NYU

Mid Program Project Presentation

Fall 2024 Data Science Bootcamp

Topic: Music Recommendation System

Group 2 Members:
Thomas Zhang, Martin Chen, Kehan Lin, David Ren, Ruining Li

Table of Contents

NYU

Objectives

- **Goal**: Recommend songs based on Spotify data, creating personalized listening experience.
- **Focus**: Analyze song characteristics (e.g., audio features, genre), understand user behavior, and apply machine learning techniques
- **Tasks**:
  - **Midterm:** EDA & Data Visualization
  - **Final:** machine learning & recommendation system algorithm

PART 01

Exploratory Data Analysis

Import Packages

- **Downloading and reading datasets directly from Kaggle:**
  - os & kagglehub
- **Data processing and manipulation:**
  - numpy & pandas
- **Data visualization:**
  - matplotlib.pyplot & seaborn
  - plotly (express, graph_objects, make_subplots)
  - Yellowbrick.target (Feature Correlation Plot)

Datasets

kaggle

- **Source:** Spotify Dataset on Kaggle
- **Datasets:**
  - data.csv: 170653 rows, 19 columns
  - data_by_genres.csv: 2973 rows (genres), 14 columns
  - data_by_artist.csv: 28680 rows (artists), 16 columns
  - data_by_year.csv: 100 rows (1921~2020), 14 columns
  - data_w_genres.csv: 28680 rows, 15 columns

	object				int				float		normalized float (0-1)							dummy		others	
	id	name	artists	release_date	year	duration_ms	popularity	key	loudness	tempo	valence	acousticness	danceability	energy	instrumentalness	liveness	speechness	mode	explicit	genres	count
df	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
genres						✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
artist			✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓
year					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
genre			✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓

Variable Classification

- **Sound/Audio Features:** valence, acousticness, danceability, energy, liveness, instrumentalness, speechness
- **Music Composition:** key, mode, loudness, tempo
- **Track Metadata:** id, name, artists, release_date, year, explicit, duration_ms
- **Engagement:** popularity (A measure of the track's popularity, typically based on streams, likes, or other engagement metrics on the platform.)

Variable Definitions

Valence: the musical "positiveness" conveyed by a track.

Acousticness: the likelihood that a track is acoustic.

Danceability: describes how suitable a track is for dancing, based on a combination of tempo, rhythm stability, beat strength, and overall regularity.

Energy: a perceptual measure of intensity and activity in the track.

Liveness: the presence of an audience in the recording

Instrumentalness: the likelihood that a track has no vocals

Speechiness: the presence of spoken words in a track.

Variable Definitions

Explicit: whether a track contains explicit content (1 = explicit, 0 = not explicit).

Year: when the track was released.

Artist: The name(s) of the artist(s) performing the track.

Duration_ms: The duration of the track in milliseconds.

release_date: The date when the track was released.

model: Indicates the modality (major or minor) of the track, with 1 for major and 0 for minor.

name: The title of the track.

Variable Definitions

key: The musical key of the track (e.g., C major, G minor), represented as an integer (0 = C, 1 = C♯/D ♭ , 2 = D, etc.).

id: A unique identifier for the track on Spotify.

loudness: The overall loudness of a track in decibels (dB).

tempo: The speed or pace of a track, measured in beats per minute (BPM).

Variables by Data Type

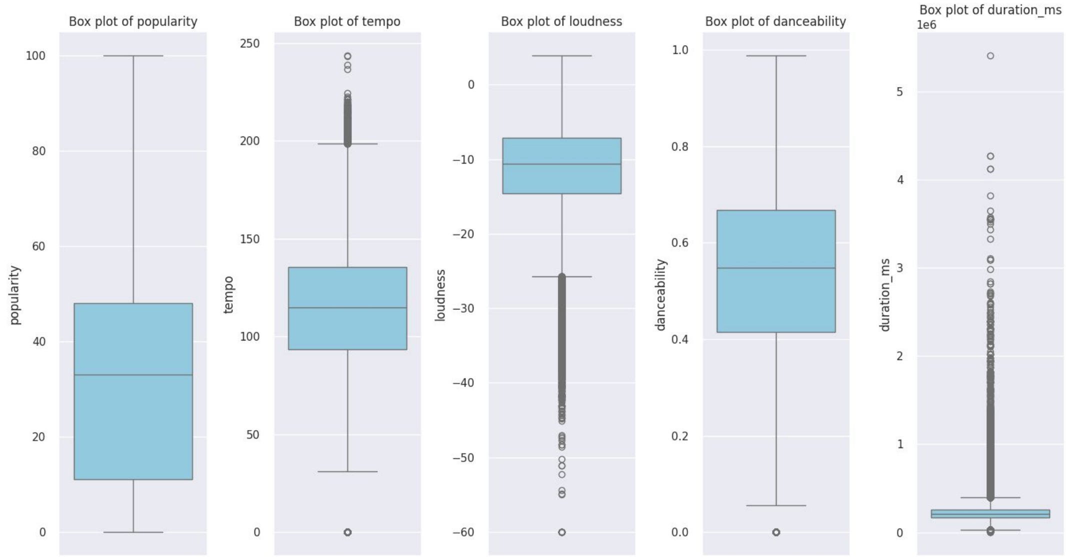
- **Object:** id, name, artists, release_date
- **Integer:** duration_ms, popularity, key
- **Float:** loudness, tempo
- **Normalized Float (0~1):** valence, acousticness, danceability, energy, liveness, speechness, instrumentalness
- **Dummy variables (0/1):** mode, explicit

Variables Processing

- **Potential Target Variable:** Popularity
- **Variables to be dropped:** id
- **Variables to be modified:** duration_ms, loudness, tempo, mode, explicit

Imbalanced data/Outliers

- **Invalid values (N/A):** No null values in all datasets
- **Out-of-bound values (Outliers):** Several variables has outliers
- **Duplicates:** ~543 in data.csv, none in other datasets



NYU

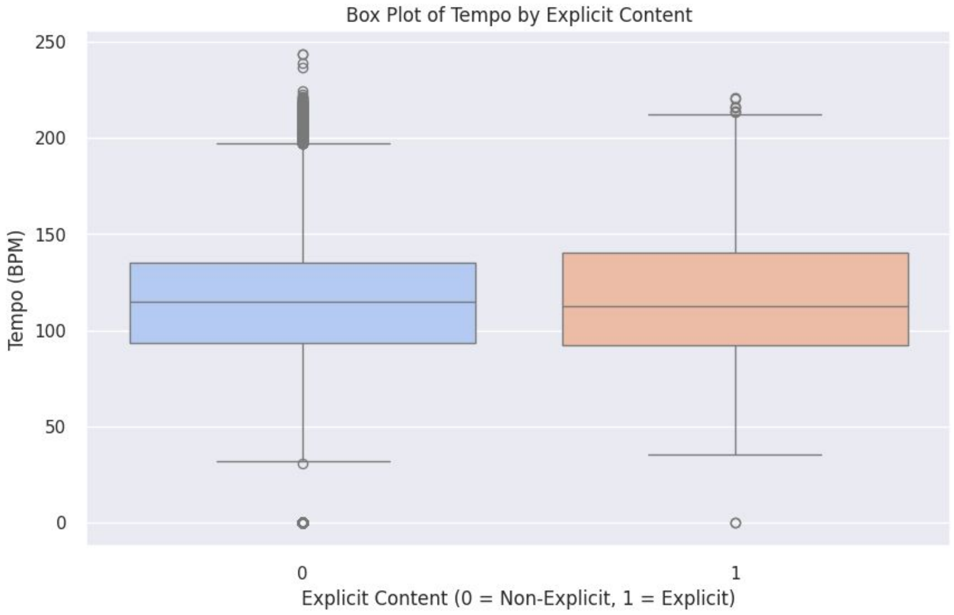
Handlements and Motivations

Motivations:

- Distortion of Centroids
- Increased Cluster Count
- Increased Computational Complexity

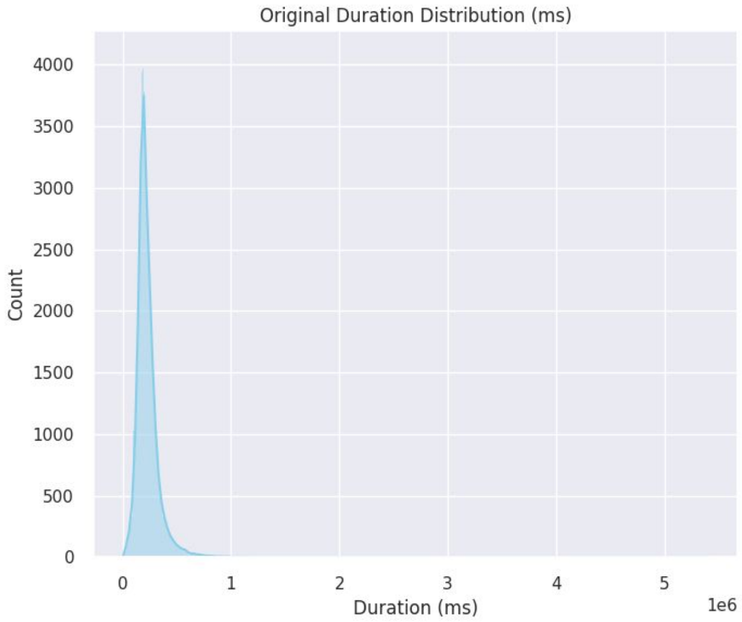
Methods:

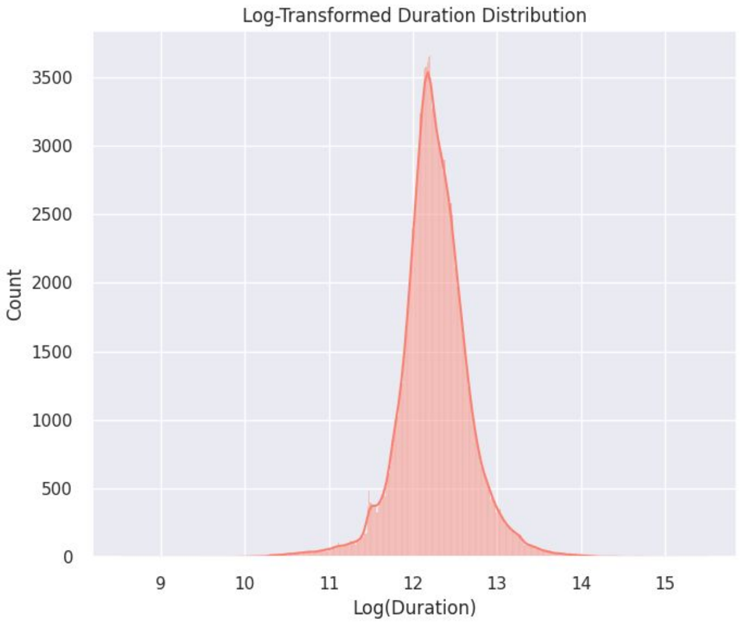
- Justification: Anova check categorical patterns
- Minimum Distortion: If pattern present, handle subgroup separately
- Handle: Capping using quantiles 1% to 99%



ANOVA p-value: 7.62e-07

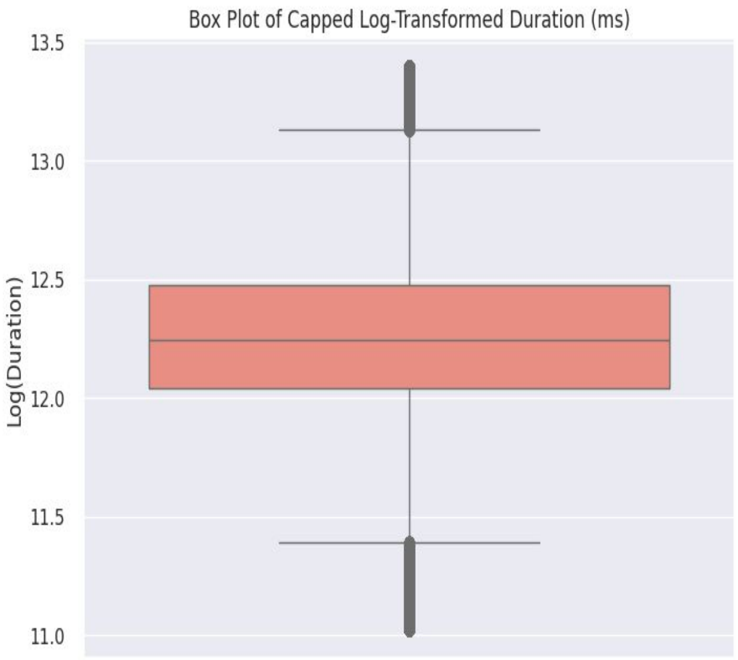
Addressing Outliers: Variable Duration





NYU

Log Transformation for Duration



Discovery: Wide range of track lengths

Possibility: Classical, or Live recordings may be longer

Measured in **milliseconds,** exaggerating the issue

To Cope:

Log Transform: Variance Stabilized correct right skewness

Frequency Change: Change to minutes

Outlier Handlement

Addressing Duplicates

After careful examination, we have noticed:

```
m = spotify_reduced[spotify_reduced.duplicated()]
```

```
m[m['artists'].astype(str) == "['Kendrick Lamar']"]
```

	valence	year	acousticness	artists	danceability	duration_ms
18090	0.0985	2012	0.0152	['Kendrick Lamar']	0.587	310720

Insights: The inherent duplicates may counting incorrect

Given size of duplicates are small

We keep them,avoid dropping unique values

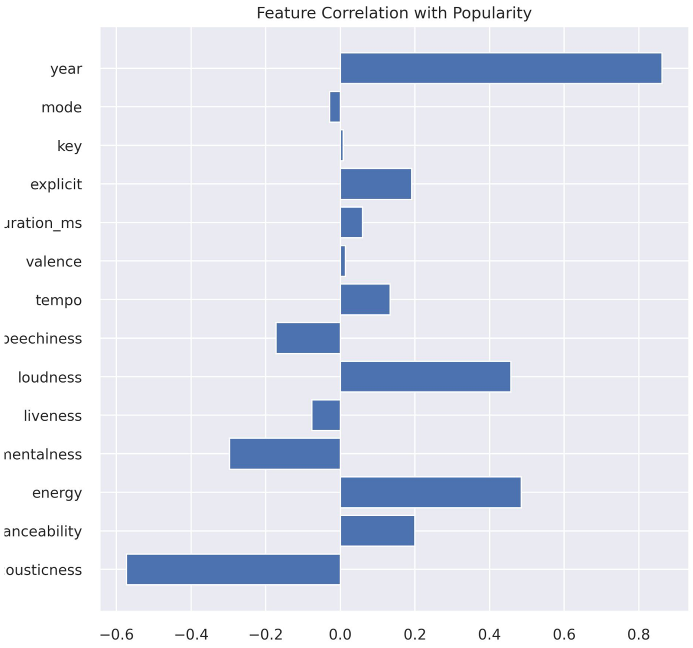
NYU

PART 02

Data Visualization

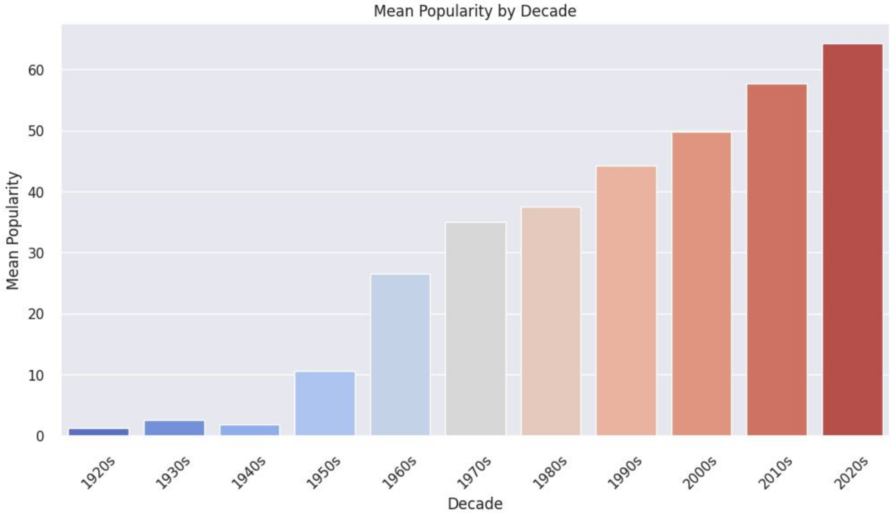
Feature Correlation with Popularity

- **Plotted from Data.csv**
- **Feature Correlation Plot**
- **Why not heatmap?**
  - Too many numeric variables (int, float)
  - Better demonstration of the relationships between numeric variables and Popularity (target variable)



Popularity by Decade

- **Plotted from data.csv**
- **Barplot**
  - Numerical comparison across decades
- **The mean popularity increases every decade…**
  - which makes sense
- **However, it is biased…**
  - from today's perspective



NYU

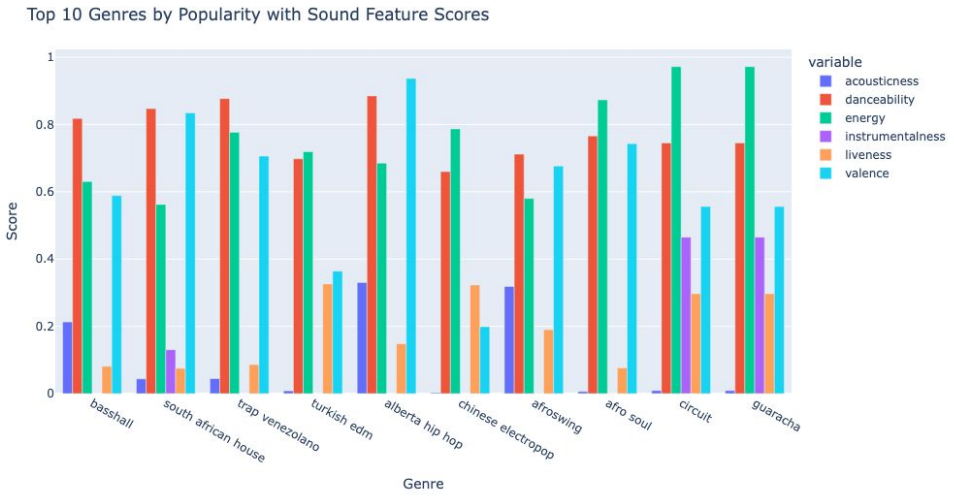
Sound Features Revolution

- **Plotted from data_by_year.csv**
- **Use plotly (go, make_subplots) for interactive graphs**
- **Conclusions:**
  - Energy and danceability increase over time
  - Acousticness and instrumentalness decrease over time
  - Liveness is relatively low and stable
  - Valence fluctuates (no clear pattern) but relatively high

Top 10 Genres

- **Plotted from data_by_genres.csv**
- **By popularity**
- **With Sound Features**
- **Used plotly (px)**
- **Matches the trends observed from "Sound Features Revolution"**

- Energy and danceability high
- Acousticness and instrumentalness low
- Liveness generally low
- Valence generally high



NYU

Data_by_artist & data_w_genres

- **Have the same row number (28680 rows)**
- **Could potentially be merged by the count function**
    - assign weights for score calculation
    - for further ranking and recommendations

PART 03

Next Steps

Problems to be Solved

- **Problems with Dataset/Models:**
  - e.g. Popularity Skewness

- **Problems with Recommendation System:**
  - Cold Start (e.g., What songs should we recommend for a user without knowing his/her preferences)

Thanks for Listening!

NYU