



TASK

Exploratory Data Analysis on the Movies Data Set

[Visit our website](#)

Introduction

Movies dataset contains $4803 \text{ rows} \times 20 \text{ columns}$, which contain the data of budget, revenue, title and other information on individual movie. It also contains some data in columns with JSON format. The exploratory data analysis (EDA) will focus on the profile and the characteristics for success of the movies.

DATA CLEANING

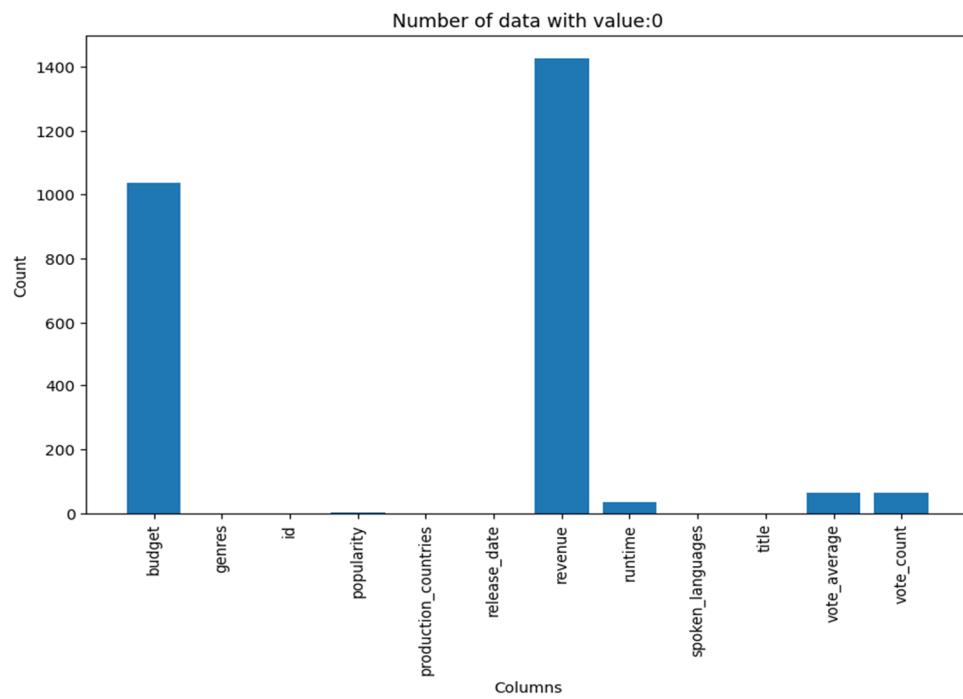
Irrelevant information: Variables: 'homepage', 'keywords', 'original_language', 'original_title', 'overview', 'production_companies', 'status', 'tagline' were identified to be irrelevant to the EDA, they were dropped.

Duplicate rows: None was found in the dataset.

Null data: None was found in the dataset.

Data with JSON format was translated, variables were casted and modified to appropriate data types to be analysed.

MISSING DATA



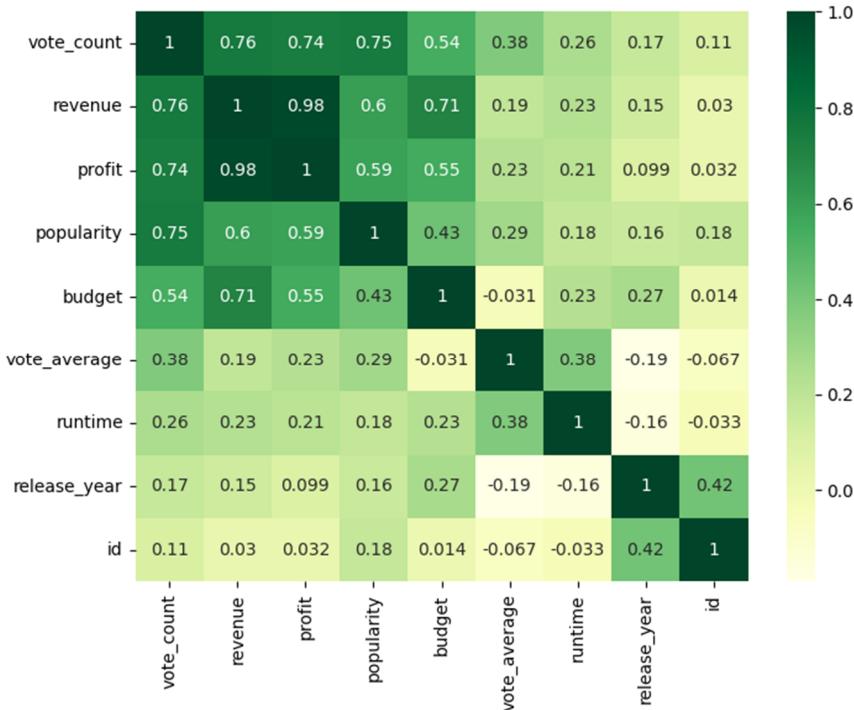
Null data or missing values: 1037 rows with zero budget (21.6%) and 1427 rows with zero revenue (29.7%) were identified, they were dropped as it will have huge impact for the following EDA on profit. Vote average and vote count could be zero, while runtime could be short and close to zero so the data were kept instead of dropping.

Dataset with $3229 \text{ rows} \times 12 \text{ columns}$ resulted after dropping.

DATA STORIES AND VISUALISATIONS

Exploratory Data Analysis

Heatmap (Numerical variables):



Top 5 most expensive and 5 cheapest movies:

To explore the most expensive movies highlights, ROI (return on investment) and profit margin calculated to give better insights. ROI and profit margin are indicators to tell if the movies are profitable:

ROI (Return on Investment) in % = $(\text{Current Value of Investment} - \text{Cost of Investment}) / \text{Cost of Investment}$ = $(\text{profit} / \text{budget}) \times 100$

Profit Margin in % = $((\text{Revenue} - \text{Cost}) / \text{Revenue}) \times 100$

Most Expensive Movies:

	Title	Budget	Revenue	Profit	ROI	Profit Margin	Vote Average	Popularity
Pirates of the Caribbean: On Stranger Tides	380000000	1045713802	665713802	175.187843	63.661185	6.4	135.413856	
Pirates of the Caribbean: At World's End	300000000	961000000	661000000	220.333333	68.782518	6.9	139.082615	
Avengers: Age of Ultron	280000000	1405403694	1125403694	401.929891	80.076899	7.3	134.279229	
Superman Returns	270000000	391081192	121081192	44.844886	30.960628	5.4	57.925623	
John Carter	260000000	284139100	24139100	9.284269	8.495522	6.1	43.926995	

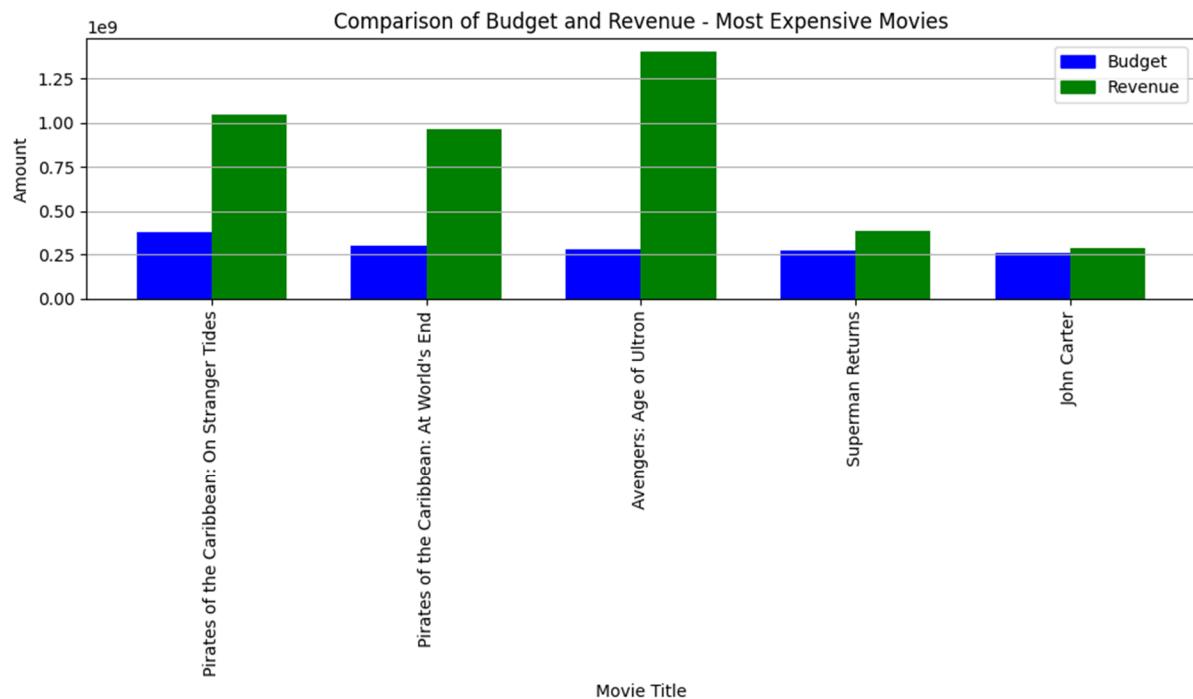
Cheapest Movies:

	Title	Budget	Revenue	Profit	ROI	Profit Margin	Vote Average	Popularity
Modern Times	1	8500000	8499999	8.499999e+08	99.999988	8.1	28.276480	
A Farewell to Arms	4	25	21	5.250000e+02	84.000000	6.2	1.199451	
Split Second	7	5	-2	-2.857143e+01	-40.000000	5.7	4.857028	
Bran Nue Dae	7	7	0	0.000000e+00	0.000000	5.2	0.058456	
The Prophecy	8	16	8	1.000000e+02	50.000000	6.4	11.818333	

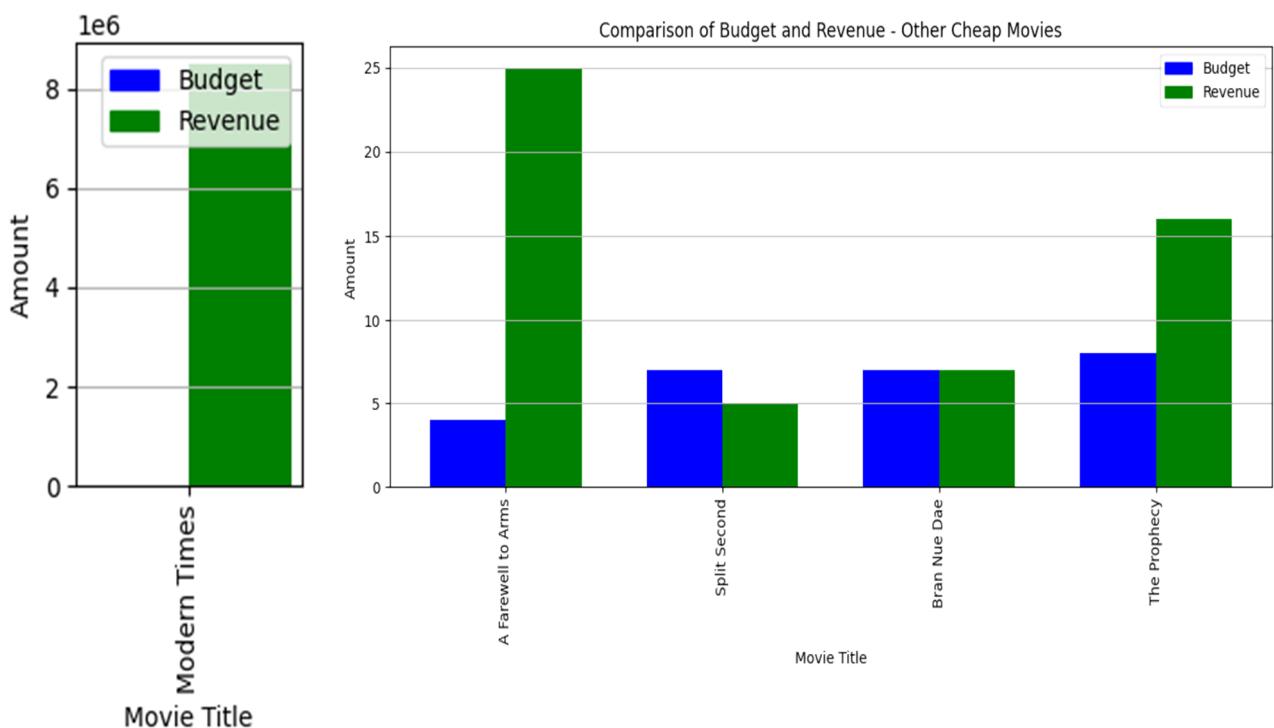
There are huge difference of budget and revenue on the expensive movies and cheapest movies. The budget and revenue of expensive movies reached ($\times 10^9$) while cheapest movies could have just single value for them. They will be illustrated on the different graphs:

1. Budget vs Revenue on the most expensive movies and cheapest movies:

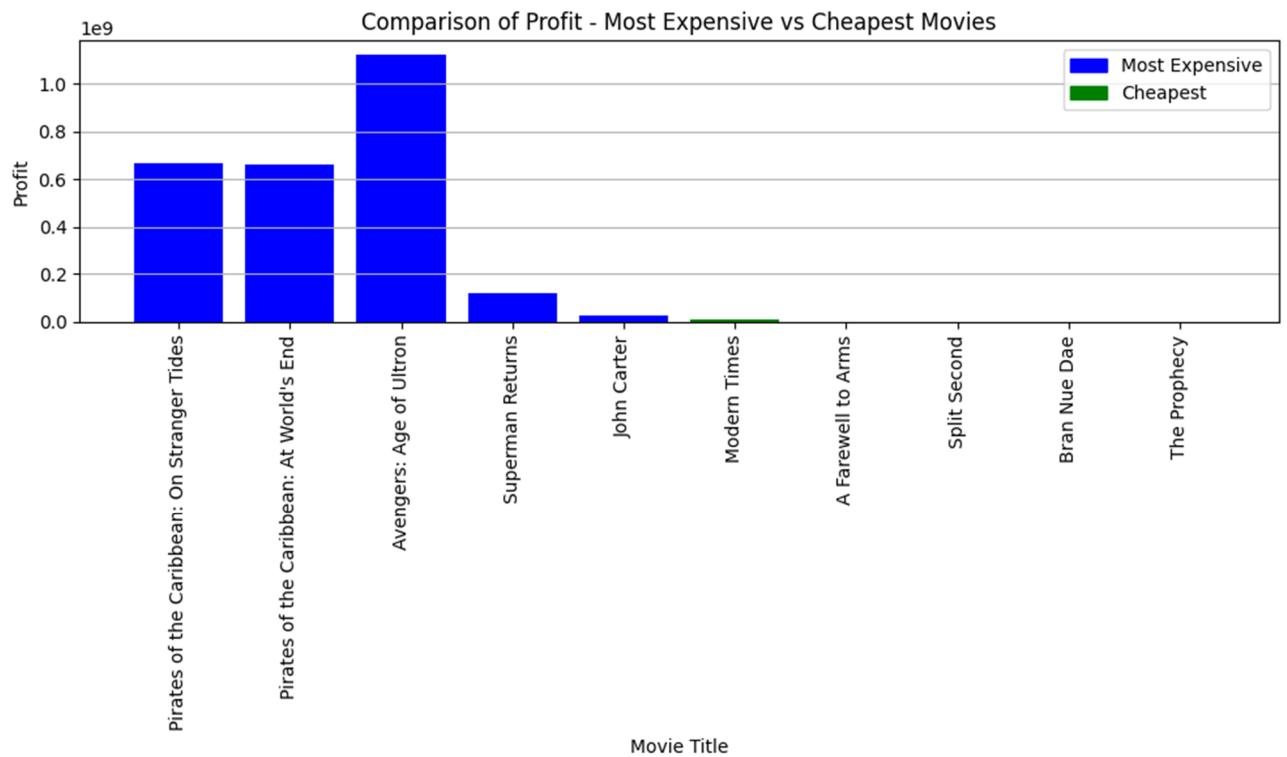
Top 5 expensive movies:



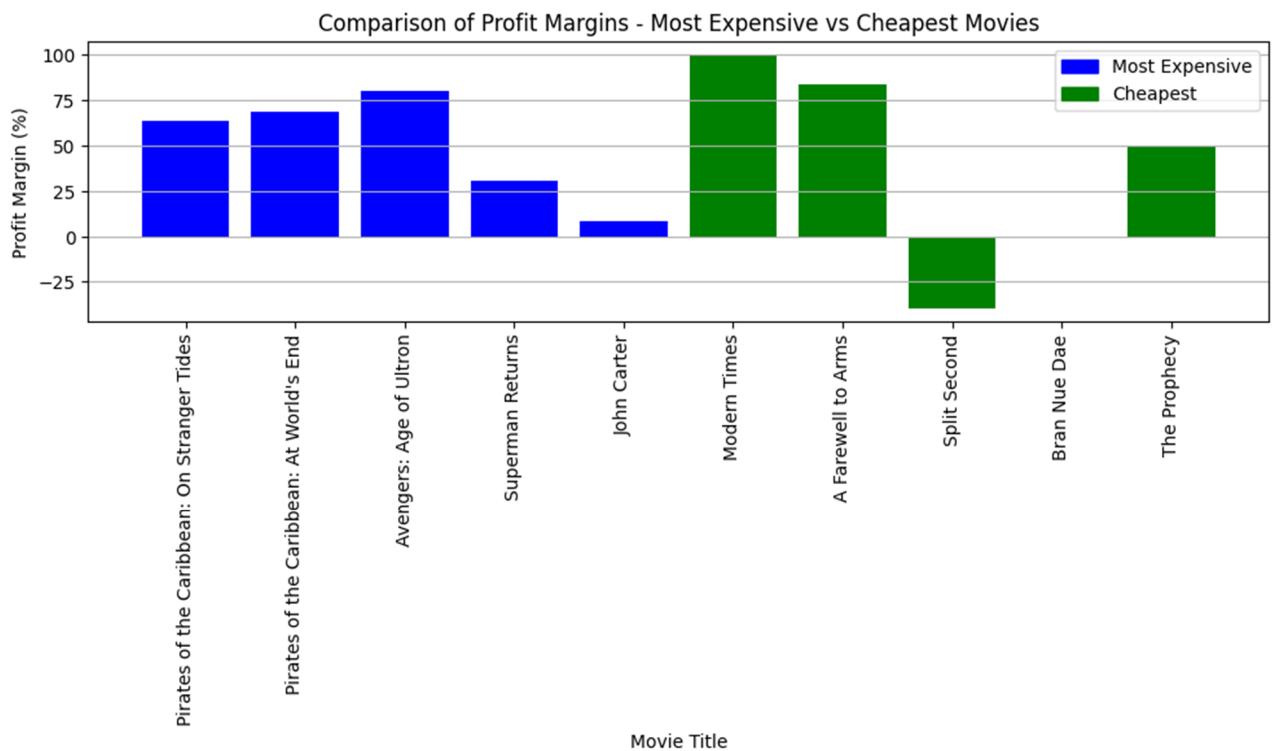
Top 5 cheapest movies:



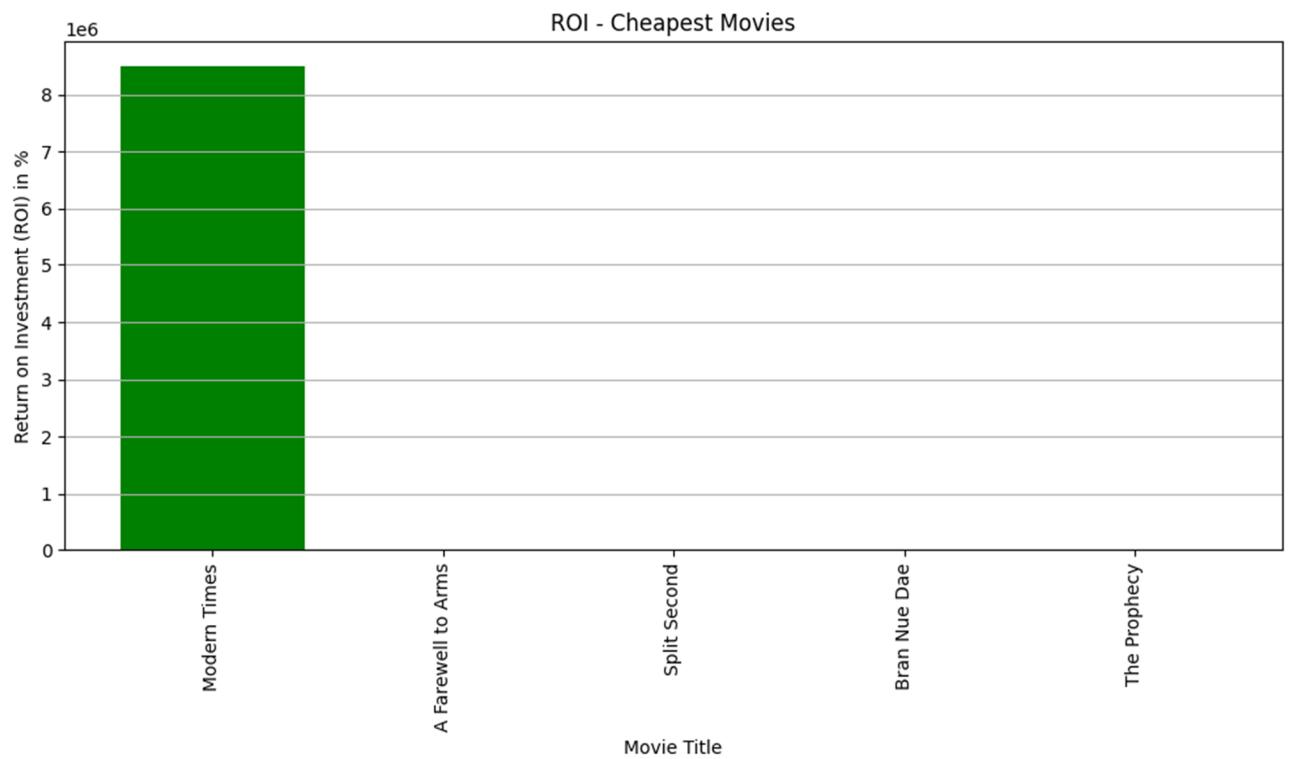
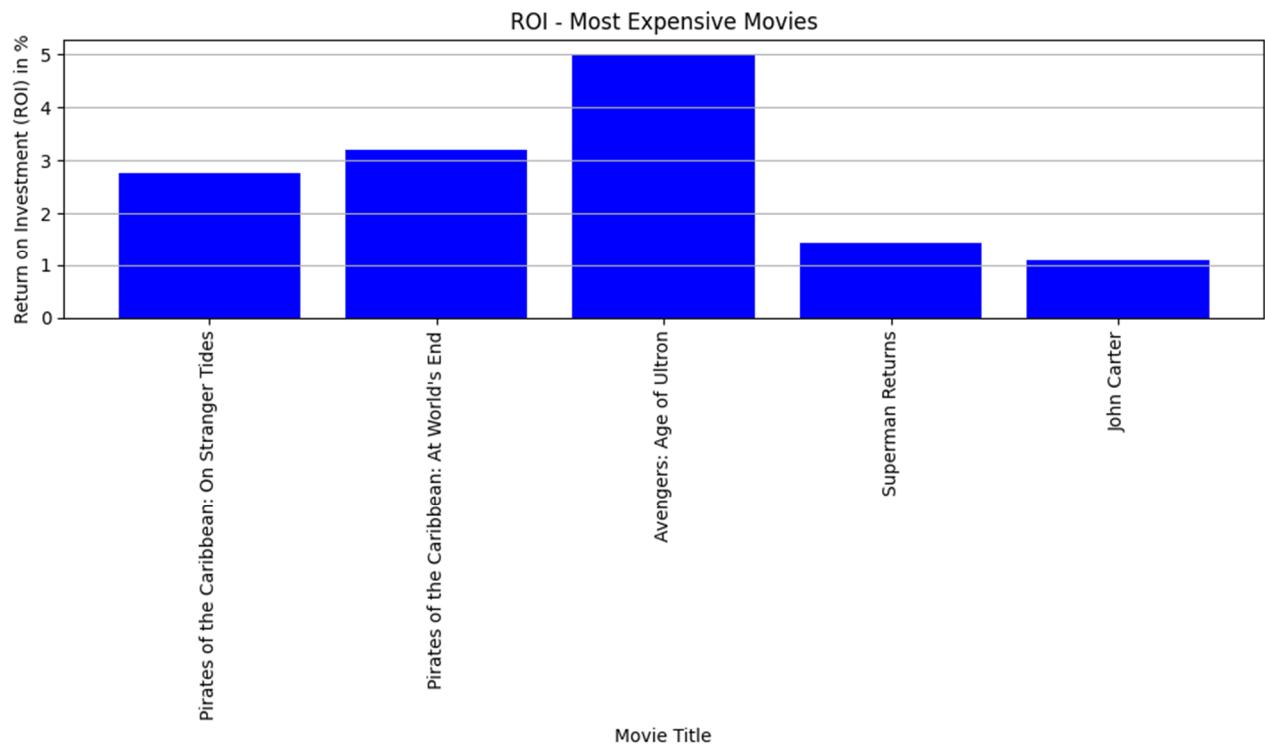
2. Profit on the most expensive movies and cheapest movies:



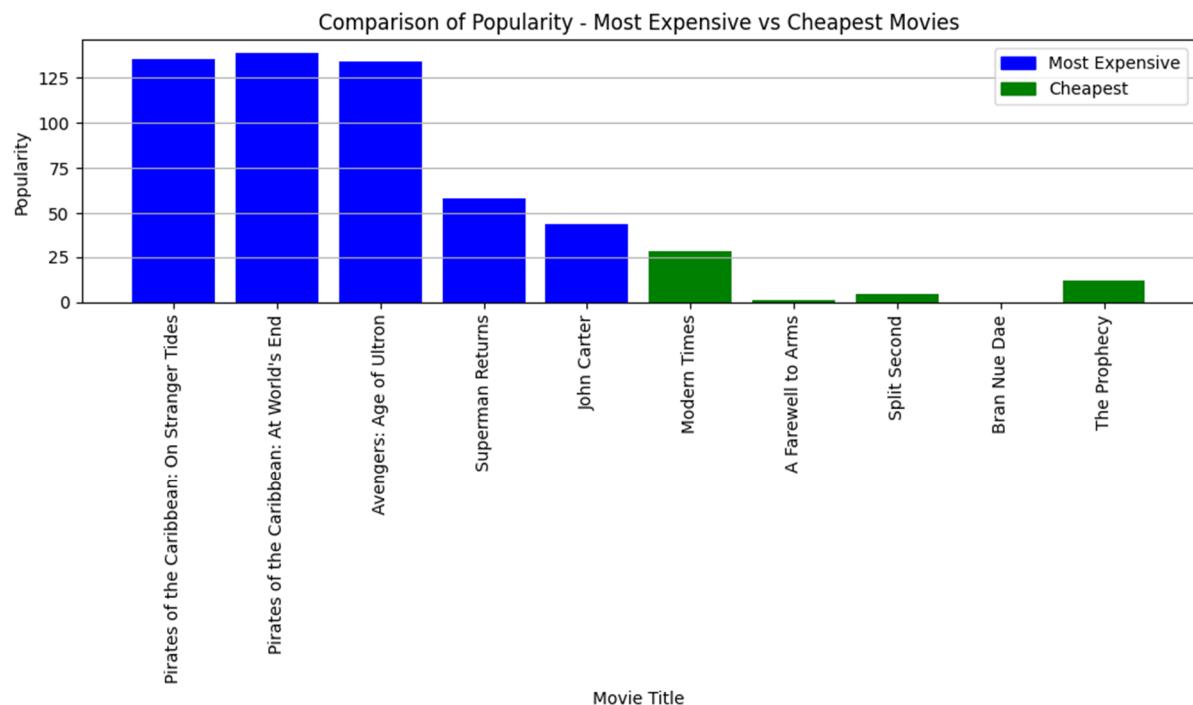
3. Profit margin (%) on the most expensive movies and cheapest movies:



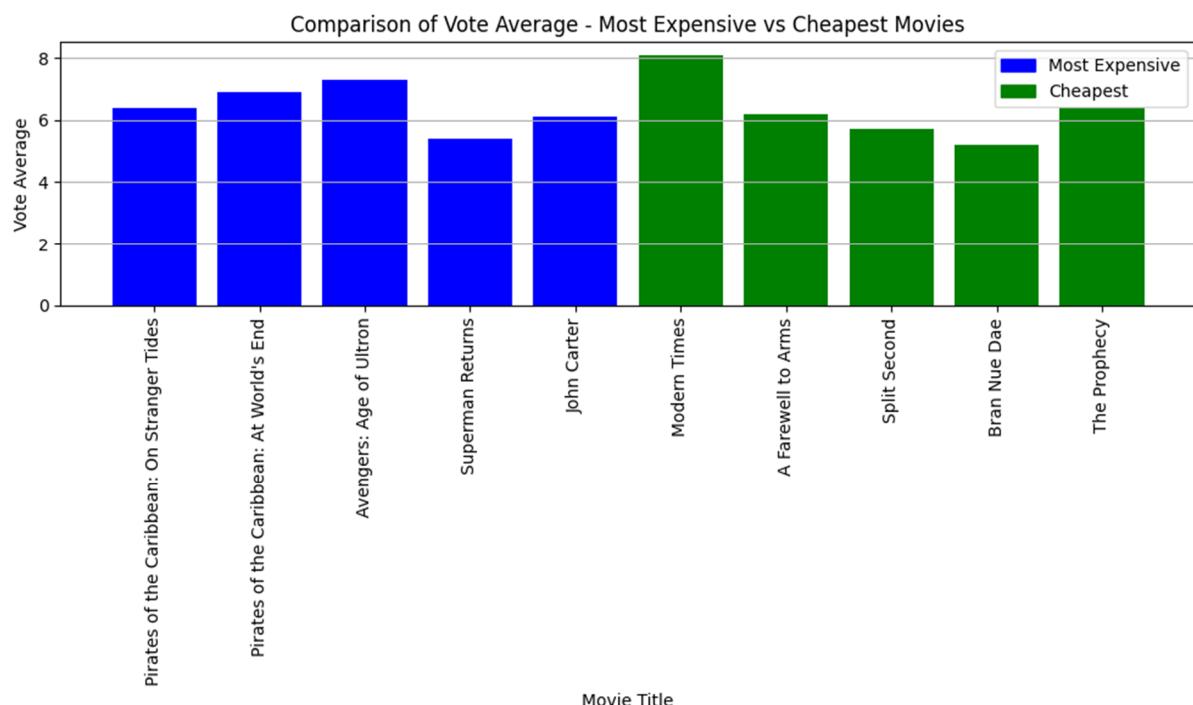
4. ROI (Return on Investment) in % on the most expensive movies and cheapest movies:



5. Popularity on the most expensive movies and cheapest movies:



6. Vote average on the most expensive movies and cheapest movies:



To see if some movies are worth the money spent on them based on their performance and revenue generated.

How do the most expensive and cheapest movies compare? Exploring the most expensive movies highlights if some movies are worth the money spent on them based on their performance and revenue generated.

This is not an easy question to answer because performance of a movie could be defined in different ways based on different opinions and indicators. Here we may look at **1.revenue, 2.profit, 3.profit margin, 4.ROI , 5.popularity and 6.vote average** to decide if they are worth the money (budget).

Based on 1.revenue,2.profit,3.profit margin and 4.ROI:

Expensive movies:

The most expensive movies listed are "Pirates of the Caribbean: On Stranger Tides," "Pirates of the Caribbean: At World's End," "Avengers: Age of Ultron," "Superman Returns," and "John Carter." These films had enormous budgets, ranging from \$260 million to \$380 million.

"Pirates of the Caribbean: On Stranger Tides" and "Pirates of the Caribbean: At World's End" are the two most expensive movies (\$380 million and \$300 million) in the list. While they have high budgets, they also generated significant revenue (\$1,045 million and \$961 million), resulting in substantial profits. Their profit margins are relatively good (64% & 69%), indicating that they were successful ventures.

"Avengers: Age of Ultron" had a high budget (\$280 million) but generated even higher revenue (\$1,405 million), leading to a significant profit (\$1.13 billion). Its profit margin (80%) is quite impressive, suggesting that it was a highly successful movie in terms of financial returns.

Despite the high budgets of expensive movies, the profitability and return on investment (ROI) of these movies vary significantly. The film that generated the highest profit among the most expensive movies was "Avengers: Age of Ultron" with a staggering profit of \$1.13 billion. It achieved an ROI of 401 % (compared with previous two movies :175%, 220%) indicating its financial success.

However, "Superman Returns" had a substantial budget (\$270 million), but its revenue(\$391 million), profit (\$121 million), profit margin (30%), ROI (44.84%) are relatively lower than the previous movies, which suggests that despite its high production cost, it might not be worth the money in terms of revenue. "John Carter" had a considerable budget (\$260 million), but its revenue (\$284 million), profit (\$24 million) and profit margin (8%) are relatively low. This

indicates that the movie struggled to recoup its production costs and generate substantial profits. Again, it may not worth the money.

The most expensive movies vary in their success, with some generating significant profits and others underperforming.

The cheapest movies:

Comparing these expensive movies to the cheapest movies listed, such as "Modern Times," "A Farewell to Arms," "Split Second," "Bran Nue Dae," and "The Prophecy," we can see a stark contrast in terms of budget and profitability. The cheapest movies had significantly lower budgets, ranging from \$1 to \$8.

Interestingly, the film "Modern Times" stands out among the cheapest movies. It had an extremely low budget of only \$1 but generated a substantial profit of \$8.5 million, resulting in an impressive ROI of 849,999,900%, a substantial profit margin of nearly 100%. This highlights the exceptional performance and profitability of this particular movie despite its minimal budget.

Other cheapest movies had mixed outcomes, with some achieving high profit margins despite their minimal budgets, while others resulted in losses or break-even scenarios. I

In summary, exploring the most expensive and cheapest movies reveals a wide range of financial outcomes. While some expensive movies justify their budgets with high profitability and ROI, others may not perform as well. On the other hand, some of the cheapest movies manage to deliver exceptional returns on their minimal investments. These comparisons emphasize the importance of evaluating a movie's financial performance in relation to its budget and revenue generated to determine if it was worth the money spent on its production.

Based other performance (5.Popularity and 6.vote average) :

Among the most expensive movies, "Avengers: Age of Ultron" stands out with a high popularity score of 134 and a relatively high vote average of 7.3. This indicates that the film was well-received by the audience and had a significant level of popularity.

The "Pirates of the Caribbean" movies also performed well in terms of popularity, with both "On Stranger Tides" and "At World's End" having scores above 130. However, their vote

average ratings are slightly lower compared to "Avengers: Age of Ultron," ranging from 6.4 to 6.9.

In contrast, "Superman Returns" and "John Carter" had lower popularity scores, indicating a relatively lower level of audience interest. Their vote average ratings were also moderate, ranging from 5.4 to 6.1.

Moving on to the cheapest movies, "Modern Times" stands out with a high popularity score of 28, suggesting that it has gained recognition and interest despite being a low-budget film. Additionally, it received a high vote average rating of 8.1, indicating positive audience reception.

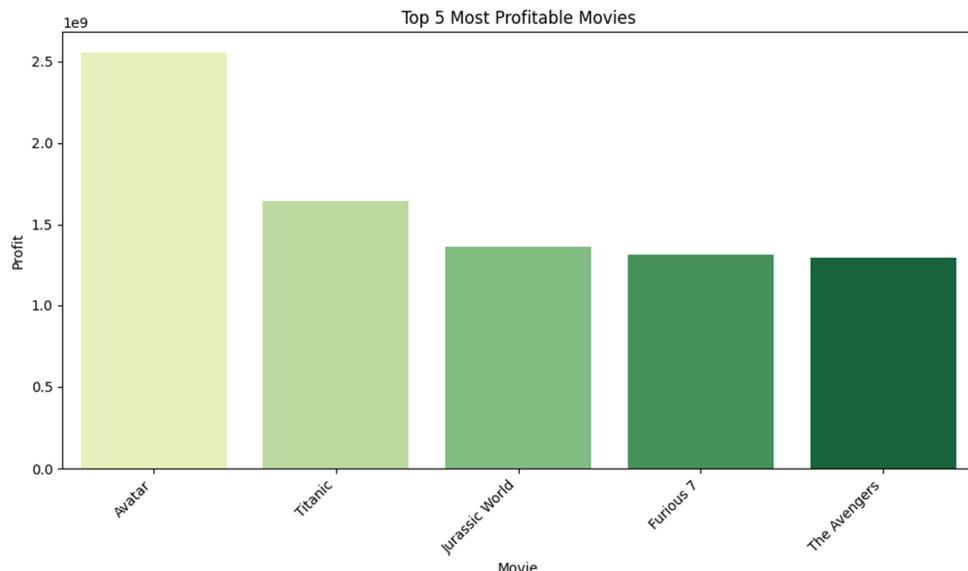
The remaining cheapest movies, including "A Farewell to Arms," "Split Second," "Bran Nue Dae," and "The Prophecy," have relatively lower popularity scores and varied vote average ratings, ranging from 1.2 to 6.4. These films may have had a more limited reach or diverse audience reactions.

Overall, when considering popularity and vote average, the most expensive movies generally had higher popularity scores, suggesting a broader audience appeal. Meanwhile, the cheapest movies had a mix of lower and moderate popularity and vote average scores, indicating a more varied performance in terms of audience reception.

Balancing all contributions:

In conclusion, "Avengers: Age of Ultron" and "Modern Times" definitely are top two movies stand out from the expensive movies and cheapest movies, worth its budget for production based on their overall performance we have discussed above. "Pirates of the Caribbean: On Stranger Tides" and "Pirates of the Caribbean: At World's End" have also remarkable performance with its budget.

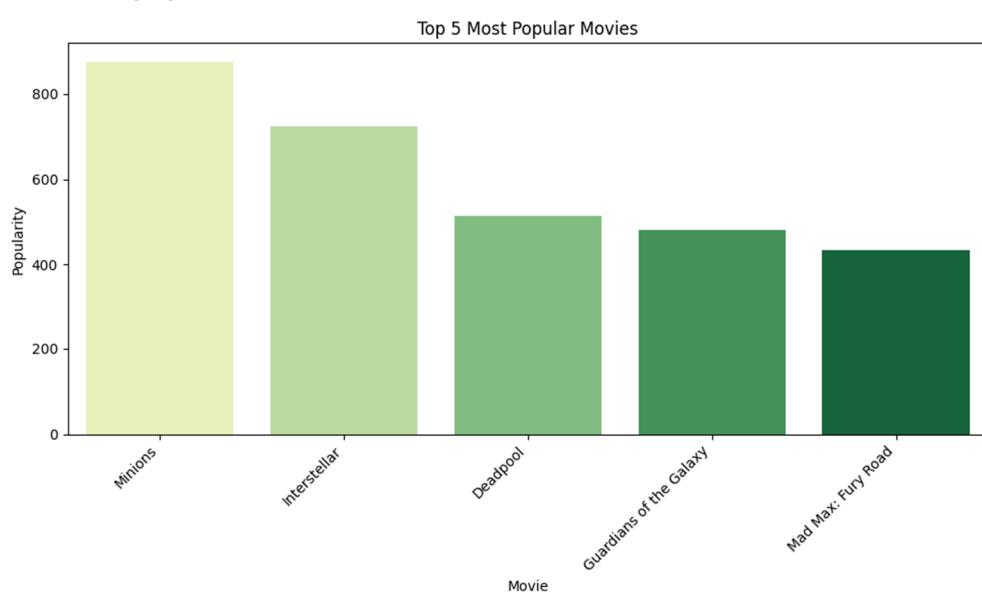
5 most profitable movies:



The movies with minimum and maximum profits:

	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count
0	237000000	['Action', 'Adventure', 'Fantasy', 'Science Fiction']	19995	150.437577	['United States of America', 'United Kingdom']	2009-12-10 00:00:00	2787965087	162.000000	['English', 'Español']	Avatar	7.200000	11800
13	255000000	['Action', 'Adventure', 'Fantasy', 'Western']	57201	49.046956	['United States of America']	2013-07-03 00:00:00	89289910	149.000000	['English']	The Lone Ranger	5.900000	2311
<hr/>												
release_year	profit	release_decade										
2009	2550965087	2000										
2013	-165710090	2010										

The most popular movies:

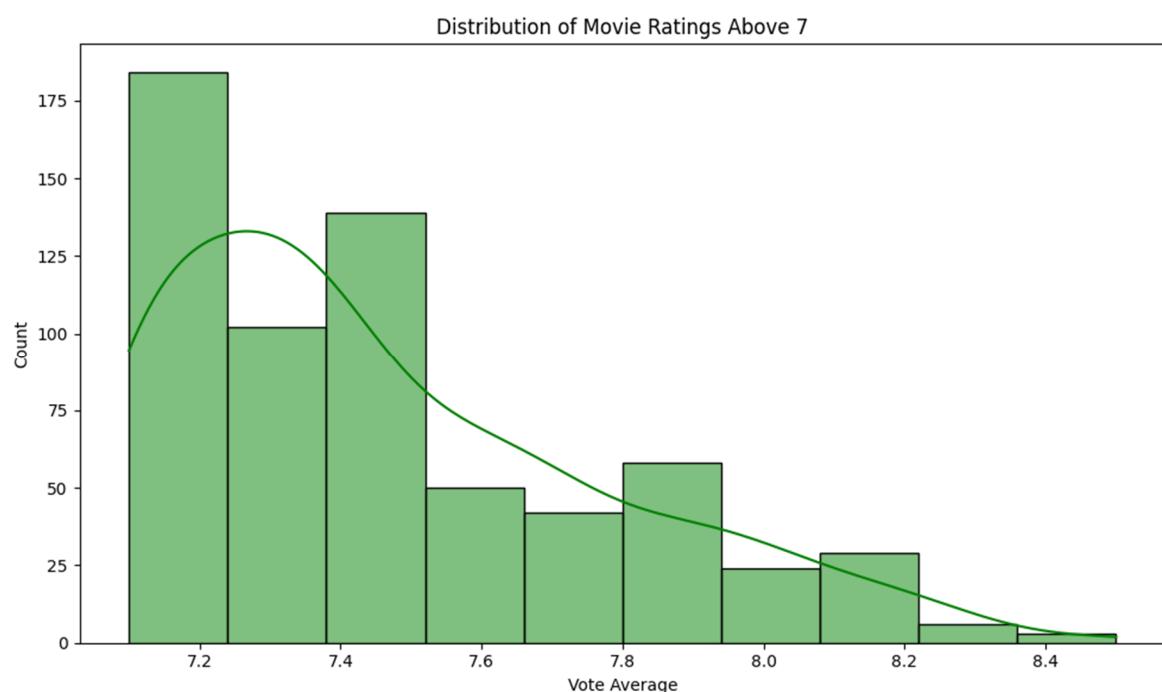


Movies which are rated above 7:

There are 637 movies with rating above 7, the top 5 are as follow:

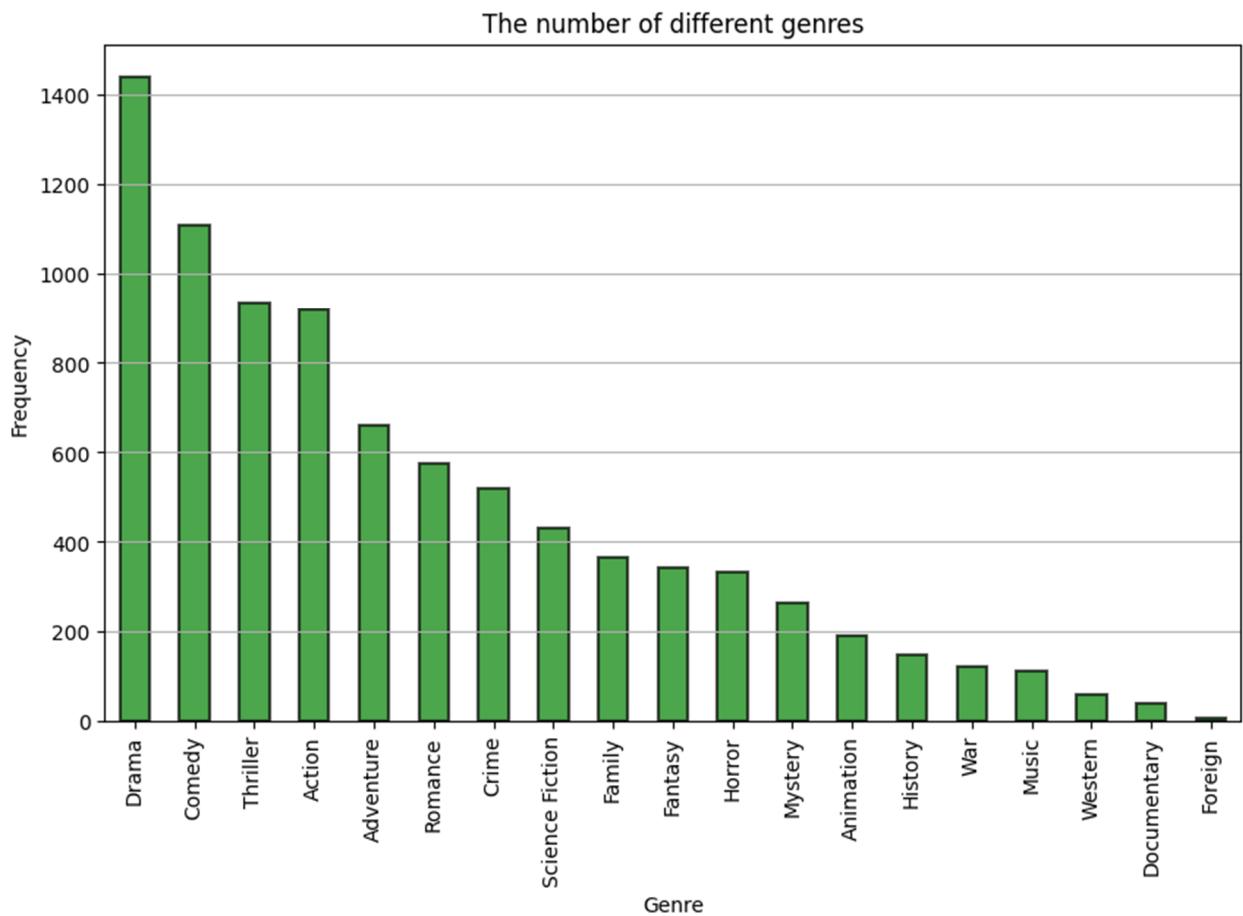
	title	vote_average
1881	The Shawshank Redemption	8.5
2970	There Goes My Baby	8.5
3337	The Godfather	8.4
1818	Schindler's List	8.3
3232	Pulp Fiction	8.3

The distribution of the rating (>7):



Most successful genres:

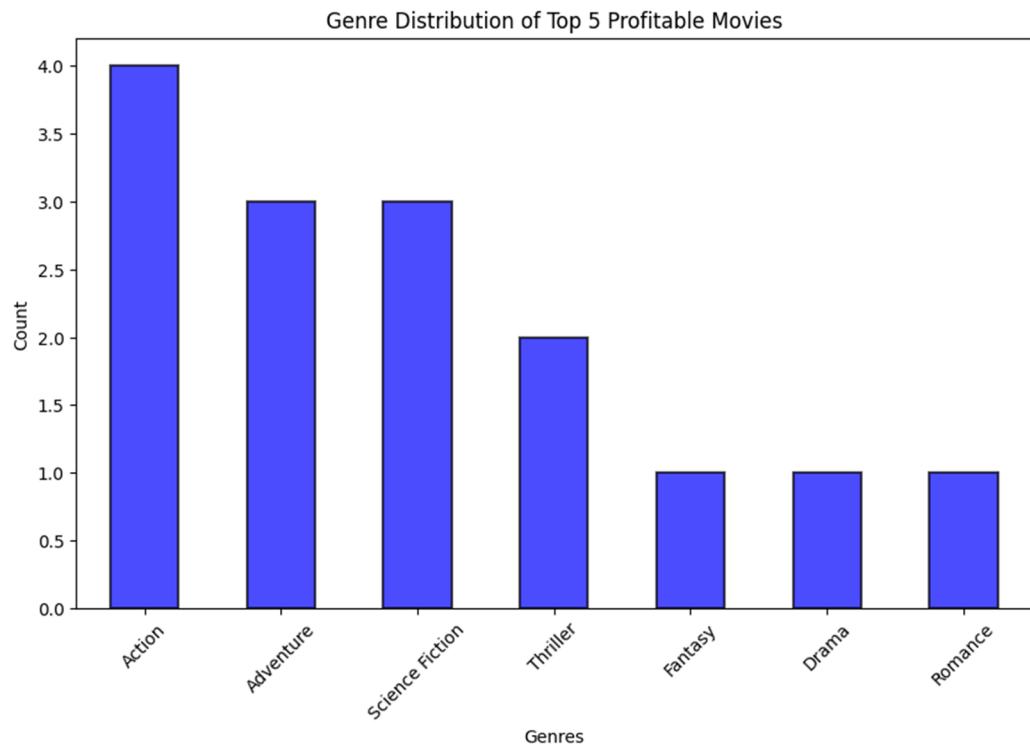
Number of different genres in the dataset:



The top 5 highest number of genres in the dataset: Drama 1441, Comedy 1110, Thriller 935, Action 918, Adventure 661.

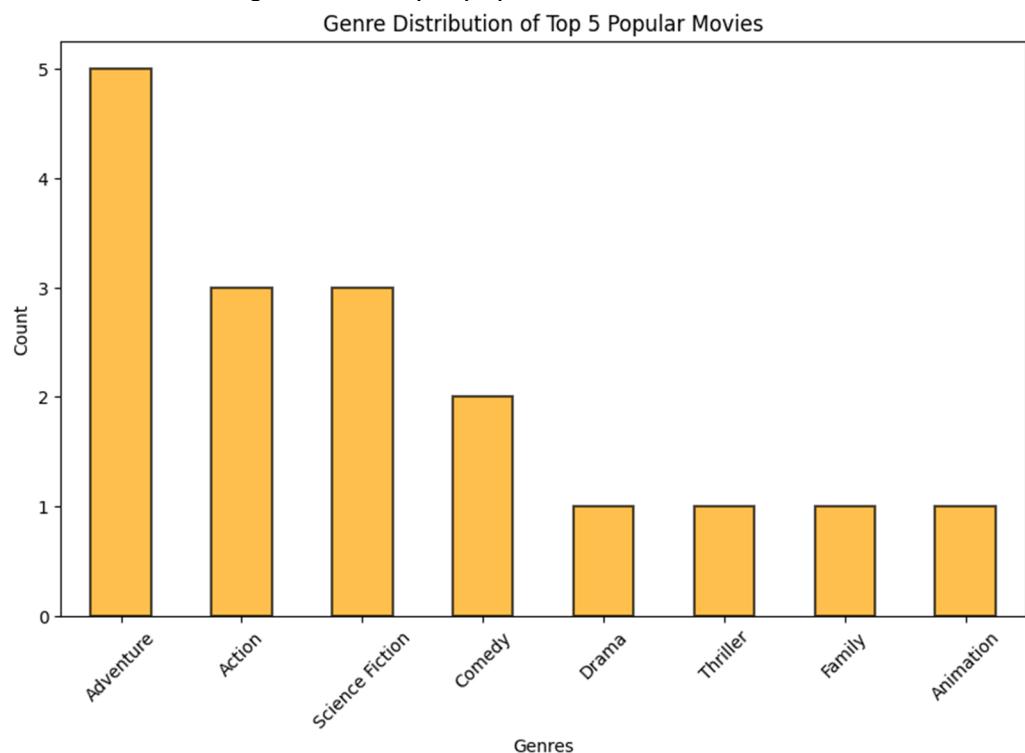
The genres of most profitable, popular and highly rated movies are identified as followings (1-3):

1. The number of genres for top 5 profitable movies:



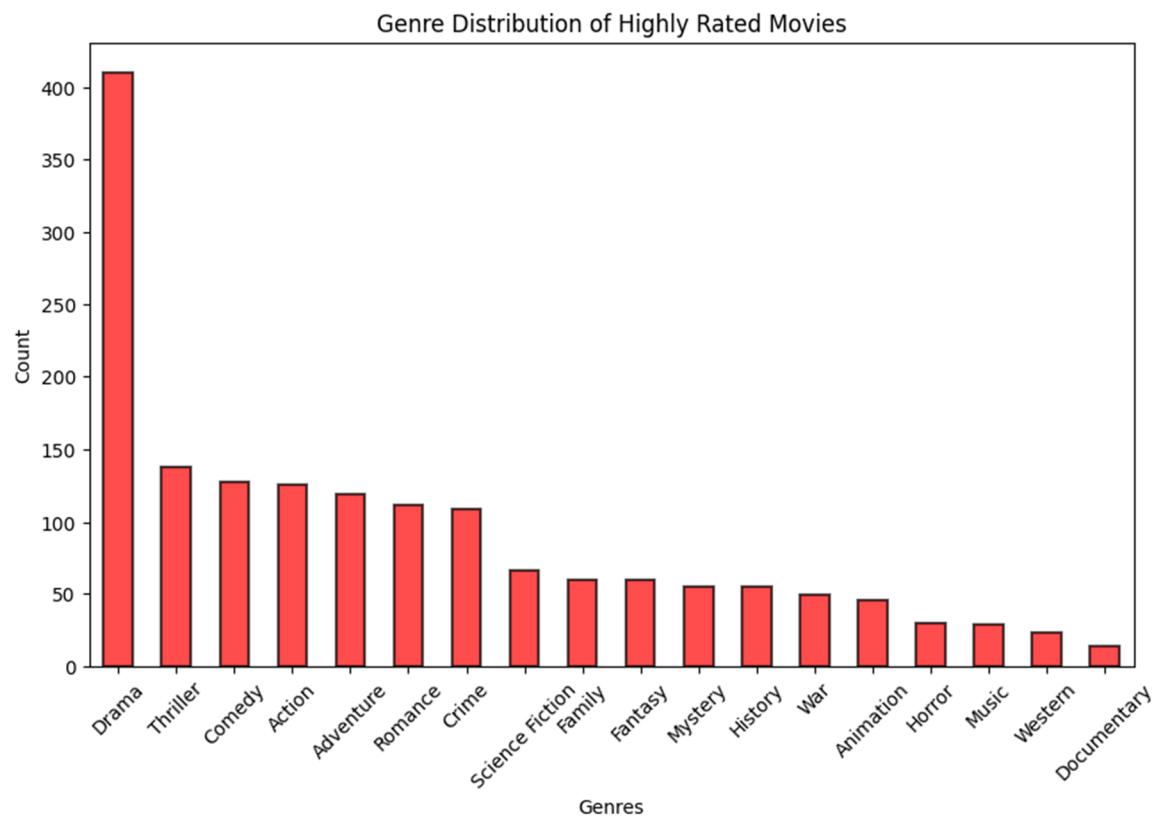
Top 5 profitable movies: Action, Adventure, Science Friction, Thriller, Fantasy.

2. The number of genres for top 5 popular movies:



Top 5 popular movies: Adventure, Action, Science Friction, Comedy, Drama

3. The number of genres for highly rated (>7) movies:

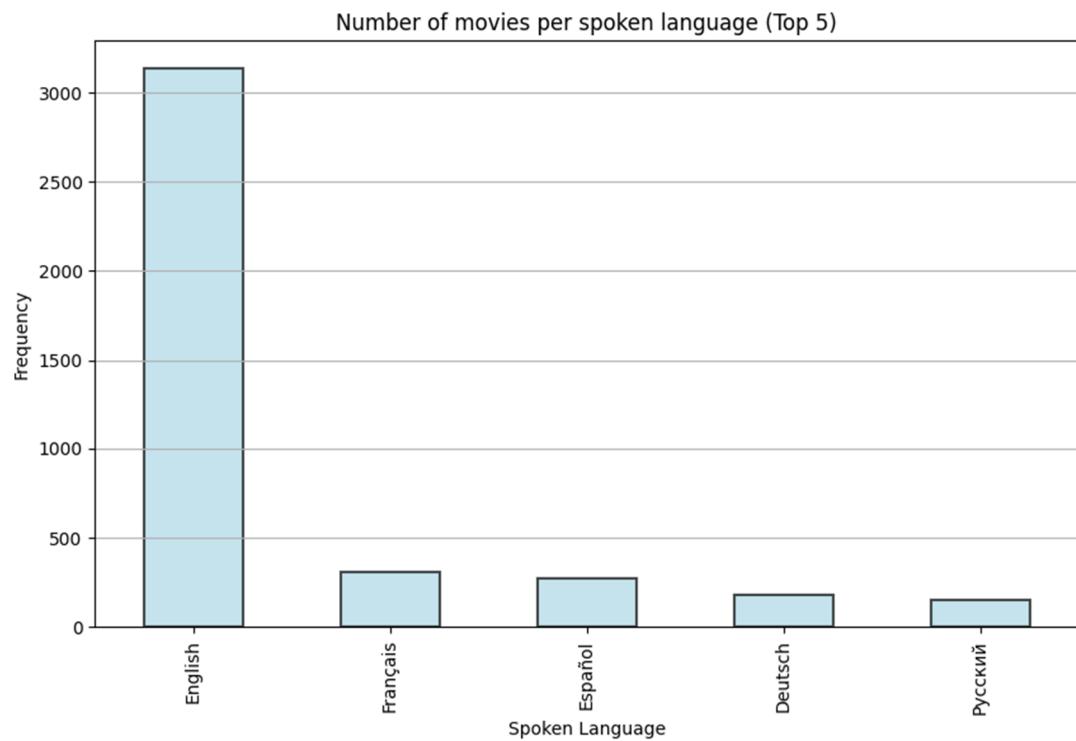


Top 5 highly rated (>7) movies: Drama, Thriller, Comedy, Action, Adventure.

When we look at the 5 highest number of genres in each bar chart: 1., 2., and 3. 6 out of 7 of those genres are included in the top 5 highest number of genres in the dataset: Drama 1441, Comedy 1110, Thriller 935, Action 918, Adventure 661. So, obviously movie production companies understand the market's demand. They clearly know which genres of movies is more profitable and popular in the market, so the production of movies with same kind of genres were made.

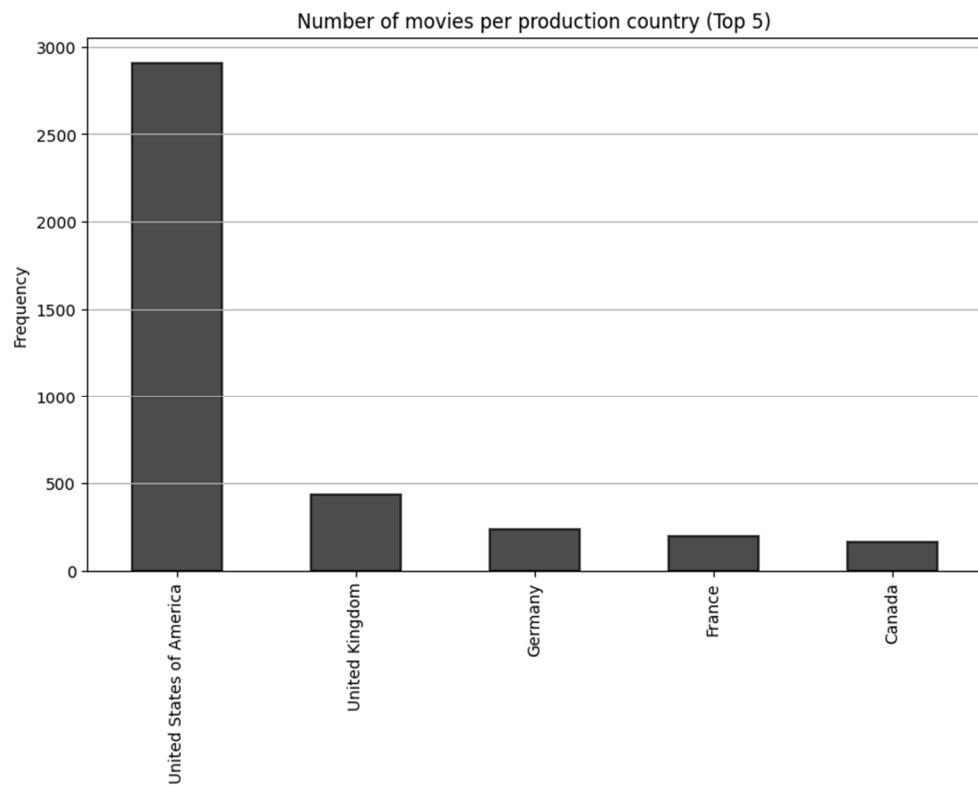
Other findings:

1. The top 5 languages of the movies:



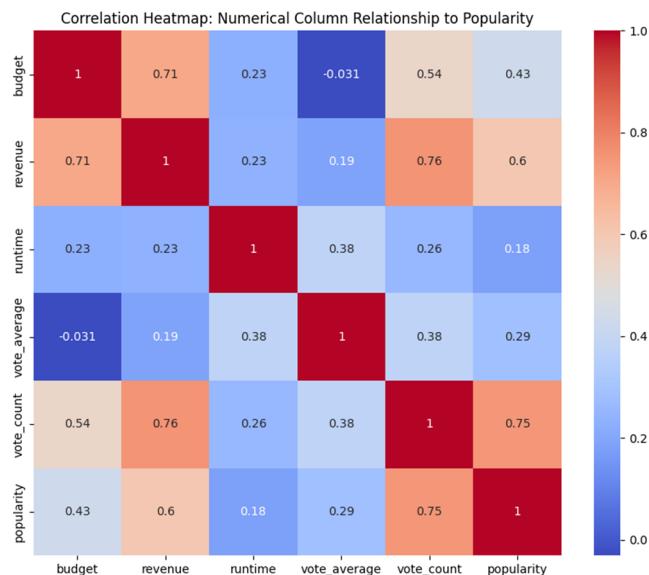
English is the dominant language of the movies. It is not surprising as English is the international language, and we can also see the next chart which country is dominant on movie production to understand the reason.

2. The top 5 countries of movie production:



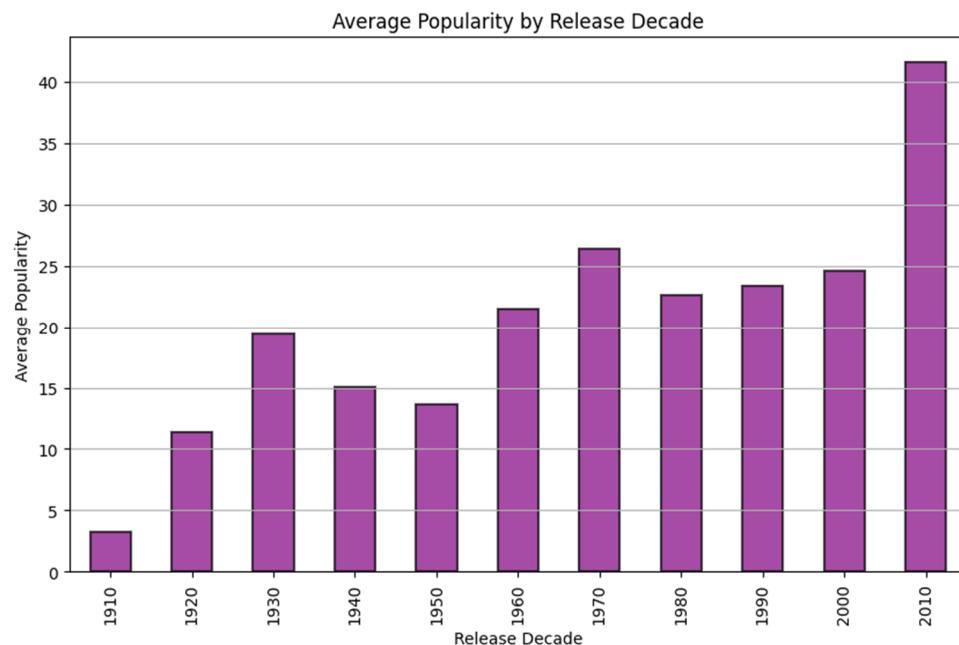
U.S is the dominant country for movie production. The production from U.S is far more (5 times or more) than U.K and other countries. It also makes sense when we look at the result that English is the dominant language of movies as the official language of U.S. and UK is English.

3. Heatmap for variables related to popularity:



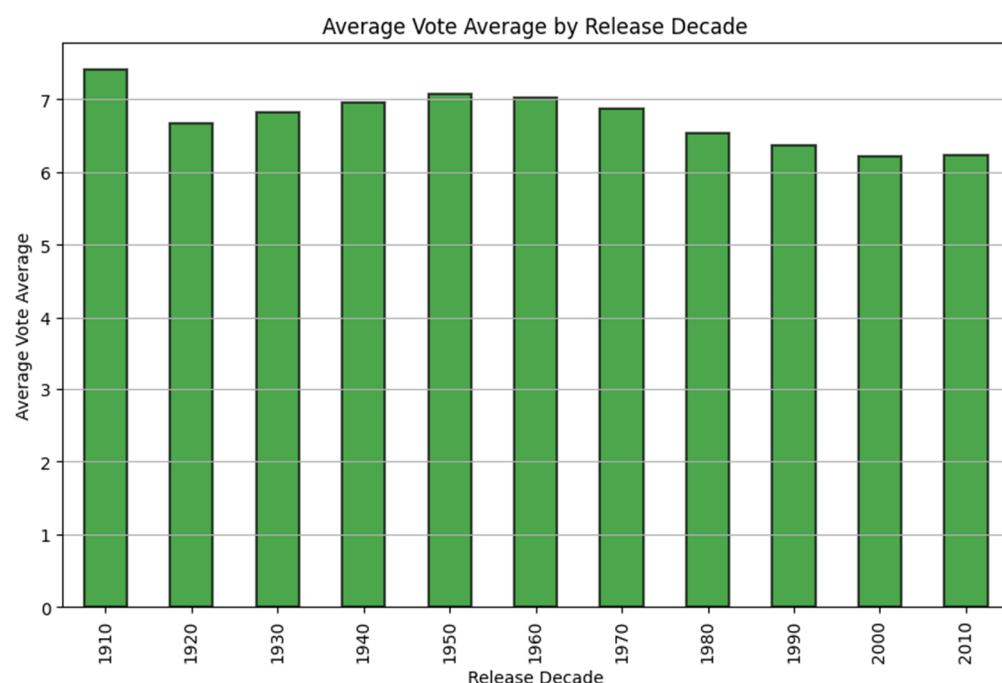
We can see that people like or are more willing to watch the movies with high budget (positive correlation with popularity).

4. Popularity with years:



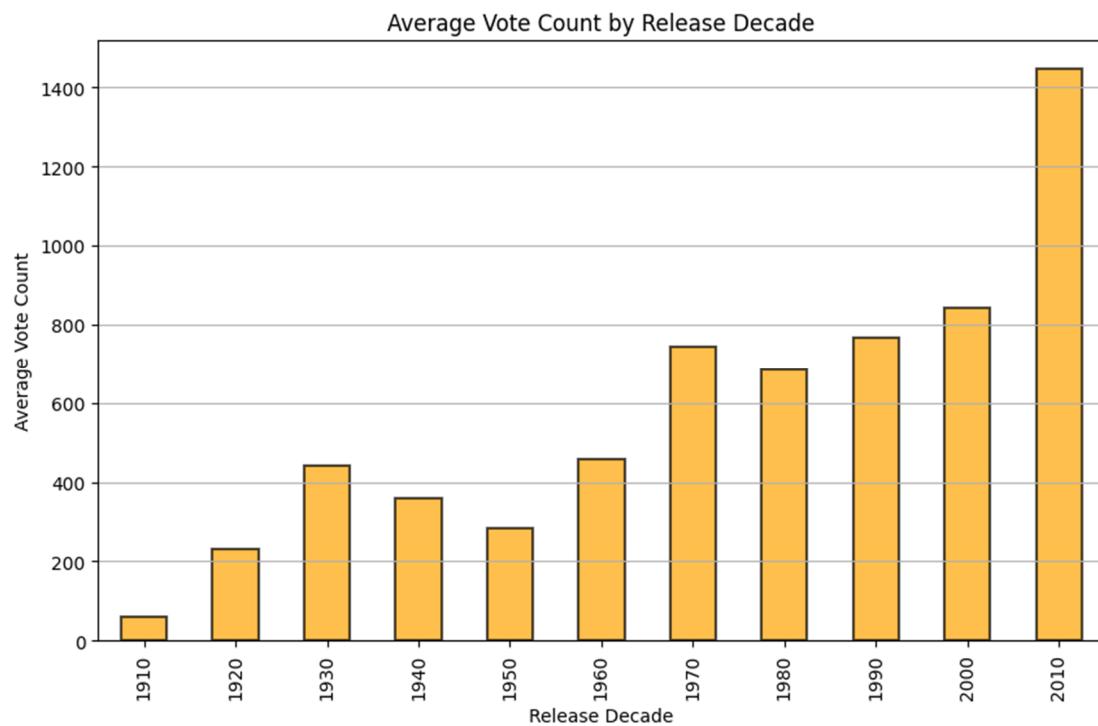
The is an increasing trend of the number of people (popularity) to watch movies compared with the past.

5. Rating with years:



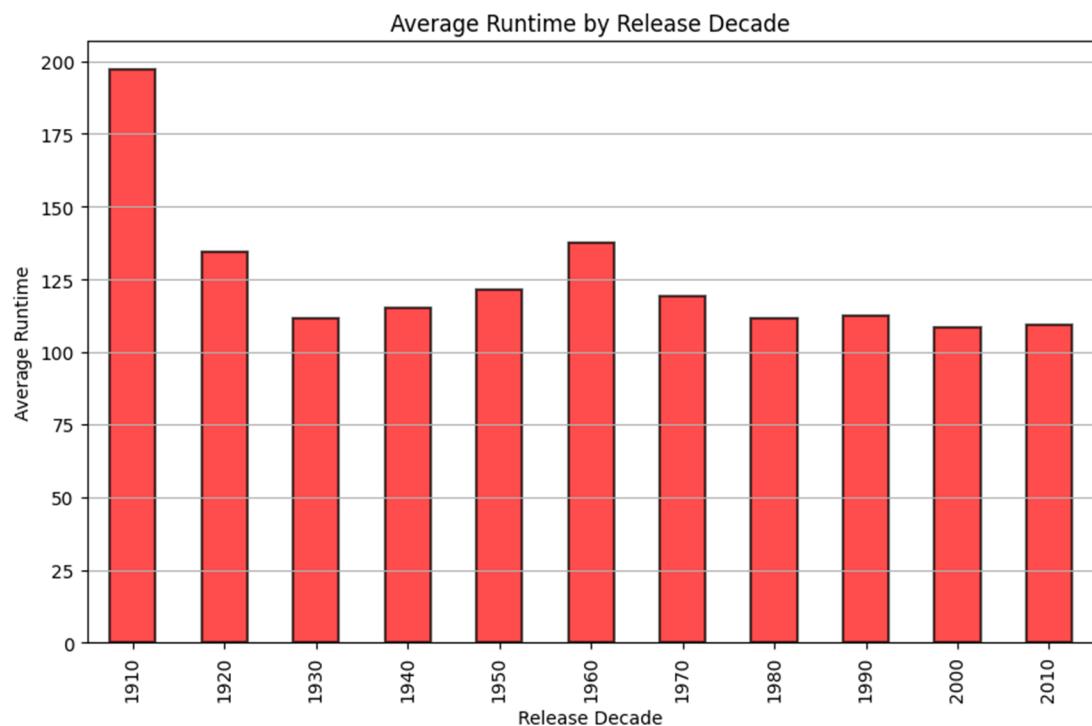
The average vote has been steady through the years from 1910 to 2010 (ranging from 6-7).

6. Vote count with years:



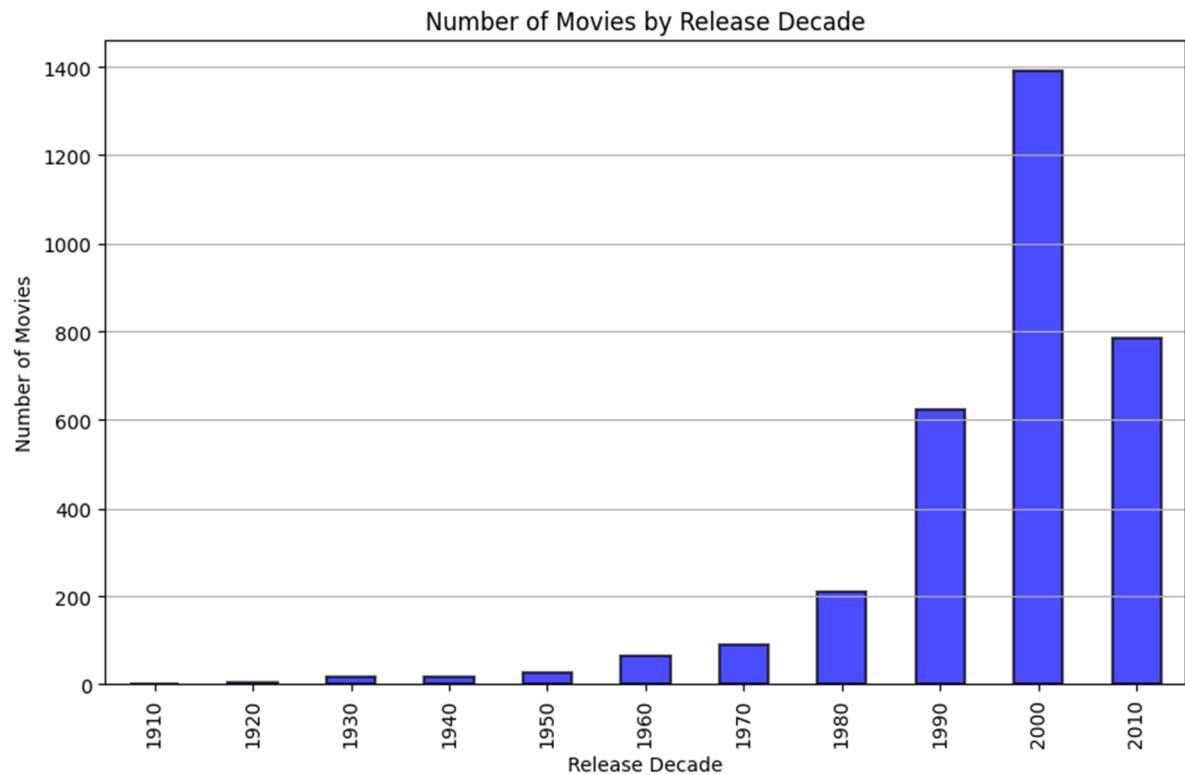
More people are willing to vote or rate the movies compared with the past.

7. Average runtime with years:



The average runtime has been kept around 2 hours.

8. Numbers of movie productions with years:



There is a sudden increase of movie production from 1990-2000, more investigation can be done to explore the reasons behind.

THIS REPORT WAS WRITTEN BY : Chu, Alex Wai Leung

