Trending Youtube Video Statistics (113 Countries)

Data source and collection:

This is an external data source, collected by a Kaggle user directly from the Youtube webpage. While not published by an official agency and not fully verifiable as trustworthy, the user describes their collection method as a daily query collection directly from the Youtube API, which should deem it reliable and authentic enough for the purposes of this exercise.

Data content:

The dataset includes basic information for the 50 most watched videos each day on the Youtube platform across 113 different countries, including daily ranking, movement trends, view counts, likes, comments, descriptions, key words used as "tags", publishing data and language, among other descriptive and numerical metrics.

Data limitations & ethics:

Considering the method of collection and the relatively low reliability of the source in comparison to, for example, a governmental agency, there is a possibility that the data is not fully accurate, and therefore we should exercise caution when deriving any conclusions from it. The size of each country as well as the popularity of the platform, their level of freedom of access to the internet and the regulations on what content can be accessed may be factors impacting the videos most watched in each country, which means that any insights obtained may not inform any deeper analysis on the preferences of the population at large.

Relevance:

The dataset fulfils the relevant criteria, it is up to date, contains more than 2,000 rows, at least 3 continuous variables and 3 categorical variables (including non-anonymized columns), a time-dependent variable (since the data is updated daily) and a geographical object in the format of the countries.

I have selected this data to work with due to a combination of a personal interest in the topic, as an avid consumer of content published on the platform, and a general curiosity for how different or similar the trends in consumption of varied internet/social media content could be across the world. I believe the tools used and the type of analysis conducted with this dataset would also be transferrable to other industries and contexts, since the structure of the dataset would be similar for a variety of businesses looking to analyse their products' popularity in different regions and the factors that might influence it.

Data Profile:

Variables	Description	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous	Data Type
title	Title of the YouTube video	Qualitative	Nominal	str
channel_name	Name of the YouTube channel	Qualitative	Nominal	str
daily_rank	Rank of the video on the respective day	Quantitative	Discrete	int
daily_movement	Change in the rank compared to the previous day	Quantitative	Discrete	int
weekly_movement	Change in the rank compared to the previous week	Quantitative	Discrete	int
snapshot_date	Date of the data collection	Qualitative	Ordinal	date
country	Country for which the trending videos are recorded	Qualitative	Nominal	str
view_count	Number of views for the video	Quantitative	Discrete	int
like_count	Number of likes for the video	Quantitative	Discrete	int
comment_count	Number of comments for the video	Quantitative	Discrete	int
description	Description of the video	Qualitative	Nominal	str
thumbnail_url	URL of the video thumbnail image	Qualitative	Nominal	str
video_id	ID of the video	Qualitative	Nominal	str
channel_id	ID of the YouTube channel	Qualitative	Nominal	str
video_tags	Tags associated with the video	Qualitative	Nominal	str
kind	Type of video	Qualitative	Nominal	str
publish_date	Date when the video was published	Qualitative	Ordinal	date
language	Language of the video	Qualitative	Nominal	str

Cleaning documentation

Dataset	Action
trending_yt_videos	52518 'nan' values found on the 'language' column. Ignored, as the lack of language id might have significance and it is not crucial for all steps of the analysis
trending_yt_videos	50420 'nan' values found on the 'video_tags' column. Ignored, as the lack of tags might have significance and it is not crucial for all steps of the analysis.
Yt_video_tags	**Subset of the data created with unique pairings of terms found in the 'video_tags' column and respective video_id for ease of analysis**
yt_video_tags	2662163 duplicates found (likely from the same videos entering the trending charts several times) and deleted since the goal is to have a unique pairing video_id vs video_tag for each tag
yt_video_tags	4632 'nan' values found on 'video_tag' column. Rows were deleted, as the purpose of the subset is to have a clean pairing of video_id and video_tag for each tag and that is not relevant in case the tags do not exist
trending_yt_videos	Column 'langauge' renamed to 'language' to correct spelling.

Questions to explore:

- What tags and potential categories identified from them are associated with popular Youtube videos?
- What is the average length of time between the time a video is published and when it reaches the peak of its popularity?
- Do videos with particular tags generate a larger number of comments than others?
- Are videos in a certain language more popular than others?
- Do the answers to these questions differ depending on the country in which the video became popular?
- Are there videos that are popular across several countries? If so, do they have anything in common?