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 in the Table 1, aggregated the dates for each group (`vide_id` and `category_id`).

Table 1: Original data example

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 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`.

After all of these changes our modified 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. We have obtained the number of observations we have for each country and visualized them in the Table 2.

Table 2: Number of videos per country

country_id	n
CA	24427
DE	29627
FR	30581
GB	3272
IN	16307
JP	12912
KR	15876
MX	33513
RU	34282
US	6351

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.

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

In the table 3 we have summarized some important statistics related to the measures of centrality, variability and shape for the quantitative variables within the dataset.

Table 3: 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	23739.88	83.93	10519.16

Once we have analyzed the table above, we can conclude that all of the YouTube videos within the dataset at least contain 200 views, not occurring the same for the other three variables where data can have 0 likes, 0 dislikes or 0 comments.

Taking a look to the measures of centrality we obtain an idea of how it is the variability, for example, the variability is much larger for the variable `likes` than the variable `views` since the difference between the mean with respect the minimum value and the maximum value of the variable `views` is considerably smaller than the same calculation for the variable `likes`. In order to verify this assumption we just need to check the value of the standard deviation (a measure of variability), for the case we have just commented the value for the variable `views` is 4617.47 in contrast with the value of the variable `likes` that is 149734.44, which is considerably much larger.

The other other two statistics provide a measure of the shape of the distribution. In particular, the skewness gives an idea of how much shifted the distribution is with respect a Normal distribution. In our case, the table 3 shows that all variables have a positive skewness, meaning that those distributions are right-skewed. It is worthy to say that this insight can be approximately acquired looking at the mean and the 1st and the 3rd quantiles. If the mean is more or less in the middle between the quantiles, then it is more likely to have a skewness close to 0. In our case, the four variables have the mean outside the interval between the 1st and the third quantile. Lastly, it is the kurtosis statistic that measures how heavy are the tails of the distribution. As it is occurred in the skewness, the Normal distribution is taken as the golden standard and the kurtosis is compared with respect this distribution. Therefore, for those variables -`views`, `likes`, `dislikes` and `comments_count`- with a value greater than 3, the distribution is *leptokurtic* (heavier tails than a Normal distribution). Let's check all of this visually thanks to the plots in Figure 1.

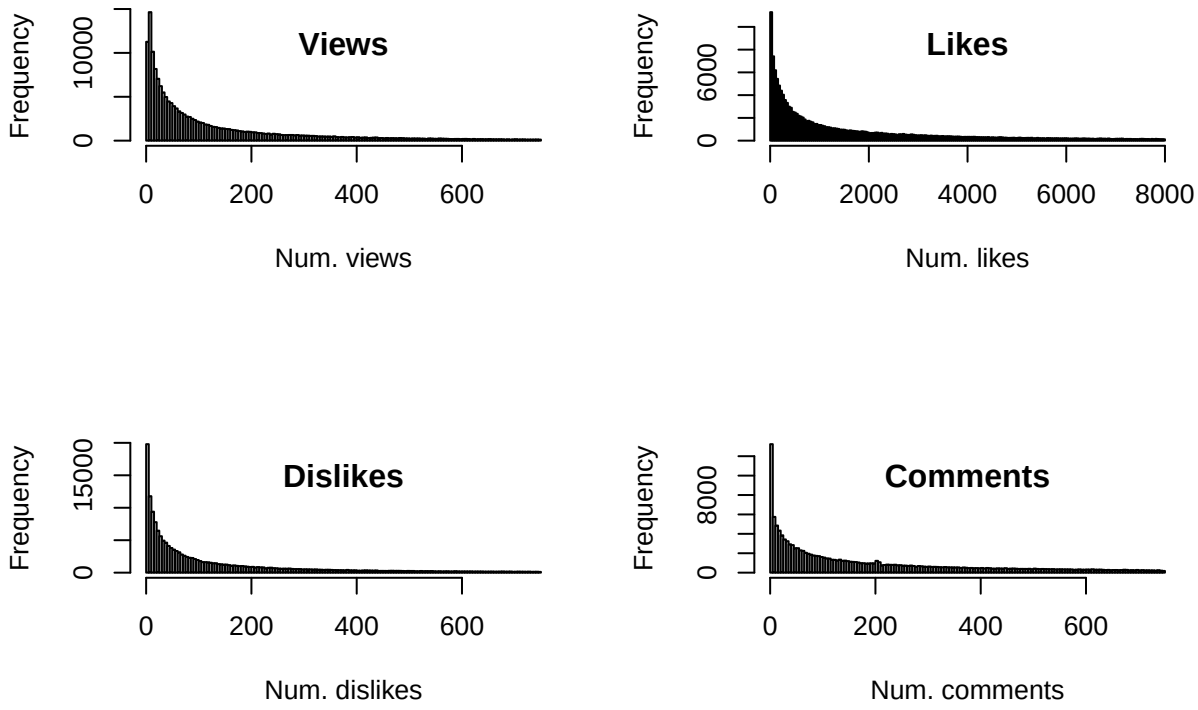

Figure 1: Histograms

In the plot above, we can see clearly the characteristic that was pointed by the kurtosis value, distributions are *leptokurtic*, having heavier tails. It is noticeable as well the fact that there is not any variable that follows a Normal distribution. We will see more about this in the following point.

After describing some distribution details of each of the variables, now it is time to see if they are related to each other. In the following pairs chart, represented in the Figure 2, we can see the relationship between each quantitative variable with others.

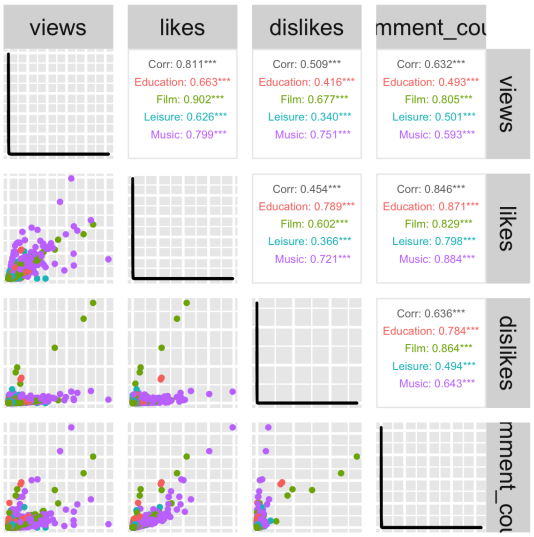

Figure 2: Correlation information

All variables are positive correlated, having that `views` and `likes` are highly correlated (0.811) since the higher of views the more likely that video is liked by the viewers. This logic works also for the number of comments and the number of likes, that's the reason the correlation is high as well (0.846). Moreover, we have shown the correlation split by the different category ids and it is clear that this correlation value depends directly on the category of the video.

Once we have analyzed the quantitative variables, we can move forward and describe a little bit the qualitative variables. For doing this, we are using the frequency tables showed in Table 4.

Table 4: Frequency tables

category_id	**Freq**
Education	43023
Film	79334
Leisure	55391
Music	17834

continent_id	Freq
America	60202
Asia	44386
Europe	90994

comments_disabled	Freq
0	189632
1	5950

The three categorical variables are unbalanced, then there are not the same number of observations for each factor. As we could expect, there are more videos with the comments enabled. It is curious that there are more videos in Europe or America than in Asia, the reason might make sense since YouTube (belonging to Google) is a western company. In case of the number of videos by category, the one that has more videos is the `Film` category, followed by `Leisure`, `Education` and lastly `Music`.

2 Model selection

2.1 Selecting a continous random variable

In this section we have selected the variable `views` with the purpose of estimating its parameters. The reason to choose this variable is based on the importance of knowing how are distributed the visualization of the top 200 YouTube videos. As we could see in the histograms visualized in Figure 1 and commented before, the distribution of this variable is right-skewed so it is far from being distributed with a Normal. To achieve normality, or at least being closer, we could apply the log transformation to our variable. Notice that in this case, since the values are greater than zero, no offset is needed.

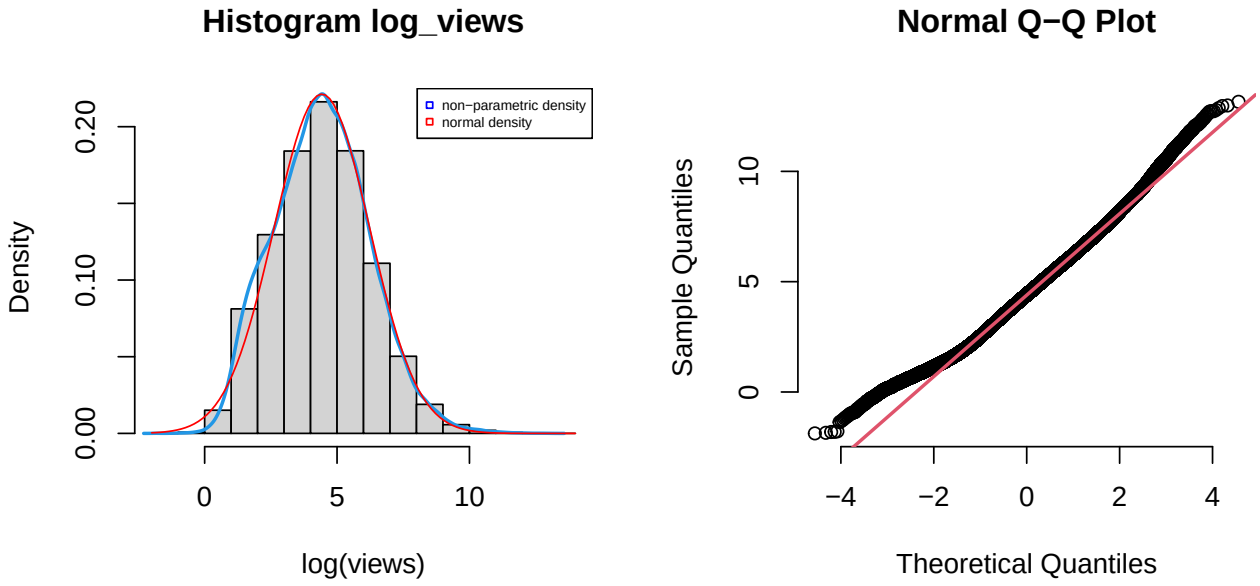

Figure 3: Log transformation of views variable

In the Figure 3 it is possible to see two different plots that checks whether normality is achieved or not. The plot on the left shows in total three components: two non-parametric components (the histogram of the distribution after applying the transformation and the blue line showing the density distribution of our data) and the third one that is a red line which shows the theoretical Normal density taking into account a estimation of the parameters of our data (sample mean and sample standard deviation). As we can see, the blue line fits almost perfectly the red line, except for the tails where lines are slightly different. This difference is also noticeable in the Q-Q plot, a chart that compares the sample quantiles with the theoretical Normal ones, showing that in the extremes of the sample the distribution is slightly different than a Gaussian but in general the distribution is adjusted perfectly. In conclusion, the `log_views` variable adapts smoothly to a Normal distribution therefore the untransformed variable `views` contains data that can be assumed to follow a log-normal distribution. Then, we have that:

$$X_{views} \sim \text{Lognormal}(\mu, \sigma^2)$$

where μ and σ^2 are the parameters of interest. Having its probability density function as it follows:

$$f(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right)$$

In the following points we will estimate the parameters of interest of the distribution we have assumed for our variable `views`.

2.2 Estimate the model parameters

In this point we will estimate the parameters of interest of $X_{views} \sim \text{Lognormal}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, $\theta_1 = \mu$ and $\theta_2 = \sigma^2$, using two different approaches: method of moments and maximum likelihood estimation.

2.2.1 Method of moments

First, let's define the population moments as: $\alpha_r = \alpha_r(\theta_1, \theta_2) \triangleq \mathbb{E}[X^r]$ and the sample moments as $\text{a}_r \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i^r$. Then, the moments estimators of θ_1 and θ_2 are the solutions of the 2 equations obtained by equating 2 population moments to the corresponding sample moments: $\alpha_{r_2}(\theta_1, \theta_2) = \text{a}_{r_2}$.

Recall that the log-normal distribution has $\mathbb{E}[X] = e^{\mu+\sigma^2/2}$ and the normal distribution for the log transformation has $\mu = \mathbb{E}[\log X]$, then we can formulate the parts of our system of equations. In the population we have that $\alpha_{1_{\mathcal{N}}} = \mu = \mathbb{E}[\log X]$ and $\alpha_{1_{\mathcal{LN}}} = \exp\{\hat{\mu}_{mm} + \hat{\sigma}^2_{mm}/2\} = \mathbb{E}[X]$ and in the sample side we have that $\text{a}_{1_{\mathcal{N}}} = \frac{1}{n}\sum_{i=1}^{n} \log x_i = \overline{\log x}$ and $\text{a}_{1_{\mathcal{LN}}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, then equating $\alpha_{1_{\mathcal{N}}} = \text{a}_{1_{\mathcal{N}}}$ and $\alpha_{1_{\mathcal{LN}}} = \text{a}_{1_{\mathcal{LN}}}$ we get to our system of equations:

$$\begin{cases} \hat{\mu}_{mm} = \overline{\log x} \\ e^{\hat{\mu}_{mm} + \hat{\sigma}^2_{mm}/2} = \bar{x} \end{cases}$$

Resolving analytically the second equations we have that:

$$\begin{aligned} e^{\hat{\mu}_{mm}} e^{\hat{\sigma}^2_{mm}/2} &= e^{\frac{1}{n}\sum_{i=1}^{n} \log x_i} e^{\hat{\sigma}^2_{mm}/2} \\ &= e^{\sum_{i=1}^{n} \log x^{1/n}} e^{\hat{\sigma}^2_{mm}/2} \\ &= e^{\log x_1^{1/n}} \dots e^{\log x_n^{1/n}} e^{\hat{\sigma}^2_{mm}/2} \\ &= x_1^{1/n} \dots x_n^{1/n} e^{\hat{\sigma}^2_{mm}/2} \\ &= \prod_{i=1}^{n} x_i^{1/n} e^{\hat{\sigma}^2_{mm}/2} \\ &= \text{geo}(x) e^{\hat{\sigma}^2_{mm}/2} \end{aligned}$$

2.2.2 Maximum likelihood

3 Apendix

3.1 Code

```
knitr::opts_chunk$set(echo = TRUE, fig.align = "center")
library(dplyr)
library(stringr)
library(moments)
library(kableExtra)
library(GGally)
load("dataset.RData")
# 1.

## Show original data: dates per video_id and country_id

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

knitr::kable(head(diff_dates), caption = "Original data example") %>%
  row_spec(0, bold=TRUE) %>%
  kable_styling(latex_options = "HOLD_position", font_size = 7)


# Applying changes to the original data. Change type of variables (date, factor)

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$comments_disabled[is.na(data$comments_disabled)] <-
  ifelse(data[is.na(data$comments_disabled),]$comment_count==0, 1, 0)
data$views = data$views/1000


# Visualize the number of videos per country

count <- data %>%
  group_by(country_id) %>%
  summarise(n = n())

knitr::kable(count, caption = "Number of videos per country") %>%
  row_spec(0, bold=TRUE) %>%
```

```
  kable_styling(latex_options = "HOLD_position", font_size = 7)


# Simplifying categorical variables: category_id and country_id

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_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


# Statistics table for the quantitative variables

stats_quant <- t(summary(final_data[,5:8]))
stats_data  <- t(sapply(((5:8)-4), function(i) {
  sapply(1:6, function(j) {
    str_trim(str_split(stats_quant[i,], ":")[[j]][2])
  })
}))
colnames(stats_data) <- drop(t(sapply(1:6, function(i) {
    str_trim(str_split(stats_quant[1,], ":")[[i]][1])
})))
rownames(stats_data) <- colnames(final_data[,5:8])
rownames(stats_data)[4] <- "comment"

skewness <- round(c(skewness(final_data$views), skewness(final_data$likes),
              skewness(final_data$dislikes),
              skewness(final_data$comment_count)), digits=2)
kurtosis <- round(c(kurtosis(final_data$views), kurtosis(final_data$likes),
              kurtosis(final_data$dislikes),
              kurtosis(final_data$comment_count)), digits=2)
sd <- round(c(sd(final_data$views), sd(final_data$likes),
              sd(final_data$dislikes),
              sd(final_data$comment_count)), digits=2)
```

```r
stats_data <- cbind(stats_data, sd, skewness, kurtosis)

knitr::kable(stats_data, digits = 2, caption = "Stats Table") %>%
  column_spec(1, bold=TRUE) %>% row_spec(0, bold=TRUE) %>%
  kable_styling(latex_options = "HOLD_position", font_size = 7)

# Histograms for the quantitative variables, selecting different breaks parameter
par(mfrow=c(2,2))
hist(final_data$views[final_data$views >= 0 & final_data$views < 750],
     breaks=seq(0, 750, by=5), main="", xlab="Num. views")
title("Views", line=-1.5)
hist(final_data$likes[final_data$likes >= 0 & final_data$likes < 8000],
     breaks=seq(0, 8000, by=40), main="", xlab="Num. likes")
title("Likes", line=-1.5)
hist(final_data$dislikes[final_data$dislikes >= 0 & final_data$dislikes < 750],
     breaks=seq(0, 750, by=5), main="", xlab="Num. dislikes")
title("Dislikes", line=-1.5)
hist(final_data$comment_count[final_data$comment_count >= 0 &
                                final_data$comment_count < 750],
     breaks=seq(0, 750, by=5), main="", xlab="Num. comments")
title("Comments", line=-1.5)

# Generating pairs plot with correlation values and save the image

ggpairs(final_data, columns = 5:8,
        mapping = ggplot2::aes(colour=final_data$category_id),
        upper = list(continuous = wrap("cor", size = 1.5)),
        lower = list(continuous = wrap("points", size=0.5),
                     combo = wrap("dot", size=0.5))) +
    theme(axis.line=element_blank(),
          axis.text=element_blank(),
          axis.ticks=element_blank())

ggsave(file="corr.png", dpi=400)

knitr::include_graphics("corr.png")

# Frequency tables for the qualitative variables

cat_table <- data.frame(table(final_data$category_id))
colnames(cat_table)[1] <- "category_id"
cont_table <- data.frame(table(final_data$continent_id))
colnames(cont_table)[1] <- "continent_id"
comm_table <- data.frame(table(final_data$comments_disabled))
colnames(comm_table)[1] <- "comments_disabled"
knitr::kable(list(cat_table, cont_table, comm_table), caption = "Frequency tables") %>%
  column_spec(1, bold=TRUE) %>% row_spec(0, bold=TRUE) %>%
  kable_styling(latex_options = "HOLD_position", font_size = 7)

# 2.

# Create variable log_views
```

```r
final_data <- final_data %>% mutate(
  log_views = log(views)
)

# Plot the histogram and the QQ-plot

par(mfrow = c(1,2))
hist(final_data$log_views, probability = TRUE, main="Histogram log_views",
     xlab = "log(views)")
lines(density(final_data$log_views), col = 4, lwd = 2)
curve(dnorm(x, mean=mean(final_data$log_views), sd=sd(final_data$log_views)),
      add=TRUE, col="red")
legend('topright', legend=c('non-parametric density', 'normal density'),
       col = c('blue', 'red'), pch = 0, cex = 0.5)
qqnorm(final_data$log_views, pch = 1, frame = FALSE)
qqline(final_data$log_views, col = 2, lwd = 2)
```