Assignment 2: Data Preparation & Machine Learning

1. Objective

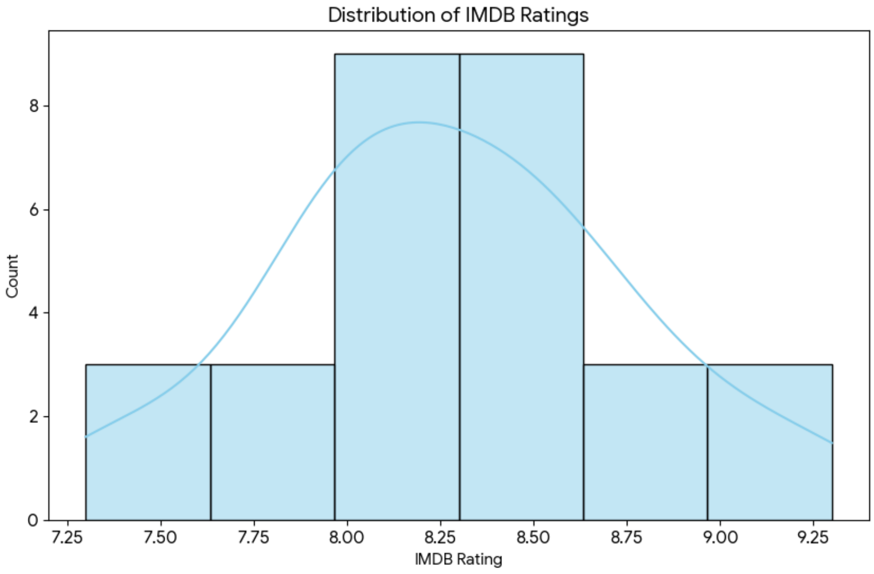
The objective of this assignment is to clean and prepare the movie dataset created in Assignment 1 for machine learning, and then apply a foundational ML algorithm (Linear Regression) to predict movie ratings.

2. Data Preprocessing & Feature Engineering

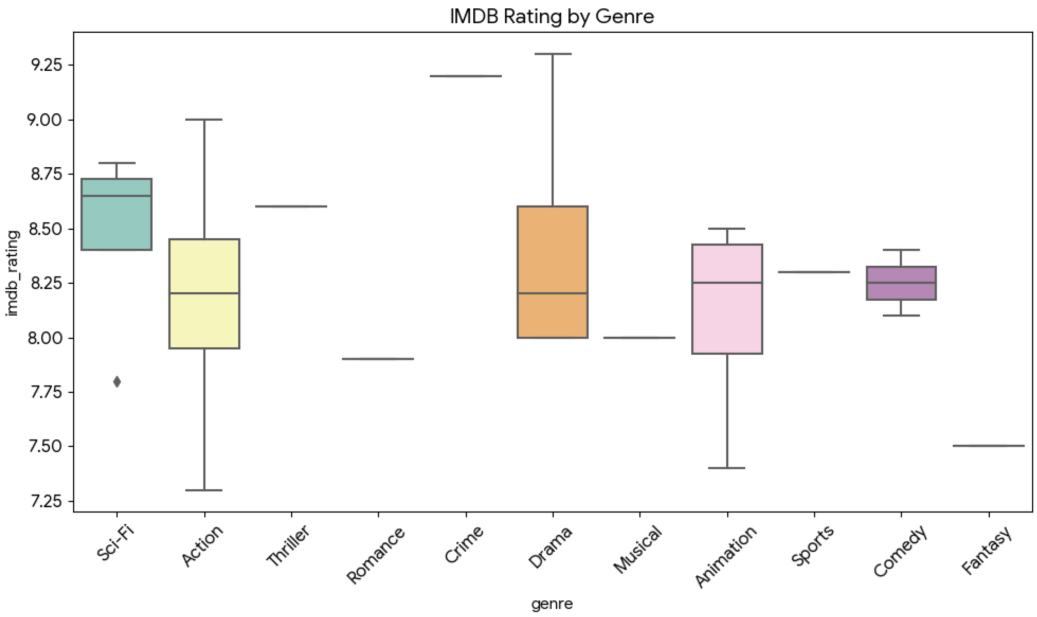
- Data Cleaning: Verified no missing values were present.
- Feature Selection: Dropped 'movie_id' and 'movie_title' as they are unique identifiers.
- Feature Engineering: Created 'movie_age' from 'release_year' (2024 - release_year).
- Encoding: Applied One-Hot Encoding to the 'genre' categorical variable.

3. Exploratory Data Analysis

Distribution of Target Variable (IMDB Rating):



Relationship between Genre and Rating:


IMDB Rating by Genre
imdb_rating
9.25
9.00
8.75
8.50
8.25
8.00
7.75
7.50
7.25
Sci-Fi
Action
Thriller
Romance
Crime
Drama
Musical
Animation
Sports
Comedy
Fantasy
genre

4. Model Selection

A Linear Regression model was selected as the foundational algorithm. The data was split into 80% training and 20% testing sets.

5. Model Evaluation

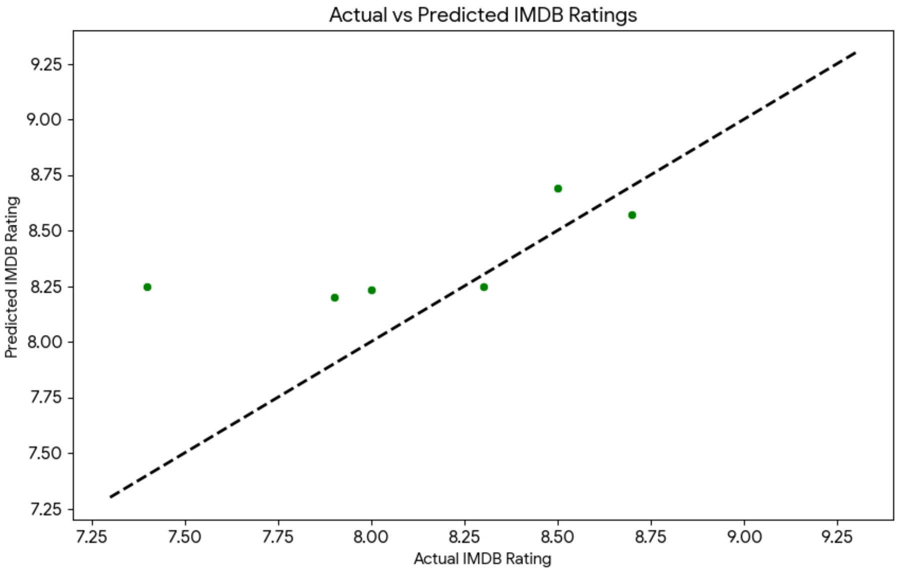
Mean Absolute Error (MAE): 0.2926

Mean Squared Error (MSE): 0.1537

Root Mean Squared Error (RMSE): 0.3920

R2 Score: 0.1566

Actual vs Predicted Values:



6. Conclusion

The model provides a baseline for predicting movie ratings. Given the small dataset size, the variance is high, but the pipeline demonstrates the essential steps of ML: cleaning, encoding, splitting, training, and evaluating.