What Predicts The Popularity Of Ted Talks? An Analysis (and Adventure in Data Engineering)

A Consulting Project for Math 6627 (2/3)

Benjamin Smith

20 March 2022

Contents

List of Figures

List of Tables

Introduction

(Quoted from the SSC website)

TED spreads ideas, primarily via short talks that can be accessed on the internet. As noted on its website, TED was initiated in 1984 as a conference where technology, entertainment, and design ideas were shared. As of present, TED Talks cover topics ranging from science to business to global issues.

The following analysis focuses on the use of inferential techniques to analyze the data. The questions addressed in this analysis are:

1. What characteristics of TED Talks predict their popularity?
2. What different ways could the popularity of TED Talks be measured?
3. Do the characteristics that predict popularity change over time?
4. Do the characteristics that predict popularity differ based on the theme of the TED Talks?

The Data

The data was made available by Kaggle. Using by use of the `naniar` R package, Figure 1 shows that there very little data missing in this data set. As such there is no treatment applied to the data and it is used as is.

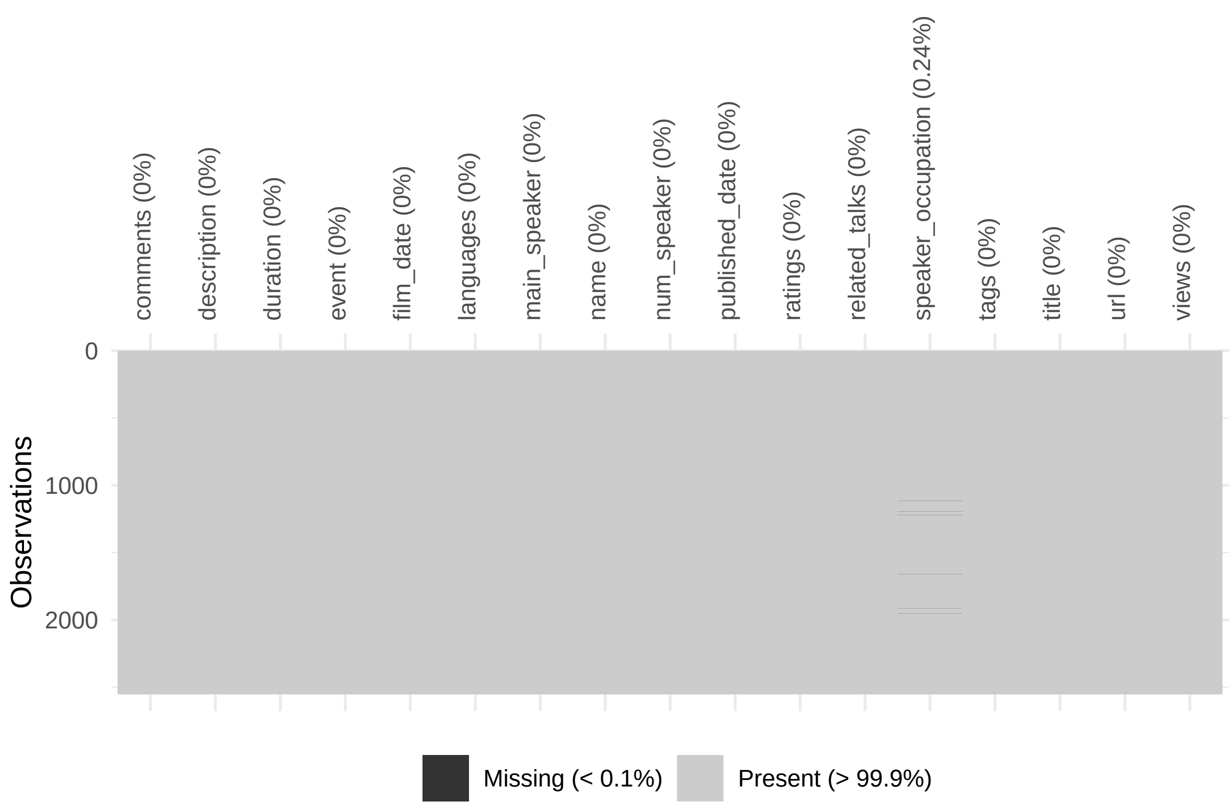

Figure 1: There is little to no missing data in this dataset

The challenge with this data set lie in the `ratings`, `related_talks` and `tags` fields. These fields are `.json` files which were inserted into the .csv file. An example of data contained in an individual `ratings` observation is listed below:

```
[{'id': 3, 'name': 'Courageous', 'count': 139}, {'id': 2, 'name': 'Confusing', 'count':
↪  25}, {'id': 1, 'name': 'Beautiful', 'count': 48}, {'id': 9, 'name': 'Ingenious',
↪  'count': 31}, {'id': 21, 'name': 'Unconvincing', 'count': 35}, {'id': 11, 'name':
↪  'Longwinded', 'count': 21}, {'id': 8, 'name': 'Informative', 'count': 218}, {'id':
↪  10, 'name': 'Inspiring', 'count': 113}, {'id': 22, 'name': 'Fascinating', 'count':
↪  44}, {'id': 25, 'name': 'OK', 'count': 51}, {'id': 23, 'name': 'Jaw-dropping',
↪  'count': 35}, {'id': 24, 'name': 'Persuasive', 'count': 112}, {'id': 7, 'name':
↪  'Funny', 'count': 9}, {'id': 26, 'name': 'Obnoxious', 'count': 11}]
```

For the `related_talks` field, an example of the data contained in an individual observation is:

```
[{'id': 127, 'hero':
↪  'https://pe.tedcdn.com/images/ted/5cd871dcf27ba4288021c2bfe6a3f6796dab2538_2880x1620.jpg',
↪  'speaker': 'Ngozi Okonjo-Iweala', 'title': 'Want to help Africa? Do business here',
↪  'duration': 1213, 'slug': 'ngozi_okonjo_iweala_on_doing_business_in_africa',
↪  'viewed_count': 1044183}, {'id': 1929, 'hero':
↪  'https://pe.tedcdn.com/images/ted/82bbf525e7b13a879e6b7299303ec510f7ceb9fb_1600x1200.jpg',
↪  'speaker': 'Michael Metcalfe', 'title': 'We need money for aid. So let’s print it.',
↪  'duration': 864, 'slug': 'michael_metcalfe_we_need_money_for_aid_so_let_s_print_it',
↪  'viewed_count': 756965}, {'id': 584, 'hero':
↪  'https://pe.tedcdn.com/images/ted/98530_800x600.jpg', 'speaker': 'Paul Collier',
↪  'title': 'New rules for rebuilding a broken nation', 'duration': 994, 'slug':
↪  'paul_collier_s_new_rules_for_rebuilding_a_broken_nation', 'viewed_count': 406525},
↪  {'id': 1196, 'hero':
↪  'https://pe.tedcdn.com/images/ted/7bb5389d0360ef7905de6b6a017b7ce836ad673d_800x600.jpg',
↪  'speaker': 'Rory Stewart', 'title': 'Time to end the war in Afghanistan', 'duration':
↪  1202, 'slug': 'rory_stewart_time_to_end_the_war_in_afghanistan', 'viewed_count':
↪  659270}, {'id': 270, 'hero':
↪  'https://pe.tedcdn.com/images/ted/1cffd7f06b5754232bc90a0ca15b1339487d7200_2400x1800.jpg',
↪  'speaker': 'Paul Collier', 'title': 'The \"bottom billion\"', 'duration': 1011,
↪  'slug': 'paul_collier_shares_4_ways_to_help_the_bottom_billion', 'viewed_count':
↪  990214}, {'id': 2806, 'hero':
↪  'https://pe.tedcdn.com/images/ted/f26393b438dfc2ed8c8ae66d0c7291ac08629153_2880x1620.jpg',
↪  'speaker': 'Jim Yong Kim', 'title': \"Doesn't everyone deserve a chance at a good
↪  life?\", 'duration': 1332, 'slug':
↪  'jim_yong_kim_doesn_t_everyone_deserve_a_chance_at_a_good_life', 'viewed_count':
↪  1341183}]
```

For the individual `tags` field, an example of the data contained in an individual observation is:

```
['business', 'corruption', 'culture', 'economics', 'entrepreneur', 'global development',
↪  'global issues', 'investment', 'military', 'policy', 'politics', 'poverty']
```

The present structure of the data has nested .json fields. For the data to be usable, it needs to be un-nested and expanded.

Data Engineering

The detailed line-by-line code for extracting the data is in the code appendix. So to discuss the issue more generally, the data needed to be extracted and converted from `.json` form to a data-frame. Surprisingly, `jsonlite` package was unable to parse the strings successfully. In lieu of this the `yaml` package was used.

Before employing the `yaml` package the data needed to be converted into a format that is easier to read. This involved removing and replacing recurring instates of forward slashes (a common escape tag) and converting utf-8 encoded characters into latin-ascii format[1]. In particular the `stringr` package was used for cleaning the `.json` strings (using `str_remove_all`) and the `stringi` package was used to convert the encoding from utf-8 to latin-ascii (by using `stri_trans_general`).

This resulted in a transformed data set which had 2550 observations of 17 variables to having 268156 observations of 17 variables. The data is used in this form for the last two questions in the analysis as it is in "long" form. For the first two questions the data needs to be pivoted to "wide" form and have all categorical variables be assigned as dummy variables. For this, the `dplyr`(a variety of functions) and `tidyr` (in particular `pivot_wider`) packages were employed. The resulting data set returned to having 2550 observations, but now having 433 variables with all the additional variables being from the extracted from the nested `.json` in the `ratings`, `related_talks` and `tags`fields.

Analysis

Characteristics of TED Talks Which Predict Popularity

For determining which characteristics predict popularity of a given TED talk, one of the paths of least resistance lies in employing LASSO regression. For this model, the variable of interest which measures popularity would intuitively be the number of views accumulated by a given TED talk.

After doing 10-fold cross validation, it was determined that that MSE is minimized when $\lambda_{views} = 38914.59$ (see figure 2). Table 1 shows the non-zero sparse estimates. Generally speaking, TED talks that are more recent and are available in multiple languages get more views. In terms of ratings (while it can be argued that this speaks more general to positive ratings), TED talks that are primarily rated as informative, beautiful and funny are have more views. In terms of tags, shows relating to drones, magic and body language have more views, while shows relating to philosophy, personality and statistics (shockingly) have less views.

[1] This was an issue which was specific to using the Windows operating system, however on a Mac or Linux operating system there was no such issue. For consistency across machines this was applied.

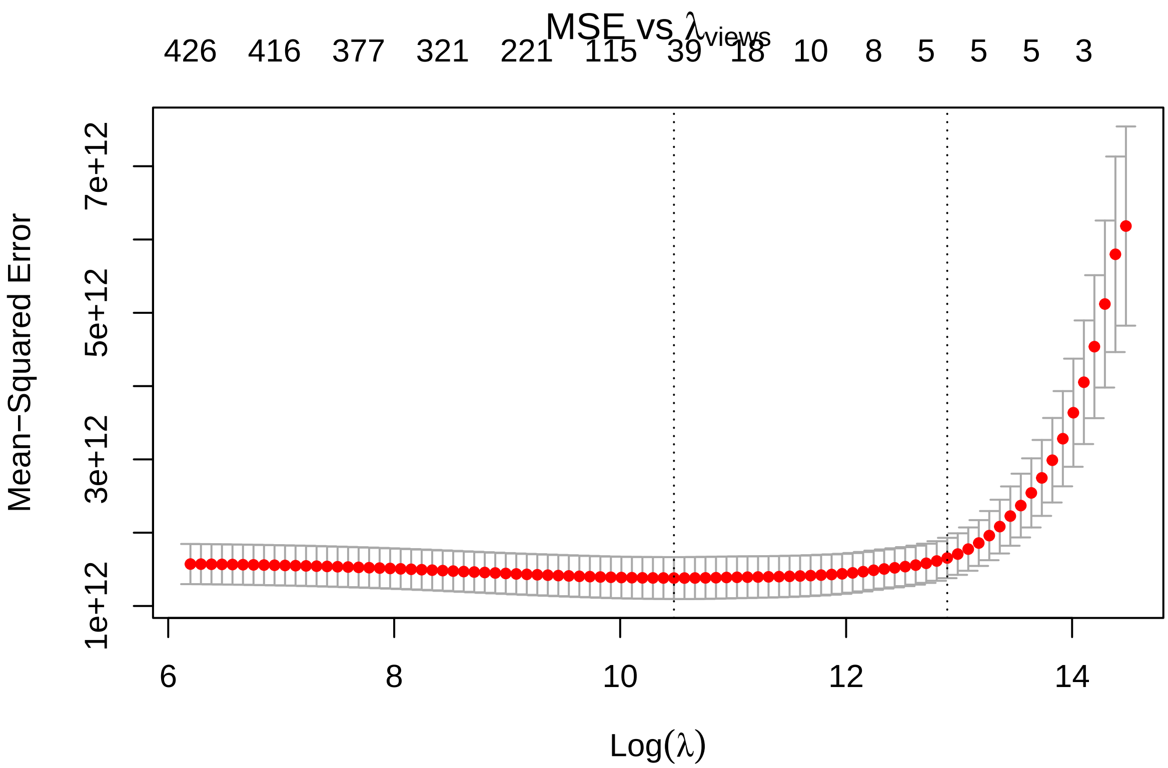

Figure 2: MSE vs λ_{views}

Table 1: Sparse Estimates for Ted Talk Popularity (in terms of Views)

	s0
(Intercept)	-4.072118e+06
duration	1.347866e+02
languages	1.369185e+04
published_date	2.782800e-03
rating_Funny	7.808418e+02
rating_Beautiful	1.057308e+02
rating_Ingenious	6.187135e+02
rating_Courageous	4.821501e+02
rating_Confusing	1.421666e+03
rating_Informative	9.573936e+02
rating_Fascinating	6.414673e+02
rating_Unconvincing	-5.311859e+01
rating_Jaw-dropping	6.090154e+01
rating_OK	5.851244e+03
rating_Inspiring	3.022947e+02
tag_global issues	-9.985131e+03
tag_science	-1.463787e+04
tag_performance	3.104004e+05
tag_politics	-9.376492e+03
tag_Google	-7.971599e+04
tag_statistics	-3.856399e+05
tag_potential	8.406557e+04
tag_consciousness	-2.845858e+05
tag_philosophy	-1.094232e+05
tag_wunderkind	1.655430e+05
tag_youth	2.018383e+02
tag_relationships	2.567798e+05
tag_aging	5.979625e+04
tag_flight	2.625136e+05
tag_photography	7.802916e+04
tag_robots	1.327036e+04
tag_success	2.107773e+05
tag_language	-5.691995e+04
tag_live music	1.949449e+05
tag_self	-8.405243e+04
tag_meme	-5.193469e+04
tag_sociology	-5.418072e+04
tag_human origins	-8.355908e+04
tag_drones	1.229892e+05
tag_magic	7.789958e+05
tag_personality	-1.324996e+05
tag_prison	4.615531e+04
tag_fashion	5.111021e+05
tag_body language	7.952739e+05
tag_advertising	-7.460081e+04
tag_speech	1.005945e+04

Different Ways Popularity Of TED Talks Can Be Measured

From simple inspection of the data, the three possible ways that the popularity of a Ted Talk can be measured would be in terms of number of views, ratings and comments. Since number of views were explored in the previous section, in this section we will focus on ratings and comments.

Ratings

From the data set it was determined that there are 14 unique rating tags for each TED talk. Table 2 shows the unique rating tags and the manual classification assigned to them. Since "OK" is an ambiguous term it is not given a good or bad assignment. A "Good/Bad Ratio" is calculated by looking at the ratio of the number of positive and negative reviews. With this ratio, tables 3 and 4 show the top 10 worst and best ted talks as classified by this ratio.

By visual inspection, it can be seen that the good/bad ratio is not indicative of engagement in terms of views[2] or comments.

Table 2: Unique Rating Tags accross all Ted Talks

Rating Tag	Classification
Funny	Good
Beautiful	Good
Ingenious	Good
Courageous	Good
Longwinded	Bad
Confusing	Bad
Informative	Good
Fascinating	Good
Unconvincing	Bad
Persuasive	Good
Jaw-dropping	Good
OK	Ambiguos
Obnoxious	Bad
Inspiring	Bad

Table 3: Top 10 Worst Ted Talks

name	Good/Bad Ratio	views	comments	published_date
Daniel Libeskind: 17 words of architectural inspiration	0.1412942	784642	423	2009-07-01 01:00:00
John Maeda: My journey in design	0.4345550	241858	26	2009-01-06 05:08:00
Pete Alcorn: The world in 2200	0.4710018	493966	126	2009-06-08 01:00:00
Richard Ledgett: The NSA responds to Edward Snowden's TED Talk	0.5082927	1191342	440	2014-03-21 00:46:29
Raghava KK: What's your 200-year plan?	0.5119760	811778	56	2012-07-04 14:16:42

[2]However, from the sparse estimates in the LASSO model for predicting number of views, more positive reviews appear to effect the number of views on a given TED talk

name	Good/Bad Ratio	views	comments	published_date
Kelli Jean Drinkwater: Enough with the fear of fat	0.5132450	1594248	326	2016-10-28 16:55:49
Fields Wicker-Miurin: Learning from leadership's missing manual	0.5567766	956175	55	2009-11-18 09:17:00
David Rockwell: A memorial at Ground Zero	0.5755814	404402	14	2007-06-12 05:11:00
Jakob Trollback: A new kind of music video	0.5810277	480377	68	2008-04-03 01:14:00
Susan Lim: Transplant cells, not organs	0.6083569	620231	273	2011-04-15 18:47:00

Table 4: Top 10 Best Ted Talks

name	Good/Bad Ratio	views	comments	published_date
Jack Horner: Where are the baby dinosaurs?	27.90625	1063288	78	2012-02-09 15:59:58
Ed Yong: Zombie roaches and other parasite tales	25.86413	1624605	173	2014-03-26 15:05:29
Rodrigo Canales: The deadly genius of drug cartels	20.49474	2225283	286	2013-11-04 16:01:14
Sebastian Wernicke: Lies, damned lies and statistics (about TEDTalks)	20.40919	2212944	279	2010-04-30 08:59:00
Marcus Byrne: The dance of the dung beetle	19.34286	1003863	72	2012-12-13 16:00:50
James Veitch: This is what happens when you reply to spam email	19.20331	20475972	150	2016-01-08 16:03:40
Apollo Robbins: The art of misdirection	19.12000	15283242	285	2013-09-13 15:02:39
Blaise Agüera y Arcas: How PhotoSynth can connect the world's images	18.92850	4772595	260	2007-05-27 00:37:00
Ben Goldacre: What doctors don't know about the drugs they prescribe	18.86826	2228138	380	2012-09-27 15:01:44
Jennifer 8. Lee: The hunt for General Tso	18.71429	1285775	84	2008-12-24 01:00:00

Comments

If we want to define popularity in terms of engagement, the number of comments on a given TED talk can be indicative. As in the case of predicting the number of views, LASSO regression is applied with the response variable being the number of comments on a given TED talk.

After doing 10-fold cross validation, it was determined that that MSE is minimized when $\lambda_{comments} = 5.742292$ (see figure 3). Table 5 shows the sparse estimates produced. In particular TED talks with tags about atheism, religion and G-d are some of the largest predictors. This is an intuitive result as for many these topics are controversial and will bring about much engagement in the form of comments.

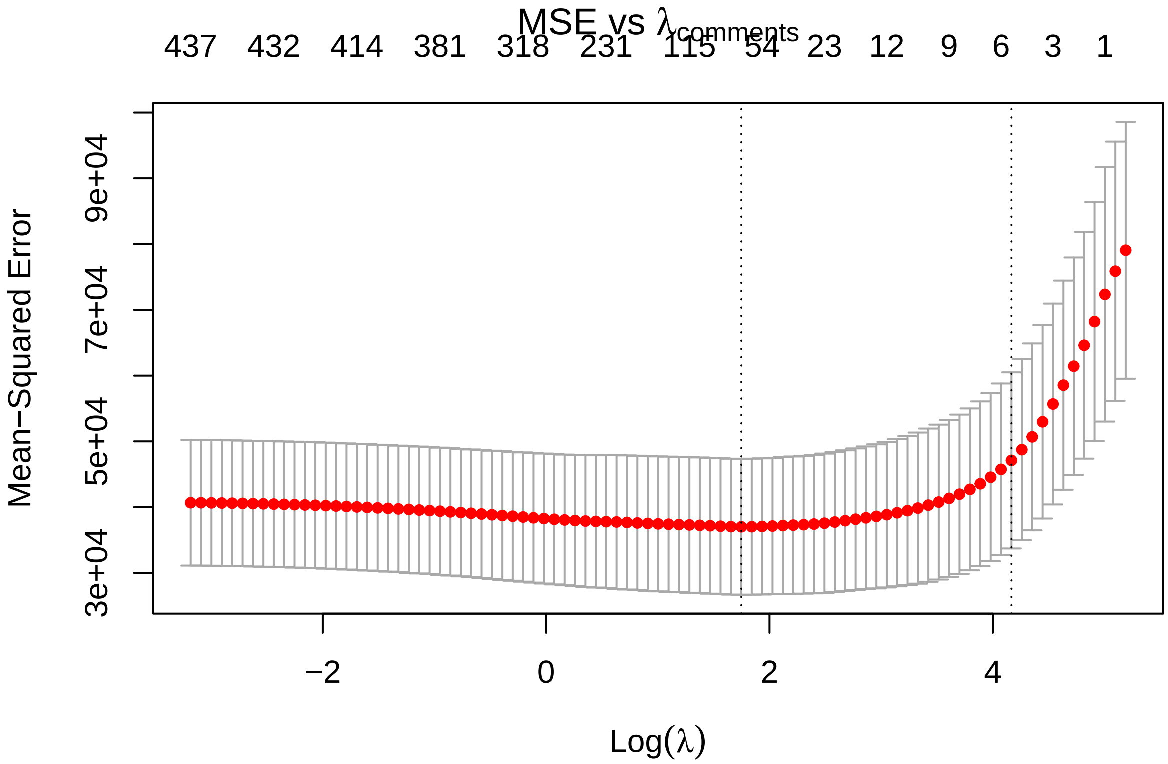

Figure 3: MSE vs $\lambda_{comments}$

Table 5: Sparse Estimates for Ted Talk Popularity (in terms of comments)

	s0
(Intercept)	-13.8205661
duration	0.0307682
languages	2.3330879
published_date	0.0000000
rating_Ingenious	0.1061314
rating_Courageous	0.1562817
rating_Fascinating	0.0170527
rating_Unconvincing	0.3975905
rating_Persuasive	0.1793041
rating_Jaw-dropping	0.0152816
rating_Obnoxious	0.4138278
tag_culture	18.1625423
tag_global issues	4.8005621
tag_science	6.0682422
tag_activism	-14.1120791
tag_politics	14.2064918
tag_Africa	-6.5881978
tag_Asia	20.9994775
tag_Google	-23.9416065
tag_motivation	-32.7677643
tag_Christianity	-167.5055312
tag_God	194.6708619
tag_atheism	961.5909912
tag_humor	-18.7171107
tag_religion	114.6727052
tag_architecture	-17.2424120
tag_consciousness	67.2292684
tag_philosophy	24.7800148
tag_happiness	-10.7264205
tag_leadership	-39.9985690
tag_nature	-4.0756278
tag_community	-5.4490951
tag_communication	-19.0190913
tag_choice	-32.8085377
tag_personal growth	-18.1285473
tag_faith	-50.1179268
tag_success	-27.9351874
tag_work	-8.9979603
tag_evolutionary psychology	69.4856062
tag_work-life balance	-17.7415122
tag_apes	-132.4431576
tag_self	-14.8765642
tag_china	102.3609495
tag_energy	11.8737559
tag_adventure	-3.6914367
tag_String theory	47.9024655
tag_big bang	11.3096334
tag_society	-15.8508392
tag_beauty	-5.8238615

	s0
tag_identity	-4.0311250
tag_morality	49.5967214
tag_fear	-28.6068239
tag_wind energy	2.5587885
tag_productivity	-9.0342923
tag_agriculture	37.6329893
tag_neuroscience	23.4460328
tag_money	7.2093860
tag_Anthropocene	19.6748356
tag_novel	124.4857261
tag_feminism	9.9820810
tag_nuclear weapons	79.6142764
tag_bullying	24.4004534
tag_deextinction	21.8371304

Characteristics predicting popularity over time

For seeing how the characteristics which predict popularity change over time, an exploratory approach is taken. Figures 4 and 5 demonstrate this. Since there are hundreds of tags, the top 5 tags with the most views and comments for each year were selected. In terms of views, talks with the tags “business”, “culture”, “entertainment”,“science”,“technology” and “TEDx” have held some consistency over the years. In terms of comments, contrary to the LASSO regression results, it is found that talks with the tags “culture”, “global issues” and “technology” proved to be the talks with the most comments.

It can thus be understood that the large LASSO estimates relating to the magnitude of the individual effect of a given tag while the figures 4 and 5 show the top tags among the TED talks available.

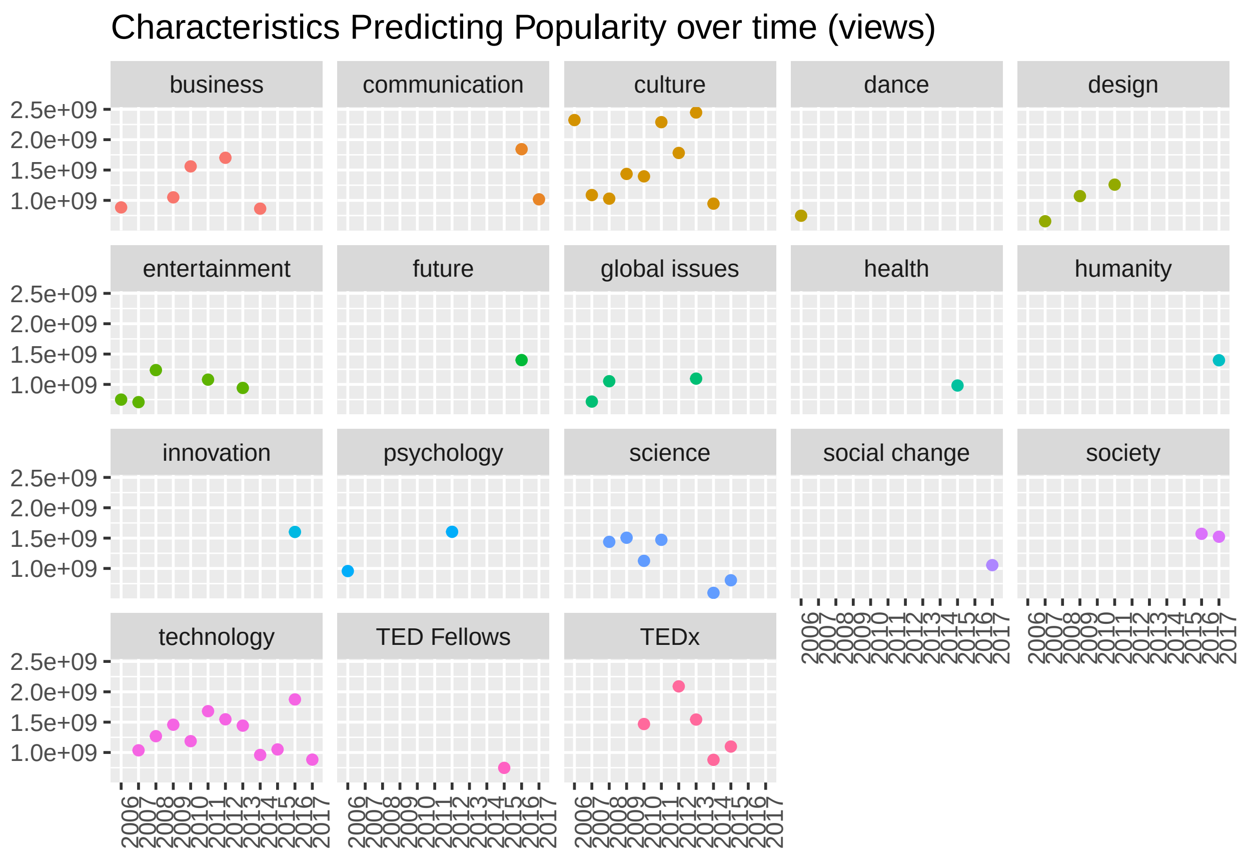

Figure 4: Characteristics Predicting Popularity over time in terms of views

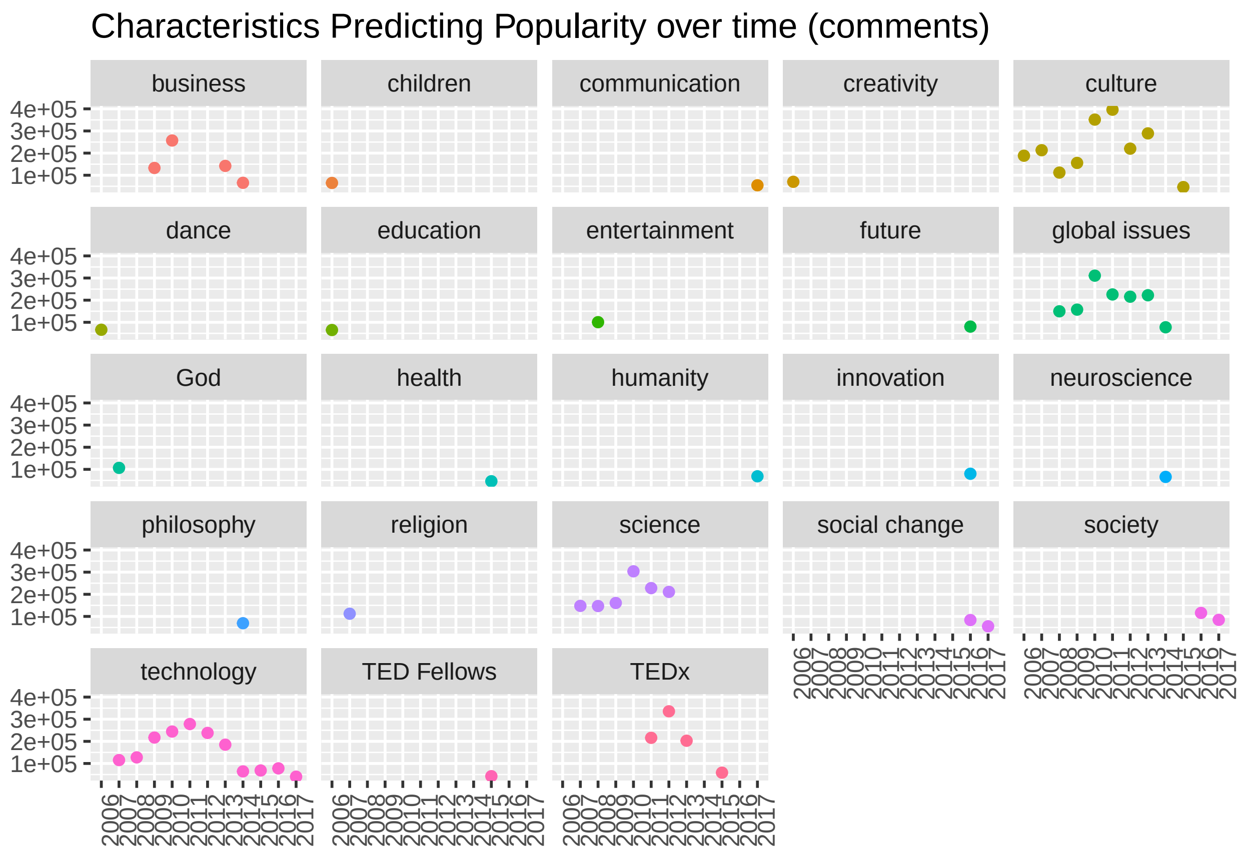

Figure 5: Characteristics Predicting Popularity over time in terms of comments

Characteristics That Predict Popularity Based On The Theme Of The TED Talk

To see which characteristics that predict popularity based on the theme of a given TED talk, a mixed model structure can be adopted. The model used is:

$$Y = X\beta + Z\mathrm{b}$$

Where Y is the response variable of interest (views or comments) and X is the design matrix for the fixed effects Z is the design matrix of the random effects and β and b are the fixed and random effects vectors.

The fixed effects are:

- Talk duration (`duration`)
- Tag (`TagValue`)
- Number of languages the talk is available in (`languages`)
- For the views model - the number of comments
- For the comments model- the number of views
- Tag interaction with the other fixed effects (two way interactions)

The random effects are assigned to the main speaker of the talk.

Since there are hundreds of tags available, the data for each model is filtered to the tags which appeared to get the most engagement as shown in the previous section. Tables 6 and 7 show the fixed effects summaries from the views and comments models.

For the views model, it is found that longer talks related to psychology and science have a significant[3] positive relationship with views while talks relating to other tags which are longer have a negative or a non-significant relationship[4] with views. Talks which are translated into more languages has a significant positive effect for views in talks related to culture, global issues, psychology, science, while other talks translated into more languages either had a negative or non-significant relationship with views. Talks with more comments which related to communication, dance, social change and TEDx have a significant positive relationship with views while other talks have a negative or non-significant relationship with views.

For the comments model, it is found that longer talks related to culture, G-d, philosophy and religion have a significant positive relationship with number of comments while talks relating to other topics have a negative or a non-significant relationship with the number of comments. Talks which are translated in more languages relating to communication, culture, G-d, humanity, philosophy, religion, social change and society has a significant positive relationship with number of comments a talk receives while other tags have a negative or non-significant relationship with the number of comments received. Talks with more views related to children, creativity, dance, education, global issues and religion have a significant positive relationship with the number of comments while other tags have a negative or non-significant relationship with the number of comments on a given talk.

Table 6: Fixed effects of the views mixed model

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-3.706875e+06	95347.89520	75804	-38.8773699	0.0000000
duration	1.735638e+03	54.45275	75804	31.8741951	0.0000000
tagValuecommunication	7.071076e+05	131600.46403	75804	5.3731392	0.0000001
tagValueculture	-3.791367e+05	101218.46293	75804	-3.7457263	0.0001800
tagValuedance	-8.333399e+05	287311.08789	75804	-2.9004795	0.0037270
tagValuedesign	8.474041e+05	103400.26455	75804	8.1953766	0.0000000

[3]Having a p-value of less than 0.05.
[4]Having a p value greater than 0.05

	Value	Std.Error	DF	t-value	p-value
tagValueentertainment	7.276037e+05	109105.25121	75804	6.6688240	0.0000000
tagValuefuture	8.123840e+05	114215.33104	75804	7.1127407	0.0000000
tagValueglobal issues	4.522387e+05	95270.87686	75804	4.7468727	0.0000021
tagValuehealth	5.804644e+04	122855.02570	75804	0.4724792	0.6365862
tagValuehumanity	6.837653e+05	113828.39179	75804	6.0069834	0.0000000
tagValueinnovation	7.953551e+05	110279.88864	75804	7.2121499	0.0000000
tagValuepsychology	-1.402262e+06	129773.70104	75804	-10.8054389	0.0000000
tagValuescience	-5.589092e+04	93155.64450	75804	-0.5999735	0.5485257
tagValuesocial change	8.484756e+05	110317.97292	75804	7.6911823	0.0000000
tagValuesociety	1.096184e+06	103875.30262	75804	10.5528866	0.0000000
tagValuetechnology	7.596702e+05	88587.18077	75804	8.5753962	0.0000000
tagValueTED Fellows	1.126998e+06	155316.66195	75804	7.2561298	0.0000000
tagValueTEDx	3.058269e+05	97557.18950	75804	3.1348473	0.0017201
languages	1.203193e+05	2036.29593	75804	59.0873329	0.0000000
comments	3.269992e+03	61.94161	75804	52.7915140	0.0000000
duration:tagValuecommunication	-7.448601e+02	85.65844	75804	-8.6956991	0.0000000
duration:tagValueculture	-3.266486e+02	62.22886	75804	-5.2491497	0.0000002
duration:tagValuedance	-1.231117e+02	186.30807	75804	-0.6607963	0.5087449
duration:tagValuedesign	-4.184884e+02	63.56767	75804	-6.5833539	0.0000000
duration:tagValueentertainment	-5.232168e+02	68.73591	75804	-7.6119873	0.0000000
duration:tagValuefuture	-4.439324e+02	70.62145	75804	-6.2860843	0.0000000
duration:tagValueglobal issues	-3.264002e+02	60.45641	75804	-5.3989353	0.0000001
duration:tagValuehealth	-3.242299e+01	81.53956	75804	-0.3976351	0.6909003
duration:tagValuehumanity	-5.062720e+02	71.60771	75804	-7.0700773	0.0000000
duration:tagValueinnovation	-2.961098e+02	77.57997	75804	-3.8168337	0.0001353
duration:tagValuepsychology	4.228786e+02	84.49134	75804	5.0049929	0.0000006
duration:tagValuescience	2.747452e+02	60.04547	75804	4.5756187	0.0000048
duration:tagValuesocial change	-6.019794e+02	69.09053	75804	-8.7129070	0.0000000
duration:tagValuesociety	-7.627092e+02	71.81968	75804	-10.6197808	0.0000000
duration:tagValuetechnology	-4.409539e+02	55.88277	75804	-7.8906948	0.0000000
duration:tagValueTED Fellows	3.692569e+01	138.03617	75804	0.2675073	0.7890793
duration:tagValueTEDx	-2.619173e+02	76.06716	75804	-3.4432376	0.0005751
tagValuecommunication:languages	-5.225462e+03	3325.07075	75804	-1.5715339	0.1160629
tagValueculture:languages	3.015650e+04	2326.55177	75804	12.9618887	0.0000000
tagValuedance:languages	1.533547e+04	5988.44999	75804	2.5608420	0.0104438
tagValuedesign:languages	-1.577909e+04	2471.42704	75804	-6.3846069	0.0000000
tagValueentertainment:languages	-7.209536e+03	2584.75661	75804	-2.7892515	0.0052843
tagValuefuture:languages	-1.059869e+04	3112.54582	75804	-3.4051519	0.0006616
tagValueglobal issues:languages	6.647680e+03	2295.22547	75804	2.8963080	0.0037769
tagValuehealth:languages	5.748961e+03	3104.86938	75804	1.8515950	0.0640879
tagValuehumanity:languages	-6.001093e+03	3195.91046	75804	-1.8777414	0.0604204
tagValueinnovation:languages	-1.376655e+04	2687.84876	75804	-5.1217730	0.0000003
tagValuepsychology:languages	4.658742e+04	2913.24552	75804	15.9915872	0.0000000
tagValuescience:languages	1.096380e+04	2251.22775	75804	4.8701423	0.0000011
tagValuesocial change:languages	-1.482568e+04	3051.46026	75804	-4.8585538	0.0000012
tagValuesociety:languages	-1.416986e+04	2915.00988	75804	-4.8610006	0.0000012
tagValuetechnology:languages	-1.002310e+04	2136.38132	75804	-4.6916266	0.0000027
tagValueTED Fellows:languages	-3.013088e+04	3615.96102	75804	-8.3327444	0.0000000
tagValueTEDx:languages	-3.225885e+03	2220.82611	75804	-1.4525610	0.1463498
tagValuecommunication:comments	1.589423e+03	126.10501	75804	12.6039650	0.0000000
tagValueculture:comments	-6.675631e+02	63.90025	75804	-10.4469547	0.0000000
tagValuedance:comments	3.797328e+03	96.13438	75804	39.5002065	0.0000000

	Value	Std.Error	DF	t-value	p-value
tagValuedesign:comments	-4.526271e+02	85.06522	75804	-5.3209425	0.0000001
tagValueentertainment:comments	-1.445626e+02	116.37145	75804	-1.2422510	0.2141478
tagValuefuture:comments	-3.484828e+02	157.81089	75804	-2.2082306	0.0272312
tagValueglobal issues:comments	-1.584602e+03	68.16836	75804	-23.2454201	0.0000000
tagValuehealth:comments	-1.920227e+02	122.68201	75804	-1.5652069	0.1175386
tagValuehumanity:comments	4.461750e+01	193.68624	75804	0.2303597	0.8178129
tagValueinnovation:comments	-4.199268e+02	120.34742	75804	-3.4892876	0.0004846
tagValuepsychology:comments	-1.415880e+03	79.15485	75804	-17.8874731	0.0000000
tagValuescience:comments	-2.041610e+03	66.54184	75804	-30.6816012	0.0000000
tagValuesocial change:comments	1.019973e+03	106.04238	75804	9.6185448	0.0000000
tagValuesociety:comments	8.075945e+01	167.69963	75804	0.4815720	0.6301114
tagValuetechnology:comments	-2.182060e+02	78.25390	75804	-2.7884354	0.0052976
tagValueTED Fellows:comments	-3.421850e+02	196.81776	75804	-1.7385878	0.0821114
tagValueTEDx:comments	6.247005e+02	80.42092	75804	7.7678854	0.0000000

Table 7: Fixed effects of the comments mixed model

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-300.5858206	13.7904216	78159	-21.7967100	0.0000000
duration	0.1513540	0.0081839	78159	18.4940626	0.0000000
tagValuechildren	21.3793520	19.5547427	78159	1.0933078	0.2742621
tagValuecommunication	-65.7608807	19.4657554	78159	-3.3782856	0.0007297
tagValuecreativity	51.3471187	18.1231134	78159	2.8332394	0.0046091
tagValueculture	-185.1848110	15.4810573	78159	-11.9620261	0.0000000
tagValuedance	25.0851992	42.4521739	78159	0.5909049	0.5545858
tagValueeducation	56.5174765	18.3824506	78159	3.0745344	0.0021090
tagValueentertainment	37.4334202	16.8737066	78159	2.2184468	0.0265272
tagValuefuture	22.2249687	17.5626070	78159	1.2654709	0.2057063
tagValueglobal issues	-32.3395209	14.7451339	78159	-2.1932335	0.0282935
tagValueGod	-1297.0162749	38.4700154	78159	-33.7149924	0.0000000
tagValuehealth	22.3776320	18.6731766	78159	1.1983838	0.2307713
tagValuehumanity	-41.4809291	17.9264257	78159	-2.3139543	0.0206728
tagValueinnovation	57.9612810	17.1973784	78159	3.3703556	0.0007511
tagValueneuroscience	41.0715970	26.7820572	78159	1.5335490	0.1251447
tagValuephilosophy	-216.9157222	26.2950737	78159	-8.2492913	0.0000000
tagValuereligion	-988.1673883	27.0011814	78159	-36.5971909	0.0000000
tagValuescience	118.6514361	14.3482446	78159	8.2694043	0.0000000
tagValuesocial change	-50.5181648	16.6331157	78159	-3.0372040	0.0023886
tagValuesociety	-47.6033783	15.6919913	78159	-3.0336098	0.0024173
tagValuetechnology	29.4105677	13.7249782	78159	2.1428499	0.0321282
tagValueTED Fellows	198.9226350	23.6694172	78159	8.4042050	0.0000000
tagValueTEDx	-1.2779932	15.0917074	78159	-0.0846818	0.9325146
languages	10.2861284	0.3124239	78159	32.9236249	0.0000000
views	0.0000578	0.0000009	78159	62.0413025	0.0000000
duration:tagValuechildren	-0.0208126	0.0131165	78159	-1.5867405	0.1125755
duration:tagValuecommunication	0.0262415	0.0126152	78159	2.0801491	0.0375151
duration:tagValuecreativity	-0.0260181	0.0124793	78159	-2.0848995	0.0370817
duration:tagValueculture	0.0956168	0.0092163	78159	10.3747465	0.0000000
duration:tagValuedance	-0.0568943	0.0283864	78159	-2.0042831	0.0450432
duration:tagValueeducation	-0.0106075	0.0112952	78159	-0.9391090	0.3476777
duration:tagValueentertainment	-0.0196821	0.0103085	78159	-1.9093157	0.0562250

	Value	Std.Error	DF	t-value	p-value
duration:tagValuefuture	-0.0095586	0.0108843	78159	-0.8781982	0.3798389
duration:tagValueglobal issues	0.0084614	0.0091892	78159	0.9207957	0.3571599
duration:tagValueGod	0.8279863	0.0336792	78159	24.5845141	0.0000000
duration:tagValuehealth	-0.0348963	0.0122977	78159	-2.8376201	0.0045463
duration:tagValuehumanity	-0.0061303	0.0109558	78159	-0.5595477	0.5757896
duration:tagValueinnovation	-0.0453190	0.0120497	78159	-3.7610040	0.0001694
duration:tagValueneuroscience	-0.0517792	0.0191480	78159	-2.7041500	0.0068494
duration:tagValuephilosophy	0.1310418	0.0149949	78159	8.7391056	0.0000000
duration:tagValuereligion	0.5818564	0.0198860	78159	29.2595534	0.0000000
duration:tagValuescience	-0.1828987	0.0090770	78159	-20.1497657	0.0000000
duration:tagValuesocial change	-0.0079097	0.0103266	78159	-0.7659550	0.4437054
duration:tagValuesociety	-0.0068622	0.0106426	78159	-0.6447837	0.5190693
duration:tagValuetechnology	-0.0239876	0.0084853	78159	-2.8269609	0.0047004
duration:tagValueTED Fellows	-0.0776909	0.0220273	78159	-3.5270321	0.0004205
duration:tagValueTEDx	-0.0162040	0.0116046	78159	-1.3963422	0.1626154
tagValuechildren:languages	-1.9135176	0.4759869	78159	-4.0201060	0.0000582
tagValuecommunication:languages	1.8677064	0.4942242	78159	3.7790667	0.0001575
tagValuecreativity:languages	-2.6316512	0.4203869	78159	-6.2600691	0.0000000
tagValueculture:languages	3.4150707	0.3653444	78159	9.3475380	0.0000000
tagValuedance:languages	-1.7509996	0.9155935	78159	-1.9124204	0.0558259
tagValueeducation:languages	-3.2947188	0.4500920	78159	-7.3201015	0.0000000
tagValueentertainment:languages	-1.1416296	0.4084233	78159	-2.7952116	0.0051878
tagValuefuture:languages	0.6553030	0.4887823	78159	1.3406848	0.1800267
tagValueglobal issues:languages	0.6836310	0.3654813	78159	1.8704950	0.0614188
tagValueGod:languages	24.8419794	1.5098150	78159	16.4536577	0.0000000
tagValuehealth:languages	0.3768343	0.4725207	78159	0.7974979	0.4251643
tagValuehumanity:languages	1.9344518	0.5058463	78159	3.8241885	0.0001313
tagValueinnovation:languages	-1.2295255	0.4318263	78159	-2.8472685	0.0044108
tagValueneuroscience:languages	0.5541349	0.6875189	78159	0.8059923	0.4202498
tagValuephilosophy:languages	5.2136997	0.7047766	78159	7.3976624	0.0000000
tagValuereligion:languages	14.7510802	0.7532711	78159	19.5826972	0.0000000
tagValuescience:languages	0.5163750	0.3569929	78159	1.4464574	0.1480530
tagValuesocial change:languages	1.9363024	0.4521904	78159	4.2820508	0.0000185
tagValuesociety:languages	1.8156901	0.4413701	78159	4.1137586	0.0000390
tagValuetechnology:languages	0.1832093	0.3387449	78159	0.5408475	0.5886142
tagValueTED Fellows:languages	-4.1066248	0.5832756	78159	-7.0406245	0.0000000
tagValueTEDx:languages	0.6187117	0.3456264	78159	1.7901170	0.0734390
tagValuechildren:views	0.0000198	0.0000012	78159	16.0978630	0.0000000
tagValuecommunication:views	-0.0000141	0.0000014	78159	-10.0915911	0.0000000
tagValuecreativity:views	0.0000177	0.0000011	78159	15.5242308	0.0000000
tagValueculture:views	0.0000003	0.0000010	78159	0.3318356	0.7400143
tagValuedance:views	0.0000208	0.0000014	78159	14.4608624	0.0000000
tagValueeducation:views	0.0000189	0.0000012	78159	15.8802034	0.0000000
tagValueentertainment:views	0.0000019	0.0000015	78159	1.3084426	0.1907271
tagValuefuture:views	-0.0000391	0.0000042	78159	-9.2666052	0.0000000
tagValueglobal issues:views	0.0000105	0.0000017	78159	6.0461723	0.0000000
tagValueGod:views	-0.0000093	0.0000078	78159	-1.1908103	0.2337317
tagValuehealth:views	-0.0000116	0.0000017	78159	-6.8117813	0.0000000
tagValuehumanity:views	-0.0000139	0.0000025	78159	-5.6456628	0.0000000
tagValueinnovation:views	-0.0000067	0.0000021	78159	-3.2731984	0.0010638
tagValueneuroscience:views	-0.0000241	0.0000033	78159	-7.3998639	0.0000000
tagValuephilosophy:views	-0.0000133	0.0000049	78159	-2.7185984	0.0065574

	Value	Std.Error	DF	t-value	p-value
tagValuereligion:views	0.0000630	0.0000046	78159	13.7308411	0.0000000
tagValuescience:views	0.0000022	0.0000012	78159	1.7722025	0.0763648
tagValuesocial change:views	-0.0000050	0.0000013	78159	-3.7714362	0.0001624
tagValuesociety:views	-0.0000083	0.0000020	78159	-4.1951393	0.0000273
tagValuetechnology:views	-0.0000190	0.0000014	78159	-13.9637252	0.0000000
tagValueTED Fellows:views	-0.0000356	0.0000060	78159	-5.8819976	0.0000000
tagValueTEDx:views	-0.0000038	0.0000010	78159	-3.7292278	0.0001922

Conclusion

After dealing with reshaping of the data, the use of LASSO regression and mixed models proved to be a straight forward approach for analyzing this data set.

The use of a positive/negative ratings ratio as a measure on its own was not a clear measure for popularity in terms of views or comments. In terms of views and comments positive ratings have a positive relationship with the number of views/comments on a given talk. In terms of tags, shows relating to drones, magic and body language have more views, while shows relating to philosophy, personality and statistics (shockingly) have less views. In terms of comments, TED talks with tags about atheism, religion and G-d are some of the largest predictors.

By using a mixed model, the interactions with the tags provide insight into which talks receive more views/comments.

References

1. Statistical Society of Canada, What Predicts The Popularity Of Ted Talks?, https://ssc.ca/en/case-study/case-study-2-what-predicts-popularity-ted-talks
2. TED, https://www.ted.com
3. Kaggle. TED Talks, Data about TED Talks on the TED.com website until September 21st, 2017. Rounak Banik, https://www.kaggle.com/rounakbanik/ted-talks

Code Appendix

```
library(tidyverse)
library(stringr)
library(stringi)
library(glmnet)
library(lme4)
library(jsonlite)
library(nlme)
library(yaml)
library(tidyr)
dt <- readr::read_csv("./ted_main.csv")

# No missing data
naniar::vis_miss(dt) + theme(axis.text.x = element_text(angle = 90))
```

```r
######################## Data Engineering ##

ratingsDf <- dt$ratings %>%
    lapply(function(x) read_yaml(text = x)) %>%
    lapply(function(x) do.call(rbind, x))

# Using a for loop because vectorizing is hard. Don't make
# fun of me

for (i in 1:length(ratingsDf)) {

    ratingsDf[[i]] <- ratingsDf[[i]] %>%
        as.data.frame.matrix() %>%
        transmute(ratingID = id, ratingTag = name, count = count,
            url = rep(dt$url[i], length(ratingsDf[[i]][, 1])))

}

ratingsDf <- do.call(rbind, ratingsDf) %>%
    unnest()

# Manually editing JSON Files to be read into R The titles
# may differ as such
setwd("Proj2JsonFiles")
for (i in 1:length(dt$related_talks)) {

    write(dt$related_talks[i] %>%
        str_remove_all("(?<=\\w)\\'(?=\\w)") %>%
        str_remove_all("’") %>%
        str_remove_all("\\\\'") %>%
        str_replace("é", "e") %>%
        stri_trans_general("latin-ascii"), paste0(i, ".json"))
}

relatedTalksJSON <- list()

for (i in 1:length(dt$related_talks)) {
    tryCatch({
        relatedTalksJSON[[i]] <- read_yaml(paste0(i, ".json"))
    }, error = function(e) {
        tryCatch({
            relatedTalksJSON[[i]] <- read_yaml(text = dt$related_talks[i] %>%
                str_remove_all("\\'"))
        }, error = function(f) {
            relatedTalksJSON[[i]] <- read_yaml(text = dt$related_talks[i] %>%
                str_remove_all("\\\"") %>%
                str_remove_all("\\\\'") %>%
                stri_trans_general("latin-ascii"))
        }, finally = i)
    }, finally = i)
}
```

```r
for (i in 1:length(relatedTalksJSON)) {

    relatedTalksJSON[[i]] <- relatedTalksJSON[[i]] %>%
        lapply(function(y) as.data.frame(y) %>%
            mutate(url = dt$url[i]))


}

relatedTalksJSONDf <- list()
for (i in 1:length(relatedTalksJSON)) {
    relatedTalksJSONDf[[i]] <- do.call(rbind, relatedTalksJSON[i])
}

relatedTalksJSONDf <- relatedTalksJSONDf %>%
    lapply(function(x) t(x))
relatedTalksJSONDf <- do.call(rbind, do.call(rbind, relatedTalksJSONDf)) %>%
    transmute(related_talks_id = id, related_talks_hero = hero,
        related_talks_speaker = speaker, related_talks_title = title,
        related_talks_duration = duration, related_talks_slug = slug,
        related_talks_view_count = viewed_count, url = url)


tagsDf <- dt$tags %>%
    lapply(function(x) read_yaml(text = x) %>%
        as_tibble())

for (i in 1:length(tagsDf)) {
    tagsDf[[i]] <- tagsDf[[i]] %>%
        transmute(tagValue = value, url = rep(dt$url[i], length(tagsDf[[i]][,
            1])))
}

tagsDf <- do.call(rbind, tagsDf)


# Need to fix dates
fullyJoinedDf <- ratingsDf %>%
    left_join(tagsDf, by = "url") %>%
    left_join(dt, by = "url") %>%
    select(!c(description, tags, related_talks, ratings)) %>%
    mutate(film_date = lubridate::as_datetime(film_date), published_date =
     ↪  lubridate::as_datetime(published_date))


# Working dataframe
dtt <- fullyJoinedDf %>%
    select(!ratingID) %>%
    pivot_wider(names_from = ratingTag, values_from = count,
        names_prefix = "rating_") %>%
    pivot_wider(names_from = tagValue, values_from = tagValue,
        names_prefix = "tag_")
```

```r
dtt <- dtt %>%
    mutate(across(names(dtt)[grepl("tag_", names(dtt))], ~ifelse(!is.na(.x),
        1, 0)))


############ Analysis #

y <- dtt$views
X <- data.matrix(dtt[, -which(names(dtt) %in% c("views"))])


# perform k-fold cross-validation to find optimal lambda
# value
cv_model_views <- cv.glmnet(X, y, alpha = 1)
# find optimal lambda value that minimizes test MSE
best_lambda <- cv_model_views$lambda.min

# produce plot of test MSE by lambda value
plot(cv_model_views, main = expression("MSE vs " * lambda[views] *
    ""))


best_model_views <- glmnet(X, y, alpha = 1, lambda = best_lambda)

as.matrix(coef(best_model_views), rownames) %>%
    as.data.frame.matrix() %>%
    filter(s0 != 0) %>%
    knitr::kable(caption = "Sparse Estimates for Ted Talk Popularity (in terms of
↪   Views)")

tibble(`Rating Tag` = unique(fullyJoinedDf$ratingTag), Classification = c("Good",
    "Good", "Good", "Good", "Bad", "Bad", "Good", "Good", "Bad",
    "Good", "Good", "Ambiguos", "Bad", "Bad")) %>%
    knitr::kable(caption = "Unique Rating Tags accross all Ted Talks")


# Including Good/Bad Ratio
dtt <- dtt %>%
    rowwise() %>%
    mutate(`Good/Bad Ratio` = sum(rating_Funny, rating_Beautiful,
        rating_Ingenious, rating_Courageous, rating_Informative,
        rating_Fascinating, rating_Persuasive,
↪   `rating_Jaw-dropping`)/sum(rating_Longwinded,
        rating_Confusing, rating_Unconvincing, rating_Obnoxious,
        rating_Inspiring))

dtt %>%
    group_by(`Good/Bad Ratio`) %>%
    arrange(-desc(`Good/Bad Ratio`), .by_group = T) %>%
    select(name, `Good/Bad Ratio`, views, comments, published_date) %>%
    filter(`Good/Bad Ratio` < 0.608357) %>%
    knitr::kable(caption = "Top 10 Worst Ted Talks")
```

```
dtt %>%
    group_by(`Good/Bad Ratio`) %>%
    arrange(desc(`Good/Bad Ratio`)) %>%
    select(name, `Good/Bad Ratio`, views, comments, published_date) %>%
    filter(`Good/Bad Ratio` >= 18.714285) %>%
    knitr::kable(caption = "Top 10 Best Ted Talks")


# Comments

y <- dtt$comments
X <- data.matrix(dtt[, -which(names(dtt) %in% c("comments"))])

# perform k-fold cross-validation to find optimal lambda
# value
cv_model_comments <- cv.glmnet(X, y, alpha = 1)
# find optimal lambda value that minimizes test MSE
best_lambda <- cv_model_comments$lambda.min

# produce plot of test MSE by lambda value
plot(cv_model_comments, main = expression("MSE vs " * lambda[comments] *
    ""))


best_model_comments <- glmnet(X, y, alpha = 1, lambda = best_lambda)

as.matrix(coef(best_model_comments), rownames) %>%
    as.data.frame.matrix() %>%
    filter(s0 != 0) %>%
    knitr::kable(caption = "Sparse Estimates for Ted Talk Popularity (in terms of
↪   comments)")


fullyJoinedDf %>%
    mutate(published_year = lubridate::year(published_date)) %>%
    group_by(published_year, tagValue) %>%
    summarize(total_views = sum(views)) %>%
    slice_max(order_by = total_views, n = 5) %>%
    ggplot() + geom_point(mapping = aes(x = as.factor(published_year),
    y = total_views, color = tagValue)) + facet_wrap(~tagValue) +
    ggtitle("Characteristics Predicting Popularity over time (views)") +
    theme(legend.position = "none", axis.text.x = element_text(angle = 90),
        axis.title.y = element_blank(), axis.title.x = element_blank())

fullyJoinedDf %>%
    mutate(published_year = lubridate::year(published_date)) %>%
    group_by(published_year, tagValue) %>%
    summarize(total_comments = sum(comments)) %>%
    slice_max(order_by = total_comments, n = 5) %>%
    ggplot() + geom_point(mapping = aes(x = as.factor(published_year),
    y = total_comments, color = tagValue)) + facet_wrap(~tagValue) +
```

```r
    ggtitle("Characteristics Predicting Popularity over time (comments)") +
    theme(legend.position = "none", axis.text.x = element_text(angle = 90),
        axis.title.y = element_blank(), axis.title.x = element_blank())

set.seed(6627)
library(nlme)
fit_views <- lme(views ~ duration + tagValue + languages + comments +
    duration * tagValue + languages * tagValue + comments * tagValue,
    data = fullyJoinedDf %>%
        filter(tagValue %in% c("culture", "psychology", "business",
            "entertainment", "dance", "technology", "global issues",
            "design", "science", "TEDx", "health", "TED Fellows",
            "communication", "innovation", "society", "future",
            "humanity", "social change")), random = ~1 | main_speaker)


sum_views <- summary(fit_views)

knitr::kable(sum_views$tTable, caption = "Fixed effects of the views mixed model")

set.seed(6627)

fit_comments <- lme(comments ~ duration + tagValue + languages +
    views + duration * tagValue + languages * tagValue + views *
    tagValue, data = fullyJoinedDf %>%
    filter(tagValue %in% c("culture", "creativity", "dance",
        "children", "education", "science", "technology", "religion",
        "God", "global issues", "entertainment", "business",
        "TEDx", "philosophy", "neuroscience", "health", "TED Fellows",
        "society", "social change", "future", "innovation", "humanity",
        "communication")), random = ~1 | main_speaker)


sum_comments <- summary(fit_comments)


knitr::kable(sum_comments$tTable, caption = "Fixed effects of the comments mixed model")
```