YouTube Trending Report

Danping Liu

12/3/2020

Abstract

YouTube is one of the most visited websites. There are millions of users watching YouTube videos every day. In this project, I take a look into the videos on different country's daily trending and build a mixed effect logistic model to predict if a first shown trending video could trend for more than one day. From analyzing the model, I found that traveling or events videos and videos on United Kingdom's trending are most likely to trend longer.

Introduction

The dataset I am using is YouTube Trending data by Mitchell J on Kaggle. This dataset contains videos' data such as view count, likes count, and comment count on 10 countries' trending from November 14, 2017 to June 14, 2018. Most videos could only be on trending once, I would like to know that given the video's first trend day's data, is this video going to trend again, or keep trending.

Method

Data Cleaning

The original data is stored in different countries, then I integrate them into one. There are 375942 observations and 18 variables in total, each row contains a video's data from a country's trending rank on a specific day.

I omitted some unneeded columns and did some mutations so that each row represents each video's data on a country's trending. Some videos appear more than once because they get into multiple countries' trend ranking. Since I am taking likes, dislikes, and comment count into consideration, and about 96.6% of data is comment enabled and ratings enabled, I decided to not include the comment and ratings disabled data.

Then I found that all numerical variables are extremely left-skewed, so I did a log transformation on them. I add a new column trend_longer indicates whether the video trend for more than one day.

	trend_days	log_first_views	log_first_likes	log_first_dislikes	log_first_comments
	Min. : 1.000	Min. : 4.762	Min. : 0.000	Min. : 0.000	Min. : 0.000
	1st Qu.: 1.000	1st Qu.:10.062	1st Qu.: 5.969	1st Qu.: 3.178	1st Qu.: 4.143
	Median : 1.000	Median :11.273	Median : 7.419	Median : 4.443	Median : 5.557
	Mean : 1.761	Mean :11.244	Mean : 7.400	Mean : 4.465	Mean : 5.468
	3rd Qu.: 2.000	3rd Qu.:12.415	3rd Qu.: 8.806	3rd Qu.: 5.694	3rd Qu.: 6.845
	Max. :38.000	Max. :18.781	Max. :15.313	Max. :13.693	Max. :13.717

EDA

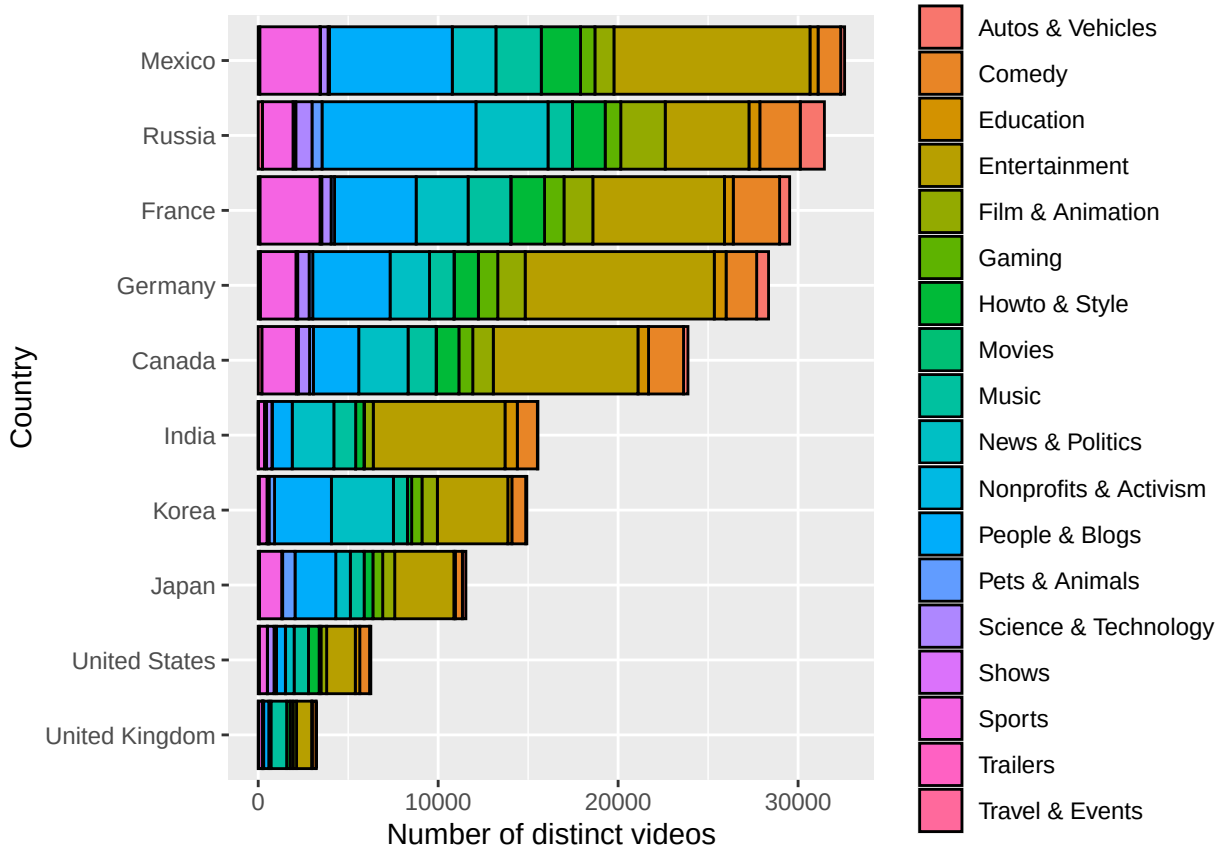


Figure 1: Video distribution

From the graph we can see that Mexico has the largest number of distinct videos on trending, meaning that trending in Mexico updates very frequently and each video tends to trend very shortly. Videos in the United Kingdom tend to trend longer. Entertainment is the most popular category in most countries.

Model

$$trend\ longer \sim$$
$$log(first\ day's\ view\ counts) * log(first\ day's\ likes) * log(first\ day's\ dislikes) *$$
$$log(first\ day's\ comment\ counts) + (1|category) + (1|country)$$

The multilevel binomial logistic model I use to predict trend_long, which means whether a video could keep trending on the second day, with predictors of log(first day's view counts), log(first day's likes), log(first day's dislikes) and log(first day's comment counts) and their interaction terms. The model has varying intercepts among category and country.

Model Check

I use a binned residual plot to assess the overall fit of my model.

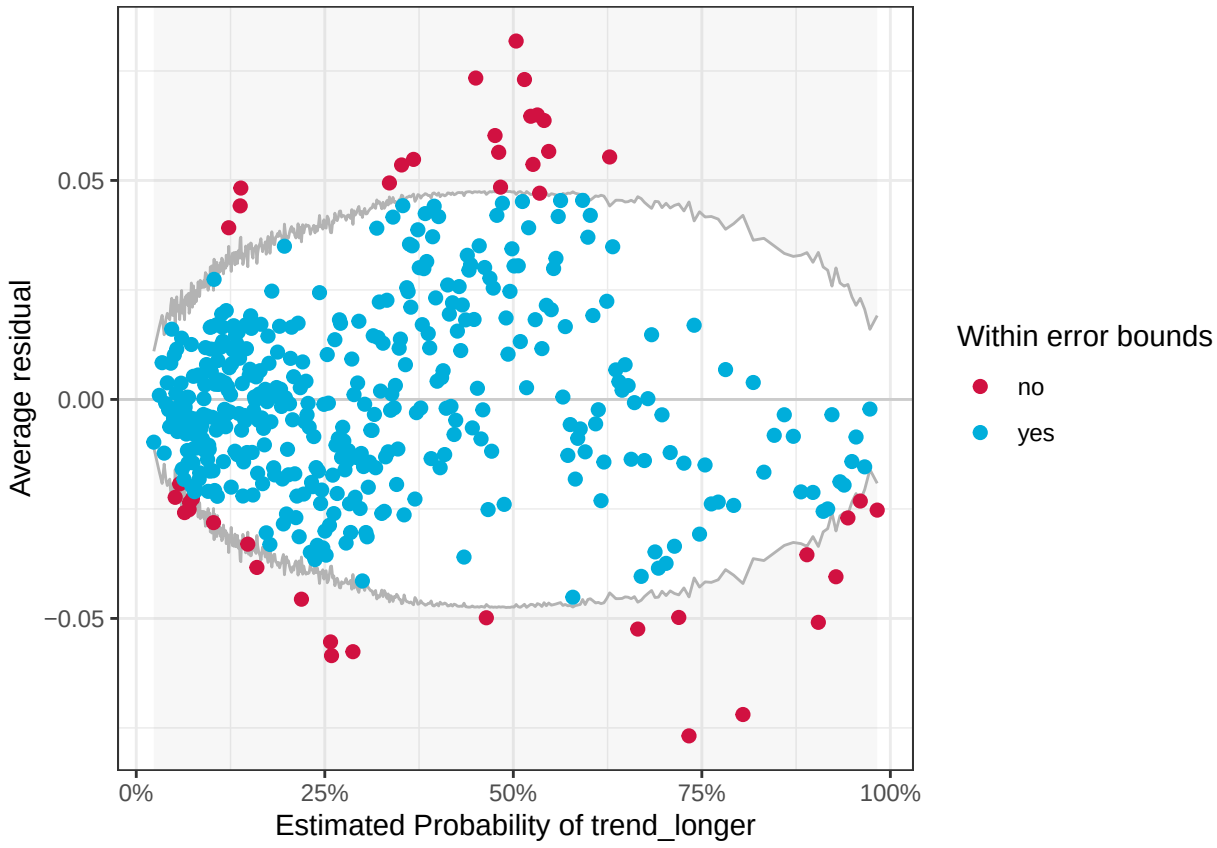


Figure 2: Binned residual plot

```
## Warning: About 90% of the residuals are inside the error bounds (~95% or higher would be good).
```

There are a few outliers, but most of the residuals are within the confidence limits, and they do not show an obvious pattern.

Result

First, let us take a look at the odds ratios for the fixed effects.

	trend_longer		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.11	0.03 – 0.33	**<0.001**
log_first_views	1.18	1.07 – 1.29	**<0.001**
log_first_likes	0.60	0.51 – 0.70	**<0.001**
log_first_dislikes	1.57	1.34 – 1.84	**<0.001**
log_first_comments	0.51	0.43 – 0.60	**<0.001**
log_first_views * log_first_likes	1.03	1.01 – 1.04	**0.004**
log_first_views * log_first_dislikes	0.94	0.93 – 0.96	**<0.001**
log_first_likes * log_first_dislikes	1.10	1.07 – 1.12	**<0.001**
log_first_views * log_first_comments	1.08	1.06 – 1.10	**<0.001**
log_first_likes * log_first_comments	1.12	1.10 – 1.14	**<0.001**
log_first_dislikes * log_first_comments	0.97	0.95 – 1.00	**0.035**
(log_first_views * log_first_likes) * log_first_dislikes	1.00	1.00 – 1.00	**0.004**
(log_first_views * log_first_likes) * log_first_comments	0.99	0.99 – 0.99	**<0.001**
(log_first_views * log_first_dislikes) * log_first_comments	1.00	0.99 – 1.00	**0.004**
(log_first_likes * log_first_dislikes) * log_first_comments	1.00	1.00 – 1.00	0.137
(log_first_views * log_first_likes * log_first_dislikes) * log_first_comments	1.00	1.00 – 1.00	**<0.001**

Figure 3: fixed effct odds ratios

The fixed effect odds ratios tell us that if a video is viewed more and gets more dislikes on the first trend day, it is more likely to keep trending. If a video gets more likes and comments on the first trend day, it is less likely to keep trending. The interaction effects are not that significant.

The random effect odds ratios tell us that:

- For category:
  - Videos in these categories tend to trend for more than one day: travel & events, trailers, pets & animals, people & blogs, nonprofits & activism, music, movies, gaming, film & animation, entertainment, comedy.
  - Videos in these categories tend to trend for only one day: sports, science & technology, news & politics, howto & style, education, autos & vehicles.

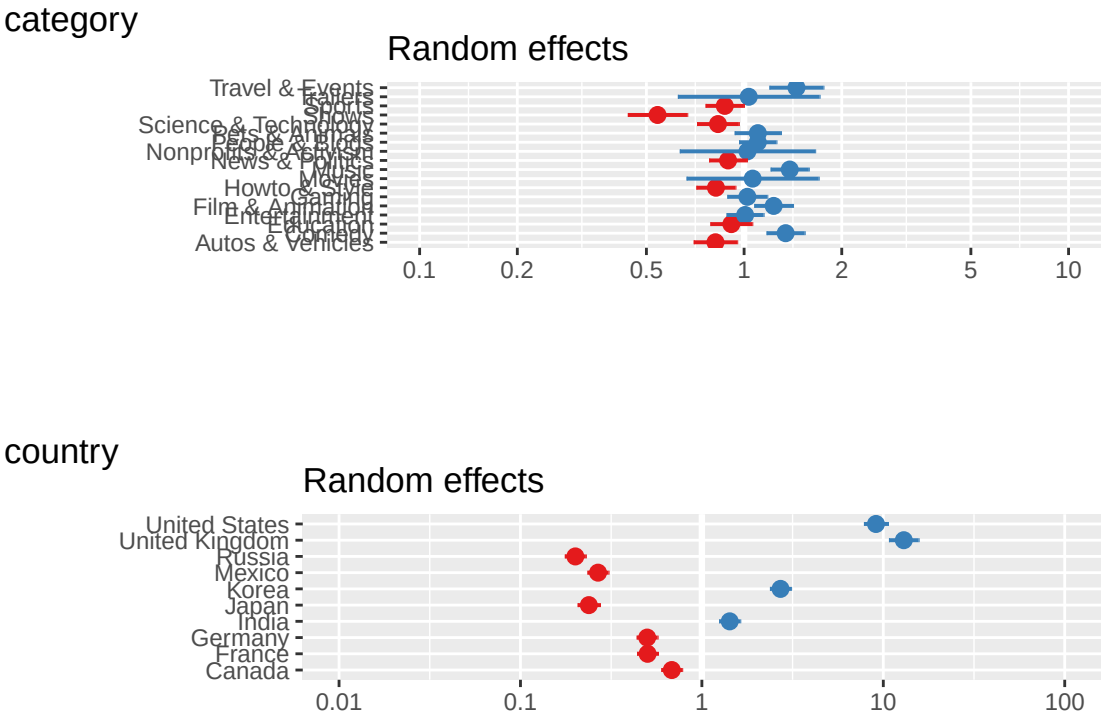


Figure 4: Random effect odds ratios

  - Among these categories, travel & events videos are most likely to trend for more than one day, shows are least likely to trend for more than one day.
  - The effect of trailers, nonprofits & activism, and movies differ a lot in different countries.

- For country:
  - Videos in these countries tend to trend for more than one day: United States, United Kingdom, Korea, India.
  - Videos in these countries tend to trend for more than one day: Russia, Mexico, Japan, Germany, France, Canada.
  - Among these countries, videos on the United Kingdom's trending are most likely to trend for more than one day, videos on Russia's trending are least likely to trend for more than one day.

Discussion

Other than multilevel binomial logistic regression, I tried several other models. I also tried different dependent variables because my initial purpose is to predict how long a video could trend. In the final version, I use trend_longer as the output. But trend_longer only shows if the video trend for more than 1 day. There are also lots of videos that trend for 2 or 3 days. I tried to use trend days as the dependent variable. I also tried to use multinomial logistic regression, with an output of 1 meaning that the video trend for 1 day, 2 meaning that the video trend for more than 1 day but within 7 days, And 3 meaning that video trend for more than a week. These tries have all failed because of the large number of videos that trend for only once. I tried to change my way of thinking to use log(final trend day's view count) as output. And the models run well. But the model itself is not helpful for me to analyze my problem. The numerical predictors in my model are all highly correlated, so I add interaction terms between each of them. But I think next time I should probably add more predictors. For example, I am doing sentiment analysis recently, so I could use the sentiment score for the description or the tags for the videos.

Appendix

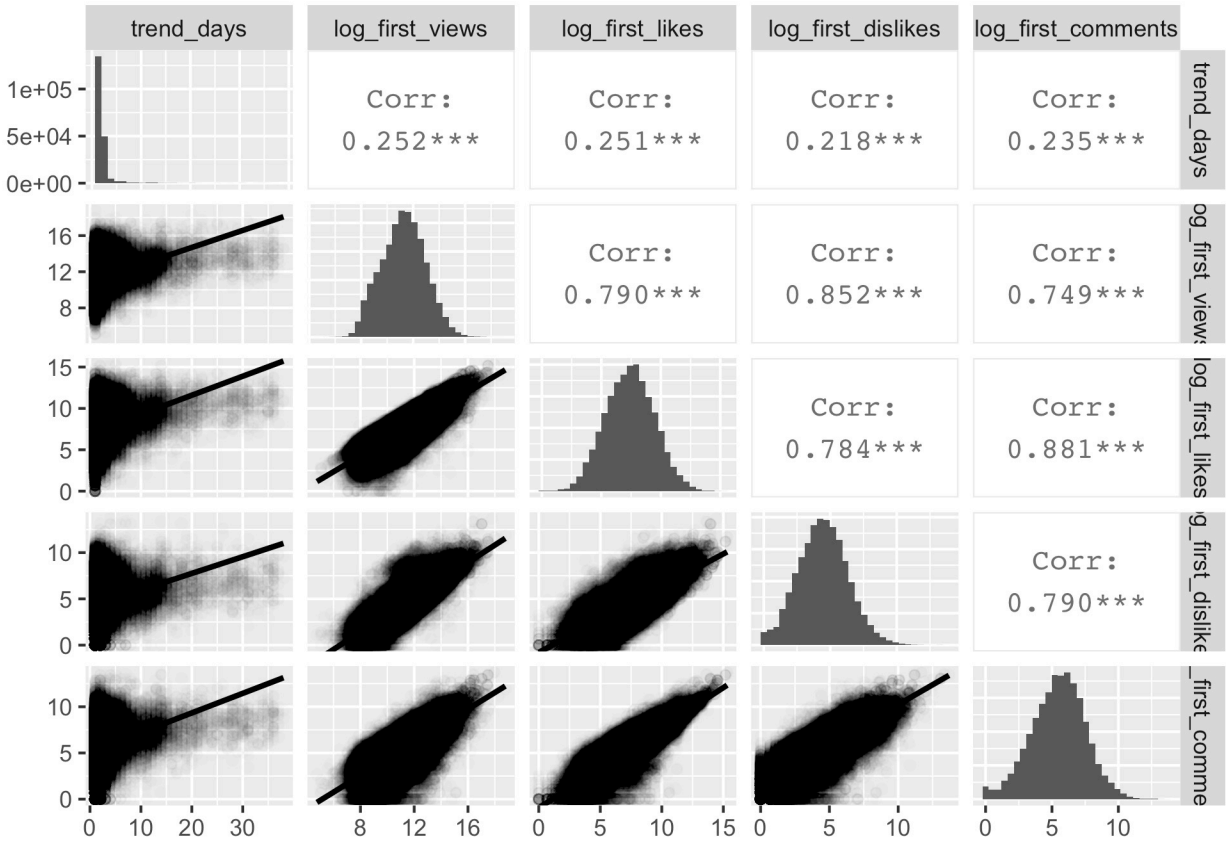
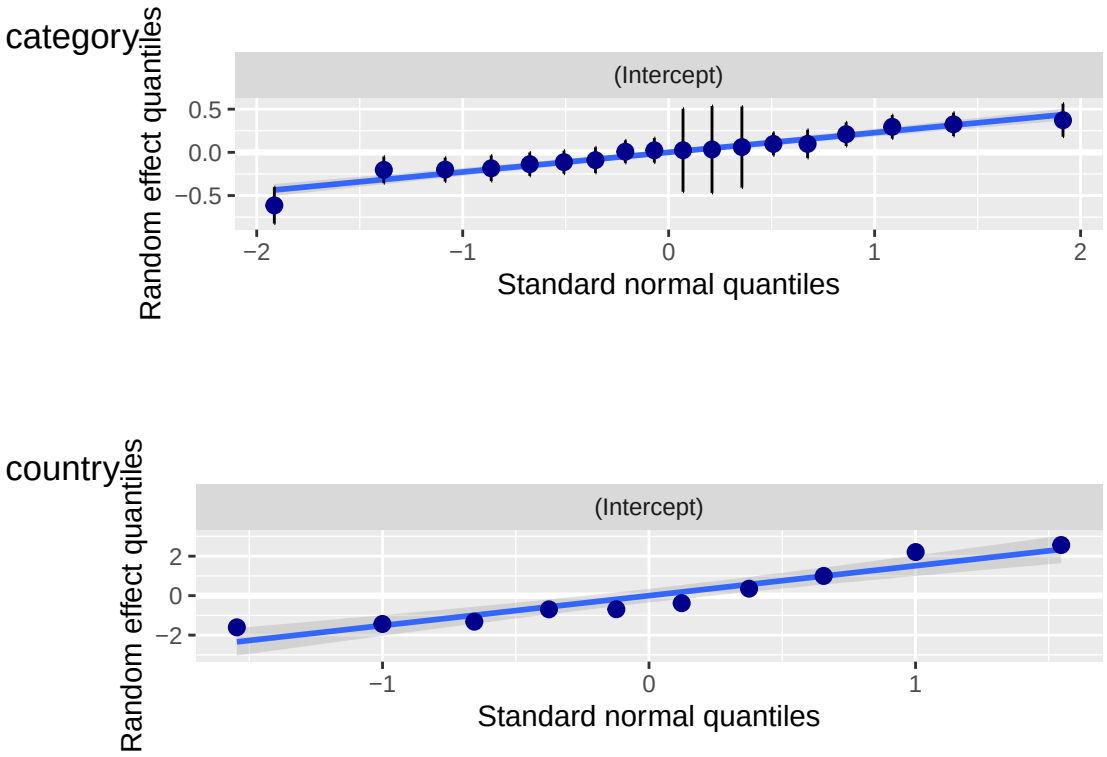


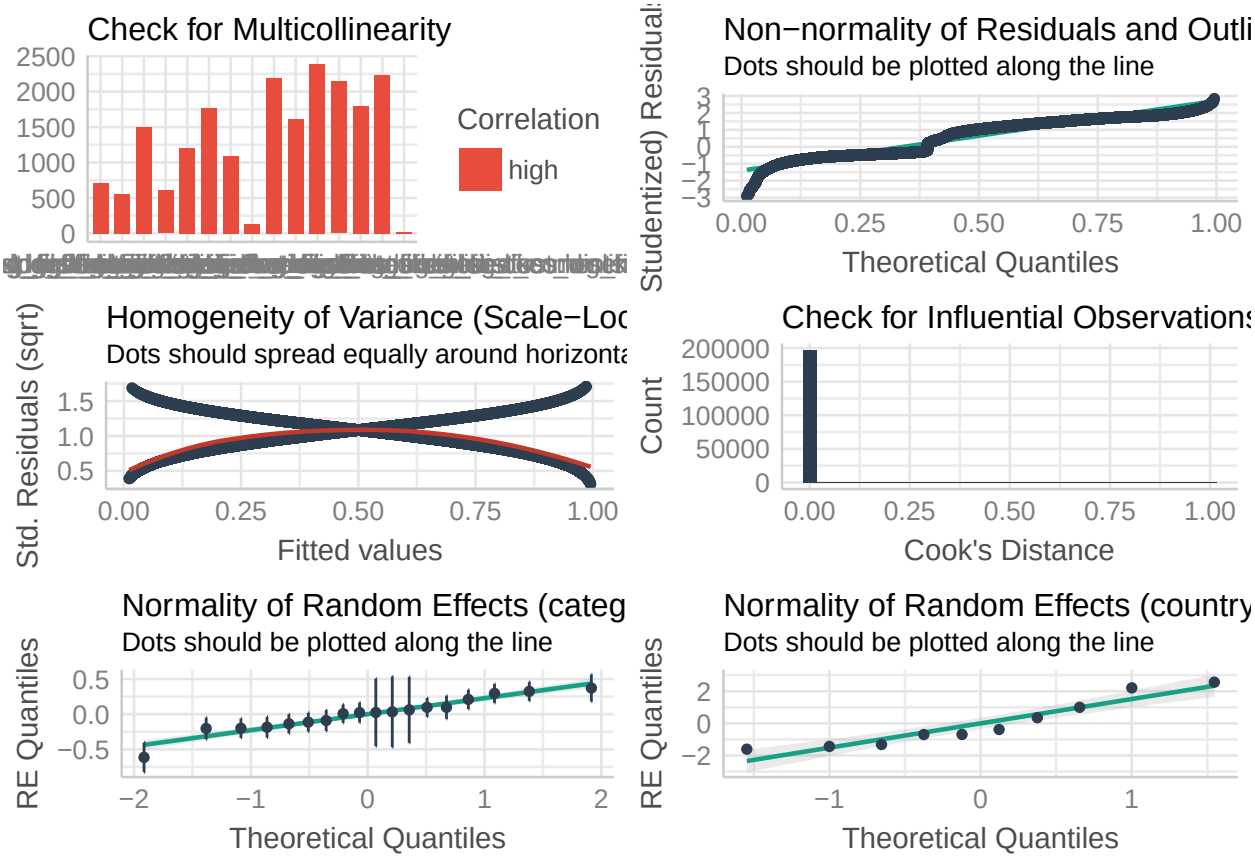
Figure 5: numerical data ggpairs

From the ggpair plot we can see that after log transformation, most of the numerical data seems to be more normalized. View count, likes, dislikes, and comment counts are highly correlated with each other. The left plots are not very clear because most videos trend for only one day. But we can still see that videos that trend longer tend to have a better performance on their first trend day.

Checking normality for random effects using Q-Q plot



Most of the points are close to the normality assumption line, but the tails deviate a little.



Bibliography

- Alex Couture-Beil (2018). rjson: JSON for R. R package version 0.2.20. https://CRAN.R-project.org/package=rjson
- Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm
- Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2020). GGally: Extension to 'ggplot2'. R package version 2.0.0. https://CRAN.R-project.org/package=GGally
- Douglas Bates and Martin Maechler (2019). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-18. https://CRAN.R-project.org/package=Matrix
- Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
- Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
- Hadley Wickham (2020). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.0. https://CRAN.R-project.org/package=forcats
- Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr
- Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra
- Kirill Müller and Hadley Wickham (2020). tibble: Simple Data Frames. R package version 3.0.4. https://CRAN.R-project.org/package=tibble
- Lionel Henry and Hadley Wickham (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr
- Lüdecke D (2020). *sjPlot: Data Visualization for Statistics in Social Science*. R package version2.8.6, <URL: https://CRAN.R-project.org/package=sjPlot>.
- Lüdecke, Makowski, Waggoner & Patil (2020). Assessment of Regression Models Performance. CRAN. Available from https://easystats.github.io/performance/
- Makowski, D., Lüdecke, D., & Ben-Shachar, M.S. (2020). Automated reporting as a practical tool to improve reproducibility and methodological best practices adoption. CRAN. Available from https://github.com/easystats/report. doi: .
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.
- Yihui Xie (2020). tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents. R package version 0.27.