

# Comparing Critics to the General Public: Who is a better predictor of a movie's success?

Jacob Inwald and Ollie Jones

## 1 Overview

This report analyzes the relationship between different aspects of a movie and its gross income. These aspects include director list, actor list, runtime and so on. To investigate these relationships, we looked at a comprehensive dataset compiled by the movie ranking website IMDB.com. This dataset includes information about a movie like the director, the actors, the genre.

We found ..

## 2 Introduction

**Context and motivation** We explore this dataset using data analytics,

**Previous work** Brief description of any previous work in this area (e.g., in the media, or scientific literature or blogs).

E.g. Recent surveys show that most students prefer final projects to final exams [3].

**Objectives** What questions are you setting out to answer? We set out to investigate the relationship between various factors and the rating or gross revenue of a movie. There are various factors that can come into play affecting the rating a movie gets or the success of that movie at box office. As such, we will only focus on a few of these factors, looking specifically at: Director, Actors, Genre and Runtime. We will investigate the effect these factors have on movie reception, and attempt to determine which are most impactful to the success of a movie.

### 3 Data

**Data provenance** We used two datasets to aid our investigation: an IMDB movie dataset, containing various details about movies made from 2006-2016; and a larger movie dataset, containing details about movies released on or before July 2017. The IMDB dataset has a bit of a contentious past as it was scraped off of the movie ranking website IMDB.com, and is actually only a sample dataset from a much larger dataset that has every movie made from 2006-2016 that is in the IMDB. The large movie dataset was collated using data from TMDB (The Movie DataBase) and grouplens.org; but we only use the part that was obtained from TMDB as it is used to find Director and Actor experience.

Both datasets were downloaded from kaggle.com, a data science platform that enables users to access and share datasets. These datasets are shared underneath the CC0 1.0 Universal Public Domain Dedication, as such we will use them underneath fair use.

**Data description** The IMDB movie dataset has 12 columns, all either containing string values or floating point values. The only odd column is the Genre column which contains the different genres that can be applied to a particular movie. The genres that are in this dataset are the arbitrary genres:

- Action
- Adventure
- Sci-Fi
- Mystery
- Horror
- Thriller
- Animation
- Comedy
- Family
- Fantasy
- Drama
- Music
- Biography
- Romance
- History
- Crime
- Western
- War
- Musical
- Sport

Column Name	Description
Rank	The rank the movie has in the IMDB database
Title	The name of the movie
Genre	The genres that apply to the movie, there can be anywhere from 1-3 genres. A genre can be any from: Action, Adventure, Sci-Fi, Mystery, Horror, Thriller, Animation, Comedy, Family, Fantasy, Drama, Music, Biography, Romance, History, Crime, Western, War, Musical, Sport
Description	The description of the movie
Director	The person who directed the movie
Actors	The lead roles in the movie
Year	The year the movie was released
Runtime (Minutes)	The runtime in minutes of the movie
Rating	The mean rating of the movie, taken from IMDB.com
Votes	The amount of users that voted on a movie to give it that rating
Revenue (Millions)	The gross income the movie made at the US box office
Metascore	The rating of movie, determined using aggregated weighted critic scores

Figure 1: The different columns in the IMDB-Movie-Data.csv file

The other dataset used had a bit of a more odd structure. We used the crew.csv file obtained from

**Data processing** How you have processed the dataset, e.g., cleaning, removing missing values, joining tables.

### 4 Exploration and Analysis

A data science analysis of the paper, including:

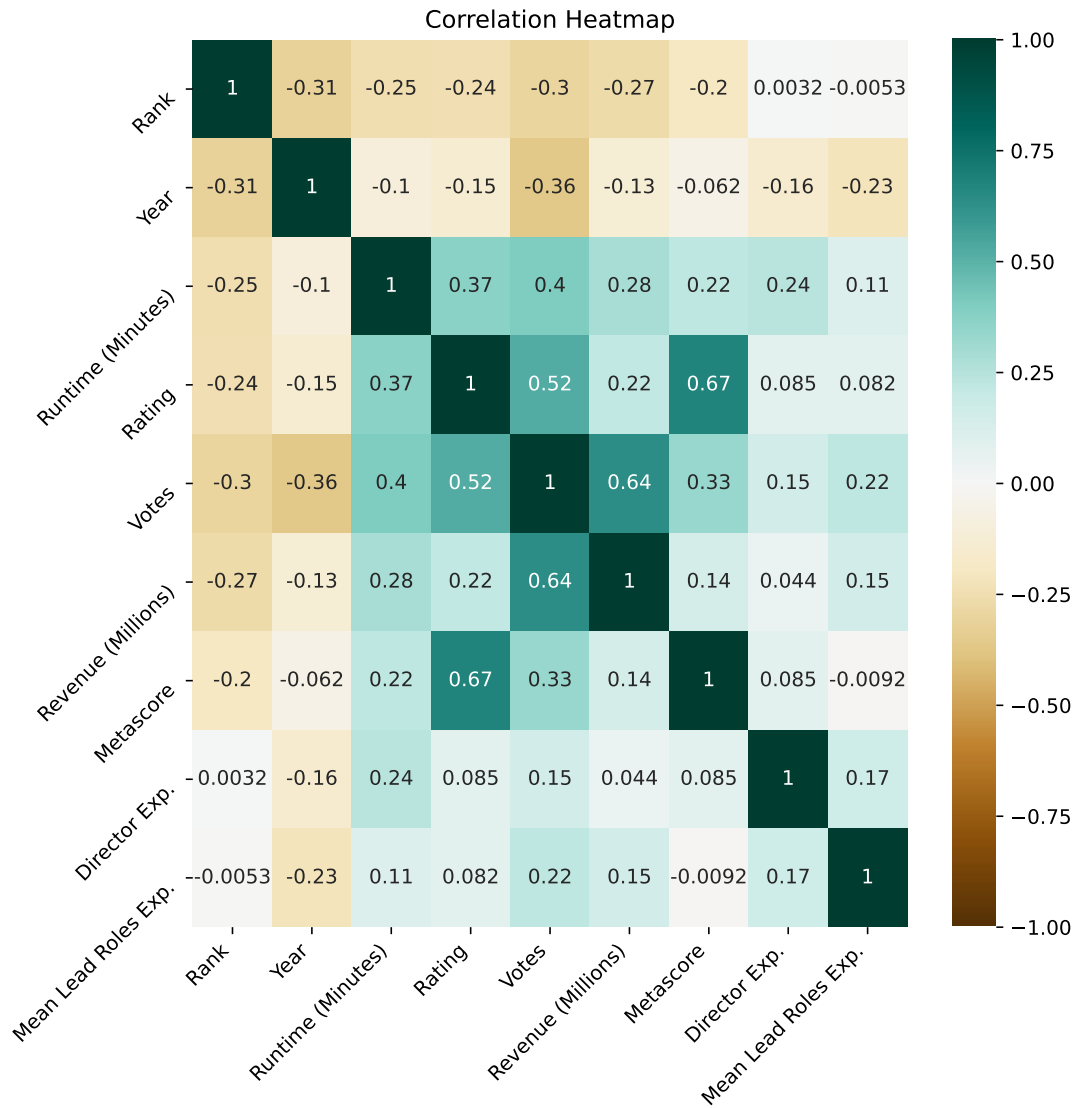


Figure 2: Demonstration figure. This caption explains more about the figure. Note that the font size of the labels in the plot is 9pt, which is obtained by the settings as shown in the Jupyter notebook.

- Visualisations (for example Figure 2) and tables (for example Table 1). Please make sure that all figures and tables are referred to in the text, as demonstrated in this bullet point.
- Interpretation of the results
- Description of how you have applied one or more of the statistical and ML methods learned in the FDS to the data
- Interpretation of the findings

You can use equations like this:

$$\bar{x} = \sum_{i=1}^n x_i \quad (1)$$

or maths inline:  $E = mc^2$ . However, you do not need to reexplain techniques that you have learned in the course – assume the reader understands linear regression, logistic regression K-nearest neighbours etc. Remember to explain any symbols use, e.g. “ $n$  is the number of data points and  $x_i$  is the value of the  $i$ th data point.”.

## 5 Discussion and Conclusions

### Summary of findings

### Evaluation of own work: strengths and limitations

**Comparison with any other related work** E.g. “Anscombe has also demonstrated that many patterns of data can have the same correlation coefficient” [1].

Wikipedia can also be cited but it is better if you find the original reference it for a particular claim in the list of references on the Wikipedia page, read it, and cite it.

The golden rule is always to cite information that has come from other sources, to avoid plagiarism [2].

### Improvements and extensions

## References

- [1] Francis J Anscombe. “Graphs in statistical analysis”. In: *The American Statistician* 27.1 (1973), pp. 17–21.
- [2] Wikipedia contributors. *Plagiarism – Wikipedia, The Free Encyclopedia*. Last accessed 22 July 2004. 2004. URL: <https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350>.
- [3] Phil Space. “Why oh why must I do this project?” In: *The Daily Post* (2021). Retrieved on 28 February 2021. URL: <https://www.dailypost.com>.

Table 1: Excerpt from Scottish Index of Multiple Deprivation, 2016 edition. <https://simd.scot>. You may put more information in the caption.

Location	Employ- ment	Illness	Attain- ment	Drive Primary	Drive Secondary	Crime	...
Macduff	10	95	5.3	1.5	6.6	249	...
Kemnay	3	40	5.3	2.4	2.4	168	...
Hilton	0	10	6.3	2.2	3.0	144	...
Ruchill	8	130	4.9	1.7	5.6	318	...
Belmont	2	50	6.1	3.1	3.2	129	...
...	...	...	...	...	...	...	...