YouTube Statistics

Daniel Setiawan
NYC Data Science Academy, 11/18

Brief Introduction

- As YT, TikTok, and other video websites evolve, our mindset on what captures viewer attentions must evolve as well
- As our consumer base grows, we need to keep up with it using data:
  - Understanding specific trends by country as crucial to business success
  - Understand how timing can play a huge factor in product marketing

Let’s talk data.

- 10 datasets of trending Youtube videos from around the globe
  - ~ 8 months worth of data, from 2017-2018
  - Each dataset contains ~ 40K observations
  - Canada, Germany, France, Great Britain, India, Japan, Korea, Mexico, Russia, US
- Additional dataset containing detailed categories
  - Missing from original datasets

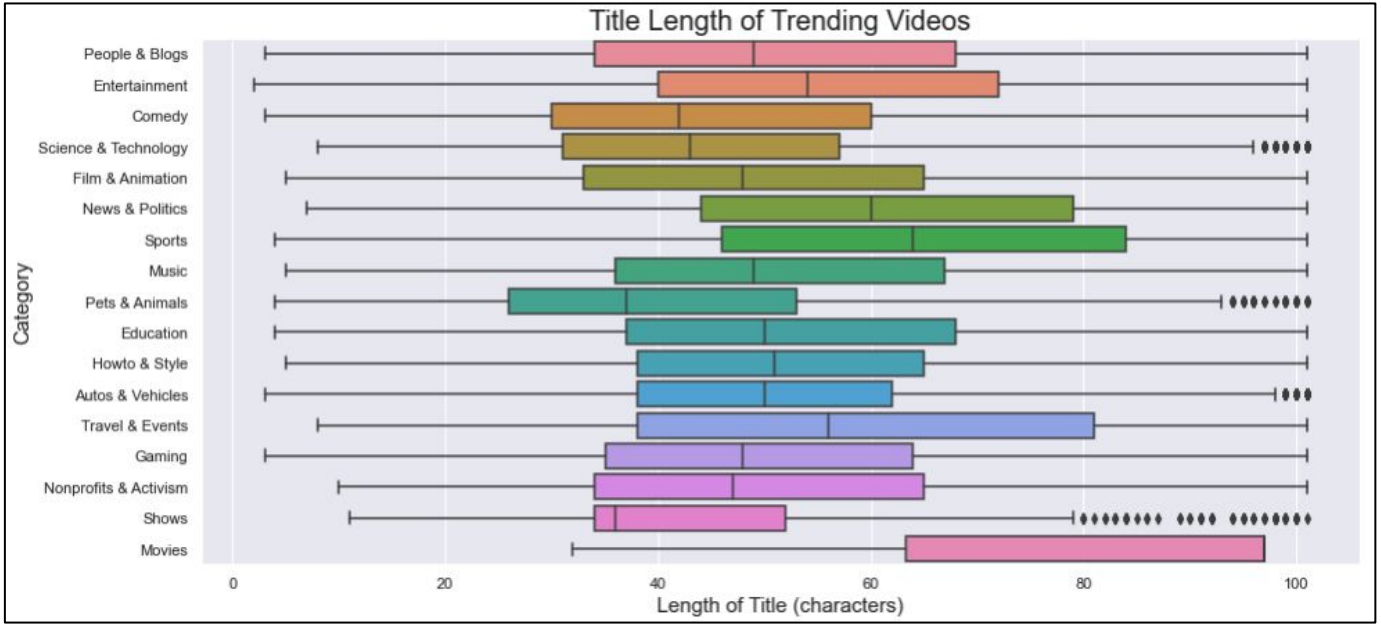
Objective

- Which categories should a company invest/allocate resources towards?
- Could we use viral data to predict consumer buying trends?
- What infrastructure requisites are needed to upkeep the data center?
- How does the energy consumption vary geographically?
- How does category-viewership differ between regions, and how can we use this data to make smart business decisions?

Let's start with the basics.

We're also able to capture attentiveness from the title lengths of these videos

→ Successful video titles can give a company insight on improving their product marketing.



- Density is

What can we learn from this?

- Viewers have different thresholds for different categories.
- Helps us understand what captures the viewers attention by category
- E-sport events are likely to reel more people in when headlines are ~ 50 letters long

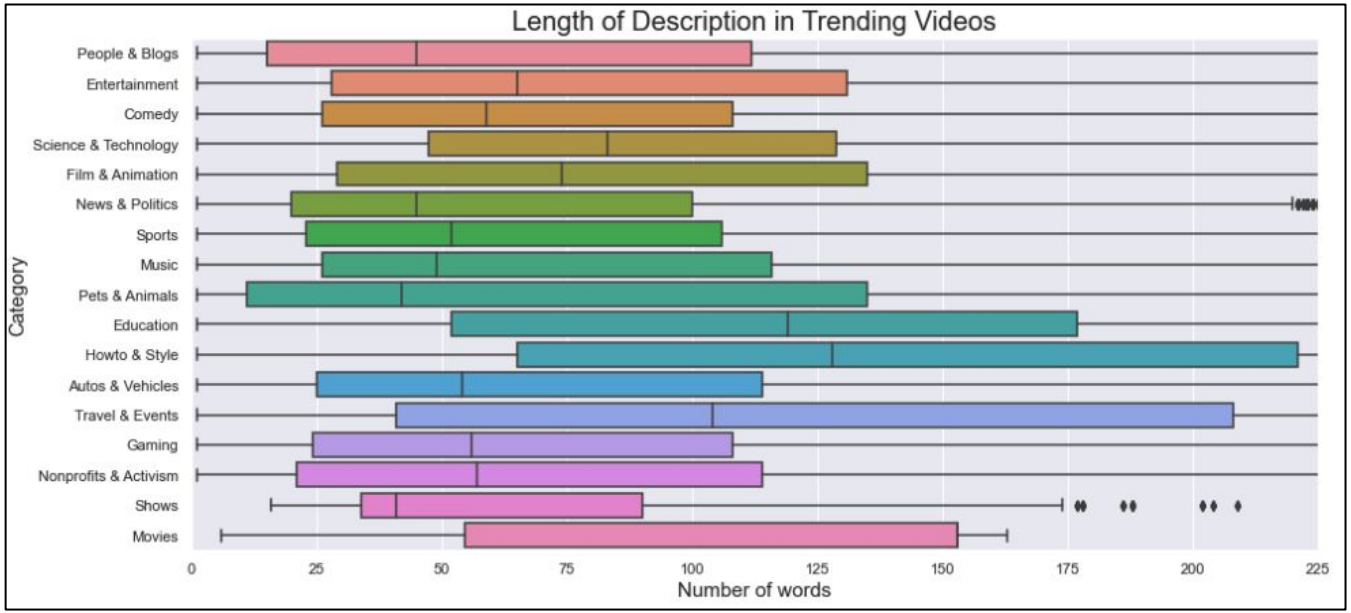
Let's start with the basics.

Likewise, we're able to look into the description length to see how many words are optimal in a product ad.

→ Too many words could risk a word blob, leading to a decrease in product sales, whereas too little may not give enough information. It's a delicate balancing act.

Note:

This metric alone is insufficient to drive a business pivot. It needs to be tied with additional data not currently available in this dataset.

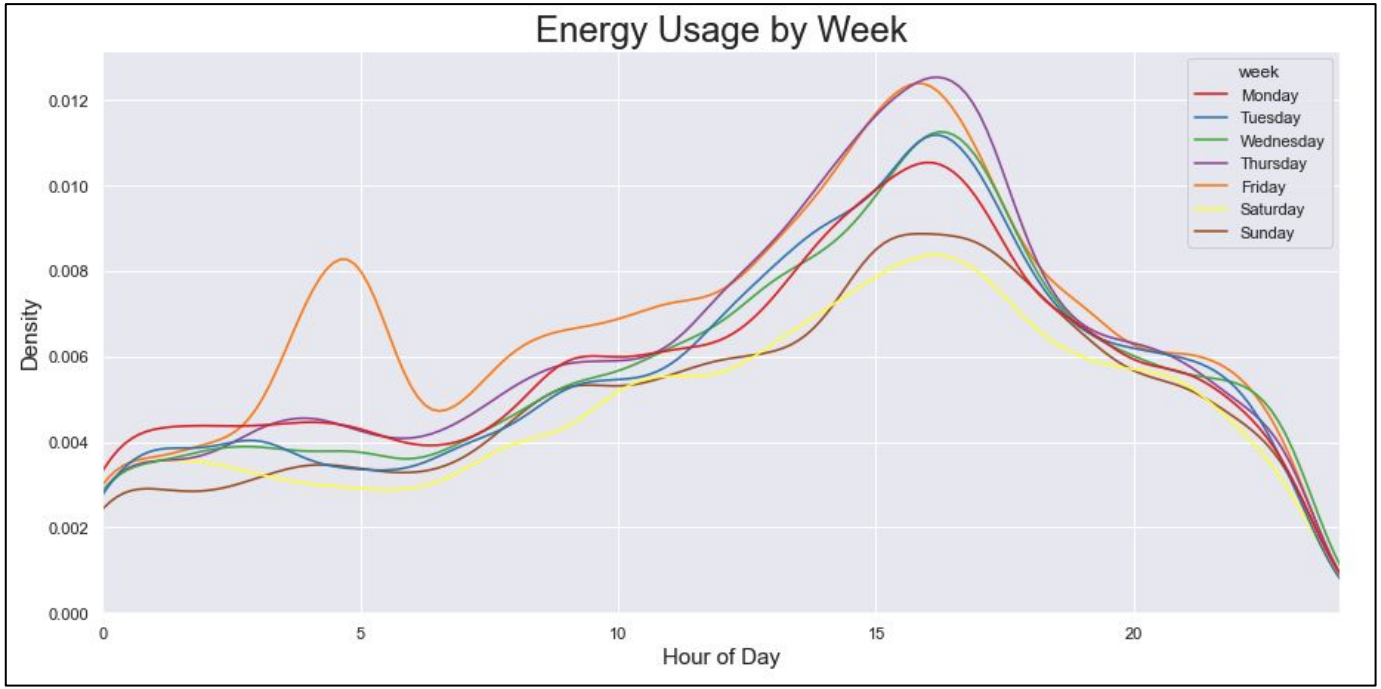


What can we learn from this?

- Viewers have different thresholds for different categories.
- Helps us understand what captures the viewers attention by category
- E-sport events are likely to reel more people in when headlines are ~ 50 letters long

Energy Consumption

Understanding how energy is consumed on a daily basis is vital to company operations



Insight into:

→ When are consumers more likely to buy certain products? (eg: most relaxed)

→ What type of bandwidth support is required throughout the day / week?

What can we learn from this?

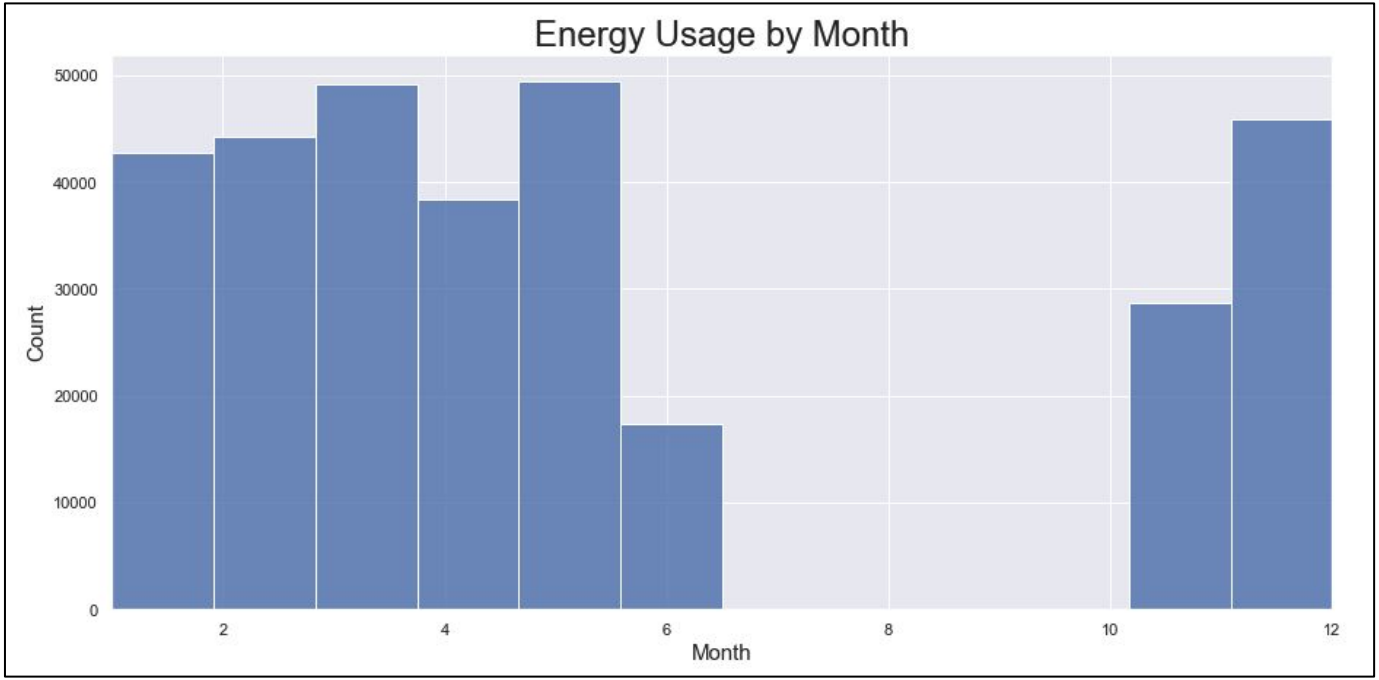
- Fridays see significant usage ~5AM
- During off-work hours, we expect triple the energy consumption from when the day first started
- Energy is less consumed on the weekends
- Coupling energy usage with viewers 'relax time' allow you to predict consumer buying habits

Energy Consumption

Understanding weekly trends isn't enough. It needs to be coupled with additional timeseries data.

Note:

→ Unfortunately, datasets did not have viable data between July to October.



What can we learn from this?

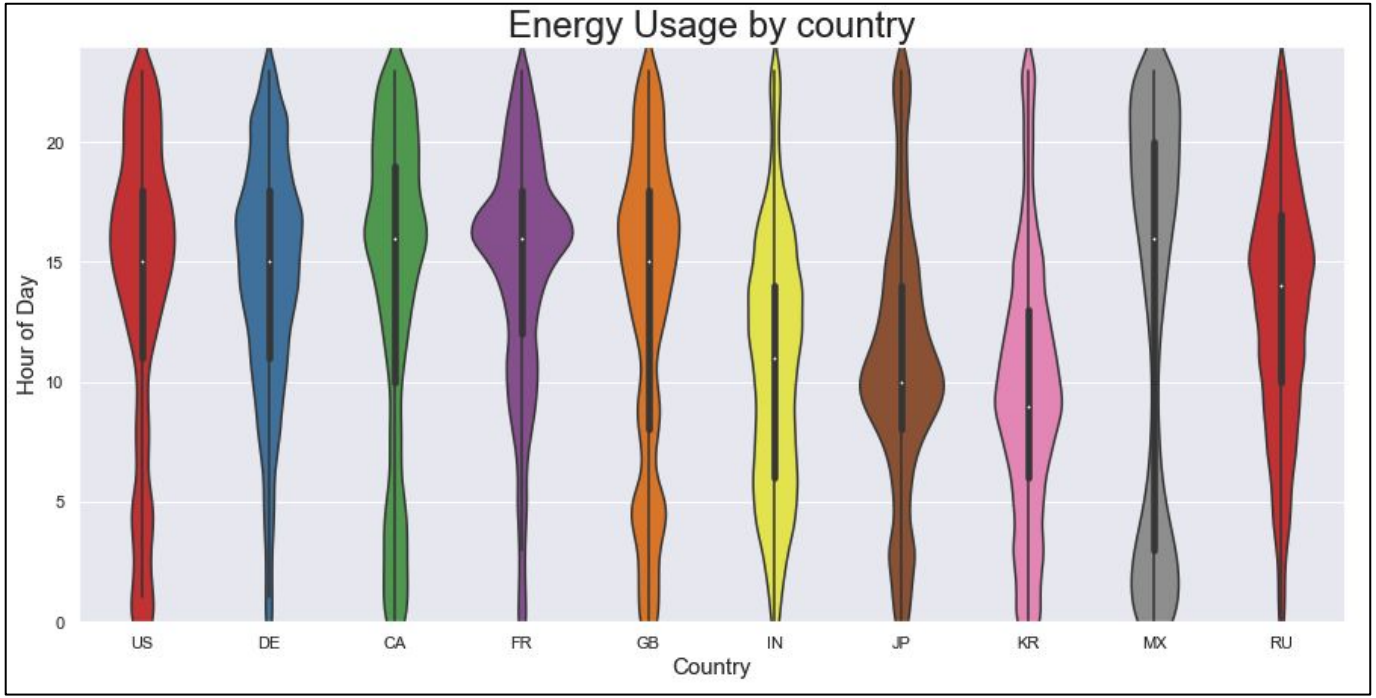
- Unable to make concrete conclusions without access to missing data
- But we can speculate that energy usage is lower during the summer months
  - Travel plans?
  - Spring break
  - Thanksgiving
- Marketing campaigns should target months with higher energy usage

Energy Consumption

We'd be foolish if we think energy consumption isn't dynamic by region.

Note:

→ Unfortunately, datasets did not have viable data between July to October.

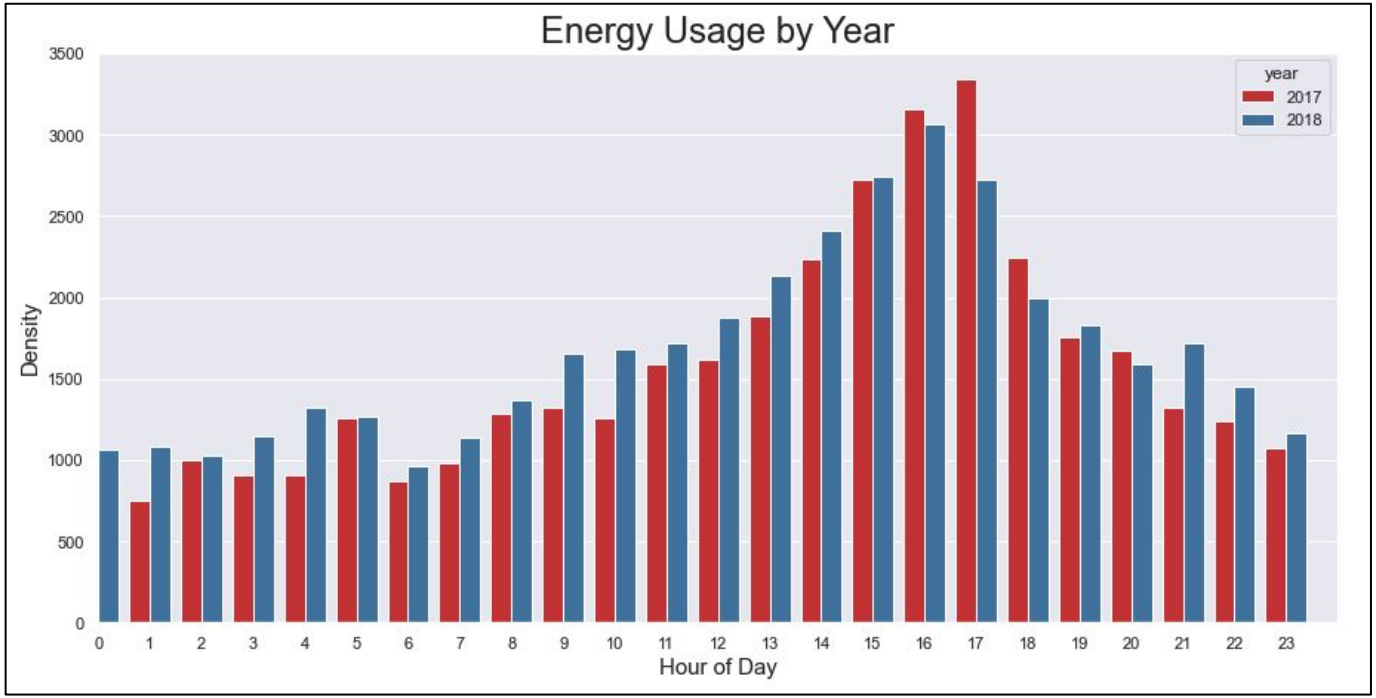


What can we learn from this?

- Data centers in France should be equipped to handle the high energy influx during peak hours
- Products targeting asian consumers should take into account their inactivity during normal peak hours
- Indian power grids is more vulnerable to overheating and should use more resilient parts.

Growth by Trends

If only we had more data...



Note:

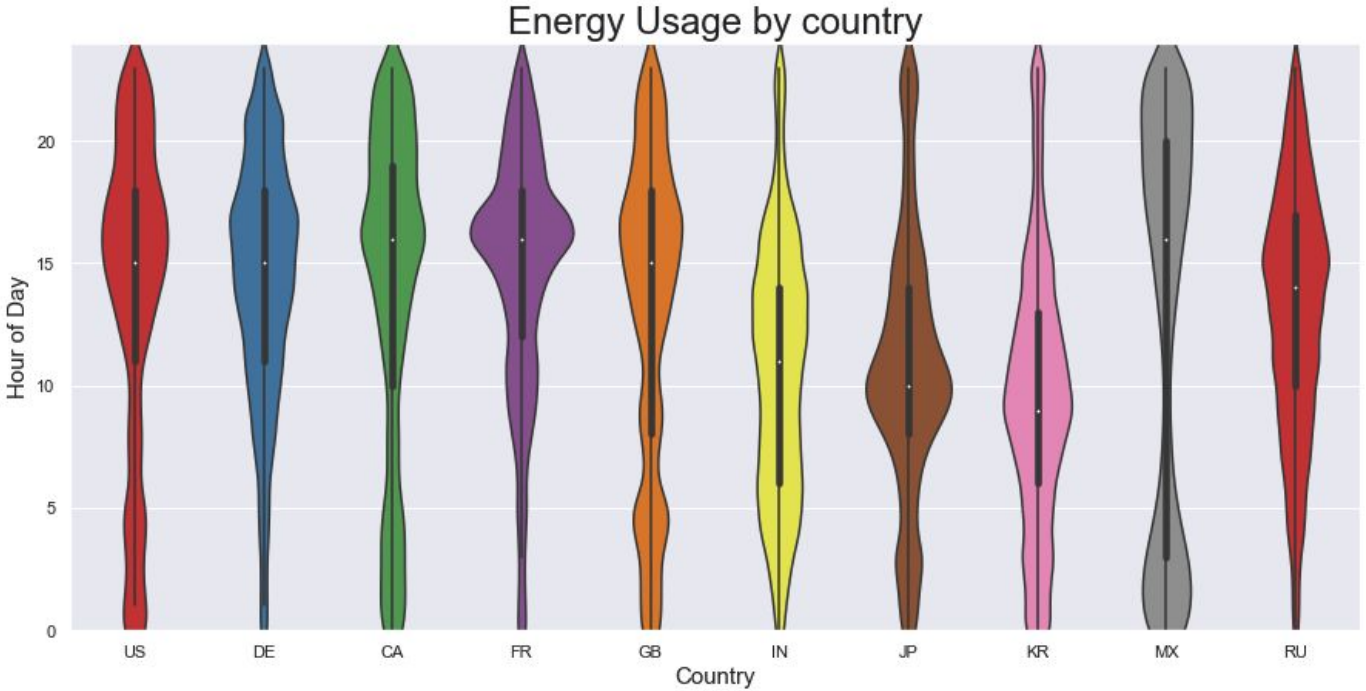
→ Unfortunately, data only captures 2017-2018

→ Data is normalized to ensure that months in data is reflected

What can we learn from this?

- We can use annual usage data to predict growth patterns
- We see higher late night energy usage in 2018 as opposed to 2017
- With more data, we can understand the shift and dependency towards electronics
- Gain insights on which platform is best suited for product advertisement

Publish time metric



- usage in france is much higher than the rest of the world ~4PM
- # the peak usage in Japan/Korea is ~9AM, whos countries are known for their nightlife
- # Mexico peak hours start at 7PM and continues to the early AM
- # usage in India is fairly consistent throughout the day, but takes a sharp decline ~8PM
-

Conclusion

- Which categories should a company invest/allocate resources towards?
  - Changes based on region, but filmmaking / music are top
- Could we use viral data to predict consumer buying trends?
  - Products targeting asian consumers should take into account their inactivity during normal peak hours
- What infrastructure requisites are needed to upkeep the data center?
  - Data centers in France should be equipped to handle the high energy influx during peak hours
- How does the energy consumption vary geographically?
  - Indian power grids is more vulnerable to overheating and should use more resilient parts.
- How does category-viewership differ between regions, and how can we use this data to make smart business decisions?

Future Work

- Find additional data to cover the gap between missing months
- Utilize updated YT data to see how the trend is affected by the recession and CoVID
- Scrape data from other social media platforms (TikTok, ...) to see the shift in demand from one platform to another
- Utilize techniques to see which age groups are frequently using the platform based on videos and consumer purchases, to better predict how people behave, based on region.