## 



Taylor Swift Stream Prediction

15.03.2024

**─**

Arun, Wei Ming

[Lab Exercise 7 Sample](https://docs.google.com/document/u/0/d/1M6n7XuQ3SByzWOPsgO_SmO6IQtCTG-EsGj2E461XZf4/edit) [Canva Slides](https://www.canva.com/design/DAGATz5ltyM/8nzbBmvleU0j76xFnNcPDQ/edit?utm_content=DAGATz5ltyM&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)

# Problem Statement

Predict future stream counts based on song features (e.g., danceability, energy, speechiness).

# 1. Sample Collection and Practical Motivation

## General Description

There are 530 data points

There are 18 columns (including serial number) in the original dataset.

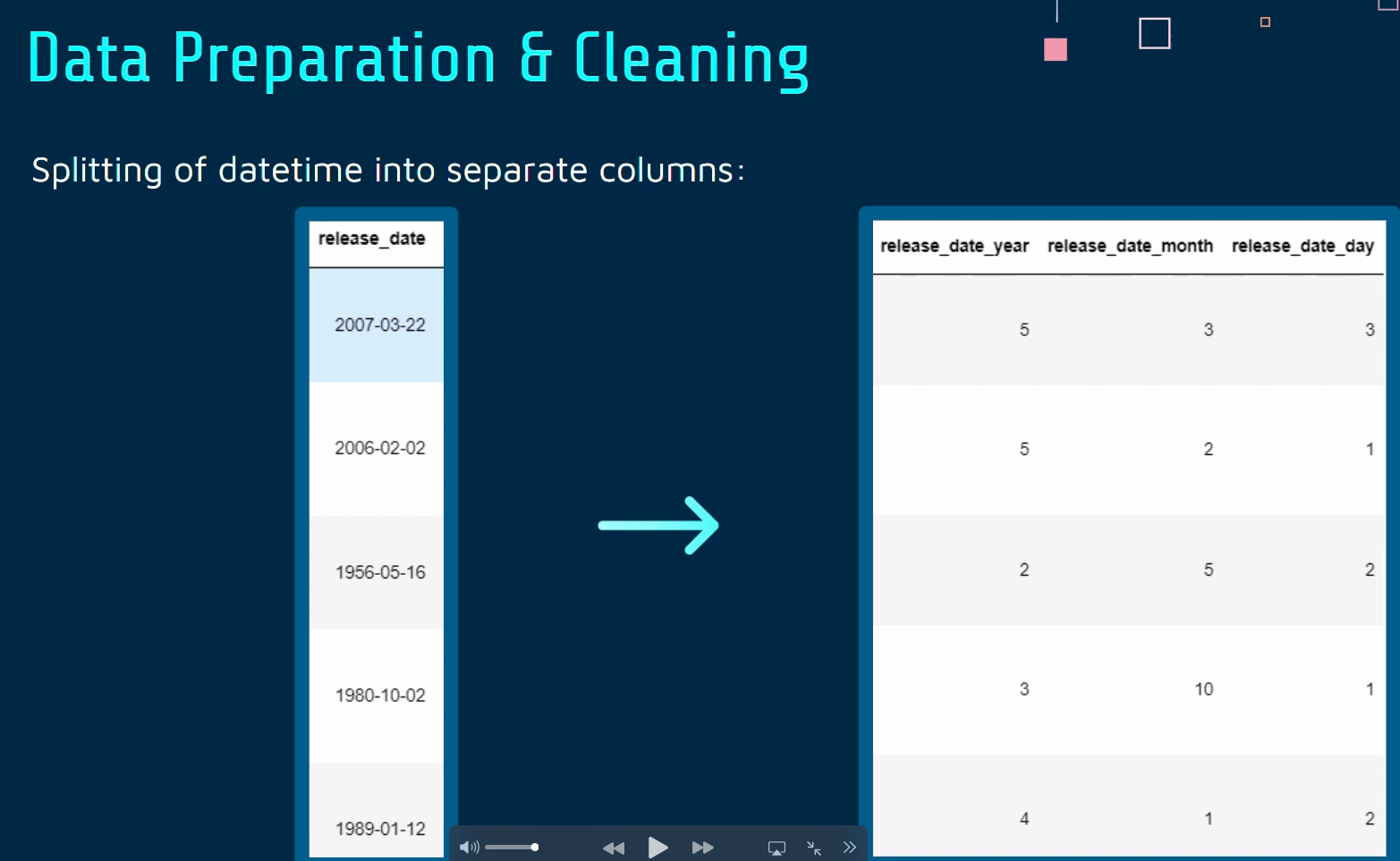
## Existing Elements for Prediction (12 elements + serial number)

* **Acousticness**: A measure of how acoustic the song is.
* **Danceability**: Describes how suitable a track is for dancing.
* **Energy**: A measure of intensity and activity.
* **Instrumentalness**: Indicates the absence of vocals.
* **Liveness**: Detects the presence of an audience in the recording.
* **Loudness**: The overall loudness of a track in decibels.
* **Speechiness**: Identifies the presence of spoken words in a track.
* **Tempo**: The overall estimated tempo of a song in beats per minute (BPM).
* **Valence**: Describes the musical positiveness conveyed by a track.
* **Popularity**: The popularity of the song on Spotify (used with caution, as noted).
* **Release Date**: The date the song was made available on Spotify.
* **Duration**: The total length of the track, in milliseconds.
* (Serial Number)

## Newly Created Elements for Prediction (2 elements) //use statistical exploration tools to check which data to be dropped

* **Mood Score**: A new feature combining valence, energy, and tempo to represent the song's overall mood. This score might indicate a song's potential to be streamed based on its emotional and energetic appeal. A simple way to create this could be by averaging these features, or by more complex weighted combinations if certain aspects are deemed more important.
* **Rhythm Index**: Another new feature combining danceability and tempo to reflect how the song's rhythm aligns with listener activities or preferences. This index could help identify songs that are more likely to be played in certain settings or activities. Like the Mood Score, it could be a simple average or a more nuanced combination depending on the importance of each feature.

## Elements Used for Prediction (9 elements)



The final list of elements used for prediction includes both the existing elements and the newly created elements:

* Acousticness
* Instrumentalness
* Liveness
* Loudness
* Speechiness
* Release Date
* Duration
* **Mood Score** (Newly Created)
* **Rhythm Index** (Newly Created)

The Popularity feature, although listed, should be used with caution primarily for exploratory analysis rather than direct prediction, to avoid circular reasoning as it is a result of streaming behavior.

# 2. Visualization and Pattern recognition (Combined Steps)

# 2. Data Preparation and Problem Formulation (not showing)

# To show relations or correlations among these features and potential stream counts (not directly available in the dataset), you can use Pandas for data manipulation and preparation. For correlation, use the .corr() method in Pandas to find pairwise correlations among the numerical features.

# 3. Exploratory Analysis and Statistical Description (graph/relation)

# Use histograms, scatter plots, and box plots to understand the distribution and relationships of your features. matplotlib and seaborn are great libraries for this purpose.

# 4. Analytic Visualization and Pattern Recognition (show how they are related)

# For visualizing patterns, consider plotting feature importance or using pair plots to identify relationships between song features and their popularity as a proxy for streams.

# 5. Algorithmic Optimization and Machine Learning

# For predicting future stream counts, you can start with regression models since your target variable (streams) is continuous. A good starting point would be:

# Linear Regression: To establish a baseline.

# Random Forest Regressor: For capturing non-linear relationships without much hyperparameter tuning.

# Gradient Boosting Machines (e.g., XGBoost, LightGBM): To potentially improve prediction accuracy by capturing complex patterns in the data.

# 6. Information Presentation and Statistical Inference

# Visualize the importance of different features in predicting stream counts. This could involve plotting feature importances from tree-based models or coefficients from linear models.

# 7. Consideration and Intelligent Decision

# Evaluate the models based on metrics like RMSE (Root Mean Square Error) or MAE (Mean Absolute Error) to understand their accuracy. Use these insights to make informed decisions about song promotion strategies or to understand what song features are most predictive of streaming success.

# 

# 

# Rubrics

* ~~10% for coming up with your own problem definition based on a dataset~~
* **Cultural Impact Studies**: Offers a lens to study cultural trends and shifts in music preferences over time.
* ~~10% for data preparation and cleaning to suit the problem of your choice~~
* 20% for exploratory data analysis/visualization to gather relevant insights
  + ~~EDA visualisation~~
  + EDA Comments & Comparisons
* 20% for the use of machine learning techniques to solve specific problems
  + ~~Machine Learning~~
  + Comments and improvement if needed
* 20% for the presentation of data-driven insights and the recommendations
  + Recommendations
  + Presentation and Statistical Inference
* 10% for the quality of your final team presentation and overall impressions
* ~~10% for learning something new and doing something beyond this course~~
  + ~~Random Forest Regressor Model~~
  + ~~XGBoost Regressor Model~~

# Milestones

| Event | Due Date | Status | Allocation | Link | Remarks |
| --- | --- | --- | --- | --- | --- |
| Learn Github | Mar 16, 2024 | Completed |  |  |  |
| Step 1 | Mar 23, 2024 | In Progress |  |  |  |
|  | Mar 16, 2024 | Not Started |  |  |  |
|  | Date | Not Started |  |  |  |
|  | Date | Not Started |  |  |  |
|  | Date | Not Started |  |  |  |

# References

## Data pipeline



## Video

INTRODUCTION ON WHY WE ARE TAKING THE TOPIC

MOTIVATION

REAL WORLD APPLICATION

ELEMENTS THAT AFFECTS OUR OUTCOME

INTRODUCING THE DATASET

INTRODUCING THE LEGITIMACY OF OUR DATASETS

EXPLAIN THE NUMBER OF DATAPOINTS

ANY ANOMALIES

DATA ANALYSIS - VISUAL REPRESENTATION

KEY FEATURES OF DATASET

MACHINE LEARNING