YOUTUBE TRENDS

DSC 423 Data Regression and Analysis

3/16/20

Baha Gharbi

Evelina Ramoskaite

Stephen Kim

Contents

Introduction

YouTube Background

YouTube is an American video-sharing platform founded on February 14, 2005 and headquartered in San Bruno, California, founded by three former PayPal employees, Chad Hurley, Steve Chen, and Jawed Karim. Google bought the site in November 2006 for US$1.65 billion; and YouTube now operates as one of Google's subsidiaries. The total number of people who use YouTube is currently over 1,300,000,000. More than 300 hours of video are uploaded to YouTube every minute and over five billion videos are viewed on YouTube every single day. In an average month, 8 out of 10 18-49-year-olds watch YouTube.

As the second most visited website in the world behind Google.com, people have recognized that the amount of traffic generated by posting media can be monetized, and YouTube creators are able to make a comfortable living just by posting videos. In 2019, the highest earning creator was an 8-year-old boy named Ryan Kaji, who created videos of himself opening boxes of toys. People enjoyed these videos so much that he created other sources of income such as toy stores, clothing brands, podcasts, etc. Kaji ended up being the highest earning YouTuber at $26 million dollars that year. While not everyone can make that kind of money, a substantial amount can be made and that it is possible for any person to start their journey of earning income through this avenue.

Research Topics Chosen

While there are many more possible subjects, three different topics were chosen for this paper. Stephen Kim opted to find if there is a correlation between the type of category a video is in and its popularity, or the number of views, in the United States. Baha Gharbi researched the influence of comments on the number of likes in different countries. Evelina Ramoskaite took another angle and chose to find whether there is a correlation between markers of controversy and video of popularity.

Possible topics of research for this dataset include:

1. Are certain genres more likely to be trending in different countries?
2. How many videos from this channel will be selected as trending videos next week?
3. Which channel will be most popular in category 1 next month?
4. How does comment involvement influence view count?
5. How do video titles and descriptions influence popularity?
6. Are longer descriptions more effective?
7. How does language influence popularity abroad?
8. Are there any common keywords within the titles, descriptions, or tags among trending videos?

Data Set Used

The dataset, taken from https://www.kaggle.com/datasnaek/youtube-new, includes several months of data on daily trending YouTube videos. According to the dataset creator, an automated script was run at 9AM GMT every day, which took about 30 seconds to collect all of the data.

Scope

The exact dates of the study are between 11/14/2017 and 05/31/2018. The data included represents ten different countries: USA, Great Britain, Russia, Mexico, Korea, Japan, Germany, Canada, France, and India.

Variables

The dataset includes the following variables:

Category	Description
video_id	ID code given by YouTube
trending_date	Date of the trending video
title	Title of the video given by the creator
channel_title	Title of channel video is a part of
category_id	ID code of the video category chosen by the creator
publish_time	Date that the video was published
tags	Searchable tags assigned by the creator
views	Number of views
likes	Number of likes
dislikes	Number of dislikes
comment_count	Number of comments
thumbnail_link	Thumbnail picture of the video
comments_disabled	True if comments are disabled, False otherwise
ratings_disabled	True if ratings are disables, False otherwise
video_error_or_removed	True if there is a video error or the videowas removed, False otherwise
Description	Full video description given by the creator

Figure 1 Variables in the dataset.

The dataset is observational, as none of the variables are manipulated, just recorded as they are seen. Each of the variables can be considered explanatory variables, except for the video_id, views, and thumbnail_link, meaning there are 13 independent variables that may explain variations in the response variable. The response variable is the outcome variable whose value is predicted or explained by the explanatory variables. In this study the response, or dependent, variable is the number of views, as it depicts the popularity of the video.

Cleaning the Data

Initial procedures used to clean the data included deleting blank cells and rows, removing any “Deleted videos,” which created duplicate entries, and finally finding and removing all the rows affected by the new line character. The CSV files were also converted to either ANSI or Unicode UTF-8 to view the original characters of the ten different languages. The category ID, representing the different genres of videos, were converted from JSON format to tables in Excel.

Genre Trends

Stephen Kim

YouTube video creators are given the option to select a category when uploading a video. This gives YouTube the ability to sort all the videos accordingly, and for people to be able to search for the videos they want to see.

Importance

While it might seem like a trivial matter, it is important to select the correct category so that the proper advertisers may find videos that may be synonymous with their brands, especially if the video starts to get popular, or is considered trending.

Advertising is one of the larger sources of income for YouTube creators, through the site itself, as well as external sources like Google AdSense. Monetizing a YouTube channel through advertisements has proven to have substantial benefits, and by growing viewer traffic and attracting advertisers, people can begin their journey of making money through posting videos.

One analysis of the YouTube trending video data involves tracking the different genres of the trending videos. In particular, the United States, which is the largest user of YouTube in the world, and also has the most money made by creators. By researching the records of all the trending videos over a six month period, it may be possible to find patterns or even predict what kind of videos have a higher chance of becoming viral in the US in the future. The qualification of becoming viral is to have at least 5 million hits within a three to seven day period, with the most viral video in history being the “Kony 2012” with 34,000,000 views within three days and 100,000,000 views in six days.

Variable Selection

The top trending videos around the world fall under 32 categories which are listed on the following page:

ID	Category
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
18	Short Movies
19	Travel & Events
20	Gaming
21	Videoblogging
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism
30	Movies
31	Anime/Animation
32	Action/Adventure
33	Classics
34	Comedy
35	Documentary
36	Drama
37	Family
38	Foreign
39	Horror
40	Sci-Fi/Fantasy
41	Thriller
42	Shorts
43	Shows
44	Trailers

Figure 2 Categories in the Data Set

Creating a frequency diagram of the categories in each of the different countries reveals that there are 18 categories with trending videos among the 10 countries. In particular, it can be noticed that the “Entertainment” category has the most overall trending videos, and is the largest category in all but two of the genres.

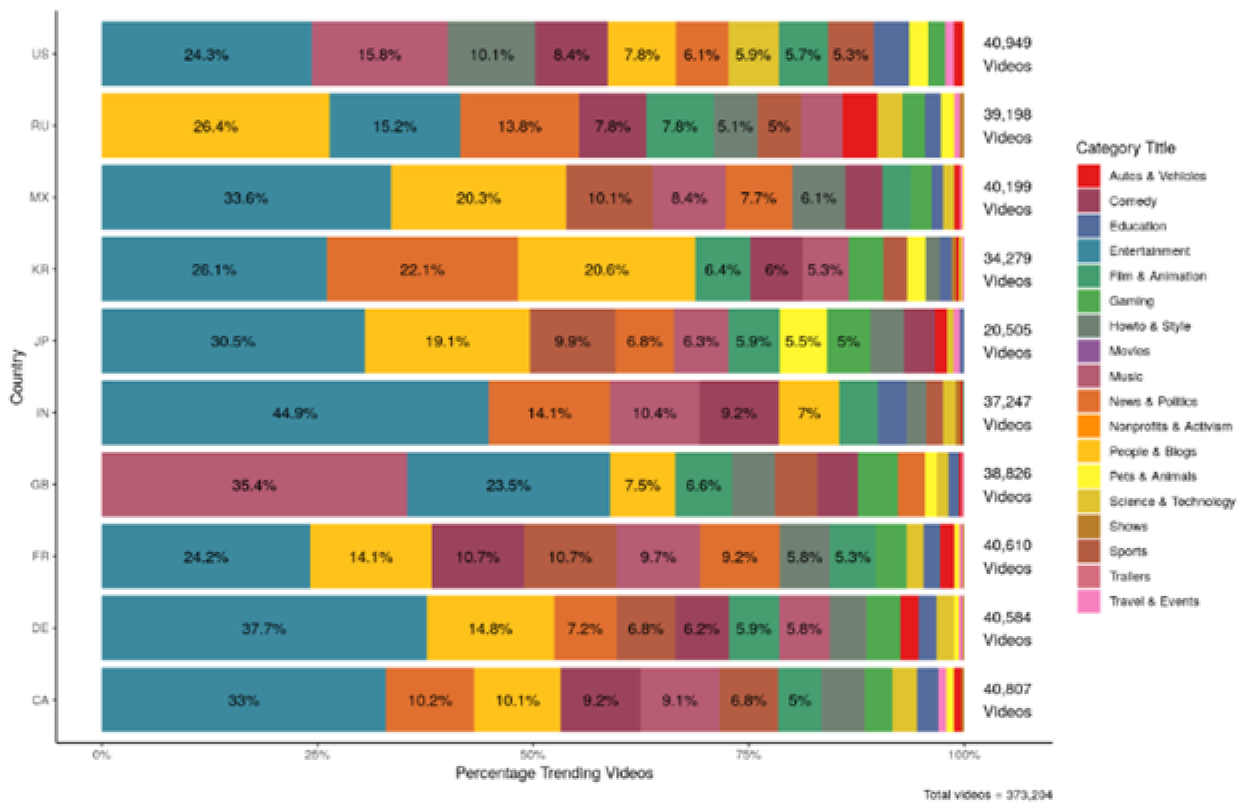


Figure 3 Visualization the most popular videos, divided by category for each county in the dataset.

Further inspection of the data reveals that certain categories are not considered to be trending at all during the six month dataset in certain countries. By tracking the amount of views, likes, dislikes, and trending videos for each category, it can be seen that some of the categories are populated by "0's," meaning several categories were never trending during the dataset's time period. This allows for the removal of those categories, further cleaning the dataset to create more accurate models, as is shown by the following tables:

ID	Category	Views	Likes	Dislikes	Videos
1	Film & Animation	2784070071	105033908	3767973	2344
2	Autos & Vehicles	650024670	23219694	701668	383
10	Music	8451285585	309359258	23461649	6471
15	Pets & Animals	1315915762	44295541	3094078	919
17	Sports	2316362895	89923157	12203359	2173
18	Short Movies	0	0	0	0
19	Travel & Events	351757502	12684881	6318	401
20	Gaming	815308153	32942379	1500740	816
21	Videoblogging	0	0	0	0
22	People & Blogs	3879038392	131137378	8503058	3209
23	Comedy	5388139019	142366546	6244676	3457
24	Entertainment	20253394764	691830817	25146631	9962
25	News & Politics	9279743892	258380376	10530040	2486
26	Howto & Style	18920579850	535158869	23146739	4165
27	Education	9321395922	269898584	13328092	1655
28	Science & Technology	12285110784	375682154	18816250	2400
29	Nonprofits & Activism	322100544	9664605	195390	56
30	Movies	0	0	0	0

Figure 4 Categories removed from the model

31	Anime/Animation	0	0	0	0
32	Action/Adventure	0	0	0	0
33	Classics	0	0	0	0
34	Comedy	0	0	0	0
35	Documentary	0	0	0	0
36	Drama	0	0	0	0
37	Family	0	0	0	0
38	Foreign	0	0	0	0
39	Horror	0	0	0	0
40	Sci-Fi/Fantasy	0	0	0	0
41	Thriller	0	0	0	0
42	Shorts	0	0	0	0
43	Shows	337542347	1342160	570835	56
44	Trailers	0	0	0	0

Figure 5 Categories removed from the model

This dataset can be reduced to a cleaner version, focusing only on the videos that do trend during the duration of the observations. By choosing the independent variables that are trending during the time period of the dataset, the number of categories can be reduced to 16.

ID	Category	Views	Likes	Dislikes	Videos
1	Film & Animation	2784070071	105033908	3767973	2344
2	Autos & Vehicles	650024670	23219694	701668	383
10	Music	8451285585	309359258	23461649	6471
15	Pets & Animals	1315915762	44295541	3094078	919
17	Sports	2316362895	89923157	12203359	2173
19	Travel & Events	351757502	12684881	6318	401
20	Gaming	815308153	32942379	1500740	816
22	People & Blogs	3879038392	131137378	8503058	3209
23	Comedy	5388139019	142366546	6244676	3457
24	Entertainment	20253394764	691830817	25146631	9962
25	News & Politics	9279743892	258380376	10530040	2486
26	Howto & Style	18920579850	535158869	23146739	4165
27	Education	9321395922	269898584	13328092	1655
28	Science & Technology	12285110784	375682154	18816250	2400
29	Nonprofits & Activism	322100544	9664605	195390	56
43	Shows	337542347	1342160	570835	56

Figure 6 Final list of the categories included in the model.

Creating sorted frequency diagrams for each dependent variable, as shown on the following pages, depicts the categories that appear to be the most popular among the trending videos. Now it is clear that the entertainment category has the most views, likes, dislikes, and total videos when compared to all the other trending categories. The other categories are not always in the same order in each of the graphs, though generally in the same area, indicating that the quantity of one of the variables may or may not have an impact on the other variables. There is also quite the disparage between the top categories and the lower range categories, indicating that certain trends may exist.

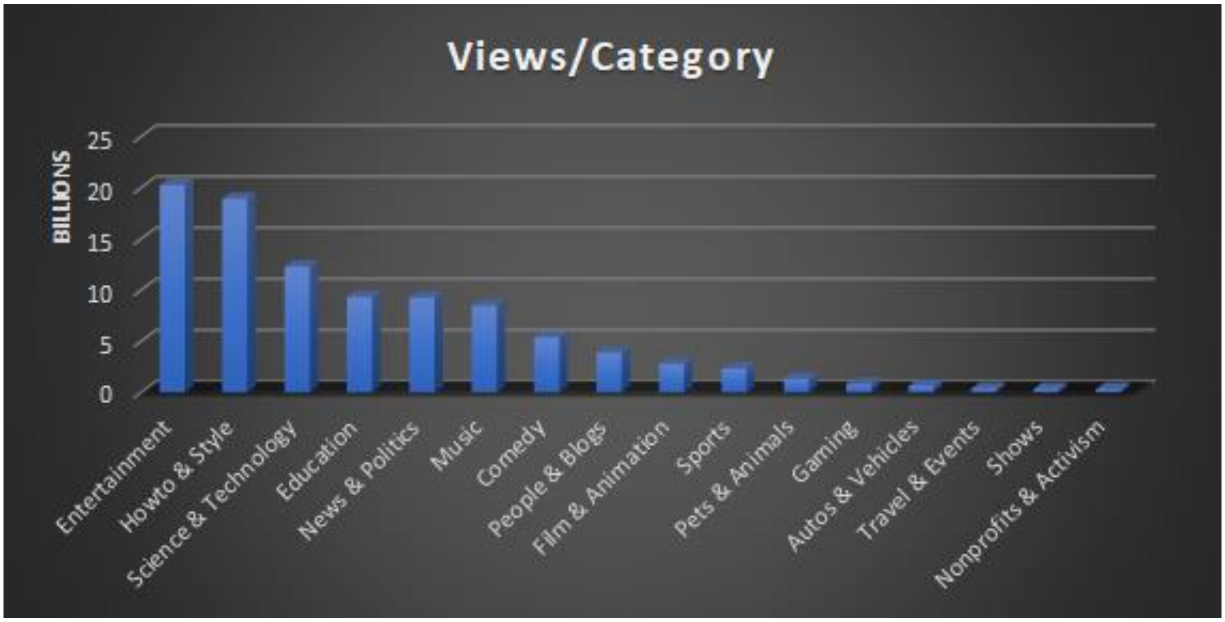


Figure 7 Views by category

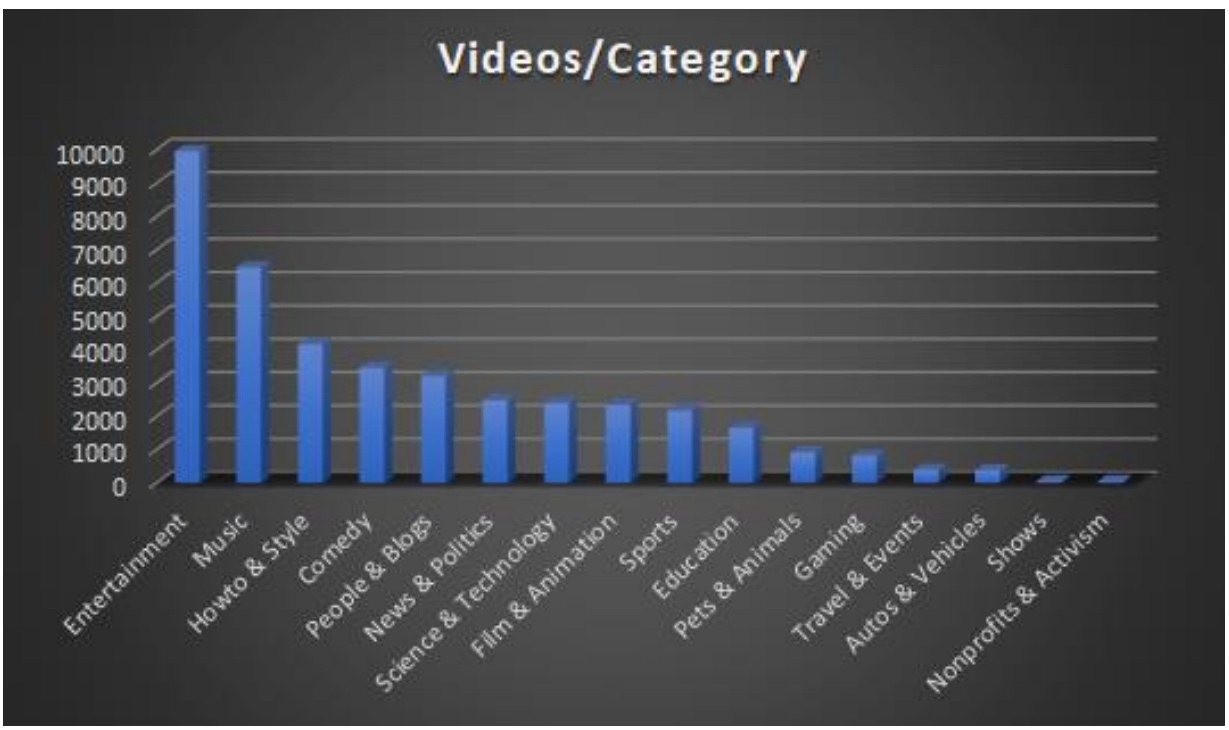


Figure 8 Video count by category

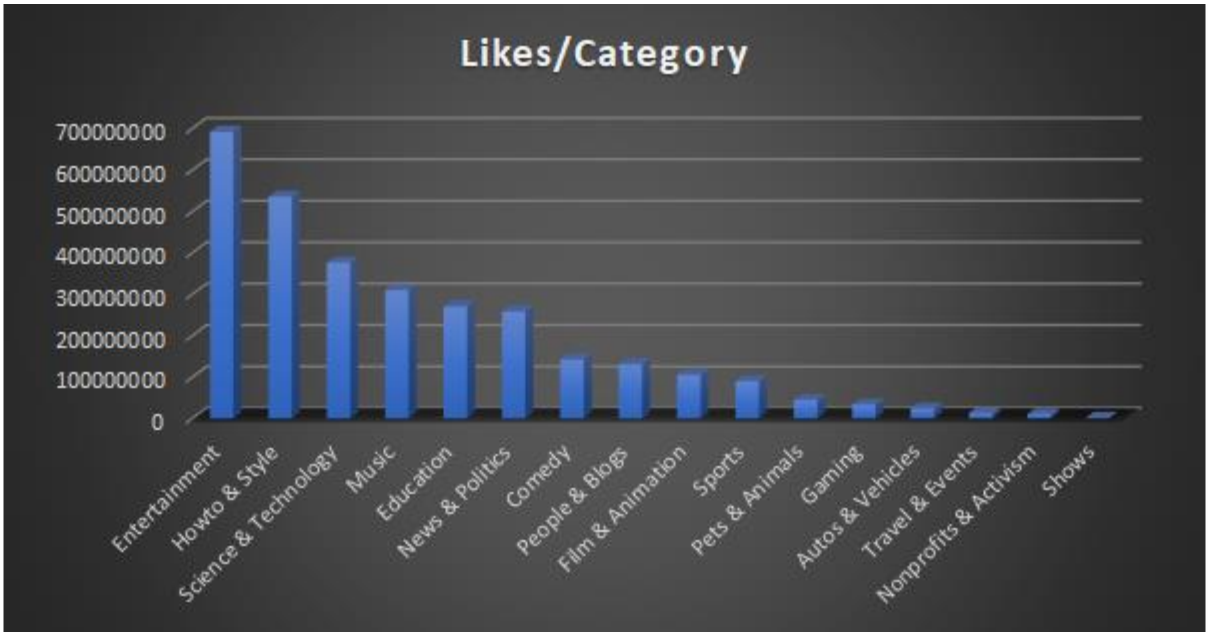


Figure 9 Likes by category.

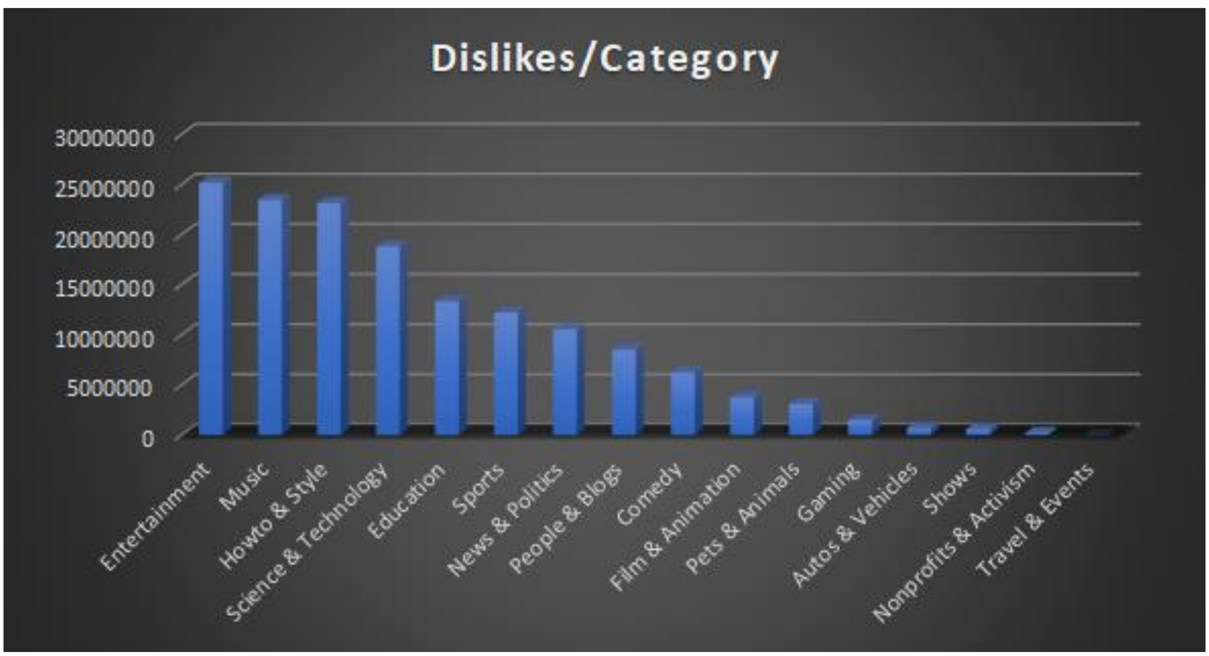


Figure 10 Dislike count by category.

One area to draw attention to is that for each bar graph, the same six categories are consistently at the far right of each graph: Pets & Animals, Gaming, Autos & Vehicles, Shows, Nonprofits & Activism, and Travel & Events. Looking back at the table of categories, those six categories all have less than 1000 videos, representing a small portion of all the trending videos. It is clear that these categories hold little weight to the data, and to clean the data even further, they are removed from the dataset.

ID	Category	Views	Likes	Dislikes	Videos
1	Film & Animation	2784070071	105033908	3767973	2344
10	Music	8451285585	309359258	23461649	6471
17	Sports	2316362895	89923157	12203359	2173
22	People & Blogs	3879038392	131137378	8503058	3209
23	Comedy	5388139019	142366546	6244676	3457
24	Entertainment	20253394764	691830817	25146631	9962
25	News & Politics	9279743892	258380376	10530040	2486
26	Howto & Style	18920579850	535158869	23146739	4165
27	Education	9321395922	269898584	13328092	1655
28	Science & Technology	12285110784	375682154	18816250	2400

Figure 11 Final Cleaned list of Categories used in the model.

Now that a reasonable dataset has been reached, and initial observations have been made, the data can be checked for multicollinearity. This last step of cleaning the data involves using the original dataset, which included over 40,000 recorded videos, and creating dummy variables for the different qualitative variables. Since there are 10 categories, there are *k - 1*, or 9, dummy variables, with the last category omitted:

Film & Animation = c_1 => 1 if true; 0 if false

Music = c_2 => 1 if true; 0 if false

Sports = c_3 => 1 if true; 0 if false

People & Blogs = c_4 => 1 if true; 0 if false

Comedy = c_5 => 1 if true; 0 if false

Entertainment = c_6 => 1 if true; 0 if false

News & Politics = c_7 => 1 if true; 0 if false

HowTo & Style = c_8 => 1 if true; 0 if false

Education = c_9 => 1 if true; 0 if false

Science & Technology = omitted

Model Validation

To check for multicollinearity, the Variance Inflation Factor(VIF) must be calculated for each of the independent variables. By running a regression model for each independent variable against all other independent variables, the Coefficient of Determination (R^2) is found for each specific variable. Using the equation for VIF,

$$VIF = 1/(1 - R^2)$$

each variable can be checked for multicollinearity as shown in the results below created through Excel. Each independent variable, both quantitative and qualitative are shown along with their VIF and R^2 value when regressed against the other independent variables.

	VIF	IV Corr R^2
Intercept		
Education	1.615	0.381
HowTo & Style	2.426	0.588
News & Politics	1.904	0.475
Entertainment	3.805	0.737
Comedy	2.235	0.553
People & Blogs	2.156	0.536
Sports	1.480	0.324
Music	3.088	0.676
Film & Animation	1.866	0.464
Dislikes	1.296	0.228
Likes	1.328	0.247

Figure 12 VIF Values for each of the categories in the model.

The calculated VIF for each independent variable, as shown in the previous output is below 10. Various authors maintain that, in practice, a severe multicollinearity problem exists if the largest VIF for the β variables is greater than 10 or if the R^2 is greater than 0.90. Since no VIF or R^2 reaches that amount, it is implied that the independent variables are not multicollinear; however, some authors will argue that a VIF of 3 or 5 may be signs for multicollinearity, and since there are a couple variables (Entertainment and Music) that show a VIF above 3, further checks should be made on the dataset.

By creating the correlation matrix of the r values of each variable when compared to all other variables, it can be seen once again that there exists little correlation besides “likes” and “dislikes” relating to views. The most correlation among the categories is highlighted below, and belongs to any of the category’s relationship with “Entertainment.”

	Views	Dislikes	Likes	Education	HowTo & Style	News & Politics	Entertainment	Comedy	People & Blogs	Sports	Music	Film & Animation
Views	1.000											
Dislikes	0.511	1.000										
Likes	0.849	0.476	1.000									
Education	0.090	0.035	0.079	1.000								
HowTo & S	0.098	0.025	0.079	-0.076	1.000							
News & Pi	0.045	0.005	0.031	-0.057	-0.094	1.000						
Entertainr	-0.033	-0.026	-0.019	-0.130	-0.213	-0.161	1.000					
ComedyD	-0.038	-0.023	-0.048	-0.069	-0.113	-0.085	-0.192	1.000				
People &	-0.050	-0.012	-0.047	-0.066	-0.108	-0.082	-0.185	-0.098	1.000			
SportsDur	-0.036	0.004	-0.030	-0.040	-0.066	-0.050	-0.112	-0.059	-0.057	1.000		
MusicDur	-0.069	-0.001	-0.056	-0.098	-0.162	-0.122	-0.276	-0.146	-0.140	-0.085	1.000	
Film & An	-0.043	-0.020	-0.035	-0.056	-0.091	-0.069	-0.156	-0.083	-0.079	-0.048	-0.118	1.000

Figure 13 Correlation Matrix

The correlation matrix shows that the categories are generally uncorrelated, so a first order model can be created.

First Order Model

The model for regression which can be created from the new dataset is:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 c_1 + \beta_4 c_2 + \beta_5 c_3 + \beta_6 c_4 + \beta_7 c_5 + \beta_8 c_6 + \beta_9 c_7 + \beta_{10} c_8 + \beta_{11} c_9 + \varepsilon$$

Once all the individual variables are included, the equation can be interpreted as:

Total Views = β_0 + β_1(likes) + β_2(dislikes) + β_3(Film & Animation) + β_4(Music) + β_5(Sports) + β_6(People & Blogs) + β_7(Comedy) + β_8(Entertainment) + β_9(News & Politics) + β_{10}(HowTo & Style) + β_{11}(Education) + ε

By running a regression analysis, the following tables are reached through Excel.

Regression Statistics	
Multiple R	0.859
R Square	0.739
Adj R Square	0.739
Std Error	3889808.010
Observations	37392.000

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	11.000	15992425188.000	1453856835.000	9608.715	0.000
Residual	37380.000	5655820655.000	1513060.000		
Total	37391.000	21648245843.000			

Figure 14 Analysis of Variance

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*VIF*	*IV Corr R^2*
Intercept	880946.947	80544.126	10.937	0.000	723078.248	1038815.645		
Education	340439.605	124253.649	2.740	0.006	96899.041	583980.168	1.615	0.381
HowTo & Style	223448.273	99786.437	2.239	0.025	27864.117	419032.428	2.426	0.588
News & Politics	76448.745	111386.567	0.686	0.493	-141871.984	294769.473	1.904	0.475
Entertainment	-689979.019	88757.203	-7.774	0.000	-863945.573	-516012.465	3.805	0.737
Comedy	-426972.366	103822.841	-4.113	0.000	-630467.985	-223476.748	2.235	0.553
People & Blogs	-802390.760	105439.227	-7.610	0.000	-1009054.539	-595726.981	2.156	0.536
Sports	-1023605.799	135937.194	-7.530	0.000	-1290046.430	-757165.168	1.480	0.324
Music	-918014.810	93440.733	-9.825	0.000	-1101161.212	-734868.408	3.088	0.676
Film & Animation	-880922.082	113350.888	-7.772	0.000	-1103092.935	-658751.230	1.866	0.464
Dislikes	39.742	0.856	46.439	0.000	38.065	41.420	1.296	0.228
Likes	25.080	0.099	253.923	0.000	24.887	25.274	1.328	0.247

The R^2 value is 0.739, suggesting that about 74% of the data can be explained by the first order linear model created. The overall F- value is very high, as well as the overall Significance F value, or the p-value. Each individual p-values are basically 0, except for the News & Politics category. However, the coefficients are incredibly large, and the Sum of Squares, as well as the

Sum of Squared Errors, also show that while the Coefficient of Determination is a decent number, the model itself is not very accurate.

Interaction Model

Perhaps a second model is needed to more accurately show the correlation between the categories and total views. By using an interaction model, likes and dislikes are combined into one category, a new category called "Total Opinions." Each category is multiplied by the total number of opinions to represent the interaction. Now the model will carry the following equation:

x_1x_2 = Total Opinions c_i = Category => 1 if True; 0 if False

$E(y) = \beta_0 + \beta_1x_1x_2 + \beta_2c_1x_1x_2 + \beta_3c_2x_1x_2 + \beta_4c_3x_1x_2 + \beta_5c_4x_1x_2 + \beta_6c_5x_1x_2 + \beta_7c_6x_1x_2 + \beta_8c_7x_1x_2 + \beta_9c_8x_1x_2 + \beta_{10}c_9x_1x_2 + \varepsilon$

When each of the variables are included, the final model is as follows:

Total Views = β_0 + β_1(Total Opinions) + β_2(Film & Animation)(Total Opinions) + β_3(Music)(Total Opinions) + β_4(Sports)(Total Opinions) + β_5(People & Blogs)(Total Opinions) + β_6(Comedy)(Total Opinions) + β_7(Entertainment)(Total Opinions) + β_8(News & Politics)(Total Opinions) + β_9(HowTo & Style)(Total Opinions) + β_{10}(Education)(Total Opinions) + ε

These purpose of looking into these interactions are in hopes that they will capture a better model of the dataset. After running a regression analysis on the new model, the Excel Data Analysis Pack shows that the results are as follows:

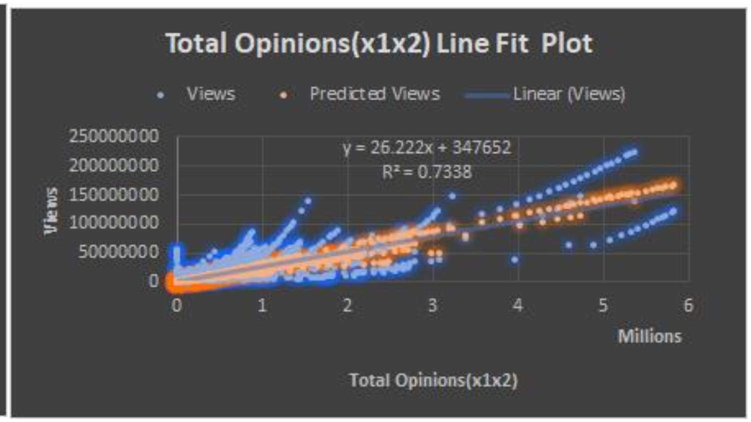
Regression Statistics	
Multiple R	0.867783
R Square	0.753047
Adj R Square	0.752981
Standard Error	3781752
Observations	37392

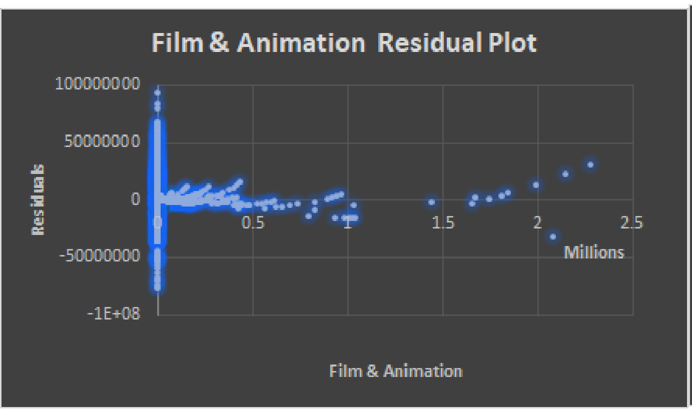
ANOVA					
	df	SS	MS	F	Significance F
Regression	10	1.63021E+18	1.63E+17	11398.79	0
Residual	37381	5.3461E+17	1.43E+13		
Total	37391	2.16482E+18			

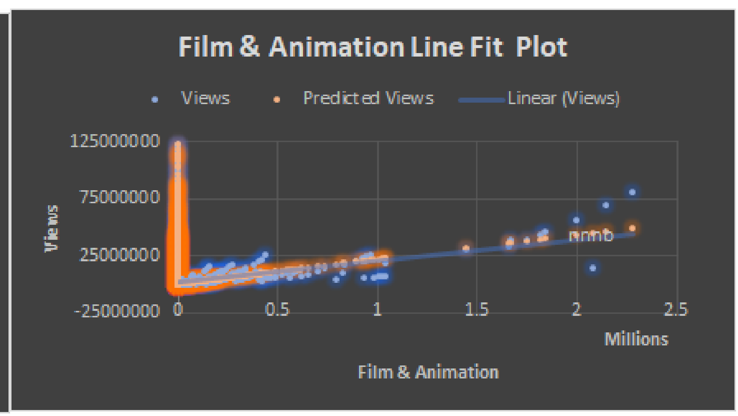
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	462582.235	20804.37928	22.23485	8E-109	421805.0806	503359.39
Total Opinions(x1x2)	26.1403661	0.224817448	116.2737	0	25.69971774	26.581014
Film & Animation	-4.6725272	0.53175923	-8.78692	1.6E-18	-5.71478987	-3.630264
Music	-1.9714026	0.314323883	-6.27188	3.61E-10	-2.58748604	-1.355319
Sports	-9.9572715	0.622287924	-16.0011	1.95E-57	-11.1769729	-8.73757
People & Blogs	-12.381585	0.491002706	-25.2169	3.8E-139	-13.3439636	-11.41921
Comedy	-3.3430083	0.614101883	-5.44374	5.25E-08	-4.54666489	-2.139352
Entertainment	-5.2332576	0.298178264	-17.5508	1.11E-68	-5.8176952	-4.64882
News & Politics	5.12809448	0.352317174	14.55533	7.31E-48	4.437543146	5.8186458
HowTo & Style	2.74550484	0.267598433	10.25979	1.15E-24	2.221004568	3.2700051
Education	2.51900195	0.277110735	9.090236	1.04E-19	1.975857301	3.0621466

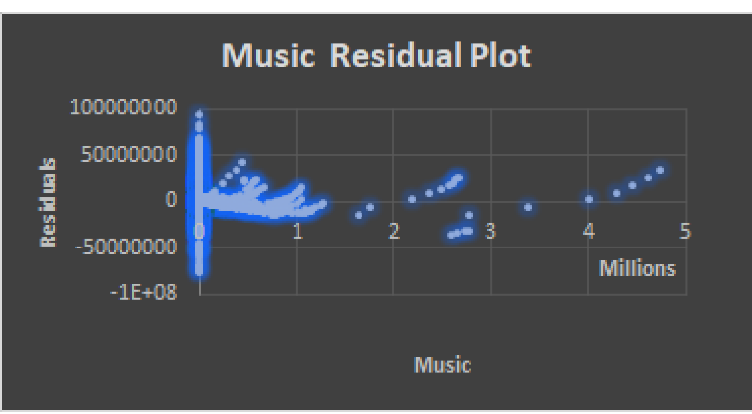
Immediately, this model appears to show better results. The Correlation of Determination(R^2) is a little higher than the first model at 0.753, the F-Value is higher, and the overall p-value is below 0.05. The coefficients look more palatable, although the negative correlation still appears to show an error in the model. It should be the case that each video posted no matter what category it is should have a positive number of views, especially since these are all the most trending videos. The sum of squares (SS) and Sum of squared errors (MSE) are once again very high, however the model is still better than the regular first-order model. A closer look at the residual and line plot graphs for each category will paint a clearer picture of what is happening.

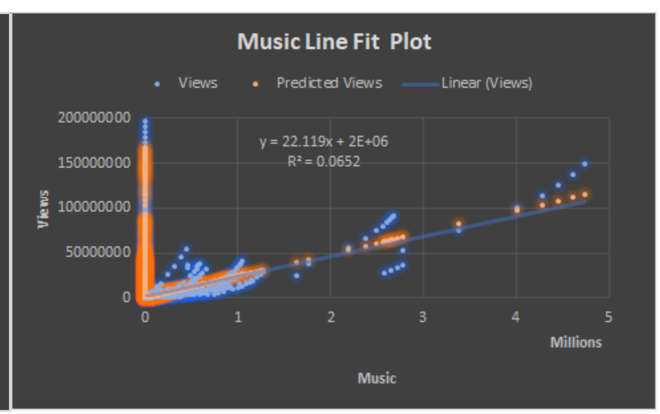
Residual Plots

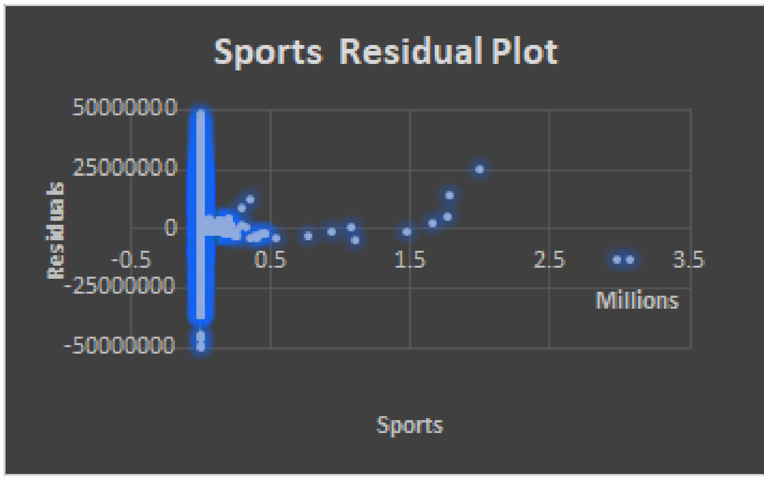


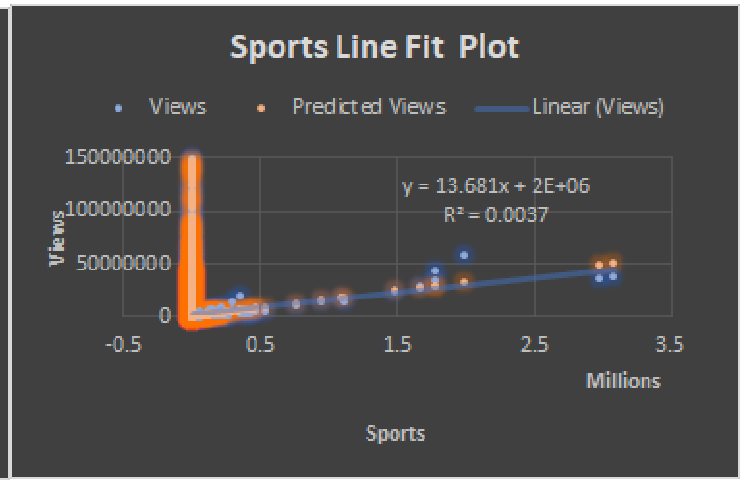


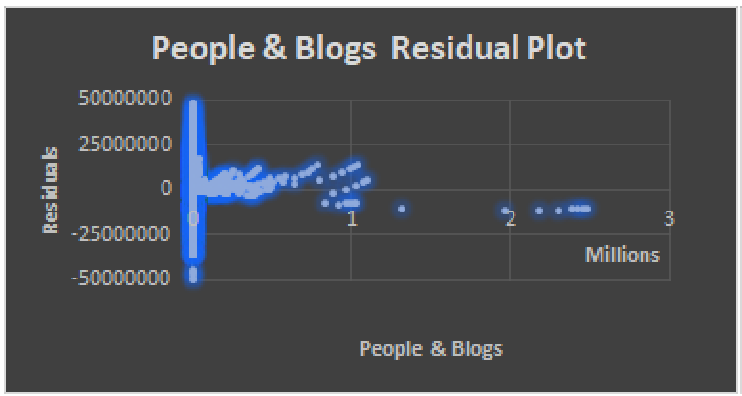


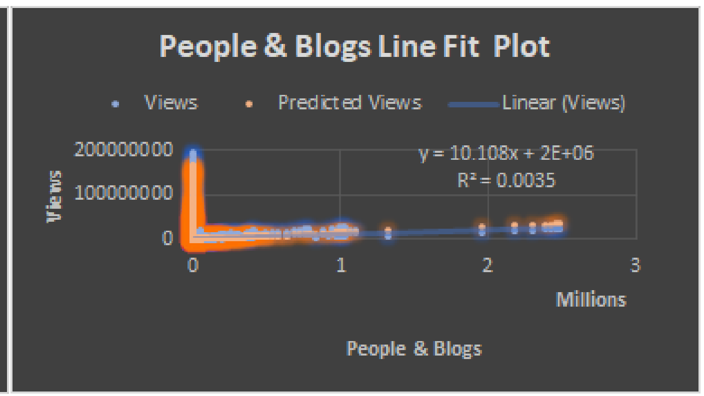


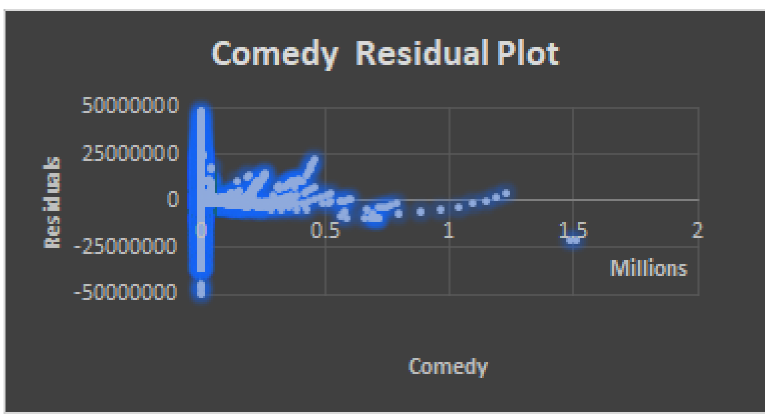


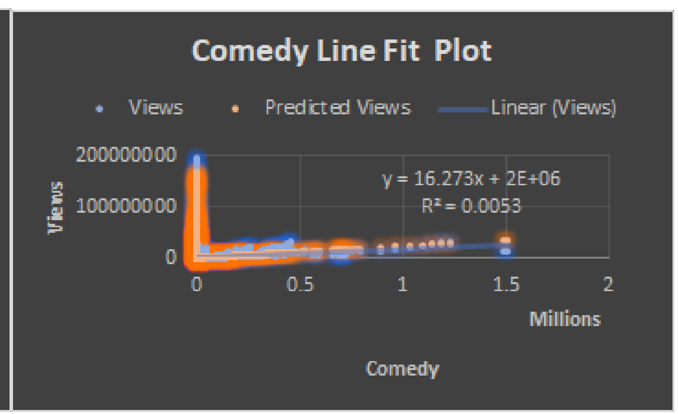

Sports Residual Plot
50000000
25000000
0
-25000000
-50000000
-0.5
0.5
1.5
2.5
3.5
Millions
Residuals
Sports

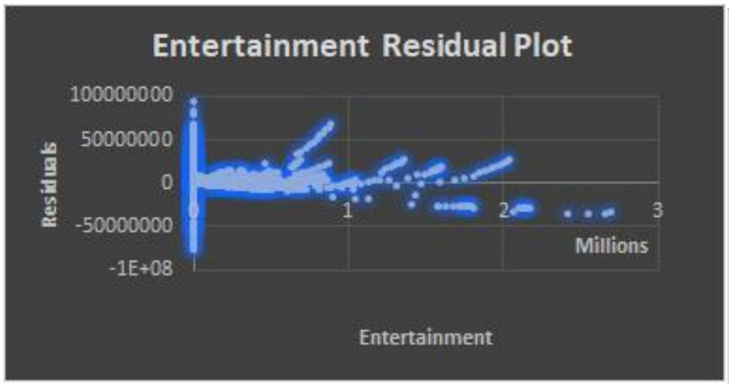

Sports Line Fit Plot
Views
Predicted Views
Linear (Views)
150000000
100000000
50000000
0
-0.5
0.5
1.5
2.5
3.5
y = 13.681x + 2E+06
R² = 0.0037
Millions
Views
Sports

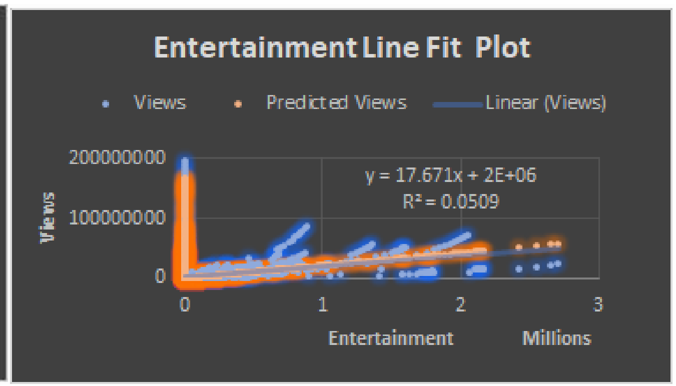

People & Blogs Residual Plot
50000000
25000000
0
-25000000
-50000000
0
1
2
3
Millions
Residuals
People & Blogs


People & Blogs Line Fit Plot
Views
Predicted Views
Linear (Views)
200000000
100000000
0
0
1
2
3
y = 10.108x + 2E+06
R² = 0.0035
Millions
Views
People & Blogs

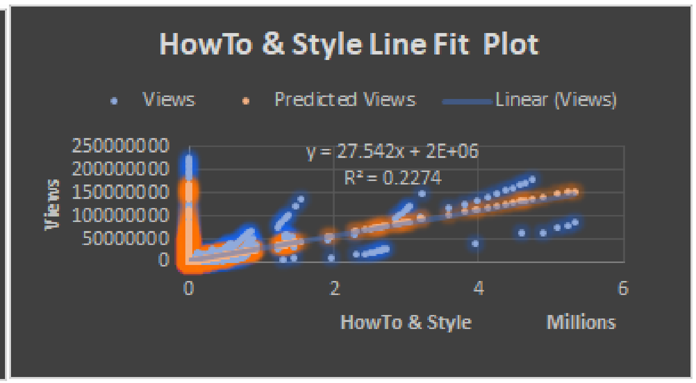

Comedy Residual Plot
50000000
25000000
0
-25000000
-50000000
0
0.5
1
1.5
2
Millions
Residuals
Comedy

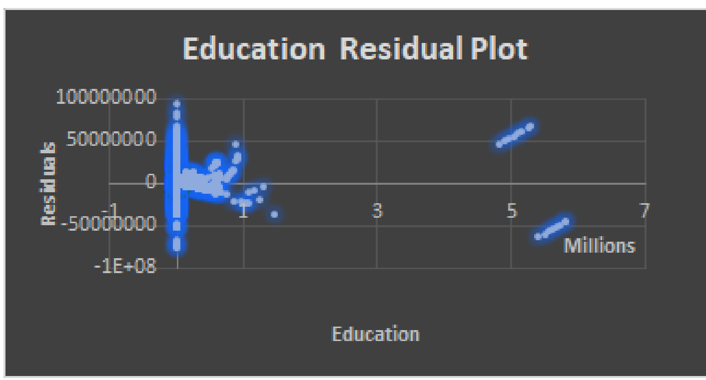

Comedy Line Fit Plot
Views
Predicted Views
Linear (Views)
200000000
100000000
0
0
0.5
1
1.5
2
y = 16.273x + 2E+06
R² = 0.0053
Millions
Views
Comedy

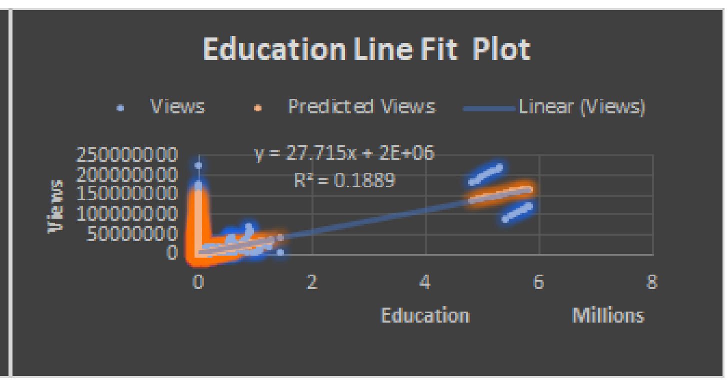

Entertainment Residual Plot
100000000
50000000
0
-50000000
-1E+08
Residuals
1
2
3
Millions
Entertainment

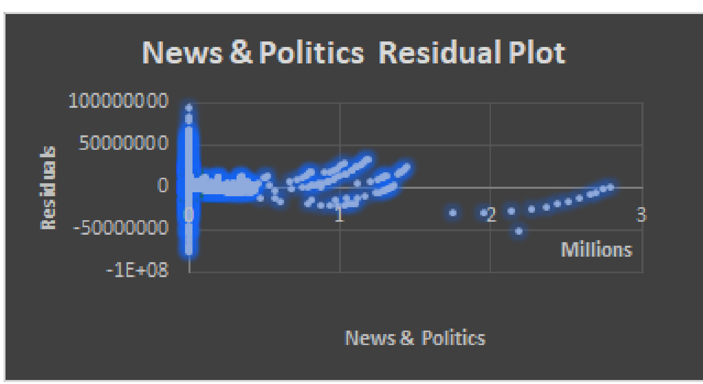

Entertainment Line Fit Plot
Views
Predicted Views
Linear (Views)
200000000
100000000
0
Views
y = 17.671x + 2E+06
R² = 0.0509
0
1
2
3
Entertainment
Millions

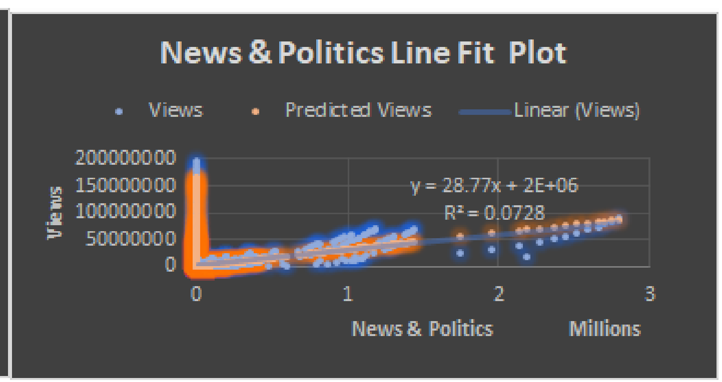

HowTo & Style Residual Plot
150000000
100000000
50000000
0
-50000000
-1E+08
-1.5E+08
Residuals
2
4
6
Millions
HowTo & Style


HowTo & Style Line Fit Plot
Views
Predicted Views
Linear (Views)
250000000
200000000
150000000
100000000
50000000
0
Views
y = 27.542x + 2E+06
R² = 0.2274
0
2
4
6
HowTo & Style
Millions


Education Residual Plot
100000000
50000000
0
-50000000
-1E+08
Residuals
-1
1
3
5
7
Millions
Education


Education Line Fit Plot
Views
Predicted Views
Linear (Views)
250000000
200000000
150000000
100000000
50000000
0
Views
y = 27.715x + 2E+06
R² = 0.1889
0
2
4
6
8
Education
Millions


News & Politics Residual Plot
100000000
50000000
0
-50000000
-1E+08
Residuals
0
1
2
3
Millions
News & Politics


News & Politics Line Fit Plot
Views
Predicted Views
Linear (Views)
200000000
150000000
100000000
50000000
0
Views
y = 28.77x + 2E+06
R² = 0.0728
0
1
2
3
News & Politics
Millions

The previous graphs of residual plots portray unequal variances for different settings of the independent variables and are considered to be heteroscedastic. The overall residual plot of the total number of opinions could be considered to represent a multiplicative model with a general “cone” shape of residual variability; the number of views increases as the estimated number of opinions per video increases. The presence of heteroscedasticity means that the ordinary least squares line estimators may not be the best linear unbiased estimators and their variance is not the lowest of all other unbiased estimators.

For many of the graphs, it can be seen that as the x variable increases, the variance often increases as well. One reason for this may be that the most trending videos have much more likes and dislikes when compared to the other trending videos. The graphs also take into account the unused values for each variables, which is likely what is making the coefficients look strange. A slight change to the model may show more accurate results by removing the total opinions variable, but keeping its interaction with all the qualitative variables.

x_1x_2 = Total Opinions c_i = Category => 1 if True; 0 if False

$E(y) = \beta_0 + \beta_1c_1x_1x_2 + \beta_2c_2x_1x_2 + \beta_3c_3x_1x_2 + \beta_4c_4x_1x_2 + \beta_5c_5x_1x_2 + \beta_6c_6x_1x_2 + \beta_7c_7x_1x_2 + \beta_8c_8x_1x_2 + \beta_9c_9x_1x_2 + \varepsilon$

Total Views = β_0 + β_1(Film & Animation)(Total Opinions) + β_2(Music)(Total Opinions) + β_3(Sports)(Total Opinions) + β_4(People & Blogs)(Total Opinions) + β_5(Comedy)(Total Opinions) + β_6(Entertainment)(Total Opinions) + β_7(News & Politics)(Total Opinions) + β_8(HowTo & Style)(Total Opinions) + β_9(Education)(Total Opinions) + ε

This model brings the regression analysis results shown in the following tables:

Regression Statistics	
Multiple R	0.814697258
R Square	0.663731622
Adjusted R Square	0.663650663
Standard Error	4412890.016
Observations	37392

ANOVA					
	df	SS	MS	F	Significance F
Regression	9	1.43686E+18	1.59651E+17	8198.351	0
Residual	37382	7.27962E+17	1.94736E+13		
Total	37391	2.16482E+18			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	769643.9341	24080.05473	31.96188476	3.6E-221	722446.3659	816841.5
Film & Animation	20.92361714	0.564837227	37.04362275	5.1E-295	19.81652067	22.030714
Music	23.81847632	0.259886401	91.64956779	0	23.30909184	24.327861
Sports	15.79365867	0.678598882	23.27392382	5.7E-119	14.46358624	17.123731
People & Blogs	13.18539187	0.512281914	25.73854652	7.9E-145	12.18130526	14.189478
Comedy	21.75832433	0.670851811	32.4338758	1.3E-227	20.44343636	23.073212
Entertainment	20.30623396	0.235308011	86.2963988	0	19.8450238	20.767444
News & Politics	30.83891231	0.320059879	96.35357119	0	30.21158616	31.466238
HowTo & Style	28.62500403	0.173352121	165.1263556	0	28.28522911	28.964779
Education	28.49667062	0.191294159	148.967803	0	28.12172882	28.871612

The R^2 value has decreased to 0.664, meaning 66.4% of this model is representative of the dataset; the F-value has decreased as well to 8198, and the overall p-value remains lower than 0.05. This model seems to be very similar to the past models, except the coefficients of all the variables looks closer to what is expected. A look at the coefficient correlation this time shows different results than before:

	Film & Animation	Music	Sports	People & Blogs	Comedy	ıtertainme	News & Politics	HowTo & Style	Education
Film & An	1								
Music	-0.007294296	1							
Sports	-0.003076764	-0.00433	1						
People &	-0.006033574	-0.00849	-0.00358	1					
Comedy	-0.008406308	-0.01183	-0.00499	-0.009784225	1				
Entertainr	-0.014215881	-0.02	-0.00844	-0.016546073	-0.02305	1			
News & Pı	-0.007258218	-0.01021	-0.00431	-0.008447947	-0.01177	-0.0199	1		
HowTo & !	-0.008160978	-0.01148	-0.00484	-0.009498681	-0.01323	-0.02238	-0.011426645	1	
Education	-0.004570411	-0.00643	-0.00271	-0.005319568	-0.00741	-0.01253	-0.00639929	-0.007195218	1

Now it appears that none of the categories are correlated to each other, not even the entertainment category. This shows that for this model, the result of one variable has nothing to do with the other categorical variables. Each variable has a p-value below 0.05 and high t values suggest that the coefficients are good predictors. So inputting all the coefficients into the model:

Total Views = 769643.934 + 20.924(Film & Animation)(Total Opinions) + 23.818(Music)(Total Opinions) + 15.794(Sports)(Total Opinions) + 13.185(People & Blogs)(Total Opinions) + 21.758(Comedy)(Total Opinions) + 20.306(Entertainment)(Total Opinions) + 30.839(News & Politics)(Total Opinions) + 28.625(HowTo & Style)(Total Opinions) + 28.497(Education)(Total Opinions) + ε

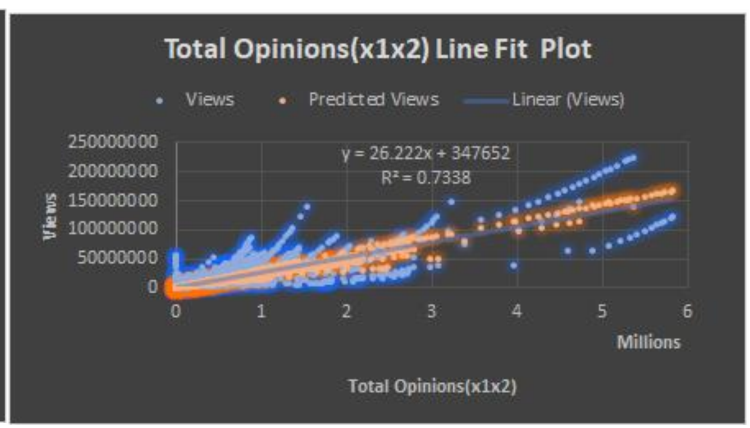
According to the equation, although entertainment has the most overall views, likes, and dislikes, it is closer to the median when it comes to views per total opinions and the category with 20.306 views per opinion in the category. Other categories such as News & Politics (30.839), HowTo & Style (28.625), and Education (28.497) have more views per opinion, suggesting that when videos in those categories become trending, people are more opinionated and more interested. So perhaps even though the amount of videos are less for those categories, the ones that become trending are very popular, and that may be another direction for someone to consider when creating YouTube videos.

Conclusion

Looking at the three models, it is hard to make any concrete conclusions that relate the category to the number of views with the data given. However, it should be noted that Entertainment is the largest category in the United States, and the negative correlation with the other categories might mean that since there are so many Entertainment videos trending, it stymies the chances for other videos to trending to some degree. Another point to take away is that the News & Politics, How To & Style, and Education videos always have the highest coefficients, which may mean that videos in those categories might be fewer but more impactful per video.

Possible Improvements

In future experiments, it would be nice to have data over a longer period of time so monitor trends and test predictive hypotheses. Another model that could be considered as well is a piecewise model that could possibly be of the second order. Looking at the residual and line fit plots of the total number of opinions, it appears to have several different trendlines.



So perhaps a more accurate model could be created by splitting up the dataset into section such as 0 to 1.5 million, 1.5 to 3.25 million, 3.25 to 5.25 million, and 5.25 million to 6 million. Breaking it up into those sections, could possibly bring better results, as the data would also be less heteroscedastic. In conclusion, people should probably make videos that start in the Entertainment category, and once more comfortable, venture out to different categories.

The Influence of Comments on the Number of Likes

Baha Gharbi

In this analysis, I am trying to find the influence of the comment section on the number of likes in 10 different countries. First, I started by cleaning the original data and focusing only on the variables needed for my research. I created ten cleaner and more focused datasets extracted from the original dataset.

Variables

I only focused on 3 variables:

- The response variable: the number of likes in each video
- The explanatory variables:
    - One quantitative variable: the number of comments in each video
    - One qualitative variable: whether the comment section was disabled or not: 1 if True and 0 if False.

First-Order Model

Then, I fitted the data into a first order multiple regression model where:

The number of likes (y)=Beta0+Beta1*(the number of comments)+Beta2*(Comment Disabled)+error

Using SAS, I implemented the data into the model using the following code:

Code Used

```
DATA Comments;
INFILE "C:\Users\mgharbi\Desktop\Project.txt" firstobs=2 dlm='09'x;
INPUT Likes CD CC;
PROC PRINT DATA=Comments;
RUN;
PROC REG DATA=Comments;
MODEL Likes = CD CC;
RUN;
```

United States

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.383513E15	6.917565E14	37184.6	<.0001
Error	40943	7.616754E14	18603311266		
Corrected Total	40945	2.145188E15			

Root MSE	136394	R-Square	0.6449
Dependent Mean	74272	Adj R-Sq	0.6449
Coeff Var	183.64109		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	32977	696.67984	47.33	<.0001
CD	1	-11532	5465.75522	-2.11	0.0349
CC	1	4.90956	0.01801	272.53	<.0001

According to the data provided by SAS, the model that can be used for the USA model:

- The number of observations: 40945

- The number of likes=32977+4.9095*the number of comments-11532*Comments disabled+error
- The adjusted R-square is equal to 0.6449 which means that %64.5 of the variation in the number of likes is explained by the model and that is acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

    - Null hypothesis: Beta = 0

    - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv=0.0349<0.05.

→ there is a negative relationship between ability to comment and the number of likes.

India

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.148772E14	1.074386E14	29160.0	<.0001
Error	37349	1.376105E14	3684448845		
Corrected Total	37351	3.524877E14			

Root MSE	60700	R-Square	0.6096
Dependent Mean	27083	Adj R-Sq	0.6096
Coeff Var	224.12692		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13797	324.56858	42.51	<.0001
CC	1	5.09535	0.02114	241.08	<.0001
CD	1	-10999	1779.19033	-6.18	<.0001

According to the data provided by SAS, the model that can be used for the INDIA model:

- The number of observations: 37351
- The number of likes=13797+5.0935*the number of comments-10999*Comments disabled+error
- The adjusted R-square is equal to 0.6096 which means that %60.96 of the variation in the number of likes is explained by the model and that is acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

    - Null hypothesis: Beta = 0

    - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

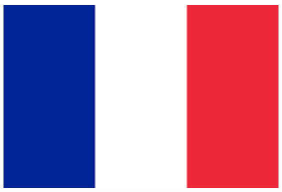
→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a negative relationship between ability to comment and the number of likes.

France

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.235836E14	1.117918E14	52869.7	<.0001
Error	40721	8.610372E13	2114479527		
Corrected Total	40723	3.096873E14			

Root MSE	45983	R-Square	0.7220
Dependent Mean	17389	Adj R-Sq	0.7220
Coeff Var	264.44206		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7827.39459	232.39038	33.68	<.0001
CC	1	5.27514	0.01623	325.04	<.0001
CD	1	-4808.58131	1559.64658	-3.08	0.0020

According to the data provided by SAS, the model that can be used for the FRANCE model:

- The number of observations: 40723
- The number of likes=7827.39+5.275*the number of comments-4808*Comments disabled+error
- The adjusted R-square is equal to 0.7220which means that %72.2 of the variation in the number of likes is explained by the model and that is very acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero

- Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv=0.002<0.05.

→ there is a negative relationship between ability to comment and the number of likes.

Korea

:

Number of Observations Read	34561
Number of Observations Used	34561

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.032406E14	2.016203E14	99024.4	<.0001
Error	34558	7.036236E13	2036065836		
Corrected Total	34560	4.73603E14			

Root MSE	45123	R-Square	0.8514
Dependent Mean	12188	Adj R-Sq	0.8514
Coeff Var	370.21476		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2016.57130	245.63138	8.21	<.0001
CC	1	5.02206	0.01129	445.00	<.0001
CD	1	-87.59674	2011.17081	-0.04	0.9653

According to the data provided by SAS, the model that can be used for the KOREA model:

- The number of observations: 34560
- The number of likes=2016.57+5.022*the number of comments-87.59*Comments disabled+error
- The adjusted R-square is equal to 0.8514which means that %85.14 of the variation in the number of likes is explained by the model and that is very good.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

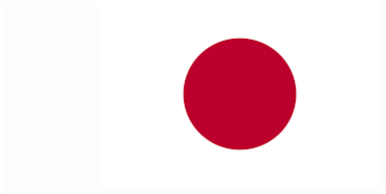
→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we fail to reject the null hypothesis since the Pv=0.9653>0.05.

→ there is no relationship between the ability to comment and the number of likes.

Japan


:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.275691E14	6.378454E13	78315.7	<.0001
Error	20519	1.671177E13	814453621		
Corrected Total	20521	1.442808E14			

Root MSE	28539	R-Square	0.8842
Dependent Mean	8059.97832	Adj R-Sq	0.8842
Coeff Var	354.07830		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1753.75216	207.16451	8.47	<.0001
CC	1	5.27601	0.01333	395.67	<.0001
CD	1	-66.69700	786.70820	-0.08	0.9324

According to the data provided by SAS, the model that can be used for the JAPAN model:

- The number of observations: 20521
- The number of likes=1753.75+5.27*the number of comments-66.69*Comments disabled+error
- The adjusted R-square is equal to 0.8842which means that %88.42 of the variation in the number of likes is explained by the model and that is very good.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we fail to reject the null hypothesis since the Pv=0.9324>0.05.

→ there is no relationship between the ability to comment and the number of likes.

Russia

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.063017E14	5.315086E13	51267.6	<.0001
Error	40736	4.223241E13	1036734253		
Corrected Total	40738	1.485341E14			

Root MSE	32198	R-Square	0.7157
Dependent Mean	12435	Adj R-Sq	0.7157
Coeff Var	258.92870		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4476.91387	163.69145	27.35	<.0001
CC	1	4.52920	0.01415	320.01	<.0001
CD	1	-3148.46001	1001.50096	-3.14	0.0017

According to the data provided by SAS, the model that can be used for the RUSSIA model:

- The number of observations: 40738
- The number of likes=4476.91+4.529*the number of comments-3148.46*Comments disabled+error
- The adjusted R-square is equal to 0.7157which means that %71.57 of the variation in the number of likes is explained by the model and that is very acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we reject the null hypothesis since the Pv=0.0017<0.05.

→ there is a negative relationship between the ability to comment and the number of likes.

Mexico

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.902721E14	9.513606E13	50812.0	<.0001
Error	40435	7.570705E13	1872314802		
Corrected Total	40437	2.659792E14			

Root MSE	43270	R-Square	0.7154
Dependent Mean	15864	Adj R-Sq	0.7154
Coeff Var	272.75301		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5849.27896	218.68668	26.75	<.0001
CC	1	4.92046	0.01544	318.73	<.0001
CD	1	-2052.98805	2074.38781	-0.99	0.3223

According to the data provided by SAS, the model that can be used for the MEXICO model:

- The number of observations: 40437
- The number of likes=5849.278+4.92*the number of comments-2052.98*Comments disabled+error
- The adjusted R-square is equal to 0.7154which means that %71.54 of the variation in the number of likes is explained by the model and that is very acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
  - Null hypothesis: all parameters are equal to zero
  - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we fail to reject the null hypothesis since the Pv=0.3223>0.05.

→ there is no relationship between the ability to comment and the number of likes.

Denmark

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.075879E14	1.537939E14	54306.1	<.0001
Error	40831	1.156328E14	2831984308		
Corrected Total	40833	4.232206E14			

Root MSE	53216	R-Square	0.7268
Dependent Mean	21876	Adj R-Sq	0.7268
Coeff Var	243.26240		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8128.73139	270.24878	30.08	<.0001
CC	1	4.97007	0.01509	329.39	<.0001
CD	1	-3882.27928	1669.03057	-2.33	0.0200

According to the data provided by SAS, the model that can be used for the DENMARK model:

- The number of observations: 40831
- The number of likes=8128.73+4.97*the number of comments-3882.27*Comments disabled+error
- The adjusted R-square is equal to 0.7268which means that %72.68 of the variation in the number of likes is explained by the model and that is very acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we reject the null hypothesis since the Pv=0.02<0.05.

→ there is a negative relationship between the ability to comment and the number of likes.

Great Britain

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.642376E15	1.321188E15	24200.2	<.0001
Error	38913	2.12442E15	54594106295		
Corrected Total	38915	4.766796E15			

Root MSE	233654	R-Square	0.5543
Dependent Mean	134520	Adj R-Sq	0.5543
Coeff Var	173.69506		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68022	1234.92348	55.08	<.0001
CC	1	5.13792	0.02339	219.66	<.0001
CD	1	-42690	9025.39981	-4.73	<.0001

According to the data provided by SAS, the model that can be used for the GREAT BRITAIN model:

- The number of observations: 38913
- The number of likes=68022+5.13*the number of comments-42690*Comments disabled+error
- The adjusted R-square is equal to 0.5543which means that %55.43 of the variation in the number of likes is explained by the model and that is acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
    - Null hypothesis: all parameters are equal to zero
    - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

- Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we reject the null hypothesis since the Pv=0.02<0.05.

→ there is a negative relationship between the ability to comment and the number of likes.

Canada

:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.037374E14	2.518687E14	47662.4	<.0001
Error	40878	2.160168E14	5284426863		
Corrected Total	40880	7.197542E14			

Root MSE	72694	R-Square	0.6999
Dependent Mean	39583	Adj R-Sq	0.6999
Coeff Var	183.65115		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13633	372.02836	36.65	<.0001
CC	1	5.14424	0.01667	308.63	<.0001
CD	1	514.30399	3033.57909	0.17	0.8654

According to the data provided by SAS, the model that can be used for the CANADA model:

- The number of observations: 40880
- The number of likes=13633+5.14*the number of comments+514.3*Comments disabled+error
- The adjusted R-square is equal to 0.6999which means that %69.99 of the variation in the number of likes is explained by the model and that is acceptable.

After interpreting the model, I ran two hypothesis tests for both, the overall model and for each parameter.

- F-test:
  - Null hypothesis: all parameters are equal to zero
  - Alternative Hypothesis: at least one parameter is different than zero

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05

→ The model is very adequate

- T-test:

  - Null hypothesis: Beta = 0

  - Alternative Hypothesis: Beta =/= 0

For Beta1:

→ Using Alpha=0.05, we can reject the null hypothesis since the Pv<0.0001<0.05.

→ there is a positive relationship between the number of comments and the number of likes.

For Beta2:

→ Using Alpha=0.05, we fail to reject the null hypothesis since the Pv=0.8654>0.05.

→ there is no relationship between the ability to comment and the number of likes.

Conclusion:

All 10 overall models for all the 10 countries are adequate.

- That in every data of every country, there is a positive relationship between the number of comments and the number of likes.

- In the Data of Canada, Mexico, Japan, and Korea, we fail to reject H0→ the ability to comment does not contribute to the number of likes.

- In the Data of USA, India, France, Denmark, Russia, and GB:

→ we reject H0: B1

→ there is a negative relationship between the ability to comment and the number of likes, which means that the user is less likely to like a video if he is not able to comment.

Measures of Controversy and View Counts

Evelina Ramoskaite

In this analysis, I am examining the utility of using controversy as a predictor of video popularity. I have noticed that videos detailing dramatic events and quarrels between influencers tend to make it to the trending page frequently and wanted to explore this observation in an empirical way.

Three different applications of a first-order model are examined, including a domestic application of videos in the United States, an international analysis, and a genre-specific application of a controversy-focused model.

Variables

The dependent variable is the total view count of each video. The models utilize two dummy variables, which indicate whether comments and ratings are disabled. The remaining variables are quantitative in nature.

The RatingsDisabled and CommentsDisabled betas were chosen because when influencers choose to disable feedback, it is usually because they are avoiding negative backlash. Someone who knows they are uploading a universally pleasing video is much less likely to feel the need to censor feedback. It is common for content creators to disable comments after uploading their videos and discovering that the feedback is too negative. A 1 signifies that the ratings or comments are disabled, and a 0 signifies that feedback is enabled.

PercentDislike represents the percentage of the like/dislike rating that is comprised of dislikes. A controversial video would have more dislikes, or a mix or likes and dislikes.

The DisliketoComment ratio is the number of dislikes divided by the number of comments. A higher ratio would be indicative of more negative engagements with the video.

The Comment_to_views ratio is the number of comments divided by the number of views. A higher number signifies that a larger portion of viewers are discussing the video.

Limitations

Because the data set available focuses only on top trending videos, the analysis does not offer a complete look at the relationships amongst the average video. I am observing those few popular outliers and determining if markers of controversy have any utility in predicting view counts amongst these videos.

There is some data that would have been insightful but impractical to gather or outside of the intended application of my model. For instance, subscriber counts would have explained a sizeable amount of the variation but were not available in the data set. They could be obtained but would not be relevant because the trending statistics were time-specific.

Domestic Analysis: USA

The United States is the first data set that I applied my model to. Most of the popular content on YouTube comes from US-based creators, and the platform is very popular in its home country.

Hypothesis

Null hypothesis: all parameters are equal to zero

Alternative Hypothesis: at least one parameter is different than zero

Regression

Model: MODEL1
Dependent Variable: views

Number of Observations Read	40946
Number of Observations Used	40946

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.578369E17	2.859456E17	8477.98	<.0001
Error	40942	1.380893E18	3.372804E13		
Corrected Total	40945	2.23873E18			

Root MSE	5807584	R-Square	0.3832
Dependent Mean	2360956	Adj R-Sq	0.3831
Coeff Var	245.98444		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1482652	34891	42.49	<.0001
CommentsDisabled	1	1404687	233675	6.01	<.0001
PercentDislike	1	-2641647	282186	-9.36	<.0001
comment_count	1	122.22464	0.76714	159.32	<.0001

Equation:

Views =1482652 + 1404687(Comments Disabled) -2641647(Percent Dislike) +

122.22(Comment Count)

There is a 1,404,687 increase in views associated with comments being disables. For each percentage increase in people disliking a video, the view count decreases by 2,641,647. For each additional comment on the video, the view count is estimated to increase by 122.

Model Adequacy

F-Test:

The F-value of the model is very high at 8477, suggesting that the model is significantly different from the alternative hypothesis.

T-Test:

The P values are all below 0.001. At a 95% significance level, this suggests that all of these values are relevant in the United States model, since $P < 0.05$.

Multicollinearity:

All of the parameters used have a VIF below 2.5, suggesting that multicollinearity is not a prudent issue in this model. The highest VIF values are for the CommentsDisabled and RatingsDisabled dummy variables, which would be expected to have a relationship.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	1472908	35402	41.61	<.0001	.	0
comment_count	1	122.21700	0.77138	158.44	<.0001	0.99955	1.00045
CommentsDisabled	1	4297639	865539	4.97	<.0001	0.40750	2.45399

/38/sasexec/submissions/d4c545fd-8f71-4572-8087-0ab4980b17c1/results

Results: Program 1

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
RatingsDisabled	1	-94420	705346	-0.13	0.8935	0.40723	2.45560
PercentDislike	1	-2493378	292228	-8.53	<.0001	0.99786	1.00214
DislikeCommentRatio	1	932.33903	901.63498	1.03	0.3011	0.99989	1.00011

The correlation matrix does not show any values that would lead me to question the adequacy of the model, given that some variables are expected to have a relationship.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	comment_count	CommentsDisabled	RatingsDisabled	PercentDislike	DislikeCommentRatio
comment_count	1.00000 40946	-0.02828 <.0001 40946	-0.01382 0.0052 40946	0.01218 0.0137 40946	-0.00338 0.4969 40384
CommentsDisabled	-0.02828 <.0001 40946	1.00000 40946	0.31923 <.0001 40946	0.08950 <.0001 40946	-0.00150 0.7638 40384
RatingsDisabled	-0.01382 0.0052 40946	0.31923 <.0001 40946	1.00000 40946	-0.04197 <.0001 40946	-0.00184 0.7119 40384
PercentDislike	0.01218 0.0137 40946	0.08950 <.0001 40946	-0.04197 <.0001 40946	1.00000 40946	0.00976 0.0499 40384
DislikeCommentRatio	-0.00338 0.4969 40384	-0.00150 0.7638 40384	-0.00184 0.7119 40384	0.00976 0.0499 40384	1.00000 40384

Conclusion

These tests lead me to conclude that the model is adequate. I reject the null hypothesis for the comments disabled, percent dislike, and comment count betas. The other variables in the initial model were found to be irrelevant.

The model explains about 38.3% of the variation in the number of views with a 95% confidence interval. This model does not provide a high degree of certainty when predicting future view counts but can give users an idea of which videos will become more popular with relatively little information.

Code used

To import Data

```
proc import datafile="/folders/myfolders/data/USAData.csv"
out=USAData dbms=csv replace;
getnames=yes;
run;
```

To run Regression

```
ods noproctitle;
ods graphics / imagemap=on;
proc import datafile="/folders/myfolders/data/USAData.csv"
out=USAData dbms=csv replace;
getnames=yes;
run;
```

Multicollinearity Check

```
proc corr;
var comment_count CommentsDisabled RatingsDisabled PercentDislike dislikeCommentRatio;
run;

proc reg;
model views = comment_count CommentDisabled RatingDisabled DislikePercent
dislikeCommentRatio / vif tol collin;
run;
```

International Analysis: Korea

Different viewers across the world likely respond to videos in a distinct way. To gauge the utility of this modeling method across different cultures, the data in Korea was also analyzed. An Eastern country was chosen to examine if the factors driving video popularity there are unique from the west.

Hypothesis

Null hypothesis: all parameters are equal to zero

Alternative Hypothesis: at least one parameter is different than zero

Regression

Number of Observations Read	34561
Number of Observations Used	34561

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.332313E17	3.330782E16	16214.8	<.0001
Error	34556	7.098376E16	2.054166E12		
Corrected Total	34560	2.042151E17			

Root MSE	1433236	R-Square	0.6524
Dependent Mean	424992	Adj R-Sq	0.6524
Coeff Var	337.23809		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	240334	7972.70622	30.14	<.0001
RatingsDisabled	1	-115344	40101	-2.88	0.0040
DislikeCommentRatio	1	3943.66169	1517.62064	2.60	0.0094
comment_count	1	91.27243	0.35850	254.60	<.0001
CommentsDisabled	1	163157	65492	2.49	0.0127

Equation:

Views = 229,822 + 163,157 (Comments Disabled) – 115,344(Ratings Disabled) +

3,943.66(Dislike Comment Ratio) + 91.27(Comment Count)

Model Adequacy

F-Test:

The F-value of 16214 is high, suggesting that the model is significantly different from the alternative hypothesis.

T-Test:

The P values are all below 0.001. At a 95% significance level, this suggests that all of these values are relevant in the Korean model, since P< 0.05. However, the P values are slightly higher than they are in the USA- specific model. In both models, the percentage of dislikes was found to be statistically insignificant. For the initial Korean model, the PercentDislike variable had a P vlalue of 0.07, which lead me to exclude it from the final regression.

Multicollinearity:

All of the VIF values are relatively low, which leads me to believe that the amount of multicollinearity is not a cause of concern. In this aspect, the model is adequate.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	229822	9937.32028	23.13	<.0001	.	0
comment_count	1	91.28628	0.35857	254.58	<.0001	0.99922	1.00078
CommentsDisabled	1	157044	65581	2.39	0.0166	0.94863	1.05415
RatingsDisabled	1	-103957	40612	-2.56	0.0105	0.92712	1.07861
PercentDislike	1	123540	69716	1.77	0.0764	0.96112	1.04045
DislikeCommentRatio	1	3632.63638	1527.68981	2.38	0.0174	0.98628	1.01391

The RatingsDisabled and CommentsDisabled variables were also highly correlated with each other, with the exact same correlation coefficient of 0.22. There is also some multicollinearity between the PercentDislike and CommentsDisabled variables. While the PercentDislike variable was not included in the model, the correlation suggests that people are more likely to feel negative sentiments towards videos that content creators felt a need to censor in this way.

Pearson Correlation Coefficients, N = 34561 Prob > |r| under H0: Rho=0					
	comment_count	CommentsDisabled	RatingsDisabled	PercentDislike	DislikeCommentRatio
comment_count	1.00000	-0.01154 0.0320	-0.01430 0.0079	-0.02020 0.0002	-0.00465 0.3869
CommentsDisabled	-0.01154 0.0320	1.00000	0.22046 <.0001	0.01627 0.0025	-0.01242 0.0210
RatingsDisabled	-0.01430 0.0079	0.22046 <.0001	1.00000	-0.15148 <.0001	-0.02083 0.0001
PercentDislike	-0.02020 0.0002	0.01627 0.0025	-0.15148 <.0001	1.00000	0.11621 <.0001
DislikeCommentRatio	-0.00465 0.3869	-0.01242 0.0210	-0.02083 0.0001	0.11621 <.0001	1.00000

Conclusion

There is enough evidence to reject the null hypothesis. About 65.24% of the variation in the number of views can be explained with the model. The relevant variables include whether ratings and comments are disabled, the percentage of dislikes, and the dislike to comment ratio. This was the only instance where the dislike percentage was statistically relevant. Multicollinearity was not an issue in this model. In conclusion, the model is adequate and there is sufficient proof to reject the null hypotheisis.

Code Used

Importing Data

```
proc import datafile="/folders/myfolders/data/KoreaData.csv"
out=KoreaData dbms=csv replace;
getnames=yes;
run;
```

Running Regression

```
ods noproctitle;
ods graphics / imagemap=on;

proc reg data=WORK.KORADATA alpha=0.05 plots(only)=(diagnostics residuals
		observedbypredicted);
	model views=DislikeCommentRatio PercentDislike comment_count CommentsDisabled
		RatingsDisabled /;
	run;
```

Multicollinearity Check

```
proc corr;
var comment_count CommentsDisabled RatingsDisabled PercentDislike DislikeCommentRatio;
run;
proc reg;
model views = comment_count CommentsDisabled RatingsDisabled PercentDislike
DislikeCommentRatio / vif tol collin;
run;
```

Genre-Specific Analysis : The USA Beauty Community

The beauty community is one of the most prevalent groups in the YouTube ecosystem. It is also one of the most lucrative.

The most popular content creators are based in the United States, although some are in Europe. It is well known for its intrapersonal drama between influencers, and some quarrels have become relevant enough to not only trend on the platform but make international news headlines. Most notably an argument between James Charles and Tatti Westbrook, has been discussed in outlets like the Business Insider, ABC, and Vox.

Jeffree Star, an influencer with a $ 75 million cosmetics line, regularly appears on the trending page with dramatic videos about his business challenges and YouTube drama. He is especially known in the community for having animosity with many brands and influencers. This makes him an object of controversy, and therefore, interest.

Dozens of channels exist with the sole purpose of following drama that occurs between beauty guru personalities on the platform. This analysis examined 1,572 videos with the word "beauty" included in the list of each video's tags.

Model Applications

Consumers rely heavily on the opinions of content creators to guide their buying decisions, and brands understand this. Makeup advertising has moved from traditional routes to influencer-focused marketing. Makeup brands form affiliate-link deals with content creators, giving them a share of the profits for makeup sold. They also collaborate to launch exclusive limited-edition products and send out PR packages with thousands of dollars of merchandise to influencers in the hope of gaining positive exposure.

Cosmetics brands can use this analysis to better predict which creator collaborations will procure more exposure for their products at the lowest cost. It can also help to identify less established influencers that have not yet gone viral but are likely to grow in the future. These people are more likely to accept lower profit margins to work with a big brand.

Hypothesis

Null hypothesis: all parameters are equal to zero

Alternative Hypothesis: at least one of the parameters is different than zero

Regression

Model: MODEL1
Dependent Variable: views

Number of Observations Read	1571
Number of Observations Used	1547
Number of Observations with Missing Values	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.291748E15	7.63916E14	678.25	<.0001
Error	1543	1.737876E15	1.126297E12		
Corrected Total	1546	4.029624E15			

Root MSE	1061271	R-Square	0.5687
Dependent Mean	1179025	Adj R-Sq	0.5679
Coeff Var	90.01264		

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

RatingDisabled =	0
CommentDisabled =	0

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1352784	41871	32.31	<.0001
comment_count	1	80.73625	1.81621	44.45	<.0001
RatingDisabled	0	0	.	.	.
dislikeCommentRatio	1	-245545	56175	-4.37	<.0001
CommentViewRatio	1	-128031898	3905527	-32.78	<.0001
CommentDisabled	0	0	.	.	.

Equation:

Views = 1,352,784 - 245,545(DislikeCommentRatio) – 128,031,898(CommentViewRatio) +

80.73(CommentCount)

Model Adequacy

F-Test:

The F-value of 678 is high, suggesting a significant difference from the alternative hypothesis.

T-Test:

The RatingDisabled, DislikePercentage, and CommentsDisabled variables were found to be insignificant in the beauty community regression. Of the variables that are significant, all of them have a P-value of <0.001, which is well below the 0.5 threshold of a 95% confidence interval.

Multicollinearity:

All of the VIF values are relatively low, which leads me to believe that the amount of multicollinearity is not a cause of concern. In this aspect, the model is adequate. The values are higher than they were in the other two scenarios, but still close enough to 1. If they were above 2.5, then I would exclude some variables for multicollinearity.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	880551	51020	17.26	<.0001	.	0
comment_count	1	33.47536	1.42613	23.47	<.0001	0.95848	1.04332
CommentDisabled	0	0	.	.	.	.	.
RatingDisabled	0	0	.	.	.	.	.
DislikePercent	1	-28676	8923.87012	-3.21	0.0013	0.56297	1.77630
dislikeCommentRatio	1	196823	95828	2.05	0.0402	0.54637	1.83026

There is some multicollinearity, but it is between variables where this would be expected. When comments are disabled, the comment count is low, so the correlation is at -0.46. Likewise, the dislike-to-comment ratio is correlated with the dislike percentage. In conclusion, the multicollinearity is not a cause of concern for the genre-specific analysis.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations					
	comment_count	CommentDisabled	RatingDisabled	DislikePercent	dislikeCommentRatio
comment_count	1.00000 1571	-0.04656 0.0651 1571	-0.04656 0.0651 1571	-0.10501 <.0001 1547	-0.20046 <.0001 1547
CommentDisabled	-0.04656 0.0651 1571	1.00000 1571	1.00000 <.0001 1571	. . 1547	. . 1547
RatingDisabled	-0.04656 0.0651 1571	1.00000 <.0001 1571	1.00000 1571	. . 1547	. . 1547
DislikePercent	-0.10501 <.0001 1547	. . 1547	. . 1547	1.00000 1547	0.66049 <.0001 1547
dislikeCommentRatio	-0.20046 <.0001 1547	. . 1547	. . 1547	0.66049 <.0001 1547	1.00000 1547

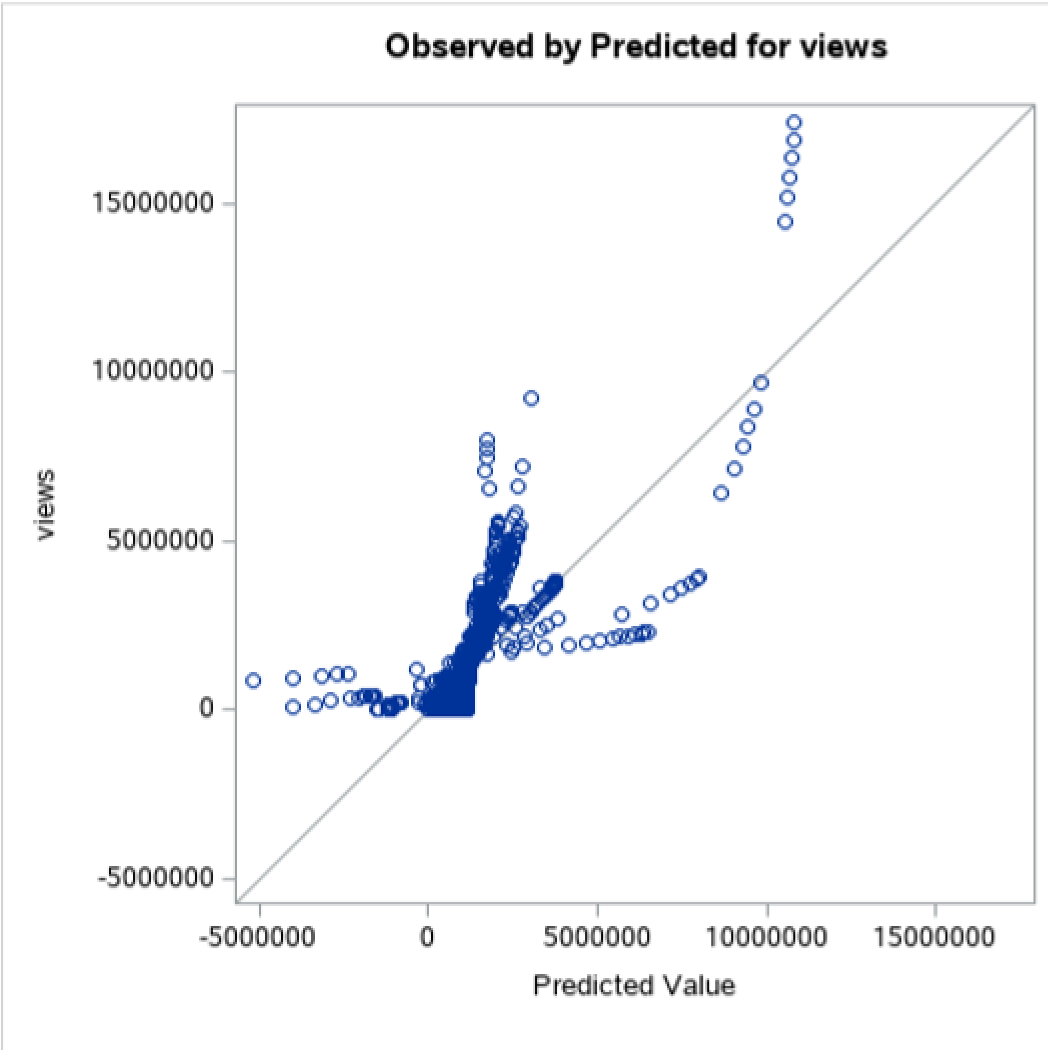
Conclusion

There is enough evidence to reject the null hypothesis. The comment count, dislike-to-comment ratio, and comment-to-view ratio are all significant predictors of view counts in the beauty community. There is more evidence that engagement is a predictor of high view counts than controversy in this genre, contrary to my initial hypothesis.

Potential for Improvement

The beauty community dataset seems to follow a logarithmic trend more than the other categories. I believe that when looking at genre-specific models, there is more variance in the type of trends we observe. Models applied to smaller subsets of the YouTube ecosystem should be evaluated on a case-by-case basis. More/different variables or data transformations may be required for the best results.

In this specific case, the beauty community is known to have a high degree of subscriber loyalty and in-fighting. People often tend to attatch to a few content creators and express a feeling of attachment to them after watching their videos for years. Subscriber counts and the percentage of new viewers that subscribe is probably a strong indicator of view counts.



Code Used

import

```
proc import datafile="/folders/myfolders/data/BeautyCommunityData12.csv"
out=beautydata dbms=csv replace;
getnames=yes;
run;
```

Regression

```
ods noproctitle;
ods graphics / imagemap=on;
```

```
proc reg data=WORK.beautydata alpha=0.05 plots(only)=(diagnostics residuals
		observedbypredicted);
	model views=comment_count RatingDisabled dislikeCommentRatio CommentViewRatio
		DislikePercent CommentDisabled /;
	run;
quit;
run;
```

Multicollinearity

```
proc corr;
var comment_count CommentDisabled RatingDisabled DislikePercent dislikeCommentRatio;
run;

proc reg;
model views = comment_count CommentDisabled RatingDisabled DislikePercent dislikeCommentRatio /
vif tol collin;
run;
```