PROJECT 1

Explanations for the coding used are inserted as comments in the R Scripts for each task.

Task 2

Movies are a form of art, transforming over the century from having harmful implications into valuable commodities that promote cultural interests and national identity (Kindem 2000), raise awareness on an array of issues and further possess the ability to influence a society. The movie industry is an integral part of the global entertainment market and boosts a country's economy, as many individuals regard movies as a form of escapism, something to distract them from the day-to-day stress of reality. Certainly, the production of a film in a certain area has a multiplier effect on the local economy factoring from the jobs (actors, directors, technicians, make-up artists etc.) and money it generates (Motion Picture Association n.d.-a). To provide a clearer picture its importance, a research report by the Motion Picture Association (MPA) reported that the American film and television industry alone has more than 122,000 businesses, besides being responsible for 2.4 million job openings and $186 billion in total wages (Motion Picture Association 2021). Having said so, it would be beneficial to analyse movies from different genres to grasp an understanding of their characteristics and success in the movie industry.

In this article, simple analyses are conducted to make comparisons between movies/feature films from two popular genres, namely the romance and horror genres. According to a study by Bioglio and Pensa, the romance and horror genres are the third and seventh most produced movie genres respectively (Bioglio & Pensa 2018). The romance and horror movie data sets are scraped from IMDb, which is one of the most famous movie review platforms out there. The first three pages of each genre, amounting to 150 movie entries each are used to build the data sets. The movies were sorted in descending order of number of votes before web scraping was carried out to ensure that famous movies are analysed. In my personal opinion, the number of votes is one of the important metrics to gauge a movies popularity as it means that the movie has resonated with a huge crowd and generated widespread interest. However, a huge number of votes does not mean that the movie was successful because the reviews could go either way, good or bad. Therefore, the data set consists of other variables such as Title, Release Year, Certification, Runtime (minutes), User Rating, Synopsis, Number of Votes and Gross (in millions of US Dollars), to better understand the differences between the two movie genres. It is also worth noting that a movie can

be labelled with more than one genre tag, the second reason for choosing the romance and horror genres as they rarely overlap. Even so, there were four instances where movies belonged to both genres. After the data cleaning process of removing missing entries and overlapping genres, the romance and horror data sets were left with 139 and 134 entries respectively. Be reminded that the data sets consist of the top 139 and 134 movies from the romance and horror genres based on the number of votes.

To compare between top movies from the romance and horror genres, simple descriptive statistics and exploratory data analysis are carried out on selected variables. Let's first compare and have a look at the movie runtime distribution between genres. Table 1 below displays some important descriptive statistics for the runtime variable of both movie genres. This is followed by a boxplot in Figure 1, graphically displaying the runtime spread.

Table 1 Descriptive statistics for the runtime of romance and horror movies

Genre	Number of Entries	Mean	Standard Deviation	Median	Minimum	Maximum	Range	Skewness	Quartile 1 (Q1)	Quartile 3 (Q3)
Romance	139	118	22.7	118	77	238	161	1.76	104.00	127
Horror	134	107	17.3	105	80	191	111	1.42	95.25	116

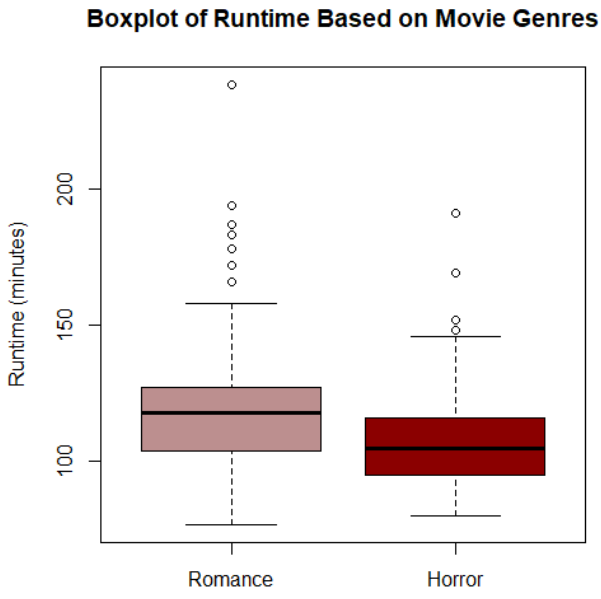


Figure 1 Boxplot of runtime based on romance and horror movies

Looking at Figure 1, there are quite a few outliers for the runtime spread of both genres, especially for the romance genre at a maximum runtime of 238 minutes, which could heavily distort the data distribution. There are no outliers outside the lower inner fences because if a film is too short, it wouldn't be categorised as a movie. It appears that even though both genres have runtimes that are skewed to the right, romance movies are slightly more skewed, which can be explained by the higher number of outliers that are movies with long runtimes. Observing the median, we gather that romance movies have a median runtime of 118 minutes, whereas horror movies have a median runtime of 105 minutes, 11.02% lower. Romance movie runtimes range from 104 to 127 minutes for the middle 50% of data, while horror movies have lower values for both Q1 and Q3. Therefore, 25% of romance movies run longer than 127 minutes, and 25% of horror movies run longer than 116 minutes. Horror movies having shorter runtimes in general is quite a reasonable finding if we think about it, since horror movies are usually focused on the scares, thrill and gimmicks instead of character development. Thus, horror movies would need to spend less time explaining the characters and their backgrounds or relationships as compared to romance movies, where character development of the male and female leads are essential to the plot. Besides, horror movies usually try to keep the audience on their toes with the suspense, and it would be hard to keep the adrenaline or tension going on for prolonged periods.

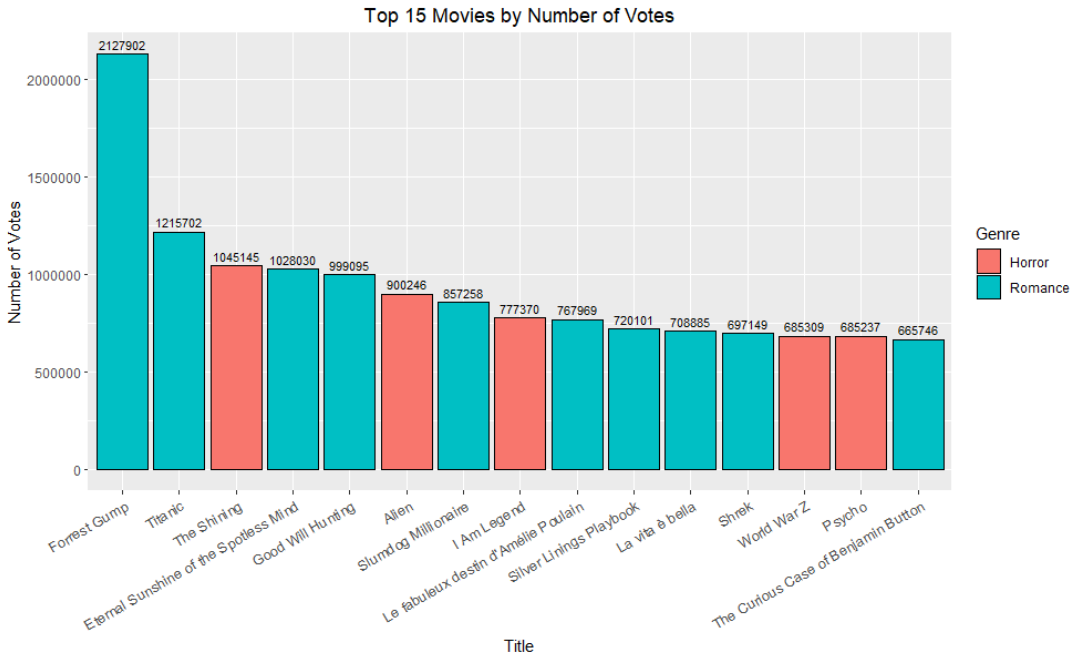


Figure 2 Bar chart of the top 15 movies based on the number of votes, colour coded according to genre

Next, the top 15 movies based on the number of votes are observed, where the bars are also colour coded according to their genres. The first clear observation gained from Figure 2 is that there are more romance movies in the top 15 as compared to horror movies, taking up 10 out of 15 spots. Noticeably, there is a very clear distinction between the first and second spot as Forrest Gump has 2,127,902 votes. This is around 75.03% votes higher than Titanic, one of the greatest movies of all time, following the fact that it was re-released again in theatres worldwide in February 2023 for its 25th anniversary. Having much more votes does not in any way mean that Forrest Gump was the most successful movie, but it does mean that it is the movie that gained the most traction among IMDb users. Look at how the bar charts start to run flat after the ninth or so spot, as movies have almost similar number of votes. Noteworthily, The Shining, a 1980 movie claims it spot as the most voted horror movie with 1,045,145 votes. By only looking at the top 15, the results seem to indicate that romance movies are gaining more buzz amongst IMDb users. This could be due to the demographic of the audience. For example, Statista (2023) states that 77% of female adults in the United States (US) watch romance movies, whereas only 55% of men watch it. Horror movies on the other hand have a wider male audience, where 57% of them watch horror and only 47% of females do. As we are all aware, it is human nature that women are more interested in giving feedbacks and thus cast their votes as compared to men, leading to the higher votes for romance movies. According to Cision Gorkana (2016), a site for comments and reviews called Feefo, found that 58% of customer reviews about a brand were written by women.

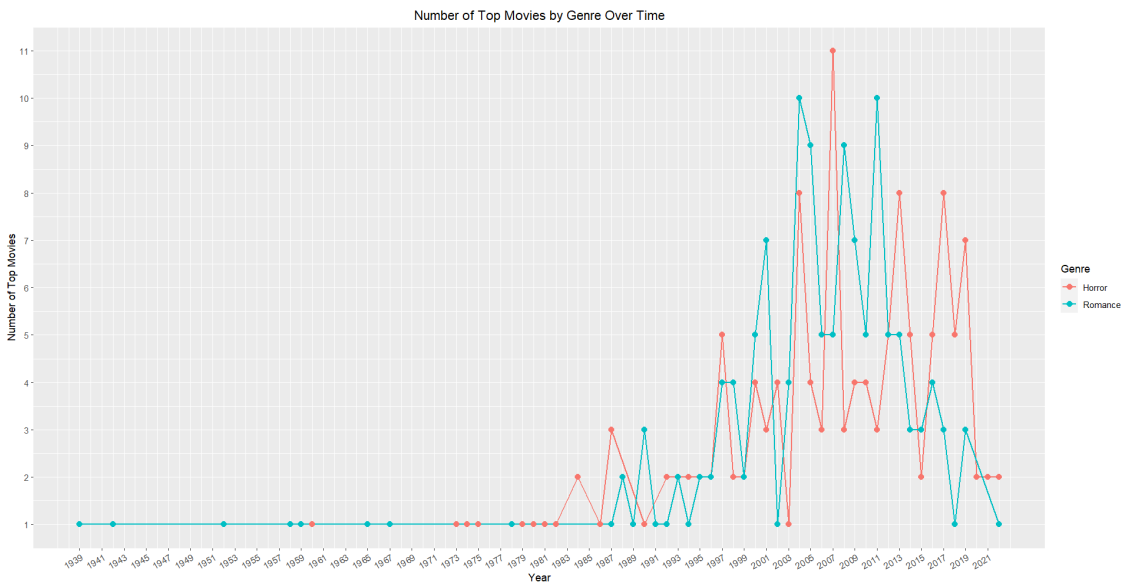


Figure 3 Line plot of the number of top movies released by genre from 1939 to 2022

Figure 3 displays a line plot of the number of top movies released for public viewing by genre over time. Note that the years without any points indicate that no movies were released that year. The earliest romance movie in the top 139 was released in 1939, whereas the earliest horror movie in the top 134 was released in 1960. In these years, only one movie each from the romance and horror genres made it to the current top movies list. Only one movie was getting released year after year, until 1984 and 1988, where two movies were released for the horror and romance genres respectively. Evidently, the years 2004 to 2017 showed the most releases in top movies, especially the year 2011, where a total of 11 horror movies were released. The highest number of romance movie releases on the other hand totalled to 10, in the years 2004 and 2011 respectively. Interestingly, both movie genres showed a steep decrease in the release of movies in the year 2020 onwards. The romance genre had no movie releases at all in 2020 and 2021, with only one release in 2022. For horror movies, there were only two movie releases per year for three consecutive years, taking a huge plunge from its seven movie releases in 2019. The drop in top movie releases can in fact be attributed to the COVID-19 pandemic which took the whole world by surprise. To paint a clearer picture, a study by Kim 2021) mentioned that COVID-19 caused extensive concern among Koreans about going to the movies. As a result, movie demand and ticket sales took a deep plunge, leading to distributors postponing the release of some of their potentially successful films.

Besides that, lockdowns and bans imposed during the pandemic could have caused a complete halt to movie production, leading to the drop in both movie genre releases for three consecutive years.

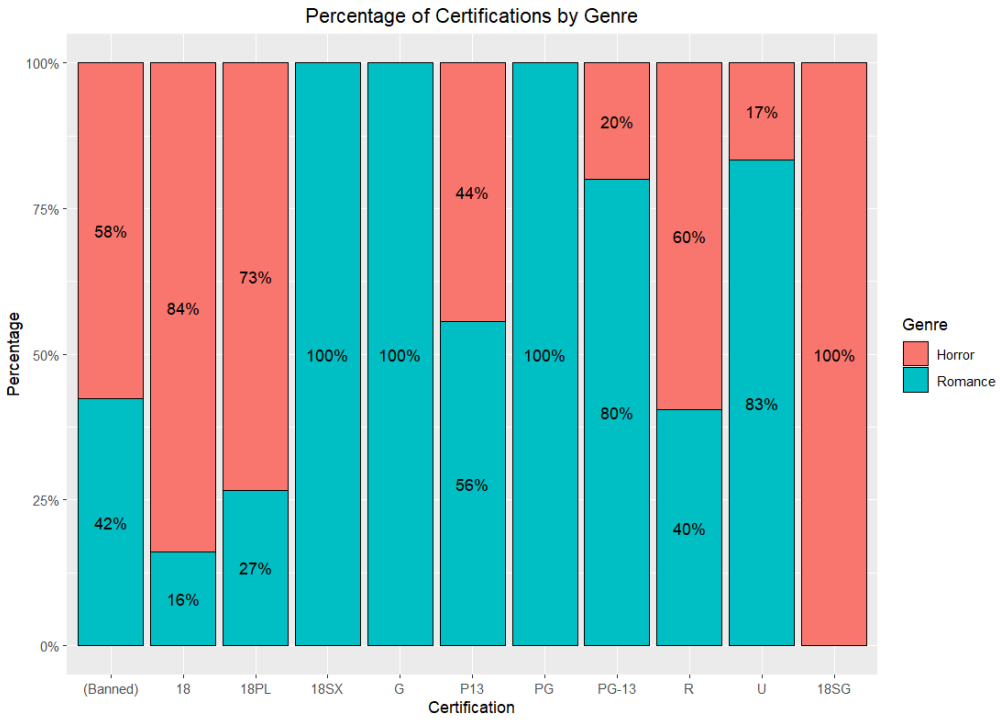


Figure 4 Stacked bar chart of the percentage of certifications by genre

Table 2 A list of certifications and their description

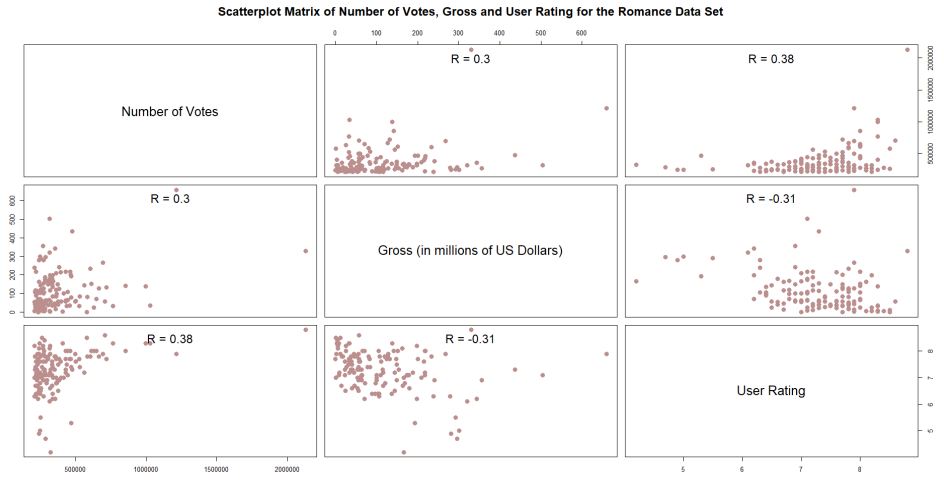
Certification	Description
(Banned)	Actively deemed unsuitable for local distribution
18	Passed only for persons 18 and over
18PL	For 18+ with a combination of two or more elements
18SX	For 18+ with non-excessive sex scenes
G	Suitable for all
P13	Parental guidance required for audiences under the age of 13
PG	Parental guidance
PG-13	Parental guidance suggested for children under 13 years of age
R	Restricted. Children under 17 require accompanying parent or legal guardian
U	General viewing for all ages

18SG	For 18+ with non-excessive violent/horrifying scenes

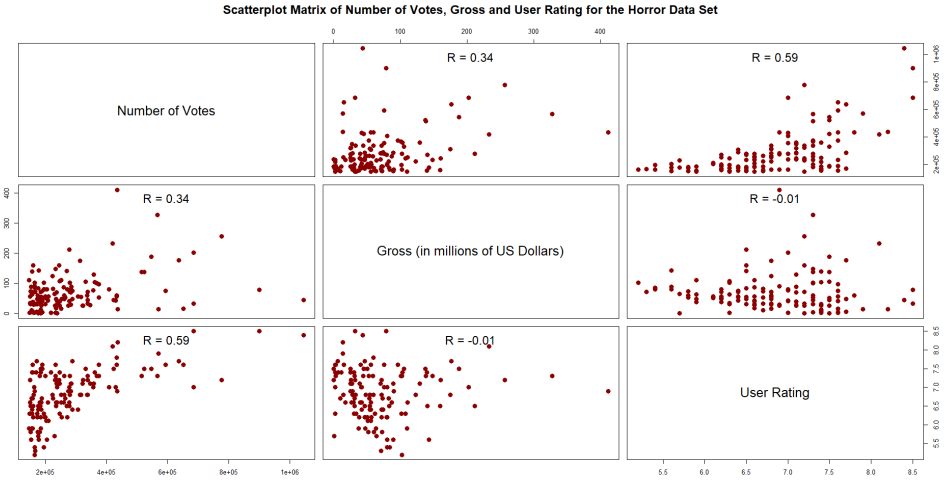
Source: IMDb.com, Inc. 2022

In Figure 4, the percentage of certifications by genre are listed across the eleven certifications available. Additionally, Table 2 provides general descriptions for each of the certifications available. Different certifications could be used for similar audiences because various countries usually have their own content rating systems, each with their own set of unique standards and regulations. Furthermore, the age suitability of movies is determined by organisations like the Motion Picture Association of America (MPAA) (Motion Picture Association n.d.-b) and the British Board of Film Classification (BBFC) (British Board of Film Classification n.d.) based on the substance of the movies. Both these organisations have their own regulations and standards for providing content ratings, therefore it's possible that various movies will receive different ratings under each system.

An initial glance at Figure 4 tells us that most movies with the "18" label are more comprised by horror movies than romance movies. This is especially true for the *18SG* label, where all of its components are horror movies. This does not come as a surprise following the fact that *18SG* movies are labelled that way due to horrifying or violent scenes, something that horror movies are famous for. It is important to keep underaged audiences away from *18SG* movies as it was even studied that children's aggressive reactions are heightened when they watch violent films (Bandura et al. 1963). Romance movies might not be as susceptible to violent scenes as to be labelled *18SG*. Next, *(Banned)*, *18*, *18PL* and *R* movies are also mostly comprised by horror movies. These movies might be too disturbing or offensive for the consumption of teenagers, to the point that some movies were even banned for consumption of the general public. A lot of horror movies are not suitable for kids or teenagers as they are not mature enough to handle the content and it might influence or affect them in the long run, even causing unnecessary harm to their emotional development. Romance movies on the other hand comprise fully the *18SX* label, as we know that romance movies tend to display explicit or sexual content, which could disrupt the social development and moral compass of underage individuals. Interestingly, it can be concluded that romance movies are more suitable for younger audiences, except for the ones rated *18SX*. Nevertheless, most of them are rated *U*, *G*, *P13*, *PG* and *PG-13*. On the contrary, horror movies may be more suitable for matured audiences due to disturbing content.



(a)



(b)

Figure 5 Scatterplot matrices of the number of votes, gross and user rating for (a) romance movies and (b) horror movies

Moving on to the final analysis in comparing movie genres, scatterplot matrices of three variables, namely User Rating, Number of Votes and Gross (in millions of US Dollars) are observed. Figure 5 shows two scatterplot matrices, one for each of the romance and horror genres respectively along with their Pearson correlation coefficient values. This can help us to find any interesting patterns in the data. The strongest correlation in both the genres can be seen between the Number of Votes and User Rating variables, with values of 0.38 (low positive correlation) and 0.59 (moderate positive correlation) for the romance and horror genres respectively. Apparently,

when the number of votes increases, the user ratings for the movies also increase. Although, this does not imply a causal relationship and external factors should be taken into consideration for a more thorough analysis. The positive relationship for the horror movies is much stronger, implying that as more people become aware of the movies and vote on them, the ratings converge to a more "true rating" since the sample size is much bigger. The ratings are more representative of the movie's true quality with higher votes, hence leading to the increase in ratings. With lower votes, the individuals could be biased and mess up the ratings, as can be seen in the lower left panel, where the ratings of movies with lesser votes are more scattered.

The relationship between the number of votes and gross for both movie genres is positive and similar, albeit it being quite weak. This indicates that there is a tendency for the gross to increase as the number of votes increases but the degree of this relationship is not very strong. Many movies with a high number of votes still have a low gross. Furthermore, notice how some movies that have a low number of votes possess high gross, which could be due to powerful marketing strategies. There could be other factors at play such as the popularity of the actors, time/season of the release date, competition and so on that could also contribute to the increase or decrease in gross. Nonetheless, there are still tendencies for movies with high buzz among users to gain more ticket sales as more people are aware of it through word-of-mouth and therefore curious to watch it. Lastly, the relationship between user ratings and gross is quite unexpected as both genres show negative correlations. The correlation of -0.01 for horror movies however is negligible, as it just indicates that there is no consistent relationship between the two variables. Conversely, there is a low negative correlation for romance movies amounting to -0.31. One would expect the gross to increase as user ratings increase but that is not the case for romance movies. A possible explanation could be the fact that romance movies tend to be released during special occasions such as Valentine's Day or Christmas, leading to high competition as many romance movies are released at the same time. This could lead to less movie goers for a certain movie but high ratings as a small number of individuals enjoy the movie they watch and rate them. On the other hand, some movies with low ratings can gain high gross maybe due to promotional strategies and the likes of it.

Also worth noting, from the scatterplot matrices, it seems that more romance movies have a higher gross. By computing the gross median for both genres, we obtain $83.82M and $54.90M

for romance and horror movies respectively. The medians for user ratings on the other hand are 7.3 for romance and 6.85 for horror. We could say that romance movies are more successful in terms of their gross box office earnings by a big margin. In terms of user ratings, romance movies also fare better, which could be due to their appeal and strong emotional impact. Not to mention, romance movies are more age friendly than horror movies as even eager teenagers could watch them. As stated earlier, more women would leave reviews for romance movies and thus lead to a more genuine user rating. This could be one of the reasons why more romance movies are produced compared to horror movies (Bioglio & Pensa 2018), since there is a bigger chance to receive higher earnings and popularity.

In conclusion, the analyses carried out on movies from the romance and horror genres have revealed some riveting insights on variables such as movie runtime, release year, number of votes, certifications, user ratings and gross. Firstly, it is clear that the length of a movie is an important factor in shaping the genre of movies, as horror movies tend to have shorter runtimes than romance movies. This can be attributed to a horror movie's intention to intensify the impact of scares, which might not be as effective over longer periods of time. Next, it is observed that there are more romance movies in the top 15 (based on number of votes) than there are horror movies, mainly due to the demographics of the audience. As more women enjoy romances and men enjoy horror, the number of votes for romance movies could be higher as women are keener on leaving their reviews and opinions on the movies they watched. Interestingly, the number of top movie releases for both the romance and horror genres dropped in the year 2020 onwards, right around the time the COVID-19 pandemic hit. Lower movie sales, delayed releases and halt in movie productions are big factors caused by the pandemic.

Besides that, the analysis reveals that horror movies are usually targeted at mature audiences due to violence, horrifying and disturbing content, which is not suitable for younger individuals. Romance movies are more age-friendly except for movies that are labelled for their explicit sexual content. It was also noted that the correlations between the number of votes, gross and user ratings vary between genres. Horror movies have a much stronger positive correlation between the number of votes and user rating than romance movies, whereas romance movies have a more negative correlation between user ratings and gross. However, both genres show similar correlations for the number of votes and gross. Moreover, simple median computations revealed

that romance movies are more successful than horror movies in terms of higher gross earnings and user ratings. Many factors could come into play leading to these findings, such as marketing strategies, famous stars, movie quality, competition and so on. These factors can be analysed further to gain in depth insights.

All in all, this article highlights the distinctive qualities of the romance and horror movie genres, besides emphasising on the importance of taking various variables and factors into account to better comprehend the differences. The findings of this article are more on the surface level as it only involves exploratory data analysis (EDA). To further understand which genre is more successful, more robust and complex techniques should be implemented on a myriad of other variables.

Task 4

A word cloud of song lyrics is a graphical representation of the most frequent terms appearing in a song, where more frequent terms are displayed bigger in size. Word clouds may help to give a rapid overview of the textual contents of several songs, including the intentions of the musicians, since analysing lyrics can be a tedious and time-consuming task (Burch et al. 2016). In Task 4, word clouds are generated based on three songs with differing themes, namely "The Scientist" by Coldplay, "Better Than Revenge" by Taylor Swift and "Royals" by Lorde. "The Scientist" is about heartbreak and wanting second chances in a relationship, as a man who is deeply intertwined with his work neglects his girlfriend. When he realises how important she is in his life, he wants to start over with her. Next, we have "Better Than Revenge", very revenge themed as the lead singer warns the woman who stole her boyfriend from her. Taylor Swift reminds her rival that there is nothing she does better than getting revenge. Lastly is "Royals", a song that mostly criticises materialism and wealth. Lorde emphasises on how people can be happy in their own fantasy and simplicity even without wealth. Table 3 below lists down terms that appear at least five times in all three song lyrics, along with their frequencies in descending order. Figure 6 is the word cloud generated from all the song lyrics, where terms with the same frequency have the same colour.

Table 3 Terms that appear at least five times in all three song lyrics and their frequencies in descending order

Term	Frequency	Term	Frequency	Term	Frequency	Term	Frequency
Rule	18	keep	6	royal	6	never	5
think	10	kind	6	said	6	run	5
better	8	let	6	thing	6	start	5
know	7	love	6	back	5	tell	5
come	6	mind	6	care	5		

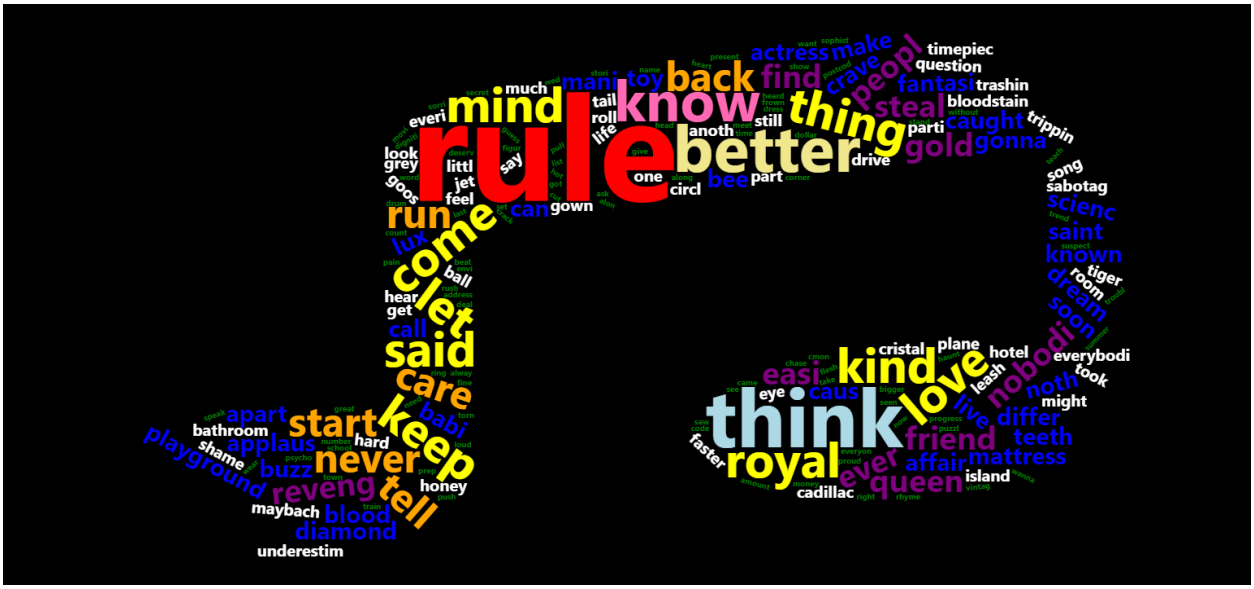


Figure 6 Word Cloud of three song lyrics, namely "The Scientist", "Better Than Revenge" and "Royals"

We might think that the most frequent terms in a song would be their title, but this is not the case for the three songs analysed. This is especially true for the song "The Scientist", as the term *scienc*, the word for science, only appears three times in the lyrics (blue in the word cloud). It could be that "Scientist" is only used as a creative expression in the title by the songwriters because scientists are usually very busy individuals who tend to ignore the people around them. The term *reveng*, for revenge appears four times, whereas the term *better* appears eight times, the third most common term. On the other hand, the term *royal* appears only six times (in yellow). However, we notice how the term *rule* appears most frequently (18 times), which is very relevant to the song "Royals" as it indicates the ruling of royalty. *Rule* appears many times in the chorus of the song. This term portrays the controlling tone of "Royals" as Lorde repeats on and on again about how she wants to rule and take charge of her own simplistic life. Next, both the terms *think* (10 times) and *better* are from the song "Better Than Revenge". These terms represent how Taylor Swift constantly bad-mouths, criticises and is sarcastic to her former boyfriend and rival. *Better* also indicates empowerment and dominance as Taylor shows superiority by wanting to take revenge on them.

Evidently, the song "The Scientist" is drowned by the other two songs as one of its most used terms is *said*, appearing only six times. Overall, the emotional tone dominating the three songs is mostly aggressive, controlling, critical and sarcastic, dominated by the two female singers. The lyrics are most likely fuelled by feelings of hatred, hurt or betrayal. Unfortunately, the sad and melancholic tone of Coldplay's song does not stick out. Even so, it could just mean that "The Scientist" uses a variety of terms without much repetition. We do see positive terms such as *kind* and *love*, but they should be understood in their full context as *kind* is actually used to portray "type" in "Royals", whereas half of the time the term *love* is used to describe a love affair. Generally, many terms are used in a negative context as two out of three songs lean towards negativity.

Table 4 Most frequent terms in each of the three song lyrics in descending order

The Scientist by Coldplay		**Better Than Revenge by Taylor Swift**		**Royals by Lorde**	
Term	**Frequency**	**Term**	**Frequency**	**Term**	**Frequency**
said	6	think	10	rule	18
back	5	better	8	kind	6
tell	5	keep	6	let	6
come	4	mind	6	royal	6
easi	4	thing	6	care	5
nobodi	4				
start	4				

(a) (b) (c)

Figure 7 Separate word cloud of three song lyrics, namely (a) "The Scientist", (b) "Better Than Revenge" and (c) "Royals"

In order to compare between the songs more easily, multiple word clouds are constructed, one for each song. The results obtained are similar to what has been explained earlier. In the first word cloud for "The Scientist", the term *said* appears the most, a total of six times. This term portrays how the singer goes on and on again about how his feelings of heartbreak are hard to handle. The terms *back* and *tell* on the other hand shows his desperation in wanting to go back to the start of their relationship, besides telling his significant other how sorry he is. Even the terms *come*, *easi* (easy), *nobodi* (nobody), *start*, *love*, *hard* and *shame* are heavily themed with heartbreak and second chances. Besides the terms explained earlier, *keep*, *mind* and *thing* are the third most common terms in Taylor Swift's "Better Than Revenge". Taylor constantly repeats the line "She should keep in mind", a warning of vengeance to come for the girl who stole her boyfriend. *sabotag* (sabotage) and *underestim* (underestimate) also portray feelings of revenge. The word cloud shows how revengeful and even sarcastic (terms like *saint* and *actress*) Taylor is in putting down her enemies. The final song "Royals" has a big gap in the frequency of its first and second most common terms. The margin is so big that the song just completely emphasises on themes of class and social status. The criticism of wealth by Lorde is further portrayed by terms such as *diamond*, *lux* (luxe or luxury), *maybach* (German luxury car brand) and *cadillac* (American luxury car brand) among others. Also worth noting, terms like *care*, *differ* (different) and *fantasi* (fantasy) show that people like Lorde should not care about wealth and that they crave a different kind of fantasy to be happy. Clearly, all three word clouds show differing themes through their most frequently used terms. Although, it can be said that the contrast between "The Scientist" and the other two songs are more apparent as there is not much tones of sadness in "Better Than Revenge" and "Royals. To sum up, each song focuses on heartbreak and second chances, revenge and sarcasm, besides criticism of material wealth, social classes and authenticity respectively.

In general, many terms in the overall word cloud are used in a negative context as two out of three songs lean towards negativity. Word clouds are able to provide readers with a gist of what a song is all about before they listen to it. This provides a speedy way to grasp the prominent themes of songs that might not be apparent at first. They can also be used to compare between songs, identifying key themes and any similarities or differences by constructing multiple word clouds. As shown in this task, separate word clouds for each song help us to clearly see the themes of the songs and distinguish between the works of different artists. Nevertheless, it would be a good thing for readers to understand the context of the songs in order to better comprehend the

emotional tone expressed by singers in the lyrics. This is due to the fact that certain terms may be used in a whole different context than what is imagined, as was portrayed by this simple analysis. In conclusion, word clouds can be used as effective instruments to analyse and make sense of song lyrics by parties such as researchers, music enthusiasts and so on, who are intrigued by music.

References

Bandura, A., Ross, D. & Ross, S.A. 1963. Imitation of film-mediated aggressive models. *Journal of Abnormal and Social Psychology* 66(1): 3–11.

Bioglio, L. & Pensa, R.G. 2018. Identification of key films and personalities in the history of cinema from a Western perspective. *Applied Network Science* 3(1): 50.

British Board of Film Classification. n.d. British Board of Film Classification (BBFC). https://www.bbfc.co.uk/ [4 May 2023].

Burch, M., Fluck, T., Freund, J., Walzer, T., Kloos, U. & Weiskopf, D. 2016. Lyrics word clouds. *Proceedings of the International Conference on Information Visualisation*, pp. 51–56.

Cision Gorkana. 2016. Women aged 25 to 34 "most likely" to post customer reviews. https://www.gorkana.com/2016/08/women-aged-25-to-34-most-likely-to-post-a-customer-review/ [4 May 2023].

IMDb.com, Inc. 2022. IMDb | Help Center. https://help.imdb.com/article/contribution/titles/certificates/GU757M8ZJ9ZPXB39?ref_=helpart_nav_27# [5 May 2023].

Kim, I.K. 2021. The impact of social distancing on box-office revenue: Evidence from the COVID-19 pandemic. *Quantitative Marketing and Economics* 19(1): 93–125.

Kindem, G.A. 2000. *The International Movie Industry*. G. A. Kindem (ed.). Carbondale: SIU Press.

Motion Picture Association. 2021. The American motion picture and television industry | Creating jobs, trading around the world.

Motion Picture Association. (n.d.-a). Driving economic growth. https://www.motionpictures.org/what-we-do/driving-economic-growth/ [30 April 2023].

Motion Picture Association. (n.d.-b). Welcome to FilmRatings.com. https://www.filmratings.com/ [5 May 2023].

Statista. 2023. Most popular movie genres among adults in the United States as of December 2018, by gender. https://www.statista.com/statistics/254115/favorite-movie-genres-in-the-us/ [5 May 2023].