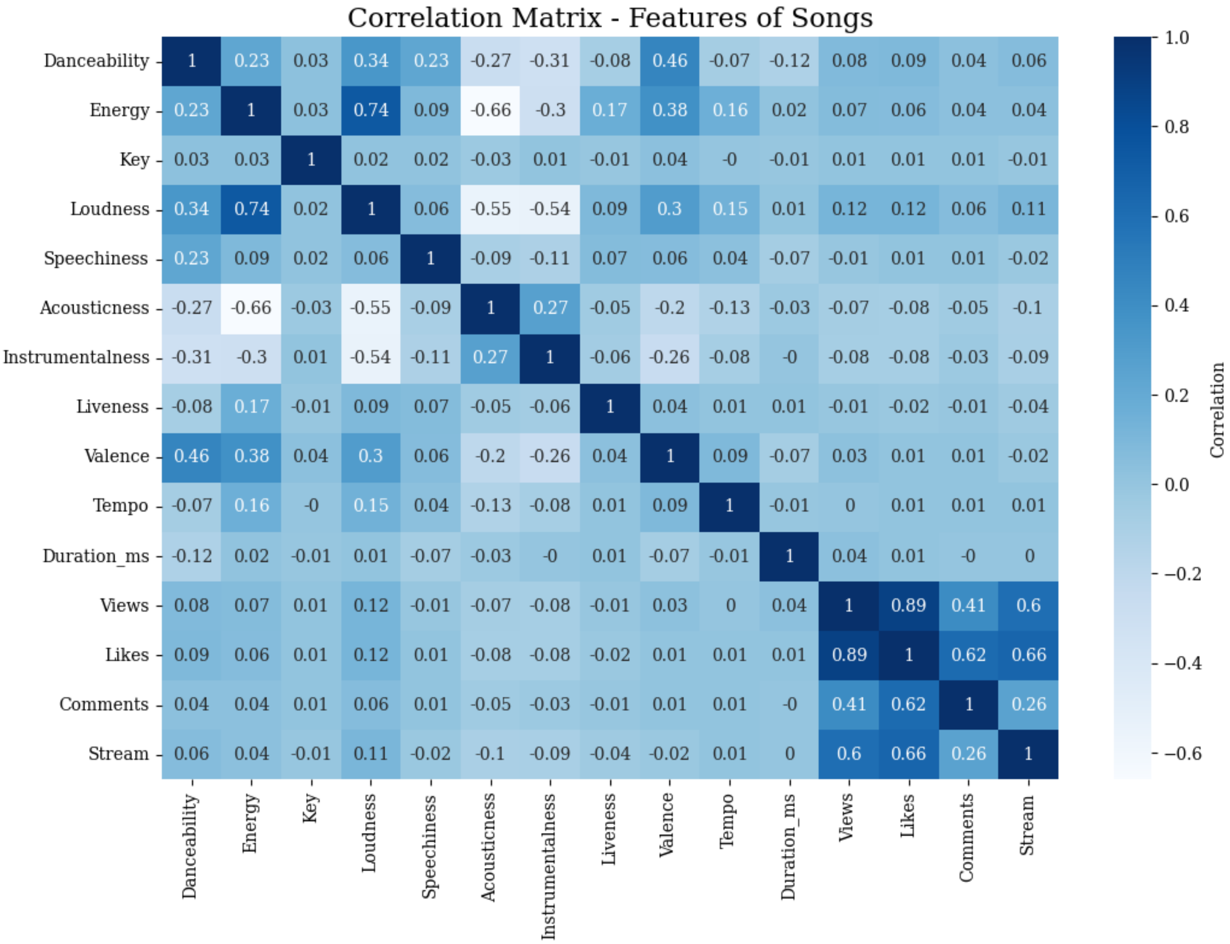
The classification report shows the precision, recall, f1-score, and support metrics for the logistic regression model.

- Precision measures the percentage of correctly predicted positive instances out of all positive predictions made by the model. In this case, the precision for predicting unlicensed songs is 0, indicating that the model did not correctly predict any unlicensed songs.
- Recall measures the percentage of correctly predicted positive instances out of all actual positive instances. The recall for predicting licensed songs is 1, indicating that the model correctly predicted all licensed songs.
- The f1-score is 0.83, which indicates that the model is accurate in predicting licensed songs but less accurate in predicting unlicensed songs.
- The overall accuracy of the model is 0.71, meaning that it correctly predicted the licensing status of 71% of the songs.

The macro avg and weighted avg metrics indicate that the model performs significantly better in predicting licensed songs than unlicensed songs. This discrepancy could be attributed to the imbalanced dataset, where the number of licensed songs is much higher than the number of unlicensed songs. To improve the model's performance in predicting unlicensed songs, it may be necessary to balance the dataset or adjust the model's parameters.

7.3. Linear Regression Analysis and Correlation Matrix

Do the features tempo, volume, energy, danceability and others play any part in making up the popularity of a track? There is a belief that how catchy a song is or how easy it is to dance to, will make a song become well-received by the population. Uncovering this question will certainly enable artists or music producers to find out the formula for creating mainstream content.



	Danceability	Energy	Key	Loudness	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Duration_ms	Views	Likes	Comments	Stream
Danceability	1	0.23	0.03	0.34	0.23	-0.27	-0.31	-0.08	0.46	-0.07	-0.12	0.08	0.09	0.04	0.06
Energy	0.23	1	0.03	0.74	0.09	-0.66	-0.3	0.17	0.38	0.16	0.02	0.07	0.06	0.04	0.04
Key	0.03	0.03	1	0.02	0.02	-0.03	0.01	-0.01	0.04	-0	-0.01	0.01	0.01	0.01	-0.01
Loudness	0.34	0.74	0.02	1	0.06	-0.55	-0.54	0.09	0.3	0.15	0.01	0.12	0.12	0.06	0.11
Speechiness	0.23	0.09	0.02	0.06	1	-0.09	-0.11	0.07	0.06	0.04	-0.07	-0.01	0.01	0.01	-0.02
Acousticness	-0.27	-0.66	-0.03	-0.55	-0.09	1	0.27	-0.05	-0.2	-0.13	-0.03	-0.07	-0.08	-0.05	-0.1
Instrumentalness	-0.31	-0.3	0.01	-0.54	-0.11	0.27	1	-0.06	-0.26	-0.08	-0	-0.08	-0.08	-0.03	-0.09
Liveness	-0.08	0.17	-0.01	0.09	0.07	-0.05	-0.06	1	0.04	0.01	0.01	-0.01	-0.02	-0.01	-0.04
Valence	0.46	0.38	0.04	0.3	0.06	-0.2	-0.26	0.04	1	0.09	-0.07	0.03	0.01	0.01	-0.02
Tempo	-0.07	0.16	-0	0.15	0.04	-0.13	-0.08	0.01	0.09	1	-0.01	0	0.01	0.01	0.01
Duration_ms	-0.12	0.02	-0.01	0.01	-0.07	-0.03	-0	0.01	-0.07	-0.01	1	0.04	0.01	-0	0
Views	0.08	0.07	0.01	0.12	-0.01	-0.07	-0.08	-0.01	0.03	0	0.04	1	0.89	0.41	0.6
Likes	0.09	0.06	0.01	0.12	0.01	-0.08	-0.08	-0.02	0.01	0.01	0.01	0.89	1	0.62	0.66
Comments	0.04	0.04	0.01	0.06	0.01	-0.05	-0.03	-0.01	0.01	0.01	-0	0.41	0.62	1	0.26
Stream	0.06	0.04	-0.01	0.11	-0.02	-0.1	-0.09	-0.04	-0.02	0.01	0	0.6	0.66	0.26	1

Figure 28. Confusion Matrix - Features of Songs

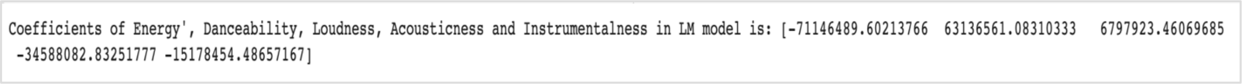
The above correlation matrix displays the pairwise correlations between the features of songs in a dataset, as well as their corresponding reach on both YouTube and Spotify. The matrix values range from -1 to 1, where negative values indicate an inverse correlation, positive values indicate a direct correlation, and the absolute value indicates the strength of the correlation. The heatmap visualization uses a color scale to show the strength of the correlation, with darker colors indicating stronger correlations.

Upon examining the correlation matrix, it is evident that the Views and Likes (0.89) of video songs on YouTube are highly correlated with each other. Additionally, the Loudness and Energy (0.74) of the songs' features are also strongly correlated. This suggests that when one of these features increases, the other feature also tends to increase.

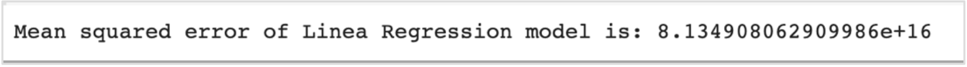
Conversely, features such as Acousticness and Energy of songs (-0.66) and Loudness (-0.54) display a strong negative correlation, which implies that as one feature increases, the other tends to decrease.

Following the correlation matrix above, YouTube Views prediction and Spotify Streams based on songs features are explored. *(Note: the attributes with less than 0.05 correlation with the dependent variables will not be included)*

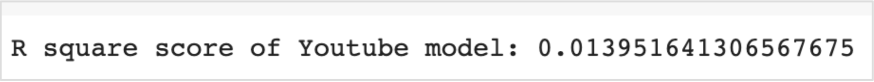
YouTube View Linear Regression model result:



```
Coefficients of Energy', Danceability, Loudness, Acousticness and Instrumentalness in LM model is: [-71146489.60213766  63136561.08310333   6797923.46069685
 -34588082.83251777 -15178454.48657167]
```



```
Mean squared error of Linea Regression model is: 8.134908062909986e+16
```



```
R square score of Youtube model: 0.013951641306567675
```

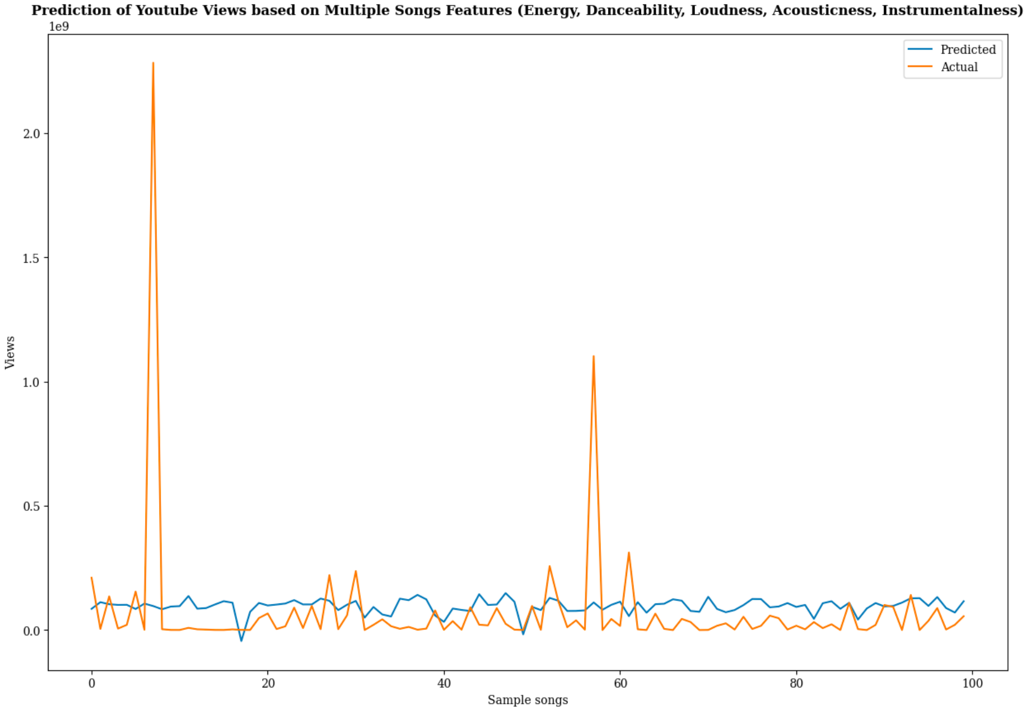
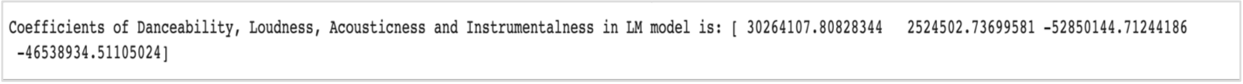
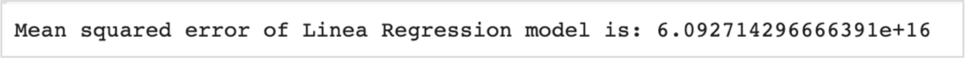


Figure 29. Prediction of YouTube Views based on Energy, Danceability, Loudness, Acousticness, and Instrumentalness of songs

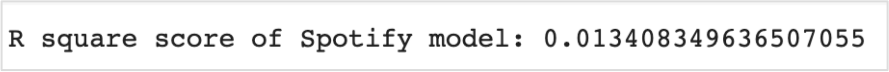
Spotify Stream Linear Regression model result:



```
Coefficients of Danceability, Loudness, Acousticness and Instrumentalness in LM model is: [ 30264107.80828344   2524502.73699581 -52850144.71244186
 -46538934.51105024]
```



```
Mean squared error of Linea Regression model is: 6.092714296666391e+16
```



```
R square score of Spotify model: 0.013408349636507055
```

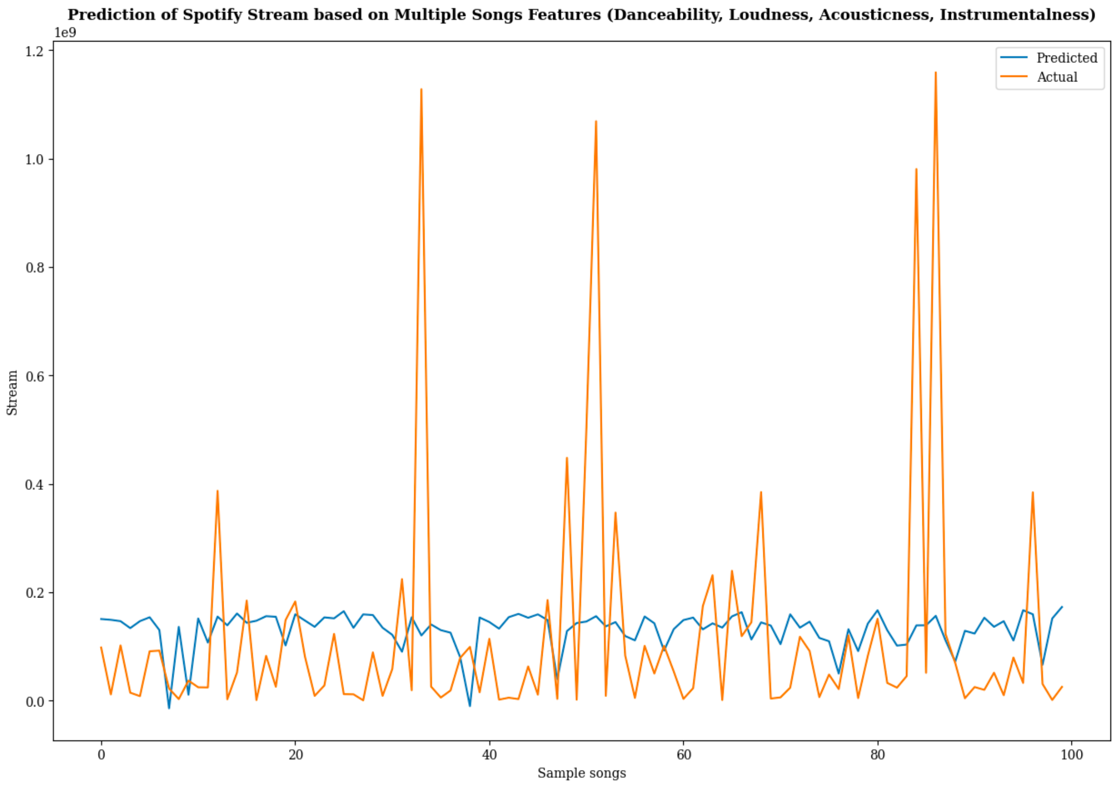


Figure 30. Prediction of Spotify Streams based on Energy, Danceability, Loudness, Acousticness, and Instrumentalness of songs

Reading the statistic of the two models, it can be concluded that:

- The prediction ability of YouTube model is around 1.4% and Spotify model is 1.3%
- The mean squared error values are both very high
- The graphs also demonstrate that even though the model can make some predictions matching the test data, the accuracy is very low.

8. CONCLUSION

The analysis of 18,862 unique songs present in the dataset revealed that songs on Spotify had more views (59.5%) when compared to YouTube. The presence of an official video significantly influenced a song's popularity, with 77% of songs on YouTube having official videos. Energy and loudness showed weak or no correlation with Spotify streams and YouTube views suggesting other factors play a more significant role in a song's popularity on both platforms.

It was evident that Ed Sheeran was the top-performing artist on YouTube with 15.5 billion views, closely followed by CoComelon, a children's content channel, with 14.6 billion views. On Spotify, Post Malone emerged as the most successful artist with 15.3 billion streams, while Ed Sheeran ranked second with 14.4 billion streams. These findings highlight differences in the best-performing artists between the platforms, with Spotify favoring artists from the US and UK, while YouTube includes artists from various countries and languages.

The logistic regression analysis examining the impact of licensing on YouTube views encountered challenges in accurately predicting unlicensed songs, likely due to an imbalanced dataset with more licensed songs. The overall accuracy of the model in predicting licensing status was 71%. By applying paired t-test, it was confirmed that there is a significant difference in audience engagement and consumption patterns between Spotify and YouTube. Strong positive correlations were observed between views and likes on YouTube, indicating high viewer engagement and appreciation for the video content. Linear regression models were built to predict YouTube views and Spotify streams based on song features. However, when predicting song popularity based on song features, such as danceability, energy, and loudness, the models exhibited low prediction abilities and high mean squared error values, suggesting limited accuracy. This implies that factors beyond these characteristics contribute more significantly to a song's popularity on both platforms.

In summary, this analysis provides valuable insights into the differences between Spotify and YouTube in terms of audience engagement and consumption. It highlights the importance of official videos and licensing in driving a song's popularity, as well as the variations in best-performing artists and top-viewed songs between the platforms. However, predicting song popularity based on song features remains challenging.

9. RECOMMENDATIONS

Based on the analysis findings, the following recommendations are suggested for the entertainment industry, artists, and channels:

- **Focus on creating official videos:** Given that the presence of an official video significantly influenced a song's popularity on YouTube, it is essential to prioritize creating high-quality visual content to accompany their music. Investing in professional music videos can help attract more views and engage viewers on the platform.
- **Consider cross-platform promotion:** Recognizing the differences in audience engagement and consumption patterns between Spotify and YouTube will help to maximize their reach and exposure. It is advisable to develop cross-platform promotion strategies that cater to the unique characteristics and preferences of each platform's user base.
- **Explore licensing opportunities:** Licensing plays a crucial role in the success of songs on YouTube. Ensuring the songs are properly licensed will allow for wider distribution and

monetization. This involves obtaining the necessary permissions and rights addressing any potential copyright or authorization issues.

- **Consider diverse factors for popularity prediction:** Apart from features of the songs, it is important to consider additional factors, such as marketing efforts, audience targeting, and promotion strategies, to enhance the potential success of their songs on the respective platforms.
- **Utilize viewer engagement metrics:** Strong positive correlations were observed between views and likes on YouTube, indicating the importance of viewer engagement. Therefore, by active monitoring and leveraging viewer engagement metrics gauge the reception and popularity of their songs. Encouraging viewers to like, comment, and share their content can help boost engagement and visibility.
- **Explore international markets:** The analysis highlighted the differences in best-performing artists between Spotify and YouTube, with Spotify favoring artists from the US and UK while YouTube encompassed artists from various countries and languages. Artists and channels should consider exploring international markets and diversifying their audience base to capitalize on the global reach of platforms like YouTube.
- **Continuously analyze and adapt:** The music industry and digital platforms are dynamic and ever evolving. It is crucial for the entertainment industry to continually analyze trends, audience preferences, and platform algorithms. By staying informed and adapting their strategies accordingly, they can effectively navigate the changing landscape and maximize their chances of success on Spotify and YouTube.
- **Balanced Dataset:** To address the challenges faced in accurately predicting unlicensed songs' impact on YouTube views, by using appropriate sampling method such as Random

oversampling or Random undersampling to analyze imbalanced dataset, it would further help to predict licensed and unlicensed modelling with respect to its views. A dataset with an equal number of licensed and unlicensed songs, as well as songs with and without official videos will help improve the accuracy of the prediction model and provide a more comprehensive understanding of the reach and impact of songs on both YouTube and Spotify.

Overall, understanding the nuances of audience engagement, platform-specific dynamics, and the impact of different factors on song popularity can help artists and channels make informed decisions and optimize their strategies for greater visibility and success on Spotify and YouTube.

10. REFERENCES

Spotify and Youtube. (2023b, March 20). Kaggle. Retrieved on April 24, from https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

GeeksforGeeks. (2022). How to Find Drop duplicate columns in a Pandas DataFrame. *GeeksforGeeks*. Retrieved on April 24, from https://www.geeksforgeeks.org/how-to-find-drop-duplicate-columns-in-a-pandas-dataframe/

Zach. (2021). How to Drop Duplicate Columns in Pandas (With Examples). *Statology*. Retrieved on April 24, 2023 from https://www.statology.org/pandas-drop-duplicate-columns/

Merge, join, concatenate and compare — pandas 2.0.1 documentation. (n.d.). Retrieved on April 24, 2023 from https://pandas.pydata.org/docs/user_guide/merging.html

Kleppen, E. (2023, April 5). How To Find Outliers in Data Using Python (and How To Handle Them). *CareerFoundry*. Retrieved on April 24, 2023 from https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/

Stack Overflow. *Compare two lists in python and return matches*. (n.d.). Retrieved on April 24, 2023 from https://stackoverflow.com/questions/1388818/how-can-i-compare-two-lists-in-python-and-return-matches

Northeastern University, ALY6015 - Intermediate Analytics (2023). *Module 2 – Chi Square and ANOVA*. Retrieved on April 24, 2023.

Python, R. (2022). Logistic Regression in Python. *realpython.com*. https://realpython.com/logistic-regression-python/#:~:text=The%20logistic%20regression%20function%20%F0%9D%91%9D(%F0%9D%90%B1)%20is%20the%20sigmoid%20function,that%20the%20output%20is%200.

GeeksforGeeks. (2022b). Create a correlation Matrix using Python. *GeeksforGeeks*. https://www.geeksforgeeks.org/create-a-correlation-matrix-using-python/

GeeksforGeeks. (2023a). Linear Regression Python Implementation. *GeeksforGeeks*. https://www.geeksforgeeks.org/linear-regression-python-implementation/

Narkhede, S. (2021, June 15). Understanding Confusion Matrix - Towards Data Science. *Medium*. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

PyCoach. (2022b, January 4). A Simple Guide to Linear Regression using Python - Towards Data Science. *Medium*. https://towardsdatascience.com/a-simple-guide-to-linear-regression-using-python-7050e8c751c1 *sklearn.linear_model.LogisticRegression*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

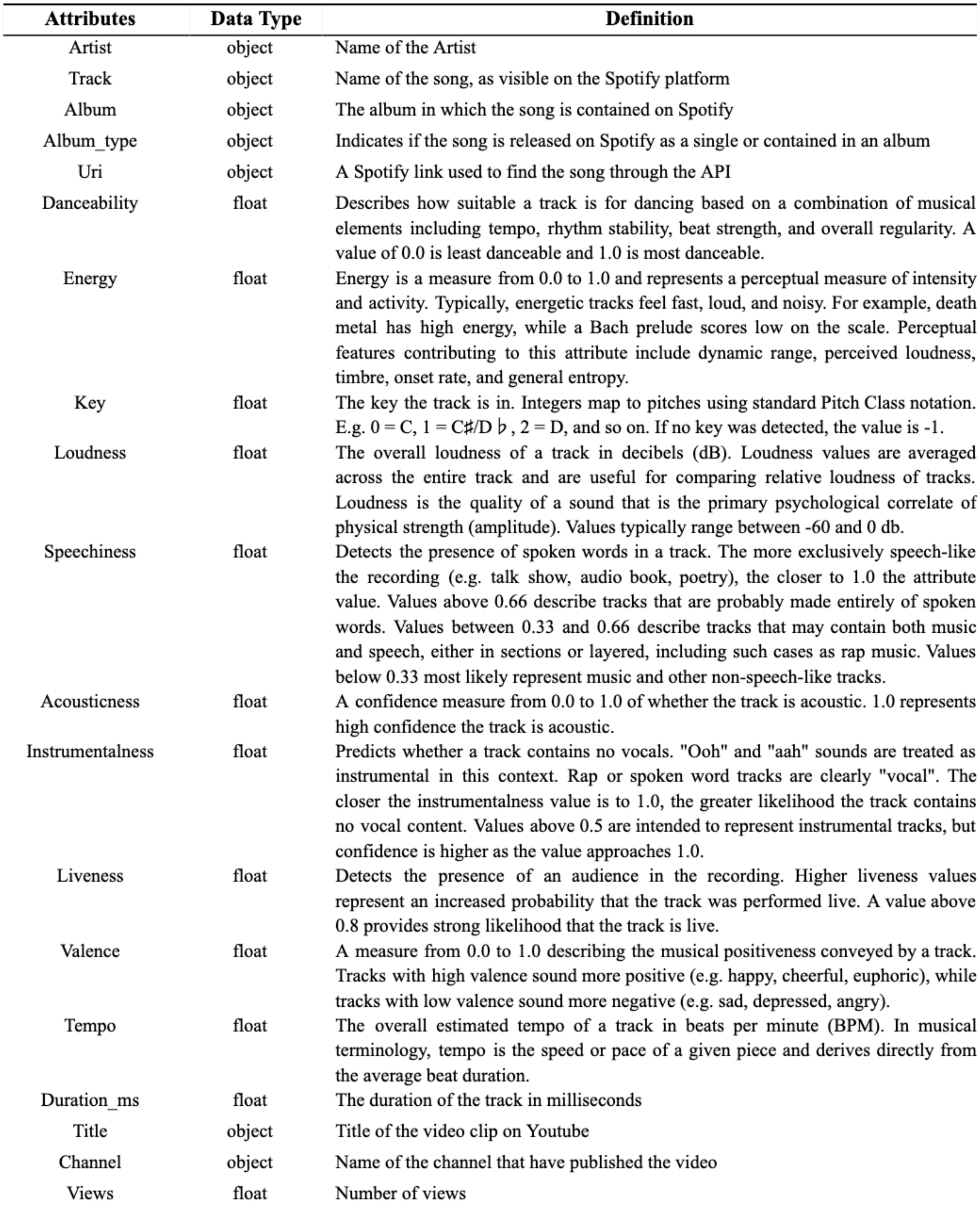
GeeksforGeeks. (2021a). Bar Plot in Matplotlib. *GeeksforGeeks*. https://www.geeksforgeeks.org/bar-plot-in-matplotlib/

Matplotlib Pie Charts. (n.d.). https://www.w3schools.com/python/matplotlib_pie_charts.asp

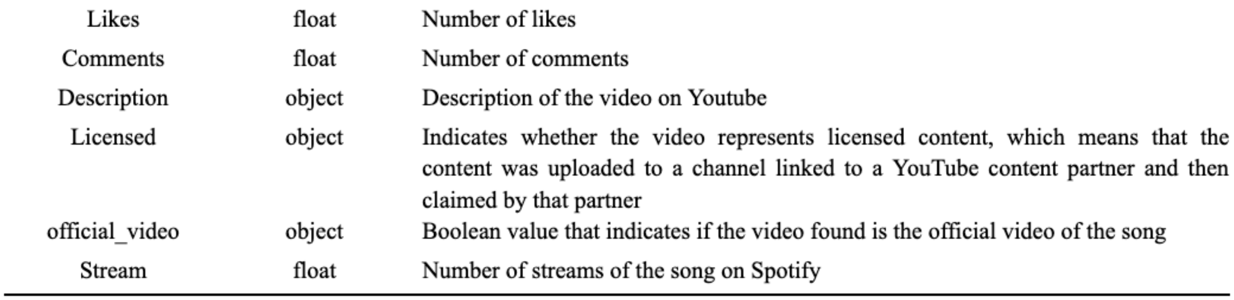
Scatter plot — Matplotlib 3.7.1 documentation. (n.d.). https://matplotlib.org/stable/gallery/shapes_and_collections/scatter.html

11. APPENDIX

Table 1: Data type and definition of attributes



Attributes	Data Type	Definition
Artist	object	Name of the Artist
Track	object	Name of the song, as visible on the Spotify platform
Album	object	The album in which the song is contained on Spotify
Album_type	object	Indicates if the song is released on Spotify as a single or contained in an album
Uri	object	A Spotify link used to find the song through the API
Danceability	float	Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	float	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Key	float	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
Loudness	float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
Speechiness	float	Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Instrumentalness	float	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Liveness	float	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
Valence	float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
Tempo	float	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
Duration_ms	float	The duration of the track in milliseconds
Title	object	Title of the video clip on Youtube
Channel	object	Name of the channel that have published the video
Views	float	Number of views



Likes	float	Number of likes
Comments	float	Number of comments
Description	object	Description of the video on Youtube
Licensed	object	Indicates whether the video represents licensed content, which means that the content was uploaded to a channel linked to a YouTube content partner and then claimed by that partner
official_video	object	Boolean value that indicates if the video found is the official video of the song
Stream	float	Number of streams of the song on Spotify