Top Trending YouTube 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 for a specific country. 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 trending videos for that country.

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 updated 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. Each country does not contain the same period of time and it also depends on the video itself. Let's see the data groupped, aggregating the dates for each group:

```
diff_dates <- youtube %>%
  group_by(country_id, video_id) %>%
  mutate(trending_dates = paste0(trending_date, collapse = "|")) %>%
  distinct(trending_dates) %>%
  data.frame
```

video_id	country_id	trending_dates
n1WpP7iowLc	CA	17.14.11|17.15.11|17.16.11|17.17.11
0dBIkQ4Mz1M	CA	17.14.11|17.15.11|17.16.11|17.17.11|17.18.11
5qpjK5DgCt4	CA	17.14.11|17.15.11|17.16.11|17.17.11
d380meD0W0M	CA	17.14.11|17.15.11|17.16.11|17.17.11
2Vv-BfVoq4g	CA	17.14.11|17.15.11
0yIWz1XEeyc	CA	17.14.11

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 are for Canada, US, Germany, FR, UK, India, South Korea, Mexico and Russia 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 (categorical variable)
- `country_id` from integer to factor class (categorical variable)
- `comments_disables` from character to factor class (binary variable)

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 removing the variables that we are not going to use 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$views = data$views/1000
```

Our final dataset contains 207,148 rows and 8 variables.

1.3 Description of the population

1.4 Description of the variables

The datasets has 8 dimensions, these variables are:

- `video_id` (character variable): unique identifier for a YouTube video
- `category_id` (categorical variable): unique identifier for video's category
- `view` (continuous variable): number of views, measured in miles, up to the record's date for a specific video and country
- `likes` (discrete variable): number of likes up to the record's date for a specific video and country
- `dislikes` (discrete variable): number of dislikes up to the record's date for a specific video and country
- `comments_count` (discrete variable): number of comments up to the record's date for a specific video and country
- `comments_disabled` (binary variable): whether the comments are enabled or not
- `country_id` (categorical variable): unique identifier for the country where the video is hosted (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`)).