1. Introduction

The IMDb Ratings dataset provides user ratings and vote counts for a wide range of movies and TV shows. The dataset is crucial for understanding audience preferences, analyzing rating distributions, and conducting data-driven decision-making in the entertainment industry.

This report aims to **profile, assess, clean, and transform** the dataset to ensure high data quality and readiness for analysis.

2. Dataset Overview

The IMDb Ratings dataset contains **1,541,709** records and **3 columns**, providing details about **average user ratings and the number of votes received for various movie and TV show titles**.

Dataset Structure

Column Name	**Data Type**	**Description**
tconst	**String**	Unique IMDb identifier for a title
averageRating	**String**	IMDb average rating (1-10 scale)
numVotes	**String**	Number of votes the title received

3. Data Profiling & Quality Assessment

Missing Values Check

- **No missing values** were found in any column.
- **No \N placeholders** were detected.

Data Type Issues

- **averageRating & numVotes were initially stored as strings**, which is incorrect.
- **Solution:** Convert averageRating to **FLOAT**, and numVotes to **INTEGER**.

Duplicate Records Check

- tconst is **unique** across all records.
- **No duplicate records** found.

Statistical Summary & Key Insights

Column	**Min**	**Max**	**Mean**	**Standard Deviation**
averageRating	**1.0**	**10.0**	~6.9	Moderate variance
numVotes	**1**	**Millions**	**Few thousand**	Highly skewed

- **Right-skewed distribution**: Most movies have **low votes**, while a few titles receive **millions of votes**.
- **Outliers detected**: Some titles have very few votes, which may not be reliable.

4. Field-Level Summary

Each column is analyzed for its distribution, uniqueness, and validity.

tconst (Unique Title Identifier)

- **100% unique values**
- No null or duplicate values
- Used as the **primary key** for joining with other IMDb datasets.

averageRating (IMDb Rating: 1-10 Scale)

- **Range: 1.0 - 10.0**
- **Mean rating ~6.9**, indicating most movies are rated between **5.5 and 8.0**.
- **Outliers detected**: Some titles have **extreme ratings (1.0 or 10.0)**, which could indicate bias.

numVotes (Number of Votes per Title)

- **Highly skewed distribution**: Some movies have only **a few votes**, while blockbusters have **millions**.
- **Outliers detected**: Titles with fewer than **100 votes** may not be reliable.

5. Data Quality Assessment

Quality Check	**Assessment**	**Resolution**
Missing Values	No missing values	No action needed
Incorrect Data Types	averageRating & numVotes as strings	Convert to float & integer
Duplicate Records	No duplicates found	No action needed

Quality Check	**Assessment**	**Resolution**
Outliers in numVotes	High variance in votes	Consider filtering movies with < 100 votes

6. Data Cleaning & Transformation Plan

Convert Data Types

- Convert averageRating **String → FLOAT**.
- Convert numVotes **String → INTEGER**.

Handle Outliers

- **Low vote counts (numVotes < 10)** may indicate **unreliable ratings**.
- **Recommended Action:** Filter out titles with fewer than **100 votes**.