Investigating potential predictors of view count for trending Youtube videos using linear regression

Timothy Lanthier

3/12/2022

Introduction

For content creators on Youtube, their income is largely determined by the popularity of their videos. The most obvious source of income for creators is revenue from ads. The more views your video gets, the more ads are served and the more a creator might get paid. Additionally, higher view counts on videos may result in sponsorship deals for creators which serve an additional source of income. Hence, it is important for creators on Youtube to maximize the number of views that they get on videos. Of course the success of a Youtube video depends on the content in the video itself, but how much is the view count of a video associated with other aspects such as the title or time of upload?

In this paper, we will explore how certain characteristics of trending Youtube videos are associated with view count. We wish to find what aspects of a video such as length of the title or time of upload may be associated with the view count a Youtube video.

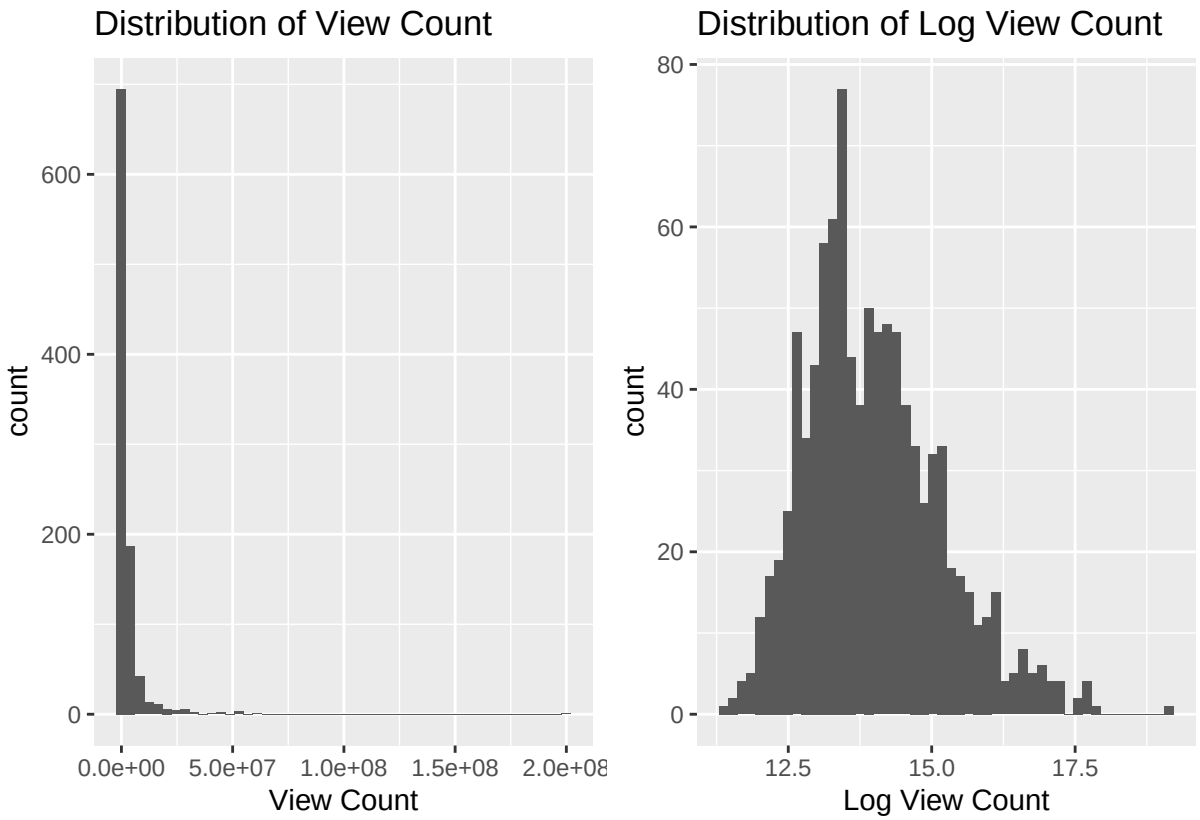
The dataset was obtained through Kaggle and includes as observations all trending videos from August of 2020 to now from a variety of regions. For this analysis, we will only be using trending videos from the US. Note that the dataset we will be using was accessed from Kaggle on February 7, 2022, so we will only include trending Youtube videos up to that date. From this dataset, we have taken a random sample of 1000 videos and have generated some additional variables from the dataset. We also used the Youtube API with the provided video and channel IDs to gain additional potential predictors such as video length and subscriber counts for the channels. Our feature generation procedure left us with a few different missing values. There were some missing subscriber counts as well as video lengths due to videos and channels no longer existing. Since there were very few of those observations, we have removed them from our dataset. Additionally, some videos had no description leading to missing values. So for those videos we imputed the description length of be 0 characters.

We also removed variables such as `likes`, `dislikes`, and `trending_age` since those values would not be observed until after the video is uploaded. Since we want to look at variables which can only be observed before the video is uploaded, those will not be important for our analysis. This leaves us with a dataset with 973 observations and the following 20 variables:

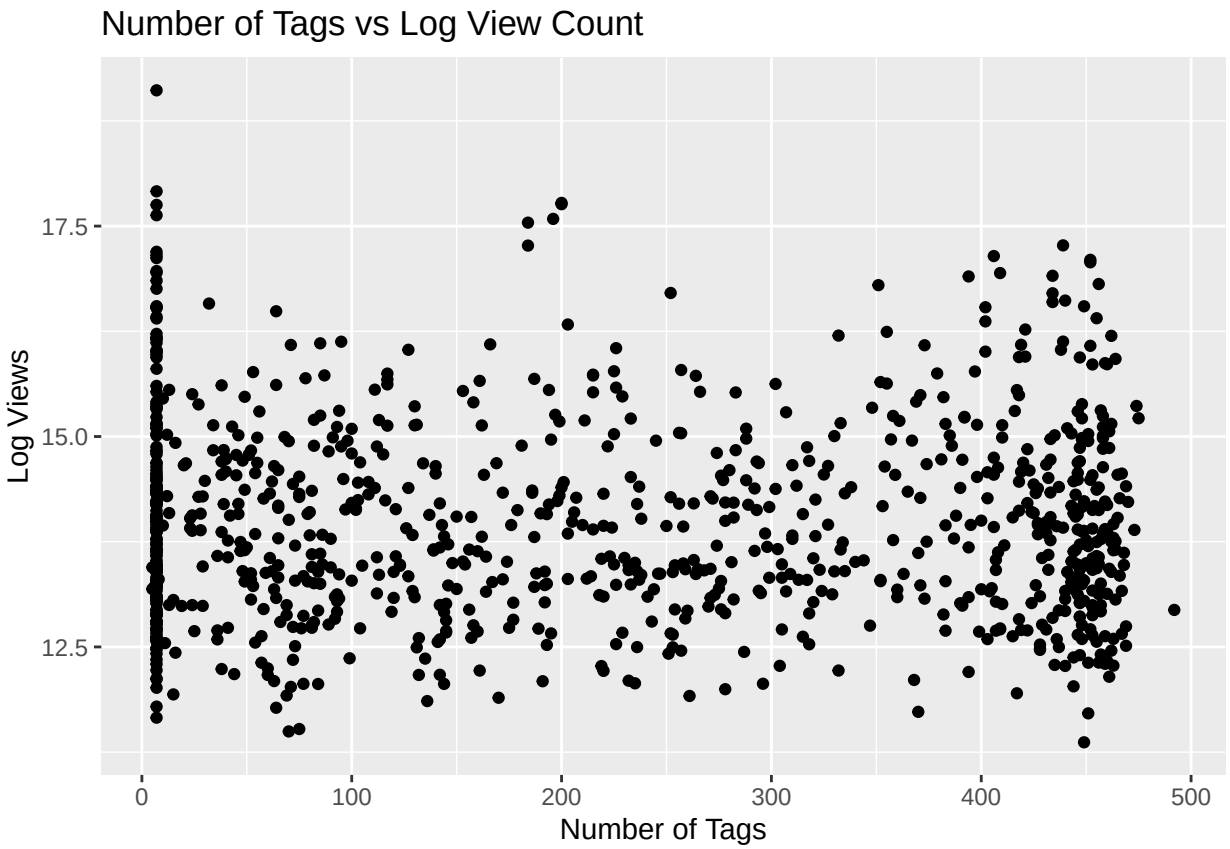
Variable	Description
publishedAt	date and time of upload (PST)
view_count	number of views
comments_disabled	comments are disabled (T/F)
ratings_disabled	ratings are disabled (T/F)
num_tags	number of tags in video
num_caps	number of capital letters in title
num_exc	number of exclamation points in title
num_qm	number of question marks in title
num_period	number of periods in title
num_dollar	number of dollar signs in title

Variable	Description
title_length	length of title in characters
desc_length	length of description in characters
channel_length	length of channel name in characters
weekday_published	weekday of video upload
day_published	day of month video was uploaded
hour_published	hour of day video was published
video_length	length of video in seconds
subscriberCount	number of subscribers on publisher's channel
videoCount	number of videos on publisher's channel
category	video category

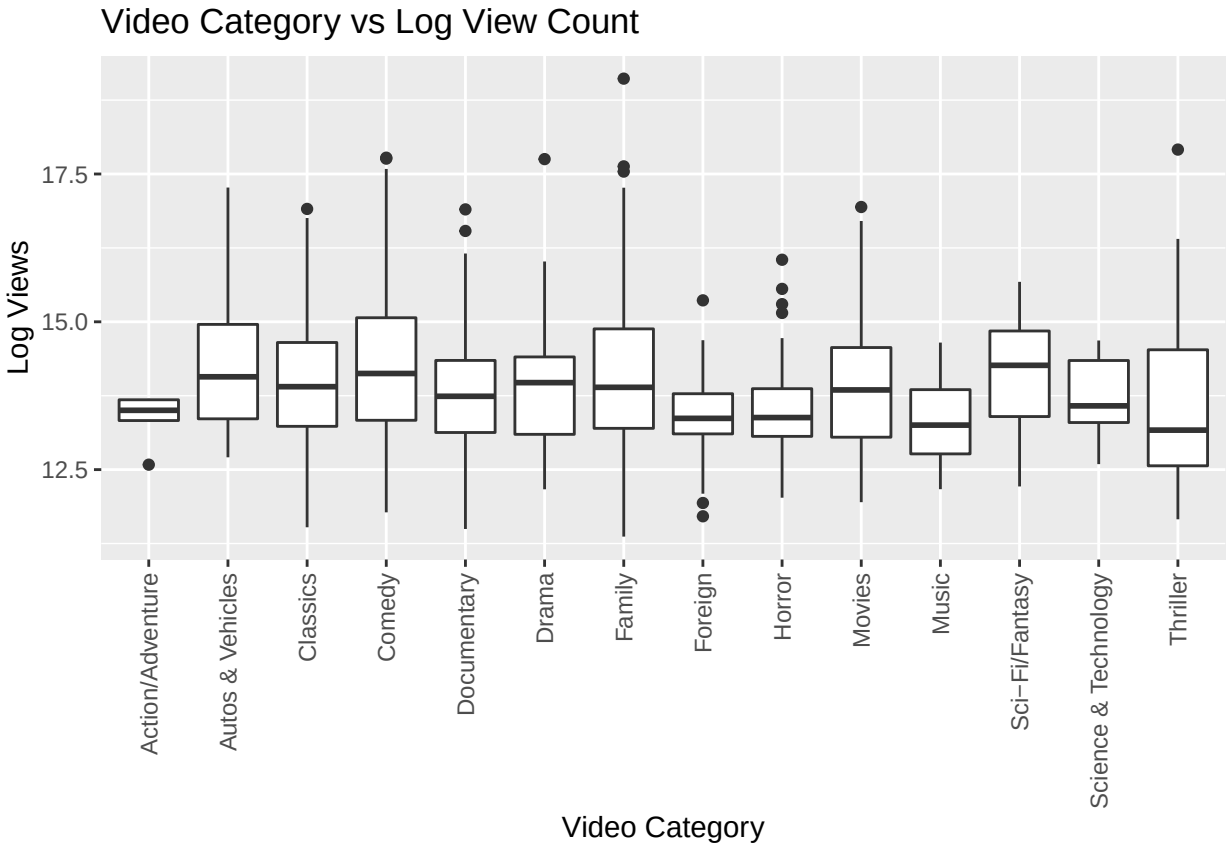
Looking at the response variable of view count, we find that the distribution for view count has a heavy positive skew. Since we wish to conduct inference on our model, we will instead be using the logarithm of view count so as to have a more normally distributed response variable. The distributions for both variables are shown below.



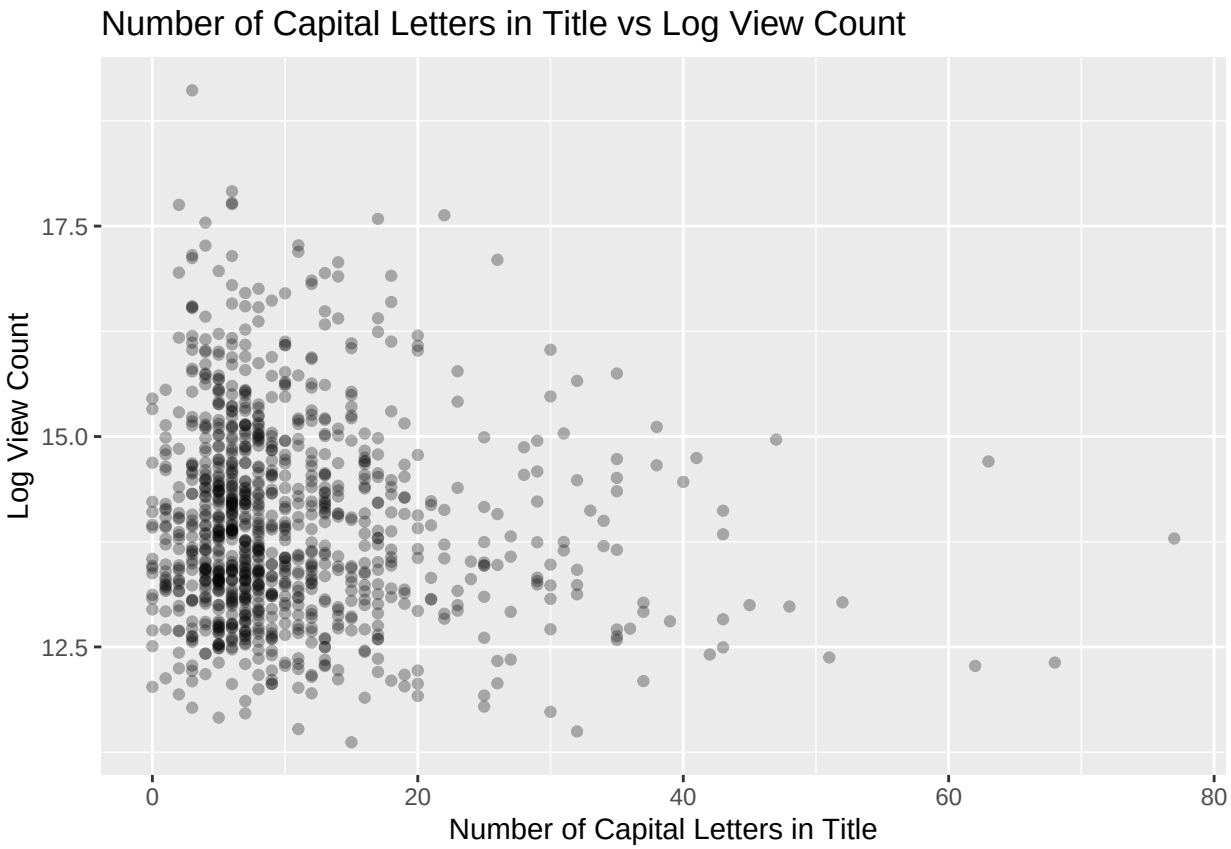
Interestingly, looking at the plots for log view count against the number of tags, we see what appears to be a completely random scatter. So there appears to be no relationship between log view count and the number of tags a video has. This is quite surprising as I would expect a large number of tags might result in the video getting suggested to more users and thus more people viewing the video.



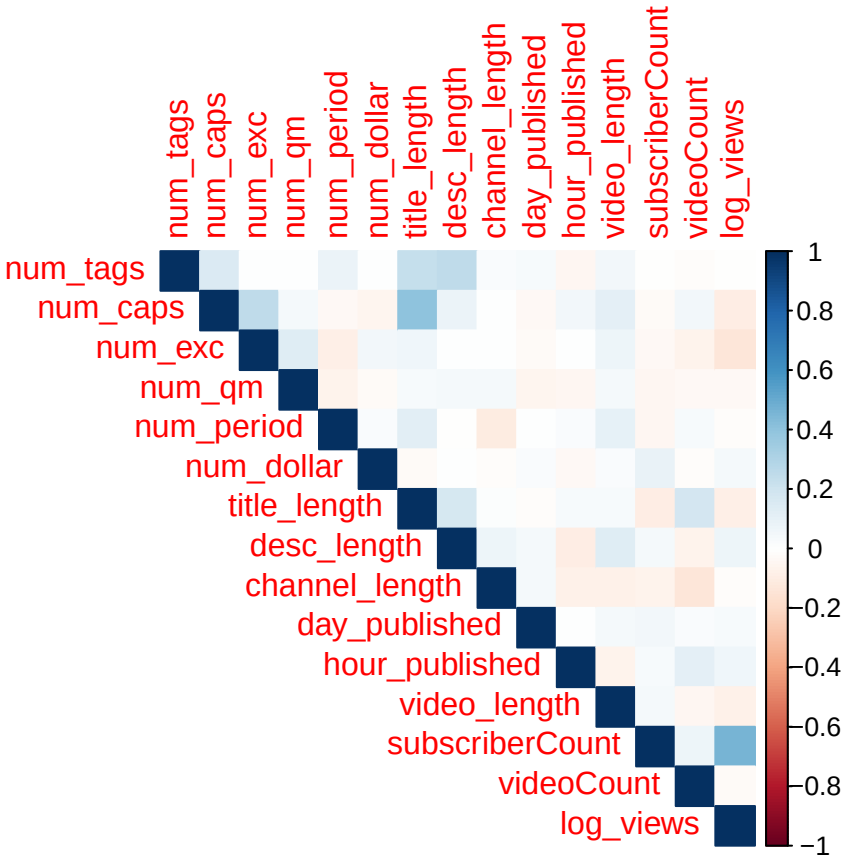
Meanwhile looking at the box plots for log view count across the categories of videos, we find that there are some large differences in log view count across the different video categories. While the variability varies between the categories, just looking at the means, we have comedy and SciFi/Fantasy videos appears to be the most popular while foreign, horror, and music videos appear to have the lowest mean log view counts. So video category may be useful in our model. What seemed unusual is that videos in the Music category have such a low mean log view count. Just based off my own observations of trending videos, it seems that music videos which hit trending tend to have pretty high view counts.



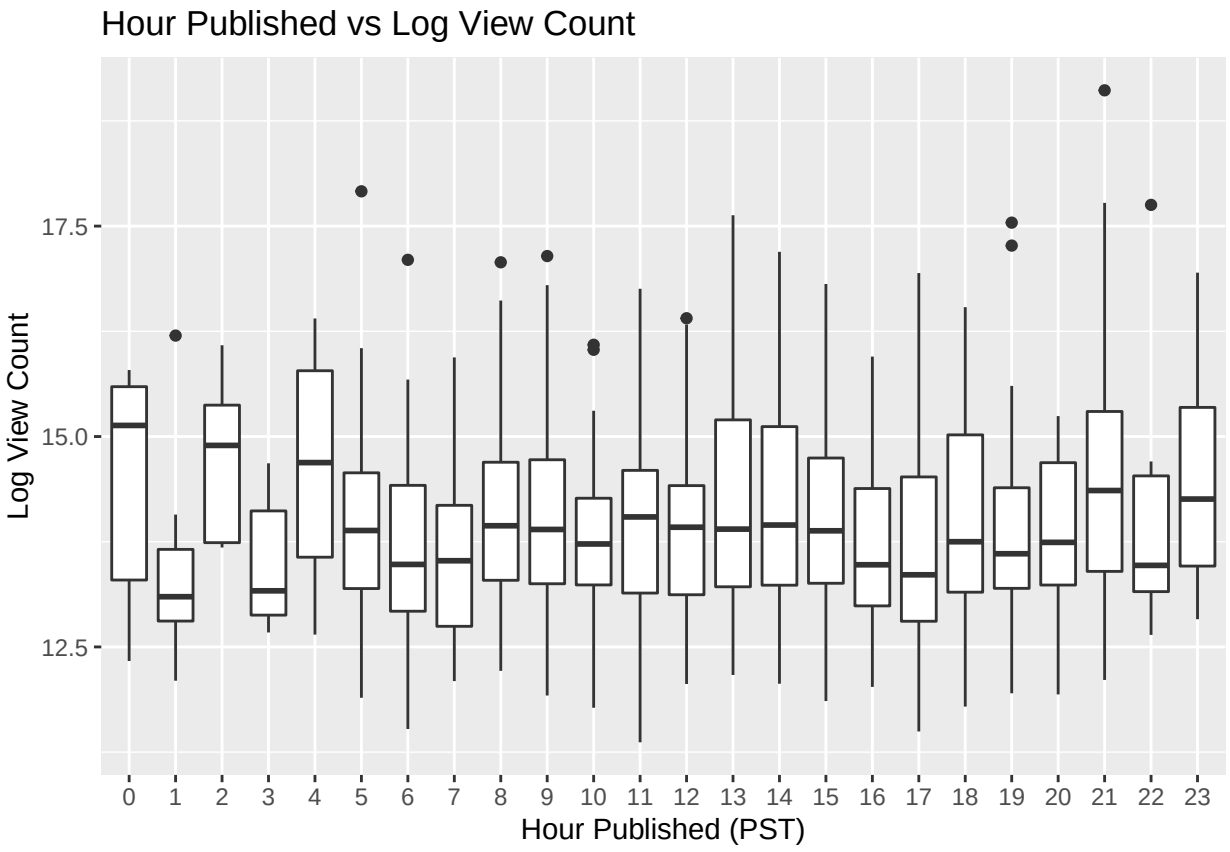
Next we will take a look at the number of capital letters in the title. The plot for the log view count against the number of capital letters in the title is shown below. As we can see, there it looks like there may be a weak negative linear relationship between log view count and the number of capital letters in the title. We will investigate this variable later.



While we will not include all of the plots from our exploratory data analysis, we identified quite a few variables that we thought may be useful for our final model. In addition to those mentioned above, these included the variables `subscriberCount`, `comments_disabled`, `ratings_disabled`, `num_caps`, `num_exc`, `num_period`, `num_dollar`, `desc_length`, `hour_published`, `weekday_published`, `video_length`.



Looking at the correlation plot shown above, we can see that `subscriberCount` has a moderate correlation with log view count. We also find that `num_caps` and `num_exc` have a weak correlation with log views. Additionally we see that the hour published has a very weak correlation. But from our exploratory data analysis, we found that there was quite a bit of variability of log view count for the different hours published so we will investigate its use further. Note that `hour_published` has a very weak relationship with log view count. But looking at the plot below, we see that there is quite a bit of variation in mean log view count across the different hours. So while there doesn't appear to be a strong linear relationship, `hour_published` may be useful as a categorical variable.



Regression Analysis

As mentioned earlier, rather than using view count as our response variable, we have decided to use log view count since we intend on conducting inference on our model. For constructing our final model, we started by creating a model using all of the variables we identified as potentially important:

`subscriberCount`, `category`, `comments_disabled`, `ratings_disabled`, `num_caps`, `num_exc`, `num_period`, `num_dollar`, `desc_length`, `hour_published`, `weekday_published`, `video_length`

We built 2 models using these variables. One model had `num_exc`, `num_period`, and `num_dollar` treated as categorical since there were so few values and the other model treated them as integer values. Our categorical model had an R^2 of 0.285 and AIC of 2842.164. Our integer model had an R^2 of 0.273 and AIC of 2836.423. Since the R^2 values between the 2 models are quite close, we decided to choose the model with the above variables treated as integers due to having a lower AIC.

Next, we checked whether or not to treat `hour_published` as numeric. As we noticed the correlation table, the correlation between the hour published and log view count is extremely weak. But we found in our data analysis that the log view count across different hours varied quite a bit, just not in a linear fashion with respect to hour.

We tested 3 different models: one with `hour_published` treated as numeric, one with `hour_published` treated as categorical with 24 categories, and one with `hour_published` treated as categorical with 4 categories, each one being a 6 hour interval. We found that our numeric model and model with 4 categories to have similar values for R^2 of around 0.274 while our model with 24 categories for `hour_published`, we got an R^2 of 0.294. However, the AIC and BIC of our last model was significantly greater than that of our other 2 models. So we decided to choose the model with the best AIC being the model with 4 categories for hour published. We called this variable with 4 categories `time_of_day`.

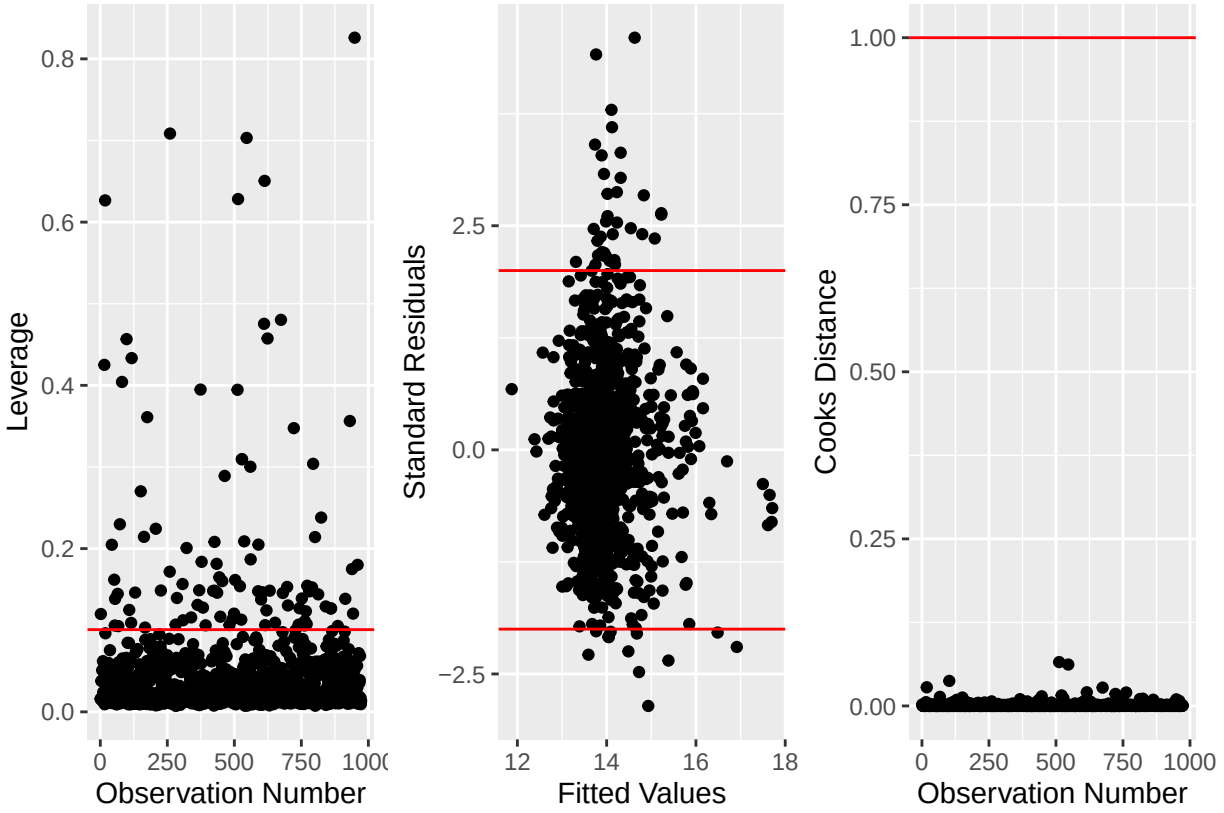
Next we ran backwards selection on this model using AIC. This gave us a model with `time_of_day`, `subscriberCount`, `category`, `num_caps`, `num_exc`, `desc_length`, and `video_length` as the best model. Just to make sure we didn't miss any important variables, we added back in some of the variables to check whether there would be an improvement, but none of the variables led to any significant increase in R^2. We noted that there may be some interactions between `category` and other variables, so we checked some potential interactions between `category` and found that interactions between `time_of_day` and `category`, `category` and `subscriberCount`, and `category` and `desc_length` to yield significant increases in R^2 for our model. We then checked all possible models with these different interaction terms and chose the model with the lowest AIC.

This gave us the final model which includes the predictor variables `time_of_day`, `subscriberCount`, `category`, `num_caps`, `num_exc`, `desc_length`, `video_length` and the interactions between `category` and both `desc_length` and `subscriberCount`. The coefficients for our final model are shown on the following page.

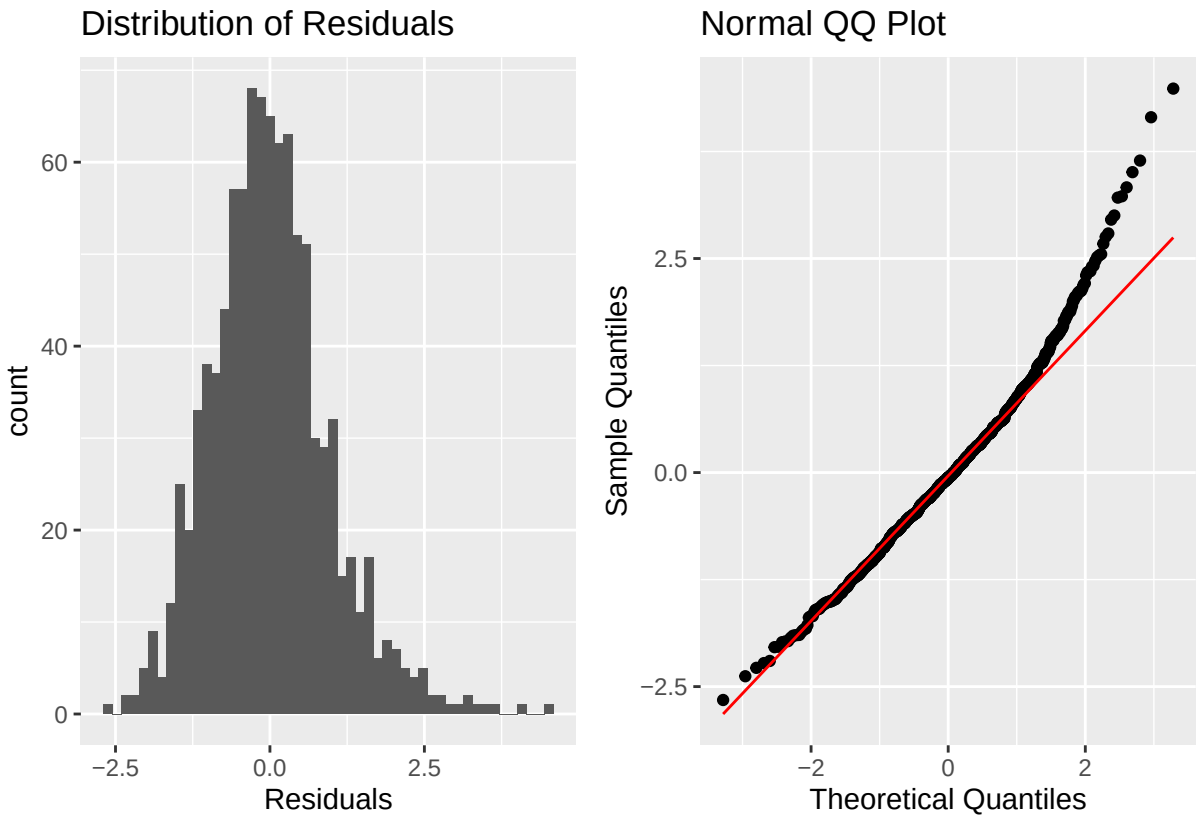
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	12.4384	1.8957	6.5615	0.0000	8.7181	16.1588
time_of_day6am_to_11am	-0.3359	0.1293	-2.5971	0.0096	-0.5896	-0.0821
time_of_day12pm_to_5pm	-0.2037	0.1346	-1.5134	0.1305	-0.4678	0.0604
time_of_day6pm_to_11pm	-0.0533	0.1471	-0.3622	0.7173	-0.3419	0.2354
subscriberCount	0.0000	0.0000	0.4889	0.6250	0.0000	0.0000
categoryAutos & Vehicles	1.7420	1.9107	0.9117	0.3621	-2.0077	5.4918
categoryClassics	1.1654	1.8968	0.6144	0.5391	-2.5571	4.8879
categoryComedy	1.4781	1.8968	0.7792	0.4360	-2.2444	5.2006
categoryDocumentary	1.4545	1.9017	0.7648	0.4446	-2.2776	5.1866
categoryDrama	1.4597	1.9070	0.7654	0.4442	-2.2829	5.2023
categoryFamily	1.9525	1.8963	1.0296	0.3035	-1.7691	5.6741
categoryForeign	0.8946	1.9106	0.4682	0.6397	-2.8550	4.6443
categoryHorror	1.4060	1.9253	0.7302	0.4654	-2.3726	5.1845
categoryMovies	1.5703	1.9000	0.8265	0.4087	-2.1584	5.2991
categoryMusic	0.6198	1.9424	0.3191	0.7497	-3.1921	4.4318
categorySci-Fi/Fantasy	1.8965	1.9170	0.9893	0.3228	-1.8657	5.6587
categoryScience & Technology	1.4210	2.0042	0.7090	0.4785	-2.5123	5.3543
categoryThriller	1.3602	1.9134	0.7109	0.4773	-2.3949	5.1152
num_caps	-0.0085	0.0039	-2.2193	0.0267	-0.0161	-0.0010
num_exc	-0.1957	0.0597	-3.2767	0.0011	-0.3130	-0.0785
desc_length	0.0002	0.0010	0.2012	0.8406	-0.0017	0.0021
video_length	-0.0001	0.0001	-2.4690	0.0137	-0.0002	0.0000
subscriberCount:categoryAutos & Vehicles	0.0000	0.0000	-0.4016	0.6881	0.0000	0.0000
subscriberCount:categoryClassics	0.0000	0.0000	-0.3898	0.6968	0.0000	0.0000
subscriberCount:categoryComedy	0.0000	0.0000	-0.4539	0.6500	0.0000	0.0000
subscriberCount:categoryDocumentary	0.0000	0.0000	-0.4213	0.6736	0.0000	0.0000
subscriberCount:categoryDrama	0.0000	0.0000	-0.3245	0.7456	0.0000	0.0000
subscriberCount:categoryFamily	0.0000	0.0000	-0.4514	0.6518	0.0000	0.0000
subscriberCount:categoryForeign	0.0000	0.0000	-0.4017	0.6880	0.0000	0.0000
subscriberCount:categoryHorror	0.0000	0.0000	-0.4340	0.6644	0.0000	0.0000
subscriberCount:categoryMovies	0.0000	0.0000	-0.4578	0.6472	0.0000	0.0000
subscriberCount:categoryMusic	0.0000	0.0000	-0.3193	0.7496	0.0000	0.0000
subscriberCount:categorySci-Fi/Fantasy	0.0000	0.0000	-0.3640	0.7160	0.0000	0.0000
subscriberCount:categoryScience & Technology	0.0000	0.0000	-0.3913	0.6957	0.0000	0.0000
subscriberCount:categoryThriller	0.0000	0.0000	-0.3946	0.6932	0.0000	0.0000
categoryAutos & Vehicles:desc_length	-0.0001	0.0010	-0.0818	0.9348	-0.0020	0.0018
categoryClassics:desc_length	0.0000	0.0010	0.0056	0.9955	-0.0019	0.0019
categoryComedy:desc_length	0.0001	0.0010	0.1243	0.9011	-0.0018	0.0020
categoryDocumentary:desc_length	-0.0002	0.0010	-0.1587	0.8739	-0.0021	0.0017
categoryDrama:desc_length	-0.0005	0.0010	-0.4828	0.6293	-0.0024	0.0014
categoryFamily:desc_length	-0.0003	0.0010	-0.3571	0.7211	-0.0022	0.0015
categoryForeign:desc_length	-0.0001	0.0010	-0.0560	0.9554	-0.0020	0.0019
categoryHorror:desc_length	-0.0002	0.0010	-0.2094	0.8341	-0.0022	0.0017
categoryMovies:desc_length	-0.0001	0.0010	-0.1138	0.9094	-0.0020	0.0018
categoryMusic:desc_length	-0.0001	0.0010	-0.0523	0.9583	-0.0020	0.0019
categorySci-Fi/Fantasy:desc_length	-0.0004	0.0010	-0.4018	0.6879	-0.0023	0.0015
categoryScience & Technology:desc_length	-0.0003	0.0010	-0.3092	0.7572	-0.0023	0.0017
categoryThriller:desc_length	-0.0004	0.0010	-0.3709	0.7108	-0.0023	0.0015

R^2	AIC	BIC
0.3396	2777.482	3016.621

Even though `subscriberCount` has a very large p-value, we found excluding it and its interactions from the models resulted in a considerable drop in R^2 of over 0.2. We also investigated whether there were any influential points in our dataset. While we found there were quite a few observations which lied outside our threshold for leverage and standardized residuals, looking at the cook's distance, we found that removing these observations would have little effect on our model. The plots for leverage, standardized residuals, and cook's distance are shown below.

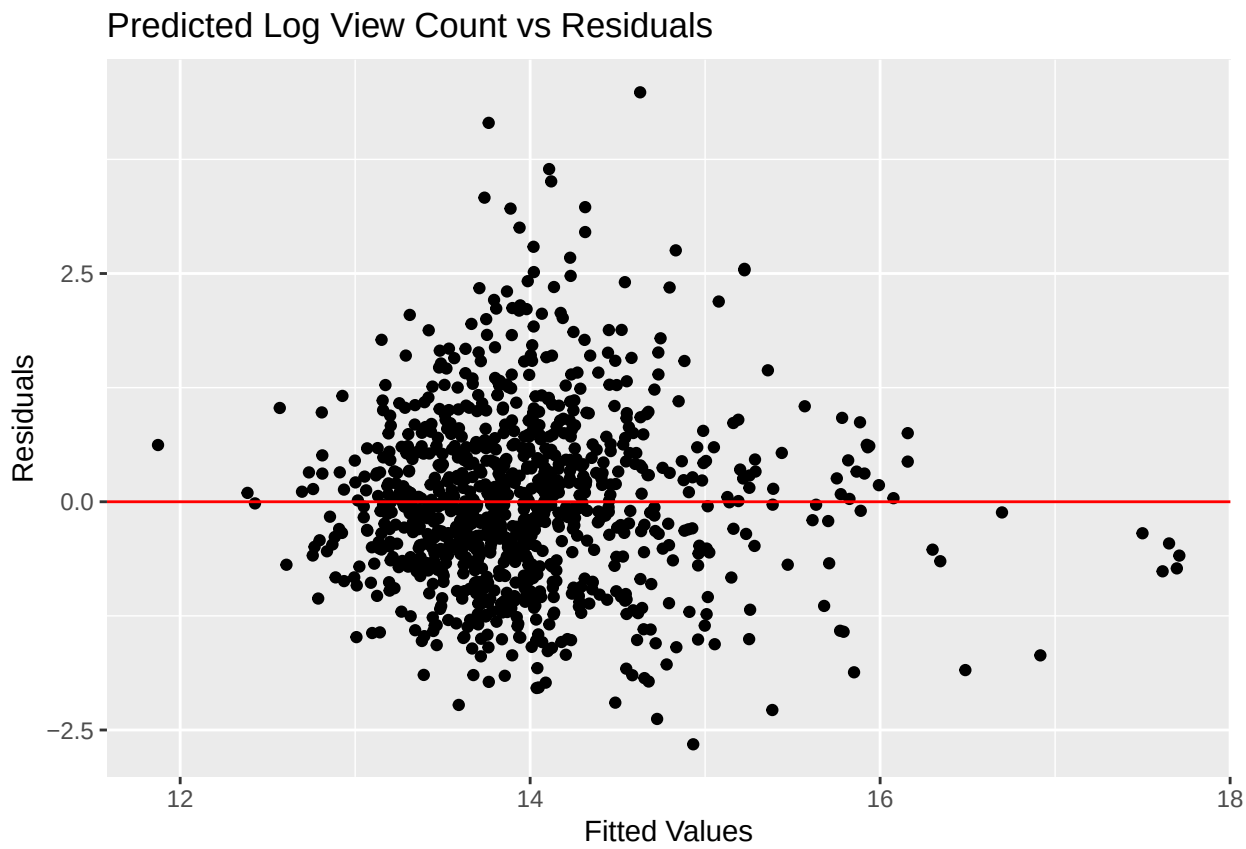


Seeing as our model is a multiple linear regression model, we have quite a few assumptions which needed to be checked if we wish to conduct inference on the coefficients. The first assumption is that our response variable is normally distributed. We intended to use `view_count` as our response variable, but seeing as `view_count` was heavily skewed, we decided to use `log_views` as our response variable. As we noted before, the distribution for `log_views` seems to be approximately normal. We also have the following distribution and QQ plot of the residuals.



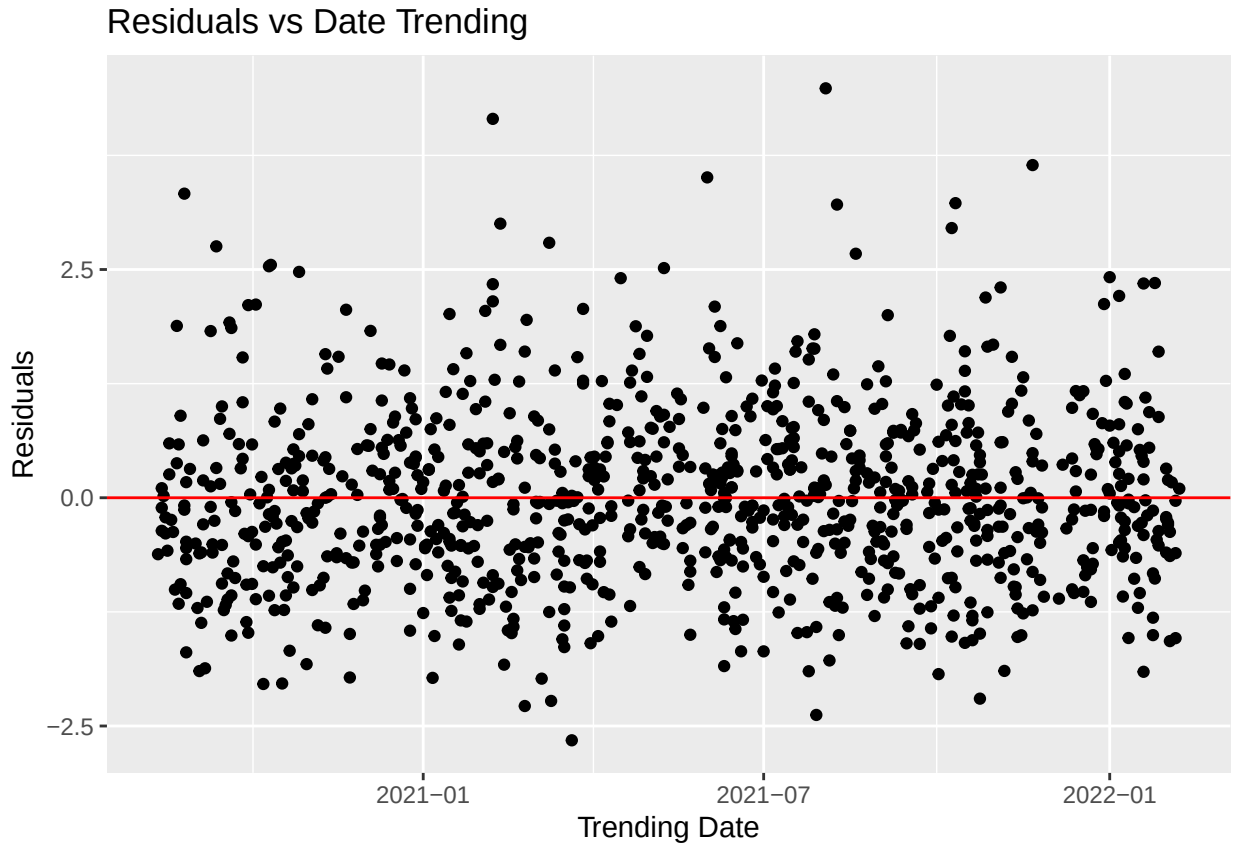
As we see the residuals seem to be normally distributed, although we have a longer right tail. This is reflected in our QQ plot as the residuals don't appear to follow the normal distribution for the larger residuals. So the normality assumption is partially satisfied.

The next assumption we need to check is that the regression variance is constant. We can check this by looking at our predictions versus the residuals shown in the plot below. As we can see, for middling predicted log view count, we have a lot of variability in the residuals. Meanwhile for very large or small predicted log views, we have considerably less variability in the residuals. Hence the constant variance assumption has been violated.



We also must check that `log_views` has a linear relationship with the predictor variables used in our model. Looking at the above plot, there isn't a clear pattern in our plot so it looks like a linear relationship may be appropriate. We also found that looks at the plots of the residuals against each of the predictors, there was no discernable pattern. The only exception was for `num_exc` where the average residuals seemed to be higher for videos with more exclamation points in the title. That being said, it seems that the linearity assumption was satisfied.

The final assumption is the independence assumption. Plotting the residuals against the time the videos were trending (and subsequently added to the dataset), we find that the residuals are randomly scattered around 0. So the independence assumption has been satisfied. Additionally,



Since the constant variance assumption has been violated and the normality assumption is only partially satisfied, it would be inappropriate to conduct inference on the coefficients of our model.

Discussion

As mentioned in the previous section, since the assumptions of our model (linearity, constant variance, independence, and normality) have not all been satisfied, we should be cautious when interpreting the coefficients of our model. Since we have so many interaction terms, interpreting the coefficients becomes quite difficult. Our model was of the form

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

where y is the view count. Since we are predicting the log view count if we wanted to interpret the coefficients in terms of the actual view count, we would need to exponentiate our model. That is

$$y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots)$$

Looking at our model, very few of the coefficients are significantly different from 0. In our final model, only the upload time of day, number of capital letters in the title, and number of exclamation points in the title seem to be statistically significant. We would also claim that subscriber count is statistically significant. Even though the confidence interval for `subscriberCount` include 0 and we have a large p-value, this only occurs due to our inclusion of interaction terms. Without the interaction terms, subscriber count is statistically significant. Additionally, if we were to remove subscriber count and its associated interaction terms from the model, our R^2 would drop from 0.3395 to 0.1134. So subscriber count is very important. While we cannot see it in our model above, the coefficient for subscriber count is a very small positive number. So holding all else constant, this means that an increase in subscriber count is associated with an increase in log view count (and subsequently an increase in view count).

While the coefficients for the interaction terms seem to be very close to 0 and aren't statistically significant, our model with those interactions removed has a much smaller R^2. Also, since subscriber counts and description lengths tend to be pretty large, while the coefficients are small the effect of these interaction terms are larger than the coefficients may indicate.

From our analysis, we find most of the variables are not very helpful in explaining the differences in log view count across videos. Most notably, subscriber count is very important. Trending videos uploaded by channels with a larger subscriber counts tend to have larger view counts. But the content creator isn't directly in control of the subscriber count at the time of uploading. Additionally the number of capital letters in the title is useful. It seems that a larger number of capital letters in the title is associated with smaller view counts. Interpreting the coefficient, holding all else constant, by increasing the number of capital letters in the title by 1 we would expect the view count of the video to be multiplied by a factor of $e^{-0.0085} = 0.992$. So avoiding titles with all capital letters would be advised if we wanted to get as many views as possible. We also found that the number of exclamation points in the title is negatively associated with the log view count (and thus the total view count). So content creators may want to avoid putting many exclamation points in the title. So it seems that titling strategies like putting titles in all caps with many exclamation points is not associated with increased viewership. Rather it should be avoided. According to our model, uploading at times between 6 am and 11 pm is associated with a decrease in log views from uploading between 12 am and 5 am. The largest decrease in log view count is uploading between 6 am and 11 am and the smallest decrease is when uploading between 6 pm and 11 pm with a decrease in log view count of only -0.053 from our baseline upload time. So the optimal upload times seem to be during late nights and very early mornings.

Limitations

There are quite a few limitations of our analysis. First off, our model is built based off of a random sample of trending videos between August of 2020 and February of 2022. Seeing that this is the case, it may be inappropriate to use conclusions based off the model in the future. If Youtube's algorithm for selecting trending videos were to change our conclusions would no longer be relevant. Furthermore the assumptions of the model were not all satisfied. Specifically, the constant variance assumption was violated and the normality assumption was only partially satisfied. Since this is the case, the findings that were discussed in the previous section should be taken with a grain of salt.

Conclusion

As we have seen, we have built a model using the predictor variables `time_of_day`, `subscriberCount`, `category`, `num_caps`, `num_exc`, `desc_length`, `video_length`, and the interactions of `subscriberCount` and `desc_length` with `category` for predicting log view count. From this model we found the time of day of upload, subscriber count, number of exclamation points in the title, and number of capital letters in the title were significant predictors of log view count. We found there was a positive association between log view count and subscriber count. Also, there was a negative association between log view count and the number of capital letters and exclamation points in the title. Finally, we found that the optimal times to maximize log view count would be during late night and early morning hours in the PST time zone. But due to not all assumptions being satisfied, all conclusions should not be taken at face value. So we cannot conclude that these are the true associations between the variables we have mentioned.

If I were to run this analysis again, the main thing I would have changed is the dataset used. Our dataset in our analysis only includes trending videos. I'd be interested to see what our conclusions would be if we were to run this analysis on a random sample of all Youtube videos. I would suspect predictors which were less important in our model such as `category` and `num_tags` would be more important in these cases.