MONASH BUSINESS SCHOOL

ETC5513 Assignment 4: IMDB Data Analysis

Mr Will Nguyen
Master of Actuarial Science

Ms Wang Xue
Master of Business

Mr Rahul Bharadwaj Mysore Venkatesh
Master of Business Analytics

Mr Aryan Jain
Master of Business Analytics

Report for
Monash University

12 June 2020

Department of Econometrics & Business Statistics

(03) 9905 2478
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

1 Introduction

Due to the effects of Covid-19, there has been a significant increase in demand for in-home entertainment. One particular form of this entertainment that continues to thrive due to streaming services such as Netflix and Stan is watching movies. This report will be analysing a large dataset from the popular reviews website, IMDb. It will attempt to answer questions such as what periods experienced a significantly higher quality of movies and which streaming services offer the greatest variety of movies.

```
## Rows: 81,273
## Columns: 22
## $ imdb_title_id         <fct> tt0000574, tt0001892, tt0002101, tt0002130, t...
## $ title                 <fct> "The Story of the Kelly Gang", "Den sorte drÃ...
## $ original_title        <fct> "The Story of the Kelly Gang", "Den sorte drÃ...
## $ year                  <int> 1906, 1911, 1912, 1911, 1912, 1919, 1913, 191...
## $ date_published        <fct> 1906-12-26, 1911-08-19, 1912-11-13, 1911-03-0...
## $ genre                 <fct> "Biography, Crime, Drama", "Drama", "Drama, H...
## $ duration              <int> 70, 53, 100, 68, 60, 85, 120, 120, 55, 121, 5...
## $ country               <fct> "Australia", "Germany, Denmark", "USA", "Ital...
## $ language              <fct> "", "", "English", "Italian", "English", "Ger...
## $ director              <fct> "Charles Tait", "Urban Gad", "Charles L. Gask...
## $ writer                <fct> "Charles Tait", "Urban Gad, Gebhard SchÃ¤tzle...
## $ production_company    <fct> J. and N. Tait, Fotorama, Helen Gardner Pictu...
## $ actors                <fct> "Elizabeth Tait, John Tait, Norman Campbell, ...
## $ description           <fct> "True story of notorious Australian outlaw Ne...
## $ avg_vote              <dbl> 6.1, 5.9, 5.2, 7.0, 5.7, 6.8, 6.2, 6.7, 5.5, ...
## $ votes                 <int> 537, 171, 420, 2019, 438, 709, 241, 187, 211,...
## $ budget                <fct> $ 2250, , $ 45000, , , , ITL 45000, ROL 40000...
## $ usa_gross_income      <fct> , , , , , , , , , , , , , , , , , , , , , ,...
## $ worlwide_gross_income <fct> , , , , , , , , , , , , , , , , , , , , , ,...
## $ metascore             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ reviews_from_users    <dbl> 7, 4, 24, 28, 12, 11, 6, 3, 7, 9, 9, 16, 8, 2...
## $ reviews_from_critics  <dbl> 7, 2, 3, 14, 5, 9, 4, 1, 1, 9, 29, 7, 22, 2, ...
```

Table 1: *Metascore summary*

	metascore
	Min. : 1.00
	1st Qu.: 43.00
	Median : 56.00
	Mean : 55.76
	3rd Qu.: 69.00
	Max. :100.00
	NA's :68551

2 William's Section

According to IMDb, metascore is a single number that represents the overall critical opinion about a movie. Its legitimacy is derived from applying a weighted average to the reviews of the world's most respected critics. That said, metascore is still an extremely vague concept - particularly when it comes to the threshold of what metascore can be considered good.

Therefore, it is important to first consider the distribution of metascore within the data set. This is summarised by Table 1. The most noteworthy features are that the median and mean are very close at 56 and 55.76 respectively and nearly 70,000 movies have not been assigned a metascore. Consequently, any analysis related to the metascore variable must be treated with caution as less than 20% of the movies are represented.

As expected, the scatter plot depicted in Figure 1 is not particularly informative and potentially misleading. Although the spread is increasing, one can still identify a negative correlation which suggests that the quality of movies has degraded over time. However, it is important to recognise that this dataset was taken from IMDb which means that sampling was far from random. IMDb has much more incentive to review recently released movies. Hence, when older movies are reviewed, it is likely because it is still immensely popular (i.e. a 'classic').

The aforementioned 'classics bias' is further confirmed by both Table 2 and Figure 2. This list is dominated by the most famous movies from the 1900's such as The Wizard of Oz, Citizen Kane and The Godfather which were all assigned perfect metascores. Moreover, seventeen of the twenty listed movies were released before 1975 and only one was released after 2000.

Assuming that metascore is a reliable representation of a movie's quality, it is easy to see why arguments have been made that movies produced in recent times are not as good. Contrarily, a counterargument can be made that there were just as many bad movies made during previous generations that have

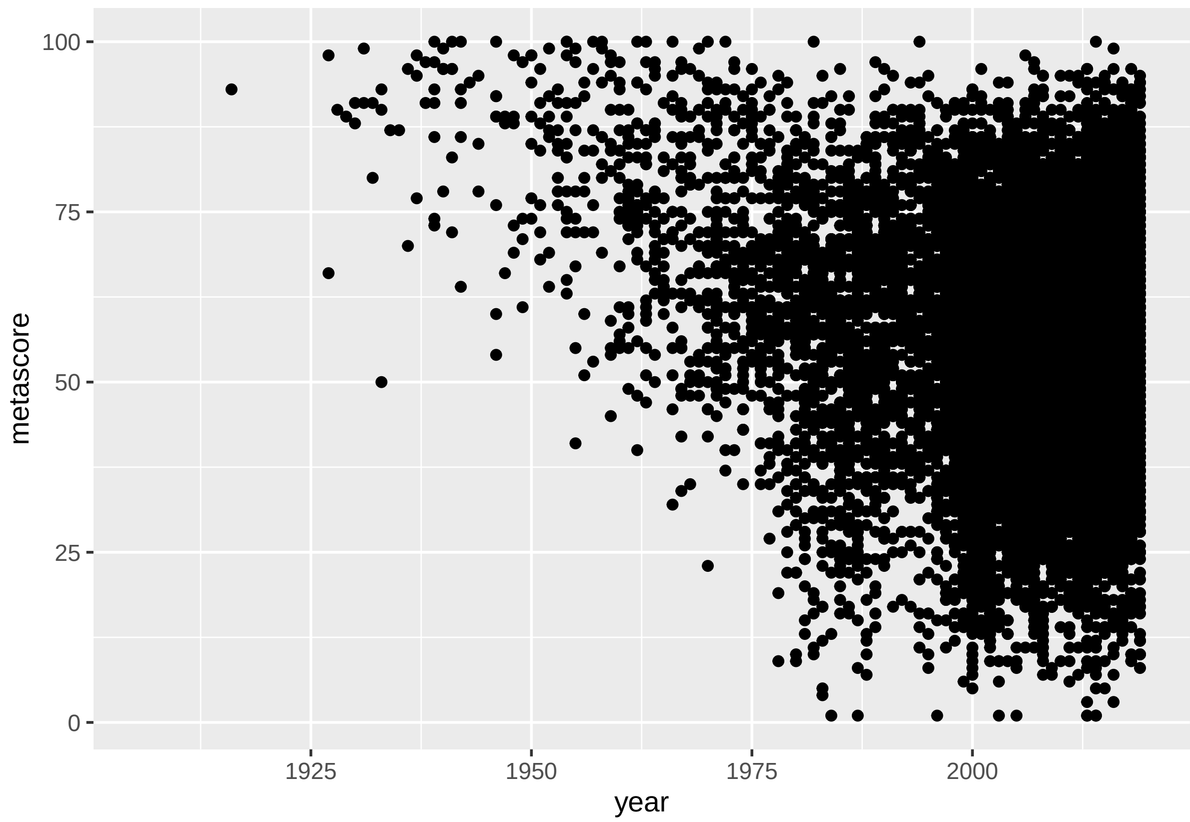

Figure 1: *metascore vs time*

Table 2: *The top 20 movies arranged by metascore*

title	year	metascore
The Wizard of Oz	1939	100
Citizen Kane	1941	100
Casablanca	1942	100
Notorious	1946	100
Viaggio in Italia	1954	100
Rear Window	1954	100
Sweet Smell of Success	1957	100
Vertigo	1958	100
Lawrence of Arabia	1962	100
Il gattopardo	1963	100
Au hasard Balthazar	1966	100
Il conformista	1970	100
The Godfather	1972	100
Fanny och Alexander	1982	100
Trois couleurs: Rouge	1994	100
Boyhood	2014	100
City Lights	1931	99
Pinocchio	1940	99
Singin' in the Rain	1952	99
The Night of the Hunter	1955	99

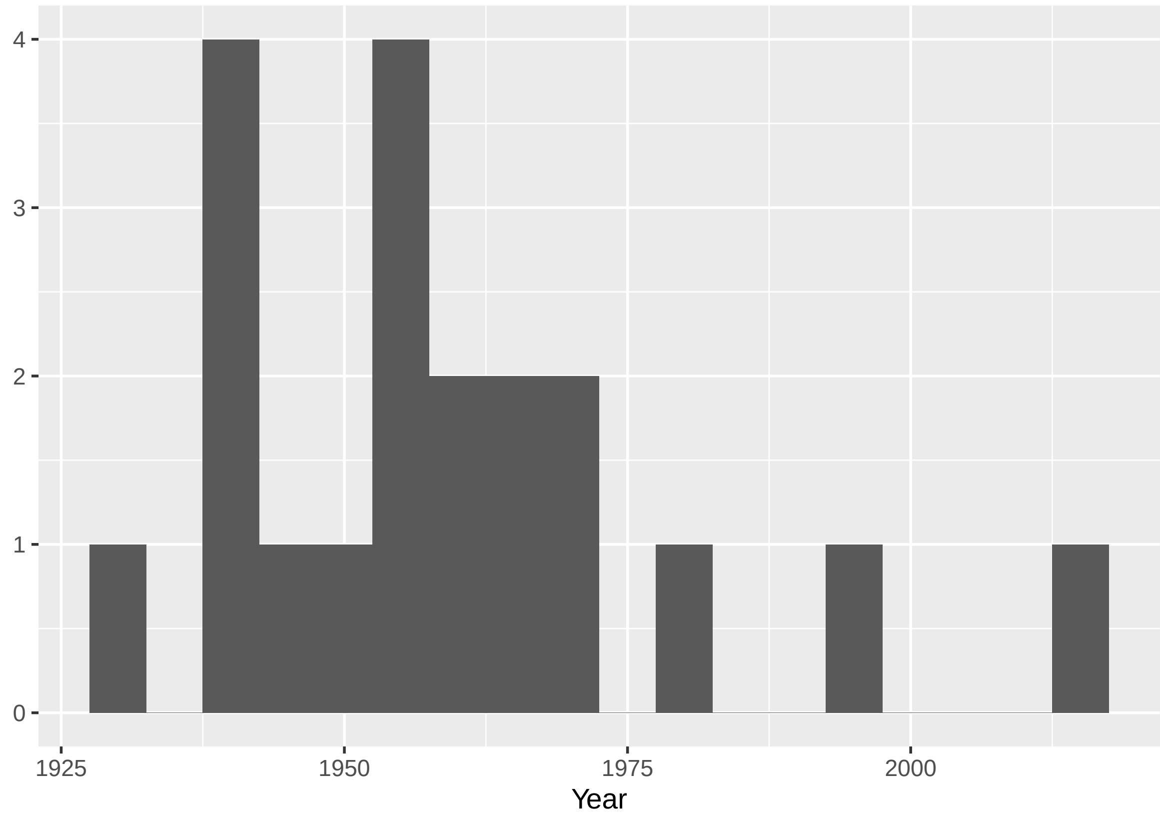

Figure 2: *Time distribution of top twenty movies*

been forgotten or overshadowed by the classics. Due to the clear bias within this dataset, it is still unclear which argument is stronger.

3 Xue Wang's Section

3.1 Language

3.1.1 The proportion of top 100 movie language

The figure 3 shows the proportion of the number top 100 languages. It can clearly be seen that the English movies account for almost half of all movies. French, Spanish, Japanese, Italian almost have the same proportion with the number of above 2500. It is very clear to see the situation about the distribution.

3.1.2 Number of top 9 languages between 2009 and 2019

The figure 4 shows the number of different languages movies between 2009 and 2019. It can be clearly seen that the largest number of movie languages is English, it has an increasing trend between 2009 and 2017 and will drop to 600 in 2019. The number of Korean language movies have the minimum quantity within the 9 languages, it shows a decreasing trend during the ten years. The

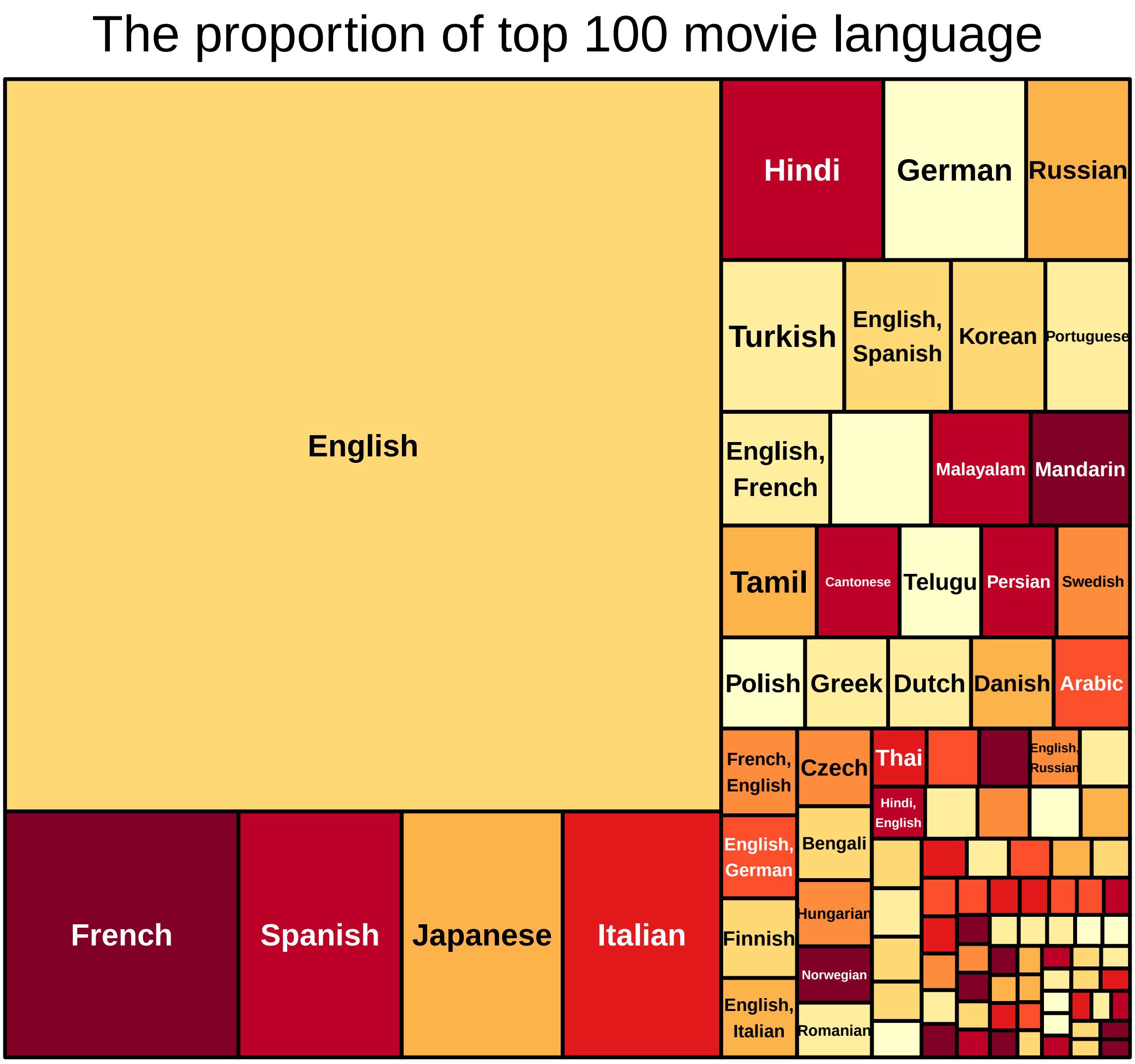

Figure 3: *The proportion of top 100 movie language*

number of French,German, Japanese and Italian, they almost maintained the level from 2009 to 2017, and then sharply decreased to 2019.

3.2 Genres

3.2.1 The proportion of top 100 movie genres

The figure 5 shows the proportion of top 100 movie genres, from the figure, we can see that the drama and comedy almost occupied 30% of total movies, followed by the horror, comedy and romance, drama and romance, thriller.etc almost have the same proportion, this picture clearly shows the proportion of different movie genres.

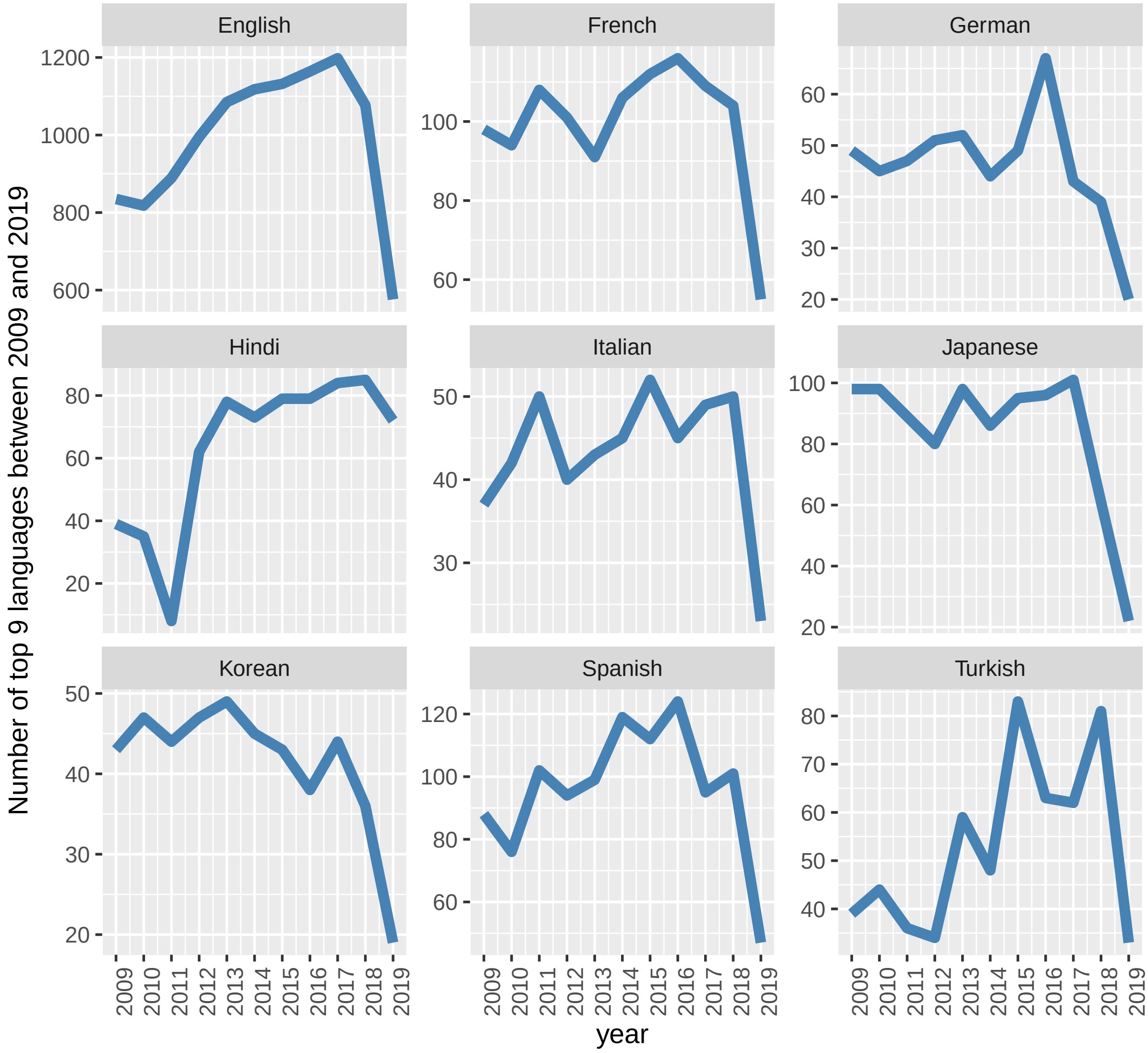

Figure 4: *Top 9 languages between 2009 and 2019*

3.2.2 Number of top 10 genres of movies published from 2009 to 2019

The figure 6 graph shows the change in the number of different movie genres in ten years, from this figure, it can be seen that,the number of dramas and comedies almost exceeds 100 every year, and dramas almost reach 500 every year. When it comes to the line chart, the number of almost all movie genres peaked in 2017 and then fell to 2019. This is because the movie industry returns to the content ontology and enters the transition period from quantity to quality.

3.2.3 The largest number of movies genres in each country

Table 3 show the largest number of movies genres in each country in the 21st century, according to this table, we can see the the largest number of movies in many countries is drama, According Table

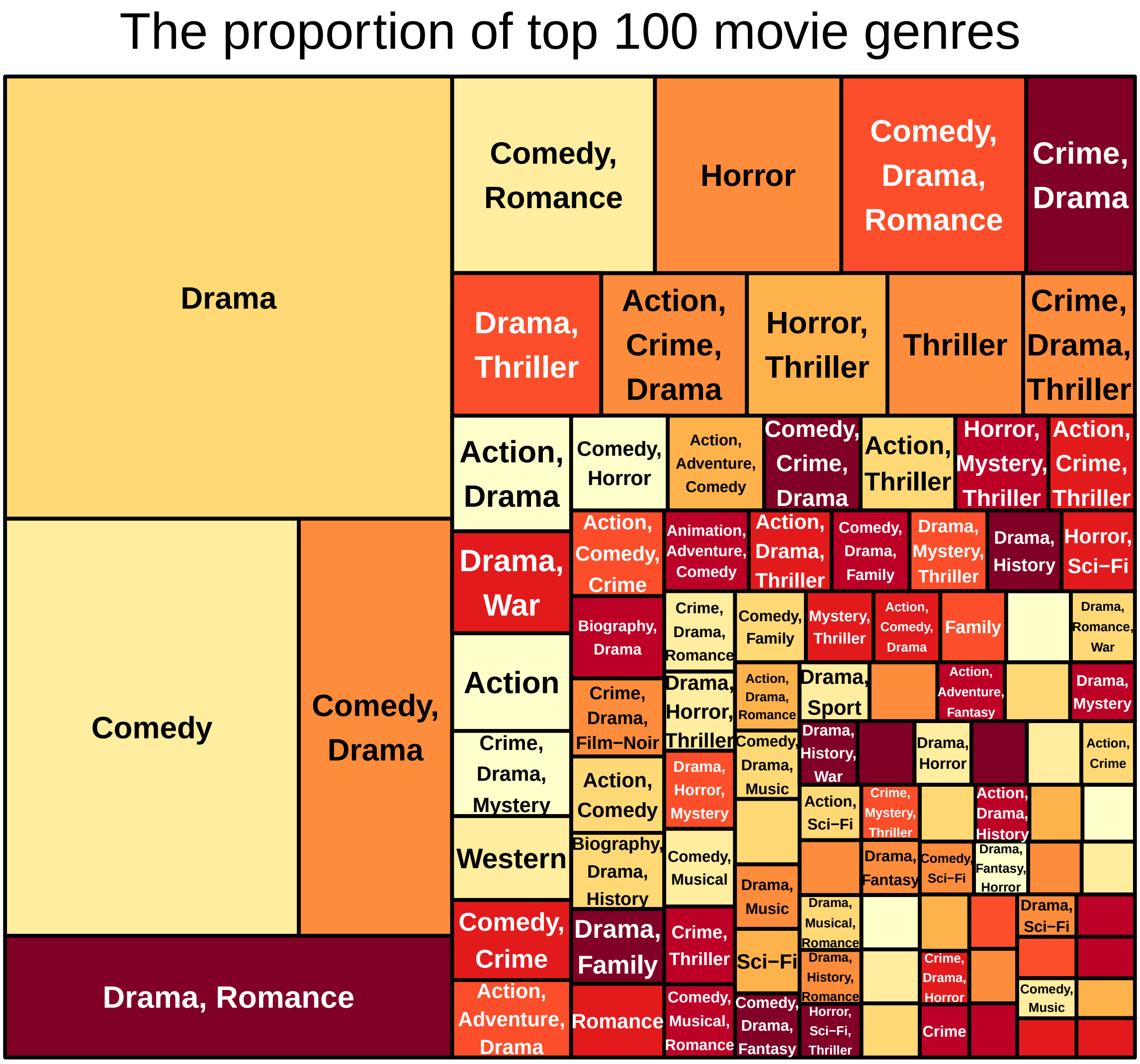

Figure 5: *The proportion of top 100 movie genres*

Table 3: *The largest number of movies genres in each country*

country	genre	n
USA	Drama	1252
India	Drama	436
France	Drama	334
Italy	Comedy	295
Japan	Drama	267
Canada	Drama	217
Germany	Drama	215
UK	Drama	171
Turkey	Comedy	166
Spain	Drama	163

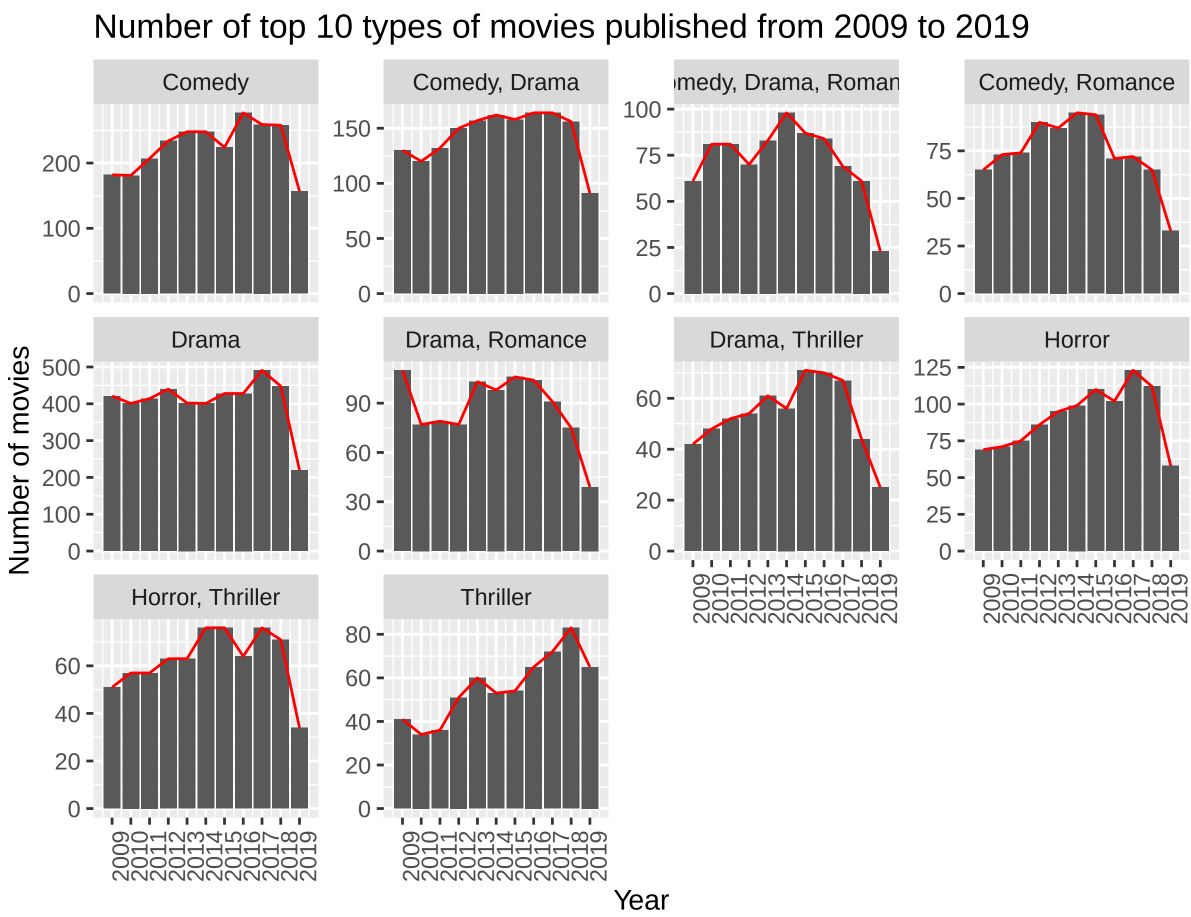

Figure 6: *Number of top 10 types of movies published from 2009 to 2019*

Table 4: *The largest number of movies genres in each country*

genre	total	total_movies	prop_genre
Drama	6534	11103	0.5884896
Comedy	914	11103	0.0823201
Comedy, Drama	290	11103	0.0261191
Drama, Romance	239	11103	0.0215257
Comedy, Drama, Romance	121	11103	0.0108980

4 , it shows the drama has the largest proportion with 58.84%, However, the second place comedy only reached 8%.

Table 5: *Sample Gross Data Table*

Title	Budget	USA_Gross
Das Cabinet des Dr. Caligari	$ 18000	$ 8811
The Four Horsemen of the Apocalypse	$ 800000	$ 9183673
Metropolis	DEM 6000000	$ 1236166
City Lights	$ 1500000	$ 19181
Modern Times	$ 1500000	$ 163577
Snow White and the Seven Dwarfs	$ 1499000	$ 184925486
Gone with the Wind	$ 3977000	$ 200852579
Mr. Smith Goes to Washington	$ 1900000	$ 144738
La rÃ¨gle du jeu	FRF 5500500	$ 273641
The Wizard of Oz	$ 2777000	$ 24790250

Table 6: *Sample Review Data Table*

Title	User_Reviews	Critic_Reviews
The Story of the Kelly Gang	7	7
Den sorte drÃ¸m	4	2
Cleopatra	24	3
L'Inferno	28	14
From the Manger to the Cross; or, Jesus of Nazareth	12	5
Madame DuBarry	11	9
Quo Vadis?	6	4
Independenta Romaniei	3	1
Richard III	7	1
Atlantis	9	9

4 Rahul's Section.

4.1 Movie Grosses and Popularity Analysis

- What is the Budget and USA Gross for Avenger and Spider-Man movie franchises?
- What is the Number of user reviews and critic reviews for the same, implying their popularity?
- What is the gross of the most popular movies of all-time?

Tidyverse (Wickham et al. (2019)) is used to clean the data and kableExtra (Zhu (2019)) is used to display tables.

The table 5 shows a sample of the gross data used for the analysis. The main columns needed for Gross analysis are Title, Budget, and USA_Gross.

The table 6 shows a sample of the review data used for the analysis. The main columns needed for Gross analysis are Title, User_Reviews, and Critic_Reviews.

Using the table contents, we plot data for a particular movie franchise. The franchises selected for this anlysis are Avengers and Spider-Man. We display plots to analyse how much the franchise spent for the movie and how much it grossed. Also a plot for the user reviews and critic reviews are displayed to see which were the most popular among all the movies in the franchise. We also plot the same for all the movies in the dataset.

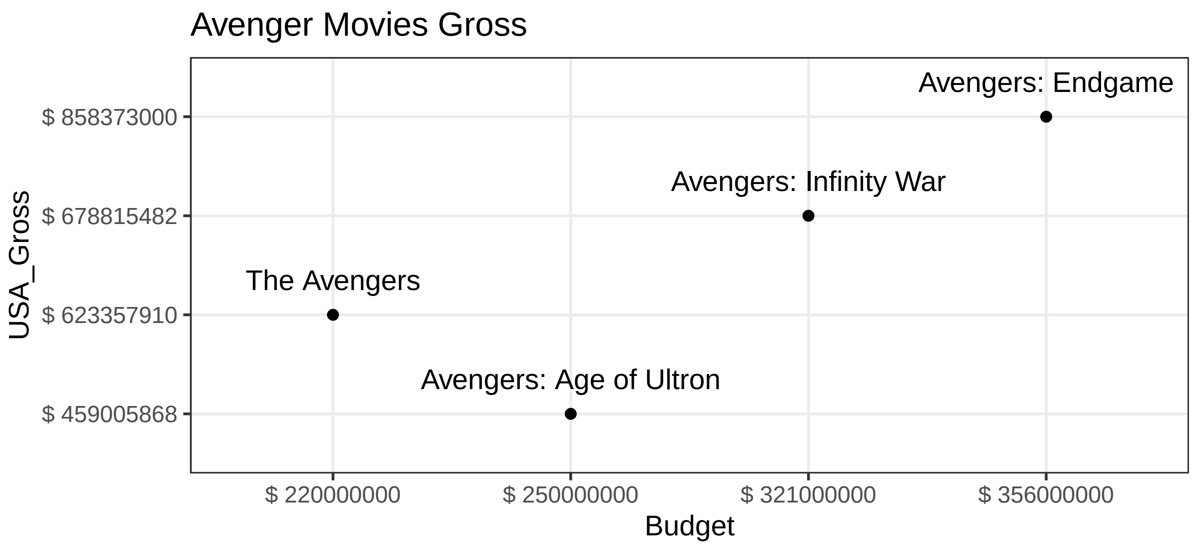

Figure 7: *Avengers Budget v/s Gross*

From the figure 7, it can be seen that all Avenger movies grossed more than the amount spent. Avengers: Age of Ultron grossed relatively lesser than the other movies. Endgame was the highest grossing Avenger movie.

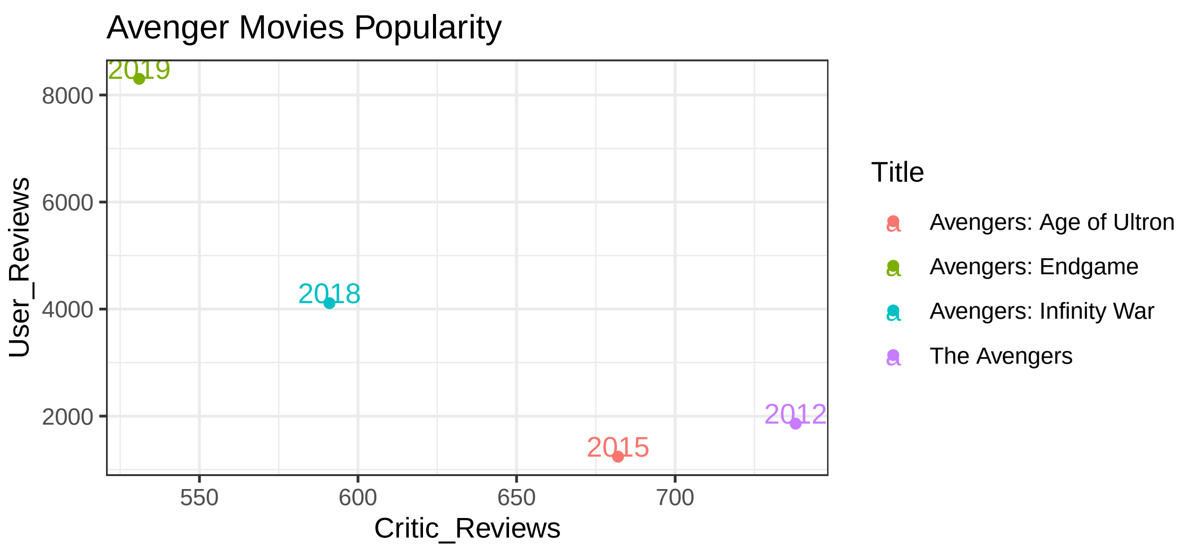

Figure 8: *Avengers Popularity*

The figure 8 displays how many reviews was given by users and critics. This can be a measure of how much the movies were talked about. The year tags show the timeline. It is evident that the first avenger movie in 2012 attracted the attention of critics with a high number of critic reviews. As time passed by, the movies become more popular among users which shows an increse in the fanbase till 2019. The only exception to this is a dip in user reviews for Age of Ultron which justifies why it grossed lesser than other movies.

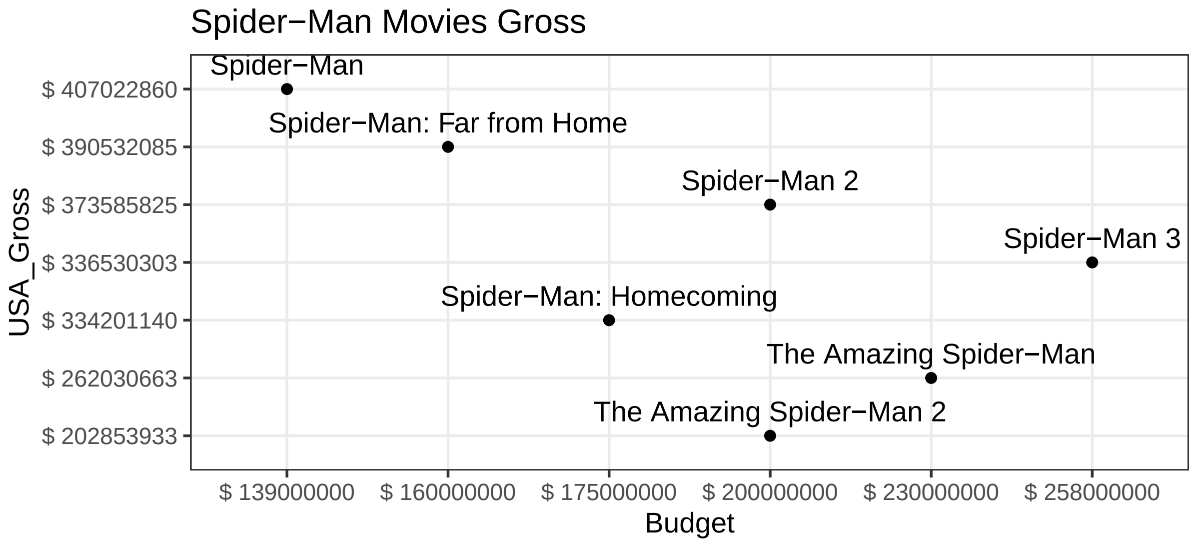

Figure 9: *Spider-Man Budget v/s Gross*

The figure 9 shows the budget and gross of Spider-Man movies. Spider-Man grossed the highest and the sequels Spider-Man 2 and Spider-Man 3 grossed lesser showing a decline in the gross. The Amazing Spider-Man seems to have grossed better than its sequel The Amazing Spider-Man 2. These movies has grossed far less than the first 3 movies. The Homecoming and Far from Home movies seem to have made a great comeback in terms of its gross compared to the fourth and fifth movies. Far from Home has grossed more and spent less which makes it the second best Spider-Man movie after the first ever one in the franchise.

The figure 10 shows a graph for the number of reviews by critics and users over a period of time named by the years for each points. The first ever movie was most talked about by both fans and users in 2002. We see a dip in user reviews for 2004 and the movie in 2007 was a little better but the first ever was the best recieved. The fourth and fifth movies in 2012 and 2014 was talked about most by critics but never really kicked off among fans. This justifies its low grosses. The final sixth and seventh movies in 2017 and 2019 have made a great comeback after this dip and the latest one during 2019 has managed to almost equal the fanbase like it was when the franchise started.

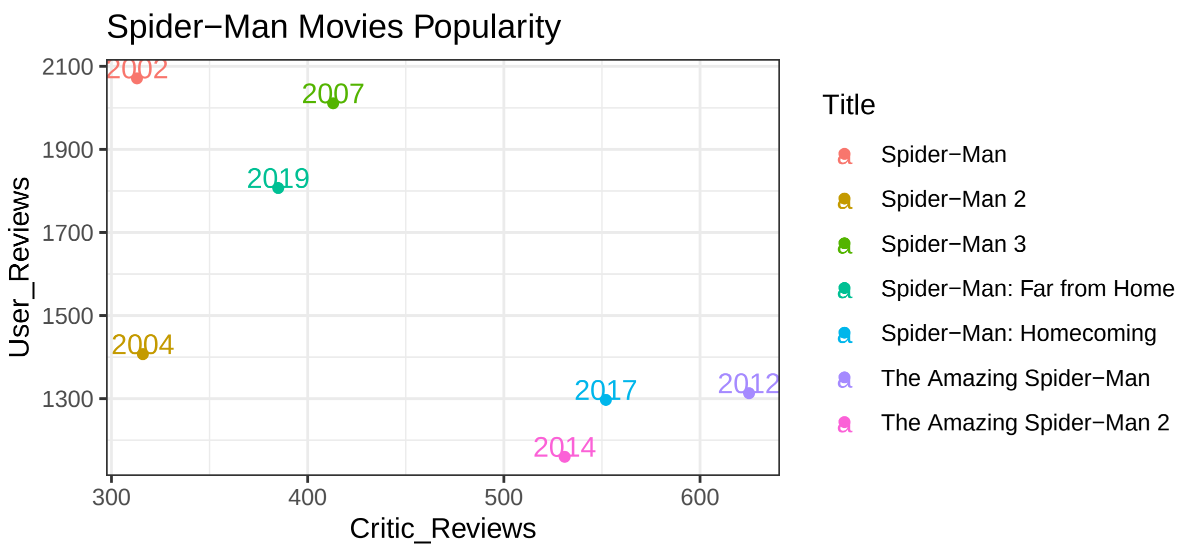

Figure 10: *Spider-Man Popularity*

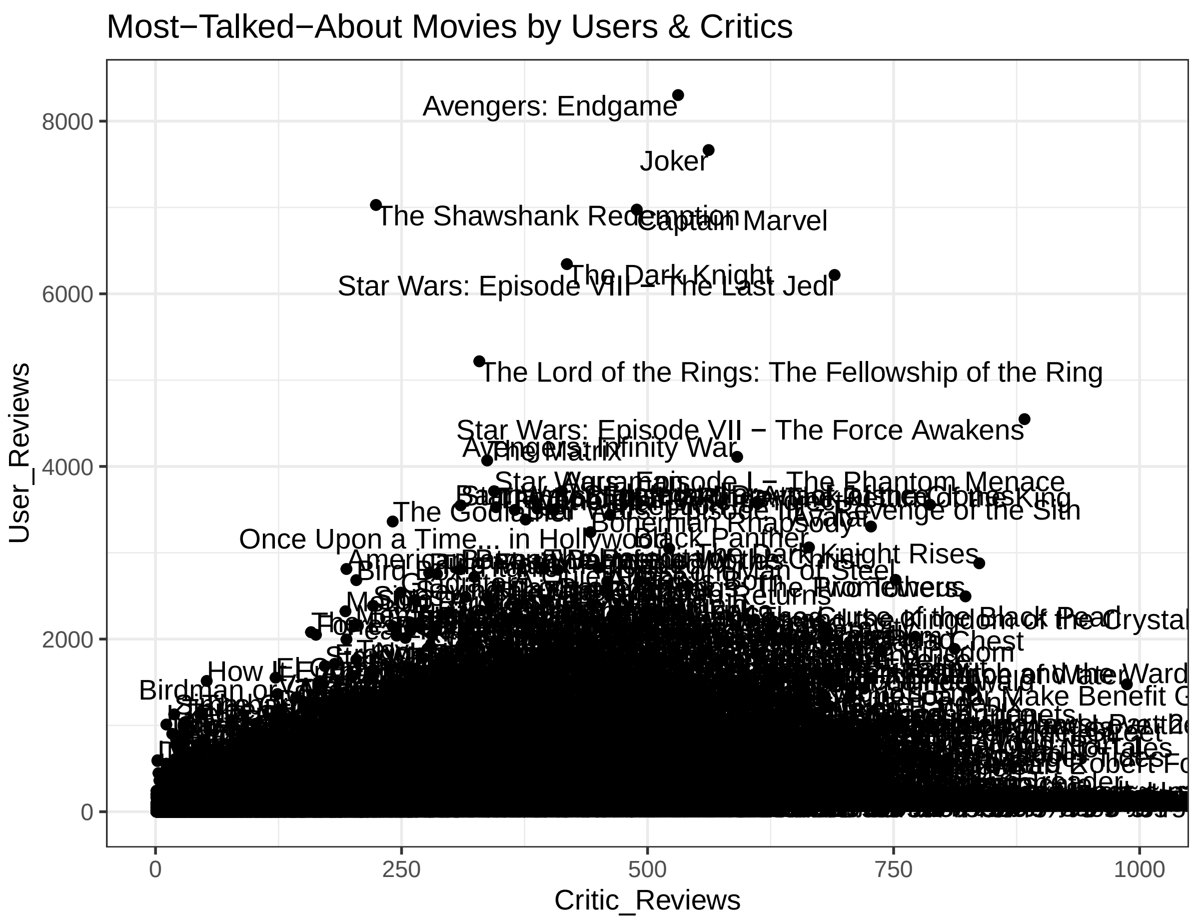

Figure 11: *All-time Popularity*

The figure 11 looks shabby when you first look at it but it has some insights that can be drawn from it. The black cluster at the left bottom are the movies that have a low number of reviews by both users and critics. These are the least popular movies. The outliers are clearly visible in this plot with Endgame being the most popular among fans around the globe. This plot is used to pick the most popular movies to be used to compare the budget and gross. It is not necessary for a movie to be good, to be most talked about. So the next plot serves as a verification to confirm if the movie was talked about because it was good, or bad.

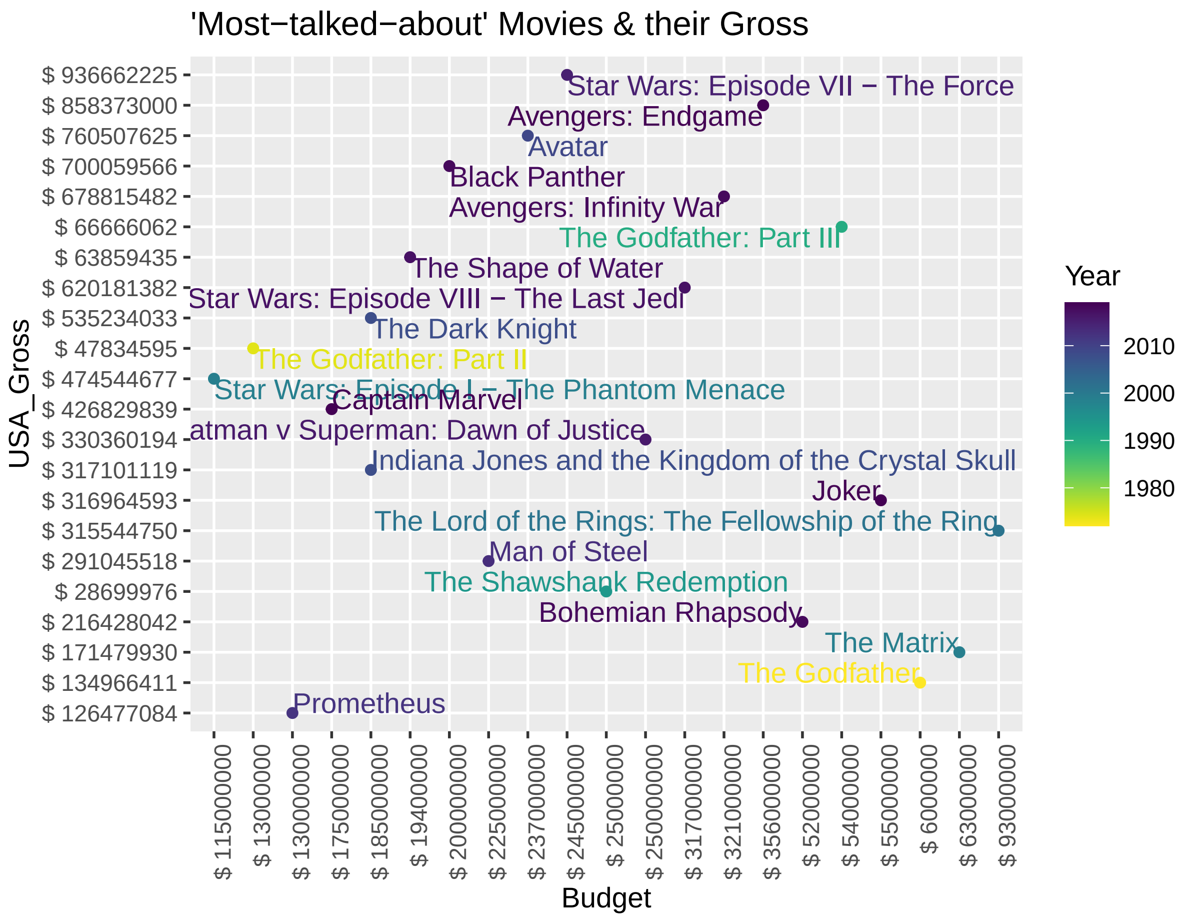

Figure 12: *Some all-time Popular Movies and their Gross*

The figure 12 shows the budget and gross of the most popular movies from over a century. We can compare the production cost, which is the budget, and the gross to ensure that it was most talked about because it was good. This plot also has information about the decade it was released in. This was to ensure there was no bias between the movies since currency value is always changing. The values are discrete and shows how much was spent and how much was grossed for that time. The gradient is gven to year using Viridis.

5 Aryan's Section.

5.1 Director Analysis

5.1.1 Most Successful Director

Lets take a look at the some of the most successful directors from our dataset. Now, there are limitless ways to measure success but for our test, we're only looking at people with at least 4 movies under their belt and a Metascore upwards of 80. And conveniently, we get only 10 names.

Table 7: *Top Grossing Movies*

Director	Total Movies	Metascore
Damien Chazelle	4	87
Patrick Wang	4	86
Paul Thomas Anderson	8	84
Spike Jonze	4	84
Steve McQueen	4	84
Mike Leigh	13	82
Alfonso CuarÃ³n	8	81
Nuri Bilge Ceylan	8	81
Bong Joon Ho	7	81
Andrey Zvyagintsev	5	81

We get some interesting results from Table 7, as on one hand there is Mike Leigh with as much as 13 movies but a lower average Metascore. On the other hand, there are directors like Damien Chazelle and Patrick wang with 4 movies each and an average Metascore of 87 and 86 respectively.

5.2 Top Grossing Movie

In this section, we'll look at the top grossing movie of the most successful directors that we established from Table 7.

Table 8: *Director Top grossing movies*

Director	Highest Grossing Movie	Total Revenue
Alfonso CuarÃ³n	Harry Potter and the Prisoner of Azkaban	$796,093,802
Damien Chazelle	La La Land	$446,092,357
Steve McQueen	12 Years a Slave	$187,733,202
Bong Joon Ho	Gisaengchung	$109,012,467
Spike Jonze	Where the Wild Things Are	$100,086,793
Paul Thomas Anderson	There Will Be Blood	$76,181,545
Mike Leigh	Mr. Turner	$22,179,785
Andrey Zvyagintsev	Vozvrashchenie	$8,482,993
Nuri Bilge Ceylan	Kis Uykusu	$4,018,705
Patrick Wang	In the Family	$101,934

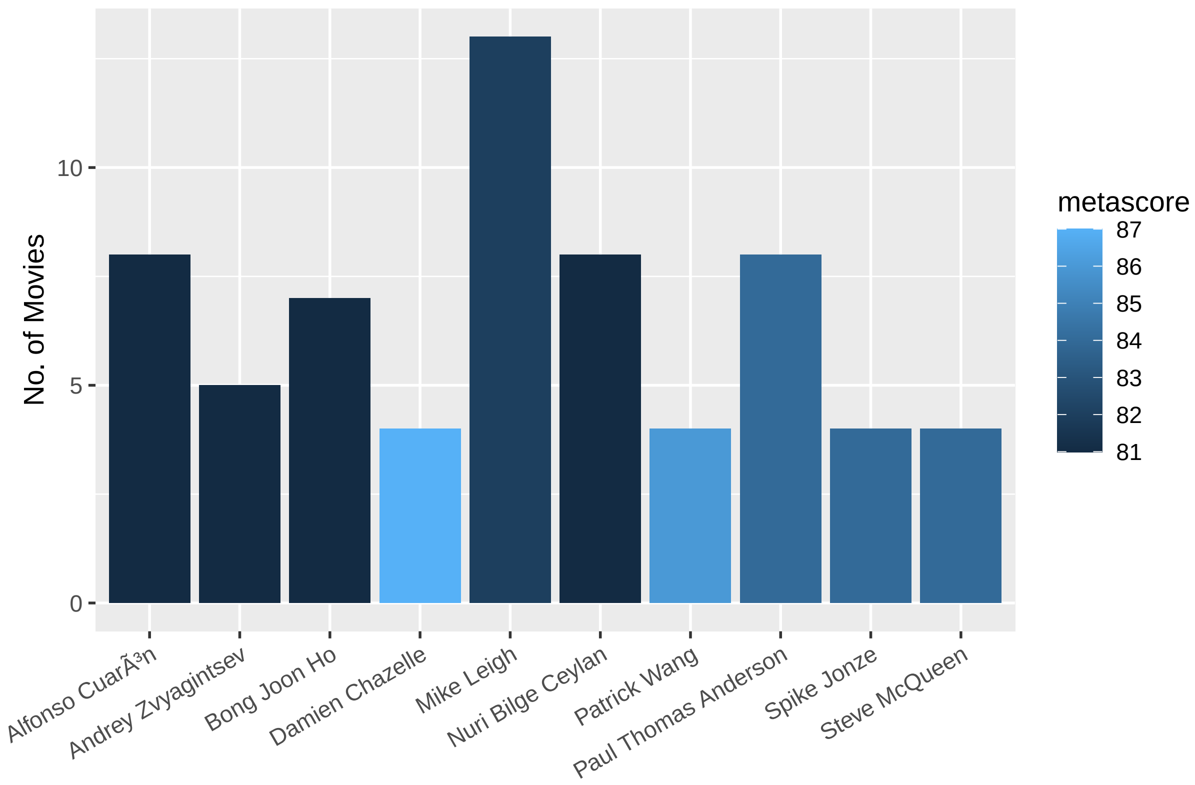

Figure 13: *Total number of movies for Top Directors*

It can be observed from Table 8 that the *Alfonso Cuarón* is the director of the top grossing movie *Harry Potter and the Prisoner of Azkaban* which made $796,093,802 worldwide. Followed by, *Damien Chazelle*, the director of *La La Land* grossing $446,092,35. Suprisingly, *Patrick Wang*, with an average metascore of 86 and a total of 4 movies made the least at just $101,934.

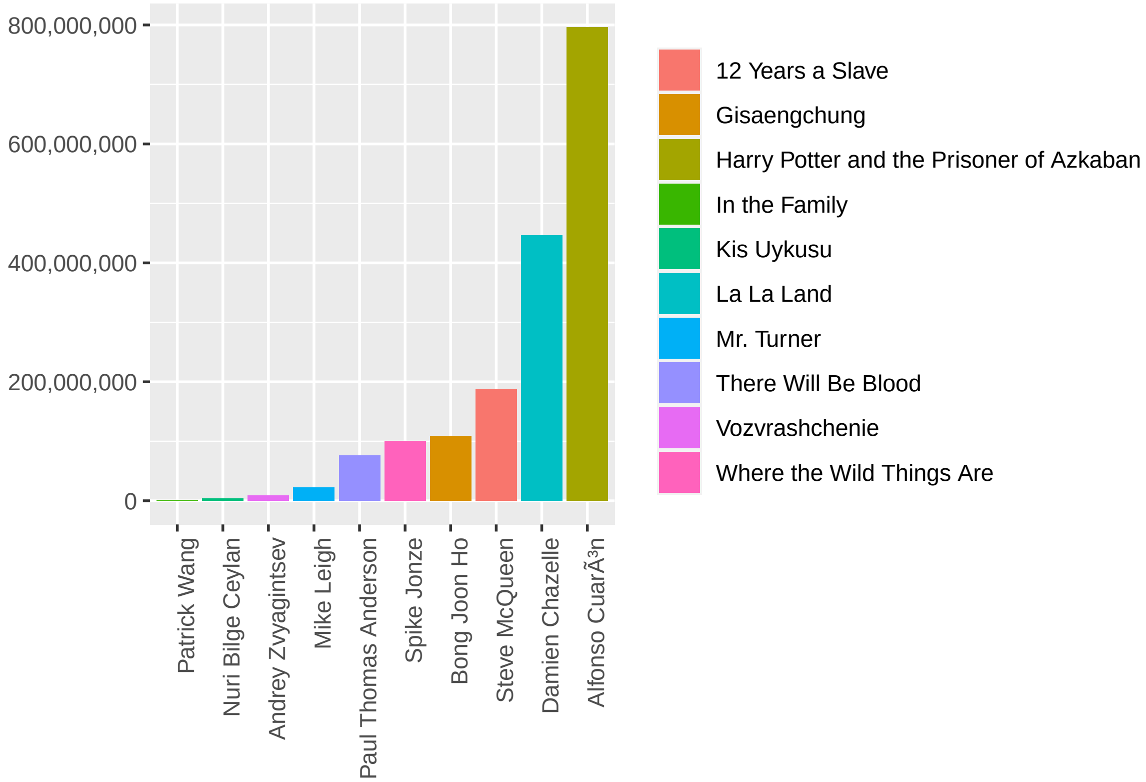

Figure 14: *Top Grossing Movies*

5.3 Movie Length and Language Analysis

5.3.1 Let's look at some of the common movie languages.

Table 9: *Number of movies per language*

Language	Number of Movies
507	34519
multi-lingual	15188
1854	3777
3770	2640
2875	2605
2743	2565
2501	1930
2206	1702
3541	1235
4133	1229

From the Table 9, it is quite clear that most movies are released in either only *English* or in *multi-lingual* format.

Table 10: *Average movie duration for all languages*

Language	Avg. Movie Duration
1854	99.38734
2206	96.43831
2501	140.40466
2743	98.94191
2875	102.96161
3541	99.78219
3770	97.35833
4133	96.79170
507	93.50190
multi-lingual	106.46787

5.3.2 Is there a pattern in Movie length for various languages?

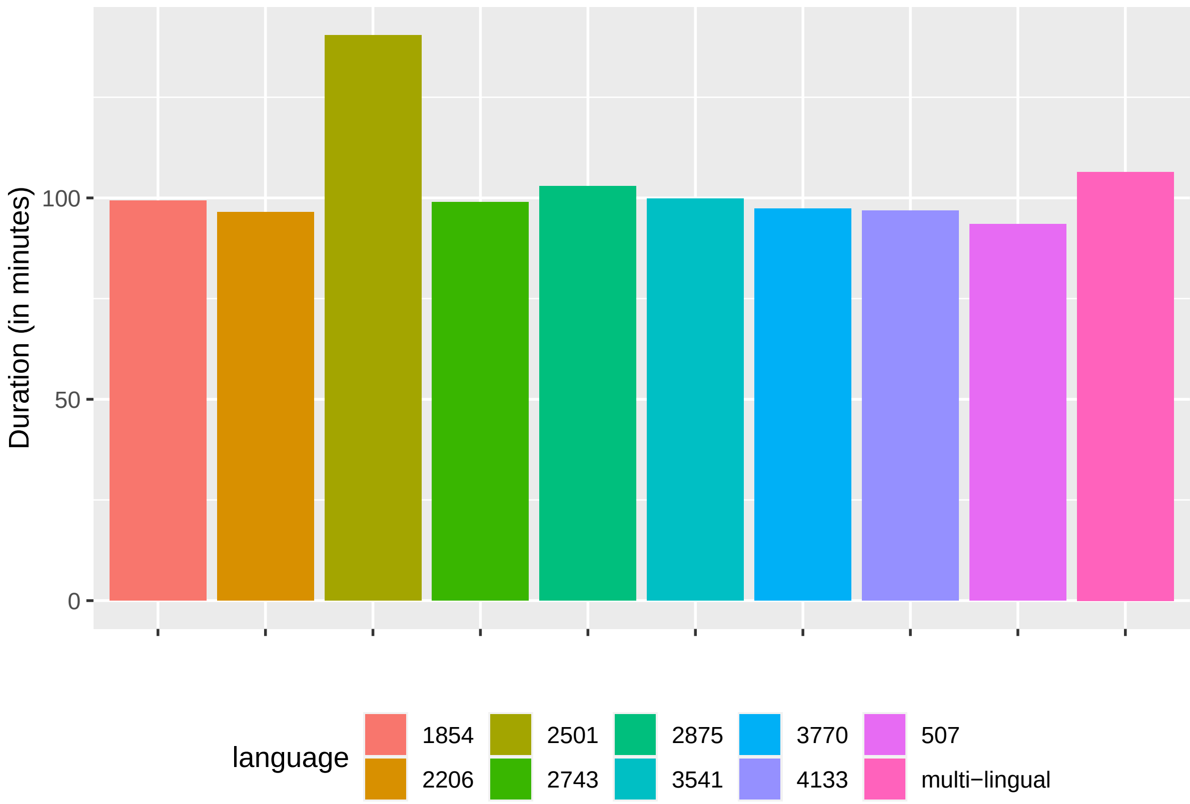

Figure 15: *Movie Duration for different Languages*

From the Figure 15, we can infer that *Hindi* movies are usually much longer than other languages clocking at an average runtime of 140 minutes. While, *English* movies are usually the shortest at just 93 minutes.

6 Conclusion

William's Section Conclusion

There appears to be a strong correlation between movies that have been given extremely high ratings and the mid 1950's. However, due to the inherent bias within this data set, it is still impossible to identify which periods truly experienced a higher quality of movie production.

Xue's Section Conclusion

In conclusion, the English movie has the largest proportion of language movies within ten years, the drama movie is the favorite movie in all the country.

Rahul's Section Conclusion

In conclusion, we can see a general trend that a movie that is most talked about by the users or fans has grossed the most amount of money compared to lesser popular movies. The number of reviews by users can generally be used to predict if it was a box office hit.

Aryan's Section Conclusion

For Director analysis, it is clear from Figure 1 and Figure 2 that there is no one single way to measure success and that all directors in the list are doing very good in their own ways.

And from the Movie length analysis, we can infer that Hindi Movies are usually significantly longer than any other kind of movies. Also, english movies are among the shortest. We also concluded that movie language has an impact on movie length.

7 Citations

tidyverse Wickham et al. (2019)

dplyr Wickham et al. (2020)

kableExtra Zhu (2019)

treemap Tennekes (2017)

viridis Garnier (2018)

scales Wickham and Seidel (2020)

References

Garnier, S (2018). *viridis: Default Color Maps from 'matplotlib'*. R package version 0.5.1. https://CRAN.R-project.org/package=viridis.

Tennekes, M (2017). *treemap: Treemap Visualization*. R package version 2.4-2. https://CRAN.R-project.org/package=treemap.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H, R François, L Henry, and K Müller (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr.

Wickham, H and D Seidel (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. https://CRAN.R-project.org/package=scales.

Zhu, H (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra.