IMDB Movies Data Analysis

Gerald Yuen

2022

# project SUMMARY

The data is from IMDB with around 5000 movies from 1916 to 2016. The dataset is from Kaggle back in 2017.

My goal is to find interesting facts with the dataset while showing analytic skills.

# data source

https://www.kaggle.com/datasets/carolzhangdc/imdb-5000

# Limitations and Ethics

* Data was created 5 years ago, so data might not be up to date.
* Only 5000 movies are included.
* Data is not official from IMDB so result might vary.

# Data Cleaning and Data Consistency Checks

* Removing unnecessary columns
* Updated “movie\_title” rows
* Check for blank space and replaced with “N/A”
* Removed duplicate values

# Data Profile

* Number of rows and columns in the cleaned dataset: 4920 x 19

# Column details:

| **Variable Name** | **Description** | **Datatype** | **Time Variant** |
| --- | --- | --- | --- |
| movie\_title | Title of the Movie | Qualitative/Nominal | No |
| duration | Duration in minutes | Quantitative/Continuous | No |
| title\_year | Year of the movie | Quantitative/Discrete | Yes |
| director\_name | Name of the Director of the Movie | Qualitative/Nominal | No |
| actor\_1\_name | Primary actor starring in the movie | Qualitative/Nominal | No |
| actor\_2\_name | Other actor starring in the movie | Qualitative/Nominal | No |
| actor\_3\_name | Other actor starring in the movie | Qualitative/Nominal | No |
| language | English, Arabic, Chinese, French, German, Danish, Italian, Japanese etc | Qualitative/Nominal | No |
| country | Country where the movie is produced | Qualitative/Nominal | No |
| content\_rating | Content rating of the movie | Quantitative/Discrete | No |
| num\_critic\_for\_reviews | Number of reviews | Quantitative/Continuous | No |
| director\_facebook\_likes | Number of likes of the Director on his Facebook Page | Quantitative/Continuous | No |
| budget | Budget of the movie in Dollars | Quantitative/Continuous | No |
| gross | Gross earnings of the movie in Dollars | Quantitative/Continuous | No |
| genres | Film categorization like ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’ | Qualitative/Nominal | No |
| cast\_total\_facebook\_likes | Total number of facebook likes of the entire cast of the movie | Quantitative/Continuous | No |
| imdb\_score | IMDB Score of the movie on IMDB | Quantitative/Continuous | No |
| color | Film colorization. ‘Black and White’ or ‘Color’ | Qualitative/Nominal | No |
| Plot\_keywords | Keywords describing the movie plot | Qualitative/Nominal | No |

# Questions to explore:

* What is the average time of a movie?
* Who are the top directors and how much money does their movies made?
* What genre are the most popular?
* For the highest gross earning of the movies, how are the IMDB scores?