Music recommendation system

Gayatri Yendamury, MTech (AI)(gayatriy@iisc.ac.in)
Bagi Shirisha, MTech (AI), bagishirisha@iisc.ac.in
Priyabrata Samantaray, MTech (ECE), priyabratas@iisc.ac.in
Vineet Kumar Tripathi, MTech (ECE), vineetkt@iisc.ac.in

3 December 2024

1 Problem Statement

Music streaming platforms offer millions of tracks, but users often struggle to find songs that match their choice. Existing recommendation systems are limited, often suggesting irrelevant content, leading to user disappointment and disengagement.Providing accurate song recommendations is crucial for keeping users engaged with the platform. Better recommendations mean happier users, who are more likely to listen music longer ,explore more and remain loyal to the service. This, in turn, increases user retention and revenue for the platform.

A smarter, data-driven recommendation system is needed to enhance personalization and improve user experience by offering more accurate music suggestions.

2 Dataset description

2.1 Dataset

We have used the public Spotify dataset available in Kaggle (`spotify Dataset`).

The datasets used are `data.csv` and `data_by_genre.csv`.

Dataset columns description :
The dataset has below mentioned feature columns:

- **valence**: Indicates the musical positiveness of a track, where high values represent happier and more positive sounds, and low values indicate sadder or negative tones.
- **year**: The release year of the track.It is useful for analyzing trends over time.
- **acousticness**: Measures how acoustic a track sounds. High values suggest the track is more acoustic, while low values imply more electronic or synthetic sounds.
- **artists**: Names of the artists involved in the track.It helps to group songs by same artist.
- **danceability**: Reflects how suitable a track is for dancing. High values denote easier dance rhythms, while low values indicate more complex or irregular beats.
- **duration_ms**: Duration of the track in milliseconds.The longer durations values are associated with extended compositions or live recordings.
- **energy**: It Represents the intensity and activity level of a track. Higher values suggest energetic tracks, while lower values indicate calm or mellow tracks.
- **explicit**: Indicates if the track has explicit lyrics (1 : yes, 0 : no).
- **id**: Unique Spotify ID for the track.
- **instrumentalness**: Predicts the likelihood of a track being instrumental. High values (close to

1.0) suggest little or no vocals, while low values indicate vocal presence.

- **key**: The musical key of the track, represented by numbers (0 to 11).
- **liveness**: It estimates the presence of a live audience. High values suggest live performance , while low values indicate studio recordings.
- **loudness**: Overall loudness of the track in decibels. Higher values mean louder tracks, while lower values are quieter.
- **mode**: Indicates modality (1 = major, 0 = minor); major often sounds happier, while minor is generally more somber.
- **name**: The title of the track.
- **popularity**: Popularity score on Spotify (0 to 100). High values indicate popular tracks, while low values indicate less popular tracks.
- **release_date**: The date the track was released.It helps to analyze time-based trends.
- **speechiness**: Measures the presence of spoken words in the track. High values indicate more speech-like content (e.g., podcasts), while low values suggest minimal or no speech.
- **tempo**: Tempo of the track in beats per minute (BPM). High values indicate faster tracks, while low values indicate slower songs.

This dataset provides a range of audio features that describe the musical characteristics and popularity metrics for each track, essential for building a recommendation system.

3 Data Collection and Pre-Processing

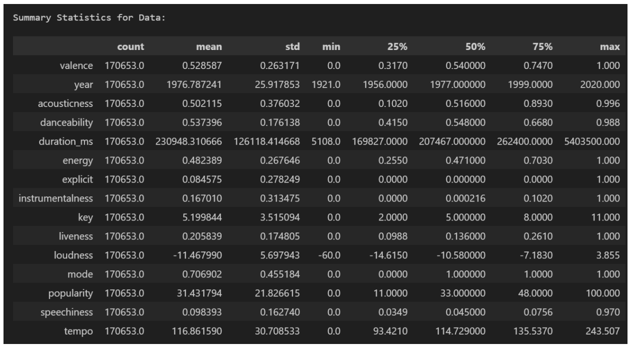
The dataset was read into a pandas dataframe for further analysis.

3.1 Assessing Missing Data and checking for duplicate values

The dataset is well prepared .There were no missing values found in any of the columns.Also , no duplicate rows were present.

3.2 Descriptive statistics of features

Descriptive statistics (like mean, standard deviation, minimum, maximum, and quartiles) provide insights into data distributions.It helps in identifying central tendency, spread, and outliers across multiple datasets in one go. Shown below is a snippet of output statistics using Pandas library command describe.



Summary Statistics for Data:

	count	mean	std	min	25%	50%	75%	max
valence	170653.0	0.528587	0.263171	0.0	0.3170	0.540000	0.7470	1.000
year	170653.0	1976.787241	25.917853	1921.0	1956.0000	1977.000000	1999.0000	2020.000
acousticness	170653.0	0.502115	0.376032	0.0	0.1020	0.516000	0.8930	0.996
danceability	170653.0	0.537396	0.176138	0.0	0.4150	0.548000	0.6680	0.988
duration_ms	170653.0	230948.310666	126118.414668	5108.0	169827.0000	207467.000000	262400.0000	5403500.000
energy	170653.0	0.482389	0.267646	0.0	0.2550	0.471000	0.7030	1.000
explicit	170653.0	0.084575	0.278249	0.0	0.0000	0.000000	0.0000	1.000
instrumentalness	170653.0	0.167010	0.313475	0.0	0.0000	0.000216	0.1020	1.000
key	170653.0	5.199844	3.515094	0.0	2.0000	5.000000	8.0000	11.000
liveness	170653.0	0.205839	0.174805	0.0	0.0988	0.136000	0.2610	1.000
loudness	170653.0	-11.467990	5.697943	-60.0	-14.6150	-10.580000	-7.1830	3.855
mode	170653.0	0.706902	0.455184	0.0	0.0000	1.000000	1.0000	1.000
popularity	170653.0	31.431794	21.826615	0.0	11.0000	33.000000	48.0000	100.000
speechiness	170653.0	0.098393	0.162740	0.0	0.0349	0.045000	0.0756	0.970
tempo	170653.0	116.861590	30.708533	0.0	93.4210	114.729000	135.5370	243.507

Figure 1: descriptive statistics for numerical feature columns of dataset

Below are some observations from descriptive statistics:

1. **Balanced Features:** Balanced Features: Valence, Acousticness, Danceability, and Energy show a diverse range of musical characteristics, suggesting varied tracks, while Instrumentalness and Speechiness indicate a vocal-dominant dataset, with most tracks being music-focused rather than speech-heavy.
2. **Popularity:** With 75% of the tracks having a popularity score below 48, there is a notable scarcity of highly popular tracks. This highlights an opportunity for the recommendation system

to explore and promote lesser-known but potentially valuable tracks.

3. **Year :** We have data samples from 1921 to 2020 in this dataset.

4. **loudness :** The loudness feature ranges from -60.0 to 3.855, encompassing both positive and negative values. To facilitate better analysis and ensure that all features are on a similar scale, it would be beneficial to normalize this feature.

3.3 Visualizing Outliers

Outliers are data samples that falls on the far left or right side of the ordered data. Generally, the outliers fall more than the specified distance from the first and third quartile (IQR: Interquartile Range i.e. outliers are greater than Q3 + (1.5 * IQR) or less than Q1 - (1.5 * IQR).

Shown below are the boxplots for different numerical features of the dataset(figure 2):

Below are the conclusions drawn from the boxplot analysis:

- **duration_ms and tempo**: Songs with very high durations (like live performances) or unique tempos (extremely slow or fast) are often valid data points and may be retained depending on analysis needs.

- **instrumentalness, liveness and speechiness** : It is expected that the dataset will show a greater concentration of values near 0 for instrumentalness, liveness, and speechiness. This aligns with the intention to recommend tracks that are primarily music-focused and feature significant vocal elements.

- **loudness**: Extremely low loudness values may represent specific song genres or recording qualities and may need further investigation to decide if they should be capped or retained.

- **explicit** : This is a categorical feature and it has only few data samples with explicit feature value as 1.However, it might help in personalised song recommendations.Hence, its better to retain this feature outliers.

- **danceability** : There are 143 outliers in Data and 11 outliers in Genre Data, with low danceability values.These outliers can be capped.

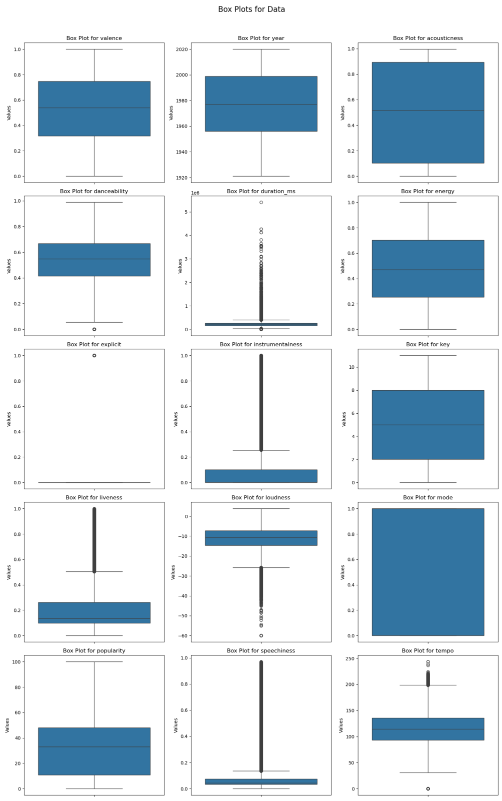


Figure 2: boxplot of numerical features

3.4 Outlier Handling

For outlier handling , Outlier Capping is chosen for its simplicity, effectiveness, and ability to retain data integrity. This method minimizes the impact of extreme values while preserving the dataset for accurate, interpretable analysis.

The column danceability had 143 outliers in Data and 11 outliers in Genre Data. After outlier capping , the column danceability has no outliers as plotted in below boxplot(figure 3). All other features are retained with outliers , as they have too many outliers

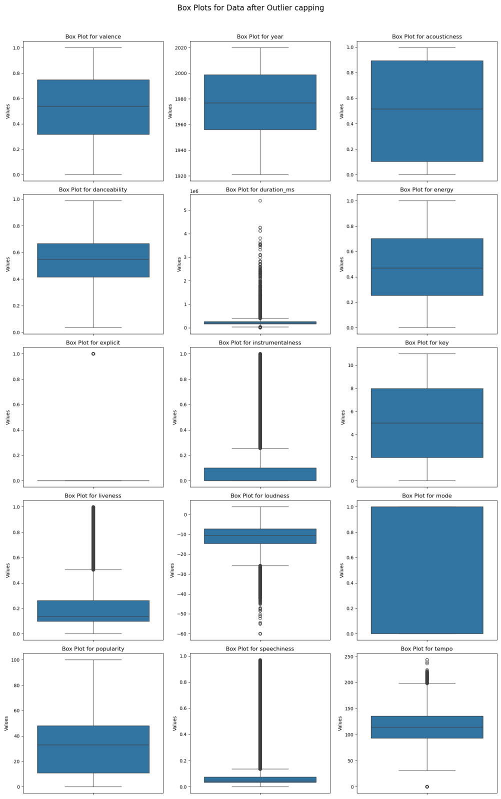


Figure 3: boxplot after outlier capping of numerical features

and they can provide important information for music recommendation system.

3.5 Feature Engineering

To enhance interpretability and enable more effective analysis, we created new categorical features and transformed existing ones. These engineered features will make it easier to identify trends and improve compatibility with models that benefit from normalized or categorical data.

3.5.1 Add new music feature categorical columns

Columns valence_category, acousticness_category, danceability_category, energy_category, liveness_category, instrumentalness_category, speechiness_category, tempo_category,loudness_category, popularity_category, key_category, duration_ms_category have been added to provide a clearer understanding of the numerical values, based on their relative positions within the overall range (e.g. less than 25%, 50%, 75%, or 100%).

3.5.2 Convert or Normalize existing columns

- To address the negative values in the 'loudness' column, we created a new column, 'loudness_scaled', which has been normalized to a range of 0 to 1.

- The column 'duration_min' is added after converting 'duration_ms' to minutes and duration_ms is discarded.

- The column 'release_decade' is added after converting 'year' to corresponding decade.

- Category Columns 'mode_category', 'explicit_category' are added which helps in mapping numerical values i.e. 0 and 1 to the corresponding meaning.This will help in further data analysis.

3.5.3 Removing unecessary columns

Column 'release_date' is inconsistent and thus discarded as the release year information is already present in 'year' column.

3.5.4 Adding columns based on artists

New columns 'artist_count','artist_category' are added to the dataframe.This will help in data analysis based on artist count. This might also support in music recommendation by capturing user preferences for solo versus group performances.

Finally , the pre-processed dataset is stored as `data_cleaned.csv` and `genre_data_cleaned.csv`

4 Data Exploration

4.1 Observations based on popularity

4.1.1 Least popular songs and the similarity in their features

The least popular songs (having popularity score between 0 to 10) have common features like very low energy,low danceability , low valence- emotionally negative or neutrals, low tempo , low key-typically quieter tonal center,moderate live presence,mostly instrumental and from release decade 1940's.

4.1.2 Most popular songs and the similarity in their features

The song 'Dakiti' is the most popular song with popularity score of 100.It has high energy , neutral emotions and high danceability. The most popular songs (having popularity score between 90 to 100) have common features like moderate energy,high danceability , low valence- emotionally negative or neutrals, high tempo,very low live presence,no instrumental elements, primarily vocal driven and from release decade 2020's.

4.1.3 Duration of the Songs in most popular Genres

The genre 'south african house' has highest song duration and 'alberta hip hop' has lowest song duration , as shown in below bar graph.

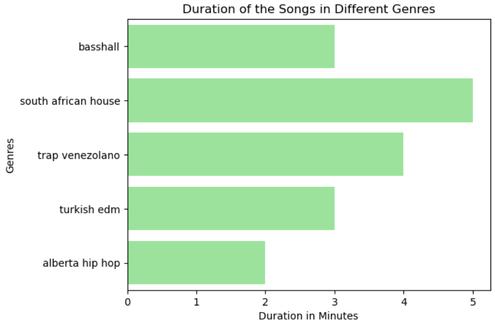


Figure 4: song duration across multiple genres

4.2 Univariate Analysis

4.2.1 Distribution of Numerical Features

The feature distributions give insights into the characteristics of the dataset.Shown below are the distribution of numerical features of the dataset.

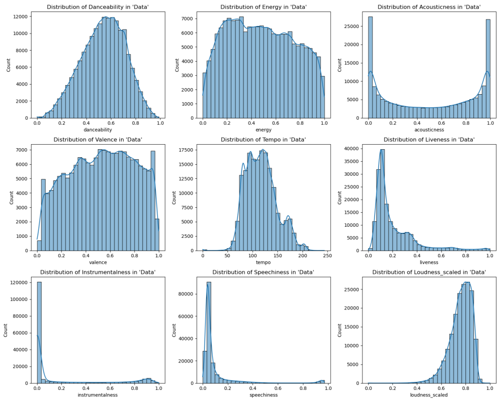


Figure 5: distribution of numerical features

From above observations we can say that the dataset represent a variety of musical genres with an emphasis on moderately energetic, moderate tempo , mostly high loudness , studio-produced tracks with some vocal content .

4.2.2 Distribution of Categorical Features

We have plotted the distribution of categorial features of newly added columns such as valence category, acousticness category, danceability category, energy category, etc.. in order to understand the spread of the dataset with respect to each music feature. The series of bar charts below (figure 6) provides a breakdown of various categorized features in the dataset, displaying the distribution of songs across different features.

These distributions suggest that the dataset is diverse across multiple musical features, allowing for a broad range of recommendation possibilities. The dataset is slightly skewed towards recommending songs which are vocal and studio recorded songs , and with no explicit content and high mode. However, this aligns with the objective of music recommendation which focuses more on recommending studio recorded songs suitable for wide range of audience.

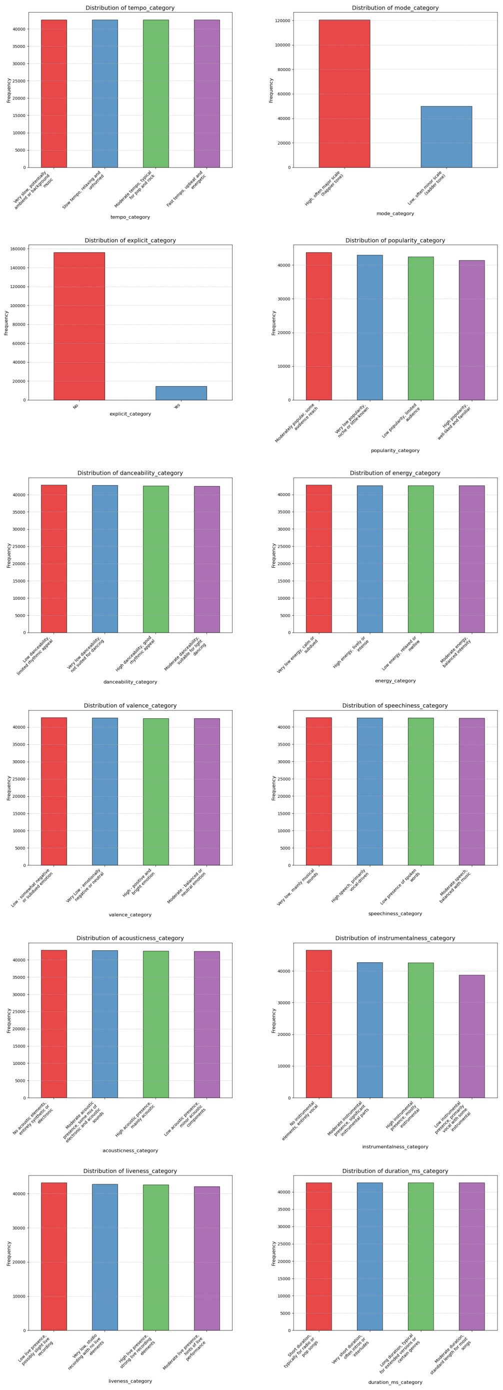


Figure 6: distribution of categorical features

Below shown pie chart shows the distribution of songs across two emotional categories(major and minor scale) based on their mode.

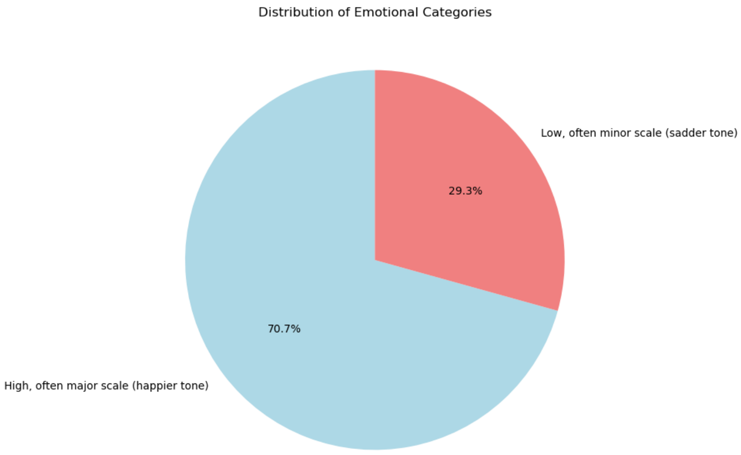


Figure 7: pie chart of emotional categories

From above pie chart , we can say that category 'High, often major scale(happier tone)' represents 70.7% of the songs. The major scale is typically associated with a happier or more uplifting mood, so a large portion of the dataset may be perceived as positive or cheerful.
The category 'Low, often minor scale (sadder tone)' represents 29.3% of the songs. The minor scale is often linked to a sadder or more melancholic tone, so this smaller portion of the dataset might have a more somber or introspective mood. This distribution indicates that the dataset is weighted towards songs with a “happier” emotional tone, which may reflect a preference for major key compositions in the dataset.

4.3 Bivariate Analysis - Numerical vs. Numerical Relationships

4.3.1 Correlation Analysis - Heatmap of Numerical Feature Correlations

The correlation heat map is generated for all numerical features as shown below (figure 8):

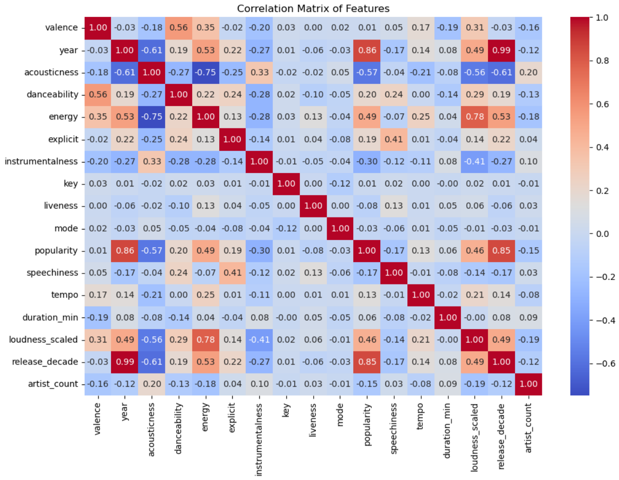


Figure 8: correlation heat map of all numerical features

We can see below observations from this correlation heatmap:

- **Valence and Danceability**: There is a moderate positive correlation between valence and danceability (0.56). This suggests that songs with a higher valence (happier or more positive mood) tend to also be more danceable. This makes sense, as upbeat, positive songs are often more suitable for dancing.
- **Year and Popularity**: There is a strong correlation between year and popularit (0.86) indicating that newer songs are much more likely to be popular in this dataset.
- **Energy and Loudness_scaled**: Energy and loudness have a strong positive correlation (0.78), which is expected as louder tracks often convey a higher energy level.
- **Acousticness and Energy**: Acousticness is negatively correlated with energy (-0.75), indicating that acoustic songs tend to have lower energy levels.
- **Acousticness and Year**: There is a moderate negative correlation between acousticness and year (-0.61). This suggests that newer songs tend to be less acoustic, possibly due to the increased use of electronic production techniques in recent music.

4.3.2 Heatmap - Audio Features by Popularity Category

This heatmap shows the average values of various audio features across different popularity categories as shown below(figure 9).

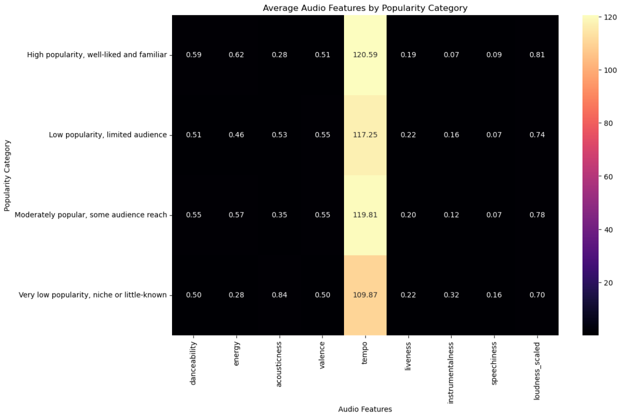


Figure 9: heat map of average audio features by popularity

We can see below observations from this heatmap:

- **Danceability and Energy:** : Popular tracks have higher danceability (0.59 vs. 0.50) and energy (0.62 vs. 0.28), indicating that upbeat and energetic songs are more likely to be popular.
- **Acousticness and Tempo:** : Popular songs are less acoustic (0.28 vs. 0.84) and tend to have a slightly faster tempo (120.59 BPM vs. 109.87 BPM), suggesting a preference for digitally enhanced, fast-paced tracks.

- **Loudness**: Higher loudness values (0.81 vs. 0.70) in popular songs indicate that dynamic and engaging tracks resonate better with listeners.

4.3.3 Violin Plot - Distribution of Popularity by Tempo Category

The violin plot displays the distribution of popularity across different tempo categories as shown below.

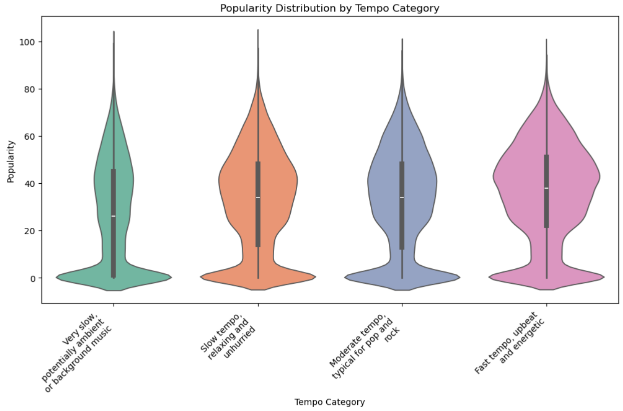


Figure 10: Violin plot of popularity across different tempo

Fast-tempo songs have a wide spread in popularity, with a slight concentration around moderate popularity levels. This shows that while fast, energetic music can reach high popularity, it also varies widely in its appeal. Overall, moderate and fast-tempo categories appear to align more with popular music, while slower tempos may have more specialized or selective appeal.

4.4 Trend analysis

Below shown graphs were plotted to observe music trend over the years.

- **Year vs Duration** : There was an increase in Duration in 1940s-1970s which reflects the influence of rock, album-oriented rock, and experimental genres that allowed for longer

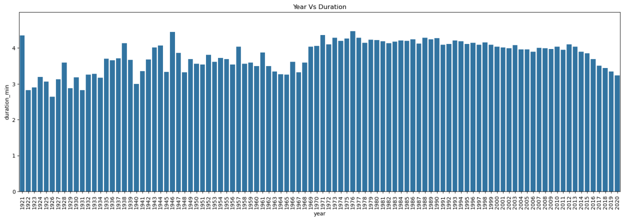


Figure 11: song duration across the years

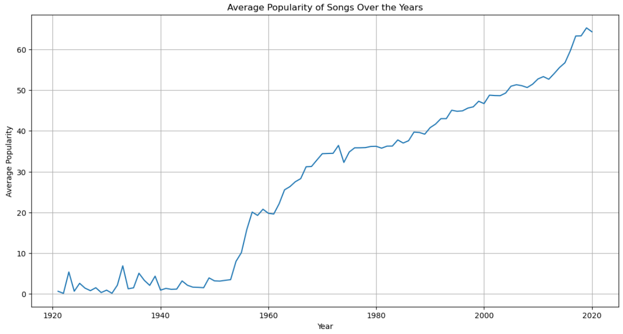


Figure 12: song popularity across the years

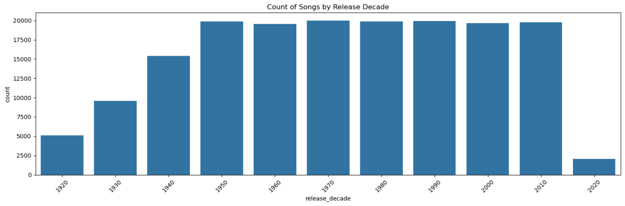


Figure 13: number of songs released across the years

tracks.Later there was a decrease in Duration 1990s-2020s likely tied to the shift towards pop and streaming, where shorter tracks are more commercially viable.

- **Average Popularity of Songs Over the Years** : Song popularity has seen steady growth over time, with significant boosts during periods of technological advancement in music distribution and accessibility.i.e. between 1980s to 2020s.

- **Count of Songs by Release Decade** : The music industry has sustained high levels of song production from the 1950s onwards, with significant growth in earlier years as recording technology and distribution developed.In 2020s there is a noticeable drop in song count, likely due to incomplete data for this decade or the limited number of years covered within it so far.

4.5 Analysis based on artists

4.5.1 Distribution of artist category

Below distribution graph of artists category was plotted (figure 14).We observe that the dataset contains highest number of data samples from Solo tracks followed by duet, small group and choir.This indicates that audience is more aligned towards solo tracks.

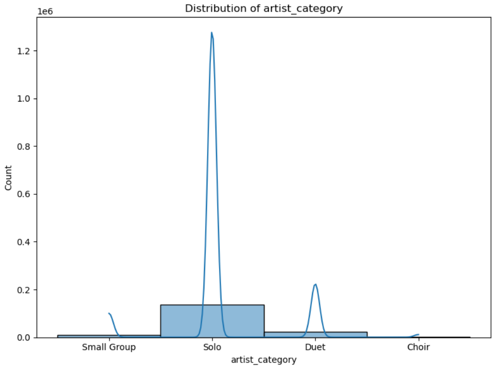


Figure 14: distribution of artist category

4.5.2 Average audio features by artist category

The average music features for different artist types was plotted as shown below(figure 15)

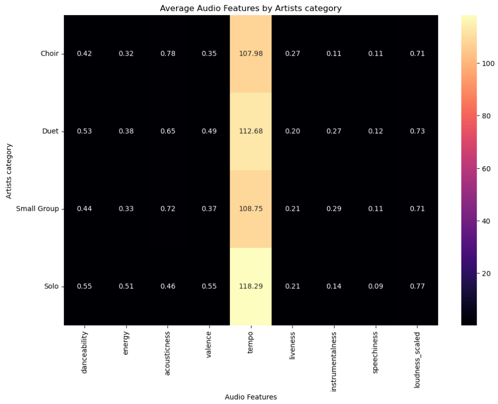


Figure 15: distribution of artist category

- **danceability and loudness_scaled** : The average danceability and loudness_scaled are almost the same across all artist categories.
- **energy and tempo** : The average energy and tempo are highest for solo tracks followed by duet.
- **acousticness and instrumentalness**:The average acousticness and instrumentalness is the highest in choir and small group.This is as expected as group songs usually have more acoustics present , unlike solo tracks.