**INFSCI 2415 Fall 2022**

**Information Visualization**

Xiaoming Shen

University of Pittsburgh

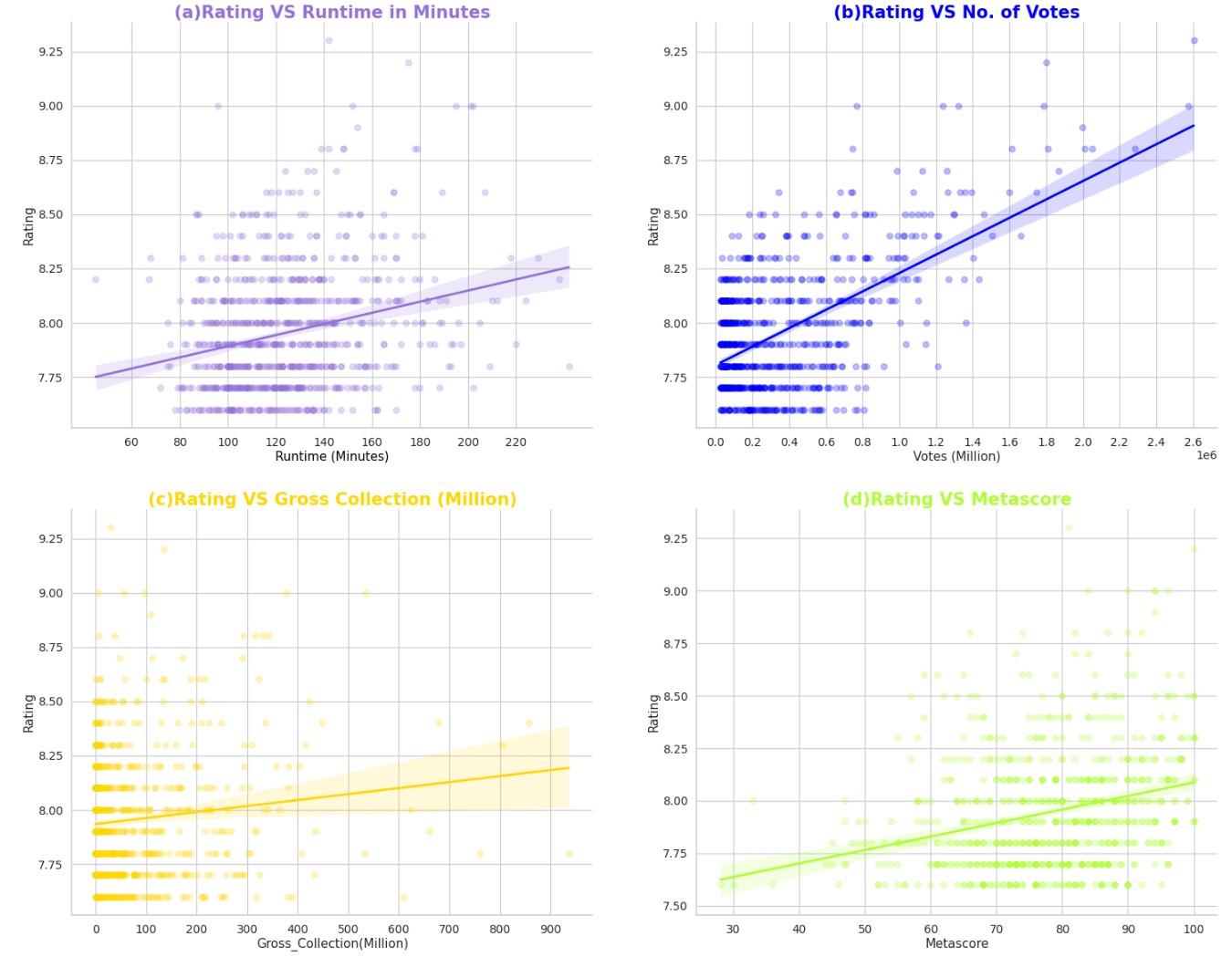
INFSCI 2415

Final Report

**Analysis of IMDb Data: Which Factors Affecting High-Rating Films Most?**

IMDb is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, ratings and critical reviews. Among all the storage types, films are definitely at the heart of IMDb. To see which factor has the most impact on film ratings, this figure was painted by Python.

**Four Factors Influencing the Ratings of a Film in IMDb**



In this figure, It contains four sub-charts which are related to the runtime, quantity of votes, total box office until 2022, and score in Metacritic of a film recorded in IMDb. X-axis represents different Influencing factors separately. And Y-axis in all the four sub-charts represents films rating. Each light colored circle is a record of movie in IMDb. With similar rating and attribute of impact, the circles will overlap and make some areas become darker than others

Edward Yang, a renowned Taiwan director, has said since the invention of movies, human life has been extended at least three times. So I selected the purple, a color of dream, to represent the runtime of films. Blue represents logic and ration. People evaluate a movie usually for some reasons. Besides, gold can reflect the commercial value of a work. And if a work get score more than 60 points, its background in Metacritic will go green.

Here come to some noticeable findings:

* Only the quantity of votes shows a relatively strong correlation to the rating of a film.
* Those films which get the highest ratings didn’t get the best performance of gross collection.
* Although the data set records top 1000 films in IMDb, some of them can’t even get 60 points in Metacritic, another authoritative rating website.
* A film with longer runtime does not necessarily mean it's going to get a higher rating in IMDb.

In my midterm report, I used a data set only containing films on Netflix. However, Netflix is not a traditional film company. So it’s better to focus on a professional movie review website if my research topic is the film industry. Finally, I found this data set about top 1000 films on IMDb. Except for the movie ratings and attributes mentioned in the figure above, this set also provides the year of release, genre and even specific certificate. In order to show the correlation better, I selected the linear regression model, and arranged four charts at once.

There are hundreds of standards to judge whether a movie is good. For those avid moviegoers, it’s common to rely on referring ratings on film review sites like IMDb or Metacritic. However, the most relevant factor to movie ratings, in this figure, is the quantity of votes which has nothing to do with the movie itself. The figure also shows the subjectivity of film evaluation. A highly reviewed film tends not to do well at the box office, and a same movie could receive totally different reviews in two evaluation medias. Therefore, I think for individuals, the best way to judge a movie is to form your own criteria instead of relying on movie review sites.

Github link: https://github.com/Ashenming/Visualization-Final-Project.git