Analysis of the Top 250 Movies on IMDb- A Web Scraping Project

IMDb (Internet Movies Database) is an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDb which began as a fan-operated movie database has grown to be one of the most popular and trusted source for tv, movies reviews etc content.

This project explores data of a list of the top 250 movies on IMDb. It seeks to understand the characteristics of these movies / by what criteria IMDb uses to qualify them as 'TOP'.

This project can be sub-divided into:

- Data Gathering
- Data processing / Cleaning
- Data Exploration/ Visualization
- Findings and Conclusions

Data Gathering

The data for this project was scraped from IMDb website. The data spanned across five pages with each page consisting of information about 50 movies. Data was gathered through the following processes:

- A list was created to store each url
- A request to the web through the "request library" which produced a successful response (<Response [200]>)
- Each url was then looped through to get necessary information through the "BeautifulSoup Library".
- Information concerning some movies were not on the website, so they were replaced with empty value.
- Lastly, the final dataframe was saved to a csv file.

Final Dataframe contained 9 columns and 250 rows

- Movie name: name of the movie
- year: year it was released
- runtime: duration of movies in min
- genre: category of the film
- rating: average rating by the voters (over 10)
- metascore: score given by metacritic (over 100)

- votes: number of users who gave ratings
- gross: gross revenue of the movie in dollars(M)
- rank: rank of the movie on the list from 1-250

Data Processing/Cleaning

This data did not need a lot of cleaning as it was done alongside scrapping. The data however contained null values which were pieces of data that were not on the website. Also, the year was converted to a datetime datatype.

Data Analysis/Visualization

The data was explored using basic statistics methods and visualizations.

Findings:

- The data spanned movies released in a wide range of years (1921 – 2022). The year with the highest top movies was 1995.
- The top 250 movies are high rated movies with a minimum of 8 and a maximum of 9.3. 75% of the data are rated around 8.
- The above cannot really be said for the metascore. The metascore has a minimum of 55 and a maximum of 100.
- The gross revenue of more than half of the movies are below $50M and 25% of the movies are well above $140M with the maximum over $800M. These movies do well in revenue considering that most of these movies are from way old back where inflation had not taken over, so the money had a lot of value.
- Majority of the movies are above 100 min and below 175 min duration.
- High ranked movies have higher ratings.
- There is a positive correlation between ratings and votes which is no surprise since the ratings are the average of the total ratings the voters gave.
- Votes and rank have the same relationship has rank and ratings. This can also be understood from the relationship between ratings and votes.
- There is no significant relationship between rankings and the metascore.

Conclusion

The top 250 movies on IMDb are mostly dependent on the ratings which are done by voters(users) compared to the metascore which are done by the metacritics (movie critics).

Limitation

More data would have been better to make concrete conclusions regarding the relationships between variables.