CMPT 353 Final Project: Trending YouTube Categories and the Pandemic

CMPT 353 - D100

Kaveh Alemi and Mike Thai

December 11, 2020

Project Introduction

In this paper, we will investigate and explore trending YouTube video categories in Canada, Great Britain, and the USA during the years of 2018 and 2020. We have designed our analysis such that the 2018 data represents a timeframe not experiencing a pandemic, and conversely, the 2020 data representing a timeframe undergoing an active pandemic. In addition, we chose these three geographical regions in order to limit the data to the Western World and eliminate any cultural confounding variables. Furthermore, our analysis will aim to answer the following two questions:

1) Has the frequency and view count of trending YouTube categories significantly changed in 2018 versus 2020 in Canada, Great Britain, and the USA?

2) Has the ratio of likes versus views of trending YouTube categories changed significantly in 2018 versus 2020 in Canada, Great Britain, and the USA?

Defining YouTube and Trending Categories

YouTube is an online video-sharing website, where users are able to upload videos, comment on videos, and like or dislike videos. Although the platform is much more robust than this short definition, we will skip explaining the semantics since almost everyone on the planet is familiar with this platform. Moreover, YouTube experiences massive amounts of video uploads every single day, and in order to showcase the most popular videos currently being watched, YouTube created a Trending Videos Page. YouTube has been very secretive about how videos become trending and how they are defined as popular. Many have speculated that each geographical region receives their own list of trending videos on a daily basis. It is unknown how this list is compiled but many believe it is generated by Machine Learning algorithms that evaluate parameters such as video views, number of comments, and the ratio of likes to dislikes. Since YouTube has not directly explained how videos become trending, we will make the following assumptions about these videos:

- Users in the same country or geographical location receive the same list of trending videos.

- Once a video becomes trending, all of the respective video parameters are recorded in the dataset exactly once.

- There is no fixed count of videos that must become trending on a given day.

In addition, every trending video is placed in one of the following 15 video categories:

1. Pets & Animals
2. Travel & Events
3. Autos & Vehicles
4. Education
5. Science & Technology
6. Gaming
7. How to & Style
8. Film & Animation
9. Sports
10. Comedy
11. Music
12. News & Politics
13. People & Blogs
14. Entertainment
15. Nonprofits & Activism (USA only)

Explaining The Data

We will be using two Kaggle datasets where each contains data pertaining to trending YouTube videos from 2018 and 2020. In our initial analysis of the datasets, we noticed that there were some discrepancies between the data from each year. The 2018 dataset only contained data of trending videos in the months of January to April of 2018. The 2020 dataset only contained data of trending videos in the months of August to October of 2020. It is apparent that both datasets cover the same length of time, four months each, but don't cover the same exact months in their respective years. Although seasonality could definitely play a factor in which videos become trending, we will largely ignore this confounding variable and treat each dataset as a sample representing their respective year.

(i) Data Cleaning:

Both datasets have the same relational structure, where each row corresponds to an instance of a trending video and columns correspond to video features. The features that we will be focusing on are Video Views, Video Likes, and Video Categories. Moreover, both datasets are fairly clean, but there were some values that were from the wrong year which had to be filtered out. The 2018 data had two additional trending categories, *Movies* and *Shows*, that were not in the 2020 data. We filtered these two categories out of the 2020 dataset in order to keep each dataset consistent. After all of the data cleaning, the 2018 dataset had approximately 22,000 rows and the 2020 dataset had approximately 17,000 rows, with both containing the following features: Video Category, Video Views, and Video Likes.

(ii) Feature Engineering:

In order to answer the second question mentioned in the *Project Introduction*, we need to create a new feature called Like-To-View Ratio (L-T-V) which is equivalent to Video Likes divided by Video Views.

$$L - T - V \ Ratio = \frac{Video\ Likes}{Video\ Views}$$

We created this feature in both of the 2018 and 2020 datasets.

Hypothesis

We have based our hypotheses directly on the two questions proposed in the *Project Introduction* and plan to investigate each thoroughly. For question 1, our analysis will aim to disprove the following null hypothesis:

- H_0 : The frequency and view count distribution of trending YouTube categories is the same in 2018 versus 2020 in Canada, Great Britain, and the USA.

Similarly, for question 2, our analysis will aim to disprove the following null hypothesis:

- H_0 : The *L-V-T Ratio* of trending YouTube categories is the same in 2018 versus 2020 in Canada, Great Britain, and the USA.

* Please note that the significance of all statistical tests will be assessed on an alpha level (α) of 0.05.

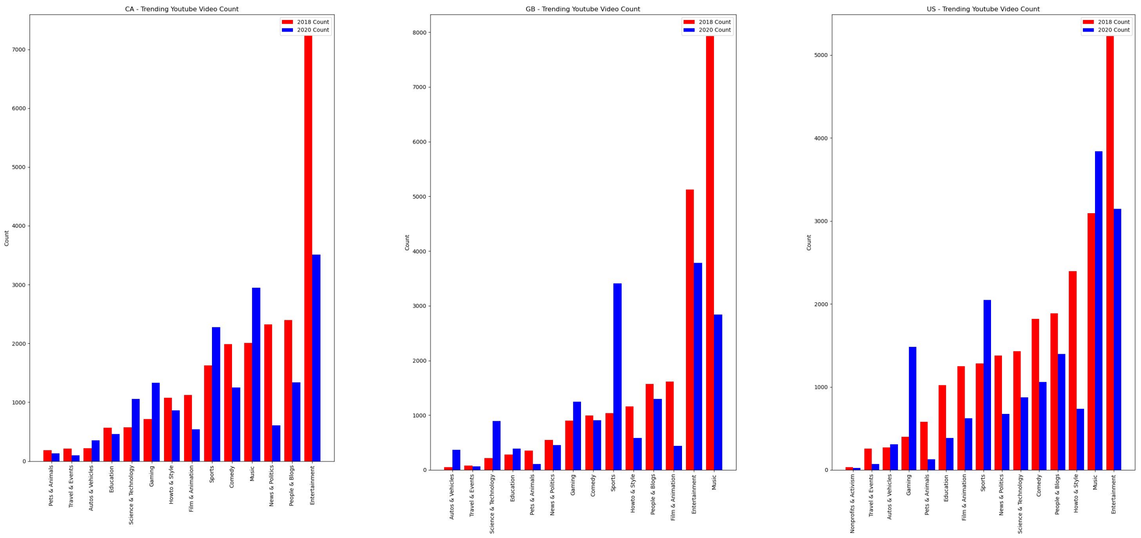
Analysis

(1) Has the Frequency and View Count of Trending YouTube Categories Significantly Changed in 2018 Versus 2020 in Canada, Great Britain, and the USA?

(i) A Look at the Data

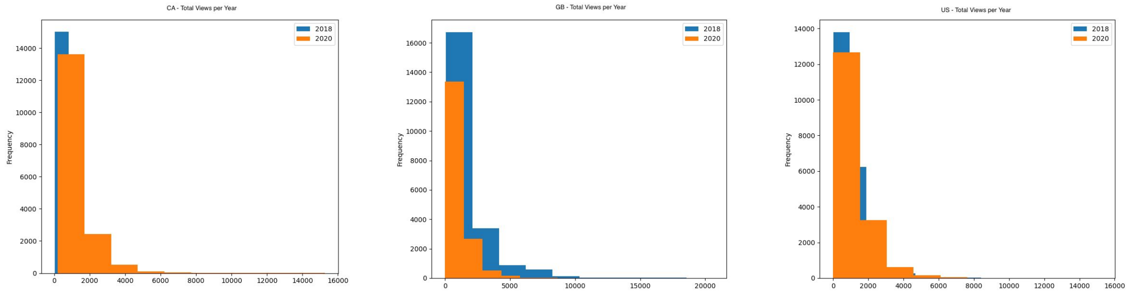
Before we perform any statistical tests to answer the question at hand, we believe it is worthwhile to create some plots to help us paint a picture of the data. First, let's look at how frequently a video category became trending in 2018 versus 2020 for each of the three geographical regions. We grouped by year and category, and then counted every occurrence of a video for each region and derived the following bar plots.

Figure 1. Barplot of Trending Category Frequencies per Region



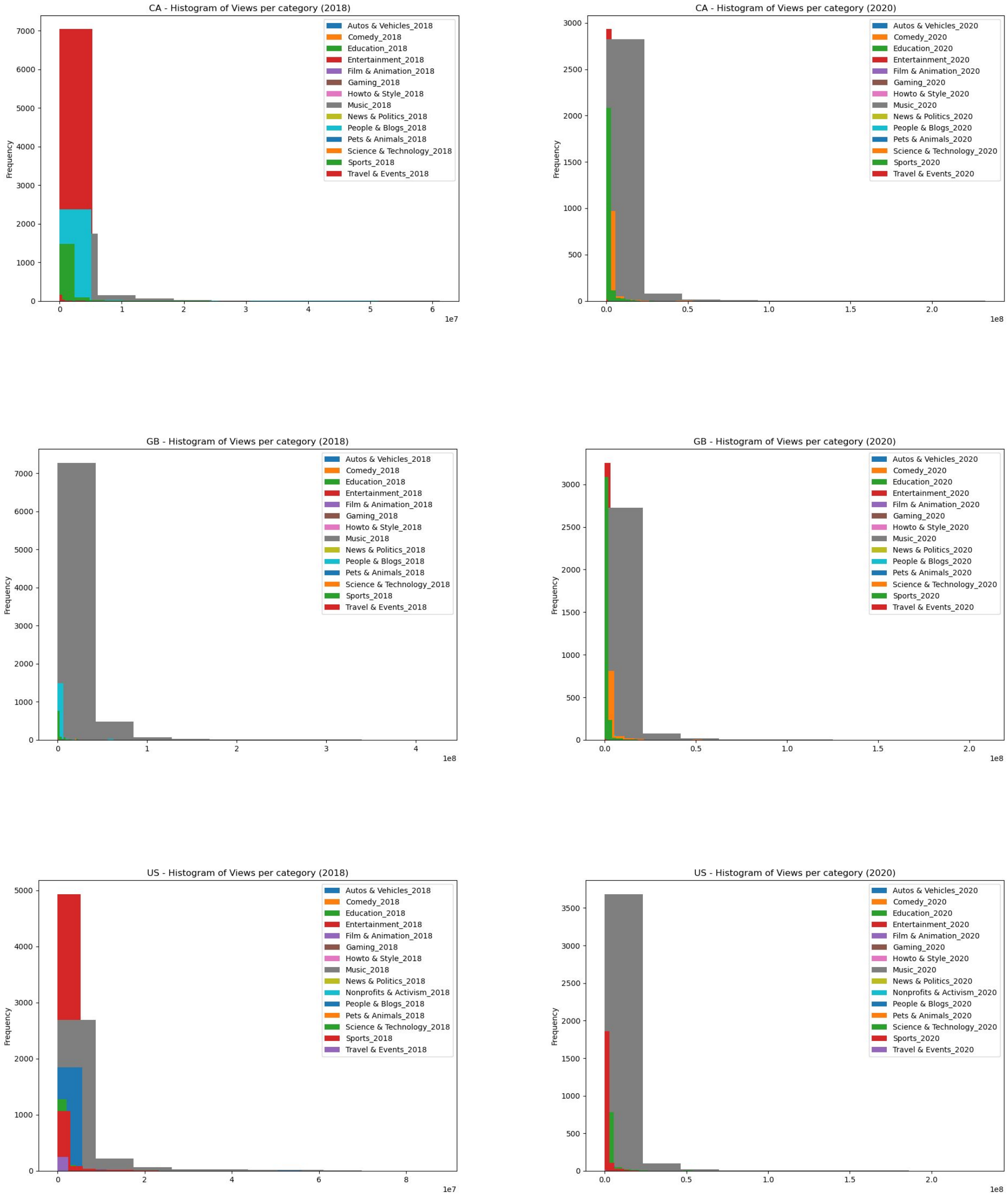
In Canada, Entertainment has been the most frequent category in both of 2018 and 2020 but is much more frequent in 2018. Both In Great Britain and the USA, Music was most frequent in 2018 and Entertainment was most frequent in 2020. Also, it appears that in all three regions Sports became much more frequent in 2020, while News and Politics decreased in trending frequency. Moreover, let's now consider the second part of the question and explore the distribution of total trending views across all categories for each region. We derived the following plots.

Figure 2. Histogram of Total Trending Views per Year for Each Region



From the above plots, we can see that none of the three regions have a normal distribution for total trending views for either 2018 or 2020. Let's now be more specific and consider the distribution of trending views for each category of each region.

Figure 3. Histogram of Trending Views per Category for Each Region
(Left: 2018, Right: 2020)



It certainly appears that the distribution of trending views for each category is also not normal for either 2018 or 2020 for all three regions. Furthermore, we should take note of this non-normality since in the next section we will conduct parametric statistical analysis.

(ii) Statistical Analysis:

We begin our statistical analysis by first looking into the independence of the frequency of trending categories between 2018 and 2020. For each region, we will create a contingency table that has 2018 and 2020 as columns and has video categories as rows. Next, we will apply a Chi-Squared test to each table to assess the independence of the frequencies of each year. We obtained the following results:

Table 1. Chi-Squared Test

Country	**P-Value**	**Chi Test Statistic**
Canada	< 0.0000001	2993.27
Great Britain	< 0.0000001	5047.88
USA	< 0.0000001	3098.69

The Chi-Squared test was significant for all three regions given that the p-values were all less than our alpha of 0.05. This tells us that year is not independent of the frequency of trending categories. An alternative conclusion can be that 2018 and 2020 have different frequencies of trending YouTube categories for all three of Canada, Great Britain, and the USA.

Let's now investigate potential differences between total trending views between 2018 and 2020 for all three regions. In the prior section, *A Look at the Data*, we observed that none of the three regions had normally distributed total trending views for either of 2018 or 2020. As a result, we will have to apply the non-parametric Mann-Whitney U test to decide if the distribution of total trending views is larger or smaller between 2018 and 2020. We obtained the following results.

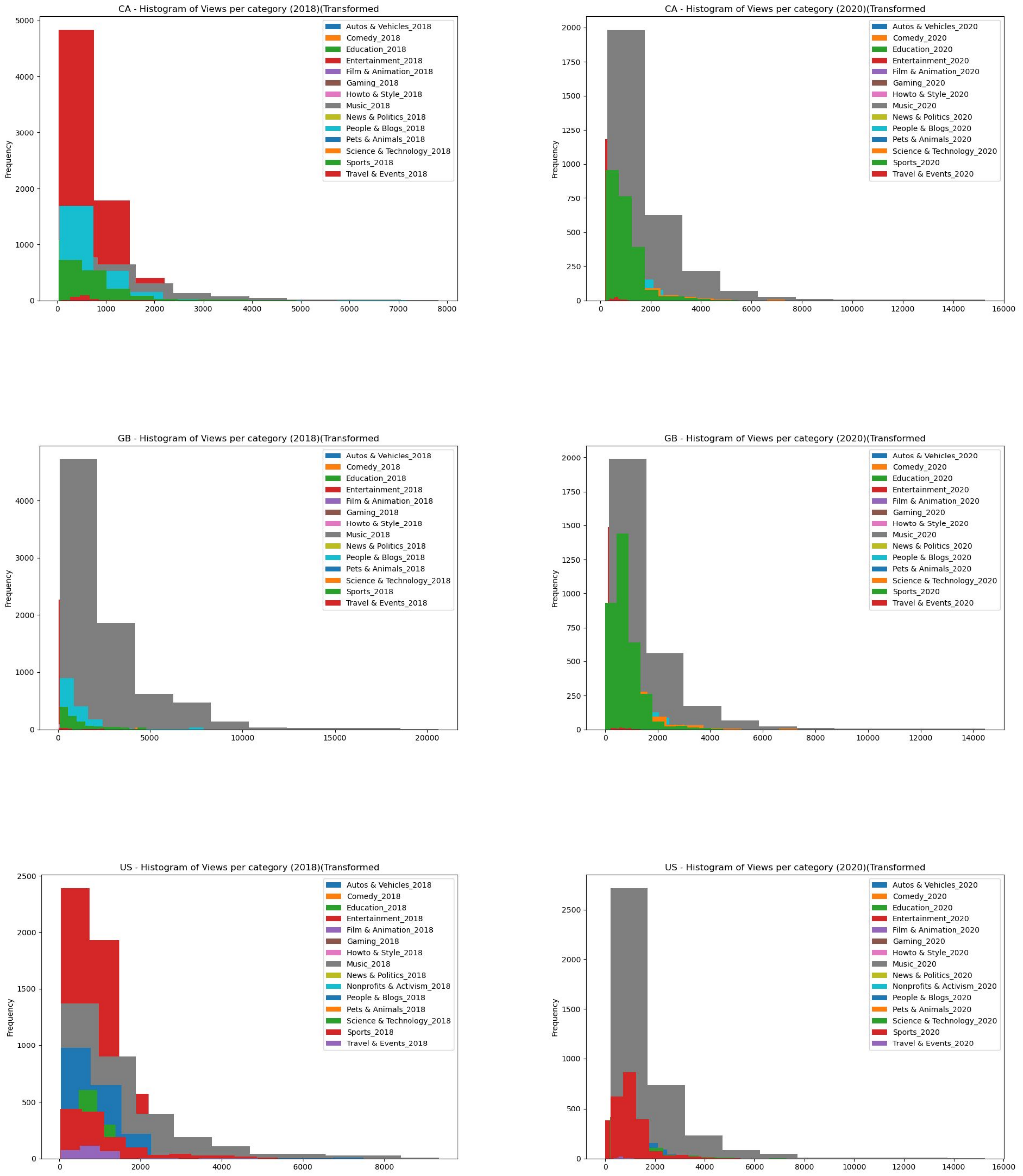
Table 2. Mann-Whitney U Test

Country	**P-Value**	**U Statistic**
Canada	< 0.0000001	111188337.00
Great Britain	< 0.0000001	214731350.50
USA	< 0.0000001	142397210.50

The U test showed significance for all three regions given that the p-values were all less than the alpha-level of 0.05. We can conclude that the total trending views are not equal between the years of 2018 and 2020 for Canada, Great Britain, and the USA.

Let's now compare the views of each trending category in 2018 and 2020. For each region, we will group by year and category in order to obtain the mean view count of each group. Ideally, we want to first apply a One-Way ANOVA to see if there is a difference in these group means. If the ANOVA shows significance, we would then like to perform Post Hoc analysis to better quantify these differences. Moreover, before conducting this analysis, we should acknowledge that none of these groups meet the normality condition required for all of these tests since none passed Pearson's Normality Test. As a result, we should make an effort to normalize the data by applying transformations to it. After applying three sets of transformations, the square-root function yielded the best results but still could not pass Pearson's Normality Test. The following plots are the distribution of views per category with the square root function applied to the data.

Figure 4. Histogram of Transformed Views per Category for Each Region (Left: 2018, Right: 2020)



Although not entirely Normal, we will still apply a One-Way ANOVA to the untransformed grouped means and will leave the reader to assess how confident they should be in our results. We obtained the following results.

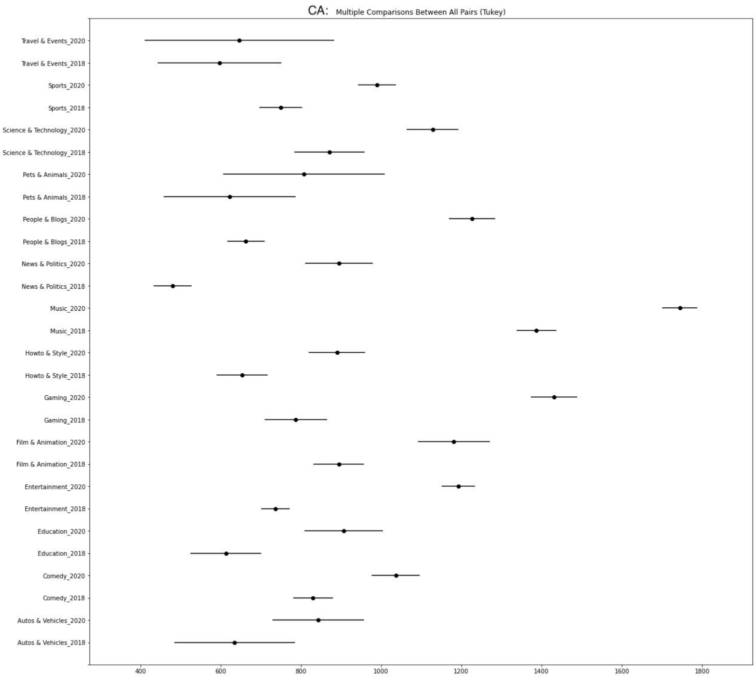
Table 3. One-Way ANOVA Test

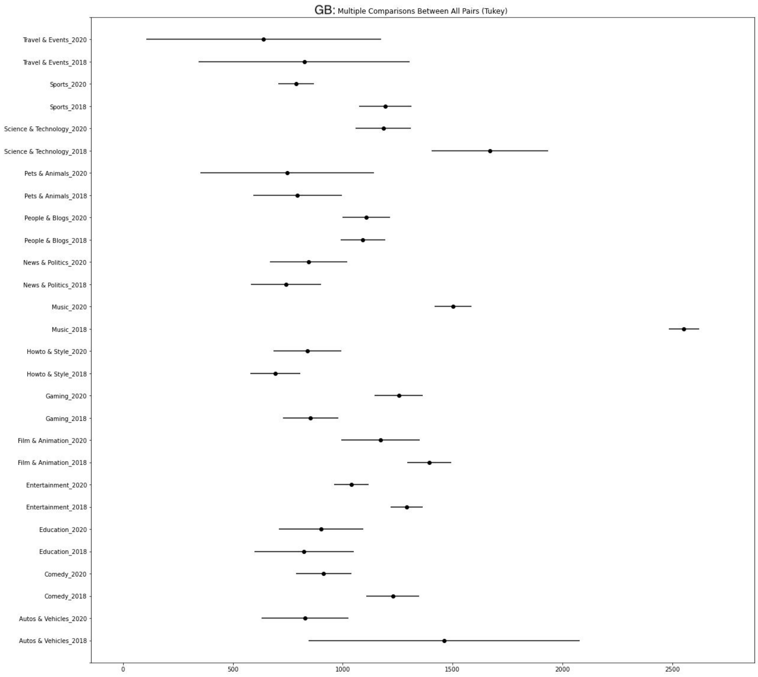
Country	**P-Value**	**F Statistic**
Canada	< 0.0000001	140.60
Great Britain	< 0.0000001	277.10
USA	< 0.0000001	258.10

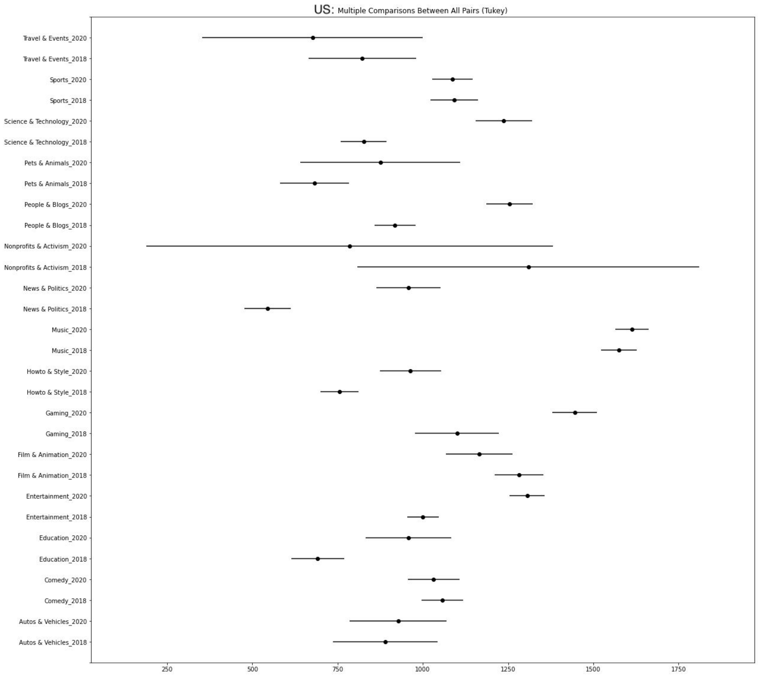
The ANOVA showed significance for all three regions given that the obtained p-values were all less than our alpha level of 0.05. This indicates that in each region there is a difference between the group means of total views per category across all of 2018 and 2020.

Next, we will begin the Post Hoc Analysis by applying Tukey's HSD to the group means to better understand the individual differences existing between each of the groups. The following are plots obtained from applying Tukey's HSD to the trending categories of each region:

Figure 5. Plots of Tukey's HSD for Views per Category of Each Region




GB: Multiple Comparisons Between All Pairs (Tukey)
Travel & Events_2020
Travel & Events_2018
Sports_2020
Sports_2018
Science & Technology_2020
Science & Technology_2018
Pets & Animals_2020
Pets & Animals_2018
People & Blogs_2020
People & Blogs_2018
News & Politics_2020
News & Politics_2018
Music_2020
Music_2018
Howto & Style_2020
Howto & Style_2018
Gaming_2020
Gaming_2018
Film & Animation_2020
Film & Animation_2018
Entertainment_2020
Entertainment_2018
Education_2020
Education_2018
Comedy_2020
Comedy_2018
Autos & Vehicles_2020
Autos & Vehicles_2018
0
500
1000
1500
2000
2500


US: Multiple Comparisons Between All Pairs (Tukey)
Travel & Events_2020
Travel & Events_2018
Sports_2020
Sports_2018
Science & Technology_2020
Science & Technology_2018
Pets & Animals_2020
Pets & Animals_2018
People & Blogs_2020
People & Blogs_2018
Nonprofits & Activism_2020
Nonprofits & Activism_2018
News & Politics_2020
News & Politics_2018
Music_2020
Music_2018
Howto & Style_2020
Howto & Style_2018
Gaming_2020
Gaming_2018
Film & Animation_2020
Film & Animation_2018
Entertainment_2020
Entertainment_2018
Education_2020
Education_2018
Comedy_2020
Comedy_2018
Autos & Vehicles_2020
Autos & Vehicles_2018
250
500
750
1000
1250
1500
1750

For Canada, every single category except Travel & Events, Pets & Animals, and Auto & Vehicles had a statistically significant increase in viewership in 2020. More specifically, News & Politics, People & Blogs, and Gaming saw the largest increase in viewership in 2020.
For Great Britain, Gaming is the only category that saw some statistically significant increase in viewership in 2020. Categories like Sports, Science & Technology, Music, Entertainment, and Comedy all saw a statistically significant decrease in viewership in 2020.
Lastly, for the USA, Science & Technology, People & Blogs, News & Politics, Gaming, and Entertainment all saw a large statistically significant increase in viewership in 2020. Categories such as How To & Styles and Education saw only a minimal statistically significant increase in viewership in 2020.

(iii) Conclusions

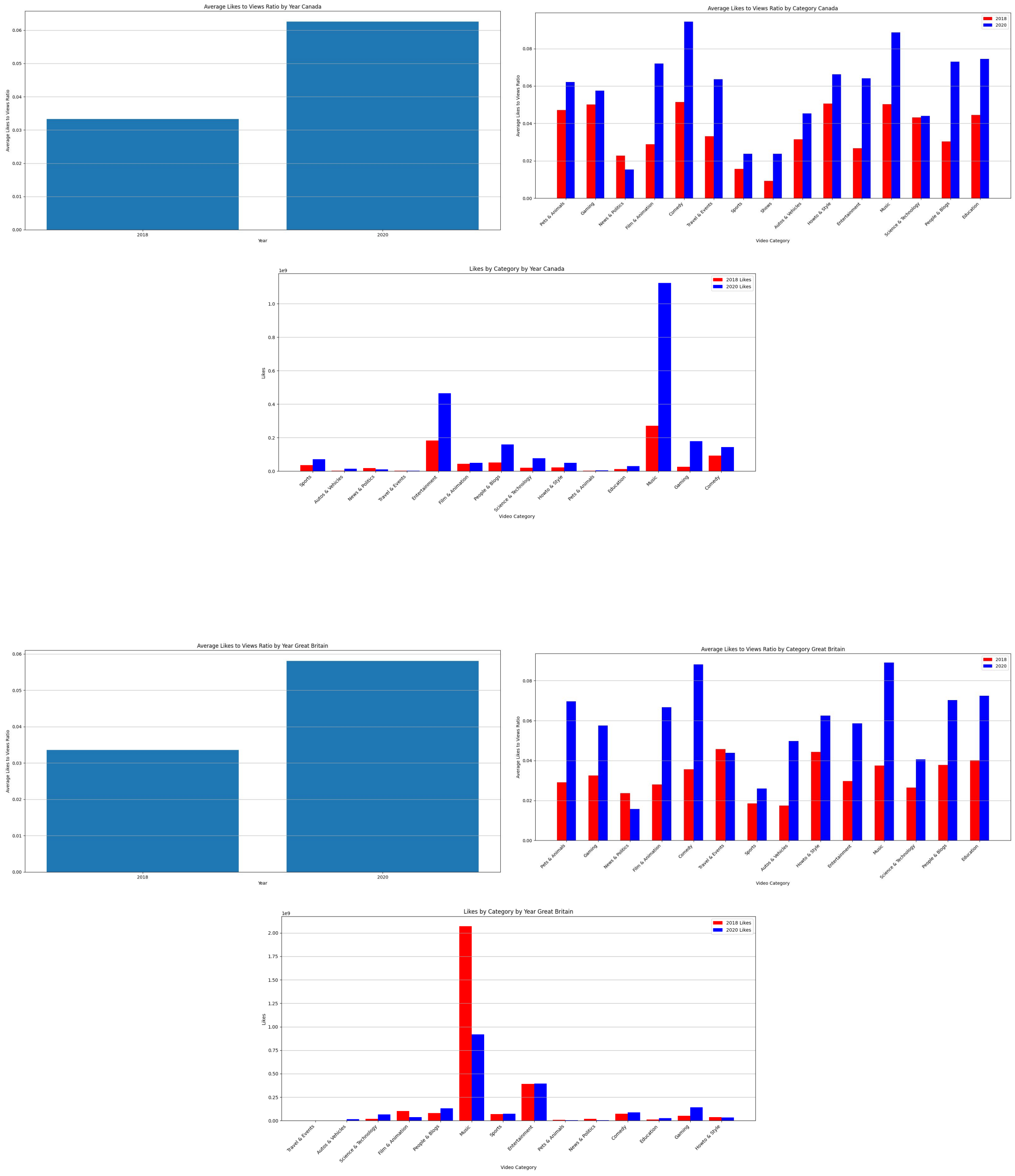
From our analysis, we can confidently conclude that 2018 and 2020 had different frequencies of trending YouTube categories for Canada, Great Britain, and the USA. One potential reason for why video frequencies changed in 2020 could be the presence of the COVID-19 pandemic. Since this is an analysis and not an experiment we can’t conclude a causal relationship between these two variables but instead will acknowledge that a potential correlation could exist between them.
From the non-parametric tests, we can conclude that the total trending views have changed in 2020 in comparison to 2018, for all three of the regions. More specific to categories, Canada and the USA saw an increase in viewership in most categories and Great Britain saw a decrease in viewership in most categories. The Post Hoc analysis showed us that Gaming is the only category that saw a statistically significant increase in viewership in all three of the regions. This increase could very well be correlated with the pandemic, where people tuned into watching Gaming videos to feel a social connection with the person playing the game. One more notable observation is that the 2020 increase in viewership for News & Politics in the USA and Canada could very well be correlated with the fact that the US had a planned federal election in November of 2020. Those affected closest by this election, the USA and Canada, could have purposefully increased their viewership of News & Politics related YouTube videos to be more aware of the election.

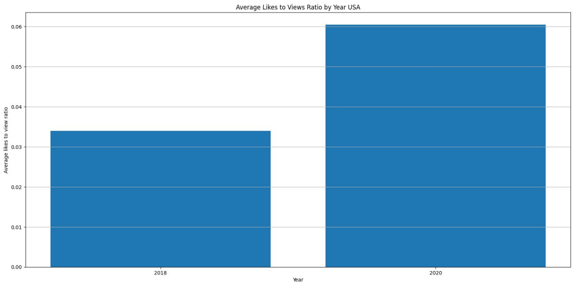
(2) Has the Ratio of Likes Versus Views of Trending YouTube Categories Changed Significantly in 2018 Versus 2020 in Canada, Great Britain, and the USA?

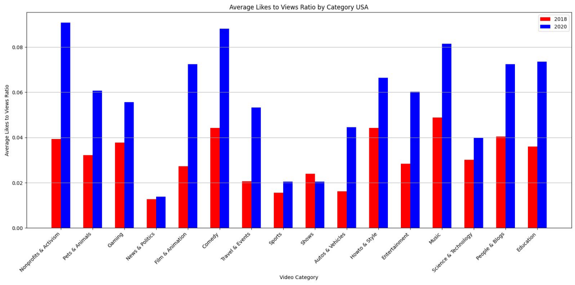
(i) A Look at The Data

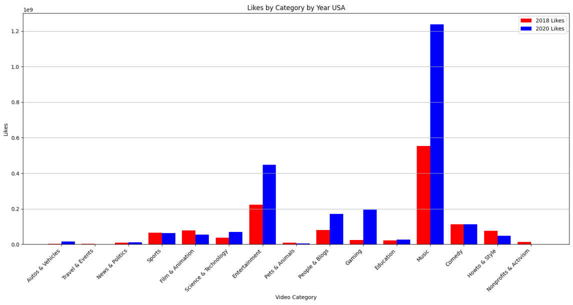
Again, before conducting any statistical tests, let's first visualize the data. First, we created plots to better understand each region's average L-T-V ratio by year, average L-T-V ratio by category, and the number of likes a trending video received by category.

Figure 6. Average LTV Ratio by Year and Category, and Total Likes Per Category (Top: Canada, Middle: Great Britain, Bottom: USA)









The first plot (average L-T-V per year) shows that the average L-T-V ratios have increased in 2020 for all three of the regions. Moreover, the following table highlights the exact numerical values for each region's L-T-V ratio for 2018 and 2020.

Table 4. Average L-T-V Ratio by Country

Country	**Average L-T-V Ratio of 2018**	**Average L-T-V Ratio of 2020**
Canada	0.033	0.062
Great Britain	0.033	0.058
USA	0.033	0.060

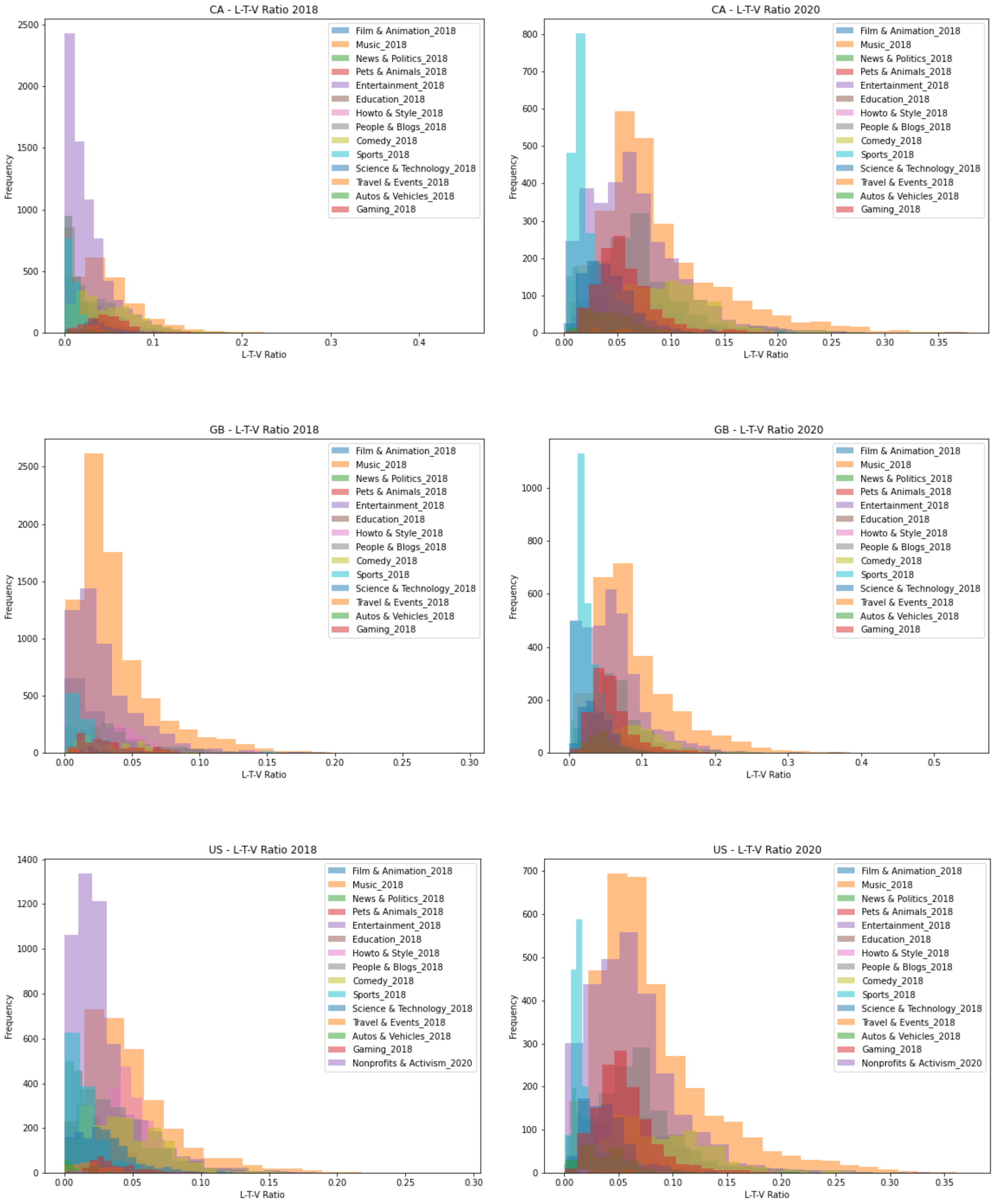
From the second plot (average L-T-V ratio by category), we can see that the L-T-V ratios appear to be different for each category of 2018 and 2020 for each region. In Canada, Film & Animation, Comedy, Education, People & Blogs, Entertainment, and Music all had an L-T-V ratio that approximately doubled from 2018 to 2020. Interestingly, News & Politics was the only category that saw a decreasing in L-T-V ratio in 2020. In Great Britain, we see a similar trend but with more categories whose L-T-V ratio approximately doubled from 2018 to 2020. Again, News & Politics was the only category with an L-T-V ratio that decreased. In the USA, we see an even stronger increasing trend, where every category saw an increase in L-T-V ratio from 2018 to 2020.

From the third graph (likes per category), we can observe that categories such as Music and Entertainment have received significantly more likes in comparison to all other categories. In Canada and the USA, these two categories saw significant increases in likes in 2020, while in Great Britain, likes for Music dropped significantly and stayed relatively the same for Entertainment. Now that we have gained more insight about our data, we will perform statistical analysis on the L-T-C ratio and answer the question at hand.

(ii) Statistical Analysis

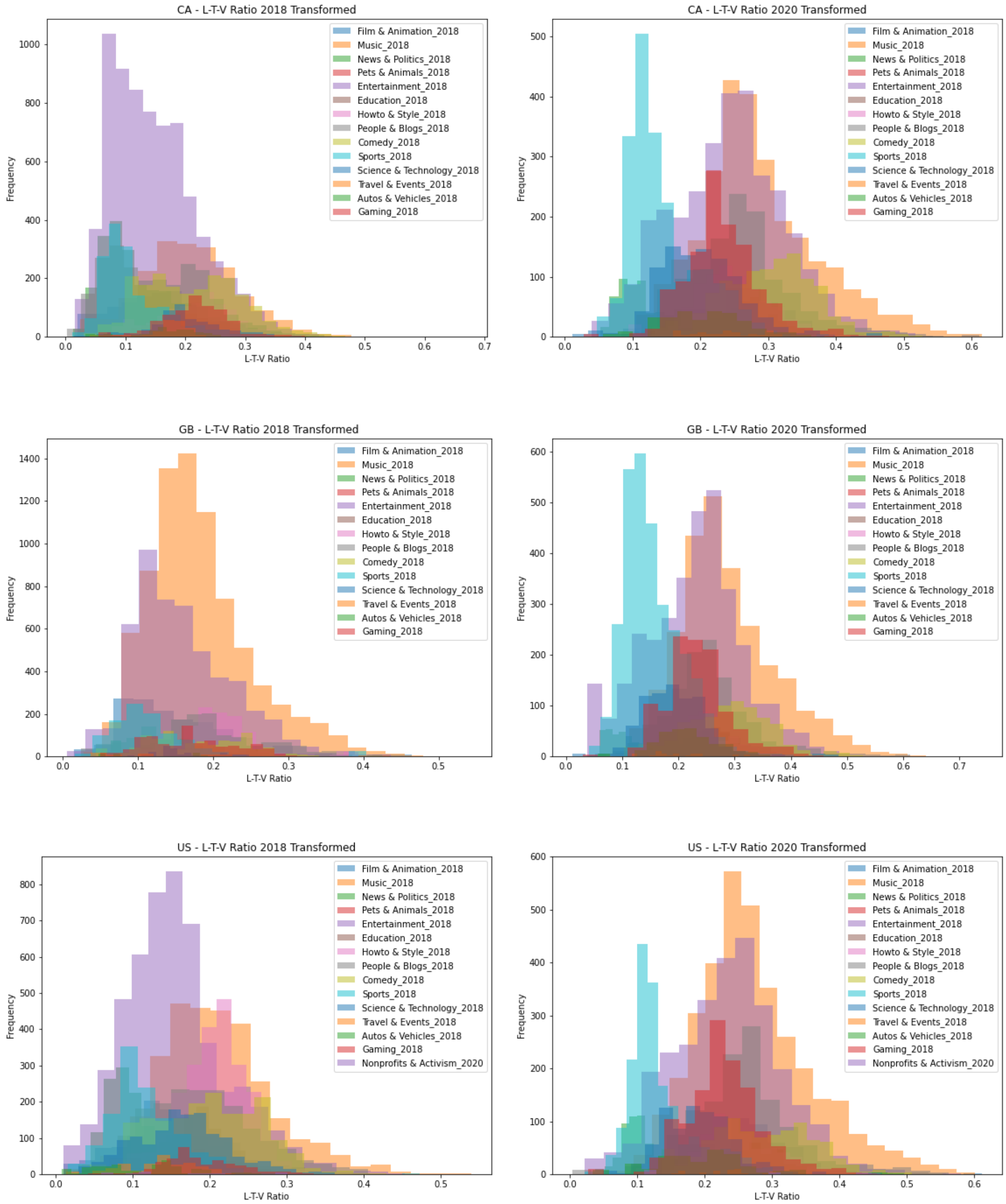
To answer the question at hand, we have to determine if the mean L-T-V ratio of any category differs in 2018 versus 2020 in all three of the regions. Similar to question 1, we want to apply a One-Way ANOVA to the L-T-V ratio of each category and year for each region. Before doing so, it is necessary to analyze the data to see if it meets the normality requirement of the ANOVA test.

Figure 7. Histogram of L-T-V Ratio per Category and Year for Each Region (Left: 2018, Right: 2020)



From *Figure 7*, it becomes evident that the L-T-V ratio of each category is right-skewed. This data fails Pearson's Normality Test which means it would be inappropriate to perform an ANOVA test. After applying transformations to the data, more specifically, applying the square-root function to each L-T-V ratio value, we obtained the following results.

Figure 8. Histogram of Transformed L-T-V Ratio per Category and Year for Each Region (Left: 2018, Right: 2020)



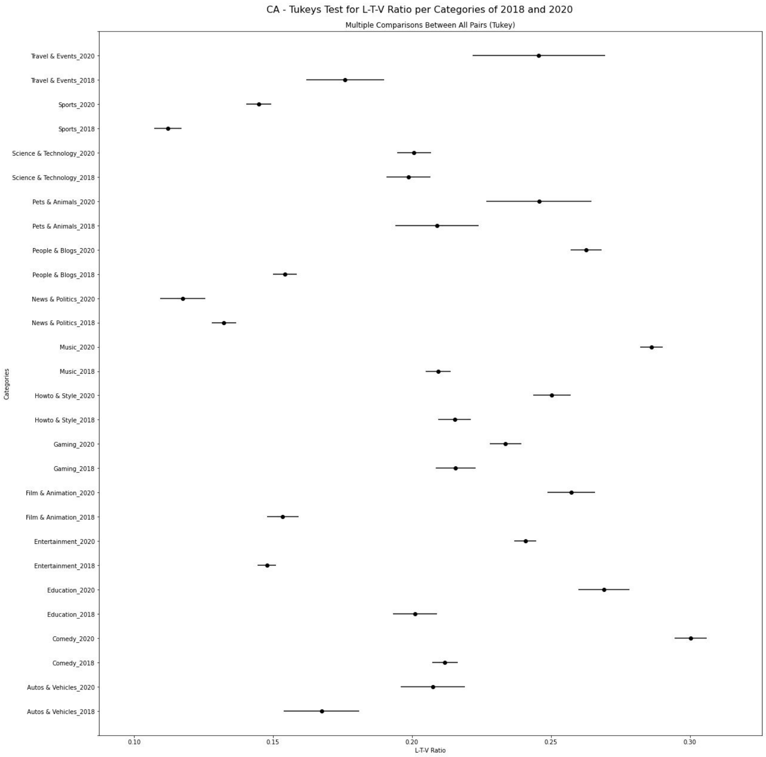
We can see that the transformation made the data much closer to a normal-distribution, hence why it was able to pass Pearsons Normality Test. As a result, we applied a One-Way ANOVA and obtained the following results.

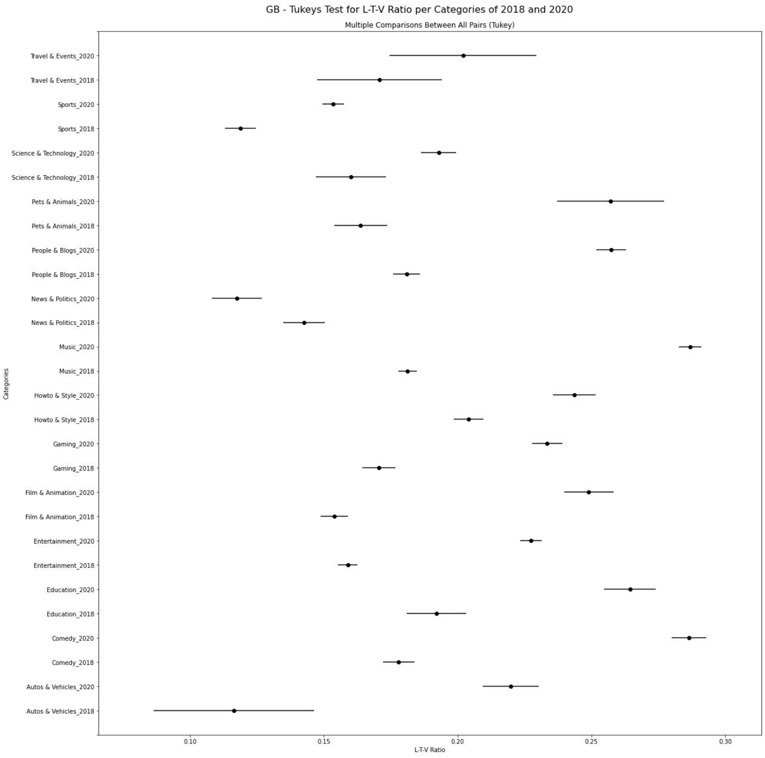
Table 5. One-Way ANOVA Test

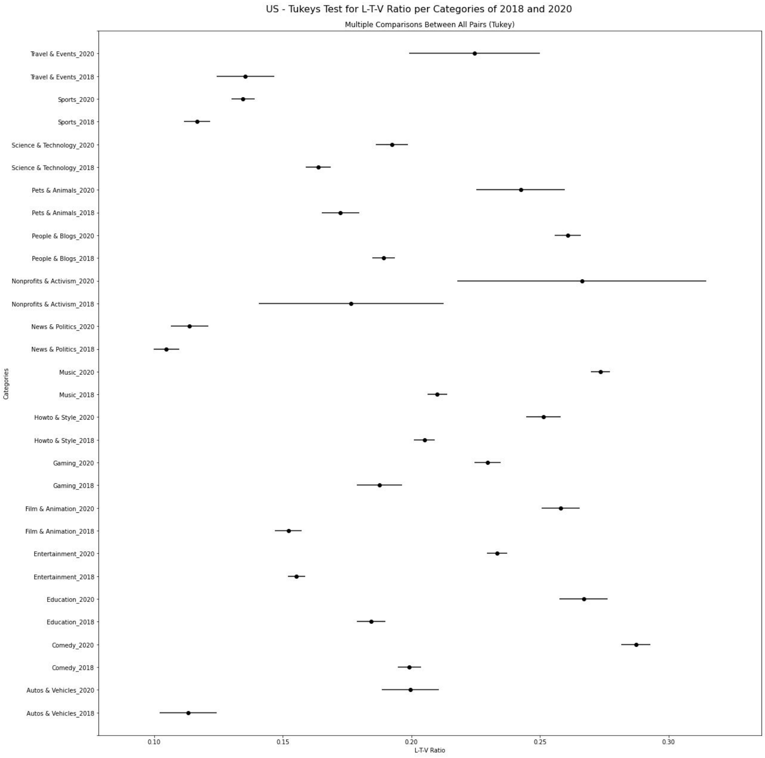
Country	**P-Value**	**Chi Test Statistic**
Canada	0.00000	773.10
Great Britain	0.00000	549.60
USA	0.00000	744.80

The ANOVA test showed significance for all three regions given that the obtained p-values were all less than our alpha level of 0.05. This indicates that in each region there is a difference between the square-root of the group means of L-T-V ratios for each category of 2018 and 2020. Next, as part of our Post Hoc analysis, we will apply Tukey's HSD to gain insight about the differences of the transformed group means. The following plots were obtained from applying Tukey's HSD to the square-root of the L-T-V ratios of each category of each region.

Figure 9. Plots of Tukey's HSD for Transformed Views per Category of Each Region




GB - Tukeys Test for L-T-V Ratio per Categories of 2018 and 2020
Multiple Comparisons Between All Pairs (Tukey)
Travel & Events_2020
Travel & Events_2018
Sports_2020
Sports_2018
Science & Technology_2020
Science & Technology_2018
Pets & Animals_2020
Pets & Animals_2018
People & Blogs_2020
People & Blogs_2018
News & Politics_2020
News & Politics_2018
Music_2020
Music_2018
Howto & Style_2020
Howto & Style_2018
Gaming_2020
Gaming_2018
Film & Animation_2020
Film & Animation_2018
Entertainment_2020
Entertainment_2018
Education_2020
Education_2018
Comedy_2020
Comedy_2018
Autos & Vehicles_2020
Autos & Vehicles_2018
Categories
0.10
0.15
0.20
0.25
0.30
L-T-V Ratio


US - Tukeys Test for L-T-V Ratio per Categories of 2018 and 2020
Multiple Comparisons Between All Pairs (Tukey)
Travel & Events_2020
Travel & Events_2018
Sports_2020
Sports_2018
Science & Technology_2020
Science & Technology_2018
Pets & Animals_2020
Pets & Animals_2018
People & Blogs_2020
People & Blogs_2018
Nonprofits & Activism_2020
Nonprofits & Activism_2018
News & Politics_2020
News & Politics_2018
Music_2020
Music_2018
Howto & Style_2020
Howto & Style_2018
Gaming_2020
Gaming_2018
Film & Animation_2020
Film & Animation_2018
Entertainment_2020
Entertainment_2018
Education_2020
Education_2018
Comedy_2020
Comedy_2018
Autos & Vehicles_2020
Autos & Vehicles_2018
Categories
0.10
0.15
0.20
0.25
0.30
L-T-V Ratio

In both Canada and Great Britain, every single category except Science & Technology and News & Politics had a statistically significant increase in L-T-V ratio. For both of these regions, News & Politics saw a drop in L-T-V ratio and Travel & Events saw no significant difference in L-T-V ratio. Moreover, In the USA, every single category except News & Politics had a statistically significant increase in L-T-V ratio. Specifically, News & Politics saw no significant difference in L-T-V ratio between 2018 and 2020.

(iii) Conclusions

From our analysis, we can conclude that for all three regions the L-T-V ratios among trending categories have changed between 2018, a year without a pandemic, and 2020, a year undergoing an active pandemic. We saw from our analysis that almost every region's trending categories saw an increase in L-T-V ratio in 2020. Moreover, the only category that saw a decrease in L-T-V ratio in all three of the regions was News & Politics. This can likely be attributed to the federal election in the USA in November of 2020. Given that this election has been extremely partisan, it is entirely possible that politically affiliated individuals are viewing opposing political news videos and not liking them. This cause has most likely affected the other two regions given that the USA has a strong global influence through it's news media outlets.

A potential reason for the increases in L-T-V ratio in 2020 can be that the total viewership of videos have increased in 2020, as seen from our analysis in part 1, and with increasing views, we expect more likes in more neutral categories. Furthermore, a higher L-T-V ratio might be interpreted as viewers expressing more positivity towards videos categories in 2020, in comparison to 2018. Moreover, it could also be possible that the pandemic has shifted viewership patterns and individuals exposed to newer categories are more likely to like them. One such example is the Education category, where we see an increase in L-T-V ratio in 2020 which could be due to the shift of learning from the classroom to an online environment.

Limitations

The one major limitation of this project is definitely the quality of the data that was used for the analysis. The first drawback of the data is that each dataset only included four months of the year rather than all of the data pertaining to the entire year. This alone has prevented us from minimizing the chances of erroneous errors showing up in the data, which could very well result in our analysis to be slightly or very skewed. The second drawback of the data is that there is no overlap in months between the 2018 and 2020 datasets since each covers a separate set of four months. This has prevented us from being able to control for seasonality which could very well introduce seasonal confounding variables resulting in our results to be skewed. In addition, the statistical tests that were used to assess viewership specific to video categories all had strict Normality requirements that our data failed to meet. This is clearly not ideal but since our datasets were fairly large we proceeded with using these tests and left the reader to assess their confidence in our results.

If we had more time we would have investigated the potential correlation between the 2020 trending viewership and COVID-19. We would have considered pandemic related data such as daily new cases, daily deaths, and daily hospitalizations. Analyzing these parameters could potentially give us more insights about the shifts in viewership trends for the different categories.

In retrospect, we should have spent more time sourcing our data and maybe even considered syndicated datasets that contain more thorough samples. If we had better data we would definitely be more confident in our conclusions and potentially have a more insightful analysis.

Project Experience Summary

(i) Kaveh Alemi

- Acquired YouTube dataset (50k rows), cleaned dataset to production quality, and data mined for interesting insights.
- Crafted hypothesis about frequency and viewership of trending categories in 2018 and 2020 for Canada, Great Britain, and the USA.
- Tested hypothesis using several parametric and non-parametric statistical tests using the SciPy package in Python.
- Performed in-depth Post Hoc analysis to gain insights about viewership patterns of trending categories.

(ii) Mike Thai

- Performed Extract-Load-Transform process on YouTube Video data to clean and filter out unwanted data for the use of statistical analysis.
- Generated meaningful visualizations of data that facilitate the explanation and interpretation of characteristics and trends in the data.
- Performed parametric and non-parametric statistical tests using SciPy in Python to determine differences in means for categorical data.

References

[1] Trending YouTube Video Statistics, 2018, Mitchell J, https://www.kaggle.com/datasnaek/youtube-new

[2] YouTube Trending Video Dataset (updated daily), 2020, Rishav Sharma, https://www.kaggle.com/rsrishav/youtube-trending-video-dataset

[3] YouTube video Categories list FAQs and solutions, 2019, TechPostPlus, https://techpostplus.com/youtube-video-categories-list-faqs-and-solutions