# MUSE

## Group 16

Adib Menchali 771031
Lorenzo Conti 760361

# Table of contents

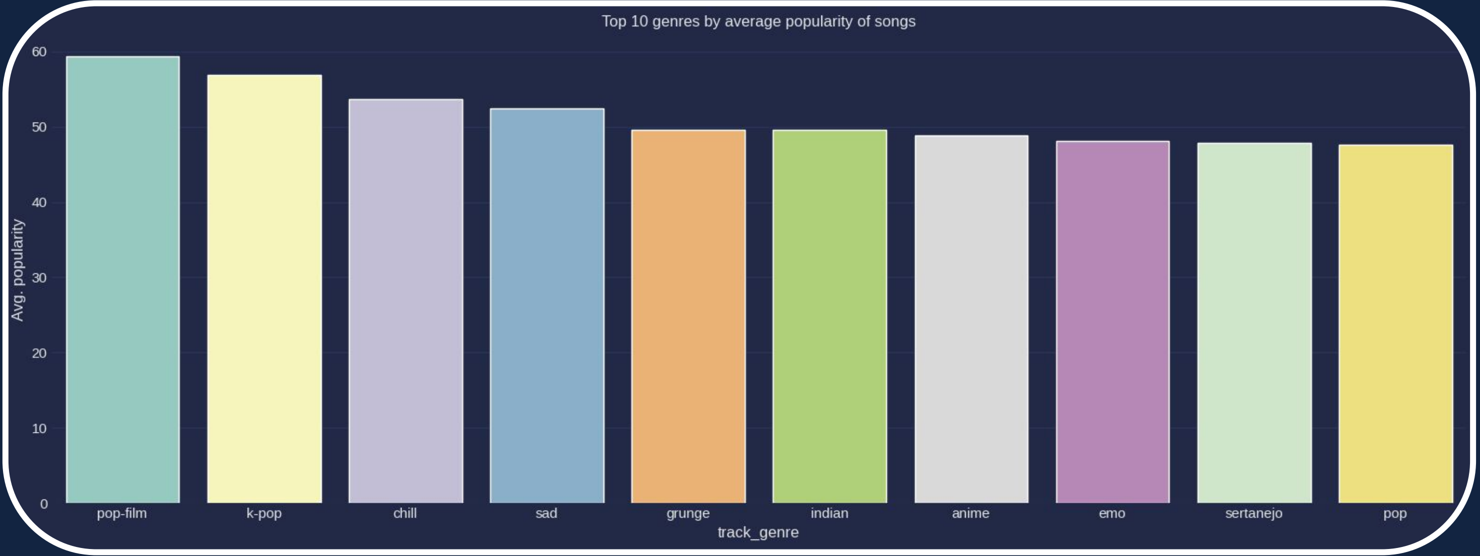
1. Introduction
2. Data preprocessing
3. Exploring the data
4. Building the models
5. Wrap up

# Data preprocessing

Dropping null values.
Dropping duplicates: 450 duplicated rows.
Aggregating the track genre to avoid data redundancy.
Feature selection: (Dropping track_id, energy)

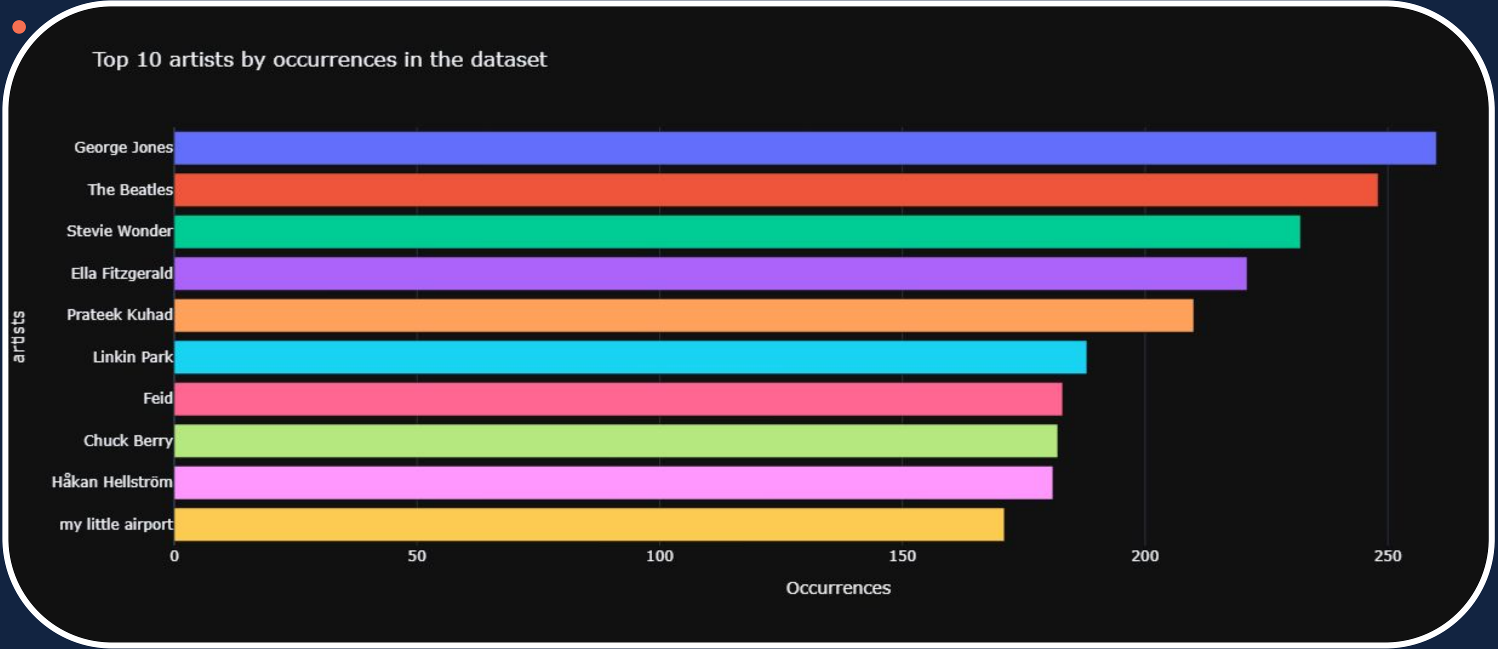
# Exploring the data

# What are the most popular Genres in our data?



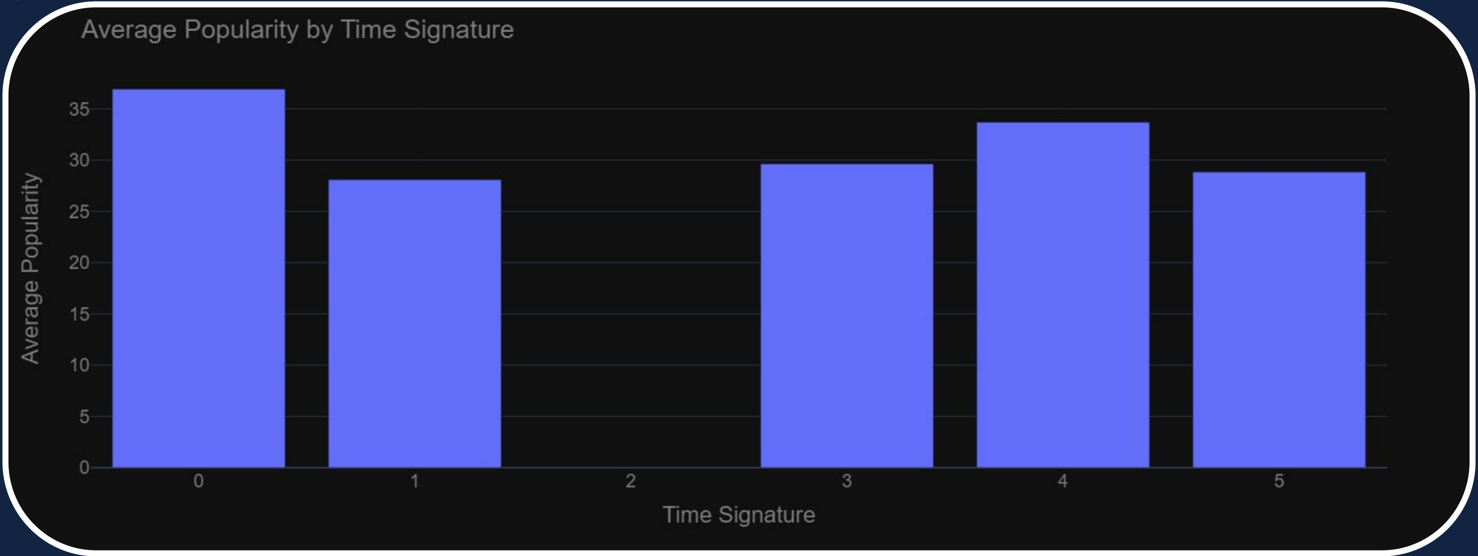
- Pop-film is the most popular genre in our dataset followed by k-pop and chill.

# Who are the most influential Artists in our data?



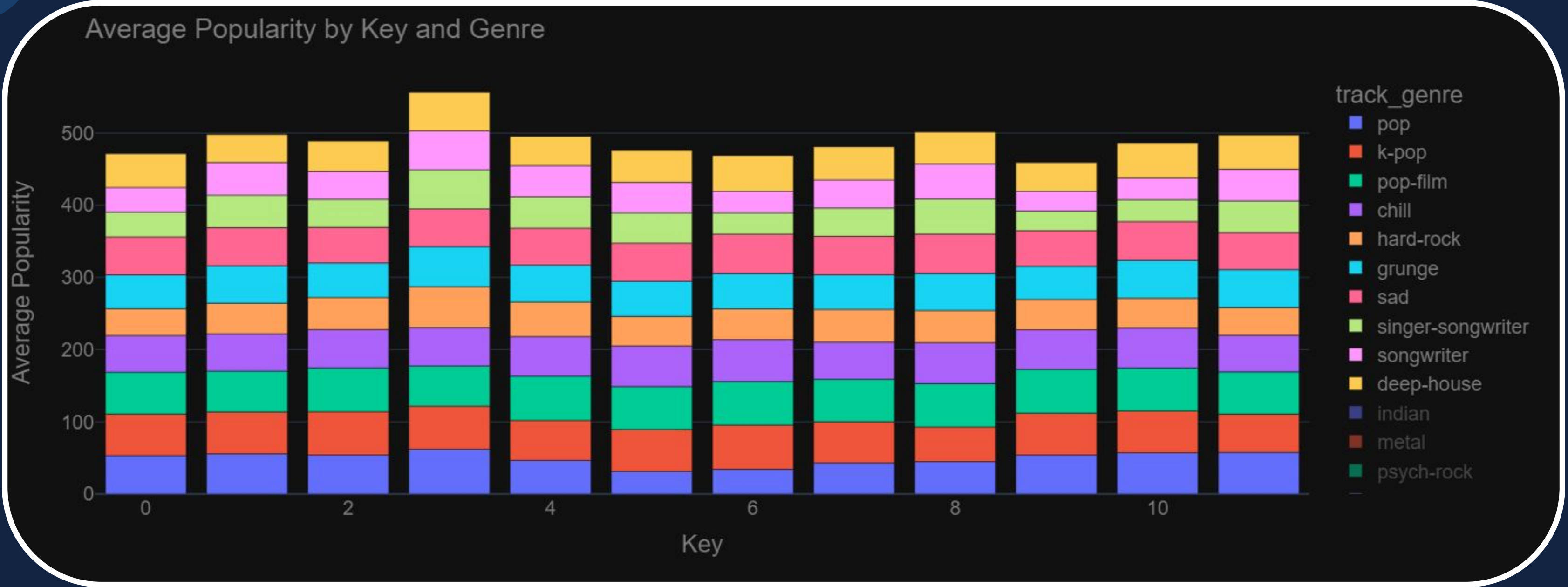
- George Jones, The Beatles, Stevie Wonder, Ella Fitzgerald, Prateek Kuhad are the most influential artists with over 200 songs produced.

# Does Time Signature affect popularity?



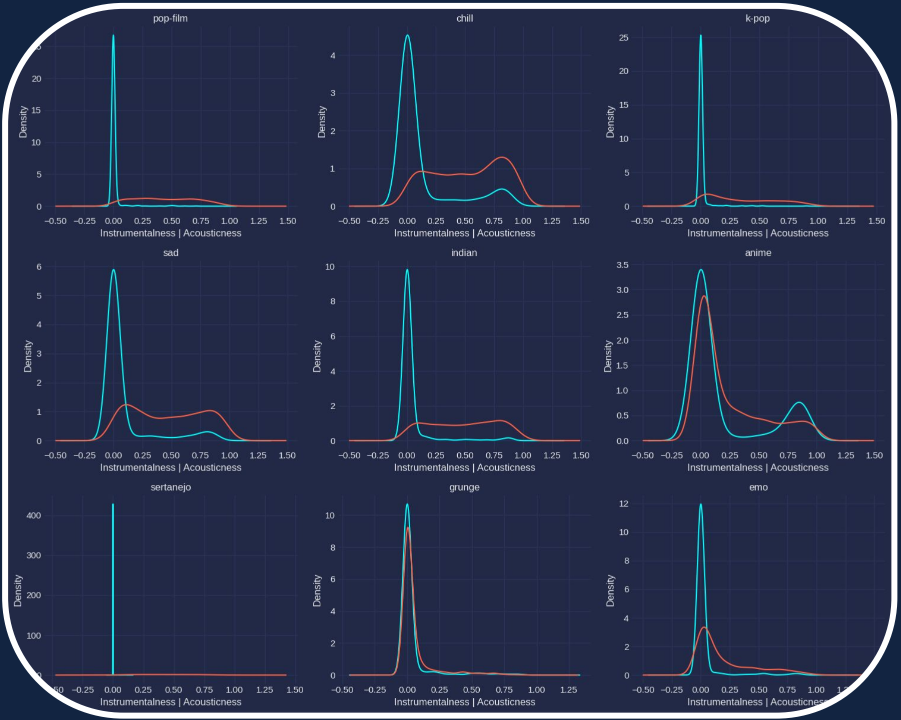
- Time signature 0 corresponds to white/brown noise.
- The data doesn’t contain tracks with Time Signature 2.

# Does the track Key affect popularity?



- Key 3 has the highest average popularity.
- All keys seem to be used across all genres.

# Can we identify genres by visualizing their features?

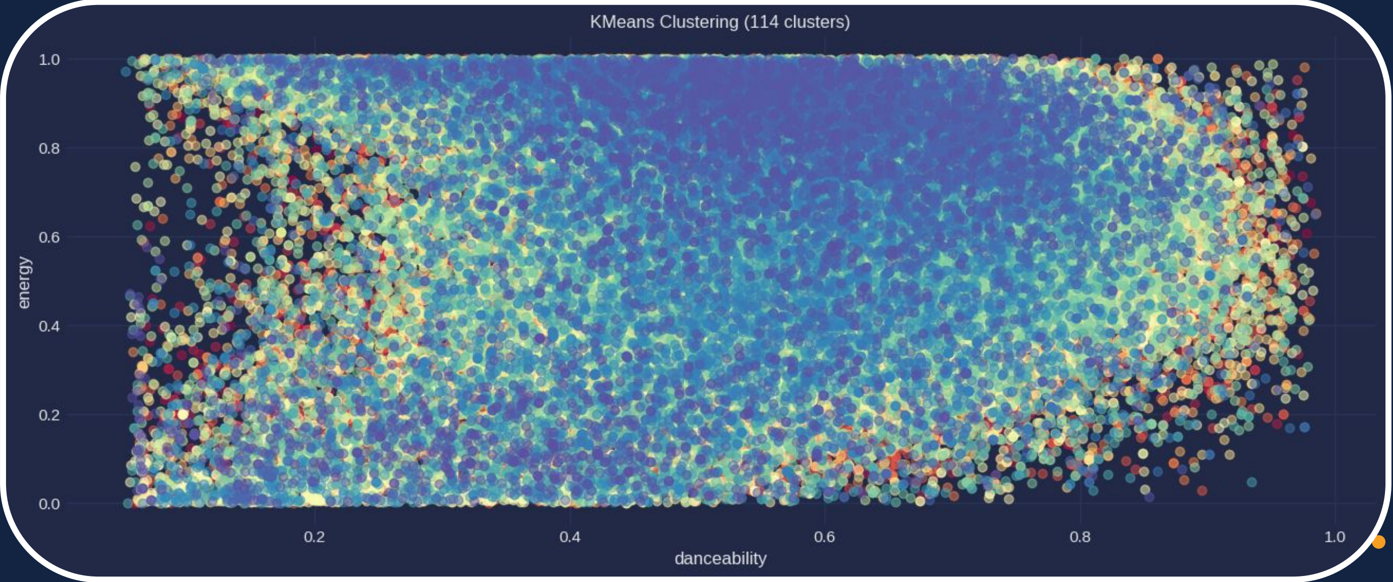


- It’s difficult to pinpoint the genre just by looking at the distribution of its music features.

# Clustering the data using K-means

Independent features: 'danceability', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'time_signature'

K = 114 clusters



# Clustering the data using K-means

**Adjusted Rand Index (ARI):** A metric that measures the similarity between two data clusterings.

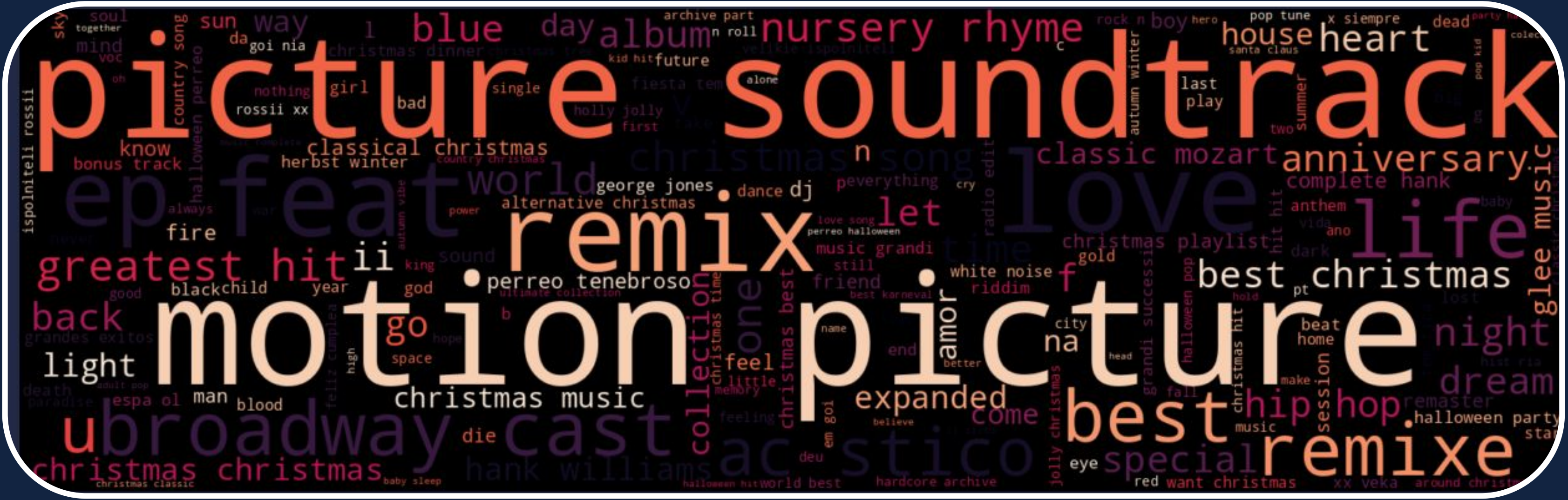
**Adjusted Rand Index: 0.010877535748024246**

**An ARI close to 0 indicates that the clustering does not align well with the track_genre labels.**

**> This is not a good technique to infer the genre of a track.**

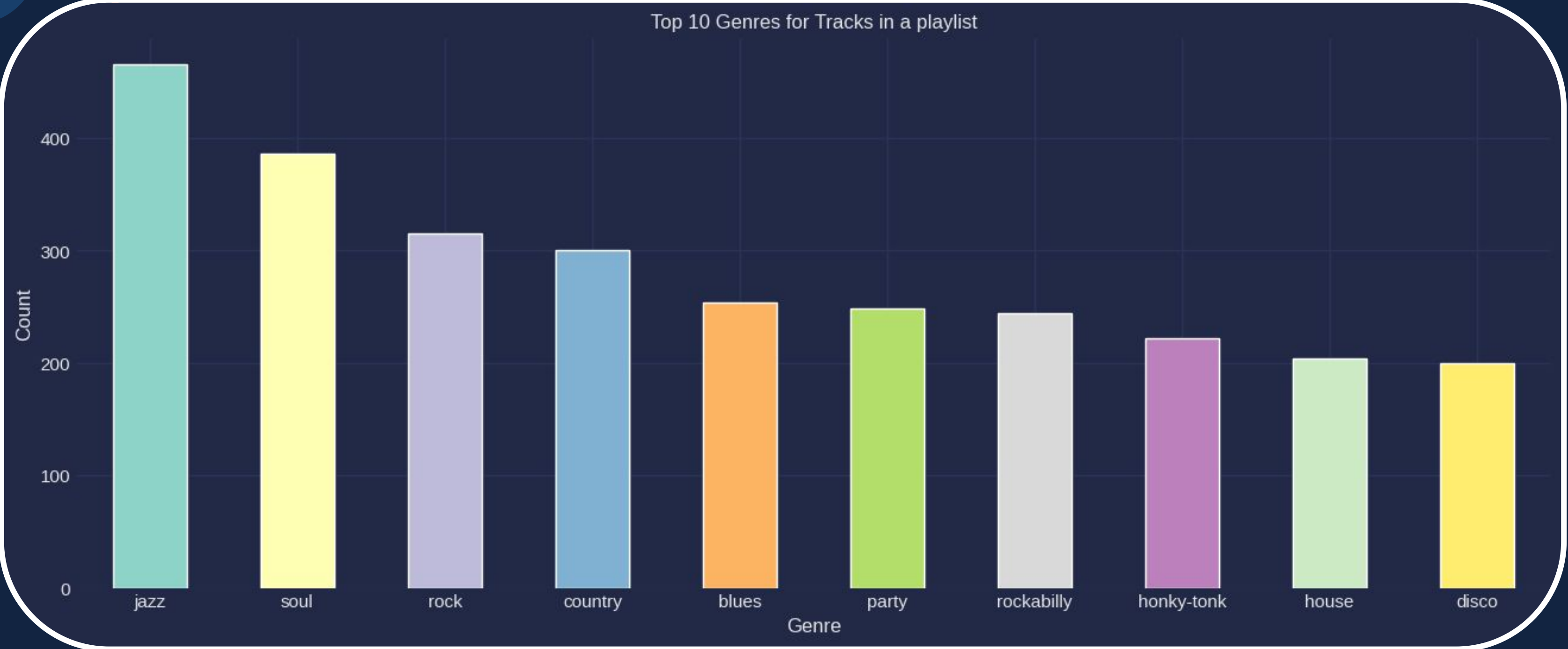
# Natural Language Processing

Using **NLP** to conduct **topic modeling** (Latent Dirichlet Allocation) on album names



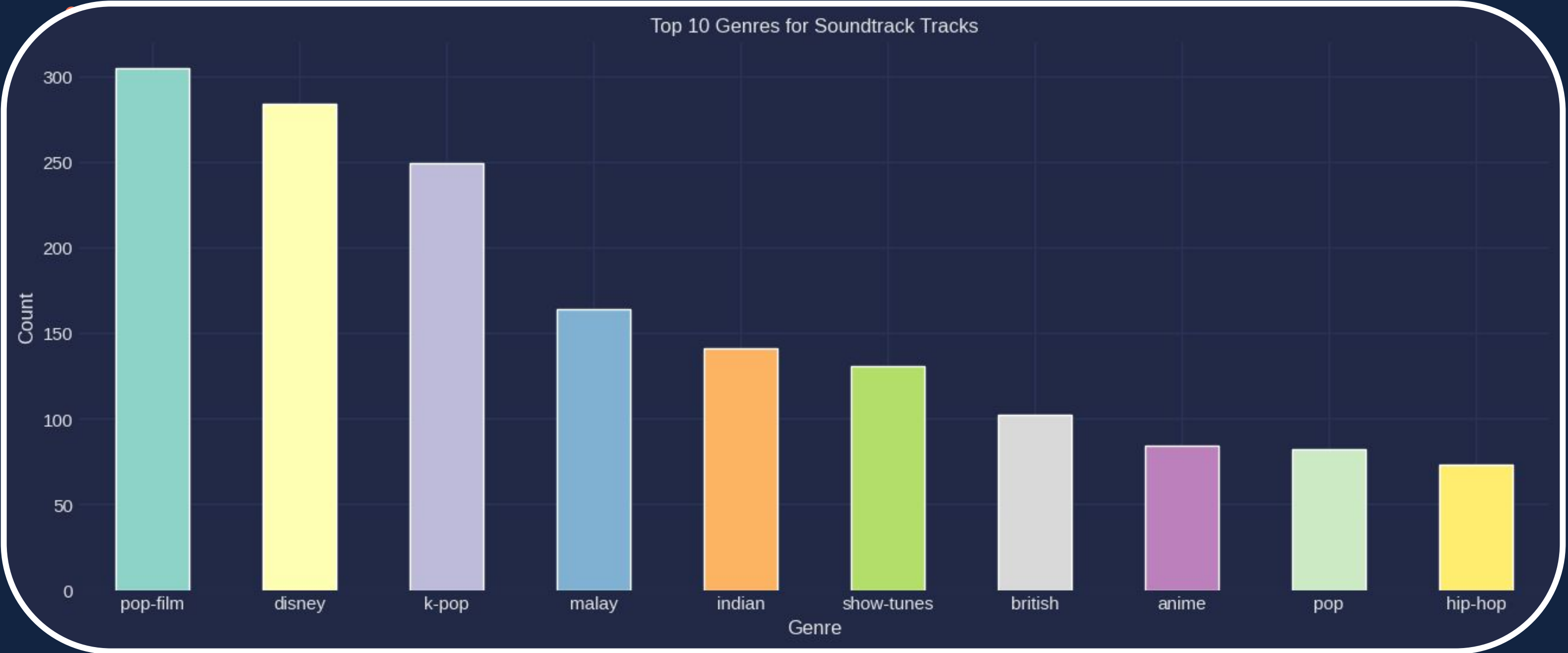
- Our dataset contains tracks that are found in playlists and soundtracks
- Feature extraction: Playlist, Soundtrack (Binary variables)

# What track genres are being added to playlists the most?



- Jazz is the most popular genre for songs contained in playlists, followed by soul and rock.

# What track genres are being used in soundtracks the most?



- Pop-film, disney and k-pop are the most popular tracks genres that are being used in soundtracks.

# Building the models

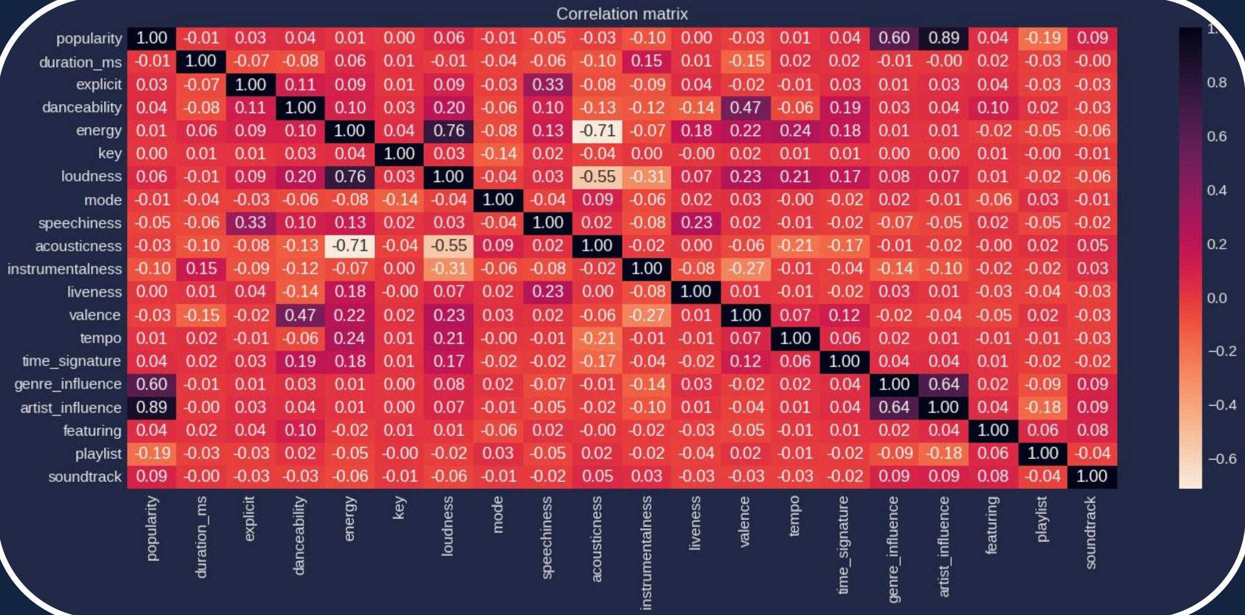
# Methodology

Extracting new features from the artists and genre columns: **Artist_influence** and **Genre_influence**. These variables store the average Popularity for each artist and genre respectively.

**Approach 1:**
Feeding the whole pre-processed data to the model.

**Approach 2:**
Filtering the data based on a popularity threshold that minimizes the correlation between the dependent feature ‘Popularity’ and ‘Artist_influence’.

# Approach 1: Correlation Matrix



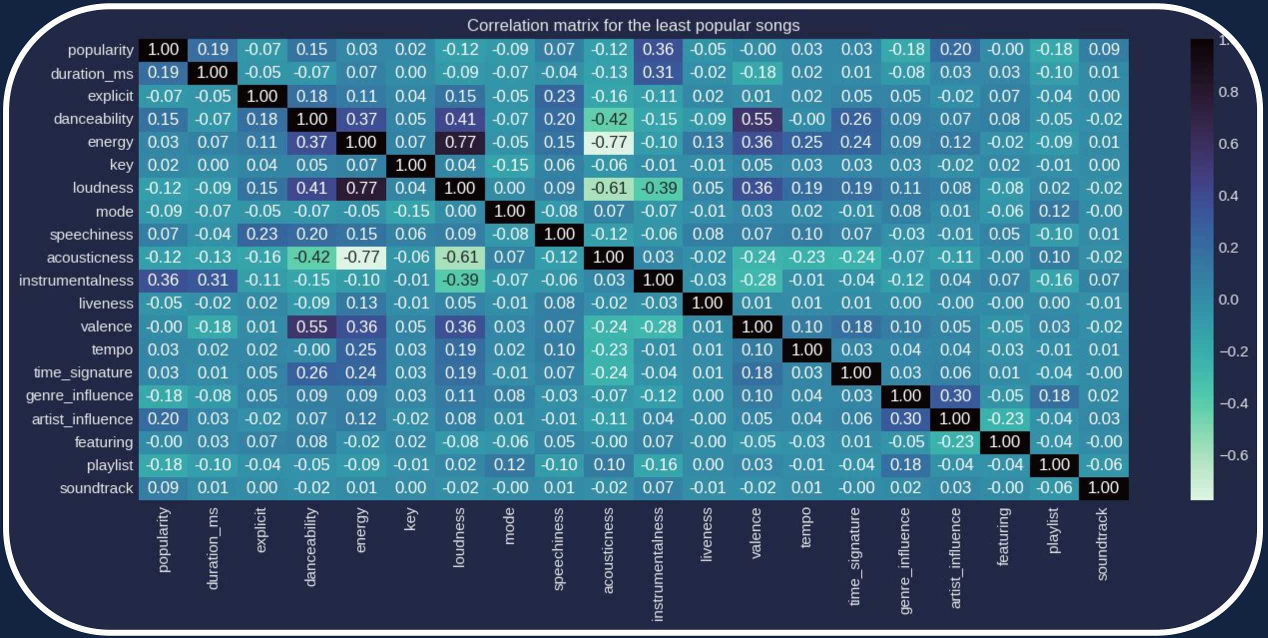
Correlation matrix

| | popularity | duration_ms | explicit | danceability | energy | key | loudness | mode | speechiness | acousticness | instrumentalness | liveness | valence | tempo | time_signature | genre_influence | artist_influence | featuring | playlist | soundtrack |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| popularity | 1.00 | -0.01 | 0.03 | 0.04 | 0.01 | 0.00 | 0.06 | -0.01 | -0.05 | -0.03 | -0.10 | 0.00 | -0.03 | 0.01 | 0.04 | 0.60 | 0.89 | 0.04 | -0.19 | 0.09 |
| duration_ms | -0.01 | 1.00 | -0.07 | -0.08 | 0.06 | 0.01 | -0.01 | -0.04 | -0.06 | -0.10 | 0.15 | 0.01 | -0.15 | 0.02 | 0.02 | -0.01 | -0.00 | 0.02 | -0.03 | -0.00 |
| explicit | 0.03 | -0.07 | 1.00 | 0.11 | 0.09 | 0.01 | 0.09 | -0.03 | 0.33 | -0.08 | -0.09 | 0.04 | -0.02 | -0.01 | 0.03 | 0.01 | 0.03 | 0.04 | -0.03 | -0.03 |
| danceability | 0.04 | -0.08 | 0.11 | 1.00 | 0.10 | 0.03 | 0.20 | -0.06 | 0.10 | -0.13 | -0.12 | -0.14 | 0.47 | -0.06 | 0.19 | 0.03 | 0.04 | 0.10 | 0.02 | -0.03 |
| energy | 0.01 | 0.06 | 0.09 | 0.10 | 1.00 | 0.04 | 0.76 | -0.08 | 0.13 | -0.71 | -0.07 | 0.18 | 0.22 | 0.24 | 0.18 | 0.01 | 0.01 | -0.02 | -0.05 | -0.06 |
| key | 0.00 | 0.01 | 0.01 | 0.03 | 0.04 | 1.00 | 0.03 | -0.14 | 0.02 | -0.04 | 0.00 | -0.00 | 0.02 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | -0.00 | -0.01 |
| loudness | 0.06 | -0.01 | 0.09 | 0.20 | 0.76 | 0.03 | 1.00 | -0.04 | 0.03 | -0.55 | -0.31 | 0.07 | 0.23 | 0.21 | 0.17 | 0.08 | 0.07 | 0.01 | -0.02 | -0.06 |
| mode | -0.01 | -0.04 | -0.03 | -0.06 | -0.08 | -0.14 | -0.04 | 1.00 | -0.04 | 0.09 | -0.06 | 0.02 | 0.03 | -0.00 | -0.02 | 0.02 | -0.01 | -0.06 | 0.03 | -0.01 |
| speechiness | -0.05 | -0.06 | 0.33 | 0.10 | 0.13 | 0.02 | 0.03 | -0.04 | 1.00 | 0.02 | -0.08 | 0.23 | 0.02 | -0.01 | -0.02 | -0.07 | -0.05 | 0.02 | -0.05 | -0.02 |
| acousticness | -0.03 | -0.10 | -0.08 | -0.13 | -0.71 | -0.04 | -0.55 | 0.09 | 0.02 | 1.00 | -0.02 | 0.00 | -0.06 | -0.21 | -0.17 | -0.01 | -0.02 | -0.00 | 0.02 | 0.05 |
| instrumentalness | -0.10 | 0.15 | -0.09 | -0.12 | -0.07 | 0.00 | -0.31 | -0.06 | -0.08 | -0.02 | 1.00 | -0.08 | -0.27 | -0.01 | -0.04 | -0.14 | -0.10 | -0.02 | -0.02 | 0.03 |
| liveness | 0.00 | 0.01 | 0.04 | -0.14 | 0.18 | -0.00 | 0.07 | 0.02 | 0.23 | 0.00 | -0.08 | 1.00 | 0.01 | -0.01 | -0.02 | 0.03 | 0.01 | -0.03 | -0.04 | -0.03 |
| valence | -0.03 | -0.15 | -0.02 | 0.47 | 0.22 | 0.02 | 0.23 | 0.03 | 0.02 | -0.06 | -0.27 | 0.01 | 1.00 | 0.07 | 0.12 | -0.02 | -0.04 | -0.05 | 0.02 | -0.03 |
| tempo | 0.01 | 0.02 | -0.01 | -0.06 | 0.24 | 0.01 | 0.21 | -0.00 | -0.01 | -0.21 | -0.01 | -0.01 | 0.07 | 1.00 | 0.06 | 0.02 | 0.01 | -0.01 | -0.01 | -0.03 |
| time_signature | 0.04 | 0.02 | 0.03 | 0.19 | 0.18 | 0.01 | 0.17 | -0.02 | -0.02 | -0.17 | -0.04 | -0.02 | 0.12 | 0.06 | 1.00 | 0.04 | 0.04 | 0.01 | -0.02 | -0.02 |
| genre_influence | 0.60 | -0.01 | 0.01 | 0.03 | 0.01 | 0.00 | 0.08 | 0.02 | -0.07 | -0.01 | -0.14 | 0.03 | -0.02 | 0.02 | 0.04 | 1.00 | 0.64 | 0.02 | -0.09 | 0.09 |
| artist_influence | 0.89 | -0.00 | 0.03 | 0.04 | 0.01 | 0.00 | 0.07 | -0.01 | -0.05 | -0.02 | -0.10 | 0.01 | -0.04 | 0.01 | 0.04 | 0.64 | 1.00 | 0.04 | -0.18 | 0.09 |
| featuring | 0.04 | 0.02 | 0.04 | 0.10 | -0.02 | 0.01 | 0.01 | -0.06 | 0.02 | -0.00 | -0.02 | -0.03 | -0.05 | -0.01 | 0.01 | 0.02 | 0.04 | 1.00 | 0.06 | 0.08 |
| playlist | -0.19 | -0.03 | -0.03 | 0.02 | -0.05 | -0.00 | -0.02 | 0.03 | -0.05 | 0.02 | -0.02 | -0.04 | 0.02 | -0.01 | -0.02 | -0.09 | -0.18 | 0.06 | 1.00 | -0.04 |
| soundtrack | 0.09 | -0.00 | -0.03 | -0.03 | -0.06 | -0.01 | -0.06 | -0.01 | -0.02 | 0.05 | 0.03 | -0.03 | -0.03 | -0.03 | -0.02 | 0.09 | 0.09 | 0.08 | -0.04 | 1.00 |

- Popularity is highly correlated with Artist influence and Genre influence, 0.89 and 0.60 respectively.

- The music features don’t seem to correlate well with our target variable.

# Approach 2: Correlation Matrix



Correlation matrix for the least popular songs

| | popularity | duration_ms | explicit | danceability | energy | key | loudness | mode | speechiness | acousticness | instrumentalness | liveness | valence | tempo | time_signature | genre_influence | artist_influence | featuring | playlist | soundtrack |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| popularity | 1.00 | 0.19 | -0.07 | 0.15 | 0.03 | 0.02 | -0.12 | -0.09 | 0.07 | -0.12 | 0.36 | -0.05 | -0.00 | 0.03 | 0.03 | -0.18 | 0.20 | -0.00 | -0.18 | 0.09 |
| duration_ms | 0.19 | 1.00 | -0.05 | -0.07 | 0.07 | 0.00 | -0.09 | -0.07 | -0.04 | -0.13 | 0.31 | -0.02 | -0.18 | 0.02 | 0.01 | -0.08 | 0.03 | 0.03 | -0.10 | 0.01 |
| explicit | -0.07 | -0.05 | 1.00 | 0.18 | 0.11 | 0.04 | 0.15 | -0.05 | 0.23 | -0.16 | -0.11 | 0.02 | 0.01 | 0.02 | 0.05 | 0.05 | -0.02 | 0.07 | -0.04 | 0.00 |
| danceability | 0.15 | -0.07 | 0.18 | 1.00 | 0.37 | 0.05 | 0.41 | -0.07 | 0.20 | -0.42 | -0.15 | -0.09 | 0.55 | -0.00 | 0.26 | 0.09 | 0.07 | 0.08 | -0.05 | -0.02 |
| energy | 0.03 | 0.07 | 0.11 | 0.37 | 1.00 | 0.07 | 0.77 | -0.05 | 0.15 | -0.77 | -0.10 | 0.13 | 0.36 | 0.25 | 0.24 | 0.09 | 0.12 | -0.02 | -0.09 | 0.01 |
| key | 0.02 | 0.00 | 0.04 | 0.05 | 0.07 | 1.00 | 0.04 | -0.15 | 0.06 | -0.06 | -0.01 | -0.01 | 0.05 | 0.03 | 0.03 | 0.03 | -0.02 | 0.02 | -0.01 | 0.00 |
| loudness | -0.12 | -0.09 | 0.15 | 0.41 | 0.77 | 0.04 | 1.00 | 0.00 | 0.09 | -0.61 | -0.39 | 0.05 | 0.36 | 0.19 | 0.19 | 0.11 | 0.08 | -0.08 | 0.02 | -0.02 |
| mode | -0.09 | -0.07 | -0.05 | -0.07 | -0.05 | -0.15 | 0.00 | 1.00 | -0.08 | 0.07 | -0.07 | -0.01 | 0.03 | 0.02 | -0.01 | 0.08 | 0.01 | -0.06 | 0.12 | -0.00 |
| speechiness | 0.07 | -0.04 | 0.23 | 0.20 | 0.15 | 0.06 | 0.09 | -0.08 | 1.00 | -0.12 | -0.06 | 0.08 | 0.07 | 0.10 | 0.07 | -0.03 | -0.01 | 0.05 | -0.10 | 0.01 |
| acousticness | -0.12 | -0.13 | -0.16 | -0.42 | -0.77 | -0.06 | -0.61 | 0.07 | -0.12 | 1.00 | 0.03 | -0.02 | -0.24 | -0.23 | -0.24 | -0.07 | -0.11 | -0.00 | 0.10 | -0.02 |
| instrumentalness | 0.36 | 0.31 | -0.11 | -0.15 | -0.10 | -0.01 | -0.39 | -0.07 | -0.06 | 0.03 | 1.00 | -0.03 | -0.28 | -0.01 | -0.04 | -0.12 | 0.04 | 0.07 | -0.16 | 0.07 |
| liveness | -0.05 | -0.02 | 0.02 | -0.09 | 0.13 | -0.01 | 0.05 | -0.01 | 0.08 | -0.02 | -0.03 | 1.00 | 0.01 | 0.01 | 0.01 | 0.00 | -0.00 | -0.00 | 0.00 | -0.01 |
| valence | -0.00 | -0.18 | 0.01 | 0.55 | 0.36 | 0.05 | 0.36 | 0.03 | 0.07 | -0.24 | -0.28 | 0.01 | 1.00 | 0.10 | 0.18 | 0.10 | 0.05 | -0.05 | 0.03 | -0.02 |
| tempo | 0.03 | 0.02 | 0.02 | -0.00 | 0.25 | 0.03 | 0.19 | 0.02 | 0.10 | -0.23 | -0.01 | 0.01 | 0.10 | 1.00 | 0.03 | 0.04 | 0.04 | -0.03 | -0.01 | 0.01 |
| time_signature | 0.03 | 0.01 | 0.05 | 0.26 | 0.24 | 0.03 | 0.19 | -0.01 | 0.07 | -0.24 | -0.04 | 0.01 | 0.18 | 0.03 | 1.00 | 0.03 | 0.06 | 0.01 | -0.04 | -0.00 |
| genre_influence | -0.18 | -0.08 | 0.05 | 0.09 | 0.09 | 0.03 | 0.11 | 0.08 | -0.03 | -0.07 | -0.12 | 0.00 | 0.10 | 0.04 | 0.03 | 1.00 | 0.30 | -0.05 | 0.18 | 0.02 |
| artist_influence | 0.20 | 0.03 | -0.02 | 0.07 | 0.12 | -0.02 | 0.08 | 0.01 | -0.01 | -0.11 | 0.04 | -0.00 | 0.05 | 0.04 | 0.06 | 0.30 | 1.00 | -0.23 | -0.04 | 0.03 |
| featuring | -0.00 | 0.03 | 0.07 | 0.08 | -0.02 | 0.02 | -0.08 | -0.06 | 0.05 | -0.00 | 0.07 | -0.00 | -0.05 | -0.03 | 0.01 | -0.05 | -0.23 | 1.00 | -0.04 | -0.00 |
| playlist | -0.18 | -0.10 | -0.04 | -0.05 | -0.09 | -0.01 | 0.02 | 0.12 | -0.10 | 0.10 | -0.16 | 0.00 | 0.03 | -0.01 | -0.04 | 0.18 | -0.04 | -0.04 | 1.00 | -0.06 |
| soundtrack | 0.09 | 0.01 | 0.00 | -0.02 | 0.01 | 0.00 | -0.02 | -0.00 | 0.01 | -0.02 | 0.07 | -0.01 | -0.02 | 0.01 | -0.00 | 0.02 | 0.03 | -0.00 | -0.06 | 1.00 |

- Correlation is reduced for the extracted features and increased for the original music features.

- Instrumentalness, Duration, and Danceability are the most correlated with Popularity out of the original features.

# Feature Engineering

Feature extraction: Artist_influence, Genre_influence, Playlist, Soundtrack, Featuring.

Identifying significant categorical variables:
1st approach: Using ANOVA, we determined that all the categorical features except for Mode are significant in predicting Popularity.

2nd approach: All categorical features except for Featuring are significant in predicting Popularity.

# The Models

**Algorithms: Linear Regression, Random Forest, XGBoost**
**Evaluation metrics: MSE, RMSE, R-Squared**

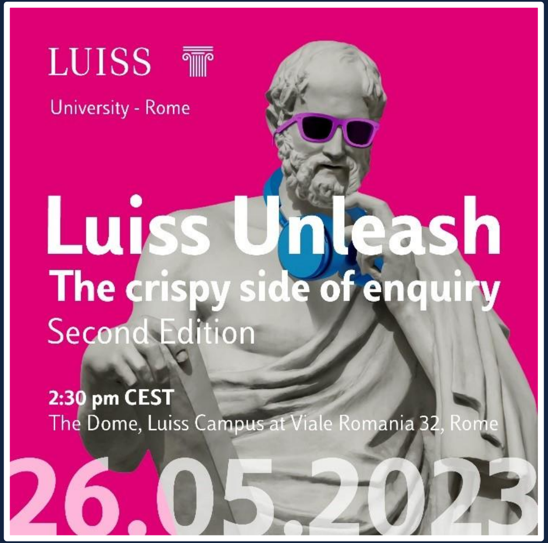
| | Approach 1 | | | Approach 2 | | |
|---|---|---|---|---|---|---|
| | MSE | RMSE | $R^2$ | MSE | RMSE | $R^2$ |
| Linear Regression | 78.77 | 8.875 | 79% | 5.728 | 2.393 | 15% |
| Random Forest | 86.870 | 9.320 | 77% | 2.752 | 1.659 | 60% |
| XGBoost | 71.769 | 8.472 | 81% | 2.578 | 1.606 | 62% |

# Wrap up

- Audio features are not sufficient in predicting popularity due to their overlap across different tracks and genres.
- Genre is the most important predictor of popularity outside of Artist influence.
- The most popular genres used in soundtracks are pop-film, disney, and k-pop.
- The most popular genres added to playlists are jazz, soul, and rock.
- Less popular tracks tend to be added to more playlists for marketing purposes.

- Up and coming artists should be inclined to produce pop-film, chill, or sad songs. They should also collaborate with other musicians since featurings seem to gain more popularity on average.

# Luiss Unleash



LUISS

University - Rome

Luiss Unleash

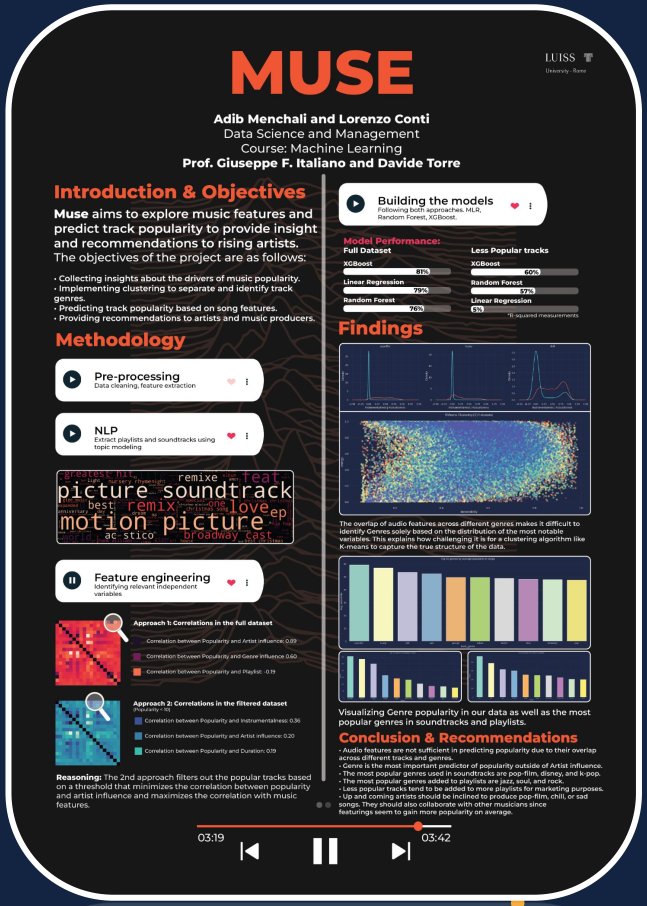
The crispy side of enquiry

Second Edition

2:30 pm CEST

The Dome, Luiss Campus at Viale Romania 32, Rome

26.05.2023



LUISS
University - Rome

# MUSE

Adib Menchali and Lorenzo Conti
Data Science and Management
Course: Machine Learning
Prof. Giuseppe F. Italiano and Davide Torre

## Introduction & Objectives

**Muse** aims to explore music features and predict track popularity to provide insight and recommendations to rising artists.

The objectives of the project are as follows:

- Collecting insights about the drivers of music popularity.
- Implementing clustering to separate and identify track genres.
- Predicting track popularity based on song features.
- Providing recommendations to artists and music producers.

## Methodology

Pre-processing
Data cleaning, feature extraction

NLP
Extract playlists and soundtracks using topic modeling

Feature engineering
Identifying relevant independent variables

**Approach 1: Correlations in the full dataset**

- Correlation between Popularity and Artist influence: 0.89
- Correlation between Popularity and Genre influence 0.60
- Correlation between Popularity and Playlist: -0.19

**Approach 2: Correlations in the filtered dataset**
(Popularity < 10)

- Correlation between Popularity and Instrumentalness: 0.36
- Correlation between Popularity and Artist influence: 0.20
- Correlation between Popularity and Duration: 0.19

**Reasoning:** The 2nd approach filters out the popular tracks based on a threshold that minimizes the correlation between popularity and artist influence and maximizes the correlation with music features.

Building the models
Following both approaches. MLR, Random Forest, XGBoost.

**Model Performance:**

| Full Dataset | | Less Popular tracks | |
|---|---|---|---|
| XGBoost | 81% | XGBoost | 60% |
| Linear Regression | 79% | Random Forest | 57% |
| Random Forest | 76% | Linear Regression | 5% |

*R-squared measurements

## Findings

The overlap of audio features across different genres makes it difficult to identify Genres solely based on the distribution of the most notable variables. This explains how challenging it is for a clustering algorithm like K-means to capture the true structure of the data.

Visualizing Genre popularity in our data as well as the most popular genres in soundtracks and playlists.

## Conclusion & Recommendations

- Audio features are not sufficient in predicting popularity due to their overlap across different tracks and genres.
- Genre is the most important predictor of popularity outside of Artist influence.
- The most popular genres used in soundtracks are pop-film, disney, and k-pop.
- The most popular genres added to playlists are jazz, soul, and rock.
- Less popular tracks tend to be added to more playlists for marketing purposes.
- Up and coming artists should be inclined to produce pop-film, chill, or sad songs. They should also collaborate with other musicians since featurings seem to gain more popularity on average.

# Thank you!