# Analysis of Cross-Country Influence on Trending Youtube Videos

Muskan Goyal(006)
Kritika Rana(014)
Sanya Gupta (028)
Purva Gulati (039)
Muskan Sethi (050)
Kavya Arora (071)

Indira Gandhi Delhi Technical University for Women, Delhi, India
aiproject.team2.2020@gmail.com

**Abstract.** YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they're not the most-viewed videos overall for the calendar year". Top performers on the YouTube trending list are music videos (such as the famously virile "Gangam Style"), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for.

The dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.
EDIT: Now includes data from RU, MX, KR, JP and IN regions (Russia, Mexico, South Korea, Japan and India respectively) over the same time period.
Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.
The data also includes a category-id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

**Keywords:** First keyword · Second keyword · Another keyword.

## 1 Introduction

YouTube represents one of the largest scale and sophisticated industrial recommendation systems in existence. YouTube recommendations are responsible for helping more than a billion users discover personalised content from an ever-growing corpus of videos. One major aspect of this recommendation system is

a list of the top trending videos maintained on the platform. According to Variety magazine, "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes).

YouTube provide access to diverse cultural products from all over the world, making it possible to test theories that the Web facilitates global cultural convergence.

How the factors influence whether a video will be trending or not also varies country wise due to cross cultural differences. In order to understand this influence we investigate the consumption of trending videos across 10 countries namely,
1. Canada
2. Denmark
3. France
4. Great Britain
5. India
6. United States
7. Mexico
8. Japan
9. South Korea
10. Russia

The investigation will consist of an exploratory analysis of the data to reveal various relations between attributes and to analyse how each attribute influences the trending characteristics of a video. Through this analysis we aim to find how different attributes vary across countries and address why some YouTube videos are globally consumed while others are limited to a single country, despite the existence of a technological infrastructure for global cross-cultural communication.

**Problem Statement.** *To find out whether cross-cultural differences(country) influence video popularity.*

## 2 Related Work

We have analysed time taken for a video to trend by using trending date and publish time as our basis of analysis. In addition to this, we've used regression to analyse our data set in many ways and extract useful results from it which help us to make our results much more precise.

## 3 Methodology

### 3.1 Dataset Description

Source: `https://www.kaggle.com/`
Dataset: `https://www.kaggle.com/datasnaek/youtube-new`

Method adopted for collection of this data:This dat aset was collected using the YouTube API
Table 1 describes the data set through counts of some key entities involved in the data set

| **Details** | **Description** |
|---|---|
| Number of instances | 37352 |
| Number of attributes | 16 |
| Whether labeled or unlabeled | Unlabeled |
| Type of label information (if present) | N/A |
| Number of unique videos | 24427 |

**Table 1.** Details of the dataset.

The data set comprises of 16 data attributes.
Table 2 describes attributes of data.

| **Data Attributes** | **Brief Explanation** |
|---|---|
| Video ID(Numeric) | Unique ID to identify the video |
| Trending date(Date) | Date on which video made it to the trending list |
| Title(Categoric) | Name of the video |
| Channel$_t itle(Categoric)$ | Name of channel which posted the video |
| Category$_i d(Numeric)$ | Specific to region |
| Publish$_t ime(Date)$ | Time at which video was uploaded |
| Tags(Categoric) | Words and phrases used to give YouTube context about a video |
| Views(Numeric) | Number of views on the video |
| Likes(Numeric) | Number of likes on the video |
| Dislikes(Numeric) | Number of dislikes on the video |
| Comment$_c ount(Numeric)$ | Number of comments on the video |
| Thumbnail$_l ink(Categoric)$ | Thumbnail link of YouTube video |
| Comments$_d isabled(Boolean)$ | Whether to or not disable comment option |
| Ratings$_d isable(Boolean)$ | Whether to or not disable rating option |
| Video$_e rror_o r_r emoved(Boolean)$ | Either error in playing video or video has been removed from the platform |
| Description(Categorical) | Brief description of the video |

**Table 2.** Details of Data Attributes.

### 3.2 Data Pre-processing

**3.2.1 Data Files Combining** Data was available in the form of 10 .csv files, each file holding the data for one country. These files were combined to form one dataset, with the addition of an addition attribute - 'region'.

**3.2.2 Data Transformations** Data of all attributes was of the type : string. The following changes were done -
1. All numeric columns (Category ID,Views,Likes,Dislikes,Comment Count ) was transformed into type : int.
2. Trending Date column was transformed into datetime format.
3. Publish Time column (consisting of both publish date and publish time) was divided into 2 columns : Publish Date and Publish Time , in the datetime format.

### 3.3 Proposed Approach

**First we focus on exploratory analysis to understand the data. Further,we use the concept of Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). Hypothesis function for Linear Regression :
y= a1 + a2*x
While training the model we are given :
x: input training data (univariate)
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best a1 and a2 values.
a1: intercept
a2: coefficient of x
Once we find the best a1 and a2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

**Algorithm 1** Linear Regression Algorithm

**procedure** (N)
**for** i=1 to n **do** read xi,yi $i \leftarrow i + 1$
**end for**
$sumX \leftarrow 0$ $sumX2 \leftarrow 0$ $sumY \leftarrow 0$ $sumXY \leftarrow 0$
**for** i=1 to n **do**
sumX=sumX+Xi
sumX2=sumX2+Xi*Xi
sumY=sumY+Yi
sumXY=sumXY+Xi*Yi
$i \leftarrow i + 1$
**end for**
For a and b of y = a + bx:
b = (n * sumXY - sumX * sumY)/(n*sumX2 - sumX * sumX)
a = (sumY - b*sumX)/n
Print a,b
**end procedure**

## 4 Exploratory Data Analysis

We will be analysing the data through exploration of the attributes. The aim is to establish patterns, check hypothesis and to spot anomalies with the help of summary statistics and visualisation.
Exploratory Data Analysis is carried out for each attribute

### 4.1 Views, Likes, Dislikes and Comments

We analyse how mean values for the attributes vary country-wise

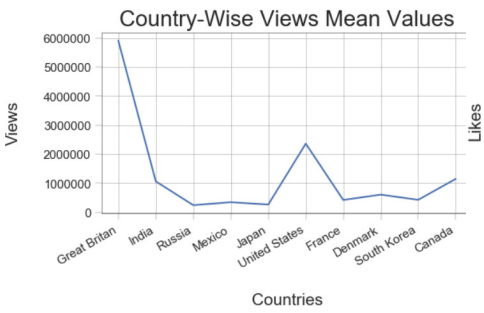



**Fig. 1.** Majority countries have Average Views value of $\leq$ 1 million views.

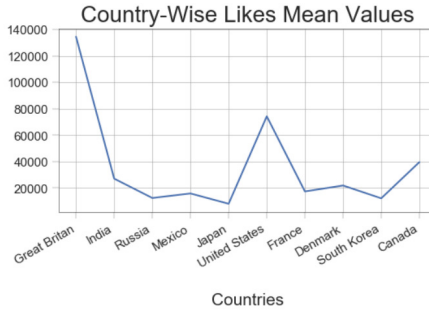



**Fig. 2.** Majority countries have Average Likes value of $\leq$ 50,000 Likes.

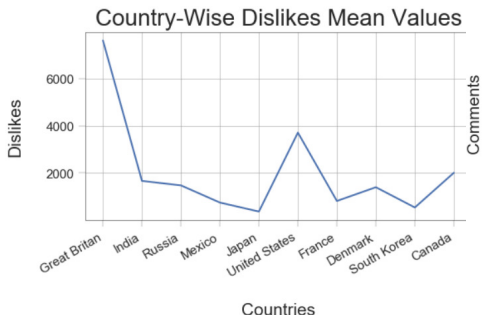



**Fig. 3.** Majority countries have Average Dislikes value of $\leq$ 2000 Dislikes.

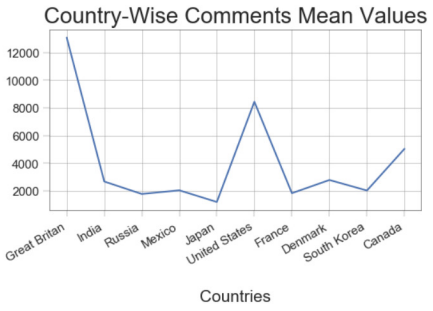



**Fig. 4.** Majority countries have Average Comment Count value of $\leq$ 4000 Comments.

### 4.2 Time to Trend (trending_time - publish_time)

We analyse how the number of days it takes for a video to trend varies country-wise

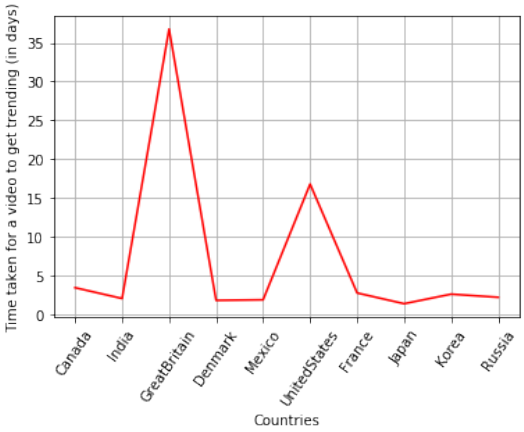



**Fig. 5.** On an average, it takes 7.5 days for a video to trend

*Observations* : Videos from Great Britain take the maximum time to trend followed by United States.

### Correlation Between Attributes

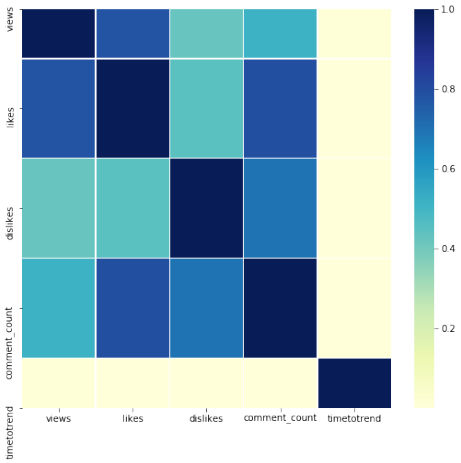



**Fig. 6.** Correlation Matrix

*Observations* :

1. There is a strong positive correlation between the number of likes and the number of comments of trending videos: As one of them increases, the other increases, and vice versa.

2. There is a strong positive correlation also between the number of views and the number of likes.
3. There is a slightly weaker correlation between the number of dislikes and the number of comments.
4. There is no strong correlation between timetotrend and the other attributes.
5. However, timetotrend is most related to number of views when compared to number of views, likes, dislikes and comments.

### 4.3 Categories

We analyse how categories vary country-wise which will help us identify trends which are popular in different regions and also help us see how trends vary across countries.

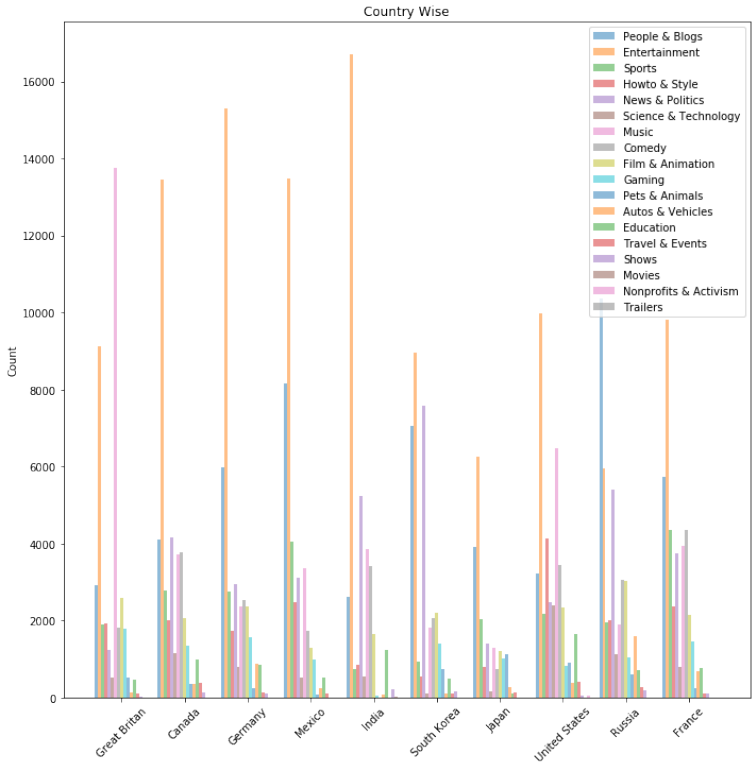



**Fig. 7.** Entertainment is the most popular category across countries with Great Britain and Russia as exceptions where Music and People and Blogs are the most popular categories respectively.

### 4.4 Tags

We analyse the tags and identify the top 10 most common tags which are used in the video description.

| Tag | Frequency |
|---|---|
| '[none]' | *22349* |
| "2018" | *5302* |
| "funny" | *4053* |
| "comedy" | *3019* |
| "news" | *2601* |
| "2017" | *2473* |
| "video" | *1985* |
| "show" | *1814* |
| "television" | *1689* |
| "tv" | *1512* |
| "music" | *1438* |

**Table 3.** Top 10 most common tags - globally.

*Observations*: Most video descriptions do not have any tags.

### 4.5 Title

We analyse the titles to find the top 15 most common words used in trending video titles.

| Character | Frequency |
|---|---|
| '-' | 114594 |
| '—' | 110541 |
| '2018' | 27992 |
| 'The' | 22449 |
| 'de' | 20671 |
| '&' | 15091 |
| 'a' | 10835 |
| '/' | 10805 |
| 'the' | 10802 |
| 'Episode' | 10244 |
| 'in' | 9396 |
| 'A' | 9178 |
| 'of' | 8553 |
| 'Video' | 8540 |
| '2017' | 8327 |

**Table 4.** Top 15 most common characters in video titles - globally

### 4.6 comments_disabled, ratings_disabled

We analyse how user participation affects views.

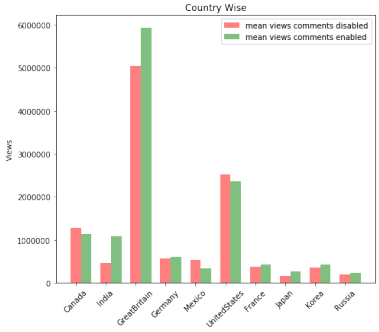



**Fig. 8.** Videos for which comments are enabled get more views.

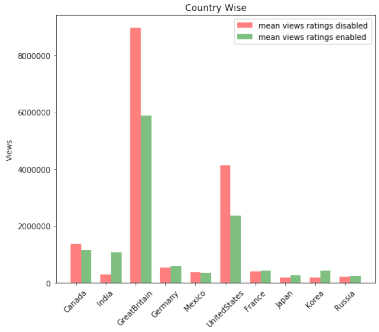



**Fig. 9.** Videos for which ratings are enabled get more views.

*Observations*: Videos with comments enabled and ratings enabled yield more views which verifies our hypothesis that a trending video is a direct result of user participation.

### 4.7 Publish Time

We analyse publish time to identify the time slot with the highest number of uploads and to see how publish time affects user participation and user engagement.

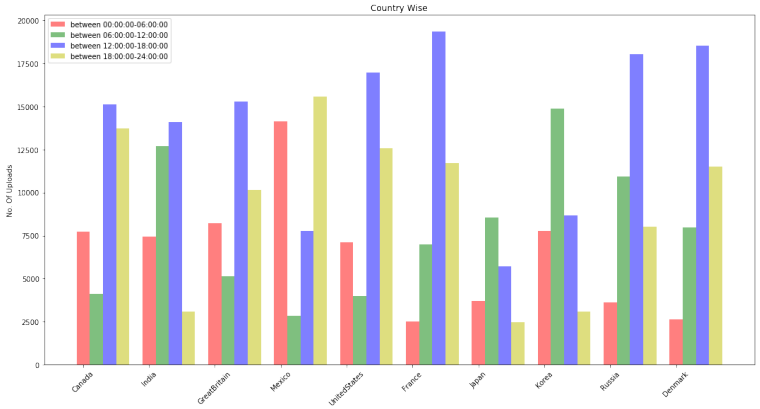



**Fig. 10.** The most popular publish time slot is between 12:00 to 18:00

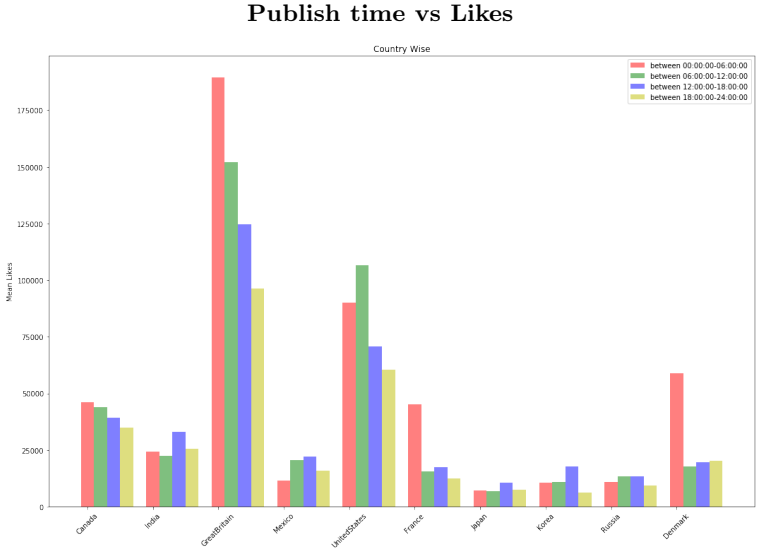



**Fig. 11.** 00:00 to 06:00 time slot garners most likes

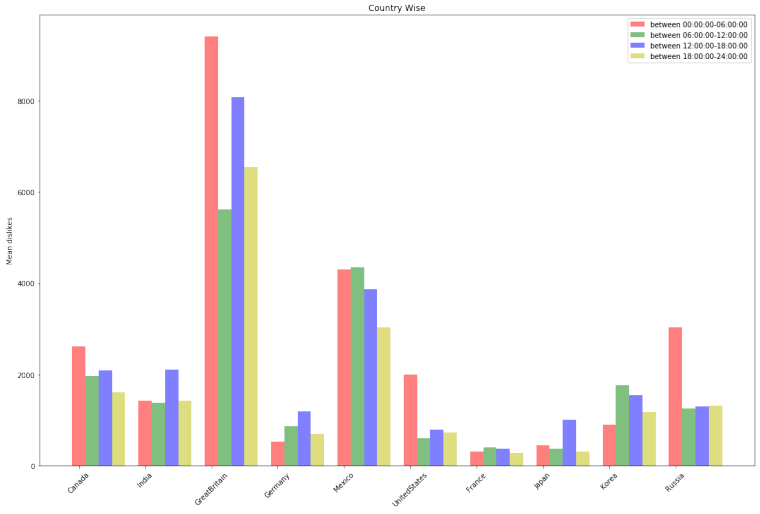



**Fig. 12.** 00:00 to 06:00 time slot garners most dislikes

**Publish Time Vs. Comments**

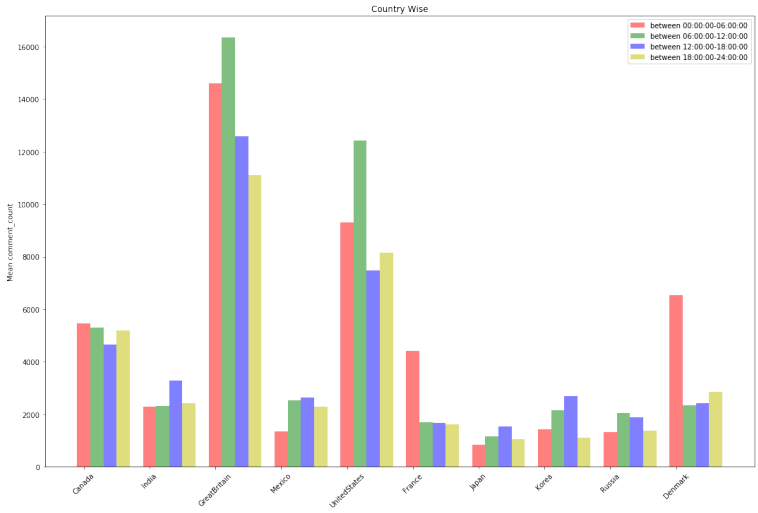



**Fig. 13.** 00:00 to 06:00 time slot and 12:00 to 18:00 time garner most comments

**Publish Time Vs. Views**

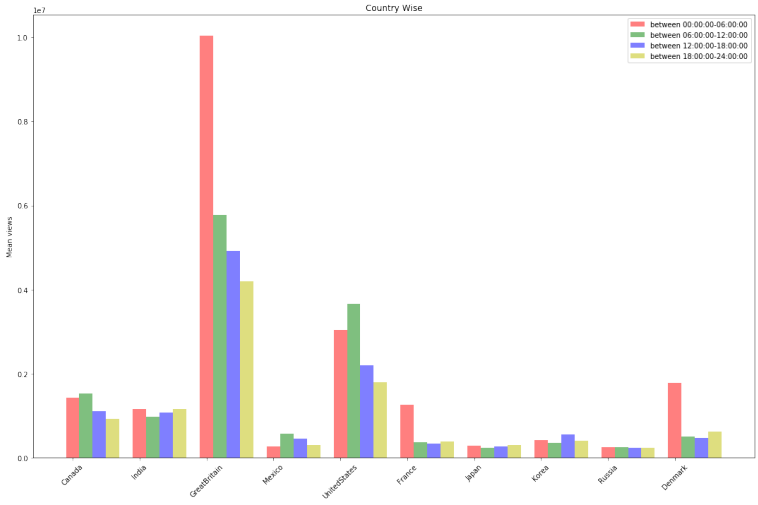



**Fig. 14.** 00:00 to 06:00 time slot garners most views

# 5 Linear Regression

## 5.1 Global Linear Regression

**5.1.1 Model Description :** There are 2 Linear Regression Models considered and compared.

| Model | Inputs/Features | Target |
|---|---|---|
| MODEL 1 | likes,dislikes,comment count, time to trend,publish hour, category ID, tags_count , comments_disabled, ratings_disabled ,video_error_or_removed | views |
| MODEL 2 | likes,dislikes,comment count,time to trend,publish hour, category ID, tags_count , comments_disabled, ratings_disabled , video_error_or_removed, region | views |

**Table 5.** Model 1 is exclusive of region , Model 2 is inclusive of region

**5.1.2 Comparison of Regression Summary :** Comparison of OLS Regression Results of Model 1 and Model 2.

**MODEL 1**

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | log_views | **R-squared:** | 0.796 |
| **Model:** | OLS | **Adj. R-squared:** | 0.796 |
| **Method:** | Least Squares | **F-statistic:** | 1.036e+05 |
| **Date:** | Mon, 06 Apr 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 13:07:40 | **Log-Likelihood:** | -2.4128e+05 |
| **No. Observations:** | 211925 | **AIC:** | 4.826e+05 |
| **Df Residuals:** | 211916 | **BIC:** | 4.827e+05 |
| **Df Model:** | 8 | | |
| **Covariance Type:** | nonrobust | | |

**MODEL 2**

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | log_views | **R-squared:** | 0.831 |
| **Model:** | OLS | **Adj. R-squared:** | 0.831 |
| **Method:** | Least Squares | **F-statistic:** | 6.127e+04 |
| **Date:** | Mon, 06 Apr 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 13:07:54 | **Log-Likelihood:** | -2.2153e+05 |
| **No. Observations:** | 211925 | **AIC:** | 4.431e+05 |
| **Df Residuals:** | 211907 | **BIC:** | 4.433e+05 |
| **Df Model:** | 17 | | |
| **Covariance Type:** | nonrobust | | |

**Fig. 15.** R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model..

*Observations :* R-squared Value of Model 1 = 0.796
R-squared Value of Model 2 = 0.831
Hence, in Model 2 (inclusive of region variable) , the data is fitted much closer to the Regression Line than Model 1. (exclusive of region variable)

Therefore, addition of the region variable in the inputs improves the model.

**5.1.3 Comparison of Regression Coefficients :** Comparison of Regression Coefficients of Features of Model 1 and Model 2.

**MODEL 1**

| | Features | Weights |
|---|---|---|
| 0 | category_id | -0.05 |
| 1 | comments_disabled | 0.00 |
| 2 | ratings_disabled | -0.00 |
| 3 | video_error_or_removed | 0.01 |
| 4 | timetotrend | 0.19 |
| 5 | hour | -0.08 |
| 6 | tag_counts | 0.01 |
| 7 | log_likes | 0.52 |
| 8 | log_dislikes | 1.01 |
| 9 | log_comments | -0.06 |

**MODEL 2**

| | Features | Weights |
|---|---|---|
| 0 | category_id | -0.08 |
| 1 | comments_disabled | 0.00 |
| 2 | ratings_disabled | 0.00 |
| 3 | video_error_or_removed | 0.01 |
| 4 | timetotrend | 0.16 |
| 5 | hour | -0.04 |
| 6 | tag_counts | 0.00 |
| 7 | log_likes | 0.53 |
| 8 | log_dislikes | 1.00 |
| 9 | log_comments | -0.12 |
| 10 | region_Denmark | -0.16 |
| 11 | region_France | -0.19 |
| 12 | region_Great Britan | -0.11 |
| 13 | region_India | -0.02 |
| 14 | region_Japan | -0.07 |
| 15 | region_Mexico | -0.20 |
| 16 | region_Russia | -0.31 |
| 17 | region_South Korea | 0.02 |
| 18 | region_United States | -0.07 |

**Fig. 16.** Higher the absolute value of weights(regression coefficients) of a feature, higher is its influence on the Target value(views).

*Observations :* 1. In both Models, comments_disabled , ratings_disabled , video_error_or_removed , tag_counts - these features have Regression Coefficients as 0.0 - 0.1 , hence they have negligible influence on views.
2. A negative Regression Coefficient indicates that as the value of the Feature decreases, the value of Target increases.
3. A positive Regression Coefficient indicates that as the value of the Feature increases, the value of Target increases.
4. In both the models, dislikes has the highest Regression Coefficient.

### 5.2 Test Results Comparison at different Configurations

**5.2.1 Model Description :** 1. From above results, it is observed that Model inclusive of region is better.
2. Moreover, comments_disabled , ratings_disabled , video_error_or_removed , tag_counts don't influence the target value.

Hence, the final model taken :
1. Inputs = likes,dislikes,comments,timetotrend,publish hour,category ID
2.Target = Views

Now, we will test the best train-test split configuration for the Model by evaluating the error values at different configurations.

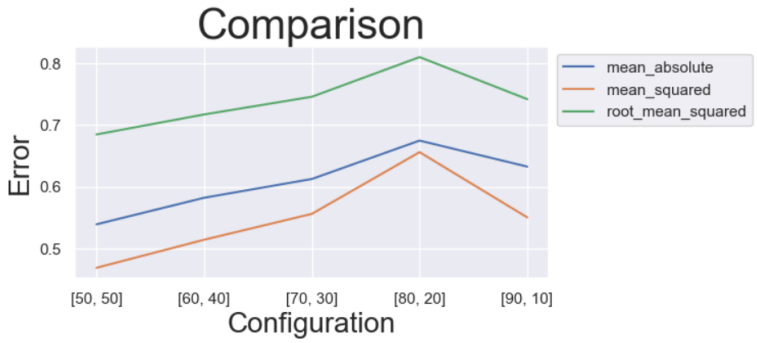



**Fig. 17.** Mean Absolute error,Mean Squared error and Root Mean Squared error for different configurations

**5.2.2 Observations :** From the above graphical representation which shows the mean absolute error,mean squared error and root mean squared error for different configurations, Linear Regression is giving least Mean absolute error, least mean squared error and root mean squared error for the configuration 50:50.

### 5.3 Country-Wise Linear Regression

**5.3.1 Model Description :** The following values of Input,Target and Train-Test Split have been taken for Linear Regression Models of all countries :
1. Inputs = likes,dislikes,comments,timetotrend,publish hour,category ID
2.Target = Views
3. Train-Test Split = 50:50

Now , we will compare Regression Coefficients of all features, country-wise.

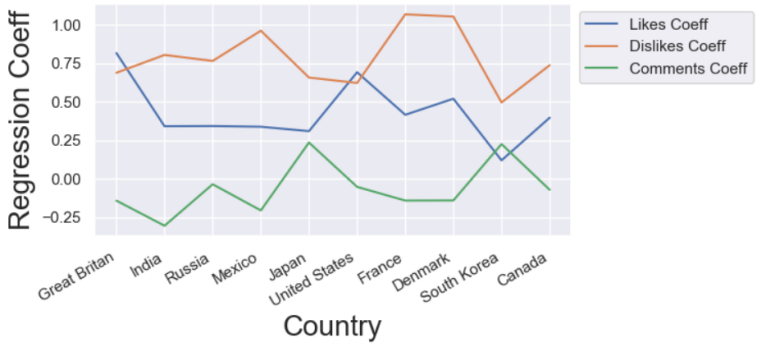



**Fig. 18.** Regression Coeff. for Likes, Dislikes , Comment Count

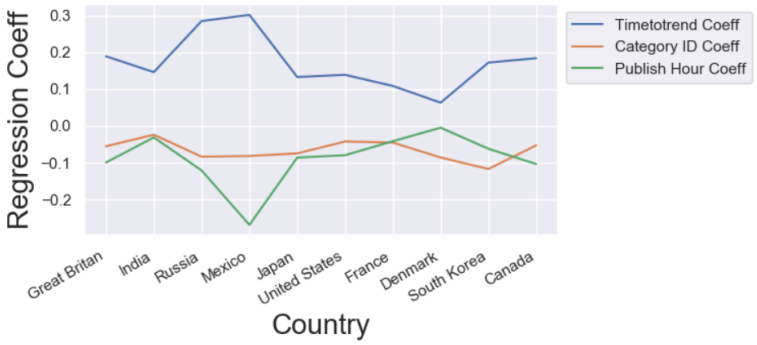



**Fig. 19.** Regression Coeff. for TimeToTrend , Category ID , Publish Hour

**5.3.2 Observations :** 1. Great Britain has the highest value for Likes Regression Coefficient, hence number of likes has the largest influence on number of views in Great Britain.
2. France has the highest value for Dislikes Regression Coefficient.
3. Japan & South Korea have the highest value for Comment Count Regression Coefficient.
4. Mexico has the highest value for TimeToTrend Regression Coefficient.
5. India has the highest value for Category ID Regression Coefficient.
6. Denmark has the highest value for Publish Hour Regression Coefficient.

# 6 Appendix

## 6.1 Results of analysis of the following attributes of a trending YouTube video

- Number of Views
- Number of Likes
- Number of Dislikes
- Number of Comments
- Time taken to trend

### Visual Analysis of all attributes

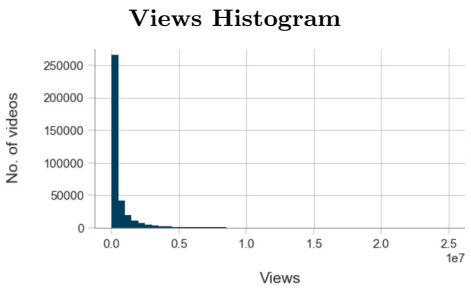



**Fig. 20.** Note that majority of the trending videos have 1 million views or less.

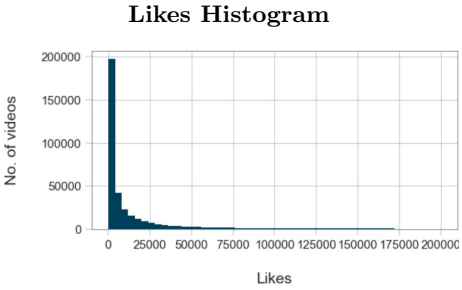



**Fig. 21.** Note that majority of the trending videos have 50,000 likes or less.

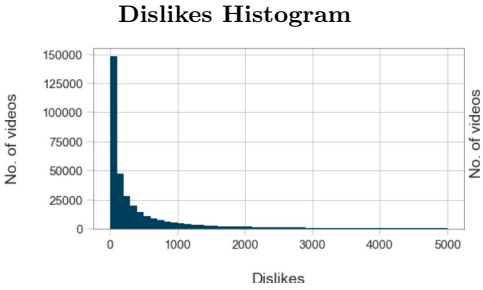



**Fig. 22.** Note that majority of the trending videos have 2000 dislikes or less.

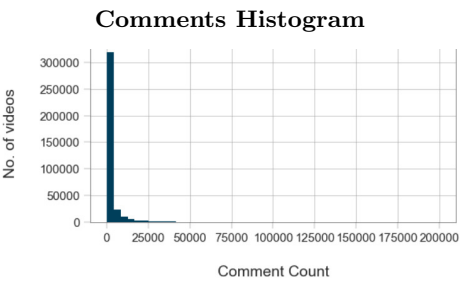



**Fig. 23.** Note that majority of the trending videos have 4000 comments or less.

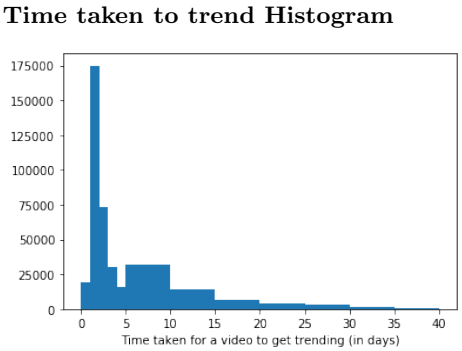



**Fig. 24.** Note that majority of the trending videos take 1 day to get trending.

## 6.2 Country-wise ratio of likes:views, dislikes:views , comments:views

*Please note:*
1. Likes:views = x means that on average, x of total viewers like the video.
2. Dislikes:views = x means that on average, x of total viewers dislike the video.
3. Comments:views = x means that on average, x of total viewers leave a comment on the video.

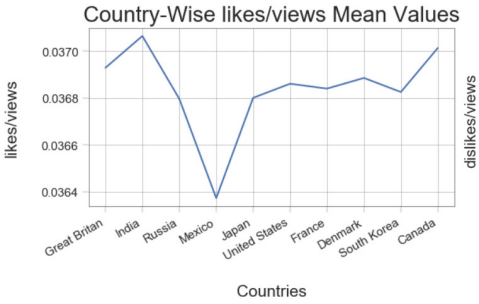



**Fig. 25.** India , on average, has the highest ratio of likes/views.(>0.0370)

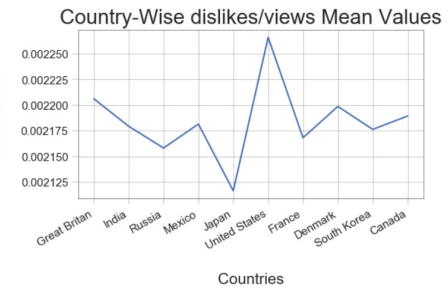



**Fig. 26.** US , on average , has the highest ratio of dislikes/views.(>0.002250))

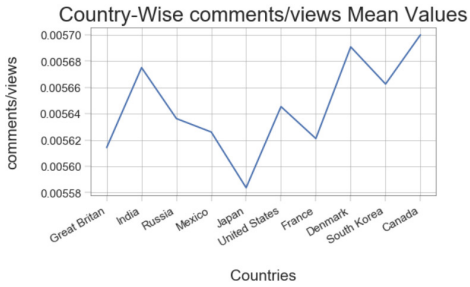



**Fig. 27.** Canada , on average , has the highest ratio of comments/views.(0.00570)

### 6.3 Global Analysis

**RESULT 1 : Analysis of Publish Time of videos Globally**

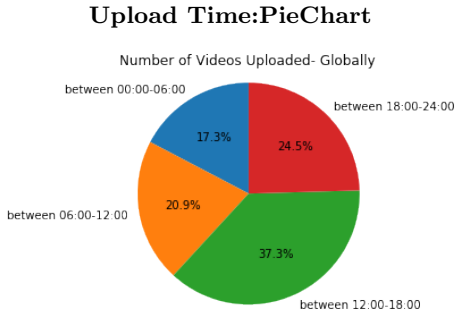



**Fig. 28.** Note that majority of the videos are uploaded b/w 12:00-18:00.

**RESULT 2 : Analysis of Likes on videos Globally**

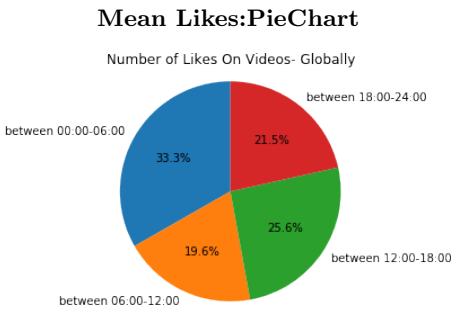



**Fig. 29.** Note that a video gets maximum likes in the time slot of 00:00-06:00.

**RESULT 3 : Analysis of Dislikes on videos Globally**

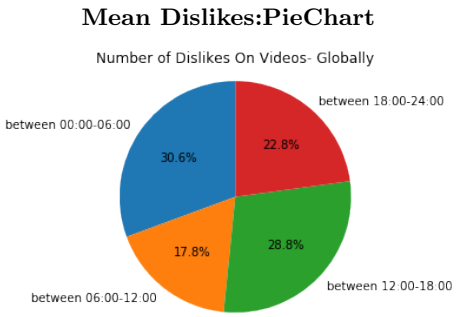



**Fig. 30.** Note that a video gets maximum dislikes in the time slot of 00:00-06:00.

**RESULT 4 : Analysis of Comments on videos Globally**

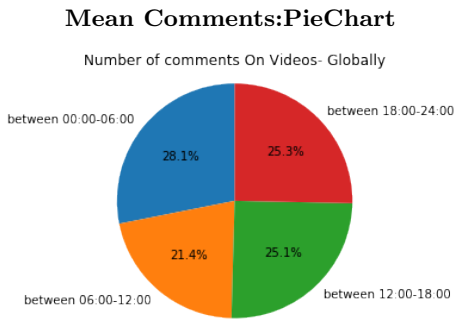



**Fig. 31.** Note that a video gets maximum comments in the time slot of 00:00-06:00.

## 6.4 Results on Comments Disabled and Ratings Disabled

**Views Vs. comments disabled and ratings disabled**
There are 4 possible cases:
1. Both ratings and comments are disabled
2. Ratings are disabled but comments are enabled
3. Ratings are enabled but comments are disabled.
4. Both ratings and comments are enabled

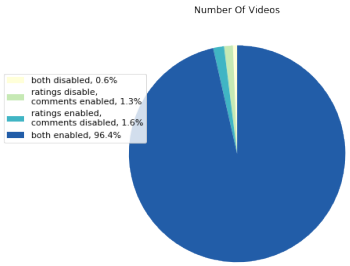



**Fig. 32.** Number of Videos

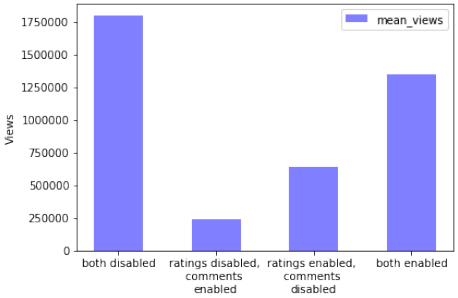



**Fig. 33.** Mean Views for each case

Since number of instances is much greater for case4(both enabled), mean views doesn't give us an accurate picture which may lead to faulty results. In order to deal with that, we consider total views and max views.

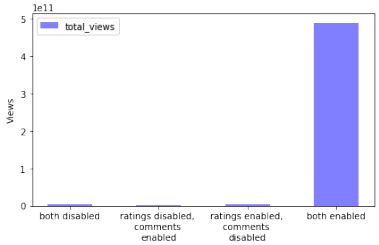



**Fig. 34.** Total Views

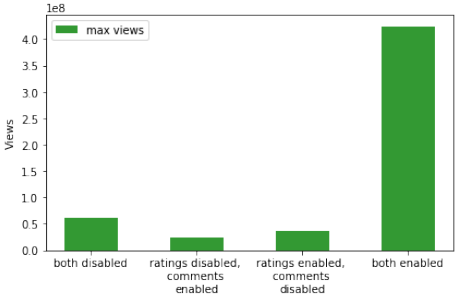



**Fig. 35.** Max Views

Ratio of total number of views on videos with both comments and ratings disabled to total number of views on videos with both ratings and comments enabled = 0.008683448502212093
This implies that, number of views on videos with both comments and ratings disabled ¡ number of views on videos with both comments and ratings enabled.

Ratio of max number of views on videos with both comments and ratings disabled to max number of views on videos with both ratings and comments enabled = 0.1468378050584913 This implies that, max number of views on videos with both comments and ratings disabled ¡ max number of views on videos with both comments and ratings enabled.

### Time to Trend Vs. Comments disabled and Ratings disabled

Time to Trend is the the difference between the date on which the video first got trending and the date on which the video was published. The aim of this analysis is to see the extent to which user participation affects how long it will take for a video to get trending.

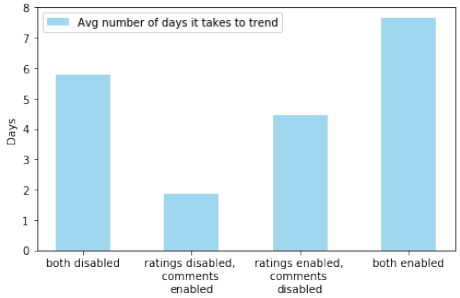



**Fig. 36.** Average number of days it takes for a video to get trending.

### 6.5 Analysis of the Title containing Capitalized words

Results of analysis on basis of two factors:

- No, as in the tag does not contain Capital Words
- Yes,as in the tag contains Capital Words

**RESULT: Observations :**

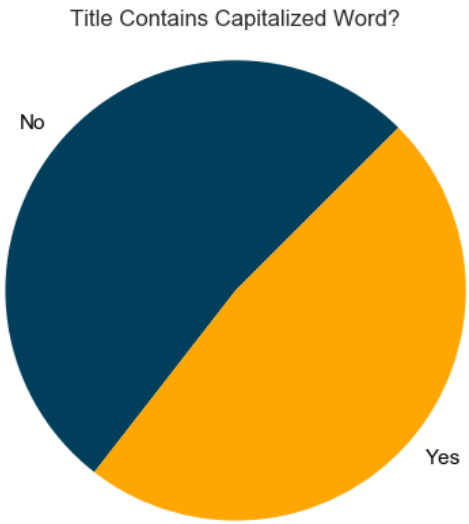
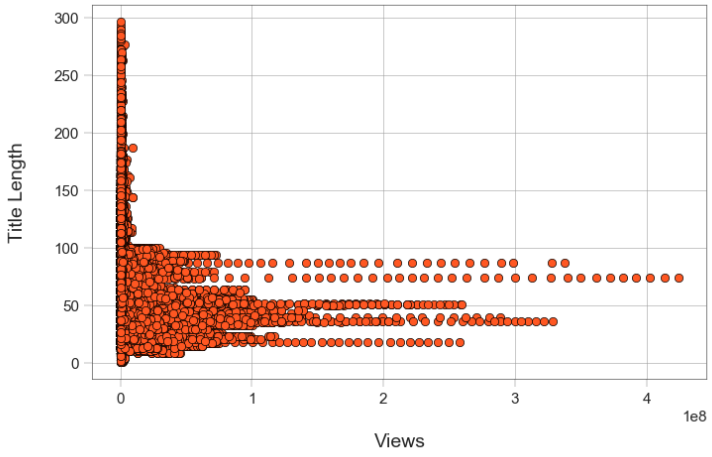



**Fig. 37.** This graph tells us about whether the title contains Capital words or not.

False 0.52
True 0.48
It can be seen that 52 percent of trending video titles contain at least a capitalized word and 48percent of the video titles does not contain any Capital word.

**Analysis between Title length and Number of views.**
A scatter plot between title length and number of views is drawn to see the relationship between these two variables.

By looking at the scatter plot, it can be said that there is no relationship between the title length and the number of views.
However videos that have 100,000,000 views and more have title length between 33 and 65 characters approximately.




## 7 Future Work and Conclusion

Finally write the future work and conclusion.

## References