VERSION 1.1
NOVEMBER 25, 2018

YouTube Video Statistics and Analysis

- EXPLORING VIDEO ANALYSIS

PRESENTED BY: *ANANTHARAMAN CHANDAR*
GOWRISANKAR ARUMUGAM
ASHOK RAMASAMI

ITMD529 ADVANCED DATA ANALYTICS
ILLINOIS INSTITUTE OF TECHNOLOGY
- SCHOOL OF APPLIED TECHNOLOGY

CONTENTS

Document Change History 5

Who Can Refer 6

Introduction 6

Expected Outcomes 6

Dataset 6

Analytics Plan **7**

Research Questions **7**

Hypothesis Testing **7**

Data Definition **7**

Project Plan **8**

Project Requirements **8**

Domain Knowledge **8**

Technical Knowledge **8**

System and Software Requirements **8**

Project Constraints **9**

Project Assumptions **9**

General Assumptions **9**

Technical Assumptions **9**

Work Breakdown Structure **10**

Phase - I **10**

Phase – II **11**

Phase – III **12**

Team Roles and Responsibilities **13**

Phase – I and Phase - II **13**

Solution Development and Implementation Report **14**

Data Dictionary **14**

Column Description **14**

Raw Data Quality **16**

Statistics Distribution **17**

Columns to be Considered **17**

Reason for Elimination **18**

Quartile Distribution Based on Country **18**

Canada **18**

France **18**

Denmark **19**

Great Britain **19**

United States **19**

Exploratory Data Analysis of YouTube Data based on Countries [USA] **20**

USA Count of Movies Based on Category **20**

Total Views/Likes/Dislikes/Comments based on Category Id **20**

Scatter Plot Analysis for Sum of Views **21**

Scatter Plot Analysis for Sum of Likes **21**

Scatter Plot Analysis for Sum of Dislikes **22**

Scatter Plot Analysis for Sum of Comments **22**

Average Views/Likes/Dislikes/Comments based on Category Id **23**

Scatter Plot for Sum of Channels based on Category ID **24**

Bar chart for Count of Comments/ratings/videos **25**

Linear Trendline for Count of Ratings Disabled **26**

Linear Trendline for Count of Videos Removed **27**

Linear Trendline for Count of Comments Disabled **28**

Model Development and Implementation **29**

Cluster Map **29**

Scatter Plot **30**

Principal Component Analysis **31**

Exploratory Data Analysis of YouTube Data based on each Country **32**

USA Exploratory Data Analysis **32**

Canada Exploratory Data Analysis **35**

France Exploratory Data Analysis **37**

Great Britain Exploratory Data Analysis **40**

Germany Britain Exploratory Data Analysis **42**

Clustering and Segmentation Analysis **45**

USA Clustering and Segmentation Analysis **45**

Canada Clustering and Segmentation Analysis **48**

France Clustering and Segmentation Analysis **51**

Great Britain Clustering and Segmentation Analysis **54**

Germany Clustering and Segmentation Analysis **57**

Sentiment Analysis **60**

USA Sentiment Analysis **60**

Canada Sentiment Analysis **63**

France Sentiment Analysis **66**

Great Britain Sentiment Analysis **69**

Germany Sentiment Analysis **72**

Time Series Analysis **75**

USA Time Series Analysis **75**

Canada Time Series Analysis **77**

France Time Series Analysis **80**

Great Britain Time Series Analysis **82**

Germany Time Series Analysis **85**

References **87**

Document Change History

SNO	Date	Owner	Version	Description
1.	10/27/2017	Ananth, Gowri, Ashok	v1.0	Initial Draft v1.0
2.	11/25/2017	Ananth, Gowri, Ashok	v1.1	Final documentation and Project summary

Who Can Refer

This document can be referred by the following persons as a part of ITMD527 Data Analytics

- ✓ Robert Henkins
- ✓ Anantharaman Chandar
- ✓ Gowrisankar Arumugam
- ✓ Ashok Ramasami

Introduction

YouTube video statistics dataset has almost 2L records from with each video provides us information about statistic and sentiment. Each video has a unique video id that will be tagged along with likes/dislikes/comments/views. With this information, we can track the most popular videos, predict the videos that are highly popular among the users, maximum views of the video based on its category, most search video keywords.

Expected Outcomes

This project outcome is to provide the following Predictive Analysis

- ✓ How do you know if your videos are doing well?
- ✓ Why is it important to understand who your viewers are?
- ✓ Look at your channel's YouTube Analytics Interaction reports from the last six months.
- ✓ Name two reports that are most important for your channel
- ✓ What surprising things have you learned about your audience or viewership based on these reports?

Dataset

The YouTube Dataset has been collected from Kaggle.com. This dataset contains information on default videos based on demography.

Reference: (**https://bit.ly/2DAPfBk**).

Analytics Plan

Research Questions

Below are some of the research questions which will be analyzed and presented to the Higher Management for Decision making.

- ✓ How is my channel doing?
- ✓ Who's watching my channel?
- ✓ Understanding your video reach on YouTube
- ✓ How much money am I making?
- ✓ How engaged is my audience?
- ✓ Is a single video responsible for the drop?
- ✓ Was the content that was responsible for the drop seasonal or topical to what was happening in the world?

Hypothesis Testing

Assumptions to be considered for Hypothesis testing to start the project analysis

- ✓ H_0 = How well the video is reached to the end users?
- ✓ H_0 =How are the videos watched across various countries?
- ✓ H_0 = Does each country tradition plays a significance role in the YouTube video
- ✓ H_0 =Commonly watched video based on the countries?
- ✓ H_0 =Sentiment analysis should be used to predict video reach?
- ✓ H_0 =Trending dates of the video based on each country?

Data Definition

Please refer the below sections for detailed data definition.

- ✓ Raw Data's data definition can be found in Data Dictionary

Project Plan

Project Requirements

Below are the requirements which are required to proceed with YouTube Data Analysis project.

Domain Knowledge

- ✓ Domain Knowledge about the project background
- ✓ Picking your Dataset which matches the project from a trusted source
- ✓ Understanding the Raw Dataset and its corresponding data types
- ✓ Ability to interpret and infer the records matching the project background
- ✓ Preparing the scope of the project along with future enhancements
- ✓ Ability to prepare a summary report of what are all the findings clearly
- ✓ Preparing what are all the analysis that will be put forth for this project
- ✓ Documenting all your findings, analysis, code snippets and screenshots clearly
- ✓ Time Management

Technical Knowledge

- ✓ Programming Knowledge in Python
- ✓ Proficient in Microsoft Excel such as calculations, graphing tools, pivot tables, formulas and functions
- ✓ How to prepare Exploratory Data Analysis

System and Software Requirements

- ✓ A working laptop or PC or Tablet
- ✓ Jupyter Notebook
- ✓ Latest R Studio
- ✓ Microsoft Excel
- ✓ Microsoft Work
- ✓ Snipping Tool
- ✓ Internet connection

Project Constraints

This project will not be moved forward if the below constraints occur during this project analysis

- ✓ Not getting a valid or relevant Dataset from a trusted site
- ✓ Too many missing values and outliers
- ✓ Minimal records inside the dataset i.e. <100
- ✓ Unable to infer or identify a dependent variable
- ✓ If the dataset does not pass minimum criteria such as Hypothesis, goodness fit test etc.
- ✓ Unable to present what you are going to analyze or predict

Project Assumptions

Below are some of the assumptions which will be implemented during this project. Additional assumptions will be added and implemented as and when needed and will be documented precisely

General Assumptions

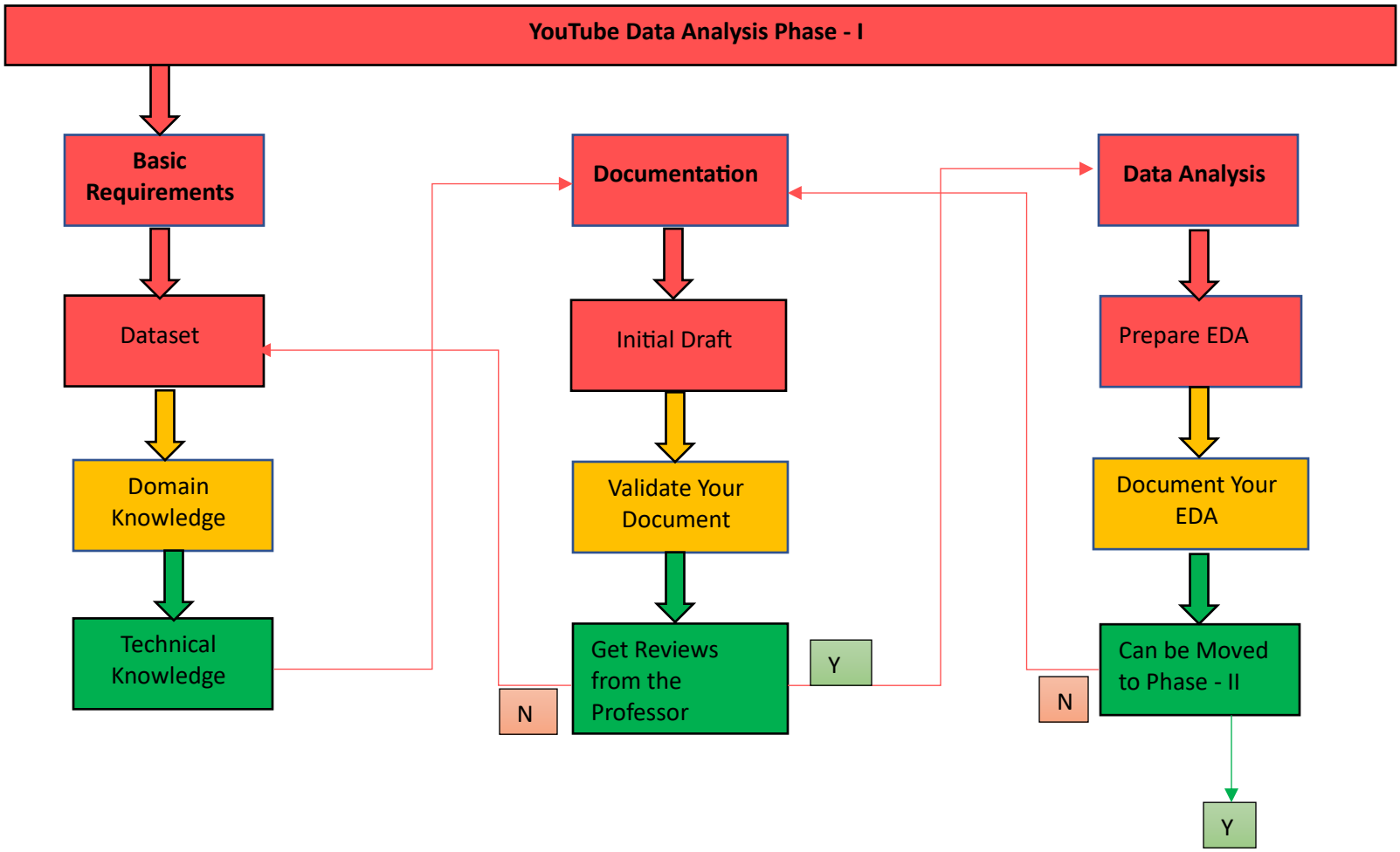
- ✓ Formatting the dataset such as moving the columns front and back if necessary to understand, sort a column by ascending or descending
- ✓ Changing column names for a better understanding of your dataset

Technical Assumptions

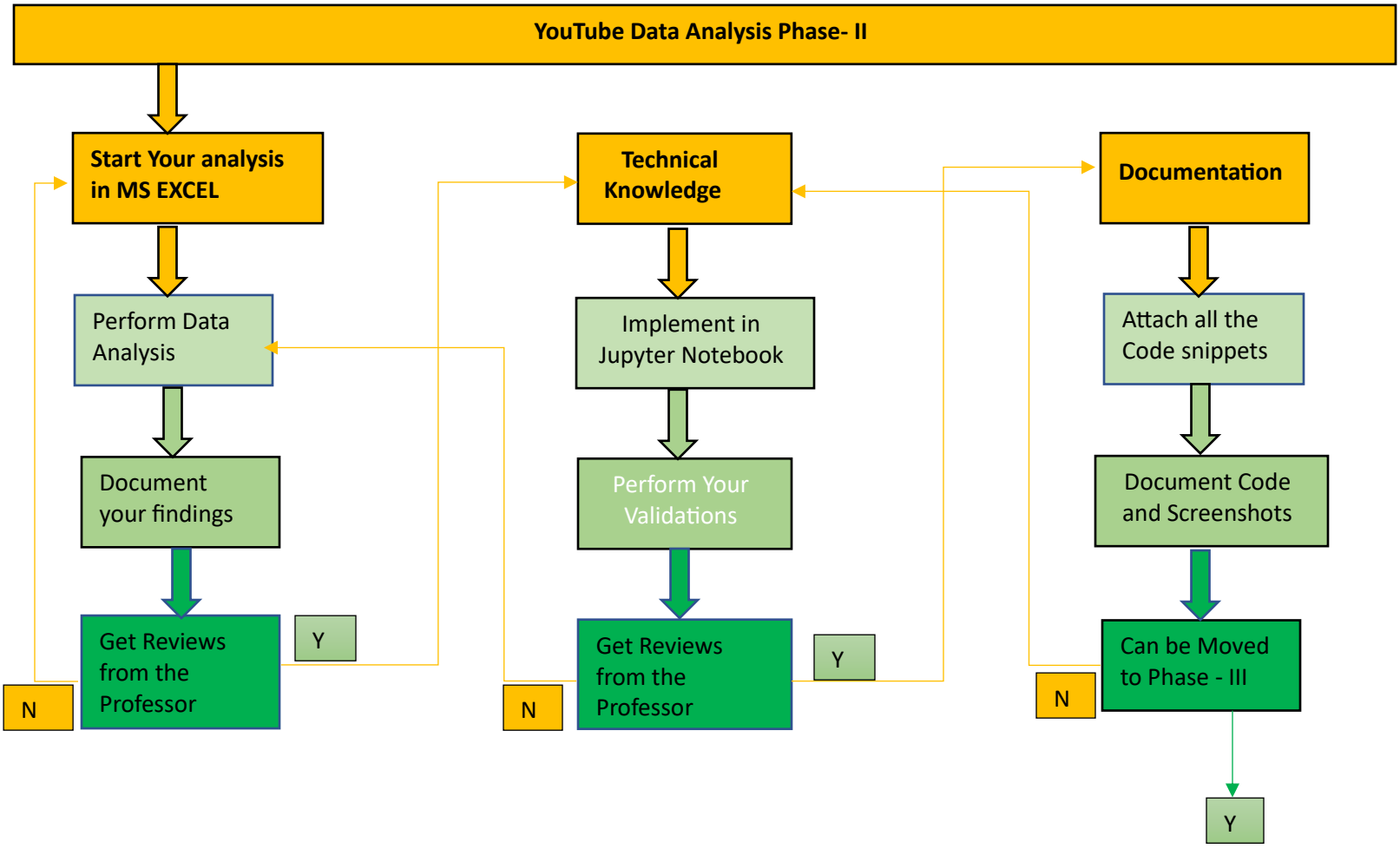
- ✓ Provide a similar variable name for all the columns from the dataset in jupyter notebook which can be a self-explanatory when anyone reads
- ✓ Mapping column names from the dataset with its corresponding variables in jupyter notebook
- ✓ Introduce dummy variable wherever necessary
- ✓ Eliminating the columns and variables which does not impact the analysis
- ✓ Standardize the values if necessary, such as 21-50 will be updated as 1, 51-100 will be updated as 2

Work Breakdown Structure

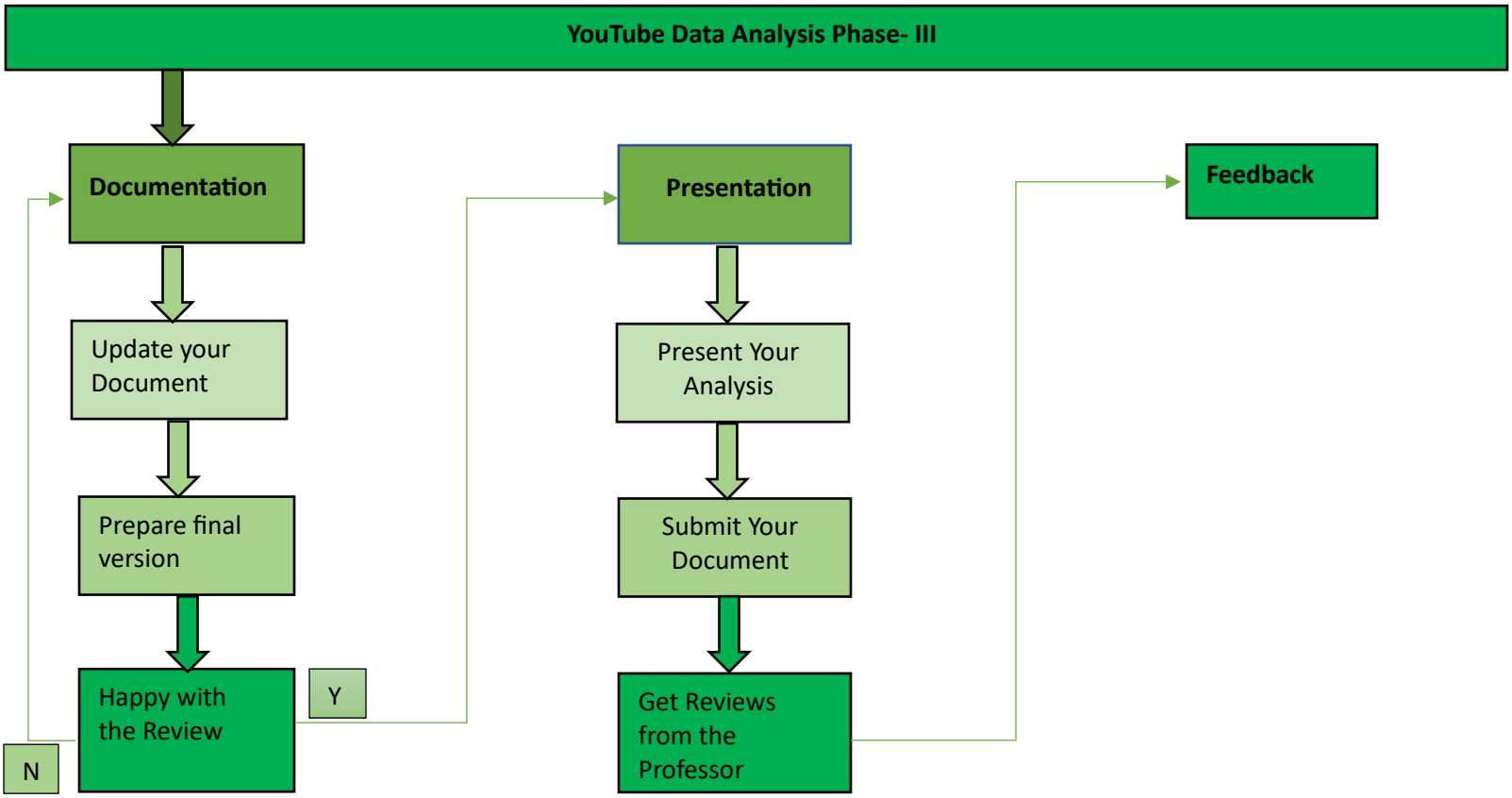
Phase - I



Phase – II


YouTube Data Analysis Phase- II
Start Your analysis in MS EXCEL
Perform Data Analysis
Document your findings
Get Reviews from the Professor
Y
N
Technical Knowledge
Implement in Jupyter Notebook
Perform Your Validations
Get Reviews from the Professor
Y
N
Documentation
Attach all the Code snippets
Document Code and Screenshots
Can be Moved to Phase - III
N
Y

Phase – III


YouTube Data Analysis Phase- III
Documentation
Update your Document
Prepare final version
Happy with the Review
Y
N
Presentation
Present Your Analysis
Submit Your Document
Get Reviews from the Professor
Feedback

Team Roles and Responsibilities

Phase – I and Phase - II

CWID	Name	Responsibilities
A20403439	**Ananth**	Identify Dataset
		Review with Team members
		Project Summary Document
		Solution Development and Report
		Prepare EDA, Project Plan, Analytics Plan
		Document Project Report Phase - I
		Time Series Analysis
		Documentation of Time Series Analysis
		Final draft documentation and review

CWID	Name	Responsibilities
A20400590	**Gowrisankar Arumugam**	Identify Dataset
		Document Project Summary
		Prepare Project Plan
		Prepare EDA, Project Plan, Analytics Plan
		Document Project Report Phase - II
		Sentiment Analysis
		Documentation of Sentiment Analysis
		Final draft documentation and review

CWID	Name	Responsibilities
A20401032	**Ashok Ramasami**	Discuss the dataset
		Document Project Summary
		Prepare Analytics Plan
		Prepare EDA, Project Plan, Analytics Plan
		Review Project Report Phase – I and Phase-II
		Clustering and Segmentation
		Documentation of Clustering and Segmentation
		Final draft documentation and review

Solution Development and Implementation Report

Data Dictionary

Column Datatypes					
Column Name	**CANADA**	**DENMARK**	**FRANCE**	**GREAT BRITAN**	**USA**
video_id	TEXT	TEXT	TEXT	TEXT	TEXT
trending_date	DATE	DATE	DATE	DATE	DATE
title	TEXT	TEXT	TEXT	TEXT	TEXT
channel_title	TEXT	TEXT	TEXT	TEXT	TEXT
category_id	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
publish_time	DATETIME	DATETIME	DATETIME	DATETIME	DATETIME
tags	TEXT	TEXT	TEXT	TEXT	TEXT
views	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
likes	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
dislikes	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
comment_count	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
thumbnail_link	TEXT	TEXT	TEXT	TEXT	TEXT
comments_disabled	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN
ratings_disabled	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN
video_error_or_removed	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN
description	TEXT	TEXT	TEXT	TEXT	TEXT

Column Description

Column Description	
Column Name	**Description**
video_id	ID associated with each video
trending_date	Date where it reached highest views
title	Title of the video

channel_title	Artist who published this video
category_id	Category of the video
publish_time	Date of release
tags	Key words for Hashtags
views	Total No.of Views
likes	Total No.of Likes
dislikes	Total No.of Dislikes
comment_count	Total No.of Comments
thumbnail_link	Image of the video
comments_disabled	Comments can be posted or not
ratings_disabled	Ratings can be posted or not
video_error_or_removed	Video still available
description	Description of the video

Raw Data Quality

Missing Values					
Column Name	**CANADA**	**DENMARK**	**FRANCE**	**GREAT BRITAN**	**USA**
video_id	0	0	0	0	0
trending_date	0	0	0	0	0
title	0	0	0	0	0
channel_title	0	0	0	0	0
category_id	0	0	0	0	0
publish_time	0	0	0	0	0
tags	0	0	0	0	0
views	0	0	0	0	0
likes	0	0	0	0	0
dislikes	0	0	0	0	0
comment_count	0	0	0	0	0
thumbnail_link	0	0	0	0	0
comments_disabled	0	0	0	0	0
ratings_disabled	0	0	0	0	0
video_error_or_removed	0	0	0	0	0
description	1296	1552	2912	612	570

Statistics Distribution

Statistics						
Description	**CANADA**	**DENMARK**	**FRANCE**	**GREAT BRITAN**	**USA**	**Total**
Total Records	40881	40840	40724	38916	40949	202310
Blanks	1296	1552	2912	612	570	6942
Missing Values %	3.17%	3.80%	7.15%	1.57%	1.39%	3.43%
Data Quality %	96.83%	96.20%	92.85%	98.43%	98.61%	96.57%

Columns to be Considered

Columns to be Considered					
Column Name	**CANADA**	**DENMARK**	**FRANCE**	**GREAT BRITAN**	**USA**
video_id	Y	Y	Y	Y	Y
trending_date	Y	Y	Y	Y	Y
title	Y	Y	Y	Y	Y
channel_title	Y	Y	Y	Y	Y
category_id	Y	Y	Y	Y	Y
publish_time	Y	Y	Y	Y	Y
tags	Y	Y	Y	Y	Y
views	Y	Y	Y	Y	Y
likes	Y	Y	Y	Y	Y
dislikes	Y	Y	Y	Y	Y
comment_count	Y	Y	Y	Y	Y
~~***thumbnail_link***~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~	~~NA~~
comments_disabled	Y	Y	Y	Y	Y
ratings_disabled	Y	Y	Y	Y	Y
video_error_or_removed	Y	Y	Y	Y	Y
description	Y	Y	Y	Y	Y

Reason for Elimination

Column Name	*Reason for Elimination*
~~***thumbnail_link***~~	*Not necessary to show each video icon image*

Quartile Distribution Based on Country

Canada

CANADA					
Quartile Range	**category_id**	**views**	**likes**	**dislikes**	**comment_count**
Minimum	1	733	0	0	0
First Quartile	20	143902	2191	99	417
Median	24	371204	8780	303	1301
Third Quartile	24	963302	28717	950	3713
Maximum	43	137843120	5053338	1602383	1114800

France

FRANCE					
Quartile Range	**category_id**	**views**	**likes**	**dislikes**	**comment_count**
Minimum	1	223	0	0	0
First Quartile	17	16974.5	338	18	56
Median	23	73721	1892.5	83	235
Third Quartile	24	270808.75	7969.5	335	841
Maximum	44	100911567	4750254	1353661	1040912

Denmark

DENMARK					
Quartile Range	**category_id**	**views**	**likes**	**dislikes**	**comment_count**
Minimum	1	518	0	0	0
First Quartile	20	27068.75	533	29	79
Median	24	119277	2699	134	376
Third Quartile	44	113876217	4924056	1470386	1084435
Maximum	44	113876217	4924056	1470386	1084435

Great Britain

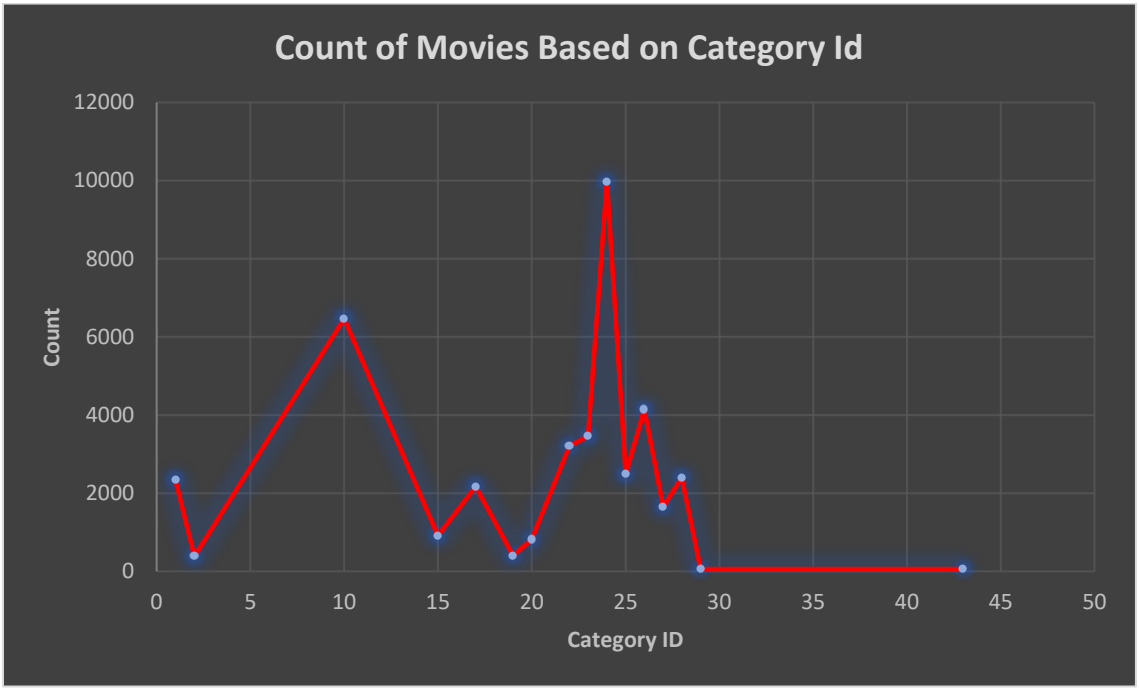
GREAT BRITIAN					
Quartile Range	**category_id**	**views**	**likes**	**dislikes**	**comment_count**
Minimum	1	851	0	0	0
First Quartile	10	251527.25	5897	200	679
Median	20	981889	25182.5	821	2478
Third Quartile	24	3683628.5	114089.3	3357.5	9241.5
Maximum	43	424538912	5613827	1944971	1626501

United States

USA					
Quartile Range	**category_id**	**views**	**likes**	**dislikes**	**comment_count**
Minimum	1	549	0	0	0
First Quartile	17	242329	5424	202	614
Median	24	681861	18091	631	1856
Third Quartile	25	1823157	55417	1938	5755
Maximum	43	2.25E+08	5613827	1674420	1361580

Exploratory Data Analysis of YouTube Data based on Countries [USA]

USA Count of Movies Based on Category

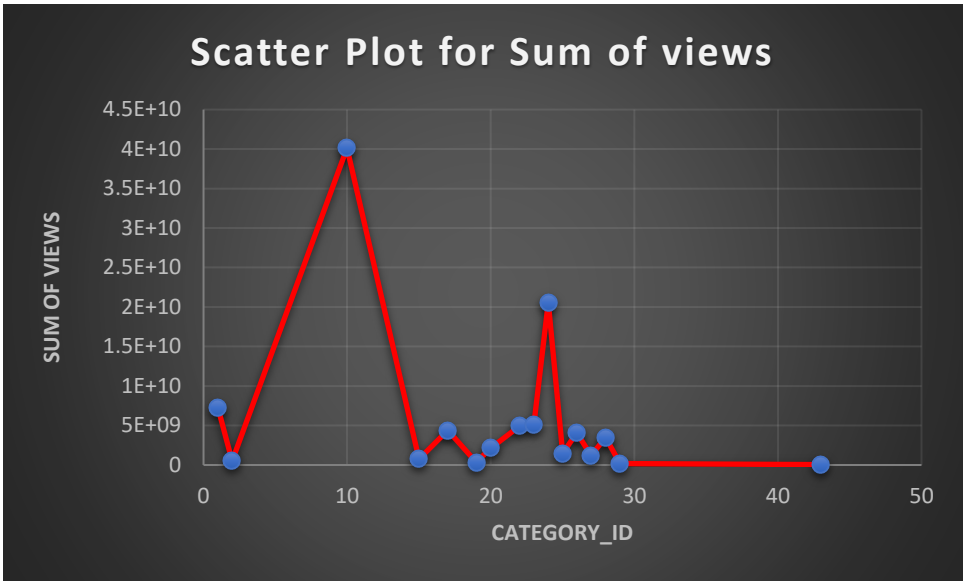


Total Views/Likes/Dislikes/Comments based on Category Id

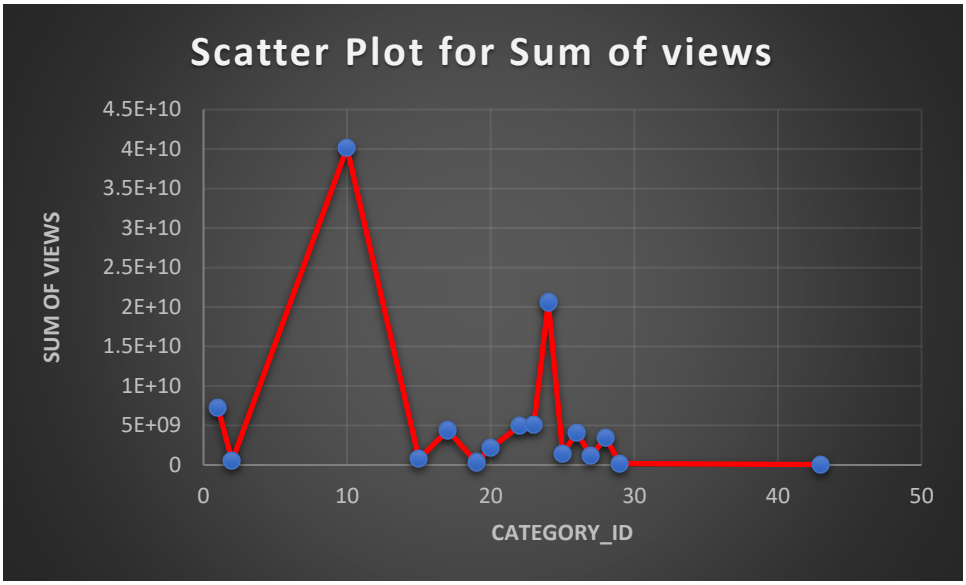
Category_ID	Sum of views	Sum of likes	Sum of dislikes	Sum of comment_count
1	7284156721	165997476	6075148	17887060
2	520690717	4245656	243010	784447
10	40132892190	1416838584	51179008	125296396
15	764651989	19370702	527379	2660705
17	4404456673	98621211	5133551	11192155
19	343557084	4836246	340427	911511
20	2141218625	69038284	9184466	14740713
22	4917191726	186615999	10187901	24778032
23	5117426208	216346746	7230391	22545582

24	20604388195	530516491	42987663	73566498
25	1473765704	18151033	4180049	6039433
26	4078545064	162880075	5473899	23149550
27	1180629990	49257772	1351972	5442242
28	3487756816	82532638	4548402	11989926
29	168941392	14815646	3310381	4808797
43	51501058	1082639	24508	95117
Grand Total	**96671770152**	**3041147198**	**151978155**	**345888164**

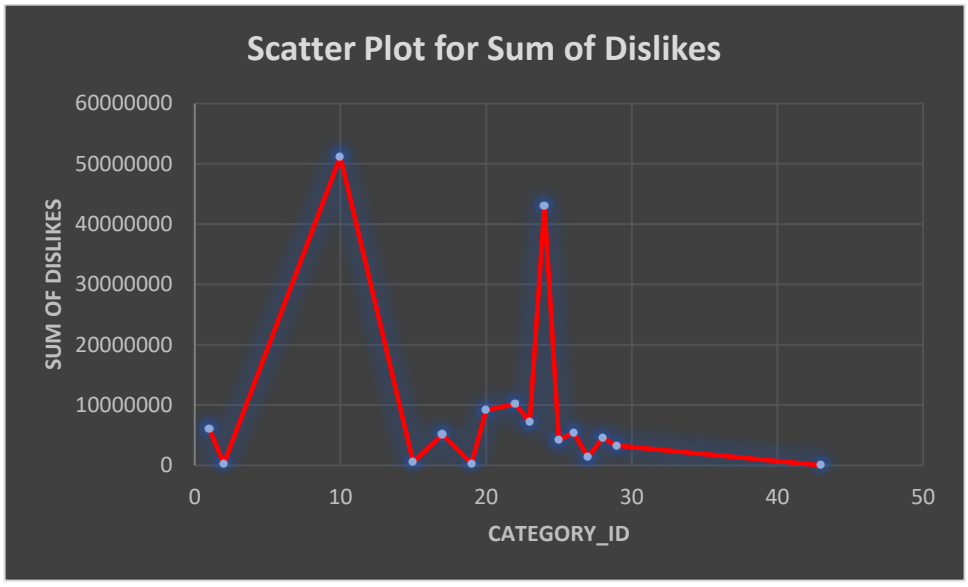
Scatter Plot Analysis for Sum of Views



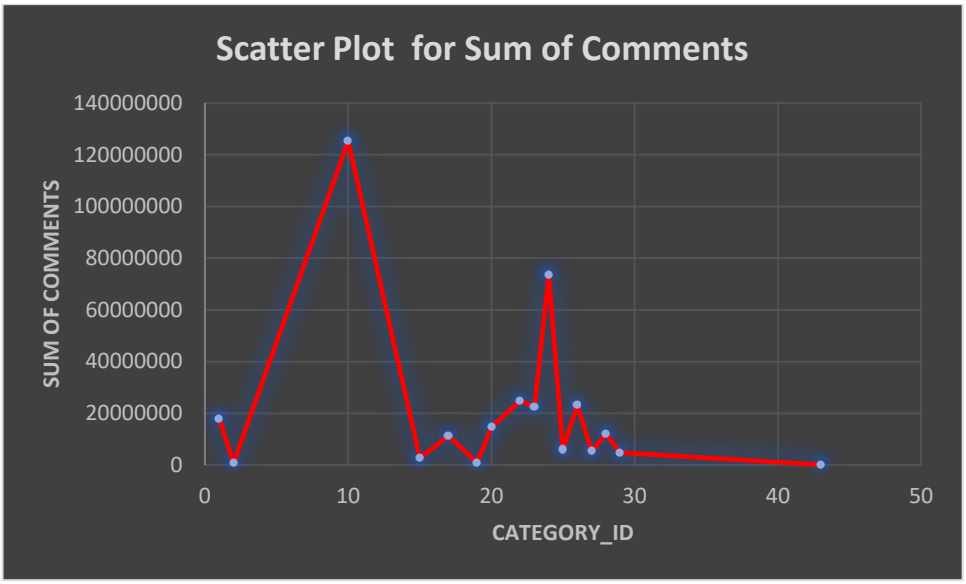
Scatter Plot Analysis for Sum of Likes



Scatter Plot Analysis for Sum of Dislikes



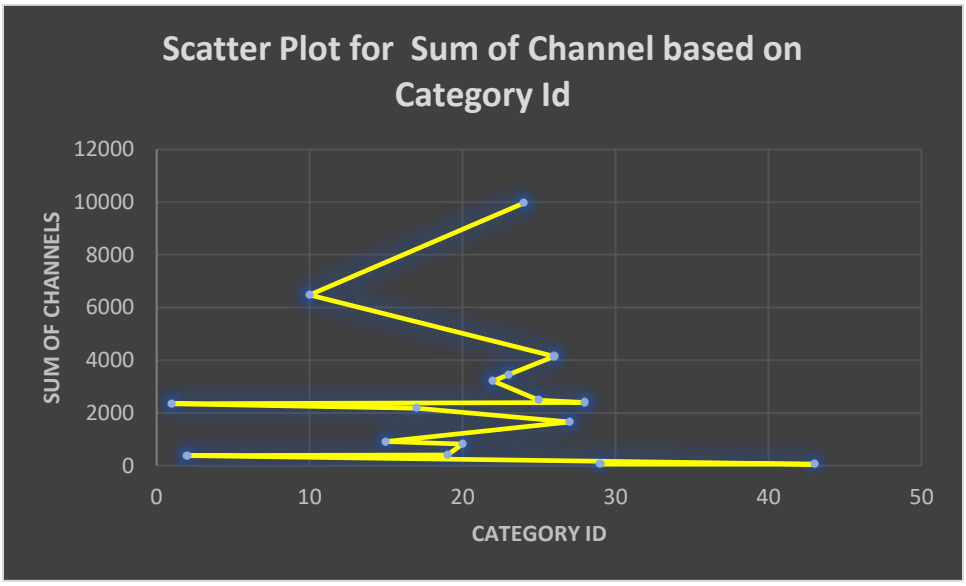
Scatter Plot Analysis for Sum of Comments



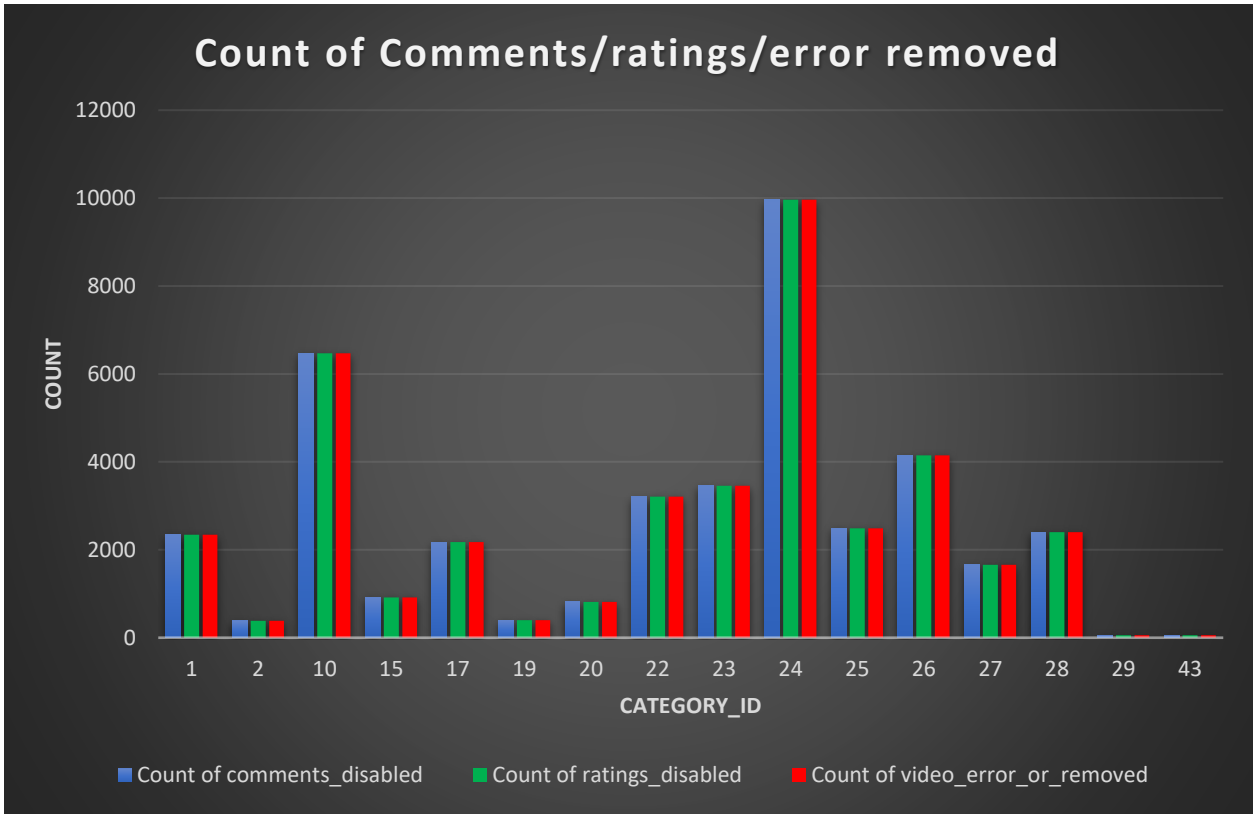
Average Views/Likes/Dislikes/Comments based on Category Id

Category_Id	Average of views	Average of likes	Average of dislikes	Average of comment_count
1	3106250.20	70787.84	2590.68	7627.74
2	1355965.41	11056.40	632.84	2042.83
10	6201003.12	218918.20	7907.76	19359.76
15	831143.47	21055.11	573.24	2892.07
17	2025969.03	45363.94	2361.34	5148.19
19	854619.61	12030.46	846.83	2267.44
20	2620830.63	84502.18	11241.70	18042.49
22	1531835.43	58135.83	3173.80	7719.01
23	1480308.42	62582.22	2091.52	6521.72
24	2067883.20	53243.33	4314.30	7383.23
25	592587.74	7298.36	1680.76	2428.40
26	983730.12	39286.08	1320.28	5583.59
27	712940.82	29745.03	816.41	3286.38
28	1452626.75	34374.28	1894.38	4993.72
29	2963884.07	259923.61	58076.86	84364.86
43	903527.33	18993.67	429.96	1668.72
Grand Total	**2360784.64**	**74266.70**	**3711.40**	**8446.80**

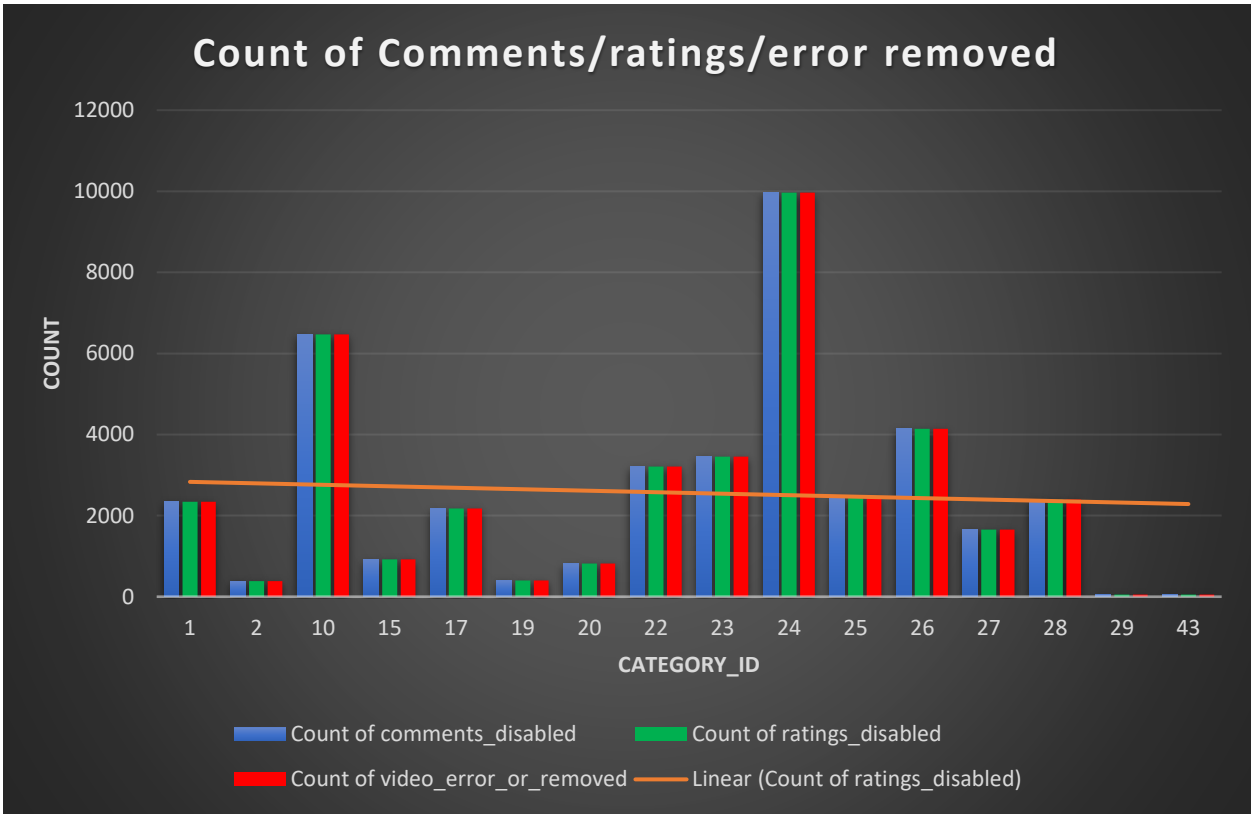
Scatter Plot for Sum of Channels based on Category ID



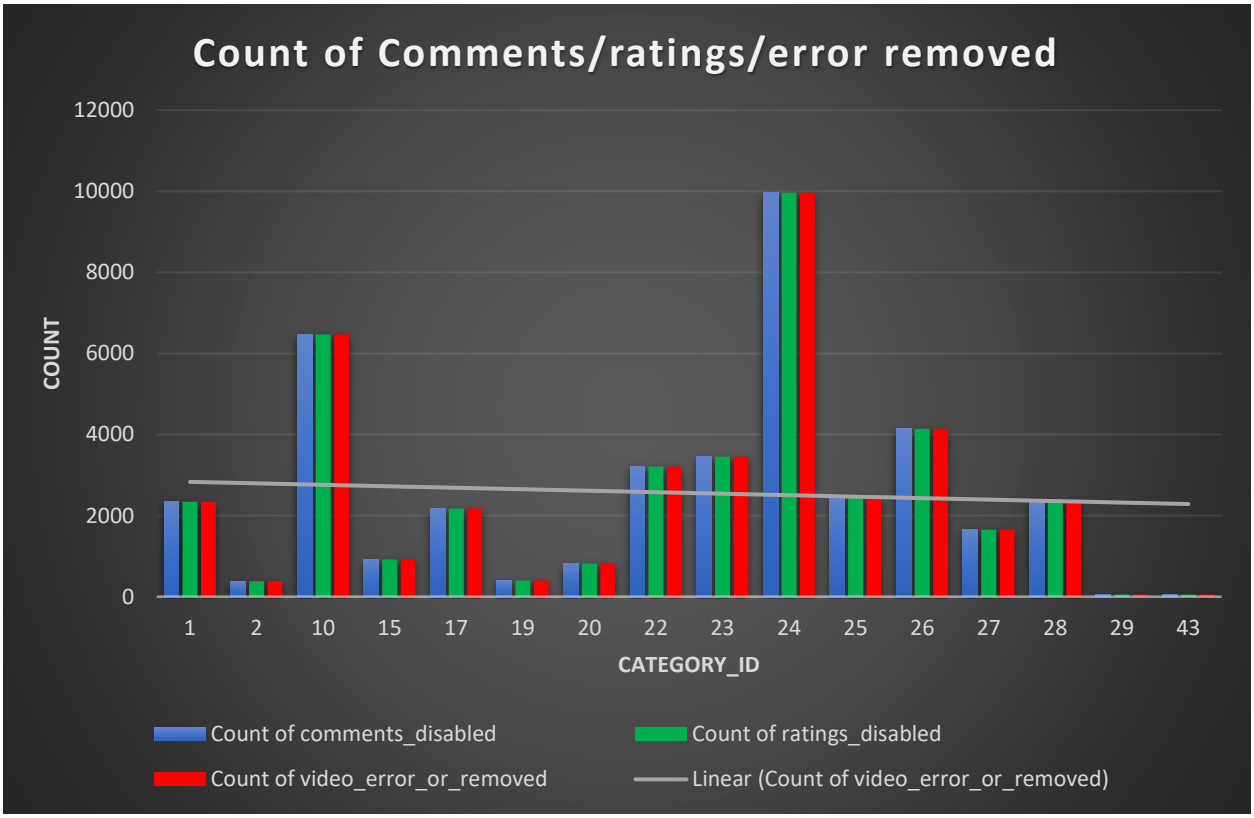
Bar chart for Count of Comments/ratings/videos



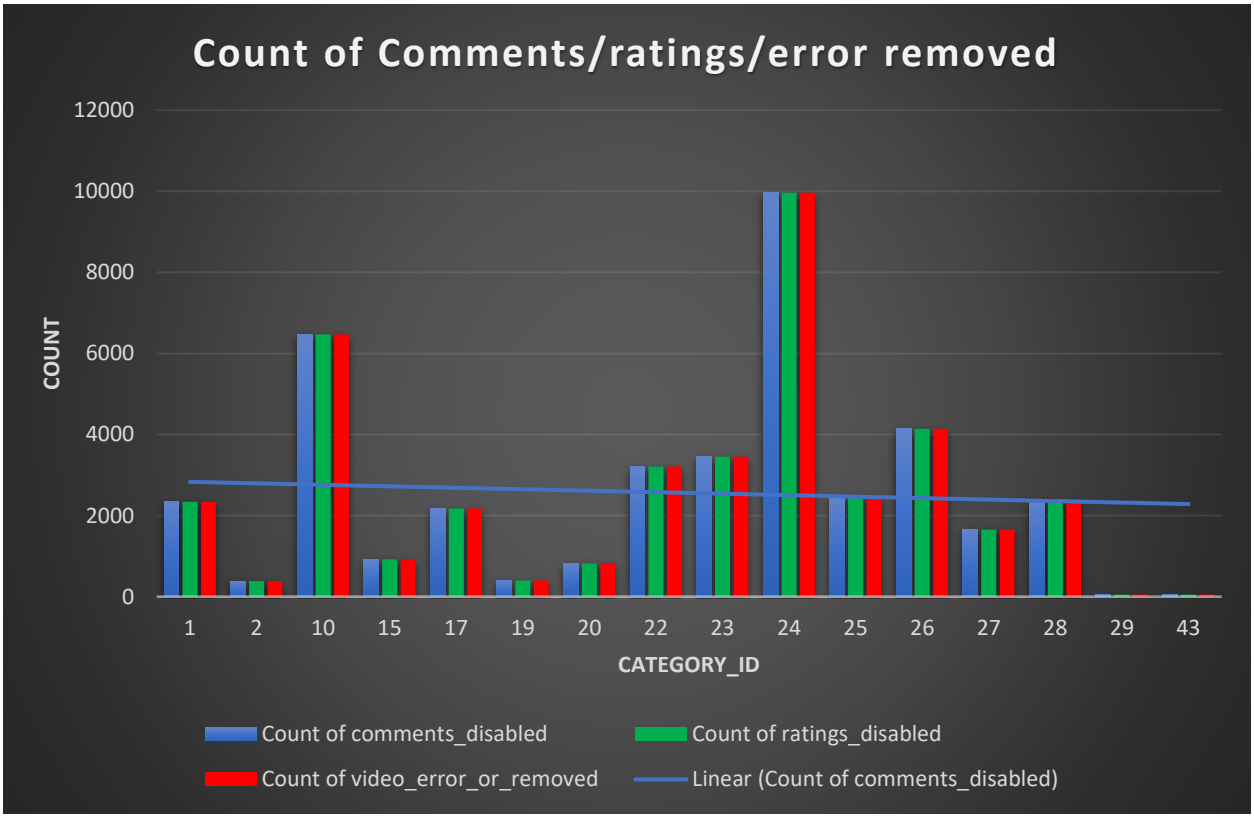
Linear Trendline for Count of Ratings Disabled



Linear Trendline for Count of Videos Removed



Linear Trendline for Count of Comments Disabled



Model Development and Implementation

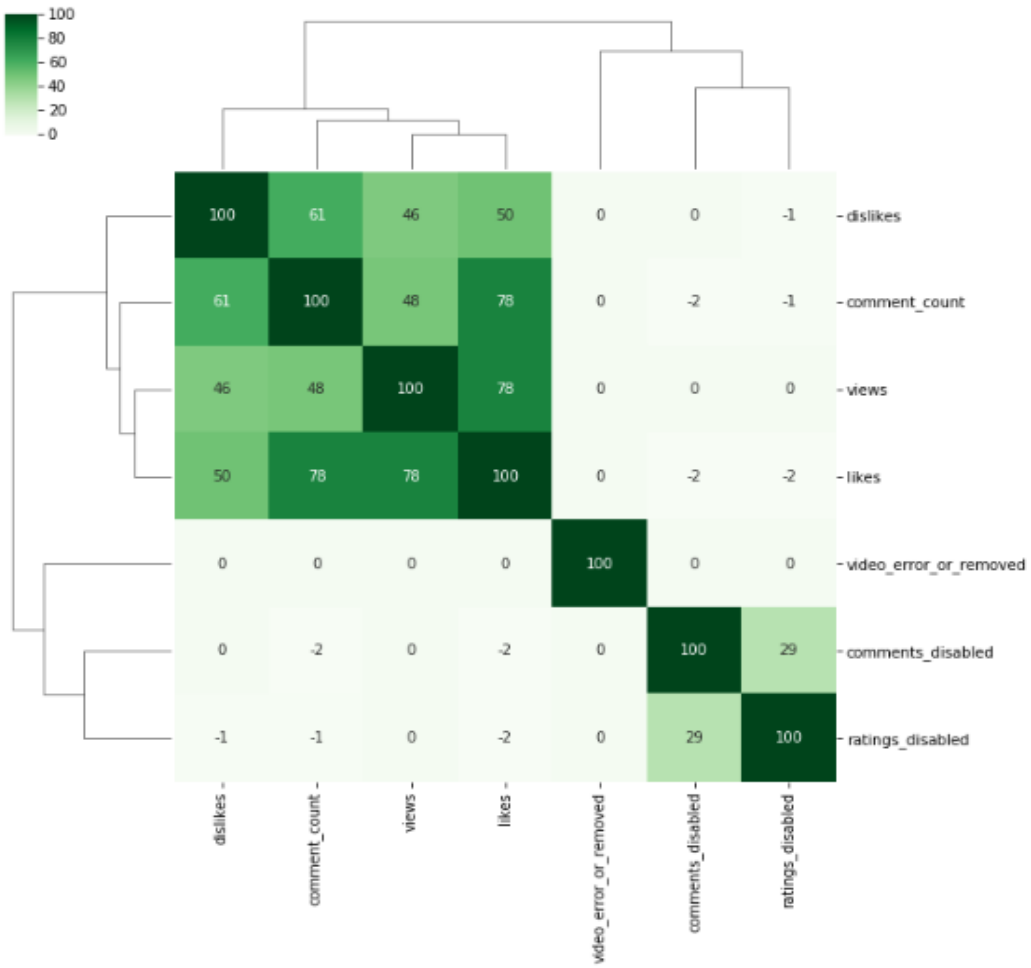
Cluster Map

Description: Clustering is a type of pre-writing that allows a writer to explore many ideas as soon as they occur to them. Like brainstorming or free associating, clustering allows a writer to begin without clear ideas.

Attributes Taken: Dislikes/Comments_Count/Views/Video removed/Comments disabled/Ratings disabled

Code: Clustering.py

Results:



Conclusion:

Based on the scatterplot and the cluster map it is evident that all the columns except video_error_or_removed should be retained. The new data set will be without the above-mentioned column.

Next Step: Perform scatter plot analysis for the remaining columns to check for the correlation.

Scatter Plot

Description: A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are color-coded, one additional variable can be displayed

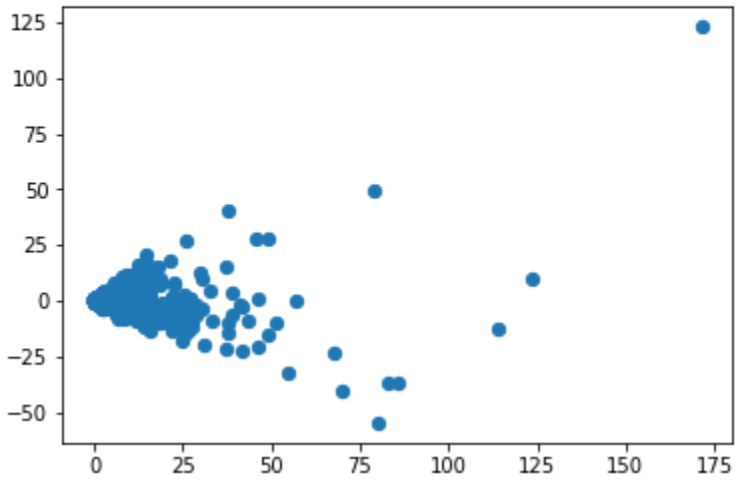
Attributes Taken: trending_Date/Channel_title/Category_Id/Publish_date/Publish_Time/Views/likes/dislikes/comments_count/comments_disabled/ratings_disabled/country

Code: ScatterPlot.py

ScatterPlot.py

Results:

```
# plot the standardized data
plt.scatter(Y_sklearn[:,0], Y_sklearn[:,1], s=40)
plt.show()
```



Conclusion:

This shows the extent of correlation between the columns trending_Date/Channel_title/Category_Id/Publish_date/Publish_Time/Views/likes/dislikes/comments_count/comments_disabled/ratings_disabled/country

Principal Component Analysis

Description: Scatter plot of principal component analysis (PCA) Component 1 versus PCA Component 2 scores. Each point is represented by a symbol denoting its analytical cluster and a line connecting it to the cluster centroid

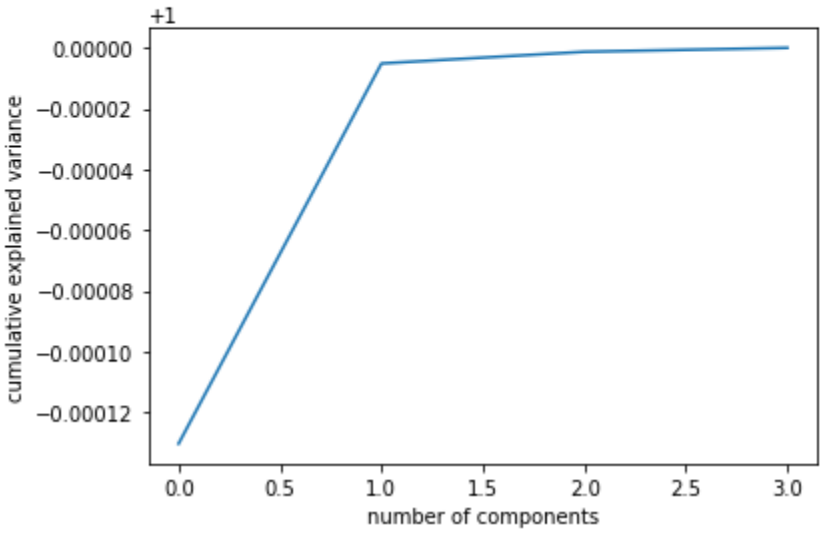
Attributes **Taken**: trending_Date/Channel_title/Category_Id/Publish_date/Publish_Time/Views/likes/dislikes/comments_count/comments_disabled/ratings_disabled/country

Code: R_Code_Regression.R

PAC.py

Results:

```
# plot the PCA components
pca = PCA().fit(X)
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance')
plt.show()
```



Conclusion:

Each dot in this plot represents one community. Looking at the red dot out by itself to the right, you may conclude that this particular dot has a very high value for the first principal component and we would expect this community to have high values for the attributes taken

Exploratory Data Analysis of YouTube Data based on each Country

Description: In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task

USA Exploratory Data Analysis

Uniqueness of each Columns: Uniqueness is a state or condition wherein someone or something is unlike anything else in comparison.

No.of Nulls in each Column: Often, data can have missing values due to a variety of reasons, for example with survey data, some observations may not have been recorded

Log Distribution of Views/Comments/Dislikes/likes: In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

Total number of records based on the category: To display the number of rows that are in the dataset based on the category

Code: ADA_EDA_US.ipynb

ADA_EDA_US.ipynb

Results:

Uniqueness of each Columns

```
In [4]:
## Identift the uniqueness

print(df_YT_US.nunique())
```

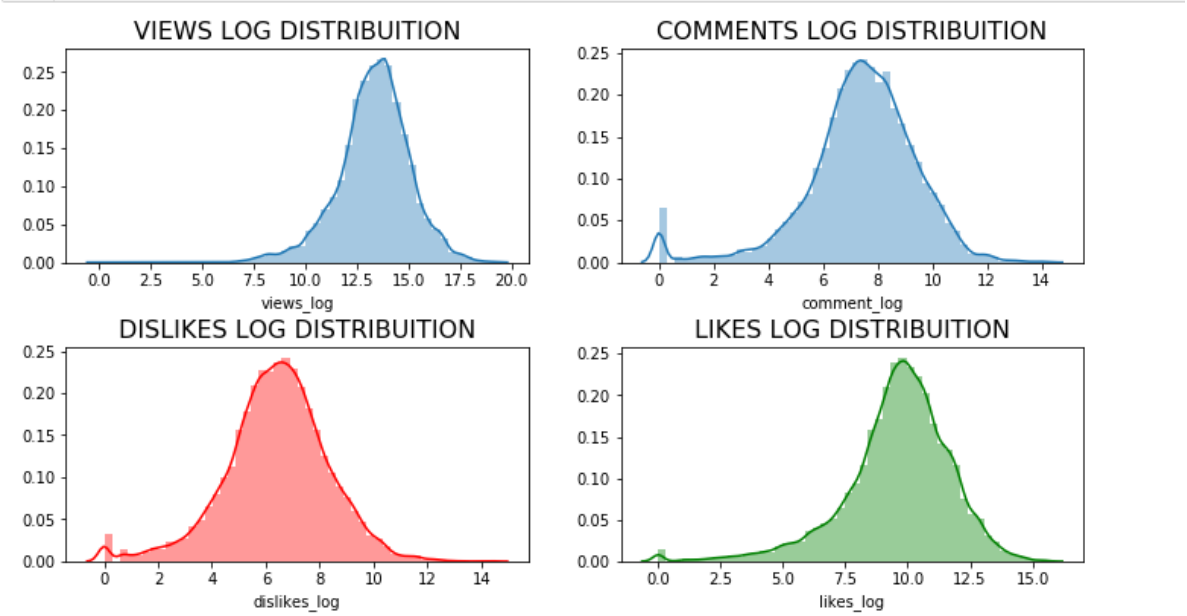
```
video_id                  6283
trending_date              207
title                     6456
channel_title             2209
category_id                 17
publish_time              6271
tags                      6059
views                    40479
likes                    29850
dislikes                  8516
comment_count            13773
thumbnail_link            6354
comments_disabled            6
ratings_disabled             5
video_error_or_removed       6
description               6903
dtype: int64
```

No.of Nulls in each Column:

```
In [5]:
# No of Nulls
df_YT_US.info()
```

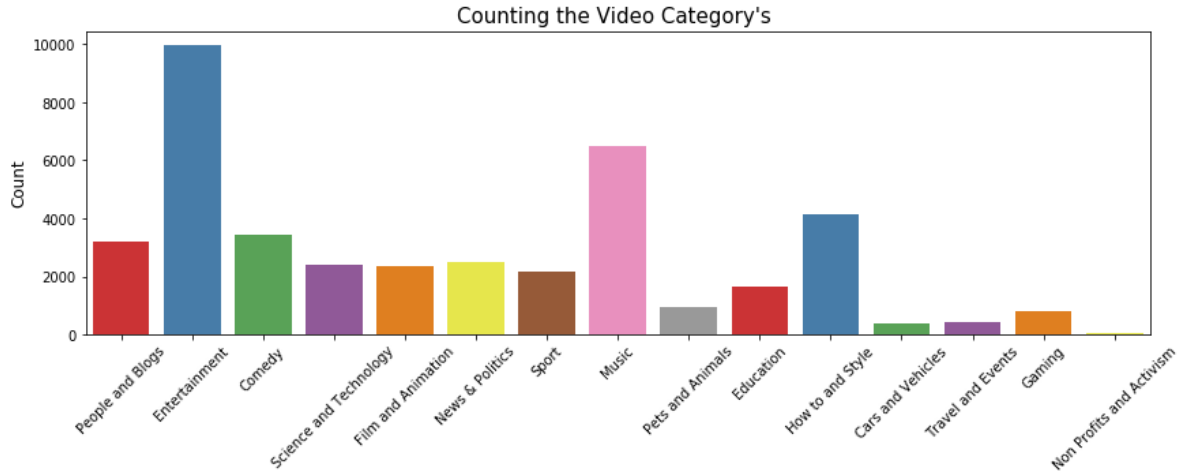
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40955 entries, 0 to 40954
Data columns (total 16 columns):
video_id                  40955 non-null object
trending_date             40955 non-null object
title                     40955 non-null object
channel_title             40955 non-null object
category_id               40955 non-null int64
publish_time              40955 non-null object
tags                      40955 non-null object
views                     40955 non-null int64
likes                     40955 non-null int64
dislikes                  40955 non-null int64
comment_count             40955 non-null int64
thumbnail_link            40955 non-null object
comments_disabled         40955 non-null object
ratings_disabled          40955 non-null object
video_error_or_removed    40955 non-null object
description               40385 non-null object
dtypes: int64(5), object(11)
memory usage: 3.3+ MB
```

Log Distribution of Views/Comments/Dislikes/likes:



Total number of records based on the category:

```
Category Name count
Entertainment       9965
Music               6472
How to and Style    4146
Comedy              3457
People and Blogs    3210
Name: category_name, dtype: int64
```



Conclusion:

In general, life is sub-optimal and so are the data sets we work with. Therefore, to add that during EDA, one detects all kinds of singularities/anomalies/outliers/etc. Do not ignore these signals. Sometime these things tell and reveal super important insights and provide deeper understanding of the data sets' nature.

Canada Exploratory Data Analysis

Uniqueness of each Columns: Uniqueness is a state or condition wherein someone or something is unlike anything else in comparison.

No.of Nulls in each Column: Often, data can have missing values due to a variety of reasons, for example with survey data, some observations may not have been recorded

Log Distribution of Views/Comments/Dislikes/likes: In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

Total number of records based on the category: To display the number of rows that are in the dataset based on the category

Code: ADA_EDA_CA.ipynb

ADA_EDA_CA.ipynb

Results:

Uniqueness of each Columns

In [4]:
```
## Identify the uniqueness

print(df_YT_CA.nunique())
```

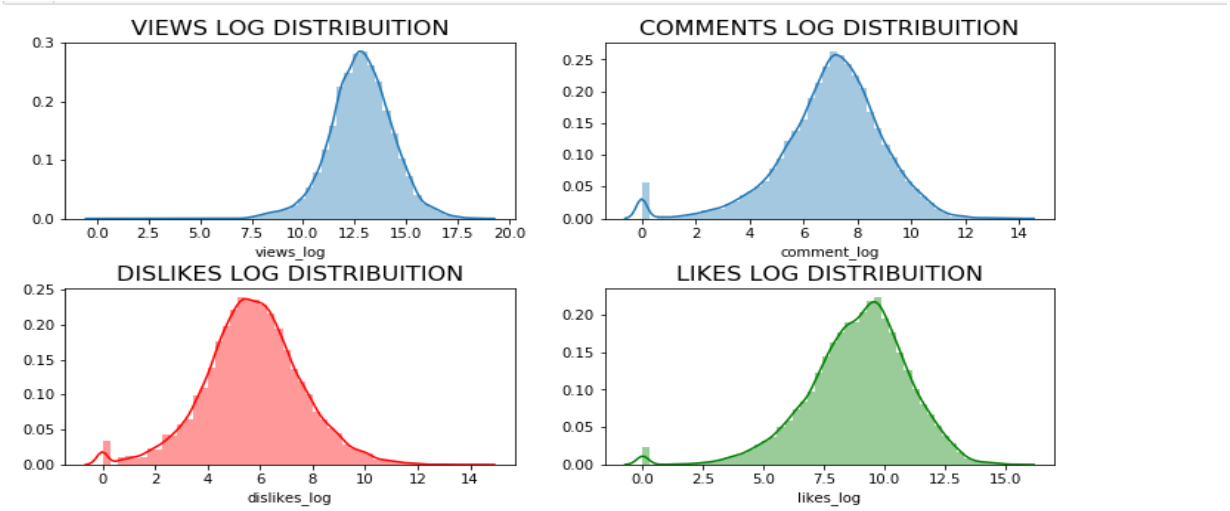
```
Unnamed: 0                  1
video_id                24105
trending_date             206
title                   24574
channel_title            5077
category_id                18
publish_time            23615
tags                    20159
views                   40171
likes                   24676
dislikes                 6241
comment_count           11172
thumbnail_link          24424
comments_disabled           6
ratings_disabled            6
video_error_or_removed      6
description             22343
Special Characters          1
dtype: int64
```

No.of Nulls in each Column:

```
In [5]:  1 #No of nulls
         2 df_YT_CA.info()
```

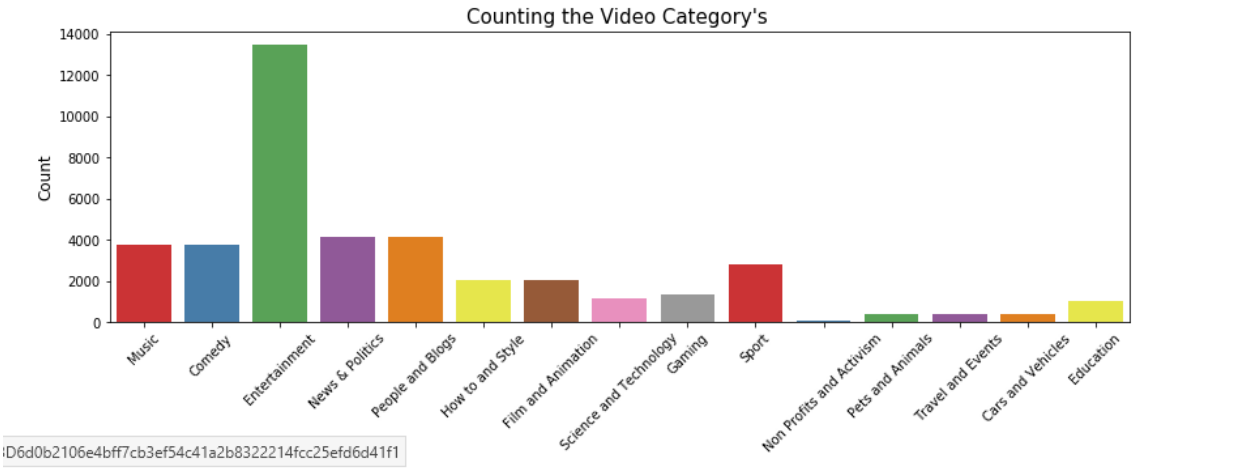
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40887 entries, 0 to 40886
Data columns (total 18 columns):
Unnamed: 0                1 non-null object
video_id                  40882 non-null object
trending_date             40882 non-null object
title                     40882 non-null object
channel_title             40882 non-null object
category_id               40887 non-null int64
publish_time              40887 non-null object
tags                      40887 non-null object
views                     40887 non-null int64
likes                     40887 non-null int64
dislikes                  40887 non-null int64
comment_count             40887 non-null int64
thumbnail_link            40887 non-null object
comments_disabled         40887 non-null object
ratings_disabled          40887 non-null object
video_error_or_removed    40887 non-null object
description               39591 non-null object
Special Characters        40882 non-null float64
dtypes: float64(1), int64(5), object(12)
memory usage: 3.7+ MB
```

Log Distribution of Views/Comments/Dislikes/likes:



Total number of records based on the category:

```
Category Name count
Entertainment       13453
News & Politics      4159
People and Blogs     4105
Comedy               3773
Music                3731
Name: category_name, dtype: int64
```



D6d0b2106e4bff7cb3ef54c41a2b8322214fcc25efd6d41f1

Conclusion:

In general, life is sub-optimal and so are the data sets we work with. Therefore, to add that during EDA, one detects all kinds of singularities/anomalies/outliers/etc. Do not ignore these signals. Sometime these things tell and reveal super important insights and provide deeper understanding of the data sets' nature.

France Exploratory Data Analysis

Uniqueness of each Columns: Uniqueness is a state or condition wherein someone or something is unlike anything else in comparison.

No.of Nulls in each Column: Often, data can have missing values due to a variety of reasons, for example with survey data, some observations may not have been recorded

Log Distribution of Views/Comments/Dislikes/likes: In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

Total number of records based on the category: To display the number of rows that are in the dataset based on the category

Code: ADA_EDA_FR.ipynb

ADA_EDA_FR.ipynb

Results:

Uniqueness of each Columns

```
In [3]: 1 print(df_YT_FR.nunique())
```

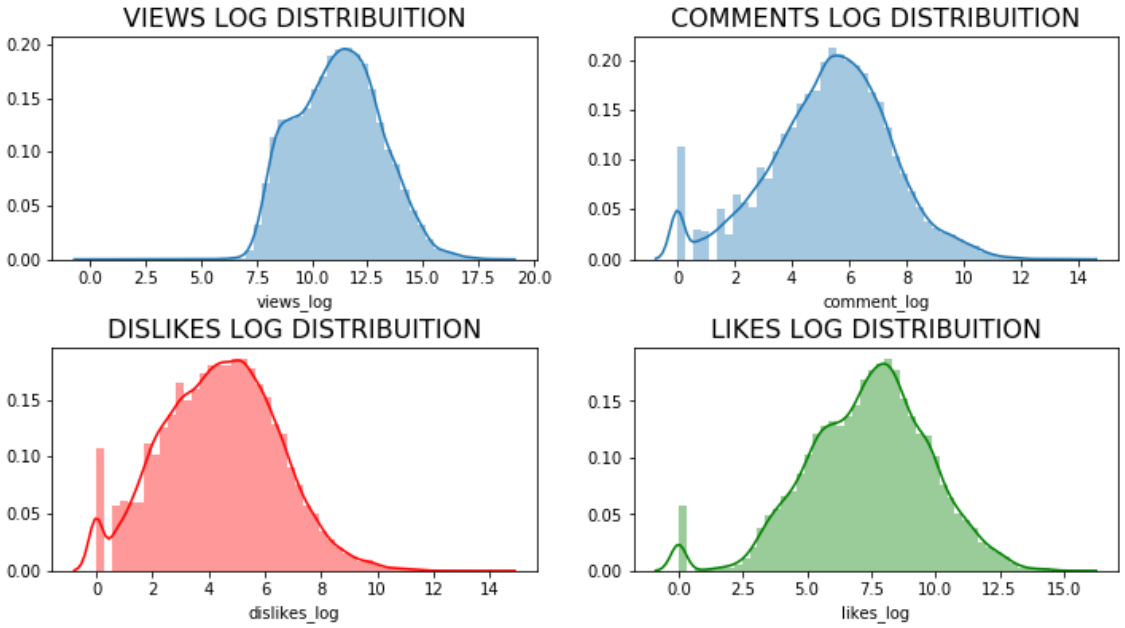
```
video_id                  30183
trending_date               207
title                     30548
channel_title              6681
category_id                  19
publish_time              29236
tags                      22907
views                     36446
likes                     15620
dislikes                   3736
comment_count              5832
thumbnail_link            30574
comments_disabled             6
ratings_disabled              6
video_error_or_removed        5
description               25006
dtype: int64
```

No.of Nulls in each Column:

```
In [4]: 1 df_YT_FR.info()
```

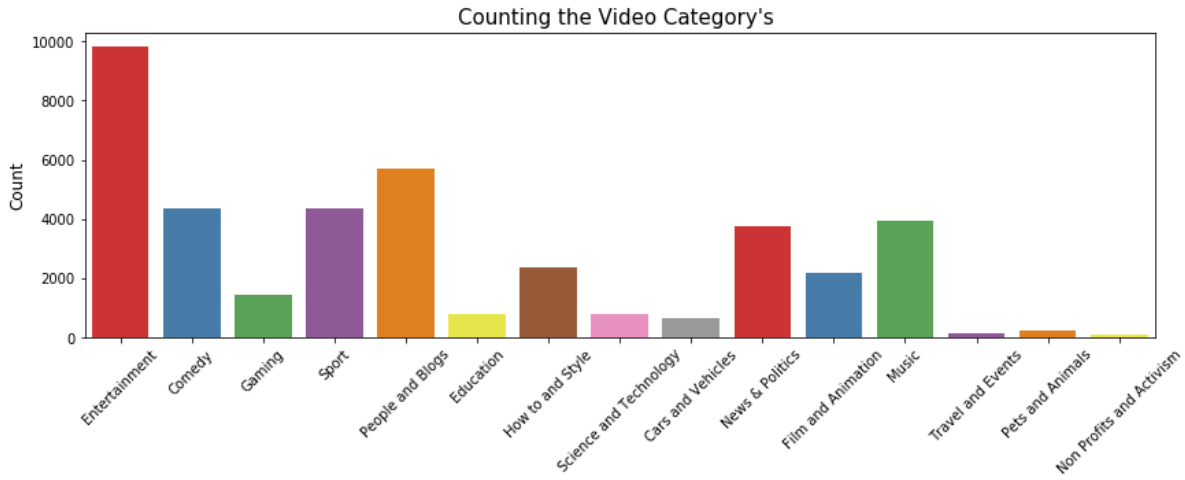
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40730 entries, 0 to 40729
Data columns (total 16 columns):
video_id                  40730 non-null object
trending_date             40730 non-null object
title                     40730 non-null object
channel_title             40730 non-null object
category_id               40730 non-null int64
publish_time              40730 non-null object
tags                      40730 non-null object
views                     40730 non-null float64
likes                     40730 non-null float64
dislikes                  40730 non-null int64
comment_count             40730 non-null int64
thumbnail_link            40730 non-null object
comments_disabled         40730 non-null object
ratings_disabled          40730 non-null object
video_error_or_removed    40730 non-null object
description               37818 non-null object
dtypes: float64(2), int64(3), object(11)
memory usage: 3.3+ MB
```

Log Distribution of Views/Comments/Dislikes/likes:



Total number of records based on the category:

```
Category Name count
Entertainment       9820
People and Blogs    5719
Comedy              4344
Sport               4343
Music               3946
Name: category_name, dtype: int64
```



Conclusion:

In general, life is sub-optimal and so are the data sets we work with. Therefore, to add that during EDA, one detects all kinds of singularities/anomalies/outliers/etc. Do not ignore these signals. Sometime these things tell and reveal super important insights and provide deeper understanding of the data sets' nature.

Great Britain Exploratory Data Analysis

Uniqueness of each Columns: Uniqueness is a state or condition wherein someone or something is unlike anything else in comparison.

No.of Nulls in each Column: Often, data can have missing values due to a variety of reasons, for example with survey data, some observations may not have been recorded

Log Distribution of Views/Comments/Dislikes/likes: In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

Total number of records based on the category: To display the number of rows that are in the dataset based on the category

Code: ADA_EDA_GB.ipynb

ADA_EDA_GB.ipynb

Results:

Uniqueness of each Columns

```
In [3]: 1 print(df_YT_GB.nunique())
```

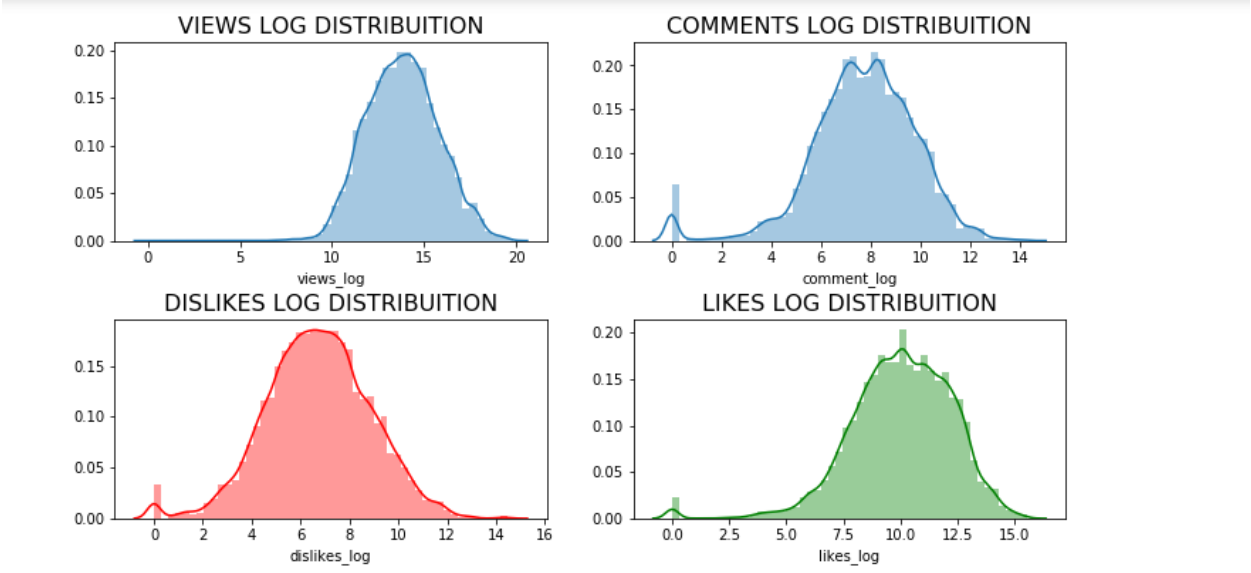
```
video_id                 3238
trending_date             207
title                    3371
channel_title            1626
category_id                17
publish_time             3252
tags                     3124
views                   38397
likes                   30558
dislikes                11093
comment_count           15779
thumbnail_link           3274
comments_disabled           6
ratings_disabled            6
video_error_or_removed      6
description              3706
dtype: int64
```

No.of Nulls in each Column:

```
In [4]:
1  # No of Nulls
2  df_YT_GB.info()
```

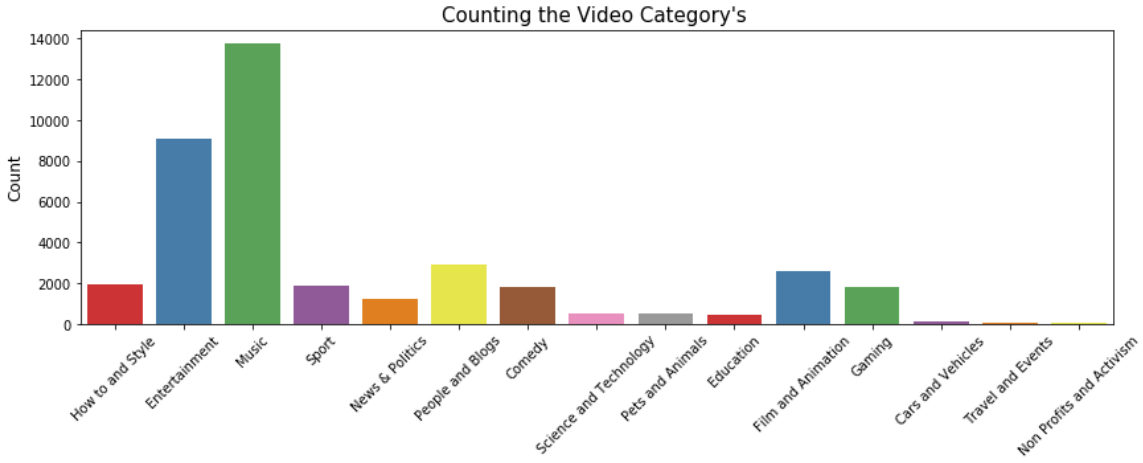
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38922 entries, 0 to 38921
Data columns (total 16 columns):
video_id                  38922 non-null object
trending_date             38922 non-null object
title                     38922 non-null object
channel_title             38922 non-null object
category_id               38922 non-null int64
publish_time              38922 non-null object
tags                      38922 non-null object
views                     38922 non-null float64
likes                     38922 non-null float64
dislikes                  38922 non-null float64
comment_count             38922 non-null float64
thumbnail_link            38922 non-null object
comments_disabled         38922 non-null object
ratings_disabled          38922 non-null object
video_error_or_removed    38922 non-null object
description               38310 non-null object
dtypes: float64(4), int64(1), object(11)
memory usage: 3.1+ MB
```

Log Distribution of Views/Comments/Dislikes/likes:



Total number of records based on the category:

```
Category Name count
Music                13755
Entertainment         9125
People and Blogs      2926
Film and Animation    2578
How to and Style      1928
Name: category_name, dtype: int64
```



Conclusion:

In general, life is sub-optimal and so are the data sets we work with. Therefore, to add that during EDA, one detects all kinds of singularities/anomalies/outliers/etc. Do not ignore these signals. Sometime these things tell and reveal super important insights and provide deeper understanding of the data sets' nature.

Germany Britain Exploratory Data Analysis

Uniqueness of each Columns: Uniqueness is a state or condition wherein someone or something is unlike anything else in comparison.

No.of Nulls in each Column: Often, data can have missing values due to a variety of reasons, for example with survey data, some observations may not have been recorded

Log Distribution of Views/Comments/Dislikes/likes: In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

Total number of records based on the category: To display the number of rows that are in the dataset based on the category

Code: ADA_EDA_DE.ipynb

ADA_EDA_DE.ipynb

Results:

Uniqueness of each Columns

```
In [3]: 1 print(df_YT_DE.nunique())
```

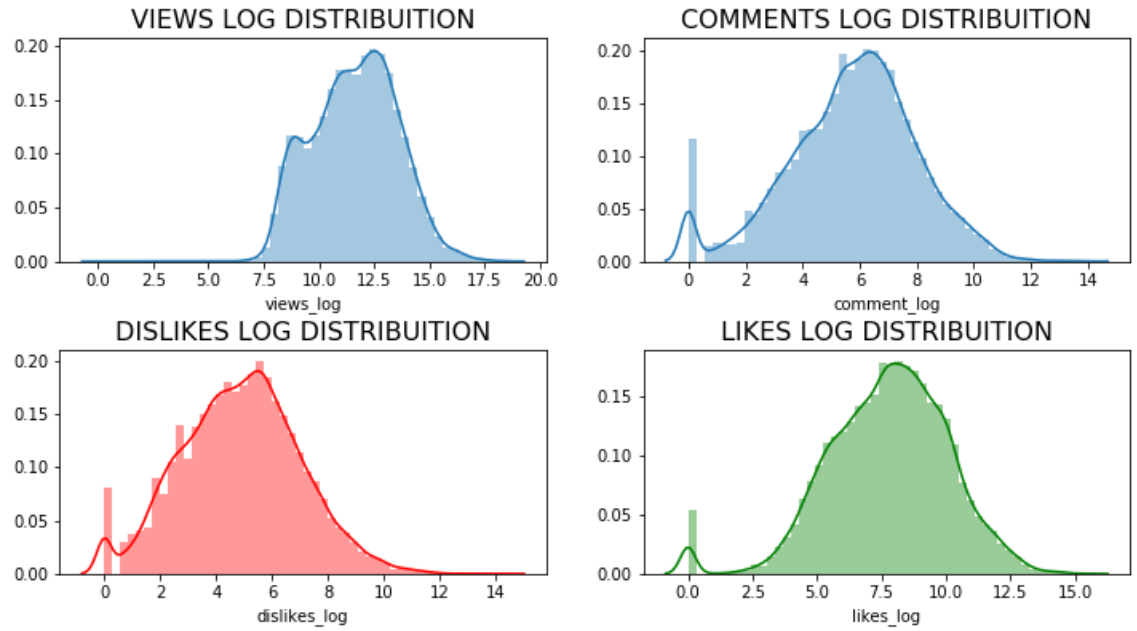
```
video_id                  29241
trending_date               207
title                     29681
channel_title              6086
category_id                  19
publish_time              28263
tags                      23559
views                     37919
likes                     17793
dislikes                   5108
comment_count              7579
thumbnail_link            29628
comments_disabled             6
ratings_disabled              6
video_error_or_removed        6
description               25615
dtype: int64
```

No.of Nulls in each Column:

```
In [4]: 1 # No of Nulls
        2 df_YT_DE.info()
```

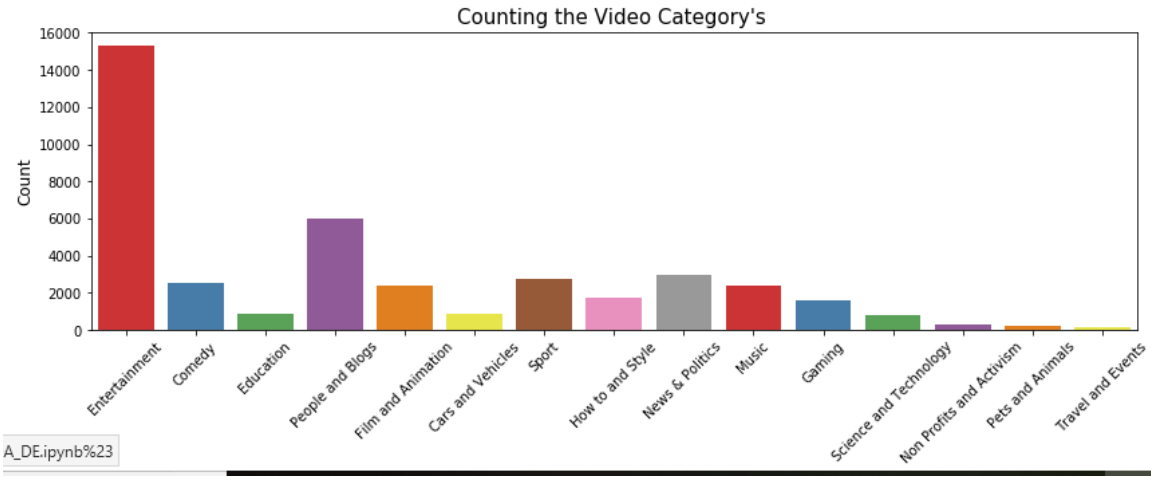
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40846 entries, 0 to 40845
Data columns (total 16 columns):
video_id                  40846 non-null object
trending_date             40846 non-null object
title                     40846 non-null object
channel_title             40846 non-null object
category_id               40846 non-null int64
publish_time              40846 non-null object
tags                      40846 non-null object
views                     40846 non-null float64
likes                     40846 non-null int64
dislikes                  40846 non-null int64
comment_count             40846 non-null int64
thumbnail_link            40846 non-null object
comments_disabled         40846 non-null object
ratings_disabled          40846 non-null object
video_error_or_removed    40846 non-null object
description               39294 non-null object
dtypes: float64(1), int64(4), object(11)
memory usage: 3.3+ MB
```

Log Distribution of Views/Comments/Dislikes/likes:



Total number of records based on the category:

```
Category Name count
Entertainment       15293
People and Blogs     5988
News & Politics      2935
Sport                2752
Comedy               2534
Name: category_name, dtype: int64
```



A_DE.ipynb%23

Conclusion:

In general, life is sub-optimal and so are the data sets we work with. Therefore, to add that during EDA, one detects all kinds of singularities/anomalies/outliers/etc. Do not ignore these signals. Sometime these things tell and reveal super important insights and provide deeper understanding of the data sets' nature.

Clustering and Segmentation Analysis

Description: In today's competitive world, it is crucial to understand customer behavior and categorize customers based on their demography and buying behavior. This is a critical aspect of segmentation that allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies.

USA Clustering and Segmentation Analysis

Frequency of YouTube data based on the category: A frequency distribution is an overview of all distinct values in some variable and the number of times they occur. That is, a frequency distribution tells how frequencies are distributed over values. Frequency distributions are mostly used for summarizing categorical variables.

Analyzing No.of Clusters before normalizing: Normalization usually improves the results, but if your question is if normalization ALWAYS improve the clustering results, the answer is NOT. Normalization could even degrade performance. Think, for instance, in an array of spherical clusters of points with centers along a line. Normalization would transform spherical clusters into elliptical ones, which could be problematic for the clustering algorithm.

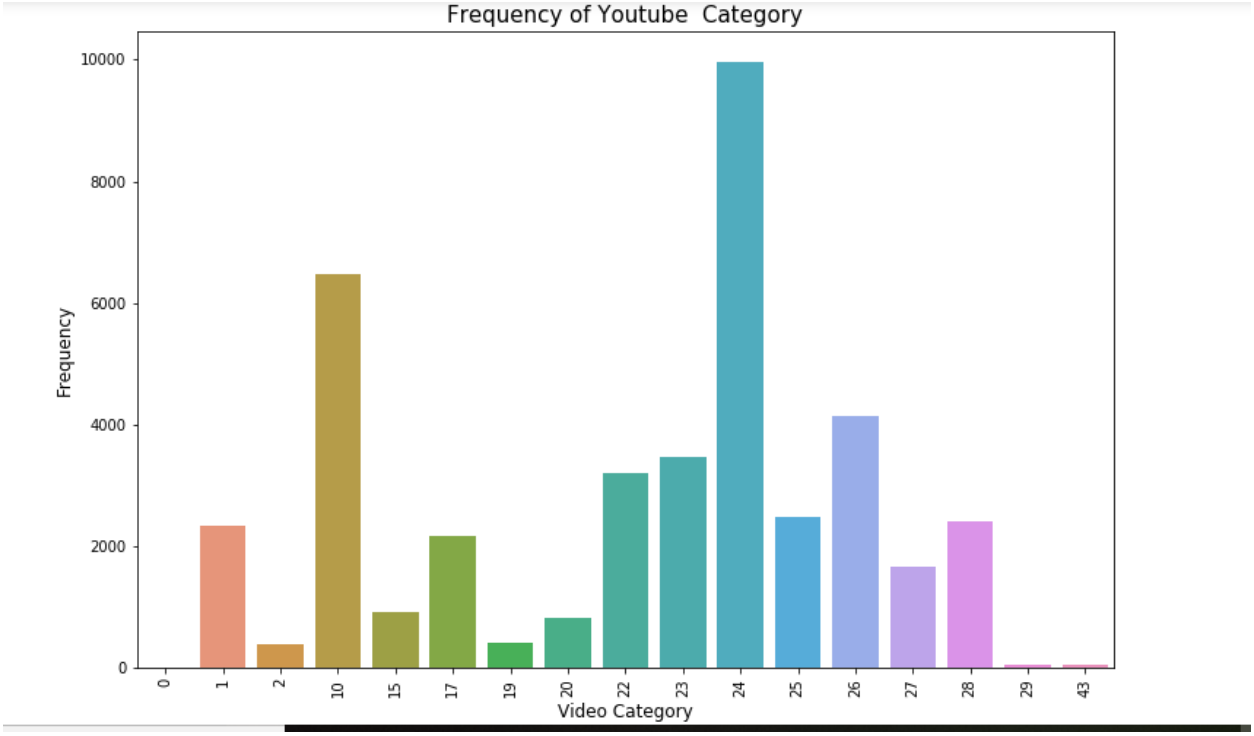
Analyzing no.of Clusters after normalizing: Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So, it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

Likes/Comments/Views in a 3D graph to show data distribution: In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points

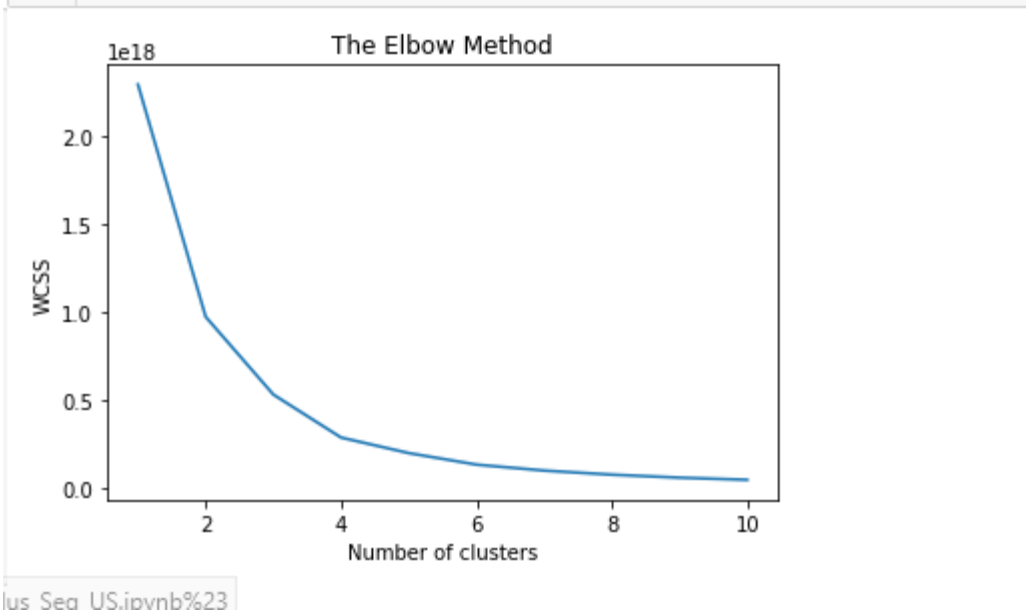
Code: ADA_Clus_Sef_US.ipynb

ADA_Clus_Seg_US.ipynb

Frequency of YouTube data based on the category:

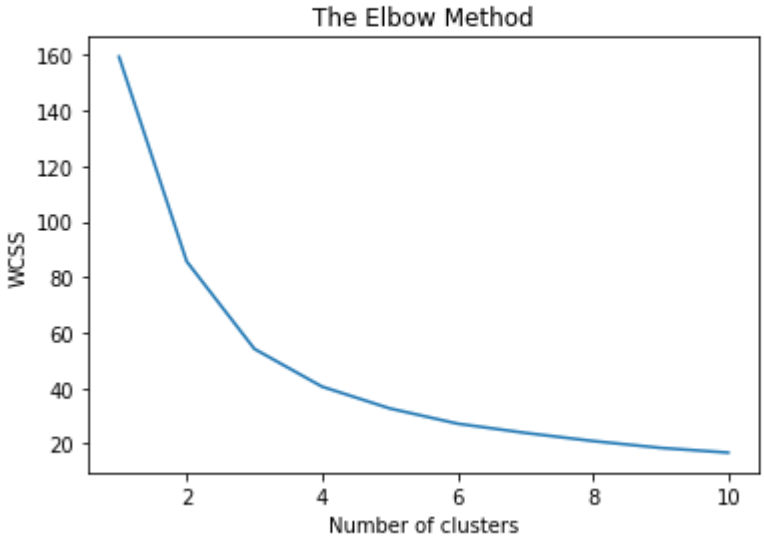


Analyzing No.of Clusters before normalizing:

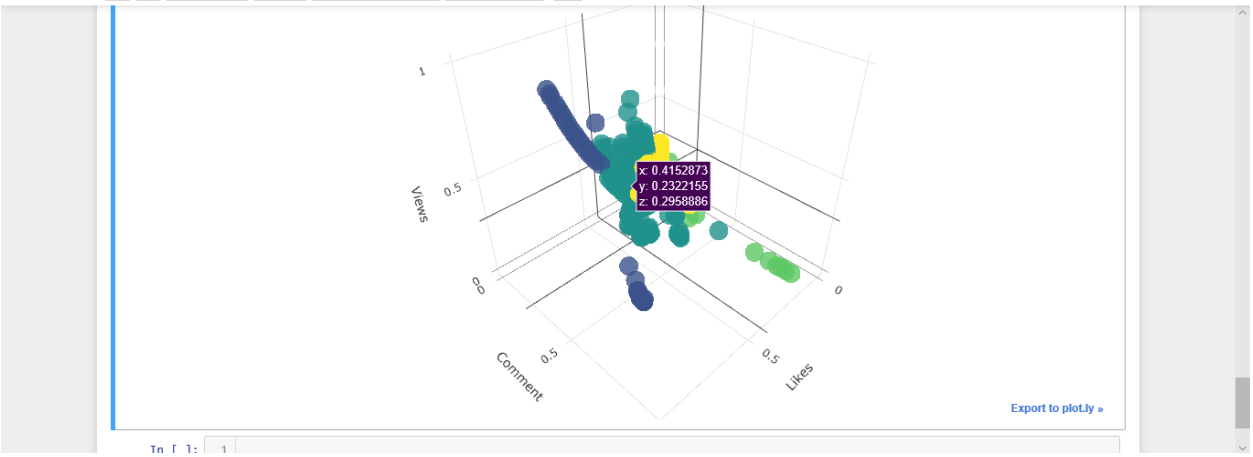


Analyzing no.of Clusters after normalizing

```
11 plt.ylabel('WCSS')
12 plt.show()
13
```



Likes/Comments/Views in a 3D graph to show data distribution. No.of Clusters taken = 5



Conclusion: Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments. To this purpose, one needs to spend time visualizing and understanding the data within each of the selected segments.

Canada Clustering and Segmentation Analysis

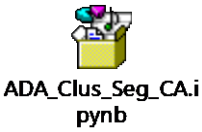
Frequency of YouTube data based on the category: A frequency distribution is an overview of all distinct values in some variable and the number of times they occur. That is, a frequency distribution tells how frequencies are distributed over values. Frequency distributions are mostly used for summarizing categorical variables.

Analyzing No.of Clusters before normalizing: Normalization usually improves the results, but if your question is if normalization ALWAYS improve the clustering results, the answer is NOT. Normalization could even degrade performance. Think, for instance, in an array of spherical clusters of points with centers along a line. Normalization would transform spherical clusters into elliptical ones, which could be problematic for the clustering algorithm.

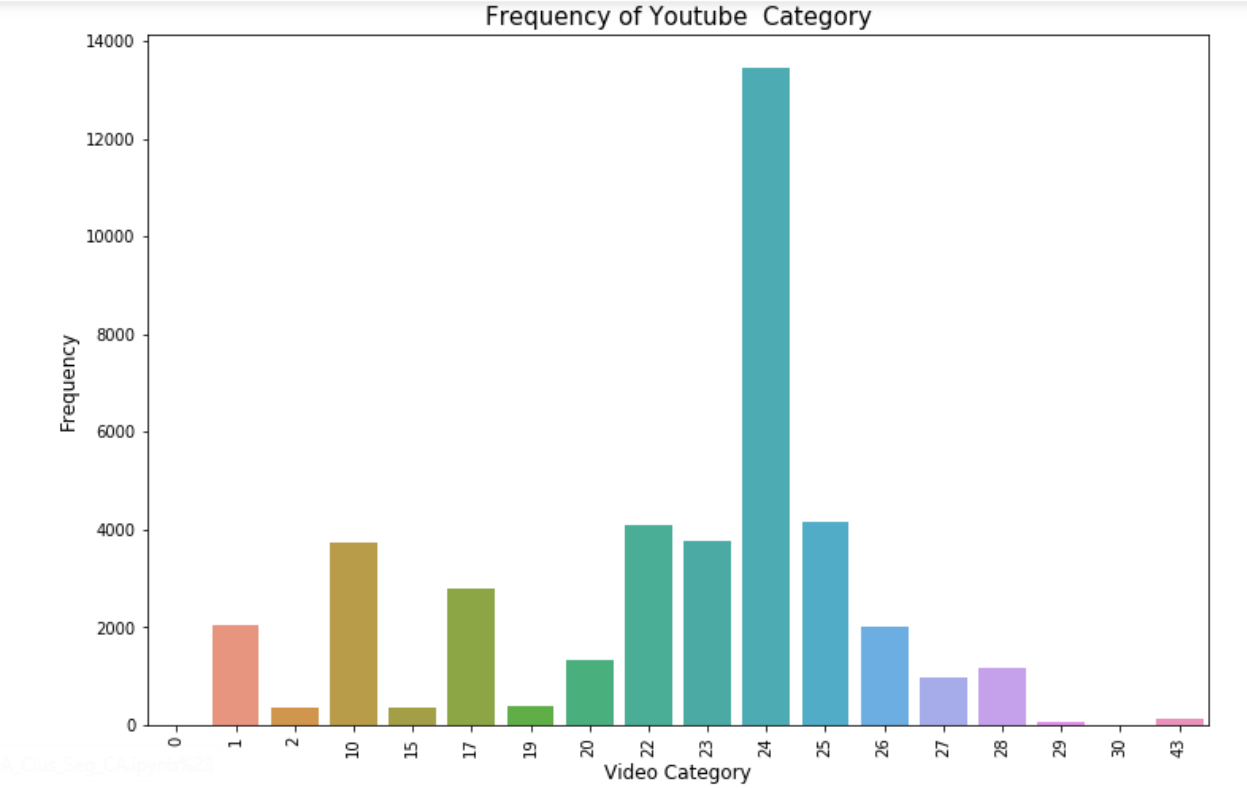
Analyzing no.of Clusters after normalizing: Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So, it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

Likes/Comments/Views in a 3D graph to show data distribution: In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points

Code: ADA_Clus_Sef_CA.ipynb

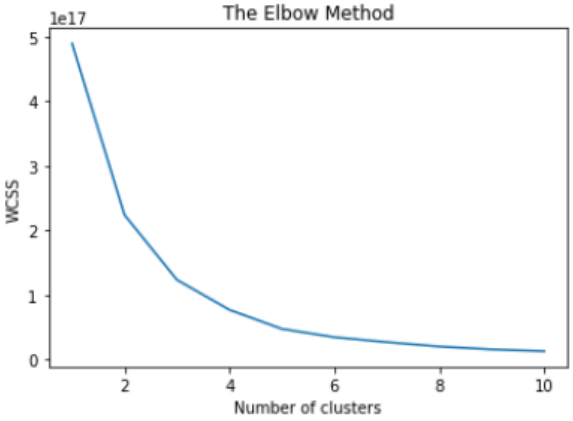


Frequency of YouTube data based on the category:

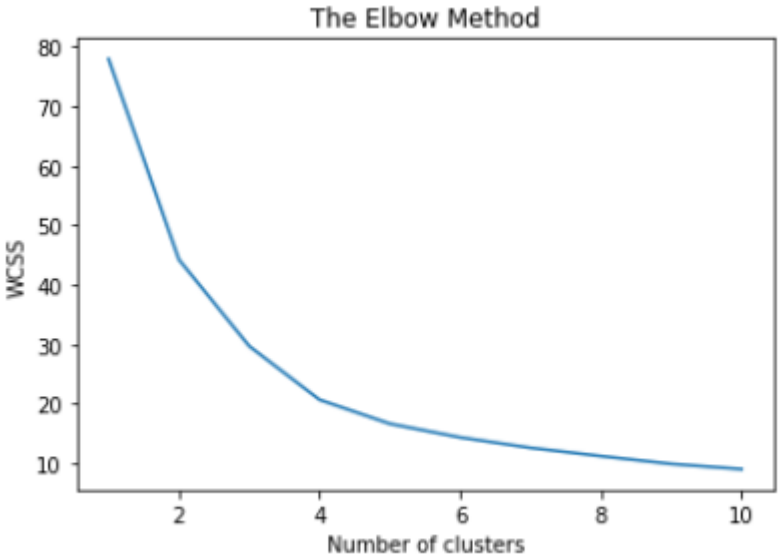


Analyzing No.of Clusters before normalizing:

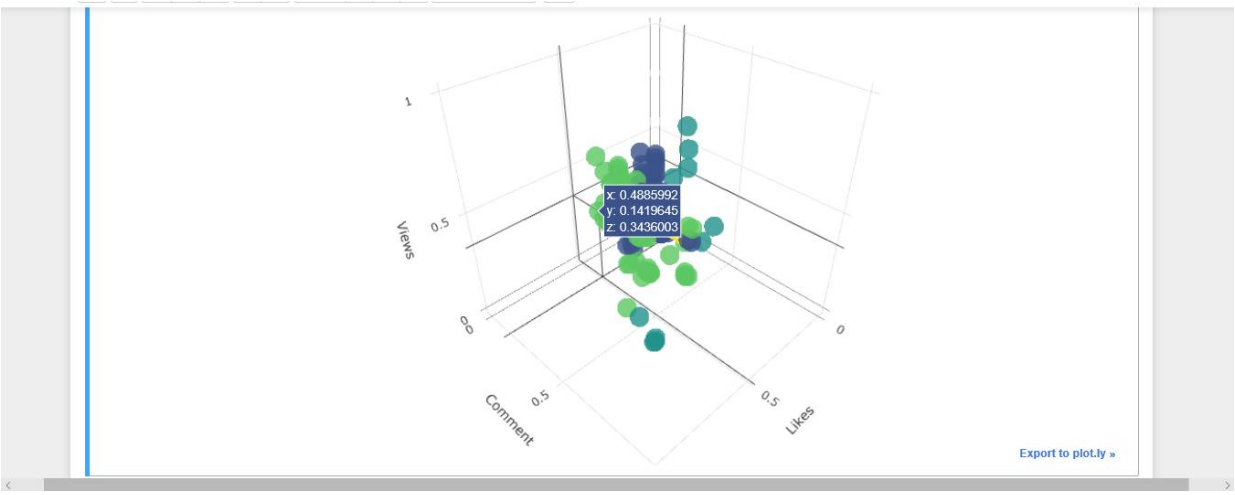
```
from sklearn.cluster import KMeans
wcss = []
for i in range(1,11):
    kmeans = KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10,random_state=0)
    kmeans.fit(cluster)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,11),wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



Analyzing no.of Clusters after normalizing



Likes/Comments/Views in a 3D graph to show data distribution. No.of Clusters taken = 5



Conclusion: Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments. To this purpose, one needs to spend time visualizing and understanding the data within each of the selected segments.

France Clustering and Segmentation Analysis

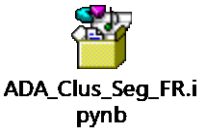
Frequency of YouTube data based on the category: A frequency distribution is an overview of all distinct values in some variable and the number of times they occur. That is, a frequency distribution tells how frequencies are distributed over values. Frequency distributions are mostly used for summarizing categorical variables.

Analyzing No.of Clusters before normalizing: Normalization usually improves the results, but if your question is if normalization ALWAYS improve the clustering results, the answer is NOT. Normalization could even degrade performance. Think, for instance, in an array of spherical clusters of points with centers along a line. Normalization would transform spherical clusters into elliptical ones, which could be problematic for the clustering algorithm.

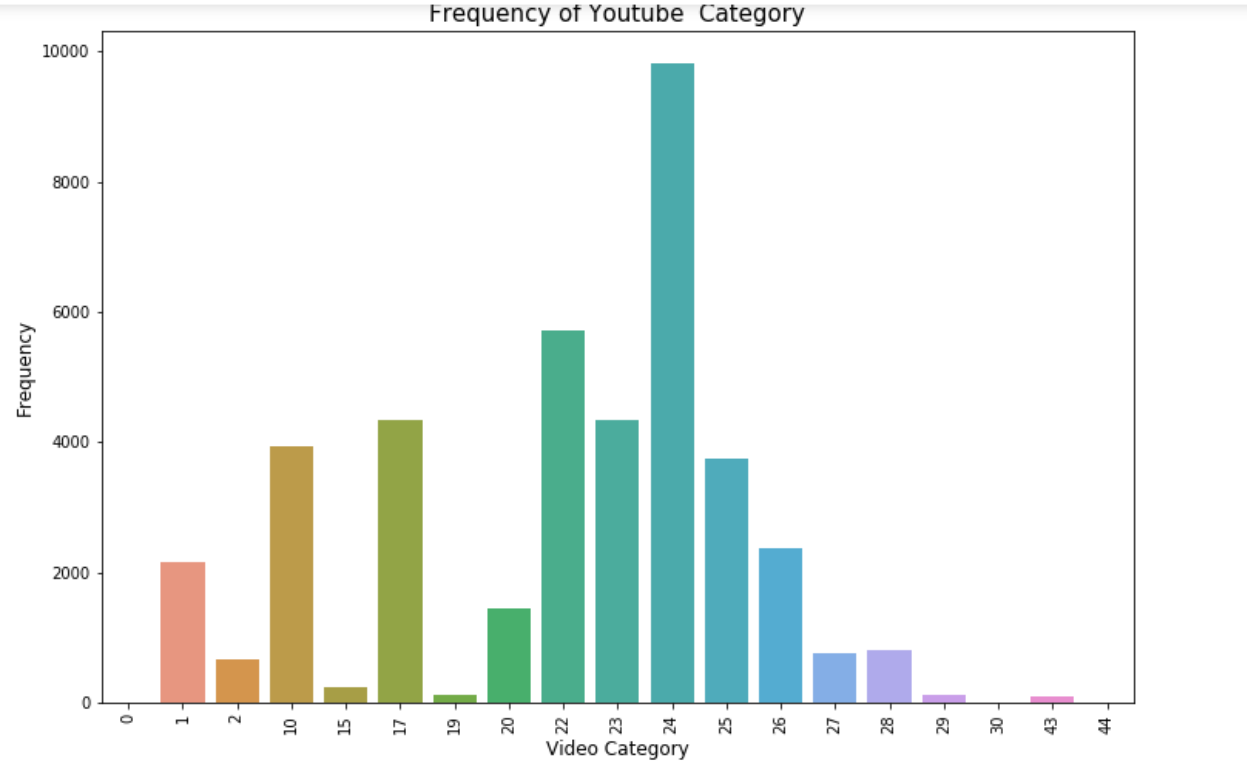
Analyzing no.of Clusters after normalizing: Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So, it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

Likes/Comments/Views in a 3D graph to show data distribution: In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points

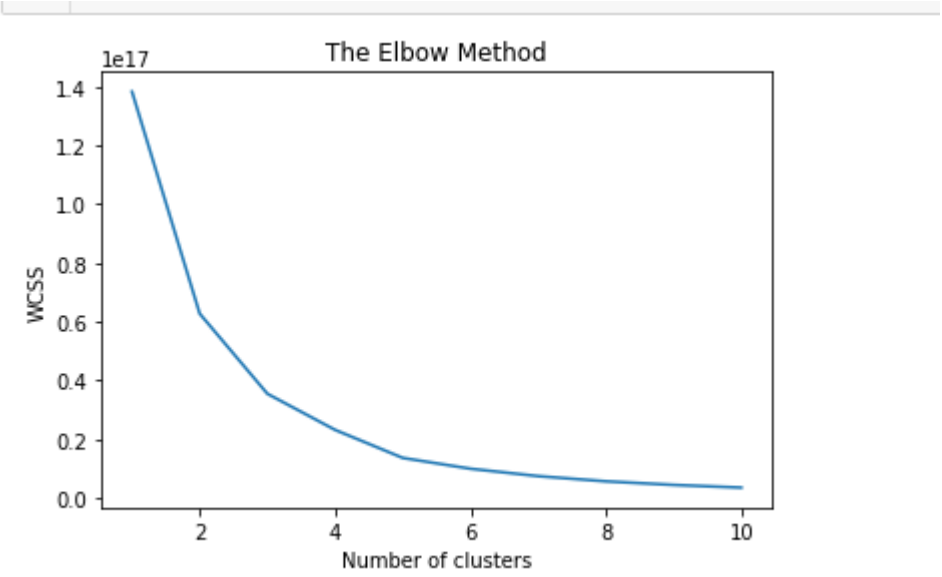
Code: ADA_Clus_Sef_FR.ipynb



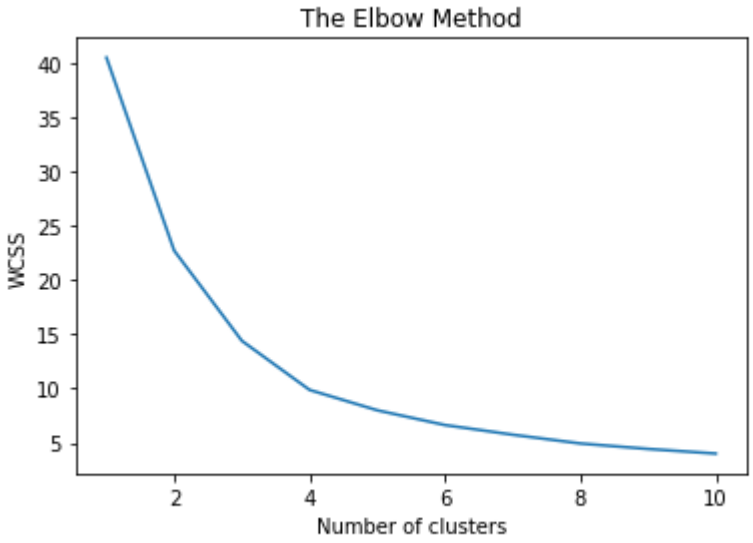
Frequency of YouTube data based on the category:



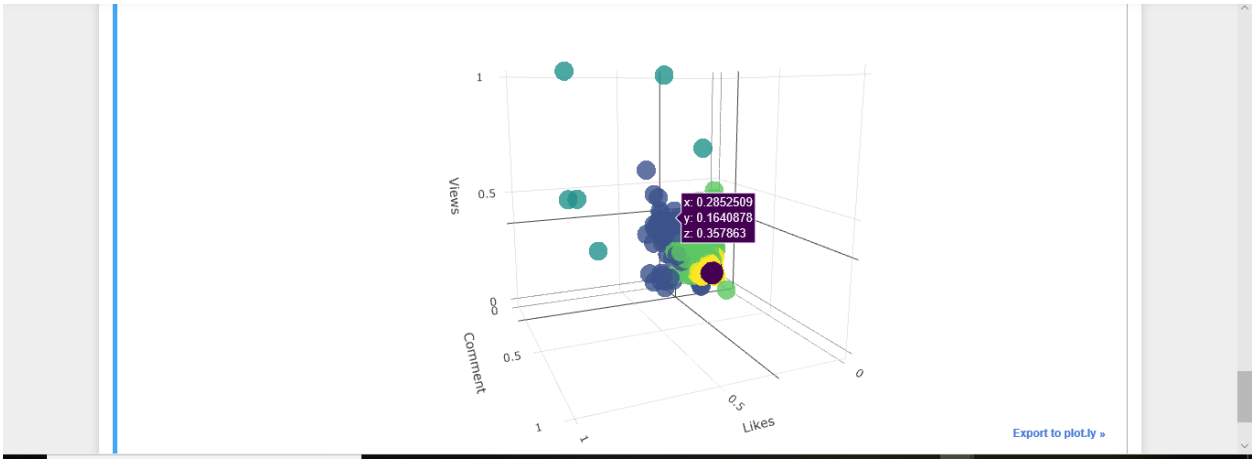
Analyzing No.of Clusters before normalizing:



Analyzing no.of Clusters after normalizing



Likes/Comments/Views in a 3D graph to show data distribution. No.of Clusters taken = 5



Conclusion: Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments. To this purpose, one needs to spend time visualizing and understanding the data within each of the selected segments.

Great Britain Clustering and Segmentation Analysis

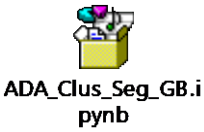
Frequency of YouTube data based on the category: A frequency distribution is an overview of all distinct values in some variable and the number of times they occur. That is, a frequency distribution tells how frequencies are distributed over values. Frequency distributions are mostly used for summarizing categorical variables.

Analyzing No.of Clusters before normalizing: Normalization usually improves the results, but if your question is if normalization ALWAYS improve the clustering results, the answer is NOT. Normalization could even degrade performance. Think, for instance, in an array of spherical clusters of points with centers along a line. Normalization would transform spherical clusters into elliptical ones, which could be problematic for the clustering algorithm.

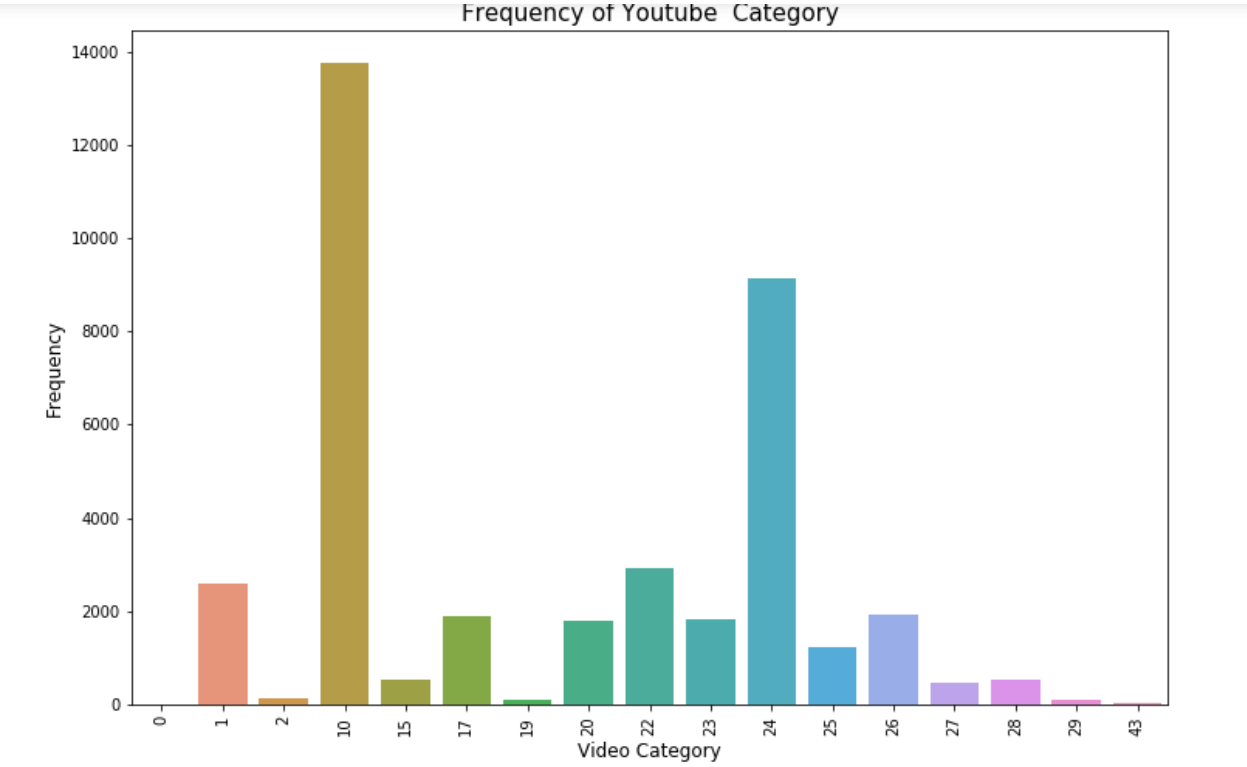
Analyzing no.of Clusters after normalizing: Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So, it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

Likes/Comments/Views in a 3D graph to show data distribution: In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points

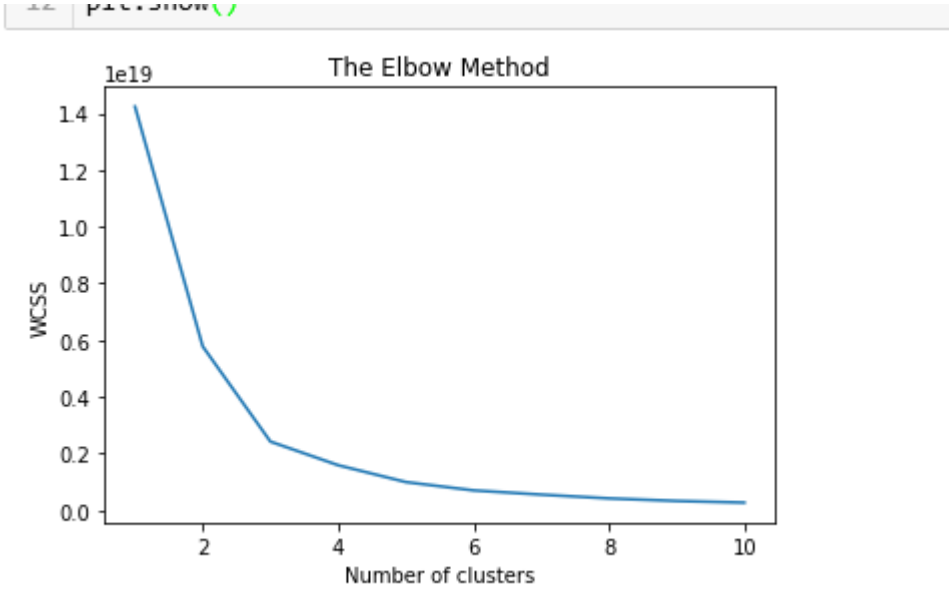
Code: ADA_Clus_Sef_GB.ipynb



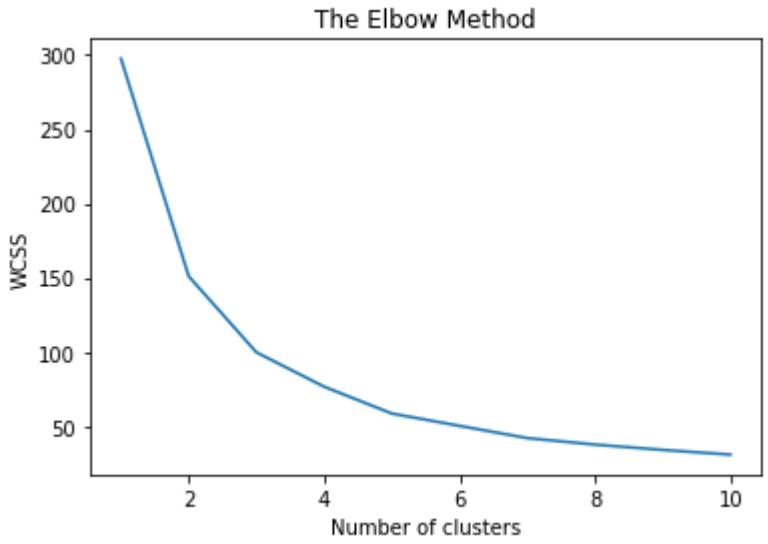
Frequency of YouTube data based on the category:



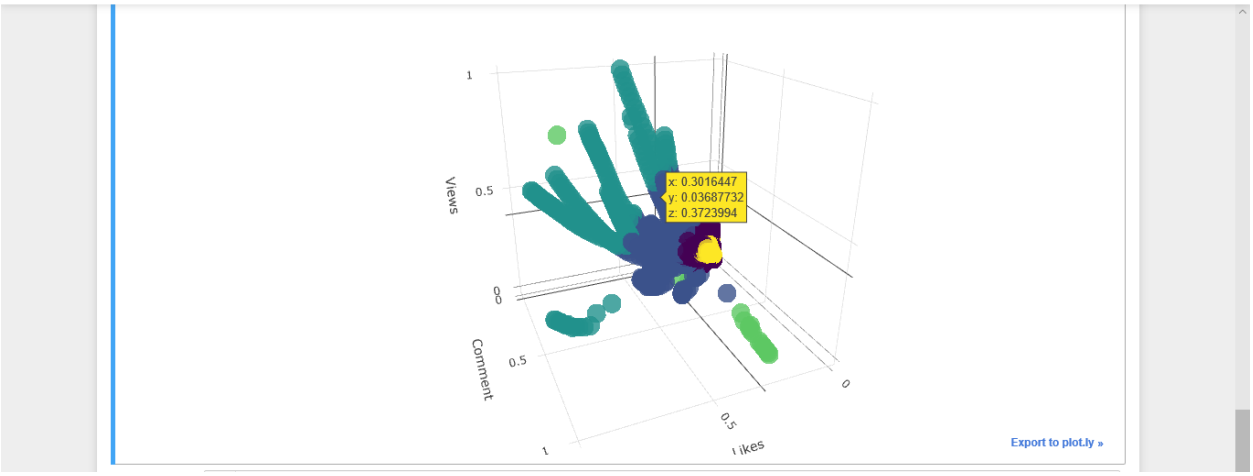
Analyzing No.of Clusters before normalizing:



Analyzing no.of Clusters after normalizing



Likes/Comments/Views in a 3D graph to show data distribution. No.of Clusters taken = 5



Conclusion: Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments. To this purpose, one needs to spend time visualizing and understanding the data within each of the selected segments.

Germany Clustering and Segmentation Analysis

Frequency of YouTube data based on the category: A frequency distribution is an overview of all distinct values in some variable and the number of times they occur. That is, a frequency distribution tells how frequencies are distributed over values. Frequency distributions are mostly used for summarizing categorical variables.

Analyzing No.of Clusters before normalizing: Normalization usually improves the results, but if your question is if normalization ALWAYS improve the clustering results, the answer is NOT. Normalization could even degrade performance. Think, for instance, in an array of spherical clusters of points with centers along a line. Normalization would transform spherical clusters into elliptical ones, which could be problematic for the clustering algorithm.

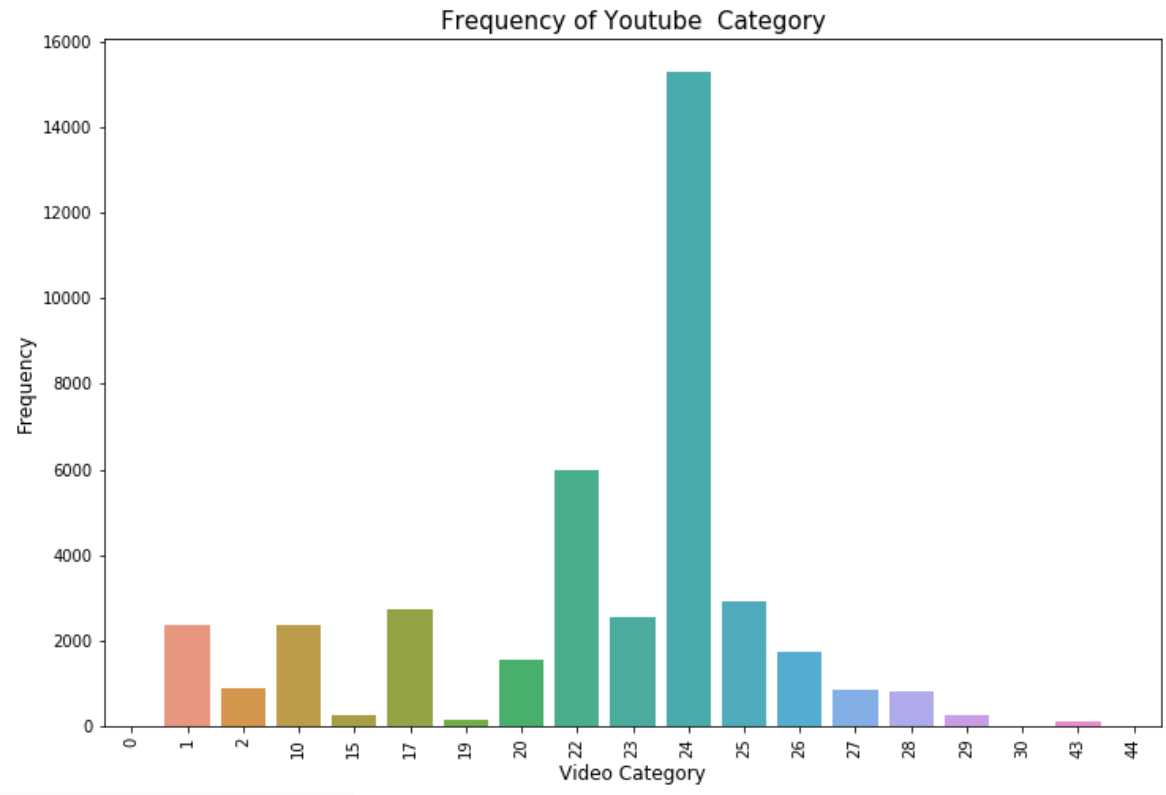
Analyzing no.of Clusters after normalizing: Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms. So, it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

Likes/Comments/Views in a 3D graph to show data distribution: In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points

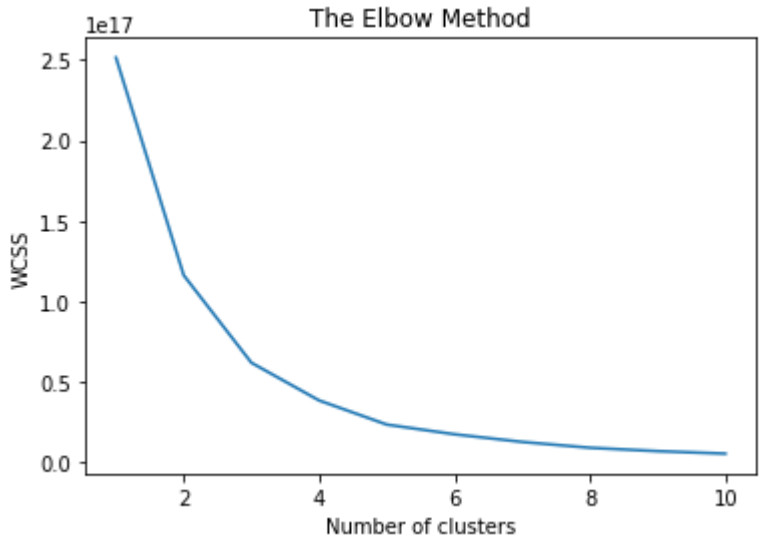
Code: ADA_Clus_Sef_DE.ipynb

ADA_Clus_Seg_DE.ipynb

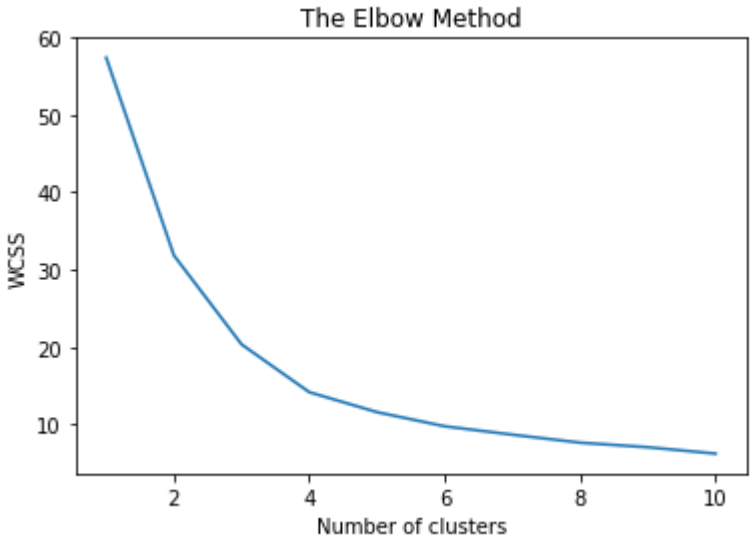
Frequency of YouTube data based on the category:



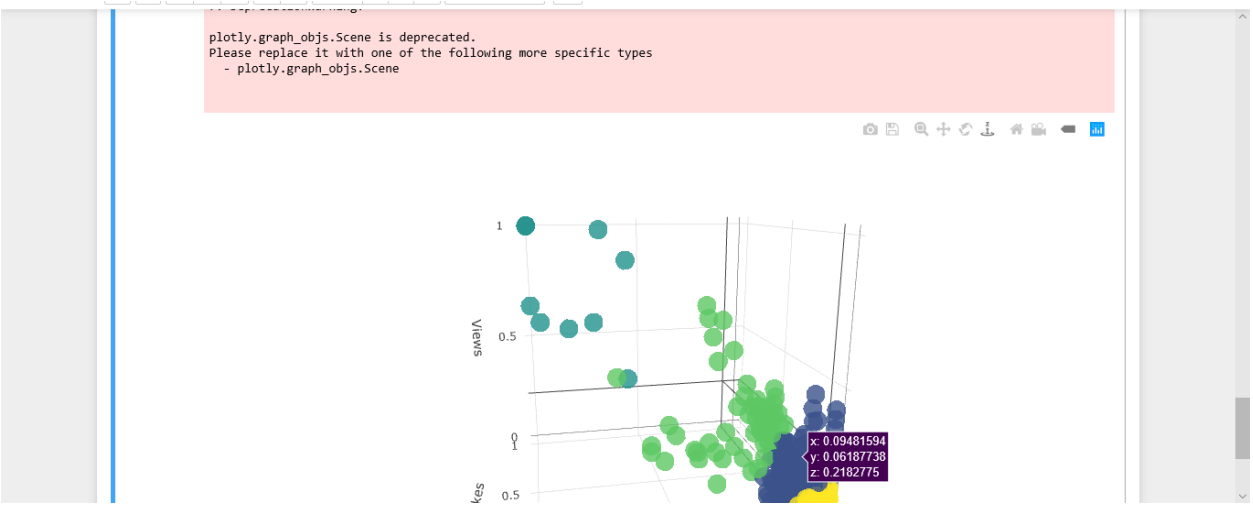
Analyzing No.of Clusters before normalizing:



Analyzing no.of Clusters after normalizing



Likes/Comments/Views in a 3D graph to show data distribution. No.of Clusters taken = 5



Conclusion: Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments. To this purpose, one needs to spend time visualizing and understanding the data within each of the selected segments.

Sentiment Analysis

Description: Opinion mining refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information

USA Sentiment Analysis

Positive Sentiment Analysis for Video Tags column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Video Tags column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

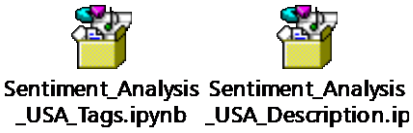
Distribution of Positive/Negative/Neutral based on the Video Tags column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention.

Positive Sentiment Analysis for Description column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Description column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color

Distribution of Positive/Negative/Neutral based on the Description column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention

Code: Sentiment_Analysis_USA_Description.ipynb/ Sentiment_Analysis_USA_Tags.ipynb



Positive Sentiment Analysis for Video Tags column:

In [3]:
```
# show imported records
df_positive = comm[comm.pol==1]
df_positive.head()
```

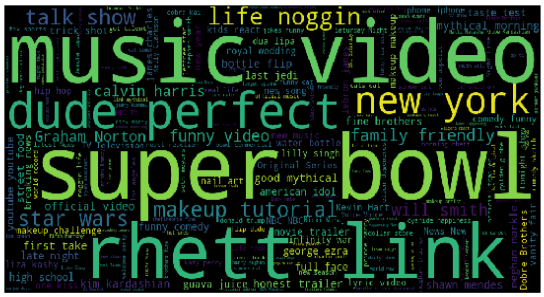
Out[3]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033	5339
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good mythical morning"|"...	343168	10172	666
6	39idVpFF7NQ	17.14.11	Roy Moore & Jeff Sessions Cold Open - SNL	Saturday Night Live	24	2017-11-12T05:37:17.000Z	SNL|"Saturday Night Live"|"SNL Season 43"|"Epi...	2103417	15993	2445
8	jr9QtXwC9vc	17.14.11	The Greatest Showman | Official Trailer 2 [HD]...	20th Century Fox	1	2017-11-13T14:00:23.000Z	Trailer|"Hugh Jackman"|"Michelle Williams"|"Za...	826059	3543	119
10	9wROjiENDW8	17.14.11	Dion Lewis' 103-Yd Kick	NFL	17	2017-11-	NFL|"Football"|"offense"|"defense"|"afc"|"nfc"...	81377	655	25

Word Cloud for Video Tags column:

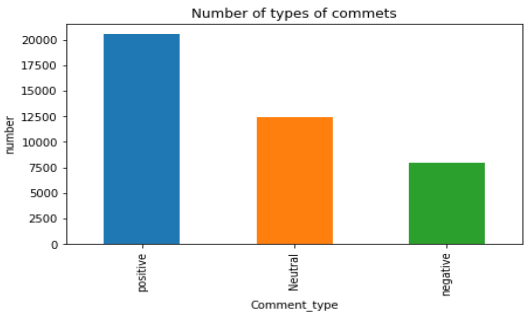
In [4]:
```
k= (' '.join(df_positive['tags']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[4]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [5]:
```
comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');

```



Positive Sentiment Analysis for Description column:

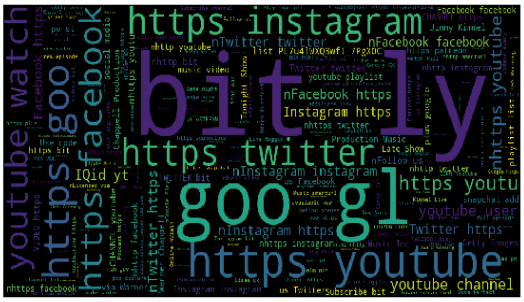
Out[2]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	2418783	97185	6146
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033	5339
4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095731	132235	1989
5	gHZ1Qz0KiKM	17.14.11	2 Weeks with iPhone X	iJustine	28	2017-11-13T19:07:23.000Z	ijustine|"week with iPhone X"|"iphone x"|"appl...	119180	9763	511
6	39idVpFF7NQ	17.14.11	Roy Moore & Jeff Sessions Cold Open - SNL	Saturday Night Live	24	2017-11-12T05:37:17.000Z	SNL|"Saturday Night Live"|"SNL Season 43"|"Epi...	2103417	15993	2445

Word Cloud for Video Tags column:

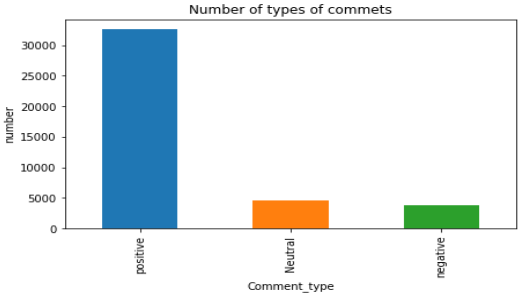
In [3]:
```
k= (' '.join(df_positive['description']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')

```

Out[3]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [7]:
```
comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');
```



Canada Sentiment Analysis

Positive Sentiment Analysis for Video Tags column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Video Tags column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

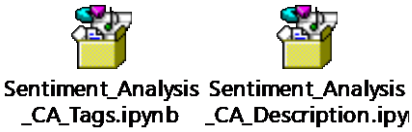
Distribution of Positive/Negative/Neutral based on the Video Tags column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention.

Positive Sentiment Analysis for Description column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Description column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color

Distribution of Positive/Negative/Neutral based on the Description column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention

Code: Sentiment_Analysis_CA_Description.ipynb/ Sentiment_Analysis_CA_Tags.ipynb



Positive Sentiment Analysis for Video Tags column:

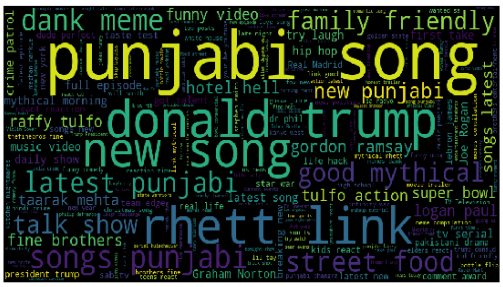
Out[2]:

	Unnamed: 0	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes
1	NaN	0dBIkQ4Mz1M	17.14.11	PLUSH - Bad Unboxing Fan Mail	iDubbbzTV	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	1014651	127794
2	NaN	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146035
4	NaN	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	33523622	1634130
6	NaN	_uM5kFfkhB8	17.14.11	Vanoss Superhero School - New Students	VanossGaming	23	2017-11-12T23:52:13.000Z	Funny Moments|"Montage video games"|"gaming"|"...	2987945	187464
9	NaN	43sm-Owl-cy4	17.14.11	Finally Sheldon is winning	Sheikh Musa	22	2017-11-	God|"Sheldon Cooper"|"Young Sheldon"	505161	4135

Word Cloud for Video Tags column:

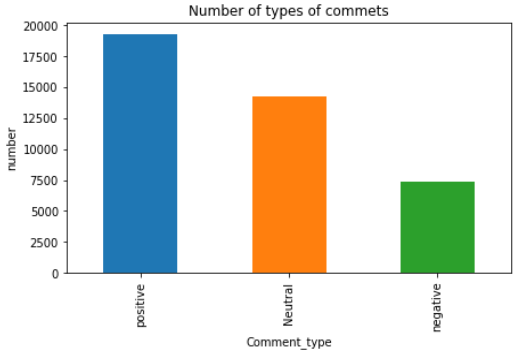
In [3]:
```
k= (' '.join(df_positive['tags']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')

```

Out[3]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [4]:
```
comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');

```



Positive Sentiment Analysis for Description column:

```
from ipykernel import kernelapp as app
```

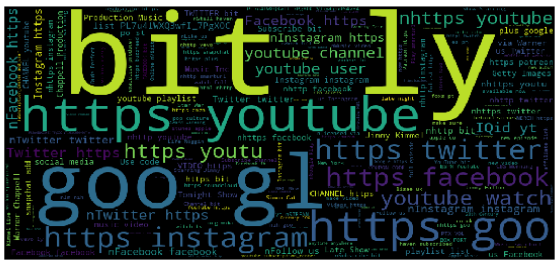
Out[2]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	com
4	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033	5339	
5	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good mythical morning"|"...	343168	10172	666	
6	39idVpFF7NQ	17.14.11	Roy Moore & Jeff Sessions Cold Open - SNL	Saturday Night Live	24	2017-11-12T05:37:17.000Z	SNL|"Saturday Night Live"|"SNL Season 43"|"Epi...	2103417	15993	2445	
8	jr9QtXwC9vc	17.14.11	The Greatest Showman | Official Trailer 2 [HD]...	20th Century Fox	1	2017-11-13T14:00:23.000Z	Trailer|"Hugh Jackman"|"Michelle Williams"|"Za...	826059	3543	119	
9			Dion								

Word Cloud for Video Tags column:

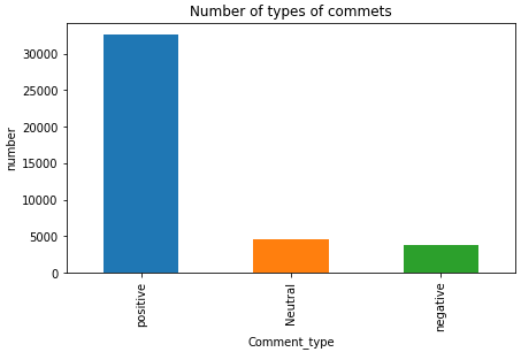
In [9]:
```
### Show description
k= (' '.join(df_positive['description']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[9]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [10]:
```
comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');

```



France Sentiment Analysis

Positive Sentiment Analysis for Video Tags column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Video Tags column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

Distribution of Positive/Negative/Neutral based on the Video Tags column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention.

Positive Sentiment Analysis for Description column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Description column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color

Distribution of Positive/Negative/Neutral based on the Description column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention

Code: Sentiment_Analysis_FR_Description.ipynb/ Sentiment_Analysis_FR_Tags.ipynb

Positive Sentiment Analysis for Video Tags column:

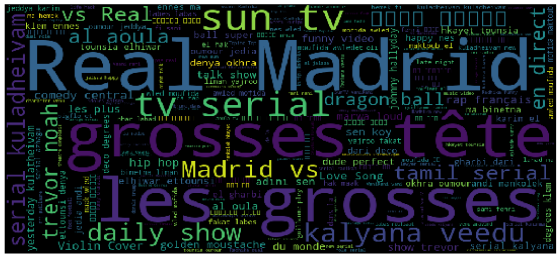
Out[2]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislike
5	teXaL6GdQRk	17.14.11	STRANGER JOKES : Jokes de Papa avec les teens ...	Le Jeu, C'est Sérieux	23	2017-11-13T15:48:57.000Z	Stranger Jokes"|"Jokes de Papa"|"Stranger Thin...	141253.0	14354.0	2(
6	nduL7G_gJoY	17.14.11	De retour dans le Manoir hanté avec le Grand J...	silent jill	24	2017-11-12T19:00:08.000Z	fantome"|"esprits"|"spiritisme"|"hanté"|"ouija...	187654.0	9286.0	13(
8	GBVxEpQr8R8	17.14.11	ON VOUS DÉVOILE NOTRE VRAI SALAIRE	Mcfly & Carlito	24	2017-11-12T08:59:25.000Z	mcfly"|"carlito"|"golden moustache"|"fatshow"|...	2340941.0	200598.0	60
17	tsMw-VMUtNU	17.14.11	Kid Barely Avoids Getting Run Over by Trailer ...	Pirateay	24	2017-11-11T18:38:02.000Z	near accident"|"safety"|"danger"|"volvo brakes...	79611.0	56.0	

Word Cloud for Video Tags column:

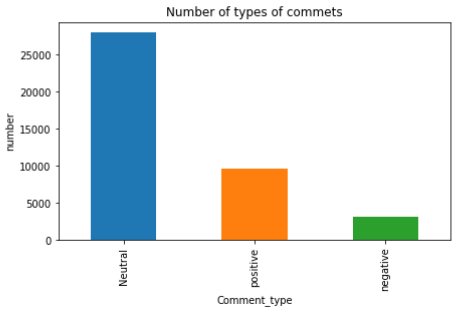
In [3]:
```
k= (' '.join(df_positive['tags']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[3]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [4]:
```

comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');
```



Positive Sentiment Analysis for Description column:

Out[2]:

likes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	video_error_or_removed	description	pol
576	1161	https://i.ytimg.com/vi/Yo84eqYwP98/default.jpg	False	False	False	Le jeu de société: https://goo.gl/hhG1Ta\n\nGa...	1.0
1381	2419	https://i.ytimg.com/vi/nduL7G_gJoY/default.jpg	False	False	False	Bonsoir à tous, \nJe tenais beaucoup à retourn...	1.0
6018	7575	https://i.ytimg.com/vi/GBVxEpQr8R8/default.jpg	False	False	False	Nouvelle vidéo tous les dimanches matins 10h. ...	1.0
0	0	https://i.ytimg.com/vi/LhjhGOBIoHM/default.jpg	True	True	False	Jérémy Ferrari - On n'est pas couché 11 novemb...	1.0

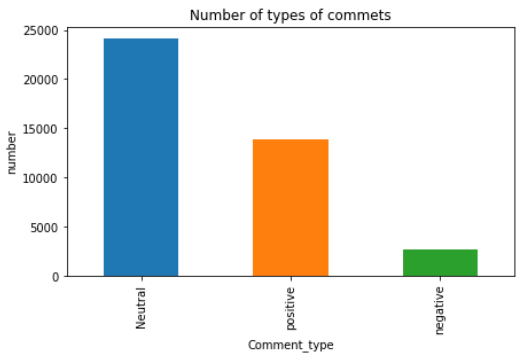
Word Cloud for Video Tags column:

In [3]:
```
k= (' '.join(df_positive['description']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[3]: (-0.5, 999.5, 499.5, -0.5)

Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [4]:
```
    comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');
```



Great Britain Sentiment Analysis

Positive Sentiment Analysis for Video Tags column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Video Tags column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

Distribution of Positive/Negative/Neutral based on the Video Tags column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention.

Positive Sentiment Analysis for Description column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Description column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color

Distribution of Positive/Negative/Neutral based on the Description column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention

Code: Sentiment_Analysis_GB_Description.ipynb/ Sentiment_Analysis_ GB _Tags.ipynb


Sentiment_Analysis_GB_Tags.ipynb


Sentiment_Analysis_GB_Description.ipy

Positive Sentiment Analysis for Video Tags column:

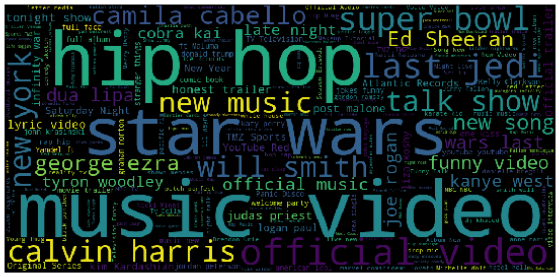
Out[3]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	
1	3s1rvMFUweQ	17.14.11	Taylor Swift: ... Ready for It? (Live) - SNL	Saturday Night Live	24	2017-11-12T06:24:44.000Z	SNL|"Saturday Night Live"|"SNL Season 43"|"Epi...	1053632.0	2
5	AumaWl0TNBo	17.14.11	How My Relationship Started!	PointlessBlogVlogs	24	2017-11-11T17:00:00.000Z	pointlessblog|"pointlessblogtv"|"pointlessblog...	1182775.0	5
6	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	33523622.0	163
8	LMCuKltaY3M	17.14.11	Elbow - Golden Slumbers (John Lewis Advert 2017)	ElbowVEVO	10	2017-11-10T08:00:01.000Z	Elbow|"Golden"|"Slumbers"|"Polydor"|"Alternative"	154494.0	
10	ONQ-fAp5X64	17.14.11	CAN BABIES DO GYMNASTICS? **World Record**	Nile Wilson	17	2017-11-11T10:30:00.000Z	nile wilson|"nile wilson gymnastics"|"nile wil...	306724.0	

Word Cloud for Video Tags column:

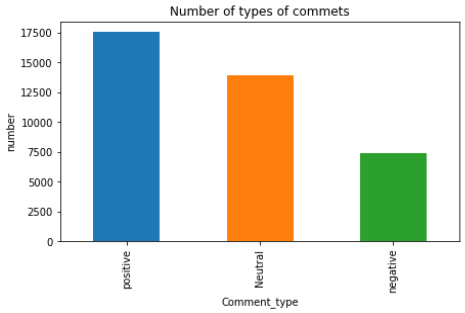
In [4]:
```
k= (' '.join(df_positive['tags']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[4]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [5]:
```
comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');
```



Positive Sentiment Analysis for Description column:

Out[2]:

/s	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	video_error_or_removed	description	pol
.0	55681.0	10247.0	9479.0	https://i.ytimg.com/vi/Jw1Y-zhQURU/default.jpg	False	False	False	Click here to continue the story and make your...	1.0
.0	25561.0	2294.0	2757.0	https://i.ytimg.com/vi/3s1rvMFUweQ/default.jpg	False	False	False	Musical guest Taylor Swift performs … Ready for...	1.0
.0	787420.0	43420.0	125882.0	https://i.ytimg.com/vi/n1WpP7iowLc/default.jpg	False	False	False	Eminem's new track Walk on Water ft. Beyoncé i...	1.0
.0	193.0	12.0	37.0	https://i.ytimg.com/vi/PUTEiSjKwJU/default.jpg	False	False	False	Salford drew 4-4 against the Class of 92 and F...	1.0
.0	30.0	2.0	30.0	https://i.ytimg.com/vi/rHwDegptbI4/default.jpg	False	False	False	Dashcam captures truck's near miss with	1.0

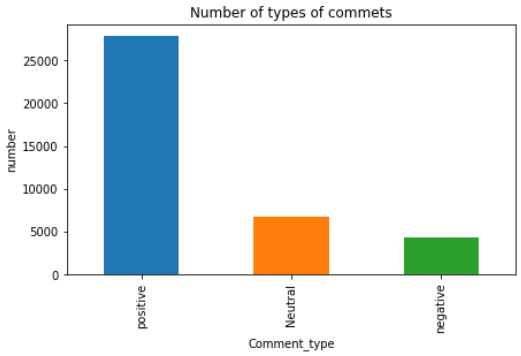
Word Cloud for Video Tags column:

In [3]:
```
### Show description
k= (' '.join(df_positive['description']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[3]: (-0.5, 999.5, 499.5, -0.5)

Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [4]:
```
comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');
```



Germany Sentiment Analysis

Positive Sentiment Analysis for Video Tags column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Video Tags column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

Distribution of Positive/Negative/Neutral based on the Video Tags column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention.

Positive Sentiment Analysis for Description column: Positive sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject.

Word Cloud for Description column: A tag cloud (word cloud, or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color

Distribution of Positive/Negative/Neutral based on the Description column: Negative, positive or neutral sentiment in general means the attitude or opinion one expressed within a given post towards a specific subject. It's based on algorithms evaluating whether the words included in a post are related to positive, negative or neutral emotions. Sentiment analysis is used by a majority of social media monitoring tools such as Social Mention

Code: Sentiment_Analysis_DE_Description.ipynb/ Sentiment_Analysis_DE_Tags.ipynb

Positive Sentiment Analysis for Video Tags column:

Out[2]:

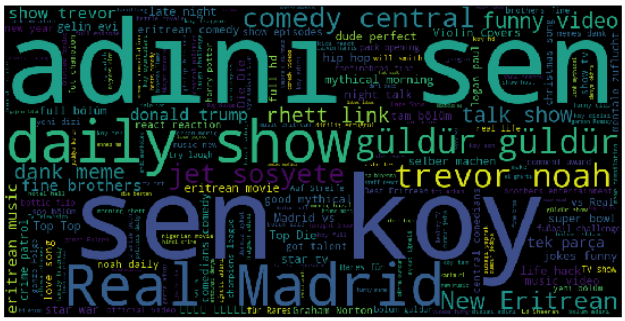
	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	like
8	GHct2dGNLks	17.14.11	Antoine leiht Auto von Ernährungsberaterin	TeddyComedy	23	2017-11-12T11:53:45.000Z	Antoine Auto|"Antoine Boot Camp"|"Antoine Ernä...	369173.0	1695
18	riV8xuBqUQ0	17.14.11	Duell der Giganten 2.0 | inscopelifestyle	InscopeLifestyle	22	2017-11-12T16:34:53.000Z	inscopelifestyle|"VLOG"|"CarVlog"|"lamborghini...	113961.0	889
21	KLxP8VxZjIk	17.14.11	Portuguese trawler nets 'prehistoric shark'	BJ Magazine	25	2017-11-11T13:09:16.000Z	magazine|"prehistoric shark"|"live fossil"|"sc...	91914.0	1
24	PK8lHszeXNk	17.14.11	WOW 😱 Die Beste Wasserdicht SmartWatch Nur für...	Cool Mobile	22	2017-11-12T19:27:27.000Z	smartwatch|"review"|"android"|"apple watch"|"m...	27356.0	70
27	hg0OwRhQpGE	17.14.11	10 Unglaubliche Entdeckungen in der Antarktis!	TopWelt	24	2017-11-11T17:00:08.000Z	topwelt|"top welt"|"unterhaltung"|"fakten"|"wi...	165276.0	408

In [ ]: 1

Word Cloud for Video Tags column:

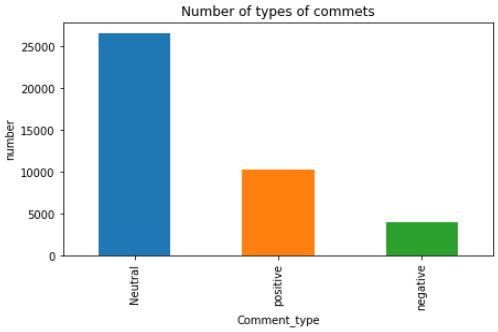
```
In [3]:
### Show word cloud
k= (' '.join(df_positive['tags']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

```
Out[3]: (-0.5, 999.5, 499.5, -0.5)
```



Distribution of Positive/Negative/Neutral based on the Video Tags column:

```
In [4]:

comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');

```



Positive Sentiment Analysis for Description column:

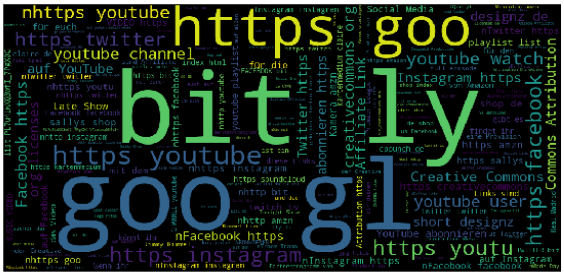
Out[2]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes
2	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	2418783.0	97190
5	xapGFgWqtg4	17.14.11	Geld verdienen mit Online-Umfragen? Geht das w...	Die Allestester	22	2017-11-13T01:49:24.000Z	[none]	32709.0	3093
8	GHct2dGNLks	17.14.11	Antoine leiht Auto von Ernährungsberaterin	TeddyComedy	23	2017-11-12T11:53:45.000Z	Antoine Auto|"Antoine Boot Camp"|"Antoine Ernä...	369173.0	16953
10	2hu_evXPpMM	17.14.11	Dagi Bee wird Heiraten | Coldmirror bekommt Eh...	HerrNewstime	24	2017-11-12T16:33:18.000Z	Dagi Bee|"Heiraten"|"Coldmirror"|"YouTube"|"Tr...	228574.0	11349
12	2Zp-Qm3wJkA	17.14.11	JP Performance - Quetschen wir den Japaner ma...	JP Performance	2	2017-11-13T10:30:01.000Z	V8|"VMAX"|"Topspeed"|"Prüfstand"|"JP Performan...	465883.0	19928

Word Cloud for Video Tags column:

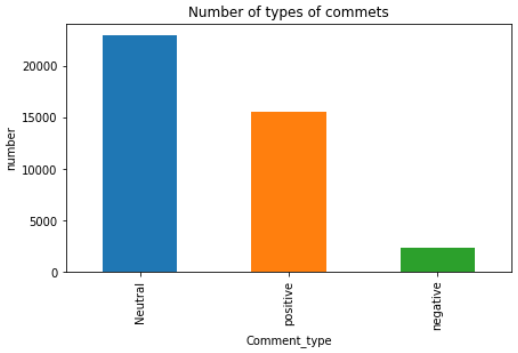
In [3]:
```
k= (' '.join(df_positive['description']))

wordcloud = WordCloud(width = 1000, height = 500).generate(k)
plt.figure(figsize=(15,5))
plt.imshow(wordcloud)
plt.axis('off')
```

Out[3]: (-0.5, 999.5, 499.5, -0.5)



Distribution of Positive/Negative/Neutral based on the Video Tags column:

In [4]:
```

comm['pol'].replace({1:'positive',0:'Neutral',-1:'negative'}).value_counts().plot(kind='bar',figsize=(7,4));
plt.title('Number of types of commets');
plt.xlabel('Comment_type');
plt.ylabel('number');
```



Time Series Analysis

Description: A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.

USA Time Series Analysis

Heat map of View/Likes/Dislikes/Comments count based on category: A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century

Histogram analysis of likes/dislikes/Views: Analyze the histogram to see whether it represents a normal distribution. Once you have plotted all the frequencies on the histogram, your histogram would show a shape. If the shape looks like a bell curve, it would mean that the frequencies are equally distributed. The histogram would have a peak

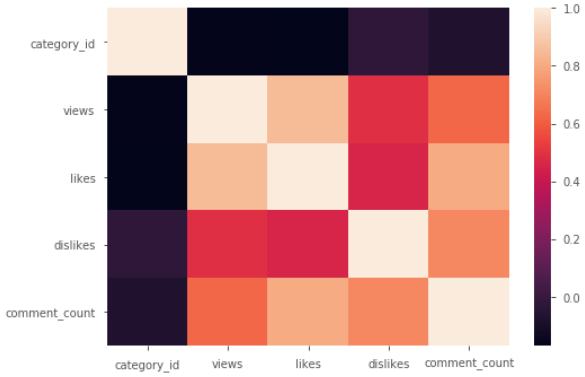
Time Series of US YouTube dataset based on the Category and Date [People and Blogs]: A series of values of a quantity obtained at successive times, often with equal intervals between them.

Code: TimeSeries_USA.ipynb

TimeSeries_USA.ipynb

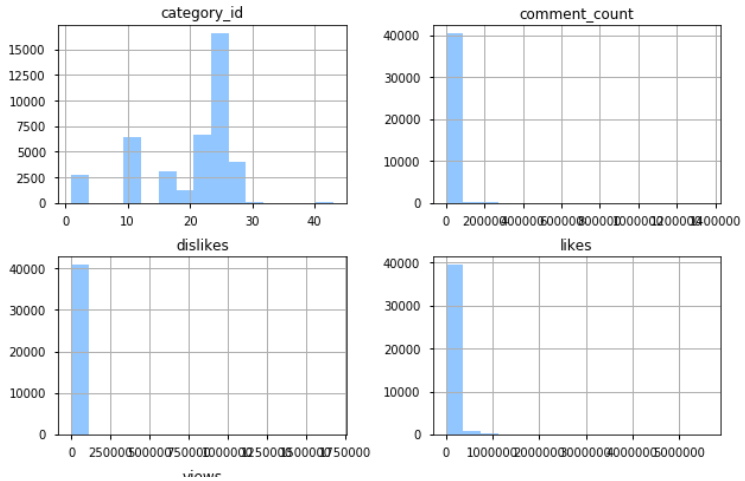
Heat map of View/Likes/Dislikes/Comments count based on category:

```
In [9]:
1
2 with plt.style.context(('ggplot')):
3     sns.heatmap(df.corr())
```



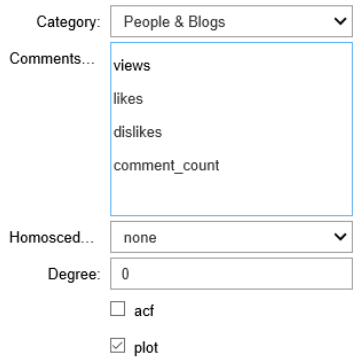
Histogram analysis of likes/dislikes/Views:

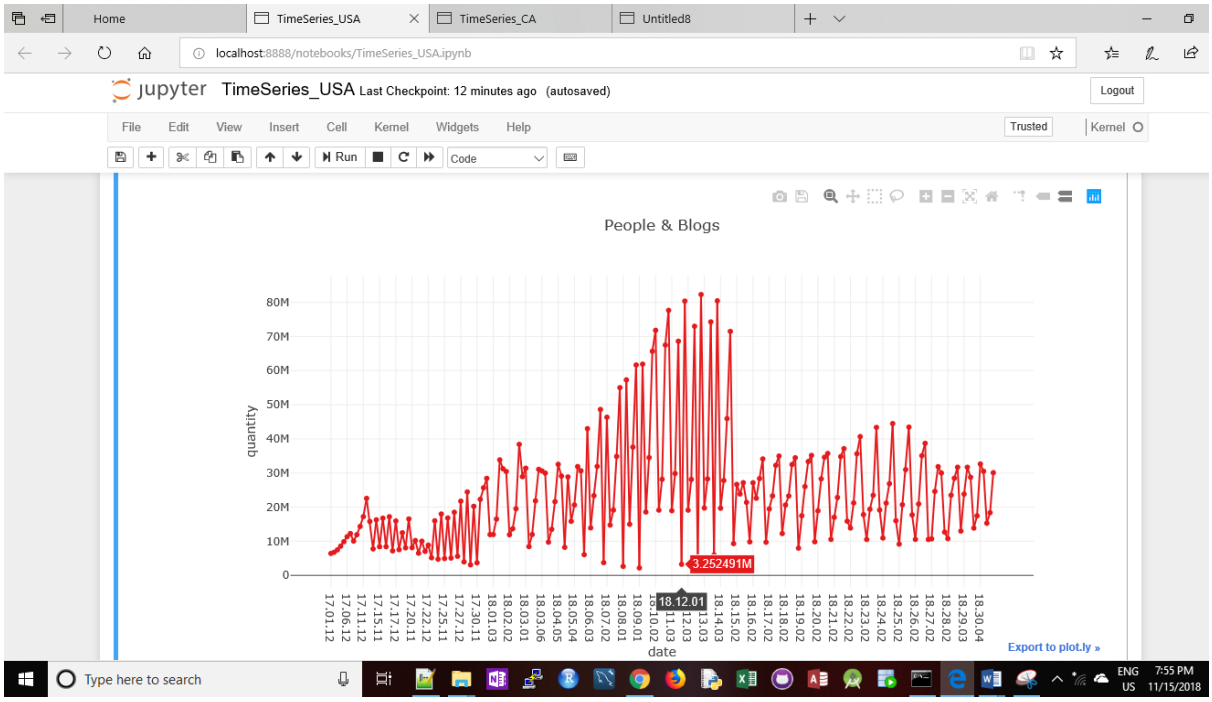
```
In [10]:
with plt.style.context(('seaborn-pastel')):

#Histogramas de las variables
    df.hist(figsize=(10,10), bins=15)


color=cl.scales["4"]["qual"]["Set1"]
```



Time Series of YouTube dataset based on the Category and Date:

```
    description='Degree:',
    disabled=False)


widgets.interactive(proyect_yt, columns=columns, category=category, tend=t,
                    Homoscedasticity=["none","inv", "sqrt","log", "Isqrt", "Ilog"])
```





Conclusion: Good to see that the videos have been linearly increasing over time and will be keep growing during the mid of the month for People and Blogs category.

Canada Time Series Analysis

Heat map of View/Likes/Dislikes/Comments count based on category: A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century

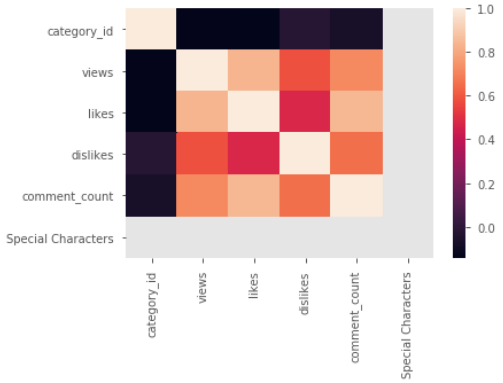
Histogram analysis of likes/dislikes/Views: Analyze the histogram to see whether it represents a normal distribution. Once you have plotted all the frequencies on the histogram, your histogram would show a shape. If the shape looks like a bell curve, it would mean that the frequencies are equally distributed. The histogram would have a peak

Time Series of US YouTube dataset based on the Category and Date[Entertainment]: A series of values of a quantity obtained at successive times, often with equal intervals between them.

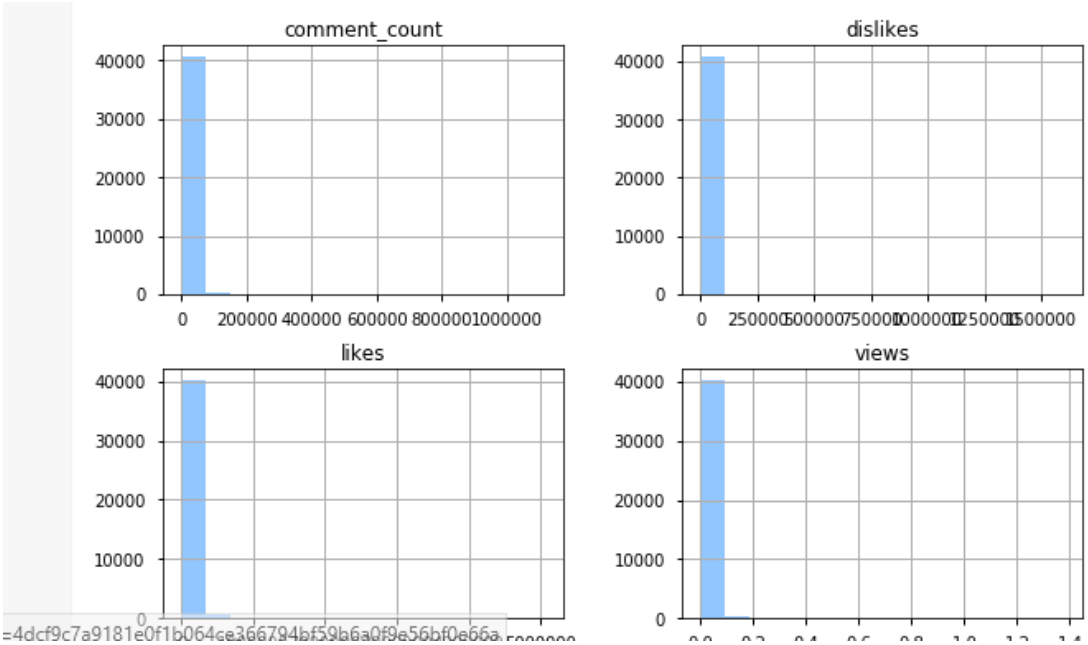
Code: TimeSeries_CA.ipynb

Heat map of View/Likes/Dislikes/Comments count based on category:

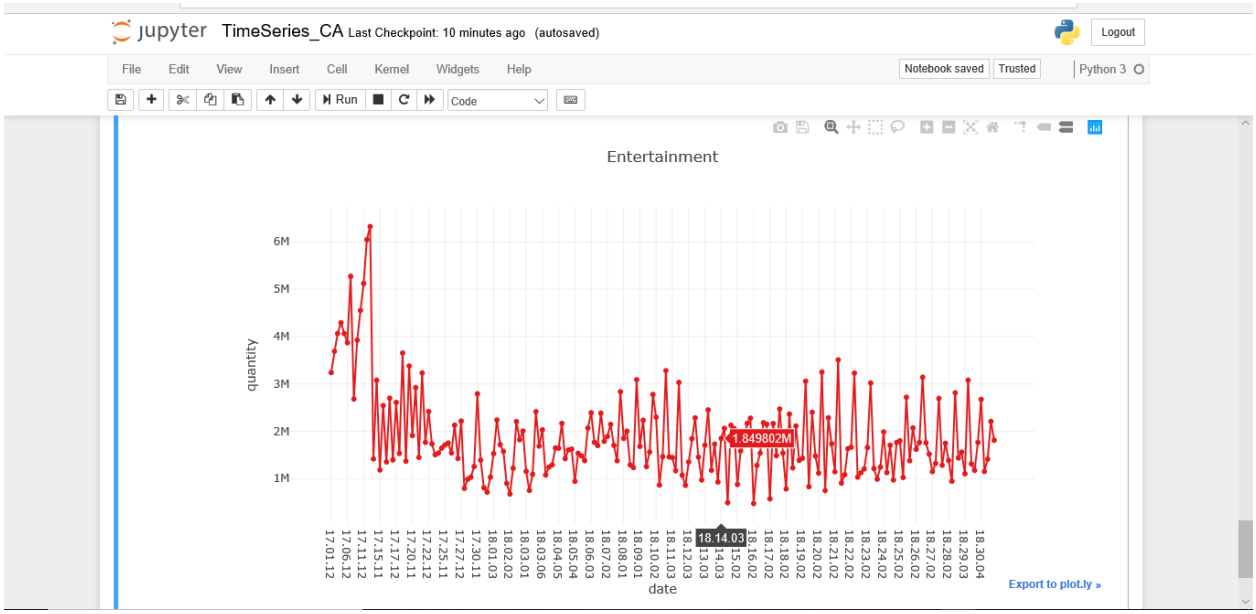
```
In [5]:
1 with plt.style.context(('ggplot')):
2     sns.heatmap(df.corr())
3
```



Histogram analysis of likes/dislikes/Views:



Time Series of YouTube dataset based on the Category and Date:



Conclusion: Good to see that the videos have been linearly increasing during the start of every month for Entertainment Category.

France Time Series Analysis

Heat map of View/Likes/Dislikes/Comments count based on category: A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century

Histogram analysis of likes/dislikes/Views: Analyze the histogram to see whether it represents a normal distribution. Once you have plotted all the frequencies on the histogram, your histogram would show a shape. If the shape looks like a bell curve, it would mean that the frequencies are equally distributed. The histogram would have a peak

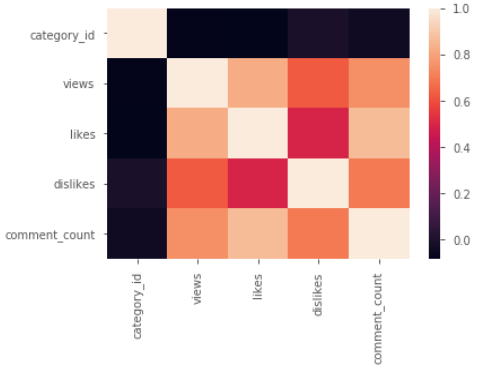
Time Series of US YouTube dataset based on the Category and Date [Music]: A series of values of a quantity obtained at successive times, often with equal intervals between them.

Code: TimeSeries_FR.ipynb

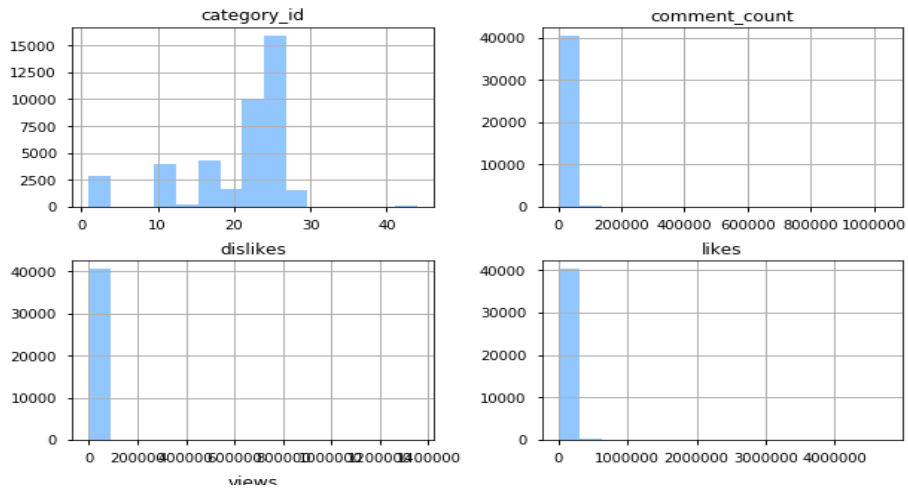
TimeSeries_FR.ipyn
b

Heat map of View/Likes/Dislikes/Comments count based on category:

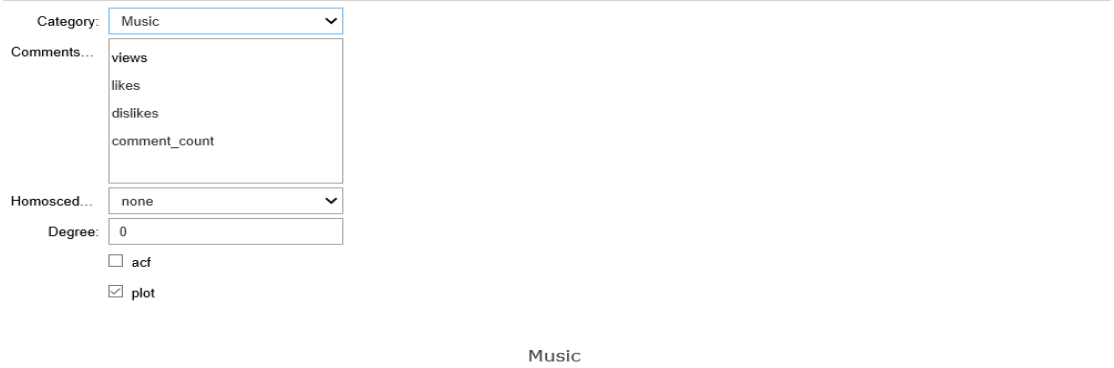
```
In [4]:
1 with plt.style.context(('ggplot')):
2     sns.heatmap(df.corr())
```

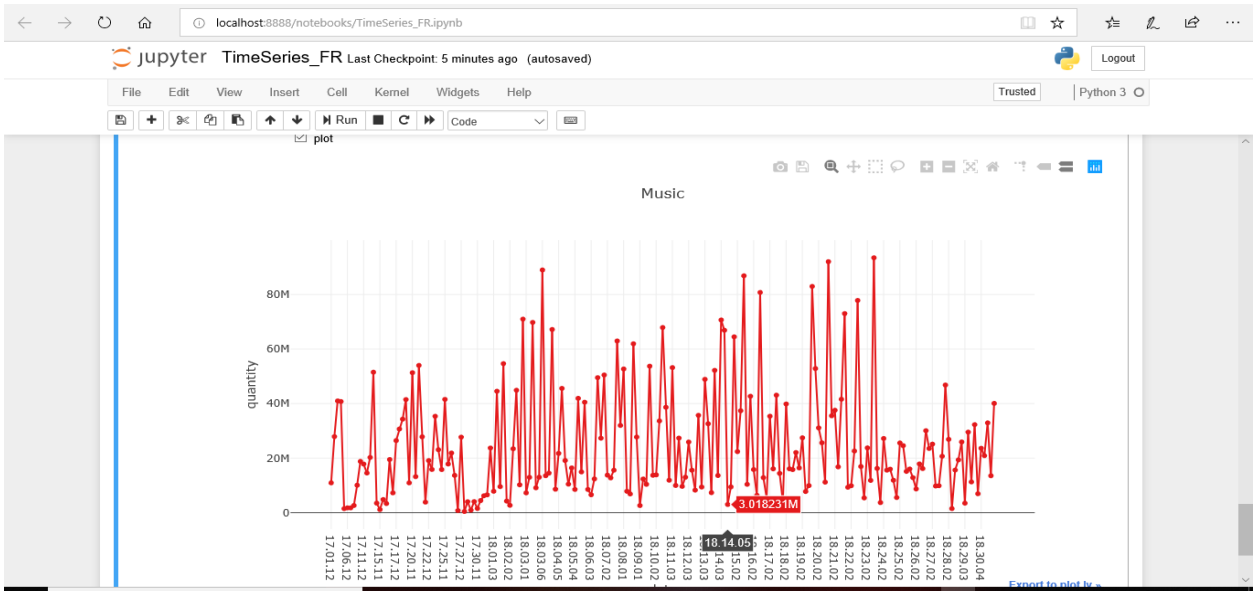


Histogram analysis of likes/dislikes/Views:



Time Series of YouTube dataset based on the Category and Date:





Conclusion: Good to see that the videos have been constant throughout the month Music Category.

Great Britain Time Series Analysis

Heat map of View/Likes/Dislikes/Comments count based on category: A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century

Histogram analysis of likes/dislikes/Views: Analyze the histogram to see whether it represents a normal distribution. Once you have plotted all the frequencies on the histogram, your histogram would show a shape. If the shape looks like a bell curve, it would mean that the frequencies are equally distributed. The histogram would have a peak

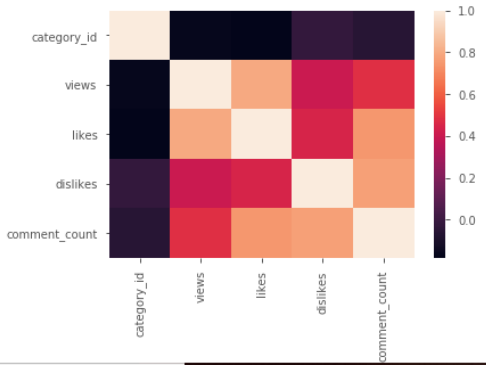
Time Series of US YouTube dataset based on the Category and Date [Science and Technology]: A series of values of a quantity obtained at successive times, often with equal intervals between them.

Code: TimeSeries_GB.ipynb

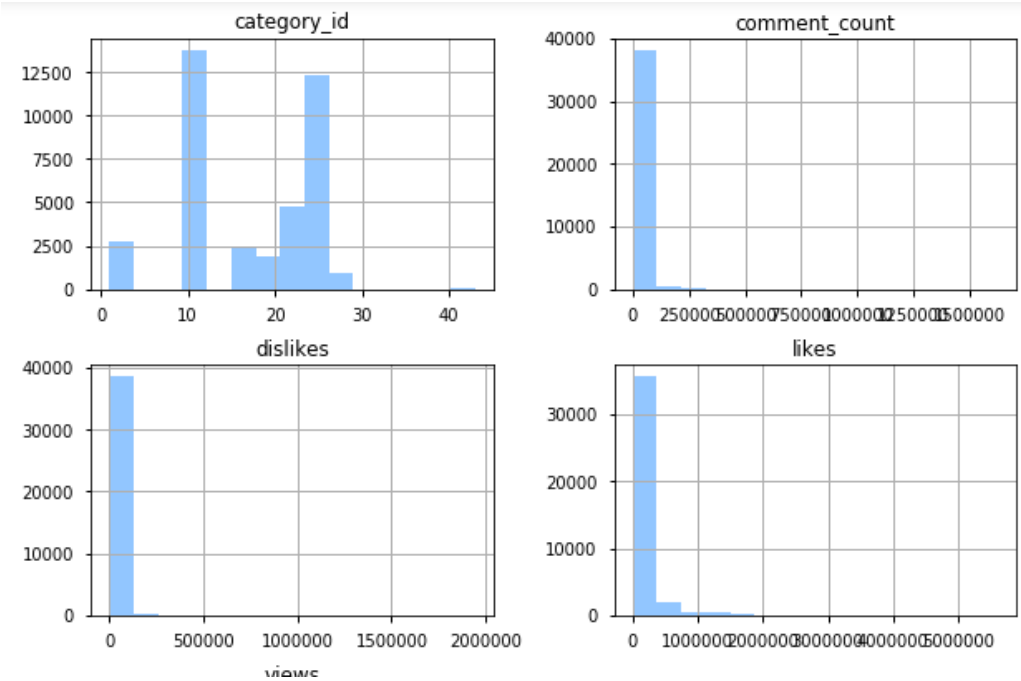
TimeSeries_GB.ipyn
b

Heat map of View/Likes/Dislikes/Comments count based on category:

```
In [4]:
1 with plt.style.context(('ggplot')):
2     sns.heatmap(df.corr())
```



Histogram analysis of likes/dislikes/Views:



Time Series of YouTube dataset based on the Category and Date:

```
68
69 t=widgets.IntText(
70     value=0,
71     description='Degree:',
72     disabled=False)
73
74
75 widgets.interactive(proyect_yt, columns=columns, category=category, tend=t,
76                     Homoscedasticity=["none","inv", "sqrt","log", "Isqrt", "Ilog"])
```

Category: Science & Technology

Comments...
views
likes
dislikes
comment_count

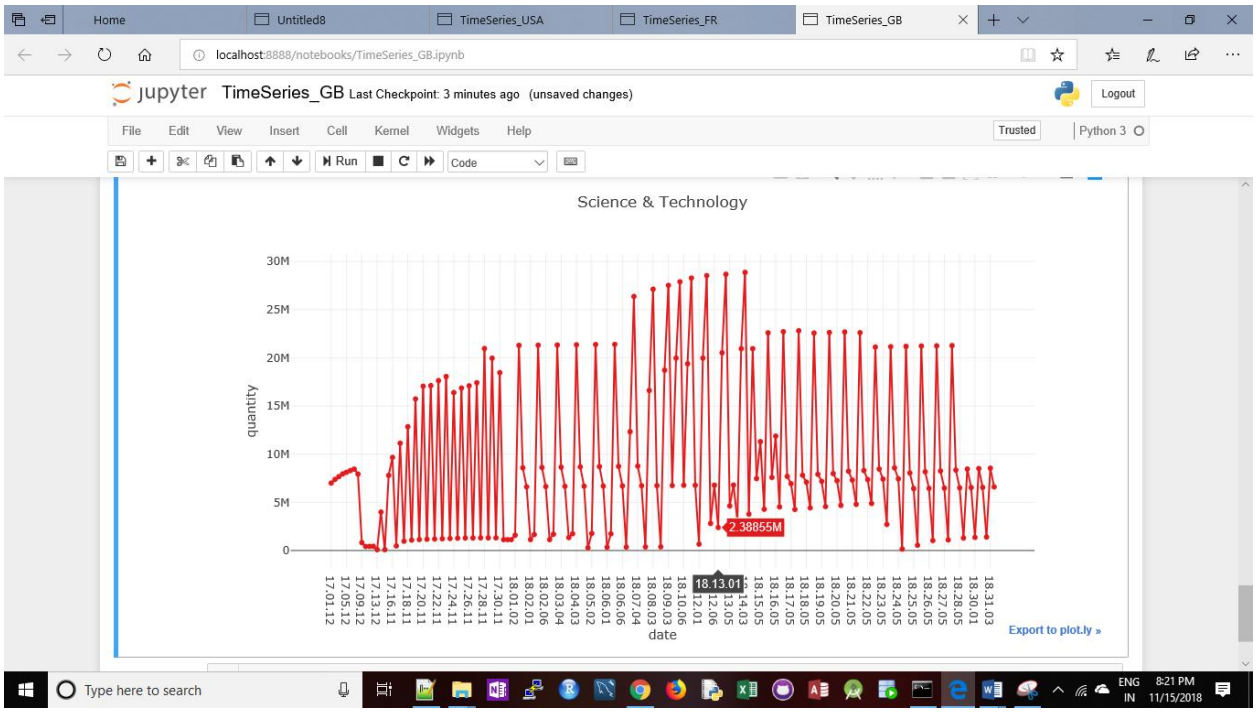
Homosced... none

Degree: 0

acf

plot

Science & Technology



Conclusion: Good to see that the videos have been constant throughout the month Science and Technology Category.

Germany Time Series Analysis

Heat map of View/Likes/Dislikes/Comments count based on category: A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. "Heat map" is a newer term but shading matrices have existed for over a century

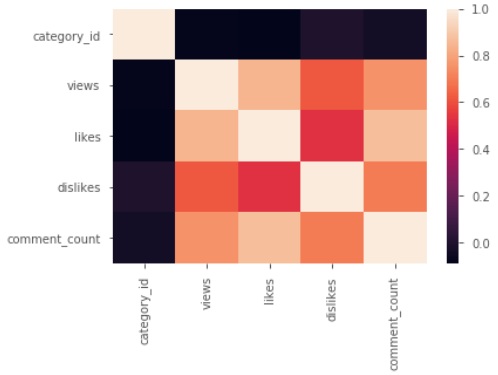
Histogram analysis of likes/dislikes/Views: Analyze the histogram to see whether it represents a normal distribution. Once you have plotted all the frequencies on the histogram, your histogram would show a shape. If the shape looks like a bell curve, it would mean that the frequencies are equally distributed. The histogram would have a peak

Time Series of US YouTube dataset based on the Category and Date [Education]: A series of values of a quantity obtained at successive times, often with equal intervals between them.

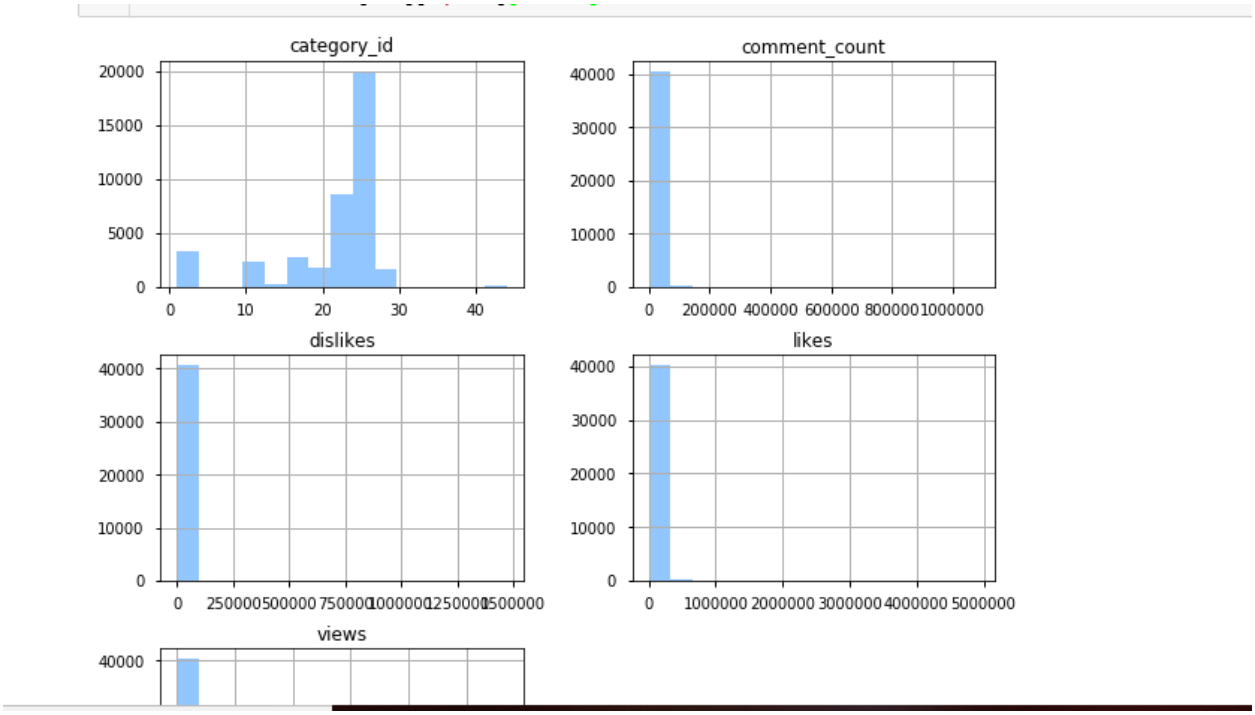
Code: TimeSeries_DE.ipynb

Heat map of View/Likes/Dislikes/Comments count based on category:

```
In [4]:
1
2 with plt.style.context(('ggplot')):
3     sns.heatmap(df.corr())
```



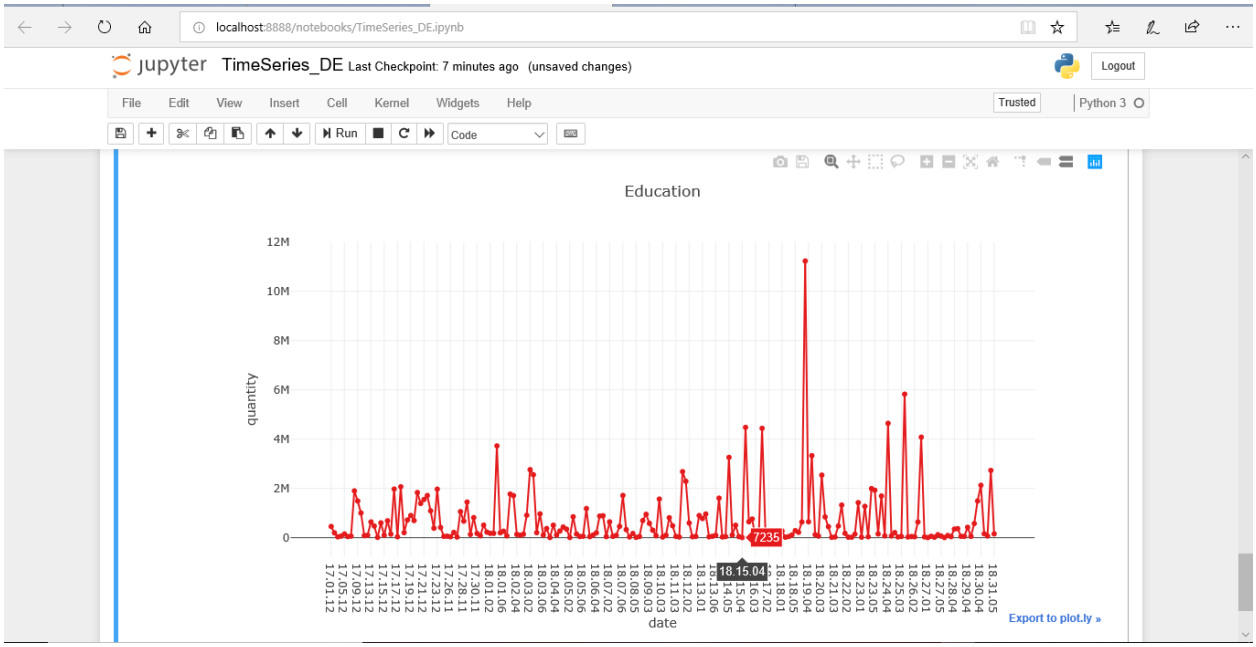
Histogram analysis of likes/dislikes/Views:



Time Series of YouTube dataset based on the Category and Date:

```
t=widgets.IntText(
    value=0,
    description='Degree:',
    disabled=False)


widgets.interactive(proyect_yt, columns=columns, category=category, tend=t,
                    Homoscedasticity=["none","inv", "sqrt","log", "Isqrt", "Ilog"])
```

Category: Education
Comments...: views, likes, dislikes, comment_count
Homosced...: none
Degree: 0
acf
plot



Conclusion: Good to see that the videos is low when compared to other Time Series interpretation for Education category.

References

The YouTube Dataset has been collected from Kaggle.com
Link: (**https://bit.ly/2DAPfBk**).