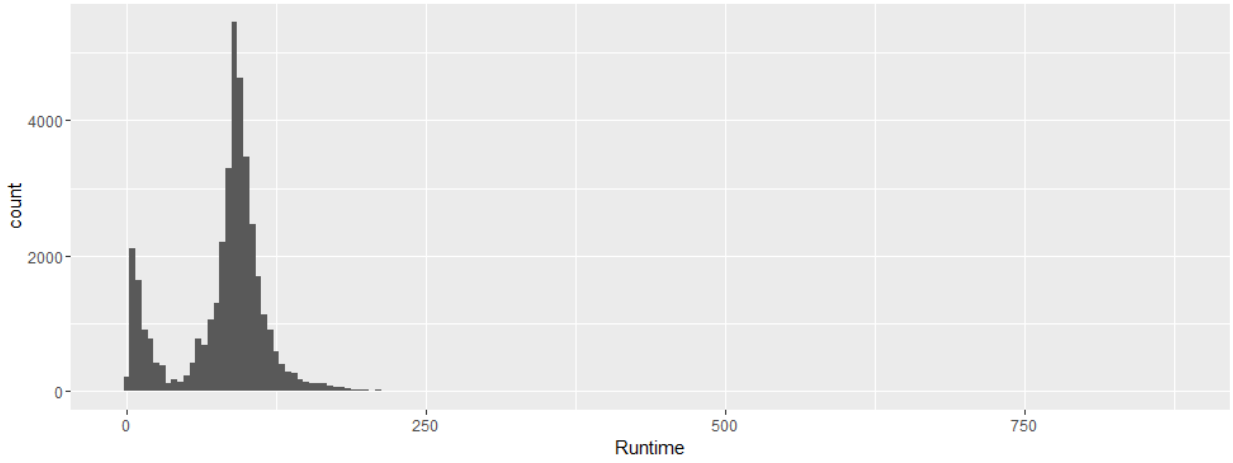
# DAVA Project 1 Report

Daniel Rozen, drozen3

1. Removed 789 rows from the data set.

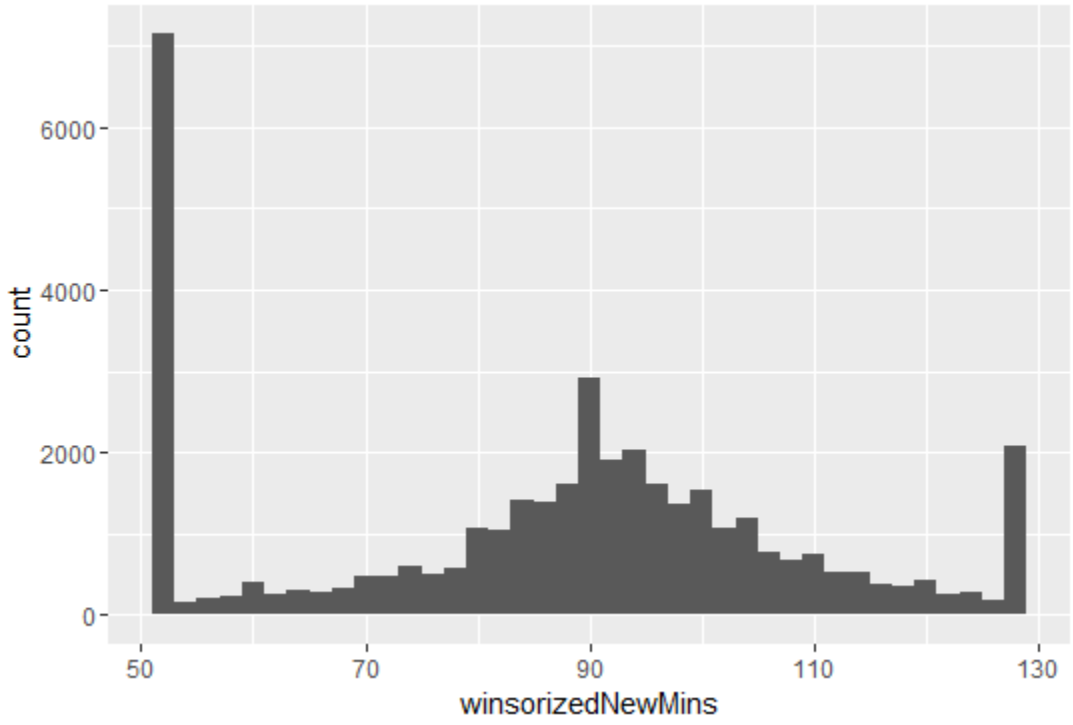
2. Investigate and describe the distribution of Runtime:

Plotting a histogram of Runtime in minutes with a bin-width of 5 minutes, gives us the following plot:



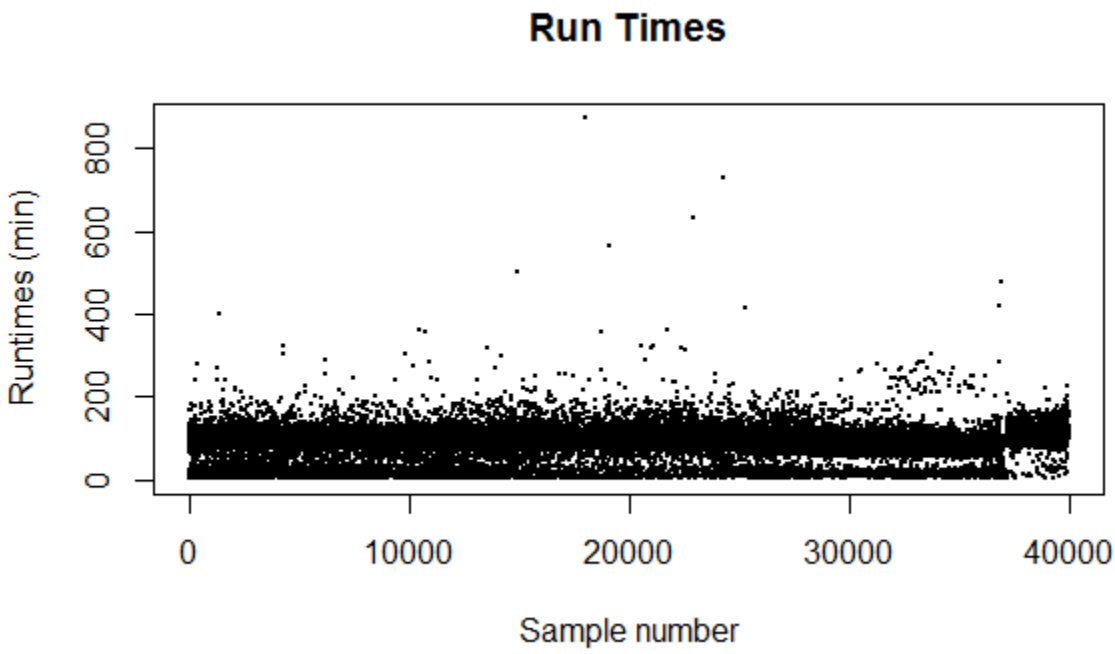

We find a multimodal distribution with a small peak at 10 minutes and then leveling off to the right with a Gaussian like curve. Then there's a 2nd large peak at around 100 minutes with a Gaussian like curve.

If we winsorize the outliers with the winsorize function with a bin-width of 4 we find:

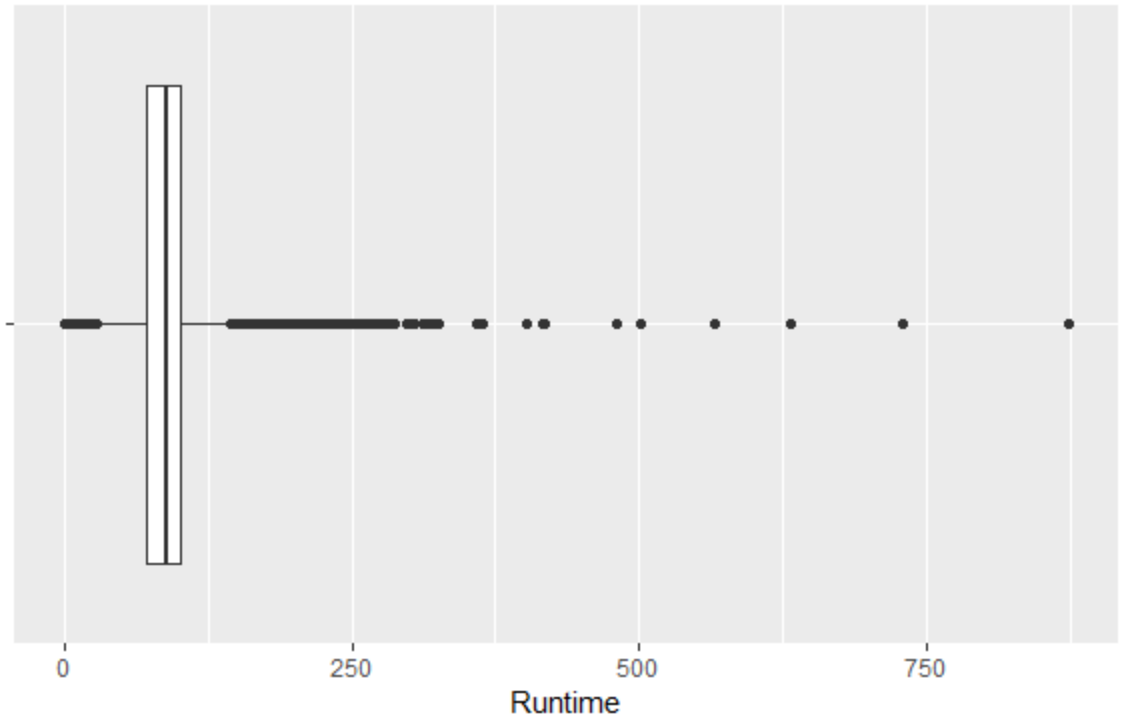



There are also a number of outliers past the 250 minutes mark.

The following scatterplot also reveals the multimodal distribution, with a smaller lower band at around 10 minutes and a thicker upper band at around 100 minutes.

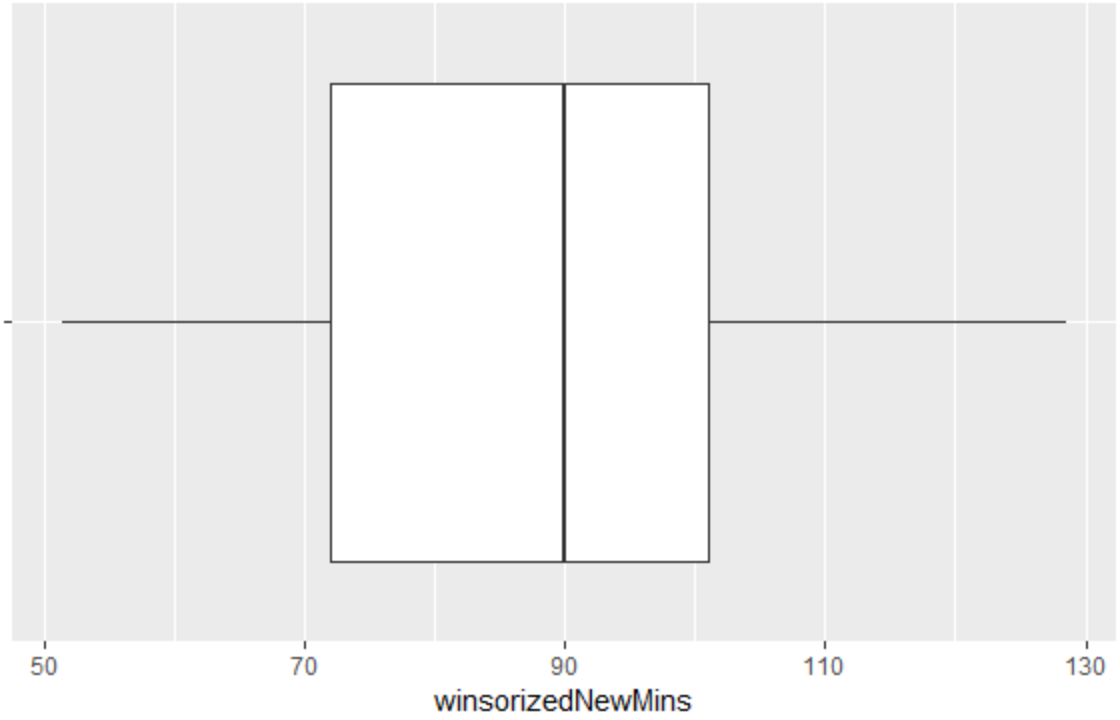



Plotting a boxplot:



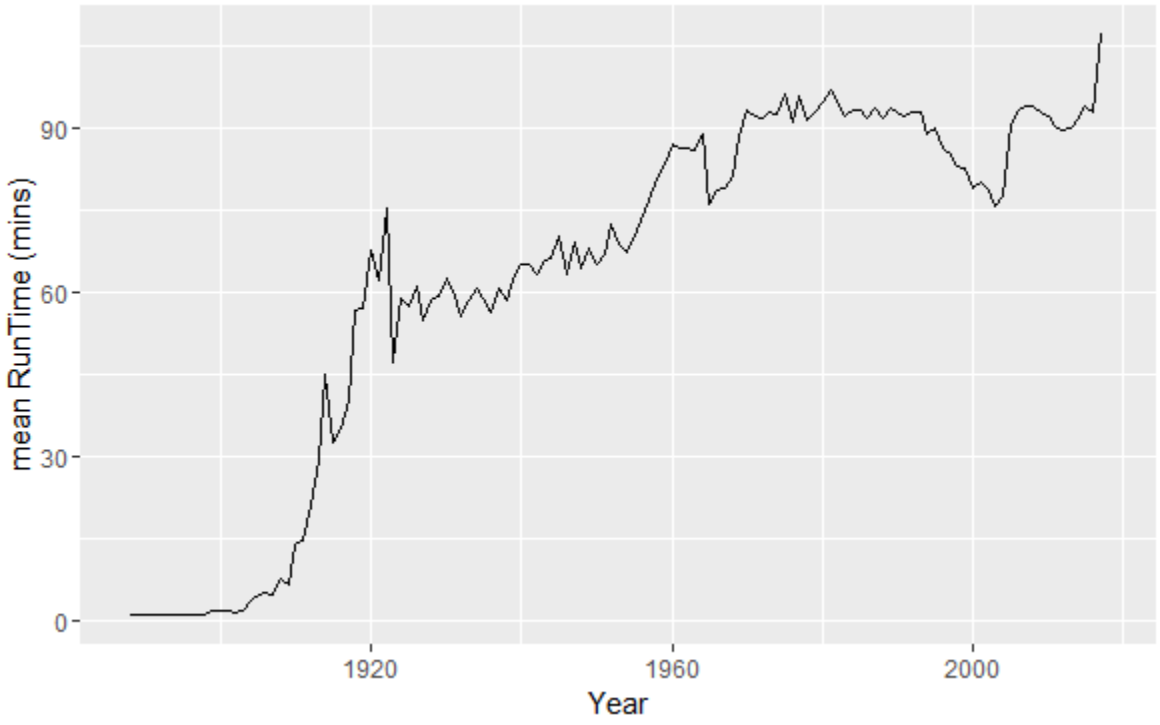

This shows us that the distribution is skewed to the right with a median of around 100 minutes. We also see many outliers to the right.

Winsorizing the data we get:



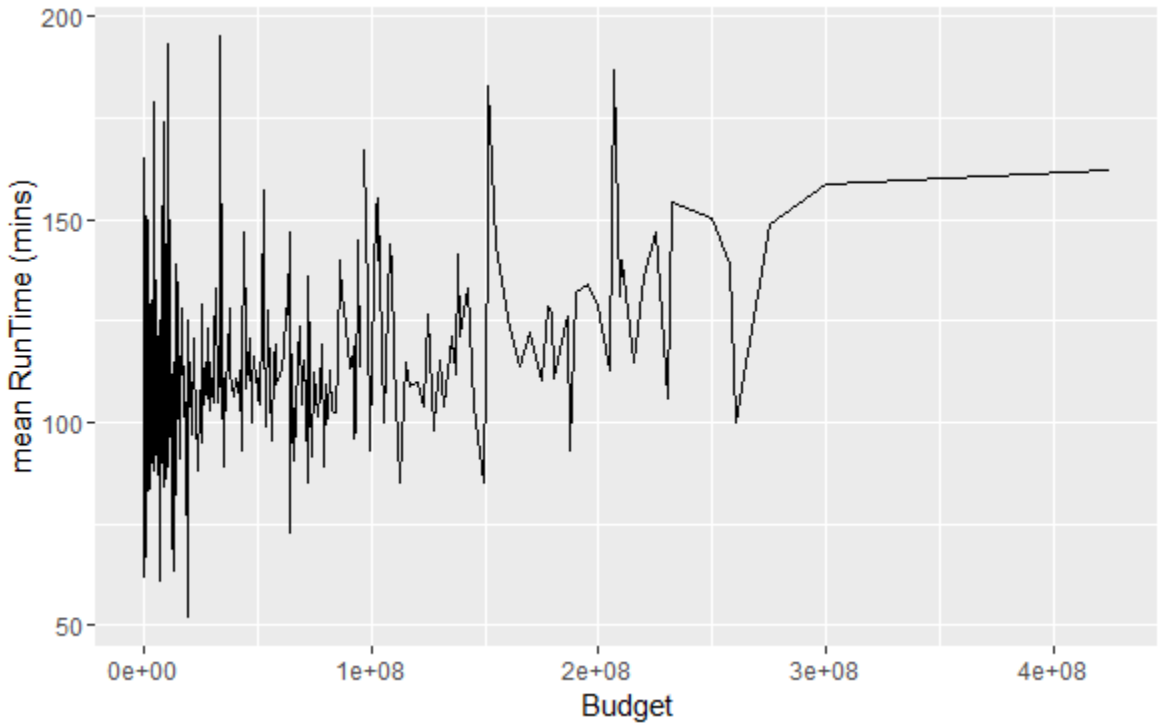

Here it's much easier to see the median at 90 minutes with the right skew.

Plotting mean Runtime vs. Year gives us the following graph:




We see that the mean runtime starts off very low at the beginning with a large fast increase from around 1910 till 1920, a small dip and then a general steady increase.

Plotting mean Runtime vs. Budget gives us the following graph:

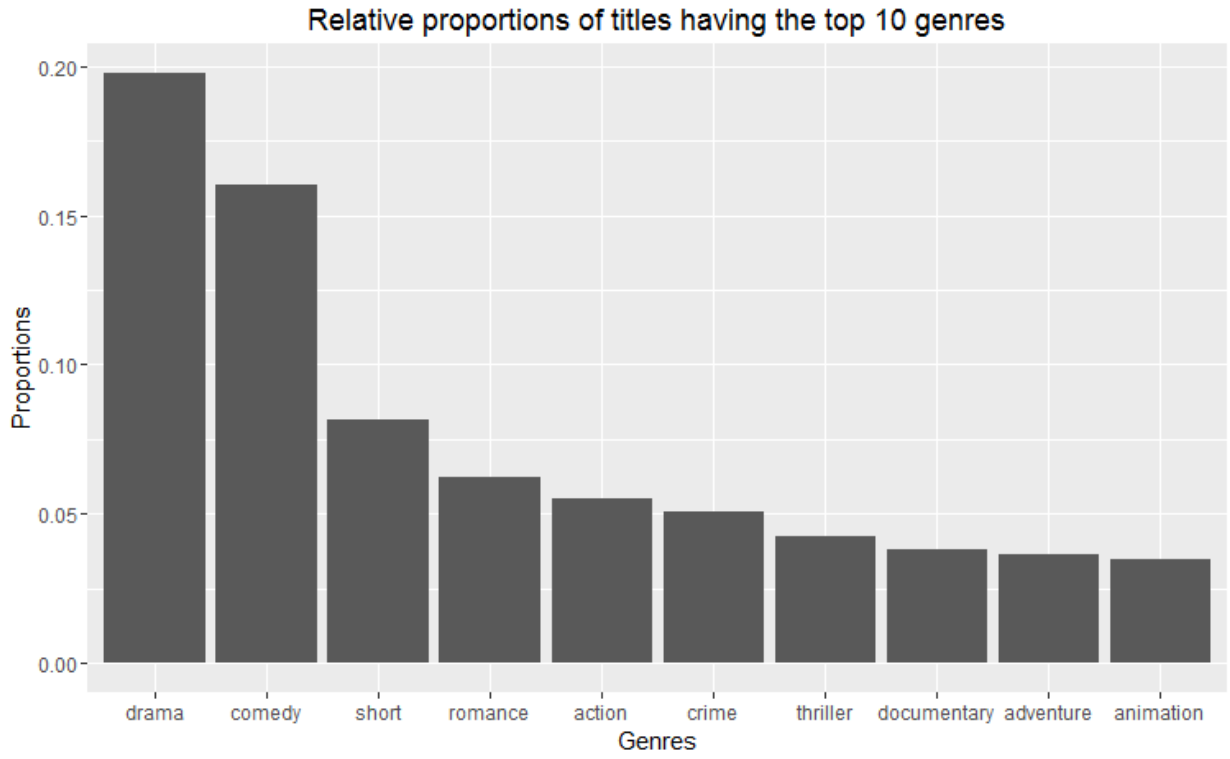



Here we can see the runtime is quite volotile but with a slightly increasing trend with increasing budget

3.

#Graph and describe the relative proportions of titles having the top 10 genres

I convert to proportions in percentages by dividing by the top 10 total and not the entire total of

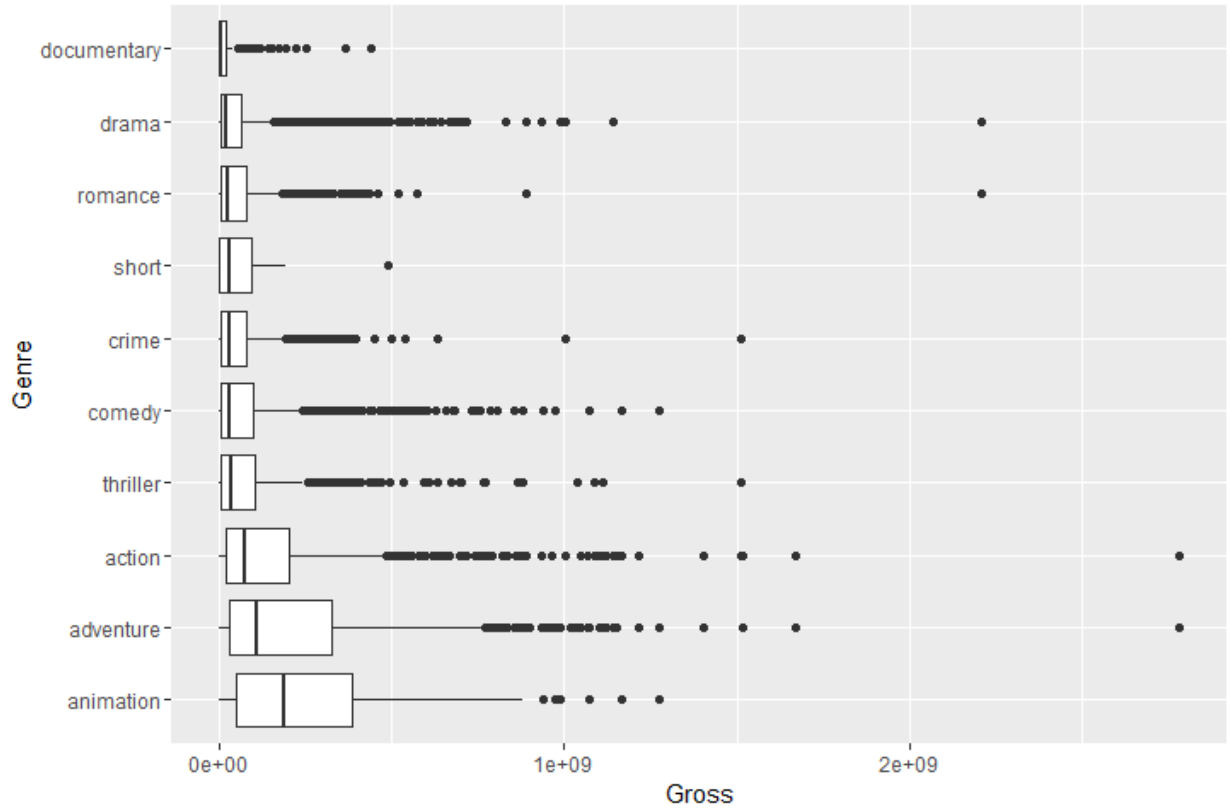



Top 10 proportion table:

| | |
|---|---|
| drama | 0.19807162 |
| comedy | 0.16047810 |
| short | 0.08138184 |
| romance | 0.06213546 |
| action | 0.05511634 |
| crime | 0.05073251 |
| thriller | 0.04221465 |
| documentary | 0.03810559 |
| adventure | 0.03656937 |
| animation | 0.03482084 |

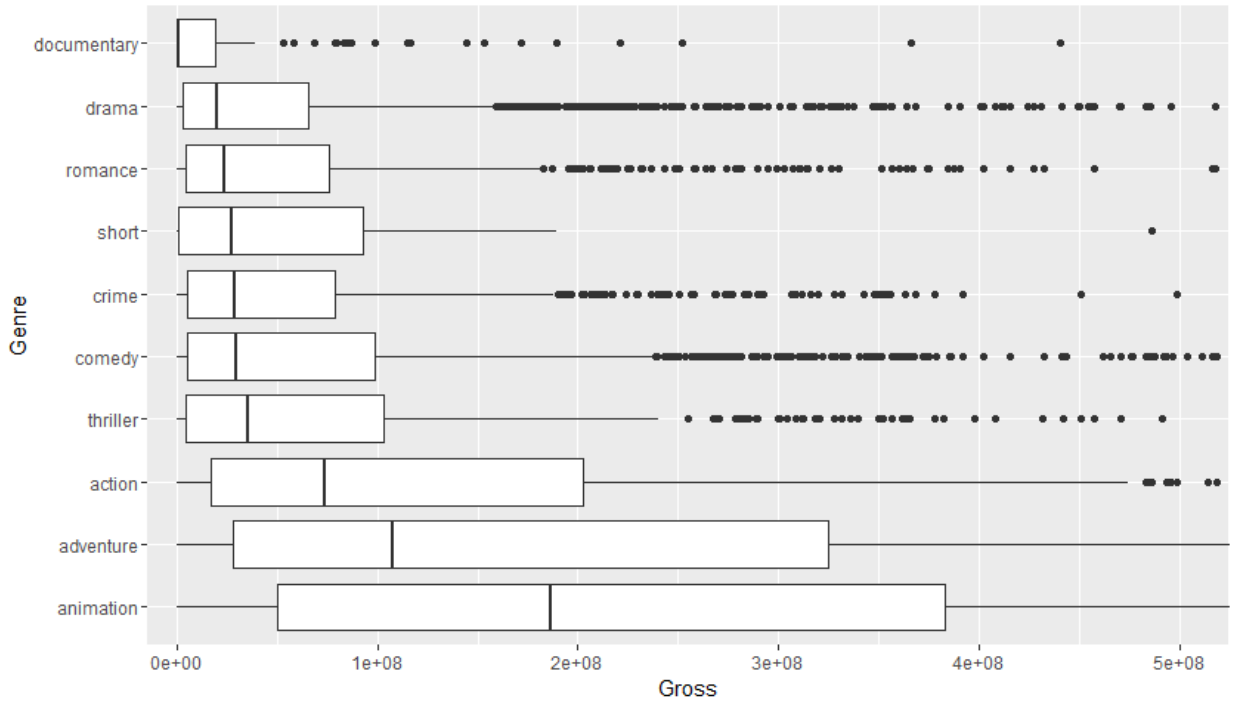
Viewing the above graph and table we find that drama and comedy have much larger proportions than the rest of the genres, followed by short, and then smoothly decreasing till the end of the genres.

Plotting multiple box plots of Gross revenue vs. Genre we find:

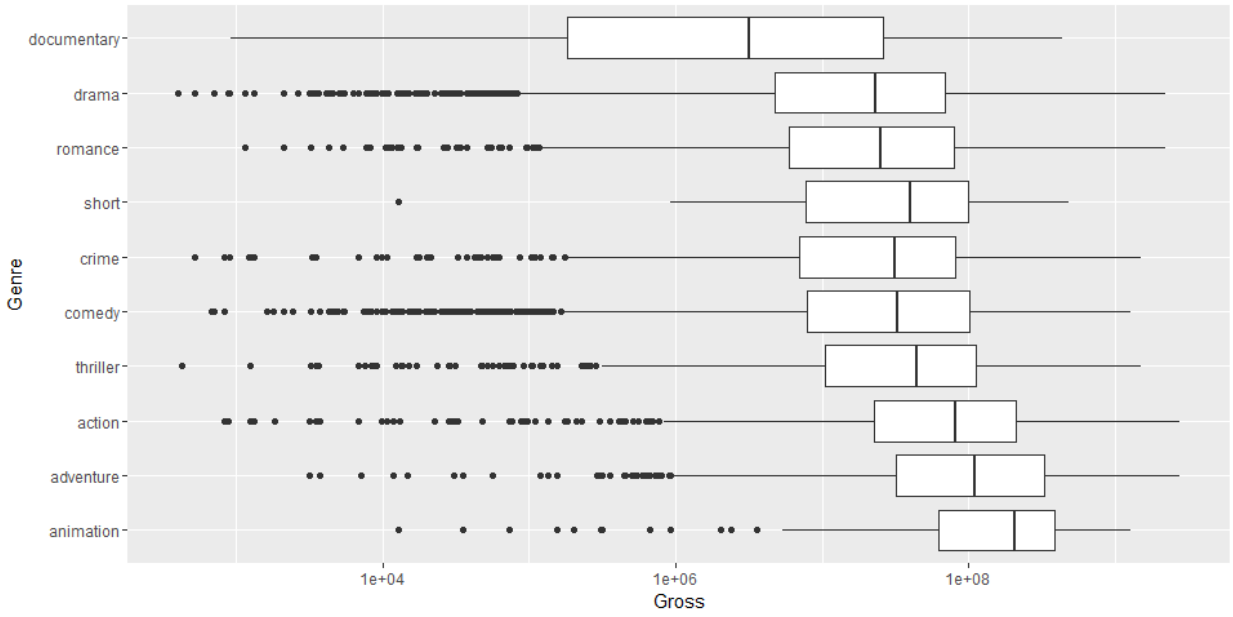



Outliers are making it difficult for comparison. We do see clearly that Animation has by far the highest median, followed by adventure and action. These top 3 genres (in terms of Gross) also have the greatest spread.

Zooming in we find:




Plotting on a log x-scale we find:




With the above 2 plots we can see documentary has a much lower median Gross than the rest.

Drama, romance, crime, comedy and thriller are all more or less similar in terms of median and spread.

Short has a higher median and much less of a spread beyond the IQR, with shorter whiskers and only 1 outlier in comparison to long whiskers and many outliers for the rest.

4.

*What is your precise removal logic and how many rows did you end up removing?*

Initially I removed all rows that didn't match. But that removed too much.

rowsRemoved= 815
percentRowsRemoved = 17.88065

So my later logic was to prevent rows that had NA in the released column from being removed. Which improved slightly.

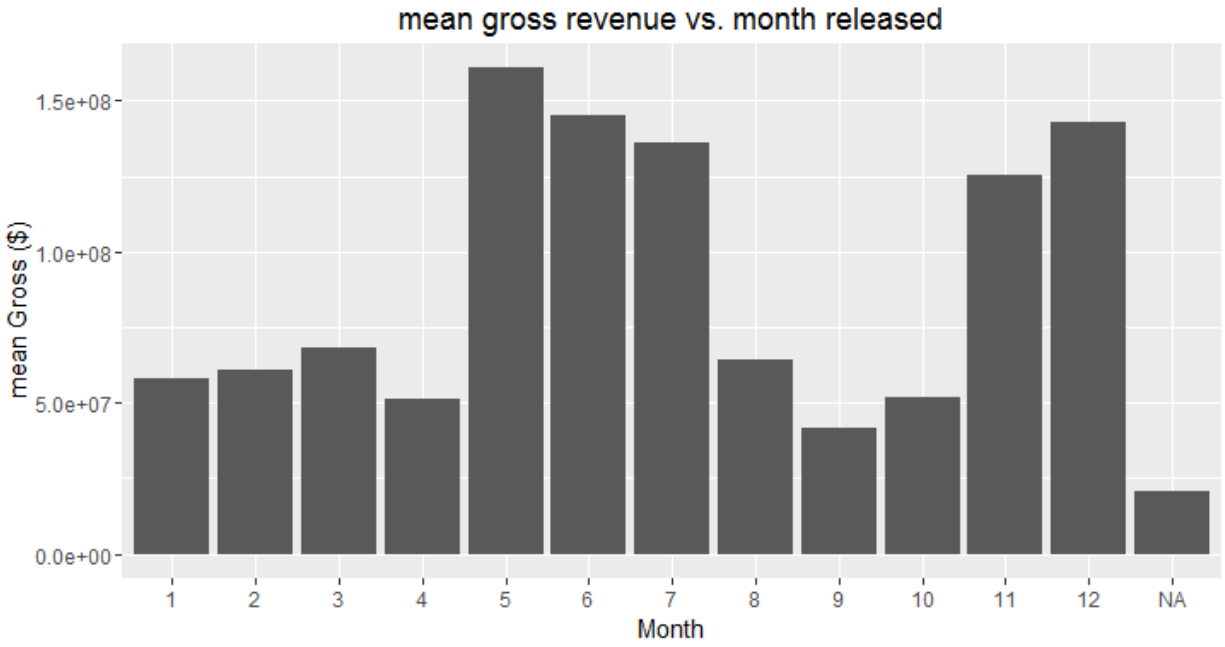
rowsRemoved = 770
percentRowsRemoved = 16.89337

Finally I changed my logic to removing rows which had Released years that came more than 2 years after the Year column with the following drastically improved results. Once again I prevented rows that had NA in the released column from being removed.

rowsRemoved = 20
percentRowsRemoved = 0.4387889

My precise removal logic was that sometimes movies are delayed in being released up until an average of 2 years. Therefore within this threshold should be assumes to be good data. Beyond the threshold could be bad data. I also avoided removing rows with NA's in the Released-Year column as for those there was no way to know if there is bad data based on the discrepancy between Year and Released-Year.

5. *Investigate the relationship between release date and gross revenue and comment on what tim es of year are most high revenue movies released in.*

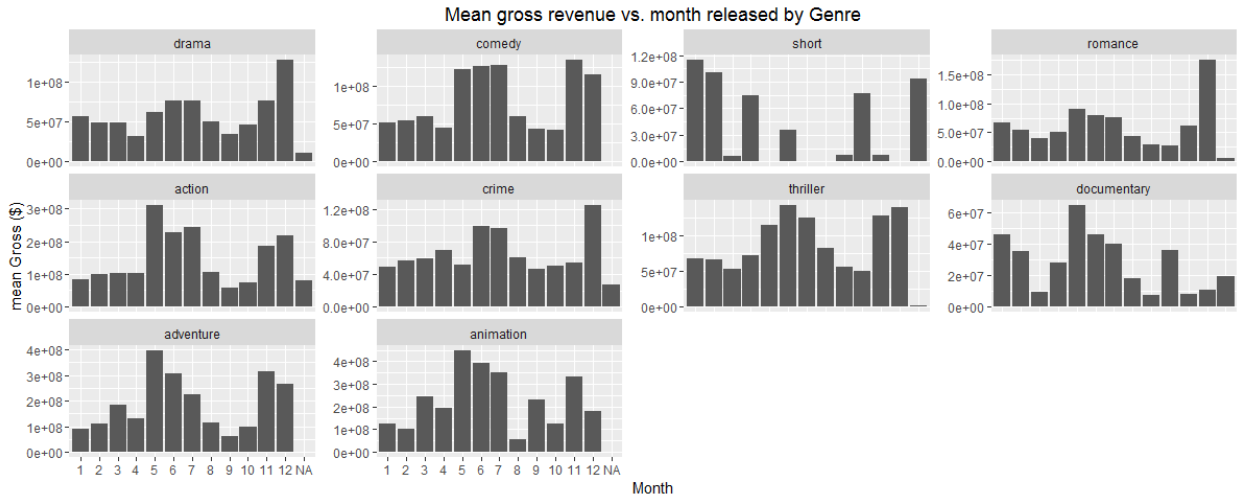



| | ReleasedMonth | mean.Gross |
|---|---|---|
| 1 | 1 | 57836688 |
| 2 | 2 | 61128507 |
| 3 | 3 | 68021453 |
| 4 | 4 | 51247655 |
| 5 | 5 | 161114758 |
| 6 | 6 | 145353909 |
| 7 | 7 | 136138074 |
| 8 | 8 | 64131995 |
| 9 | 9 | 41734009 |
| 10 | 10 | 51597923 |
| 11 | 11 | 125460536 |
| 12 | 12 | 143038741 |

We can see clearly from the above graph and table that that May, June July have the greatest gross revenue, which makes sense as they are summer months when people are on break. Also November and December have high revenues which are also break times.

*Does your answer changes for different genres?*

Plotting bar charts for the top 10 genres:




We see that the answer doesn't change for most genres. However it does change for short and documentary. Documentary is highest around May June period, but also January and February. Short is high in January and February but very small during the summer months.

To analyze the reliability of these graphs, I computed the number of samples per genre

```
> genreCount
     variable nrow
1       drama 2249
2      comedy 1713
3       short   21
4     romance  784
5      action 1025
6       crime  792
7    thriller  705
8 documentary  132
9   adventure  818
10  animation  206
```

*Based on the data, can you formulate a genre-based recommendation for release date that is likely to increase the title's revenue? If you have a recommendation motivate it with the appropriate disclaimers, or otherwise explain why you are unable to produce a recommendation.*

For all of the Genres except for short and documentary, I would recommend a release date of May, June, July, or December.

Specifically I would recommend romance, drama, and crime to be released in December. For Documentary I would recommend May.

However since short has such a small number of samples, I wouldn't be able to rely on these statistics very heavily. Similarly with documentary and animation, I'm not sure if these are enough samples, however they do follow the general trend of all of the movies combined together.

Disclaimer on all the above: data sets ranged between 21 and 2249, which may not be sufficiently large to be accurately relied upon.

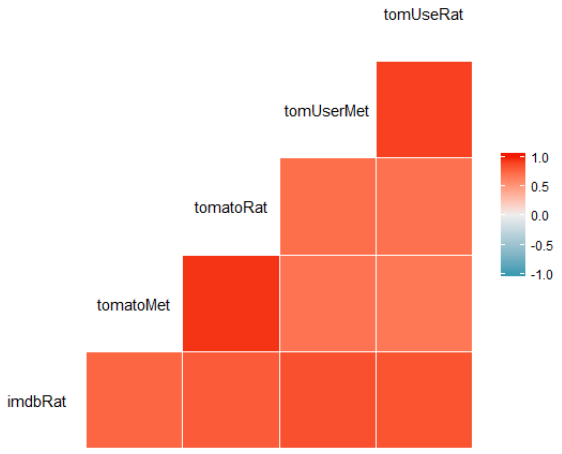
6. *investigate the pairwise relationships between these different descriptors using graphs.*

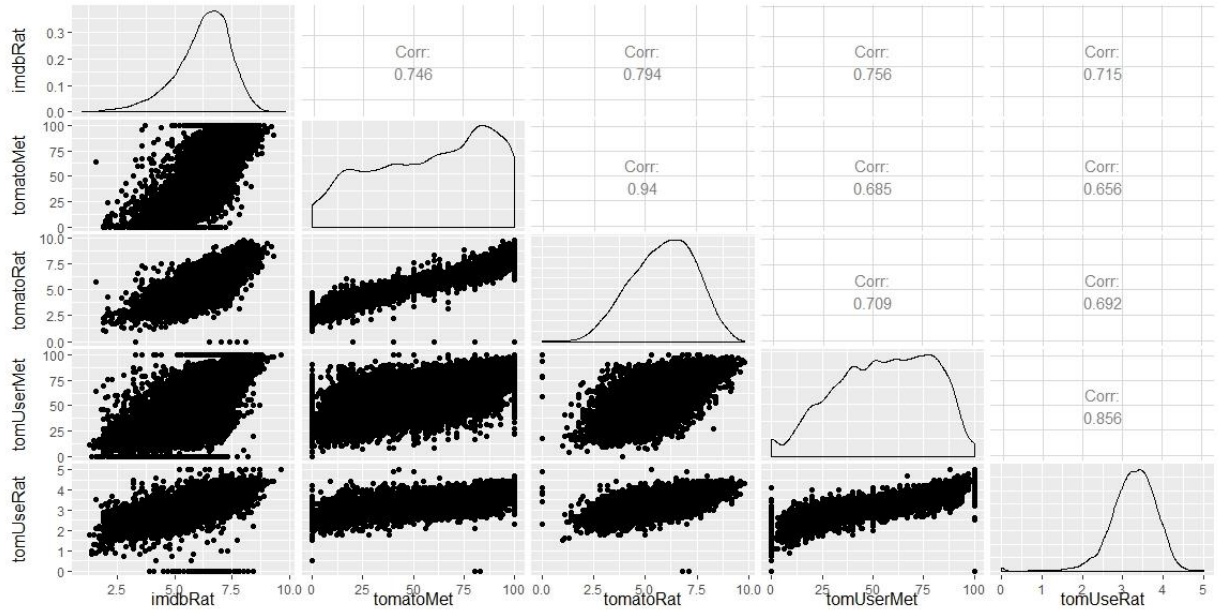
I used all of the data including all of the rows with missing Gross values for higher accuracy.

For correlation categorization, I used the following guideline from http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

> Correlation is an effect size and so we can verbally describe the strength of the correlation using the guide that Evans (1996) suggests for the absolute value of *r*:
>
> - .00-.19 "very weak"
> - .20-.39 "weak"
> - .40-.59 "moderate"
> - .60-.79 "strong"
> - .80-1.0 "very strong"

I renamed the columns for easier readability and produced the following pairwise correlation plots





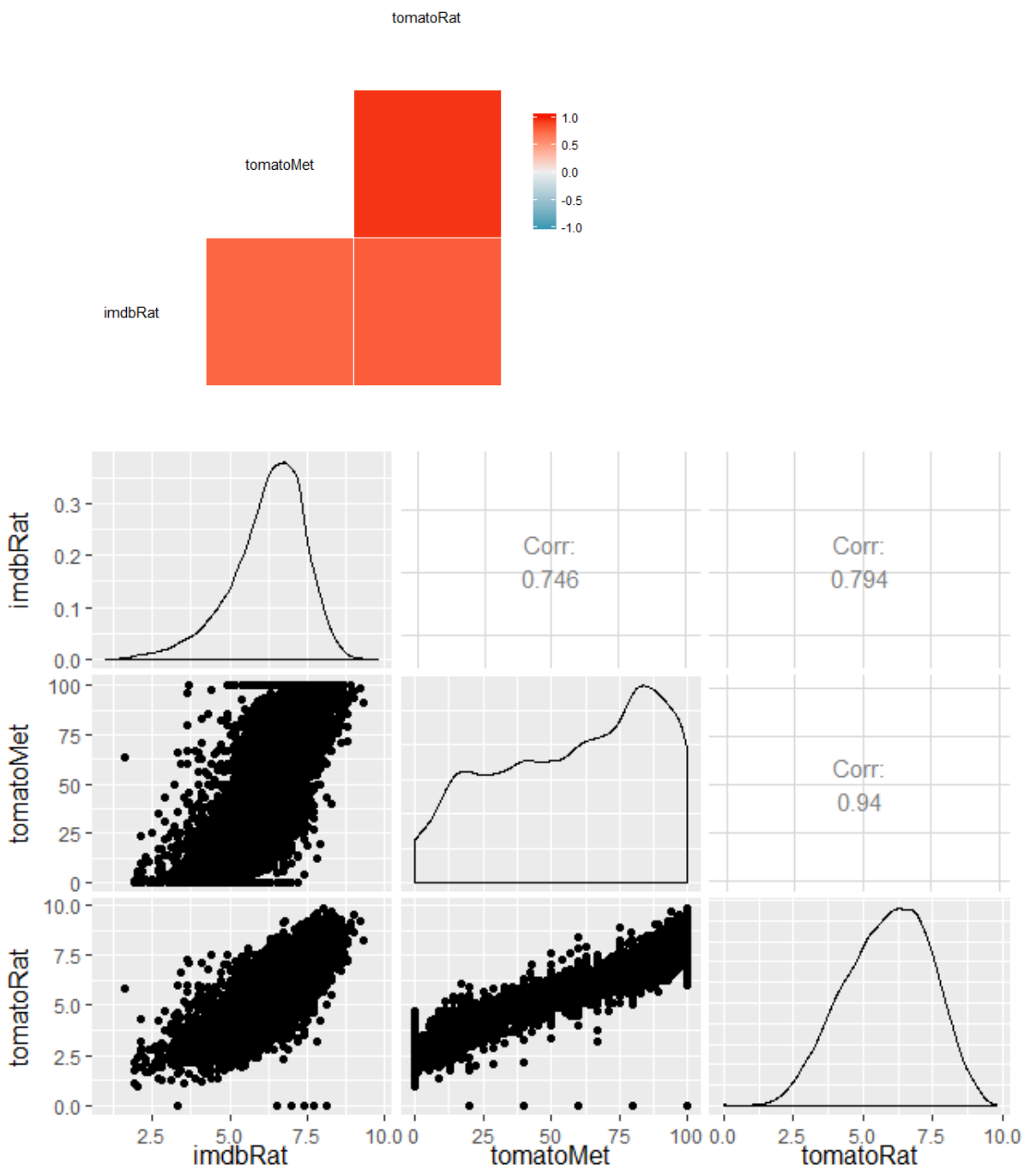

We see from the above graphs that Tomato Meter and Tomato rating have an extremely strong positive correlation of 0.94 because there are both ratings provided by Critics.

TomatoUserMeter and TomatoUserRating also have strong positive correlations because they are both ratings provided by Users.

Imdb Rating has a positive correlation with the other meters and ratings. This makes sense since the movie critics should be giving movie ratings that the general public agrees to. tomatoMeter and tomatoRating represents critic opinion whereas the remaining meters and ratings represents the public opinion.

However, it is interesting that the lowest correlations are with tomatoMeter and tomatoRatings vs. TomatoUserMeters and tomatoUserRatings, which suggests a greater discrepancy between tomato critics and tomato users than that with tomato critics and Imdb users. However they all still have strong positive correlations.

In order to make the investigation of pairwise relationships more manageable, I initially reduced my investigation to imdbRating, tomatoMeter, tomatoRating, tomatoUserMeter, and tomatoUserRating. Anything that wasn't a rating or part of imdb or tomato I removed. Eg. I removed votes, images, reviews, etc. and produced the following plots:



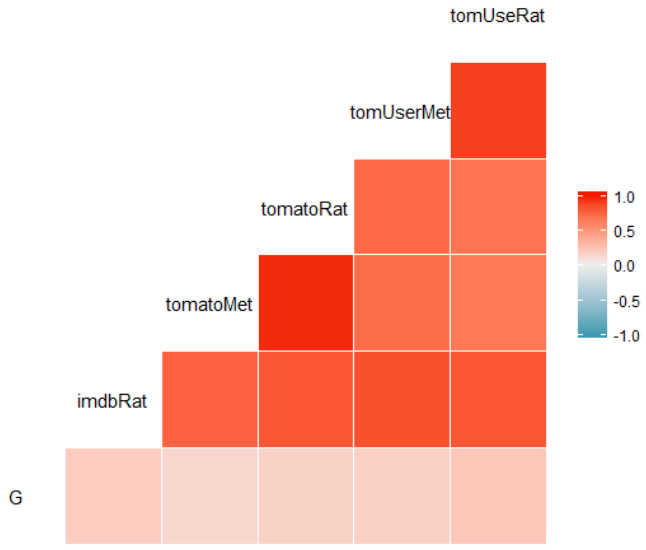

*Comment on similarities and differences between the user ratings of IMDb and the critics ratings of Rotten Tomatoes.*

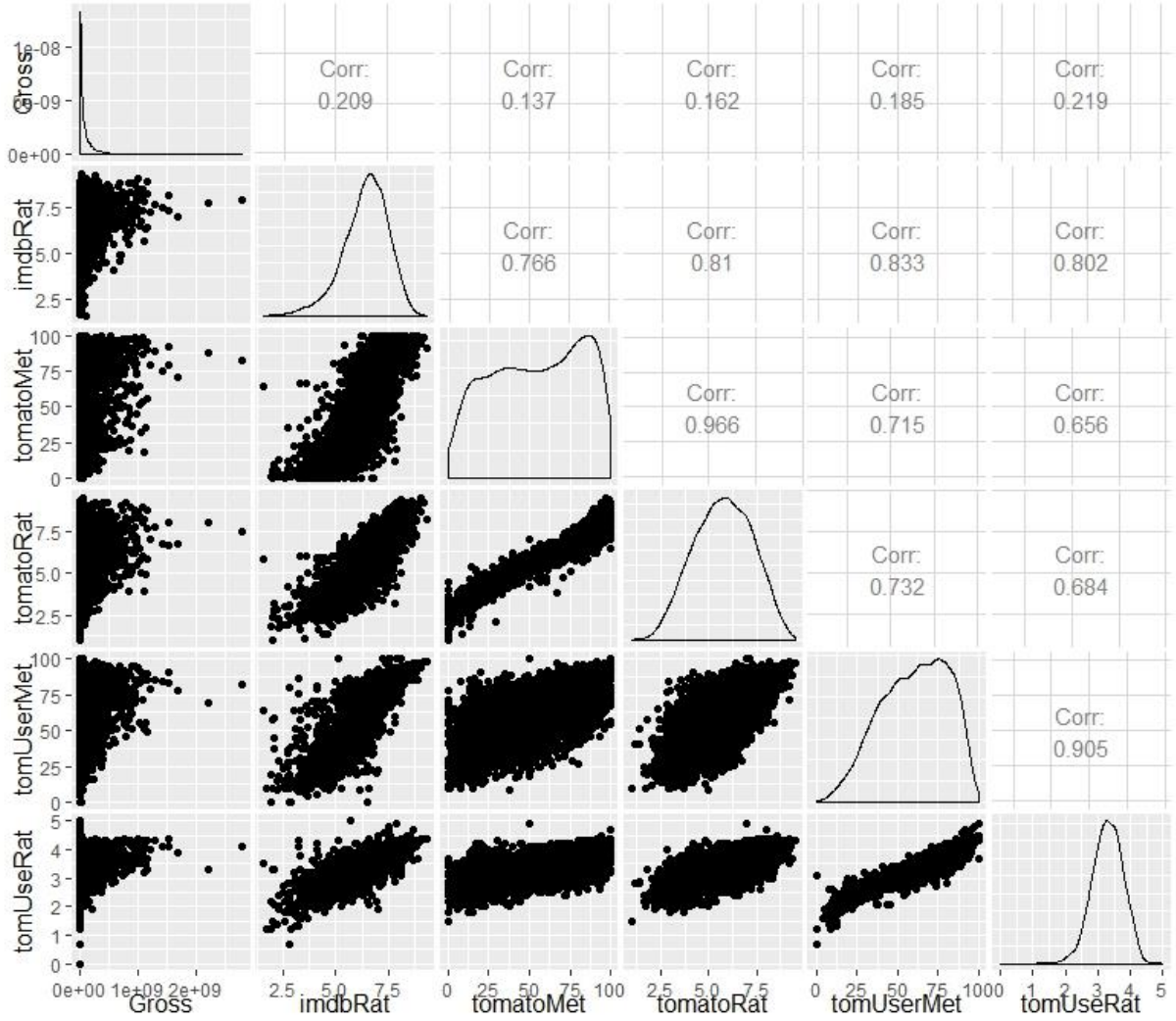
Imdb Rating has a strong positive correlation with the critics tomatoMeter and tomatoRatings. This makes sense since the movie critics should be giving movie ratings that the general public agrees to. tomatoMeter and tomatoRating represents critic opinion whereas the remaining meters and ratings represents the public opinion.

*Comment on the relationships between these variables and the gross revenue.*

*Which of these ratings are the most highly correlated with gross revenue (use the R function cor and remove rows with missing values)?*

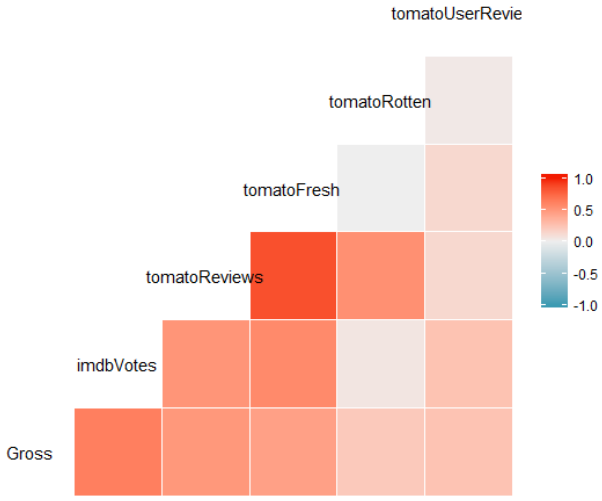
For easy readability, we graph the correlations of the meters/ratings variables with Gross:

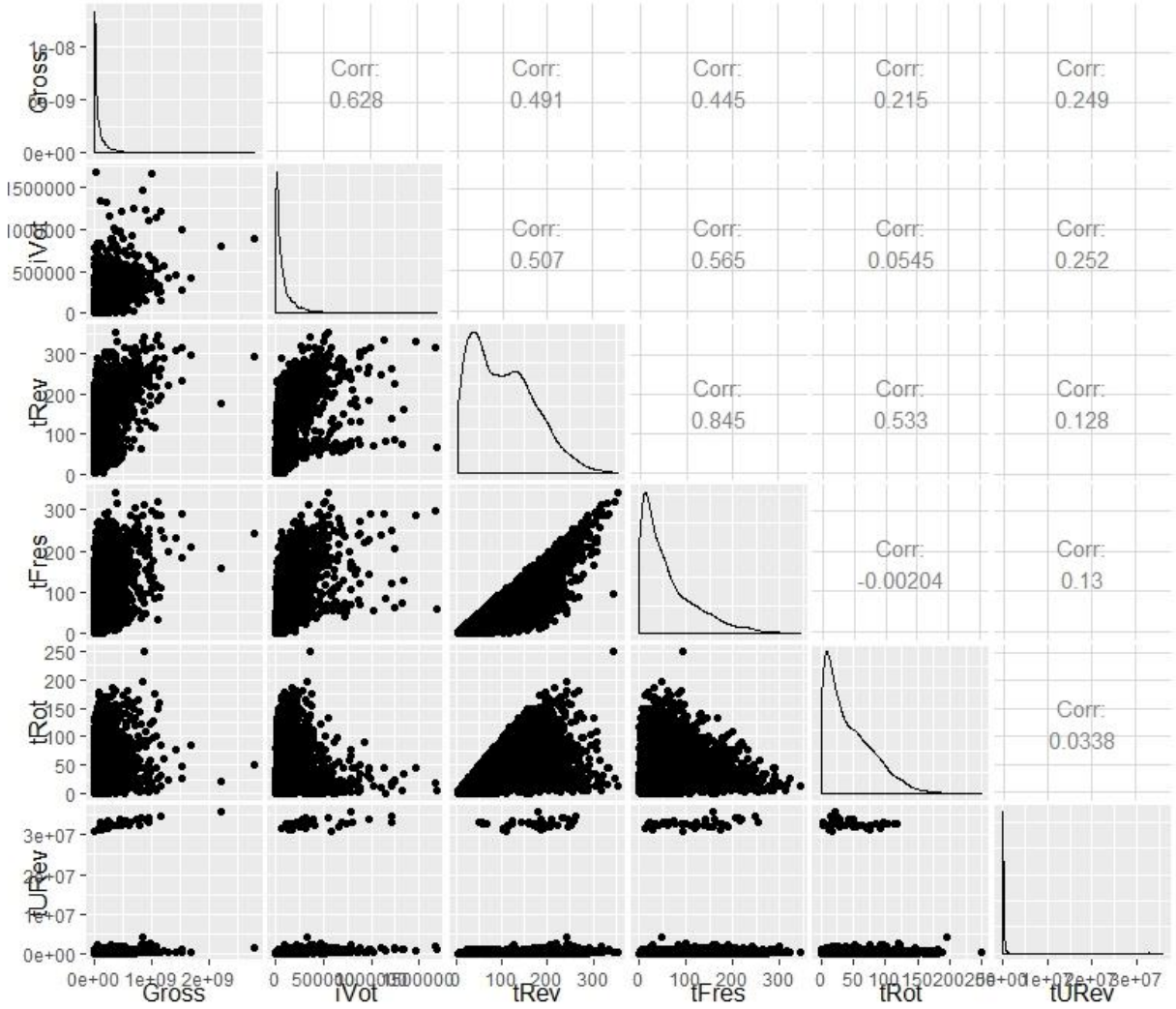





We see from the above that imdb Rating and tomatoUserRating have weak positive correlations with gross, whereas the rest of very weak positive corresponding correlations with Gross. This shows that viewers are not very much influenced by ratings in deciding whether to view a movie or not. This could be because viewers are curious to see movies they're interested in for themselves despite the ratings of others.

For easy readability, now we graph the correlations of the non-meters/ratings variables with Gross:


tomatoUserRevie
tomatoRotten
tomatoFresh
tomatoReviews
imdbVotes
Gross
1.0
0.5
0.0
-0.5
-1.0


Gross
lVot
tRev
tFres
tRot
tURev
Corr: 0.628
Corr: 0.491
Corr: 0.445
Corr: 0.215
Corr: 0.249
Corr: 0.507
Corr: 0.565
Corr: 0.0545
Corr: 0.252
Corr: 0.845
Corr: 0.533
Corr: 0.128
Corr: -0.00204
Corr: 0.13
Corr: 0.0338

Here we see that Gross has the highest strong positive correlation with imdbVotes. This makes sense since users should only be voting on movies they've seen. Therefore the more people have seen the movie, which results in higher Gross, the more people should be voting. However, not everyone who sees the movie will vote. Next highest is tomatoes reviews and tomatoFresh, which also makes sense because more reviews implies more watchers which implies higher Gross. I assume this is also the reason for a weak positive correlation for tomatoesRotten. Even though the movies got negative reviews, more negative reviews implies more people saw the movies.

*7.*

*How did you construct your conversion mechanism?*

I constructed a function as the following based on an upper threshold, which was determined with quantiles:

```
> quantile(unlist(xsum),seq(0, 1, length.out = 41))
   0%  2.5%    5%  7.5%   10% 12.5%   15% 17.5%   20% 22.5%
    0     0     0     0     0     0     0     1     1     1
  25% 27.5%   30% 32.5%   35% 37.5%   40% 42.5%   45% 47.5%
    1     2     2     2     3     3     4     4     4     5
  50% 52.5%   55% 57.5%   60% 62.5%   65% 67.5%   70% 72.5%
    5     6     7     7     8     9    10    11    13    15
  75% 77.5%   80% 82.5%   85% 87.5%   90% 92.5%   95% 97.5%
   17    19    22    26    31    37    45    58    79   127
 100%
  548
```

I chose the 87.5% quantiles = 37 as the upper threshold since 0's represented roughly the 1st 15% Therefore I deterimend from the "some" quantile range 87.5%-17.5% / (100%-87.5%) ("many" quantile range) = 5.6 gave us close enough to the desired ratio of 5.

```
convertFun = function(x) {
  upperThreshold = 37
  if (x == 0) {x = "none"}
else if (x >=1 & x<=upperThreshold) {x="some"}
else if (x > upperThreshold) {x= "many"}
}
```

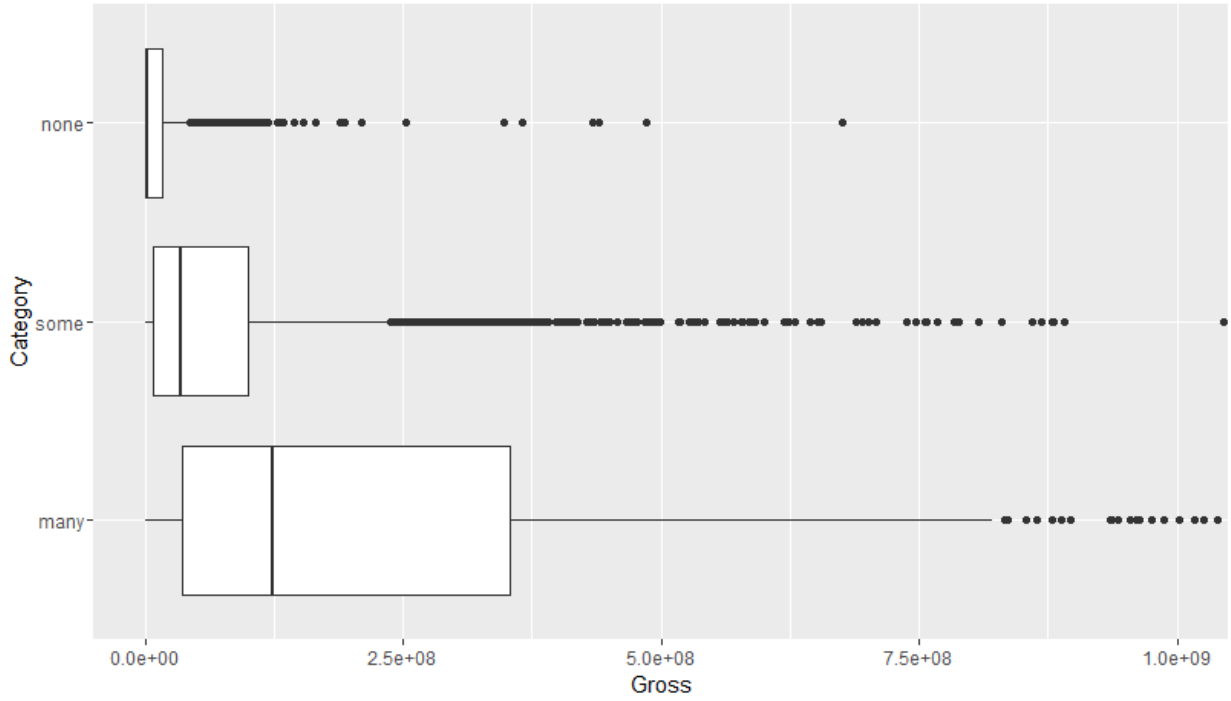
This function converted 0 award counts to "none", x >=1 & x<=upperThreshold to "some" and x > upperThreshold to "many"

This created a ratio of some/many of 5.71 with counts of

```
 many none some
1  561  772 3205
```

*How does the gross revenue distribution changes across these three categories.*

which




We see none has a very small distribution and a very low median. This makes sense since movies with 0 awards probably won't make many sales.

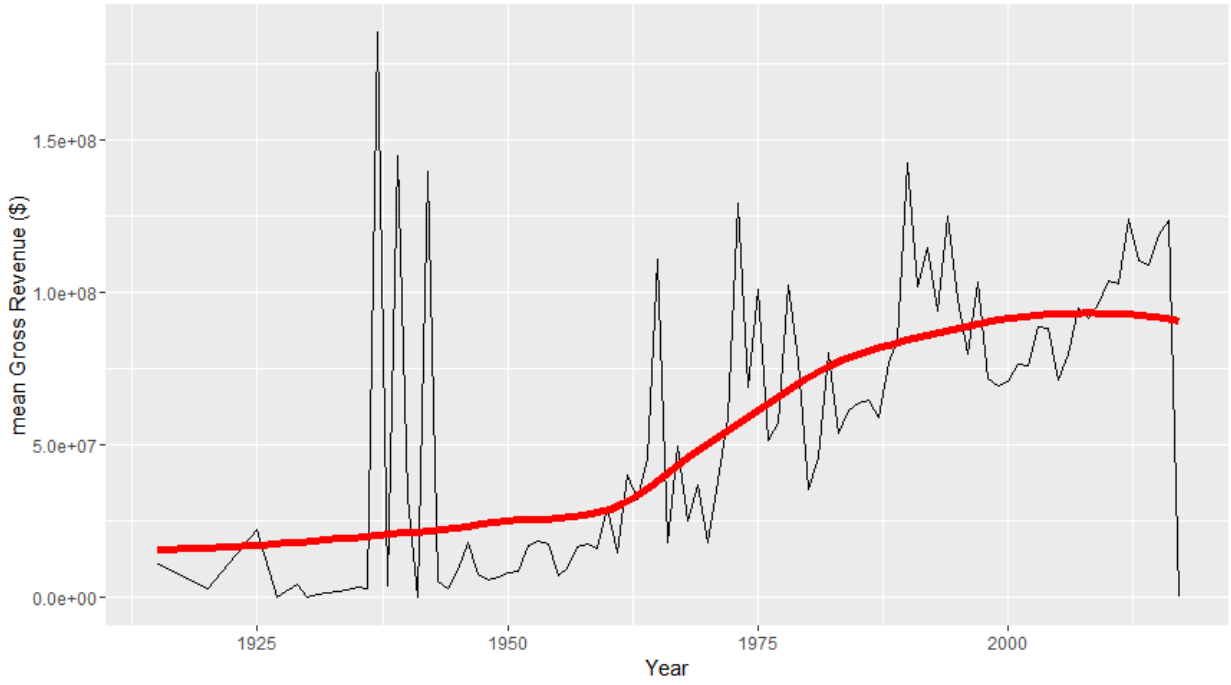
Some awards has a higher median and a greater distribution.

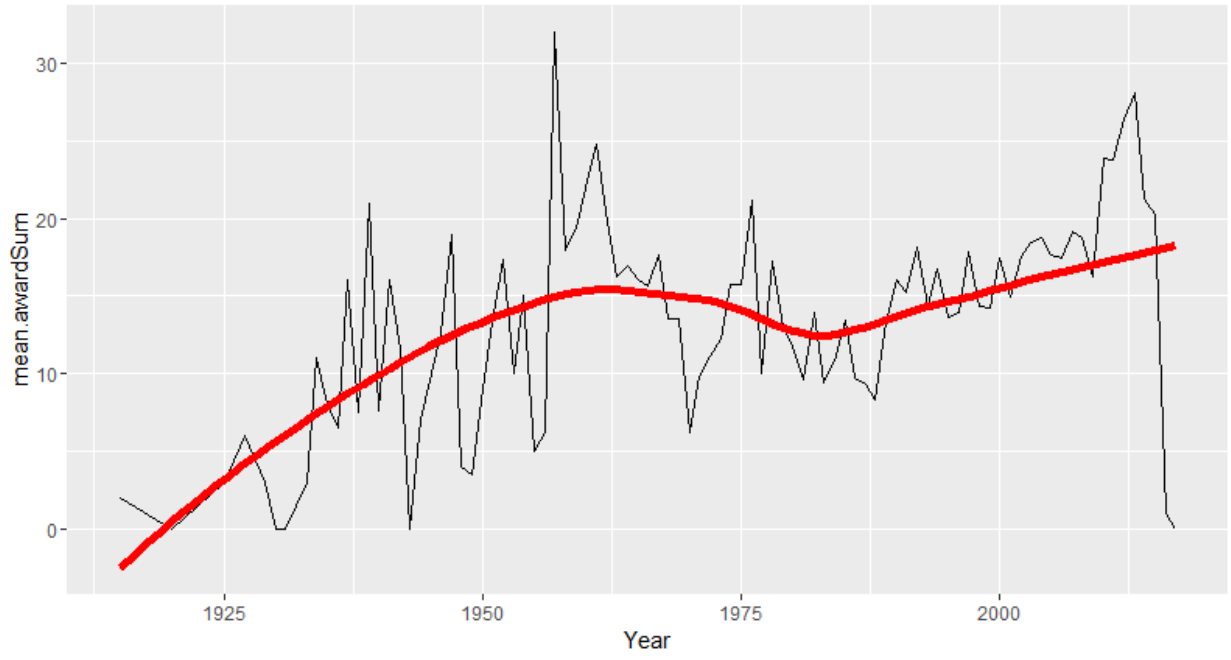
Many awards as a much higher median and a much greater distribution. This makes sense since there's a greater range of awards received in the many category which ranges from 38 to 548 awards than the some category which only ranges from 1 to 37 awards. I'm assuming that gross increases with the number of awards received (this will be further investigated in section 8 below).

*8. Come up with two new insights (backed up by the data and graphs) that are expected, and one new insight (backed up by data and graphs) that is unexpected at first glance and do your best to motivate it. By "new" here I mean insights that are not an immediate consequence of one of the above assignments.*

**Expected insight 1**

**Mean Gross Revenue per year**

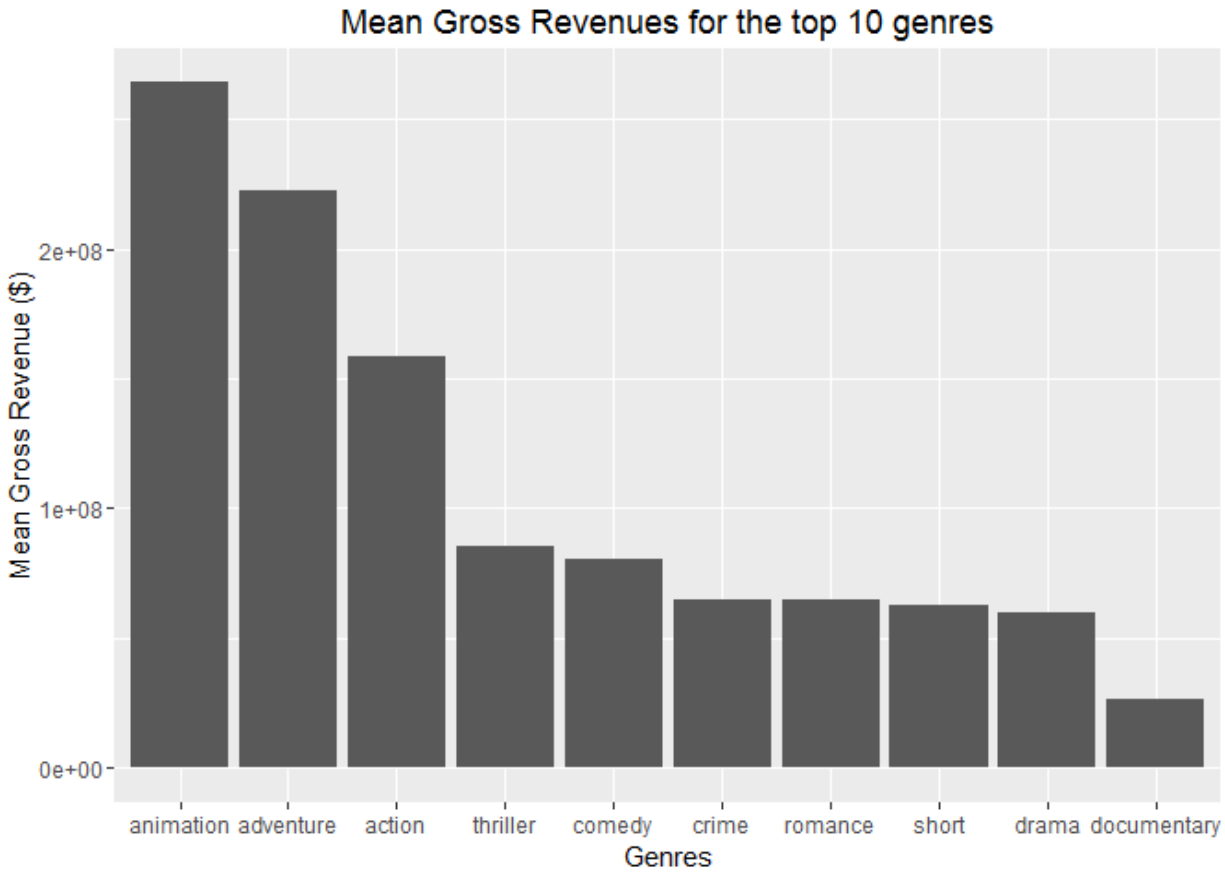


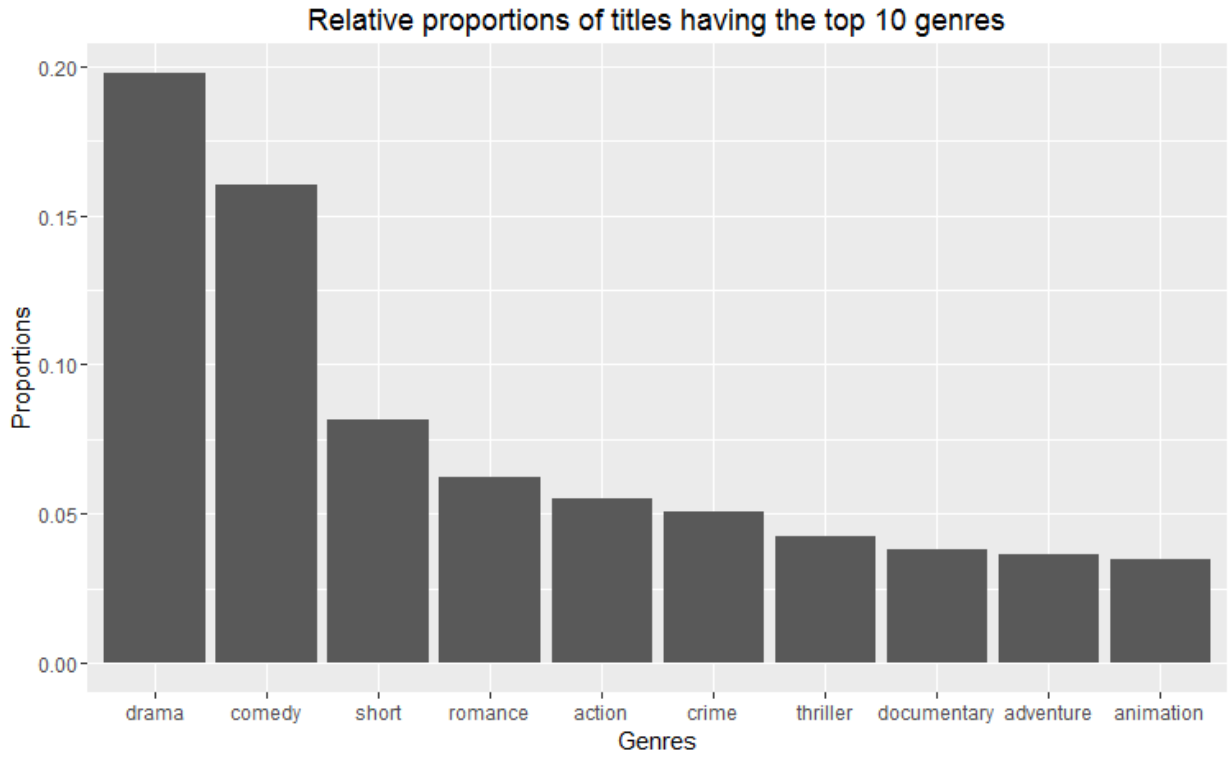




We see that Mean Gross revenue increases steadily from years 1900 till around 1960 and then sharply increases and then tapers off around 1990. This makes sense since there was less diposable income in earlier years to be spent on movies before 1960. Also perhaps movies really gained popularity then with color and improved technologies. It tapers off after around 1990 because disposable income and movie popularity didn't increase as much as the years increase.

Also we see a few high jumps certain years. Assuming the data is correct, this could be due to the fact that certain years had many best sellers.

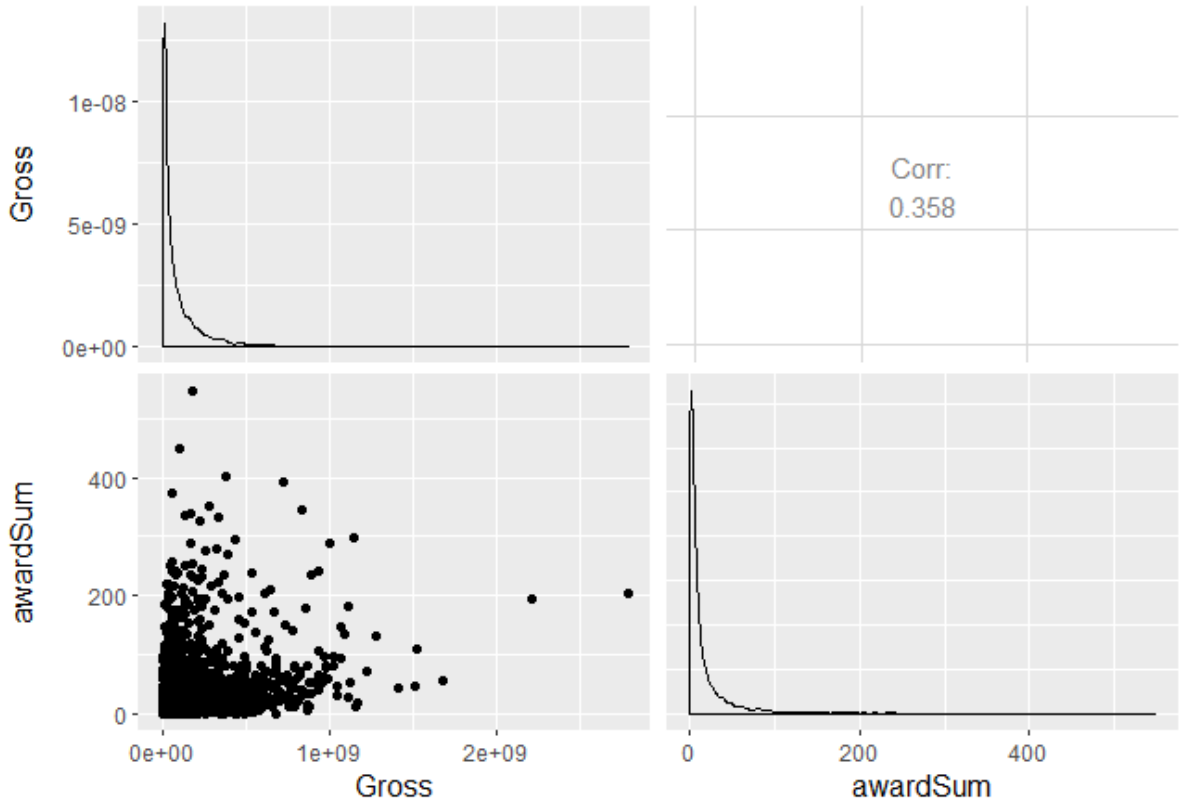
**Expected insight 2**






We see that even though animation and documentary has the smallest relative proportions of titles having the top 10 genres, their mean gross revenue is much greater than the rest. This was to be expected because more quantity doesn't imply more quality, usually it's the opposite. We also see for action, which has 3rd highest mean revenue, whereas it's much lower in proportion.

**Unexpected insight**




I would've expected Gross revenue to be positively correlated with the number of awards received, since better movies should receive more awards, and better movies should make more sales. We do find it to be positively correlated.

However it turns out it's only weakly positively correlated with correlation = 0.358. Perhaps this is because perhaps the award givers and viewers who want to see the movie are of different opinions. Perhaps they have different criteria of what makes a good movie.