Delanie Dahm, Melanie Gradeler

Project Report

Analyzing Top 250 Movie Data

1. Introduction

Have you ever wondered which movies stay ranked in IMBDs top 250 movies list overtime? Or which of those movies won the most awards? In this project we aim to analyze IMBD data to see what factors contribute to their top 250 movies list. We will be using a 2021 list obtained from Kaggle titled *IMBD Top 250 Movies Dataset*[1] as well as scraping the most current list from *IMBD.com*[2] and scraping additional metrics of the movies not available in our downloaded dataset. Using this data, we aim to compare how the list has changed since 2021, determined which movies maintained their rank and explore correlations of these new insights.

2. Data

2.1 IMBD Overview

The Internet Movie Database (IMBD) is one of the biggest public platforms providing users with comprehensive data and information on their favorite movies. *IMBD.com* provides it users with data such as reviews, awards, cast information, production details, and variety of ratings. These ratings include, the IMBD rating (based on worldwide user rating), the IMBD rank (based on IMBD rating), the metascore (based on a combination of user and critic reviews) and popularity score (measured level of engagement with the title on IMBD and seasonal relevance). These metrics are widely used and recognized as benchmarks for the overall success of a movie. We plan to use these to measure the change of a movies' public perception from 2021 to 2024.

2.2 Cleaning IMBD Kaggle 2021 Data

We obtained our first dataset in the form of a *.csv file* found on Kaggle.com titled "*IMBD Top 250 Movies Dataset*". We renamed this dataset "*2021_List_Top_250_Movies_Cleaned.csv*" after cleaning, accounting for any necessary filtering, data type changes and removal of columns unnecessary for our analysis. This dataset was previously scraped from IMBD.com in 2021 based on ratings, and information such as title, release year, director, and cast. None of the data was altered or modified in any way at the time of collection and was collected in accordance with IMBD's terms of use. The dataset provides a snapshot of the highest rated movies in 2021 to gain insights on the movie industry, trends, and popular genres at the time.

2.3 Cleaning IMBD Kaggle 2024 Data

Our second dataset was created by scraping IMBD's 2024 Top 250 Movies list. The data collected during this process includes all listed movie titles, release years, runtimes, ratings and URLs. We identified some key metrics left out of the 2021 dataset that we would like to further observe including an updated rank, metascore, popularity score and number of Oscars won by each movie. Due to the steps required to scrape additional details, the raw data will be split into

[1] IMBD Top 250 Movies Dataset

[2] IMBD.com

two separate CSV files. The first file, titled "*top_250_raw.csv*" will contain all the information about the movies found directly on the list e.g title, rank, rating, runtime, and year. We merged the raw scraped "top_250_raw.csv" with its corresponding details csv by creating a structured and clean version of the data. This involves splitting, reordering, and formatting the columns to make the data suitable for analysis while ensuring proper tracking of ranks and ratings. The second file, titled "*top_250_details_raw.csv*" will include additional details from each movie's individual page, such as the metascore, popularity score, and the number of Oscars won. We processed this information to create a clean version of the data. This process includes restructuring, reordering, and formatting columns to ensure the data is well suited for analysis, with proper tracking of ratings and ranks. After cleaning "*top_250_cleaned.cvs*" and "*top_250_details_cleaned*", we **merged horizontally** using an **inner join** to match corresponding URLs. We used an inner join here as all the URLs match and there is no additional data to be concerned about. This created a new simplified and clean data file titled "*2024_List_Top_250_Movies_Cleaned_Merged.csv*". We do not plan to update box office revenue as of 2024 due to the continuous change in gross revenue. We will do any revenue calculations upon 2021 data to ensure consistent and accurate information.

2.4 Merging Cleaned Datasets

Now that we have two clean sets of data, how will we combine them to perform our analysis? We recognize that some of the movies ranked in 2021 may no longer be ranked in 2024 and the same can be said for the inverse as the past analysis could not account for movies that did not exist at the time. To ensure we are merging our data accurately we aimed to simplify our approach by merging "*2021_List_Top_250_Movies_Cleaned.csv*" to "*2024_List_Top_250_Movies_Cleaned_Merged.csv*" **horizontally**, with an **inner join** using the title, runtime, and year columns. This type of join ensures only matching attributes are aggregated and the final dataset includes only movies that are present in both the 2021 and 2024 lists. It also takes additional steps to reorder and clean the columns so they are easier to see for the viewer. This new file is titled "2021_2024_List_Union.csv" also contains null values for any numerical columns that did not have the requested information available and will be used to perform our final analysis.

Delanie Dahm, Melanie Gradeler

Table 1 Data Dictionary:

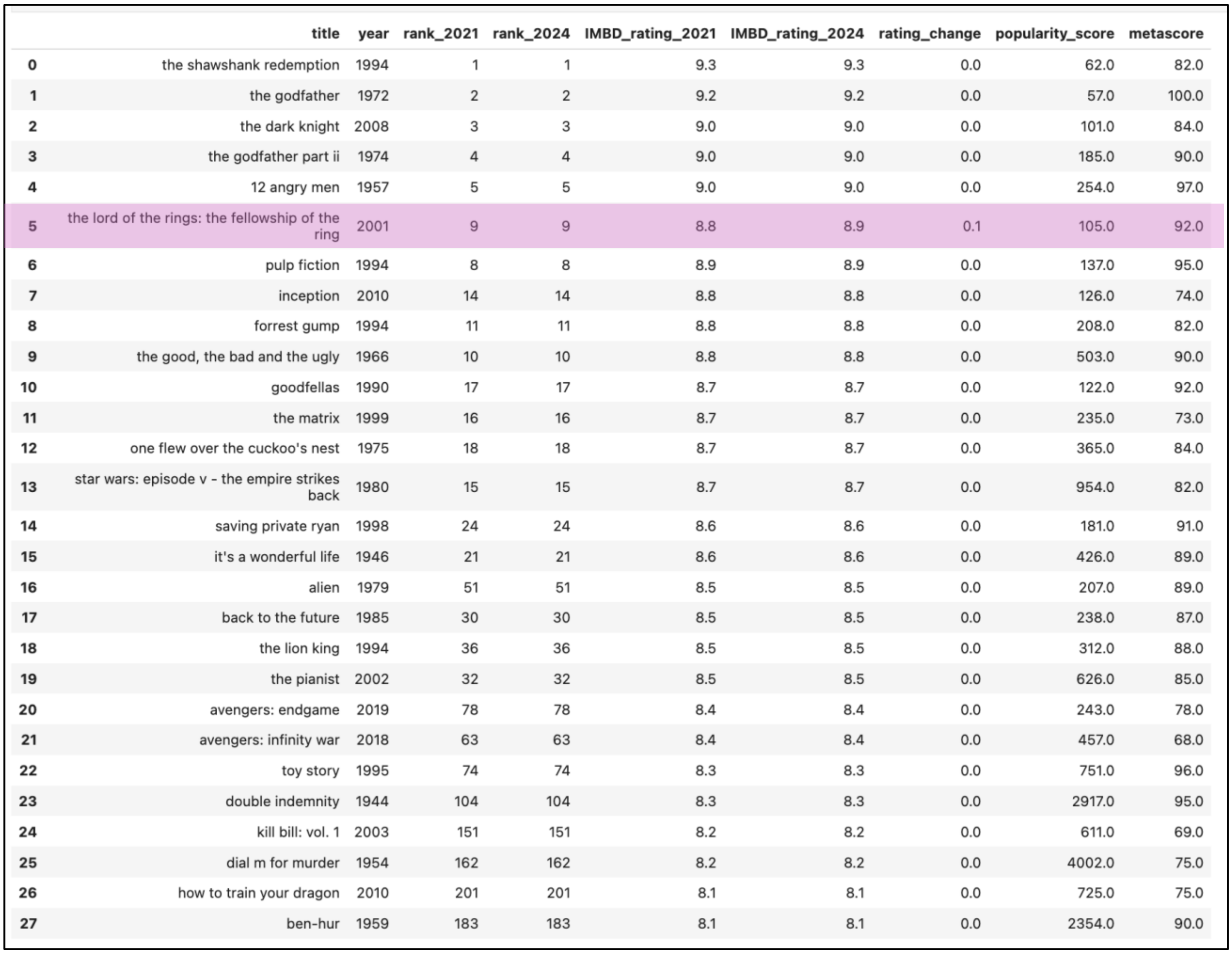
Field	***Type***	***Source***	***Description***
rank_2021	Integer	Kaggle dataset	This is the ranking of each movie in 2021 (based on IMBD user rating)
rank_2024	Integer	IMBD.com	This is the ranking of each movie in 2024 (based on IMBD user rating)
title	Object	Both	Title of the movie
year	Integer	Both	Year the movie was released
run_time	Integer	Both	The duration of the movie in minutes
IMBD_rating_2021	Float	Kaggle dataset	User rating of the movie on IMBD scale (1-10)
IMBD_rating_2024	Float	IMBD .com	User rating of the movie on IMBD scale (1-10)
average_rating	Float	both	Average of the user rating on the IMBD scale (1-10) between 2021 and 2024.
genre	Object	Kaggle dataset	Genres of the movie
box_office	Float	Kaggle dataset	Total box office revenue collection across the world (in $ USD)
budget	Float	Kaggle dataset	The total cost of producing the movie (in $ USD)
certificate	Object	Kaggle dataset	Movie Certificate (PG-13, R, PG, etc.)
tag_line	Object	Kaggle dataset	A slogan/catchphrase used to promote the movie
cast	Object	Kaggle dataset	List of main actors and actresses or anyone that appeared as part of the cast in the movie
directors	Object	Kaggle dataset	The person responsible for overseeing the creative aspects of the film
writers	Object	Kaggle dataset	List of writers that created and wrote the script/ story of the movie.
popularity_score	Integer	IMBD.com	Additional ranking encompassing present user interaction with the movies their relevance. This a separate list updated weekly that changes based on season trends and consumer behavior. It ranks all movies for how watched, relevant and discussed they still are at the time of viewing the list. The higher the popularity score, the lower it appears on this list, meaning it was less popular during the time/season the user is looking at the specific movie.
metascore	Integer	IMBD.com	Numeric score that combines of user and critic ratings and reviews (1-100)
oscars	Object	IMBD.com	Number of Oscars won by the movie
url	Object	IMBD.com	Link to each movie’s page pulled from the list

3. Analysis

For our analysis, we performed everything within a Jupyter notebook titled "*Mgradeler_Ddahm _T250_Movies_Analyzing_P02.ipynb*". This would be the third file to run in order to complete the project from start to finish. Within this notebook you will find a brief introduction followed by sections of code separated by each research question posed. There is a brief explanation of each question followed by the corresponding code blocks. We used a multitude of tools to answer these questions including additional filtering, correlation calculations, and plots for better visibility. We will be describing and analyzing the findings of each question paragraphs below.

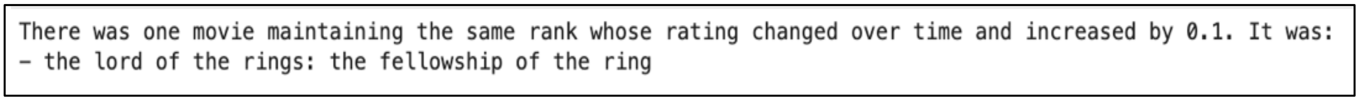
For our analysis, we wanted to see which of the movies that were ranked in 2021 and are still ranked in 2024. We aimed to compare their current to previous ratings and ranks, correlations between different scores, Oscars or budgets and which genres and were more successful in terms of profit. Our research questions are as follows:

3.1 Which movies maintained the same IMBD rank in 2024 and 2021? Of these movies, how much did their IMBD ratings change at all?



	title	year	rank_2021	rank_2024	IMBD_rating_2021	IMBD_rating_2024	rating_change	popularity_score	metascore
0	the shawshank redemption	1994	1	1	9.3	9.3	0.0	62.0	82.0
1	the godfather	1972	2	2	9.2	9.2	0.0	57.0	100.0
2	the dark knight	2008	3	3	9.0	9.0	0.0	101.0	84.0
3	the godfather part ii	1974	4	4	9.0	9.0	0.0	185.0	90.0
4	12 angry men	1957	5	5	9.0	9.0	0.0	254.0	97.0
5	the lord of the rings: the fellowship of the ring	2001	9	9	8.8	8.9	0.1	105.0	92.0
6	pulp fiction	1994	8	8	8.9	8.9	0.0	137.0	95.0
7	inception	2010	14	14	8.8	8.8	0.0	126.0	74.0
8	forrest gump	1994	11	11	8.8	8.8	0.0	208.0	82.0
9	the good, the bad and the ugly	1966	10	10	8.8	8.8	0.0	503.0	90.0
10	goodfellas	1990	17	17	8.7	8.7	0.0	122.0	92.0
11	the matrix	1999	16	16	8.7	8.7	0.0	235.0	73.0
12	one flew over the cuckoo's nest	1975	18	18	8.7	8.7	0.0	365.0	84.0
13	star wars: episode v - the empire strikes back	1980	15	15	8.7	8.7	0.0	954.0	82.0
14	saving private ryan	1998	24	24	8.6	8.6	0.0	181.0	91.0
15	it's a wonderful life	1946	21	21	8.6	8.6	0.0	426.0	89.0
16	alien	1979	51	51	8.5	8.5	0.0	207.0	89.0
17	back to the future	1985	30	30	8.5	8.5	0.0	238.0	87.0
18	the lion king	1994	36	36	8.5	8.5	0.0	312.0	88.0
19	the pianist	2002	32	32	8.5	8.5	0.0	626.0	85.0
20	avengers: endgame	2019	78	78	8.4	8.4	0.0	243.0	78.0
21	avengers: infinity war	2018	63	63	8.4	8.4	0.0	457.0	68.0
22	toy story	1995	74	74	8.3	8.3	0.0	751.0	96.0
23	double indemnity	1944	104	104	8.3	8.3	0.0	2917.0	95.0
24	kill bill: vol. 1	2003	151	151	8.2	8.2	0.0	611.0	69.0
25	dial m for murder	1954	162	162	8.2	8.2	0.0	4002.0	75.0
26	how to train your dragon	2010	201	201	8.1	8.1	0.0	725.0	75.0
27	ben-hur	1959	183	183	8.1	8.1	0.0	2354.0	90.0

Figure 1: Movies that maintained their rank from 2021 to 2024

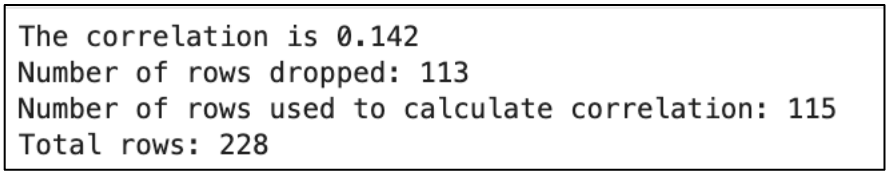


```
There was one movie maintaining the same rank whose rating changed over time and increased by 0.1. It was:
- the lord of the rings: the fellowship of the ring
```

Figure 2: Movie with rating Changes

We identified the movies that maintained the same IMBD rank in both 2021 and 2024 to show the possible consistency over time. To achieve this, we filtered the dataset to include only movies with identical ranks in both years. We made an additional column to calculate the change in ratings, providing a measure of how each movie evolved over the two-year period. The results show key metrics such as title, rank IMBD ratings for 2021 and 2024, popularity scores, and meta scores, to offer a comprehensive view of the movies' performances. Several iconic movies such as The Shawshank Redemption and The Godfather maintained their high IMBD ranks between 2021 and 2024 reflecting their popularity and critical acclaim. An interesting insight we found was that only 28 movies maintained their rank within two years. And of those movies, "*The Lord of the Rings: The Fellowship of the Ring*" was the only rating that changed. It increased by 0.1 stars which seems like a significant jump for a movie released in 2001.

3.2 Is there a correlation between metascore and number of Oscars?



```
The correlation is 0.142
Number of rows dropped: 113
Number of rows used to calculate correlation: 115
Total rows: 228
```

Figure 3: Correlation Calculation of Metascore v. Oscars

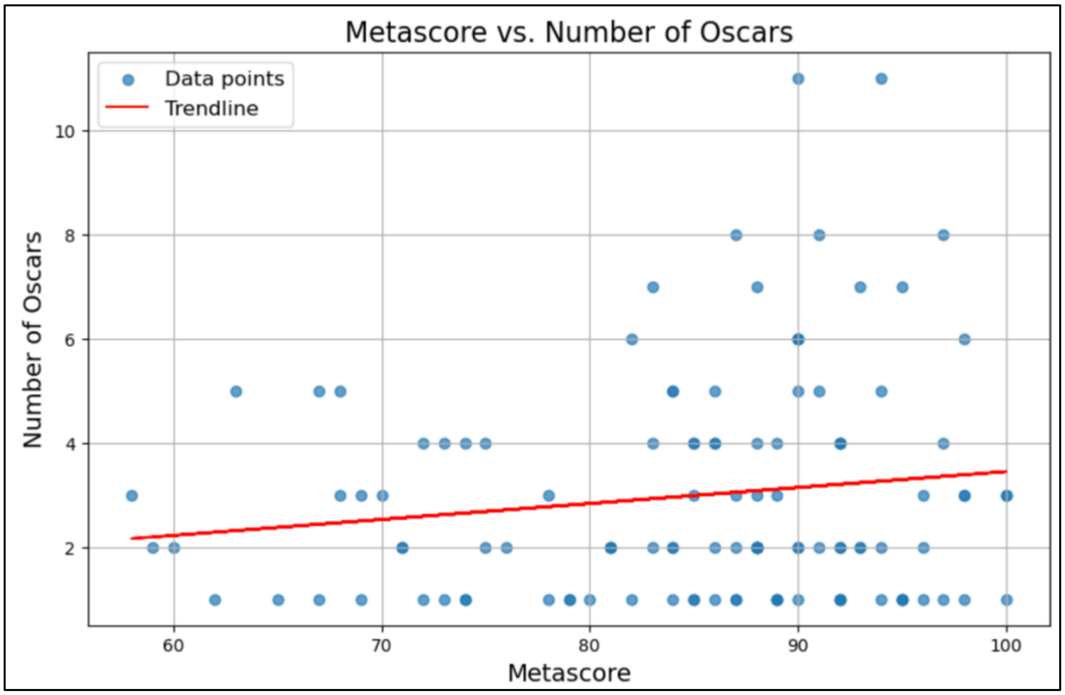
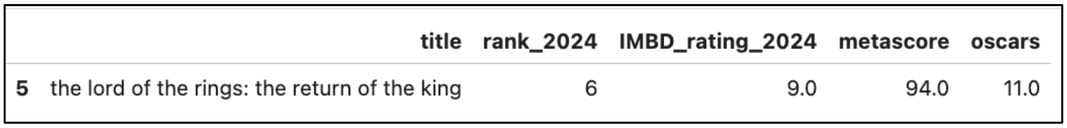


Figure 4: Scatterplot of Metascore vs. Number of Oscars

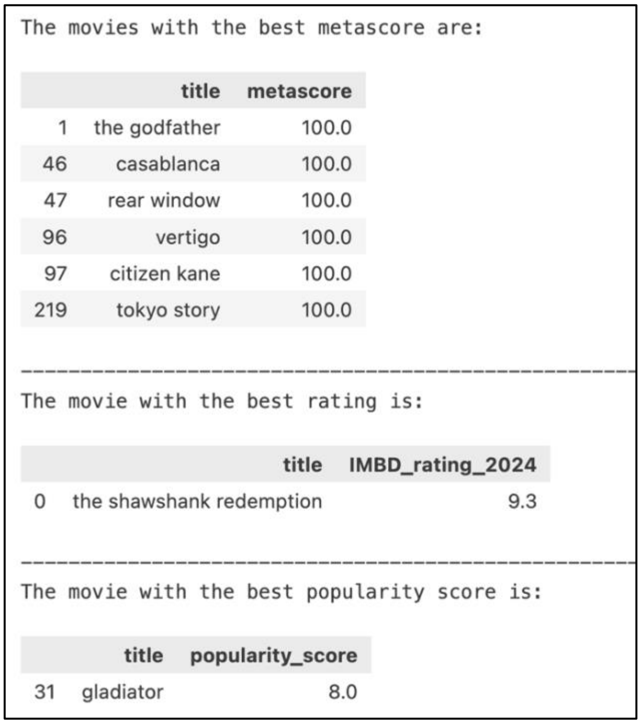


	title	rank_2024	IMBD_rating_2024	metascore	oscars
5	the lord of the rings: the return of the king	6	9.0	94.0	11.0

Figure 5: Movie with the highest Metascore and Oscars won

We aimed to explore the potential relationship between a movie's metascore and the number of Oscars it won. After cleaning the data to address missing values, we computed the correlation coefficient, which was found to be 0.142. This value indicates a weak positive correlation, suggesting that while higher metascores are somewhat associated with an increased number of Oscars, the relationship is not particularly strong. We used a scatterplot with a trendline to represent the relationship and further validate the correlation calculation. This graph shoes that most movies cluster around lower numbers of Oscars, no matter their metascore, though the upward trendline confirms a slight positive association. We wanted to see which one of the outliers in this plot was to see if we could potentially identify a pattern at the end of this analysis. The outlier was once again found to be "*The Lord of the Rings: The Fellowship of the Ring*" with 11 Oscar's won and a metascore of 94.

3.3 __Which movies have the best scores across all three scoring methods (popularity score, metascore, rating)?__



```
The movies with the best metascore are:
```

	title	metascore
1	the godfather	100.0
46	casablanca	100.0
47	rear window	100.0
96	vertigo	100.0
97	citizen kane	100.0
219	tokyo story	100.0

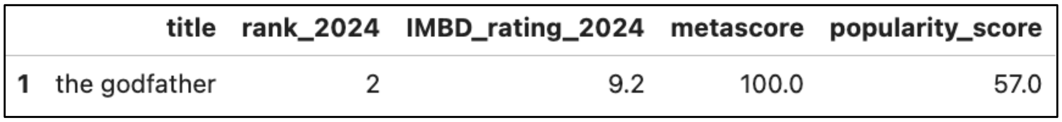
```
--------------------------------------------------------
The movie with the best rating is:
```

	title	IMBD_rating_2024
0	the shawshank redemption	9.3

```
--------------------------------------------------------
The movie with the best popularity score is:
```

	title	popularity_score
31	gladiator	8.0

Figure 6: Top Movies for each Scoring Metric

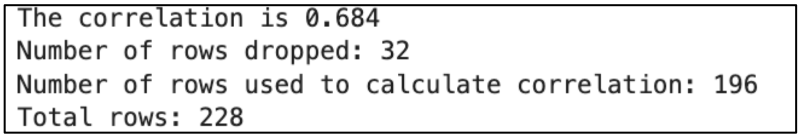


	title	rank_2024	IMBD_rating_2024	metascore	popularity_score
1	the godfather	2	9.2	100.0	57.0

Figure 7: Top Movie across all three metrics

We identified which movies have the best scores across all three scoring methods individually (popularity score, metascore, and rating) and the top movie across all three metrics. These metrics provide a comprehensive view of a movie’s quality, popularity, and critical acclaim. A lower popularity score indicates higher popularity (see data dictionary for context), while higher meta scores and IMBD ratings reflect strong performance in those respective areas. There are a total of six movies with the highest metascore of 100: “*The Godfather”*, “*Casablanca”*, “*Rear Window*”, “*Vertigo*”, “*Citizen Kane*”, and “*Tokyo Story*”. “The *Shawshank Redemption*” had the highest IMBD rating of 9.3 which makes sense as it is ranked #1 in our list. Finally, “*Gladiator*” scored the best for popularity score at 8.0 meaning right now this is in the top 8 most visited movie titles on IMBD. This could be due to the recent release of Gladiator 2 with viewers wanting insights on the original before seeing the sequel in theaters. After these insights, we were left wondering which movie performed the best across all three types. Our findings show that “*The Godfather*” scores the best over all three metrics with a metascore of 100, a IMBD rating of 9.2, and a popularity score of 57. This movie is a total classic, so it is unsurprising to see its overall performance and even though it is an older release from 1972, it clearly remains relevant today with such a low popularity score.

3.4 Is there a correlation between budget and box office revenue? Which movie had the highest box office revenue and what was its budget



```
The correlation is 0.684
Number of rows dropped: 32
Number of rows used to calculate correlation: 196
Total rows: 228
```

Figure 8: Correlation Calculation Box Office Revenue v. Production Budget

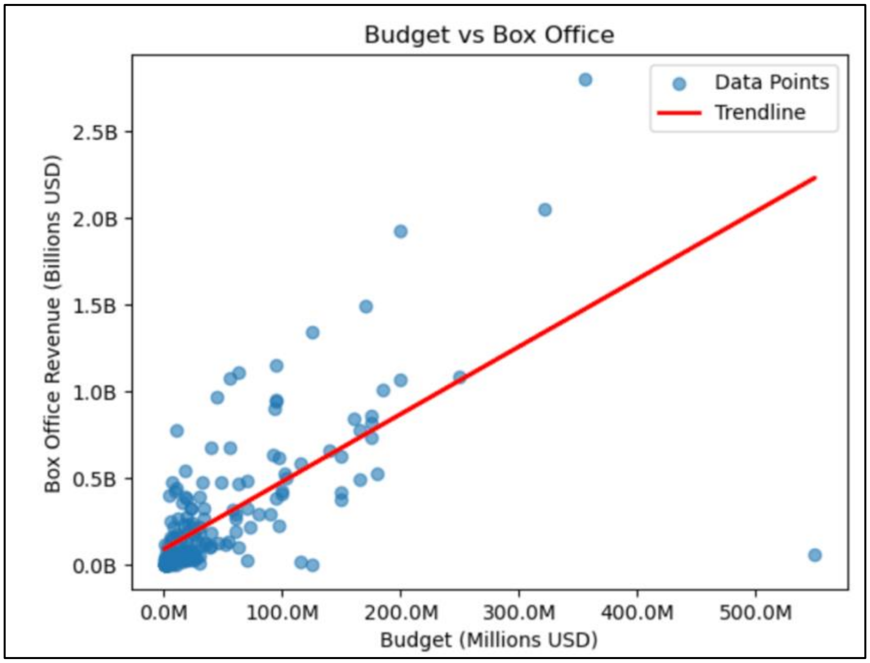
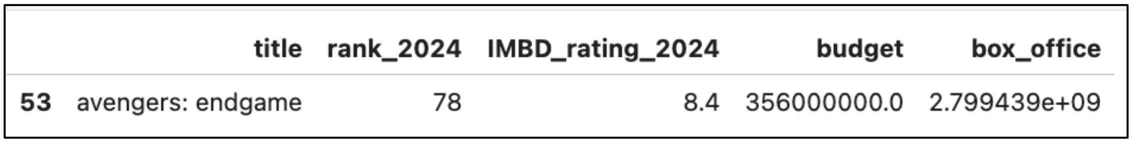
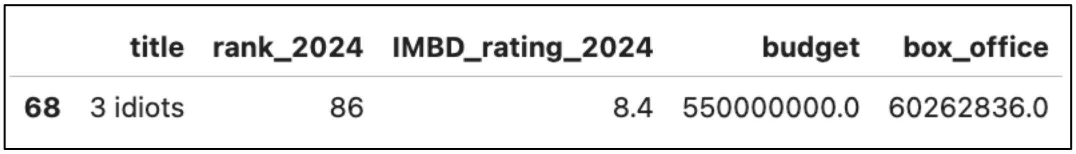


Figure 9: Budget vs Box Office Revenue



	title	rank_2024	IMBD_rating_2024	budget	box_office
53	avengers: endgame	78	8.4	356000000.0	2.799439e+09

Figure 10: Outlier 1 Most Profitable movie



	title	rank_2024	IMBD_rating_2024	budget	box_office
68	3 idiots	86	8.4	550000000.0	60262836.0

Figure 11: Outlier 2 Least Profitable movie

We really wanted to focus on the impact a budget has on a movies box office performance to determine if higher budgets lead to higher earnings. We cleaned the dataset to address missing values, ensuring we have accurate calculations. Then we made a new column to calculate the profit by subtracting production budgets from box office revenues. To further investigate this relationship, we computed a correlation coefficient of 0.684, indicating that there is a strong positive correlation. We created a scatterplot to visualize this data with a trendline to illustrate the upward trend. Additionally, we were curious what some of the outlying movies were. Particularly, we identified "*Avengers: Endgame*" as the most profitable movie with an initial budget of $356,000,000 million and producing over $2.8 billion in box office revenue. On the flip side we wanted to identify the least profitable movie, "3 idiots" an Indian movie that had a production budget of $550,000,000 million but sadly only produced $60,262,836 in box office revenue. This finding shows that while higher budgets often lead to higher revenues, outliers and exceptions still exist, further emphasizing the importance of factors beyond budget alone.

3.5 Which genre was the most profitable overall?

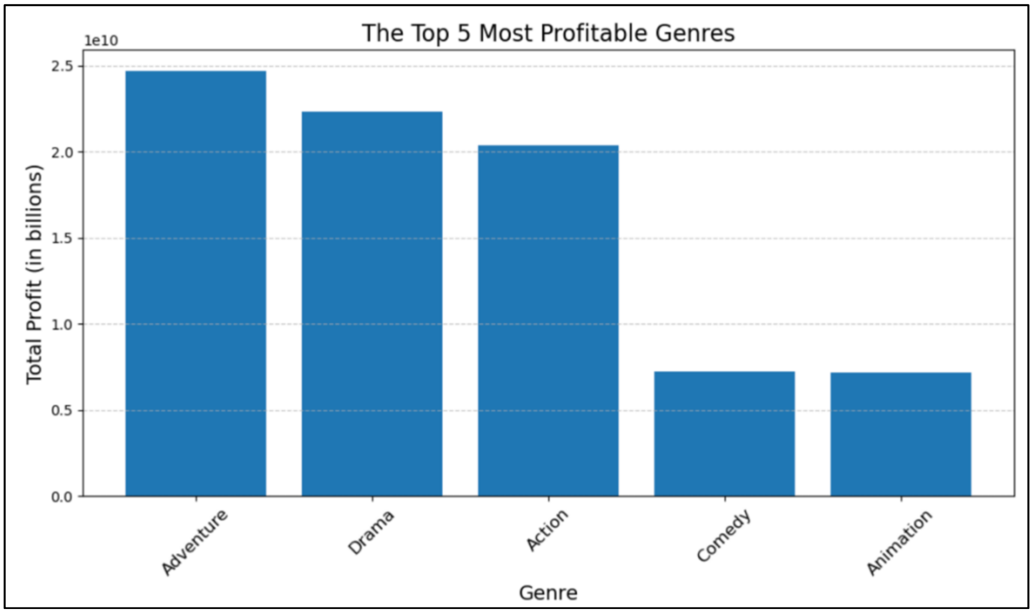
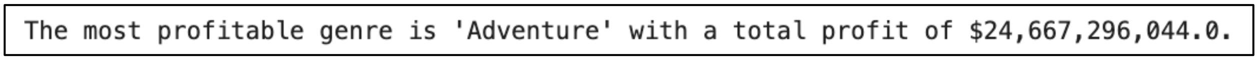


Figure 11: Bar Graph of Most Profitable Genres



```
The most profitable genre is 'Adventure' with a total profit of $24,667,296,044.0.
```

Figure 12: Most Profitable Genre

We wanted to determine profitability each movie genre, revealing valuable insights into financial performance across the film industry. After cleaning the data to address missing values and grouping by genre, profit was calculated as the difference between box office revenue and budget. Since many movies have multiple genres, the genre column was expanded to ensure each genre's contribution was accurately accounted for. The results show that Adventure movies are the most profitable at approximately $24.67 billion, followed by Drama, Action, Comedy, and Animation. The bar chart showcases the significant lead Adventure holds over other genres, while Drama and Action also demonstrate notable profitability, showing that there is a constant appeal from big audiences. The outcome of this analysis is not surprising given the types of movies showing success in previous questions.

4. <u>Conclusion</u>

In this project, we analyzed key aspects of the IMBD Top 250 Movies, list focusing on rank changes, correlations between metrics, and profitability insights. By using datasets from 2021 and 2024, our goal was to find patterns, trends, and contributing factors to a movie's success. Below are the main findings based on the analysis questions posed:

1. *Which movies maintained the same IMBD rank in 2021 and 2024, and how did their ratings change?*

   Only 28 movies maintained their rank, with "*Lord of the Rings: The Fellowship of the Ring*" being the only one to experience a rating change, increasing by 0.1 stars.

2. *Is there a correlation between metascore and the number of Oscars won?*

   A weak positive correlation (0.142) was found, with "The Lord of the Rings: *The Fellowship of the Ring*" standing out as an outlier with 11 Oscars and a metascore of 94.

3. *Which movies scores the best cross popularity score, metascore, and IMBD ratings? Which movie scored the best overall?*

   "*The Godfather*" was the top performer across all three metrics, with a perfect metascore of 100, an IMBD rating of 9.2, and a popularity score of 57.

4. *Is there a correlation between budget and box office revenue? Which movie has the highest revenue? Lowest revenue?*
   A strong correlation (0.684) was identified, with "*Avengers: Endgame*" achieving the highest box office revenue.

5. *Which genre was the most profitable overall?*
   Adventure films were the most profitable, generating approximately $24.67 billion in revenue, followed by Drama and Action.

Overall, the findings of this analysis were very interesting to us, there were a few movies that showed up successful in a few metrics more than once and their genres and scores aligned well with additional findings later. Our analysis faced limitations, such as not having access to 2024 budget updates, as well as access limitations to previous budget and revenue metrics resulting in null values for those respective titles. Had we had all this information, some of the calculations may be slightly different though we believe our missing data did not pose a big risk to the accuracy of our results. Also, for the new metrics in 2024, we only had access to those metrics in real time. We couldn’t analyze the possible change in scores from 2021 to 2024 except for rating since this was the only data available to us at the time.

Some future work that could be performed on this data could be to explore a regression across all variables. This could be useful to find which variables have the most significant impact on box office revenue or a different target variable. Something else that could be interesting to visit, would be to look at this project a few years from now and pull new insights again from a new timeframe and compare the data in a similar manner. This could show how much they've changed across four or five years instead of just two and potentially trying to fix some of the limitations we faced by having access to the current data we pulled in future years. One last idea would be to resurface some of the dropped columns and analyze common directors or cast members.