Content Categorization and Viewer Behavior: A Statistical Examination of YouTube Genres

By Brenna Somdahl-Sands, Ethen Kantu, Natalia Morales Flores, and Soulai Vang

Introduction

YouTube has become a prominent platform in which creators from diverse backgrounds and with diverse intentions choose to share their work. We initiated this research project after noticing compelling content on certain YouTube channels that, despite their potential for wider acclaim, haven't received the recognition they merit, or on the other hand, seem to receive more recognition than what one would expect. This observation sparked our interest in understanding the factors that influence the popularity of a YouTube channel. Our research question for this project poses the inquiry of **What is the relationship between a YouTuber's category and their popularity, as measured by the number of subscribers and video views?** Our goal with this research is to uncover the primary elements that contribute to a YouTube channel's success and to identify if a channel's category and content type are influential factors. From an individual content creator perspective, insights into popularity factors can contribute to strategy for audience engagement and the enhancement of visibility and impact of the desired content. From a broader point of view, understanding these relationships can contribute to the understanding of media culture and digital trends and influence. We plan to conduct a comprehensive analysis to identify strategies for boosting a channel's popularity. The dataset for this study has been sourced from Kaggle.

Data and Exploratory Data Analysis

Using the Global YouTube Statistics from kaggle.com, we have multiple relevant variables to look at in order to answer our main research question. In the dataset, it includes variables such as *subscribers*, *video views*, and *category*. These main three variables will be the basis of our research question, with *subscribers* and *video views* being our outcome variables, while *category* will be our main predictor variable. Extending off of our main research question, we are also interested in seeing if other predictor variables could help explain the number of subscribers and video views that each YouTube channel has. To this end, we include the *population* of the country of origin of the channel. This dataset was obtained through the use of 'various reputable sources', collected and compiled by Nidula Elgiryewithana. However, it is to be noted that the specific resources as to how the data was collected are unknown to us and would need further research into the methods as to how it was gathered.

Our provided numerical summaries count the total number of YouTube channels in each category (Table 1), which we will later use in our statistical models as a predictor variable, along with the mean and standard deviation of each of our relevant quantitative variables (Table 2). Our provided data visualizations (Figure 1-6) showcase our selected linear regression models.

Methods:

To address our research inquiry, we employed linear and multilinear regression models to elucidate the interplay between various variables. Throughout our investigation, we experimented with several regression models, each designed to capture distinct relationships. The models we considered are outlined below:

1. **Expected Subscribers (in thousands) Given Category Parsing:**

$$E[Subscribers|Category] = \beta_0 + \beta_1 Comedy + \beta_2 Education + \beta_3 Entertainment + \beta_4 FilmAndAnimation + \beta_5 Gaming + \beta_6 HowToAndStyle + \beta_7 Music + \beta_8 NewsAndPolitics + \beta_9 Other + \beta_{10} PeopleAndBlogs$$

This model explores how the number of subscribers is expected to vary with changes in the category variable.

H_0: There is no significant relationship between the number of subscribers and the YouTube channel's category.
H_1: There is a significant relationship between the number of subscribers and the YouTube channel's category.

The covariates shown above in our model statement related to our categories will be repeated in future model statements.

2. **Expected Video Views (in millions) Given Category Parsing:**

$$E[Views|Category] = \beta_0 + \beta_1 Comedy + \ldots + \beta_{10} PeopleAndBlogs$$

Here, we examine the relationship between video views and the category variable using a similar linear regression framework.

H_0: There is no significant relationship between the number of video views and the YouTube channel's category.
H_1: There is a significant relationship between the number of video views and the YouTube channel's category.

3. **Expected Subscribers for the Last 30 Days (in thousands) Given Category Parsing:**

$$E[Subscribers30|\, Category] = \beta_0 + \beta_1 Comedy + \ldots + \beta_{10} PeopleAndBlogs$$

This model specifically focuses on the anticipated number of subscribers in the last 30 days, considering the category.

H_0: There is no significant relationship between the number of subscribers in the last 30 days and the YouTube channel's category.
H_1: There is a significant relationship between the number of subscribers in the last 30 days and the YouTube channel's category.

4. **Expected Video Views for the Last 30 Days (in millions) Given Category Parsing:**

$$E[Views30|Category] = \beta_0 + \beta_1 Comedy + \ldots + \beta_{10} PeopleAndBlogs$$

Similar to the third model, this examines the expected video views for the last 30 days based on the category.

H_0: There is no significant relationship between the number of video views in the last 30 days and the YouTube channel's category.
H_1: There is a significant relationship between the number of video views in the last 30 days and the YouTube channel's category.

5. **Expected Subscribers (in thousands) Given Population (in millions):**

$$E[Subscribers|Population] = \beta_0 + \beta_1 Population$$

In this instance, we explore the relationship between the number of subscribers and the population, considering the population variable as a factor.

H_0: There is no significant relationship between the number of subscribers and the population of the YouTube channel's country of origin.
H_1: There is a significant relationship between the number of subscribers and the population of the YouTube channel's country of origin.

6. **Subscriber Count based on Population and Category:**

$$E[Subscribers|Population, Category] = \beta_0 + \beta_1 Population + \beta_2 Comedy + \ldots + \beta_{11} PeopleAndBlogs$$

In this instance, we explore the relationship between the number of subscribers, population and category, considering the population and category variables as a factor.

H_0: There is no significant relationship between the number of subscribers and the population of the YouTube channel's country of origin when considering the influence of the category.
H_1: There is a significant relationship between the number of subscribers and the population of the YouTube channel's country of origin when considering the influence of the category.

7. **Subscriber Count based on Population, Category, and Interaction Term:**

$$E[Subscribers| Population, Category] = \beta_0 + \beta_1 Population + \beta_2 Comedy + \ldots + \beta_{11} PeopleAndBlogs + \beta_{12} Population * Comedy + \ldots + \beta_{21} Population * PeopleAndBlogs$$

In this instance, we explore the relationship between the number of subscribers, population and category, with an interaction term between population and category.

H_0: There is no significant effect between the population of a YouTube channel's country of origin and its category on the number of subscribers.
H_1: There is a significant effect between the population of a YouTube channel's country of origin and its category on the number of subscribers.

Results:

For the linear regression section, we examined the relationship between category and the four popularity indicators. On average, the number of video views in the millions when YouTube videos have the category of education is 15 billion video views, for music is 15 billion video views, for people and blogs is 6 billion video views, and for gaming is 7.6 billion video views. For the video views over the last 30 days, on average YouTube videos in the category of education revived 186 million video views, music received 179 million video views, people and blogs received 143 million video views, and gaming received 72 million video views over the past 30 days. The total number of subscribers when the YouTube video is in the category education averaged 265 thousand subscribers, music averaged 257 thousand subscribers, People and blogs averaged 211 thousand subscribers, and gaming averaged 209 thousand subscribers. On average the number of subscribers gained in the last 30 days in thousands when the YouTube video is in the category of education is 3 thousand subscribers, music is 2 thousand subscribers, People and blogs is 4 thousand subscribers, and gaming is 2 thousand subscribers. The population of the country of origin of a YouTube channel was also examined to see the effect on the subscribers of the channel. The interval (232,199) provides a range of possible values for the true mean of subscribers in the millions for countries with a population of 0. The interval (0.057, 0.006) provides a range of plausible values for the average true difference in millions of subscribers for a 1 million increase in the population of origin to also examine the significance of the difference of popularity by channel type hypothesis tests were also conducted.

On average the number of subscribers in thousands when the YouTube video is in the category of education is 253 thousand subscribers, Music is 247 thousand subscribers, People and Blogs is 207 thousand subscribers and Gaming is 198 thousand subscribers when the YouTube channels are from countries with a population of 0. For every increase of 1 million in the population there is an average increase of 28 subscribers when the video category is the same.

Channels with the category of Education have on average 287 thousand subscribers when their country of origin's population is 0, for every 1 million people increase in the country of origin's population there is an average of 41 subscribers decrease in subscribers. Channels in the category of Gaming have an average of 214 thousand subscribers when their country of origin's population is 0, for every 1 million people increase in the country of origin's population there is an average 78 subscriber decrease in subscribers. Channels in the category of Music have an average of 211 thousand subscribers when their country of origin's population is 0, for every 1 million people increase in the country of origin's population there is an average 103 subscriber increase in subscribers. Channels with the category of People and Blogs have an average of 239 thousand subscribers when their country of origin's population is 0, for every 1 million people increase in the country of origin's population there is a 74 subscriber decrease in subscribers.

For this analysis, we will use a p-value threshold of 0.05 to show if the plot is significant to our research question. A Low p-value (which for use will be anything less than 0.05): This will indicate strong evidence against our null hypothesis. In this case, we will reject the null hypothesis and conclude that there is a statistically significant effect or difference. For each of the relationships observed between category and the popularity of the video (based on the parameters set above), there is enough evidence to conclude that there is a relationship between the popularity of a Youtuber and the category of the YouTuber.

Conclusion

It is of note that our study is limited in the fact that we primarily only focused on a channel's category as the main predictor of how popular a YouTube channel is (based on the total number of subscribers and video views). We would need to conduct further research into other possible predictor variables such as when the channel was created, or the shift to other media platforms (e.g. TikTok, Instagram). There is also the possibility of people subscribing to a channel with multiple accounts on YouTube and bots artificially increasing the number of subscribers on YouTube. Also, different generations have different watching habits for example, a child with relatively little object permanence would be more likely to rewatch videos increasing the views. Also specifically for channels such as education, it is likely that the children, who are the consumers of the media are less likely to be the ones subscribing to it.

The origins of this data are ambiguous and therefore we don't know if this data is sourced ethically. It may come from a breach of privacy of the creator's YouTube analytics. It is difficult then to make conclusions and base behavior off of findings of a dataset that has unknown origins so we do not know how reliable the data is.

Looking at our results, on average, it is predicted that Music and Education YouTube channels have the highest number of total subscribers across all categories, averaging 265 thousand and 257 thousand total subscribers respectively. However, when we look at the number of subscribers that YouTube channels have gained in the last 30 days, we can see how their growth compared to other categories is slower than other categories. The top category for growth is People & Blogs, with an average growth rate of 4 thousand subscribers in the last 30 days.

For total video views for YouTube channels, Education and Music categories average about 15 billion total video views on their channels, which corresponds to the total category of YouTube channels that also have the highest number of total subscribers. In the last 30 days, this trend of Education and Music YouTube channels seems consistent, as Education and Music YouTube channels have an average of 186 million and 179 million total video views in the last 30 days.

When looking at our multiple logistic regression model, we can see how there is a positive increase in subscribers when we do not include an interaction term. However, when we do include an interaction term, the country of origin's population has an effect on the relationship between category and channel popularity as measured in the total number of subscribers that each channel has. These tests and conclusions are fairly reliable because given the large sample size of the dataset it is unlikely that a type two error has occurred. However, due to the amount of regressions run, there is a 26.5% probability that a type 1 error occurred.

Based on a previous similar research article titled "YouTube channels, uploads, and views: A statistical analysis of the past 10 years" by Mathias Bärtl, he concludes that YouTube channels that are most popular follow a 'rich-get-richer' phenomenon since channels that already tend to get many views are bound to get more views in the future. This follows our conclusions of channel categories such as music channels being some of the top channels on YouTube.

Given these results, for content creators, if one wants to maximize their popularity then the channel category matters. Through this analysis, we expect that on average, Music and Education videos will be more popular than Gaming or People and Blogs. Therefore to maximize popularity between the four categories it would be beneficial to create a Music or Education channel over a Gaming or People and Blogs channel.

References

Bärtl, M. "YouTube channels, uploads, and views". *Convergence: The International Journal of Research into New Media Technologies*, 24(1), 16–32, 2018, https://doi.org/10.1177/1354856517736979

Elgiriyewithana, Nidula. "Global YouTube Statistics 2023." *www.kaggle.com*, 2023, www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023.

Appendix:

Channel Category	
Category	**Number**
Comedy	69
Education	45
Entertainment	241
Film & Animation	46
Gaming	94
How to & Style	40
Music	202
News & Politics	26
People & Blogs	132
Other	44
Nan	46

Table 1. Number of channels in each category of channel

Quantitative Variables		
Variable Name	Mean	Standard Deviation
Video Views (Millions)	11000	14000
Subscribers (Millions)	23	18
Video views for the last 30 days* (Millions)	176	416
for the last 30 days* (Millions)	.3	.6
population* (Millions)	430	473

* denotes that ‘NaN’ values were excluded from the mean calculation

Table 2. The mean and standard deviation of video views, subscribers, number of new video views for the last 30 days, number of new subscribers gained in the last 30 days, and the population of the county where the YouTube channel originated.

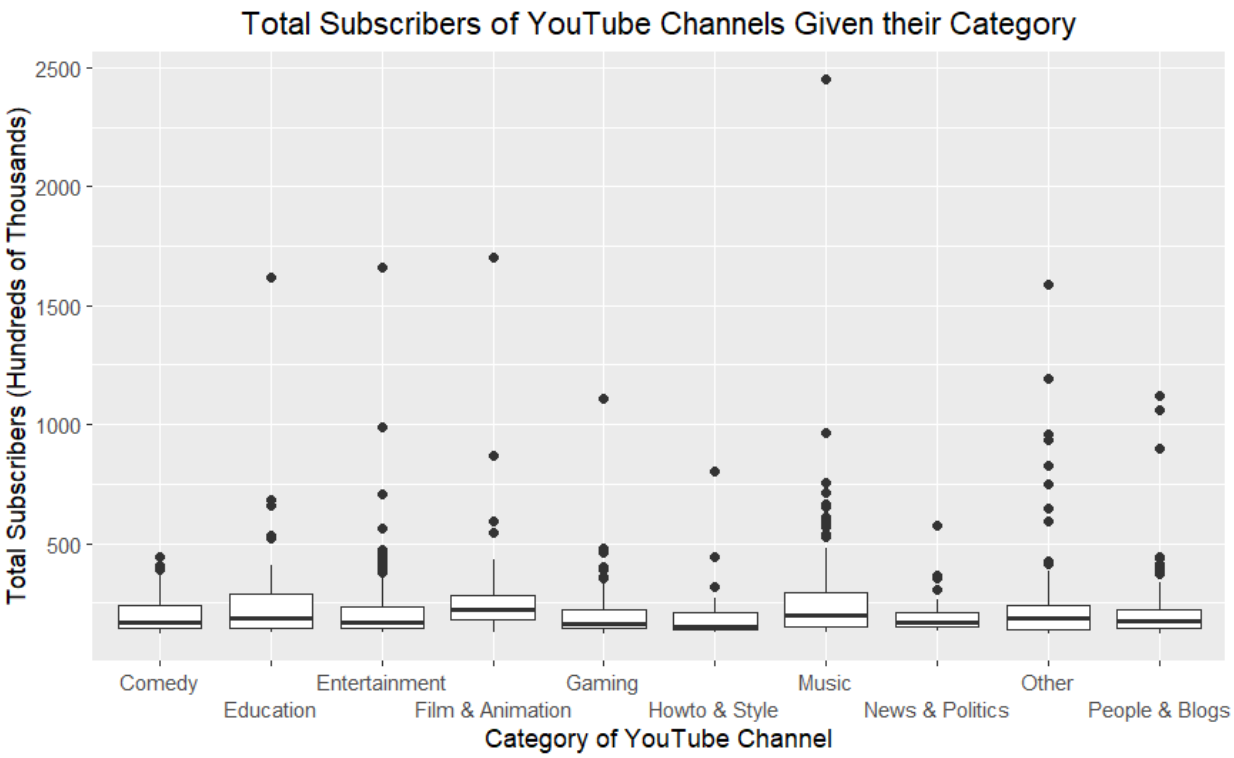

Figure 1. A box plot showing the differences between categories and total number of subscribers

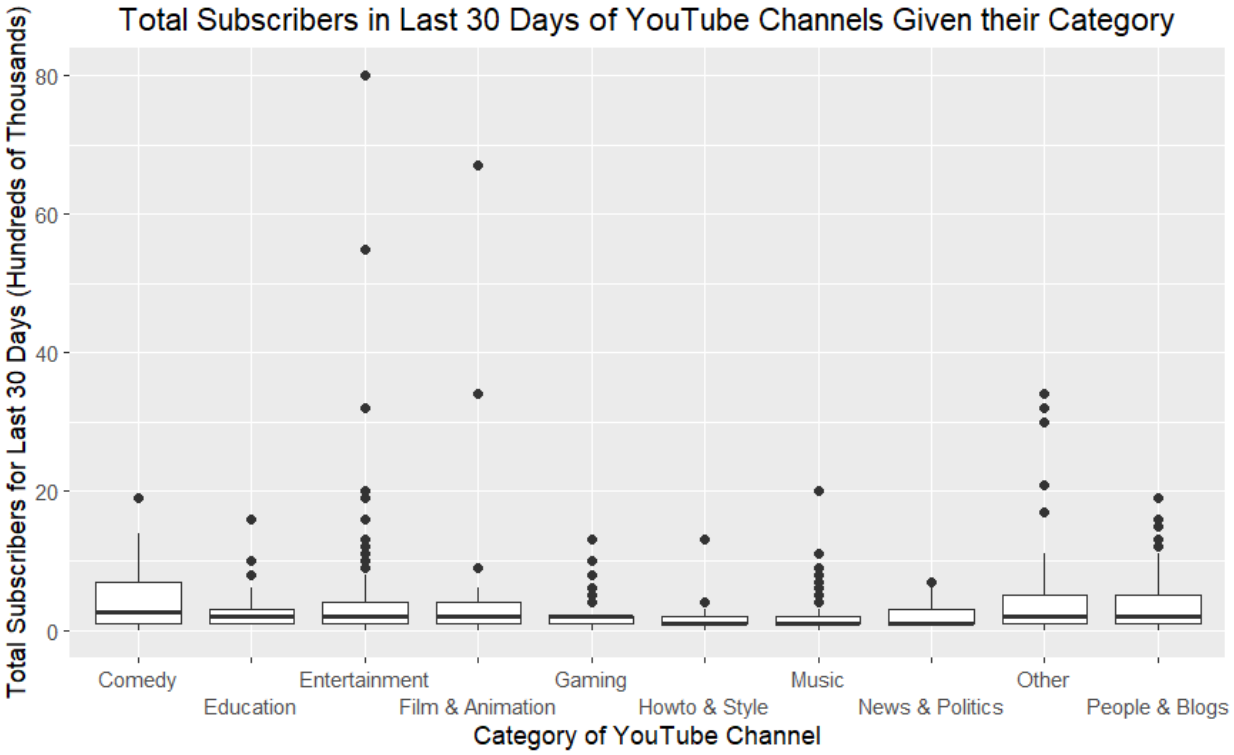

Figure 2. A box plot showing the differences between categories and the total number of subscribers for the past 30 days

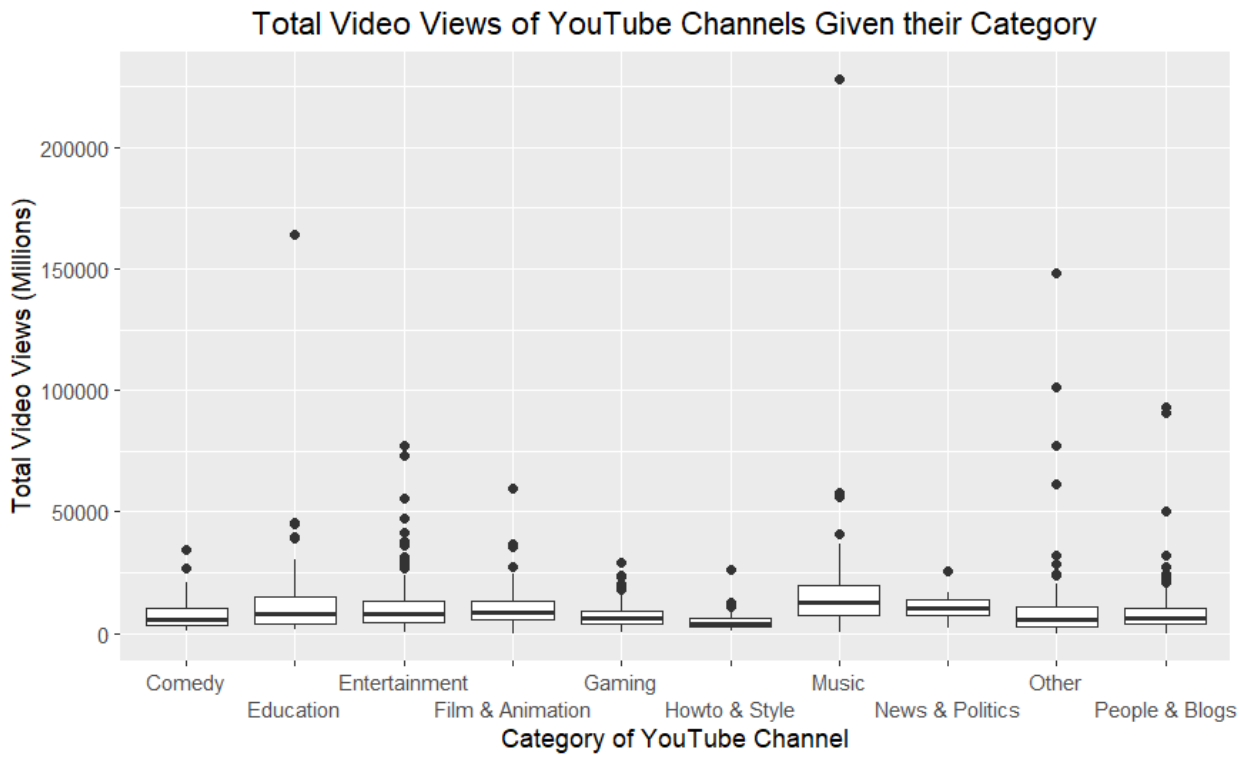

Figure 3. A box plot showing the differences between categories and the total amount of video views

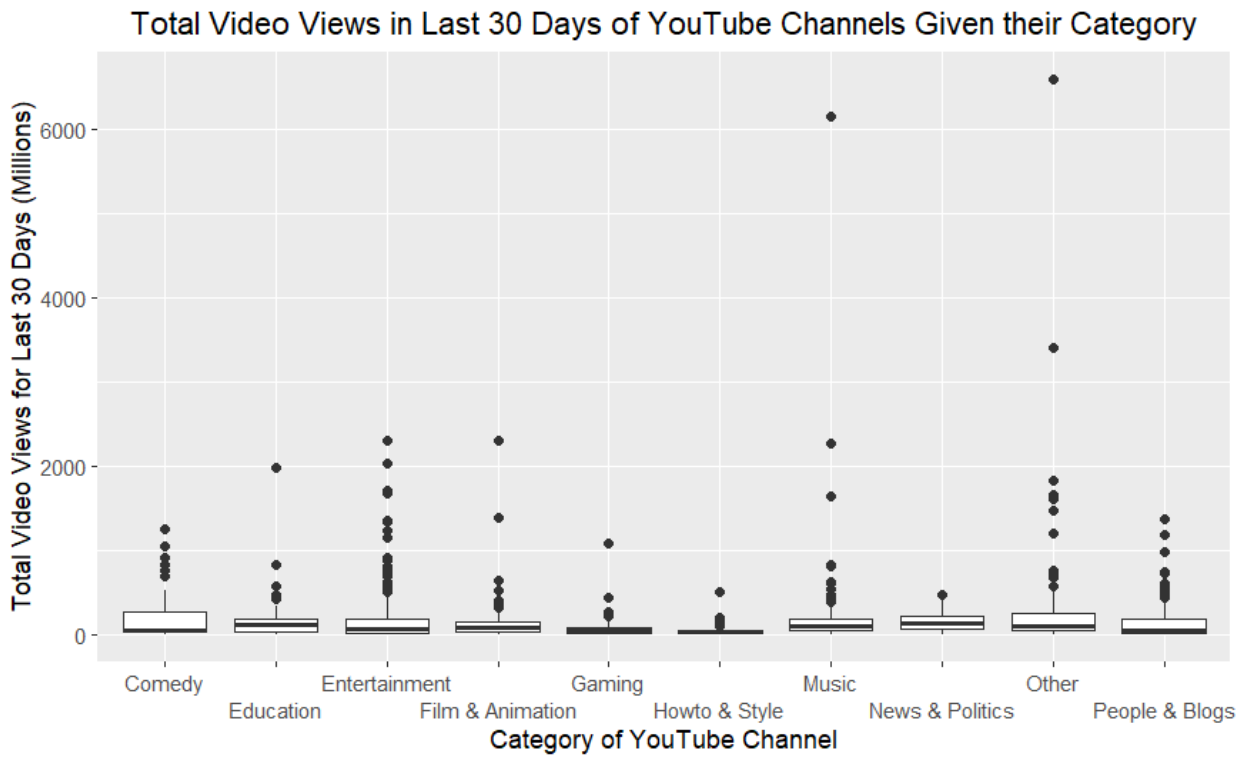

Figure 4. A box plot showing the differences between categories and the total amount of video views for the past 30 days

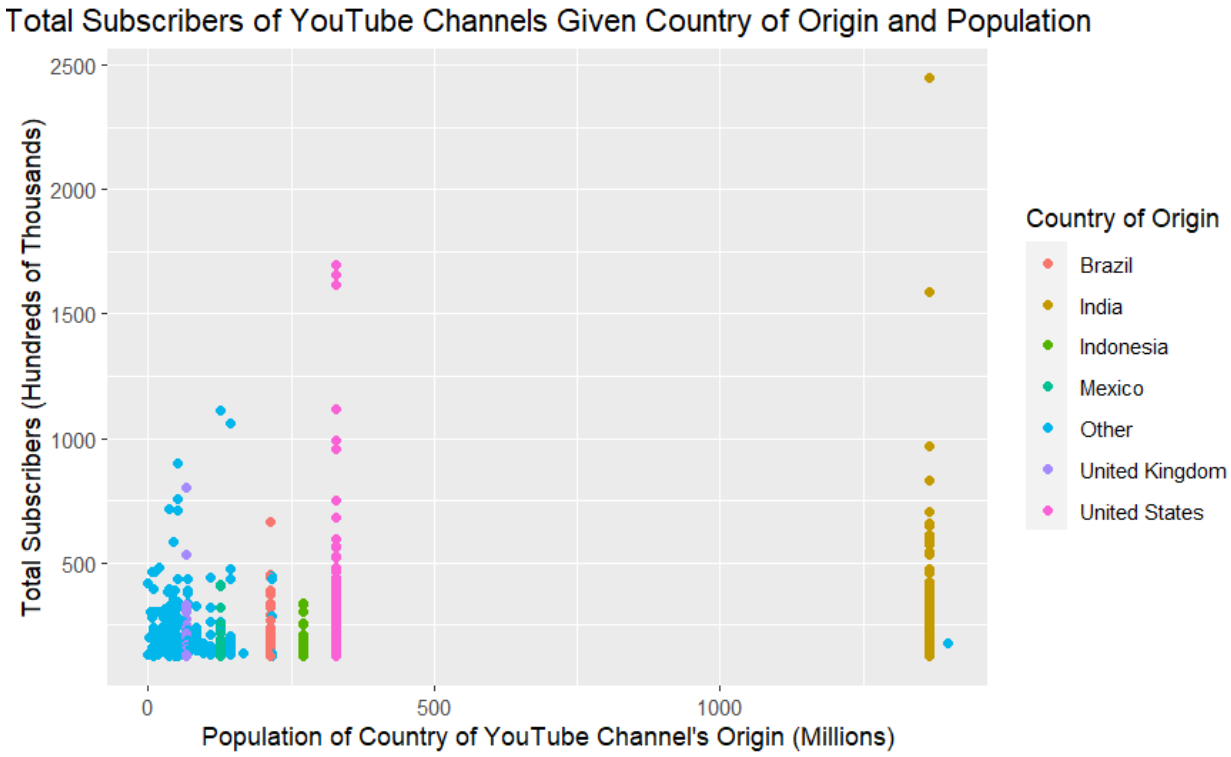

Figure 5. A scatterplot showing the differences between the total number of subscribers a YouTube channel has and its country of origin's population

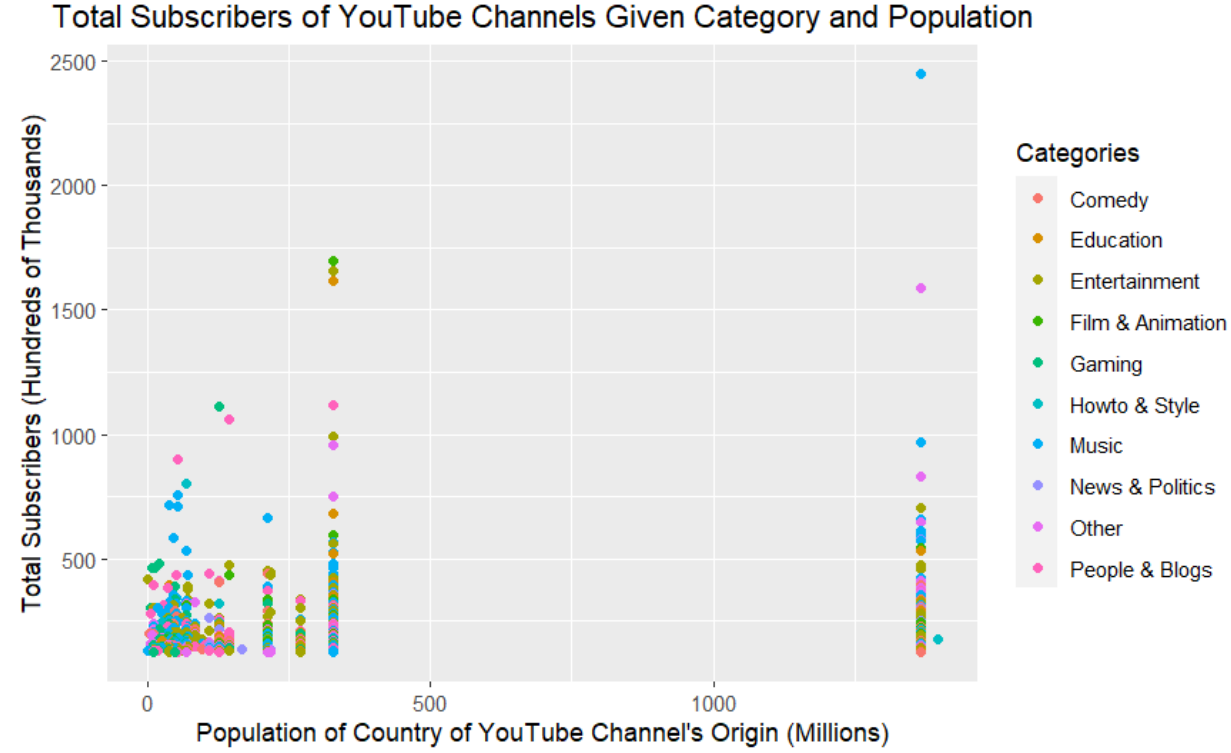

Figure 6. A scatterplot showing the differences between the total number of subscribers a YouTube channel has and its category based on the country of origin's population

Effect of Category of the Videos on the number of subscribers (in thousands)				
Category	Estimate	Standard Error	Test Statistic	P value
Education	265	33.375	1.923	0.01169
Gaming	209	27.612	0.264	
Music	257	24.287	2.304	
People and Blogs	211	25.875	0.361	

*four of the ten categories are present in this table

Effect of the category of the videos on the number of views of a video (in millions):				
Category	Estimate	Standard Error	Test Statistic	P value
Education	15480.3	2649.2	2.834	1.42E-07
Gaming	7634.4	2191.7	-0.154	
Music	492.4	1927.8	3.88	
People and Blogs	6355.9	2053.9	0.787	

*four of the ten categories are present in this table

Effect of the category of the videos on the number of subscribers in the last 30 days (in thousands)				
Category	Estimate	Standard Error	Test Statistic	P value
Education	3.0883	1.3886	-1.017	0.008006
Gaming	2.1865	1.2113	-1.91	
Music	2.0092	1.0706	-2.327	
People and Blogs	3.8644	1.1228	-0.566	

*four of the ten categories are present in this table

Effect of the category of the videos on the number of views of the video in the last 30 days (in thousands)				
Category	Estimate	Standard Error	Test Statistic	P value
Education	185863494	82426525	0.007	0.01418
Gaming	71636921	68127708	-1.669	
Music	179028215	60039673	-0.105	
People and Blogs	143220073	63847312	-0.659	

*four of the ten categories are present in this table

Effect of the population (in millions) of a country on the number of subscribers (in thousands)				
	Estimate	Standard Error	Test Statistic	P Value
Intercept	215.19913	8.15241	26.397	0.01516
Slope	0.03104	0.01276	2.433	

*four of the ten categories are present in this table

Effect of Category of the Videos on the number of subscribers (in thousands) accounting for Country Population				
Category	Estimate	Standard Error	Test Statistic	P value
Education	253.42467	34.87119	1.936	0.004388
Gaming	197.51449	29.4432	0.393	
Music	247.05532	25.31387	2.415	
People and Blogs	164.97739	27.73014	0.756	
Population Slope	0.0278	0.01303	2.134	

*four of the ten categories are present in this table

Effect of Country Population on the Relationship Between Category of the Video and Subscribers (in thousands)				
Category	Estimate	Standard Error	Test Statistic	P value
Education	287.0751	51.39077	1.868	8.95E-06
Gaming	214.5811	38.7964	0.606	
Music	211.19533	34.08708	0.591	
People and Blogs	239.21621	37.41178	1.287	
Slope Education	-0.02382	0.06517	-0.629	
Slope Gaming	-0.06131	0.09506	-0.825	
Slope Music	0.11964	0.05123	2.001	
Slope People and Blogs	-0.05686	0.05865	-1.262	

*four of the ten categories are present in this table