Sentiment Analysis of Canadian YouTube Viewership

Kaveh Alemi

STAT 403 - D100: Intermediate Sampling and Experimental Design

April 25, 2021

Abstract

YouTube is one of the most visited websites globally and, as a result, collects a considerable amount of user data. This project aims to analyze the viewer sentiment of the different Canadian YouTube video categories from the past six months. ANOVA and Tukey's HSD are the main statistical methods used to analyze the difference in viewer sentiment between the video categories. Furthermore, this project concludes by finding significant differences between the viewer sentiment of each Canadian YouTube video category. The *News & Politics* video category received the most negative sentiment, and the *Comedy* video category received the most positive sentiment.

Keywords: Canadian YouTube Video Categories, Multiple Group Comparison Methods, YouTube User Sentiment Analysis

Table of Contents

Introduction

YouTube has over two billion users and flawlessly streams one billion hours' worth of digital video every day (YouTube, 2021). This enormous user traffic paired with YouTube's sophisticated software infrastructure spells out one word: data. It is unknown exactly how much user data YouTube collects, but we can be assured that it is a lot (Smith, 2020). The consequence of all of this data is that the possibilities of analyzing YouTube's userbase are virtually endless. Moreover, this project focuses on analyzing a subsection of YouTube's user-related data. We will specifically analyze the viewer sentiment shared towards the 14 Canadian YouTube video categories to draw conclusions about Canada's YouTube viewership patterns.

The first step of this project was acquiring YouTube-related datasets. Since YouTube does not share its data openly, finding credible datasets was challenging. However, they do offer one open-access database that is hosted and updated on Kaggle.com (refer to *Appendix Section i* for instructions on accessing the database). This database contains datasets specific to trending videos from a certain number of countries, including Canada. A video is labelled as trending if it gains many views in a short amount of time. Moreover, the specific dataset used for this project contains all Canadian trending YouTube videos uploaded in the last three months of 2020 and the first three months of 2021. This dataset contains approximately 31 thousand rows, where each row corresponds to one trending video. Each video contains ten attributes, but in this project, we will only work with the following four: upload year, video category, total views, and like count (refer to *Appendix Section ii* for a snippet of the dataframe used for the main analysis).

After obtaining the data, we then began to structure the analysis. The acquired dataset will be used as the experiment's data sample. This sample will be used to make inferences about the population, which is all YouTube videos uploaded in the last three months of 2020 and the first three months of 2021 in Canada. The following figure visualizes the relationship between the experiment's sample and population:

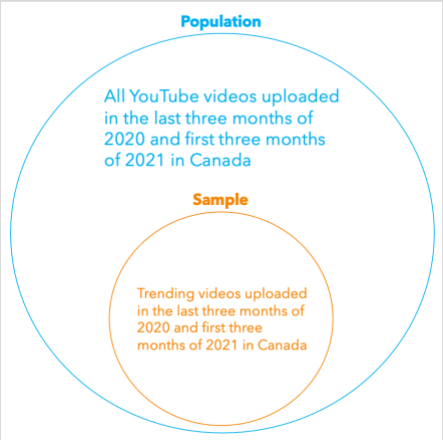


Figure 1: Structure of Analysis

It is evident that this data sample is not a simple random sample and instead is a convenient sample. This aspect of the experiment is likely to introduce some bias into the experiment and will be further addressed in the *Limitations and Future Work* section.

The focus of this project is to analyze the viewer sentiment shared towards the 14 Canadian YouTube video categories. In the context of this project, viewer sentiment is defined as the general attitude shared towards a YouTube video. We will quantify viewer sentiment as the ratio between a video's total like count and total view count, which will be formally called the Like-To-View ratio or LTV Ratio for short. The equation defining the LTV Ratio can be seen below:

$$LTV\ Ratio = \frac{Total\ Like\ Count}{Total\ View\ Count}$$

To expand on our definition of viewer sentiment, a high LTV Ratio will be indicative of positive viewer sentiment, and a low LTV Ratio will be indicative of negative viewer sentiment. Moreover, since the LTV ratio did not come with the original data, it had to be calculated for every row using data engineering.

Furthermore, the primary purpose of this project is to thoroughly answer the following question: *Is there a difference in the Like-To-View ratio of the 14 Canadian YouTube video categories from the last three months of 2020 and the first three months of 2021?* We will employ various visualizations and statistical tests on our sample data to adequately answer this question.

Method

i) Exploratory Data Analysis

To gain intuition about the sample data, we plotted the following figure which contains a boxplot for the LTV Ratio of each Canadian video category from the past six months:

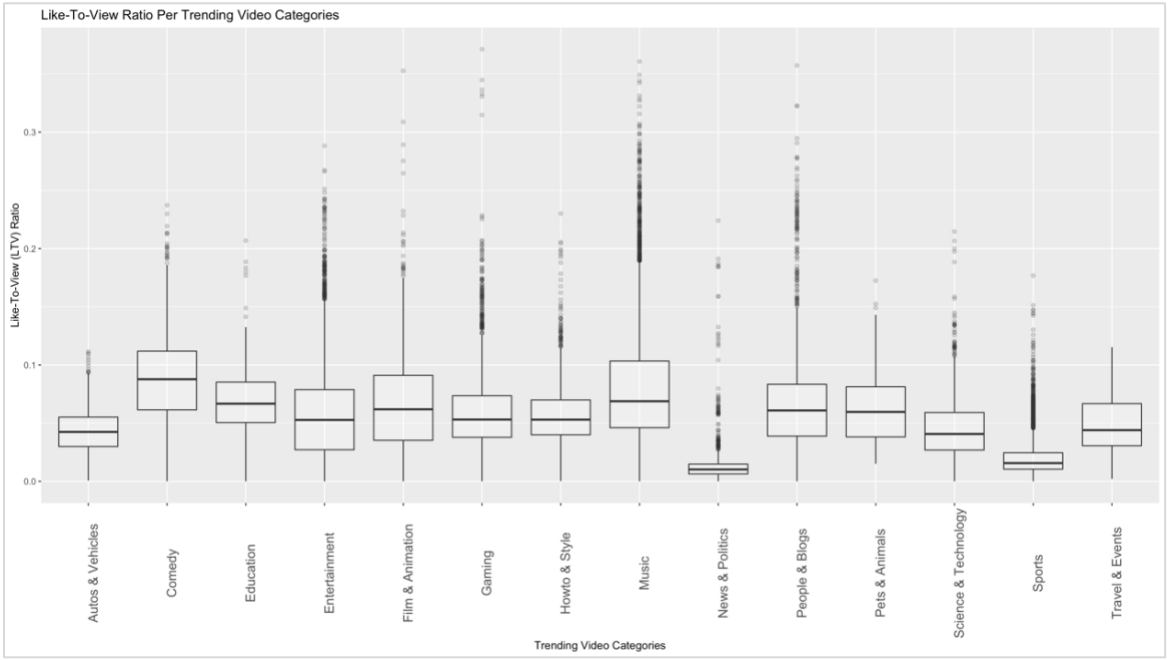


Figure 2: Boxplot of LTV Ratio of Canadian Video Categories

The above figure shows a visual difference in the median LTV Ratio of the 14 video categories. In addition, the sizes of the boxplots also differ, indicating a difference in the spread of the LTV Ratios. Moreover, the two categories, *Comedy* and *News & Politics*, appear to have the highest and lowest LTV Ratios, respectively.

The next step of the exploratory data analysis consisted of investigating to see if the upload year variable is extraneous. If the upload year variable has a significantly different LTV Ratio for each of its values, then its effect must be accounted for in the main statistical test. By accounting for this extraneous effect, we can be more certain that the difference in LTV Ratio is due to video category rather than upload year. Moreover, we first created a boxplot of the LTV Ratio for each video category and upload year. This figure can be seen below:

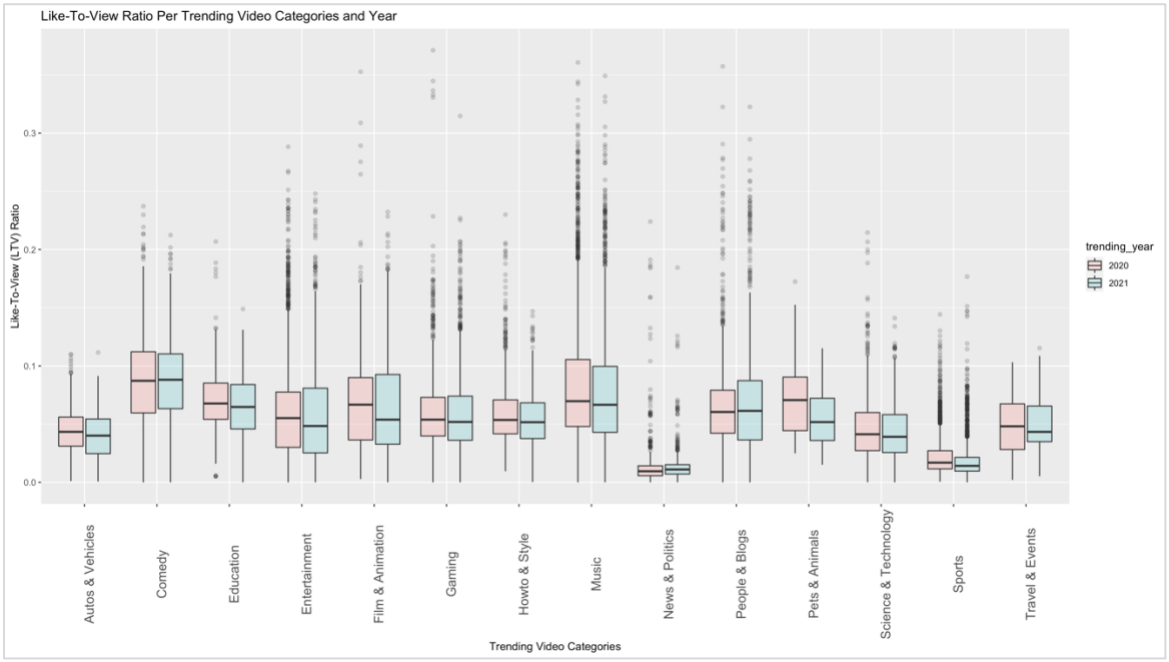


Figure 3: Boxplot of LTV Ratio for Each Video Category and Upload Year

The above figure shows a difference in the median LTV Ratio between the two upload years of some video categories. For example, the *Pets & Animals* category appears to have a noticeable difference in LTV Ratio between the 2020 and 2021 upload years. Conversely, *News & Politics* appears to have a similar LTV Ratio between the 2020 and 2021 upload years.

In addition to the above boxplot, we calculated a confidence interval for the LTV Ratio of the 2020 and 2021 upload years. These two intervals contain a 95% confidence range for the mean LTV Ratio of Canadian videos uploaded in 2020 and 2021. The equation used to calculate the intervals can be seen below:

$$\bar{Y}_i \pm t_{0.025,(n_i-1)}\left(\frac{\hat{\sigma}_i}{\sqrt{n_i}}\right) \; where \; i \in \{2020, 2021\}$$

In the above equation, $\bar{Y}_i$ is the sample mean of upload year *i*, $t_{0.025,(n_i-1)}$ is the T critical value of upload year *i*, n_i is the sample size of upload year *i*, and $\hat{\sigma}_i$ is the sample standard deviation of upload year *i*. The following figure visualizes these two confidence intervals:

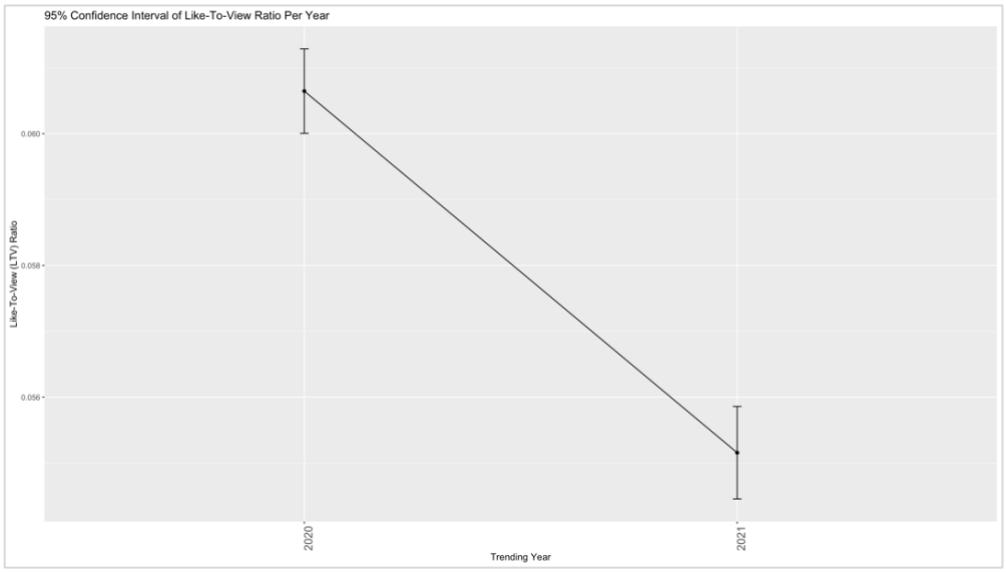


Figure 4: 95% CI for LTV Ratio of Each Upload Year

In the above figure, it is apparent that the two confidence intervals do not overlap, which entails a significant difference in the LTV Ratio of videos uploaded in 2020 and 2021. As a result, in the main statistical test, we will mitigate the extraneous effect introduced by the upload year variable to fully capture the effect of video category on LTV Ratio.

ii) Statistical Testing

The explanatory data analysis influenced us to choose the Randomized Block Design (RBD) as the experiment's main design. In this design, the primary factor will be the video category, the blocking factor will be the upload year, and the response will be the LTV Ratio. Moreover, we will test to see if there exists a significant difference in the LTV Ratio of the 14 video categories while blocking on the upload year. The following equation defines the RBD model:

$$Y_{i,j} = \mu + \alpha_i + B_j + \epsilon_{ij}$$

In the above equation, $Y_{i,j}$ is the LTV Ratio, μ is the grand mean, α_i is the effect of the i^{th} video category, B_j is the effect of the j^{th} upload year (block), and ϵ_{ij} is the random error term assumed to follow $N(0, \sigma^2)$. Furthermore, the hypotheses of interest are:

$$H_0\colon \mu_1 = \cdots = \mu_{14}$$

$H_A\colon \mu_i \neq \mu_j$ for at least one pair of unique video categories where $i, j \in \{1, \ldots, 14\}$

We will use a standard two-way ANOVA to model the decomposition of variance. The template for the ANOVA table can be seen below:

Variation	SS	DF	MS	F	P
Video Category	SS_{VC}	$t-1$	MS_{VC}	MS_{VC}/MS_{Er}	$P(F_{t-1,(b-1)(t-1)} > F_{obs})$
Upload Year	SS_{UY}	$b-1$	MS_{UY}		
Error	SS_{Er}	$(t-1)(b-1)$	MS_{Er}		
Total	SS_{To}	$N-1$	MS_{To}		

Table 1: Two-way ANOVA Table Template for Decomposition of Variance

iii) Post-hoc Analysis

If the two-way ANOVA shows a significant difference in the LTV Ratio of the video categories, a follow-up post-hoc analysis will be performed. We will conduct a multiple group comparisons technique to better understand the specific differences in the LTV Ratio of each pair of categories. We will use a commonly used test called the Tukey-Kramer test.

This test first considers a pair of video categories *i* and *j*, and then constructs a 95% confidence interval for the difference in their mean LTV Ratios. If this confidence interval contains 0, then video categories *i* and *j* are 95% likely to have the same mean LTV Ratio; if the interval does not contain 0 then the mean LTV Ratios are 95% likely to differ. Moreover, the equation used to construct the Tukey-Kramer confidence interval can be seen below:

$$\bar{y}_i - \bar{y}_j \pm \frac{q_{0.05,k,N-k}}{\sqrt{2}} \hat{\sigma}_\epsilon \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \qquad i,j \in \{1,\dots,14\}$$

Where $\bar{y}_i$ and $\bar{y}_j$ are the mean LTV Ratios of the i^{th} and j^{th} video categories, $q_{0.05,k,N-k}$ is the critical value obtained from the studentized range distribution, $\hat{\sigma}_\epsilon$ is the standard deviation of the LTV Ratio of the entire sample, and n_i and n_j are the sample sizes of the i^{th} and j^{th} video categories. Furthermore, we will apply the Tukey-Kramer test for all category pairs to assess every potential difference in the LTV Ratio. This will require a total of $\binom{14}{2} = 91$ confidence intervals to be constructed and analyzed.

If you would like to run the source code that was used in this section refer to *Appendix Section iii* for instruction on how to correctly run the code. If you like to view the source code that was used for this section refer to *Appendix Section v*.

Results

In this section we will evaluate and interpret the results of the two-way ANOVA and the Tukey-Kramer test.

i) Two-way ANOVA

We calculated the two-way ANOVA in R using the *aov* function. The following table contains the test results:

Variation	SS	DF	MS	F	P
Video Category	13.24	13	1.0181	714.49	2×10^{-16}
Upload Year	0.024	1	0.0674		
Error	44.90	31512	0.0014		
Total	58.164	31526	1.0869		

Table 2: Two-way ANOVA Results for Decomposition of Variance

This table contains a very large F-value, and a corresponding p-value that is very close to zero. Therefore, at a standard alpha level of 5%, we can reject the null hypothesis given that the p-value is less than 0.05. By rejecting the null hypothesis, we can conclude that the 14 Canadian YouTube video categories from the past six months have different mean LTV Ratios. An equivalent conclusion is that in the past six months, there have been different levels of viewer sentiment shared towards the 14 Canadian YouTube video categories.

ii) Tukey-Kramer Test

We applied the Tukey-Kramer test to the sample data using R's *TukeyHSD* function. The output of this test is a table that contains a confidence interval for the difference in mean LTV Ratio of every pair of video categories. Given that this table is very large, we have not included it in this section and instead encourage the reader to refer to *Appendix Section iv*. Moreover, by closely inspecting the test table, we can make the following inferences.

First, every confidence interval for the *News & Politics* category is of the form $(\bar{y}_{News\ \&\ Comedy} - \bar{y}_j)$, and is entirely negative. This indicates that this specific category has the lowest LTV Ratio among all of the 14 video categories. As a result, we can conclude that the *News & Politics* category has received the most negative viewer sentiment among all Canadian YouTube video categories from the past six months. Furthermore, this conclusion intuitively makes sense because internet users are likely to share negative attitudes towards opposing political topics and discussions. Also, this conclusion agrees with the findings of a paper called *Exploring User Responses to Entertainment and Political Videos* written by researchers at the University of Amsterdam (Möller et al., 2019).

Second, every confidence interval for the *Comedy* category that is of the form $(\bar{y}_{Comedy} - \bar{y}_j)$ is entirely positive. Also, every confidence interval that is of the form $(\bar{y}_j - \bar{y}_{Comedy})$ is entirely negative. This observation indicates that the *Comedy* video category has the highest LTV Ratio among all of the categories. From this observation, we can conclude that the *Comedy* category has received the most positive viewer sentiment among all Canadian YouTube video categories from the past six months. This conclusion intuitively makes sense because people tend to react positively towards things that make them laugh like comedy videos.

If you like to view the source code that was used for this section refer to *Appendix Section v*.

Conclusions and Discussion

i) Conclusion

The question stated at the end of the *Introduction* section was adequately answered in the *Methods* and *Results*. The following is a summary of the answer to this question:

There is a statistically significant difference in the Like-To-View ratio of the 14 Canadian YouTube video categories uploaded in the last three months of 2020 and the first three months of 2021. The *News & Politics* video category received the lowest LTV Ratio, and the *Comedy* video category received the highest LTV Ratio. Furthermore, the specific difference in LTV Ratio between every pair of video categories can be seen in *Appendix Section iv*.

ii) Substantive Takeaway

Although this project was mainly designed to answer a question that emerged out of curiosity, the reader should take away one key observation from the findings of this paper. Firstly, we concluded that YouTube videos receive and emanate different viewer sentimental reactions. As a result, we can speculate that YouTube could be able to manipulate its viewers to maximize its own best interest. As an example, YouTube could change its video recommendation algorithm to prioritize videos with high LTV Ratios in order to generate positive sentiment amongst its users. Once the positive sentiment has been created, advertisements could be shown to users knowing full-well that a positive reaction is highly probable. Furthermore, the implications of manipulating recommendation algorithms are a developing area of study, and we will direct the reader to the paper *Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects* authored by researchers at the ISR institute (Adomavicius et al., 2013)

iii) Limitations and Future Work

The first limitation of this project is its use of convenient sampling rather than simple random sampling. Due to a lack of reputable public datasets, we had to resort to using a trending video sample rather than a simple random sample of YouTube videos. The downside of this sample is that it is not representative of all Canadian YouTube videos because of the following two reasons. First, large amounts of viewers could have an underlying bias towards favouring some videos over others when it comes to driving up their trending status. Second, the YouTube

algorithm could show bias when labelling certain videos as trending. Moreover, both of these factors entail that trending videos are not entirely representative of all YouTube videos. Furthermore, future work with YouTube data should use random sampling techniques for maximum reduction of data bias.

The second limitation of this project is that the sample data doesn't entirely meet the underlying assumptions of the two-way ANOVA model. Firstly, the Normal Q-Q Plot seen below shows a large deviance away from the normal line at theoretical quantiles greater than 2. This deviation indicates a right-skew in the normal shape of the experimental errors (ϵ_{ij}), which violates the normality assumption of the ANOVA model. Secondly, the Residuals vs Fitted Plot shows an increase in residual variance as the fitted values gradually increase. This increase of variance violates the homoscedasticity assumption of the ANOVA model.

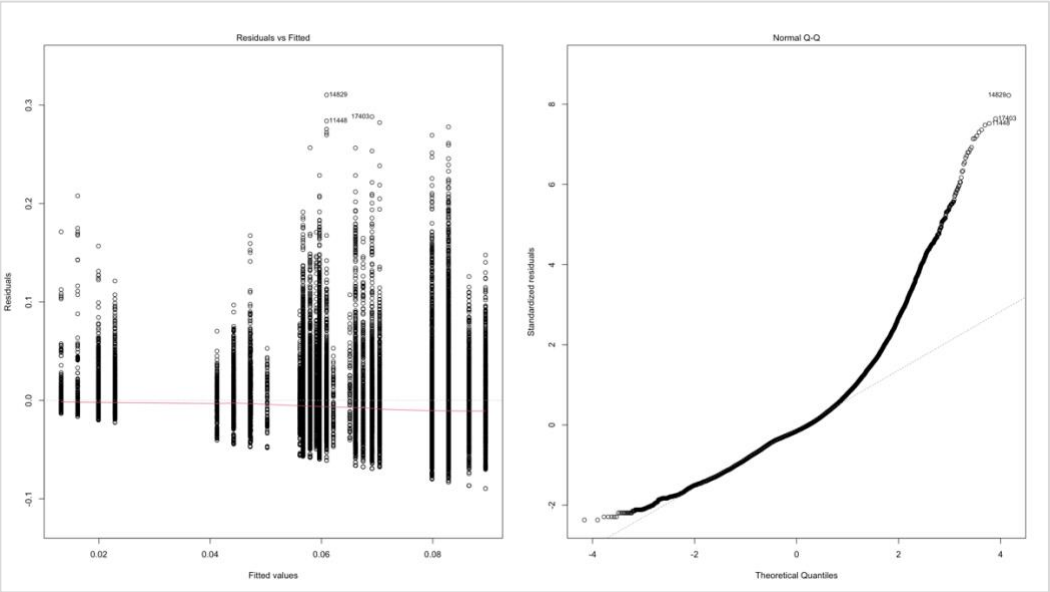


Figure 5: Two-way ANOVA Assumption Plots

Furthermore, future researchers working with YouTube-related data should consider applying transformations to their data to remove the effects of outliers and high-leverage points. These transformations can aid in meeting the underlying assumptions of the model.

References

Adomavicius, G., Bockstedt, J., Curley, S., & Zhang, J. (2013). Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research, 24*(4), 956-975. Retrieved April 16, 2021, from http://www.jstor.org/stable/24700286

Möller, A. M., Kühne, R., Baumgartner, S. E., & Peter, J. (2019). Exploring User Responses to Entertainment and Political Videos: An Automated Content Analysis of YouTube. *Social Science Computer Review, 37*(4), 510–528. https://doi.org/10.1177/0894439318779336

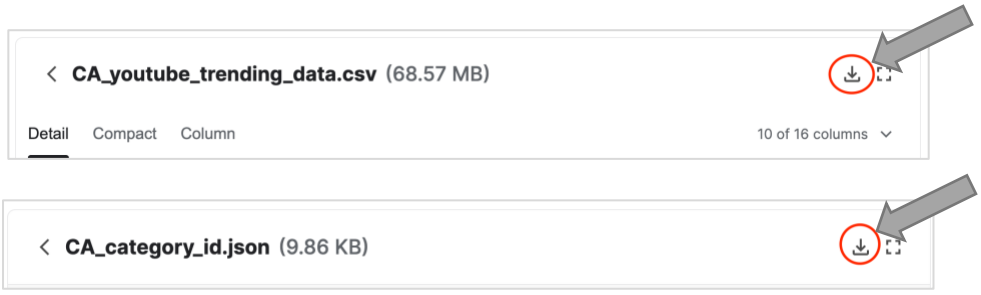
Smith, D. (2020, June 28). *Google collects a frightening amount of data about you*. Retrieved from cnet: https://www.cnet.com/how-to/google-collects-a-frightening-amount-of-data-about-you-you-can-find-and-delete-it-now/

YouTube. (2021). *YouTube in numbers*. Retrieved from Youtube About: https://www.youtube.com/intl/en-GB/about/press/#:~:text=Over%202%20billion%20logged%2Din,time%20comes%20from%20mobile%20devices.

Appendix

Section i: Acquiring the Data

1) Navigate to the following two Kaggle pages:
    a. CA_youtube_trending_data
    b. CA_category_id
2) Click on the specific download button as seen below:



Section ii: Snippet of Dataframe Used in the Analysis

	Category Name	Trending Year	LTV Ratio
1	Film & Animation	2020	0.075
3	Film & Animation	2021	0.0109
4	Film & Animation	2021	0.0552
6	Film & Animation	2020	0.00296
7	Film & Animation	2020	0.1161

Section iii: Setting up the Source Code Used for the Analysis

1) After you have downloaded the data files, place the two files in a folder called "Data" in the same directory as the R code. The directory name should look something like this:

"/STAT 403/Project /Data/"

2) Unzip the file with the following name:

"CA_youtube_trending_data.csv.zip"

3) Place the R code in the parent folder of the *"data"* folder.

4) Before running the scripts, make sure that the directory names in the *setwd* function in the R code is the same as the directory names on your computer.

5) NOTE: Since the data is constantly getting updated you will most likely not see the exact same outputs as the ones in this paper and presentation.

6) Run the script called *"Alemi_Kaveh_301309335_Rcode_1_Data_Processing.R"* to prepare, process, and save the data.

7) Run the script called *"Alemi_Kaveh_301309335_Rcode_2_Data_Analysis.R"* for the analyzing the data.

Section iv: Tukey-Kramer Results Table

	Mean Difference	Lower Bound Difference	Upper Bound Difference	P-Value
Comedy-Autos & Vehicles	0.045	0.040	0.051	0.000
Education-Autos & Vehicles	0.025	0.018	0.031	0.000
Entertainment-Autos & Vehicles	0.015	0.010	0.020	0.000
Film & Animation-Autos & Vehicles	0.026	0.020	0.032	0.000
Gaming-Autos & Vehicles	0.016	0.011	0.021	0.000
Howto & Style-Autos & Vehicles	0.015	0.009	0.021	0.000
Music-Autos & Vehicles	0.039	0.034	0.044	0.000
News & Politics-Autos & Vehicles	-0.028	-0.034	-0.022	0.000
People & Blogs-Autos & Vehicles	0.025	0.019	0.030	0.000
Pets & Animals-Autos & Vehicles	0.021	0.010	0.031	0.000
Science & Technology-Autos & Vehicles	0.003	-0.003	0.009	0.872
Sports-Autos & Vehicles	-0.022	-0.027	-0.017	0.000
Travel & Events-Autos & Vehicles	0.006	-0.004	0.015	0.821
Education-Comedy	-0.020	-0.026	-0.015	0.000
Entertainment-Comedy	-0.030	-0.033	-0.027	0.000
Film & Animation-Comedy	-0.019	-0.024	-0.014	0.000
Gaming-Comedy	-0.029	-0.032	-0.026	0.000
Howto & Style-Comedy	-0.030	-0.035	-0.026	0.000
Music-Comedy	-0.007	-0.010	-0.003	0.000
News & Politics-Comedy	-0.074	-0.078	-0.069	0.000
People & Blogs-Comedy	-0.021	-0.024	-0.017	0.000
Pets & Animals-Comedy	-0.025	-0.035	-0.015	0.000
Science & Technology-Comedy	-0.042	-0.046	-0.038	0.000
Sports-Comedy	-0.067	-0.070	-0.064	0.000
Travel & Events-Comedy	-0.040	-0.049	-0.031	0.000
Entertainment-Education	-0.010	-0.014	-0.005	0.000
Film & Animation-Education	0.001	-0.005	0.007	1.000
Gaming-Education	-0.009	-0.014	-0.004	0.000
Howto & Style-Education	-0.010	-0.016	-0.004	0.000
Music-Education	0.014	0.009	0.019	0.000
News & Politics-Education	-0.053	-0.059	-0.047	0.000
People & Blogs-Education	-0.000	-0.005	0.005	1.000
Pets & Animals-Education	-0.004	-0.015	0.007	0.989
Science & Technology-Education	-0.022	-0.027	-0.016	0.000
Sports-Education	-0.046	-0.051	-0.041	0.000
Travel & Events-Education	-0.019	-0.029	-0.010	0.000
Film & Animation-Entertainment	0.011	0.007	0.015	0.000
Gaming-Entertainment	0.001	-0.002	0.004	0.994
Howto & Style-Entertainment	-0.000	-0.004	0.003	1.000
Music-Entertainment	0.023	0.021	0.026	0.000
News & Politics-Entertainment	-0.044	-0.048	-0.040	0.000
People & Blogs-Entertainment	0.009	0.006	0.012	0.000
Pets & Animals-Entertainment	0.005	-0.005	0.015	0.898
Science & Technology-Entertainment	-0.012	-0.016	-0.009	0.000
Sports-Entertainment	-0.037	-0.039	-0.034	0.000
Travel & Events-Entertainment	-0.010	-0.018	-0.001	0.013
Gaming-Film & Animation	-0.010	-0.014	-0.006	0.000
Howto & Style-Film & Animation	-0.011	-0.016	-0.006	0.000
Music-Film & Animation	0.012	0.008	0.017	0.000
News & Politics-Film & Animation	-0.055	-0.060	-0.049	0.000
People & Blogs-Film & Animation	-0.002	-0.006	0.003	0.997
Pets & Animals-Film & Animation	-0.006	-0.016	0.005	0.881
Science & Technology-Film & Animation	-0.023	-0.028	-0.018	0.000
Sports-Film & Animation	-0.048	-0.052	-0.043	0.000
Travel & Events-Film & Animation	-0.021	-0.030	-0.011	0.000

Howto & Style-Gaming	-0.001	-0.005	0.003	0.997
Music-Gaming	0.022	0.020	0.025	0.000
News & Politics-Gaming	-0.045	-0.049	-0.040	0.000
People & Blogs-Gaming	0.008	0.005	0.012	0.000
Pets & Animals-Gaming	0.004	-0.006	0.014	0.981
Science & Technology-Gaming	-0.013	-0.017	-0.009	0.000
Sports-Gaming	-0.038	-0.041	-0.035	0.000
Travel & Events-Gaming	-0.011	-0.020	-0.002	0.004
Music-Howto & Style	0.024	0.020	0.028	0.000
News & Politics-Howto & Style	-0.043	-0.048	-0.038	0.000
People & Blogs-Howto & Style	0.010	0.006	0.014	0.000
Pets & Animals-Howto & Style	0.006	-0.005	0.016	0.876
Science & Technology-Howto & Style	-0.012	-0.016	-0.007	0.000
Sports-Howto & Style	-0.036	-0.040	-0.033	0.000
Travel & Events-Howto & Style	-0.009	-0.019	-0.000	0.041
News & Politics-Music	-0.067	-0.071	-0.063	0.000
People & Blogs-Music	-0.014	-0.017	-0.011	0.000
Pets & Animals-Music	-0.018	-0.028	-0.008	0.000
Science & Technology-Music	-0.036	-0.039	-0.032	0.000
Sports-Music	-0.060	-0.063	-0.058	0.000
Travel & Events-Music	-0.033	-0.042	-0.024	0.000
People & Blogs-News & Politics	0.053	0.049	0.057	0.000
Pets & Animals-News & Politics	0.049	0.038	0.059	0.000
Science & Technology-News & Politics	0.031	0.027	0.036	0.000
Sports-News & Politics	0.007	0.003	0.011	0.000
Travel & Events-News & Politics	0.034	0.025	0.043	0.000
Pets & Animals-People & Blogs	-0.004	-0.014	0.006	0.987
Science & Technology-People & Blogs	-0.022	-0.026	-0.018	0.000
Sports-People & Blogs	-0.046	-0.049	-0.043	0.000
Travel & Events-People & Blogs	-0.019	-0.028	-0.010	0.000
Science & Technology-Pets & Animals	-0.017	-0.028	-0.007	0.000
Sports-Pets & Animals	-0.042	-0.052	-0.032	0.000
Travel & Events-Pets & Animals	-0.015	-0.028	-0.002	0.010
Sports-Science & Technology	-0.025	-0.028	-0.021	0.000
Travel & Events-Science & Technology	0.003	-0.007	0.012	1.000
Travel & Events-Sports	0.027	0.018	0.036	0.000

Section v: Source Code Used for the Analysis

```r
# Figure 2
LTV_boxplot <- ggplot(data = video_df, aes(x = categoryName, y = like_view_ratio)) +
  ggtitle("Like-To-View Ratio Per Trending Video Categories") +
  xlab("Trending Video Categories") +
  ylab("Like-To-View (LTV) Ratio") +
  theme(axis.text.x = element_text(size = 13, angle = 90)) +
  geom_boxplot(alpha=0.2)

LTV_boxplot

# Figure 3
LTV_Year_boxplot <- ggplot(data = video_df,
                           aes(x = categoryName, y = like_view_ratio, fill = trending_year)) +
  ggtitle("Like-To-View Ratio Per Trending Video Categories and Year") +
  xlab("Trending Video Categories") +
  ylab("Like-To-View (LTV) Ratio") +
  geom_boxplot(alpha=0.2) +
  theme(axis.text.x = element_text(size = 13, angle = 90))

LTV_Year_boxplot

# 95% confidence interval of  LTV ratio for 2020 and 2021 and upload year
summary_stats <- plyr::ddply(video_df, "trending_year", function(x) {

  oneci <- t.test(x$like_view_ratio)$conf.int
  names(oneci) <- c("lcl","ucl")
  return(oneci)

})
summary_stats$mean = rowMeans(summary_stats[ , c(2,3)], na.rm=TRUE)

# Figure 4
ltv_ci_plot <- ggplot(data = summary_stats, aes(x = trending_year, y = mean)) +
  geom_point() +
  geom_line(group = 1)+
  xlab("Trending Year") +
  ylab("Like-To-View (LTV) Ratio") +
  ggtitle("95% Confidence Interval of Like-To-View Ratio Per Year")+
  geom_errorbar(aes(ymin = lcl, ymax = ucl), width=.02,color="black") +
  theme(axis.text.x = element_text(size = 13, angle = 90))

ltv_ci_plot

# ANOVA
LTV_cat_anova_test = aov(like_view_ratio ~ categoryName + trending_year, data = video_df)
summary(LTV_cat_anova_test)


# Tukey Kramer Test
posthoc <- TukeyHSD(x=LTV_cat_anova_test, conf.level=0.95)
(posthoc)
```