UNIVERSIDAD TORCUATO DI TELLA

UNIVERSIDAD TORCUATO DI TELLA, BUSINESS SCHOOL
MASTER IN MANAGEMENT & ANALYTICS

Monetization opportunity sizing of YouTube Trending

Causal inference modeling and counterfactual analysis with YouTube public API data

Abstract

YouTube is the biggest, most popular and most important video platform in the world. It receives millions of daily visits that generate billions of dollars in advertisement revenue which gets split between YouTube and "youtubers". "YouTube Trending" is a feature that gives videos an exposure boost by featuring them in a dedicated sector within the platform. These videos, which are capped at 50 per day per country, are automatically categorized as trending—based on their performance and other variables—and then exposed as such, boosting their discoverability within the application. In this thesis we measure how the number of video views of the platform increases due to this exposure boost. Higher number of views leads to higher ad revenue, which leads to higher value to YouTube and video creators. Our objective is to size how much of that value can be captured by YouTube by monetizing Trending. In order to perform this analysis we model the problem as an observational study to calculate Trending effect on views by using causal inference techniques. We also define the addressable market in which we will use the average treatment effect obtained by reverse-engineering YouTube's highly strict trending assignment criteria with a machine learning classification model. After that, we run a counterfactual analysis to estimate the willingness-to-pay creators have for each of the videos in the addressable market. Finally, we use this information to compute the price that maximizes net revenue under a scheme with no price discrimination and the net revenue YouTube could make under a first-degree price discrimination scheme.

Student: Chiaro, Franco.

Thesis advisor: Camporino, Maximiliano. PhD.

Date: May 26th, 2021.

UNIVERSIDAD TORCUATO DI TELLA, BUSINESS SCHOOL
MASTER IN MANAGEMENT & ANALYTICS

Dimensionamiento de la oportunidad de monetización de YouTube Trending

Modelado de inferencia causal y análisis contrafáctico con datos de la API pública de YouTube

Resumen

YouTube es la plataforma de vídeos más grande, popular e importante del mundo. Recibe millones de visitas diarias que generan miles de millones de dólares en ingresos publicitarios que se dividen entre YouTube y los "youtubers". “YouTube Trending” es una función que aumenta la exposición de los vídeos al presentarlos en un sector específico dentro de la plataforma. Estos vídeos, que tienen un límite de 50 por día, por país, son categorizados automáticamente como tendencias—según su rendimiento y otras variables—y luego expuestos como tales, aumentando su visibilidad dentro de la aplicación. En esta tesis medimos cómo aumenta la cantidad de vistas de videos de la plataforma debido a este incremento de exposición. Una mayor cantidad de vistas implica mayores ingresos publicitarios, lo que implica un mayor valor para YouTube y para los creadores de videos. Nuestro objetivo es medir cuánto de ese valor puede ser capturado por YouTube monetizando Trending. Para realizar este análisis, modelamos el problema como un estudio observacional para calcular el efecto de tendencia en las vistas de vídeos utilizando técnicas de inferencia causal. También definimos el mercado direccionable en el que utilizaremos el efecto promedio de tratamiento obtenido, haciendo ingeniería inversa sobre el criterio de asignación de tendencias altamente estricto que utiliza YouTube con un modelo de clasificación de aprendizaje automático. Después de eso, realizamos un análisis contrafáctico para estimar la disposición a pagar que tienen los creadores para cada uno de los videos en el mercado direccionable. Finalmente, usamos esta información para calcular el precio que maximiza los ingresos netos bajo un esquema de precios sin discriminación y los ingresos netos que YouTube podría generar bajo un esquema de precios con discriminación de primer grado.

Alumno: Chiaro, Franco.

Director: Camporino, Maximiliano. PhD.

Fecha de entrega: 26 de Mayo del 2021.

1. Introduction

1.1 YouTube Advertisement

YouTube is an online video platform owned by Google that has 2+ billion monthly users. It is localized in over 100 countries and registers one billion hours of video watched daily.[1] It is the second most-visited website in the world and has registered more than 500 hours of video content uploaded to its servers every minute in 2019.[2]

YouTube made US$15.1 billion in advertisement revenue in 2019.[3] It leverages Google AdSense to promote ads in its videos, which is a program through which website publishers in the Google Network of content sites serve text, images, video, or interactive media advertisements that are targeted to the site content and audience. YouTube's revenue also comes from YouTube Premium, but ads are still the main contributor.[4]

Not all videos have advertisements. In order to get ads in a video, the creator has to open an account and turn on *account monetization* and upload a video after that. There is a 45/55 split for all ad revenue generated, meaning YouTube keeps 45% of all YouTube ads revenue while creators the remaining 55%.[5]

Ad revenue generated for every 1,000 ad views can be calculated as the product of the CPM (*cost per Mile,* which expresses the revenue every 1,000 ad views) and the number of ad views:

$$Ad\ Revenue\ =\ CPM\ *\ Ad\ views$$

Equation 1.1: Ad revenue per 1,000 ad views

CPM can be understood as the reflection of the income the creator shares with YouTube. RPM (*revenue per Mile)* is used to express the creator's income considering the platform's retention. It is challenging to calculate YouTube CPM ad rates, but it is estimated that the average CPM in YouTube is US$18. With an average conversion rate from video view to ad view of 22%, RPM is estimated to be US$4.[6] This can be summarized as:

$$Ad\ Revenue\ =\ CPM\ *\ Video\ views\ *\ Conversion\ Rate$$

Equation 1.2: Ad revenue per 1,000 ad views as a function of video views

$$YouTube\ Ad\ Revenue\ =\ CPM\ *\ Video\ views\ *\ Conversion\ Rate\ *\ YouTube\ Share$$

Equation 1.3: YouTube ad revenue per 1,000 ad views as a function of video views

$$Creator\ Ad\ Revenue\ =\ CPM\ *\ Video\ views\ *\ Conversion\ Rate\ *\ Channels\ Share$$

Equation 1.4: Creators ad revenue per 1,000 ad views as a function of video views

---

[1] *YouTube for Press*, 2020. YouTube

[2] James Hales, May 2019. *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*. Tubefilter.

[3] Mountain View, California, February 2020. *Alphabet Announces Fourth Quarter and Fiscal Year 2019 Results*. Alphabet.

[4] YouTube, 2020. How does YouTube make money?

[5] Eric Rosenberg, June 2020. *How YouTube Ad Revenue Works*. Investopedia.

[6] Werner Geyser, August 2020. *How Much do YouTubers Make? — A YouTuber's Pocket Guide [Calculator]*. Influencer Marketing Hub.

The corollary is that the higher the number of video views, the higher the number of ad views; the higher the number of ad views, the higher the advertisement revenue; the higher the advertisement revenue, the higher the amount of money both YouTube and creators make.

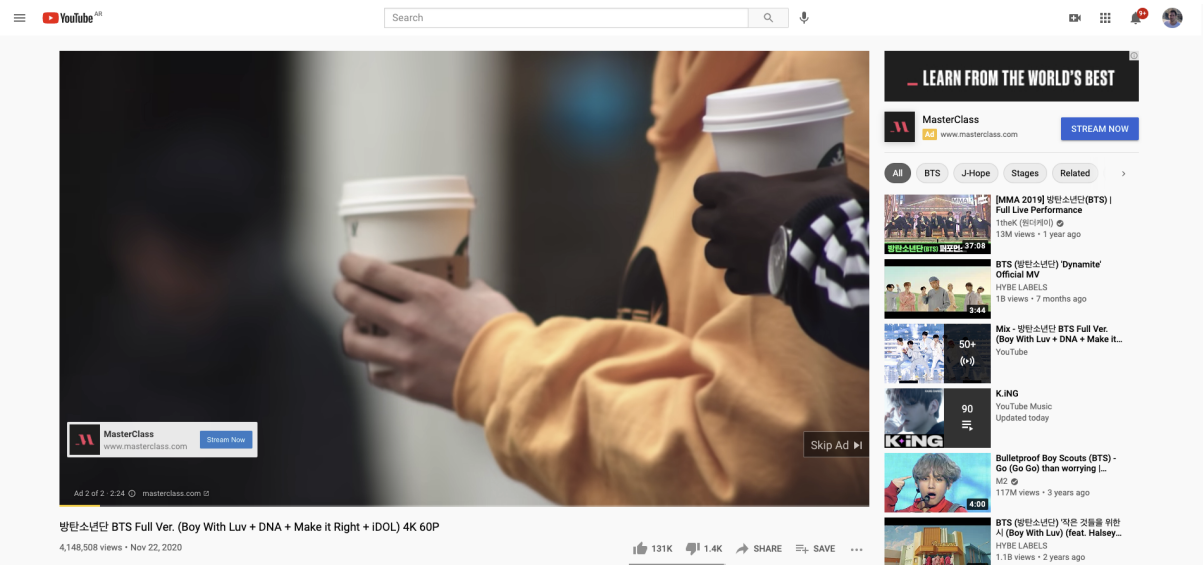

Figure 1.1: AdSense video in the middle of a YouTube video.

1.2 YouTube Trending

YouTube automatically selects videos on a daily basis and categorizes them as trending. Trending videos are exposed on a different section, significantly increasing video discoverability. Figure 1.2 shows the entry point to Trending (right below the Home icon in the left hand panel) and how this section looks.[7] The feature is available in the desktop web, mobile web, Android and iOS applications and it shares the same user interface and logic in each of these platforms.

[7] Layout was updated at the beginning of 2021. Implications of this redesign will be addressed in chapter 7.

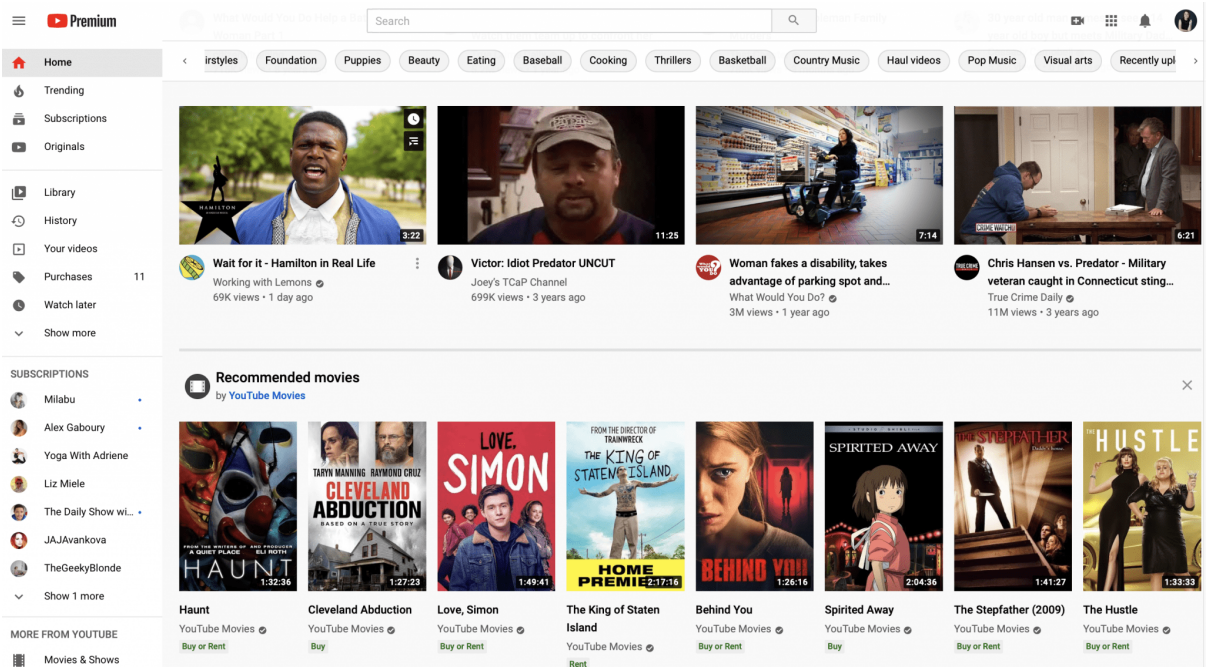

Figure 1.2: Desktop YouTube Trending videos.

There is a maximum number of 50 trending videos at a time per country. These are selected by a YouTube in-house algorithm on a daily basis, whose way of working is unknown to the public. However, by doing some reverse engineering we observed that videos selected as trending are brand new—i.e. they have less than 20 days since uploaded to the platform, most of them being 2 days-old—and have a considerably large number of views at the moment of being tagged as trending: the average number of views for trending videos is 205,000 compared to 16,000 for those non-trending videos with the same age. These seem to be two key variables at the moment of classifying videos between trending and not trending.

We model YouTube's users funnel in the following way: (1) Users enter YouTube, (2) they look for a video, (3) they find a video, (4) they play the video. We merge steps 1 to 3 and call it "users discover a video". We can then think that the probability of a user playing a video is equal to the probability of that user discovering a video multiplied by the conditional probability of the user playing the video once he or she found it:

$$P(user\ plays\ video) = P(user\ discovers\ video) * P(user\ plays\ video\ /\ user\ discovers\ video)$$

Equation 1.5: Probability of a user playing a video.

The main hypothesis we want to prove in this thesis, indirectly, is that being marked as "trending" increases the probability of being discovered by users, and therefore increases the probability of being played. We later use the result obtained to validate that hypothesis to prove that the YouTube Trending feature increases the total number of video views on the platform at large.

As a final comment, we note that the effect found cannot be extrapolated to a different setup. That is, if YouTube changed its layout, e.g. changed its trending cap from 50 videos to 100, the results obtained throughout this work would no longer be valid.

1.3 The business case

Assuming the main hypothesis is true (i.e. the YouTube Trending feature increases the number of views on the YouTube platform at large), we can conclude that the Trending feature increases ad revenue in the system. In other words, even though Trending is free, it generates revenue for YouTube. But at the same time it generates revenue for creators, at least those whose videos make it to the Trending section.

The purpose of this work is to understand if YouTube can capture some of the revenue creators are getting from Trending by charging a fee. By doing this, there will be a population of creators whose videos will fall out from Trending due to an unwillingness to pay for such functionality. This means that the total number of views of the system may drop and therefore so may the ad revenue.

This dilemma generates a tradeoff between Trending fee-driven revenue and Trending ad-driven revenue. We want to find the fee amount, i.e. price, that generates a breakeven between the revenue created by charging for retaining placement on Trending and the ad revenue lost by those videos that fall from Trending because their creators are not willing to pay to keep their spot. In effect, we want to find the price that implies the following state of equilibrium:

$$Trending\ fee\ revenue\ =\ Ad\ revenue\ lost$$

From YouTube's standpoint, Trending revenue can be broken down into the number of paying-user trending videos (PUTV) multiplied by the fee per video (from now on, the "price") and the ad revenue lost as the number of non-paying-user trending ad views lost (for simplification, ad views lost) multiplied by the CPM over 1,000, corrected by the revenue share correspondent to YouTube:

$$PUTV\ *\ Price\ =\ Ad\ views\ lost\ *\ \frac{CPM}{1000} *\ 45\%$$

Equation 1.6: Breakeven condition for YouTube

The price that solves this equation is the minimum price YouTube can charge for the Trending feature without losing revenue, assuming a determined level of paying users. If YouTube charges below that minimum price, the ad revenue lost by the incremental views of those videos that should have been trending but were not, will be higher than the revenue generated from the monetization of the feature.

From a creator standpoint, one would expect that the price of Trending must be smaller than the revenue generated by it:

$$Price <\ Ad\ revenue\ generated$$

The right side of this expression can be broken down into the number of ad views the video gained due to Trending's effect multiplied by the CPM over 1,000, corrected by the revenue share correspondent to creators:

$$Price <\ Ad\ views\ gained\ *\ \frac{CPM}{1000} *\ 55\%$$

Equation 1.7: Creators condition to convert to Trending paying-user

An important simplifying assumption is that creators behave in a rational way, so if this inequality is not met, they do not pay for Trending. We are also assuming that there is no external revenue for creators tied to video views (e.g. a company advertising a product might be willing to "lose" money in YouTube in order to get more exposure and boost offline sales).

Net revenue for YouTube will be defined as the difference between Trending fee revenue and ad revenue lost:

$$YouTube\ Net\ Revenue\ =\ Trending\ fee\ revenue\ -\ Ad\ revenue\ lost$$

Equation 1.8: YouTube Net Revenue

High prices make more creators not pay for Trending, making the number of ad views lost increase, demanding a higher price to YouTube in order to remain breakeven. On the contrary, low prices makes more creators pay for Trending, making the minimum price decrease. In this thesis we will find the price that maximizes YouTube net revenue taking into account the tradeoff and constraints presented.

1.4 Methodology

As expressed above, we want to maximize net revenue with respect to Trending price. If we formalize this statement as an optimization problem, what we want is to:

$$Maximize\ \{PUTV\ x\ Price\ -Ad\ views\ lost\ *\ \frac{CPM}{1000} *\ 45\%\}$$

Following the reasoning of the previous section, we model ad views lost as:

$$Ad\ views\ lost\ =\ \sum^{videos} f(Price, creator)$$

$$Where\ f(Price, creator) =\ Trending\ effect\ on\ ad\ views, if$$
$$Price \geq\ Ad\ views\ gained\ *\ \frac{CPM}{1000} *\ 55\%$$
$$Else, f(Price, creator) =\ 0$$

For both cases:

$$Trending\ effect\ on\ ad\ views\ =Ad\ views\ gained$$

The missing piece of this problem is $f(Price, creator)$. If we dig into that formula, we realize that what we need to find is the effect Trending feature has on ad views. Once we have this value, we can use historical data and run a counterfactual analysis to calculate the price that maximizes revenue. We understand "counterfactual analysis" as the process of calculating the ad views a trending video would gain from one week to another in the absence of the exposure boost Trending provides and, similarly, the ad views a non-trending video would gain in the same time span in the presence of this exposure.

Something that has been omitted so far is that there might be videos that did not make it to Trending section due to the 50 videos cap but share the same characteristics as actual trending videos and would have made it to Trending if the cap was slightly higher, for instance.

This is something to bear in mind when calculating ad views lost: we will evaluate a video's characteristics when defining the addressable market in which we run the counterfactual analysis, including any videos that could have been trending, but weren't.

Our procedure to find the optimal price is as follows:

1. Estimate average Trending effect on ad views
2. Define the addressable market
3. Model creators' willingness to pay
4. Find the optimal price

The time frame in which we evaluate ad revenue lost is one week. As mentioned above, most trending videos are roughly 2 days old when uploaded. This suggests that they will remain trending just for the day or for a couple of days; in a minority of the cases will a video trend for more than a week.

1.4.1 Trending effect estimation

In the first step, we employ a variety of causal inference techniques to compute the average effect of Trending on ad views. Notably, there are variables that affect both the probability of being selected as trending and the views the video gains after a week (e.g. video views and active days), which we need to control in order to size the isolated effect of Trending properly, i.e. in an unbiased manner.

The effect we infer is over the delta (difference) of video views from one week to the next.

1.4.2 Addressable market definition

Here we select the subset of videos that we will use to run the counterfactual analysis.

To belong to this group, a video must be trending, or looks like one. We use a machine learning algorithm to define the latter.

1.4.3 Willingness-to-pay calculation

In this stage we use the effect calculated in the previous step to compute the maximum price creators would be willing to pay for each of their videos (given our stated assumptions).

1.4.4 Optimal price calculation

Finally, we simulate scenarios with different prices to the one which maximizes net revenue (given results of the previous steps).

We also evaluate a scenario in which there is not an unique price for all creators but, on the contrary, a bespoke price for each, based on their willingness to pay (i.e. first-degree price discrimination).

2. The data

2.1 Dataset setup

All the data used for this thesis was gathered from the YouTube public API, YouTube Data API v3.[8] This API has multiple endpoints that expose metadata and statistics for YouTube videos and channels.

Building the dataset is not trivial. It requires multiple steps, as: (1) the API limits the number of requests allowed per day; (2) the API has multiple endpoints; different data are extracted from different endpoints; once extracted, we must post-process then consolidate the data into a single dataframe; (3) we must manually generate a sample of non-trending videos (the API does not support directly pulling random videos); (4) trending videos have their own endpoint, which requires a separate pull execution and later consolidation; (5) when querying the API for statistics about a given video, it returns statistics about that video at that point in time; as such, in order to get statistics for videos and channels at two different moments (i.e. beginning of week N and N+1) we must make three distinct calls to the API and then consolidate: two calls at time T (beginning of week N) to get video and channel statistics and one final call one week later to get updated video statistics.

If not for this complexity, we would have built a dataset with daily statistics for numerous weeks (week N-1, week N and the first day of week N+1). This would have permitted us to extract information about the natural trend of these statistics using time series techniques—which would have proven valuable at the moment of inferring Trend effect. Unfortunately, owing to this complexity, we did not pursue these techniques.

2.2.1 Dataset structure

Our resulting dataset has the following structure:

- Each row represents a video
- Each video has its metadata, video statistics snapshot at time T and T+1 week, its channel metadata and channel statistics at time T as well.

The fields of this data are defined as follows:

Field name	***Field description***	***Data type***
videoId	Video unique id	key
categoryId	Video category id	categorical
channelId	Channel unique id	key
videoTitle	Title of the video	string
videoDescription	Description of the video	string

[8] https://developers.google.com/youtube/v3/docs/playlists

liveBroadcastContent	Indicates if the video is an upcoming/active live broadcast	boolean
publishedAt	Date in which the video was uploaded	date
tags	List of keyword tags associated with the video	string
viewCount	Number of views the video has	numerical
commentCount	Number of comments the video has	numerical
dislikeCount	Number of dislikes the video has	numerical
likeCount	Number of likes the video has	numerical
dimension	Indicates whether the video is available in 3D or in 2D	categorical
duration	Length of the video in ISO 8601 format[9]	string
licensedContent	Indicates whether the video represents licensed content	boolean
license	Video's license[10]	categorical
projection	Specifies the projection format of the video[11]	categorical
uploadStatus	The status of the uploaded video[12]	categorical
madeForKids	Indicates whether the video is designated as child-directed	boolean
embeddable	Indicates whether the video can be embedded on another website	boolean
channelSubscribers	Number of subscribers the video's channel has	numerical
channelViews	Sum of video views the video's channel has	numerical
channelVideos	Number of videos the video's channel has	numerical
videoTrend	Indicates whether the video is trending or not	boolean

Figure 2.1: Dataset fields

Numerical fields correspond to video and channel statistics at the moment of hitting the API. For that reason, these fields actually appear twice in the dataset: one with data corresponding to 02/16/2021 (time T) and the other to 02/23/2021 (T+1 week)—5PM ET in both cases. The value of videoTrend corresponds to the snapshot taken on 02/16/2021, 5PM ET. Figures 2.2, 2.3, 2.4 and 2.5 show a preview of the dataset.

[9] https://en.wikipedia.org/wiki/ISO_8601#Durations

[10] Valid values for this property are: *creativeCommon* and *youtube.*

[11] Valid values for this property are: *360* and *rectangular.*

[12] Valid values for this property are: *deleted, failed, processed, rejected,* and *uploaded.*

	videoId	categoryId	channelId	videoTitle	videoDescription
0	Yzf06dGJ_wg	20	UCogK8QXzia_Cbejl5uk9LkQ	Hoty Gaming Vs. Mekel (Youtuber 1v1) - Rainbo...	Can we take on the champ of 1v1s?? \nHoty's Cha...
1	eTMDwsXdeCE	10	UCDXmIU0uUvO_gQ0D6t--hOA	H.O.T.Y.	Provided to YouTube by EMPIRE\n\nH.O.T.Y. · Sl...
2	xnf_8A83ITo	22	UCSQLE6jT9w0YsHQ0TjORfSg	Wo Kam Jo Sirf South Korea main Hoty hain I 1...	Wo Kam Jo Sirf South Korea main Hoty hain I 1...

Figure 2.2: Preview of dataset used (1/4)

liveBroadcastContent	publishedAt	tags	viewCount	commentCount	dislikeCount	likeCount
none	2019-08-21T19:15:01Z	["'Rainbow Six Siege Mekel Mekel Siege R6'", "...	270414.0	131.0	131.0	7217.0
none	2015-10-31T06:10:06Z	["'Slim Thug Hogg Life'", " 'Vol. 3: Hustler o...	186862.0	94.0	94.0	1805.0
none	2020-09-26T07:37:04Z	["'South korea'", " 'facts about south korea'"...	32411.0	44.0	44.0	556.0

Figure 2.3: Preview of dataset used (2/4)

dimension	duration	licensedContent	projection	uploadStatus	madeForKids	license
2d	PT24M3S	True	rectangular	processed	False	youtube
2d	PT4M5S	True	rectangular	processed	False	youtube
2d	PT4M39S	True	rectangular	processed	False	youtube

Figure 2.4: Preview of dataset used (3/4)

embeddable	channelSubscribers	channelViews	channelVideos	videoTrend
True	357000.0	40081995.0	370.0	False
True	12000.0	37396244.0	1343.0	False
True	632000.0	93273372.0	151.0	False

Figure 2.5: Preview of dataset used (4/4)

2.2.2 Dataset construction

We begin building this dataset by generating a random sample of search queries—the same queries one would type into YouTube's search bar when looking for a video. The API requires this query as an input in order to retrieve metadata information.

The method we used for creating these random search queries is shown in Figure 2.5, which simply consists of generating random strings. The same figure shows the function used to hit the YouTube API. These requests retrieve lists of videos and their associated metadata. Given the API's daily request limitation, we extract and save video IDs on a daily basis then later merge and deduplicate them once we reach a sufficient number of observations. Finally, we use this larger list of video IDs to hit the metadata and statistics endpoints to construct our dataset.

```
import string
import random

k = 0
randomVideos = []
nextPage_token = None


while k < 100:
 random1 = ''.join(random.choice(string.ascii_uppercase + string.digits) for _ in range(4))
 random2 = ''.join(random.choice(string.ascii_uppercase + string.digits) for _ in range(5))
 random_search_query = random1 + '|' + random2

 snippets = youtube.search().list(part='snippet', maxResults=7500,order='relevance' ,
q=random_search_query, type='video').execute()
  j = 0
 for i in snippets['items']:
   randomVideos.append(snippets['items'][j])
   j += 1

 k += 1
```

Figure 2.6: Script used for retrieving random videos from YouTube API.

We understand that generating search queries merging random letters instead of picking random words from a dictionary: (1) Reduces language bias at a low cost. With a word-based approach we would need to create a weighted random selection of words method in order to generate unbiased search queries from dictionaries –which would require one dictionary per known language. (2) Reduces word-centric bias. Search queries generated with words from dictionaries would have actual words; an urban expression (e.g. slang language), a fictional character name or a brand, for example, would not appear in a dictionary. Therefore, using a word-based approach would reduce the probability of retrieving videos in which the search key words are not actual words.

It is worth mentioning that the request to the YouTube API shown in Figure 2.6 requires as an input the order in which the result will be presented. In our case we select *relevance*, which is a metric whose definition YouTube does not publically share. We acknowledge that using an "order" criterion introduces bias to the dataset, meaning it removes a certain level of

randomness. However, the number of videos retrieved on each request barely hits the maximum amount of results, so this bias should not be a meaningful issue.

In parallel, we hit the trending videos endpoint to retrieve the videos that are trending at time T.[13] This list of IDs is later merged into the bigger video ID list, deduped and differentiated from the videos that are not trending. Once we have that list of IDs, we hit the following endpoints to start populating the dataset:

- *Video*, to add video and channel metadata
- *Video statistics*, to add the video's statistics. We hit this endpoint twice in order to populate these fields at moment T and T+1 week.
- *Channel statistics,* to add the channel's statistics. We hit this endpoint once in order to populate these fields at time T.

2.2.3 Feature engineering

We perform a set of transformations to create new variables that (a) fit our model's required format and (b) facilitate the work of these models, increasing the chances of obtaining higher performance, both for the inference and prediction tasks. The variables created are shown in the following table:

Field name	***Field description***	***Data type***
activeDays	Day difference between 02/16/2021 and video publish date	numerical
durationInSeconds	Duration in seconds of the video	numerical
titleLength	Length of video title	numerical
descriptionLength	Length of video description	numerical
titleLanguage	Language of video title[14]	categorical
descriptionLanguage	Language of video description	categorical
hasDescription	Indicates whether the video has a description	boolean
tagCount	Indicates the number of tags the video has	numerical
hasTag	Indicates whether the video has at least a tag	boolean
commentsEnabled	Indicates whether comments are enabled for that video	boolean
likesEnabled	Indicates whether likes are enabled for that video	boolean
dislikesEnabled	Indicates whether dislikes are enabled for that video	boolean
viewsToLikes	Video views to video likes ratio	numerical
viewsToDislikes	Video views to video dislikes ratio	numerical

[13] YouTube API allows one request per country. For this thesis we retrieved trending videos from 20 countries which account for the largest number of views of the world. ISO 3166-1 alpha-2 country codes of these countries are: US, GB, DE, CA, FR, RU, MX, KR, JP, IN, AR, CO, CL, BR, ES, ID, AU, ZA, NG, PK.

[14] We used https://pypi.org/project/langdetect/ library to detect string language.

viewsToComments	Video views to video comments ratio	numerical
viewsToChannelSubscribers	Video views to video channel subscribers ratio	numerical
viewsToChannelVideos	Video views to video channel videos ratio	numerical
viewsToChannelViews	Video views to video channel views ratio	numerical
deltaViews	Views difference from moment T to T+1 week	numerical
logDeltaViews	Logarithm (base e) of (delta views + 1) [15]	numerical
logViewCount	Logarithm (base e) of (video views + 1)	numerical
logLikeCount	Logarithm (base e) of (video likes + 1)	numerical
logDislikeCuount	Logarithm (base e) of (video dislikes + 1)	numerical
logCommentCount	Logarithm (base e) of (video dislikes + 1)	numerical
logChannelSubscribers	Logarithm (base e) of (video channel subscribers + 1)	numerical
logChannelViews	Logarithm (base e) of (video channel views + 1)	numerical
logChannelVideos	Logarithm (base e) of (video channel videos + 1)	numerical

Figure 2.7: Variables added to the dataset.

Owing to the high dispersion amongst the values in video statistics, we transform them into logarithms. As shown in the descriptive analysis in the next section, metrics like video views go from zero to more than a billion, having most of the views concentrated near zero.

2.1 Descriptive analysis

The share of trending videos at a moment in time is completely negligible. If taking into account all the countries in the world and assuming all countries have YouTube and Trending, the maximum number of possible trending videos is 9750. YouTube has hundreds of millions of active videos uploaded that are not trending.

Our dataset has 64,399 videos. From those, trending videos represent 1.1%.

We will start by analyzing some basic statistics of the variables in our dataset:

- *categoryId:* There are 16 different categories. categoryId = 22 has the largest share of videos (43%), followed by categoryId = 20 (13%) and categoryId = 10 (9%)
- *channelId:* There are 51,857 distinct channels. 80.5% of videos belong to different channels.
- *videoTitle:* 97.5% of the titles in the sample are distinct.
- *videoDescription:* 91.5% of the titles in the sample are distinct. 25% of videos do not have a description.
- *liveBroadcastContent:* This field does not provide any valuable information. All videos but 35 have the same value (“none”).

[15] We sum 1 to the actual stat to avoid log(0), which is undefined. We select 1 to avoid having negative values.

- *tags:* Almost 50% of videos do not have tags. From the remaining half, 90% have distinct sets of tags.
- *viewCount:* 5% of videos have 1 or 0 views. 75% of them have less than 1,090 views. The maximum number of views of the sample is 1.5 billion.
- *commentCount, dislikeCount, likeCount:* The three follow a similar distribution than video views.
- *dimension:* This field does not provide any valuable information. All videos but three have the same value (“2d”)
- *licensedContent:* this value is False for 81% of the videos in the sample.
- *projection:* This field does not provide any valuable information. All videos but 16 have the same value (“rectangular”)
- *uploadStatus:* This field does not provide any valuable information. All videos but 35 have the same value (“processed”)
- *madeForKids:* Only 5% of the videos of the sample are marked as True.
- *license:* 99% of videos in the sample have a YouTube license
- *embeddable:* Almost 99% of videos in the sample are embeddable.
- *channelSubscribers*: 5% of channels have no subscribers. 25% have less than 5 while 75% have less than 1,770. The channel with the most subscribers has 172 million.
- *channelViews, channelVideos:* They follow a similar distribution than channel subscribers.
- *activeDays:* This variable follows a distribution with an exponential-like shape, having a maximum of 5,625 days
- *durationInSeconds:* This variable also follows a distribution with an exponential-like shape but with a way higher concentration around zero and an extremely higher range: 43,155 seconds.
- *titleLength:* This variable follows a distribution with a lognormal-like shape, with a mode around 20 characters. Title field has a cap on 100 characters, so there is a slight second mode at the very right of the distribution because of that.
- *descriptionLength:* This variable follows a distribution with an exponential-like shape and a high concentration of videos around zero. It has a range of 10,000 characters.
- *titleLanguage*: 56 distinct title langauges identified. Most of them (36%) are in english. The remaining distribution looks pretty much even.
- *descriptionLanguage*: 56 distinct description langauges identified. Most of them (36%) are in english. The remaining distribution looks pretty much even.
- *hasDescription:* 75% of videos in the sample have description.
- *tagCount, hasTag:* 52.5% of videos in the sample have tags. In general they have less than 10, but the majority has just 1. Extreme values reach up to 108 tags.
- *commentsEnabled:* comments are enabled in 98% of videos of the sample.
- *likesEnabled:* likes are enabled in 98% of the videos of the sample.
- *dislikesEnabled:* dislikes are enabled in 98% of the videos of the sample.
- *viewsToLikes:* 25% of videos have this ratio in zero. 75% have it smaller than 0.05 while the maximum is 6.21
- *viewsToDislikes:* dislikes are way less common. 95% of videos have this ratio below 0.01, while the maximum is 1.00.

- *viewsToComments:* it follows a similar distribution than the variable above.
- *channel stats ratios:* all channel stats divided by video views follow an exponential-like distribution with a very high concentration around 0.
- *deltaViews*: Almost 62% of videos show no new views from one week to the other. The rest of the videos follow a distribution with an exponential-like shape and a maximum of almost 50 million new views.

We now take a look at the existing correlations between the variables of the dataset. In Figure 2.8 we observe some strong relationships between video views and likes, dislikes and comments, which sounds intuitive. Also intuitive, but worth pointing out, the more views a video has, the bigger the number of views one week after. We can also see that the correlation between being trending and delta view supports our hypothesis.

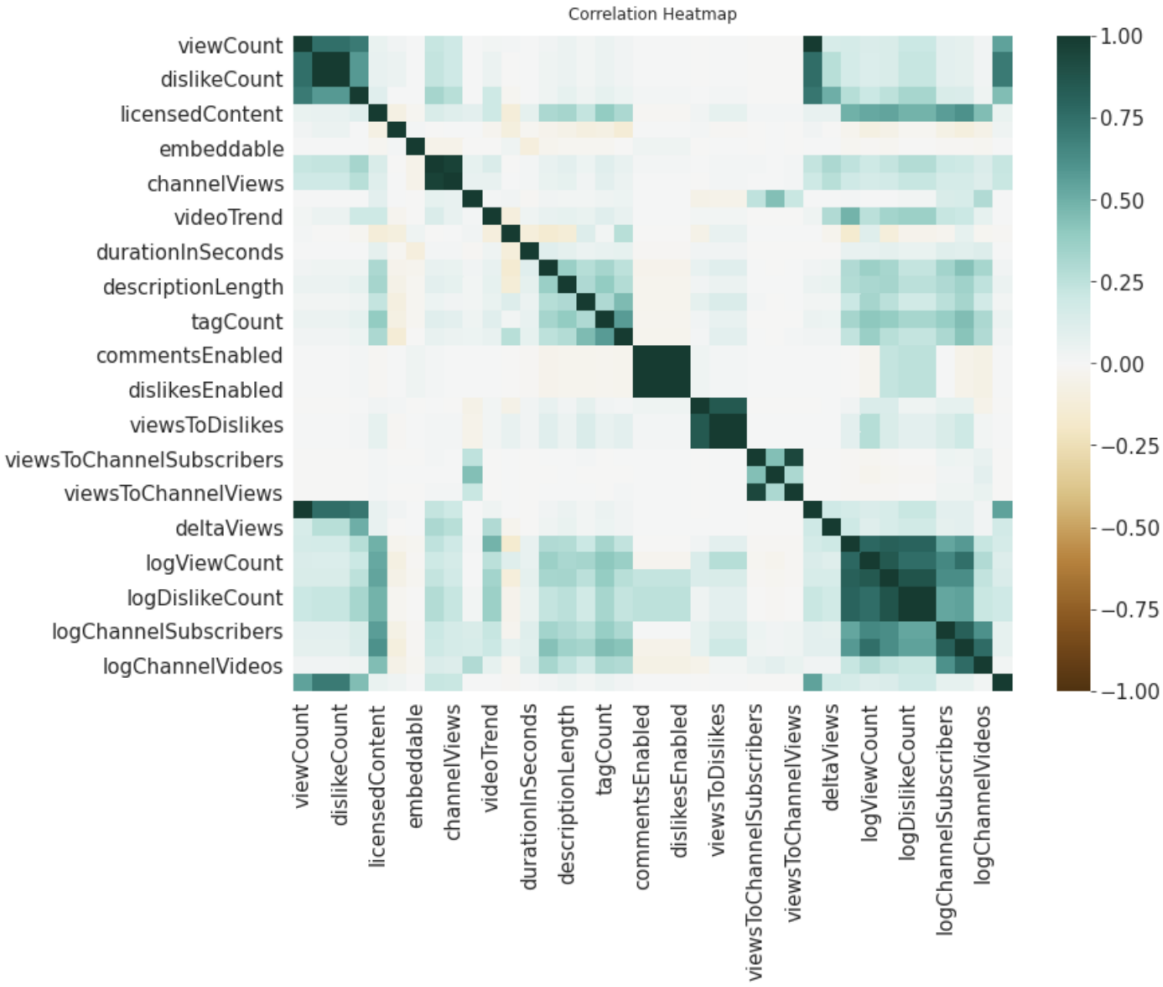

Figure 2.8: Pearson's correlation heatmap.

When we start comparing trending videos with non-trending ones, we observe two key differences, which were previously mentioned: (1) trending videos have higher views at time T and (2) they are quite new in the platform. This can be observed in Figure 2.9, Figure 2.10 and Figure 2.11.

Some non-trending videos show a huge average number of views for some specific "active days". It is pretty likely that these groups of active days include either

a viral video or a former trending video (or both!). However, trending videos are more popular than non-trending in most cases.

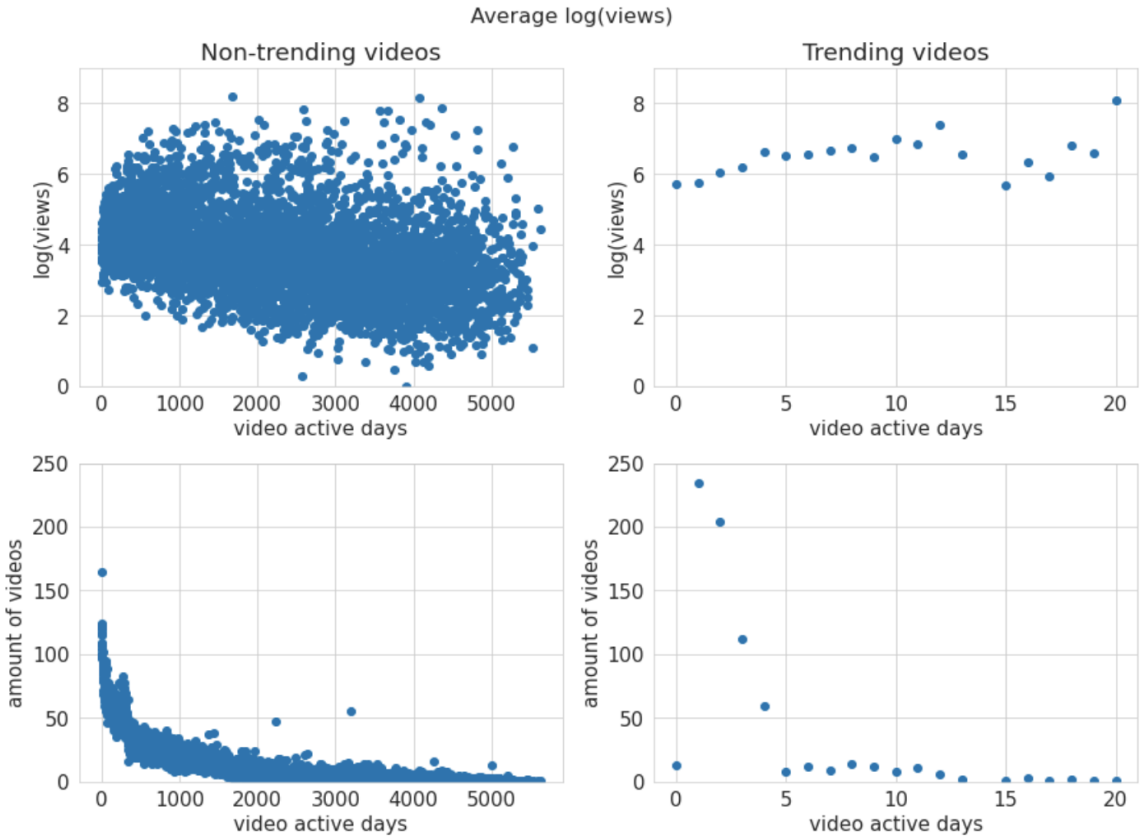

Figure 2.9: Average views by active days (up). Active days distribution (down)

Not only the number of views is higher for trending videos but also their delta views (Figures 2.10 and 2.12).

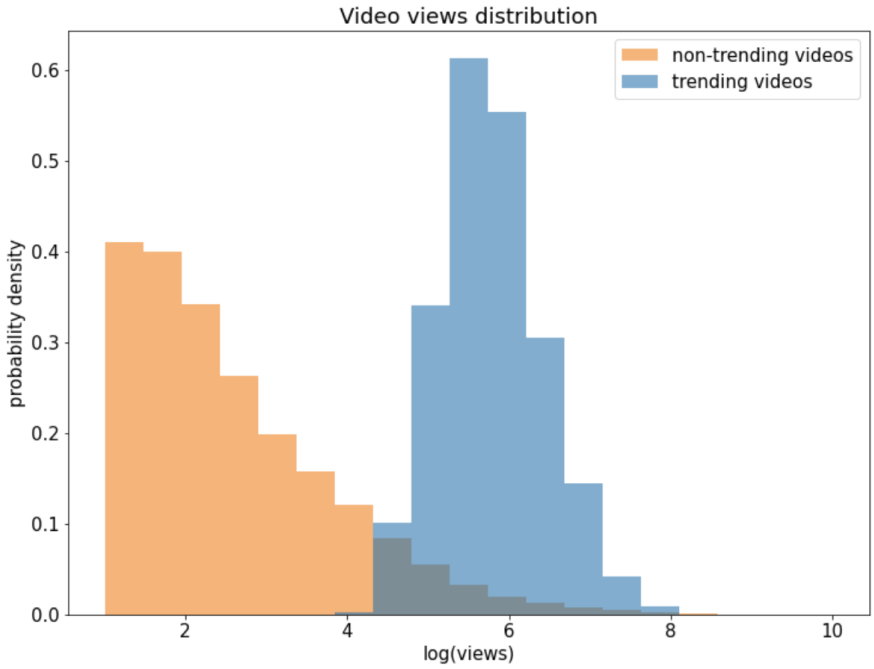

Figure 2.10: $log_{10}(views)$ distribution by type of video.

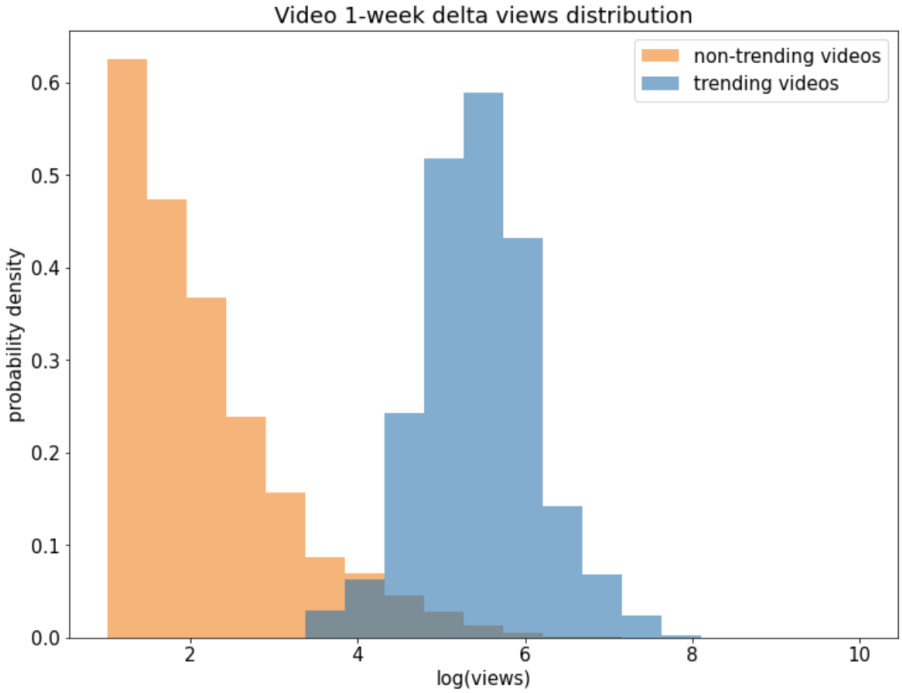

Figure 2.11: $log_{10}(delta\ views)$ distribution by type of video.

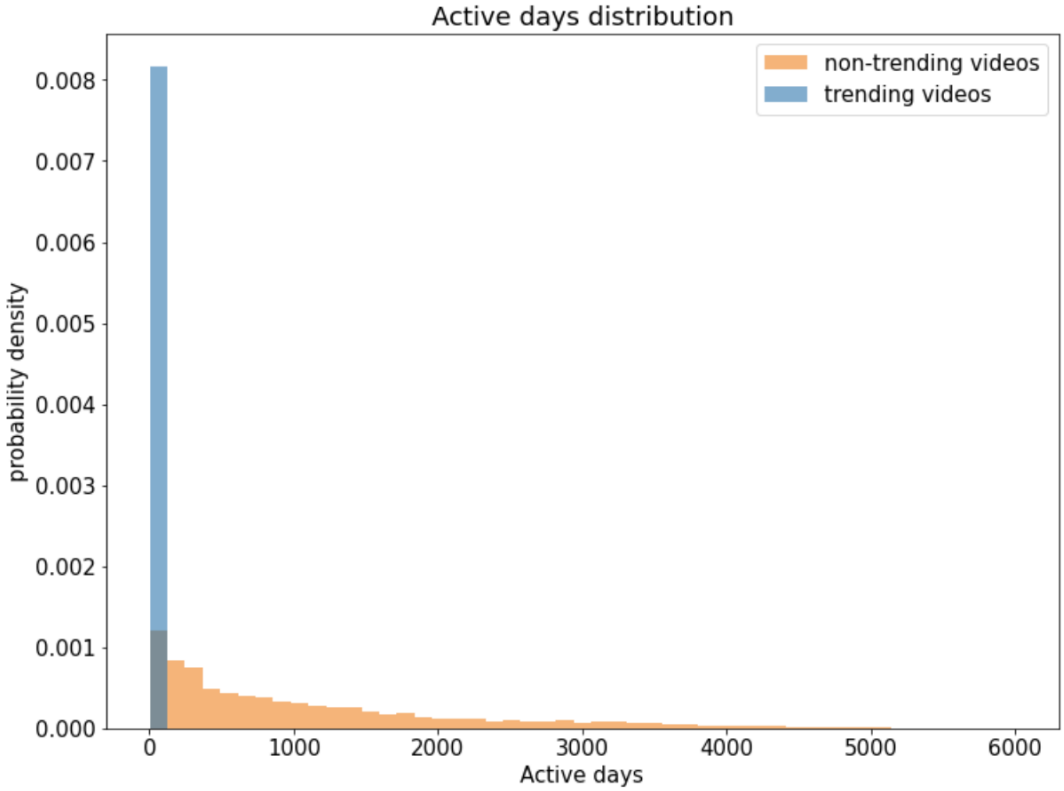

Figure 2.12: Active days distribution by type of video.

Figure 2.12 shows something we have mentioned earlier and it is naturally intuitive: trending videos are young videos. They are videos that have a big level of novelty. On the contrary, non-trending videos gather both new videos and videos that have been uploaded to

the platform even in the year in which YouTube was created (2005, 26 years ago, equivalent to ~6000 days ago).

So far we have seen there is an intuitive correlation between the number of views a video has and the incremental number of views it has the week after. Given the exponential-like distribution of these two variables, we apply logarithm to both of them and graph them in a scatter plot, as shown in figure 2.13:

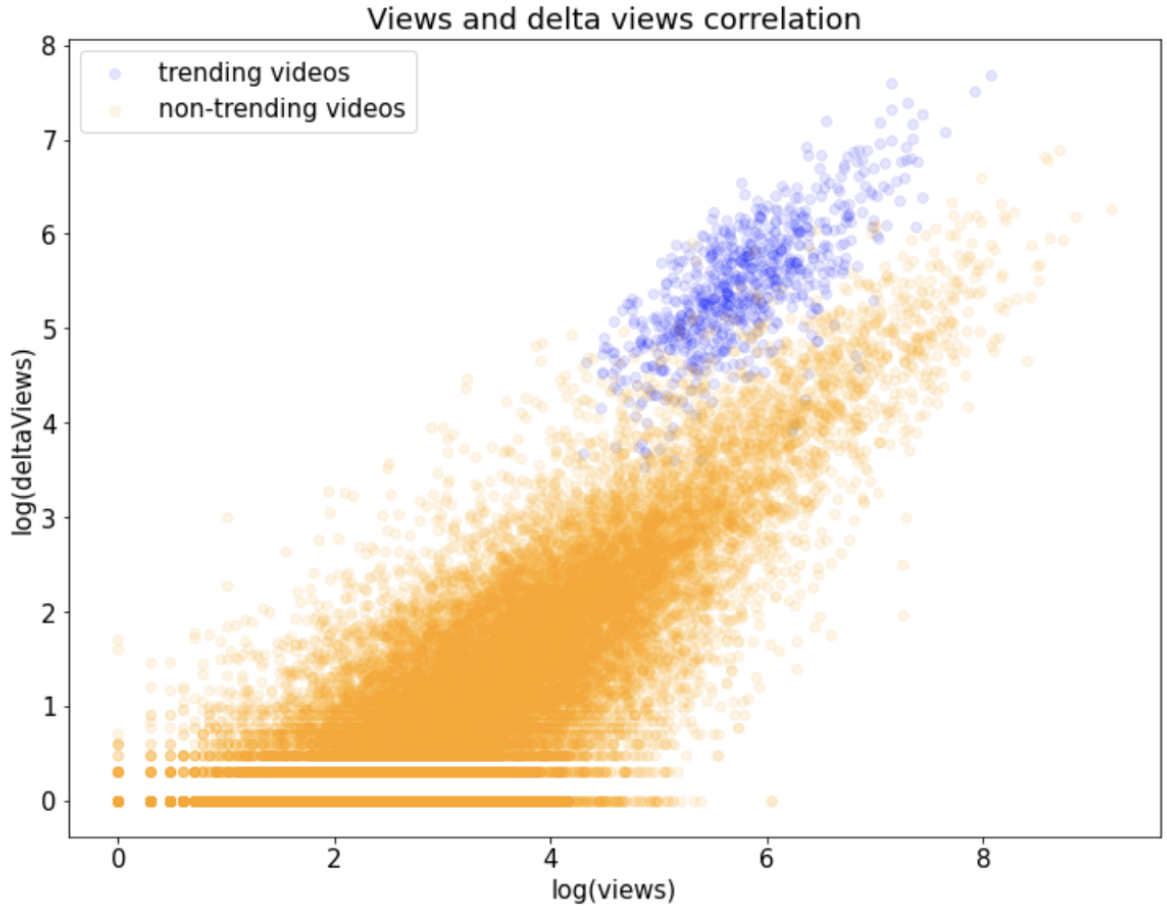

Figure 2.13: $log_{10} - log_{10}$ views and delta views correlation.

What we see in this chart is very insightful at the moment of selecting our inference model. One main conclusion here is that there seems to be a strong linear correlation between these variables when logarithm is applied to both. Moreover, we observe again the difference in delta views between trending and non-trending videos. Since there is logarithm applied on both sides, we have to take into account that the delta views difference between these types of videos increases exponentially in absolute terms as video views are higher. It is also worth mentioning how the distribution of non-trending delta views breaks as views decrease.

3. Trending effect estimation: causal inference analysis

In this section we try to estimate the average effect that being featured as trending has on the views a video receives over a week. We frame this problem with our data as follows: for each video we estimate what is the impact of being a trend at time T into the number of views that video will receive in the period between T and T+1 week. The main challenge of this problem is how to distinguish the views driven by the effect of being trending from the number of views that a video would receive in the absence of that feature.

When framing the problem this way, we find ourselves in an observational study[16] in which we are trying to find the average treatment effect (ATE) over a given population, where the treatment is being trending, the effect is the delta views it drives and the population the YouTube videos. An observational study is one in which the objective is to draw inference from a sample of the population and the independent variables are not under the control of the researchers. The latter is what differentiates observational studies from experiments. Formally, we can write this in the following way:

$$r_{T_i} = r_{C_i} + \tau$$

Equation 3.1: Additive treatment effect model.

where r_{T_i} represents the delta views of video i with the treatment (i.e. being featured in Trending section), r_{C_i} represents the delta views of the same video unexposed to the treatment (i.e. not being featured in Trending) and τ the ATE. This is an additive treatment effect model and assumes observations do not interfere with each other, and that the treatment raises the response of a unit by a constant number τ (i.e. being featured as trending adds a constant number of views regardless of the nature of the video).

From what we discussed in the previous sections, this is probably not the most adequate model for our problem. As we expect that being featured as trending increases exposure of a video (therefore, its discoverability) by a constant number of users without modifying its likelihood of being played once discovered, the effect Trending has over views has to be proportional to the number of views it would have had if unexposed (Equation 1.5). In this sense, a multiplicative treatment effect model looks like the best fit for our case:

$$r_{T_i} = \delta r_{C_i}$$

Equation 3.3: Multiplicative treatment effect model.

Here we assume that the effect of the treatment is proportional to the counterfactual state: the higher δ is, the bigger the difference in delta views a video will have if it is treated or not. More important, the higher r_{C_i} is, the higher the effect. By taking logarithms, we can bring this model to an additive setting:

$$log(r_{T_i}) = log(r_{C_i}) + log(\delta)$$

Equation 3.4: Multiplicative treatment effect model transformed into an additive setting.

[16] W. G. Cochran and S. Paul Chambers, 1965. *The Planning of Observational Studies of Human Populations.* Series A (General) Vol. 128, No. 2, Journal of the Royal Statistical Society.

In the following section we explore and compare the two types of models. It is worth making explicit that since this is an observational study, an observation (i.e. a video) is either treated or not (i.e. trending or non-trending), so we actually do not have a counterfactual observation to calculate ATE. We use statistical inference methods to close this gap: for each video we model what the number of views would be for each state (treatment and control) and then estimate the ATE by computing the difference between them.

3.1 Naive estimation

Since we are dealing with an observational study, the treatment has not been assigned randomly across the population like it would have happened in a controlled experiment or a randomized controlled trial (informally called AB test). Under this setting, is pretty likely that we have confounders[17] and overt bias[18] that we need to address at the moment of estimating the ATE.

A confounder is a variable that influences both the dependent variable and independent variable, causing spurious correlations. Confounders introduce bias when omitted from the model, known as omitted variable bias. On the contrary, overt bias is one that can be seen in the data at hand, even when not omitting confounders; this happens when the distribution of covariates between treatment and control differ between them.

An example of overt bias in this case can be the one introduced by the number of views: trending videos naturally have more views, on average, than non-trending ones and, as we have seen, having more views positively correlates with delta views. Views also correlate with being trending, so there's a need of controlling the ATE by this variable. When doing so on a regular regression, the effect controlled by it is flawed due to the small overlap between their distributions (overt bias). Omitting this variable, however, drives omitted variable bias.

Ignoring these disclaimers, we proceed to calculate the ATE as if this study was an AB test. In an AB test, from the total number of videos that fulfills YouTube's requirements for being selected as trending, only 50% would be actually tagged as such, and in a random fashion. By doing this, we make sure that control and treatment groups (group A and group B, correspondingly) belong to the same population, therefore there are no biases to account for at the moment of calculating the ATE. In this setup, in order to calculate ATE we run the following linear regression (we could have performed a mean difference test as well):

$$delta\ views_i = \beta_0 + \beta_1 X_i + \varepsilon$$

Equation 3.5: Single-variable regression model taking an additive effect approach.

where X_i indicates whether or not the video i is trending (boolean variable), β_1 the incremental average number of views being trending adds to $delta\ views_i$, β_0 the average number of delta views a non-trending video has and ε random noise. The result of this model is shown in Figure 3.1:

[17] Paul R. Rosenbaum, December 1991. *Discussing Hidden Bias in Observational Studies*. Annals of Internal Medicine.

[18] Paul R. Rosenbaum. *Overt Bias in Observational Studies*. Springer Series in Statistics book series (SSS).

OLS Regression Results

Dep. Variable:	deltaViews	R-squared:	0.086
Model:	OLS	Adj. R-squared:	0.086
Method:	Least Squares	F-statistic:	6049.
Date:	Tue, 23 Feb 2021	Prob (F-statistic):	0.00
Time:	23:14:05	Log-Likelihood:	-9.1492e+05
No. Observations:	64400	AIC:	1.830e+06
Df Residuals:	64398	BIC:	1.830e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1928.5597	1418.105	1.360	0.174	-850.927	4708.047
videoTrend[T.True]	1.048e+06	1.35e+04	77.777	0.000	1.02e+06	1.07e+06

Omnibus:	238354.964	Durbin-Watson:	1.910
Prob(Omnibus):	0.000	Jarque-Bera (JB):	195566780683.424
Skew:	80.469	Prob(JB):	0.00
Kurtosis:	8538.580	Cond. No.	9.56

Figure 3.1: Single-variable linear regression results.[19]

The coefficient of this regression is statistically significant (|P| < 0.05).[20] In a controlled experiment, like an AB test, we could interpret this coefficient as follows: Trending feature gives, on average, 1.05 million more views to videos. We could get the same conclusion by observing Figure 3.2. We know, however, that due to all the possible biases that live in observational studies this conclusion is most likely untrue.

[19] https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html python library used for OLS.

[20] In this thesis we will conclude a result is statistically significant if the p-value of the estimator is lower than 5%.The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

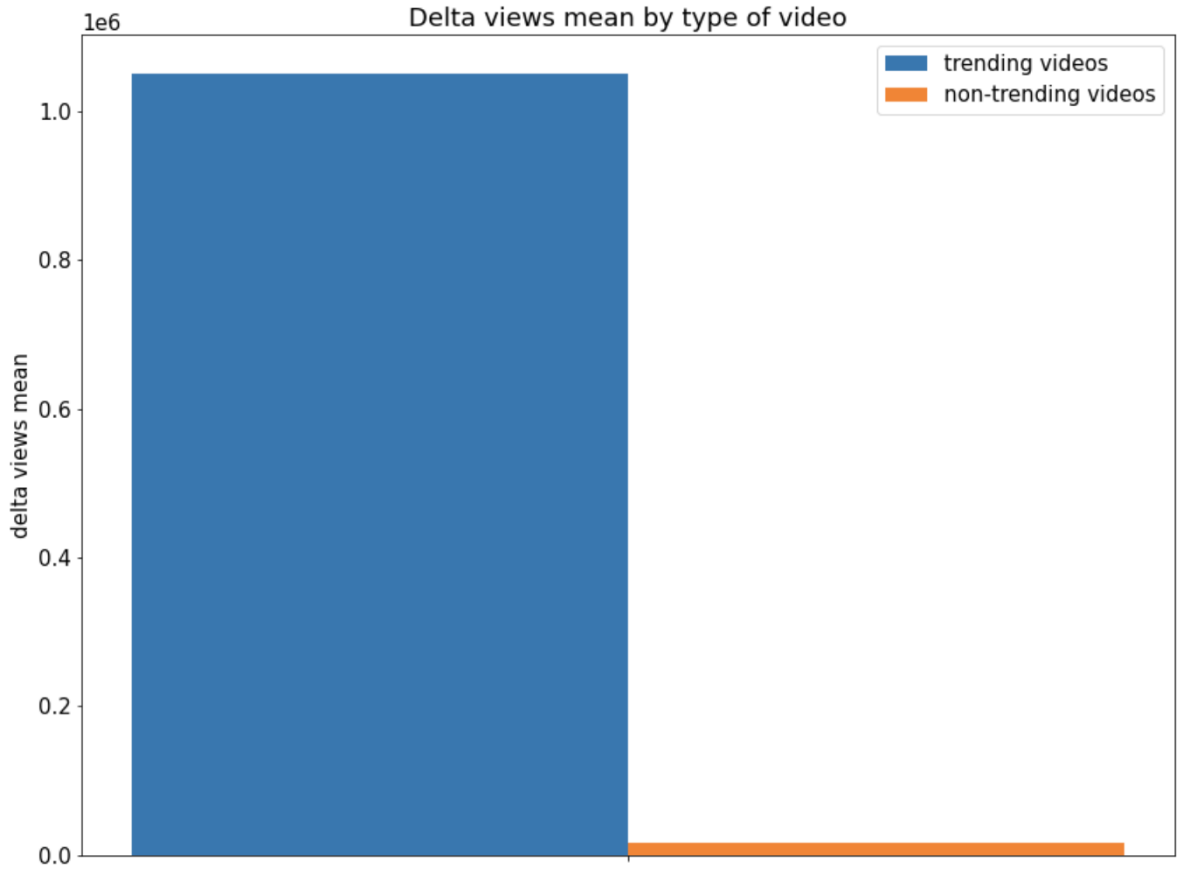

Figure 3.2: Average delta views by type of video.

As mentioned in the previous section, if we want to model a multiplicative effect (instead of an additive one) we change $delta\ views_i$ of the regression for $log(delta\ views_i)$ and end up solving the following linear regression:

$$log(delta\ views_i) = \beta_0 + \beta_1 X_i + \varepsilon$$

Equation 3.6: Single-variable regression model taking a multiplicative effect approach.

This regression also draws a statistically significant conclusion, as seen in Figure 3.3 and backed up by figure 3.4:

OLS Regression Results

Dep. Variable:	logDeltaViews	R-squared:	0.238
Model:	OLS	Adj. R-squared:	0.238
Method:	Least Squares	F-statistic:	2.007e+04
Date:	Sat, 17 Apr 2021	Prob (F-statistic):	0.00
Time:	18:20:53	Log-Likelihood:	-1.3965e+05
No. Observations:	64399	AIC:	2.793e+05
Df Residuals:	64397	BIC:	2.793e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2363	0.008	147.433	0.000	1.220	1.253
videoTrend[T.True]	11.2890	0.080	141.652	0.000	11.133	11.445

Omnibus:	31124.959	Durbin-Watson:	1.164
Prob(Omnibus):	0.000	Jarque-Bera (JB):	160065.429
Skew:	2.380	Prob(JB):	0.00
Kurtosis:	9.082	Cond. No.	9.56

Figure 3.3: Single-variable linear log-regression results.

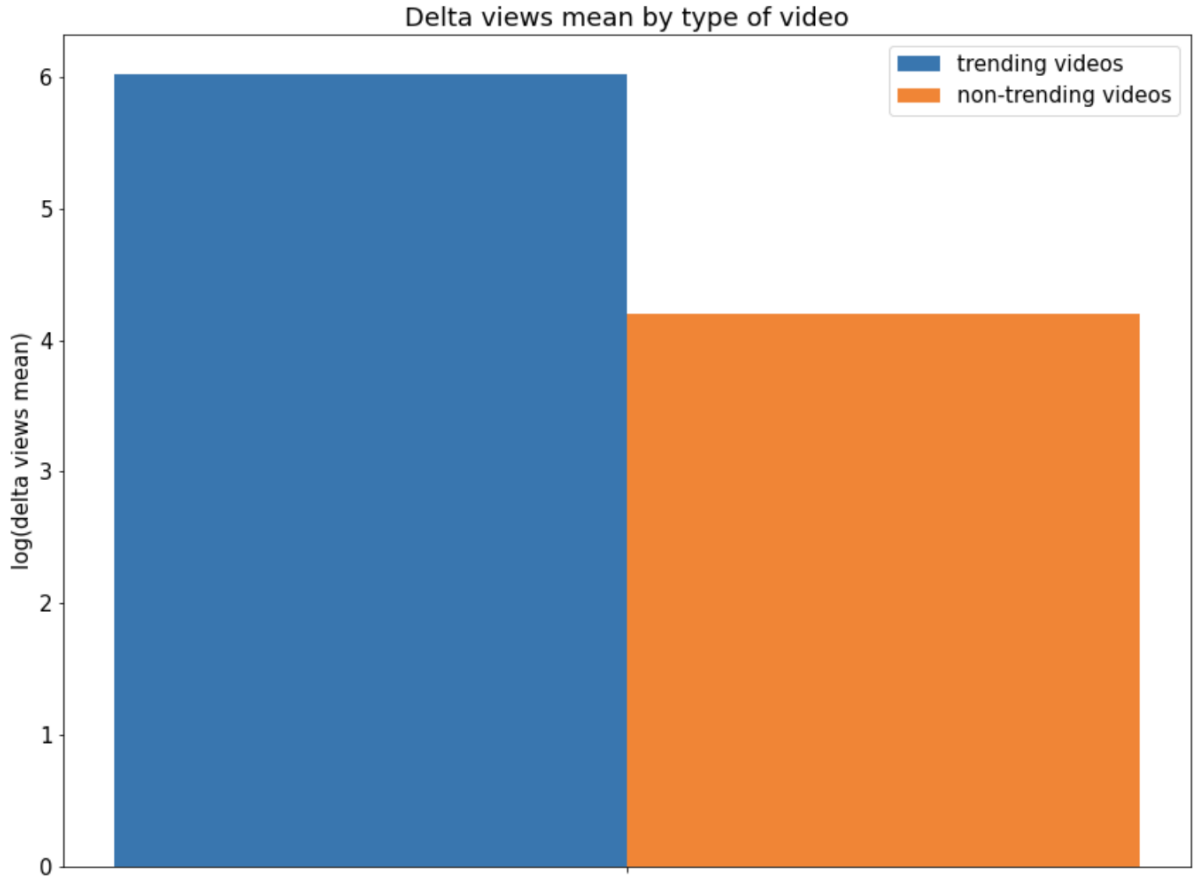

Figure 3.4: Average $log_{10}(delta\ views)$ by type of video.

We can interpret the result of this regression as follows: average log(delta views) of trending videos is higher than that for non-trending videos by 11.3 units. On top of the descriptive analysis conclusions earlier shared, the statistics of these two regressions also suggest than modeling this problem with a multiplicative approach better represents the real

system: R-squared for the log-regression is 2.8 times higher than that for the regular regression, meaning that the trend variable on this model linearly explains variance almost 3 times better. On top of that, AIC (Akaike information criterion) for the log-regression is ~85% smaller, giving more support to this statement.

3.2 Multivariable Linear Regression

In order to remove bias introduced by omitted variables (i.e. confounders), we add covariates to the regression to control for all these excluded effects. We can visualize this phenomenon with the following scenario:

> Viral videos get way more views than regular videos. If we compare delta views between viral and non-viral videos, we should expect to see a huge difference in favor of viral videos. Now, it is very likely that viral videos will end up on the trending section of YouTube. If we do not control covariates and interpret the coefficient of a simple regression as the effect of being trending over views, we will conclude that this effect is drastically bigger than it should. This is because by omitting confounders the ATE of our model ends up absorbing that uncontrolled effect generating spurious correlations.

When building the multivariable linear regression model to control ATE by covariates we need to make sure we do not introduce variables that generate multicollinearity. Multicollinearity refers to a situation in which more than two explanatory variables in a multivariable regression model are highly linearly related; avoiding this is one of the assumptions for OLS (Ordinary Least Squares) regression. The conclusion one can make with an OLS model that has multicollinearity[21] can be totally misleading since it may not affect the accuracy of the model as much, but we might lose reliability in determining the effects of individual features in our model—and that can be a problem when it comes to interpretability. For instance, it is very likely that the sign of the estimator ends up being in the opposite direction than it should: in our case, the estimator of trending may end up being negative if we do not account for multicollinearity!

To start with, we do not include any of the views-based ratios added to our dataset since they have an indirect correlation with video views, one of the most important variables based on our domain knowledge. Having a high number of views is a requisite for being trending, and apparently (and intuitively), the more views a video has, the higher the delta views will be. We are not stating this as truth, but it is logical to think that way and the descriptive analysis provides evidence to think this hypothesis is reasonable. For this reason, we want it to be in our model.

In addition to these variables, we will also ditch covariates whose VIF (Variance Inflation Factor) is higher than 5. VIF score of an independent variable represents how well the variable is explained by other independent variables: the higher the score, the stronger the multicollinearity. It is predicted by taking a variable and regressing it against every other variable:

[21] Jamal I. Daoud. 2017. *Multicollinearity and Regression Analysis.* Journal of Physics, IOP Publishing Ltd.

$$VIF = \frac{1}{1 - R^2}$$

Equation 3.7: VIF formula as a function of R-squared.

A rule of thumb for interpreting VIF states that: (1) VIF = 1 implies no correlation; (2) between 1 and 5, the variable is moderately correlated; (3) greater than 5, it is highly correlated.[22]

After doing this, we get rid of *commentCount* and *channelSubscribers*. It should not be a surprise these variables were selected after analyzing Figure 2.8.

3.2.1 Linear-linear regression

Now we have the set of variables we want to use on our model. We will start with a linear-linear regression: numerical variables used are not transformed with a logarithm. In other words, we build an additive treatment effect model. The regression model has this form:

$$delta\ views_i = \beta_0 + \beta_1 X_{1i} + \sum_{j>1}^{covariates} \beta_j X_{ji} + \varepsilon$$

Equation 3.8: Multivariable regression model taking an additive effect approach.

where X_{ji} is the value of the variable j for the video i and β_j the estimator for each of those variables. $\beta_1$is the ATE while $\beta_{j\neq1}$are the estimators that control for covariates. Result of this model is found in Figure 3.5.

[22] Dodge, Y., 2008. *The Concise Encyclopedia of Statistics*. Springer.

OLS Regression Results

Dep. Variable:	deltaViews	R-squared:	0.410
Model:	OLS	Adj. R-squared:	0.409
Method:	Least Squares	F-statistic:	301.5
Date:	Sat, 17 Apr 2021	Prob (F-statistic):	0.00
Time:	19:52:50	Log-Likelihood:	-9.0082e+05
No. Observations:	64399	AIC:	1.802e+06
Df Residuals:	64250	BIC:	1.803e+06
Df Model:	148		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.279e+05	1.21e+05	1.061	0.289	-1.08e+05	3.64e+05
videoTrend[T.True]	5.498e+05	1.16e+04	47.369	0.000	5.27e+05	5.73e+05
likesEnabled[T.True]	2.856e+04	3.02e+04	0.947	0.344	-3.06e+04	8.77e+04
dislikesEnabled[T.True]	2.856e+04	3.02e+04	0.947	0.344	-3.06e+04	8.77e+04
licensedContent[T.True]	-3.018e+04	3464.905	-8.711	0.000	-3.7e+04	-2.34e+04
embeddable[T.True]	2.436e+04	1.04e+04	2.350	0.019	4040.018	4.47e+04
hasDescription[T.True]	262.7058	7893.039	0.033	0.973	-1.52e+04	1.57e+04
hasTag[T.True]	-1573.8944	3328.663	-0.473	0.636	-8098.076	4950.288
commentsEnabled[T.True]	2.856e+04	3.02e+04	0.947	0.344	-3.06e+04	8.77e+04
madeForKids[T.True]	-5131.5622	5324.984	-0.964	0.335	-1.56e+04	5305.411
C(liveBroadcastContent)[T.none]	-2.346e+05	1.95e+05	-1.202	0.229	-6.17e+05	1.48e+05
C(liveBroadcastContent)[T.upcoming]	-9.267e+04	1.4e+05	-0.664	0.507	-3.66e+05	1.81e+05
C(categoryId)[T.10]	2.741e+04	7976.211	3.436	0.001	1.18e+04	4.3e+04
C(categoryId)[T.15]	8425.8347	1.48e+04	0.570	0.569	-2.05e+04	3.74e+04
C(categoryId)[T.17]	-1.038e+04	9611.593	-1.080	0.280	-2.92e+04	8454.181
C(categoryId)[T.19]	1.163e+04	1.24e+04	0.939	0.348	-1.26e+04	3.59e+04
C(categoryId)[T.2]	1.217e+04	9981.198	1.219	0.223	-7395.452	3.17e+04
C(categoryId)[T.20]	1.074e+04	7713.550	1.392	0.164	-4379.562	2.59e+04
C(categoryId)[T.22]	7411.4970	7292.812	1.016	0.310	-6882.421	2.17e+04
C(categoryId)[T.23]	-1.587e+04	1.06e+04	-1.496	0.135	-3.67e+04	4928.366
C(categoryId)[T.24]	-1.854e+04	8024.445	-2.310	0.021	-3.43e+04	-2810.120
C(categoryId)[T.25]	5948.1843	1.05e+04	0.565	0.572	-1.47e+04	2.66e+04
C(categoryId)[T.26]	1.357e+04	1e+04	1.355	0.176	-6063.543	3.32e+04
C(categoryId)[T.27]	1.021e+04	8518.427	1.199	0.231	-6486.499	2.69e+04
C(categoryId)[T.28]	1.108e+04	9785.352	1.132	0.258	-8099.583	3.03e+04
C(categoryId)[T.29]	9613.8703	1.46e+04	0.659	0.510	-1.9e+04	3.82e+04
C(categoryId)[T.44]	3.539e+04	2.88e+05	0.123	0.902	-5.29e+05	6e+05
C(dimension)[T.3d]	-4510.8312	1.66e+05	-0.027	0.978	-3.31e+05	3.22e+05
C(projection)[T.rectangular]	-7762.7864	7.22e+04	-0.108	0.914	-1.49e+05	1.34e+05
C(uploadStatus)[T.uploaded]	-1.434e+05	1.46e+05	-0.979	0.328	-4.31e+05	1.44e+05
C(license)[T.youtube]	-3274.6641	1.18e+04	-0.277	0.782	-2.65e+04	1.99e+04
C(titleLanguage)[T.ar]	-3815.1773	1.53e+04	-0.249	0.804	-3.39e+04	2.63e+04
C(titleLanguage)[T.bg]	-1836.1075	1.94e+04	-0.095	0.925	-3.98e+04	3.62e+04
C(titleLanguage)[T.bn]	-7723.0461	2.72e+04	-0.284	0.777	-6.11e+04	4.56e+04
viewCount	-0.0234	0.000	-103.051	0.000	-0.024	-0.023
likeCount	-4.965e+04	4.53e+04	-1.097	0.273	-1.38e+05	3.9e+04
likeCount:likesEnabled[T.True]	4.965e+04	4.53e+04	1.097	0.273	-3.9e+04	1.38e+05
dislikeCount	-4.964e+04	4.53e+04	-1.097	0.273	-1.38e+05	3.9e+04
dislikeCount:dislikesEnabled[T.True]	4.966e+04	4.53e+04	1.097	0.272	-3.9e+04	1.38e+05
channelViews	2.128e-05	4.8e-07	44.324	0.000	2.03e-05	2.22e-05
channelVideos	-0.0858	0.053	-1.627	0.104	-0.189	0.018
activeDays	0.6051	1.220	0.496	0.620	-1.787	2.997
titleLength	204.5085	58.120	3.519	0.000	90.594	318.423
descriptionLength	6.3535	1.810	3.510	0.000	2.805	9.902
tagCount	-387.2528	142.966	-2.709	0.007	-667.467	-107.039
durationInSeconds	-1.1127	0.623	-1.785	0.074	-2.334	0.109

Omnibus:	207410.205	Durbin-Watson:	1.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	75238162295.015
Skew:	53.116	Prob(JB):	0.00
Kurtosis:	5297.168	Cond. No.	7.78e+16

Figure 3.5: Multivariable linear lin-lin regression results. Estimators for some of the title and description languages were omitted in the figure.

From these results we can conclude that there is a positive ATE: being trending, apparently, increments weekly views by 0.55 million units. This is half the value we would get if we did not control by any covariate (Figure 3.1), meaning that there were definitely confounders we were excluding. One of those, as expected, is the number of views the video has at time T: this variable is positively related both with the delta views and with being trend, so it is expected to see a reduction in the ATE obtained when adding it to the model. By adding the covariates the model also fits the data considerably better (adjusted R-squared of 0.41 vs 0.09, 4.6 times higher). We repeat this analysis but using a log-log approach.

3.2.2 Log-log regression

As we have been seeing throughout this work, a multiplicative treatment effect model seems more suitable to calculate the ATE. This justifies applying logarithm to delta views, but we also think it convenient for channel and video statistics based on what we have seen in the descriptive analysis (Figure 2.13). To transform the former lin-lin regression model we apply logarithms to delta views and to all video and channel statistics: views, comments, likes, dislikes, channel subscribers, channel views, channel videos:

$$log(delta\ views_i) = \beta_0 + \beta_1 X_{1i} + \sum_{j>1}^{stats} \beta_j log(X_{ji}) + \sum_{j>stats}^{covariates} \beta_j X_{ji} + \varepsilon$$

Equation 3.9: Multivariable regression model taking a multiplicative effect approach.

In order to avoid a mathematical indetermination we sum 1 unit to all statistics before applying logarithms. Since the distribution of our variables has now changed due to the application of logarithms, we need to verify VIF once again. After running the VIF analysis, we drop the following covariates: *logCommentCount, logChannelViews, logLikeCount, logDislikeCount, logChannelVideos.* Running the regression with the remaining variables drops the following results:

OLS Regression Results

Dep. Variable:	logDeltaViews	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.759
Method:	Least Squares	F-statistic:	1403.
Date:	Sat, 17 Apr 2021	Prob (F-statistic):	0.00
Time:	20:41:33	Log-Likelihood:	-1.0244e+05
No. Observations:	64399	AIC:	2.052e+05
Df Residuals:	64253	BIC:	2.065e+05
Df Model:	145		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.8136	0.872	6.671	0.000	4.105	7.522
videoTrend[T.True]	6.2224	0.048	130.905	0.000	6.129	6.316
licensedContent[T.True]	0.2757	0.016	17.303	0.000	0.244	0.307
embeddable[T.True]	-0.0439	0.043	-1.026	0.305	-0.128	0.040
hasDescription[T.True]	0.1777	0.033	5.451	0.000	0.114	0.242
hasTag[T.True]	0.0907	0.014	6.597	0.000	0.064	0.118
likesEnabled[T.True]	0.0767	0.012	6.190	0.000	0.052	0.101
dislikesEnabled[T.True]	0.0767	0.012	6.190	0.000	0.052	0.101
commentsEnabled[T.True]	0.0767	0.012	6.190	0.000	0.052	0.101
madeForKids[T.True]	0.0745	0.022	3.395	0.001	0.031	0.117
C(liveBroadcastContent)[T.none]	-7.0970	0.806	-8.809	0.000	-8.676	-5.518
C(liveBroadcastContent)[T.upcoming]	-1.9930	0.576	-3.458	0.001	-3.123	-0.863
C(categoryId)[T.10]	0.0147	0.033	0.445	0.656	-0.050	0.079
C(categoryId)[T.15]	-0.1411	0.061	-2.312	0.021	-0.261	-0.021
C(categoryId)[T.17]	-0.4351	0.040	-10.961	0.000	-0.513	-0.357
C(categoryId)[T.19]	-0.3664	0.051	-7.160	0.000	-0.467	-0.266
C(categoryId)[T.2]	-0.0983	0.041	-2.385	0.017	-0.179	-0.018
C(categoryId)[T.20]	-0.2861	0.032	-8.952	0.000	-0.349	-0.223
C(categoryId)[T.22]	-0.1792	0.030	-5.933	0.000	-0.238	-0.120
C(categoryId)[T.23]	-0.0538	0.044	-1.227	0.220	-0.140	0.032
C(categoryId)[T.24]	-0.0491	0.033	-1.483	0.138	-0.114	0.016
C(categoryId)[T.25]	-0.7420	0.043	-17.080	0.000	-0.827	-0.657
C(categoryId)[T.26]	-0.2826	0.041	-6.835	0.000	-0.364	-0.202
C(categoryId)[T.27]	-0.2085	0.035	-5.923	0.000	-0.277	-0.139
C(categoryId)[T.28]	-0.1123	0.040	-2.779	0.005	-0.191	-0.033
C(categoryId)[T.29]	-0.4425	0.060	-7.345	0.000	-0.561	-0.324
C(categoryId)[T.44]	-1.4497	1.190	-1.218	0.223	-3.782	0.882
C(dimension)[T.3d]	-0.2332	0.687	-0.339	0.734	-1.580	1.114
C(projection)[T.rectangular]	0.2315	0.298	0.777	0.437	-0.352	0.816
C(uploadStatus)[T.uploaded]	-3.3023	0.605	-5.459	0.000	-4.488	-2.117
C(license)[T.youtube]	-0.0589	0.049	-1.207	0.228	-0.155	0.037

logViewCount	0.5458	0.002	252.428	0.000	0.542	0.550
logChannelSubscribers	-0.0163	0.002	-9.709	0.000	-0.020	-0.013
activeDays	-0.0005	5.15e-06	-99.326	0.000	-0.001	-0.001
titleLength	-0.0027	0.000	-10.870	0.000	-0.003	-0.002
descriptionLength	3.108e-05	7.48e-06	4.155	0.000	1.64e-05	4.57e-05
tagCount	0.0026	0.001	4.427	0.000	0.001	0.004
durationInSeconds	-1.766e-05	2.58e-06	-6.852	0.000	-2.27e-05	-1.26e-05

Omnibus:	5772.809	Durbin-Watson:	1.607
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14919.808
Skew:	0.523	Prob(JB):	0.00
Kurtosis:	5.113	Cond. No.	1.41e+16

Figure 3.6: Multivariable log-log regression results. Estimators of title and description languages were omitted in the figure.

What we see in Figure 3.6 is consistent with the results of the lin-lin regression: (1) the ATE is statistically significant and (2) trending regression estimator is reduced by ~50% when controlled by covariates (6.2 units vs 11.3 units). In terms of absolute delta views, however, this reduction is massively bigger than the one observed with the lin-lin approach.

What is also consistent with the naive ATE calculation is that the log-log regression model shows better statistics: the adjusted R-squared for the log-log multivariable regression is 85% higher than that for the lin-lin multivariable regression—and it is actually pretty high — and its AIC is ~89% smaller.

Based on these results, we chose the log-log approach for calculating ATE.

3.2.3 Flaws of OLS regression models

We try to remove bias with OLS regression models. The concern with this approach is that control group observations may be very different from the treated units. In other words, the distribution of covariates may have no overlap between treatment and control, and this can lead to misleading results and conclusions. Regression models estimate the counterfactual of a sample in regions in which observations have similar characteristics and differ on the treatment. Lack of similar videos between trending and non-trending groups can harm the estimation. We have seen in Figures 2.9 and 2.10, for instance, that two of the key variables of the system have very little overlap between traetment and control.

Figure 3.7 shows the level of overlap between the numerical covariates. *Y* represents *deltaViews* and x_0 to x_{15} our numerical variables. The normalized difference is calculated with the following formula and is a rough estimate of how far the distributions of two variables are:

$$Normalized\ difference\ = \frac{\bar{X}_{T-}\bar{X}_C}{(\sigma_T^2 + \sigma_C^2)/2}$$

Equation 3.10: Normalized difference between treatment and control covariates.

where $\bar{X}_T$ is the sample mean of the variable x for the treatment, $\bar{X}_C$ is the sample mean of variable x for the control, σ_T the sample standard deviation of the variable x for the treatment and σ_C the sample standard deviation of the variable x for the control.

Normalized differences close to zero mean the level of overlap is high. This is not the case for many of the numerical covariates of our dataset.

Variable	Controls (N_c=63686) Mean	S.d.	Treated (N_t=713) Mean	S.d.	Raw-diff
Y	1928.590	58625.459	1050164.377	3358067.873	1048235.787

Variable	Controls (N_c=63686) Mean	S.d.	Treated (N_t=713) Mean	S.d.	Nor-diff
X0	291321.432	8968299.696	2056241.289	6589409.579	0.224
X1	1089.193	1090.276	2.893	2.902	-1.409
X2	185.264	6226.556	3346.760	14949.474	0.276
X3	1919.392	40039.946	96847.673	261645.094	0.507
X4	183158.825	2849590.399	4134347.736	11493293.275	0.472
X5	118469644.851	2330448411.069	1748478787.680	8402887786.114	0.264
X6	2143.515	22007.700	4019.533	22070.334	0.085
X7	717.793	1892.405	1308.328	2764.770	0.249
X8	39.472	23.694	54.886	23.606	0.652
X9	380.494	746.009	934.997	843.801	0.696
X10	6.306	10.540	16.888	14.087	0.851
X11	0.017	0.221	0.058	0.045	0.254
X12	-0.031	0.182	0.001	0.002	0.251
X13	-0.031	0.182	0.001	0.002	0.251
X14	231.940	12157.803	4.450	9.864	-0.026
X15	184.011	6719.155	0.011	0.058	-0.039

Figure 3.7: Overlap between treatment and control numerical covariates.

Lack of overlap is not the only flaw OLS regression models have for inference. Regression approaches involve fine tuning (e.g. transformations of variables and variable selection). One is looking at the results when fine tuning, potentially leading to researcher bias, which is definitely our case.[23]

As a consequence of these issues, we proceed to complement the bare OLS regression model with matching techniques.

3.3 Propensity Score & Matching

In an ideal scenario we would have two parallel worlds: one in which all applicable videos have the treatment and other in which they do not. The estimation of the ATE would be very simple then: you would just need to compute the delta views difference between treated and not treated versions of each video then calculate the average.

The reality is that such a scenario does not exist, but we can try to approach it as much as we can. We can try to compare trending and non-trending videos that look alike. Maybe they do not have the exact same number of views, or the same title, but they are still very similar. This technique is called Propensity Score Matching. It requires calculating the probability of observations of being part of the treatment group based on their covariates (we call this probability the propensity score) and then matching treatment and control observations based on those probabilities: observations between treatment and control that have similar scores are paired together.[24]

3.3.1 Propensity Score Matching

We will start this section with an important disclaimer: this technique works if there is a good degree of overlap between treatment and control propensity scores. If the overlap is

[23] Santiago Gallino, August 2020. Universidad Torcuato Ditella. Observational Studies lecture.

[24] Paul R. Rosenbaum & Donalds B. Rubin, December 1983. *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.* The American Statistician.

very low—or null—it will be hard or impossible to generate a sufficient number of pairs of matches. We already know that the covariate overlap is low, but that does not necessarily mean that the propensity scores will follow the same behavior. On the other hand, if the scores between the two groups are distributed similarly, it is likely that this technique will have no effect at all, as the matching will keep all the samples. The sweet spot lives in between these scenarios.

After successfully applying propensity score matching one should see that the distribution of covariates between matched control and treatment observations have a decent degree of overlap. To get a good sense of how good the matching worked we could recalculate the statistics of Figure 3.7 with the matched observations; if the matching worked we should see a normalized difference close to zero for all covariates. Of course, there are formal hypothesis tests to statistically validate two distributions belong to the same population.

In our case, propensity score is the conditional probability of a video being part of the treatment (i.e. being trending):

$$P(treatment/X = x_i)$$

Equation 3.11: Conditional probability of a video being trending.

where x_i represents the covariates of video i. We use a logistic regression to estimate this probability. The logistic model for our problem looks as it follows:

$$P(treatment = 1/X = x_i) = \frac{1}{1-e^{-(\beta_0+\sum_{j>1}^{covariates}\beta_j X_{ji})}}$$

Equation 3.12: Multivariable logistic regression.

A Maximum Likelihood Estimation is used for estimating the parameters in the linear expression of the logistic model.[25] We continue using the same nomenclature as the regressions analyzed above; observe that the component referred to the treatment is not present in the linear expression of the formula, since it is what we are trying to predict.

Once we have the propensity scores we check how much they overlap. We use Figure 3.8 to do a first sanity check. In this graph we can see the propensity score distribution before matching observations, i.e. a comparison between treatment and control propensity scores distributions. At a first glance, we see that the overlap is almost null. It is enough to conclude that a posterior matching approach cannot be executed. If we proceed to calculate the ATE for a lin-lin regression model matching observations with the propensity score anyway, we see that the effect is not statistically significant (p-value equals 0.723). We will need to use another technique for calculating the ATE.

[25] Scott A. A Czepiel. *Maximum Likelihood Estimation of Logistic Models: Theory and Implementation.*

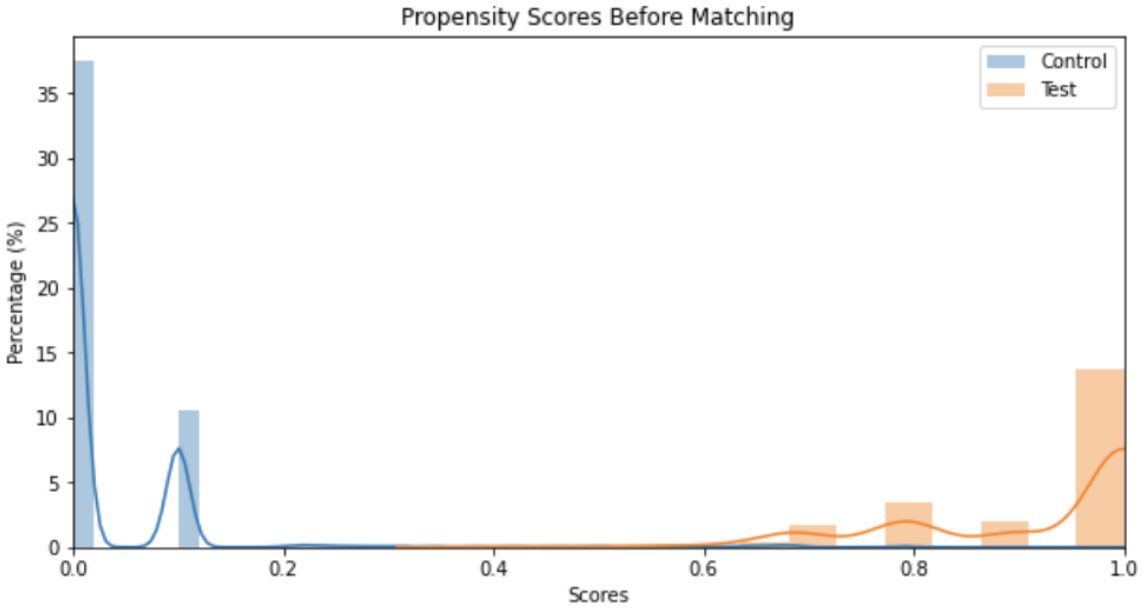

Figure 3.8: Propensity score distribution for trending and non-trending videos (Test & Control, correspondingly).

There is an interesting insight that pops out of this chart: trending videos are unique. They are significantly different from those that are not trending. This is something that we bring up again in chapter 4 when defining the addressable market. Before jumping into that problem we still need to solve this one: calculating the ATE.

3.3.2 Propensity Score Weighting

So far, we have concluded that a bare OLS regression model is flawed for our problem and the same with propensity score matching. What drives these flaws is the high level of difference between trending and non-trending videos. For the former model, this difference results in incorrect estimations and for the latter, it makes creating a sufficient number of matching observations impossible.

We can combine the propensity score matching together with the OLS regression model to reduce this problem. This can be done by performing a weighted linear regression on the data, with each point weighted by the inverse of the propensity score. The result is the propensity score weighting:

> *"Weighting by this quantity [inverse propensity score] creates a pseudopopulation in which the distributions of confounders among the exposed and unexposed are the same as the overall distribution of those confounders in the original total population. If the distributions of confounders are the same within each exposure group, then there is no longer an association between the confounders and exposure, making the exposed and unexposed exchangeable."*[26]

[26] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Sturmer, M. Alan Brookhart, and Marie Davidian, November 17, 2010. *Doubly Robust Estimation of Causal Effects*. American Journal of Epidemiology.

A weighted least squares (WLS)[27] regression consists of an ordinary least squares (OLS) regression in which each of the components of the expression the model tries to minimize is weighted:

$$arg_{\beta} min \ \sum_{i=1}^{n} w_i \ | \ y_i - \sum_{j=1}^{m} x_{ij} \beta_j \ |^2,$$

Equation 3.14: WLS cost function.

where $w_i \ > 0$ is the weight of video i. In our case, the weight introduced comes from the propensity score already calculated. It is applied to the variable we are trying to explain, $log(delta \ views)$, and its formula depends on whether the video is trending:

$$w_i = \frac{1}{PS} \ if \ X \ = 1 \ and \ w_i = \frac{1}{1 - PS} \ if \ X \ = \ 0$$

Equation 3.15: WLS weights as a function of the propensity score.

The motivation for using the propensity-score-based weight is correcting (1) omitted variable bias and (2) overt bias. If the regression is already correctly specified, weighting it will bias estimators and perform poorly. In our case we have enough evidence to conclude that an OLS regression with the variables we have cannot be correctly specified. This technique, by the way, aims to balance the distribution of both populations without getting rid of data.

As we can imagine, most weights are 1 or close to 1 (Figure 3.14). However, we can see there are some extreme weights, and those correspond to observations that (1) are not trending but look like a trending video and (2) are trending but look like a non-trending video. The regression model will try to learn more out of these outliers.

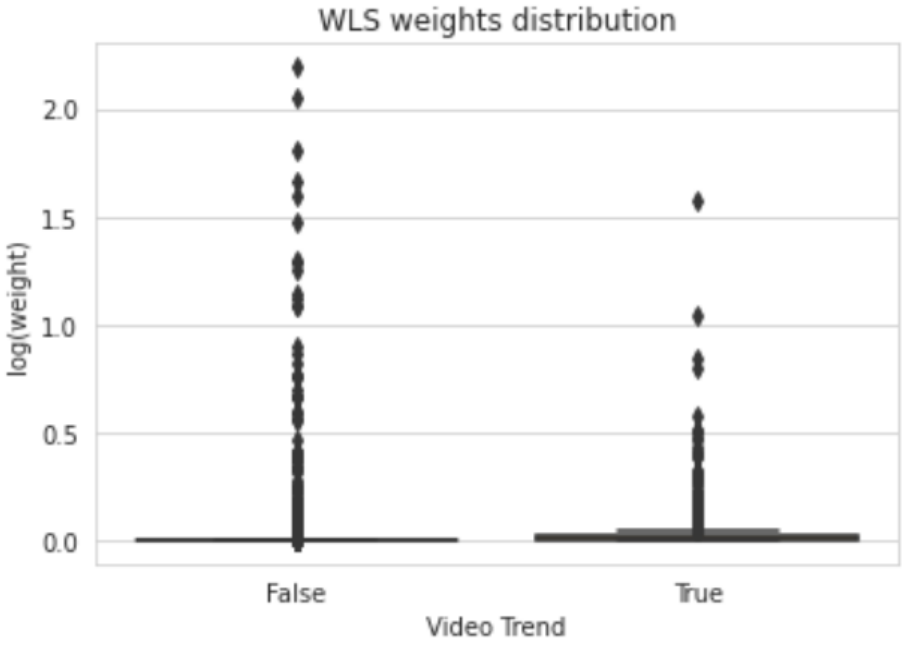

Figure 3.14: $log_{10}(w_i)$ distribution by type of video.

[27] David A. Freedman, Ricard A. Berk. November 2008. *Weighting Regressions by Propensity Scores.* Ensemble methods for Data Analysis in the Behavioral, Social and Economics Sciences.

The result of this weighted regression is shown in Figure 3.15:

WLS Regression Results

Dep. Variable:	logDeltaViews	R-squared:	0.759
Model:	WLS	Adj. R-squared:	0.759
Method:	Least Squares	F-statistic:	6350.
Date:	Sun, 02 May 2021	Prob (F-statistic):	0.00
Time:	22:23:53	Log-Likelihood:	-1.0851e+05
No. Observations:	64399	AIC:	2.171e+05
Df Residuals:	64366	BIC:	2.174e+05
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5656	0.133	4.257	0.000	0.305	0.826
videoTrend	5.5655	0.048	115.911	0.000	5.471	5.660
logViewCount	0.5812	0.002	251.567	0.000	0.577	0.586
logChannelSubscribers	-0.0071	0.002	-3.914	0.000	-0.011	-0.004
activeDays	-0.0006	5.45e-06	-103.909	0.000	-0.001	-0.001
titleLength	-0.0033	0.000	-13.078	0.000	-0.004	-0.003
descriptionLength	-1.702e-05	8.09e-06	-2.103	0.035	-3.29e-05	-1.16e-06
tagCount	0.0023	0.001	3.563	0.000	0.001	0.004
durationInSeconds	-2.472e-05	2.79e-06	-8.867	0.000	-3.02e-05	-1.93e-05
licensedContent	0.3072	0.017	17.748	0.000	0.273	0.341
embeddable	0.0001	0.046	0.003	0.998	-0.091	0.091
hasDescription	-0.2151	0.014	-15.425	0.000	-0.242	-0.188
hasTag	0.1219	0.015	8.176	0.000	0.093	0.151
likesEnabled	0.1028	0.014	7.605	0.000	0.076	0.129
dislikesEnabled	0.1028	0.014	7.605	0.000	0.076	0.129
commentsEnabled	0.1028	0.014	7.605	0.000	0.076	0.129
madeForKids	0.0491	0.024	2.045	0.041	0.002	0.096
live	2.9268	0.498	5.878	0.000	1.951	3.903
none	-3.7282	0.445	-8.384	0.000	-4.600	-2.857
upcoming	1.3671	0.315	4.341	0.000	0.750	1.984
processed	1.8615	0.361	5.157	0.000	1.154	2.569
uploaded	-1.2959	0.311	-4.164	0.000	-1.906	-0.686

Omnibus:	68733.721	Durbin-Watson:	1.654
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58097559.779
Skew:	4.534	Prob(JB):	0.00
Kurtosis:	149.865	Cond. No.	1.42e+16

Figure 3.15: Multivariable log-log weighted regression results. Some were omitted in the figure due to relevance purposes.

We see that $videoTrend$ estimator is statistically significant and has a magnitude of 5.5655. This is 10% smaller than the one we get without weighting (6.2224) and means that on average, being trending increases weekly $log(views)$ by ~5.6 units.

3.3.3 Doubly Robust Weighted Estimator

This inference technique has the same spirit as the propensity score weighting. It uses regression modeling and the inverse propensity score as weights and it actually solves the same flaws the other model does. We decided to calculate the ATE with this model to (1) try a novel and sophisticated inference technique and (2) add more data points to make a robust decision with picking the ATE we use downstream.

To calculate the doubly robust weighted estimator we start by estimating the parameter of two regression models: one for the treatment group and another one for the control group:

$$log(delta\ views_{Ti}) = \beta_0 + \sum_{j>1}^{stats} \beta_j log(X_{Tji}) + \sum_{j>stats}^{covariates} \beta_j X_{Tji} + \varepsilon$$

$$log(delta\ views_{Ci}) = \beta_0 + \sum_{j>1}^{stats} \beta_j log(X_{Cji}) + \sum_{j>stats}^{covariates} \beta_j X_{Cji} + \varepsilon$$

Equation 3.13: Multivariable regression models for treated and untreated observations taking a multiplicative approach.

We call these two predictive response functions as $\hat{Y}_1$and $\hat{Y}_0$ respectively. Also for simplicity, we call PS to the propensity score and Z to the covariates. Figure 3.9 shows how to calculate the treatment effect on individual observations. Note that the "Exposed" and "Unexposed" are analogue to treatment and control.

Table 1. Equations for the Expected Response Under Exposed (DR_1) and Unexposed (DR_0) Conditions for Each Individual in the Population[a]

	DR_1	DR_0
General form	$\frac{Y_{X=1}\times X}{PS} - \frac{\hat{Y}_1\left(X-PS\right)}{PS}$	$\frac{Y_{x=0}(1-X)}{1-PS} + \frac{\hat{Y}_0\left(X-PS\right)}{1-PS}$
Among $X=1$	$\frac{Y_{X=1}}{PS} - \frac{\hat{Y}_1\left(1-PS\right)}{PS}$	$\hat{Y}_0$
Among $X=0$	$\hat{Y}_1$	$\frac{Y_{x=0}}{1-PS} - \frac{\hat{Y}_0\times PS}{1-PS}$

Abbreviations: DR, doubly robust; PS, propensity score.
[a] $PS = p(X=1|\mathbf{Z})$; X = exposure; $Y_{X=0}$ and $Y_{X=1}$ = observed outcome among individuals with $X=0$ and $X=1$, respectively; $\hat{Y}_0 = E(Y|X=0,\mathbf{Z})$ = predicted outcome given $X=0$; $\hat{Y}_1 = E(Y|X=1,\mathbf{Z})$ = predicted outcome given $X=1$.

Figure 3.9: Individual treatment effect calculation. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology. November 17, 2010

To calculate the ATE we compute DR_0 and DR_1 for every video using the formula of Figure 3.9. Then we take the average for each of those variables; the difference between these averages constitutes the ATE. Based on all the evidence gathered so far, we decide to use a log-log model: (1) $Y, \hat{Y} = log(delta\ views)$ and (2) $Z_{stats} = log(stats)$. Result of this model is presented in figure 3.10.

```
Treatment Effect Estimates: Weighting

                Est.       S.e.          z      P>|z|      [95% Conf. int.]
--------------------------------------------------------------------------------
         ATE    1.495      0.558      2.678      0.007      0.401      2.588
```

Figure 3.10: ATE using doubly robust estimation.[28]

[28] https://causalinferenceinpython.org/ python library used to calculate ATE.

This result is statistically significant and a priori suggests that the decision of not using the ATE calculated via the OLS regression was correct: the ATE obtained through this more robust model is almost half the value than the one obtained with the bare regression, which is good since it suggests that is removing omitted variable and overt bias that tend to inflate estimations.

Far from getting comfortable with this result, we proceed to analyze the distribution of DR_0 and $DR_1$for trending and non-trending videos. We should be comfortable using the average with this approach if there's sufficient overlap between them.

Figures 3.11 and 3.12 show how different DR_0 and DR_1distributions are when breaking them out by type of video. If we calculate $DR_1 - DR_0$ for every video and plot its distribution we can get a sense of Trending effect for each of them (Figure 3.13). This model suggests that the effect is negative for non-trending videos and considerably positive for trending videos. Such a conclusion goes against our modeling hypothesis, that states that the effect of this feature should be the same amongst all videos.

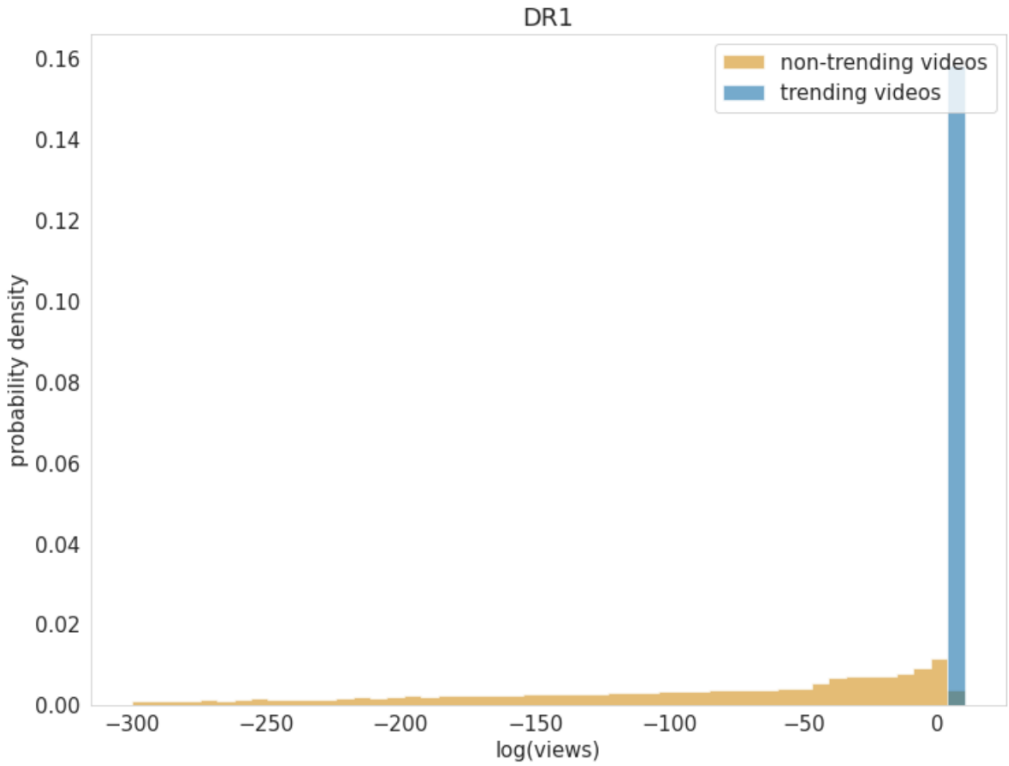

Figure 3.11: Expected response (delta views) distribution when treated.

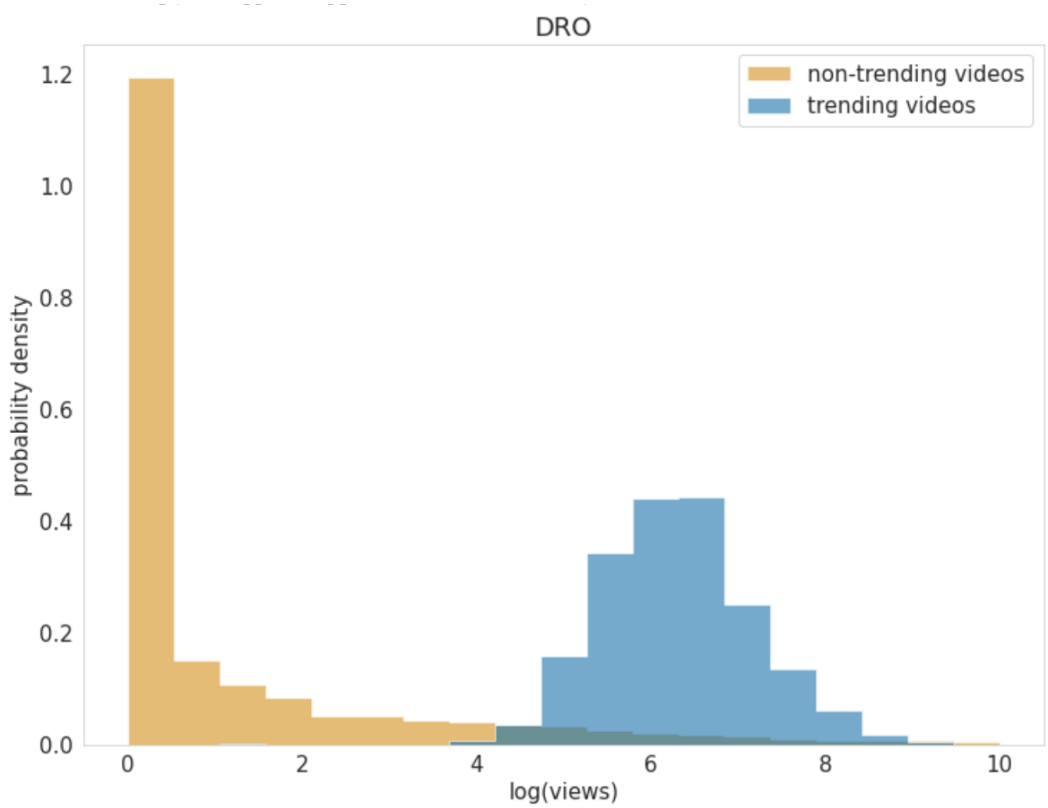

Figure 3.12: Expected response (delta views) distribution when untreated.

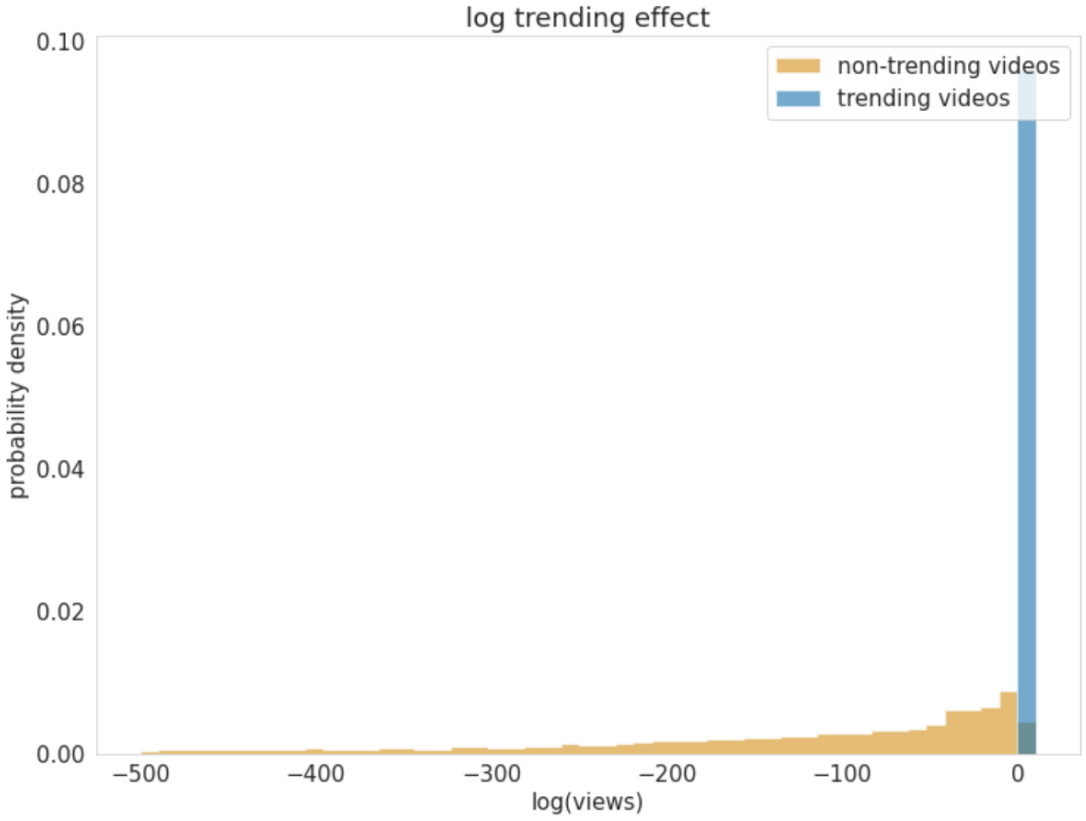

Figure 3.13: Estimated treatment effect (delta views) distribution.

Due to how different trending and non-trending videos are we knew beforehand that the estimated treatment effect using this technique could differ between trending and non-trending observations, but even when comparing non-trending videos that are very similar to trending ones we observe that the effect is radically different between them. To determine similarity between videos we used a machine learning classifier which is explained in chapter 4.

The main driver of this flaw is the huge difference in active days between trending and non-trending videos. Most trending videos have around 2 days since published, whereas non-trending videos have a somewhat homogeneous distribution with a massive range (up to thousands of days). Regression coefficients for active days of $\hat{Y}_1$ and $\hat{Y}_0$ are -0.1425 and -0.0005 respectively. This makes $\hat{Y}_1$decrease its value rapidly as active days increase when the video is not trending, resulting on very low DR_1. Moreover, $DR_1$for non-trending videos is negative and with a high absolute value because of this reason.

3.4 ATE interpretation, usage and hypothesis validation

3.3.1 Interpreting and using ATE

Figure 3.14 summarizes the ATE obtained with each inference technique together with its confidence intervals. Observe how big the confidence interval is for the doubly robust model compared to the regressions.

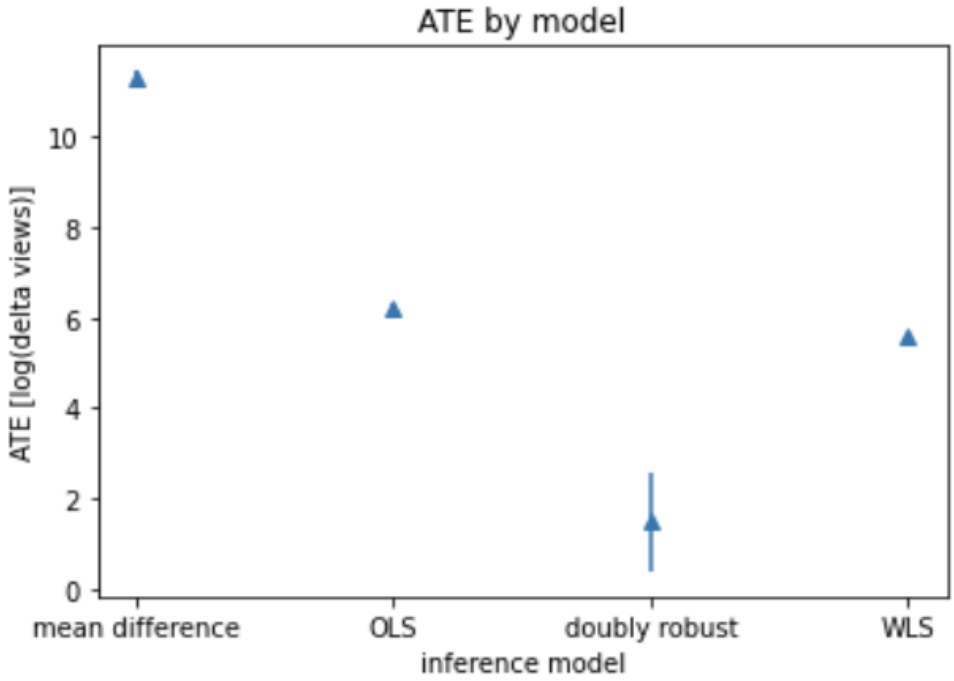

Figure 3.14: ATE confidence intervals by inference model, using a confidence level of 95%.

We decide to use the ATE calculated through the WLS regression to compute the optimal price for YouTube Trending. We proceed to explain how to interpret this value and how to use it.

Due to the fact we applied logarithms to *delta views* (natural logarithms), the ATE we have corresponds to a multiplicative estimation model. Our original formulation yields that on average:

$$ln(delta\ views_{T_i}) = ln(delta\ views_{C_i}) + ATE$$

Doing some mathematical manipulation we get:

$$delta\ views_{T_i} = delta\ views_{C_i} * e^{ATE}$$

When building the counterfactual scenarios we need to apply the ATE for both trending and non-trending videos. Whether the expression e^{ATE} multiplies or divides the real $delta\ views_i$ will depend on whether or not the video is trending. Based on this, we build the counterfactual delta views function, which retrieves the delta views a video of our dataset would have if it the treatment was applied (or not applied, in case it is a trending video):

$$counterfactual\ delta\ views\ (X) = \begin{cases} delta\ views * e^{ATE} & if\ X = 0 \\ \dfrac{delta\ views}{e^{ATE}} & if\ X = 1 \end{cases}$$

Equation 3.17: Counterfactual delta views formula.

where $X = 0$ if the video is not trending, $X = 1$ if it is and $ATE = 5.5655$. We compute the counterfactual delta views for each of the videos on our dataset and calculate the difference with the original delta views: the result is the effect on views Trending has (Figure 3.5). We do not take into account videos whose original delta views are equal to zero when plotting the trending effect, given that anything multiplied by zero will continue being zero. This set of videos represent 36.5% of the dataset we used in the thesis. We can interpret this in the following way: there are videos that no matter whether they are featured in trending they will not get any views from one week to the other; this should be the case for very old videos (so, in essence, they are not trending) or very bad ones. We acknowledge that it is a very strong statement to say that they will get zero views, but for our modeling purpose it is fair to assume so.

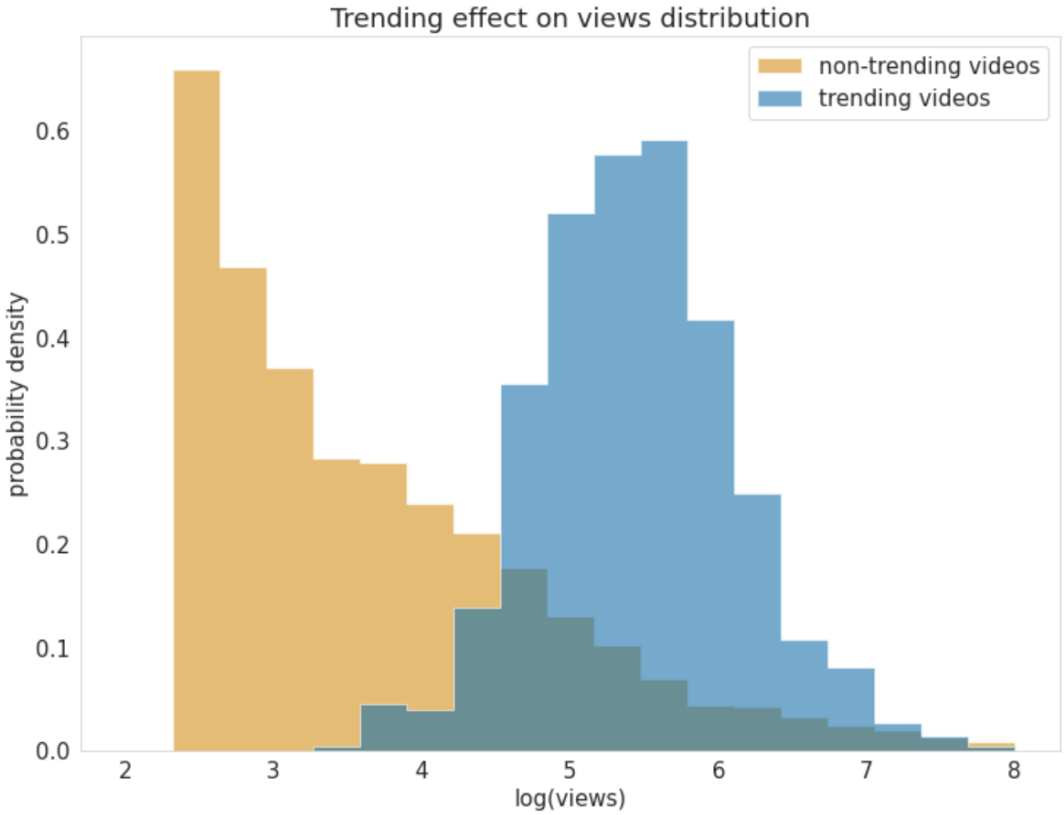

Figure 3.15: Estimated treatment effect distribution. Videos with delta views = 0 are not included.

3.3.2 Hypothesis validation

So far, we have gathered statistically significant evidence to conclude that being trending at the a given day increases the the number of *individual* views generated throughout the following week.

For us to validate the main hypothesis of this thesis, *"trending feature increases the total number of views of YouTube",* we calculate the incremental number of views YouTube would generate in a given week without the existence of trending and later compare it with the current scenario:

$$delta\ views_{no\ Trending} = \sum_{i}^{non-trending} delta\ views_i + \sum_{j}^{trending} \frac{delta\ views_j}{e^{5.5655}}$$

This expression can also be written as:

$$delta\ views_{no\ Trending} = \sum_{i}^{non-trending} delta\ views_i + \frac{1}{e^{5.5655}} \sum_{j}^{trending} delta\ views_j$$

Equation 3.18: Delta views if Trending does not exist.

The expression for the current situation, on the other hand, can be written as:

$$delta\ views_{Trending} = \sum_{i}^{non-trending} delta\ views_i + \sum_{j}^{trending} delta\ views_j$$

Equation 3.19: Breakdown of current delta views.

With these 2 expressions we prove that $delta\ views_{Trending} > delta\ views_{no\ Trending}$:

$$\sum_{i}^{non-trending} delta\ views_i + \sum_{j}^{trending} delta\ views_j > \sum_{i}^{non-trending} delta\ views_i + \frac{1}{e^{5.5655}} \sum_{j}^{trending} delta\ views_j$$

$$\sum_{j}^{trending} delta\ views_j > + \frac{1}{e^{5.5655}} \sum_{j}^{trending} delta\ views_j$$

$$e^{5.5655} > 1$$

Equation 3.20: Mathematical demonstration of Trending's positive effect on YouTube's video views.

It is important to mention that the aggregation used to build the counterfactual scenario is true only if we assume that the trending effect over individual videos has no effect over the delta views of the remaining videos of the platform. This is also a base assumption for any of the inference models we have used in this work. This assumption can also be reframed as the funnel we introduced early in this work—which is the framework of this thesis (Equation 1.5)—if adding that increasing the discoverability of a video (which is what Trending does) does not affect either the discoverability probability nor the conditional probability expression that follows for other videos. In other words, Trending does not cannibalize views from other videos.

This strong assumption can be intuitively justified. There are three types of YouTube users (excluding creators): (1) we have those users who entered YouTube knowing what they want to see, (2) those who entered without knowing, (3) and those who were redirected from an external URL to a specific video or came across an embedded YouTube video outside the platform. Trending feature does not interfere with the user journey of persona (1) and (3): persona (1) goes to the search bar and looks for the video he or she wants to see while persona (3) lands directly to that video. Trending feature never interfere in this process. However, it is possible that after watching (or not) the videos they looked for (or landed to) they might stick around in the platform and be impacted by Trending's effect. This is not considered cannibalization. The user journey of persona (2), on the contrary, will probably be affected by Trending feature since he or she is wandering across the platform looking for a video that might be interesting to watch, but there is no explicit reason to imply Trending would cannibalize video views from the homepage or category navigation. We are not saying it is not possible, but since there are no strong reasons to imply so, we will neglect it for simplifying the modeling of this thesis.

3.5 Inference model validation

We finish this inference section validating the assumptions of the WLS model. In order to study that a regression is suitably defined for a dataset, it is usual to analyze how the residuals look like. Remember that the main assumptions of a linear regression model are: (1) linearity: the relationship between covariates and the mean of the response is linear; (2) homoscedasticity: the variance of residuals is the same for any value of the covariates; (3) independence: observations are independent of each other; (4) normality: for any fixed value of the covariates, the response is normally distributed.

We have addressed and checked assumptions (1) and (3) across this chapter and the descriptive analysis section. We can validate the remaining assumptions by validating another set of linear regression assumptions related to residuals. These are: (1) normality assumption: errors are normally distributed; (2) Zero mean assumption: they are normally distributed around zero; (3) homoscedasticity (already mentioned above); (4) independent error assumption: there is no correlation between the residuals and the predicted values, or among the residuals themselves.

$$e_i = y_i - \hat{y}_i$$

Equation 3.16: Regression residual.

We want to check that e_i follows a normal distribution with $E(e_i) = 0$ and that there is no correlation between $\hat{y}_i$ and e_i. We study Figure 3.16 and 3.17 to validate that. In the first one we observe that the distribution of residuals is symmetric, centered in 0 and has a normal distribution shape.

The second chart shows that e_i is centered in 0 across the different values of $\hat{y}_i$. We also observe some irregularities in this chart that are worth mentioning even though we decided not to alter our decision of using the ATE calculated through this model because of them. On the one hand, we observe two clouds of values: the one at the right corresponds to trending videos who have the highest predicted $\hat{y}_i$. Residuals for these look healthy. For the cloud at the left we observe some odd correlation at the extremes, i.e. for videos with very low $\hat{y}_i$ and very high $\hat{y}_i$; the vast majority of observations, however, have their e_i averaging 0 and with an homogeneous distribution concentrated in that value. It is very likely that this odd behavior at the extremes is what drives the low but statistically significant correlation between the residual and the prediction (-0.09 Pearson correlation with a p-value lower than 0.05).

The difference in variance between the dotted clouds and the odd correlations in the extremes of the left cloud violate the homoscedasticity assumption. This does not bias the estimation of the coefficient anyways, so it is not dangerous to use the ATE calculated downstream; it affects the p-value instead. In a future work we would try to correct this modeling flaw but for the time being we are comfortable with the transformations done and weighting used.

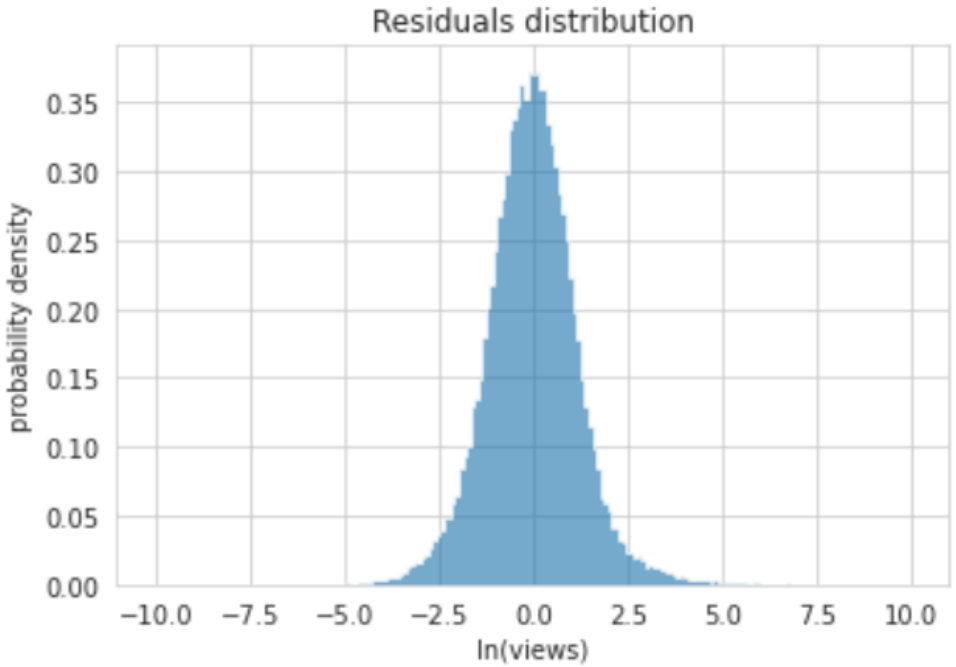

Figure 3.16: Distribution of e_i for WLS regression.

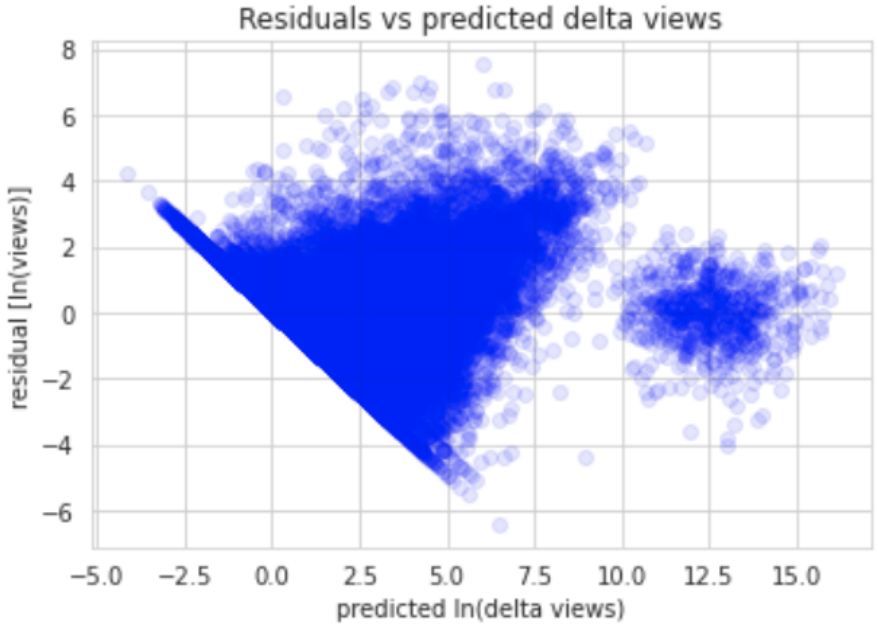

Figure 3.17: Correlation between e_i and $\hat{y}_i$ for WLS regression.

4. Addressable market definition

We have our formula to calculate delta views in counterfactual scenarios, i.e. calculate how many views YouTube will gain or lose in a week if a video j was not trending (in case it originally was) or if a video i was trending (in case it originally was not).

However, we cannot apply this formula to all YouTube videos as YouTube cannot offer the possibility of being featured in Trending to any video. As we have seen in the data, YouTube is very demanding at the moment of picking trending videos and this has a reason to be: the conditional probability $P(user\ plays\ video\ /\ user\ discovers\ video)$ has to be *high* for Trending videos. If Trending section starts showing uninteresting videos it is very likely that the ATE will decrease in the mid to long term.

This forces us to limit the scope of the counterfactual analysis to a particular subset of videos. This subset of videos from our dataset includes those that are currently trending at time T and those that look like trending videos, i.e. videos that YouTube could have chosen as trending based on their characteristics.

To define this we use a machine learning classification model. Those videos whose probability of being trending is higher than a certain threshold are considered as part of our addressable market. For sure, the threshold used is high (0.95). On top of that we do a further evaluation of how much the distributions overlap: YouTube is really strict when selecting trending so we have to do the same when defining the addressable market.

For instance, the available number of trending slots of our sample is 1,000 but only ~70% of them are being occupied. What we can conclude is that not putting videos whose conditional probability of being watched once found are really high could harm the positive effect Trending has.

4.1 Logistic Regression with L2 regularization

The model we use to classify videos is the same used for calculating the propensity score earlier but with a modification. Given we are building a predictive model, we need to make sure that it does not overfit to the data. Overfitting[29] is fitting the parameters and hyperparameters of your predictive model excessively to the training data. As a consequence, when new data is exposed, the results of the model will have an important level of error compared to the one had over the training set, as it learned to be extremely accurate in very specific situations, learning from the noise instead of capturing the signal.

The parameters of the logistic regression are the β_i of the linear expression that the model learns through the Maximum Likelihood Estimation method. The more variables you introduce to the model (from now on, features), the more β_i the model has to learn and the more likely these coefficients will be fine-tuned to replicate the exact responses each observation has in the training set. This is the overfitting risk this model has.

[29] Xue Ying, February 2019. *An Overview of Overfitting and its Solutions.* Journal of Physics Conference Series.

In order to define the variables we need to keep in order to avoid overfitting, we use a regularization technique.[30] Regularization techniques are ways of penalizing the solutions that the model finds by their complexity. If we punish solutions by how complex they are, we end up favoring simpler ones that are likely to generalize better without throwing away any of our variables before fitting the model.

Logistic regression works by finding the set of parameters β_i that minimize a loss function. Of course, $loss\ function = f(\beta_i)$. When using L2 regularization, we add an extra component to the cost function that penalizes the magnitude of β_i. As a consequence, the higher they are, the higher the cost function. So the model, when minimizing this cost function, has to deal with the tradeoff of its two components.

As a result, the β_i of some variables will be comparably smaller than those obtained if the regularization was not made. The logic here is that the larger the coefficients, the more complex the model is. This prevents overfitting.

When building predictive models it is a good practice to fold a random subset of observations of the training set for testing purposes. This testing set is later used to check how good your model fits to unseen data. In our case we build a testing set with 20% of the total number of videos of our dataset.

We train the model with the training set and build the ROC curve (receiving operating characteristics curve), shown in Figure 4.1. Here you can also see the AUC (area under the ROC curve), which is ~1.00 (0.998). AUC is a metric that ranges between 0 and 1. The farther AUC is from 0.5 the better predictor the model is, with 1 being a perfect predictor and 0 being a perfect anti-predictor. It should not be a surprise that the model performs so well since we are dealing with such an easy classification problem: it is pretty clear from the descriptive analysis that the distribution of some of the variables are very different between trending and non-trending videos. This helps the model to learn how to distinguish them better.[31]

[30] Tulrose Deori, July 2020. *Implement Logistic Regression with L2 Regularization from scratch in Python.* Towards data science.

[31] Hossin, M. and Sulaiman, M.N., March 2015. *A Review On Evaluation Metrics For Data Classification Evaluations.* International Journal of Data Mining & Knowledge Management Process (IJDKP). Vol. 5, No.2.

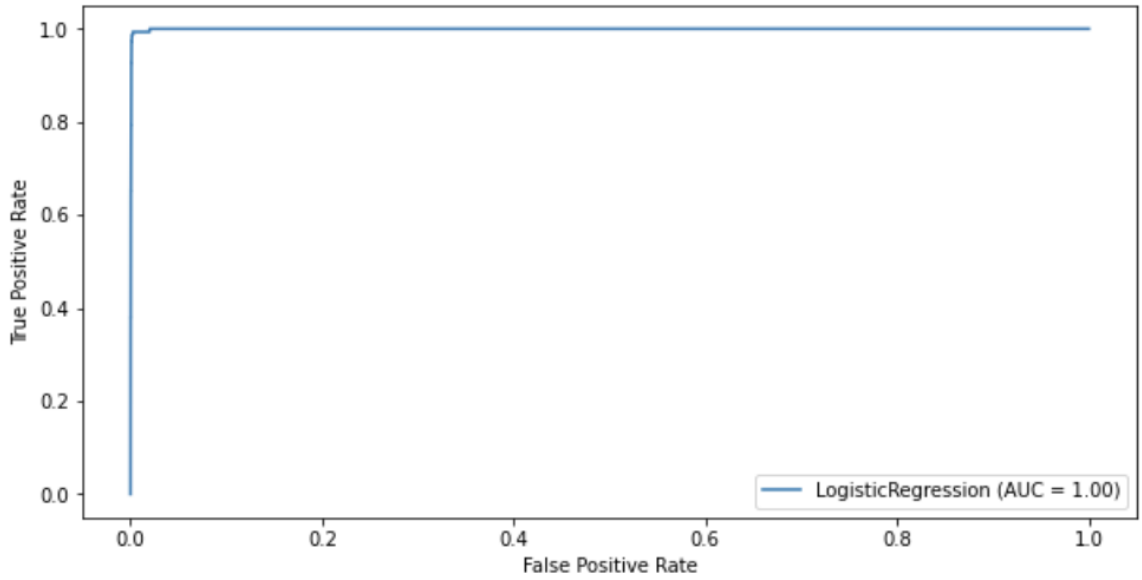

Figure 4.1: Logistic regression + L2 regularization model ROC.

Before moving forward we clarify that by using L2 regularization there is a hyperparameter introduced into the cost function which we decided not to optimize. The best way of learning the optimal hyperparameter is through k-fold cross validation. Given the model obtained is *almost* perfect performance without hyperparameter optimization—in terms of AUC—we decide not to spend time on it.

The output of this classification model is a score between 0 and 1 for each video: the higher the score, the more likely it corresponds to a trending video. We want to consider non-trending videos as part of the addressable market only if they are similar enough to trending videos, and for us that means that their classification score is higher than 0.95. This threshold was defined in an iterative process in which we looked for one that assured a high level of overlap between false positives and true positives distribution of covariates (the higher, the better) and also a high recall.

Formally speaking, the addressable market used for the pricing model is composed of: true positives, false negatives and false positives. Figure 4.2 shows how these are distributed across our dataset.

		Prediction outcome	
		Non-trending	*Trending*
Ground truth	*Non-trending*	98.7%	0%
	Trending	0.2%	1.1%

Figure 4.2: Logistic regression + L2 regularization model confusion matrix.

We can see that this model has an accuracy of 99.8%, a recall of 100% and a precision of 84.6%. This last result suggests there are a few non-trending videos that are actually *very* similar to trending videos. As a reminder:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

Equation 4.1: Classification model performance metrics.

where *T* stands for *true* and *F* for *false*, *P* for *positive* and *N* for *negative*.

If we compare the distribution of active days and video views between FP and TP we see that they are comparably similar. All FP videos have less than 25 days being active on the platform, which is a huge close to one of the main characteristics of trending videos. The distribution of video views of FP has the same shape of trending videos, but it is a little shifted to the left (Figure 4.3).

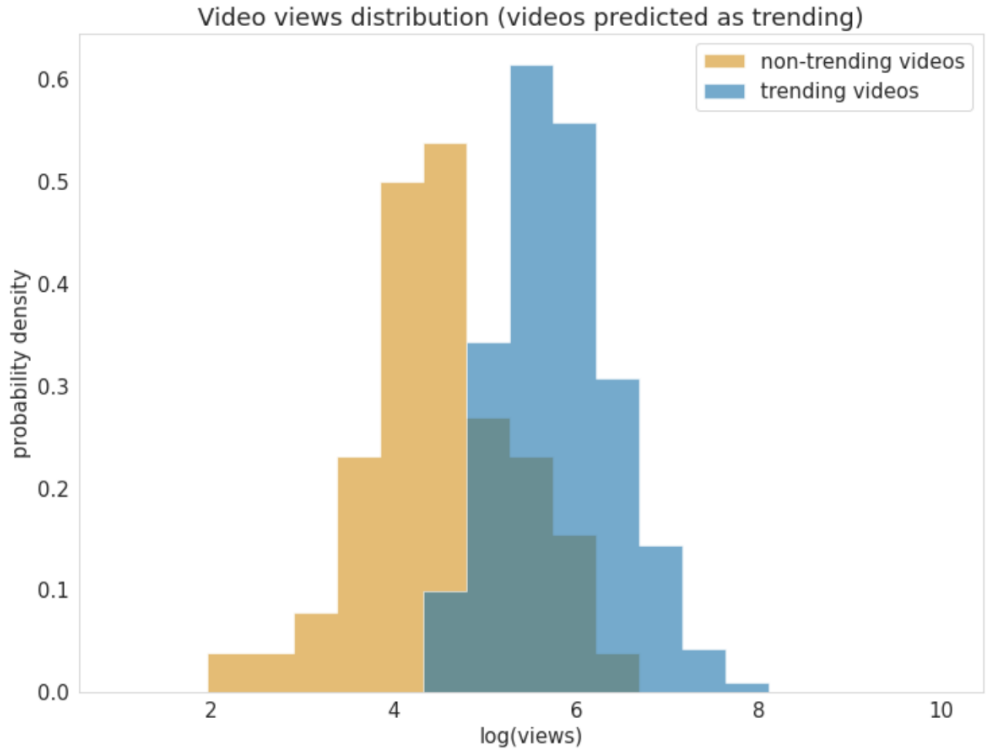

Figure 4.3: $log_{10}(views)$ distribution by type of video for videos predicted as trending.

If we check Figure 4.4, we observe that the distribution of FP's active days is way closer to the one of trending videos when comparing it with the distribution of this variable when looking at the whole population of non-trending videos (Figure 2.12).

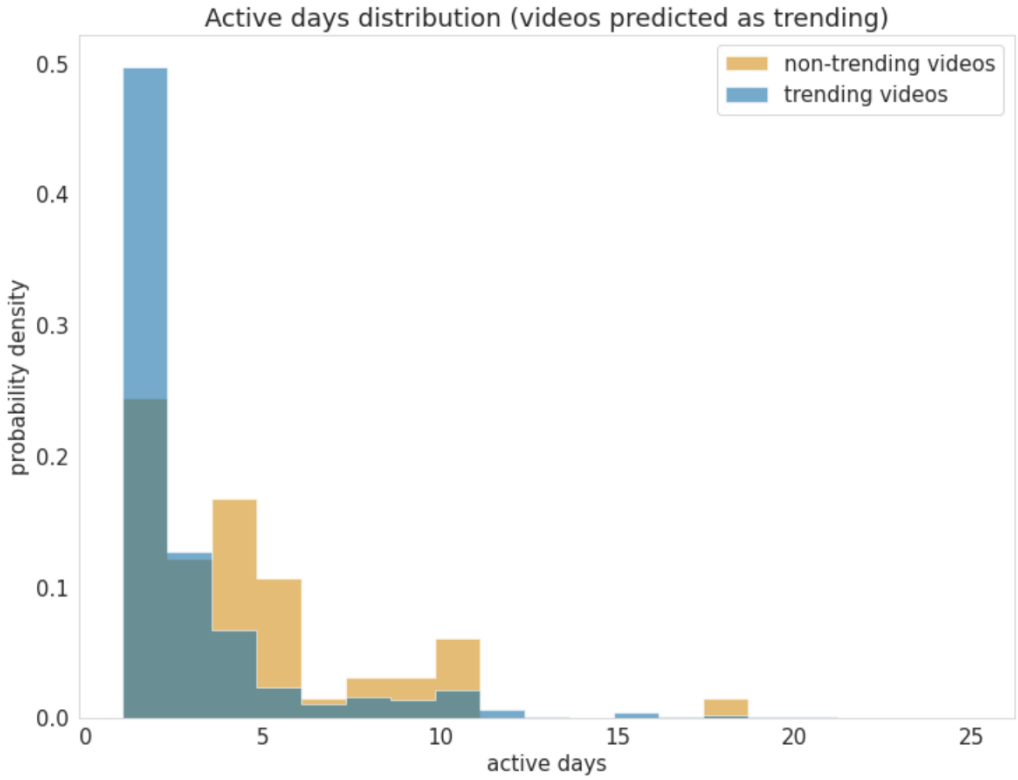

Figure 4.4: Active days distribution by type of video for videos predicted as trending

Even though the classification model did a very good job identifying non-trending videos that are very similar to trending ones, we opt to take a conservative call and not include them in our addressable market. Key variable distributions look pretty similar and the overlap between variables is lower than before (Figure 4.5) but there is still a significant difference between them and we cannot conclude they belong to the same population. The addressable market we will use for the rest of the analysis will be limited to real trending videos. Worst case scenario, the price we end up coming up with is lower than the *real* optimal one, but probably nothing that could not be addressed with pricing techniques once the monetization feature is live.

The fact that YouTube keeps empty slots also supports the decision of sticking to the actual trending videos when defining the addressable market. Having access to YouTube's Trending algorithm would solve these estimations, but we know that is not possible. Having a richer dataset could also help to come up with a model that generates false positives more similar to actual trendings. Something we have not explored but has a good fit with this particular problem is the use of unsupervised machine learning models to cluster videos into groups; one can expect that trending videos would be categorized as part of the same cluster, along with a few non-trending videos, defining a potential addressable market.

Variable	Controls (N_c=55) Mean	S.d.	Treated (N_t=709) Mean	S.d.	Raw-diff
Y	62287.018	153066.223	1054112.530	3366895.126	991825.512

Variable	Controls (N_c=55) Mean	S.d.	Treated (N_t=709) Mean	S.d.	Nor-diff
X0	158124.309	448105.374	2049342.746	6594413.633	0.405
X1	4.109	3.414	2.869	2.864	-0.394
X2	265.473	542.944	3350.494	14987.628	0.291
X3	6982.000	13533.651	97294.360	262307.422	0.486
X4	1085303.509	1808653.700	4154515.989	11522456.381	0.372
X5	277200262.673	690657458.851	1757410263.274	8425736869.644	0.248
X6	2955.873	10153.985	4036.882	22131.273	0.063
X7	688.982	842.873	1303.226	2762.291	0.301
X8	68.255	23.620	54.784	23.593	-0.571
X9	786.382	777.926	937.017	844.823	0.185
X10	17.418	17.074	16.945	14.097	-0.030
X11	0.058	0.039	0.058	0.045	0.008
X12	0.003	0.004	0.001	0.002	-0.520
X13	0.003	0.004	0.001	0.002	-0.520
X14	48.287	222.318	4.448	9.882	-0.279
X15	0.344	1.321	0.011	0.058	-0.356

Figure 4.5: Overlap between false positives and trendings.

5. Pricing

5.1 Willingness to pay

Assuming creators act on a rational way and that there is no information asymmetry between YouTube and them (i.e. both know what the effect of Trending is), the owner of a video selected as trending by the platform should be willing to pay for that functionality if the value it retributes is higher than its price. These are strong assumptions but are necessary for determining a ballpark for the optimal price.

For a given video, we define the the number of views trending could provide as:

$$Video\ views\ gained = delta\ views_{Trending} - delta\ views_{non\ Trending}$$

Equation 4.2: Theoretical formula to obtain video views generated by Trending.

For the videos of our dataset, we can obtain this value by applying the counterfactual delta views function defined in Equation 3.17. To compute the number of video views gained we grab each video and use this expression:

$$Video\ views\ gained = |\ delta\ views_{Real} - counterfactual\ delta\ views\ (x)\ |$$

Equation 4.3: Video views gained as a function of factual delta views and the counterfactual delta views function.

Now that we have the number of views Trending adds in a week to each of the videos in our dataset, we transform it into the number of ad views by applying its conversion rate (22%):

$$Ad\ views\ gained = |\ delta\ views_{Real} - counterfactual\ delta\ views\ (x)\ | * 22\%$$

Equation 4.4: Ad views gained as a function of factual delta views and the counterfactual delta views function.

To transform these ad views gained to the revenue gained, we use the CPM and the revenue share creators get:

$$Ad\ revenue\ gained = |\ delta\ views_{Real} - counterfactual\ delta\ views\ (x)\ | * 22\% * \frac{CPM}{1000} * 55\%$$

Equation 4.5: Ad revenue gained as a function of factual delta views and the counterfactual delta views function.

To conclude, we state that creators will be willing to pay for a trending position if the price for it is lower than the ad revenue gained, which follows this expression:

$$Price < |\ delta\ views_{Real} - counterfactual\ delta\ views\ (x)\ | * 22\% * \frac{CPM}{1000} * 55\%$$

Equation 4.6: Creators condition to convert to Trending paying-user based on the counterfactual delta views function.

We use this mathematical entity to check for every video in our addressable market what is the maximum price their creators will be willing to pay for Trending. We can see the

distribution of this variable in Figure 5.1. If we look at the frequency instead of the probability density, we get the inverse demand curve (Figure 5.2).

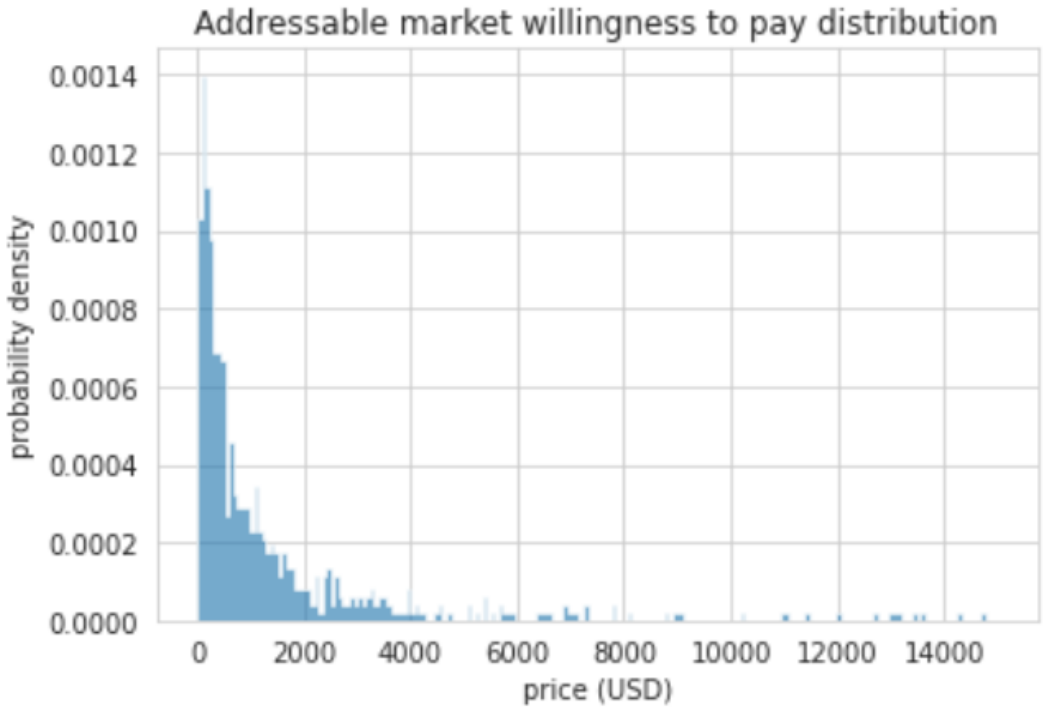

Figure 5.1: Willingness to pay distribution of the addressable market.

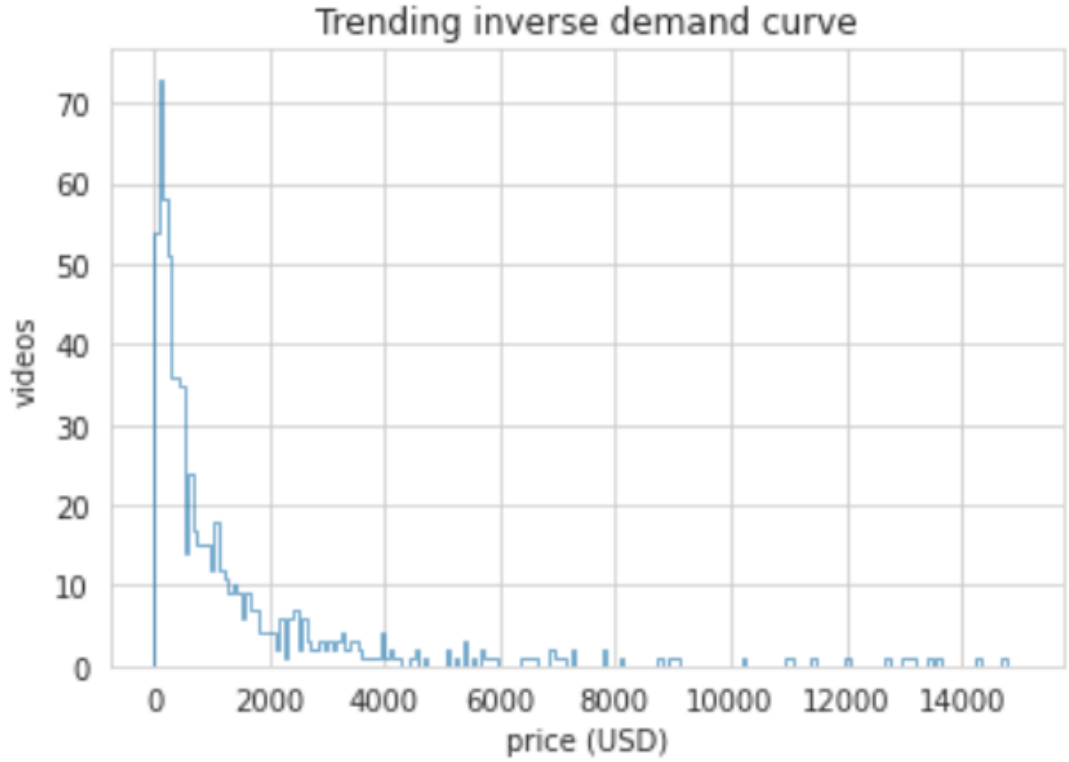

Figure 5.2: Addressable market inverse demand curve.

5.2 Optimal price calculation

Let us bring back again the optimization problem we are trying to solve:

$$Maximize\ \{PUTV\ x\ Price\ -Ad\ views\ lost\ *\ \frac{CPM}{1000} *\ 45\%\}$$

Modelling ad views lost as:

$$Ad\ views\ lost\ =\ \sum^{videos} f(Price, creator)$$

$$Where\ f(Price, creator) =\ Trending\ effect\ on\ ad\ views, if$$

$$Price \geq\ Ad\ views\ gained\ *\ \frac{CPM}{1000} *\ 55\%$$

$$Else, f(Price, creator) =\ 0$$

For both cases:

$$Trend\ effect\ on\ ad\ views\ =Ad\ views\ gained$$

where $PUTV$ is paying-user trending videos. At this point, we have all the information needed to calculate the maximum net revenue.

Since the size of the mathematical problem is not huge, we will calculate the net revenue by trying all the distinct willingness to pay previously calculated as potencial prices. Figure 5.3 shows the estimated average weekly net revenue generated by Trending monetization when trying these different prices. The price YouTube should charge for Trending would be the one that generates the highest net revenue.

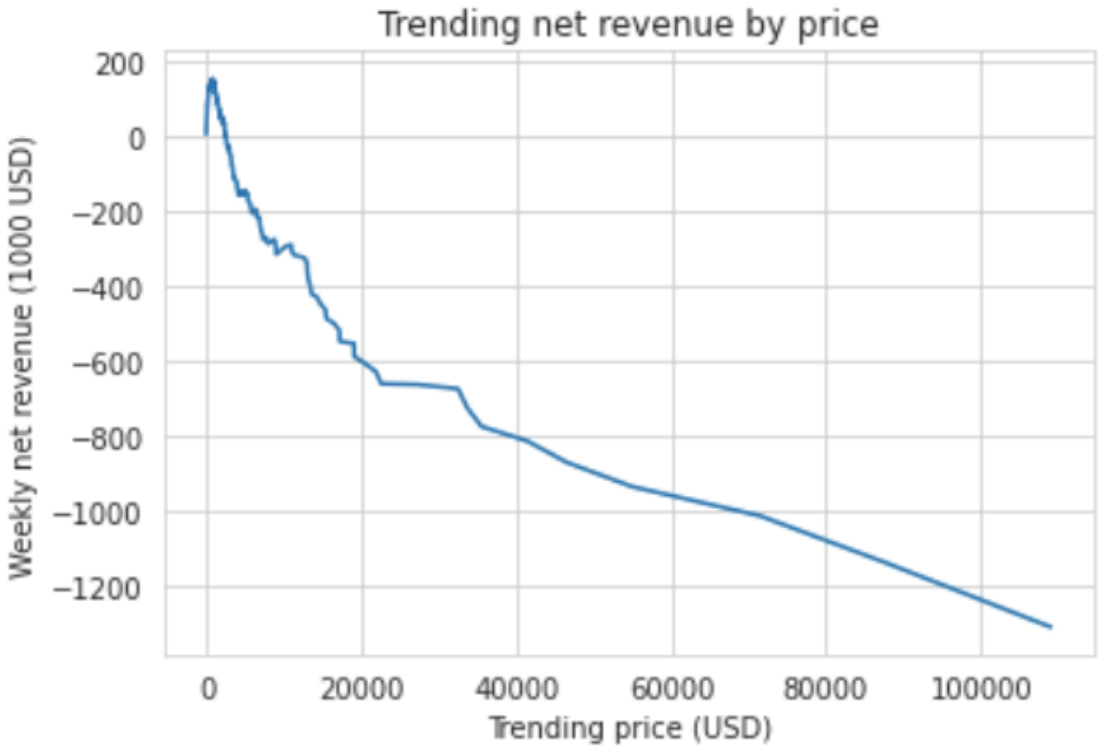

Figure 5.3: Estimated weekly net revenue for Trending as a function of its price.

The price that maximizes weekly revenue is US$856, generating an average incremental revenue of US$154,500 per week (US$8 million per year, 0.04% of YouTube's yearly revenue reported in 2020) driven by the trending videos of the 20 countries studied. It is worth noticing how elastic the demand curve is and how it impacts net revenue: there are a good number of videos (~40%) that would pay more than the optimal price. However, putting a price higher than the optimal one can lead to revenue losses driven by the high opportunity cost advertisement revenue represents. From a price of US$2560 onward (3 times the optimal price) net revenue starts being negative. There will still be some videos that can make profit with such a high price, but the majority would not, so they would not pay for the feature and the incremental ad revenue of those would be lost.

With this optimal price, only 415 trending slots out of 1,000 would be occupied (this means there would be a 41.5% occupation rate). There are two thoughts that come out from this insight: (1) the reduced set of trending videos may boost Trending effect more than what we estimated. This is because users would browse through a feed with less content, increasing the discoverability of the videos that made it to that feed; (2) there is a wasted opportunity to capture value when thinking of the unoccupied ~60%. With a single price strategy, there is not too much to be done; by discriminating by price we can definitely make the most out of Trending section. We analyze this concept in the following section.

5.3 Price fine tuning

By observing this analysis, YouTube should conclude that: (1) there is an opportunity to capture part of the revenue being delivered to trending videos through Trending feature; (2) monetizing this feature represents a high risk in terms of ad revenue opportunity cost.

US$856 is an average price that comes out from statistical models that have statistical errors. YouTube should not consider this *the* price to launch the monetized version of the feature. It should be used to understand the level of magnitude of the launch price and the business opportunity behind Trending monetization as well as a key input for further pricing research.

Pricing research can be divided into two stages: (1) offline research and (2) online research. The former consists of applying user research techniques (such as focus groups, surveys, interviews, etc.) to understand their *real* willingness to pay for this feature. Here YouTube should focus on understanding the risk information asymmetry brings. Everyone should be willing to pay a certain amount of money if they will be getting more in exchange; the problem might be that creators (1) will not know how many incremental views they will make at the moment of being selected as trending by YouTube and (2) will not be able to identify how many views were driven by Trending after being trending for a week. When monetizing the feature YouTube should bring this information upfront somehow.

During this first research phase YouTube should also try to size what percentage of the incremental ad revenue generated by trending creators would be willing to leave to YouTube, and how. It is not the same to go through a payment funnel once your video has been selected as trending than just letting YouTube discount Trending price from the advertisement revenue that corresponds to that video a week after. The former experience

adds cognitive load to creators, a psychological barrier to purchase such functionality and the risk of not getting the return YouTube might promise. The latter experience, on the contrary, is seamless, user friendly for creators and mitigates performance risk, since it allows charging by real performance.

Charging for Trending by discounting a percentage of ad revenue transferred to creators brings up the possibility of price discrimination. Such a pricing approach removes the need of having one single price for all videos, but instead looks to charge every particular case based on what the creator is willing to pay. Weekly net revenue for the addressable market studied can be boosted up to US$1,670,000 if Trending could be monetized through a first-degree price discrimination scheme like the one suggested (this is a yearly incremental revenue of US$87 million, 0.4% of YouTube's yearly revenue reported in 2020). This means almost a 11 times in rease in net revenue compared to the scenario in which no price discrimination is feasible. With this approach, however, the price is no longer a static fee but a percentage of incremental ad revenue: this 11 times increase would be possible if (1) the whole addressable market is willing to pay for Trending and (2) the percentual fee is 100% (i.e. all the incremental ad revenue driven by Trending is given to YouTube).

After doing this research and defining the monetization strategy, which includes the product strategy and pricing strategy (static fee vs price discrimination), YouTube should start a second phase of price tuning using online information. A commonly used technique consists of launching different prices on different regions and then studying which price brought better results. That gives very valuable information to the business to better shape the final price. AB testing prices in production is also an alternative to optimize them, but one has to be very careful with the legal implications this has (it is not allowed in some markets) and also with the negative impact it can have in a company's image. Drawing conclusions from AB tests also requires a significant volume of data points. In this particular problem we know volume is low, so YouTube may need to run very long AB tests in order to get statistically significant results.

5.4 Addressable market expansion

We decided not to include any video to the addressable market that was not originally trending in our dataset, but what if we do?

We repeat the same willingness to pay analysis and optimization exercise to a broader set of videos. This set includes the false positives we previously decided not to include. We do not include all false positives anyway; instead we just keep those that belong to the right tail of the distributions shared in Figure 4.3 and 4.4. For that we apply the following rules over the false positives: we just keep videos that have more than 50,000 views and less than 4 active days. By doing this, we obtain a set of videos that fit better in the trending distribution, as observed in Figure 5.4 and 5.5.

Note that the distribution of both variables for the non-trending subset of videos are odd. This is because the size of this cohort is really small: only 5 false positive videos meet all the requirements specified.[32] This, of course, is not good: the dataset we use has all YouTube

[32] Three of these videos belong to influencers targeting young audiences; the forth is a russian TopDog fight; the fifth is an Indian Reality TV episode.

trending videos (for the 20 countries being studied) but not all non-trending ones, given the latter implies having hundreds of millions of videos in the dataset. That means that the subset of false positives we ended up with does not represent the whole population of videos on YouTube that look similar to trending videos. Goes without saying that if we narrow down that subset, the small cohort of videos we obtain is far from representative of all the non-trending videos that could be categorized as trending by YouTube. With all this said, we continue with the analysis acknowledging all the bias these assumptions introduce.

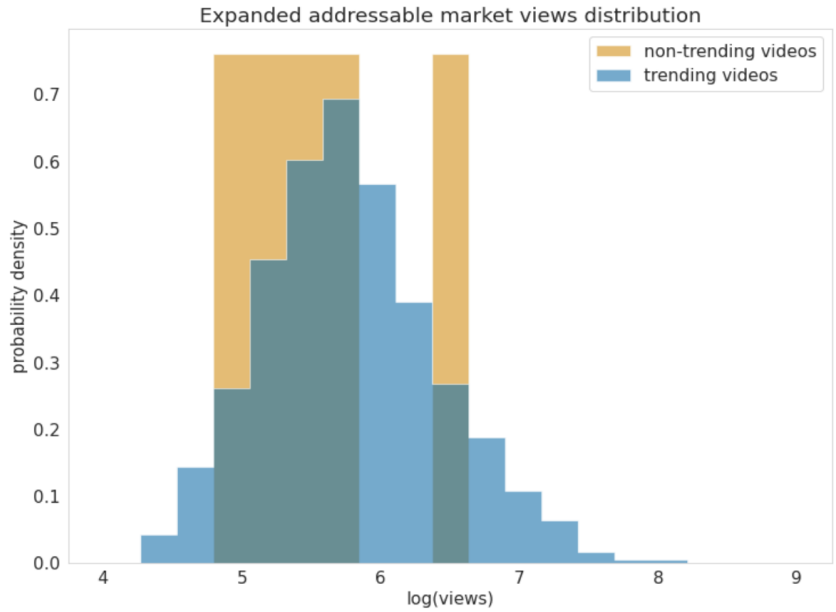

Figure 5.4: $log_{10}(views)$distribution of expanded addressable market.

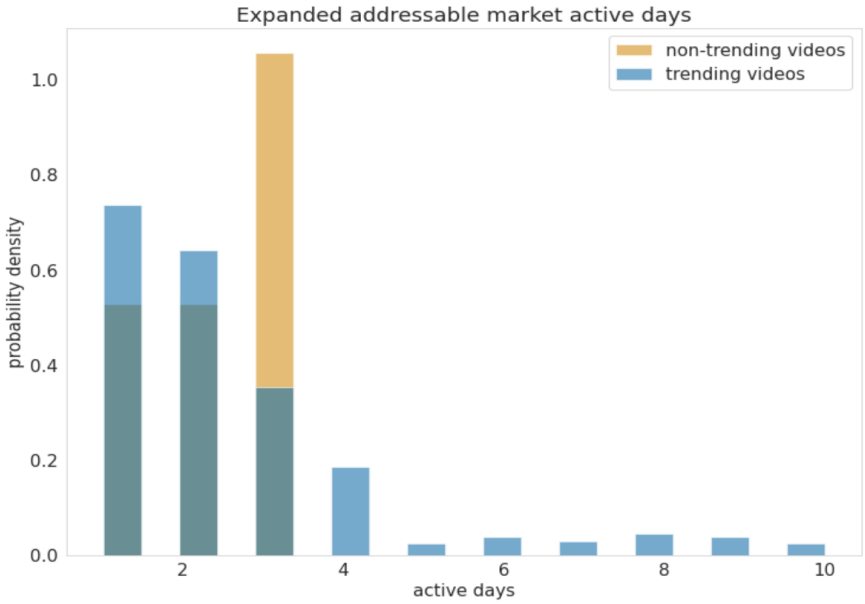

Figure 5.4: Active days distribution of expanded addressable market.

Moving forward, in order to size the market we need to grab these five videos and bootstrap them at a rate of 200x each in order to obtain a number of videos representative of the real world. As mentioned above, the number of trending videos is comprehensive, but not

the number of non-trending ones. This 200x correction factor comes from comparing the estimated daily video upload rate versus the one we have in our sample (the number of videos with a given value of active days is equal to the upload video rate of our dataset for a specific date). We calculate the estimated daily video upload rate with the average duration of videos (we calculate it with the data we already have) and the information shared at the beginning of this work which said that, on average, 500 hours of content is uploaded to YouTube per minute.

By doing this we obtain an expanded addressable market of ~1700 videos where ~1000 are these five false positive multiplicated 200 times each. Figure 5.5 and 5.6 show how the demand for Trending looks like when doing this addressable market expansion. As expected, introducing the bootstrap of the false positives breaks what would be a regular demand behavior: the higher the price of a good, the lower the demanded quantity of it. It appears to be that one of the false positives seems to have a pretty high trending effect, therefore its willingness to pay is high. When bootstrapping it almost ~200 times, we obtain what is on the charts.

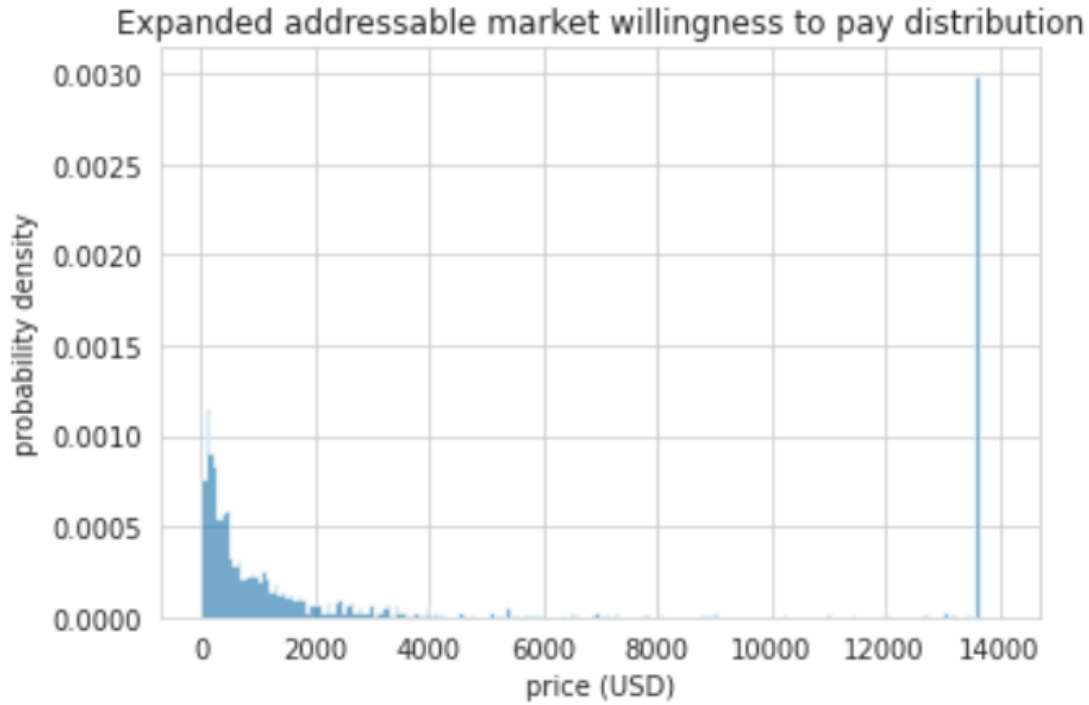

Figure 5.1: Willingness to pay distribution of the expanded addressable market.

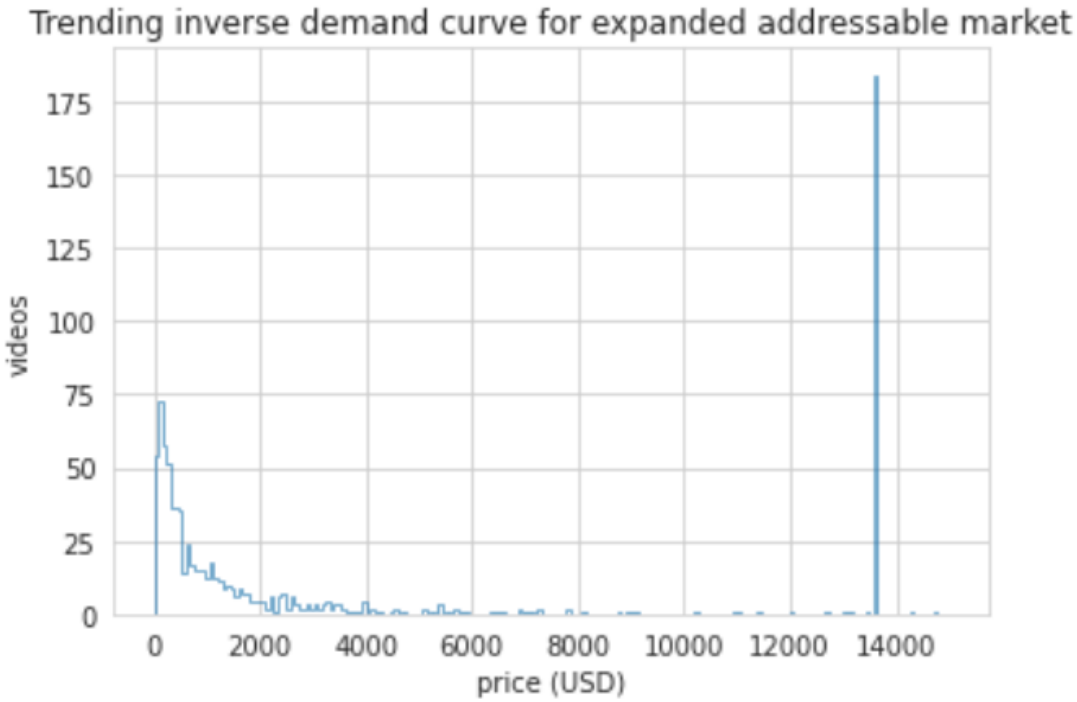

Figure 5.3: Estimated weekly net revenue for Trending as a function of its price for the expanded addressable market.

If we do not stop here and continue the analysis with this subnormal demand curve, we obtain an optimal trending price with the same subnormal characteristics (Figure 5.4): US$301,590. The weekly net revenue in this scenario is US$104,040,500. By year, this is 5.4 trillion dollars: 30 times the revenue Alphabet[33] reported in 2020. It goes without saying how absurd this is. With more data and with better extreme value theory techniques we could get better estimations for non-trending videos and obtain reasonable results.

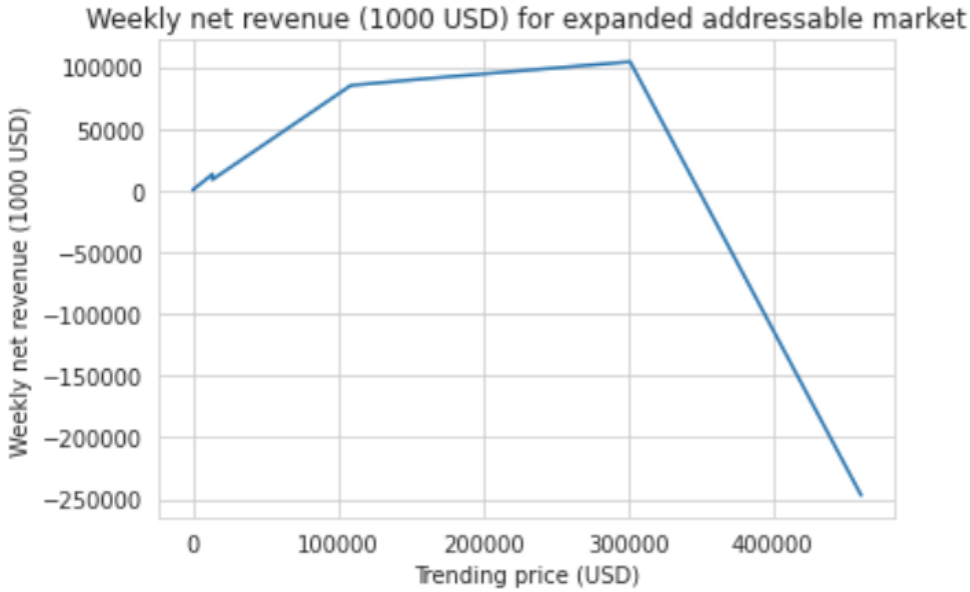

Figure 5.4: Estimated weekly net revenue for Trending as a function of its price for the expanded addressable market.

[33] Parent company of Google and several former Google subsidiaries.

6. Conclusion

YouTube Trending is a feature that increases the number of views a video gets without cannibalizing video views from other sections of the platform. As a consequence, it increases the total number of video views of the whole platform, increasing the advertisement revenue it generates. This incremental revenue means more money for YouTube and for the creators of the videos that make it to Trending. This increment on video views is proportional to the number of views the video naturally gets: videos that are more likely to be seen by people have a higher boost on views when being featured in Trending.

Trending videos are unique, which makes it really hard to find non-trending videos with the same characteristics. The machine learning classifier we built suggests a very small subset of videos as potential trendings. Even though they look very similar, they still come from different populations. Since YouTube's policy for selecting trending videos is very strict, we do not include them as part of the addressable market and we just keep those who YouTube actually selected as such for the counterfactual analysis. By doing this we take a more conservative position when sizing the economic opportunity monetizing Trending means to YouTube.

In the counterfactual analysis we calculate how many views the Trending feature generates to trending videos. With that information we calculate the rational willingness to pay creators should have for those videos, demonstrating that YouTube can capture part (or all) of the value creators make with this feature. YouTube can leverage different ways of capturing that value; some may have higher efficiency than others in terms of how much percentage of the market's willingness to pay can be captured.

With a non-discriminatory pricing scheme we have calculated a potential incremental market of U$154,500 per week for the market sample studied (equivalent to U$8 million per year). On the other extreme, with a first-degree discriminatory pricing scheme YouTube could make up to 11 times more net revenue than this (U$1,670,000 per week, U$87 million per year). What is the best pricing schema and product strategy for launching Trending monetization to production is something YouTube should fine tune with offline and online research and analysis.

Charging for Trending also brings along a high quota of risk. Empty slots on the trending section carries along a huge advertisement opportunity cost when there are potential trending videos out there that were not monetized. How YouTube approaches creators when charging this feature is key: non-paying creators can mean a significant loss of money for the platform.

Something important to bear in mind is that these net revenues calculated are just an educated estimation of the business opportunity of monetizing Trending. We should expect that the real market size will be in the same order of magnitude as the ones calculated in our analysis but it will definitely be different. There are five reasons to conclude this: (1) the data used in this thesis just covers one week of early 2021. Seasonality and the natural noise of the system is being totally ignored; (2) we calculate Trending effect through an observational study instead of an AB test. Observational studies are not perfect; it is impossible to conclude that even after taking all the considerations we have taken we have not omitted a variable that

is introducing some level of bias or that the modeling decisions we took are not capturing the dynamics of the system perfectly; (3) we use statistical models to infer average effects, whose robustness is tied to the level of noise of the system: in a very noisy system, using averages to extrapolate unseen observations will not give the most representative results. A more complex statistical solution for this issue is using heterogeneous treatment effect modeling, which does not assume that the effect is the same for all observations but instead takes into consideration the fact that the effect varies per subject; (4) we make strong assumptions about creators willingness to pay. We assume they will act in a rational way and will count on the sufficient information to make the best informed decision possible; (5) we assume zero cannibalization of video views. Although there are good reasons to accept this assumption, we do not demonstrate it so it is not necessarily true: there might exist some level of cannibalization that we are ignoring.

We could have obtained more representative results if we counted with more relevant attributes of videos. Ideally, more covariates would have permitted performing a propensity matching score inference: there might be a set of video features that generates a higher overlap between trending and non-trending videos, and these features are most likely related to the content of videos: we think that image and audio related features would have been key for this analysis. Even if the new features were not enough for actual matching, they could have increased the fit of the WLS regression model, which would also have improved the estimation of trending effect. Keep in mind that the goal of causal inference is modelling a "what if" scenario, and the more complete the dataset we have, the more precise our estimation can be. Things like the growth speed of a views curve should be highly predictive of future views and is surely used by YouTube for deciding trending allocations.

7. YouTube latest update

Trending has changed during the second quarter of 2021. The data used for this thesis was generated with the version prior to this update. Therefore, all the conclusions and recommendations made in this work are no longer directly applicable to what YouTube currently has.

7.1 New Explore section

Trending entry point can no longer be found on YouTube's homepage. The video platform has replaced this entry point with one for entering to this new section called Explore (Figure 7.1).

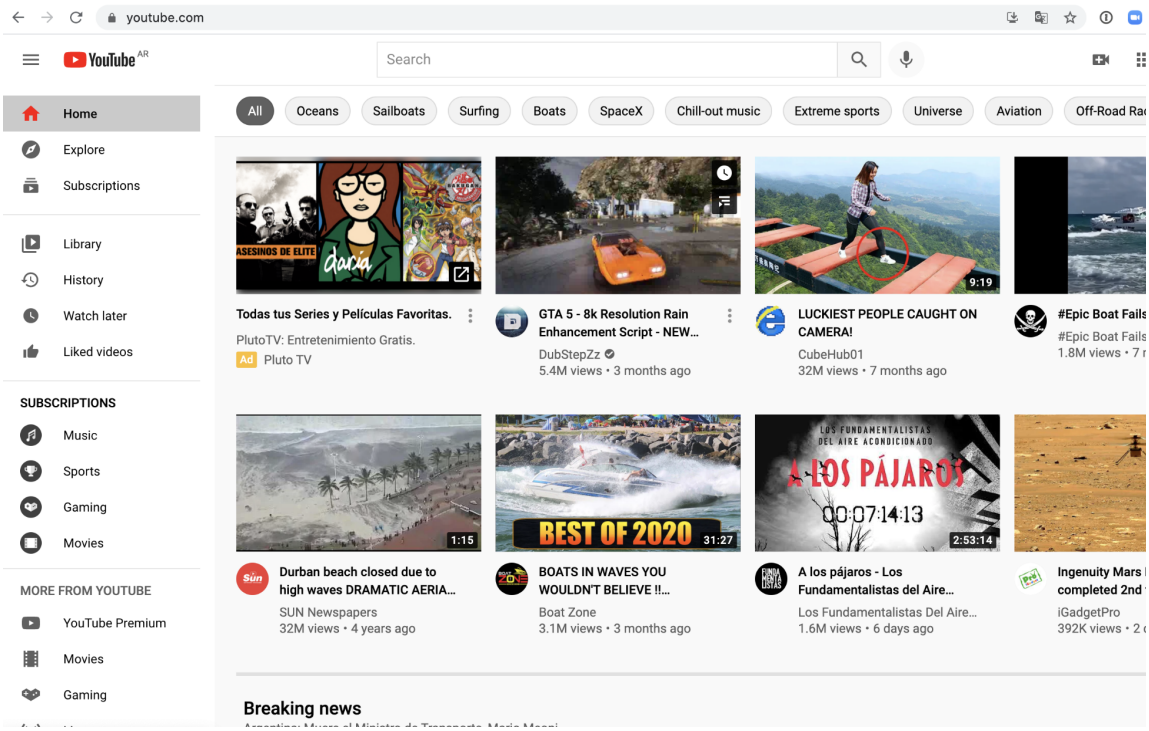

Figure 7.1. YouTube homepage in Q2 2021.

When entering to Explore you can see that there are now multiple categories with recommendations from YouTube (Figure 7.2). One of those categories is Trending, and it is selected by default when entering to Explore. The remaining categories (Music, Gaming, News, Movies, Learning, Live, Sports) redirect users to the “Subscription” sections when clicking on them. A user can enter these category-based sections both through “Explore” as we just mentioned or directly by clicking on them on the left hand panel.

Changes around Trending do not stop there. Once you are in the Explore section you will see 50 most trending videos of your region by default. However, if you click on the Trending icon at the top of the screen, you will get more trending videos organized by category (Figure 7.3). This increases the available trending slots, going from 50 per region to 200, but it is probable that the most trending videos of Music, Gaming & Movies are repeated in the Now tab, which shows the most trending videos of the platform regardless of their category.

There is another slight modification that can also drive user behavior modifications. If you compare Figure 7.2 with Figure 1.2 you can observe that SUBSCRIPTION content in the left hand panel has also changed. Subscription options displayed are now related to classic and mainstream categories (Music, Sports, Gaming, Movies) whereas on the previous version you would see ad-hoc & specific topics.

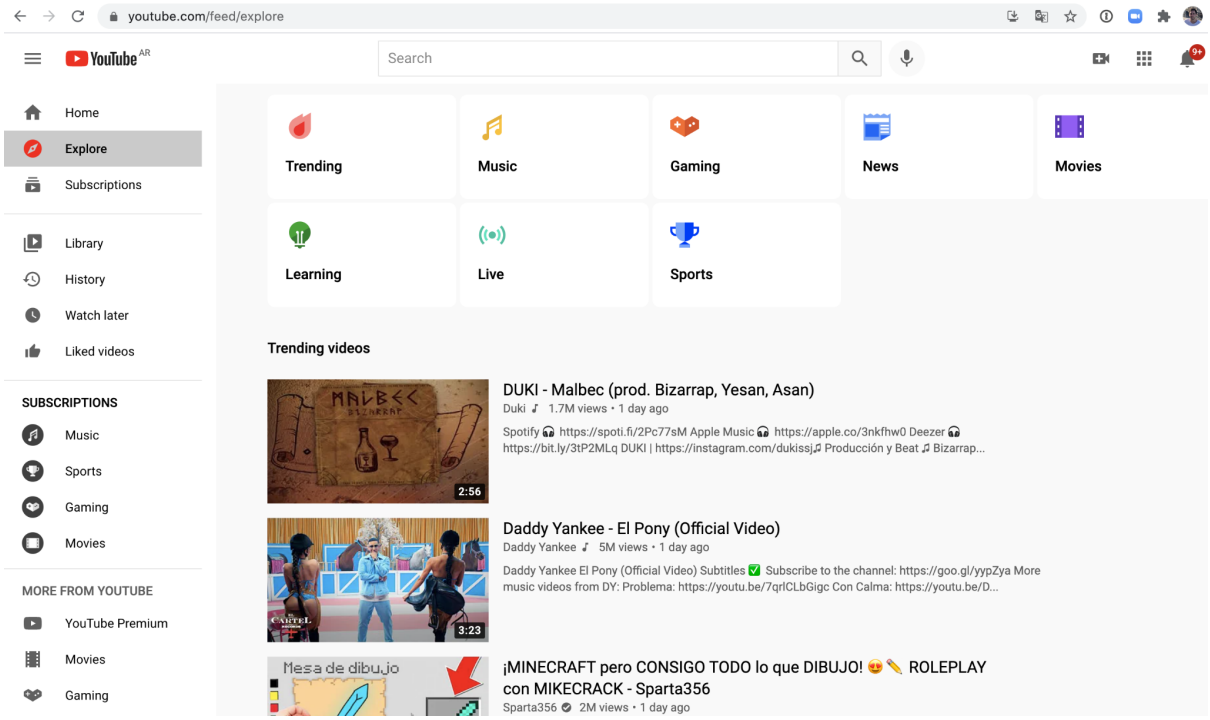

Figure 7.2: New "Explore" section.

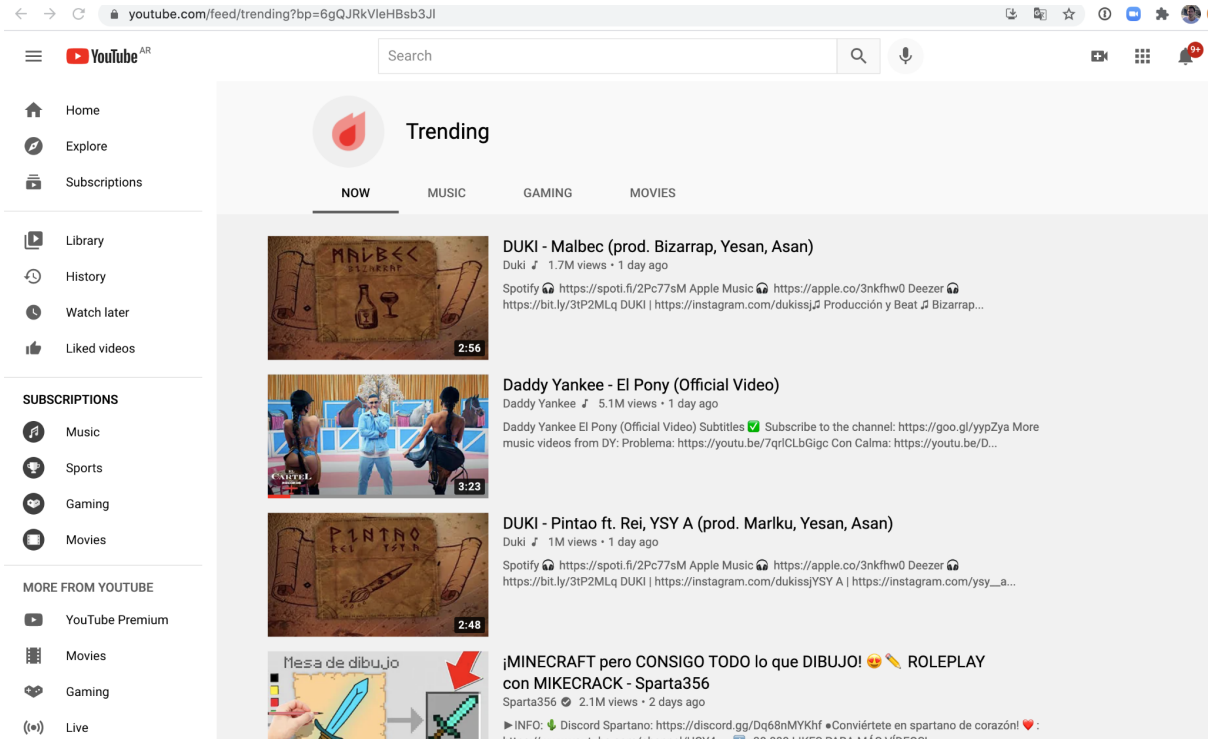

Figure 7.3: New "Trending" section.

7.2 Impact on thesis findings and conclusions

There are three variables that are impacted by this latest update: (1) trending discoverability, (2) conditional probability of trending videos being watched and (3) video views cannibalization.

Small user interface changes can have a big impact in user behavior, especially in consumer applications like YouTube. Conversion from homepage impression to trending/explore click might have been heavily impacted by the modification of the icon and the copy: prior to the change, there was a fire icon that said “Trending”. Now there is a compass that says “Explore”. There are three possible scenarios here: (1) the icon and copy update did not change user inflow to this section, (2) more users are now entering to this section due to this change or (3) less users are entering to Trending/Explore. Points (2) and (3) have a direct impact on the probability of a trending video being discovered, whether it is in a negative or positive direction.

A bigger number of available trending slots implies that YouTube’s trending selection criteria had to be modified. Since the number of trending videos is higher, the level of strictness the platform has towards the selection of these types of videos must have decreased. This implies that the average conditional probability of a video being played once selected must have decreased as well. So here YouTube is dealing with a tradeoff between increasing the discoverability of a higher number of videos (200 videos vs 50 videos) and decreasing their average conditional probability of being watched. Assuming YouTube’s rationale for this feature was increasing the overall number of video views of the platform, we can imagine that the net result of this tradeoff is positive. However, taking into account that now there is more competition within this section on top of the fact that the average propensity to play a video of this section once found is probably lower, the impact Trending may now have individually for each video is probably lower as well. This means that the value delivered to creators that have trending videos is probably lower too, as it might be their willingness to pay for such a feature. We cannot conclude that because of this, the potential net revenue of this feature —if monetized— is lower: revenue is equal to price by quantity, and with this configuration the quantity can be four times bigger than with the older one.

Last of all, we have the cannibalization effect between trending videos and subscription videos. Users that get into the trending section are exposed with entry points to these subscription categories. This reduces the probability of a trending video of being discovered, impacting its probability of being played. This reduces the effect of Trending over video views. Nevertheless, video views lost in the trending section may be captured by this other section; it may not even be a zero-sum game: it may be the case that users leaving Trending because they clicked a subscription category increase their probability of playing a video. The reality is that we do not know how this modification affects the former user dynamics between users and the platform, but we acknowledge that continuing assuming no cannibalization effects with it is no longer accurate, regardless of whether or not the overall net result in terms of total video views ends up being positive.

8. Bibliography and resources

1. *YouTube for Press*, 2020. YouTube. *https://www.youtube.com/intl/en-GB/about/press/*

2. James Hales, May 2019. *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute.* Tubefilter. *https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/*

3. Mountain View, California, February 2020. *Alphabet Announces Fourth Quarter and Fiscal Year 2019 Results.* Alphabet. *https://abc.xyz/investor/static/pdf/2019Q4_alphabet_earnings_release.pdf?cache=05bd9fe*

4. YouTube, 2020. *How does YouTube make money? https://www.youtube.com/howyoutubeworks/our-commitments/sharing-revenue/#:~:text=YouTube%27s%20main%20source%20of%20revenue,%2C%20channel%20memberships%2C%20and%20merchandise*

5. Eric Rosenberg, June 2020. *How YouTube Ad Revenue Works.* Investopedia. *https://www.investopedia.com/articles/personal-finance/032615/how-youtube-ad-revenue-works.asp*

6. Werner Geyser, August 2020. *How Much do YouTubers Make? — A YouTuber's Pocket Guide [Calculator].* Influencer Marketing Hub. *https://influencermarketinghub.com/how-much-do-youtubers-make/*

7. *Developers Google.* YouTube Data API. *https://developers.google.com/youtube/v3/docs/playlists*

8. *ISO 8601.* Wikipedia. *https://en.wikipedia.org/wiki/ISO_8601#Durations*

9. *langdetect 1.0.8.* Pypi. *https://pypi.org/project/langdetect/*

10. W. G. Cochran and S. Paul Chambers, 1965. *The Planning of Observational Studies of Human Populations.* Series A (General) Vol. 128, No. 2, Journal of the Royal Statistical Society.

11. Paul R. Rosenbaum, December 1991. *Discussing Hidden Bias in Observational Studies.* Annals of Internal Medicine.

12. Paul R. Rosenbaum. *Overt Bias in Observational Studies.* Springer Series in Statistics book series (SSS).

13. *Statsmodels v0.12.2. https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html*

14. Bikaramjit Mann and Evan Wood. May 2012. *Confounding in Observational Studies Explained.* The Open Epidemiology Journal. Department of Medicine, University of Calgary, Canada. University of Calgary, Canada. Department of Medicine, University of British Columbia, Canada. BC Centre for Excellence in HIV.AIDS, Canada.

15. Jamal I. Daoud. 2017. *Multicollinearity and Regression Analysis.* Journal of Physics, IOP Publishing Ltd.

16. Dodge, Y. 2008. *The Concise Encyclopedia of Statistics*. Springer.

17. Paul R. Rosenbaum & Donalds B. Rubin, December 1983. *Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score.* The American Statistician.

18. Scott A. A Czepiel. *Maximum Likelihood Estimation of Logistic Models: Theory and Implementation.*

19. Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Sturmer, M. Alan Brookhart, and Marie Davidian, November 17, 2010. *Doubly Robust Estimation of Causal Effects.* American Journal of Epidemiology.

20. *Causal Inference in Python. https://causalinferenceinpython.org/*

21. David A. Freedman, Ricard A. Berk. November 2008. *Weighting Regressions by Propensity Scores.* Ensemble methods for Data Analysis in the Behavioral, Social and Economics Sciences.

22. Xue Ying, February 2019. *An Overview of Overfitting and its Solutions.* Journal of Physics Conference Series.

23. Tulrose Deori, July 2020. *Implement Logistic Regression with L2 Regularization from scratch in Python.* Towards data science.

24. Hossin, M. and Sulaiman, M.N., March 2015. *A Review On Evaluation Metrics For Data Classification Evaluations.* International Journal of Data Mining & Knowledge Management Process (IJDKP). Vol. 5, No.2.