STAT - S670: Final Project

IMDB RATINGS DATA ANALYSIS

Team: Nidhi Vraj Sadhuvala, Sai Charan Chintala

INTRODUCTION

Based on the massive movie information available on the internet, it would be fascinating to learn what aspects contribute to a film's success and determine how uniform or polarized the opinion of a movie is. It's always been fun to experiment with the combination of user ratings for movies and detailed movie metadata. As a result, we'd like to investigate which types of films are more popular or have a higher IMDB rating. We also wish to illustrate the study's findings to make them more intuitive. The response variable in this study is IMDB rating, while the rest of the components in the IMDB movie dataset are analyzed to focus on operating predictions.

We will be exploring the IMDB Movies Datasets taken from https://www.kaggle.com/.

STATEMENT OF GOALS

Through a series of plots and inferences, we want to investigate the findings of the following questions:

1. *Relationship between the length of a movie(runtime) and its rating on IMDB.*
2. *Relationship between the count of votes and ratings.*
3. *Relationship between rating and movies across all years.*
4. *Are the highest-rated movies the ones with the most votes?*
5. *Do the movie genres have an impact on ratings?*
6. *Does language have an impact on movie ratings?*
7. *Does a top successful director affect the movie ratings?*

DATA DESCRIPTION

This dataset contains 14 columns for 8451 shows from various countries spanned across 89 years between 1916 and 2005. When we first looked at the Dataset, we noticed that only 4 of the columns contained numerical values, while the rest were categorical, meaning that the data for those columns was stored in a labeled manner.

The description of the variables in the dataset is as follows:

Categorical Variables:

- **Title:** describes name of the show
- **Year:** the year in which the movie/show was released
- **Kind:** kind of TV show i.e., tv mini series, movie, episode
- **Genre:** refers to the type of story being told and is decided by the playwright
- **Country:** Countries In Which The Shows Were Released
- **Language:** describes in which language the show was shot
- **Cast:** the group of actors who acted in the show
- **Director:** director of the show
- **Writer:** person who writes the scripts
- **Composer:** names of music director

Numerical Variables:

- **Rating:** On Internet Movie Database(IMDB), all films are given an overall rating out of ten; IMDB ratings are based on the votes of the website's users.
- **Vote:** All registered IMDB users can submit a single rating – a number between 1 and 10 for any film on the website. These votes are then re-jigged so that certain demographics (newly-registered users, for example) don't disproportionately influence the overall rating of the film.
- **Runtime:** duration of the show

title	year	kind	genre	rating	vote	country	language
My Big Phat Hip Hop Family	2005	movie	['Comedy', 'Music']	2.3	145	['United States']	['English']
Alone in the Dark	2005	movie	['Action', 'Horror', 'Sci-Fi']	2.4	45019	['Canada', 'Germany', 'United St...	['English']
The Adventures of Sharkboy and Lava...	2005	movie	['Action', 'Adventure', 'Comedy', 'Family', 'Fantasy', 'Sci...	3.7	35417	['United States']	['English']
Cold and Dark	2005	movie	['Crime', 'Horror', 'Thriller']	3.9	986	['United Kingdom']	['English', 'German']
Target of Opportunity	2005	movie	['Action', 'Drama']	3.9	233	['Bulgaria']	['English']
Kisna: The Warrior Poet	2005	movie	['Drama', 'Musical', 'Romance']	4.5	1366	['India']	['English', 'Hindi']
Pit Fighter	2005	movie	['Action']	4.5	1144	['United States']	['English']

DATA CLEANING:

We deleted outliers (97 quartile runtime values) that could potentially skew the distribution in this dataset by removing redundant information such as observations with numerical null values. While analyzing the data, we encountered unrealistic and false runtime for some movies. These movies have a very large

runtime. We considered them as outliers and removed them. For the sake of this study, we solely looked at movies.

title	year	kind	genre	rating	vote	country	language	cast	director	composer	writer	runtime
0	0	0	0	316	316	0	0	0	0	0	0	1653

These values depict the number of N/A values in that particular column. The variables 'runtime', 'rating' and 'vote' have 1653, 316 and 316 null values respectively which were removed from the data.

EXPLORATORY DATA ANALYSIS

Firstly, we want to investigate from simple histograms of "movies Vs year", "movies Vs rating", "movies Vs runtime", "rating Vs runtime", and "movies Vs vote".

The function qplot() plots the distribution of movies over time, as shown in the diagram below:

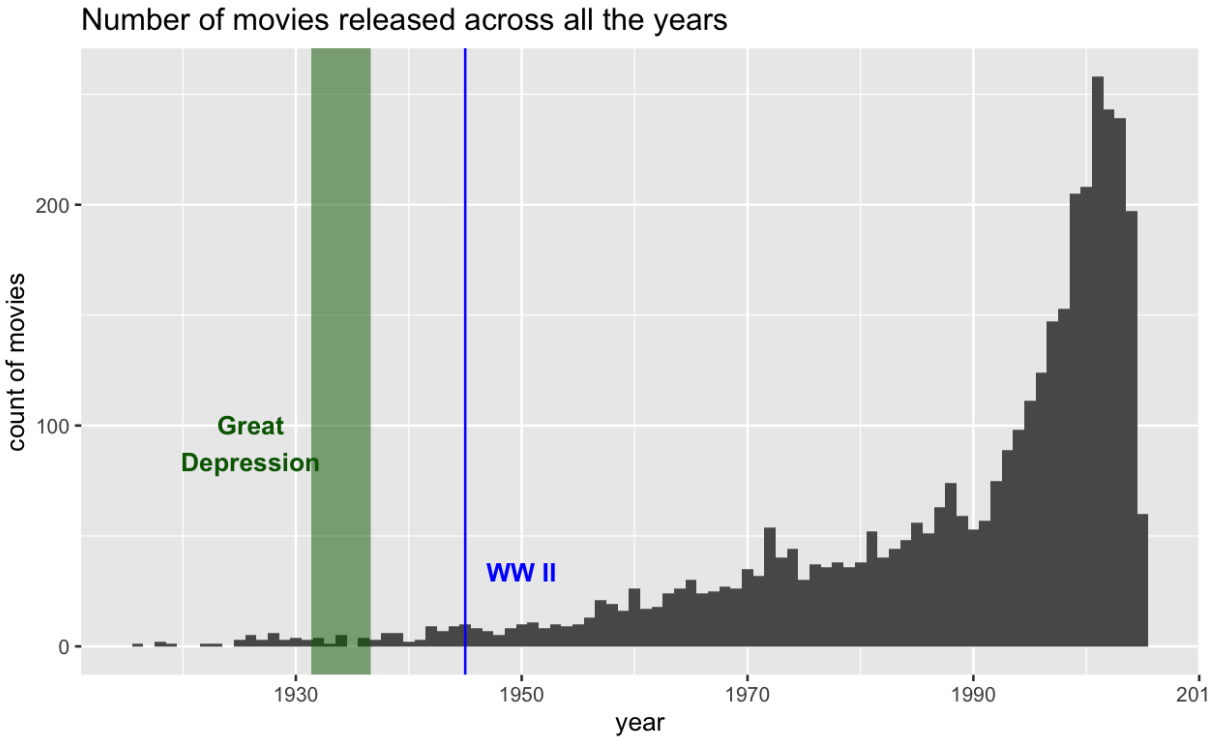


Fig. 2

Fig. 2 depicts the relationship between count of movies across all years:

- The number of movies released is less during the **Great Depression** period (1931 – 1939) as compared to the movies released before and after the period whereas **World War II** (1939 – 1945) didn't have much effect on the show's count.
- We can also infer that the number of movies being released are exponentially increasing over time.

Fig.3 and Fig. 4 describes the ratings and runtime of the movies:

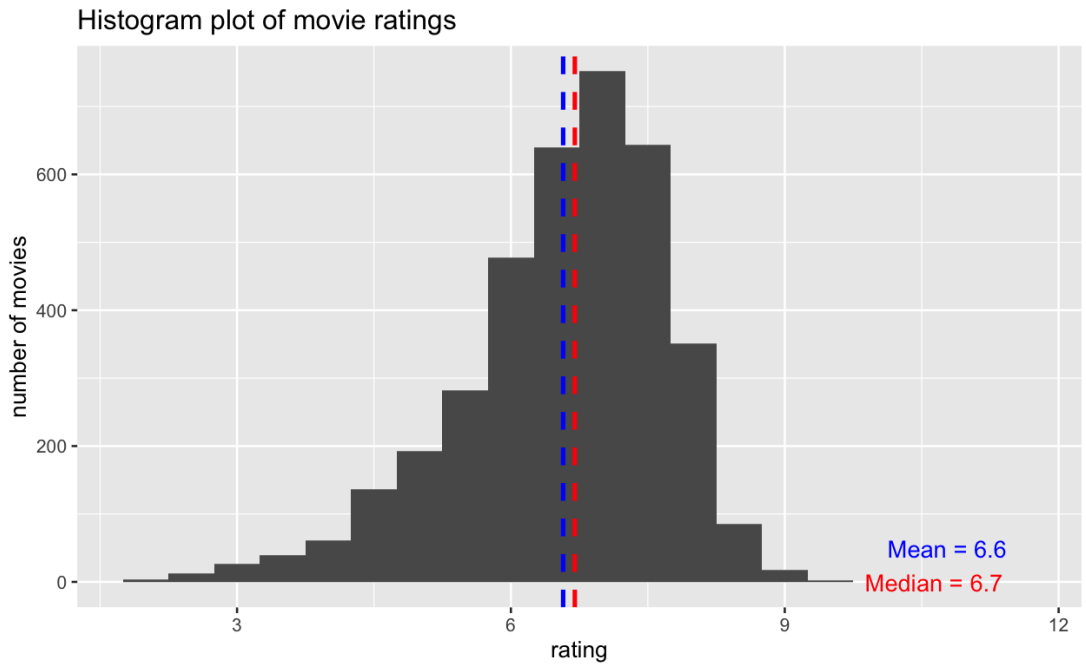


Fig. 3

- This distribution is left-skewed because it attains a peak towards the right side and has a tail towards the left side. Furthermore, the mean best depicts the skewing. In general, if the data distribution is skewed to the left, the mean is lower than the median, as shown in Fig. 3.
- The peak is approximately 750, implying that the majority of the films are rated 7.
- We can also see that 70% of the films have a rating between 6 to 8.

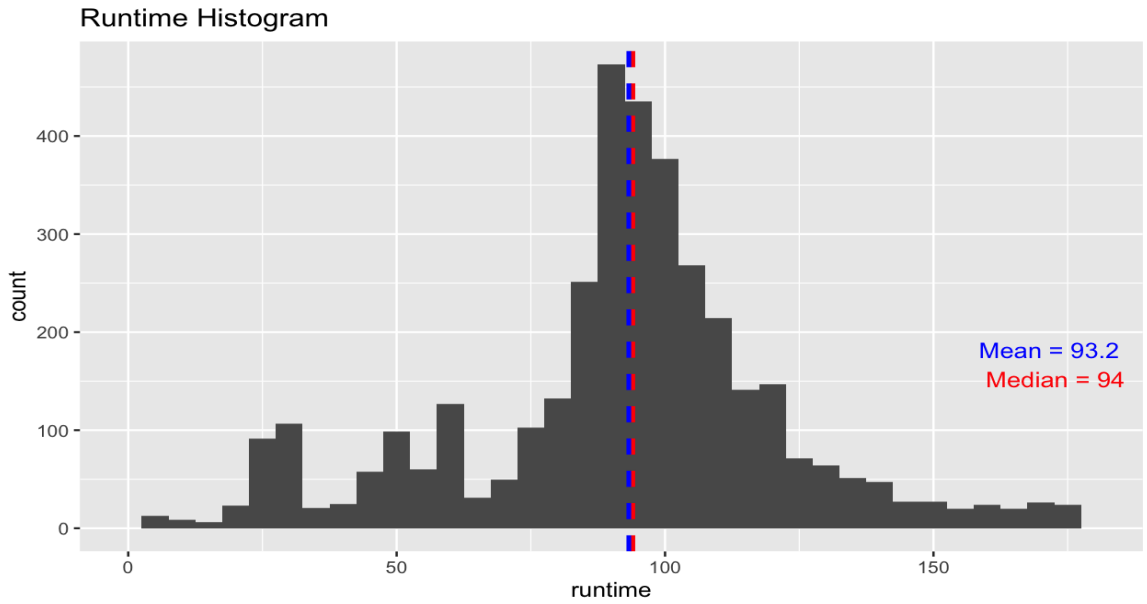


Fig. 4

In the above graph, the x-axis represents the runtime of a movie in minutes . Among all the movies released over the years the number of movies having runtime between 90 and 120 minutes is very high which explains that it is the most acceptable average standard length of a movie by the people and movies below or above this time range are less willingly accepted by people . The distribution is approximately normal as the mean and median are almost equal. We can also infer that the count of movies are very less with duration greater than 120 minutes.

Relationship between Rating and Runtime of a movie:

As per below Fig.5, most of the shows have a duration between 1.5 to 2 hours. Most of these shows have ratings between 6 and 8. However, as the duration increases, distribution of the ratings decreased.

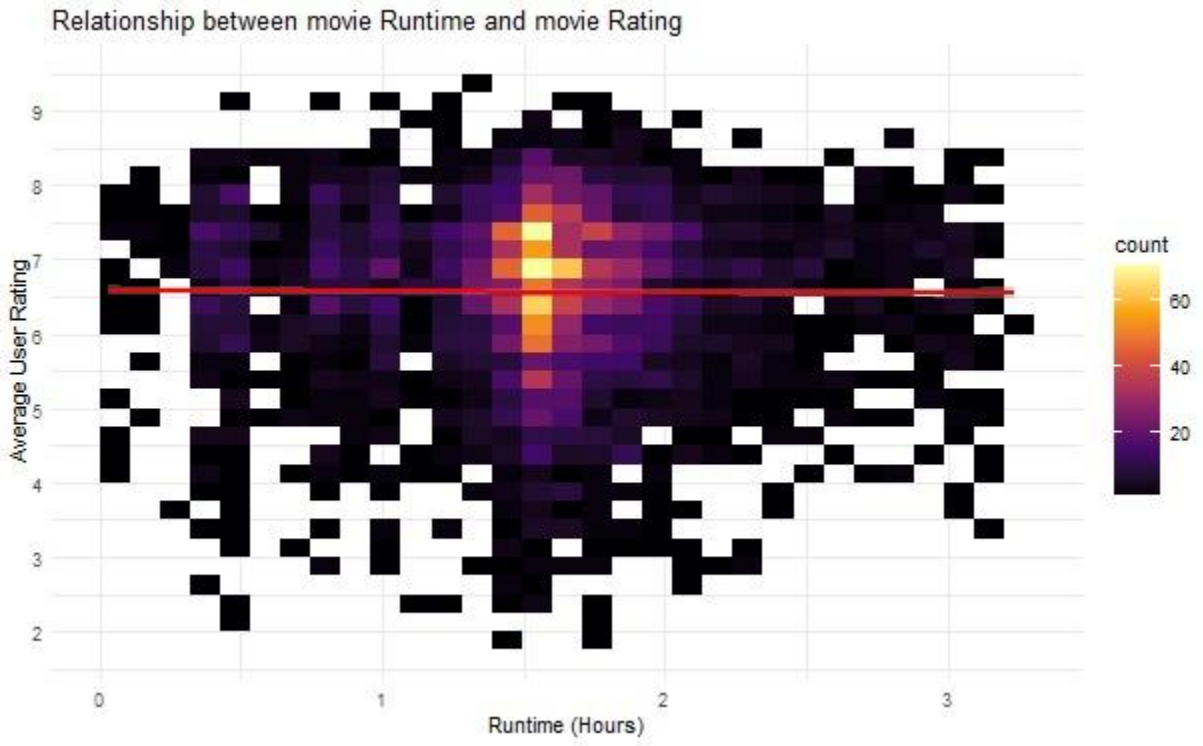


Fig. 5

Plot describing votes received for all movies:

In Fig. 6, the bin size for the histogram below is 500. We may depict that about 550 films have votes ranging from 0-500, with the number rapidly decreasing. The distribution is heavily skewed to the right, which means that just a few films receive the most votes.

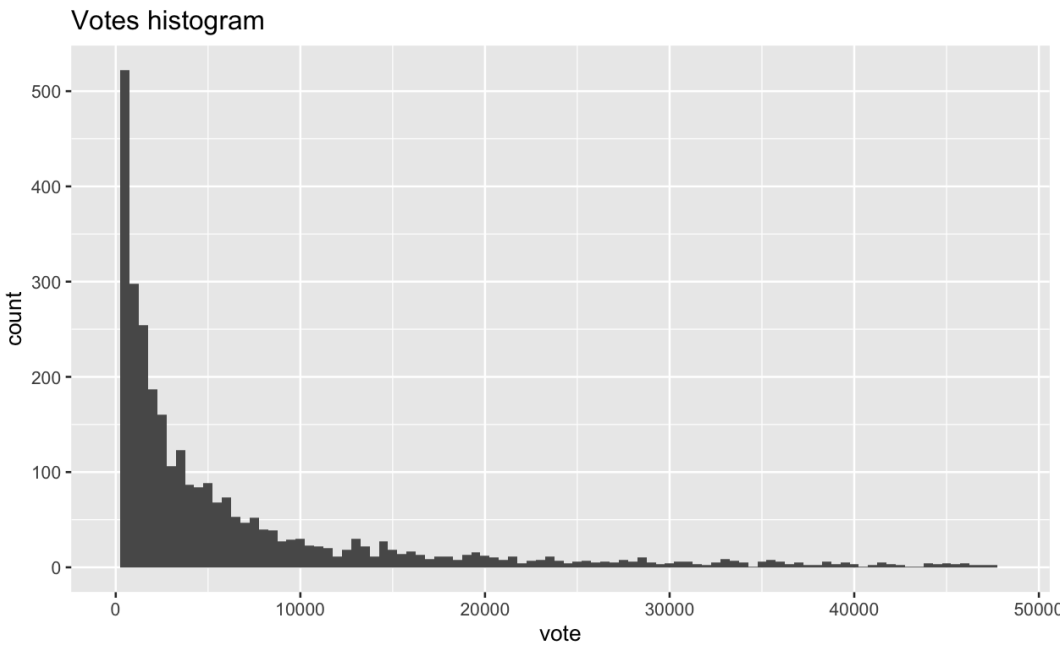


Fig. 6

Plot describing relationship between votes Vs rating:

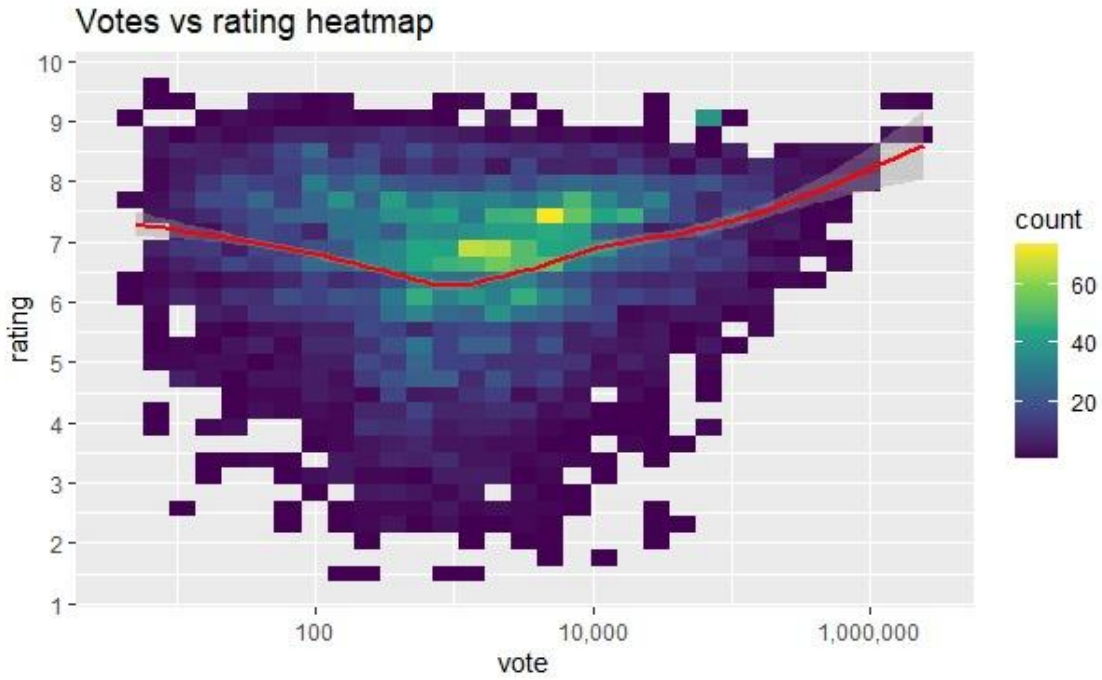
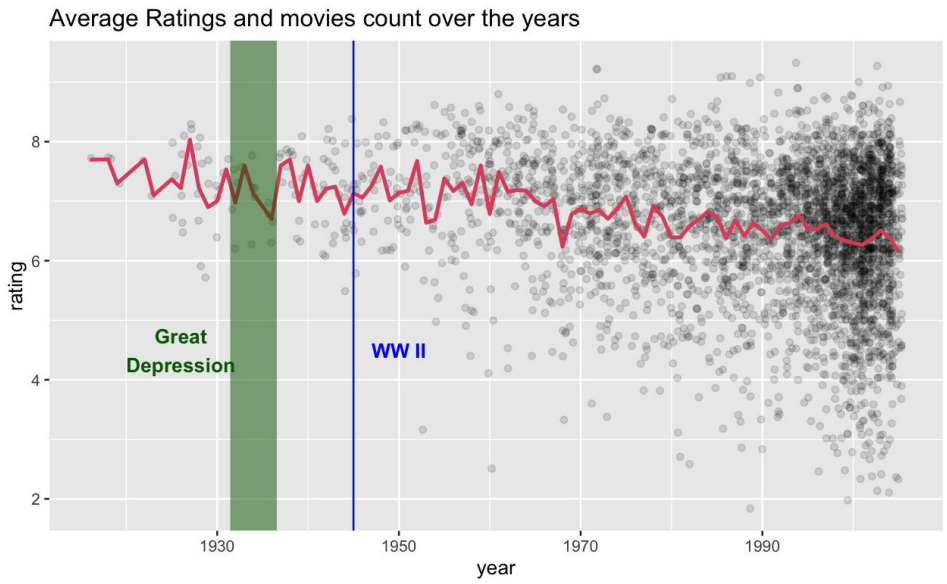


Fig. 7

As per Fig. 7, if we observe closely more movies have votes between 1000 and 10000 with average rating between 6 to 8. As the number of votes is increasing, distribution of the ratings is constrained to a particular limited range(higher values).

Plot describing the analysis w.r.t movie ratings and historical events:



We wanted to analyze if the movie ratings were affected by any conflictive events that happened in some particular years. So we have considered two major historical events which have impacted the world like the Great depression and World War-II. But from the above analysis we can see that the average of the movie ratings did not differ much during the events which conclude that the movie ratings are mostly ineffective of any historical event.

"Loess Model" depicting the analysis of movie ratings over time:

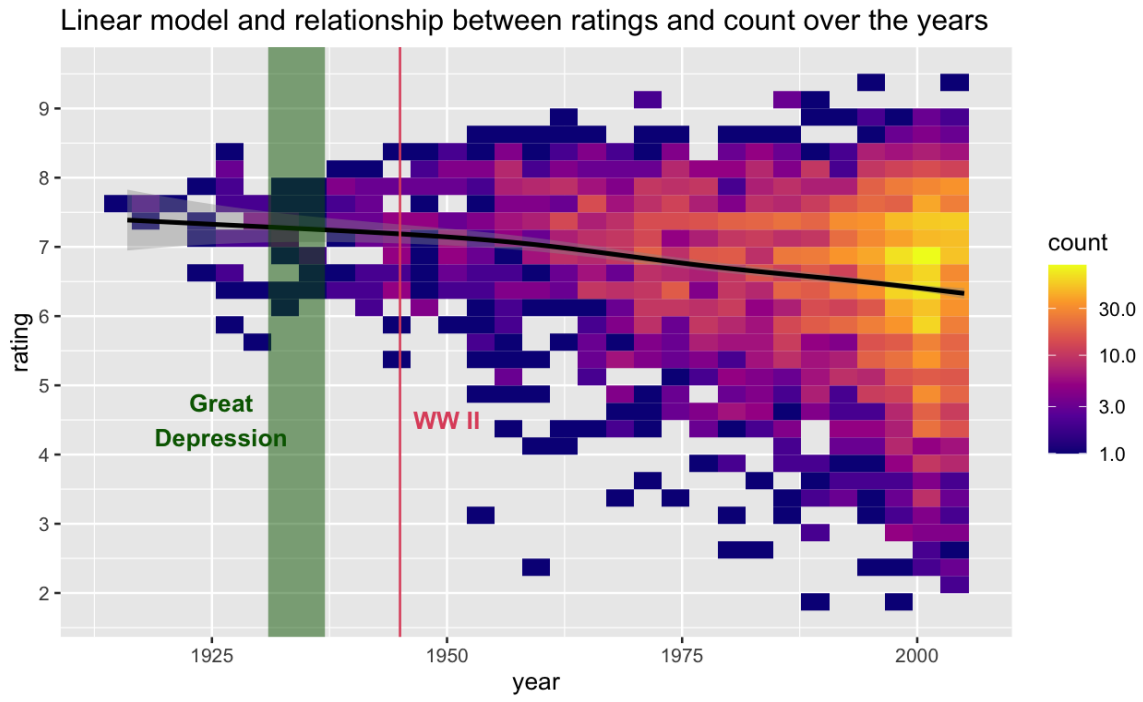
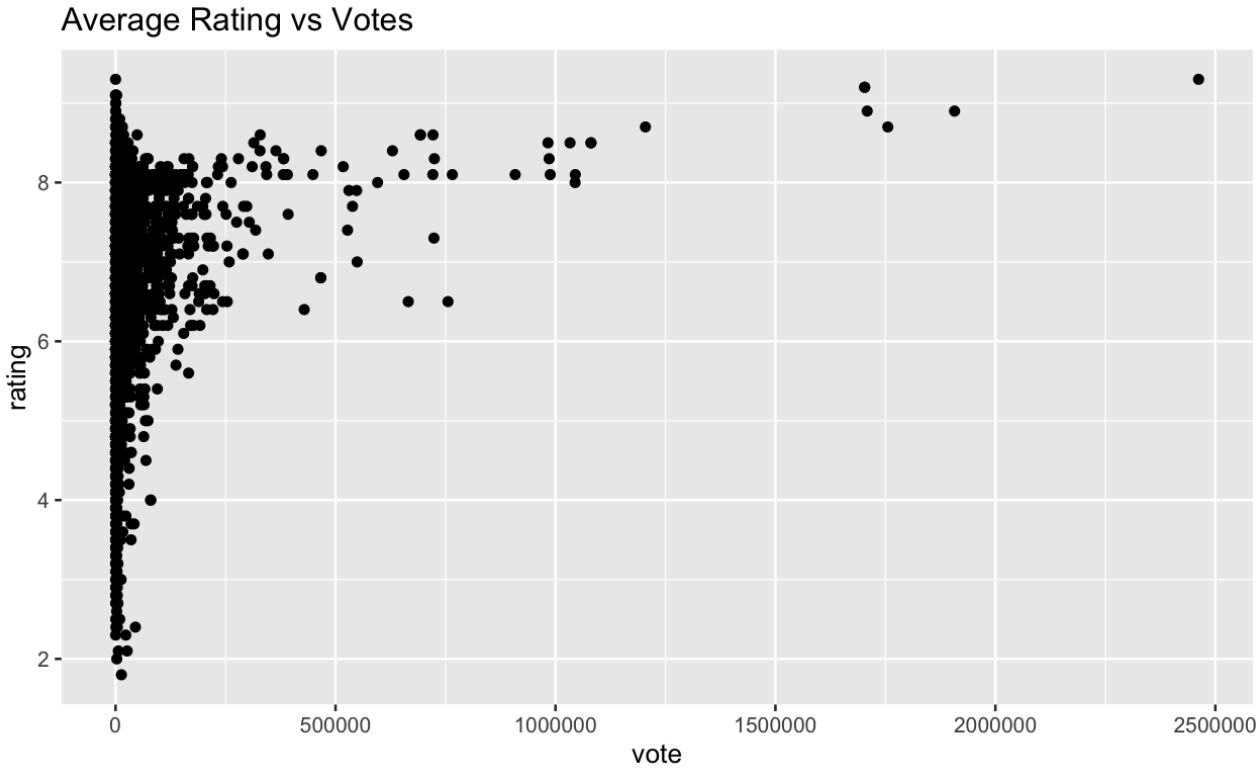


Fig. 9

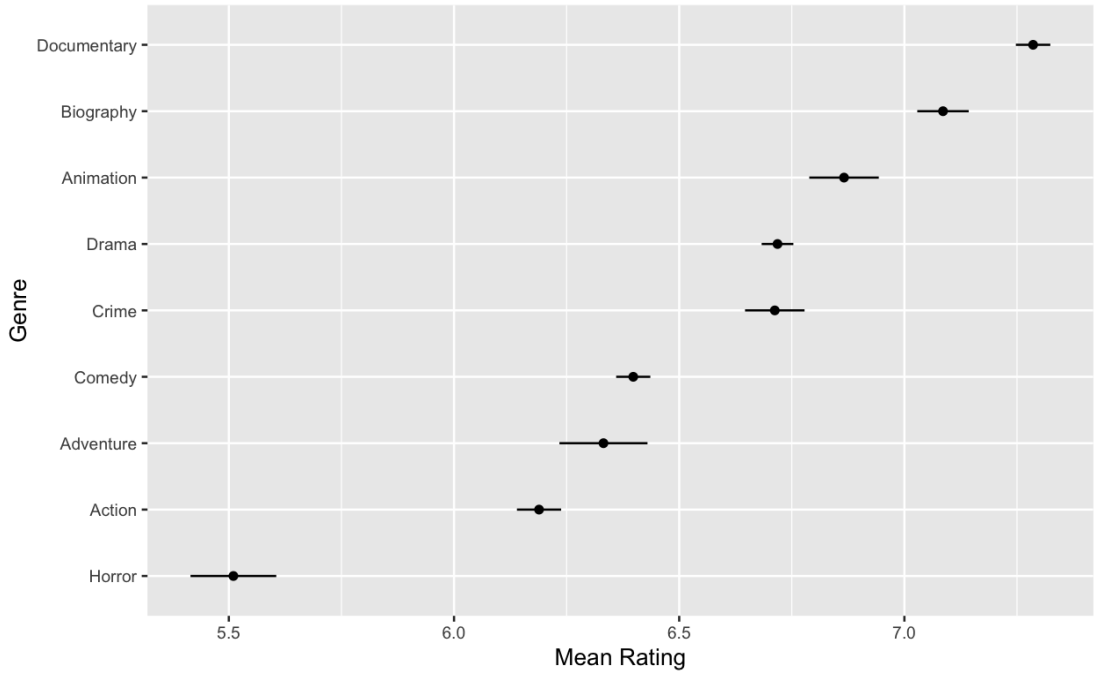
Furthermore, the distribution in the above graph over the time showcases that the count of movies has increased over the years. In order to analyze the movie ratings over time we have applied the loess model to the following distribution and have visualized that the average ratings for movies in a year were deflecting between the points 6 and 8. Also, the decrease in the loess line depicts that the ratings have slightly decreased in recent years. This concludes that the movies in the early years were better than in the recent years.

Plot depicting average ratings w.r.t count of votes:



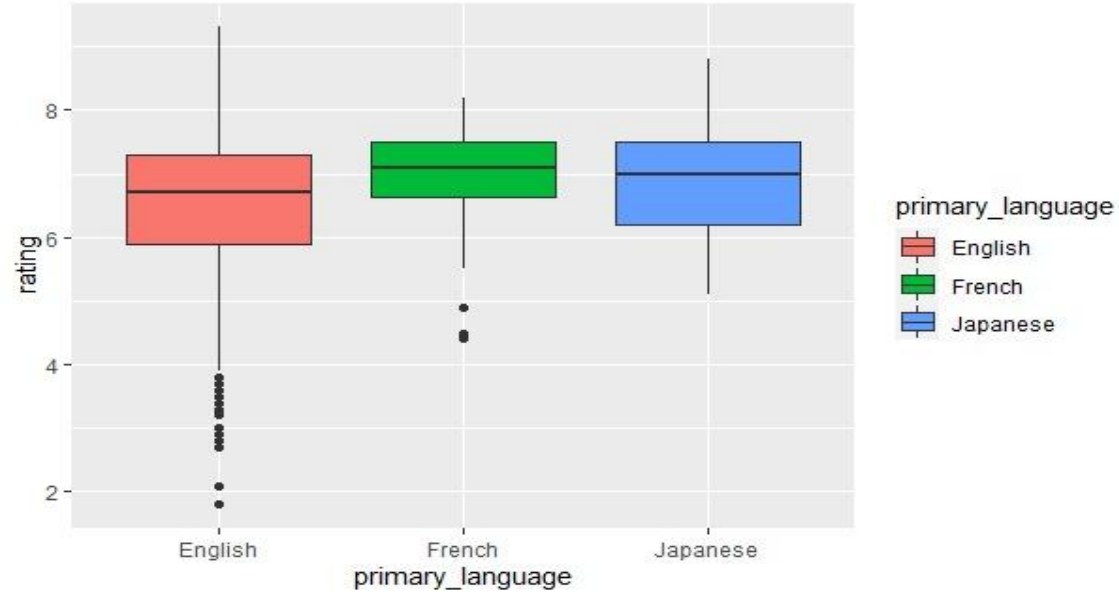
Using the above graph we wanted to understand the interest of people in voting. So, we have implemented a graph depicting the number of votes and ratings for a particular movie; each data point in the above graph depicts a particular movie. Since a large set of movies are centered to a particular range describing that on an average most movies have 0 to 250000 votes. Also, the farthest data point on the right side of the graph tells us that if a movie is very good then it would receive an unevenly high number of votes making it an exceptional case. Here in this case the farthest data point represents the movie "The Shawshank Redemption" which has received about 2500000 votes and a highest rating of 9.3. Thus, we conclude that as the votes of a movie increases the rating is also most likely to be high.

Analysis Based upon Genres:



We have considered the genres that have a count of movies greater than 150 being set as a threshold . For the above plot The above plot explains that the Documentary Genre has the highest ratings when compared to other genres which explains it is a more preferred genre. Also, the lines over the data point describe the standard deviation of that particular genre and since the SD for the documentary is less it represents that there is a less variation in ratings and most likely people tend to have the same opinion. The horror genre movies have the least rating which describes it as the least preferred genre but the SD is high meaning that people have differing opinions which explains the spread in the ratings.

Analysis of ratings based upon languages:



Language being one of the major factors for a particular person to watch a movie and since we live in a multilingual world, languages play a major role for ratings. For the following analysis we have considered the three major primary languages across the world in movies being English, French and Japanese. From the above plot, we understand that the French movies have higher average ratings when compared to other two languages. But as the spread of the French average rating is less when compared to other languages we could say that the rating population is very less in french or the people's opinion upon the French movies could be the same. From the above points, we can conclude that French movies have better films compared to English and Japanese. The spread for English movies is high compared to other two languages. We also know that the number of English movies are more, which might be the reason for a higher spread of ratings.

Analysis to find the top most successful director:-

	primary_director	avg_ratings	avg_votes	success_rate	list_titles
1	Frank Darabont	9.300000	2461873.00	22895418.9	The Shawshank Redemption
2	Quentin Tarantino	8.500000	1476103.50	12546879.8	c("Pulp Fiction", "Kill Bill: Vol. 1")
3	Irvin Kershner	8.700000	1204367.00	10477992.9	Star Wars: Episode V - The Empire Strikes Back
4	Lana Wachowski	7.750000	1110929.50	8609703.6	c("The Matrix Revolutions", "The Matrix")
5	Gore Verbinski	8.000000	1044629.00	8357032.0	Pirates of the Caribbean: The Curse of the Black Pearl

We have found out the top successful director by considering the product of the average ratings and the average votes achieved by a director. The director named "Frank Darabont" has turned out to have the highest success rate and he directed the movie "The Shawshank Redemption". Also, this director achieved the highest number of ratings and votes.

CONCLUSION:

1. If the runtime of the movie is outside the range 90 to 120 minutes then the ratings are most likely to decrease.
2. As the count of the votes increases for a movie the ratings also tend to be higher.
3. The average IMDB ratings for the movies have decreased over the years.
4. The movies with the highest ratings tend to have the highest number of votes. Example: “The Shawshank Redemption”.
5. The genre does affect the ratings and the documentary genre seems to have the highest rating when compared to others and SD explains that the variation in voter ratings differ from genre to genre.
6. The different movie languages have different average ratings. But French movies tend to have the highest rating and English movies have a large variation since it has the highest movie count.
7. A successful director does affect the number of votes received and ratings achieved. We found that “Frank Darabont” is the top most successful director and his movie “The Shawshank Redemption” received the highest votes and ratings till today.

LIMITATIONS:

1. As the number of movie categories are more, we cannot replace missing values of categorical variables with mode.For instance, we have different categorical variables such as cast, director, writer which wouldn’t make sense if we replace those values with modes.
2. We encountered many languages with fewer movies. But, we have confined our data to the top 3 major languages.
3. While analyzing data based upon directors and languages, some movies have multiple directors and are released in many languages. Since the combinations for these variables are many, we just considered the primary director and primary language for our analysis. We were not able to analyze multiple directors and languages at the same time.