YouTube Video Statistics

UCLA Data Science Intensive Capstone

Matt Eisenbrei

August 31, 2019

Contents

- Goals and Approach
- Initial Data Assessment & EDA
- Quantitative Analysis
  - Regression
  - Model Selection
  - Best Model for Views
- Discussion

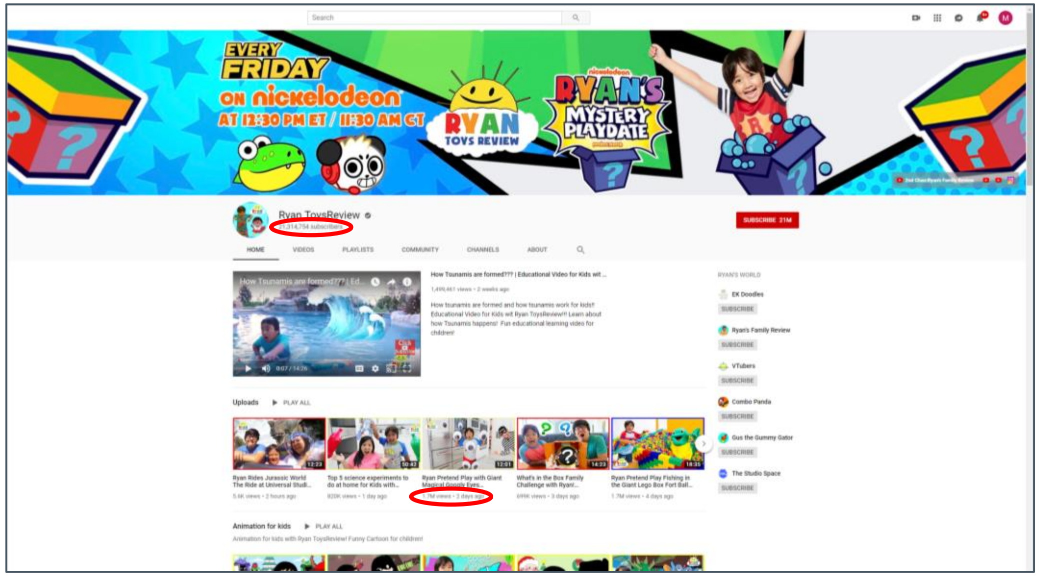
- The domain name "YouTube.com" was activated on February 14th, 2005
- The first YouTube video was uploaded on April 23rd, 2005
- YouTube was purchased by Google for $1.7B in Nov. 2006
- The first advertisements launched in August, 2007
- In an average month, 8 out of 10 18 to 49-year-olds watch YouTube
- The platform has over 1.9B monthly users
- YouTube is the world's second largest search engine and second most visited site after Google
- The platform has also launched in over 91 countries
- Ryan Kaji, better known as Ryan ToysReview, is the highest earning YouTuber, bringing in $22m in 2018

All YouTube data on this slide courtesy of Brandwatch
https://www.brandwatch.com/blog/youtube-stats/

Goals & Approach

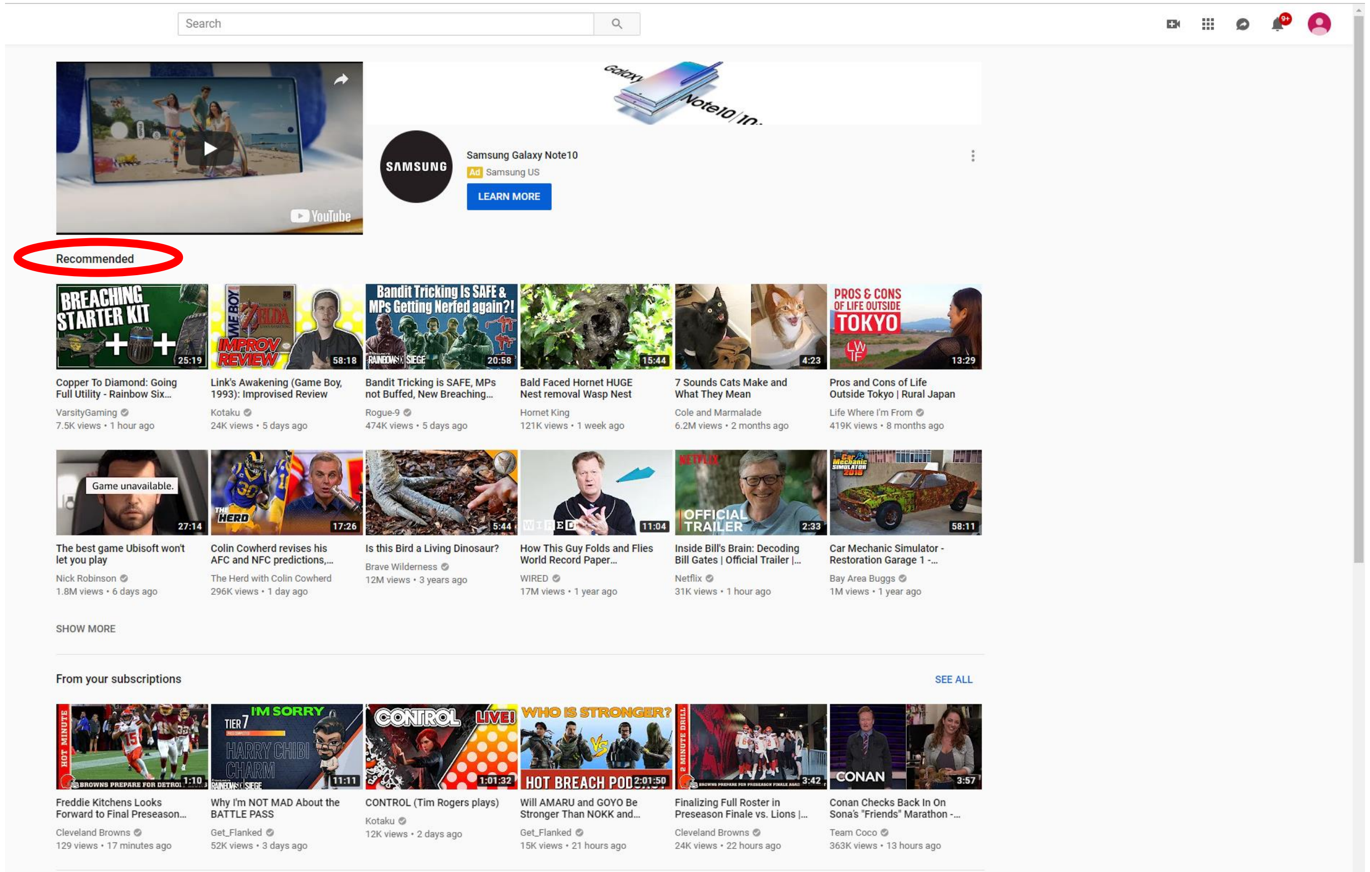
Analysis Goals

- YouTube's business model is based on ad revenue tied to video views
- For content creators, this means that content with higher views drives more ad revenue, which in turn provides higher compensation
- My primary goal in analyzing the data is to better understand which variables have the most significant impact on views in the US
- Is there a recipe for success?
- Is it possible to accurately predict video views?



Content Creator Goal

- Maximize views
- Create content that is highly “recommended” to maximize views (algorithm)



The Data: Overview

- The YouTube data set is a daily listing of top trending videos by country
- It is available via Kaggle: https://www.kaggle.com/datasnaek/youtube-new
- It includes data for 10 countries (as .csv files) as well as separate column metadata sets by country (as .json files)
- My analysis was limited to US video data due to time constraints
  - Comparing US data to international data would be interesting for future analyses
- I chose this data set for three reasons:
  1. Personal interest and curiosity about the subject matter; I use the platform daily
  2. Wide variety of data types – quantitative, text, time series, and image
  3. Real world application: maximizing views = maximizing creator compensation

The Data: Starting Set

- The primary US data file is ~41K observations across 16 variables
- I kept most variables and introduced more to improve the story-telling
- I removed 656 observations with technical issues or comments disabled
- There were no “subscription” or “share” variables which was surprising

```
> str(us.data)
'data.frame':	40949 obs. of  16 variables:
 $ video_id              : Factor w/ 6282 levels "-0CMnp02rNY",..: 378 319 711 4337 1755 2418 454 3797 3063 5155 ...
 $ trending_date         : Factor w/ 205 levels "17.01.12","17.02.12",..: 14 14 14 14 14 14 14 14 14 14 ...
 $ title                 : Factor w/ 6455 levels "'Avengers: Infinity War' Cast Tours Los Angeles w/ James Corden",..: 6046 5527 4441 4068 2640 161 4625
 256 5352 6305 ...
 $ channel_title         : Factor w/ 2207 levels "12 News","1MILLION Dance Studio",..: 332 1109 1651 768 1424 890 1681 464 4 2126 ...
 $ category_id           : int  22 24 23 24 24 28 24 28 1 25 ...
 $ publish_time          : Factor w/ 6269 levels "2006-07-23T08:24:11.000Z",..: 318 287 271 291 269 323 256 274 297 295 ...
 $ tags                  : Factor w/ 6055 levels "#guitar #musiciseverywhere #jammin #meme #funny #deeppurple #pinkfloyd",..: 4673 3071 4300 4442 4521 24
87 4808 64 5563 5782 ...
 $ views                 : int  748374 2418783 3191434 343168 2095731 119180 2103417 817732 826059 256426 ...
 $ likes                 : int  57527 97185 146033 10172 132235 9763 15993 23663 3543 12654 ...
 $ dislikes              : int  2966 6146 5339 666 1989 511 2445 778 119 1363 ...
 $ comment_count         : int  15954 12703 8181 2146 17518 1434 1970 3432 340 2368 ...
 $ thumbnail_link        : Factor w/ 6352 levels "https://i.ytimg.com/vi/-_jlqATo9eo/default.jpg",..: 447 388 780 4407 1824 2487 523 3866 3132 5225 ...
 $ comments_disabled     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ ratings_disabled      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ video_error_or_removed: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ description           : Factor w/ 6902 levels "","'A curious cat helps his owner with home improvements.'\\nWe're releasing a NEW BLACK & WHITE episod
e every wee"| __truncated__,..: 4844 4286 6380 6122 2630 6262 1611 2724 2973 1872 ...
> 
```

The Data: Additional Variables & Expanded Set

- I expanded the category ID, date, and descriptive text variables to be able to later run regression against them for a more robust model
- I removed a portion of the variables after filtering due to limited value

Key: Removed Variable | Expanded Variable | New Variable

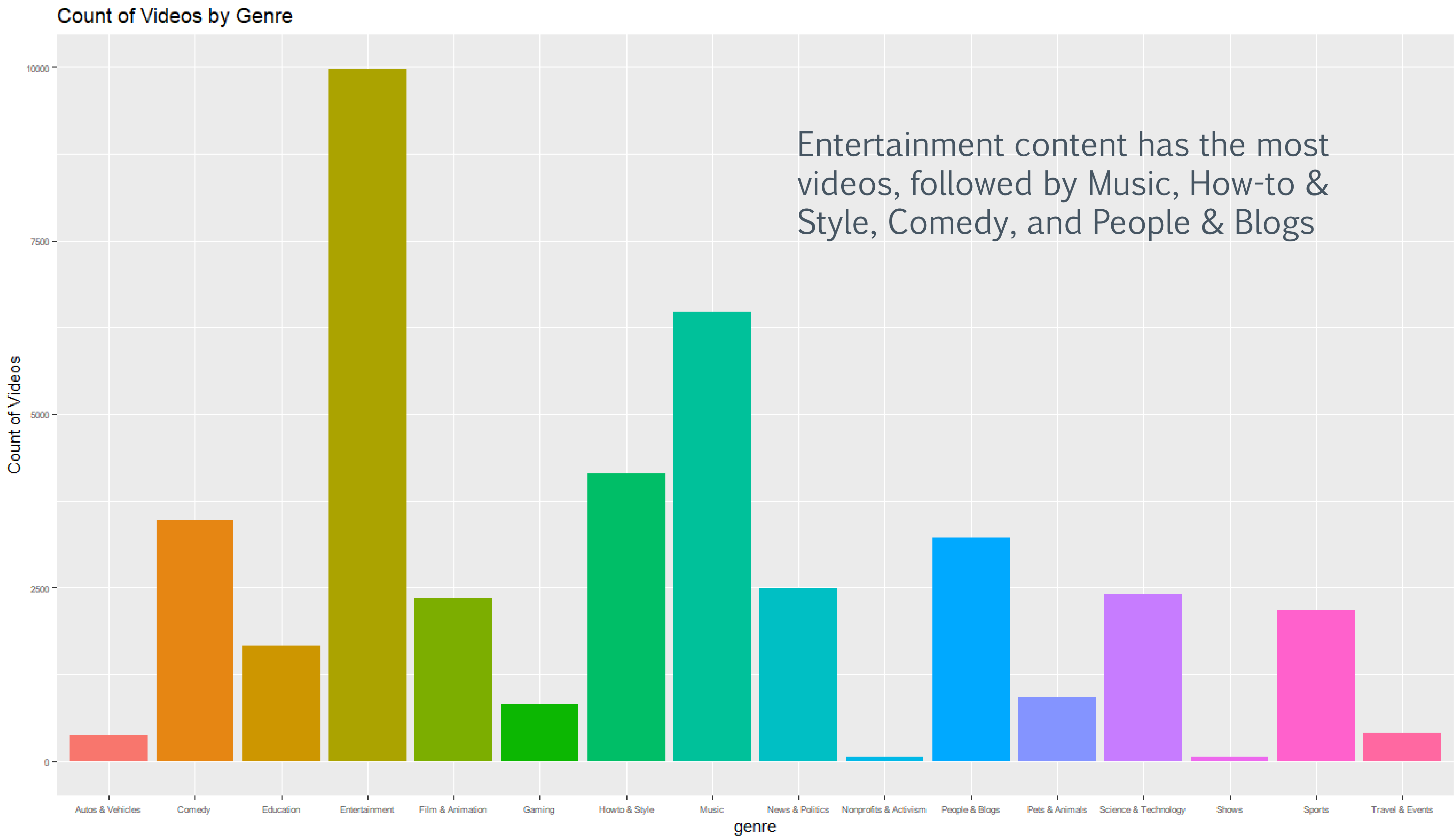
```
> str(us.data)
'data.frame':	40949 obs. of  24 variables:
 $ category_id           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ video_id              : Factor w/ 6282 levels "-0CMnp02rNY",..: 2373 3438 4798 5527 804 5445 1700 2801 6143 3438 ...
 $ trending_date         : Date, format: "0018-02-26" "0018-03-12" "0018-01-17" "0018-04-26" ...
 $ title                 : chr  "Honest Trailers - Justice League" "Everything Wrong With Birdman In 13 Minutes Or Less" "BEIRUT | Official Trailer" "Apple Dessert that looks like a real
apple!  No mold challenge" ...
 $ channel_title         : chr  "Screen Junkies" "CinemaSins" "Bleecker Street" "How To Cook That" ...
 $ publish_time          : Factor w/ 6269 levels "2006-07-23T08:24:11.000Z",..: 4141 4371 2580 5405 5875 722 3238 1535 1478 4371 ...
 $ tags                  : chr  "screenjunkies|\"screen junkies\"|\"honest trailers\"|\"honest trailer\"|\"justice league\"|\"dc movies\"|\"wond"| __truncated__ "birdman|\"bird man\"|\"mi
chael keaton\"|\"cinemasins\"|\"cinema sins\"|\"everything wrong with\"|\"eww\"|\"movi"| __truncated__ "bleecker street|\"bleecker street media\"|\"bleecker street films\"|\"bleecker stree
t movies\"|\"movies\"|\"fil"| __truncated__ "apple pie|\"dessert\"|\"apple shaped dessert\"|\"apple pen\"|\"balloon dessert\"|\"amazing dessert\"|\"fruit de"| __truncated__ ...
 $ views                 : int  2633079 1080795 5186780 259340 12887949 153898 705015 228745 28013 945670 ...
 $ likes                 : int  69999 23939 510 9059 338755 7551 37505 1848 2374 21927 ...
 $ dislikes              : int  2377 950 1774 177 6727 1028 1413 41 15 832 ...
 $ comment_count         : int  10501 2282 665 934 19502 1904 16044 144 102 2131 ...
 $ thumbnail_link        : Factor w/ 6352 levels "https://i.ytimg.com/vi/-_jlqATo9eo/default.jpg",..: 2442 3507 4868 5597 873 5515 1769 2870 6213 3507 ...
 $ comments_disabled     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ ratings_disabled      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ video_error_or_removed: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ description           : chr  "Somewhere between the awful Suicide Squad and the really good Wonder Woman, there's a movie that is determined "| __truncated__ "Birdman is a pretty great
 movie, with a lot of unique touches. We love it. But like all movies, it's got sins, "| __truncated__ "Official Site: http://www.BeirutMovie.com\\nLIKE us on Facebook: http://www.faceboo
k.com/BeirutTheMovie\\nFOLLO"| __truncated__ "Watch Next: https://www.youtube.com/watch?v=W0qQKNvOKtg&list=PLPTOYU_OVLHx2zhtX3nVim5Y852Z-4-gO\\nRecipe: https"| __truncated__ ...
 $ genre                 : Factor w/ 31 levels "Action/Adventure",..: 11 11 11 11 11 11 11 11 11 11 ...
 $ trending_weekday      : Factor w/ 7 levels "Friday","Monday",..: 2 2 7 5 1 5 4 2 5 1 ...
 $ publish_date          : Date, format: "2018-02-20" "2018-02-27" "2018-01-11" "2018-04-20" ...
 $ publish_date_weekday  : Factor w/ 7 levels "Friday","Monday",..: 6 6 5 1 6 7 3 6 2 6 ...
 $ title.sentiment       : num  1 -0.265 0 0.337 0.212 ...
 $ channel_title.sentiment: num  0 0 0 0 0 0 0 0 0 0 ...
 $ tags.sentiment        : num  1.942 -0.323 0.2 1.134 0.561 ...
 $ description.sentiment : num  0.3117 0.2411 -0.0329 0.2703 0.1687 ...
> 
```

Approach

- Reviewed the .csv data set and added category data from the .json files
- Removed select variables and observations. Added new variables
- Incorporated some upfront text analysis for regression
- Assessed correlation between variables (especially likes, dislikes, and comment count) and with views specifically
- Performed EDA to get additional context for the base data set
- Ran regression against key variables, including one case with a key variable included and another with the same variable omitted
- Identified the optimal explanatory model using best subset selection, forward selection, backward selection, and cross-validation

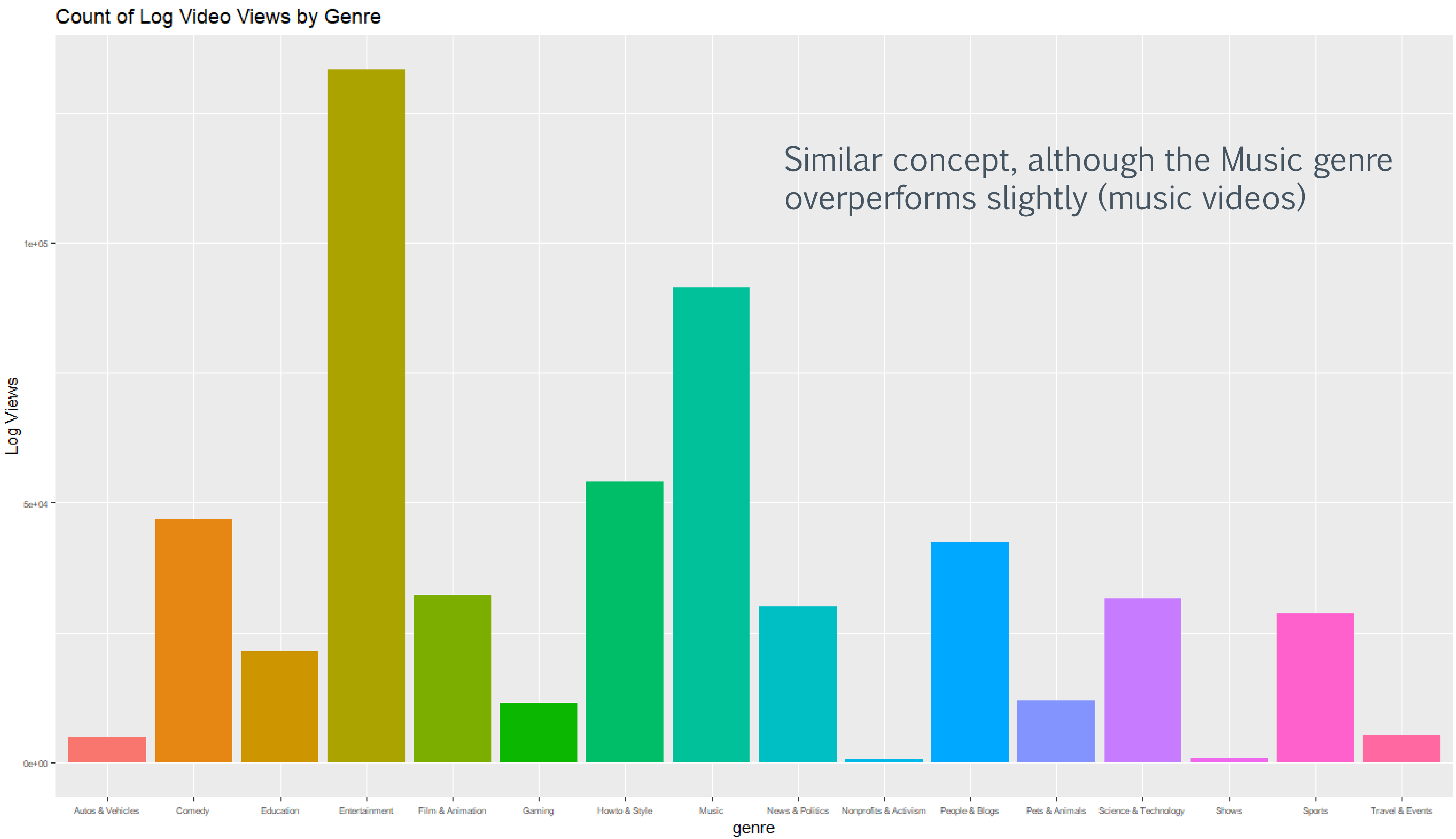
Initial Data Assessment & EDA

Which Genres Have the Most Videos?



Entertainment content has the most videos, followed by Music, How-to & Style, Comedy, and People & Blogs

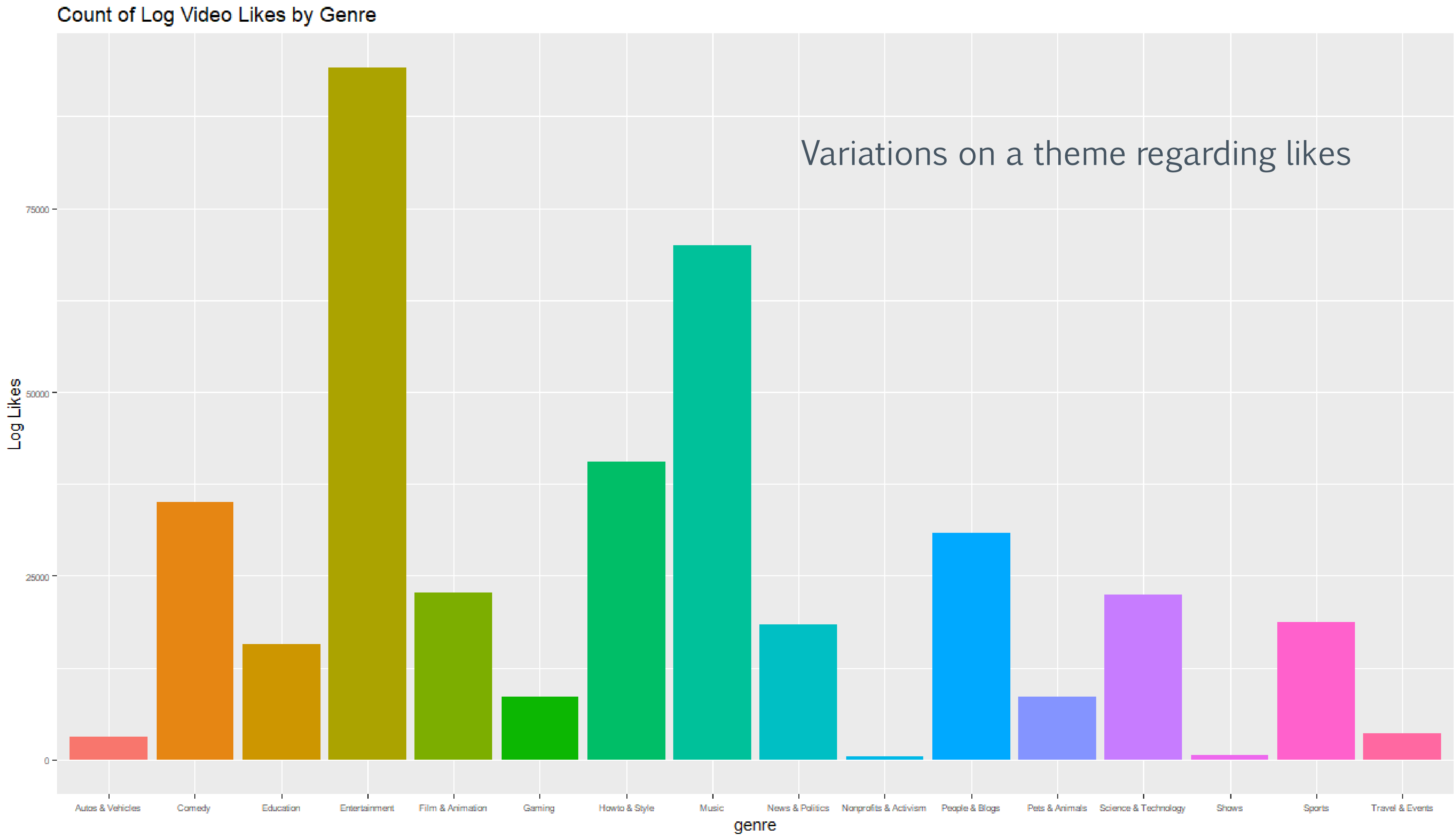
Which Genres Have the Most Views?



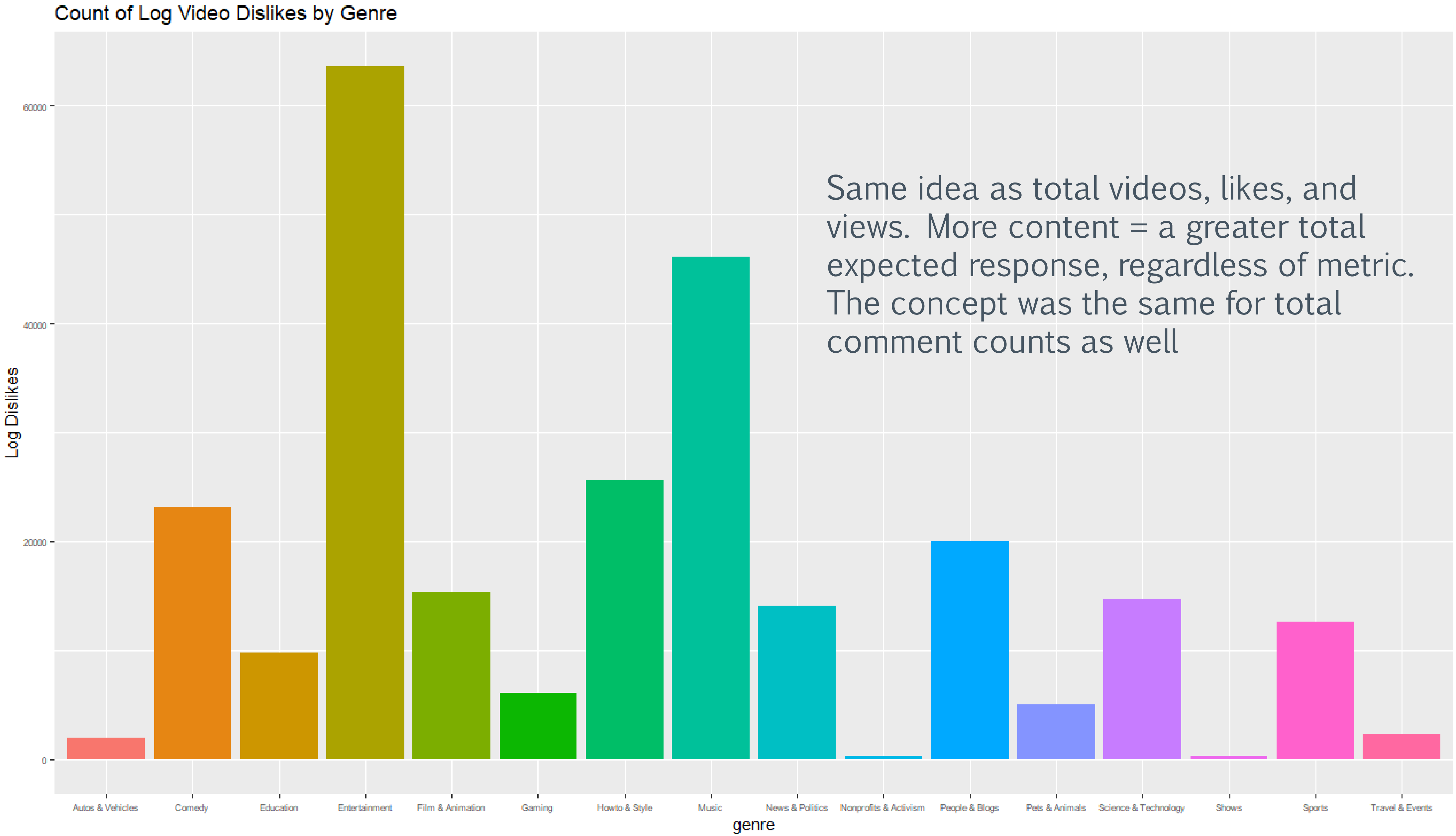
Similar concept, although the Music genre overperforms slightly (music videos)

Which Genres Have the Most Likes?

Variations on a theme regarding likes

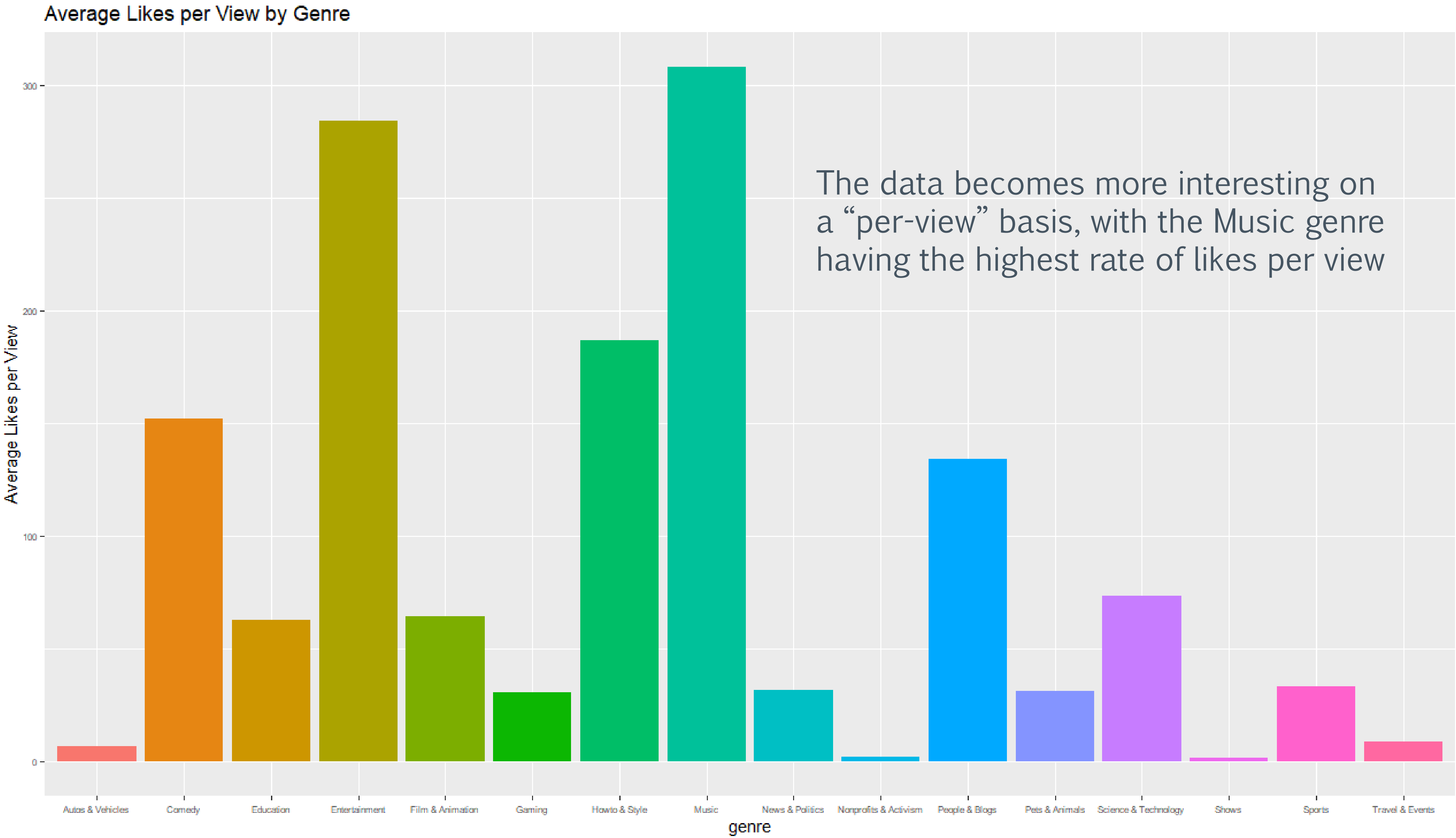


Which Genres Have the Most Dislikes?



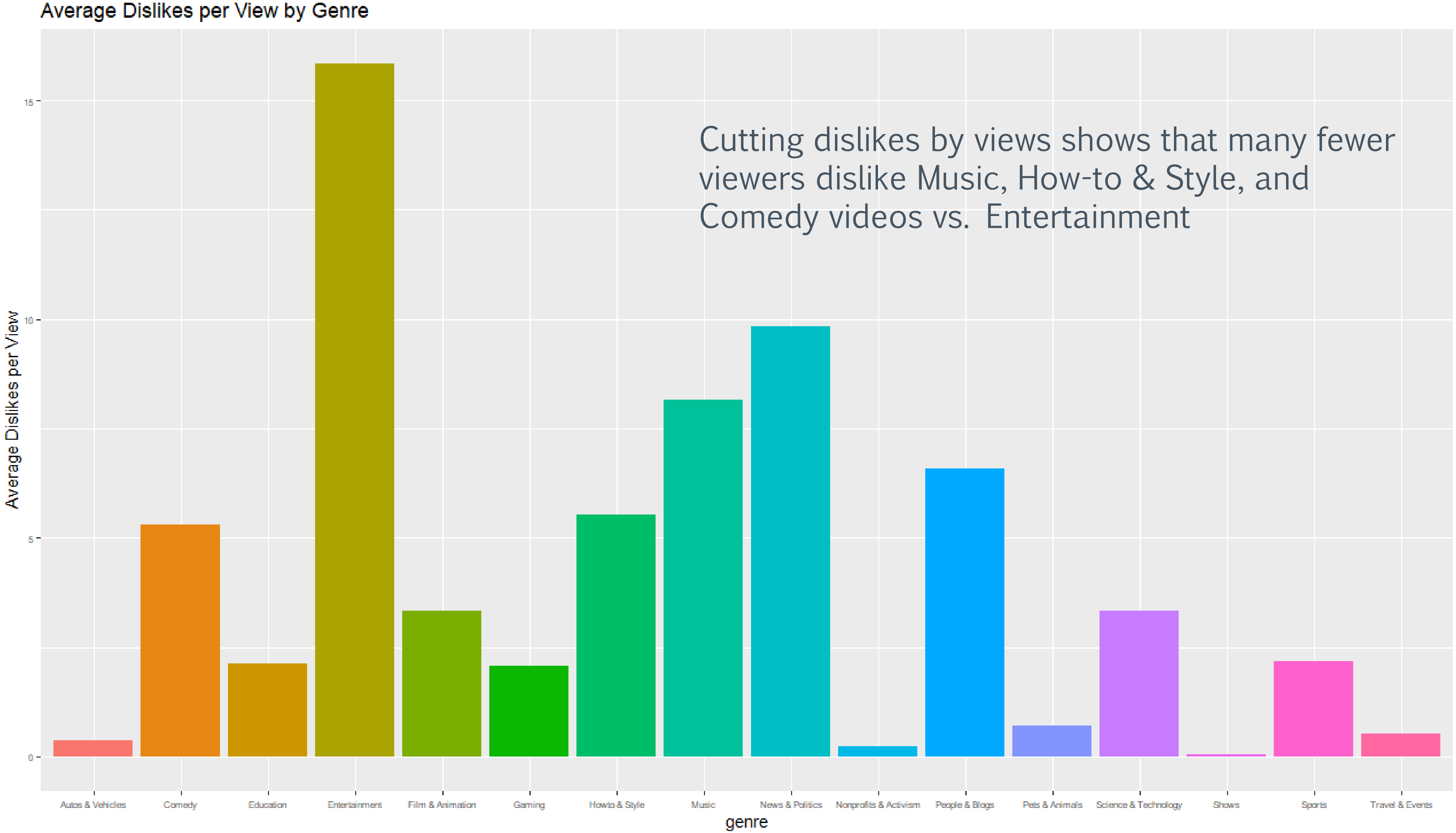
Same idea as total videos, likes, and views. More content = a greater total expected response, regardless of metric. The concept was the same for total comment counts as well

Which Genres Have the Most Likes per View?



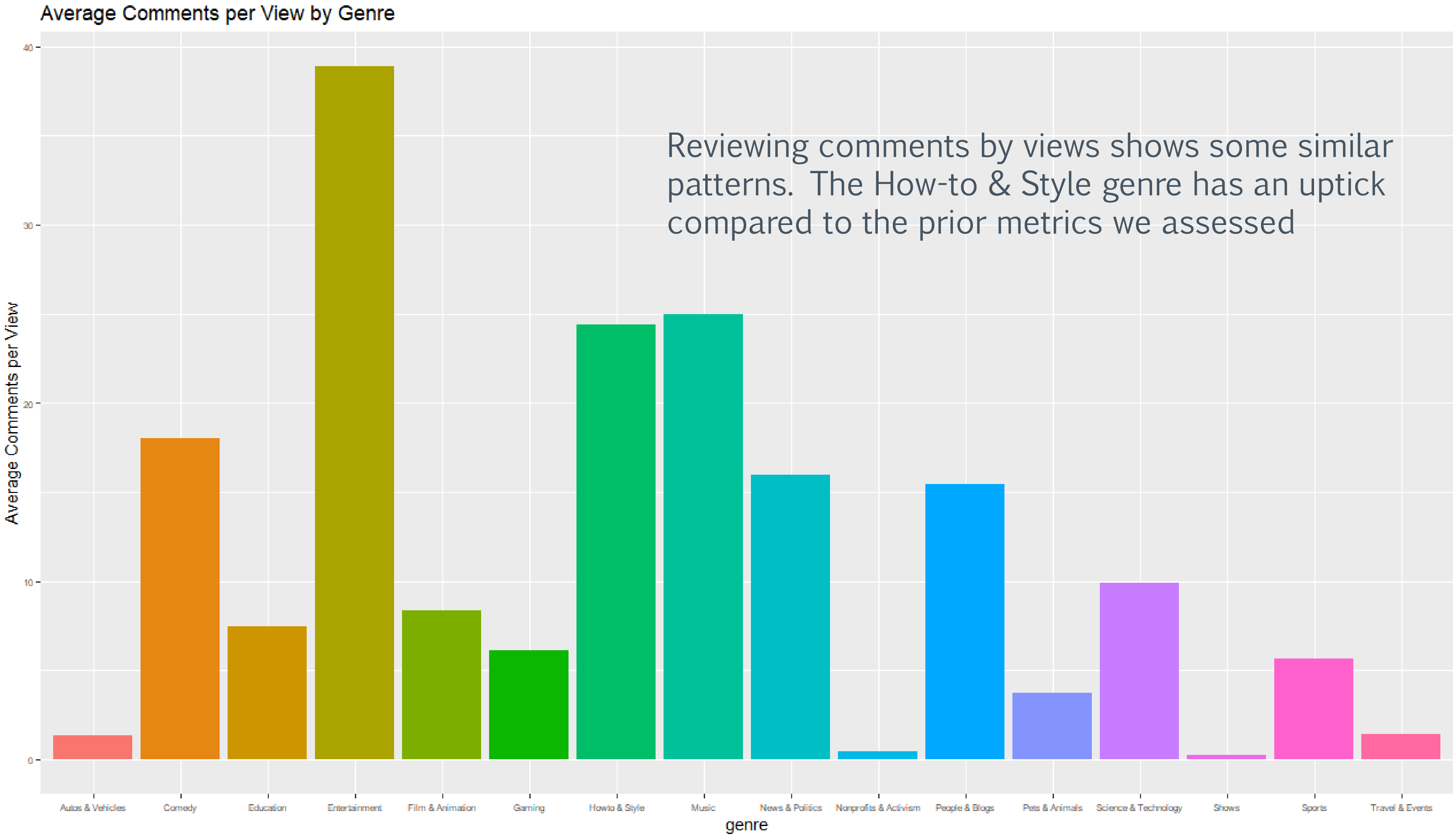
The data becomes more interesting on a “per-view” basis, with the Music genre having the highest rate of likes per view

Which Genres Have the Most Dislikes per View?



Cutting dislikes by views shows that many fewer viewers dislike Music, How-to & Style, and Comedy videos vs. Entertainment

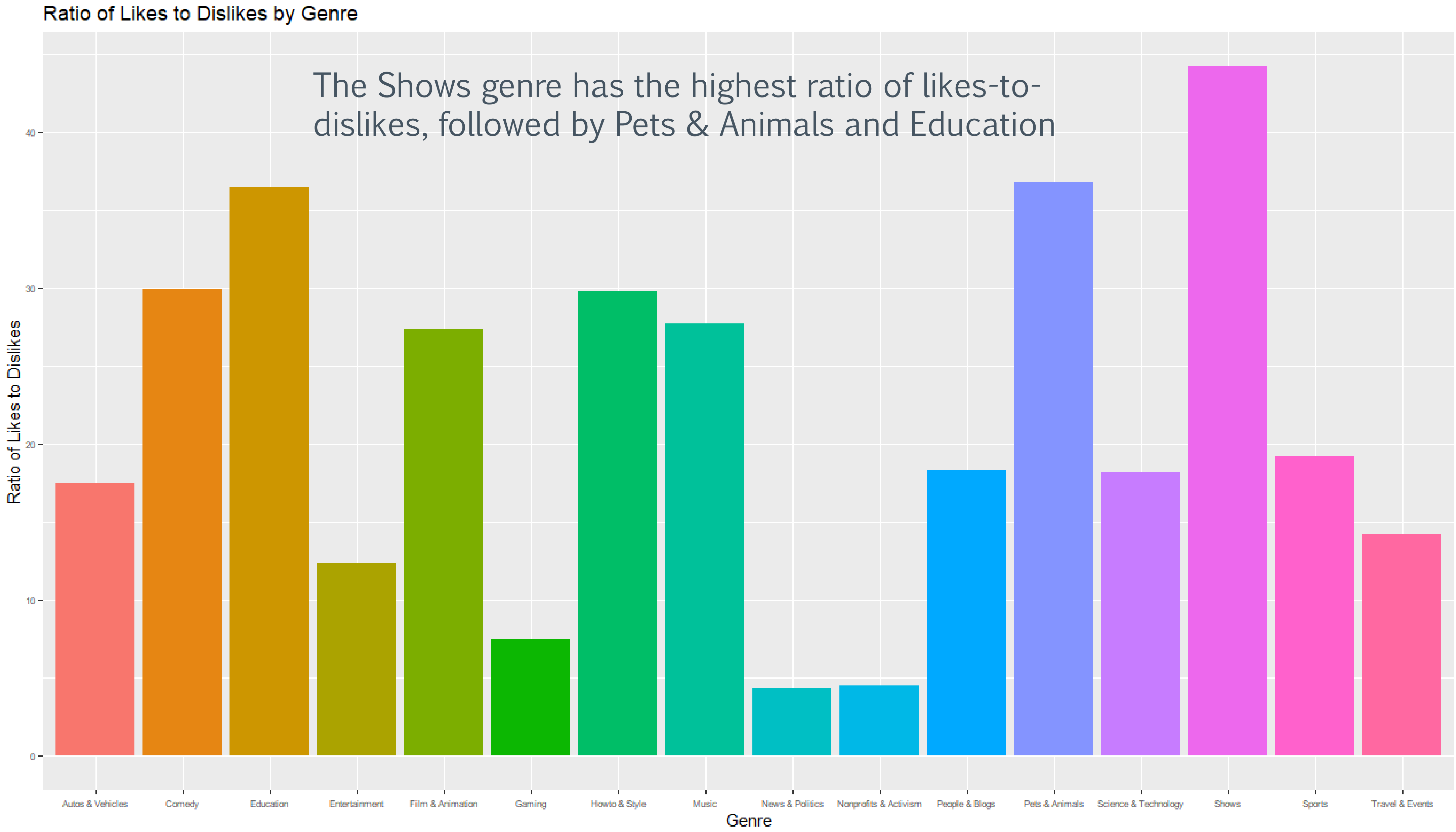
Which Genres Have the Most Comments per View?



Reviewing comments by views shows some similar patterns. The How-to & Style genre has an uptick compared to the prior metrics we assessed

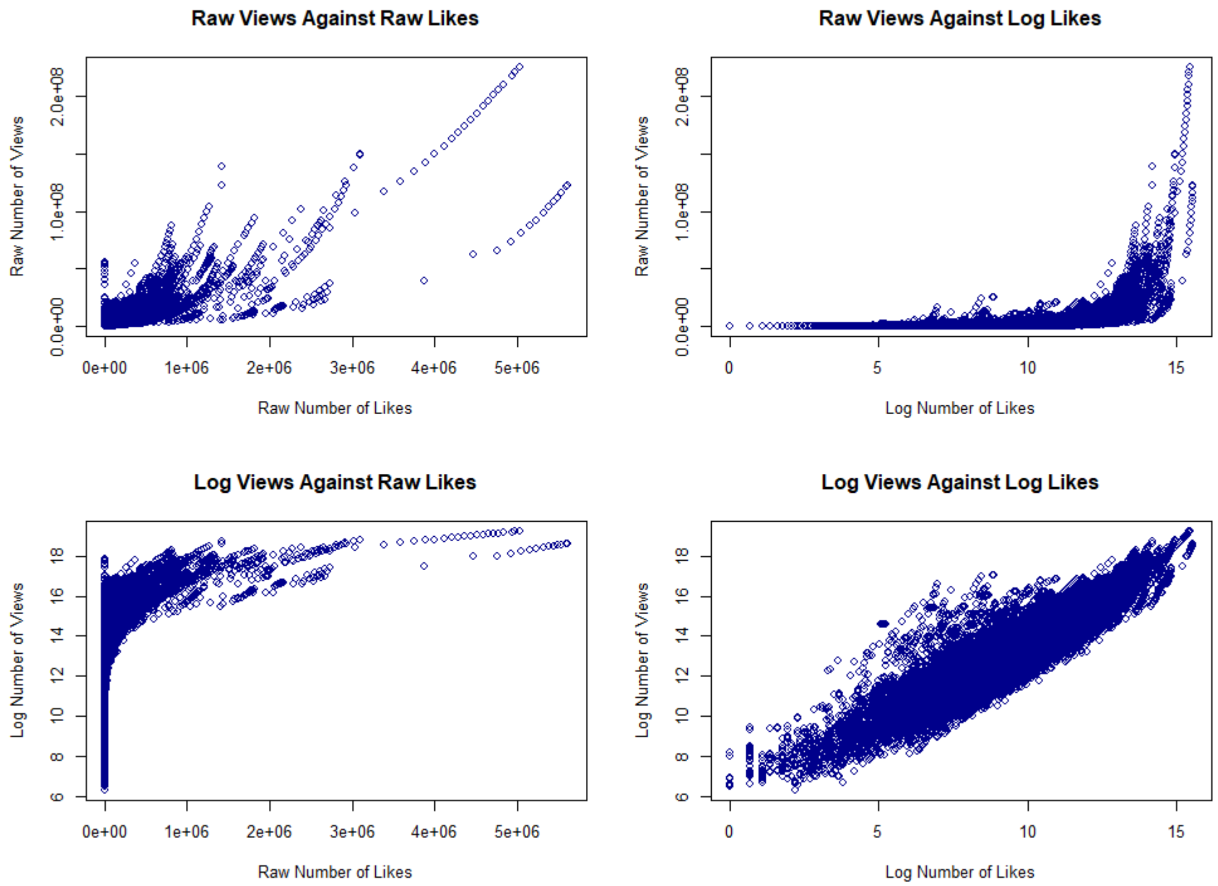
Which Genres Have the Highest Likes-to-Dislikes Ratios?

The Shows genre has the highest ratio of likes-to-dislikes, followed by Pets & Animals and Education

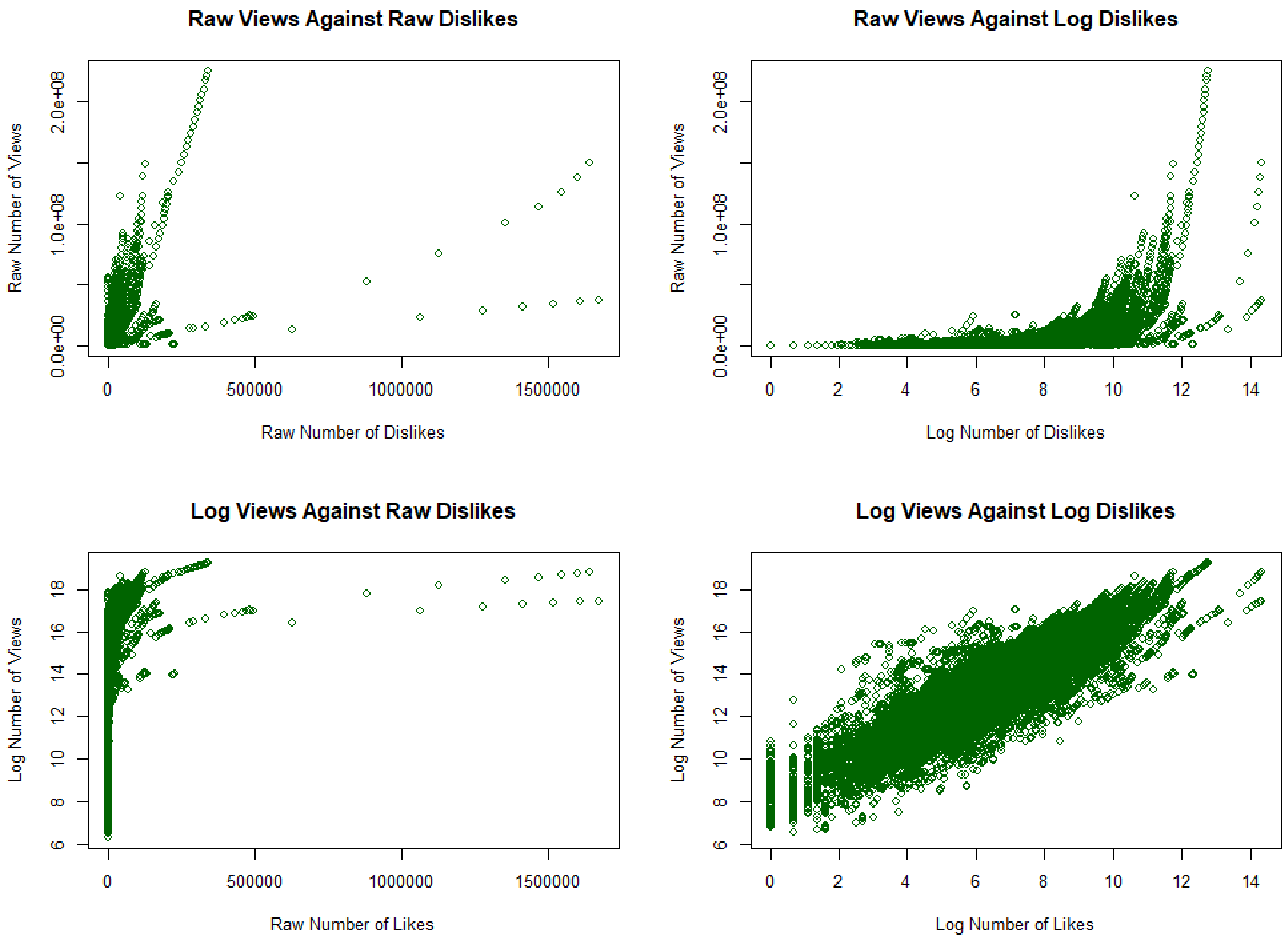


Evaluating Likes Against Views

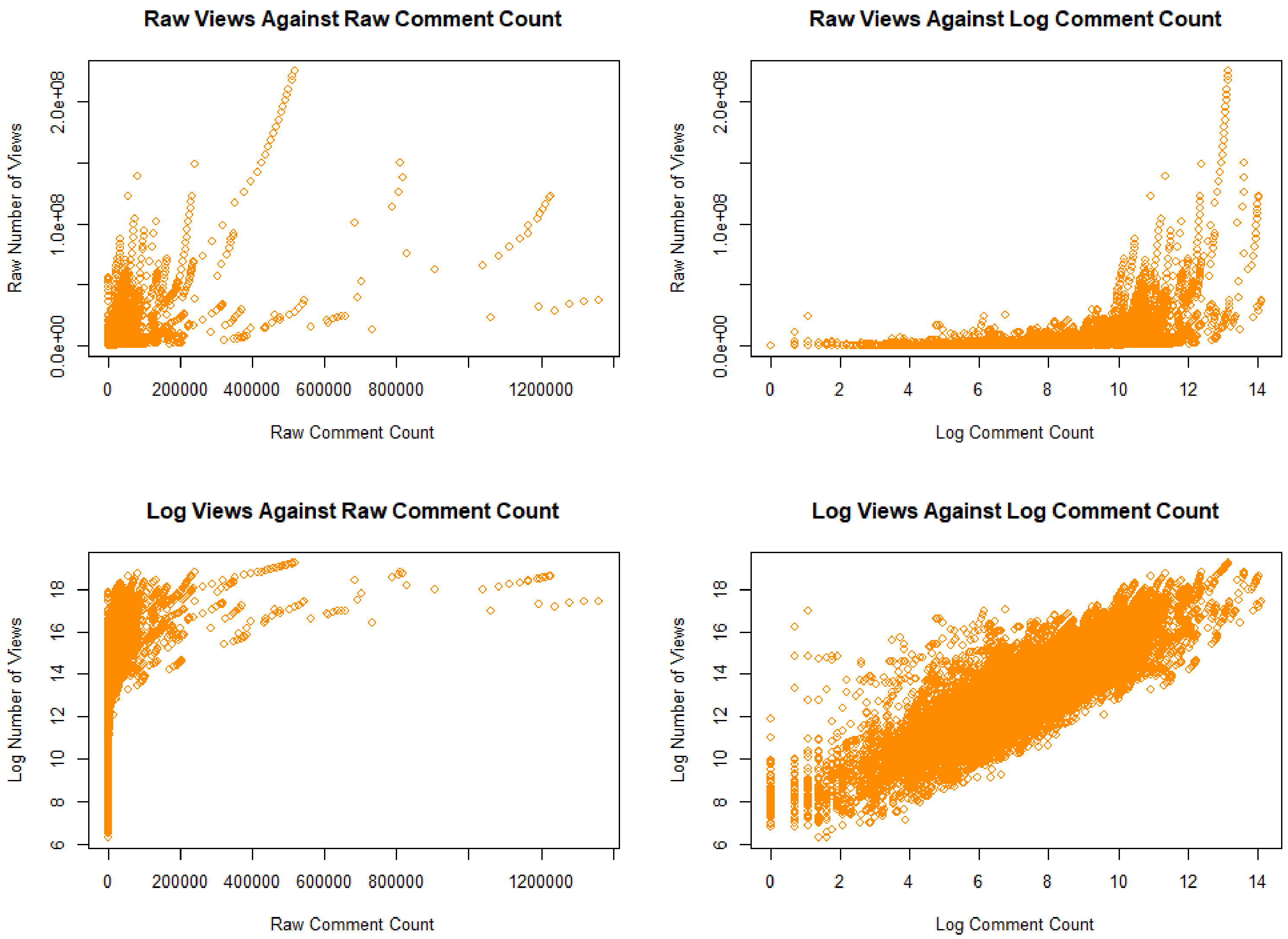
- The raw data plot led to me investigate log charts, at least with the log of views
- Plotting log/log produced a plot with a linear correlation (% change vs. % change)



Evaluating Dislikes Against Views

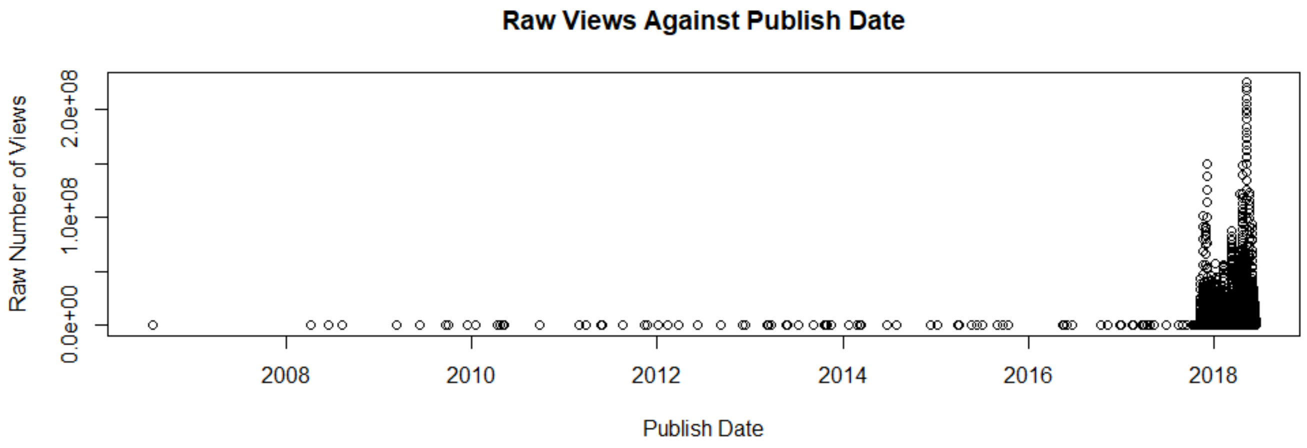


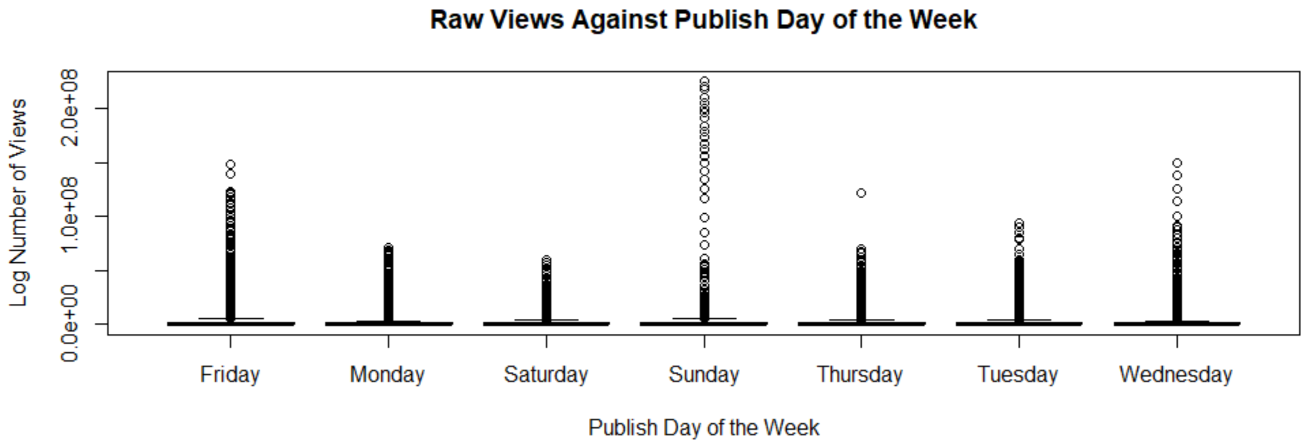
Evaluating Comment Count Against Views



Evaluating Publishing Time Against Views

- Most of the data in the set is for videos published in 2018
- Sunday is the most popular day to publish new videos
  - Drive Monday morning views
- Saturday is the least popular day to publish new videos
  - Likely harder to capture viewer attention on the weekend

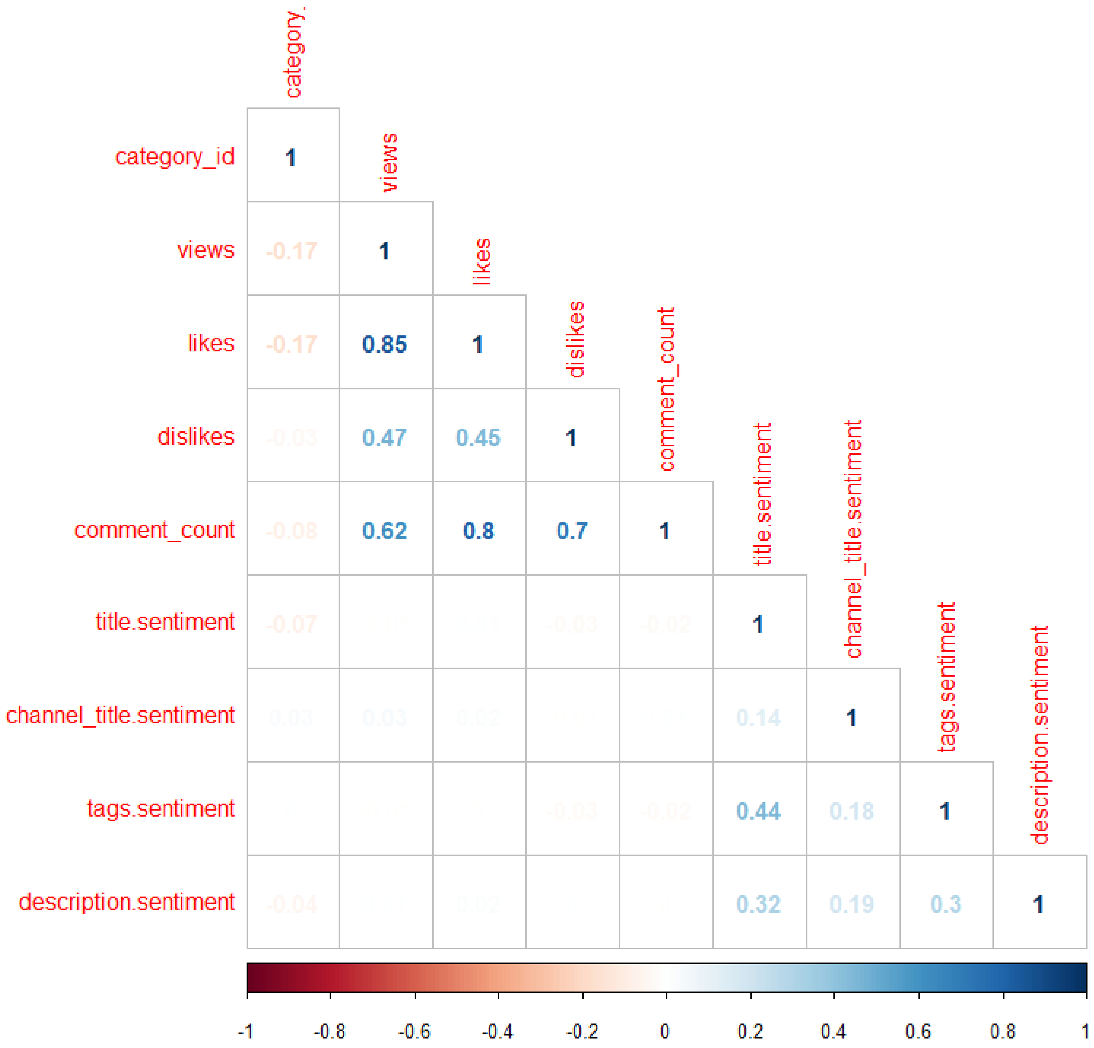




Quantitative Analysis

Correlation Assessment

- As expected, there is a strong correlation between likes and views (.85), but also between likes and comment_count (.80)
- Views and dislikes had a weaker correlation (.47) which was predictable
- Title.sentiment and tags.sentiment had some correlation (.44) which is not surprising since tags by nature summarize content key themes



Linear Regression Results

- We start by analyzing three key variables: likes, dislikes, and comment count
- All three variables are highly statistically significant
- Surprisingly, comment count has a strongly negative affect on views!

```
> model1=lm(views~likes+dislikes+comment_count,data=us.data.filter)
> summary(model1)

Call:
lm(formula = views ~ likes + dislikes + comment_count, data = us.data.filter)

Residuals:
      Min        1Q    Median        3Q       Max
-39610573   -385209   -171146    178202  86661340

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.156e+05  1.774e+04   12.15   <2e-16 ***
likes          3.556e+01  1.273e-01  279.30   <2e-16 ***
dislikes       8.311e+01  8.384e-01   99.13   <2e-16 ***
comment_count -9.765e+01  9.758e-01 -100.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3381000 on 40289 degrees of freedom
Multiple R-squared:  0.792,	Adjusted R-squared:  0.792
F-statistic: 5.113e+04 on 3 and 40289 DF,  p-value: < 2.2e-16
```

Linear Regression Results: Likes Only

- Given the very high t-value for likes, I assessed it directly against views. Alone, it explains about 73% of variation to the fitted regression line

```
> model2=lm(views~likes,data=us.data.filter)
> summary(model2)

Call:
lm(formula = views ~ likes, data = us.data.filter)

Residuals:
      Min        1Q    Median        3Q       Max 
-67472176   -422292   -249184     99565  99901610 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.966e+05  2.020e+04   14.68   <2e-16 ***
likes       2.745e+01  8.328e-02  329.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3855000 on 40291 degrees of freedom
Multiple R-squared:  0.7295,	Adjusted R-squared:  0.7295 
F-statistic: 1.087e+05 on 1 and 40291 DF,  p-value: < 2.2e-16
```

Linear Regression Results: Comment Count Only

- Given the negative coefficient for comment count, I assessed it directly against views. The coefficient was positive which is logical. I also checked for collinearity and found a moderate amount based on VIF scores

```
> model4=lm(views~comment_count,data=us.data.filter)
> summary(model4)

Call:
lm(formula = views ~ comment_count, data = us.data.filter)

Residuals:
      Min        1Q    Median        3Q       Max
-130043215  -1250654   -999653   -310887 160738463

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.311e+06  2.967e+04   44.17   <2e-16 ***
comment_count 1.221e+02  7.670e-01  159.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5807000 on 40291 degrees of freedom
Multiple R-squared:  0.3862,	Adjusted R-squared:  0.3862
F-statistic: 2.535e+04 on 1 and 40291 DF,  p-value: < 2.2e-16

> vif(model3)
                            GVIF Df GVIF^(1/(2*Df))
likes                   3.366670  1        1.834849
dislikes                2.130074  1        1.459477
comment_count           4.936731  1        2.221876
genre                   1.365585 15        1.010440
trending_weekday        1.018752  6        1.001549
publish_date_weekday    1.125906  6        1.009931
title.sentiment         1.333748  1        1.154880
channel_title.sentiment 1.084158  1        1.041229
tags.sentiment          1.345860  1        1.160112
description.sentiment   1.203213  1        1.096911
```

Potential interaction effect?

Moderate collinearity

Regression Using a Larger Data Set (with Comment_Count)

- The larger data set includes genre types, days of the week, and sentiment. Adjusted R^2 is .797
- The most significant predictors are the variables already discussed; however a few others are also quite significant:
  - Comedy genre
  - Music genre
  - Nonprofits & Activism genre
  - Content title sentiment
  - Channel title sentiment
- “Trending day of week” variables were all statistically insignificant
- Tuesday and Thursday publish date variables were the only significant days of the week

```
Call:
lm(formula = views ~ ., data = us.data.1)

Residuals:
     Min        1Q    Median        3Q       Max 
-40731740   -601027   -129997    358184  86228808 

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                  1.367e+06  1.823e+05    7.499 6.56e-14 ***
likes                        3.617e+01  1.324e-01  273.214  < 2e-16 ***
dislikes                     8.372e+01  8.302e-01  100.849  < 2e-16 ***
comment_count               -1.000e+02  9.801e-01 -102.079  < 2e-16 ***
genreComedy                 -1.517e+06  1.812e+05   -8.372  < 2e-16 ***
genreEducation              -1.388e+06  1.908e+05   -7.274 3.56e-13 ***
genreEntertainment          -7.333e+05  1.753e+05   -4.184 2.87e-05 ***
genreFilm & Animation       -1.003e+05  1.856e+05   -0.541 0.588782    
genreGaming                 -7.194e+05  2.084e+05   -3.452 0.000557 ***
genreHowto & Style          -1.182e+06  1.796e+05   -6.580 4.76e-11 ***
genreMusic                  -1.632e+06  1.779e+05   -9.173  < 2e-16 ***
genreNews & Politics        -9.173e+05  1.859e+05   -4.936 8.02e-07 ***
genreNonprofits & Activism  -4.293e+06  4.924e+05   -8.719  < 2e-16 ***
genrePeople & Blogs         -1.426e+06  1.819e+05   -7.838 4.69e-15 ***
genrePets & Animals         -9.363e+05  2.045e+05   -4.579 4.68e-06 ***
genreScience & Technology   -6.686e+05  1.854e+05   -3.607 0.000310 ***
genreShows                  -7.369e+05  4.753e+05   -1.550 0.121056    
genreSports                 -4.778e+05  1.863e+05   -2.564 0.010348 *  
genreTravel & Events        -4.966e+05  2.397e+05   -2.071 0.038331 *  
trending_weekdayMonday       2.325e+04  6.271e+04    0.371 0.710757    
trending_weekdaySaturday    -6.243e+04  6.208e+04   -1.006 0.314600    
trending_weekdaySunday      -4.673e+04  6.265e+04   -0.746 0.455683    
trending_weekdayThursday    -1.457e+04  6.263e+04   -0.233 0.816072    
trending_weekdayTuesday      1.943e+04  6.221e+04    0.312 0.754779    
trending_weekdayWednesday   -3.523e+03  6.271e+04   -0.056 0.955194    
publish_date_weekdayMonday   2.892e+04  5.970e+04    0.484 0.628038    
publish_date_weekdaySaturday -6.160e+04  7.040e+04   -0.875 0.381606    
publish_date_weekdaySunday  -1.968e+05  6.988e+04   -2.816 0.004861 ** 
publish_date_weekdayThursday -2.705e+05  5.740e+04   -4.712 2.46e-06 ***
publish_date_weekdayTuesday -2.341e+05  5.831e+04   -4.015 5.97e-05 ***
publish_date_weekdayWednesday 6.412e+03  5.810e+04    0.110 0.912121    
title.sentiment             -6.796e+05  7.783e+04   -8.731  < 2e-16 ***
channel_title.sentiment      8.268e+05  9.524e+04    8.681  < 2e-16 ***
tags.sentiment              -4.695e+04  4.656e+04   -1.008 0.313296    
description.sentiment       -2.180e+05  1.065e+05   -2.047 0.040663 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3340000 on 40258 degrees of freedom
Multiple R-squared:  0.7971,	Adjusted R-squared:  0.797 
F-statistic:  4653 on 34 and 40258 DF,  p-value: < 2.2e-16
```

Regression Using a Larger Data Set (without Comment_Count)

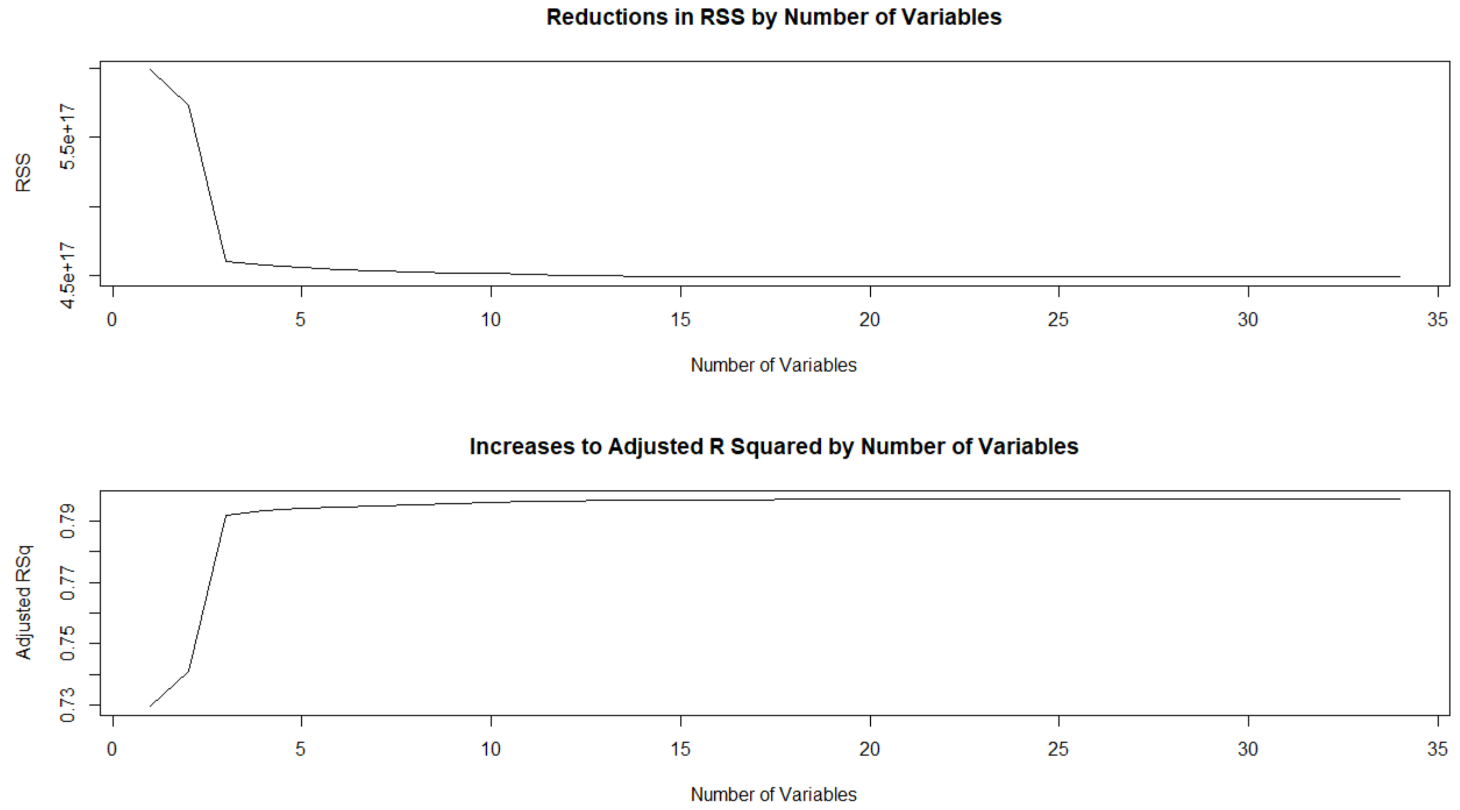
- Removing comment_count dropped adjusted R^2 from .797 to .7444
- The t-value of dislikes decreased. The Nonprofits & Activism genre t-score grew
- So, decision time: should we keep the comment_count variable or remove it?
  - It seems logical that more comments would lead to more views (so drop it, right?)
  - However, adjusted R^2 fell by 6.6%
  - I decided to execute the model improvement and optimal model analyses using both scenarios (including it and dropping it) to understand how the results might differ

```
Call:
lm(formula = views ~ ., data = us.data.2)

Residuals:
      Min        1Q    Median        3Q       Max
-63126337   -583215   -191808    246348  98430187

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.254e+06  2.045e+05   6.129 8.94e-10 ***
likes                       2.576e+01  9.465e-02 272.122  < 2e-16 ***
dislikes                    2.979e+01  7.185e-01  41.467  < 2e-16 ***
genreComedy                -1.314e+06  2.033e+05  -6.462 1.05e-10 ***
genreEducation             -1.273e+06  2.140e+05  -5.948 2.74e-09 ***
genreEntertainment         -6.128e+05  1.966e+05  -3.116  0.00183 **
genreFilm & Animation       9.330e+04  2.082e+05   0.448  0.65406
genreGaming                -9.675e+05  2.338e+05  -4.138 3.50e-05 ***
genreHowto & Style         -1.193e+06  2.015e+05  -5.917 3.30e-09 ***
genreMusic                 -7.940e+05  1.994e+05  -3.983 6.81e-05 ***
genreNews & Politics       -8.965e+05  2.085e+05  -4.299 1.72e-05 ***
genreNonprofits & Activism -6.983e+06  5.517e+05 -12.657  < 2e-16 ***
genrePeople & Blogs        -1.353e+06  2.041e+05  -6.627 3.48e-11 ***
genrePets & Animals        -8.965e+05  2.294e+05  -3.908 9.31e-05 ***
genreScience & Technology  -6.553e+05  2.080e+05  -3.151  0.00163 **
genreShows                 -6.706e+05  5.333e+05  -1.257  0.20861
genreSports                -3.105e+05  2.090e+05  -1.485  0.13751
genreTravel & Events       -4.989e+05  2.690e+05  -1.855  0.06364 .
trending_weekdayMonday      3.712e+04  7.035e+04   0.528  0.59778
trending_weekdaySaturday   -6.296e+04  6.965e+04  -0.904  0.36603
trending_weekdaySunday     -3.923e+04  7.029e+04  -0.558  0.57679
trending_weekdayThursday    4.356e+03  7.027e+04   0.062  0.95057
trending_weekdayTuesday     4.075e+04  6.980e+04   0.584  0.55936
trending_weekdayWednesday   9.654e+03  7.035e+04   0.137  0.89085
publish_date_weekdayMonday -3.244e+04  6.698e+04  -0.484  0.62813
publish_date_weekdaySaturday -1.262e+05  7.899e+04  -1.598  0.11010
publish_date_weekdaySunday -1.167e+05  7.839e+04  -1.489  0.13657
publish_date_weekdayThursday -1.551e+05  6.439e+04  -2.410  0.01598 *
publish_date_weekdayTuesday -2.085e+05  6.542e+04  -3.187  0.00144 **
publish_date_weekdayWednesday -7.956e+04  6.518e+04  -1.221  0.22222
title.sentiment            -5.451e+05  8.731e+04  -6.243 4.33e-10 ***
channel_title.sentiment     8.627e+05  1.069e+05   8.074 7.00e-16 ***
tags.sentiment              9.771e+03  5.223e+04   0.187  0.85161
description.sentiment      -1.856e+05  1.195e+05  -1.554  0.12030
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3747000 on 40259 degrees of freedom
Multiple R-squared:  0.7446,    Adjusted R-squared:  0.7444
F-statistic:  3557 on 33 and 40259 DF,  p-value: < 2.2e-16
```

Model Improvement and Optimal Model Including the Comment_Count Variable

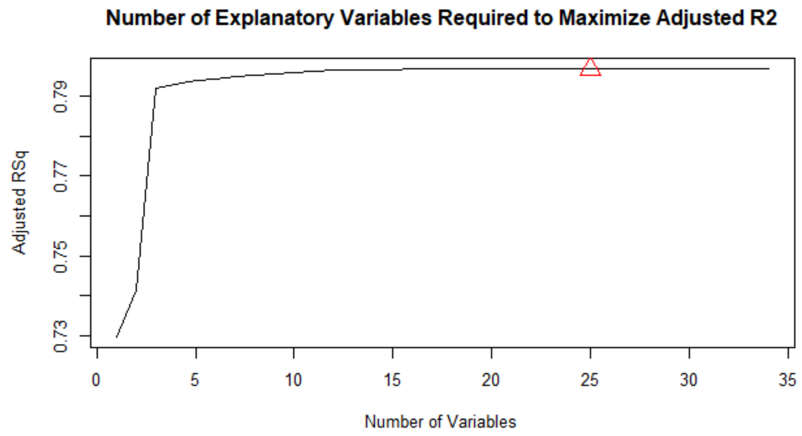
Model Improvement

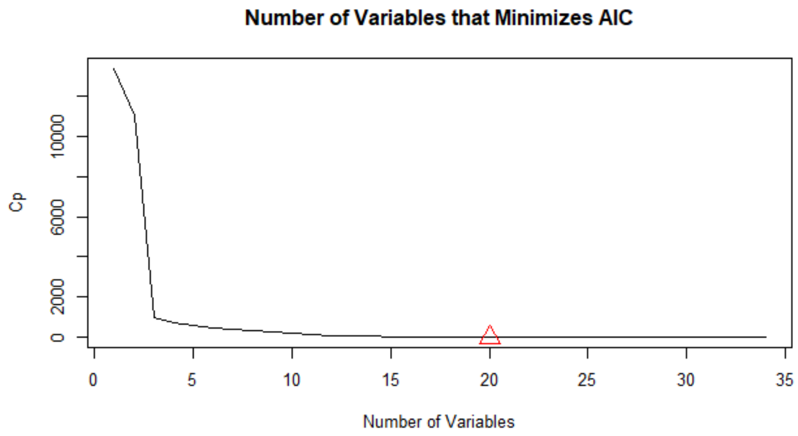
- I next sought to determine whether it was possible to improve the simple linear model by comparing the results of best subset selection, forward selection, and backward selection

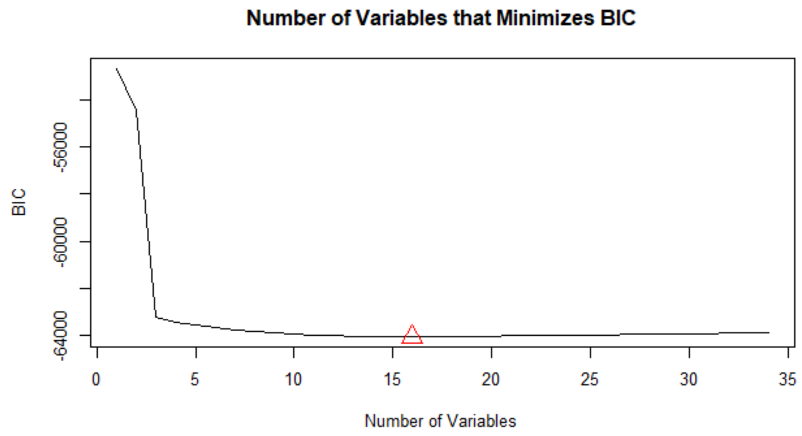


Model Improvement

- I evaluated the number of variables required to maximize Adj. R^2 and minimize both AIC and BIC







Stepwise Selection

- I used stepwise selection to identify more restrictive models that still had high explanatory power with a lower likelihood of over-fitting on the test set
- Forward and backward stepwise selection revealed that at least the first 6 common variables had the same coefficients

```
> coef(regfit.full.yt,6)
              (Intercept)                     likes                  dislikes             comment_count       genreEntertainment 
             1.112291e+05              3.560300e+01              8.304395e+01             -9.748289e+01             4.080328e+05 
genreNonprofits & Activism    trending_weekdayMonday 
            -3.225255e+06              4.046112e+04 
> coef(regfit.fwd.yt,6)
              (Intercept)                     likes                  dislikes             comment_count       genreEntertainment 
             1.112291e+05              3.560300e+01              8.304395e+01             -9.748289e+01             4.080328e+05 
genreNonprofits & Activism    trending_weekdayMonday 
            -3.225255e+06              4.046112e+04 
> coef(regfit.bwd.yt,6)
              (Intercept)                     likes                  dislikes             comment_count       genreEntertainment 
             1.112291e+05              3.560300e+01              8.304395e+01             -9.748289e+01             4.080328e+05 
genreNonprofits & Activism    trending_weekdayMonday 
            -3.225255e+06              4.046112e+04 
```

Optimal Model

- The last step in the model evaluation process was to use cross-validation. The cross-validation approach selected a 6-variable model
- The included genres had the largest coefficients, although the raw number of likes, dislikes, and comments have a major impact on views in the final model given the high values for these variables in the data

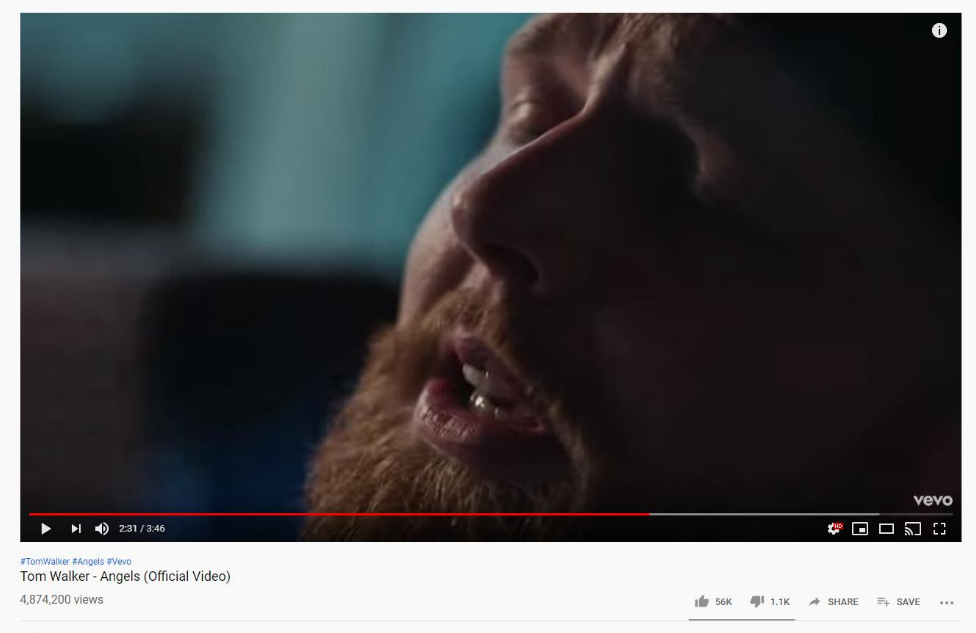
```
> coef(reg.best,6)
            (Intercept)                   likes                dislikes           comment_count      genreEntertainment
           1.112291e+05            3.560300e+01            8.304395e+01           -9.748289e+01            4.080328e+05
genreNonprofits & Activism   trending_weekdayMonday
          -3.225255e+06            4.046112e+04
```

Using the Optimal Model - Hypothetical

- For comparison with the data set that later excludes comment_count, we look at a hypothetical video with 1,000 likes, 500 dislikes, 30 comments, in the Music genre
- The model expects 185,429 views
  - hypothetical.views= (1.112291e+05)+((3.560300e+01)*1000) + ((8.304395e+01)*500) + ((-9.748289e+01)*30)

Using the Optimal Model – Music Video

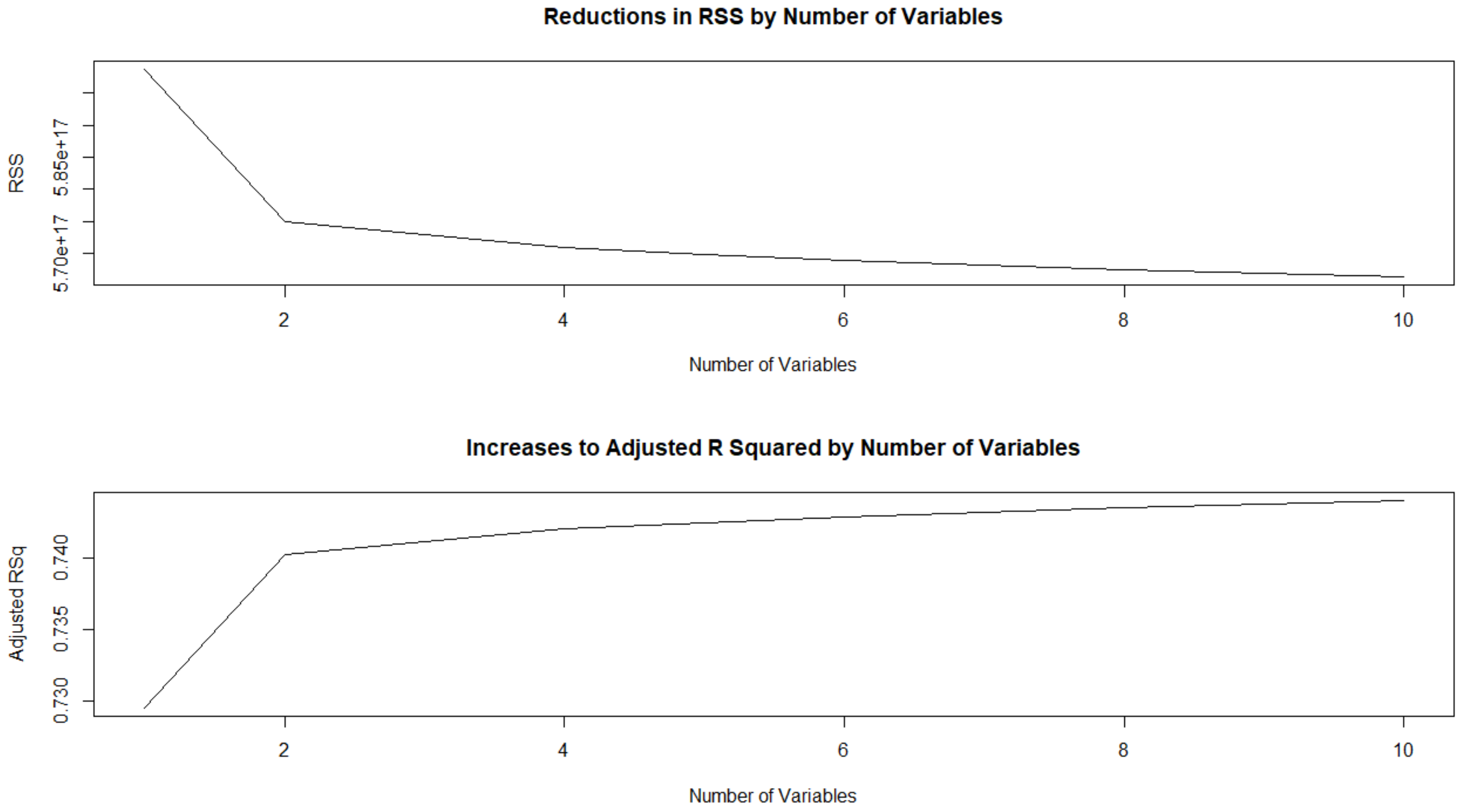
- We test a moderately successful music video from 2018
  - "Angels" by Tom Walker
- The model expects 2,100,227 views
  - (1.112291e+05)+((3.560300e+01)*56000) + ((8.304395e+01)*1100) + ((-9.748289e+01)*986)
- Actual views were 4,874,200
- So not a great prediction



Model Improvement and Optimal Model Dropping the Comment_Count Variable

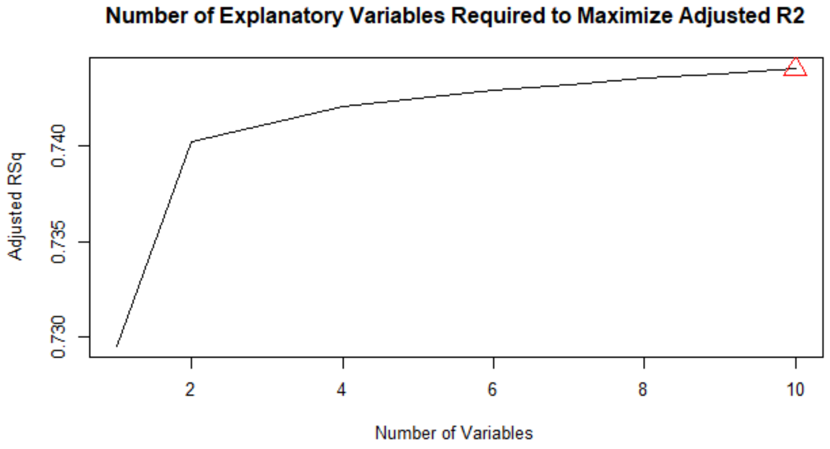
Model Improvement

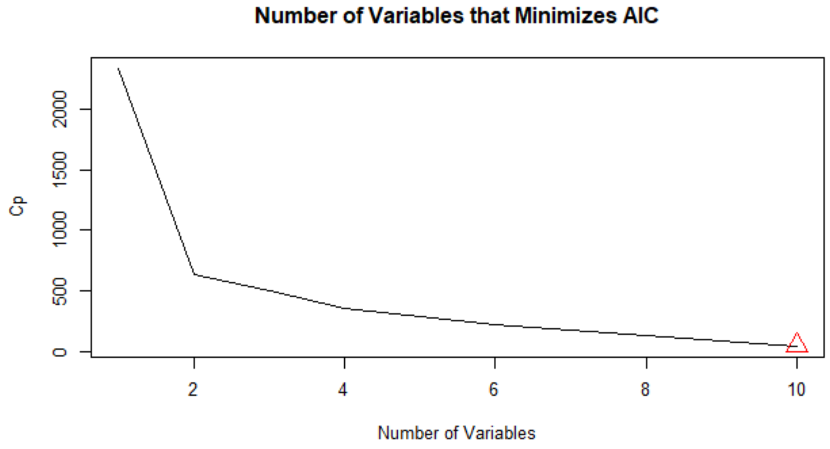
› I next determined how to improve the simple linear model by comparing the results of best subset selection, forward selection, and backward selection

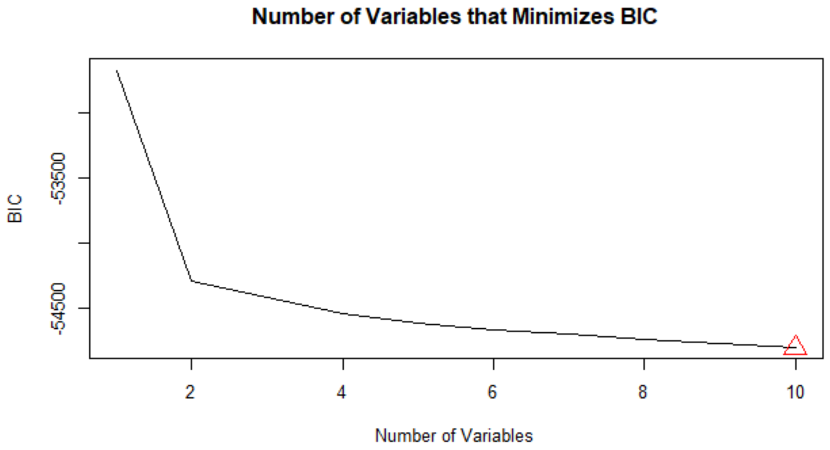


Model Improvement

- I evaluated the number of variables required to maximize Adj. R^2 and minimize both AIC and BIC







Stepwise Selection

› Forward and backward stepwise selection revealed that at least the first 10 common variables had the same coefficients (vs. 6 when including comment_count)

```
> coef(regfit.full.yt1,10)
             (Intercept)                    likes                 dislikes      genreEntertainment      genreHowto & Style
            3.001120e+05             2.579793e+01             2.973625e+01            2.549486e+05           -3.652998e+05
genreNonprofits & Activism              genreShows    trending_weekdaySunday  trending_weekdayThursday           genreThriller
           -6.173997e+06             1.084551e+05            -4.472987e+04            2.352725e+02            0.000000e+00
           genreTrailers
            0.000000e+00
> # Maximum Adjusted R2
> coef(regfit.full.yt1,10)
             (Intercept)                    likes                 dislikes      genreEntertainment      genreHowto & Style
            3.001120e+05             2.579793e+01             2.973625e+01            2.549486e+05           -3.652998e+05
genreNonprofits & Activism              genreShows    trending_weekdaySunday  trending_weekdayThursday           genreThriller
           -6.173997e+06             1.084551e+05            -4.472987e+04            2.352725e+02            0.000000e+00
           genreTrailers
            0.000000e+00
> # Minimum AIC
> coef(regfit.full.yt1,10)
             (Intercept)                    likes                 dislikes      genreEntertainment      genreHowto & Style
            3.001120e+05             2.579793e+01             2.973625e+01            2.549486e+05           -3.652998e+05
genreNonprofits & Activism              genreShows    trending_weekdaySunday  trending_weekdayThursday           genreThriller
           -6.173997e+06             1.084551e+05            -4.472987e+04            2.352725e+02            0.000000e+00
           genreTrailers
            0.000000e+00
> |
```

Optimal Model

- The last step in the model evaluation process was to use cross-validation. The cross-validation approach selected a 3-variable model

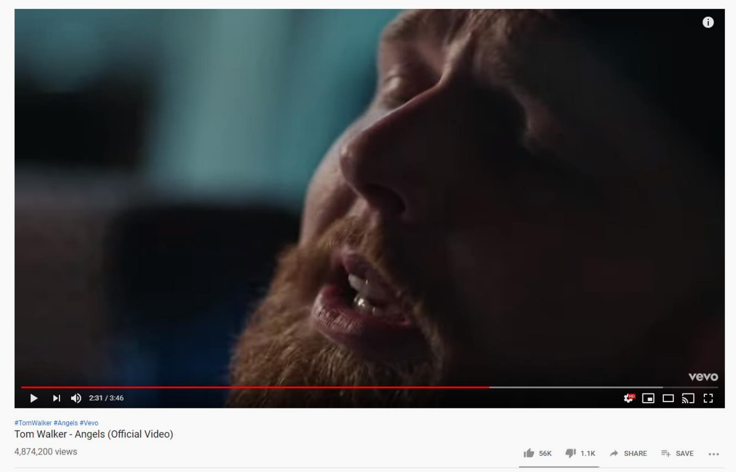
```
> coef(reg.best1,3)
        (Intercept)                  likes               dislikes trending_weekdaySunday 
       319092.17332               25.78769               29.36622           -48254.61414 
> hypothetical.views.3b=(319092.17332)+((25.78769)*1000)+((29.36622)*500)+(-48254.61414)
```

Using the Optimal Model - Hypothetical

- Same as the prior example, we look at a hypothetical video with 1,000 likes, 500 dislikes, 30 comments, in the Music genre
- The model expects 311,308 views
  - hypothetical.views= (319,092) + (25.78769*1000) + (29.36622*500)
- These 311,308 views compare to the 185,429 views from the model which includes comment_count, a difference of 125,879 views
- How can we attribute the difference? We check the regression results:
  - The intercepts differed by 200,079
  - The likes coefficients differed by -9.815 (x1000) = -9,815
  - The dislikes coefficients differed by -53.678 (x500) = -26,839
  - The comment_count coefficient differed by -97.4829 (x30) = -2,924
- Removing comment_count as a variable increased the resulting video view estimate by 68% in this hypothetical example

Using the Optimal Model – Music Video

- We again test "Angels" by Tom Walker for consistency
- The model expects 1,795,505 views
  - (319,092) + (25.78769*56000) + (29.36622*1100)
- Actual views were (again) 4,874,200
- This was somewhat worse than the model which included comment_count
- Neither model was very accurate unfortunately using this example
- A less popular video might offer a more accurate prediction



Thank You

Questions?