MULTIDISCIPLINARY DESIGN REPORT

On

STATISTICAL ANALYSIS ON TRENDING YOUTUBE VIDEOS

Submitted in partial fulfillment of

Degree of Bachelor of Technology in Computer Science & Engineering

Submitted by
Batch - 17

Name: AAKASH S	**Name: SB SRIDHAR**	**Name: G SANTHOSH**
Reg. No: RA1711003040077	**Reg. No: RA1711003040132**	**Reg. No: RA1711003040086**

Submitted to

B.S. VIDHYASAGAR B.E., M.E.,(Ph.D)

Asst.Professor (O.G)

Department of Computer Science

DEPT. OF COMPUTER SCIENCE & ENGINEERING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

VADAPALANI CAMPUS, CHENNAI

BONAFIDE CERTIFICATE

Certified that this project report "**STATISTICAL ANALYSIS OF TRENDING YOUTUBE VIDEOS"** is the bonafide work of the following three students:

AAKASH S (Reg.No. : RA1711003040077)

SB SRIDHAR (Reg.No. : RA1711003040132)

G SANTHOSH (Reg.No. : RA1711003040086)

who carried out the project work under my supervision.

SIGNATURE OF THE GUIDE

B.S.VIDHYASAGAR B.E.,M.E.,(Ph.D)

Asst.Professor(O.G.)

Department of Computer Science

SRM Institute of Science and Technology

SIGNATURE OF THE HOD

Dr.S Prasanna Devi,

B.E.,M.E.,Ph.D.,PGDHRM.,

PDF(IISc)

Professor

Department of Computer Science

SRM Institute of Science and Technology

ACKNOWLEDGEMENT

We are grateful to our beloved Chancellor, **Dr.T.R.Pachamuthu**, SRM IST, for providing us with requisite infrastructure throughout the course.

We would like to extend our gratitude and heartfelt thanks to our respected Dean, **Dr.K.Duraivelu,**B.E, MBA, M.E., Ph.D. for supporting us.

We want to thank our Head of the Department, **Dr.S.Prasanna Devi**, B.E., M.E., PhD, PGDHRM, PDF(IISc) for her constant guidance and support.

Finally, we would like to thank **B.S. VIDHYASAGAR B.E., M.E.,(Ph.D)**,Assistant Professor, CSE for all her support and guidance. Without her help, it would have been really hard for us to accomplish our ideas and finish the project.

TABLE OF CONTENTS

LIST OF FIGURES

Trending Youtube Videos Analysis

(Daily statistics for tending videos on youtube)

Introduction

YouTube is the most famous and most utilized video platform on the planet today. YouTube has a rundown of trending videos that is refreshed always. Here we will utilize Python with certain bundles like Pandas to break down a dataset that was gathered. For every one of those days, the dataset contains information about the trending videos of that day. It contains information about more than 40,000 trending videos. We will break down this information to get bits of knowledge into YouTube trending videos, to perceive what is normal between these videos. Those bits of knowledge may likewise be utilized by individuals who need to build prevalence of their recordings on YouTube.

Abstract

Unlike popular videos, which would have already achieved high view numbers by the time they are declared popular, YouTube trending videos represent content that targets all viewers' attention over a relatively short time, and has the potential of becoming famous. Despite the importance and visibility, YouTube trending videos have not been studied or analyzed thoroughly. In this paper, we present our analysis on measuring, analyzing, and comparing key aspects of YouTube trending videos. Our study is based on collecting and monitoring high-resolution time-series of the views and other related statistics of YouTube videos.

Objective

Analyze Trending Videos from Youtube.

-How many views do our trending videos have? Do most of them have a large number of views? Is having a large number of views required for a video to become trending?

-Which video remained the most on the trending-videos list?

-How many trending videos contain a fully-capitalized word in their titles?

-What are the lengths of trending video titles? Is this length related to the video becoming trendy?

-How are views, likes, dislikes, comment count, title length, and other attributes correlate with (relate to) each other? How are they connected?

-What are the most common words in trending video titles?

-Which video category (e.g. Entertainment, Gaming, Comedy, etc.) has the largest number of trending videos?

-When were trending videos published? On which days of the week? at which time of the day?

Objective

Unlike popular videos, which would have already achieved high view numbers by the time they are declared popular, YouTube trending videos represent content that targets all viewers' attention over a relatively short time, and has the potential of becoming famous. Despite the importance and visibility, YouTube trending videos have not been studied or analyzed thoroughly.

Possible uses from this dataset could include:

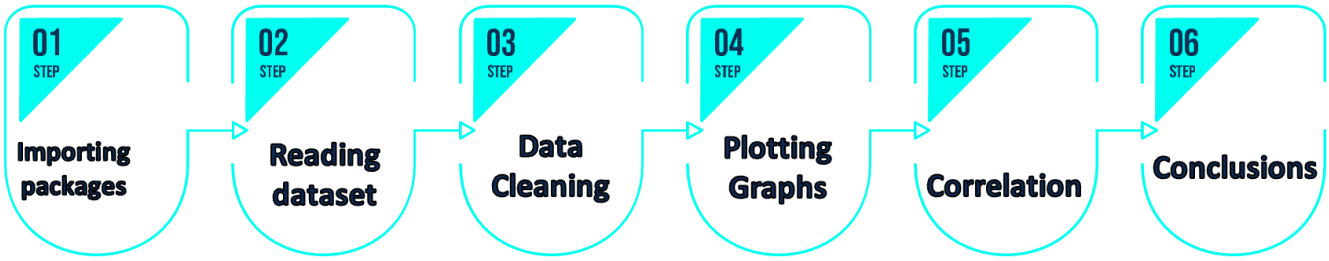
-Sentimental analysis in a variety of forms

-Categorising YouTube videos based on their comments and statistics

-Analysing what factors affect how popular a YouTube video will be

-Statistical analysis over time

System Architecture



Process

-Importing some packages

-Reading the dataset

-Getting a feel of the dataset

-Data cleaning

-Dataset collection years

-Describtion of numerical columns

-Views histogram

-Likes histogram

-Comment count histogram

Process

-Views histogram

-Likes histogram

-Comment count histogram

-Description on non-numerical columns

-How many trending video titles contain capitalized word?

-Video title lengths

-Correlation between dataset variables

-Which video category has the largest number of trending videos?

-Trending videos and their publishing time

-Conclusions

Packages & Modules

Scikit-learn (sklearn): Contains a number of state-of-the-art machine learning algorithms that were used throughout this project. Scikit-learn is considered the most prominent Python library for machine learning and depends on two other Python packages, Numpy and SciPy (5).

OS: The OS module in Python provides a way of using operating system dependent functionality. The functions that the OS module provides allows you to interface with the underlying operating system that Python is running on - be that Windows, Mac or Linux (3).

Pandas: This is a Python library for data wrangling and analysis. Pandas provides a great range of methods to modify and operate on this table; in particular, it allows SQL-like queries and joins of tables (5). This library is mainly used in data pre-processing.

Numpy: This is one of the fundamental packages for scientific computing in Python. It contains functionality for multidimensional arrays, high-level mathematical functions such as linear algebra operations. In scikit-learn, the Numpy array is the fundamental data structure (5).

Scipy: This is a collection of functions for scientic computing in Python. It provides, among other functionalities, advanced linear algebra routines, mathematical functions, and statistical distributions (5).

Matplotlib: This is considered the primary scientific plotting library in Python. It provides functions for making publication-quality visualizations such as line charts, histograms, scatter plots and so on (5).

Seaborn: This is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures. Seaborn aims to make visualization a central part of exploring and understanding data. It's dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally

perform the necessary semantinc mapping and statistical aggregation to produce informative plots (6). Seaborn's functionality include but is not limited to:

- A dataset-oriented API for examining relationships between multiple variables.
- Specialized support for using categorical variables to show observations or aggregate statistics.
- Options for visualizing univariate or bivariate distributions and for comparing them between subsets of data.

Itertools: Collection of tools for handling iterators. Simply put, iterators are data types that can be used in a for loop. The most common iterator in Python is the list(9).

```
#Importing Dependencies
import os

import sys
print("Python version: {}".format(sys.version))

# Importing panas
import pandas as pd
print("Pandas version: {}".format(pd.__version__))

# Importing numpy
import numpy as np
print("Numpy version: {}".format(np.__version__))

# Importing matplotlib for plotting
import matplotlib as plt
print("MatplotLib version: {}".format(plt.__version__))

# importing seaborn for plotting
import seaborn as sns
sns.set_style('whitegrid')
sns.set()
print("Seaborn version: {}".format(sns.__version__))
```

Hardware/Software Requirements

Windows System Requirements:-

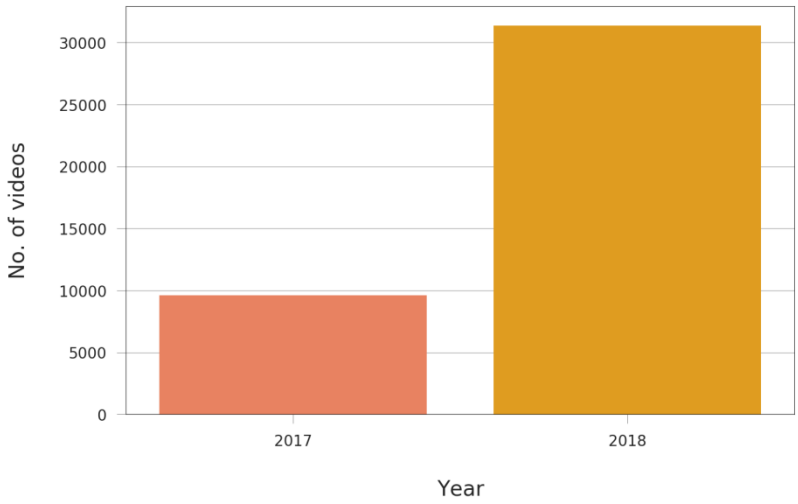
- **RAM**: 8 GB
- **Processor**: Intel® Core i5 1.6 Ghz or AMD Ryzen 5, with SSE2 technology.
- **Hard Disk Space**: 10 GB (minimum) free space available.
- **Screen Resolution**: 1024 x 768 or higher.

Linux System Requirements

- **Operating System:** Ubuntu 14.04 or 16.04, CentOS 6.9 or 7.5, RHEL 6.9 or 7.5.
- **RAM:** 8 GB.
- **Processor:** Intel® Core i5 1.6 Ghz or AMD Ryzen 5, with SSE2 technology.
- **Hard Disk Space:** 10 GB (minimum) free space available.

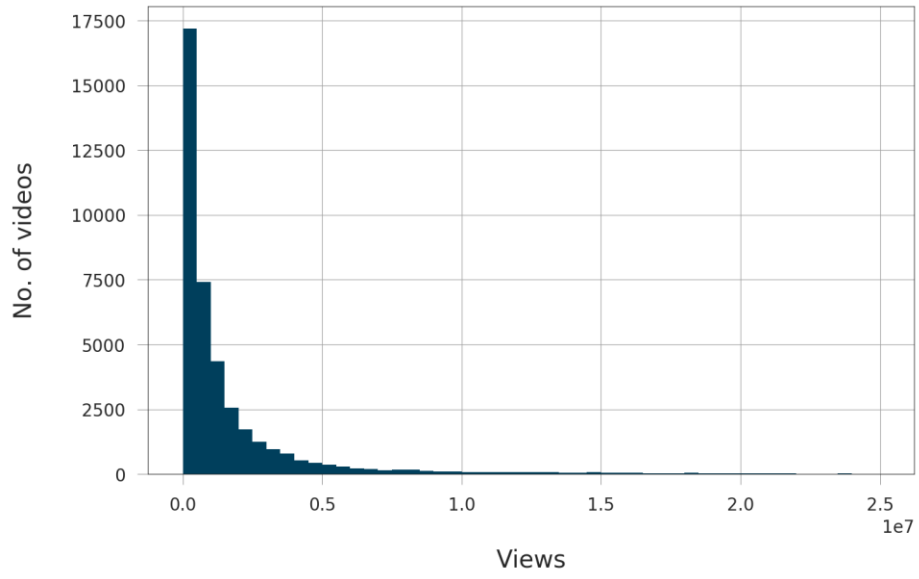
Data Collection Years

The above dataset used has the collection of trending videos from the year 2017-18 and the graph is plotted between the Year and the No of videos.



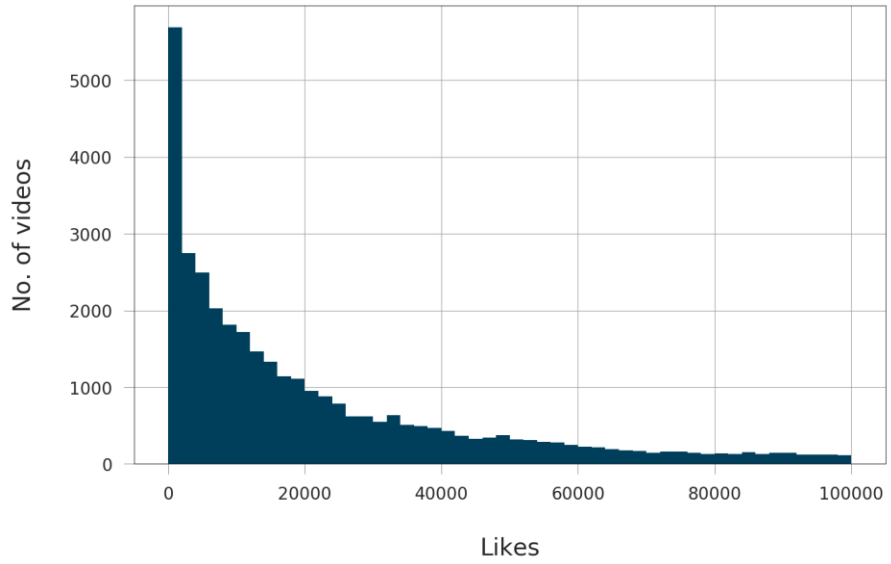
Views Histogram

Plotting a graph to analyze between the Views and the No of videos uploaded in YouTube uploaded in 2017-18 to take a look at its distribution:



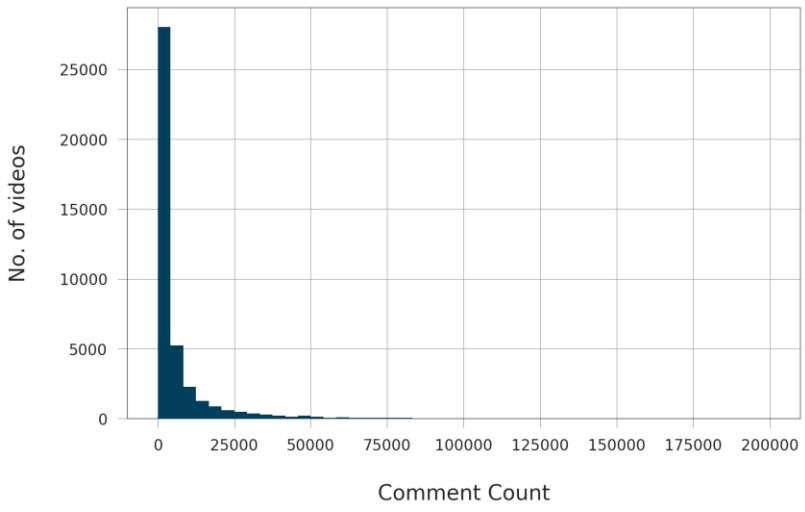
Likes Histogram

Plotting a graph to analyze between the Likes and the No of videos uploaded in YouTube uploaded in 2017-18 to take a look at its distribution:



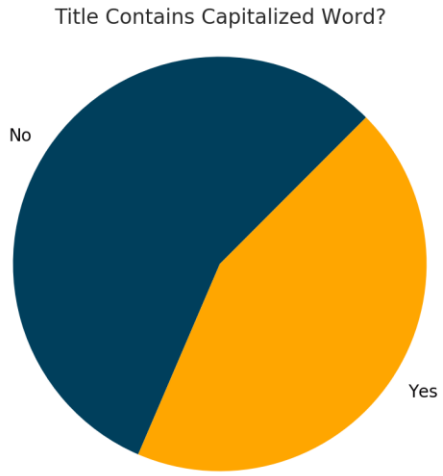
Comment Count Histogram

Plotting a graph to analyze between the Comment Count and the No of videos uploaded in YouTube uploaded in 2017-18 to take a look at its distribution:



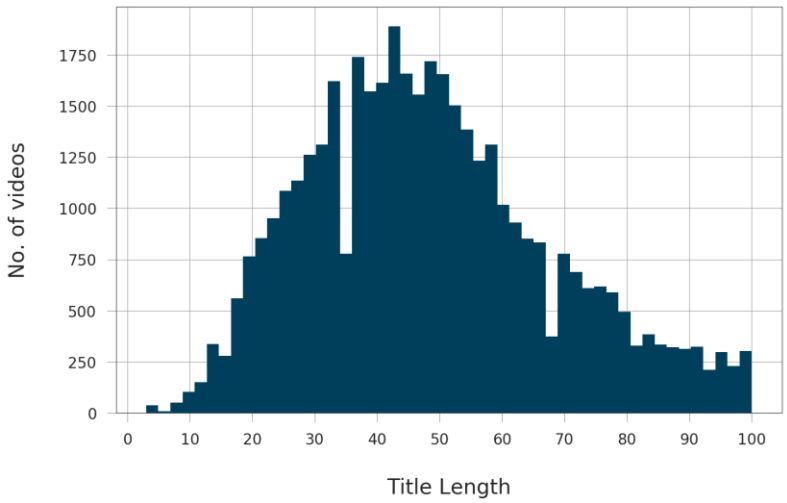
How many trending video titles contain capitalized word?

Now we want to see how many trending video titles contain at least a capitalized word. To do that, we will add a new variable to the dataset whose value is True if the video title has at least a capitalized word in it, and False otherwise

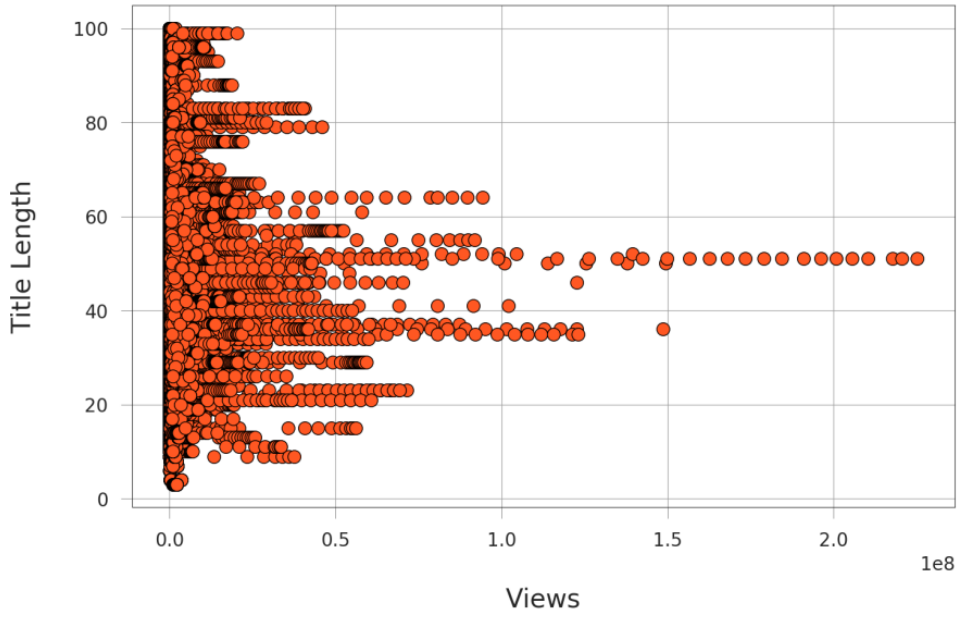


Video title lengths

Let's add another column to our dataset to represent the length of each video title, then plot the histogram of title length to get an idea about the lengths of trending video titles



We can see that title-length distribution resembles a normal distribution. Now let's draw a scatter plot between title length and number of views to see the relationship between these two variables

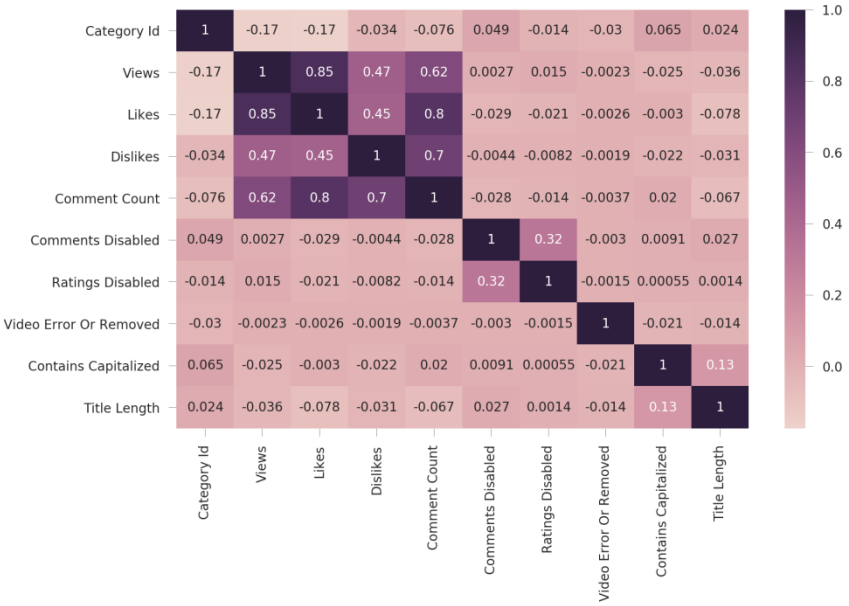


By looking at the scatter plot, we can say that there is no relationship between the title length and the number of views

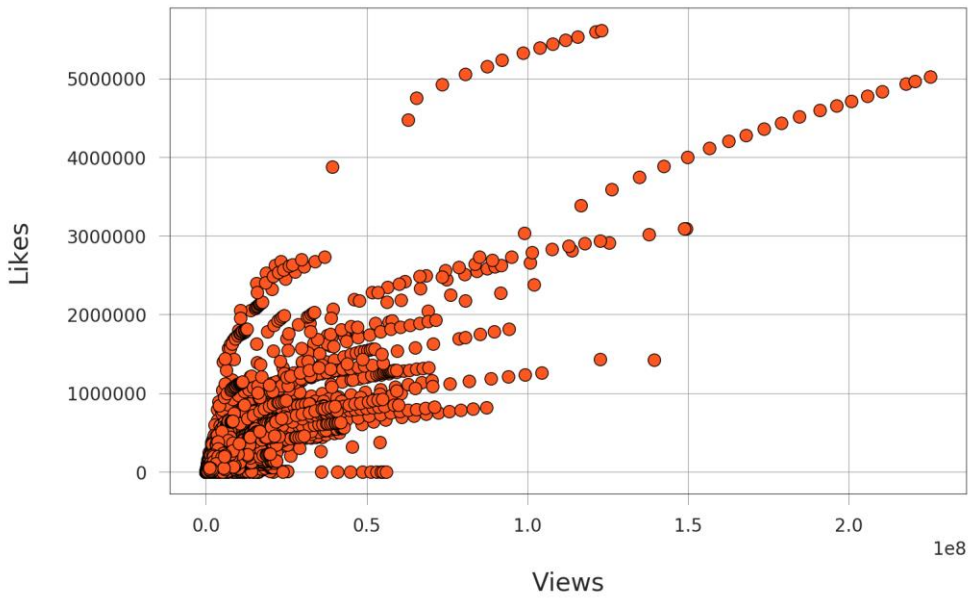
Correlation between dataset variables

Now let's see how the dataset variables are correlated with each other: for example, we would like to see how views and likes are correlated, meaning do views and likes increase and decrease together? Does one of them increase when the other decrease and vice versa ? Or are they not correlated?

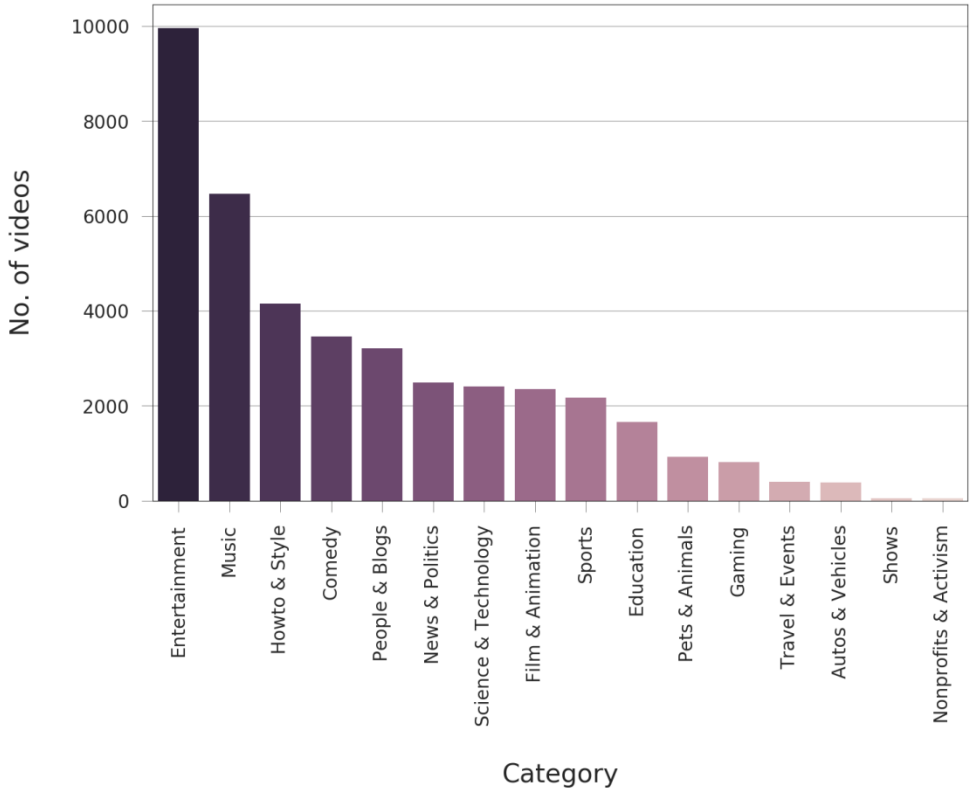
Correlation is represented as a value between -1 and +1



There is some positive correlation between views and dislikes, between views and comment count, between likes and dislikes. The above say that views and likes are highly positively correlated. Let's verify that by plotting a scatter plot between views and likes to visualize the relationship between these variables

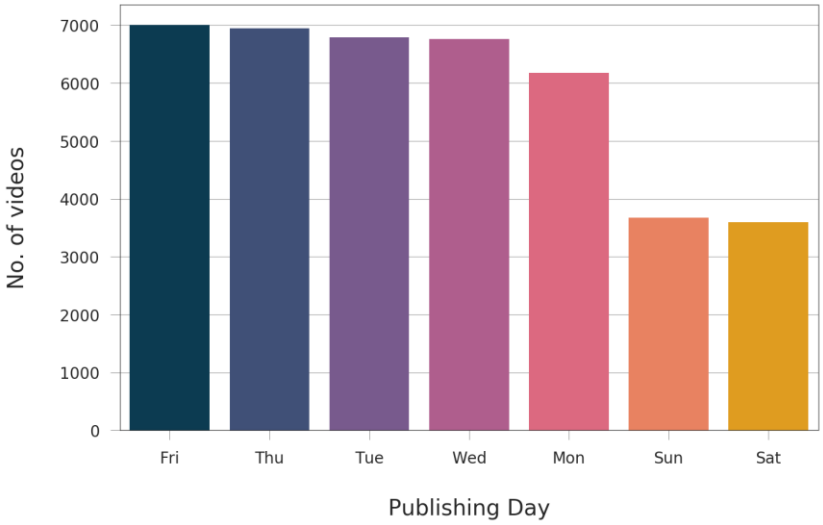


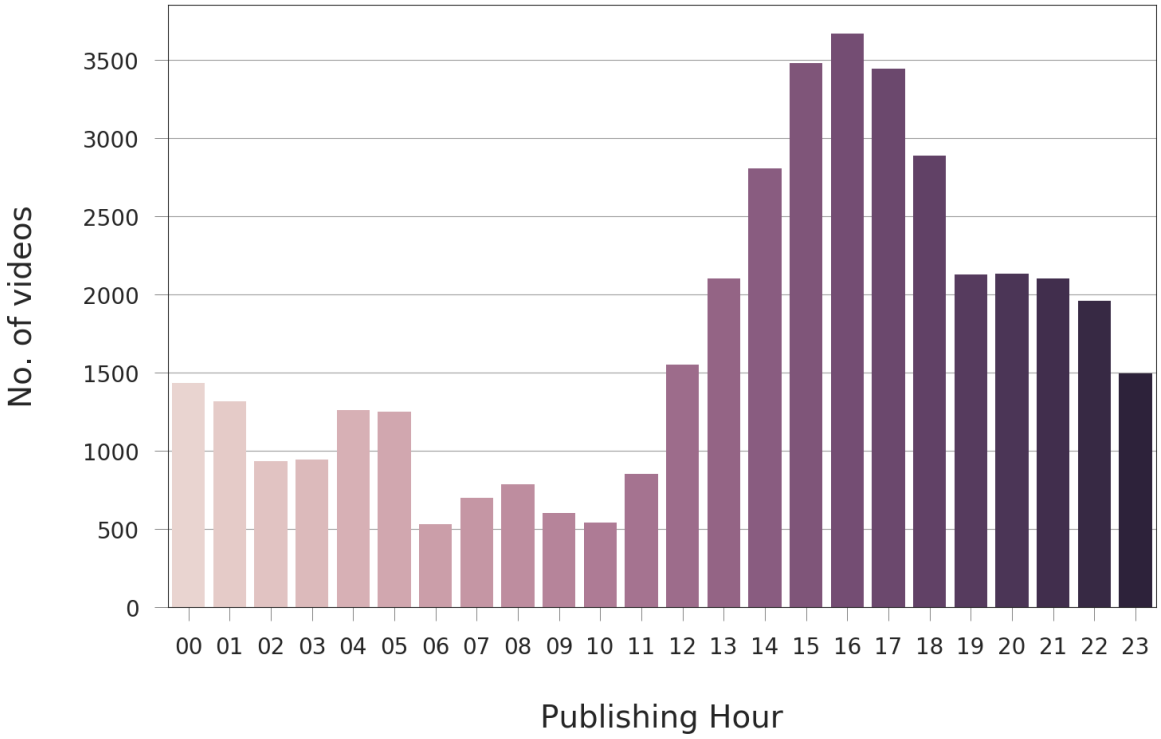
Which video category has the largest number of trending videos?



Category plays a major role in determining the trending videos. We see that the Entertainment category contains the largest number of trending videos among other categories

Trending Videos and their publishing time





Creating Model

-The correlation analysis made in the previous section resulted in likes and views being the two most correlated hyperparameters and therefore we can infer that they also make the most impact on the status of a video being kept as trending.

-Since we already have ratios of views, next we will also have ratios of likes. I didn't create label encods for the category_id columns simply because they were not very highly correlated with any other variables in the original dataset.

-Even certain categories such as music videos might have dominated the dataset, this simply implies that people like to watch music videos on Youtube but without necessarily having an impact on whether it keeps trending. The machine learning models used will only be Linear Models.

```
from sklearn import tree
from sklearn.tree import DecisionTreeRegressor
from sklearn import metrics
from sklearn.linear_model import LinearRegression
```

-In order to make sure that the machine learning models don't simply become good at memorizing the the introduced data, the final dataset will have to be split into two : 1) Training Data: Used to build the machine learning model. 2) Testing Data: Used to assess how well the model works.

```
from sklearn.model_selection import train_test_split
```

Conclusions and Applications

Nowadays, an increasing number of people post their videos on YouTube, and it is interesting to know whether a video is popular is dependent on its category and the cultural background of viewers. We plan to use the acquired dataset to analyze the composition and popularity associated with different factors of online videos on YouTube. We'd like to dig in deeper to elaborate the relationship between them. To be more specific, some videos are highly controversial because of its content or types. Also, we want to show how the cultural divergence affects people's likes and the overall most popular video types to shed light on how YouTubers are supposed to refine their videos to get more subscribers, and recommend popular channels of particular video genres.

Here are the some of the conclusions we extracted from the analysis:

We analyzed a dataset that contains information about YouTube trending videos The dataset was collected in the year 2017-18. It contains about 40,000 video entries. Some videos may appear on the trending videos list on more than one day. The There is a strong positive relation between the number of views and the number of likes of trending videos: As one of them increases, the other increases, and vice versa. There is a strong positive correlation also between the number of likes and the number of comments, and a slightly weaker one between the number of dislikes and the number of comments.

This Analysis helps for the people who choose Youtube as their carrer where they can prepare themselves on what basis Youtube makes a video famous and add in the trending list. Hence prepare their videos on the famous contents out there in the world.

Since the Trending Youtube Video dataset includes only top trending videos, there is a potential for survivor bias in the results. Videos with similar characteristics, that did not trend, are not included in the dataset and thus will not be studied.

For convenience, we assume that frequent words in each category has higher impact on the number of views. However, some less frequent words that appear in only a few videos might also have great impact (by the assumption of tf-idf). In our study, we only focus on more general cases that covers the majority videos of the category.

Regarding future work, our group will try more advanced techniques on video statistics analysis, such as more regression models and learning algorithms. We will also employ more language models in the realm of Natural Language Processing, such as word2vec, topic models and tf-idf. Besides working on numerical and string-based data, we will extend this analysis to image data by encoding thumbnails.

References

-Ammar Alyousfi (Kernel Author) using Python programming for determining the Trending Youtube Videos and analysis of it.

https://www.kaggle.com/ammar111/youtube-trending-videos-analysis

-Youtube information API documentation: https://www.researchgate.net/deref/https%3A%2F%2Fdevelopers.google.com%2Fyoutube%2F2.0%2Freference

-Youtube Statistics: https://www.researchgate.net/deref/http%3A%2F%2Fwww.youtube.com%2Fyt%2Fpress%2Fstatistics.html

-Trending YouTube Video Statistics:

https://www.kaggle.com/datasnaek/youtube-new/home

- The entire dataset contains 5 csv files and 5 json files(for 5 different countries), including various kind of information like video titles, channels, video categories, publish time, number of views, number of likes and dislikes, etc.

-Mitchell Jolly(Data Snaek) (2019). Trending YouTube Video Statistics Daily Statistics for Trending Youtube Videos, Retrieved from https://www.kaggle.com/datasnaek/youtube-new

-Author Unknown . scikit-learn

5.9. Transforming the prediction target (y), Retrieved from: https://scikitlearn.org/stable/modules/preprocessing_targets.html#preprocessing-targets

-Base Paper

https://www.researchgate.net/publication/266262149_Trending_Videos_Measurement_and_Analysis