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 [6].

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 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[5], containing various details about movies made from 2006-2016; and a larger movie dataset[2] (TMD), 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 (TMD) 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 datsets 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 arbritary genres:

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

The column summary is shown in Figure 1.

Column Name	Description	Data Type
Rank	The rank the movie has in the IMDB database	Integer
Title	The name of the movie	String
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, Thriller, Animation, Comedy, Family, Fantasy, Drama, Music, Romance, History, Crime, Western, War, Musical, Sport, Horror Mystery, Biography	Genre Category
Description	The description of the movie	String
Director	The person who directed the movie	String
Actors	The lead roles in the movie	String
Year	The year the movie was released	Integer
Runtime (Minutes)	The runtime in minutes of the movie	Integer
Rating	The mean rating of the movie, taken from IMDB.com	Float
Votes	The amount of users that voted on a movie to give it that rating	Integer
Revenue (Millions)	The gross income the movie made at the US box office	Float
Metascore	The rating of movie, determined using aggregated weighted Critics scores	Integer

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

The TMD dataset used had a bit of a more odd structure. We only used the crew.csv file that is a small part of the larger dataset, so we only explain the structure of crew.csv. This file was composed of three columns: cast, crew and id. The odd part of the dataset is that the cast and crew columns are composed of .json files, and as such we needed to parse out the useful data from these json files to actually get useable data. The column summary is shown in Figure 2.

Data processing To cleanup the TMD dataset, we made and used a python script that went through all the entries of the dataset and extracted two things; number of movies each actor appeared in, and number of movies each director directed. These two sets of data were saved to .csv files, named actor_counts.csv and director_counts respectively.

Column Name	Description	Data Type
cast	The cast list of the movie, including all actors who appeared in it.	.json file
crew	The entire crew list of the movie, including all the people who made it.	.json file
id	The movie id. This is used in the larger dataset to connect data together	Integer

Figure 2: The different columns in the credits.csv file

Using this newly acquired data, we merged it with the original IMDB dataset. In the merging process the Description column was dropped, and the Director and Actor columns were replaced with the Director Exp. and Mean Lead Roles Exp. This resulted in a merged_movie_data.csv file with structure shown in Figure 3.

Column Name	Description	Data Type
Rank	See fig 1	Integer
Title	See fig 1	String
Genre	See fig 1	Genre
Director Exp.	The number of movies that the director of the movie has made	Float
Mean Lead Roles Exp.	The mean number of movies that the lead actors have been in	Float
Year	See fig 1	Integer
Runtime (Minutes)	See fig 1	Integer
Rating	See fig 1	Float
Votes	See fig 1	Integer
Revenue (Millions)	See fig 1	Float
Metascore	See fig 1	Integer

Figure 3: The different columns in the merged data set

To check this data was properly normalised we made a histogram plot of all the numeric variables, shown in Figure 4. As expected, a few variables did not appear to be normally distributed, namely: Revenue (Millions), Votes, Runtime (Minutes), Director Exp., and Rating.

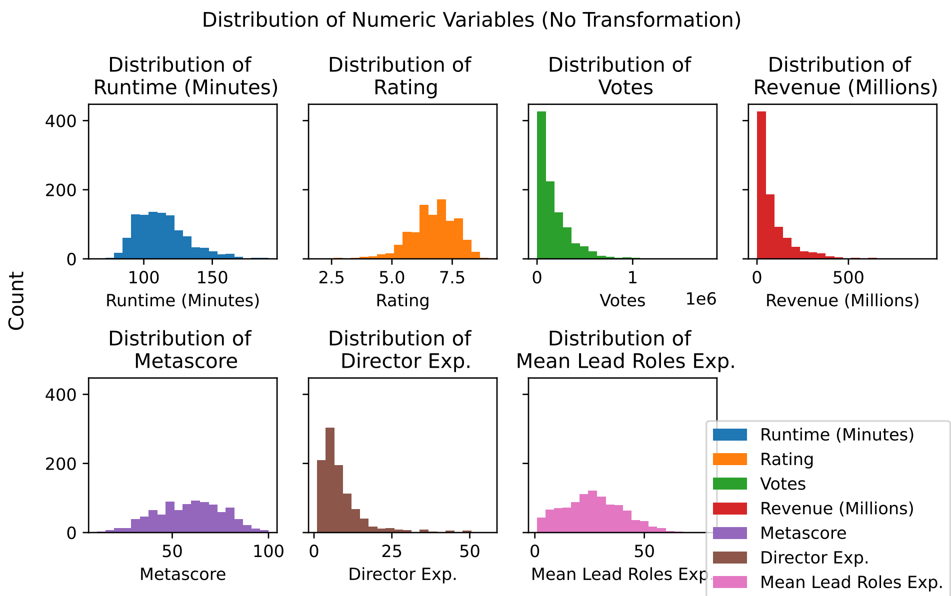

Figure 4: The distributions of the numeric variables in the merged dataset

As shown in the plot, Revenue (Millions) and Votes are severly right skewed, implying an exponential distribution; Runtime (Minutes) and Director Exp. appear to be less severly right-skewed, implying a lognormal distribution; Rating seems to be left-skewed. The transformations for these variables that gave the best approximations to a normal distribution were:

- Revenue (Millions) : Cube Root transform
- Votes : Cube Root transform
- Runtime (Minutes) : Log transform
- Director Exp. : Log transform
- Rating : Square transform

Figure 5 shows the transformed and normalised numeric variables. The label has the p-values from testing whether the transformed distribution is normal, using the Kolmogorov-Smirnov test[3] for goodness of fit. One interesting note is that altough the Director Exp. column fails the Kolmogorov-Smirnov test, there is clearly missing data; around 3 columns are missing from the histogram shown. With this in mind, and as the histogram does follow the normal distribution curve, there is sufficient evidence to assume ln(Director Exp.) has a normal distribution.

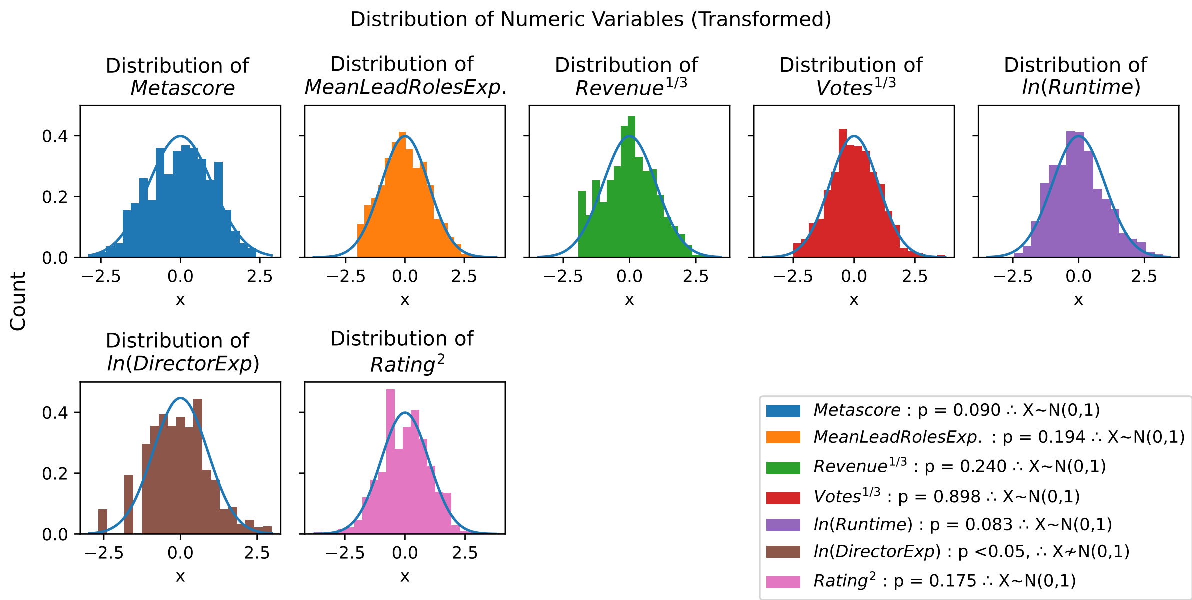

Figure 5: The distributions of the numeric variables in the merged dataset

This normalised data was used for the rest of our investigations later on.

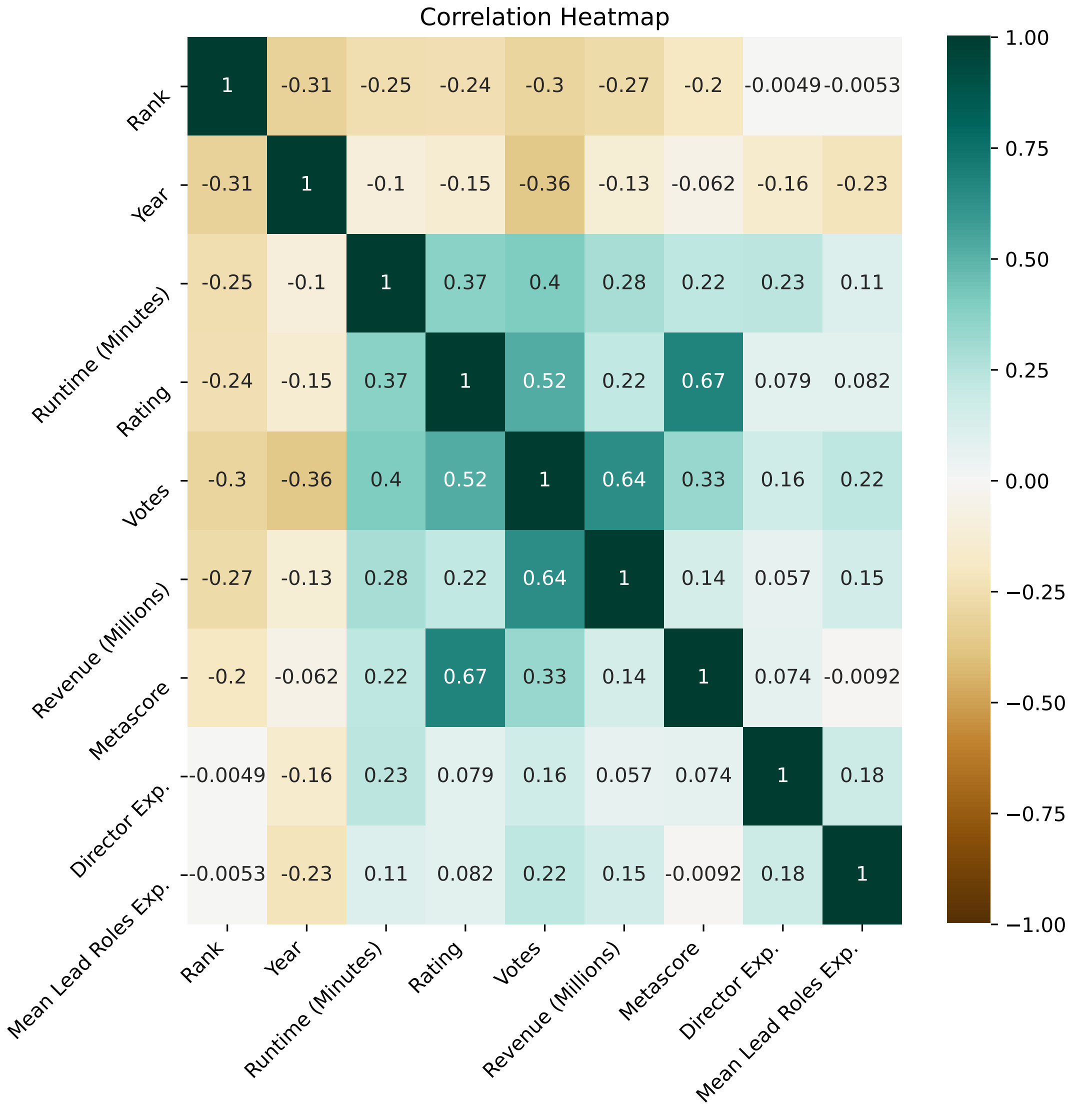

Figure 6: Demonstration figure. This caption explains more about the figure. Note that the font size of the labels in the plot is 9pt, which is obtained by the settings as shown in the Jupyter notebook.

4 Exploration and Analysis

A data science analysis of the paper, including:

- Visualisations (for example Figure 6) 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 ore more of the statistical and ML methods learned in the FDS to the data
- Interpretation of the findings

You can use equations like this:

$$\bar{x} = \sum_{i=1}^{n} x_i \tag{1}$$

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 ith 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 [4].

Improvements and extensions

References

[1] Francis J Anscombe. "Graphs in statistical analysis". In: *The American Statistician* 27.1 (1973), pp. 17–21.

[2] Rounak Banik. *The Movies Dataset*. 2018. URL: `https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download`.

[3] Laha Chakravarti and Roy. *Handbook of Methods of Applied Statistics*. Vol. 1. John Wiley and Sons, 1967, pp. 392–394.

[4] GONZÁLEZ. *Plagiarism – Wikipedia, The Free Encyclopedia*. Last accessed 22 July 2004. 2004. URL: `https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350`.

[5] Ivan Gonzalez. *1000 IMDB movies (2006-2016)*. Scraped from https://IMDB.com. 2023. URL: `https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016`.

[6] 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	Employment	Illness	Attainment	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	…
…	…	…	…	…	…	…	…