Descriptive and Predictive Data Analysis for Movies

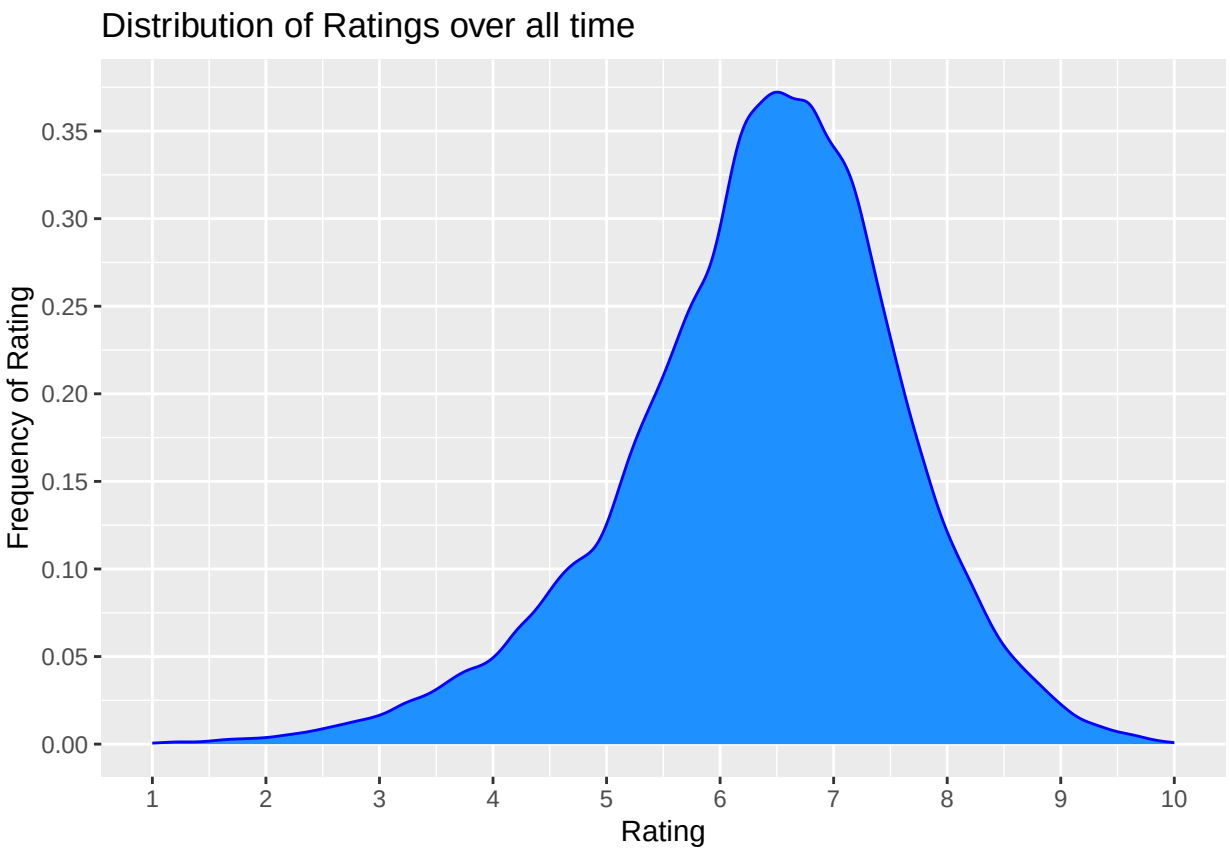
To start Analysis, we got our files from the IMDB Database, which can be found at datasets.imdbws.com

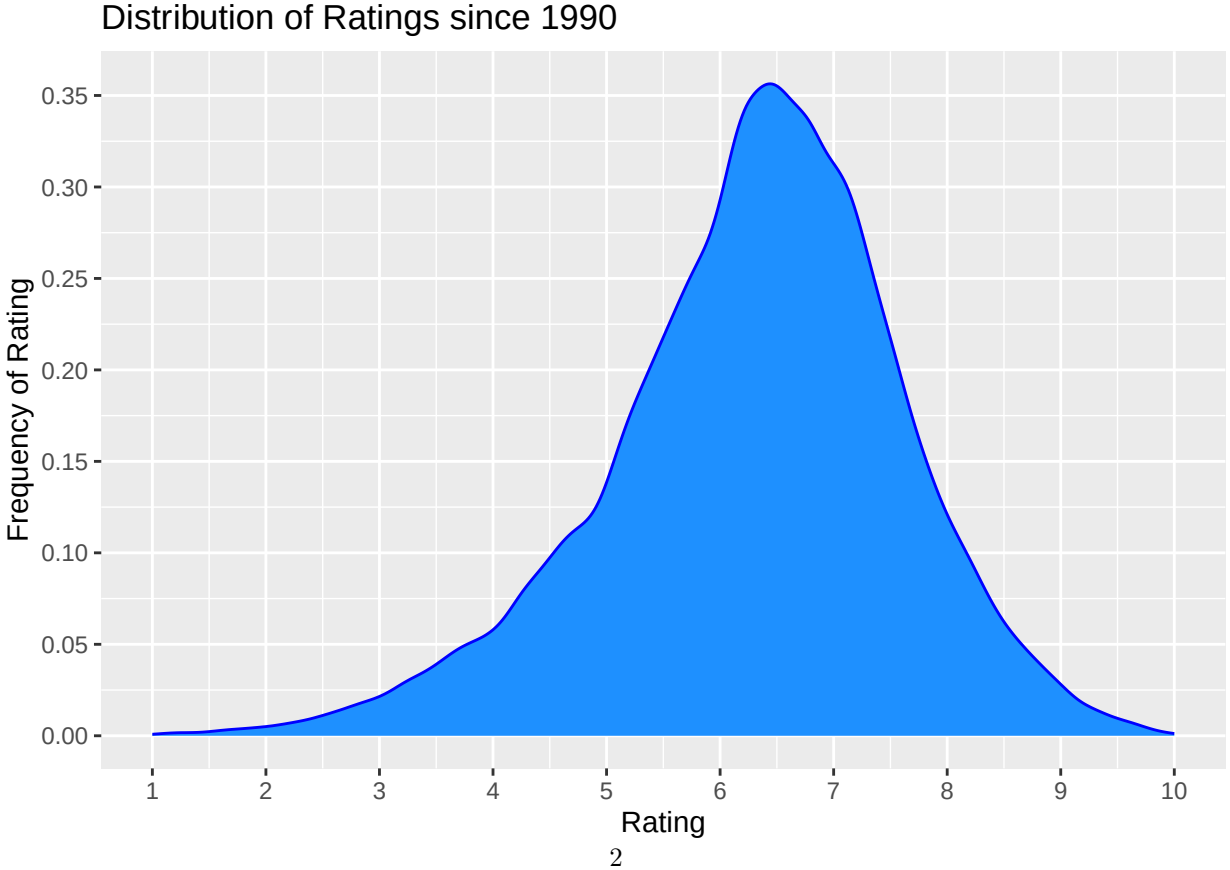
We got 3 different data sets, one for the movies basic info, one for ratings of movies, and one for writers of movies. After downloading we combine data into single object and clean data based on our specifications. We want only movies with runtime over 10 minutes and under 250. This is to exclude shorts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

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

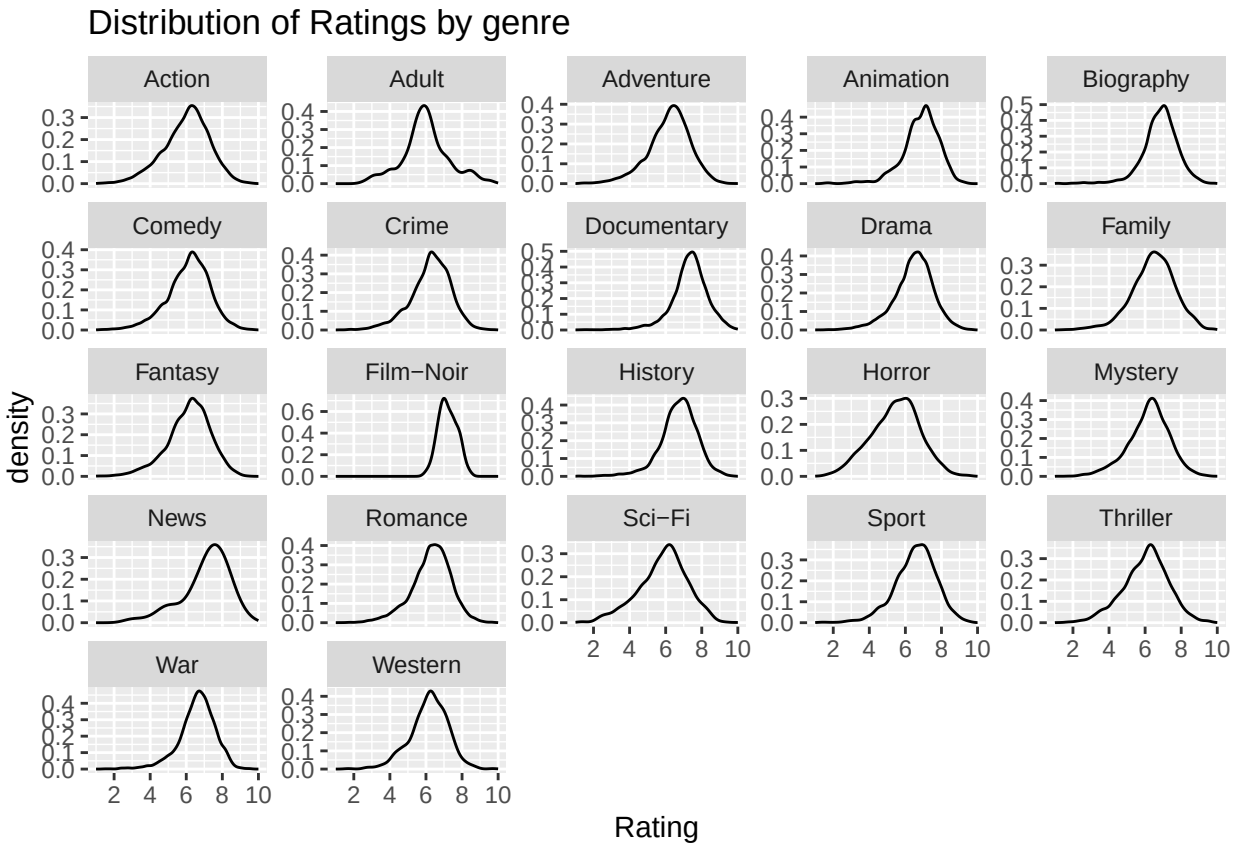




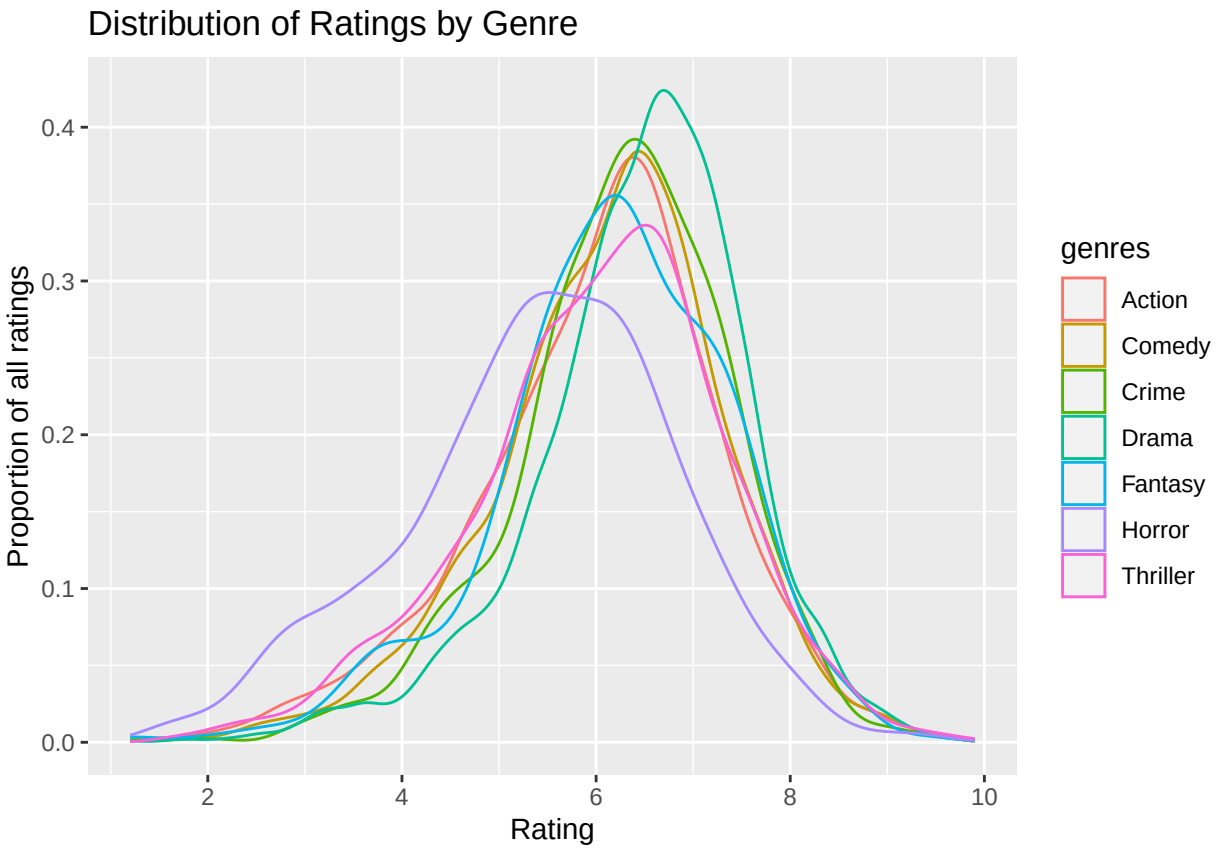
As we can see, most common ratings for movies are about 6.5. The distribution almost changed none when made to reflect more modern movies. The only visible change is a slightly more spread average rating, the peak at 6.5 lowers a bit in the modern version and that is reflected throughout the rest of the graph

Ratings based on Genre

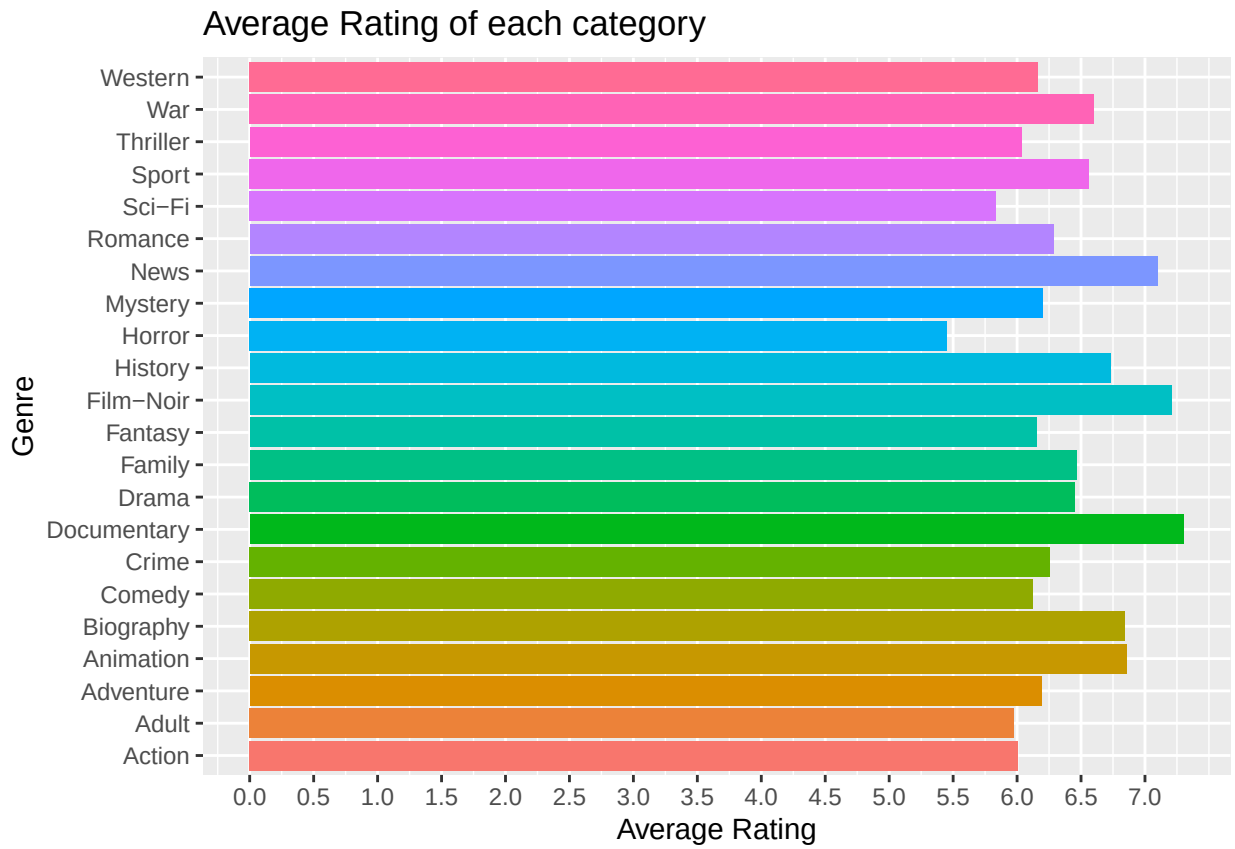
We then wondered what the most and least popular genres are



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

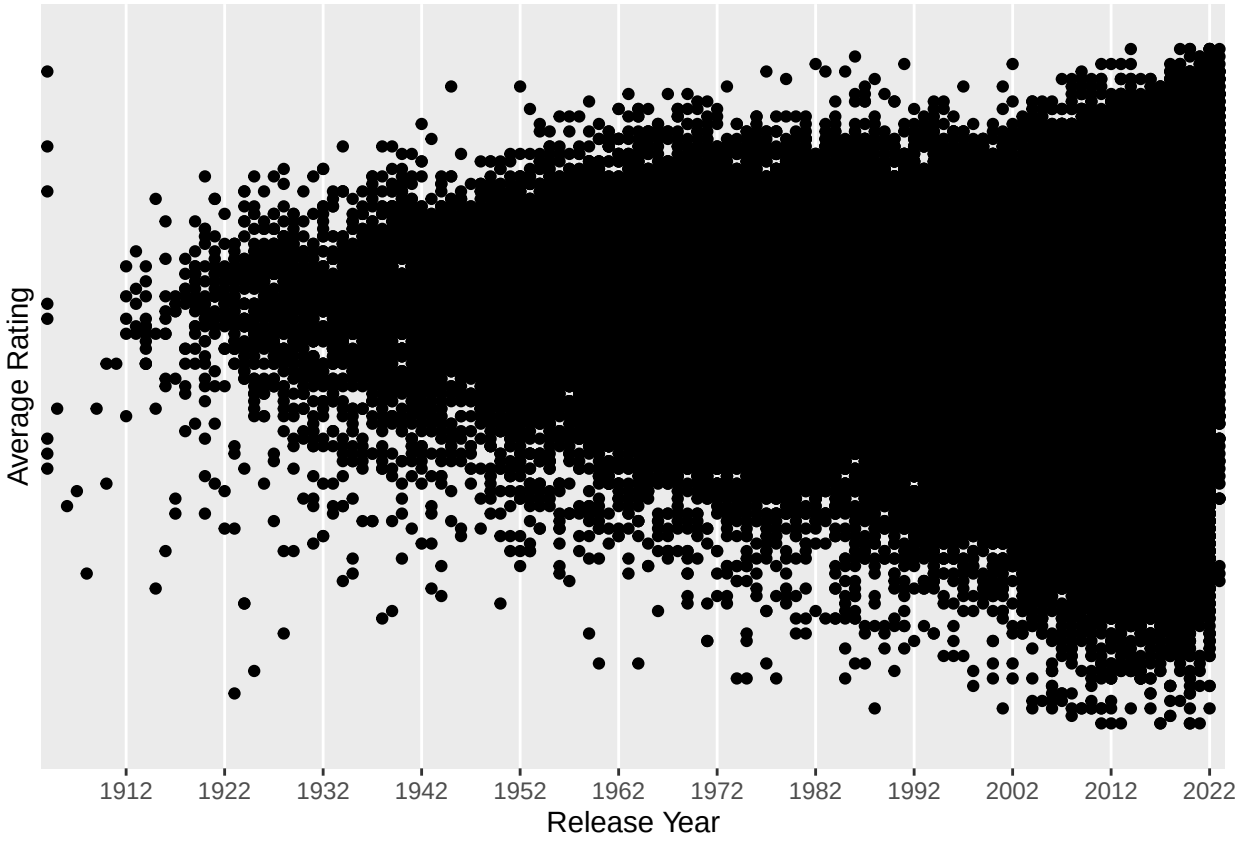


From this we can learn that Horror is more disliked, and more varied in rating in general, and the most consistently liked genre is Drama. We can also conclude that the most consistently average categories are Action, Comedy, Thriller, Fantasy, and Crime.


Average Rating of each category
Western
War
Thriller
Sport
Sci−Fi
Romance
News
Mystery
Horror
History
Film−Noir
Fantasy
Family
Drama
Documentary
Crime
Comedy
Biography
Animation
Adventure
Adult
Action
Genre
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Average Rating

Linear Regression Model

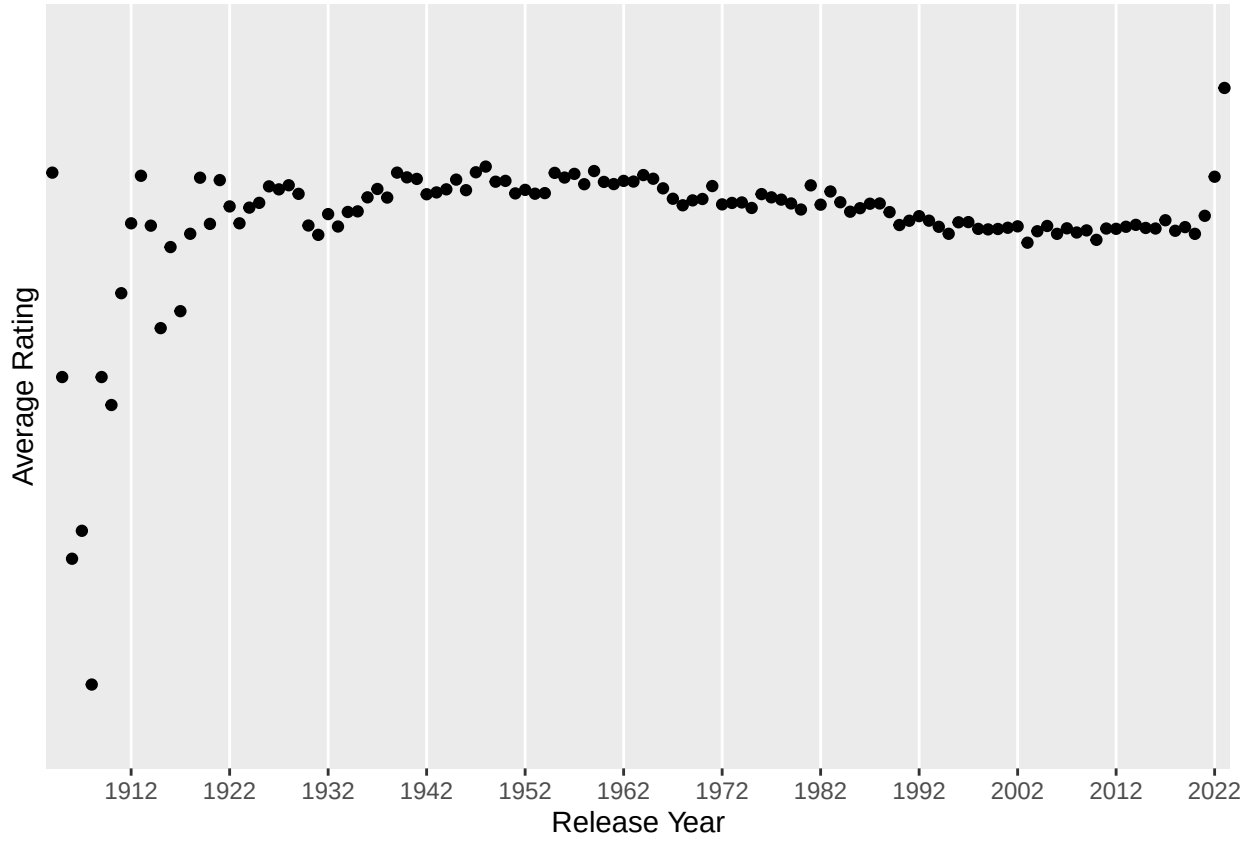
Plotting Release Year Against Average Movie Rating



This is sort of "cone shaped".This is caused by an increase in population variance over time. As times get more modern, there are simply more movies to give 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.

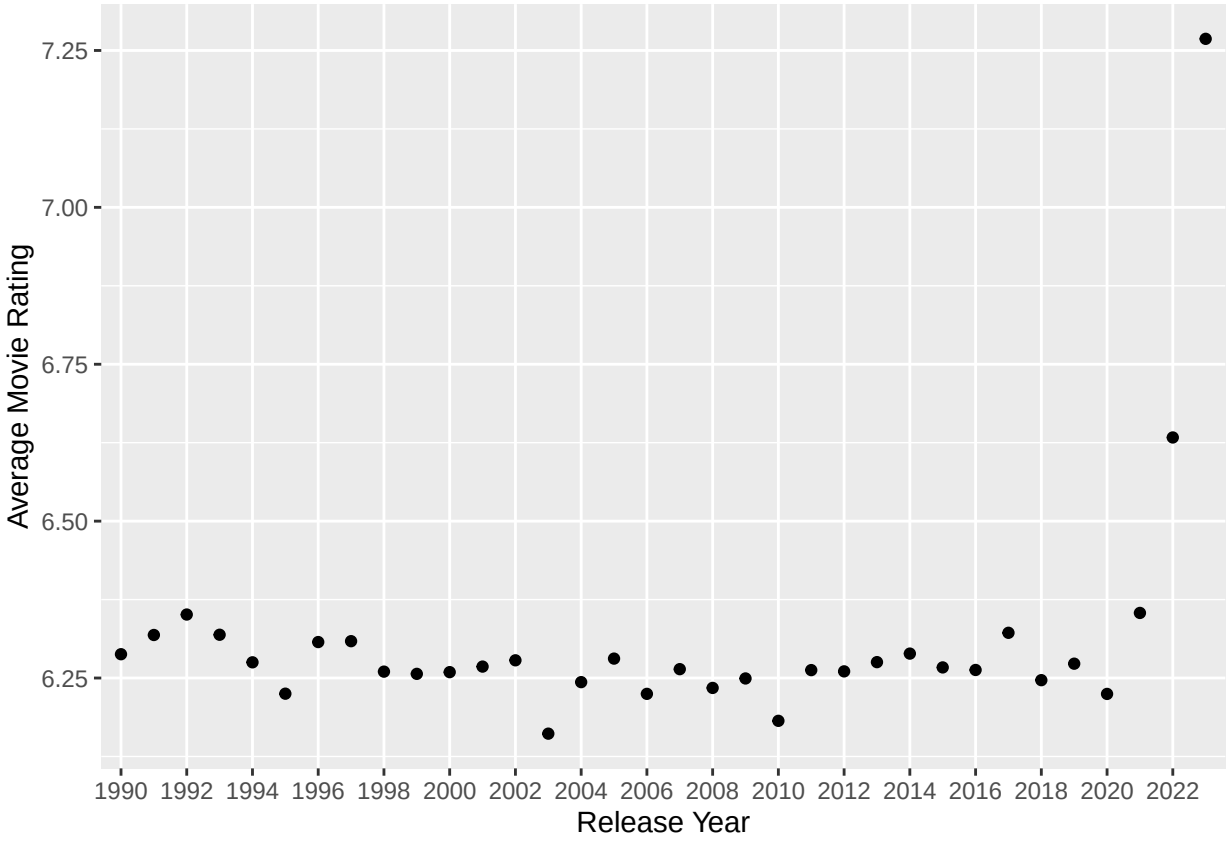
Plotting Average Rating of ALL Movies for a year vs Release year



These data are a lot easier to understand.

From this we can see that when movies were first tracked for ratings, they started out pretty low, 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

Ratings of Modern Movies over Time



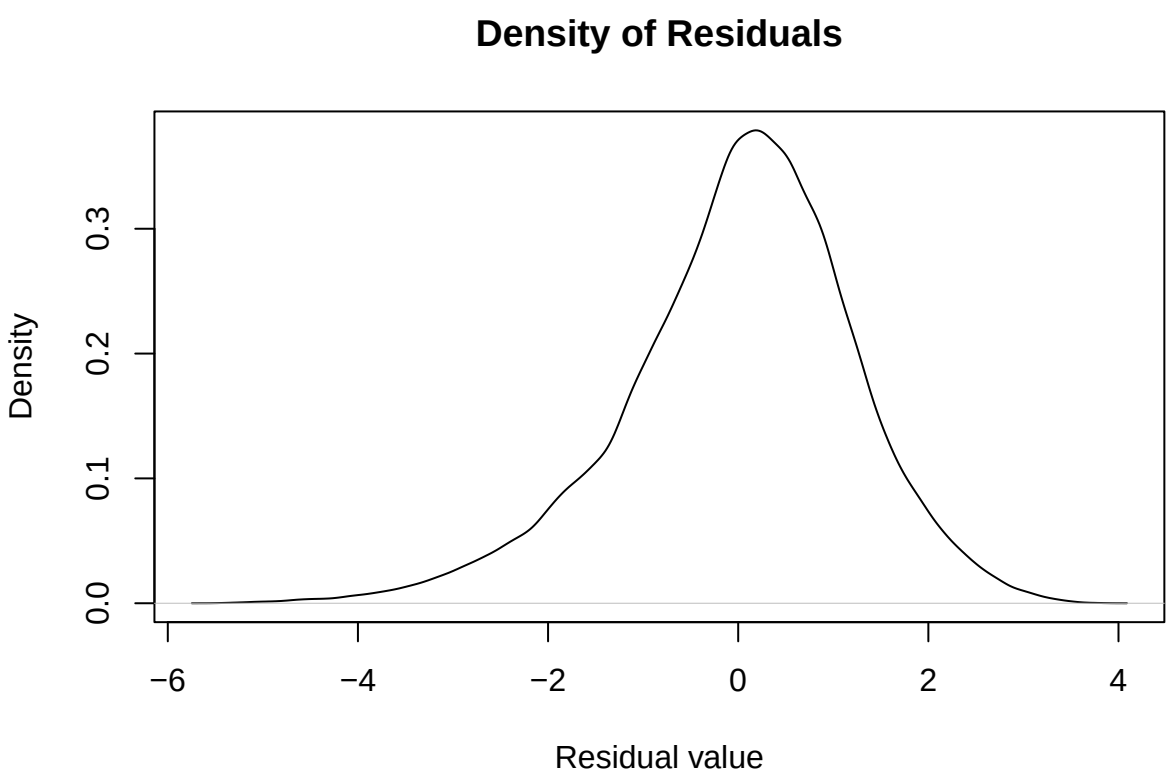
You can see here that for some reason modern movies seem to be on a trend upward in terms of average rating directly following around 2020. This could be because of the small number of movies that have come out in 2023 so far, but that would not accound for the abnormally high average movie rating for 2022.

Finding the residuals of average rating plotted against start year

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
```

These residuals actually say it is ok for predictive analysis. Look at what 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 one here.

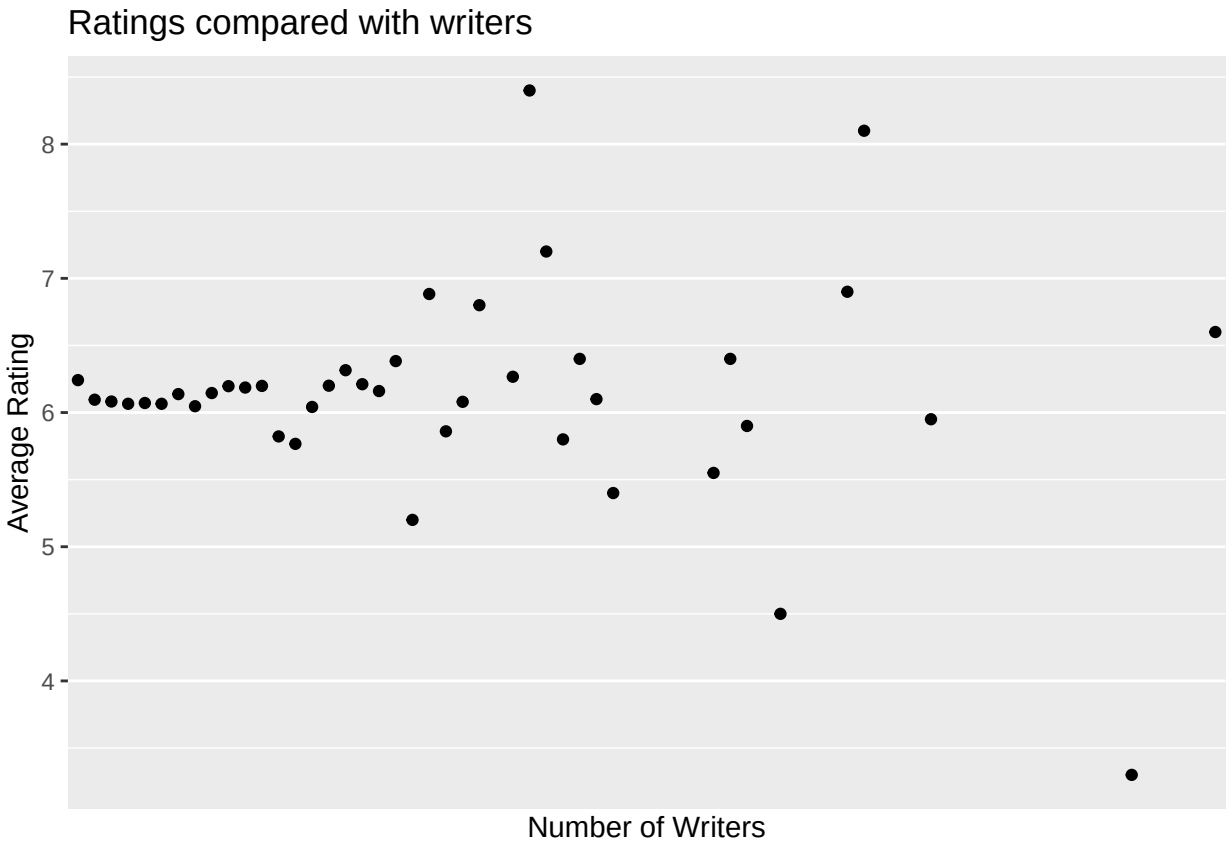
Now on to the R-squared values:

```
## [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 recieves.

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.

Made by Christian Lane and Ryan Heiert