Assignment 2
YouTube Trending Videos Analysis

Mohammad Hossein Basouli

April 17, 2025

Abstract

In this analysis we will try to uncover hidden relationships and patterns behind a video sharing platform in order to be able to get insights for future decission for growing the related products. We start off by forming different questions about the data, and then through answering these questions grow our understanding of the data. This analysis would lead us to find impactful and correlated factors with some variables like engagement metrics, discover seasonal patterns in trending date and publish date of videos, significane of overall popularity of different video categories on appearance on the top most trending videos and so much more.

Introduction

Background: Analysis of relationships & and patterns in video streaming and sharing applications has been a matter of great interest since the beginning of development of these kinda services. It matters so much to be able to answer questions like why some special kind of videos get more views and the others don't, or what factors affects engagement of users on different videos and so on.

Objectives:

1. Explaining the data; where it's coming from, an explanation of the features, etc. in **A Description of the Data** and **Features** sub sections.

2. Preprocess the data in order to handle *integration* and *inconsistency* issues in **Data Preprocessing** sub section. And finally add additional features to work with in **Feature Engineering** sub section.

3. We will start our analysis in **Exploratory Data Analysis (EDA)** section by providing different plots of the data and analysing various aspects of it.

4. After answering our initial questions, it's time to form additional questions about the dataset to gain a broader view of it.

5. And at the end, we shall summarize our understanding of the data in **Connecting the Dots** sub section.

Initial Questions:

1. How are engagement metrics (views, likes and dislikes) distributed overall and across different video categories?

2. Which YouTube channels and video categories trend the most in each country and globally?

3. Are there seasonal or day-of-week patterns in trending videos? How does the upload day and time impact video engagement?

4. Do controversial videos, defined by a high dislike ratio, receive more engagement than universally liked ones?

5. How do video tags influence engagement, and which tags are most commonly used in trending videos?

6. How does the length of a video title impact engagement levels?

Data

A Description of the Data

The dataset has been gathered from the set of trending YouTube videos across 10 different regions in 2 different years (2017/18). The cumulative dataset contains 373,204 rows (trending video records) and 21 columns (features).

Features

Important numerical features include: *views, likes, dislikes, and comment_count.* Important categorical and datetime features include: *trending_date, title (title of the video), channel_title, publish_time, tags, comments_disabled, ratings_disabled, country, category_title.*

Data Preprocessing

Handling Missing Values: The only column that contains missing values is *description* which we shall not use in the course of our analysis. Thus we remove this column along with other unnecessary columns.

Duplicated Values: Since the dataset contains multiple records related to a single video for different *trending dates* in each *country*, we have to be cautious about this fact in our future analysis, thus we shall only keep those with the maximum views. Also we must remove the records which have the same *video_id, country* and *trending_date* since these records are duplicates.

Feature Engineering

We shall add multiple new features in our dataset in order to capture new information.

- Add engagement_score & dislike_rate: $engagement\_score = \frac{likes+dislikes+comment_count}{views}$, $dislike\_rate = \frac{dislikes}{likes+dislikes}$
- Add day_of_trend and season_of_trend from trending_date.
- Add season_of_publish, day_of_publish and hour_of_publish from publish_date.

Exploratory Data Analysis (EDA)

Visualization

Distribution of Engagement Metrics: If we have a look at overall distribution of **engagement metrics** on Figure 1 and their distributions across different **video categories** on Figure 2, we can say that all of the **engagement metrics** have a long-tail distribution on the right and most of the data are distributed over a very narrow range to the left. And also among

different **video categories**, *Music* and *Entertainment* seem to have a higher **engagement metrics** in general.

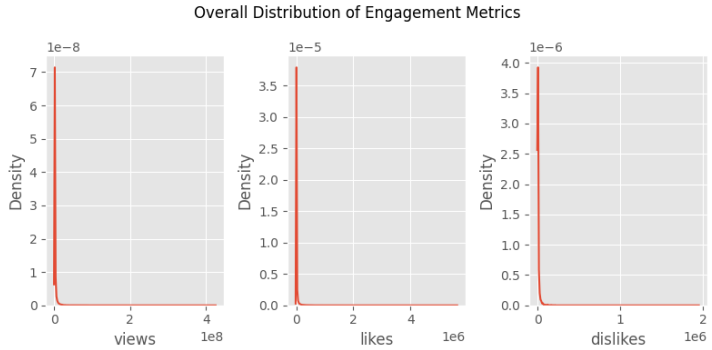

Figure 1: Overall Distribution of Engagement Metrics

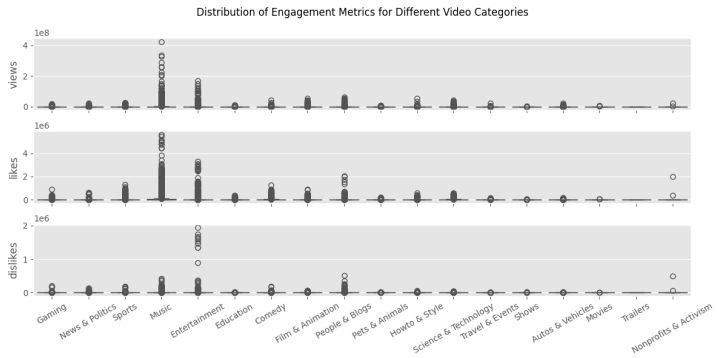

Figure 2: Distribution of Engagement Metrics Across Different Video Categories

Effect of Day/Hour of Publish on Engagment Score: By looking at Figure 3, we don't see a strong relationship between these factors. Thus we shall say that these features are independent.

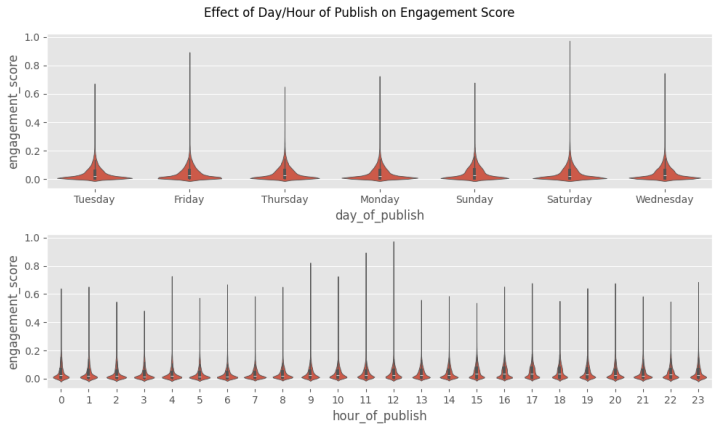

Figure 3: Effect of Day/Hour of Publish on Engagment Score

Effect of Dislike Ratio on Engagement Score: If we have a look at Figure 4, we can say that *universally-liked* group has a little higher **engagement score**. But we will investigate this relationship more, later in **Further Analysis** sub section.

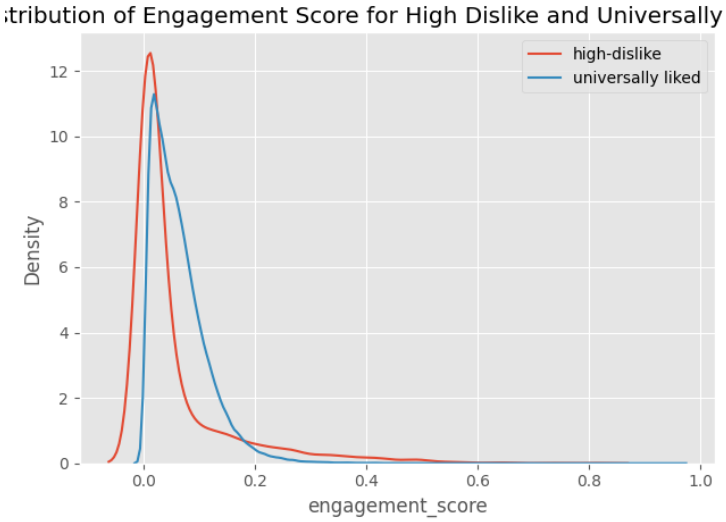

Figure 4: Effect of Dislike Ratio on Engagement Score

Relationship between Tag and Engagement Score: Figure 5 tells us that some **tags** have a better **engagement score** overally. e.g. *review* and *music* have a significatly higher **engagement_scores** medians.

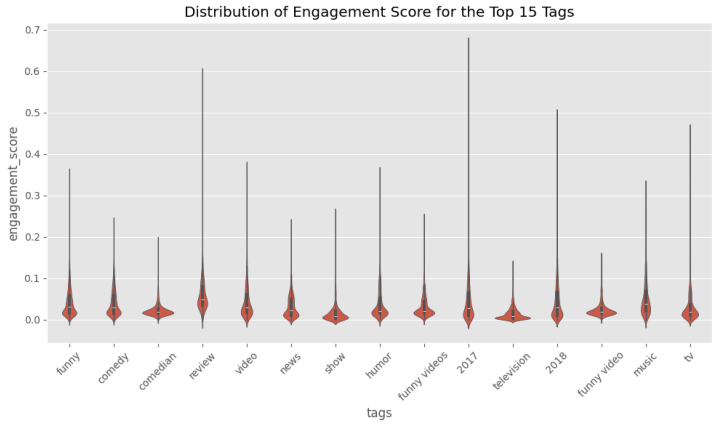

Figure 5: Relationship between Tag and Engagement Score

Correlation between Title Length & Engagement Score: By looking at Figure 6 we can say that, there isn't a strong relationship between these two features.

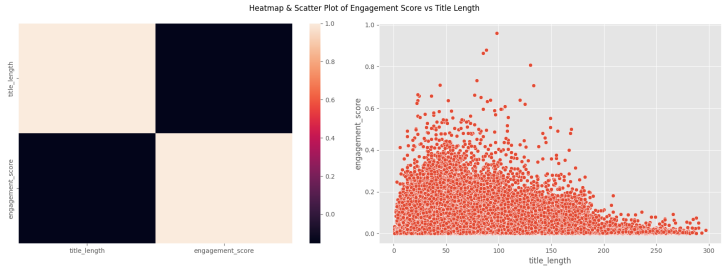

Figure 6: Correlation between Title Length & Engagement Score

Most Viewed Video, Channel and Video Category across Different Countries and Globally: If we look at Figure 7 we can easily see the most trending **video**, **channel** and **video category** in each country. Some interesting points worth noting here:

- A single category, channel and video, namely *Entertainment, Youtube Spotlight, Youtube Rewind: The Shape of 2017 | #YouTubeRewind* has dominated the most viewed **categories**, **channels** and **videos** by appearing in six regions!
- Only two **videos categories** appear in this list, meaning that most viewed videos in different regions have most likely the same **video category**.

- A single video stands as the most trending video in six different *regions*.

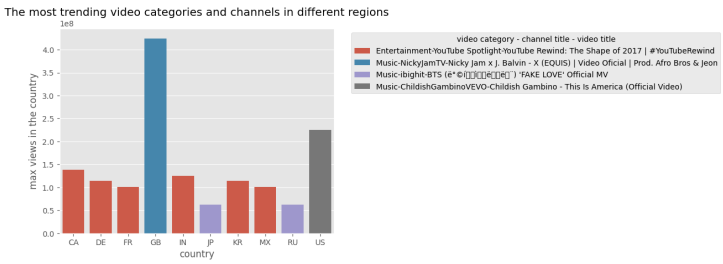

Figure 7: Most Viewed Video, Channel and Video Category across Different Countries and Globally

Seasonal/Daily Patterns in Trending Date: Figure 8 shows that we have a significant difference in the number of trending videos in each season. e.g. *Spring* and *Winter* have so many more trending videos in them.

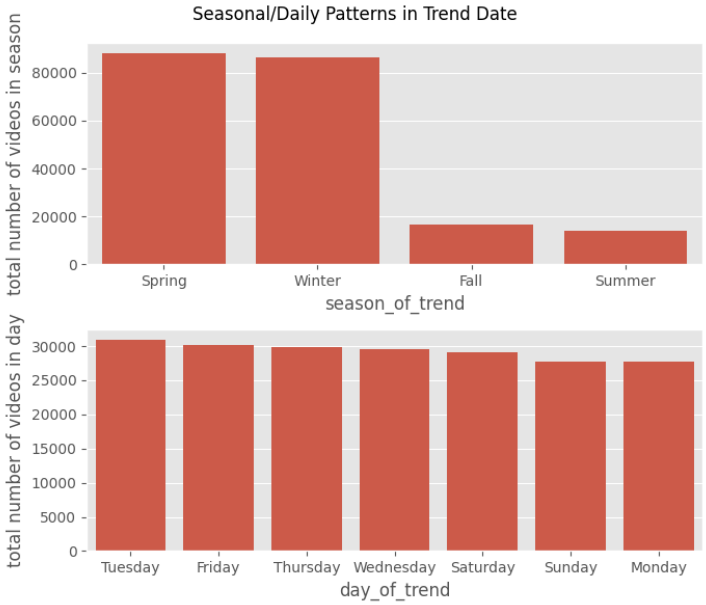

Figure 8: Seasonal/Daily Patterns in Trending Date

Top 10 Tags: Figure 9 shows the top 10 **tags** and their number of occurances.

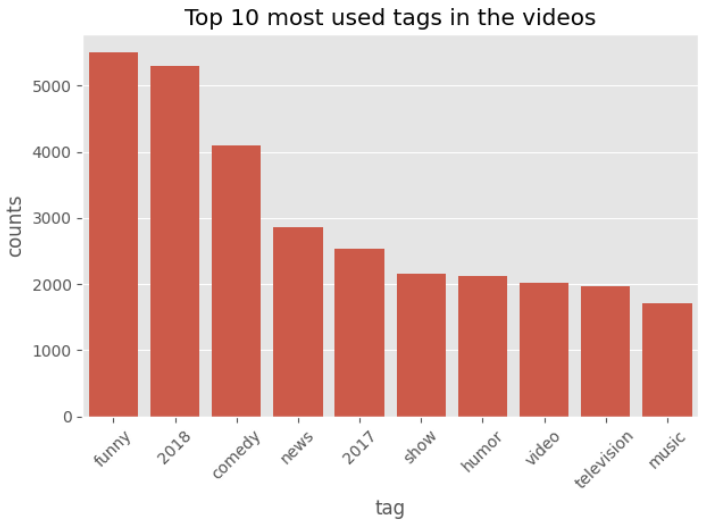

Figure 9: Top 10 Tags

Furthur Analysis

Now we want to investiagte the dataset more to gain more insights.

Frequency Distribution of the Video Categories: If we have a look at Figure 10 we can observe each of the categories along with their frequency percentages.

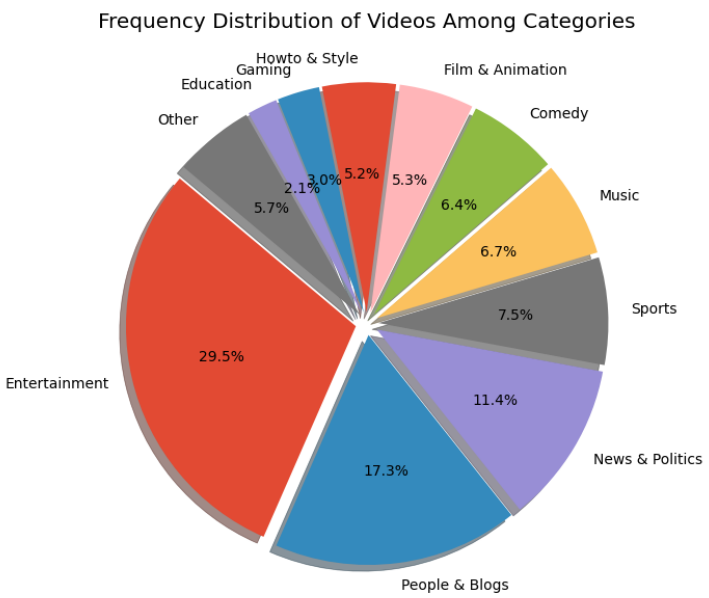

Figure 10: Frequency Distribution of the Different Video Categories

Top 5 Most Trending Categories per Each Year: By looking at Figure 11 we find out that there has been a huge increase in the number of usages for some categories such as *Entertainment, People & Blogs and News & Politics etc.*. Also we find out that *Music* has not been of great interest in the year 2017 but has been replaced with *Comedy* category in the year 2018.

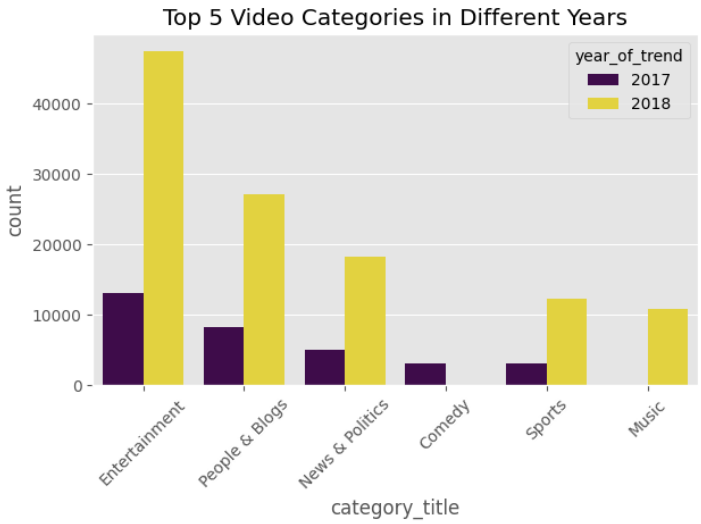

Figure 11: Top 5 Categories in Each Year

Seasonal/Daily Pattterns in Publish Time: Figure 12 shows us a significant difference across different **seasons** in the number of **publishes**. e.g. *Spring* and *Winter* have many more *publishes* compared to other **seasons**. But there isn't that much of difference in different **days of week** in terms of **number of publishes**

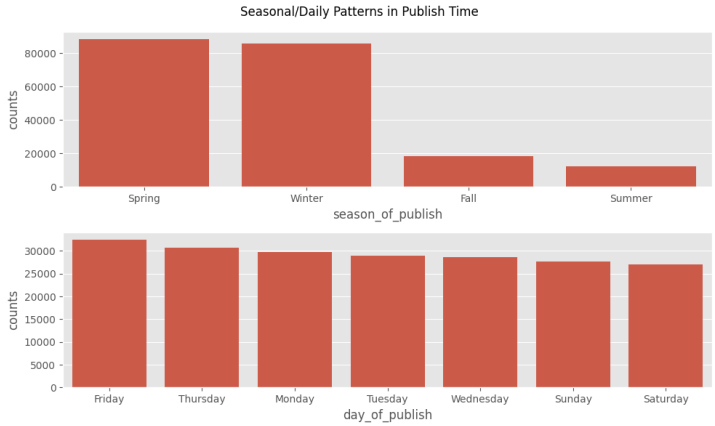

Figure 12: Seasonal/Daily Patterns in Publish Time

Effect of Comments Being Disabled on Likes: If we have a look at Figure 13, we can say that there is a significant difference between these two groups of data. Meaning that video with *comments enabled* get more **likes** in general.

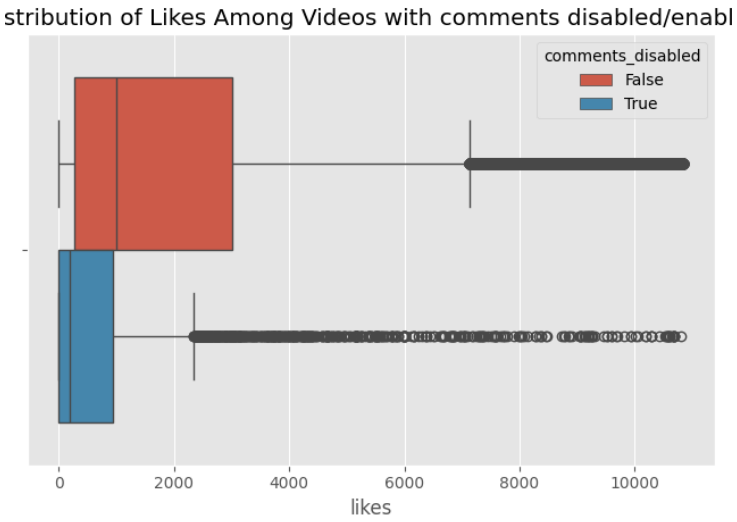

Figure 13: Effect of Comments Being Disabled on Likes

Top 10 Categories in Having Videos with Comments Disabled: By having a look at Figure 14, we see that the **categories** *Entertainment, News & Politics, Education and Film & Animation* have took the lead in terms of having more videos with *comments disabled*. It's is really interesting to see this huge amount of videos with *comments disabled* in *News & Politics* **category**.

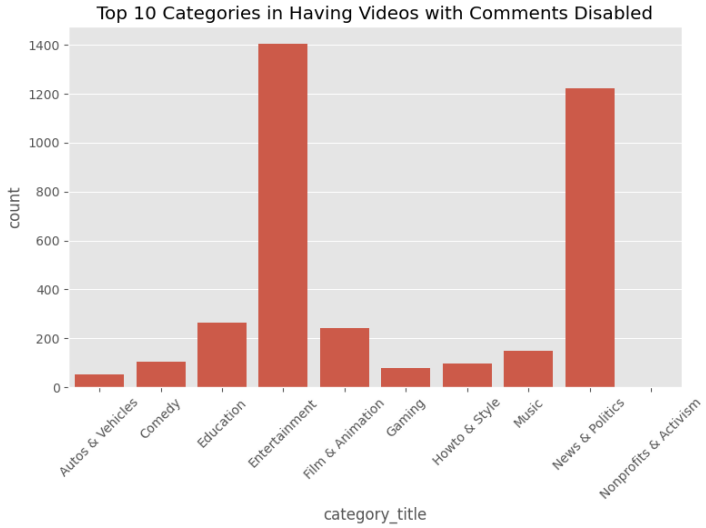

Figure 14: Top 10 Categories in Having Videos with Comments Disabled

Effect of Dislike Ratio on Views: Figure 15 shows that as we move from *low* **dislike ratios** to *higher* **dislikes ratios**, we will find fewer and fewer videos with an extremely high amount of views, meaning that videos with extremely high amount of views have at least a moderately good like ratio.

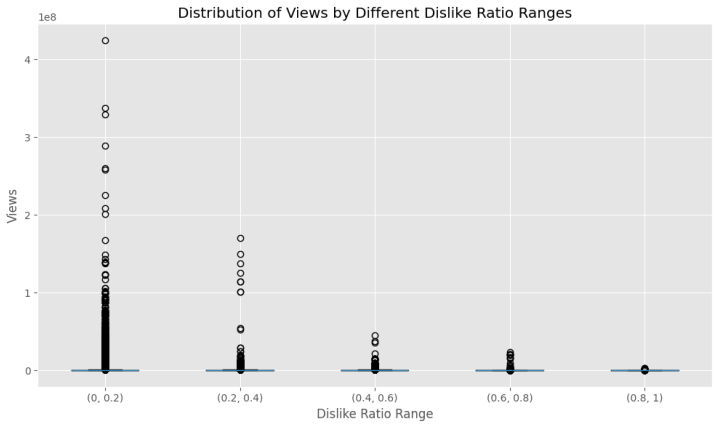

Figure 15: Effect of Dislike Ratio on Views

Effect of Dislike Ratio on Engagement Score: By looking at Figure 16 we can see a nice pattern unraveling. We observe that videos with extremely low or high dislike ratio are in a slightly better situation in terms of **engagement score** comapred to the videos with a more moderate dislike ratio.

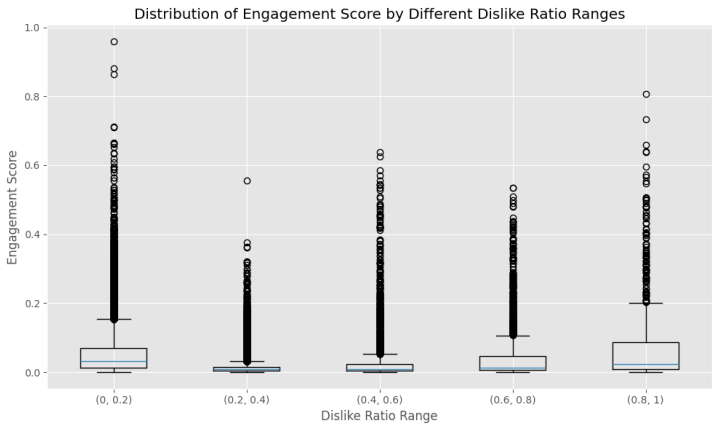

Figure 16: Effect of Dislike Ratio on Engagement Score

Connecting the Dots

Now we will try to connect the insights that we have taken from our analysis and combine them to deepen our understanding of the relationships and patterns that exists in the data.

Infuental Factors for Engagement Score: Based on the prior analysis on Figure 16, Figure 5, Figure 4 and Figure 9, we find out that the variables **dislike ratio** and **tags** have significant impacts on the variable **engagement score**. For exmple very high or very low **dislike ratios** comes with more **engagement scores** as well. Or videos with some certain **tags** like *review* and *music* appear in videos which have higher engagement scores in general, whereas some others like *television, show* and *comedian*. Also it worth mentioning that frequency of occurance of a tag doesn't mean that it will gain higher **engagement scores**. e.g. the **tags** *funny, 2018, comedy, news* and *2017* despite of being at the top of most used *tags* list, have a moderate **engagement score** in general.

Difference of Development in Different Categories and Hidden Impact of Each on Videos Being Trended: From our earlier analysis on Figure 7, Figure 10 and Figure 11, we can say that some categories have been developed so much more than the others over the course of these two year (2017-18); e.g. *Entertainment* has got 4 times more popular, *People & Blogs* has got 3 times more popular, whereas *Comedy* has been kicked out of the list. And also despite *Music* tag has been placed at 5-th place in terms of total occurances, it stands as the category of most trended video in one thirds of the regions, meaning that this category appears on few videos but it only appears on the most trending ones.

Infuental Factors in Engagement Metrics: From what was saw in our prior analysis on Figure 2, Figure 13 and Figure 15, we find out that some influential factors on **engagement metrics** are **dislike ratio**, **comments being enabled/disabled for the video** and **video category**. e.g. The **video categories** *Music* and *Entertainment* are the two have higher values **engagement metrics** in general. Or videos with **comments being enabled** receive higher amounts of **like** in general compared to those which have it turned off. or as the **dislike ratio** of a video increases, its likelihood of having a very large amoung of views decreases dramatically.

Similar Seasonal Pattern in Trending Date and Publish Date: If we turn back to the analysis that we have done on Figure 8 and Figure 12

we can see that there is a very similar *seasonal* pattern between **trending date** and **publish date**. Both of these variables seem to happen often on the same *seasons* more often.

Special Categories Are Not Open to Discussions: By looking back at our analysis on Figure 14, we can bring more insights on the table. e.g. the videos in the **category** *News & Politics* don't want to hear comments and have discussions about their videos whereas almost none of the videos in the **category** *Nonprofits & Activism* have their comments disabled and want to hear about users' comments and concerns about the video.