IMDb Film Data Analysis by Moe Malik

Introduction

IMDb - shorthand for the Internet Movie Database - is an online database of all things film, TV, and to a lesser extent, video games. The database has more than 4 million records of entertainment, along with almost 8 million records of the individuals who have worked on the content. Be it a silent film classic from the 1920s, or a video game dud from 2016, chances are you could find information about it on IMDb.com. As of March 12th 2017, it is 55th most popular website in the world.

Courtesy of the IMDB 5000 Movie Dataset on Kaggle, we have access to data on 5000 film titles for analysis.

What we're interested in exploring and investigating as we dive into the data is the following:

What variables correlate with a low or high IMDb score?

IMDb scores are a rating of a given film, rated on a scale from 0.0 to 10.0. As per the IMDb FAQ, the rating is calculated by taking all the individual votes cast by registered IMDb users, and then calculating a weighted average (IMDb does not close how/when/why scores are weighted).

As a big movie buff who use to have IMDb as his Firefox's homepage for 2 years, IMDb scores were at one point *the* way I determined whether a film was worth my time. In particular, I combed through the IMDb Top 250 every week - a list of the top 250 rated films on the website - looking for a new, quality film to watch. However, I never really took a close look if there were any characteristics were frequently shared among highly rated films. Analyzing this IMDb dataset from Kaggle is our chance to do just that.

Data Summary and Wrangling

Let's start off by taking a look at a summary of the dataset.

```
##              color              director_name  num_critic_for_reviews
##                 :  19                    : 104   Min.   :  1.0
##   Black and White: 209   Steven Spielberg:  26   1st Qu.: 50.0
##  Color          :4815   Woody Allen     :  22   Median :110.0
##                         Clint Eastwood  :  20   Mean   :140.2
##                         Martin Scorsese :  20   3rd Qu.:195.0
##                         Ridley Scott    :  17   Max.   :813.0
##                         (Other)         :4834   NA's   :50
##     duration     director_facebook_likes actor_3_facebook_likes
##  Min.   :  7.0   Min.   :    0.0         Min.   :    0.0
##  1st Qu.: 93.0   1st Qu.:    7.0         1st Qu.:  133.0
##  Median :103.0   Median :   49.0         Median :  371.5
##  Mean   :107.2   Mean   :  686.5         Mean   :  645.0
##  3rd Qu.:118.0   3rd Qu.:  194.5         3rd Qu.:  636.0
##  Max.   :511.0   Max.   :23000.0         Max.   :23000.0
##  NA's   :15      NA's   :104             NA's   :23
##           actor_2_name  actor_1_facebook_likes     gross
##  Morgan Freeman :  20   Min.   :     0        Min.   :      162
##  Charlize Theron:  15   1st Qu.:   614        1st Qu.:  5340988
##  Brad Pitt      :  14   Median :   988        Median : 25517500
##                 :  13   Mean   :  6560        Mean   : 48468408
##  James Franco   :  11   3rd Qu.: 11000        3rd Qu.: 62309438
##  Meryl Streep   :  11   Max.   :640000        Max.   :760505847
##  (Other)        :4959   NA's   :7             NA's   :884
```

```
##                  genres                actor_1_name
##  Drama               : 236   Robert De Niro   :  49
##  Comedy              : 209   Johnny Depp      :  41
##  Comedy|Drama        : 191   Nicolas Cage     :  33
##  Comedy|Drama|Romance: 187   J.K. Simmons     :  31
##  Comedy|Romance      : 158   Bruce Willis     :  30
##  Drama|Romance       : 152   Denzel Washington:  30
##  (Other)             :3910   (Other)          :4829
##                     movie_title   num_voted_users
##  Ben-HurÂ                  :   3   Min.   :      5
##  HalloweenÂ                :   3   1st Qu.:   8594
##  HomeÂ                     :   3   Median :  34359
##  King KongÂ                :   3   Mean   :  83668
##  PanÂ                      :   3   3rd Qu.:  96309
##  The Fast and the FuriousÂ :   3   Max.   :1689764
##  (Other)                   :5025
##  cast_total_facebook_likes         actor_3_name  facenumber_in_poster
##  Min.   :     0                          :  23   Min.   : 0.000
##  1st Qu.:  1411           Ben Mendelsohn:   8   1st Qu.: 0.000
##  Median :  3090           John Heard    :   8   Median : 1.000
##  Mean   :  9699           Steve Coogan  :   8   Mean   : 1.371
##  3rd Qu.: 13756           Anne Hathaway :   7   3rd Qu.: 2.000
##  Max.   :656730           Jon Gries     :   7   Max.   :43.000
##                           (Other)       :4982   NA's   :13
##                                                                                       plot_keywords
##                                                                                              : 153
##  based on novel                                                                              :   4
##  1940s|child hero|fantasy world|orphan|reference to peter pan                                :   3
##  alien friendship|alien invasion|australia|flying car|mother daughter relationship:   3
##  animal name in title|ape abducts a woman|gorilla|island|king kong                           :   3
##  assistant|experiment|frankenstein|medical student|scientist                                 :   3
##  (Other)                                                                                     :4874
##                                               movie_imdb_link
##  http://www.imdb.com/title/tt0077651/?ref_=fn_tt_tt_1:   3
##  http://www.imdb.com/title/tt0232500/?ref_=fn_tt_tt_1:   3
##  http://www.imdb.com/title/tt0360717/?ref_=fn_tt_tt_1:   3
##  http://www.imdb.com/title/tt1976009/?ref_=fn_tt_tt_1:   3
##  http://www.imdb.com/title/tt2224026/?ref_=fn_tt_tt_1:   3
##  http://www.imdb.com/title/tt2638144/?ref_=fn_tt_tt_1:   3
##  (Other)                                             :5025
##  num_user_for_reviews     language         country       content_rating
##  Min.   :   1.0       English :4704   USA      :3807   R        :2118
##  1st Qu.:  65.0       French  :  73   UK       : 448   PG-13    :1461
##  Median : 156.0       Spanish :  40   France   : 154   PG       : 701
##  Mean   : 272.8       Hindi   :  28   Canada   : 126            : 303
##  3rd Qu.: 326.0       Mandarin:  26   Germany  :  97   Not Rated: 116
##  Max.   :5060.0       German  :  19   Australia:  55   G        : 112
##  NA's   :21           (Other) : 153   (Other)  : 356   (Other)  : 232
##      budget            title_year   actor_2_facebook_likes   imdb_score
##  Min.   :2.180e+02   Min.   :1916   Min.   :     0         Min.   :1.600
##  1st Qu.:6.000e+06   1st Qu.:1999   1st Qu.:   281         1st Qu.:5.800
##  Median :2.000e+07   Median :2005   Median :   595         Median :6.600
##  Mean   :3.975e+07   Mean   :2002   Mean   :  1652         Mean   :6.442
##  3rd Qu.:4.500e+07   3rd Qu.:2011   3rd Qu.:   918         3rd Qu.:7.200
```

```
## Max.   :1.222e+10   Max.   :2016   Max.   :137000         Max.   :9.500
## NA's   :492         NA's   :108    NA's   :13
##  aspect_ratio   movie_facebook_likes
## Min.   : 1.18   Min.   :     0
## 1st Qu.: 1.85   1st Qu.:     0
## Median : 2.35   Median :   166
## Mean   : 2.22   Mean   :  7526
## 3rd Qu.: 2.35   3rd Qu.:  3000
## Max.   :16.00   Max.   :349000
## NA's   :329
```

Great. So we see the dataset has 5043 observations across 28 different film related variables.

Before we can continue with the analysis, there's some cleaning we should do.

Regarding the content ratings, I noticed two things in the summary - there were 303 instances of apparently blank ratings, and that there were "Other" ratings not shown. Let's see what the "Other" consist of.

```
##
##            Approved         G        GP         M     NC-17 Not Rated
##       303        55       112         6         5         7       116
##    Passed        PG     PG-13         R     TV-14      TV-G     TV-MA
##         9       701      1461      2118        30        10        20
##     TV-PG      TV-Y     TV-Y7   Unrated         X
##        13         1         1        62        13
```

It should be first noted that MPAA ratings for films you'd see in the cinema (in the United States) are G, PG, PG-13, R, and very rarely, NC-17. "GP", "M", and "X" are old MPAA ratings that preceded the current rating system. While outdated, they still give us a sense of what a film's content was so they should be kept. However, there are some problematic cases we should be aware of:

1. It looks like the data included instances of TV ratings, meaning this data likely includes made-for-television movies - TV-Y, TV-Y7, TV-G, TV-PG, TV14, and TV-MA.
2. 'Approved', 'Passed', "Not Rated", and 'Unrated' lacks information on a film's content.
3. The films with a blank entry for rating. Looking up the films from the dataset with no rating on IMDb, I came across mainly television series and foreign films.

To ensure we'll be able to accurately analyze the relationship between IMDb scores and content_ratings, we should remove the aforementioned problematic cases from the dataset.

```
##
##     G    GP     M NC-17    PG PG-13     R     X
##   112     6     5     7   701  1461  2118    13
```

Looks like we have only ratings related to theatrical films now.

I also want to dive into one more categorical variable we'll be working with later during analysis - Country.

```
##
##                                 Afghanistan            Argentina
##                      0                    1                    3
##                  Aruba            Australia              Bahamas
##                      1                   50                    1
##                Belgium               Brazil             Bulgaria
##                      3                    6                    1
##               Cambodia             Cameroon               Canada
##                      0                    0                  101
##                  Chile                China             Colombia
```

```
##                    1                   21                    1
##       Czech Republic              Denmark   Dominican Republic
##                    3                    7                    1
##                Egypt              Finland               France
##                    0                    0                  115
##              Georgia              Germany               Greece
##                    1                   86                    1
##            Hong Kong              Hungary              Iceland
##                   16                    2                    1
##                India            Indonesia                 Iran
##                    7                    1                    3
##              Ireland               Israel                Italy
##                   10                    2                   15
##                Japan                Kenya           Kyrgyzstan
##                   14                    0                    1
##                Libya               Mexico          Netherlands
##                    1                   14                    4
##             New Line          New Zealand              Nigeria
##                    1                   12                    0
##               Norway        Official site             Pakistan
##                    5                    1                    0
##               Panama                 Peru          Philippines
##                    1                    1                    0
##               Poland              Romania               Russia
##                    2                    1                    5
##             Slovakia             Slovenia         South Africa
##                    1                    0                    7
##          South Korea         Soviet Union                Spain
##                   10                    1                   29
##               Sweden          Switzerland               Taiwan
##                    3                    1                    1
##             Thailand               Turkey                   UK
##                    5                    0                  392
## United Arab Emirates                  USA         West Germany
##                    0                 3446                    3
```

It looks as if we have several instances of countries with '0' for their value - This would have happened if rows that included these values as countries were removed with a subset. In this case, we'll have to re-factor the column like we did with content_ratings. Let's do that and then see how summary of the data looks again.

```
##
##        Afghanistan          Argentina              Aruba
##                  1                  3                  1
##          Australia            Bahamas            Belgium
##                 50                  1                  3
##             Brazil           Bulgaria             Canada
##                  6                  1                101
##              Chile              China           Colombia
##                  1                 21                  1
##     Czech Republic            Denmark Dominican Republic
##                  3                  7                  1
##             France            Georgia            Germany
##                115                  1                 86
##             Greece          Hong Kong            Hungary
##                  1                 16                  2
```

```
##           Iceland            India        Indonesia
##                 1                7                1
##              Iran          Ireland           Israel
##                 3               10                2
##             Italy            Japan       Kyrgyzstan
##                15               14                1
##             Libya           Mexico      Netherlands
##                 1               14                4
##          New Line      New Zealand           Norway
##                 1               12                5
##     Official site           Panama             Peru
##                 1                1                1
##            Poland          Romania           Russia
##                 2                1                5
##          Slovakia     South Africa      South Korea
##                 1                7               10
##      Soviet Union            Spain           Sweden
##                 1               29                3
##       Switzerland           Taiwan         Thailand
##                 1                1                5
##                UK              USA     West Germany
##               392             3446                3
```

Refactoring the column removed the porblematic cases.

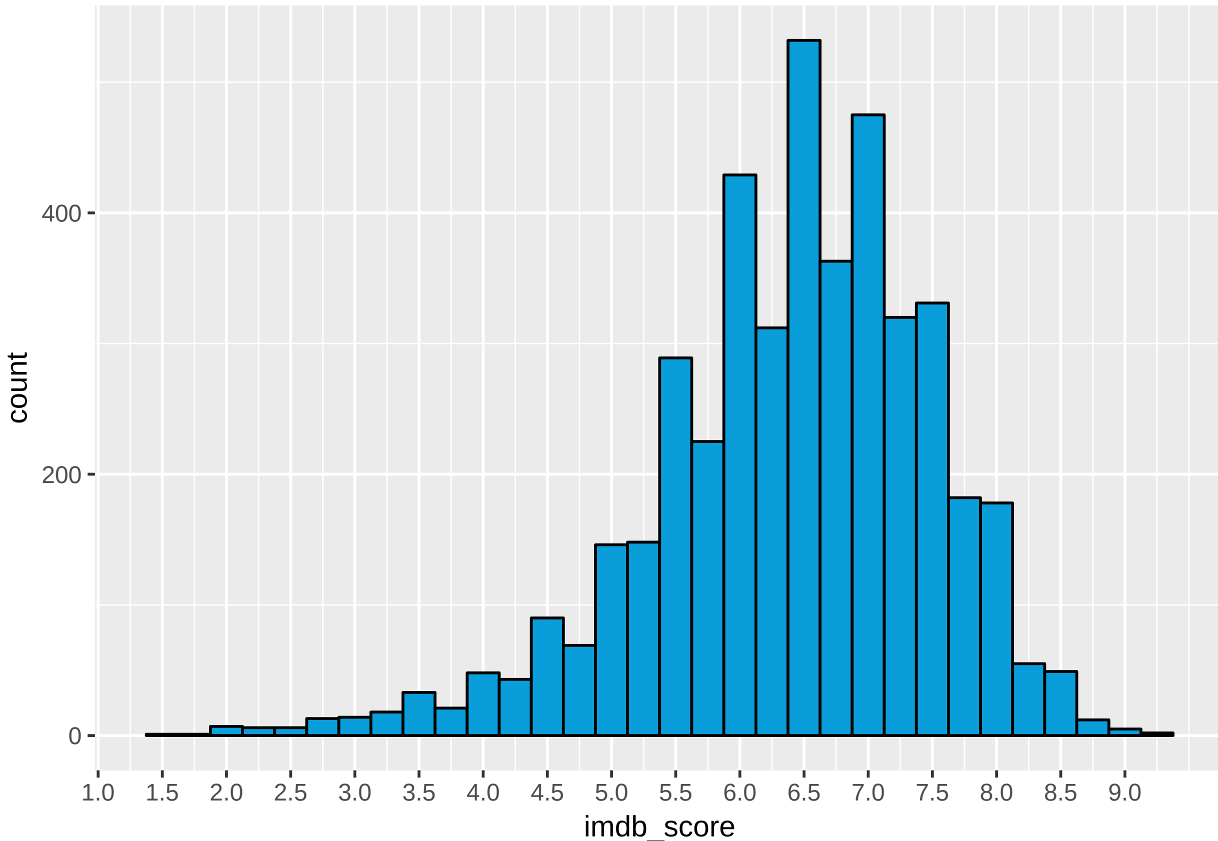
The last concern that should be noted is that most of the columns include NA values. However, we'll deal with the NA values on a plot-by-plot basis. In some cases, the NA values won't show up (in the case of the histogram). In other cases where they might (such as a facet wrap), we'll remove them within the code for the plot.

Now that we've finished data cleaning, let's start off the exploration with a univariate analysis of several of the variables.

Univariate Plots

We'll first look at a histogram of IMDb scores.

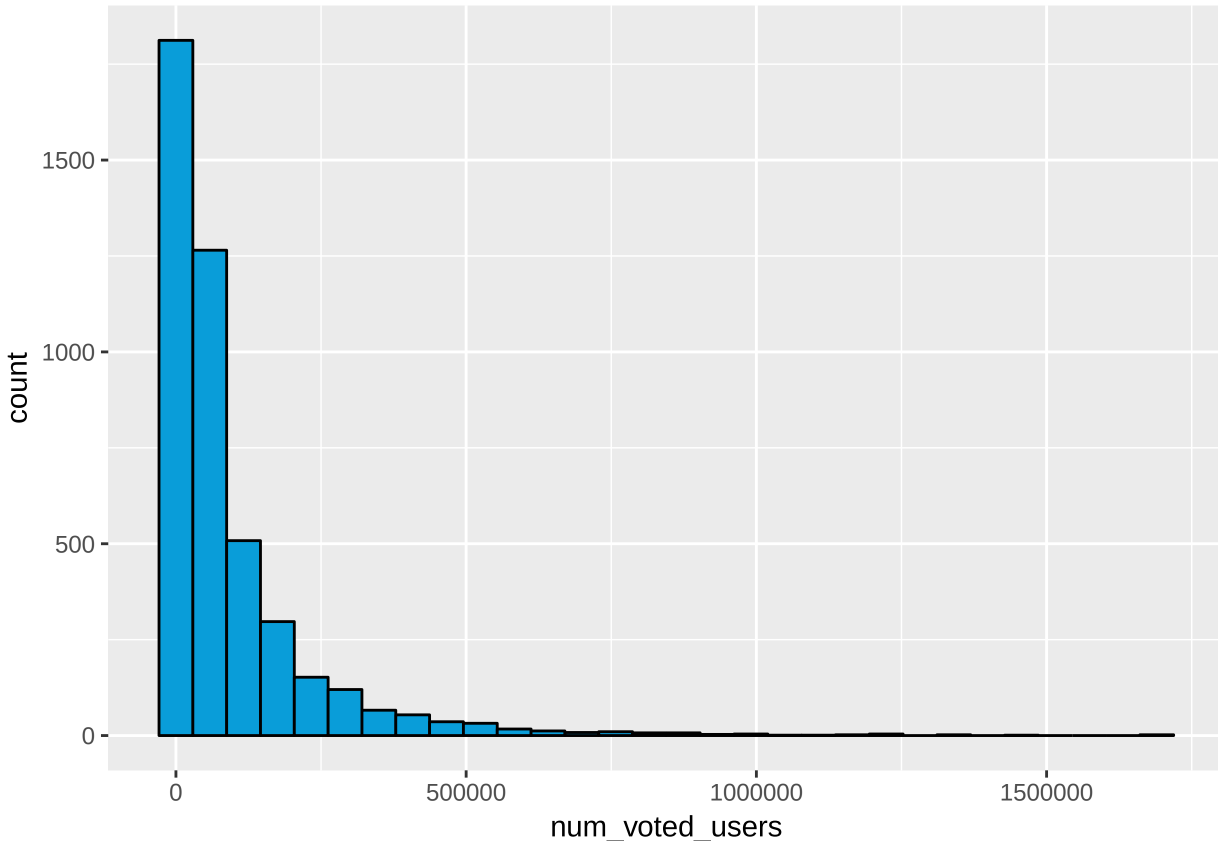
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.600   5.800   6.500   6.402   7.200   9.300
```



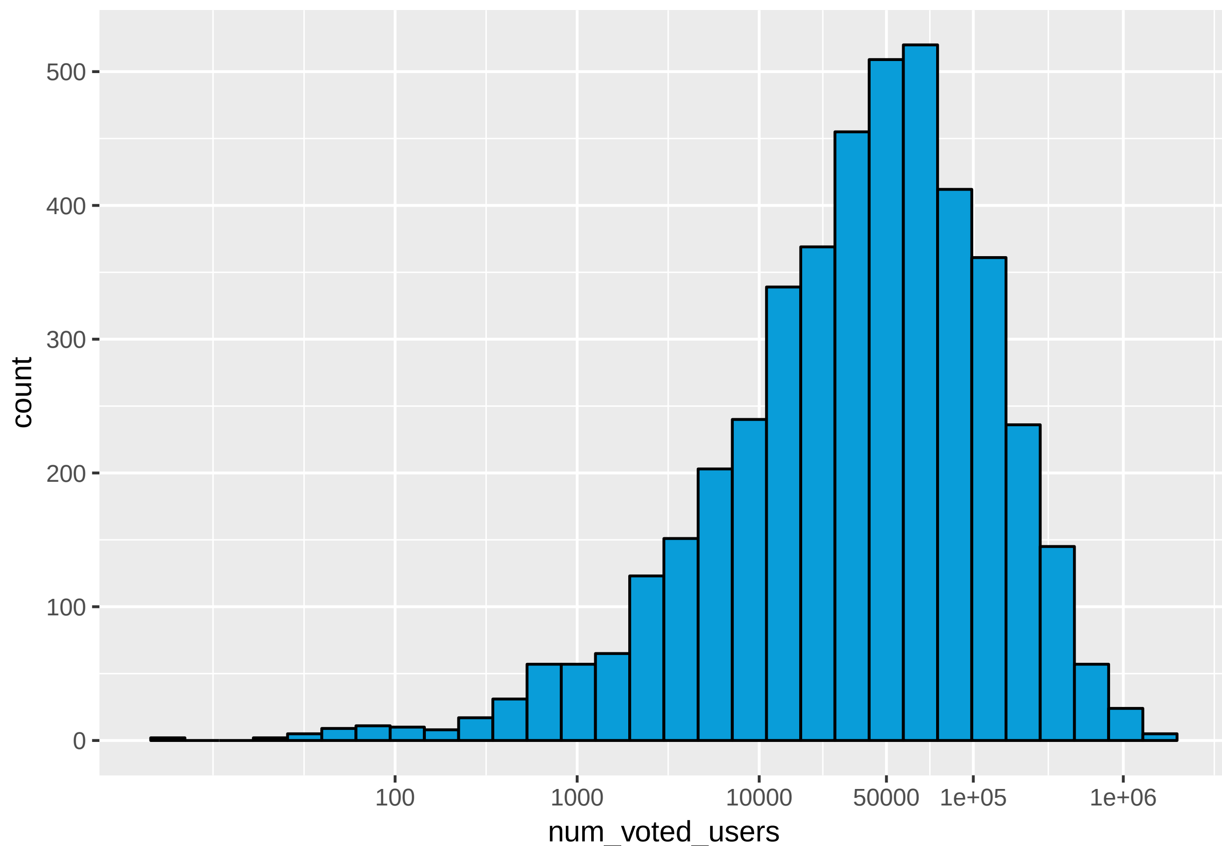
It appears we have a normal distribution that is slightly negatively skewed. The peak is at an IMDb score of 6.5, which is the summary of the data shows to be the median. This is consistent with my experience of visiting the website frequently through out the years. Films in the 7.0 - 8.0 range are the films generally anyone could consider as a great movie, while the 8.0+ tier are what many would call as the all-time greats or classics. Films in that range are usually in the IMDb top 250 or right outside it.

Let's take a look at a histogram of the number of users who voted for each film.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       6   12700   42390   92850  108800 1690000
```



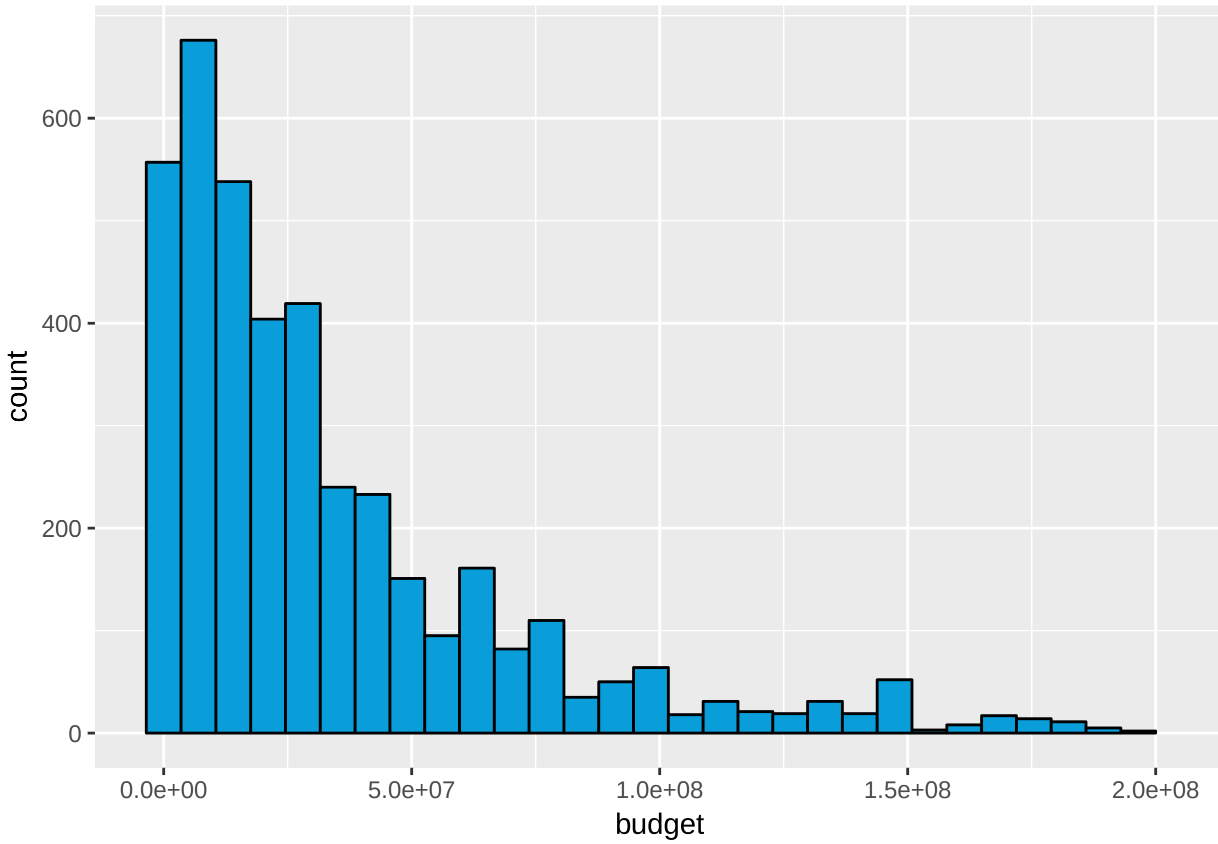
The size of the outliers are resulting in a highly skewed distribution. In order to reduce the skew and better see the distribution, we should apply a log transformation.



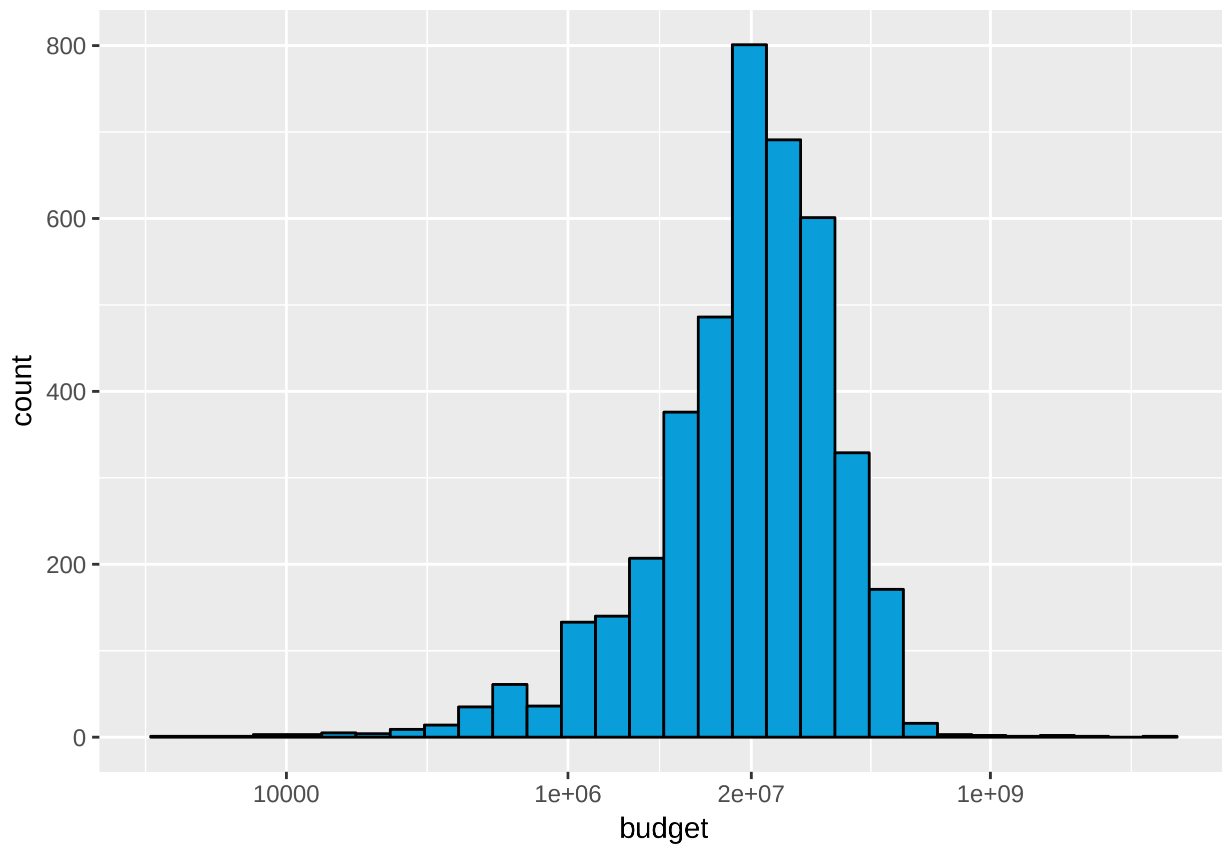
The median number of users who voted - 42,390 - is much easier to see once we apply the transformation.

Let's take a look at film budgets.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 1.100e+03 8.500e+06 2.108e+07 4.298e+07 5.000e+07 1.222e+10       289
```



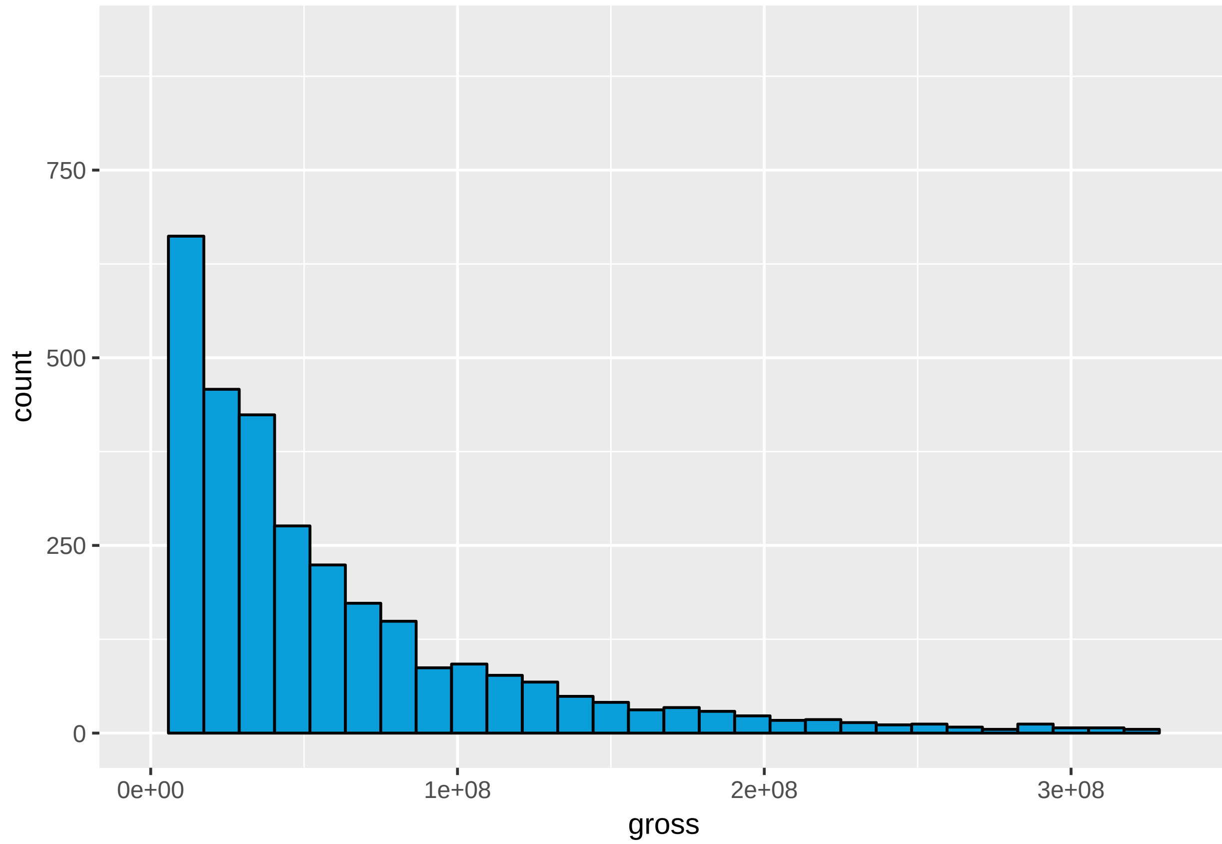
I removed the top .1% of the data due to the size of the outlier (12,220,000,000), but the histogram still warrants a log transformation in order to better visualize the data distribution of film budgets.



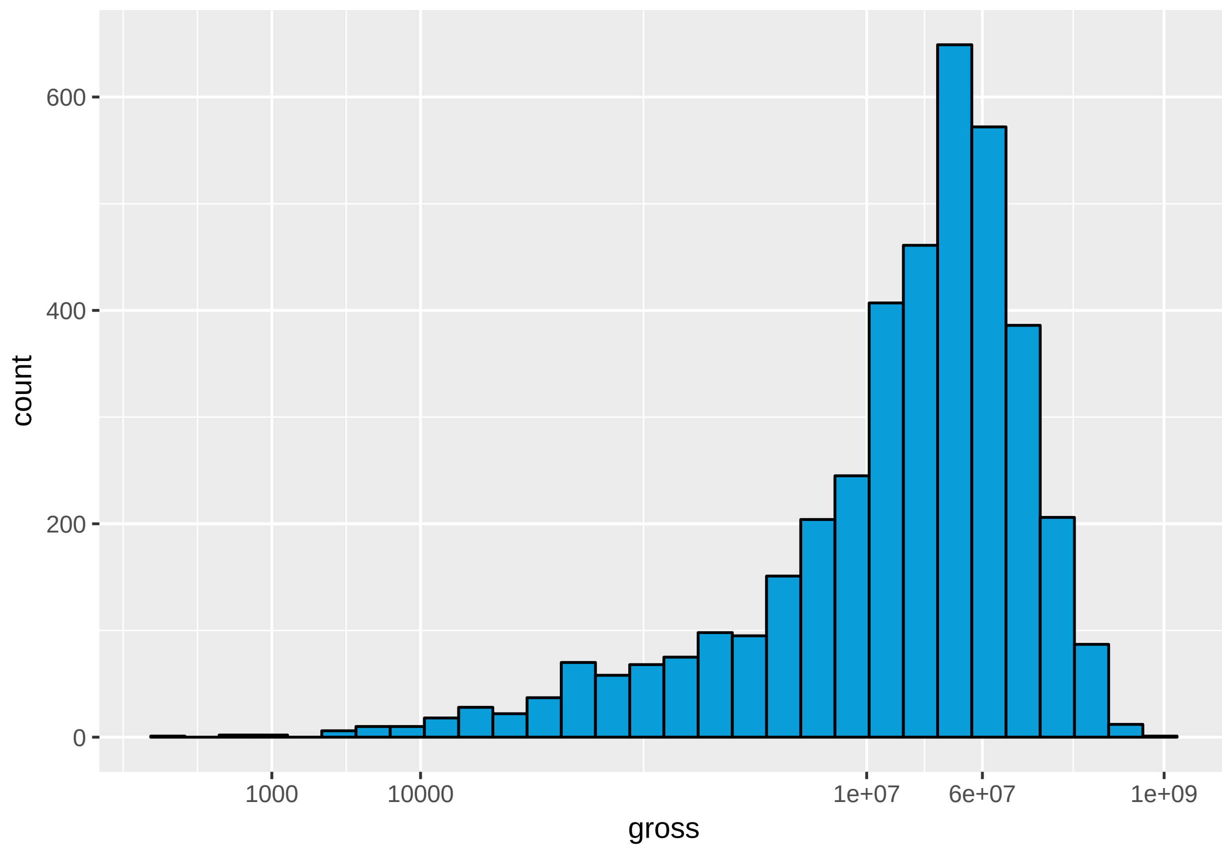
The transformation gives us a normal distribution that is slightly negatively skewed. We're able better see the median budget of 21,080,000 and how it's positioned relative to the rest of the data.

Let's also take a look at the amount of money these films have made in the box office - their gross.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##       162   6712000  27280000  50330000  64600000 760500000       442
```



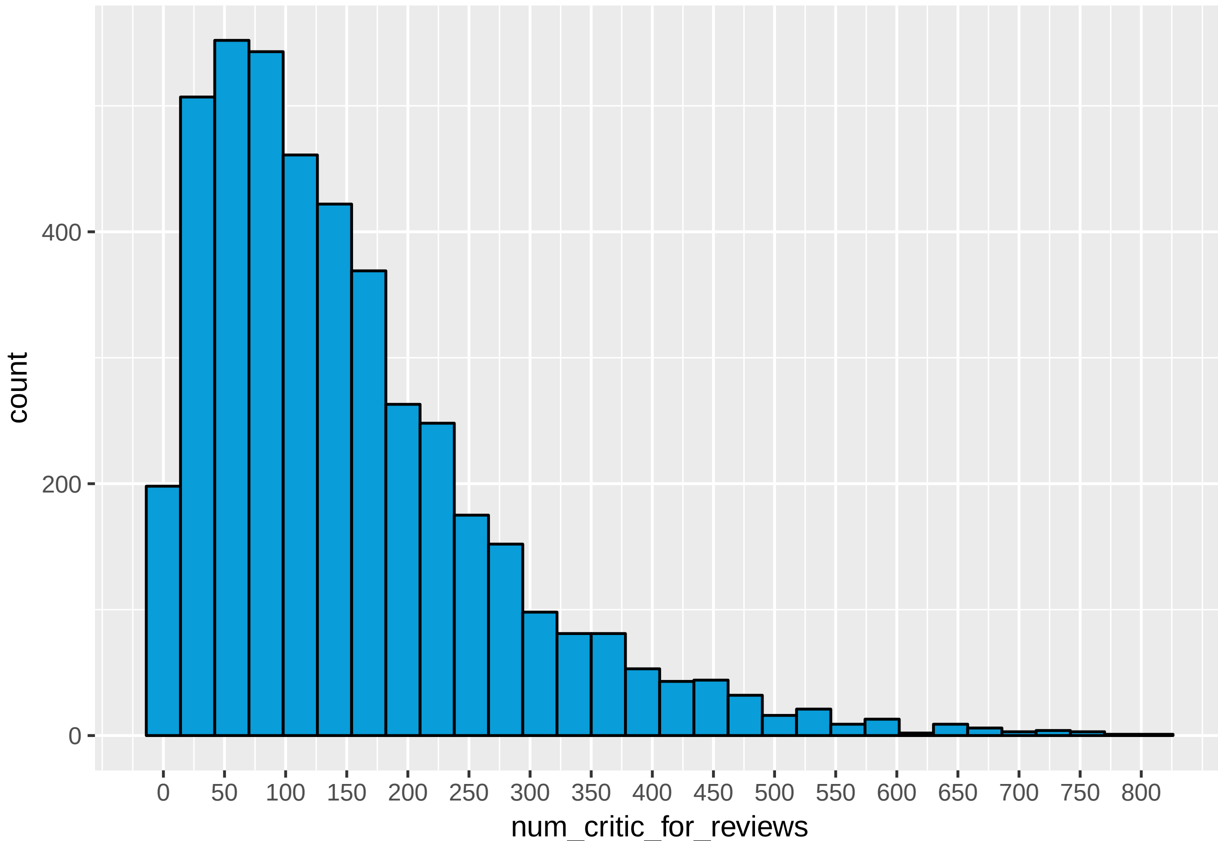
Similarly to the budget distribution, removing the top .1% of the data to decrease the positive skew wasn't enough to effectively visualize the distribution. This also calls for a log transformation.



With the transformation, the median gross of 27,280,000 is much clearer. The distribution is normal but negatively skewed.

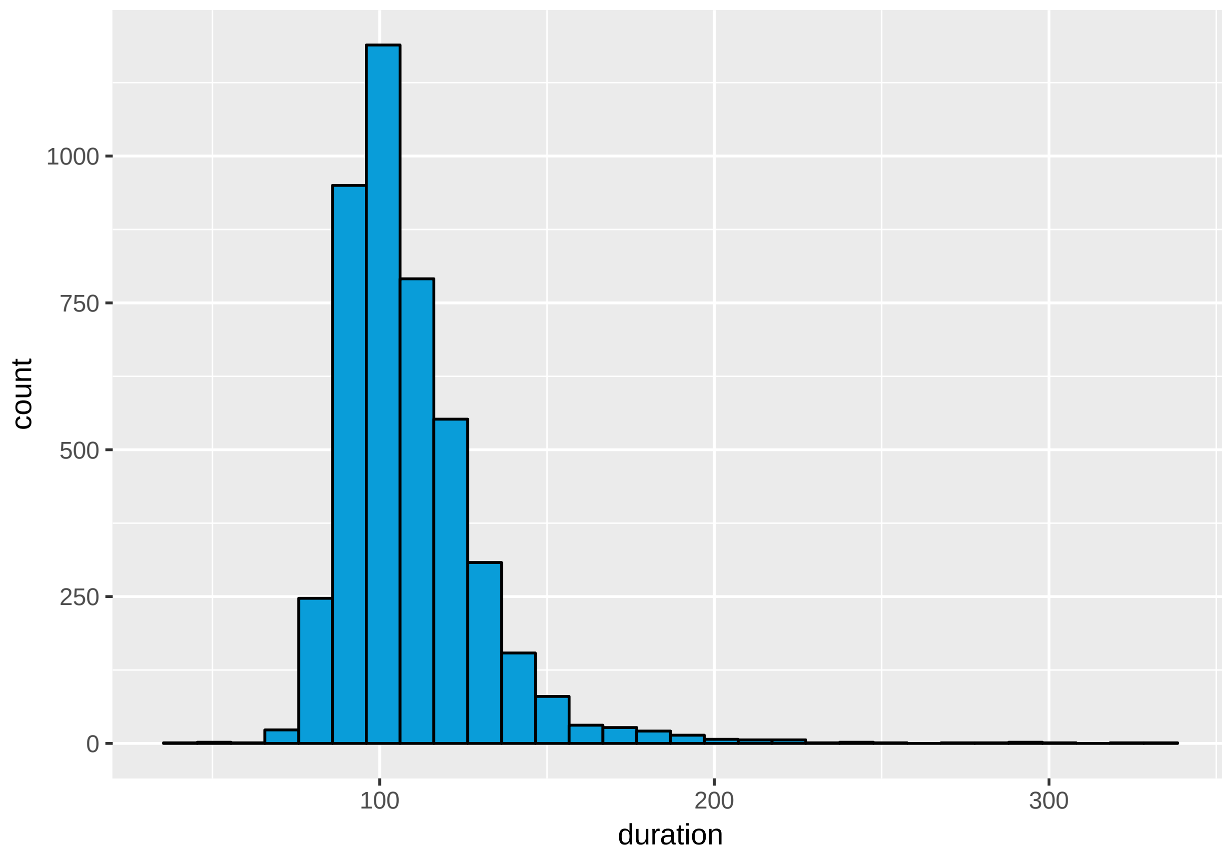
Next, we'll look at the number of reviews from critics each film received.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     1.0    63.0   123.0   152.5   210.0   813.0      13
```



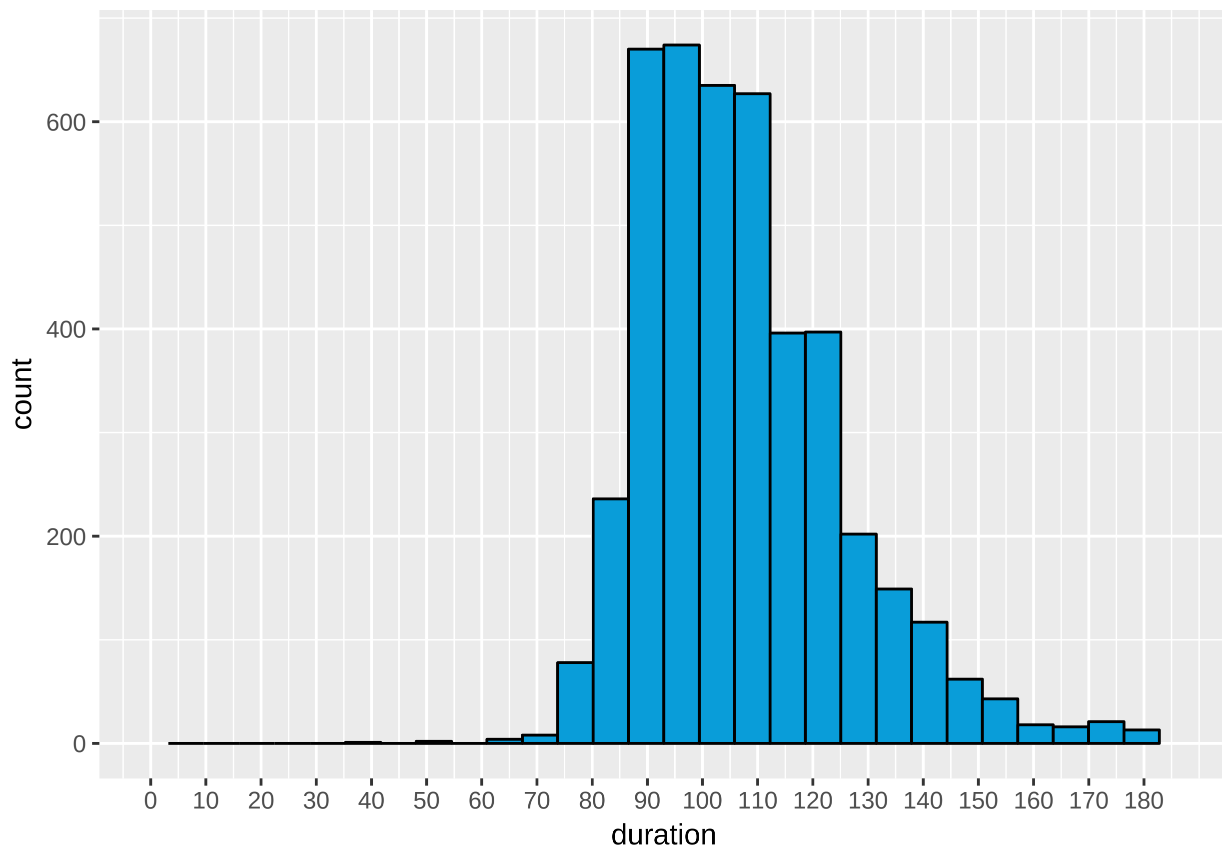
We have a positively skewed distribution of number of reviews from critics. The median is 123 reviews.

Let's now take a look at the duration (the length of film).



Let's remove the top 1% of the data to get a better look at the distribution. A log transformatio won't be needed.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    37.0    94.0   104.0   108.7   118.0   330.0       2
```



We have a normal distribution that gives a clear idea of the distribution of running times. The median duration is 104 minutes.

However, to enhance our analysis later in this project, I want to bucket duration times into a couple groups.

1. 0 - 60 minutes = short
2. 60 - 120 minutes = regular
3. 120 - 180 minutes = long
4. 180 - 330 (max value) minutes = very long

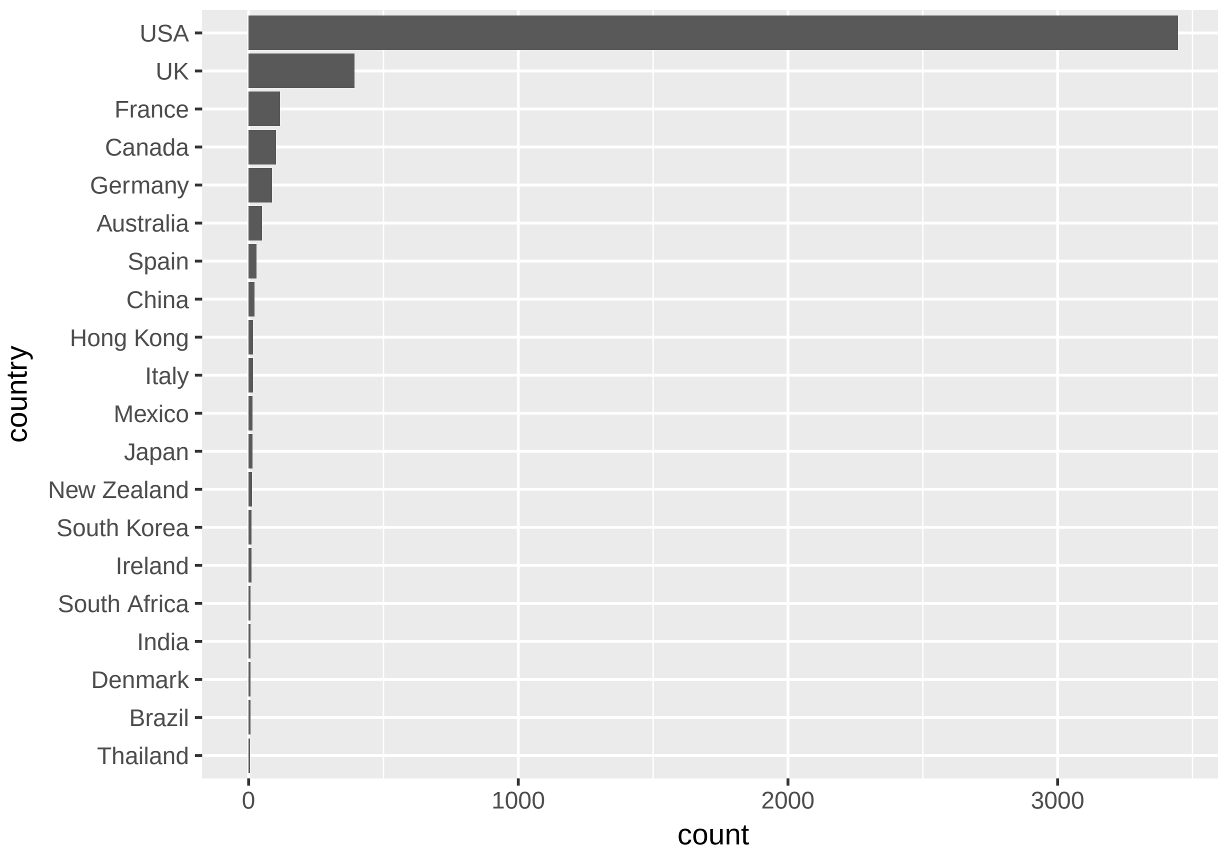
Let's do that and see the count of each.

```
breaks_format <- c(37, 60, 120, 180, 330) #This creates the buckets that we
#want for the categorization

imdb.2$duration.group <- cut(imdb.2$duration, breaks_format,
                             labels=c("short", "regular", "long", "very long"))

summary(imdb.2$duration.group)
```

```
##     short   regular      long very long      NA's
##         2      3451       913        54         3
```

So we see a majority of the films (3451) land in the 60-120 minute "regular" category", but still a sizable chunk (913) land in the 120-180 minute"long"" category.

What countries are these films coming from?

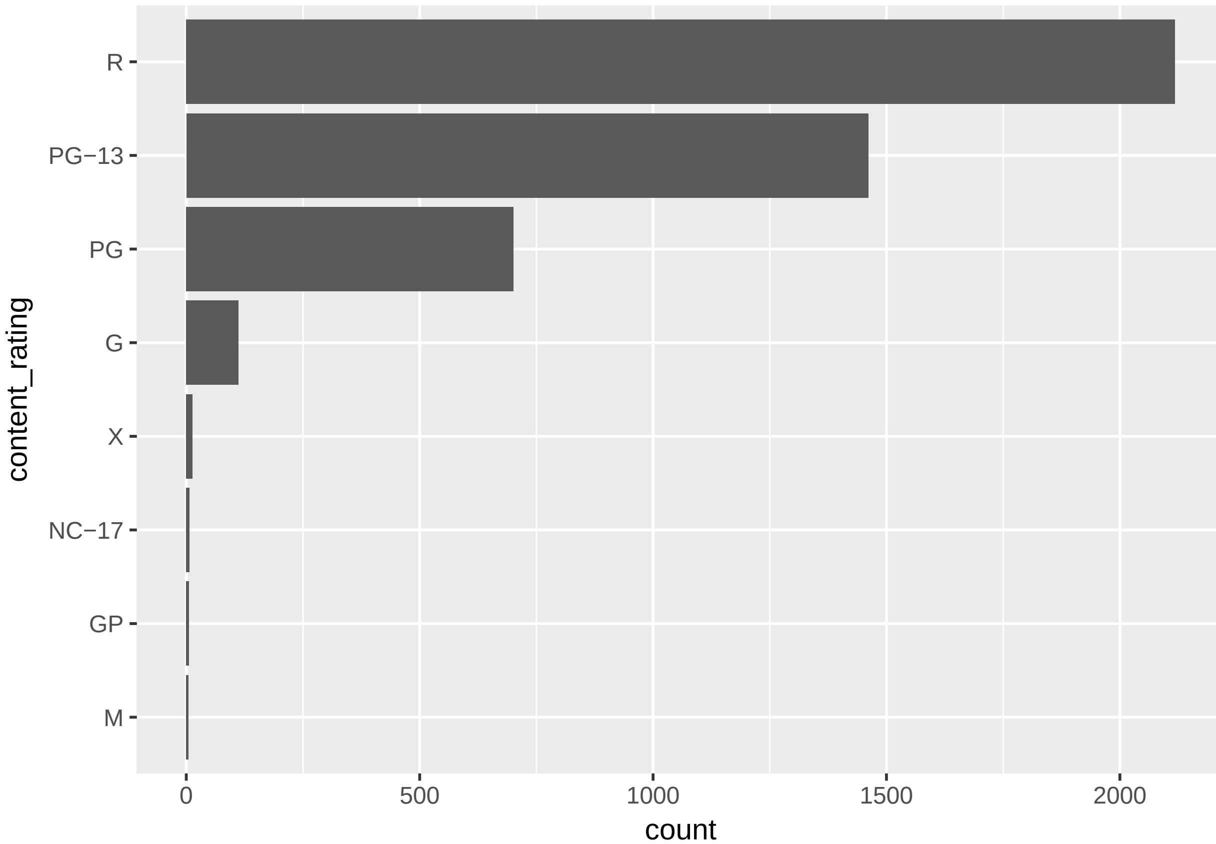


Considering the number of countries (50+), I decided to show a snapshot of the top 20 countries instead.

Unsurprisingly - considering the dominance of Hollywood in international cinema - an overwhelming amount of the films come from the United States, with 3446 films. The second most is the United Kingdom (UK) with 392. Thailand sits at the bottom of this list with 5. If the graph opened up further, we'd see 30+ countries with 5 films or less.

With the content rating cleaned up from earlier, let's take a look at the count of films for each rating.

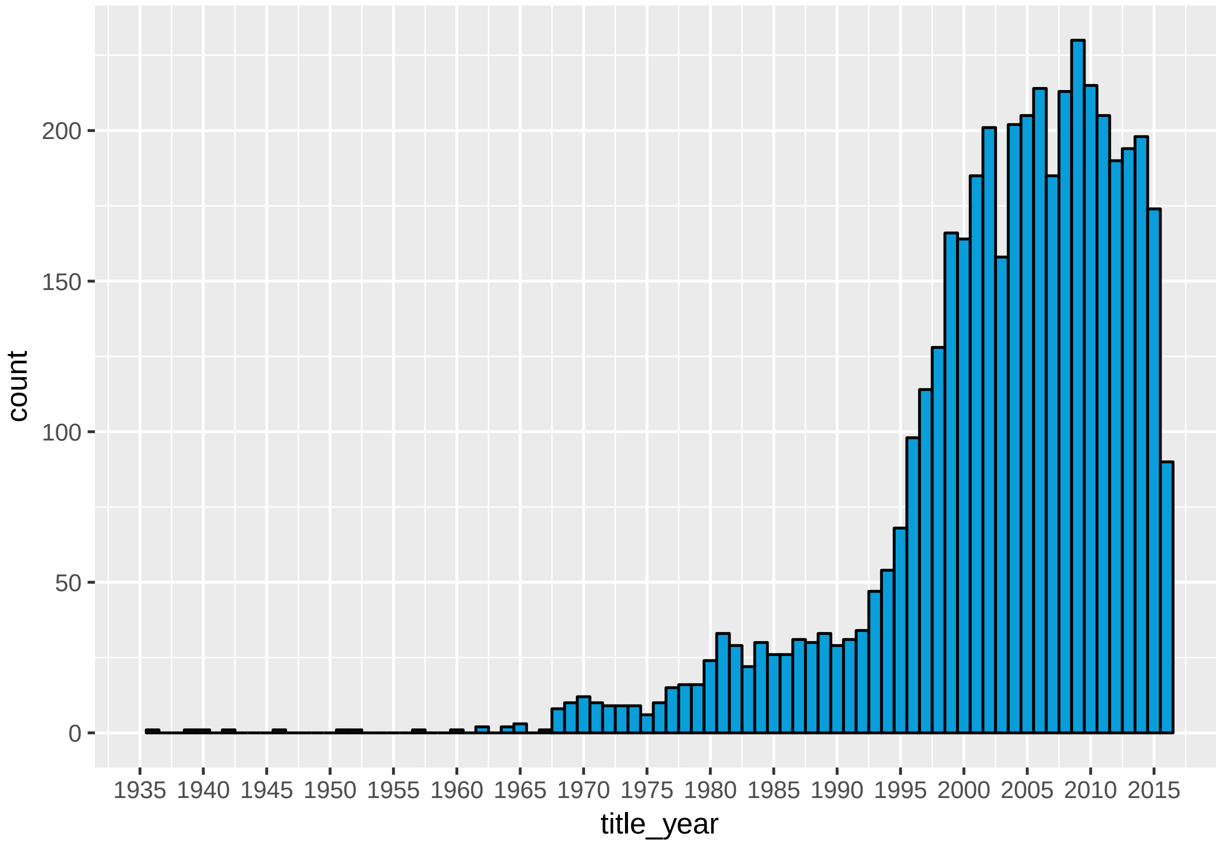
```
##     G    GP     M NC-17    PG PG-13     R     X
##   112     6     5     7   701  1461  2118    13
```



I was a bit surprised with this. I would have expected there to be more PG-13/PG films, since by the very nature of their rating, it means the films are more accessible to a wider audience. Only movie goers who are 17 and older - unless they're accompanied by an adult 21 or older - are allowed to watch rated R films. Yet there are much more R films than any other rating. Why this is so would be worth investigating further in a future follow-up to the project.

Finally, this dataset contains films across almost a century.Let's create a histogram that gives a sense of the number of films released each year.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1936    1999    2005    2003    2010    2016
```



A negatively skewed distribution, this isn't surprising to see. Naturally we would see more films produced over time, especially as film grew in popularity as a source of entertainment, business, and a field of study. We have a median year of 2005.

Univariate Analysis

What is the structure of your dataset?

The original dataset has 5043 movies in this dataset with 28 variables. However, initial investigation into the dataset's structure revealed that a portion of the database contained data on non-theatrical films, such as made-for-TV films, and the occasional TV series. Removing them from the dataset left us with 4423 films with 28 variables. Most of the dataset is continuous, but some of it is categorical as well.

What is/are the main features of interest in your dataset?

The main feature of interest in this dataset is the IMDb score. I'm interested to see which features of a film are correlated with IMDb scores. Can we take a look at certain figures and attributes of a film and get an idea of how a film is performing on IMDb?

What other features in the dataset do you think will help support your investigation into feature(s) of interest?

The features I believe will might be linked to a film's IMDb score are a film's gross, year of release, number of users voted, and number of reviews from critics.

Did you create any new variables from existing variables in the dataset?

I did not create any new variables from existing variables in this dataset.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the datat to tidy, adjust, or change the form of the data? If so, why did you do this?

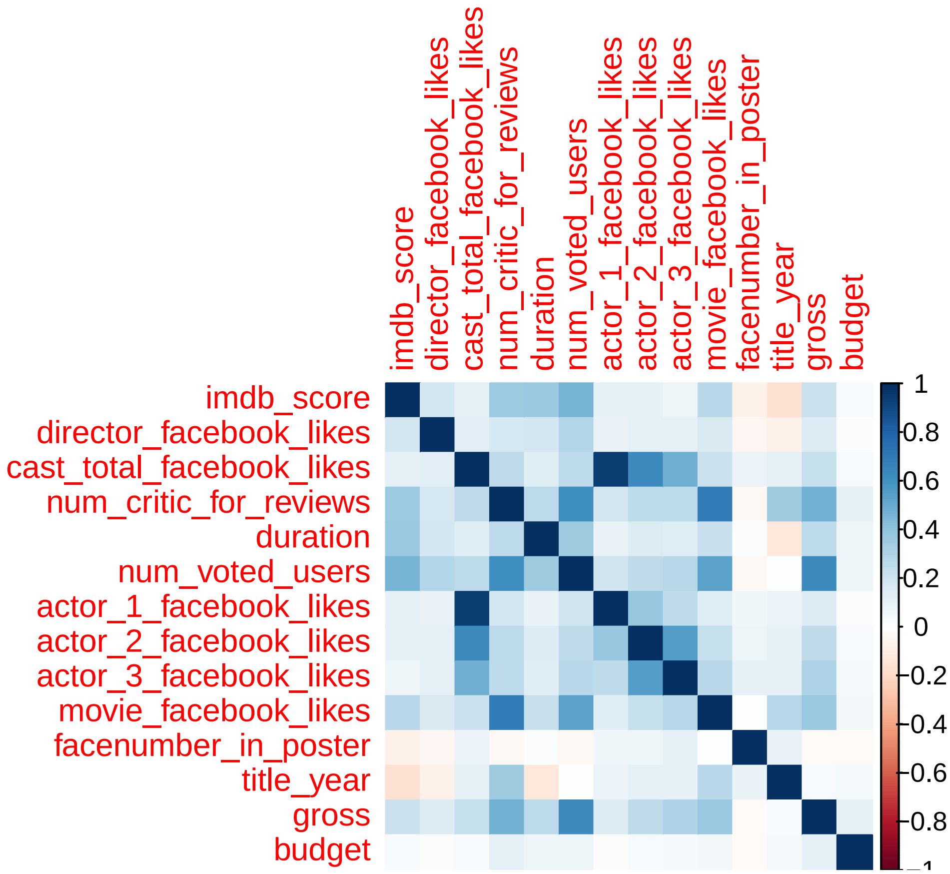
A distribution that took me by surprise was the distribution of content ratings - I would not have expected rated R films to take the greatest share.

I log-transformed the histograms for number of voted users, budget, and gross. As each of these charts were skewed, it was difficult to accurately see how most of the data was generally distributed outside of the outliers. Log transforming them gave a much clearer picture of what was happening.

Additionally, I created new data frames for the count plots of content rating and country - a column for each value and a second column for the count of the value. This allowed me to then to easily chart the number of instances of each value, sorted by largest to smallest.

Bivariate Plots

Let's start off this section with a corr plot. We'll look at the correlation between the 14 continuous variables of the dataset.



From this visualization, here are some observations we can make:

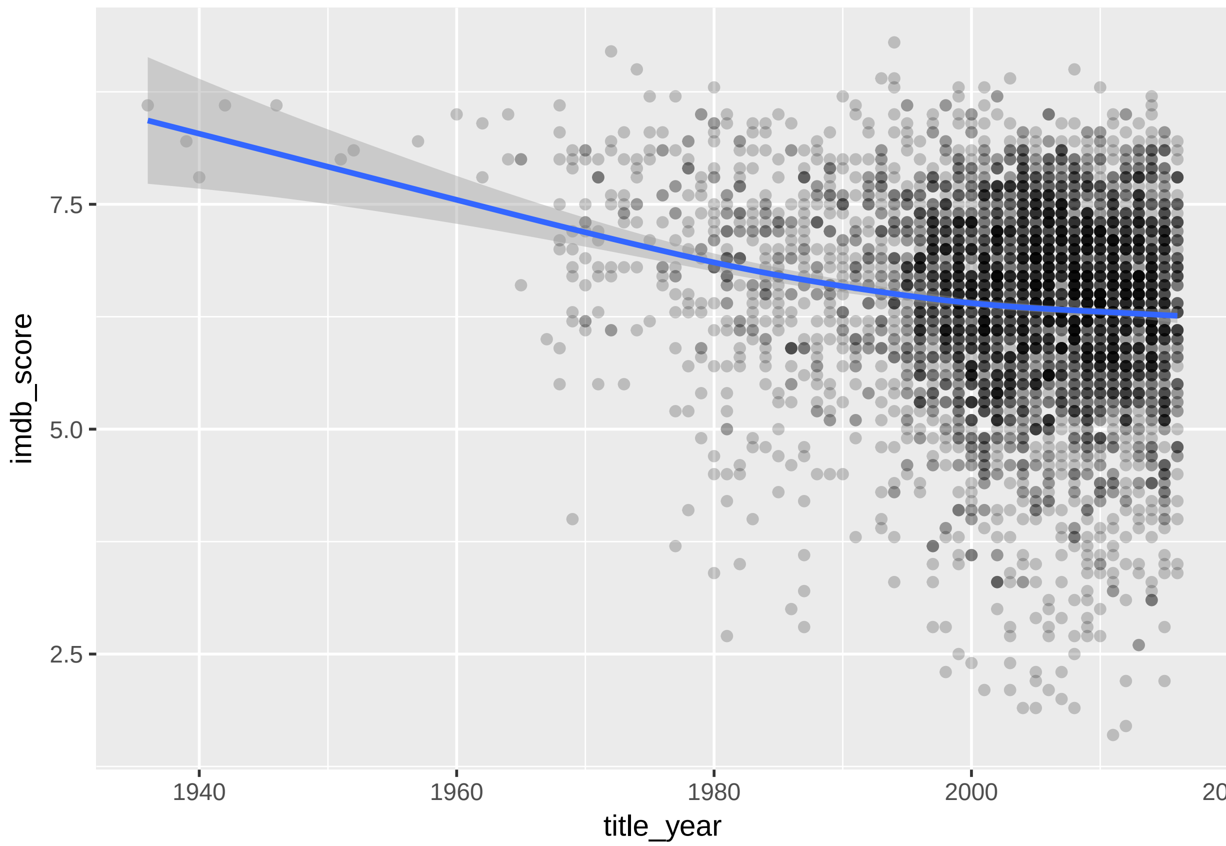
1. The variables most correlated with IMDb scores are number of critic reviews (0.365), the duration of a film (0.37), and the number of voted users (0.462).
2. The year a film was released was a weak negative correlation of -0.159. I would have expected a stronger negative correlation - as the number of films increase with the passing of time, so does the number of average/bad films. But by extension, so does the number of good films released each year, so that likely is plays a role in why the effect isn't too large.
3. There is very little correlation between budget and IMDb score (0.036). Considering the number of expensive blockbuster duds that come out of year, it's no surprise to see that you can't buy your way into quality.
4. On the flip-side, a film's performance in the box office appears to have a much higher correlation than budget (0.217). There's sense here - better the film, more people will be keen on seeing it.

Other observations not related to IMDb score that caught my interest:

1. The strongest correlation is between the number of Facebook likes for the cast's most popular actor (actor_1_likes) - popularity determined by number of Facebook likes of each cast member - and the total of Facebook likes of the cast (0.947). The number of Facebook likes for the second and third most popular actors in the cast is also fairly strongly correlated with the Facebook likes of the total cast (0.638 and 0.486 respectively).
2. The number of voted users is fairly strongly correlated with both number of critics reviews (0.613) and gross (0.633), which can be expected. The more a film rakes in the box office, the more popular it is, and thus the more votes. And if a lot of users are voting for it, then chances are so are critics.
3. Number of voters also has a fairly strong correlation with the number of Facebook likes a film has, with 0.531.
4. There's a positive but weak correlation between duration and number of voted users and duration (0.35). It's possible this is due to the link that they're both positive and weakly correlated with IMDb

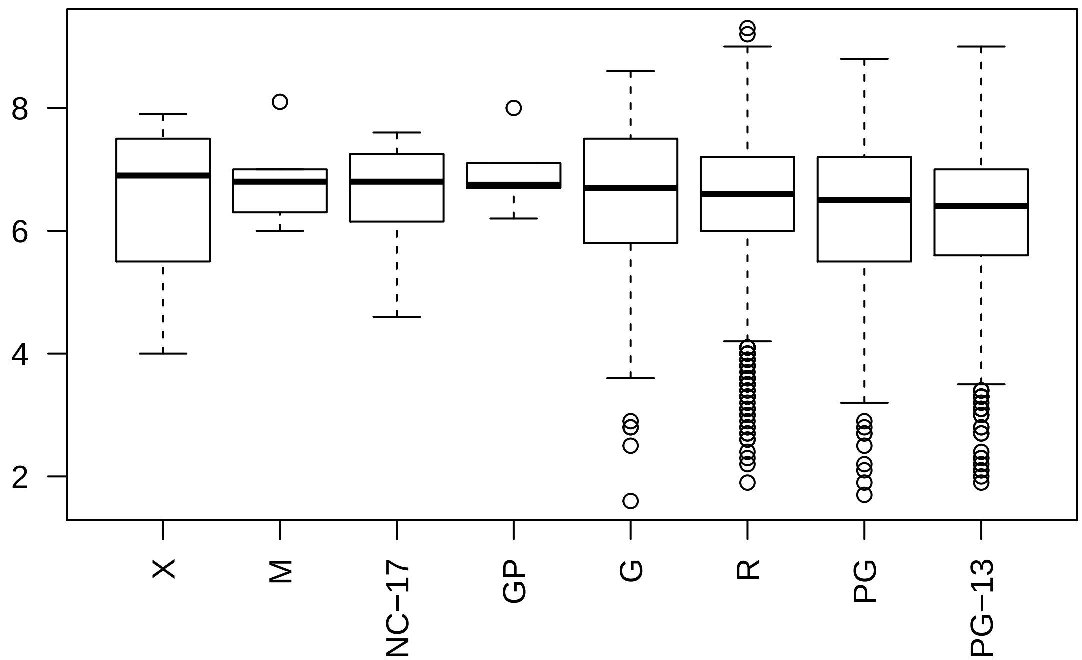
score. The longer the film, the better the score, and thus the more people who have voted for it.

Let's dive into some of these relationships we just observed with bivariate plots.



The negative correlation between imdb_score and title_year that we saw in the corr plot is shown here. Again, this could be due to the fact there are more movies being produced now then there were 20+ years ago. Thus, we're seeing greater instances of bad to average films, and as a result a downtrend in IMDb scores, which is reflected by the line.

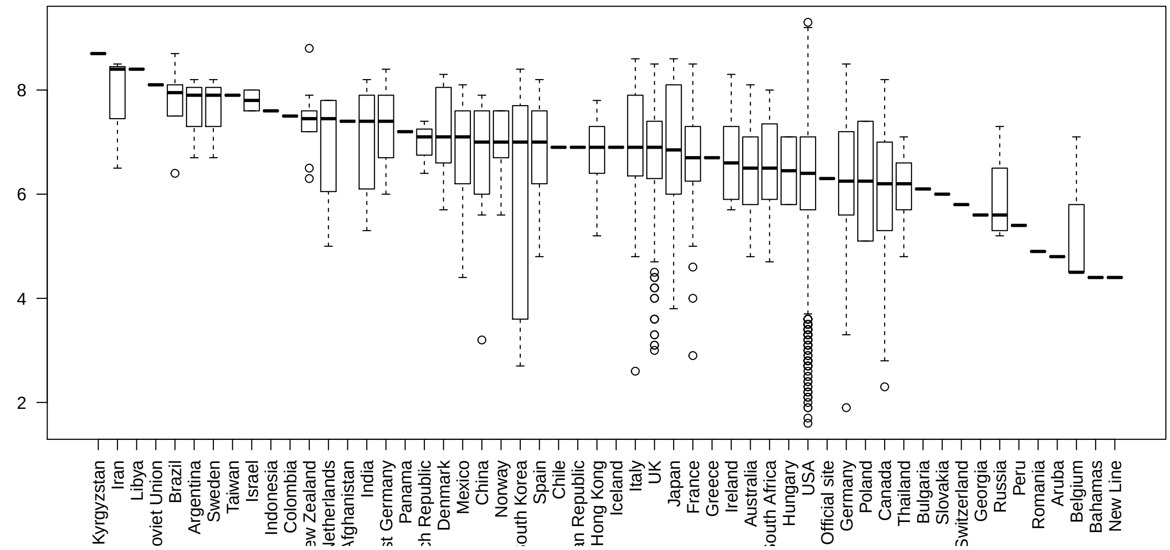
Let's visualize the relationship between content ratings and IMDb scores and see if there's a pattern.



The box plot is ordered from the largest median value to the smallest. Two observations:

1. It appears the outdated rating system - X, M, NC-17, GP - have higher median scores than the current rating system. Likely due to the fact there are less films in the IMDb system with the rating system and thus less bad films.
2. Among the crop of films with the current rating system (G, PG, PG-13, R), G has the highest median, while PG-13 has the lowest. But due to the substantial disparity in the number of data points (G only has 112, while R has 2118), the difference shouldn't be overstated.
3. Across the board, there isn't a sizable difference in median IMDb scores to be found between content ratings.

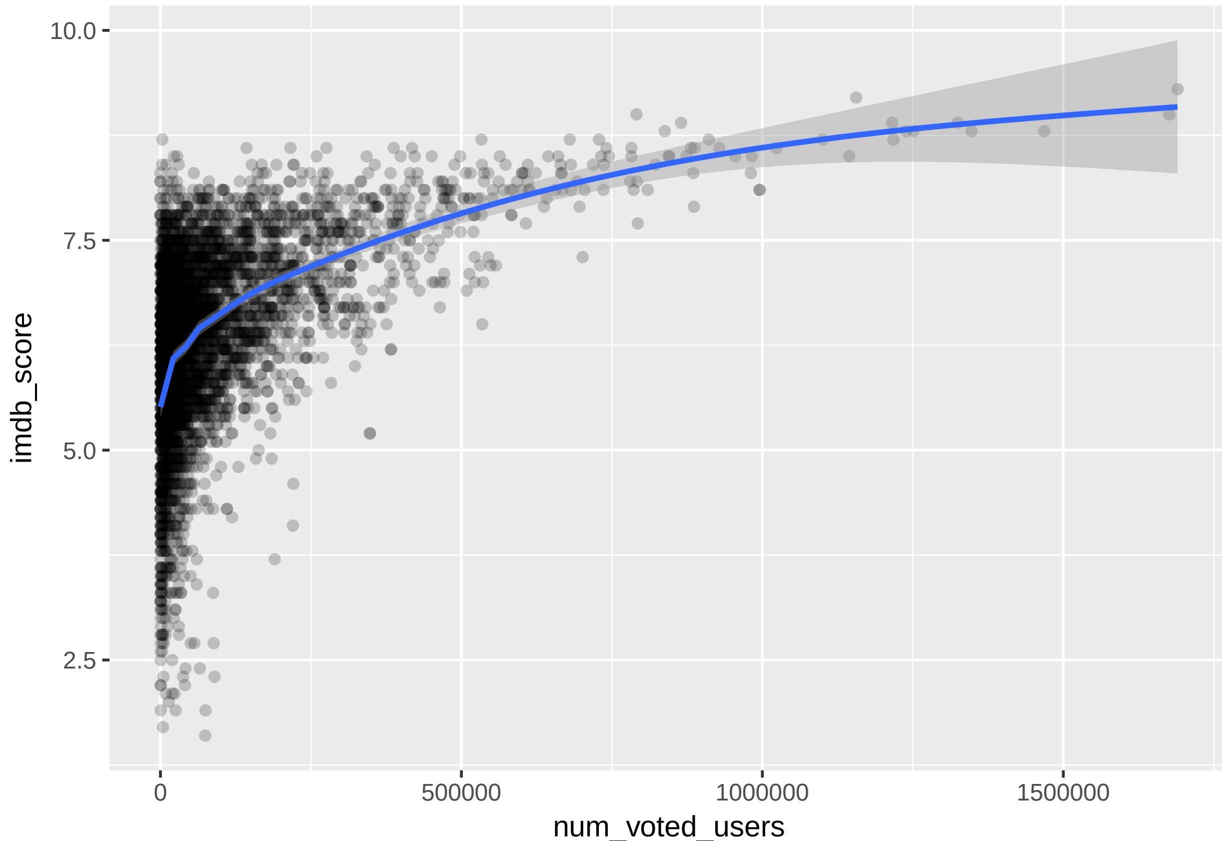
We'll output another box plot to see the relationship between country and IMDb score. Do countries generally tend to perform the same, or are there some countries is noticeably lower or higher IMDb scores relative to the rest of the world?



While it's difficult to see if there's a pattern, we can definitely see that some countries perform better or worse than others.

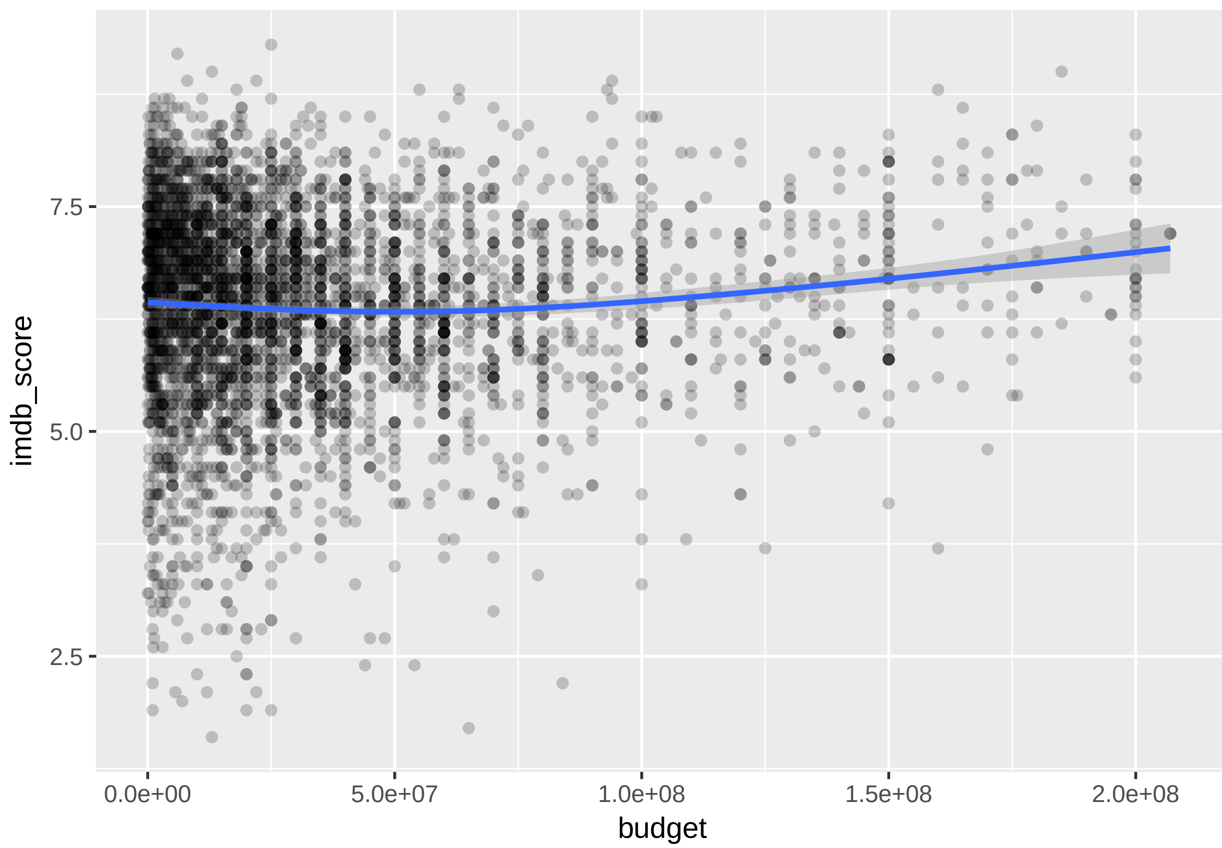
1. While the US and the UK produce the largest number of films, they do not have the highest median. The UK appears in the middle of the chart, with a median IMDB score of 6.9. The US is the lower third of the chart with a median score of 6.40.
2. The highest median belongs to Kyrgyzstan (8.70), with Iran (8.40) and Libya (8.40) closely behind it. It appears they the countries have a small but high quality number of films.
3. The bottom four countries, and the only countries with less than a 5.0 IMDb median score, are Romania (4.90), Aruba (4.80), Belgium (4.50), and the Bahamas (4.40).

Let's move on towards scatter plots for our quantitative data. We'll look at how imdb_score compares with the number of users who have voted towards a film's rating.



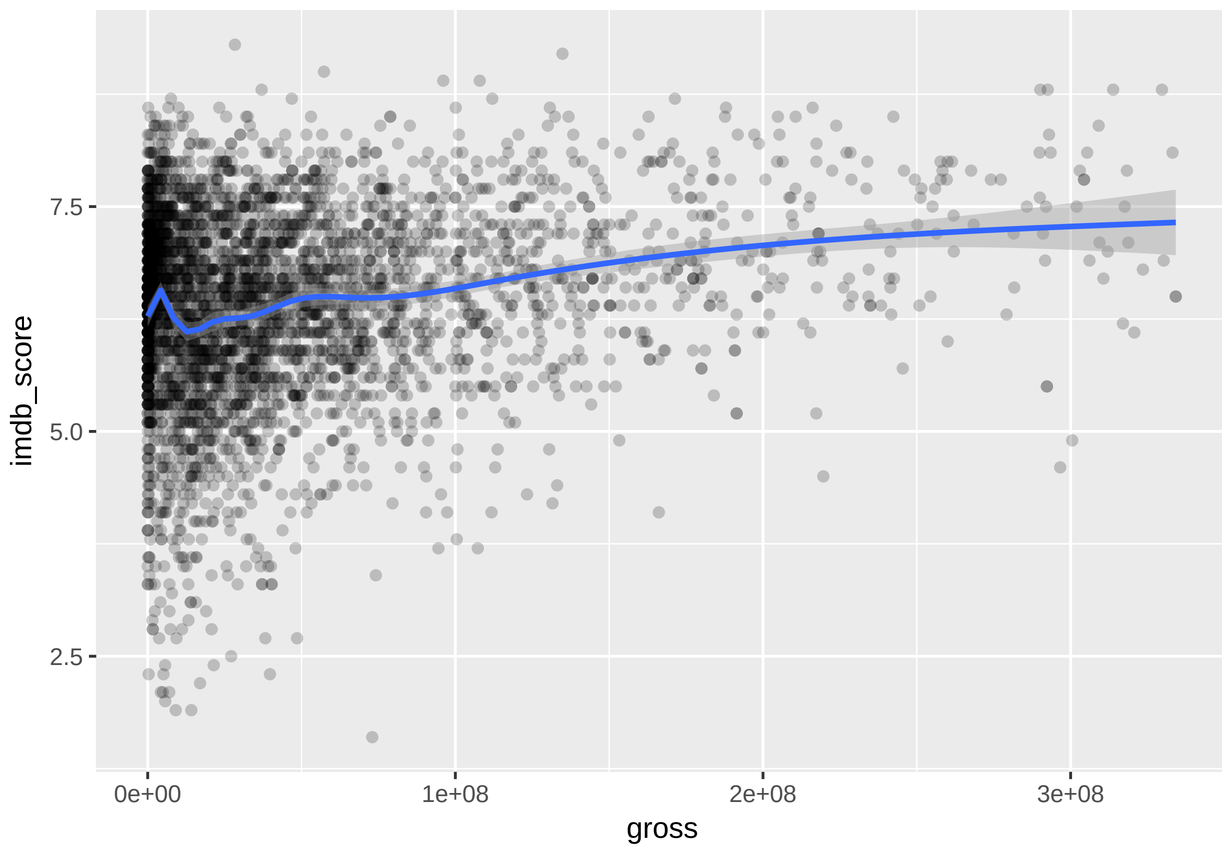
Out of all the variables in the data, num_voted_user had the highest correlation with imdb_score with a correlation of 0.462. You can clearly see that positive correlation in the plot. While data points for high scoring films appear across all counts of voted users, they are almost exclusively the only data points shown when the count of users increases. Still, it is interesting you're not seeing more instances of films with high number of voted users but low IMDb scores. It seems that users are more likely to come out in full force to vote for a film the better it is.

We'll see how imdb_score correlate with budget.



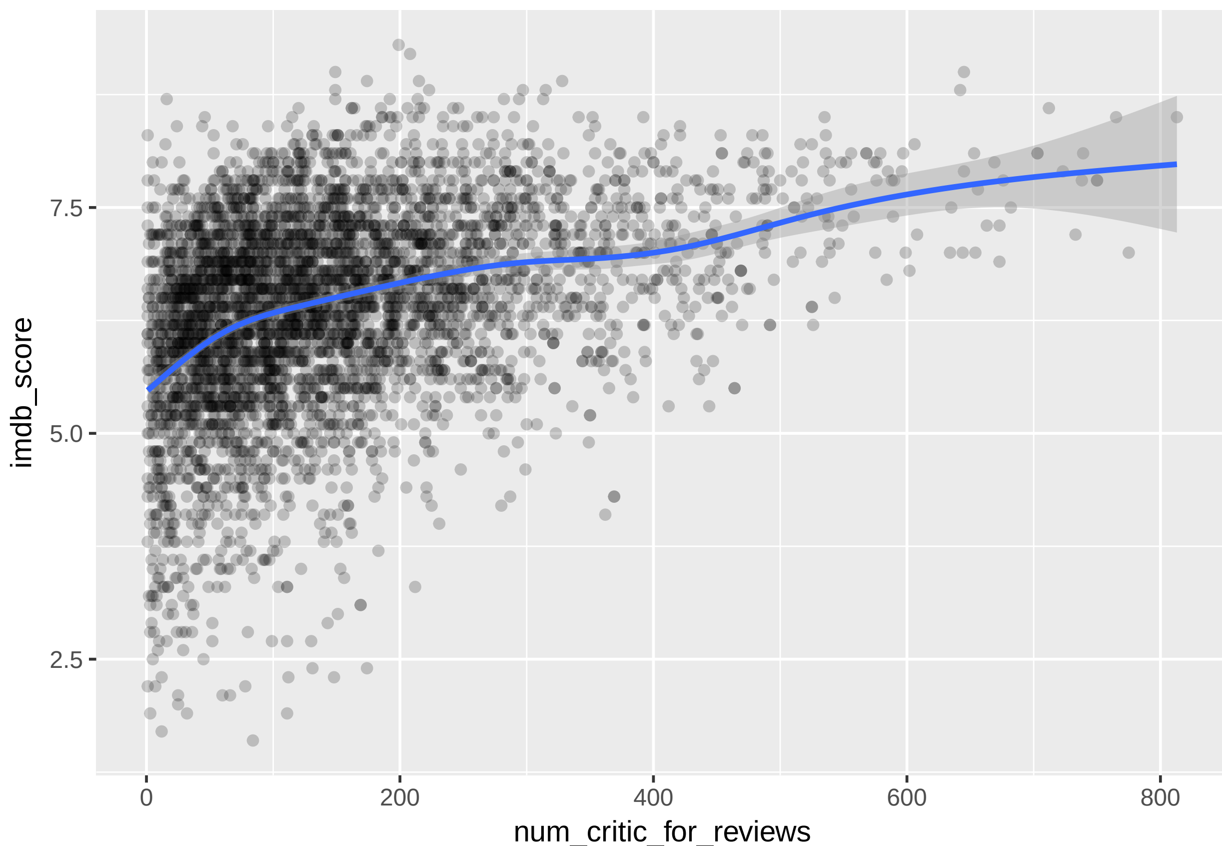
While there are less instances of films scoring near the 2.5 mark the higher you go up in budget, there does not appear to be a substantial correlation between budget and imdb_score. High scoring and low scoring films come in all types of budgets.

Let's move on to gross and imdb_score.



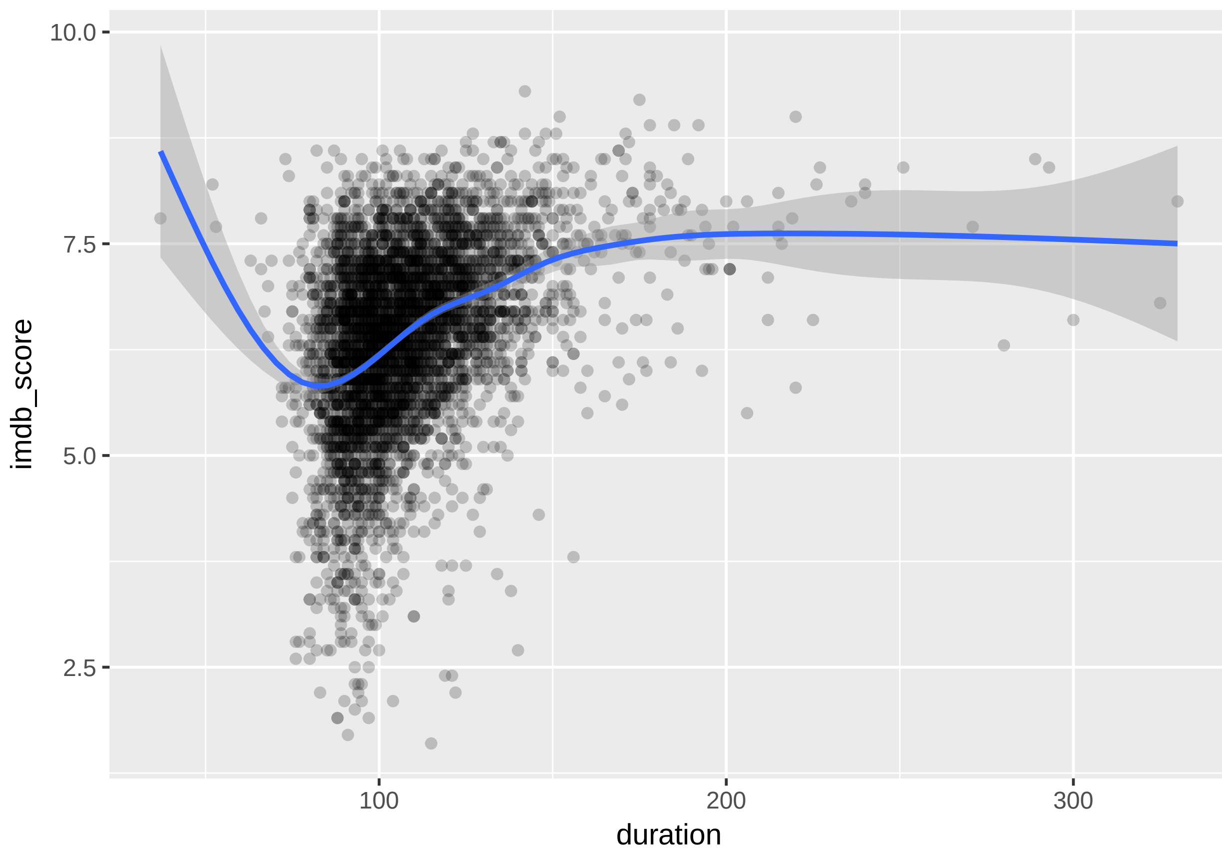
It appears that gross on the other hand indeed has a stronger (albeit still weak) correlation with imdb_score than budget. This makes sense, since a film's performance in a box office is an outcome of how much movie-goers are enjoying it.

Next, we'll see how the number of reviews received by critics relates with imdb_score.



Similarly to imdb_score's relationship with num_voted_users, the score goes up as number of reviews from critics goes up as well. A weak but positive correlation (0.365).

Finally, let's visualize the relationship between duration times and imdb scores.



Duration is also positively correlated (0.37) with IMDb scores. However, there a small group of data points before the 90 minute duration mark of films with small running times but imdb_scores.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in this dataset?

IMDb score had a clear, positively correlated relationship with number of critic reviews (0.365), the duration of a film (0.37), and the number of voted users (0.462). While it's not surprising to see that more reviews and more votes can have a noticeable correlation with a film's score, it's surprising that duration was as well. This would be interesting to investigate in a future update to the project.

The most noticeable negatively correlated (albeit weak) relationship it had with a feature was year of a film's release. IMDb scores dropped over time.

The most interesting plot for me was how median IMDb scores shifted from country to country. My gut instinct before the data was plotted was that the US would be in the higher end of scores, but it actually placed in the lower third of the chart. However, most of these films performing much better than US have significantly smaller sample sizes. It would be interesting to see how this chart would look if we ensured equal number of films from each country.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

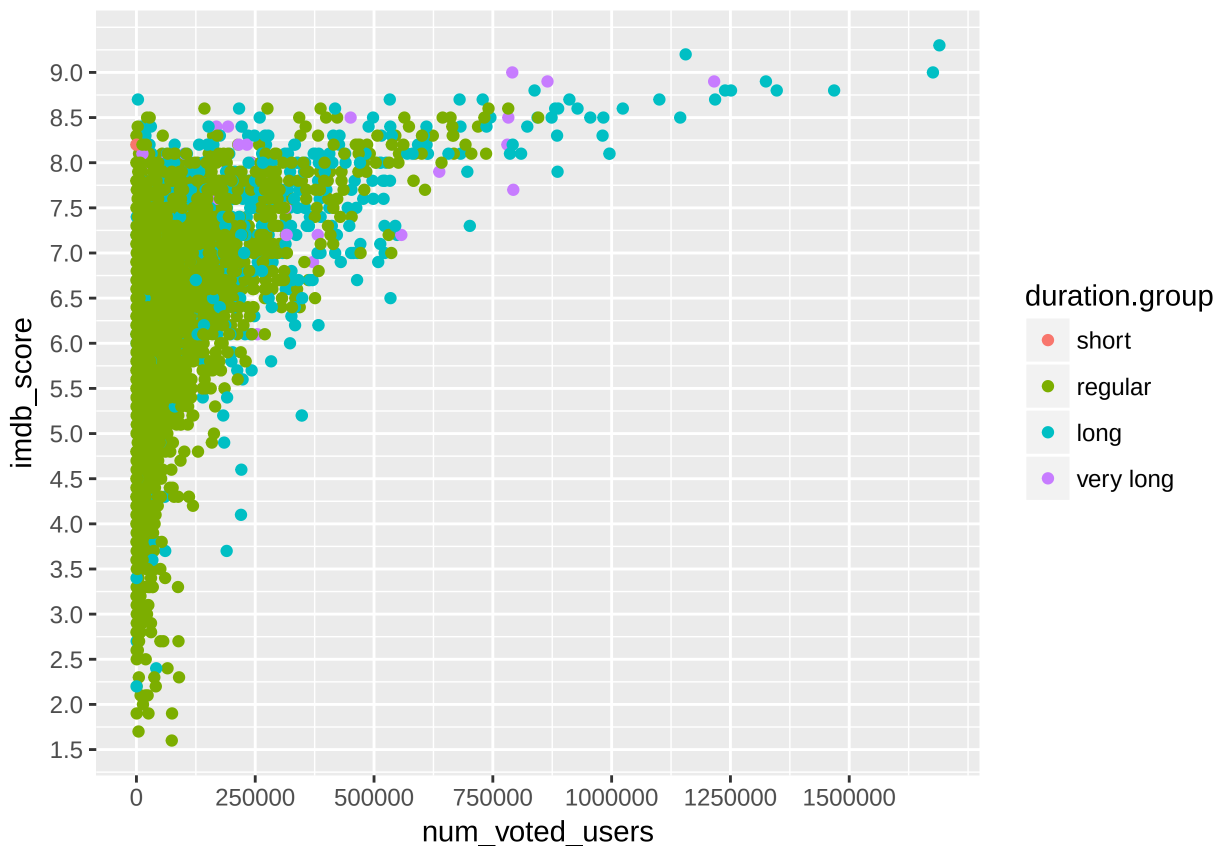
Number of voted users appears to share a weak to fairly strong positive correlation with the most number of variables (6) - IMDb score, number of reviews from critics, gross, duration, and the number of Facebook likes a film has.

What was the strongest relationship you found?

While I didn't investigate this with a scatter plot, the corr plot analysis at the start revealed that the strongest correlation in the dataset was between the number of Facebook likes for the film's most popular actor and the total Facebook likes of the cast.

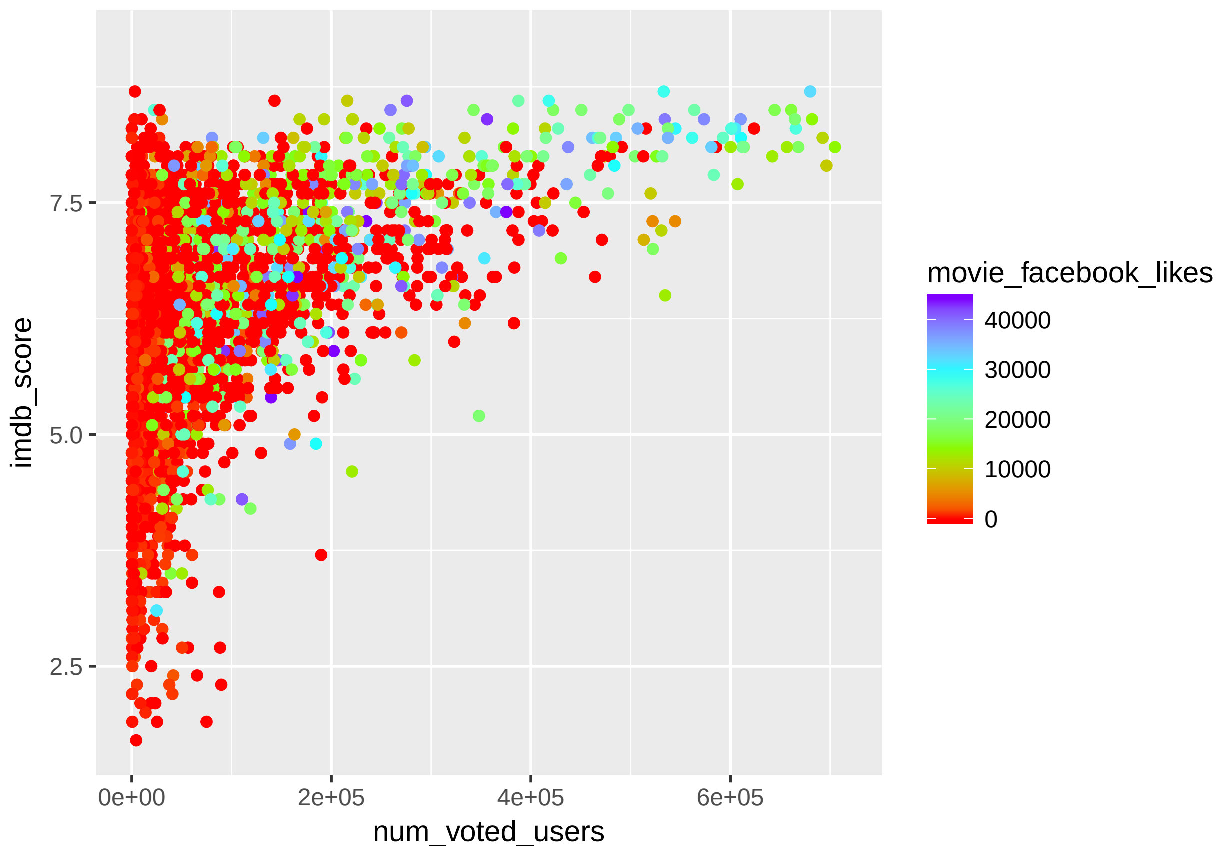
Multivariate Plots

We know that duration and number of voted users are positively correlated (0.35), so let's see how adding in duration as color to the num_voted_use/imdb_score plot looks.



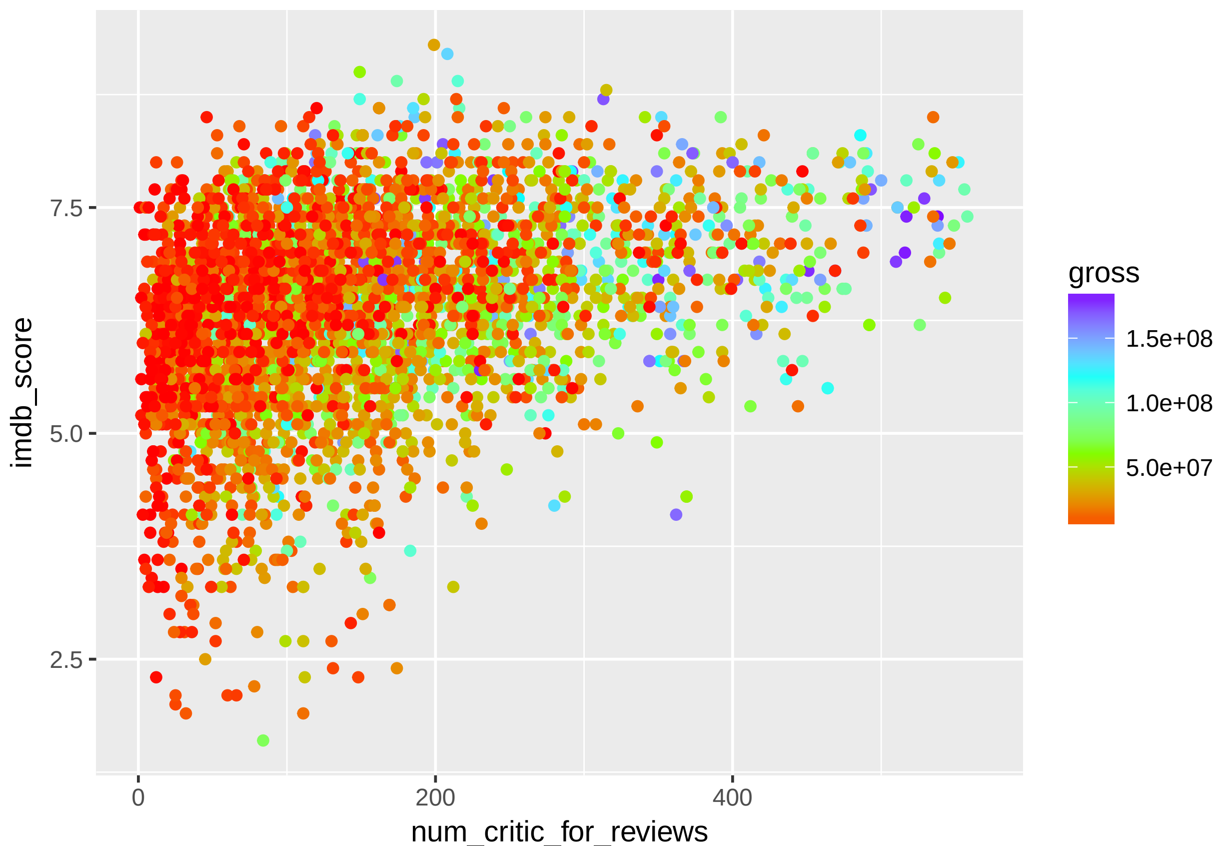
We can see their correlation here - duration generally increases with number of users who have voted. As the chart crosses the 250,000 mark for number of voted users, we increasingly see almost exclueively "long" and "very long" films. Additionally, you can see the correlation between duration and imdb_scores, with long and very long films appearing mostly in the 6.0 and above portion of the chart.

Let's replace duration with the number of Facebook likes a movie has as the third variable.



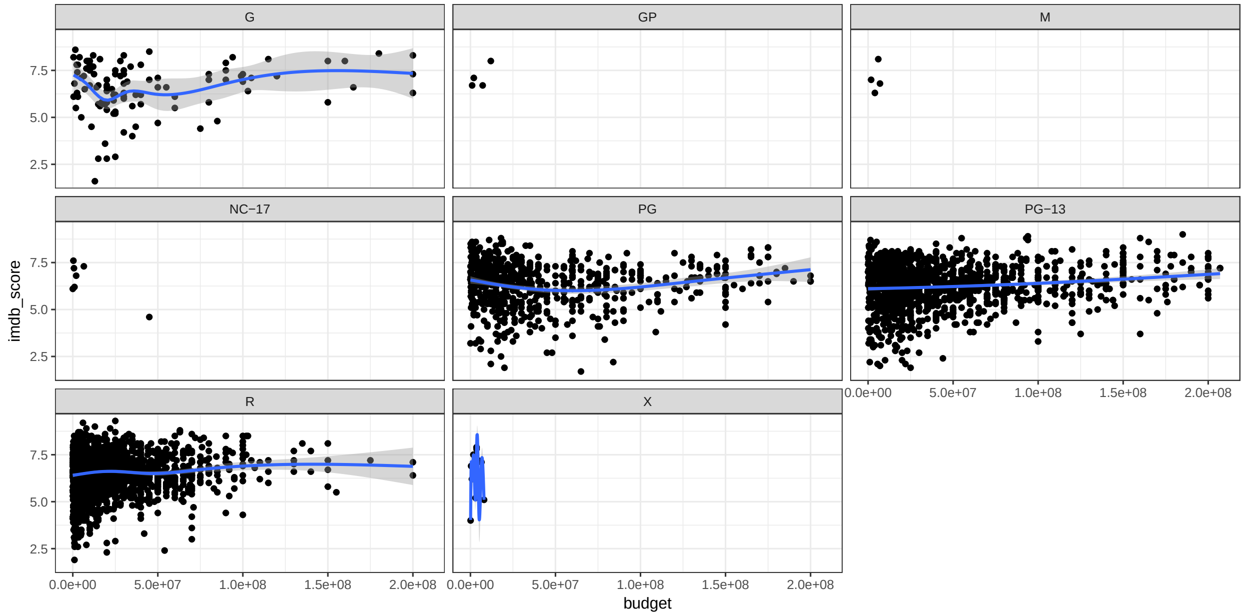
We're seeing a similar relationship between Facebook likes and number of voted users (0.531) as the relationship in the chart above. However, the correlation between imdb_score and movie_facebook_likes isn't as strong (0.277148858).

According to the corr plot we made earlier, the number of reviews from critics and gross have a correlation of 0.474. Let's visualize that within our imdb_score / num_critic_for_reviews scatter plot.



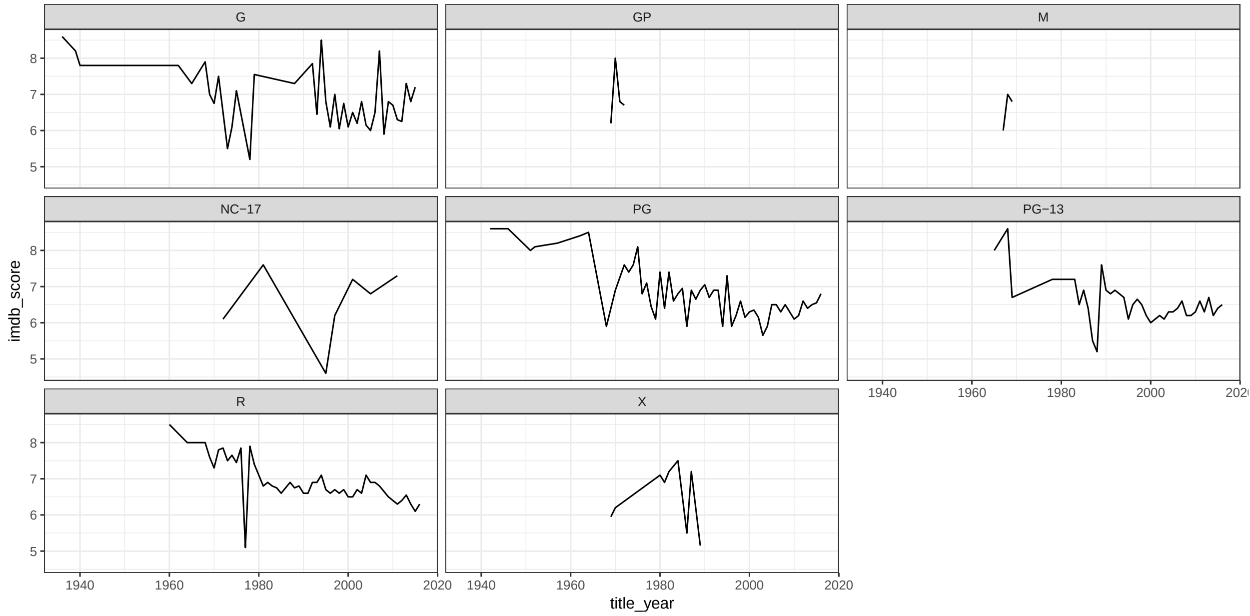
We can see gross' positively correlated relationship with both num_critic_for_reviews and imdb_score.

Earlier, we visualized the lack of correlation between budget and imdb_score. Is this relationship constant across all values of content_rating? Let's find out.



It appears the relationship between budget and imdb_scores is consistent across the content_ratings.

Finally, let's take a look again at the relationship between imdb_score and title_year, but broken down by content_rating.



Interesting. While the charts are consistent with the aggregated chart from earlier, G's chart stands out a bit. While IMDb scores is clearly still negatively correlated with the passing of time, the magnitude of the decrease appears to be less for G - especially since G's average IMDb score has received a substantial boost over the last couple years.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

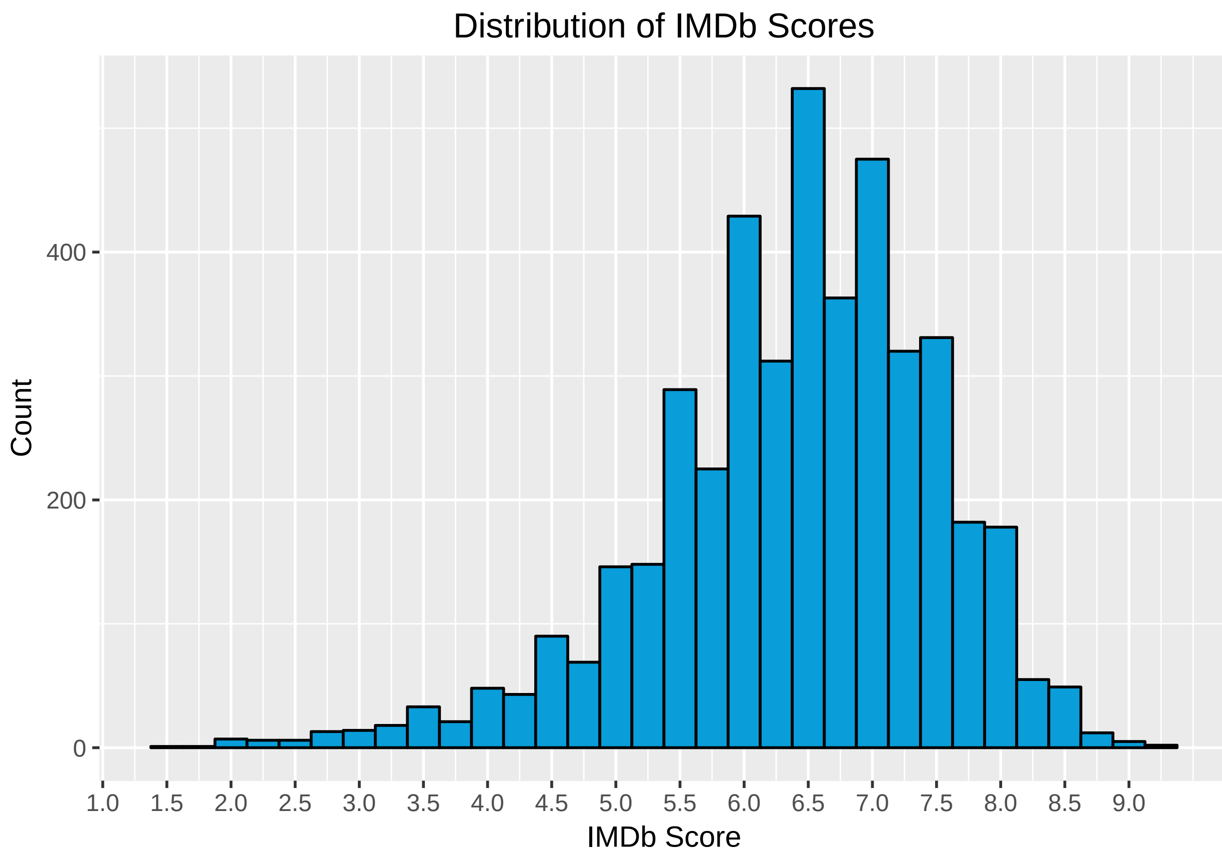
When looking at the relationship between IMDb and number of voted users, we saw number of voted users fairly strong positive correlation with duration and number of Facebook movie likes, and the two features' positive correlation with imdb_scores.

Were there any interesting or surprising interactions between features?

What surprised me was breaking down the relationship between title year and imdb_score across each of the content ratings. While they were essentially consistent to what we see in the bi variate analysis, G stood out as it currently had imdb_scores that were closer to the median imdb_scores of the older era of cinema compared to the rest of the ratings.

Final Plots and Summary

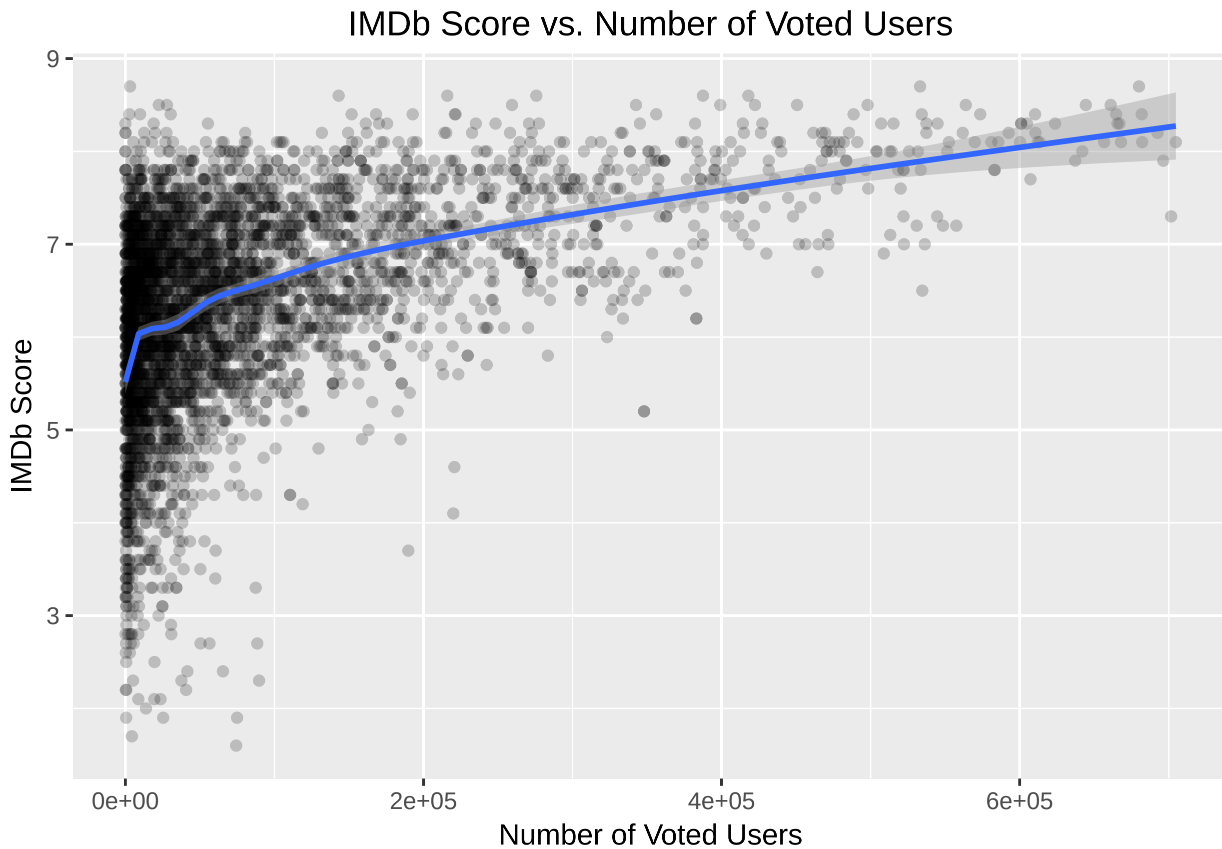
Plot One



Description One

This histogram gives a normal but slightly negatively skewed distribution of IMDb scores. The 1.0 to 5.0 portion of the chart contain what are known to be just bad films. Considering that most of these films aren't in this range, a film would have be pretty unpleasant to land here. The 5.0 to 6.5 IMDb score range are films that range from average to pretty good. 6.5 to 8.0 IMDb score range brings you films that go anywhere from "pretty good" to "great". This category of film likely represents most of the films you've loved. The 8.0 and onward range is the creme of the crop. The films that usually land in the IMDb's Top 250. The films that are timeless classics and referenced in any film class - i.e. The Godfather, 2001: A Space Odyssey, Singing in the Rain.

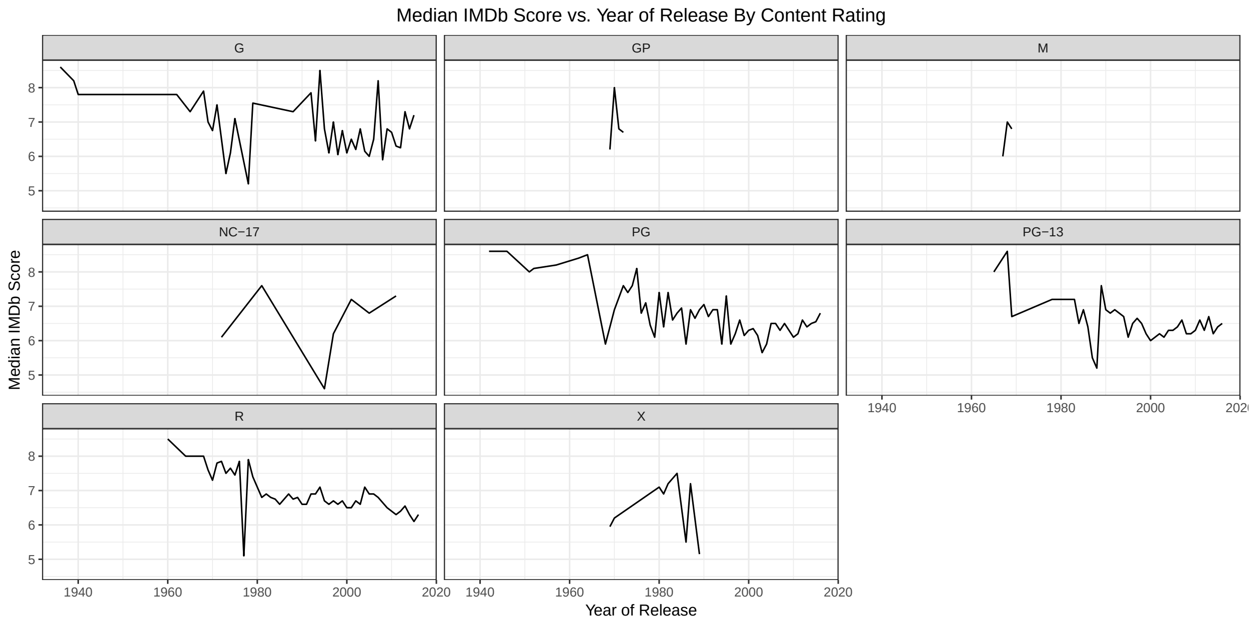
Plot Two



Description Two

Number of voted users has the strongest correlation with IMDb scores, with a Pearson's R of 0.462. In this scatter plot, we clearly see that's the case. Films with high IMDb scores (7.0 and above) exist across the entire spectrum of num_voted_users. However, as num_voted_users increases, the instances of low IMDb scores decrease, along with an increasing average of IMDb scores.

Plot Three



Description Three

While our Bivariate Analysis between IMDb score and the year of a film's release revealed a negative correlation, breaking down the trend across content ratings revealed one stand-out insight. While G rated films generally saw median IMDb scores lower than any of the other ratings, the last several years have seen G rated films with the highest median IMDb scores. Additionally, while films saw a peak in median scores early on in their history that films today haven't been able to match, G rated film's recent median scores are getting close to those scores of the past. Are we seeing a renaissance of high quality G rated films at the moment?

Reflection

The original dataset had 5043 movies in this dataset with 28 variables. However, initial exploration revealed there were non-theatrical films in the dataset, such as made-for-TV films and TV series. We were left with 4423 films and 28 variables after removing them.

I began the exploration with a simple univariate analysis, to look at the structure and distribution of the film features. Occasionally I had to apply a log transformation in order to develop a clearer picture of the distribution. Moving forward, I determined that the IMDb score would be the main feature of interest in this project. Meaning, I wanted to see what variables correlated with a high or low IMDb score.

In the following Bivariate Analysis section, I then compared the relationship between IMDb scores and the features I highlighted in the univariate section. The most important findings were that IMDb scores had weak to fairly strong positive correlations with number of reviews from critics, duration, and number of users who voted.

Finally, in our Multivariate Analysis section, I decided to see how certain features related to one another while looking at my main feature of interest (imdb_scores). The one interesting finding in this section was finding out that while IMDb scores and the year of a film's release was still negative across all content_ratings, rated G films currently had median IMDb scores higher than the rest of the categories and close to the median scores of older films.

In the future, there are a variety of ways this project could be improved on or expanded:

1. Earlier in my analysis, I noticed that there were a lot more rated R films than PG 13, and wondered why this was so. Outside research led me to a 2013 article from the The Wrap that discussed this exact phenomenon - TheWrap: If PG-13 Is the Moneymaker, Why Is Hollywood Cranking Out So Many R-Rated Movies? Two main reasons popped up: 1. Rated R films allow full creativity and thus tend to be the award winners. PG-13 films are made to generate big time money. 2. Rated R films not performing too well in the box office aren't always an overt concern because Video-on-Demand/streaming is where the real revenue is. With this in mind, I'd love to scrape data on VOD sales and awards won for each film and see if the data backs it up.
2. I'd approach this project with building a predictive model for IMDb scores in mind. As most of the variables I looked at in this project weren't variables that are available prior to a film's release (i.e. gross, number of voted users), I'd take into account different variables, such as movie_facebook_likes.
3. I'd be interesting to see demographic breakdowns of a film's cast and director (and even include writer, producer, etc.). Not only would it give us a sense of demographic representation in cinema over time, but we can then find relationships between demographics and awards, genres, box office performance, and more.