IMDB Project

xiaofan xia

11/20/2019

Overview

Introduction

This project focuses on digging into the IMDB dataset. My main objective of this project is to find out how much the movie's reputation (IMDB rating) is influenced by other related variables and if any difference in IMDB rating and number of voters regarding to different movie genres or release year.

Motivation

I am a very big fan of many kinds of movies and TV series. During my spare time, I always choose to watch movies and TV shows to relax. Also, many friends of mine are working on the film industry right now. Therefore, I generate some questions about movies and its rating system, and that is why I choose this dataset and this project. I wish I could generate some models to predict the ratings of movies during this project.

Dataset

IMDB Dataset. (https://www.imdb.com/interfaces/) This dataset contains 6 different tsv files and the dataset is refreshed daily. I'll just stick to the dataset I downloaded by Nov.6.2019. The dataset contains total 36 variables and more than 900,000 observations.

Research Questions

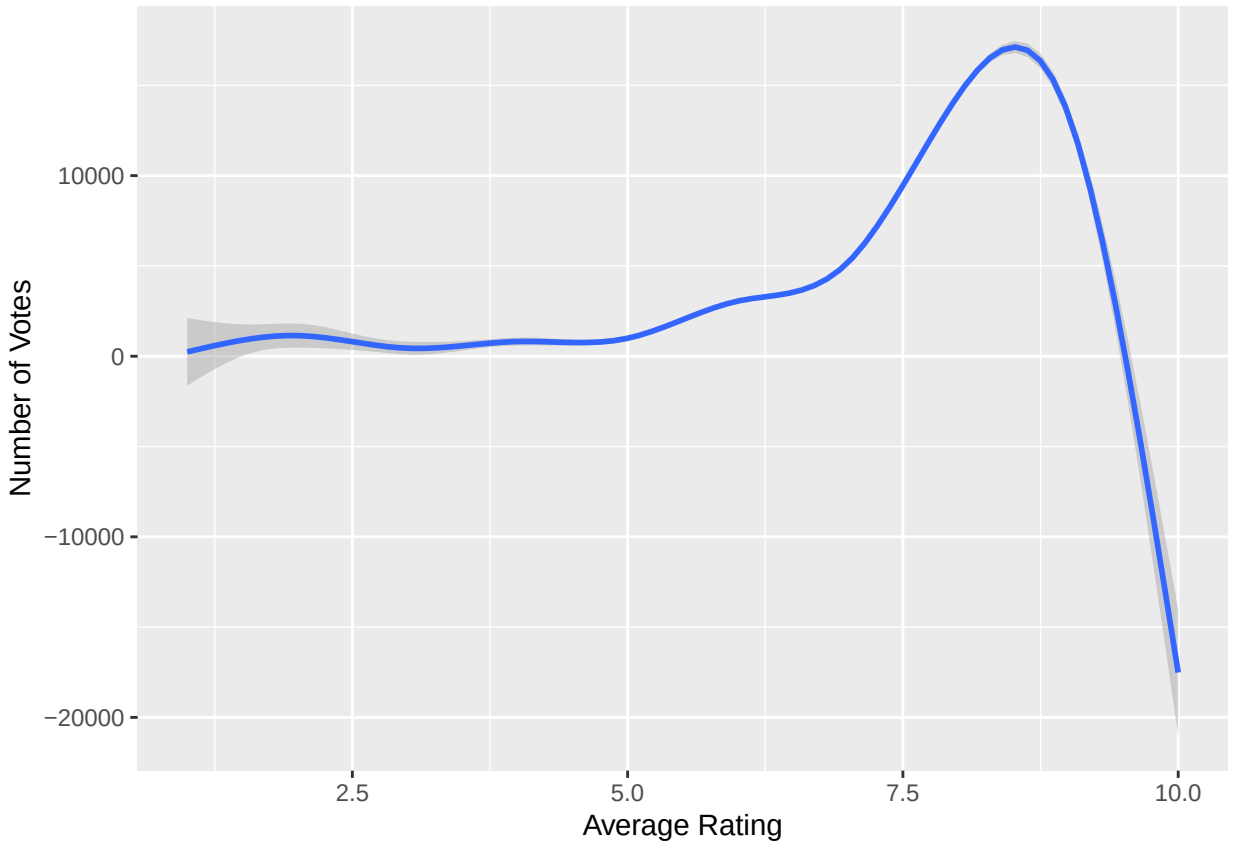
1. What's the relationship between the popularity of casts and the reputation of the product? If the primary cast of the product is known for more titles, does it means his or her work get higher attentions (more votes) or higher ratings?
2. Is there any difference in the ratings of the product regarding to different genres?
3. Generally, is there any difference in the average ratings regarding to the release year?
4. Is there any difference in above questions regarding to different language or country of the product?
5. Generate a model to predict the ratings of the product if it is possible.

EDA

```
## [1] 6.088718
```

```
## [1] 4312.774
```

```
## [1] 220.015
```


10000
0
−10000
−20000
Number of Votes
2.5
5.0
7.5
10.0
Average Rating