Course: DS 5610
Team Number: Group 9
Team Member: Kelly Chen, Cameron Rondeau, Henry Savich, Zhuoyi Zhan

IMDB Movie Data Analysis

Executive Summary

IMDB movies are abbreviated from the Internet movie database, and it actively gathers information from filmmakers and studios. It is currently the most popular and authoritative source for movies and media content. We hope through our exploration, we can have a deeper understanding of the movie industry and provide recommendations and implementation feedback to filmmakers. The main goal of this data analysis is to find out the factors that contribute to a high-rated movie.

After viewing different movie datasets including Netflix and global movie datasets, we finally locked on the IMDB movie dataset because it contains all the necessary variables we want to explore, and we also believe that its authority and authenticity enables our data analysis conclusions to be more reliable and persuasive.

In the data analysis, we explored our data through two main topics: movie content and movie production process. We developed two branches under each of the topics to dig into the data and circle around our main question: What contributes to a high-rated movie? With the information and summary we explore with the datasets, filmmakers could have a comprehensive understanding of how the movie would be rated, and they can also switch their emphasis during movie production under the consciousness that a certain factor may add more weight to the ratings. The findings and conclusions from the data exploration could help filmmakers and movie industry professionals frame their decisions and directions during the movie production process.

Dataset description

The dataset we are exploring is the IMDB movies extensive dataset from Kaggle which is uploaded by author, Stefano Leone. IMDB is the most popular database of information regarding movies and TV series. It aggregates many attributes such as movie plot description, cast, production crew, ratings, and user and critic reviews. Stefano Leone has put together this four-part dataset with 86955 movies that have more than 100 votes. The IMDB movies dataset contains 86955 movies and attributes such as genre, year, country, average rating, movie description, budget, etc. The IMDB names dataset contains 297705 cast members with personal information like name and birth year. The "IMDB ratings" file contains ratings for the 85855

movies with demographic dissection. It shows average voting for different age groups and genders. The IMDB title_principals dataset links movie titles with crew roles in movies.

We analyzed three of the four IMDB datasets: IMDB movies, IMDB names, and IMDB title_principals.

The columns that we are interested in IMDB movies are:

- **Imdb_title_id**: title id on IMDB
- **Year**: year of release(between 1894 to 2020)
- **Genre**: movie genre(comedy, Drama, Romance, Action, etc)
- **Country**: movie country
- **Language**: movie language
- **Description**: plot description
- **Avg_vote**: average vote (between 1 to 9.9)
- **Budget**: budget(in each country's currency)

The columns that we are interested in IMDB names are:

- **Imdb_name_id**: name ID on IMDB
- **Name**: cast member name

The columns that we are interested in IMDB title_principals are:

- **Imdb_title_id**: title ID on IMDB
- **Imdb_name_id**: name ID on IMDB
- **Category**: category of the job done by the cast member(actor, actress, director, producer, etc)

Data cleaning process

We did not combine the four files because they have different variables. When we want the values from more than one file, we join them with keys like imdb_title_id and imdb_name_id. The missing values are already expressed as NA so we handle the missing value by omitting the null value. To note that genre is a special variable that needs transformation. A movie can have multiple genres and that are separated by space. To explore individual genres in the subsequent analysis, genres columns are split up to have one genre in a row.

To facilitate our analysis on the relationship between actors, budgets, and average ratings of movies, we filter the column "country" for "USA" to select USA movies and the column "language" for "English" to select English movies. Since movies only in the USA do not require conversion of currency and can be easily adjusted for inflation over the year. We also set the

standard of highly rated movies with average votes higher than 8.0 because the IMDB rating is in ascending scale from 1 to 10.

Variable Transformations

Actor Rating

In order to examine the relationship between specific actors and the ratings of the films they were in, we needed to design a variable that describes the distribution of ratings of a certain actor's films. Described generally, the actor rating we used is the mean average rating of the movies an actor starred in, but there were more choices involved in formulating the metric. First, on a movie-by-movie basis, we had to decide what metric best described how good a movie was. Our choices for this were either the average rating, the median rating, or a percent of ratings above a certain threshold. The reason we chose average rating is because we felt it was important to account for the differences between a 9 and a 10, or a 1 and a 2, which a median or a threshold wouldn't do.

To connect actors to movies, we had to use the join functions from the tidyverse package to relate the title_principals.csv, names.csv, and movies.csv tables. Once we decided to use the *avg_vote* variable for individual movies, we then had to decide how we would aggregate the ratings across an individual actor. Again we had similar choices, in that we could take the mean o**r** median of *avg_vote*, or report the % of movies with *avg_vote* above a certain threshold. Here we discarded the idea of a threshold because we thought it was most intuitive that actor ratings would be on the same scale as movie ratings. As it turned out, because the observations were already generated by taking an average, the distribution of *avg_vote* for an actor was fairly symmetrical, so whether or not we used the mean or median didn't make much of a difference. Ultimately we decided to use the mean to present a metric generated from all the data. The end result is that actor rating is an *average of averages*.

Adjusted Budget

Another important question we wanted to investigate was how the decision of how much to invest in a movie affected it's reception. As movie budget was already a variable in the movies csv, it may seem like a simple question of generating a correlation between budgets and ratings, but there is a complication. The IMDb contains movies dated back to the 19th century, so due to inflation we can't compare movie budgets (with any real implication) across the data set. To solve this problem we instead used the budget adjusted to 2020 dollars. To do this we used a package called PriceR. Once we had adjusted the budgets, we were able to use standard numerical methods to detect correlations between budgets, movie ratings, and actor ratings.

Genre Analysis

We want to know if the genre of a movie affects the rating. For example, are certain genres good at getting a higher rating? To understand the relationship between genre and rating, we can look at the distribution of different genres. In this analysis, the variables needed are genre and average vote in IMDB movie.csv. There are no missing values in both variables. In the genre column, each value contains all genres of a movie.
In order to explore whether individual genres can affect the rating of movies, we need to split up rows with more than one genre so that all genres of a movie are listed separately.

Genre	Total count	Average vote
Drama	13827	5.93
Comedy	9972	5.68
Romance	5131	6.04
Horror	4590	4.49
Thriller	4382	5.01
Action	4273	5.25
Crime	4267	5.81
Adventure	2854	5.71
Mystery	2012	5.59
Sci-Fi	1774	4.76
Family	1285	5.66
Fantasy	1175	5.34
Western	1127	6.00
Musical	902	6.14
Music	810	6.03
Biography	704	6.14
War	647	6.14
Film-Noir	647	6.64

Sport	492	5.88
History	439	6.37
Animation	432	6.26
Documentary	1	7.50
News	1	6.40

Figure 1

After filtering for movies in the USA and grouping the movies by genre, there are a total of 23 genres. The most produced genre is Drama with a total of 13827 movies while there is only one movie in the genre of documentary and news. The genre documentary has the highest average vote, 7.50, and horror movies have the lowest average vote, 4.49. The two have a rating difference of 3. Since genres have a large difference in average voting, we can say that genre is an important factor in deciding the rating.

For visualization, we used the "ggplot" library in R. To get a better view of vote distribution, we selected the ten most produced movie genres. This graph of grouped violin plots with stacked box plots shows the vote distribution of the ten most produced genres. We are able to see the range of average votes and the size of the vote number.

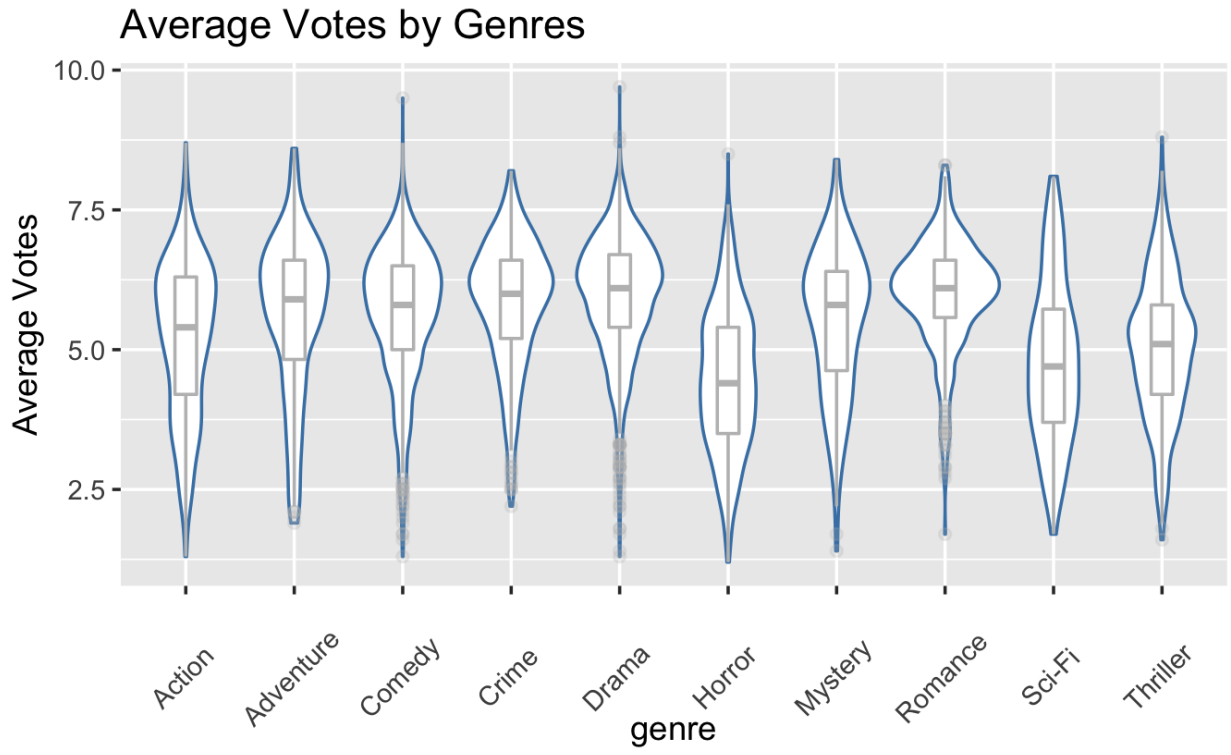


Figure 2

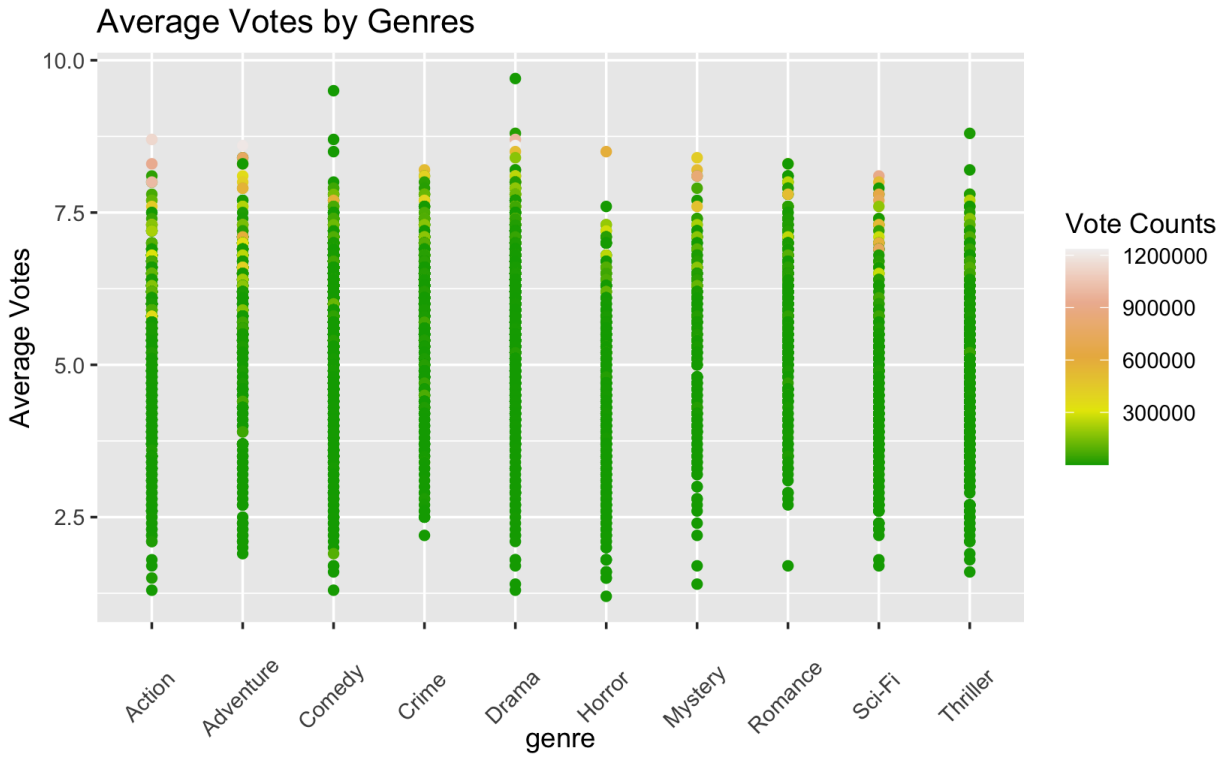


Figure 3

Figure 2 shows that among the ten most produced genres, Drama and Comedy have larger spans in the score. From the shape of the violin plot, action, adventure, comedy, crime, drama, mystery, and thriller have received much more votes beyond this genre's median. Romance has the highest average vote and Horror has the lowest average vote. Comedy, crime, drama, romance have more outliers at the lower rating portion than at the upper rating portion. Figure 3 is a dot plot showing the distribution of votes by genre with color indicating the number of people who voted. It is more obvious that high-rated movies have more people voting. Observed from the color of the dots, action, adventure, mystery, drama, and Sci-Fi have more movies that are rated by more than 60000 users.

IMDB user voting is on a scale from 1 to 10. Normally, people consider movies that are rated higher than 8.0 as excellent, between 6.0 to 8.0 as good, and below 6.0 as dissatisfactory. The movie genres are produced disproportionately. Drama movies have the largest quantity and it has the most movies that are rated 8.0. But we want to learn more about the percentage of movies higher than 8.0 in all genres in order to see if any genre outperforms the others and is better at scoring high ratings. Which genre of movie has the highest percentage of rating greater than 8.0?

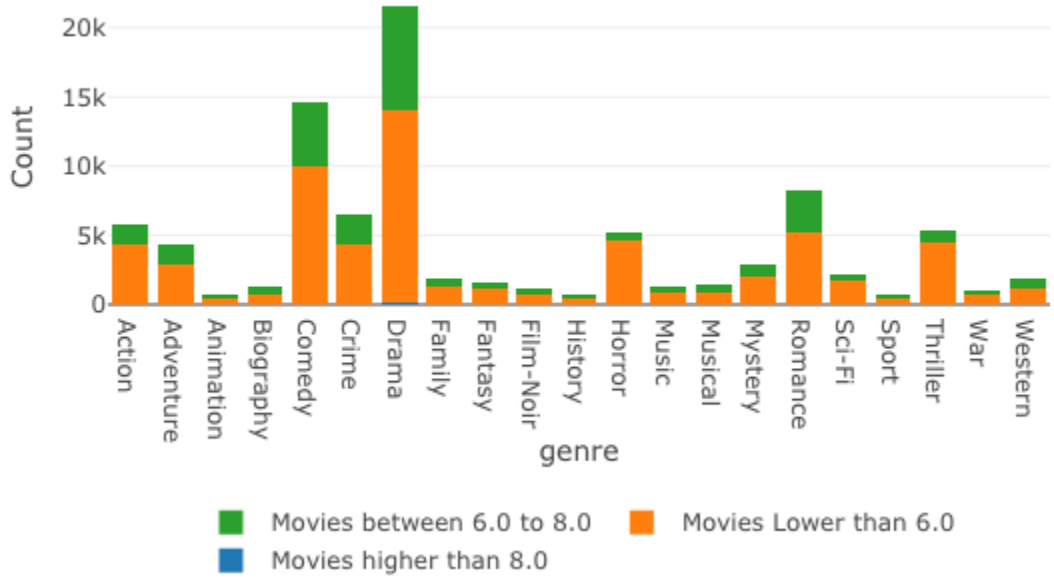


Figure 4

Figure 4 shows that all genres have most of its movies voted under 6.0. The percentage of movies vored higher than 8.0 is so insignificant that can hardly be seen from this visualization. Only at the bottom of the Drama column can we see a little of blue that indicates voting more than 8.0.

So we filter out the movies that voted more than 8.0 in each genre and calculate the percentage of movies higher than 8.0 in all the genre categories. Then we used bar charts to visualize the ranking of percentages.

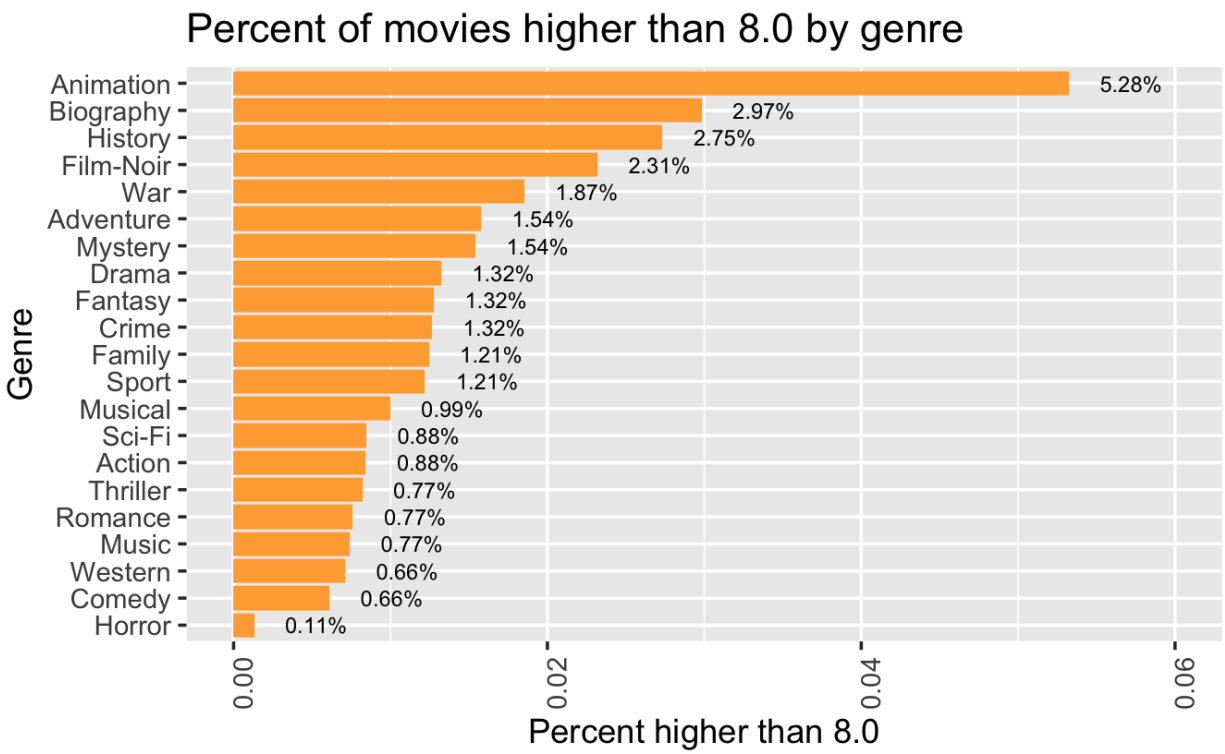


Figure 5

It is shown in Figure 5 that Animation has the highest percentage of movies rated higher than 8.0, which is 5.28%, followed by biography, history, film-noir. We can see that they are not the most produced genres. shown in figure 1, they are the least produced in all genres but they have a higher percentage of movies being excellent. Horror movies have the lowest possibility of rating above 8.0 and are much lower than other genres. Comedy and Romance, the second and third most-produced genres, only have 0.66% and 0.77% of great movies.

To sum up, the genre is a good contributor to movie ratings because genres have differences in their average rating. Drama movies have the highest average rating and horror movies have the lowest average rating. Action, adventure, mystery, drama, and Sci-Fi have received more users' votes than other genres. Animation has the highest percentage of movies rated above 8.0 while horror movies have the lowest percentage of movies above 8.0.

Sentiment Analysis

To get a deeper understanding of the content of the movie, we took a step into the movie's description and attempted to analyze the sentiments involved in the top-rated movie to see if certain sentiments can contribute to a higher score. Sentiments are crucial elements in the movie contents, and they are the most direct emotions that represent the movie. Through the sentiment

analysis, we could know if certain emotions could help attribute to a higher rating than other sentiments. With the information in their head, filmmakers and movie professionals could find out a direction that most likely contributes to excellent artwork.

The first step in sentiment analysis that we did is to tokenize the movie description. We selected IMDB movie id and movie description and used the unnest_token function to get the separate word of each movie description. We also assured that the IMDB movie id is a unique variable with no repetitive title ids by employing the assert function. Once we tokenized the description, we began to remove the stop words from the word tokens, since we don't want our produced conclusion to contain too many pronouns or meaningless words to interfere with our analysis. After removing the stop words, the next step is to get the sentiment data. We employed the nrc lexicon to help us categorize all the tokenized words from the movie description. The nrc lexicon is a text-sentiment package from the tidytext package which could categorize words in a binary fashion into different emotions and sentiments. The next step is to employ a left join to help us categorize the movie description tokenized words. After combining the data, we then use count to see how many of the same sentiments had been conveyed by the movie description. The ggplot helped us visualize the top sentiments contained in the top-rated movie by plotting the bar chart.

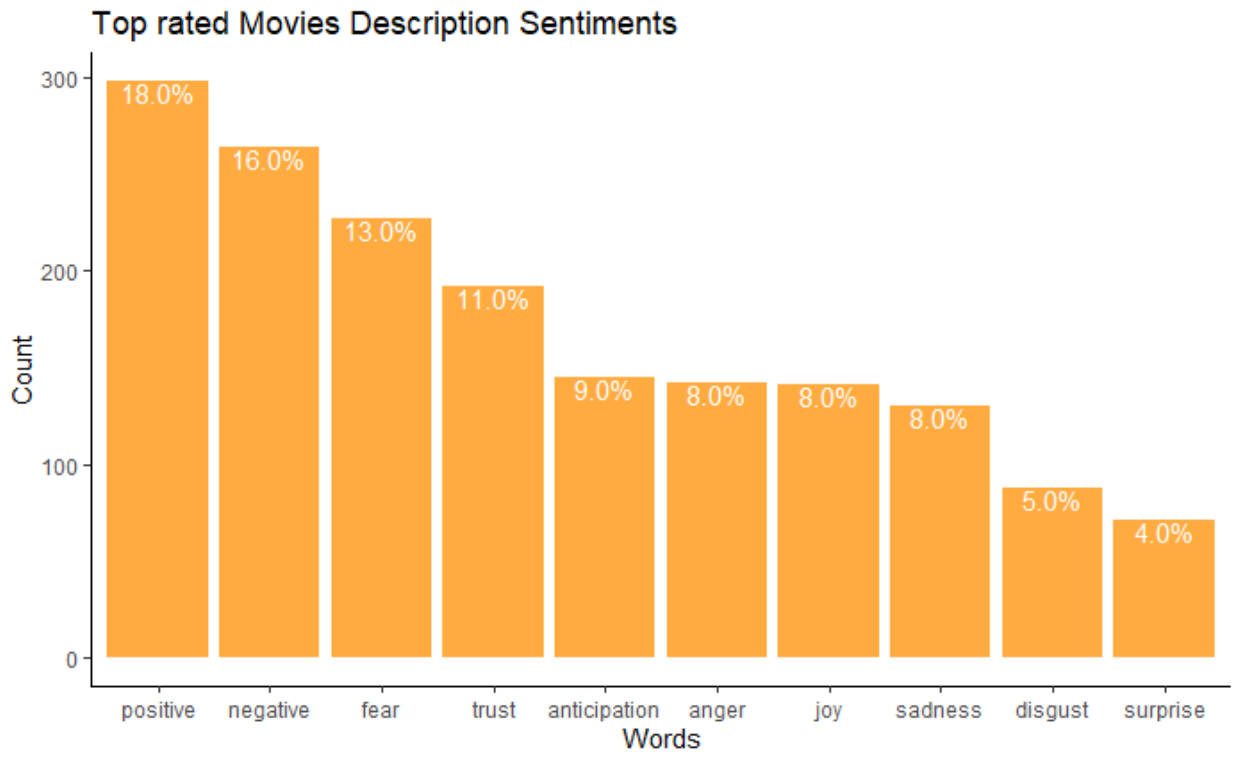


Figure 6

From the bar chart (Figure 6), we can see that the most common sentiment in the top-rated movie is positive sentiments (18.0 %). The positive emotions take up the most positions in all the movie descriptions, while the negative take up 16 percent. We can conclude that the more positive sentiment that a movie description conveys, the more likely the movie is a high-rated movie. The top three categories of sentiments that we would suggest the filmmakers include in their films would be: Positive, Negative, and Fear. Whether the sentiment is a direct positive or a direct negative, we can see that high-rated movies tend to have a strong trend in the positive sentiments. The difference between the positive and negative sentiments is about 2 percent, so we cannot conclude if negative sentiments would bring down the overall score of the movie. What we do encourage for the filmmakers is that to convey a strong emotional tie to the movie and also include those feelings into the movie description, with a positive attitude of the movie content, the movie is more likely to get a higher score.

Common words with different genres

Since we are digging into the genre analysis, we would like to see what kind of words appear most with different genres. First, we would like to see the most common words in all movie descriptions and check if a specific genre contains more relative words. The way we did is just employing the data we have cleaned after anti joining the stop words and making a count of each word. In ggplot2 package, we also group the results in a bar chart by their genre types. Since most of the movies are categorized as “Drama”, the most common description words have been taken large positions by the drama movies. The most common word in a high-rated movie is “life”, with 16 frequencies from the drama movie and 6 frequencies from the romance movie. The top three most common words in the high-rated movie are “life”, “woman”, and “love”.

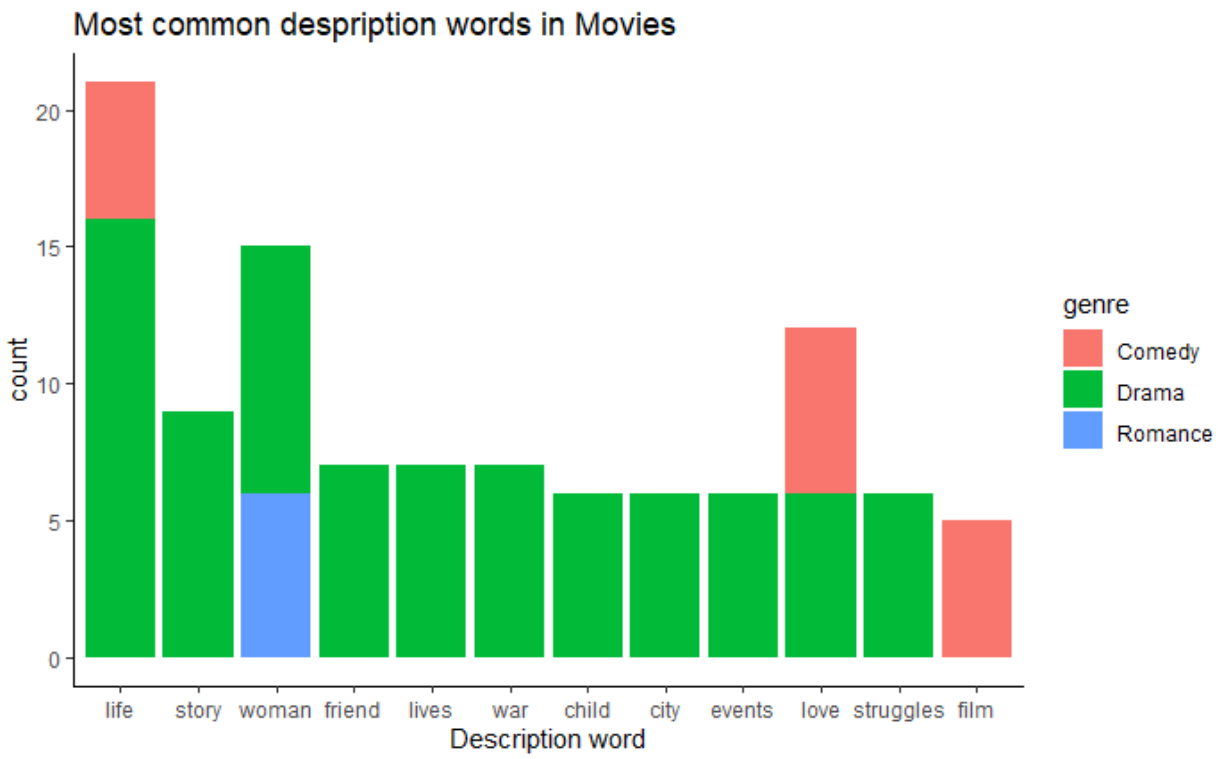
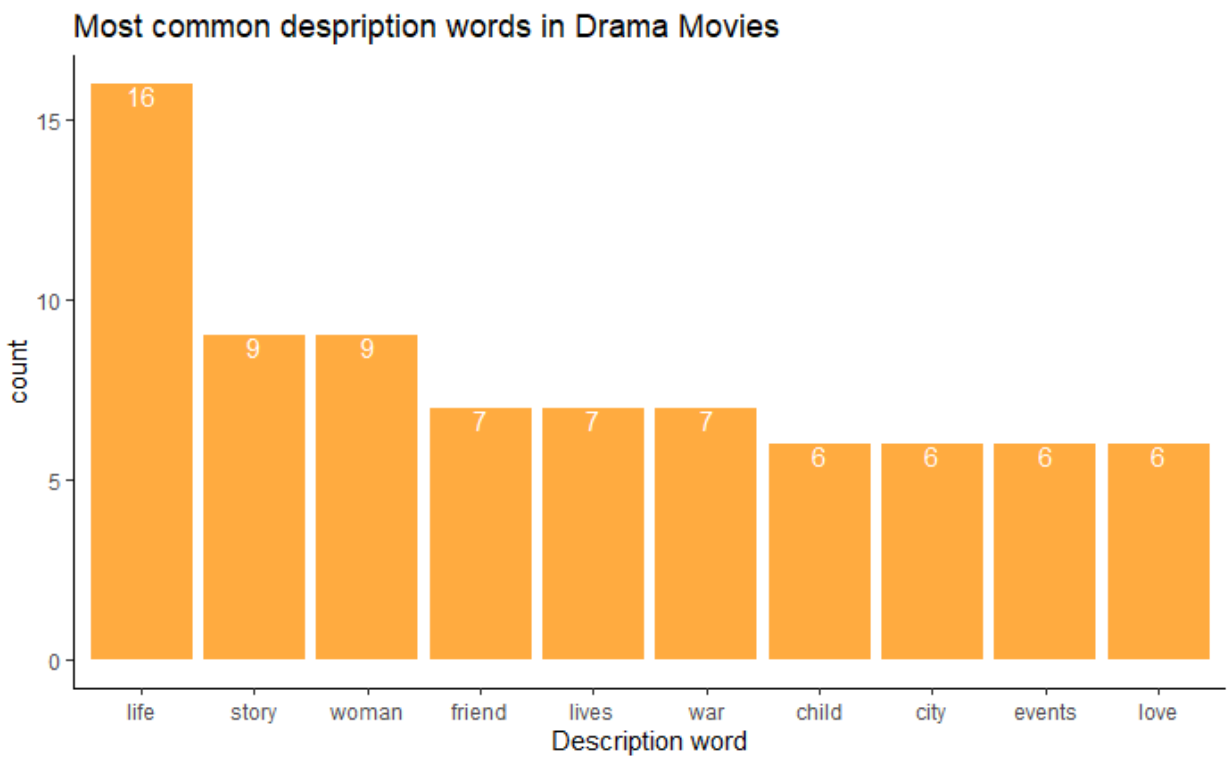
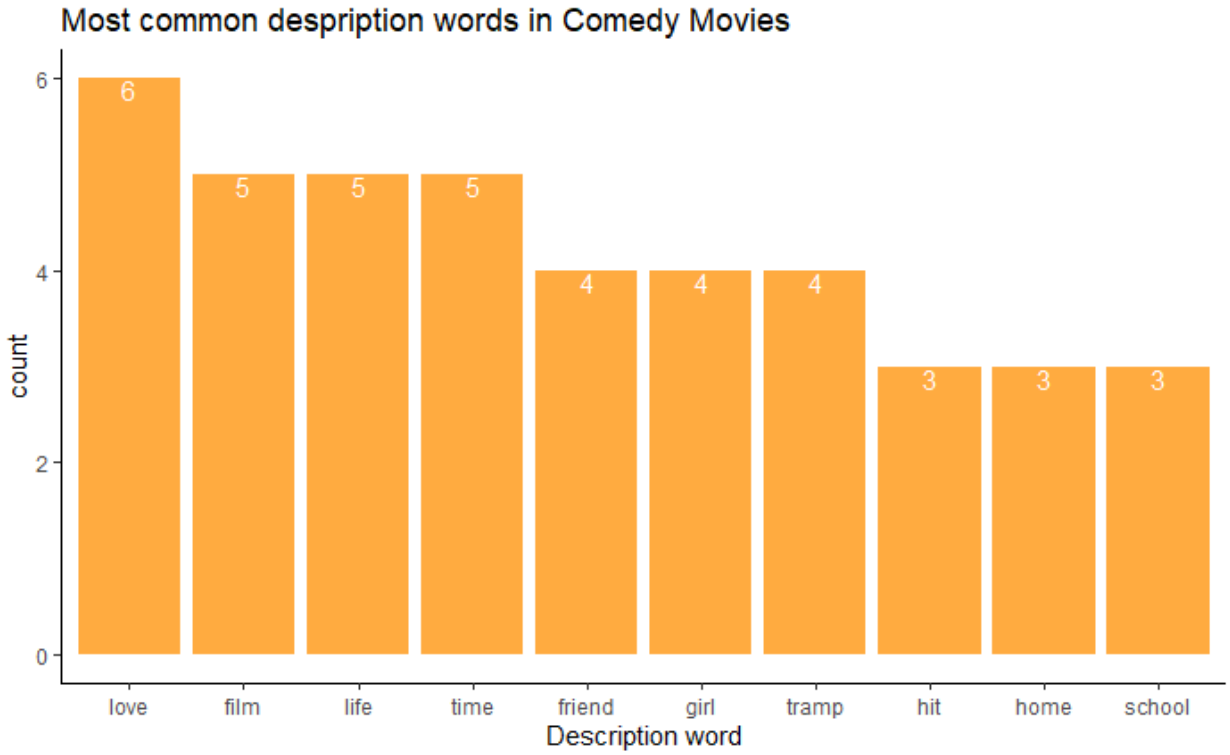
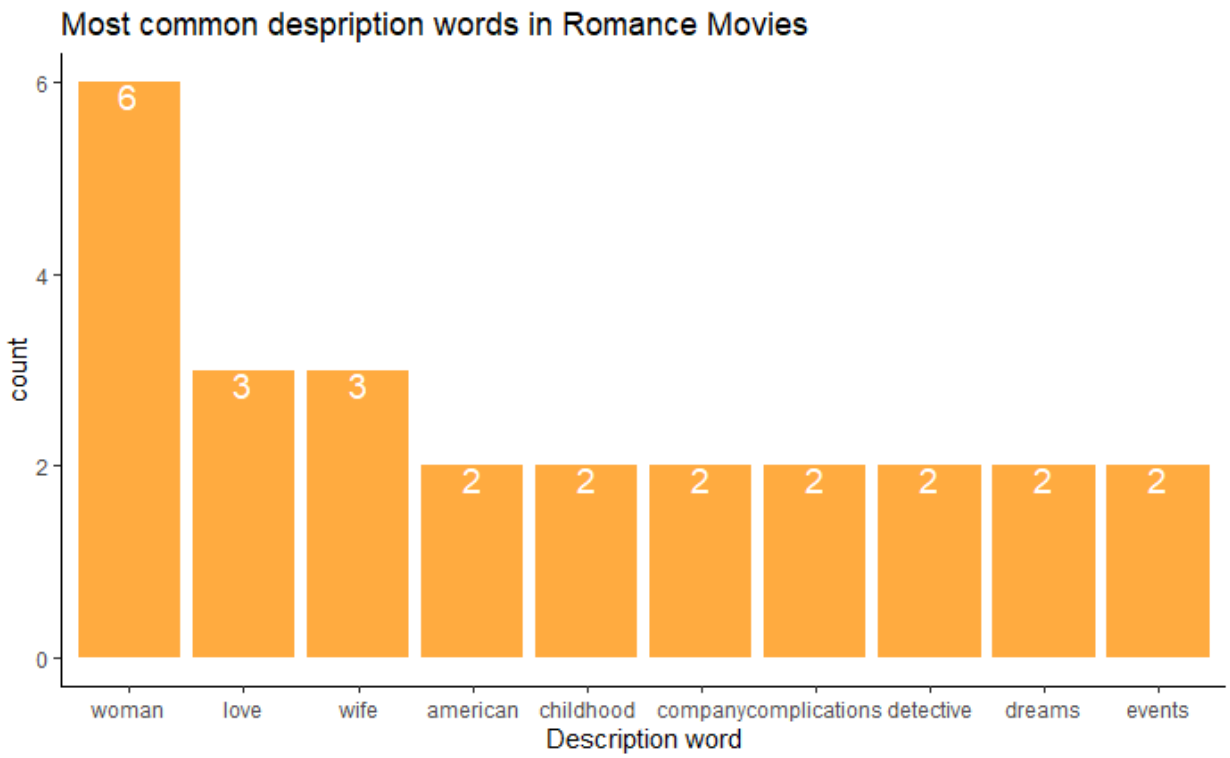


Figure 7

We also dig into the top three movie types to see if there are certain words have the highest frequency in the movie description. We filtered out each genre's specific description words and conducted the analysis over the frequency. We found out that: In high-rated drama movie, the most common words are "life", "story", "woman"; In high-rated comedy movie, the most common words are "love", "film", "life" and "time"; In high-rated romance movie, the most common words are "woman", "love" and "wife".


Most common despription words in Drama Movies
count
Description word
16
9
9
7
7
7
6
6
6
6
life
story
woman
friend
lives
war
child
city
events
love
0
5
10
15


Most common despription words in Comedy Movies
count
Description word
6
5
5
5
4
4
4
3
3
3
love
film
life
time
friend
girl
tramp
hit
home
school
0
2
4
6



Figures 8,9,10

The conclusion we can draw from the common words analysis is that those keywords might be directly attributed to the high score of the movie. In other words, the audience might be more willing to see the related content of the certain description in different categories of the movie. For filmmakers and movie professionals, we strongly recommend them to include those keywords as certain elements in their movie content design when necessary.

Budget Analysis

In addition to analyzing movie content, we also wanted to explore how different aspects of a movie's production may affect its rating. There are many different things that go into making a movie, such as scripts, sets, costumes, special effects, actors, directors, etc., all of which play an integral part in the quality of a movie. In order to explore how production as a whole affects movie ratings, we decided to look at a movie's production budget. This budget is what pays for every element that goes into making a movie, and it's a great metric to quantify the resources that a movie has in the production process. By exploring budgets, we can figure out if having more money actually leads to the creation of higher rated movies. With these results, filmmakers can determine if they really need a large budget to produce a good movie.

Before we can accurately explore the relationship between production budgets and movie ratings, we must transform the existing budget variable. As mentioned earlier in the variable transformation section, the existing variable was a character containing the currency sign before the actual budget amount. In order to be able to make calculations and compare budgets, we had to strip the currency sign from each value and convert to a numeric. After this, there was still some transformation needing to be done. Since we had budget data on movies going back to 1912 (with a mean year of 1993), we had to adjust each budget amount for inflation. Using 2020 as our base year, we now had a standard budget variable that we could make meaningful comparisons with, regardless of the year the movie was made.

Once we had our inflation adjusted budget variable, we wanted to determine the relationship between the budget and the movie rating, which can be seen below in Figure 11.

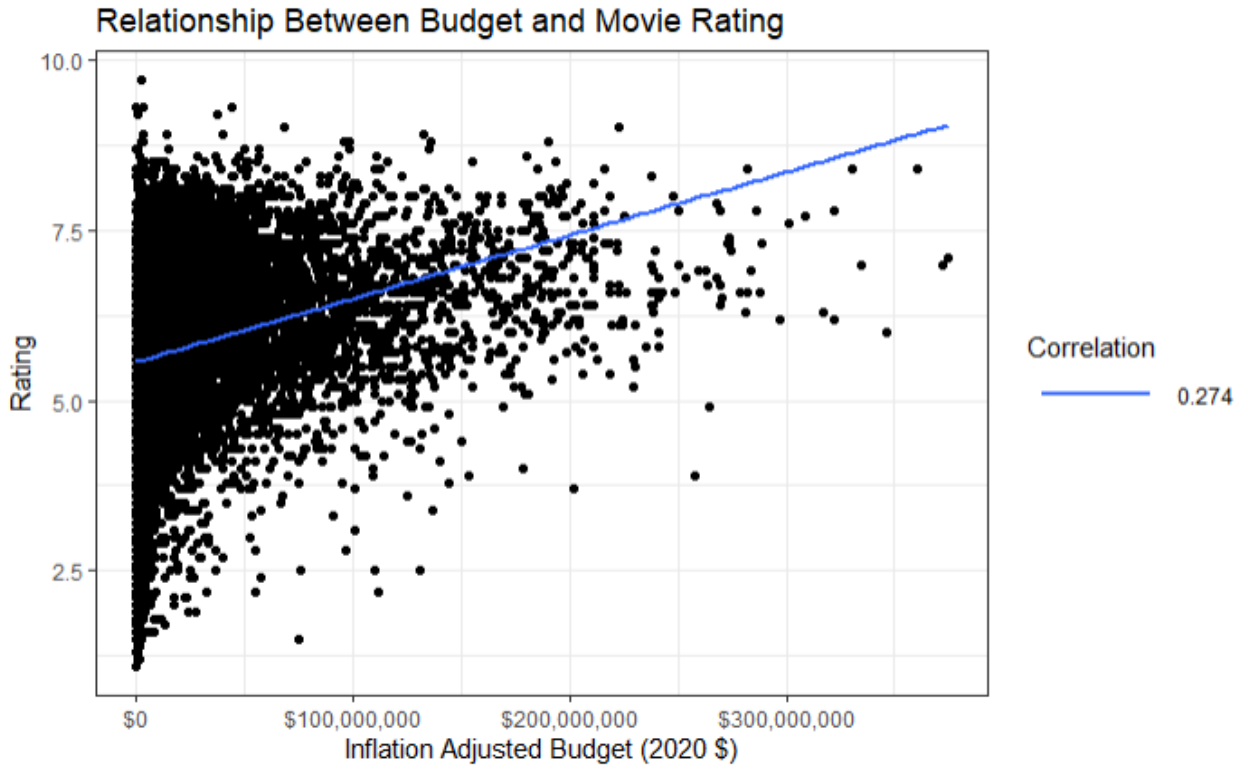


Figure 11

As you can see, there is a positive correlation between the two variables, but the correlation is not overly strong - the r-value is 0.274. However, as you can notice from Figure 11, the vast majority of movies have very small budgets, particularly in the < $50,000,000 range. In this budget range, there is an extremely high disparity in movie ratings. There are many low budget movies with very low ratings, and there are also many low budget movies with very high ratings. But, as movie budgets start to increase, the number of low rated movies starts to decrease significantly. Once a movie has a budget >$300,000,000, there are virtually no movies (for which

we have budget data on) with a rating below 6.0. So, this goes to show that a low budget movie can be really good or really bad, but a high budget movie is almost always going to be above average.

Figure 12 below is a different view of the same scatterplot shown in Figure 11. Here, we split the adjusted budget amounts into four groups: <$50M, $50M - $125M, $125M - $225M, and >$225M. The purpose of doing this was to try and quantify the interpretation Figure 11. As I mentioned, low budget movies have a big disparity in movie ratings, so we would expect an average closer to 5.0. Alternatively, very high budget movies have much fewer lower rated movies, so I would expect that average to be much higher.

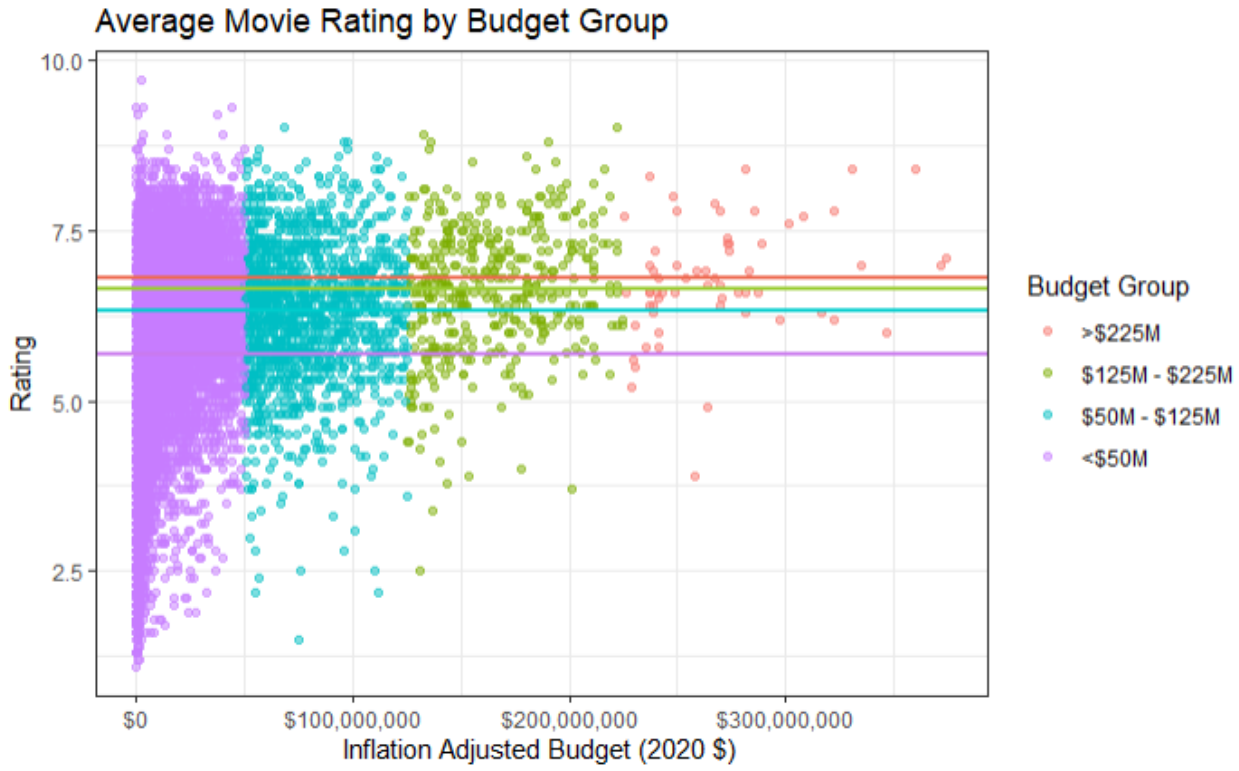


Figure 12

As expected, the lowest budget group in purple has the lowest average rating (5.69), and the highest budget group in red has the highest average rating (6.80). This backs up the idea that an increase in budget leads to a higher probability of creating a good movie.

Another thing that should be considered when looking at movie budgets is the movie runtime. Very long movies have more scenes and cost more money to make by default, whereas very short movies require much less money to make. Because of this, we also wanted to standardize the movie budgets to account for duration, and determine how, if at all, this affects

the relationship with movie rating. The new variable we created is each movie's inflation adjusted budget divided by its runtime.

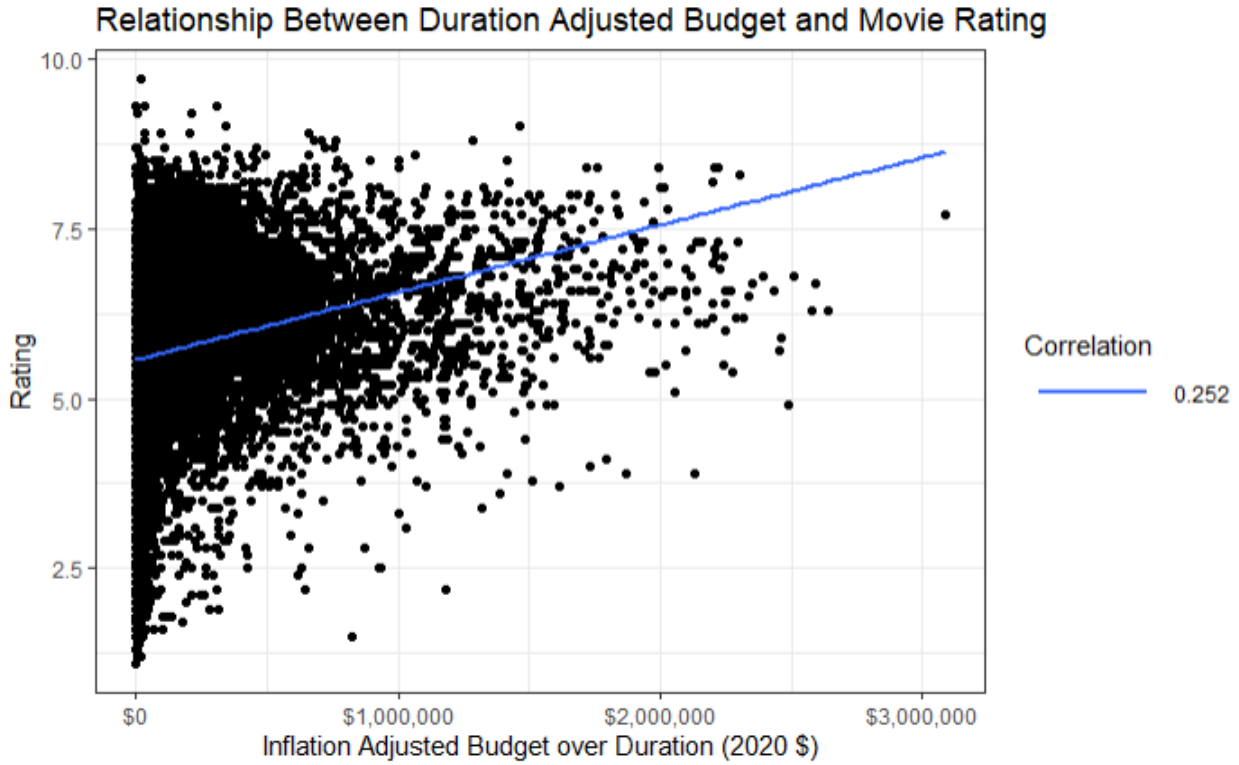


Figure 13

As you can see from Figure 13 and the correlation coefficient, the relationship between movie budget and movie rating is slightly weaker after adjusting the budget to account for movie duration. There could be two reasons for this. One is that short films with low movie ratings now have much higher adjusted budgets, which adds more movies in the bottom right quadrant of the plot and causes the right half of the regression line to get pulled down. The second reason may be that very long movies tend to have higher ratings, so their new adjusted budget is less than before, adding more movies to the top left quadrant and pulling the left side of the regression line up. Both scenarios lead to a slightly flatter regression line, but even still, there is still a positive relationship between budget and movie rating.

In conclusion, you don't need a high budget in order to create a high rated movie, however, having a high budget increases the probability that you do. The biggest takeaway for a filmmaker is that there is no indication of whether a movie with a small budget will be rated high or low. However, movies with higher budgets are much more likely to be rated highly. It's also important for filmmakers to know that the duration of their film does not noticeably change the effect of the production budget on the rating.

Actor Analysis

Continuing our analysis of movie production, we also wanted to see if there was a relation between the actors and/or actresses in a film and the reception of that film. Oftentimes in movie marketing, the acting cast is used as one of the main selling points, so it is important to ask if the cast actually does affect the quality of the movie.

So that we don't go reading too much into random noise, the first thing we wanted to confirm was that the distribution of actor rating, as described in the variable transformation section, was significantly different from the scenario where actors starred in completely random movies. To do this we ran a simulation of theoretical actors (as many as we had real actors, n = 1,424) starring in random English movies in the data set, and then calculated the actor ratings for these fictional actors/actresses. The result is figure 14:

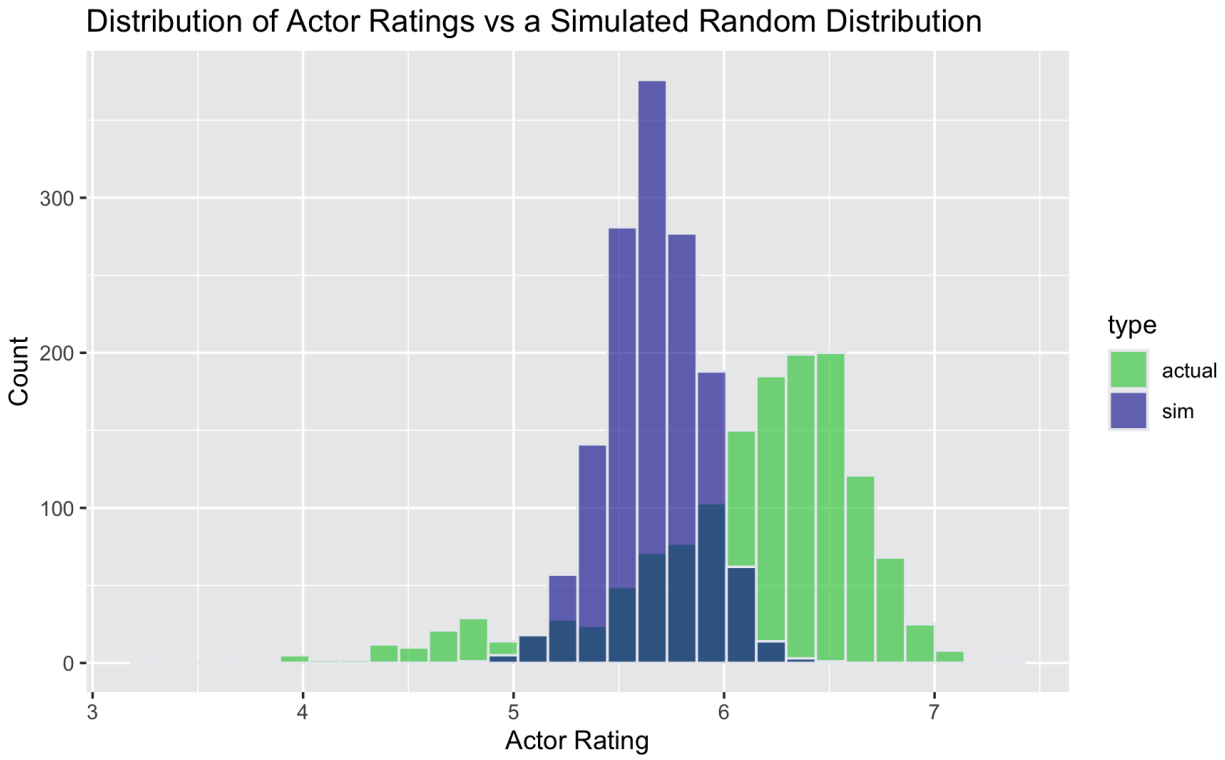


Figure 14

Above, the blue lines represent a distribution of actor ratings if actors starred in random movies. We can see that it looks to be approximately normal around 5.5, which is the average rating of movies in the data set. The green bars are the actual distribution of actor ratings, which we can see is a left-skewed distribution that is densest around 6.5. The difference between the blue and green distributions helps us visually confirm that **individual actors are related to the ratings of the movies they star in.** Most of the actors had a higher actor rating than the average

movie rating, which may be due to the fact that we filtered our data to only observe actors who were in 20 or more movies, which often means they are better actors/actresses than those who may have ended their career after only a handful of movies. Interestingly, we also see actors/actresses who's actor rating is significantly lower than what we would expect from a random sample of movies. These might be actors who made a career out of starring in very-low budget movies, so as we can see from the budget analysis, they would have a lower than average movie rating.

To further investigate the relationship between actor ratings and movie budgets, we can graph them against each other, with each point representing an individual actor. Here, we calculate the budget for each actor similarly to how we calculated the actor rating. An actor's budget is just the average budget of every English language movie they have been in that has a budget in the data set. One drawback of this method is that only about 40% of movies in the dataset have a listed budget, so we are calculating the average budget from a relatively low selection of all the movies an actor has been in.

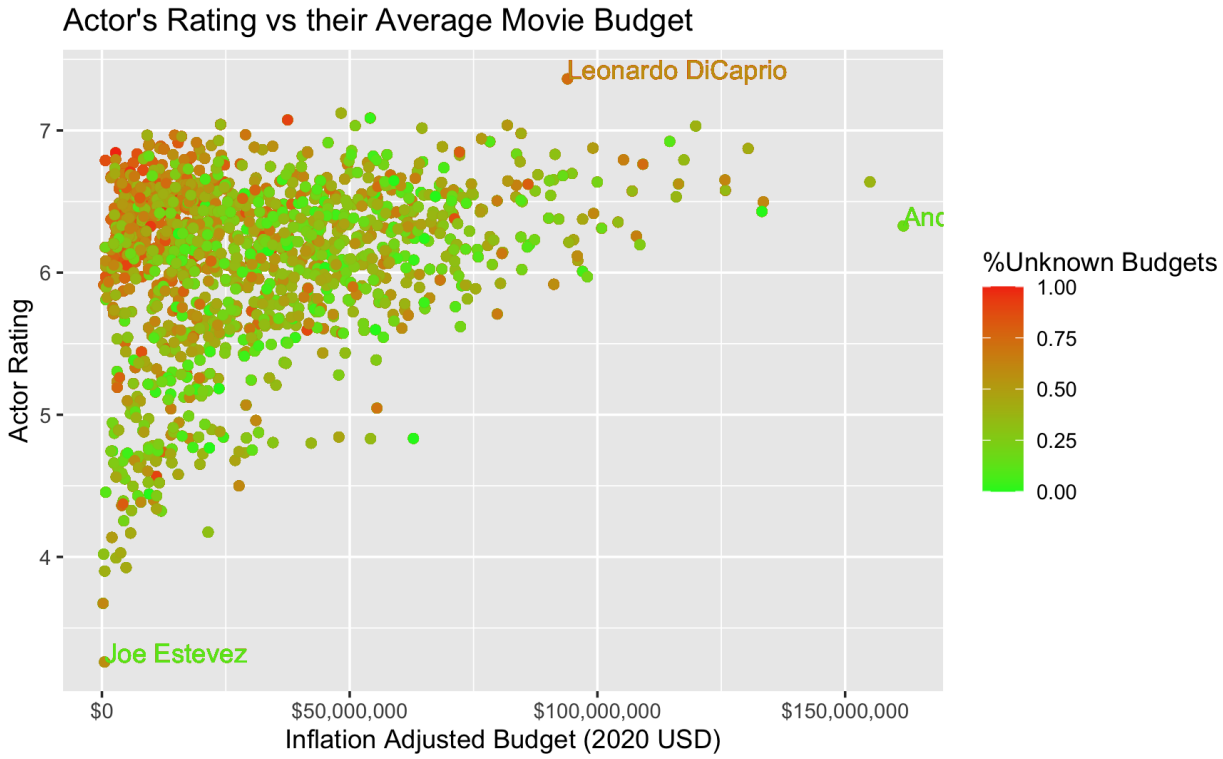


Figure 15

In figure 15 we use color to represent how much missing budget data there is for an individual actor. We can see that actors with a higher average budget are in a higher percentage of movies with known budgets, which probably reflects the fact that the bigger a movie's budget,

the more likely it is to be known. That means that we might expect the redder points to be farther left on this graph if we had full data.

What we see is a graph that looks a lot like the graph of budget vs movie rating. On the low end of budgets actors have a full range of actor ratings, but as the average budget gets higher, the range of actor ratings narrows and the average actor rating gets higher. We can conclude that a high average budget for an actor is predictive of a high actor rating, but a low average budget does not exclude an actor from having a high actor rating.

Limitations

There are certain limitations associated with our data analysis. First, the number of votes can have a large impact on the rating. Some movies only have 100 people to vote for while some movies have over 2000 ratings. The difference in the sample sizes could have a large impact on the average ratings, and we have no way to fully authenticate the ratings. Second, since drama movies have the most percentage among all movies, the sentiment analysis or the common word analysis may fully depend on the same category. It is difficult for us to see the trends in other categories besides drama movies. It is also due to the reason that most movies are categorized as drama movies along with other movie types as well. Third, language is a barrier for us during analysis. We only conducted analysis on USA movies because all of the budgets are in dollars and the related content is in English. Our conclusions might be only applied to the movies produced in the United States. Four, it is hard to set up the threshold of a high-rated movie. In our data analysis, we set 8.0 as a high rating, but we would not conclude that ratings under 8.0 are bad movies. There are many sample size differences included, and since we have no way to authenticate the dataset, we are not sure if 8.0 is a high score for the movie. Moreover, there is missing information for the budget analysis, so our conclusions might be directly impacted by that information. If we had a chance to possess more time with the datasets, we would like to take more time on the budget variable transformation, so that we can transform variables using currency rates to uniform our dataset. We can also conduct the correlation between sentiments and ratings to see if those two variables have a direct correlation. We would also be willing to compare the results with missing information and without missing information to check if our conclusions still stand. Besides common data analysis tools, we can also employ statistical methods to validate if the sample sizes have a large impact on the ratings.

Conclusion

Through our multiple analyses on the IMDB data set, we were able to determine what features of a movie contribute to a high rating. Specifically, we explored data related to movie content and movie production, and our findings should be very helpful to any filmmaker who is trying to understand what goes into a high-rated movie. In our content exploration, we completed

a genre analysis, as well as a sentiment analysis on the movie descriptions of the highest rated movies. Our genre analysis showed that documentaries have the highest average rating among all genres, while animated movies have the highest percentage of movies with a rating greater than 8.0. Our sentiment analysis showed that the more positive sentiment a movie conveyed, the more likely it was to receive a higher rating. We also determined that the most common words in high-rated movie descriptions are “life”, “woman”, and “love”, so incorporating those themes into your movie content will likely lead to high ratings. The next exploration we wanted to conduct was centered around movie production. We wanted to figure out how movie budgets and casting decisions affected the movie rating. First, after adjusting each budget amount for inflation, we were able to show that big budget movies have a much higher probability of being rated highly. We found that low budget movies can be rated highly or lowly, which doesn’t provide much guidance for filmmakers, but as your budget increases, you are much less likely to create a bad movie. Finally, our actor analysis was able to show that there is a strong correlation between actors and movie ratings. So, casting decisions are extremely consequential, and it’s very important that filmmakers try to cast as many actors and actresses with high ratings as possible. With each of these analyses, filmmakers should have a much better sense of what features of a movie will lead to a higher rating, and following these metrics should lead to an overall improvement to the movie-making industry.

Reference

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv