INNOVISION HACKATHON

TEAM NAME: sachdevarishav449

TEAM MEMBERS:

- RISHAV SACHDEVA (sachdevarishav449@gmail.com)
- HANSAWANI SAINI (hansawani07@gmail.com)

Source Code:

https://github.com/Rishav9911/Innovision-Hackathon

Problem Statement

On the basis of given Music Extravaganza dataset :

I) Divide the songs into various categories (pop, rock, country etc.)

II) What are the factors that contribute the most to the views of the song.

Data Pre-Processing

Handling Missing Values

Numerical features with missing values were replaced with their mean for the sake of consistency, while "Unknown" was assigned to categorical features to prevent data loss and allow for smooth processing.

Removing Duplicate Rows

Identifying and removing any duplicate rows, if they existed, was done to eradicate data redundancy from the dataset so that the repeated observations would not affect any analysis or visualization.

Outlier Detection and Removal

The IQR method compared entries and identified those that seemed outliers. The classified entries, which were outliers for at least three features were suppressed for the benefit of data fortification.

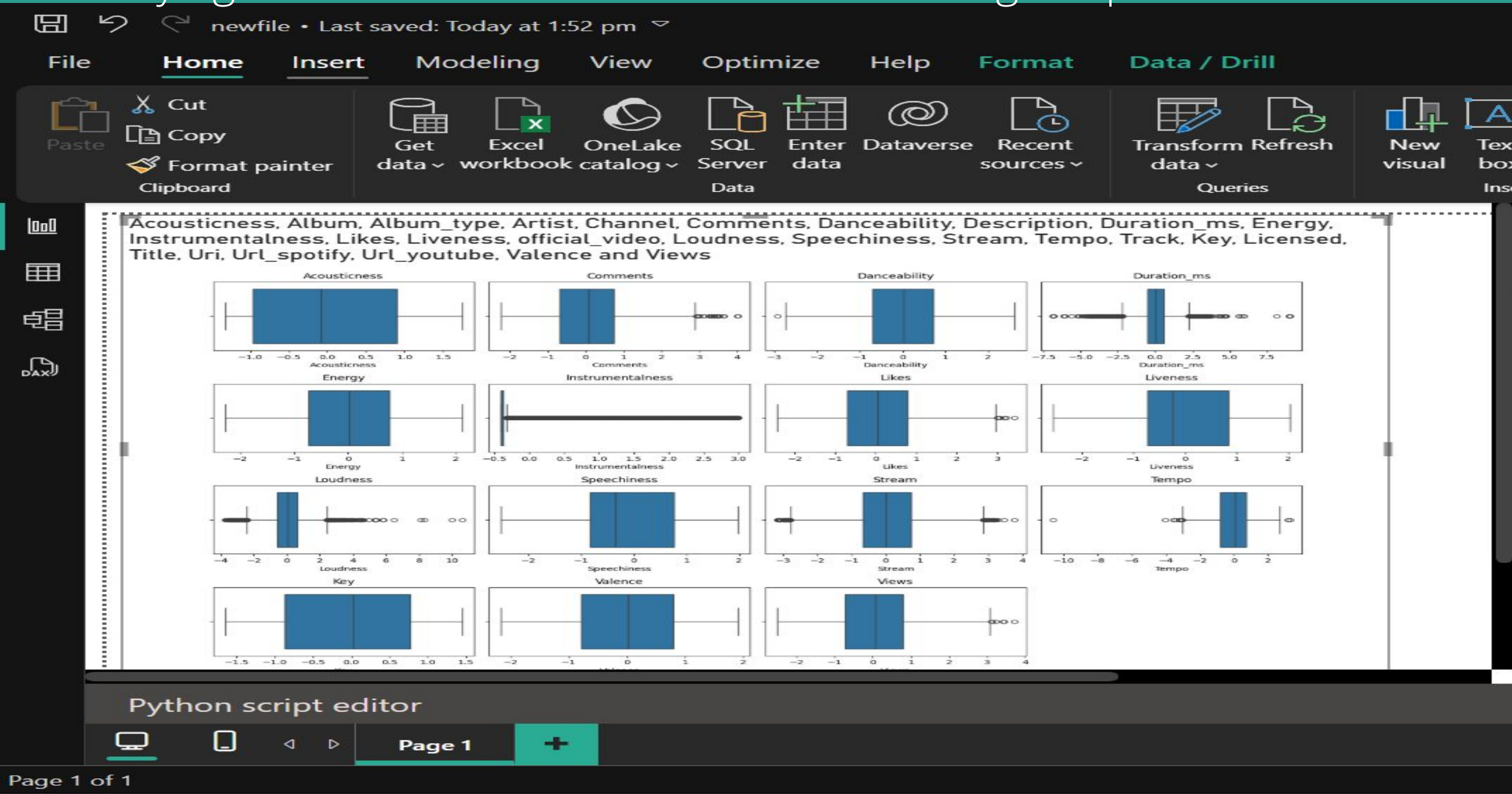
Boxplot Creation for Outliers

Boxplots were generated for numerical features to **visualize and confirm the presence of outliers** before making any removals, ensuring that only extreme values were considered.

Data Pre-Processing

Skewness Correction	Correlation & Redundancy Check	Categorical Feature Processing	Moving on to next step
Skewness may impact model performance; therefore, positively skewed numerical features underwent log and power transformations to improve the normality of their distributions.	A heatmap was used to identify highly correlated features, and columns with **correlation > 0.85** were checked for redundancy. No identical features were found.	On the other hand, this procedure was a safeguard against computation problems. The categorical features were ensured to be in string format, therefore avoiding dtype=object errors in the GPU-based processing.	With this we are done with the data preprocessing and can move on to the other two problem statements

Verifying outliers in the numerical features using Boxplots in **PowerBI**



I) Divide the songs into various categories (pop, rock, country etc.)

Classification of songs according to genre is performed through **Gaussian Mixture Models** (GMMs) in order to identify natural clusters of audio features. This unsupervised technique enables the model to adjust itself better into the music's complex patterns. These clusters will be made from l**oudness, energy, speechiness, and such like patterning characteristics**. Then, a **rule-based mapping** would assign music genres based on the mean of different features across clusters.

Step 1: Feature Standardization

Actions Taken:

1. **Data Cleaning & Handling Missing Values**:
   - Missing values in numerical features could distort clustering results.
   - Used mean imputation to replace missing values while maintaining feature distributions.
2. **Removing Duplicates**:
   - Eliminated duplicate records to prevent bias in the clustering process.
3. **Feature Selection**:
   - Chose key audio features that define musical characteristics, ensuring meaningful clustering:
     - Danceability (suitability for dancing)
     - Energy (perceived intensity and activity)
     - Key (musical key of the track)
     - Loudness (overall loudness in dB)
     - Speechiness (presence of spoken words)

Step 1: Data Cleaning & Feature Standardization

- Acousticness (confidence that the track is acoustic)
- Instrumentalness (whether the song is instrumental)
- Liveness (presence of a live audience)
- Valence (positivity of the song)
- Tempo (BPM - beats per minute)
- Duration (length of the track in milliseconds)

4. **Feature Standardization:**

- Used StandardScaler to transform all features to have zero mean and unit variance.
- Prevented features with larger numerical ranges (e.g., Loudness in dB) from dominating others (e.g., Valence).

Intermediate Conclusion:
The dataset is now clean, complete, and uniformly scaled, ensuring fair influence of all features during clustering.

Step 2: Optimal Number of Clusters for GMM

Actions Taken:

- Implemented **Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC)**:
  - Both are statistical metrics used to evaluate model complexity.
  - Lower BIC/AIC scores indicate a better balance between model fit and complexity.
- Tested a range of clusters (2 to 14).
- Selected the optimal number of clusters based on the lowest BIC value.

Intermediate Conclusion:

The number of clusters is chosen based on BIC minimization, ensuring that the model is neither underfitting nor overfitting. The **optimal number of clusters is 14.**

Step 3: Applying GMM for Clustering

Actions Taken:

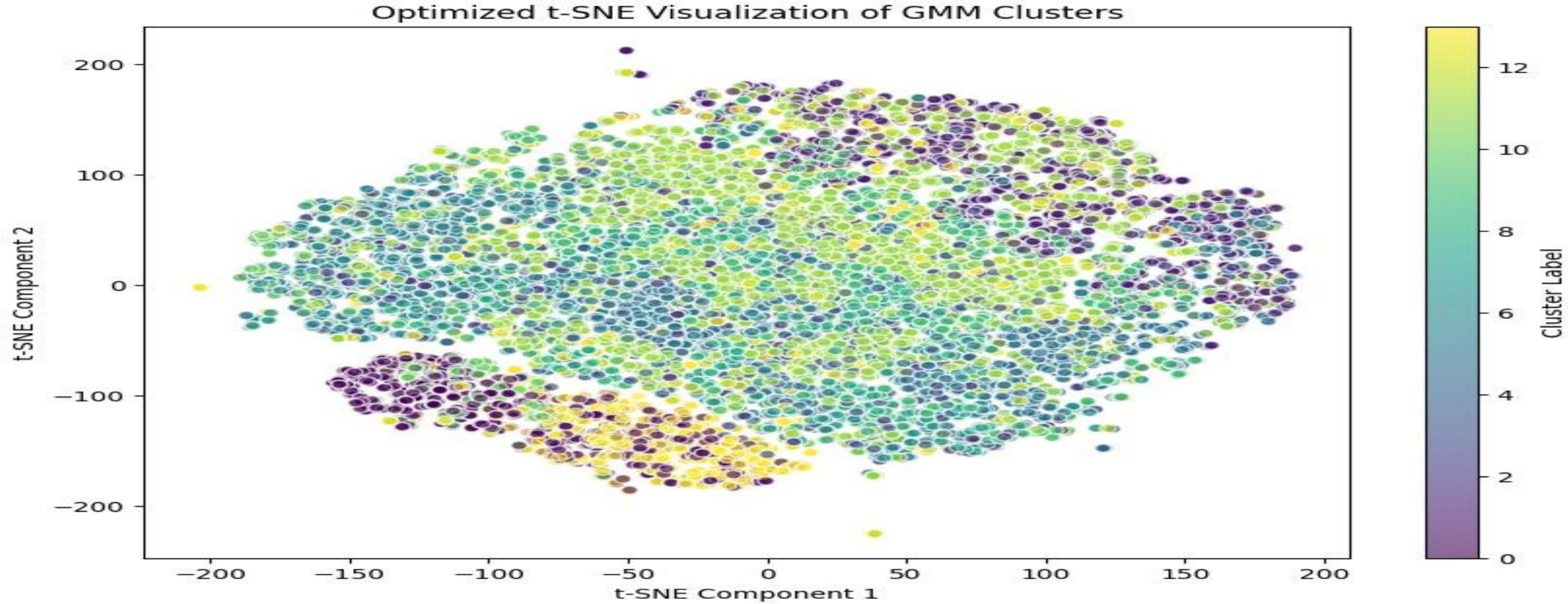
- Fit a **Gaussian Mixture Model (GMM)** using the optimal cluster count.
- Generated probability distributions for each song belonging to different clusters.
- Assigned **each song to the cluster with the highest probability** out of all the cluster probabilities for that song.

Intermediate Conclusion:

Each song is now assigned to a cluster with a probabilistic membership, allowing for **soft clustering meaning a song can belong to multiple genres** instead of using k-means that will assign the song to just a single genre

Visualisation using t-SNE:

t-SNE is used to visualize high-dimensional audio data, preserving local relationships. Key features are selected, and PCA reduces dimensions to 7 for faster processing. Optimized t-SNE runs with perplexity = 10, 3000 iterations, and learning rate = 200, using PCA initialization for better clustering. It reveals song groupings based on GMM, highlighting musical patterns.



Observation	Explanation
Distinct Cluster Formation (Bottom Left)	Purple & yellow: unique song group, possibly with distinct tempo, energy, or danceability.
Densely Mixed Region (Center & Right)	Green, blue, purple: many songs share similar audio characteristics, showing soft boundaries between clusters.
Few Outliers (Edges & Top)	Songs with extreme feature values (e.g., very high loudness, tempo, or instrumentalness), which may require separate analysis.

Step 4: Cluster Summary Analysis

Actions Taken:

1. Calculated Average Feature Values for Each Cluster:
   - Computed **mean values of key features (e.g., Loudness, Energy, Danceability) for each cluster.**
   - This helps in explain the nature of each cluster, identifying whether they contain energetic, acoustic, instrumental, or dance-heavy songs.

Intermediate **Conclusion:**

Each cluster now has their distinct summary of statistical characteristics, laying the foundation for **cluster to genre mapping** based on these average feature values.

Cluster	Danceability	Energy	Key	Loudness	Speechiness
0	-0.818036	-0.717622	0.048396	-1.220520	-0.428024
1	0.528699	-0.001718	0.087939	0.050209	1.809098
2	-0.261988	-0.444286	0.078459	-0.426755	-0.907664
3	0.081451	0.715835	0.086108	0.463608	0.165675
4	0.019902	0.018784	-0.034136	-0.194779	-0.707228
5	0.051668	0.420475	0.359293	0.425042	-0.171416
6	-0.657933	-1.102263	-0.086052	-0.552149	-0.877075
7	-0.833687	1.398370	-0.029974	0.858652	0.046354
8	-0.037535	0.118949	-1.081438	0.272684	-0.782452
9	0.345628	0.198883	-1.365000	0.351131	0.422263
10	-0.874775	-1.549546	-0.140811	-1.410193	-0.836256
11	0.200410	0.108561	0.654126	0.259972	-0.102088
12	0.457457	0.113491	0.066928	0.031331	0.655307
13	0.017354	0.268231	-0.028079	-0.184389	-0.021129

Cluster	Acousticness	Instrumentalness	Liveness	Valence	Tempo
0	0.525330	2.994580	-0.255148	-0.825313	-0.211275
1	-0.060479	-0.404866	0.143464	0.131040	-0.066424
2	0.252314	-0.392090	-0.145585	-0.015638	-0.082890
3	-1.097182	-0.028356	0.039924	0.045535	0.215289
4	-0.060560	-0.306145	-0.152846	0.056643	-0.053081
5	-0.137843	-0.404803	1.403172	0.392473	0.060939
6	1.318134	-0.404635	-0.063474	-0.565591	-0.139153
7	-1.309661	-0.395073	0.247603	-0.188803	0.522336
8	-0.258374	-0.404605	-0.471019	0.104964	-0.069828
9	-0.216385	-0.404767	0.134001	0.121203	0.133793
10	1.532047	1.011281	-0.250370	-0.887017	-0.331613
11	-0.114419	-0.404764	-0.541147	0.185209	0.003357
12	-0.046633	-0.391237	-0.009175	0.179014	0.050074
13	-0.248941	1.464875	-0.055723	-0.007420	0.072505

Step 5: Cluster-to-Genre Mapping

Overview of the Mapping Approach:

- The genre classification was based on distinct musical characteristics extracted from audio features.
- Each feature provides key information about the song's style, instrumentation, and overall feel.
- We developed **threshold-based rules to categorize songs into different genres based on patterns observed in their clusters.**

Feature	Condition	Interpretation	Mapped Genre(s)
Loudness & Energy	High (> 0.3 & > 0.6)	Powerful, high-energy music	Rock, Heavy Metal, Punk Rock, EDM
	Low (< -0.5)	Soft, calm, orchestral sound	Classical, Ambient, Acoustic

Speechiness	High (> 0.4)	Spoken-word content (e.g., rap vocals)	Hip-Hop, Rap, R&B
	Low (< 0.4)	More melodic or instrumental tracks	Rock, Pop, Dance
Acousticness & Instrumentalness	High Acousticness (> 0.6) & Low Loudness (< -0.5)	High acoustic presence, soft sound	Classical, Orchestral, Ambient
	High Instrumentalness (> 1.0) & High Acousticness (> 0.5)	Mostly instrumental tracks	Jazz, Blues, Instrumental
Danceability & Tempo	High Danceability (> 0.5) & High Tempo (> 0.4)	Upbeat, rhythmic, dance-oriented tracks	EDM, House, Techno, Dance Pop
	Low Danceability	Less structured rhythm for dancing	Indie, Alternative, Rock
Liveness	High Liveness (> 0.3) & High Energy (> 0.4)	Indicates a live performance or concert recording	Live, Alternative, Jam Band

Result

Examples after classifying songs into genres:

Track Name	Genre assigned by our approach	True genre as per the Web
She's My Collar (feat. Kali Uchis)	Alternative / Indie / Soft Rock	Alternative / Indie
By the Way	Rock / Alternative Rock / Classic Rock	Alternative/ Indie/ Rock
One Step Closer	Hard Rock / Heavy Metal / Punk Rock	Metal
Karma Police	Alternative / Indie / Soft Rock	Alternative / Indie
Pehla Nasha	Classical / Orchestral / Ambient	Old Indian/ Instrumental

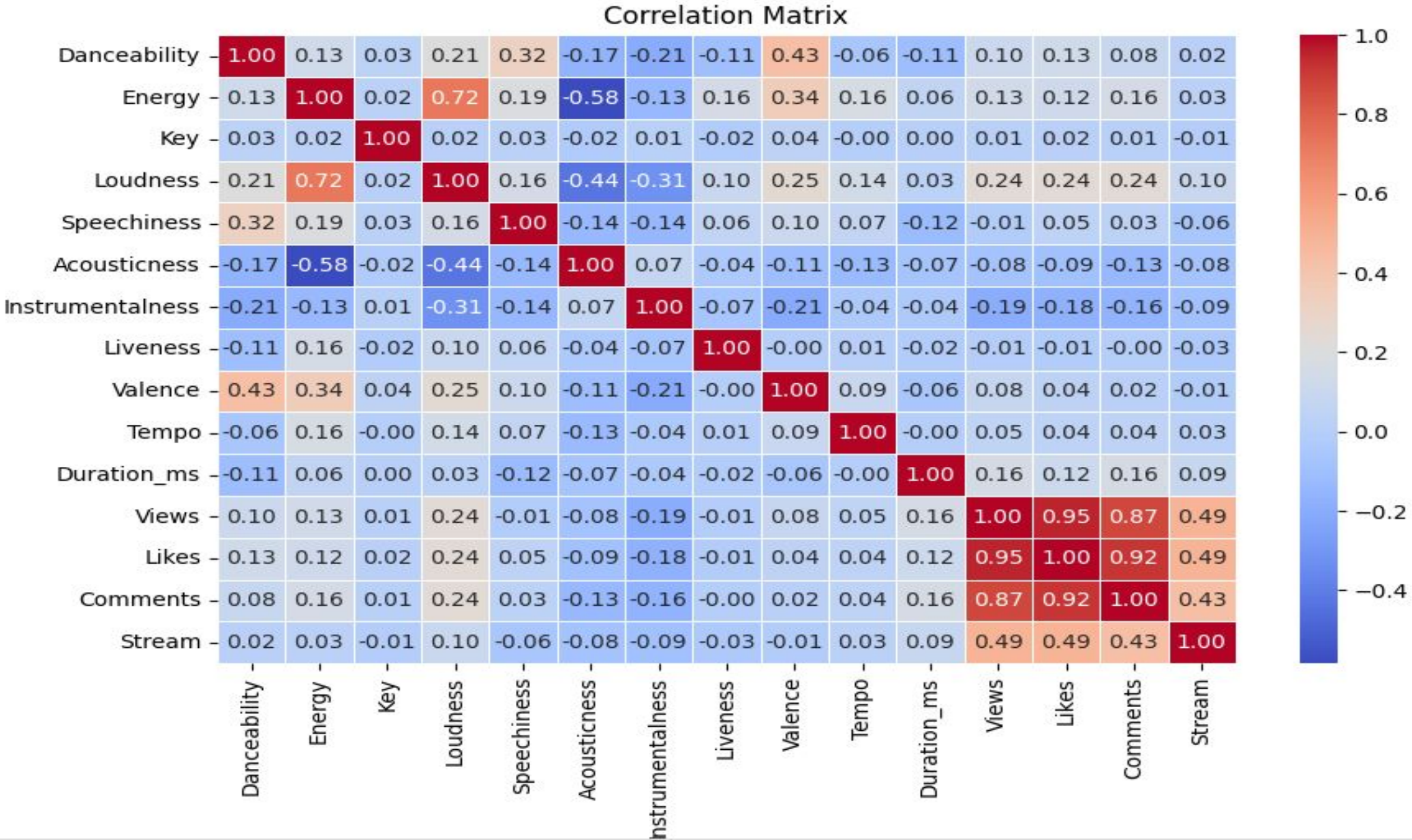
II) What are the factors that contribute the most to the views of the song?

We will conduct **correlation analysis** for **numerical features** and their **linear relationship with Views.**
To ascertain feature importance through potential linear and **non-linear** dependencies, we will apply **Random Forest**. Hence, a **combination of both methods** will be used for optimization.
For **categorical features**, we will use **aggregation method** to analyse impact of them on Views. Coupled with that, **boxplots** will visualize the variations across the different categories in a broad analysis.

Correlation Analysis

- Filtered Numerical Features: Only **numerical columns** were considered to ensure correct correlation calculations.
- Computed **Correlation Matrix**: **Because** it will measure the strength of relationships between the features.
- Identified Key Correlations: Extracted the top features most correlated with Views. The top features were likes, comments, and streams with all of them having correlation probability of 0.5 or more.
- Visualized with a **Heatmap**: Displayed correlation strengths to spot highly related features.
- Set Selection Threshold: Features with **correlation > 0.2 were shortlisted for Random Forest model training.**

Purpose: Identify important numerical factors influencing Views for better model performance.


Correlation Matrix
Danceability 1.00 0.13 0.03 0.21 0.32 -0.17 -0.21 -0.11 0.43 -0.06 -0.11 0.10 0.13 0.08 0.02
Energy 0.13 1.00 0.02 0.72 0.19 -0.58 -0.13 0.16 0.34 0.16 0.06 0.13 0.12 0.16 0.03
Key 0.03 0.02 1.00 0.02 0.03 -0.02 0.01 -0.02 0.04 -0.00 0.00 0.01 0.02 0.01 -0.01
Loudness 0.21 0.72 0.02 1.00 0.16 -0.44 -0.31 0.10 0.25 0.14 0.03 0.24 0.24 0.24 0.10
Speechiness 0.32 0.19 0.03 0.16 1.00 -0.14 -0.14 0.06 0.10 0.07 -0.12 -0.01 0.05 0.03 -0.06
Acousticness -0.17 -0.58 -0.02 -0.44 -0.14 1.00 0.07 -0.04 -0.11 -0.13 -0.07 -0.08 -0.09 -0.13 -0.08
Instrumentalness -0.21 -0.13 0.01 -0.31 -0.14 0.07 1.00 -0.07 -0.21 -0.04 -0.04 -0.19 -0.18 -0.16 -0.09
Liveness -0.11 0.16 -0.02 0.10 0.06 -0.04 -0.07 1.00 -0.00 0.01 -0.02 -0.01 -0.01 -0.00 -0.03
Valence 0.43 0.34 0.04 0.25 0.10 -0.11 -0.21 -0.00 1.00 0.09 -0.06 0.08 0.04 0.02 -0.01
Tempo -0.06 0.16 -0.00 0.14 0.07 -0.13 -0.04 0.01 0.09 1.00 -0.00 0.05 0.04 0.04 0.03
Duration_ms -0.11 0.06 0.00 0.03 -0.12 -0.07 -0.04 -0.02 -0.06 -0.00 1.00 0.16 0.12 0.16 0.09
Views 0.10 0.13 0.01 0.24 -0.01 -0.08 -0.19 -0.01 0.08 0.05 0.16 1.00 0.95 0.87 0.49
Likes 0.13 0.12 0.02 0.24 0.05 -0.09 -0.18 -0.01 0.04 0.04 0.12 0.95 1.00 0.92 0.49
Comments 0.08 0.16 0.01 0.24 0.03 -0.13 -0.16 -0.00 0.02 0.04 0.16 0.87 0.92 1.00 0.43
Stream 0.02 0.03 -0.01 0.10 -0.06 -0.08 -0.09 -0.03 -0.01 0.03 0.09 0.49 0.49 0.43 1.00
Danceability
Energy
Key
Loudness
Speechiness
Acousticness
Instrumentalness
Liveness
Valence
Tempo
Duration_ms
Views
Likes
Comments
Stream
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Random Forest Model

Why Random Forest?

1. **Robustness to Overfitting** – Unlike single decision trees, Random Forest uses multiple trees (bagging) to reduce variance and improve generalization.

2. **Captures Non-Linear Relationships** – Unlike correlation or linear regression, Random Forest detects complex feature interactions and non-linear dependencies.

How Were Features Selected?

To prevent **overfitting** and improve **model efficiency**, only the most relevant features were chosen:

- **Features with correlation > 0.2** (ensuring strong relationship with views).
- **Removed redundant/weak predictors** (e.g., speechiness, liveness had negligible correlation).
- **Balanced dataset with 80% training & 20% testing** to evaluate generalization.

Key Observation from Feature Importance

- Likes have the highest importance, while Comments and Streams have very low importance.
- This is because Comments and Streams are highly correlated with Likes.
- Random Forest prioritizes one feature if other features are highly correlated with the given feature and thus reducing their importance scores.

Why Do Comments & Streams Have Low Importance?

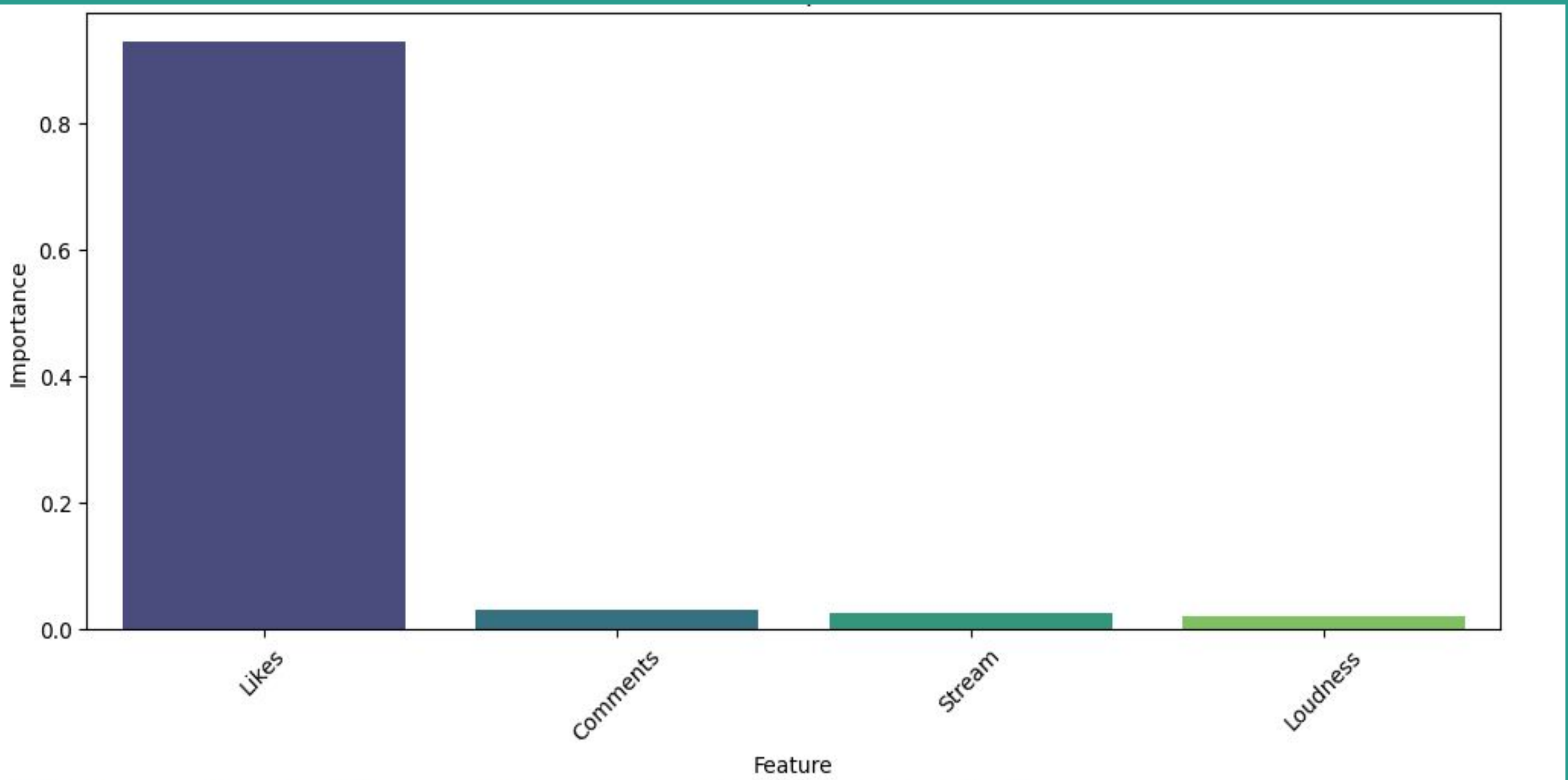
❖ **Feature Correlation Effect:**

- **Comments and Streams strongly correlate with Likes** (likely >0.8).
- Since Likes already explain most of the variance, the model does not need to rely on Comments and Streams.

Model Performance

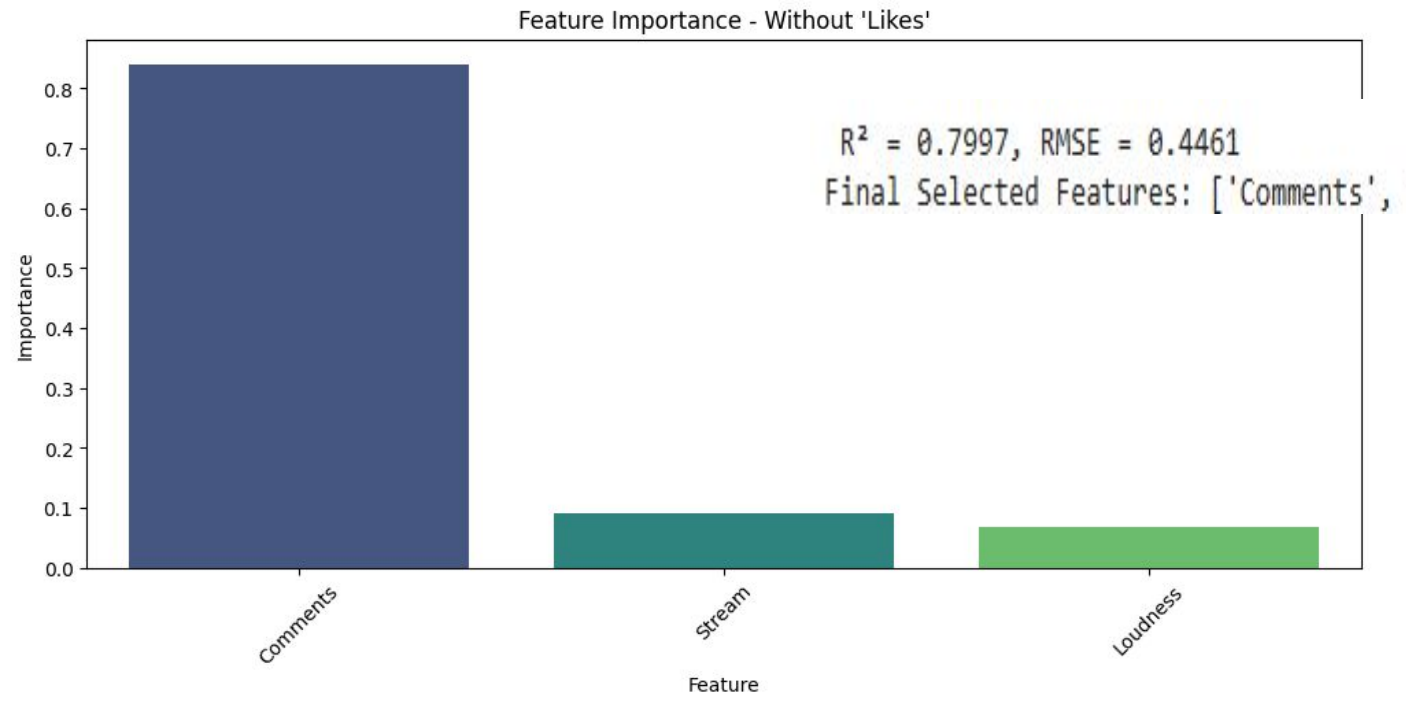
- **R^2 = 0.9230**
- **RMSE = 0.2765**
  (low error)

Feature Importance Chart




```
Final Selected Features: ['Loudness', 'Comments', 'Stream', 'Instrumentalness', 'Likes']
```

- **If we remove the likes column** from the Random Forest model training, then Random Forest would give more **importance to the next best feature, i.e. comments.**
- Though after **removing the likes column, RMSE increases** showing that the **Likes feature impact Views feature** more than Comments feature, meaning we can't remove the Likes feature.



R^2 = 0.7997, RMSE = 0.4461
Final Selected Features: ['Comments', 'Loudness', 'Stream', 'Instrumentalness']

Aggregation & Boxplots

Analysis Approach

To understand how categorical features influence the number of Views, we aggregated the Views for each category and calculated their range (max - min). A **larger range** indicates that different categories within that **feature significantly affect Views**, while a **smaller range** suggests little to no impact.

Findings & Insights

1. Strong Impact on Views (Large Variation in Range > 4)
   - Artist (4.74), Album (5.89), Channel (5.28), Track (6.22), and Title (6.22)
   - These features show **large differences in Views across their subcategories**, meaning they significantly impact how well a song performs.

2. Moderate Impact on Views (Range ~1.5 - 1.7)
   - Licensed (1.58) & Official Video (1.73)
   - Songs that are officially licensed or have an official video tend to perform slightly better, but the impact is not as strong as other factors like Artist or Album.

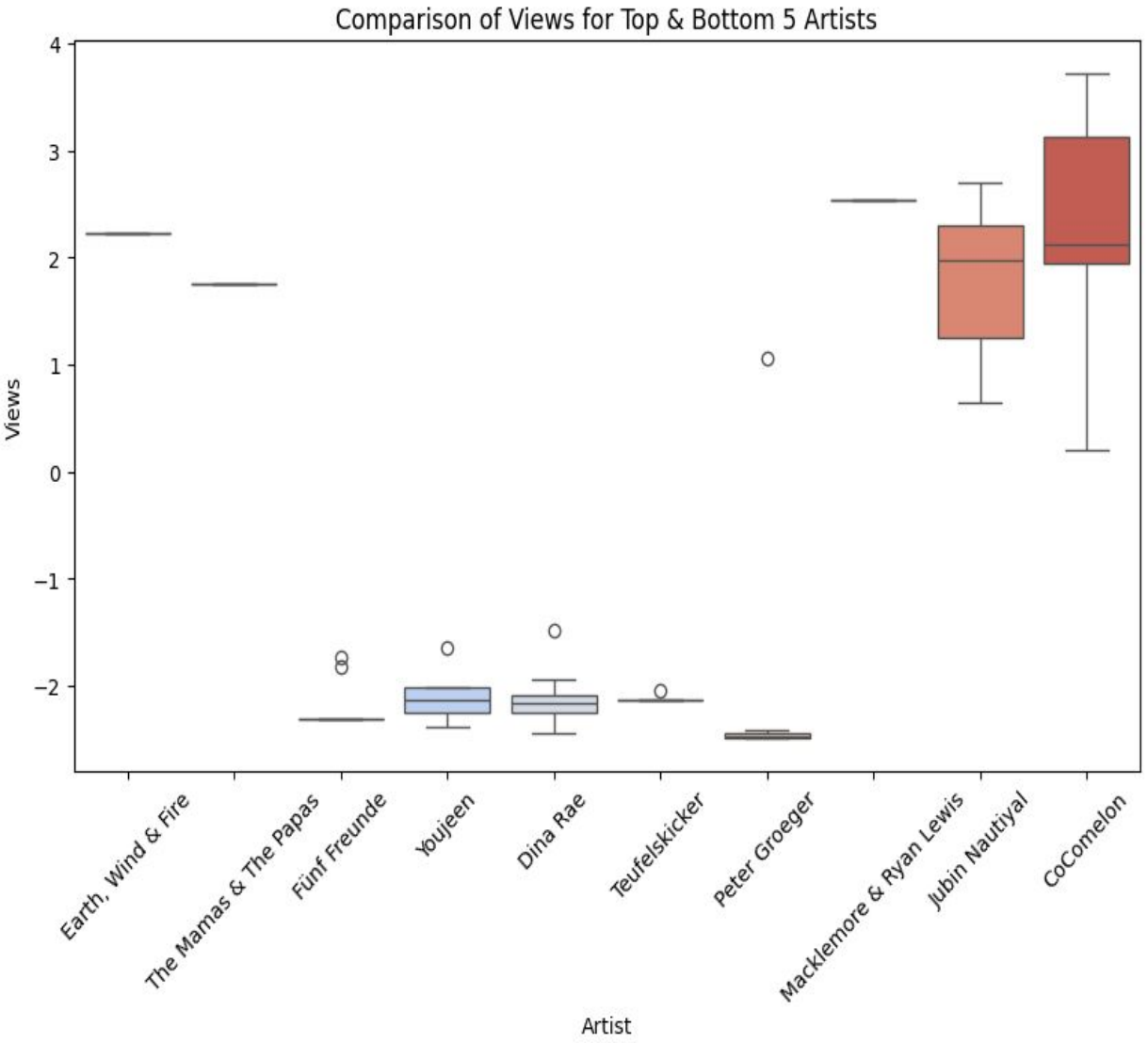
3. Minimal Impact on Views (Range ≈ 0.1)

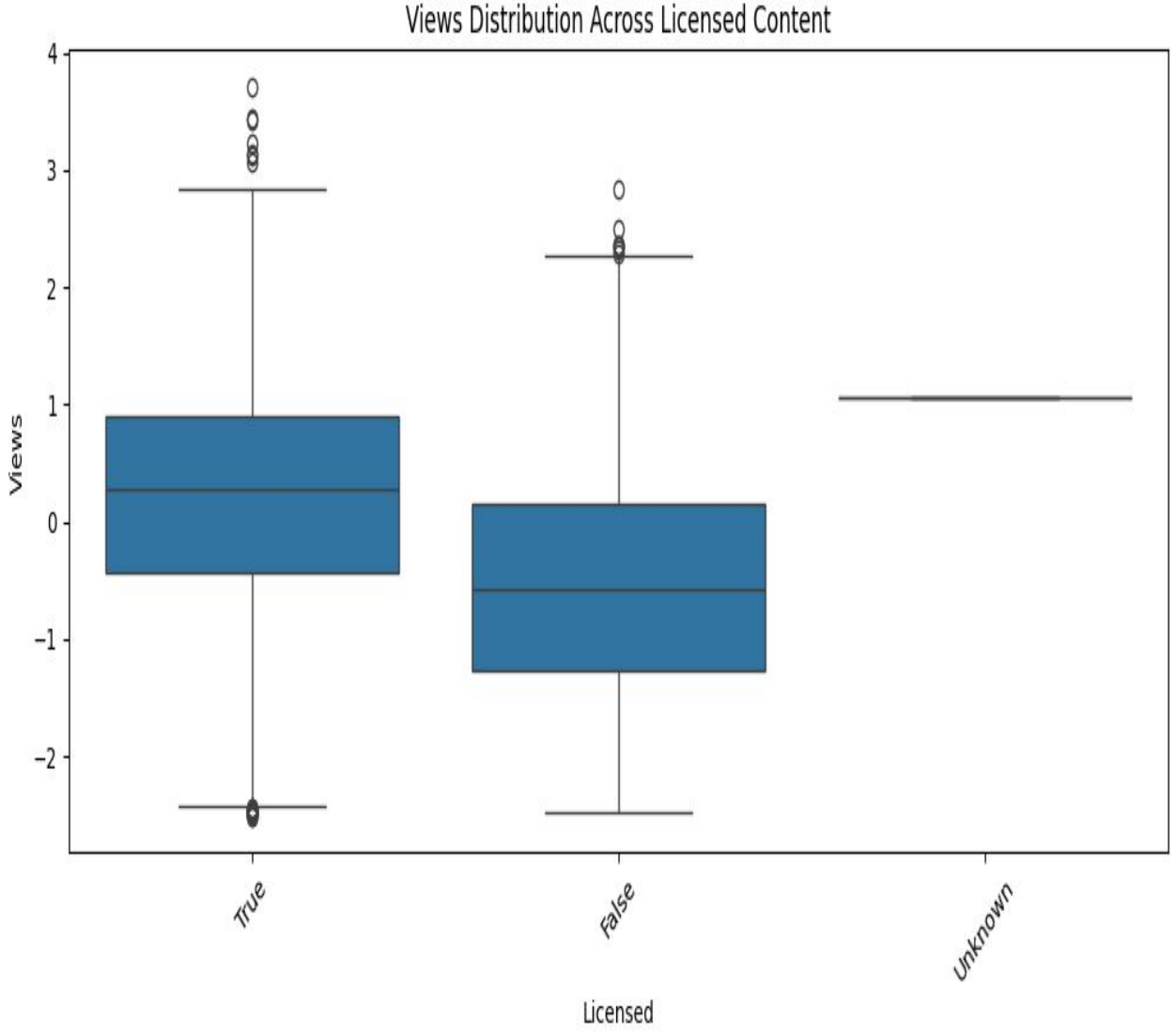
- Album Type (0.13)
- Whether a song is part of an album or a single **barely affects Views**, as the variation in Views is extremely low.

Boxplots for visualization

- We used **Power BI boxplots** to visualize how **Views vary across different categorical features** like Artist, Album, and Channel. For features with **too many unique values**, we selected only the **top 5 and bottom 5 categories** for better clarity.
- The visualizations **clearly showed** that **Artist, Album, and Channel strongly impact Views**, while features like **Album Type and Licensed** had little to no effect.

Boxplots for the categorical features





Result

1. Numerical Features Impact on Views
   - **Likes & Comments have the strongest influence on Views.**
   - Streams have some impact, but other numerical features contribute much less.

2. Categorical Features Impact on Views
   - **Artist, Album, Track, Title, and Channel play the most crucial role in determining Views.**
   - Licensing and Official Videos have some influence but are not dominant factors.
   - Album Type has an almost negligible impact on Views.

Thank you

Source Code:

https://github.com/Rishav9911/Innovision-Hackathon