Mini Project 4 - Data Analytics

Trevor Wylezik

November 24, 2024

Abstract

In a world that is becoming more and more technical, data is one of the most important aspects of a company or government. However, collecting all of this data has to be put somewhere and eventually sifted through. Data analytics is the process of using tools and technologies to collect insight from data to solve problems. For example, the insurance industry has specific teams of actuaries to perform data analytics on the insurance claims incurred to summarize the data, clear the noise, and update their prices accordingly. In this project, we will investigate two large datasets from iMDb and Rotten Tomatoes on movies. These datasets will include key information like the title of the movie, genre, rating certificate, critics/audience scores, and much more. Using tools like excel and python, we can find trends in the data to understand the movies industry at a higher level. Graphs and charts will be produced to have a convenient way to understand these trends.

1 Introduction

In this report, we will investigate two large datasets from iMDb and Rotten Tomatoes. These are movie review companies that keep track of the movies they review and their critics scores as well as the audience scores for the movies. First, we will explore the datasets in excel to understand the big picture. Then, the data will be cleaned in python to produce a more convenient dataset to work with. This will be done for both datasets. Finally, certain topics can be investigated and charts can be produced to help investigate the differences between the two companies and trends within their own data.

This project is organized as follows: in Section 2, we will highlight the problem statement. In section 3, we will introduce the datasets and go deeper in to the cleaning of the dataset. Section 4 and 5 will be the final results and charts of the project and a discussion about the results.

2 Problem Statement

The goal of this project is to investigate the movies that have been reviewed by iMDb and Rotten Tomatoes to see if there are discrepancies between the two companies. Each company individually will also be examined to compare the critics and audience members.

3 Methodology

In this section, we will dive in to the data collection/cleaning and creation of charts. Two different datasets will be investigated in this project: iMDb and Rotten Tomatoes. Between both datasets, they did not collect the same exact data. For example, Rotten Tomatoes has collected the studio of the movies but iMDb has not. This means we have to break each dataset in to their own parts and can't just clean both in the same way.

3.1 Rotten Tomatoes

3.1.1 An overview of the data

Looking at the Rotten Tomatoes dataset, the following information has been collected:

(I) Movie title

(II) Certificate (PG, PG-13, R, etc.)

(III) Director

(IV) Runtime

(V) Studio

(VI) Rotten Tomatoes Critics Score

(VII) Rotten Tomatoes Audience Score

(VIII) Number of Rotten Tomatoes Critic Reviews

(IX) Number of Rotten Tomatoes Audience Reviews

There are 2,000 movies that have been reviewed with the above information collected. An image is attached below to help visualize this.

	A	B	C	D	E	F	G	H	I
1	Title	Certificate	Director	Runtime	Studio	RT Critics Score	RT Critic Reviews	RT Audience Score	RT Audience Reviews
2	The Angry Birds Movie 2	PG (for rude humor and	Thurop Van Orman	100 minutes	Columbia Pictures	73%	107	84%	4,023
3	Legend Of The Demon (	NR	Kaige Chen	129 minutes	Well Go USA	91%	11	37%	74
4	Dora and the Lost City (	PG (for action and some	James Bobin	102 minutes	Paramount Pictures	84%	148	88%	6,715
5	Luce	R (for language through	Julius Onah	109 minutes	NEON	91%	151	77%	284
6	Good Boys	R (for strong crude sexu	Gene Stupnitsky	95 minutes	Universal Pictures	80%	237	86%	13,007
7	Danger Close	R (for sequences of war	Kriv Stenders	118 minutes	Saban Films	67%	30	67%	90
8	Tel Aviv on Fire	NR	Sameh Zoabi	97 minutes	Cohen Media Group	90%	58	86%	28
9	Brian Banks	PG-13 (for thematic con	Tom Shadyac	99 minutes	Bleecker Street	61%	69	97%	1,831
10	The Farewell	PG (for thematic materia	Lulu Wang	98 minutes	A24	98%	322	87%	2,490
11	Angel Has Fallen	R (for violence and langu	Ric Roman Waugh	114 minutes	Lionsgate	39%	178	93%	14,666
12	Them That Follow	R (for some disturbing vi	Britt Poulton, Dan Ma	98 minutes	1091	59%	91	47%	147
13	After the Wedding	PG-13 (for thematic mat	Bart Freundlich	110 minutes	Sony Pictures Classics	44%	153	77%	196
14	Official Secrets	R (for language)	Gavin Hood	112 minutes	IFC Films	83%	156	89%	361
15	The Peanut Butter Falcc	PG-13 (for thematic con	Tyler Nilson, Michael	93 minutes	Roadside Attractions	96%	200	96%	4,704
16	Paradise Hills	NR	Alice Waddington	94 minutes	Samuel Goldwyn Films	64%	61	82%	330
17	Adopt a Highway	NR	Logan Marshall-Greer	81 minutes	RLJE Films	68%	31	73%	67
18	Cubby	NR	Mark Blane, Ben Man	83 minutes	Breaking Glass Picture	70%	10	62%	16
19	Drive	NR	Tarun Mansukhani	118 minutes	Netflix	0%	7	13%	15
20	The King	R (for some strong violer	David MichÃ´d	133 minutes	Netflix	71%	136	83%	1,794
21	Klaus	PG (for rude humor and	Sergio Pablos	98 minutes	Netflix	94%	63	97%	1,837
22	Earthquake Bird	R (for some sexuality, fu	Wash Westmoreland	106 minutes	Netflix	50%	40	55%	173
23	American Son	NR	Kenny Leon	90 minutes	Netflix	50%	24	47%	282
24	Philophobia: Or the Fea	NR	Aaron Burt	85 minutes	Gravitas Ventures	100%	10	79%	14

Figure 1: The first 23 rows of the Rotten Tomatoes dataset

3.1.2 Cleaning the dataset

As seen from Figure 1, the data is not a form that is convenient for analysis. For example, the director column will have multiple names on rare occasions. Since this can throw off the analysis, we will reduce this column to only having one director.

Overall, this will be done for many of the columns such as Certificate, Director, RT Critics Score, and RT Audience Score. A shortened example table showing the original and desired clean dataset is shown below:

Original Data:

Title	Certificate	Director	Critics Score	Audience Score
Light of My Life	R (for violence)	Casey Affleck	34%	44%
Plus One	NR	Jeff Chan, Ari Gold	88%	81%

We want to take the original data and remove any justifications for the certificate of the movie, extra directors, and percantage symbols.

Cleaned Data:

Title	Certificate	Director	Critics Score	Audience Score
Light of My Life	R	Casey Affleck	34	44
Plus One	NR	Jeff Chan	88	81

This cleaned data set is referenced below in Figure 2:

	A	B	C	D	E	F	G	H	I	J
1		Title	Certificate	Director	Runtime	Studio	RT Critics Score	RT Critic Reviews	RT Audience Score	RT Audience Reviews
2	0	The Angry Birds Movie 2	PG	Thurop Van Orman	100	Columbia Pictures	73	107	84	4023
3	1	Legend Of The Demon Ca	NR	Kaige Chen	129	Well Go USA	91	11	37	74
4	2	Dora and the Lost City of	PG	James Bobin	102	Paramount Pictures	84	148	88	6715
5	3	Luce	R	Julius Onah	109	NEON	91	151	77	284
6	4	Good Boys	R	Gene Stupnitsky	95	Universal Pictures	80	237	86	13007
7	5	Danger Close	R	Kriv Stenders	118	Saban Films	67	30	67	90
8	6	Tel Aviv on Fire	NR	Sameh Zoabi	97	Cohen Media Group	90	58	86	28
9	7	Brian Banks	PG-13	Tom Shadyac	99	Bleecker Street	61	69	97	1831
10	8	The Farewell	PG	Lulu Wang	98	A24	98	322	87	2490
11	9	Angel Has Fallen	R	Ric Roman Waugh	114	Lionsgate	39	178	93	14666
12	10	Them That Follow	R	Britt Poulton	98	1091	59	91	47	147
13	11	After the Wedding	PG-13	Bart Freundlich	110	Sony Pictures Classics	44	153	77	196
14	12	Official Secrets	R	Gavin Hood	112	IFC Films	83	156	89	361
15	13	The Peanut Butter Falcon	PG-13	Tyler Nilson	93	Roadside Attractions	96	200	96	4704
16	14	Paradise Hills	NR	Alice Waddington	94	Samuel Goldwyn Films	64	61	82	330
17	15	Adopt a Highway	NR	Logan Marshall-Green	81	RLJE Films	68	31	73	67

Figure 2: The first 16 rows of the cleaned Rotten Tomatoes dataset

3.2 iMDb

3.2.1 An overview of the data

Looking at the iMDb dataset, the following information has been collected:

(I) Movie Poster

(II) Movie Title

(III) Year

(IV) Certificate

(V) Runtime

(VI) Genre

(VII) iMDb Critics Score

(VIII) iMDb Metascore

(IX) Director

(X) Cast

(XI) Movie Description

(XII) iMDb Critic Reviews

(XIII) Review Title

(XIV) Review

There are 10,000 movies that have been reviewed with the above information collected. An image is attached below to help visualize this.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Poster	Title	Year	Certificate	Runtime	Genre	iMDb Critics Score	iMDb Metascore	Director	Cast	iMDb Audience Reviews	Description	iMDb Critic Reviews	Review Tit	Review	
2	https://m.medi	The Idea of You	2023	R	115	Comedy, Dra	6.4	67	Michael Showalt	Anne Hath	28,744	SolÃ¨ne, a 40-y	166	Hypocrisy	This film, as well a	
3	https://m.medi	Kingdom of the	2023	PG-13	145	Action, Adve	7.3	66	Wes Ball	Owen Tea	22,248	Many years aft	183	A phenom	I'm a big fan of all	
4	https://m.medi	Unfrosted	2023	PG-13	97	Biography, C	5.5	42	Jerry Seinfeld	Isaac Bae	18,401	In 1963 Michig	333	not funny	Pretty much the w	
5	https://m.medi	The Fall Guy	2023	PG-13	126	Action, Come	7.3	73	David Leitch	Ryan Gosl	38,953	A down-and-ou	384	Everything	Just got out of the	
6	https://m.medi	Challengers	2023	R	131	Drama, Rom	7.7	82	Luca Guadagnin	Zendaya, I	32,517	Tashi, a forme	194	Watch "Mi	This is a tough on	
7	https://m.medi	Abigail	2023	R	109	Horror, Thrill	6.8	62	Matt Bettinelli-Ol	Tyler Gille	27,284	After a group o	168	Underwhe	Everything in this r	
8	https://m.medi	Civil War	2023	R	109	Action, Adve	7.5	75	Alex Garland	Kirsten Du	64,014	A journey acro	610	Meander:	I am huge fan of w	
9	https://m.medi	Twisters	2023			Action, Adventure, Thriller			Lee Isaac Chung	Daisy Edgar-Jones, Glen Powell, Da	An update to th		0			
10	https://m.medi	Anyone But You	2023	R	103	Comedy, Ror	6.1	52	Will Gluck	Sydney Sw	82,215	After an amazi	373	Pure esca	This was like an R	
11	https://m.medi	The Ministry of	2023	R	120	Action, Dram	7	57	Guy Ritchie	Henry Cav	21,084	The British mil	117	Fun Movie	As a fan of both He	
12	https://m.medi	Dune: Part Two	2023	PG-13	166	Action, Adve	8.7	79	Denis Villeneuve	TimothÃ©	401,659	Paul Atreides u	1,741	Ladies an	This is the kind of	
13	https://m.medi	Furiosa: A Mad	2023	R	148	Action, Adve	8.3	83	George Miller	Anya Tayl	259	The origin stor	3	Being vast	I had the wonderfu	
14	https://m.medi	Gojira -1.0	2023	PG-13	124	Action, Adve	8	81	Takashi Yamaza	Minami Ha	71,109	Post war Japar	526	Maybe the	I'm a	
15	https://m.medi	Ghostbusters: F	2023	PG-13	115	Adventure, C	6.2	46	Gil Kenan	Paul Rudd	39,558	When the disc	403	Loved it	If you want to hate	
16	https://m.medi	Love Lies Bleedi	2023	R	104	Action, Adve	6.8	77	Rose Glass	Anna Bary	20,636	Gym manager	95	Refreshing	Fantastic movie, g	
17	https://m.medi	The Judge	2023	A	141	Crime, Dram	7.4	48	David Dobkin	Robert Do	204,222	Big-city lawyer	338	Beautiful a	I don't usually writ	
18	https://m.medi	Mother of the Bi	2023		88	Comedy, Dra	4.9	39	Mark Waters	Brooke Sh	5,081	Lana's daughte	70	Poor actir	I knew it was a pre	

Figure 3: The first 17 rows of the iMDb dataset

3.2.2 Cleaning the dataset

First, we can remove a few of the columns that don't apply to the analysis. These columns include poster, cast, description, review title, and review.

The genres in the iMDb dataset have the same issue as the director column in the Rotten Tomatoes dataset from Figure 1. So, the excess genres after the first will be removed.

The certificate column in the iMDb dataset is in terms of a rating system from India. So, a conversion is required to turn it in to the U.S. system that the Rotten Tomatoes dataset uses. In order to do this, we can refer to a table that shows equivalences between the India and U.S. certificate ratings. In python, we can use a string to string replacement. For example, this would replace a string "UA 13+" with "PG-13". This conversion code is shown below:

```
# Turn rating certificate systems over to U.S.
str_2_str = {
    'U': 'G',
    'UA': 'PG',
    'UA 7+': 'PG',
    'UA 13+': 'PG-13',
    'UA 16+': 'R',
    'A': 'R',
    'S': 'NC-17'
    }

IMDB_df['Certificate'] = IMDB_df['Certificate'].replace(str_2_str)
```

Figure 4: The python code for India to U.S. certificate rating conversion

This cleaned data set is referenced below in Figure 5:

	A	B	C	D	E	F	G	H	I	J	K
1		Title	Year	Certificate	Runtime	Genre	iMDb Critics Score	iMDb Metascore	Director	iMDb Audience Reviews	iMDb Critic Reviews
2	0	The Idea of You	2023	R	115	Comedy	6.4	67	Michael Showalter	28744	166
3	1	Kingdom of the Planet of the Apes	2023	PG-13	145	Action	7.3	66	Wes Ball	22248	183
4	2	Unfrosted	2023	PG-13	97	Biography	5.5	42	Jerry Seinfeld	18401	333
5	3	The Fall Guy	2023	PG-13	126	Action	7.3	73	David Leitch	38953	384
6	4	Challengers	2023	R	131	Drama	7.7	82	Luca Guadagnino	32517	194
7	5	Abigail	2023	R	109	Horror	6.8	62	Matt Bettinelli-Olpin	27284	168
8	6	Civil War	2023	R	109	Action	7.5	75	Alex Garland	64014	610
9	7	Twisters	2023			Action			Lee Isaac Chung		0
10	8	Anyone But You	2023	R	103	Comedy	6.1	52	Will Gluck	82215	373
11	9	The Ministry of Ungentlemanly War	2023	R	120	Action	7	57	Guy Ritchie	21084	117
12	10	Dune: Part Two	2023	PG-13	166	Action	8.7	79	Denis Villeneuve	401659	1741
13	11	Furiosa: A Mad Max Saga	2023	R	148	Action	8.3	83	George Miller	259	3
14	12	Gojira -1.0	2023	PG-13	124	Action	8	81	Takashi Yamazaki	71109	526
15	13	Ghostbusters: Frozen Empire	2023	PG-13	115	Adventure	6.2	46	Gil Kenan	39558	403
16	14	Love Lies Bleeding	2023	R	104	Action	6.8	77	Rose Glass	20636	95
17	15	The Judge	2023	R	141	Crime	7.4	48	David Dobkin	204222	338
18	16	Mother of the Bride	2023		88	Comedy	4.9	39	Mark Waters	5081	70
19	17	IF	2023	PG	104	Animation	6.8		John Krasinski	906	15
20	18	Megalopolis	2023		133	Drama			Francis Ford Coppola		1
21	19	Late Night with the Devil	2023	R	93	Horror	7.1	72	Cameron Cairnes	43902	278
22	20	Laapataa Ladies	2023		122	Comedy	8.5		Kiran Rao	20957	209

Figure 5: The first 20 rows of the cleaned iMDb dataset

3.3 Merging the data sets

We can compare the two datasets more closely by looking at all of their common movies. We can achieve this by joining the two datasets based on the movie title. This will effectively be an anchor for all other data points to fill in to a new table based on that anchor.

However, we do need to remove some duplicate columns that are in both datasets. In python, we can achieve this by using a data-frame merge function shown below:

```
# Remove duplicate columns (Allows us to merge cleanly)
IMDB_df = IMDB_df.drop(['Runtime', 'Director', 'Certificate', 'Unnamed: 0'], axis=1)
RT_df = RT_df.drop(['Unnamed: 0'], axis=1)

# Merge the datasets
Merged_df = pd.DataFrame.merge(IMDB_df, RT_df, on='Title')
```

Figure 6: A python function that merges dataframes based on a shared column

4 Results

4.1 Five Number Summary

A five number summary has been produced for each of the critic's and audience's scores.

Five Number Summary:

Data	Minimum	25th Percentile	Median	75th Percentile	Maximum
RT Critics	0	42	71	88	100
RT Audience	5	43	61	76	100
iMDb Critics	13	58	65	72	97
iMDb Metascore	1	45.5	58	71	100

Initially, it may be confusing looking at the iMDb Metascore, since it has no equivalent name in the Rotten Tomatoes dataset. On top of this, it doesn't seem to closely align with any of the other five number summaries. Another area to investigate is the mean and standard deviations of these data point. In the next subsection, we can visualize all of these values using box plots.

4.2 Box Plots

In Figure 7 there are four different box plots. Each vertical line represents a specific value. Going from left to right, they represent the minimum value, 25th percentile, median, 75th percentile, and maximum value. The mean of each box plot is also displayed by a green triangle.

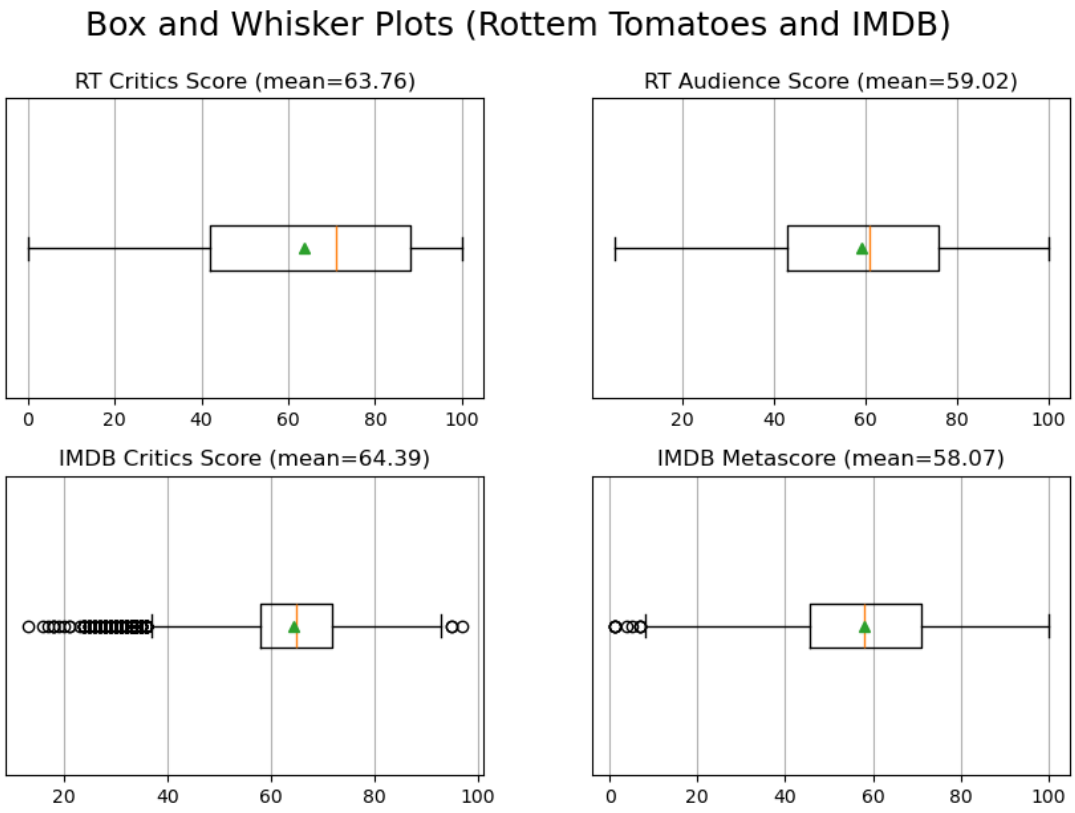

Figure 7: Four box plots of Rotten Tomato and iMDb review scores (with means)

Looking at the box plots, there are a few key takeaways. First, the iMDb metascore was confusing in the last subsection, however it seems like it is comparable to the Rotten Tomatoes audience score. This is due to the means having similar values.

Second, it seems like the Rotten Tomatoes movie scores have a larger standard deviation, as there are no detected outliers in their box plots. In other words, their ratings are more spread out across all possibilities. On the other hand, iMDb seems to have a smaller standard deviation as the boxes are smaller (25th and 75th percentiles). There are also visible outliers towards the lower end of the scores. This suggests that iMDb is more favorable to movies, but will occasionally step out of their normal trends and rate some movies on the low end (36 and below for critics).

4.3 iMDb Pie Chart

A pie chart is useful for gauging relative sizes of categories. With a pie chart (Figure 8), we can visualize which movie genres have the most reviews from iMDb.

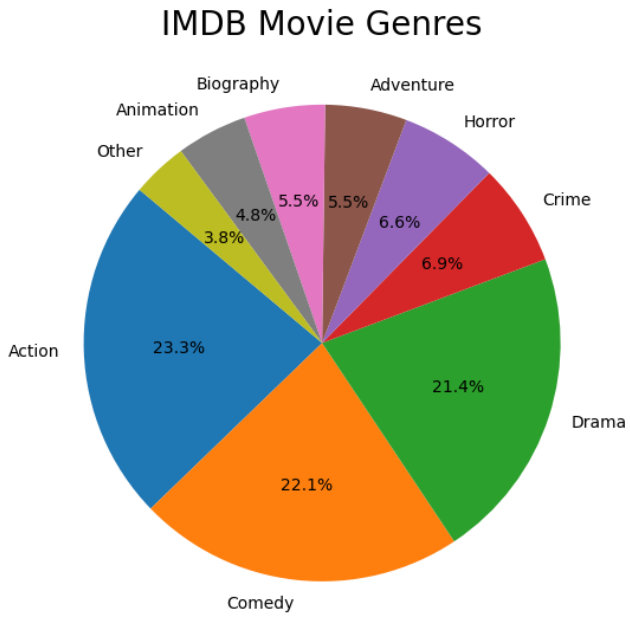

Figure 8: A pie chart of iMDb movie genre counts

It is clear to see that action, comedy, and drama movies are very common. Crime, horror, adventure, biography, and animation movies are also somewhat common. Finally, there are many movie genres in the other category are rare to see getting a review from iMDb. The following genres are rare: documentary, fantasy, thriller, mystery, sci-fi, romance, western, musical, film-noir, family, history, music, sport, and war.

Note: There is no chart for Rotten Tomatoes since their dataset doesn't include genres.

4.4 iMDb Stacked Histogram

Another idea to investigate is how iMDb had reviewed movies with time. The stacked histogram in Figure 9 shows the number of each genre for every year 2000-2025.

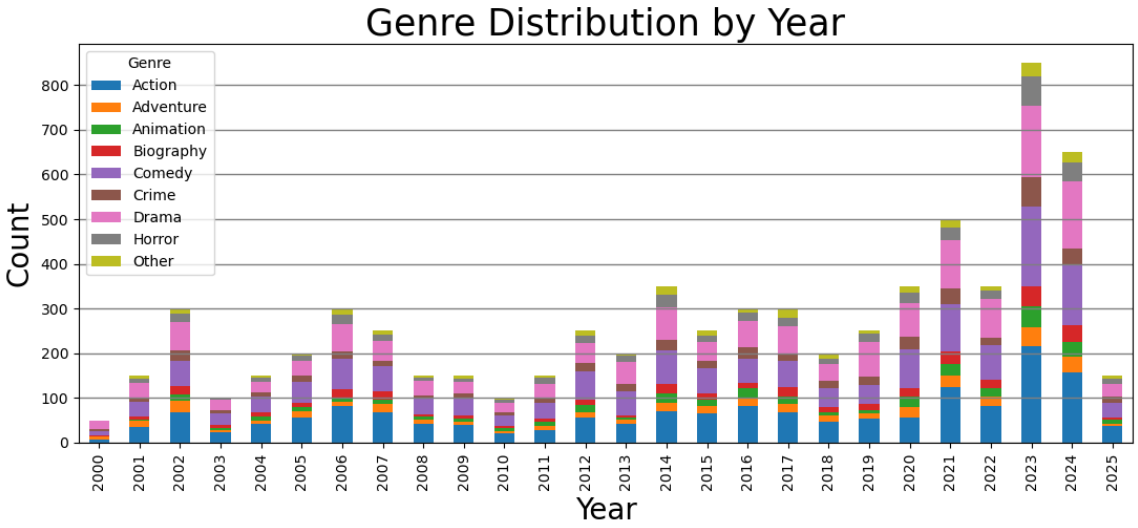

Figure 9: A pie chart of iMDb movie genre counts

Overall, it seems that the amount of movies has gone up overtime. What is interesting though, are the peaks in years like 2002, 2006, 2014, 2021, 2023, and 2024. For 2023, it is suspected that this is due to movie production being halted from COVID-19, and a surge of movies releasing once the pandemic slowed down.

It is surprising to see how the proportion of movies over the years has not changed that much. It seems like the amount of each genre is increasing/decreasing proportional to the total number of movies.

Something to note, though, is that its unclear if this is due to the number of movies being released, or just the proportion that iMDb is covering. Unfortunately, this is a flaw in the dataset.

4.5 Rotten Tomatoes Bar Chart

In the U.S., movies are given ratings that tell the audience the intended audience of the movie. For example, G is the lowest that means everyone of all ages can watch it. On the higher end, there is rated R movies where someone under the age of 18 must be accompanied by an adult to watch the movie. And finally, there is NR which stands for not rated.

We can visualize how Rotten Tomatoes rates movies based on their movie certificates in Figure 10 by using a bar chart.

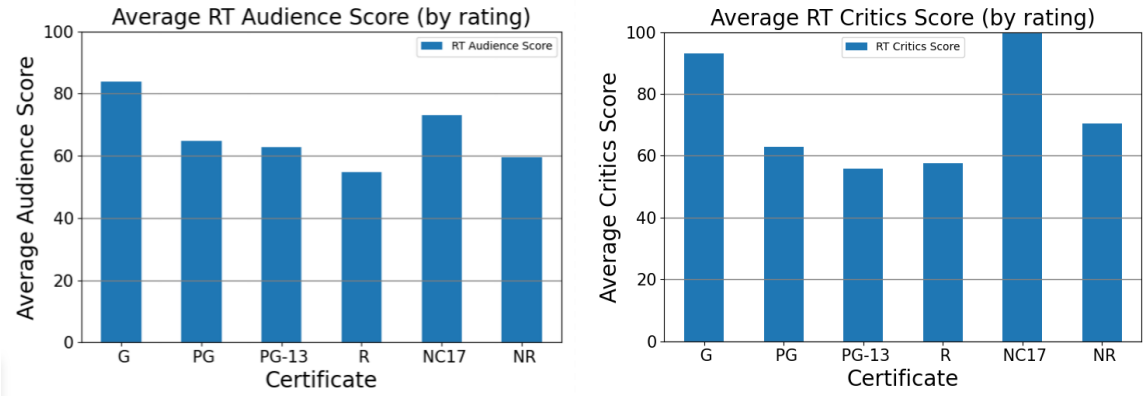

Figure 10: A bar chart of Rotten Tomatoes scores by movie certificate

There are a few things to point out about this chart. Overall, it seems like the ratings of movies decrease as the movie becomes more mature. This may seem inconsistent, though, due to the NC17 rated movies having an average RT critics score of 100. Peaking behind the curtain, though, and we can see where this discrepency is coming from:

Rotten Tomatoes Critics Score by Certificate:

Certificate:	**G**	**PG**	**PG-13**	**R**	**NC17**	**NR**
Avg. RT Critics Score	93.17	62.97	55.76	57.69	100	70.32
Avg. RT Critics Score	84	64.76	62.81	54.90	73	59.54
# of Movies	6	143	316	656	1	978

It is clear to see that G and NC17 rated movies are uncredible, due to them having a sample size of 6 and 1 respectively. This is almost nothing compared to the R rated movies having a sample size os 656.

4.6 Merged Dataset Box Plot

We can refine our analysis more by only focusing on movies that both companies has reviewed. This new merged data set contains 741 movies, which is credible enough to create a box plot.

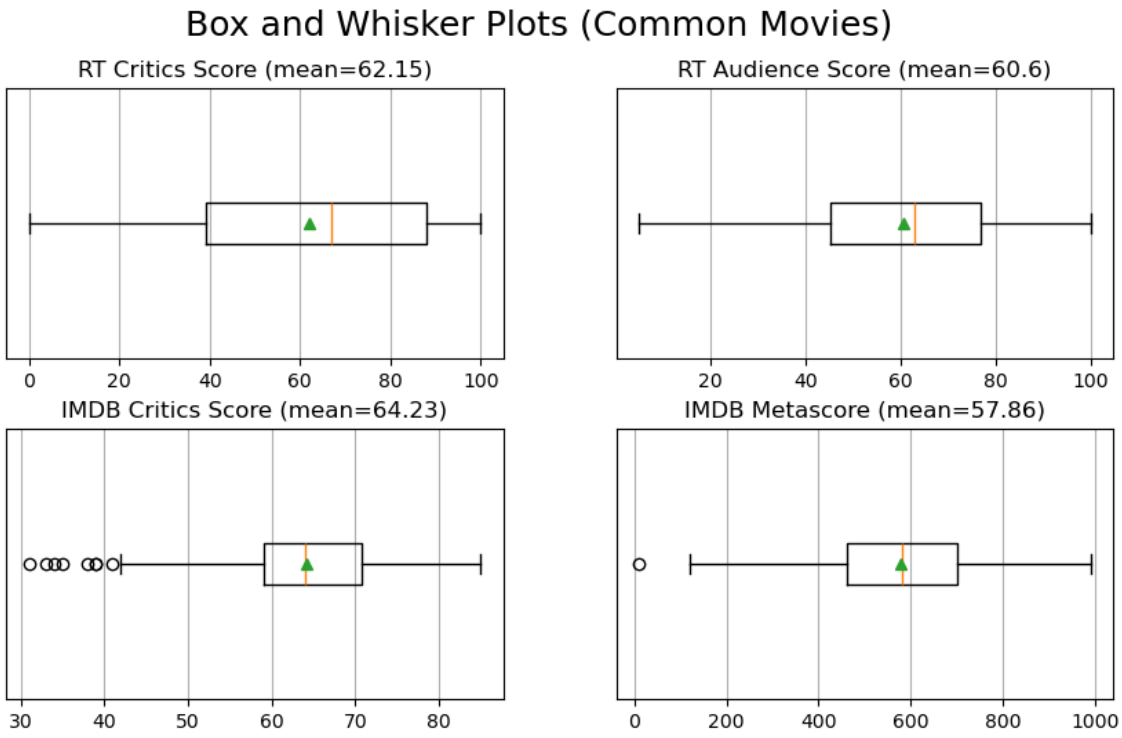

Figure 11: A box plot of the merged data set comparing Rotten Tomatoes and iMDb

Compared to the previous box plot in Figure 7, there is little to no change in the box plots. Due to the merging of the datasets reducing the number of movies, it looks like iMDb has fewer outliers, which is to be expected. The five number summaries and means have no practical difference.

5 Discussion/Conclusion

Overall, the Rotten Tomatoes and iMDb datasets performed similarly to each other. It seems that Rotten Tomatoes and iMDb's critics are more favorable to movies while the audience is less favorable to them. It is hard to tell why that may be, but it may have to do with critics having more of an eye for artistic ability, while a general audience may not be thinking about so many factors when watching a movie. On top of this, it is very apparent that iMDb is way may consistent with critics and metascore's than Rotten Tomatoes. This is evident by looking at both box plot versions and seeing the length of each box and whisker.

Another interesting thing to look at is the breakdown of certificate rating and Rotten Tomatoes scores. Overall, the initial hypothesis was that critics and audience scores rating would be down as movies got more mature, more-so for the critics score. But, this was not seen at all. It seems that both the critics and audience liked NR rated movies more than anything other than G rated movies. The only exception to this are the NC17 movies, but those can be ignored since they only have a sample size of one.

6 Resources

1. Rotten Tomatoes Dataset

2. iMDb Dataset