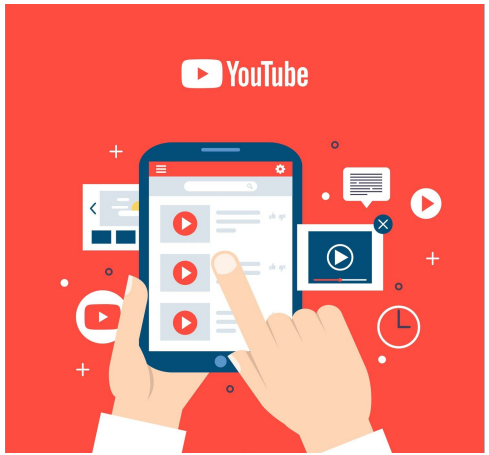
Analyzing Youtube Trending Videos

Rachel He

Tina Song

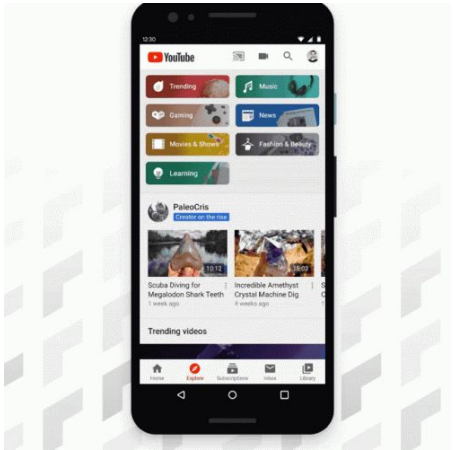
Project Goal

YouTube is one of the most influential digital media platforms, shaping entertainment, news, and online culture. Understanding what makes a video trend can provide valuable insights for content creators, marketers, and media analysts. Our project aims to analyze trending YouTube videos in the U.S. using interactive data visualizations to identify patterns in video popularity, engagement metrics, and content categories.



Project Motivation

Data visualization is a powerful tool for this analysis, allowing us to present complex trends in a clear and engaging way. Interactive bar charts, scatter plots, and time-series visualizations can help reveal patterns in trending videos, highlight key engagement factors, and track changes in YouTube's trending landscape over time. By making data-driven insights more accessible, this project can benefit content creators, marketers, and media analysts seeking to understand YouTube's evolving ecosystem.



Key Questions

1. Which categories appeared most frequently in YouTube trending videos for a given year?

2. Is there a relationship between the number of tags used in a video and its engagement rate?

3. How have YouTube trending video views changed over time across different categories?

4. Which videos were the most popular each year based on views and likes, and how do they compare across categories?

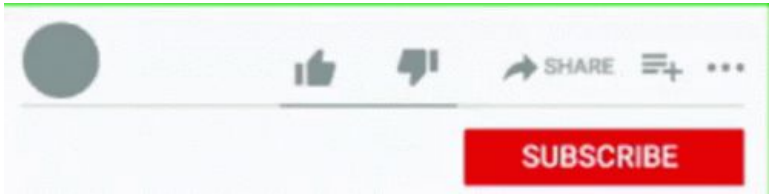
Approach

To address these questions, we designed an interactive notebook in Vega-Lite within Observable.

Our intended audience include:

- General internet users interested in media consumption trends
- Content creators who want to understand how videos trend
- Digital marketers and researchers analyzing online engagement

We assume our readers have a basic understanding of YouTube but may not be familiar with data visualization techniques. They are reading our page to gain insights into YouTube's trending patterns.



Dataset Overview

We are using the YouTube Trending Video Dataset from Kaggle (link). This dataset is considered reliable because, according to the author, it was collected directly from YouTube's API, ensuring accurate data about trending videos.

Our dataset consists of 268,787 rows and 16 columns, capturing trending YouTube videos in the U.S. from 2020 to 2024.

video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date
3C66w5Z0ixs	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51xNrqAzg	Brawadis	22	2020-08-12T00:00:00Z

tags	view_count	likes	dislikes	comment_count
brawadis|prank|basketball|skits|ghost|funny videos|vlog|vlogging|NBA|browadis|challenges|bmw i8|faze rug|faze rug brother|mama rug and papa rug	1514614	156908	5855	35313

thumbnail_link	comments_disabled	ratings_disabled	description
https://i.ytimg.com/vi/3C66w5Z0ixs/default.jpg	FALSE	FALSE	SUBSCRIBE to BRAWADIS ▶ http://bit.ly/SubscribeToBrawadis FOLLOW ME ON SOCIAL ▶ Twitter: https://twitter.com/Brawadis ▶ Instagram: https://www.instagram.com/brawadis/ ▶ Snapchat: brawadis Hi! I'm Brandon Awadis and I like to make dope vlogs, pranks, reactions, challenges and basketball videos. Don't forget to subscribe and come be a part of the BrawadSquad!

Dataset Integrity

To ensure data reliability and consistency,

we conducted the following data preprocessing steps in Azure SQL:

- We added a new column, engagement_rate, calculated as: (likes+comments)/views
- We dropped columns deemed irrelevant to our analysis
- We joined a JSON file (link) that contained category IDs and their corresponding category names
- We counted the number of tags associated with each video and created a new column (tag_count) to capture this information
- We converted publishedAt and trending_date into proper datetime formats

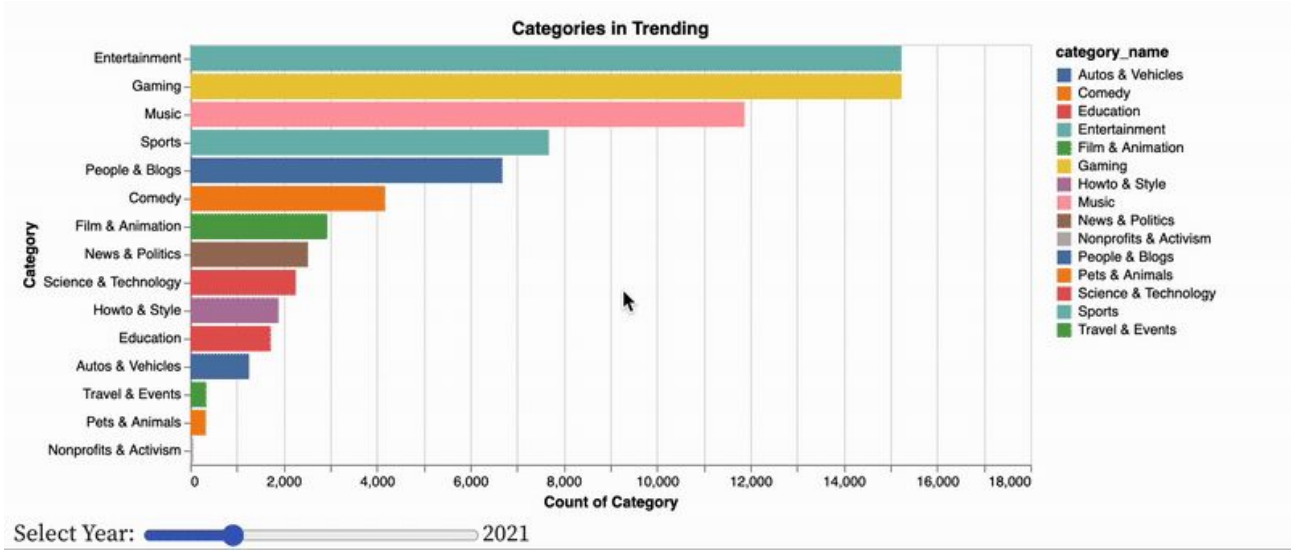
There were no missing values in the columns we kept

Our cleaned dataset (link) consists of 268,787 rows and 11 columns.

Column	Why drop it?
channelId	channelTitle is more readable and useful
categoryId	Use category_name instead (already mapped from JSON)
tags	Use tag_count instead
dislikes	YouTube removed public dislikes (no longer relevant)
thumbnail_link	Just an image link, not useful for analysis
comments_disabled	Not relevant since we focus on engagement metrics
ratings_disabled	Most videos enable ratings, so not very useful
description	Long text, not useful for structured analysis

https://observablehq.com/d/624d61066e35a893

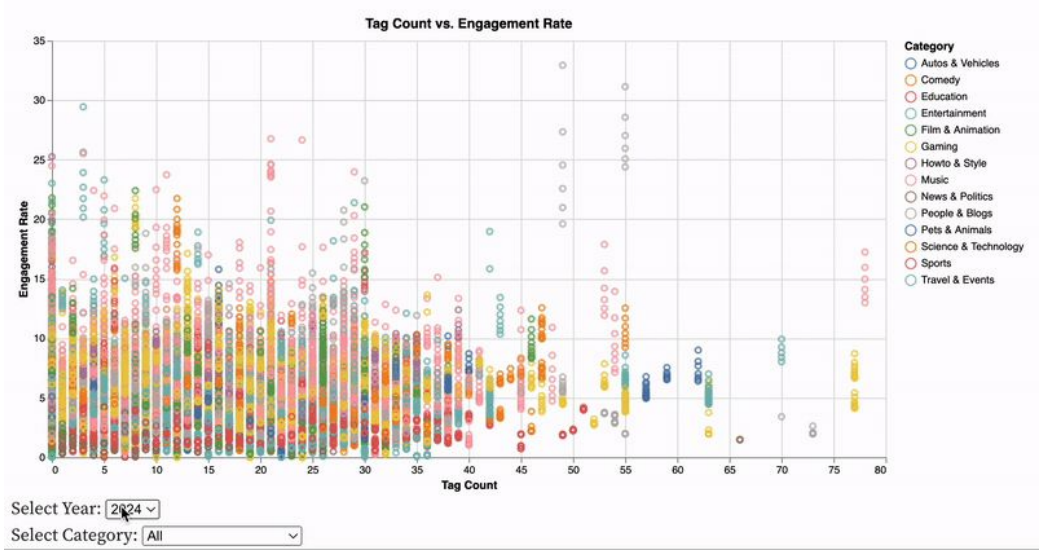
Visualization 1



Question: Which categories appeared most frequently in YouTube trending videos for a given year?

- The bar chart shows which video categories trend the most, with a year-based filter allowing users to explore trends over time.
- Sorted bars, clear color coding, and tooltips make it easy to compare categories, while a fixed x-axis scale ensures consistency across years.

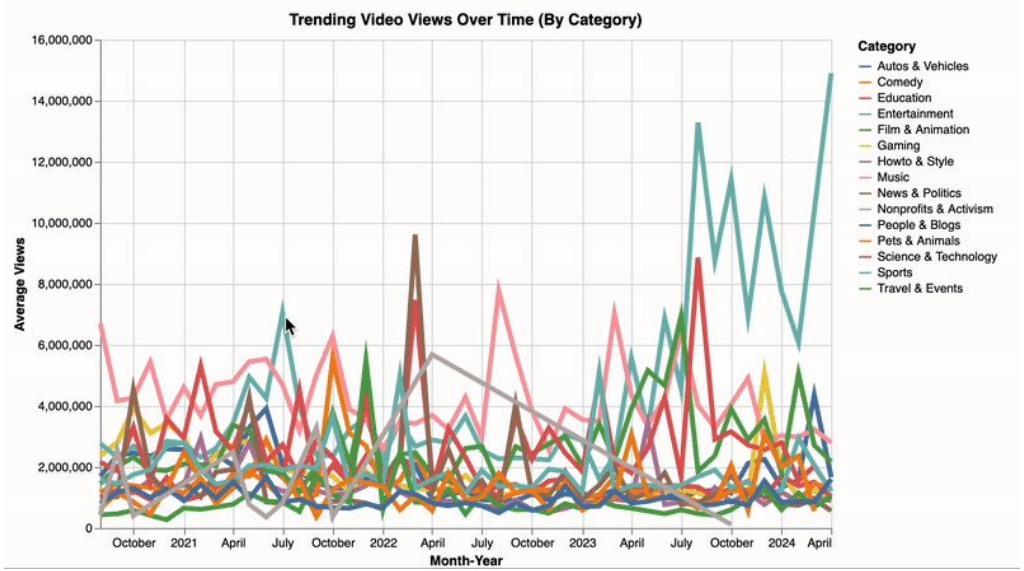
Visualization 2



Question: Is there a relationship between the number of tags used in a video and its engagement rate?

- The scatter plot shows the relationship between tag count and engagement rate, with filters for year and category to explore trends dynamically.
- Interactive filters let users focus on specific years or categories
- color coding and tooltips provide clear insights into engagement patterns.

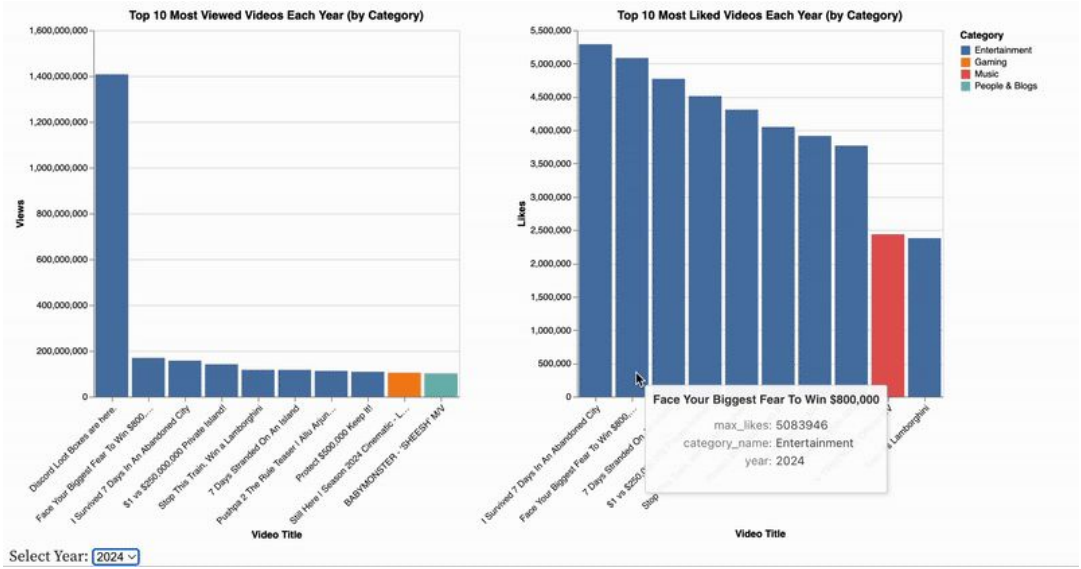
Visualization 3



Question: How have YouTube trending video views changed over time across different categories?

- The line chart tracks how trending video views change over time for different categories, helping analyze long-term patterns.
- It aggregates average views by month, uses color coding to distinguish categories, and includes interactive selection for deeper exploration.

Visualization 4



Question: Which videos were the most popular each year based on views and likes, and how do they compare across categories?

- These bar charts highlight the top 10 most viewed and most liked videos for each year, categorized by content type, helping analyze which videos gain the most traction.
- The side-by-side comparison of views and likes provides a comprehensive look at popularity
- Interactive filtering by year allows for trend analysis over time.

Conclusion

Limitations & Future Work

- Our dataset only includes videos that appeared in the trending section, meaning it does not represent all YouTube videos. Future work could analyze non-trending videos for a broader perspective.
- Expanding beyond U.S.-only data to include global trends could offer insights into regional content preferences and engagement behaviors.
- Our analysis was primarily descriptive, identifying historical trends. In future work, machine learning models could be applied to predict which videos are likely to trend based on their metadata, tags, and engagement patterns.
- We calculated engagement rate based on likes and comments but did not include watch time or shares, which could provide additional insights. Future work could integrate YouTube API watch-time data for a more comprehensive engagement analysis.

Key Contributions

In summary, we built an interactive notebook in Vega-Lite within Observable to explore trending YouTube videos in the U.S. from 2020 to 2024. Our notebook enhances the reader’s ability to explore and interpret trends dynamically, rather than relying on static charts.

Reflection

Throughout this project, we explored trending YouTube videos in the U.S. from 2020 to 2024 using an interactive Vega-Lite notebook in Observable. This experience provided valuable insights into data cleaning, feature engineering, and visualization design, strengthening our ability to derive meaningful conclusions from large datasets.

One of the most important lessons we learned was the significance of data preprocessing and integrity. A clean dataset makes building visualizations much easier and more effective.

Another key takeaway was the importance of feature engineering in enhancing analysis. These new metrics—engagement_rate and tag_count—allowed us to derive deeper insights beyond raw data.

Additionally, we realized the value of interactive data storytelling in making data exploration more engaging and insightful. By allowing users to filter data by year and category, we enabled a dynamic and flexible approach to analysis, rather than presenting static findings. This interactivity enhances user engagement.