**Communicate Data Funding (USA YouTube dataset)**

**Introduction**

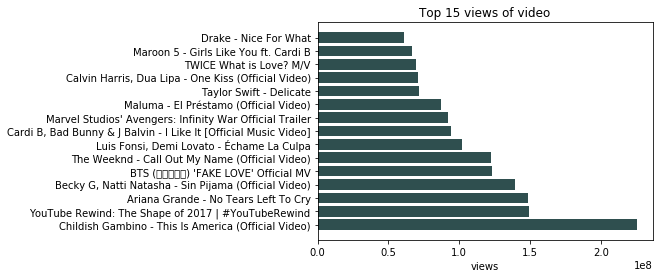
The datasets I used are from www.kaggle.com -- Trending YouTube Video Statistics, I chose datasets for USA only. The files contain a CSV file for Videos information such as video id, title, tags, channel, some feature dates but excluding category, which is contained in the other JSON file. So, at the very beginning, after check shape and info of the data frame, I read the JSON file and extract the category of each video by indicating the keys. Then here comes the completed data frame youtube\_df, which has 40949 records for videos mainly from 2017-2018 with 16 different categories. However, there is a column named trending\_date, making records duplicated, which influences the ranking of videos (i.e, the top 10 most liked videos could be same one). So I drop the duplicated records, keeping one with latest trending\_date, name the data frame as *youtube,* which is the main data frame I will use later on.

**Exploratory Analysis**

**Q1: which videos are the most viral?**

The best indicator for virality would be the counting on views. I sort the views to see top 10 most viewed video in the dataset. As I said in the introduction, if we use youtube\_df, magically, the top ten videos are same one -- *Childish Gambino - This Is America (Official Video)*. Well, it’s because the video is so popular, it trends a long time and the views are keeping increasing, so we need use the data frame without duplicate here.

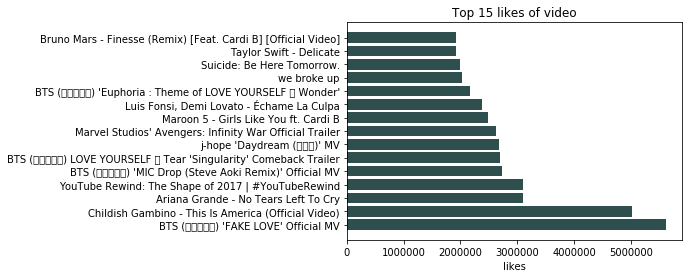
Here’s the outcome,

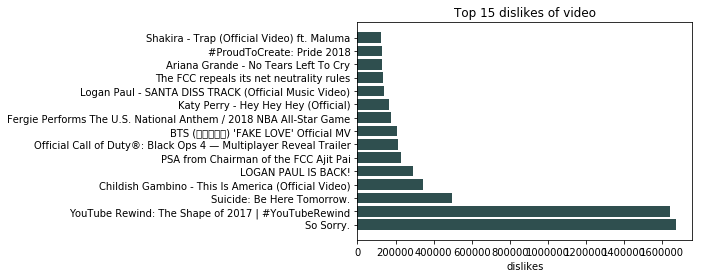


The views of *This Is America* is about 22 million, it’s really well-known. Also, it’s not hard to see, there majority of top 15 views of video is music. I could guess that Music is one of the most popular categories or could be the most. We will have a look at it later.

**Q2: which videos are the most liked/disliked?**

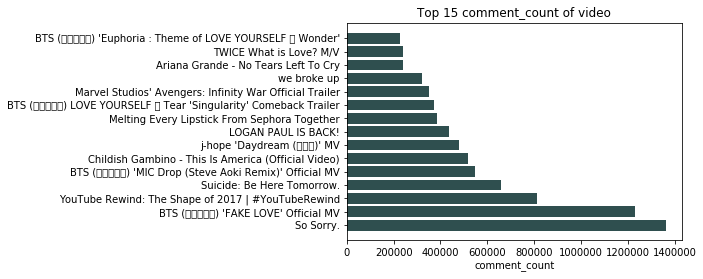
To define the popularity or not, we better look at the counting on likes/dislikes.



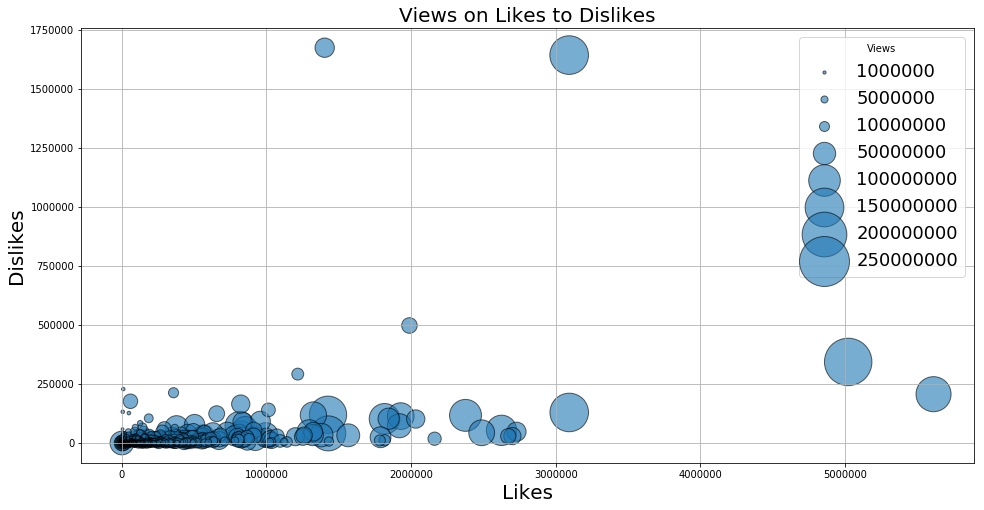


*This Is America* is the 2nd of top 15 likes and 4th of top 15 dislikes. The key finding here is the views really indicate on virality only, if we want to know if it’s popular or not, we need to check counting on likes as well.

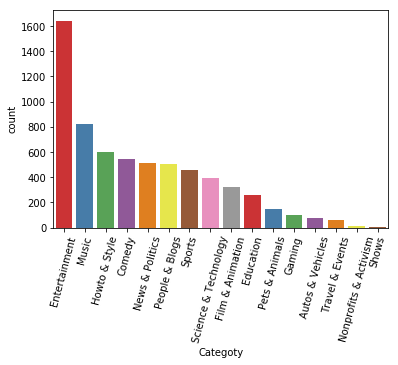
Furthermore, we could look at the top 15 comments. The comments could be positive, negative and neutral, so it can’t help to decide the popularity as well.



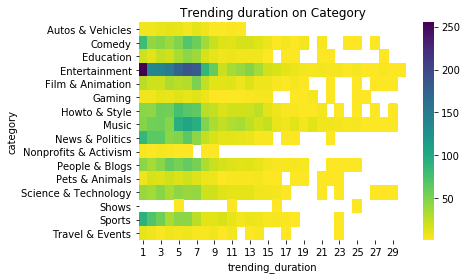
Let’s have a look at the views impact on Likes vs. Dislikes: the views impact does exist. for majority videos, views sizes are medium, more likes do lead more dislikes, but in a small amount compares to likes.



**Q3: Which categories are most influenced/popular?**



The bar chart shows the most popular is Entertainment (not Music!), then comes with Music, Howto & Style…However, we need to look at the trending time for each category to define the popularity.

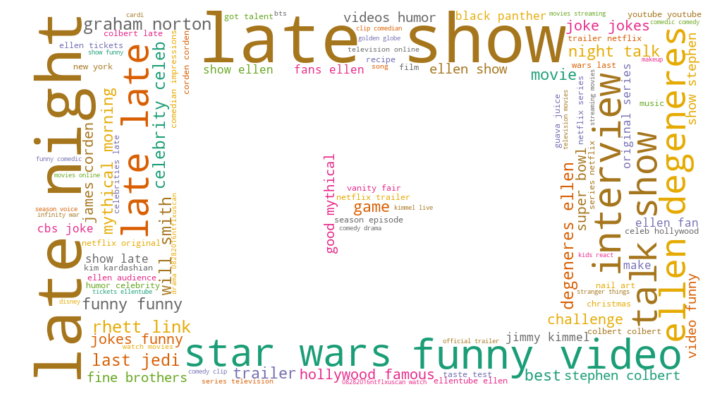


As was expected, the Entertainment is the most popular category, then I also can conclude that Music, Howto & Style, News & Politics and Comedy are also welcomed.

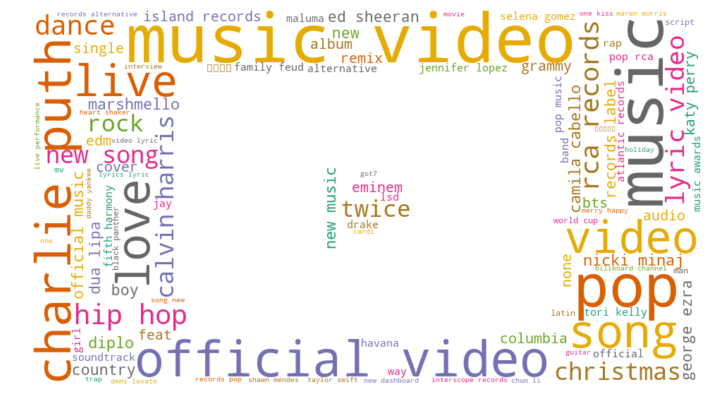
**Q4: How will you introduce each category to others?**

I would definitely extract the key words in videos’ tags, make a word cloud, with higher frequency words in larger font size. Here I attach the word cloud for some popular categories.

Entertainment: The key words are late, late, late. It seems like American people are really into watch YouTube at night and into funny videos. It makes sense, watching funny videos or talk show with family or friends together in a comfortable coach at night could be the best way to forget daily tiredness. On October 17th, 9PM, YouTube was down for about half hour, tons of people were crying at that time. From the accident we can tell over half of the country were watching YouTube at night.



Music: Official video and music video are most frequent. I could conclude that majority singers/bands would like to publish their new songs through YouTube.

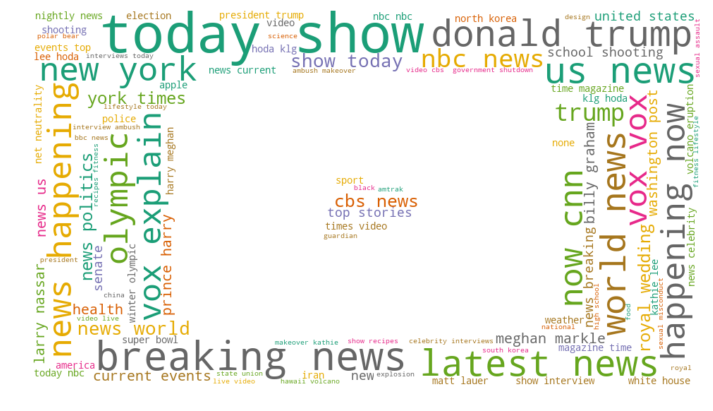


Howto & Style: Not to be surprised, this category is so influenced. Make-up, DIY, food recipe, tutorial…if you need a channel to learn something you can’t, search it on YouTube!



Comedy: Talk show is main stream in this category.

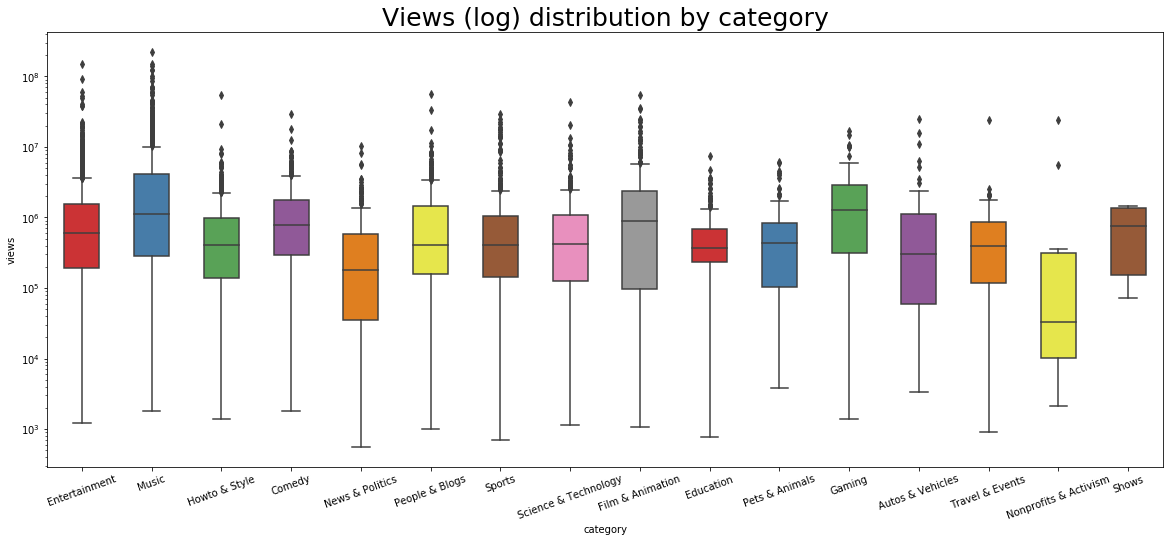
News & Politics: It all about news, breaking news, world news, latest news, along with some political topics.

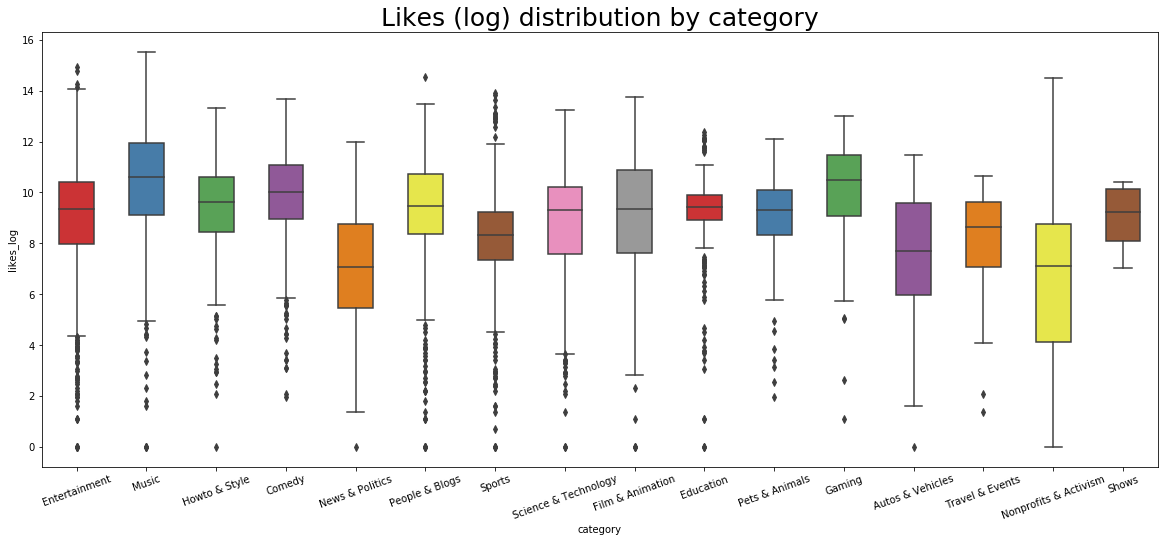


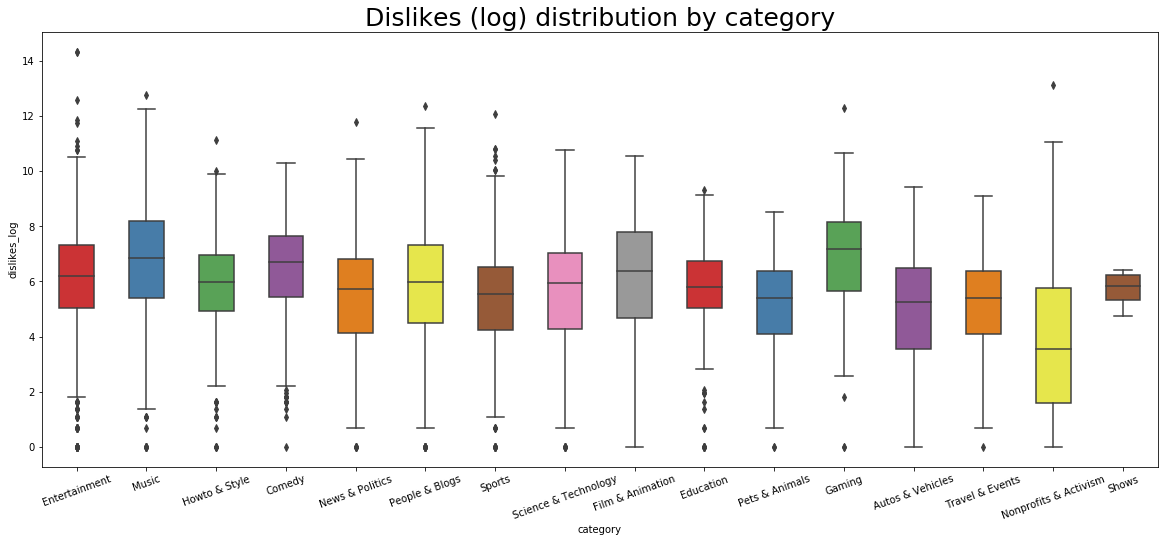
**Q5: How’s audiences’ reflection on different categories?**

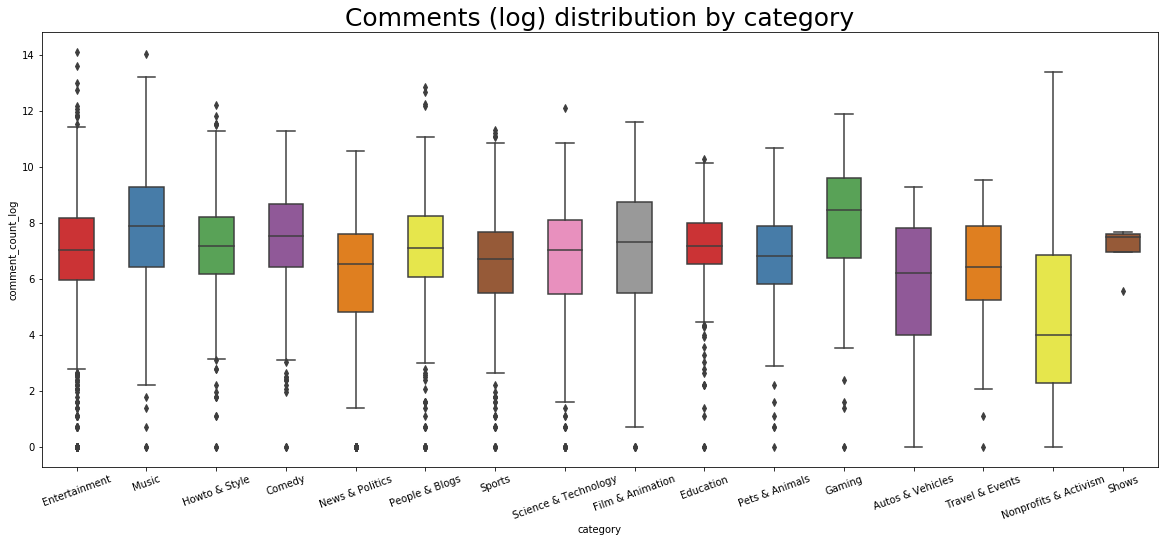
Audiences’ reflection delivered by views, likes, dislikes, comments. Here I apply boxplot, which is meaningful when one is categorical variable, the other is numeric variable.

The category is ordered by counting on videos. The medians are in random pattern, up and down, which means the views doesn’t related to the category that much. There are only 4 videos in Shows category, however the median of views is still higher than some of other categories.



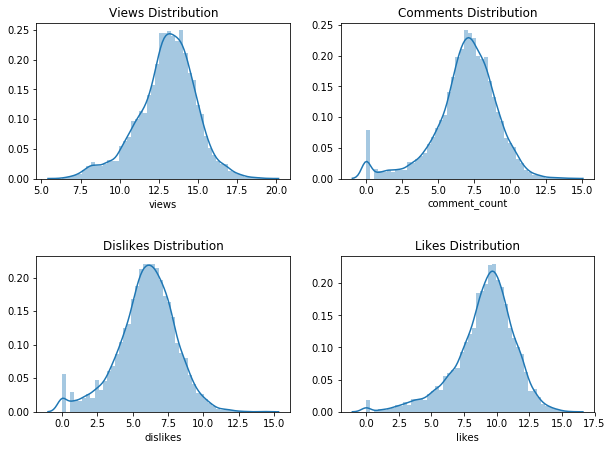
So, the likes, dislikes, comments don’t get much influence by category either. 





**Q6: What are the distributions for 4 different variables (views, like, dislike, comments)?**

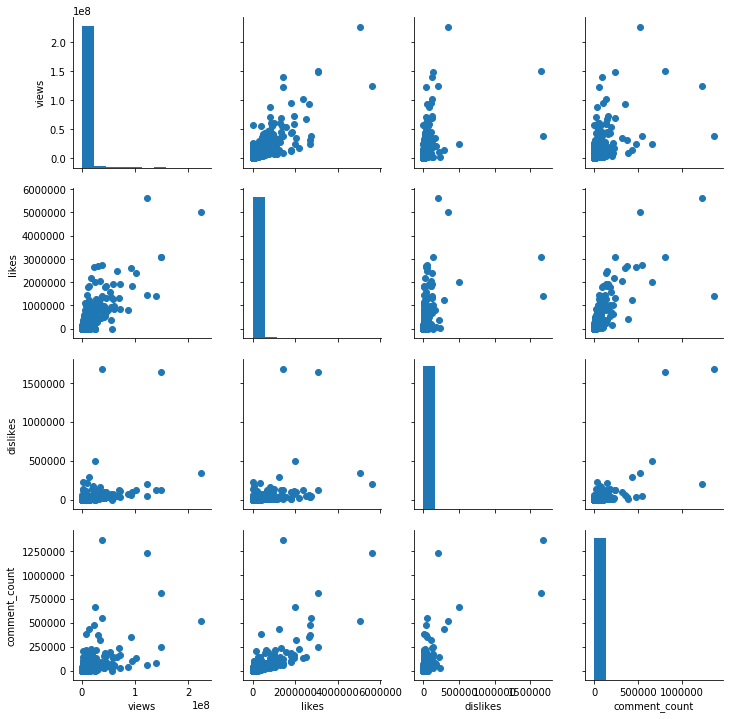
All shapes are belled, which is normal distribution.

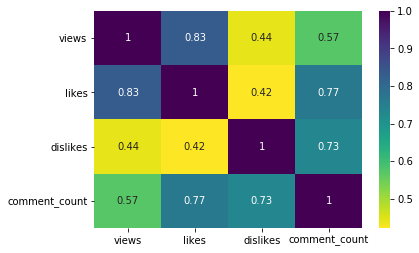


**Q7: Do the 4 variables linear related to each other?**

Combine both plots, we could conclude that

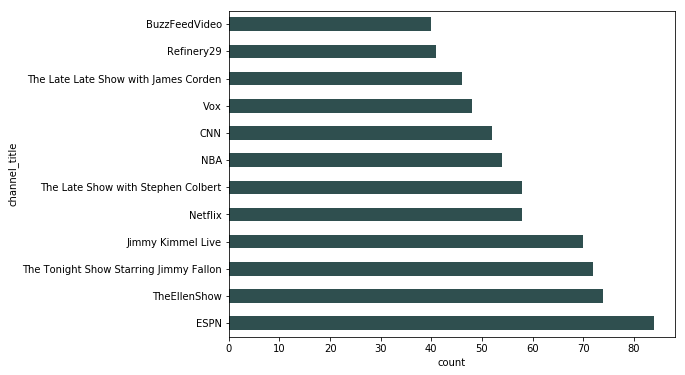
1. likes and views are highly correlated
2. likes and comments are highly correlated
3. comments and views are moderately correlated
4. dislikes and views are low correlated
5. dislikes and likes are low correlated





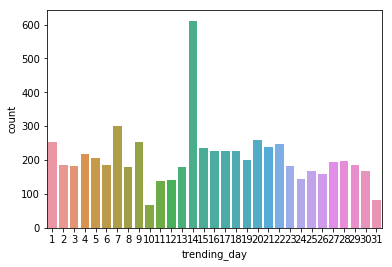
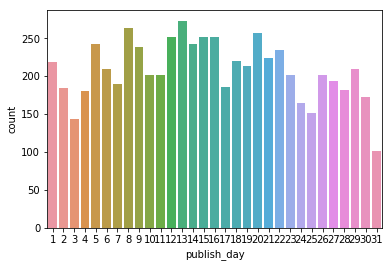
**Q8: Who’s the most influenced channel on YouTube?**

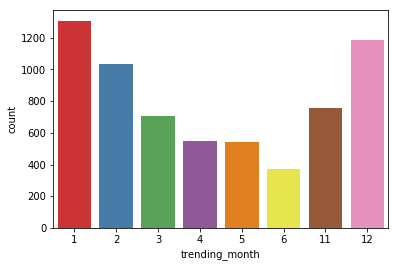
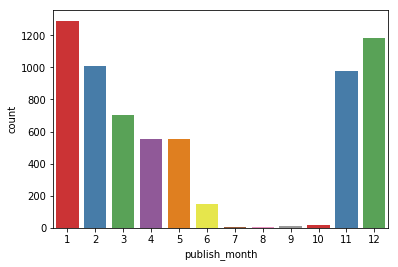
I define the influenced channel by if the videos they published are over 40. There are only 12 channels with published videos more than 40, out of 2199 channels.



**Q9: When would be the best time to publish a video?**

There are lots of videos uploaded every day, and 15th is the lucky day for videos to trend. Based on month, winter is the hot season for publishing, I assume this is due to holiday season. People would like to YouTube and chill at home, publisher of course would like to publish. So that, winter is also trending season.





**Conclusion**

The counting on views decides the virality of a video, plus counting on likes decide the popularity of a videos. The counting on dislikes and comments don’t influence that much on the virality and popularity. In this dataset, the most welcomed category are Entertainment, Music, Howto & Style, News & Politics, Comedy and Sports, and the most influenced channel are almost from above category. The best publish time for a youtuber would be winter, either day, while better before 15th.

**Comment**

I really enjoy this project. This could be my favorite project while learning at Udacity. It’s also my last project. Trying to be a data analyst, I always regard plots and graphs as the best tool to explain insights. Through this chapter and the project, I know how to choose a better plot, how to make use of proper maker and color, how to polish graph for a more accurate explanation. I appreciate everything here I learn.

**Resource:**

<https://www.datacamp.com/community/tutorials/wordcloud-python>

<https://www.kaggle.com/datasnaek/youtube-new>

<https://stackoverflow.com/questions/18674064/how-do-i-insert-a-column-at-a-specific-column-index-in-pandas>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.to_datetime.html>