MOVIE RECOMMENDER SYSTEMS

Final Report

December 2023

TABLE OF CONTENTS

TABLE OF FIGURES

1. OVERVIEW

1.1. The story of film

This section objective is narrating the history, trivia and facts behind the world of cinema through the lens of data. Extensive Exploration Data Analysis is performed on Movie Metadata about Movie Revenues, Casts, Crews, Budget, etc. through the years. Two predictive models are built to predict movie revenues and movie success. Through these models, what features have the most significant impact in determining revenue and success could be discovered later.

1.2. Movie recommender systems

This part is concentrated on building multiple kinds of recommendation engines, named the Simple Generic Recommenders, the Content Based Filter and the User Based Collaborative Filter. The performance of the systems is evaluated in both a qualitative and quantitative manner.

2. THE CLIENT

The first section of the project does not have a definitive client. But some of the analysis performed in this part could be used in the Movie Making Business (Streaming Providers, Producers, etc). The Movie Succeed and Revenue Prediction Models can give valuable insights into the features that actually determine the end class and value respectively.

The Movie Recommender System is useful to any business that makes money via recommendations. This includes Amazon, Netflix, Hotstar, etc. Giving good recommendations directly entails one or many of the following:

1. Customers who buy a particular product or service leading to increased revenue or sales.
2. Customers who use the platform more frequently due to the quality and relevance of content shown to them.
3. Better user experience. Customers spend less time on searching and more time on watching. The pain of discovery is eliminated.

3. THE DATA

The data used in this project has been obtained from 2 sources: MovieLens and The Movie Database (TMDB).

MovieLens has a publicly available full dataset containing approximately 33,000,000 ratings and 2,000,000 tag applications applied to 86,000 movies by 330,975 users between January 09, 1995 and July 20, 2023. Includes tag genome data with 14 million relevance scores across 1,100 tags. This dataset was generated on July 20, 2023. A small subset of the dataset, containing 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users between March 29, 1996 and September 24, 2018. The subset was generated on September 26, 2018. [1]

One of the files contains the TMDB ID of every movie listed in the MovieLens dataset. Using this ID, the metadata, credits and keywords of all 86,000 movies were obtained by running a script that requested and parsed data from TMDB Open API. The data collected was initially in the JSON format but was converted into CSV files using Python's Pandas Library. [2]

The following files were used in the project from both MovieLens subset database and TMDB:

1. **movies_metadata.csv:** The file containing metadata collected from TMDB for over 86,000 movies. Data includes budget, revenue, date released, genres, etc.
2. **credits.csv:** Complete information on credits for a particular movie. Data includes Director, Producer, Actors, Characters, etc.
3. **keywords.csv:** Contains plot keywords associated with a movie.
4. **links_small.csv:** Contains the list of movies.
5. **ratings_small.csv:** Contains 100,000 ratings on 9,000 movies from 600 users. The main dataset used for building the Collaborative Filter.

4. DATA COLLECTION

The MovieLens full and subset dataset is publicly accessible at the GroupLens website. The dataset includes in the file genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv and tags.csv. More details about the contents and use of all these files are in details in README.txt file.

TMDB website suggests signing up for an API Key. This can allow individual access to data at 3 endpoints. Each endpoint gives details about the movie, its casts and crews information and plot keywords. Three separate scrapers were written to hit each endpoint and collect this data for

all 86,000 movies. Since TMDB has a restriction in the 50 requests per second range, this task took a day to execute.

All the data collected was in the form of JSON which demanded more processing

5. DATA WRANGLING

5.1. Overview

This section describes the various data cleaning and data wrangling methods applied on the Movie datasets to make it more suitable for further analysis. The following sections are divided based on the procedures followed.

5.2. Conversion to csv files

The data obtained from scraping was in the form of stringified JSON. This had to be converted into CSV Files to enable easier parsing and subsequent upload to public platforms such as Kaggle.

5.3. Removing unnecessary features

Some features such as the Backdrop Path, Adult and IMDB ID were unnecessary attributes and were dropped to reduce the dimensions of the dataset.

5.4. Cleaning

The dataset had a lot of features which had 0s for values it did not possess. These values were converted to NaN. Some features were still in the form of a Stringified JSON Object. They were converted into Python Dictionaries using Python's json library. These were further reduced into lists since we did not have a need for ID, timestamp and other attributes.

The dataframe was exploded wherever the analysis demanded it (for instance, genres and production countries).

Finally, most of the features were converted into a Python basic type (integer, string, float) by removing all the unclean values. The date string was converted into a Pandas Datetime and from it, the month, year and day of release of every movie was extracted.

6. EXPLORATORY DATA VISUALIZATION AND ANALYSIS

In this section, the various insights produced through descriptive statistics and data visualization is presented.

6.1. Production countries

The Full MovieLens Dataset consists of movies that are overwhelmingly in the English language (more than 53,800). However, these movies may have shot in various locations around the world. It would be interesting to see which countries serve as the most popular destinations for shooting movies by filmmakers, especially those in the United States of America and the United Kingdom.

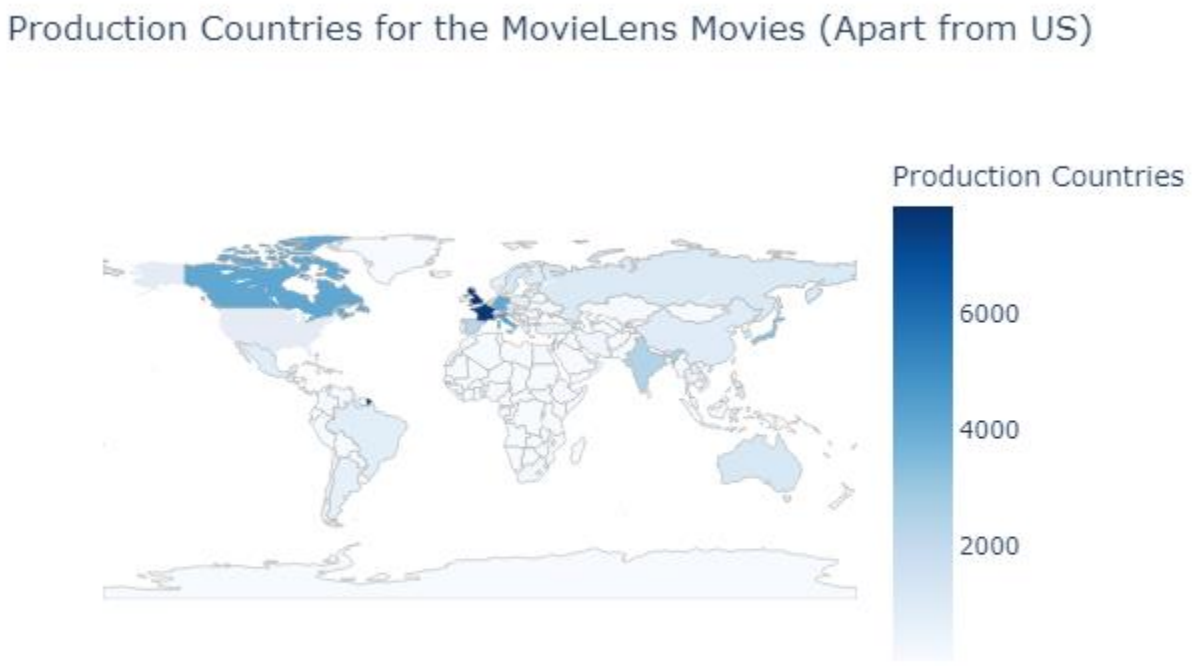

Figure 6.1. Production Countries for the MovieLens Movies (Apart from US).

Unsurprisingly, the United States is the most popular destination of production for movies given that the dataset largely consists of English movies.

Europe is also an extremely popular location with the UK, France, Germany and Italy in the top 5.

Japan and India are the most popular Asian countries when it comes to movie production.

6.2. Franchise movies

The Star Wars Franchise is the most successful movie franchise raking in more than 8.785 billion dollars from 9 movies.

The James Bond Movies come in a separately second with a 7.816 billion dollars but the franchise has significantly more movies compared to the others which is 26 and therefore, a much smaller average gross.

In contrast, the The Avengers has only 4 movies but it grosses 7.776 billion dollars then it gets extremely high average gross.

6.3. Production companies

Warner Bros is the highest earning production company of all time earning a staggering 78.3 billion dollars from close to 700 movies. Universal Pictures and 20th Century Fox are the second and the third highest earning companies with 77 billion dollars and 59 billion dollars in revenue respectively.

Marvel Studios has produced the most successful movies, on average. This is not surprising considering the amazing array of movies that it has produced in the last few decades: The Avengers, Iron man 3, Captain America: Winter Soldier, Avengers: Infinity War, Avengers: Endgame etc.

Pixar with an average gross of 551 million dollars comes in second with movies such as Up, Finding Nemo, Inside Out, Wall-E, Ratatouille, the Toy Story Franchise, Cars Franchise, etc. under its banner.

6.4. Movie title wordcloud

Figure 6.2. Movie title wordcloud

The word Love is the most commonly used word in movie titles. Girl, Man and Life are also among the most commonly occuring words. This encapsulates the idea of the ubiquitious presence of romance in movies pretty well.

6.5. Original languages

There are 122 languages represented in the dataset. As expected, English language films form the overwhelmingly majority. French and Italian movies come at a very distant second and third respectively.

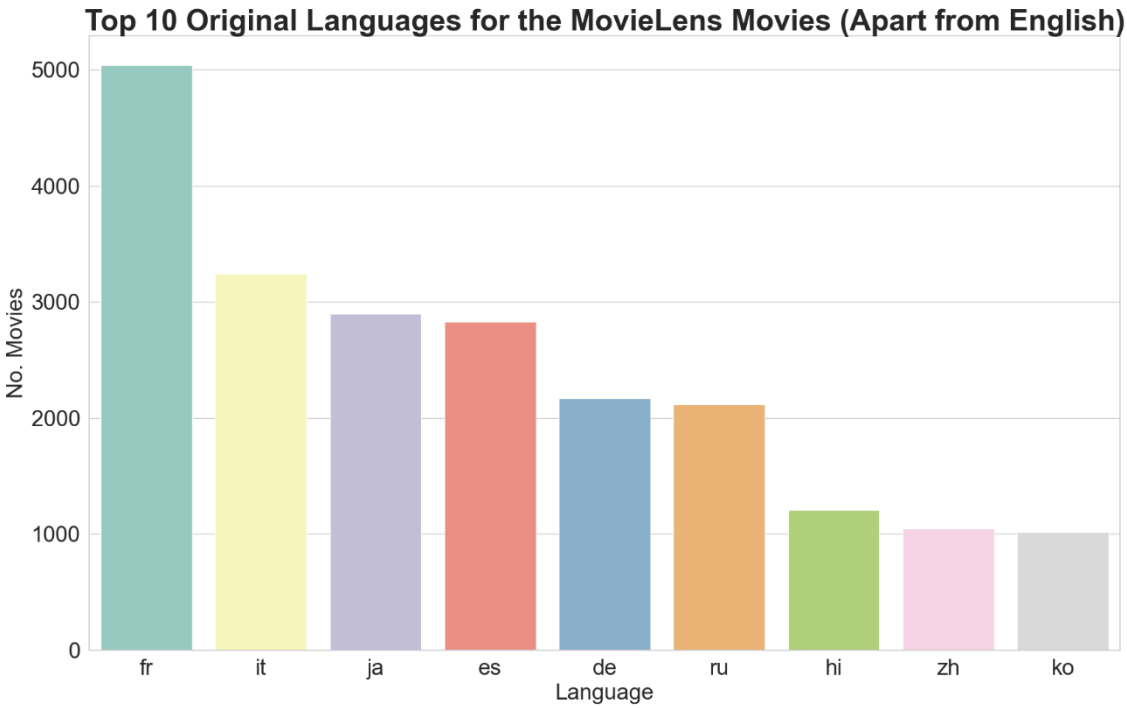

Figure 6.3. Top 10 Original Languages for the MovieLens Movies (Apart from English).

As mentioned earlier, French and Italian are the most commonly occurring languages after English. Japanese and Hindi form the majority as far as Asian Languages are concerned.

6.6. Popularity, vote average and vote count

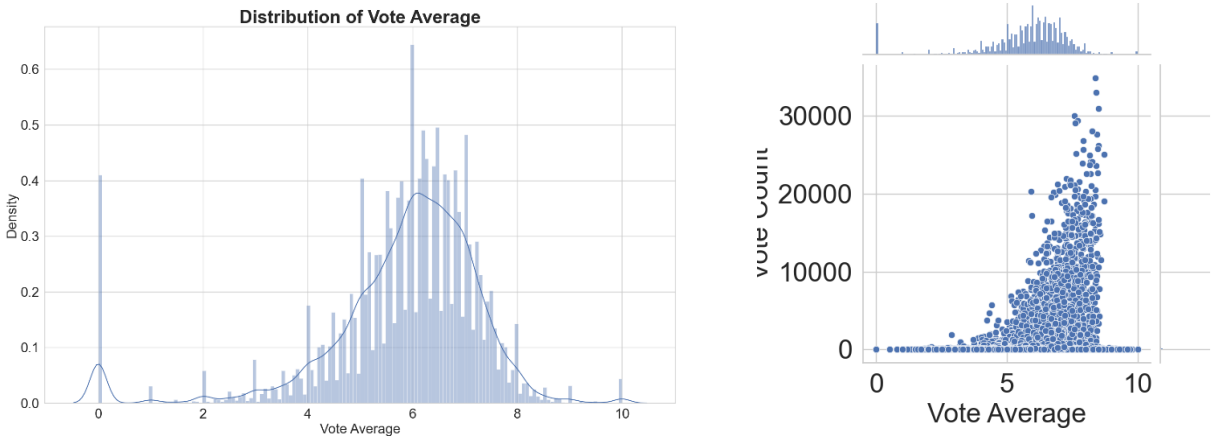

Figure 6.4. Distribution of Vote Average, Vote count - Vote average correlation

Oppenheimer is the most popular movie by the TMDB Popularity Score. Fast X and Mission: Impossible - Dead Reckoning Part One, two extremely successful action movies come in second and third respectively.

Inception and Interstellar, two critically acclaimed and commercially successful Christopher Nolan movies figure at the top of our chart.

It appears that TMDB Users are extremely strict in their ratings. The mean rating is only a 5.286 on a scale of 10. Half the movies have a rating of less than or equal to 6.

The Shawshank Redemption and The Dark Knight are the two most critically acclaimed movies in the TMDB Database. Interestingly, they are the top 2 movies in IMDB's Top 250 Movies list too. They have a rating of over 9 on IMDB as compared to their 8.7 and 8.5 TMDB Scores respectively.

There is a very small correlation between Vote Count and Vote Average. A large number of votes on a particular movie does not necessarily imply that the movie is good.

6.7. Movie release dates

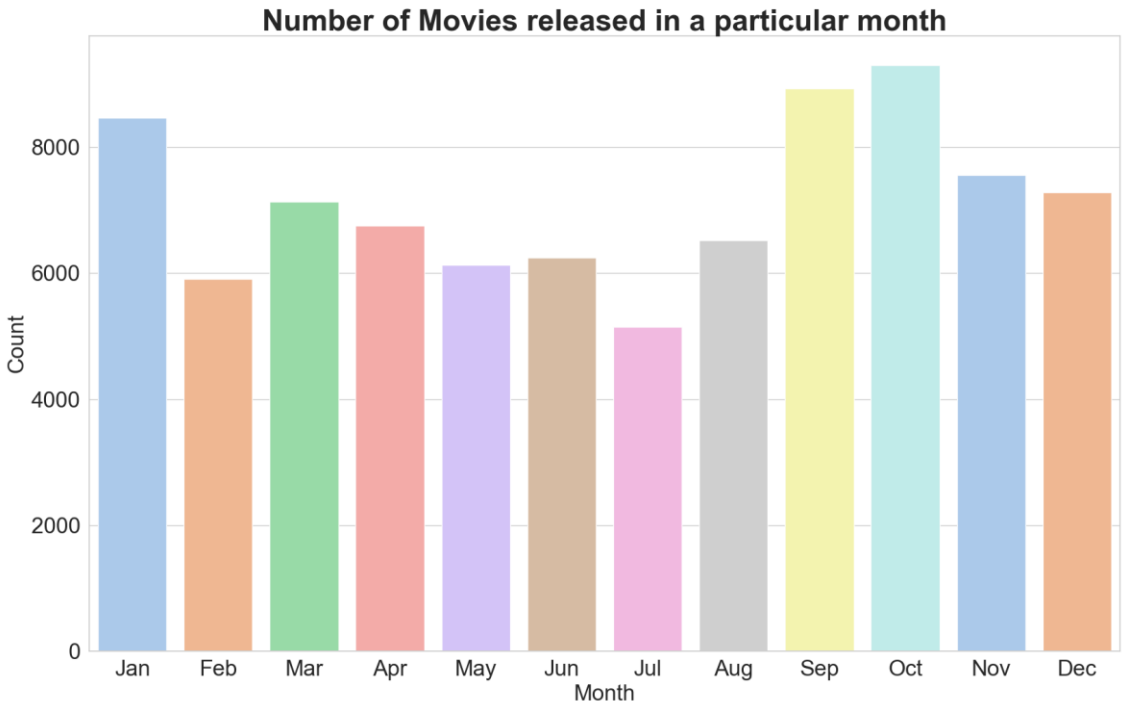

Figure 6.5. Number of Movies released in a particular month

It appears that October is the most popular month when it comes to movie releases.

Awards season: Many prestigious film awards, such as the Oscars, take place early in the following year. Studios release their best films towards the end of the year to be fresh in the minds of voters.

Holiday season: The holiday period, including Thanksgiving and Christmas, is a popular time for people to go to the movies. Studios release their most anticipated films during this time to take advantage of increased audience interest.

Box office performance: Historically, movies released in the fall and winter have performed well at the box office, as people tend to spend more time indoors during these months.

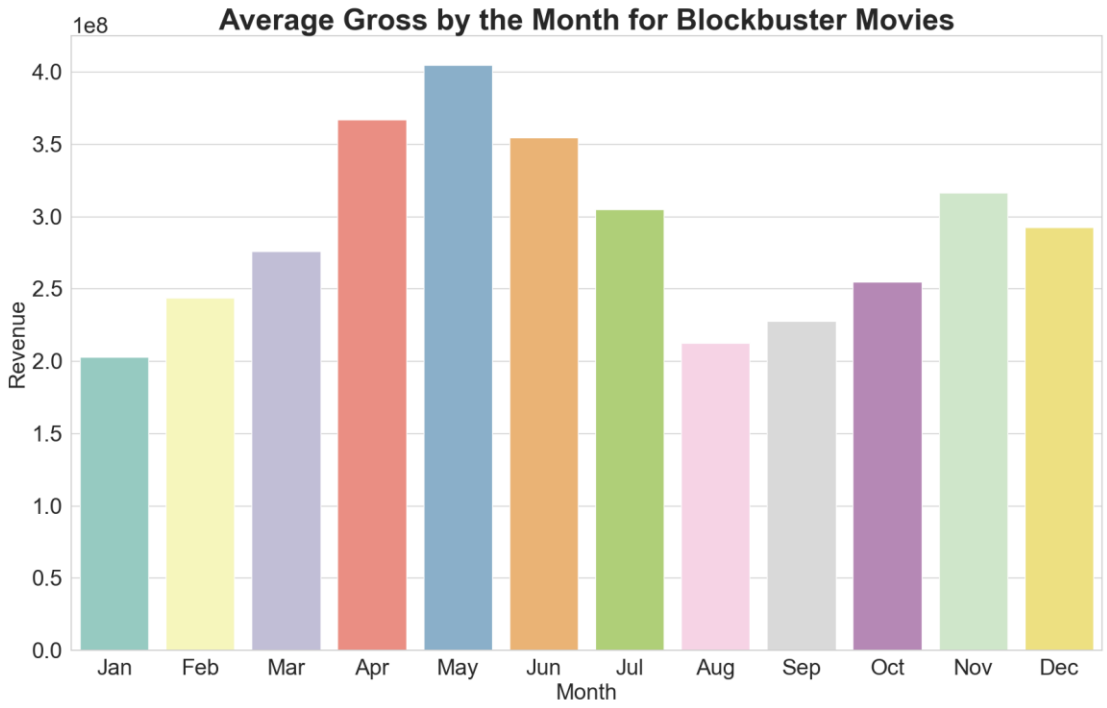

Figure 6.6. Average Gross by the Month for Blockbuster Movies

The months of April, May and June have the highest average gross among high grossing movies. This can be attributed to the fact that blockbuster movies are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment.

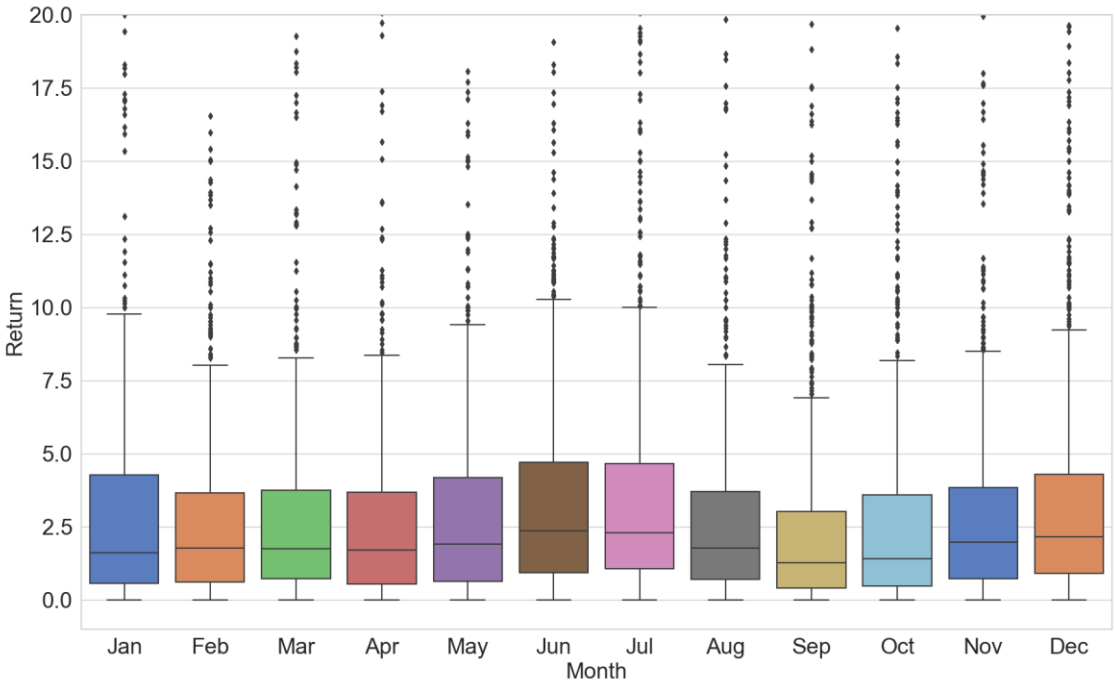

Figure 6.7. Month - Return boxplot

The months of April, May and June have the highest average gross among high grossing movies. This can be attributed to the fact that blockbuster movies are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment.

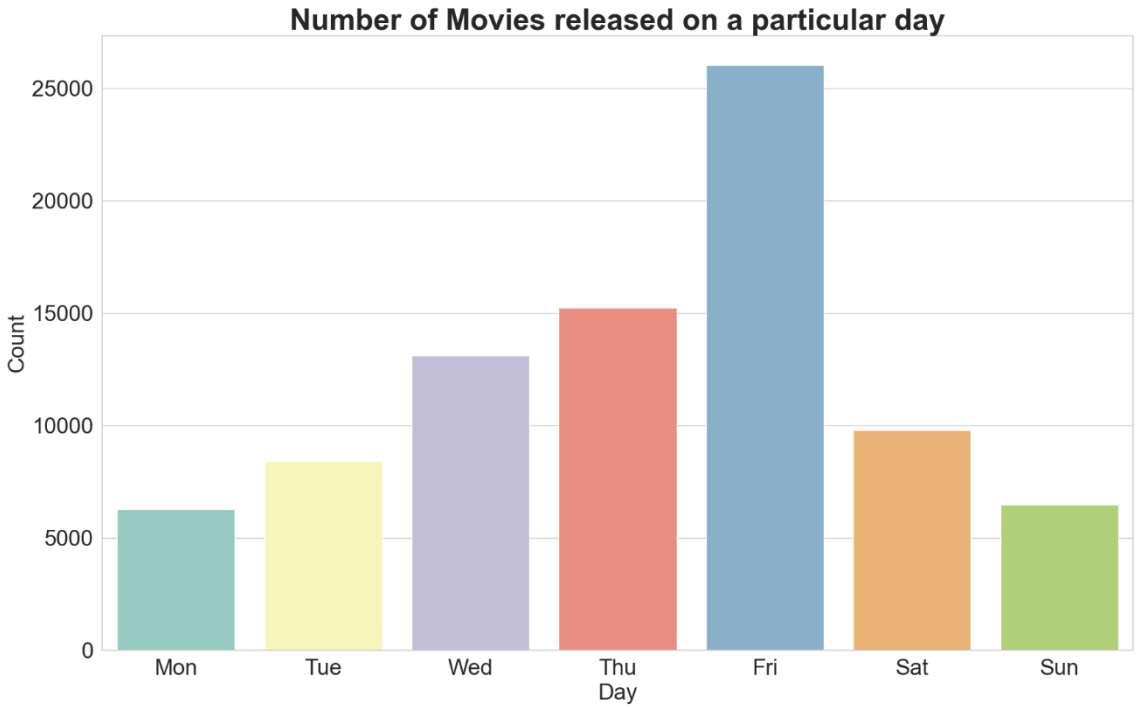

Figure 6.8. Number of Movies released on a particular day

Friday is clearly the most popular day for movie releases. This is understandable considering the fact that it usually denotes the beginning of the weekend. Sunday and Monday are the least popular days and this can be attributed to the same aforementioned reason.

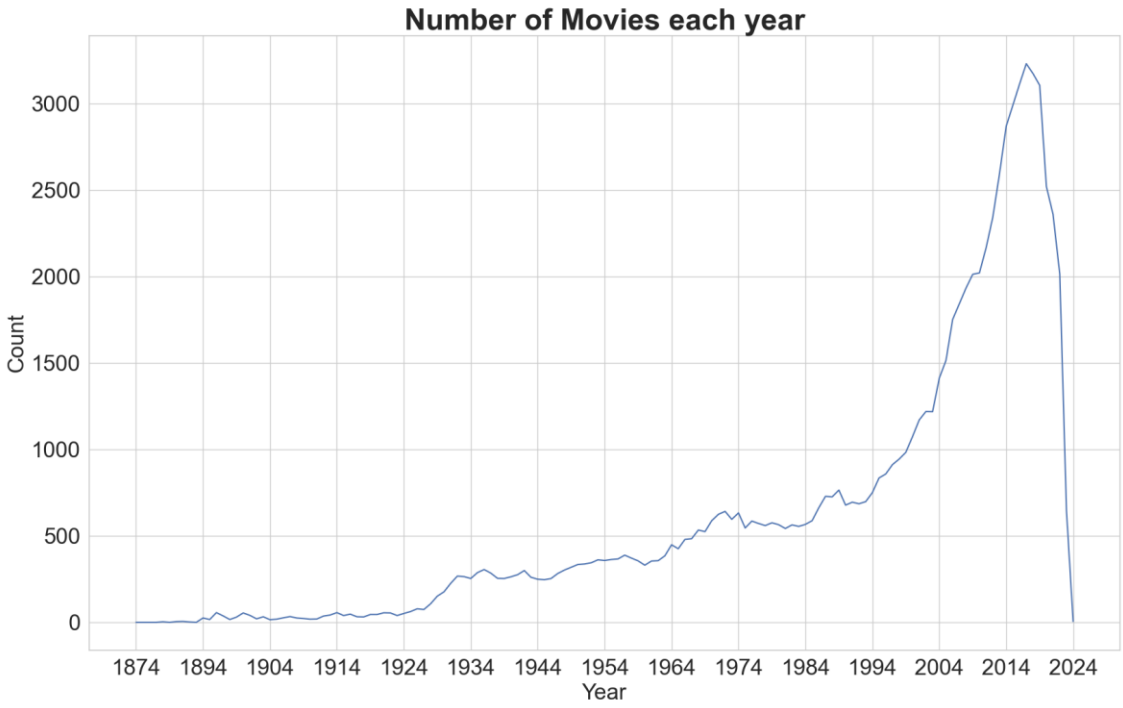

Figure 6.9. Number of Movies each year

It is remarkably noticed that there is a sharp rise in the number of movies starting the 1990s decade. However, it is entirely possible that recent movies were oversampled for the purposes of this dataset.

6.8. Spoken languages

The movie with the most number of languages, Train Station.

Train Station follows a single character, known only as "The Person in Brown", played by 40 actors who vary in age, gender, ethnicity and sexual orientation. Along the character's journey, they are presented with a series of choices - some minor, some life-altering. Cities include Berlin, Bogota, Dubai, Jakarta, Los Angeles, Singapore, Tehran and 20 others across five continents. Train Station unites cultures and breaks language barriers, reminding us that we all live in the same world full of diversity, options and consequences.

6.9. Runtime

The average length of a movie is about 1 hour and 30 minutes. The longest movie on record in this dataset is a staggering 12,480 minutes (or 208 hours) long.

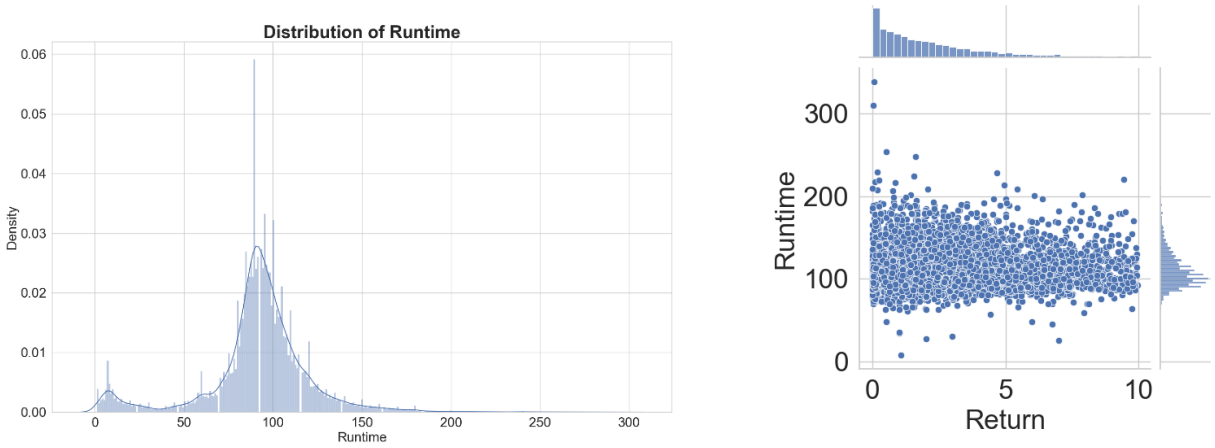

Figure 6.10. Distribution of Runtime, Return - Runtime correlation

There seems to be relationship between the two quantities. The duration of a movie is independent of its success.

However, this might not be the case with duration and budget. A longer movie should entail a higher budget.

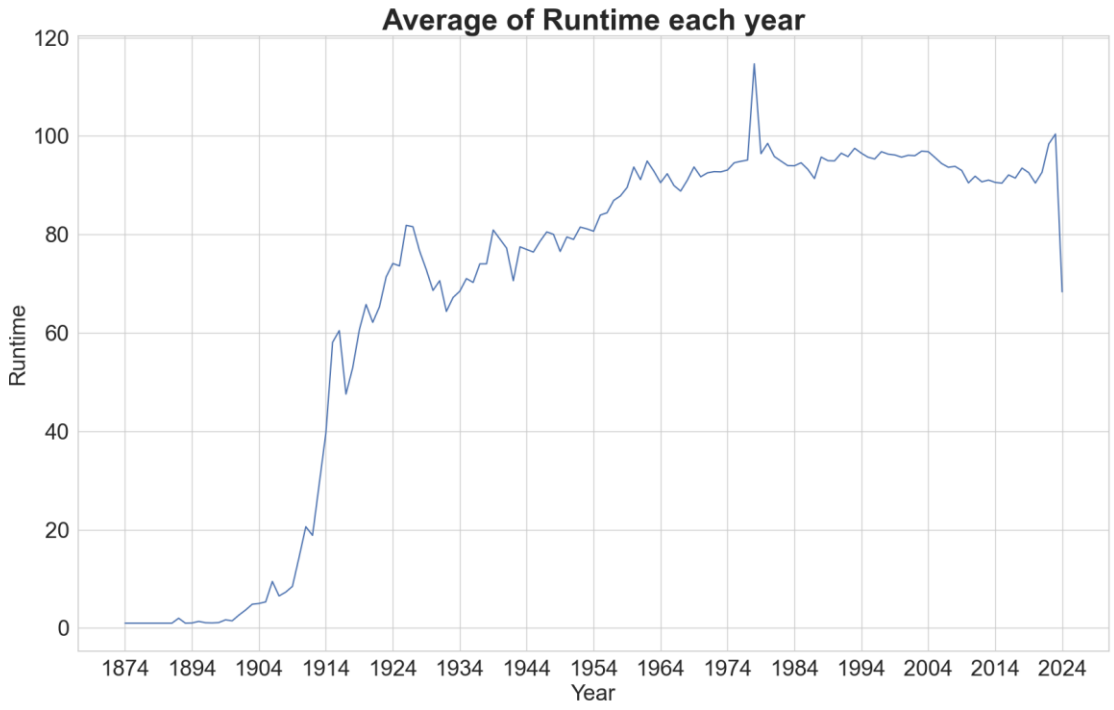

Figure 6.11. Average of Runtime each year

It would be noticed that films started hitting the 60 minute mark as early as 1914. Starting 1924, films started having the traiditonal 90 minute duration and has remained more or less constant ever since.

It can be seen that every movie in this list were filmed in the late 1890s and the beginning of the 20th century. All these movies were one minute long.

Almost all the entries in the above chart are actually miniseries and hence, do not count as feature length films. It cannot gather too much insight from this list of longest movies as there is no way of distinguishing feature length films from TV Mini Series from the dataset (except, of course, by doing it manually).

6.10. Budget

Budgets are expected to be a skewed quantity and also heavily influenced by inflation. Nevertheless, it would be interesting to gather as much insights as possible from this quantity as budget is often a critical feature in predicting movie revenue and success.

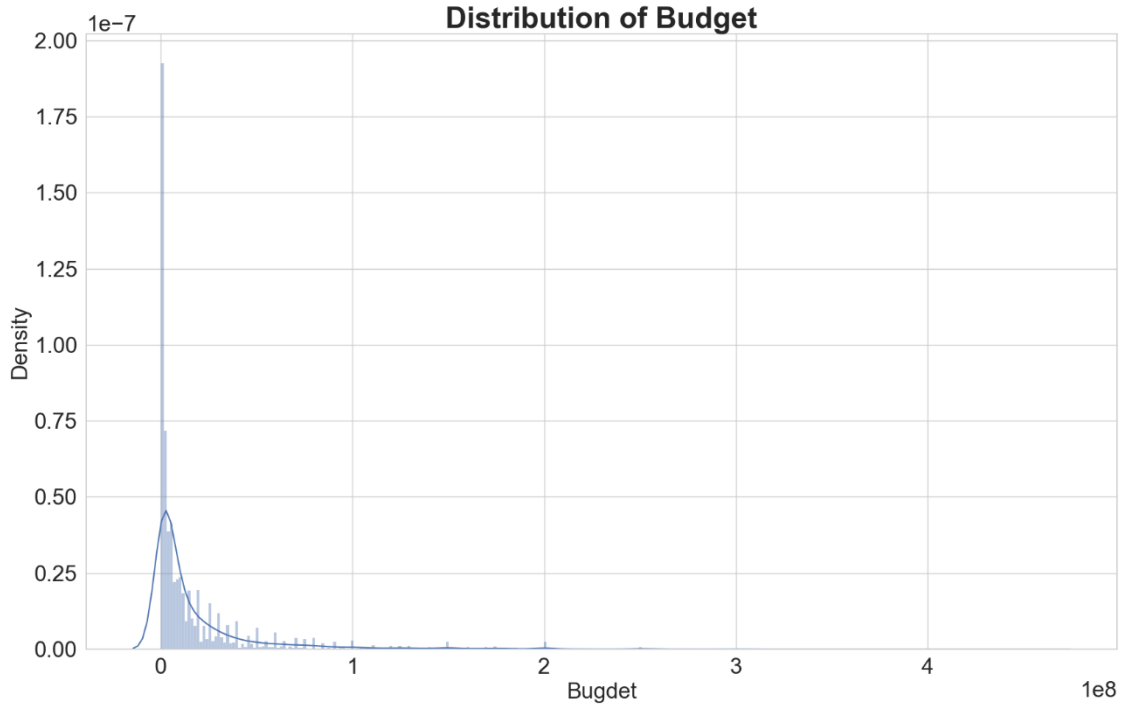

Figure 6.12. Distribution of Budget

The distribution of movie budgets shows an exponential decay. More than 75% of the movies have a budget smaller than 250 million dollars.

Avatar: The Way of Water film - the sequel to Avatar (2009) and the second installment in the Avatar film series - occupy the top spot in this list with a staggering budget of over 460 million dollars. All the top 10 most expensive films made a profit on their investment except for The Flash which managed to recoup about 90% of its investment, taking in a 264 million dollars on a 300 million dollar budget.

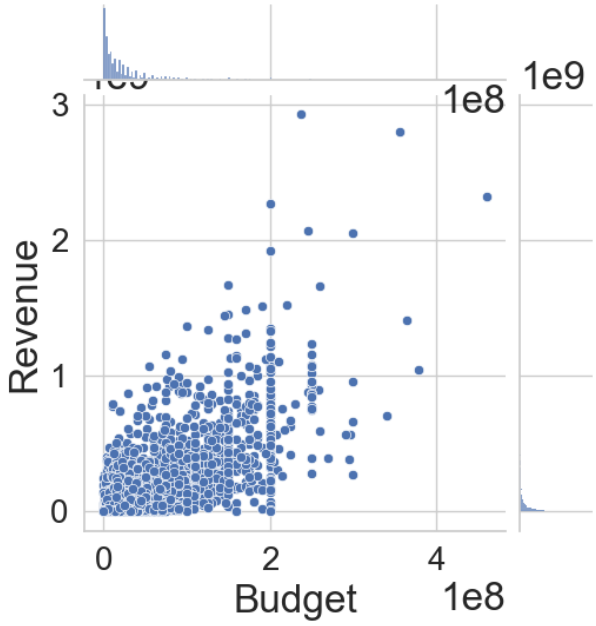

Figure 6.13. Budget - Revenue correlation

Two quantities indicates a very strong correlation.

6.11. Revenue

The mean gross of a movie is 52.8 million dollars whereas the median gross is much lower at 8.3 million dollars, suggesting the skewed nature of revenue. The lowest revenue generated by a movie is just 1 dollar whereas the highest grossing movie of all time has raked in an astonishing *2.92 billion dollars.*

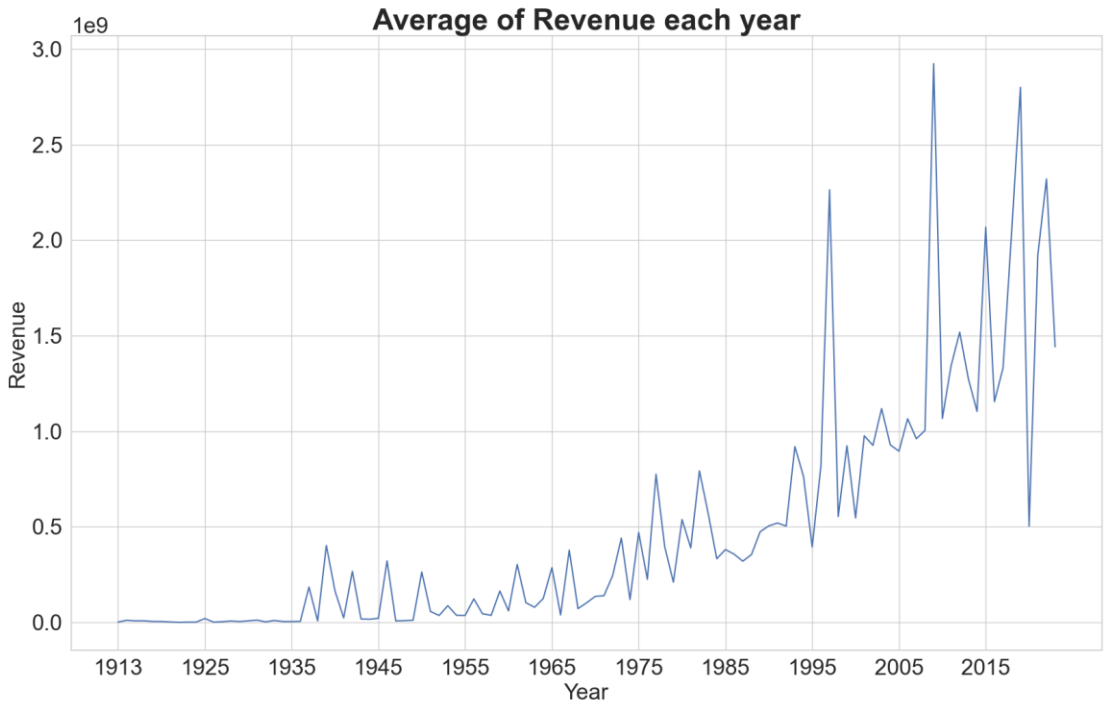

Figure 6.14. Average of Revenue each year

As can be seen from the figure, the maximum gross has steadily risen over the years. The world of movies broke the 1 billion dollar mark in 1997 with the release of Titanic. It took another few years to break the 2 billion dollar mark with Avatar. Both these movies were directed by James Cameron.

6.12. Correlation matrix

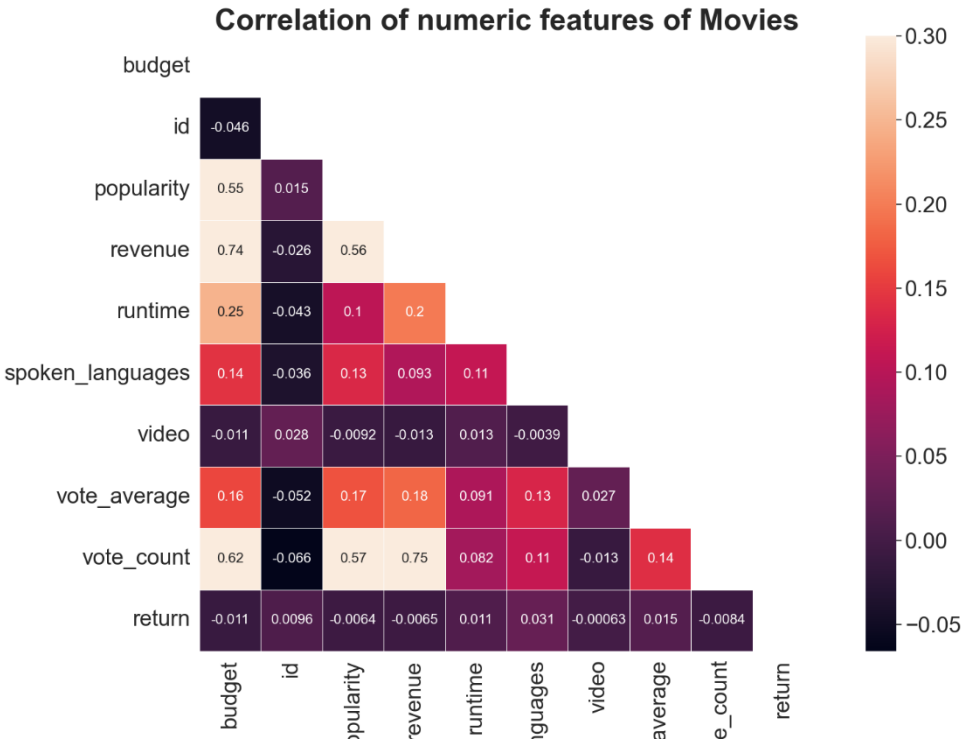

Figure 6.15. Correlation of numeric features of Movies

6.13. Genres

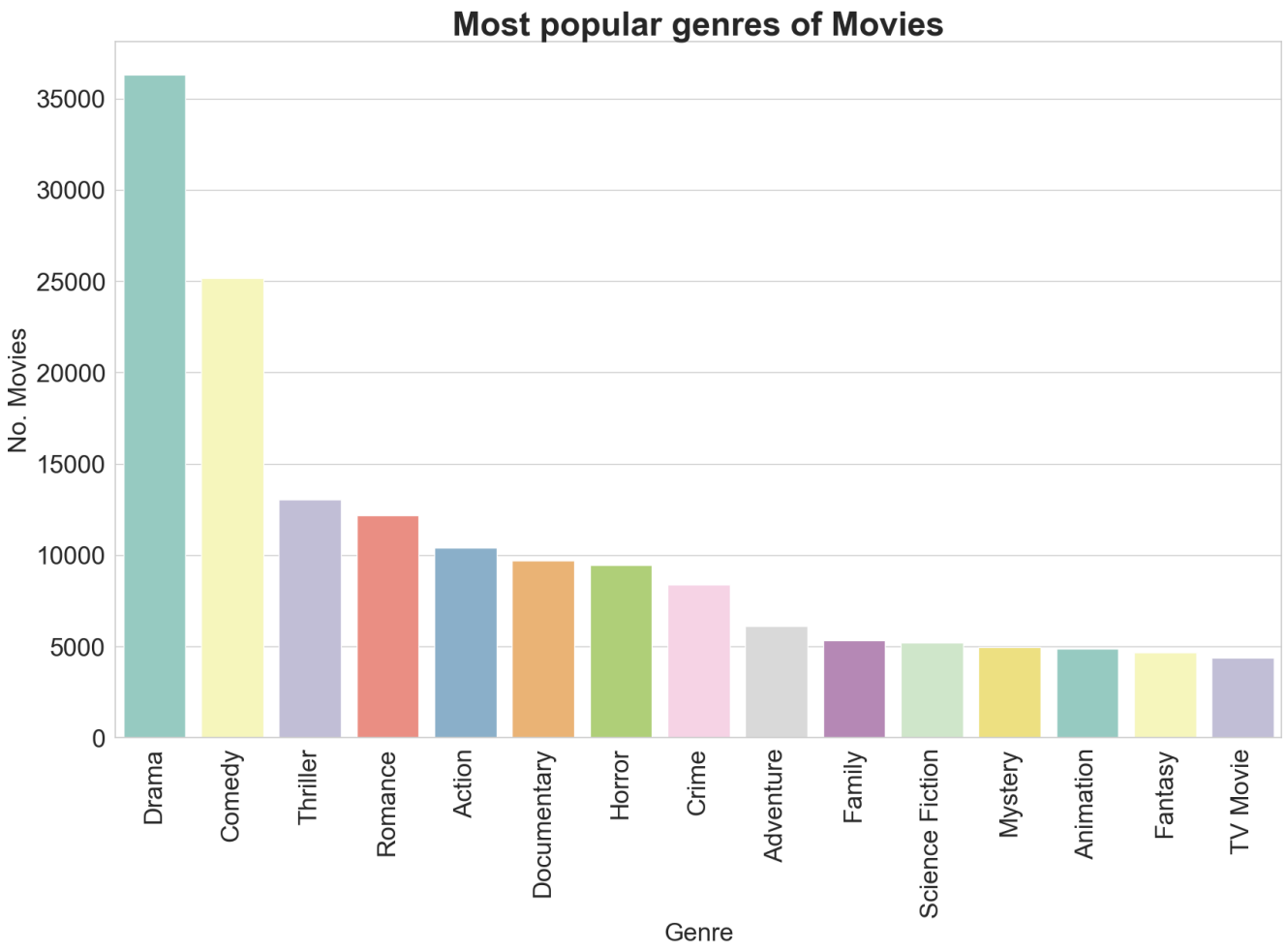

Figure 6.16. Most popular genres of Movies

Drama is the most commonly occurring genre with almost half the movies identifying itself as a drama film. Comedy comes in at a distant second with more than 25% of the movies having adequate doses of humor. Other major genres represented in the top 10 are Thriller, Romance, Action, Documentary, Horror, Crime, Adventure, Family.

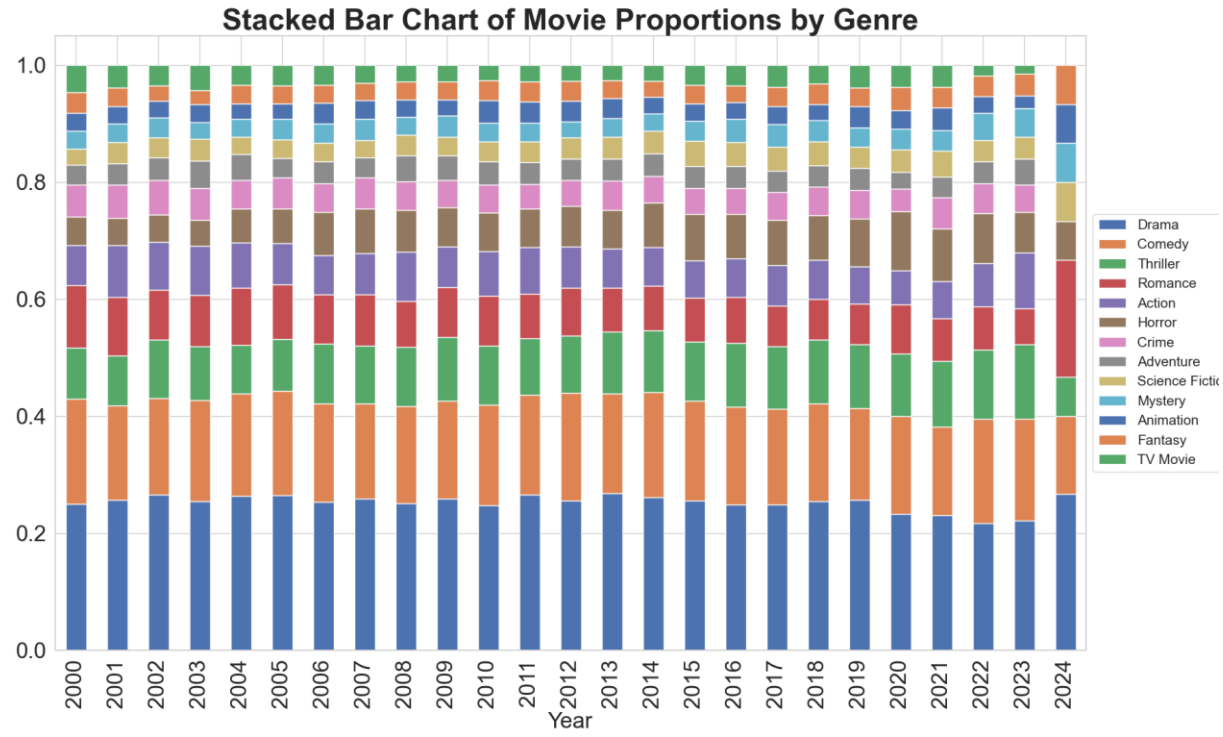

Figure 6.17. Stacked Bar Chart of Movie Proportions by Genre.

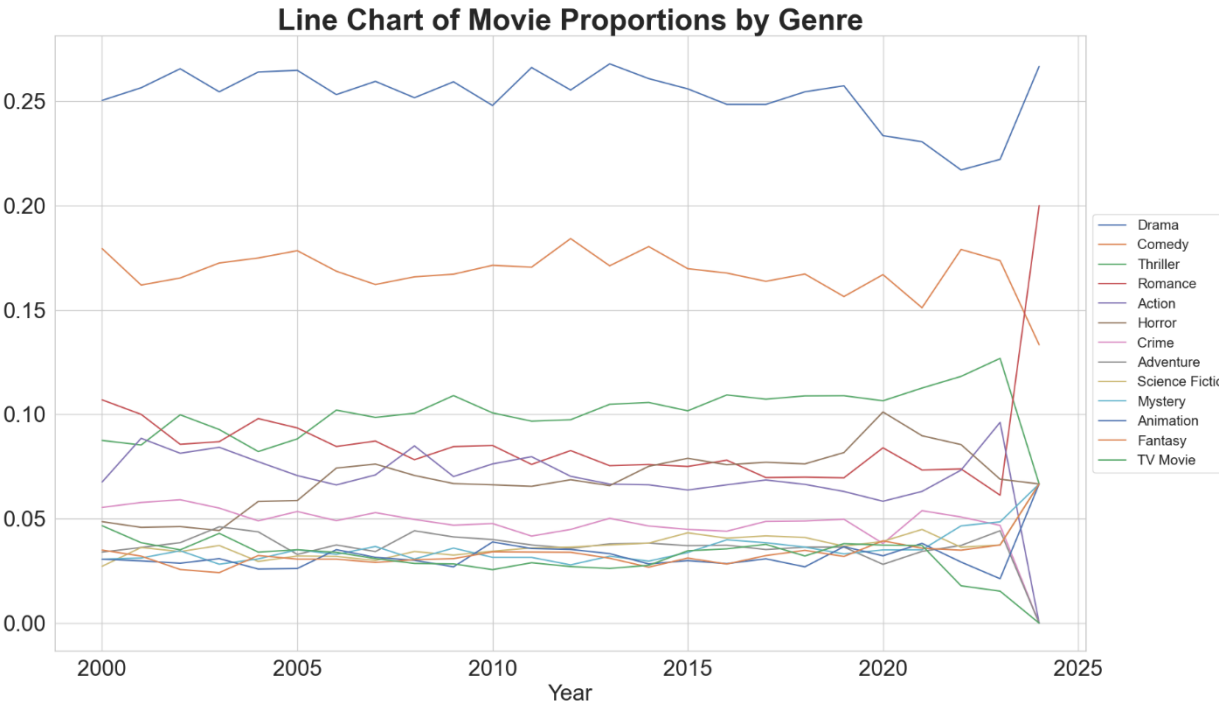

Figure 6.18. Line Chart of Movie Proportions by Genre.

The proportion of movies of each genre has remained fairly constant since the beginning of this century except for Drama. The proportion of drama films has fallen by over 5%. Romance movies have enjoyed a astonishing increase in their share.

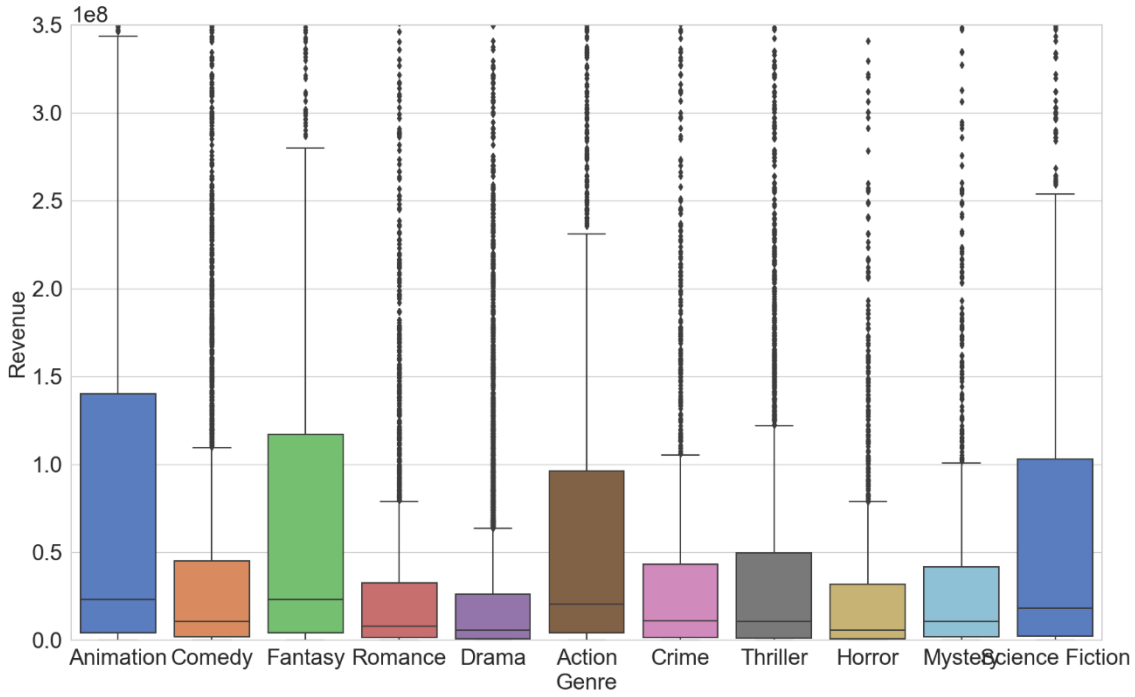

Figure 6.19. Genre – Revenue boxplot

Animation movies has the largest 25-75 range as well as the median revenue among all the genres plotted. Fantasy and Science Fiction have the second and third highest median revenue respectively.

6.14. Cast and crew

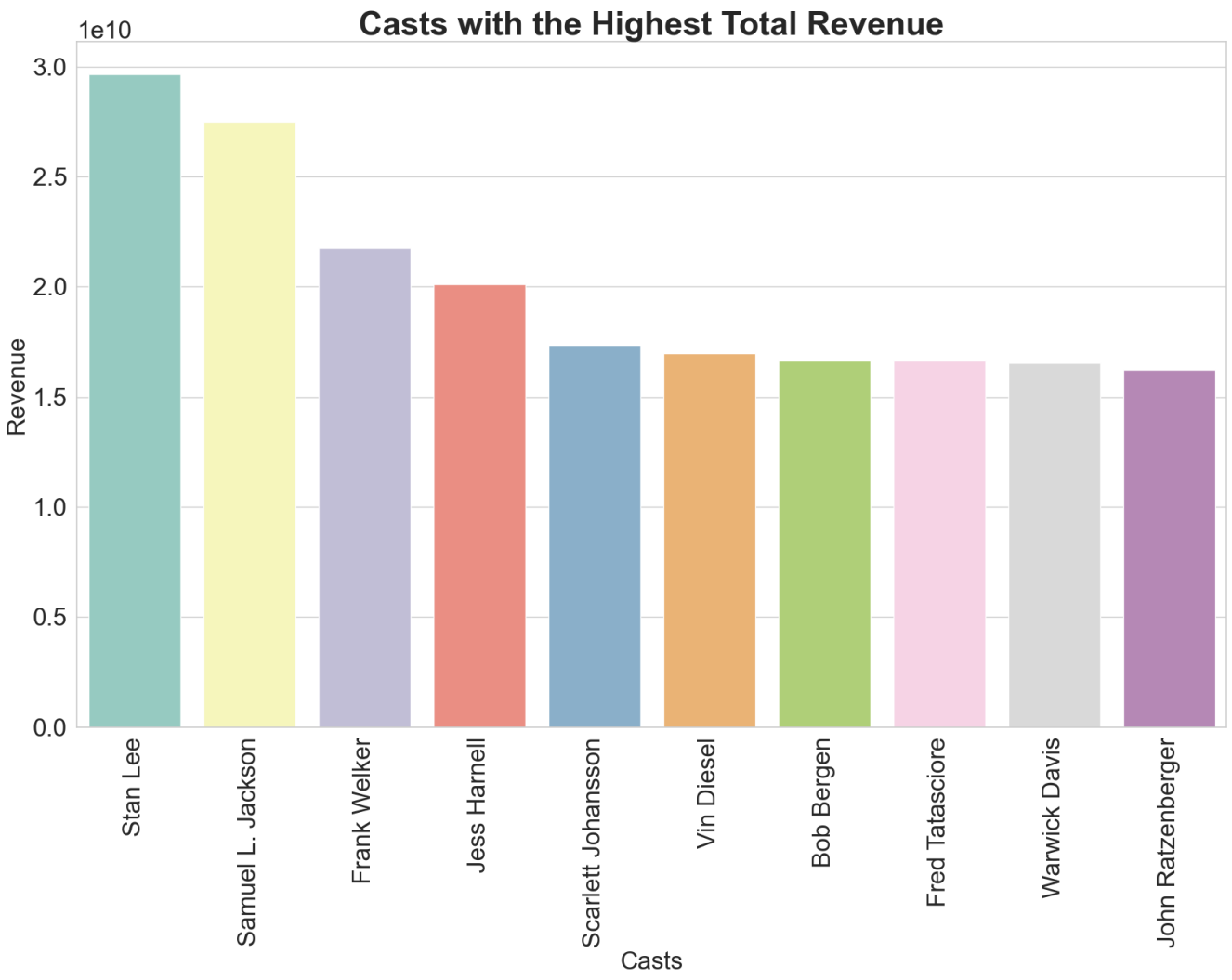

Figure 6.20. Casts with the Highest Total Revenue

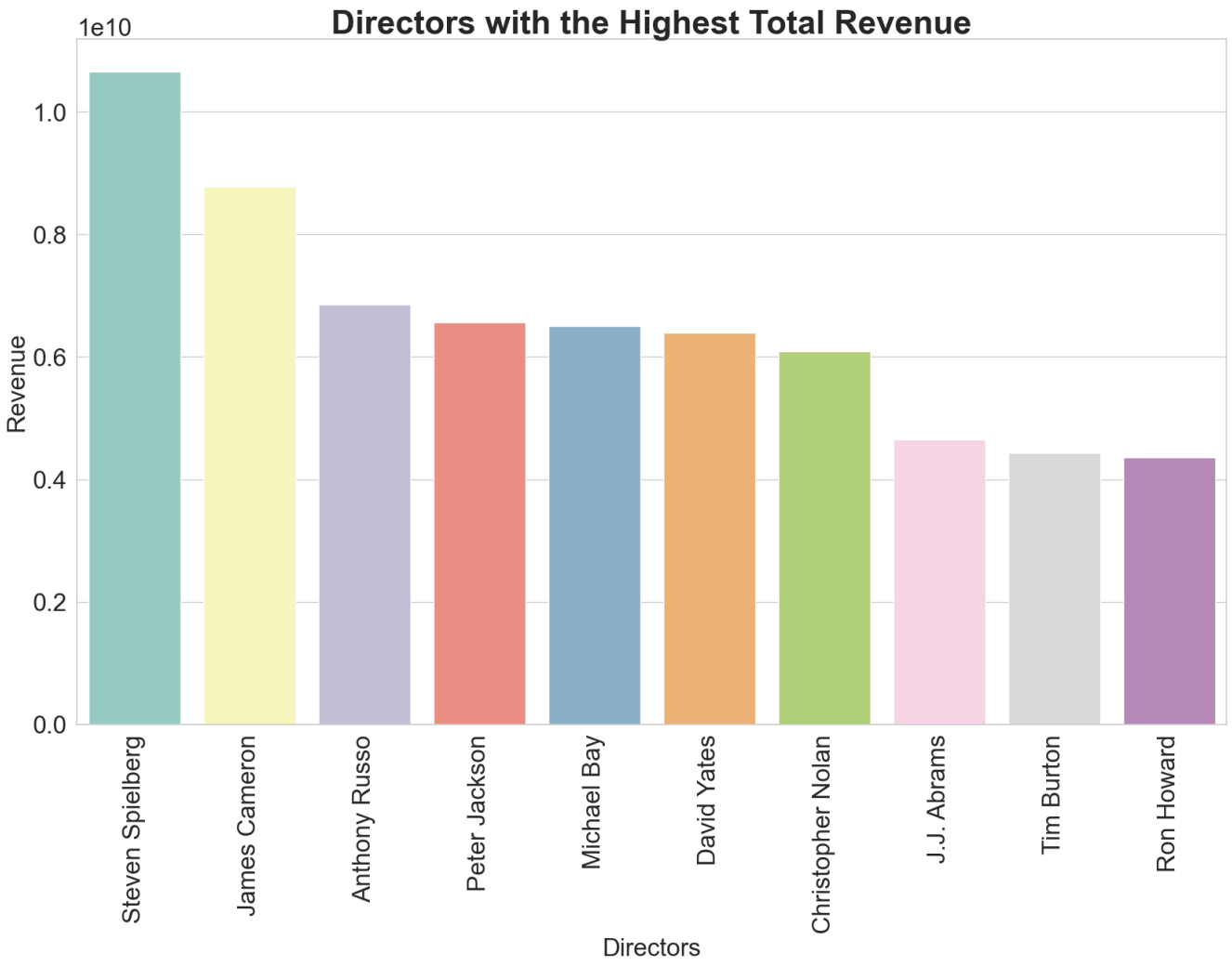

Figure 6.21. Directors with the Highest Total Revenue

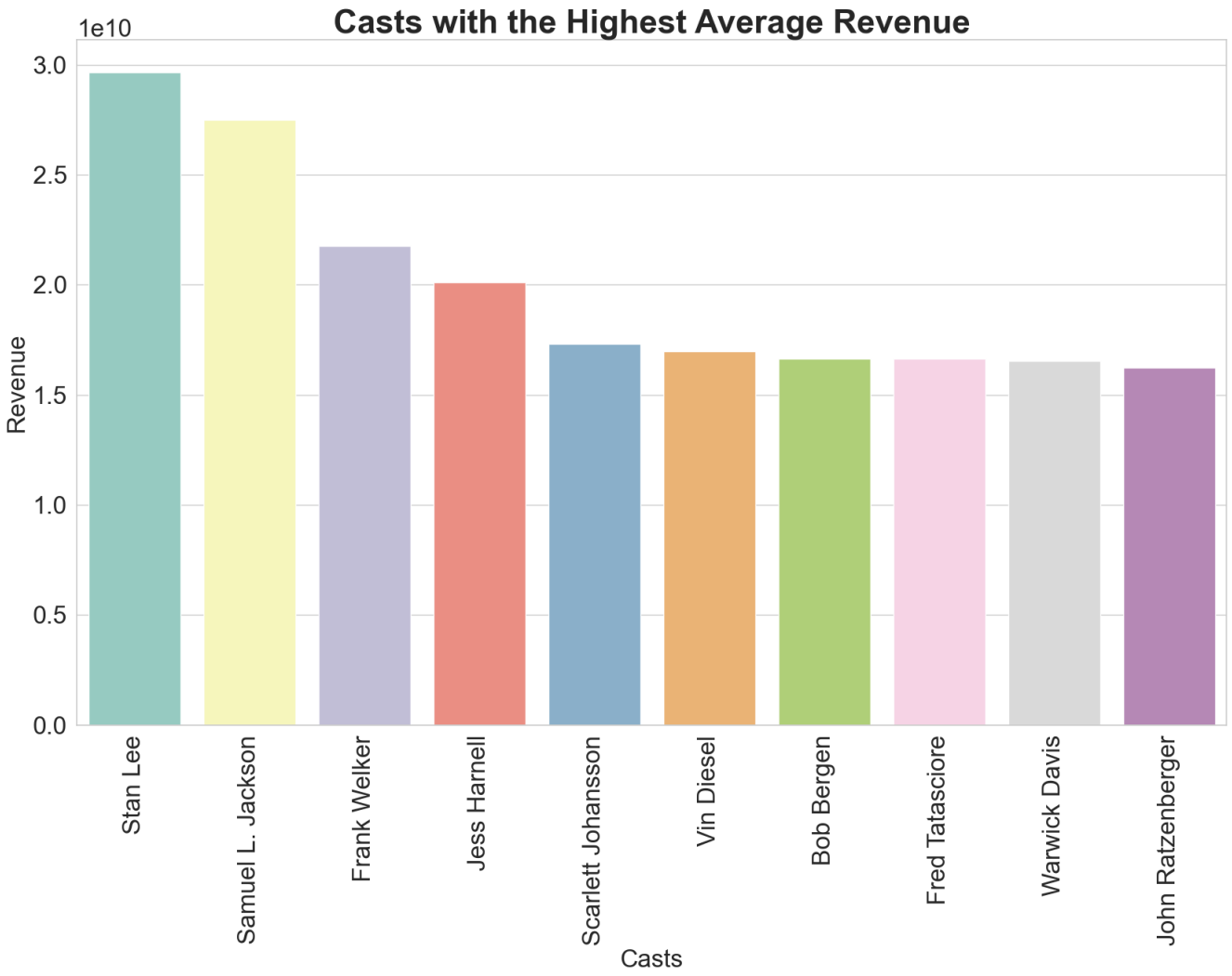

Figure 6.22. Casts with the Highest Average Revenue

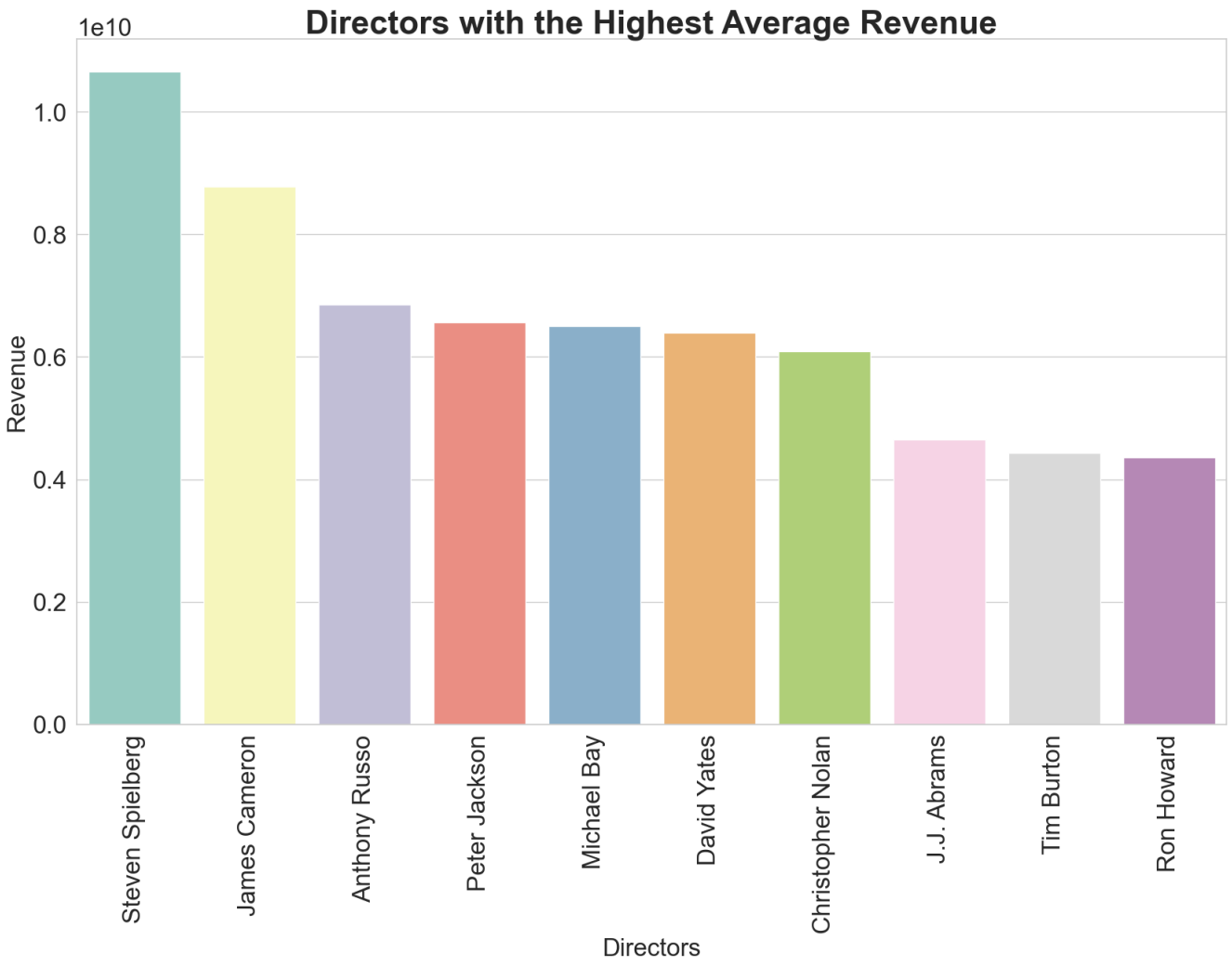

Figure 6.23. Directors with the Highest Average Revenue

7. REGRESSION: PREDICTING MOVIE REVENUES

Predicting Movie Revenues is an extremely popular problem in Machine Learning which has created a huge amount of literature. Most of the models proposed in these papers use far more potent features than what possess at the moment. These include Facebook Page Likes, Information on Tweets about the Movie, YouTube Trailer Reaction (Views, Likes, Dislikes, etc.), Movie Rating (MPCAA, CBIFC) among many others.

To compensate for the lack of these features, there are going to be a few cheats. TMDB's Popularity Score and Vote Average will be being used as features in model to assign a nuerical value to popularity. However, it must be kept in mind that these metrics will not be available when predicting movie revenues in the real world, when the movie has not been released yet.

7.1. Feature engineering

1. belongs_to_collection will be turned into a Boolean variable. 1 indicates a movie is a part of collection whereas 0 indicates it is not.
2. genres will be converted into number of genres.
3. homepage will be converted into a Boolean variable that will indicate if a movie has a homepage or not.
4. original_language will be replaced by a feature called is_english to denote if a particular film is in English or a Foreign Language.
5. production_companies will be replaced with just the number of production companies collaborating to make the movie.
6. production_countries will be replaced with the number of countries the film was shot in.
7. month will be converted into a variable that indicates if the month was a holiday season.
8. day will be converted into a binary feature to indicate if the film was released on a Friday.

7.2. Model

The model chosen for regression is the Gradient Boosting Regression. Coefficient of Determination is 0.7213 which is a pretty score for the basic model that have built.

7.3. Feature importances

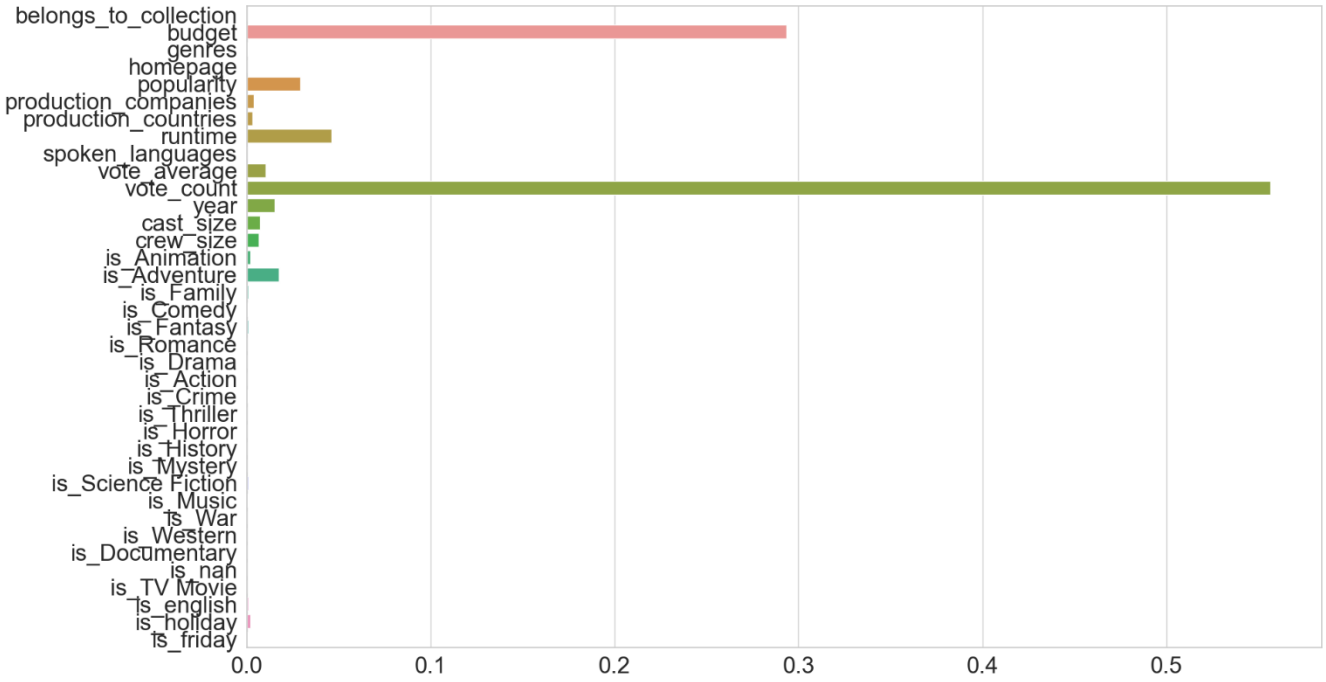

Figure 7.1. Regression model.

It has been noticed that vote_count, a feature was cheated with, is the most important feature to the Gradient Boosting Model.

This goes on to show the importance of popularity metrics in determining the revenue of a movie. Budget was the second most important feature followed by Runtime and Popularity.

8. CLASSIFICATION: PREDICTING MOVIE SUCCESS

As with the regression model, there are a few cheats and use features that may not be available in the real world for the lack of other useful popularity metrics.

Extensive analysis of data has already been performed and haven't been done a lot with respect to determining factors that make a movie a success. Attempt at doing that in this section and follow it up by building a model.

8.1. Model

Gradient Boosting Classifier has an accuracy of 78%. Again, this model can be improved upon through hyperparameter tuning and more advanced feature engineering but since this is not the main objective of this project, this will be skipped.

8.2. Feature importances

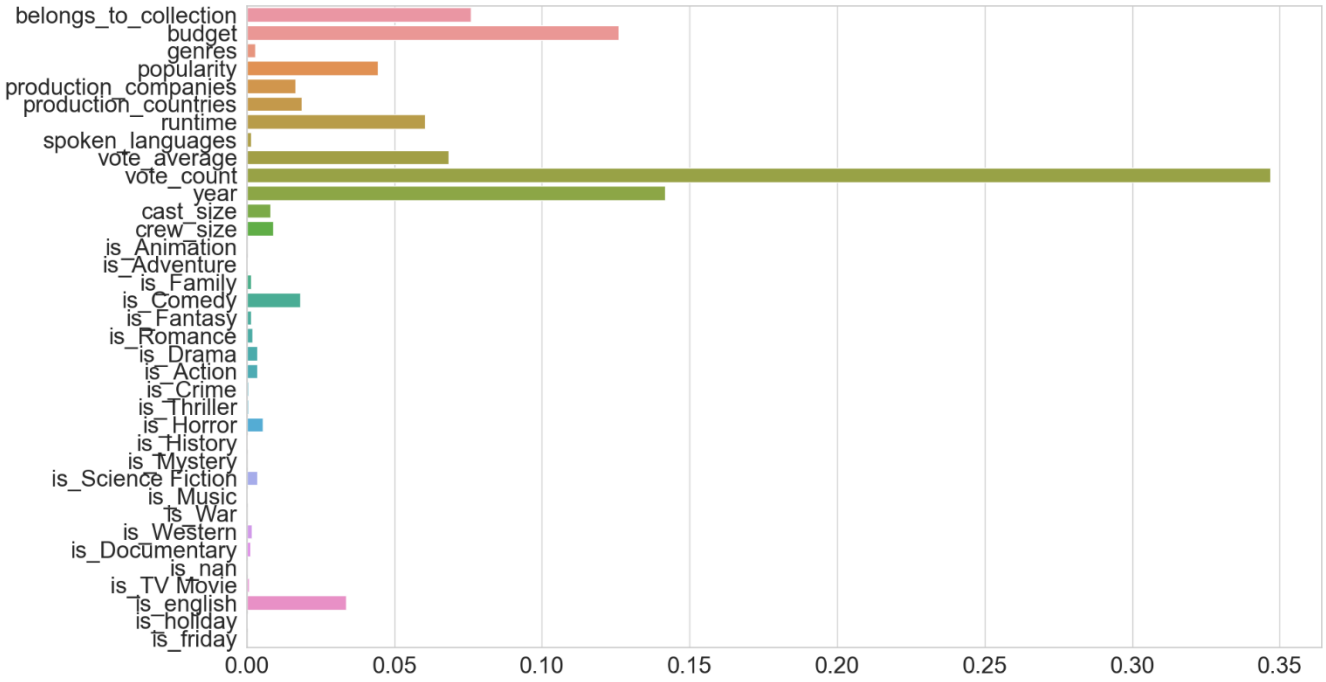

Figure 8.1. Classification model.

It is clear to see that vote _count is once again the most significant feature identified by the Classifier. Other important features include Year, Budget, Belongs to collection and Popularity. With this, it will conclude this discussion on the classification model and move on to the main part of the project.

9. RECOMMENDATION SYSTEMS

9.1. The simple recommender

The Simple Recommender offers generalized recommendations to every user based on movie popularity and (sometimes) genre. The basic idea behind this recommender is that movies that are more popular and more critically acclaimed will have a higher probability of being liked by the average audience. This model does not give personalized recommendations based on the user.

The implementation of this model is extremely trivial. All have to do is sort movies based on ratings and popularity and display the top movies of the list. As an added step, pass in a genre argument to get the top movies of a particular genre.

The next step is to determine an appropriate value for *m*, the minimum votes required to be listed in the chart. Using 95th percentile as cutoff. In other words, for a movie to feature in the charts, it must have more votes than at least 95% of the movies in the list.

An overall Top 250 Chart is built and defined a function to build charts for a particular genre.

	title	genres	popularity	vote_average	vote_count	year	weighted_rating
14848	Inception	[Action, Science Fiction, Adventure]	118.271	8	34845	2010	7.936020
20995	Interstellar	[Adventure, Drama, Science Fiction]	158.041	8	33043	2014	7.932622
12167	The Dark Knight	[Drama, Action, Crime, Thriller]	94.471	8	31010	2008	7.928328
24844	Avengers: Infinity War	[Adventure, Action, Science Fiction]	211.354	8	28055	2018	7.921010
2848	Fight Club	[Drama]	73.879	8	27610	1999	7.919777
292	Pulp Fiction	[Thriller, Crime]	78.343	8	26219	1994	7.915660
351	Forrest Gump	[Comedy, Drama, Romance]	82.404	8	25752	1994	7.914181
314	The Shawshank Redemption	[Drama, Crime]	259.230	8	25045	1994	7.911842
18866	Django Unchained	[Drama, Western]	56.438	8	24935	2012	7.911466
24845	Avengers: Endgame	[Adventure, Science Fiction, Action]	134.347	8	24152	2019	7.908698
2463	The Matrix	[Action, Science Fiction]	84.441	8	24149	1999	7.908687
61336	Joker	[Crime, Thriller, Drama]	64.097	8	23731	2019	7.907136
4866	The Lord of the Rings: The Fellowship of the Ring	[Adventure, Fantasy, Action]	113.568	8	23613	2001	7.906689
7003	The Lord of the Rings: The Return of the King	[Adventure, Fantasy, Action]	107.280	8	22651	2003	7.902876
14259	Shutter Island	[Drama, Thriller, Mystery]	65.617	8	22597	2010	7.902652

Figure 9.1. Top 250 recommended movies

Three Christopher Nolan Films, Inception, Interstellar and The Dark Knight occur at the very top of the chart. The chart also indicates a strong bias of TMDB Users towards particular genres and directors.

9.2. Content based recommender

The recommender built in the previous section suffers some severe limitations. For one, it gives the same recommendation to everyone, regardless of the user's personal taste. If a person who loves romantic movies (and hates action) were to look at our Top 15 Chart, she/he wouldn't probably like most of the movies. If she/he were to go one step further and look at the charts by genre, she/he wouldn't still be getting the best recommendations.

To personalise recommendations more, an engine is going to be built that computes similarity between movies based on certain metrics and suggests movies that are most similar to a particular movie that a user liked. Since using movie metadata (or content) to build this engine, this also known as Content Based Filtering.

Build two Content Based Recommenders based on:

- Movie Overviews and Taglines
- Movie Cast, Crew, Keywords and Genre

Also, as mentioned in the introduction, using a subset of all the movies available due to limiting personal computing power available.

```
improved_recommendations('Fast & Furious')
```

[63] ✓ 0.0s Python

...

	title	vote_count	vote_average	year	weighted_rating
8487	Furious 7	10086	7	2015	6.936316
7414	Fast Five	7666	7	2011	6.917744
915	The Killer	689	7	1989	6.483301
7906	Fast & Furious 6	10096	6	2013	5.995287
3221	The Fast and the Furious	9322	6	2001	5.994921
6145	The Fast and the Furious: Tokyo Drift	6318	6	2006	5.992725
8088	Need for Speed	4101	6	2014	5.989318
7449	Takers	1170	6	2010	5.971942
6056	Running Scared	952	6	2006	5.968081
6727	Righteous Kill	1143	5	2008	5.327572

```
improved_recommendations('Twilight')
```

[64] ✓ 0.1s

	title	vote_count	vote_average	year	weighted_rating
7038	The Twilight Saga: New Moon	8706	6	2009	5.985141
7795	The Twilight Saga: Breaking Dawn - Part 2	8433	6	2012	5.984721
7196	The Twilight Saga: Eclipse	8312	6	2010	5.984527
7554	The Twilight Saga: Breaking Dawn - Part 1	8311	6	2011	5.984526
8439	Paper Towns	4901	6	2015	5.975919
6876	17 Again	4839	6	2009	5.975673
7851	Beautiful Creatures	2726	6	2013	5.962681
7411	Water for Elephants	2408	6	2011	5.959420
4848	Thirteen	1486	6	2003	5.945646
7379	Red Riding Hood	2991	5	2011	5.256491

9.3. Collaborative filtering

The Content Based engine suffers from some severe limitations. It is only capable of suggesting movies which are *close* to a certain movie. That is, it is not capable of capturing tastes and providing recommendations across genres.

Also, the engine that built is not really personal in that it doesn't capture the personal tastes and biases of a user. Anyone querying this engine for recommendations based on a movie will receive the same recommendations for that movie, regardless of who she/he is.

Therefore, in this section, it is about to use a technique called Collaborative Filtering to make recommendations to Movie Watchers. Collaborative filtering filters information by using the interactions and data collected by the system from other users. It’s based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future.

It will not be implementing Collaborative Filtering from scratch. Instead, it will use the Surprise library that used extremely powerful algorithms like Singular Value Decomposition (SVD) to minimise RMSE (Root Mean Square Error) and give great recommendations.

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8758  0.8813  0.8719  0.8703  0.8740  0.8747  0.0038
MAE (testset)     0.6730  0.6772  0.6687  0.6686  0.6737  0.6722  0.0033
Fit time          2.53    1.17    1.33    1.24    1.24    1.50    0.52
Test time         0.90    0.15    0.39    0.17    0.18    0.36    0.28

{'test_rmse': array([0.87580527, 0.88130383, 0.87187043, 0.870342  , 0.87401289]),
 'test_mae': array([0.6730075 , 0.67717802, 0.66865805, 0.66857067, 0.67374549]),
 'fit_time': (2.5314135551452637,
  1.1726486682891846,
  1.3252294063568115,
  1.2438914775848389,
  1.2426700592041016),
 'test_time': (0.9007225036621094,
  0.15437030792236328,
  0.38810062408447266,
  0.16626310348510742,
  0.18392682075500488)}
```

A mean Root Mean Sqaure Error of 0.8747 which is more than good enough for this case.

9.4. Hybrid recommender

In this section, a main purpose is to build a Simple Hybrid Recommender that brings together techniques that have been implemented in the Content Based and Collaborative Filter Based engines. This is how it will work:

- Input: User ID and the Title of a Movie
- Output: Similar movies sorted on the basis of expected ratings by that particular user.

```
hybrid_recommendations(1, 'Iron Man')
```

[80] ✓ 0.1s Python

	title	id	vote_count	vote_average	year	est
8166	Guardians of the Galaxy	118340	26869	7.905	2014	4.769449
8377	Thor: Ragnarok	284053	19720	7.594	2017	4.676347
7434	X-Men: First Class	49538	12093	7.297	2011	4.622500
4317	X2	36658	9521	7.000	2003	4.578880
7502	The Avengers	24428	29412	7.711	2012	4.526972
7459	Captain America: The First Avenger	1771	20446	6.995	2011	4.503477
8375	Black Panther	284054	21221	7.388	2018	4.488888
8096	Captain America: The Winter Soldier	100402	17866	7.670	2014	4.477583
8378	Guardians of the Galaxy Vol. 2	283995	20544	7.621	2017	4.406626
2826	X-Men	36657	10643	6.998	2000	4.392230

```
hybrid_recommendations(500, 'Iron Man')
```

[81] ✓ 0.1s Python

	title	id	vote_count	vote_average	year	est
8378	Guardians of the Galaxy Vol. 2	283995	20544	7.621	2017	3.740729
8377	Thor: Ragnarok	284053	19720	7.594	2017	3.659600
7434	X-Men: First Class	49538	12093	7.297	2011	3.501074
8372	Ant-Man	102899	18882	7.081	2015	3.488785
8376	Avengers: Infinity War	299536	28055	8.252	2018	3.433095
9107	Marvel One-Shot: Agent Carter	211387	690	7.200	2013	3.431697
8379	Captain America: Civil War	271110	21721	7.442	2016	3.402881
8096	Captain America: The Winter Soldier	100402	17866	7.670	2014	3.358924
4317	X2	36658	9521	7.000	2003	3.349765
8369	Avengers: Age of Ultron	99861	21951	7.273	2015	3.346312

It can be seen that for the Hybrid Recommender, it can get different recommendations for different users although the movie is the same. Hence, this recommendations are more personalized and tailored towards particular users.

10. CONCLUSION

In this notebook, 4 different recommendation engines have been built based on different ideas and algorithms. They are as follows:

1. **Simple Recommender:** This system used overall TMDB Vote Count and Vote Averages to build Top Movies Charts, in general and for a specific genre. The IMDB Weighted Rating System was used to calculate ratings on which the sorting was finally performed.
2. **Content Based Recommender:** This system was built by two content based engines: One that took movie overview and taglines as input. The other which took metadata such as cast, crew, genre and keywords to come up with predictions. They were also deviced a simple filter to give greater preference to movies with more votes and higher ratings.
3. **Collaborative Filtering:** The powerful Surprise Library was used to build a Collaborative Filter based on single value decomposition. The RMSE obtained was less than 1 and the engine gave estimated ratings for a given user and movie.
4. **Hybrid Engine:** All the ideas from Content and Collaborative Filterting were brought together to build an engine that gave movie suggestions to a particular user based on the estimated ratings that it had internally calculated for that user.

The code associated with this report is available at:

https://github.com/thanhnghth99/recommender_system.git

REFERENCES

[1] Department of Computer Science and Engineering at the University of Minnesota, "MovieLens," 1997. [Online]. Available: https://grouplens.org/datasets/movielens/. [Accessed 01 December 2023].

[2] "The Movie Database (TMDB)," 2008. [Online]. Available: https://www.themoviedb.org/. [Accessed 01 December 2023].