Vote Predict of Movie

Introduction

IMDb (an acronym for Internet Movie Database) is an online database of information related to films, television programs, home videos, video games, and streaming content online including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

Problem Statement

The Internet Movie Database (IMDb) is one of the world's most entertaining popular sources of movie, TV and celebrity content with over 100 million unique visitors each day. IMDb has a huge collection of movie database which includes various classification such as action, classic drama, horror and comedy The idea of our project is to web scraping the data from IMDb and form an analysis that helps the data analyst or production company decide how to go about making a new movie, and the second is to create a model to .predict the feelings of movies based on user opinions

Data Description

We are used a web scraping from IMBD and extracted data form the site for 1000 rows and 10 columns.

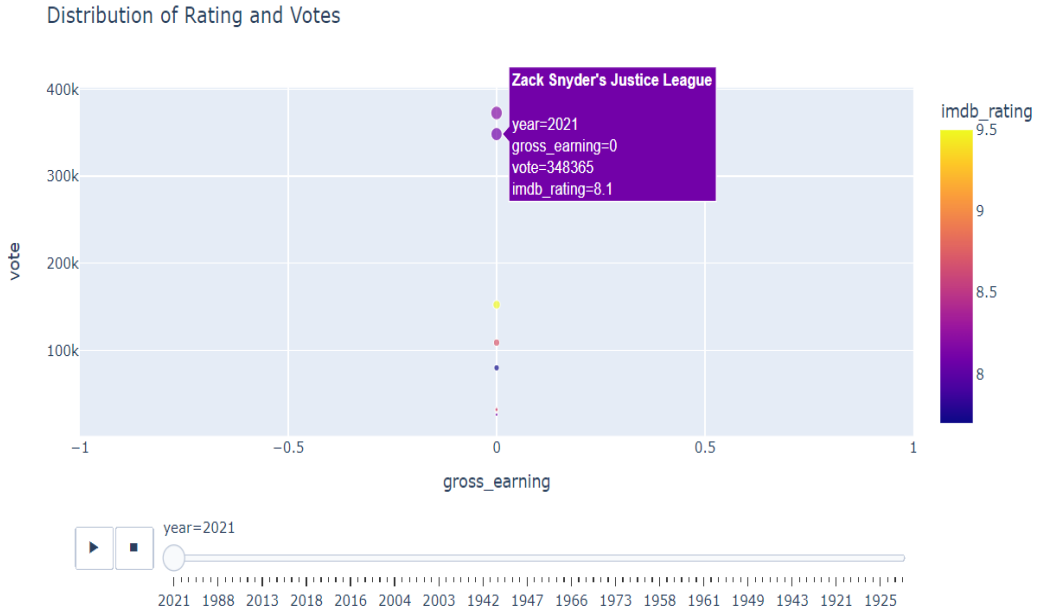
Tools

Technologies: python, Jupyter notebook

Libraries: BeautifulSoup, requests, pandas, numpy, matplotlib, seaborn, , LogisticRegression , sklearn.linear_model, sklearn.model_selection, sklearn.preprocessing

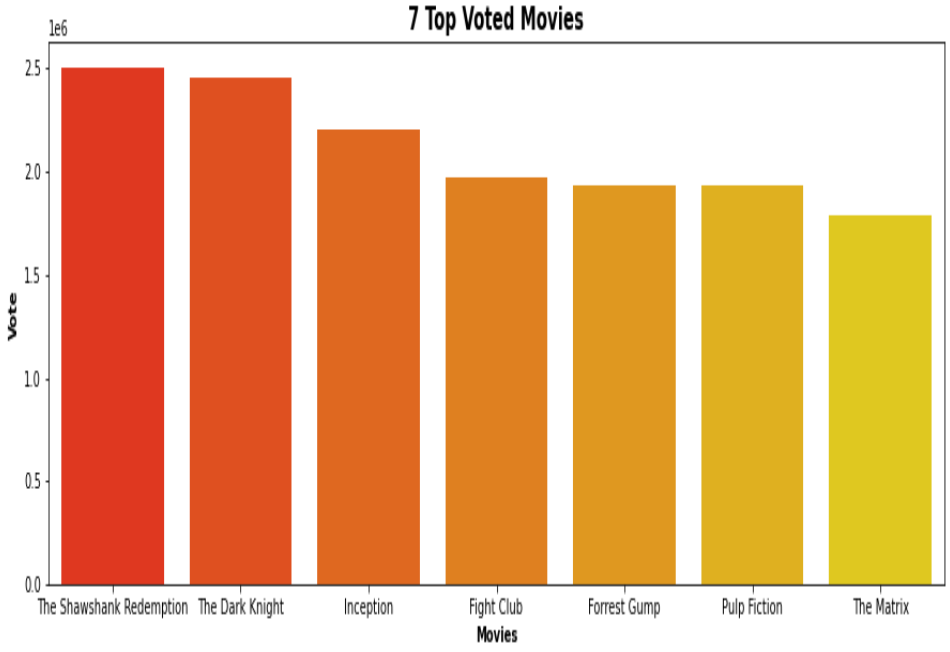
What is the distribution of Rating and Votes for each years?

We noticed in this diagram the distribution between rating and votes in each year



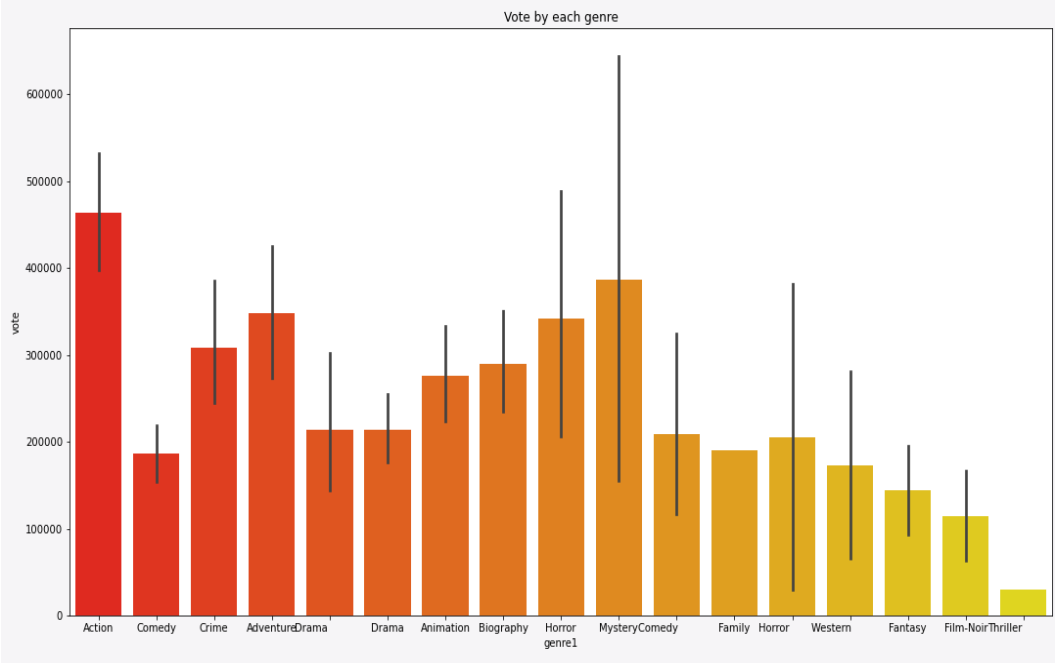
What is the top 7 Movies according to highest votes?

We noticed the best movie vote the movie Shawshank redemption, which was 25% and the lower vote movie it is the matrix which was 1.5%.



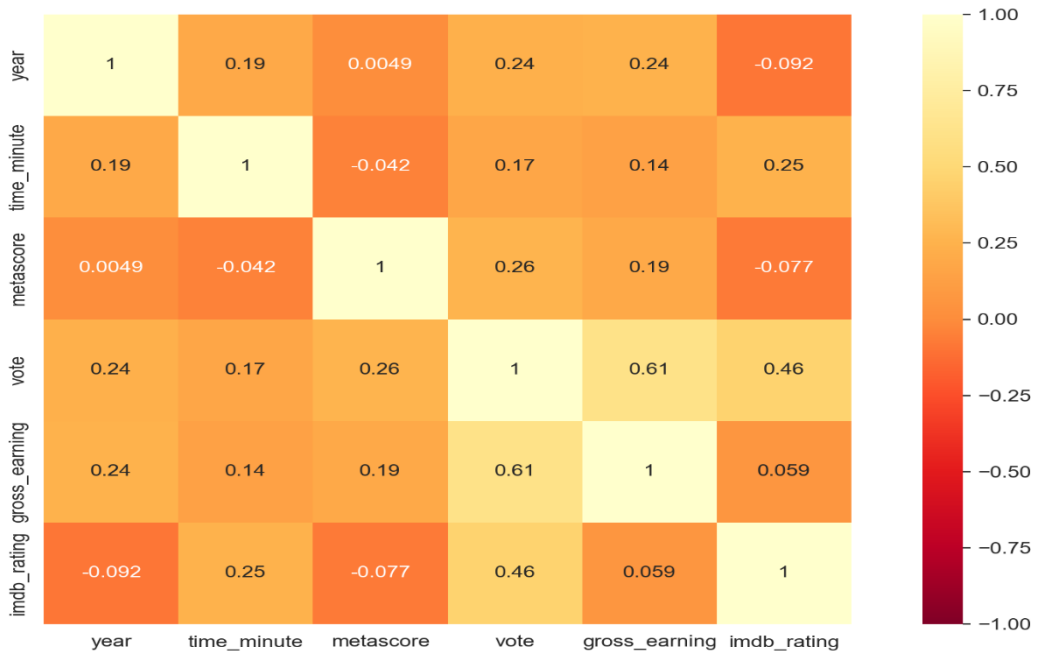
How many vote for each movies genre?

We noticed that the action vote of the films is considered the highest rating, the second rating is my story , and the third rating is adventure



Correlation

The correlation between all numeric columns was calculated. As shown in the heatmap graph there are a lot of columns that are highly correlated with another for example Gross_earning and Imdb_rating



Algorithm

- **Data Preparation**

feature Selection , splitting data , feature engineering , dummy variables , add new cloumns , Impute zero value with mean and regression

- **Models**

We have explored many regression machine learning models to select best models . The models are linear regression, K-fold linear regression, Polynomial regression, Ridge regression(alpha=0.2), andTuned RidgeRegression(Alpha=1)and lasso regression cross (alpha =1). The data was splitted into 60 percent training, 20 percent validating, and 20 percent testing. The model we selected was polynomial regression with degree 2 because it shows the best score between all models.

Train Score 0.77 and validation Score 0.63 and test Score 0.69