MONASH
BUSINESS
SCHOOL

ETC5513 Assignment 4: IMDB Data Analysis

Mr Rahul Bharadwaj Mysore Venkatesh
Master of Business Analytics

Report for
Monash University

11 June 2020

Department of Econometrics & Business Statistics

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

ABN: 12 377 614 012

Reasearch Questions

- 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?

```
## 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<c3><87>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, ...
```

- The above chunk shows the column names of the raw dataset. The columns are selected based on the research questions to be answered. The column names are cleaned by renaming the columns clearly and prepared to be useful for analysis.

Table 1: *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 2: *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

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

The table 1 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 2 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.

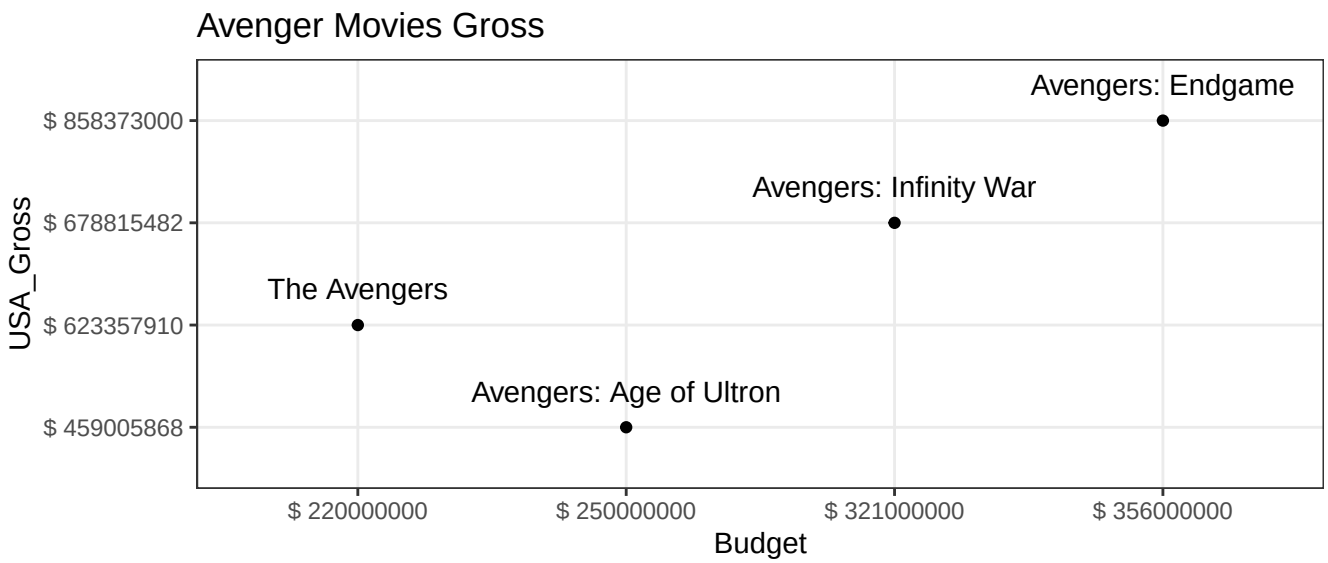

Figure 1: *Avengers Budget v/s Gross*

From the figure 1, 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.

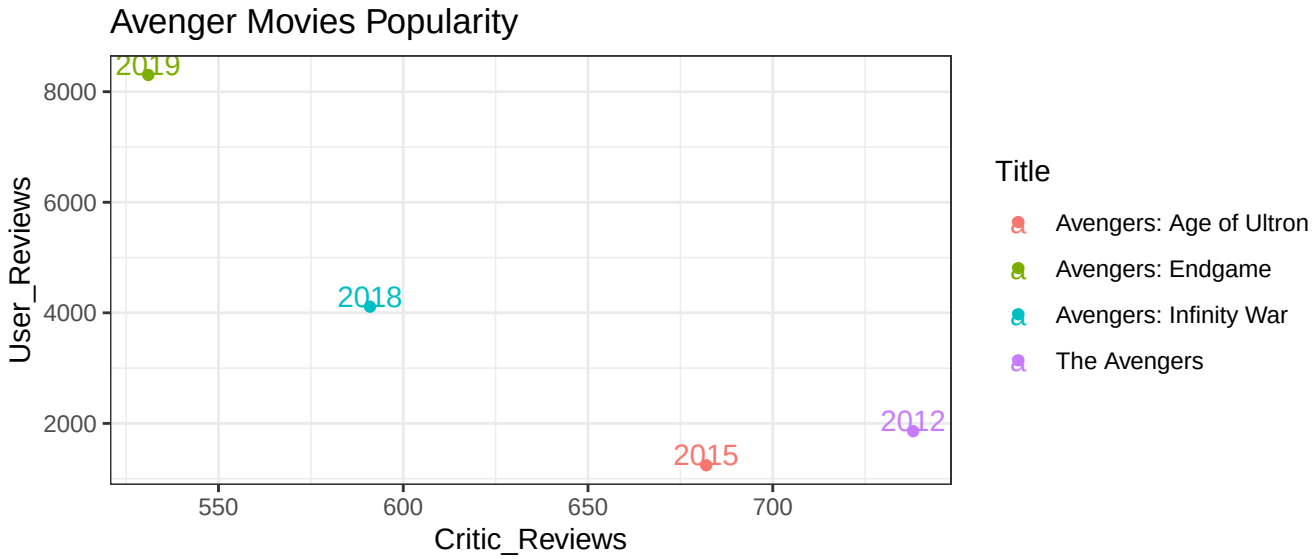

Figure 2: *Avengers Popularity*

The figure 2 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.

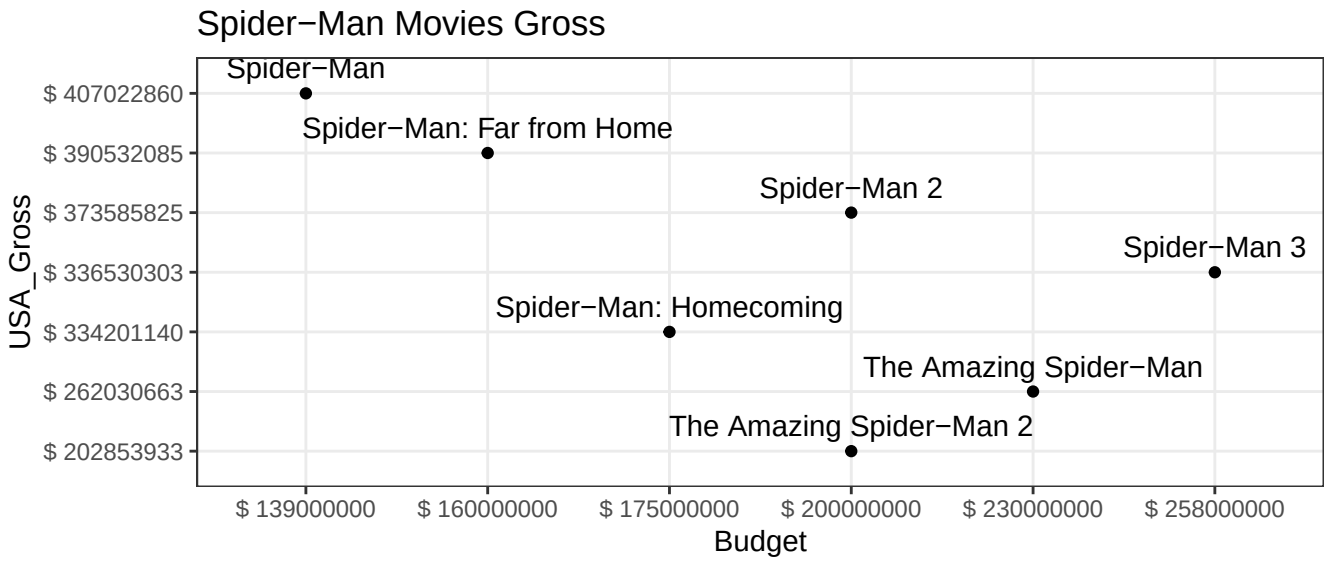

Figure 3: *Spider-Man Budget v/s Gross*

The figure 3 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 4 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.

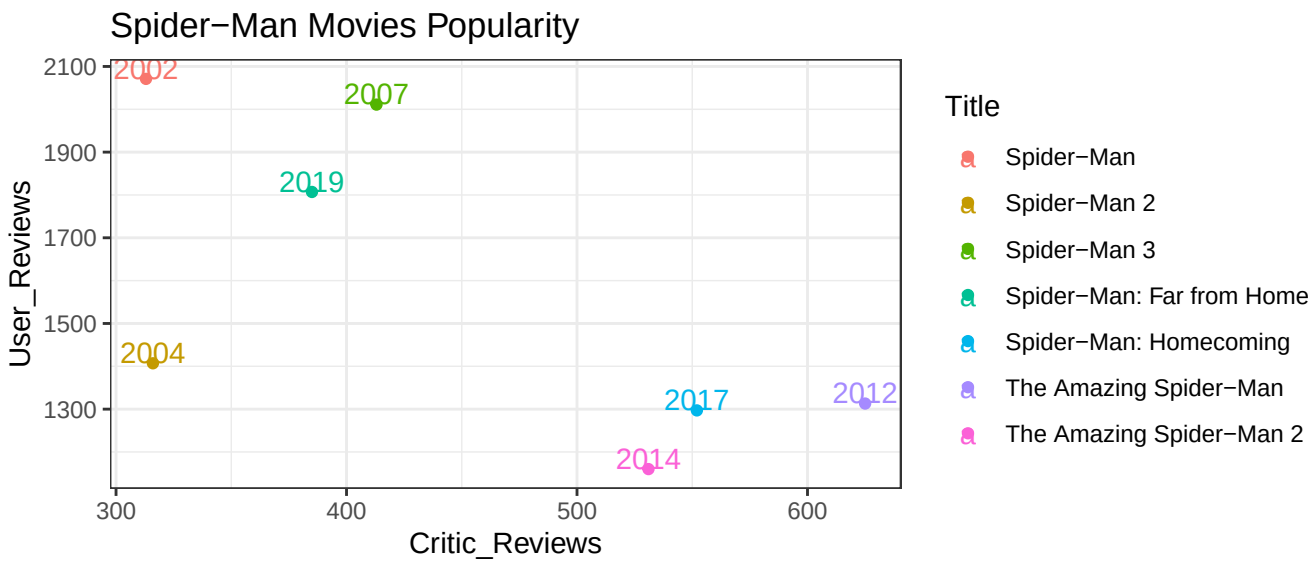

Figure 4: *Spider-Man Popularity*

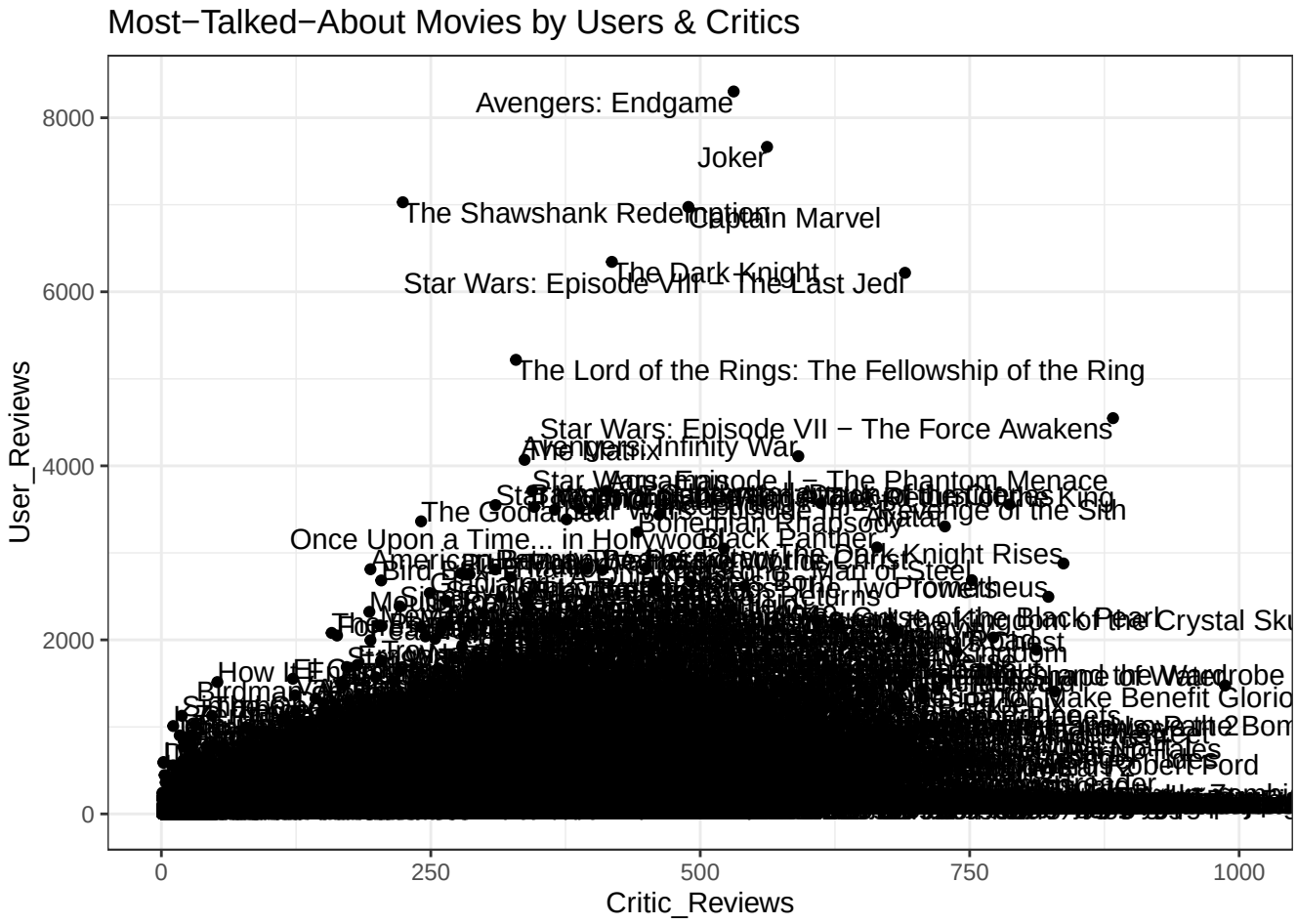

Figure 5: *All-time Popularity*

The figure 5 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.

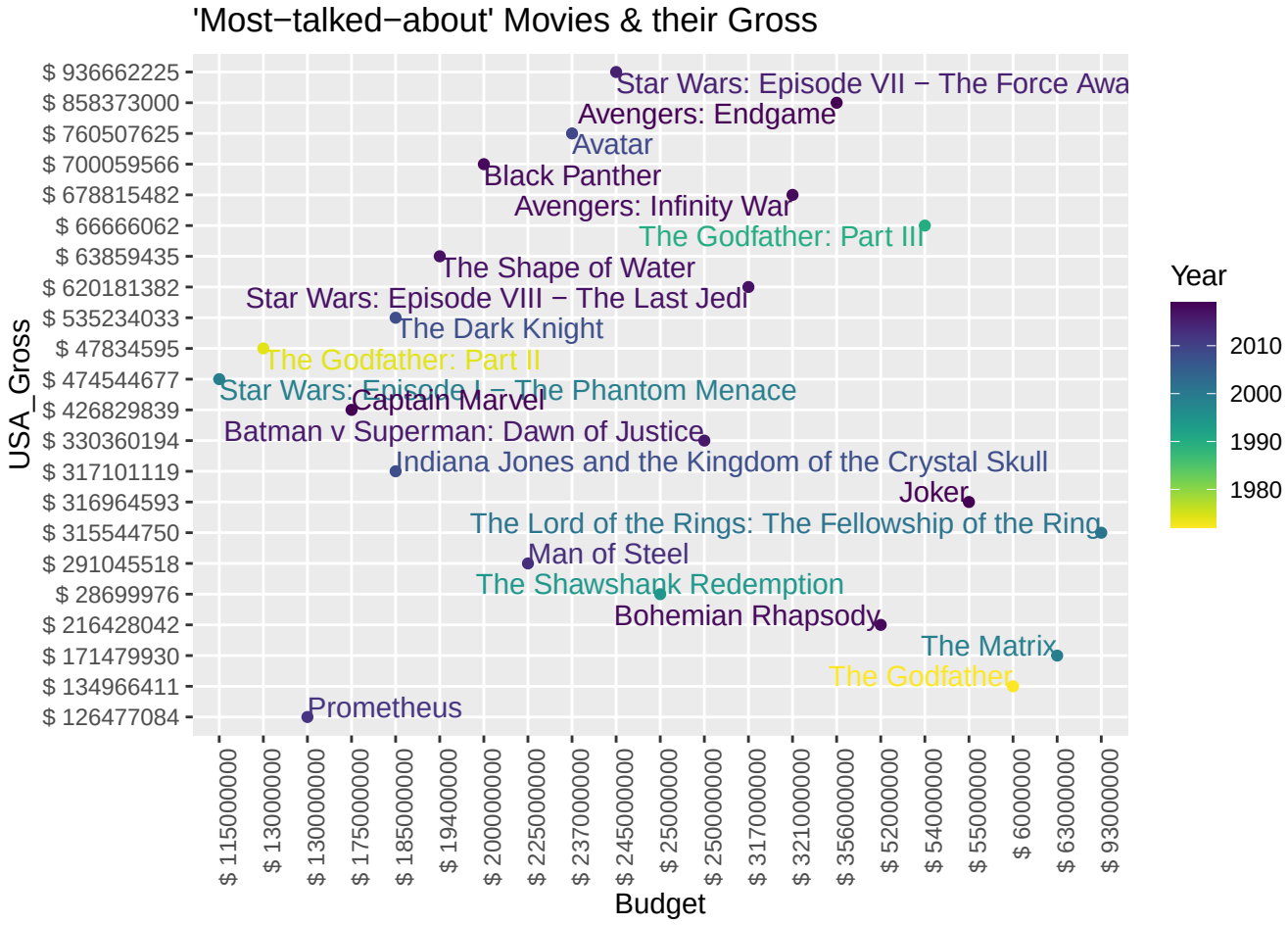

Figure 6: *Some all-time Popular Movies and their Gross*

The figure 6 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 (Garnier (2018)).

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.

References

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

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.

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.