Marketing on YouTube

05.11.2022

—

Jacob Fishman,
Jf4322

Laksh Kataria,
Lvk8525

Ujjwal Vikram Kulkarni,
uk2011

Overview

Youtube has slowly become one of the leading entertainment/media conglomerates worldwide. The entertainment/media company opened its “doors” in 2005, and has been making steady progress in captivating all types of audiences. As of today, YouTube boasts impressive numbers, such as 122 million active users daily (1), 2.6 Billion monthly active users (2), and 1 billion hours of view time per day (3). This success has made YouTube one of, if not, the most influential entertainment/media platforms in the world. One of the side effects of this success is gardening a large and broad audience, which includes people all over the world. Over 71% of Canadian citizens visit the site on a monthly basis (4), making Youtube a viable platform for advertisers. Youtube’s advertising revenue hit 28.8 billion in the year 2021 (5), and with 500 hours of new videos of content being uploaded every day, there is an untapped market in identifying popular videos before they get popular.

Our mission is to provide customers with an opportunity to effectively advertise to a larger audience by discovering videos that will captivate audiences around Canada before they reach the trending page. Our analysis uses real time data from YouTube’s Trending section to understand what videos make it to the most viewed page on the site. Our approach is to analyze the real time data to determine the most effective ways of getting videos to the trending page, such as; time of day to post, best category of videos, useful tags and descriptions, and optimal thumbnails to catch a user’s interest.

Our goal is that we want to identify the videos that will reach the largest audience before they get to the trending page. This way your company can reach more viewers for less money than if you waited until the video got to the trending page. We know that this analysis will help your company reach more audiences around Canada, and therefore help your company's advertising goals.

Why do we need Big Data?

To perform our analysis, we have made use of several techniques that are commonly used in a Big Data paradigm. Big Data is an overarching term that encompasses all kinds of data, however, the most common and technical definitions of Big Data aim to define it within the scope of the ‘Three V’s’. The ‘Three V’s’ of Big Data are Volume, which refers to the size of data that we are dealing with, Variety, which refers to the variation in the fields of the available data, and Velocity, referring to the speed at which the data is collected or accumulated.

The need for our company to utilize Big Data approaches for our analysis is due to the size and variation within the dataset at hand, two of the three “V’s”. Without these approaches, it becomes almost impossible to ingest and subsequently operate on this data using traditional data processing methods.

To ingest and operate upon the YouTube video dataset, we are making use of the Pyspark framework. The primary reason for the use of Pyspark is to take advantage of Pandas' extremely efficient capabilities of reading and loading data, while overcoming Pandas' inability to perform or support distributed system operations in cases where we might need additional processing power to support our growing data.

In simple terms, Pandas runs data processing operations on a single machine where as Pyspark runs these operations in a distributed manner, on multiple machines and hence, can give much faster computation times and processing speeds as compared to Pandas, thereby finding extensive application in the fields of Data Science, Analytics and Machine Learning.

For our specific application, the data set we used was over 835MB and included several variations in the data field, per country. This made it incredibly difficult to operate on the data using regular data processing techniques and hence, Pyspark was the go-to solution.

The Data

The Dataset that we are using for this analysis is the Youtube Trending dataset from Kaggle. The Data is split up into 9 different countries, but for our analysis we are using the Canadian dataset, given that the companies we are pitching to want to advertise to Canadian citizens. Our hope is that with the success of this project we will be able to increase our efforts to other countries, and to gain access to all of YouTube's data, not just the Trending data.

The dataset we have spans from 2020-07-27 to 2022-04-24 for all videos that were on Canada's Trending dataset. The features we have are:

video_id | title | publishedAt | channelId | channelTitle | categoryId | trending_date | tags | view_count | likes | dislikes | comment_count | thumbnail_link | comments_disabled | ratings_disabled | description

The dataset we are working with looks like:

[30]

video_id	title	publishedAt	channelId	channelTitle	categoryId	trendi
KX06ksuS6Xo	Diljit Dosanjh: C...	2020-08-11T07:30:02Z	UCZRdNleCgW-BGUJf...	Diljit Dosanjh	10	2020-08-12T00
J78aPJ3VyNs	I left youtube fo...	2020-08-11T16:34:06Z	UCYzPXprvl5Y-Sf0g...	jacksepticeye	24	2020-08-12T00
M9Pmf9AB4Mo	Apex Legends | St...	2020-08-11T17:00:10Z	UC0ZV6M2THA81QT9h...	Apex Legends	20	2020-08-12T00
3C66w5Z0ixs	I ASKED HER TO BE...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51...	Brawadis	22	2020-08-12T00
VIUo6yapDbc	Ultimate DIY Home...	2020-08-11T15:10:05Z	UCDVPcEbVLQgLZX0R...	Mr. Kate	26	2020-08-12T00
ua4QMFQATco	CGP Grey was WRONG	2020-08-11T17:15:11Z	UC2C_jShtL725hvbm...	CGP Grey	27	2020-08-12T00
gi3VMMiFHVg	Giannis Gets Ejec...	2020-08-12T02:30:32Z	UC9-OpMMVoNP5o10_...	Bleacher Report	17	2020-08-12T00
7rlwxSPUcQk	ON EST POSITIF AU...	2020-08-11T16:00:31Z	UCpWaR3gNAQGsX48c...	Tibo InShape	17	2020-08-12T00
49Z6Mv4_WCA	i don't know what...	2020-08-11T20:24:34Z	UCtinbF-Q-fVthA0q...	CaseyNeistat	22	2020-08-12T00
p7HGUZWq_8s	Doing Doja Cat's ...	2020-08-11T19:00:09Z	UCucot-Zp428OwkyR...	James Charles	24	2020-08-12T00
w-aidBdvZo8	I Haven't Been Ho...	2020-08-11T20:00:04Z	UC5zJwsFtEs9WYe3A...	Professor Live	24	2020-08-12T00
kXLn3HkpjaA	XXL 2020 Freshman...	2020-08-11T16:38:55Z	UCbg_UMjlHJg_19SZ...	XXL	10	2020-08-12T00
AcBd_RH9JSw	PASSER UNE NUIT D...	2020-08-11T10:55:22Z	UCUl7mwOyySfZzUkq...	LeBouseuh	24	2020-08-12T00
jbGRowa5tIk	ITZY "Not Shy" M/...	2020-08-11T15:00:13Z	UCaO6TYtlC8U5ttz6...	JYP Entertainment	10	2020-08-12T00
GTp-0S82guE	Time to Talk..	2020-08-11T12:04:40Z	UCCgLoMYIyP0U56dE...	Chloe Ting	26	2020-08-12T00
nt3VVyv5pxQ	Try Not To Laugh ...	2020-08-11T17:00:31Z	UCYJPby9DRCteedh5...	Smosh Pit	22	2020-08-12T00
DF5T7HJ0Xug	Barra Swapped GT3...	2020-08-11T20:14:12Z	UCXIYLgIp6DYZHjmU...	Adam LZ	2	2020-08-12T00
gPdUslndvVI	Our Farm Got Dest...	2020-08-11T23:00:06Z	UCuxlXCfVyV-i5YLL...	Cole The Cornstar	22	2020-08-12T00
I6hswz4rIrU	Rainbow Six Siege...	2020-08-11T17:13:53Z	UCBMvc6jvuTxH6TNo...	Ubisoft North Ame...	20	2020-08-12T00
9AecsACtkB4	Watch Secret Serv...	2020-08-10T22:29:23Z	UCi7Zk9baY1tvdlgx...	CTV News	25	2020-08-12T00

Our analysis is solely focused on why videos reach and stay on Youtube's Trending page.

Exploratory Data Analysis

The primary aim of our research is to provide our clients with the ability to maximize their discoverability on YouTube and with it, produce substantial improvements in return on investment (ROI) on AdSense (official Google/YouTube partner program) revenue.

One of the major factors to ensure maximum ROI on each video is to advertise on videos with the maximum interactions. Interactions on YouTube are of several different types, including likes, comments, dislikes and most importantly, views.

The AdSense program on Youtube runs in the following manner: once a particular creator has been accepted to the YouTube Partner Program they can switch on the 'monetization' feature on each one of their videos. By monetizing a video, a content creator allows different advertisers to run their advertisements on the video. Advertisements are of varying length and may or may not provide the option to skip through them. For every view the advertisement gets, the advertizer typically pays between \$0.18 to \$0.31 to Google.

Google in turn, pays around 68% of this money to the publisher of the original video. The advertisement payments are varying but the average out to somewhere between $18 to $31 per 10,000 views for a content creator.

Hence, the most quantifiable metric to measure the popularity of a video is the number of views that a video garners. There are obviously other factors that could affect the popularity of a video such as the number of subscribers on a channel. However, the most robust metric is still the number of views on a video and our analysis would largely revolve around the same.

Another feature that YouTube offers, is a list of trending videos for each day of the week, and this is the dataset that we would be working with.

Videos are classified as trending on YouTube based on certain criteria, such as new and upcoming creators or artists, the growth in the number of views on a certain video, and what part of the world views come from. This implies that a particular video with the maximum view count for a day would appear on the countries trending video page. Trending videos however, provide us with the optimal dataset to work with in order to expand discoverability. These are the videos that everyone is watching, or everyone will watch at some particular point in time, and hence, advertising on these videos will almost certainly be a guarantee for success.

1. Most Popular Video Categories:

As part of our analysis, we first aim to find the categories of videos that perform the best in terms of the average view count for that category, and the number of videos created within that category.

In order to generate and subsequently visualize this data, we first need to clean up our data set by dropping any null values. After this, we need to ensure that all values listed within the category column are in the appropriate numerical format which can be used for further analysis. Once these checks are complete, we can move forwards with visualizing the data.

```
filepath = "/content/CA_youtube_trending_data.csv"

country_statistics(filepath)
```

The available data set has categories denoted by numerical values between 1 and 30, with each number mapping out to a particular category. However, upon further analysis of the data, it was found that the actual labeled data only contained a subset of these values.

Since there was no way of artificially generating synthetic labels for the available data, we have to work with what is available.

Based on the available data, the most popular categories for Canada were as shown below

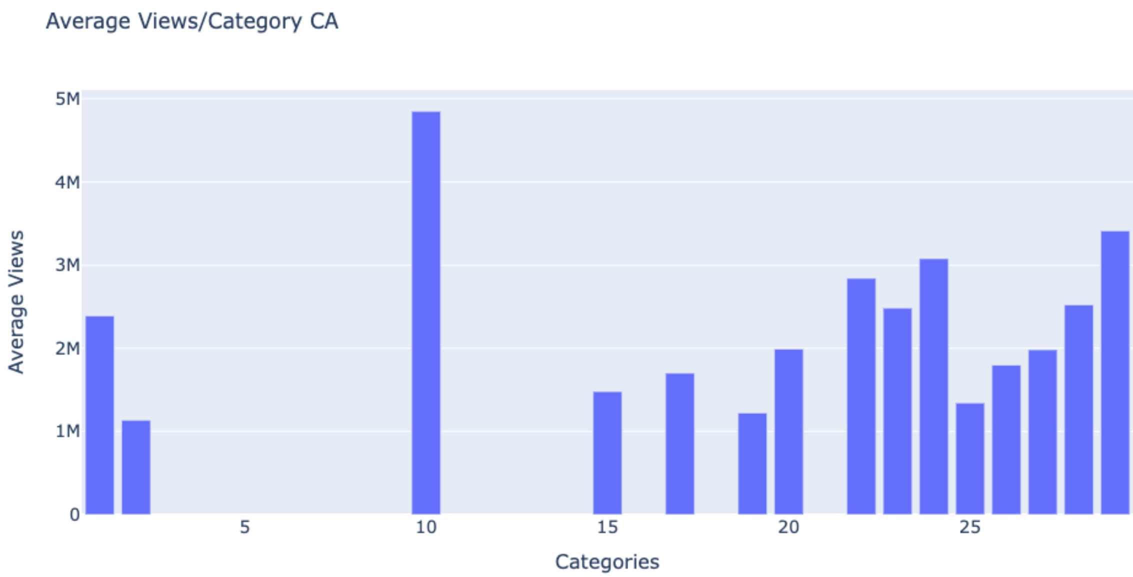


Fig: Graph depicting most popular categories by average views

```
df_final.sort(col('Average_View_Count').desc()).show()

+----------+------------------+
|CategoryID|Average_View_Count|
+----------+------------------+
|        10| 4849285.665504359|
|        29| 3412229.566037736|
|        24|3077659.5529612754|
|        22| 2841269.744994111|
|        28|2521956.3437367305|
|        23|2483340.4684210527|
|         1|2389252.1261770246|
|        20|1990565.7293740953|
|        27|1982307.7930264992|
|        26|1797695.3187198597|
|        17|1700892.5436824132|
|        15|1479093.6220689656|
|        25| 1342505.743773839|
|        19|1222084.5667022413|
|         2| 1134787.602897474|
+----------+------------------+
```

Fig: List of most popular categories by average view count

From the plot, we can clearly see that videos belonging to category 10, perform the best on average in terms of their view counts and those belonging to category 2 do not have great performance. Based on the labeled data available for Canada, category 10 is 'Comedy' and Category 2 is 'Music'.

This goes against our presumptions about categorical popularity, as we assumed Music videos would do pretty well in terms of their popularity and therefore provide the perfect marketplace for advertisement, however, the reality is that Music videos actually seldom make it on to the trending page on YouTube and hence, other avenues should be explored to advertise. We acknowledge that this may also be due to the data being misclassified, but based on the available evidence, our clients might want to steer clear of music videos when looking for advertisement opportunities!

Furthermore, we also provide a pictorial representation of the number of videos produced per category

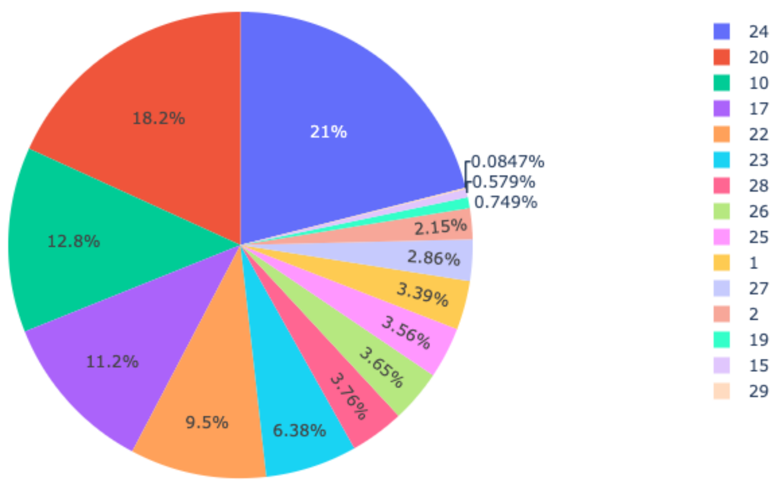


Fig: Chart showing distribution of videos per category

The plot shows that in terms of distribution, category 24 seems to be the most popular category amongst content creators. Category 24 for Canada is 'Family'. This makes sense as most YouTubers aim to produce family friendly content and we could imagine why a large number of videos from other categories such as 'Kids' or even 'Comedy' may fall into this category. In terms of the average view count however, category 24 lies third on the list.

This graph also highlights our second surprise in the analysis. As we can see from the plot, category 29 has the fewest number of videos produced. Even still, it ranks 2nd on the list of most popular categories by average view count. This implies that almost every video in that category is a guarantee to garner a large chunk of viewers. Category 29 is 'Shorts', which is a new feature implemented by YouTube. Multiple content creators could get paid from this category under a new 'Shorts Fund' created by YouTube, but the avenues for advertising on these videos are limited.

2. Popular Times:

As the second part of our analysis, we aim to find out if the time of day a video is posted has any relation with whether or not the video makes it onto the trending page. We select the entire data set available for Canada, filter down the rows containing null values, and then work on the 'publishedAt' column.

We firstly convert this column into the DateTime format by using the 'unixtime' function within the pyspark.sql.functions library. Then, we split up the values within this column to create a new column containing the corresponding hours.

```
df_new = df_new.withColumn('Time', regexp_replace('publishedAt','T',' '))
df_new = df_new.withColumn('Time_', regexp_replace('Time','Z',''))
df_new = df_new.drop('Time')
df_new = df_new.withColumn("Hour", from_unixtime(unix_timestamp(col("Time_"),"yyyy-MM-dd HH:mm:ss"),"HH"))
```

Fig: Pyspark code snippet of time transformation

We then classify the hours into four different parts of the day, i.e, Morning - 5am to 11am, Afternoon - 12pm to 4pm, Evening - 5pm to 8pm and Night - 9pm to 4am. These are the local times Based on this bifurcation, we create a pie chart depicting the most common timeframe in which a video is posted, such that the video also makes it onto the trending page.

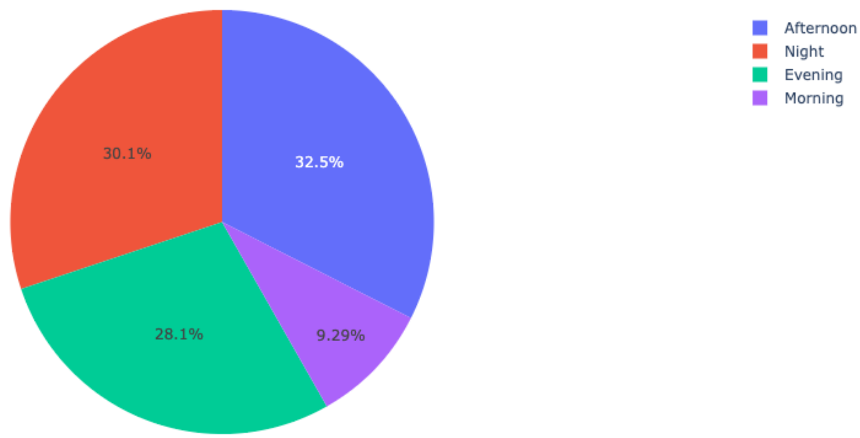


Fig: Chart showing the most common times for publishing trending videos

We can see here that most videos that make it on to the trending page in Canada are actually posted during the afternoon. This provides a great opportunity for advertising, if we keep a track of which videos are being posted during the afternoon times, and identify certain creators or artists that are most likely to make it to the trending page, this could give us a head start in identifying prospective advertisement avenues.

3. Popularity - Factors

As the next part in our analysis, we aim to identify the factors that affect the popularity of a video the most. The popularity of a video, as defined by us, depends upon the three major mediums of interaction with YoutTube videos, i.e, number of Views, Likes, and Comments. We do not consider the number of Dislikes on a video to be a useful metric because YouTube discontinued Dislikes from 2022 onwards, and hence, the dataset would be unnecessarily skewed.

Before we delve into this analysis, it is essential to understand that we use Popularity as a separate metric. This implies that just because a certain video makes it onto the trending page, does not necessarily mean that video is 'popular'. As defined earlier, trending videos are those which show the maximum growth in terms of number views over a particular day, further, the trending page also accommodates new artists and creators each day, and their content might not necessarily be the most viewed. Advertising on trending videos might be a good way to reach the maximum audience for a given day, but over the long term, advertising on videos with the maximum number of views seems to be the more robust metric.

The data set itself did not provide a binary classification between popular and not-popular videos, this classification was synthetically created by using a Pyspark UDF.

```
popularity_detector = f.udf(lambda x,y,z: (1.0 if x > (500000) and y > (10000) and z > (5000) else 0.0), returnType= FloatType())
popularity_detector_likes = f.udf(lambda x: (1.0 if x > (10000) else 0.0), returnType= FloatType())
popularity_detector_comments = f.udf(lambda x: (1.0 if x > (5000) else 0.0), returnType= FloatType())
df_sea = df_sea.withColumn('Popularity', popularity_detector('Views','Likes1','Comment_Count1') )
```

Fig: Pyspark code to show how a video was classified as popular

Intuitively, one would think that since the Popularity column has been synthetically created using three other columns from the dataset, all three of them would have equal correlation to the popularity metric, however, this was not observed.

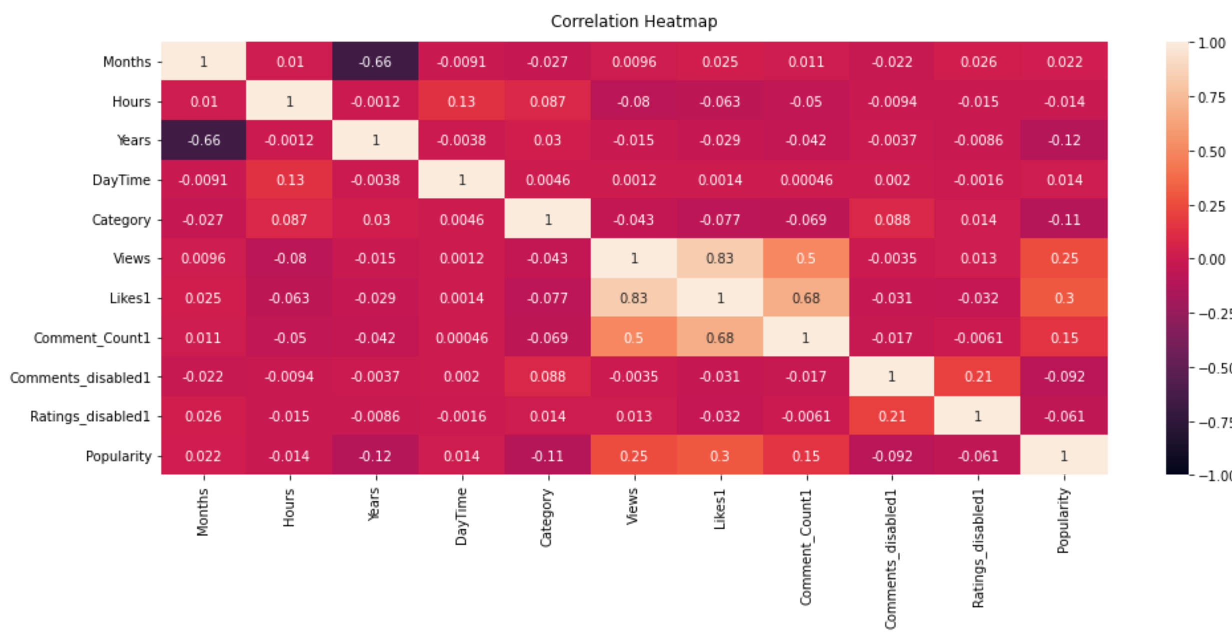


Fig: Heatmap showing feature correlation

Our intuition was somewhat correct, the features, Likes, Views and Comments do have the most correlation with popularity. However, it is observed that videos that are well liked are pushed further by YouTubes algorithm and are more likely to garner a larger number of views in the long run. Hence, while looking for long term advertisement opportunities, it would make sense to keep track of videos with a large number of likes as these will provide a good return on long time investment.

We noticed that the popularity metric we created does not correlate well with any one metric, which is expected given that it was created by using multiple metrics. However, we do observe some positive correlation between the popularity of a video and the month within which it is posted. This forms the basis for the last part of our analysis.

4. Target Months:

As the final part of our analysis, we aim to find the months during which most videos make it onto the trending page on YouTube. Our assumption is that holiday months are usually the most popular for content creators and consumers alike. AdSense revenues are also sky high during these months and hence, we expect maximum production to be during the months from October to February. Let's see if the analysis corroborates this claim.

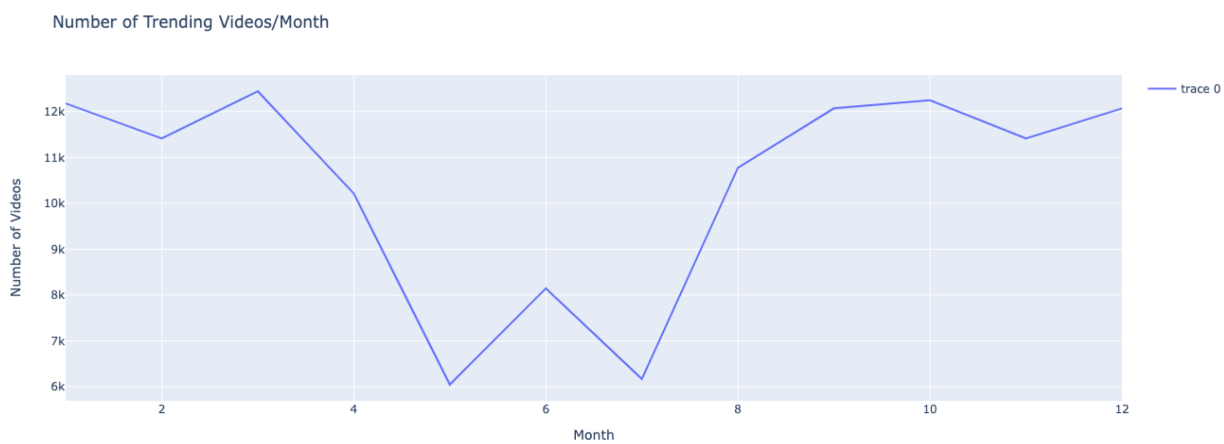


Fig: Line Chart showing video distribution by month

From the plot, we can see that it is in fact the month of March which is the most popular for uploading videos in Canada, and more specifically, videos that actually make it onto the trending page. However, this particular statistic should be viewed with a lower priority, and here is why. As we mentioned earlier, our data set runs from January of 2020 to April of 2022. This means that for the month of 2022, we only consider 4 months, i.e, January, February, March and April. This could possibly skew the data. However, even with the additional data, October still remains our second most popular month to upload a YouTube video with the best chance to make it onto trending, and the predicted trend of the holiday months being a gold mine for YouTubers and advertisers alike is maintained!

As there could be a possible skew in the data, we would choose the second most popular month by video count, i.e October for our further analysis.

The next bit of analysis we perform is we select our target month (October), and find out if there is a specific time of the day which gives a video the best possible chance to make it onto YouTubes trending page. We had observed earlier that for Canada, a majority of videos were posted between 12pm - 4pm, closely followed by night time, ie, between 8pm and 4am. This should give us some intuition as to what the expected trend should be.

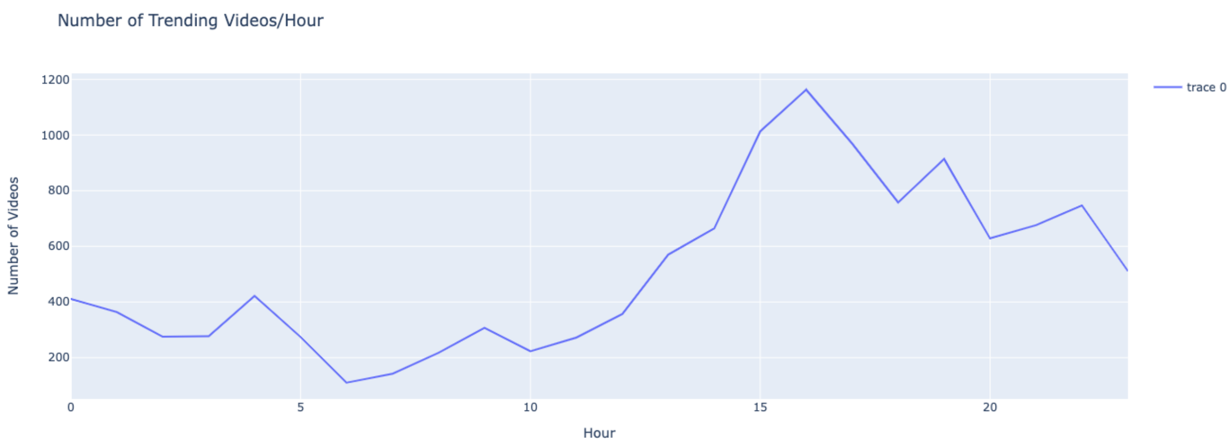


Fig: Line Chart showing video distribution by hour for October

Videos posted in the timeframe of 3pm to 5pm are the ones most likely to make it onto the trending page on YouTube which is updated daily at 00:00:00 UTC. Similarly, we can also visualize how many of these videos are actually popular as per our predefined popularity metric and compare if the trend follows a similar pattern.

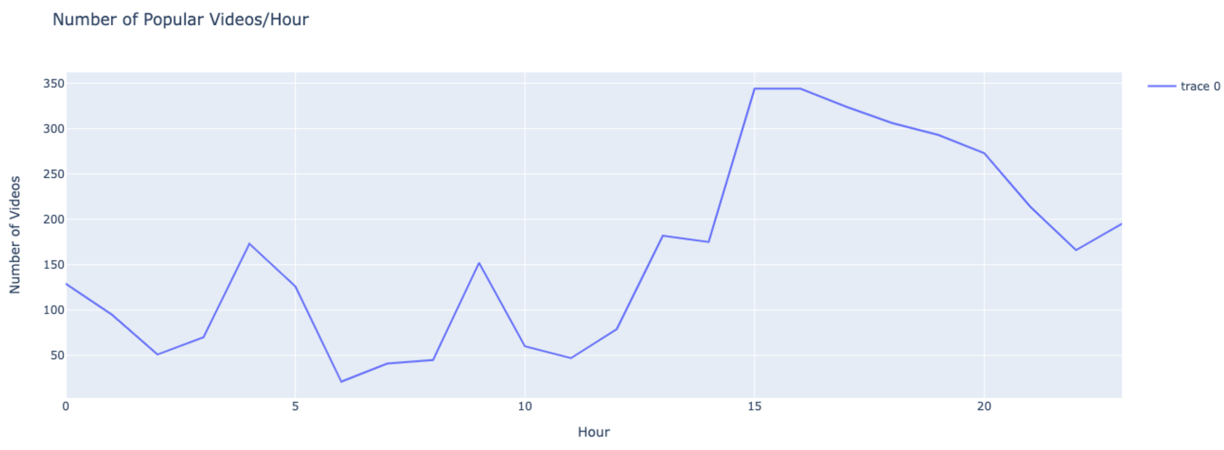


Fig: Line Chart showing popular video distribution by hour for October

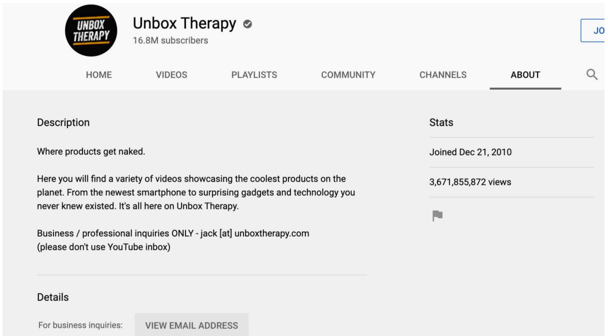
Videos posted between the times of 3pm to 8pm have the best chance of appearing on the trending page, and qualifying as generally popular videos.

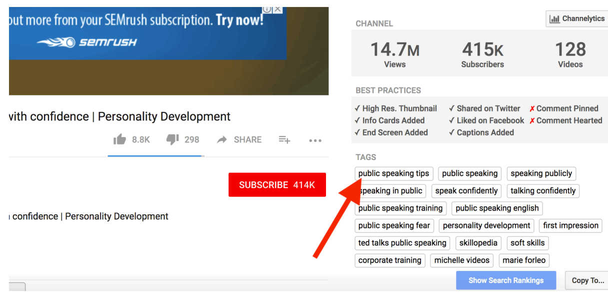
These analyses give our clients the best possible metrics for identifying advertisement opportunities and capitalizing on the same. Based on what we have learnt from these visual representations is that a video belonging to the 'Comedy' category, posted in the month of October, and in the 3pm to 8pm timeframe, might be the best possible target for a company looking to maximize their discoverability through advertisement.

These analyses however are very numerical and static, through the rest of our report, we go through other techniques such as identifying key tags and descriptions of videos which contribute to their popularity.

Tags and Description

The Tags and Description section of YouTube is a creator's ability to describe what their video is about, information about links to other videos, or the ability to promote their channel/other channels that they have. These tags and descriptions are very important to the success of a video, as they help draw people to click on a video given the short buzzwords, and encourage people to watch and stay engaged with the videos with a lively description. Throughout the datasets, we were able to do more analysis on these tags and descriptions with the aim of uncovering the perfect tags and descriptions to get viewers to watch videos.





Tags:

The tags of the videos can be anywhere from 0 tags for a video up to 50 unique tags for a single video. Our analysis of the tags for a video started by identifying the amount of unique tags in the Canadian dataset. There are 153,619 unique tags in the trending dataset. The top tag for all of the videos is the tag "funny" with 7026 unique instances. We use techniques of MinHash LSH that we learned in class in order to determine the similarities between tags of different videos. We tried many different NGram combinations, but the one that gave the most telling results was a Ngram of 2. Using this we were able to get the Jaccardian distances for the tags:

JaccardDistance	channel_Title_1	Title_1	view_count1	categoryId1	channel_Title_2	Title_2	view_count2	categoryId2
0.025000000000000022	MTV	BTS Reveals the M...	564762	24	MTV	BTS Performs "Tel...	8750155	10
0.02857142857142858	Chelsea Football ...	Tottenham 0-3 Che...	3707193	17	Chelsea Football ...	Chelsea 3-0 Aston...	5644621	17
0.02857142857142858	Cleetus McFarland	We Just Made the ...	1319203	1	Cleetus McFarland	The Freedom Facto...	921084	1
0.02857142857142858	Ed Sheeran	Ed Sheeran - Shiv...	8660548	10	Ed Sheeran	Ed Sheeran - Shiv...	426145	10
0.029411764705882236	Dude Perfect Plus	Unreleased Footag...	836888	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.029411764705882236	Dude Perfect Plus	UFC Golf Battle (...	756947	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.029411764705882236	Dude Perfect Plus	Submarine Minefie...	291478	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.029411764705882236	Dude Perfect Plus	Remote Control Ta...	462376	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.029411764705882236	Dude Perfect Plus	Chainsaw Carving ...	437812	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.029411764705882236	Dude Perfect Plus	Dude Perfect Corn...	948998	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.029411764705882236	Dude Perfect Plus	TY QUITS DUDE PER...	3852837	22	Dude Perfect Plus	Build A Raft Batt...	306796	22
0.030303030303030276	NLE CHOPPA	NLE Choppa - Jigg...	804052	10	NLE CHOPPA	NLE Choppa - Brys...	5968207	10
0.030303030303030276	NLE CHOPPA	NLE Choppa - Jigg...	804052	10	NLE CHOPPA	NLE Choppa - Done...	823010	10
0.030303030303030276	FitMC	The Mystery of 2b...	776922	20	FitMC	2b2t Griefers Jus...	916080	20
0.030303030303030276	FitMC	Why Are 2b2t Play...	829378	20	FitMC	2b2t Griefers Jus...	916080	20
0.030303030303030276	FitMC	The History of 2b...	613462	20	FitMC	2b2t Griefers Jus...	916080	20
0.030303030303030276	FitMC	2b2t's Final Neth...	980677	20	FitMC	2b2t Griefers Jus...	916080	20
0.032258064516129004	JxmyHighroller	We Thought This W...	975553	17	JxmyHighroller	The NBA is Gettin...	801618	17
0.032258064516129004	Paramount Pictures	Scream | Official...	9428997	1	Paramount Pictures	Scream (2022) - F...	2683321	1
0.033333333333333326	ScreenCrush	LOKI 1x05: Every ...	912518	1	ScreenCrush	LOKI 1x04: Every ...	699548	1
0.033333333333333326	Gucci Mane	Gucci Mane - Seri...	951642	10	Gucci Mane	Gucci Mane - Like...	733710	10
0.033333333333333326	Gucci Mane	Gucci Mane - Like...	733710	10	Gucci Mane	Gucci Mane - Bloo...	3449507	10
0.03448275862068961	Kodak Black	Kodak Black - Pur...	816341	10	Kodak Black	Kodak Black - Aug...	817087	10
0.03448275862068961	Kodak Black	Kodak Black - Pur...	816341	10	Kodak Black	Kodak Black - Hal...	922902	10
0.03448275862068961	Kodak Black	Kodak Black - Pur...	816341	10	Kodak Black	Kodak Black - Clo...	3982947	10
0.03448275862068961	Kodak Black	Kodak Black - Pur...	816341	10	Kodak Black	Kodak Black - I W...	622818	10
0.03448275862068961	JGOD	Warzone Pacific T...	256275	20	JGOD	Huge Stealth Chan...	207599	20
0.03448275862068961	JGOD	Played Warzone Pa...	415705	20	JGOD	Huge Stealth Chan...	207599	20
0.03448275862068961	Kodak Black	Kodak Black - Sup...	3685598	10	Kodak Black	Kodak Black - Pur...	816341	10
0.03448275862068961	Kodak Black	Kodak Black Ft. R...	3772972	10	Kodak Black	Kodak Black - Pur...	816341	10
0.03448275862068961	Kodak Black	Kodak Black Ft. C...	2220895	10	Kodak Black	Kodak Black - Pur...	816341	10
0.03448275862068961	Kodak Black	Kodak Black - Sen...	6418143	10	Kodak Black	Kodak Black - Pur...	816341	10

Here we can see that the tags that are the most similar are the ones from similar channels. When we filter out the Jacardian distances between videos from the same channel:

JaccardDistance	channel_Title_1	Title_1	view_count1	categoryId1	channel_Title_2	Title_2	view_count2	categoryId2
.03448275862068961	Parrot	Can I Solve this ...	932811	20	Spoke	1,037 Creepers VS...	696223	20
.03448275862068961	Parrot	Can I Solve this ...	932811	20	Spoke	1,017 Withers VS ...	672911	20
.03703703703703709	TommyInnit	I met KSI in real...	9377852	20	TommyOutit	I met George in r...	5237577	20
.040000000000000036	MoreTalkFCB	Ronald Koeman SPE...	242270	17	TalkFCB	Koeman tells Luis...	307148	17
.07017543859649122	KBS World	TWICE(트와이스) - I C...	2926854	24	KBS WORLD TV	TWICE(트와이스) -...	960996	24
.0714285714285714	JYP Entertainment	TWICE The Feels M/V	57594063	10	TWICE	TWICE The Feels C...	3836933	10
.0714285714285714	TWICE	TWICE The Feels O...	1655115	10	JYP Entertainment	TWICE The Feels M/V	57594063	10
.08333333333333337	ITZY	ITZY Performance ...	2906871	24	JYP Entertainment	ITZY Not Shy Albu...	3164828	10
.08333333333333337	JYP Entertainment	ITZY "Not Shy" M/V	6620953	10	ITZY	ITZY Performance ...	2906871	24
.08333333333333337	JYP Entertainment	ITZY "Not Shy" M/...	9073176	10	ITZY	ITZY Performance ...	2906871	24
.08333333333333337	ITZY	ITZY Not Shy Stag...	6984695	24	JYP Entertainment	ITZY Not Shy (Eng...	5674657	10
.08333333333333337	ITZY	ITZY Performance ...	2906871	24	JYP Entertainment	ITZY Not Shy (Eng...	5674657	10
.08333333333333337	Sleepy Hallow	Sleepy Hallow x S...	585063	10	Sheff G	Sheff G - Start S...	654982	10
.09090909090909094	NMIXX	[NMIXX] 占 (TANK) ...	877983	10	JYP Entertainment	NMIXX O.O M/V	41364831	10
.09090909090909094	NMIXX	[NMIXX] 占 (TANK) ...	877983	10	JYP Entertainment	NMIXX O.O M/V Teaser	2821535	10
.09999999999999998	NMIXX	[NMIXX] O.O Perfo...	1754842	10	JYP Entertainment	[NMIXX] Debut Tra...	3854729	10
.09999999999999998	Simon and Martina	What Happened to ...	392744	22	Eatyourkimchi Studio	Moving to Japan -...	190989	19
.09999999999999998	Simon and Martina	What's Inside a C...	171941	19	Eatyourkimchi Studio	Moving to Japan -...	190989	19
.1020408163265306l	ITZY	ITZY Not Shy Danc...	5553183	24	JYP Entertainment	ITZY Not Shy Albu...	3164828	10
.1020408163265306l	JYP Entertainment	ITZY "Not Shy" M/V	6620953	10	ITZY	ITZY Not Shy Danc...	5553183	24
.1020408163265306l	JYP Entertainment	ITZY "Not Shy" M/...	9073176	10	ITZY	ITZY Not Shy Danc...	5553183	24
.1020408163265306l	ITZY	ITZY Not Shy Danc...	5553183	24	JYP Entertainment	ITZY Not Shy (Eng...	5674657	10
.10344827586206895	SEVENTEEN	[Choreography Vid...	721340	10	Big Hit Labels	SEVENTEEN (세븐틴) '...	9428338	10
.10810810810810811	FIFATV	Tanzania v Congo ...	277791	17	FIFA	Egypt v Senegal |...	3039517	17
.10810810810810811	FIFATV	Tanzania v Congo ...	277791	17	FIFA	Senegal v Egypt |...	6150008	17
.10810810810810811	FIFATV	Tanzania v Congo ...	277791	17	FIFA	Algeria v Cameroo...	3544648	17
.10810810810810811	FIFATV	Djibouti v Algeri...	394556	17	FIFA	Algeria v Cameroo...	3544648	17

Where we can see that the videos with the most similar tags are the ones that are from channels of the same creators. This tells us that the tags are very unique to each channel. We can draw from this that the channels that are constantly getting to the Trending page are using their own tags that help them get to the top of their respective categorical charts.

We now looked at the tags for the top 10 channels for each category, to help us know what tags the Content creators can use to help model their tags for their videos to help get a following. The results of this data for the Canadian categories that were present in the trending data is:

categoryId	tags
15	['pets', 'animals', 'cute animals', 'wildlife', 'Animal Rescue', 'animal video', 'animals the dodo', 'dodo', 'pet videos', 'rescuing animals', 'the dodo', 'the dodo animals', 'wildlife videos', 'skit', 'dog', 'Alligator', 'Alligators', 'Shorts shelf', 'YouTube shorts', 'nature', 'animal', 'animaltherapy', 'comedy', '
29	['Ben Affleck', 'Chrissy Teigen', 'Daniel el Travieso', 'David Letterman', 'Eddie Vedder', 'Foo Fighters', 'Global Citizen', 'Global Citizen 2021', 'Global Citizen Festival', 'Global Citizen VAX Live', 'Global Citizen YouTube', 'Global Citizen concert', 'H.E.R.', 'HER', 'J Balvin', 'JLo', 'Jennifer Lopez', 'Jimmy K
28	['Apple', 'Apple Event', 'Apple Keynote', 'Apple Special Event', 'iPhone', 'AirPods', 'Mac', 'Keynote', '2021', 'iPhone 12', 'Ceramic Shield', 'iPad', 'Special Event', 'Tim Cook', 'M1', 'Introducing', 'M1 Max', 'iPad Pro', 'iMac', 'iOS', '5G', 'Apple Spring Event', 'Adaptive EQ', 'Apple silicon', 'Dolby Vision', 'M
22	['bella poarch', 'bella poarch new song', 'bella poarch tiktok', 'bella porch', 'bretman rock', 'sub urban', 'tiktok songs', 'valkyrae', '#shorts', 'TikTok #shorts', 'comedy', 'comp', 'daily', 'daily fun', 'daily tiktok', 'daily tiktoks', 'foryourpage', 'fun', 'funny', 'funny tiktok', 'funny tiktoks', 'funny videos', 'fu
27	['shorts', 'Face Mask', 'Face Mask Machine Compilation', 'asmr', 'asmr skin care', 'asmr skincare', 'crushing', 'diy', 'diy face mask', 'diy skincare', 'doctor ryan', 'dr ryan', 'dr. ryan', 'face mask asmr', 'face mask machine', 'face mask tiktok', 'face masks', 'fruit face mask', 'home face mask', 'natural face
17	['bowling', 'football', 'trick shots', 'bottle flip', 'clean', 'dp', 'dude', 'dude perfect', 'dude perfect bottle flip', 'dude perfect stereotypes', 'dude perfect water bottle flip', 'dude perfect world record', 'edition', 'family', 'family friendly', 'fidget spinners', 'ping pong', 'trick shot', 'water bottle flip', 'bu
26	['5 minute craft', '5 minute crafts', '5 minute crafts men', '5 minutes craft', 'camp', 'camping', 'camping hacks', 'fixing', 'home repair', 'how to fix', 'how to make', 'how to repair', 'minor repairs', 'outdoors', 'repairing', 'shorts', 'to fix', 'to repair', 'woodworking', '5 minute crafts for school', 'crafts', 'di
19	['1minute', 'Lesan.io', 'Nas Academy', 'Nas Daily', 'Nas Studio', 'Nasmeanspeople', 'Nuseir Yassin', 'Travel', 'people', 'è‹±èªžå-¦ç¿'', 'Disney World', 'shorts', 'Genius', 'Genius Kid', 'Inventor', 'Morocco', 'Robotics', 'Science', 'Science and Technology', 'Shorts', 'Talent', 'Avengers Campus', 'Avengers C
23	['tiktok', 'shorts', 'funny', 'comedy', 'fail', 'lol', 'Adam Waheed', 'AdamW', 'Sketch comedy', 'Youtube shorts', 'anwar', 'best shorts', 'best youtube shorts', 'hannah stocking', 'lele pons', 'room1041', 'skits', 'youtube trending', 'best tik toks', 'best tiktoks', 'tik', 'tik tok', 'tik tok compilations', 'tik tok fa
25	['news', 'News', '2020', 'guardian', 'C-SPAN', 'CSPAN', 'Telegraph', 'Donald Trump', 'White House', 'politics', 'us', 'Joe Biden', 'Republican', 'joe biden', 'trump', '2021', 'biden', 'donald trump', 'us election 2020', '2020 election', 'usa', 'scmp', 'gdnpfpnewsworld', 'Commission on Presidential Debates',
24	['#shorts', 'shorts', 'youtube shorts', 'best of tiktok', 'edits', 'experiment videos', 'family booms', 'family booms tiktok', 'funny videos', 'illusion', 'magic videos', 'shorts videos', 'tiktok trends', 'tiktok videos', 'tiktok viral', 'vfx', 'viral tiktok videos', 'viral videos', 'amazon', 'amazon prime', 'amazon p
1	['New Movie', 'New Trailer', 'Coming Soon', 'Trailer', 'stick fight', 'Official Trailer', 'minecraft animation', 'stick figure', 'Action (Movie Genre)', 'SkyDance', 'amazon prime', 'amazon prime video', 'amazon studios', 'animator vs animation', 'minecraft animations', 'minecraft vs animation', 'stickman
20	['mobile game', 'brawl stars', 'mobile battle royale', 'mobile strategy game', 'supercell game', 'battle royale', 'mobile rpg', 'minecraft', 'minecraft challenge', 'dream', 'beating minecraft', 'minecraft but challenge', 'Dream Minecraft', 'Minecraft but water rises every minute', 'challenge', 'dream M
10	['K-pop', 'YG', 'YG Entertainment', 'ì™€ì´ì§€', 'BLACKPINK', 'BLINK', 'ë¡œì œ', 'ë¦¬ì‚¬', 'ë¸”ëž™í•‘í¬', 'ë¸”ë§í¬', 'ë¸”í•‘', 'ì œë‹ˆ', 'ì§€ìˆ˜', 'Jennie', 'Jisoo', 'Lisa', 'RosÃ©', 'ë¹…ìž‡ìŠ¤', 'BLACKPINK THE ALBUM', 'BTS', 'THE ALBUM', 'ë¸”ëž™í•‘í¬ THE ALBUM', 'BANGTAN', 'BIGHIT', 'ë°©íƒ„', 'ë°©íƒ„ì†Œë…„ë‹¨', 'BLA
2	['satisfying', 'repair', 'Chevrolet', 'fast', 'shorts', 'at home', 'capron funk', 'car paint', 'chalk paint', 'chalk spray paint', 'challenge', 'charlie oliver bates', 'collosion', 'comedy', 'copart', 'corey funk', 'diy', 'drew dirksen', 'easy experiments', 'experiment', 'fix', 'funk bros', 'funny', 'gnar char', 'gnarly char

These tags are crucial for creators in these categories to get to the top view counts, and stay on the trending page with all of their new videos. These tags will be the baseline for your employees to help them accrue a following to their videos.

Descriptions

The Description place gives a creator more freedom to describe the content of their video, which means there will be more variety in the descriptions of different videos. We use the same technique that we used with the tags to see if there are any similarities between videos in the trending set. Yet again we see very similar results for the Jacarian Distances between descriptions for videos:

JaccardDistance	channel_Title_1	Title_1	view_count1	categoryId1	channel_Title_2	Title_2	view_count2	categoryId2
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	18246708	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	22756261	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	8382569	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	15605938	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	26038234	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars anima...	6118652	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	8202134	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	22756261	20	Brawl Stars	Brawl Stars: Braw...	9213036	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	8382569	20	Brawl Stars	Brawl Stars: Braw...	9213036	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	15605938	20	Brawl Stars	Brawl Stars: Braw...	9213036	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	26038234	20	Brawl Stars	Brawl Stars: Braw...	9213036	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	8202134	20	Brawl Stars	Brawl Stars: Braw...	9213036	20
0.001901140684410...	Brawl Stars	Brawl Stars Anima...	9084466	20	Brawl Stars	Brawl Stars Anima...	3768694	20
0.001901140684410...	Brawl Stars	Brawl Stars Anima...	9084466	20	Brawl Stars	Brawl Stars Anima...	5702517	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	9213036	20	Brawl Stars	Brawl Stars: Braw...	18246708	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	9213036	20	Brawl Stars	Brawl Stars Anima...	3768694	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	9213036	20	Brawl Stars	Brawl Stars anima...	6118652	20
0.001901140684410...	Brawl Stars	Brawl Stars: Braw...	9213036	20	Brawl Stars	Brawl Stars Anima...	5702517	20
0.001919385796545...	Brawl Stars	Brawl Stars: Braw...	22633604	20	Brawl Stars	Brawl Stars Anima...	6630505	20
0.001919385796545...	Brawl Stars	Brawl Stars: Braw...	23809149	20	Brawl Stars	Brawl Stars Anima...	6630505	20
0.003795066413662229	Brawl Stars	Brawl Stars: Braw...	9213036	20	Brawl Stars	Brawl Stars Anima...	9084466	20
0.004132231404958664	CouRage Shorts	The Worlds Fastes...	550612	22	CouRage Shorts	Connor Grew Up... 😥	4678495	22
0.006802721088435382	Coldplay	Coldplay - Higher...	9068687	10	Coldplay	Coldplay - Higher...	607611	10
0.007547169811320753	Clash of Clans	WANTED: Pirate Qu...	4807763	20	Clash of Clans	Pirate King Takes...	6062083	20
0.007692307692307665	Breakfast Club Po...	The LOX On Showma...	850237	24	Breakfast Club Po...	The Breakfast Clu...	995503	24
0.007692307692307665	Breakfast Club Po...	The Breakfast Clu...	995503	24	Breakfast Club Po...	Pressa Talks Toro...	161870	24
0.007692307692307665	Breakfast Club Po...	The Breakfast Clu...	995503	24	Breakfast Club Po...	Soulja Boy Goes O...	961566	24
0.00917431192660545	Kyle Exum	When a Robber Tri...	954438	23	Kyle Exum	Robbing Chick-Fil...	805413	23
0.00917431192660545	Kyle Exum	The Mom vs. Middl...	988343	23	Kyle Exum	Robbing Chick-Fil...	805413	23
0.00917431192660545	Kyle Exum	When a Robber Tri...	954438	23	Kyle Exum	The Mom vs. Middl...	988343	23
0.00917431192660545	Kyle Exum	When a Robber Tri...	954438	23	Kyle Exum	If Among Us Was a...	933581	23
0.00917431192660545	Kyle Exum	Robbing Chick-Fil...	805413	23	Kyle Exum	If Among Us Was a...	933581	23
0.00952380952380949	Brawl Stars	Brawl Stars: Braw...	22756261	20	Brawl Stars	Brawl Stars: Braw...	23809149	20
0.00952380952380949	Brawl Stars	Brawl Stars: Braw...	15605938	20	Brawl Stars	Brawl Stars: Braw...	23809149	20
0.00952380952380949	Brawl Stars	Brawl Stars: Braw...	26038234	20	Brawl Stars	Brawl Stars: Braw...	23809149	20

And when we filter out videos that are from different channels:

JaccardDistance	channel_Title_1	Title_1	view_count1	categoryId1	channel_Title_2	Title_2	view_count2	categoryId2
0.05208333333333337	SMTOWN	[STATION] TEN 텐 '...	1520815	10	SM STATION	[STATION] TEN 텐 '...	5384634	10
0.07002801120448177	Jurassic World	Jurassic World Do...	447161	24	Universal Pictures	Jurassic World Do...	9410344	24
0.0755555555555556	Miley Cyrus	Miley Cyrus - Liv...	3685672	24	Foo Fighters	Foo Fighters - Li...	439238	10
0.0755555555555556	Miley Cyrus	Miley Cyrus - Liv...	903935	10	Foo Fighters	Foo Fighters - Li...	439238	10
0.0849220103986l347	MoneyBagg Yo	Moneybagg Yo - Fr...	639078	10	MoneybaggYoVEVO	Moneybagg Yo - Fr...	772427	10
0.08571428571428574	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23	Colleen Vlogs	I CAN FINALLY TAL...	992526	24
0.08571428571428574	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23	Colleen Vlogs	Holding The Twins...	1567423	24
0.08571428571428574	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23	Colleen Vlogs	BRINGING MAISY HO...	2435850	24
0.08571428571428574	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23	Colleen Vlogs	BRINGING WESLEY H...	1973041	24
0.08571428571428574	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23	Colleen Vlogs	First Trimester N...	562426	24
0.08571428571428574	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23	Colleen Vlogs	Finding Out I'm P...	708850	24
0.08571428571428574	Colleen Vlogs	Rumors About TRIP...	1387209	24	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23
0.08571428571428574	Colleen Vlogs	Visiting my babie...	1756877	24	Colleen Ballinger	PREGNANCY MUKBANG...	928878	23
0.0861244019138756	SidemenShorts	Why Harry wore ta...	668134	22	MoreSidemen	SIDEMEN 8 YEAR AN...	2966055	22
0.0861244019138756	SidemenShorts	The Brown Variant	885290	22	MoreSidemen	SIDEMEN 8 YEAR AN...	2966055	22
0.08823529411764708	UnspeakablePlays	Testing Illegal M...	930621	20	Unspeakable	Can You Land In T...	2790011	22
0.08823529411764708	UnspeakablePlays	Testing Illegal M...	930621	20	Unspeakable	LAST To LEAVE 100...	4242056	22
0.08823529411764708	UnspeakablePlays	Testing Illegal M...	930621	20	Unspeakable	ESCAPING 100 LAYE...	3581598	22
0.08823529411764708	UnspeakablePlays	Testing Illegal M...	930621	20	Unspeakable	EVERY STEP, YOU K...	2385647	22
0.08823529411764708	UnspeakablePlays	Testing Illegal M...	930621	20	Unspeakable	I Bought A REAL T...	3318312	22
0.08823529411764708	UnspeakablePlays	Testing Illegal M...	930621	20	Unspeakable	100 Buttons but O...	4602973	22
0.08823529411764708	UnspeakablePlays	Testing Clickbait...	2560841	20	Unspeakable	Can You Land In T...	2790011	22
0.08823529411764708	UnspeakablePlays	Testing Clickbait...	2560841	20	Unspeakable	LAST To LEAVE 100...	4242056	22
0.08823529411764708	UnspeakablePlays	Testing Clickbait...	2560841	20	Unspeakable	ESCAPING 100 LAYE...	3581598	22
0.08823529411764708	UnspeakablePlays	Testing Clickbait...	2560841	20	Unspeakable	EVERY STEP, YOU K...	2385647	22
0.08823529411764708	UnspeakablePlays	Testing Clickbait...	2560841	20	Unspeakable	I Bought A REAL T...	3318312	22
0.08823529411764708	UnspeakablePlays	Testing Clickbait...	2560841	20	Unspeakable	100 Buttons but O...	4602973	22
0.08823529411764708	UnspeakablePlays	STRANDED ON AN IS...	819695	20	Unspeakable	Can You Land In T...	2790011	22
0.08823529411764708	UnspeakablePlays	STRANDED ON AN IS...	819695	20	Unspeakable	LAST To LEAVE 100...	4242056	22
0.08823529411764708	UnspeakablePlays	STRANDED ON AN IS...	819695	20	Unspeakable	ESCAPING 100 LAYE...	3581598	22
0.08823529411764708	UnspeakablePlays	STRANDED ON AN IS...	819695	20	Unspeakable	EVERY STEP, YOU K...	2385647	22
0.08823529411764708	UnspeakablePlays	STRANDED ON AN IS...	819695	20	Unspeakable	I Bought A REAL T...	3318312	22
0.08823529411764708	UnspeakablePlays	STRANDED ON AN IS...	819695	20	Unspeakable	100 Buttons but O...	4602973	22
0.08823529411764708	Unspeakable 2.0	We Put A BOAT In ...	611549	22	UnspeakablePlays	Testing Illegal M...	930621	20
0.08823529411764708	Unspeakable 2.0	We Put A BOAT In ...	611549	22	UnspeakablePlays	Testing Clickbait...	2560841	20

What we learned from this information is that there is no formula to a description that would help garner more views. Tags are the bigger driving force in a video’s success and accumulating videos. The last aspect of our analysis focuses on thumbnails, and how they impact a video's popularity.

Object Detection using AWS Rekognition

One of the interesting features in the youtube dataset was the thumbnail link column.
A video thumbnail is the viewer's first impression of the video, as it is the first thing that a user sees. A great video thumbnail can mean the difference between thousands of views and just a few. Various thumbnail items can entice the user to play a specific video. Hence, we decided to leverage this feature and used AWS Recognition to recognize objects from the thumbnails.

Rekognition is an AWS service that analyzes your photographs using deep learning. By sending an image or video to the AWS Rekognition API, you can easily incorporate Rekognition into your application. Objects, people, language, scenes, and activities will be identified by the service. Face analysis and recognition are also incredibly accurate with Amazon Rekognition. Face detection, analysis, and comparison are available for a range of applications, including user identification, cataloging, people counting, and public safety.

Our goal is to identify which objects can grab a user's interest quickly.

```
%matplotlib inline
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import boto3
import time
final = []
client=boto3.client('rekognition',region_name='us-east-1')
for i in range(114869,us_df.shape[0]):
  imageFile=str(i)+'.jpg'
  try:
    img=mpimg.imread(imageFile)
 with open(imageFile, 'rb') as image:
        %time response = client.detect_labels(Image={'Bytes': image.read()})

    print('Detected labels in ' + imageFile + ':\n')
    objects = []
    for label in response['Labels']:
        obj = label['Name']
        objects.append(obj)
  except Exception as e:
    print('Not detected labels in ' + imageFile + ':\n')
    objects.append([])
  final.append(objects)

Streaming output truncated to the last 5000 lines.
CPU times: user 6.84 ms, sys: 176 µs, total: 7.02 ms
Wall time: 253 ms
Detected labels in 123893.jpg:

CPU times: user 4.77 ms, sys: 1.34 ms, total: 6.11 ms
Wall time: 392 ms
Detected labels in 123894.jpg:

CPU times: user 4.98 ms, sys: 1.07 ms, total: 6.05 ms
Wall time: 223 ms
Detected labels in 123895.jpg:
```

Highest object count per category:

We can see that a human/person dominates the count for each category. Rekognition often gives an extensive list of synonymous objects which can cause some objects to occur together very frequently. In the given snapshot, it can be seen that person and human are categorized as different entities, even though they indicate the same thing. This analysis does not provide any detail about how the objects play a role in determining views but it does give an idea that almost every image has a human in it.

We can get a deeper understanding of the objects by ranking them and see how they are scattered for each category.

```
dfwords=df.withColumn('obj',split(col('objects'),','))\
.withColumn('object',explode(col('obj')))\
.drop('objects','obj').groupBy('object','categoryId').agg(count('object')\
.alias('count')).orderBy('count',ascending=False)
df_find = dfwords.withColumn("category", recode('categoryId', category))
df_find = df_find.select('category','object','count')
df_find.show()
```

```
+-------------+------------+-----+
|     category|      object|count|
+-------------+------------+-----+
|Entertainment|      Person|19463|
|Entertainment|       Human|18339|
|       Gaming|      Person|15843|
|       Gaming|       Human|15009|
|        Music|      Person|11581|
|        Music|       Human|10838|
|       Sports|      Person|10631|
|       Sports|       Human|10110|
|People & Blogs|     Person| 8612|
|People & Blogs|      Human| 8274|
|Entertainment|   Performer| 7814|
|Entertainment|        Face| 7003|
|Entertainment|Advertisement| 6959|
|Entertainment|      Poster| 6444|
|Entertainment|    Clothing| 6179|
|Entertainment|     Apparel| 6151|
|       Gaming|   Performer| 6128|
|       Comedy|      Person| 5991|
|Entertainment|      People| 5949|
|       Gaming|Advertisement| 5880|
+-------------+------------+-----+
only showing top 20 rows
```

Top 5 ranked objects :

After exploding the dataset, we ranked each object and extracted the top 5 ranked individual objects that appear for each category. Even though all of them have a lot of commonalities, some unique objects do stand out for each category. As you can see, in the category how to & style, females in the thumbnails attract more users. We can leverage this point to advertise female centric products in this category. In the Autos & Vehicles category, Sedan cars dominate over every type of car. Hence, this can again be useful to target a specific set of audience.

Ranking based on individual objects gives an insight as to how objects on its own dominate each category.

category	collect_set(object)
Science & Technology	[Human, Text, Electronics, Phone, Person]
Howto & Style	[Human, Clothing, Female, Apparel, Person]
Education	[Human, Text, Advertisement, Poster, Person]
People & Blogs	[Human, Face, People, Performer, Person]
Film & Animation	[Human, Text, Art, Performer, Person]
Gaming	[Human, Text, Advertisement, Poster, Person]
Travel & Events	[Human, Face, People, Person, Sitting]
Pets & Animals	[Canine, Mammal, Pet, Person, Animal]
Sports	[Human, Crowd, People, Sport, Person]
Comedy	[Human, Clothing, Face, Performer, Person]
Music	[Human, Crowd, Apparel, Performer, Person]
News & Politics	[Human, Electronics, Crowd, Performer, Person]
Entertainment	[Human, Face, People, Performer, Person]
Nature	[Human, Text, Face, [], Person]
Autos & Vehicles	[Human, Vehicle, Sedan, Person, Car]

Top ranked grouped objects:

The next we tried to see was which set of objects frequently together. This will give us a more specific idea about the videos which trend. As it can be inferred, videos containing Iphones topped the Science and Technology category, Film and Animation is dominated by Angry Birds and so on.

Thumbnails do matter when it comes to influencing views. We may deduce from the analysis that each category has a distinct set of objects that appeal to the user. Advertisement corporations can leverage their knowledge of such objects to target specific audiences.

category	collect_set(objects)
Science & Technology	[Phone,Electronics,Mobile Phone,Cell PhoneIphone]
Howto & Style	[Electronics,Phone,Mobile Phone,Cell PhonePerson,Human]
Education	[Art,Graphics,Symbol,SignLight,Lighting]
People & Blogs	[Sitting,Person,Human,IndoorsCrowd,People,Performer,Clothing,Apparel,Hair,Pants,Office,Music Band,M
Film & Animation	[Angry Birds]
Gaming	[Pac Man]
Travel & Events	[Person,Human,Sweets,ConfectioneryFood,Text,Word,Face,People,Dating,Performer,Banner,Birthday Cake,
Pets & Animals	[Golden Retriever,Canine,Animal,MammalDog,Pet,Puppy, Sitting,Person,Human,AnimalGolden Retriever,Do
Sports	[Person,Human,Sport,SportsBoxing, Crowd,Person,Human,Press Conference]
Comedy	[Person,Human,Performer,HairPeople,Poster,Advertisement,Face,Crowd,[]]
Music	[Painting,Art,Person,Human, Performer,Person,Human,HeadFace,Hair,Advertisement,Poster,Haircut,Miner
News & Politics	[Person,Human,Crowd,PerformerPress Conference]
Entertainment	[Person,Human,Face,WaterAnimal,People,Aquarium,Sea Life,Outdoors,Photography,Photo,Advertisement,Bi
Nature	[Blonde,Female,Teen,PersonGirl,Woman,Kid,Human,Child,Tie,Accessories,Accessory,Hair,Dress,Clothing,.
Autos & Vehicles	[Sports Car,Car,Vehicle,TransportationAutomobile,Coupe,Race Car,Mustang,Sedan]

Conclusion

We have seen throughout our analysis that there are a multitude of factors that contribute to the overall popularity of a Youtube video. Our company has identified qualities of videos that drive success and overall popularity, such as posting in a Family category, posting in the month of March, tags that match the top 50 per category, and thumbnails that include people. We are confident that we can identify videos that will reach the broadest audience in Canada before they become popular, thus cutting down on the expenses associated with advertising on videos, and making sure we advertise as early as possible. We will be able to test this analysis on upcoming videos throughout the year, and be able to expand our efforts as we receive more information about Youtube videos.

We are excited for the opportunity to help your company grow and to have your product reach a whole new audience of Canadian citizens. As we grow together, we will make your product one of the most popular around the world.

References

(1) *YouTube Statistics 2022 [Users by Country + Demographics]*. (n.d.). Retrieved May 7, 2022, from https://www.globalmediainsight.com/blog/youtube-users-statistics/

(2) • *Most used social media 2021 | Statista*. (n.d.). Retrieved May 7, 2022, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

(3) *You know what's cool? A billion hours*. (n.d.). Retrieved May 7, 2022, from https://blog.youtube/news-and-events/you-know-whats-cool-billion-hours/

(4) *YouTube Stats | AC Social Media*. (n.d.). Retrieved May 7, 2022, from https://www.algonquincollege.com/ac-social-media/youtube-stats/

(5) *YouTube Earnings: Video Platform Has More Revenue Than Netflix in Q4 – The Hollywood Reporter*. (n.d.). Retrieved May 7, 2022, from https://www.hollywoodreporter.com/business/digital/youtube-ad-revenue-tops-8-6b-beating-netflix-in-the-quarter-1235085391

(6) *Pandas vs PySpark DataFrame. Pandas: | by Prakash R | featurepreneur | Medium*. (n.d.). Retrieved May 7, 2022, from https://medium.com/featurepreneur/pandas-vs-pyspark-dataframe-1722cb987fbd

(7)

(8)