# HW week 12

w203: Statistics for Data Science

w203 teaching team

```
library(tidyverse)
library(ggplot2)

library(sandwich)
library(stargazer)
```

```
d <- load_and_clean(input = 'videos.txt')
```

```
## Rows: 9618 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): video_id, uploader, category
## dbl (6): age, length, views, rate, ratings, comments
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

## Regression analysis of YouTube dataset

You want to explain how much the quality of a video affects the number of views it receives on social media. In a world where people can now buy followers and likes, would such an investment increase the number of views that their content receives? **This is a causal question.**

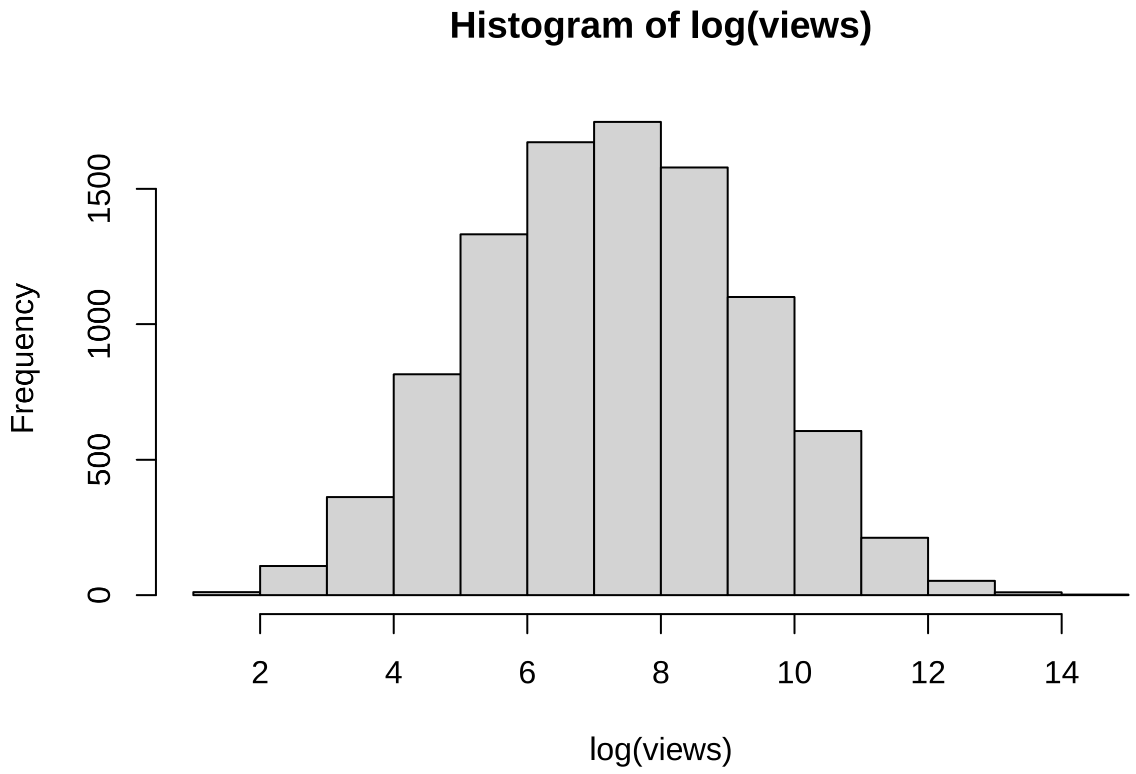
You will use a dataset created by Cheng, Dale and Liu at Simon Fraser University. It includes observations about 9618 videos shared on YouTube. Please see this link for details about how the data was collected.

You will use the following variables:

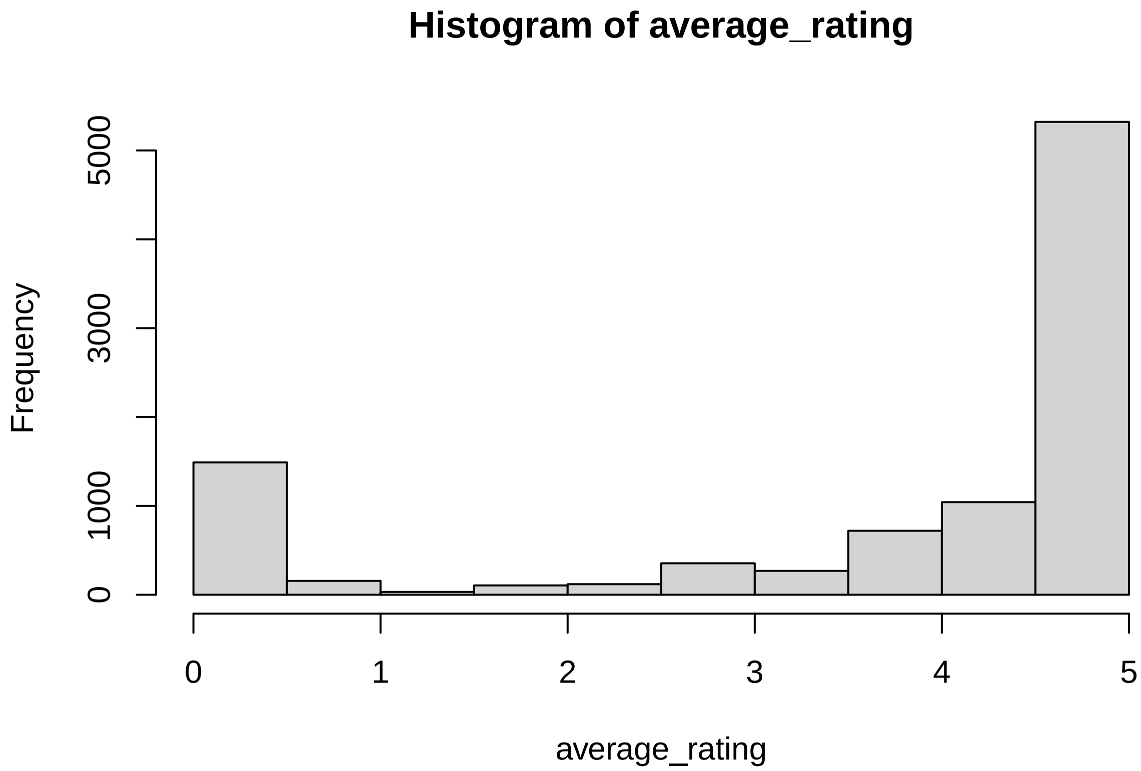
- `views`: the number of views by YouTube users.
- `average_rating`: This is the average of the ratings that the video received, it is a renamed feature from `rate` that is provided in the original dataset. (Notice that this is different from `cout_of_ratings` which is a count of the total number of ratings that a video has received.
- `length:` the duration of the video in seconds.

a. Perform a brief exploratory data analysis on the data to discover patterns, outliers, or wrong data entries and summarize your findings.

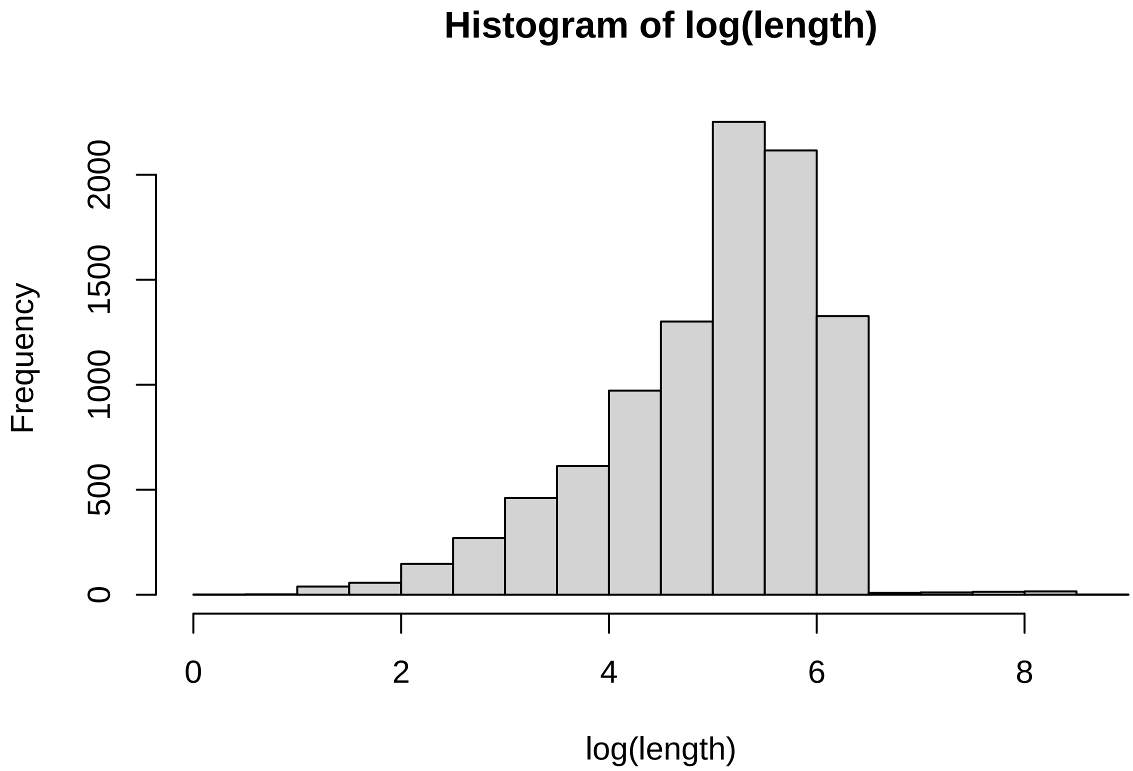
```
attach(d)
hist(log(views))
```



```
hist(average_rating)
```



```
hist(log(length))
```



> ‘What did you learn from your EDA? Cut this quoted text and describe your analysis in the quote block.’

b. Based on your EDA, select an appropriate variable transformation (if any) to apply to each of your three variables. You will fit a model of the type,

$$f(\text{views}) = \beta_0 + \beta_1 g(\text{rate}) + \beta_3 h(\text{length})$$

Where $f$, $g$ and $h$ are sensible transformations, which might include making *no* transformation.

```
model <- lm(log(views) ~ average_rating + log(length), data=d)

stargazer(model, type = 'text', se = list(get_robust_se(model)))
```

```
## 
## ===============================================
##                         Dependent variable:    
##                     ---------------------------
##                              log(views)        
## -----------------------------------------------
## average_rating               0.467***          
##                               (0.010)          
##                                                
## log(length)                  0.105***          
##                               (0.018)          
##                                                
## Constant                     5.010***          
##                                (0.088)
##
## -----------------------------------------------
## Observations                    9,609
## R2                              0.189
## Adjusted R2                     0.189
## Residual Std. Error      1.799 (df = 9606)
## F Statistic         1,121.899*** (df = 2; 9606)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

c. Using diagnostic plots, background knowledge, and statistical tests, assess all five assumptions of the CLM. When an assumption is violated, state what response you will take. As part of this process, you should decide what transformation (if any) to apply to each variable. Iterate against your model until your satisfied that at least four of the five assumption have been reasonably addressed.

1. **IID Data:** The crawler uses BFS, so all videos are not completely independent. They're probably related in topic. Next, they use BFS from the "Most viewed" or some similar popular videos list. All videos that end up here are probably on the high end of views, no video from a 1000 subscriber Youtuber that gets 5000 views per video. Not identically distributed.
2. **No Perfect Colinearity:** No perfect colinearity. In fact, very low correlation between the two input variables
3. **Linear Conditional Expectation:** Yes: look at residuals plot from plot(model).

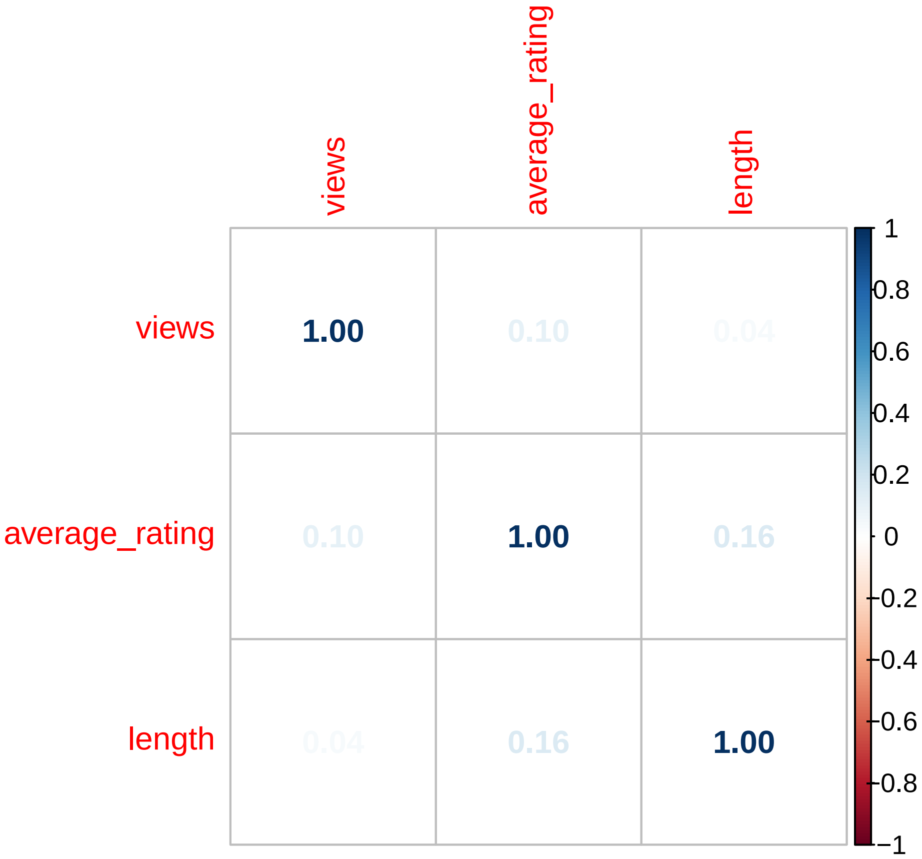
```
## remove this commented block and write code that can help you assess whether
## your model satisfied the requirement of a linear conditional expectation.

library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

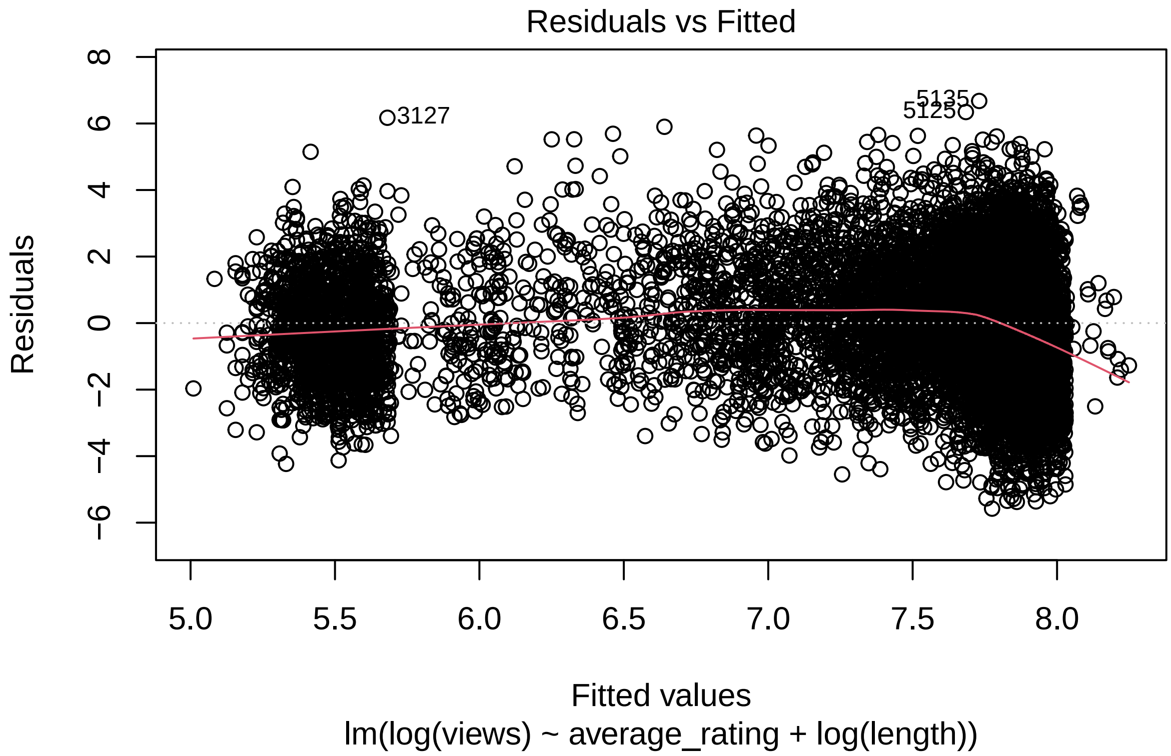
```
## corrplot 0.92 loaded
```

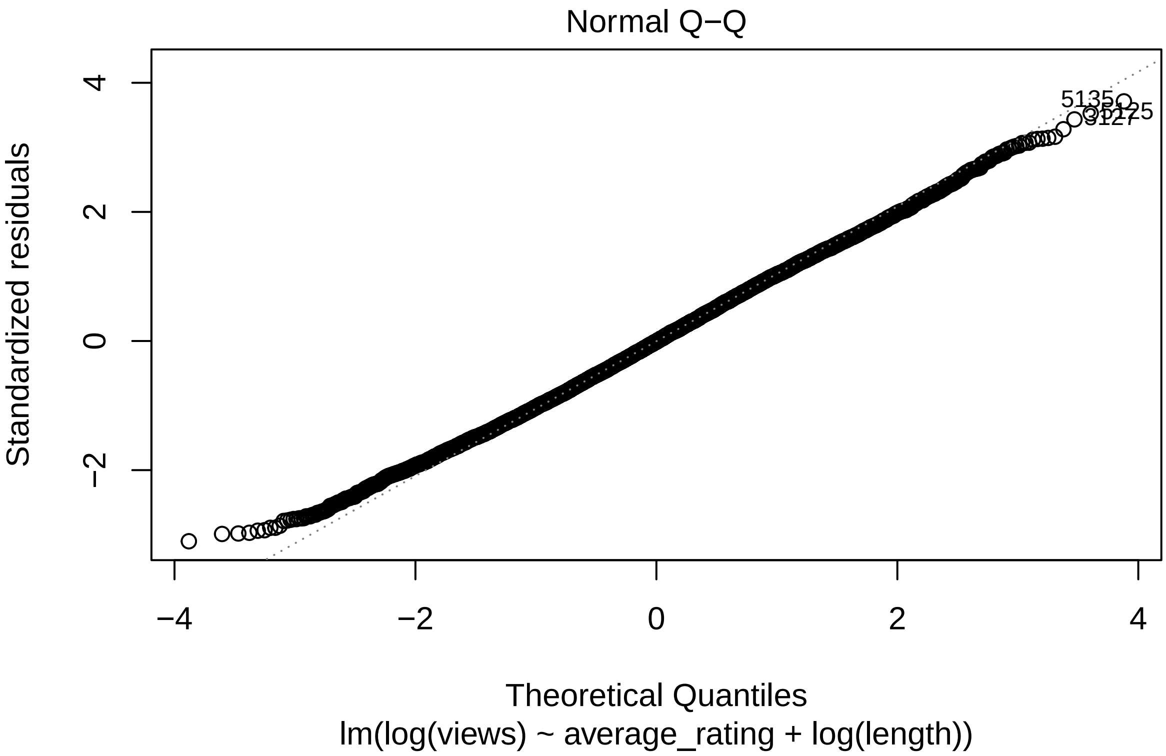
```
corrplot(abs(cor(d[c("views", "average_rating", "length")], use="pairwise.complete.obs")), method = 'nu
```

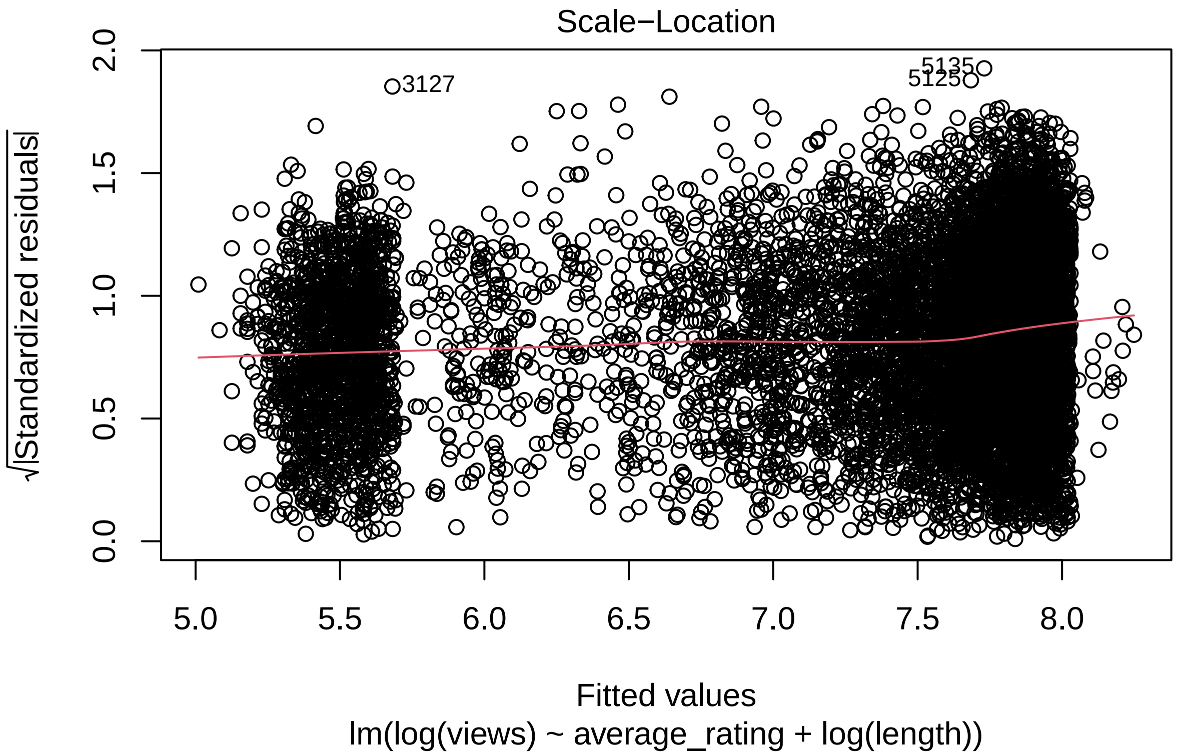


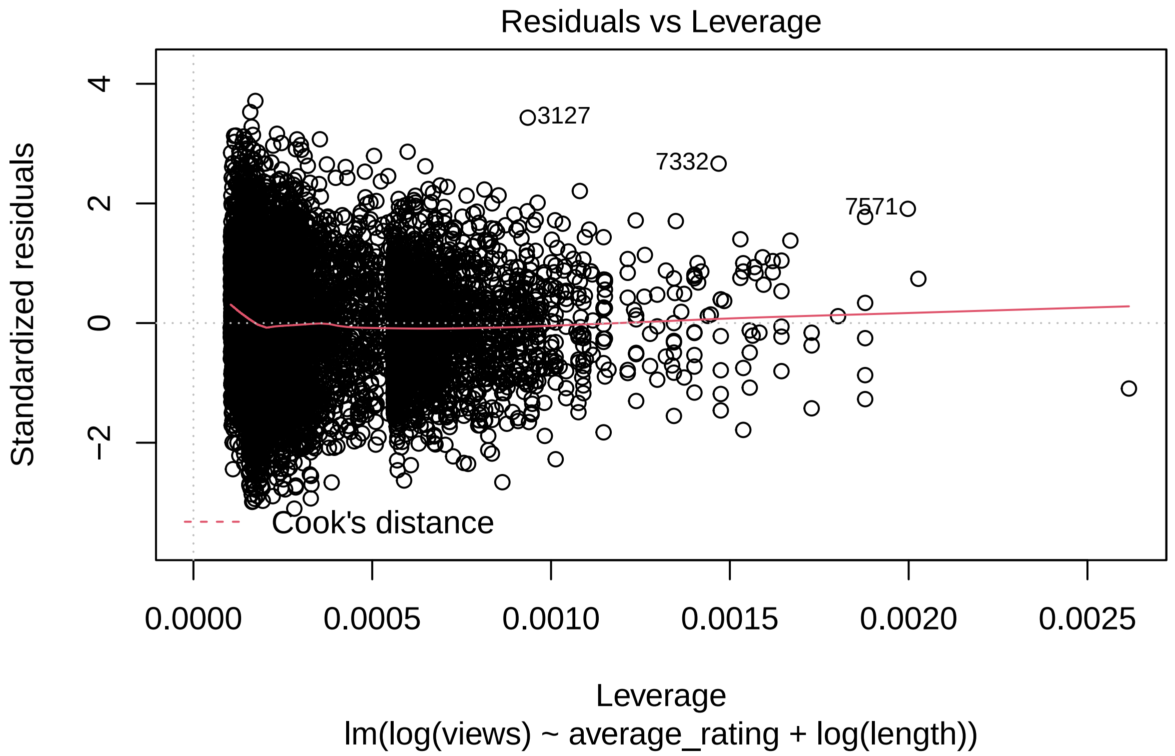
4. **Homoskedastic Errors:** I think decently homoskedastic

```
## remove this commented block and write code that can help you assess whether
## your model satisfied the requirement of homoskedastic errors
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(views) ~ average_rating + log(length))
3127
5135
5125
8
6
4
2
0
−2
−4
−6
5.0
5.5
6.0
6.5
7.0
7.5
8.0


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−2
0
2
4
5135
5125
3127
Theoretical Quantiles
lm(log(views) ~ average_rating + log(length))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
3127
5135
5125
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Fitted values
lm(log(views) ~ average_rating + log(length))



5. **Normally Distributed Errors:** Very normal errors. Look at QQ plot

```
## remove this commented block and write code that can help you assess whether
## your model satisfied the requirement of normally distributed errors
```