Data Immersion 6.1:

IMDB movies (1980-2020)

Dataset: https://www.kaggle.com/datasets/danielgrijalvas/movies

Raw summary

7668 rows (7668 movies, 200 for each year between 1985-2019)

Columns:

1. name – name of the film.
2. rating – appropriateness of film rating, i.e. PG, R, etc. Based on mpaa.
3. genre – the main genre of the film.
4. year – the year the film was released.
5. released – the month, day, year of release with country in parentheses.
6. score – the IMDB score by users. Scale = 1 – 10, 1 being terrible.
7. votes – number of votes towards the IMDB score.
8. director – name of the director of the film.
9. writer – name of the writer of the film.
10. star – main actor credited with the film.
11. country – the country where the film was produced.
12. budget – the estimated budget for the film (in U.S. dollars)
13. gross – the estimated gross worldwide revenue of the film (in U.S. dollars)
14. company – the production company of the film
15. runtime – the runtime of the film (in minutes).

Data Profile

Cleaning (done in Excel):

1. name –
    a. Came up as it was missing other columns of info, but I changed Saw: The Final Chapter to Saw 3D. The movie has recently changed its title.
2. rating –
    a. changed the 1 "Approved" result for "Tarzan the Ape Man" to R based on research on production company website.
    b. two observations "Unrated" and "Not Rated" mean same thing. Changing "Not Rated" to "Unrated" for brevity and consistency. (283 results)
    c. Imputed 9 blank results with mpaa rating found within IMDB elsewhere. Changed 68 blanks to "Unrated".
3. released –
    a. deleting this column. Many of the "release" dates/years are inconsistent with the "year" column. Also, IMDB is inconsistent in how it prioritizes this field. For some movies they release in United States primarily, but it lists "Argentina" or "Ireland" as the release date, even if the production company is from the U.S. Sometimes, it's the opposite, the film is foreign, but the release date is when it released in the U.S., sometimes years after initial release.
4. score, votes –
    a. imputed for 2 movies where the rating for some reason wasn't scraped but is listed in IMDB.
5. "The Robinsons" - Movie listed for 2020 I cannot find any record of at all. Deleting
6. director –
    a. imputed 30 directors where there was an inconsistency, listing the name as "Director" or "Directors" in the entry. Imputed first name in list of directors according to IMDB, consistent with formatting of writer and star column.
7. writer –
    a. Imputed 3 writers in blanks, 2 results they were easy to find in IMDB elsewhere, the others I found through research.
8. country –
    a. imputed 1 country for a blank. Result was available through researching production companies within IMDB.
9. budget, gross, company, runtime –
    a. imputed for "Saw 3D" and "The Wolfman" since the numbers are available in IMDB.
10. company –
    a. Imputed 14 companies for blanks. Was able to find them with a little research.
11. runtime –
    a. Imputed 2 entries. Both were simple to find either on IMDB or elsewhere with a little research.

7667 rows (7667 movies, 200 for each year between 1985-2019) – No Duplicates found

Descriptive Statistics:

Variables	time - variant/invariant	structured/unstructured	qualitative/quantitative	Qualitative: nominal/ordinal Quantitative: discrete/continuous
name	time-variant	unstructured	qualitative	nominal
rating	time-variant	structured	qualitative	nominal
genre	time-variant	structured	qualitative	nominal
year	time-invariant	structured	quantitative	continuous
score	time-variant	structured	qualitative	ordinal
votes	time-variant	structured	quantitative	discrete
director	time-invariant	unstructured	qualitative	nominal
writer	time-invariant	unstructured	qualitative	nominal
star	time-invariant	unstructured	qualitative	nominal
country	time-invariant	structured	qualitative	nominal
budget	time-invariant	structured	quantitative	discrete
gross	time-variant	structured	quantitative	discrete
company	time-invariant	unstructured	qualitative	nominal
runtime	time-invariant	structured	quantitative	continuous

Qualitative:

- name
  - value count: 7511
  - mode: ‘Anna’, ‘Fever Pitch’, ‘Hamlet’, ‘Hercules’, ‘Nobody’s Fool’, ‘Pulse’, ‘Venom’
- rating
  - value count: 10
  - mode: ‘R’
- genre
  - value count: 19
  - mode: ‘Comedy’
- score
  - value count: 72
  - mode: 6.6
- director
  - value count: 2959
  - mode: ‘Woody Allen’
- writer
  - value count: 4535
  - mode: ‘Woody Allen’
- star
  - value count: 2814
  - mode: ‘Nicolas Cage’
- country
  - value count: 59

    - mode: ‘United States’
- company
    - value count: 2390
    - mode: ‘Universal Pictures’

Quantitative:

- year
    - minimum: 1980
    - maximum: 2020
    - mean: 2000
    - median: 2000
    - mode: 1985 – 2019 all have 200 movies each.
- votes
    - minimum: 7
    - maximum: 2400000
    - mean: 88085
    - median: 33000
    - mode: 13000
- budget
    - minimum: 3000
    - maximum: 356,000,000
    - mean: 35,614,321
    - median: 20,500,000
    - mode: 20,000,000
- gross
    - minimum: 309
    - maximum: 2,847,246,203
    - mean: 78,516,440
    - median: 20,208,496
    - mode: 14,000,000
- runtime
    - minimum: 55
    - maximum: 366
    - mean: 107
    - median: 104
    - mode: 97

Limitations and Ethics:

Sourcing: IMDB data is from industry documents and submitted by industry professionals. However, some information is also user submitted which means is it not entirely accurate. Still, IMDB conducts consistency checks to maintain the integrity of their information to make sure it is as accurate as possible. Budget and Revenue numbers are estimated but they are the closest available to accurate information. Many results are missing budget and revenue information as well, so need to be mindful of this as it could lead to bias toward certain movies, genres, etc.

Collection: Through scraping IMDB. There are some inconsistencies with the data, mainly some blanks that when checking the website, look like they should have been collected. Also, it is not specified how exactly the movies were chosen for the dataset, random or otherwise.

Ethics: All information collected is open to the Public Domain. The names of the directors, writers, and stars are personal information that is associated with the films in public records, so displaying in this dataset and analysis does not violate laws.

Exploratory Questions

- What are the most profitable movies?
- What are the least profitable movies?
- What's the most profitable genre?
- What's the least profitable genre?
- What is the relationship between budget and revenue?
- Is there a positive relationship between score and votes?
- Is there a positive relationship between score and revenue?
- How has movie revenue or budget changed over time?
    - Is there a year more, or less, profitable than the others?
    - Have companies seen significant changes in revenue over time?
    - Have countries seen significant changes in revenue over time?
- How does genre and/or rating affect revenue?
- What is the breakdown of genre regarding revenue?
- What is the most profitable rating category?
- What is the most successful company?
- Who is the most successful director/writer/star?
- How does revenue breakdown regarding country?
- Which country has the most revenue?
- Do some countries or companies have a score that is significantly different from others, and why?

- Resulting Hypothesis during analysis: Profit/revenue can be maximized for a film.