DA 204o: **Data Science in Practice**

Course Project Proposal

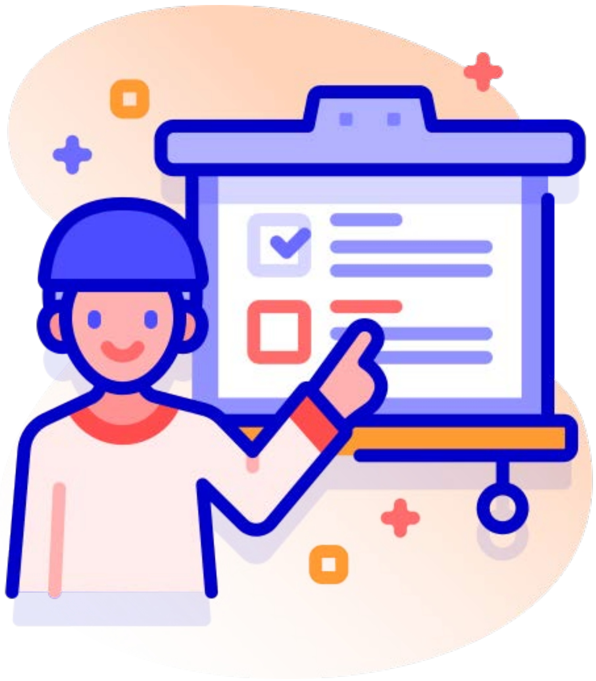
Personalized Music Recommendation System

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*

Problem Definition

- 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.

- A smarter, data-driven recommendation system is needed to :
  - enhance personalization,
  - improve user experience by offering more accurate music suggestions based,
  - keep users engaged with the platform,
  - increase user retention and revenue for the platform



Data Collections and Preparation

Reference: 01_data_collection.ipynb

➢ **Data source :** Spotify dataset (kaggle.com) : data.csv , data_by_genre.csv

Dataset columns description :

- **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.

Data.csv : (170,653 , 19)
Data_by_genre: (2973, 14)

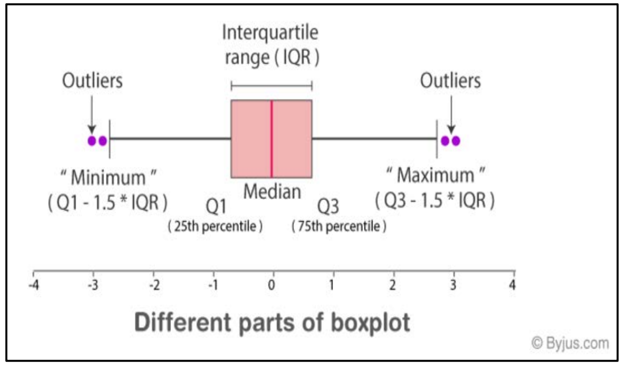
- **Features (Columns):** The dataset contains 19 features, including:
  - **Numerical:** valence, year, acousticness, danceability, duration_ms, energy, explicit, instrumentalness, key, liveness, loudness, mode, popularity, speechiness, tempo
  - **Categorical:** artists, id, name, release_date, genre
- **Size:** 170,653 records (rows).
- **Format:** CSV file with mixed data types (numerical, categorical, and text).

Data Collections and Preparation

Reference: 02_data_preprocessing.ipynb

➢ Data preprocessing steps

- **Check Data completeness:** No missing values and duplicate rows found .
- **Descriptive statistics:** Descriptive statistics performed.
- **Check Outliers:** Outliers visualized via Boxplots and outliers counted using interquartile range
- **Outlier Handling:** Outlier capping performed only on danceability as the other features had too many outliers and can provide useful information for recommendation system.
- **Feature engineering:** New music feature categorical columns added , ‘duration min’ , ‘release decade’ added . Column ‘release date’ is inconsistent and thus discarded. New columns 'artist_count’, 'artist_category' are added based on artist.




```
Outlier Summary for Data:
valence: 0 outliers
year: 0 outliers
acousticness: 0 outliers
danceability: 143 outliers
duration_ms: 9518 outliers
energy: 0 outliers
explicit: 14433 outliers
instrumentalness: 36105 outliers
key: 0 outliers
liveness: 11808 outliers
loudness: 3501 outliers
mode: 0 outliers
popularity: 0 outliers
speechiness: 23937 outliers
tempo: 1645 outliers
```

```
Outlier Summary for Data after Outlier capping:
valence: 0 outliers
year: 0 outliers
acousticness: 0 outliers
danceability: 0 outliers ✅
duration_ms: 9518 outliers
energy: 0 outliers
explicit: 14433 outliers
instrumentalness: 36105 outliers
key: 0 outliers
liveness: 11808 outliers
loudness: 3501 outliers
mode: 0 outliers
popularity: 0 outliers
speechiness: 23937 outliers
tempo: 1645 outliers
```

```
Outlier Summary for Genre Data:
mode: 496 outliers
acousticness: 0 outliers
danceability: 11 outliers
duration_ms: 153 outliers
energy: 0 outliers
instrumentalness: 117 outliers
liveness: 150 outliers
loudness: 198 outliers
speechiness: 265 outliers
tempo: 141 outliers
valence: 0 outliers
popularity: 196 outliers
key: 0 outliers
```

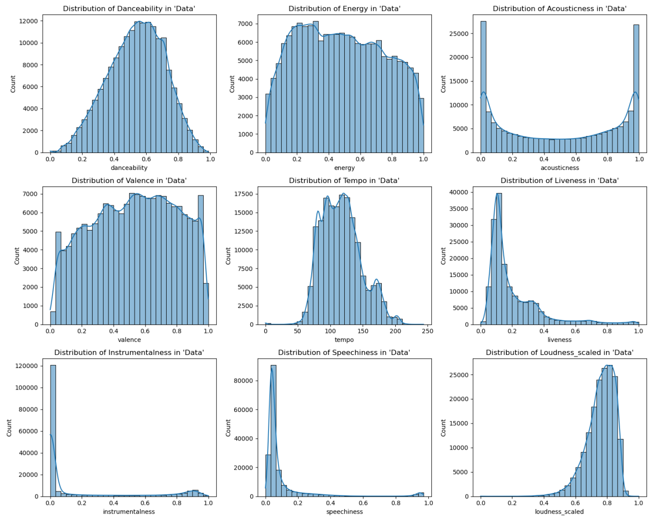
```
Outlier Summary for Genre Data after Outlier capping:
mode: 496 outliers
acousticness: 0 outliers
danceability: 0 outliers ✅
duration_ms: 153 outliers
energy: 0 outliers
instrumentalness: 117 outliers
liveness: 150 outliers
loudness: 198 outliers
speechiness: 265 outliers
tempo: 141 outliers
valence: 0 outliers
popularity: 196 outliers
key: 0 outliers
```

Data Exploration

Reference: 03_data_Exploration.ipynb

Common Features among the least popular tracks

```
Common features among the least popular tracks
{'year': 1940,
 'mode_category': 'High, often major scale (happier tone)',
 'danceability_category': 'Very low danceability, not suited for dancing',
 'energy_category': 'Very low energy, calm or subdued',
 'valence_category': 'Very Low - emotionally negative or neutral',
 'explicit_category': 'No',
 'release_decade': 1940,
 'tempo_category': 'Very slow, potentially ambient or background music',
 'key_category': 'Very low, typically quieter tonal center',
 'acousticness_category': 'High acoustic presence, mainly acoustic',
 'liveness_category': 'Moderate live presence, hints of live performance',
 'instrumentalness_category': 'High instrumental presence, mostly instrumental',
 'speechiness_category': 'High speech, primarily vocal-driven'}
```

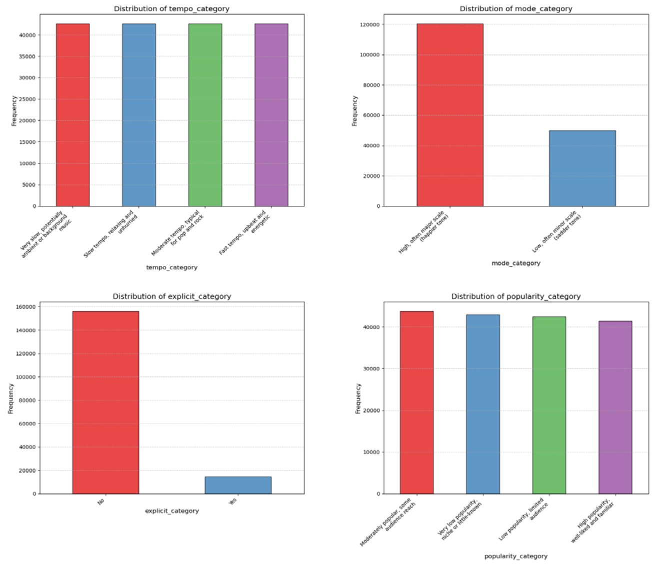


Distribution of numerical features

Common Features among the most popular tracks

```
Common features among the most popular tracks
{'year': 2020,
 'mode_category': 'Low, often minor scale (sadder tone)',
 'danceability_category': 'High danceability, good rhythmic appeal',
 'energy_category': 'Moderate energy, balanced intensity',
 'valence_category': 'Low - somewhat negative or subdued emotion',
 'explicit_category': 'No',
 'release_decade': 2020,
 'tempo_category': 'Fast tempo, upbeat and energetic',
 'key_category': 'Very low, typically quieter tonal center',
 'acousticness_category': 'Low acoustic presence, minor acoustic components',
 'liveness_category': 'Very low, studio recording with no live elements',
 'instrumentalness_category': 'No instrumental elements, entirely vocal',
 'speechiness_category': 'High speech, primarily vocal-driven'}
```

Distribution of categorical features

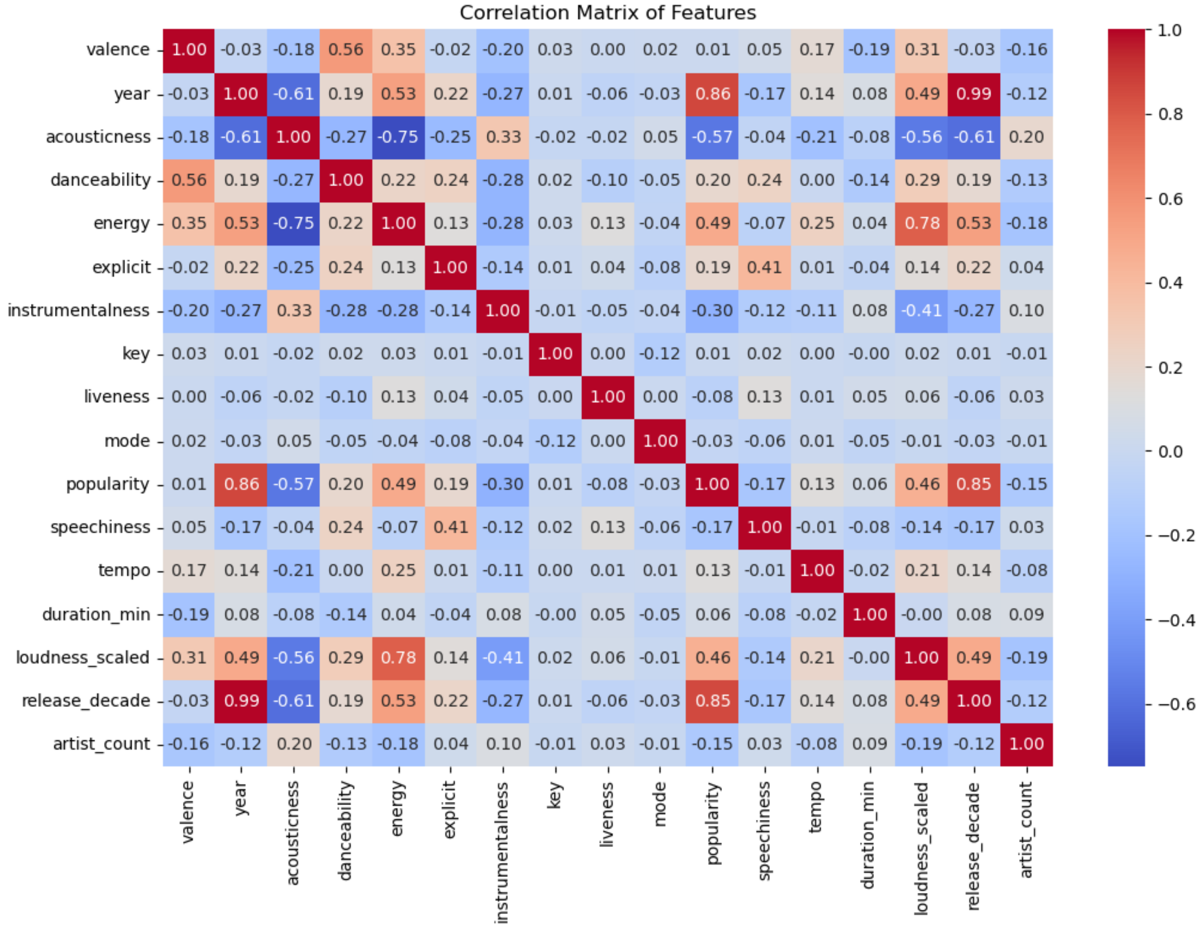


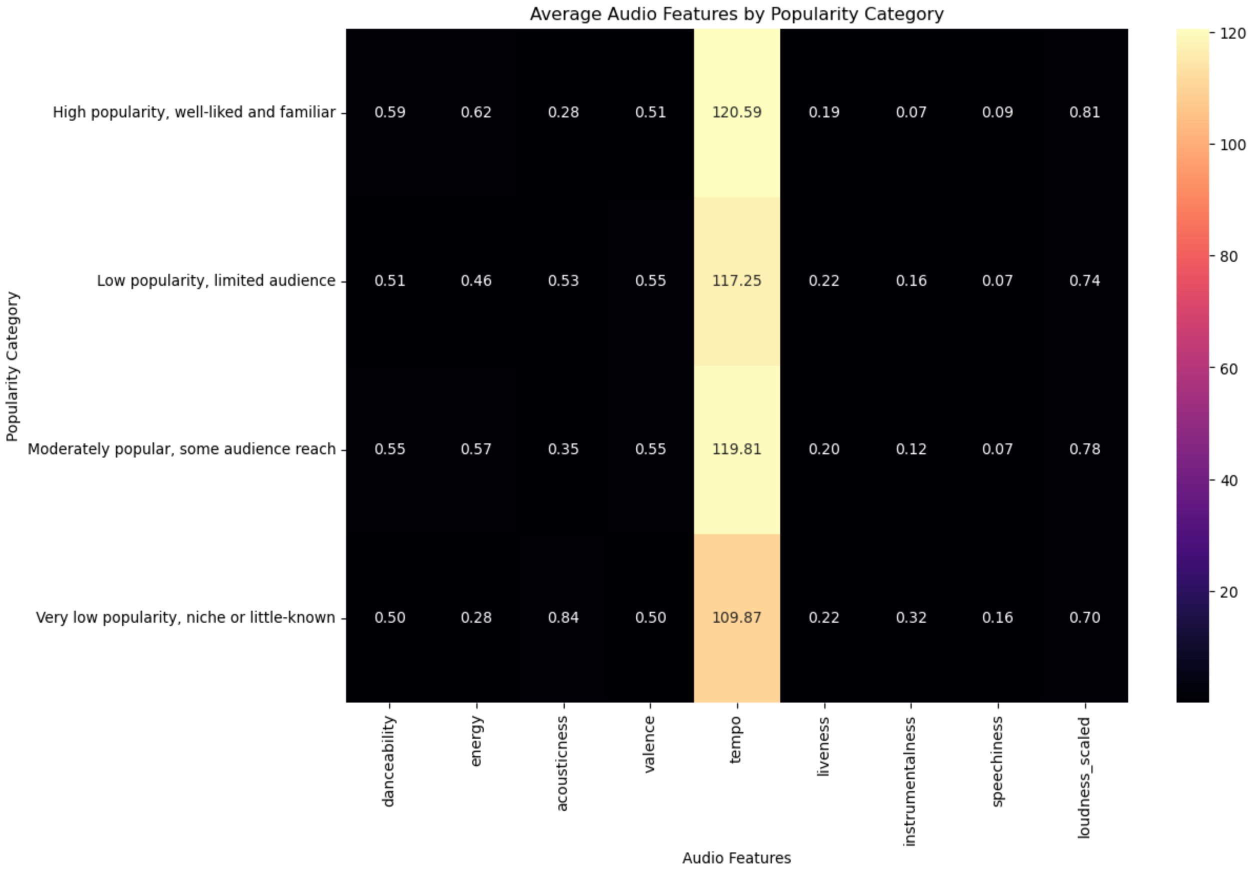
Data Exploration

Reference: 03_data_Exploration.ipynb

➢ Bivariate Analysis - Numerical vs. Numerical Relationships:

- **Correlation Analysis : Heatmap is observed for Numerical Feature Correlations as shown below. Strong correlations were found between Year and Popularity, Energy and Loudness.**

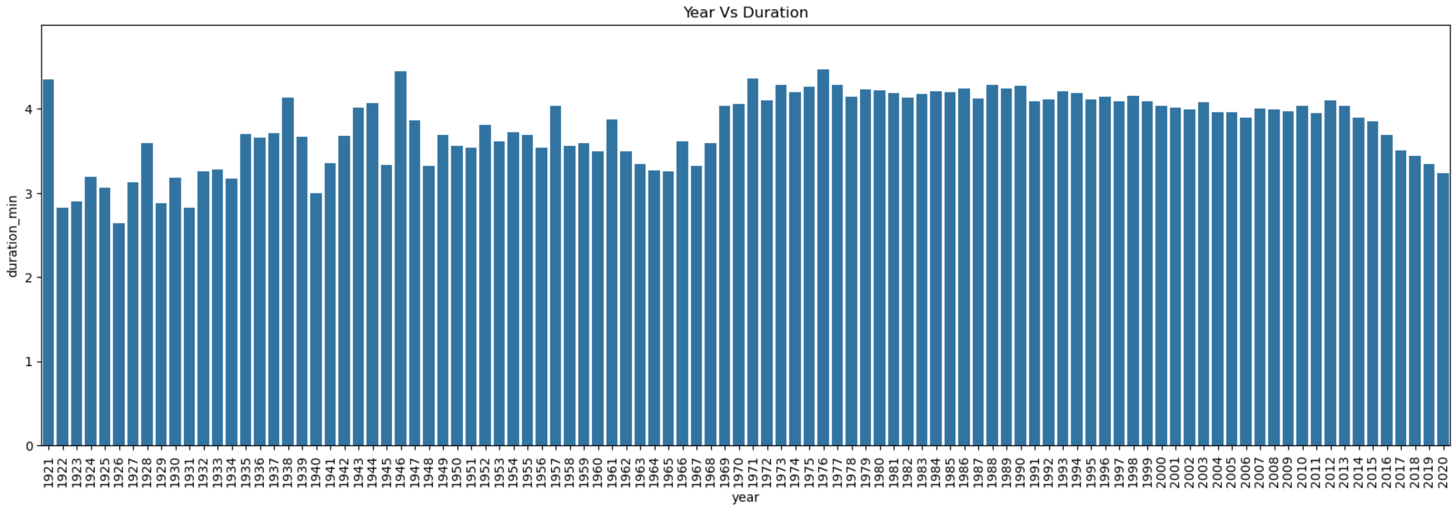


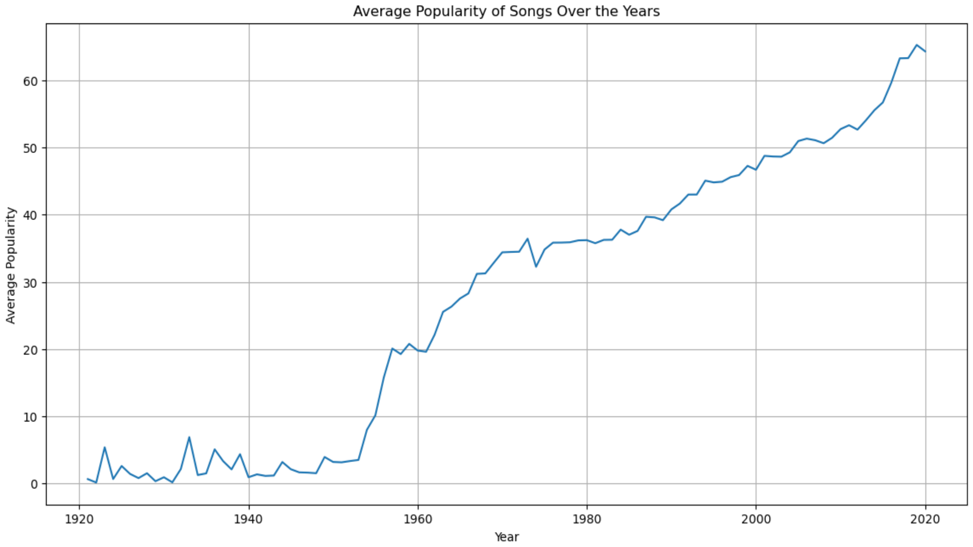


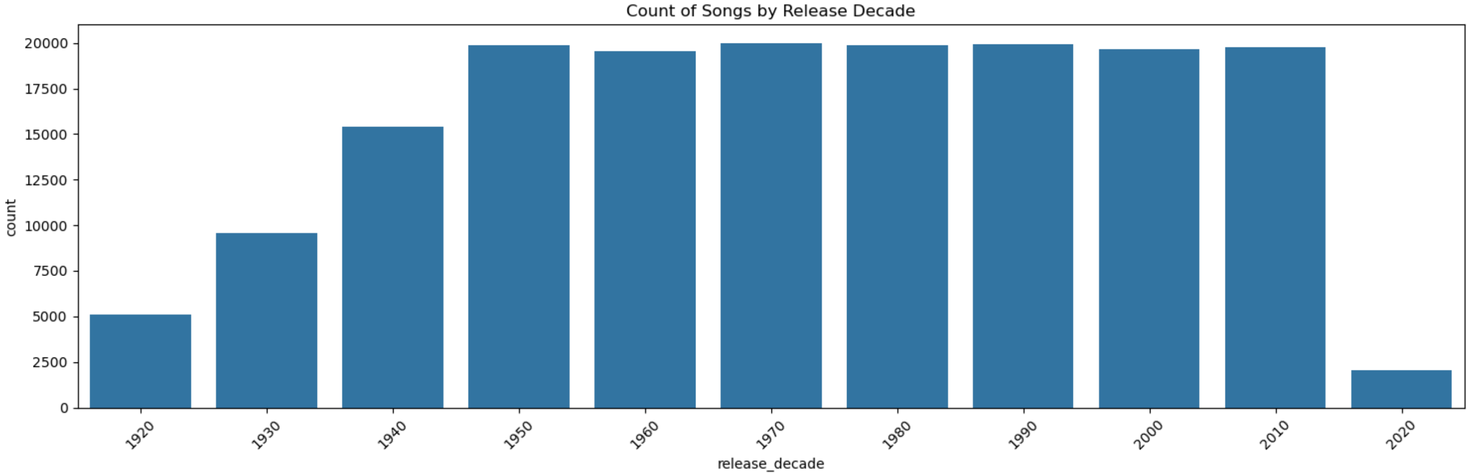
Data Exploration

Reference: 03_data_Exploration.ipynb

- **Trend analysis: Over the years , the trend of song duration , average popularity of songs , count of songs produced was observed.**



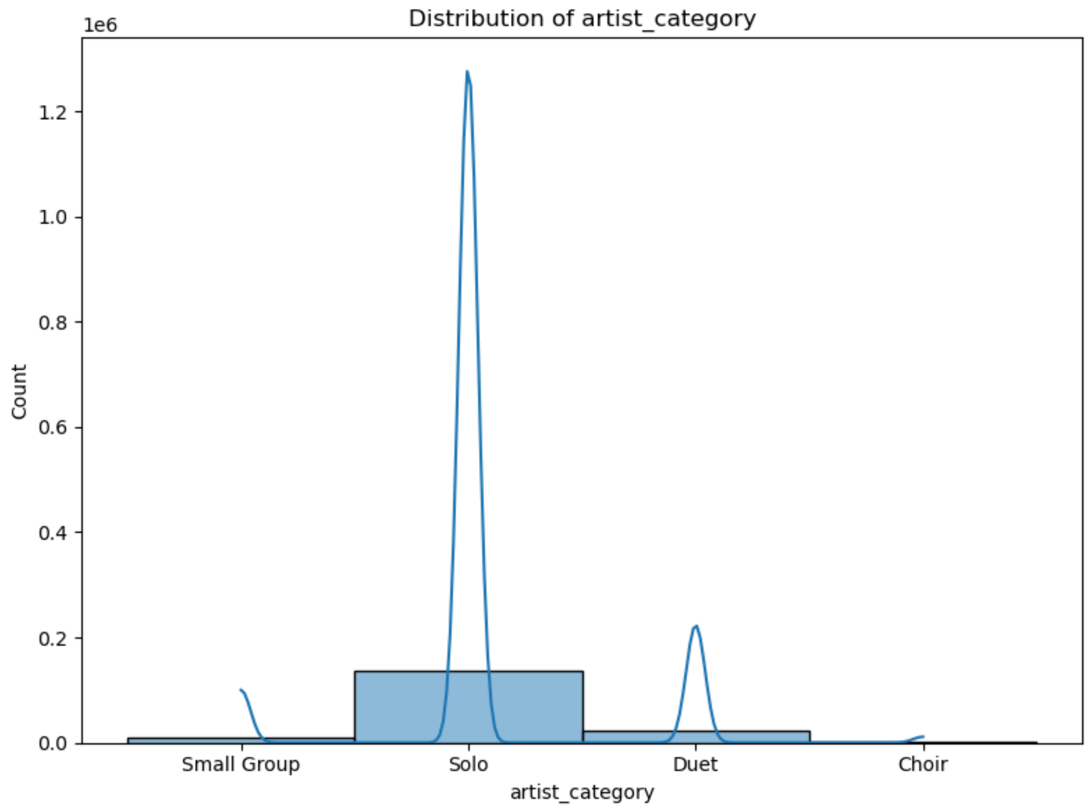


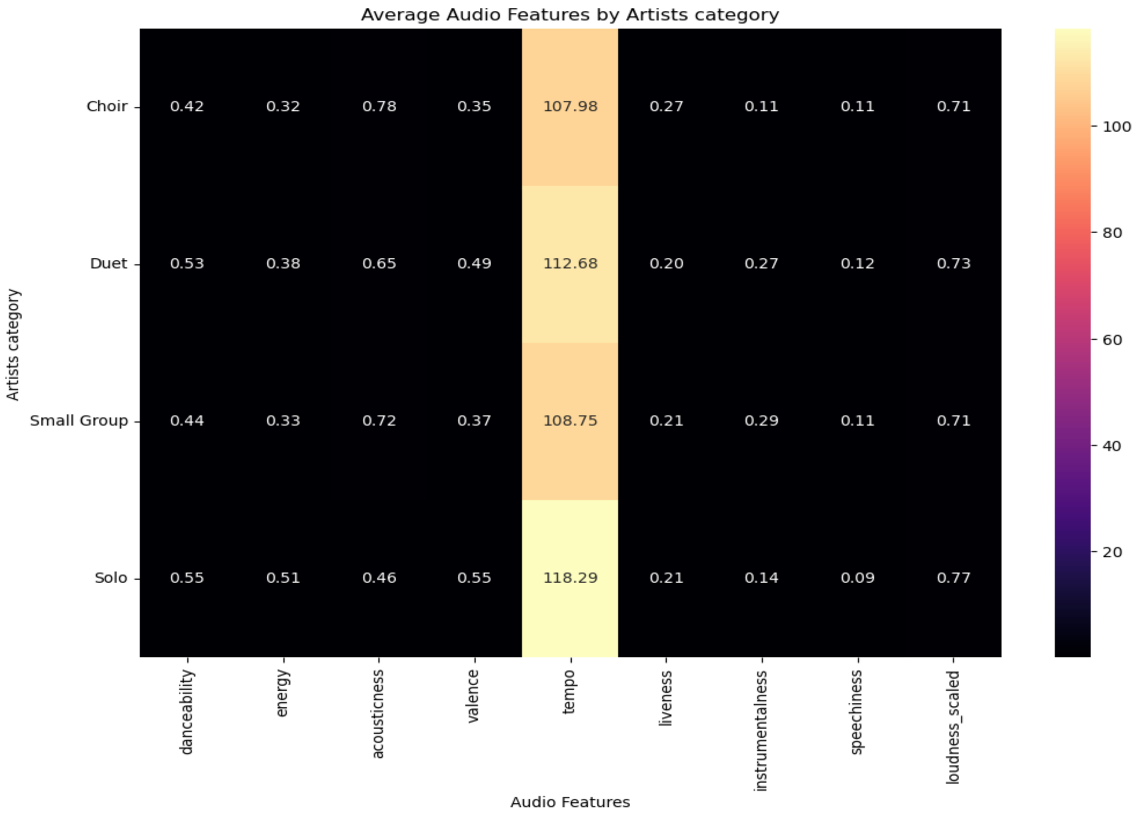


Data Exploration

Reference: 03_data_Exploration.ipynb

- **Analysis based on artists :**
  - **Distribution of artist category:** Distribution of the artist type like solo, duet, small group and choir is observed.
  - **Average audio features by artist category:** Based on the artist type , the average of music features is observed.





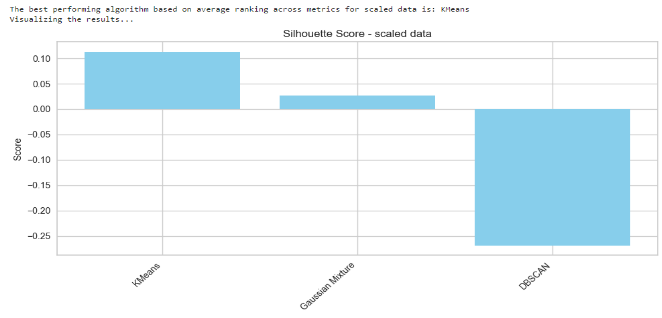
Model Selection

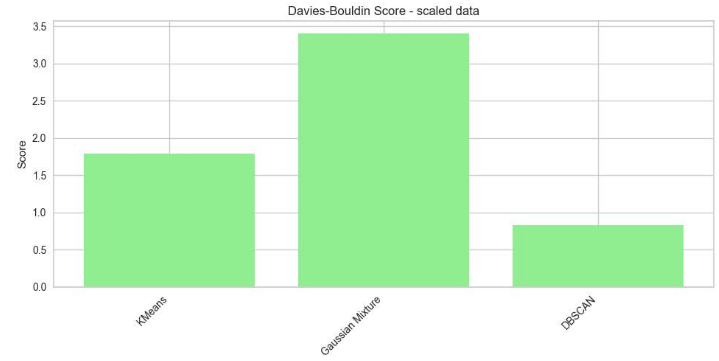
Best Algorithm: K-Means

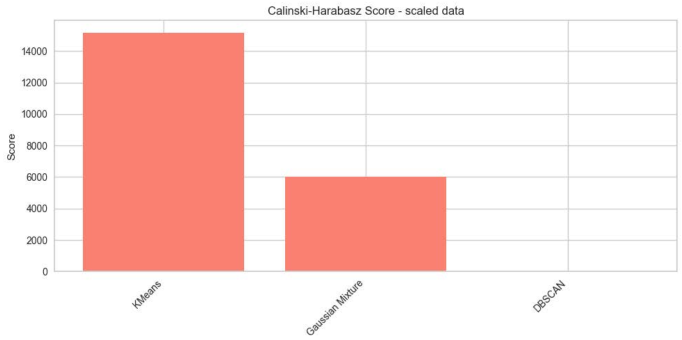
Reference: 04_Model_Selection.ipynb

Optimum cluster number found using elbow method

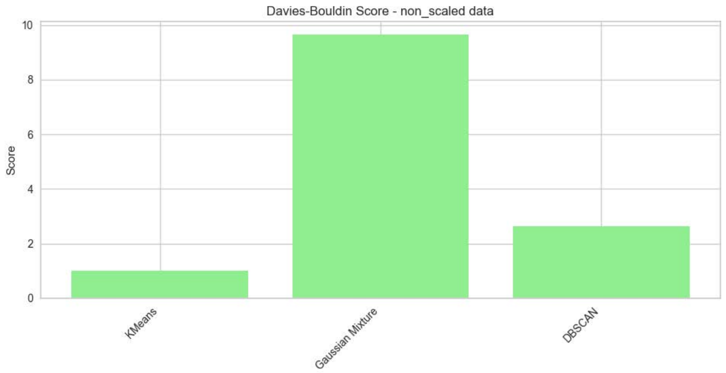
Data.csv

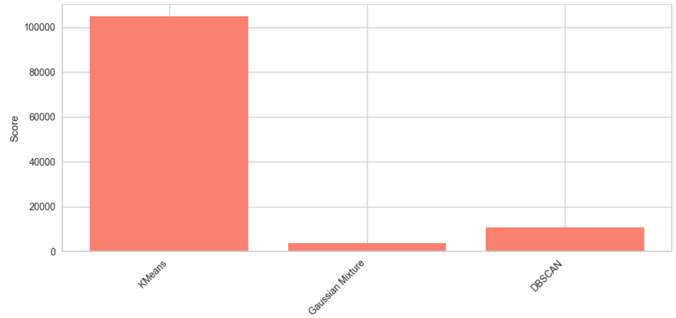






Scaled





Non Scaled

Silhouette Score

Measures the similarity of an object to its own cluster compared to other clusters.

Higher Score: Indicates a good match to its own cluster and poor match to neighboring clusters, suggesting well-defined, separate clusters.

Lower Score: Indicates overlapping clusters.

Davies-Bouldin Score

Measures the average similarity between clusters, where similarity is the ratio of the sum of intra-cluster distances to inter-cluster distances.

Higher Score: Indicates poor clustering, as clusters are either overlapping or dispersed.

Lower Score: Indicates better clustering, as clusters are more compact and better separated.

Calinski-Harabasz Score

It measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion.

Higher Score: Indicates clusters are well-separated and dense, which reflects a better clustering structure.

Lower Score: Indicates the clusters are not distinctly separated and might be too spread out.

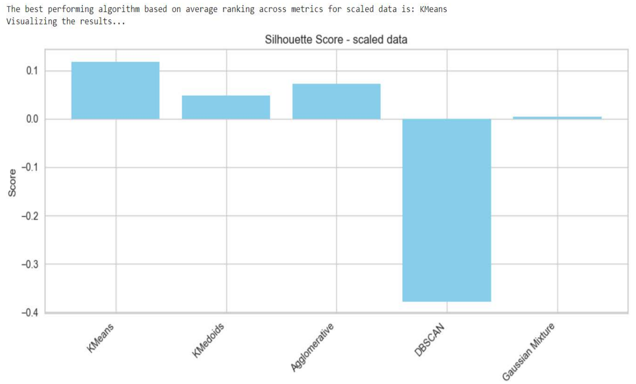
Model Selection

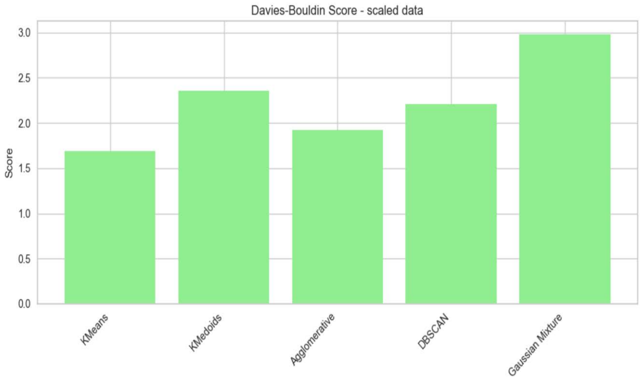
Best Algorithm: K-Means

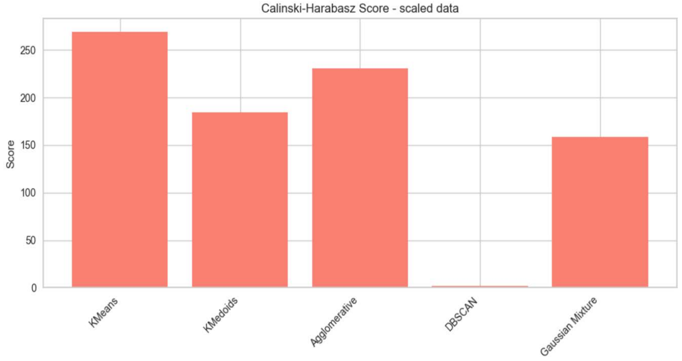
Reference: 04_Model_Selection.ipynb

Optimum cluster number found using elbow method

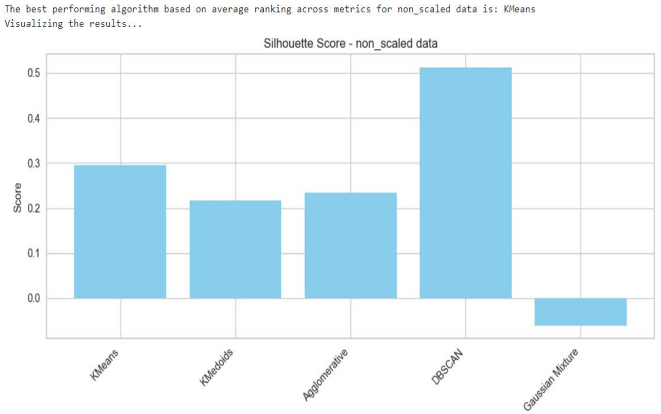
Genre_data.csv

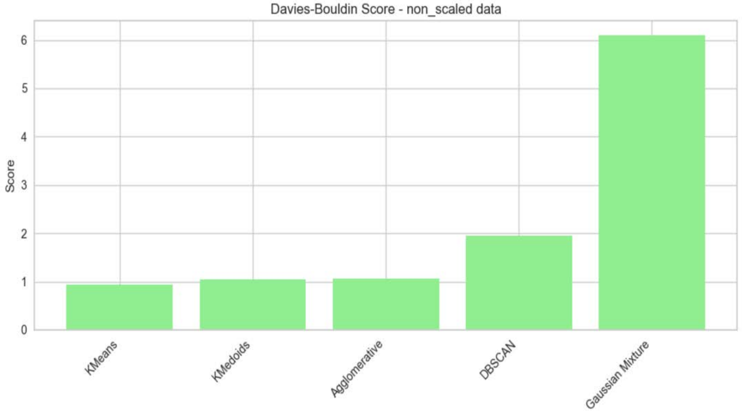


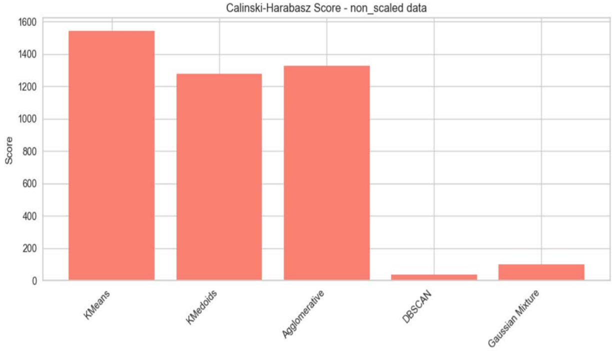




Scaled







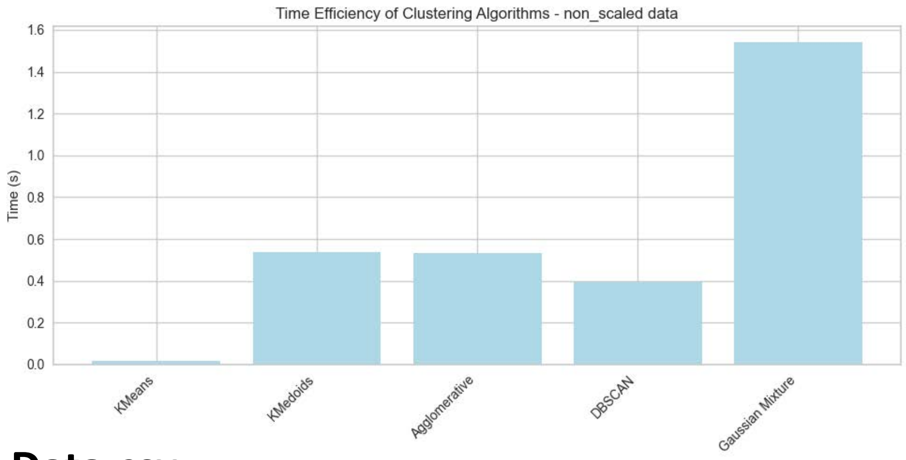
Non Scaled

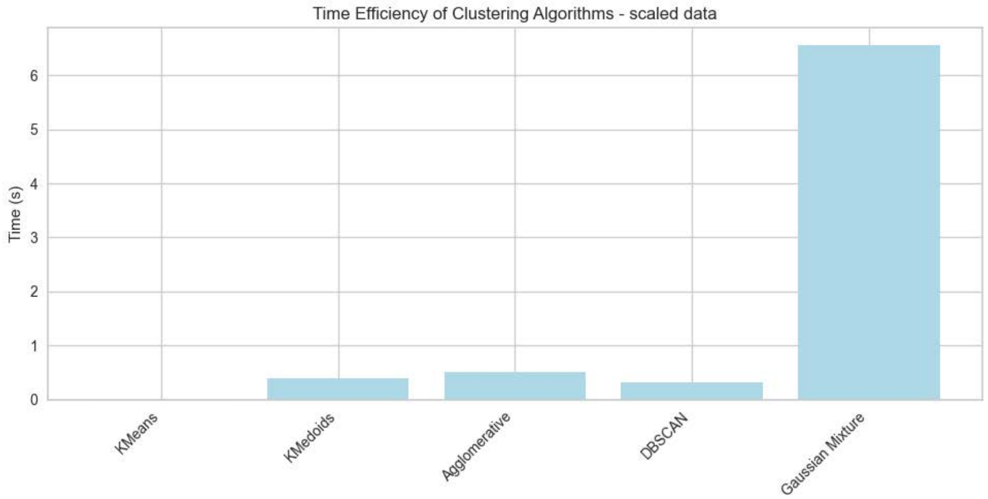
- ✓ Based on the data from the chart's, clustering performance is generally better with **non-scaled data** for this dataset.
- ✓ The non-scaled data shows higher silhouette and Calinski-Harabasz scores and lower Davies-Bouldin scores, which collectively suggest more effective and distinct clustering compared to when the data is scaled.
- ✓ This implies that scaling, in this case, might distort important relationships and distributions in the data that are crucial for effective clustering.

Model Selection

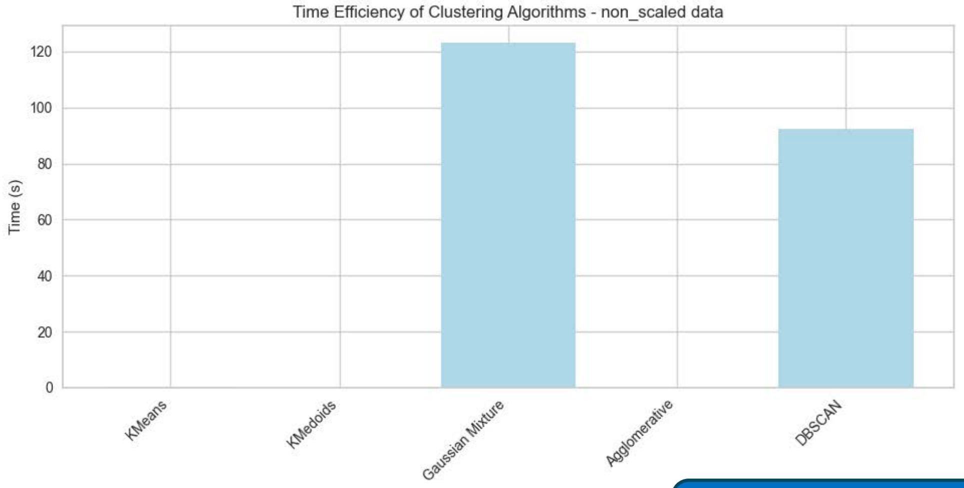
Reference: 04_Model_Selection.ipynb

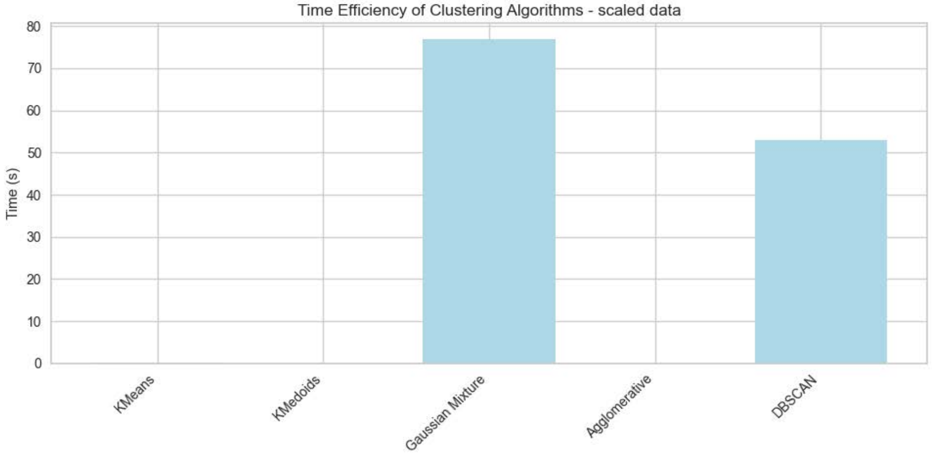
Genre_data.csv





Data.csv





Best Algorithm: K-Means

Hyperparameter Tuning

Reference: 05_hyperparameter_tuning.ipynb

Grid Search- Genre Data

Best Parameters

- Init: k-means++
- Max_iter:300
- N_clusters:13
- N_init:20
- Tol:0.001

Test data Evaluation

- Silhouette score:0.2846
- Davies-Bouldin Index:0.9270
- Calinski-Harabasz:313.9262

Grid Search Data

Best Parameters

- Init: k-means++
- Max_iter:300
- N_clusters:11
- N_init:10
- Tol:0.0001

Test data Evaluation

- Silhouette score:0.2917
- Davies-Bouldin Index:0.9754
- Calinski-Harabasz:21179.5098

t-SNE Genre Data

Best Parameters

- Early_exaggeration:24
- Learning_rate:200
- Max_iter:2000
- Perplexity:50

Test Data Evaluation

KL Divergence : 0.4223

PCA Data

Best Parameters

- PCA__n_components:2
- PCA__svd_solver:'auto'
- PCA__tol:0.0001
- PCA__whiten:True

Test Data Evaluation

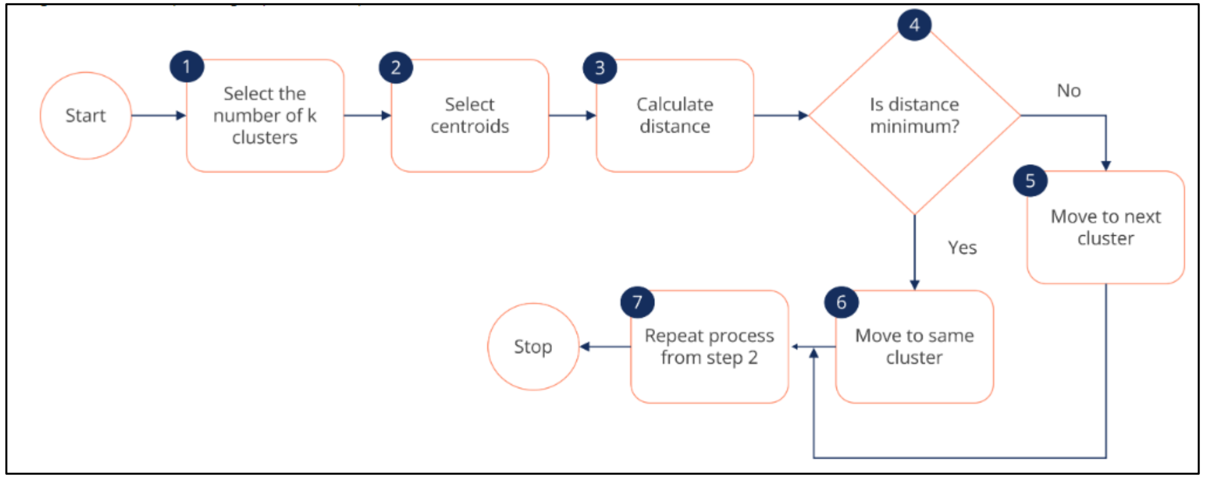
- Reconstruction error:459695.2128

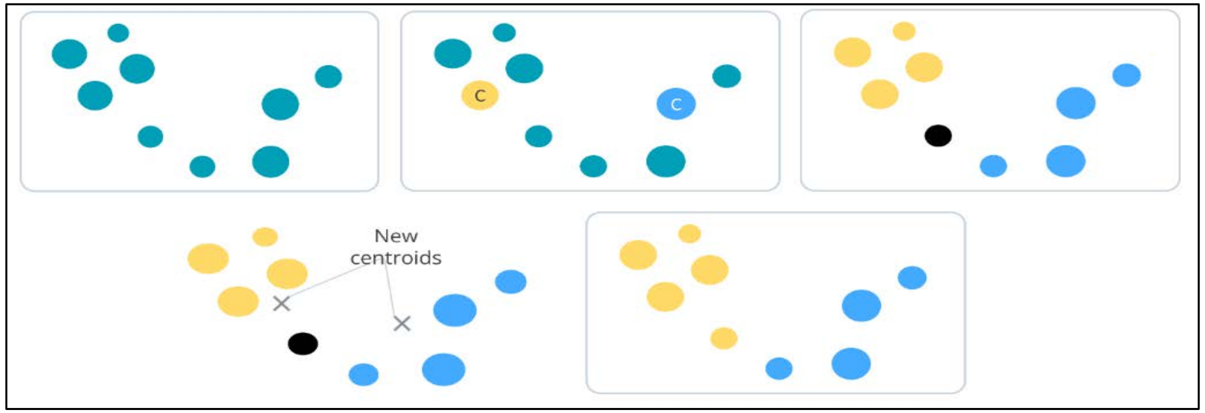
Model Training

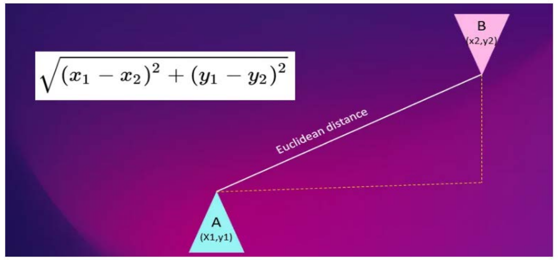
Understanding K-Means….

Reference: 06_model_training.ipynb

- ✓ Clustering is a fundamental technique in unsupervised machine learning, used to identify patterns within data by grouping similar data points together.
- ✓ The core objective of a clustering algorithm is to locate data points that share common characteristics, thus assigning them to the same cluster.

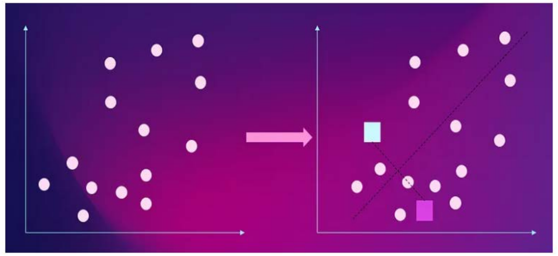


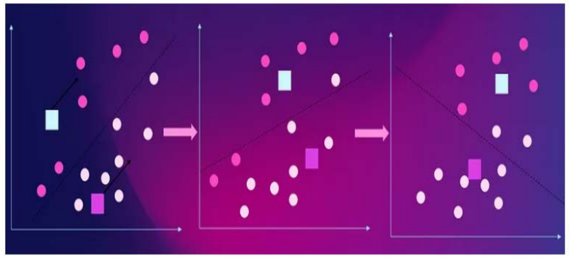




Euclidean Distance

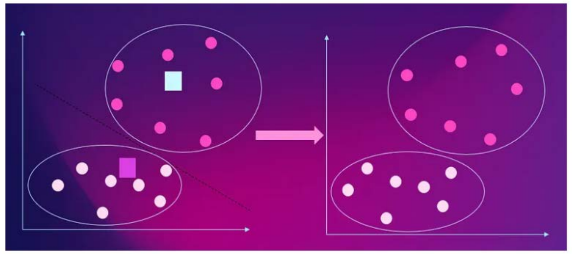
K-means in Action





Assignment & Optimization Step

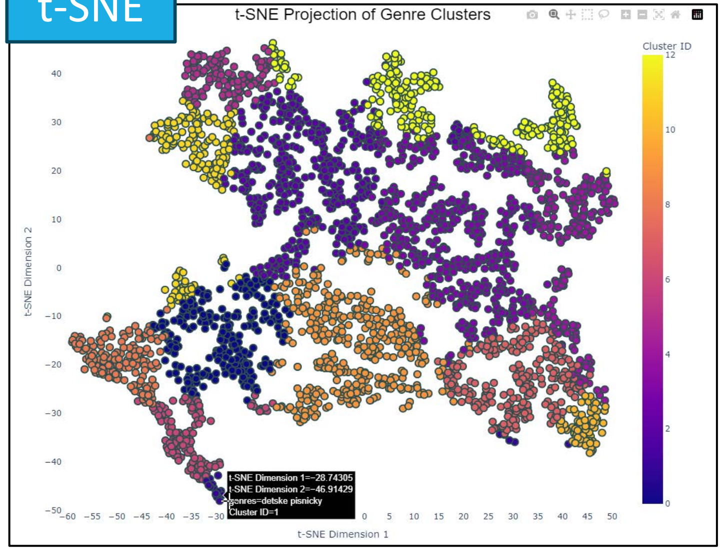
Iteration



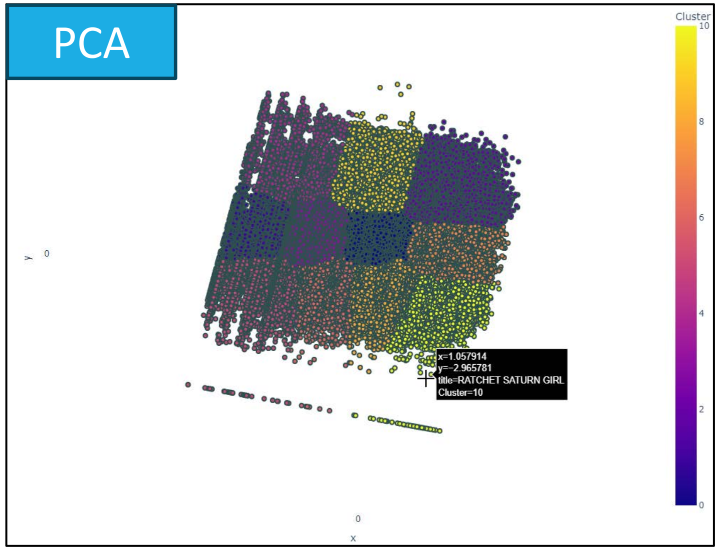
Model Training

Reference: 06_model_training.ipynb

t-SNE



PCA



- ✓ Our analysis shows that songs from the same genre group together in our dataset.
- ✓ This pattern is clear and expected, as songs within a genre typically share many features like beat, instruments, and even the era they're from.
- ✓ This grouping is great for making music recommendations. If you like a particular song, chances are you'll enjoy other songs that are close to it in our dataset.
- ✓ These songs aren't just similar in sound—they share a style and vibe that appeal to the same taste.

How We Use This for Better Recommendations:

- ✓ We can use this information to suggest songs that you are likely to enjoy. By looking at the songs you've listened to and loved, our system finds other songs from the same cluster and recommends them to you.

It's like having a smart DJ who knows exactly what you like and what to play next.

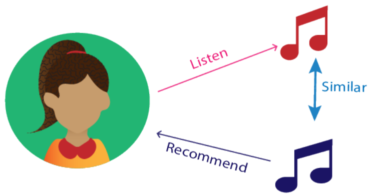
This isn't about guessing what you might like; it's about using data to find perfect matches based on your previous choices.

The result? Every song suggestion feels personal and just right for your musical preferences.

Model Deployment

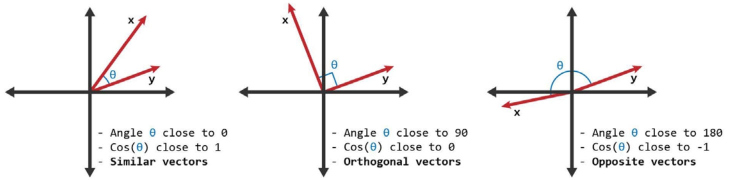
Reference: 07_model_deployment.ipynb

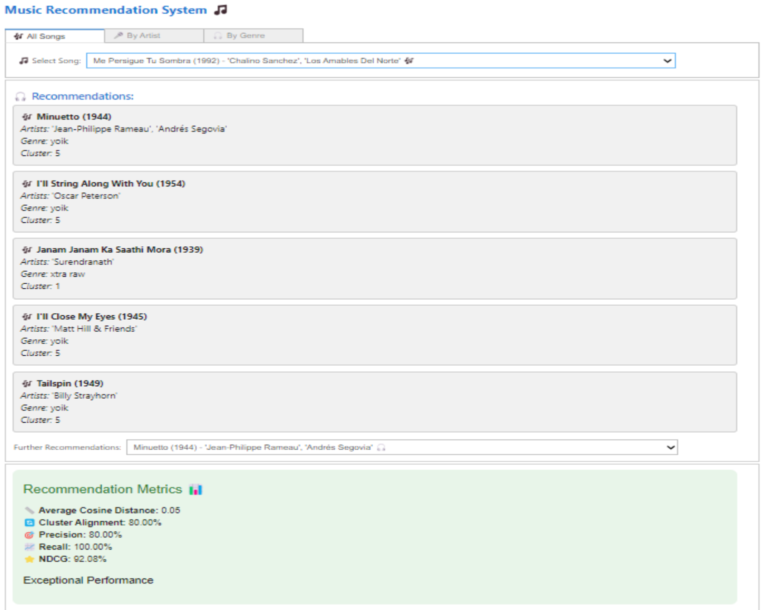
Content-based filtering recommends items based on the features and attributes of the items themselves, often comparing the similarity between item profiles to determine which items to recommend to a user.



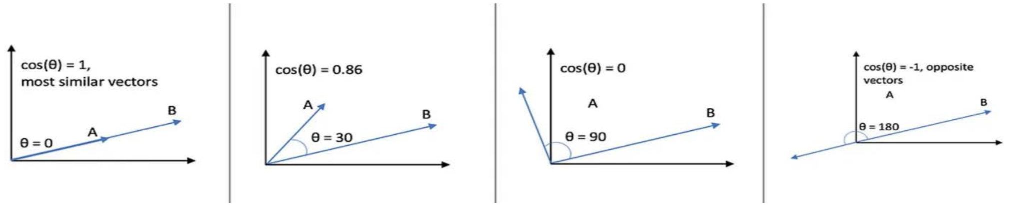
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

calculation of cosine of the angle between A and B



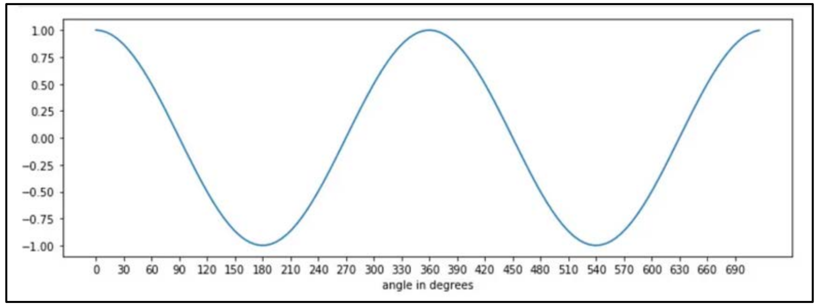


A cosine similarity is a value that is bound by a constrained range of 0 and 1. The closer the value is to 0 means that the two vectors are orthogonal or perpendicular to each other. When the value is closer to one, it means the angle is smaller and the songs are more similar.



Why cosine of the angle between A and B gives us the similarity?

If we look at the cosine function, it is 1 at theta = 0 and -1 at theta = 180, that means for two overlapping vectors cosine will be the highest and lowest for two exactly opposite vectors.



Average Cosine Distance: Measures the average cosine similarity between items in a dataset, with a value of 0 indicating perfect similarity.

Cluster Alignment: Indicates the percentage of data points that are correctly grouped into their respective clusters as per the model's predictions.

>> **Precision**: The proportion of positive identifications that were correct.

>> **Recall**: The proportion of actual positives that were correctly identified.

>> **NDCG (Normalized Discounted Cumulative Gain)**: A measure of ranking quality, assessing the effectiveness of a recommendation system by weighing the ranks of relevant items.

Current Capability

- ✓ **Data Handling**: Utilizes features like valence and danceability .. etc., shuffled and sampled for processing efficiency.
- ✓ **Recommendation Mechanics**: Employs cosine similarity to recommend songs based on nearest feature neighbors.
- ✓ **Interactive Interface**: Uses IPython widgets for dynamic song, artist, and genre selection.
- ✓ **Performance Metrics**: Measures effectiveness through precision, recall, and NDCG scores, displayed in real-time.
- ✓ **Recursive Recommendations**: Allows users to explore deeper into song recommendations iteratively.

Future Scope

- ✓ **Algorithm Optimization**: Explore deep learning for feature extraction and collaborative filtering for personalized recommendations.
- ✓ **Real-time Processing**: Adapt system to handle live data updates and scale efficiently with larger datasets.
- ✓ **Personalization**: Develop user profiles and contextual filters for mood-based recommendations.
- ✓ **Advanced Metrics**: Implement A/B testing and use engagement metrics for continuous improvement.
- ✓ **User Experience**: Enhance interface with audio previews and detailed visualizations of song features.

Data Science Canvas

Project: Personalized Music Recommendation System

Team:
1. Gayatri Yendamury
2. Bagi Shirisha
3. Priyabrata Samantaray
4. Vineet Kumar Tripathi

Problem Statement

Business Case & Value Added
Which business case should be analyzed and what added value does it generate?

Enhances user experience by providing personalized, relevant music recommendations, which drive user engagement and retention, offering a competitive edge in the music streaming market.

Data Landscape
Which data is required for this, and which is already available? Which additional data must be collected?

Required song attribute data is already available in the Spotify dataset. Additional user interaction data is optional for enhancing recommendations.

Model Selection
Which analysis methods can be considered on the basis of the specific data landscape and the business case?

Using K-Means clustering and content-based filtering.

Model Requirements
Which model requirements must be complied with in order to obtain a valid model?

The system must handle scalable feature engineering and must deliver recommendations without significant lag, ensuring data used in the model is current and relevant.

Software & Libraries
Which software should be used? Is there already a standard solution? Which libraries are used?

Python is the primary software, using libraries like Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn for data analysis and model building. No standard solution is required beyond these tools.

Skills
What skills are needed to provide the data and model development?

Proficiency in Python, data manipulation with Pandas, visualization with Matplotlib and Seaborn, machine learning with Scikit-learn, and a good understanding of recommendation algorithms.

Execution & Evaluation

Model Evaluation
Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

Success metrics will include Silhouette Score, Davies-Bouldin Score, Calinski-Harabasz Score Continuous performance monitoring is recommended to adjust models in real-time based on user feedback and interaction.

Data Storytelling
What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

Results will be presented using clear, actionable charts and summaries. Emphasis on key findings and how the model improves user experience and engagement.

Data Collection & Preparation

Data Selection & Cleansing
Which of the available data is relevant? Do the data have to be cleaned up?

Data is relevant to the current trends and user needs. Regular updates and cleaning of data sets are crucial to maintain the effectiveness of the recommendations.

Data Integration
In which system should the data from different sources be migrated?

Data from various sources (song metadata, user data, and interaction data) should be integrated into a unified database for efficient processing and analysis.

Data Collection
How and with which methods should additionally required data be collected? What properties has this data to fulfil?

Collection of song metadata, user behavior data, and preferences. Focus on cleaning the data for accuracy by addressing missing values, duplicates, and inconsistencies.

Explorative Data Analysis
Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.

Identify outliers, perform summary statistics, and visualize data distributions to understand underlying patterns. Use these insights for feature engineering and model refinement.

Adopted from: https://github.com/tomalytics/datasciencecanvas

Roles & Responsibilities

Stage	Notebook	Description	Responsibilities
Check point 1	01_data_collection.ipynb	Notebook for the task of data collection, likely involving gathering data	Bagi Shirisha Review: Priyabrata Samantaray
	02_data_preprocessing.ipynb	Notebook focused on preprocessing the collected data, which includes cleaning, outlier's detection, feature extraction, etc.	Gayatri Yendamury Review: Vineet Kumar Tripathi
	03_data_Exploration.ipynb	Notebook for exploring the data through statistical analysis and visualization to understand patterns etc.	Bagi Shirisha, Gayatri Yendamury Review: All
Check point 2	04_model_selection.ipynb	Notebook dedicated to selecting the appropriate model for the data.	Priyabrata Samantaray, Vineet Kumar Tripathi Review: Bagi Shirisha, Gayatri Yendamury
	05_hyperparameter_tuning.ipynb	Notebook for tuning the hyperparameters of the selected model to optimize performance.	Bagi Shirisha, Gayatri Yendamury Review: Priyabrata Samantaray, Vineet Kumar Tripathi
	06_model_training.ipynb	Notebook for training the model using the preprocessed and optimized data.	Priyabrata Samantaray, Vineet Kumar Tripathi Review: All
	07_model_deployment.ipynb	Notebook for deploying the trained model into a production environment or for further validation.	Priyabrata Samantaray, Vineet Kumar Tripathi Review: All