PART I

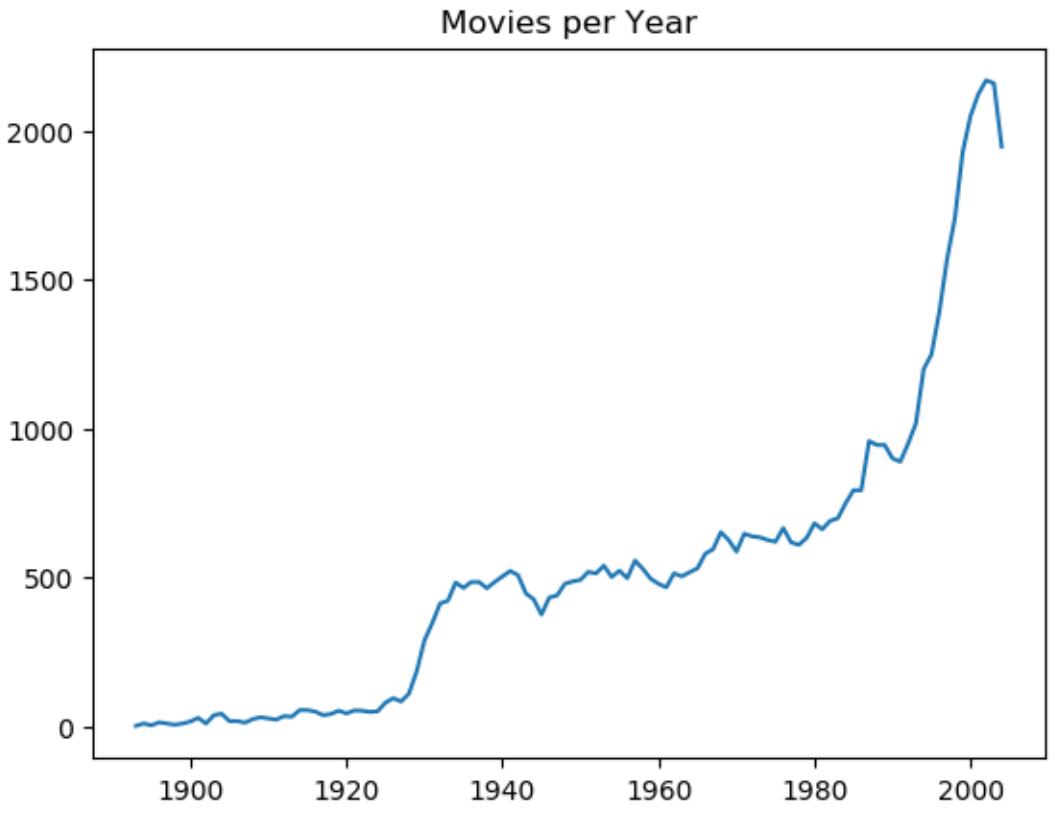
1.1 A few simple queries:

Mean IMDb Rating: 5.93

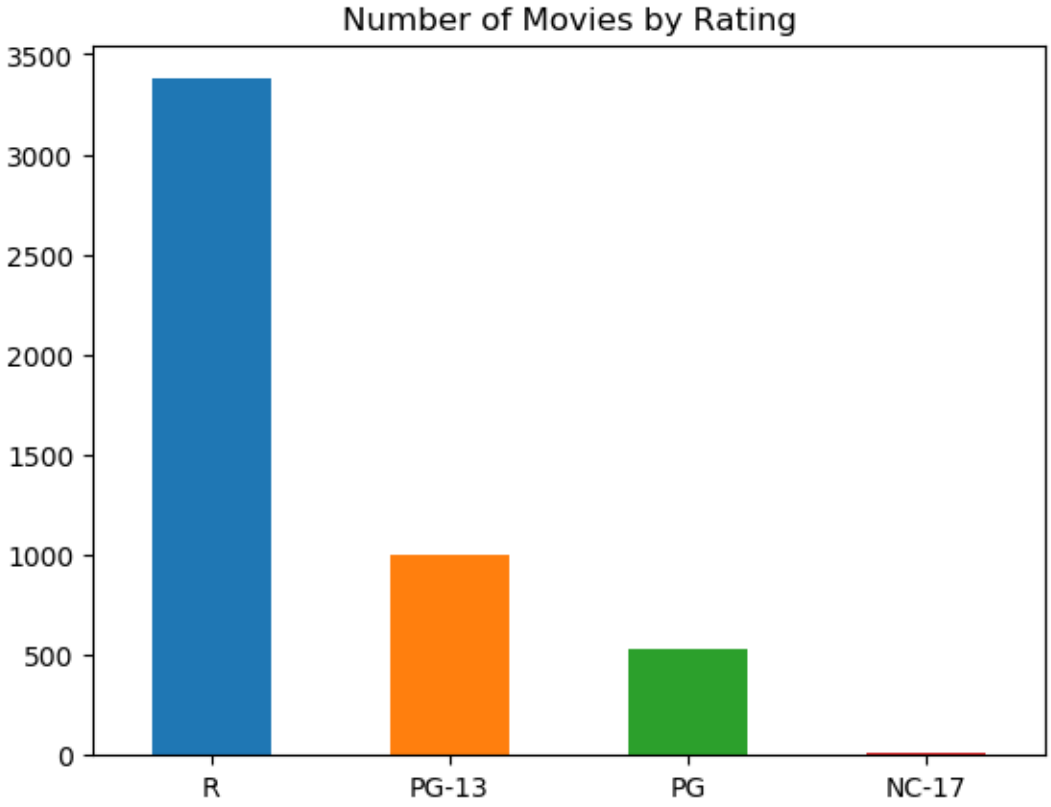
Median IMDb Rating: 6.10

First PG-13 Movie: 1945

1.2 Movies released per year:



1.3 Number of movies per MPAA rating:



PART II

2.1 What proportion of each genre is rated R?

Action movies rated 'R': 644/4688 = 13.74%

Animation movies rated 'R': 11/3690 = 0.30%

Comedy movies rated 'R': 916/17271 = 5.30%

Drama movies rated 'R': 1723/21811 = 7.90%

Documentary movies rated 'R': 67/3472 = 1.93%

Romance movies rated 'R': 440/4744 = 9.27%

Short movies rated 'R': 6/9458 = 0.06%

The above proportions are from all movies, including the ones that have not been rated.

Below are the proportions found after removing the unrated movies:

Action movies rated 'R': 644/942 = 68.37%

Animation movies rated 'R': 11/67 = 16.42%

Comedy movies rated 'R': 916/1662 = 55.11%

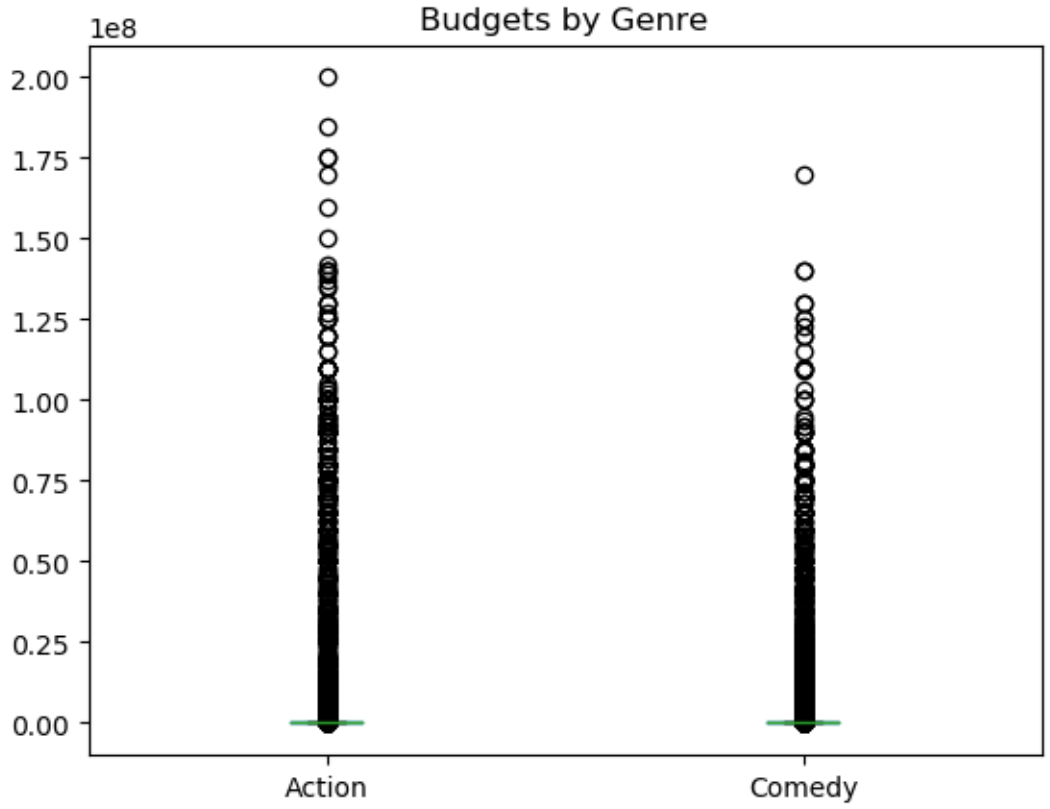
Drama movies rated 'R': 1723/2393 = 72.00%

Documentary movies rated 'R': 67/121 = 55.37%

Romance movies rated 'R': 440/754 = 58.36%

Short movies rated 'R': 6/16 = 37.50%

2.2 Comparing the costs of comedies and action movies:



The numbers on the axis are in $100,000,000.

Here is a summary of the values since the chart is hard to read:

Budgets of comedies

count	58788.000000
mean	436895.098285
std	4585209.102229
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	170000000.000000

Budgets of action movies

count	58788.000000
mean	440941.208206
std	5657433.221485
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	200000000.000000

I noticed that the 0 values are interfering. I will exclude them:

Budgets of comedies, without null or 0

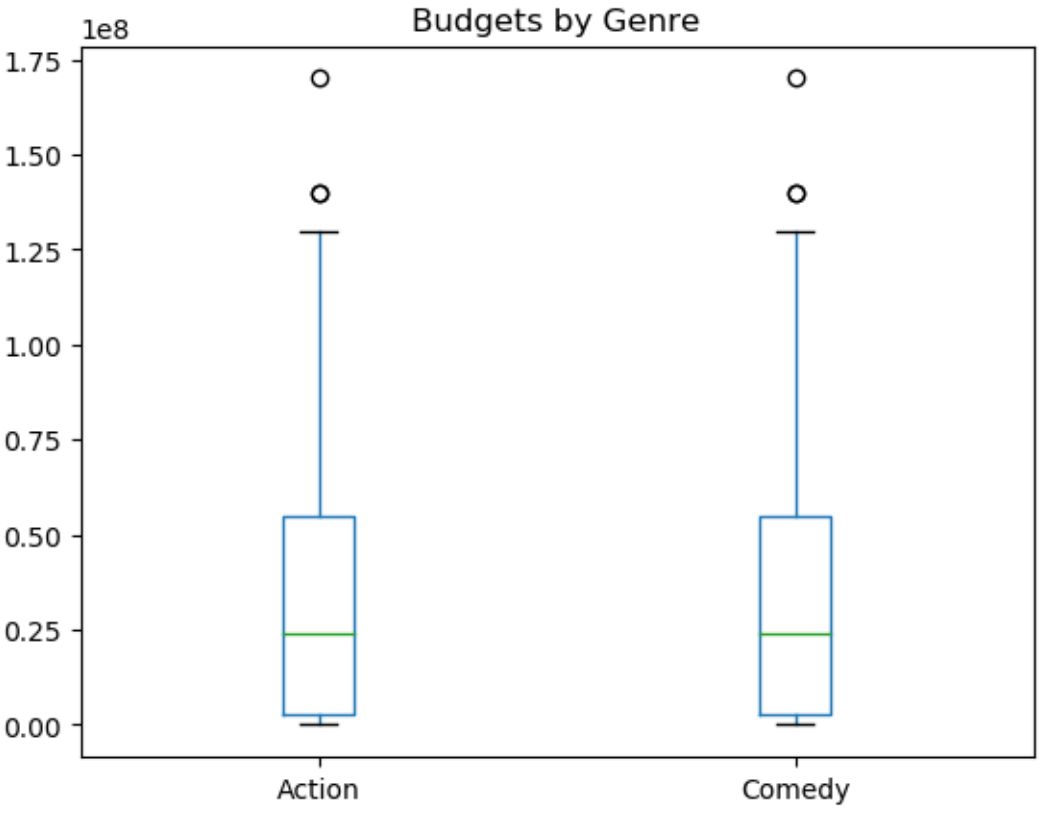
count	1748.000000
mean	14693471.989703
std	22312984.787983
min	1000.000000
25%	500000.000000
50%	4000000.000000
75%	20000000.000000

max	170000000.000000

Budgets of action movies, without null or 0

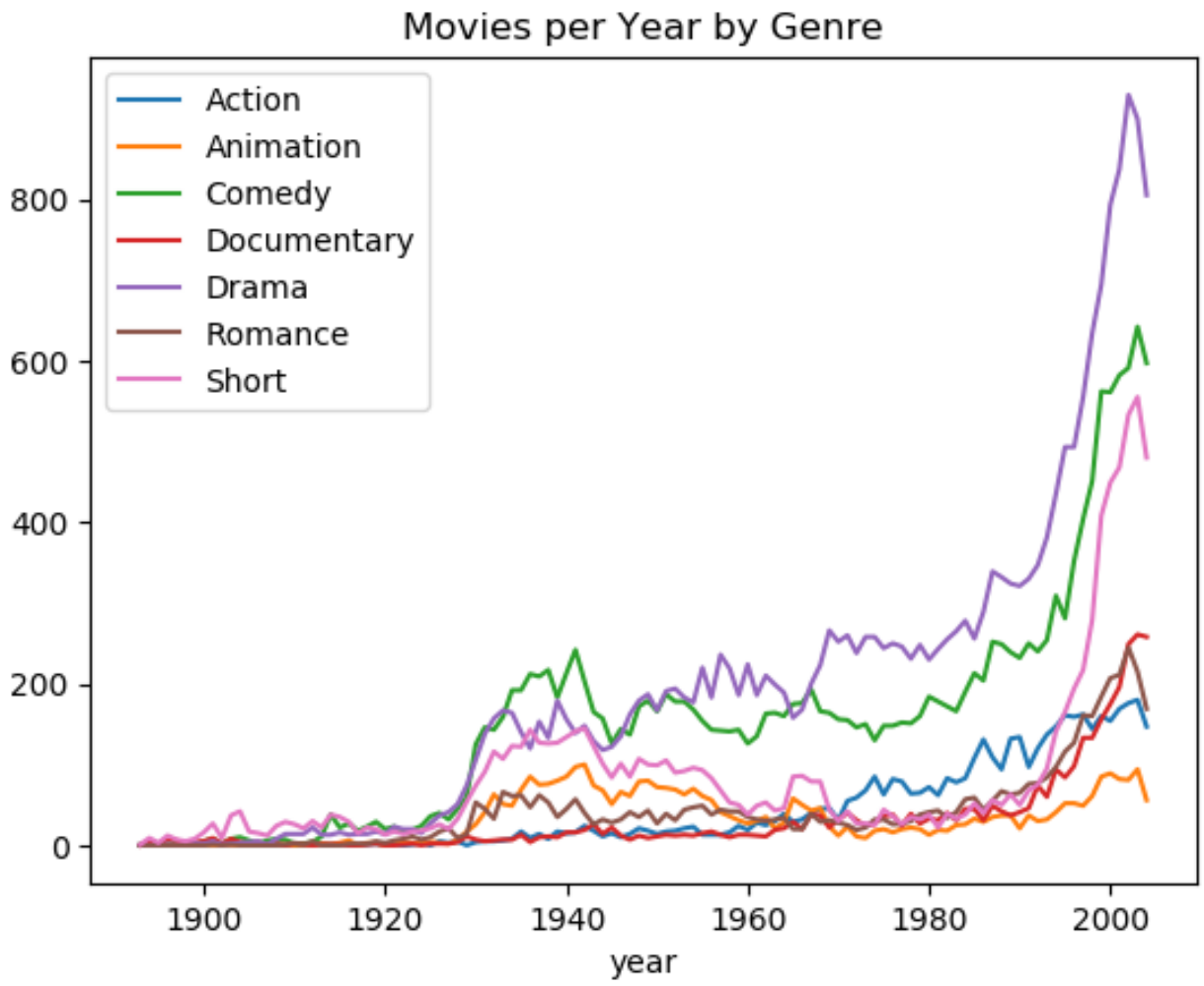
count	834.000000
mean	31081596.820144
std	36128695.419026
min	1000.000000
25%	2500000.000000
50%	16000000.000000
75%	50000000.000000
max	200000000.000000

The box plot is much easier to read now too:



There is not a significant difference.

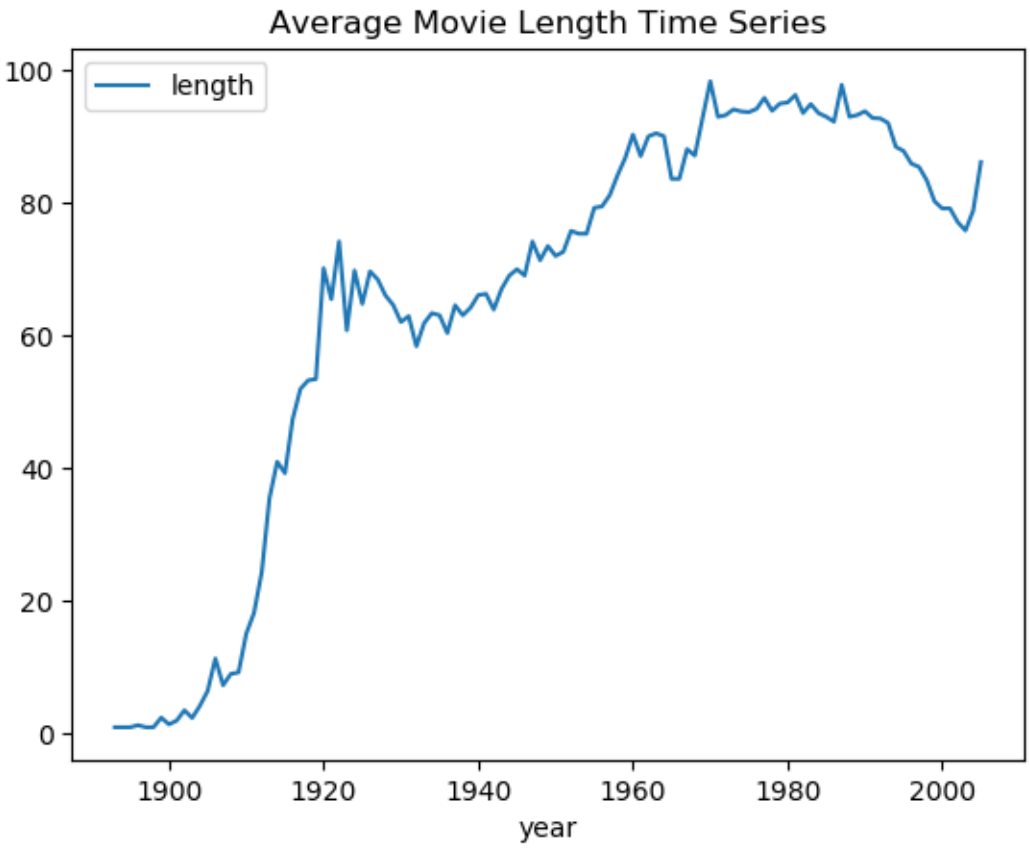
2.3 Genre popularity by year:



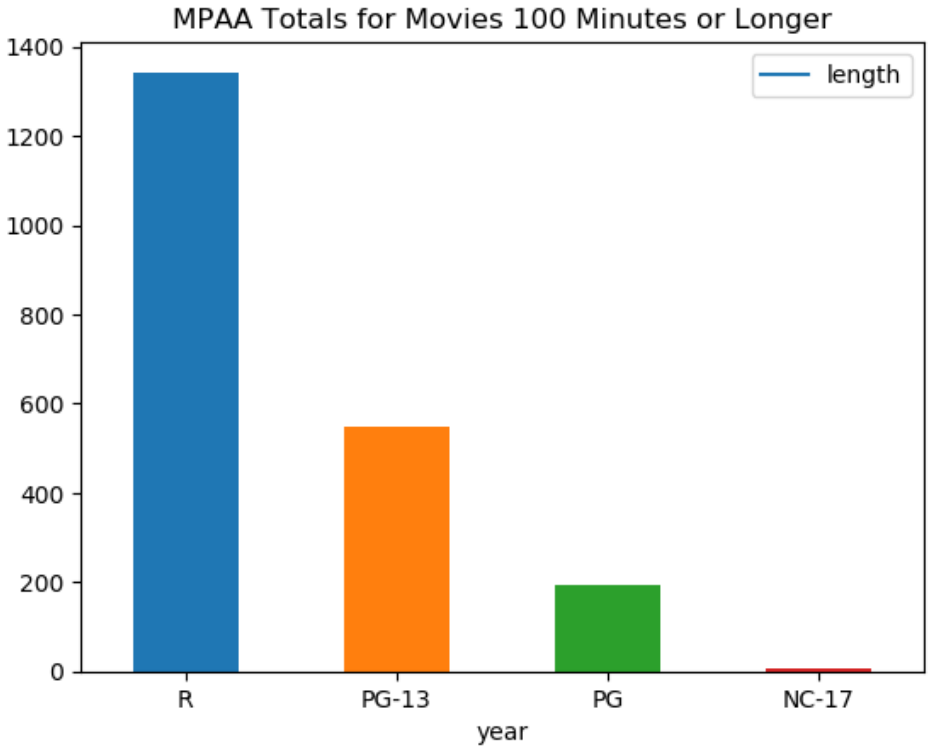
This is my favorite one, aesthetically. Meaningful observations: Short films became popular in the 1990s and remained that way. Comedies were the most popular shortly after the great depression. The numbers of action movies and documentaries were on the most steady trends.

PART III

3.1 Things about movie length:

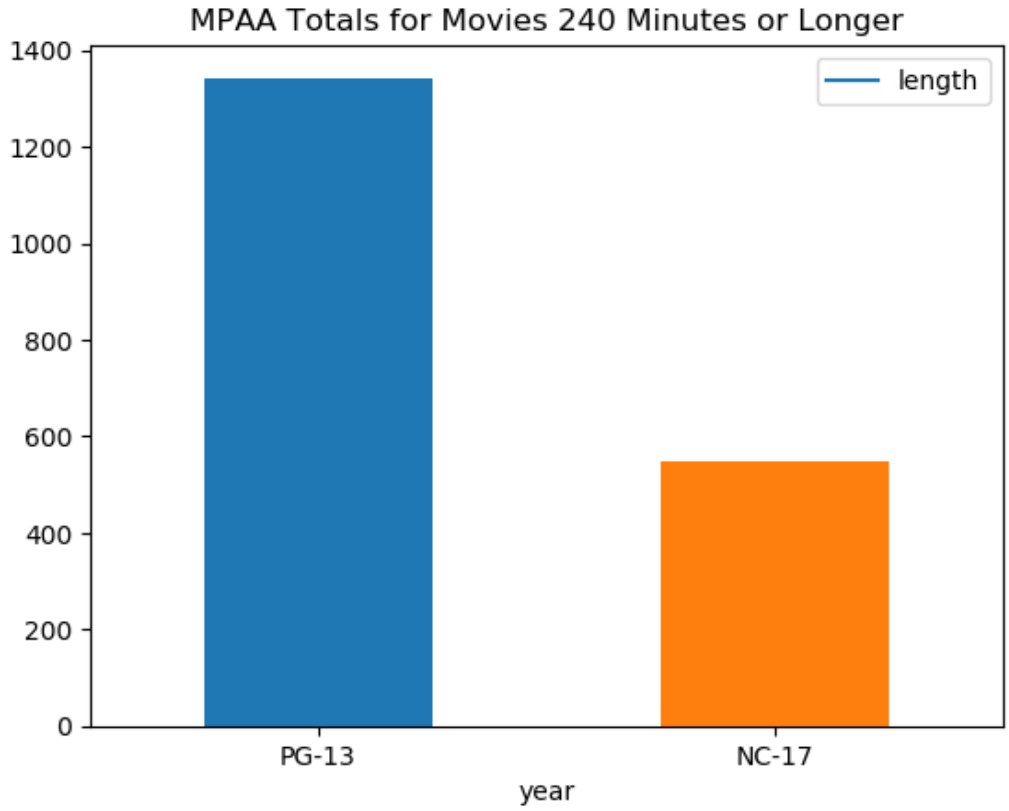


The first movies were short.



The above plot mostly just shows that there are a lot of R rated films. The ratio of PG-13 movies is slightly increased from the selection of all movies as seen in the plot in 1.3.

In the below plot, the PG-13 and NC-17 rated movies are the only ones that appear to have hit the 240 minute mark. 4 hours is a very long movie.



Of course, we are curious:

Shortest “movie” of all time: 1 minute

Longest “movie” of all time: 87 hours, 0 minutes

3.2 Length and IMDb scores:

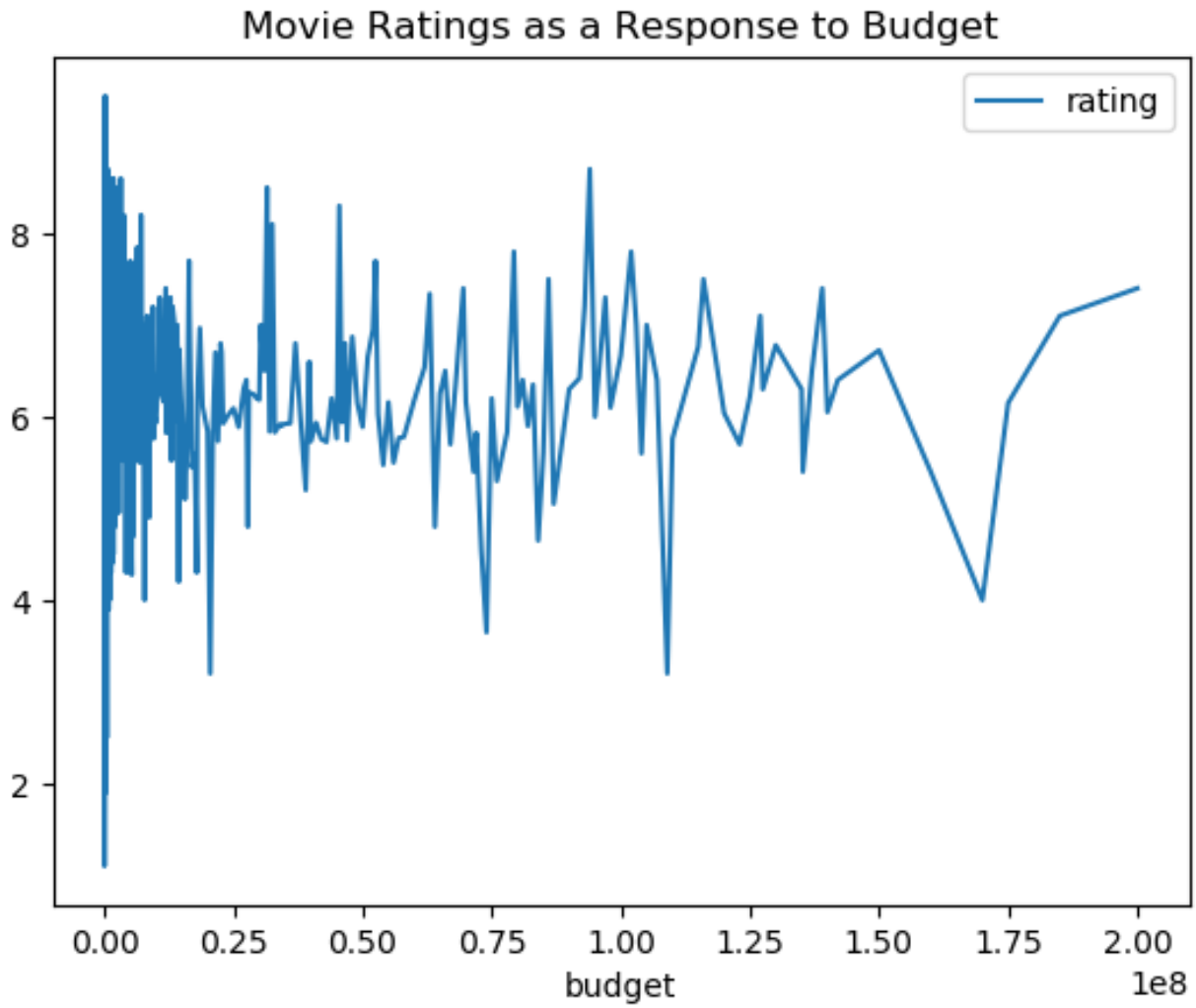
Lengths of Movies with a Perfect Score (10.0)

count	3.000000
mean	15.666667
std	12.503333
min	7.000000
25%	8.500000
50%	10.000000
75%	20.000000
max	30.000000

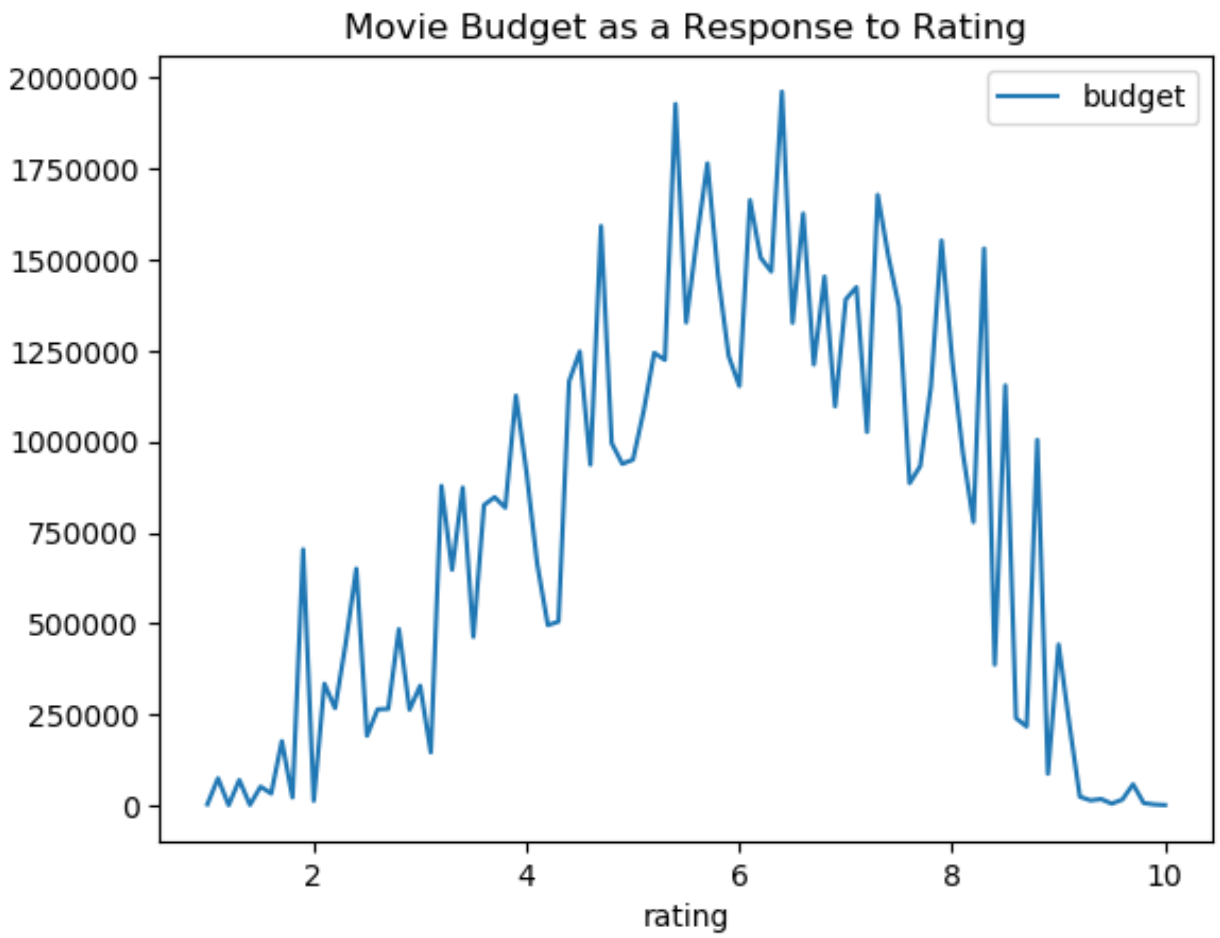
Lengths of Movies with the Lowest Score Possible (1.0)

count	106.000000
mean	64.915094
std	32.047322
min	3.000000
25%	43.000000
50%	70.000000
75%	86.750000
max	150.000000

3.3 How budget and IMDb score are correlated:



This is not super insightful, but this shows there are many low budget films.



Now that the graph has been transposed it is more obvious: There is a relationship between budget and score, but the highest rated movies are not the most expensive. This graph is my favorite for what it reveals.