What movies are performing best at the box office?

Presenter: *Setare Hajarolasvadi*

DS Flex Program: *Phase II Project*

Overview

- Objective
- Datasets
- Data Cleaning
- Data Analysis
- Results
- Recommendations
- Limitations

Objective

- Identify the highest profitable movies using several datasets and considering success factors such as movie genre, release date, etc.

$$\text{Metric: ROI\%} = \frac{\text{Gross Revenue} - \text{Production Budget}}{\text{Production Budget}} \times 100$$

Datasets

- Movie title
- Release date
- Production budget
- Gross domestic/worldwide

Missing:

- Genres, average rating

Join on

- Movie title
- Release date
- Genres
- Average rating

Missing:

- Production budget
- Domestic/worldwide gross

- Movie title
- Gross domestic/foreign
- studio

Missing:

- *Production budget*
- Genres, average rating

- Genres
- Box office
- Rating
- Studio

Missing:

- *common key for merge*

Datasets

- Movie title
- Release date
- Production budget
- Gross domestic/worldwide

Missing:

- Genres, average rating

Join on

- Movie title
- Release date
- Genres
- Average rating

Missing:

- Production budget
- Domestic/worldwide gross

- Movie title
- Gross domestic/foreign
- studio

Missing:

- *Production budget*
- Genres, average rating

- Genres
- Box office
- Rating
- Studio

Missing:

- *common key for merge*

Datasets

THE NUMBERS

- Movie title
- Release date
- Production budget
- Gross domestic/worldwide

Missing:

- Genres, average rating

Join on

THE MOVIE DB

IMDb

- Movie title
- Release date
- Genres
- Average rating

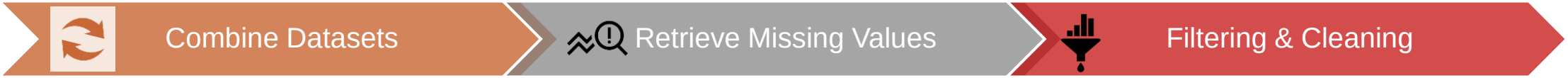
Missing:

- Production budget
- Domestic/worldwide gross

Still Missing Data?

IMDbPY

Data Preparation



- Join data from The Numbers, TMDB & IMDB using movie titles.
- Refine the merge by basic fuzzy matching logic.

- Retrieve genres & average rating from the IMDB and TMDB datasets.
- Refine the retrieval using Python's imdb module (computationally demanding).

- Remove suspicious entries.
- Remove remaining entries with missing information.

Data Analysis

Group data based on important attributes

Visualize Return On Investment (ROI) based on different metrics.

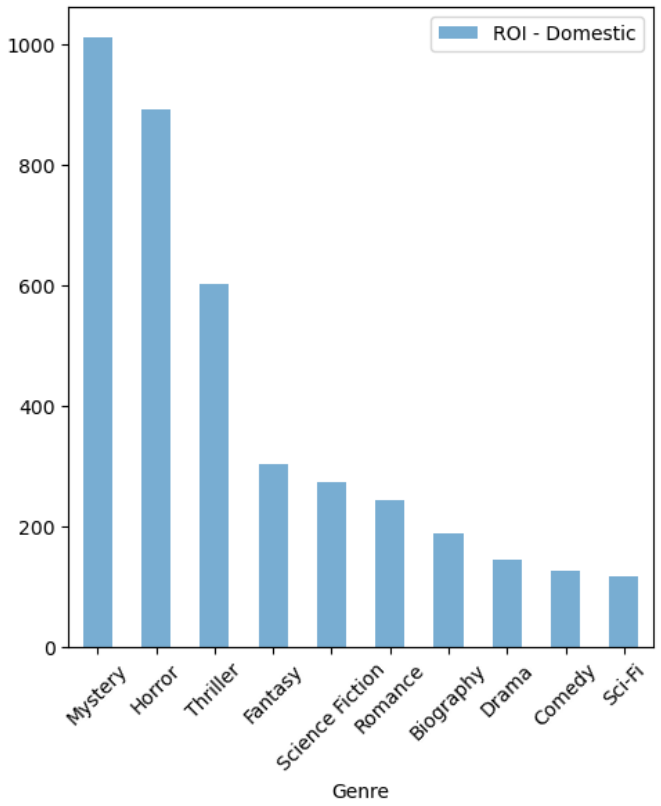
Interpret the visualizations.

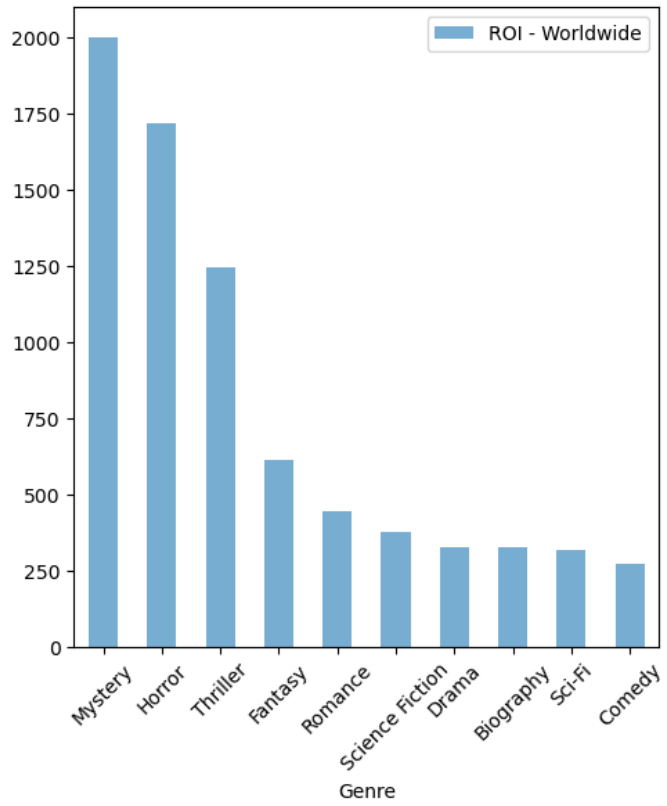
Recommend business strategies.

$$\text{ROI\%} = \frac{\text{Gross Revenue} - \text{Production Budget}}{\text{Production Budget}} \times 100$$

- 1866 entries
- 2010 to 2019

Results: Genre



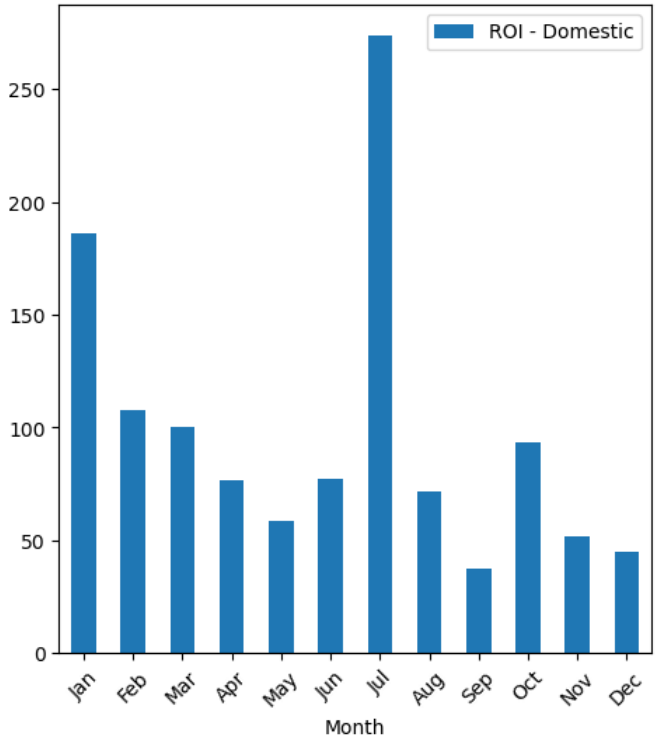


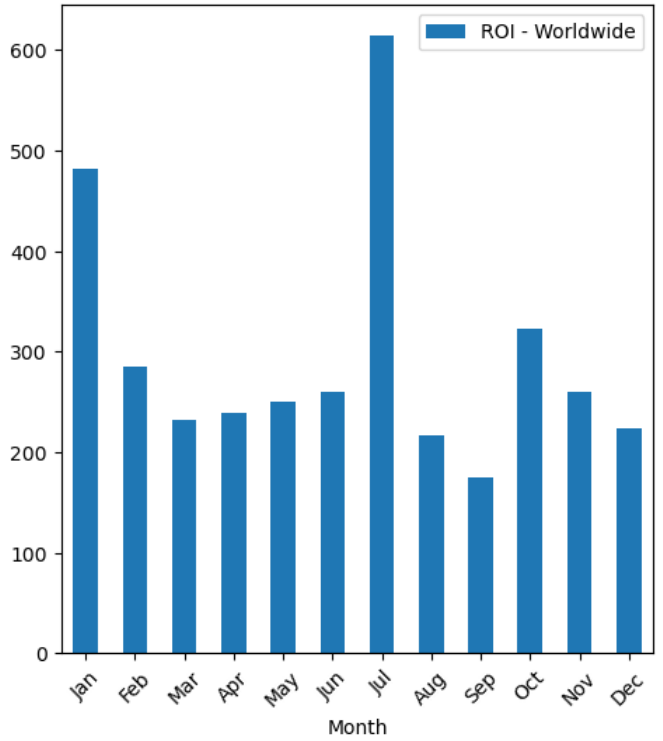
Key Takeaway(s):

The highest profitable movies belong to the *mystery*, *horror* and *thriller* genres (also *Documentary* for low to medium budget movies)

- 1866 entries
- 2010 to 2019

Results: Release Date



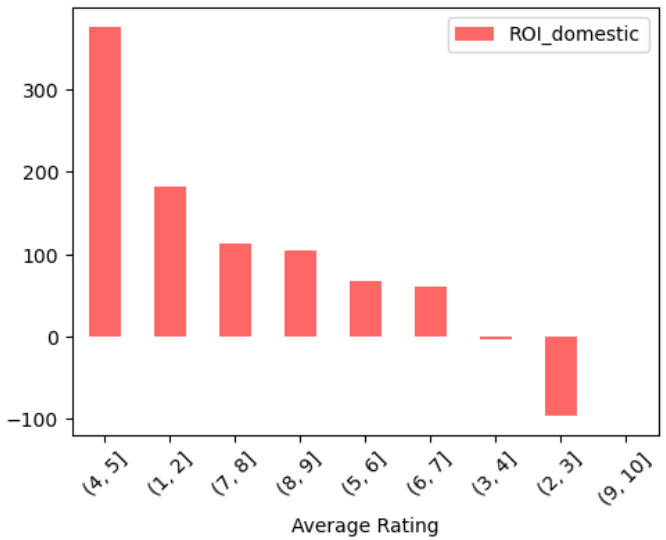


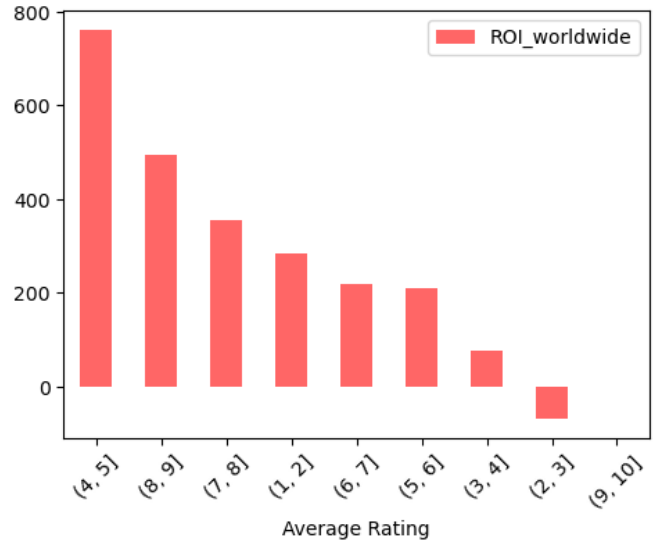
Key Takeaway(s):

Most highly profitable movies have been released in the month of *July* (followed by *January*). September is a dump month.

- 1866 entries
- 2010 to 2019

Results: Audience Average Rating





Key Takeaway(s):

Movies with average rating have the highest return on investment.

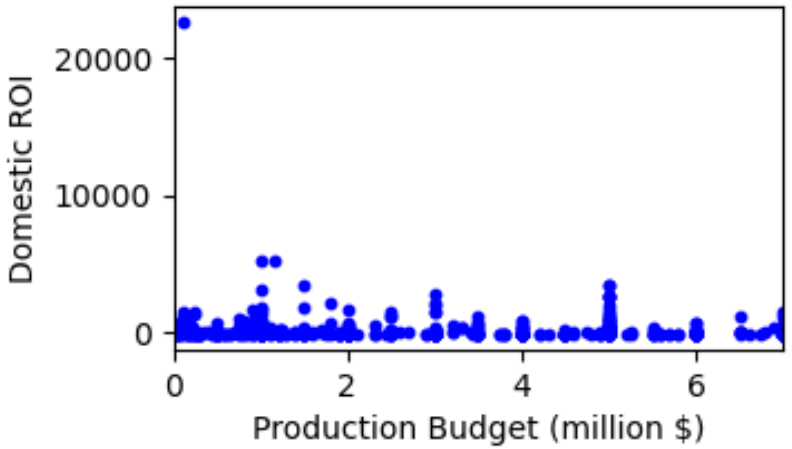
Note: Data for extremely low/high rating limited and unreliable.

Results: Production Budget

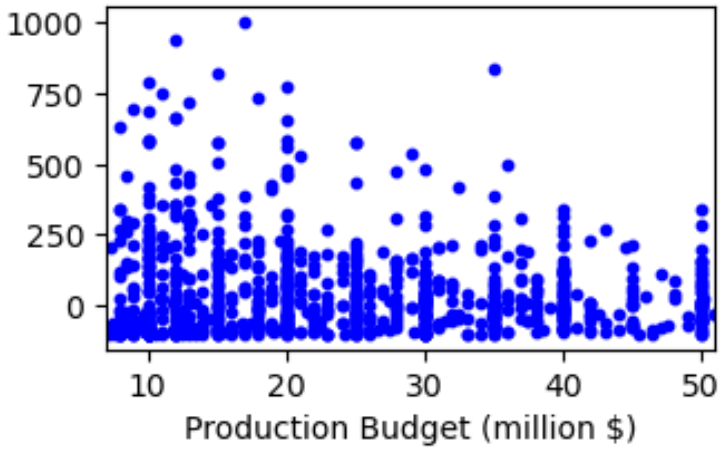
Key Takeaway(s):

Higher production budget doesn't necessarily lead to more profit.

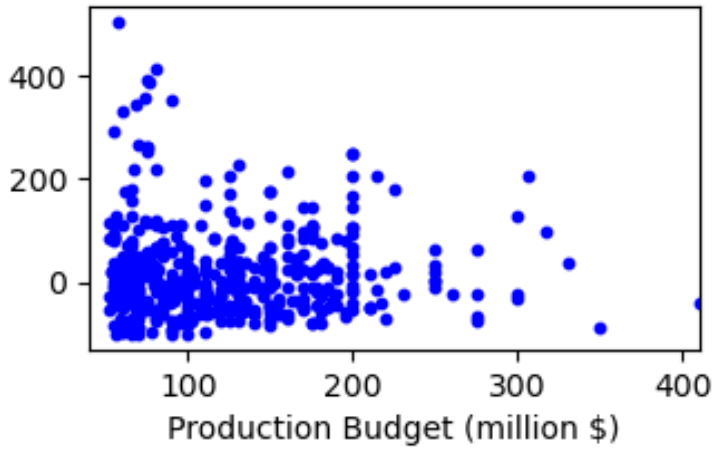
Low-Budget



Medium-Budget



High-Budget



Recommendations

1. Consider investment in *horror, mystery or thriller* movies. If budget is low to medium, consider *documentaries* as well.
2. Consider releasing the movie in *July or January* and *avoid* releasing the movie in *September*.
3. Be cautious when investing in highly-rated/acclaimed movies. Average-rated movies seem to be more profitable.

Limitations

The above study is limited in the following ways:

1. The work could benefit from optimization with fuzzy matching techniques to retrieve/keep more of the original datasets.
2. Do you have a specific budget in mind? Ask us to do an in-depth analysis on movies falling in a specific budget category. The factors driving success may be different in each category.
3. The work is an exploratory data analysis. It can benefit from building a model that can describe ROI in terms of the factors considered in a statistically-significant manner.