Descriptive and Predictive Data Analysis for Movies

By: Christian Lane and Ryan Heiert

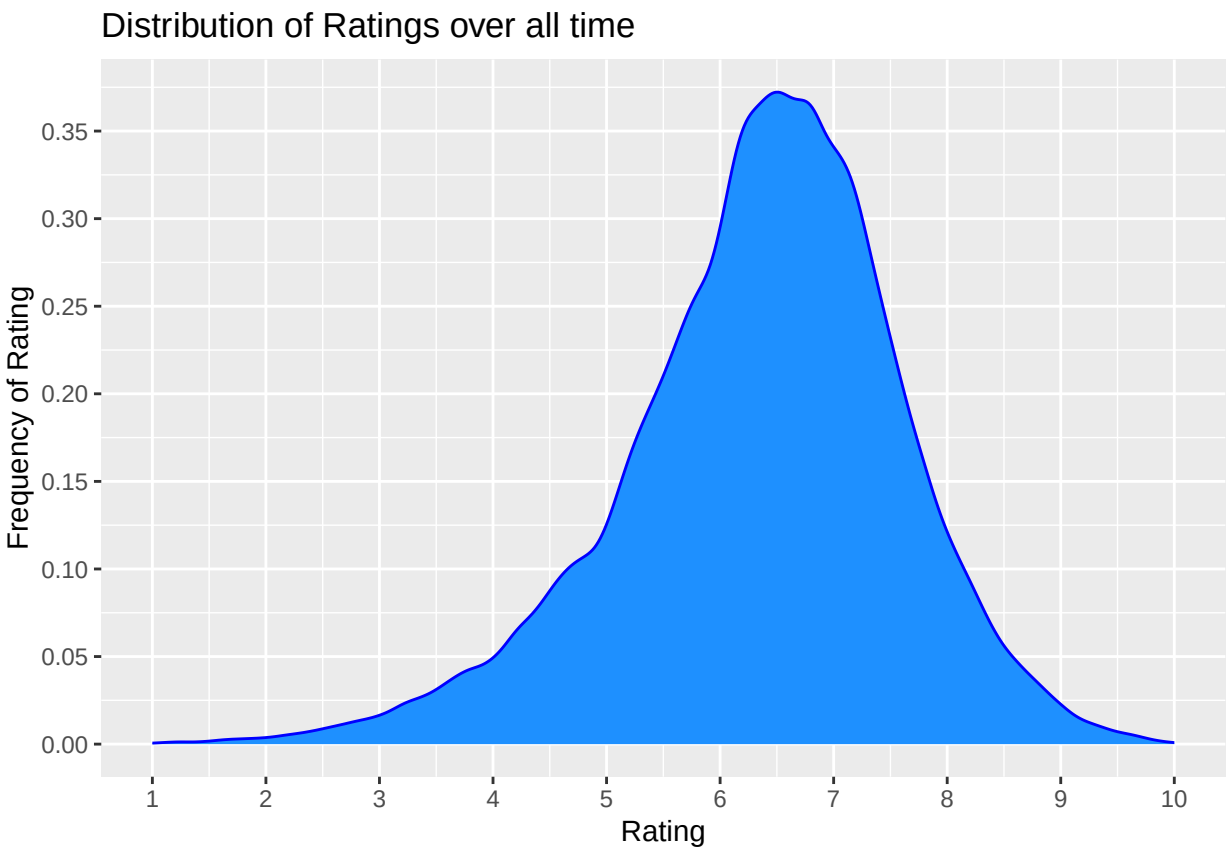
To start Analysis, we got our files from the IMDB Database, which can be found at datasets.imdbws.com

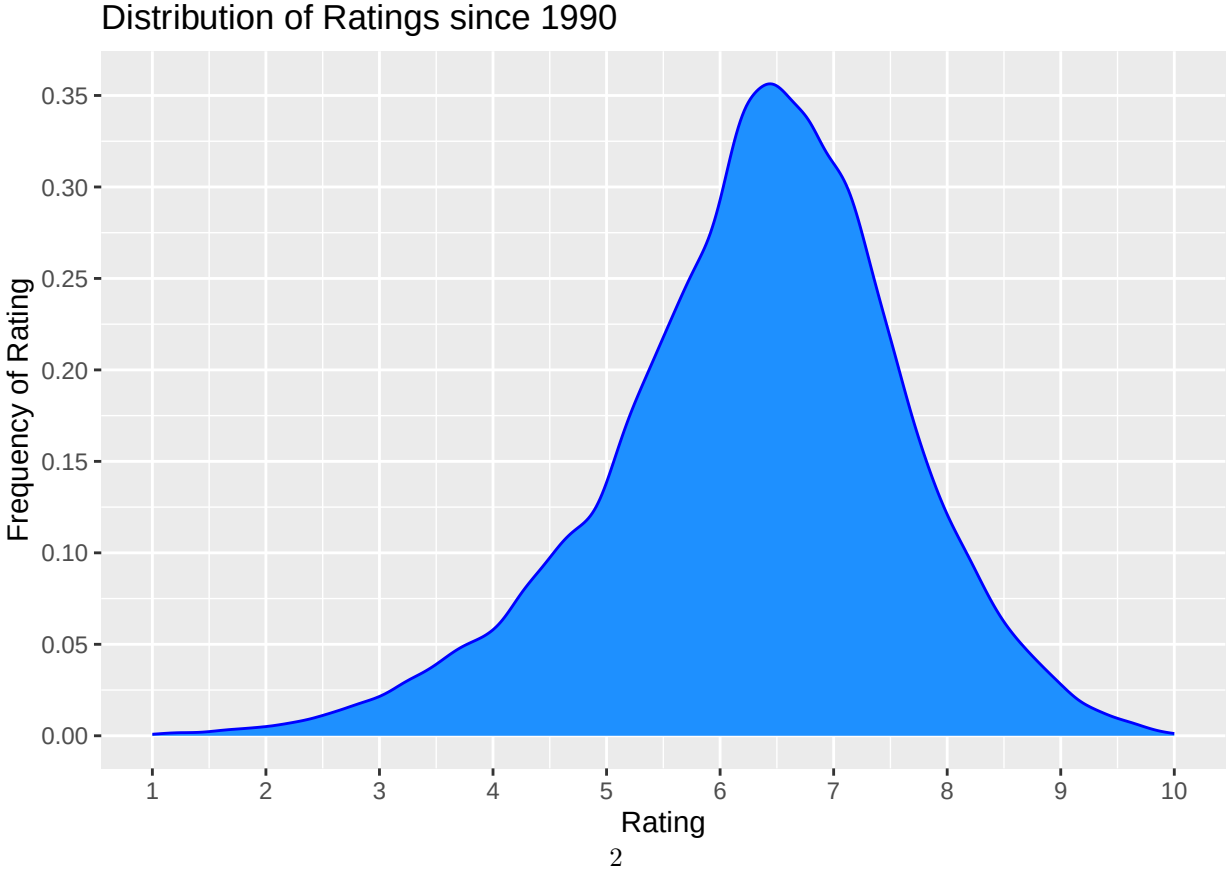
We got 3 different data sets, one for the movies basic info, one for ratings of movies, and one for writers of movies. After downloading we combine data into single object and clean data based on our specifications. We want only movies with runtime over 10 minutes and under 250. This is to exclude shorts and single entries that account for multiple movies (and so have an excessively long runtime).

Descriptive data analysis

First, we wondered what the average rating for movies overall was, and what the distribution is like for movies made since 1990 (“modern” movies).

Average Rating of Movies

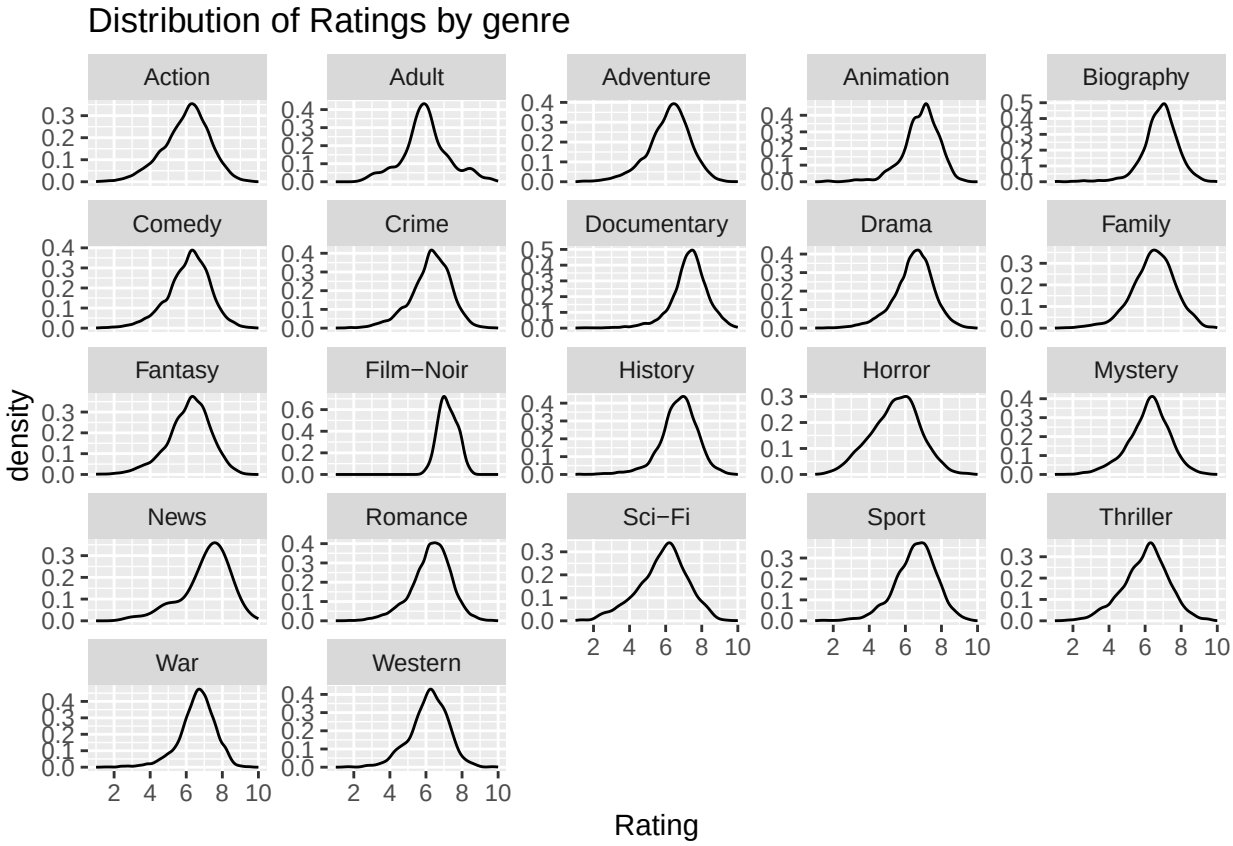




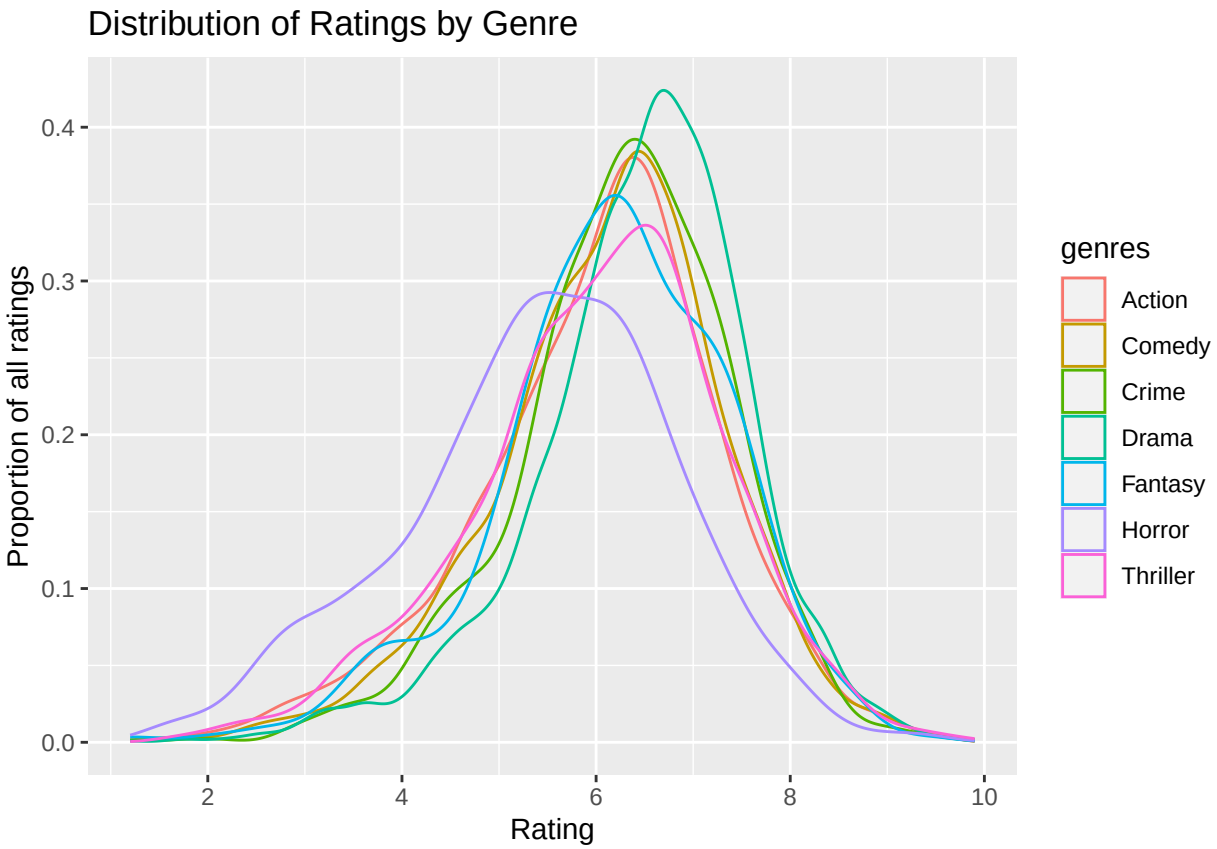
Pay close attention to the y axis scales. These distributions look almost identical but the modern movies actually have a slightly lower peak, at around the x value 6.5, which is also the most common rating.

Ratings based on Genre

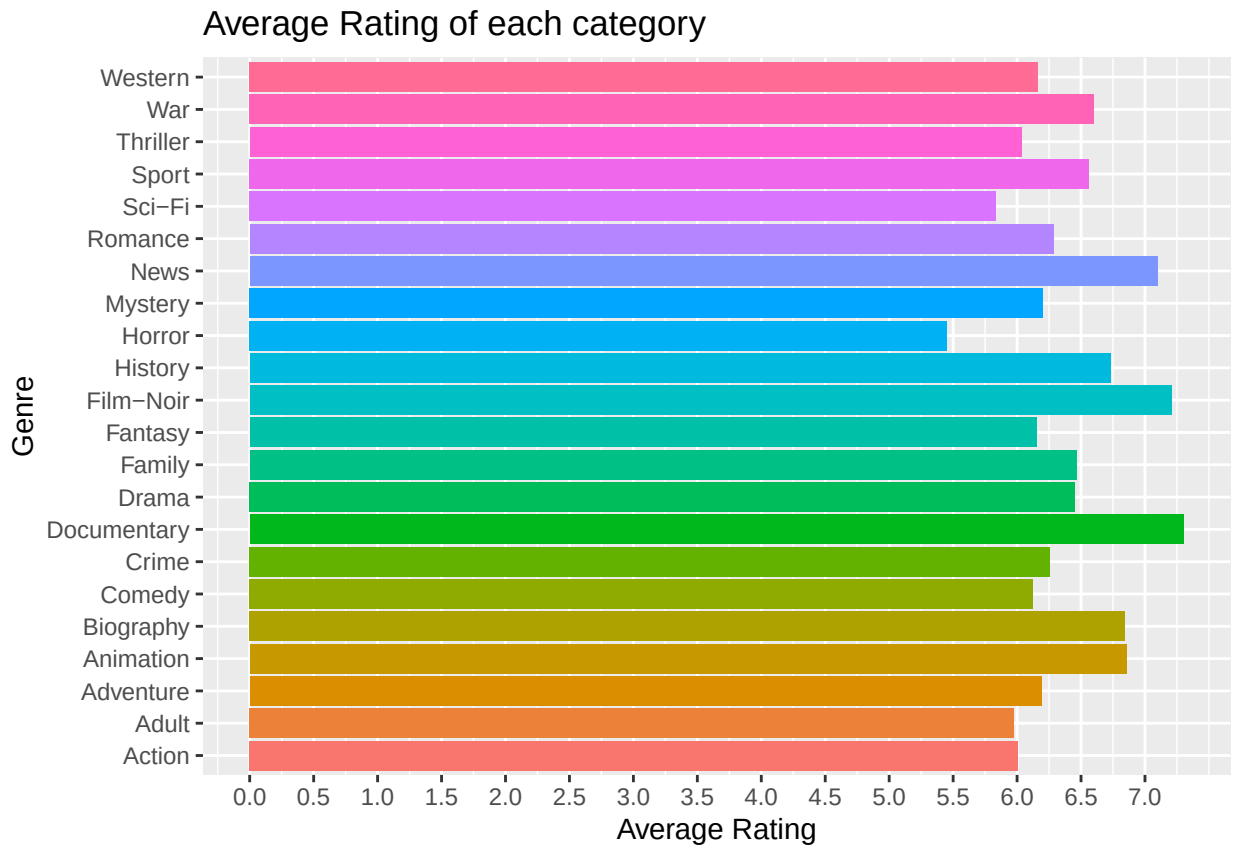
We then wondered what the most and least popular genres are



This is what that graph looks like with a few of the main genres overlayed with each other



From this we can learn that Horror is more disliked, and more varied in rating in general, and the most consistently liked genre is Drama. We can also conclude that the most consistently average categories are Action, Comedy, Thriller, Fantasy, and Crime.

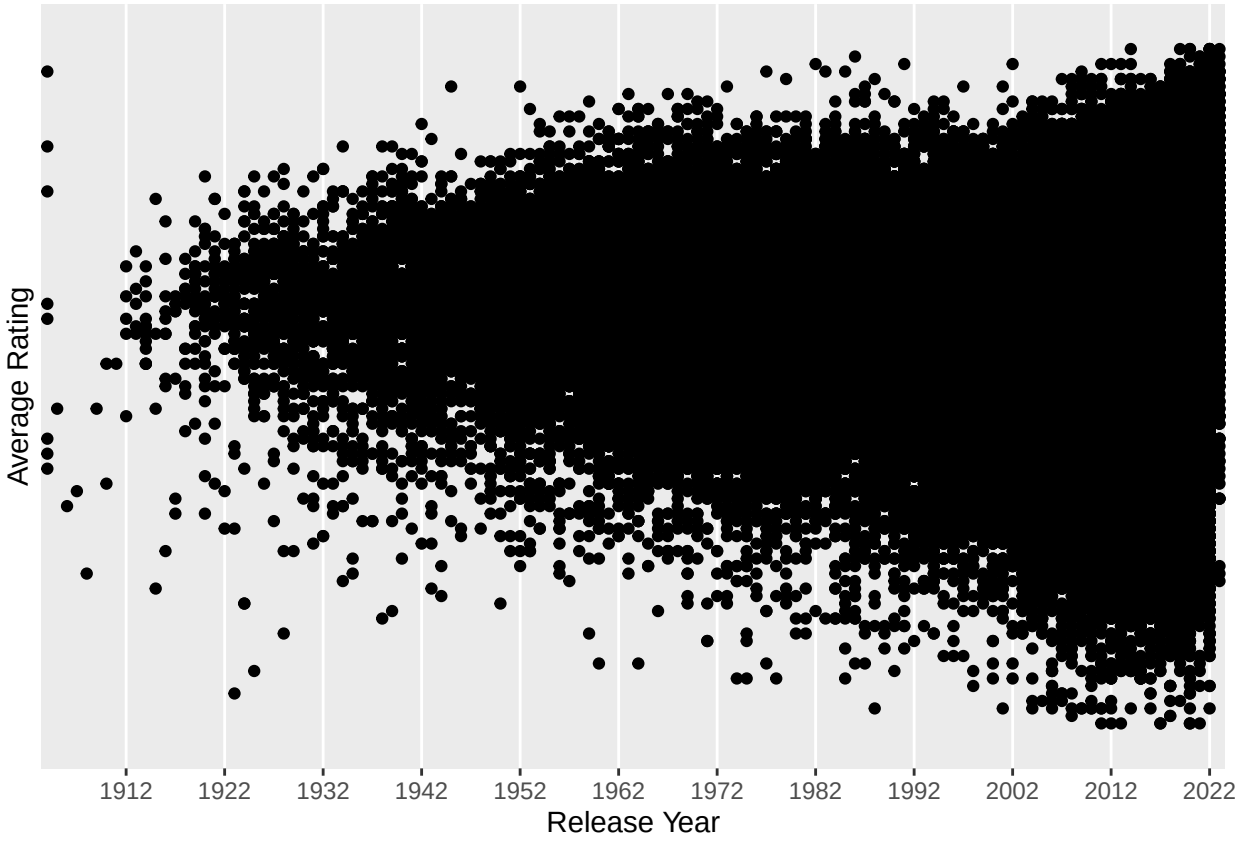


Linear Regression Model

Finding a Candidate

We first had to decide what variables we wanted to plot against each other, and why. The first thing that we came up with was plotting release year against average movie rating during that year, to see if critics had gotten more harsh over the years. We also came up with the idea of comparing the number of writers with a movie's average rating, to see if the number of writers a movie had had an influence on the ratings it received. If a movie's rating is indicative of its quality overall, then this could point to a correlation between the quality of the movie and the number of writers, which is interesting.

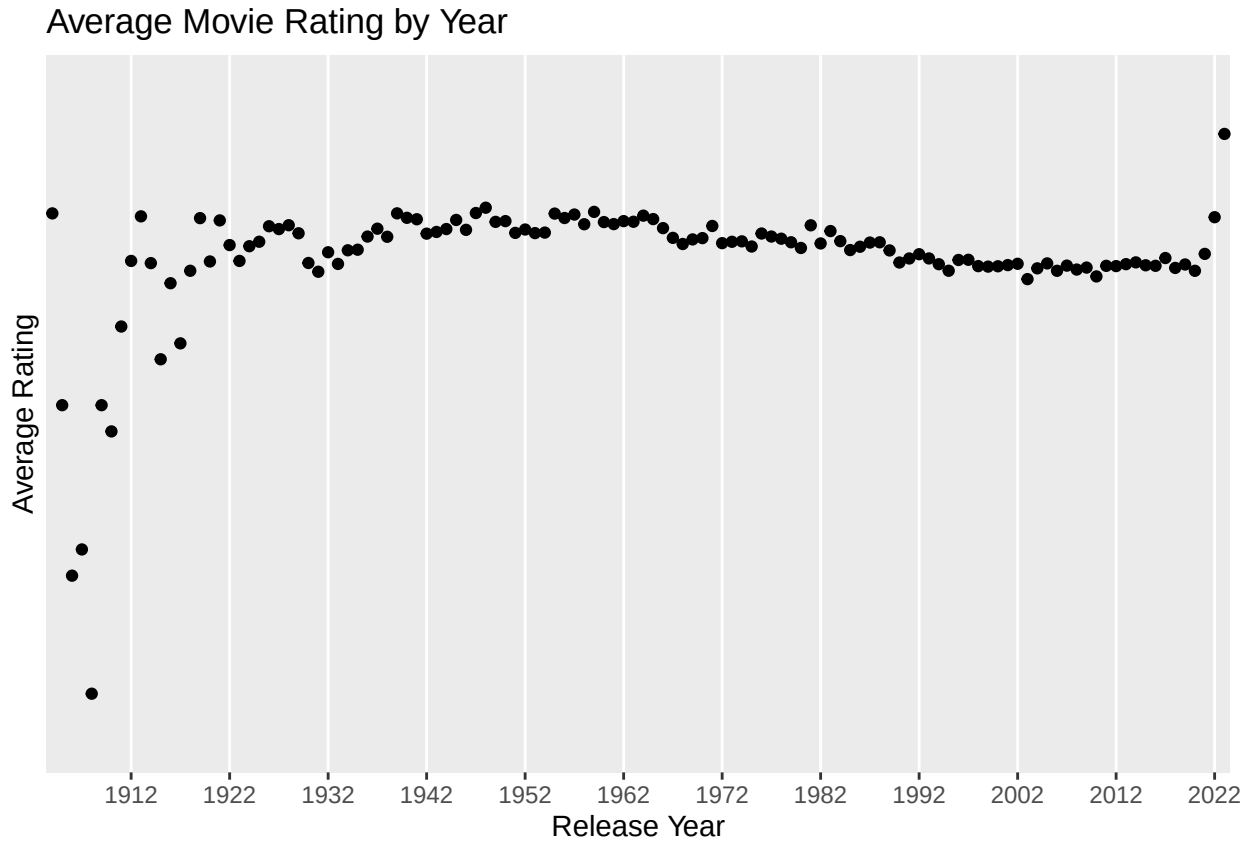
Plotting Release Year Against Average Movie Rating



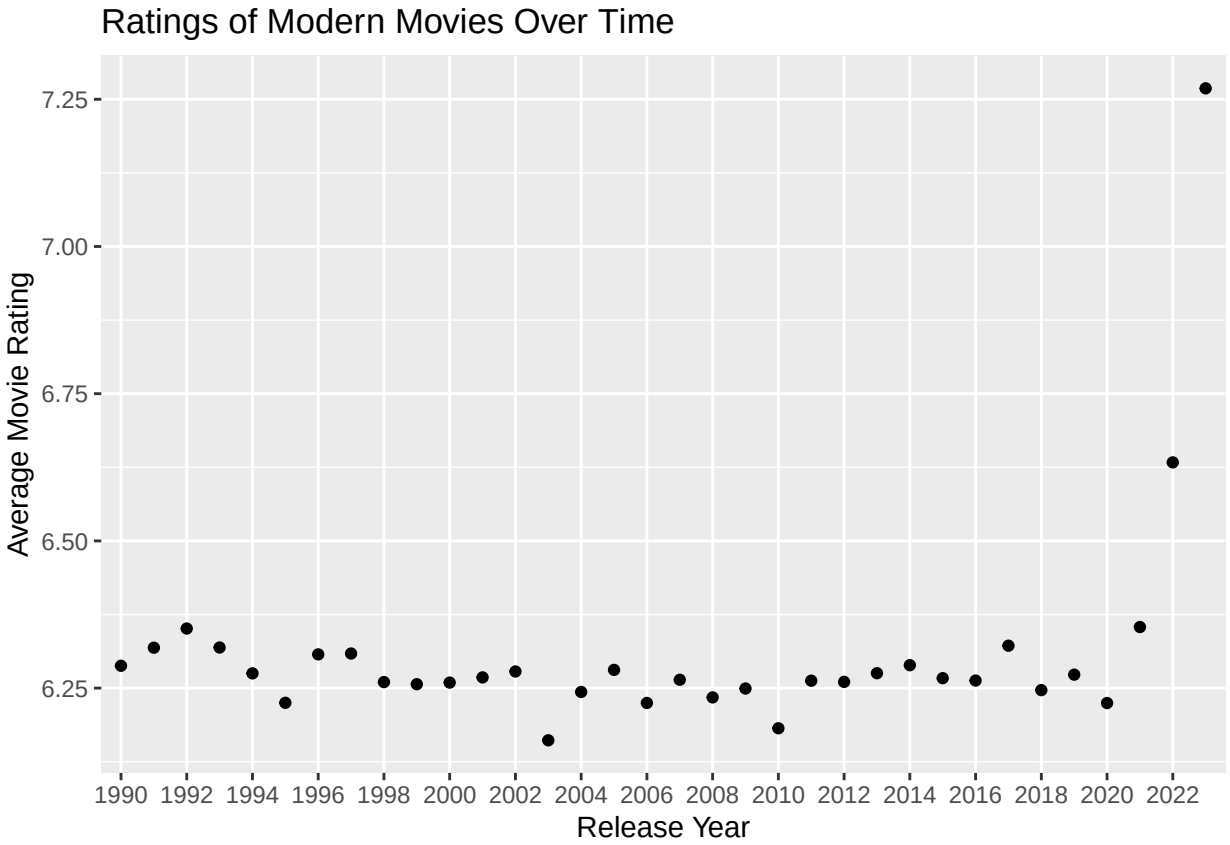
This is sort of "cone shaped". This is indicative of heteroskedasticity, which is a problem for data analysis. It is effectively an increase in the variance of the residuals of the dataset. This is caused by an increase in population variance over time. As times get more modern, there are simply more movies to give ratings, because modern conditions facilitate the development of a larger amount of ratings. This causes slight problems for the progression of our analysis.

It is possible to instead take the average of all ratings for all movies made in a given year and plot that against the year.

Plotting Average Rating of ALL Movies for a year vs Release year



This averaged data set is a lot easier to understand, and from this we may be able to find a correlation. When movies were first tracked for ratings, they started out pretty low, (around the year 1900) then seemed to level off for about 100 years with a slight downward trend. Then recently, it seems the average has gone up at what looks like an exponential rate. To explore this further, Below is a graph of this data, but modernized to include only movies that were made in 1990 or sooner.



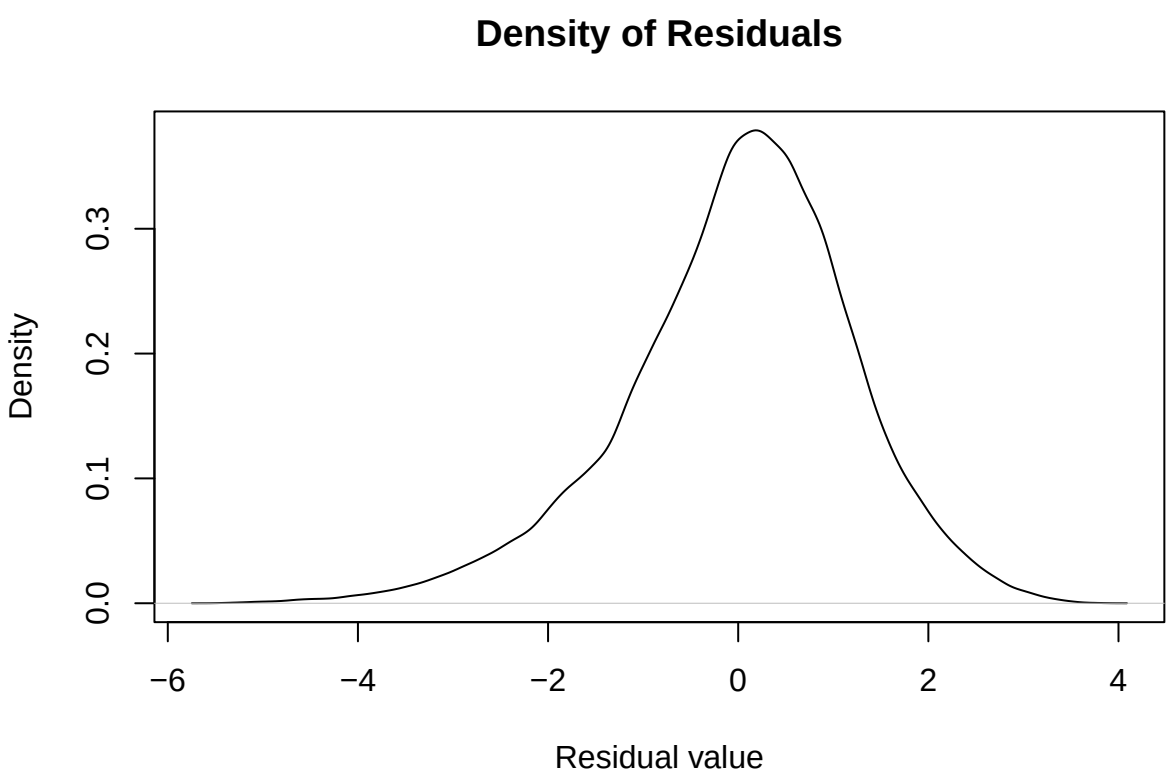
You can see here that for some reason modern movies seem to be on a trend upward in terms of average rating directly following a low in the year 2020, the same time as the pandemic. This could be because of the small number of movies that came out in 2020, and 2023's abnormally high rating may be accounted for by the relatively small number of movies that have come out in 2023 so far. This still does not account for the abnormally high average movie rating for 2022, so it is possible we are seeing a weird trend of abnormally good movies in the modern age.

Average Rating Plotted Against Movie Release Year

The residuals of this data set are useful to determine whether a naive approach is an accurate predictor in this context. Of all the residuals in the data set, these are their data:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.4333 -0.6809  0.1121  0.0000  0.8111  3.7753
```

They actually say it is OK for predictive analysis. The density plot for the residuals looks like:



Because residuals should be normally distributed in an unbiased data set, we seem to have an OK distribution here.

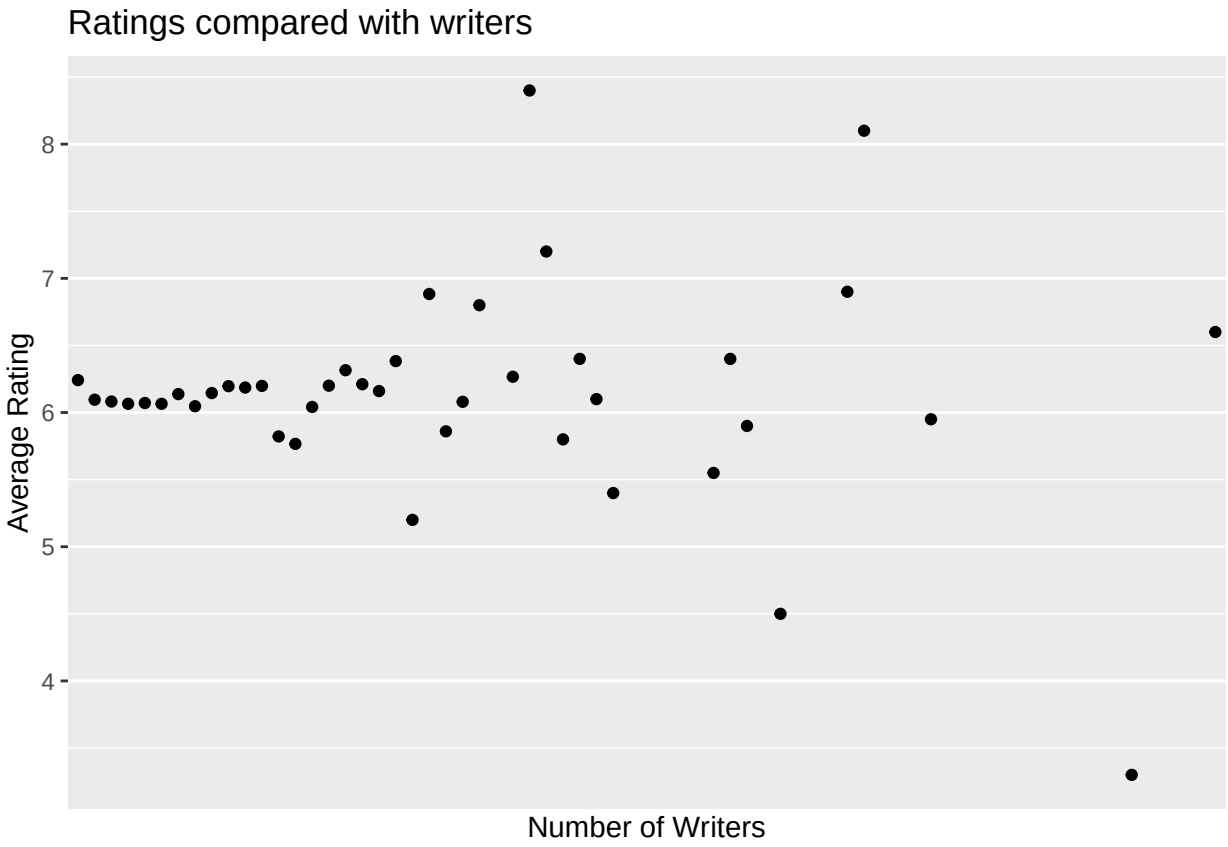
Now on to the R-squared values:

```
## [1] "R-squared: 0.0145443427733645"
```

```
## [1] "Adjusted R-squared: 0.0130914625997911"
```

As we can see, the adjusted R-squared value is 0.013, which implies that plotting these 2 variables against each other has almost no predictive validity. This makes sense if you think about it, release year should theoretically have nothing to do with the ratings a movie receives.

Plotting number of writers against rating



This scatter plot shows that there seems to be no correlation between number of writers listed and rating. From around 1-10 writers, rating stays constant at about 6.0. At around 10-20 writers, a drop followed by a weak upward trend can be seen. However, a vast majority of movies had between 1-3 writers, and very few movies listed $>$ 10 writers. As number of writers increased, the amount of data points for that value decreased significantly. We believe this resulted in more random data points, and that this trend appears only by chance.

Conclusion

Our predictive data analysis models came to multiple conclusions separately, one possibly that there is a trend in modern movies to have abnormally high ratings, another correlating the number of writers of a movie with that movie's quality. Unfortunately, not many real useful conclusions were reached following our analysis of the data.

References

The only reference we have is the IMDB Database We also used the Tidyverse and ggplot2 libraries.