Trend Videos

Statistical Inference | Master in Statistics for Data Science

Javier Muñoz Flores & Luis Ángel Rodríguez García

17-05-2022

1 Introduction

1.1 Topic and motivation

The aim of this project is to analyze what are the most favorite YouTube videos and their categories associated. These trends are measured with the following quantifiers: number of accumulated views, number of accumulated likes, number of accumulated dislikes and number of accumulated comments. Therefore, the higher the value of any of these parameters, the more likely to be in the top of the trendiest videos.

1.2 Description of the dataset

Data comes from Kaggle separate by countries and we has merged all of them into a single csv file. As it is commented in the previous website, the information was obtained through the YouTube API.

The csv file we uploaded at the beginning of the course contains 375,942 rows and 10 variables. It describes the 200 top trending YouTube videos per day for some different countries: United States of America, Canada, Mexico, Japan, South Korea, India, Russia, United Kingdom, France and Germany. Not all countries have the same interval of time. YouTube videos do not have the same period of time between each other either. Let's see the data grouped, aggregating the dates for each group:

Then it is clear that the record dates depend on the `video_id` and the `country_id`. For example, in the original data the interval times for Canada, US, Germany, FR, UK, India, South Korea, Mexico and Russia are from 14-11-2017 to 14-06-2018 and for Japan from 07-02-2018 to 14-06-2018.

Therefore, due to the fact that time-series is not allowed for the purpose of this project, we are going to consider the last registered date time observation for a specific `video_id` and `country_id`.

Besides that, we are applying a couple of tweaks to have our data ready for the analysis. First, we are going to modify the type of a few variables:

- `trending_date` from character to Date class
- `category_id` from integer to factor class
- `country_id` from integer to factor class
- `comments_disables` from character to factor class

Secondly, in order to have available a continuous variable, we are going to change the value of the variable `views` to be measured in miles. Lastly, we are filling in the missing values for the variable `comments_disables` (setting `0` when there are comments and `1` when the number of comments is zero) and removing the variables that we are not going to use afterwards such as `trending_date` and `tags`.

```
data <- youtube %>%
  mutate(
    trending_date = as.Date(trending_date, '%y.%d.%m'),
    category_id = as.factor(category_id),
    country_id = as.factor(country_id),
    comments_disabled = factor(ifelse(is.na(comments_disabled), NA,
                                      ifelse(comments_disabled=="True", 1, 0)))
    ) %>%
  group_by(video_id, country_id) %>%
  slice(which.max(trending_date)) %>%
  select(-c(trending_date, tags)) %>%
  data.frame
data$comments_disabled = ifelse(is.na(data$comments_disabled),
                                ifelse(data$comment_count==0, 1, 0),
                                data$comments_disabled)
data$views = data$views/1000
```

Our final dataset contains 207,148 rows.

1.3 Description of the population

Observing some blogs about the YouTube's figures for 2018, we can say that the total number of YouTube videos approximately was around 8 billion videos. A significantly large population size comparing with the sample size of our data that is just 207,148 observations, then the population size is totally different from the sample size.

The total number of videos in YouTube at the time where the last observation was recorded vary based on the country. Considering the number of active users per country we could approximate the total videos for each market, however this value would not be accurate at all. Let's obtain the number of observations we have for each country.

We could see in the table above that UK and US have a smaller number of observations, nevertheless it does not mean that they have less videos, it could be that they have a large number of videos repeated (consecutive days appeared in the 200 trendiest videos).

We would like to extrapolate our results (top 200 trending videos per day) to the whole population of YouTube videos. At the beginning just for the whole YouTube and later on segmented by country.

1.4 Description of the variables

The dataset has 8 dimensions, these variables are:

- `video_id` (*qualitative nominal variable*) unique identifier for a YouTube video, it is a string of characters of length 11.
- `category_id` (*qualitative nominal variable*) unique identifier for video's category. The mapping between categories and ids is the following one: Film & Animation (`1`), Autos & Vehicles (`2`), Music (`10`), Pets & Animals (`15`), Sports (`17`), Travel & Events (`19`), Gaming (`20`), People & Blogs (`22`), Comedy (`23`), Entertainment (`24`), News & Politics (`25`), Howto & Style (`26`), Education (`27`), Science & Technology (`28`), Nonprofits & Activism (`29`), Movies (`30`), Shows (`43`) and Trailers (`44`).
- `view` (*quantitative continuous variable*) number of views, measured in miles, up to the last record's date for a specific video and country.
- `likes` (*quantitative discrete variable*) number of likes up to the last record's date for a specific video and country
- `dislikes` (*quantitative discrete variable*) number of dislikes up to the last record's date for a specific video and country
- `comments_count` (*quantitative discrete variable*) number of comments up to the last record's date for a specific video and country.
- `comments_disabled` (*qualitative binary variable*) whether the comments are enabled or not, represented as `0` if it is false and `1` if it is true.
- `country_id` (*qualitative nominal variable*) unique identifier for the country where the video is hosted. It consists a string of characters of length 2. The mapping between countries and ids is the following one: United States of America (`US`), Canada (`CA`), Mexico (`MX`), Japan (`JP`), South Korea (`KR`), India (`IN`), Russia (`RU`), United Kingdom (`GB`), France (`FR`) and Germany (`DE`).

As it is indicated in the guide project, we are going to reduce the number of factors for the categorical variables we have: `category_id` and `country_id`. For the first variable, we are going to merge those factors that are intimately related to, having:

- `Film`, that includes: Film & Animation (`1`), Comedy (`23`), Entertainment (`24`), Movies (`30`), Shows (`43`) and Trailers (`44`).
- `Music`, that is just Music (`10`).
- `Education`, that is formed by News & Politics (`25`), Howto & Style (`26`), Education (`27`), Science & Technology (`28`) and Nonprofits & Activism (`29`).
- `Leisure`, that includes: Autos & Vehicles (`2`), Pets & Animals (`15`), Sports (`17`), Travel & Events (`19`), Gaming (`20`) and People & Blogs (`22`).

The number of factors of the variable `country_id` can be reduced using the continent associated instead of the country, therefore this variable will be renamed to `continent_id`. Then, we will have three different ids:

- `America`, that contains United States of America (`US`), Canada (`CA`) and Mexico (`MX`).
- `Europe`, that includes Russia (`RU`), United Kingdom (`GB`), France (`FR`) and Germany (`DE`).
- `Asia`, that contains Japan (`JP`), South Korea (`KR`) and India (`IN`).

Notice that for this last case, we should add the value of the quantifiers for the same video and different countries in the same continent.

```
# new category ids
film_categories <- c(1, 23, 24, 30, 43, 44)
education_categories <- c(25, 26, 27, 28, 29)
leisure_categories <- c(2, 15, 17, 19, 29, 22)

new_category_ids <-
  ifelse(data$category_id %in% film_categories, "Film",
         ifelse(data$category_id %in% education_categories, "Education",
                ifelse(data$category_id %in% leisure_categories, "Leisure", "Music")))

# new continent_id
new_continent_ids <-
  ifelse(data$country_id %in% c("US", "CA", "MX"), "America",
         ifelse(data$country_id %in% c("RU", "GB", "FR", "DE"), "Europe", "Asia"))
final_data <- data %>%
  rename(continent_id = country_id) %>%
  mutate(
    category_id = factor(new_category_ids),
    continent_id = factor(new_continent_ids)
  ) %>%
  group_by(video_id, continent_id, category_id) %>%
  summarize(comments_disabled=comments_disabled[1],
            across(where(is.numeric), sum), .groups="keep") %>%
  data.frame
```

After the previous transformation, we have a dataset with 195,582 observations.

In the table 1 we have summarized some important statistics related to the measures of centrality, variability and shape for the quantitative variables within the dataset.

Table 1: Stats Table

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	skewness	kurtosis
views	0.2	23.2	82.2	561.1	279.5	516003.4	4617.47	46.85	3290.39
likes	0	341	1529	17182	6420	19759060	149734.44	46.63	3537.62
dislikes	0	21	84	1037	329	5462288	20691.41	178.41	40459.19
comment	0	54	243	2141	919	4259927	0.52	0.3	1.57