**Project report:**

**By Adi Levi and Tal Blau:**

YouTube maintains a list of the [top trending videos](https://www.youtube.com/feed/trending) on the platform. YouTube uses a combination of factors including measuring user's interactions (number of views, shares, comments and likes). This dataset is a daily record of the top trending YouTube videos.

This dataset includes several months (and counting) of data on daily trending YouTube videos (above 10,000 views) between the years 2017-2018. The Data includes USA, Great Britain, Germany, Canada, and France, Russia, Mexico, South Korea, Japan and India, with up to 200 listed trending videos per day.

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. Number of records in the dataset is 40949.

The link for the dataset: <https://www.kaggle.com/datasnaek/youtube-new>

Inspiration:

we both are frequent users (everyday couple of times a day) whether its for music, tv shows, lectures of Tzenzor or how to make things yourself- YouTube plays a major part in our life. Therefore, we would like to explore and analyze this important (for us) site.

Possible uses for this dataset could include:

* Sentiment analysis in a variety of forms.
* Categorizing YouTube videos based on their comments and statistics.
* Analyzing what factors affect how popular a YouTube video will be.
* Statistical analysis over time.
* Categorize a trendy video base on the categories or even title or tags.
* Try to predict if a video is going to be viral.

We were interested in those questions because we thought that if we can define what makes a video to be trendy, we can:

* 1. Advice youtubers how to make themselves more popular.
  2. Advice companies where to advertise, when and how much base on their goal and target audience.
  3. Suggest a company a specific channel with a potential to become viral so they could even contact the channel owner so people couldn’t skip or ignore the commercial.

**Distributions of 5 variables**:

**The number of videos follow the years - Usage of YouTube:**

We group the csv file according the months in years 2017-8 to explore the uses of USA citizens in YouTube website.

We want to learn if the number of posted videos is from uniform distribution or is there a period where more videos posted there?

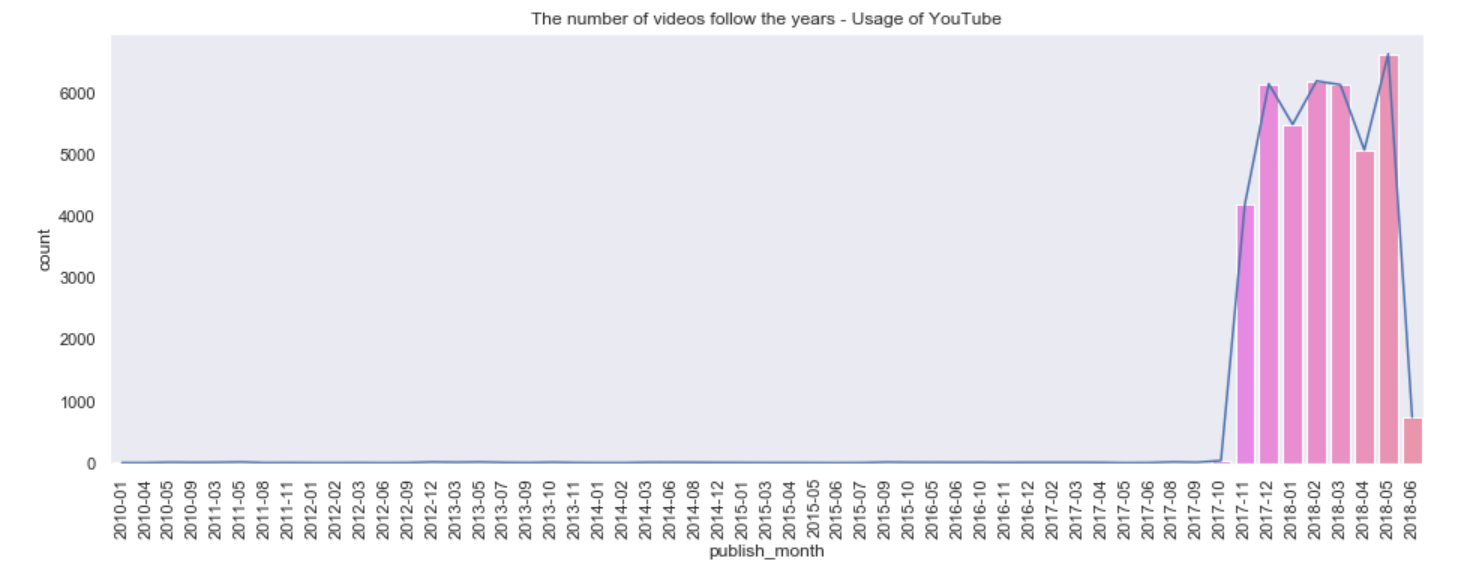
We chose to look at the timeslot between 10-2017 till 06-2018:

Figure 1

We can see the increase of usage in YouTube between 11-2017 till 05-2018. (there are a few numbers of videos between the dates 2010-10 to 2017-10.)

Between those months there is no dramatic increase or decrease we can learn from it.

In those months there is approximation to uniform distribution.

**Number of views:**

In the suggestion for the project we wanted to explore the popularity of videos in YouTube. A Good quantifier is the number of views.

The major goal for a publisher is to show your product to as many potential customers. In YouTube case, for advertisement publishers or bloggers, they want their videos to be seen with maximal number of views. Therefore, we chose to explore this feature.

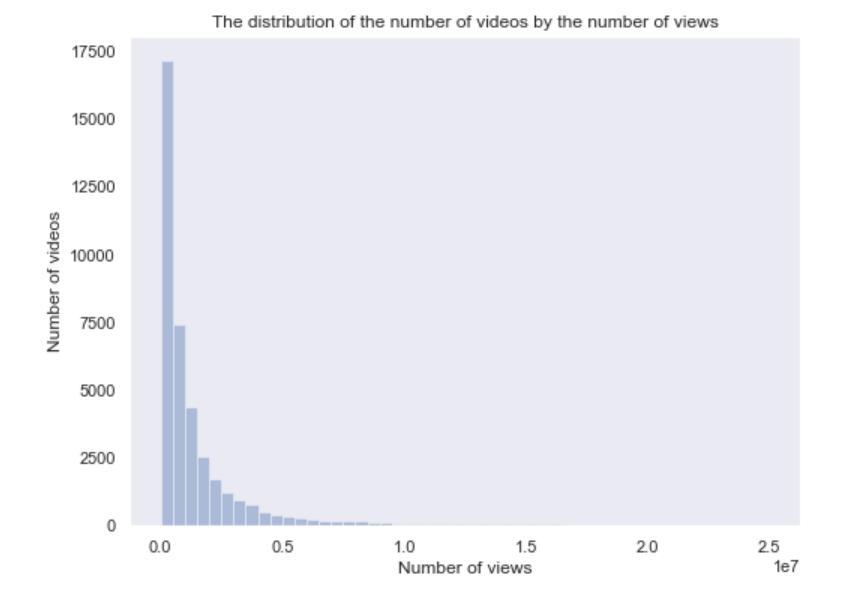
We want to see how many videos are viral. We can do that by looking at the distribution of the number of videos that are uploaded, according to the number of views they have and see the relationship between the two.

Figure 2

According to the plot we can see that the bigger the number of viewers a video has the smaller the number of videos that has a similar number of viewers as it.

We chose to delimit the histogram of the distribution because we wanted to get a closer look on the data, as well we can see that the number of videos that with larger amount of views than 25 million are barely seen in the histogram and are negligible to the amount of videos who has less views (only 1.2% from the total data)

We can have a better understanding on how the popularity of videos scatter:

From the code in Jupiter we calculated:

* 60.09% of the videos in YouTube have less than 1 million views, meaning that if I want to advertise in a big scale only 40% of the videos are relevant.
* 85% of the videos in YouTube have less than 3 million views.
* 91% of the videos in YouTube have less than 5 million views.

**Category name:**

In our project suggestion, we wanted to explore if the habits or interests for each country can be learnt from their YouTube videos. We were interested to learn how different cultures effect the type of category people upload to YouTube. In the end we would like to see the comparison to USA (our subject)

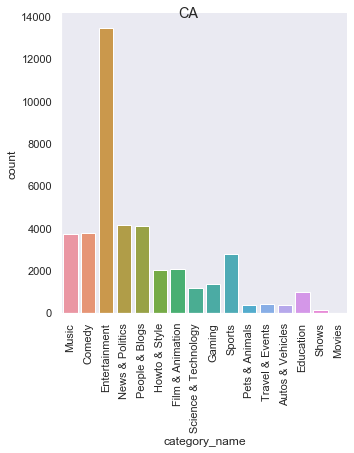
We group each category for each country and count the number of videos in each.

Figure .a

Canada:

* Entertainment is the most frequent category overshadowing by far all other categories by 3 times or more with 13000 videos.
* Following are music, comedy, people & blog, news & politics with 4000 videos.
* The least frequent category is movies.

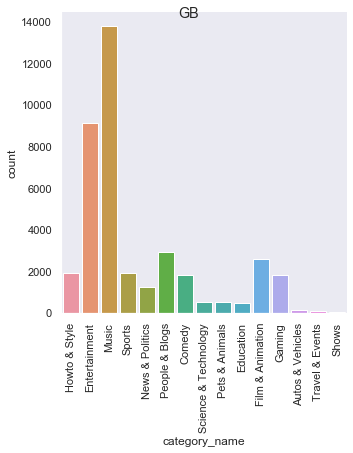


Figure .b

Great Britain:

* music is the most frequent category (14000 videos) followed up by Entertainment (9000 videos). Both categories are the major part of the videos that people from Great Britain uploads.
* The least frequent category are shows, travel & events and autos & vehicles without any category of movies.

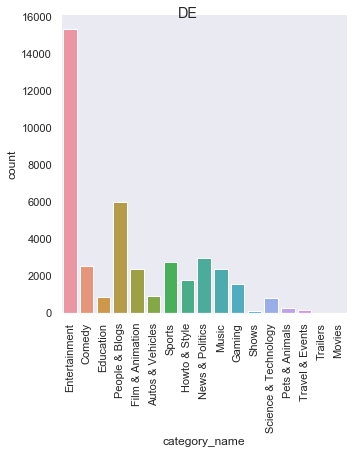
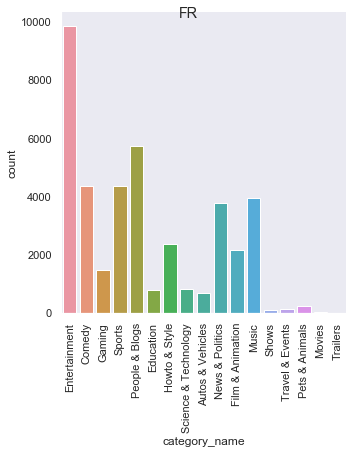


Figure .c

Denmark:

* Entertainment is the most frequent with 15000 videos.
* The second most frequent category is people & blog with 6000 videos, only 40% of the number of videos in the Entertainment category.
* The least frequent category are movies, shows, pets & animals, travel & events, science & technology and trailers.

Figure .d



France:

* Entertainment is the most frequent with 9000 videos.
* The second most frequent category is people & blog with 6000 followed up by sports, comedy, music and news & politics with 4000 videos.
* The least frequent category are movies, shows, pets & animals, travel & events and trailers.

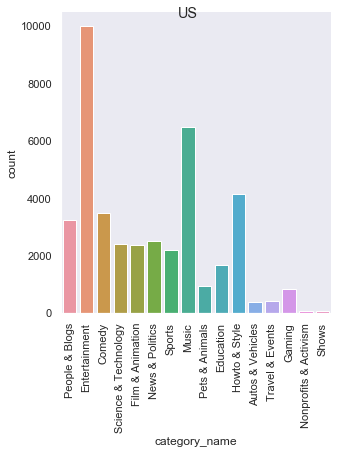


Figure .e

USA:

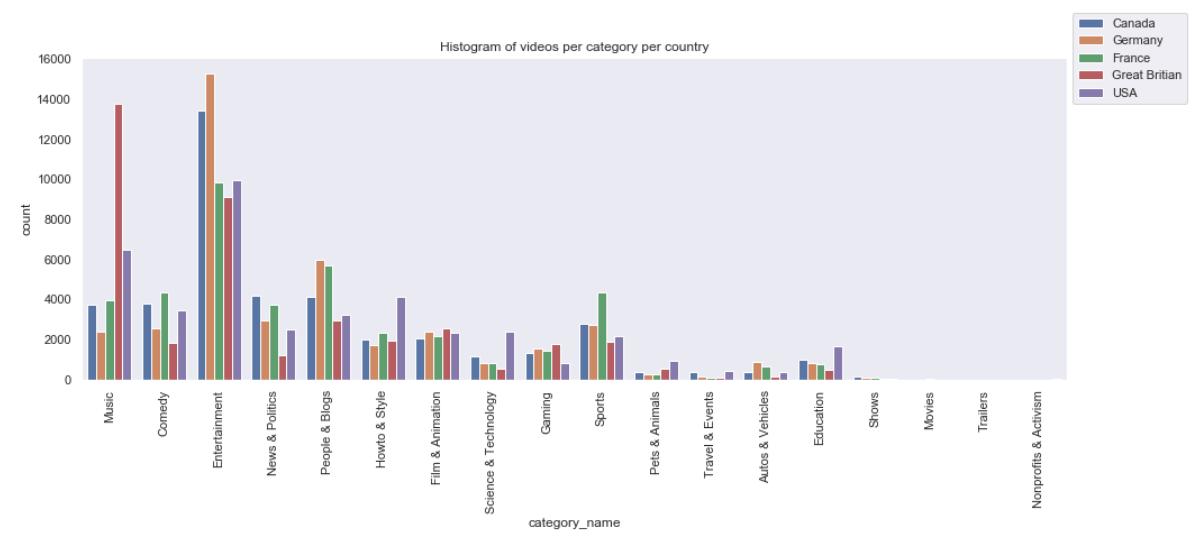
* Entertainment is the most frequent with 10000 videos.
* The second most frequent category is music with 6500.
* The third most frequent category is how to & style (4000), followed up by people & blog, comedy, film & animation, science & technology, news & politics and sports.
* The least frequent category are shows and nonprofits & activism.

Conclusions from the plots:

* We can see that there are categories in some countries that don’t appear in others like nonprofits & activism are special for USA only.
* We see that all countries have the categories of Entertainment, music, comedy, people & blog, science & technology, news & politics, film & animation, sports, pets & animals, travel & events, how to & style, education, gaming and shows.
* In all countries we observed, except Britain, the most frequent category is Entertainment.

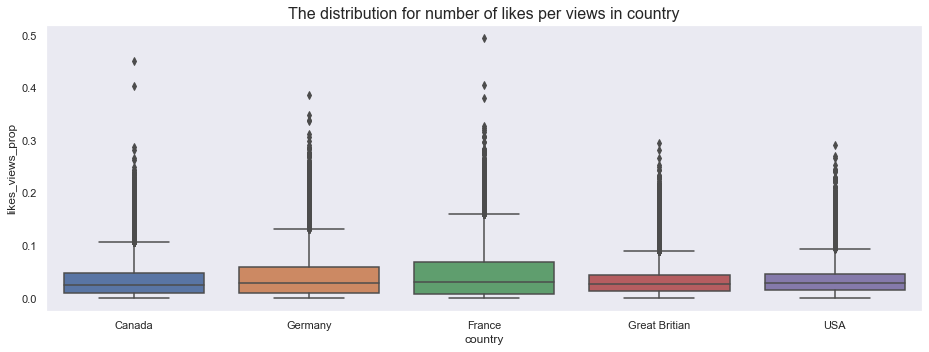
In Britain the most interested category is Music.

* In all the countries we observed we can see that the major interest is music and entertainment as we except.
* USA interested a lot about videos on How to… & Style whereas France and Germany have a lot of videos about people and blogs.
* We thought that countries with close culture will have same distribution over the categories, but here every state has different interests, so our assumption is incorrect.

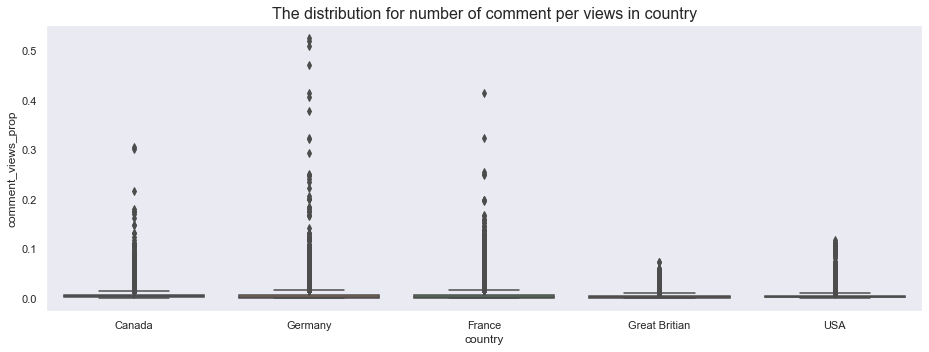


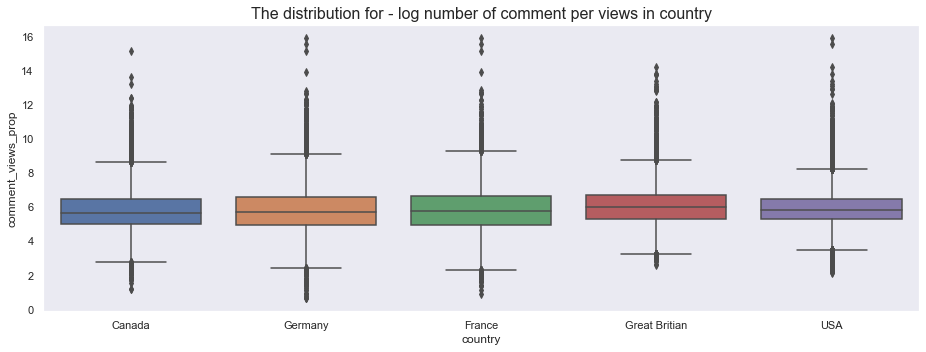
Here we can see a comparison in the categories in each country

Figure .f

here we can observe the distribution of the proportion of people that liked the video from the total number of viewers for each country. This can show us the distribution of the active viewers per country

here we can observe the distribution of the proportion of people that commented on the video from the total number of viewers for each country. This can show us the distribution of the active viewers per country



We can't see any difference in this graph, so we use -log function to keep the order but emphasize the difference.

There is no big difference between those distributions.

Also, in this parameter France and Germany are leading with Great Britain following - this country in the likes' graph was in the last place.

Conclusions (base on this data only):

* France and Germany are the most active countries that display their opinions in YouTube.
* Entertainment is the most frequent category in most countries in YouTube, followed by Music.
* People from Great Britain prefer to write comments than to do likes.

**Title:**

In order to have even better understanding about people's preferences we can look at the title of the video. This kind of feature can extract more information from the data that we don’t have in the traditional categories. From the title we can observe which topics are most common in YouTube and not only what category it falls into. We can learn better about the way people chose to advertise their videos, hidden and sub- categories that are more common than we thought. This feature can present the most frequent issues or terms in years 2017-8.

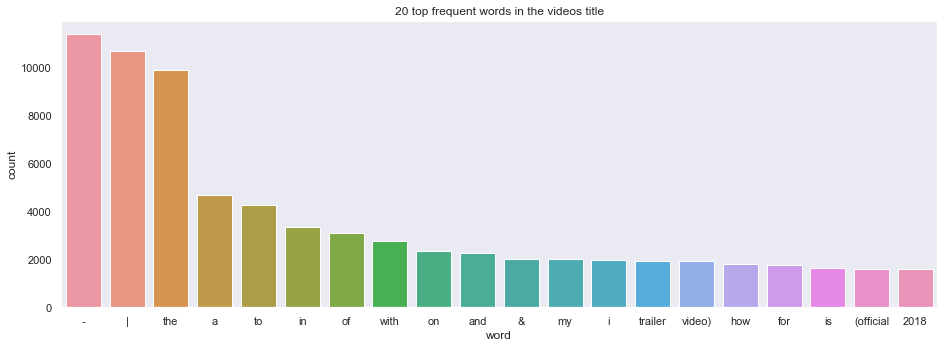
For start we show the most 20 frequent words in the titles:

Figure 4.a

Most of the words we get are stop words, we can't learn a lot from it. So, we will cut those stop words to have a better understanding of the data.

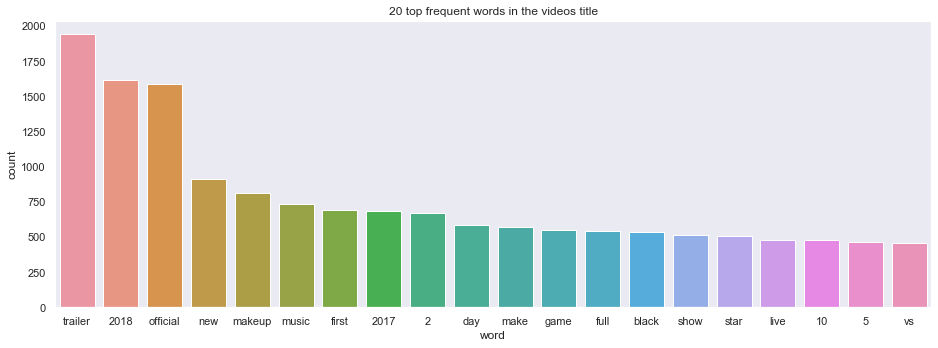


Figure 4.b

From the plot we can see that the 20 most common words and we can only assume that:

1. **trailers**- meaning that in those years we can see that many movies and tv shows released their trailers in YouTube.
2. **years 2018 and 2017**- those are the years this data is concerns with. we can see that the year 2018 is mentioned more in the titles that 2017, so we can suggest that there were more topics related to the year 2018 or more videos published in our data from the year 2018.
3. **makeup**- we can see that are many makeup related videos in this data. This is a new kind of topic we didn’t know was so common and couldn’t have spot by only looking at the categories.
4. **official**- we can infer that there are many covers to the videos that are trending and they are trending as well because you wouldn't write it's the official video if it's the only one out there or it's the first one that pops in search. Maybe official even concerns other topic like ceremonies or events.
5. **music**- it is less common for videos concerning music to write music in the title of the video than subject related video to write their subject in the title.
6. **new**- means that are new videos of the same topic. meaning there are many follow up videos.
7. **make**- there are many topics on YouTube on how to make something by yourself.
8. **game**- there many videos of gamers on the game they are playing.
9. **star/show/live**- celebrity and news related topics.  
   and more...

From those topics we can suggest to companies how to advertise themselves:

* In which category of videos to publish
* If there are many follow up videos so they can consistently advertise their product
* If there could be a similar video to this specific video.   
  and more...

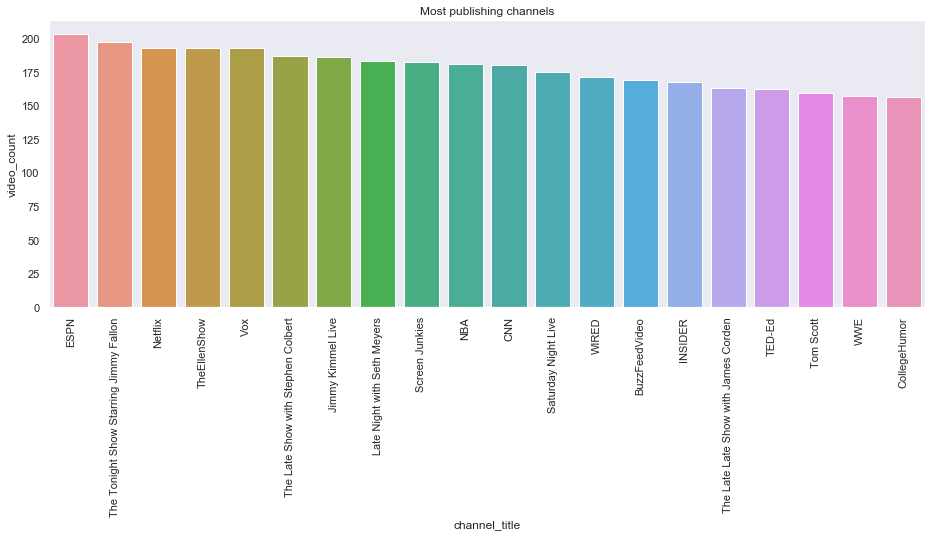
**YouTube Channels**:

Figure 5.a

we can see here the number of videos that of the top 20 most trending channels and how many videos they uploaded in the years of 2017 and 2018. Most of them are TV shows or TV channels with a lot of contents.

But as a publisher we want also to publish in channel with a lot of viewers.

For this answer we calculate the total amount of views for a channel and divide it to channel's number of videos it publishes.

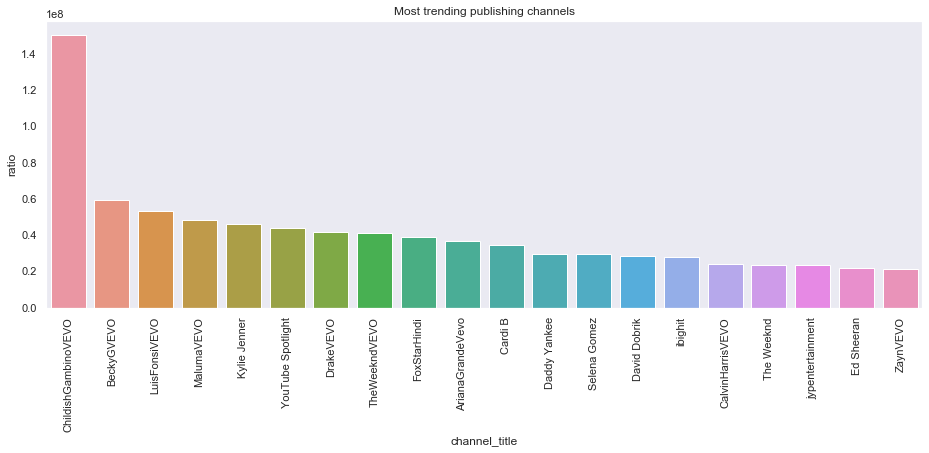
 The top 20 channels we get are:

Figure 5.b

We get a lot of channels which are from the entertainment category - which fit the number of videos from this category and the interests of USA in this category.

In the future we can separate this graph for each category and can recommend for publisher in which channel he supposed to post depends on his topic.

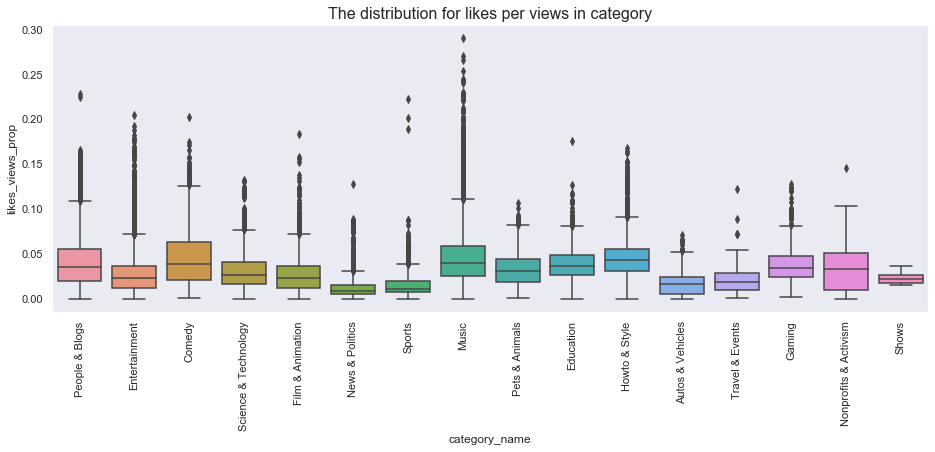
here we can observe the distribution of the proportion of people that liked the video from the total number of viewers for each category. This can show us the distribution of the active viewers per category.

Figure 5.c

We see that many people like the Music category.

But surprisingly, also the Comedy and How to & Style getting a lot of likes and views - this is a surprise because Comedy is in the 8th place in the amount of views - means this category get a lot of likes, and How to & Style got in the 11th place

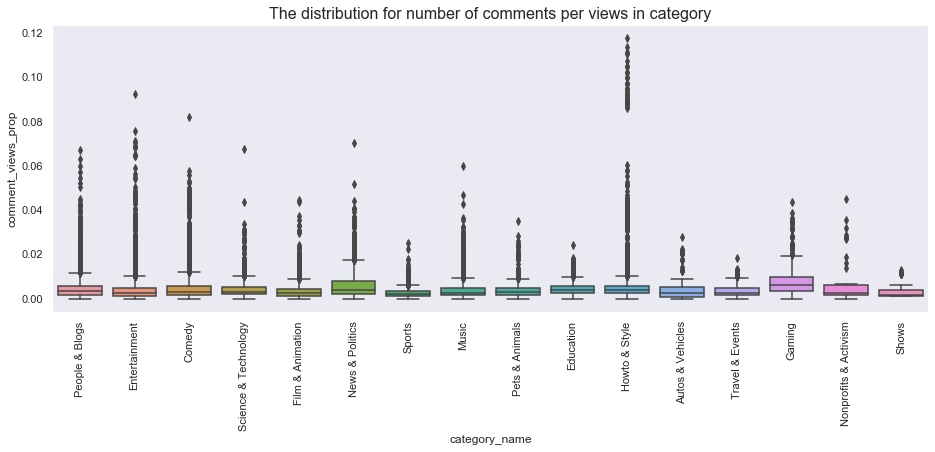
****here we can observe the distribution of the proportion of people that commented on the video from the total number of viewers for each category. This can show us the distribution of the active viewers per category

Figure 5.d

Also, here we get a surprise! The News & Politics and Gaming categories get the highest median of this proportion, if we add it to the fact that Politics was in the last place in the number of views per videos in category - means that this category gets a lot of comments. This makes sense with probably the multiple opinions in this topic.

Also, the low rank of Music category is not a surprise because, music videos don't usually have as many discussions as politics.

Conclusions (base on our data only):

* Music is the most watched category in YouTube.
* Comedy People & Blogs and Music are the most liked categories.
* Gaming & Politics are most commented categories.
* Each feature of success we chose - define different category as the most successful.

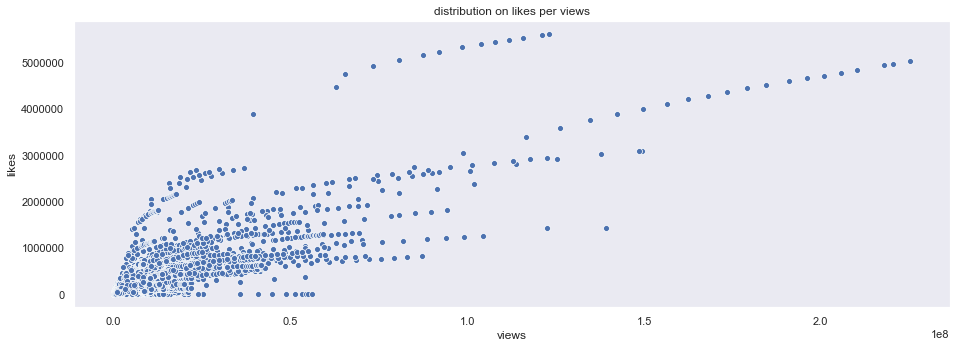
**Relations between features:**

## Likes and Views:

We want to explore which features can predict video's success except number of views.

Number of likes can be a good candidate for this. If a person likes a video, he wants to publish it, so he presses the like button.

Figure 6



We can see, as we except, there is a good relation between those features.

Where there is low views also low number of likes.

(Also, this relation is causal, because if someone likes a video, he must watch it first)

## Dislike and Views

When a video gets popular a lot of people see it, also as the number of people who like the video increase so as the number of people who don't like the video increase.

There are publishers that all they want is to be talked about and don't care about the opinion.

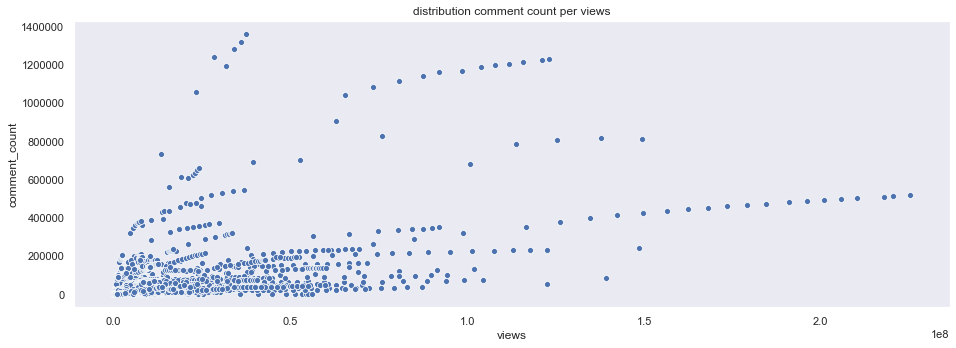
So perhaps there is a connection here also:

Figure 7

Here we see weaker connection than in the like case.

When there are low number of views also there's a low number of dislikes. In the higher numbers of views there aren’t many dislikes.

**Hypotheses Testing:**

Our main object in this project is to try to define what makes a video a trending video or even viral. Although there is a lot to ask on this subject, we tried to narrow our subject on focus on one of the most defining choices a 'youtuber' must decide- The Title.

Choosing a title is a crucial part in the video process. The title set the tone of the video, it's one of the first things people see and might even determines if it will be a success or a failure.

In the [preprocessing part](#figure4b) [[1]](#footnote-1) of the project we saw which words were the most popular in YouTube. At this part we will look at the word 'Official' (The third frequent word)

**Does the title include the word 'official' gets the video more popular?**

We chose to define video's popularity with the view feature - with bigger exposure the popularity rise.

For start we look if word's popularity is due to the bias of 'Entertainment' category's popularity in the data frame.

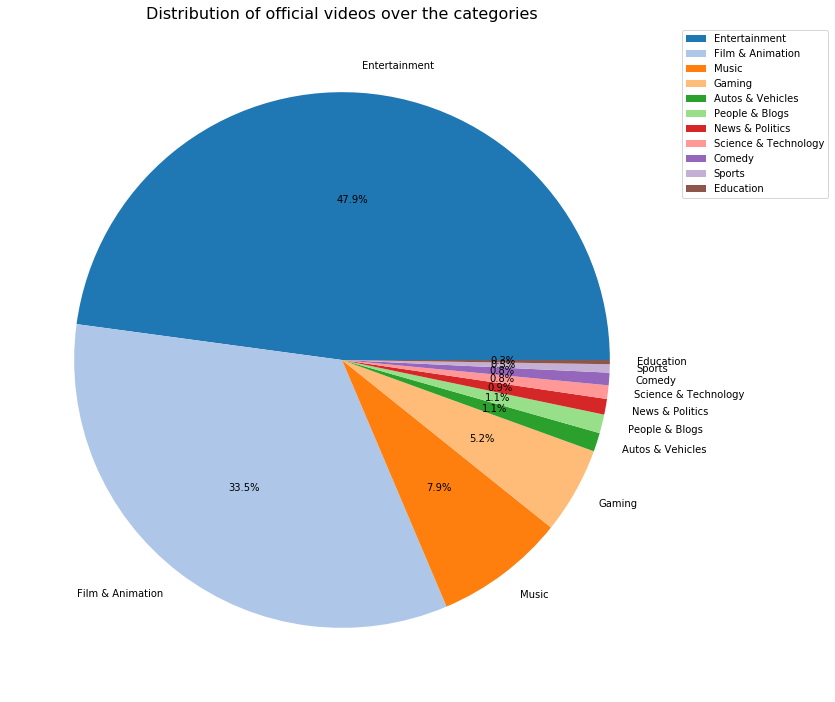
We look for distribution of videos with 'official' in their title over the different categories:

Figure 8

There is a lot of official videos in the Entertainment, Film & Animation and Music categories, which makes sense because the Entertainment is the most frequent category in YouTube.

Also, 'official' is usually recognize with; official trailer or official music video, therefore the amount in other categories.

Need also to check the proportion of official videos that there is enough data to conclusion about.

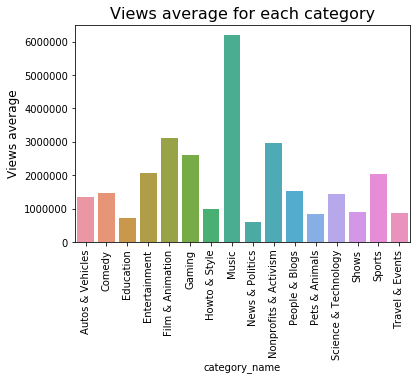
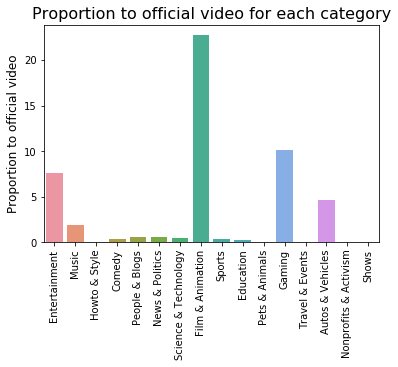


Figure 9.b

Figure 9.a

We can see that each category has a different proportion for videos with 'official' in their title and different average views for each category.

So, if we want to explore the official videos' impact on average views, we can't ignore the category feature.

We assume that the dataset represents the population of all videos in YouTube.

We define our hypotheses:

Null hypothesis: There is no different in average of views between official video or not official in YouTube

Alternative hypothesis: There is a difference in average of views between official video or not official in YouTube

Our test statistic is the difference in the average of views for all videos.

We can calculate one value of the test statistic under the null hypothesis. We shuffle our feature - official in the title or not, and check if the average of views has changed?

We can therefore create an empirical distribution for the statistic, compare to the observed value and find a p-value with reference to the original official video's proportion in each category.

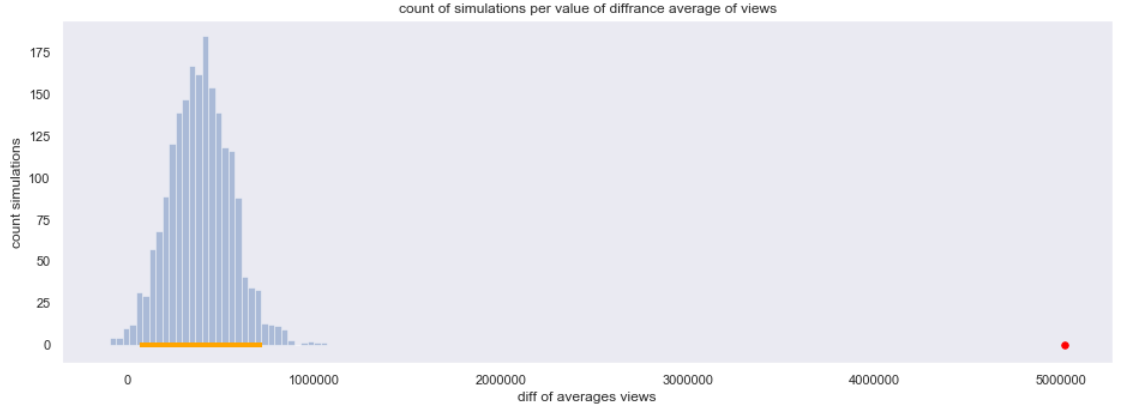


Figure 10

According the P-Value test, we got 0 as it - so for each positive statistical significance (and of course for 0.05), we can deny our null hypothesis.

Also, with the confidence interval test- we can deny our null hypothesis, 0 isn't belong to the empirical distribution given confidence interval.

Conclusion: There is a difference in average of views in videos with 'official' in the title or not.

**Predictions:**

After testing our hypnosis and having a better understanding than before of what can make our video 'go viral' we would like to try to predict base on our features if a video will have a high number of views a low number of views or a medium number of views.

First let's look at the distribution of views:

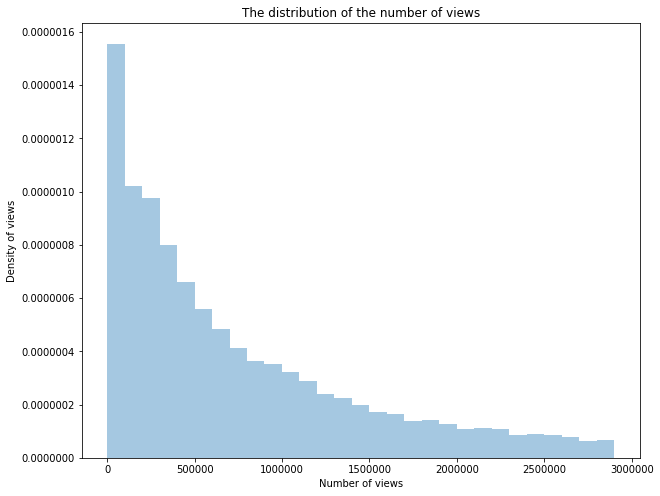


Figure 11

Next, we will look at our features and try to see where we can have a correlation strong enough that we can base our model on.

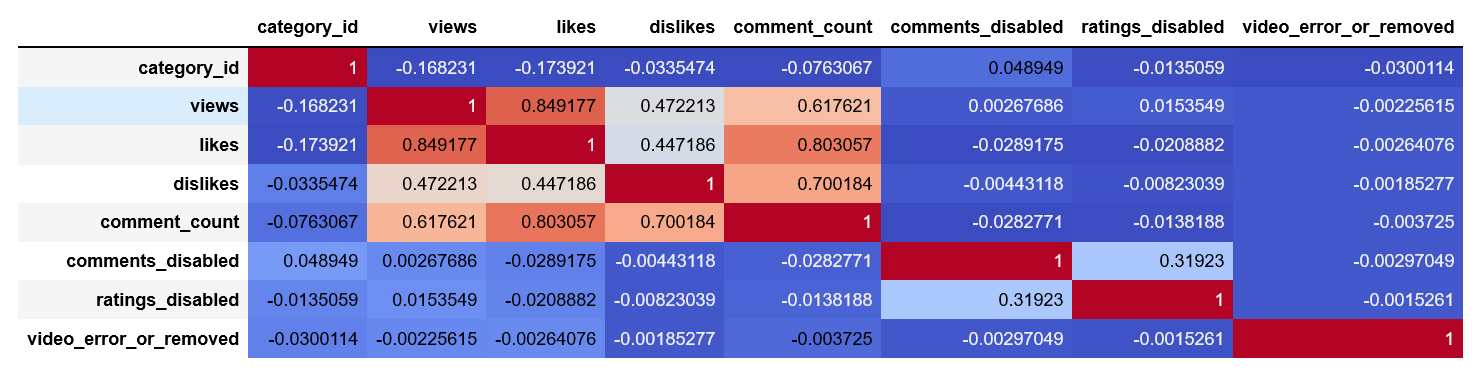


Figure 12.a

We can see that the most corelated part is between the features: views, comment count, likes and dislikes with the lowest correlation between the above.

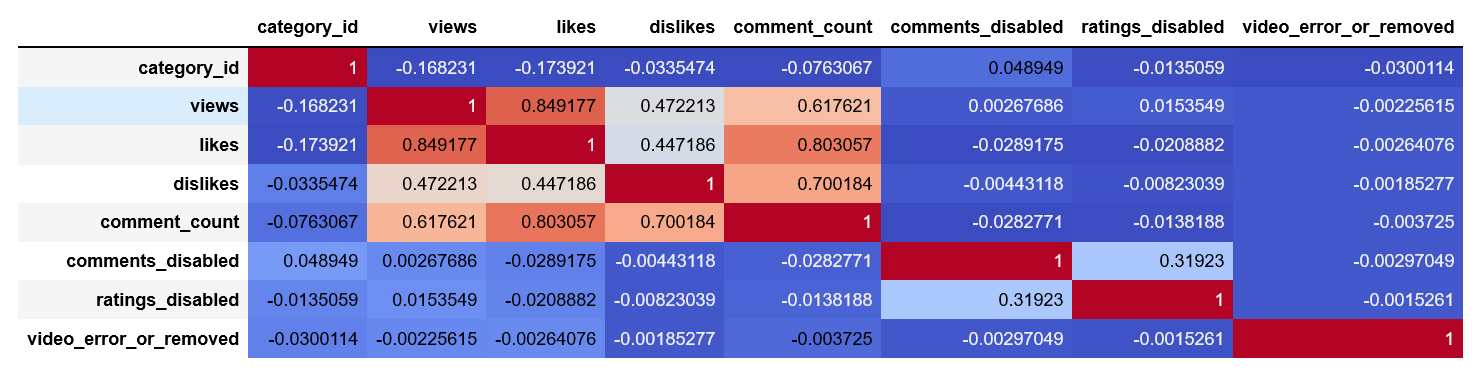
*An enlarged cut:*

Figure 12.b

In order to make classification we split our data into 3 even groups based on the number of views: high number of views, low number of views, medium number of views.

We chose to predict our views category with the features likes and comment count because they had the highest correlation and we assumed that if we had the like feature it won't make any difference to use the dislike feature as well.

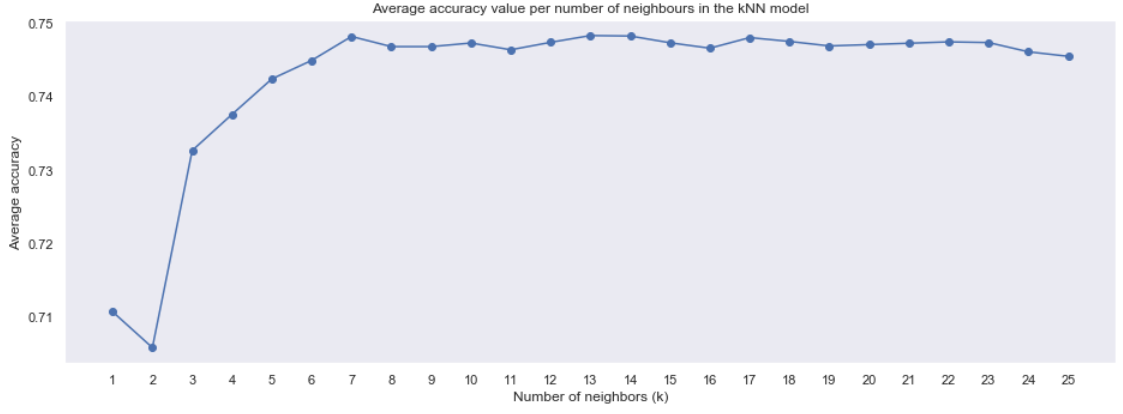
We trained our model on those features and got:

Figure 13.a

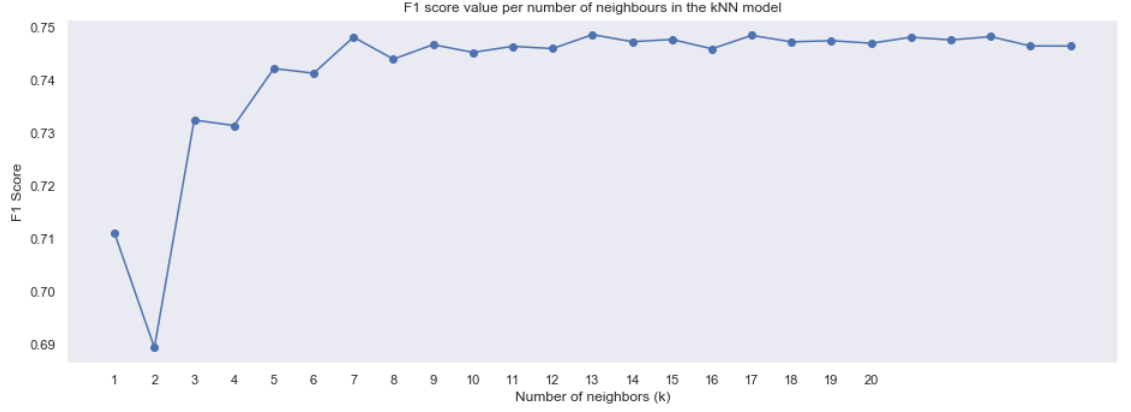


Figure 13.b

Highest accuracy is obtained for k = 13 and equals 0.7489546936043654

accuracy of the classifier is 0.7571428571428571

we weren’t satisfied with the result then we decided to improve our model and try to add the

feature dislikes. Little so we knew it will make a big difference.

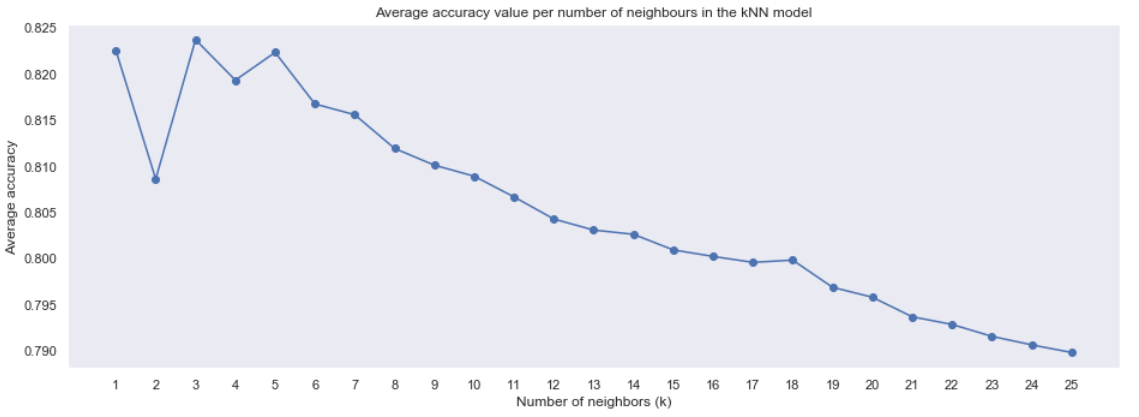


Figure 14.a

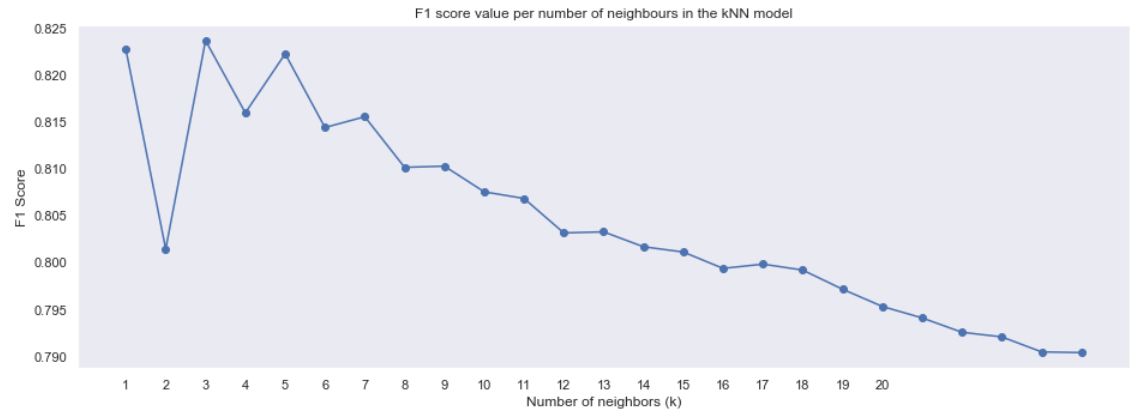


Figure 14.b

Highest accuracy is obtained for k = 3 and equals 0.825849315817776

accuracy of the classifier is 0.8367521367521368

we can see that there was **a rise in the level of accuracy of 8%** after adding the feature dislikes.

Now we can see that the saying: "There is no bad press" is true.

**Reflection:**

**Limitations:**

1. The dataset for USA – most of its videos are published in years 2017-8. That’s why we choose to cut the videos that weren't published in those years. So, all our conclusions are about those 2 years.
2. According the categories histogram: most of the videos are from Entertainment and Music categories – which also the most popular categories. According to this bias we did our conclusions – for example in the hypothesis test.
3. We assume our dataset **is** presenting the “population”- the all successful videos in YouTube USA.
4. In our calculations for proportion likes per views or number of comments per views- we assume there is a uniform distribution between those 2 features – that perhaps not the truth.

**Future Directions:**

1. Why YouTube users decide to disable possibility to comment or rate? Our dataset has a few videos that users choose those decisions. If we have more balanced data, we can explore those questions.
2. Are there specific words in the comments that are common in popular videos? Are there specific words in comments that provoking other comments? Our dataset only has the number of comments and not the comments themselves. If we could have those comments, we could answer those questions.
3. Most of our data is about videos from 2017-8, if we have more videos from other years, we could research about the distribution of videos from specific category over time, or from a specific successful channel.
4. In our dataset there are a lot of various tags that we can’t learn from all of them, because there are a few from each so, we can't make any conclusion base on such little data. If the dataset’s feature tags were more representative, we could ask and learn the same questions as we ask about title for example.

Thanks for reading our report. We hope it answered on all your expectations and kept you interested and wanting more.

1. Figure 4.b [↑](#footnote-ref-1)