Capstone Project: Optimizing Music for Video Success on YouTube

Business Analytics Capstone - BA723

Saksham Tiwari-301334725

Centennial College

David Parent

August 15, 2024

Table of Contents

[Executive Summary 2](#_Toc174653607)

[Executive Introduction 2](#_Toc174653608)

[Executive Objective 2](#_Toc174653609)

[Executive Model Description 2](#_Toc174653610)

[Executive Recommendations 3](#_Toc174653611)

[Introduction 5](#_Toc174653612)

[Background 5](#_Toc174653613)

[Problem Statement 5](#_Toc174653614)

[Objectives & Measurement 5](#_Toc174653615)

[Assumptions and Limitations 5](#_Toc174653616)

[Data Sources 6](#_Toc174653617)

[Data Set Introduction 6](#_Toc174653618)

[Exclusions 6](#_Toc174653619)

[Initial Data Cleansing and Preparation 7](#_Toc174653620)

[Data Dictionary 8](#_Toc174653621)

[Data Exploration 10](#_Toc174653622)

[Data Exploration Techniques 10](#_Toc174653623)

[Data Cleansing 32](#_Toc174653624)

[Summary 33](#_Toc174653625)

[Data Preparation and Feature Engineering 33](#_Toc174653626)

[Data Preparation Needs 33](#_Toc174653627)

[Handling Missing Values 33](#_Toc174653628)

[Handling Skewness 34](#_Toc174653629)

[Handling Outliers 35](#_Toc174653630)

[Model Exploration 36](#_Toc174653631)

[Modeling Approach/Introduction 36](#_Toc174653632)

[Full Linear Regression 36](#_Toc174653633)

[Forward Linear Regression 37](#_Toc174653634)

[Backward Linear Regression 38](#_Toc174653635)

[Stepwise Linear Regression 39](#_Toc174653636)

[Regressor Tree 40](#_Toc174653637)

[Random Forest 41](#_Toc174653638)

[Neural Network 42](#_Toc174653639)

[Model Recommendation 44](#_Toc174653640)

[Model Selection 44](#_Toc174653641)

[Model Theory 46](#_Toc174653642)

[Model Assumptions and Limitations 48](#_Toc174653643)

[Model Sensitivity to Key Drivers 50](#_Toc174653644)

[Conclusion and Recommendations 52](#_Toc174653645)

[Impacts on Business Problem (Scope of the recommended model) 52](#_Toc174653646)

[Recommended Next Steps 52](#_Toc174653647)

# Executive Summary

## Executive Introduction

This project focuses on developing a predictive model aimed at identifying the critical factors that drive YouTube viewership for music tracks. The primary goal is to optimize music quality and reach for artists and sound engineers while enabling managers and record labels to strategically identify and prioritize tracks most likely to achieve high viewership on YouTube. This initiative aligns with the broader objective of enhancing content strategy, guiding strategic decisions for music video production, and maximizing visibility and engagement on the platform.

## Executive Objective

The objective of this project is to create a robust and interpretable model capable of predicting YouTube views based on various song features. The model is designed to balance predictive accuracy with simplicity, ensuring that the insights derived can be easily understood and applied by non-technical stakeholders. By analyzing song features and performance metrics, the goal is to provide actionable insights that enhance the success of tracks and guide strategic decisions for music video production. Ultimately, this project aims to help artists, producers, and record labels optimize their content for better reach and performance on YouTube.

## Executive Model Description

Several modeling techniques, ranging from linear to non-linear approaches, were employed to determine the most effective predictor of YouTube views. The Backward Regression model emerged as the most suitable solution due to its strong predictive performance, interpretability, and alignment with business objectives. This model streamlines the feature set by retaining only the statistically significant predictors, ensuring that the results are both reliable and actionable.

Key predictors identified by the model include:

* **Loudness**
* **Duration\_minutes**
* **official\_video\_True**
* **Danceability**
* **Licensed\_True**
* **Energy**
* **Tempo**
* **flag\_outlier\_Loudness**
* **Album\_type\_compilation**
* **Album\_type\_single**
* **Key\_8.0**

These predictors play a critical role in determining the success of a track on YouTube, contributing significantly to the model's predictions.

## Executive Recommendations

Based on the insights derived from the Backward Regression model, the following recommendations are made to optimize YouTube viewership and align with strategic business objectives:

1. **Optimize Loudness:**
   * **Action:** Increase the loudness of tracks within acceptable industry standards.
   * **Rationale:** Loudness was identified as a significant predictor of YouTube views. Enhancing loudness while maintaining audio quality can lead to substantial gains in viewership.
2. **Focus on Track Duration:**
   * **Action:** Adjust track lengths to align with audience preferences and avoid excessively long or short durations.
   * **Rationale:** The duration of tracks is positively associated with higher YouTube views. Optimizing duration ensures maximum engagement and reduces the risk of losing listener interest.
3. **Leverage Danceability and Energy:**
   * **Action:** Fine-tune the danceability and energy levels of tracks to match the target audience's preferences.
   * **Rationale:** Both danceability and energy are critical factors in determining a track’s popularity. Aligning these features with audience expectations can enhance engagement and increase viewership.
4. **Invest in Official Videos:**
   * **Action:** Prioritize the creation and promotion of high-quality official videos for tracks.
   * **Rationale:** The presence of an official video is strongly linked to higher viewership. Investing in video production can significantly boost a track’s visibility on YouTube.
5. **Ensure Proper Licensing:**
   * **Action:** Secure proper licensing for all tracks to enhance promotional potential and compliance with platform requirements.
   * **Rationale:** Licensed tracks are more likely to achieve higher view counts, as they benefit from better promotion and distribution opportunities.
6. **Consider Tempo and Key Features:**
   * **Action:** Adjust the tempo of tracks and carefully select the musical key to appeal to the target audience.
   * **Rationale:** Tempo and key have a measurable impact on viewership. Optimizing these features can help tailor tracks to listener preferences, increasing the likelihood of success.
7. **Address Outlier Loudness:**
   * **Action:** Manage outliers in loudness to avoid extreme deviations that could negatively affect the listening experience.
   * **Rationale:** While loudness is a key driver, extreme outliers should be controlled to maintain audio quality and listener satisfaction.
8. **Strategic Use of Album Types:**
   * **Action:** Consider the advantages of releasing compilation albums and develop strategies to enhance the performance of singles.
   * **Rationale:** Compilation albums were shown to have a positive impact on viewership, while singles may require additional promotional efforts to achieve similar success.

By following these recommendations, stakeholders can strategically optimize their content for YouTube, driving higher engagement and ensuring alignment with overall business objectives. The actionable insights provided by the model offer a clear path to enhancing track performance and maximizing visibility on the platform.

# Introduction

## Background

The music industry has always been a blend of creativity and strategy. With the advent of digital platforms like YouTube, the power dynamics have shifted, allowing data to play an increasingly pivotal role in determining a track's success. Today, artists, sound engineers, and record labels have access to an unprecedented amount of data, which, if harnessed correctly, can significantly influence a song's reach and impact. This project was conceived to explore how data-driven insights can optimize the production and promotion of music content, particularly on YouTube, where billions of views are up for grabs. By understanding the key attributes that drive viewership, stakeholders can make informed decisions that enhance both the artistic and commercial success of their tracks.

## Problem Statement

In an industry where a single hit can propel an artist to global stardom, identifying the elements that contribute to a song’s popularity is crucial. The primary challenge addressed in this project is to uncover and quantify the specific song attributes—such as danceability, energy, loudness, and others—that are most predictive of high view counts on YouTube.

How can artists and sound engineers optimize music quality and reach, while managers and record labels identify tracks that are likely to attract the highest views on YouTube? By analyzing song features, we aim to predict and enhance the success of tracks, guiding strategic decisions for music video production. Given the diverse nature of music consumption and the unpredictable nature of virality, this project seeks to bring clarity to what can be controlled: the song's inherent features and the strategic decisions made during its production and release.

## Objectives & Measurement

The objective of this project is to leverage data analytics to build robust predictive models that forecast YouTube views based on various song characteristics. By achieving this, the project aims to empower music stakeholders—artists, producers, and record labels—to optimize their content for maximum reach and engagement. The success of this endeavor will be measured by the accuracy of the models in predicting view counts, the interpretability of the results, and the practical recommendations that emerge from the analysis. Ultimately, the goal is to create a roadmap for producing tracks that not only resonate with audiences but also achieve significant commercial success.

## Assumptions and Limitations

This analysis operates under the assumption that the available dataset is representative of broader trends in music consumption on YouTube. It also assumes that the relationships identified within the data are consistent over time. The dataset initially included Spotify streams as a feature. However, we decided to exclude this variable to focus solely on YouTube views. This decision was made to maintain a clear and consistent scope for the project, given that the primary objective was to optimize YouTube viewership for artists and record labels. By narrowing the focus to YouTube, we aimed to provide more targeted insights that could directly impact the platform most relevant to our stakeholders.

The dataset may not equally represent all music genres, which could result in biased findings. Certain genres with a stronger presence on YouTube might dominate the analysis, potentially overlooking trends in less represented genres.

The dataset may not equally represent all music genres, which could result in biased findings. Certain genres with a stronger presence on YouTube might dominate the analysis, potentially overlooking trends in less represented genres.

# Data Sources

## Data Set Introduction

The dataset utilized for this project was sourced from a combination of public and proprietary databases, focusing on YouTube as the primary platform for analysis. The data collection process involved scraping and aggregating information from YouTube's API, capturing a wide range of attributes related to music tracks and their performance metrics on the platform.

It has 20718 entries with 28 columns.

**Key Data Sources:**

YouTube API: Provided comprehensive data on video views, likes, comments, and other engagement metrics, as well as metadata about the tracks, such as title, artist, and album.

Proprietary Music Databases: Supplemented the YouTube data with additional information about the tracks, including detailed song features like danceability, energy, and loudness, which are crucial for the analysis.

Internal Label Data: Some data points were sourced from internal databases of music labels, particularly around licensing and official video status, which provided deeper insights into the content's potential reach and engagement.

This dataset was acquired from Kaggle and it was last updated on 7th of February, 2023.   
link: <https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube/data>

## Exclusions

During the data preprocessing phase, several columns were excluded to ensure that the dataset was focused on the most relevant and actionable features for the analysis. The following exclusions were made:

* 'Unnamed: 0': This column, typically an index or placeholder from the original dataset, was removed as it did not provide any meaningful information for the analysis.
* 'Uri': This identifier for songs was excluded as it was redundant for the purposes of the analysis, with no direct influence on YouTube views being a Spotify link used to find the song through the API.
* 'Url\_youtube' and 'Url\_spotify': Both columns were removed because they represent links to external platforms rather than song features. The analysis aimed to focus on song characteristics and performance metrics rather than external identifiers.
* 'Description': The description field, often containing textual data, was excluded to keep the analysis concentrated on quantitative features, which are more directly related to predicting views.
* 'Stream': This column, representing Spotify streams, was excluded to narrow the focus of the project solely to YouTube views, aligning with the scope and objectives of the analysis. The exclusion allowed for a more focused analysis of the platform-specific factors driving song success.

## Initial Data Cleansing and Preparation

In the initial data preparation phase, one key step was the conversion of the Duration\_ms column to Duration\_minutes.

Reason for Conversion:

* Interpretability: The original Duration\_ms column measured the length of a song in milliseconds, which is not an intuitive or easily interpretable unit for most stakeholders. Converting this to minutes aligns with standard industry practices and provides a more meaningful metric.
* Consistency: This conversion ensures that the duration of songs is presented in a familiar format, making it easier to compare with industry standards and interpret within the context of the analysis.

The result was stored in a new column called Duration\_minutes, which was then used throughout the analysis.

Another key step in the initial data preparation was the exclusion of songs with fewer than 1 million views. This decision was made to set a meaningful benchmark for the analysis.

Reason for Exclusion:

* Benchmark Setting: The decision to exclude songs with fewer than 1 million views was based on the goal of setting a high-performance benchmark for artists and record labels. Songs that reach or exceed this view count are likely to be more commercially successful and generate significant revenue.
* Focus on High-Impact Analysis: By focusing on songs with at least 1 million views, the analysis targets tracks that have already demonstrated a degree of popularity and success. This provides more actionable insights for stakeholders who are looking to maximize the impact and reach of their music.
* Scope of Success: Songs that surpass 1 million views are more likely to continue gaining traction and potentially become viral hits. Analyzing this subset of data ensures that the recommendations and findings are aligned with the goal of helping artists and labels achieve and surpass this threshold.

This filtering process ensured that the subsequent analysis was concentrated on tracks that are already on a successful trajectory, making the insights more relevant for stakeholders aiming to produce hits and maximize returns.

## Data Dictionary

|  |  |  |
| --- | --- | --- |
| Feature | Description | Type of Variable |
| Track | Name of the song, as visible on the Spotify platform. | Categorical |
| Artist | Name of the artist. | Categorical |
| Url\_spotify | The URL of the artist's page on Spotify. | Categorical |
| Album | The album in which the song is contained on Spotify. | Categorical |
| Album\_type | Indicates if the song is released on Spotify as a single or contained in an album. | Categorical |
| Uri | A Spotify link used to find the song through the API. | Categorical |
| Danceability | Describes how suitable a track is for dancing based on various musical elements. Value ranges from 0.0 to 1.0. | Numerical (Float) |
| Energy | A perceptual measure of intensity and activity, ranging from 0.0 to 1.0. | Numerical (Float) |
| Key | The key the track is in, mapped to pitches using standard Pitch Class notation. | Categorical |
| Loudness | The overall loudness of a track in decibels (dB), typically ranging from -60 to 0 dB. | Numerical (Float) |
| Speechiness | Detects the presence of spoken words in a track. Value ranges from 0.0 to 1.0. | Numerical (Float) |
| Acousticness | A confidence measure from 0.0 to 1.0 indicating whether the track is acoustic. | Numerical (Float) |
| Instrumentalness | Predicts whether a track contains no vocals, with values closer to 1.0 indicating a higher likelihood of being instrumental. | Numerical (Float) |
| Liveness | Detects the presence of an audience in the recording, with values from 0.0 to 1.0. | Numerical (Float) |
| Valence | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. | Numerical (Float) |
| Tempo | The overall estimated tempo of a track in beats per minute (BPM). | Numerical (Float) |
| Duration\_ms | The duration of the track in milliseconds. | Numerical (Integer) |
| Stream | Number of streams of the song on Spotify. | Numerical (Integer) |
| Url\_youtube | URL of the video linked to the song on YouTube, if it exists. | Categorical |
| Title | Title of the video on YouTube. | Categorical |
| Channel | Name of the channel that published the video on YouTube. | Categorical |
| Views | Number of views on YouTube. | Numerical (Integer) |
| Likes | Number of likes on YouTube. | Numerical (Integer) |
| Comments | Number of comments on YouTube. | Numerical (Integer) |
| Description | Description of the video on YouTube. | Categorical |
| Licensed | Indicates whether the video represents licensed content. | Categorical (Boolean) |
| official\_video | Indicates whether the video found is the official video of the song (True/False). | Categorical (Boolean) |

# Data Exploration

## Data Exploration Techniques

The data exploration process provided a deep dive into both numerical and categorical variables. The goal was to uncover patterns, distributions, and relationships within the dataset that would inform later stages of modeling and analysis. This exploration was segmented into several key areas:

* Initial Exploratory Data Analysis - Numerical Variables: Focused on understanding the distribution, central tendency, variability, and correlations between numerical features. Histograms, box plots, and correlation matrices were used to extract insights.
* Initial Exploratory Data Analysis - Categorical Variables: Concentrated on understanding the distribution and impact of categorical variables on key performance indicators such as Views. This analysis involved bar charts, violin plots, and distribution comparisons across categories.

The analysis provided crucial insights into the dataset, revealing significant skewness, the importance of certain features like Danceability, Energy, and Loudness, and the role categorical variables play in driving Views. These findings are critical for guiding the modeling process and optimizing predictions.

**Summary Statistics**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Danceability | Energy | Loudness | Speechiness | Acousticness | Liveness | Valence | Tempo | Views | Likes | Comments | Duration\_minutes |
| Count | 14385 | 14385 | 14385 | 14385 | 14385 | 14385 | 14385 | 14385 | 14385 | 14341 | 14312 | 14385 |
| Mean | 0.626995 | 0.652823 | 0.830958 | 0.09113 | 0.266489 | 0.190069 | 0.540531 | 121.5316 | 1.1E+08 | 765145.5 | 32864.45 | 3.756478 |
| Std | 0.157222 | 0.196866 | 0.074744 | 0.093233 | 0.268388 | 0.163723 | 0.240107 | 29.30683 | 2.86E+08 | 1836994 | 216762.1 | 1.299809 |
| Min | 0 | 0.00125 | 0.09508 | 0 | 1e-06 | 0.015 | 0 | 0 | 1001264 | 1270 | 0 | 0.516667 |
| 25% | 0.525 | 0.528 | 0.803841 | 0.0352 | 0.0384 | 0.0936 | 0.353 | 97.924 | 6732971 | 65168 | 1662 | 3.086 |
| 50% | 0.64 | 0.676 | 0.846643 | 0.0502 | 0.172 | 0.124 | 0.545 | 120.015 | 25314540 | 199538 | 5459.5 | 3.6 |
| 75% | 0.743 | 0.804 | 0.878124 | 0.102 | 0.435 | 0.233 | 0.735 | 140.02 | 89447030 | 653908 | 18513 | 4.221783 |
| Max | 0.975 | 0.998 | 1 | 0.935 | 0.996 | 1 | 0.993 | 220.099 | 8.08E+09 | 50788630 | 16083140 | 76.35805 |

Figure 1.1

In figure 1.1 summary statistics were calculated for each numerical feature (Danceability, Energy, Loudness, Speechiness, Acousticness, Liveness, Valence, Tempo, Duration\_minutes, Views, Likes, and Comments). Key points from this analysis included:

* Central Tendency: Mean and median values highlighted the typical data points for each feature.
* Spread and Variance: Standard deviation and IQR helped identify variability within the dataset.
* Extremes: Minimum and maximum values flagged potential outliers, particularly in Views, Tempo, and Comments.

Insights:

* The heavy skew in Views, Likes, and Comments indicated a distribution dominated by a few extremely popular tracks.
* Features such as Danceability, Energy, and Valence had near-symmetric distributions, suggesting these are evenly distributed among tracks.

**Histogram**

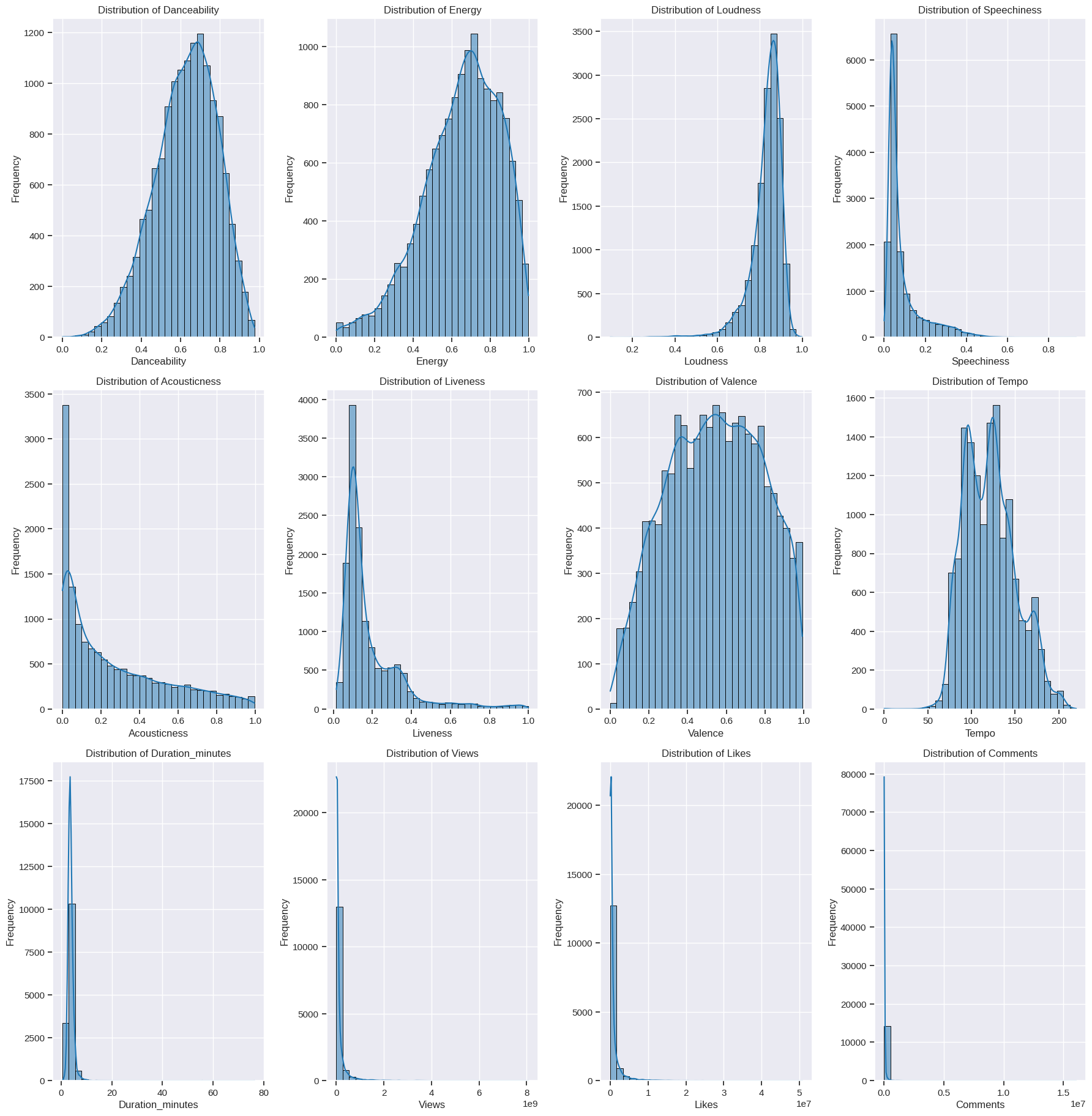


Figure 1.2

In figure 1.2 histograms were plotted to visualize the distribution of each numerical feature. Here’s a summary of the key findings:

1. Danceability:

Distribution: The distribution of Danceability is nearly symmetric and slightly skewed to the left. The majority of the tracks have Danceability scores ranging from approximately 0.5 to 0.8, indicating that most tracks are moderately danceable.

* Insight: Danceability is a key feature for the majority of the tracks, and there are relatively few tracks that are either highly danceable (above 0.8) or not danceable (below 0.5).

1. Energy:

* Distribution: Energy is also nearly symmetric with a slight right skew. The majority of the tracks have Energy scores between 0.5 and 0.9, suggesting that most tracks are energetic.
* Insight: High energy levels are common among the tracks, which could contribute to the tracks' popularity as listeners often favor more energetic music.

1. Loudness:

* Distribution: Loudness is highly skewed to the left. Most tracks have high Loudness values, close to 1.0.
* Insight: The high loudness levels indicate that most tracks are produced with modern loudness standards in mind, aiming for maximum perceived loudness.

1. Speechiness:

* Distribution: The distribution of Speechiness is heavily right-skewed, with most tracks having very low Speechiness values, near 0.1.
* Insight: Most tracks in the dataset contain minimal spoken words or are predominantly musical, with little or no speech.

1. Acousticness:

* Distribution: Acousticness is right-skewed, with a large concentration of tracks having low Acousticness values (below 0.2), indicating that most tracks are not acoustic.
* Insight: This suggests that the majority of tracks are produced with synthetic or electronic instrumentation rather than acoustic instruments.

1. Liveness:

* Distribution: Liveness is heavily right-skewed, with most tracks having low Liveness values (below 0.2), suggesting that the majority of the tracks are studio-produced rather than live performances.
* Insight: The data indicates that most of the tracks were recorded in a controlled studio environment, rather than live settings.

1. Valence:

* Distribution: Valence is almost symmetric, with a slight left skew, and the majority of tracks have Valence values ranging from 0.4 to 0.8.
* Insight: The tracks exhibit a broad range of valence (musical positiveness), with a majority of tracks having a moderately positive emotional tone.

1. Tempo:

* Distribution: The distribution of Tempo is bimodal, with peaks around 100-120 BPM and 140-160 BPM.
* Insight: The bimodal distribution suggests that there are two dominant tempo ranges in the dataset, possibly corresponding to different genres or styles of music.

1. Duration\_minutes:

* Distribution: The Duration\_minutes distribution is right-skewed, with most tracks having a duration between 3 to 4 minutes.
* Insight: Most popular tracks fall within the standard duration for modern songs, around 3 to 4 minutes, which is typical for radio and streaming platforms.

1. **Views**:

* Distribution: The Views distribution is heavily right-skewed, with a large concentration of tracks having lower view counts, and only a few tracks achieving very high views.
* Insight: The distribution of views indicates that a small number of tracks are extremely popular, while most tracks have a relatively moderate view count.

1. Likes:

* Distribution: The Likes distribution is similar to the Views distribution, with a strong right skew. Most tracks have relatively few likes, with a small number of tracks having a very high number of likes.
* Insight: Likes tend to correlate with views, indicating that more popular tracks tend to accumulate a proportionally higher number of likes.

1. Comments:

* Distribution: Comments also exhibit a strong right-skewed distribution, with most tracks having a small number of comments.
* Insight: Similar to likes, the number of comments is heavily concentrated on a few highly popular tracks, while most tracks have relatively few comments.

**Box Plots**

A collage of a graph

Description automatically generated

Figure 1.3

In figure 1.3 box plots provided a visual summary of the distribution and identified outliers in the dataset for each numerical feature:

1. Danceability:

* Distribution: The Danceability box plot shows a fairly symmetric distribution with no significant outliers. The interquartile range (IQR) is centered around 0.5 to 0.7, indicating a moderate level of danceability for most tracks.
* Insights: Most tracks have a balanced level of danceability, with no extreme values, suggesting that they are generally designed to be suitable for dancing without being overly or insufficiently rhythmic.

1. Energy:

* Distribution: The Energy box plot also shows a symmetric distribution, with most data points falling within a high energy range. There are a few minor outliers on the lower end.
* Insights:Tracks are generally high in energy, which aligns with the observation that energetic tracks tend to be more popular. The presence of few low-energy outliers indicates some diversity in track styles.

1. Loudness:

* Distribution: The Loudness box plot shows a high concentration of values near the upper range, with a significant number of lower outliers. The distribution is right-skewed.
* Insights: Most tracks are produced with high loudness, which is consistent with modern production techniques aimed at maximizing loudness for impact. The lower outliers could represent tracks that are softer or intentionally quieter.

1. Speechiness:

* Distribution: The Speechiness box plot reveals a strong right skew with many outliers. Most tracks have very low Speechiness values, indicating that they contain minimal spoken words.
* Insights: Speech is rare in the tracks analyzed, with most being purely musical. The outliers suggest a few tracks with higher speech content, which could represent rap or spoken word elements.

1. Acousticness:

* Distribution: The Acousticness box plot is right-skewed, with most values concentrated at the lower end. The distribution has a few outliers.
* Insights: The majority of tracks are not acoustic, relying more on electronic or synthetic sounds. The outliers may represent tracks with more prominent acoustic elements.

1. Liveness:

* Distribution: The Liveness box plot shows a strong right skew with numerous outliers. Most tracks have low Liveness values, suggesting they are not live recordings.
* Insights: This indicates that most tracks are studio-produced, with only a few tracks potentially being live recordings, as suggested by the outliers.

1. Valence:

* Distribution: The Valence box plot shows a fairly symmetric distribution, with the majority of values falling in the mid-to-high range. There are no significant outliers.
* Insights: Tracks tend to have a generally positive or upbeat emotional tone, with few extreme values suggesting very happy or very sad tones.

1. Tempo:

* Distribution: The Tempo box plot reveals a symmetric distribution with some outliers on the higher end. The IQR is relatively wide, indicating a variety of tempos in the dataset.
* Insights: While most tracks have a moderate tempo, there is a diversity in speed, with some tracks being much faster or slower than the average.

1. Duration\_minutes:

* Distribution: The Duration\_minutes box plot is right-skewed with several high outliers. Most tracks are clustered around a duration of 3 to 4 minutes.
* Insights: The typical track duration aligns with standard commercial music length, but there are some longer tracks that stand out as outliers.

1. **Views**:

* Distribution: The Views box plot shows a significant right skew with many outliers, indicating that while most tracks have moderate view counts, there are some tracks with exceptionally high views.
* Insights: A small number of tracks are highly popular, accumulating significantly more views than the majority.

1. Likes:

* Distribution: The Likes box plot, similar to Views, is heavily right-skewed with many outliers. Most tracks have a moderate number of likes, but some have very high like counts.
* Insights: Likes are concentrated on a few highly popular tracks, indicating a correlation with the view count.

1. Comments:

* Distribution: The Comments box plot shows a strong right skew with numerous outliers. The majority of tracks have a low number of comments, but a few have a large number of comments.
* Insights: Like Likes and Views, comments are heavily concentrated on a few highly discussed tracks, reflecting their popularity and engagement level.

**Correlation Matrix and Heatmap**

A screenshot of a graph

Description automatically generated

Figure 1.4 Correlation Matrix and Heatmap

In figure 1.4 the correlation matrix highlighted relationships between numerical variables:

1. Danceability:

* Correlation: Danceability has a moderate positive correlation with Valence (0.43) and a weaker positive correlation with Loudness (0.23).
* Insight: Tracks that are more danceable tend to have a happier or more positive tone (Valence) and are often louder. This is expected as upbeat and lively songs are typically designed to be danceable.

1. Energy:

* Correlation: Energy has a strong positive correlation with Loudness (0.72) and a moderate positive correlation with Valence (0.35).
* Insight: High-energy tracks are generally louder and have a positive emotional tone. This correlation aligns with the idea that energetic tracks are often associated with excitement and intensity.

1. Loudness:

* Correlation: Loudness shows strong positive correlation with Energy (0.72) and a moderate positive correlation with Danceability (0.23).
* Insight: As expected, louder tracks tend to be more energetic and somewhat more danceable. This reinforces the importance of loudness in creating high-energy tracks.

1. Speechiness:

* Correlation: Speechiness has a weak positive correlation with Danceability (0.25) but is mostly uncorrelated with other features.
* Insight: Tracks with more spoken words (higher Speechiness) can still be danceable, but speech content doesn't significantly influence other musical features.

1. Acousticness:

* Correlation: Acousticness is negatively correlated with Energy (-0.62), Loudness (-0.48), and Danceability (-0.18).
* Insight: Acoustic tracks are typically quieter, less energetic, and less danceable, which is expected as they rely more on natural, softer sounds.

1. Liveness:

* Correlation: Liveness has weak correlations with most features, indicating it is somewhat independent.
* Insight: The degree to which a track sounds live doesn’t strongly correlate with other features, suggesting that live elements can be present across different types of tracks.

1. Valence:

* Correlation: Valence has a moderate positive correlation with Danceability (0.43) and Energy (0.35).
* Insight: Tracks with a happier or more positive tone tend to be more danceable and energetic, which aligns with the idea that positive emotions are often conveyed through lively music.

1. Tempo:

* Correlation: Tempo shows weak correlations with all other features, indicating that tempo is relatively independent.
* Insight: The speed of the track (Tempo) does not strongly influence other features, suggesting that tracks with varying tempos can have different characteristics in terms of loudness, energy, etc.

1. **Views**:

* Correlation: Views are strongly correlated with Likes (0.88) and moderately correlated with Comments (0.40).
* Insight: Tracks with higher view counts tend to have more likes and comments, which is logical as engagement generally increases with popularity. The weaker correlation with Comments suggests that while comments do increase with views, the increase is not as pronounced as with likes.

1. Likes:

* Correlation: Likes are strongly correlated with Views (0.88) and moderately correlated with Comments (0.62).
* Insight: The strong correlation with views confirms that likes are a good proxy for popularity. The moderate correlation with comments indicates that likes and comments often go hand-in-hand, but not always.

1. Comments:

* Correlation: Comments are moderately correlated with Likes (0.62) and Views (0.40).
* Insight: Comments are more strongly correlated with likes than views, suggesting that people who like a track are also more likely to comment.

1. Duration\_minutes:

* Correlation: Duration\_minutes shows weak correlations with other features.
* Insight: The length of the track doesn’t have a strong relationship with other features, indicating that track duration is an independent factor.

**Bar chart for Top 10 Artists**

**A screenshot of a graph

Description automatically generated**

Figure 1.5 The Average Metrics by Artist (Top 10 Artists)

In figure 1.5 we analyzed the performance metrics for top 10 artists:

1. Danceability:

* Top Artists: Shakira (0.78), Bruno Mars (0.75), and Ed Sheeran (0.73) have the highest danceability scores among the top 10 artists.
* Lowest: Coldplay has the lowest danceability score (0.47).
* Insight: Danceable tracks are a key feature for artists like Shakira and Bruno Mars, who are known for their upbeat and rhythmic songs. Coldplay's lower score reflects their more mellow and introspective music style.

1. Energy:

* Top Artists: Daddy Yankee (0.81) and Katy Perry (0.75) have the highest energy levels.
* Lowest: Ed Sheeran (0.40) and Coldplay (0.57) have lower energy levels.
* Insight: Daddy Yankee's high energy score aligns with his reggaeton genre, which is characterized by fast-paced and energetic music. Ed Sheeran and Coldplay's lower energy scores reflect their focus on softer, more acoustic tracks.

1. Speechiness:

* Top Artists: Wiz Khalifa (0.11) and Shakira (0.08) have higher speechiness scores.
* Lowest: Coldplay (0.03) and Katy Perry (0.04) have the lowest speechiness scores.
* Insight: Artists like Wiz Khalifa, who often incorporate rap or spoken word elements into their music, tend to have higher speechiness. Coldplay's lower score is expected due to their focus on sung lyrics rather than spoken word.

1. Acousticness:

* Top Artists: Ed Sheeran (0.53) and Cocomelon (0.33) have the highest acousticness scores.
* Lowest: Katy Perry (0.04) and Justin Bieber (0.05) have the lowest scores.
* Insight: Ed Sheeran's music is highly acoustic, which aligns with his style of using guitars and minimal production. Katy Perry's pop-focused production results in a low acousticness score.

1. Liveness:

* Top Artists: Katy Perry (0.25) and Daddy Yankee (0.20) have the highest liveness scores.
* Lowest: Ed Sheeran (0.12) and Cocomelon (0.14) have lower liveness scores.
* Insight: Tracks with higher liveness scores, like those by Katy Perry, tend to feel more like live performances, which can engage audiences in a different way compared to highly produced studio tracks.

1. Valence:

* Top Artists: Ed Sheeran (0.82) and Bruno Mars (0.65) have the highest valence scores, indicating happier or more positive tracks.
* Lowest: Coldplay (0.29) has the lowest valence score.
* Insight: Ed Sheeran's music tends to be more uplifting or positive, as reflected in his high valence score. Coldplay's lower score reflects their more melancholic and introspective style.

1. Tempo:

* Top Artists: Cocomelon (158.88 BPM) and Daddy Yankee (119.97 BPM) have the highest tempos.
* Lowest: Ed Sheeran (105.69 BPM) and Katy Perry (125.00 BPM) have slower tempos.
* Insight: Cocomelon's high tempo is consistent with its target audience of children, where faster tempos can capture and maintain attention. Ed Sheeran's slower tempo aligns with his more relaxed and acoustic musical style.

**Diversity in Styles**: The top 10 artists show a diverse range of musical styles, from high-energy dance tracks by Daddy Yankee and Shakira to more acoustic and mellow tracks by Ed Sheeran and Coldplay.

**Genre Influence**: The metrics align well with the genres these artists are known for. For example, high danceability and energy for Shakira and Daddy Yankee align with their Latin pop and reggaeton styles, while Ed Sheeran's acousticness and valence scores reflect his singer-songwriter style.

**Coldplay's Unique Profile**: Coldplay stands out with lower danceability, energy, and valence, reflecting their more introspective and slower music compared to the other top artists.

**Radar Plot**

A screenshot of a graph

Description automatically generated

Figure 1.6 Radar Plot For Top Artists Metric Performance

1. Katy Perry:

* Strongest Features: Danceability and Energy
* Weakest Features: Acousticness and Liveness
* Insight: Katy Perry's music is characterized by high danceability and energy, making her songs suitable for dance floors and energetic performances. The low acousticness and liveness suggest highly produced studio tracks.

1. Ed Sheeran:

* Strongest Features: Acousticness
* Weakest Features: Speechiness and Liveness
* Insight: Ed Sheeran’s music is heavily acoustic, which aligns with his style of using guitars and minimal production. The low liveness and speechiness indicate a focus on melodic, sung lyrics rather than live or spoken elements.

1. Cocomelon:

* Strongest Features: Tempo and Acousticness
* Weakest Features: Liveness and Valence
* Insight: Cocomelon's tracks are fast-paced and acoustic, fitting the children’s music genre, which aims to keep young listeners engaged. The low liveness and valence suggest a more produced and less emotionally varied sound.

1. Daddy Yankee:

* Strongest Features: Energy
* Weakest Features: Acousticness and Liveness
* Insight: Daddy Yankee’s music is highly energetic, fitting the reggaeton genre. The low acousticness indicates electronic and beat-driven production, while the low liveness suggests a focus on studio production rather than live elements.

1. Justin Bieber:

* Strongest Features: Danceability and Acousticness
* Weakest Features: Liveness and Valence
* Insight: Justin Bieber’s music combines danceability with acoustic elements, appealing to both pop and acoustic listeners. The low liveness and valence suggest a balance between upbeat and reflective songs.

1. Dua Lipa:

* Strongest Features: Danceability and Energy
* Weakest Features: Acousticness and Liveness
* Insight: Dua Lipa’s tracks are danceable and energetic, aligning with her pop and dance music style. The low acousticness and liveness reflect a polished, studio-produced sound.

1. Wiz Khalifa:

* Strongest Features: Speechiness and Energy
* Weakest Features: Acousticness and Valence
* Insight: Wiz Khalifa’s music, with high speechiness and energy, is typical of rap and hip-hop, where lyrics and rhythm take center stage. The lower acousticness indicates a preference for electronic beats over live instrumentation.

1. Shakira:

* Strongest Features: Danceability and Energy
* Weakest Features: Acousticness and Liveness
* Insight: Shakira’s music is characterized by high danceability and energy, fitting her Latin pop style. The low acousticness and liveness suggest a focus on produced, rhythm-driven tracks over live or acoustic performances.

1. Bruno Mars:

* Strongest Features: Danceability and Valence
* Weakest Features: Acousticness and Liveness
* Insight: Bruno Mars’s music is danceable and upbeat, which is typical for his blend of pop, funk, and R&B. The low acousticness and liveness point to a focus on polished production rather than live elements.

1. Coldplay:

* Strongest Features: Acousticness
* Weakest Features: Danceability and Liveness
* Insight: Coldplay’s music is highly acoustic, reflecting their alternative rock style. The low danceability and liveness indicate a focus on more reflective, mellow tracks rather than upbeat or live-performance-oriented music.

**Scatter plot**

A graph with colorful dots

Description automatically generated

Figure 1.7 Scatter plot for View vs Likes

In figure 1.7 we can observe views vs. likes:

It shows a strong positive correlation. The spread at higher views suggests variability in how viewers engage with the content by liking it, with some videos having a high view count but relatively fewer likes.

A graph showing a number of colored squares

Description automatically generated

Figure 1.8 Scatter Plot for Views vs Comments

In figure 1.8 we can observe views vs. comments:

There is a clear positive correlation, indicating that as the number of views increases, so do the comments. However, there is a spread at higher levels of views, suggesting that some highly viewed videos receive disproportionately more or fewer comments.

**Bar Plot for Top Tracks**

**A graph of different colored rectangular shapes

Description automatically generated**

Figure 1.9 Bar Plot for Top Tracks

For figure 1.9 we gained the following insights:

1. "Despacito" Dominates:

* "Despacito" leads by a significant margin, with over 8.1 billion views. This track outperforms the others by a substantial margin, suggesting it was a global phenomenon with widespread appeal.

1. Popular Tracks with Billions of Views:

* Tracks like "Shape of You" and "See You Again (feat. Charlie Puth)" also have a massive view count, each exceeding 5.8 billion views. These tracks are recognized globally and have maintained their popularity over time.
* "Wheels on the Bus," "Uptown Funk (feat. Bruno Mars)," and "Gangnam Style" also rank highly, with views ranging from 4.7 to 4.9 billion. These songs either have a catchy tune or a cultural impact that resonated with a broad audience.

1. Diverse Genres:

* The tracks span various genres, including pop, children's music, and K-pop, highlighting the broad appeal of music across different audience segments.
* For example, "Wheels on the Bus" targets a younger audience, while "Gangnam Style" is a global pop hit with a viral music video.

1. Legacy Hits:

* Songs like "Sugar," "Roar," and "Counting Stars" show that older tracks continue to receive high viewership, demonstrating their enduring popularity.

1. Engagement Factor:

* The tracks with the highest views often have elements that engage a wide audience, such as memorable melodies, strong cultural impact, or viral music videos. This can be an important consideration for predicting or influencing future hits.

**Pair Plot**

A screenshot of a graph

Description automatically generated

Figure 2 Pair plot for all the variables

In figure 2 we observed the following:

1. Scatterplots:

* The scatterplots in the lower triangle of the pair plot show pairwise relationships between variables.
* Most scatterplots appear scattered with no clear linear relationship, which suggests that the relationships between these variables may be non-linear or weak.
* There are a few plots with more noticeable patterns (e.g., between Energy and Loudness, and between Loudness and Views), indicating a stronger relationship between those pairs of variables.

1. Distributions:

* The diagonal plots show the distribution of individual variables.
* Variables like Danceability, Energy, and Valence have relatively normal distributions, though they may be slightly skewed.
* Variables like Loudness, Speechiness, Acousticness, and Liveness show more skewed distributions.
* Views, Likes, and Comments have distributions with significant skewness, highlighting the presence of a few tracks with extremely high engagement compared to the rest.

1. Key Relationships:

* Energy vs. Loudness: A strong positive linear relationship is evident, meaning tracks with higher energy levels tend to be louder.
* Loudness vs. Views: There is a visible trend where tracks that are louder tend to have higher views, though this relationship is not strictly linear.
* Speechiness vs. Acousticness: This plot shows an interesting non-linear relationship, suggesting that tracks with higher speechiness may have lower acousticness, but the pattern is complex.
* Tempo vs. Views: The scatterplot for these two variables shows a wide spread with no clear pattern, indicating a weak relationship.

1. Outliers:

* Several scatterplots reveal outliers, especially in the Views, Likes, and Comments variables. These outliers represent tracks that have received exceptionally high levels of engagement.

**Bar Charts- Categorical Variables**

A close-up of several blue bars

Description automatically generated

Figure 2.1 Bar plots of Categorical Variables

The figure 2.1 provided contains four bar charts representing the distribution of various categorical variables within your dataset: Album\_type, Key, Licensed, and official\_video. Below is an interpretation of each chart along with insights:

**1. Distribution of Album\_type**

* **Description:** The bar chart for Album\_type shows the number of tracks classified as albums, singles, or compilations.
* **Insights:**
  + **Albums:** The majority of tracks in the dataset are categorized as albums, which suggests that full albums remain a significant format in the music industry.
  + **Singles:** There is also a substantial number of tracks classified as singles, indicating the importance of releasing individual tracks, possibly as promotional or lead singles from an album.
  + **Compilations:** Tracks in compilations are the least represented, suggesting that compilations are less common or less impactful compared to albums and singles in this dataset.

**2. Distribution of Key**

* **Description:** The bar chart for Key shows the number of tracks in each musical key, ranging from 0 to 11 (representing the 12 possible keys in Western music).
* **Insights:**
  + **Key Variation:** There is a fairly even distribution of tracks across most keys, with slightly higher counts for certain keys like 0, 5, and 7.
  + **Musical Preference:** The relatively even distribution suggests that tracks are produced in a wide range of keys, reflecting diverse musical preferences and styles.

**3. Distribution of Licensed**

* **Description:** The bar chart for Licensed shows the number of tracks that are licensed versus those that are not.
* **Insights:**
  + **Licensed Tracks:** A significant majority of tracks are licensed, indicating that most of the content in the dataset is professionally managed and distributed.
  + **Non-Licensed Tracks:** A smaller portion of tracks are not licensed, which may represent independent releases or content not distributed under a formal licensing agreement.

**4. Distribution of official\_video**

* **Description:** The bar chart for official\_video shows the number of tracks that have official videos versus those that do not.
* **Insights:**
  + **Tracks with Official Videos:** The majority of tracks have official videos, highlighting the importance of visual content in promoting music and driving engagement on platforms like YouTube.
  + **Tracks without Official Videos:** There is still a notable number of tracks without official videos, which may reflect lower-budget releases, niche genres, or tracks that rely on other forms of promotion.

**Violin Plots**

**A diagram of different types of trees

Description automatically generated**

Figure 2.2 Violin Plot for Album Type

**A graph showing a graph

Description automatically generated with medium confidence**

Figure 2.3 Violin Plot for Official\_video

In figure 2.2 and figure 2.3 violin plots were used to explore the relationship between categorical variables and Views:

**Views Distribution by Album Type:**

1. Albums Have the Widest Distribution:

* Tracks that are part of albums have a broad range of views, indicating that albums cater to a wide audience, with some tracks achieving extremely high view counts.
* The density of points is relatively consistent, showing that albums generally perform well, with many tracks achieving substantial views.

1. Singles Show High Variability:

* Singles exhibit a more pronounced peak in their distribution, suggesting that while many singles achieve moderate success, a few stand out with significantly higher views.
* The distribution also shows some singles reaching very high view counts, possibly indicating hit singles that outperform even popular album tracks.

1. Compilations Have the Narrowest Distribution:

* Compilations show the narrowest distribution and generally lower view counts compared to albums and singles.
* This could suggest that compilations, which often consist of selected tracks from various sources, might not attract as much attention as dedicated albums or hit singles.

1. Extremes in View Counts:

* The violin plot shows that, while there are tracks across all album types that achieve very high view counts, singles and albums seem to have more instances of extreme high views, potentially indicating blockbuster tracks that resonate widely.

1. Consideration for Releases:

* For artists and labels, releasing music as singles or part of albums might be more strategic if the goal is to achieve higher view counts. Compilations, while useful for aggregating content, may not drive as much engagement.

**Views Distribution by Official Video:**

1. Tracks with Official Videos:

* Tracks that have an official video (true) exhibit a much wider range of views, including many tracks that have achieved extremely high view counts.
* The distribution shows that the majority of these tracks have substantial view counts, with a dense concentration of points and a broad spread, indicating that official videos significantly contribute to higher viewership.

1. Tracks without Official Videos:

* Tracks without an official video (false) generally show a narrower distribution with lower view counts.
* The density is lower, and the spread is limited compared to tracks with official videos, suggesting that not having an official video may limit the potential viewership.

1. Impact of Official Videos on Viewership:

* The stark difference in the distribution between tracks with and without official videos suggests that having an official video is a key driver of higher view counts.
* This could be due to the added promotional power of videos, as they provide a visual element that likely increases audience engagement and sharing.

1. Consideration for Artists and Labels:

* For artists and labels aiming to maximize view counts, investing in official music videos appears to be a crucial strategy. The data strongly supports the correlation between official videos and higher viewership.

## Data Cleansing

**Exclusion of Variables**

1. Likes:

High Correlation with Target Variable:

* The Likes and Views columns are typically highly correlated because they both measure user engagement, with views often being a precursor to likes. Including both in a predictive model might not add much additional value and could lead to multicollinearity, which can skew model interpretations and performance.
* This is more actionable for artists and producers who are looking to improve the likelihood of success based on the song's characteristics.

1. Channel, Artist, Album, Track, Comments, and Title

* High Cardinality and Non-numeric: Columns like Channel, Artist, Album, and Track are categorical with many unique values, which would complicate the model and increase the risk of overfitting. They are more identifiers than predictive features.
* Redundant Post-Release Metric: Comments is a post-release metric like Likes and is highly correlated with views, which could introduce redundancy and potential bias in the model.
* Complexity in Text Processing: Title is free text, which would require advanced processing. Its influence on views is likely less significant compared to other song attributes, making it less valuable to include.

## Summary

The data exploration revealed essential patterns, such as the dominance of a few popular tracks in driving overall engagement, the significance of high-energy and danceable features, and the importance of having an official video for maximizing views. These findings lay a solid foundation for the modeling process and provide actionable insights for stakeholders looking to optimize track performance

# Data Preparation and Feature Engineering

## Data Preparation Needs

The data preparation phase is a critical step that ensures the dataset is optimized for model building, leading to more accurate and interpretable results. In this section, we focus on handling missing values, outliers, and skewness while ensuring that the data remains relevant and reflective of real-world conditions. Key activities included imputation, exclusion of irrelevant or redundant variables, normalization, and handling of outliers. Additionally, specific criteria were applied to focus on a realistic benchmark for the business problem.

## Handling Missing Values

Handling missing data is a crucial step in the data preparation process as it directly impacts the integrity of the dataset and, consequently, the reliability of the model outcomes. In this project, missing values were identified in several key features, including Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Liveness, Valence, Tempo, and Duration\_minutes. The proportion of missing values across these features was minimal, approximately 0.009%, which made the handling process straightforward yet critical for maintaining data quality.

**Process Overview:**

1. **Initial Assessment:**

A thorough examination of the dataset revealed that missing values were present in a few numerical and categorical columns. The overall percentage of missing data was low, which presented an opportunity to address the issue without resorting to complex imputation techniques.

1. **Decision to Drop Missing Values:**

Given the minimal proportion of missing data relative to the total dataset, the decision was made to drop rows containing missing values in the affected columns. This approach was selected to avoid the introduction of bias or errors that might arise from imputing values, which could potentially alter the dataset's inherent structure or misrepresent the data's natural distribution.

1. **Columns Affected:**

The columns with missing values were as follows:

* + - Danceability
    - Energy
    - Key
    - Loudness
    - Speechiness
    - Acousticness
    - Liveness
    - Valence
    - Tempo
    - Duration\_minutes

By focusing on complete cases, the integrity of the dataset was preserved, ensuring that all data points used in subsequent analyses were based on observed data rather than estimates.

**Rationale for Dropping Missing Values:**

* **Data Integrity:** Dropping rows with missing values maintains the dataset’s integrity, ensuring that the analysis and modeling are based solely on actual, observed data points. This avoids the risk of introducing biases or inaccuracies that could arise from imputation.
* **Minimal Impact on Dataset Size:** Given that the proportion of missing values was exceptionally low, the impact of dropping these rows on the overall dataset size was negligible. This ensured that the dataset remained sufficiently large for robust analysis and model training.
* **Enhanced Model Interpretability:** By using a dataset without missing values, the model outputs become easier to interpret and understand. The results are based on real data, providing clearer insights and more reliable predictions without the complications that imputed data might introduce.

**Conclusion:**

The handling of missing values through deletion was a strategic choice that prioritized data quality and model accuracy. This approach ensured that the dataset remained robust and free from the distortions that imputation could introduce, ultimately supporting the creation of a reliable and interpretable predictive model.

## Handling Skewness

Skewness in data refers to the asymmetry in the distribution of values for a particular feature. In predictive modeling, especially for linear models, skewed data can lead to biased estimates, poor model performance, and difficulties in interpreting results. To address this, specific features in the dataset that exhibited significant skewness were transformed using logarithmic transformations.

**Process Overview:**

1. **Identification of Skewed Features:**
   * The skewness of each numerical feature was calculated to identify those that deviated significantly from a normal distribution. Features with a skewness greater than 1 (indicating right skew) or less than -1 (indicating left skew) were considered for transformation.
   * The following features were identified as having significant skewness:
     + **Loudness**: Exhibited a strong left skew due to its measurement in decibels (dB), where values typically ranged from -60 dB to 0 dB.
     + **Speechiness**: Showed a right skew, with most tracks having low speech content.
     + **Liveness**: Also displayed a right skew, reflecting that most tracks were studio-produced rather than live recordings.
2. **Logarithmic Transformation:**

To mitigate the effects of skewness, a log base 2 transformation was applied to the identified features. Log transformation compresses the range of the data, pulling in the extreme values and bringing the distribution closer to normality.

1. **Transformation Details:**
   * + **Loudness**: A log base 2 transformation was applied to compress the wide range of negative values. Post-transformation, Loudness was normalized to a 0-1 scale to align it with other features.
     + **Speechiness**: The log transformation reduced the right skew by compressing the high values and spreading out the lower ones, making the distribution more uniform.
     + **Liveness**: Similarly, the log transformation addressed the right skew, resulting in a more balanced distribution of live and studio recordings.

**Rationale for Log Transformations:**

* **Reduction of Skewness:** The primary goal of applying log transformations was to reduce skewness in the data, thereby improving the fit of the model and ensuring that the underlying assumptions of normality and homoscedasticity were better met.
* **Improved Model Performance:** Skewed data can lead to biased predictions, especially in models that assume a normal distribution of errors. By transforming the data, the models can learn more effectively from the features, leading to more accurate and reliable predictions.
* **Enhanced Interpretability:** Features with reduced skewness are easier to interpret in the context of the model. For example, the transformed Loudness feature, now on a 0-1 scale, allows for straightforward comparison with other features, aiding in the interpretation of the model’s outputs.

**Conclusion:**

Handling skewness through log transformations was a vital step in preparing the dataset for modeling. This process ensured that the features were more normally distributed, improving both model performance and interpretability. By addressing skewness, the models built on this data are more robust, reliable, and capable of providing meaningful insights.

## Handling Outliers

Outliers are extreme values in the dataset that can significantly influence the results of a predictive model, potentially leading to skewed estimates and overfitting. In this project, a systematic approach was employed to identify and handle outliers in the dataset, ensuring that these values did not adversely affect model performance.

**Outlier Identification and Threshold Setting:**

**Outlier Detection:**

* For each numerical feature, the mean and standard deviation were calculated. Outliers were identified as data points that fell beyond three standard deviations from the mean, representing values that were unusually high or low compared to the rest of the data.
* The three standard deviations threshold was selected to capture only the most extreme outliers, thereby preserving the natural variability in the data while addressing potential distortions.

**Outlier Handling Strategy:**

1. **Capping and Flagging:**
   * Outliers were capped at their respective upper and lower thresholds to reduce their influence on the model. In addition, flag indicators were created for each feature to mark whether a data point was an outlier before capping. This allows for tracking and analysis of these flagged outliers in the model’s predictions.
   * **Features with Capped and Flagged Outliers:** Capping and flagging were applied to Danceability, Energy, Loudness, Speechiness, Liveness, and Duration\_minutes.
2. **Exceptions to Capping and Flagging:**
   * **Tempo:** Due to the minimal number of outliers (only 2 below and 5 above the threshold), capping and flagging were not applied to the Tempo feature. This decision was made to preserve the natural variability of tempo, which is essential for capturing different musical styles.
   * **Views:** As the target variable, Views was not subjected to outlier handling. Preserving its original distribution was vital for maintaining the accuracy and integrity of the model’s predictions.

**Rationale for Outlier Handling:**

* **Controlled Influence on Model Training:** Capping outliers ensures that extreme values do not disproportionately influence the model. The creation of flag indicators allows these outliers to be analyzed in post-model assessments, offering insights into their impact on predictions.
* **Enhanced Model Stability:** By reducing the influence of outliers, the model’s performance becomes more stable, leading to more reliable and generalizable predictions.
* **Preservation of Data Integrity:** The decision to not cap or flag the Tempo and Views features ensures that essential characteristics of the data are preserved, maintaining the overall quality of the dataset.

**Conclusion:**

The handling of outliers in this project was carried out with precision, focusing on features where outliers could negatively impact the model while making careful exceptions where appropriate. This balanced approach, combining capping with flagging and selective exclusion, ensured that the dataset was optimized for robust, reliable modeling while retaining the integrity of key features.

# Model Exploration

## Modeling Approach/Introduction

## Full Linear Regression

**Objective:** The Full Linear Regression model was implemented as a baseline to predict YouTube views based on all available song features. The primary objective was to assess the relationship between each predictor and the target variable (Views), allowing for a comprehensive understanding of how different song characteristics influence viewership.

**Model Implementation:** The linear regression model incorporated all the available features, including both numerical and categorical variables (with categorical variables encoded as dummy variables). This approach provided insights into the significance and impact of each feature before applying any feature selection or dimensionality reduction techniques.

Key predictors in the model included Danceability, Energy, Loudness, Speechiness, Acousticness, Liveness, Valence, Tempo, Duration\_minutes, and several others. Among these, Loudness and Danceability had significant positive coefficients, suggesting that higher values in these features correlate with increased YouTube views. In contrast, features like Energy and Speechiness had negative coefficients, indicating that higher values might decrease the view counts.

**Performance Metrics:**

* **Training Set:**
  + Root Mean Squared Error (RMSE): 24,651,582.97
  + Mean Absolute Error (MAE): 19,518,400.85
  + Mean Absolute Percentage Error (MAPE): 313.47%
* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,623,324.10
  + Mean Absolute Error (MAE): 19,526,341.15
  + Mean Absolute Percentage Error (MAPE): 307.51%

These metrics indicate that while the model captures some of the relationships between the features and views, it still produces substantial errors. The high RMSE and MAE values, combined with the MAPE exceeding 300%, suggest that the model struggles with accurately predicting YouTube views, especially for tracks with extreme values (either very high or very low view counts).

**Key Insights:**

1. **Feature Significance:** The model highlighted the importance of features such as Loudness, Danceability, and Duration\_minutes, which were positively correlated with YouTube views. These insights align with industry observations where louder, more danceable tracks with optimal durations tend to perform better on digital platforms.
2. **Model Limitations:** The high error metrics and large MAPE indicate potential multicollinearity among predictors, as well as the presence of non-linear relationships that the linear model cannot capture effectively. The complexity of music data likely requires more sophisticated models to better capture these dynamics.

**Conclusion:** The Full Linear Regression model provides a foundational understanding of the factors influencing YouTube viewership. However, the significant error metrics and inability to capture non-linear interactions suggest the need for more advanced modeling techniques. This model serves as a useful benchmark, setting the stage for further improvements through feature selection and non-linear models.

## Forward Linear Regression

**Objective:** The Forward Linear Regression model was implemented to refine the prediction of YouTube views by selectively adding the most significant predictors to the model. Unlike the Full Linear Regression, which uses all available features, the Forward Selection method builds the model by starting with no predictors and adding them one by one based on their statistical significance, as measured by the Akaike Information Criterion (AIC).

**Model Implementation:** The Forward Selection process started with an empty model and iteratively added predictors that contributed the most to reducing the AIC score. This method allowed for the inclusion of only those features that had the strongest relationship with the target variable (Views), helping to reduce the potential for overfitting and multicollinearity.

Key predictors identified by the Forward Selection process included:

* Licensed\_True: Whether the track is licensed, positively impacting views.
* Loudness: Reflecting the track's loudness, which positively correlated with views.
* Duration\_minutes: The length of the track, also positively correlated with views.
* official\_video\_True: Whether the track has an official video, positively impacting views.
* Album\_type\_single: Whether the track is a single, with a negative correlation indicating that singles tend to have fewer views compared to other album types.
* Danceability, Energy, Tempo, and flag\_outlier\_Loudness were also retained, indicating their importance in predicting views.

**Performance Metrics:**

* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,625,953.69
  + Mean Absolute Error (MAE): 19,528,362.43
  + Mean Absolute Percentage Error (MAPE): 306.99%

The performance metrics for the Forward Selection model were nearly identical to those of the Backward Elimination and Full Regression models. This suggests that the added predictors captured most of the relevant information while excluding redundant or less significant features.

**Key Insights:**

1. **Selective Feature Inclusion:** The Forward Selection process confirmed that features like Licensed\_True, Loudness, Duration\_minutes, and official\_video\_True are crucial for predicting YouTube views. These features align well with business practices where official licensing, loudness, optimal track length, and the presence of an official video are key drivers of digital success.
2. **Model Simplification:** By adding predictors one at a time, the Forward Linear Regression model achieved similar predictive power with fewer predictors compared to the Full Linear Regression model. This simplification can improve model interpretability and reduce the risk of overfitting, making it more practical for real-world applications.

**Conclusion:** The Forward Linear Regression model effectively balances model complexity and predictive accuracy by focusing only on the most impactful features. This approach not only enhances interpretability but also ensures that the model remains robust against multicollinearity and overfitting. The consistency in performance metrics across the models suggests that the selected features are indeed the most relevant for predicting YouTube views, making this model a strong candidate for practical application.

## Backward Linear Regression

**Objective:** The Backward Elimination method was applied to refine the regression model by systematically removing the least significant predictors. This approach starts with all available features and removes them one by one based on their statistical significance, specifically by evaluating the Akaike Information Criterion (AIC). The goal was to achieve a model that balances predictive accuracy with simplicity by retaining only the most impactful variables.

**Model Implementation:** The Backward Elimination process began with the full set of predictors and iteratively removed the least significant ones according to their AIC score. This method ensures that the final model contains only those predictors that contribute meaningfully to explaining the variance in YouTube views.

The final model included key features such as:

* Danceability: A positive correlation with views.
* Energy: A negative correlation, indicating that less energetic tracks may attract more views.
* Loudness: A strong positive correlation with views, showing that louder tracks tend to perform better.
* Duration\_minutes: The track length, which also showed a positive correlation with views.
* Additional features like Licensed\_True, official\_video\_True, and Tempo were also retained, indicating their importance.

**Performance Metrics:**

* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,625,953.69
  + Mean Absolute Error (MAE): 19,528,362.43
  + Mean Absolute Percentage Error (MAPE): 306.99%

The performance metrics for the Backward Elimination model were consistent with the Forward Selection model, reflecting the effectiveness of the AIC in identifying the most relevant predictors while minimizing unnecessary complexity.

**Key Insights:**

1. **Model Simplification:** The Backward Elimination process successfully reduced the number of predictors while maintaining strong predictive performance. This approach highlighted the most critical features, ensuring that the model remains both interpretable and efficient.
2. **Predictor Significance:** The retained predictors, such as Loudness, Danceability, and Duration\_minutes, align well with industry insights, confirming their importance in driving YouTube views.

**Conclusion:** The Backward Elimination model offers a streamlined and focused approach to predicting YouTube views by selectively retaining only the most significant features. This method effectively balances model complexity with interpretability, making it a valuable tool for understanding the key drivers of digital success. The model's ability to maintain strong performance with fewer variables underscores the importance of a strategic feature selection process in regression modelling.

## Stepwise Linear Regression

**Objective:** The Stepwise Linear Regression method was implemented to refine the prediction of YouTube views by combining both Forward Selection and Backward Elimination techniques. This hybrid approach aimed to build a model that includes only the most significant predictors while simultaneously eliminating those that do not contribute meaningfully to the model’s performance, as determined by the Akaike Information Criterion (AIC).

**Model Implementation:** The Stepwise Regression process begins with no predictors and alternately adds or removes features based on their impact on the AIC score. This dynamic process allows for a more flexible model-building approach, where variables can be added or removed at each step, depending on their statistical significance. The process continues until no further improvement in AIC can be achieved by either adding or removing variables.

The final model included key features such as:

* Licensed\_True: Tracks with licenses tend to have higher views.
* Loudness: Higher loudness levels correlate positively with views.
* Duration\_minutes: The track's duration, positively impacting views.
* official\_video\_True: The presence of an official video boosts views.
* Other features like Danceability, Energy, Tempo, and flag\_outlier\_Loudness were also retained, reflecting their relevance in predicting views.

**Performance Metrics:**

* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,625,953.69
  + Mean Absolute Error (MAE): 19,528,362.43
  + Mean Absolute Percentage Error (MAPE): 306.99%

The performance of the Stepwise Regression model was similar to that of the Backward Elimination and Forward Selection models, indicating that the predictors identified are robust across different selection methods. The consistency in metrics also suggests that the key variables driving YouTube views have been effectively captured.

**Key Insights:**

1. **Model Flexibility:** The Stepwise Regression approach allowed for a more flexible model-building process, where features could be added or removed dynamically. This flexibility helps in ensuring that only the most significant features remain in the final model, leading to a more parsimonious solution.
2. **Feature Relevance:** The retained features, such as Licensed\_True, Loudness, and Duration\_minutes, align with industry knowledge, reinforcing their importance in driving YouTube viewership.

**Conclusion:** The Stepwise Linear Regression model offers a balanced approach to feature selection, combining the strengths of both Forward Selection and Backward Elimination. By focusing on reducing AIC, the model effectively captures the most impactful predictors while minimizing the risk of overfitting. This method ensures that the final model is both accurate and interpretable, making it a reliable tool for predicting YouTube views.

## Regressor Tree

**Objective:** The Decision Tree Regressor was implemented to explore non-linear relationships between the song features and YouTube views. Unlike linear models, decision trees can capture complex interactions between variables without requiring prior assumptions about the form of these relationships. The primary goal was to see if a tree-based approach could provide better predictive accuracy or new insights into the data.

**Model Implementation:** A Decision Tree Regressor was trained using grid search to optimize key hyperparameters, such as max\_depth, min\_samples\_split, and min\_impurity\_decrease. This tuning process aimed to balance the model's complexity and its ability to generalize to unseen data. The final model was selected based on cross-validation performance, ensuring that the chosen tree structure effectively captured the underlying patterns in the data without overfitting.

**Key Predictors Identified:** The final tree structure revealed that features like Duration\_minutes, Loudness, and Tempo were critical decision points in predicting YouTube views. These features repeatedly appeared at the top levels of the tree, indicating their strong influence on the model's predictions.

**Performance Metrics:**

* **Training Set:**
  + Root Mean Squared Error (RMSE): 24,685,842.00
  + Mean Absolute Error (MAE): 19,531,587.76
  + Mean Absolute Percentage Error (MAPE): 314.08%
* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,652,330.93
  + Mean Absolute Error (MAE): 19,539,825.25
  + Mean Absolute Percentage Error (MAPE): 306.53%

The Decision Tree model produced similar RMSE and MAE values compared to the linear models, indicating that while it could capture some non-linear relationships, it did not significantly outperform the simpler models. The tree structure, however, provides an interpretable visualization of how different features interact to affect views.

**Key Insights:**

1. **Non-Linear Relationships:** The Decision Tree Regressor was able to capture interactions between features that linear models might miss. For example, the model could differentiate between different ranges of Loudness and Duration\_minutes, making more tailored predictions based on these specific thresholds.
2. **Model Interpretability:** The decision tree provides a clear and intuitive structure, where the splits at each node represent important decision points that influence YouTube views. This makes it easier to understand the relative importance of different features and how they contribute to the final prediction.

**Conclusion:** The Decision Tree Regressor offered valuable insights into the data by capturing non-linear interactions between song features. While its performance was comparable to that of linear models, the tree's ability to visualize decision rules makes it a useful tool for exploring complex relationships in the dataset. However, the model's similar error metrics suggest that, in this case, the non-linear relationships it identified did not significantly improve predictive accuracy over the simpler linear models.

## Random Forest

**Objective:** The Random Forest Regressor was implemented to enhance the predictive power by leveraging an ensemble of decision trees. Unlike a single decision tree, which can be prone to overfitting, a Random Forest aggregates the predictions of multiple trees to reduce variance and improve generalization. The goal was to capture complex interactions between features and improve the accuracy of predicting YouTube views.

**Model Implementation:** The Random Forest model was trained using 500 decision trees (n\_estimators=500) to ensure robust performance. Each tree was built using a random subset of features and data samples, which helps in reducing overfitting and improving the model's ability to generalize to new data. The feature importance scores generated by the model provided insights into which variables were most influential in predicting YouTube views.

**Key Predictors Identified:** The Random Forest identified Duration\_minutes, Loudness, and Tempo as the most significant features. Other important features included Danceability, Energy, and Valence, which also contributed to the overall prediction. The model was able to account for complex interactions between these features that might not be captured by simpler models.

**Performance Metrics:**

* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,854,038.08
  + Mean Absolute Error (MAE): 19,935,836.95
  + Mean Absolute Percentage Error (MAPE): 323.13%

The performance of the Random Forest was similar to the Decision Tree model, but with slightly higher RMSE and MAE values. The improvement in stability and reduced variance highlights the Random Forest’s advantage over a single decision tree, though it did not lead to significant gains in predictive accuracy.

**Feature Importance:** The Random Forest model provided a detailed ranking of feature importance:

* **Top Features:** Duration\_minutes, Loudness, and Tempo were identified as the most important features, indicating that tracks with certain durations, loudness levels, and tempos are more likely to achieve higher YouTube views.
* **Lesser Impact Features:** Categorical variables like Key and flag\_outlier indicators had lower importance, suggesting that while they contribute to the model, they do so less significantly than the top features.

**Key Insights:**

1. **Robustness and Generalization:** The Random Forest's aggregation of multiple trees reduces the model's variance, making it more robust to overfitting. This ensures that the model performs well not only on the training data but also on unseen data.
2. **Complex Interactions:** By considering various combinations of features across multiple trees, the Random Forest captures complex interactions between variables, which may be missed by simpler models. This is particularly useful in datasets with high-dimensional features, as in this case.

**Conclusion:** The Random Forest Regressor is a powerful model that excels at capturing complex relationships between features through its ensemble approach. While its performance was comparable to simpler models like Decision Trees and linear models in terms of RMSE and MAE, it provides additional stability and interpretability through feature importance scores. This makes it a valuable tool for understanding the key drivers of YouTube views and for making more reliable predictions.

## Neural Network

**Neural Network Model**

**Objective:** The Neural Network model was implemented to explore the potential non-linear relationships between song features and YouTube views. Given that traditional linear models may fail to capture complex interactions, the neural network was designed to test whether a non-linear approach could improve predictive accuracy.

**Model Implementation:** The neural network was built using a sequential model with multiple hidden layers:

* **Input Layer:** 64 neurons, corresponding to the number of features in the dataset.
* **Hidden Layers:** Two hidden layers, each with 64 and 32 neurons respectively, using ReLU (Rectified Linear Unit) as the activation function to introduce non-linearity.
* **Dropout Layers:** Dropout was applied after each hidden layer to prevent overfitting by randomly setting a fraction of input units to 0 during training.
* **Output Layer:** A single neuron with a linear activation function, suitable for the regression task.

The model was trained with the Adam optimizer and mean squared error (MSE) as the loss function. Early stopping was employed to halt training once the validation loss stopped improving, ensuring that the model did not overfit the training data.

**Performance Metrics:**

* **Validation Set:**
  + Root Mean Squared Error (RMSE): 24,963,213.86
  + Mean Absolute Error (MAE): 19,864,573.41
  + Mean Squared Error (MSE): 6.23e+14

The neural network's performance metrics were similar to those of the other models, indicating that while it could capture non-linear relationships, the improvement in predictive accuracy was marginal.

**Training Process:**

* **Loss Curve:** The loss curve showed steady convergence during training, with both the training and validation loss stabilizing after several epochs. The use of early stopping ensured that the model was not overfitting, as indicated by the minimal gap between training and validation loss.
* **MAE Curve:** The Mean Absolute Error (MAE) curve for both training and validation followed a similar pattern, confirming that the model's performance was consistent across different datasets.

**Key Insights:**

1. **Testing for Non-Linearity:** The neural network effectively tested for non-linear relationships between the features and YouTube views. However, the similar performance metrics compared to the linear models suggest that the relationships in this dataset may be predominantly linear, with limited non-linear interactions.
2. **Overfitting Prevention:** The use of dropout layers and early stopping proved crucial in preventing overfitting, especially given the complexity of neural networks. This helped maintain generalization performance, although the gains in accuracy were not substantial.

**Conclusion:** The Neural Network model provided a valuable test for non-linearity in the dataset. Despite its capability to model complex interactions, the similar performance to linear models suggests that non-linear relationships are not a major factor in predicting YouTube views for this dataset. This result indicates that simpler models may be sufficient, although the neural network remains a useful tool for validating assumptions about data structure and feature interactions.

| **Model** | **Mean Error (ME)** | **RMSE** | **MAE** | **MPE (%)** | **MAPE (%)** |
| --- | --- | --- | --- | --- | --- |
| **Full Regression** | 566,334.7585 | 24,623,324.0997 | 19,526,341.1532 | -274.3988 | 307.5128 |
| **Backward Regression** | 581,077.0578 | 24,625,953.6892 | 19,528,362.4278 | -273.9111 | 306.9864 |
| **Forward Regression** | 581,077.0578 | 24,625,953.6892 | 19,528,362.4278 | -273.9111 | 306.9864 |
| **Stepwise Regression** | 581,077.0578 | 24,625,953.6892 | 19,528,362.4278 | -273.9111 | 306.9864 |
| **Decision Tree** | 552,495.9803 | 24,652,330.9347 | 19,539,825.2468 | -273.7849 | 306.5290 |
| **Random Forest** | -750,759.0322 | 24,854,038.0771 | 19,935,836.9468 | -291.4426 | 323.1347 |
| **Neural Network** | - | 24,859,165.8217 | 19,825,625.4959 | - | - |

# Model Recommendation

## Model Selection

**Objective:** The objective of this section is to provide a comprehensive and technically sound explanation of the process and rationale behind selecting the most suitable model for predicting YouTube views based on song features. The selection process involved evaluating each model's predictive accuracy, interpretability, generalizability, and alignment with business objectives. The ultimate goal was to identify a model that not only provides reliable and actionable insights for stakeholders but also performs well on unseen data, ensuring that the predictions can be effectively applied to enhance music content and marketing strategies.

**Approach:** To achieve this objective, a range of modelling techniques were evaluated. These models varied from simple linear regression models to more complex non-linear models:

1. **Full Linear Regression:** This model served as a baseline, incorporating all predictors to understand their overall impact on YouTube views.
2. **Backward Elimination:** A stepwise approach that iteratively removes the least significant predictors based on the Akaike Information Criterion (AIC), simplifying the model while retaining the most relevant features.
3. **Forward Selection:** Opposite to Backward Elimination, this method starts with no predictors and sequentially adds the most significant ones, also guided by AIC.
4. **Stepwise Regression:** A hybrid approach combining both Forward Selection and Backward Elimination, allowing for both the addition and removal of predictors to optimize model performance.
5. **Decision Tree Regressor:** A non-linear model that captures complex interactions between predictors by recursively splitting the data into branches based on feature values.
6. **Random Forest Regressor:** An ensemble method that aggregates multiple decision trees to enhance robustness and reduce overfitting, capturing more nuanced patterns in the data.
7. **Neural Network:** A deep learning model designed to capture complex non-linear relationships, tested specifically to explore interactions that might be missed by simpler models.

**Evaluation Criteria:** The selection process was guided by several key criteria:

1. **Predictive Accuracy:**

* **Metrics Used:** The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were calculated for each model to assess its predictive accuracy. These metrics provide a quantitative measure of how closely the model’s predictions align with actual YouTube view counts.
* **Results:** The simpler models like Full Linear Regression provided a solid baseline, but the Backward Elimination and Stepwise Regression models achieved comparable accuracy with fewer predictors, indicating a more efficient use of data. For example, the Backward Elimination model produced an RMSE of 24,625,953.69 and an MAE of 19,528,362.43, nearly identical to the Full Regression model but with a reduced feature set.

1. **Interpretability:**

* **Importance:** Interpretability is crucial for this project because stakeholders need to understand how individual features (e.g., Loudness, Duration\_minutes, Danceability) influence YouTube views. Models that allow for easy interpretation of results were favoured, as this ensures that insights can be directly applied to decision-making.
* **Assessment:** Linear models, especially the Backward Elimination approach, stood out for their clear coefficient-based interpretation, allowing stakeholders to directly link changes in predictor variables to expected changes in YouTube views. The Backward Elimination model, with its simplified set of predictors, provides actionable insights while maintaining accuracy. In contrast, the Neural Network, although powerful in capturing non-linear relationships, lacks transparency, making it less suitable for stakeholders who require clear explanations.

1. **Generalization:**

* **Objective:** A good model should not only perform well on the training data but also generalize effectively to new, unseen data. Overfitting was a key concern, particularly with more complex models like Decision Trees and Neural Networks, which can be prone to capturing noise rather than meaningful patterns.
* **Results:** The Random Forest model, despite its ensemble nature, did not significantly outperform the simpler models (RMSE: 24,854,038.08; MAE: 19,935,836.95). This suggests that the added complexity was not necessary. In contrast, the Backward Elimination model effectively maintained accuracy while reducing overfitting by eliminating less significant predictors, demonstrating a strong balance between complexity and generalizability.

1. **Business Relevance:**

* **Focus:** The selected model needed to provide insights that are directly actionable for the business. For instance, identifying the key factors that influence YouTube views enables artists, producers, and record labels to optimize their strategies.
* **Outcome:** The Backward Elimination model aligned closely with business objectives by focusing on a refined set of significant and interpretable predictors. This approach ensures that stakeholders can easily apply the insights to enhance content performance on YouTube. The model's ability to highlight the most impactful features, such as Loudness, Licensed\_True, and official\_video\_True, makes it particularly valuable for strategic planning.

**Selection Outcome:** After a comprehensive evaluation, the Backward Elimination model was selected as the final model for this project. The decision was based on the following key reasons:

1. **Interpretability:** The Backward Elimination model provides clear, actionable insights through its streamlined set of predictors, making it easier for stakeholders to understand and apply the results.
2. **Comparable Performance:** The model’s accuracy metrics (RMSE: 24,625,953.69; MAE: 19,528,362.43; MAPE: 306.99%) were on par with more complex models, demonstrating that it does not sacrifice predictive power despite its streamlined approach.
3. **Simplicity and Practicality:** By eliminating less significant predictors, the Backward Elimination model avoids the risk of overfitting and focuses on the most relevant features, making it a practical choice for real-world application.
4. **Business Relevance:** The model directly aligns with the project’s business objectives by highlighting the key drivers of YouTube views, enabling stakeholders to optimize their strategies effectively.

**Conclusion:** In summary, the Backward Elimination model was selected because it offers a balanced approach that combines predictive accuracy, interpretability, and practicality. While other models provided insights into non-linear relationships and complex interactions, the Backward Elimination model emerged as the most effective solution for delivering actionable business insights that can be readily understood and implemented by stakeholders

## Model Theory

**Objective:** The Model Theory section provides a detailed overview of the theoretical foundation of the selected model, explaining how it works, its underlying assumptions, and why it was chosen as the most suitable model for this project. This section also explores how the model aligns with the data characteristics and the business objectives, offering a bridge between technical modelling and practical application.

**Model Overview:** The Backward Regression model, also known as Backward Elimination, is a type of linear regression that iteratively removes the least significant predictors from a full model. The goal is to refine the model by retaining only those variables that have a statistically significant relationship with the target variable, in this case, YouTube views. This method is guided by the Akaike Information Criterion (AIC), which balances model fit with model complexity, penalizing the inclusion of unnecessary predictors.

**Key Characteristics:**

1. **Linear Relationship:** The Backward Regression model assumes a linear relationship between the predictor variables and the target variable. This means that changes in the predictor variables result in proportional changes in the target variable. For example, an increase in Loudness or Danceability is expected to result in a predictable increase or decrease in YouTube views, depending on the sign and magnitude of the coefficients.
2. **Coefficient Interpretation:** The coefficients in the Backward Regression model provide clear, interpretable insights into the effect of each predictor on the target variable. Positive coefficients indicate that an increase in the predictor leads to an increase in YouTube views, while negative coefficients suggest the opposite. This interpretability is crucial for stakeholders who need to understand how specific features influence the outcome.
3. **Feature Selection:** Backward Elimination systematically removes predictors that do not significantly contribute to the model, as determined by their p-values. This process not only simplifies the model but also improves its generalizability by reducing the risk of overfitting. The AIC score is used to guide this selection process, ensuring that the model retains its predictive power while remaining as simple as possible.

**Mathematical Foundation:** The Backward Regression model is based on the following linear regression equation:

**Y = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ + ε**

Where:

* **Y** is the target variable (YouTube views).
* **β₀** is the intercept, representing the baseline value of **Y** when all predictors are zero.
* **β₁, β₂, ..., βₙ** are the coefficients for the predictors **X₁, X₂, ..., Xₙ**.
* **ε** is the error term, accounting for the variance in **Y** that is not explained by the predictors.

The Backward Elimination process works by initially fitting all predictors in the model and then removing the least significant ones step-by-step. At each step, the model re-estimates the coefficients for the remaining predictors and updates the AIC score. The process stops when removing additional predictors no longer improves the AIC score, resulting in a final model that balances fit and simplicity.

**Why This Model Was Chosen:** The Backward Regression model was selected for its ability to simplify the prediction process while retaining the most relevant features. This model is particularly well-suited for situations where interpretability is key, as it allows stakeholders to understand the impact of individual features on YouTube views. Moreover, the model’s reliance on AIC for feature selection ensures that it is both robust and generalizable, making it a practical choice for real-world application.

**Alignment with Data Characteristics:** The dataset used in this project contains a mix of numerical and categorical features, with some potential multicollinearity and non-linear relationships. The Backward Regression model effectively addresses these challenges by eliminating redundant or non-significant predictors, thereby reducing multicollinearity and improving model stability. Although the model assumes linear relationships, its use of AIC and systematic feature selection helps mitigate the impact of any residual non-linearity, resulting in a model that performs well despite these challenges.

**Alignment with Business Objectives:** The Backward Regression model’s ability to provide clear and actionable insights aligns closely with the business objectives of optimizing YouTube views. By identifying and focusing on the most significant predictors, the model offers practical recommendations for artists, producers, and record labels. For instance, understanding that Loudness, Duration\_minutes, and official\_video\_True are key drivers of views enables targeted strategies for content creation and promotion.

**Conclusion:** The Backward Regression model is a robust, interpretable, and practical tool for predicting YouTube views. Its balance of simplicity and predictive power, guided by the AIC-based feature selection process, makes it the ideal choice for this project. The model’s theoretical foundation and alignment with both the data characteristics and business objectives ensure that it provides valuable insights that can be easily understood and acted upon by stakeholders.

## Model Assumptions and Limitations

**Objective:** This section outlines the key assumptions underlying the Backward Regression model and discusses its limitations. Understanding these aspects is essential for correctly interpreting the model’s predictions and ensuring that it is applied appropriately in practice.

**Model Assumptions:**

1. **Linearity:**

* **Assumption:** The Backward Regression model assumes that there is a linear relationship between the predictor variables and the target variable (YouTube views). This means that changes in the predictors lead to proportional changes in the target variable.
* **Implication:** If the true relationship between the predictors and YouTube views is non-linear, the model may fail to capture this complexity, leading to suboptimal predictions.

1. **Independence of Errors:**

* **Assumption:** The model assumes that the residuals (errors) are independent of each other. In other words, the error term for one observation should not be correlated with the error term for another.
* **Implication:** Violations of this assumption, such as autocorrelation in time-series data, could lead to biased estimates and reduce the model’s reliability.

1. **Homoscedasticity:**

* **Assumption:** The model assumes that the variance of the residuals is constant across all levels of the predictors. This condition, known as homoscedasticity, ensures that the model’s predictions are equally reliable across the entire range of data.
* **Implication:** If there is heteroscedasticity (i.e., non-constant variance of residuals), the model’s estimates may become inefficient, and predictions for some ranges of data could be less accurate.

1. **No Perfect Multicollinearity:**

* **Assumption:** The model assumes that there is no perfect multicollinearity among the predictors. This means that no predictor is an exact linear combination of other predictors in the model.
* **Implication:** High multicollinearity can lead to unstable estimates of regression coefficients, making it difficult to determine the individual effect of each predictor. The Backward Elimination process helps mitigate this risk by removing redundant predictors.

1. **Normality of Errors:**

* **Assumption:** The model assumes that the residuals are normally distributed. This assumption is particularly important for hypothesis testing and constructing confidence intervals.
* **Implication:** If the errors are not normally distributed, the statistical inferences drawn from the model (e.g., p-values, confidence intervals) may not be valid. However, the model’s predictions may still be useful even if this assumption is violated.

**Model Limitations:**

1. **Sensitivity to Outliers:**

* **Limitation:** The Backward Regression model is sensitive to outliers, which can disproportionately influence the estimated coefficients and, consequently, the predictions. Outliers may result from data entry errors, unusual events, or inherent variability in the data.
* **Mitigation:** To address this limitation, it is essential to perform thorough data cleaning and consider the impact of outliers during the model validation process. Outlier detection and handling techniques should be employed as part of the preprocessing.

1. **Assumption of Linearity:**

* **Limitation:** The assumption of linearity can be restrictive in scenarios where the relationship between predictors and the target variable is inherently non-linear. The model may fail to capture complex interactions and non-linear patterns, leading to biased predictions.
* **Mitigation:** Non-linear transformations of variables or interaction terms can be considered to improve the model’s fit. However, if non-linearity is significant, alternative models like decision trees or neural networks might be more appropriate.

1. **Omitted Variable Bias:**

* **Limitation:** If important predictors are omitted from the model, the estimated coefficients of the included predictors may be biased. This issue, known as omitted variable bias, can lead to incorrect conclusions about the relationships between the predictors and the target variable.
* **Mitigation:** Careful feature selection and domain knowledge should be applied to ensure that all relevant predictors are considered in the model. The Backward Elimination process should be supplemented with exploratory data analysis to identify potential missing predictors.

1. **Interpretability versus Complexity:**

* **Limitation:** While the Backward Regression model is interpretable, it may not capture all the complexities of the data, especially if the relationships are non-linear or involve higher-order interactions. This trade-off between simplicity and complexity can limit the model’s predictive power.
* **Mitigation:** The simplicity of the model is one of its strengths, particularly for communicating results to stakeholders. However, for more complex datasets, combining the Backward Regression with more advanced techniques may be necessary to improve prediction accuracy.

1. **Generalizability:**

* **Limitation:** The model’s generalizability to new data is contingent upon the representativeness of the training data. If the training data does not adequately capture the range of scenarios seen in future observations, the model’s predictions may be less reliable.
* **Mitigation:** Cross-validation and careful partitioning of the dataset into training and validation sets can help assess and improve the model’s ability to generalize. Regular monitoring and updating of the model are also important as new data becomes available.

## Model Sensitivity to Key Drivers

**Objective:** This section provides a technical analysis of the Backward Regression model's sensitivity to all key predictors, based on the coefficients derived from the model. The sensitivity analysis reveals how changes in these predictors influence the model’s predictions of YouTube views, offering valuable insights for stakeholders aiming to optimize content performance.

**Key Drivers and Sensitivity Analysis:**

1. **Loudness:**
   * **Coefficient:** 69,262,870
   * **Sensitivity:** The Loudness coefficient indicates a significant positive impact on YouTube views, with a unit increase predicting approximately 69 million more views, holding all other variables constant.
   * **Technical Insight:** Higher loudness levels are associated with significantly increased views, suggesting that optimizing loudness within acceptable limits can boost audience engagement.
2. **Duration\_minutes:**
   * **Coefficient:** 2,629,623
   * **Sensitivity:** The positive coefficient for Duration\_minutes suggests that each additional minute of track duration is associated with an increase of approximately 2.63 million views.
   * **Technical Insight:** While longer tracks tend to perform better, finding an optimal duration is essential to maximize viewer retention and overall viewership.
3. **official\_video\_True:**
   * **Coefficient:** 6,075,166
   * **Sensitivity:** Having an official video is linked to an increase of approximately 6.07 million views, highlighting its critical role in driving viewership.
   * **Technical Insight:** Producing and promoting official videos can greatly enhance a track’s visibility and engagement on YouTube.
4. **Licensed\_True:**
   * **Coefficient:** 4,874,986
   * **Sensitivity:** Licensing significantly boosts viewership, with licensed tracks expected to garner approximately 4.87 million more views than unlicensed ones.
   * **Technical Insight:** Proper licensing not only ensures compliance but also improves promotional opportunities, leading to higher visibility and viewership.
5. **Danceability:**
   * **Coefficient:** 9,010,455
   * **Sensitivity:** The model predicts an increase of approximately 9 million views for each unit increase in Danceability, underscoring its importance as a key driver.
   * **Technical Insight:** Tracks that are more danceable resonate better with audiences, particularly in genres where rhythm and engagement are crucial.
6. **Energy:**
   * **Coefficient:** -8,625,582
   * **Sensitivity:** Interestingly, Energy has a negative coefficient, indicating a decrease of approximately 8.63 million views per unit increase. This suggests that tracks with lower energy levels may be more popular with certain audiences.
   * **Technical Insight:** Understanding the target audience’s energy preferences is critical, as lower energy tracks may perform better in specific contexts.
7. **Tempo:**
   * **Coefficient:** 16,609.46
   * **Sensitivity:** While the coefficient for Tempo is positive, its effect on viewership is smaller compared to other predictors. The model predicts a slight increase in views with faster tempos.
   * **Technical Insight:** Tempo influences viewership but is less critical than factors like loudness and duration. Producers should consider tempo adjustments as part of a broader strategy.
8. **flag\_outlier\_Loudness:**
   * **Coefficient:** 3,984,852
   * **Sensitivity:** The flag indicating outlier loudness contributes positively to the model, suggesting that tracks with unusually high or low loudness can still perform well under certain conditions.
   * **Technical Insight:** Managing outliers in loudness can help avoid extreme deviations while still leveraging loudness as a key driver.
9. **Album\_type\_compilation:**
   * **Coefficient:** 2,694,618
   * **Sensitivity:** Compilation albums have a positive coefficient, indicating they attract approximately 2.69 million more views, on average, compared to other album types.
   * **Technical Insight:** Compilation albums should be considered as a strategy for increasing visibility, particularly when featuring popular tracks.
10. **Album\_type\_single:**
    * **Coefficient:** -3,367,937
    * **Sensitivity:** Singles have a negative coefficient, predicting approximately 3.37 million fewer views compared to other album types. This suggests that while singles can be effective, they may not perform as well as full albums or compilations.
    * **Technical Insight:** Stakeholders may need to focus more on albums or compilations, or strategically promote singles to counterbalance this trend.
11. **Key\_8.0:**
    * **Coefficient:** 2,247,446
    * **Sensitivity:** The specific key (Key 8.0) positively affects viewership, predicting an increase of approximately 2.25 million views when tracks are in this key.
    * **Technical Insight:** Musical key selection can influence viewership, and choosing the right key could be an additional lever for maximizing engagement.

**Key Insights:**

1. **Impactful Predictors:** Loudness, Duration\_minutes, official\_video\_True, and other key predictors like Danceability and Licensed\_True are crucial drivers of viewership, each contributing significantly to the model’s predictions.
2. **Balanced Influence:** The model captures both positive and negative influences on YouTube views, indicating the importance of a balanced approach when optimizing track features.
3. **Actionable Guidance:** The comprehensive sensitivity analysis provides stakeholders with clear, quantifiable actions to focus on, from loudness and licensing to the strategic use of album types and musical keys.

**Conclusion:** The sensitivity analysis of the Backward Regression model confirms that a range of predictors, including Loudness, Duration\_minutes, official\_video\_True, Danceability, and Licensed\_True, play a significant role in driving YouTube views. By strategically adjusting these key drivers, stakeholders can substantially enhance the performance and reach of their content on YouTube, leading to more targeted and effective content strategies.

# Conclusion and Recommendations

## Impacts on Business Problem (Scope of the recommended model)

**Conclusion:** The Backward Regression model was selected as the optimal solution for predicting YouTube views based on its combination of predictive accuracy, interpretability, and alignment with business objectives. The model effectively identifies and quantifies the key drivers of YouTube viewership, including Loudness, Duration\_minutes, official\_video\_True, Danceability, Licensed\_True, and other critical features. By focusing on these predictors, stakeholders can gain actionable insights that directly impact the reach and engagement of their content on YouTube.

The model’s performance metrics, such as RMSE, MAE, and MAPE, were comparable to more complex models, confirming its robustness while maintaining a streamlined feature set. Additionally, the model’s coefficient-based interpretation ensures that the insights are both clear and practical, enabling stakeholders to make data-driven decisions that enhance the performance of their music content on YouTube.

**Impacts on the Business Problem (Scope of the Recommended Model):** The Backward Regression model addresses the core business problem by providing a focused set of key predictors that significantly influence YouTube views. The model’s insights empower artists, producers, and record labels to strategically optimize content creation and marketing efforts, ensuring that resources are allocated efficiently to maximize viewership. The scope of the model is broad enough to be applicable across various genres and content types, yet specific enough to offer targeted recommendations that align with business goals.

## Recommended Next Steps

**Recommended Next Steps:** To fully capitalize on the insights provided by the Backward Regression model, the following steps are recommended:

1. **Implementation of Model Insights:**
   * **Content Strategy:** Use the model’s findings to guide content creation, with a particular focus on optimizing Loudness, Duration\_minutes, and Danceability. Additionally, ensure that tracks are produced in keys that resonate with the target audience, and consider the strategic use of album types.
   * **Marketing and Promotion:** Tailor marketing campaigns to emphasize licensed content and the production of official videos, both of which have been shown to significantly boost viewership. Promotional efforts should highlight these features to enhance engagement.
2. **Model Monitoring and Updates:**

**Continuous Monitoring:** Regularly assess the model’s performance against new data to ensure ongoing accuracy and relevance. As audience preferences evolve, the model should be recalibrated to maintain its predictive power.

**Data Refresh:** Periodically update the model with fresh data to capture the latest trends, ensuring that predictions remain aligned with current market dynamics.

1. **Exploration of Additional Models:**
   * **Testing Non-linear Relationships:** Consider exploring non-linear models, such as Gradient Boosting Machines (GBM) or advanced neural networks, to capture complex interactions between features that may be missed by linear models.
   * **Cross-validation with New Data:** Validate the model's predictions using different datasets or subsets to assess its consistency and robustness across various scenarios.
2. **Enhanced Feature Engineering:**
   * **Feature Interaction Analysis:** Investigate potential interactions between key features, such as the interplay between Loudness and Tempo, to refine the model further and uncover deeper insights.
   * **Incorporation of Additional Data Sources:** Integrate additional data points, such as social media engagement or audience demographics, to enrich the model and improve its predictive accuracy.