Youtube Views Linear Regression

Corbin Couger

2/10/2023

Introduction

Throughout my childhood and young adult life, I have been fasicnated with Youtube. It is becasue there is so much entertainment on there, which has been great, because throughout my life I have gained/lost hobbies, interests, or ideas and Youtube has been there to provide help for when I get into new hobbies. I have also always known that Youtube is a profession for some of those who are on the platform. You can even make millions of dollars by just creating videos. This isn't surprising to me since the entertainment industry is huge. What is surprising is Youtubers are now making so much money that some are making as much as (or even making more) than the big name actors. The 'Youtube Algorythm' is something that has been researched a lot and many people have their different ideas and mindsets on it. Basically people have constantly been trying to crack the code about how to make a lot of money through the platform. So, What makes a Youtube video gain a lot of views? Views are a big part in what brings revenue in for Youtubers, through ads. Now yes, Youtubers make money from sponsorships, collaborations with brands, etc. but a big driver of revenue is the view count. This is the topic I will be exploring in this paper. This is interesting because it could be helpful in determining what can make a Youtube video get a lot of views, or at least answer the question. Again views = money and money is very intersting to a lot of people. Now, why is this a data science question? Well, there are plenty of reasons but the main reasons are; there are lots of data points about each Youtube video and these different variables could explain it's view count and using these variables would be a rational way to answer this topic.

Research Questions

1. What variables included in my Youtube dataset (if any) determine view count?
2. Are likes vs. dislikes a better determinating factor for more views?
3. Do the amount of views increase based on the date the video was published/went viral?
4. Does the category of the video increase or decrease view count?
5. Could I predict how many views my video would get?
6. What attributes should a Youtube content creator keep in mind to help them gain a large view count on their video?

How I plan to address this topic

Since I have a dataset containing multiple variables for thousands of trending Youtube videos, I will get into the dataset later on, I can create a multiple linear model using this data. I plan to first look at each attribute and how it correlates, or relates in general, to view count. I plan to create the model with the most significant variables and attempt to predict some cases.

How will this plan address my topic?

This plan should answer my question of: What makes a Youtube video gain a lot of views? Now, 'a lot' of views is definitely subjective, so I will determine that later on in this report, but my plan will help determine relationships view count has with other attributes a Youtube video has. There could be no relationships between these attributes and view count and that will also help me understand this question further. To make sure I do this, I will address each research question through the data and create a concise yet useful result.

Data Sets

I've explored numerous sites for datasets on Youtube videos and all are collected, researched and used in similar ways so I will give a gist of 3 datasets I have found below.

1. YouTube Trending Video Dataset (updated daily)
   https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset

This dataset is a compilation of months worth of trending Youtube videos that is separated (by files) into different regions of the world. I personally, if I use this dataset, will be utilizing the USA trending videos. This dataset some interesting variables such as; date trending, category(videos category), likes (count of likes), dislikes (count of dislikes), comments (count of comments), channel title, and whether or not the video has comments or ratings enabled/disabled. This data was a modification/recreation of a similar dataset and it was made to be easier to use. These video attributes are something Youtube keeps track of and the creator had just mined the data from Youtube.

2. YouTube Videos and Channels Metadata
   https://www.kaggle.com/datasets/thedevastator/revealing-insights-from-youtube-video-and-channe

This dataset is similar to the previous, but this one contains more info on the channels and how their views ratio with different attributes. For example variables are like; totalviews/channelelapstedtime (how many views the channel's video has based on the total run time of all the channel's videos combined), like/subscriber ration, views/subscriber ratio, subscriber count, view count, etc. This is an intersting one as it looks at some different attributes. This data was collected from Youtube and transformed to look at the different ratios between views/subribers and the other attributes. One of the most, what I presume to be, significant variables in this dataset is the totalviews/channelelapsed time. This will help me determine if the amount of time in videos determines the amount of views.

3. Youtube Dataset https://www.kaggle.com/datasets/arbazmohammad/youtube-dataset

This dataset is similar to the first as, but this is not specific to just 'trending' videos. This dataset has similar variables such as; channel name, view count on video, likes, dislikes, category, video length, etc. This dataset was collected for performing analysis on these different variables to see relationships.

Packages

Here are the packages I believe I will need, more might come along and I will address those then.

1. ggplot2 (visualizing the data)
2. ggm (possibly for the pcor() function)
3. more

R has a lot of built in functions that are already useful for the analysis I want to perform. This is a good thing as it will be nice to not need to think of which packages could help.

What types of plots or tables?

When looking at relationships I will begin by looking at the correlation matrix of the variables and see which correlate the most. I will then go into creating boxplots for each variable, see which ones have outliers, address those outliers. I will be creating histograms to check normality and then I will also be creating summary and anova tables of my model(s) I create to compare them. There will be more as I go, but those are the ones I will make sure to get created.

What I need to learn to answer my research questions?

I need to discover the relationships (if any) between the variables in the dataset I choose. I need to learn the significance of each variable and the influencial data points within the population. I will need to create a sample and learn if that sample is bias or if the sample is a good one. Many preliminary analysis will be done in this report and will be a big part of what builds into the actual modeling. In the model, I will need to learn which variables make the best model to help me determine the answers to my research questions.

Data Importing & Cleaning

The data I will be using for this project is the 3rd dataset. I am doing this because it isn't using just viral videos. This is just a set of Youtube videos and the different variabes I described. I will import this data, look at the outliers, look at the variables that might need some transforming (data types) and then turn this new dataset into a new file.

```
setwd("C:/Users/corbi/Dropbox/Masters/Winter 2022/DSC520/DSC520 R
Code/FinalProject")
raw_video_df = read.csv("videos-stats.csv")
head(raw_video_df)
```

```
##   X
## 1 0
## 2 1
## 3 2
## 4 3
## 5 4
## 6 5
##
Title
## 1                                            Apple Pay Is Killing the Physical
Wallet After Only Eight Years | Tech News Briefing Podcast | WSJ
## 2
The most EXPENSIVE thing I own.
## 3
My New House Gaming Setup is SICK!
## 4 Petrol Vs Liquid Nitrogen | Freezing Experiment |
à´ªàµ†à´Ÿàµ\215à´°àµ‹à´³à´¿à´¨àµ† à´\220à´¸àµ\215 à´†à´•àµ¾àµ»
à´ªà´±àµ\215à´±àµ\201à´®àµ‹ | M4 Tech |
## 5
Best Back to School Tech 2022!
## 6
Brewmaster Answers Beer Questions From Twitter | Tech Support | WIRED
##      Video.ID Published.At Keyword Likes Comments   Views
## 1 wAZZ-UWGVHI   2022-08-23    tech  3407      672  135612
## 2 b3x28s61q3c   2022-08-24    tech 76779     4306 1758063
## 3 4mgePWWCAmA   2022-08-23    tech 63825     3338 1564007
## 4 kXiYSI7H2b0   2022-08-23    tech 71566     1426  922918
## 5 ErMwWXQxHp0   2022-08-08    tech 96513     5155 1855644
## 6 18fwz9Itbvo   2021-11-05    tech 33570     1643  943119
```

```
str(raw_video_df)
```

```
## 'data.frame':    1881 obs. of  8 variables:
##  $ X           : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Title       : Factor w/ 1854 levels "$1 Burger vs $10,000 Burger!",..:
226 1565 1189 1289 328 390 1496 854 67 1649 ...
##  $ Video.ID    : Factor w/ 1869 levels "--hxd1CrOqg",..: 1637 401 203 981
607 107 910 159 1660 1350 ...
```

```
##  $ Published.At: Factor w/ 757 levels "2007-07-16","2007-12-11",..: 756
757 756 756 741 525 689 740 547 716 ...
##  $ Keyword     : Factor w/ 41 levels "animals","apple",..: 38 38 38 38 38
38 38 38 38 38 ...
##  $ Likes       : num  3407 76779 63825 71566 96513 ...
##  $ Comments    : num  672 4306 3338 1426 5155 ...
##  $ Views       : num  135612 1758063 1564007 922918 1855644 ...
```

```r
dim(raw_video_df)
```

```
## [1] 1881    8
```

```r
summary(raw_video_df)
```

```
##        X       
##  Min.   :   0  
##  1st Qu.: 470  
##  Median : 940  
##  Mean   : 940  
##  3rd Qu.:1410  
##  Max.   :1880  
##                
##                
Title
##  What is Machine Learning?
:   4
##  ASMR MUKBANG ì§\201ì ‘ ë§Œë“  ì–‘ë…\220 ì¹\230í‚¨ë¨¹ë°©!
ìŠ¤í…Œì\235´í\201¬ ì§œíŒŒê²Œí‹° ë \210ì‹œí”¼ &amp; ë¨¹ë°© FRIED CHICKEN AND
BLACK BEAN NOODLES EATING SOUND!                    :   3
##  TiÃ«sto - The Business (Lyrics)
:   3
##  20 Minecraft Block Facts You Maybe Didn&#39;t Know
:   2
##  ASMR Gaming ðŸ\230´ Fortnite 1 Kill = 1 Trigger Relaxing Mouth Sounds
ðŸŽ®ðŸŽ§ Controller Sounds + Whispering ðŸ’¤
:   2
##  ASMR MUKBANG ì§\201ì ‘ ë§Œë“  íƒ\200í‚¤ìŠ¤ ëŒ\200ì\231• ê°\200ëž\230ë–¡
ë–¡ë³¶ì\235´ ë¶\210ë‹ë³¶ì\235Œë©´ ì¹\230ì¦\210ìŠ¤í‹± í•«ë\217„ê·¸ ë¨¹ë°©
&amp; ë \210ì‹œí”¼ FIRE NOODLES AND Tteokbokki EATING SOUND!:   2
##  (Other)
:1865
##         Video.ID        Published.At              Keyword    
##  2FYvHn12pOQ:   2   2022-08-24: 288   asmr            :  50  
##  4mgePWWCAmA:   2   2022-08-23: 183   cnn             :  50  
##  5q87K1WaoFI:   2   2022-08-22:  39   crypto          :  50  
##  7eh4d6sabA0:   2   2022-08-20:  38   cubes           :  50  
##  96mrgd8-3yE:   2   2022-08-21:  35   data science    :  50  
##  kkOweffr3II:   2   2022-08-17:  23   game development:  50  
##  (Other)    :1869   (Other)   :1275   (Other)         :1581  
##      Likes             Comments          Views          
##  Min.   :      -1   Min.   :    -1   Min.   :2.500e+01  
##  1st Qu.:    2672   1st Qu.:   199   1st Qu.:8.452e+04
##  Median :   14787   Median :   814   Median :5.917e+05
##  Mean   :  170061   Mean   :  7863   Mean   :1.161e+07
##  3rd Qu.:   60906   3rd Qu.:  3378   3rd Qu.:2.805e+06
##  Max.   :16445558   Max.   :732818   Max.   :4.034e+09
##  NA's   :2          NA's   :2        NA's   :2
```

```
# Looking at the data and looking at my questions I've introduced, I am able
to determine now what I think to be the variables I will use for my analysis.
I am going to remove the first two columns (X, and Title) these two wont be
needed as I can always reference the video title with the video ID. I also
see that Likes and Comments have a minimum of -1, that just doesn't seem
possible to have -1 comments or likes so I will remove those values, These
two variables along with views have 2 NAs, those could end up being the same
rows so I will take out those as well.

raw_video_df = subset(raw_video_df, raw_video_df$Likes > -1)
raw_video_df = subset(raw_video_df, raw_video_df$Comments > -1)

# Looks like doing these two subsets got rid of the NAs as well. Now to
removing the first 3 columns, then adding in a variable that will have a
numeric version of 'keyword'. I will also convert the Published.At variable
from a date variable to the, day of week (dow).

kwn = c(as.numeric(raw_video_df$Keyword))

raw_video_df = raw_video_df[,4:8]

raw_video_df = cbind(raw_video_df, kwn)

require(lubridate)
```

```
## Loading required package: lubridate
```

```
## Warning: package 'lubridate' was built under R version 3.6.3
```

```
## 
## Attaching package: 'lubridate'
```

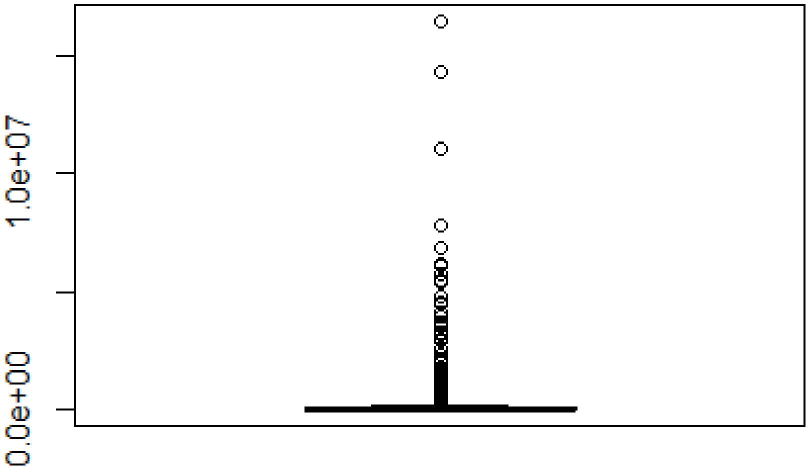
```
## The following objects are masked from 'package:base':
## 
##     date, intersect, setdiff, union
```

```
dow = wday(raw_video_df$Published.At)
```

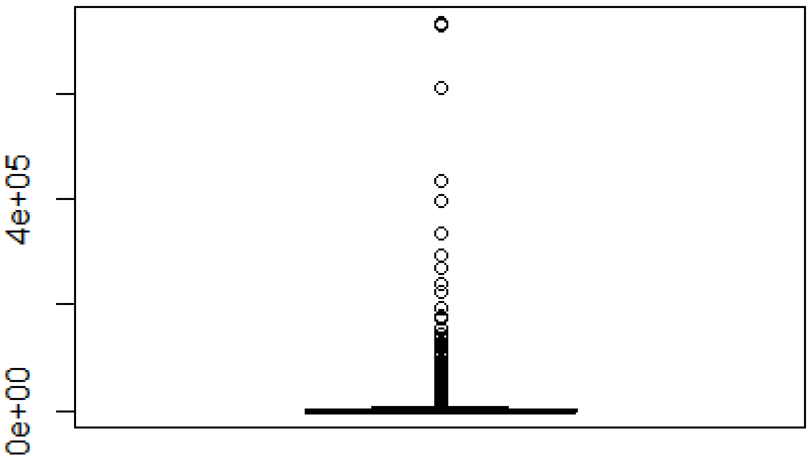
```
## Warning: tz(): Don't know how to compute timezone for object of class
## factor; returning "UTC". This warning will become an error in the next
## major version of lubridate.
```

```
raw_video_df = cbind(raw_video_df, dow)
# Now I will take a look at the outliers for Likes, Comments and Views
```

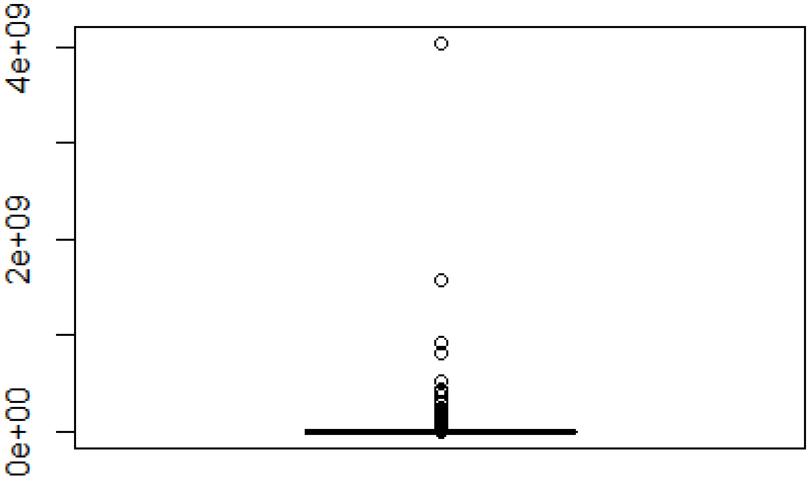
```r
boxplot(raw_video_df$Likes)
```



```r
boxplot(raw_video_df$Comments)
```



```r
boxplot(raw_video_df$Views)
```



```
# I can see that each of these variables have outliers. I will go ahead and
keep them in because for the analysis I want to perform, a linear regression
model, I want to see the extremes and see how they affect the model.

# So what does the new dataset look like?
head(raw_video_df)

##   Published.At Keyword Likes Comments   Views kwn dow
## 1   2022-08-23    tech  3407      672  135612  38   3
## 2   2022-08-24    tech 76779     4306 1758063  38   4
## 3   2022-08-23    tech 63825     3338 1564007  38   3
## 4   2022-08-23    tech 71566     1426  922918  38   3
## 5   2022-08-08    tech 96513     5155 1855644  38   2
## 6   2021-11-05    tech 33570     1643  943119  38   6

dim(raw_video_df)

## [1] 1870    7

summary(raw_video_df)

##      Published.At              Keyword         Likes
##  2022-08-24: 287   asmr            :  50   Min.   :       0
##  2022-08-23: 181   cnn             :  50   1st Qu.:    2697
##  2022-08-22:  39   crypto          :  50   Median :   14942
##  2022-08-20:  37   cubes           :  50   Mean   :  170869
##  2022-08-21:  35   data science    :  50   3rd Qu.:   61074
##  2022-08-17:  23   game development:  50   Max.   :16445558
##  (Other)   :1268   (Other)         :1570
##     Comments            Views                kwn             dow
##  Min.   :     0.0   Min.   :2.500e+01   Min.   : 1.00   Min.   :1.000
##  1st Qu.:   200.2   1st Qu.:8.469e+04   1st Qu.:11.00   1st Qu.:3.000
##  Median :   819.0   Median :5.908e+05   Median :21.00   Median :4.000
##  Mean   :  7889.8   Mean   :1.166e+07   Mean   :21.22   Mean   :3.966
##  3rd Qu.:  3402.5   3rd Qu.:2.809e+06   3rd Qu.:32.00   3rd Qu.:5.000
##  Max.   :732818.0   Max.   :4.034e+09   Max.   :41.00   Max.   :7.000
##

#Now I will save this new dataset as a new file

write.csv(raw_video_df, file = "ytvideodf.csv")
```

The Plan

Now that I've taken a quick look at the dataset and did some inital cleaning, I can now reflect on the questions I've came up with and discuss how I plan to answer those questions. Like I mentioned, I want to discover which of these attributes determine if a video gets more views, or less. To do this I will be looking at the 'views' variable and its relationship with the other variables through correlation and creating a linear regression model. So far, the information is not self-evident and I believe this type of analysis would get me the answers I'm looking for. Along with linear regresion, I could also determine a number for 'a lot' of views and create a variable that says whether or not (True or False) a video has 'a lot' of views. This could be thrown into some logistic regression and then determine relationships that way for gaining 'a lot' of views.

If I go the logistic regression route, I would create that new variable and use it as my dependent variable. I think it would be nice to possibly do both a linear and logistic regression models to see how they might differ. The one challenge is determining what 'a lot' of views is. With whatever route I decide to go, I will dice the data into Train (80% of the data) and Test (other 20% of the data). I will create the model with the train, and use the test, to test my predictions of the model.

After creating the model, I'm going to summarize the model into a summary table to determine which variables (if any) are significant enough to generating a high view count. Looking at the model of the data in this way will help me answer my questions. I will be creating a correlation matrix to look at how the correlations of the variables come out. In my tests, I will be using a confusion matrix to determine my model's accuracy. There will be few visuals I create regarding the modeling, but in preliminary analysis, I will be looking at how views, graphically, relate to other variables.

So, that's the plan! Next, I will create a linear regression model and look through its summary and find the most significant variables as well as its accuracy. Then I might dive into the logistic regression of views, the only set back is determing what to set the 'a lot' of views number at.

```
setwd("C:/Users/corbi/Dropbox/Masters/Winter 2022/DSC520/DSC520 R
Code/FinalProject")
df = read.csv("ytvideodf.csv")
set.seed(111111)
# First I'll create a train and test data sets.
require(caTools)

## Loading required package: caTools

## Warning: package 'caTools' was built under R version 3.6.3

split = sample.split(df, SplitRatio = .8)
train = subset(df, split == "TRUE")
test = subset(df, split == "FALSE")
head(train)

##   X Published.At Keyword  Likes Comments   Views kwn dow
## 1 1   2022-08-23    tech   3407      672  135612  38   3
## 2 2   2022-08-24    tech  76779     4306 1758063  38   4
## 4 4   2022-08-23    tech  71566     1426  922918  38   3
## 6 6   2021-11-05    tech  33570     1643  943119  38   6
## 7 7   2022-06-13    tech 135047     9367 5937790  38   2
## 8 8   2022-08-07    tech 216935    12605 4782514  38   1
```

```
# Now I will look at the dimensions of both the train and test dataframes
dim(train)

## [1] 1402    8

dim(test)

## [1] 468   8

# Now I will look at each of the Pearson's correlation for each of the
variables relationships.
cor(train[,4:8], method = 'pearson')

##                  Likes   Comments       Views          kwn         dow
## Likes     1.000000000 0.87033846  0.80493512 -0.006499246  0.06611609
## Comments  0.870338462 1.00000000  0.69025901  0.011579045  0.07205661
## Views     0.804935117 0.69025901  1.00000000 -0.036506945  0.01884325
## kwn      -0.006499246 0.01157904 -0.03650694  1.000000000 -0.02861610
## dow       0.066116088 0.07205661  0.01884325 -0.028616095  1.00000000

# Looking at this correlation matrix, I see that Views and Likes have a high
correlation, as well as the Comments and Likes have a high correlation. This
makes sense though, if a video gains more views, more people are inclined to
like.

# Now I will create a linear regression model using views as the dependent
variable.

ytlm = lm(Views ~ Likes + Comments + Keyword + dow, train)
# Now I will look at the summary of my model:
summary(ytlm)

##
## Call:
## lm(formula = Views ~ Likes + Comments + Keyword + dow, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -573797689    -407430    2492390    7448561 1710283230
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.275e+07  1.332e+07  -0.957 0.338864
## Likes                    1.412e+02  4.610e+00  30.626  < 2e-16 ***
## Comments                -6.077e+01  9.251e+01  -0.657 0.511374
## Keywordapple             8.759e+06  1.716e+07   0.511 0.609765
## Keywordasmr              1.216e+07  1.661e+07   0.732 0.464477
## Keywordbed              -7.797e+06  1.680e+07  -0.464 0.642740
## Keywordbiology           1.063e+07  1.681e+07   0.632 0.527416
## Keywordbusiness          1.168e+07  1.679e+07   0.696 0.486731
## Keywordchess             1.197e+07  1.681e+07   0.712 0.476586
## Keywordcnn              1.520e+07  1.659e+07   0.916 0.359686
## Keywordcomputer science  1.177e+07  1.672e+07   0.704 0.481394
## Keywordcrypto            1.408e+07  1.664e+07   0.846 0.397456
## Keywordcubes            -1.613e+07  1.652e+07  -0.976 0.329010
## Keyworddata science      1.386e+07  1.662e+07   0.834 0.404632
## Keywordeducation         1.059e+07  1.999e+07   0.530 0.596234
## Keywordfinance           1.445e+07  1.743e+07   0.829 0.407384
## Keywordfood              7.059e+06  1.672e+07   0.422 0.672949
## Keywordgame development  1.321e+07  1.662e+07   0.795 0.426786
## Keywordgaming            1.210e+07  1.706e+07   0.710 0.478041
## Keywordgoogle            6.236e+07  1.696e+07   3.678 0.000245 ***
## Keywordhistory          -1.205e+07  1.656e+07  -0.728 0.466997
## Keywordhow-to           -6.570e+05  1.668e+07  -0.039 0.968593
## Keywordinterview         7.564e+06  1.662e+07   0.455 0.649149
## Keywordliterature        1.390e+07  1.693e+07   0.821 0.411690
## Keywordlofi              7.733e+06  1.729e+07   0.447 0.654816
## Keywordmachine learning  1.129e+07  1.671e+07   0.676 0.499276
## Keywordmarvel           -7.111e+06  1.661e+07  -0.428 0.668601
## Keywordmathchemistry     3.884e+06  2.350e+07   0.165 0.868762
## Keywordminecraft         6.968e+06  1.652e+07   0.422 0.673319
## Keywordmovies            1.539e+07  1.730e+07   0.889 0.374103
## Keywordmrbeast          -1.975e+08  1.686e+07 -11.718  < 2e-16 ***
## Keywordmukbang           7.659e+06  1.704e+07   0.450 0.653081
## Keywordmusic            -1.085e+06  1.675e+07  -0.065 0.948364
## Keywordnews              1.490e+07  1.759e+07   0.847 0.396888
## Keywordnintendo          1.163e+07  1.673e+07   0.695 0.487191
## Keywordphysics           5.346e+06  1.650e+07   0.324 0.746025
## Keywordreaction          3.220e+06  1.651e+07   0.195 0.845448
## Keywordsat               1.419e+07  1.672e+07   0.848 0.396368
## Keywordsports            1.317e+07  1.673e+07   0.787 0.431351
## Keywordtech              1.073e+07  1.673e+07   0.642 0.521303
## Keywordtrolling          6.543e+06  1.661e+07   0.394 0.693739
## Keywordtutorial         -4.437e+05  1.660e+07  -0.027 0.978680
## Keywordxbox             1.076e+07  1.662e+07   0.648 0.517352
## dow                     -6.433e+05  1.013e+06  -0.635 0.525419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65770000 on 1358 degrees of freedom
## Multiple R-squared:  0.7195, Adjusted R-squared:  0.7106
## F-statistic:    81 on 43 and 1358 DF,  p-value: < 2.2e-16
```

Overall, write a coherent narrative that tells a story with the data as you complete this section. Summarize the problem statement you addressed. Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented). Summarize the interesting insights that your analysis provided. Summarize the implications to the consumer (target audience) of your analysis. Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

Results, Interpretations & Conclusion

Looking at my Linear Regression results, I can see that a few of the variables are very significant in predicting view count. The model also produced and Adjusted R-squared of .7106. This to me isn't bad! It is basically saying that my model can explain about 71% of the variance in the way Views is predicted. This to me would mean a lot for creating a Youtube video and definitely helps further the answer to my question of: What makes a Youtube video gain a lot of views? This model, though it isn't perfect, also tells me what keywords are best to throw into my video's Title or description.

Overall, my methodology for this project was to create a non-time series regression model that would help me predict the number of views my Youtube video would get. However, I say 'would get' this means I'm assuming the person who uses these results would have to have an ongoing Youtube channel where they have a good sample of videos. They would need to know their average number of likes and comments. They would then need to plug it into this model, using the significant variables, and again this model is not perfect and using the averages from their channel would also decrese the accuracy. So, yes it might be an interesting way, but it is doable with what I have accomplished with this anaylsis.

These implications are endless, and at some point, if I had my own Youtube channel, I would create a model with just my videos and compare it to the model of someone like Mr. Beast (a huge, very successful Youtuber) and see where the differences are and how I could improve. This is where it gets limiting, like I mentioned my assumptions were that someone would need a Youtube channel with a good sample of videos, and I also would like to see this model re-run with some new variables. I would want to know the videos 'duration', and I would take the 'Title' variable I initially removed and count the number of characters in it to get a variable that would show if the number of words/letters in the Title would impact view count.

These are just some final remarks as this analysis and project come to a close. The next steps for this would be for someone to use this model, perfect it by using the significant variables and tweaking the model until the most accurate one was created. This overall was a very fun project to do and am excited I was able to pull some insightful and interesting things from this data set. Youtube is incredible and being able to peek behind the curtain on a sample of its data was useful, getting access to that whole world would make it very easy to get the exact answer to my question.