Data: Trending YouTube Video Statistics

Source: kaggle.com user Mitchell J

About the data

A brief history of YouTube

Sources: engadget.com, brandwatch.com

About the data

A brief history of YouTube

- Founded in 2005 by three PayPal employees.
  - Jawed Karim, Steve Chen, and Chad Hurley
- Less than a year after the site launched in December 2005, Google acquired YouTube for around $1.65 Billion.
- Today, YouTube is the third most visited site behind Google and Facebook.

Sources: engadget.com, brandwatch.com

About the data

How the data was collected

Source: kaggle.com

About the data

How the data was collected

- Kaggle.com user Mitchell J
- 4525 trending videos
  - between November 2017 and March 2018
- YouTube API
- All the videos were on YouTube’s ‘Trending’ page

Source: kaggle.com

About the data

What is YouTube Trending? How does a video end up on the Trending page?

Source: support.google.com

About the data

What is YouTube Trending? How does a video end up on the Trending page?

- *Trending helps viewers see what's happening on YouTube and in the world.*
- *Trending considers many signals, including (but not limited to):*
    - *View count*
    - *The rate of growth in views*
    - *Where views are coming from (including outside of YouTube)*
    - *The age of the video*
- *The Trending system tries to choose videos that will be most relevant to our viewers and most reflective of the broad content on the platform.*
- *YouTube does not favor specific creators.*

Source: support.google.com

Questions

- What factors contribute to a video’s success?
- Does anything that the content creator controls have a significant influence on a video’s success?

The Variables

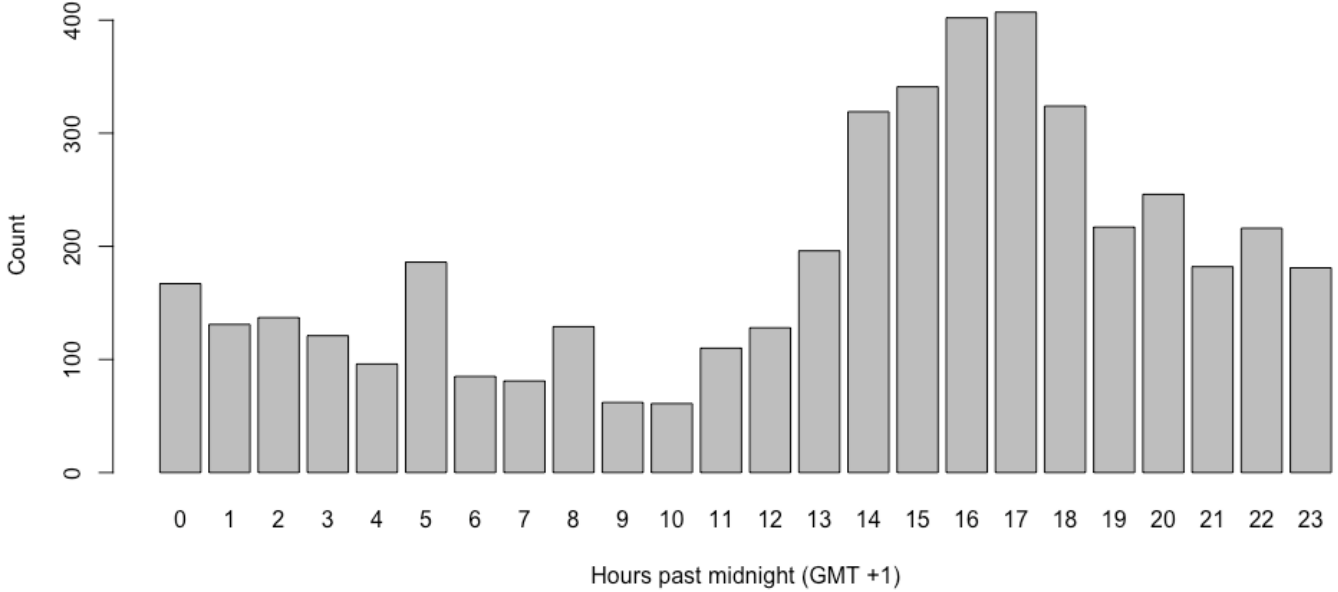
The Variables

Controllable

The Variables

Controllable

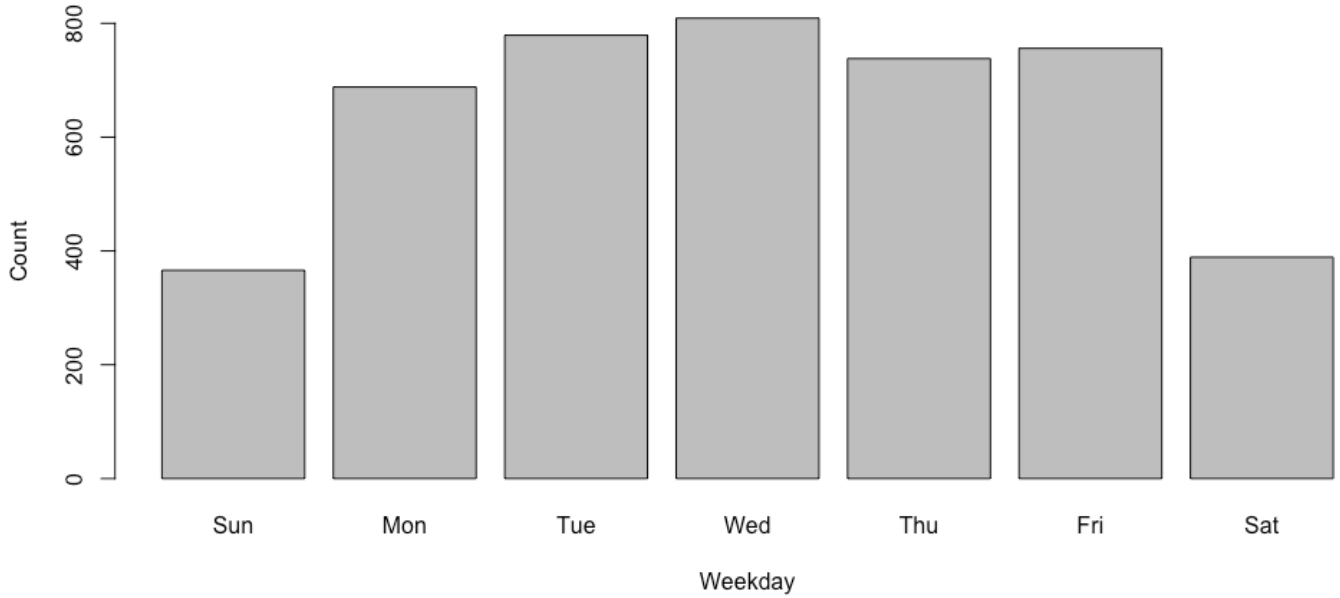
publish.hour



The Variables

Controllable

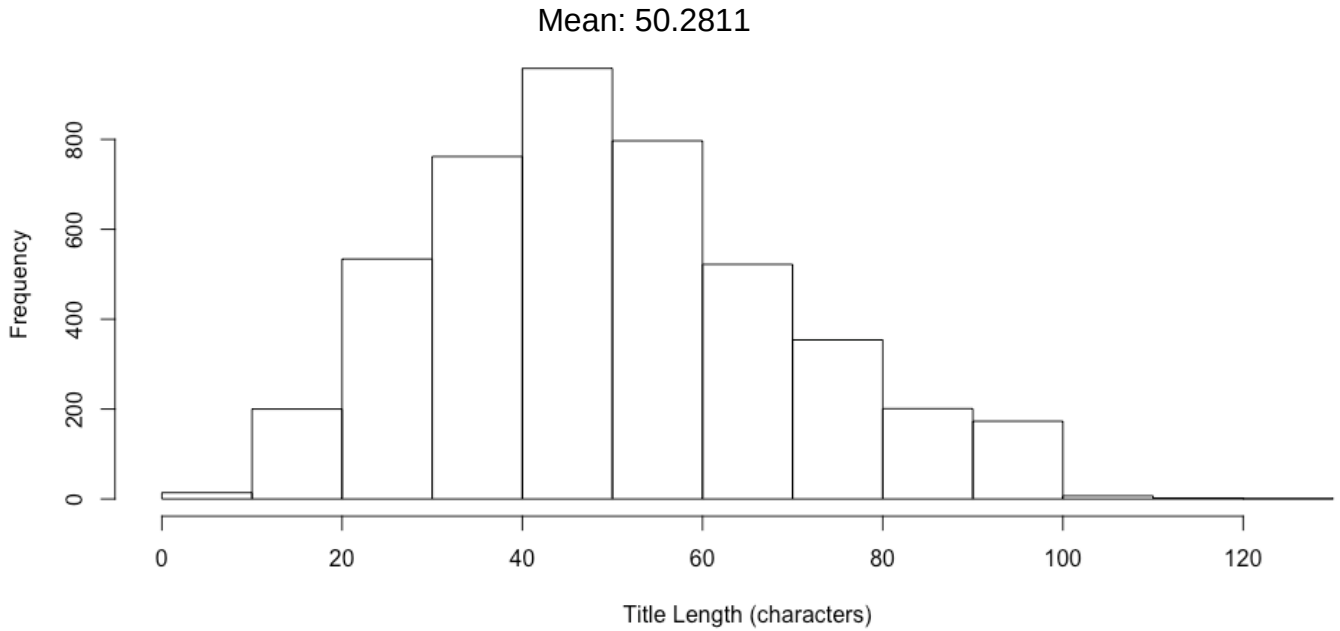
publish.hour
weekday



The Variables

Controllable

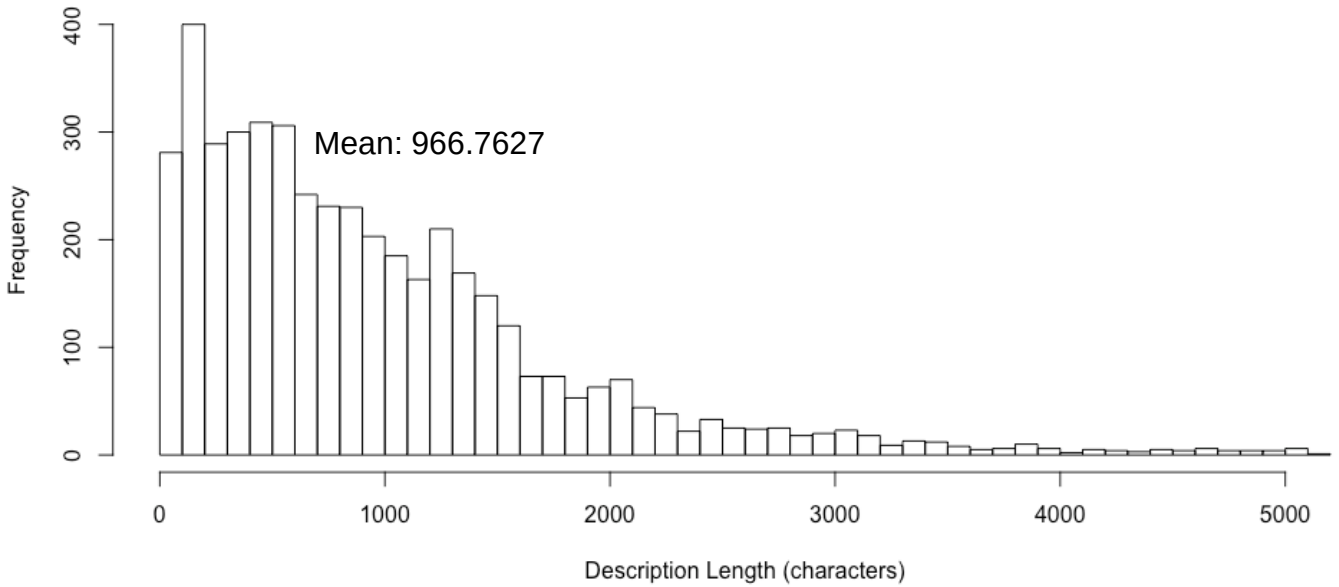
publish.hour
weekday
title.length



The Variables

Controllable

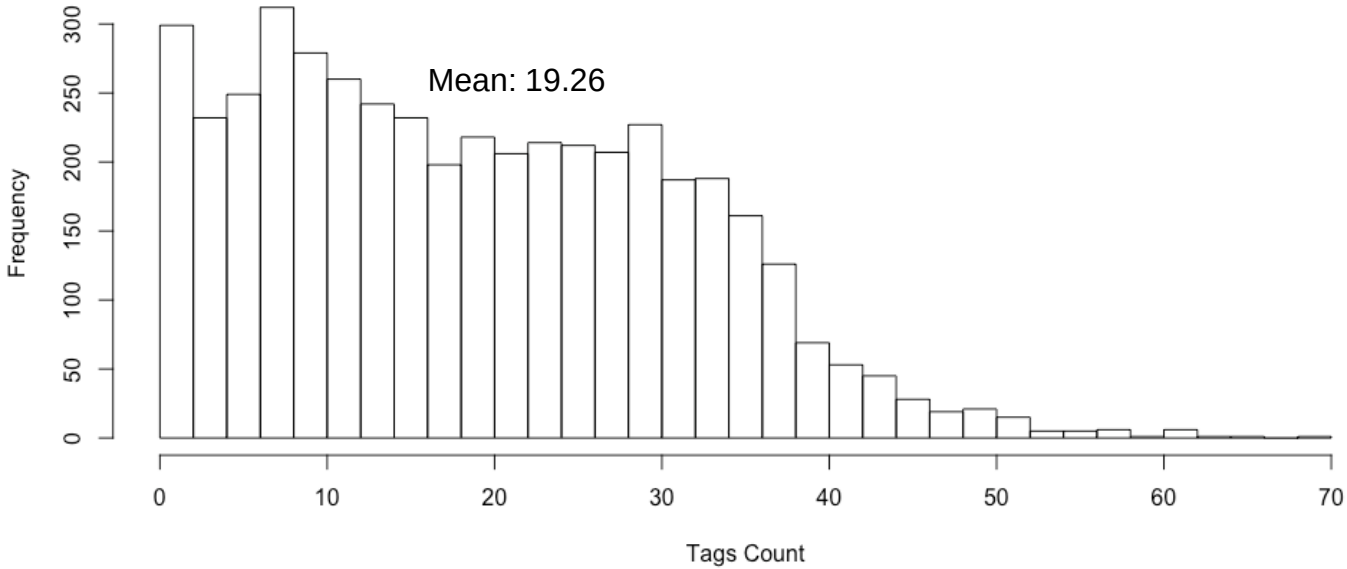
publish.hour
weekday
title.length
description.length



The Variables

Controllable

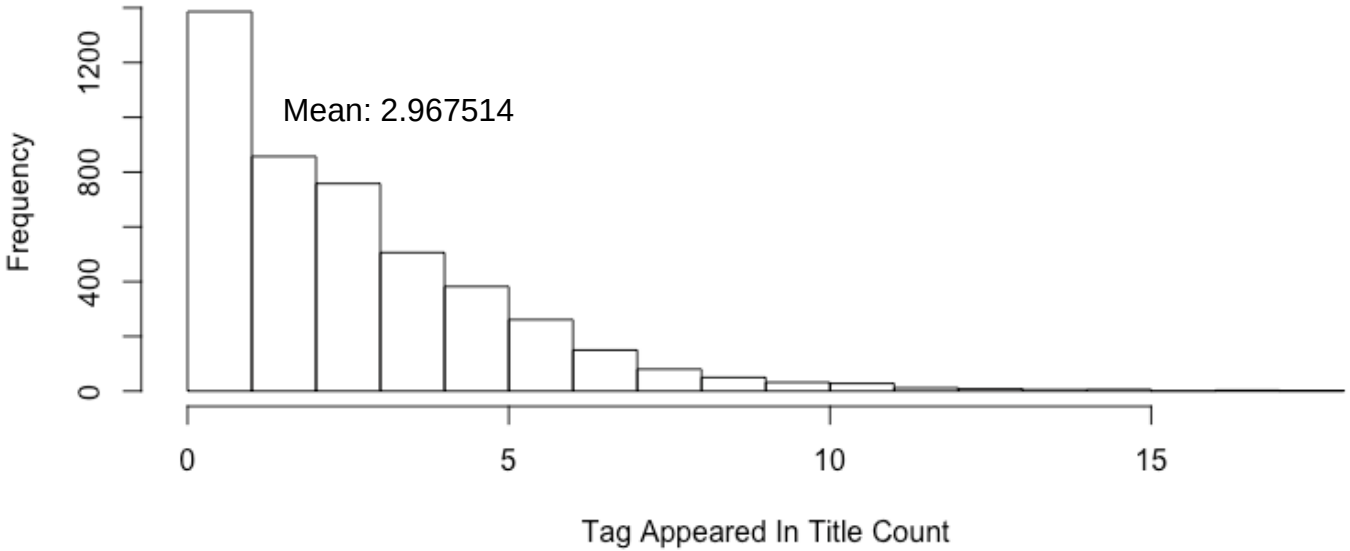
publish.hour
weekday
title.length
description.length
tags.count



The Variables

Controllable

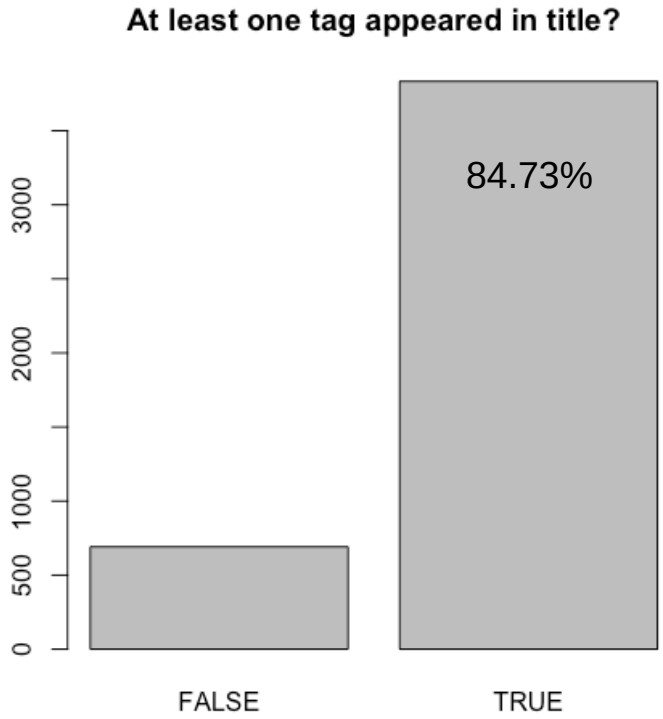
publish.hour
weekday
title.length
description.length
tags.count
tag.appeared.in.title.count



The Variables

Controllable

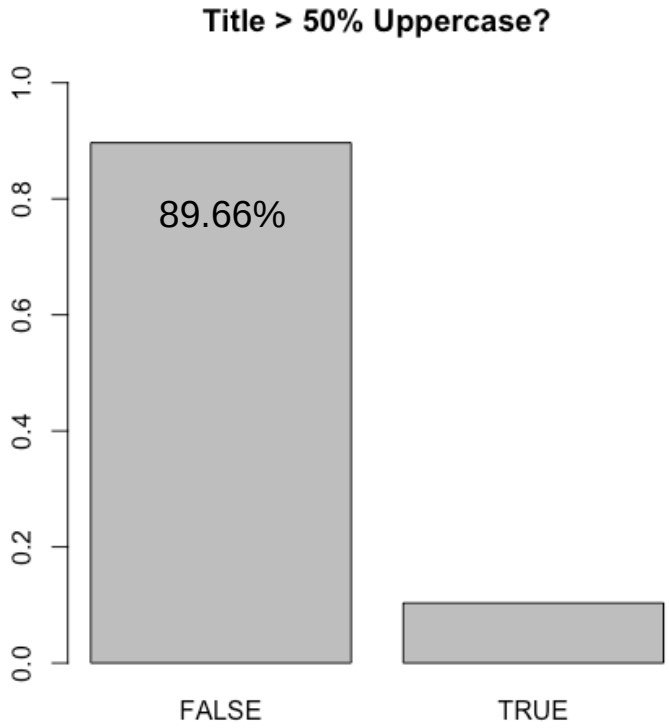
publish.hour
weekday
title.length
description.length
tags.count
tag.appeared.in.title.count
tag.appeared.in.title



The Variables

Controllable

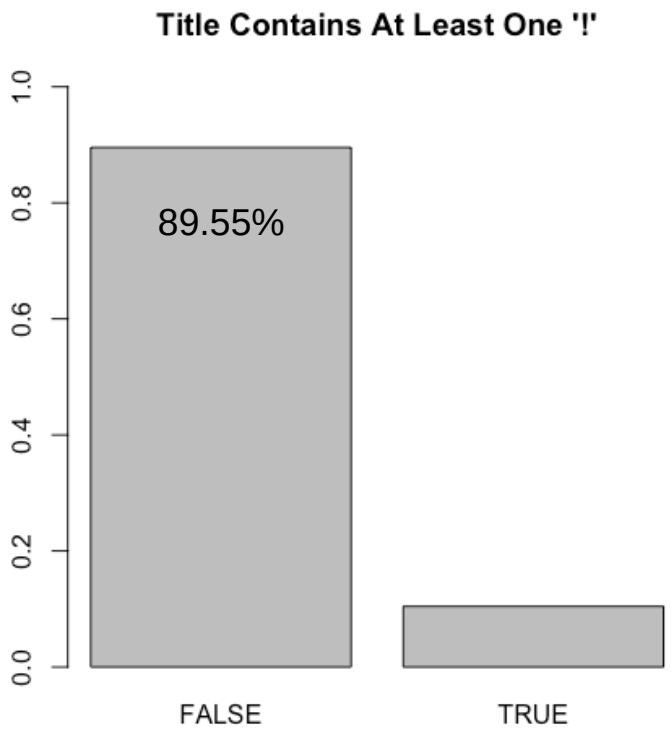
publish.hour
weekday
title.length
description.length
tags.count
tag.appeared.in.title.count
tag.appeared.in.title
caps



The Variables

Controllable

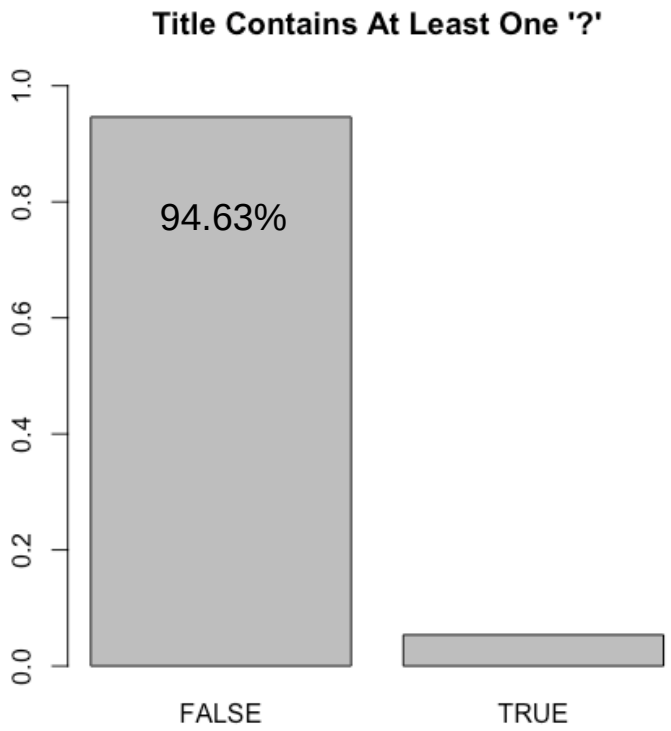
publish.hour
weekday
title.length
description.length
tags.count
tag.appeared.in.title.count
tag.appeared.in.title
caps
exclamation



The Variables

Controllable

publish.hour
weekday
title.length
description.length
tags.count
tag.appeared.in.title.count
tag.appeared.in.title
caps
exclamation
question



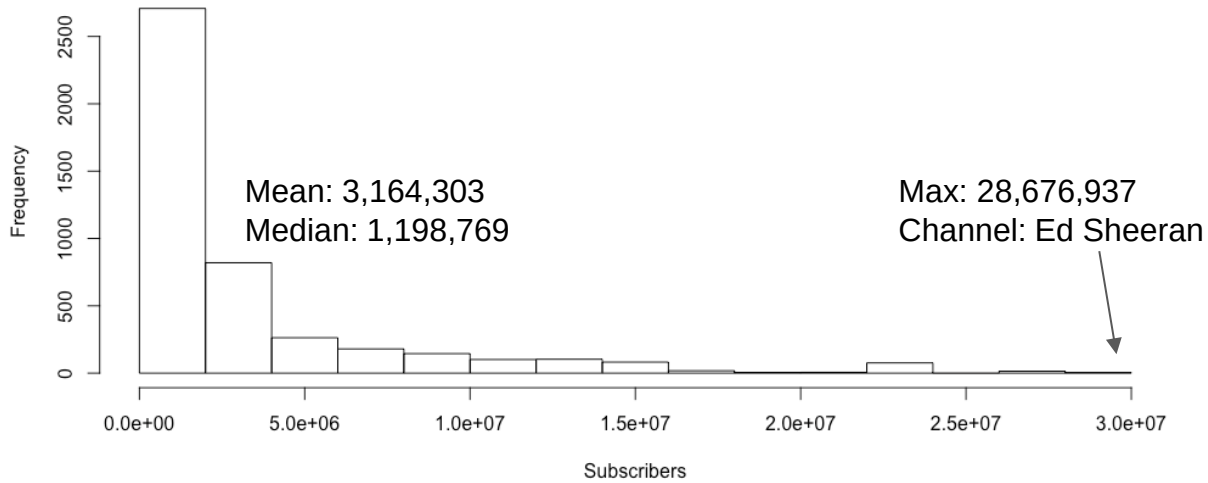
The Variables

Controllable Uncontrollable

publish.hour
weekday
title.length
description.length
tags.count
tag.appeared.in.title.count
tag.appeared.in.title
caps
exclamation
question

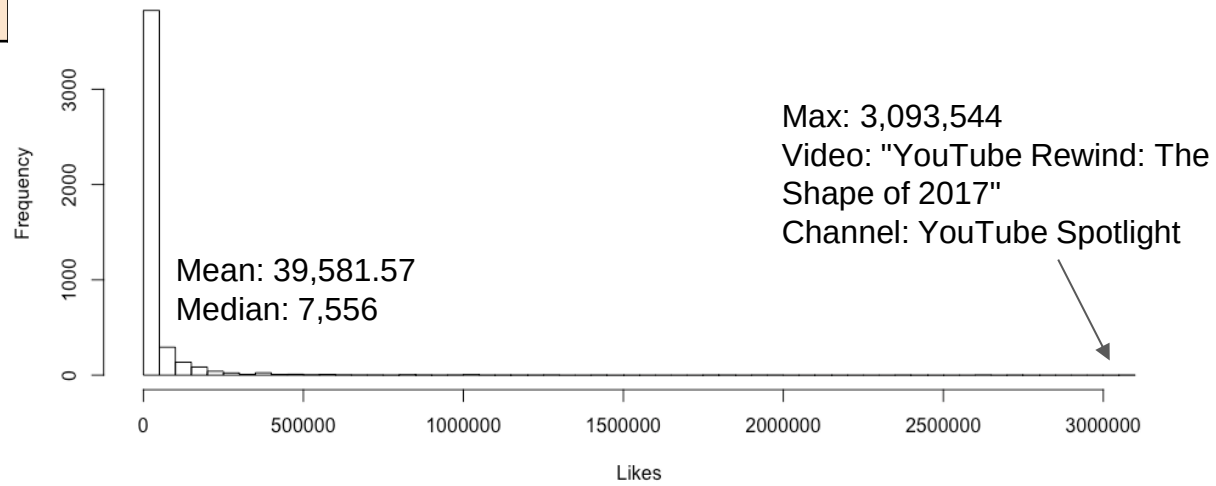
The Variables

Controllable	Uncontrollable
publish.hour	**subscribers**
weekday	
title.length	
description.length	
tags.count	
tag.appeared.in.title.count	
tag.appeared.in.title	
caps	
exclamation	
question	



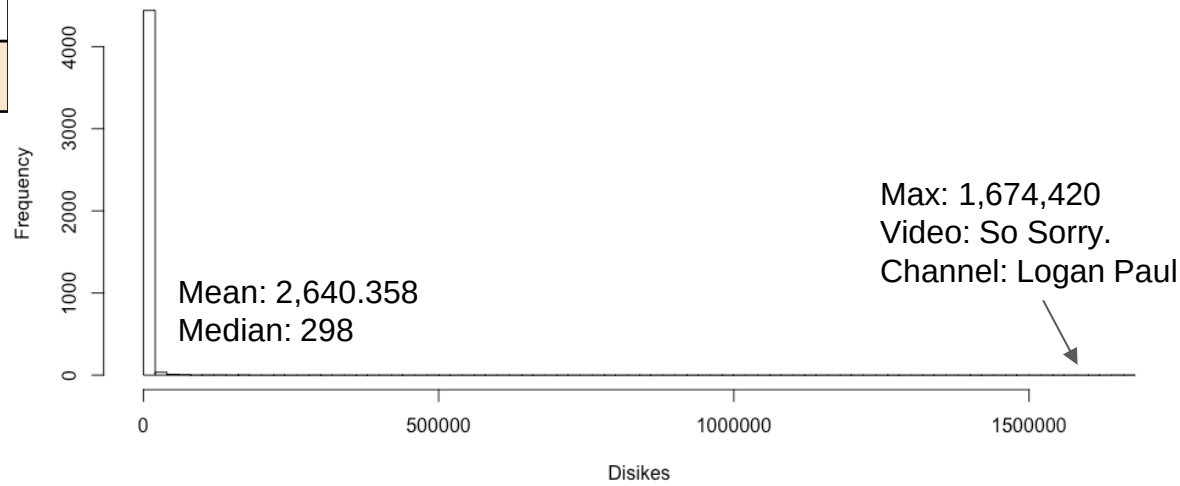
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	**likes**
title.length	
description.length	
tags.count	
tag.appeared.in.title.count	
tag.appeared.in.title	
caps	
exclamation	
question	



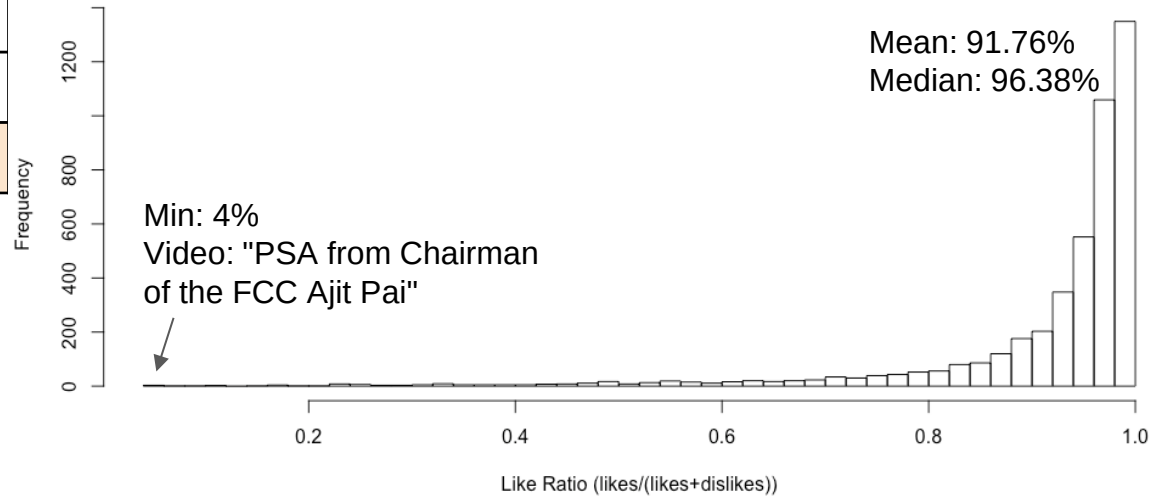
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	**dislikes**
description.length	
tags.count	
tag.appeared.in.title.count	
tag.appeared.in.title	
caps	
exclamation	
question	



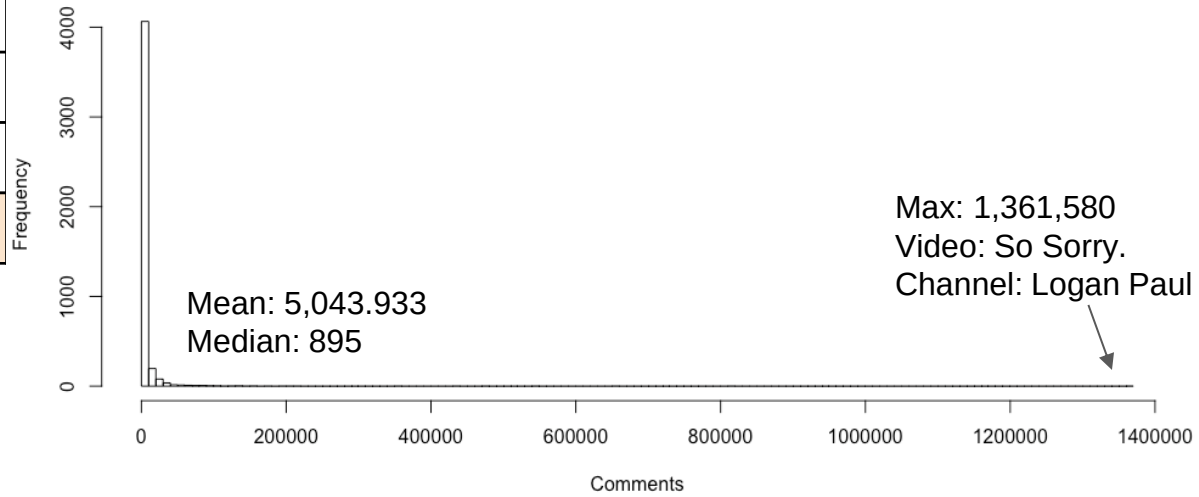
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	**like.ratio**
tags.count	
tag.appeared.in.title.count	
tag.appeared.in.title	
caps	
exclamation	
question	



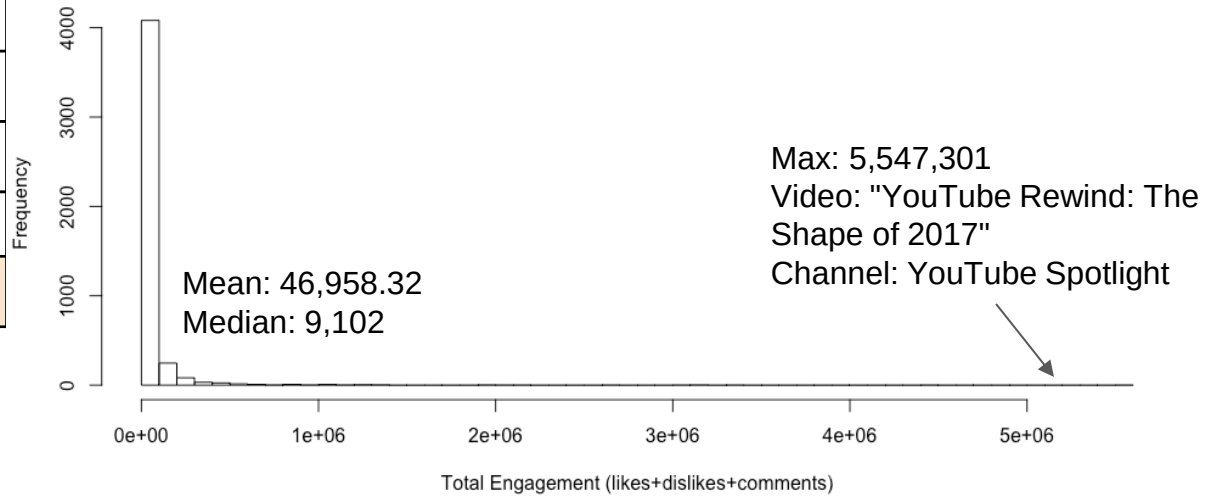
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	**comments**
tag.appeared.in.title.count	
tag.appeared.in.title	
caps	
exclamation	
question	



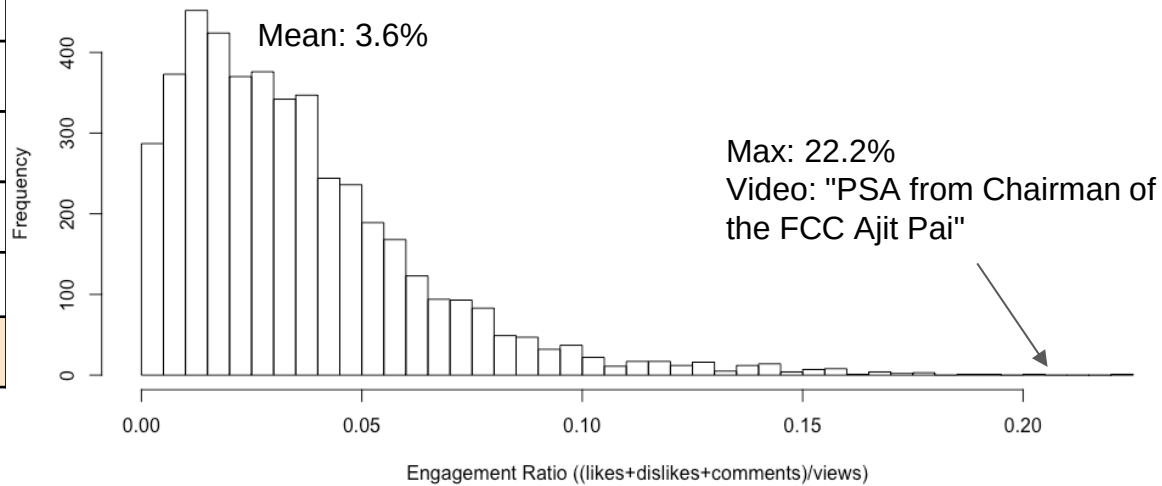
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	**total.engagement**
tag.appeared.in.title	
caps	
exclamation	
question	



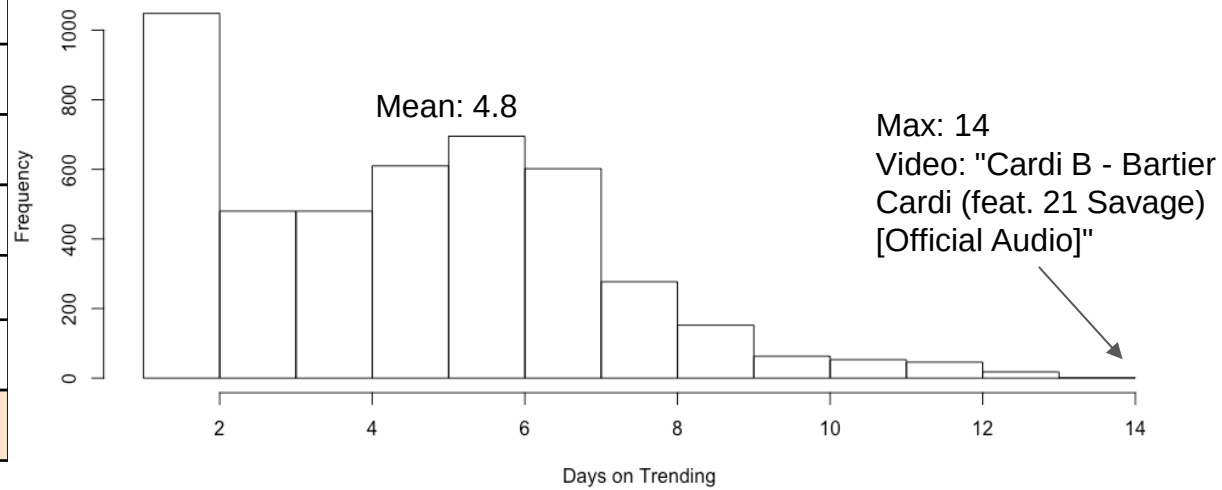
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	**engagement.ratio**
caps	
exclamation	
question	



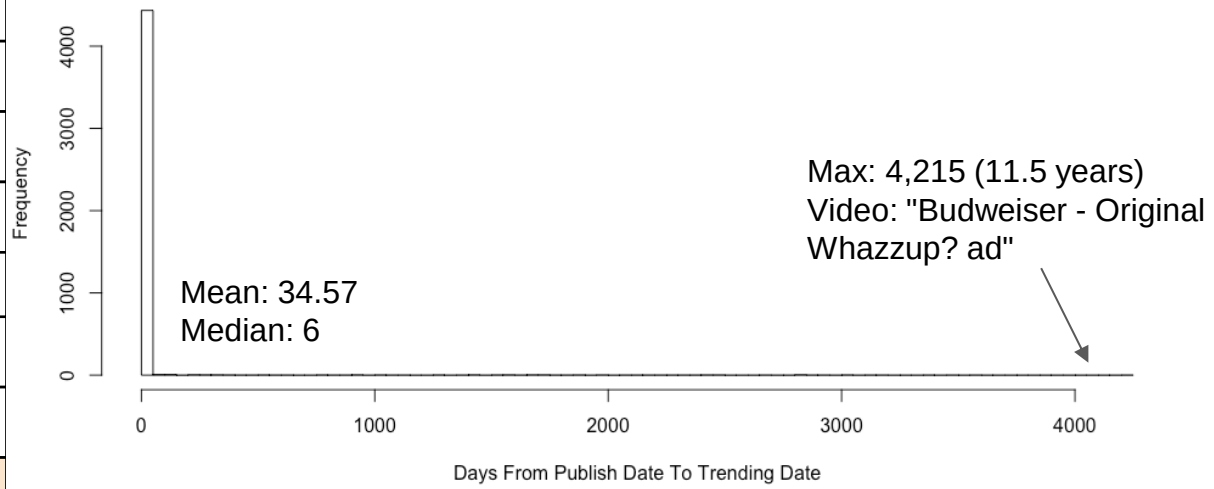
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	engagement.ratio
caps	**trend.day.count**
exclamation	
question	



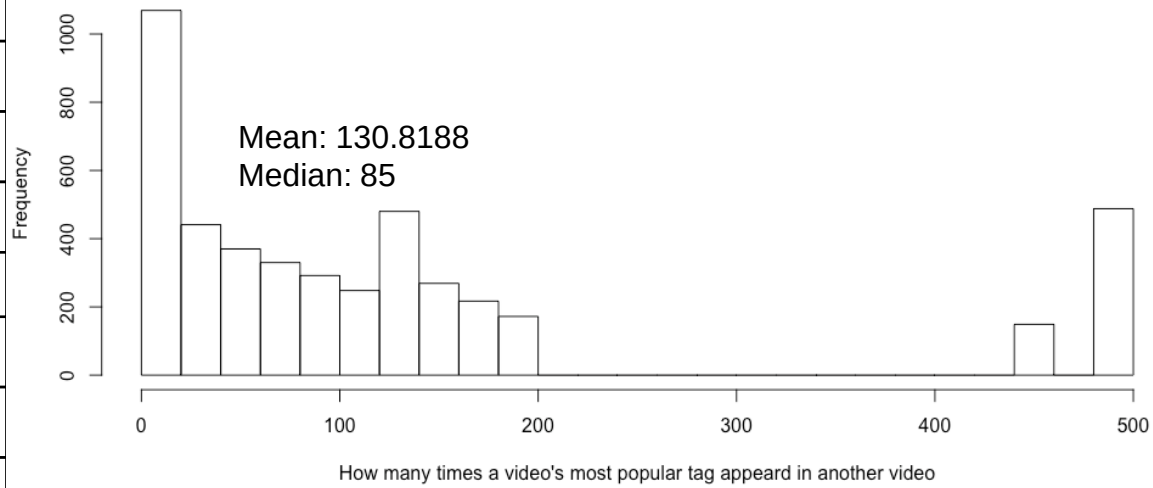
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	engagement.ratio
caps	trend.day.count
exclamation	**trend.pub.diff**
question	



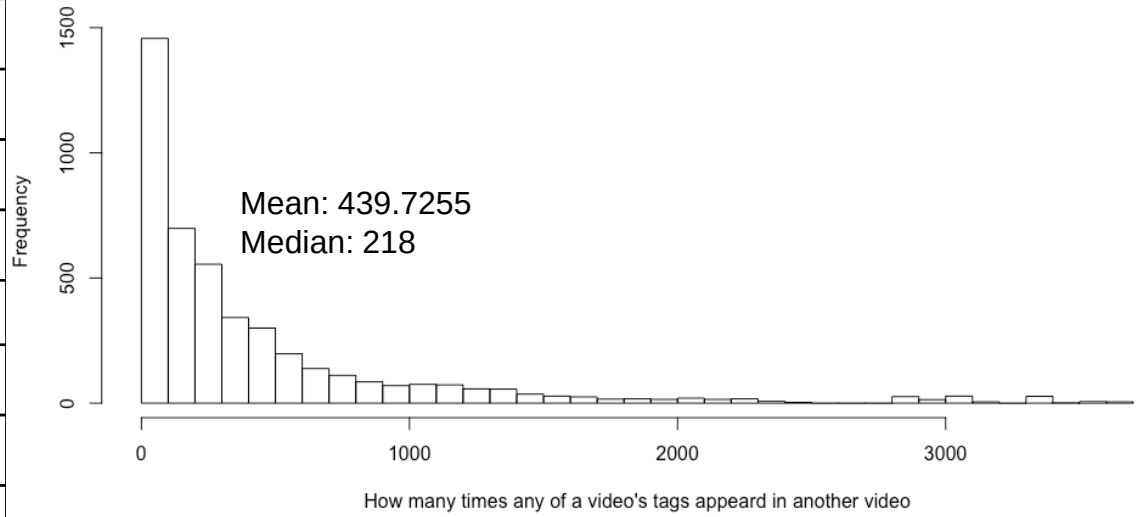
The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	engagement.ratio
caps	trend.day.count
exclamation	trend.pub.diff
question	**trend.tag.highest**



The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	engagement.ratio
caps	trend.day.count
exclamation	trend.pub.diff
question	trend.tag.highest
	trend.tag.total



The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	engagement.ratio
caps	trend.day.count
exclamation	trend.pub.diff
question	trend.tag.highest
	trend.tag.total

The Response

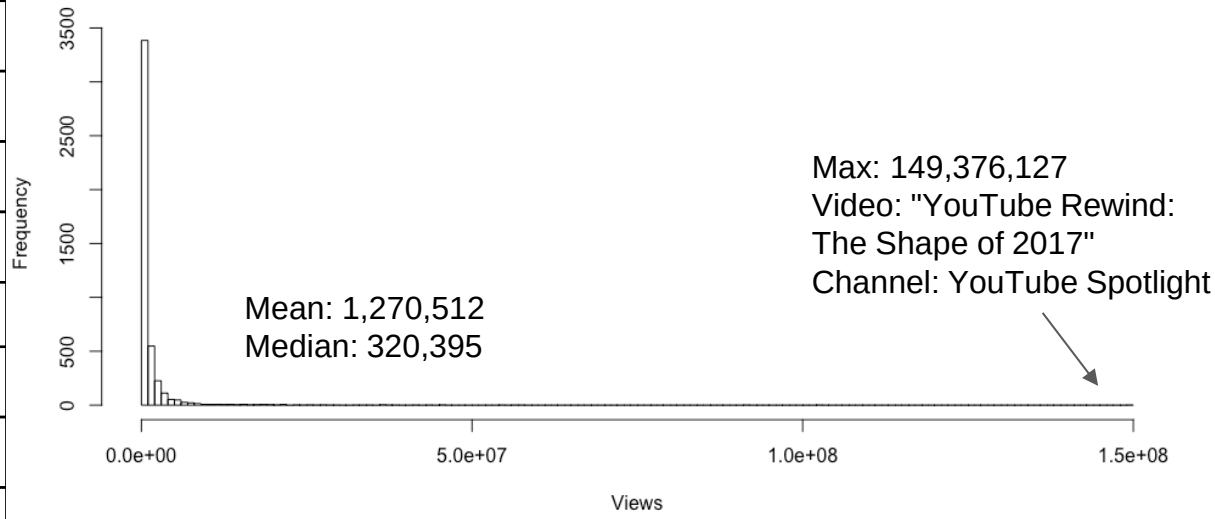
Views

The Variables

Controllable	Uncontrollable
publish.hour	subscribers
weekday	likes
title.length	dislikes
description.length	like.ratio
tags.count	comments
tag.appeared.in.title.count	total.engagement
tag.appeared.in.title	engagement.ratio
caps	trend.day.count
exclamation	trend.pub.diff
question	trend.tag.highest
	trend.tag.total

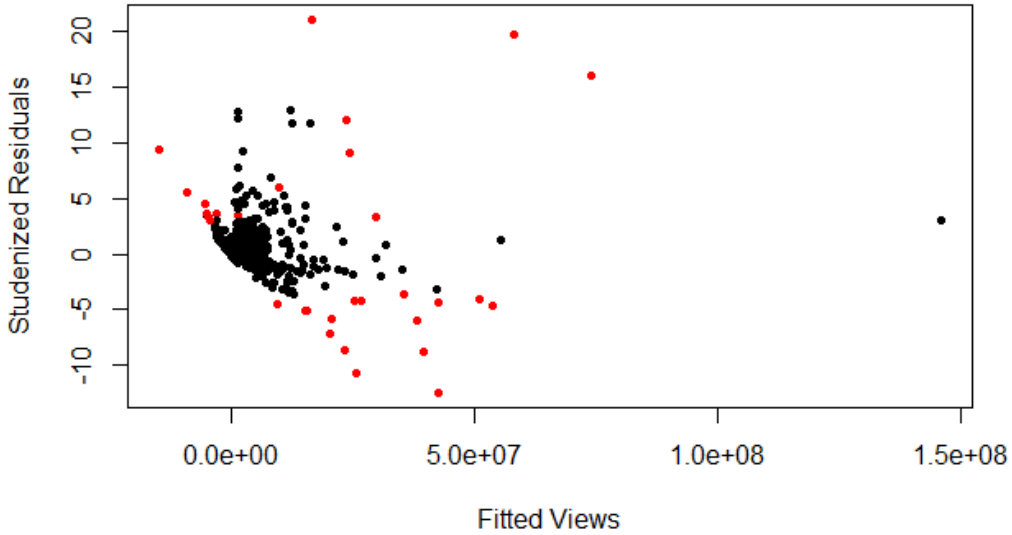
The Response

Views

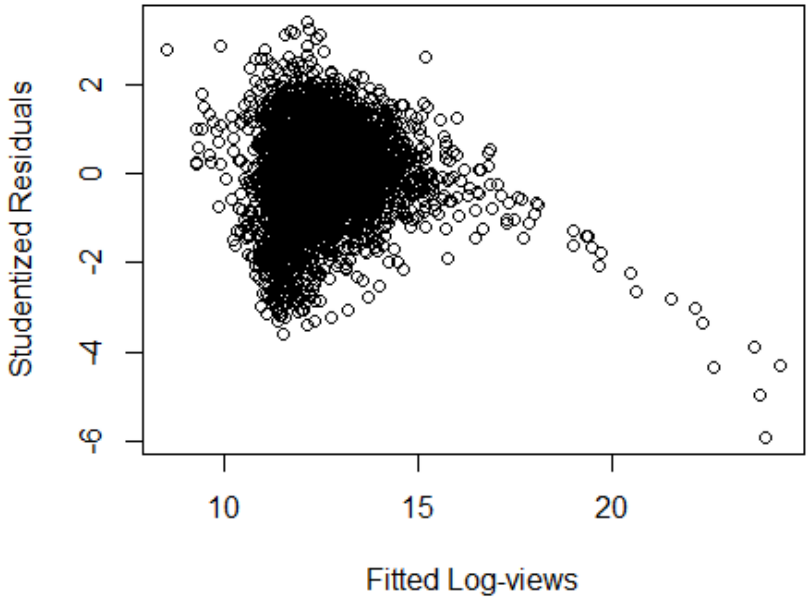


Model building

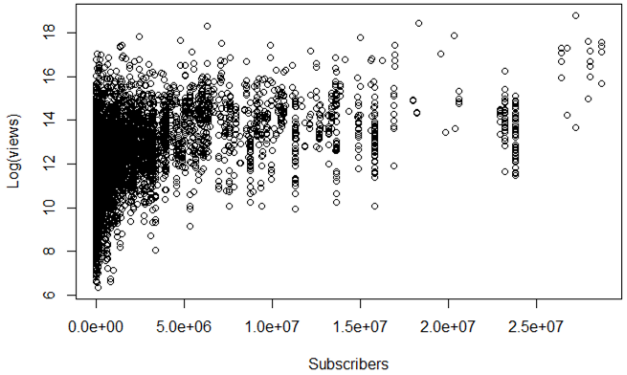
Studentized Residual Plot for untransformed dataset



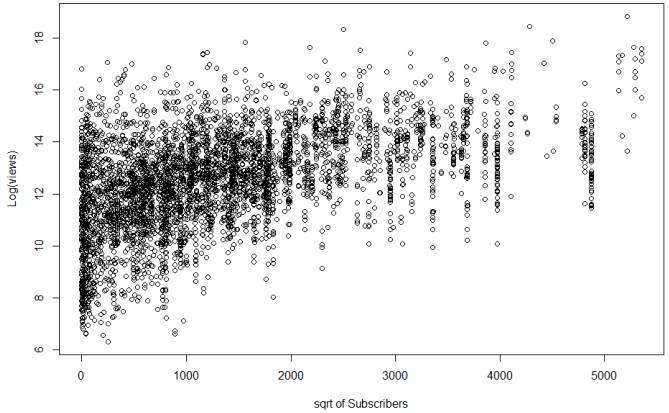
Studentized Residual Plot for log-transformed Views



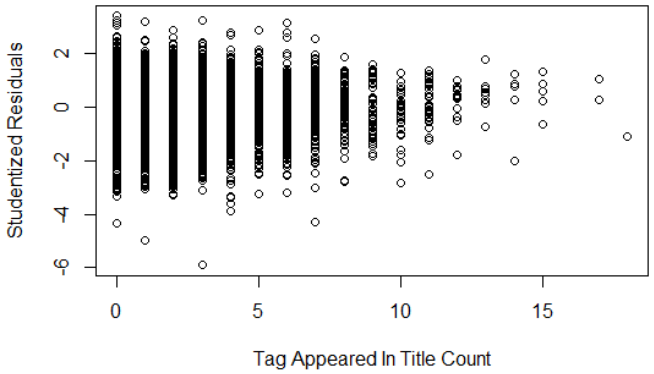
* Non-linear



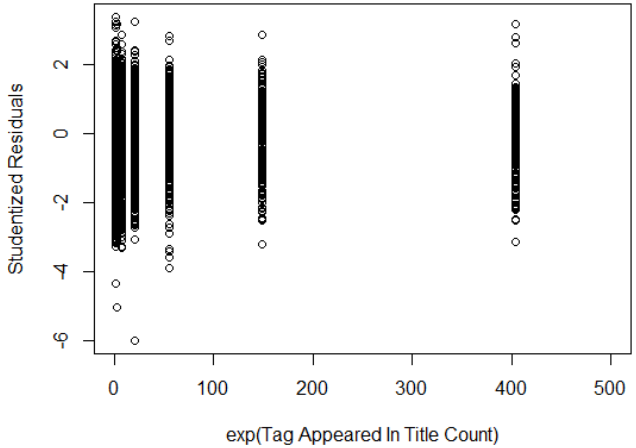
sqrt on subscribers →

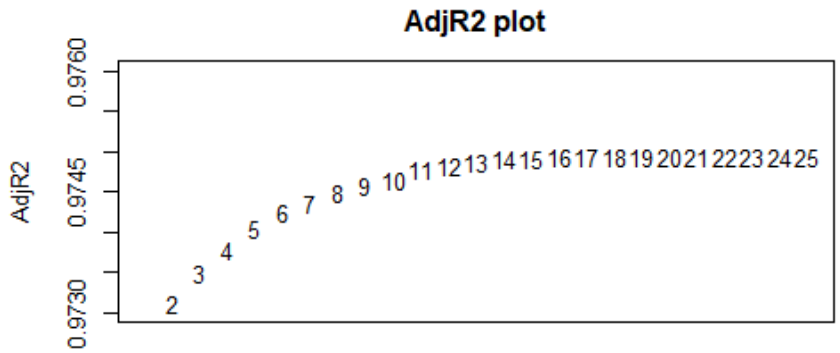


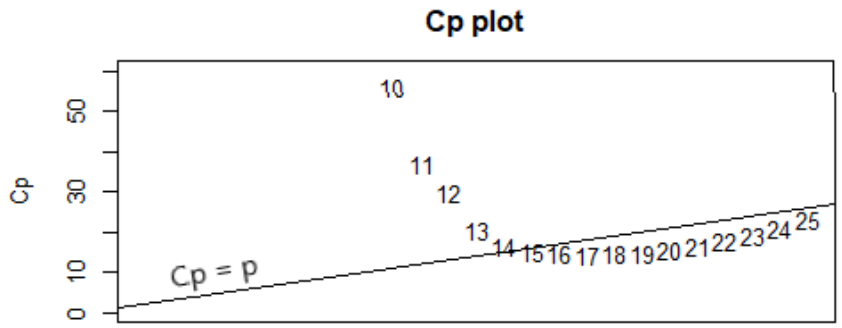
* Unconstancy

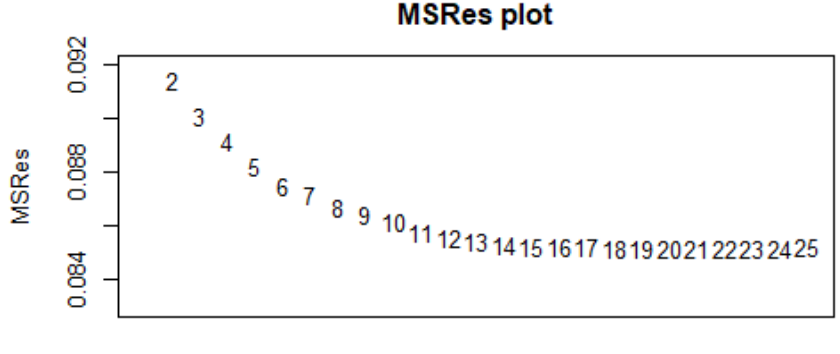


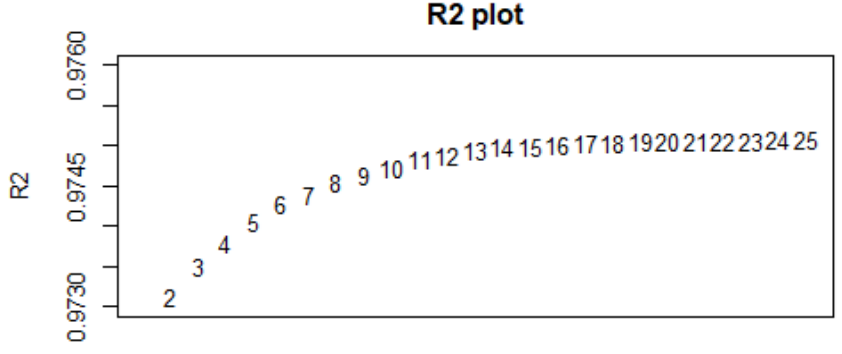
exp on tag_appeared _in_title_count →




AdjR2 plot
AdjR2
0.9730
0.9745
0.9760
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25


Cp plot
Cp
0
10
30
50
10
11
12
13
14 15 16 17 18 19 20 21 22 23 24 25
Cp = p


MSRes plot
MSRes
0.084
0.088
0.092
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25


R2 plot
R2
0.9730
0.9745
0.9760
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Subset models

k	subscribers	likes	dislikes	like ratio	comments	total engagement	engagement ratio	publish hour	title length	description length	trend publish diff	caps	trend tag highest	trend tag total	weekday	R2	AdjR2	MSRes	Cp
12	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0	0.974888828	0.974820746	0.085616608	31.7468907
12	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0.974872082	0.974803954	0.085673705	34.71109659
12	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0.974870754	0.974802622	0.085678232	34.94607207
13	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0.974952018	0.974878431	0.085420462	22.56214878
13	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0.97492197	0.974848295	0.085522933	27.88069797
13	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0.974907905	0.974834189	0.085570898	30.37019212
14	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0.974985717	0.974906558	0.085324823	18.59739589
14	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0.974970688	0.974891481	0.085376086	21.25750319
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.97497	0.974890791	0.085378432	21.37922642

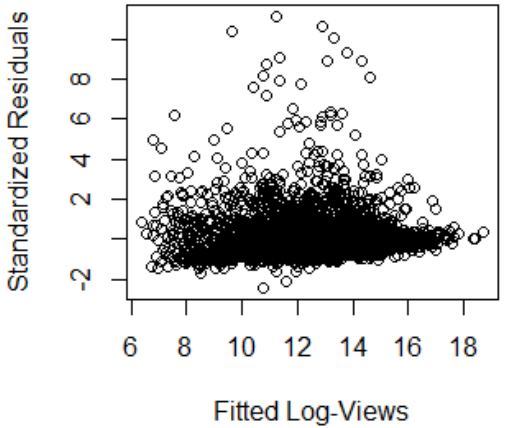
k	subscribers	likes	dislikes	like ratio	comments	total engagement	engagement ratio	publish hour	title length	description length	trend publish diff	caps	trend tag highest	trend tag total	weekday	
13	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	Forward selection (p = 0.05)
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	Backward selection (p = 0.1)

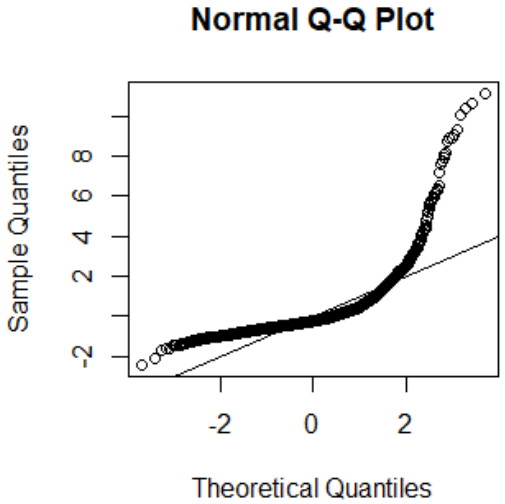
- VIF, Cp check
- t-test p-value check

- Final candidates

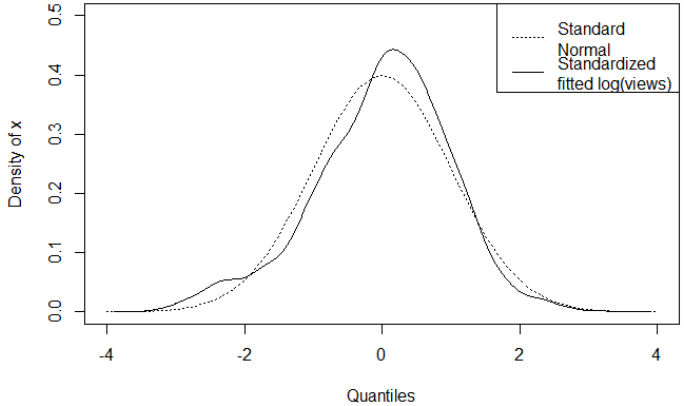
Predictors	Adjusted Rsquared	PRESS	MSRes	Cp
“subscribers”, “total_engagement”, “like_ratio”, “engagement_ratio”, “title_length”, “trend_pub_diff”, “caps”, “trend_tag_highest”, “trend_tag_total” (9 predictors)	0.9739	404.8821	0.0896	236.491
"subscribers", "total_engagement", "like_ratio", "engagement_ratio", "title_length", "trend_pub_diff", "caps", "trend_tag_highest", "trend_tag_total", "description_length" (10 predictors)	0.974	404.2452	0.0894	229.0986

Central Limit Theorem





Std Normal vs. Stdized Fitted log(views)



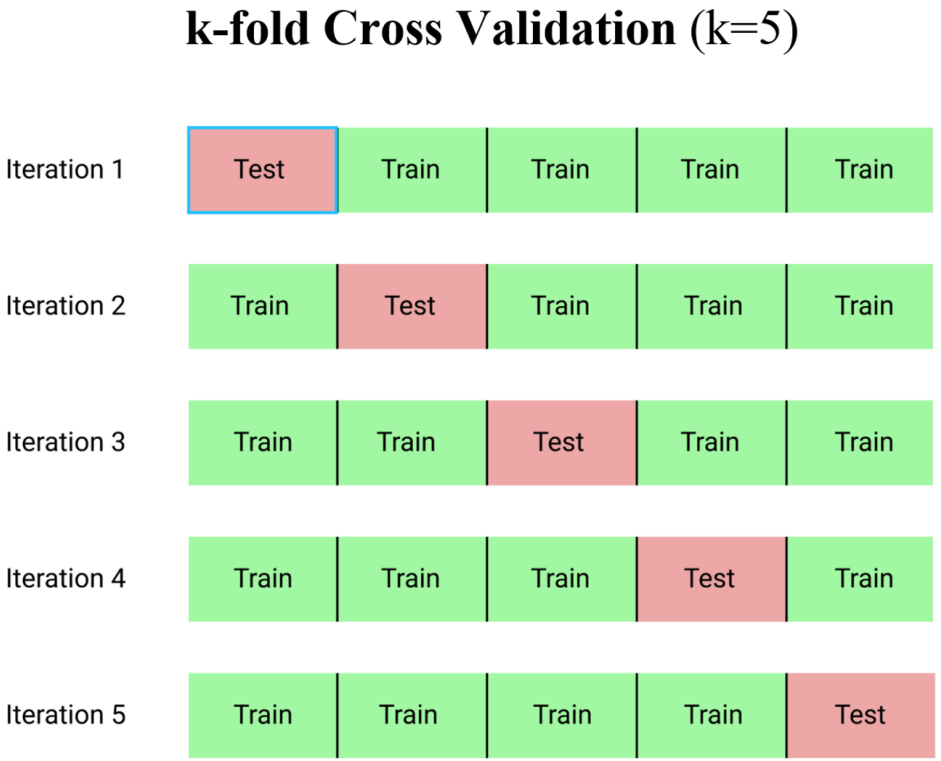
Model Validation

- The model is a valid representation of the true relationship between the predictors and the response.

- Stable estimated coefficients
- Reasonable sign
- Reasonable magnitude

k-fold Cross Validation

- Fix an integer k
  (in our case: k=10)
- Partition the whole dataset randomly into k equal-sized subsets (call folds).
- Out of k subsets, each one will retain as validation (test) dataset once and the remaining k-1 are used as estimation (train) datasets.



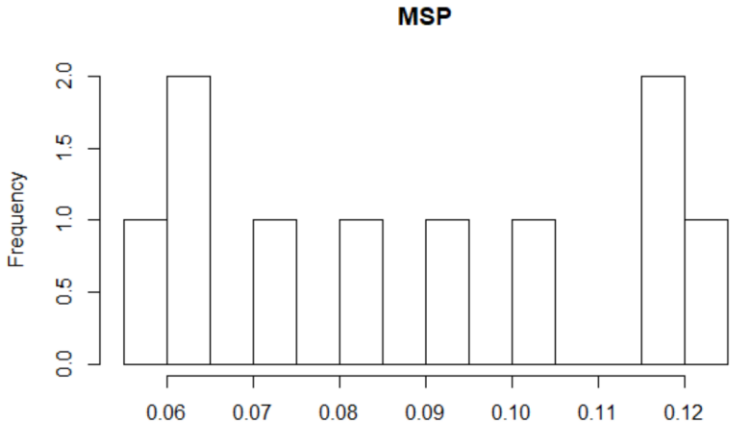
10-fold Cross Validation table

- Stable coefficients
- Mean (MSP)=0.0895 (*Our model's MSres=0.0894)*

	R^2 predict.	MSP	intercept	subscribers	like.ratio	total.engagement	engagement.ratio
1	0.978921925	0.068409232	6.161152316	-2.60E-05	-0.016289965	0.977501216	-11.99887787
2	0.955410064	0.137648917	6.143934479	-2.41E-05	-0.011194363	0.976481027	-11.86692924
3	0.977136359	0.079884089	6.14503411	-2.65E-05	-0.018616061	0.977504447	-11.98170032
4	0.972125673	0.096751401	6.125805684	-2.75E-05	-0.016487	0.97860248	-11.92082453
5	0.979111701	0.07333549	6.156568683	-2.49E-05	-0.018372489	0.97683507	-11.95910176
6	0.978178301	0.076773986	6.141665949	-2.87E-05	-0.013648478	0.979996929	-12.07820742
7	0.977613964	0.08446562	6.141675977	-2.78E-05	-0.02006612	0.977623028	-11.98463913
8	0.959274786	0.139520641	6.126507404	-2.75E-05	-0.017213111	0.976758979	-11.9458183
9	0.976974218	0.081214025	6.144755151	-2.52E-05	-0.019395615	0.976324902	-11.93543496
10	0.983588018	0.057421051	6.169986208	-2.66E-05	-0.001848395	0.977643674	-12.06432496

	title.length	description.length	trend.pub.diff	caps	trend_tag_highest	trend_tag_total
1	-0.001491293	-0.001148606	0.000106464	0.103002396	0.000147119	-0.015457969
2	-0.001290602	-0.000942835	0.000125231	0.087826373	0.000157649	-0.018876617
3	-0.001174336	-0.0011451	0.00011286	0.104505787	0.00015721	-0.016872487
4	-0.001343048	-0.001210494	0.000126782	0.097762626	0.000118484	-0.013753913
5	-0.001416802	-0.00124813	0.000121756	0.098633031	0.00014062	-0.015523488
6	-0.001401459	-0.001224772	0.000126464	0.091612512	0.000119742	-0.012574676
7	-0.001357106	-0.001138206	0.000115491	0.104102582	0.000146414	-0.013940076
8	-0.001254389	-0.001008324	0.000124743	0.103842984	0.000124025	-0.012355707
9	-0.00128918	-0.001026236	0.000143351	0.101970936	0.000138957	-0.015624595
10	-0.00130183	-0.00120676	0.000114677	0.109408418	0.000135549	-0.015922971

MSP barplot



Variation reason in MSP:

- Including influential points in the estimation or validation dataset in each iteration or not.

Mean (MSP)=0.0895
Model's MSres=0.0894

Comparison between the original coefficients and mean estimates

Regressor	intercept	subscribers	like_ratio	total_engagement	engagement_ratio	title_length	description_length	trend_publish_diff	caps	trend_tag_highest	trend_tag_total
Our model	6.145475	-0.00002648816	-0.01553134	0.9775194	-11.97265	-0.0013324007	-0.001128420	0.00012160 57	0.1002788	0.0001385073	-0.01507096
Mean of estimates	6.145708596	-0.0000265	-0.01531316	0.9775271 75	-11.97358585	-0.0013320005	-0.001129946	0.00012178 2	0.1002667 65	0.000138577	-0.01509025
Difference ratio	0.000038	0.000447	0.0142	0.000008	0.0000 7816	0.000302	0.00135	0.0014	0.000 12	0.000503	0.00128

- Max difference = 1.4% for *like_ratio*

Conclusion

Final Model:

log(views) = 6.145475 - 0.00002648816*sqrt(subscribers)

- 0.01553134*(like_ratio^3) + 0.9775194*log(total_engagement + 1)

- 11.97265*sqrt(engagement_ratio) - 0.001332407*(title_length)

- 0.001128420*sqrt(description_length) + 0.0001216057*(trend_publish_diff)

+ 0.1002788*(caps) + 0.0001385073*(trend_tag_highest)

- 0.01507096*log(trend_tag_total + 1)

5th observation in the dataset

subscribers	20,563,106
like_ratio	0.9845
total_engagement	176,384
engagement_ratio	0.0626
title_length	636
description_length	24
trend_publish_diff	7
caps	TRUE
trend_tag_highest	488
trend_tag_total	1,007

- 95% Confidence Interval for the Average number of Views:

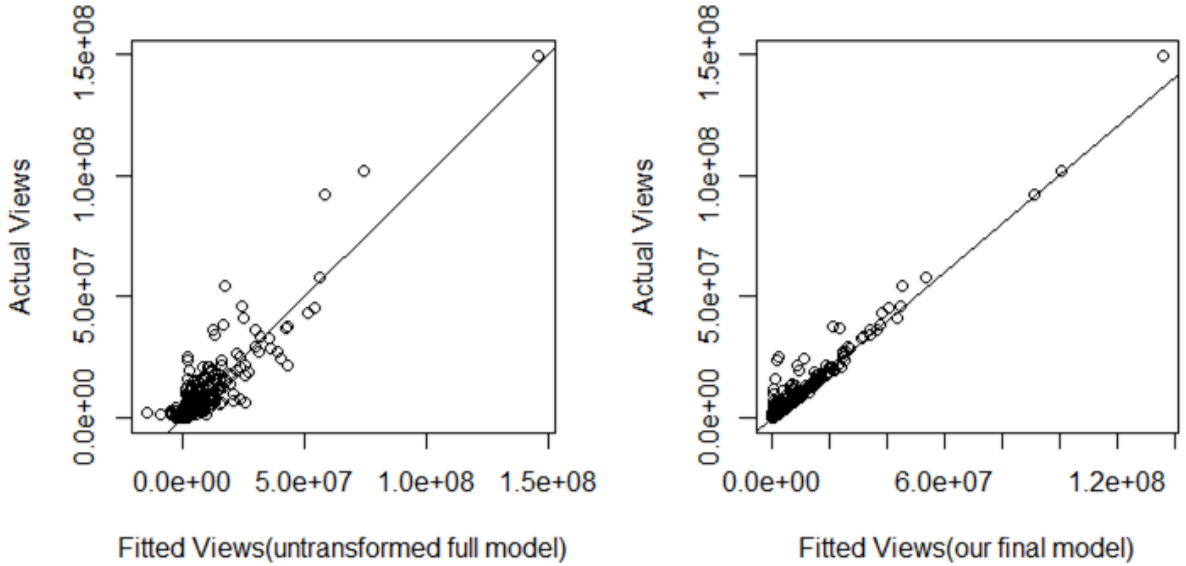
  (2,681,744 , 2,913,608)

- y = 2,819,118 views
- y_hat = 2,795,273 views

	5th observation	New observation	Difference in log(views)	Average change in views
title_length	24	14	0.01332407	0.01341323
description_length	636	536	0.002332858	0.002335581
total_engagement	176,384	177,384	0.005526314	0.005541612
caps	TRUE	FALSE	-0.1002788	0.09541482

95% prediction interval for the true coefficient of caps:
(0.07032247,0.13023513)

Comparison between the Raw Full Model and the Final Fitted Model



- root of MSRes(full model) = 1,882,863 views
- root of MSRes(fitted model) = 889,123 views

References

Montgomery, D. C., Peck, E. A., Vining, G. G. (2013), *Introduction to Linear Regression Analysis*, Hoboken, NJ: John Wiley & Sons, Inc.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Mitchell J (2018), “Trending YouTube Video Statistics: Daily statistics for trending YouTube videos”, Kaggle, Available at https://www.kaggle.com/datasnaek/youtube-new/home

Mr.SG (2018), “[Youtube Trends] About the Dataset (full details)”, Kaggle, Available at https://www.kaggle.com/sgonkaggle/youtube-trend-with-subscriber/discussion/57391

Brownlee, J. (2018), “A Gentle Introduction to k-fold Cross-Validation”, Machine Learning Mastery, Available at https://machinelearningmastery.com/k-fold-cross-validation/

Bartlett, J. (2013), “Assumptions for linear regression”, The Stats Geek, Available at http://thestatsgeek.com/2013/08/07/assumptions-for-linear-regression/

Ace X, (2016), “The History of YouTube”, Engadget, Available at https://www.engadget.com/2016/11/10/the-history-of-youtube/