Final Project: Top 1000 Movies on IMDB

Ephets Head

4/24/2022

Introduction

In this project, I will be working with a dataset of the 1000 highest-rated movies on the IMDb (International Movie Database) website. Using the information provided in this dataset, I will create and test a model that will predict the gross revenue of a film based on its other attributes. I will also conduct inference on which factors are actually useful in predicting the gross, and how some factors may be correlated with others.

Loading Data and Packages

The dataset "IMDB Movies Dataset" includes 1000 observations with 16 variables each. Some of the variables, like `Poster_link` which is simply the url to the movie poster, are not relevant to our model so they will be discarded during data cleaning. Some of the more important variables, however, I will describe below.

`Series_Title` is where the name of each movie is stored.

`Released_Year` is the year in which the movie was first released.

`Certificate` is the rating certificate that the movie is classified as.

`Runtime` is the duration of the movie in minutes.

`Genre` is the type or genre of the movie.

`IMDB_Rating` is the rating out of 10 that the movie received on the IMDb website.

`Meta_score` is the rating of the film out of 100, as calculated from the average of a large group of respected critics' reviews.

`Director` is the name of the movie's director.

`Gross` is the total money earned by that movie (the outcome variable.)