Movie Insights:
Analyzing User Reviews, Ratings, Revenue, and Popularity using Data Science Tools

Yang Shen, Yi Sun, Skye Tian, Melissa Wang

University of Michigan | SI 699 Mastery Course | April 2023

Motivation

- The factors that influence the success of movies are various. Discover patterns and relationships, maybe help predict future outcomes
- The desire to provide insights to filmmakers and studios. Help them make better decisions and improve the quality of films
- Informing marketing and advertising strategies: By understanding the factors that contribute to movie success, we can develop more effective marketing and advertising strategies that resonate with audiences.

Data

1. **Web scraping:**
   a. IMDB (https://www.imdb.com/)
   b. The Numbers (https://www.the-numbers.com/)
   c. Google Trend (https://trends.google.com/home)
   d. Metacritic (https://www.metacritic.com/)
2. **API:**
   a. IMDB

Methods

Data collection

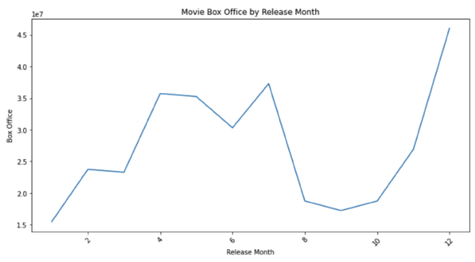
- Web scraping: Metacritic, The Numbers
- Python API, Web scraping: IMDB

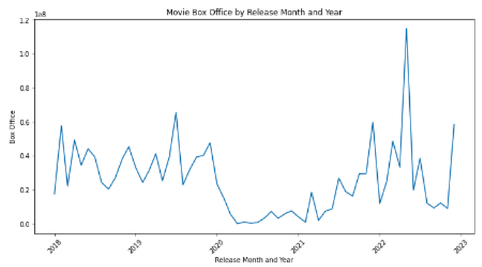
Data analysis

- **Regression analysis**
  - It is used when there is a linear relationship between two continuous variables.
- **Spearman's rank correlation coefficient**
  - It is used when analyzing two continuous variables that are not normally distributed.
- **The Kruskal-Wallis test**
  - It is used when comparing three or more groups of continuous or ordinal data. It is robust to violations of normality.

Results

- **User Rating vs. Critic Score**
  - Correlation coefficient: 0.596.
  - Highest 5 correlation coefficients genre:
    - Animation (0.764), History (0.763), Western (0.697), Adventure (0.670), Crime(0.663).
  - Lowest 5 correlation coefficients genre:
    - Music (0.494), Romance (0.520), Horror (0.526), Drama (0.528), Thriller (0.539).
- **Revenue and Release Date**
  - Time Series Analysis:
    - April, July, and December are the most beneficial month for movie. August to October is the low season.
    - The whole year of 2020 is the trough which might be due to the pandemic. Box office gradually recovered from 2021 and reached maximum in 2022.



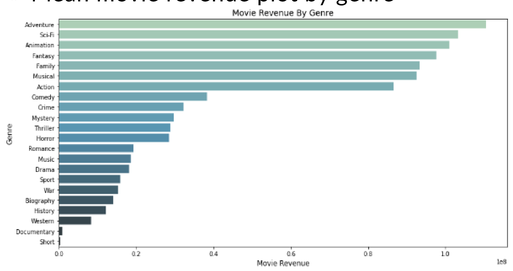


- **Revenue by month**
  - Revenue by Month using Kruskal-Wallis Test: Overall (p-value = 0.3069)
  - By Genre (all p-values > 0.05): There is no significant difference between revenue in each months by genre.
- **Movie revenue and ratings Overall**
  - Spearman's rank correlation coefficient analysis of IMDB/Critic Rating and Domestic box office: IMDB user rating and revenue: correlation=0.0973, p-value<0.05; Critic rating and revenue: correlation=-0.2062, p-value<0.05
- **Movie revenue and ratings by genres**
  - Spearman's rank correlation coefficient analysis of IMDB Rating and Domestic box office:
    - highest positive correlation: **Musical**(0.47), **family**(0.44) and **documentary**(0.42)
    - Western(-0.6) has a **negative** correlation
  - Spearman's rank correlation coefficient analysis of Critic rating and Domestic box office:
    - highest negative correlation: **Sport**(-0.43), **Western**(-0.4) and **Musical**(-0.24)
- **Movie revenue and genre**
  - **Shapiro-Wilk normality test for each genre's revenue distribution**
    - P-value < 0.05 for all genres; None of the genres' revenue have normal distribution
  - **Kruskal-Wallis test**
    - P-value<0.05; There is significant difference in revenue among genres
- **Movie revenue and runtime**
  - **Spearman analysis:** correlation=0.3965, p-value<0.05
  - By genres
    - The most positive correlation: **Science-Fic**(0.54), **Music**(0.53), and **Animation**(0.52)
    - The weakest correlation: **Western**(0.02), **Crime**(0.16), and **Musical**(0.2)

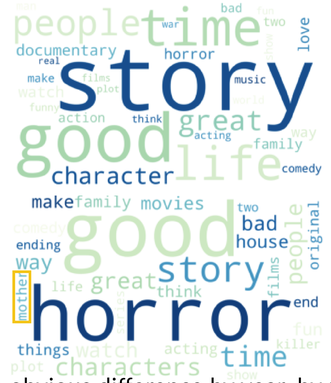
Results - more visualizations

Movie revenue

- Three most profitable genres: adventure, sci-fi, animation
- Mean movie revenue plot by genre



Top 30 most frequent words in reviews



Reviews for all movies

Reviews for horror movies, thrillers

- No obvious difference by year, by rating
- In horror and thriller movies, 'mother' occurs.

Conclusion

1. User rating and critic score are moderately positively correlated.
   a. User ratings are generally more aligned with critic scores for genres like animation, history, and western, while the opposite tends to be true for genres like music, romance, and horror.
2. There is no statistically significant difference in revenues between months.
   a. That being said, December is often considered the most advantageous month in our data sample.
3. Generally, user ratings have a slight positive correlation with revenue, while critic scores have a slight negative correlation.
4. Runtime and revenue are moderately positively correlated.
   a. Especially for sci-fi, music and animation, but not so much for western, crime and musical.
5. The most profitable genres are adventure, sci-fi and animation.