Youtube Most Liked Videos

George Cruz

12/5/2020


```
library(tidyverse)
library(scales)
library(infer)
library(psych)
library(httr)
library(jsonlite)
library(here)
library(car)
```

Introduction

youtube logo

YouTube is an American online video-sharing platform headquartered in San Bruno, California. Three former PayPal employees—Chad Hurley, Steve Chen, and Jawed Karim—created the service in February 2005. Google bought the site in November 2006 for US$1.65 billion; YouTube now operates as one of Google's subsidiaries.

YouTube allows users to upload, view, rate, share, add to playlists, report, comment on videos, and subscribe to other users. It offers a wide variety of user-generated and corporate media videos. Available content includes video clips, TV show clips, music videos, short and documentary films, audio recordings, movie trailers, live streams, and other content such as video blogging, short original videos, and educational videos.

Because a video's popularity influences on the amount of money its creators make in the platform, it is of interest to determine if any relationship exists between a video's category and its popularity. Other relationships might also be explored.

The Data

The Data Set was obtained from Kaggle. This dataset was collected using the YouTube API.

Loading the Data.

```
#Get the videos csv
raw_video_df <-
        read_csv(file="https://raw.githubusercontent.com/georg4re/ds606/main/data/USvideos.csv",quote
        = "\"")
```

```
#get the categories JSON
url <-
        paste("https://raw.githubusercontent.com/georg4re/ds606/main/data/US_category_id.json",
         sep="")
res <- GET(url)
data <- fromJSON(rawToChar(res$content))

category_df <- data$items %>%
  flatten(.) %>%
  rename(category=snippet.title)
```

Joining the data and the Categories

Because the Categories are provided in a separate JSON, we need to join them to the data frame, I will take the opportunity to remove several variables not needed in this study:

- trending_date
- channel_title
- publish_time
- comments_disabled *(although this could be used to study the relationship between the comments being enabled or disabled and the amount of likes, it is out of the scope of this endeavor.)*
- ratings_disabled
- video_error_or_removed

```
category_df <- category_df %>%
  rename(category_id = id)
category_df$category_id <- as.numeric(category_df$category_id)
```

```
video_df <- raw_video_df %>%
  left_join(category_df) %>%
  select(video_id,
         title,
         category,
         tags,
         views,
         likes,
         dislikes,
         comment_count
         )
```

A snippet

```
glimpse(video_df)
```

```
## Rows: 40,949
## Columns: 8
## $ video_id      <chr> "2kyS6SvSYSE", "1ZAPwfrtAFY", "5qpjK5DgCt4", "puqaWrE...
## $ title         <chr> "WE WANT TO TALK ABOUT OUR MARRIAGE", "The Trump Pres...
## $ category      <chr> "People & Blogs", "Entertainment", "Comedy", "Enterta...
## $ tags          <chr> "SHANtell martin", "last week tonight trump presidenc...
## $ views         <dbl> 748374, 2418783, 3191434, 343168, 2095731, 119180, 21...
## $ likes         <dbl> 57527, 97185, 146033, 10172, 132235, 9763, 15993, 236...
## $ dislikes      <dbl> 2966, 6146, 5339, 666, 1989, 511, 2445, 778, 119, 136...
## $ comment_count <dbl> 15954, 12703, 8181, 2146, 17518, 1434, 1970, 3432, 34...
```

```
knitr::kable(head(video_df,10))
```

video_id	title	category	tags
2kyS6SvSYSE	WE WANT TO TALK ABOUT OUR MARRIAGE	People & Blogs	SHANtell martin
1ZAPwfrtAFY	The Trump Presidency: Last Week Tonight with John Oliver (HBO)	Entertainment	last week tonight trump presidency"|"last week tonight donald trum
5qpjK5DgCt4	Racist Superman | Rudy Mancuso, King Bach & Lele Pons	Comedy	racist superman"|"rudy"|"mancuso"|"king"|"bach"|"racist"|"superm video"|"iphone x by pineapple"|"lelepons"|"hannahstocking"|"rudymancuso"|"inanna"|' My Driver's License | Lele Pons

video_id	title	category	tags
puqaWrEC7tY	Nickelback Lyrics: Real or Fake?	Entertainment	rhett and link"|"gmm"|"good mythical morning"|"rhett and link good morning"|"Season 12"|"nickelback lyrics"|"nickelback lyrics real or fa nickelback"|"gmm nickelback"|"lyrics (website category)"|"nickelbacl kroeger"|"canada"|"music (industry)"|"mythical"|"gmm challenge"|"
d380meD0W0M	I Dare You: GOING BALD!?	Entertainment	ryan"|"higa"|"higatv"|"nigahiga"|"i dare you"|"idy"|"rhpc"|"dares"|"
gHZ1Qz0KiKM	2 Weeks with iPhone X	Science & Technology	ijustine"|"week with iPhone X"|"iphone x"|"apple"|"iphone"|"iphone
39idVpFF7NQ	Roy Moore & Jeff Sessions Cold Open - SNL	Entertainment	SNL"|"Saturday Night Live"|"SNL Season 43"|"Episode 1730"|"Tiffany Sessions"|"Kate McKinnon"|"s43"|"s43e5"|"episode 5"|"live"|"new y night"|"host"|"music"|"guest"|"laugh"|"impersonation"|"actor"|"imp Winfrey"|"OWN"|"Girls Trip"|"The Carmichael Show"|"Keanu"|"Taylc open
nc99ccSXST0	5 Ice Cream Gadgets put to the Test	Science & Technology	5 Ice Cream Gadgets"|"Ice Cream"|"Cream Sandwich Maker"|"gadge to the Test"|"testing"|"10 Kitchen Gadgets"|"7 Camping Coffee Gadg
jr9QtXwC9vc	The Greatest Showman | Official Trailer 2 [HD] | 20th Century FOX	Film & Animation	Trailer"|"Hugh Jackman"|"Michelle Williams"|"Zac Efron"|"Zendaya"| school musical"|"hugh jackman musical"|"zac efron musical"|"music Barnum"|"Barnum and Bailey"|"Barnum Circus"|"Barnum and Bailey trailer"|"the greatest showman trailer"|"logan"|"Benj Pasek"|"Justin
TUmyygCMMGA	Why the rise of the robots won't mean the end of work	News & Politics	vox.com"|"vox"|"explain"|"shift change"|"future of work"|"automati shierholz"|"martin ford"|"rise of the robots"|"humans"|"workers"|"e income

Research question

Is it possible to predict, based on the category or a combination of other factors, the popularity of a youtube video in America?

Cases

Each observation represents a video in Youtube. There are 40,949 observations.

Data collection

Data was obtained from a Kaggle data set.

Type of study

This is an observational study based on the obervations captured in this data.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link. Data was obtained from a Kaggle data set.

Dependent Variable

What is the response variable? Is it quantitative or qualitative? The response variable will be the prediction of number of likes. It is quantitative.

Independent Variable

You should have two independent variables, one quantitative and one qualitative. Category, likes, views, comment_count. Likes, views, and comment_count are quantitative, Category is qualitative.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Summary Statistics

```
summary(video_df)
```

```
##    video_id            title             category             tags          
##  Length:40949       Length:40949       Length:40949       Length:40949      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      views               likes            dislikes        comment_count    
##  Min.   :      549   Min.   :      0   Min.   :      0   Min.   :      0  
##  1st Qu.:   242329   1st Qu.:   5424   1st Qu.:    202   1st Qu.:    614  
##  Median :   681861   Median :  18091   Median :    631   Median :   1856  
##  Mean   :  2360785   Mean   :  74267   Mean   :   3711   Mean   :   8447  
##  3rd Qu.:  1823157   3rd Qu.:  55417   3rd Qu.:   1938   3rd Qu.:   5755  
##  Max.   :225211923   Max.   :5613827   Max.   :1674420   Max.   :1361580
```

We see that the mean like per video is 74,267. We can also see other meaningful statistics in the quantitative variables that might help in our study.

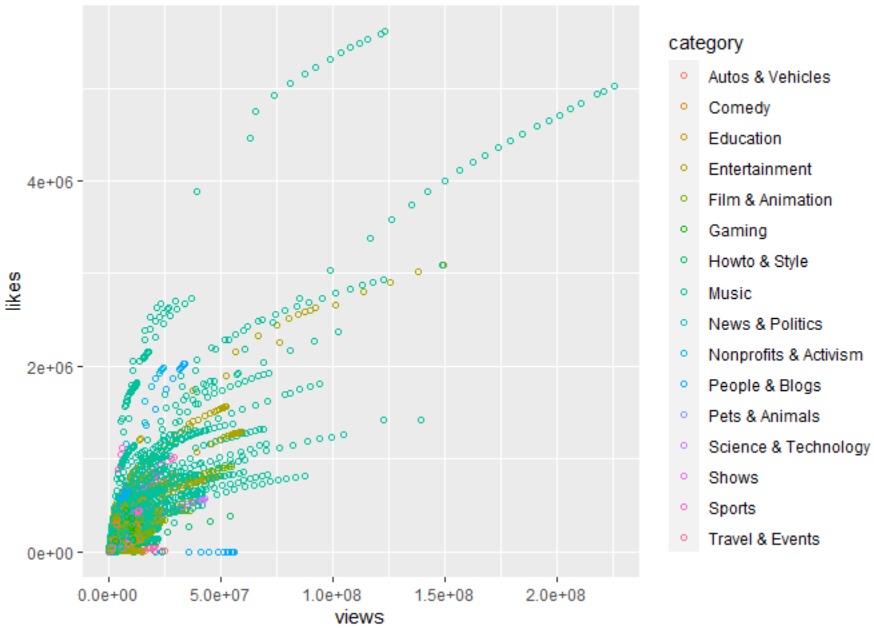
```
describe(video_df %>% select(views, likes, dislikes, comment_count))
```

```
##               vars     n      mean         sd median    trimmed       mad min
## views            1 40949 2360784.6 7394113.76 681861 1054836.27 813077.11 549
## likes            2 40949   74266.7  228885.34  18091   32156.33  23496.24   0
## dislikes         3 40949    3711.4   29029.71    631    1137.46    797.64   0
## comment_count    4 40949    8446.8   37430.49   1856    3324.49   2351.40   0
##                     max     range  skew kurtosis       se
```

```
## views         225211923 225211374 12.24   232.34 36539.66
## likes           5613827   5613827 10.92   177.82  1131.09
## dislikes        1674420   1674420 40.19  1987.08   143.46
## comment_count   1361580   1361580 19.75   532.05   184.97
```

Let's take a look at a scatter plot of views and likes:

```
ggplot(video_df, aes(x=views, y=likes, color = category)) +
    geom_point(shape=1)
```



Category Analysis

We see a clear tendency of some categories to gather more likes than others. Now, let's gather the categories and clean up their names a little:

```
video_categories <- video_df %>%
  group_by(category) %>%
  summarise(
         views_sum = sum(views),
         likes_sum = sum(likes),
         dislikes_sum = sum(dislikes)) %>%
  arrange(desc(likes_sum))
```

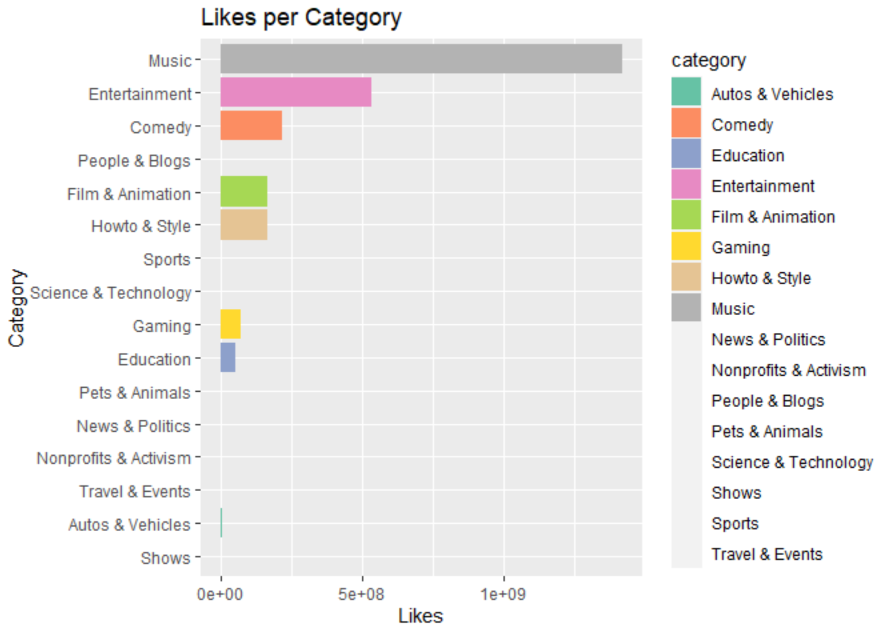
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
knitr::kable(video_categories)
```

category	views_sum	likes_sum	dislikes_sum
Music	40132892190	1416838584	51179008
Entertainment	20604388195	530516491	42987663

category	views_sum	likes_sum	dislikes_sum
Comedy	5117426208	216346746	7230391
People & Blogs	4917191726	186615999	10187901
Film & Animation	7284156721	165997476	6075148
Howto & Style	4078545064	162880075	5473899
Sports	4404456673	98621211	5133551
Science & Technology	3487756816	82532638	4548402
Gaming	2141218625	69038284	9184466
Education	1180629990	49257772	1351972
Pets & Animals	764651989	19370702	527379
News & Politics	1473765704	18151033	4180049
Nonprofits & Activism	168941392	14815646	3310381
Travel & Events	343557084	4836246	340427
Autos & Vehicles	520690717	4245656	243010
Shows	51501058	1082639	24508

```
ggplot(video_categories
         , aes(y=reorder(factor(category), likes_sum),
                             x=likes_sum,
                             fill = category)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2")+
  labs(title="Likes per Category") +
  xlab("Likes") +
  ylab("Category")
```



We can see the *Music* category seems to be the one gathering more likes. Further analysis is needed to identify and analyze the tags associated with the different videos and how the presence of these tags might help answer the initial question.

Let's take a look at the different categories in video_df:

```
cat_frequency <- table(video_df$category)%>%
        as.data.frame() %>%
        arrange(desc(Freq))
knitr::kable(cat_frequency)
```

Var1	Freq
Entertainment	9964
Music	6472
Howto & Style	4146
Comedy	3457
People & Blogs	3210
News & Politics	2487
Science & Technology	2401
Film & Animation	2345
Sports	2174
Education	1656
Pets & Animals	920
Gaming	817
Travel & Events	402

Var1	Freq
Autos & Vehicles	384
Nonprofits & Activism	57
Shows	57

Looking at this data, we can see that although Music has by far the most likes, it is not the most used category. We can probably create a proportion table for categories.

```
cat_prop <- prop.table(table(video_df$category))
knitr::kable(cat_prop %>%
  as.data.frame() %>%
  arrange(desc(Freq)))
```

Var1	Freq
Entertainment	0.2433271
Music	0.1580503
Howto & Style	0.1012479
Comedy	0.0844221
People & Blogs	0.0783902
News & Politics	0.0607341
Science & Technology	0.0586339
Film & Animation	0.0572664
Sports	0.0530904
Education	0.0404405
Pets & Animals	0.0224670
Gaming	0.0199516
Travel & Events	0.0098171
Autos & Vehicles	0.0093775
Nonprofits & Activism	0.0013920
Shows	0.0013920

Tempted to do a pie chart As visualizations go, I wanted to place a pie chart here but…the professor doesn't like Pie charts so that table should suffice.

What we will do, is probably get a proportion table for the likes per category and then formulate a null hypothesis to test with the chi-square.

```
video_categories <- video_categories %>%
  mutate(likes_prop = likes_sum /(sum(likes_sum)), total_likes=sum(likes_sum))
knitr::kable(video_categories)
```

category	views_sum	likes_sum	dislikes_sum	likes_prop	total_likes
Music	40132892190	1416838584	51179008	0.4658895	3041147198
Entertainment	20604388195	530516491	42987663	0.1744462	3041147198
Comedy	5117426208	216346746	7230391	0.0711398	3041147198
People & Blogs	4917191726	186615999	10187901	0.0613637	3041147198
Film & Animation	7284156721	165997476	6075148	0.0545838	3041147198
Howto & Style	4078545064	162880075	5473899	0.0535588	3041147198
Sports	4404456673	98621211	5133551	0.0324290	3041147198
Science & Technology	3487756816	82532638	4548402	0.0271387	3041147198
Gaming	2141218625	69038284	9184466	0.0227014	3041147198
Education	1180629990	49257772	1351972	0.0161971	3041147198
Pets & Animals	764651989	19370702	527379	0.0063695	3041147198
News & Politics	1473765704	18151033	4180049	0.0059685	3041147198
Nonprofits & Activism	168941392	14815646	3310381	0.0048717	3041147198
Travel & Events	343557084	4836246	340427	0.0015903	3041147198
Autos & Vehicles	520690717	4245656	243010	0.0013961	3041147198
Shows	51501058	1082639	24508	0.0003560	3041147198

I will also add an expected column based on the proportion_table.

```
cat_df <- cat_prop %>%
  as.data.frame()%>%
  rename(category = Var1,
         prop = Freq)

video_categories <- video_categories %>%
  left_join(cat_df) %>%
  mutate(expected = total_likes * prop)

knitr::kable(video_categories)
```

category	views_sum	likes_sum	dislikes_sum	likes_prop	total_likes	prop	expected
Music	40132892190	1416838584	51179008	0.4658895	3041147198	0.1580503	480654098
Entertainment	20604388195	530516491	42987663	0.1744462	3041147198	0.2433271	739993423
Comedy	5117426208	216346746	7230391	0.0711398	3041147198	0.0844221	256739990
People & Blogs	4917191726	186615999	10187901	0.0613637	3041147198	0.0783902	238396115

category	views_sum	likes_sum	dislikes_sum	likes_prop	total_likes	prop	expected
Film & Animation	7284156721	165997476	6075148	0.0545838	3041147198	0.0572664	174155417
Howto & Style	4078545064	162880075	5473899	0.0535588	3041147198	0.1012479	307909748
Sports	4404456673	98621211	5133551	0.0324290	3041147198	0.0530904	161455811
Science & Technology	3487756816	82532638	4548402	0.0271387	3041147198	0.0586339	178314353
Gaming	2141218625	69038284	9184466	0.0227014	3041147198	0.0199516	60675896
Education	1180629990	49257772	1351972	0.0161971	3041147198	0.0404405	122985659
Pets & Animals	764651989	19370702	527379	0.0063695	3041147198	0.0224670	68325366
News & Politics	1473765704	18151033	4180049	0.0059685	3041147198	0.0607341	184701289
Nonprofits & Activism	168941392	14815646	3310381	0.0048717	3041147198	0.0013920	4233202
Travel & Events	343557084	4836246	340427	0.0015903	3041147198	0.0098171	29855214
Autos & Vehicles	520690717	4245656	243010	0.0013961	3041147198	0.0093775	28518414
Shows	51501058	1082639	24508	0.0003560	3041147198	0.0013920	4233202

Chi-Squared

We can formulate a Null Hypothesis with our data:

- **H0** - There is no relationship between category and the number of likes a video gets
- **H1** - There is a marked relationship between category and likes.

```
k <- 16  #16 categories
df <- k - 1
chi.Sq <- 0

for(i in 1:16)
{
  chi.Sq <- chi.Sq + ((video_categories$likes_sum[i] - video_categories$expected[i])^2 /
         video_categories$expected[i])
}

p.Value <- pchisq(chi.Sq, df=df, lower.tail=FALSE)
paste('p-value is ',p.Value )
```

```
## [1] "p-value is  0"
```

Because the P-Value is so small as to approach 0, we must reject the Null Hypothesis and accept the Alternate hypothesis that says: **There is a marked relationship between category and likes**

Regression with Categorical variable

```
video_df$category.f <- factor(video_df$category)  ###Turn the category into a factor to use
        with lm

lm_output <- lm(likes ~ category.f, data = video_df)
```

Let’s take a look at the regression model:

```
lm_output
```

```
##
## Call:
## lm(formula = likes ~ category.f, data = video_df)
##
## Coefficients:
##                    (Intercept)                 category.fComedy
##                        11056.4                          51525.8
##            category.fEducation          category.fEntertainment
##                        18688.6                          42186.9
##       category.fFilm & Animation                 category.fGaming
##                        59731.4                          73445.8
##          category.fHowto & Style                  category.fMusic
##                        28229.7                         207861.8
##       category.fNews & Politics  category.fNonprofits & Activism
##                        -3758.0                         248867.2
##         category.fPeople & Blogs         category.fPets & Animals
##                        47079.4                           9998.7
##  category.fScience & Technology                  category.fShows
##                        23317.9                           7937.3
##               category.fSports        category.fTravel & Events
##                        34307.5                            974.1
```

Because we have a large number of factors, reading this is a little confusing, but let’s take a look:

- Our Intercept is: 11,056, this means that the avg likes is 11056 For a **comedy video**, we will add 51,525.8 likes For a **News & Politics** video we will remove -3,758

Let’s take a further look down the model:

```
summary(lm_output)
```

```
##
## Call:
## lm(formula = likes ~ category.f, data = video_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -259924  -49306  -28262    -861 5394909
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     11056.4    11197.9   0.987 0.323472
## category.fComedy                51525.8    11803.5   4.365 1.27e-05 ***
## category.fEducation             18688.6    12428.6   1.504 0.132673
## category.fEntertainment         42186.9    11411.7   3.697 0.000219 ***
## category.fFilm & Animation      59731.4    12080.0   4.945 7.66e-07 ***
```

```
## category.fGaming                 73445.8    13576.8   5.410 6.35e-08 ***
## category.fHowto & Style          28229.7    11705.0   2.412 0.015880 *
## category.fMusic                 207861.8    11525.4  18.035  < 2e-16 ***
## category.fNews & Politics        -3758.0    12031.4  -0.312 0.754775
## category.fNonprofits & Activism 248867.2    31147.3   7.990 1.38e-15 ***
## category.fPeople & Blogs         47079.4    11848.8   3.973 7.10e-05 ***
## category.fPets & Animals          9998.7    13331.6   0.750 0.453259
## category.fScience & Technology   23317.9    12060.2   1.933 0.053187 .
## category.fShows                   7937.3    31147.3   0.255 0.798856
## category.fSports                 34307.5    12146.7   2.824 0.004739 **
## category.fTravel & Events          974.1    15658.0   0.062 0.950397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 219400 on 40933 degrees of freedom
## Multiple R-squared:  0.08122,    Adjusted R-squared:  0.08088
## F-statistic: 241.2 on 15 and 40933 DF,  p-value: < 2.2e-16
```

We can see that some of these factors are significant: i.e. Comedy, Entertainment, Film, Gaming, Music, Nonprofit, People and Sports.

Regrettably, the adjusted R-squared is very low. This means that, even though we may have a trend, the category alone does not explain the number of likes. It seems to indicate that the number of likes are affected about 8% by the category.

We could write the prediction formula this way:

Y = b0 + b1 * category

Number of Likes = 11,056.4 + Estimate(Factor)

i.e.

We would predict that a New Film & Animation video will end up with:

number of likes = 11,056.4 + 59731.4

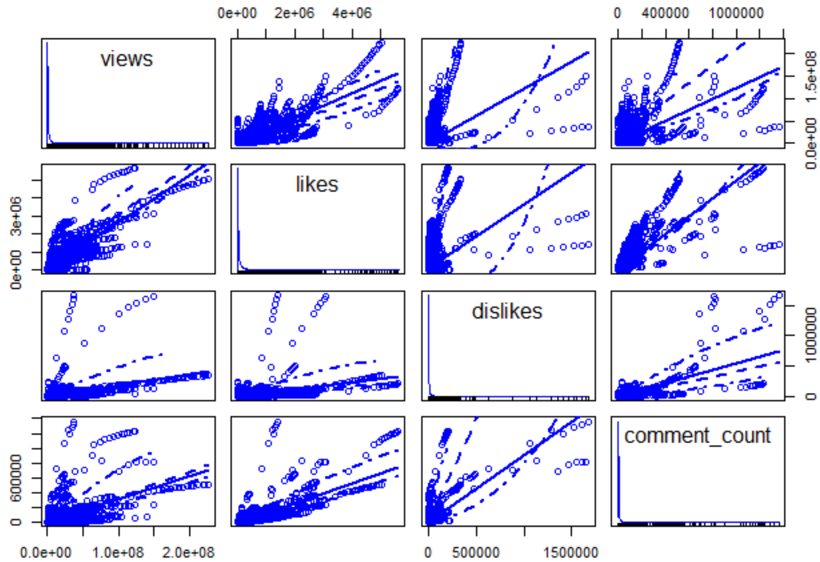
Note

As we have determined that this model is not the only or the most significant factor determining the number of likes a video will get, we have also not studied the period of time needed to gather these likes.

Exploring other factors

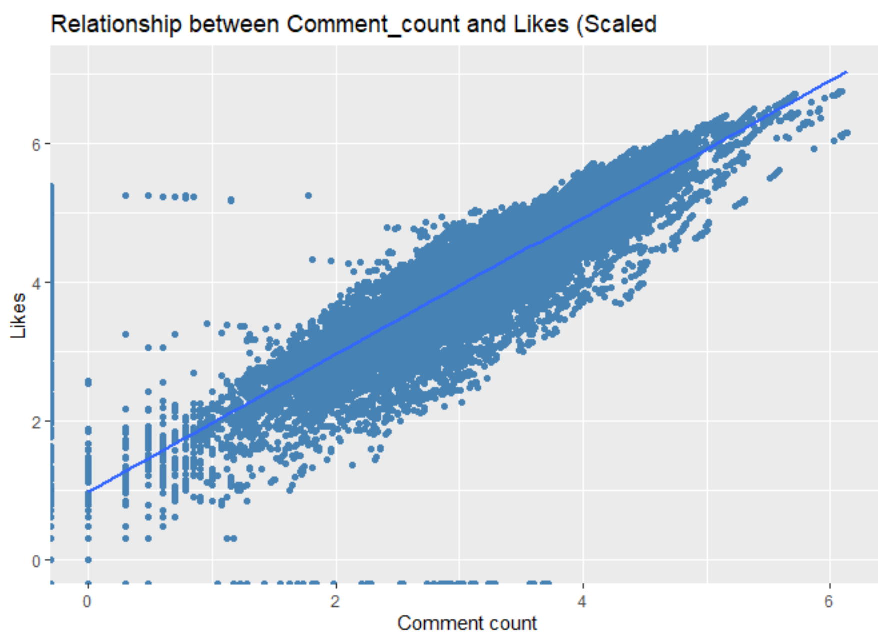
Let’s take a look at a scatterPlot Matrix of the quantitative variables in our dataset:

```
scatterplotMatrix(video_df[5:8])
```



Using Log10 to scale the number of likes and comment_count we get the following scatterplot:

```
ggplot(video_df,
       aes(x = log10(comment_count),
           y = log10(likes))) +
  geom_point(color= "steelblue") +
  geom_smooth(method = "lm")+
  labs(title="Relationship between Comment_count and Likes (Scaled") +
  xlab("Comment count") +
  ylab("Likes")
```



This shows a (expected) relationship between the number of likes and comments. Would it be possible to fit a linear regression model using comment_counts to predict how many likes a video will get?

Linear regression with Comment count.

```
comment_lm <- lm(likes ~ comment_count, video_df)
comment_lm
```

```
##
## Call:
## lm(formula = likes ~ comment_count, data = video_df)
##
## Coefficients:
##   (Intercept)  comment_count
##     32787.424          4.911
```

We get a large intercept: 32,787. But this seems to indicate that for every comment, we see an increase of about 5 likes.

Y = b0 + b1 * comment_count

```
summary(comment_lm)
```

```
##
## Call:
## lm(formula = likes ~ comment_count, data = video_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5316449   -32474   -26338    -6305  2450718
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.279e+04  6.910e+02   47.45   <2e-16 ***
## comment_count 4.911e+00  1.801e-02  272.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 136400 on 40947 degrees of freedom
## Multiple R-squared:  0.6449, Adjusted R-squared:  0.6449
## F-statistic: 7.436e+04 on 1 and 40947 DF,  p-value: < 2.2e-16
```

These values show an important relationship between the number of comments and likes. The R^2 of 0.6449 indicates that 64.5% of the likes can be explained by the comments.

So, our fitted model will be something like this:

Number of Likes = 32,787.4 + 4.911 * comment_count

Can we combine comment_count with views and dislikes?

If we try to fit a model with those variables we get:

```
multi_variate_lm <- lm(likes ~ comment_count + views + dislikes, video_df)

multi_variate_lm
```

```
##
## Call:
## lm(formula = likes ~ comment_count + views + dislikes, data = video_df)
```

```
##
## Coefficients:
##   (Intercept)  comment_count          views       dislikes
##    6680.24504        3.85149        0.01822       -2.14229
```

This indicates that for each comment, we can add about 4 likes. We will get a like for about every 100 views and each dislike will reduce about 2 likes from our total.

```
summary(multi_variate_lm)
```

```
##
## Call:
## lm(formula = likes ~ comment_count + views + dislikes, data = video_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1351640   -11911    -6696     3113  1081510
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.680e+03  4.052e+02   16.48   <2e-16 ***
## comment_count  3.851e+00  1.620e-02  237.79   <2e-16 ***
## views          1.822e-02  6.641e-05  274.30   <2e-16 ***
## dislikes      -2.142e+00  1.863e-02 -114.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77950 on 40945 degrees of freedom
## Multiple R-squared:  0.884,  Adjusted R-squared:  0.884
## F-statistic: 1.04e+05 on 3 and 40945 DF,  p-value: < 2.2e-16
```

The adjusted R^2 is 0.884 which indicates a high level of incidence in the number of likes for these three variables.

What does all this mean?

Based on our Null Hypothesis analysis we were able to identify a correlation between the video category and the number of likes attained. At the very least, we were not able to accept the null hypothesis that no relation existed. Further analysis showed that the R^2 value for such relationship was too low for us to properly fit a model that would allow us to calculate the number of likes based on the category alone.

We expanded our analysis to other variables. The **comment_count** proved to be a better predictor of likes and a multivariate regression incorporating views and dislikes gave us an adjusted R^2 of 88%. In terms of the scope of this analysis, the number of comments, views and dislikes are a better predictor than category for the number of likes a video will get.

Next Steps

In terms of practical use, the results found in this analysis do not give us a silver bullet on how to gather likes in a Youtube video. I would like to expand this analysis to incorporate the tags and the description of the video with other statistical and machine learning methods as Random Forest or SVM to gain a better understanding of the factors that better influence the number of likes a video gets.