Predicting Movies' Gross Earnings using Linear Regression

Project MVP by: Himani Kaushik

The goal of the project is to create a regression model that will predict US movies' gross after considering several features. The features include IMDb rating, certification, duration of the movie, number of votes and metascore. Anyone interested in financing or investing in the movie, including individual producers, film studios, private investors, etc.

The dataset used in the project was webscraped from IMDb website, using BeautifulSoup. There were 3000 movies with 9 features initially. After the initial data cleaning and EDA, data was analyzed to understand the correlation among different features. Also, visualization packages were used to explore relationships between different features and the target variable.

In the initial phase of the project, linear regression was explored. The data was split into 60 percent training, 20 percent validating and 20 percent testing and gave the following results:

```
Linear Regression Training Score: 0.3892859394128815
Linear Regression Validation Score: 0.46115197525657414
```