**Overview of the TMDB TV Shows Dataset**

The **TMDB TV Shows Dataset (2023)** contains detailed metadata about **150,000+ TV shows**, including information about their **genres, ratings, popularity, production companies, and more**. This dataset is useful for **analyzing trends in the TV industry, understanding viewer preferences, and making data-driven recommendations**.

**1️ Data Size and Format**

* **Size:** ~150,000 records
* **Format:** **Semi-Structured (JSON)**
  + The dataset is stored in **MongoDB**, which uses **BSON (Binary JSON)**.
  + Each document has **structured fields** (e.g., id, name, vote\_average) and **semi-structured fields** (e.g., genres, production\_companies, spoken\_languages).
  + Some fields contain **lists or nested values**, making it **semi-structured** rather than strictly structured.

**2️ Key Features/Columns**

| **Field Name** | **Description** |
| --- | --- |
| \_id | Unique MongoDB Object ID |
| id | Unique TMDB TV Show ID |
| name | Title of the TV show |
| first\_air\_date | Date when the show premiered |
| last\_air\_date | Date when the last episode aired |
| number\_of\_seasons | Total seasons of the TV show |
| number\_of\_episodes | Total episodes of the TV show |
| original\_language | Language of the show (e.g., en, es, ko) |
| vote\_average | Average rating of the show (IMDb-style) |
| vote\_count | Number of user ratings |
| popularity | Popularity score based on TMDB algorithm |
| genres | Genres (e.g., "Drama, Crime, Sci-Fi") |
| status | Current status (e.g., Returning Series, Ended) |
| networks | TV network(s) that aired the show (e.g., Netflix, HBO) |
| created\_by | Name of the creator(s) |
| origin\_country | Country where the show was produced |
| spoken\_languages | Languages spoken in the show |
| production\_companies | Studios that produced the show (e.g., Warner Bros.) |
| episode\_run\_time | Average runtime of an episode (in minutes) |

**Storing Process:**

1 Created Cluster on MongoDB - MADSClass

A screenshot of a computer

AI-generated content may be incorrect.

2 Created Collection and Table – BIG\_DATA and TMDB\_RAW to store raw data.A screenshot of a computer

AI-generated content may be incorrect.

3 Uploaded data from Kaggle Dataset to MongoDB using mongoimport:

Kaggle Dataset URL: *https://www.kaggle.com/datasets/asaniczka/full-tmdb-tv-shows-dataset-2023-150k-shows?resource=download*

*mongoimport --uri "mongodb+srv://madsclass.rgddjin.mongodb.net/BIG\_DATA" --collection TMDB\_RAW --type csv --headerline --file /Users/birajgtm/tmdb\_shows.csv --username birajgtm --password birajgtm*

A screenshot of a computer

AI-generated content may be incorrect.

4 Created new collection called TMDB\_PROCESSED to save processed data and copied RAW\_DATA to TMDB\_PROCESSED for preprocessing:A screenshot of a computer program

AI-generated content may be incorrect.

**Preprocessing:**

1 Removed unwanted columns:

A screenshot of a computer program

AI-generated content may be incorrect.

2 Convert Boolean Fields from string to Boolean

A screenshot of a computer program

AI-generated content may be incorrect.

3 Convert Date Strings (first\_air\_date, last\_air\_date) to ISODate

A screenshot of a computer program

AI-generated content may be incorrect.

4 Convert genres from String to Array

A screenshot of a computer program

AI-generated content may be incorrect.

5 Convert episode\_run\_time to integer and handle nulls

A screenshot of a computer program

AI-generated content may be incorrect.

6 Convert vote\_average to float

A screenshot of a computer program

AI-generated content may be incorrect.

7 Cap popularity values (Remove Outliers) A computer screen shot of a program code

AI-generated content may be incorrect.

8 Check and remove duplicate records

A screen shot of a computer program

AI-generated content may be incorrect.

9 Fix missing genres and origin\_country

A screenshot of a computer program

AI-generated content may be incorrect.

10 Verify Preprocessed Data

A screenshot of a computer

AI-generated content may be incorrect.

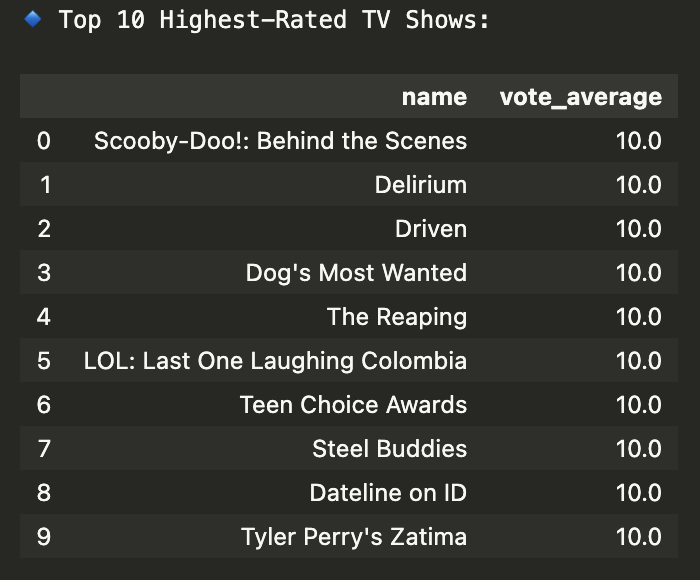
**Analysis:**

**Business Questions:**

What are the top 10 highest-rated TV shows?

How many TV shows are produced per country?

What are the most common TV show genres?

1. What are the top 10 highest-rated TV shows?
2. How many TV shows are produced per country?

A screenshot of a computer

AI-generated content may be incorrect.

1. What are the most common genres?

A screen shot of a computer

AI-generated content may be incorrect.

**Preparing and exporting data as CSV for MySql to proceed ETL Pipeline Design (SSIS):**

1 Exporting Data to CSV

**A screenshot of a computer

AI-generated content may be incorrect.**

2 Preview of the exported file:

**A screen shot of a computer program

AI-generated content may be incorrect.**