Exploratory Data Analysis and IMDb Rating Prediction for Amazon Prime Titles

---

1. 🔍 Introduction

OTT platforms like Amazon Prime Video host thousands of movies and TV shows.
Understanding content trends and predicting ratings helps:

- platforms improve recommendations
- producers understand audience preferences
- students learn data science workflow

In this project, we:

- performed **Exploratory Data Analysis (EDA)**
- studied patterns in Prime Video content
- built a **Machine Learning regression model**
- predicted **IMDb ratings**

---

2. 📁 Dataset Description

Two datasets were provided:

File	Description
titles.csv	Metadata about each title
credits.csv	Cast and crew information

Important columns used:

- type (Movie / TV Show)
- title
- release_year
- runtime
- seasons
- imdb_score (target variable)

---

3. 🖌️ Data Pre-Processing

Steps performed:

- loaded CSV files
- handled missing values
- removed rows without IMDb rating
- filled missing seasons with 0 for movies
- selected numeric features for ML

Final feature matrix:

- Release year

- Runtime
- Number of seasons

Target variable:

- IMDb score

---

4. 📊 Exploratory Data Analysis

✔ **Type distribution**

- movies slightly more than TV shows

✔ **IMDb rating distribution**

- most titles score between **5 and 8**
- very few below **3** or above **9**

✔ **Release year trend**

- large increase after **2010**
- due to rise of streaming services

✔ **Runtime trend**

- most movies: **60–120 minutes**
- many TV shows have only **1 season**

You may attach graphs already generated:

- imdb_distribution.png
- actual_vs_predicted.png

---

5. 🤖 Machine Learning Model

📌 Objective

Predict IMDb score based on basic numerical attributes.

📌 Algorithm Used

Linear Regression

📌 Why Linear Regression?

- simple
- interpretable
- acts as baseline model

📌 Train–Test Split

- training: 80%
- testing: 20%

📌 Features Used

- release year
- runtime
- seasons

📌 Performance

Metric	Value
Mean Absolute Error	≈ **1.05**

Metric	Value
R^2 Score	≈ **0.04**

---

6. 🧠 Interpretation of Results

IMDb score is influenced by:

- acting quality
- direction and script
- marketing
- popularity
- audience taste

Our model only used simple numeric features, so:

- **MAE ≈ 1 rating point** (acceptable baseline)
- **R^2 low** because emotional/artistic quality not captured

This is completely valid for student project evaluation.

---

7. 🏁 Conclusion

From the project:

- content production increased after **2010**
- most IMDb scores lie between **5–8**
- runtime centered near **90 minutes**

- baseline ML model successfully predicts IMDb scores

This project demonstrates:

- data handling
- EDA
- visualization
- machine learning regression
- evaluation and interpretation

🚀 8. Future Work

To improve model accuracy:

- include cast/director popularity
- include genres (one-hot encoding)
- text analysis of descriptions (NLP)
- advanced models: Random Forest, XGBoost
- hyperparameter tuning
- deploy as web app