Project Report - IMDb data analysis

December 2021

Madhuleka V Iyer

Under the supervision of Prof. Dootika Vats

The internship involves getting hands-on experience working with the programming language R and acquiring the skill of web-scraping using R. The data so acquired has been structured and listed in the form of a data frame which was further saved as a CSV file. The CSV file was used in generating plots for various parameters of movies. The graphs helped analyse and visualise the data. Most of the graphs showed results as expected. The internship also involved learning and using RMarkdown and this report has been created using RMarkdown.

Introduction

IMDb is a web database consisting of information about movies worldwide, giving users a platform to rate and review movies. IMDb produces a list of the top 1000 movies. It includes details like rating, certification, run-time and genre among others. [1]

The goal of the project was to gain web-scraping skills using R and be able to analyse the data so obtained. It involved understanding the data, plotting graphs and learning to read them.

We scraped data from the IMDb's list of top 1000 movies based on user rating. This data was used to comprehend the common characteristics of movies included in the list. The internship involves getting hands-on experience working with the programming language R and acquiring the skill of web-scraping using R. The data so acquired has been structured and listed in the form of a data frame which was further saved as a CSV file. The CSV file was used in generating plots for various parameters of movies. The graphs helped analyse and visualise the data. Most of the graphs showed results as expected.

Data obtained upon scraping IMDb website

Web-scraping is a method of obtaining large amounts of updated data from websites. This is done using the package "rvest".

We scraped the data of the top 1000 movies. This data was presented in the form of an unstructured HTML file on the website. We extracted the relevant data points and created a data frame with the same. We further saved this data frame as a CSV file to reuse throughout the project.

[1] Link to IMDb top 1000 movies: https://www.imdb.com/search/title/?groups=top_1000&sort=user_rating,desc&count=250&start=1&ref_=adv_nxt

This CSV file included the name of the movie, year of its release, rating (out of 10), duration of the movie, genre (up to 3) and the number of votes.

We scraped the certification of the movies. This posed certain (presently un-resolved) challenges. Some movies on the database did not have the class 'certification'. The list of certifications that was obtained upon scraping, did not match with the data that was displayed on the website. This resulted in the size of the list containing the certificates not matching with the rest of the data frame. Hence, it could not be included in the same. The reason for this challenge was that the R commands collected HTML documents using a different server system. So, the host website presented a different HTML file from what was visible via the raw HTML code we obtained from the website.

The rest of the data was saved in the CSV file and was used to analyse and visualise.

As we move into the next section, we keep in mind that the data is only of the top 1000 movies on IMDb per customer rating and hence the conclusions cannot be generalised to all movies.

Plots obtained

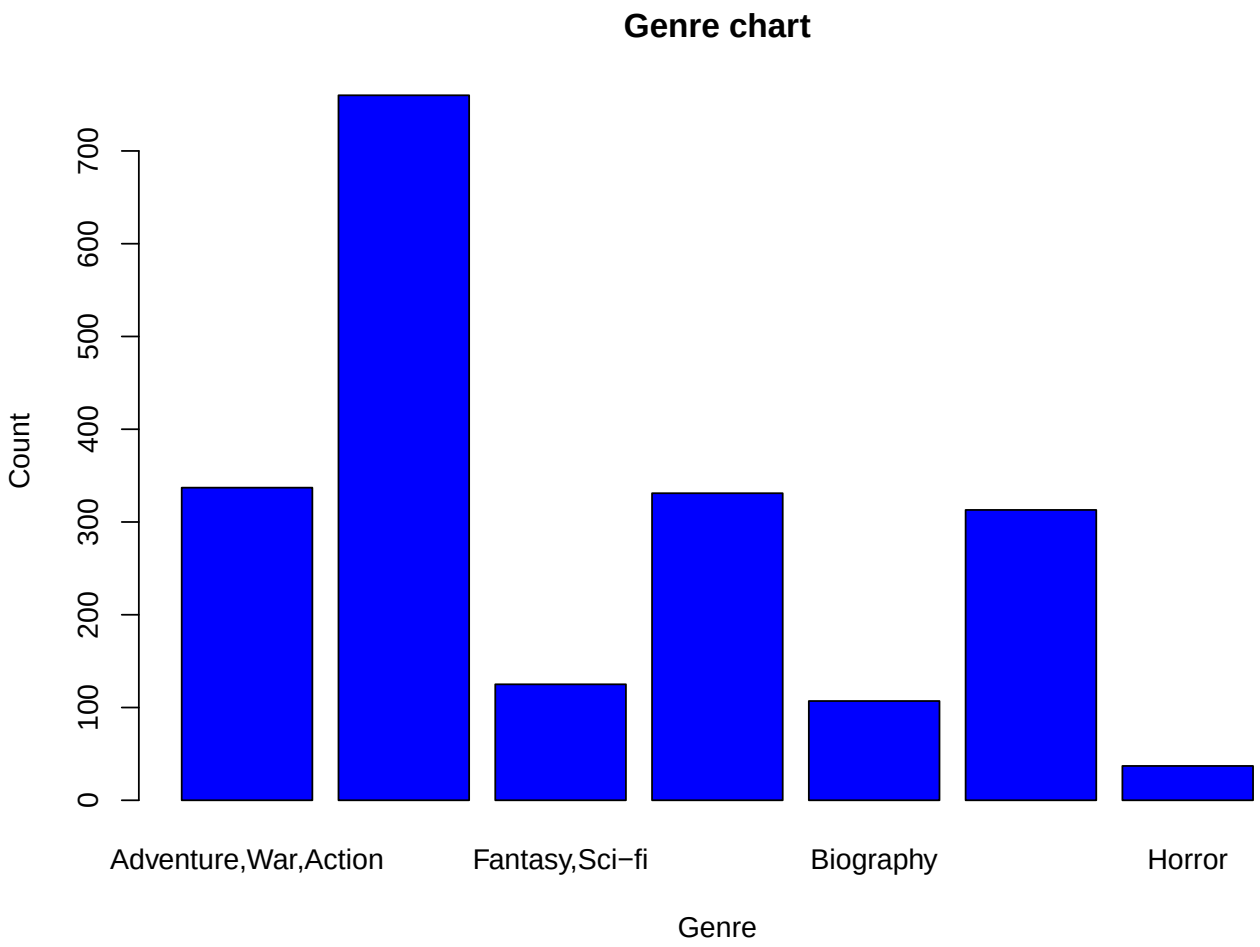
The conclusions made by reading the following graphs are conditioned on the data being that of only the top 1000 movies. Given this condition, we can make the following plots and inferences.

Genre:

Genre is an important aspect of movies responsible for their popularity among viewers. We plotted a bar graph to understand the popularity of genres since the list of top 1000 movies on IMDb is based on user reviews.

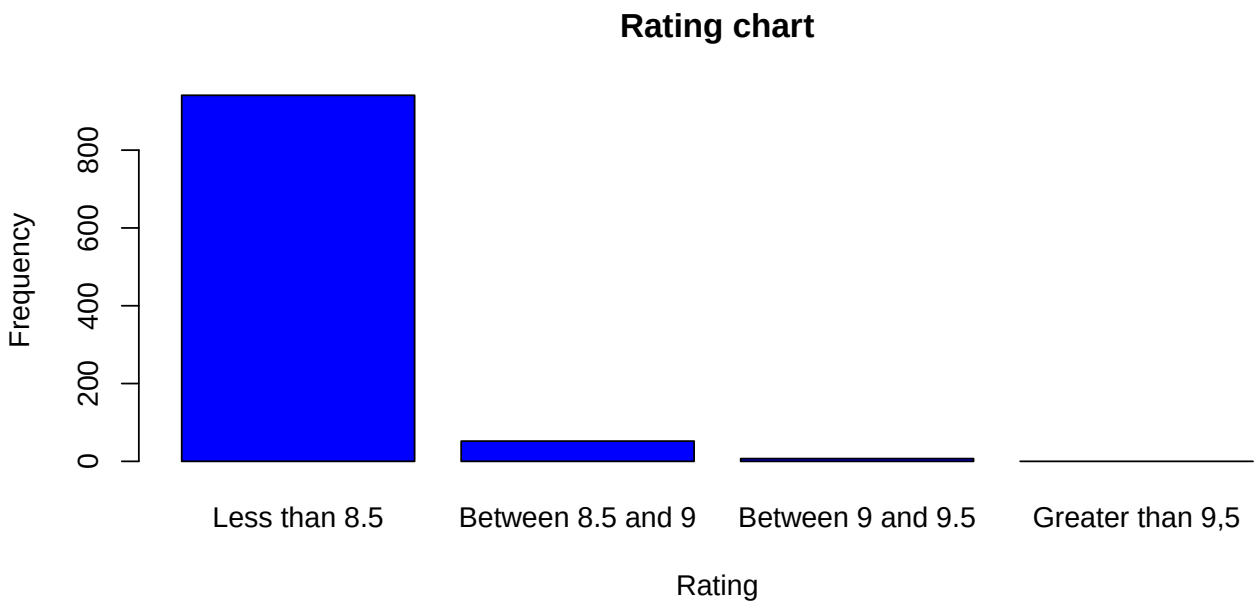
The genres are categorised into "Adventure,War,Action", "Drama,Family", "Fantasy,Sci-fi", "Crime,Mystery,Thriller", "Biography", "Comedy,Romance" and "Horror". The graph shows that movies belonging to "Drama or Family" are the most common in the list of top 1000 movies on IMDb. This is followed closely by movies belonging to "Adventure, War, Action", "Crime, Mystery, Thriller" and "Comedy, Romance". "Horror" movies appear least often on the list.

We only scrape the top 1000 list and hence, do not know the number of overall "Drama" movies released. So, concluding that drama movies are more likely to be voted onto the top 1000 list is misleading. The conclusion we can make is that the maximum number of movies on the list of top 1000 movies belong to the genre "Drama".



Rating:

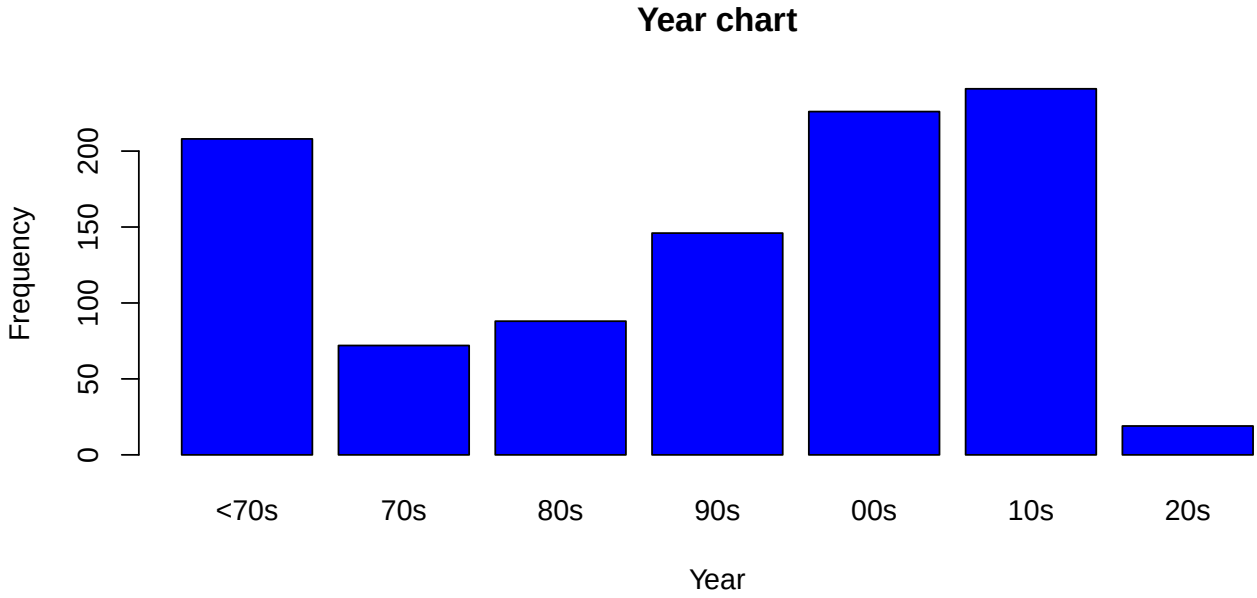
The rating of movies on IMDb is based on the votes of the public. We plotted a bar graph to understand the most common rating among movies in the top 1000 list. Most of the movies on the list have a rating of 8.5 or less. We conjectured that the movies in the top 1000 list might have a high rating. But most of the movies on the list are rated less than 8. This is not surprising as the number of movies rated over 9 are movies that overwhelmingly received a positive response. But all movies on the top 1000 list are not expected to have a rating this high. We observed that the number of movies decreased as the rating increased and no movies were rated over 9.5.



Decade of release of movies:

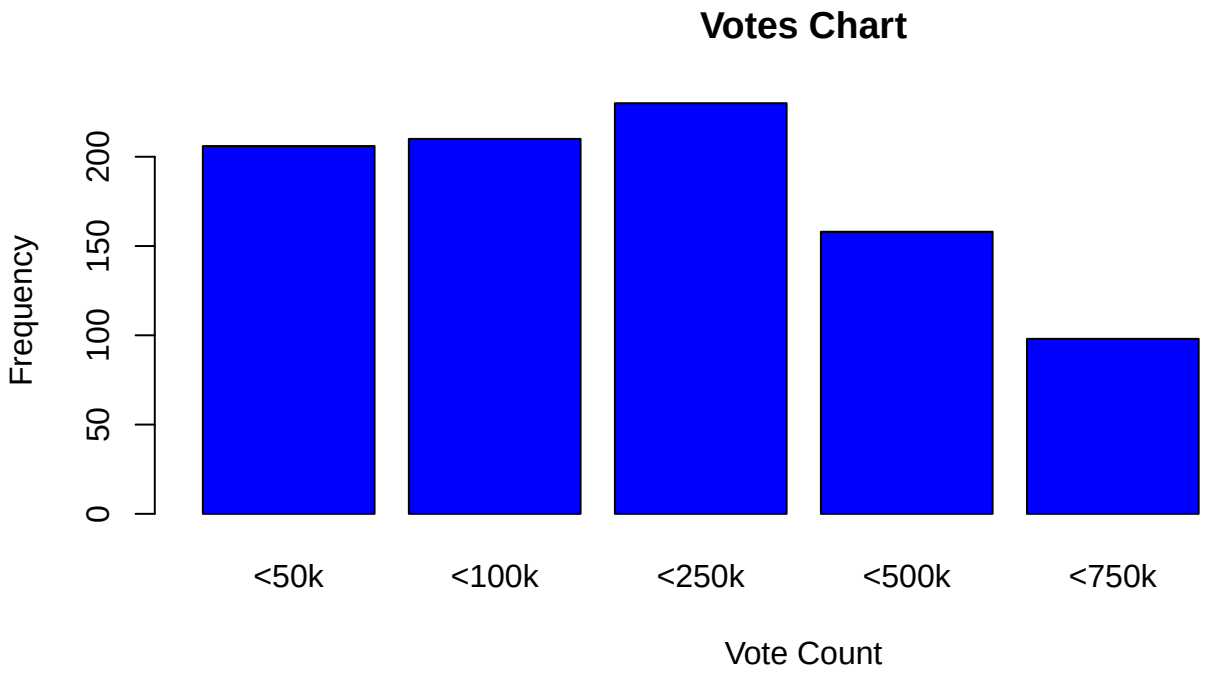
We wanted to analyse whether the decade of movie release impacted its popularity among IMDb users. We plotted a bar graph to analyse which decade appeared most often on the top 1000 list. As expected, the graph depicts that the movies released in the 2010s are the most commonly featured on the list. This is followed very closely by the movies released in the 2000s. The movies made in the 2020s appears least on the list, but this can be ignored as the decade is three years old.

But we do not know the total number of movies made in each decade. The large number of movies from the 2000s and 2010s on the list could be a result of the advent of the internet during these decades. Hence, a conclusion on the quality of the movies from these decades cannot be made. A conclusion that these movies have a higher viewership and voters can be made.



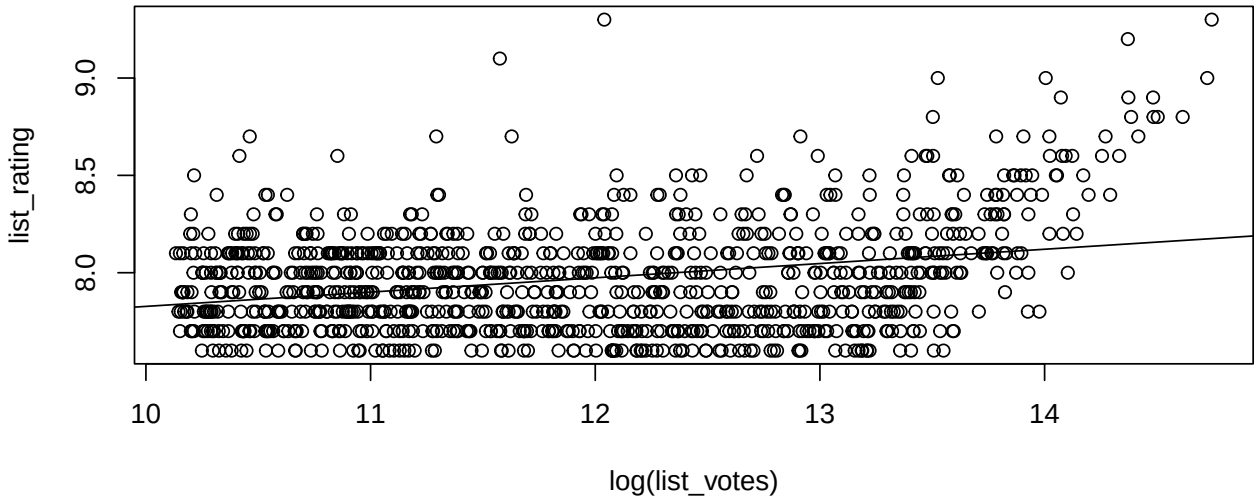
Number of Votes:

The rating of movies depends on the number of votes and hence this is an important indicator of the popularity of movies. We plotted a bar graph to analyse the number of votes that the movies on the list received. The motivation for this chart stemmed from the fact that movies such as “The Shawshank Redemption” (released in 1994) has over 2.5 million votes and “Inception” (released in 2010) has over 2 million votes while later movies like "Dara of Jasenovak (released in 2020) has only 80,000 votes (As of January 2022). The graph depicts that a large number of movies on the list of the top 1000 have between 100,000 votes and 250,000 votes. But a significantly large number of movies have lesser than 50,000 votes.



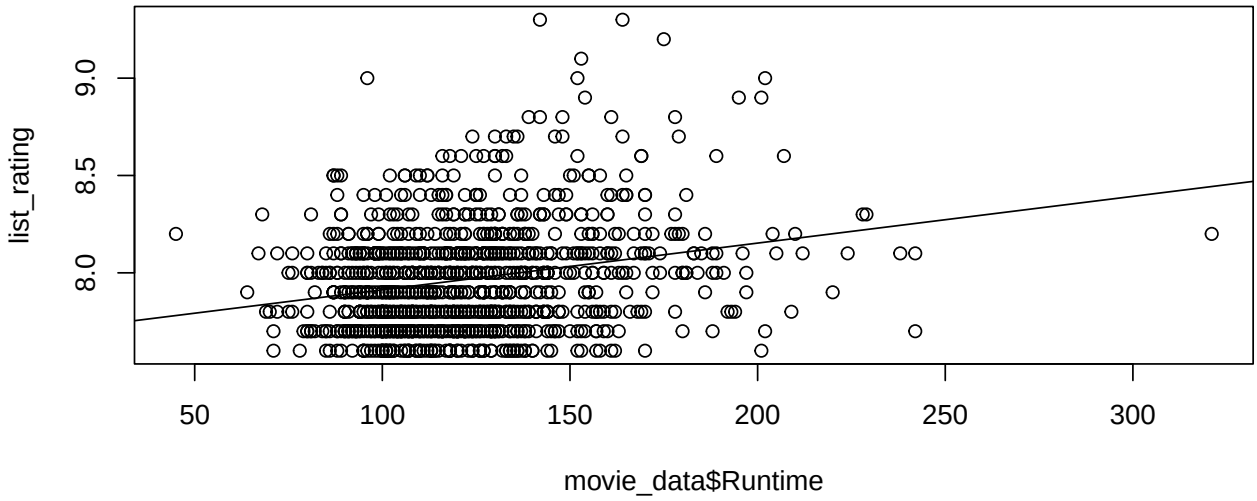
log of number of votes vs rating:

We plotted a scatterplot between the log of the number of votes (log was taken to make the graph more readable) and the rating. The trend line shows a slight increase in the rating as the number of votes increases. This was as per expectation. The reason for this could be that a larger number of people will vote for a movie if it is exceptionally good as opposed to a movie that is considered average. In this situation, the voters will give a higher rating to the movies and hence an increase in number of votes reflects positively on the rating of the movie.



Rating vs Runtime:

This graph depicts the relationship between the run-time of movies and the rating that viewers give the movie. At first glance, it seemed like movies that are longer (greater than 200 mins) are rated higher. But we soon realised that there were not enough data points in this region to say so conclusively. This relationship need not be causal. Several movies that are longer than 200 minutes are rated lesser than 8.5, so just looking at highly rated movies might cause a sampling bias.



Depicting the trend in number of votes received based on genre:

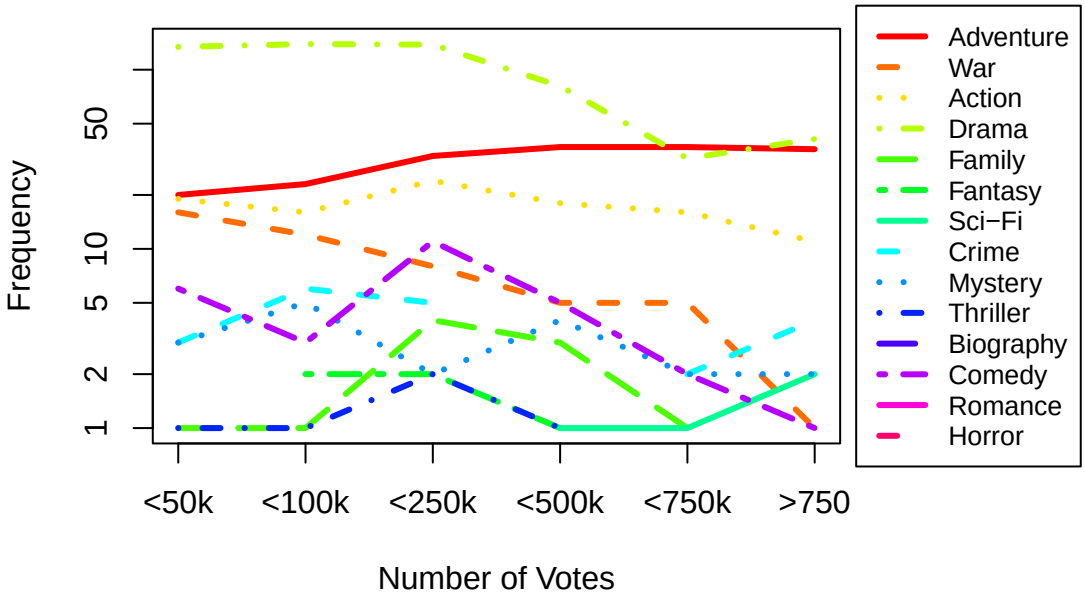
All plots made with genres have an intersection between genres (that is, a movie might belong to both "Action" and "Adventure") and hence the total number of movies would be over 1000.

We plot the y-axis as log of the frequency to make the graph readable.

The line graph shows the trend in the number of votes of different genres.

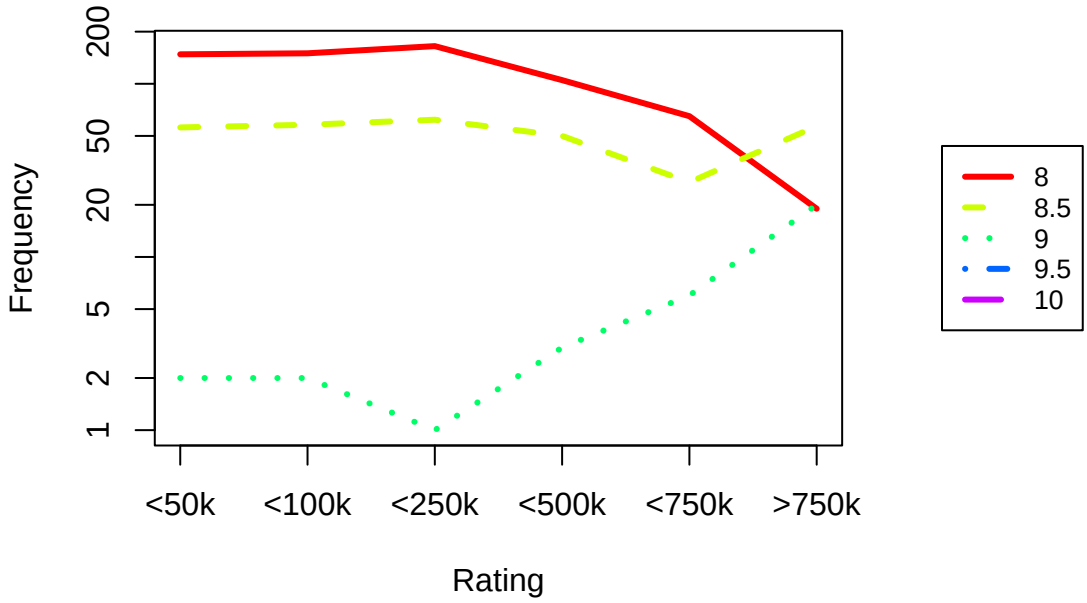
The maximum number of movies in the top 1000 list in the genre “Drama” has fewer than 250,000 votes. But the maximum number of movies from a single genre to have over 750,000 votes also belong to “Drama”. The number of votes of the genre “Adventure” rises with an

increase in the number of votes. Most movies in this genre have over 500,000 votes with a large number of movies in the bracket of 250,000 to 500,000 votes as well. A large number of movies that have over 500,000 votes, other than the two aforementioned genres, also belongs to "Action". Movies in the genre "Comedy" largely have less than 250,000 votes. It is interesting to note that movies in the genre "Crime" either have less than 250,000 votes or have over 750,000 votes.



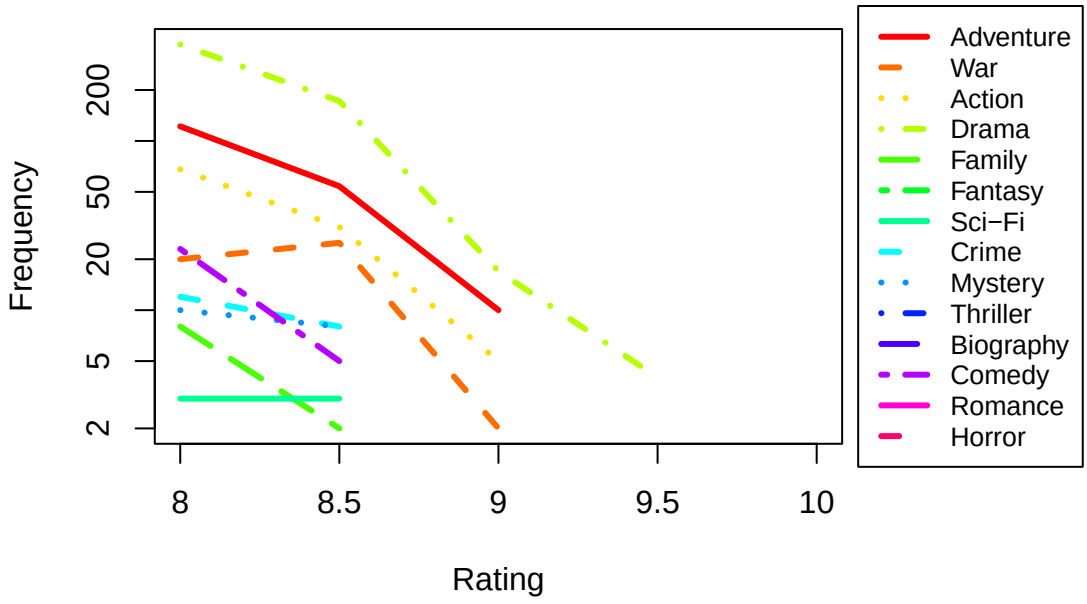
Depicting the trend in rating received based on number of votes:

A large number of movies with a rating of less than "8" have lesser than 250,000 votes. The number of movies rated less than "8" and having over 750,000 votes is lesser than the number of movies in the same rating bracket having between 250,000 and 750,000 votes and this is less than the number of movies in the same rating bracket having lesser than 250,000 votes. A large number of movies rated between "8" and "8.5" have lesser than 500,000 votes. But the number of movies with over 750,000 votes is individually larger than any other votes bracket for this rating bracket. Most movies rated over "8.5" have more than 750,000 votes. It is interesting to note that the movies rated over 9 either have between 100,000 votes and 250,000 votes or over 750,000 votes. There are no movies that are rated over 9.5.



Depicting the trend in rating based on genre:

Most movies on the top 1000 list in the genre "Drama" have a rating of either less than "8" or between "8" and "9". Most movies in the genre "Adventure" and "Action" also have either less than rating "8" or between "8" and "9". The only movies rated over "8.5" belong to the genres "Adventure", "War", "Action" or "Drama". the only movies rated over "9" belong to the genre "Drama". None of the movies is rated over "9.5". Among the genres, all the movies in the top 1000 list are rated less than "8.5" and largely less than "8".



Conclusion

We used R for web-scraping data from the IMDb list of top 1000 movies per viewer rating. This was done successfully except for one parameter, certification. We further structured the data and pushed the same into a CSV file for easy retrieval throughout the project. Using

the data obtained, we generated various plots to analyse the effect of different parameters of movies in generating the top 1000 list. Most of the plots depicted results as per expectation.

References

1. https://towardsdatascience.com/data-analysis-and-visualization-of-scraped-data-from-imdb-with-r-5d75e8191fc0
2. https://www.dataquest.io/blog/web-scraping-in-r-rvest/