

In []:

```
# libraries
import numpy as np
import pandas as pd
# !pip install altair;
import altair as alt
import datetime
# !pip install sklearn
# from sklearn.linear_model import LinearRegression
# from sklearn.metrics import r2_score
# from sklearn.preprocessing import add_dummy_feature

alt.data_transformers.disable_max_rows();
```

PSTAT 100 Project Plan Report

Sections

[Background](#)

[Data description](#)

[Initial explorations](#)

[Planned work](#)

Group Information

Group members: Harleen Kaur (Member 1), Amelia Meyer (Member 2)

Contributions:

1. Member 1: Initial data tidying, Initial Explorations
2. Member 2: Background, Data Description, Exploratory Plots, Planned Work

Background: Trending Videos on Youtube (US)



According to Cloudflare, YouTube was the 8th most visited site, and the 3rd most visited social media site in 2021. [See this cbs article for more rankings..](#) With millions of users in the US and around the world, YouTube has developed its own gravitational force, pulling people in to watch content varying from makeup tutorials to [scuba-divers solving cold cases](#).

The most popular content on YouTube is referred to as 'trending' - a term that gained popularity due to Twitter's hashtags. Luckily for us, YouTube tracks their [top trending videos](#) which, according to Variety magazine, are measured by user interactions "(number of views, shares, comments and likes)".

The purpose of our project will be to analyze the top trending videos on YouTube in the US using a dataset of the daily record of the top trending videos. Our goals are to get a better understanding of the types of videos users are interested in. Given the influence of social media on everyday life, this can give us insight into the state of the world and general trends or interests of YouTube users in the US.

[View Youtube's Trending Feed](#)

[More on how YouTube influences the topics of today](#)

Data Description

Basic Information

General Description

The dataset we will be using for our analysis is the [YouTube Trending Video Dataset](#) which documents the daily record of the top trending videos on YouTube(updated daily).

Source

Our dataset was collected by Rishav Sharma and made available through Kaggle with for the purpose of user analysis. Rishav lists possible projects as sentiment analysis, categorization based on video comments and statistics, machine learning to generate comments, predicting popularity of new videos, and general analysis over time.

Collection Methods:

The data was collected using the [YouTube API](#) as provided by Google which uses web-scraping.

Sampling Design and Scope of Inference

We'll be focusing specifically on the dataset for top trending YouTube videos in the US. The dataset includes several months of data, with up to 200 listed trending videos per day. The sample population is all YouTube videos in the US. The sample frame is the top trending YouTube videos in the US. Our scope of inference will include videos that are available to users in the US.

Data Semantics and Structure

Units and Observations

The observational units are the trending videos. One unit is one video that appeared on the top trending YouTube videos list in the US.

Variable descriptions

Name	Variable description	Type	Units of measurement
title	video title	object or string	...
date_published	date video was published	object or string	...
channel_name	name of channel video was published on	object or string	...
category	category video falls under; genre of video	object or string	...
trending_date	date video was trending	object or string	...
tags	tags attached to video; video identifiers added by video creator	object or string	...
view_count	number of views video received	int64 or integers	...
likes	number of likes video received	int64 or integers	...
dislikes	number of dislikes video received	int64 or integers	...
comment_count	number of comments on video	int64 or integers	...
comments_disabled	whether the comments were disabled	bool or true/false	...
ratings_disabled	whether the ratings were disabled	bool or true/false	...
year_published	what year the vide was published	int64 or integers	4-digit year values
month_published	what month the year was published	int64 or integers	1 or 2-digit month value
year_trending	what year the video was trending	int64 or integers	4-digit year value
month_trending	what month the video was trending	int64 or integers	1 or 2-digit month value

Example rows

```
In [ ]: # load new csv
trending = pd.read_csv('/Users/harleen/Documents/GitHub/PSTAT100FinalProject/')

# print a few example rows of dataset in tidy format
trending.drop(columns=['Unnamed: 0'], inplace=True)

trending.head(4)
```

```
Out[ ]:
```

	title	date_published	channel_name	category	trending_date	
0	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11	Brawadis	People & Blogs	2020-08-12	brawadis prank ba
1	Apex Legends Stories from the Outlands – "Th...	2020-08-11	Apex Legends	Gaming	2020-08-12	Apex
2	I left youtube for a month and THIS is what ha...	2020-08-11	jacksepticeye	Entertainment	2020-08-12	
3	XXL 2020 Freshman Class Revealed - Official An...	2020-08-11	XXL	Music	2020-08-12	xxl freshn

Initial explorations

Basic properties of the dataset

Variable summaries

Our dataset consists of 16 columns and 113391 rows. There are no missing values in our dataset. Our resulting dataset consists of the variable: title, date_published, channel_name, category, trending_date, tags, view_count, likes, dislikes, comment_count, comments_disabled, ratings_disabled, year_published, month_published, year_trending, month_trending. Trending dates range from 2020-08-12 to 2022-02-25.

As we observed our dataset, we found some cool things. The most viewed and liked video is "Butter" by BTS (a iconic KPOP songs). In fact, we see KPOP music in the top charts quite often. Curious as to which genres have the most total views, we found that music, entertainment, and gaming tend to have a highest amount of view counts.

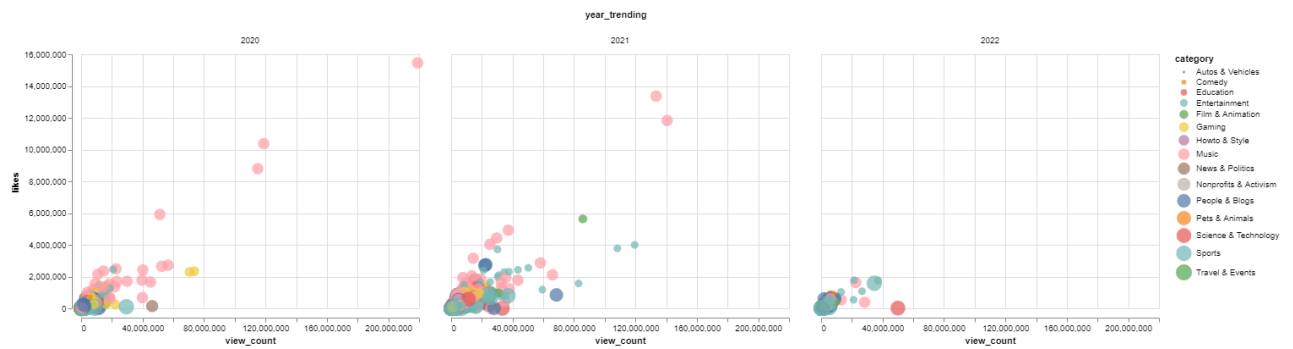
The two boolean variables in our dataset have the following allocations:

variable name	number observations with True	number of observations with False
comments_disabled	1739	112252
ratings_disabled	779	113212

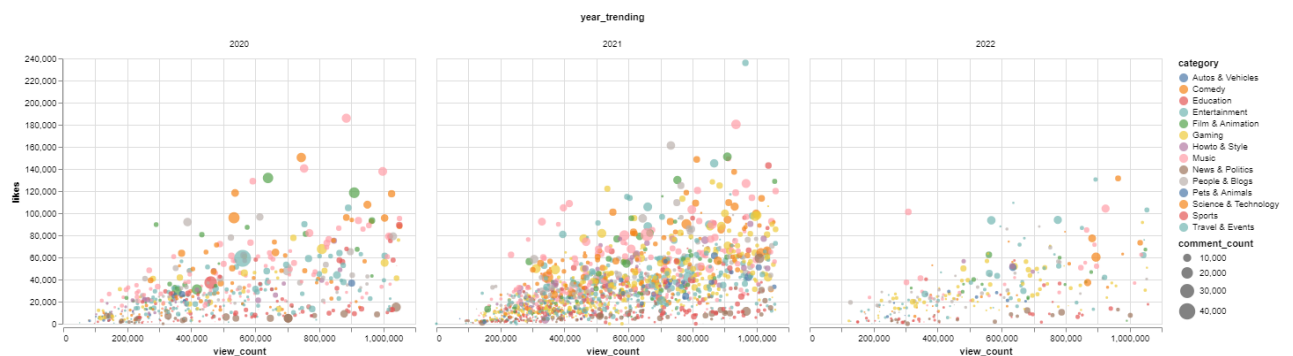
Our category variable can be represented as one of the following 15 categories:

category	number of observations
Nonprofits & Activism	88
Travel & Events	516
Pets & Animals	551
Autos & Vehicles	2073
Education	2724
Howto & Style	3599
Science & Technology	3859
News & Politics	4077
Film & Animation	4486
Comedy	6561
People & Blogs	10302
Sports	12204
Music	19558
Gaming	20437
Entertainment	22956

Exploratory analysis



The above chart panels are the number of views per video against the number of likes per video. Since our dataset is so large, we took a random sample of 5000 observations for the graphics here. We can see that the category with the most likes and views is music for 2020 and 2021, but is a toss-up between Sports and Science & Technology for 2022. We can expect our dataset to have more observations for 2020 and 2021 since we are still early on in 2022.



The above chart is a scaled plot of the number of views against the number of likes per video scaled to only look at the values that fell below the median so as to get a better view of the majority of the data.

Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel > Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Select *File > Download as > HTML*.
5. Open in Google Chrome and print to PDF on A3 paper in portrait orientation.
6. Submit to Gradescope