Bang Van Nguyen

TED Talks Analysis - Report

I. Problems and Hypotheses

Problem: TED talks may be losing their popularity!

The hypotheses I came up with include:

(1) The popularity of a TED talk depends on its speaker.

(2) TED Talks marketing strategy is not strong.

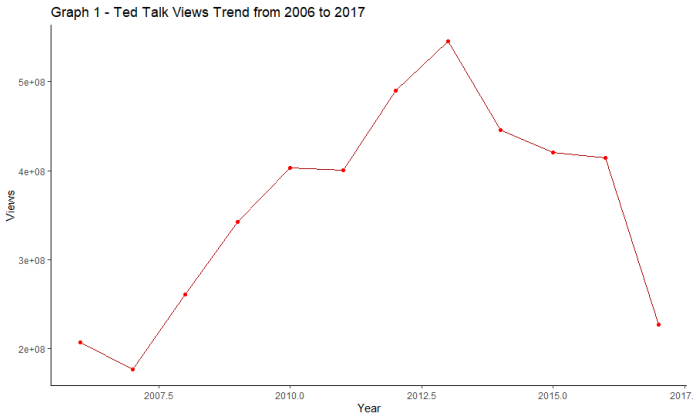
(3) The duration of a talk also affects its popularity.

(4) Topics related to technology, despite being promoted the most, are not appealing to the audience.

II. Exploratory Data Analysis

1. Confirming and analyzing the problem

I want to make sure that the observation that TED is losing its popularity is accurate and further analyze the problem. Therefore, from the original dataset, I create a graph that reflects the total number of online views for TED talks throughout the years. We can notice that from 2006 to 2013, the total number of views are increasing. From then on, the number of views consistently decreases. From 2016 to 2017, there is a sharp drop in terms of online views.



Published Year	Number of videos appeared in top 100	Average Number of Views	Average Number of Comments
2006	10	14324000	994.7
2007	3	7067660	412
2008	7	11105946	847.1429
2009	8	13025406	876.125
2010	7	15698739	1116.286
2011	13	7918454	531.3077
2012	9	13847783	810.1111
2013	22	8872865	523.6364
2014	6	9730753	333.5

2015	7	9209767	523.2857
2016	7	10053804	156
2017	1	5666038	250

Table 1 - The distribution of the top 100 most viewed TED talks throughout the years

Next, I analyze the top 100 videos with the greatest number of views. Table 1 describes how the top 100 most viewed videos are distributed from 2006 to 2017. I observe that, among the top 100 videos, there are fewer videos published after 2013 and the average views are declining in recent years. Noticeably, there is only one TED Talk in 2017 that made its way into the top 100. The average views decrease drastically over the period of 2016-2017. In 2016, the average views and comments are more than 10 million. Meanwhile, in 2017, the average views were only approximately 5.7 million. We can also see from the table that, having a high number of views does not mean provoking more comments.

From the observation above, I hope to see if there is a correlation among number of views, comments and positive or negative ratings. According to the correlation matrix, there is a strongly positive correlation graph total number of ratings and the number of positive ratings. This might suggest that people tend to provide positive ratings for a video. There is also a strongly positive correlation between views and positive ratings, which might suggest that videos with higher number of views are rated more positively and vice versa.

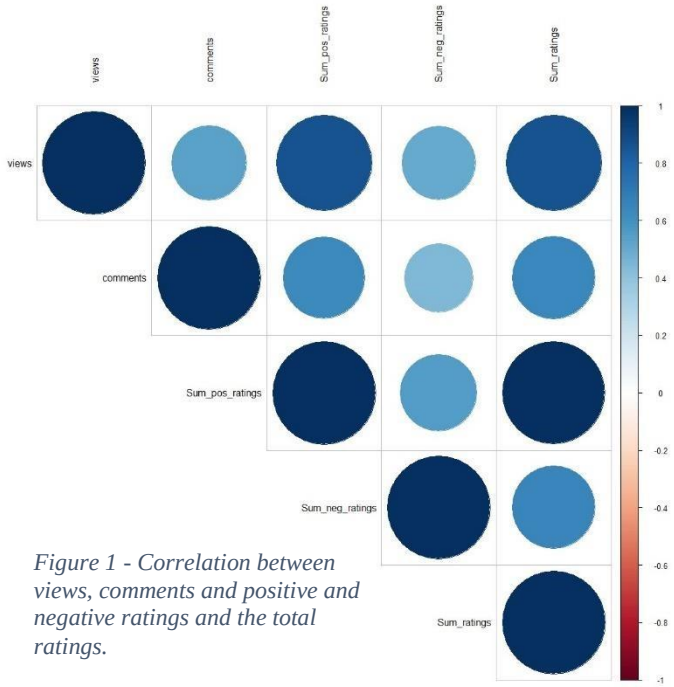
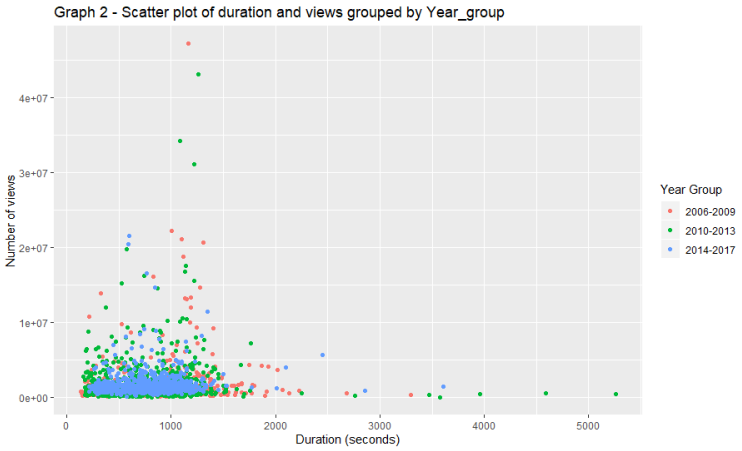
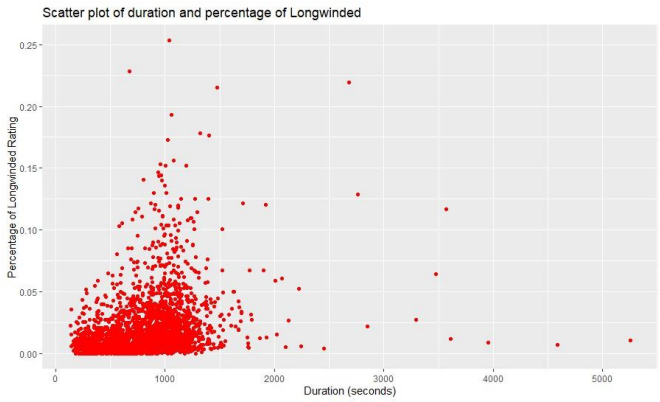


Figure 1 - Correlation between views, comments and positive and negative ratings and the total ratings.

2. Views and Duration



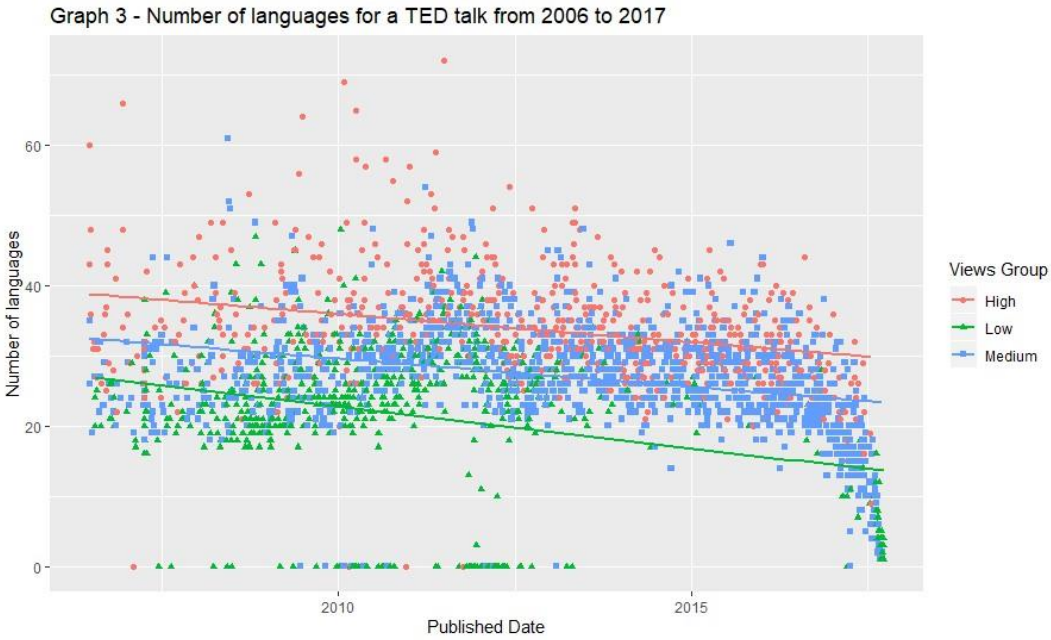
In order to analyze the relationship between the number of views and duration, the scatter plot on the left is utilized. It can be seen from the graph that for TED talks with a duration below 1500 seconds (25 minutes), the number of views is not heavily affected by its duration. In this range of length, the number of views for a video in Year Group 3 (2014-2017) is less than that of a video in Year Group 1 (2006-2009) and 2 (2010-2013). In addition, longer speeches with a duration of more than 1500 seconds do not attract a lot of viewers.



I also want to analyze the relationship between duration and ratings, especially "Longwinded", which can be seen in the figure on the left. The graph shows that the duration of a video does not affect the rating "Longwinded". Therefore, we can deduce that long speeches do not necessarily mean "Longwinded."

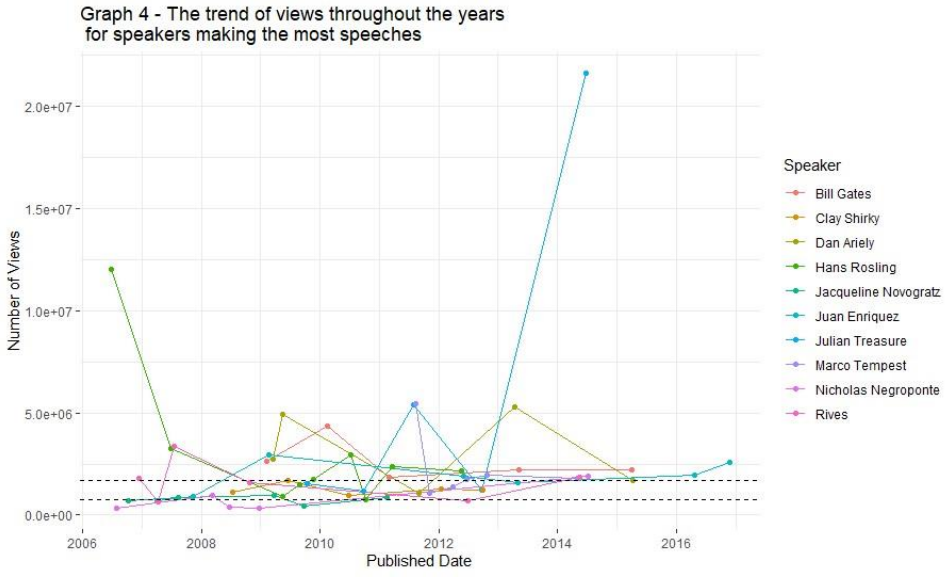
3. Views and Number of Languages

Analyzing the number of languages available for a video throughout the years with Views_group as a grouping variable, I aim to shed some light on the marketing strategy for TED:



From the scatter plot above, we can see that the number of languages for a video is decreasing throughout the years. It might also be suggested that highly viewed videos (75th percentile) tend to include more languages. This observation points out a potential strategy to increase the number of views: to increase the number of languages available for a video.

4. Views and Number of Languages



The graph above analyzes the number of views for speakers appearing in TED Talks most frequently. The upper and lower dashed lines mark the number of views' 75th percentile (which is categorized as highly-viewed) and 25th percentile respectively. We can see that, for most speakers, there is only one out of their many speeches to remarkably exceed the highly-

viewed threshold. This observation suggests that even if the audience is familiar with the speaker, it does not guarantee the success of a talk.

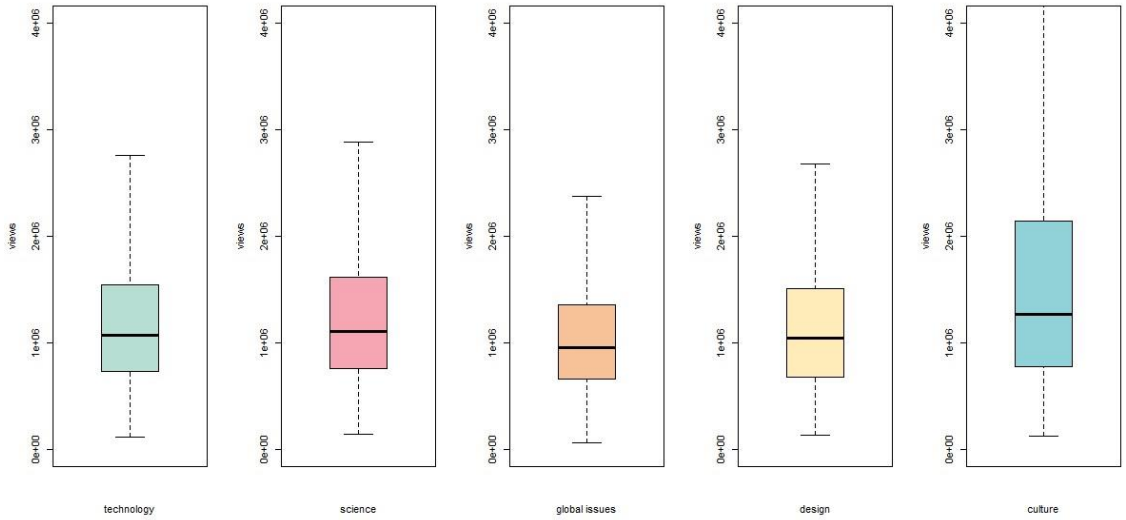
5. Views and Topics/Tags

I want to explore the most frequent topics that appear in TED talks, which are reflected in the table 2. "Technology" is the most frequent topic, which concurs with one of my hypotheses. From this finding, I hope to see how well it does in attracting viewers compared to other topics in top 5, reflected in Graph 5.

	Tags/Topics	Number of Appearance
1	'technology'	726
2	'science'	567
3	'global issues'	480
4	'design'	392
5	'culture'	380

Table 2 - Top 10 most frequent tags/topic for TED Talks

Graph 5 - Boxplot represents number of views for the top 5 tags



According to graph 5, "technology" brings about similar median number of views compared to other topics such as "science", and "design". Surprisingly, "culture" garners the highest median number of views despite being ranked top 5.

III. Fitting Models (Appendix)

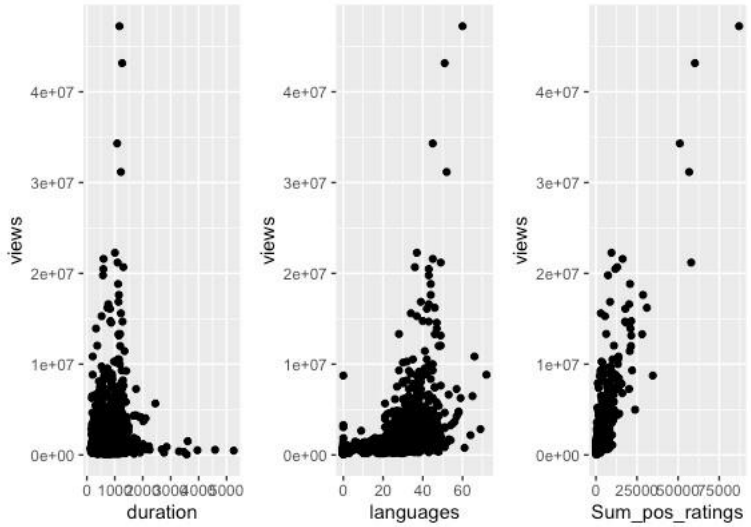
Hypothesis: There is a relationship between duration, languages, Sum_pos_ratings and views

Predictors: duration, languages, Sum_pos_ratings - **Type of variable:** quantitative

Response: views - **Type of variable:** quantitative

Type of model used:

multiple linear regression



1. Is there a relationship between the variables and the response?

Yes, there is a relationship between some of the variables and the response. From the graphs, we can notice that there is some positive correlation between Sum_pos_ratings and views and also between languages and views. There does not seem to be much of a relationship between duration and views.

2. How strong is the relationship between the predictors and the response?

The adjusted R^2 value for the model is 0.7564 which indicates that the relationship between the predictors and the response is quite strong.

3. Do the predictors contribute to the response?

Only the predictors of Sum_pos_ratings and languages contribute to the response as their Pr(>|t|) values were below 0.05 so we have evidence to say that these coefficients are different from 0; duration, with a Pr(>|t|) value of 0.615, does not contribute to the response.

4. What is the effect of each predictor on the response?

For every increase of 1 unit in Sum_pos_ratings, views increase by 5.481e+02 and for every increase in one unit of languages, views increase by 2.584e+04. Also, for every increase in one unit of duration, views decrease by 3.526e+01.

5. How accurately can we predict the response using the predictor with results from lm()?

I use RSE to measure the accuracy with which the model predicts. The RSE is 1233000 (views). That is, the number of views predicted from this model will be 1233000 units lower or higher than that in reality. We should also consider other models and compare this number to their RSE.

6. Use a model selection procedure to select the best model

In order to select the optimal model, I perform backward selection using AIC. It can be seen in the results that the final model only includes languages and Sum_pos_ratings.

7. Is the selected model the same as the model with all the variables?

After performing backward selection using AIC, the selected model contains only two predictors – languages and Sum_pos_ratings, which is different from the original model with three predictors.

8. How is the selected model different from the original model?

As mentioned above, duration has a Pr(>|t|) value that exceeds the 0.005 limit, and therefore is eliminated from the final model. Additionally, the final model with two predictors languages and Sum_pos_ratings has an AIC value of 71530.60, which is slightly less than the AIC for the original model (71532.35). This confirms that the final model is better than the original one.

IV. Conclusion

After performing EDA and fitting models, I reach a conclusion that in order to increase views and regain popularity,

1. TED should increase the number of languages available for a video to provide access for people from different countries.
2. A talk should remind audience of positive keywords such as "convincing", "persuasive", "beautiful", etc.
3. "Technology" does not bring about the highest number of views, so TED should try to also promote other topics such as culture, global issues, etc.
4. Duration does not affect the views of a video, but TED should try to keep it under 25 minutes (1500 seconds).

Appendix

```
m1=lm(views~languages+duration+Sum_pos_ratings, data=dataForModel)
> summary(m1)

Call:
lm(formula = views ~ languages + duration + Sum_pos_ratings,
    data = dataForModel)

Residuals:
      Min        1Q    Median        3Q       Max
-11697683   -422831   -111241    252323  16269659

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.159e+05  1.112e+05  -1.042    0.297
languages        2.584e+04  2.893e+03   8.934   <2e-16 ***
duration        -3.526e+01  7.009e+01  -0.503    0.615
Sum_pos_ratings  5.481e+02  7.002e+00  78.289   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1233000 on 2546 degrees of freedom
Multiple R-squared:  0.7566,   Adjusted R-squared:  0.7564
F-statistic:  2639 on 3 and 2546 DF,  p-value: < 2.2e-16

>
>
> #Perform backward selection to find the best model
> library(MASS)
> m1_subsets= stepAIC(m1, direction = "backward", trace=FALSE)
> summary(m1_subsets)

Call:
lm(formula = views ~ languages + Sum_pos_ratings, data = dataForModel)

Residuals:
      Min        1Q    Median        3Q       Max
-11674958   -424325   -110941    251910  16264248

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.574e+05  7.467e+04  -2.108   0.0351 *
languages        2.635e+04  2.708e+03   9.729   <2e-16 ***
Sum_pos_ratings  5.474e+02  6.828e+00  80.161   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1233000 on 2547 degrees of freedom
Multiple R-squared:  0.7566,   Adjusted R-squared:  0.7564
F-statistic:  3959 on 2 and 2547 DF,  p-value: < 2.2e-16

> m1_subsets$anova
Stepwise Model Path
Analysis of Deviance Table
```

```
Initial Model:
views ~ languages + duration + Sum_pos_ratings

Final Model:
views ~ languages + Sum_pos_ratings


        Step Df     Deviance Resid. Df   Resid. Dev      AIC
1                                 2546 3.872251e+15 71532.35
2 - duration  1 384953707869      2547 3.872636e+15 71530.60
```