Increasing product brand's visibility using social media

Alexandre Geraldo

Austin Brantley

Franklin Gonzales

Prashanth Saseenthar

Georgia Tech Data Science and Analytics Boot Camp 2019

Before we start

Useful links (also available on Slack)

- **GitHub repository**
  https://github.com/ageraldo1/DataScienceProject1.git

- **Jupyter notebooks**
  https://github.com/ageraldo1/DataScienceProject1/tree/master/documents/analysis/videos

  Nbviewer(online viewer)
  https://nbviewer.jupyter.org/github/ageraldo1/DataScienceProject1/blob/master/documents/analysis/videos/alex/Videos%20Analysis.ipynb

- **API source code**
  - **YouTube**
    https://github.com/ageraldo1/DataScienceProject1/blob/master/src/api/youtube/generate_video_lenght_ds.py

    https://github.com/ageraldo1/DataScienceProject1/blob/master/src/api/youtube/generate_video_lenght_ds.py

  - **Sentimental Analysis**
    https://github.com/ageraldo1/DataScienceProject1/blob/master/src/sentimental/sentimental_analysis.py

Agenda

- **Motivation & Inspirations**

- **Data Cleanup & Exploration**
  - Merging, fixing datatypes, removing duplicate records
  - New datasets (Google YouTube API, vaderSentiment)

- **Data Analysis**
  - Summary
  - Time Series
  - Trends
    - Views per category
    - Views Distribution per Category
    - Used Tags
    - Sentimental Analysis
    - Top 10 Creators & Categories
    - Videos Duration (optional)
    - Ratios per Categories (optional)

  - Correlations
    - Correlation between metrics
    - Correlation between Title Sentiment and Views
    - Correlation between Tags Sentiment and Views
    - Correlation between Total of Subscribers & Total of Uploads and Views

- **Making the Call**

- **Q & A**

Motivation & Inspirations

“If you can't explain it simply, you don't understand it well enough.”

Albert Einstein

Motivation & Inspirations

✓ Social networks are one of the fastest growing industries in the world.

✓ Social Media is crucial for Business Marketing.

✓ Data Science & Big Data Science & Data Analytics flavors mixed in one place.

✓ Entrepreneurial spirit.

✓ Several paths can be explored.

✓ Topic of easier understanding and most of the people like to talk about.

Why YouTube?

- ✓ It has a wide reach and generates plenty of traffic.
- ✓ YouTube has over 1 billion users, who spend millions of hours per day viewing videos.
- ✓ YouTube is localized in over 70 countries and is available in 76 languages.
- ✓ It has greater reach than cable in the US.
- ✓ Pay Per (Actual) View Of Your Ad.
- ✓ One of the shortest path to connect your brand to people around the world.

YouTube

How a trending dataset can help?

✓ Helps viewers see what's happening on YouTube and in the world.

✓ Trending video = Video running on steroids.

✓ Potential to connect to people on a large scale.

✓ Views = People.

✓ Video visualization is a trigger to connect people with product.

✓ Keep up the momentum.

Limitations

- ✓ Unable to identity if a person see a video more than once.
- ✓ Unable to retrieve information of removed channels.
- ✓ Unable to retrieve information of removed videos.

Data Cleanup & Exploration

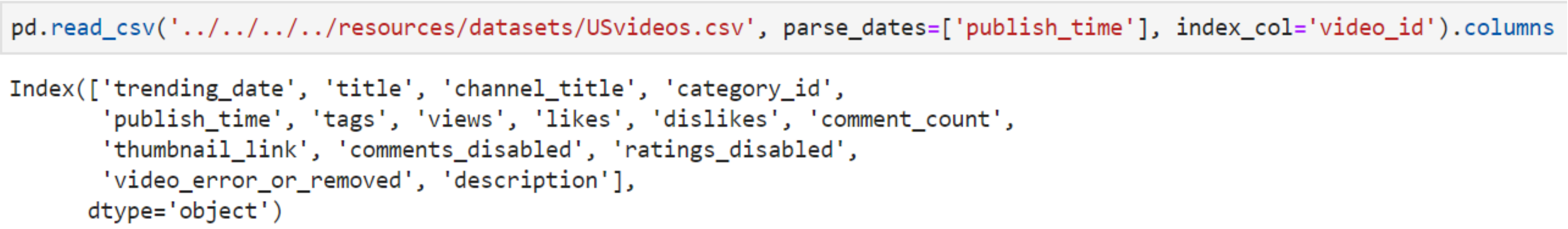
“Torture the data, and it will confess to anything.”

Ronald Coase

Merging, fixing datatypes, removing duplicate records.

Dataset : USvideos.csv

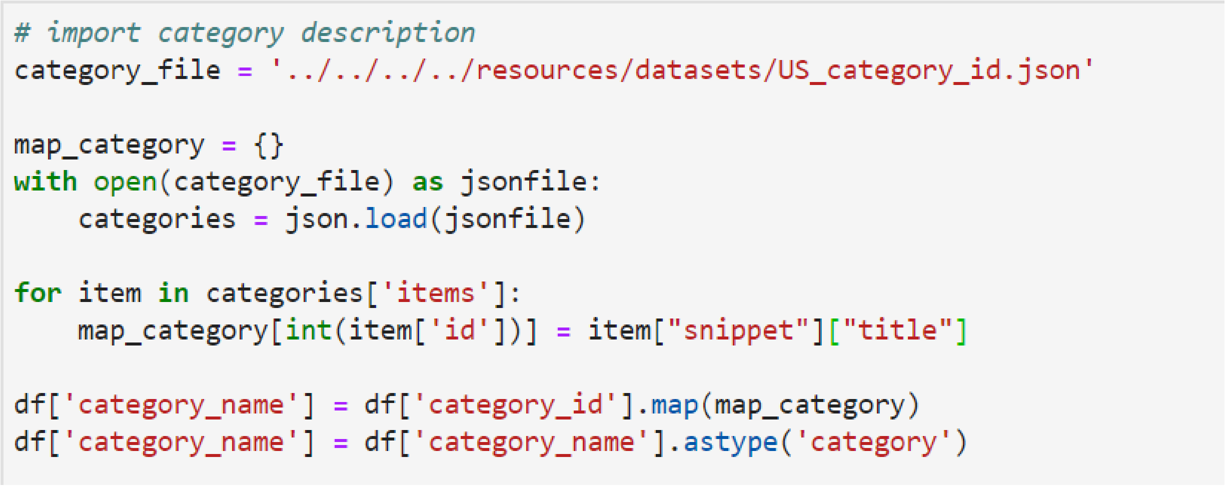
Description : Kaggle's dataset that contains a list of top trending videos of US.



```
pd.read_csv('../../../../resources/datasets/USvideos.csv', parse_dates=['publish_time'], index_col='video_id').columns
```

```
Index(['trending_date', 'title', 'channel_title', 'category_id',
       'publish_time', 'tags', 'views', 'likes', 'dislikes', 'comment_count',
       'thumbnail_link', 'comments_disabled', 'ratings_disabled',
       'video_error_or_removed', 'description'],
      dtype='object')
```

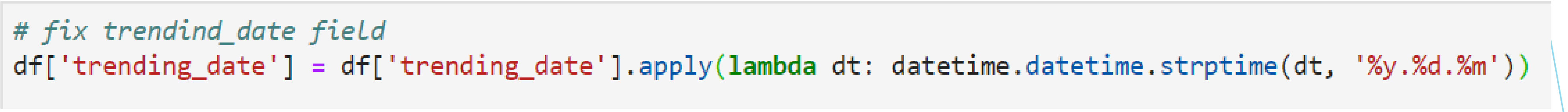
✓ **Importing Category Description**



```
# import category description
category_file = '../../../../resources/datasets/US_category_id.json'

map_category = {}
with open(category_file) as jsonfile:
    categories = json.load(jsonfile)

for item in categories['items']:
    map_category[int(item['id'])] = item["snippet"]["title"]

df['category_name'] = df['category_id'].map(map_category)
df['category_name'] = df['category_name'].astype('category')
```

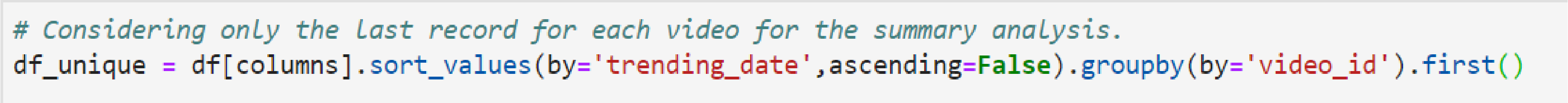
Merging, fixing datatypes, removing duplicate records.

- ✓ **Fixing trending date attribute**



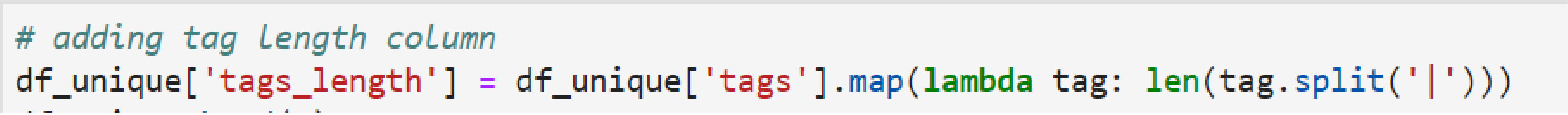
```
# fix trendind_date field
df['trending_date'] = df['trending_date'].apply(lambda dt: datetime.datetime.strptime(dt, '%y.%d.%m'))
```

- ✓ **Removing duplicate records**



```
# Considering only the last record for each video for the summary analysis.
df_unique = df[columns].sort_values(by='trending_date',ascending=False).groupby(by='video_id').first()
```

- ✓ **Adding new attributes**



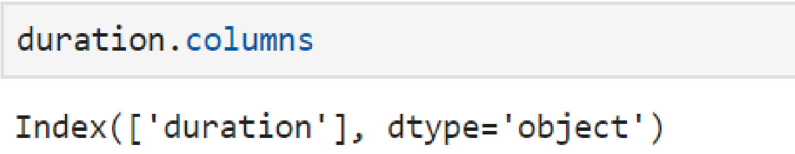
```
# adding tag length column
df_unique['tags_length'] = df_unique['tags'].map(lambda tag: len(tag.split('|')))
```

New Datasets : duration, sentimental, creators

Dataset : USvideos_duration.csv

Description : Video duration dataset created using YouTube APIs.

Source code : src/api/youtube/generate_video_lenght_ds.py



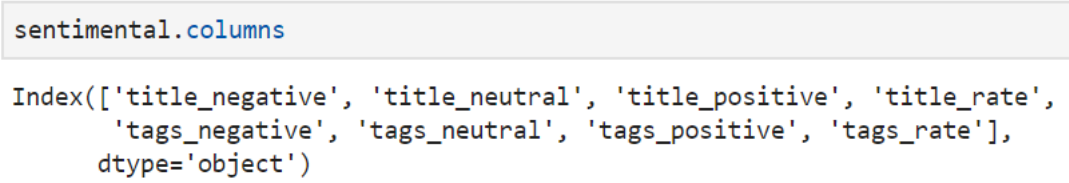
```
duration.columns
```

```
Index(['duration'], dtype='object')
```

Dataset : US_sentimental.csv

Description : Sentimental analysis of videos titles and videos tags created using Vader sentimental package.

Source code : src/api/youtube/src/sentimental/sentimental_analysis.py



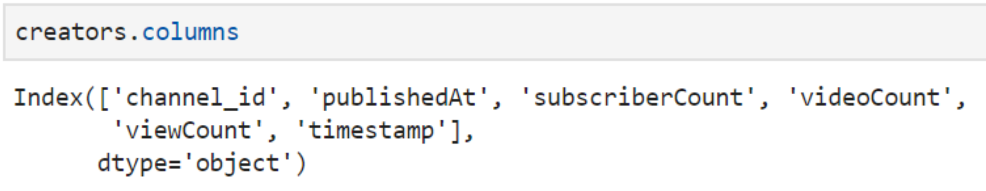
```
sentimental.columns
```

```
Index(['title_negative', 'title_neutral', 'title_positive', 'title_rate',
       'tags_negative', 'tags_neutral', 'tags_positive', 'tags_rate'],
      dtype='object')
```

Dataset : USchannels.csv

Description : channel creators dataset created using YouTube APIs.

Source code : src/api/youtube/generate_video_lenght_ds.py



```
creators.columns
```

```
Index(['channel_id', 'publishedAt', 'subscriberCount', 'videoCount',
       'viewCount', 'timestamp'],
      dtype='object')
```

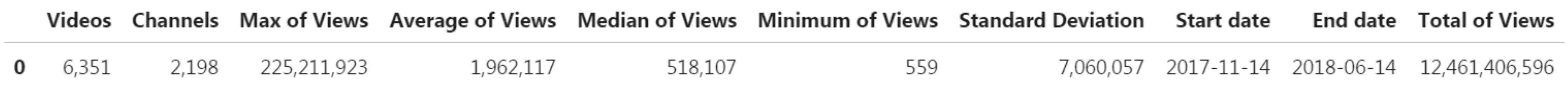
Data Analysis

“If we have data, let’s look at data. If all we have are opinions, let’s go with mine”

Jim Barksdale

Summary

Question #1 : What valuable trends can we extract from our dataset?



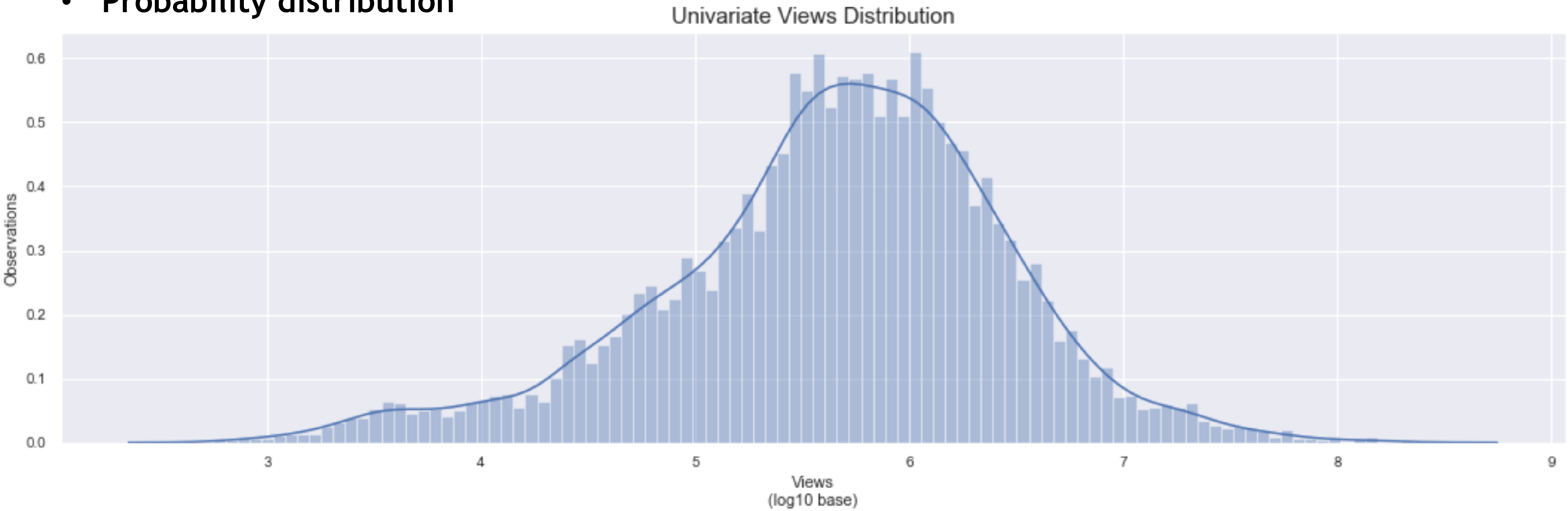
	Videos	Channels	Max of Views	Average of Views	Median of Views	Minimum of Views	Standard Deviation	Start date	End date	Total of Views
0	6,351	2,198	225,211,923	1,962,117	518,107	559	7,060,057	2017-11-14	2018-06-14	12,461,406,596

- ✓ Excellent amount of data.
- ✓ 2M of Views on Average.
- ✓ Data is spread-out by 7M of Views.
- ✓ Total number of views is almost 38 times higher than US population (2018).
- ✓ Write down the number of max visualizations (225M). We'll be using it in the end of the presentation.

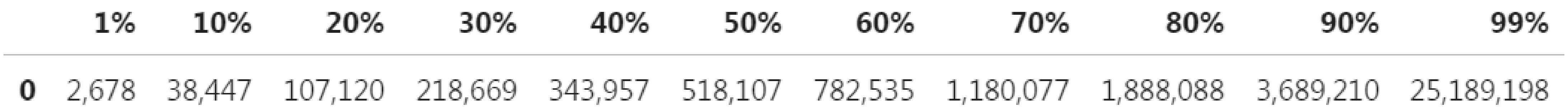
Summary

Question #2 : Can we predict the potential to reach a number X of people ?

- **Probability distribution**



- **Quantiles**



	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
0	2,678	38,447	107,120	218,669	343,957	518,107	782,535	1,180,077	1,888,088	3,689,210	25,189,198

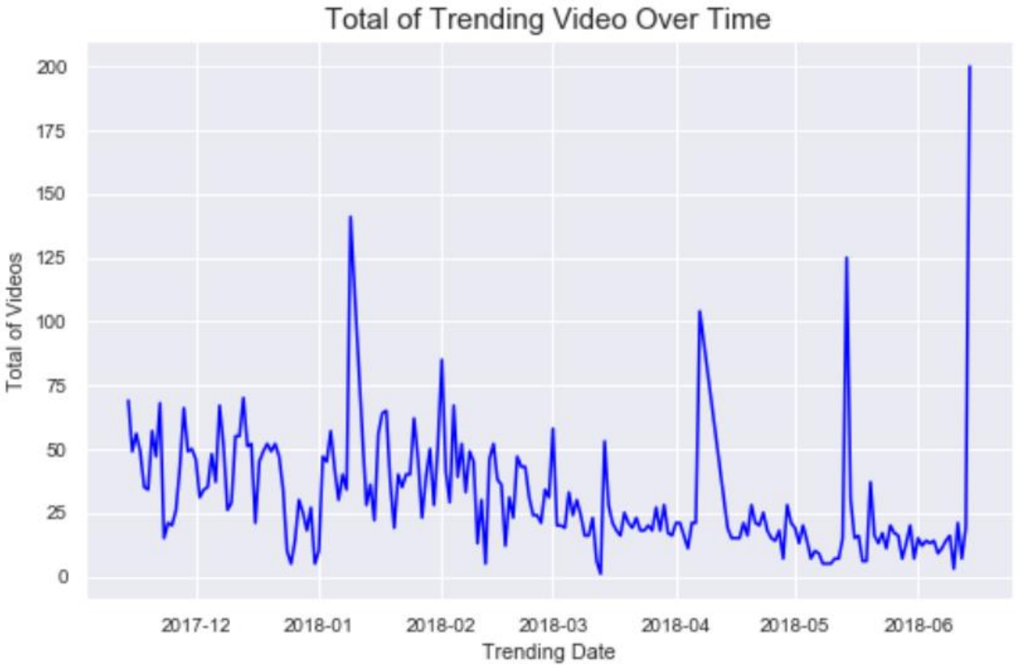
Time-Series

“If you can look into the seeds of time, and say which grain will grow and which will not, speak then unto me.”

William Shakespeare

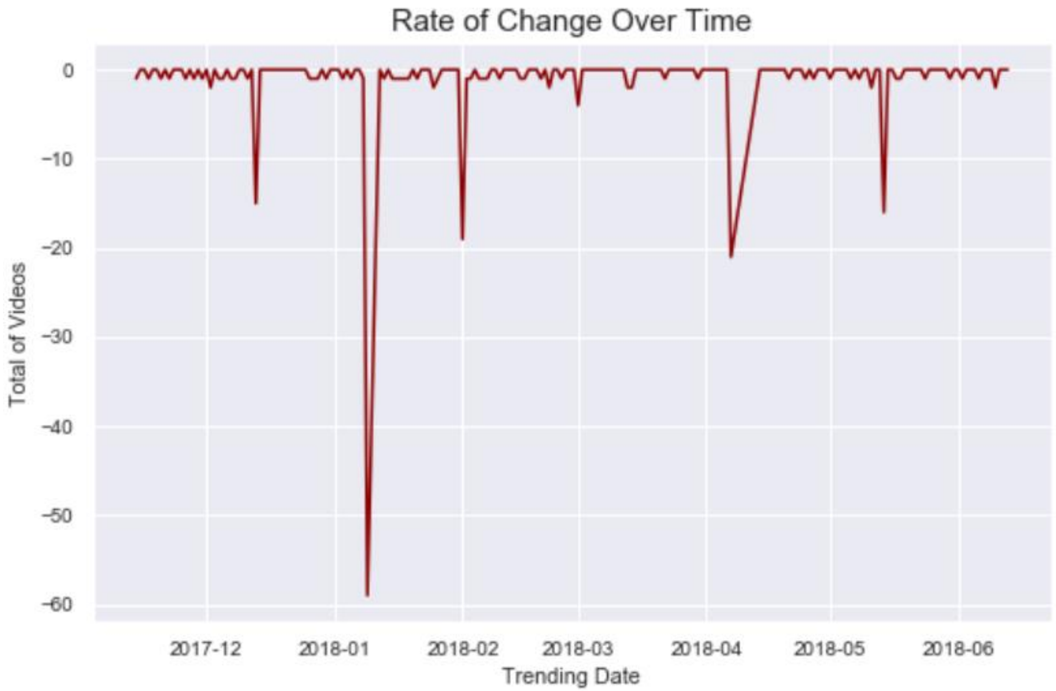
Time Series

Question #3 : How trending videos activity looks live over time?



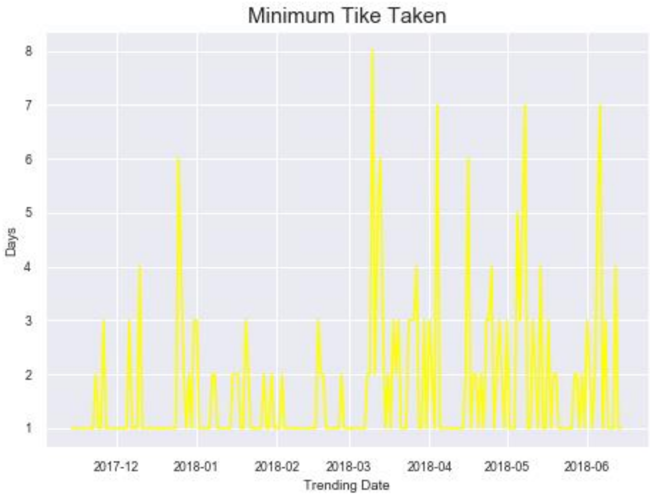
- Insights for deployment campaign dates
- Reveal activity behaviors
- Powerful insight to drive decisions

- Reveal the speed of when activity is changing.
- Reveal activity behaviors.
- Reveal maximum and minimum behaviors.

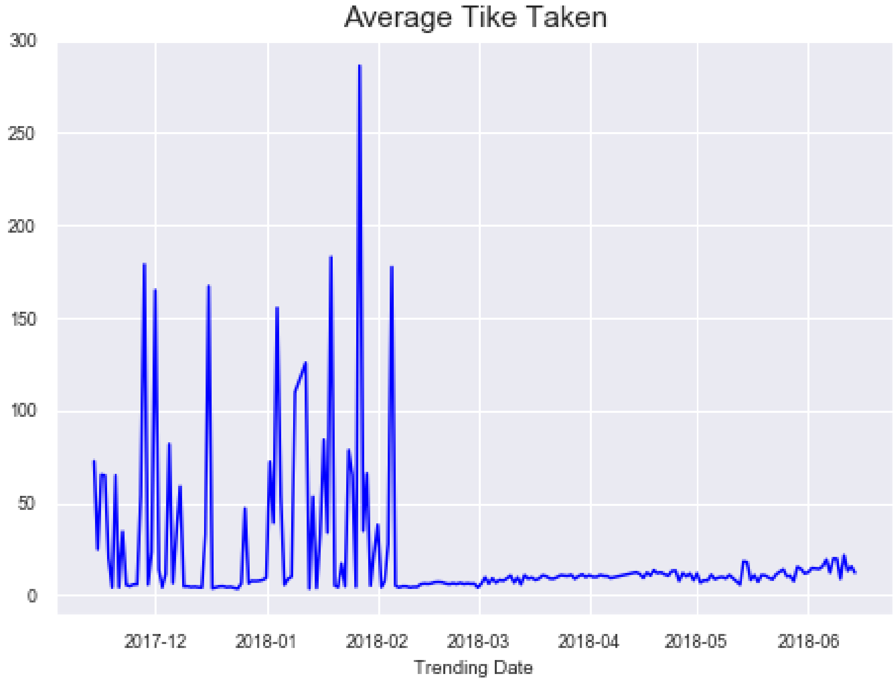


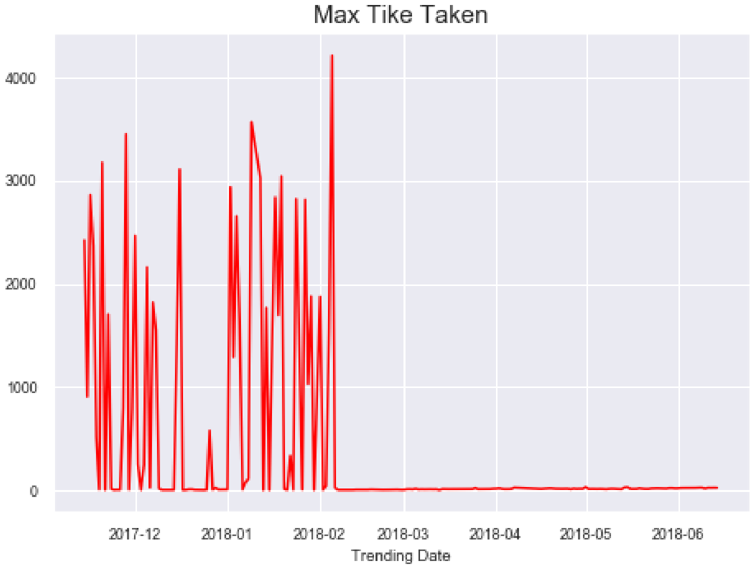
Time Series

Question #4 : How long takes for a video become a trending video ?



- ✓ Takes forever between December and February





- ✓ The “miracle” from March through June

- ✓ Christmas & End of Year side effects

Trends

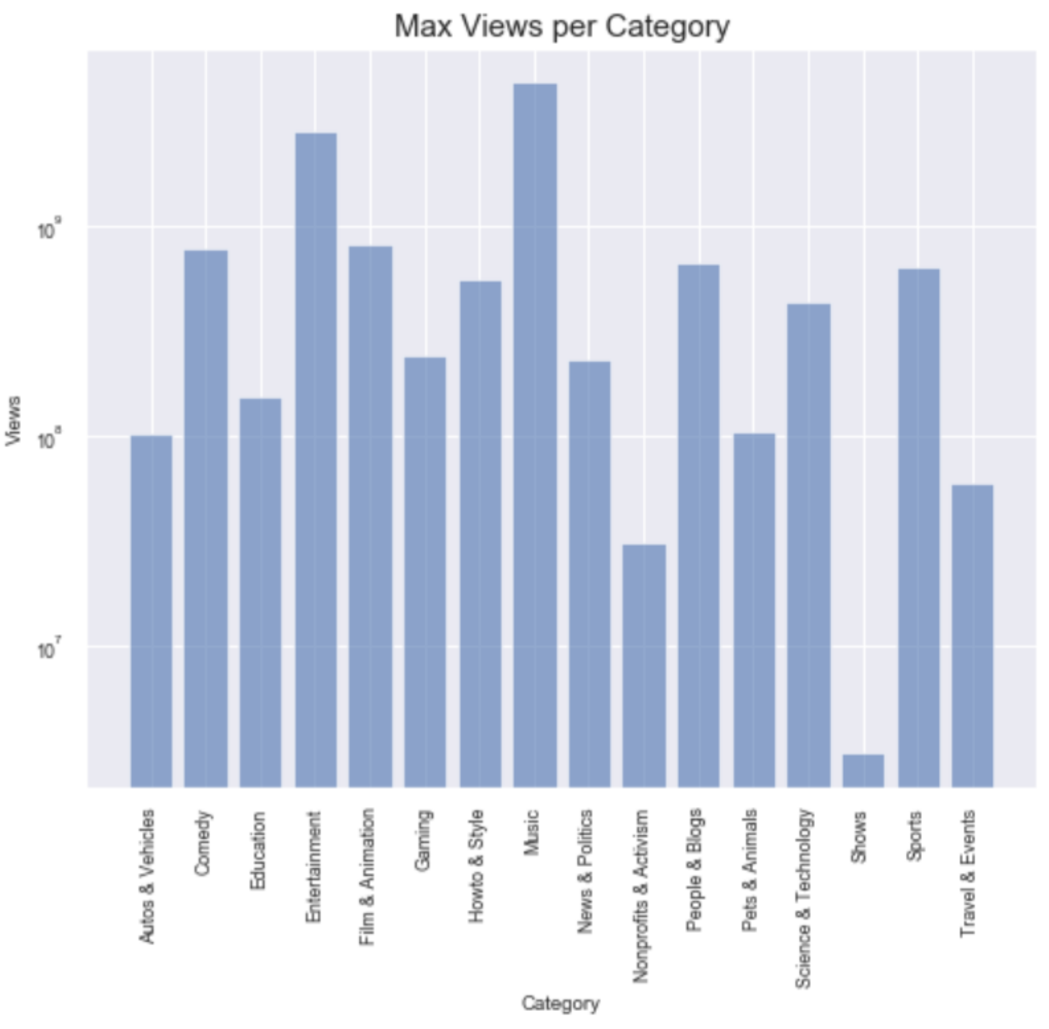
“The problem with data is that it says a lot, but it also says nothing.”

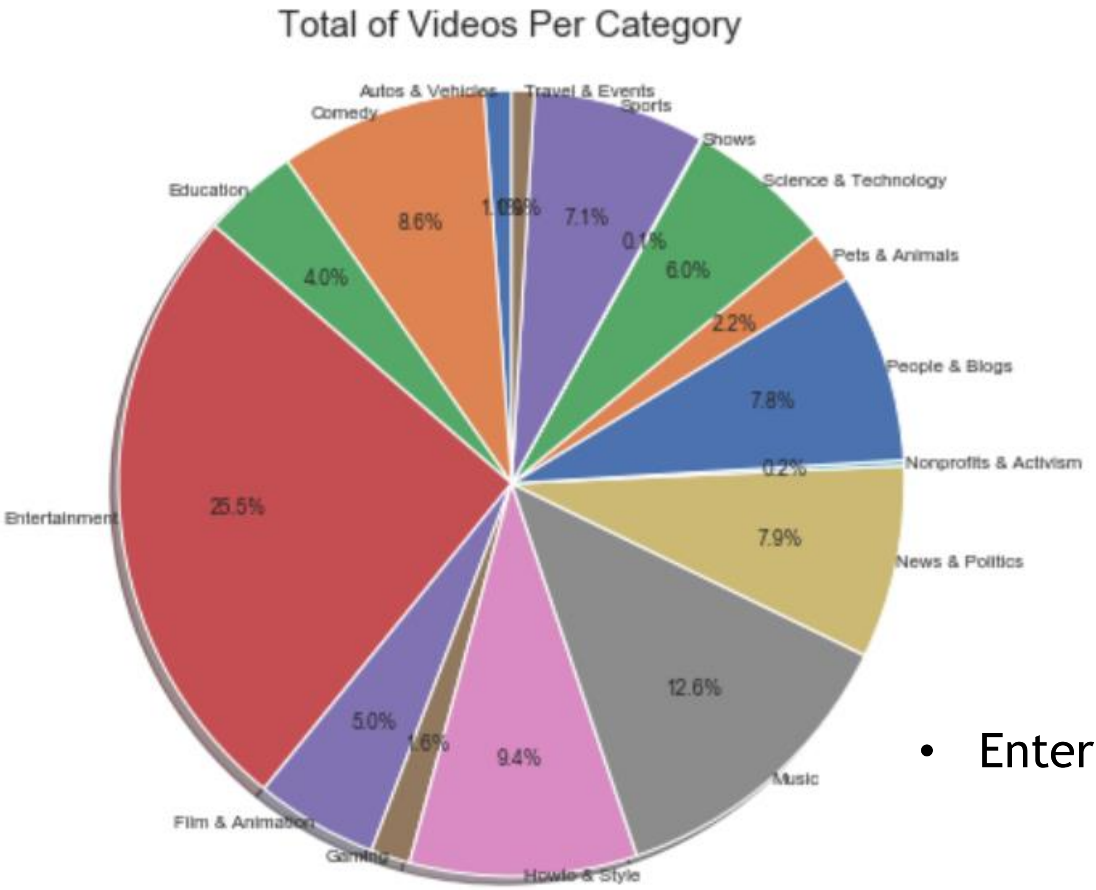
Sendhil Mullainathan

Trends

Question #5 : What are the categories dominating YouTube?

- Music and entertainment categories dominate the total of views.

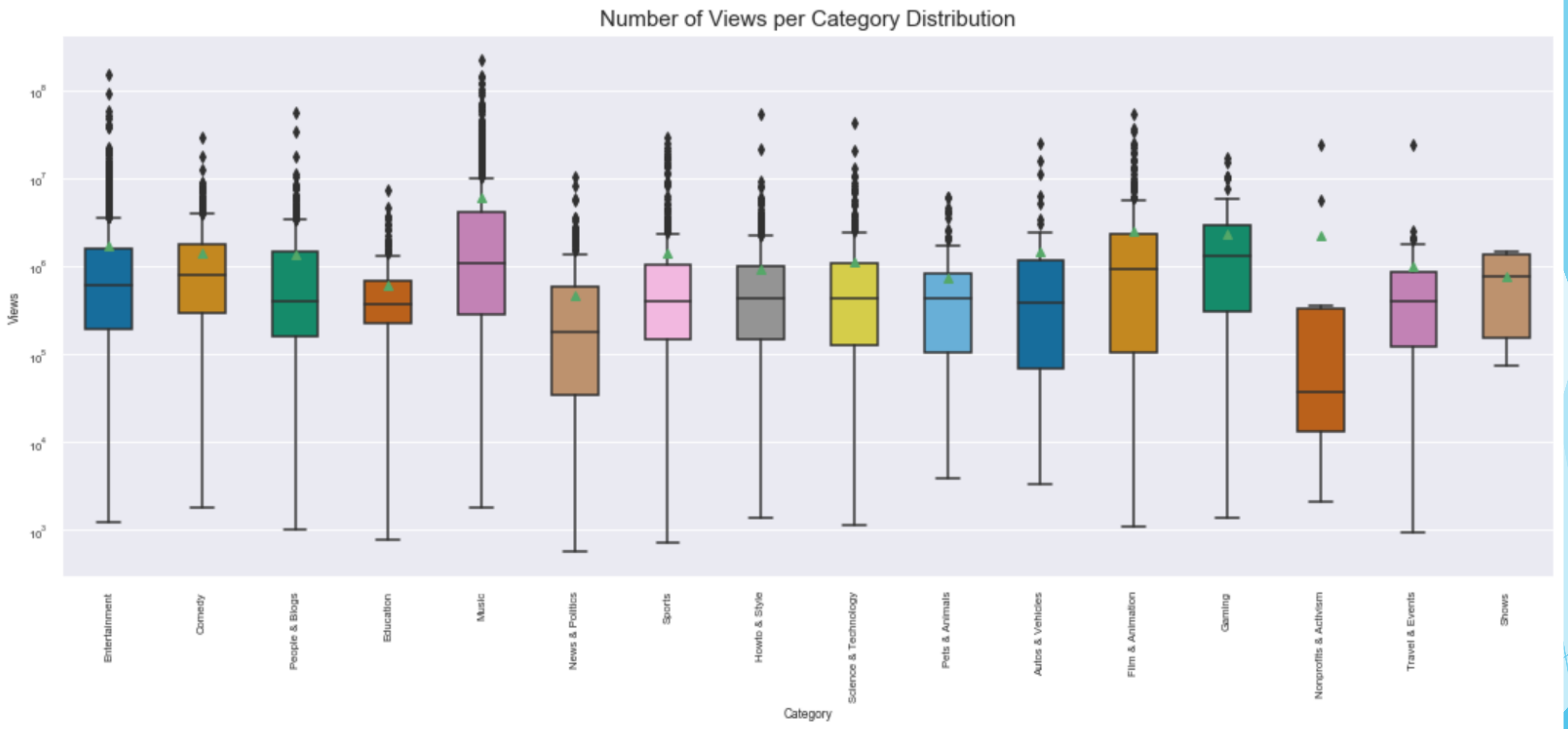




- Entertainment category dominate total of uploads.

Trends

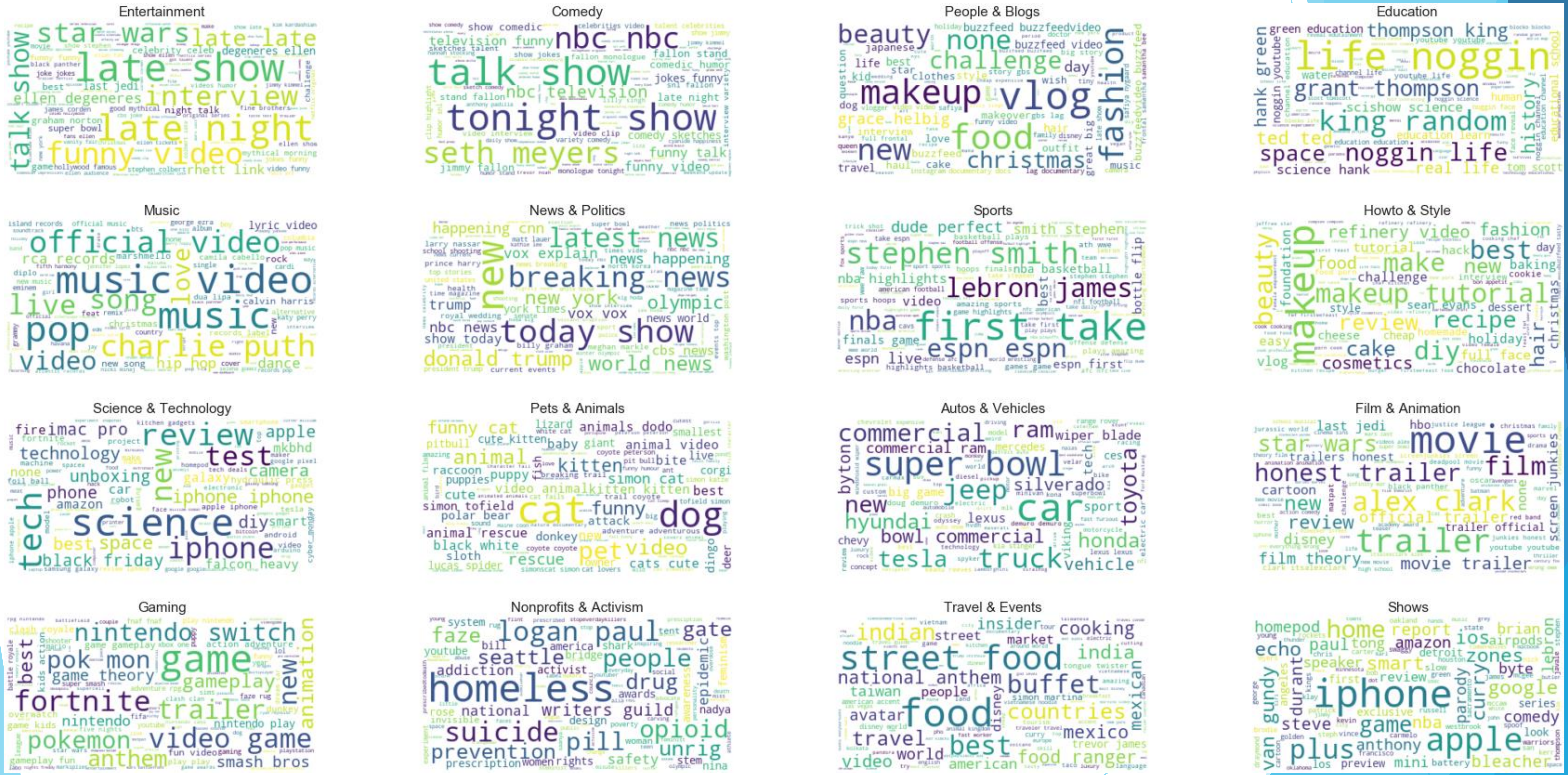
Question #6 : How spread out is the data across video categories?



- All categories except Show presents outliers.
- Show distribution seems to be a normal distribution.
- Music, Entertainment, and Film & Animations presents a considerable number of outliers.

Trends

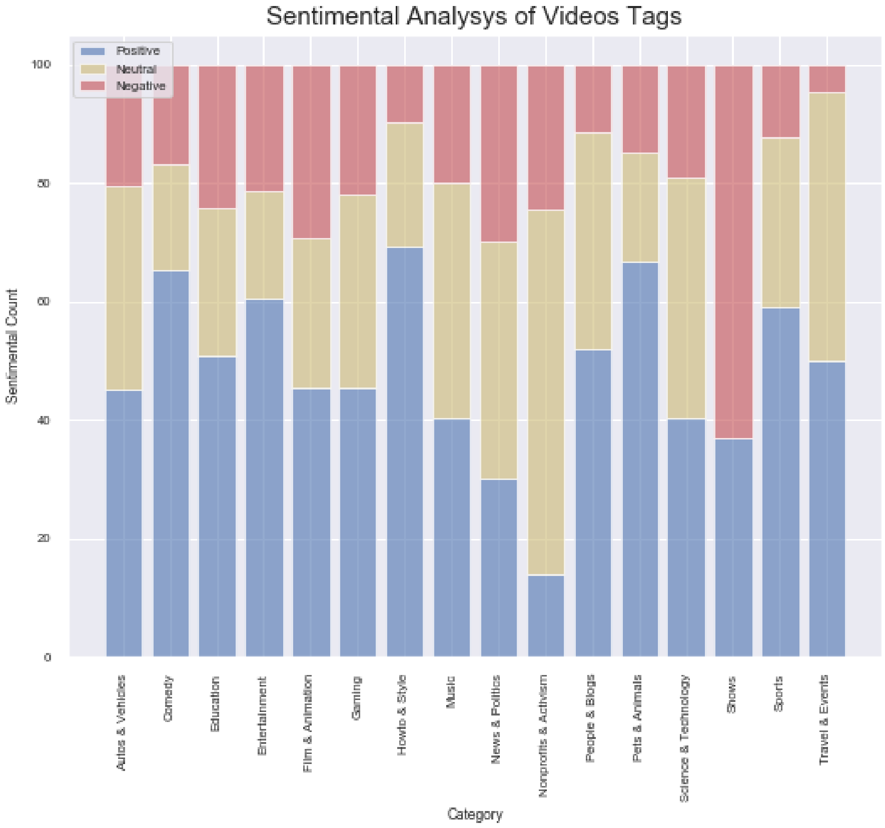
Question #7 : What are the most used tags across the categories?



- Discovering the word iPhone as one of the most used keywords for the Shows category still remains a secret for the group members.

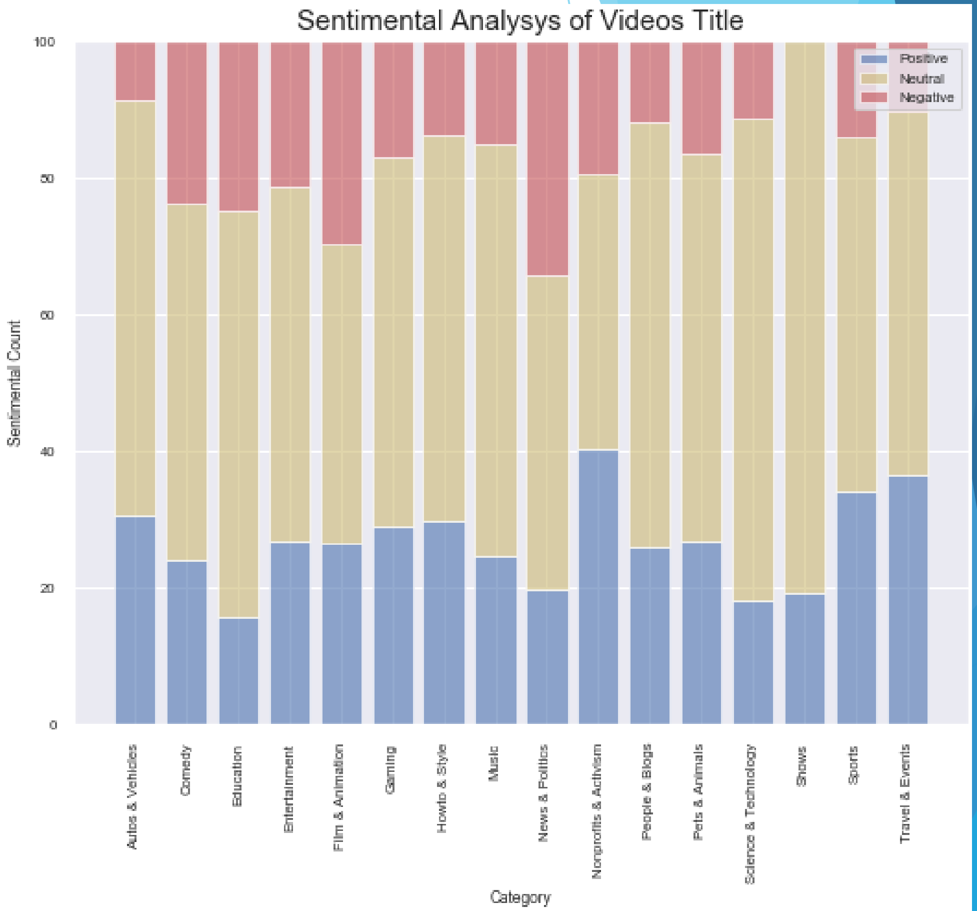
Trends

Question #8 : What is the sentimental used by titles and tags across categories?



New Question :

Is there any correction between titles and tags sentiments with the number of views?



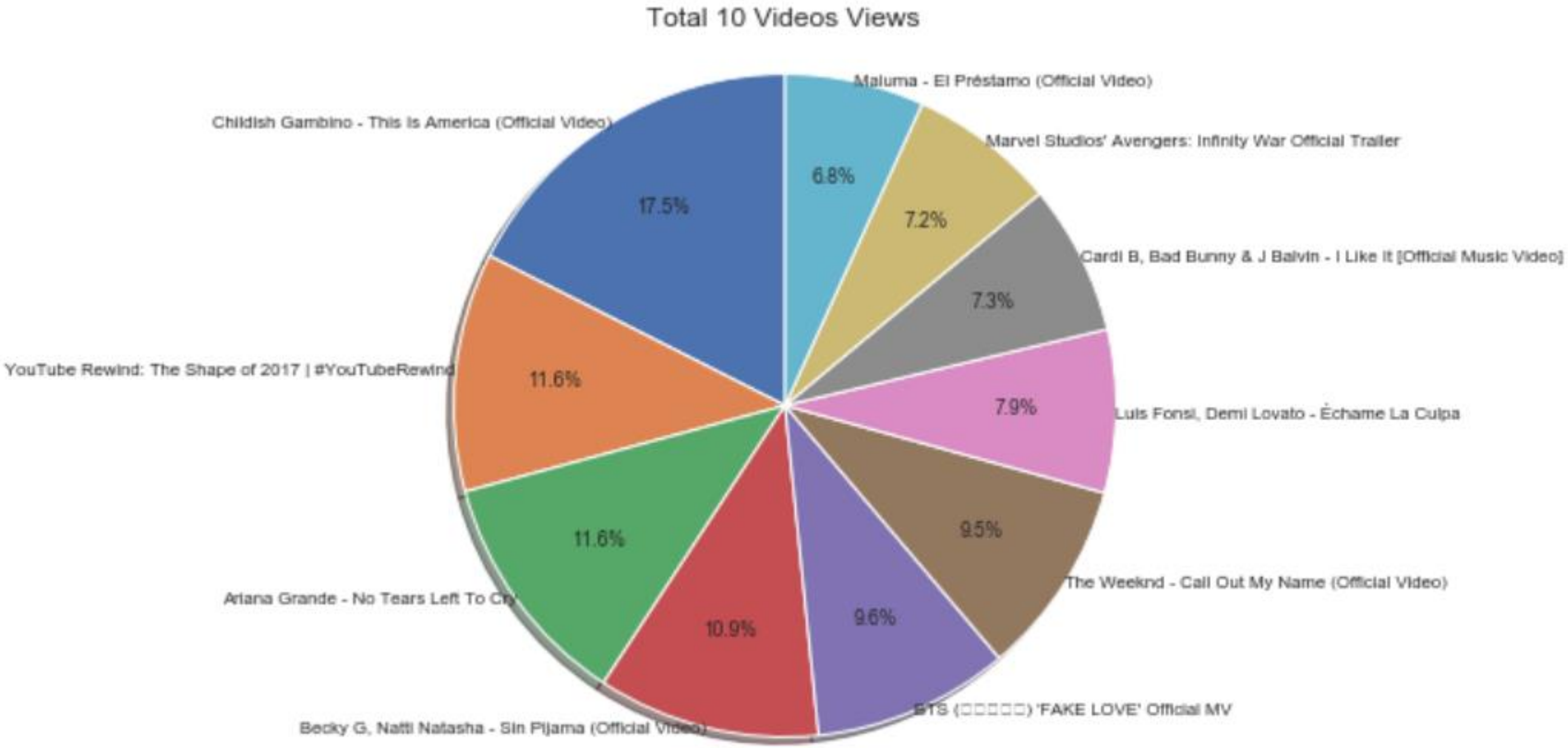
How-to & Style and Pets & Animals categories present higher positives score for tags (higher than 60%).

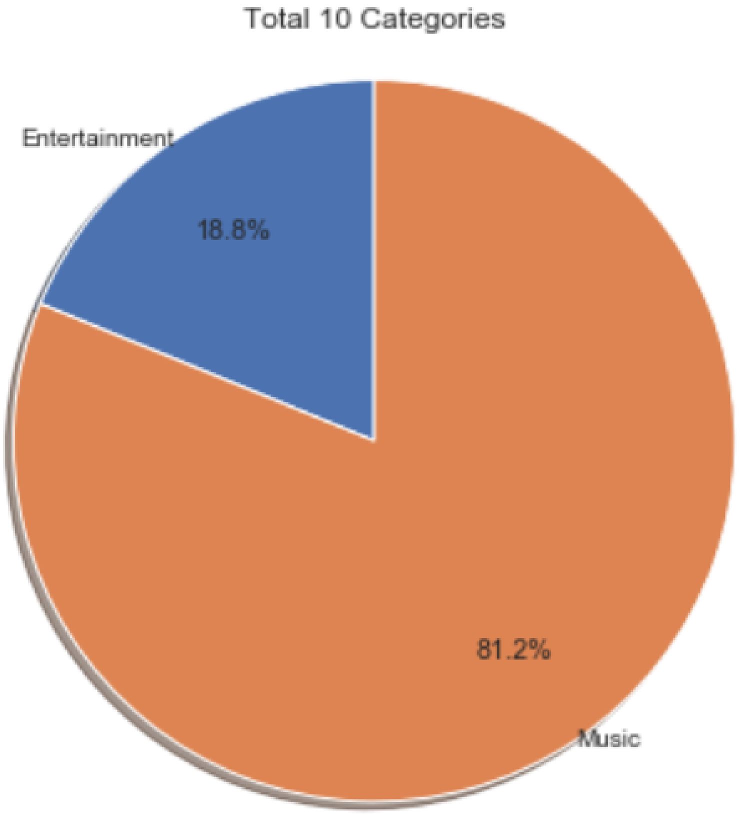
Nonprofits & Activism category presents the higher positive score but this score is below 50%.

Trends

Question #9 : Who are the Top 10 creators and what are the Top 10 categories ?

- YouTube is a territory dominated by Music & Entertainment.



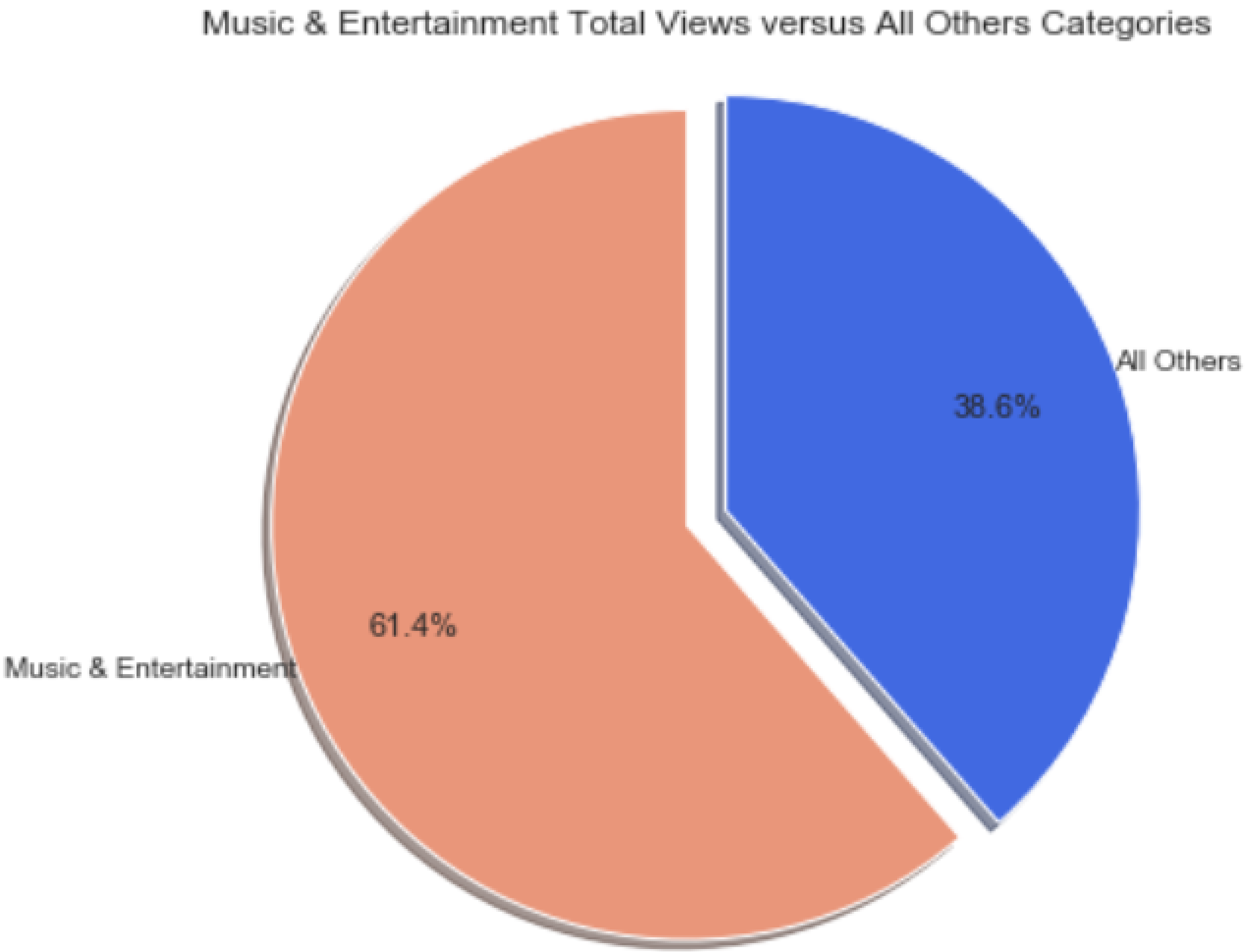


- It's clear also to see that they tend to watch the same video several times. Retaining people attention is difficult and knowing where to explore this kind of behavior can potentially lead to success.

Trends

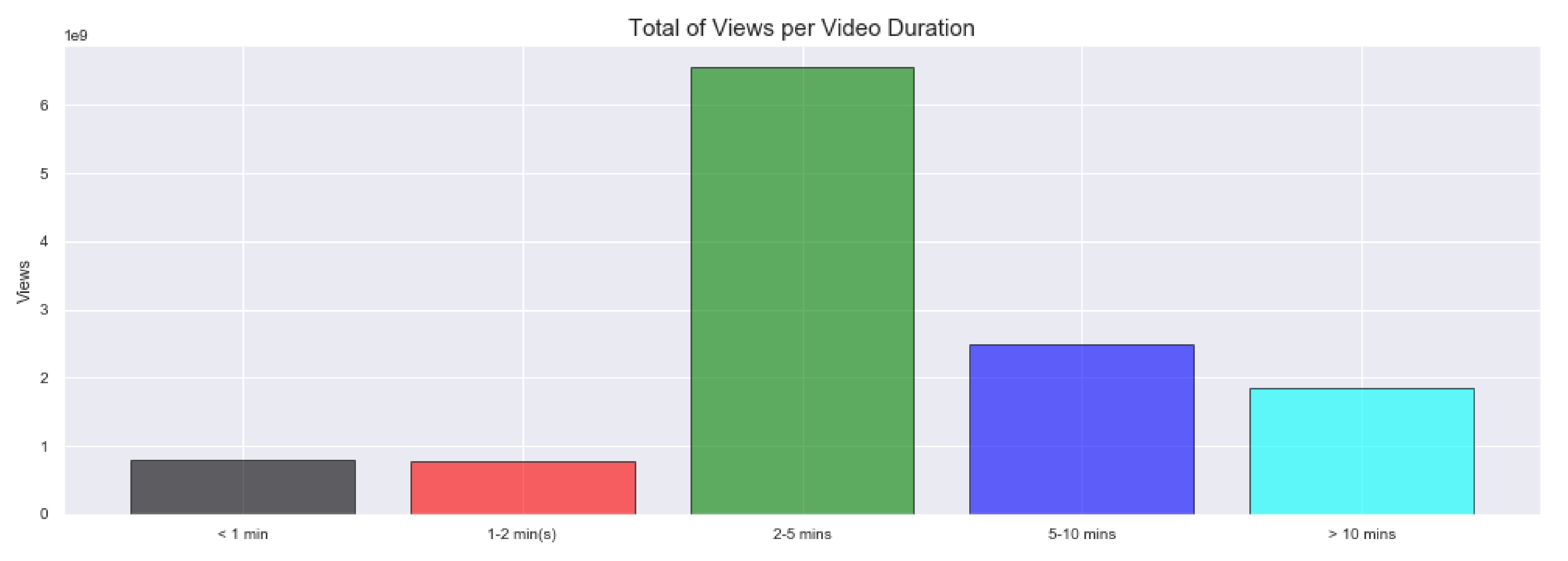
Question #10 : How powerful are Music & Entertainment combined ?

- See for yourself…



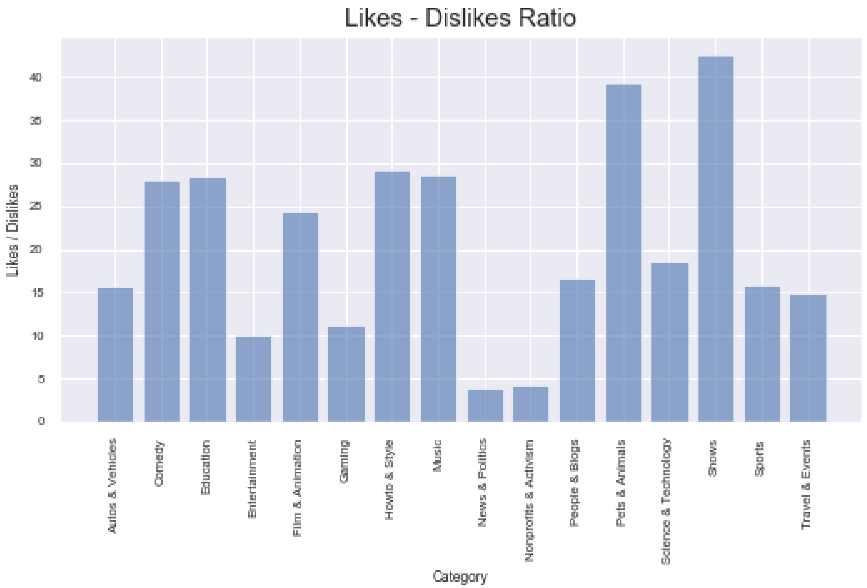
Trends

Question #11 : How long should be the video duration ? (optional)

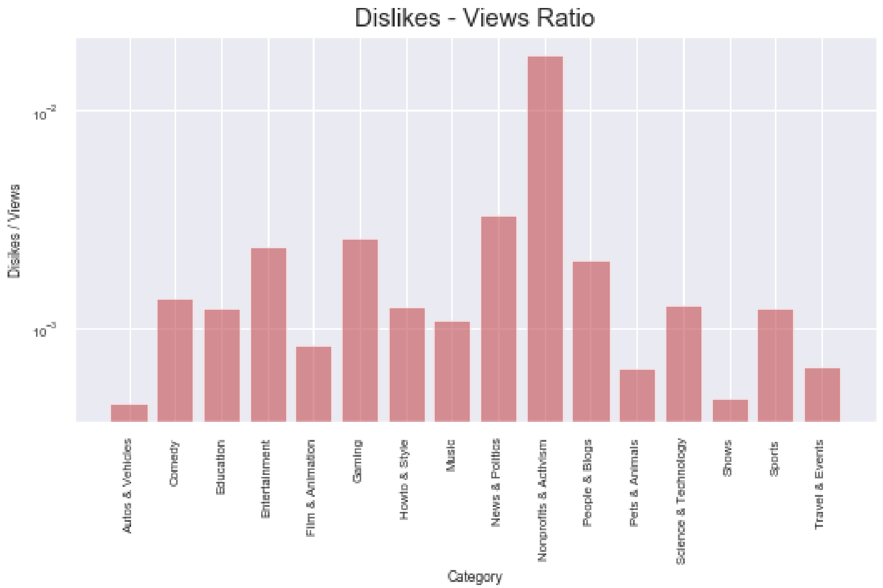


Trends

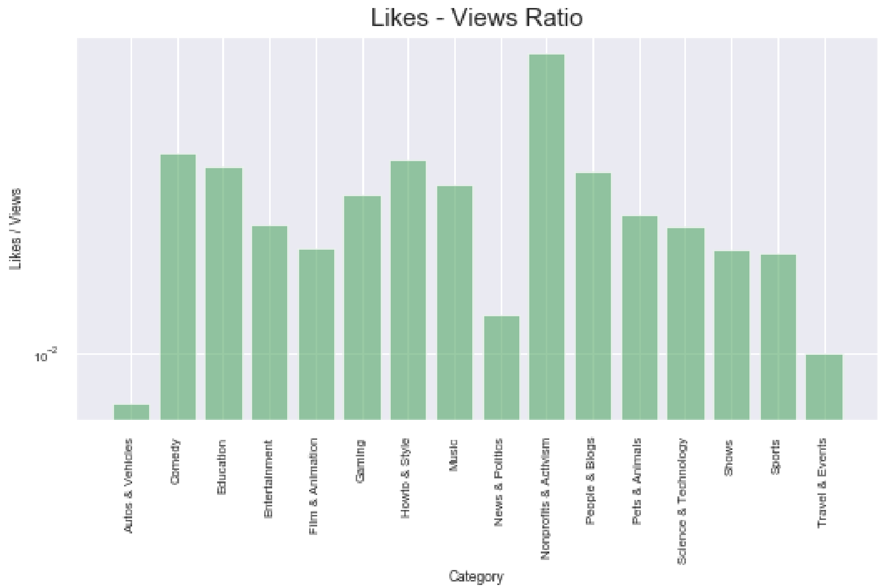
Question #12 : What are the ratios across different categories? (optional)



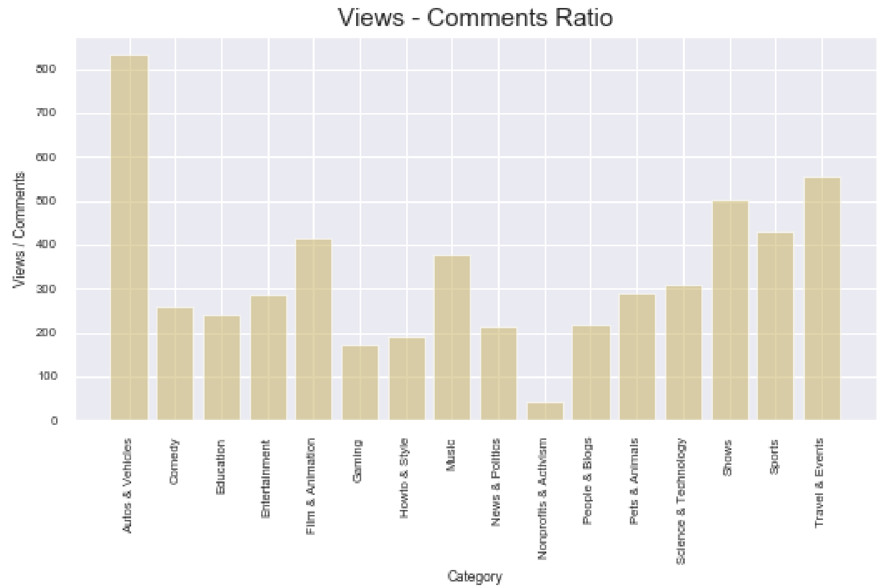
Likes - Dislikes Ratio



Dislikes - Views Ratio



Likes - Views Ratio



Views - Comments Ratio

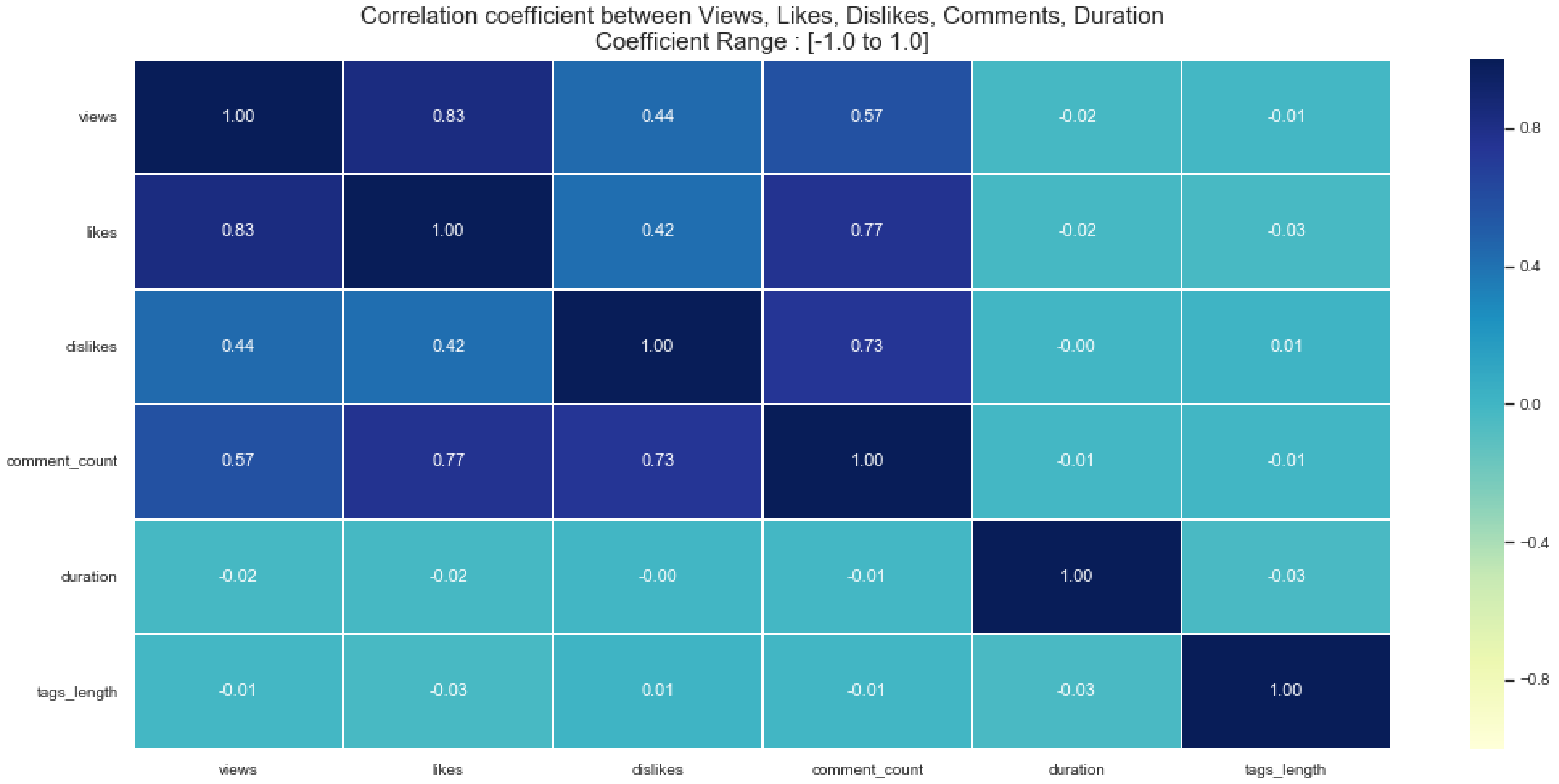
Correlations

“One of the first things taught in introductory statistics textbooks is that correlation is not causation. It is also one of the first things forgotten.”

Thomas Sowell

Correlations

Question #13 : Is there any correlation between views, likes, dislikes, comments, duration, or tags ?



Correlation coefficient between Views, Likes, Dislikes, Comments, Duration
Coefficient Range : [-1.0 to 1.0]

	views	likes	dislikes	comment_count	duration	tags_length
views	1.00	0.83	0.44	0.57	-0.02	-0.01
likes	0.83	1.00	0.42	0.77	-0.02	-0.03
dislikes	0.44	0.42	1.00	0.73	-0.00	0.01
comment_count	0.57	0.77	0.73	1.00	-0.01	-0.01
duration	-0.02	-0.02	-0.00	-0.01	1.00	-0.03
tags_length	-0.01	-0.03	0.01	-0.01	-0.03	1.00

Stronger correlations :

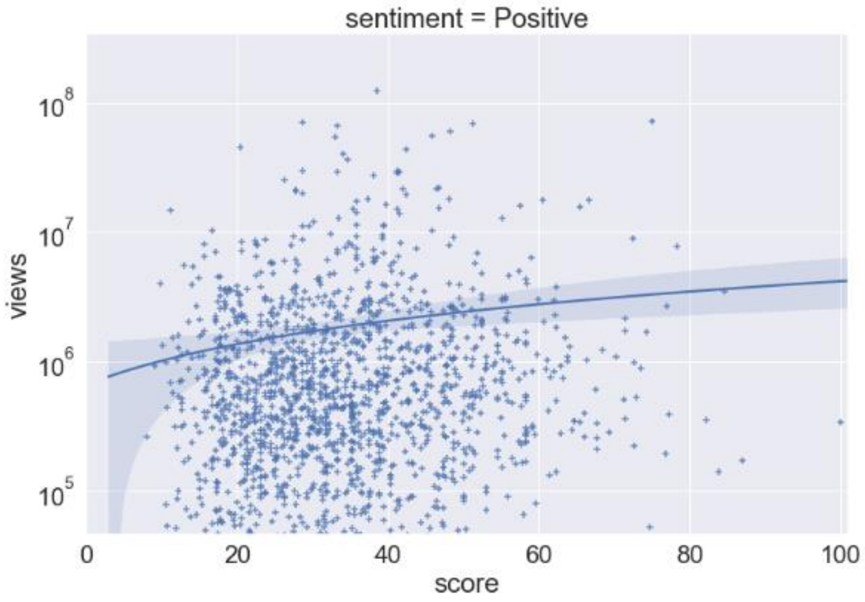
- Views and Likes
- Likes and comments
- Dislikes and comments

Weaker correlations :

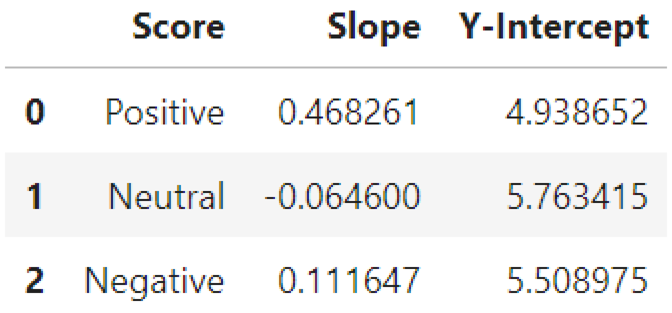
- Video duration
- Tags

Correlations

Question #14 : Is there any correlation between Title Sentiment and Views ?

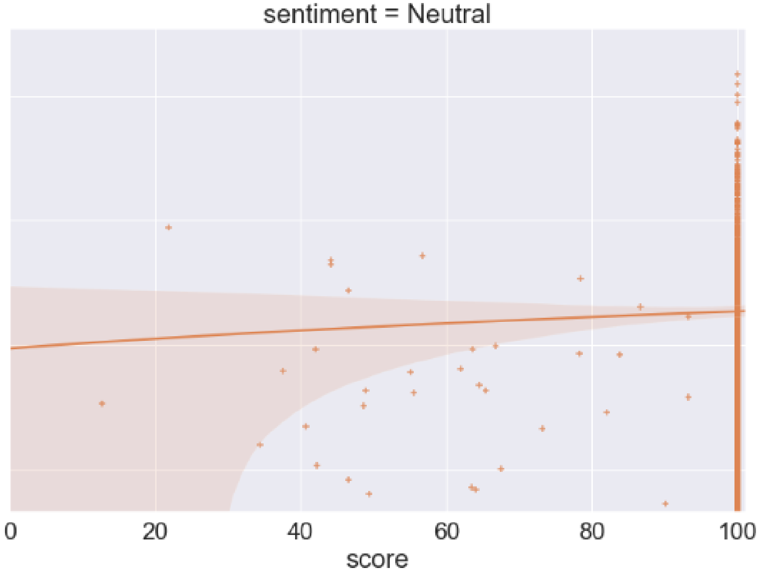


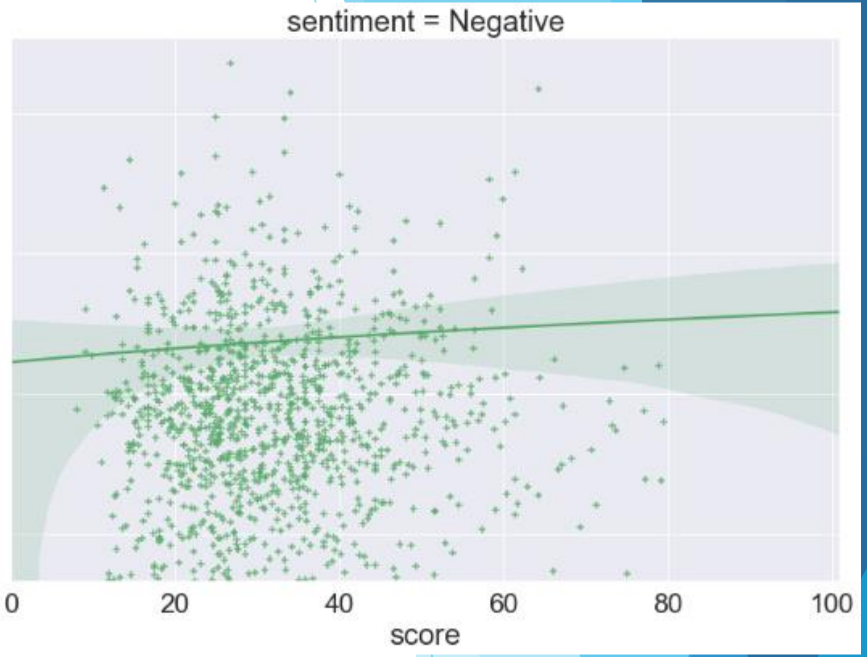
A positive correlation can be observed when the title sentiment is positive.



	Score	Slope	Y-Intercept
0	Positive	0.468261	4.938652
1	Neutral	-0.064600	5.763415
2	Negative	0.111647	5.508975

Hard to tell if there is any correlation



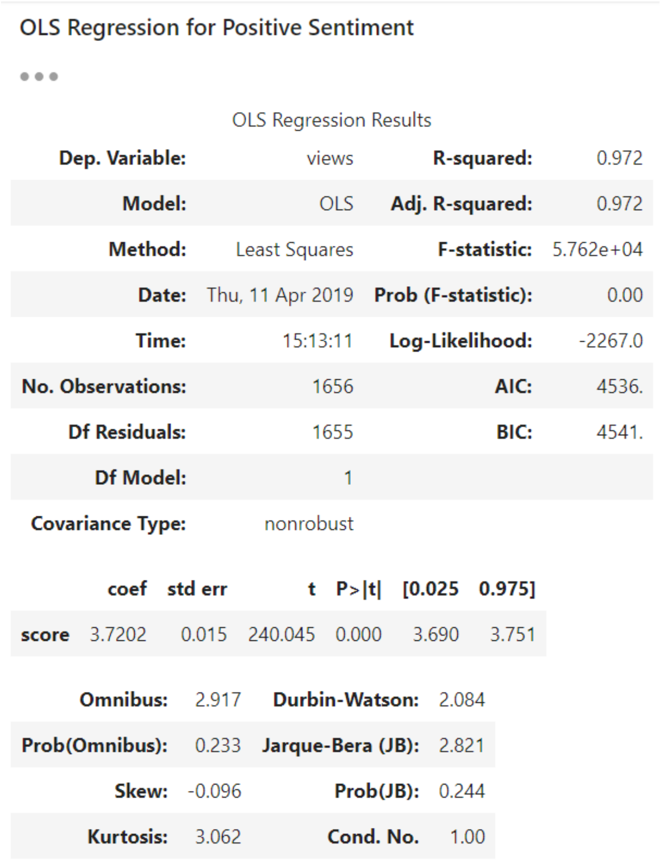


A positive correlation can be observed when the title sentiment is negative.

Correlations

Question #14 : Is there any correlation between Title Sentiment and Views ?

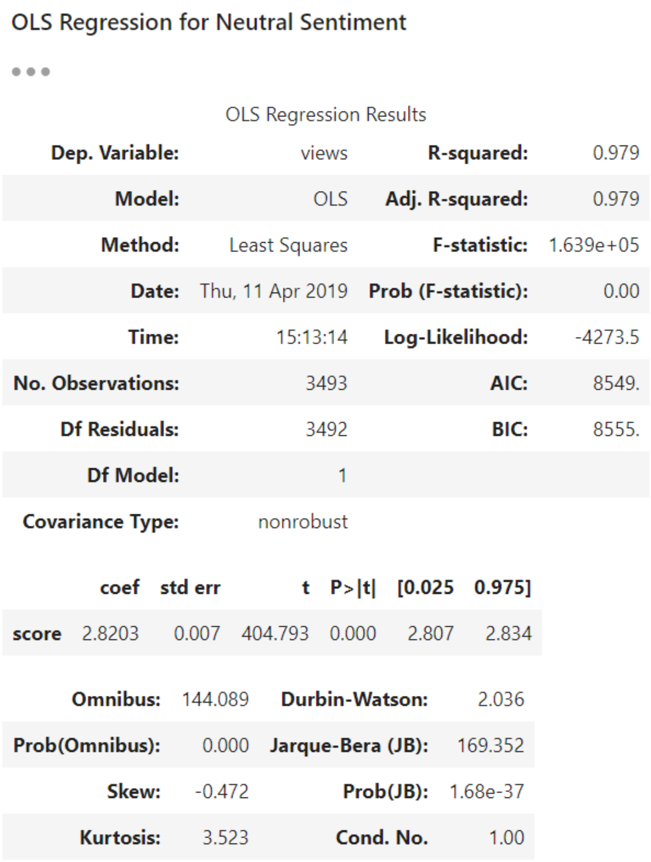
What does OLS Regression test tell us?



OLS Regression for Positive Sentiment

•••

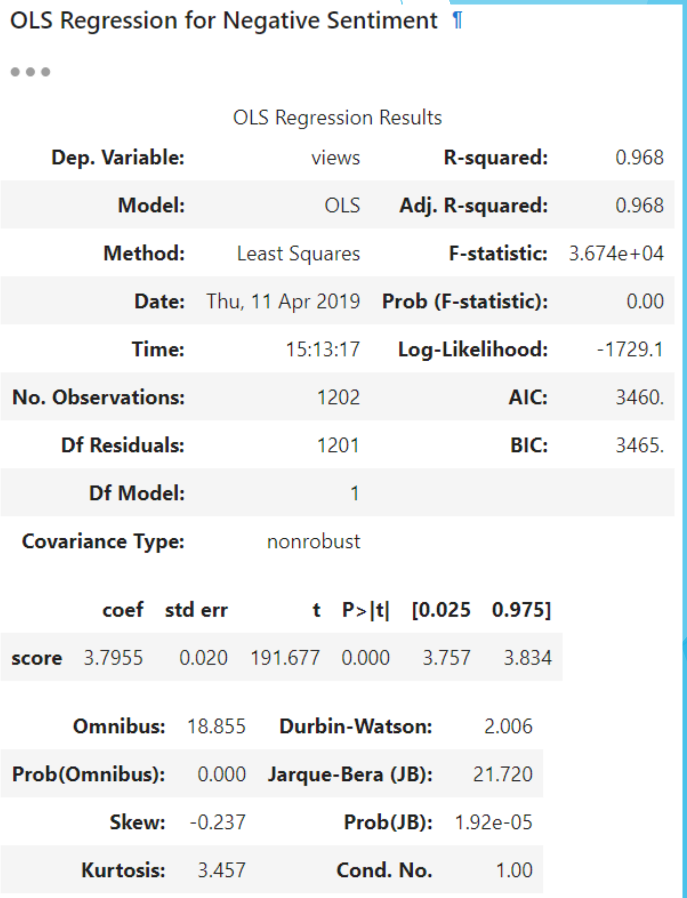
```
                          OLS Regression Results
    Dep. Variable:             views       R-squared:      0.972
           Model:                OLS   Adj. R-squared:      0.972
          Method:      Least Squares      F-statistic:  5.762e+04
            Date:   Thu, 11 Apr 2019 Prob (F-statistic):     0.00
            Time:           15:13:11   Log-Likelihood:    -2267.0
No. Observations:               1656              AIC:      4536.
    Df Residuals:               1655              BIC:      4541.
        Df Model:                  1
 Covariance Type:          nonrobust

         coef   std err       t     P>|t|   [0.025   0.975]
score  3.7202     0.015   240.045   0.000    3.690    3.751

       Omnibus:  2.917   Durbin-Watson:  2.084
Prob(Omnibus):  0.233  Jarque-Bera (JB):  2.821
          Skew: -0.096         Prob(JB):  0.244
      Kurtosis:  3.062        Cond. No.   1.00
```



OLS Regression for Neutral Sentiment

•••

```
                          OLS Regression Results
    Dep. Variable:             views       R-squared:      0.979
           Model:                OLS   Adj. R-squared:      0.979
          Method:      Least Squares      F-statistic:  1.639e+05
            Date:   Thu, 11 Apr 2019 Prob (F-statistic):     0.00
            Time:           15:13:14   Log-Likelihood:    -4273.5
No. Observations:               3493              AIC:      8549.
    Df Residuals:               3492              BIC:      8555.
        Df Model:                  1
 Covariance Type:          nonrobust

         coef   std err       t     P>|t|   [0.025   0.975]
score  2.8203     0.007   404.793   0.000    2.807    2.834

       Omnibus: 144.089   Durbin-Watson:    2.036
Prob(Omnibus):    0.000  Jarque-Bera (JB): 169.352
          Skew:  -0.472         Prob(JB):  1.68e-37
      Kurtosis:   3.523        Cond. No.     1.00
```



OLS Regression for Negative Sentiment ¶

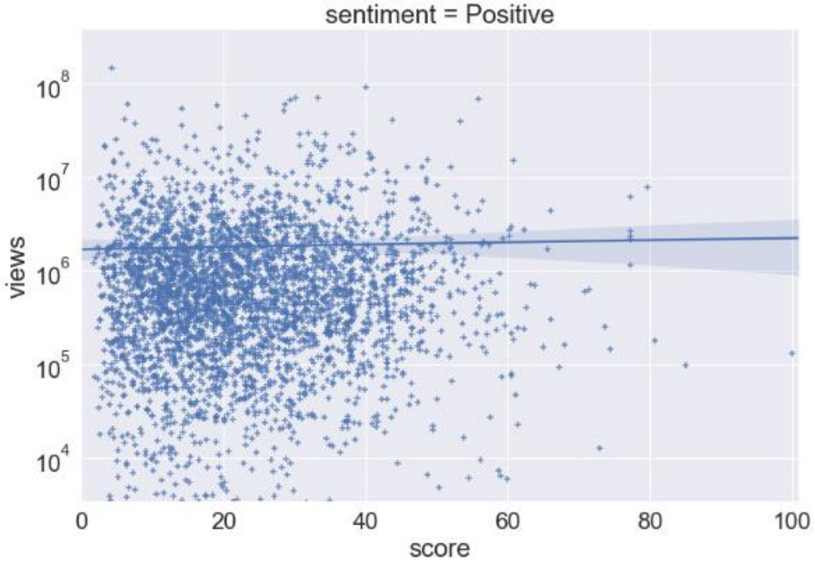
•••

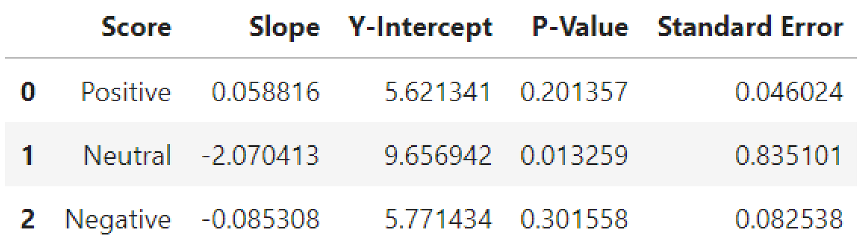
```
                          OLS Regression Results
    Dep. Variable:             views       R-squared:      0.968
           Model:                OLS   Adj. R-squared:      0.968
          Method:      Least Squares      F-statistic:  3.674e+04
            Date:   Thu, 11 Apr 2019 Prob (F-statistic):     0.00
            Time:           15:13:17   Log-Likelihood:    -1729.1
No. Observations:               1202              AIC:      3460.
    Df Residuals:               1201              BIC:      3465.
        Df Model:                  1
 Covariance Type:          nonrobust

         coef   std err       t     P>|t|   [0.025   0.975]
score  3.7955     0.020   191.677   0.000    3.757    3.834

       Omnibus: 18.855   Durbin-Watson:    2.006
Prob(Omnibus):   0.000  Jarque-Bera (JB):  21.720
          Skew: -0.237         Prob(JB):  1.92e-05
      Kurtosis:  3.457        Cond. No.     1.00
```

- R-squared score > 0.9
- P-value score < 0.001

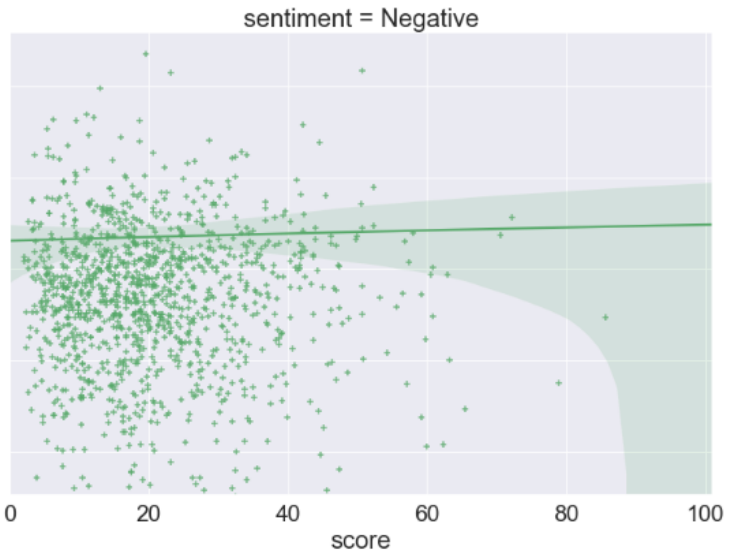
Correlations

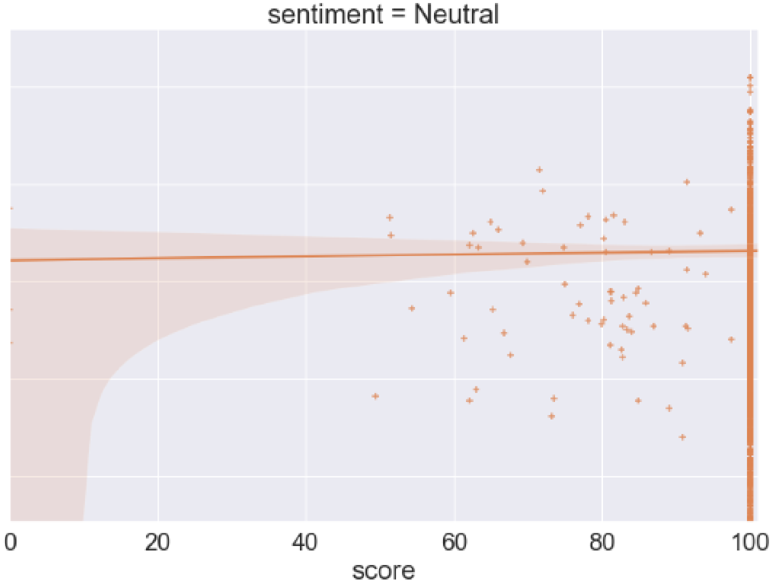
Question #15 : How about correlation between Tags Sentiment and Views ?


sentiment = Positive



	Score	Slope	Y-Intercept	P-Value	Standard Error
0	Positive	0.058816	5.621341	0.201357	0.046024
1	Neutral	-2.070413	9.656942	0.013259	0.835101
2	Negative	-0.085308	5.771434	0.301558	0.082538


sentiment = Negative

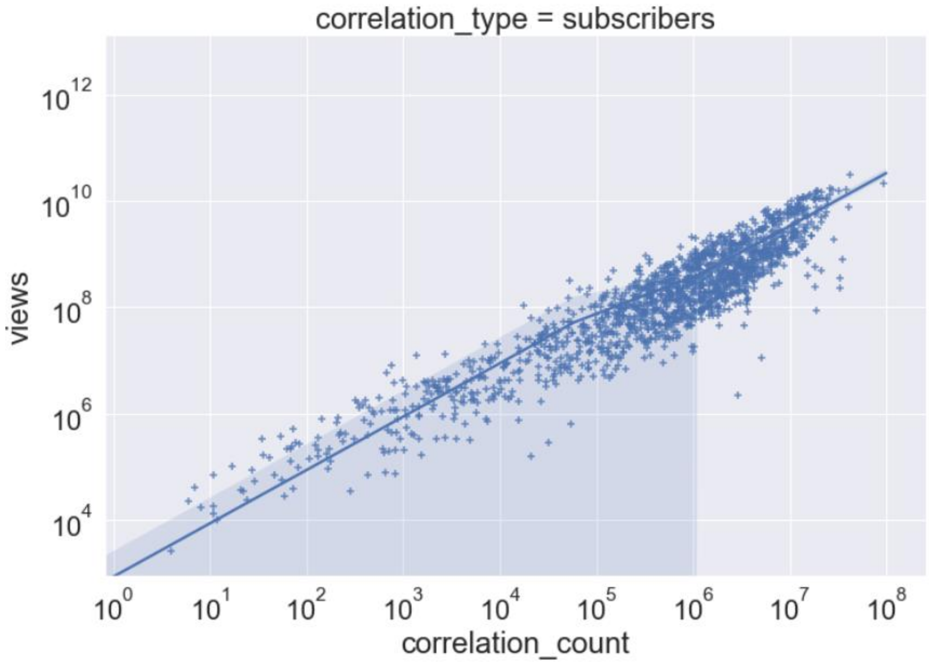

sentiment = Neutral

Not satisfactory results.

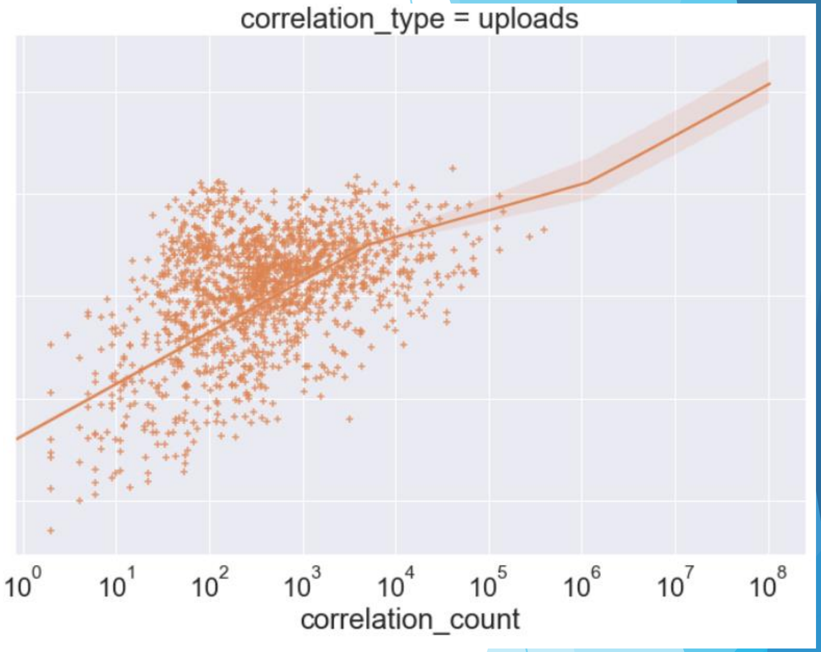
Hypothesis rejected.

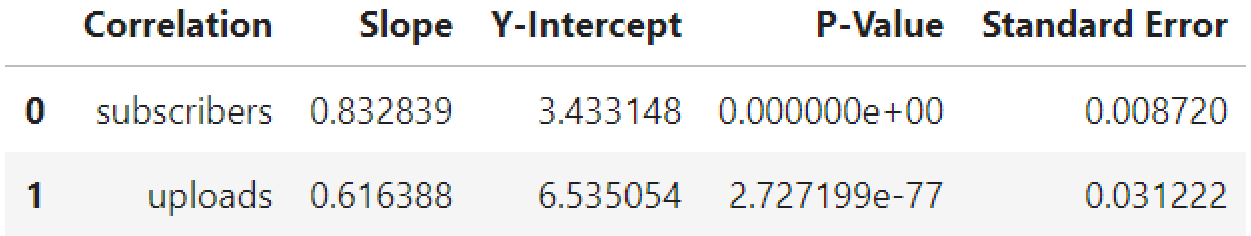
Correlations

Question #16 : Is there any correlation between Subscribers, Uploads, and Views?



Expected results.





	Correlation	Slope	Y-Intercept	P-Value	Standard Error
0	subscribers	0.832839	3.433148	0.000000e+00	0.008720
1	uploads	0.616388	6.535054	2.727199e-77	0.031222

Making the call

“The hardest choices require the strongest wills.”

Thanos

Avengers: Infinity War

Making the call

YouTube definitely can provide the means to conquer our main goal which is by increasing a product brand's visibility. Using the YouTube platform we may gain the potential to reach 518 thousands of people on average. In order to maximize the potential, we propose the following deployment plan:

Product’s brand identity

- Register a channel on YouTube and choose a channel name that correlates the name with the product brand.
- Create videos no longer than 5 min to explain everything related to the product and the target public. Adjust storyboarding over time.
- Choose the right tags for your videos based on the product's category. YouTube search algorithm relies on heavily in keywords for the video search.
- Choose carefully the words for the video titles. There is a strong correlation between title sentiment and number of visualizations.
- Publish videos between March and June. The average time for a video be marked and trend takes on average 20 days.

Channel boost

- Invest in YouTube Video Advertising Campaign. Using popular channels such as Music and Entertainment has the potential to produce good results on ROI scores.
- Boosting your channel you also boots your brand identity and customer.
- Use the same formula for other social platforms such as Facebook, Twitter, etc. All these platforms are somehow connected.
- As the number of subscribers and uploads increase over time, there is a strong correlation showing that the number of views will also increase.

Q & A

- Before we open the Q & A session, let’s review the outlier of the year for YouTube trending dataset:

Childish Gambino
This Is America (Official Video)

Did you have any doubt that the Music category will be the outlier of the year?

Q & A

Thank you all for your attention!