Predicting IMDB scores using DataScience

Phase 3 submission document

Project Title: predicting IMDB Scores

Phase 3: *Development part 1*

Topic: *Start building the predicting IMDB Scores model by loading and pre-processing the dataset*

Predicting IMDB Scores

Introduction:

- ✓ Predicting IMDb scores is a task that involves using various data-driven techniques and machine learning algorithms to estimate the likely rating or score that a movie or television show is likely to receive on the Internet Movie Database (IMDb).

- ✓ IMDb is one of the most popular and widely-used platforms for user-generated reviews and ratings of films, TV shows, and other entertainment content.

- ✓ Filmmakers and studios can use IMDb score predictions to gauge the potential reception of their upcoming releases. This information can assist in making production and marketing decisions.

- ✓ Online streaming platforms, like Netflix and Amazon Prime, use IMDb scores to recommend content to users. Accurate predictions can improve the effectiveness of these recommendations.

- ✓ IMDb score predictions can also enhance the user experience by allowing viewers to anticipate the quality of content before watching it. Ultimately, predicting IMDb scores is a valuable application of data science and machine learning, providing valuable insights to the entertainment industry and helping viewers make informed choices about the content they consume.

Given Data set:

	Title	Genre	Premiere	Runtime	IMDB Scor	Language
2	Enter the	Document	August 5,	58	2.5	English/Japanese
3	Dark Force	Thriller	August 21	81	2.6	Spanish
4	The App	Science fic	December	79	2.6	Italian
5	The Open	Horror thr	January 19	94	3.2	English
6	Kaali Khuh	Mystery	October 3(	90	3.4	Hindi
7	Drive	Action	Novembe	147	3.5	Hindi
8	Leyla Ever	Comedy	December	112	3.7	Turkish
9	The Last D	Heist film/	June 5, 20	149	3.7	English
10	Paradox	Musical/W	March 23,	73	3.9	English
11	Sardar Ka	Comedy	May 18, 2(	139	4.1	Hindi
12	Searching	Document	April 22, 2	58	4.1	English
13	The Call	Drama	Novembe	112	4.1	Korean
14	Whipped	Romantic	Septembe	97	4.1	Indonesian
15	All Becaus	Action cor	October 1,	101	4.2	Malay
16	Mercy	Thriller	Novembe	90	4.2	English
17	After the I	Document	December	25	4.3	Spanish
18	Ghost Sto	Horror an	January 1,	144	4.3	Hindi
19	The Last T	Political th	February 2	115	4.3	English
20	What Hap	Comedy	January 1,	102	4.3	Korean
21	Death Not	Horror thr	August 25	100	4.4	English
22	Hello Privi	Document	Septembe	64	4.4	English
23	Secret Obs	Thriller	July 18, 20	97	4.4	English
24	Sextuplets	Comedy	August 16	99	4.4	English
25	The Girl o	Thriller	February 2	120	4.4	Hindi
26	Thunder F	Superherc	April 9, 20	105	4.4	English
27	Fatal Affai	Thriller	July 16, 20	89	4.5	English

Necessary step to follows:

1.import Libraries:

Start by importing the necessary libraries:

Program:

```
Import pandas as pd

Import numpy as np

From sklearn.model_selection import train_test_split

From sklearn.preprocessing import StandardScaler
```

2.Load the Dataset:

Load your dataset into a Pandas DataFrame. You can typically find Predicting IMDB scores datasets in CSV format, but you can adapt this code to other formats as needed.

Program:

```
df=pd.read_csv('E:\imdb.csv')
pd.read()
```

3.Exploratory Data Analysis(EDA):

Perform EDA to understand your data better. This includes checking for missing values, exploring the data's statistics, and visualizing it to identify patterns.

Program:

```
# Check for missing values
print(df.isnull().sum())
# Explore statistics
print(df.describe())
# Visualize the data (e.g., histograms, scatter plots, etc.)
```

4. Feature Engineering:

Depending on your dataset, you may need to create new w features or transform existing ones. This can involve one-hot encoding categorical variables, handling date/time data, or scaling numerical features.

Program:

```
# Example: One-hot encoding for categorical variables

df = pd.get_dummies(df, columns=[' Avg.scores ', ' Avg. movie_name '])
```

5.Split the Data:

Split your dataset into training and testing sets. This helps you evaluate your model's performance later.

```
X = df.drop('price', axis=1) # Features

y = df['price'] # Target variable

 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

6. Feature Scaling:

Apply feature scaling to normalize your data, ensuring that all features have similar scales. Standardization (scaling to mean=0 and std=1) is a common choice.

Program:

```
scaler = StandardScaler()

 X_train = scaler.fit_transform(X_train)

X_test = scaler.transform(X_test)
```

Importance of loading and processing dataset:

Loading and preprocessing the dataset is an important first step inbuilding any machine learning model. However, it is especially important for prediction models, as IMDB datasets are often complex and noisy.

By loading and preprocessing the dataset, we can ensure that them machine learning algorithm is able to learn from the data effectively and accurately.

Challenges involved in loading and preprocessing a predicting IMDB dataset:

There are a number of challenges involved in loading and preprocessing a predicting IMDB dataset, including:

- **Handling missing values:**

  Predicting IMDB datasets often contain missing values, which can be due to a variety of factors, such as human error or incomplete data collection. Common methods for handling missing values include dropping the rows with missing values, imputing the missing values with the mean or median of the feature, or using a more sophisticated method such as multiple imputation.

- **Encoding categorical variables:**

  Predicting IMDB datasets often contain categorical features, such as the Genre, Language , and Runtime. These features need to be encoded before they can be used by machine learning models. One common way to encode categorical variables is to use one-hot encoding.

- **Scaling the features:**

It is often helpful to scale the features before training a machine learning model. This can help to improve the performance of the model and make it more robust to outliers. There are a variety of ways to scale the features, such as min-max scaling and standard scaling.

- **Splitting the dataset into training and testing sets:**

Once the data has been pre-processed, we need to split the dataset into training and testing sets. The training set will be used to train the model, and the testing set will be used to evaluate the performance of the model on unseen data. It is important to split the dataset in a way that is representative of the real world distribution of the data.

How to overcome the challenges of loading and preprocessing predicting IMDB dataset:

There are a number of things that can be done to overcome the challenges of loading and preprocessing a IMDB dataset, including:

- **Use a data preprocessing library:**

There are a number of libraries available that can help with data preprocessing tasks, such as handling missing values, encoding categorical variables, and scaling the features.

- **Carefully consider the specific needs of your model:**

The best way to preprocess the data will depend on the specific machine learning algorithm that you are using. It is important to carefully consider the requirements of the algorithm and to preprocess the data in a way that is compatible with the algorithm.

- **Validate the pre-processed data:**

It is important to validate the pre-processed data to ensure that it is in a format that can be used by the machine learning algorithm and that it is of high quality. This can be done by inspecting the data visually or by using statistical methods.

1.Loading the dataset:

- ✓ Loading the dataset using machine learning is the process of bringing the data into the machine learning environment so that it can be used to train and evaluate a model.
- ✓ The specific steps involved in loading the dataset will vary depending on the machine learning library or framework that is being used. However, there are some general steps that are common to most machine learning frameworks:

a. Identify the dataset:

The first step is to identify the dataset that you want to load. This dataset may be stored in a local file, in a database, or in a cloud storage service.

b. Load the dataset:

Once you have identified the dataset, you need to load it into the machine learning environment. This may involve using a built-in function in the machine learning library, or it may involve writing your own code.

c. Preprocess the dataset:

Once the dataset is loaded into the machine learning environment, you may need to preprocess it before you can start training and evaluating your model. This may involve cleaning the data, transforming the data into a suitable format, and splitting the data into training and test sets.

Program:

```
import pandas as pd
 import numpy as np
import seaborn as sns
 import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
 from sklearn.preprocessing import StandardScaler
from sklearn.metrics import r2_score,
mean_absolute_error,mean_squared_error
from sklearn.linear_model import LinearRegression
 from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor
 from sklearn.svm import SVR
 import xgboost as xg
 %matplotlib inline
 import warnings
warnings.filterwarnings("ignore")
 /opt/conda/lib/python3.10/site-packages/scipy/__init__.py:146: : A
NumPy version >=1.16.5 and
<1.23.0 is required for this version for SciPy  (detected version  1.23.5
warnings.warn(f"A NumPy version >={np_minversion}
and<{np_maxversion}"
```

Loading Dataset:

```
dataset = pd.read_csv('E:/imdb.csv')
```

Data Exploration:

Dataset:

Output:

	id	imdb_id	original_title	director	production	genre	cast	budget	revenue	runtime	release_year	vote_count
0	135397	tt0369610	Jurassic World	Colin Trevorrow	Universal Studios	Action	Chris Pratt	150000000	1513528810	124	2015	5562
1	76341	tt1392190	Mad Max Fury Road	George Miller	Village Roadshow Pictures	Action	Tom Hardy	150000000	378436354	120	2015	6185
2	262500	tt2908446	Insurgent	Robert Schwentke	Summit Entertainment	Adventure	Shailene Woodley	110000000	295238201	119	2015	2480
3	140607	tt2488496	Star Wars The Force Awakens	JJ Abrams	Lucasfilm	Action	Harrison Ford	200000000	2068178225	136	2015	5292
4	168259	tt2820852	Furious	James Wan	Universal Pictures	Action	Vin Diesel	190000000	1506249360	137	2015	2947
5	281957	tt1663202	The Revenant	Alejandro Gonzlez Irritu	Regency Enterprises	Western	Leonardo DiCaprio	135000000	532950503	156	2015	3929
6	87101	tt1340138	Terminator Genisys	Alan Taylor	Paramount Pictures	Science Fiction	Arnold Schwarzenegger	155000000	440603537	125	2015	2598
7	286217	tt3659388	The Martian	Ridley Scott	Twentieth Century Fox Film Corporation	Drama	Matt Damon	108000000	595380321	141	2015	4572
8	211672	tt2293640	Minions	Kyle BaldaPierre Coffin	Universal Pictures	Family	Sandra Bullock	74000000	1156730962	91	2015	2893
9	150540	tt2096673	Inside Out	Pete Docter	Walt Disney Pictures	Comedy	Amy Poehler	175000000	853708609	94	2015	3935

2.Preprocessing the dataset:

Data preprocessing is the process of cleaning, transforming, and integrating data in order to make it ready for analysis. ϖ

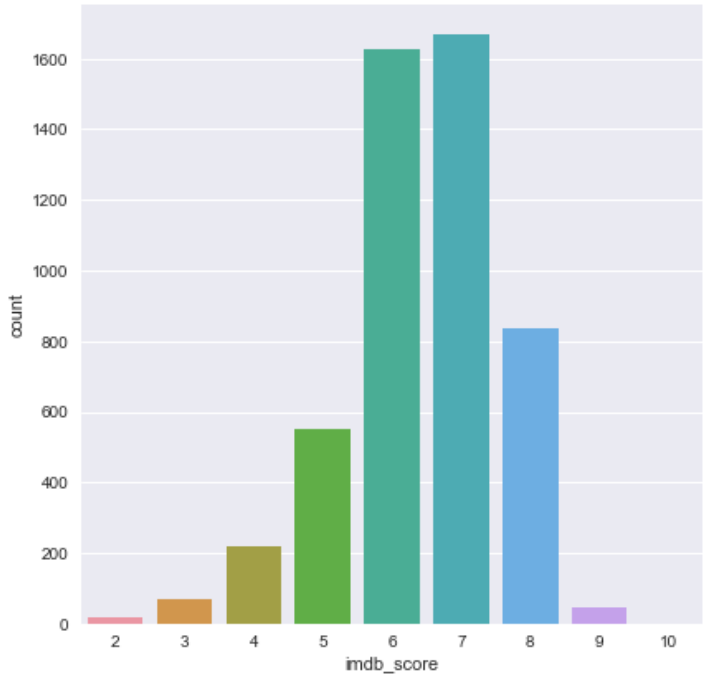
This may involve removing errors and inconsistencies, handling missing values, transforming the data into a consistent format, and scaling the data to a suitable range.

Visualisation and Pre-Processing of Data:

In[1]:

```
sns.histplot(dataset, x='Price', bins=50, color='y')
```

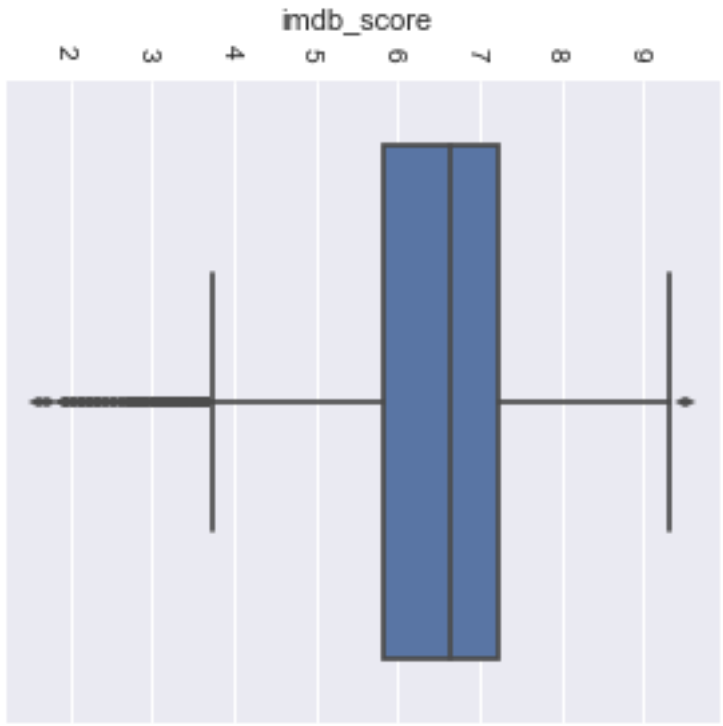
Out[1]:



In[2]:

sns.boxplot(dataset, x='Price', palette='Blues')

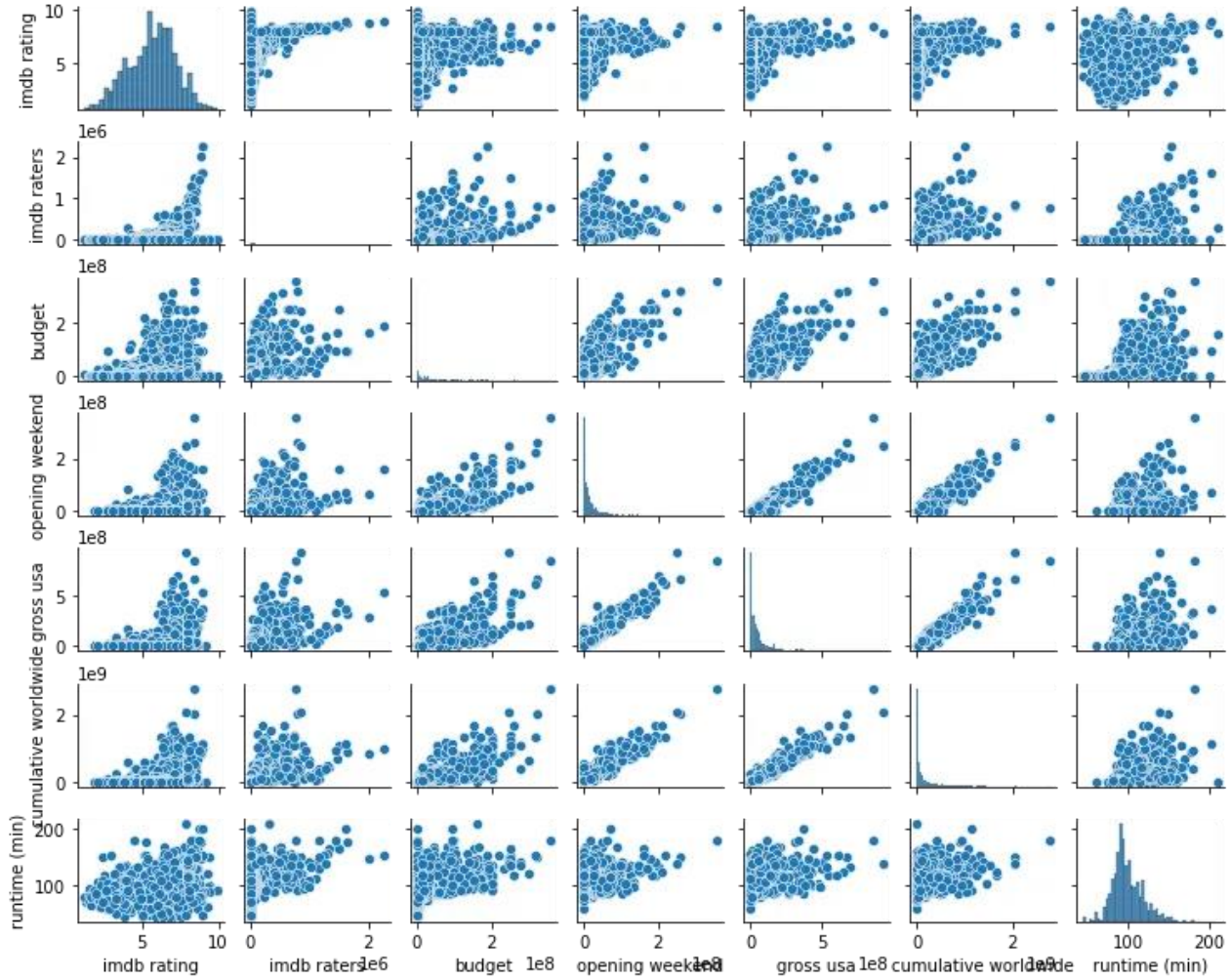
Out[2]:



In [3]:

plt.figure(figsize=(12,8))sns.pairplot(dataset)

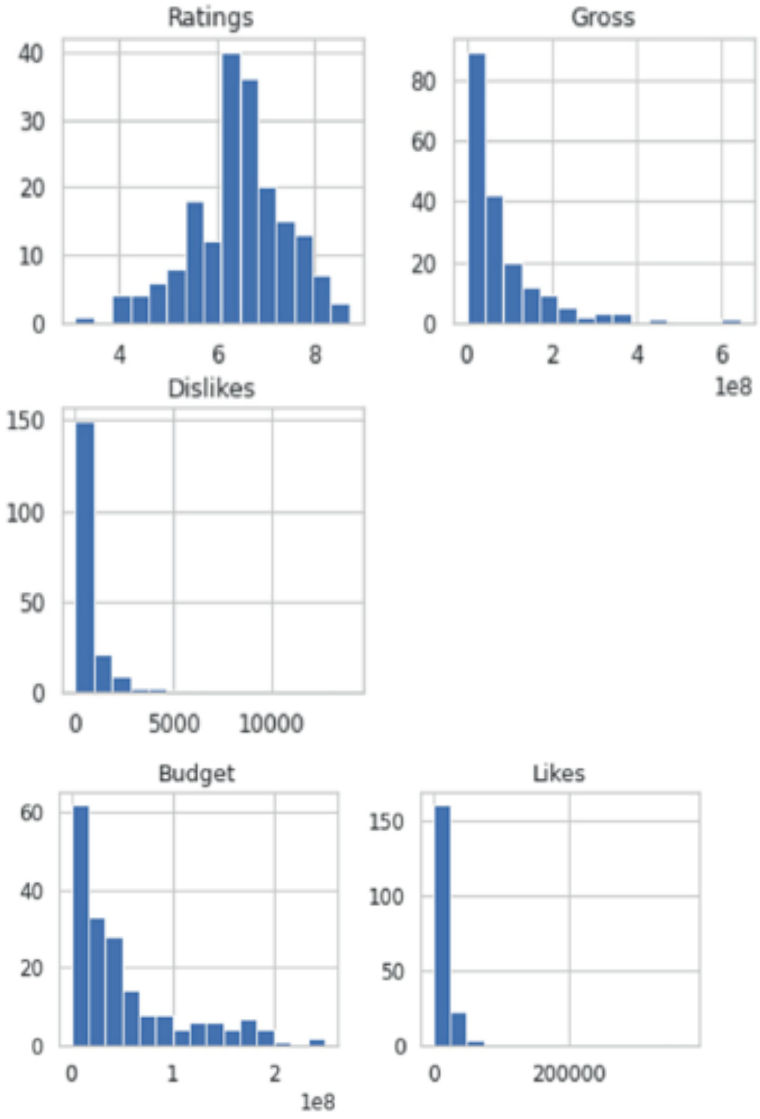
Out[3]:



In [4]:

dataset.hist(figsize=(10,8))

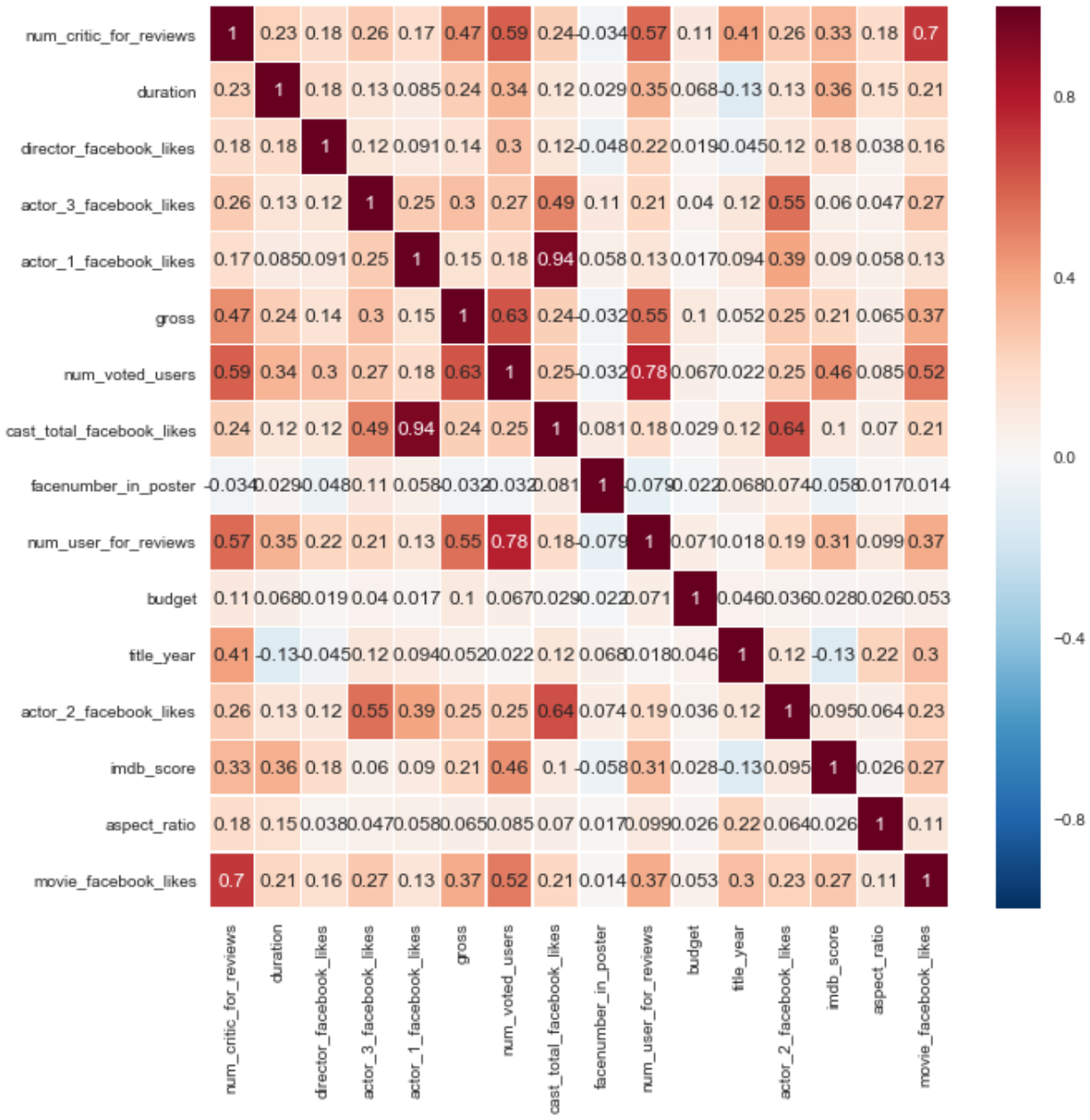
Out[4]:



In [5]:

plt.figure(figsize=(10,5))sns.heatmap(dataset.corr(numeric_only =True), annot=True)

Out[5]:



Some common data preprocessing tasks include:

- **Data cleaning:** This involves identifying and correcting errors and inconsistencies in the data. For example, this may involve removing duplicate records, correcting typos, and filling in missing values.
- **Data transformation:** This involves converting the data into a format that is suitable for the analysis task. For example, this

may involve converting categorical data to numerical data, or scaling the data to a suitable range.

- **Feature engineering:** This involves creating new features from the existing data. For example, this may involve creating features that represent interactions between variables, or features that represent summary statistics of the data.
- **Data integration:** This involves combining data from multiple sources into a single dataset. This may involve resolving inconsistencies in the data, such as different data formats or different variable names.

Conslusion:

- ✓ The accuracy of your IMDB score prediction model hinges on the quality and comprehensiveness of your dataset. A well-curated dataset is crucial for reliable predictions.
- ✓ Choosing the right features to predict IMDB scores is critical. Aspects like genre, cast, director, budget, and user reviews play a vital role in determining a movie's rating.
- ✓ Different predictive models, including linear regression, decision trees, and machine learning algorithms, can be applied based on the dataset and problem complexity.
- ✓ IMDB score predictions offer practical value for filmmakers and studios by providing early insights into a movie's potential success and helping in marketing and distribution strategies.
- ✓ Keep in mind that IMDB scores are inherently subjective and can be influenced by various factors, including user biases. Predictions should be regarded as informed estimates rather than definitive ratings.

In conclusion, predicting IMDB scores is a valuable tool for filmmakers and enthusiasts, but it should complement rather than replace human judgment. It provides data-driven insights into a movie's potential success, helping the industry make informed decisions and adapt to changing audience preferences.