YouTube Trending Videos Analysis

Elham Emami

Nov. 2019

Table of Contents

1. Introduction

Our goal is to advise our client that makes marketing videos and publish on YouTube by analyzing the YouTube datasets and creating a model that can predict the number of views for their next video. The content on YouTube covers a broad range of genres such as news, politics, movies, comedy, sports, fashion, gaming and fitness. A person looking at related videos suggested by YouTube will first see the title and the thumbnail. If more potential views can be generated with specific titles and thumbnails, a YouTuber could use this information to generate the maximum potential views with video content they worked hard on. Therefore, our goal was to create a model to predict the view count that marketing company can use to help get more vies and grow their channel.

2. Data Overview

YouTube is the most popular and most used video platform in the world today. YouTube has a list of trending videos that is updated constantly. The dataset was made from a dataset on Kaggle called "Trending YouTube Video Statistics" is being used to analyze all the important elements and features for this project. For each of those days, the dataset contains data about the trending videos of that day. It contains data for about 335000 trending videos. We will analyze this data to get insights into YouTube trending videos, to see what is common between these videos. Moreover, the main purpose of this project is to investigate what features of YouTube videos are important to attract more users and engage them by reacting on video such as commenting or liking it and predict the number of video's views based on its attributes.

More details on each video such as channel id, channel title, channel description, channel published time have been extracted from the YouTube API. There are CSV and JSON files from 9 countries. The CSV files contain the YouTube video data and the JSON files include data related to the video category. Each CSV file is imported into a separate dataframe. For reference, a sample dataframe with the Canadian video records is displayed in table 1:

video_id	trending_date	title	category_id	publish_time	channel_title	…	likes	Dislike	comment_count
n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	10	2017-11-10T17:00:03.000Z	EminemVEVO		787425	43420	125882
0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	23	2017-11-13T17:00:00.000Z	iDubbbzTV		127794	1688	13030
5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	23	2017-11-12T19:05:24.000Z	Rudy Mancuso		146035	5339	8181

Table 1 – Sample of YouTube dataframe

3. Data Wrangling

The JSON files include the video category information. The "Category_id" column of the dataframes are mapped to the JSON files to retrieve the category value of the videos and added to the dataframe. The following steps were added to clean and organize the data:

1. The description and category columns that have missing values in the dataframes are filled with the "Unavailable" string.

2. The "Trending Date" and "Published Time" columns contain a date and time value, but they are of the Object type. The columns are converted to the date type and reformatted for data analysis.
3. A “country” column is added and filled with the corresponding country to help identify the origin of each video record.

Afterward, the dataframes are merged into a single dataframe. In order to set the “video_id” column as the index column of the dataframe, the column needs to be cleaned and validated. A group by query over the video_ids is executed to identify any invalid ids. Indices are created with those ids to drop any invalid record. To ensure the “video_id” is unique in the dataframe, the dataframe is sorted in descending order based on “trending_date” and then grouped by “video_id”. The first record of each group of video_ids that has the latest "trending_date" is chosen. The purpose of this process is to ensure the uniqueness of each “video_id” as there are multiple records for each video_id with a different value of "trending_date".

Using the YouTube API, we can provide more details for the videos. One of the aspects of the video dataframe that needs to be investigated but cannot be found in the CSV file is the channel data. we can extract the channel information from the YouTube API. To get the video information through the YouTube API, requests are submitted using the API key that has been generated from the Google developer console. The first step is to prepare the query. The list of video_id is extracted from the dataframe and then the query is executed for each video id. The responses are a list of dictionaries that contain the channel details. To get more information about the channel, a new query is executed over the YouTube API requesting the channel type. The response is converted to a dataframe and then cleaned to have a unique and valid channel_id. Afterward, the channel dataframe and the video dataframe is merged based on the channel_id.

4. Exploratory Analysis

The analysis and exploration of the YouTube trending videos show that the record counts of the video dataset for each country are not equal. According to the below pie chart, only 7% of datasets are for Korea, Japan, and Great Britain, compared with Denmark that has almost 20% of the records.

The following attributes of the dataset have been analyzed to recognize what elements and features have an impact on increasing the number of video's views.

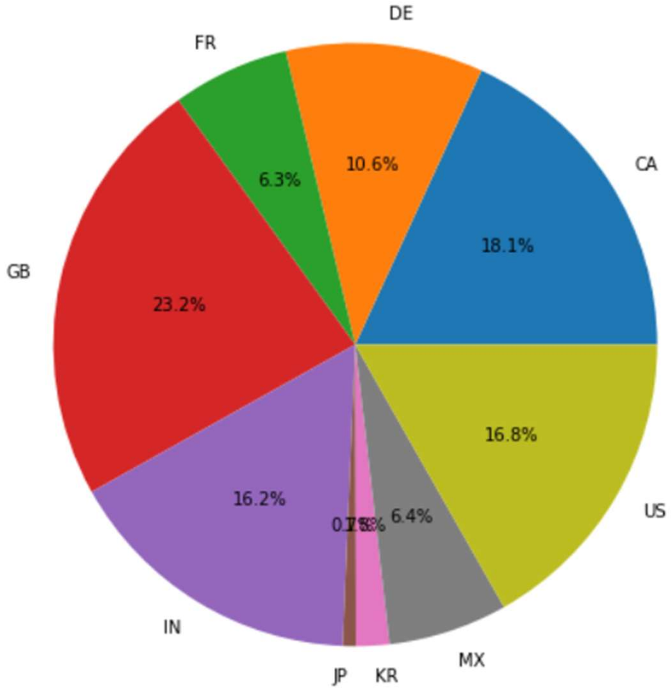


Figure 1 – YouTube views % per country

4.1 Correlation between views and video's tag and title, likes, dislike, comments:

The new attributes that might have impacts on the video's views number have been extracted such as: the number of words in the video's title, number of words in the video's tag, number of unique words in the video title and number of unique words in the video tag, number of words in title and tag that are uppercase or are in title format and the number of punctuations in the title and tag of the video.

There is a strong correlation between Views and the like counts as per correlation matrix in figure 2. There is more chance for a video to be viewed again if it has been liked previously. However, the dislike and comments number of the videos show they are not popular and cause the number of views to decrease:

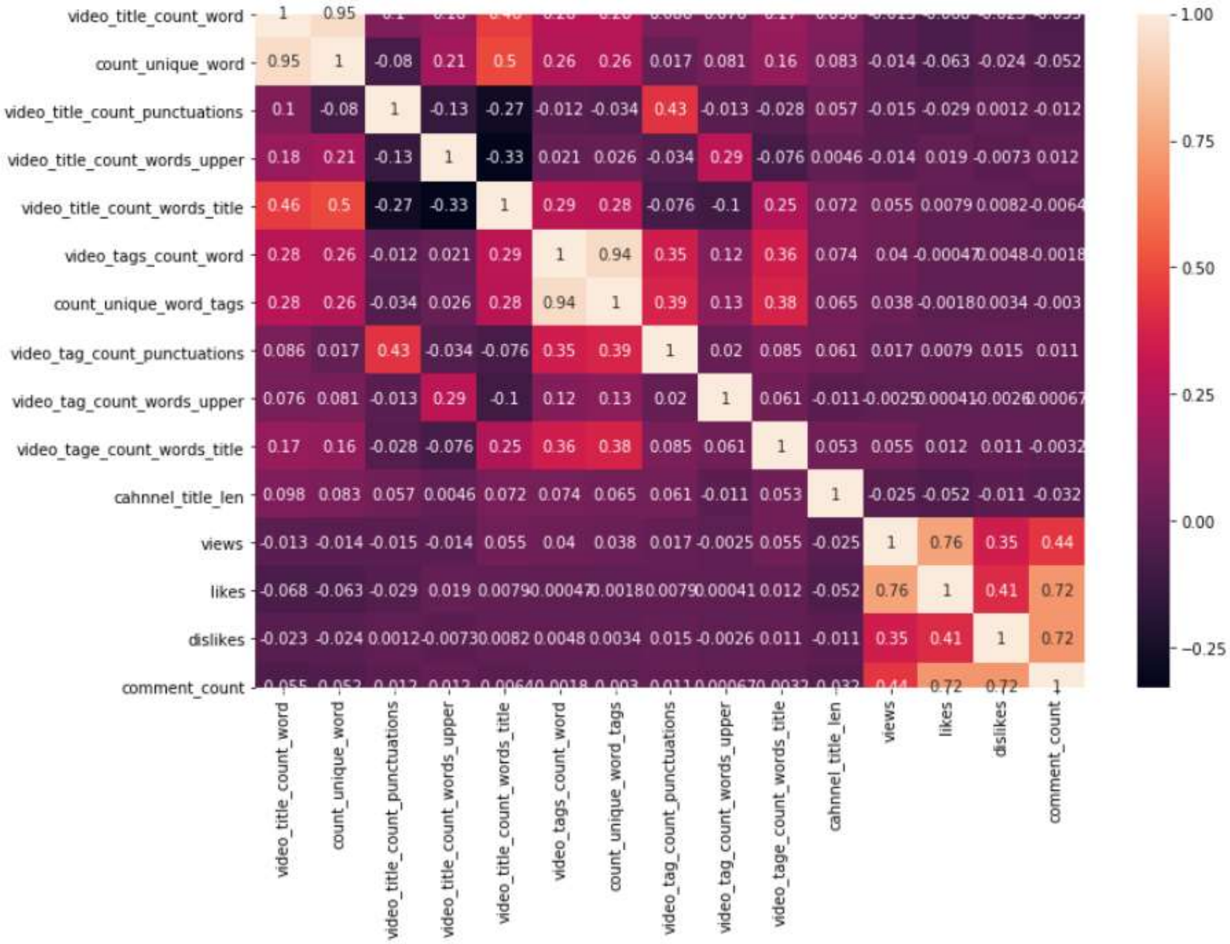


Figure 2 – Correlation matrix between YouTube dataset features

4.2 Trending date:

The graph displays that the average number of videos which has been watched on YouTube has increased dramatically over the months in Canada and Great Britain. The analysis shows most of the popular videos have been watched one Sunday in Great Britain but for most of the studied countries the videos mostly viewed on weekdays such as in the US and Denmark according to figure3.

The average views for weekdays in Canada display that Saturday has the most views among other days of the week and this number keeps decreasing till Thursday. Thus, we should consider the country for predicting the view number of views based on the day of the week.

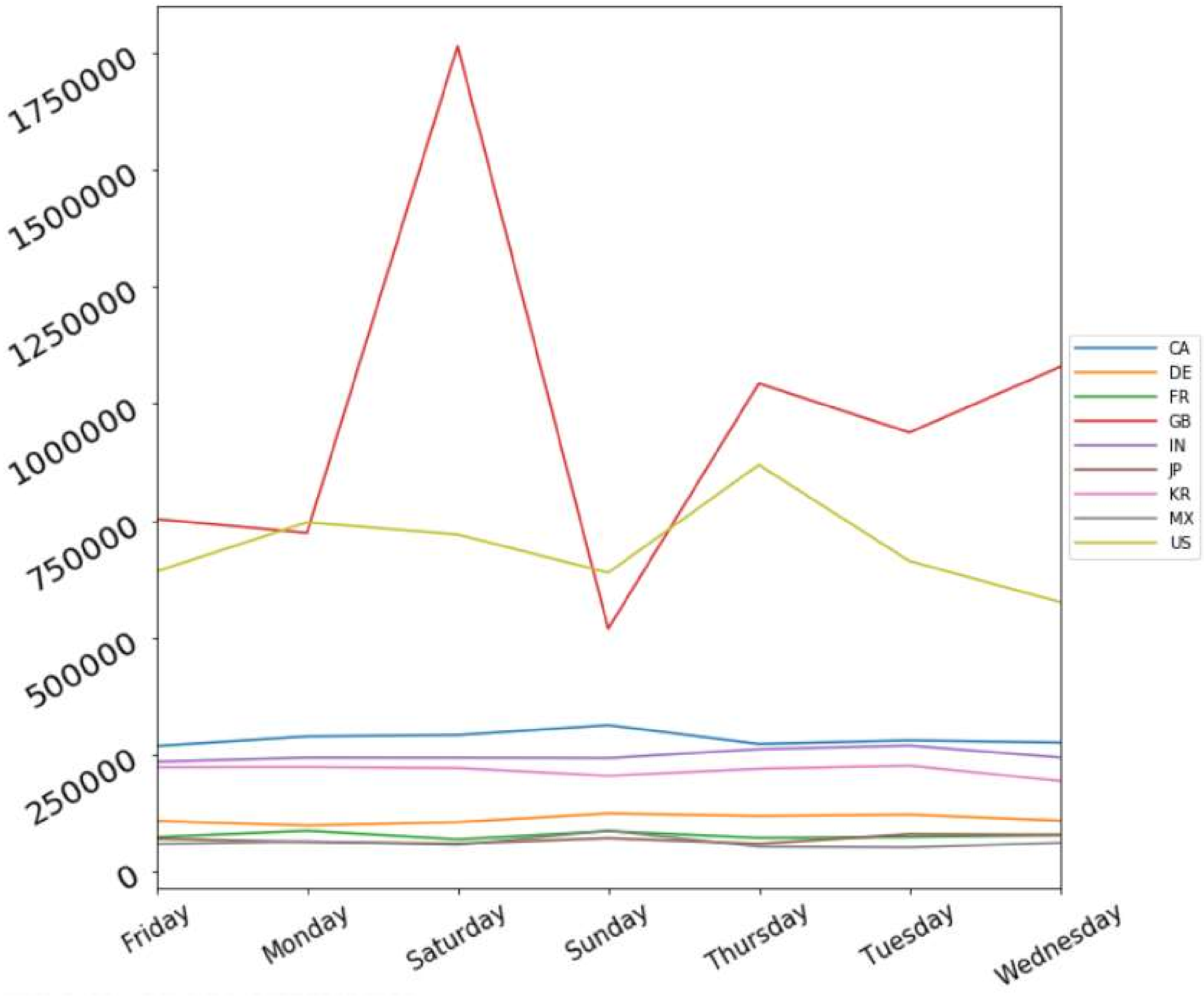


Figure 3 – YouTube video average views in weekdays

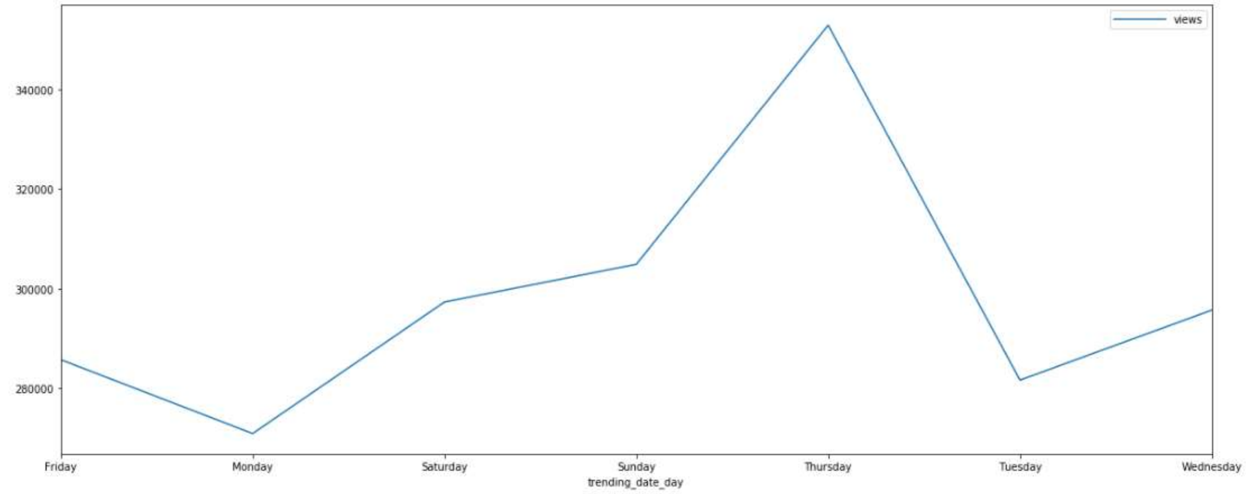


Figure 4 – YouTube trend for average vide views in days of week in Denmark

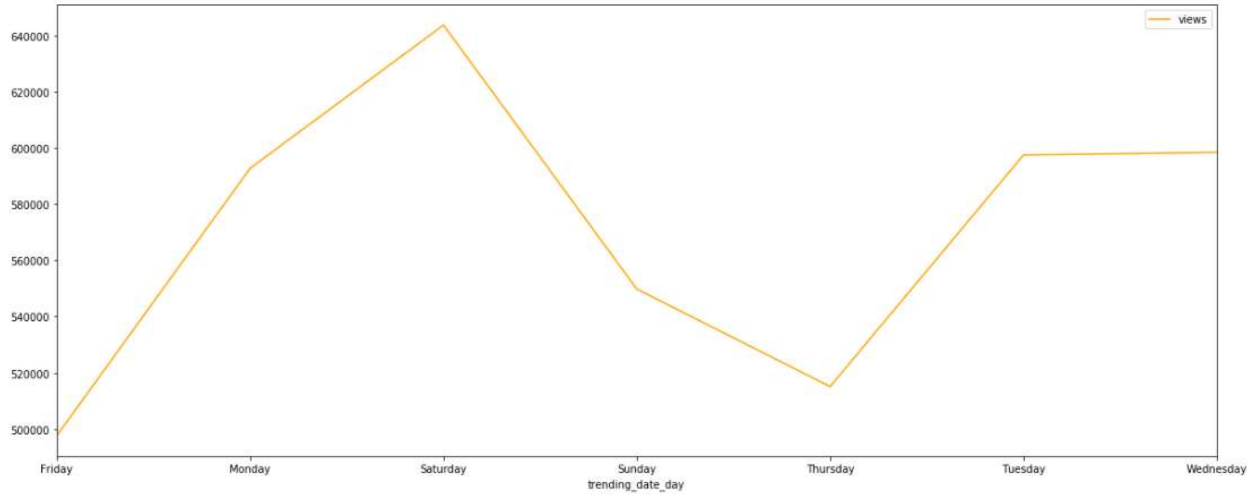


Figure 5 – YouTube trend for average vide views in days of week in Canada

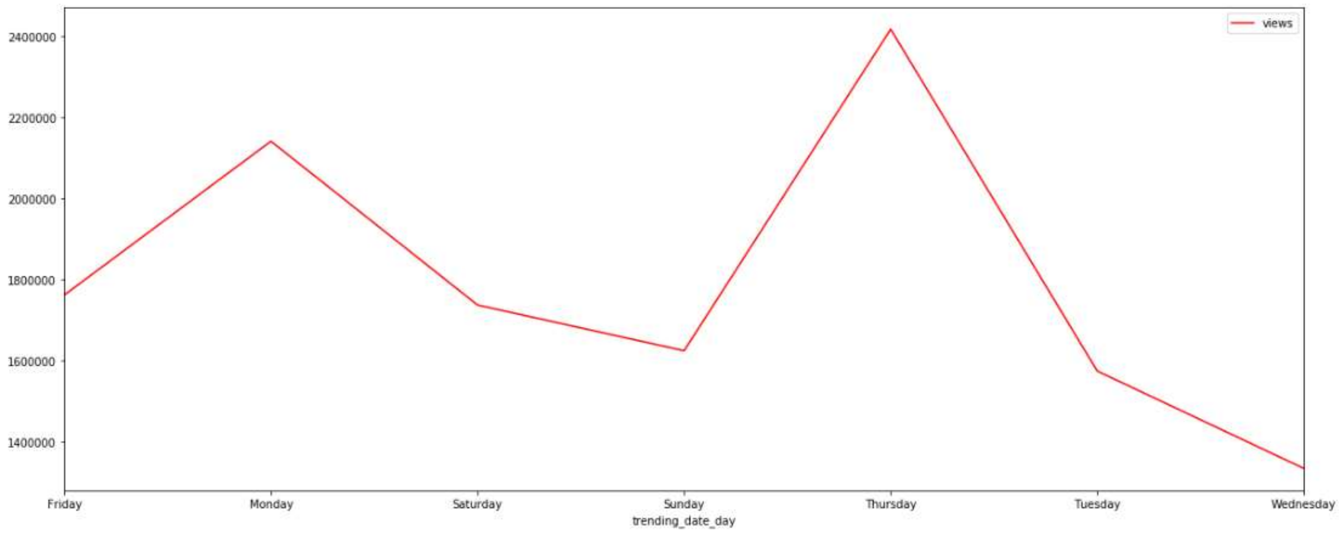


Figure 6 – YouTube trend for average vide views in days of week in United States

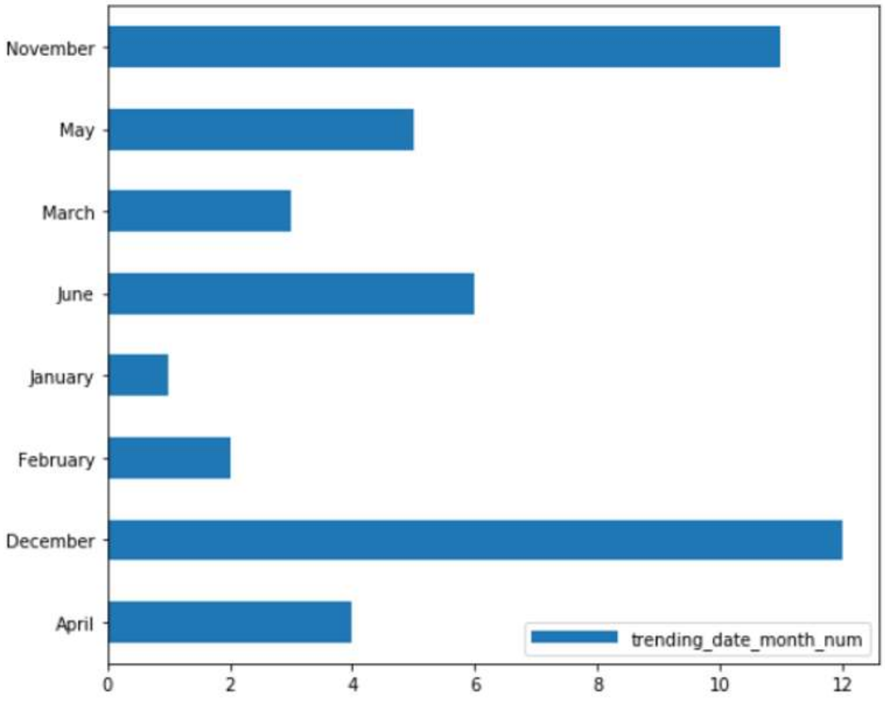


Figure 7 – YouTube trend for per months

The graph which represents the monthly views of the first half of the year 2018 shows people watch more videos in the winter months rather than Summer. We can conclude there is more possibility for a video to be top trending in winter rather than summer from the below graph.

4.3 Category:

Entertainment videos have been watched most among all other categories, but movies and music have been watched on average more than the other categories. The videos with the activism and nonprofit, sports, and political content comes next after the movie category.

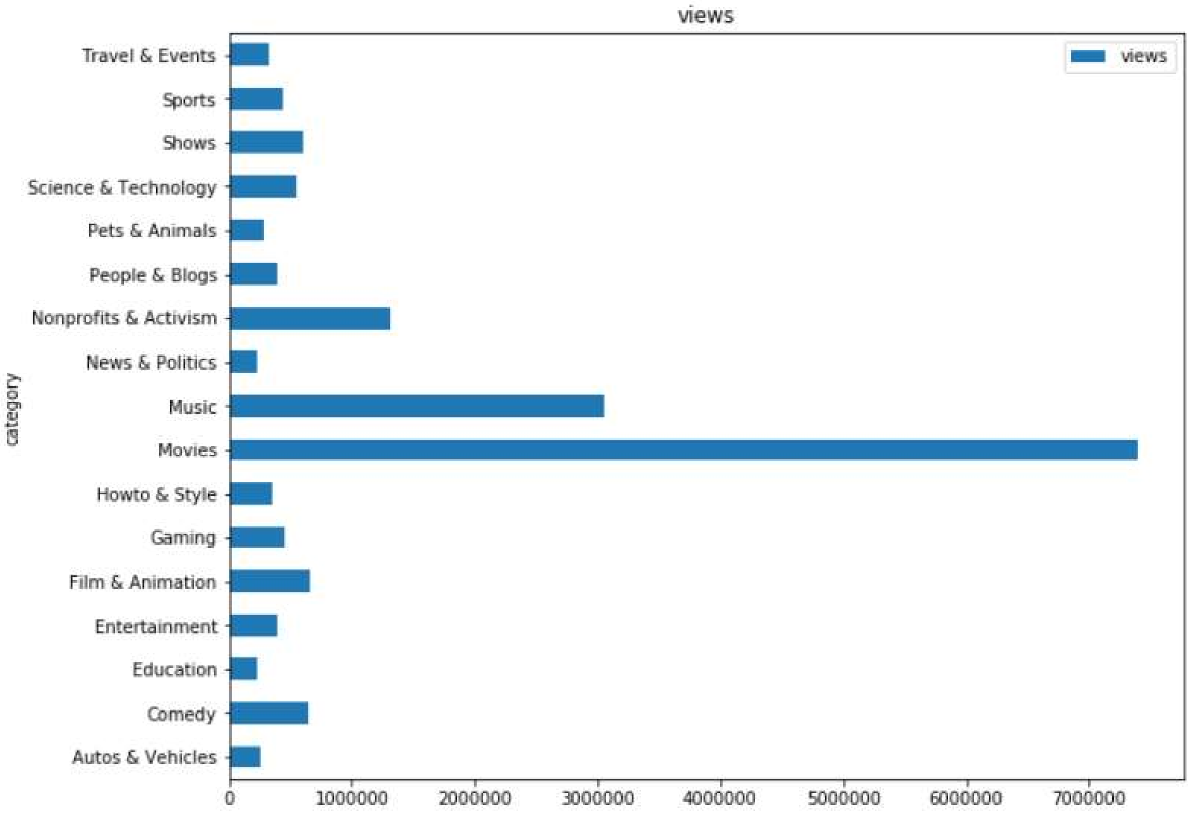


Figure 8 – YouTube average video's views for different categories

In the following graph, the most-watched videos on average in different countries have been displayed. The different categories of videos are being watched in different countries and we cannot conclude one main category is being watched mostly in all countries. According to graph 6 gaming, music and film and animations are one of the most popular video's categories in the United States. However, Canadian popular categories are music, shows and comedy.

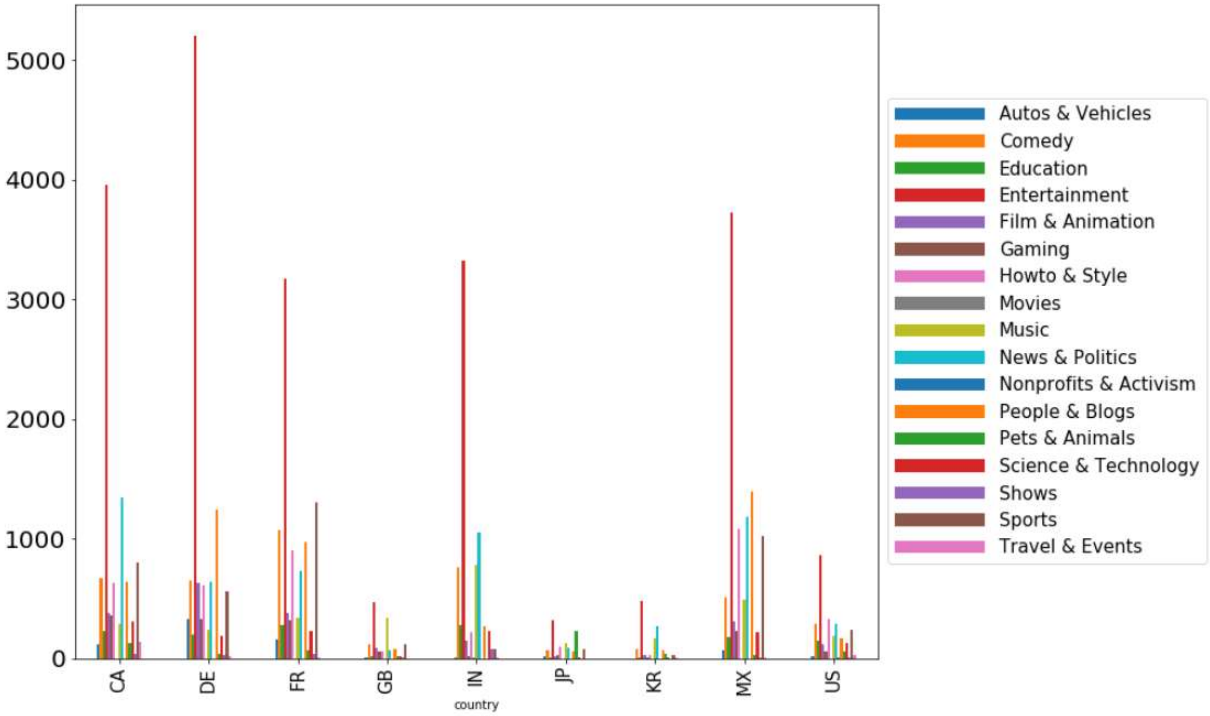


Figure 9 – Median number of views for different category from different country

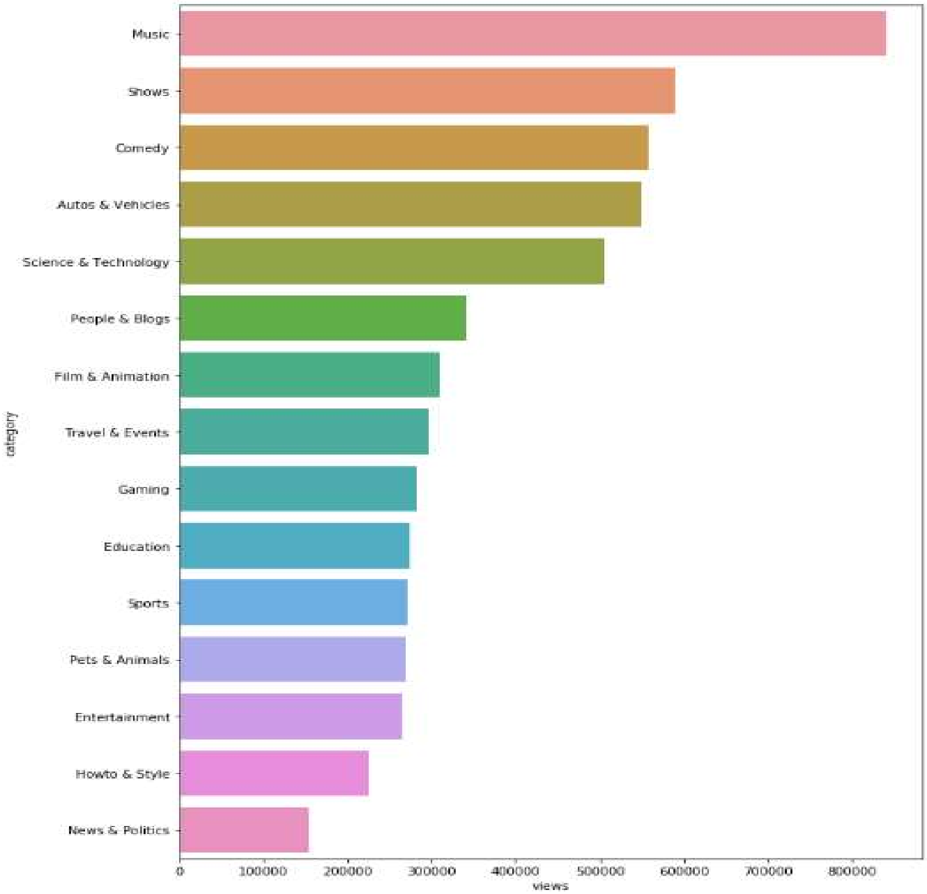


Figure 10 – Median number of views for each category - Canada

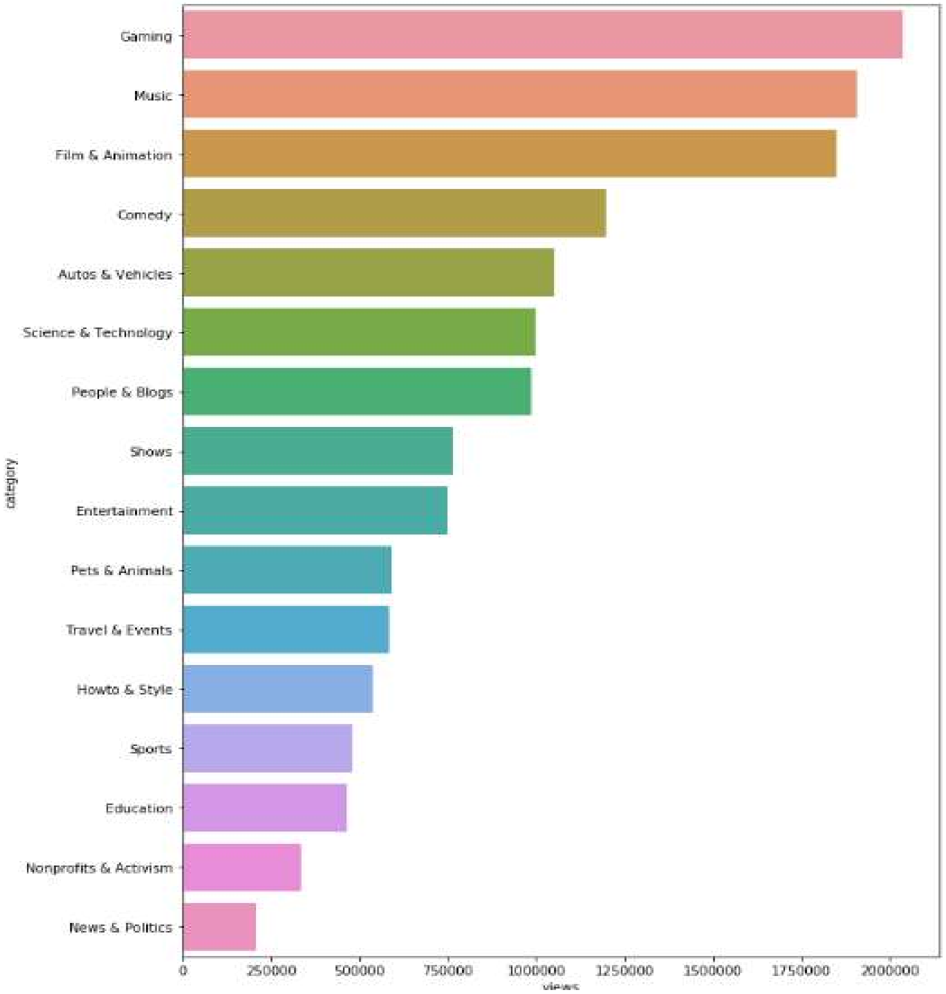


Figure 11 – Median number of views for each category - USA

The below figure shows the sentiments on the video title based on the category. As per below bar graph, video titles in news and politics, comedy and gaming had negative impacts. However, they did not have any impact on movies. The video title had a positive impact on videos with blogs and people, music and sports content.

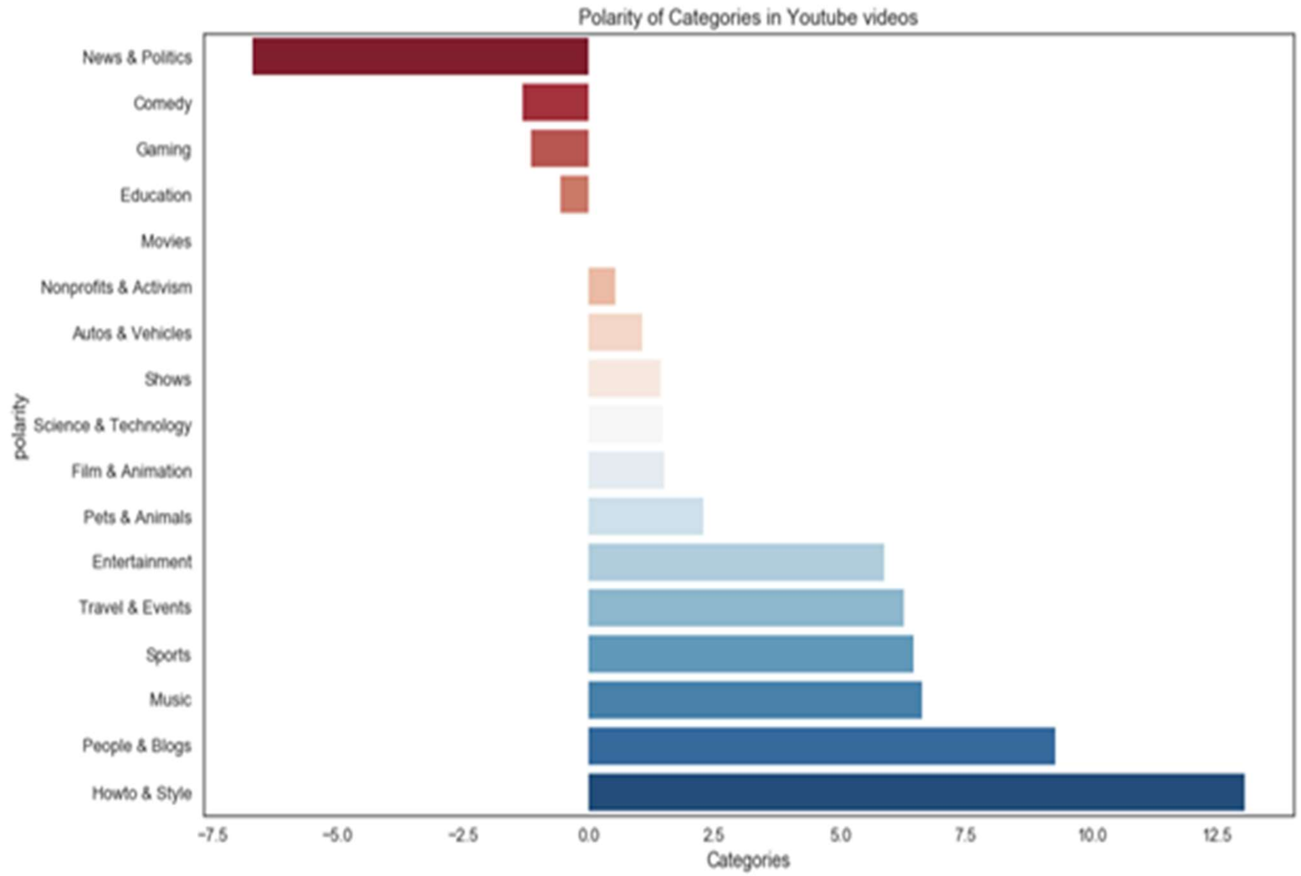
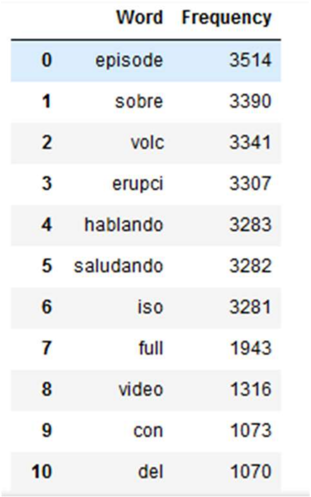


Figure 12 – Polarity of categories in YouTube data

The words of video titles have been tokenized and the frequency of top 10 title words are as following:



	Word	Frequency
0	episode	3514
1	sobre	3390
2	volc	3341
3	erupci	3307
4	hablando	3283
5	saludando	3282
6	iso	3281
7	full	1943
8	video	1316
9	con	1073
10	del	1070

Table 2 – Frequency of top 10 title words

The below plot shows the average number of views based on the most frequent words in video's title. The videos with the words such as "official", "highlights", "series", "new", "latest", "stars" have more average views.

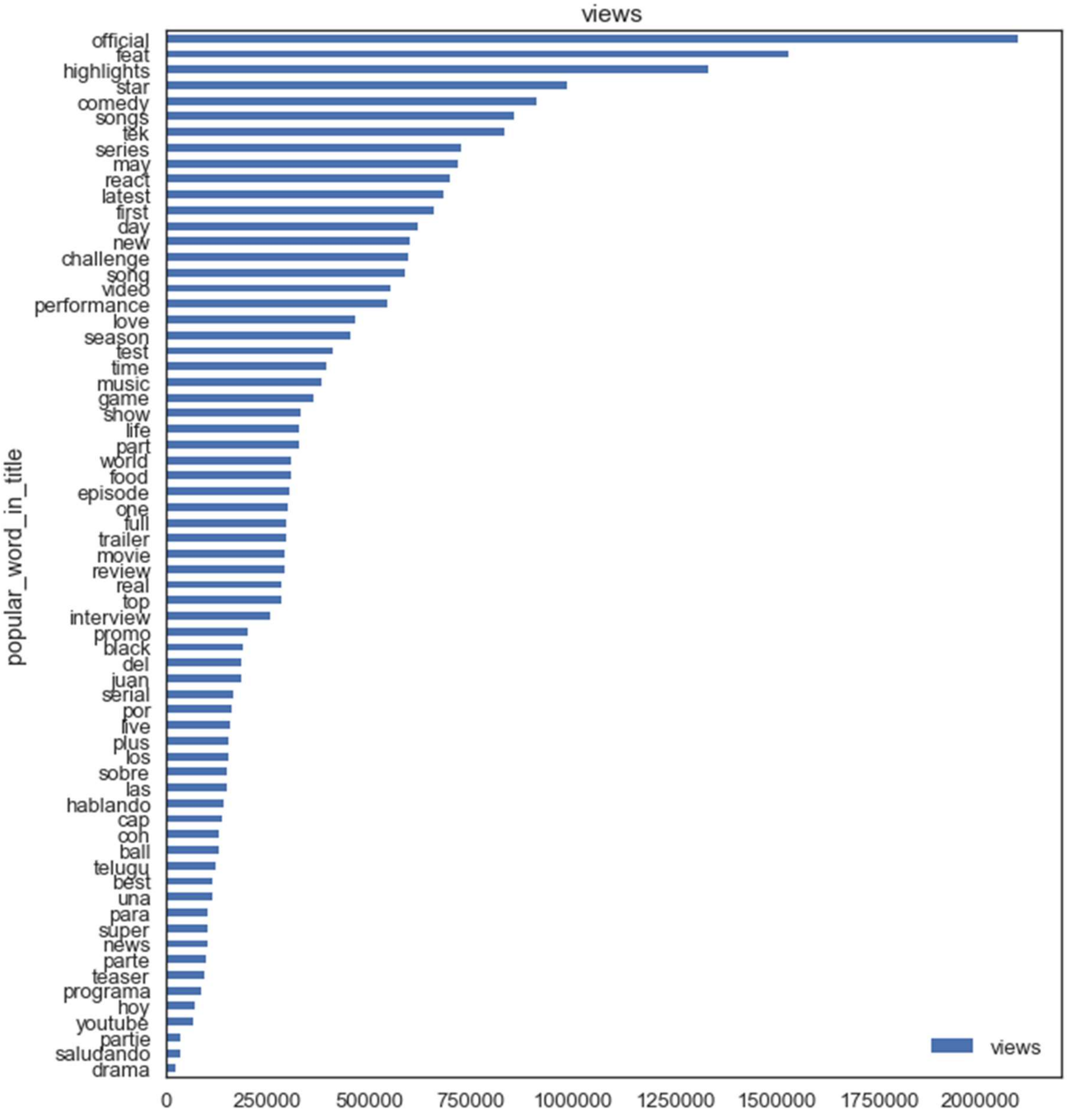


Figure 13 – Mean of views based on popular words

5. Inferential Statistics:

We are using the log of data to calculate the inferential statistics. The first hypothesis is that the views number is normally distributed. The description of the log of view number is as follow:

count	58603.000000
mean	11.766617
std	1.613539
min	6.023448
5%	9.229093
25%	10.563556
50%	11.798398
75%	12.878832
95%	14.393694
max	19.866514

Table 3 – View columns percentiles

The null hypothesis is the log of views is normally distributed. To determine this hypothesis is true or not, we will use the normality test. The p-value of the normality test with assumption the view is normal is 1.1463318993231947e-52. Since the p-value is less than alpha=0.05 then we can reject the null hypothesis which was the view number of dataset distribution is normal.

The plot for log of views histogram has been shown in below figure.

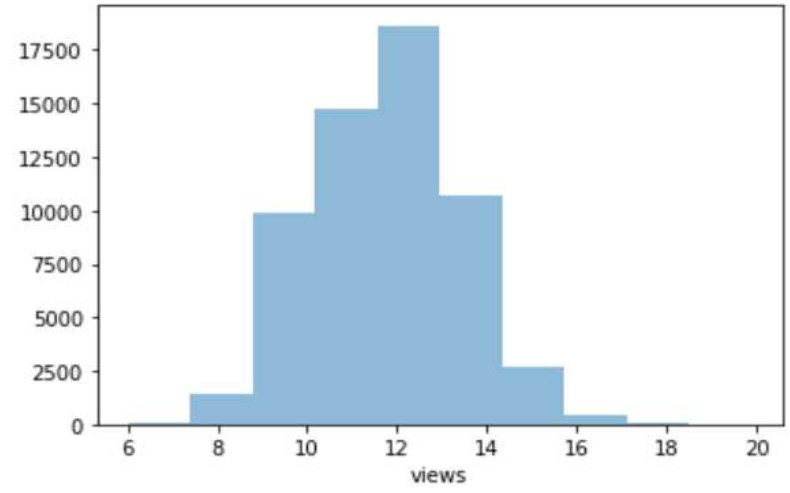


Figure 14 – The histogram of log of views

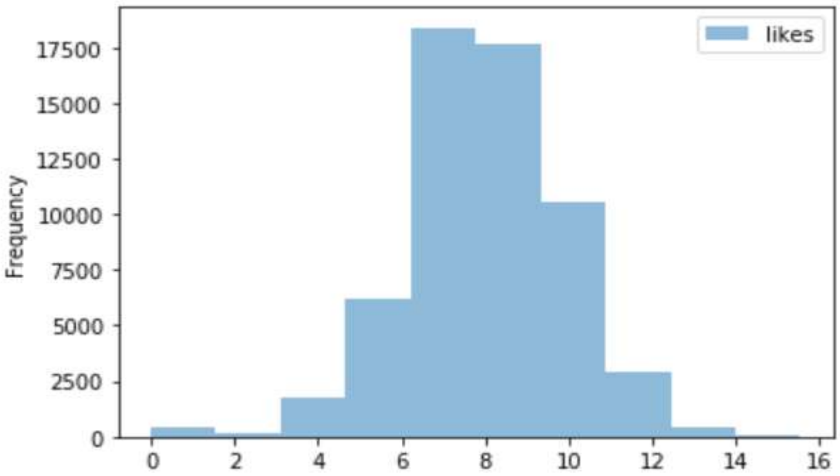


Figure 15 – The histogram of log of likes

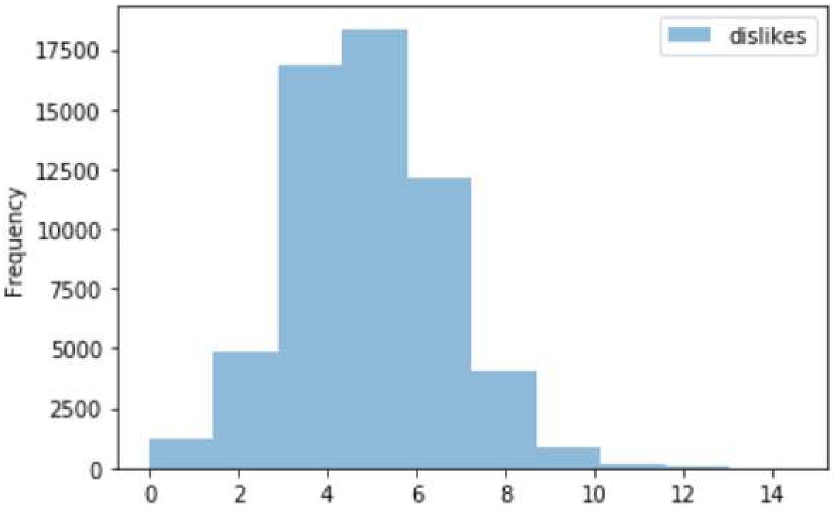


Figure 16 – The histogram of log of dislikes

The second hypothesis is to test the mean number of views is about 500000 so we will take this statement as the null hypothesis. The 99.9 confidence interval for this hypothesis is [501545.218, 587931.356]. We have calculated the T-test are (statistic=2.8527266399048683, pvalue=0.00433610169649154) and (0.4128196675207146, 0.6797387407447166) respectively. Since the used sample size is 10000 Z-test is correct and we can accept the null hypothesis which states the mean number of views is 500000.

The third test is to consider there is significant correlation between ‘likes’ and ‘dislikes’ that are very important features on the view numbers. The following plot is showing the log value of ‘likes’ feature and log value of ‘dislikes’ correlation:

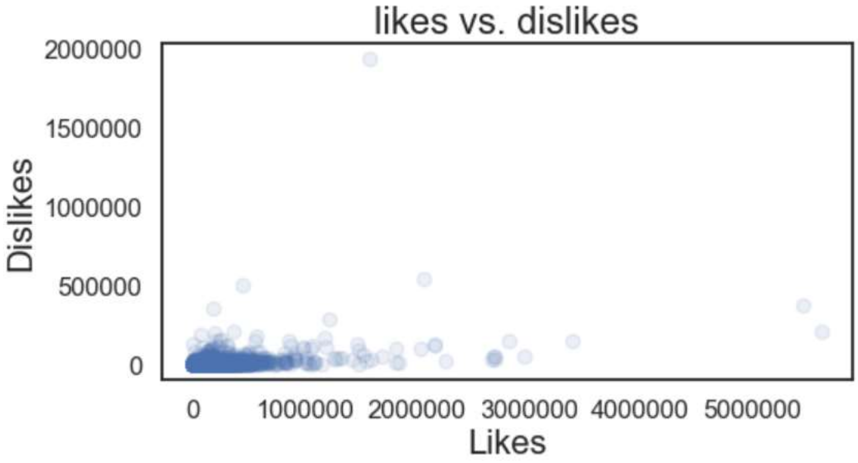


Figure 17 – likes vs. dislikes

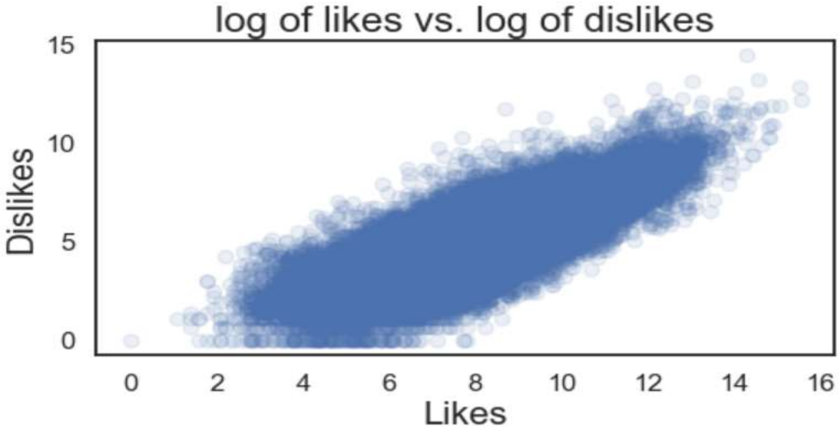
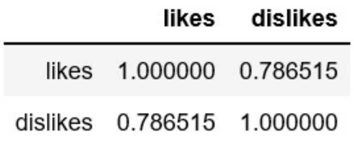


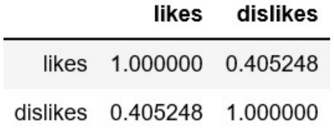
Figure 18 – log of likes vs. log of dislikes

The correlation of these two features log numbers are:



	likes	dislikes
likes	1.000000	0.786515
dislikes	0.786515	1.000000

The correlation of these two features numbers are:



	likes	dislikes
likes	1.000000	0.405248
dislikes	0.405248	1.000000

The null hypothesis is that ‘likes’ and ‘dislikes’ are uncorrelated.

Mean of likes and dislikes correlation	0.405248
Standard Error	0.00377
z-statistics value	107.4602
p-value using z-statistics	2
Margin of Error	0.00043
Confidence interval is between 0.40482 and 0.40567	

6. Machine Learning:

In this section, we will examine various machine learning models to evaluate their performance. The number of views will be used as target variable in the study. Since we want to predict the number of video views, we examine the regression models such as linear regression, random forest, gradient boosting and extreme gradient boosting.

5.1 Data Preprocessing

Data preprocessing consists of two steps. First, the data is split into the training and test datasets with 70/30 split. We have tested the machine learning algorithm on data and logarithm of data.

5.2 Performance Metrics

The performance metrics considered in this study are R2, root mean squared error (RMSE), and mean absolute percentage error (MAPE).

R^2 is a measure of the proportion of data variation explained by the model.

RMSE is the square root of the average of squared difference between the actual value and the predicted value.

MAPE is the average of the absolute percentage difference between the actual value and the predicted value.

5.3 Linear Regression

Linear model is the simplest model of the regression approach that assumes there is a linear regression between features and target values.

$$y\hat{}=\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_nx_n$$

In this equation:

- $\hat{y}$ is the predicted value.
- n is the number of features.
- x_i is the i^{th} feature value.
- θ_j is the j^{th} model parameter (including the bias term θ_0 and the feature weights θ_1, θ_2, $\cdots$, θ_n).

	R^2	RMSE	MAPE (%)
Test data	0.744344940575320	1567947	343.93
Log test data	0.7429298834704685	1.0	5.49

The model performance plot in predication of test data has been in figure 13. The residuals are normally distributed and residual shows increasing trend with the number of views. The outliers show model underpredicts views.

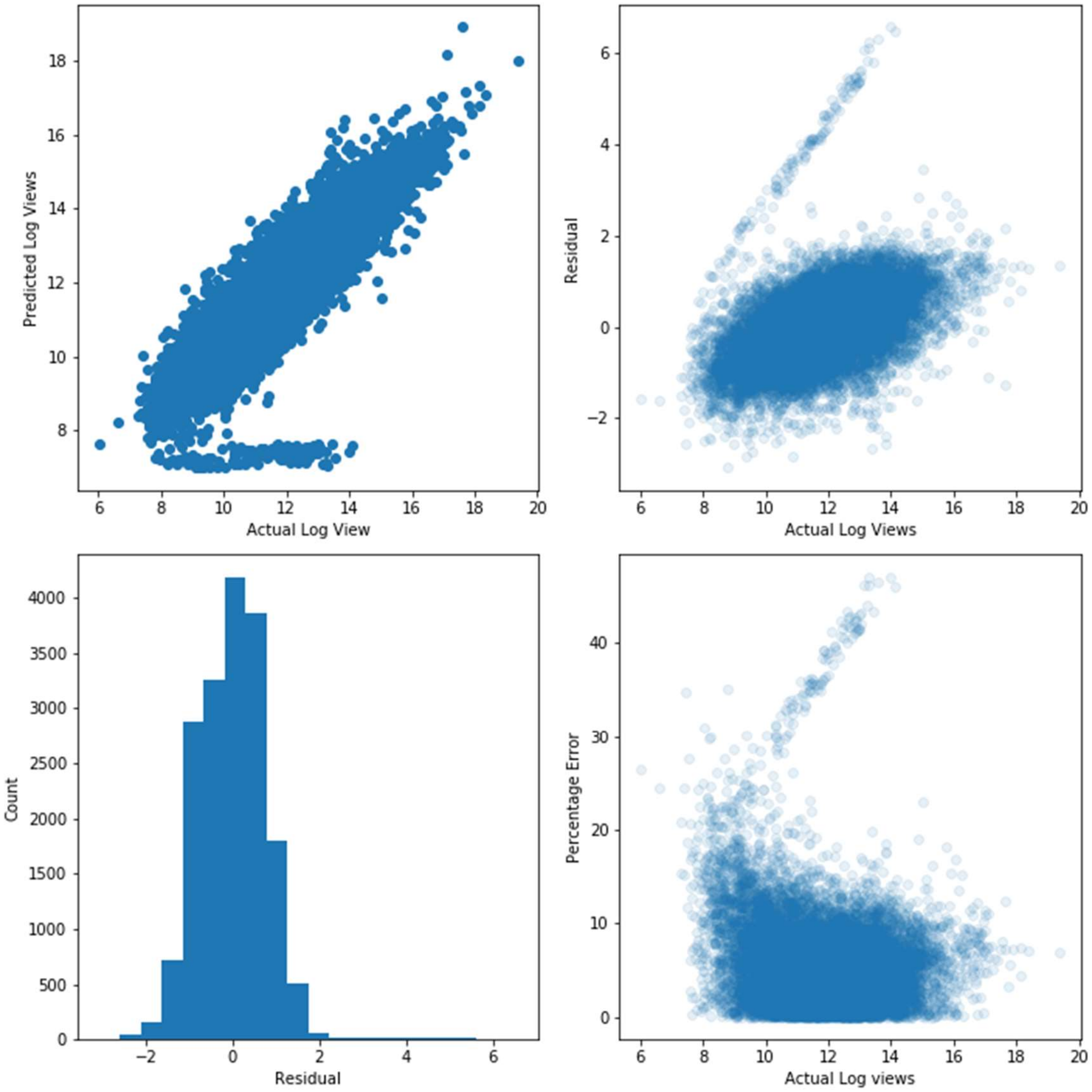


Fig 19 - Linear regression model with the test data

5.3 Random Forest

Random forest is the family of ensemble learning method that are can be used in both classification and regression. This model constructs the multitudes of decision trees in training time and provide the output class as mode of class or mean of predication for classification or regression.

The below table relates the performance metrics for random forest model applied on test data.

	R^2	RMSE	MAPE (%)
Test data	0.5018773244700341	2939401.0	69.9
Log test data	0.8305999576526518	1.0	4.28

The metrics shows that the random forest model is a significant improvement to the linear regression model.

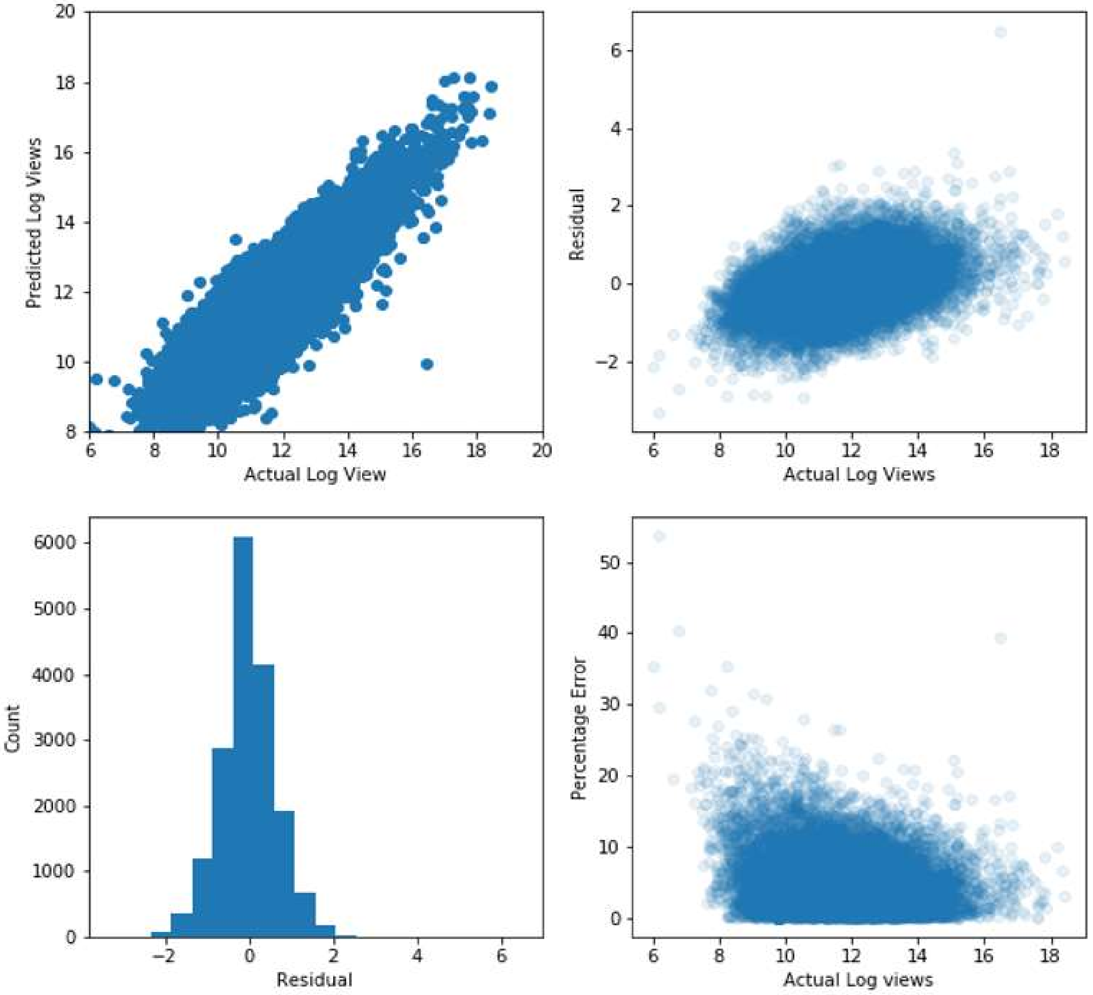


Fig 20 – Random Forest regression model with the test data

5.3 Gradient Boosting

Another very popular boosting algorithm is Gradient Boosting. Gradient Boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor. However, instead of tweaking the instance weights at every iteration like AdaBoost does, this method tries to fit the new predictor to the residual errors made by the previous predictor.

The below table shows the performance metrics for gradient model applied on test data, that the gradient model is a significant improvement to the linear regression model.

	R^2	RMSE	MAPE (%)
Test data	4531591642957319	1766771.0	140.43
Log test data	0.8350626520575262	1.0	4.39

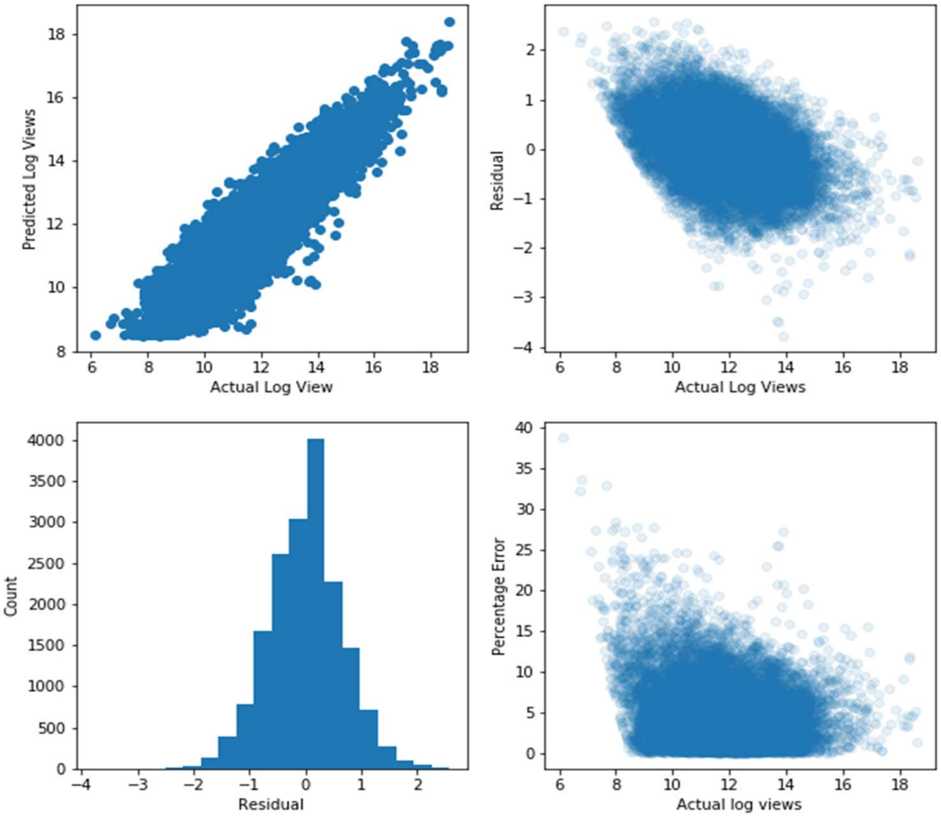


Fig 21 – Gradient Boosting regression model with the test data

5.4 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a recent advancement to the gradient boosting model. Essentially, it adds a regularization component to the algorithm to limit the complexity of trees to prevent overfitting.

Model tuning is performed with cross validation along with grid search.

	R^2	RMSE	MAPE (%)
Log test data	0.8522405976757502	1.0	4.14

The performance metrics for XGBoost and linear regression models on the test data shows that the XGBoost model is a significant improvement to the linear regression model.

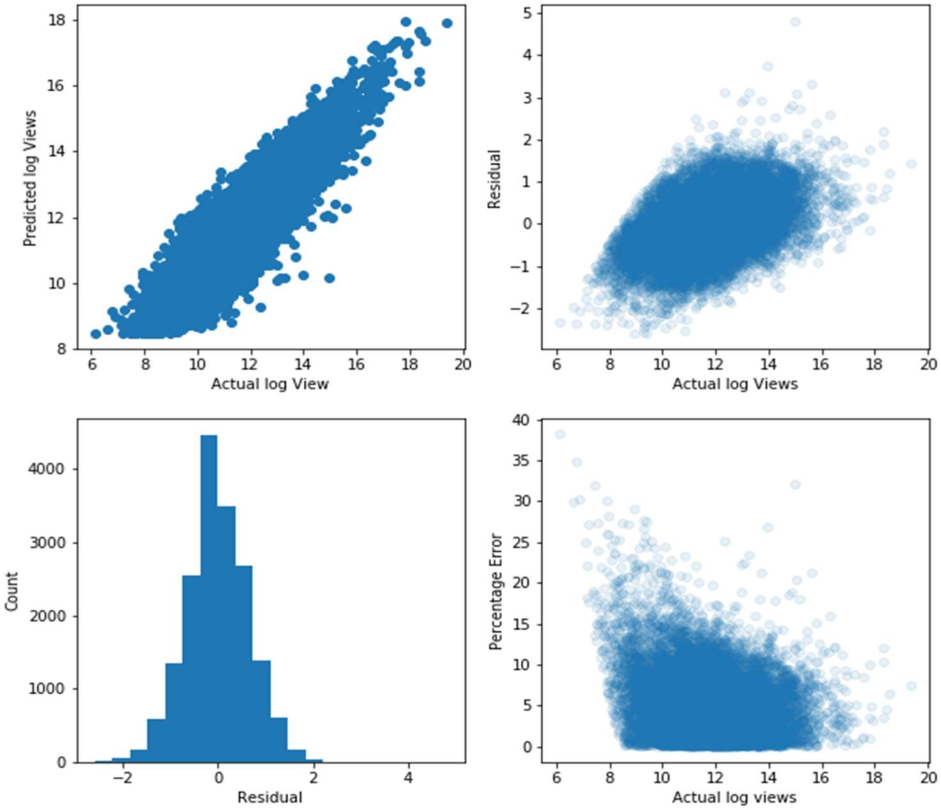


Fig 22 – Extreme Gradient Boosting regression model with the test data

5.7 Model Comparison

The performance metrics for the four models are summarized as following table. The performances of the advanced models are very close, and all of them shows significant improvement to the linear regression model. Overall, the XGBoost model shows the best result in terms of the three metrics.

	R^2	RMSE	MAPE (%)
XGBoost	0.8522405976757502	1.0	4.14
Gradient Boosting	0.8350626520575262	1.0	4.39
Random Forest	0.8305999576526518	1.0	4.28
Linear Regression	0.7429298834704685	1.0	5.49

Feature importance for the extreme gradient boosting model is shown in Figure 5-4. Interestingly, the most important features are the number of dislikes, likes and words number of the title.

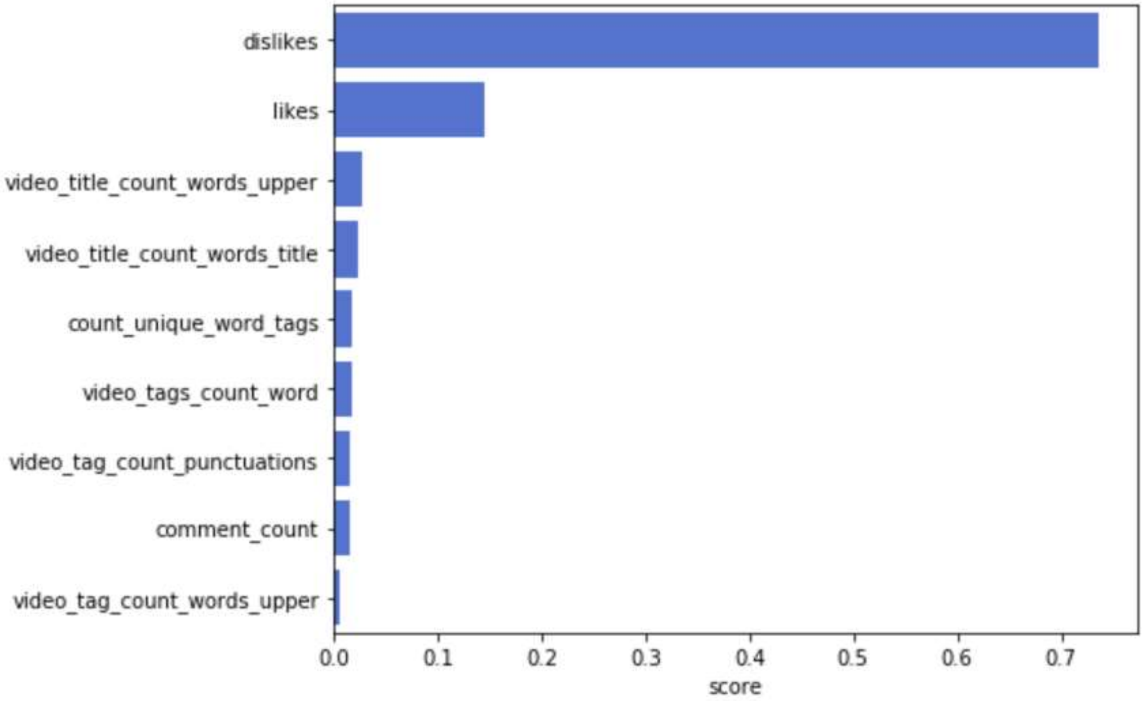


Fig 23 – Feature importance for Extreme Gradient Boosting regression

6. Recommendations and Future Work

In this project, we have examined the YouTube trending video dataset for nine countries through visualization and statistical analysis. We have developed the machine learning algorithms to predict the view number of videos based on the title of video, likes and dislike number of videos. We recommend the YouTube video publishers to ensure using the appropriate title that is related to video. Some more things that we could have tried if we had more time would include:

- Applying sentiment analysis on comments to create a more robust “user profile” that could be used as a feature
- Using neural networks that can learn hidden features