YouTube is an American video-sharing website headquartered in San Bruno, California. The service was created by three former PayPal employees. Google bought the site in November 2006 for US$1.65 billion. YouTube now operates as one of Google's subsidiaries.

YouTube allows users to upload, view, rate, share, add to favorites, report, comment on videos, and subscribe to other users. It offers a wide variety of user-generated and corporate media videos. Available content includes video clips, TV show clips, music videos, short and documentary films, audio recordings, movie trailers, live streams, and other content such as video blogging, short original videos, and educational videos.

Most of the content on YouTube is uploaded by individuals, but media corporations including CBS, the BBC, Vevo, and Hulu offer some of their material via YouTube as part of the YouTube partnership program. Unregistered users can only watch videos on the site, while registered users are permitted to upload an unlimited number of videos and add comments to videos.

Although the most viewed videos were initially viral videos the most viewed videos were increasingly related to music videos. In fact, since recently every video that has reached the top of the "most viewed YouTube videos" list has been a music video. Although the most viewed videos are no longer listed on the site, reaching the top of the list is still considered a tremendous feat. YouTube maintains a list of the top trending videos on the platform.

According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”. Top performers on the YouTube trending list are music videos (such as the famously virile “Gangam Style”), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for.

Undoubtedly trending YouTube videos is very important for the performance of the company, but what makes a video at YouTube trending? How are the users engaging over time? Are there any differences among countries? Is it possible that the performance of videos in one country affects other counties? What factors affect how popular a YouTube video will be? Those are some of the questions that the following analysis wit try to answer?

YOUTUBE TRENDING VIDEOS DATASETS

There are 2 datasets used for the purpose of the below analysis. The main dataset used is a daily record of the top trending YouTube videos. The dataset includes several months of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), 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. This dataset consists of 5 csv files:

USvideos.csv.zip

CAvideos.csv.zip

GBvideos.csv.zip

FRvideos.csv.zip

DEvideos.csv.zip

The second data set (JSON files) includes a category_id field and varies between regions. One such file is included for each of the five regions in the dataset. This dataset is much smaller than the first data set with the videos, it contains just tree columns and 156 rows. This dataset consists of 5 JSON files:

US_category_id.json

CA_category_id.json

GB_category_id.json

FR_category_id.json

DE_category_id.json

The headers in the video files are:

- video_id (Common id field to both comment and video csv files)
- title
- channel_title
- category_id (Can be looked up using the included JSON files, but varies per region so use the appropriate JSON file for the CSV file's country)
- tags (Separated by | character, [none] is displayed if there are no tags)
- views
- likes
- dislikes
- thumbnail_link
- date (Formatted like so: [day].[month])

The headers in the comments file are:

- video_id (Common id field to both comment and video csv files)
- comment_text
- likes
- replies

DATA MANIPULATION PROSSES

MERGING THE DATA-SETS¶

To create the full dataset the region's data where uploaded and merged in one data frame. The files with the category information were uploaded and merged into another data set. The category id and category name where stored in a data frame the elements of which were dictionaries. The column instead of pandas series where iteritems. As a result, there were problems to treat the data since there was not a typical pandas series. Below you can have a look of the category table.

		etag	items	kind	country
US	0	"m2yskBQFythfE4irbTIeOgYYfBU/S730Ilt-Fi-emsQJvJAAShIR6hM"	{'kind': 'youtube#videoCategory', 'etag': '"m2yskBQFythfE4irbTIeOgYYfBU/Xy1mB4_yLrHy_BmKmPBggty2mZQ"', 'id': '1', 'snippet': {'channelId': 'UCBR8-60-B28hp2BmDPdntcQ', 'title': 'Film & Animation', 'assignable': True}}	youtube#videoCategoryListResponse	US

After analyzing and merging the two datasets the final data set created contains 142,000 rows and 30 columns

FORMATING – PROCESSING DATA TYPES

The dates columns were formatted to datetime. Also, the hour, day and month info was extracted. The views, comments, likes columns were changes to integers.

DATA CLEANING¶

To make the necessary changes in format some rows had to be deleted since there was wrong information stored, fir instance date in some cases. Also after merging the two data frames non-available values were replaced with zeros (df_videos.fillna(0, inplace=True)).

Cases where there are 0 likes or comments and a lot of views (i.e. the users have decided to disable those features) have not been removed.

Cases where there were mistakes in the data ex. text in the dates columns have been removed

YOUTUBE TRENDING VIDEOS WHAT DO DATA REVEALS

ENGAGEMENT METRICS ANALYSIS

There are 4 main metrics that are taken into account to determine that a video is trending. As we can see from the table below those metrices are views, likes, comments count, dislikes.

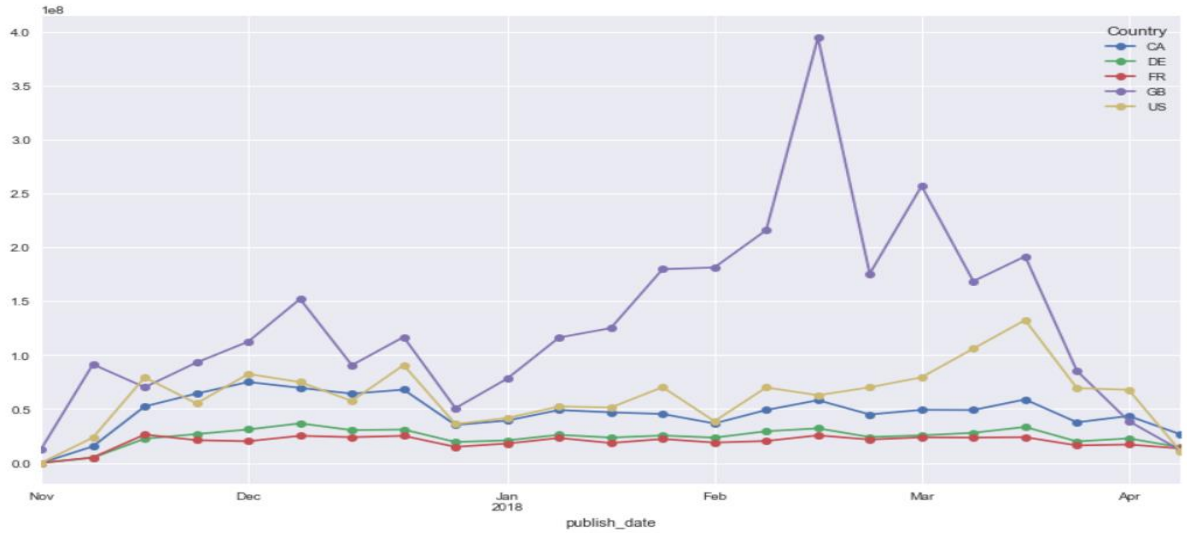
The countries in the dataset are very different in terms of numbers for those metrices. GB is the country with the highest number of all engagement metrices, then comes US and CA.

country	likes	views	comment_count	dislikes
CA	1080.0	31310.0	130.0	60.0
DE	550.0	15500.0	70.0	30.0
FR	450.0	10610.0	40.0	20.0
GB	3010.0	131440.0	310.0	190.0
US	1420.0	41230.0	180.0	80.0

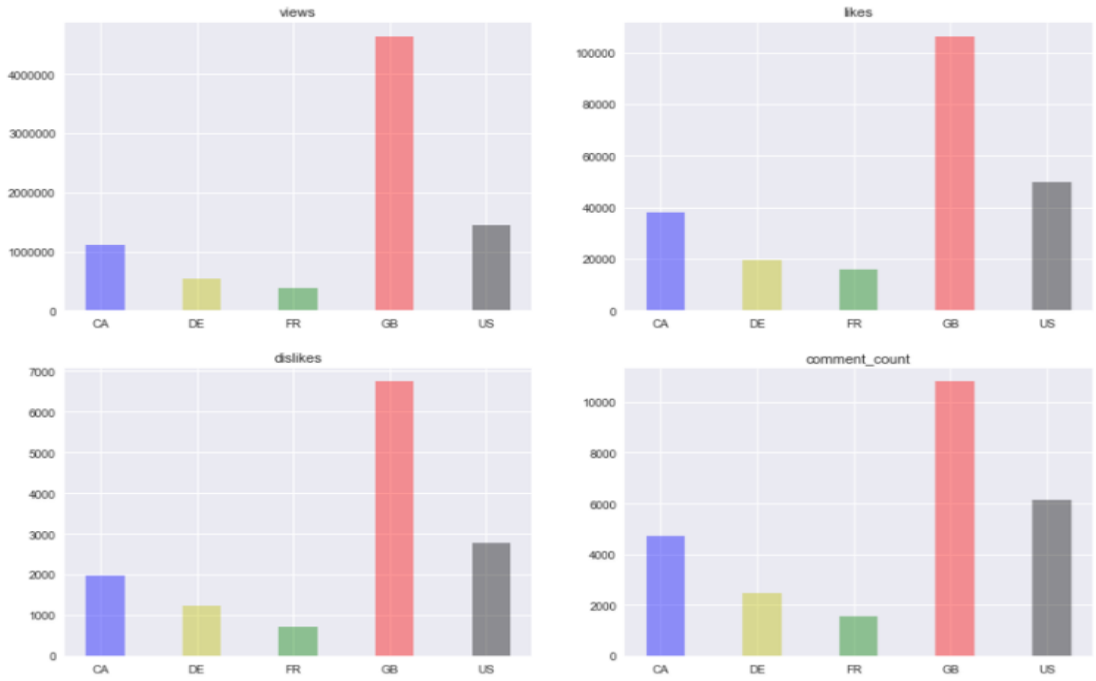
In millions

HOW ARE THOSE METRICS HAVE BEEN TRENDING? WAS THERE A CHANGE IN ANY OF THE COUNTRIES?

UK has been the country with the most likes. There is a huge increase in likes for UK starting from mid January. This could be due to data collection. Canada has shown a decrease in likes for the first months of 2018 compared to November-December 2017. US has shown a noticeable increase in likes in mid March 2018. For France and Germany there are not big fluctuations in likes. The chart below shows the trends of likes for all countries in a weekly basis from November.



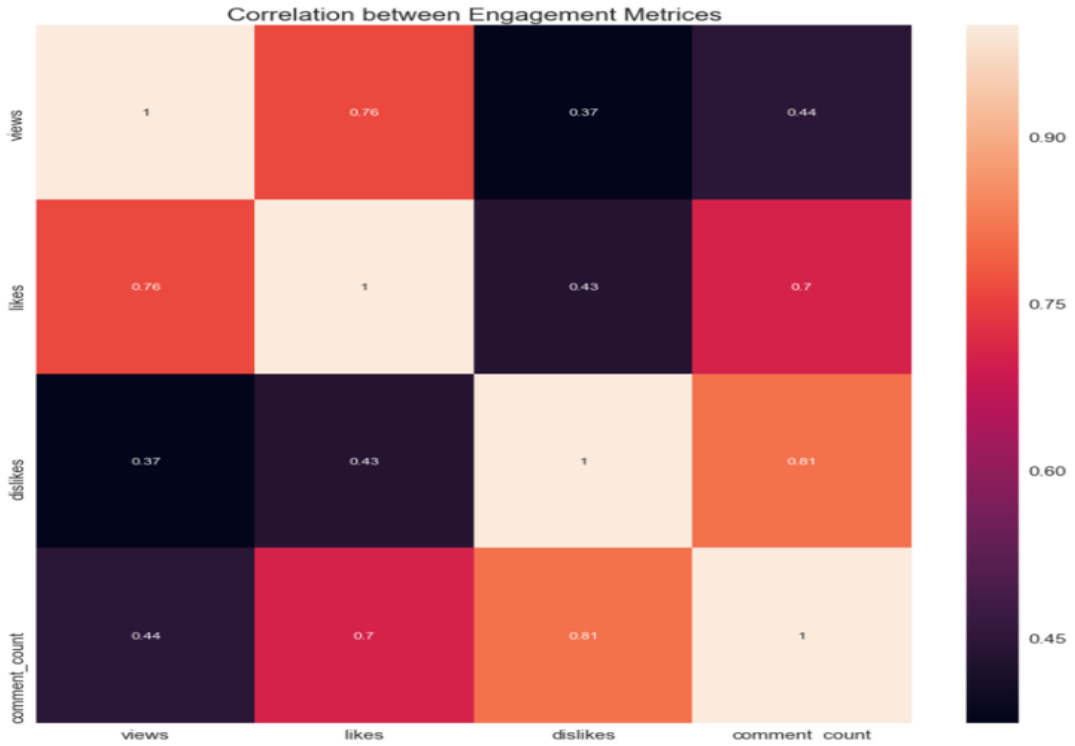
All countries share the similar trend in numbers for likes, dislikes, views and comments. One likely reason to this is due to the video's trending duration. Enduring trending videos have the advantages in getting more views, likes, dislikes and comments. GB is the country with most likes, dislikes, comments and views followed by US, Canada, Germany and France. The bar-charts below shows the total values of engagement metrics per country



IS THERE A CORRELATION BETWEEN THE ENGAGEMENT METRICS?

We observe that all engagement metrices are positively correlated!

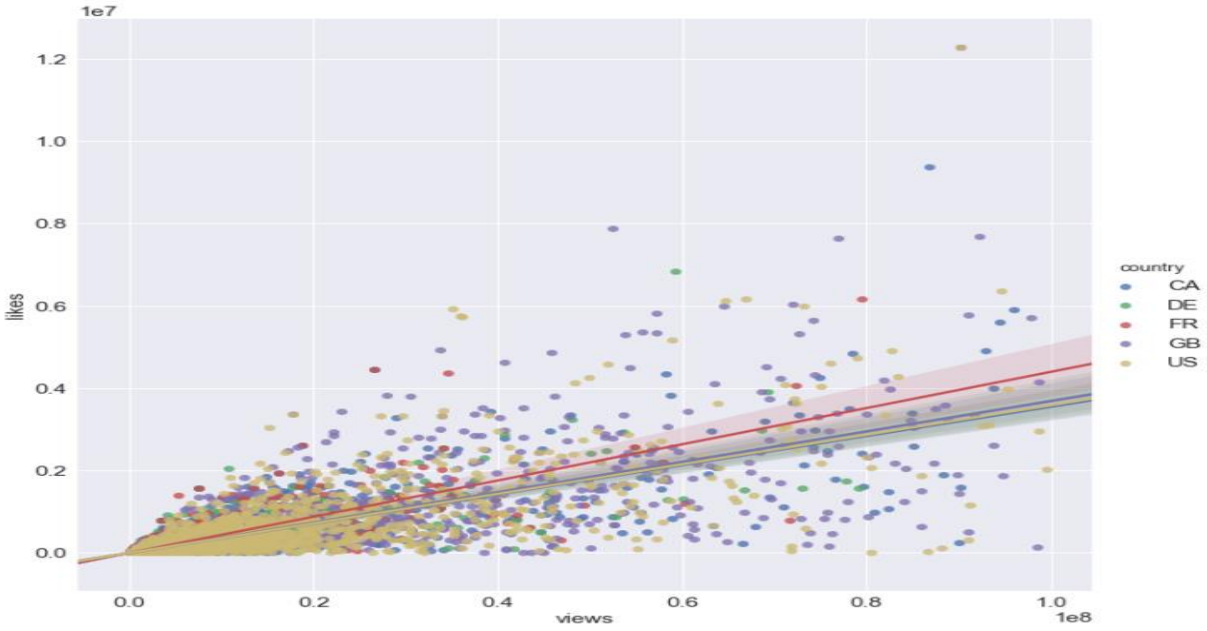
There exists a high correlation between views and comment counts (0.7) and likes and views (0.76). Views and dislikes have the lowest observed correlation among the engagement metrices. The highest correlation observed is between comments count and dislikes. Dislikes and likes and views and comment counts are slightly correlated (0.4).



DO WE OBSERVE THE SAME CORRELATION OF ENGAGEMENT METRICS AMONG COUNTRIES?

As we can see at the chart below France is the country with the highest correlation between likes and views. For all other countries the correlation between those two metrices is very close.

Similar trend is observed for the other engagement metrics combinations.

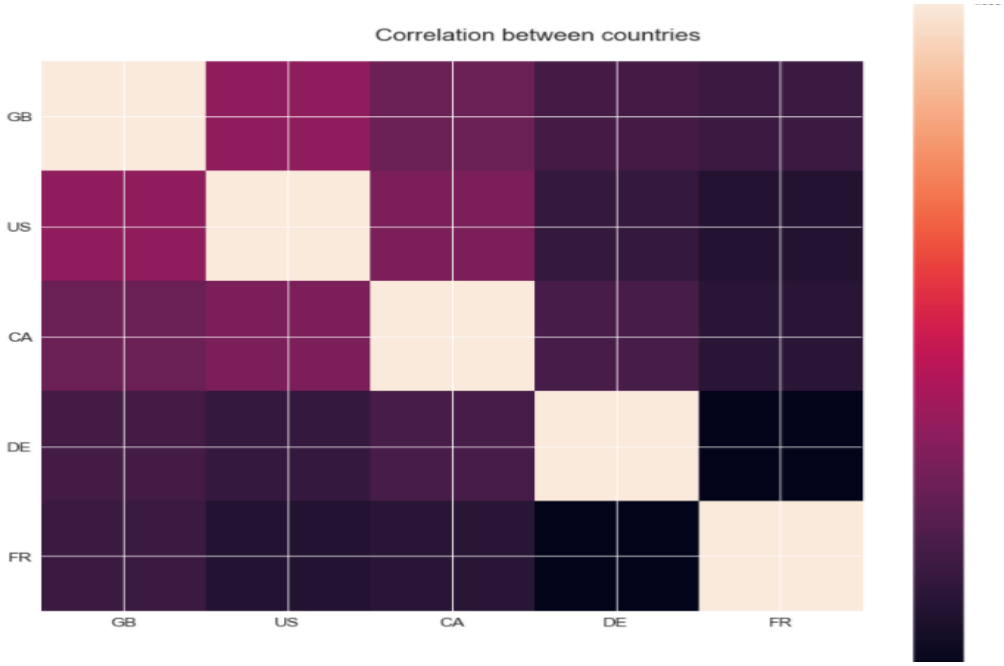


IS THERE A CORRELATION OF YOUTUBE TRENDING VIDEOS BETWEEN COUNTRIES?

Not surprisingly, the videos from United Kingdom, US and Canada are highly correlated to each other in comparison with Germany and France. This might be due to the sharing of common language in these countries.

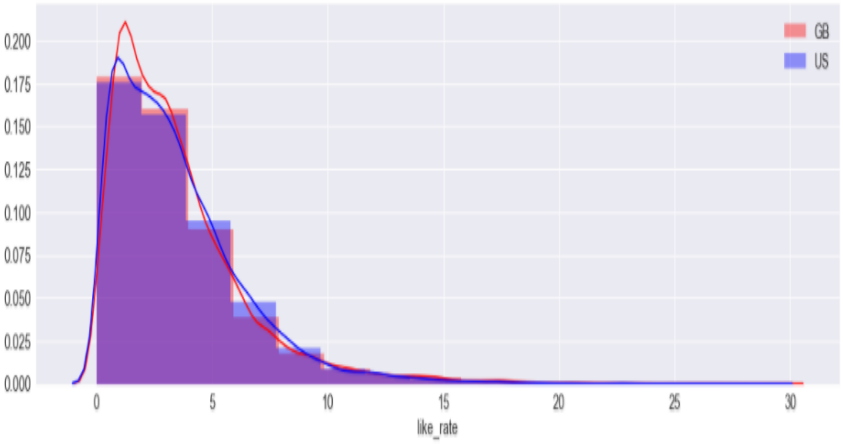
Germany and France seem to be more isolated compared to the rest as they don't follow the same trend as the English-speaking countries do.

This can also explain why UK has the highest number in long-trend videos, as it is contributing by multiple countries at the same time.



GB AND US ARE THE COUNTRIES WITH THE HIGHEST CORRELATION. THERE IS SIGNIFICANT DIFFERENCE BETWEEN UK AND US IN LIKE RATES?

It would be worth to investigate and compare those two countries in terms of likes/views. As we can see from the below bar-chart the two countries follow similar distributions although the mean for GB is higher compare to US.



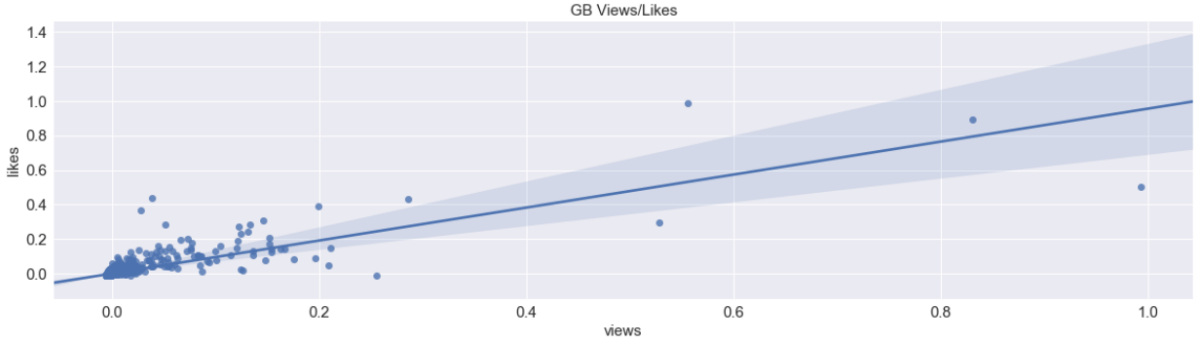
To investigate a bit more if there is significant difference between UK and US like rate means a t-test is necessary to be performed.

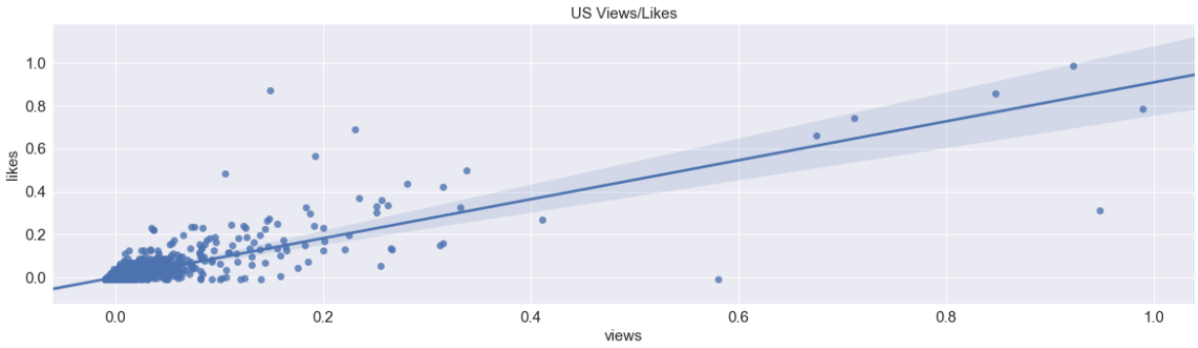
Mean likes rate UK:3.4

Mean likes rate GB:3.2

t-test results: p-value 0.001241 < 0.025 (two-tailed hypothesis)

We can reject the null hypothesis that the UK likes rate and GB likes rate means are statistically significant different

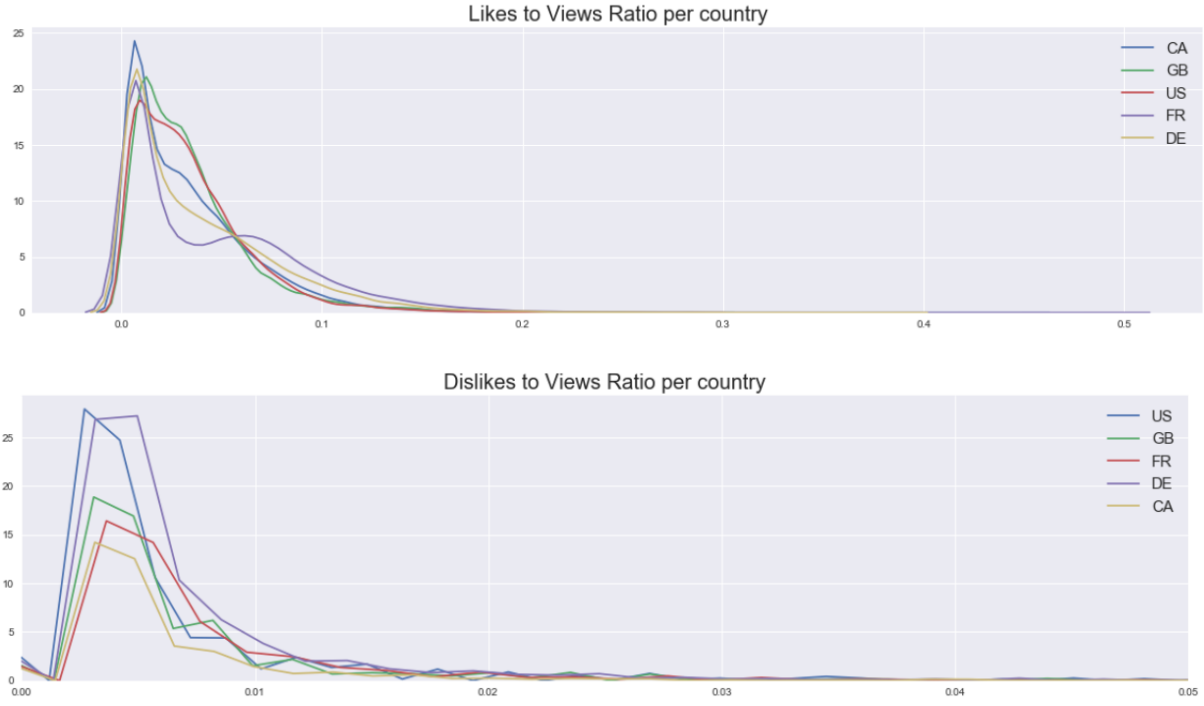


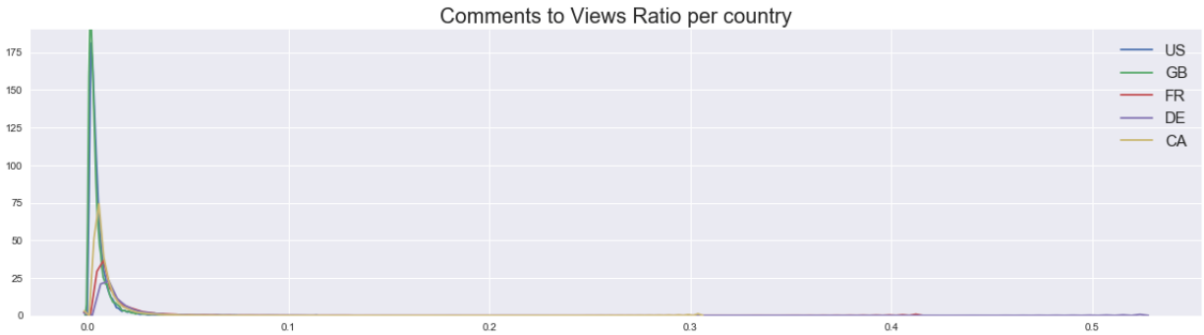


HOW DO COUNTRIES BEHAVE IN TERMS OF ENGAGEMENT RATES?

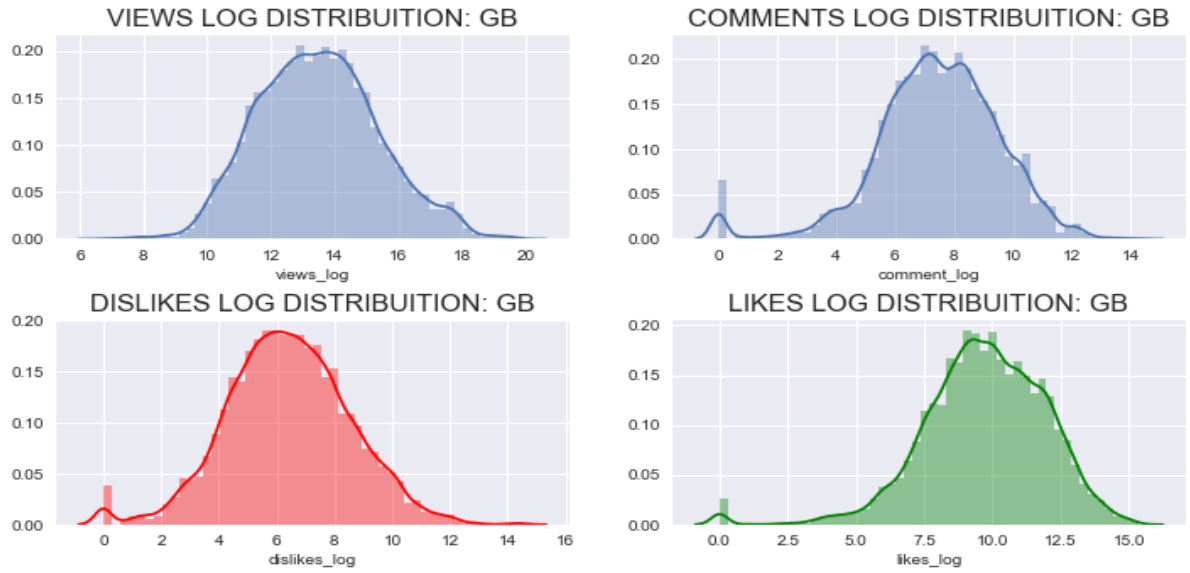
In all of the cases, ie like rates, comment rates and dislikes to likes, are highly skewed. That is, for high ratios and/or high rates of these metrics the cases we observe are very infrequent.

English speaking countries comment more! Great Britain is the country with the most likes to views followed by US. US and DE more closely resemble each other in terms of dislikes to likes ratio. Comment rates are more evident in US and in GB. The rest of countries do not engage in comments as much.





ARE THE ANGGANGEMENT METRICS NORMALY DISTRIBUTED?



To answer the question a normality test needs to be performed so we must test the following hypothesis:

H0: The data follow the normal distribution

H1: The data do not follow the normal distribution

Test results:

chi-square statistic: 66593.3
p value = 0.0

Since p value is < 0 the null hypothesis can be rejected, therefore the data do not follow the normal distribution.

YOUTUBE TRENDING VIDEOS WHAT DO DATA REVEALS

VIDEO CATEGORIES ANALYSIS

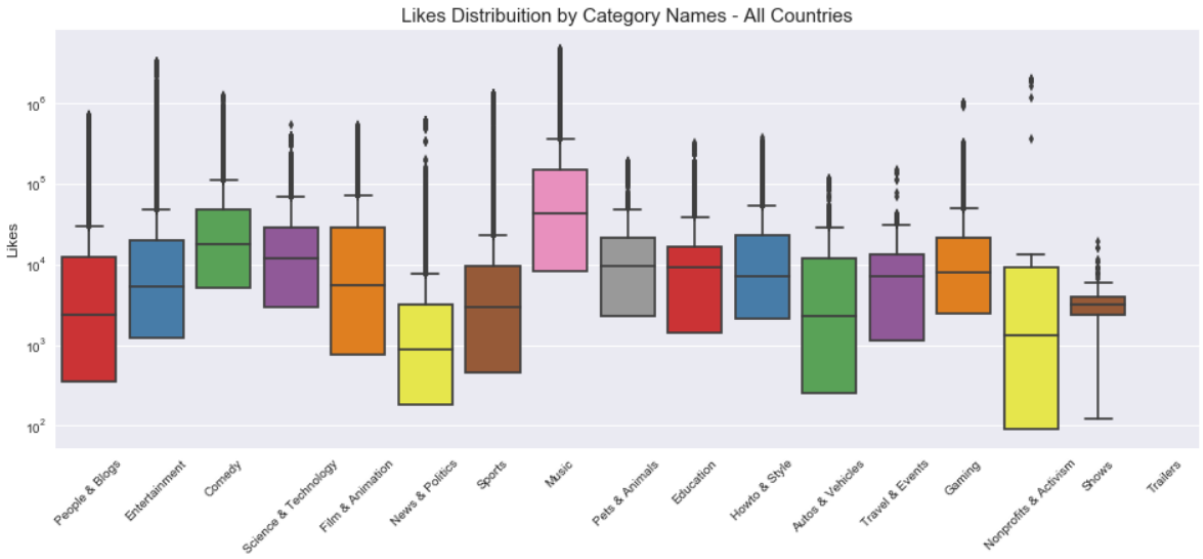
The videos of the dataset explored are clustered into 17 categories. As we will see below there is an enormous difference in terms of number of videos and engagement metrics per category and category/country.

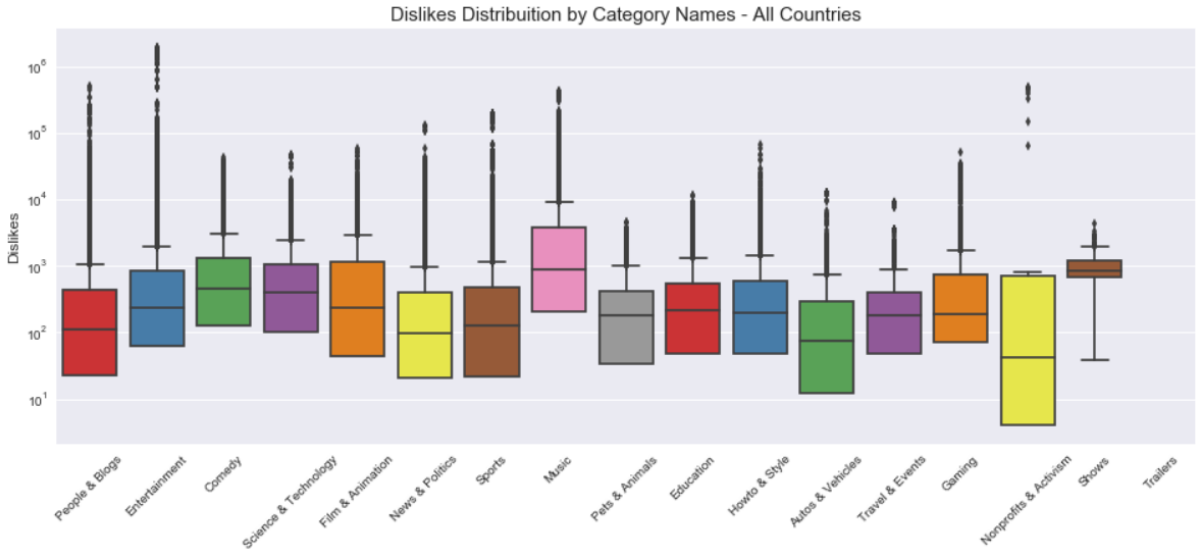
For the US the category with the highest likes/views rate is Nonprofits & Activism followed by Music and Comedy. Nonprofits & Activism category is right skewed in terms of like ratio. Shows, Gaming and News & Politics are the categories with the higher comment to views ratio. The Shows category is highly left skewed in comment rate, while News & Politics is right skewed

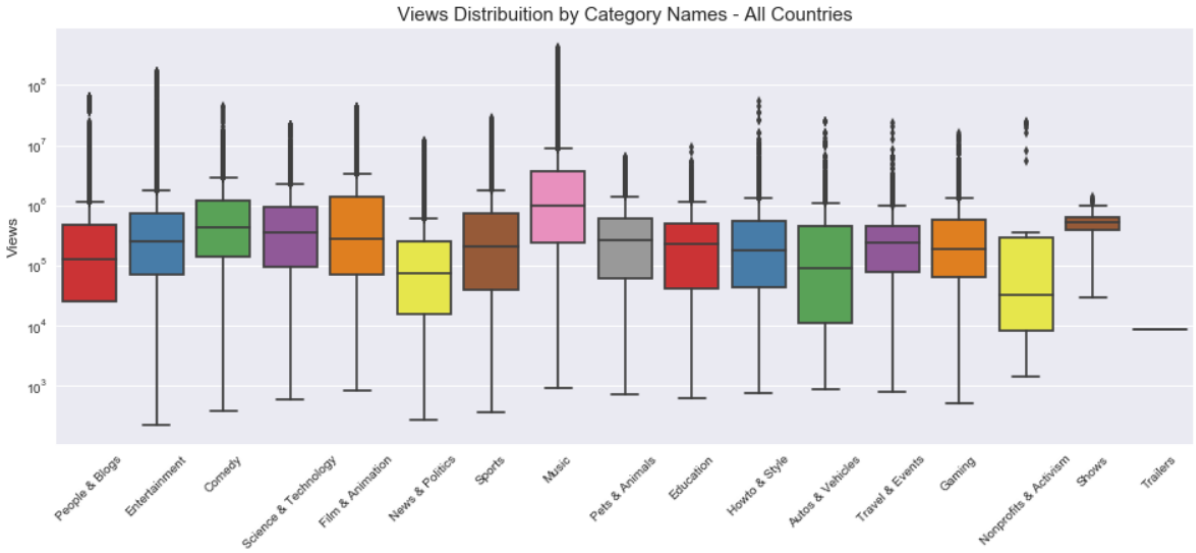
For GB comedy, Travel & Events, Travel & Events and Howto and Style are the categories with the highest likes/views rate. Travel and Events and News and Politics categories are right skewed in term of like ratio. Gaming, News & Politics, Travel and Events and Education are the categories with the higher comment to views ratio. News & Politics, Travel and Events and Education are highly right skewed in comment rate

Pets and animals is the category with the highest likes rate for France. Pets and animals and Sports are highly right skewed for Pets & Animals and Science & Technology are highly right skewed for Canada. Trailers category likes rate for Germany is zero

When it comes to all countries combined although music is the category with the most likes, music is the category with the dislikes as well. People have quite different opinions in terms of liking Nonprofits and Activism videos. Variation is quite high for People and Blogs, Autos and Film videos. Pets and Animals and Education videos are left skewed in term of likes.

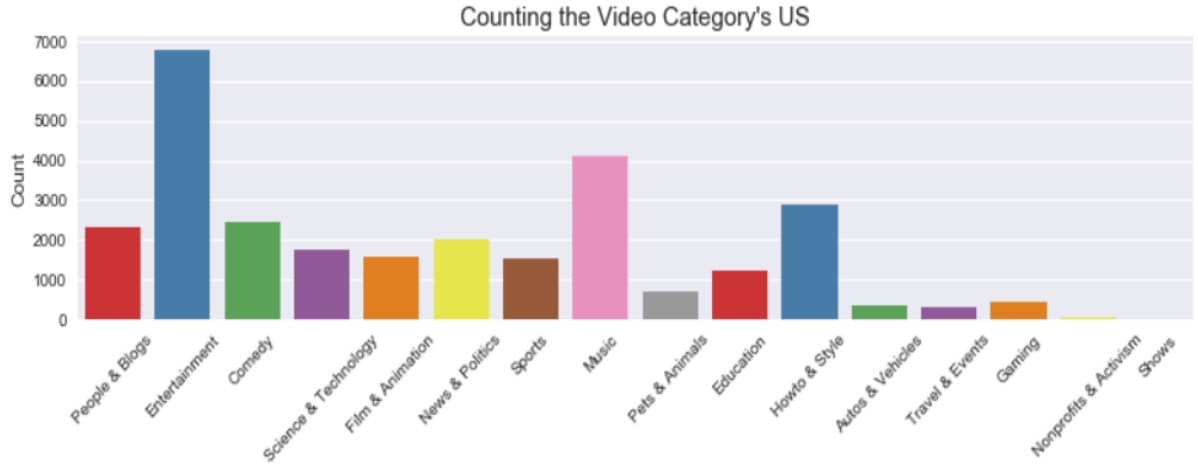


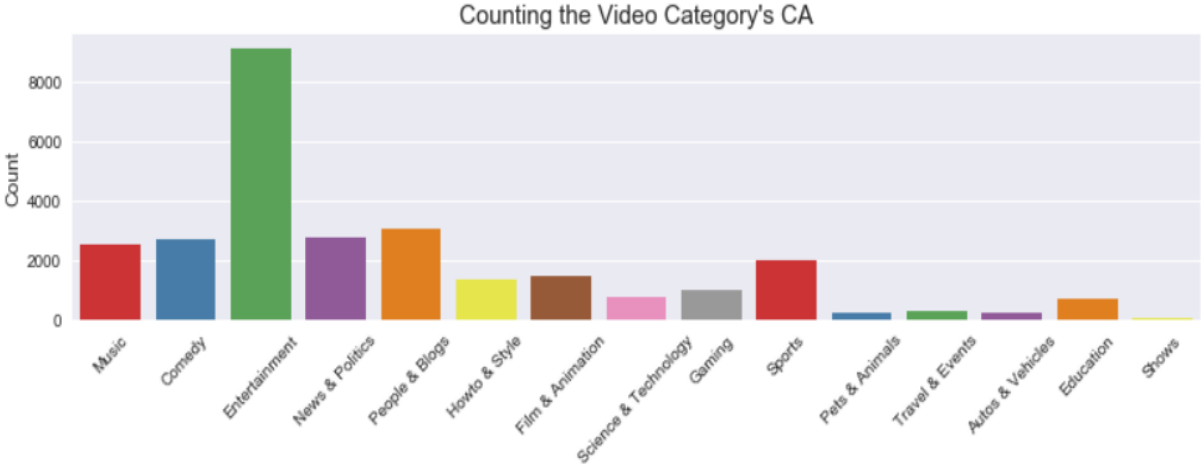


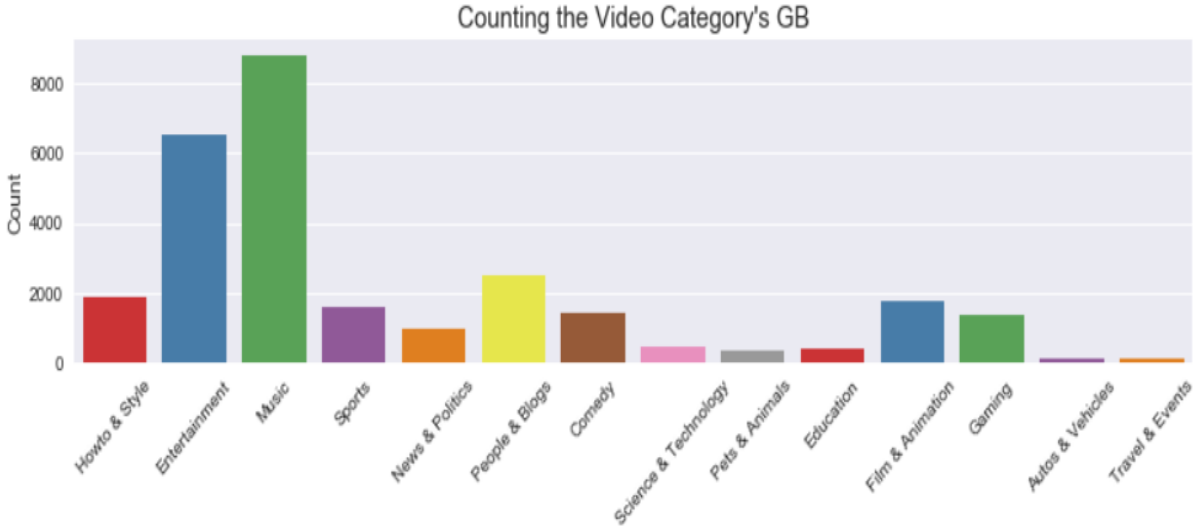


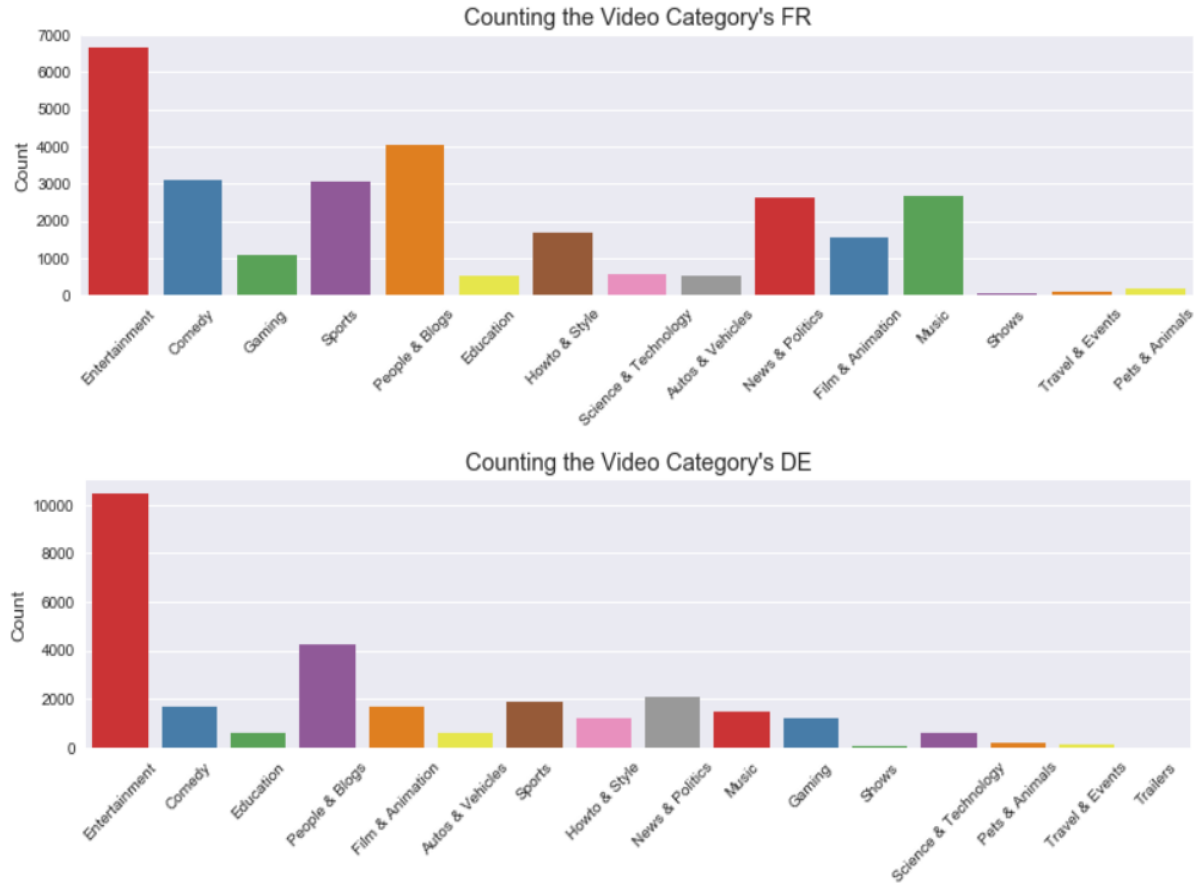
For all countries there is one prominent disliked video and for the rest of videos the number of dislikes is close to 0 million.

Top category of all countries except GB is Entertainment. For GB the top category is music. Music's videos ranked insignificantly in Canada, Germany and France compare to US and UK. Sport's videos are more popular in Canada, Germany and France. All top 8 categories in United Kingdom are entertainment-related. Show's and Activism's video get the bottom rank in all most countries

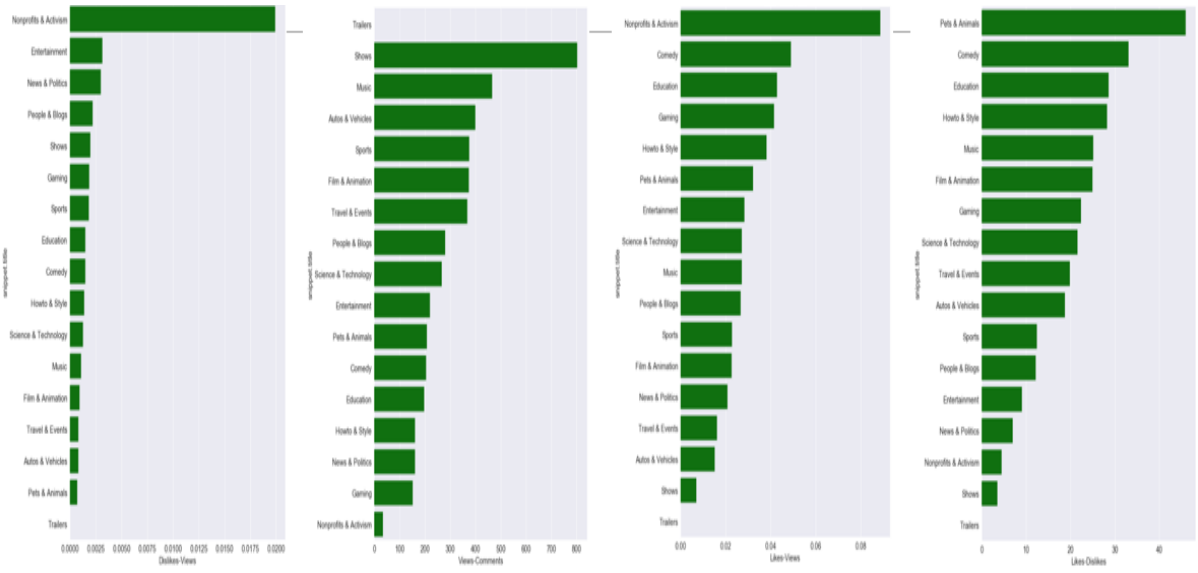

Counting the Video Category's US
Count
7000
6000
5000
4000
3000
2000
1000
0
People & Blogs
Entertainment
Comedy
Science & Technology
Film & Animation
News & Politics
Sports
Music
Pets & Animals
Education
Howto & Style
Autos & Vehicles
Travel & Events
Gaming
Nonprofits & Activism
Shows


Counting the Video Category's CA
Count
8000
6000
4000
2000
0
Music
Comedy
Entertainment
News & Politics
People & Blogs
Howto & Style
Film & Animation
Science & Technology
Gaming
Sports
Pets & Animals
Travel & Events
Autos & Vehicles
Education
Shows


Counting the Video Category's GB
Count
8000
6000
4000
2000
0
Howto & Style
Entertainment
Music
Sports
News & Politics
People & Blogs
Comedy
Science & Technology
Pets & Animals
Education
Film & Animation
Gaming
Autos & Vehicles
Travel & Events



Pets & Animals videos have highest likes-dislikes ratio. Not surprisingly, people find difficult to hate pets and animals. Non-profit & Activism's videos have lowest likes-dislike ratio and views-comments ratio. People relatively hate these videos and comment too much. Also, people still prefer implicit feedback than explicit. The ratio of views to comments is so large that only a comment written for hundreds of views.



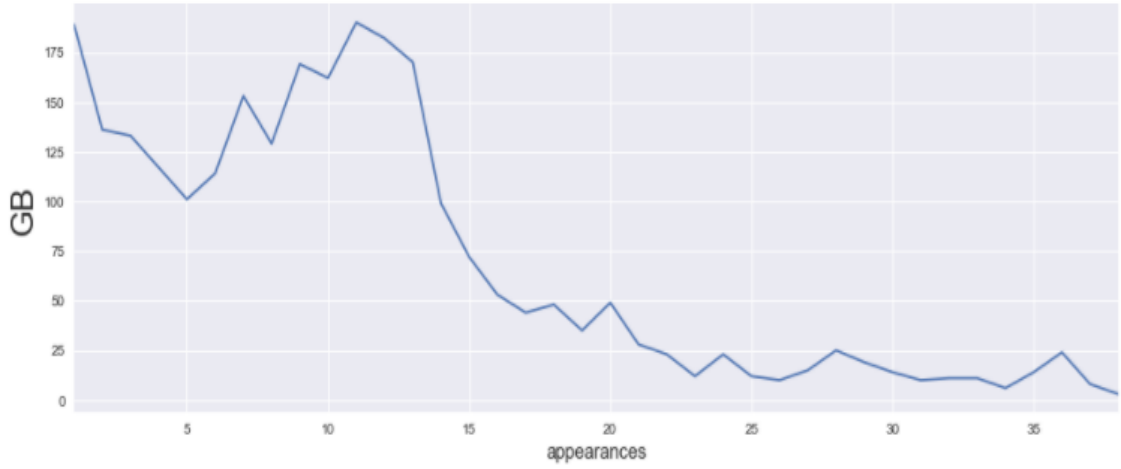
LIFETIME OF TRENDING VIDEOS

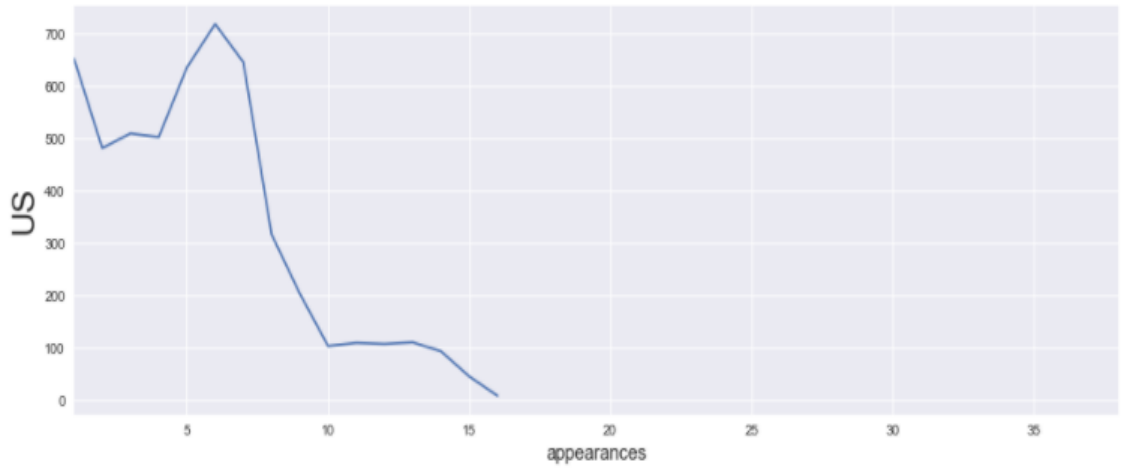
United Kingdom has the most long trended YouTube videos (in the chart below the greater the number of appearances indicate the long-last the video trend is).

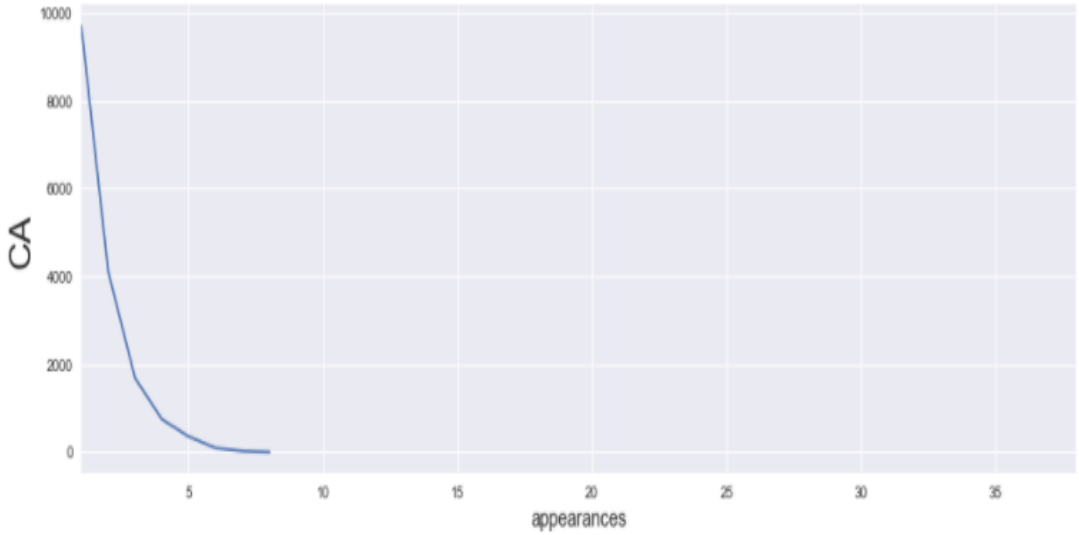
For GB most trending videos are watched for 11 days. After the 11 day there is a huge trop in the number of trending videos that keep stimulating the audience. There are some significant portions of videos though that are being watched for more that 40 days. What is interesting is that there are more videos that appear for 11 days compared to videos that appear for 5 days.

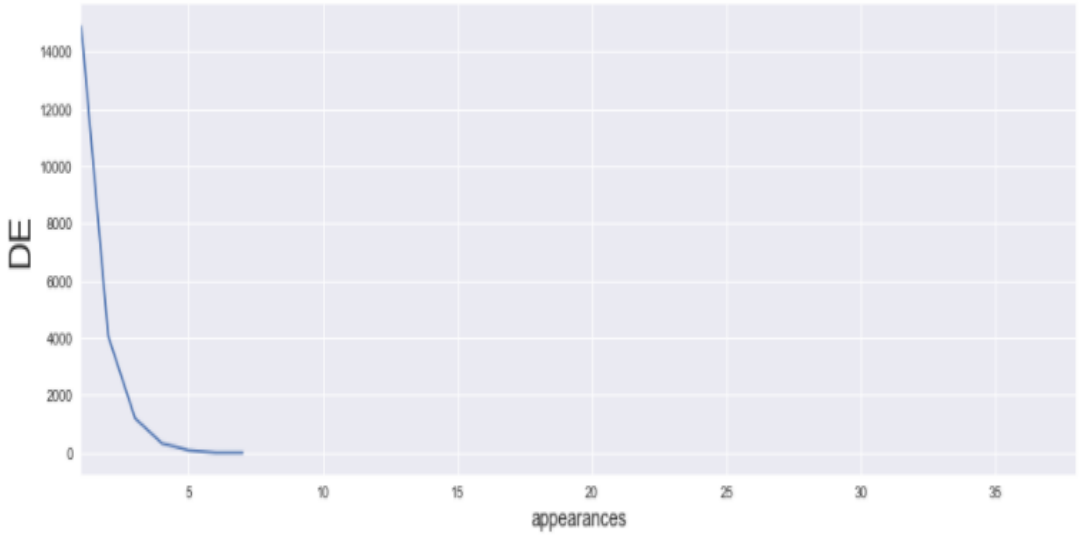
US has a different behavior, most trending videos are watched for 6 days while the maximum days observed is 16.

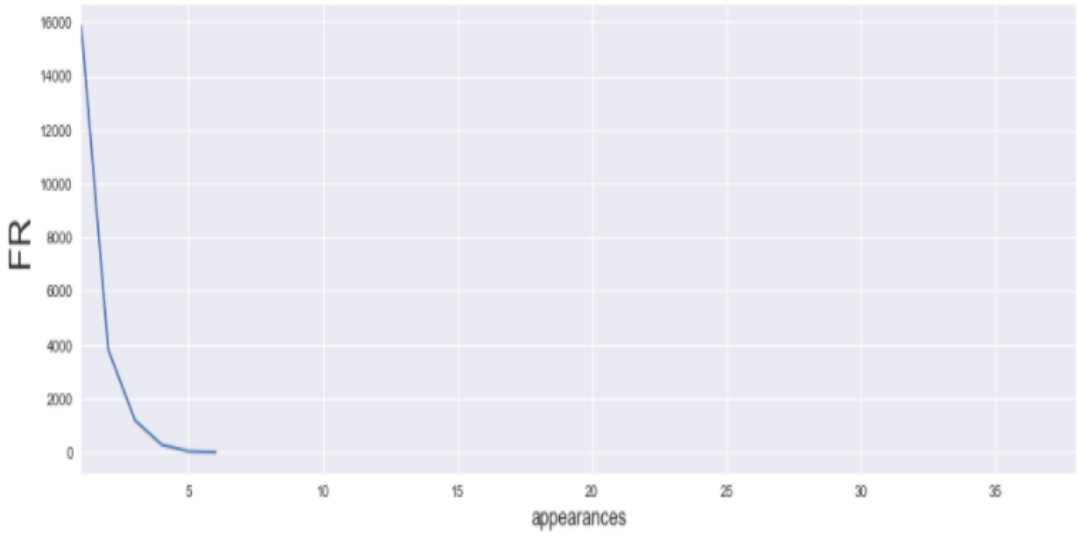
For Canada, France and Germany most videos are watched once, and the maximum days reached are close to 5.






CA
10000
8000
6000
4000
2000
0
5
10
15
20
25
30
35
appearances

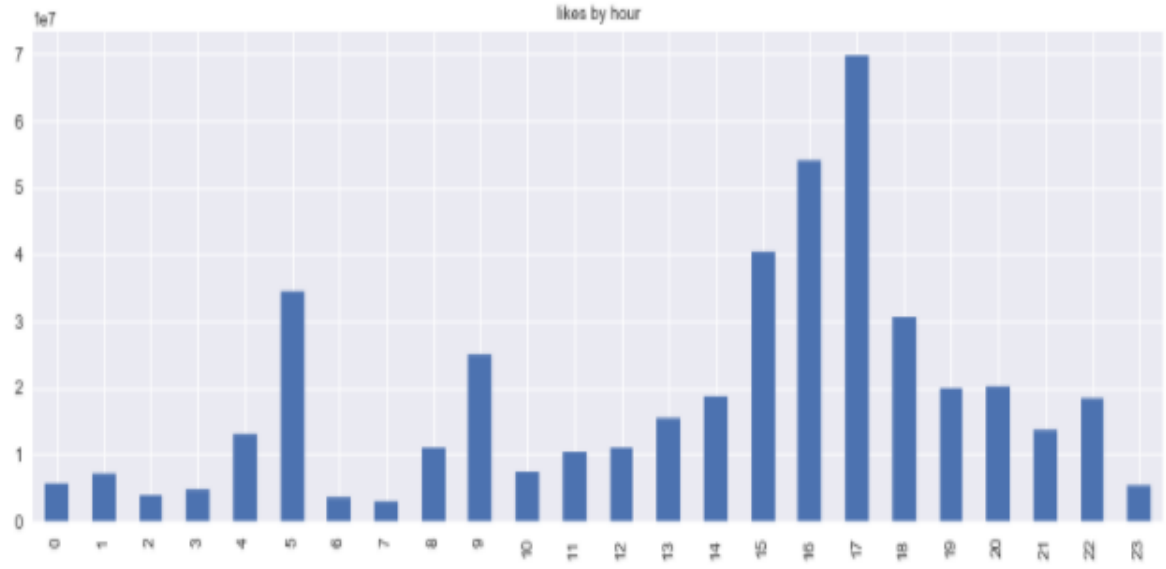

DE
14000
12000
10000
8000
6000
4000
2000
0
5
10
15
20
25
30
35
appearances


FR
16000
14000
12000
10000
8000
6000
4000
2000
0
5
10
15
20
25
30
35
appearances

OTHER INTERESTING OBSERVATIONS

After analyzing the dataset, Street food, der , food tour and Mexico are the most popular worlds in video tittles

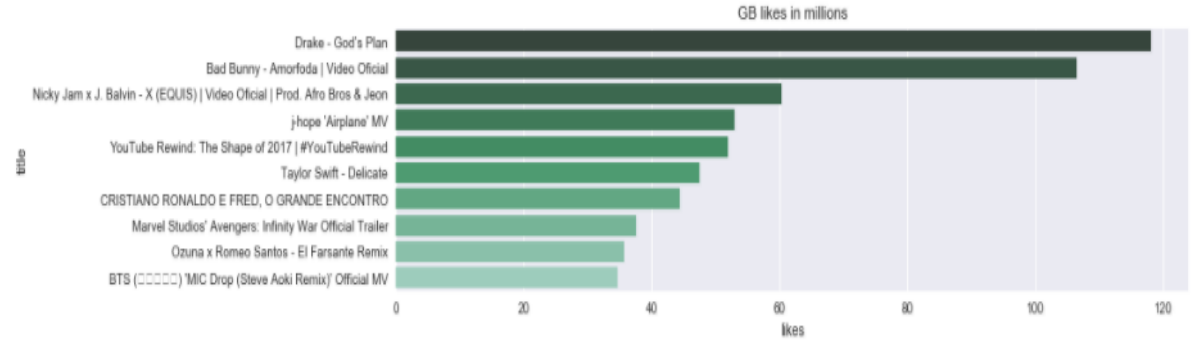
Likes are significantly higher for videos published in evening hours

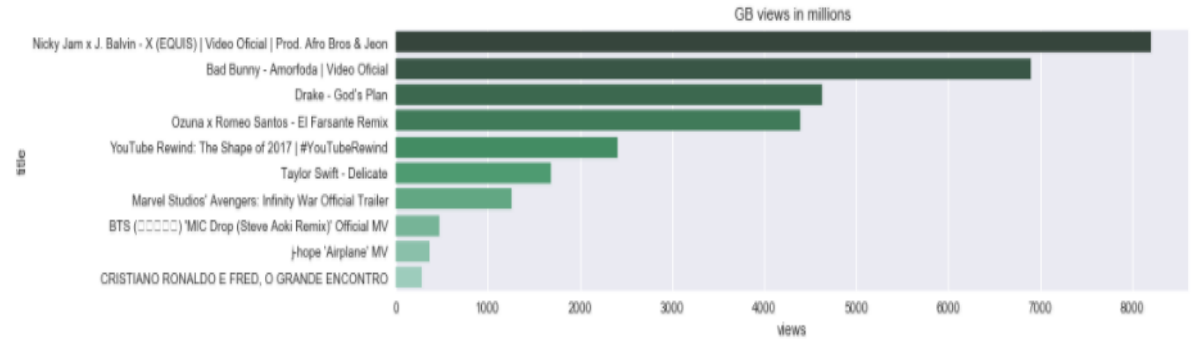


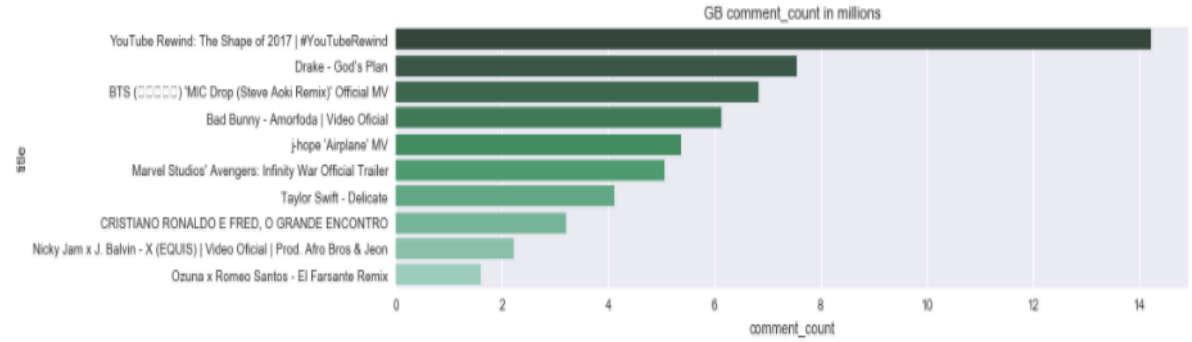
For all countries there is one prominent disliked video and for the rest 9 videos the number of dislikes is close to 0 million.

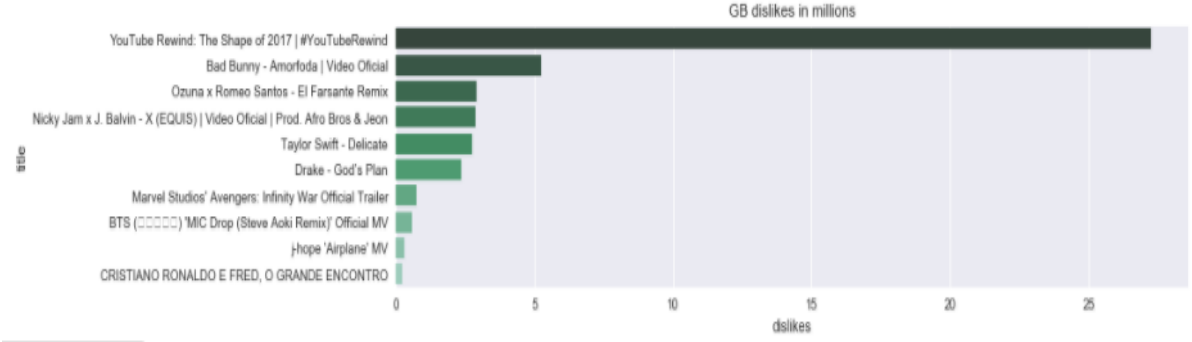
GB is the country with most likes, while the most liked video has reached 120 million likes, the second country in likes is the US and the most liked video there has only 25 million likes.

For all countries, apart from the US the difference in likes among the first 10 videos is not that big. Actually, the tenth most liked video for the US has just half of the likes compared to the most liked one.









YOUTUBE TRENDING VIDEOS K MEANS CLUSTERING

WHAT ARE THE CHARACTERISTICS OF YOUTUBE TRENDING VIDEOS?

From the analysis below we have concluded the following in terms of trending videos:

Music for UK and Entertainment for the other countries are the categories with the most likes and views.

For the US the category with the highest likes/views rate is Nonprofits & Activism followed by Music and Comedy. For GB those categories are comedy, Travel & Events and Howto and Style.

The videos from United Kingdom, US and Canada are highly correlated to each other in comparison with Germany and France.

All engagement metrics are correlated to each other.

Therefore we have some indications of what makes a video trending. To be more precise in terms of which factors can make a video successful a we will cluster the videos bases on a series of values and try to find what the common characteristics for the most viewed videos are.

NEW METRICS

As for the first step of the analysis the dataset a list of new metrics has been created

Trending date: count trending days

title: wrd count

title: If capital letters are more than small

title: has number

title: has punctuations or not

```
videos['title_len'] = videos.title.apply(lambda x: len(x.split(" ")))
videos['title_upper_case_prc'] = videos.title.apply(lambda x: 100 * len(re.findall(r'[A-Z]',x)) / len(x))
videos['title_count_letters'] = videos.title.apply(lambda x: len(x))
videos['title_has_numer'] = videos.title.apply(lambda x: (len(re.findall(r'[0-9]',x)) > 0) * 1)
videos['title__hash_exclamation'] = videos.title.apply(lambda x: len(re.findall(r'[#!]',x)) )
```

channel_title: count of posted videos

channel_title: average number of trending days per posted video

publish_time: how old is the video in days

```
videos['days_old'] = videos.publish_date.apply(lambda x: basedate.date() - x)
```

tags: Number of tags "|"

```
videos['number_of_tags'] = videos.tags.apply(lambda x: len(x.split("|")))
```

likes: Number of likes

dislikes: Number of dislikes

comment_count: Number of comments

comments_disabled: binary 0 or 1

ratings_disabled: binary 0 or 1

description: word count

description: http count

description: has #

```
videos['description_word_count'] = videos.description.apply(lambda x: len(x.split(" ")))
videos['description_http_count'] = videos.description.apply(lambda x: len(re.findall(r'http',x)) )
videos['description_hashtag_count'] = videos.description.apply(lambda x: len(re.findall(r'[#]',x)) )
```

snippet.title: category name

```
videos = pd.concat([videos, pd.get_dummies(videos['snippet.title'])], axis=1)
```

views: number of views

prc_likes

prc_dislikes

prc_comments

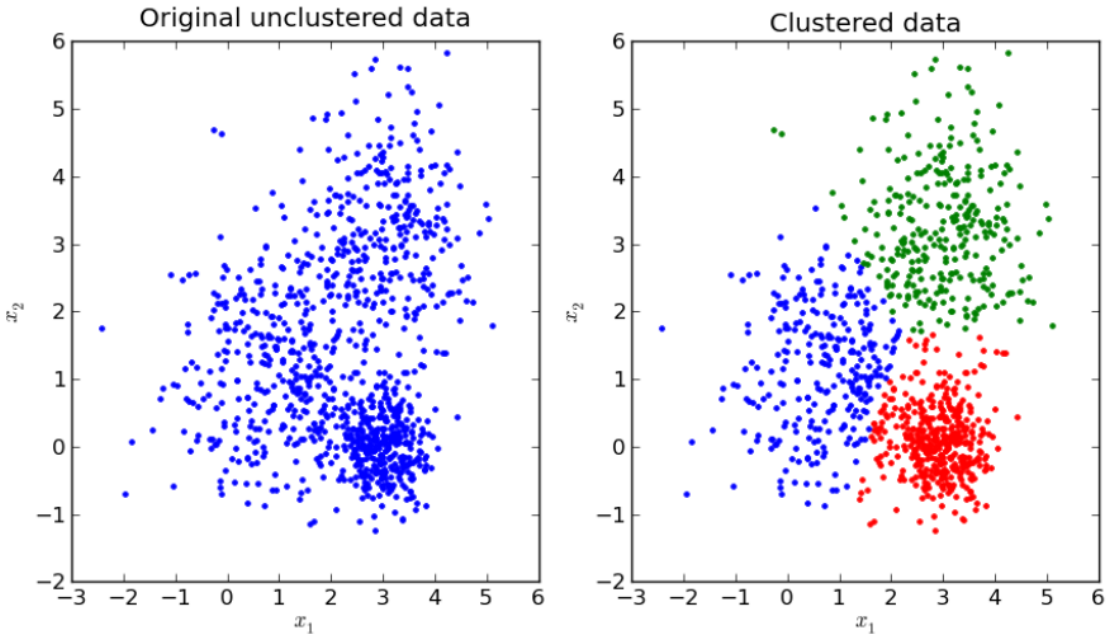
The new dataset has therefore the following columns:

Country', 'Row ID', 'video_id', 'trending_date', 'title', 'channel_title', 'category_id', 'publish_time', 'tags', 'likes', dislikes', 'comment_count', 'thumbnail_link', 'comments_disabled', 'ratings_disabled', 'video_error_or_removed', 'description', 'publish_date', 'publish_month', 'publish_day', 'publish_hour', 'id', 'snippet.title', 'country', 'views', 'prc_likes', 'prc_dislikes', 'prc_comments', 'likes_log', 'views_log', 'dislikes_log', 'comment_log', 'like_rate', 'dislike_rate', 'comment_rate', 'title_len', 'title_upper_case_prc', 'title_count_letters', 'title_has_numer', title__hash_exclamation', 'num_of_videos', 'Average_trending_days', 'days_old', 'number_of_tags', 'Autos & Vehicles', 'Comedy',

'Education', 'Entertainment', 'Film & Animation', 'Gaming', 'Howto & Style', 'Music', 'News & Politics', 'Nonprofits & Activism', 'People & Blogs', 'Pets & Animals', 'Science & Technology', 'Shows', 'Sports', 'Trailers', 'Travel & Events'

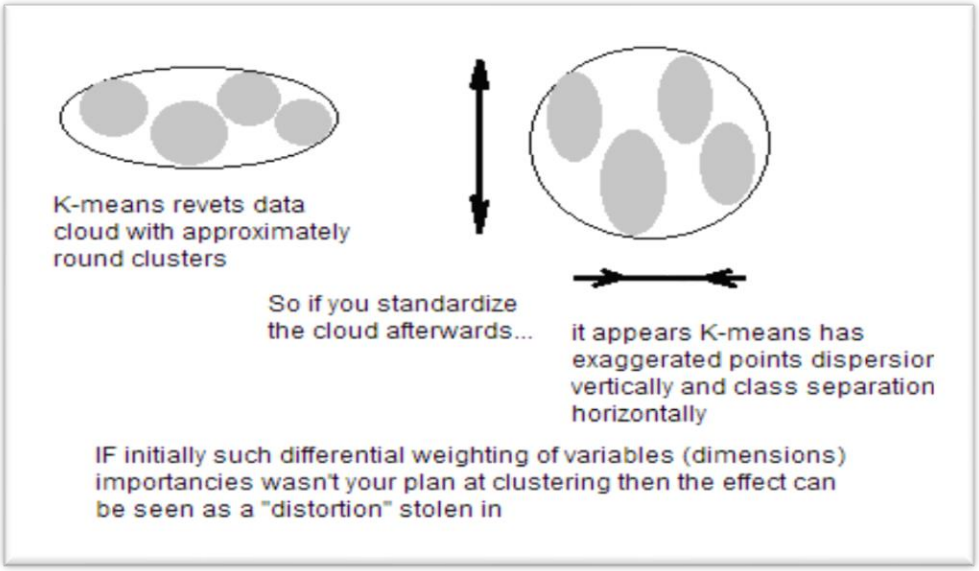
K-MEANS CLUSTERING

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid.



Determining the number of clusters in the data set – The elbow method-looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modelling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion".

Before moving to selecting the number of clusters normalization of data has to come first. If your variables are of incomparable units (e.g. height in cm and weight in kg) then you should standardize variables. Normalization is the process of scaling individual samples to have unit norm. Even if variables are of the same units but show quite different variances it is still a good idea to standardize before K-means. K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round clusters. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance, so clusters will tend to be separated along variables with greater variance.

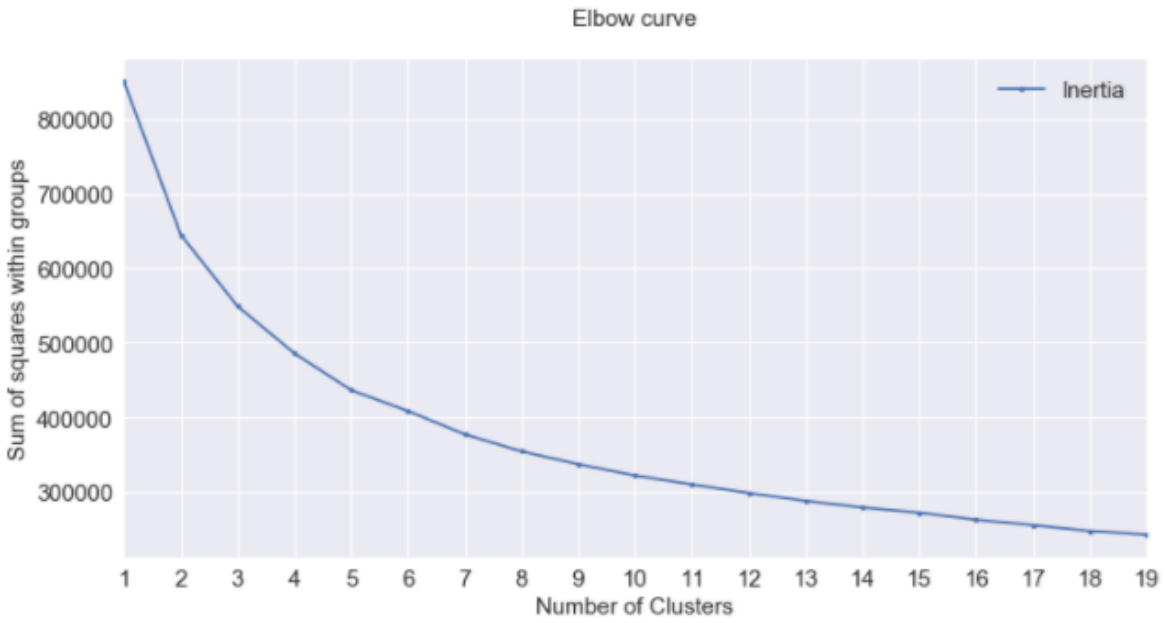


STANDARD SCALER

Standardization" typically means that the range of values are "standardized" to measure how many standard deviations the value is from its mean. Standardizing a dataset involves rescaling the distribution of values so that the mean of observed values is 0. The StandardScaler assumes your data is normally distributed within each feature and will scale them such that the distribution has a standard deviation of 1.

The mean and standard deviation are calculated for the feature and then the feature is scaled based on: xi–mean(x)/stdev(x).

From the elbow chart of the below chart we can infer that the optimal number of clusters id 7.




```
kmeans = KMeans(n_clusters=7)
kmeans.fit(df.values)

final_features_all['cluster'] = kmeans.labels_

df_clu_group_describe = final_features_all.groupby('cluster').describe(percentiles=[.1, .25, .5]).T
df_clu_group_describe.index.names = ['features', 'metric']
result = df_clu_group_describe.T
result
```

In order to find out what makes the videos of some clusters more engaging than others a profiling proses is followed by dividing the mean per cluster with the mean per metric.

```
final = round(100 * df_final.groupby('cluster_final').mean() / df_final.mean(), 0)
```

You can see part of the final dataset at the table below. The data are shorted based on number of views with the 3rd clusters having the biggest number of views. What are the characteristics of this cluster that makes the videos that belong to it more successful than the others?

cluster_final	Autos & Vehicles	Average_trending_days	Comedy	Education	Entertainment	Film & Animation	Gaming	Howto & Style	Music	News & Politics	Nonprofits & Activism	People & Blogs	Pets & Animals
2	121.0	11.0	74.0	94.0	100.0	102.0	70.0	84.0	70.0	125.0	30.0	116.0	67.0
4	92.0	20.0	129.0	160.0	81.0	70.0	145.0	172.0	126.0	102.0	169.0	95.0	168.0
5	120.0	23.0	122.0	63.0	101.0	63.0	269.0	112.0	34.0	68.0	102.0	86.0	157.0
6	103.0	30.0	71.0	37.0	123.0	188.0	58.0	50.0	71.0	55.0	98.0	108.0	16.0
1	52.0	116.0	159.0	134.0	101.0	83.0	40.0	70.0	174.0	122.0	510.0	67.0	117.0
7	55.0	84.0	72.0	72.0	137.0	134.0	56.0	76.0	137.0	28.0	510.0	76.0	7.0
3	39.0	646.0	168.0	117.0	94.0	92.0	132.0	129.0	220.0	52.0	62.0	60.0	196.0

The cluster with the highest number of views over indexes with the average trending days as expected. Also music and pets and animals are the categories that show that could more possibly have the higher number of views. The metrics that are also over indexing are: lond titles, titles with numbers, titles with high percentage of capital letters.

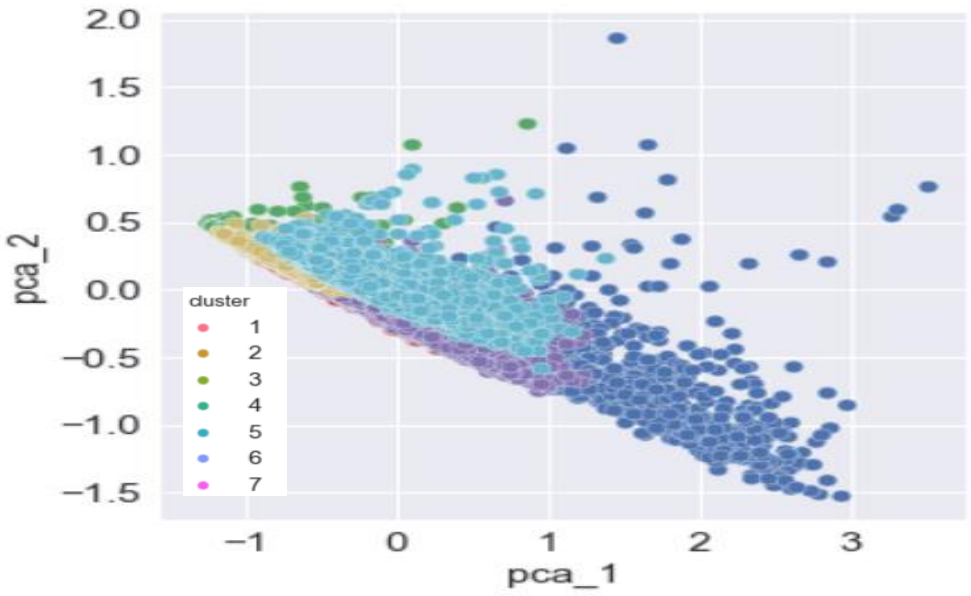
The cluster with the higher like rate over indexes with education, gaming and how to & style categories.

The cluster with the least number of views over indexes on the autos and vehicles, spots, news and politics and people and blogs.

The chart below is a representation of the clusters after applying Principal component analysis (PCA). Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. That way we can reduce the 20 dimensional data into 2 dimensions so that we can plot and hopefully understand the data better.

The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. In practice, the covariance (and sometimes the correlation) matrix of the data is constructed and the eigenvectors on this matrix are computed. The eigenvectors that correspond to the largest eigenvalues (the principal components) can now be used to reconstruct a large fraction of the variance of the original data. The original space (with dimension of the number of points) has been reduced (with data loss, but hopefully retaining the most important variance) to the space spanned by a few eigenvectors.

```
pca = decomposition.PCA(n_components=2)
pca.fit(df.drop('cluster', axis=1).values)
X = pca.transform(df.drop('cluster', axis=1).values)
```

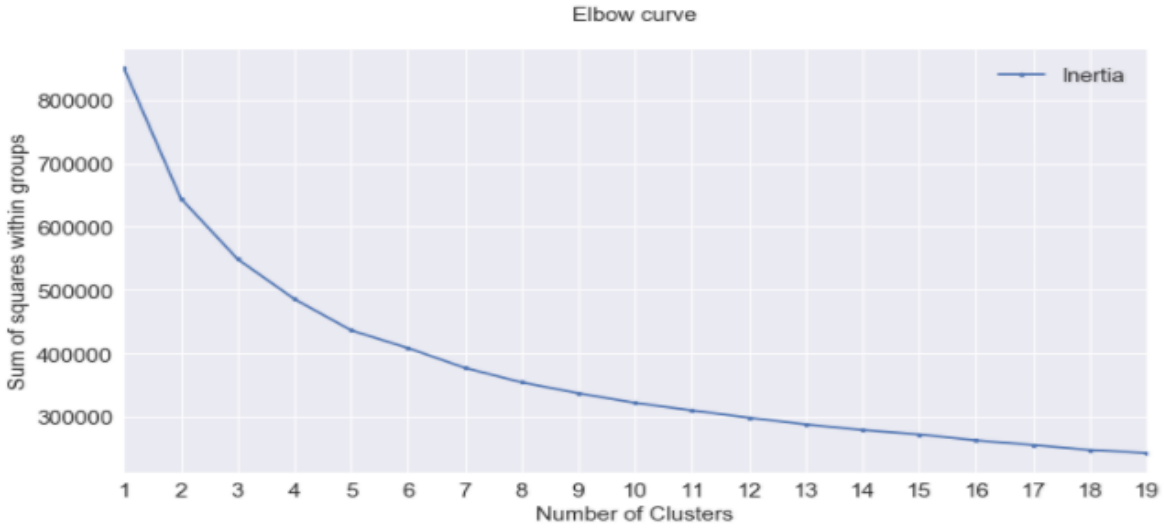


MIN-MAX SCALER

The MinMax Scaler is the probably the most famous scaling algorithm, and follows the following formula for each feature: xi–min(x)/max(x)–min(x).

```
scaler = preprocessing.MinMaxScaler()
scaled_df = scaler.fit_transform(final_features_all.fillna(0))
scaled_df_minmax = pd.DataFrame(scaled_df, columns= final_features_all.columns)
```

The elbow curve in this case of normalization is the same as for the standard scaler. So the number of optimal clusters is 7 in this case as well.



The values are spited in the 7 clusters as following:

```
round(final_features_all['cluster_minmax'].value_counts(normalize=True) * 100, 1)
```

1	29.5
0	24.3
2	14.3
3	10.5
5	8.9
6	6.4
4	6.1

The first column represents the cluster and the second how values are split in % among the 7 clusters

The cluster with the highest number of views over indexes with the average trending days. The metrics that are also over indexing are: lond titles, titles with numbers, titles with high percentage of capital letters.

The cluster with the higher like rate over indexes with education, gaming and how to & style categories.

The cluster with the least number of views over indexes on the autos and vehicles, spots, news and politics and people and blogs.

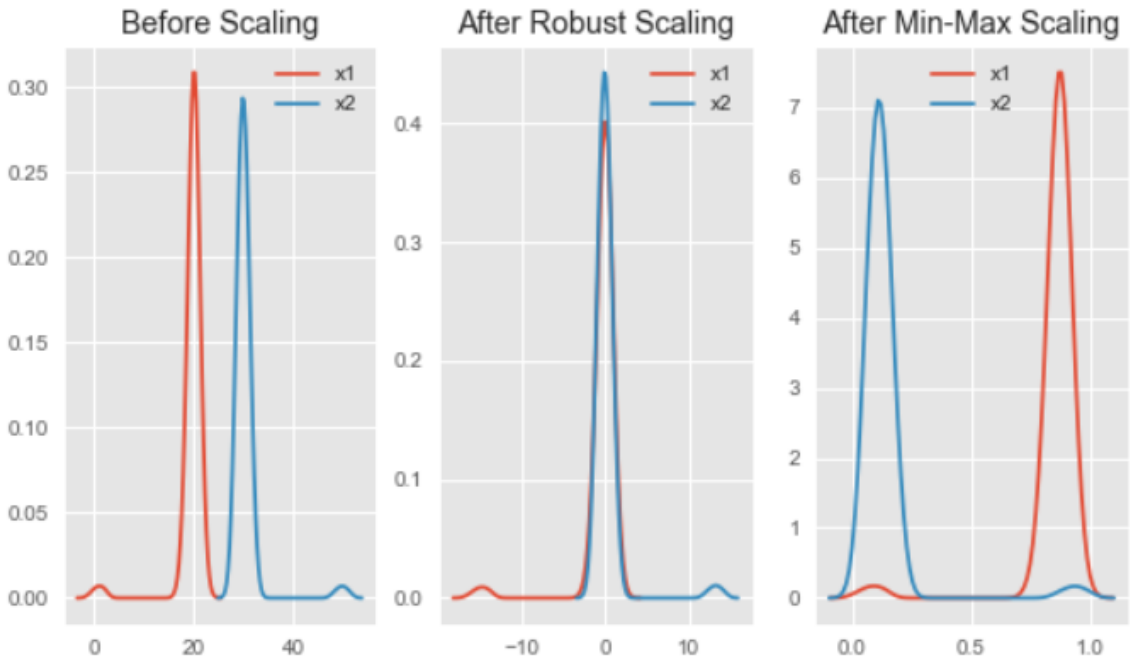
ROBUST SCALER

The RobustScaler uses a similar method to the Min-Max scaler but it instead uses the interquartile range, rather than the min-max, so that it is robust to outliers. Therefore it follows the formula: xi–Q1(x)/(Q3(x)–Q1(x))

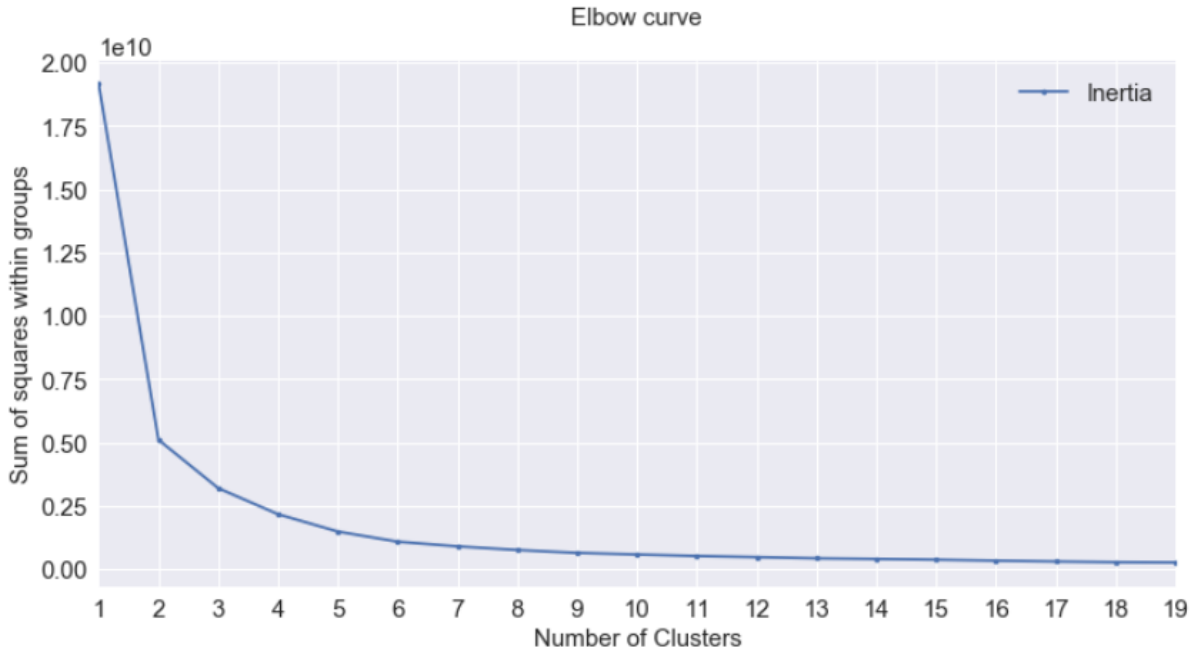
```
scaler = preprocessing.RobustScaler()
scaled_df = scaler.fit_transform(final_features_all_1.fillna(0))
scaled_df_robust = pd.DataFrame(scaled_df, columns= final_features_all_1.columns)
```

Of course this means it is using the less of the data for scaling so it's more suitable for when there are outliers in the data.

After Robust scaling, the distributions are brought into the same scale and overlap, but the outliers remain outside of bulk of the new distributions. However, in Min-Max scaling, the two normal distributions are kept separate by the outliers that are inside the 0-1 range. The charts below are an example of how the different normalization methods look like visually.



The elbow curve in this case of normalization is the same as for the standard scaler. So the number of optimal clusters is just 2 in this case.

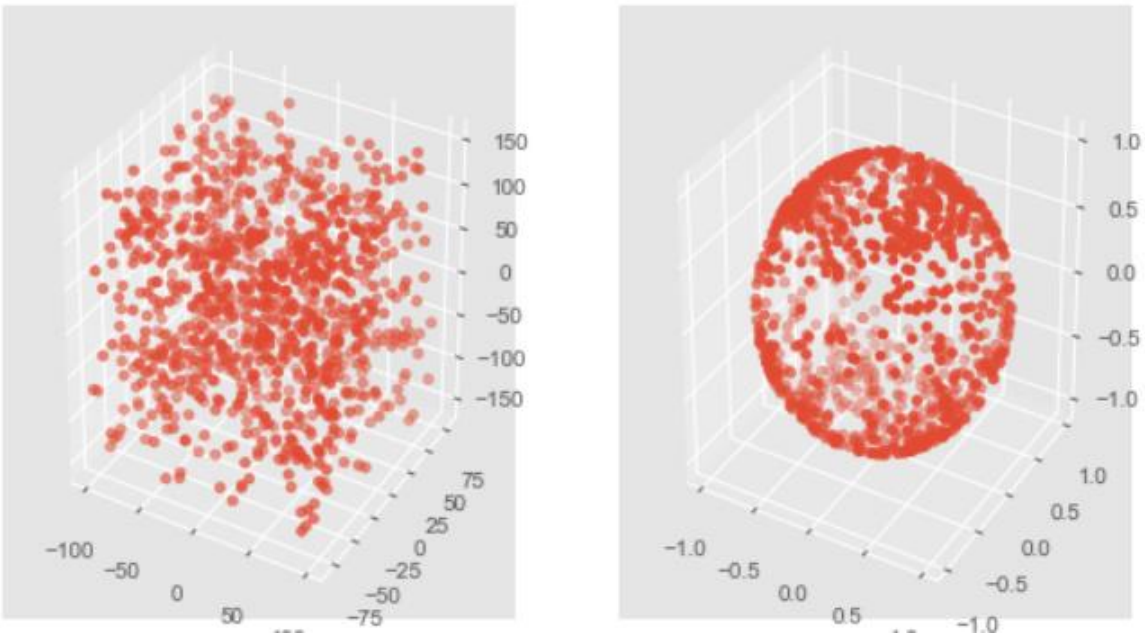


This could indicate that this method of normalization might not be the best one to be used for clustering.

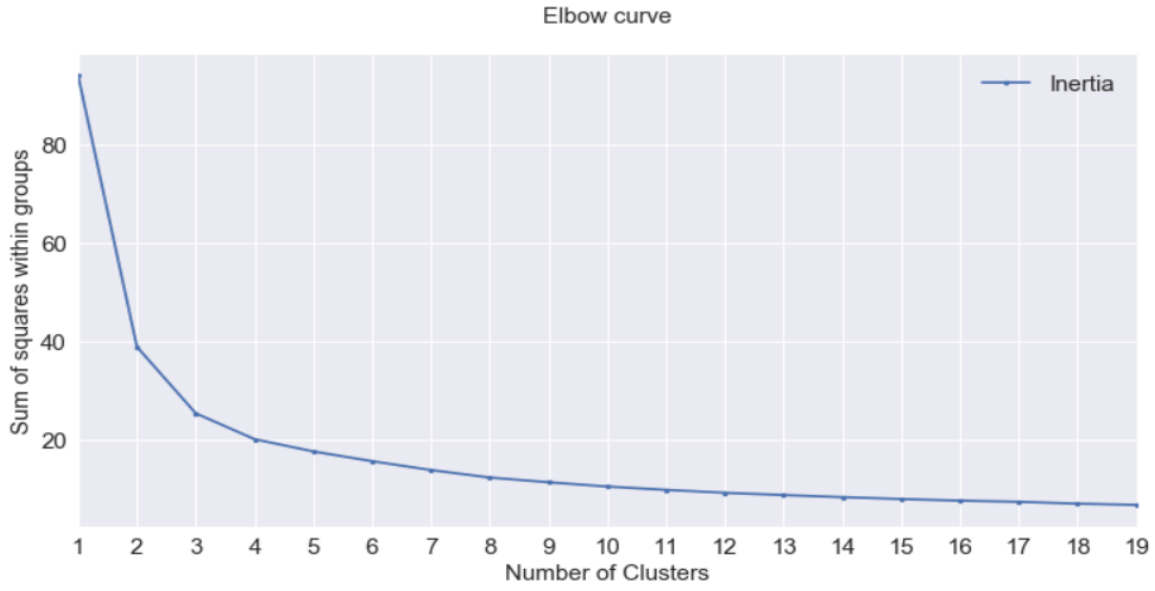
NORMALIZER

The normalizer scales each value by dividing each value by its magnitude in n-dimensional space for n number of features. Say your features were x, y and z Cartesian co-ordinates your scaled value for x would be: xi/√xi2+yi2+zi2

Each point is now within 1 unit of the origin on this Cartesian co-ordinate system. The points are all brought within a sphere that is at most 1 away from the origin at any point. Also, the axes that were previously different scales are now all one scale.



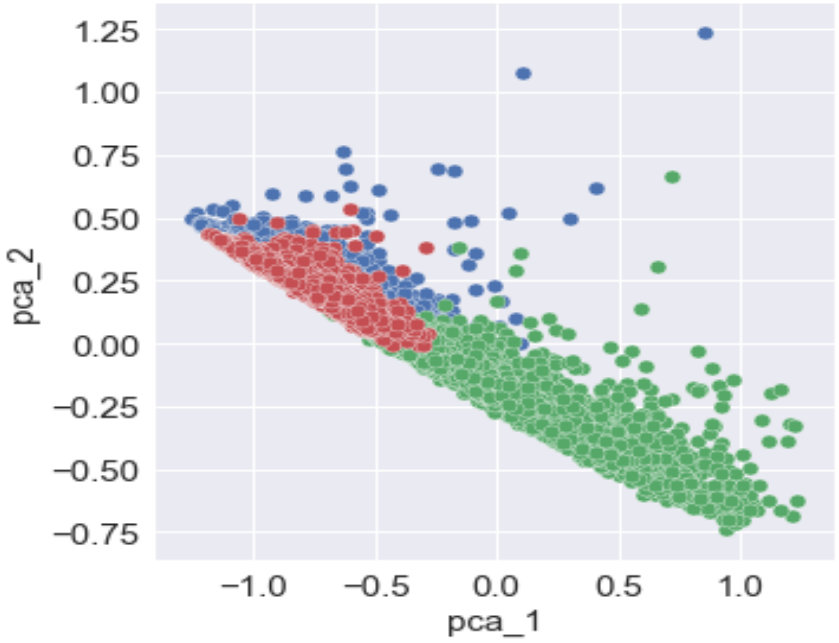
In this case of normalization the optimal number of clusters is 3 as we can see at the elbow curve at the chart below.



From the 3 clusters formed one of them gathers almost 60% of values. The second and third cluster in number of values constitute the 30% and 10% respectively.

The cluster with the most likes over indexes with the following categories: games, autos and vehicles, how to & style and education. The cluster with the higher like rate over indexes with the presence of exclamation in the title and the number of tags.

Finally the chart below is a representation of the clusters after applying Principal component analysis (PCA). Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. That way we can reduce the 20 dimensional data into 2 dimensions so that we can plot and hopefully understand the data better. As we can see the second and third cluster are not clearly separated. This could indicate that the clustering after applying this method of normalization is not the optimal.



RESULTS AND CONCLUSION

So what are key factors that can make a video to be popular?

UK is the country with the most popular videos.

Music for UK and Entertainment for the other countries are the categories with the most likes and views.

For the US the category with the highest likes/views rate is Nonprofits & Activism followed by Music and Comedy. For GB those categories are comedy, Travel & Events and How to and Style.

The videos from United Kingdom, US and Canada are highly correlated to each other in comparison with Germany and France also all engagement metrics are correlated to each other. That being said if a video has a lot of views most probably will have a high number of likes and comments.

Based on the k- means analysis on the standardized normalized data set pets and animals are the categories that show that could more possibly have the higher number of views.Long titles, titles with numbers, titles with high percentage of capital letters are a factors to success as well.