Youtube Analysis

#Youtube Dataset Analysis

```
d <- read_delim(
  file = './videos.txt',
  delim = '\t'
  )
```

```
## 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.
```

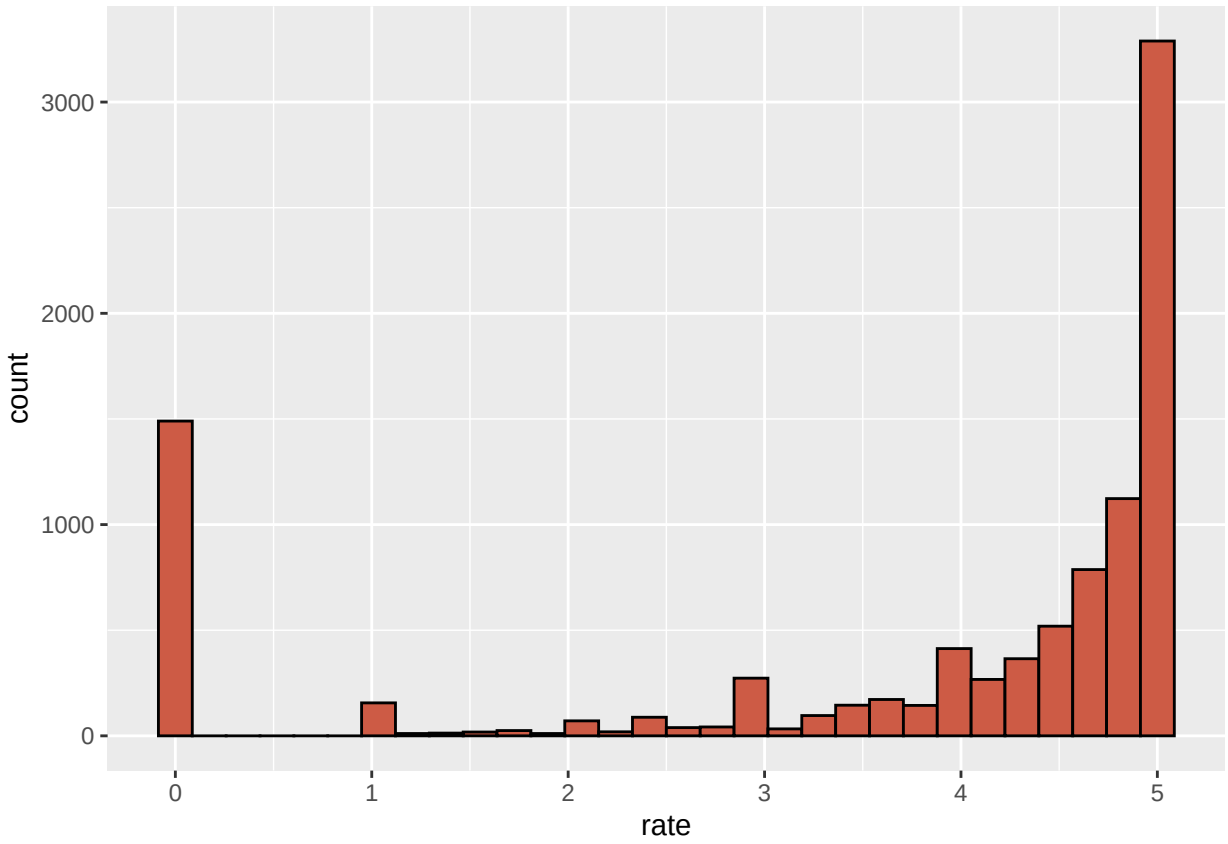
```
d <- d %>%
  mutate(
    uploader_f = factor(uploader),
    category_f = factor(category),
    )
glimpse(d)
```

```
## Rows: 9,618
## Columns: 11
## $ video_id   <chr> "9QR1tni70fo", "l1DCSqAJ740", "ZES_o3XYGjM", "4I8b40cViDE",~
## $ uploader   <chr> "BHJJYP", "musicalrox", "tessaceleste", "booloveswondergirl~
## $ age        <dbl> 1131, 1236, 1243, 1237, 1252, 1236, 1053, 1240, 1237, 1187,~
## $ category   <chr> "Comedy", "Music", "Entertainment", "Entertainment", "Comed~
## $ length     <dbl> 126, 243, 105, 278, 26, 252, 162, 37, 166, 139, 361, 243, 1~
## $ views      <dbl> 204, 1652, 898, 928, 392, 318, 749, 10, 115, 617, 37, 266, ~
## $ rate       <dbl> 3.00, 3.91, 4.48, 5.00, 1.50, 5.00, 3.00, 0.00, 2.00, 4.67,~
## $ ratings    <dbl> 2, 11, 81, 24, 8, 2, 6, 0, 1, 24, 1, 3, 52, 30, 114, 0, 1, ~
## $ comments   <dbl> 1, 4, 36, 13, 17, 3, 6, 0, 0, 17, 1, 1, 50, 17, 119, 101, 9~
## $ uploader_f <fct> BHJJYP, musicalrox, tessaceleste, booloveswondergirls, Fizz~
## $ category_f <fct> Comedy, Music, Entertainment, Entertainment, Comedy, Entert~
```

```
d %>%
  ggplot() +
  aes(rate) +
  geom_histogram(
    fill="coral3", color = "black")
```

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

```
## Warning: Removed 9 rows containing non-finite values (stat_bin).
```



What is happening with all of those zero ratings? Well, those are videos that haven't been rated at all. And, it seems that these are videos that aren't really being watched that much.

```
d %>%
  filter(rate == 0) %>%
  select(rate, ratings, views) %>%
  head()
```

```
## # A tibble: 6 x 3
##    rate ratings views
##   <dbl>   <dbl> <dbl>
## 1     0       0    10
## 2     0       0  8185
## 3     0       0    66
## 4     0       0    38
## 5     0       0    90
## 6     0       0    57
```

At this point, we've got to make a judgement call; should we keep these videos in the analysis, or not? I'm going to drop these videos that have zero ratings, and I'm going to re-save the dataset.

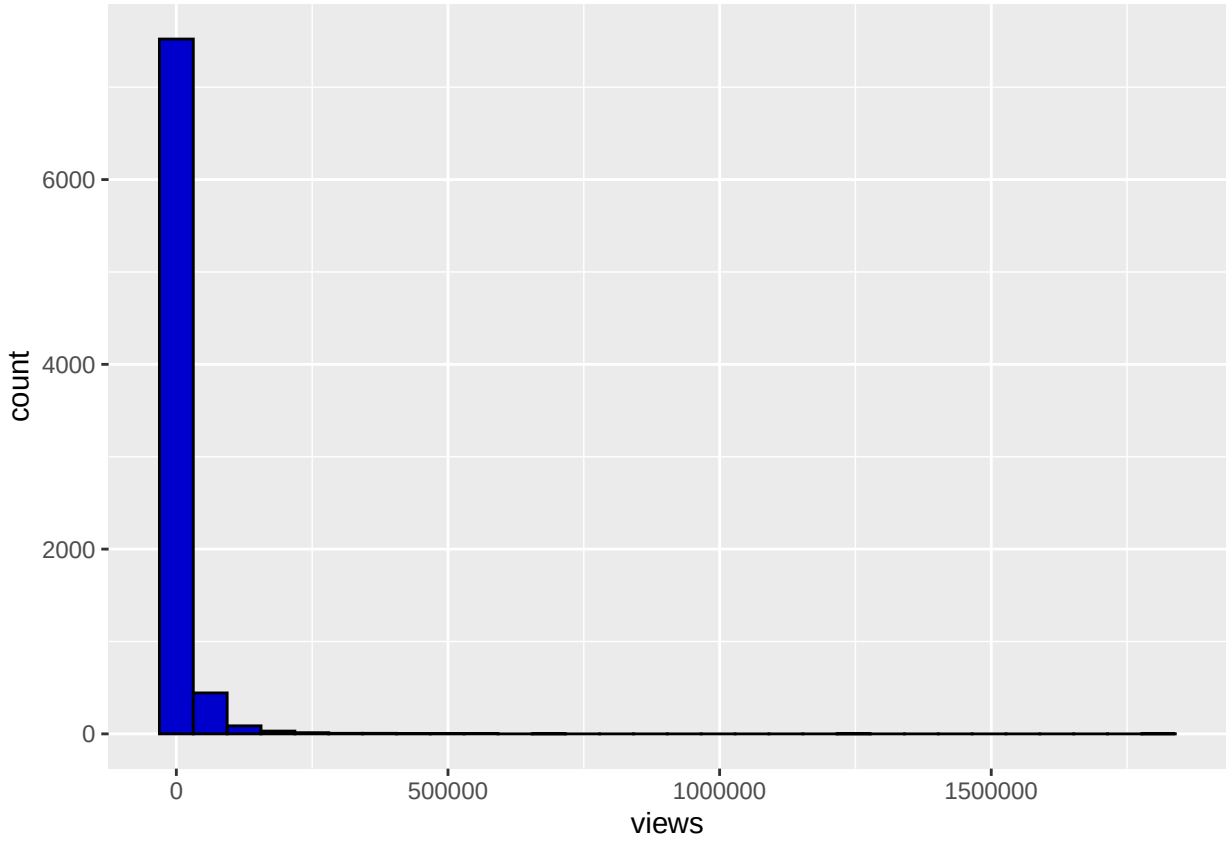
```
d <- d %>%
  filter(ratings > 0)
```

Next, I'd like to look at the distribution of views, which is the outcome variable that we're interested in. My bet is that this is going to have a long tail distribution.

```
d %>%
  ggplot(aes(x = views)) +
  geom_histogram(,
```

```
    fill="blue3", color = "black")
```

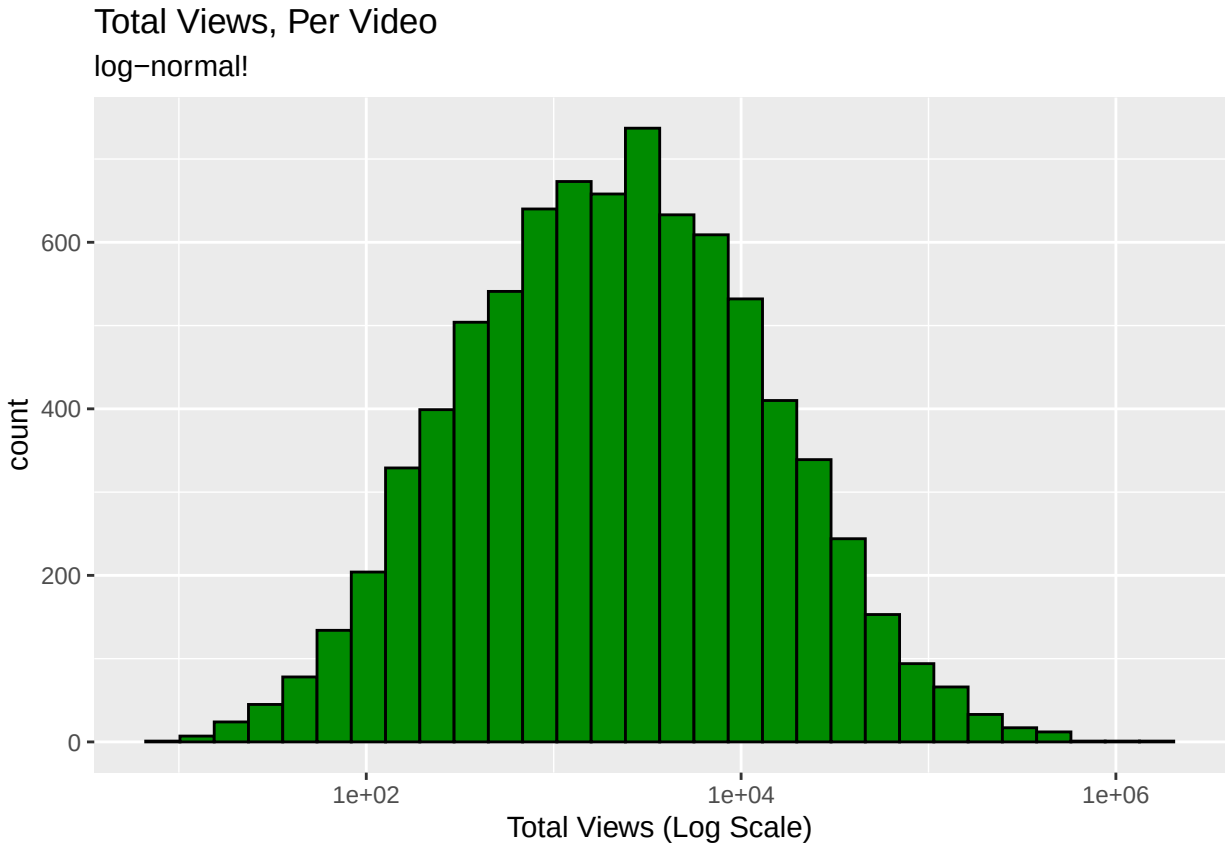
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Oh yeah. That's for sure. What does this distribution look like if we log transform it?

```
d %>%
  ggplot() +
  aes(x = views) +
  geom_histogram(,
    fill="green4", color = "black") +
  scale_x_continuous(trans = 'log10') +
  labs(
    x = 'Total Views (Log Scale)',
    title = 'Total Views, Per Video',
    subtitle = 'log-normal!'
  )
```

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



Regression analysis of YouTube dataset

- views: the number of views by YouTube users.
- rate: the average rating given by users.
- length: the duration of the video in seconds.

We want to use the `rate` variable as a proxy for video quality. we also include `length` as a control variable. Lastly we estimate the following OLS regression:

$$views = 789 + 2103 \cdot rate + 3.00 \cdot length$$

One thing that could cause significant omitted variable bias is that the category of the video has an effect on how many views the video receives. As an example, people are very willing to re-watch music videos, but they're much less likely to re-watch comedy videos.

```
d %>%
  filter(
    category %in% c(
      'Comedy', 'Education', 'Entertainment', 'Music', 'Sports')) %>%
  ggplot(aes(x = views)) +
  geom_histogram(show.legend = FALSE) +
  facet_wrap(~category_f) +
  scale_x_continuous(trans = 'log10') +
  labs(
    x = 'Total Views (Log Scale)',
    title = 'Total Views, Per Video',
    subtitle = 'log-normal!')
```

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