HW week 12

w203: Statistics for Data Science

w203 teaching team

More 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. **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.
- rate: the average rating given by users.
- 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.

ANSWER:

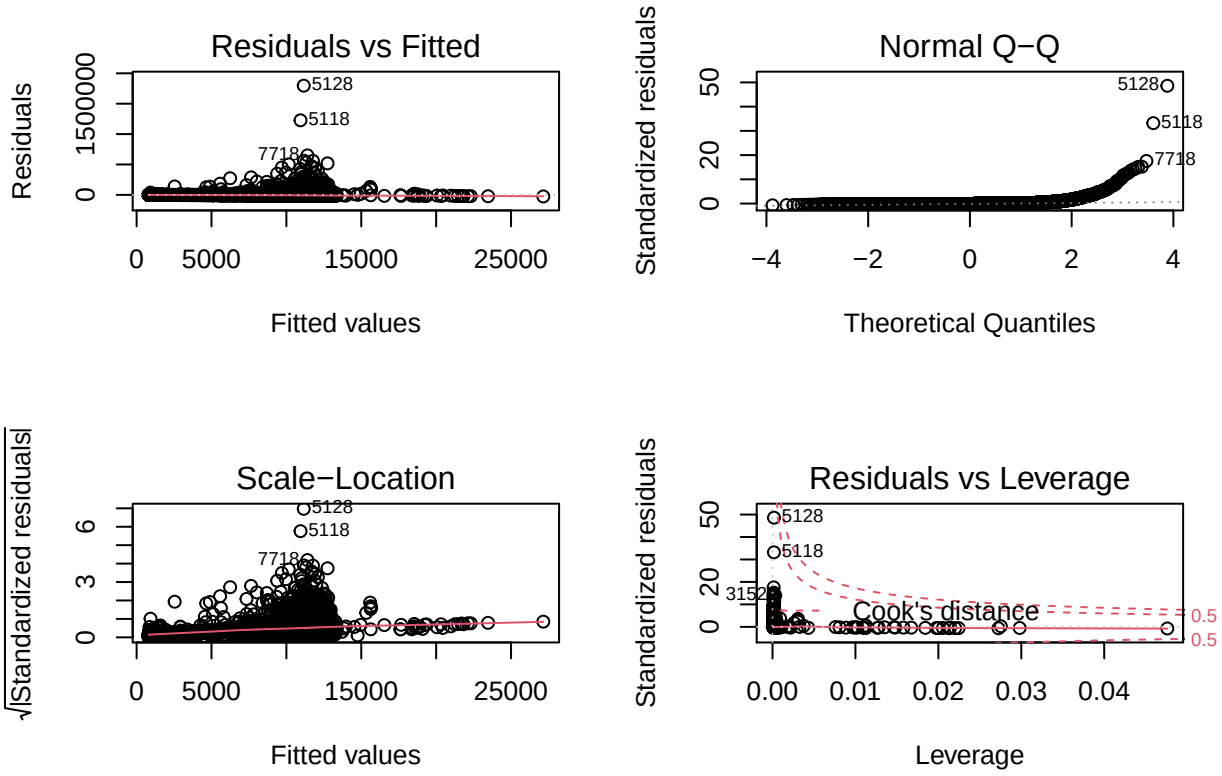
I firstly imported data from the csv file.

```
## [1] 9617
```

```
## [1] 9609
```

```
##       rate           views             length
##  Min.   :0.000   Min.   :      3   Min.   :   1
##  1st Qu.:3.400   1st Qu.:    348   1st Qu.:  83
##  Median :4.670   Median :   1453   Median : 193
##  Mean   :3.744   Mean   :   9346   Mean   : 227
##  3rd Qu.:5.000   3rd Qu.:   6179   3rd Qu.: 299
##  Max.   :5.000   Max.   :1807640   Max.   :5289
```

```
fit <- lm(df$views ~ df$rate + df$length)
par(mfrow=c(2,2))
plot (fit)
```



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 $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$ are sensible transformations.

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 tranformation (if any) to apply to each variable.