**PHASE 4 SUBMISSION**

**APPLIED DATA SCIENCE**

**PROJECT NAME-IMDB SCORE PREDICTION**

**FEATURE ENGINEERING:**

Feature engineering is an essential step when building a machine learning model to predict IMDb scores for Netflix original films. It involves creating new features or modifying existing ones to improve the model's predictive performance. Here are some feature engineering ideas for IMDb score prediction

1. Seasonality: Extract the release month or season (e.g., summer, winter) from the 'Premiere' date. Seasonality might affect IMDb scores.

2. Director Popularity: Create a feature that represents the popularity or average IMDb scores of directors. You might use external data sources or group directors into categories based on their past work.

3. Actor Popularity: Similar to directors, you can create features related to the popularity or past IMDb scores of lead actors.

4. Film Age: Calculate the age of the film by subtracting the release year from the current year. Older films might have different IMDb score trends.

5. Film Length Categories: Categorize movies by their length, such as "Short," "Medium," and "Long." You can also create binary features for specific runtime ranges.

6. Average User Rating: If available, calculate the average user rating based on user reviews or ratings.

7. Number of User Ratings: Create a feature that represents the number of user ratings or reviews. A film with more user ratings might have a more reliable IMDb score.

8. Language Categories: Create categories for the language of the films. You can group languages into categories like "English," "Non-English," "Hindi," "Spanish," etc., and encode them as binary features.

**CODING:**

**1.** import warnings

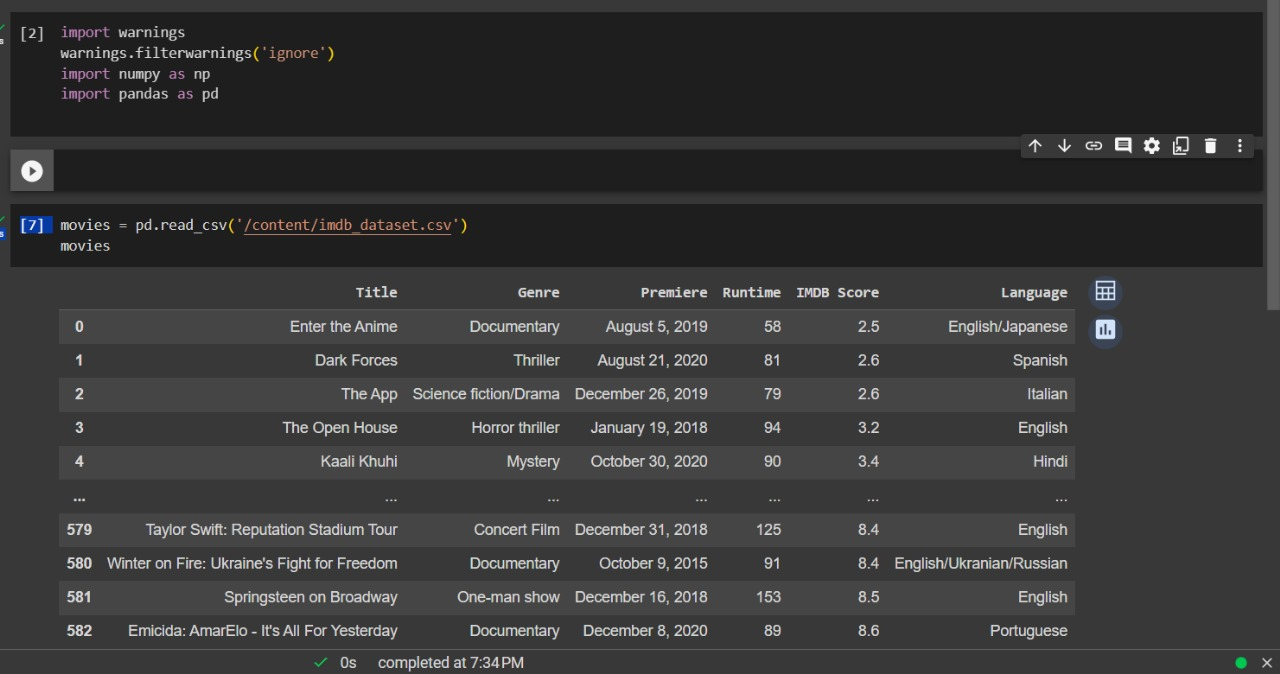
warnings.filterwarnings('ignore')

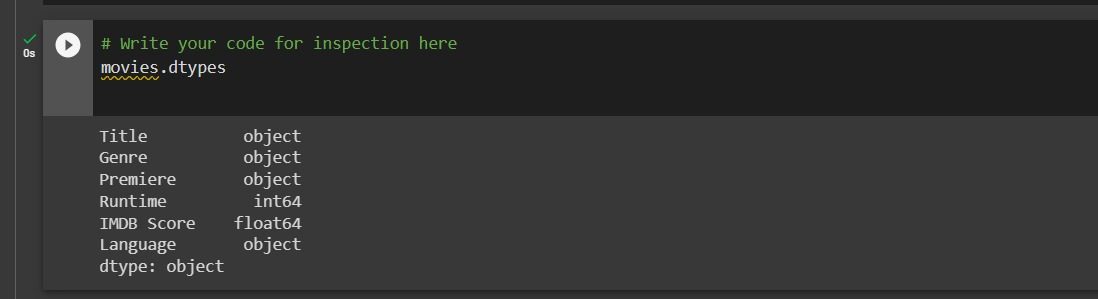
import numpy as np

import pandasaspd

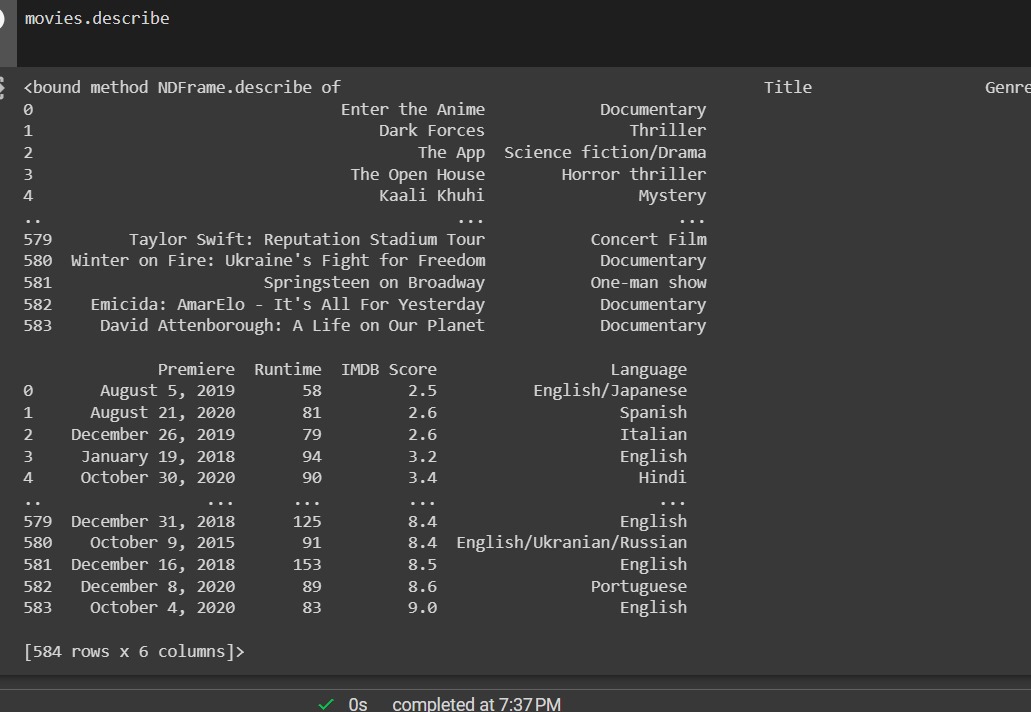
movies = pd.read\_csv('/content/imdb\_dataset.csv')

movies

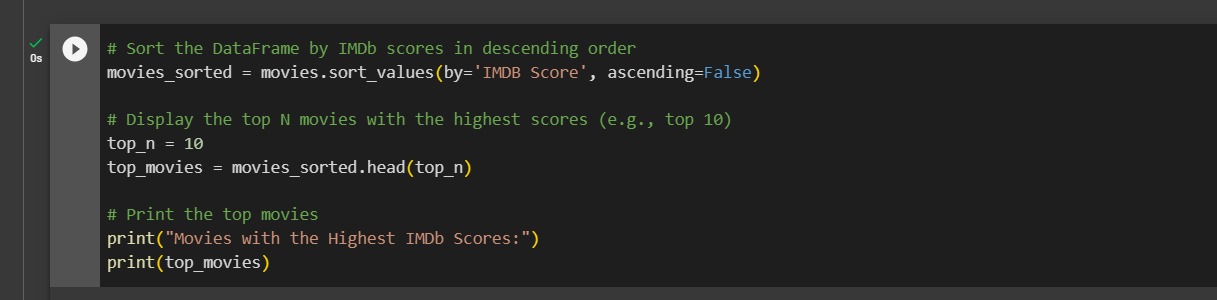


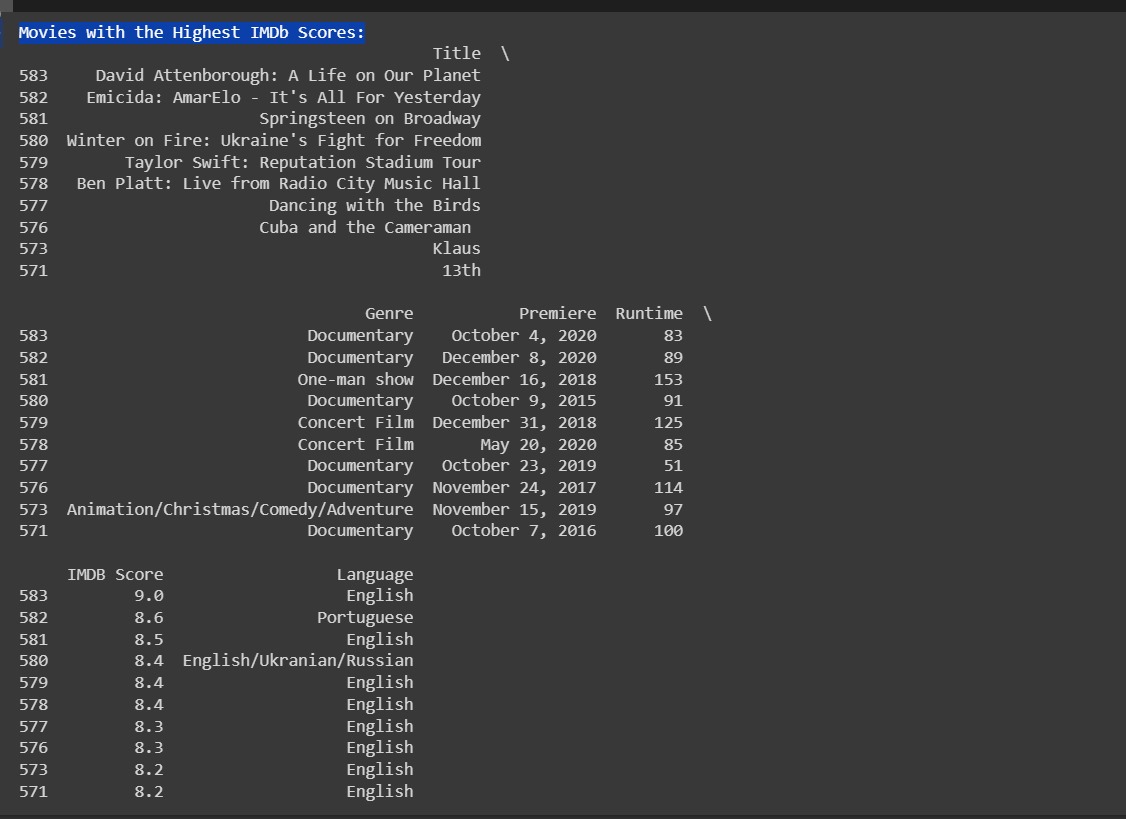
**2.** movies.dtypes

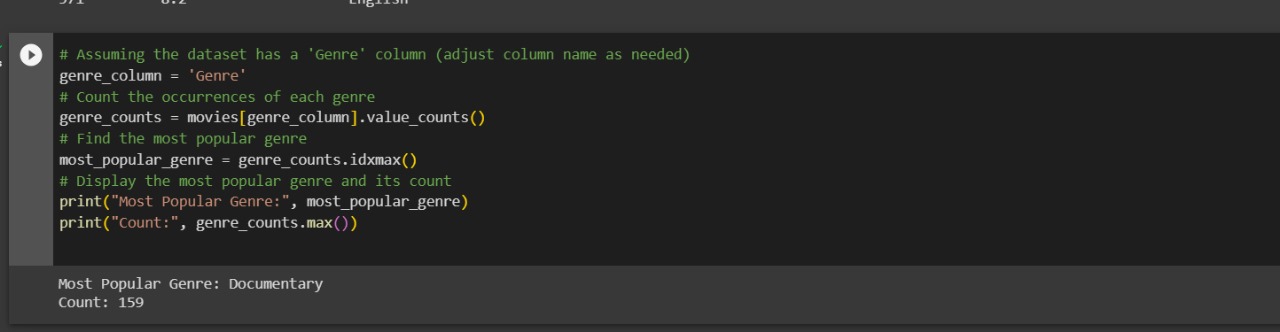
**3.** movies.describe



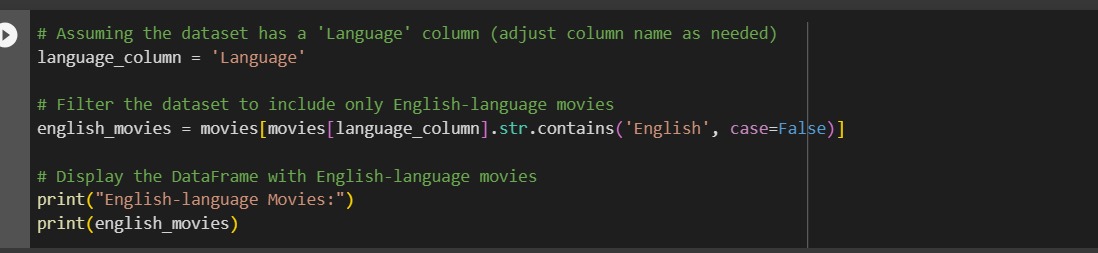
**4.** Movies with highest IMDB SCORES

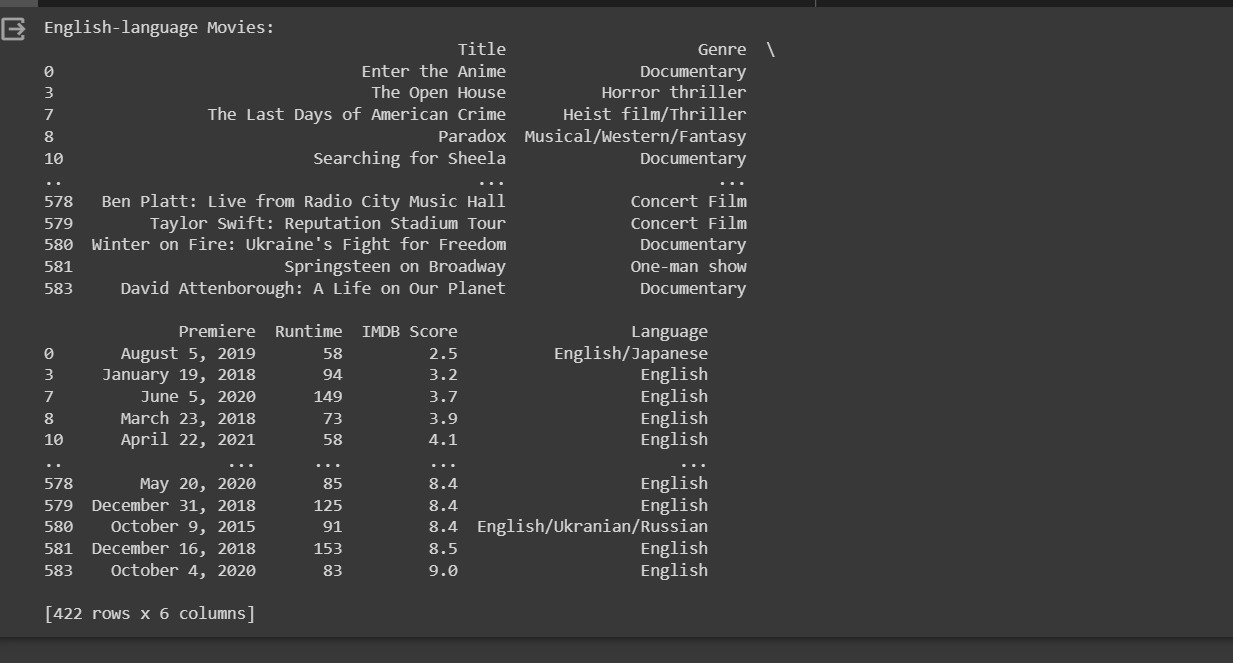




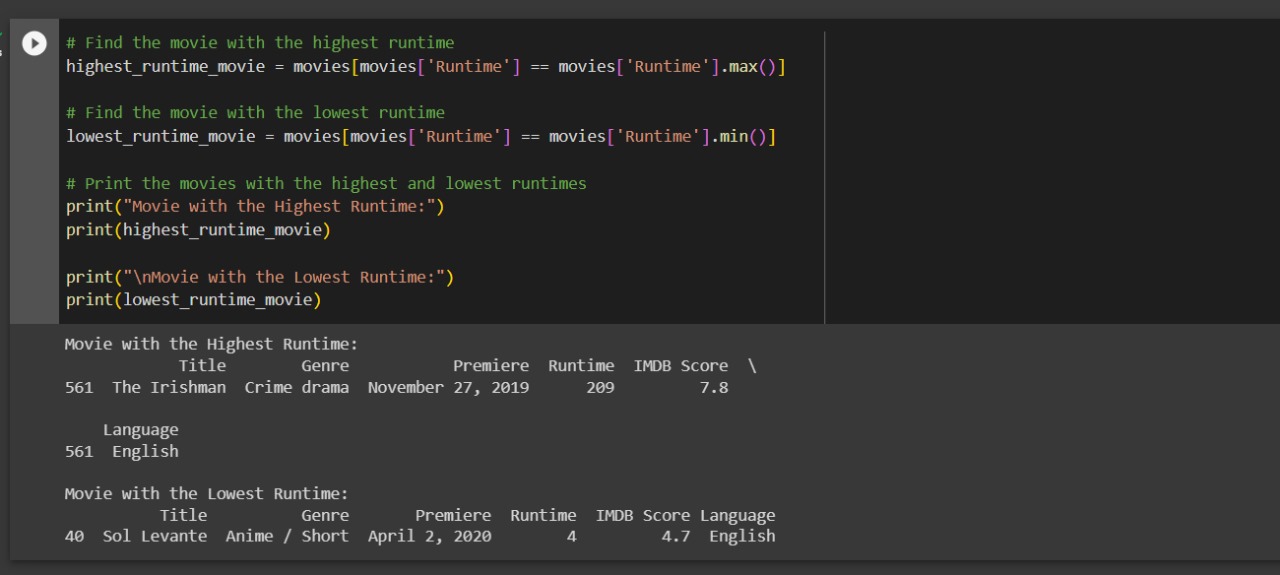
**5.** Most popular genre 

**6.** To find out English language movies

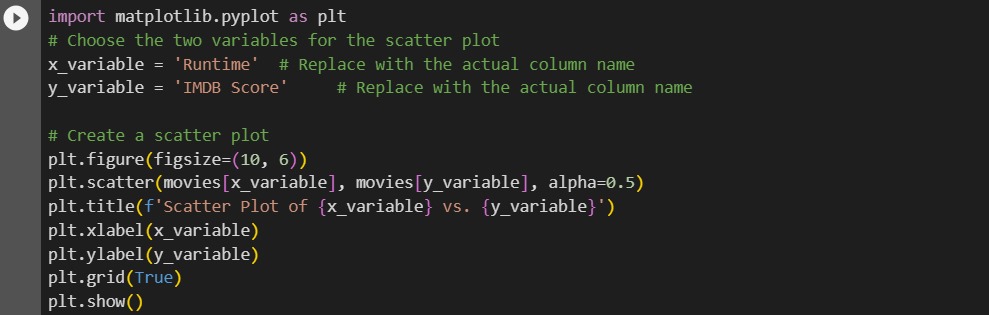


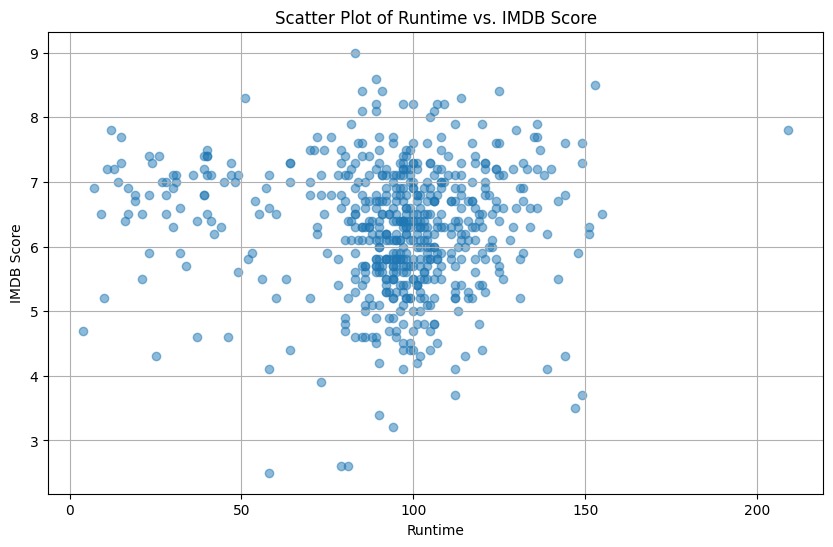


**7.** To find the movies with highest and lowest Run time

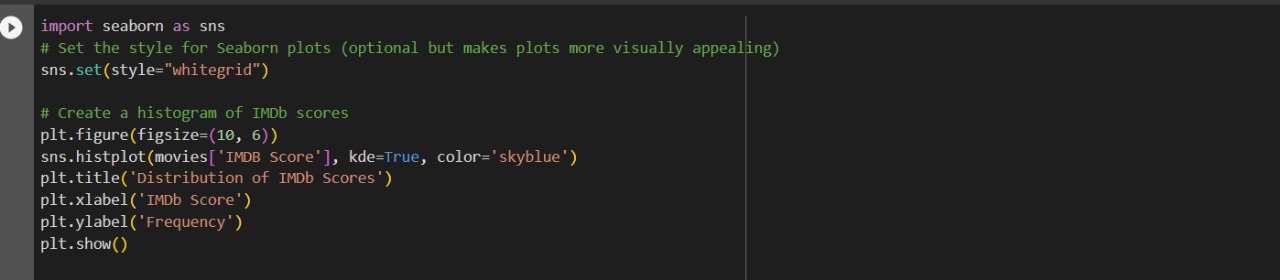


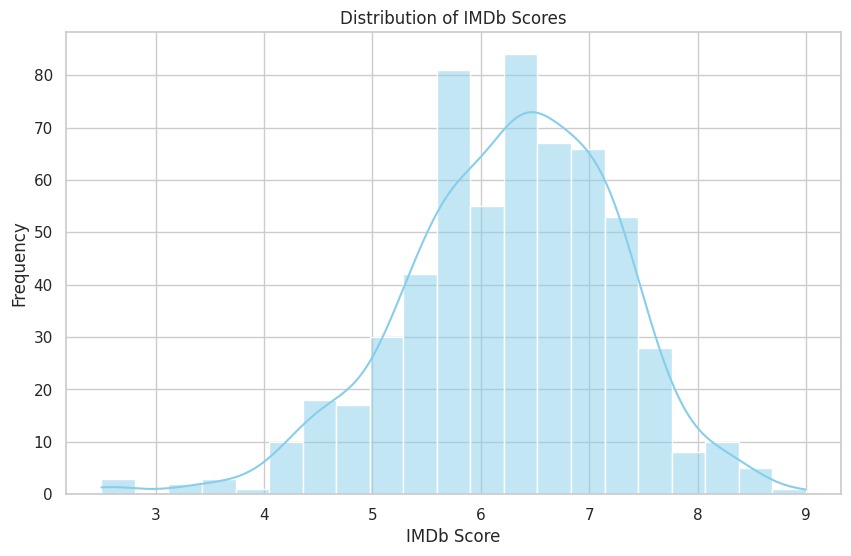
**8.** To Perform Scatter plot





**9.** To Perform histogram for IMDB Scores

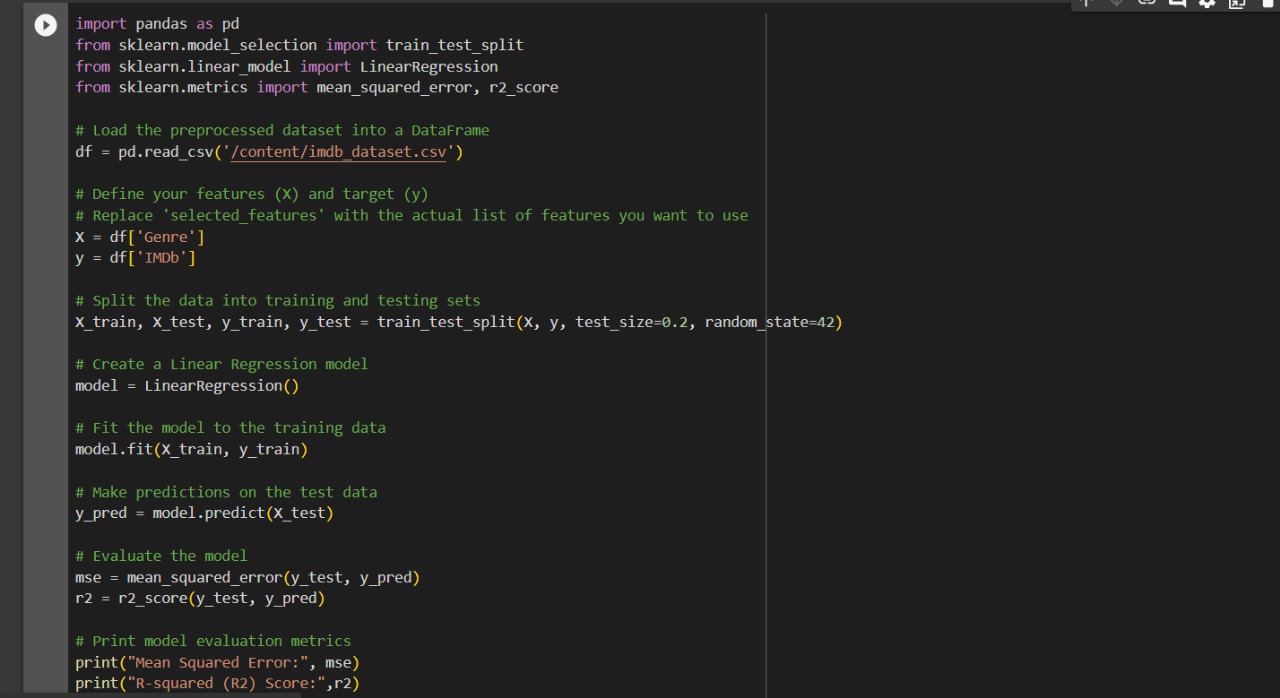




**MODEL TRAINING:**

Model Training involves selecting the right machine learning algorithms,tuning hyperparameters and training the model to predict IMDb Scores

Here we have used linear regression model to train the dataset



**Evaluation :**

Model evaluation is the process of assessing the performance of a machine learning model using various metrics to determine how well it is performing on a specific task. In the case of IMDb score prediction for Netflix original films, you would typically use regression evaluation metrics since IMDb scores are continuous values. Common regression evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared