HW 12 CLM

Nick, Blake, Peter, Reese

```
model1 <- lm(log(views) ~ rate + length, data = data)
```

- Based on some exploratory data analysis, it was discovered that the `Views` histogram was highly skewed to the right, but a histogram of `log(Views)` was relatively normally distributed. Therefore, `log(Views)` was selected as the model dependent variable.
- Additionally it was discovered that a value of 0 in the `rates` variable represents a view in which no ratings were ever selected. If left alone, this will skew the model since these values will be taken as real ratings. Therefore, all samples with a value of 0 in the `rates` column were changed to NaN values to remove this issue.

#Question 1.1

To assess IID data, we need to know about the sampling process.From the Dataset for "Statistics and Social Network of YouTube Videos" documentation, the data contains 9618 observations of videos performed by a crawler extracting information from the Youtube API. Based on this, there are several reasons the data is NOT independent of each other.

- The Youtube videos form a direct graph, where each video is a node in the graph.The edges of the graph are based on if the videos are related to each other within 20 videos.The data of the video is collected through this graph, which means that each video all has some relation to another, showing that they are NOT independent.
- Youtube videos that are related are likely inspired or based on similar topics which result in non-independence (clustering).
- A particular youtuber might make videos that are similar to each other (clustering).
- Popular videos will inspire other videomakers to make similar videos (strategic effect).

#Question 1.2

To assess perfect collinearity, we can look at our coefficients, and notice that R has not dropped any variables. We can also look at the Variance Inflation Factor of each variable.

```
model1$coefficients
```

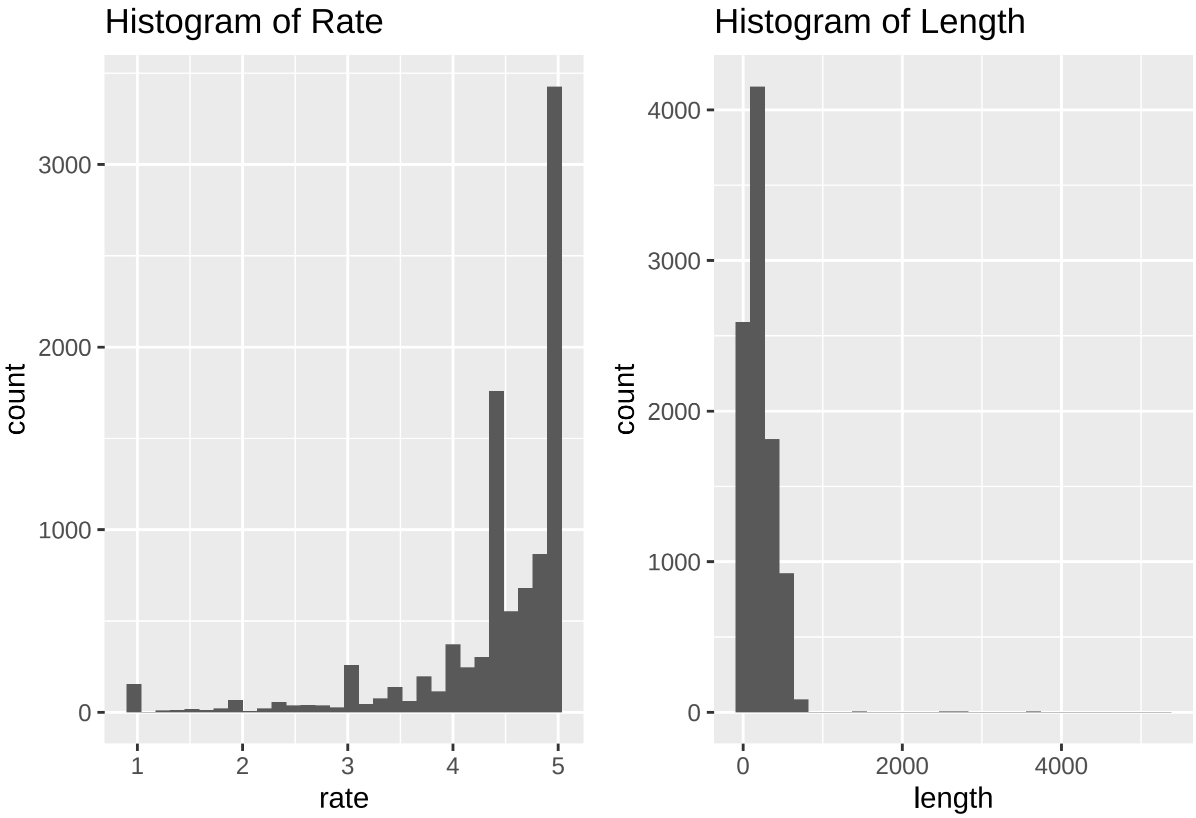
```
##  (Intercept)          rate       length 
## 6.3648152460 0.1583839523 0.0009670822
```

```
vif(model1)
```

```
##     rate   length 
## 1.011471 1.011471
```

- Since no coefficients were dropped, it suggests no perfect colinearity.
- The somewhat low VIF scores of each variable show that they are barely correlated, which also indicates no near-perfect colinearity.
- This assumption also requires that a unique BLP exists. Looking at the distributions of the variables below, there appear to be moderate skew in the `rate` values and the `length` values.However, neither skew is not to the extent where we need to be worried about extremely heavy tails.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



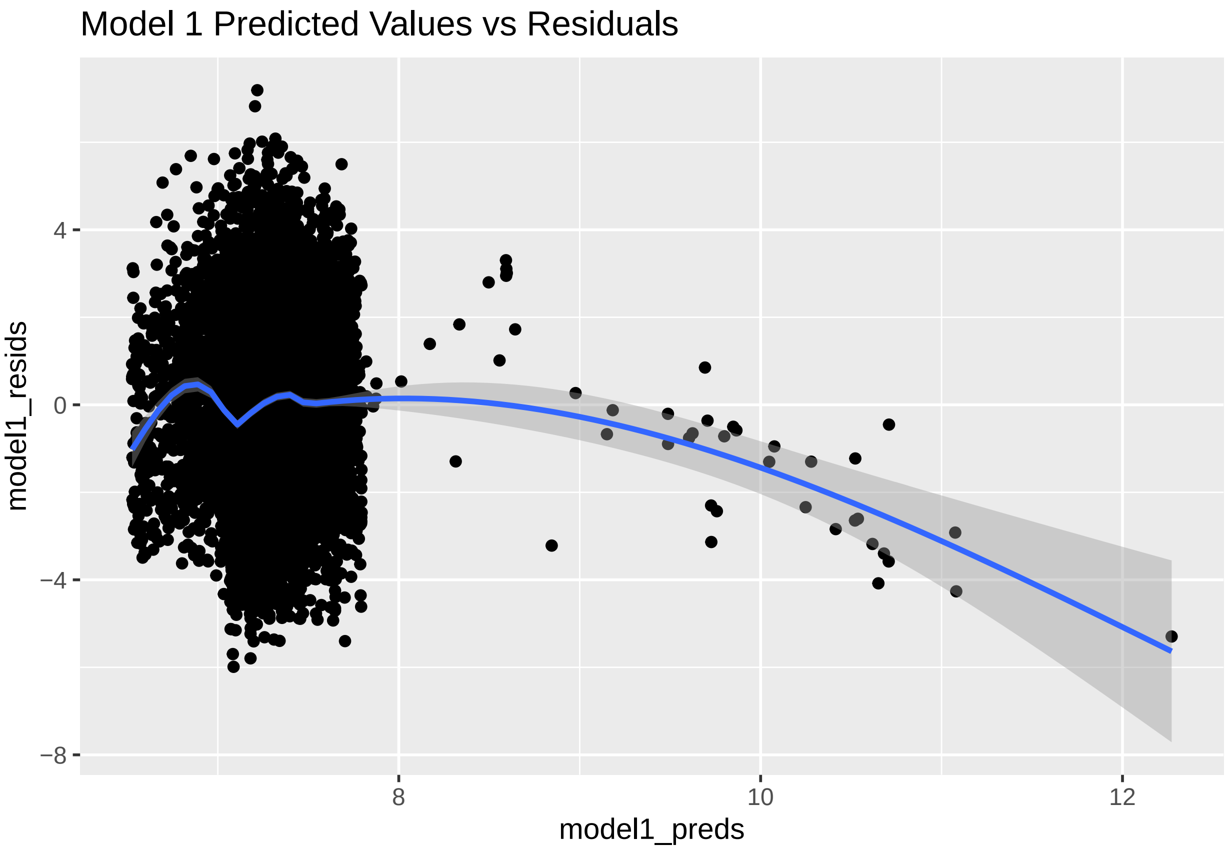
#Question 1.3

- To assess whether there is a linear conditional expectation, we've learned to look at the predicted vs. residuals of the model.

```
data <- data %>%
  mutate(
    model1_preds = predict(model1),
    model1_resids = resid(model1))

data %>% ggplot(aes(model1_preds, model1_resids)) +
  geom_point() +
  stat_smooth() +
  ggtitle("Model 1 Predicted Values vs Residuals")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

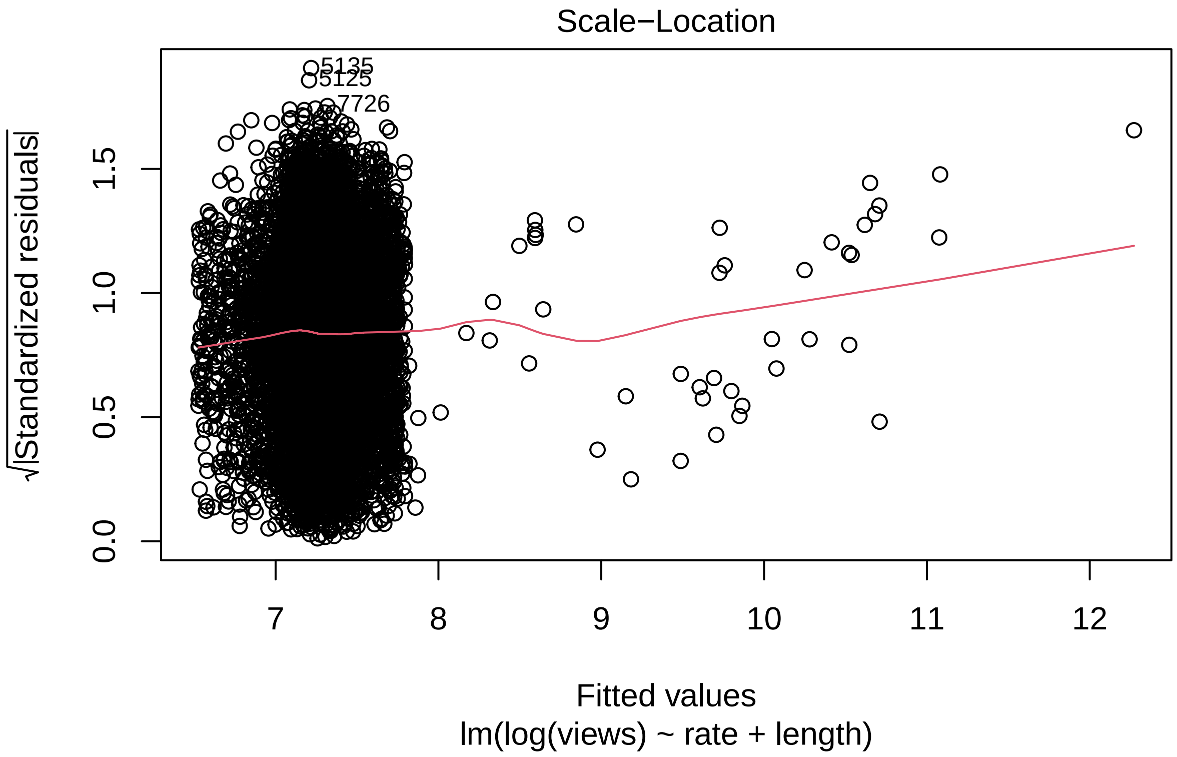


- The results show a relatively flat correlation between the prediction and the residuals, indicating linear conditional expectation.There is a non-linear section above a model prediction value of >~8, but this may be pulling the trend downward due to the high leverage of these outlier data points and may not reflect the true nature if more samples were taken in this are of the distribution.

#Question 1.4

- To assess if the model has homoskedastic errors, we can examine the residuals versus fitted plot from the scale-location plot. Homoskedasticity would look like a flat smoothing curve here. The curve is relatively flat and smooth aside from the fitted values >~8.
- However, running a Breusch-Pagan test indicates that the samples are statistically heteroskedastic.However, this is may be caused by low number of videos that performed very well (predicted y-value >~8). Therefore we can reasonable conclude that the heteroskedasticity in not severe.

```r
plot(model1, which = 3)
```



```r
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 11.067, df = 2, p-value = 0.003952
```

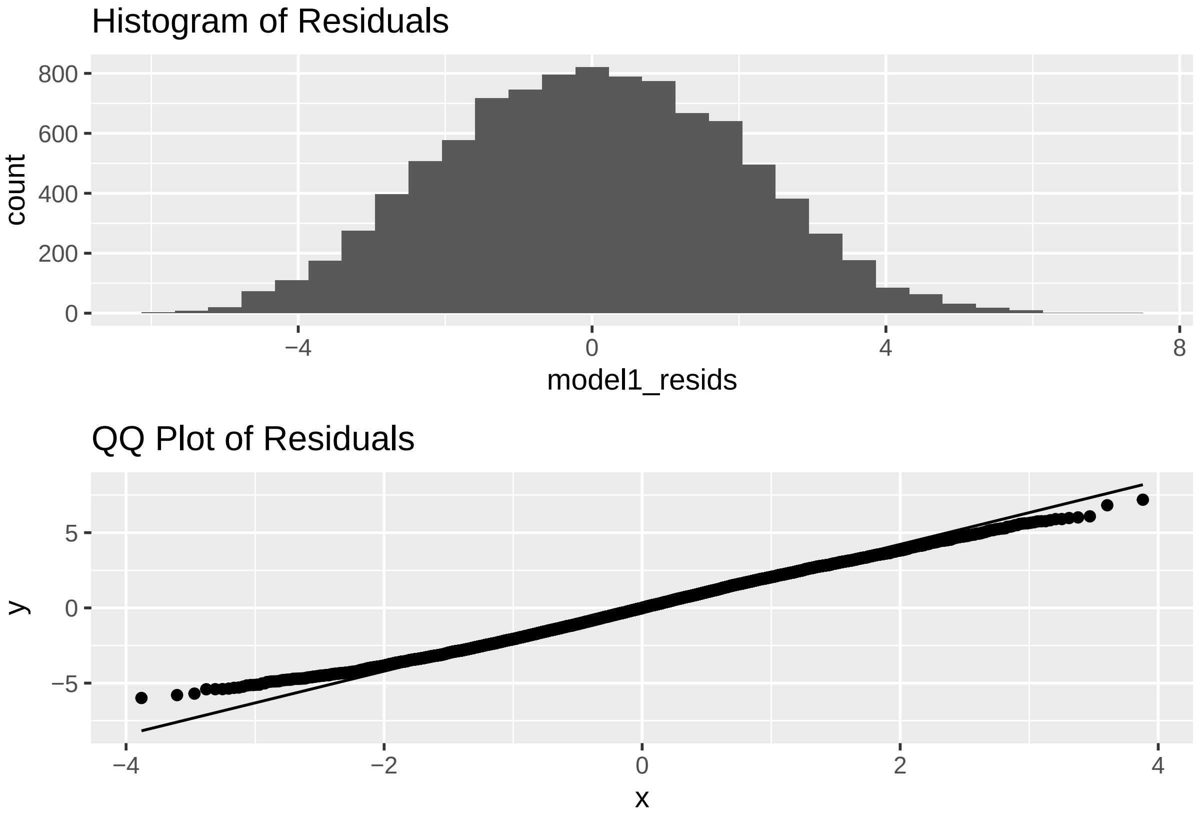
#Question 1.5

```
plot1 <- data %>%
  ggplot(aes(x = model1_resids)) +
  geom_histogram() +
  ggtitle("Histogram of Residuals")

plot2 <- data %>%
  ggplot(aes(sample = model1_resids)) +
  stat_qq() + stat_qq_line() +
  ggtitle("QQ Plot of Residuals")

plot1 / plot2
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



- The histogram of the residuals and the qqplot show signs of normally distributed errors.