IMDB Movie Dataset Analysis

Name: Mohammed Ahsan Bagsaria

Section & Roll No.: CE 85

Reg. No.: 240968484

Table Of Contents

1. Setting the Stage: The Purpose of Our Analysis (Objective of the Analysis)

Problem Statement

The movie industry is a vast and ever-expanding landscape, with hundreds of films released each year across various genres, languages, and production scales. With such a massive influx of content, organizing, analyzing, and extracting meaningful insights from movie-related data becomes increasingly challenging. Researchers, film enthusiasts, and data scientists often struggle to track key characteristics such as the performance of films, directors, etc. over long periods of time.

Identifying patterns in vast and diverse movie data can be challenging. A well-structured analysis, combined with meaningful insights, is crucial for making informed decisions in research, entertainment, and business.

Solution

To effectively analyze movie data and derive meaningful insights, we can develop a structured approach that focuses on key aspects of the dataset:

- Trend Analysis – Identify patterns over time.
- Category Performance – Compare different groups and attributes.
- Influence Assessment – Evaluate key contributors' impact.
- Audience Insights – Understand preferences and behaviors.
- Correlation Study – Examine relationships between factors.

2. Behind the Scenes: Cleaning & Preparing the Data (Data Cleaning and Preprocessing)

The Need for Pre-Processing

- To effectively and accurately analyze data and make the right decisions, the dataset needs to be checked for any irregularities.
- If found, they should be removed with efficient techniques like the IQR method, imputing, standardization, etc.
- Outliers in the data can lead to significant mis-predictions, making visualizing and analyzing data completely worthless.
- This can lead to potential losses to the companies and loss of reputation for the analysts.

Data Integration and Cleaning

The following steps were taken:

a) Data Merging:

   Three datasets were merged into one

b) Duplicate Removal:

   Duplicates were removed to ensure data accuracy

c) Missing Values:

   No missing values were found in the dataset

d) Checking for Outliers:

   No outliers were detected in IMDb Ratings or MetaScore

3. Unveiling the Dataset: Features, Relationships, and Insights (Data Exploration)

One of the most important parts of analyzing the data is determining the relation between its features. This involves using techniques like correlation heatmaps to highlight dependencies, regression plots to observe trends, and other visualization methods to uncover hidden patterns. Gaining clarity on feature relationships helps streamline insight collection and deeper analysis.

Dataset Overview

- **Entries**
  3895 movies after cleaning
- **Directors**
  2603 directors
- **Years**
  1917 to 2025 timeframe
- **Star Cast**
  3427 Star Cast
- **Genres**
  18 unique genres
- **Certificates**
  18 certificates

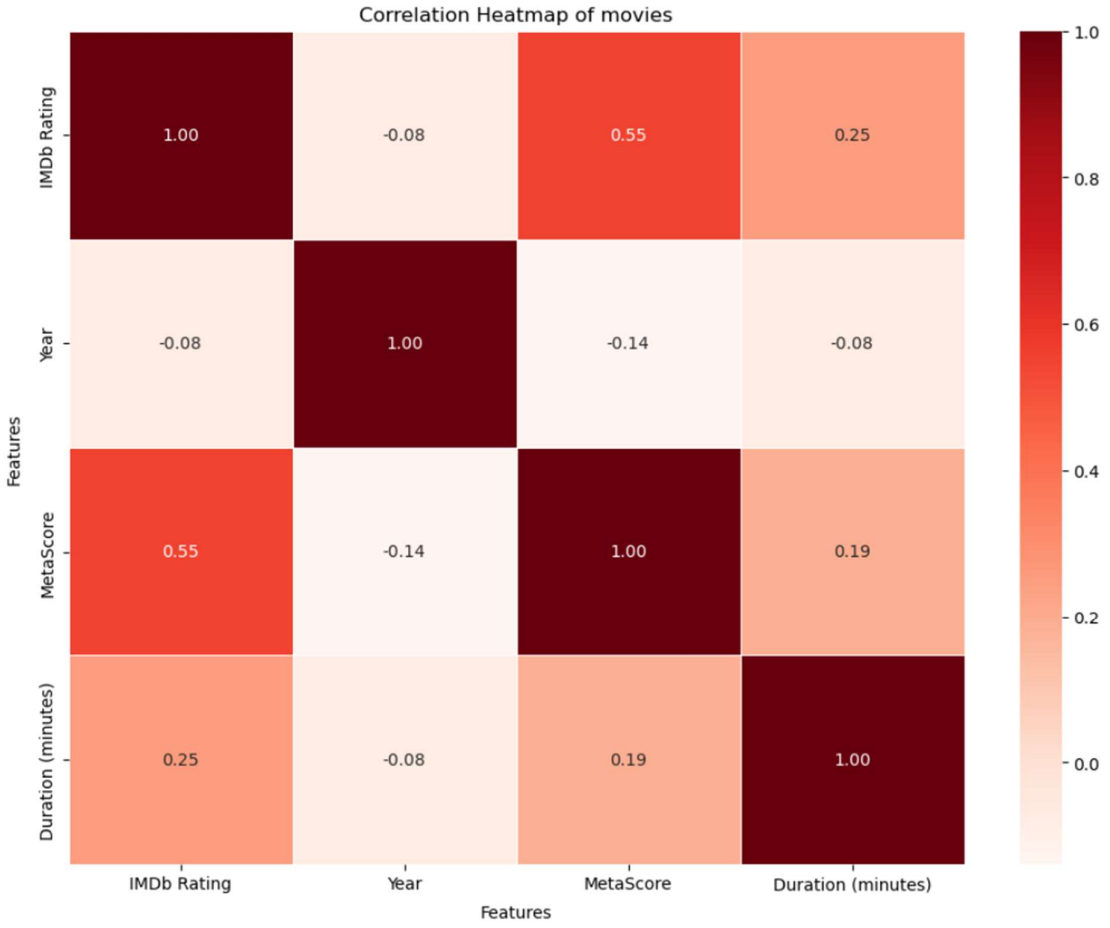
Description of the Dataset

This gives us a **5-number summary** for all the **numerical variables.** This gives us information about the average, max ratings, MetaScore etc. important for understanding the details of films produced in the given timeframe.

	IMDb Rating	Year	MetaScore	Duration (minutes)
count	3895.000000	3895.000000	3895.000000	3895.000000
mean	6.775841	2004.586393	63.502054	111.337535
std	0.873039	17.477366	12.736802	21.205658
min	3.900000	1917.000000	11.000000	46.000000
25%	6.200000	1996.000000	57.100000	98.000000
50%	6.800000	2010.000000	66.000000	114.000000
75%	7.400000	2017.000000	67.000000	116.300000
max	9.500000	2025.000000	100.000000	317.000000

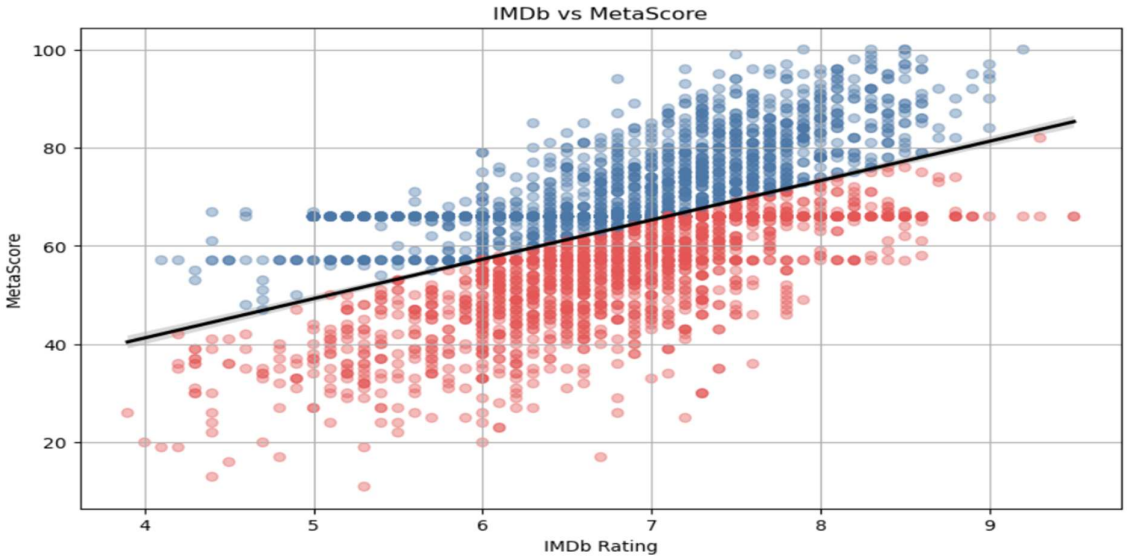
Correlation Heatmap – A Quick Way to Uncover Feature Relationships

- The correlation heatmap gives correlation factors for all features in the dataset, helping us understand underlying relations better.
- From the heatmap, we see that no two features have an absolute correlation factor greater than **0.55**, which is for **IMDb Rating** and **MetaScore**.



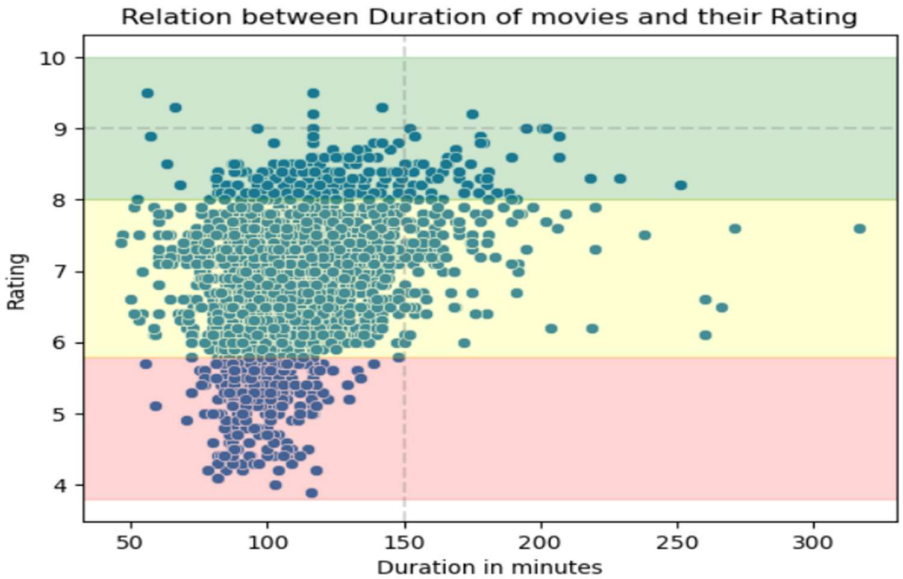
IMDb Rating and MetaScore – Understanding their relationship

- First, we must define these two concepts for deeper analysis.
- MetaScore is determined by professional critics and is fixed.
- Whereas IMDb Rating is determined by the viewers (general public) and varies.
- The regression plot helps us understand the correlation factor of 0.55 between them, clarifying that IMDb Rating and MetaScore **may have similar trends** but not always.



Movie Duration and IMDb Rating – Understanding their relationship

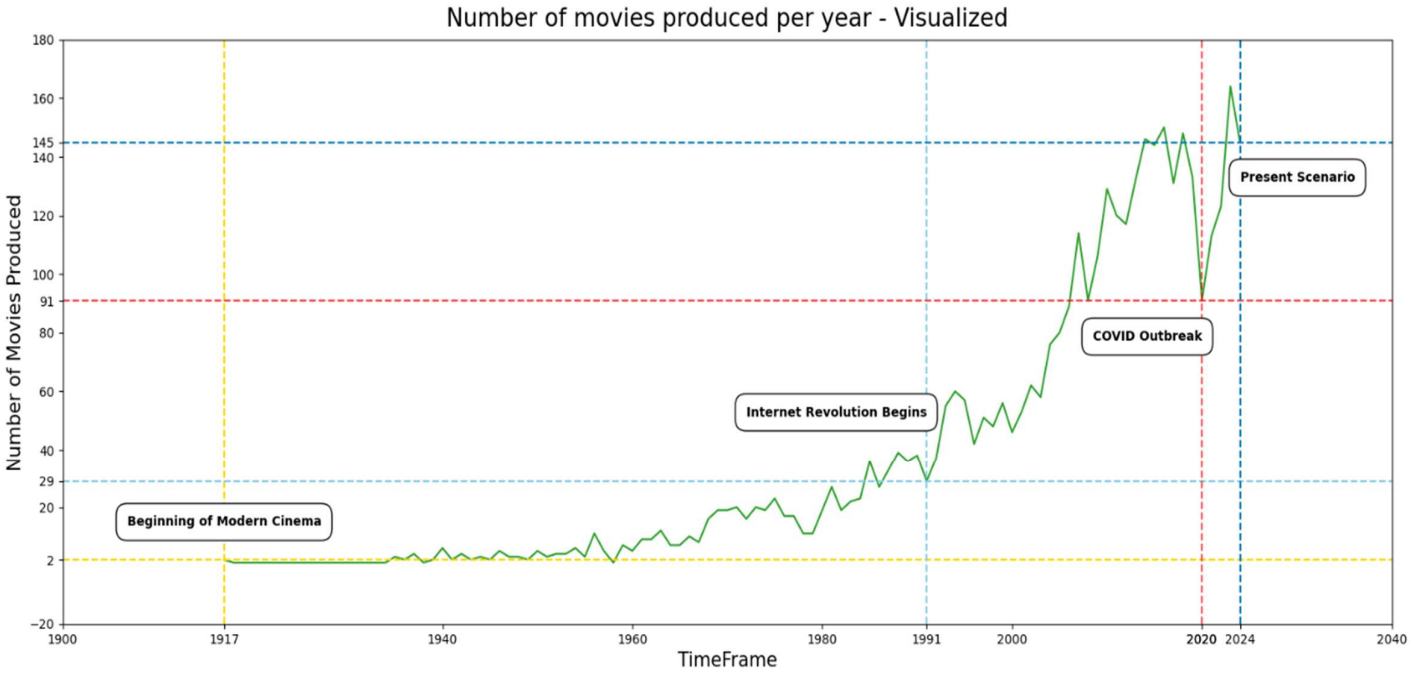
- ❖ We have seen that the correlation factor is **0.25**, indicating a **small effect** of Duration on the IMDb Rating.
- ❖ But from the graph, we can see that movies whose duration **exceeds 150 minutes** have a rating that's greater than or equal to **6**.
- ❖ But it's also important to note that the **highest rated** movies here have a duration of **less than 150 minutes.**



4. The Plot Unfolds: Visualizing Insights, Patterns & Trends (Analysis, Insights and Visualization)

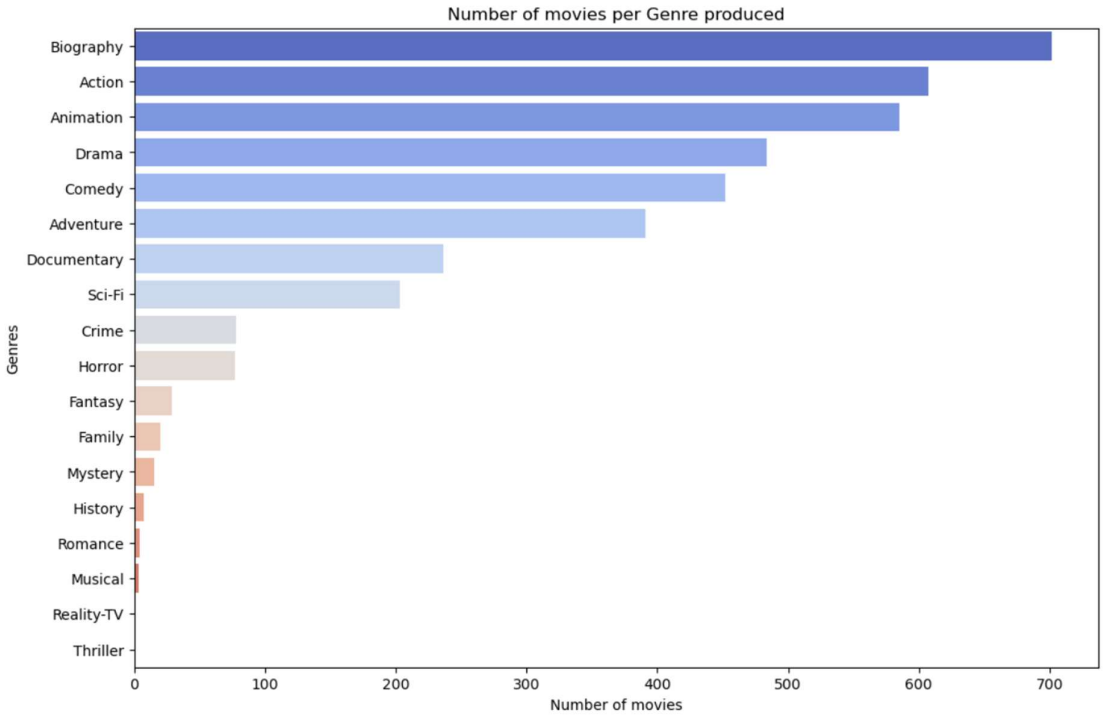
After going through the fundamental steps like Data Cleaning and Exploration, it is finally time to go through the **most interesting step: Analysis and advanced visualization.** Using Python libraries like Matplotlib and Seaborn helps in creating **aesthetically pleasing plots** which convey relevant information through variation in color palettes, etc. We'll follow a structured approach—first examining trends over the years before delving into specific insights.

4.1 Trends Over the Years – What trends have movies followed?



The plot clearly shows a **rising trend since 1917**, attributed to the increasing availability of resources, advancements in technology, and growing public demand. The slight **dip in 2020** is most likely due to coronavirus outbreak across the world.

Understanding the Production of Genres

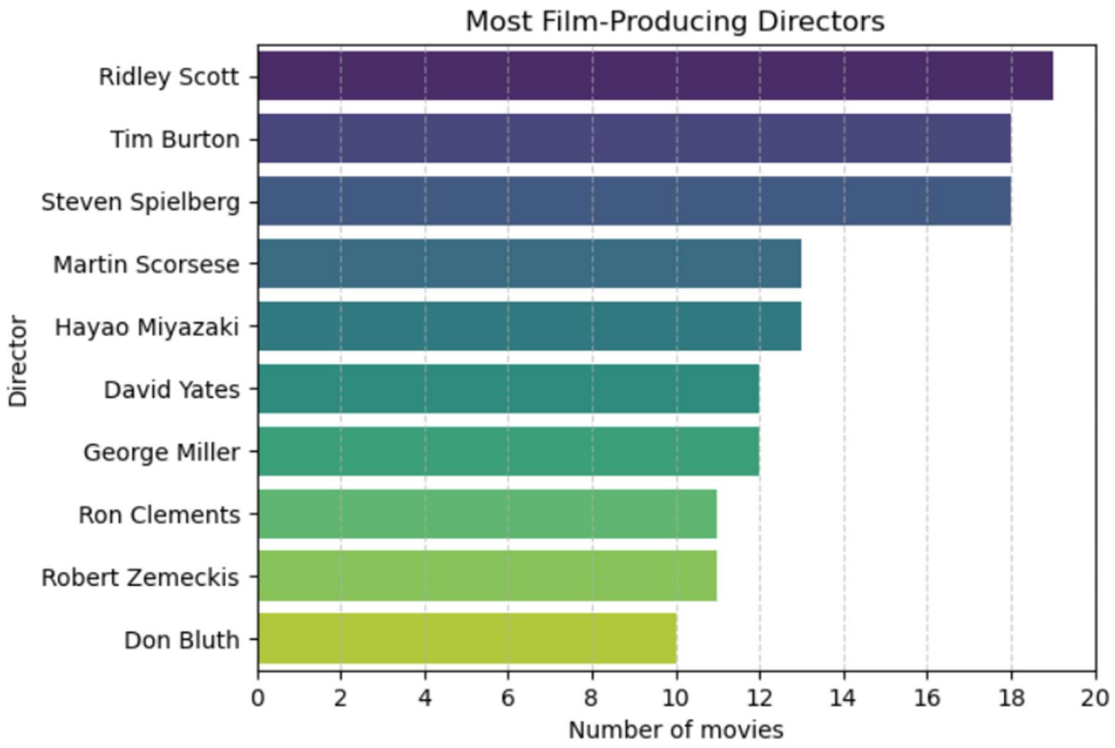


This plot gives an idea of how the **distribution of movies across Genres** really is. **Biographies** have had more than **700** movies in their category. The not-so-far runners up are **Action** and **Animation**, having around **600** movies each.

4.2 Directorial Trends: A Data Driven Comparison

Directors are the backbone of movies. The performance of movies primarily depends on how plots and scenes are structured by the directors. Thus, analyzing them is an essential step in uncovering trends.

For example, here we have the directors producing the **greatest number of movies**. All these directors have produced significant number of movies, including **Don Bluth (10)**, considering there are **2603 directors** in total.

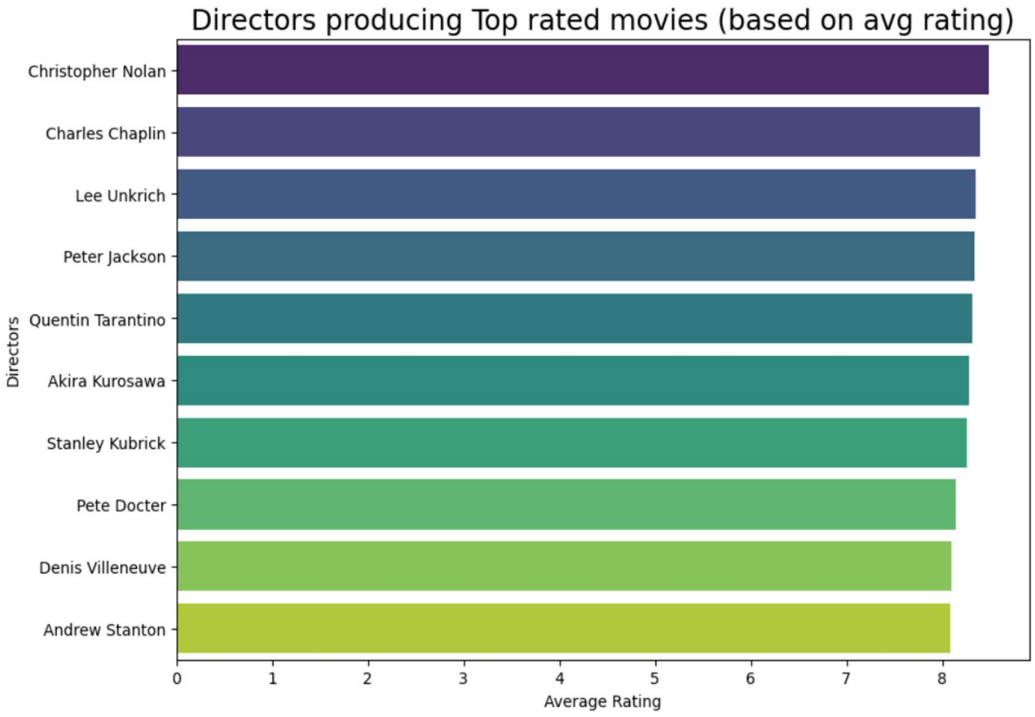


Directors Whose Films Are a Must-Watch for Viewers

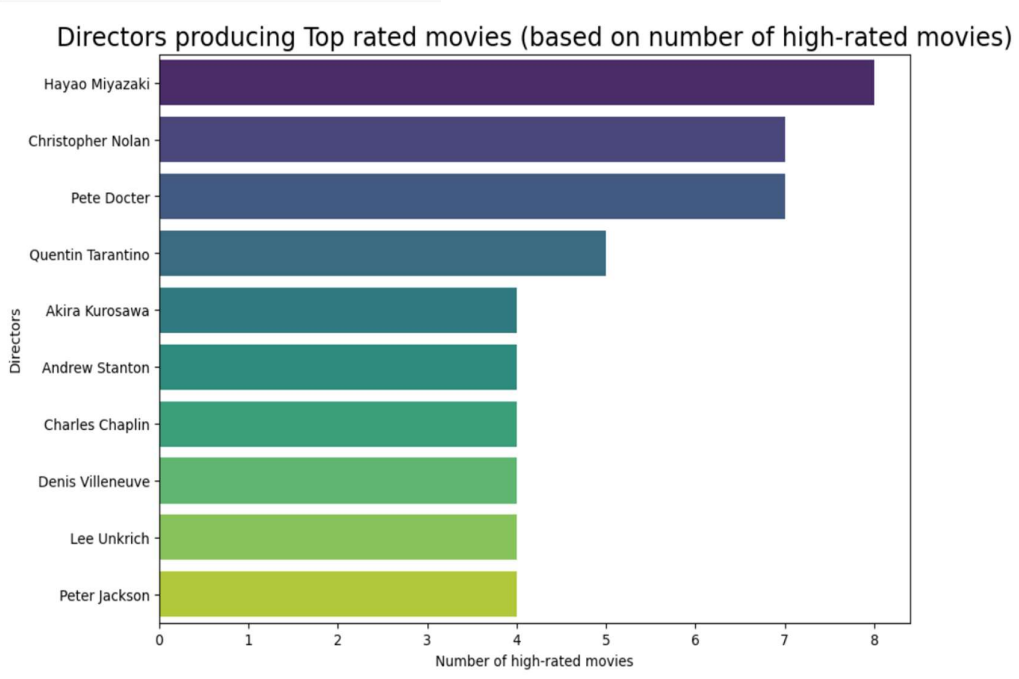
These directors are known for producing exceptional films that are loved by audiences worldwide. Here are two plots that will help you understand the impact they've had on the movie industry. The **thresholds** are as follows:

- ✓ High IMDb Rating: 8.0
- ✓ Minimum Movie count: 4
- ✓ Minimum number of High Rated movies: 3

The chart below highlights directors who've made **consistently well-regarded films**, with **Christopher Nolan's** work showing the highest average rating, followed by **Charles Chaplin and Lee Unkrich**, while **Andrew Stanton's** films, among this list, show the lowest average rating.

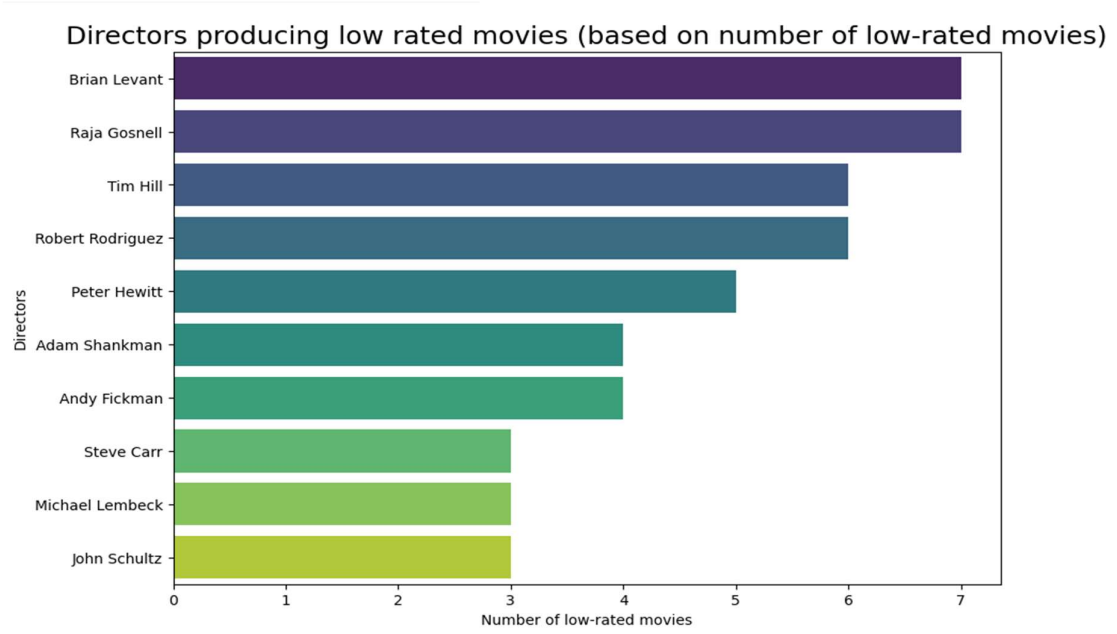


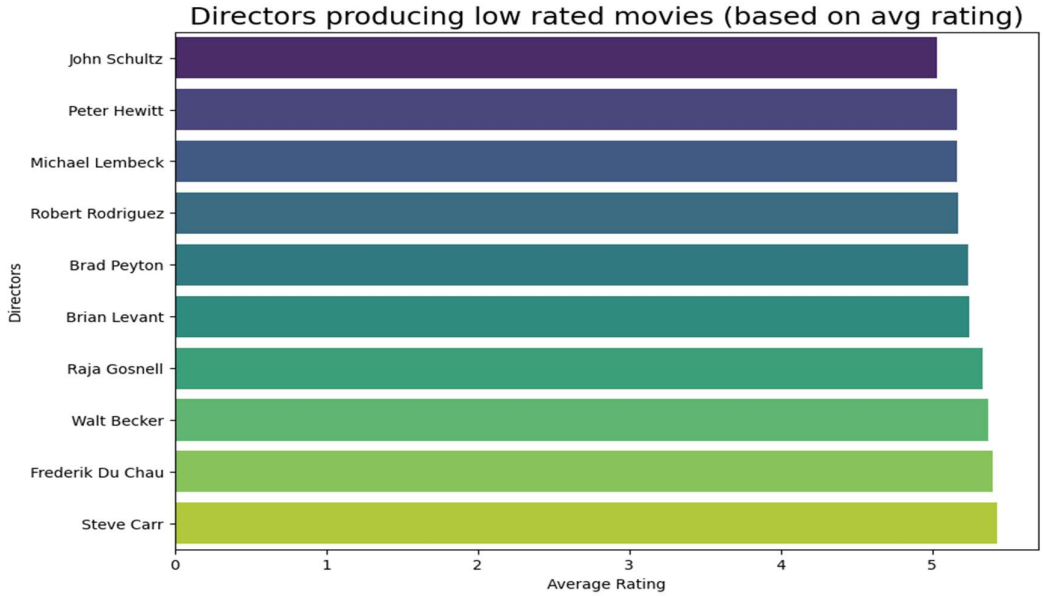
But, since Average rating can't be the sole criteria, we look at the number of successful films as well, where **Hayao Miyazaki**, the creator of **Studio Ghibli**, leads with **8 highly rated films**.



Directors Whose Films Are Not Worth Watching

This list features directors whose films haven't quite resonated with audiences or critics. Whether it's due to weak storytelling, poor execution, or uninspiring performances, their movies often miss the mark. While every filmmaker has their highs and lows, these directors have faced consistent challenges in creating standout cinema. Here are two graphs that display their poor performance:





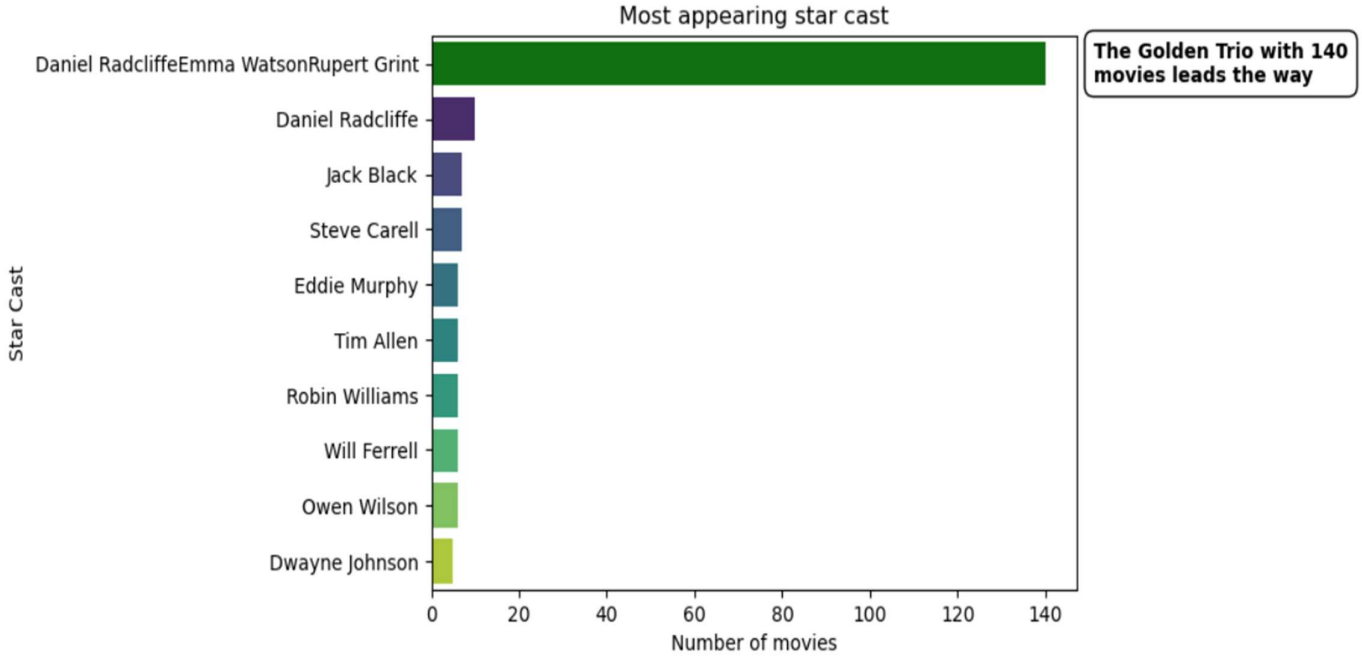
These movies have performed **very poorly** based on viewer ratings, with the lowest average rating going down to **5.033**, indicating that such movies are not worth watching.

All movies directed by **John Schultz**, **Peter Hewitt**, **Brian Levant** and **Raja Gosnell** and few others have performed very poorly. While John Schultz has the lowest average rating, Brian Levant and Raja Gosnell have the highest number of low rated movies.

4.3 Trends in Star Cast: Shaping Modern Cinema

Lately, who's in the cast can be just as important as the story itself. More and more, audiences are gravitating toward films that feature diverse talent or unexpected pairings, often driven by social media buzz or fresh faces in Hollywood. This trend is pushing filmmakers to think outside the box when casting, giving rise to unique combinations that breathe new life into familiar genres.

For instance, let's look at the Cast who've starred the greatest number of movies:

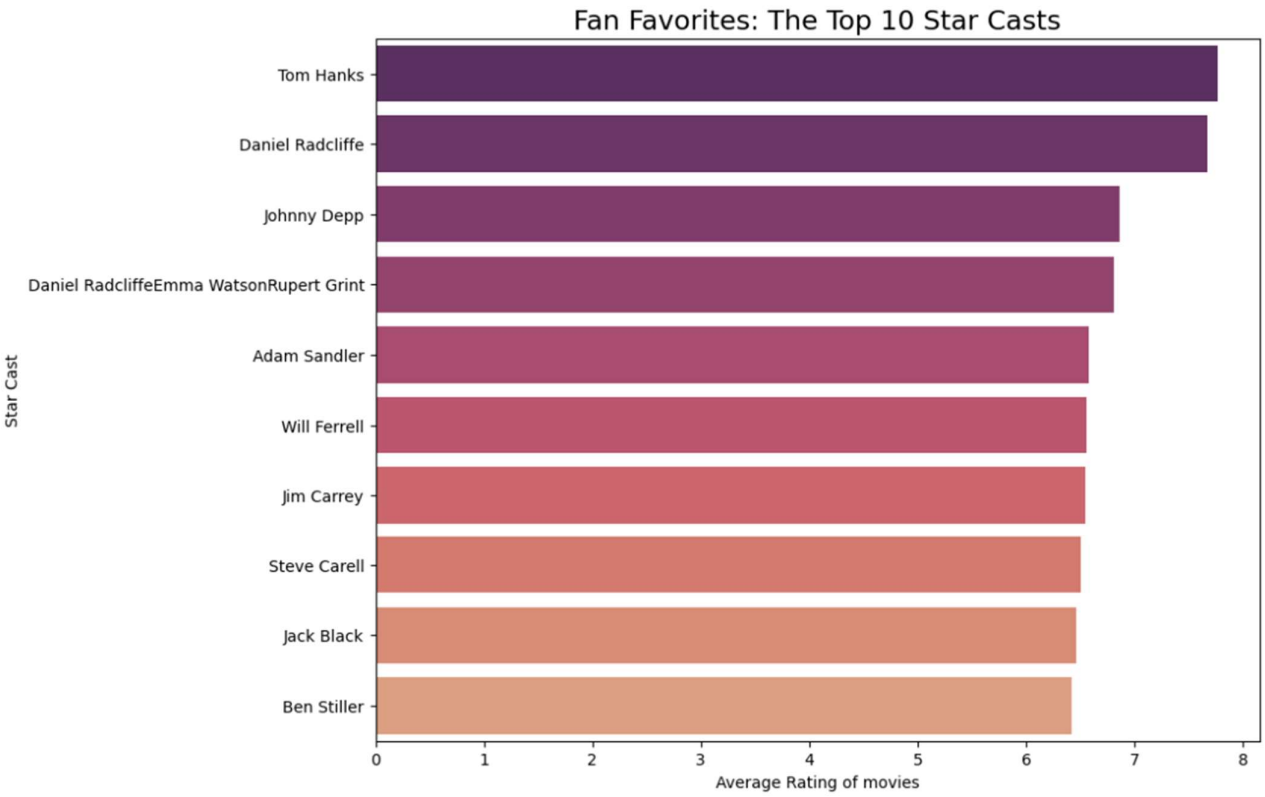


Here, we see that the Star Cast: Daniel Radcliffe, Emma Watson, and Rupert Grint, also called the Golden Trio top the list with 140 movies, far ahead than others.

Best and Worst performing Star Cast according to viewers:

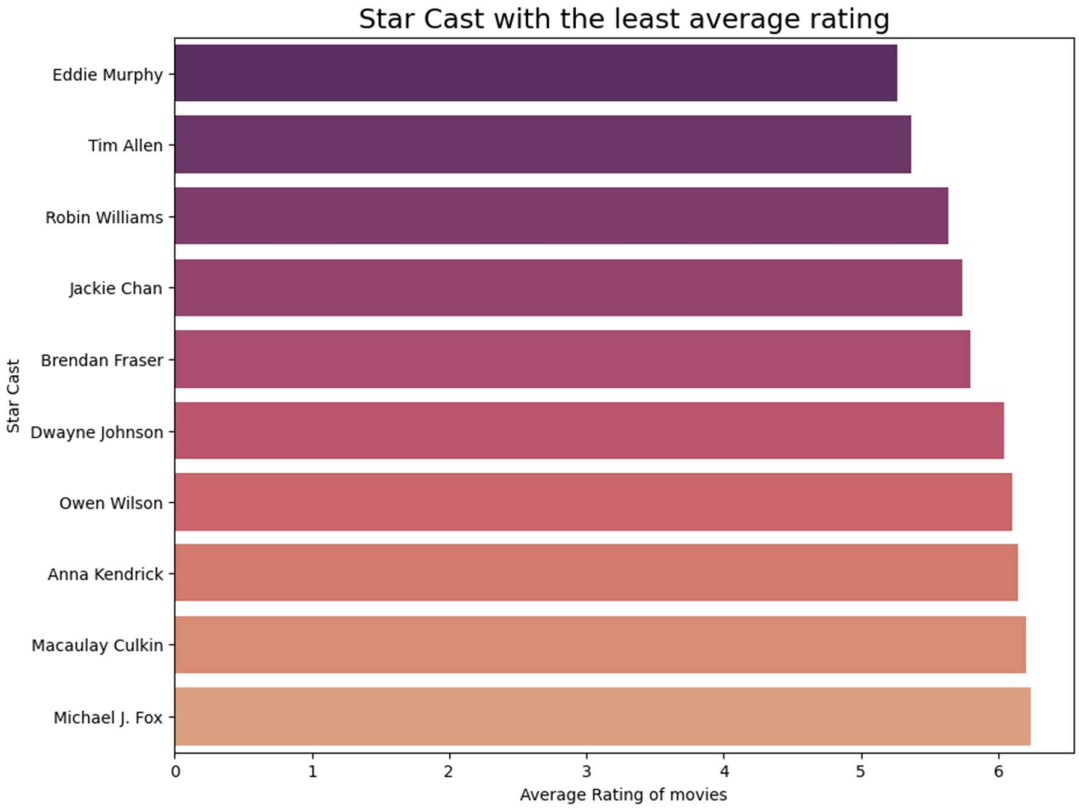
One of the main criteria for evaluating Star Cast performance is through the ratings of the movies starred by them.

The following graph shows the highest rated Star Cast, considering a minimum of 5 movies starred by them.



We see that Tom Hanks tops this list with an average rating of 7.76, followed closely by Daniel Radcliffe with 7.67. Daniel Radcliffe, Emma Watson, Rupert Grint together have performed really well with an average rating of 6.8, especially considering that they have starred 140 movies.

The following graph shows the least rated Star Cast, considering a minimum of 5 movies starred by them.

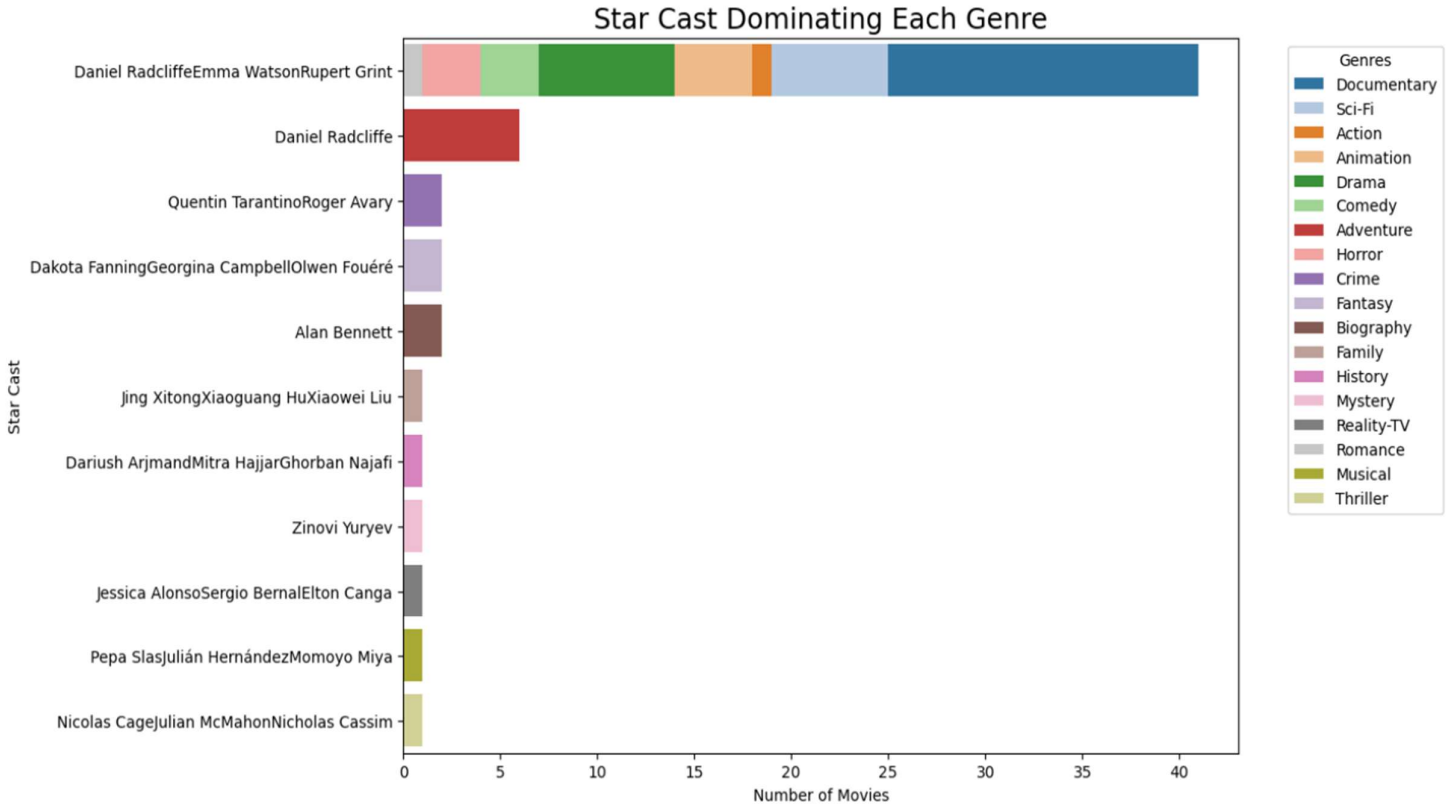


We can infer from this plot that on average, Eddie Murphy has starred the least rated movies on average (we are considering a minimum of 5 movies to get these results), followed closely by Tim Allen. Popular figures like Dwayne Johnson and Jackie Chan have also performed poorly according to this dataset.

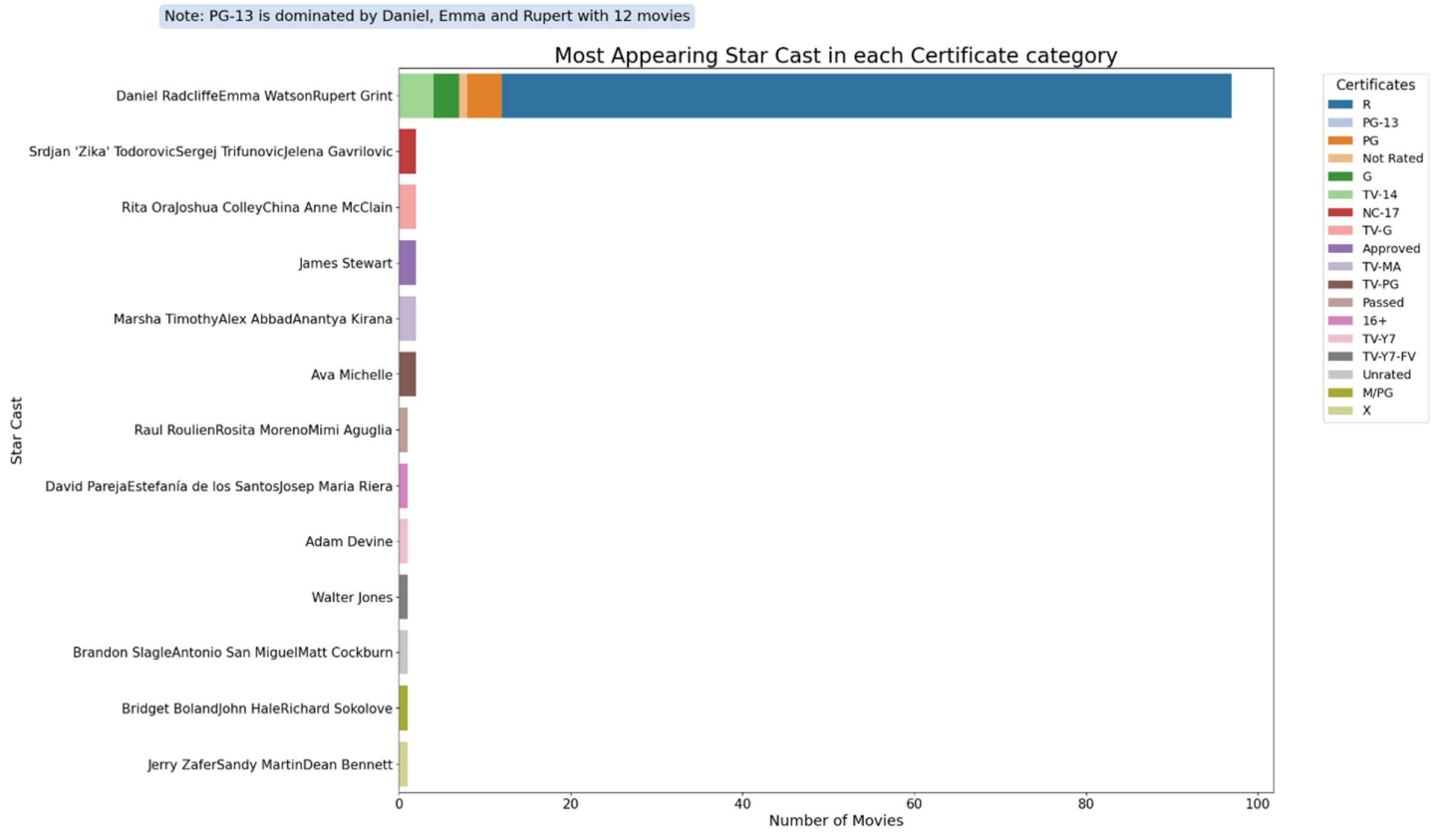
Domination of Star Cast in Genres and Certificate categories

Genre-Wise Domination:

The star cast of **Daniel Radcliffe, Emma Watson, and Rupert Grint** has made a significant impact across a variety of genres, dominating eight of them, with a particular presence in documentaries and sci-fi. This trio is the only one consistently featured in this list, underscoring their widespread influence and continued prominence in the industry.



Certificate Category-Wise Domination: We see that the same trio dominates here as well. This shows how well rounded their presence is, in different Genres and Certificate categories.

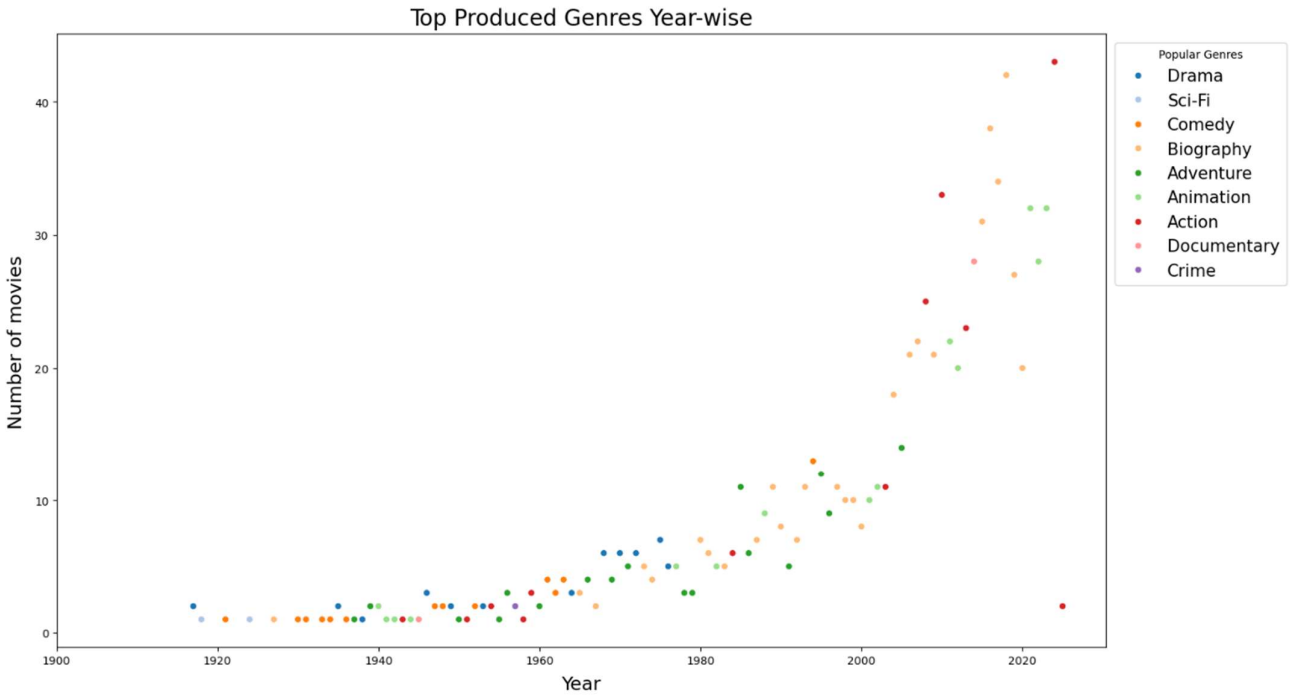


4.4 Uncovering Key Trends Across Different Genres

Here, we're going to be looking at some of the important trends across various Genres. We will also be focusing on five interesting and data-backed Genres: **Biography, Action, Animation, Sci-Fi and Comedy.**

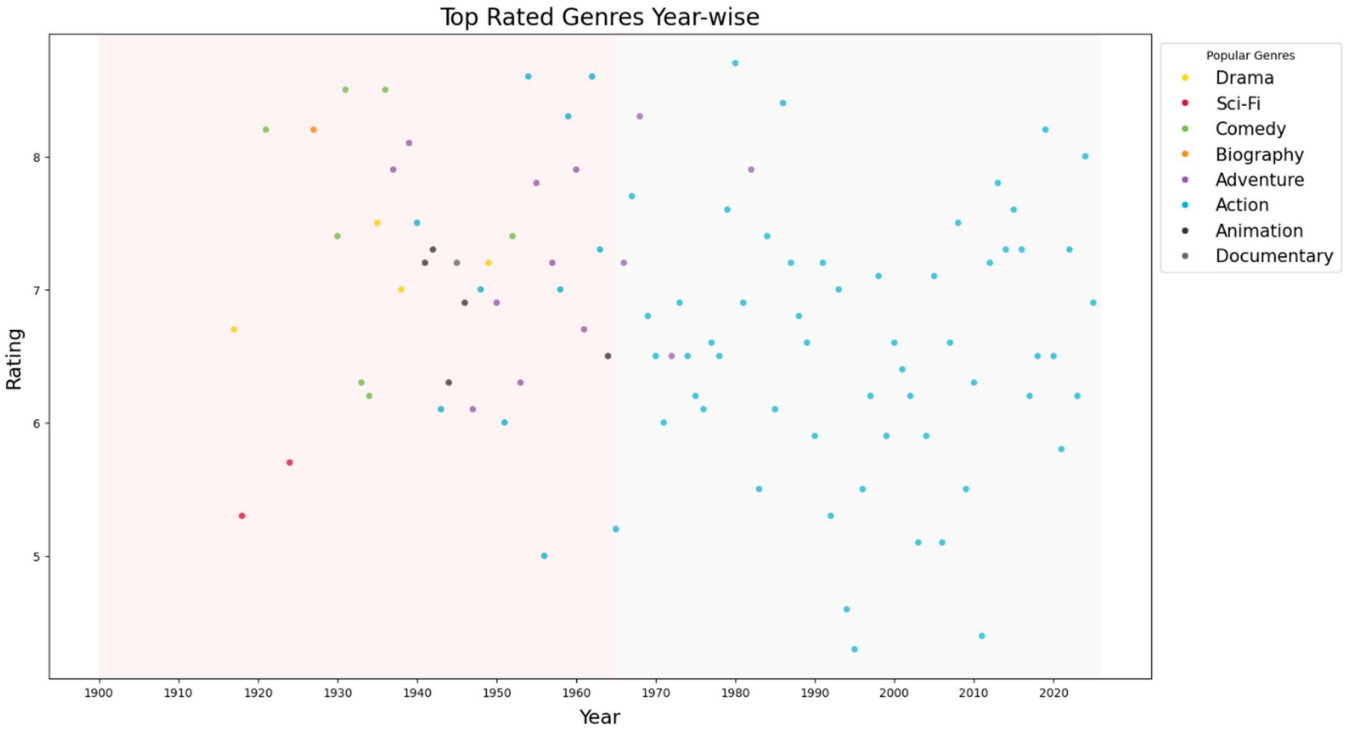
4.4.1 Trends Over the Years

First, let's look at the top produced Genres year wise in the period 1917 to 2025.



The plot indicates increasing diversity in film genres over the years. However, there is a notable rise seen in genres like Comedy, Biography, Action, Adventure, and Animation, particularly in the later part of the timeline.

Now, it’s also important to see which Genre dominates in terms of rating each year. Thus, we have another scatter plot demonstrating it:

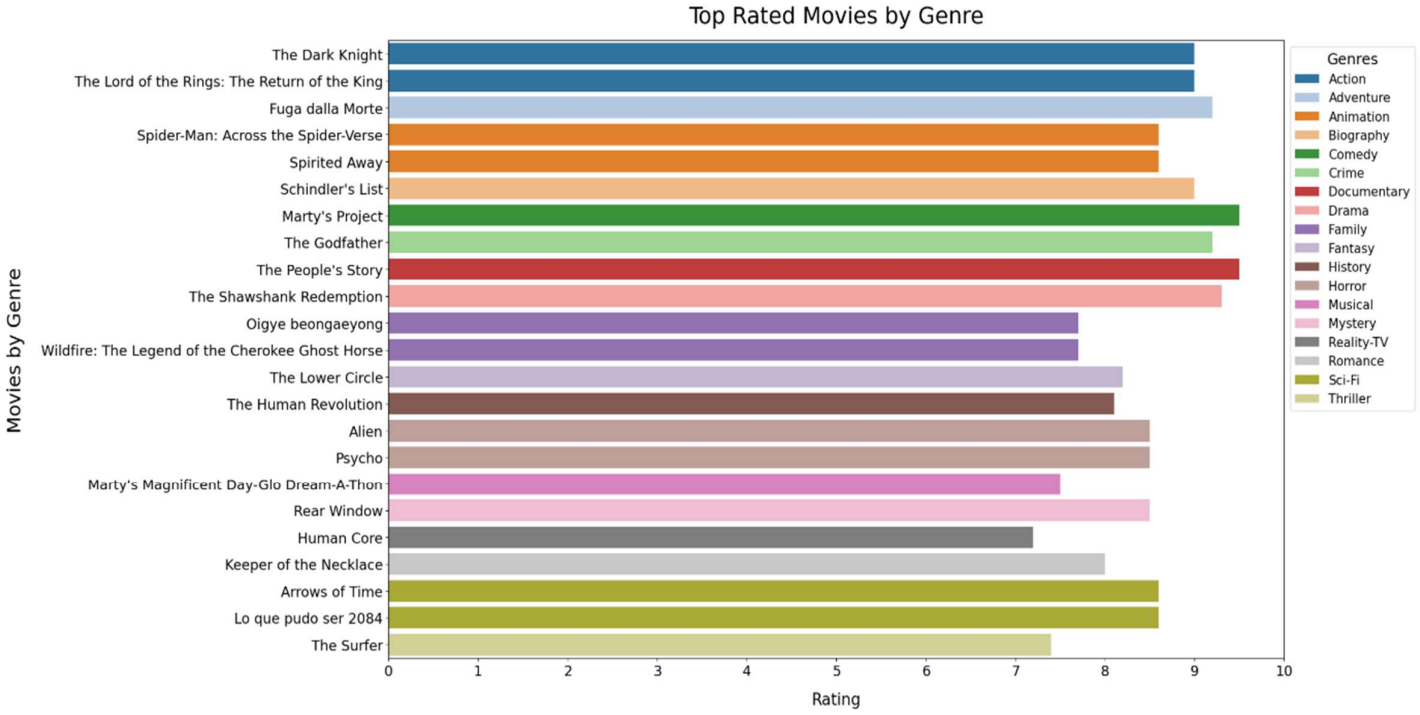


We can infer that before 1965, the Genre with the highest rating wasn't definite. But, after 1965, Action movies have dominated almost every year. The coloured regions help us understand that.

Finding the Best Movies in each Genre

Analysing trends in user ratings allows for the identification of films that are particularly well-received within their respective genres.

A bar graph with a vibrant colour palette helps us achieve just that:



Which Genres were rated the highest in 20th vs 21st Century?

20th Century

Genre	avg_rating	movie_count
Crime	7.500000	30
Documentary	7.437037	27
History	7.250000	2
Biography	7.180488	205
Animation	7.067200	125
Drama	7.005028	179
Mystery	6.950000	4
Musical	6.900000	1
Horror	6.845455	11
Action	6.659184	147
Adventure	6.644048	168
Family	6.490909	11
Comedy	6.457647	170
Romance	6.400000	2
Fantasy	6.183333	6
Sci-Fi	6.077273	44

VS

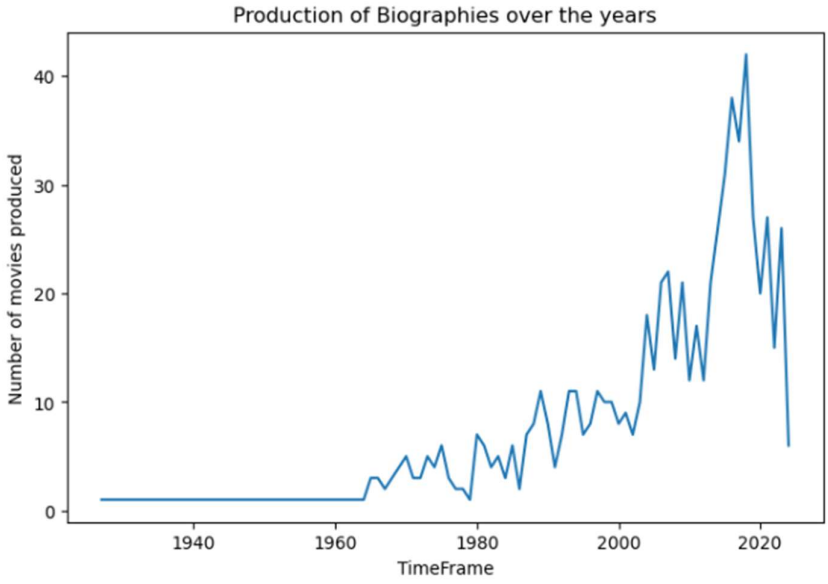
21st Century

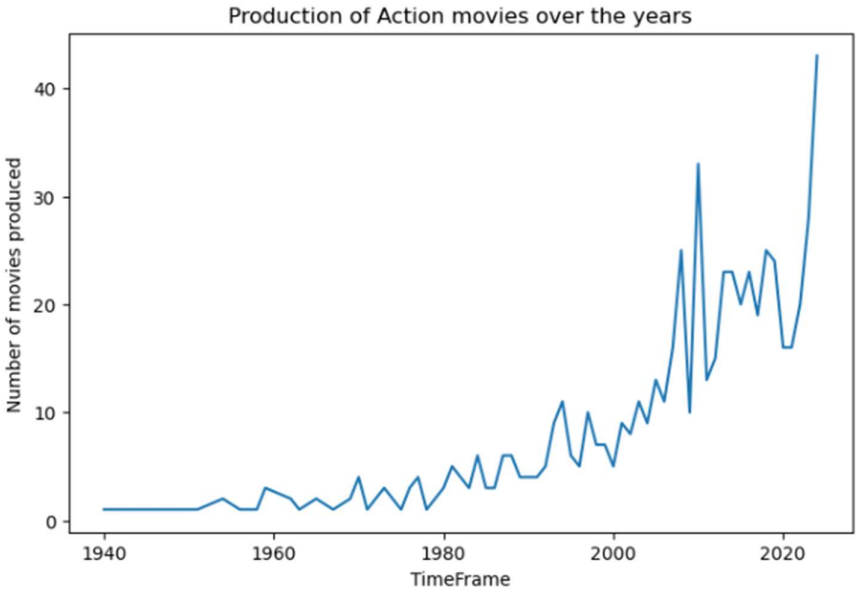
Genre	avg_rating	movie_count
Documentary	7.409091	209
Thriller	7.400000	1
Reality-TV	7.200000	1
Romance	7.000000	2
Biography	6.955131	497
History	6.820000	5
Mystery	6.818182	11
Crime	6.808333	48
Animation	6.795435	460
Drama	6.786885	305
Sci-Fi	6.659748	159
Action	6.649783	460
Fantasy	6.634783	23
Musical	6.600000	2
Adventure	6.438117	223
Comedy	6.368085	282
Horror	6.328788	66
Family	5.811111	9

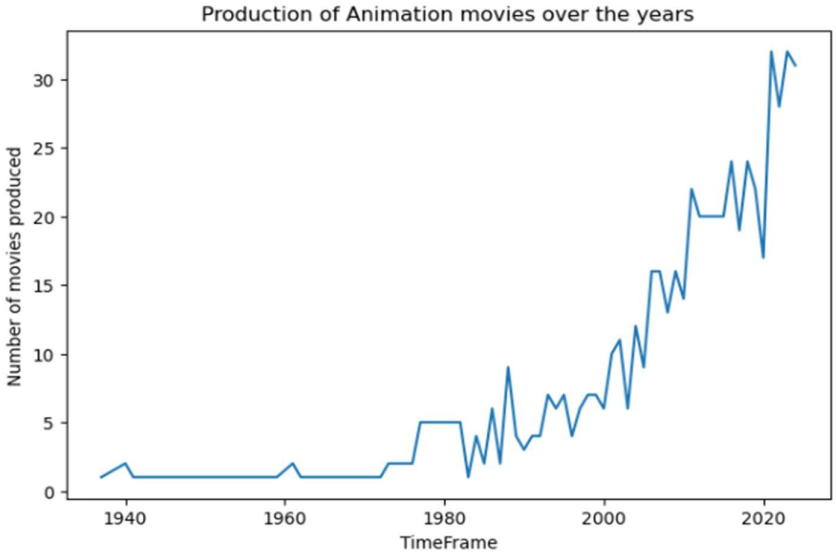
4.4.2 Five to Watch: A Closer Look at Biographies, Action, Animation, Comedy, and Sci-Fi

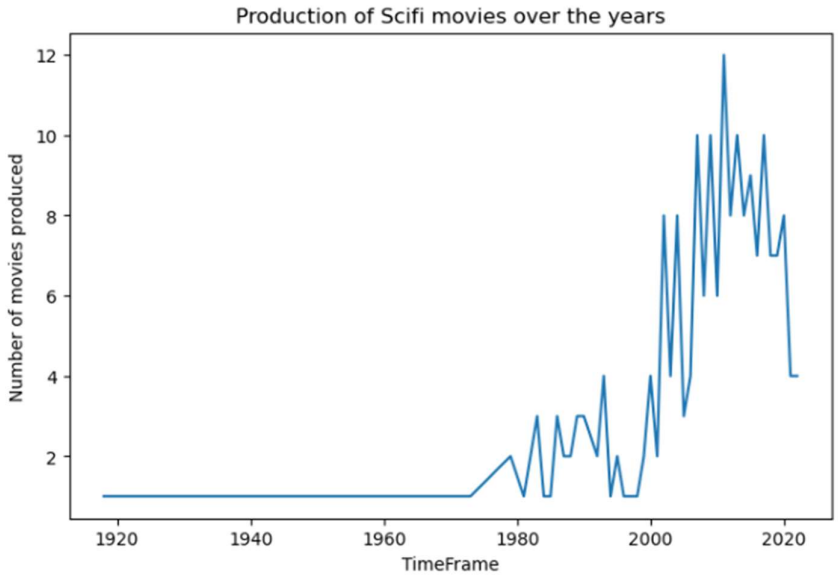
Trends of Production over the Years

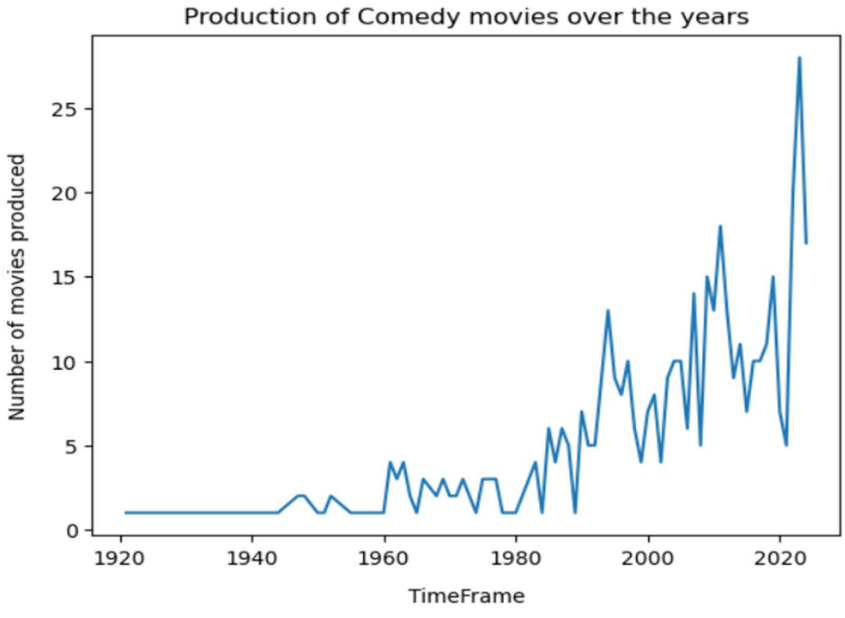
First, we'll look at the evolution of these Genres over the years, exemplifying how they've made their mark in the movie industry.





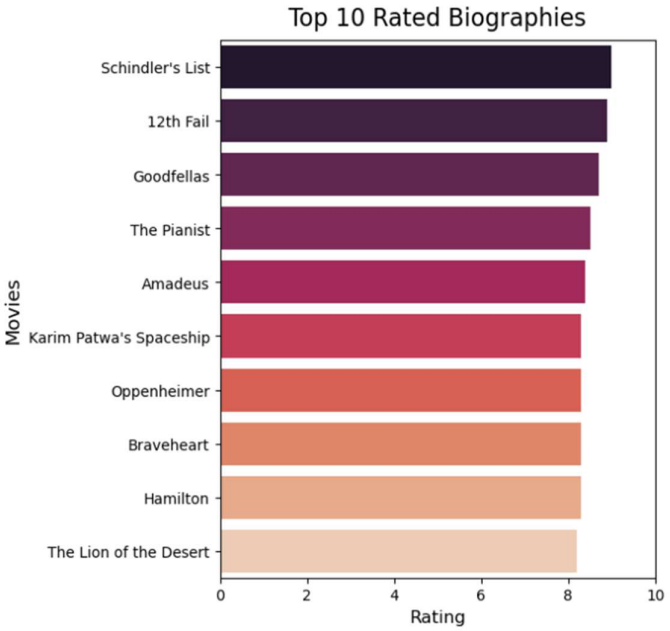

Production of Animation movies over the years
Number of movies produced
TimeFrame
0
5
10
15
20
25
30
1940
1960
1980
2000
2020

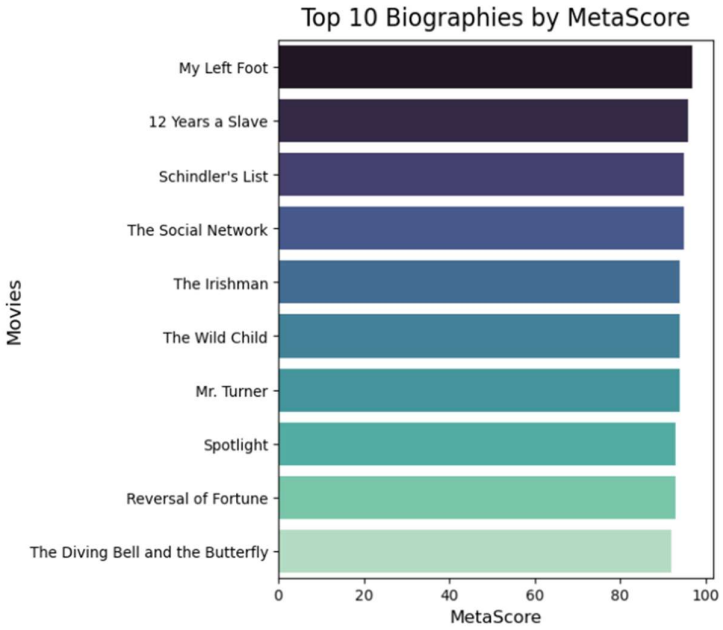

Production of Scifi movies over the years
Number of movies produced
TimeFrame
2
4
6
8
10
12
1920
1940
1960
1980
2000
2020

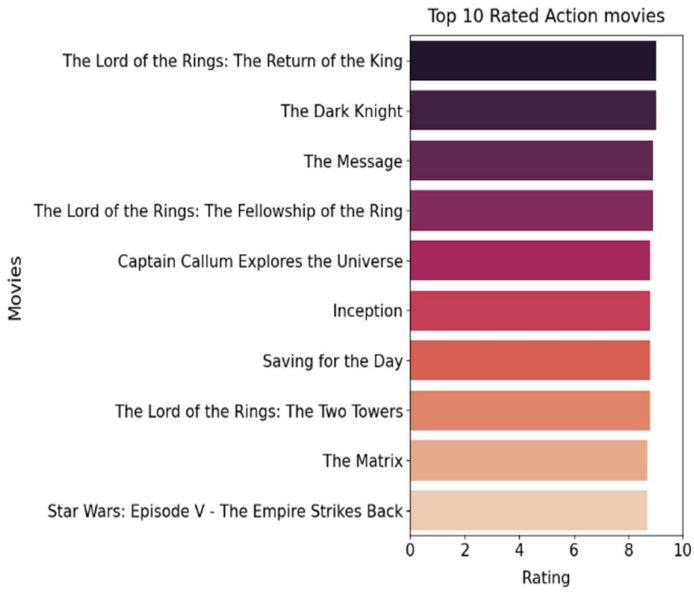

Production of Comedy movies over the years
Number of movies produced
TimeFrame
0
5
10
15
20
25
1920
1940
1960
1980
2000
2020

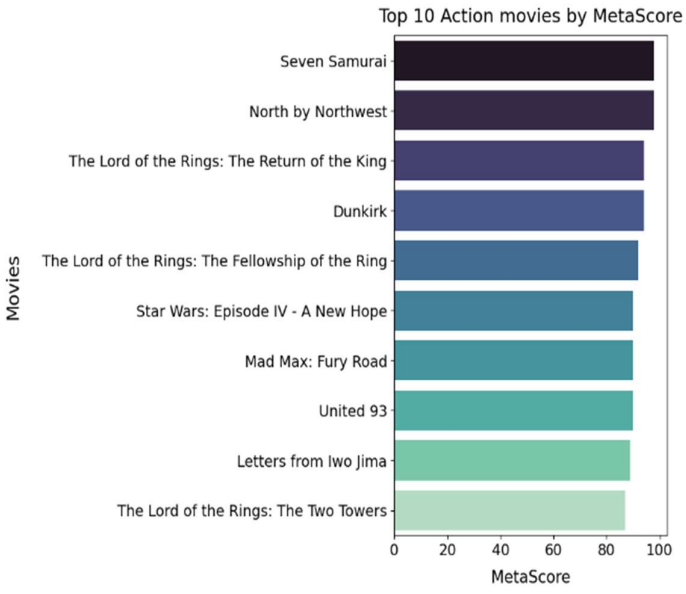
It is clear from the time series plots that the production of Biographies and Sci-Fi movies shows a dynamic trend, while that of Animation, Comedy, and Action movies shows a strong upward trend.

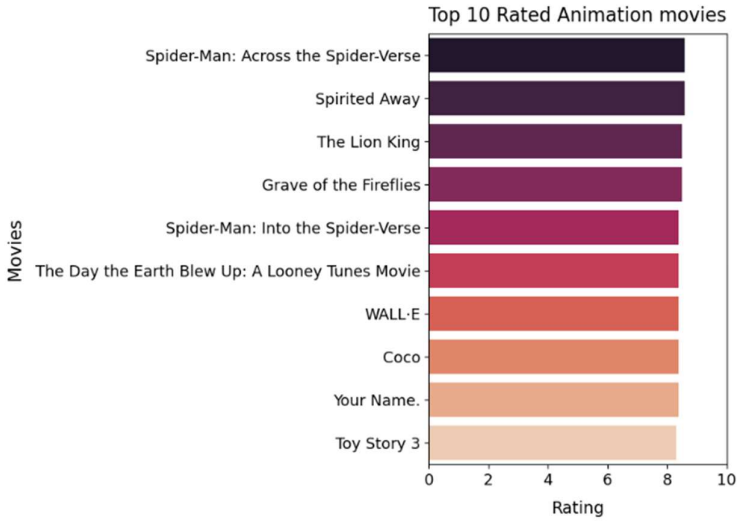
4.4.3 Finding the Best Movies per Genre

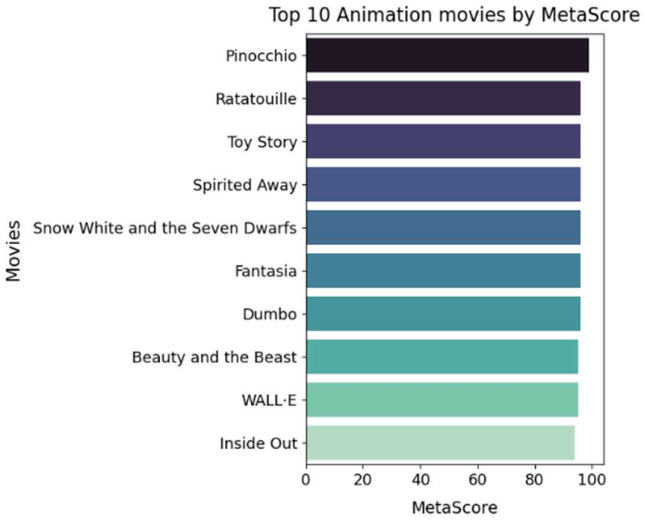


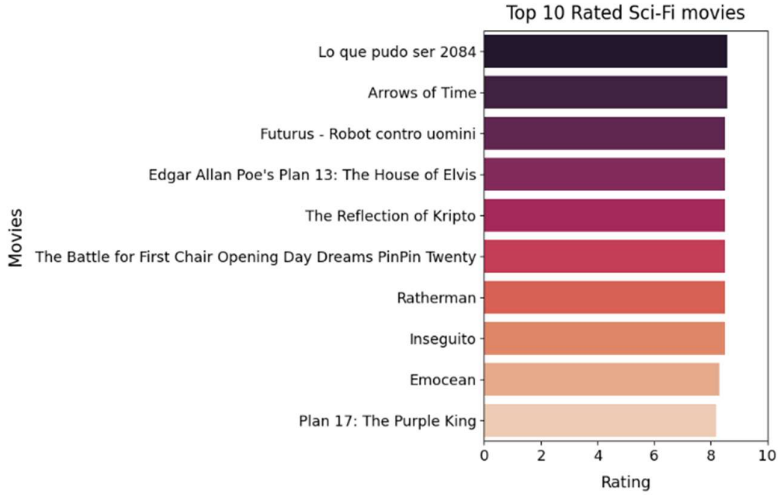


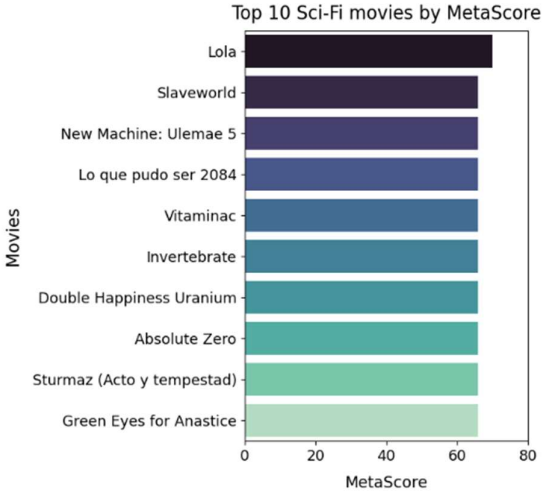


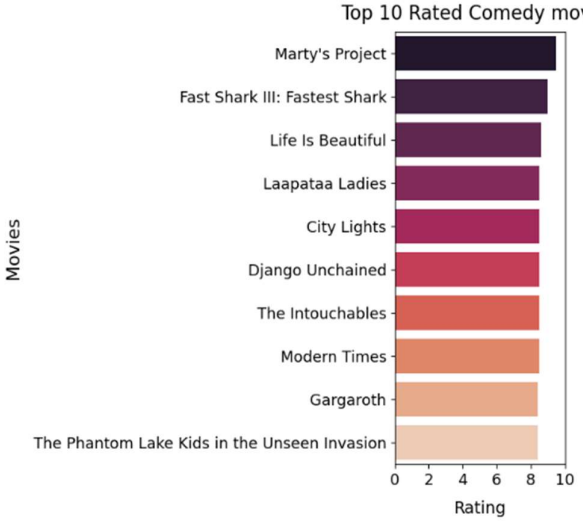


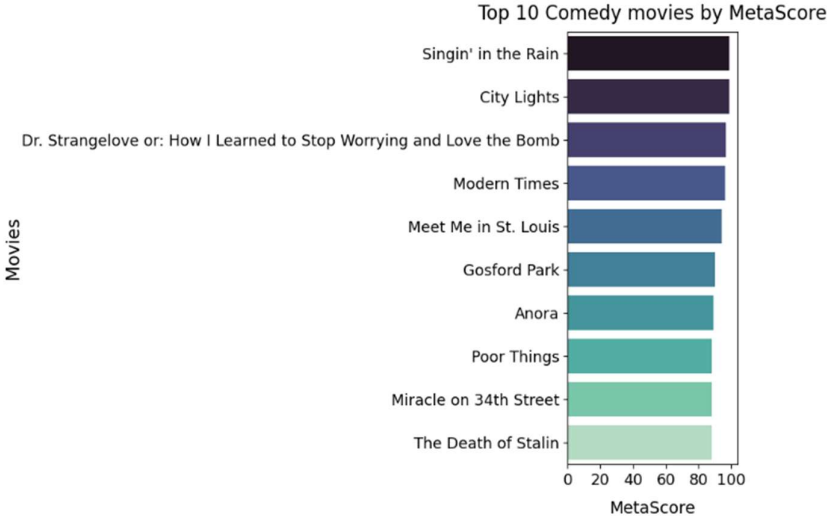

Top 10 Rated Animation movies
Spider-Man: Across the Spider-Verse
Spirited Away
The Lion King
Grave of the Fireflies
Spider-Man: Into the Spider-Verse
The Day the Earth Blew Up: A Looney Tunes Movie
WALL·E
Coco
Your Name.
Toy Story 3
Movies
0 2 4 6 8 10
Rating


Top 10 Animation movies by MetaScore
Pinocchio
Ratatouille
Toy Story
Spirited Away
Snow White and the Seven Dwarfs
Fantasia
Dumbo
Beauty and the Beast
WALL·E
Inside Out
Movies
0 20 40 60 80 100
MetaScore


Top 10 Rated Sci-Fi movies
Lo que pudo ser 2084
Arrows of Time
Futurus - Robot contro uomini
Edgar Allan Poe's Plan 13: The House of Elvis
The Reflection of Kripto
The Battle for First Chair Opening Day Dreams PinPin Twenty
Ratherman
Inseguito
Emocean
Plan 17: The Purple King
Movies
0 2 4 6 8 10
Rating


Top 10 Sci-Fi movies by MetaScore
Lola
Slaveworld
New Machine: Ulemae 5
Lo que pudo ser 2084
Vitaminac
Invertebrate
Double Happiness Uranium
Absolute Zero
Sturmaz (Acto y tempestad)
Green Eyes for Anastice
Movies
0 20 40 60 80
MetaScore


Top 10 Rated Comedy movies
Marty's Project
Fast Shark III: Fastest Shark
Life Is Beautiful
Laapataa Ladies
City Lights
Django Unchained
The Intouchables
Modern Times
Gargaroth
The Phantom Lake Kids in the Unseen Invasion
Movies
0 2 4 6 8 10
Rating


Top 10 Comedy movies by MetaScore
Singin' in the Rain
City Lights
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb
Modern Times
Meet Me in St. Louis
Gosford Park
Anora
Poor Things
Miracle on 34th Street
The Death of Stalin
Movies
0 20 40 60 80 100
MetaScore

4.5 The Story of Movie Certificates

Here, we'll be analysing the trends in movie certificate categories like **R, G, PG, PG-13**, and others, which will help significantly in the analysis of movies.

Which Certificate Categories were rated the highest in 20th vs 21st Century?

20th Century

Certificates	avg_rating	movie_count
Not Rated	7.800000	12
M/PG	7.400000	1
TV-14	7.333333	3
Approved	7.255102	49
Passed	7.250000	2
G	7.101190	168
TV-PG	7.000000	1
PG-13	6.897917	96
Unrated	6.766667	3
PG	6.735569	343
R	6.732517	449
NC-17	6.600000	4
X	6.100000	1

VS

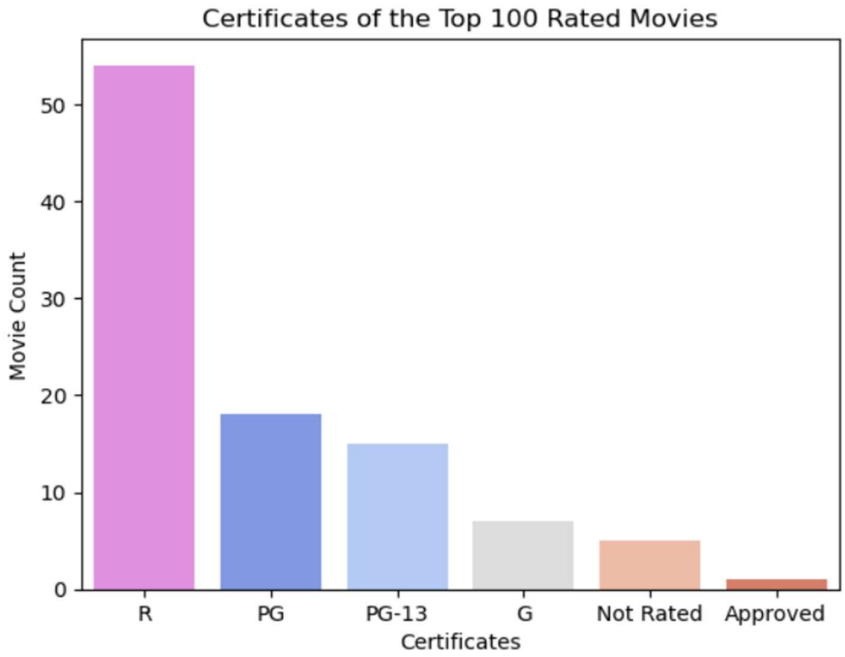
21st Century

Certificates	avg_rating	movie_count
PG-13	6.967723	505
R	6.817689	1255
G	6.812088	91
16+	6.800000	1
TV-14	6.623077	13
Not Rated	6.575728	103
PG	6.556557	732
NC-17	6.500000	5
Unrated	6.500000	5
TV-G	6.200000	11
TV-Y7	6.100000	8
TV-PG	6.048148	27
TV-MA	5.883333	6
TV-Y7-FV	5.700000	1

What is the Certification of the Top 100 Movies?

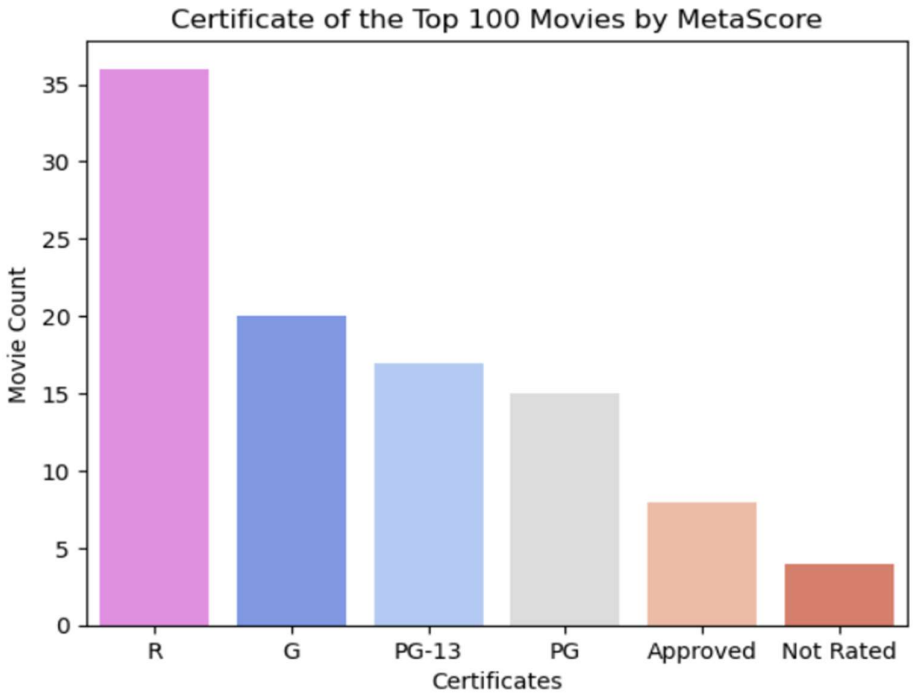
1. Based on IMDb Rating:

Certificates	count
R	54
PG	18
PG-13	15
G	7
Not Rated	5
Approved	1



2. Based on MetaScore:

Certificates	count
R	36
G	20
PG-13	17
PG	15
Approved	8
Not Rated	4



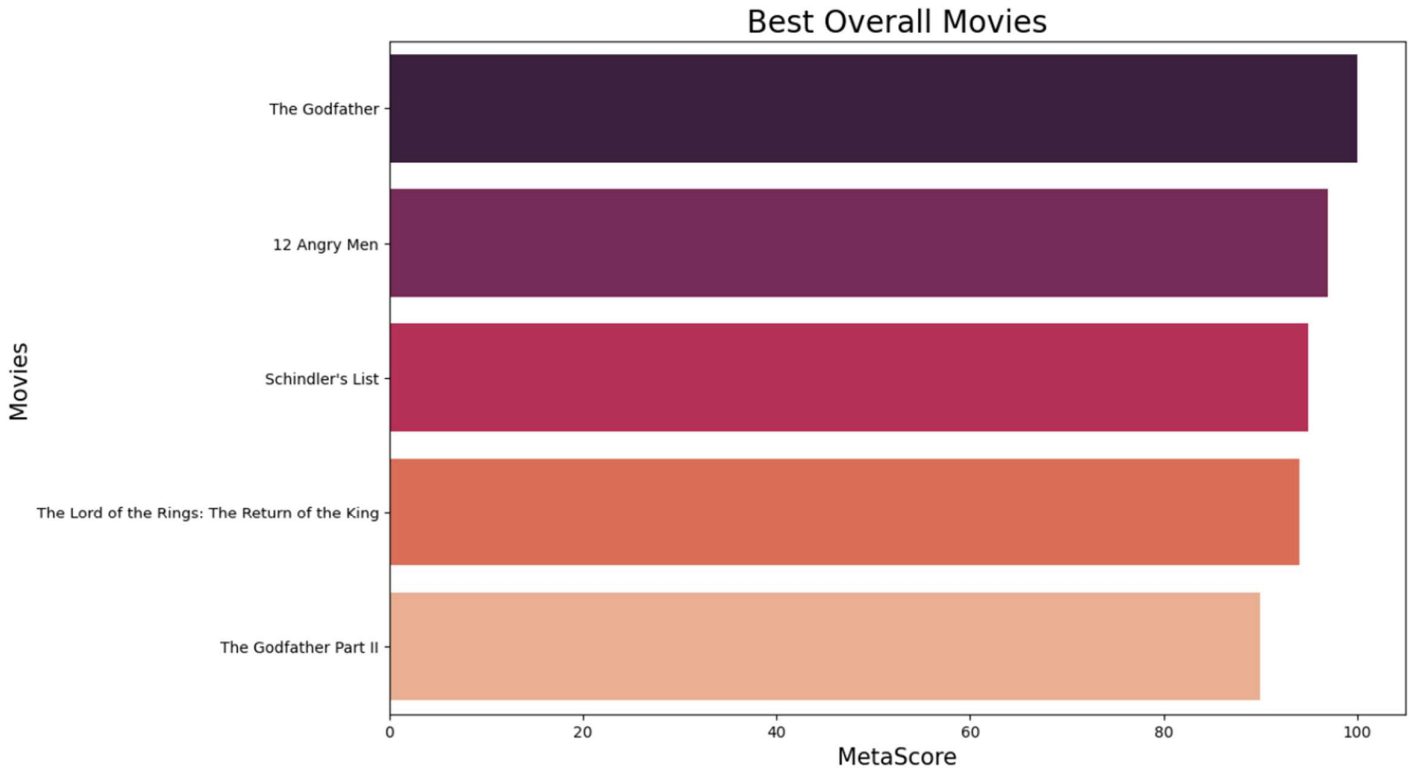
- ❖ It is very clear from the tabulations and bar plots that top movies, whether by IMDb Rating or MetaScore, tend to be R rated.
- ❖ They are followed by G, PG, and PG-13 differing by small amounts.

4.6 Identifying the Pinnacle and the Pit: A Combined Viewer & Critic Perspective

Overall Best & Worst:

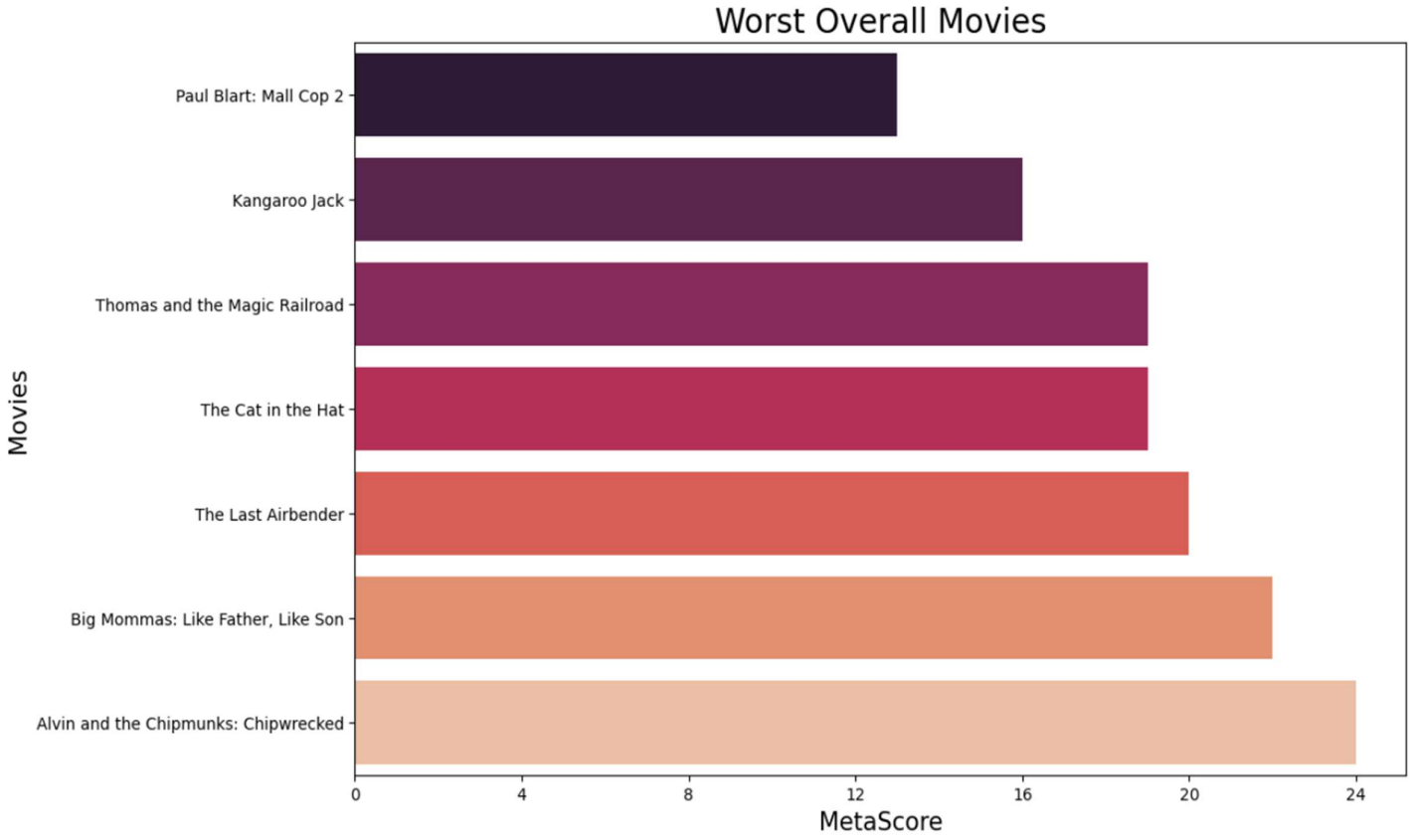
Having looked at various trends, it is now time to look at the best and worst movies, considering contributions from IMDb Rating as well as MetaScore:

Best Overall Movies:



- This list considers the best movies, having the **highest thresholds** in both fields, IMDb Rating and MetaScore (9 and 90 respectively). The movies that come in this list are by definition, exceptionally well directed and executed.
- The crime blockbuster, **The Godfather** tops the list with a perfect MetaScore of **100**, with a rating of **9.2**. Even its sequel, The Godfather Part 2 has performed really well.
- The other three movies have a rating of **9 each**, with MetaScore **>=94**, indicating **strong overall performance**.

Worst Overall Movies:



- This list shows the worst overall movies, considering both Rating and MetaScore fields, making it clear that such movies were liked by neither the audience nor the critical analysts.

- Paul Blart: Mall Cop 2 has performed the worst out of all, with a terrible MetaScore of just 13 and IMDb Rating of 4.
- Also, the Genres for these movies are surprisingly Action and Adventure, implying that the movies were very bad inherently.
- This is still not the worst individually, where the lowest rating is 3.9 for Madame Web in terms of IMDb Rating, and 11 for Nine Lives in terms of MetaScore.

Best & Worst According to IMDb Rating

Best Movie(s):

Title	IMDb Rating	Year	Certificates	Genre	Director	Star Cast	MetaScore	Duration (minutes)
Marty's Project	9.5	2010	R	Comedy	Andrew Kent	Naomi AshDavid BedfordJoanna Deering	66.0	116.3
The People's Story	9.5	2000	PG	Documentary	Steven Scaffidi	Daniel RadcliffeEmma WatsonRupert Grint	66.0	56.0

Worst Movie(s):

Title	IMDb Rating	Year	Certificates	Genre	Director	Star Cast	MetaScore	Duration (minutes)
Madame Web	3.9	2024	PG-13	Action	S.J. Clarkson	Matt SazamaBurk SharplessClaire Parker	26.0	116.0

Best & Worst According to MetaScore

Best Movie(s):

Title	IMDb Rating	Year	Certificates	Genre	Director	Star Cast	MetaScore	Duration (minutes)
The Godfather	9.2	1972	R	Crime	Francis Ford Coppola	Mario PuzoFrancis Ford Coppola	100.0	175.0
Lawrence of Arabia	8.3	1962	PG	Adventure	David Lean	Peter O'TooleAlec GuinnessAnthony Quinn	100.0	218.0
The Leopard	7.9	1963	PG	Drama	Luchino Visconti	Giuseppe Tomasi di LampedusaSuso Cecchi D'Amic...	100.0	186.0
Casablanca	8.5	1942	PG	Drama	Michael Curtiz	Julius J. EpsteinPhilip G. EpsteinHoward Koch	100.0	102.0
Rear Window	8.5	1954	PG	Mystery	Alfred Hitchcock	James StewartGrace KellyWendell Corey	100.0	112.0

Worst Movie(s):

Title	IMDb Rating	Year	Certificates	Genre	Director	Star Cast	MetaScore	Duration (minutes)
Nine Lives	5.3	2016	PG	Comedy	Barry Sonnenfeld	Kevin Spacey	11.0	87.0

5. Key Takeaways and Conclusion

Based on our analysis of the IMDB movie dataset, the following conclusions can be drawn:

- The analysis reveals a diversification of genres and an increase in overall movie production volume over the years, with genres like Comedy, Biography, Action, and Animation showing substantial growth, particularly in the later part of the timeline.
- IMDb ratings and MetaScores exhibit a general positive correlation, suggesting a degree of alignment between audience and critical reception, although discrepancies exist.
- Certain directors, such as Christopher Nolan, Charles Chaplin and Lee Unkrich have consistently produced well-regarded films, while others have faced challenges in achieving consistent critical and popular success.
- Star cast analysis indicates that certain actors and actor combinations, such as Daniel Radcliffe, Emma Watson, and Rupert Grint, have achieved significant prominence and success, impacting various genres and certificate categories.

These findings provide insights into genre trends, the relationship between critical and audience reception, director performance, and the influence of star casts within the movie industry.