TED Talks

Angela Zhai

December 9, 2018

I. Abstract

TED.com has a bunch of free and compelling videos. Most of them compress speakers' years of research into 18 minutes story-telling talks. Those videos attract many people to watch and learn. However, why, some videos are more popular than others.

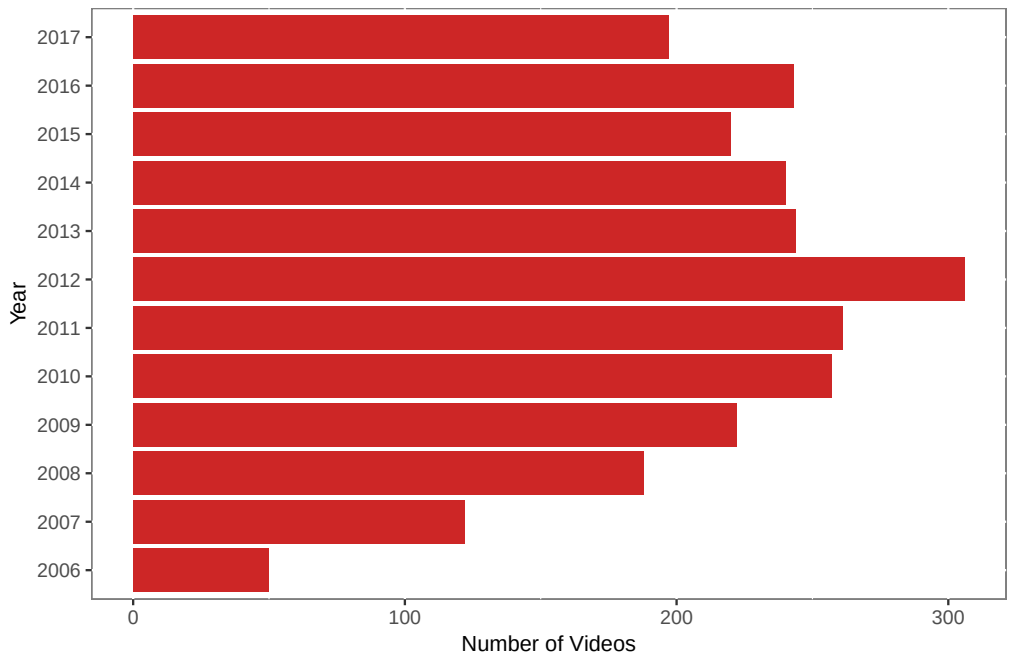
By conducting EDA and varying intercepts mixed effect model, the overall trend has shown. With more subtitle languages choice, more audiences who speak different languages could have the chance to get the speaker's ideas. Also, if there are lots of people willing to share their opinions under a video, usually more people will watch this video and join the discussion. More importantly, the talks should not be too short.

II. Background

TED is a non-profit organization offered conference and free online short videos to a global audience. They began in 1984 and focused on Technology, Entertainment, and Design. Nowadays, they cover more topics and allow free licenses to local organizers who wished to organize their own TED-like events.

TED started to post the talk videos from 2006 under their slogan "Ideas worth spreading," and they found out this is a better way to let more people hear the speaker's voice. Until September 2017, they have 2550 videos on their website (https://www.ted.com/). The videos are translated into different languages by volunteers.

Fig 1. Number of Videos Published in Each Year



"It used to be 800 people getting together once a year; now it's about a million people a day watching TED Talks online". The number of viewers per day is impressive. I would like to figure out what makes those videos popular and why some videos are more popular than others.

III. Method

1. Data source

The data was scraped from TED official Website by Rounak Banik and uploaded to Kaggle (https://www.kaggle.com/rounakbanik/ted-talks). It contains information for all videos on TED.com until September 21st, 2017.

TED.com not only contains TED Talks videos but also has a bunch of videos under "Best of the Web" and "TED Dialogue." Since the project only focused on TED Talks popularity, other videos are filtered out from the dataset and get 2437 videos in total.

TED started to offer free online videos since June 2006 and received positive responses, which let them decide to make an official website for people to watch TED videos online. There are only a few videos filmed and published before 2006. But now, 5-7 new talks will be published every week.

Fig 1.1. Histogram of Views (on log scale)

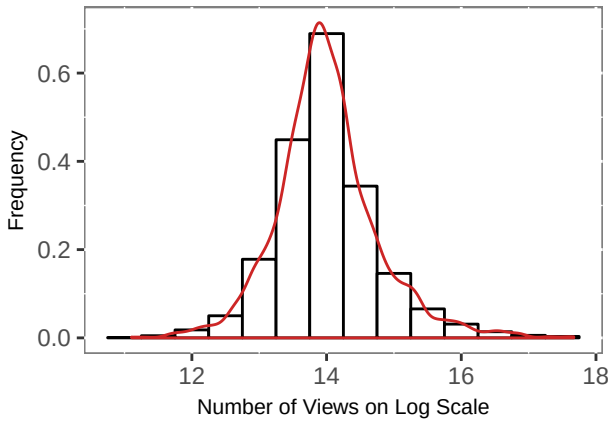


Fig 1.2. Histogram of Comments (log scale)

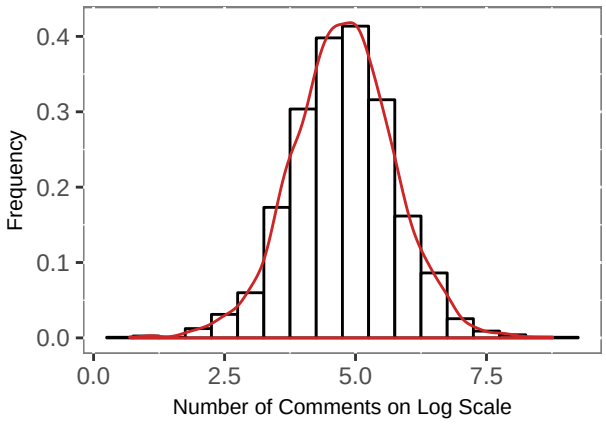


Fig 1.3. Histogram of Subtitle Languages

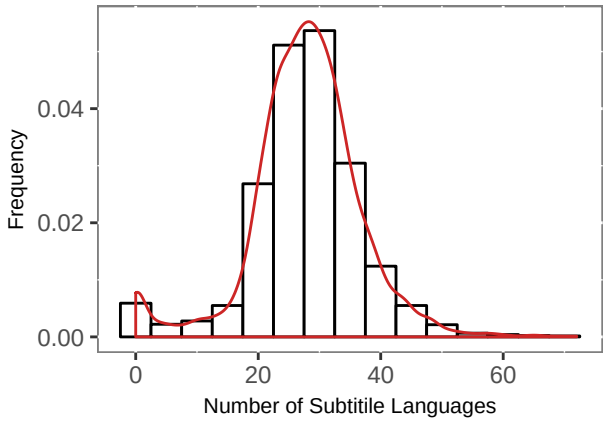


Fig 1.4. Histogram of Videos Duration (minutes)

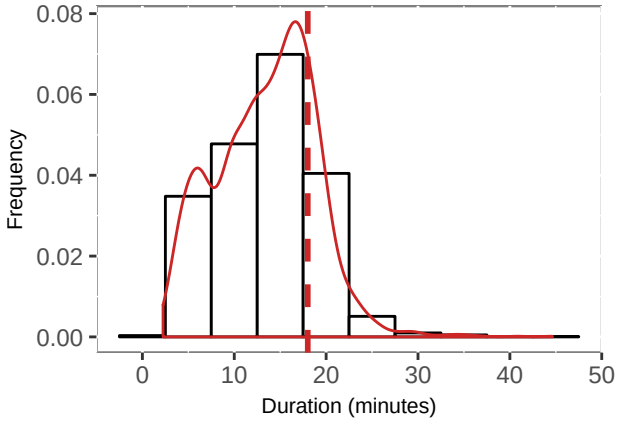


Fig 1.1-1.3 show normal distribution pattern for the number of views and number of comments (both on log scale). The number of subtitle languages also follow normal distribution pattern.

"18 minutes is long enough to be serious and short enough to hold people's attention". This is one of the reasons why TED is popular since audiences could take only a piece of time to watch the videos. Fig 1.4 shows most videos are less than 18 minutes.

Fig 2.1. Relationship between Number of Views and Comments (both on log scale)

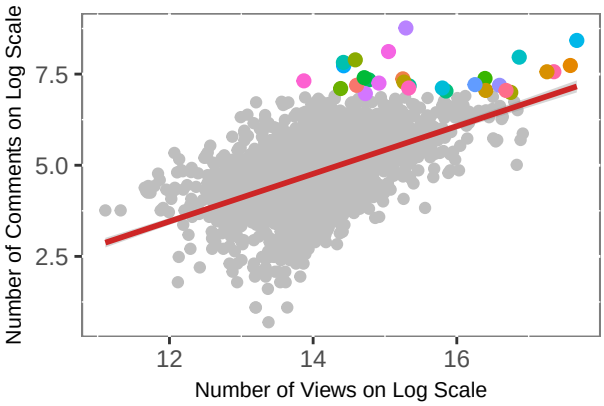


Fig 2.2. Relationship between Number of Languages and Comments (on log scale)

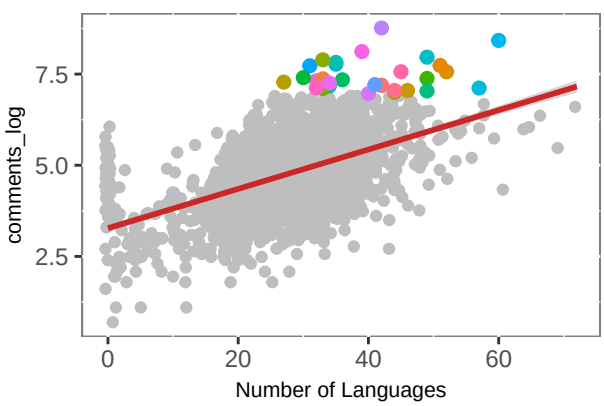


Fig 2.3. Relationship between Videos Duration and Number of Views (on log scale)

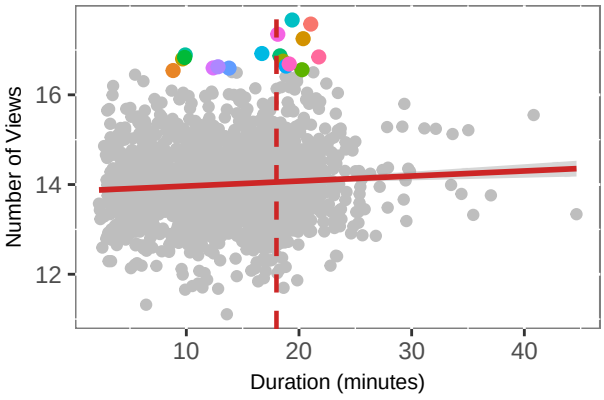


Fig 2.4. Relationship between Number of Languages and Number of Views (on log scale)

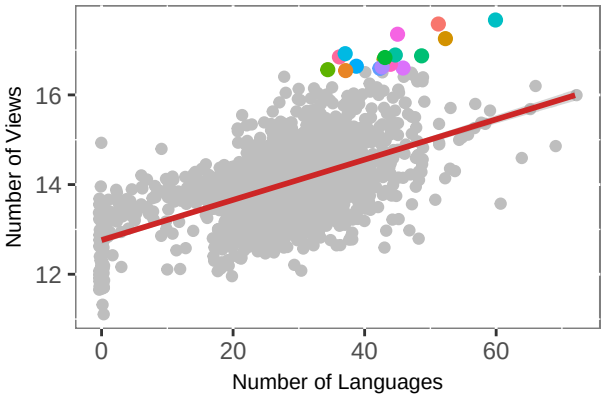


Fig 2.1 shows the relationship between the number of views and number of comments (both on log scale). Colored points in Fig 2.1-2.2 are videos with more than 1000 comments, and only 31 videos have that much comments. Obviously and naturally, the number of views has a positive relationship with the number of comments. Subtitle languages is also an essential factor for a video to become popular, since more people could understand the ideas and tend to discuss their own opinions.

The vertical dashed red line in Fig 2.3 is 18 minutes, which is TED's golden time for a speech. Colored points in Fig 2.3-2.4 are most viewed videos (viewed more than 15 million times). Those 18 videos all take around 20 minutes or less to finish, and they have translated into many languages.

2. Model used

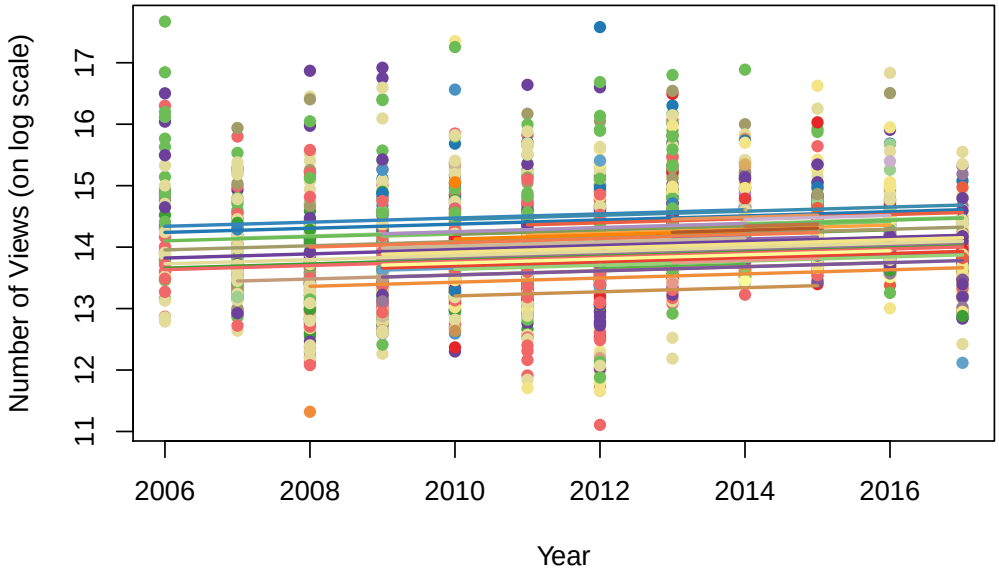
The models used for this data are linear regression model and multilevel model.

Predictors are the number of languages translated, number of comments on log scale, duration in minutes, and the year published. Outcome is the number of views on log scale.

For multilevel model, each video has a tag which is one of the tags for this video and also the most frequent tagged label through all of the videos. Those videos are grouped by tags.

Fig 4 shows the result of correlation coefficient for each tag, and there are differences among different tags.

Fig 3. Linear Regression group by Tags



IV. Result

1. Model choice

After comparing different models from no pooling to complete pooling, the model I choose is varying intercepts mixed effect model. The group-level standard deviation is not too big, and the model is close to complete pooling.

Predictors are languages(number of subtitle languages), comments_log(number of comments on log scale), duration_min(videos duration in minutes), pyear_num(published year). Outcome is views_log(number of views on log scale).

From Fig 2.2 we can tell that interaction between the number of subtitle languages and number of comments should also treat as a predictor.

Group videos into different tags. Published years are ranked from 1 to 12 (2006-2017).

```
## lmer(formula = views_log ~ languages * comments_log + duration_min +
##     factor(pyear_num) + (1 | tags), data = duration)
##                         coef.est coef.se
## (Intercept)             12.15     0.13
## languages               -0.01     0.00
## comments_log             0.06     0.03
## duration_min             0.02     0.00
## factor(pyear_num)1       0.02     0.06
## factor(pyear_num)2      -0.06     0.04
## factor(pyear_num)3      -0.29     0.03
## factor(pyear_num)4      -0.27     0.03
## factor(pyear_num)5      -0.53     0.03
## factor(pyear_num)6      -0.48     0.03
## factor(pyear_num)7      -0.21     0.03
## factor(pyear_num)8      -0.07     0.03
## factor(pyear_num)9       0.25     0.03
```

```
## factor(pyear_num)10     0.41     0.03
## factor(pyear_num)11     0.52     0.03
## languages:comments_log  0.01     0.00
##
## Error terms:
##  Groups   Name        Std.Dev.
##  tags     (Intercept) 0.11
##  Residual             0.44
## ---
## number of obs: 2437, groups: tags, 66
## AIC = 3124, DIC = 2879.1
## deviance = 2983.5
```

2. Interpretation

The model is fit to 2437 videos within 66 tags.

The estimated regression line for an average tag is:

$log(views) = 12.15 - 0.01 * languages + 0.06 * log(comments) + 0.02 * duration + 0.02 * year(2007) - 0.06 * year(2008) - 0.29 * year(2009) - 0.27 * year(2010) - 0.53 * year(2011) - 0.48 * year(2012) - 0.21 * year(2013) - 0.07 * year(2014) + 0.25 * year(2015) + 0.41 * year(2016) + 0.52 * year(2017) + 0.01 * languages * log(comments)$

which means on average, with one more language translated for the video, log(number of views) will increase 0.01∗(log(number of comments)-1);

If log(number of comments) increase one unit, log(views) will increase (0.06+0.01∗languages);

If the duration of the video last one more minute, log(number of views) will increase 0.02;

Videos published in different years have different average views, such as if the video was published in 2017, log(number of views) will be 0.52 more than the video which was published in 2006.

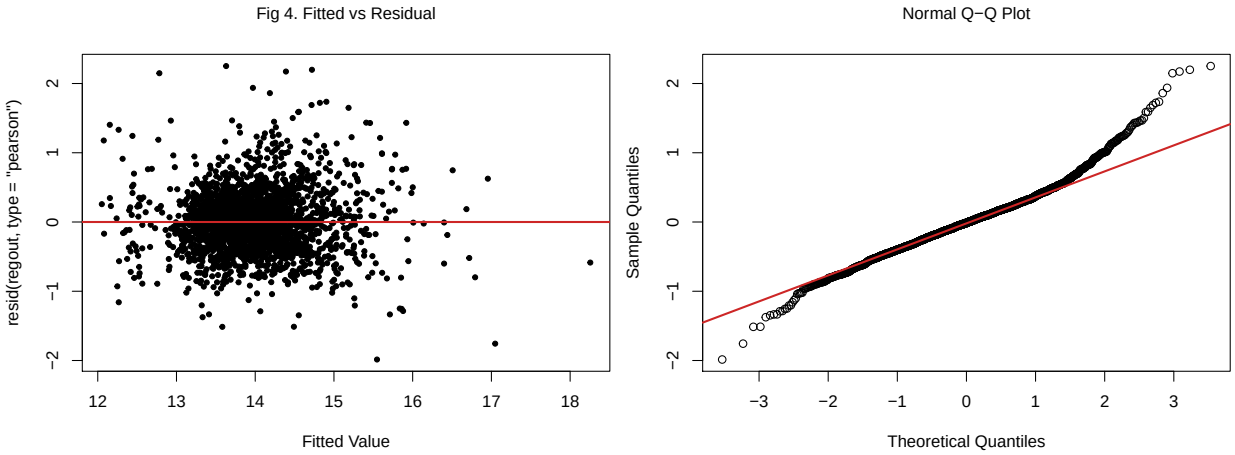
Table 1. (Intercept) by each tags

Tags	**(Intercept)**
activism	-0.0837804
adventure	-0.0089588
Africa	-0.0183912
art	0.0079001
biology	0.1155503
brain	-0.0711913
business	0.1458437
children	0.0855679
cities	-0.0335036
climate change	-0.0119213

Table 1 shows 10 tags' (Intercept) as an example, the complete table is listed in appendix.

These values indicate how much the intercept is shifted up or down for particular tags. For example, videos which tagged as "activism", the estimated intercept is 0.08 lower than average, so that the intercept value for the regression line is 12.15-0.08 = 12.07.

3. Model checking



From Fig 4, we can see that the points are pretty symmetrically distributed, tending to cluster towards the middle of the plot. They are not far away from y=0, and in general, there aren't clear patterns.

From the Normal Q-Q Plot, residual values are generally followed normal distribution. Though the data still have some extreme values.

V. Discussion

1. Implication

A popular TED video should have the following characteristics: already translated into many languages, many people left comments and discussed their opinions, and the duration of the video should not be too short.

Videos published in 2014-2017 are more popular than previous couple of years. It could because these topics are up to the minute. When scrolling through the newest videos, there are "Google Street View cars," "Me Too movement," "man-made DNA," etc. Those topics are definitely what most people care about.

From the ordered coefficient table (attached in Appendix), we can see that the videos which tagged as "entertainment" is the most popular category based on an average number of views. However, this tag contains a wide range of topics and speakers.

2. Limitation

The tags used to group videos is not so accurate, and some information missed when transforming the tags.

There is no data about comments content; however that is another question I am interested. I am curious about what the videos talked about and what the audiences discussed about. Whether most audiences agree with the speaker or they have different opinions.

3. Future direction

I would like to scrap more data from TED.com by myself and try to get comments data. Use LDA (Latent Dirichlet allocation) and NLP (Natural Language Processing) methods to explore more from the talks and comments.

VI. Acknowledgement

I would like to express my very great appreciation to Masanao for his valuable suggestion about how to build the multilevel model. This project cannot be finished without his help.

I also grateful for my friends who always supportive and generous. They gave me a lot of strength and care.

VII. Reference

Eldor, T. (2018, January 27). Data Reveals: What Makes a Ted Talk Popular? – Towards Data Science. Retrieved from https://towardsdatascience.com/data-reveals-what-makes-a-ted-talk-popular-6bc15540b995

Banik, R. (n.d.). The World of TED. Retrieved from https://www.kaggle.com/rounakbanik/ted-data-analysis

Fidelman, M. (2012, June 19). Here's Why TED and TEDx are So Incredibly Appealing (infographic). Retrieved from https://www.forbes.com/sites/markfidelman/2012/06/19/heres-why-ted-and-tedx-are-so-incredibly-appealing-inf #6a1581833b0e

Chaudhari, U. (2017, Nov 22). Why should one watch a TED talk?. Retrieved from https://www.quora.com/Why-should-one-watch-a-TED-talk

TED Speaker: Chris Anderson. (n.d.). Retrieved from https://www.ted.com/speakers/chris_anderson_ted

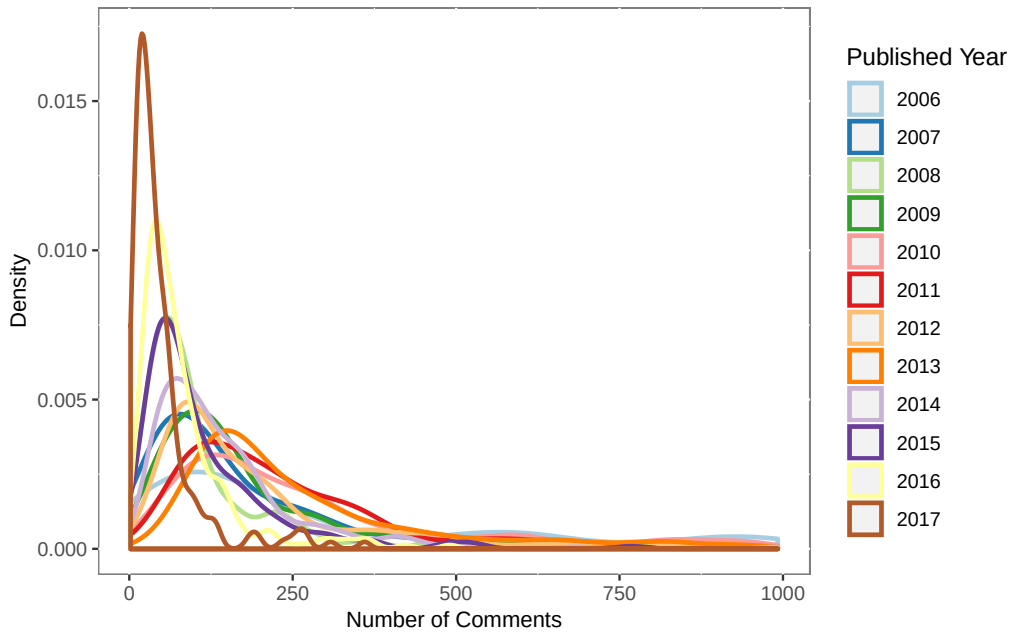
OLSEN, H. B. (2016, August 31). How TED Inspired a Whole Generation of Public Speakers. Retrieved from https://www.creativelive.com/blog/rise-of-ted-talks/

TED (conference). (2018, December 04). Retrieved from https://en.wikipedia.org/wiki/TED_(conference)

VIII. Appendix

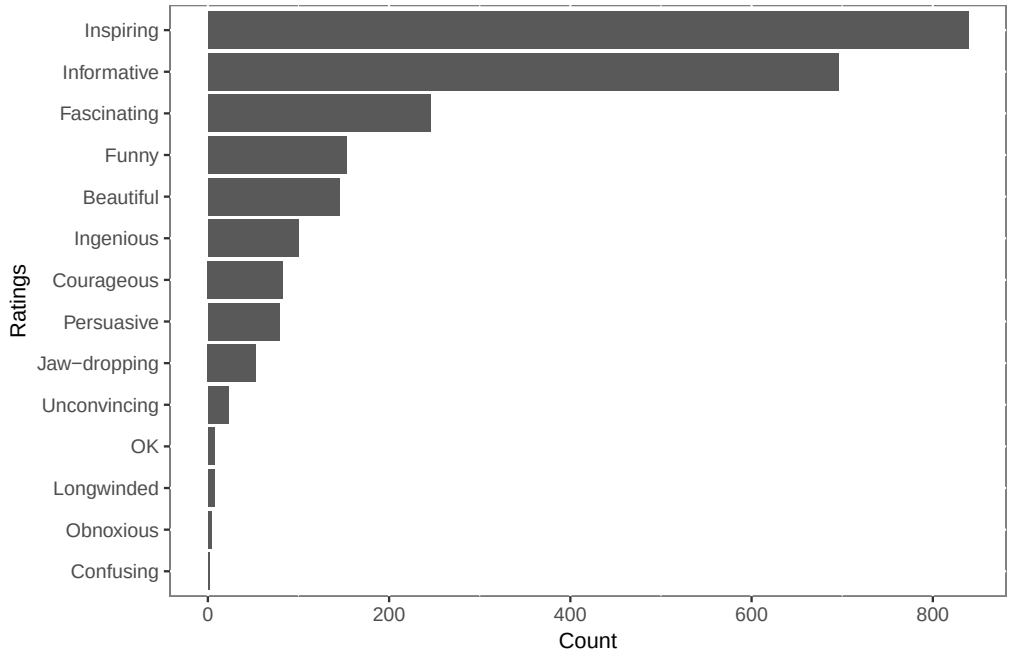
Fig 5. Density Distribution for Number of Comments by Published Year

Only focused on number of comments less than 1000 videos



Naturally, it takes time for more audiences to comment on and discuss their opinions. As Fig 3 shows that the density curve of 2017 is different from the density curve of 2011. In general, the videos which published earlier could get more comments.

Fig 6. Videos Rating Counts

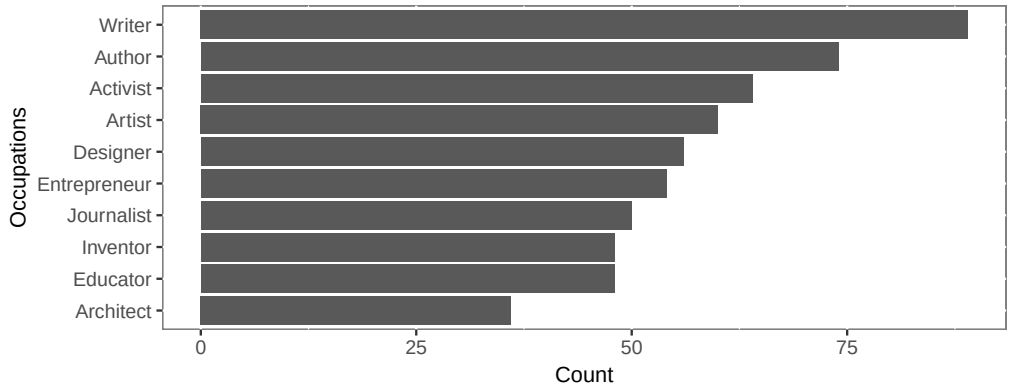


Every video has multiple ratings and counts. Here we keep the most rated one for each video and find out most talks are inspiring to most audiences.

The interesting thing is, I searched the videos which most rated is "Longwinded" and find out those videos are not long at all, except one of them. Duration of those videos is close to 18 minutes. By reading the comments, "Longwinded" is mainly because of the speaker's speaking ability, they focused on themselves or showed nervously.

Fig 7. Speakers' Occupation Count

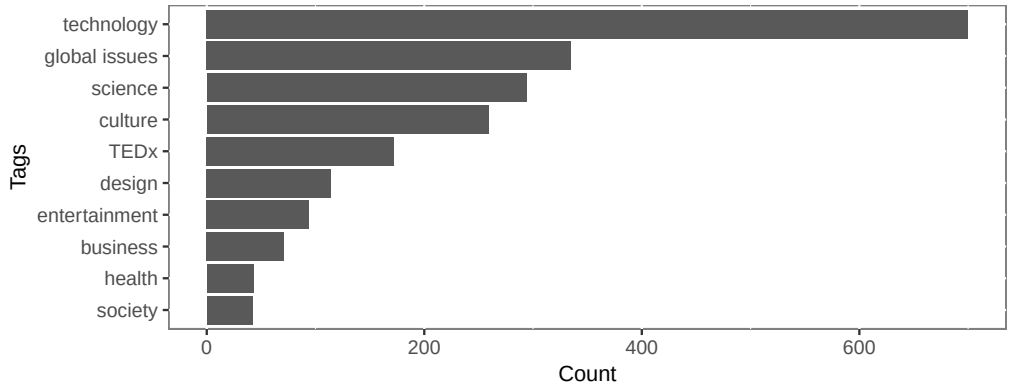
Only focused on top 10 occupations



The most frequent occupations are writers and artists. We can also see that many entrepreneurs gave the speak at TED, the reason could be a valuable way to push the company or project to the public.

Fig 8. Tags for Videos Count

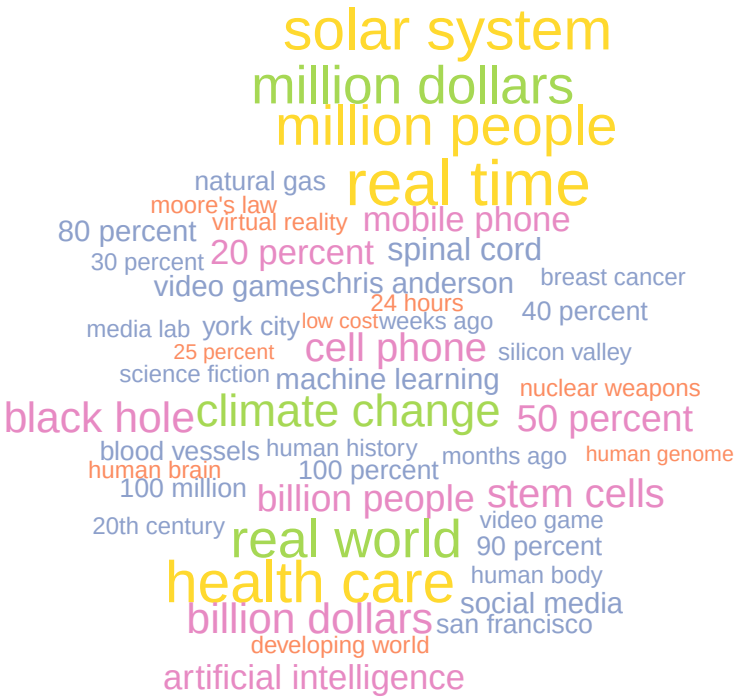
Only focused on top 10 tags



"Technology" appeared most which is one of TED's main topic.

In 2009 Chris introduced the TEDx initiative, allowing free licenses to local organizers who wished to organize their own TED-like events. More than 8,000 such events have been held, generating an archive of 60,000 TEDx talks.

Fig 9. Wordcloud for Technology Transcript



Pick the most frequent tag "Technology," and try to figure out what the speakers talked about.

Table 2. (Intercept) by tags

Tags	**(Intercept)**
entertainment	0.2317699
business	0.1458437
biology	0.1155503
humanity	0.1032408
children	0.0855679
design	0.0834242
culture	0.0684536
humor	0.0572147
nature	0.0556720
TEDx	0.0533842
invention	0.0476247
performance	0.0360828
science	0.0345319
happiness	0.0316561
personal growth	0.0305952
communication	0.0259011
math	0.0241486
identity	0.0188975
death	0.0177965
technology	0.0155269
work	0.0084725
art	0.0079001
mind	0.0063574
computers	0.0055834
medicine	0.0053790
community	0.0050684
music	0.0044622
global development	0.0005901
violence	-0.0047106
future	-0.0049511
physics	-0.0053285
philosophy	-0.0062591
space	-0.0072633
health care	-0.0075354
transportation	-0.0077465
adventure	-0.0089588
education	-0.0094849
oceans	-0.0094934
psychology	-0.0095792
Internet	-0.0111195
climate change	-0.0119213
philanthropy	-0.0131489
creativity	-0.0144074
innovation	-0.0144225
society	-0.0147001
collaboration	-0.0150770
Africa	-0.0183912
exploration	-0.0194181
women	-0.0207794

medical research	-0.0220539
data	-0.0297953
environment	-0.0304501
government	-0.0326642
cities	-0.0335036
history	-0.0396342
inequality	-0.0431872
health	-0.0431915
TED Fellows	-0.0436802
war	-0.0526382
economics	-0.0561502
politics	-0.0648260
brain	-0.0711913
activism	-0.0837804
storytelling	-0.0882987
social change	-0.1230916
global issues	-0.2338627