Visual Analytics Workflow

Dyuthi Vivek
IMT2022523
IIIT Bangalore
Bangalore, India
Dyuthi.Vivek@iiitb.ac.in

Saniya Ismail Kondkar
IMT2022128
IIIT Bangalore
Bangalore, India
Saniya.Ismail@iiitb.ac.in

Ragini Metlapalli
IMT2022029
IIIT Bangalore
Bangalore, India
Metlapalli.Ragini@iiitb.ac.in

I. Introduction

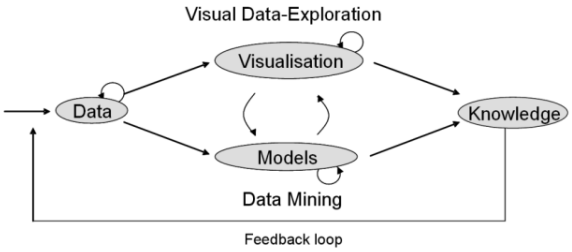

Fig 0: Kiem et al. Visual Analytics Workflow, Image courtesy: [8]

The given dataset for Assignment-1 [3] consists of 30000 songs on Spotify from the year 1957 to 2020. Our tasks for Assignment-1 involved the following:

- Task 1: Exploring the interplay of song features and music trends
- Task 2: Popularity trends over time and genre
- Task 3: Artist popularity over time and across genres

We have developed three visual analytics workflows building on our tasks in Assignment-1.

- Task 1: This goal of this task is identifying the features that affect song popularity and building a model to predict the popularity of the song.
- Task 2: The goal of this task was to extract insights into genre trends, identify patterns, and forecast genre popularity over time.
- Task 3: The goal of this task is to understand and evaluate artist popularity using artist collaboration data, song streaming metrics from different platforms and predicting future trends.

II. Dataset Info

We have used 4 datasets for the assignment.

1) Spotify 30000 [3]
   The dataset consists of 30000 songs on Spotify from the year 1957 to 2020.
   - track_id: Song unique ID
   - track_name: Song Name
   - track_artist: Song Artist
   - track_popularity: Song Popularity (0-100), where higher is better
   - track_album_id: Album unique ID
   - track_album_name: Song album name
   - track_album_release_date: Date when album was released
   - playlist_name: Playlist name
   - playlist_id: Playlist ID
   - playlist_genre: Playlist genre
   - playlist_subgenre: Playlist subgenre
   - danceability: Danceability measure (0.0-1.0)
   - energy: Energy measure (0.0-1.0)
   - key: Overall key of the track, using Pitch Class notation
   - loudness: Loudness of the track in decibels (dB)
   - mode: Modality of the track (1 = Major, 0 = Minor)
   - speechiness: Measures spoken words in a track
   - acousticness: Measures whether a track is acoustic
   - instrumentalness: Predicts whether a track contains no vocals
   - liveness: Detects the presence of an audience
   - valence: Measures the positivity of a track (0.0-1.0)
   - tempo: Estimated tempo in beats per minute (BPM)
   - duration_ms: Duration of the song in milliseconds
2) Song rankings on Spotify and other streaming platforms [6]
   The dataset contains information on some of the most streamed songs on Spotify, with additional insights from other popular streaming platforms like Apple Music, Deezer, and Shazam.
   - track_name: Name of the song
   - artist(s)_name: Name of the artist(s) performing the song
   - artist_count: Number of artists contributing to the song
   - released_year: Year the song was released
   - released_month: Month the song was released
   - released_day: Day the song was released
   - in_spotify_playlists: Number of Spotify playlists the song is featured in
   - in_spotify_charts: Rank of the song on Spotify charts
   - streams: Total number of streams on Spotify

- in_apple_playlists: Presence in Apple Music playlists
- in_apple_charts: Rank of the song on Apple Music charts
- in_deezer_playlists: Presence in Deezer playlists
- in_deezer_charts: Rank of the song on Deezer charts
- in_shazam_charts: Rank of the song on Shazam charts
- bpm: Beats per minute (tempo) of the song
- key: Key of the song
- mode: Indicates whether the song is in a major or minor mode
- danceability%: Suitability of the song for dancing
- valence%: Positivity of the song's musical content
- energy%: Perceived energy level of the song
- acousticness%: Acoustic sound presence in the song
- instrumentalness%: Proportion of instrumental content in the track
- liveness%: Presence of live performance elements
- speechiness%: Amount of spoken words in the song

3) Spotify top weekly songs with collaboration and regional data [7]
This dataset contains songs from the Spotify 'Weekly Top Songs' charts for each country from the week of 2021/02/04 to 2022/07/14. For the 'Global' charts, the data ranges from the week of 2016/12/29 to 2022/07/14.

- uri: Spotify URI for the track
- rank: Ranking of the song on the chart
- artist_names: Names of all artists who participated in the song
- artists_num: Number of artists in the song
- artist_individual: Songs with multiple artists are split into separate rows for each artist. This will be one of the artists
- artist_id: Spotify artist URI for the artist in artist_individual
- artist_genre: Artist's genre. Since many artists had multiple genres, one of those genres was chosen for each row. The code for how genres were assigned can be found at my GitHub
- artist_img: Link to the artist_individual's image
- collab: 0 if there is only one artist, 1 if there are multiple artists
- track_name: Name of the track
- release_date: Release date of the album
- album_num_tracks: Number of tracks in the album that the track is from
- album_cover: Link to the album's cover art
- source: Song's record label
- peak_rank: Highest rank the song achieved on Spotify Charts
- previous_rank: Song's rank on Spotify Charts in the previous week (in a given country)
- weeks_on_chart: Number of weeks the song was on Spotify Charts (in a given country)
- streams: Number of streams in that week
- week: Week date
- danceability: Danceability of the song (from Spotify API Docs)
- energy: Energy of the song (from Spotify API Docs)
- key: Key of the song (from Spotify API Docs)
- mode: Modality of the song (from Spotify API Docs)
- loudness: Loudness of the song (from Spotify API Docs)
- speechiness: Amount of spoken words in the song (from Spotify API Docs)
- acousticness: Acoustic sound presence in the song (from Spotify API Docs)
- instrumentalness: Proportion of instrumental content in the song (from Spotify API Docs)
- liveness: Presence of live performance elements in the song (from Spotify API Docs)
- valence: Positivity of the song's musical content (from Spotify API Docs)
- tempo: Tempo of the song (from Spotify API Docs)
- duration: Duration of the song (from Spotify API Docs)
- country: Country that the chart data is from
- region: Region the country is in
- language: Language spoken in the country. Only one value is assigned for each country, even though many speak more than one language
- pivot: When multiple artists are split into separate rows, this value takes 0 for the first artist and 1 for the rest

4) Spotify and Youtube dataset [12]
This dataset contains songs from various artists around the world, and for each song, several statistics related to its version on Spotify, as well as data regarding its official music video on YouTube, are included.

- Track: Name of the song, as visible on the Spotify platform
- Artist: Name of the artist
- Url_spotify: The URL of the song on Spotify
- Album: The album in which the song is contained on Spotify
- Album_type: Indicates if the song was released as a single or is part of an album
- Uri: A Spotify link used to find the song through the API
- Danceability: Describes how suitable the track is for dancing based on musical elements like tempo, rhythm stability, beat strength, and regularity. Values range from 0.0 (least danceable) to 1.0 (most danceable)
- Energy: A measure of intensity and activity, with values from 0.0 to 1.0. Energetic tracks feel fast, loud, and noisy, while calm tracks score lower
- Key: The key the track is in, using standard Pitch

Class notation. E.g., 0 = C, 1 = C/D, etc. If no key is detected, the value is -1
- Loudness: The overall loudness of a track in decibels (dB). Typically ranges between -60 and 0 dB
- Speechiness: Measures the presence of spoken words in the track. Values range from 0.0 (non-speech-like) to 1.0 (entirely speech-like), with values between 0.33 and 0.66 indicating a mix of music and speech (e.g., rap music)
- Acousticness: A confidence measure from 0.0 to 1.0 indicating whether the track is acoustic. A value of 1.0 means high confidence the track is acoustic
- Instrumentalness: Predicts whether a track contains no vocals. A value closer to 1.0 means the track is more likely to be instrumental
- Liveness: Detects the presence of an audience in the recording. Higher values indicate a greater likelihood the track was performed live, with values above 0.8 suggesting it was recorded live
- Valence: A measure of the musical positiveness conveyed by the track, ranging from 0.0 (negative) to 1.0 (positive). Higher values indicate happy, cheerful, or euphoric music
- Tempo: The estimated tempo of the track in beats per minute (BPM), representing the speed or pace of the piece
- Duration_ms: The duration of the track in milliseconds
- Stream: The number of streams the song has on Spotify
- Url_youtube: The URL of the official music video on YouTube, if available
- Title: The title of the video on YouTube
- Channel: The name of the channel that published the video
- Views: The number of views the video has on YouTube
- Likes: The number of likes the video has on YouTube
- Comments: The number of comments the video has on YouTube
- Description: The description of the video on YouTube
- Licensed: Indicates whether the video represents licensed content, meaning it was uploaded by a channel linked to a YouTube content partner and claimed by that partner
- Official_video: A boolean value indicating whether the video found is the official video of the song

III. Task 1

A. Summary of visual analytics workflow for task-1

1) Summary of Assignment-1: We will consider Task 1 and Task 3 in Assignment-1 to be the first iteration of the feedback loop. They focus on analyzing song popularity with respect to song features.

A heatmap of the song features and track popularity was plotted, and it was found that there is no direct correlation between the song features and track popularity. Energy and loudness show a high positive correlation, while valence and danceability show a high positive correlation.

The metrics used to analyze artist popularity in Assignment-1 Task 3 were: Average track popularity and the count of tracks above a popularity threshold. The count above threshold was introduced to quantify how many tracks an artist has with a popularity score greater than a specific threshold. This defines the artist's popularity, which is used in A3.

The following visualizations were created:

- Correlation Matrix: A heatmap to show the correlation between track popularity and song features.
- Scatter plot to show the relationship between energy and loudness.
- Scatter plot to show the relationship between danceability and valence.

The major takeaways were:

- Song features alone cannot define the popularity of a song.
- Song features like valence and danceability, taken together, can influence song popularity.
- Song features like loudness and energy, taken together, can influence song popularity.

2) Second iteration of feedback loop: Building on insights from the first iteration of the feedback loop, the second iteration incorporates additional features that potentially influence a song's popularity. Social metrics such as views, likes, and comments were extracted from the YouTube dataset and integrated into the analysis.

The objective of this iteration is to identify the key factors affecting song popularity. A predictive model was developed to classify songs into popularity categories (binned in intervals of 10). Additionally, a feature importance graph was generated to visualize the contribution of each feature.

The findings from the first iteration indicated no direct relationship between song features and popularity. However, certain features exhibited strong positive or negative correlations with each other, collectively impacting a song's popularity. In this second iteration, attributes such as artist popularity, genre, social metrics, and release year were combined with song features to better understand and predict song popularity.

The following features were added:

- `valence_danceability`, `acousticness_loudness`, `energy_loudness`, `acousticness_energy`, `valence_danceability_energy`, `acousticness_energy_loudness`
- `tempo` was classified into three categories: slow, medium, and fast.

- `mood` was classified based on valence and danceability into the following categories: `happy_energetic`, `sad_energetic`, `happy_slow`, `sad_slow`.
- `clarity` represents the interaction between loudness and acousticness.
- `popularity_weighted`, was created by calculating a weighted sum of Likes, Comments, and Views.

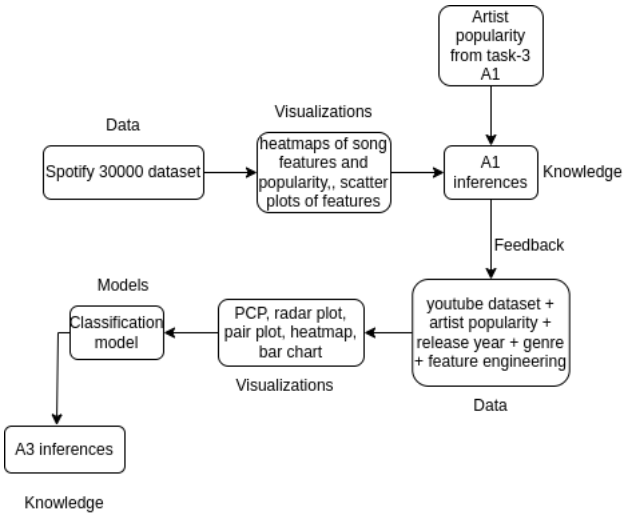

Fig 1.1: Unrolled visual analytics feedback loop diagram

B. Data Preparation

The following datasets were used for Task 3: Spotify 30000 [3], Youtube_Spotify [12]

The following data cleaning and preparation was done:

1) Spotify 30000 [3]:
   - The dataset was cleaned by dropping unnecessary columns such as track_id, track_name, track_artist, and others, which were not required for analysis or modeling.
   - The track name and artist columns were cleaned by stripping any leading or trailing whitespace and converting to lowercase for consistency.
   - Duplicate entries for tracks and artists were dropped to ensure unique records.
   - A new column for artist popularity was created based on the count of tracks with a popularity score above a specified threshold (≥ 75).
2) Parallel Coordinate Plot
   - The parallel coordinates plot for visualizing Spotify song features was developed using a combination of JavaScript and various libraries. Data preprocessing and formatting into a CSV file were performed using Pandas [1]. D3.js [14] was used to load and process the CSV data directly in the browser, while Plotly.js [15], utilizing its parcoords package, was employed to generate the plot. The brushing and axis reordering user interactions have been implemented.
3) Youtube_Spotify [12]:
   - The YouTube dataset was merged with the Spotify dataset based on track name and artist to integrate song views, likes, and comments.
   - Missing data from the YouTube dataset was handled appropriately, and rows with missing YouTube metrics were dropped.
   - A new feature for track popularity was computed using a weighted sum of likes, comments, and views.
4) Feature Engineering:
   - Several interaction features were created based on the correlations between existing features, such as valence_danceability, acousticness_loudness, energy_loudness, and others to capture relationships in the data.
   - Tempo was categorized into 'slow', 'medium', and 'fast' based on predefined tempo ranges.
   - Mood was classified based on valence and danceability, resulting in categories like 'happy_energetic' and 'sad_slow'.
   - Track duration was categorized into 'short', 'medium', or 'long' based on the track length in minutes.
5) Encoding:
   - One-hot encoding was applied to categorical features such as playlist_genre, tempo_category, mood, and duration_category to prepare the dataset for modeling.

C. Parallel Coordinate Plot

Each axis in the plot represents a different feature from the dataset. The axes include: 'track_popularity', 'genre_id', 'danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', and 'duration_ms'. Each line represents a song, and the lines are color-coded by genre (pop, rap, rock, latin, r&b, edm) using a discrete jet colormap. The Parallel Coordinate Plot (PCP), completed as part of Assignment 2, is utilized for gaining insights.

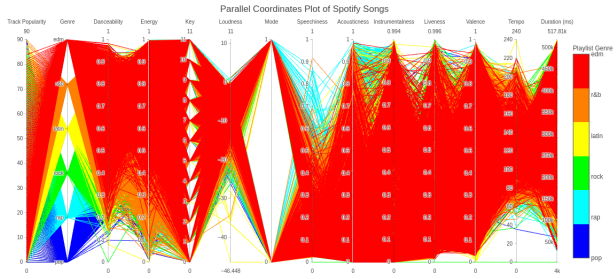

Fig 1.2: Parallel coordinates plot of Spotify song features

1) Features Across Genres:

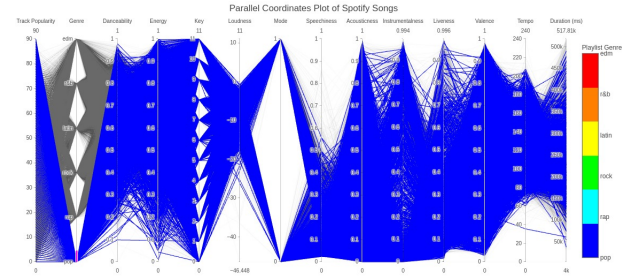

Fig 1.3: Brushing applied on the pop genre

2) Popular Song - Features and Genres: As shown in Fig. 1.4, brushing was applied to the track popularity axis to identify the most popular songs.

a) The majority of the most popular songs are from the pop and Latin genres.
b) There are relatively few highly popular songs in the rap, rock, R&B, and EDM genres.
c) The most popular songs appear to have low values for instrumentalness, liveness, and duration.
d) Valence (musical positivity) appears to be relatively low among the top songs.

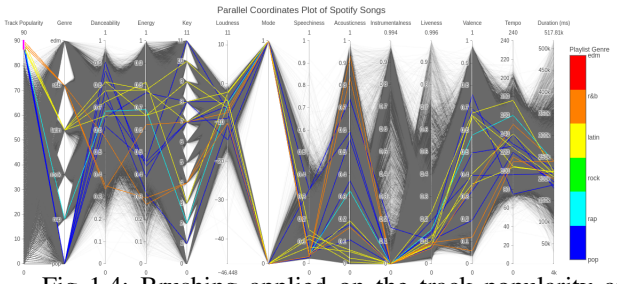

Fig 1.4: Brushing applied on the track popularity axis to select the most popular songs.

3) Speechiness: As shown in Fig 1.5, brushing was applied on the speechiness axis to select songs with high speechiness.
Songs with the highest speechiness values are predominantly rap songs, consistent with the genre's lyrical emphasis.

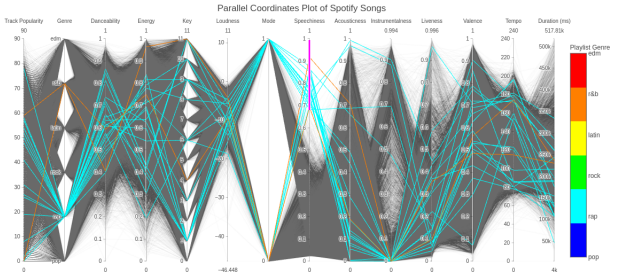

Fig 1.5: Brushing applied on the speechiness axis to select songs with high speechiness.

4) Loudness: As shown in Fig 1.6, brushing was applied on the loudness axis to select the quietest songs.
Songs with the lowest loudness values are primarily from the latin genre.

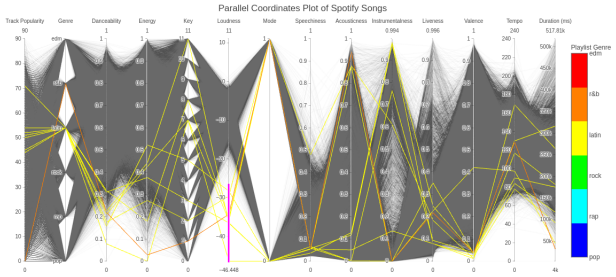

Fig 1.6: Brushing applied on the loudness axis to select the quietest songs.

D. Radar Plot

The radar chart from fig 1.7 visualizes various audio features (loudness, energy, danceability, etc.) and track popularity, compared across different music genres like EDM, Latin, Pop, R&B, Rap, and Rock. Below are some key observations:

1) Loudness and Energy: Higher loudness correlates with higher energy, with Rock showing the highest values.
2) Danceability: Rap has higher danceability, fitting its upbeat nature.
3) Acousticness and Instrumentalness: R&B shows higher acousticness, while Latin features more electronic instruments.
4) Speechiness: Rap has higher speechiness, due to its vocal focus.
5) Popularity: EDM consistently has high popularity, appealing to a broad audience.
6) Valence and Tempo: Genre-specific variations in valence and tempo reflect mood and pacing preferences.

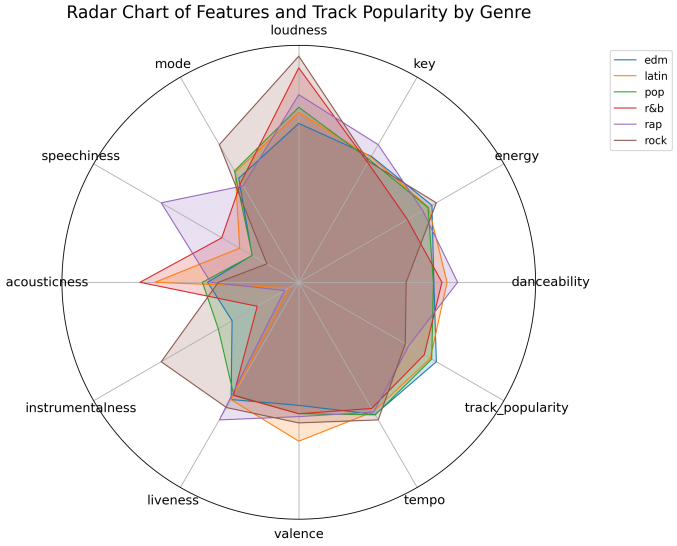

Fig 1.7: Radar Plot Showing Song Features and Track Popularity Across Genres

E. Pair_Plot

The pair plot illustrates the relationships between the following variables:

- Views: Total number of views a track has received.

- Likes: Number of likes indicating user appreciation for the track.
- Comments: Number of comments reflecting engagement and discussions about the track.
- Artist Popularity: A discretized metric representing the popularity level of the artist.
- Track Popularity: A normalized score (on a scale of 0 to 100) representing the popularity of individual tracks.

The diagonal of the pair plot shows the distribution of each variable, while the scatter plots below the diagonal display pairwise relationships between variables. This visualization is crucial for identifying correlations, outliers, and distribution patterns in the data.

1) Views, Likes, and Comments:

- There is a positive correlation between views, likes, and comments. Tracks with higher views generally receive more likes and comments, indicating that user engagement scales with exposure.
- The distributions of these metrics are highly skewed, with a majority of tracks having lower values.

2) Artist Popularity:

- Artist popularity appears to be discretized, possibly due to categorical ratings or predefined levels.
- Tracks by more popular artists tend to have higher views, likes, and comments, though the relationship is not strictly linear.

3) Track Popularity:

- Track popularity shows a moderate positive relationship with views, likes, and comments, suggesting that social media metrics significantly influence popularity scores.
- The track popularity distribution is also skewed, with most tracks not having intermediate popularity.

4

- nts,

Fig 1.8: Pair Plot Showing the relationship between social media metrics and track popularity

F. HeatMap

The heatmap (Fig 1.9) highlights the following key correlations:

The correlations observed in the heatmap reveal distinct relationships between various audio features and metrics. Positive correlations include valence with danceability, energy with valence, and energy with loudness, indicating that tracks with higher energy often exhibit higher valence and loudness. Negative correlations are observed between acousticness with both energy and loudness, suggesting that tracks with higher acousticness tend to be quieter and less energetic. Moderate correlations are seen between social media metrics (likes, comments, views) and track popularity, as well as between track duration and popularity, implying that while these factors influence popularity, their impact is not as strong as the other observed correlations.

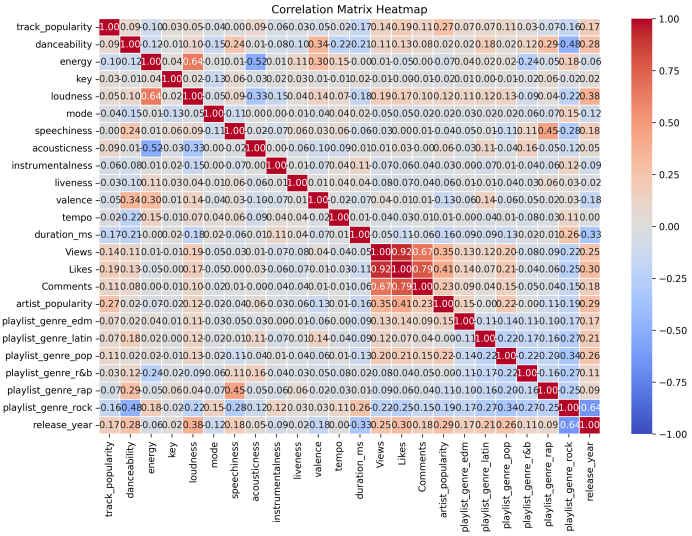

Fig 1.9: Heatmap showing relationship between the Features

1) Feature Engineering: Based on these correlations, the following new features were created:

1) valence_danceability: Capturing the high positive correlation between valence and danceability.
2) acousticness_loudness: Capturing the high negative correlation between acousticness and loudness.
3) energy_loudness: Capturing the high positive correlation between energy and loudness.
4) acousticness_energy: Capturing the high negative correlation between acousticness and energy.
5) valence_danceability_energy: Capturing positive correlation across valence, danceability, and energy.
6) acousticness_energy_loudness: Capturing negative correlation among acousticness, energy, and loudness.
7) Tempo-Related Features: Categorizing tempo as *slow*, *medium*, or *fast*.
8) Mood Classification: Classifying mood based on valence and danceability.

9) Acoustic Complexity (clarity): Combining loudness and acousticness.
10) Track Duration Features: Creating *duration_minutes* and categorizing duration as *short*, *medium*, or *long*.
11) Popularity Weighted Metric: Calculating a weighted metric using Likes, Comments, and Views.

G. Model - Predicting the Song Popularity

Our goal is to predict the song popularity using song features, artist popularity, playlist genres, and various mood and tempo metrics, *popularity_weighted* (weighted sum of social media metrics like views/comments/likes), release year, and the duration of the song.

To do this, we normalized the features using `MinMaxScaler` to ensure all features are on the same scale, aiding model convergence. The `track_popularity` column was transformed into ordinal categories (1 to 10) based on intervals of 10 for classification tasks.

1) Model 1: A Meta-Model Combining Predictions from Multiple Base Learners:

- Random Forest: An ensemble model using decision trees with bagging [16].
- Gradient Boosting: A boosting technique focusing on sequentially correcting errors of prior trees [17].
- Logistic Regression: A linear classifier acting as a simple baseline [18].
- XGBoost: An efficient implementation of gradient boosting optimized for speed and accuracy [19].

A final estimator (`RandomForestClassifier`) was used to combine predictions from these base models. The stacking classifier performed predictions on the test set and achieved an accuracy score of 61.89%.

Accuracy of Stacking Classifier: 0.6189573459715639
Classification Report for Stacking Classifier:

	precision	recall	f1-score	support
1	0.44	0.31	0.36	123
2	0.67	0.45	0.54	62
3	0.85	0.42	0.56	26
4	0.00	0.00	0.00	15
5	0.45	0.15	0.22	34
6	0.47	0.24	0.32	95
7	0.45	0.65	0.53	225
8	0.70	0.81	0.75	320
9	0.88	0.91	0.89	125
10	1.00	0.97	0.98	30
accuracy			0.62	1055
macro avg	0.59	0.49	0.52	1055
weighted avg	0.61	0.62	0.60	1055

Fig 1.10: Classification Report of Stacking Classifier

2) Model 2: Random Forest Classifier: The Random Forest Classifier was used to predict the song popularity and achieved an accuracy score of 61.04%.

Accuracy of Random Forest Classifier: 0.6104265402843602
Classification Report for Random Forest Classifier:

	precision	recall	f1-score	support
1	0.50	0.32	0.39	123
2	0.48	0.39	0.43	62
3	0.67	0.46	0.55	26
4	0.33	0.07	0.11	15
5	0.50	0.24	0.32	34
6	0.44	0.26	0.33	95
7	0.49	0.65	0.56	225
8	0.66	0.76	0.71	320
9	0.86	0.92	0.89	125
10	0.94	0.97	0.95	30
accuracy			0.61	1055
macro avg	0.59	0.50	0.52	1055
weighted avg	0.60	0.61	0.59	1055

Fig 1.11: Classification Report of Random Forest Classifier

3) Confusion Matrix for the Song Popularity Prediction: The Confusion Matrix tells us the number of points classified correctly and incorrectly in the respective classes. The points along the diagonal are predicted correctly, while the off-diagonal points are wrongly predicted.

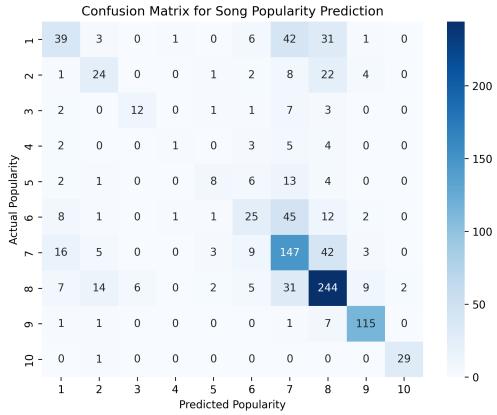

Fig 1.12: Confusion Matrix for Song Popularity Prediction

4) Reasons for Less Accuracy:

- Limited Feature Representation: The selected features might not fully capture the factors influencing song popularity, leading to reduced predictive power. There is still a need to include more features from different data sources that influence song popularity.
- Data Quality and Volume: Insufficient data points or noisy data may be the result in poor model generalization and reduced accuracy.
- Feature Interaction Complexity: The relationships between features and track popularity is non-linear and involve complex interactions, which are challenging to model effectively with the given features.

5) Important Feature in predicting the song popularity:

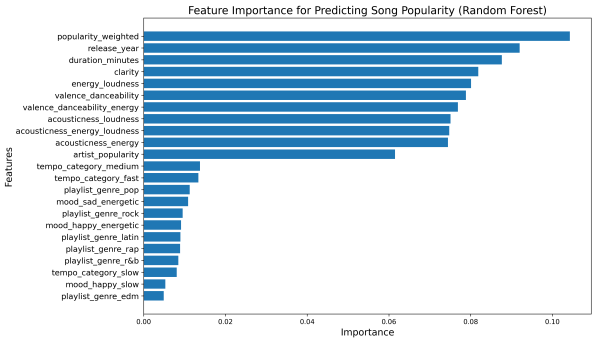

Fig 1.13: Feature Importance for Predicting Song Popularity

- popularity_weighted: The weighted sum of the views/likes/comments. It is the most influential predictor of song popularity.
- release_year and duration_minutes: Important structural features influencing popularity.
- Features like clarity, energy_loudness, and valence_danceability play a significant role in determining popularity.
- Combinations like acousticness_loudness and acousticness_energy are also key predictors.
- artist_popularity: Moderately impacts success but less than acoustic attributes.
- Tempo and Genre: Medium tempos and popular genres (pop, rock) have a higher influence compared to niche genres.
- Mood Features: Emotional and energetic moods contribute but are secondary to acoustic features.

H. Conclusion

- Correlation and Insights: Initial analysis (Assignment-1) shows no direct correlation between track popularity and song features like energy, loudness, or valence. However, combinations of features (energy and loudness, or valence and danceability) can influence popularity. In the second iteration, additional features such as social metrics (views, likes, comments) were incorporated to improve the prediction of track popularity.
- Feature Engineering: Several new features were created, including interactions between existing features (*valence_danceability*, *acousticness_loudness*), tempo classification, mood categorization, and a new feature for track popularity based on social media metrics (*popularity_weighted*).
- Visualizations: A correlation matrix and scatter plots were used to analyze relationships between song features and popularity. Parallel coordinate plots were employed to visualize the multi-dimensional relationships of song features across different genres. Radar plots compared the features and popularity of songs across various genres (EDM, Latin, Pop, R&B, Rap, Rock). Pair plots provided a deeper dive into relationships between social metrics (views, likes, comments) and track popularity.
- Analysis: The analysis emphasizes the complexity of defining song popularity. While individual song features may not directly correlate with popularity, the interaction of features—such as loudness and energy, or danceability and valence—along with social metrics (views, likes, comments), can significantly influence popularity. Additionally, genre-specific characteristics and the mood or tempo of a track also play a key role in shaping audience reception and engagement.

IV. Task 2

This task is done by building upon the foundational work of Task 2 of Assignment 1, we expanded the analysis with additional datasets and methodologies. The focus is on extracting deeper insights into the dynamics of genre trends, identifying patterns, and leveraging machine learning models for forecasting.

A. Summary of Visual Analytics workflow for task-2

The Task-2 of Assignment 1 will be considered to be taken as the first iteration of the feedback loop. It focuses on analyzing the popularity of playlist genres(Rap, Rock, R&B, Latin, Pop, EDM).

1) Summary of Assignment-1: The Spotify 30000 songs [3] has been used for this Task. The following visualizations were done:

- Distinct Count of Songs by Genre: Bar chart showing the distinct count of songs across six major genres.
- Song Releases Over Time: Line charts showing the distinct count of songs released per genre from 1957 to 2020.
- Genre Popularity Trends (2000-2020): Line charts illustrating the average popularity of songs for each genre over time.
- Top Songs in 2019: Treemaps representing top songs by average popularity across genres.
- Popularity Dynamics Within Genres: Bar chart showing average popularity for each genre and box plot showing the distribution of track popularity within each genre.
- Subgenre Contributions: Treemap showing subgenres' contributions to overall genre popularity and track counts.
- Album Popularity Distribution: Box plot analyzing song popularity distribution within albums containing more than some significant number of songs.

The major takeaways were:

- 2019 was a pivotal year for releases and popularity.
- Cross-genre tracks demonstrate broad appeal and higher success.
- Subgenres play a crucial role in driving parent genres' popularity.
- Variability in song popularity within albums.

2) Second Iteration of the feedback loop: By incorporating a new additional dataset (Spotify Weekly Top 200 Songs Streaming Data [7]) we expand upon the analysis performed in the first iteration i.e Assignment-1 by now also focusing on regional, temporal, and cross-genre dynamics of genre popularity. The goal is to build a more refined, iterative visual analytics workflow that connects the findings from Assignment 1 and provides more insights that deepen our understanding of genre dynamics. Machine learning models like Prophet are employed to predict genre popularity over time. K-means clustering was done to group countries based on their genre preferences.

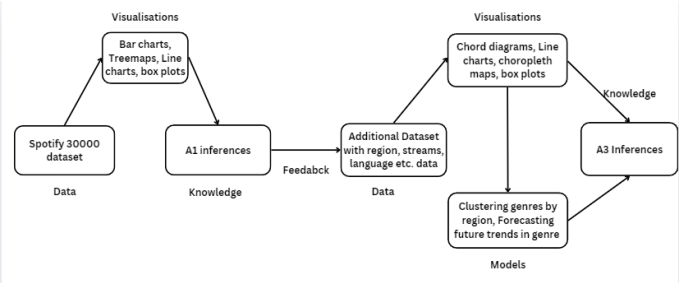

Fig 2.1: Unrolled visual analytics feedback loop diagram

B. Data Preparation

The following datasets were used for Task 2: Spotify 30000 [3] which contains detailed attributes about songs, including track popularity, audio features, and metadata such as playlist genres and subgenres and Spotify top weekly songs with regional data [7] which includes weekly rankings, streams, and regional information.

The following data cleaning and preparation was done:

1) Merging the datasets:

- Combined song-level metadata from Dataset 1 [3] with weekly performance and regional data from Dataset 2 [7].
- Cleaned the uri column in Dataset 2 to match the track_id format in Dataset 1.

2) Handling missing values:

- Dropped rows with missing values from the Spotify 30000 dataset as they were of very low percentage.
- Checked for null values after merging as well.

3) Removing Duplicates:

- Duplicate entries were present as several songs were present in multiple playlists and also as the songs were released in multiple regions.
- Duplicates were removed at various stages of preprocessing based on different subsets of columns:
  - track_id, playlist_genre, and country for unique country-level tracks.
  - track_id, month, and playlist_genre for monthly seasonal analysis.
  - track_id, playlist_genre, language, and month for language-genre relationship analysis.

C. Visualisations

To explore and understand the dynamics of music genres across different countries and regions we have the choropleth visualisations. Each map focuses on a distinct metric—popularity, song production, and streams. By analyzing all three, we aim to build a holistic view of how genres perform in terms of listener preferences, content creation, and audience engagement.

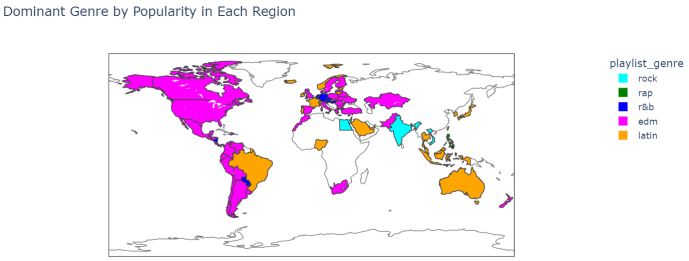

Fig 2.2: Choropleth showing Dominant Genre by Popularity

This visualization, Fig 2.2 highlights the most popular genres in different countries based on average track popularity. The results reveal several trends:

- EDM is the dominant genre in most countries globally, particularly in the Americas, Europe, and parts of Asia. Countries like Canada, the United States, and several South American nations consistently show high average popularity for EDM. This reflects the genre's widespread appeal and its association with global electronic dance culture.
- Some genres like Latin, R&B, and Rock dominate in select regions. Latin dominates in Brazil and several other countries, showcasing its strong cultural roots and the growing global influence of Latin music.
- R&B sees localized dominance in countries like Austria and Belgium, reflecting the niche but dedicated audience in European regions.
- Rock is highly popular in Egypt, Vietnam, and Hong Kong, signifying enduring appeal in certain regions despite its global decline.

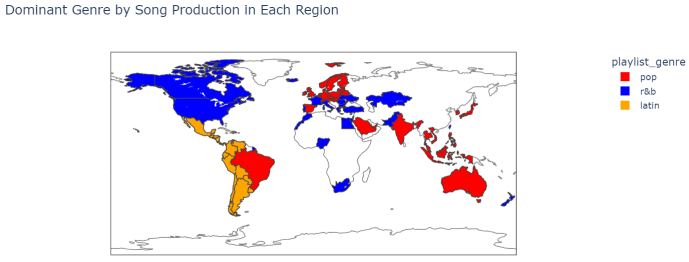

Fig 2.3: Choropleth showing Dominant Genre by Song Production

This map, Fig 2.3 visualizes the genres with the highest number of unique tracks produced in each country. It highlights

the volume of content creation.

- Latin music dominates in several South American and Central American countries, including Mexico, Nicaragua, and Panama. This aligns with the genre's rich cultural heritage and strong regional presence.
- Pop dominates in a majority of countries, particularly in Australia, much of Europe, and parts of Asia.
- Countries like South Africa and Canada show significant contributions to R&B production, reflecting strong markets for this genre.

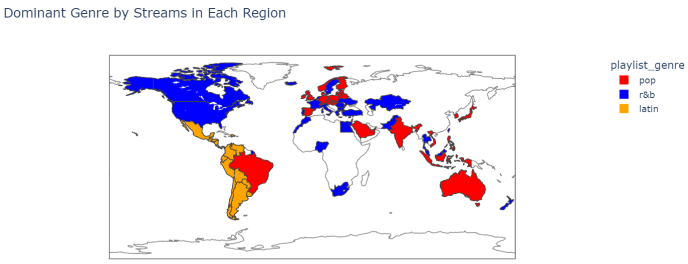

Fig 2.4: Choropleth showing Dominant Genre by Total Number of Streams

This map, Fig 2.4 shows the total number of streams per genre across countries, reflecting listener engagement and consumption.

- Pop emerges as the most streamed genre in several countries, including the United Kingdom, India, and the Philippines, suggesting its its presence on major playlists and accessibility on streaming platforms.
- Latin music dominates in many South American and Central American countries, such as Mexico and Colombia, showing high listener engagement within its regional fanbase.
- While not globally dominant, R&B achieves significant streaming numbers in countries like the United States and Canada

All three maps enables us to understand the interplay between content creation, listener preferences, and audience engagement. The differences among the metrics highlight the complexity of genre dynamics, where no single metric fully explains a genre's performance. Together, they provide a holistic view of global music trends and regional variations, allowing for informed inferences about market strategies, cultural influences, and evolving audience preferences.

The next visualization uses line charts to explore the seasonal trends in the total number of weeks each genre's songs stayed on the charts for each month. Each line represents a genre, with different colors used for clear distinction.The X-Axis represents the month (1 to 12), providing a temporal dimension. Whereas the Y-Axis shows the total weeks on the chart, indicating the cumulative engagement of tracks for each genre in that month (Fig 2.5)

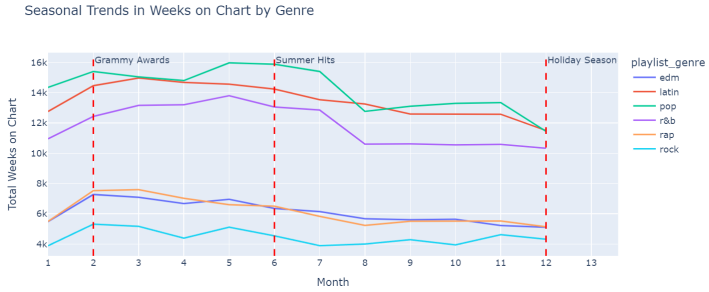

Fig 2.5: Line charts to show the seasonal trends in weeks on chart by genre

The visualization highlights seasonal patterns in the persistence of genres on the charts. It helps identify specific times of the year when certain genres dominate. Events like the Grammy Awards, summer hits, and the holiday season are marked to analyze how these impact the trends. By examining total weeks, we can infer a genre's engagement and how it sustains popularity over time.

Some of the key observations that we can get from the line charts are:

- Peaks in February and March correlate with the Grammy Awards, while summer highs align with outdoor music festivals.
- Pop and Latin maintain strong engagement across the year, reflecting their widespread appeal. Rock and EDM show narrower windows of dominance.
- The decline in performance of all genres during December could be because of the fact that during that period listeners gravitate toward holiday-themed music, such as Christmas and festive tracks. This shift leads to a decline in streams and chart positions for mainstream genres like Pop, Latin, Rap, and EDM that typically dominate other months.

The next analysis compares the performance of cross-genre tracks—those that appear in multiple genres—with solo genre tracks, which are restricted to a single genre. This is achieved through box plots for three key metrics: popularity, total streams, and weeks on chart. These visualizations offer insights into how tracks that transcend genre boundaries perform differently from those confined to a single genre.

Box plots were chosen for their ability to summarize and visually compare the distribution of data across multiple categories. They show key statistics such as the median, quartiles, and outliers, making it easy to understand trends and variations. The x-axis represents the track type (Cross-Genre or Solo Genre), while the y-axis shows the metric being analyzed (popularity, streams, or weeks on chart).

To identify cross-genre tracks, the number of unique genres associated with each track was calculated. Tracks appearing in more than one genre were labeled as Cross-Genre, while those associated with only one genre were labeled as Solo Genre.

Duplicate entries for the same track across different genres or subgenres were carefully filtered to ensure accurate analysis.

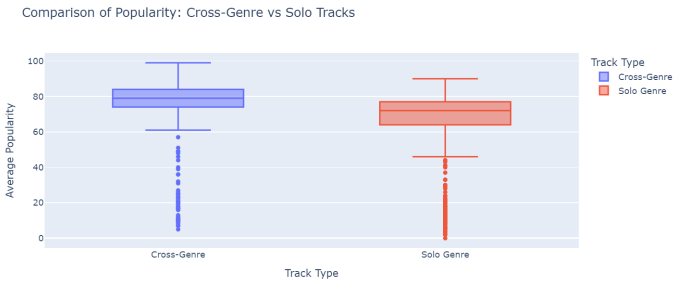

Fig 2.6: Cross-genre vs solo-genre analysis based on popularity

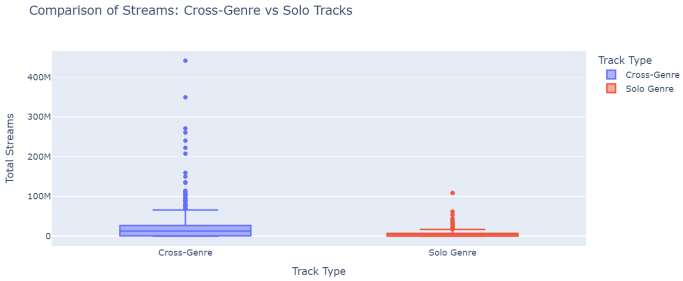

Fig 2.7: Cross-genre vs solo-genre analysis based on total streams

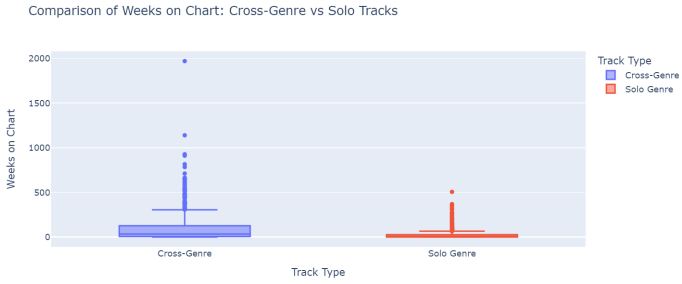

Fig 2.8: Cross-genre vs solo-genre analysis based on number of weeks on charts a track has streamed

From the three box plots (Fig 2.6, Fig 2.7, Fug 2.8) we can observe that,

- Cross-genre tracks consistently outperform solo-genre tracks in popularity, with a higher median score of 79 compared to 72. They also exhibit a broader range of high scores and fewer low-performing outliers, underscoring their stronger appeal across diverse audiences.
- With a median of 13 million streams—nearly triple that of solo-genre tracks—cross-genre tracks achieve significantly greater streaming success. Their maximum streams also far surpass those of solo tracks, highlighting their ability to captivate larger audiences.
- Cross-genre tracks dominate in chart longevity, with a median of 35.5 weeks compared to just 6 weeks for solo tracks. Cross-genre tracks maintain audience interest longer, reinforcing their lasting impact in the music industry.

This analysis quantitatively confirms observations from Assignment 1, where treemaps showed popular tracks appearing across multiple genres. While those visualizations hinted at cross-genre success, these box plots provide detailed evidence of the advantages of cross-genre tracks. They outperform solo-genre tracks in terms of popularity, streams, and weeks on chart, underscoring their broader audience reach and sustained appeal.

The findings suggest that cross-genre collaborations can be a strategic choice for artists and producers aiming for greater commercial success. By transcending genre boundaries, these tracks tap into diverse listener groups, achieving higher popularity and longer-lasting impact.

To get a deeper understanding of regional musical tastes and cultural inclinations we have clustered countries based on their genre preferences. This analysis complements the previously generated choropleths, offering a statistical perspective on how countries with similar preferences are grouped together.

To ensure reliable clustering, the data was processed in the following steps:

- Duplicate entries based on track ID, country, and playlist genre were removed to avoid biases in genre representation.
- Data was aggregated at the country-genre level, calculating the average popularity, total streams, and total weeks on the chart for each genre in each country.
- The data was reshaped into a pivot table where rows represented countries and columns represented genre-specific metrics, ensuring compatibility with clustering algorithms.
- The numerical features were scaled using the StandardScaler to normalize the data, preventing any one metric from disproportionately influencing the clustering.
- To improve the clustering process, Principal Component Analysis (PCA) was applied. This technique reduces the dataset's dimensionality by extracting the most significant features while retaining as much variance as possible.

Clustering was performed using the KMeans algorithm. To determine the optimal number of clusters:

- Inertia (WCSS): The total within-cluster sum of squares was calculated for different cluster numbers. The Elbow Method identified a point where adding more clusters yielded diminishing returns.
- Inertia decreased significantly as the number of clusters increased, with diminishing reductions after 4-5 clusters (Fig 2.9)
- Silhouette Score: This metric evaluated the quality of clustering by measuring how similar each point was to its cluster compared to others. Higher silhouette scores indicate well-separated clusters.

- Silhouette Score peaked at 2 clusters but remained acceptable for 5 clusters, chosen for a balance between interpretability and detail. (Fig 2.10)

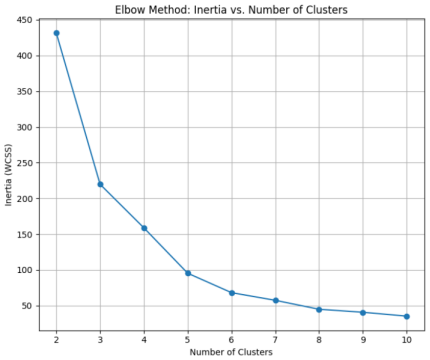

Fig 2.9: Elbow method to find optimal number of clusters

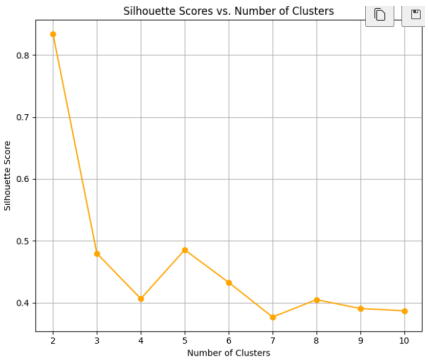

Fig 2.10: Silhouette scores to find optimal number of clusters

The PCA components enabled a clear 2D visualization of clusters and ensured the clustering algorithm focused on the most significant patterns in the data.

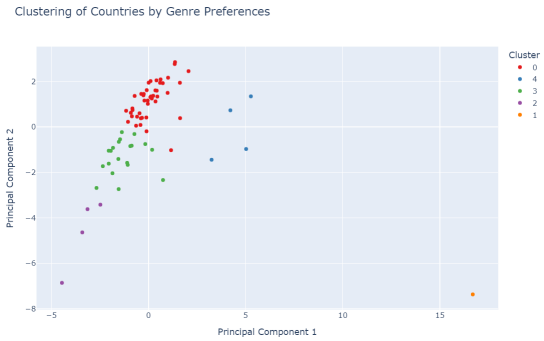

Fig 2.11: Countries clustered based on their genre preferences

With optimal number of clusters selected as 5, countries were grouped as follows:

Cluster 0 (Diverse Musical Preferences) Includes countries like Argentina, Belgium, India and Costa Rica. These countries have balanced preferences for multiple genres, such as Pop, Latin, and EDM. No single genre dominates. Audiences in these countries enjoy a variety of music styles, likely influenced by diverse cultures or exposure to global music.

Cluster 1 (Dominance in Global Trends) The US stands out as a unique cluster with a strong preference for globally dominant genres like Pop, Rap, and R&B.

Cluster 2 (Localized Preferences) Includes Belarus, Nigeria, Pakistan, and Turkey. These countries often displayed niche genre preferences, such as localized variations of EDM or regional influences, evident from their lower representation in global trends.

Cluster 3 (Preference for Global Popular Genres) Countries like Austria, Brazil, and Germany belong here. These countries favor well-known global genres, especially Pop and Latin, which are high in streams and popularity.

Cluster 4 (Modern and Streaming-Driven Preferences) Includes countries like Mexico, Canada, and Australia, reflecting their common preference of of modern high-streaming genres like EDM and pop.

The clustering reveals that musical tastes vary widely across countries, influenced by culture, exposure, and technology.

To uncover the relationships between languages and genres, chord diagrams have been plotted highlighting how music in different languages contributes to the popularity of various genres.

From the dataset, significant languages were identified based on total streams. Languages with the highest total streams were selected to ensure meaningful relationships could be observed. To improve visualization clarity, languages were grouped into regional categories (Asian and European)(Fig 2.12 and Fig 2.13). This division reduces complexity and allows for more focused insights.

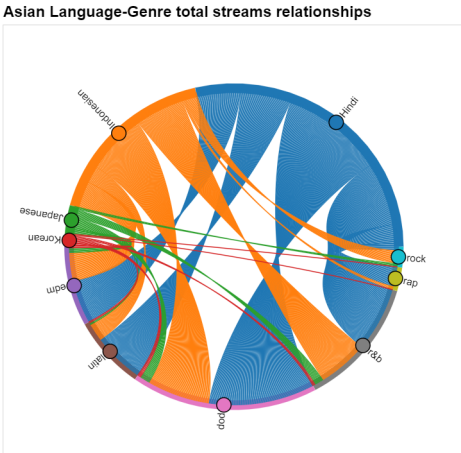

Fig 2.12: Chord diagram showing language genre relationship based on the number of streams: Asian languages

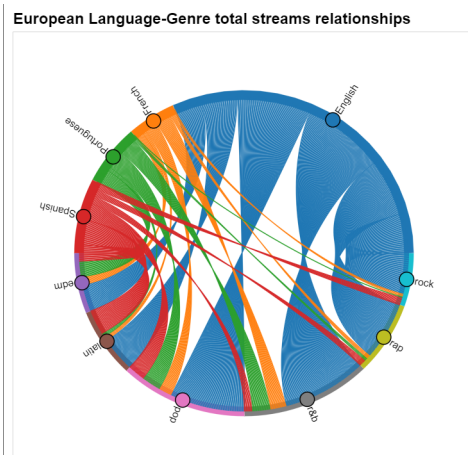

Fig 2.13: Chord diagram showing language genre relationship based on the number of streams: European languages

The nodes in the diagram represent either languages or genres each coloured differently. Links indicate the strength of the relationship between a language and a genre, quantified by total streams. A denser network of connections between a language and multiple genres suggests the versatility of the language's music or its broad appeal. For example, English's dense connections to multiple genres reflect its global dominance in the music industry. (Fig 2.10)

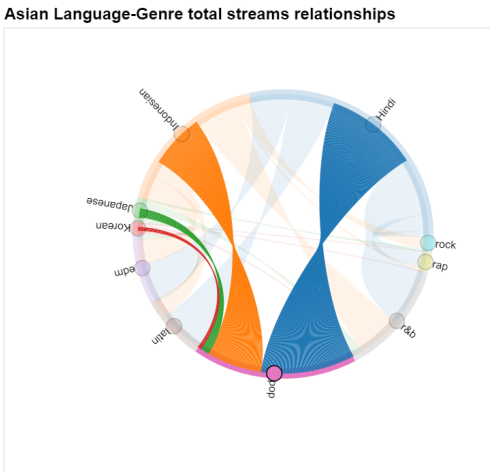

Fig 2.14: Pop genre node selected to display its connections with the European languages

Selecting a genre node displays all connected languages, revealing which languages dominate or contribute to that genre. As we can see in the Fig 2.14, among the Asian languages chosen for our visualisation, Hindi and Indonesian contribute significantly to the Pop genre compared to the other two languages.

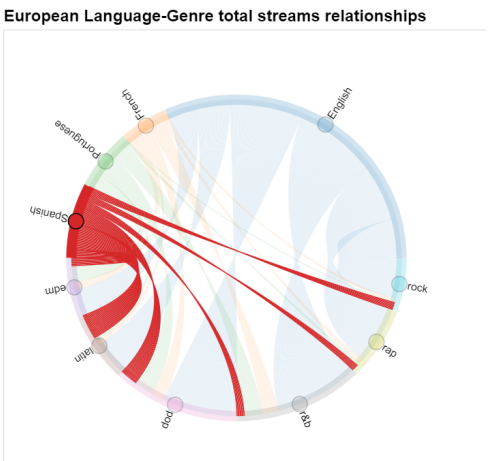

Fig 2.15: Spanish language node selected to highlight its relation to different genres

Selecting a language node shows the genres it connects to, highlighting the linguistic diversity within those genres. Selecting Spanish (Fig 2.15), highlights its connections to multiple genres like Latin and Pop. This indicates the widespread influence of Spanish-language tracks across various genres. Spanish's dominance in Latin is expected, but its presence in Pop showcases its global reach and cross-genre success.

The dense connections in these diagrams highlight successful cross-cultural trends, and with the interactive features we could get deeper exploration of specific relationships. These visualizations enrich our understanding of the multilingual nature of music and its global appeal.

D. Forecasting model

Building upon the trend line charts created in Assignment 1, which analyzed historical year-wise genre popularity trends, we focused on forecasting the future popularity trends of six music genres: rap, rock, pop, R&B, Latin, and EDM using prophet model in this task. This aims to provide a more forward-looking perspective by predicting future popularity trends for these genres.

Prophet [10] is a time series forecasting model suitable for capturing seasonal patterns and trends. The predictions help analyze how genres might evolve in the coming weeks.

We aggregated the popularity scores for each genre-week combination to calculate the average popularity. This aggregated metric formed the basis for forecasting.

Using the Prophet model, we trained the model on 80% of the data for each genre and validated it on the remaining 20%. The model captures seasonal patterns and trends in the data, making it well-suited for predicting genre popularity. For evaluation, we calculated MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error). These metrics assess the accuracy of the predictions, with lower values indicating better performance. Genre: edm - MAE: 2.31 - RMSE: 2.53 - MAPE: 2.84%

Genre: latin - MAE: 4.98 - RMSE: 6.35 - MAPE: 6.47%

Genre: pop - MAE: 2.47 - RMSE: 2.84 - MAPE: 3.07%

Genre: r&b - MAE: 3.39 - RMSE: 3.64 - MAPE: 4.16%

Genre: rap - MAE: 2.00 - RMSE: 2.42 - MAPE: 2.58%

Genre: rock - MAE: 8.27 - RMSE: 8.80 - MAPE: 10.90%

Once validated, the model was retrained on the entire dataset to forecast weekly popularity for the next year. Each genre was analyzed separately, and the results were visualized with line charts showing both historical data (as markers - red colour) and future predictions (as a smooth line) as can be seen in the images below.

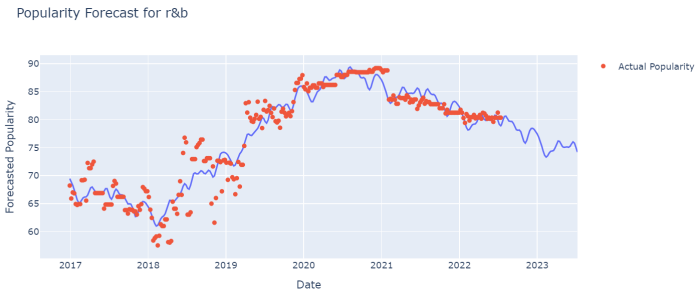

Fig 2.16: r&b popularity forecast

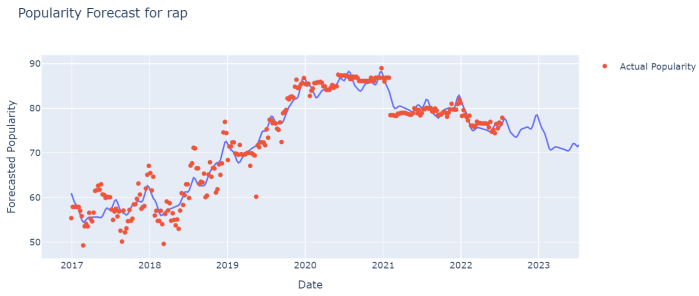

Fig 2.17: rap popularity forecast

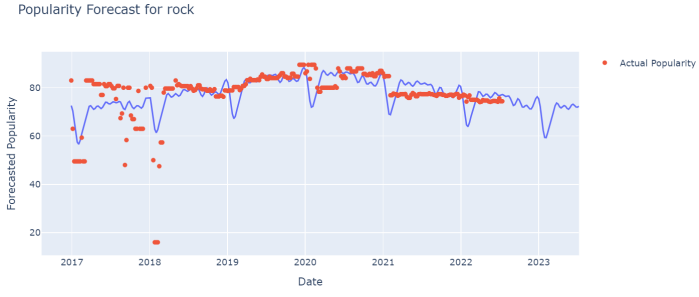

Fig 2.18: rock popularity forecast

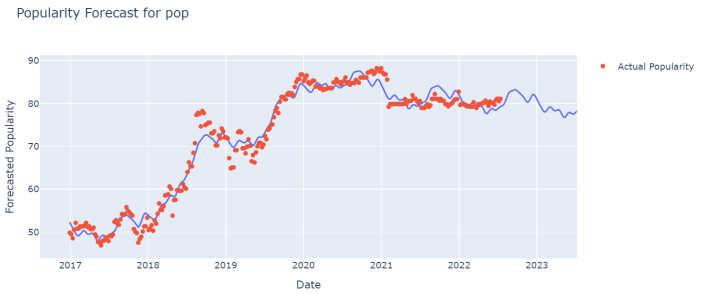

Fig 2.19: pop popularity forecast

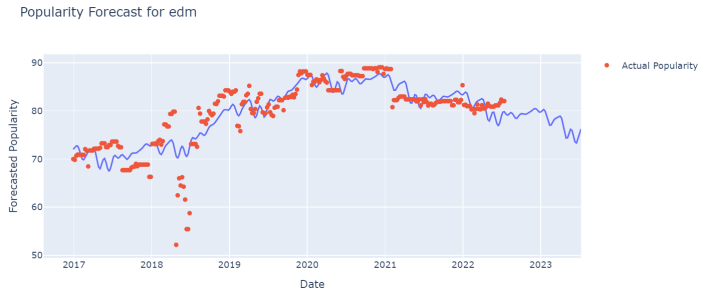

Fig 2.20: edm popularity forecast

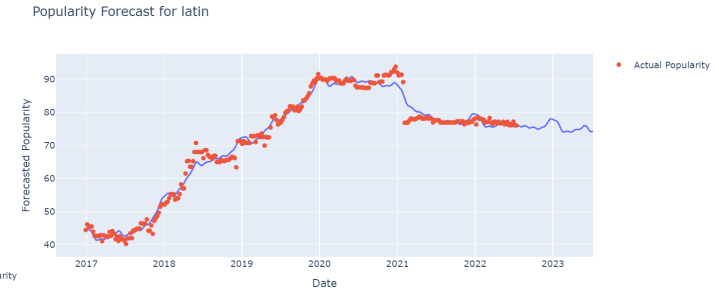

Fig 2.21: Latin popularity forecast

Pop and Latin maintain their consistent popularity and grow with occasional spikes. Our data of song popularity captures this perfectly, due to which we are able to forecast the trends in these genres.

Thus, pop and Latin show a good match between actual and predicted trends.

The visualizations for Rock and R&B revealed more irregularities and fluctuations in their historical trends. Rock genre had abrupt peaks and troughs in its historical data, making it challenging for the model to generalize these patterns effectively. Such irregular trends might arise from sudden changes in listener preferences or the influence of specific songs or artists that temporarily boosted the genre's popularity.

Despite attempts to include additional metrics, such as streams and weeks on the chart, these did not significantly improve the model's performance. While we considered track popularity as the primary metric, additional factors like social media trends could provide more context for future predictions.

E. Conclusion

The visual analytics workflow of this task provides a comprehensive understanding of global and temporal dynamics of genres by incorporating advanced visualizations, machine learning models, and feedback-driven iterations.

Some of the major takeawys include:

- Popular genres like Pop and Latin showed consistent dominance across regions and metrics, reflecting their broad global appeal.
- The analysis of cross-genre tracks highlighted their significant advantage in popularity, streams, and longevity compared to solo-genre tracks.
- Clustering revealed distinct regional differences in genre preferences, influenced by cultural and technological factors. This finding underscores the importance of tailoring content to regional tastes.
- Chord diagrams emphasized the complex relationships between languages and genres, showing how tracks transcend linguistic boundaries to achieve widespread success.
- Seasonal trends illustrated how global events and holidays impact music consumption, offering valuable insights for marketing strategies in the music industry.

V. Task 3

A. Summary of visual analytics workflow for task-3

1) Summary of Assignment-1: We will consider Task 3 in Assignment-1 to be the first iteration of the feedback loop. It focuses on analyzing artist popularity.

The metrics used to analyze artist popularity in Assignment-1 were: Average track popularity and Count of tracks above a popularity threshold. Count Above Threshold was introduced to quantify how many tracks an artist has with a popularity score greater than a specific threshold. This threshold was calculated using the 75^{th} percentile of song popularity. This helped identify artists who consistently release popular songs, rather than relying on just one or two hits. Based on the analysis done by visualizations (scatter plots), it was concluded that Count Above Threshold was a better measure of artist popularity.

The following visualizations were done:

- Bar Charts: Top artists by average popularity and count of tracks above threshold.
- Scatter Plots: Comparing metrics like track count, average popularity, and "count above threshold."
- Heatmap and Treemaps: Songs in different genres released by top artists, and genre-wise top artists
- Line Charts: Artist popularity trends over time.
- Polygon line chart: Average musical features for top artists

The major takeaways were:

- Count above threshold was found to be a better metric for artist popularity than average song popularity
- Artist popularity spiked with new releases
- Most popular artists diversify rather than focus on a single genre
- Popular artists produce songs with similar features

2) Second iteration of feedback loop: Building on the insights from the first iteration of the feedback loop, the second iteration introduces additional datasets and focuses on extending the analysis to incorporate collaborations, cross-platform popularity trends, and artist clustering.

The aim of iteration is to refine our understanding of artist success and future trends. Predictive modeling is done to understand future trends of artist clusters.

The first assignment introduced "Count Above Threshold" as a measure of popularity. The second iteration of the feedback loop expands this by analyzing how collaborations, cross-platform presence, and artist clusters influence this metric. Temporal trends of artist popularity (first loop) were expanded by incorporating multiple streaming platforms, for a view of cross-platform success. The artists were clustered based on their song features and cross-platform streaming metrics. Using Prophet time-series forecasting models, streaming trends were predicted per artist cluster rather than per individual artist.

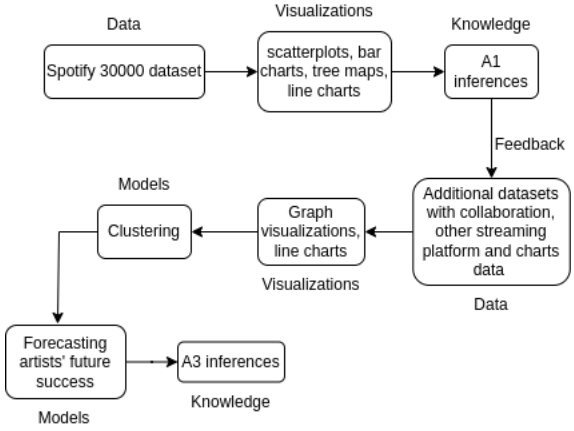

Fig 3.1: Unrolled visual analytics feedback loop diagram

B. Data Preparation

The following datasets were used for Task 3: Spotify 30000 [3], Song rankings on Spotify and other streaming platforms [6] and Spotify top weekly songs with collaboration data [7]

The following data cleaning and preparation was done:

1) Spotify 30000 [3]:
   - Songs with any missing values in any of the song features were dropped as they were a low percentage.
   - Several songs appeared in multiple playlists. These duplicate entries were dropped to ensure accurate analysis of unique song and artist popularity.
   - The "count above threshold" field was calculated for each artist using this data and saved into 'artists.csv'.
2) Cross-platform streaming metrics [6]

- Rows contained comma separated artists if a song was a collaboration. A new row was added for each artist in the collaboration, to ensure every artist is given credit for the song.
- Only the artists for which the ”count above threshold” data was available was retained.

3) Spotify top weekly songs with collaboration data [7]

- Nulls were dropped as they were a low percentage.
- Duplicates were present as the same song was released in several regions. These were dropped.
- Only the artists for which the ”count above threshold” data was available was retained.

C. Graph visualizations

Graph visualizations were chosen as collaboration networks are inherently relational, and graph visualizations are ideal for capturing and analyzing these relationships.

The data from Spotify top weekly songs with collaboration data [7] was used for this visualization. A node was created for each artist. Total streams and Count above threshold for all artists were added as attributes for each node. An edge-list was created between nodes based on collaborations between the artists.

Gephi [5] was used for the graph visualizations.

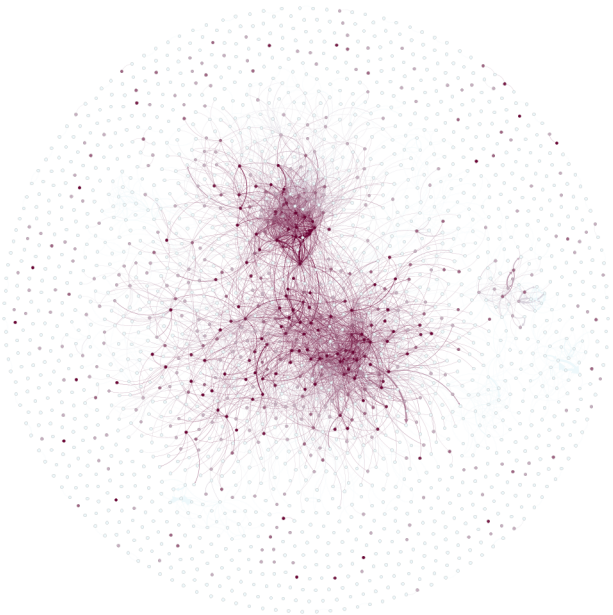

Fig 3.2: Nodes coloured based on ”count above threshold”

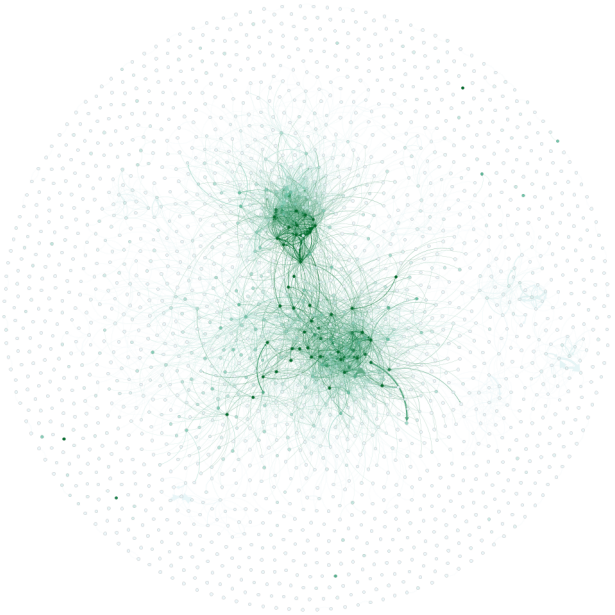

Fig 3.3: Nodes coloured based on total streams

The graphs were drawn using the Fruchterman-Reingold layout. This is a force-directed layout where nodes repel each other while edges (representing relationships) act as springs pulling connected nodes together. It results in a visually intuitive clustering, where tightly connected nodes form groups or communities, while less connected nodes remain more dispersed.

Nodes are colored based on the ”count above threshold” value using a sequential colormap in Fig 3.2 and based on total streams in Fig 3.3. Darker nodes have higher values, and lighter nodes have lower values.

Both the graphs are similar, indicating that the metrics are similar. The dense central cluster represents a highly collaborative group of artists who frequently work together. A huge number of the darker nodes belong to the densely connected cluster, indicating that most of the popular artists frequently collaborate with one another and with other lesser-known artists. The predominance of darker nodes (high values for both metrics) within this cluster indicates that collaboration amplifies popularity and streams.

Nodes on the outskirts are less connected or collaborate only within a specific subgroup, possibly representing niche or independent artists.

A modularity algorithm in Gephi was applied to detect specific communities or groups within the graph as shown in Fig 3.4. Modularity algorithms identify communities (groups of densely connected nodes) in a graph. The modularity algorithm used in Gephi is the Louvain Method for community detection [13].

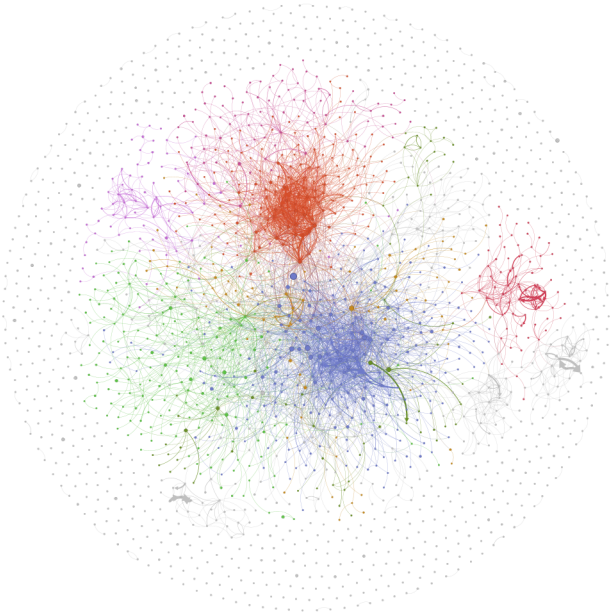

Fig 3.4: Modularity algorithms applied to detect clusters

The size of nodes in Fig 3.4 are proportional to the popularity of the artist, measured using ”count above threshold”. Each distinct color represents a community or cluster of artists who frequently collaborate with one another.

The central blue cluster is the largest, indicating a highly interconnected group of artists. The node sizes in this cluster are also large, indicating that it is a group of popular artists. The red cluster is dense but smaller than the blue, indicating another group of tightly-knit collaborators. Peripheral clusters (purple, green) suggest smaller groups of artists who collaborate within their community but have fewer connections to other clusters. The isolated red groups in the bottom-right represent independent communities with minimal interaction with the central network.

This indicates that more popular artists collaborate with other popular artists and rather than less popular artists. Likewise, artists with lesser popularity tend to collaborate with artits of similar popularity.

This is visualized better using a slightly zoomed image of the Force Atlas 2 layout of the same graph in Fig 3.5. Force Atlas 2 is a force-directed layout algorithm in Gephi that arranges network nodes by applying attractive and repulsive forces, creating a clear and aesthetically pleasing visualization of the graph’s structure. The node sizes in the blue cluster are large, indicating popular interconnected artists.

Fig 3.5: Zoomed force-atlas layout on clustered graph

The main inference is that collaboration correlates with popularity. Artists within the central cluster benefit from frequent collaborations, which amplify their reach and influence.

D. Time-series

The streaming metrics across platforms [6] dataset was used for this. The below fields are aggregated as sum: ’in_spotify_playlists’, ’in_spotify_charts’, ’streams’, ’in_apple_playlists’, ’in_apple_charts’, ’in_deezer_playlists’, ’in_deezer_charts’, ’in_shazam_charts’.

The top 10 artists are chosen based on the ”count above threshold” field and the metrics have been plotted for these in a small multiples time series line chart.

This was chosen because line charts excel at showing temporal trends, and small multiples allow for a direct comparison of the data across platforms.

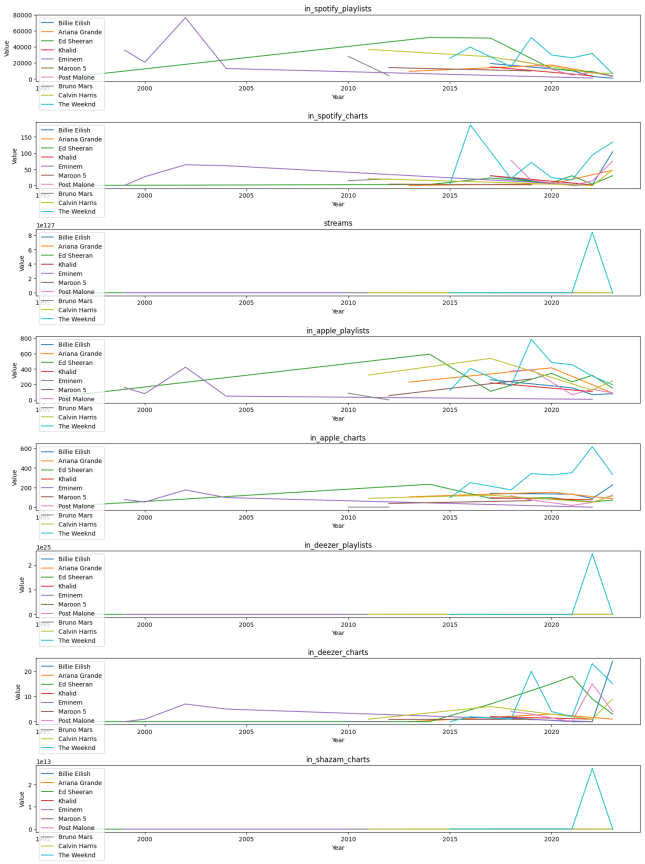

Fig 3.6: Streaming metrics across platforms time series graph

Comparing insights from the time-series line graph drawn in assignment-1 (Fig 3.8 assignment-1), both graphs highlight that artists experience spikes in popularity tied to successful album or single releases. For example, Billie Eilish and Ariana Grande in the new graph show dramatic spikes likely linked to major album drops.

The new graph introduces platform-specific performance, with metrics of songs on different platforms (Spotify, Apple Music, Deezer, and Shazam). This adds a new dimension in examining artist popularity. The older graph is more generalized and doesn't consider how specific platforms drive popularity.

The new graph emphasizes how playlist inclusions strongly correlate with chart success, particularly on platforms like Spotify. The platforms' algorithms likely amplify streams for songs included in curated playlists, especially on Spotify.

The Weeknd and Ed Sheeran show consistently high performance across various charts and playlists on platforms like Spotify, Apple Music, Deezer, and Shazam. Billie Eilish and Ariana Grande exhibit spikes in popularity, likely tied to the release of major albums or singles.

Artists have variable peaks across platforms, highlighting differing audience bases and promotional strategies.

E. Clustering

The cross-platform streaming metrics [6] dataset was used for this.

Artists were grouped using K-Means clustering based on the following:

- Musical Features: bpm, danceability, valence, energy, acousticness, instrumentalness, liveness, speechiness.
- Performance Metrics: in_spotify_playlists, in_spotify_charts, streams, in_apple_playlists, in_apple_charts, in_deezer_playlists, in_deezer_charts, in_shazam_charts.

The musical features were considered for clustering because in Fig 3.7 in Assignment-1, we observed that top artists had similar musical features. This indicates a correlation between musical features and popularity.

The features were aggregated using mean. They were scaled using StandardScaler [9].

The optimal number of clusters was chosen to be 5 using the elbow method.

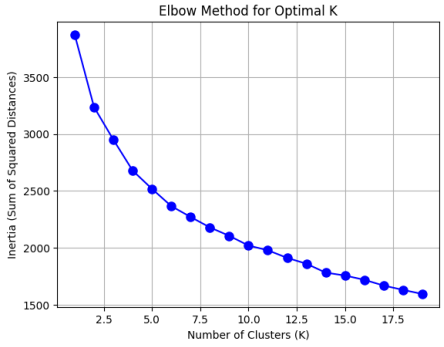

Fig 3.7: Elbow method for K-Means

Principal component analysis was done to reduce the data into 2 dimensions for visualizing the clusters. High-dimensional data is difficult to interpret; PCA reduces dimensionality while preserving as much variance as possible, making scatter plots ideal for visualization.

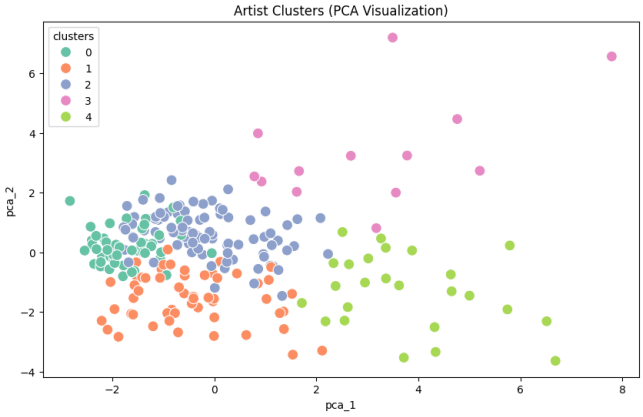

Fig 3.8: Clusters visualized using PCA in a scatter plot

A heatmap was created to visualize the average scaled features of each cluster.It was chosen because heatmaps effectively visualize aggregate metrics across clusters, allowing comparisons of multiple variables at a glance.

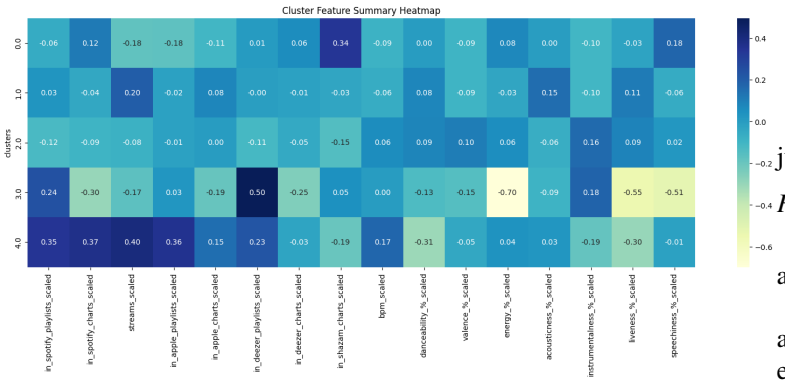

Fig 3.9: Heatmap for cluster features

The rows represent clusters, while the columns correspond to the features.

On observing the heatmap, the following insights can be made:

Cluster 4 includes many globally renowned artists such as 50 Cent, Adele, Arctic Monkeys, Avicii, Coldplay, Daft Punk, Guns N' Roses, Hozier, Imagine Dragons, Lewis Capaldi, Linkin Park, Nate Dogg, Nirvana, Queen, Radiohead, Rihanna, Snoop Dogg, Tears For Fears, The Chainsmokers, The Killers, The Police. There are some notable similarities that could explain why they may have been grouped together in a clustering analysis.

- These artists span multiple genres, including rock, pop, hip-hop, and electronic.
- Most have had timeless appeal, with music that continues to perform well on streaming platforms long after release.
- Artists such as Queen, Nirvana, Radiohead, and Guns N' Roses are legendary, with long-lasting influence in the music industry.
- Chart-topping artists like Adele, Rihanna, and Coldplay have seen massive success globally.
- Their music tends to remain in circulation across playlists and charts, even decades after release.

On examining cluster 2, it appears to include a diverse range of artists from multiple genres such as pop, Latin, hip-hop, rock, EDM, and R&B. Some of the artists are Ariana Grande, BTS, BLACKPINK, Bad Bunny, Doja Cat, Dua Lipa, Ed Sheeran, Justin Bieber, Lady Gaga, Ozuna, Post Malone, Shakira, The Weeknd, Sam Smith, Marshmello, Troye Sivan.

- This cluster stands out due to prevalent cross-genre collaborations. Examples include:
  - Ed Sheeran, who has collaborated with artists like Travis Scott (hip-hop) and Bad Bunny (Latin).
  - Post Malone, bridging genres through collaborations with Justin Bieber, Ozuna, and Swae Lee.
  - Shakira, known for collaborations with artists ranging from Rihanna (pop/R&B) to Maluma (Latin reggaeton).
  - Doja Cat, who mixes rap, pop, and R&B, with collaborators like SZA, Lil Nas X, and The Weeknd.
- Many of these artists are chart-toppers and have seen huge commercial success in recent years.
- Genres represented include pop, Latin, hip-hop, rock, EDM, and R&B, showcasing the versatility of these artists.

Based on this it is clear that clustering has been done not just on genre, but various other dimensions are involved.

F. Forecasting

We intend to predict future trends of artist popularity using a forecasting model.

A forecasting model may face challenges in cases where artists release new music in a year, as we can't predict the exact details of future releases. As observed in Fig 3.6, the artists see spikes of popularity tied to new releases. Also, we do not have enough data for single artists.

To handle this, instead of forecasting streams for individual artists, we forecast for clusters. We have sufficient data for the clusters. This smooths out individual artist variability and provides generalized trends for similar groups of artists.

The clusters created in the previous section using the cross-platform streaming metrics [6] dataset was used for this. A different forecasting model is created for each cluster. We try to forecast the song streams for the artists in each cluster as a whole.

The prophet [10] forecasting model was used.

Since we observed that collaboration correlates with popularity from the graph visualizations earlier (Fig 3.2, 3.3, 3.4, 3.5), the number of collaborations per year was added as an external regressor [11] for forecasting. This data was obtained from the Spotify top weekly songs with collaboration data [7] as used for the graph visualizations.

The data was scaled using StandardScaler [9]. The last 2 years were used for testing and the rest for training.

The performance of the forecasting models has been evaluated using two common metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

- Cluster 0 - MAE: 2.101892881273614, RMSE: 2.9725254193526065
- Cluster 1 - MAE: 0.6906212393875406, RMSE: 0.7105764724800105
- Cluster 2 - MAE: 1.245682847130813, RMSE: 1.6409534446122744
- Cluster 3 - MAE: 0.0, RMSE: 0.0
- Cluster 4 - MAE: 0.0, RMSE: 0.0

The errors for the last two clusters are 0 since the amount of data available for those clusters is limited.

Cluster 0 has a higher MAE and RMSE values, indicating that the predictions are less accurate.

The dataset does not account for important variables such as social media influence, cultural trends, or external promotion efforts, which could impact the popularity and trends of artists or songs. So, we might not be able to improve the accuracy with the existing data.

The forecasts for the next 5 years for each cluster have been plotted below. Line charts effectively communicate changes and trends over time, making them the best choice for visualizing forecasting.

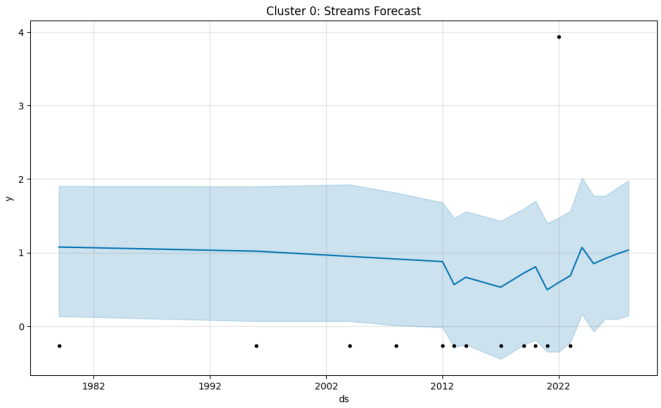

Fig 3.10: Cluster 0 streams forecast

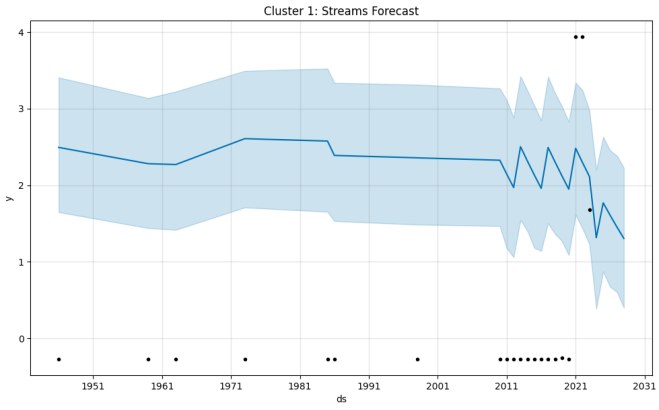

Fig 3.11: Cluster 1 streams forecast

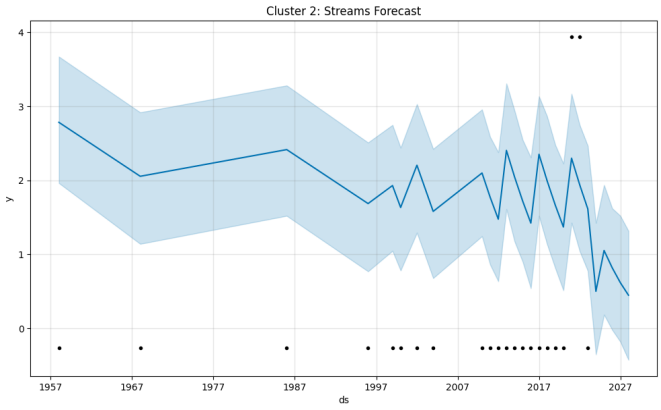

Fig 3.12: Cluster 2 streams forecast

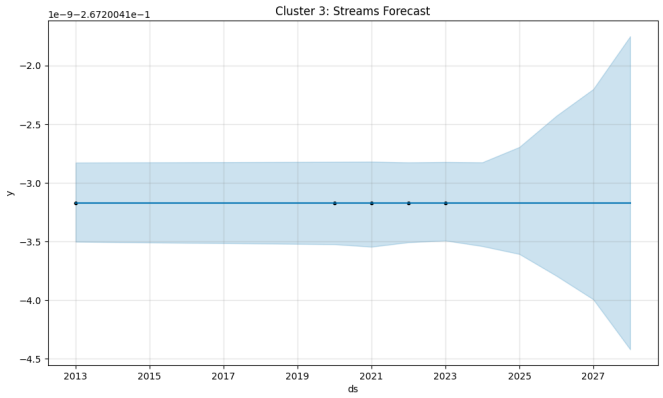

Fig 3.13: Cluster 3 streams forecast

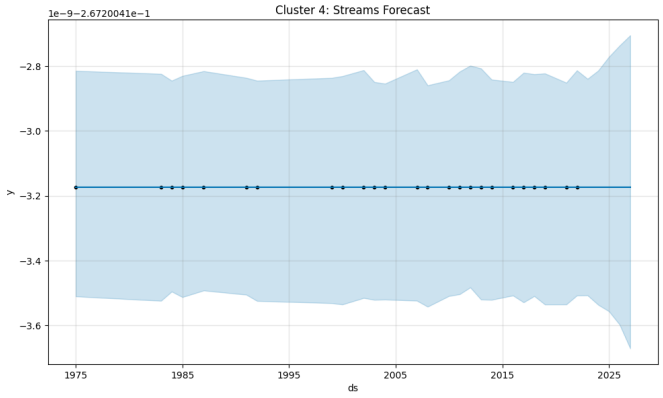

Fig 3.14: Cluster 4 streams forecast

The line indicates the predicted values (mean of the forecast). The blue area around the line in the chart represents the uncertainty interval or confidence interval of the forecast. Scaled values have been used for the streams in the y-axis, since we are just interested in observing the trend.

An interesting observation is that the forecast for Cluster 4 (Fig 3.14) shows a remarkably stable trend over time, with minimal variation in the predicted values. There is no significant upward or downward trajectory, indicating a steady popularity level for this cluster. This aligns with our observation from Fig 3.9 in the clustering section that most of the artists in this cluster have had timeless appeal, with music that continues to perform well on streaming platforms long after release. These artists do not have spikes of popularity like other artists.

As observed previously in cluster 2, the artists are recent chart toppers. Thus, we can see spikes and variability in the forecasts for cluster 2 (Fig 3.12).

G. Conclusions

This workflow revealed that artist popularity is multifaceted and is influenced by collaboration, cross-genre experimentation, and platform-specific trends. Some of the major takeaways are listed below.

- Collaboration influences popularity: Visualizations of artist networks revealed a strong correlation between collaboration and popularity. Artists in highly interconnected

clusters, particularly those in the central network, tend to have higher "count above threshold" values and total streams.

- Diverse genres and cross-platform success: Popular artists often diversify their genres, collaborating across different styles, as evident from clusters like Cluster 2 and Fig 3.6 in Assignment-1.
- Spikes in popularity tied to releases: Time-series data in Fig 3.6 in Assignment-3 and Fig 3.8 in Assignment-1 highlighted that contemporary artists experience sharp spikes in popularity tied to major releases. These spikes are often amplified by playlist inclusions on platforms like Spotify.
- Clustering based on musical features and streaming metrics: It revealed distinct groups. Popular artists are not confined to a single genre but often span multiple dimensions of success.
- Forecasting trends by clusters: Predictive modeling for artist clusters, rather than individual artists, provided insights into generalized trends. Legacy artists in Cluster 4 are predicted to maintain steady popularity, while others exhibit variability tied to potential new releases.

VI. Authors' contributions

Data preparation and cleanup required for each specific task was done individually by the respective team member.

- Saniya Ismail Kondkar: Task 1
- Ragini Metlapalli: Task 2
- Dyuthi Vivek: Task 3

References

[1] Pandas documentation https://pandas.pydata.org/docs/

[2] Matplotlib documentation https://matplotlib.org/

[3] Spotify 30000 dataset https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs?select=spotify_songs.csv

[4] NumPy documentation. https://numpy.org/doc/stable/

[5] Gephi documentation. https://gephi.org/users/

[6] Song rankings on Spotify and other streaming platforms. https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs

[7] Spotify top weekly songs with collaboration and regional data. https://www.kaggle.com/datasets/yelexa/spotify200

[8] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, et al. Visual Analytics: Definition, Process and Challenges. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North, editors, Information Visualization - Human-Centered Issues and Perspectives, volume 4950 of Lecture Notes in Computer Science, pages 154–175. Springer, 2008. https://link.springer.com/chapter/10.1007/978-3-540-70956-5_7

[9] Scikit-learn documentation. https://scikit-learn.org/stable/

[10] Prophet model documentation. https://facebook.github.io/prophet/docs/quick_start.html#python-api

[11] External regressors in forecasting. https://exploratory.io/note/kanaugust/Prophet-External-Predictors-Extra-Regressors-IeU4CHl5co#:~:text=You%20can%20add%20'External%20Predictors,on%20the%20past%20Sales%20data.

[12] Spotify and Youtube dataset. https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

[13] Louvain method for community detection. https://en.wikipedia.org/wiki/Louvain_method

[14] D3.js documentationhttps://d3js.org/getting-started

[15] Plotly.js documentation https://plotly.com/javascript/

[16] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

[17] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. https://projecteuclid.org/euclid.aos/1013203451

[18] Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley. https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+2nd+Edition-p-9780471356325

[19] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://dl.acm.org/doi/10.1145/2939672.2939785

Appendix

For the sake of completeness, the report for Assignment-1 has been added as an appendix from the following page.

Spotify Dataset Visualization

Dyuthi Vivek
IMT2022523
IIIT Bangalore
Bangalore, India
Dyuthi.Vivek@iiitb.ac.in

Saniya Ismail Kondkar
IMT2022128
IIIT Bangalore
Bangalore, India
Saniya.Ismail@iiitb.ac.in

Ragini Metlapalli
IMT2022029
IIIT Bangalore
Bangalore, India
Metlapalli.Ragini@iiitb.ac.in

I. Introduction

The primary objective of this analysis is to understand what makes a song, artist, or playlist popular. By analyzing different facets of music data, including song features, artist popularity, and genre trends, we aim to answer whether a song's popularity is driven more by the artist's brand or by the intrinsic features of the track. Additionally, we explore how artists, songs, and albums evolve in popularity over time and across genres.

The given dataset consists of 30000 songs on Spotify from the year 1957 to 2020. We have broken down our analysis into three subtasks:

- Task 1: Exploring the interplay of song features and music trends
- Task 2: Popularity trends over time and genre
- Task 3: Artist popularity over time and across genres

II. Dataset Info

- track_id: Song unique ID
- track_name: Song Name
- track_artist: Song Artist
- track_popularity: Song Popularity (0-100), where higher is better
- track_album_id: Album unique ID
- track_album_name: Song album name
- track_album_release_date: Date when album was released
- playlist_name: Playlist name
- playlist_id: Playlist ID
- playlist_genre: Playlist genre
- playlist_subgenre: Playlist subgenre
- danceability: Danceability measure (0.0-1.0)
- energy: Energy measure (0.0-1.0)
- key: Overall key of the track, using Pitch Class notation
- loudness: Loudness of the track in decibels (dB)
- mode: Modality of the track (1 = Major, 0 = Minor)
- speechiness: Measures spoken words in a track
- acousticness: Measures whether a track is acoustic
- instrumentalness: Predicts whether a track contains no vocals
- liveness: Detects the presence of an audience
- valence: Measures the positivity of a track (0.0-1.0)
- tempo: Estimated tempo in beats per minute (BPM)
- duration_ms: Duration of the song in milliseconds

III. Data Preparation

Before delving into the analysis, the following data cleaning and preparation was necessary:

- Handling missing data: Songs with any missing values in any of the song features were dropped as they were a low percentage.
- Removal of duplicates: Several songs appeared in multiple playlists. These duplicate entries were dropped to ensure accurate analysis of unique song and artist popularity.
- Removal of duplicates from albums: Removed duplicate combinations of track_id and track_album_id to ensure each track was counted only once per album and then filtered the albums to only include those with more than one track.

IV. Task 1

We explored how various audio features, such as energy, loudness, valence, and danceability, influence track popularity. We discovered that no single feature strongly correlates with popularity. Instead, deeper patterns emerged in the relationships between these features.

Through two sub stories—"Energy and Loudness – Driving the Beat" and "Mood Matters – Valence and Danceability"—we explored how these characteristics interacted across genres and evolved over time. This journey uncovers valuable insights into how music's structure and mood shape its impact across different genres and eras.

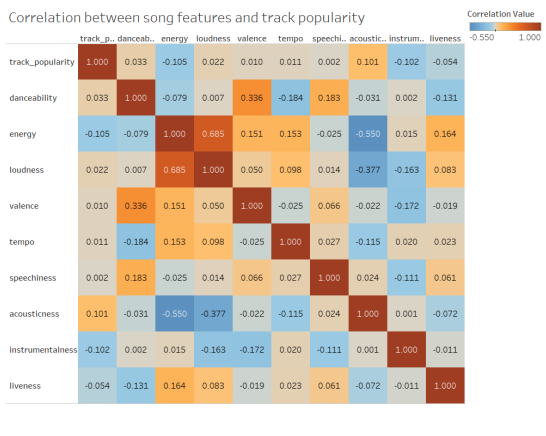

Fig 1.1: HeatMap to show correlation between track popularity and Song Features

The heatmap is chosen to visualize correlations because it effectively highlights relationships between variables using colour gradients, with cold colours representing weaker correlations and warm colours indicating stronger ones.

The heatmap shows that track popularity has a low correlation with all features, meaning individual track attributes don't strongly influence popularity. However, the heatmap also highlights other significant correlations:

- Valence and Danceability show a high positive correlation, suggesting that tracks that are more danceable tend to have a happier mood.
- Energy and Loudness are highly correlated, meaning tracks with higher loudness are typically more energetic.

A. *Mood Matters – Valence and Danceability*

This substory focuses on the relationship between valence (emotional positivity), danceability, and acousticness. It examines how these features interact across different genres and evolve over time, revealing key insights into how mood and rhythm affect a song's characteristics.

Key Question: How do valence, danceability, and acousticness interplay, and what trends can we observe over time?

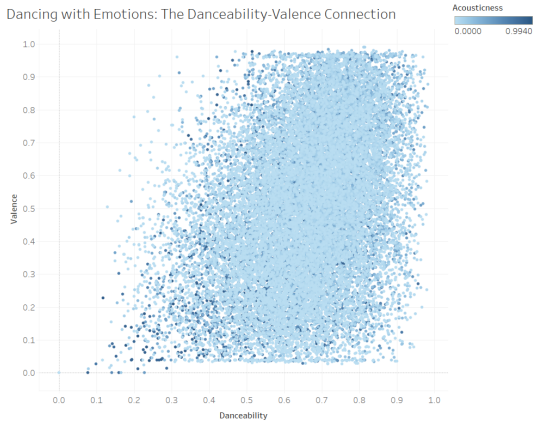

Fig 1.2: Scatter plot to show relation between danceability and valence

The scatter plot was chosen to clearly show the relationship between valence and danceability, with color used to represent acousticness. The marks (points) effectively display individual tracks, while the channel (color) highlights the distribution of acousticness without overwhelming the main trend.

The scatter plot shows a positive relation between valence and danceability. This means that as the emotional positivity of a track increases, its danceability tends to increase as well. Tracks that are happier often encourage movement and energy.

Acousticness is evenly spread across the scatter plot, with no clear influence on the valence-danceability relationship. Mood and danceability seem to operate independently of a track's acoustic nature. While acousticness has a low correlation with these features, it adds an extra layer of understanding about the production style of the music, helping to differentiate between more organic and synthetic tracks.

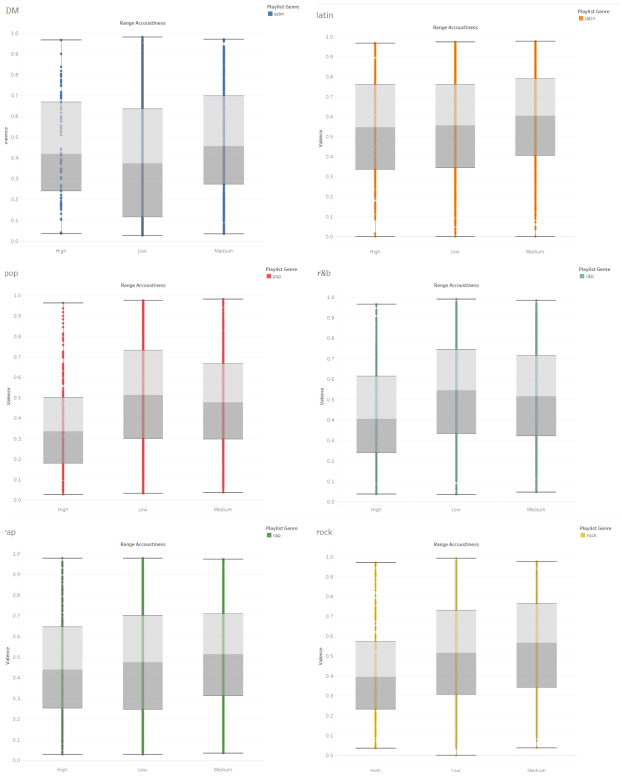

Fig 1.3: Box plot of valence range of acousticness for various genres

The box plot is chosen to effectively display the distribution, medians, and variability of valence across acousticness categories. The marks (boxes and whiskers) provide a clear view of central tendency, spread, and outliers, while different colors for genres for easy identification.

The box plot shows that across all genres, the general trend is that medium acousticness corresponds to the highest valence, suggesting that a balanced acoustic profile tends to be linked with more positive moods in music. However, the pattern of valence differs slightly depending on the genre:

EDM: Valence follows the pattern *Medium* > *High* > *Low*. This could be because EDM often blends synthetic sounds

with acoustic elements, and tracks with higher acousticness might strike a better balance between energy and emotion.

Latin, Rap, Rock: The trend is *Medium* > *Low* > *High*. Medium acousticness here might offer a mix of organic and synthetic elements that listeners find appealing, while high acousticness, which is more acoustic or traditional, might not resonate as strongly with modern listeners.

Pop, R&B: The pattern is *Low* > *Medium* > *High*. In these genres, tracks with low acousticness (more synthetic production) are more emotionally positive. This could be tied to the modern production styles in Pop and R&B, where synthetic sounds dominate and are linked to upbeat, high-energy tracks.

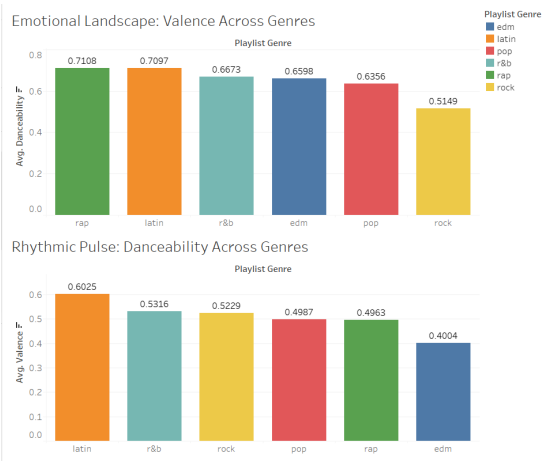

Fig 1.4: Bar chart of avg valence and danceability across genres

The bar chart was chosen for its ability to clearly compare average danceability and valence across genres. The marks (bars) effectively represent the average values for each genre, while the channel (color) distinguishes between genres given for better visuals.

The above bar chart shows that rap has the highest average danceability, followed by Latin, R&B, EDM, Pop, and Rock. This suggests that Rap and Latin music are more rhythmically engaging, while Rock, with its slower tempo, scores lower on danceability.

In terms of valence, Latin music ranks highest, followed by R&B, Rock, Pop, Rap, and EDM, which has the lowest valence. Latin music's positive emotional tone contrasts with EDM's moodier vibe. These trends highlight how genres differ in both their physical engagement

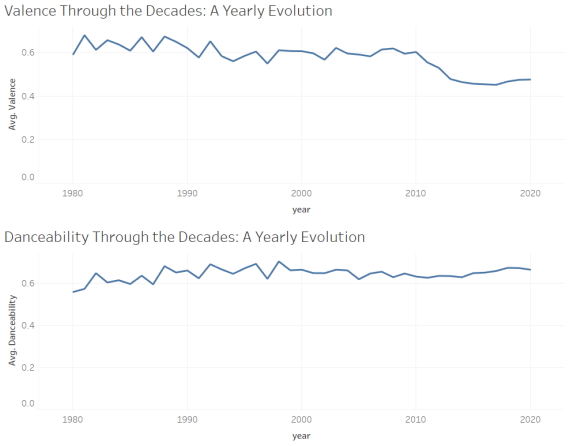

Fig 1.5: Yearly trend of avg valence and danceability

The line chart was chosen for its ability to effectively display changes over time. The marks (lines) track year-wise average valence and danceability.

- Year-Wise Average Valence: After 2000, the average valence of music noticeably decreases, indicating that music has become less positive in emotional tone. This shift may be influenced by the rise of introspective genres like alternative rock, hip-hop, and electronic music, as well as societal and cultural changes.
- Year-Wise Average Danceability: Post-2000, average danceability stabilizes, showing little variation. Unlike earlier years with fluctuating trends, music's danceability has remained relatively consistent.

B. Energy and Loudness – Driving the Beat

This substory explores how energy and loudness are highly correlated across various genres. It looks at trends over time, showing how energy levels have evolved in genres such as EDM, Pop, R&B, Latin, and Rap.

Key Question: How do energy and loudness interact across genres and over time?

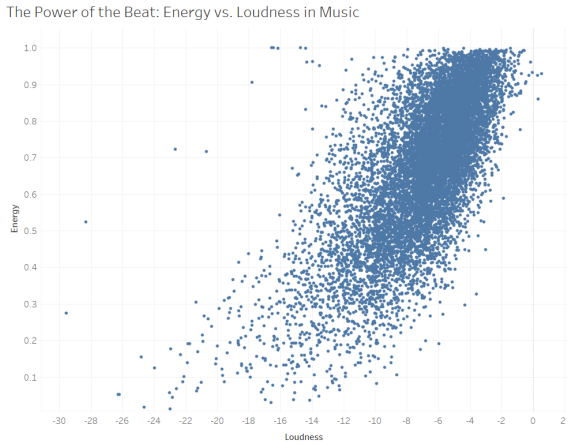

Fig 1.6: Scatter plot to show relationship between energy and loudness

The scatter plot is used here because it effectively shows the relationship between two continuous variables—energy and loudness. The marks (points) represent individual data points for each track, while the channel (position on the axes) displays the correlation between energy and loudness.

Each point represents a track, and the spread of points allows us to observe the overall trend, as well as any outliers or clusters. The scatter plot shows a positive correlation between energy and loudness. As the loudness of a track increases, so does its energy, suggesting a strong link between these two features. Louder tracks are generally more energetic, and vice versa.

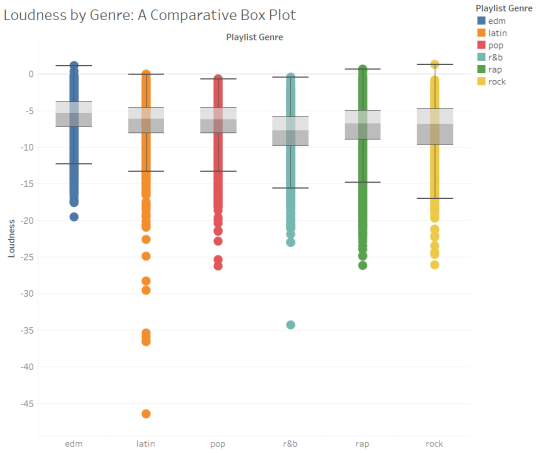

Fig 1.7: Box plot of loudness across genres

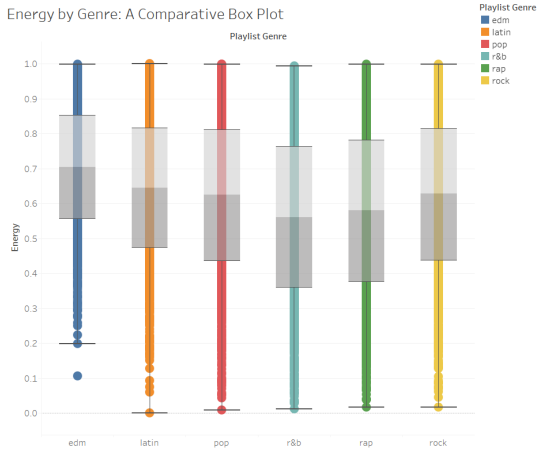

Fig 1.8: Box plot of energy across genres

The box plot is chosen here because it effectively compares the distribution of energy and loudness across multiple genres. The marks (boxes and whiskers) show the distribution, median, and range, while the channel (position on the y-axis) highlights variations in energy and consistency in loudness across genres.

The box plot reveals that the median loudness is almost the same across genres, but median energy varies significantly. EDM has the highest median energy, followed by Latin, Rock, and Pop, while Rap and R&B have lower median energy.

Genres like EDM and Latin music are more energetic, while genres like Rap and R&B tend to be less so. This suggests that different genres use loudness and energy in different ways to create their characteristic sounds.

Fig 1.9: Yearly trend of avg loudness and energy for different genres

The line chart uses lines to represent average energy and loudness trends over time, with position on the x-axis (years) and y-axis (energy levels) allowing for clear visualization of temporal changes and genre-specific patterns.

- EDM: Energy peaks in 1995 and then almost same
- Latin: A similar peak in energy is seen then same trend
- Pop: Energy drops sharply around 2000 and then remains relatively constant.
- R&B: Energy drops sharply in 1985.
- Rap: Energy increases sharply in 1985 and then remains constant.

For EDM, the peak in 1995 could be due to the rise of high-energy subgenres like Trance and Techno, which were later replaced by more melodic or mainstream EDM styles. Pop experienced a shift towards softer, more vocal-focused music around 2000, explaining the drop in energy. R&B in the mid-1980s moved towards smoother, slower styles, while Rap saw an increase in energy during the same period, likely due to the rise of more aggressive rap styles such as Hardcore Rap.

V. Task 2

The focus of this Task 2 is to analyze how genres have evolved in popularity over time. We explore the number of

songs released, the trends in popularity over different periods, and the top songs in recent years across various genres to understand the dynamics of genre-based trends, pinpoint key drivers, and observe shifts in audience preferences.

The first visualization is a bar chart which focuses on the volume of song production by genre. It highlights which genres are the most prolific in terms of releases, giving insight into the breadth of music produced.

This visualization was chosen because it provides a clear and straightforward comparison of the distinct count of songs across genres. By using a bar chart, the data clearly communicates which genres are most active in producing new music.

Each bar is assigned a different color to easily distinguish between the genres. The colors are bold to ensure clear visual separation, allowing for quick comparison.

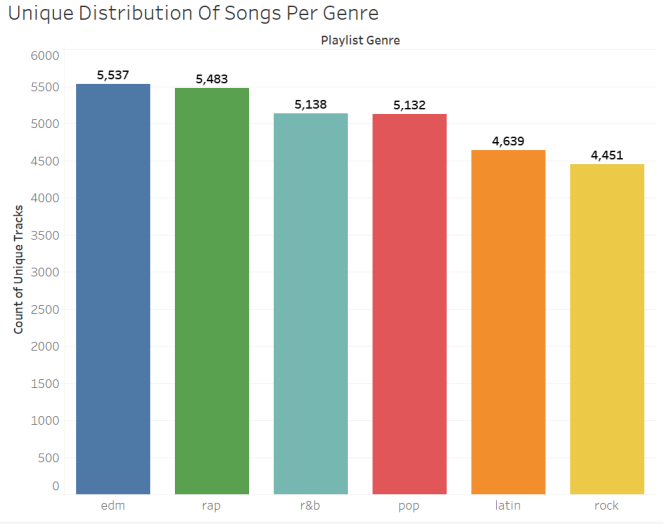

Fig 2.1: Distinct Count of Songs by Genre

- Rap and EDM stand out, Pop and R&B also show a high volume of tracks
- Rock shows lower production levels compared to other genres
- Fewer unique songs in Rock aligns with its general decline, as evidenced by their lower number of releases and decreasing popularity trends as seen in fig-2.2 and fig-2.3 to be discussed later.

It is important to understand not just the size of each genre's song library, but also the temporal trends that contributed to their evolution. To help us dive deeper into how song production has evolved over time, we have six individual line charts that plot the distinct count of unique tracks released in each genre from 1957 to 2020.

Line Charts were chosen because they effectively capture changes in song releases over time, making it easy to identify both gradual and sudden shifts in the volume of production for each genre. Each genre has its own dedicated chart to avoid clutter and provide clear insights into individual genre trends.

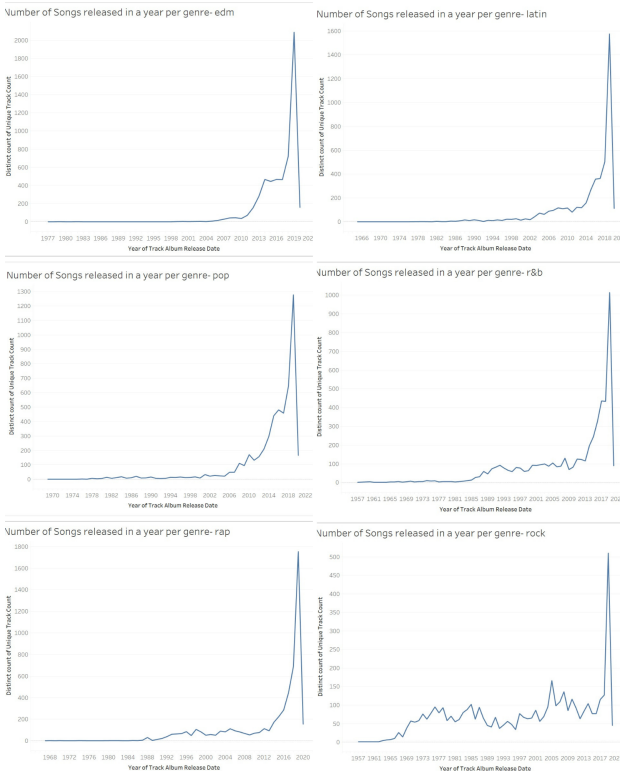

Fig 2.2: Number of Songs Released Per Year by Genre

- Across all genres, 2019 stands out as a peak year for releases and then there is a dip in 2020 which could likely be attributed to the pandemic, reflecting a slowdown in music production across the board.
- Genres like EDM, Latin, and Rap saw significant growth after 2000, reflecting their rise to mainstream prominence during the digital age.
- Both pop and R&B experienced steady increases in releases throughout the 2000s and 2010s, reinforcing their long-standing appeal in mainstream music.
- Rock has experienced less growth compared to other genres, reflecting a possible shift toward more contemporary genres.

After analyzing the number of songs released per genre, we now turn our focus to how the popularity of songs has evolved over time, specifically from the year 2000 onwards across each genre. In the previous visualizations, we observed that the number of songs before 2000 was relatively insignificant, so we have considered only post-2000 songs for this analysis.

A line chart was chosen to illustrate the average popularity of songs in each genre over time. This helps visualize the growth or decline in popularity for each genre, showing which genres have been gaining or losing popularity over the years. Lines: Represent the average popularity of tracks by year for each genre.

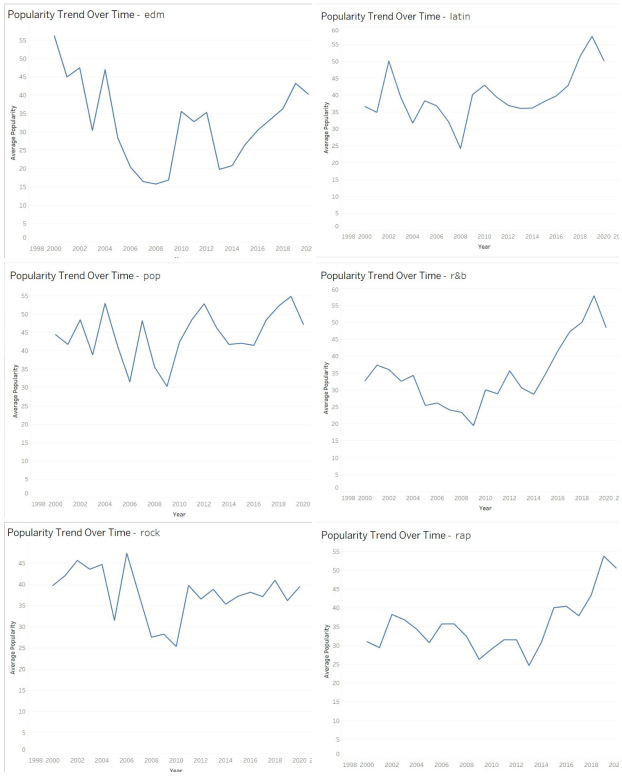

Fig 2.3: Average Popularity of Songs Released per Year (2000-2020) - Line charts illustrating the rise and fall in average popularity across six major genres

It is clear from the line charts that 2019 was a pivotal year across nearly every genre, with notable peaks in popularity for EDM, Latin, Pop, and Rap. This could possibly be because of the global rise of streaming platforms and increased access to music across different regions and genres. These line charts reveal the following trends about each of the genres,

- Pop: Strong Growth and Consistently High Popularity
- Rap: Steady Rise with a Popularity Peak in 2019
- R&B: Gradual Popularity Increase Over Time
- Latin: Rapid Popularity Surge in Recent Years
- Rock: Moderate Popularity with Consistent Releases
- EDM: Growth in Track Count but Lower Popularity

By examining the top tracks across different genres in 2019, we can see which specific songs drove this surge in music production and shaped the landscape of popular music. Treemaps were chosen to visualize the top songs (filtered by average track popularity scores greater than 90) for each genre, with average track popularity driving both the size and colour of the tiles.

The color scheme moves from light to dark tones within each treemap, where darker shades represent higher track popularity. This allows the viewer to quickly spot the most popular tracks within each genre.

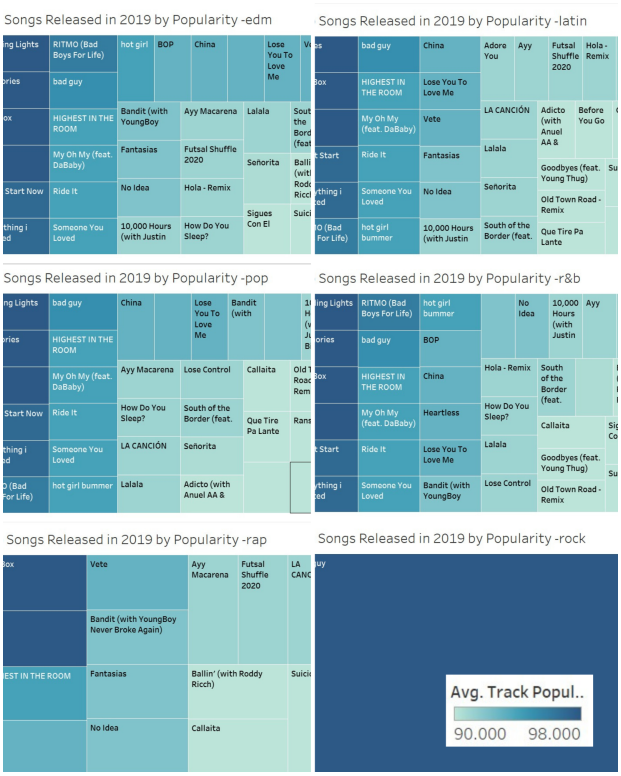

Fig 2.4: Treemaps of Top Songs in 2019 by Popularity

- One interesting observation is the recurrence of the same songs across multiple genres. This happens because the genre field used in this analysis is based on the playlist genre, meaning that a single track can appear on different playlists labeled under various genres.
- It is evident from the treemaps that tracks like “The Box,” “Blinding Lights,” “Tusa,” and “Bad Guy” were popular across multiple genres.
- This indicates that these tracks transcended specific genres and were popular enough to be included in multiple playlists with different genre tags. The cross-genre success of these tracks reflects their broad appeal, appealing to a diverse audience regardless of genre boundaries.

The next visualizations in fig-2.5 and fig-2.6 give us insight into the broader picture of how different genres fare in terms of average track popularity. A Bar chart is used to show the average track popularity across six genres. A Box plot is also plotted to show the distribution of track popularity within each genre, alongside the average popularity, which helps us in better understand the popularity dynamics at the genre level.

for the fig-2.5 Bars represent the average popularity of tracks within each genre. Color is used to distinguish the different genres.

for the fig-2.6 Boxes represent the range of popularity scores (from minimum to maximum) within each genre. Color is used to differentiate the genres.

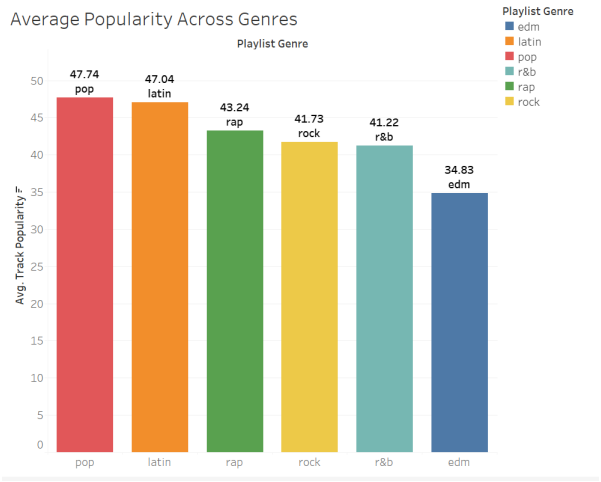

Fig 2.5: Bar Chart of Average Popularity Across Genres

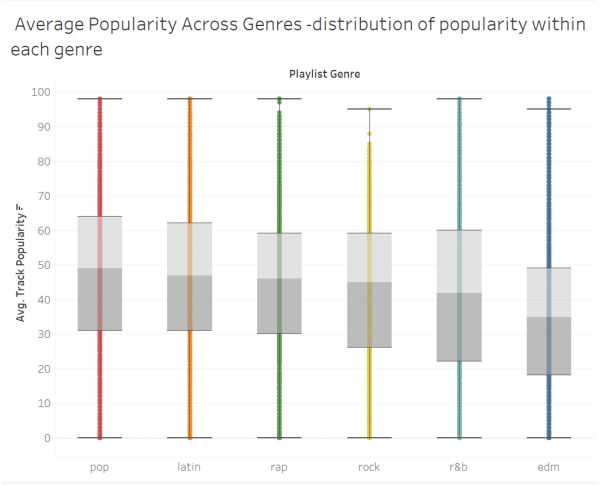

Fig 2.6: Box Plot Showing Popularity Distribution

- We can observe that in Rock, there is a significant gap between tracks with moderate popularity and those achieving higher popularity. The highest popularity track in Rock is "bad guy" with an average popularity of 95, but no other Rock songs surpass an average popularity of 90. (this was also seen in treemap of the fig-2.4) which suggests that while the genre remains consistent, it is not producing as many highly popular tracks as Pop or Latin.
- EDM's lower average popularity reflects that while the genre is prolific in terms of the number of songs released (it has the highest number of unique tracks as can be seen in fig-2.1), these tracks often cater to niche audiences rather than achieving widespread mainstream success.

The treemap visualization in fig-2.7 explores the relationship between genres and their subgenres in terms of average popularity and count of tracks. This gives us insight into which subgenres contribute the most to a genre's overall success.

Treemaps were selected to show the relative popularity of top songs within each genre. Treemaps allow for visualizing both track count (size of the tiles) and popularity (color intensity) at the same time.

Each tile represents a subgenre, with the size indicating the number of tracks in that subgenre and the color indicating the average popularity where darker shades indicate higher popularity. The colors range within a defined spectrum to maintain clarity

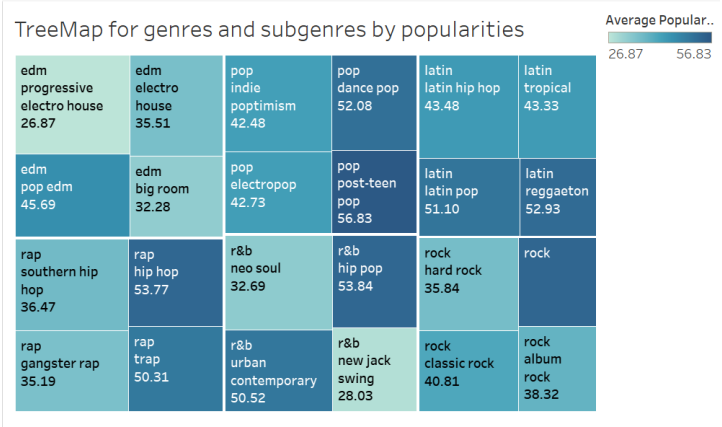

Fig 2.7: Treemap of Genres and Subgenres by Popularity

- The treemap represents the hierarchical structure of genres and subgenres in terms of both popularity and track count.
- This visualization builds on the insights from the bar chart and box plot of fig-2.5 and fig-2.6 by giving us a more detailed view of the internal diversity within each genre, helping us understand which subgenres are key to driving a genre's popularity and which ones contribute more to track volume without necessarily achieving mainstream success.
- Subgenres like Post-Teen Pop, Dance Pop, Reggaeton, and Latin Pop are the primary drivers of that high average popularity of pop and latin genres.
- Also the analysis about Rock and EDM seen in fig-2.5 and fig-2.6 applies here as well in their subgenres too.

This visualization focuses on the distribution of song popularity within albums, specifically albums containing more than 13 songs. The goal is to analyze how track popularity varies within these larger albums and identify trends in their average popularity.

A box plot seemed as an ideal choice to show the distribution. Boxes and Whiskers: Show the range of popularity within albums. The median, quartiles, and outliers can be clearly identified to assess if albums have standout songs or consistent popularity across all tracks.

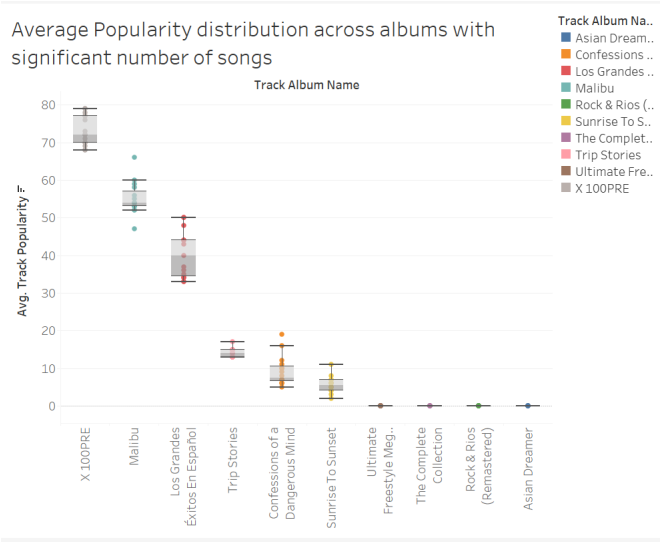

Fig 2.8: Box Plot of Album Popularity Distribution

- The box plots reveal significant variation in the distribution of song popularity across albums. Some albums exhibit a narrow range of song popularity, indicating that all the songs within the album have consistent popularity. Others show a wider spread, meaning that while some tracks are highly popular, others are less so.
- Interestingly, some albums had an average popularity of 0, meaning that all songs within these albums had 0 track popularity. This could be due to niche or less-streamed albums that didn't gain much traction.

VI. Task 3

The objective of this task was to analyze artist popularity measured by different metrics and observe how artist popularity evolved over time and across genres.

The top artists have been plotted using 2 different metrics. Figure 3.1 plots the top 20 artists measured using the artists' average track popularity.

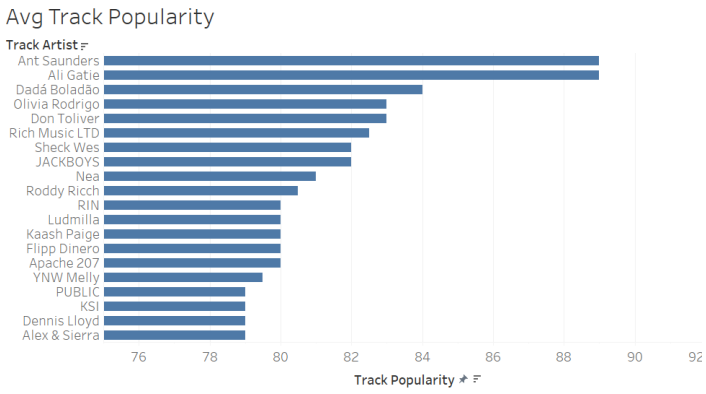

Fig 3.1: Bar Chart for top 20 artists based on average track popularity.

A calculated field named "Count Above Threshold" was introduced to quantify how many tracks an artist has with a popularity score greater than a specific threshold. This threshold was calculated using the 75^{th} percentile of song popularity. This helped identify artists who consistently release popular songs, rather than relying on just one or two hits. Figure 3.2 plots the top 20 artists with the highest count above threshold values.

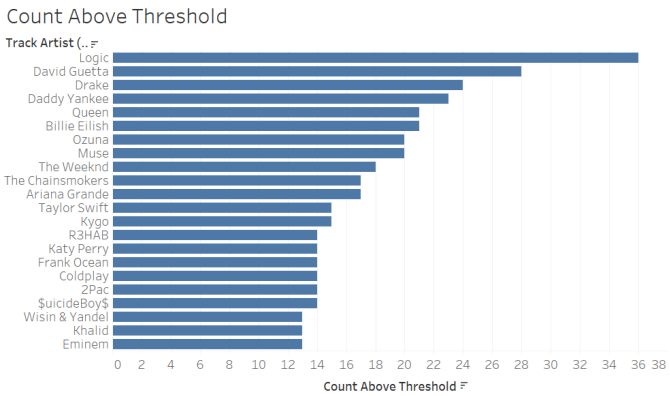

Fig 3.2: Bar Chart for top 20 artists based on count above threshold

Marks and channels used for 3.1 and 3.2 are Bars, grouped by artist and length of the bar shows popularity/count, with artists labeled on the Y-axis. Bar charts make it easy to compare the top artists based on the selected measure (average popularity, count above threshold)

To compare these methods for measuring artist popularity, two scatter plots were created.

- Count Above Threshold vs Track Count per Artist (assuming popularity is defined by count above threshold)
- Average Track Popularity vs Track Count per Artist (assuming popularity is defined by average track popularity)

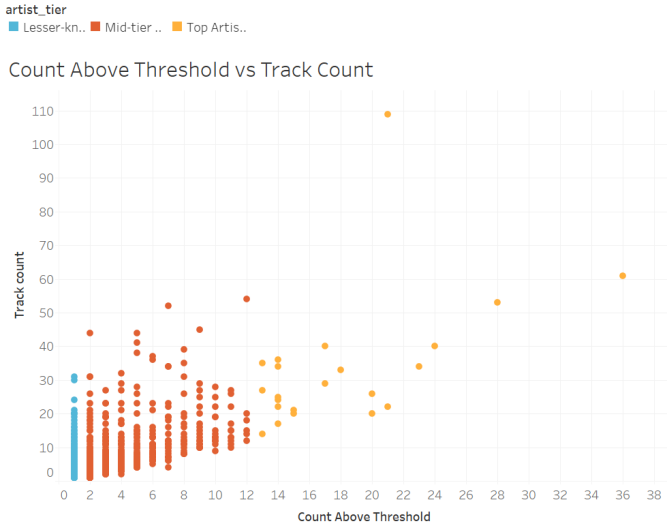

Fig 3.3: Scatter plot for count above threshold vs track count per artist

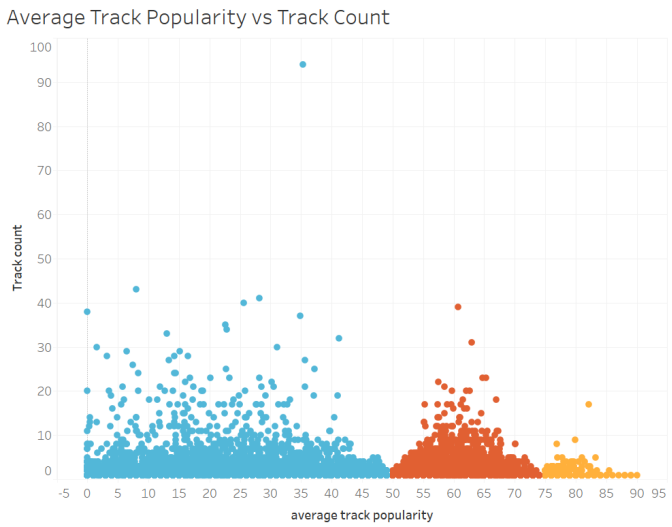

Fig 3.4: Scatter plot for average track popularity vs track count per artist

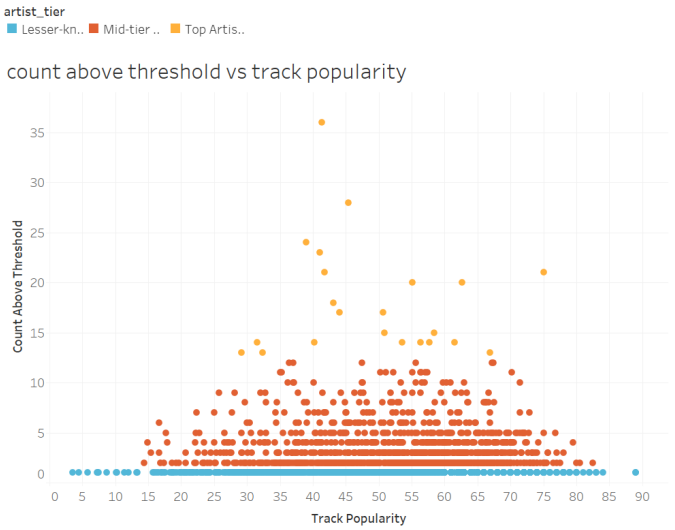

Fig 3.5: Scatter plot for relationship between the two different popularity measures

In each graph the artists have been split into lesser-know artists, mid-tier artists and top artists for the ease of visualization.

For count above threshold (Fig 3.3)

- count above threshold $\geq$ 13 - top tier artist
- count above threshold $\geq$ 2 - mid tier artist
- count above threshold $<$ 2 - lesser known artist

For average track popularity (Fig 3.4)

- average track popularity $\geq$ 75 - top tier artist
- average track popularity $\geq$ 50 - mid tier artist
- average track popularity $<$ 50 - lesser known artist

In figure 3.4 it is evident that the track count of top artists is low. This indicates that average track popularity can be misleading, as artists with a low track count but high average popularity may have one or two hit songs skewing their popularity. Count above threshold as a metric of popularity highlights artists who consistently release popular tracks, giving a more reliable picture of long-term popularity.

A plot displaying the relationship between two different popularity measures is shown in Fig 3.5 below.

As observed in the plot, the average track popularity of the most popular artists with a huge count above the threshold has median around 50. This suggests that popular artists have many tracks that are popular but not all of them stand out exceptionally. This could mean that artists' popularity is more consistent rather than being driven by the popularity of a few outstanding hits. Artists like Billie Eilish produce a large number of popular tracks and the average popularity of her tracks are also high; indicating a strong and sustained popularity in the industry.

Therefore, for the rest of the visualizations, we will use count above threshold as a metric for artist popularity.

Marks and channels used for 3.3, 3.4, 3.5 are Points, grouped by artist and position on the X and Y axes for the artist's popularity/track count, grouped by tiers. Each tier is represented by a colour. Scatter plots clearly show trends and outliers, highlighting relationships between popularity measures and track count.

To understand how the top artists gained popularity, we explored whether these top artists concentrated on a single genre or diversified their music across multiple genres.

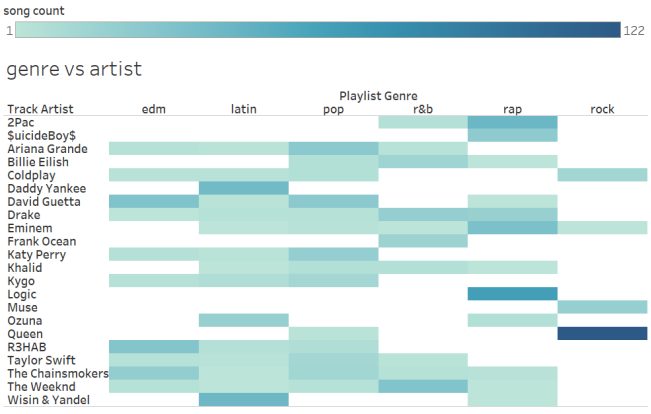

Fig 3.6: Heatmap for songs released genre-wise by top artists

From the heatmap in Fig 3.6, we observe that there are two types of artists among the top artists:

- **Artists focusing on one genre:** If an artist has a high concentration of songs in a single genre, it suggests they have built their popularity through specialization, catering to a specific audience or mastering a particular style.
- **Artists spreading across multiple genres:** On the other hand, artists with songs in multiple genres might have gained popularity through versatility, attracting a wider range of listeners by branching into different music styles.

The heatmap also indicates that most of the artists have gained popularity by spreading across multiple genres rather than focusing on a single genre.

Marks and channels used for 3.6 are Colour gradient, grouped by artist and genre, darker colour intensity represents more songs. Heatmaps effectively show song distribution across genres, emphasizing concentration of releases by artists. The color intensity makes it easy to identify which artists have a larger presence in specific genres.

The following treemaps (Fig 3.6.1 and Fig 3.6.2) visualize how the top songs are split amongst artists in a given genre. The size of each block represents the number of top songs released by the artist in that genre.

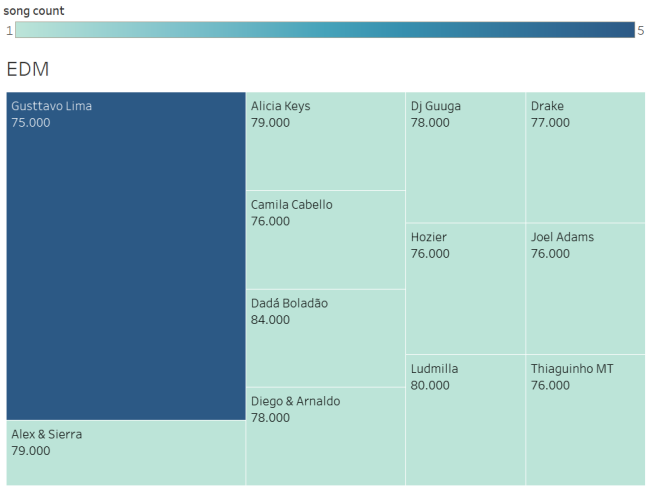

Fig 3.6.1: Treemap for top songs in EDM

From figure 3.6.1, the genre EDM is observed to be dominated by one artist.

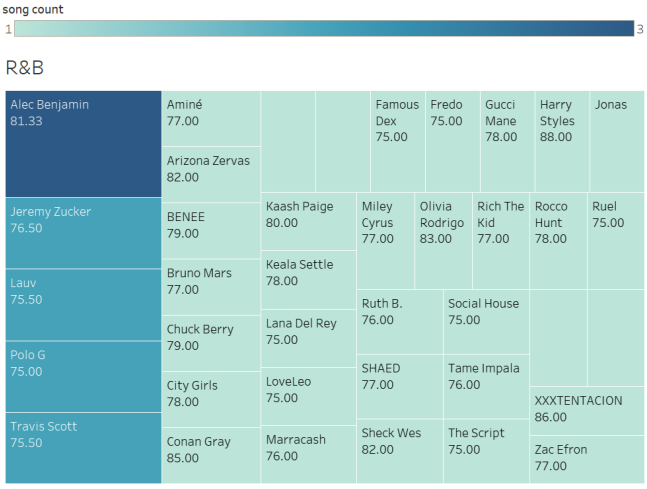

Fig 3.6.2: Treemap for top songs in R&B

Figure 3.6.2 indicates that the popular songs in the R&B genre come from many different artists.

Marks and channels used for 3.6.1 and 3.6.2 are coloured blocks representing artists' song counts and colour intensity reflects the number of top songs by an artist in a genre. Treemaps give an overview of artists' dominance within genres.

There isn't a single artist dominating the genre, meaning that R&B listeners may be drawn to a variety of artists, each contributing hits. This indicates a more competitive genre where multiple artists can achieve success, and popularity may be tied more to individual songs rather than artist dominance. In contrast, EDM seems to be dominated by a few artists, which suggests a concentration of popularity in EDM.

Some artists with top songs in specific genres aren't part of the top 20 artists overall suggests that genre-specific success doesn't always translate into overall popularity for the artist. An artist might be highly popular within a niche genre but may not make it into the general top artists list due to a smaller audience size of that genre or less appeal across genres.

Apart from genres, we tried to find a correlation between the features of songs by the top artists and their popularity.

To explore this, we visualized common musical features—acousticness, danceability, and loudness—across the top artists. This provided insight into whether certain features contribute to an artist's popularity, offering another dimension to understand their success beyond just the genre.

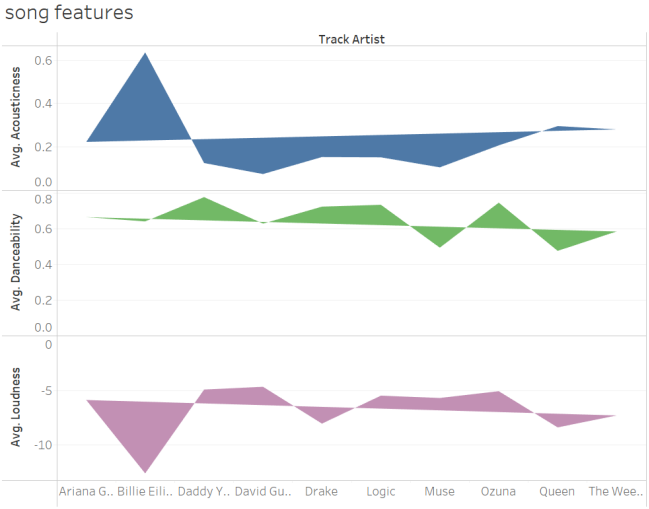

Fig 3.7: Plot for average musical features for top artists

Marks and channels used for 3.7 are Polygons, with vertices representing feature averages (acousticness, danceability, loudness) and different colours used for different features. Polygons allow easy comparison of multiple musical features for different artists Most of the top artists exhibited similar values for these features, suggesting that they tend to follow certain patterns or trends in music production that are widely appealing.

However, one notable outlier was Billie Eilish, whose musical feature profile differed from the rest. Her tracks exhibited a distinctive combination of extremely low loudness and high acousticness compared to her peers. This reinforces the idea that while many artists share common musical characteristics that define popular music trends, some, like Billie Eilish, stand out by breaking away from these patterns, potentially making their music more distinctive and appealing to listeners.

The line chart below (Fig 3.8) tracks the popularity of artists over time, measuring their popularity year by year. The average track popularity of the songs released by the artists in a specific year is plotted against the years.

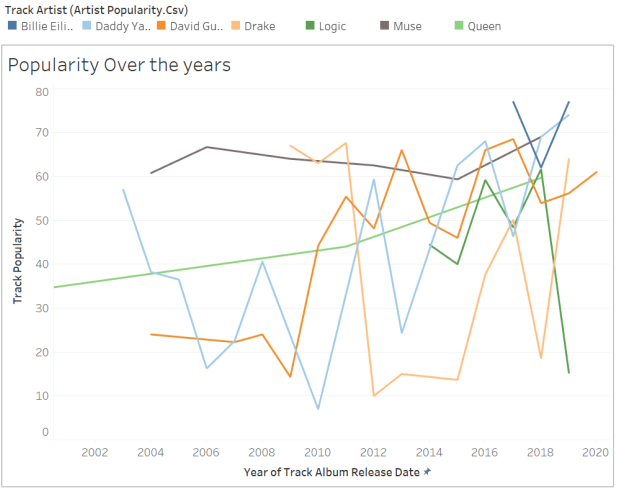

Fig 3.8: Line chart of artist popularity over time

Line charts are ideal for visualizing changes over time, helping us see trends in how artists maintain or lose popularity. By focusing on a few top artists, this chart avoids overcrowding and clearly shows shifts over years.

Artists experience spikes in popularity in certain years, tied to a successful album release or hit single. Some artists maintain steady popularity over time (such as the artists Queen and Muse in Fig 3.8), while others experience brief periods of popularity tied to a specific release.

The plot below examines the album release dates correspond with the sum of Count Above Threshold, grouped by artist tier (mid-tier, popular, not-popular).

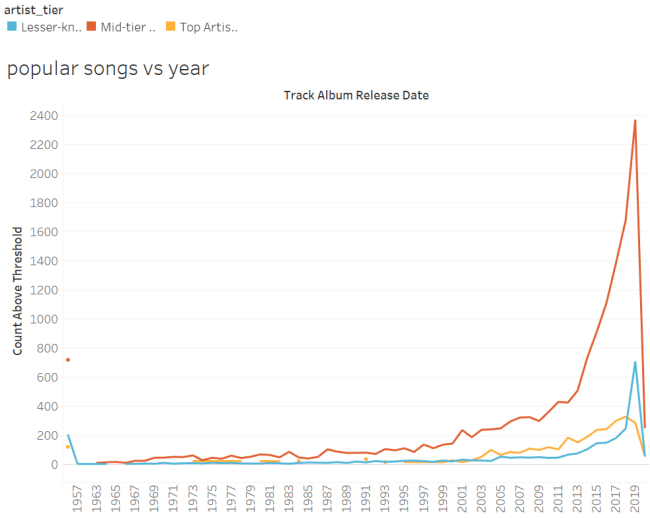

Fig 3.9: Popular songs released year-wise, grouped by artist popularity

The graph indicates that the most number of popular songs come from large number of mid-tier artists. A large peak in was viewed in 2019. This indicates that 2019 was a golden year for music, especially the mid-tier artists, where they collectively released many popular tracks.

Marks and channels used for 3.8 and 3.9 are lines, with points representing artist popularity over time or the number of popular songs per year. Colour is used in 3.9 to differentiate between different artist tiers. Line charts are best for showing trends over time

VII. Authors' Contributions

Data preparation and cleanup was done by all the team members.

- Saniya Ismail Kondkar: Task 1
- Ragini Metlapalli: Task 2
- Dyuthi Vivek: Task 3

VIII. Conclusion

In summary, the analysis of song features, artist popularity, and genre trends reveals a complex interplay of factors driving modern music success. While audio features like

energy, loudness, valence, and danceability influence track popularity, no single feature strongly correlates with it. Instead, deeper patterns emerge: since 2000, music has become less emotionally positive but more consistent in danceability, with genres like Latin showing high positivity, while EDM is more subdued. Energy and loudness are closely tied, particularly in EDM. Artist popularity, measured by count above threshold, provides a clearer picture of long-term success, emphasizing consistency across multiple tracks over average popularity. Notably, 2019 was a peak year for production and popularity across genres. However, external factors like social media, marketing, and cultural shifts also play significant roles in shaping an artist's or song's success in today's global, cross-genre music landscape.