Youtube Clickbait Title Report

Does clickbait play into views?

Oscar Casas

12/6/2021

1 Introduction

For centuries individuals have looked for new sources of entertainment. From books to television, the sources were usually in control of what people had available. Since the inception of the internet however, that has completely changed and anyone with a camera and an internet connection can create their own content. Now that this pivot has been created, specifically on YouTube, content creators need to be informed on what is the best way to compete on their platform. For this study, as YouTube content creators, we will focus specifically on YouTube. One of the key terms that have been floating around YouTube has been the term "click bait". This is the idea that individuals upload videos with misleading titles to lure viewers. While that specific idea may be outside the scope of our analytical powers due to the high amount of volume of videos that would need to be validated, the question does arise. How do titles and other YouTube characteristics such as ratings affect the amount of views videos get? We explore this question through the use of a multitude of variables to analyze what variable has the strongest effect on views received. The variables considered: a) ratings, which is the proportion of likes to all dislikes and likes. This was chosen to see if the engagement after someone clicked on a video affected the amount of people who watched it after. b) proportion of uppercase letters. Oftentimes videos will have bold titles with lots of uppercase letters to emphasize or catch attention, it will be interesting to see how they affect viewership. c) whether titles include question marks or exclamation points. Titles often use these syntax to emphasize or create a luring title. d) title length. As title length increases less is seen in the thumbnail so that may affect clicks as well. These variables are based on our input data which consists of over 40,000 trending YouTube videos titles, likes and dislikes and their number of views. While many different types of variables and measurements could be used to study what affects video views, these were chosen as the prime focus to see what content creators should do at the time of posting in relation to creating their titles.

2 Research Design

2.1 Research Question

How does the title of a trending YouTube video affect the number of views that it gets?

2.2 Method

Our research design will be based on a quantitative approach, specifically an experimental one, in which our research question will be answered by conducting statistical modeling experiments, with computer programming, to determine which variables (associated with the title) relate to higher viewership in YouTube videos.

2.3 Design

We performed a longitudinal retrospective study on our Kaggle dataset with "Trending Videos" as our target population. By operationalizing variables related to trending video titles, we tested for linear relationships between title-based variables and video views. After completing our exploratory analysis, we designed Ordinary Least Squares regression models using transformed and un-transformed variables. We tested our model coefficients for significance at a 5% level and compared each model to a reduced model that used only an intercept term. By making comparisons to a reduced model using an F-Test, we determined whether the models we created were statistically significant at predicting views of trending videos.

2.4 Data

Our data was taken from the website **Kaggle** and it contains several variables on videos viewed within the United States version of YouTube's trending section, collected from January 2017 through to May 2018. The variables found within the dataset are: video ID, date of trending, title, channel name, category, time published, tags, number of views, likes and dislikes.

Kaggle DataSet Link: https://www.kaggle.com/datasnaek/youtube-new?select-USvideos.csv=

2.5 Variables

Since our research question was focused on the effect video titles have on viewership, we narrowed our efforts to variables related to titles.

1. **Title Length:** Anecdotally, our prior experience told us that titles of an excessive length make us unlikely to watch a video. Many people feel that excessive length is indicative of overselling a video. We hypothesized that more concise titles would receive more views than their larger counterparts. To operationalize Title Length into a metric variable, we counted the number of characters in each video title.

2. **Uppercase Proportion:** Our hypothesis was that the proportion of characters in a title that were uppercase could serve as a rough estimate of the level of title "clickbait." Under this assumption, we predicted that a positive relationship would exist between the proportion of uppercase letters and views.

3. **Contains a Question Mark:** Many videos try to catch the attention of viewers by hooking them with cliffhanger questions (e.g. "Will she fall out of the helicopter??? Watch and see."). Our hypothesis was that these video titles may lure more people to watch a video because they get hooked by the title along with an interesting thumbnail picture.

4. **Contains an Exclamation Mark:** As in the case of the previous variable, an exclamation mark is often used in an attempt to capture the viewer's attention with sensationalized content. Subsequently, we hypothesized that the presence of an exclamation mark would lead to more views.

5. **Contains an Ellipses:** Again, the use of an ellipses is another sensationalist tactic to hook potential viewers with clickbait. Similarly to the previous two variables, our hypothesis was that there would likely be more views as a result of the use of ellipses.

6. **Rating:** YouTube videos don't come with metric ratings by default, however, they come with a count of likes and a count of dislikes. We suspected that the total number of likes and dislikes is highly correlated with the number of views that a video received. To standardize these variables into a metric that scales evenly across videos, we created a Rating variable that compares the proportion of likes and dislikes for each video. Our hypothesis was that a higher proportion of likes to dislikes (higher Rating) would lead to more video views. The Rating variable was determined as follows:

$$Rating = \frac{Likes}{Likes + Dislikes}$$

7. **Views:** Our independent variable. The number of views received by a trending video at a given point in time.

3 Large Sample Assumptions

1. **IID Data Sampling** Our dataset contains a list of daily YouTube trending videos over the course of several months. According to Variety magazine, "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they're not the most-viewed videos overall for the calendar year". We cannot make claims over the entire population of YouTube videos using this dataset, however, we do believe that each of the videos is representative of the "Trending Videos" distribution. The dataset is the population of trending videos over a period of time rather than a point-in-time snapshot. We can expect that the habits and preferences of YouTube viewers and maybe even the trending videos algorithm could change over time, meaning that the distribution that we are targeting may not be static. This could be an argument that the dataset is not IID, but we are choosing to ignore that and assume that we are targeting the same population distribution for all of our data.
2. **A unique Best Linear Predictor exists with variance > 0** A unique BLP exists when:
   (a) None of the variables in the dataset are perfectly colinear
   (b) $X^T X$ is invertible, meaning that no dependent variable can be written as a linear combination of the other dependent variables

   We know that none of our variables are perfectly colinear because they are all separate aspects of a YouTube title. Likewise, R automatically drops perfectly colinear variables when they're detected. None of our dependent variables can be written as a linear combination of the others.

4 Data Cleaning and Transformation

Here we will load, clean and transform our dataset to better suit our analysis.

4.1 Loading libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

4.2 Loading DataSet

```
# Random Seed
set.seed(5)
# Loading DataSet and Assigning it to Variable
videos <- read.csv('youtubevideos.csv')
# Eliminating Null Values
videos <- na.omit(videos)
# Eliminating 0 values from DataSet
videos <- videos[videos$likes > 0 & videos$dislikes > 0,]
#Adding ID for each row
videos$ID <- seq.int(nrow(videos))
# Preview of DataFrame
head(videos)
```

```
##   trending_date                                                        title
## 1      17.14.11                                WE WANT TO TALK ABOUT OUR MARRIAGE
## 2      17.14.11 The Trump Presidency: Last Week Tonight with John Oliver (HBO)
## 3      17.14.11          Racist Superman | Rudy Mancuso, King Bach & Lele Pons
## 4      17.14.11                                  Nickelback Lyrics: Real or Fake?
## 5      17.14.11                                          I Dare You: GOING BALD!?
## 6      17.14.11                                             2 Weeks with iPhone X
##            channel_title category_id             publish_time
## 1           CaseyNeistat          22 2017-11-13T17:13:01.000Z
## 2        LastWeekTonight          24 2017-11-13T07:30:00.000Z
## 3           Rudy Mancuso          23 2017-11-12T19:05:24.000Z
## 4 Good Mythical Morning          24 2017-11-13T11:00:04.000Z
## 5               nigahiga          24 2017-11-12T18:01:41.000Z
## 6               iJustine          28 2017-11-13T19:07:23.000Z
##
## 1
## 2
## 3
## 4 rhett and link|"gmm"|"good mythical morning"|"rhett and link good mythical morning"|"good mythical
## 5
## 6
##     views  likes dislikes comment_count comments_disabled ratings_disabled ID
## 1  748374  57527     2966         15954             FALSE            FALSE  1
## 2 2418783  97185     6146         12703             FALSE            FALSE  2
## 3 3191434 146033     5339          8181             FALSE            FALSE  3
## 4  343168  10172      666          2146             FALSE            FALSE  4
## 5 2095731 132235     1989         17518             FALSE            FALSE  5
## 6  119180   9763      511          1434             FALSE            FALSE  6
```

4.3 Creating New Columns

In the following script we create Binary and Boolean columns that establish if there is a question mark, exclamation mark or ellipses in a title, as well as a column that holds how many characters (including blank spaces) there are in a title, a "rating" column which is the ratio of likes to total likes and dislikes in a video, and another one which holds the numeric proportion of capital letters within a title.

```
# Creating Column for Length of Title as Numeric Value
videos['title_length'] = nchar(videos$title)
# Creating Column for rating as Numeric Value
videos['rating'] = videos['likes']/(videos['dislikes'] + videos['likes'])
# Creating Column for Proportion of Capital Letters in Title
videos['uppercase_proportion'] = str_count(videos$title, "[A-Z]")/
  str_count(videos$title)
# Creating Column for Boolean Value if Title Contains Question Mark
videos['question'] = grepl('?', videos$title, fixed=TRUE)
# Creating Column for Boolean Value if Title Contains Exclamation Mark
videos['exclamation'] = grepl('!', videos$title, fixed=TRUE)
# Creating Column for Boolean Value if Title Contains Ellipses
videos['ellipses'] = grepl('...', videos$title, fixed=TRUE)
# Creating Column for Binary Value if Title Contains Question Mark
videos['binary_question'] = ifelse(videos$question, 1, 0)
# Creating Column for Binary Value if Title Contains Exclamation Mark
videos['binary_exclamation'] = ifelse(videos$exclamation, 1, 0)
# Creating Column for Binary Value if Title Contains Ellipses
videos['binary_ellipses'] = ifelse(videos$ellipses, 1, 0)

# Eliminating 0 values from New Columns
videos <- videos[videos$title_length > 0 & videos$uppercase_proportion > 0,]
# New form of DataSet
head(videos)
```

```
##   trending_date                                                    title
## 1      17.14.11                              WE WANT TO TALK ABOUT OUR MARRIAGE
## 2      17.14.11 The Trump Presidency: Last Week Tonight with John Oliver (HBO)
## 3      17.14.11          Racist Superman | Rudy Mancuso, King Bach & Lele Pons
## 4      17.14.11                                Nickelback Lyrics: Real or Fake?
## 5      17.14.11                                        I Dare You: GOING BALD!?
## 6      17.14.11                                           2 Weeks with iPhone X
##            channel_title category_id             publish_time
## 1          CaseyNeistat          22 2017-11-13T17:13:01.000Z
## 2       LastWeekTonight          24 2017-11-13T07:30:00.000Z
## 3          Rudy Mancuso          23 2017-11-12T19:05:24.000Z
## 4 Good Mythical Morning          24 2017-11-13T11:00:04.000Z
## 5              nigahiga          24 2017-11-12T18:01:41.000Z
## 6              iJustine          28 2017-11-13T19:07:23.000Z
##
## 1
## 2
## 3
## 4 rhett and link|"gmm"|"good mythical morning"|"rhett and link good mythical morning"|"good mythical
## 5
## 6
##     views  likes dislikes comment_count comments_disabled ratings_disabled ID
## 1  748374  57527     2966         15954             FALSE            FALSE  1
```

```
## 2 2418783  97185     6146         12703             FALSE            FALSE  2
## 3 3191434 146033     5339          8181             FALSE            FALSE  3
## 4  343168  10172      666          2146             FALSE            FALSE  4
## 5 2095731 132235     1989         17518             FALSE            FALSE  5
## 6  119180   9763      511          1434             FALSE            FALSE  6
##   title_length    rating uppercase_proportion question exclamation ellipses
## 1           34 0.9509695            0.8235294    FALSE       FALSE    FALSE
## 2           62 0.9405212            0.1774194    FALSE       FALSE    FALSE
## 3           53 0.9647293            0.1509434    FALSE       FALSE    FALSE
## 4           32 0.9385495            0.1250000     TRUE       FALSE    FALSE
## 5           24 0.9851815            0.5000000     TRUE        TRUE    FALSE
## 6           21 0.9502628            0.1428571    FALSE       FALSE    FALSE
##   binary_question binary_exclamation binary_ellipses
## 1               0                  0               0
## 2               0                  0               0
## 3               0                  0               0
## 4               1                  0               0
## 5               1                  1               0
## 6               0                  0               0
```

4.4 Exploratory Set and Test Set

In order to prevent our analysis to contaminate the results of our linear model, we will perform the EDA on a randomly selected sample of 70% of our population.

```
#Exploratory Set, 70% of our DataSet
vid <- videos[sample(nrow(videos), 28664),]
#Test Set, 30% of our DataSet
test <- videos[!(videos$ID  %in% vid$ID),]
```

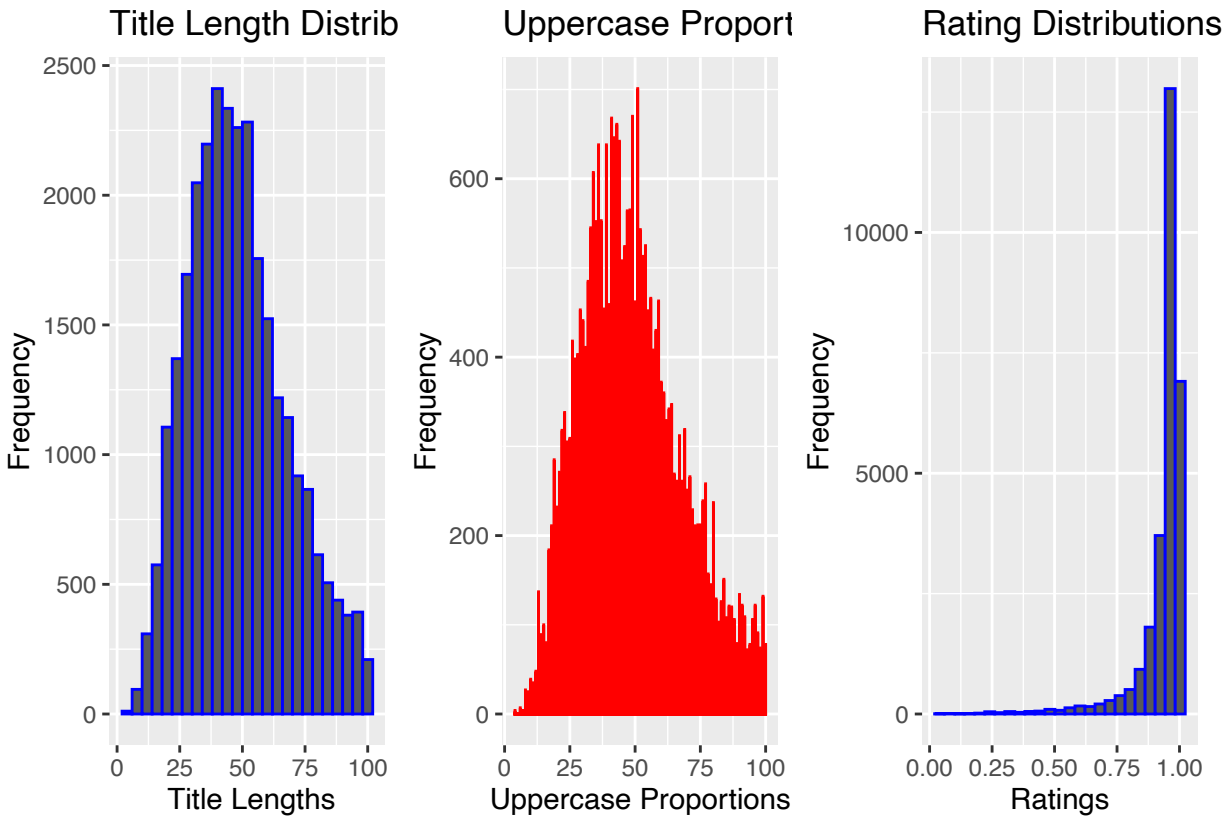
5 Exploratory Analysis

Here we will view how the behavior of various independent variables affect the amount of views a video receives. Given the observed behavior in the following plots, log of views generally demonstrates linear behavior with our **X** variables. Similarly, ratings shows linear behavior with log of views. However, it is unclear if title length and uppercase proportions exhibit desirable behavior with or without logarithmic transformations.

5.1 Distribution of Variables

Before going in-depth about the behavior of the **Y** variable given our **X** variables, we want to examine our independent variables separately from the dependent variable. Below we can observe the distributions of our numeric **X** variables and see that the uppercase proportion and title length distributions are very nearly normally distributed with some left tail skewness. This gives us reason to believe there will be much variation within these values, that there will be a tendency toward fewer uppercase letters in titles and shorter title lengths. Conversely, our ratings variable does not behave normally and is very heavily skewed towards higher ratings. Due to their behaviors, title length and uppercase proportions seem like prime candidates for our base linear model implementation.

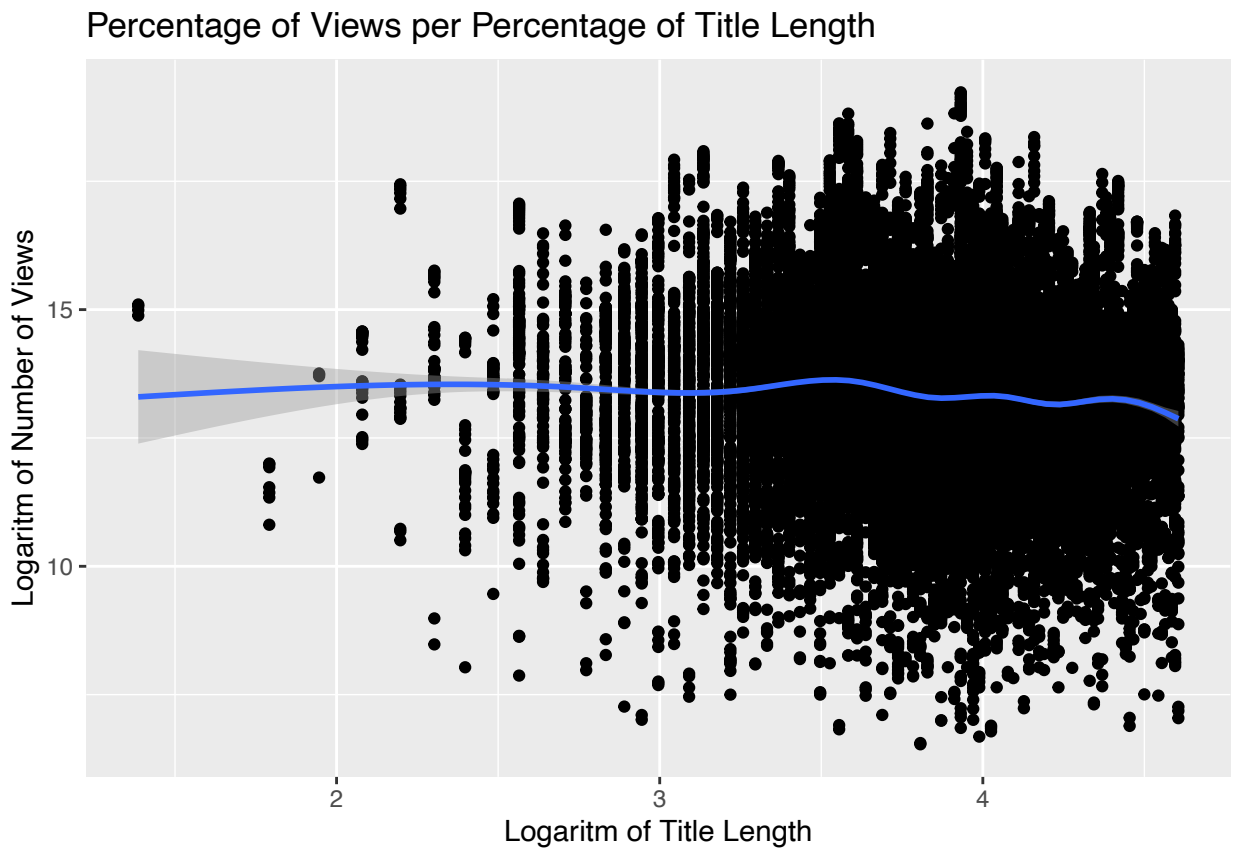
```
##   minimum_title_length minimum_uppercase_proportion minimum_rating
## 1                    4                   0.01162791     0.04011648
```



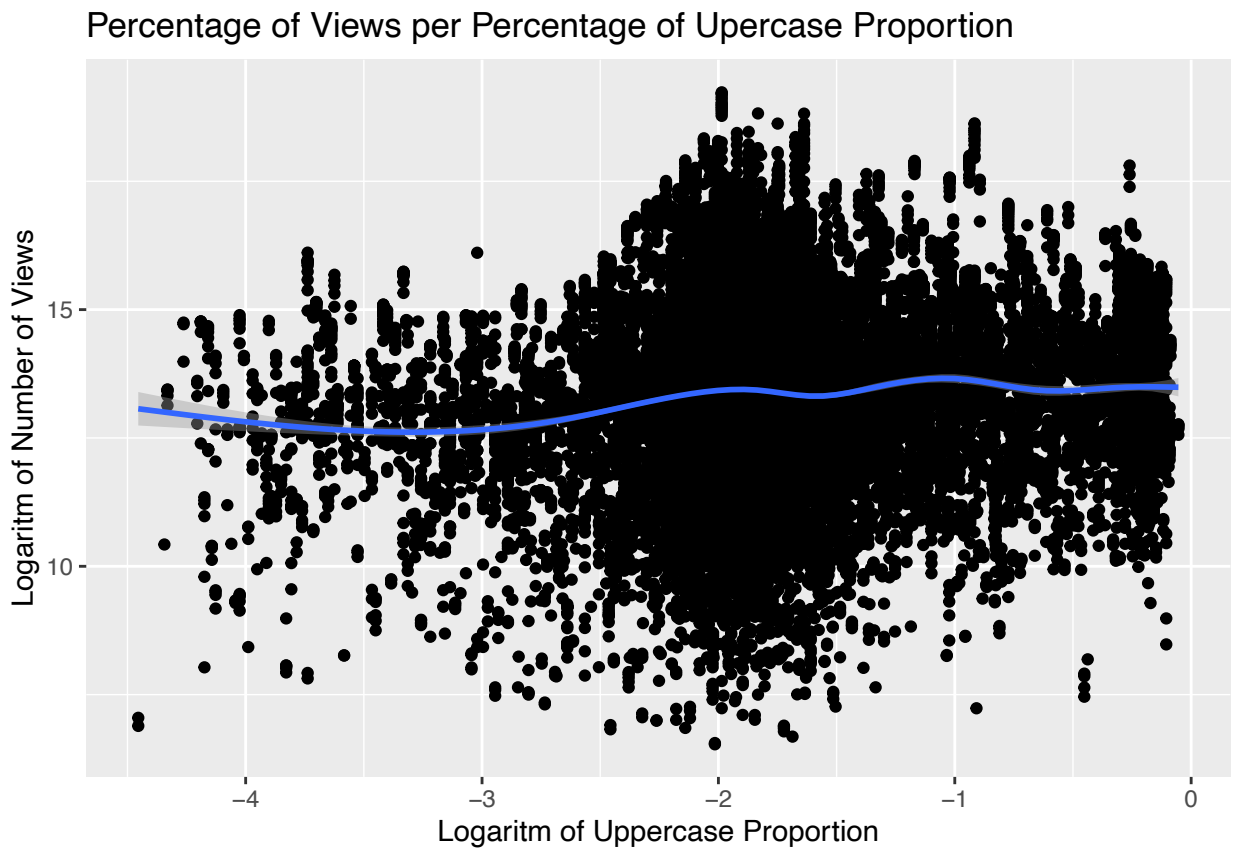
5.2 Log X vs Log Y

Within the following Plots, the logarithmic behavior of the **X** variables and how they affects the logarithmic behavior of our **Y** variable, is shown. This is done to determine if the percent change in **X** has a somewhat linear relationship with the percent change in **Y**.

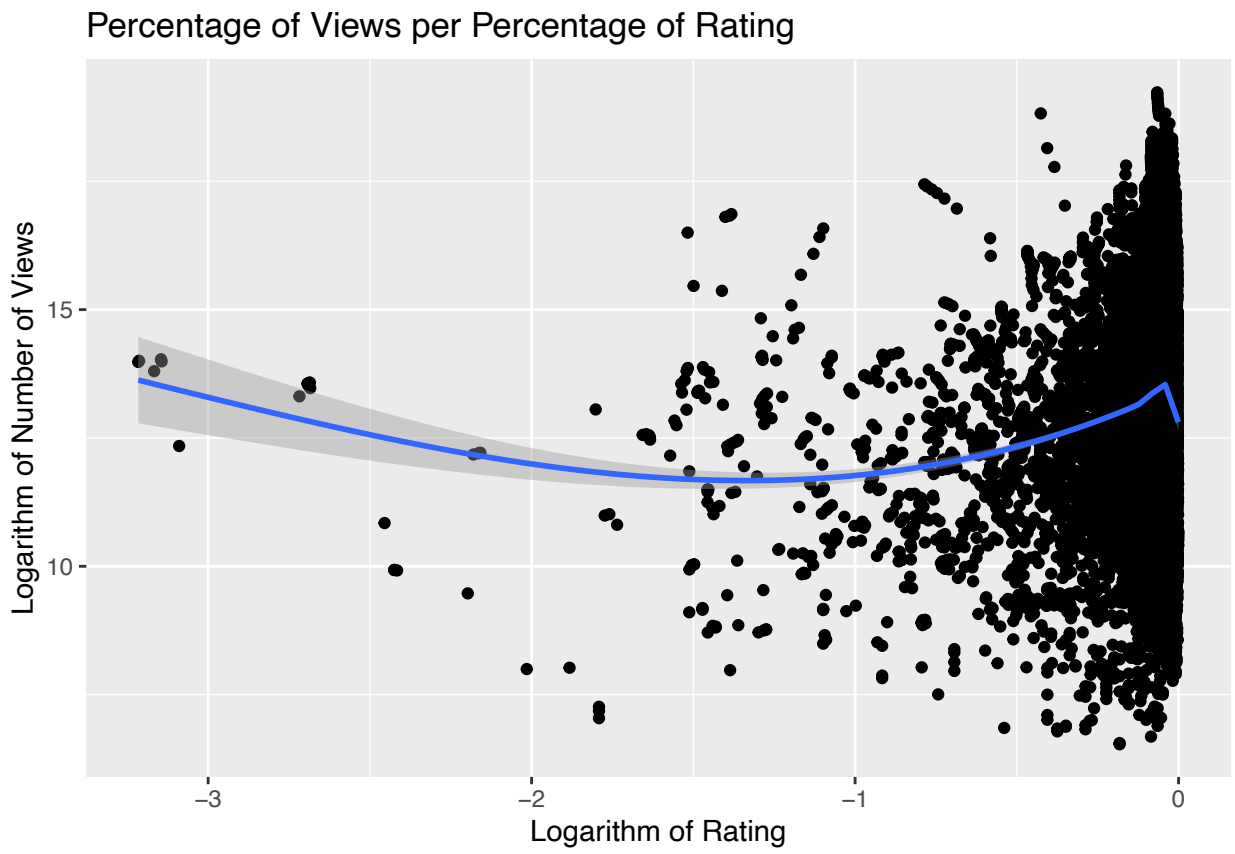
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




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



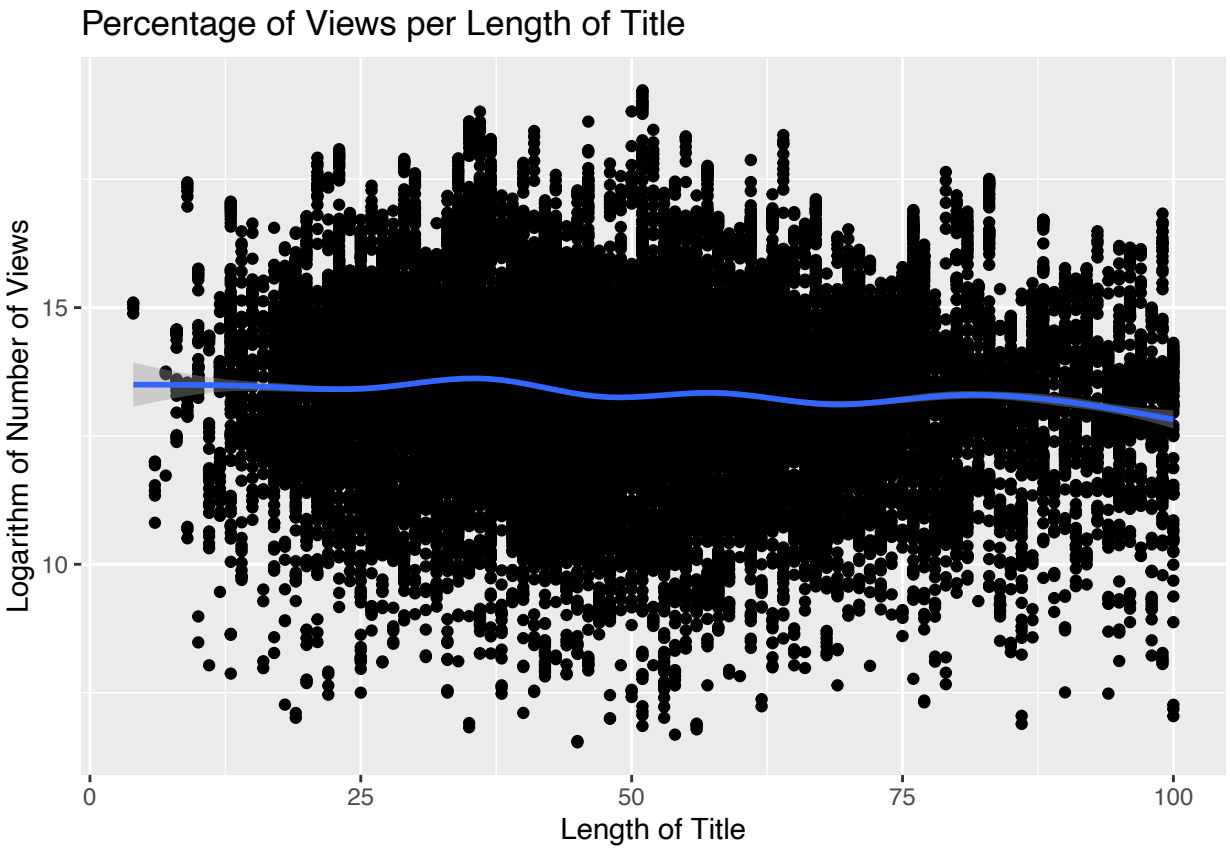

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

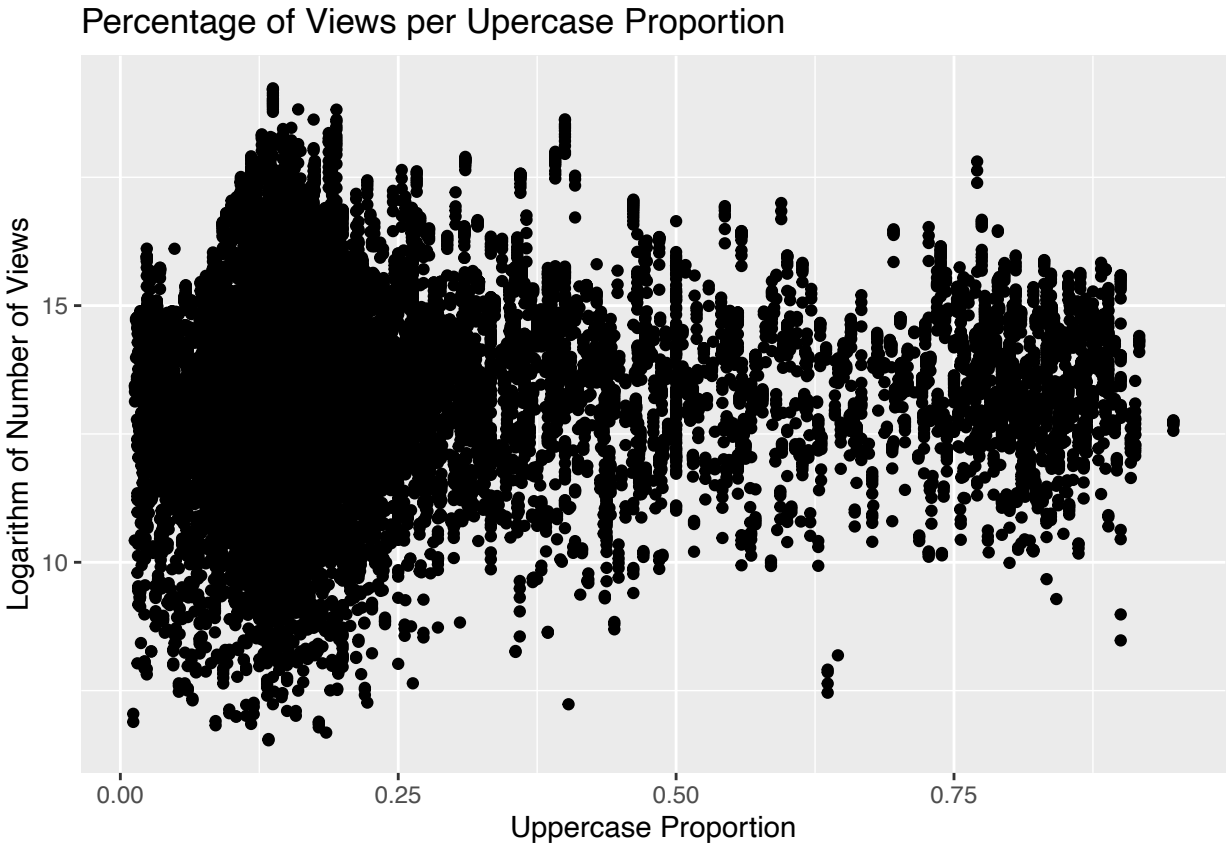


5.3 X vs Log Y

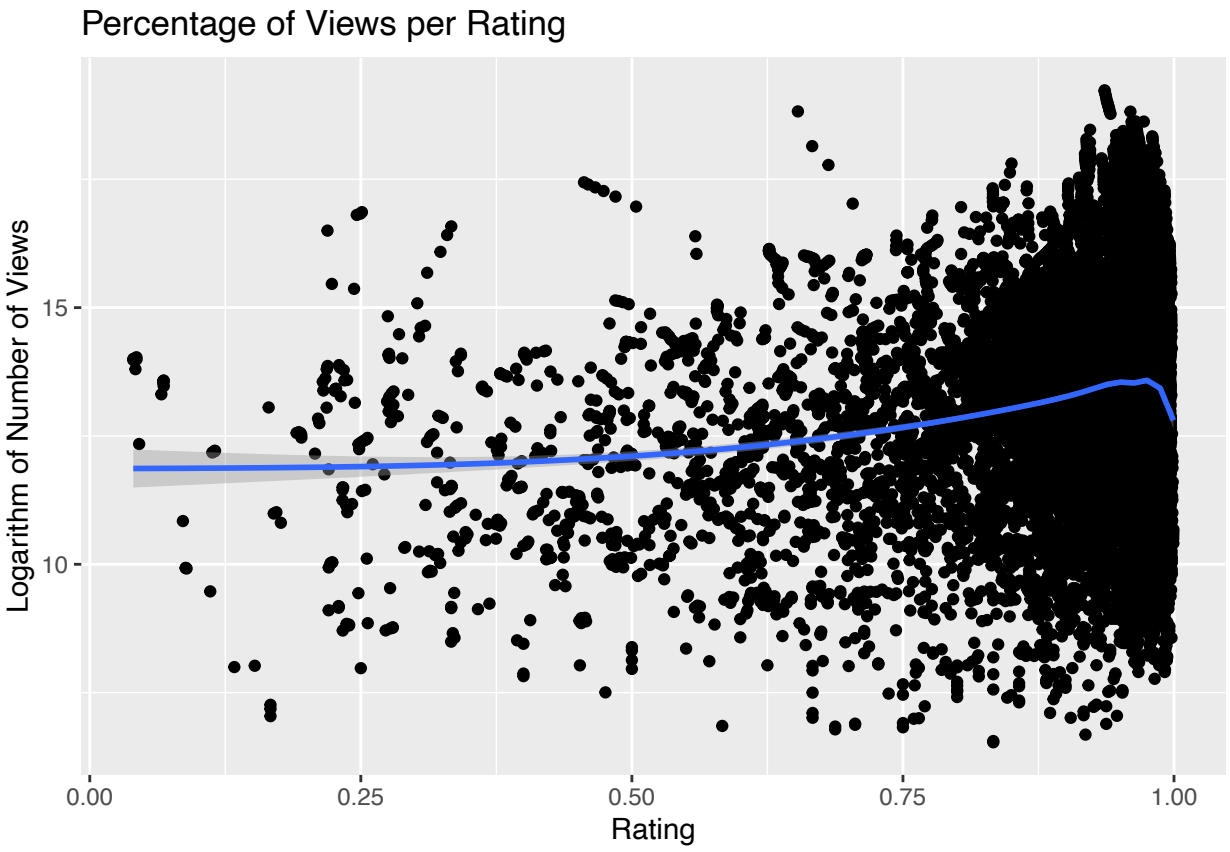
Following, we observe the behaviors of our **X** variables and how they affect the logarithmic behavior of our **Y** variable. The purpose is to confirm whether the changes in **X** have a somewhat linear relationship with the percent change in **Y**.

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


Percentage of Views per Length of Title
Logarithm of Number of Views
15
10
0
25
50
75
100
Length of Title



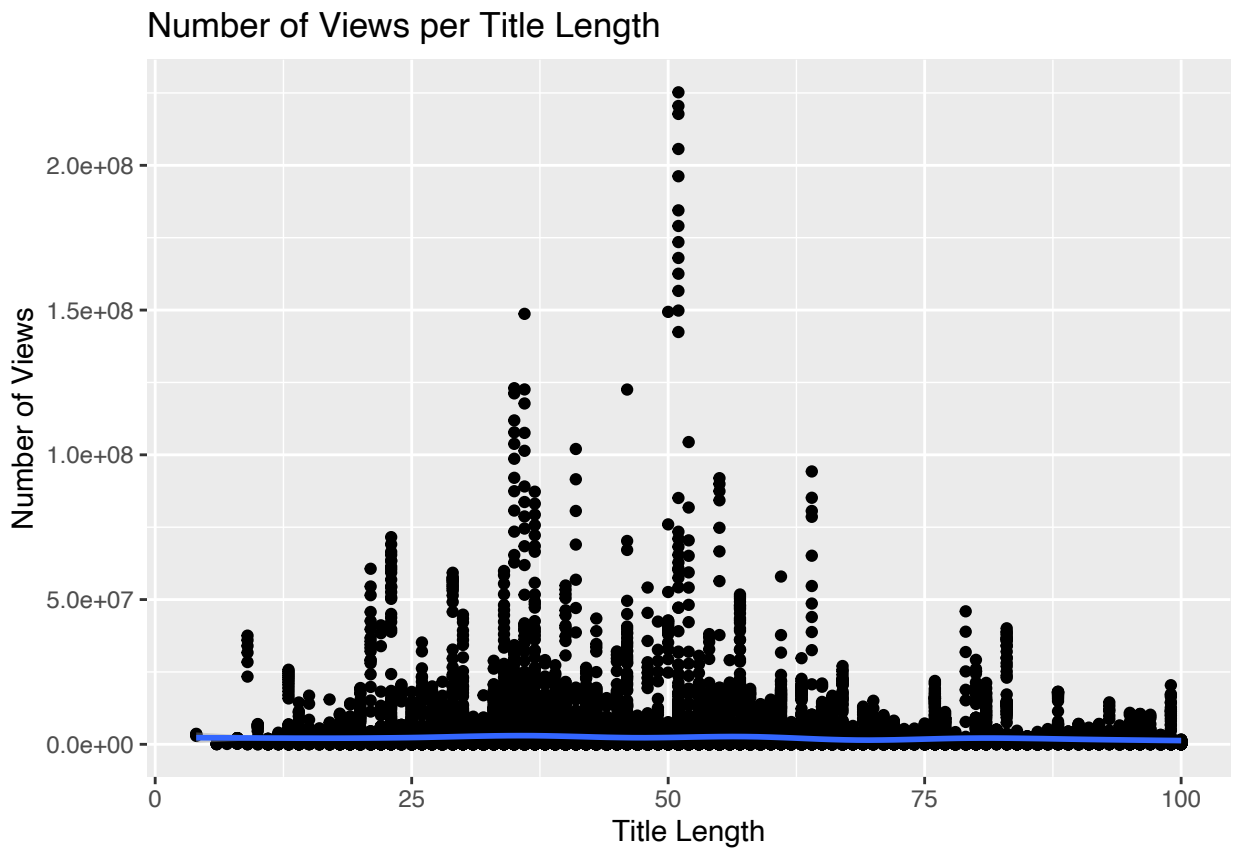

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



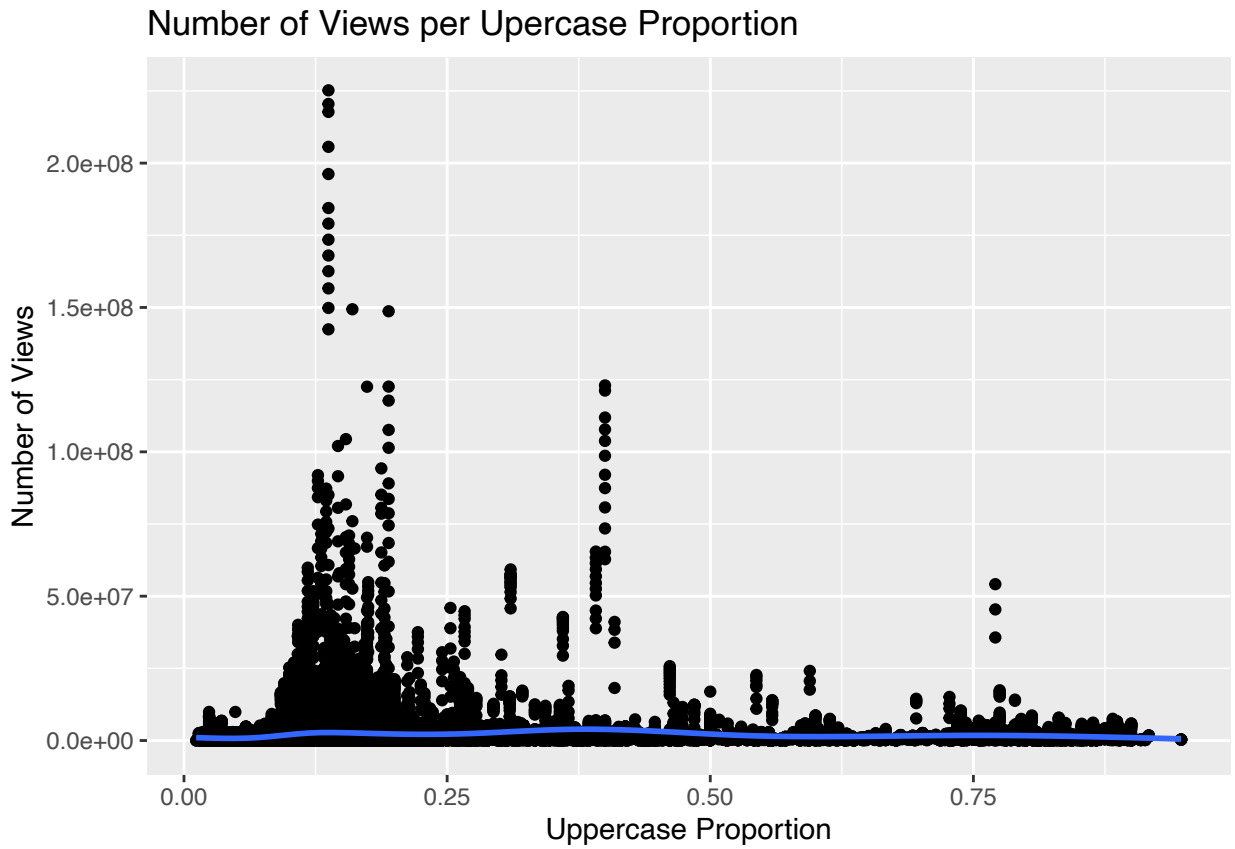
5.4 X vs Y

We also attempt to discern how the behavior of the **X** variables affect the behavior of the **Y** variable. Our objective is to ascertain if the changes in **X** have a linear relationship with the change in **Y**.

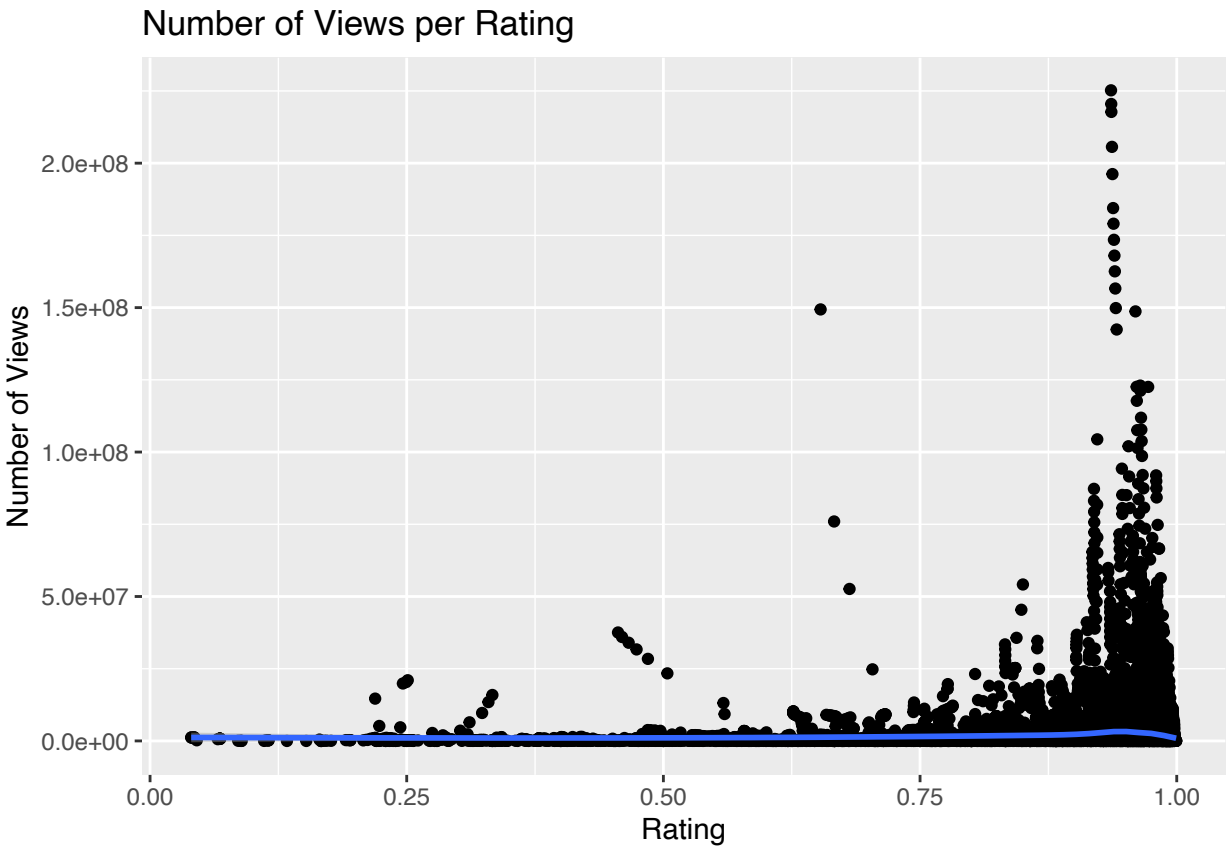
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




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



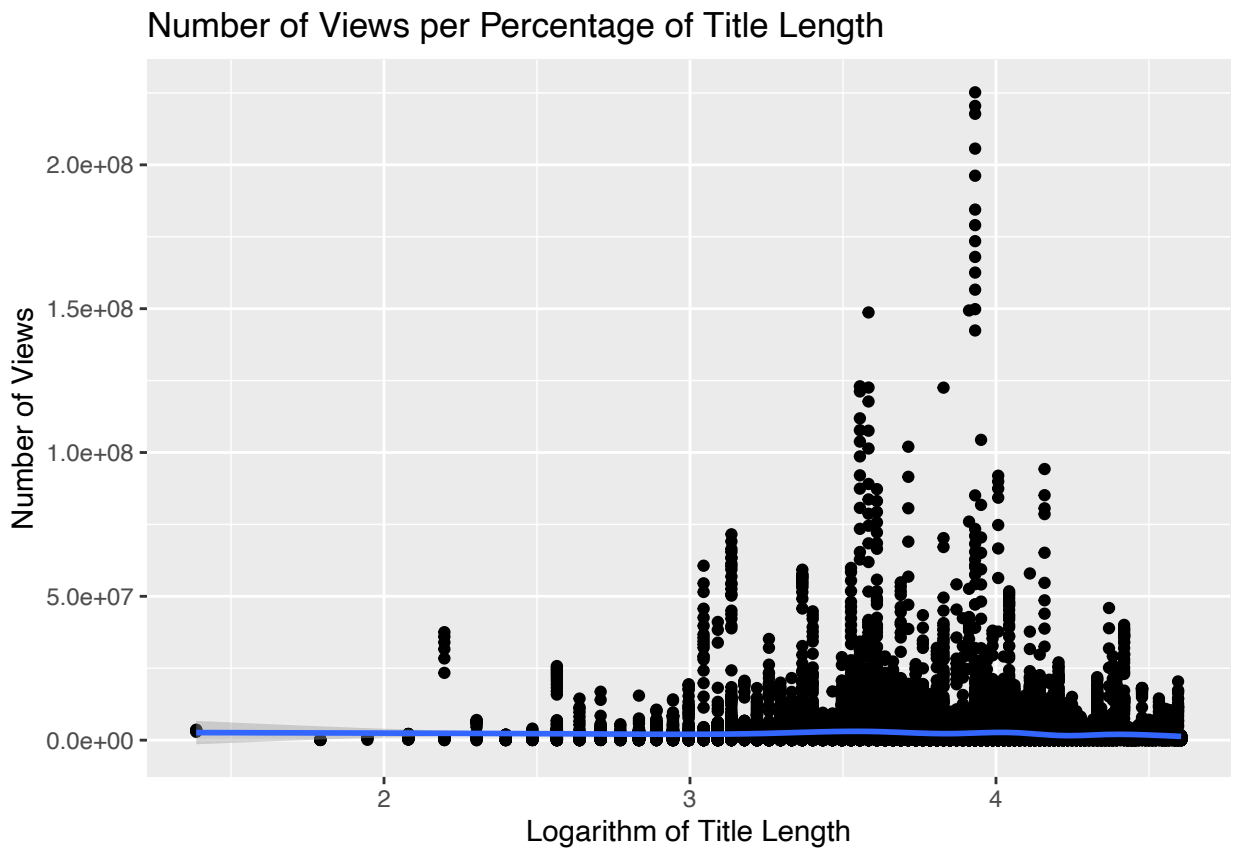

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



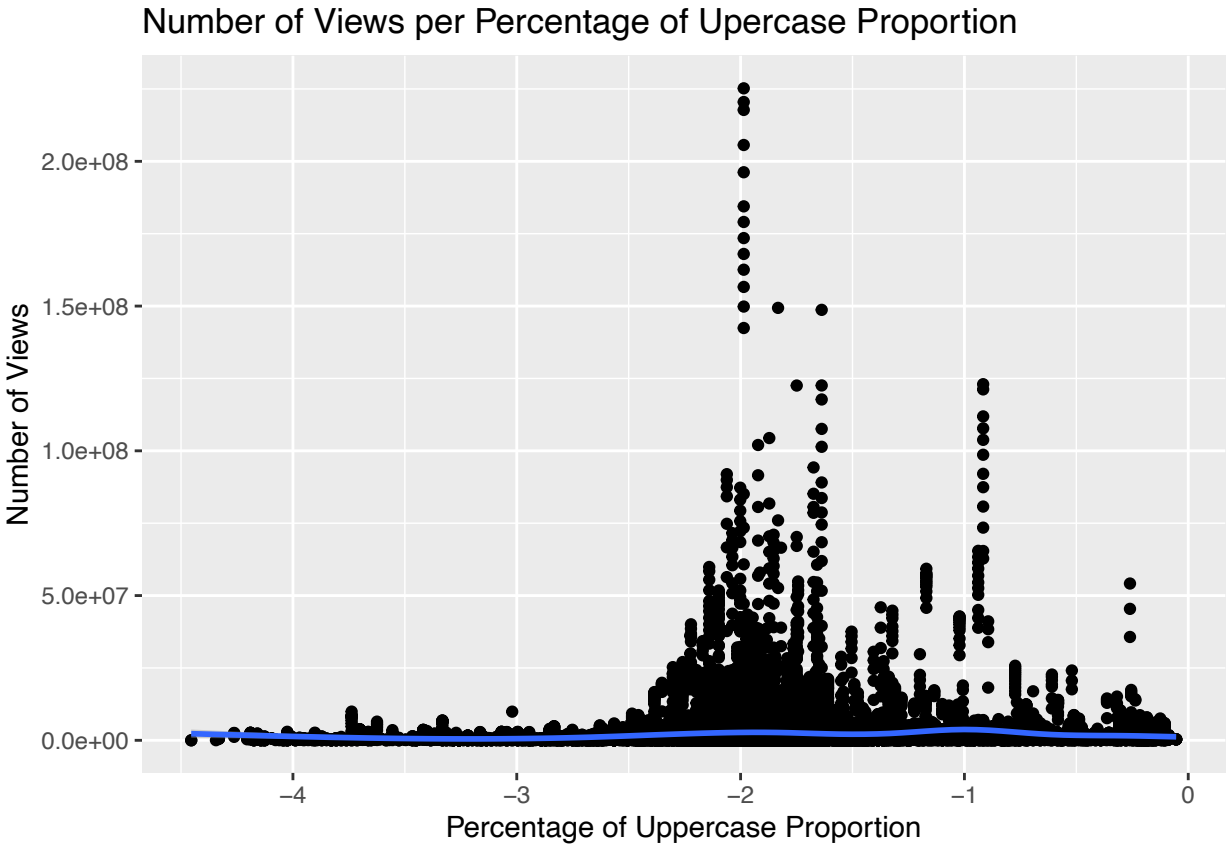
5.5 Log X vs Y

Finally, we observe the logarithmic behavior of our **X** variables and how they affect the **Y** variable. This is done to determine if the percent change in **X** has a linear relationship with the change in **Y**.

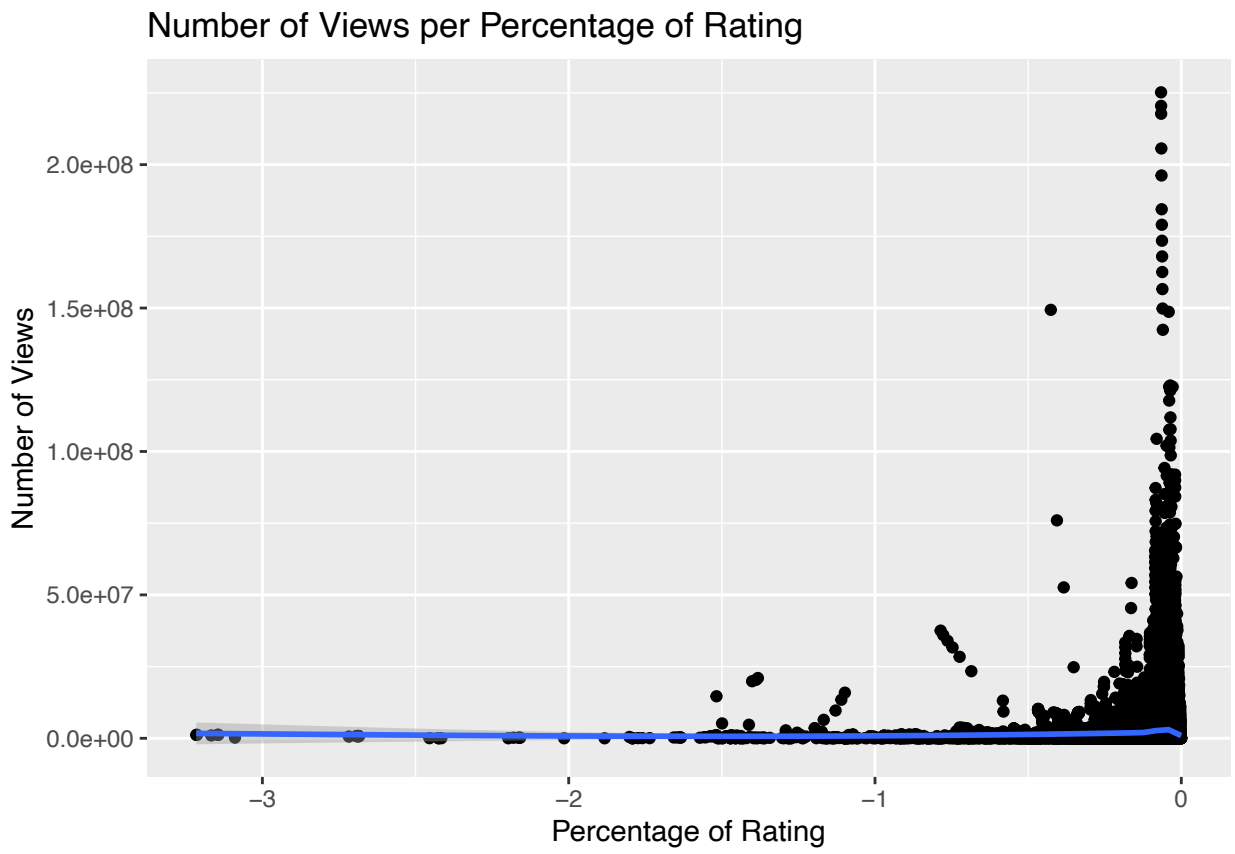
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




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




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



6 Linear Models and Testing

Given the nature of our research question, we believe that the key variables at our disposal to determine the effect of a video's title on the view count will be the length of the title and the proportion of uppercase words within the title. An appropriate number of uppercase words might make the video seem important (leading to more views) but too many (or the misuse of uppercase) might make the video appear to have less legitimacy, leading to less views. Similarly, too short a title may not adequately entice a user to view a video but too long might seem like a hassle to read and detract from viewership.

Given our EDA, we have reason to believe that log of views would be a better **Y** variable for our model than just views. However, we were unable to determine if there is any need to transform the **X** variables, title_length and uppercase_proportion. Because of this, we will construct two models. **Model 1** will have the logarithmic transformation for uppercase_proportion and title_length, while **Model A** wont.

Model 1:

$$log(Views) = \beta_0 + \beta_1 log(TitleLength) + \beta_2 log(UppercaseProportion)$$

```
##
## Model 1
## ================================================
##                              Dependent variable:
##                          ---------------------
##                                  log(views)
## ------------------------------------------------
## log(title_length)                  -0.2***
```

```
##                                    (0.02)
##
## log(uppercase_proportion)          0.2***
##                                    (0.01)
##
## Constant                           14.5***
##                                     (0.1)
##
## ------------------------------------------------
## Observations                       28,664
## R2                                  0.01
## Adjusted R2                         0.01
## Residual Std. Error           1.7 (df = 28661)
## F Statistic                152.9*** (df = 2; 28661)
## ================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Model A:

$$log(Views) = \beta_0 + \beta_1 TitleLength + \beta_2 UppercaseProportion$$

```
##
## Model A
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                              log(views)
## ------------------------------------------------
## title_length                  -0.01***
##                                (0.001)
##
## uppercase_proportion           0.4***
##                                 (0.1)
##
## Constant                       13.6***
##                                (0.03)
##
## ------------------------------------------------
## Observations                   28,664
## R2                              0.01
## Adjusted R2                     0.01
## Residual Std. Error       1.7 (df = 28661)
## F Statistic           104.0*** (df = 2; 28661)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

6.1 Key Model 1 and A MSR comparisons

To decide which model better fits our data, we will compare **Model 1** and **Model A**'s mean square residual (**MSR**). As seen below, **Model 1** has a slightly smaller **MSR** value than **Model A**, which would lead us to believe it is better model. Therefore, **Model 1** will serve as our starting model.

```
# Function to calculate MSR of models
calculate_msr <- function(model) {
  msr <- mean(resid(model)^2)
```

```
  return(msr)
}
# MSR Model 1
calculate_msr(model_1)
```

```
## [1] 2.74632
```

```
# MSR Model 2
calculate_msr(model_A)
```

```
## [1] 2.755614
```

6.2 Additional Models

After creating our base model, we added to it by creating new models and comparing their effectiveness. We formed an additional model that takes into account the grammar found in titles, including the use of question marks, exclamation points and ellipses. These punctuations may serve to entice viewership or, perhaps due to their overuse throughout the years, detract from views. Our third model added to this by using a ratings variable which attempts to capture the quality of the video as a proportion of likes over total likes and dislikes. The rating helped to make the coefficients of our more title-based variables more accurate by removing any influence from how enjoyable the video is. Because our EDA demonstrated a more linear relationship between log of views and un-transformed ratings, we did not transform the ratings variable. Finally, we created a fourth model which is a mirror of **model 3** but without the log of views transformation. This fourth model determined if our transformation choices were adequate.

Base Model:

$$log(Views) = \beta_0 + \beta_1 log(TitleLength) + \beta_2 log(UppercaseProportion)$$

Model 2:

$$log(Views) = \beta_0 + \beta_1 log(TitleLength) + \beta_2 log(UppercaseProportion) +$$
$$\beta_3 BinaryQuestionMark + \beta_4 BinaryExclamationMark + \beta_5 BinaryEllipses$$

```
##
## Model 2 Regression Table
## ================================================
##                              Dependent variable:
##                          -----------------------
##                                  log(views)
## ------------------------------------------------
## log(title_length)                 -0.2***
##                                   (0.02)
##
## log(uppercase_proportion)         0.2***
##                                   (0.02)
##
## binary_question                    -0.1*
##                                   (0.04)
##
## binary_exclamation                 0.1**
##                                   (0.03)
##
## binary_ellipses                     0.1
```

```
##                                      (0.1)
##
## Constant                            14.5***
##                                      (0.1)
##
## -----------------------------------------------------
## Observations                        28,664
## R2                                   0.01
## Adjusted R2                          0.01
## Residual Std. Error            1.7 (df = 28658)
## F Statistic                 62.8*** (df = 5; 28658)
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Model 3:

$$log(Views) = \beta_0 + \beta_1 log(TitleLength) + \beta_2 log(UppercaseProportion) +$$
$$\beta_3 BinaryQuestionMark + \beta_4 BinaryExclamationMark + \beta_5 BinaryEllipses + \beta_6 Rating$$

```
##
## Model 3 Regression Table
## =====================================================
##                                Dependent variable:
##                            ---------------------------
##                                    log(views)
## -----------------------------------------------------
## log(title_length)                   -0.2***
##                                      (0.02)
##
## log(uppercase_proportion)           0.1***
##                                      (0.02)
##
## binary_question                      -0.1*
##                                     (0.04)
##
## binary_exclamation                   0.03
##                                     (0.03)
##
## binary_ellipses                       0.1
##                                      (0.1)
##
## rating                              2.5***
##                                      (0.1)
##
## Constant                            11.9***
##                                      (0.1)
##
## -----------------------------------------------------
## Observations                        28,664
## R2                                   0.03
## Adjusted R2                          0.03
## Residual Std. Error            1.6 (df = 28657)
## F Statistic                 172.0*** (df = 6; 28657)
## =====================================================
```

```
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

Model 4:

$$Views = \beta_0 + \beta_1 TitleLength + \beta_2 UppercaseProportion +$$
$$\beta_3 BinaryQuestionMark + \beta_4 BinaryExclamationMark + \beta_5 BinaryEllipses + \beta_6 Rating$$

```
##
## Model 4 Regression Table
## ================================================
##                             Dependent variable:
##                            ---------------------
##                                    views
## ------------------------------------------------
## log(title_length)                -329,357.5***
##                                   (101,661.6)
##
## log(uppercase_proportion)        203,210.3***
##                                   (68,064.1)
##
## binary_question                  -522,125.6***
##                                   (186,536.9)
##
## binary_exclamation               -922,056.9***
##                                   (138,971.4)
##
## binary_ellipses                  -982,751.4***
##                                   (361,698.4)
##
## rating                          2,143,548.0***
##                                   (430,916.7)
##
## Constant                        2,127,021.0***
##                                   (595,091.1)
##
## ------------------------------------------------
## Observations                        28,664
## R2                                   0.004
## Adjusted R2                          0.003
## Residual Std. Error        7,420,292.0 (df = 28657)
## F Statistic                 17.7*** (df = 6; 28657)
## ================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

6.3 Comparing MSR of All models

Similar to our previous test, we can compare the **MSR** of our models to determine their effectiveness. Given that **Model 4** is not using a logarithmic transformation as the others are, its **MSR** is not comparable to the other models. From the following results it is clear that **Model 3** appears to have the best fit of our compared models, likely due to the inclusion of a proxy for video quality.

```
#MSR of Model 1
calculate_msr(model_1)
```

```
## [1] 2.74632
```

```
#MSR of Model 2
calculate_msr(model_2)
```

```
## [1] 2.74554
```

```
#MSR of Model 3
calculate_msr(model_3)
```

```
## [1] 2.679164
```

6.4 MSRE and Percent Error

In order to best compare all models, we will use their Root Mean Squared Error (**RMSE**) values over their **Y** means, as well as the **average percent difference** between their predicted values for the test set and the actual expected value. These metrics will allow us to scale our residual variation against transformed Y values to allow model comparisons over both views and the log of views. As seen below, **Model 4** presented a much higher **average percent difference** and **RMSE/Mean**, which justifies our choice of using the log of views rather than un-transformed views as the **Y** variable. In contrast, all other models have rather low **RMSE/Mean** and **average percent difference**; particularly, **Model 3** has the lowest **RMSE** and percent difference which leads us to choose it as our best linear model.

```
#calculating RMSE function
calculate_rmse <- function(model) {
  msr <- (calculate_msr(model))^(1/2)
  return(msr)
}
# Model 1 MSRE/Mean
calculate_rmse(model_1)/mean(log(test$views))
```

```
## [1] 0.1237328
```

```
# Model 2 MSRE/Mean
calculate_rmse(model_2)/mean(log(test$views))
```

```
## [1] 0.1237152
```

```
# Model 3 MSRE/Mean
calculate_rmse(model_3)/mean(log(test$views))
```

```
## [1] 0.1222106
```

```
# Model 4 MSRE/Mean
calculate_rmse(model_4)/mean(test$views)
```

```
## [1] 3.103512
```

```
# Percent Errors for Model 1
percent_error_1 <- abs(100*(predict(model_1, newdata = test) - log(test$views))/
      log(test$views))
# Average Errors for Model 1
mean(percent_error_1)
```

```
## [1] 10.03326
```

```
# Percent Error for Model 2
percent_error_2 <- abs(100*(predict(model_2, newdata = test) - log(test$views))/
      log(test$views))
# Average Errors for Model 2
mean(percent_error_2)
```

```
## [1] 10.02624
```

```
# Percent Error for Model 3
percent_error_3 <- abs(100*(predict(model_3, newdata = test) - log(test$views))/
      log(test$views))
# Average Errors for Model 3
mean(percent_error_3)
```

```
## [1] 9.905946
```

```
# Percent Error for Model 4
percent_error_4 <- abs(100*(predict(model_4, newdata = test) - test$views)/
      test$views)
# Average Errors for Model 4
mean(percent_error_4)
```

```
## [1] 1948.445
```

6.5 Model 3 Coeficient Test

Now that we have chosen our model, we will run a coefficient test to determine which **X** variables are found to be statistically significant in explaining the variation of the **Y** variable. By the results, it seems that the binary variables, binary_exclamation and binary_ellipses, do not demonstrate any statistical significance. Because of this, we will create a 5th model (**Model 5**) that will be **Model 3** without the 2 previously mentioned variables and use an **F test** in order to determine if one model explains the variation in the dependent variable significantly better than the other. Given the results of the **F test**, we were unable to reject the null hypothesis and thus we will assume **Model 5** explains the population's behavior just as well as **Model 3** with the use of less variables. Ergo, we will utilize **Model 5** as our preferred model.

```
# Coefficient Test for Model 1
coeftest(model_3, vcov = vcovHC(model_3))
```

```
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                11.872808   0.146844 80.8534 < 2.2e-16 ***
## log(title_length)          -0.161045   0.022199 -7.2546 4.132e-13 ***
## log(uppercase_proportion)   0.149983   0.013930 10.7670 < 2.2e-16 ***
## binary_question            -0.077341   0.035516 -2.1776   0.02944 *
## binary_exclamation          0.031555   0.027311  1.1554   0.24793
## binary_ellipses             0.059206   0.060614  0.9768   0.32869
## rating                      2.533075   0.111975 22.6219 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model Without Insignificant Values from Model 2
model_5 <- lm(log(views) ~ log(title_length) + log(uppercase_proportion) +
                binary_question + rating, data = vid)
# F Test Between Model 4 and Model 1
anova(model_3, model_5, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: log(views) ~ log(title_length) + log(uppercase_proportion) +
##     binary_question + binary_exclamation + binary_ellipses +
##     rating
```

```
## Model 2: log(views) ~ log(title_length) + log(uppercase_proportion) +
##     binary_question + rating
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1  28657 76796
## 2  28659 76800 -2    -4.437 0.8279  0.437
```

6.6 Model 5 Results

Finally, we view the **average percent error**, **MSR** and **MSRE/MEAN** values of our final **model 5**. The results yielded good predictive behavior for our model given its low **MSR**, **MSRE/MEAN** and **average percent error**, which are very similar to **model 3**'s values (as we expected). We cap off our modeling by conducting an F-test between **model 5** and reduced model using only an intercept term. The results of our test show us that our model is significantly better at explaining the variation in the dependent variable than just a constant expectation. The F-test tells us that our model has some predictive power at explaining the dependent variable that would be better than using just the expectation of **Y**.

```
# MSR of Model 5
calculate_msr(model_5)
```

```
## [1] 2.679318
```

```
# Percent Errors for Model 5
percent_error_5 <- abs(100*(predict(model_5, newdata = test) - log(test$views))/
      ifelse(predict(model_5, newdata = test) > log(test$views),
             predict(model_5, newdata = test ), log(test$views)))
# Mean Percent Difference for Model 5
mean(percent_error_5)
```

```
## [1] 8.855587
```

```
# MSRE/Mean for Model 5
calculate_rmse(model_5)/mean(log(test$views))
```

```
## [1] 0.1222141
```

```
restricted <- lm(log(views) ~ 1, data = vid)
#F-Test for Model 4 against an intercept term
anova(restricted, model_5, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: log(views) ~ 1
## Model 2: log(views) ~ log(title_length) + log(uppercase_proportion) +
##     binary_question + rating
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  28663 79560
## 2  28659 76800  4    2760.4 257.52 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7 Linear Model Results

Now that we have determined our linear model and its effectiveness, we can speak to the statistical and practical significance of the **X** variables found within it. Testing the coefficients for **model 5** shows that all **X** variables are statistically significant (with the slight exception of the binary question variable). Given the logarithmic transformation of our **Y** variable, views, the coefficients refer to percent change on the **Y** axis. All else equal, for every percent increase in character length within a title there will be about a 0.16 percent

decrease to viewership, for every percentage of uppercase proportion of letters in the title there will be an increase of 0.15 percent to viewership, if there is a question mark in the video there will be a decrease of about 7.7% to viewership, and for every percentage of likes (from total likes and dislikes) there will be an increase of 0.25% to viewership. While all are statistically significant, in a practical sense most variables do not appear to heavily contribute to viewership, excluding the binary indicator for question marks.

Final Model:

$$log(Views) = \beta_0 + \beta_1 log(TitleLength) + \beta_2 log(UppercaseProportion) + \beta_3 BinaryQuestionMark + \beta_4 Rating$$

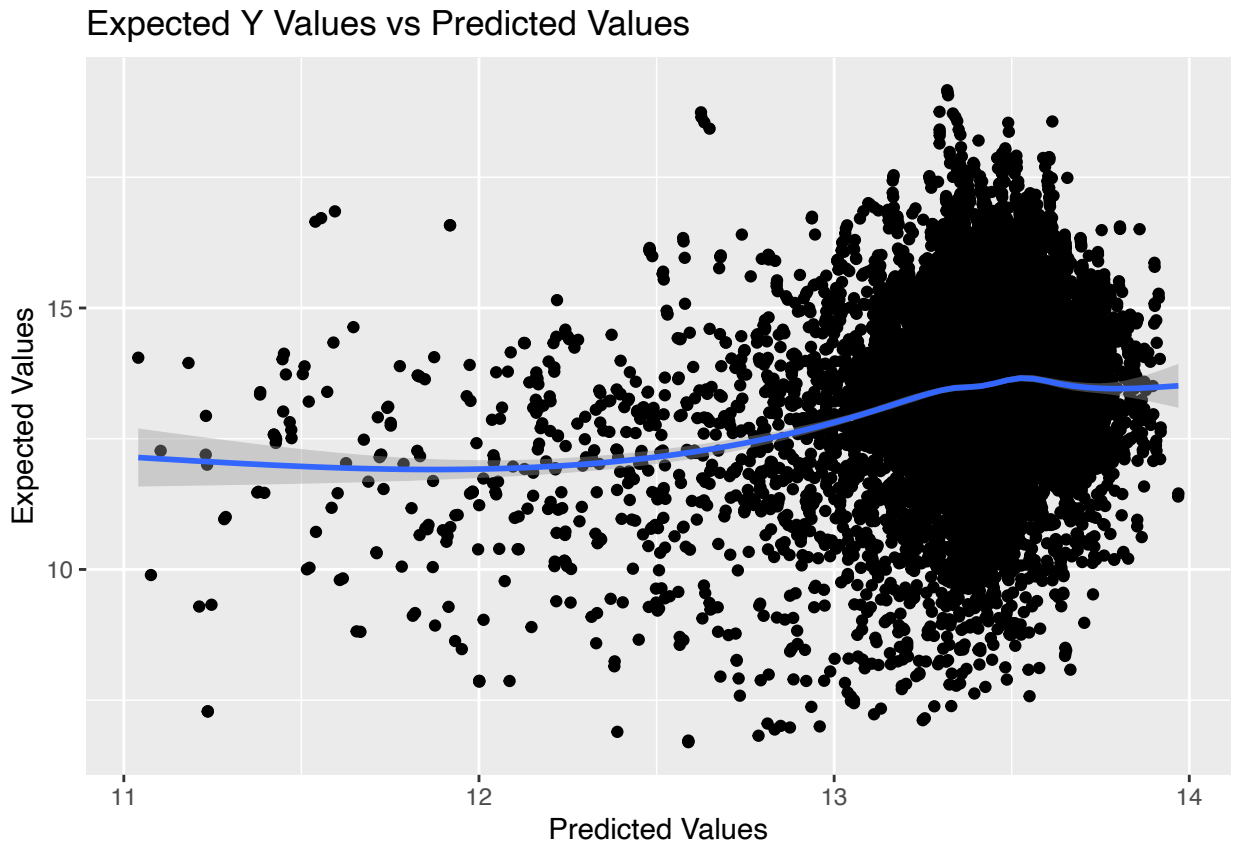
```
##
## t test of coefficients:
##
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                11.874778   0.146804 80.8884 < 2.2e-16 ***
## log(title_length)          -0.159759   0.022167 -7.2069 5.864e-13 ***
## log(uppercase_proportion)   0.153520   0.013635 11.2592 < 2.2e-16 ***
## binary_question            -0.076558   0.035456 -2.1592   0.03084 *
## rating                      2.537298   0.111836 22.6876 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Model 5
## ================================================
##                              Dependent variable:
##                            ----------------------
##                                   log(views)
## ------------------------------------------------
## log(title_length)                   -0.2***
##                                     (0.02)
##
## log(uppercase_proportion)           0.2***
##                                     (0.01)
##
## binary_question                      -0.1*
##                                     (0.04)
##
## rating                              2.5***
##                                      (0.1)
##
## Constant                            11.9***
##                                      (0.1)
##
## ------------------------------------------------
## Observations                        28,664
## R2                                   0.03
## Adjusted R2                          0.03
## Residual Std. Error            1.6 (df = 28659)
## F Statistic                257.5*** (df = 4; 28659)
## ================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

We can see from the **Expected Y Values vs Predicted Values** graph that there is a lot of variation in our expected results and that the majority of values are focused near the right extreme of the plot. Thought

the trend seems somewhat linear, the model isn't a great fit.

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



8 Omitted Variables

There are a few potential omitted variables to consider with the final model specification selected to regress on views. The three selected omitted variables are quality of the video, level of suggestiveness/incendiary material in the thumbnail, and level of suggestiveness/incendiary material in the title.

Video quality With video quality, the video quality presumably has a positive relationship with views and also with rating. Since both relationships are positive, and the measured coefficient for rating is positive, we can conclude that the direction of bias is away from zero. As this omitted variable is quite subjective in nature, as quality doesn't have a clear means of measurement, we can assume that as one example time spent on development and editing of the video can serve as a proxy. If this data were available on the YouTube platform, we would be able to more accurately predict views with the inclusion of the video quality variable.

Clickbait thumbnail For the level of suggestiveness/incendiary material in the thumbnail, or what some may refer to as "clickbait," we assume that the level of clickbait in the thumbnail has a positive correlation with views and a negative correlation with rating. Therefore, as the measured coefficient for rating is positive, the direction of bias is towards zero. This can be explained by the idea that people are more incentivized to click on the video initially, but are left disappointed by the misleading content in the thumbnail and are thus more likely to leave a dislike. If we were able to utilize a computer vision model to produce a boolean value on whether the video does or does not contain clickbait material, we would be able to produce a predictive model closer to the true regression model for views. Though this certainly is an important variable to consider, our results are likely still valid, particularly as the

context we're operating in is for trending videos selected by the YouTube algorithm, which optimizes for watch time and view satisfaction.

Clickbait title The level of suggestiveness/incendiary material in the title is likely similar in most facets to the aforementioned thumbnail variable, meaning that the omitted variable also has a direction of bias towards zero. In order to collect data for this measure, we would likely need to develop an NLP solution to help us identify such titles in context. However, it is unlikely that the current omission of this variable would have a significant impact on our results, as we have included imperfect measures of this feature through variables such as uppercase_proportion and binary_question.

9 Conclusion

While it remains to be seen whether the final specified model is the best predictor of views, our findings demonstrated that a significant relationship exists between variables relating to title syntax and number of views among trending US YouTube videos. However, it is important to note that we cannot necessarily extrapolate our results to a broader context than the data used, as our data is limited to trending YouTube videos in the US from January 2017 to May 2018. That being said, our research produced some interesting insights, as the coefficients estimated in the final model suggest a relationship with the dependent variable counter to what we anticipated in the case of the existence of a question mark. Moreover, we found that title_length, binary_question, and rating were all statistically significant after regressing views on each of the variables.