Youtube Video Trending

Kamal Rimal
Georgia State University
Atlanta, Georgia, USA
krimal1@student.gsu.edu

Yash Shah
GeorgiaState University
Atlanta, Georgia, USA
yshah4@student.gsu.edu

Hama Bako
Georgia State University
Atlanta, Georgia, USA
hbako1@student.gsu.edu

***Abstract*—https://www.kaggle.com/datasnaek/youtube-new.T he datasets we are going to study consists of YouTube videos that are trending on a specific date in different developed countries such as the USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, India, and Japan. The different attributes of this particular dataset are video_id, trending_date, title, channel title, category, publish time, tags, number_of_views,number_of_likes,number_of_dislikes,comme nts_count,thumbnail_link,comments_disables,ratings_disables, video_error_or_removed, and description.**

Keywords—videos, likes, dislikes, ratings, views, categories

I. Introduction

We are going to study the behavioral pattern of the population in these different countries and see how they vary and compare to each other, observe how the trend spreads, analyze how different categories of videos get different ranges of views. We also see what makes a video trend and if they are a channel that consistently produces trending videos.

II. Methodology

To finish our project we we followed our proposal to determine the questions we created. The procedure has different phases.

III. Research Questions

The main goal of our project is to provide the research efforts focusing on the all kind of data analysis on the YouTube videos and get some fruitful information based on the videos.

- Question 1: What will the category of a video based on the title of the video?
- Question 2: Based on the category of the videos, can we analyze and how many views the category has ?
- Question 3: We are going to see if there are any patterns with the specific day of the week and a category of videos that are trending.
- Question 4: Is there a consistent ratio of likes to dislikes of the trending videos, compared to other countries?
- Question 5: How long does a video take to become trending on average?
- Question 6: Is there a pattern in videos becoming trending in the US first and later, the trend spread to the rest of these developed nations for example?
- Question 7: Are the channels that have recurring trending videos, and is there a pattern weekly, monthly?

IV. Data Acquisition

The dataset we used for our project from Kaggle which is available for free to use. The dataset is 500MB and there it has two different files which are jason and csv.We preprocess the dataset using python and the IDE we have used is Spyder. Dataset is pretty clean so we just remove some extra column whichever we have not used it.

V. Analysis and Result

A. Question 1: What will the category of a video based on the title of the video?

For this question we used Machine Learning to predict the category of the video based on a video title.
First we used the json file to import the categories that were not part of the csv files. Then created a dataframe called “categories”. There are two columns 1. Category_# and 2. Category.
Next was training the data using Naive Bayes. Teh accuracy we got was around 0.75.

The final step was testing with a made up video title. Here is an example with the title “Kobe Bryant” and the predicted category was sports.
Look at the references tab to see the resources used.

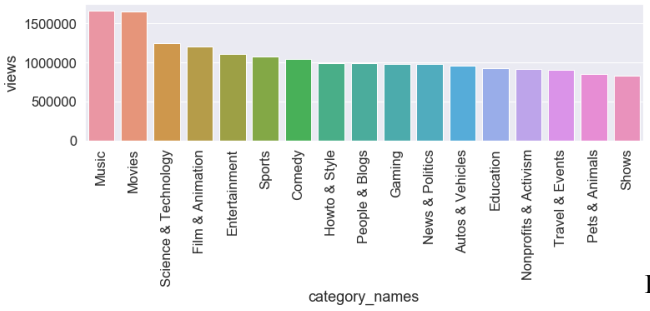
```
#Creates a title to test
Titles = ["Kobe Bryant"]
```

```
#Testing

#Inserting above title into classifier model
Titles_enumeration = countVectorizer.transform(Titles)

#Naive Bayes Model
PredictNB = NaiveBayesModel.predict(Titles_enumeration)
PredictNB

#Output will be an array of numbers. Iterate through the Category Dictionary (from JSON file) to find "title"
NameList = []
for Category_ID in PredictNB:
    CatChecker = [x for x in category_dict if x["id"] == str(Category_ID)]
    if CatChecker:
        NameList.append(CatChecker[0]["title"])

#Mapping the values to the Titles we want to Predict
TitleDataFrameNB = []
for i in range(0, len(Titles)):
    TitlestoCat = {'Title': Titles[i],  'Category': NameList[i]}
    TitleDataFrameNB.append(TitlestoCat)

#Changing the dictionary to a dataframe
PredictDFnb = pd.DataFrame(PredictNB)
TitleDFnb = pd.DataFrame(TitleDataFrameNB)
PreTitle_Aticipate = pd.concat([PredictDFnb, TitleDFnb], axis=1)
PreTitle_Aticipate.columns = (['Categ_ID', 'Predicted Category', 'Made-Up Video Title'])
Title_Aticipate = PreTitle_Aticipate.drop(['Categ_ID'],axis=1)
LastClassifer = Title_Aticipate.columns.tolist()
LastClassifer = LastClassifer[-1:] + LastClassifer[:-1]
Title_Aticipate= Title_Aticipate[LastClassifer]


Title_Aticipate
```

	Made-Up Video Title	Predicted Category
0	Kobe Bryant	Sports

B. Question 2: Based on the category of the videos, can we analyze and how many views the category has ?

For this question we used a barplot to plot all the categories vs the total amount of views the category had. Music had the most views overall. Movies had the second most views overall. Science and technology had the third most views overall. Film and animation had the fourth most views overall. Entertainment had the fifth most views overall. Sports had the fifth most views overall. Comedy had the sixth most views overall. Howto and style had the seventh most views overall. People and blogs had the eighth most views overall. Gaming had the ninth most views overall. News and politics had the tenth most views overall. Autos and vehicles had the eleventh most views overall. Education had the twelfth most views overall. Nonprofits and activism had the thirteenth most views overall. Travel and events had the fourteenth most views overall. Pets and animals had the fifteenth most views overall. Shows had the sixteenth most views overall.



C. Question 3: We are going to see if there are any patterns with the specific period or day of the week and a category of videos that are trending.

Two periods were picked to answer that question. The month of December 2017 (The celebration associated with Christmas and New Year) and the month of January 2018 (The fresh start, resolutions, and goal oriented aspect of the beginning of the year).

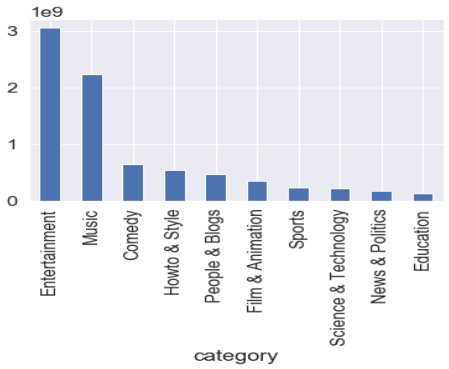


Fig. Bar Chart of December 2017 top 10 youtube video categories in terms of views

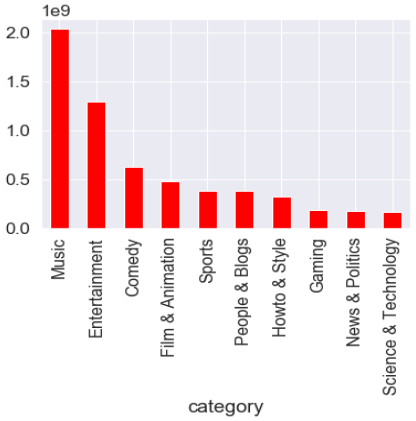


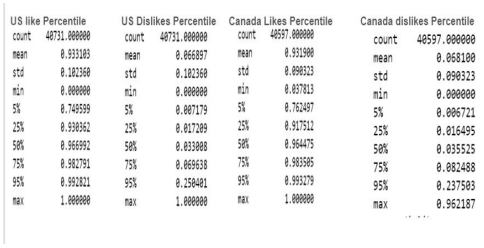
Fig. Bar Chart of January 2018 top 10 youtube video categories in terms of views.

As we can observe, there is a decrease of about 1 billions views for the music category and a decrease of about half a billion views for the entertainment category whereas the sport the sport category sees an increase of about a quarter of a billion views.
Physical exercise and take care of our body is one of the goals that keeps coming back every year for people and this increase shows the transition in the mindset of people.

D. Question 4: Is there a consistent ratio of likes to dislikes of the trending videos, compared to other countries?

The trending videos of USA and Canada has taken to see the ratio of the like and dislike between two different countries. The overall likes to dislikes ratio of the all country is different. Most of the videos has more likes than dislikes. After that the USA's videos compare with the Canada's

video to find out the proper ratio of the likes and dislikes. The likes and dislikes looks like consistent because both countries has more videos in likes sides then dislikes. In the research the videos between 12/2017 to 6/2018 is compared to compare. In both countries the trend of likes and dislikes is little different. In the USA the likes trend goes up by the different date scenario than Canada. In Canada the likes trend is up and down in all aspects.



US like Percentile		US Dislikes Percentile		Canada Likes Percentile		Canada dislikes Percentile	
count	40731.000000	count	40731.000000	count	40597.000000	count	40597.000000
mean	0.933103	mean	0.066897	mean	0.931900	mean	0.068100
std	0.102360	std	0.102360	std	0.090323	std	0.090323
min	0.000000	min	0.000000	min	0.037813	min	0.000000
5%	0.749599	5%	0.007179	5%	0.762497	5%	0.006721
25%	0.930362	25%	0.017209	25%	0.917512	25%	0.016495
50%	0.966992	50%	0.033008	50%	0.964475	50%	0.035525
75%	0.982791	75%	0.069638	75%	0.983505	75%	0.082488
95%	0.992821	95%	0.250401	95%	0.993279	95%	0.237503
max	1.000000	max	1.000000	max	1.000000	max	0.962187

Fig. USA and Canada's Likes Dislikes Percentile

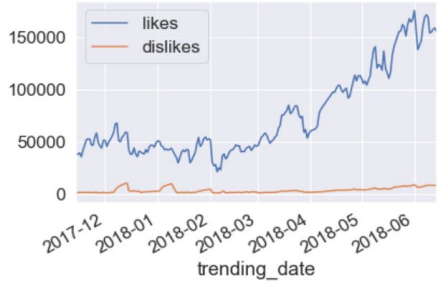


Fig. Table plot of the USA's likes Dislikes Trending Ratio

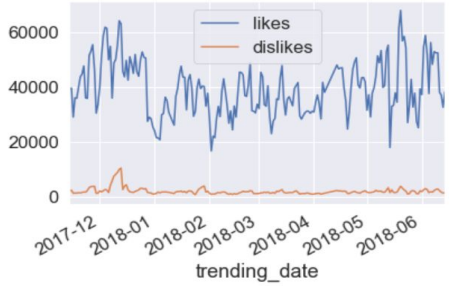
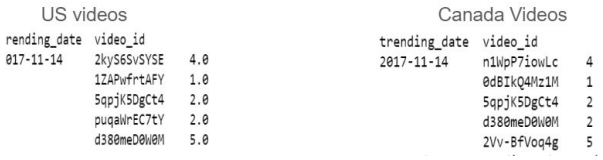


Fig.Table plot of a Canada's Likes dislikes Ratio

E. Question 5: How long does a video take to become trending on average?

In our finding the average time is at least one day to trend the videos. There are different videos which trends for multiple days. In this research the video takes at least one day to trend and become the trending videos. We take a sample of the two counties to see the average time of trending. In both countries it takes at least one day to trend the videos.



US videos

rending_date	video_id	
017-11-14	2kyS6SvSYSE	4.0
	1ZAPwfrtAFY	1.0
	5qpjK5DgCt4	2.0
	puqaWrEC7tY	2.0
	d380meD0W0M	5.0

Canada Videos

trending_date	video_id	
2017-11-14	n1NpP7iowLc	4
	0dBIkQ4Mz1M	1
	5qpjK5DgCt4	2
	d380meD0W0M	2
	2Vv-BfVoq4g	5

Fig. The trending status of the videos in the USA and Canada

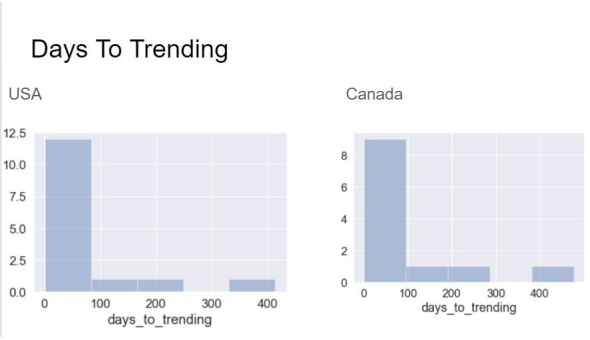


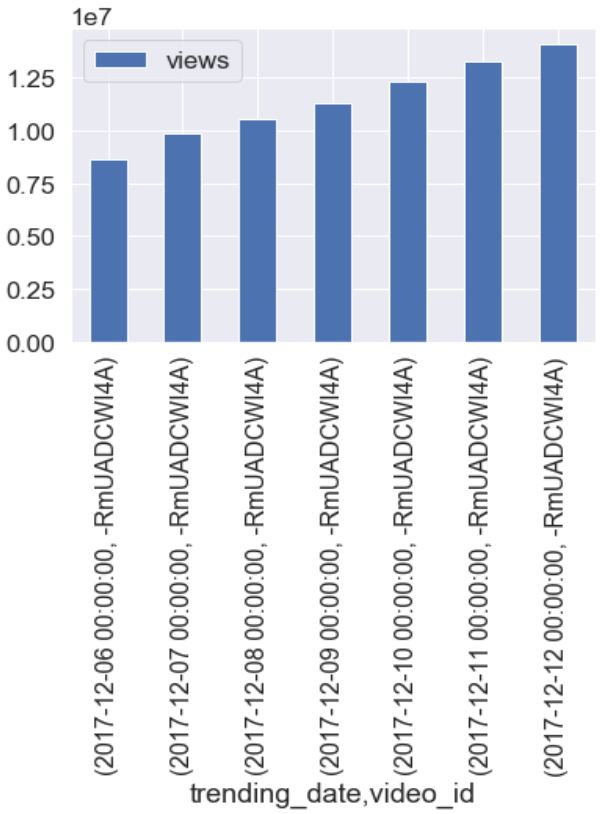
Fig. Distplot of trending videos

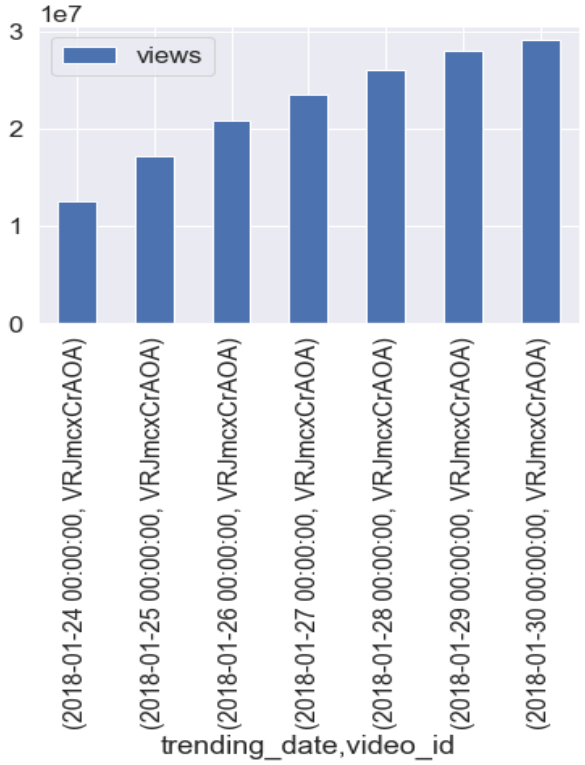
F. Question 6: Is there a pattern in videos becoming trending in the US first and later, the trend spread to the rest of these developed nations for example?

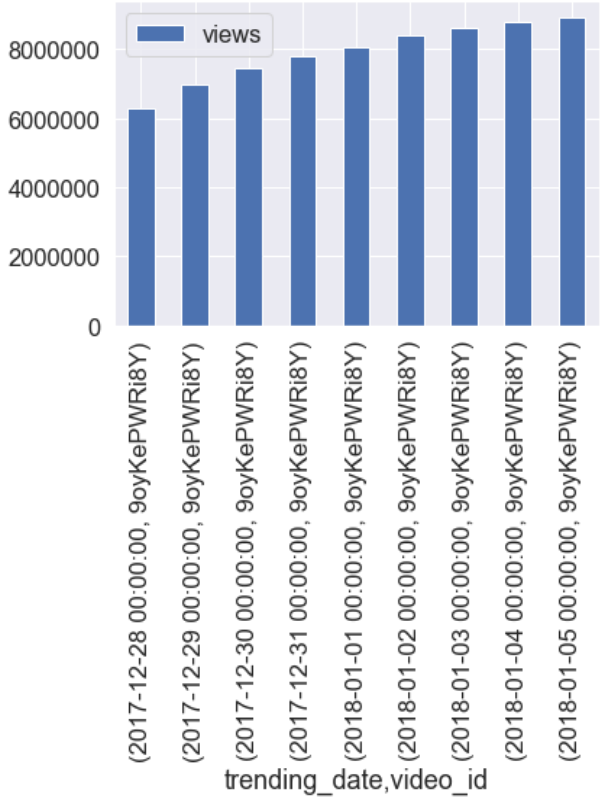
To answer this question, the top 10 videos that are trending in the US, Canada, Great Britain, France, Germany, and India were selected.
No good conclusion were drawn from this question because no real pattern was found. Also the visualisation were meaningless because I could assimilate the date correctly to make a good visualisation.

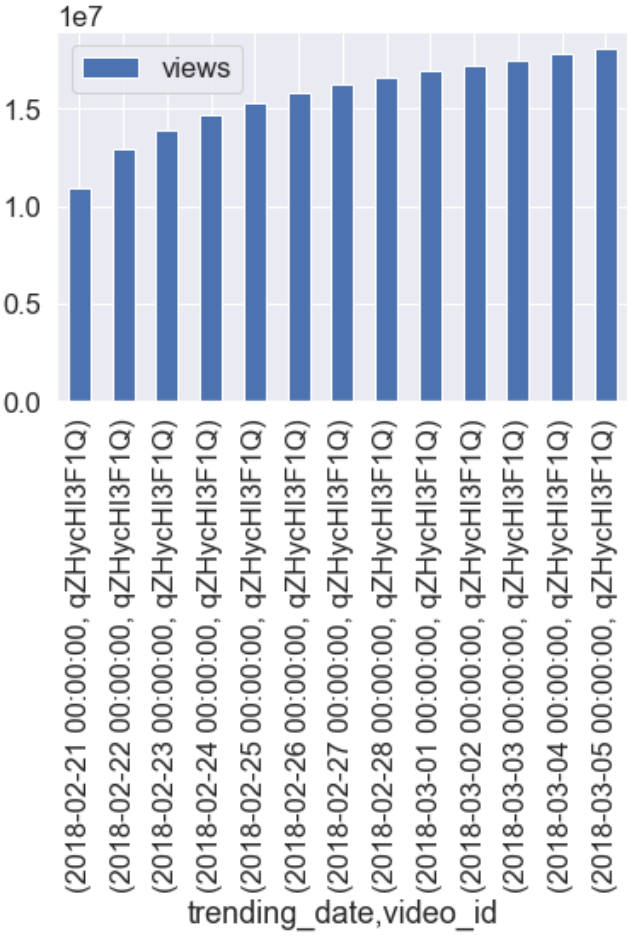
G. Question 7: Are the channels that have recurring trending videos, and is there a pattern weekly, monthly?

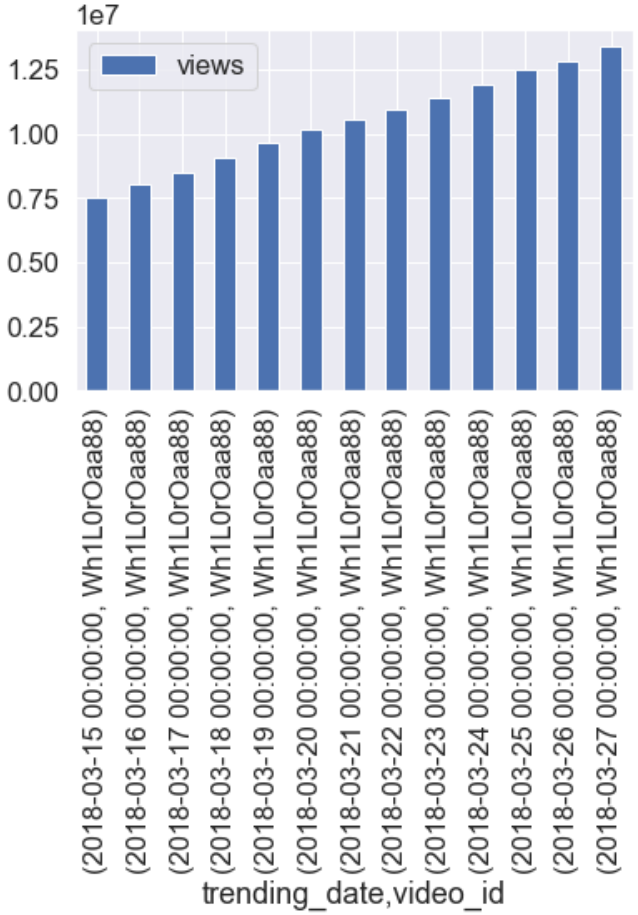
To answer this question, the top 3 channels were selected.. One channel in particular had a pattern that was interesting to see. The channel is called Dude Perfect and had at least one video that was trending each month from December 2017 to June 2018. It is the third channel with the most views.

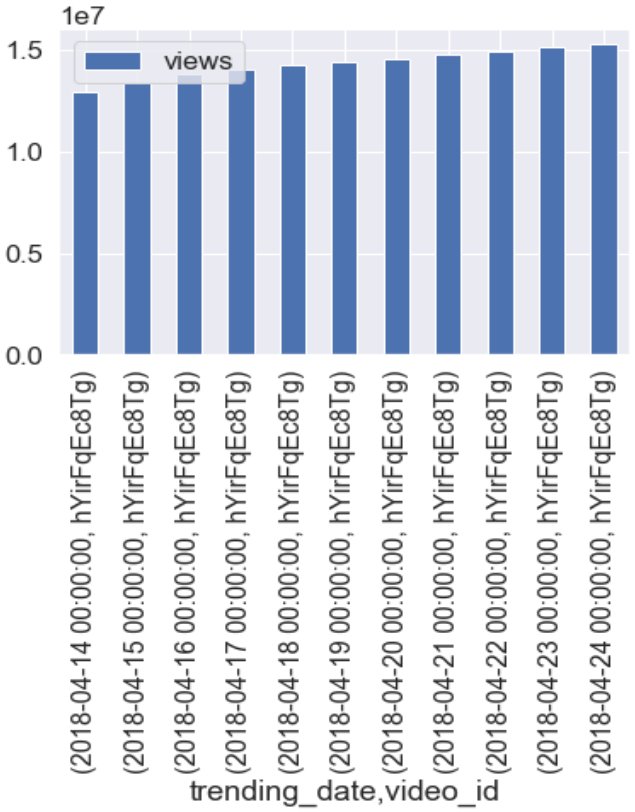

1e7
views
1.25
1.00
0.75
0.50
0.25
0.00
(2017-12-06 00:00:00, -RmUADCWI4A)
(2017-12-07 00:00:00, -RmUADCWI4A)
(2017-12-08 00:00:00, -RmUADCWI4A)
(2017-12-09 00:00:00, -RmUADCWI4A)
(2017-12-10 00:00:00, -RmUADCWI4A)
(2017-12-11 00:00:00, -RmUADCWI4A)
(2017-12-12 00:00:00, -RmUADCWI4A)
trending_date,video_id

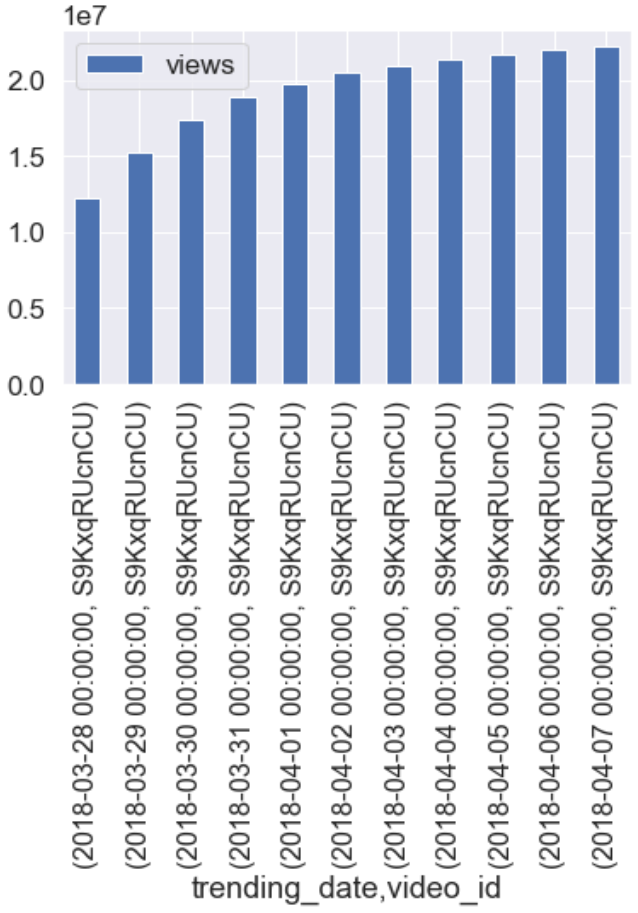

1e7
views
3
2
1
0
(2018-01-24 00:00:00, VRJmcxCrAOA)
(2018-01-25 00:00:00, VRJmcxCrAOA)
(2018-01-26 00:00:00, VRJmcxCrAOA)
(2018-01-27 00:00:00, VRJmcxCrAOA)
(2018-01-28 00:00:00, VRJmcxCrAOA)
(2018-01-29 00:00:00, VRJmcxCrAOA)
(2018-01-30 00:00:00, VRJmcxCrAOA)
trending_date,video_id

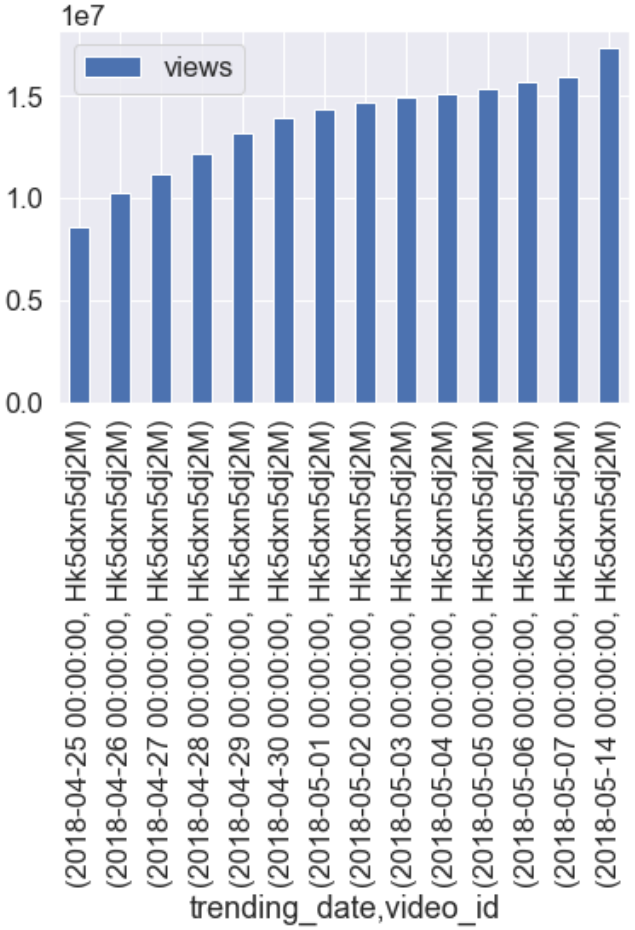

views
8000000
6000000
4000000
2000000
0
(2017-12-28 00:00:00, 9oyKePWRi8Y)
(2017-12-29 00:00:00, 9oyKePWRi8Y)
(2017-12-30 00:00:00, 9oyKePWRi8Y)
(2017-12-31 00:00:00, 9oyKePWRi8Y)
(2018-01-01 00:00:00, 9oyKePWRi8Y)
(2018-01-02 00:00:00, 9oyKePWRi8Y)
(2018-01-03 00:00:00, 9oyKePWRi8Y)
(2018-01-04 00:00:00, 9oyKePWRi8Y)
(2018-01-05 00:00:00, 9oyKePWRi8Y)
trending_date,video_id

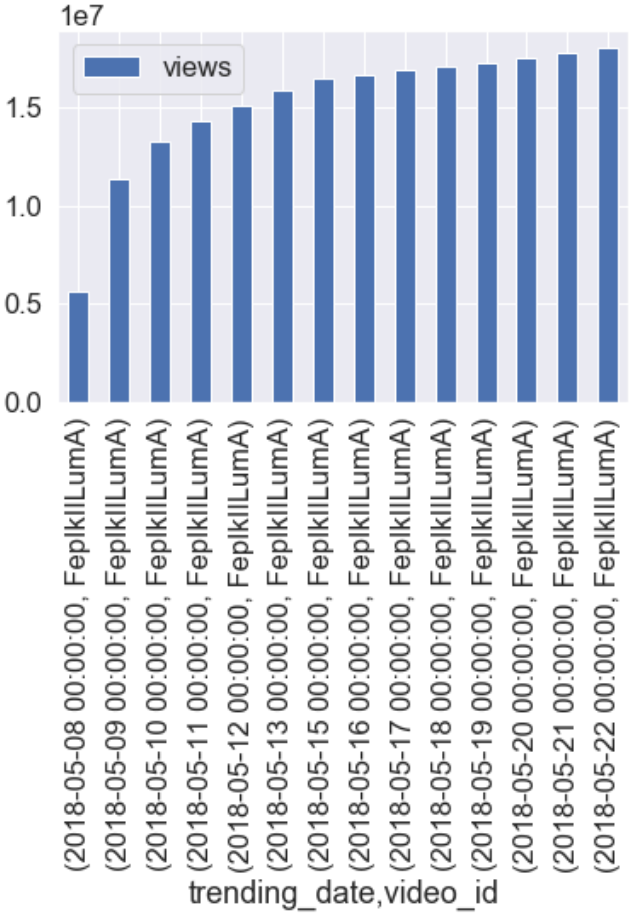

1e7
views
1.5
1.0
0.5
0.0
(2018-02-21 00:00:00, qZHycHI3F1Q)
(2018-02-22 00:00:00, qZHycHI3F1Q)
(2018-02-23 00:00:00, qZHycHI3F1Q)
(2018-02-24 00:00:00, qZHycHI3F1Q)
(2018-02-25 00:00:00, qZHycHI3F1Q)
(2018-02-26 00:00:00, qZHycHI3F1Q)
(2018-02-27 00:00:00, qZHycHI3F1Q)
(2018-02-28 00:00:00, qZHycHI3F1Q)
(2018-03-01 00:00:00, qZHycHI3F1Q)
(2018-03-02 00:00:00, qZHycHI3F1Q)
(2018-03-03 00:00:00, qZHycHI3F1Q)
(2018-03-04 00:00:00, qZHycHI3F1Q)
(2018-03-05 00:00:00, qZHycHI3F1Q)
trending_date,video_id

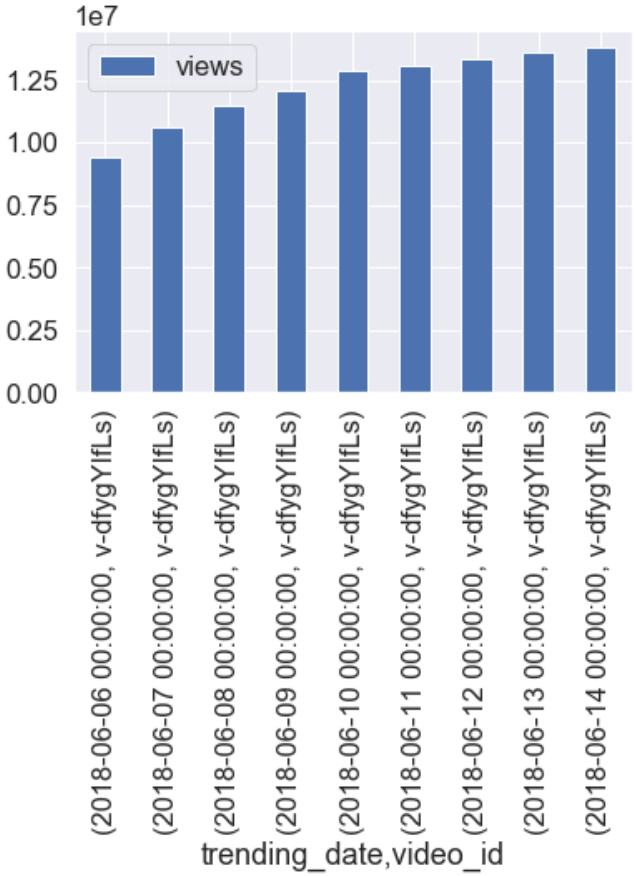

1e7
views
1.25
1.00
0.75
0.50
0.25
0.00
(2018-03-15 00:00:00, Wh1L0rOaa88)
(2018-03-16 00:00:00, Wh1L0rOaa88)
(2018-03-17 00:00:00, Wh1L0rOaa88)
(2018-03-18 00:00:00, Wh1L0rOaa88)
(2018-03-19 00:00:00, Wh1L0rOaa88)
(2018-03-20 00:00:00, Wh1L0rOaa88)
(2018-03-21 00:00:00, Wh1L0rOaa88)
(2018-03-22 00:00:00, Wh1L0rOaa88)
(2018-03-23 00:00:00, Wh1L0rOaa88)
(2018-03-24 00:00:00, Wh1L0rOaa88)
(2018-03-25 00:00:00, Wh1L0rOaa88)
(2018-03-26 00:00:00, Wh1L0rOaa88)
(2018-03-27 00:00:00, Wh1L0rOaa88)
trending_date,video_id

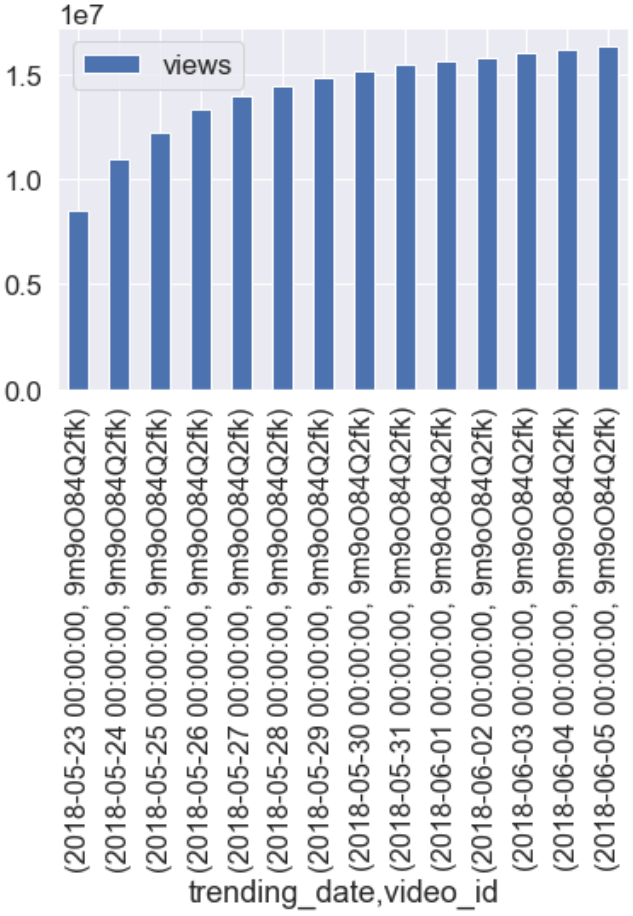

1e7
views
1.5
1.0
0.5
0.0
(2018-04-14 00:00:00, hYirFqEc8Tg)
(2018-04-15 00:00:00, hYirFqEc8Tg)
(2018-04-16 00:00:00, hYirFqEc8Tg)
(2018-04-17 00:00:00, hYirFqEc8Tg)
(2018-04-18 00:00:00, hYirFqEc8Tg)
(2018-04-19 00:00:00, hYirFqEc8Tg)
(2018-04-20 00:00:00, hYirFqEc8Tg)
(2018-04-21 00:00:00, hYirFqEc8Tg)
(2018-04-22 00:00:00, hYirFqEc8Tg)
(2018-04-23 00:00:00, hYirFqEc8Tg)
(2018-04-24 00:00:00, hYirFqEc8Tg)
trending_date,video_id


1e7
views
2.0
1.5
1.0
0.5
0.0
(2018-03-28 00:00:00, S9KxqRUcnCU)
(2018-03-29 00:00:00, S9KxqRUcnCU)
(2018-03-30 00:00:00, S9KxqRUcnCU)
(2018-03-31 00:00:00, S9KxqRUcnCU)
(2018-04-01 00:00:00, S9KxqRUcnCU)
(2018-04-02 00:00:00, S9KxqRUcnCU)
(2018-04-03 00:00:00, S9KxqRUcnCU)
(2018-04-04 00:00:00, S9KxqRUcnCU)
(2018-04-05 00:00:00, S9KxqRUcnCU)
(2018-04-06 00:00:00, S9KxqRUcnCU)
(2018-04-07 00:00:00, S9KxqRUcnCU)
trending_date,video_id


1e7
views
1.5
1.0
0.5
0.0
(2018-04-25 00:00:00, Hk5dxn5dj2M)
(2018-04-26 00:00:00, Hk5dxn5dj2M)
(2018-04-27 00:00:00, Hk5dxn5dj2M)
(2018-04-28 00:00:00, Hk5dxn5dj2M)
(2018-04-29 00:00:00, Hk5dxn5dj2M)
(2018-04-30 00:00:00, Hk5dxn5dj2M)
(2018-05-01 00:00:00, Hk5dxn5dj2M)
(2018-05-02 00:00:00, Hk5dxn5dj2M)
(2018-05-03 00:00:00, Hk5dxn5dj2M)
(2018-05-04 00:00:00, Hk5dxn5dj2M)
(2018-05-05 00:00:00, Hk5dxn5dj2M)
(2018-05-06 00:00:00, Hk5dxn5dj2M)
(2018-05-07 00:00:00, Hk5dxn5dj2M)
(2018-05-14 00:00:00, Hk5dxn5dj2M)
trending_date,video_id







VIII. References

Yanpapadakis. “Trending YouTube Video Metadata Analysis.” *Kaggle*, Kaggle, 25 Mar. 2018, https://www.kaggle.com/yanpapadakis/trending-youtube-video-metadata-analysis/notebook.

https://github.com/PrashantMohan9/Youtube-Category-Prediction/blob/master/Final_Version.ipynb

https://medium.com/@brianasaba07/analysing-the-youtube-trending-videos-dataset-951f4f4edc33

https://github.com/chefdarek/Youtube_Category_Predictor/blob/master/Youtube_Category%20_Labeling_Models.ipynb

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html