Unsupervised learning
class of 2020

Submitted by: Tal Garber 322853524 & Tal Trinquart 322893785,
undergraduate math students
Date: 30.04.2020
Github: Here

Table of contents

Introduction

The data set we've worked on is a daily record of the top trending YouTube videos between the years 2017-2018 (with some exceptions).

For each Trending video, we have its trending date, title, channel title, category, publish time, tags, number of views,likes, dislikes and comments.

With all this information, and many python tools (such as: pandas, scikit-learn, matplotlib, seaborn, and much more), we tried to find interesting facts and conclusions using various data science algorithms we've learned during the course.

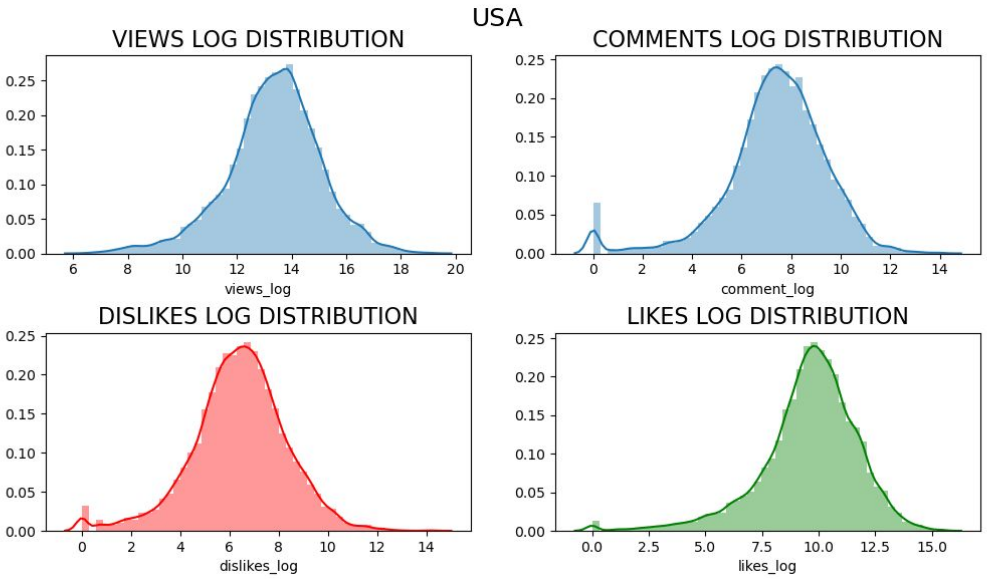
We aimed to find the most successful categories of videos in relation to all of the features that we have, as well as finding conclusions about the nature of the videos and the way the likes and videos behave together.

Important to notice:

* To run the code from Github, read the README file here first.
* The PCA algorithms work only on the USA, Great Britain, Germany, Canada, and France data sets, as the other are newer and formatted differently.
* The DBScan implementation of Scikit-Learn is very heavy and we couldn't run it with the whole data, so we used 40% sample to run on.
* We fixed the random seed to 0.

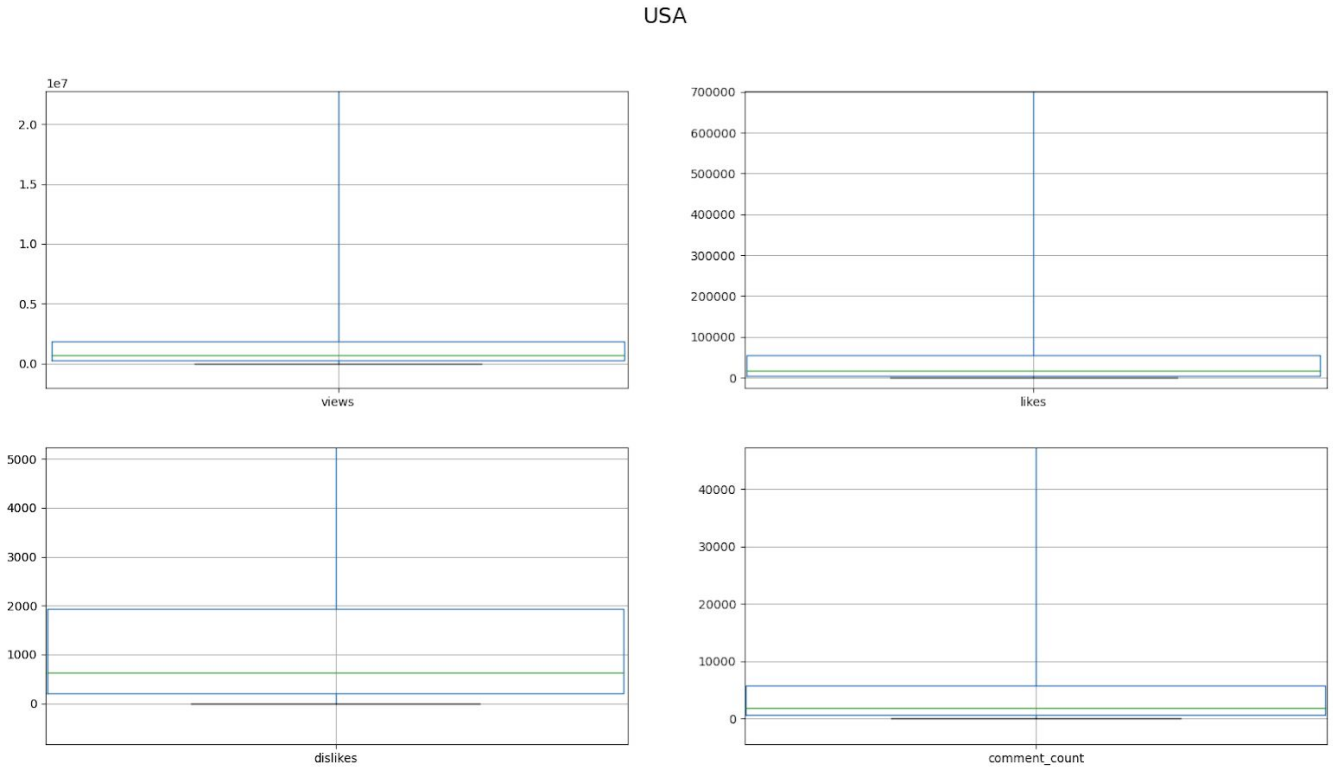
Data Statistics

Data Distribution



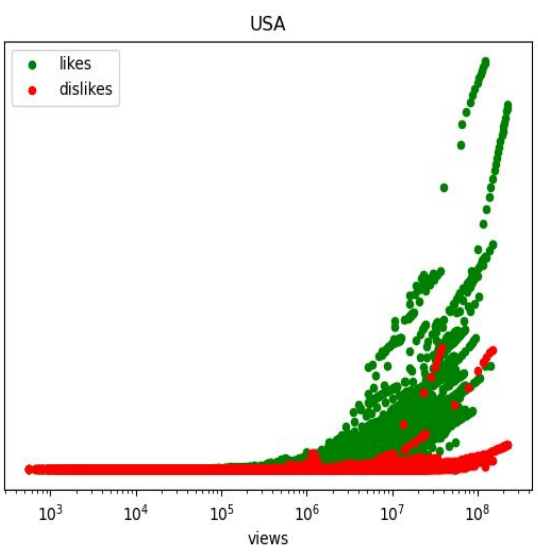
• As we can infer from the image above, our 4 main numerical features: views, comments, dislikes and likes, has Log-Normal distribution.

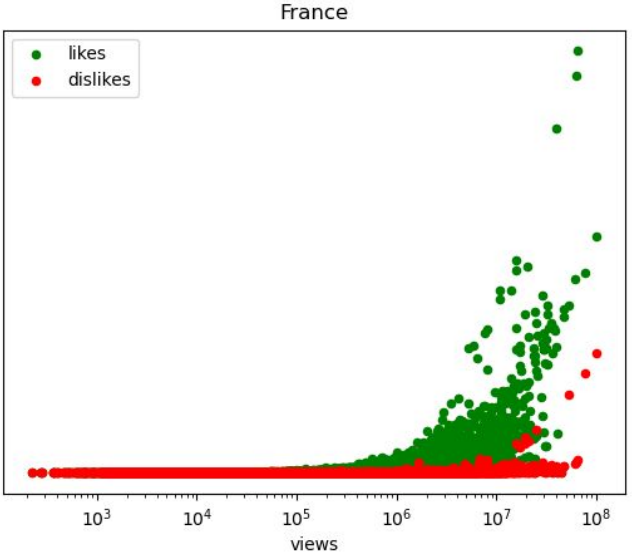
BoxPlots



• Boxplots for USA views, likes, dislikes and comment count. Since the median is really low, and the max is really high, we can't see the whisker of the max value. The values of views are: Median: $700,000$; Min: 550; Max:$225,212,500$

Likes and dislikes vs views





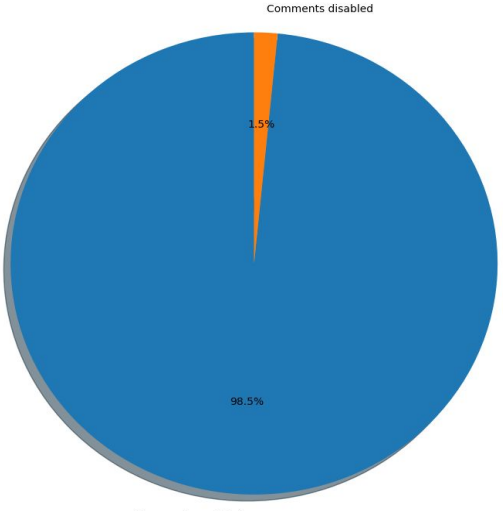
• Here, we can see that the likes and dislikes are exponential relative to the views in both the USA and France.

Word cloud of the titles

Most Popular Words in Title (USA)

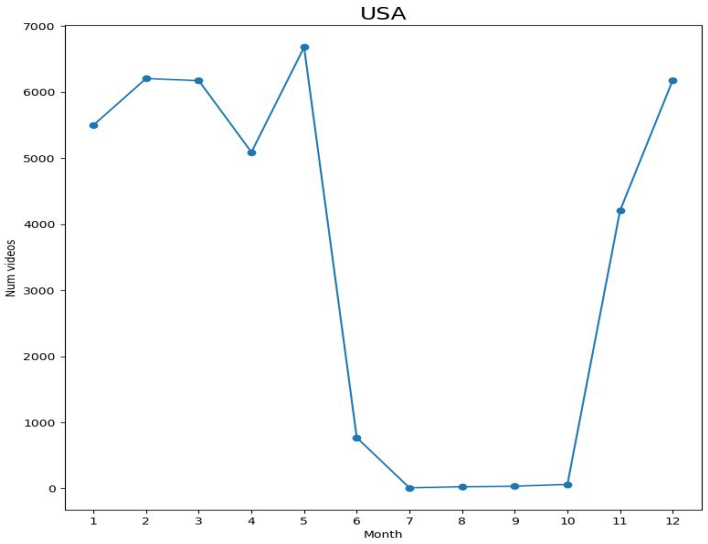
• We can see that the most popular words in titles are those who appear in music videos and trailers - “Official”, “Trailer”, “Video”, “Lyrics”, “Teaser”, ETC.

Comments enabled vs disabled

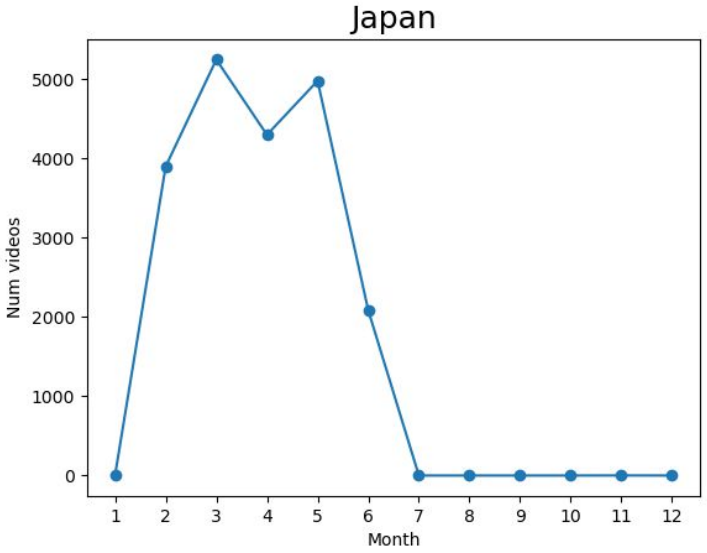


• In the USA, as well as in other countries, most of the of the videos have the comments enabled, and only a small percentage don’t.

Publishing months

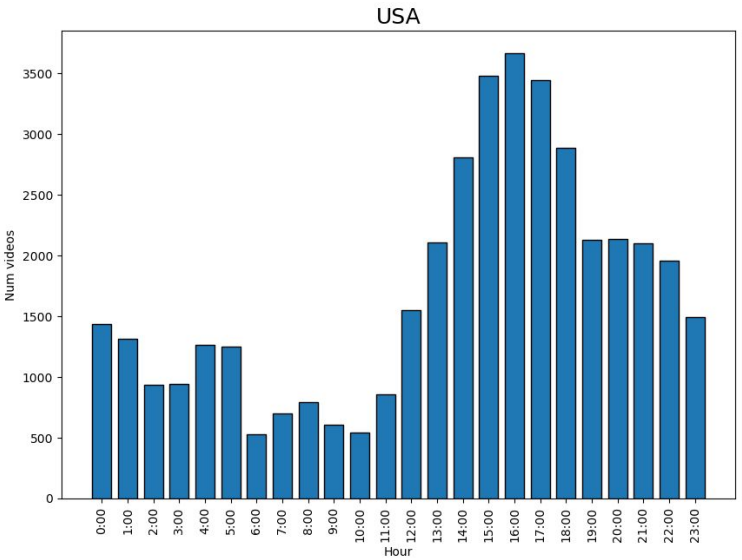


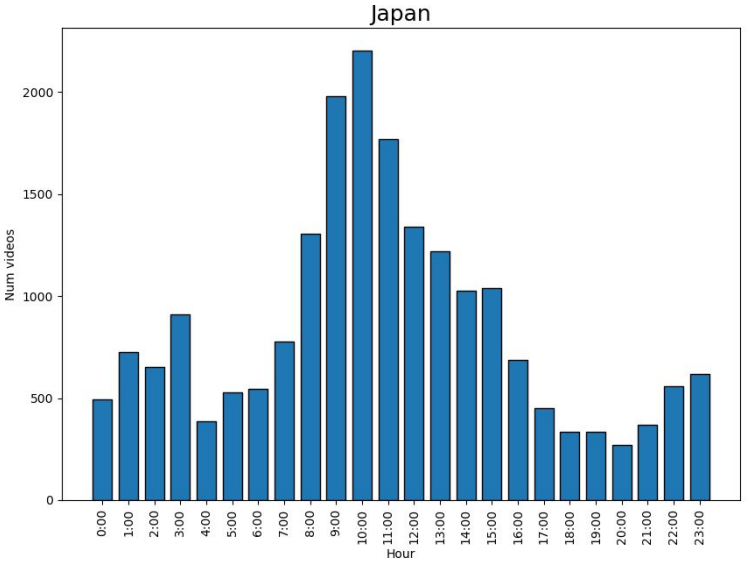
• In the USA, most of the trending videos are published in this first half of the month, and in the last quarter. These are times of christian holidays.



• In comparison, Japan doesn't have the christmas time bump of trending videos.

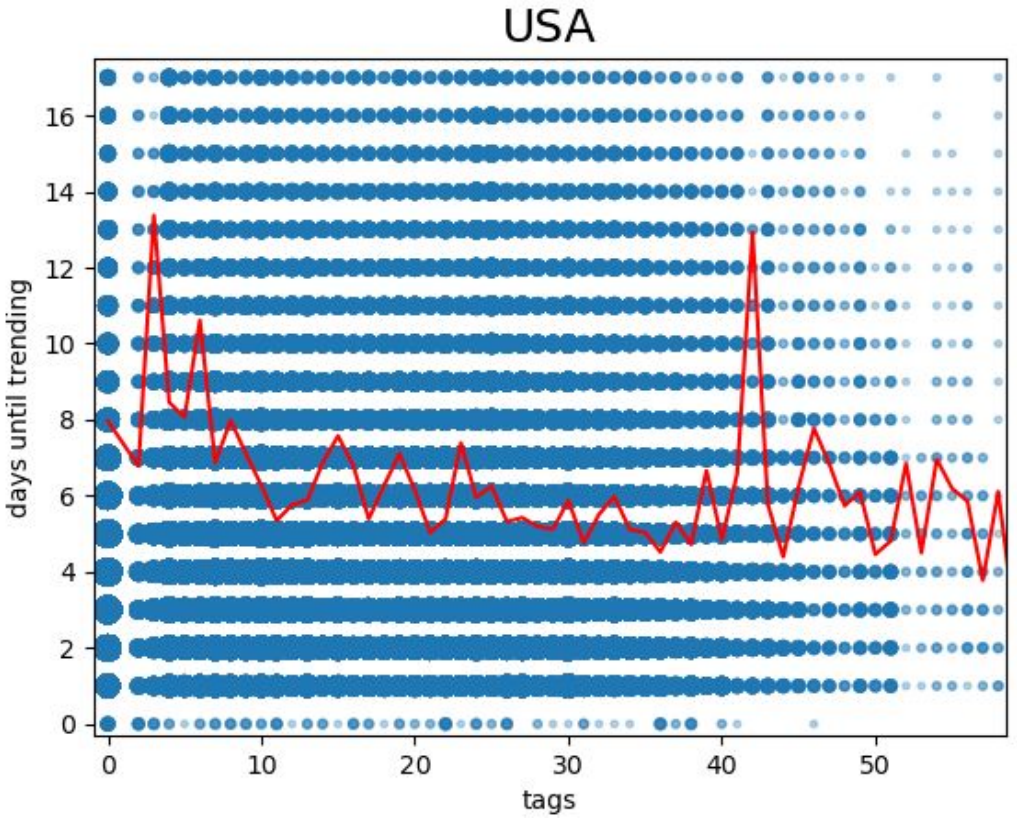
Publishing hours





• When is the best time of the day to upload videos in the USA? It's 16:00, according to our data.
On the other hand, in Japan the best time is actually in the morning, at 10:00.

Days until trending by number of tags

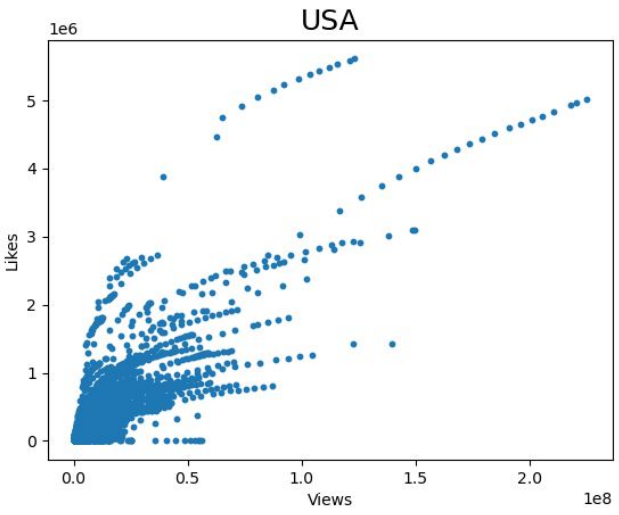


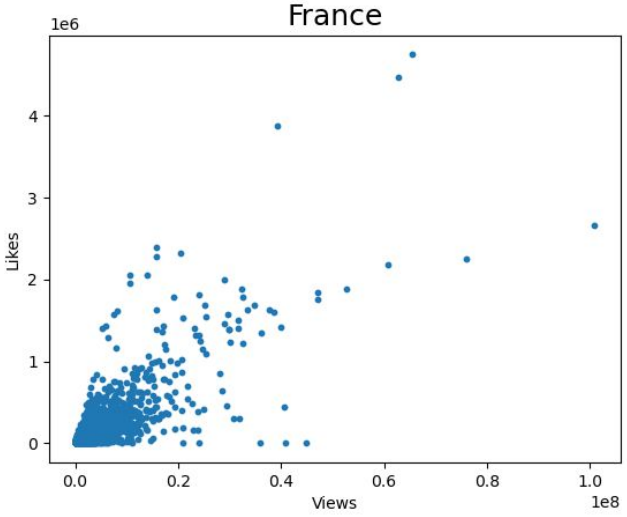
• This graph (zoomed to the interesting part) shows us the relation between the number of tags and the number of days it took for the video to hit trending. The bigger the dot, the more videos are there. As we can see using the red line - a weighted average of each column with the size of the dot (num. videos), most of the videos with 10 to 40 tags reach trending in about 5 days, and there are 2 spikes where videos get more delayed until they hit trending. Content creators should avoid these points. The tags are supposed (as described by youtube) to expose the video to more people (that are interested in these tags) - which means the more the merrier, yet we can see that still most videos have just 0-20 tags and not more.

Clustering

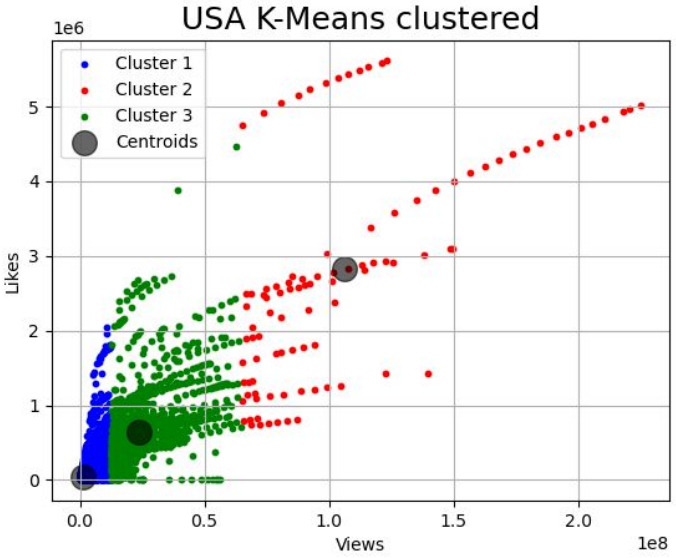
• For this part we've tried to use various methods to cluster the part of the data that describes the relation between views and likes of videos, and after that to compare results between different countries.

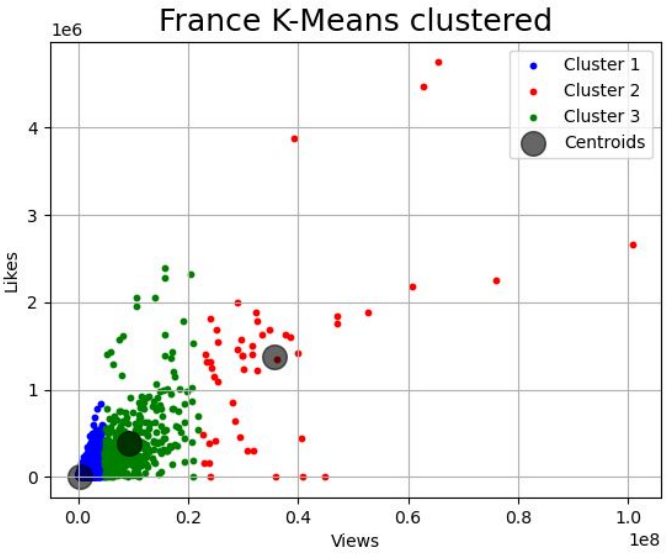
• Let's compare the result between USA and France:
The likes as a function of views:

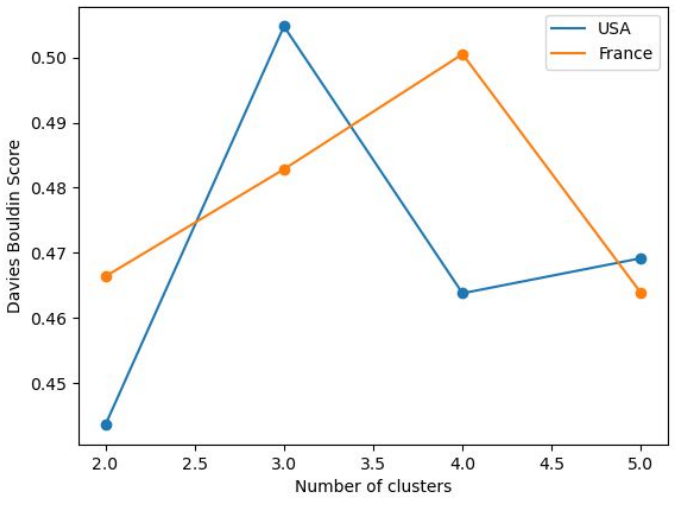




K-Means

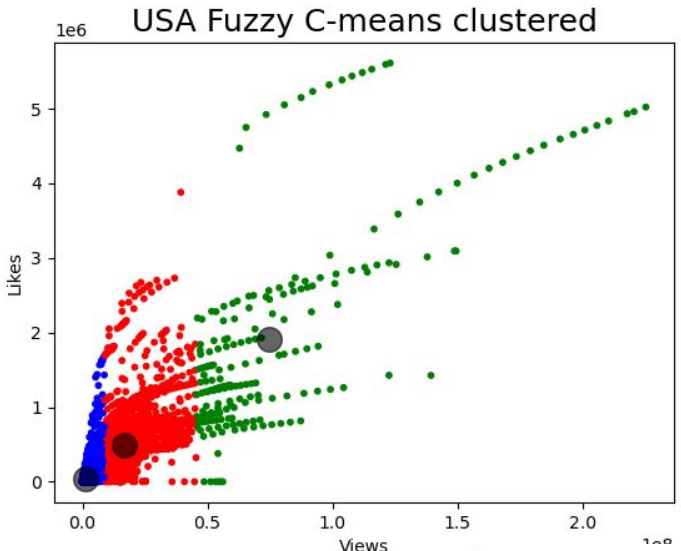


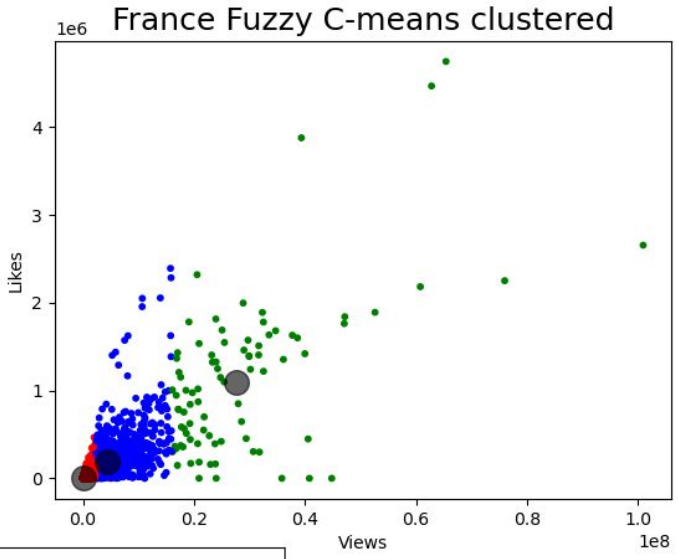


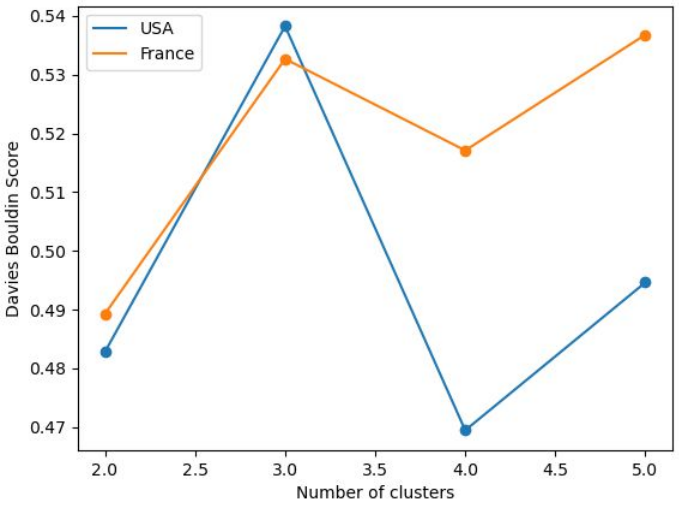


• We chose 3 clusters because we found it as the “sweet spot” between good clustering for conclusion, and good DB score. We initialized the algorithm with seed 0 for the random generator. The results of USA and France are very similar- one cluster of high density, another of less dense points, and a third of very spread dots. We can conclude that most videos on trending have rather little views with little likes, and as the views increase, the likes spread more in 2 stages - cluster 2 and cluster 3. We can also see that in the last cluster - the high views videos, the likes spread much more; there are high views videos with little to none likes, and high views videos with huge amount of likes.

Fuzzy C-Means

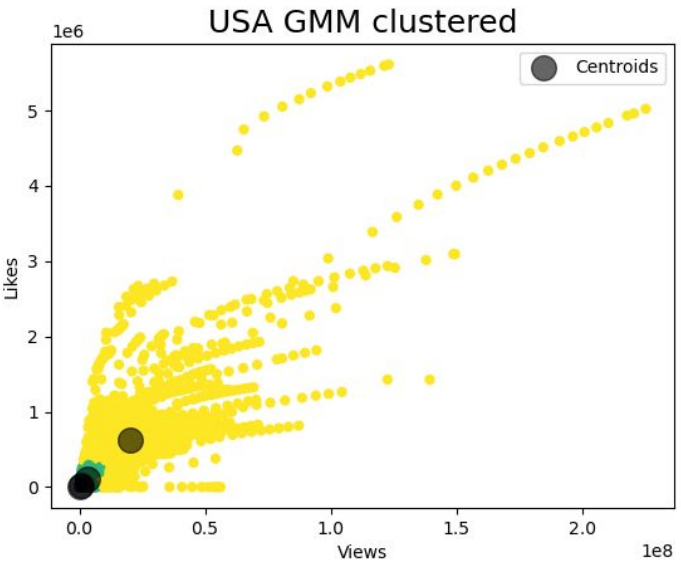


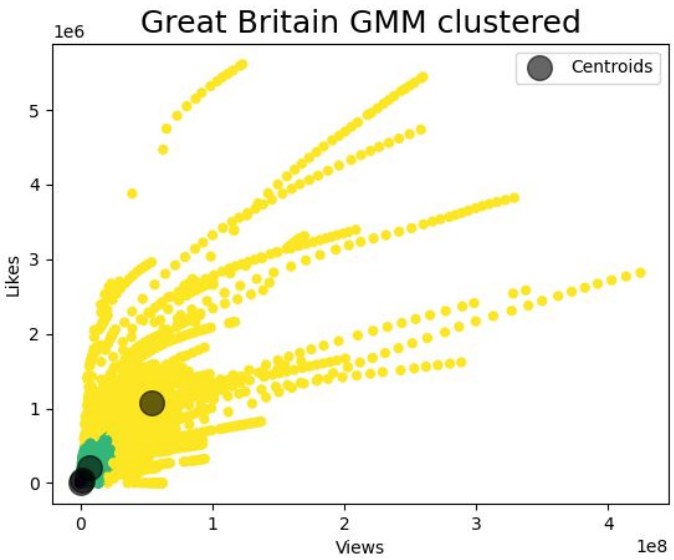


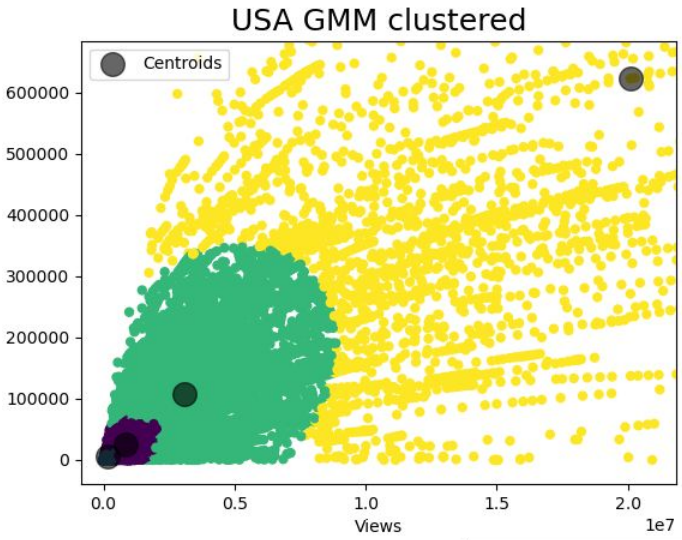


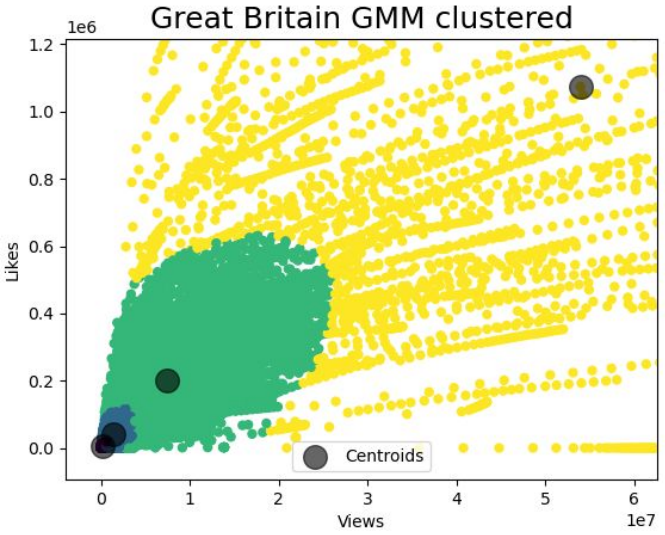
• We can see that Fuzzy C-Means gave us results that are very similar to those of K-Means - 3 stages of spread of videos. We can see the difference between the two algorithms in (for example) the small line of dots in the USA plot (left side red, right side green) - the line can belong to either the red cluster or the green cluster, it’s a ‘soft point’, as opposed to the stiffed K-Means which out the entire line in the green cluster.

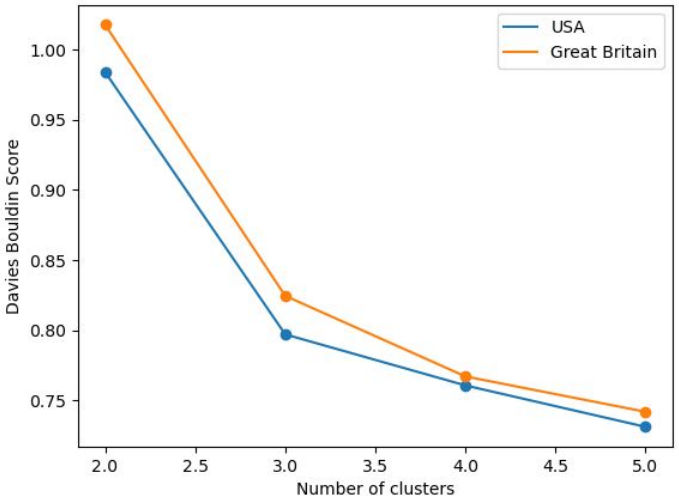
Gaussian Mixture Model





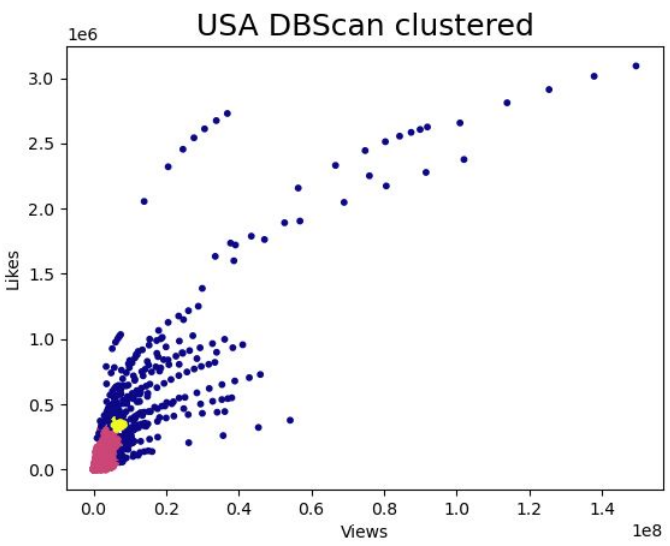


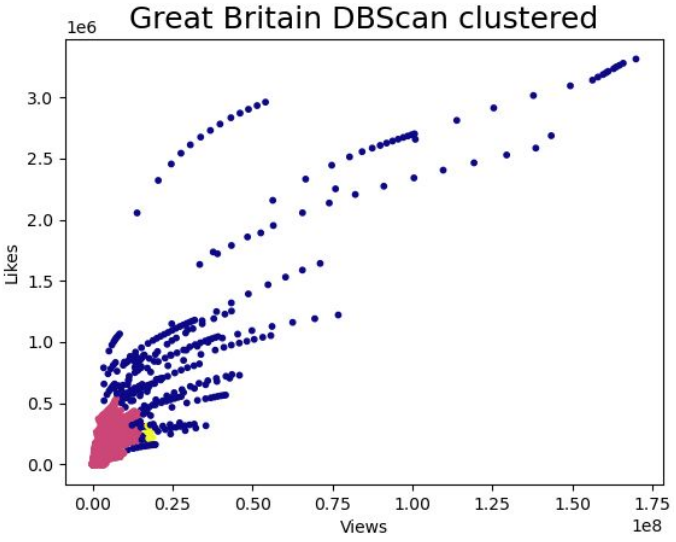




• For the GMM algorithm, we used the data of USA and Great Britain to show how similar they are. Here, we chose 4 clusters that are ellipsoid-shaped as the algorithm works. The spread is shown here by these 4 ellipsoids - each one larger than before. We can see that the clusters of Great Britain looks just like those of USA, we'll see that again in the Statistical Tests part. As a comparison to K-Means, we see that the difference is in the clusters' shape, but the general division is still similar.

DBScan





• DBScan algorithm is density based, and it shows us something very interesting. Although the other algorithms show us that the likes VS views are very similar between USA and GB, DBScan shows us that there is a difference in the density of the points. DBScan still gave us 3 clusters - but they are very different. There is a very dense cluster, and a cluster that isn't really dense. In USA you can see that the first cluster is really smaller than the one in GB, and the second one is bigger - which means in the USA there are less videos with low likes and views, and much more videos with high views and spread of likes (low amount and big amount). In addition, We can see that each one has a third cluster - a small high density one inside another cluster. It shows an irregularity in this distribution. In the USA it's videos that have about $10,000,000$ views and $400,000$ likes, and in GB it's in $24,000,000$ views and $250,000$ likes.

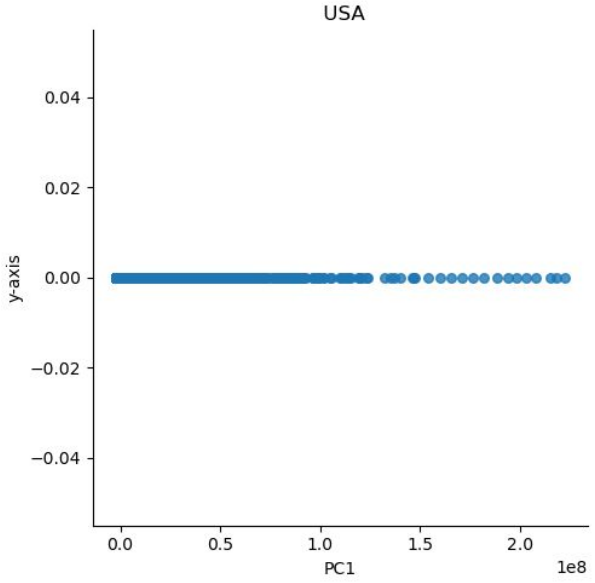
Conclusions

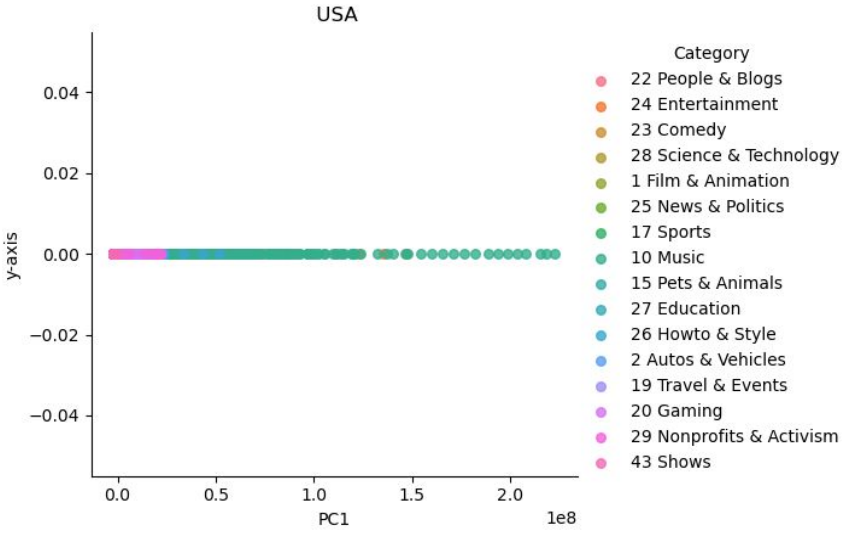
• We can see a similarity along all the results - the data is divided to 3 - 4 clusters that show how the spread of the likes grow as the views get larger. The algorithms shape the clusters a bit differently, which gives us some additional conclusion (E.G. the small cluster in DBScan). We can also see how similar the spread is between different countries - but how the density vary as well.

• We used the Davies-Bouldin index method to evaluate the clustering (For the USA data). The results are: Kmean: 0.504; FCM: 0.538 ; GMM: 0.760 ; DBScan: 1.11
We can infer from these results that Kmeans gives us the best clustering results for the USA data. As described above, FCM is also really similar and got a very close result. This evaluation doesn't work well with DBScan, which is why we got such a high score.

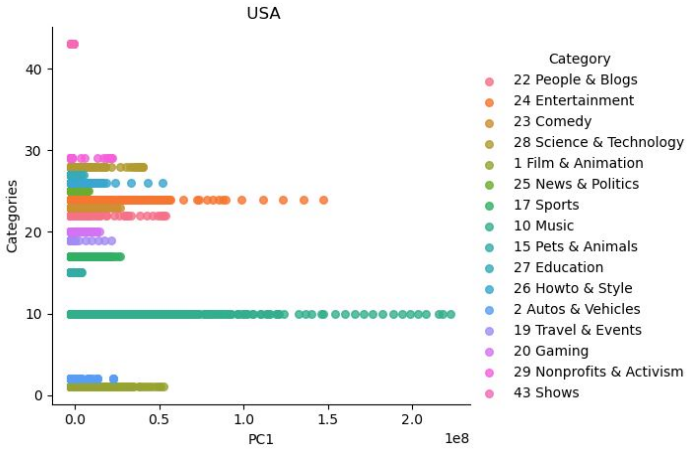
Principal Components Analysis

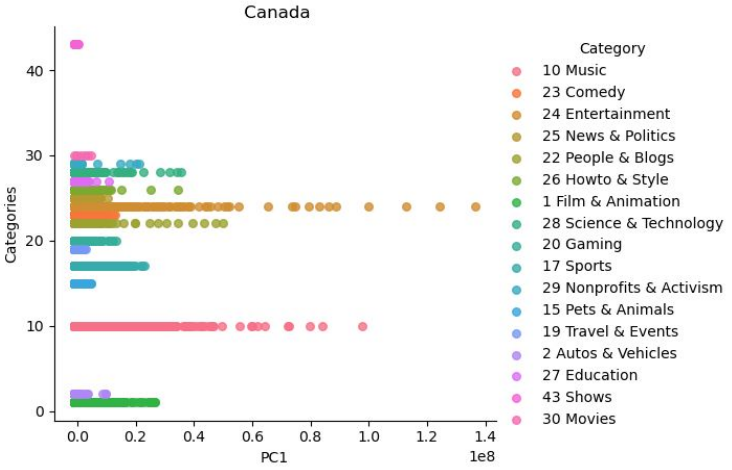
• In the following PCA plots we managed to determine what kind of videos are the most popular.
We used PCA to reduce 4 dimensional space made of the views,likes,dislikes and comments count features, into 1 dimensional vector, and then we colorized each video with different color that represents it's category.





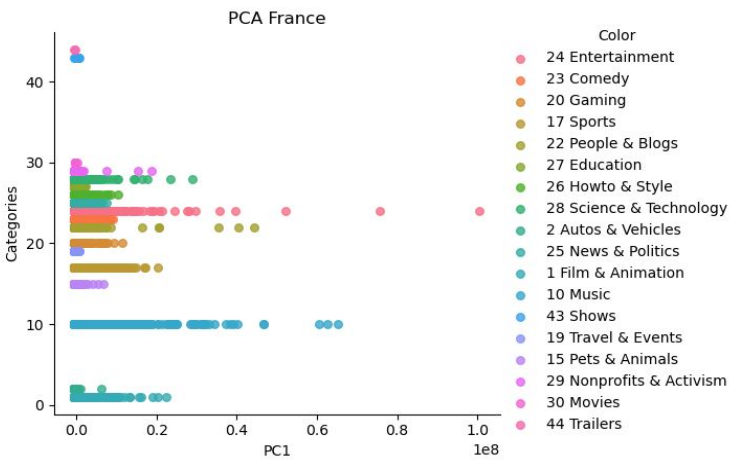
• In order to see the differences more clearly, we spread the vector with categories references:

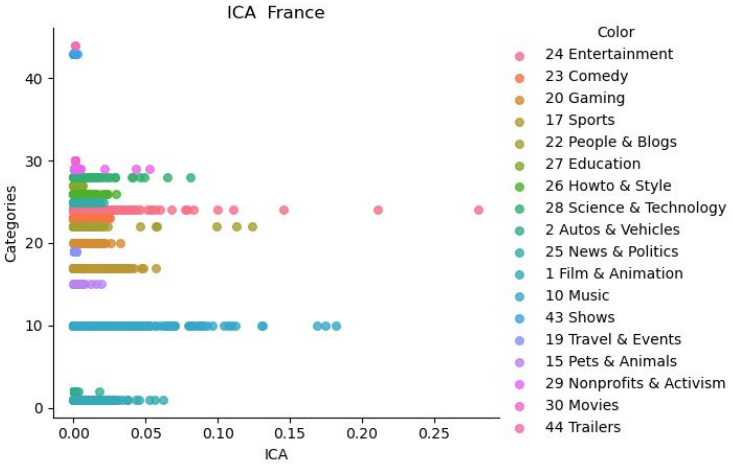


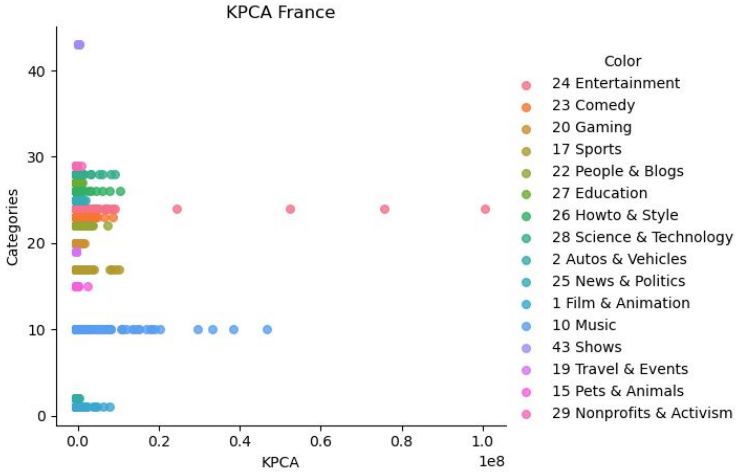


• In these graphs we can see that in the USA, the most popular videos are the music videos; but in Canada, music videos are the second most popular, and the first are the entertainment videos.
Moreover, we can see that the higher we go on the USA music graph, gaps between videos gets slightly bigger, unlike Canada's entertainment videos, where the gaps between each video are much bigger. From this result we can understand that there are more music videos that reach high amount of like and views then entertainment ones

PCA, ICA and KPCA

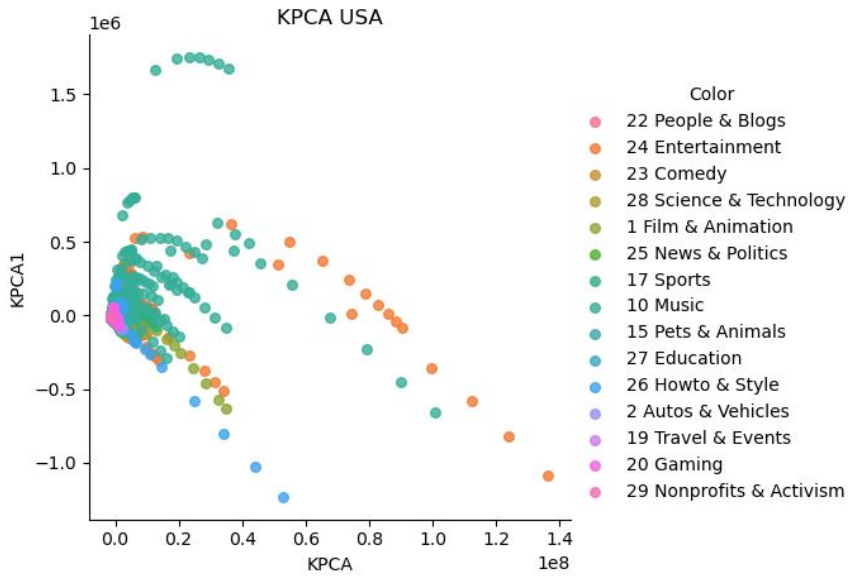




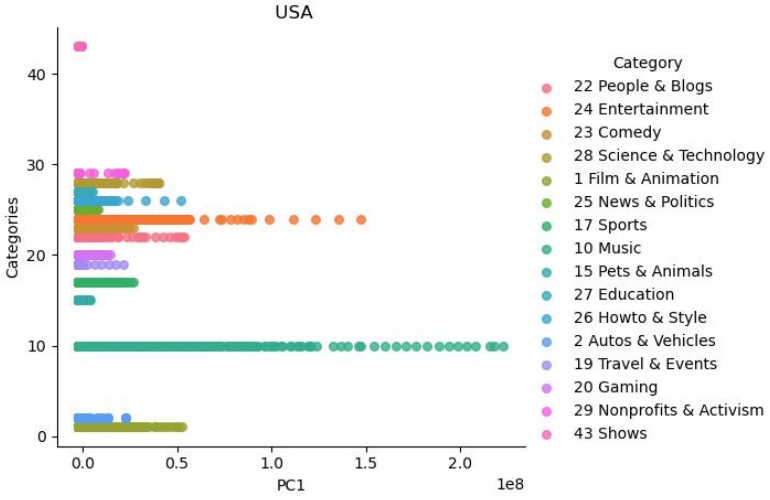


* In the KPCA we set the kernel to linear
* KPCA implementation of Scikit-Learn is so we used 20% of the data to run on

• Like the previous example on USA and Canada, we did the same on France using 3 types of algorithms this time.
As we can see from the graphs, all 3 looks very similar.
Since PCA algorithms helps us when we want to find reduced-rank representation of the data, and similarly, ICA helps when you want to find a representation of the data as independent sub-elements. From the result that the graphs are almost similar, we can infer that our elements were enough independent so that there will be almost no differences in the algorithms.



• In the following graph, we can see 2 dimensional space who got reduced from: views, likes, dislikes and comments count using kernel PCA algorithm with linear kernel.
We can see on the bottom right, the Entertainment videos which have more views compared to other videos categories. But we can also see some peeks from Music, News & Politics and Howto & Style. These peeks have the same slope, meaning they have the same ratio of views:likes. Additionally, in the top left corner we can see a unique group of Music videos that has the opposite views:likes ratio. In other words, these are videos that have a lot of likes but small ratio of views.



• We saw from the 1 dimensional representation, that music videos are the most popular.
But using the 2 dimensional graph, we see that there are actually more Entertainment videos with more views and likes overall, then why music videos are still in the first place?
The algorithm recognized that likes and dislikes has more weight than views, and as we can see from the unique peek in the 2d graph, there are music videos with bigger likes:views ratio.

Statistical Tests - Student's T-Test

• We built a code to implement Student's T-Test with 2 samples (Because obviously the whole array of countries differ in the data, so we'll compare just pairs of countries) . The formula we used is as follows: $\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, for $\bar{x_1}$, $\bar{x_1}$, s_1, s_2, n_1, n_2 being the mean, standard deviation and number of points for the sample 1 and sample 2 respectively. After calculation our T value, we want to see what's the probability for it: $2 * CDF(t, \min(n_1, n_2) - 1)$, for CDF being the Cumulative Distribution Function, and the 2 for the two tails of the normal distribution.
Lastly, we check if the probability is greater or smaller than our α, which is our significance level. If it's smaller, then we can reject our null hypothesis H_0 and accept the alternative H_1. If not, we failed to reject H_0.

Test #1 - views in the USA vs France & Great Britain

• Let's define our null hypothesis to be $u_1 = u_2$ for USA & France views mean. We chose our significance level to be 5. Our program gave us the result of 2.0 which is greater than our $\alpha = 0.05$, thus we can't reject our , meaning USA and France have very similar numbers of views

On the other hand, let's try the same with USA & Great Britain, with the same significance level. The result we got converges to 0, which is smaller than our $\alpha = 0.05$, so we can reject our and accept the alternative, meaning USA & Great Britain views are different.

Test #2 - likes in the USA vs France & Great Britain

• Repeating the process above, with the likes instead of views, we got that the same results - accepting the alternative hypothesis H_1 for USA vs. Great Britain, and not being able to reject the H_0 for USA vs. France.

Conclusions

• We can see that these results match our graphs and clustering of views vs. likes. The clusters of USA and France are very similar in shape and size (and so we couldn't reject the hypothesis that the mean is different), while USA and Great Britain clusters are very different in density (in DBScan) and in the size of the middle cluster in the GMM clustering plot.