TED TALK'S ANALYSIS

Pavani Komati | CSE 5243 | 3rd Feb'18

Table of Contents

1.INTRODUCTION

- The TED Talks dataset contains a total of 2550 entries and the following 17 features.
- Source: https://www.kaggle.com/rounakbanik/ted-talks/data

```
Data columns (total 17 columns):
comments              2550 non-null int64
description           2550 non-null object
duration              2550 non-null int64
event                 2550 non-null object
film_date             2550 non-null int64
languages             2550 non-null int64
main_speaker          2550 non-null object
name                  2550 non-null object
num_speaker           2550 non-null int64
published_date        2550 non-null int64
ratings               2550 non-null object
related_talks         2550 non-null object
speaker_occupation    2544 non-null object
tags                  2550 non-null object
title                 2550 non-null object
url                   2550 non-null object
views                 2550 non-null int64
dtypes: int64(7), object(10)
memory usage: 338.8+ KB
```

- The number of first level comments made on the talk.
- A blurb of what the talk is about.
- The duration of the talk in seconds.
- The TED/TEDx event where the talk took place.
- The Unix timestamp of the filming.
- The number of languages in which the talk is available.
- The first named speaker of the talk.
- The official name of the TED Talk. (title and the speaker)
- The number of speakers in the talk.
- The Unix timestamp for the publication of the talk
- A stringified dictionary of the various ratings of the talk
- A list of dictionaries of recommended talks to watch next.
- The occupation of the main speaker.
- The themes associated with the talk.
- The title of the talk
- The URL of the talk.
- The number of views on the talk.

2.PRE-PROCESSING

It is observed that the film_date and published_date are in Unix timestamp format, which has to be converted to generic date format for easy visualization.

Converting the given unix timestamps to general date format.

	comments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	ratings	
0	4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	1140825600	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...	'http
1	265	With the same humor and humanity he exuded in ...	977	TED2006	1140825600	43	Al Gore	Al Gore: Averting the climate crisis	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 544}, {'i...	'http

	film_date	published_date
0	2006-02-25	2006-06-27 00:11:00
1	2006-02-25	2006-06-27 00:11:00
2	2006-02-24	2006-06-27 00:11:00
3	2006-02-26	2006-06-27 00:11:00
4	2006-02-22	2006-06-27 20:38:00

Also, we can see that each entry in rating column has different kinds of ratings. We can process the data to save information explicitly regarding different type of ratings.

Sample rating cell entry:

[{'id': 22, 'name': 'Fascinating', 'count': 84}, {'id': 1, 'name': 'Beautiful', 'count': 330}, {'id': 25, 'name': 'OK', 'count': 30}, {'id': 10, 'name': 'Inspiring', 'count': 100}, {'id': 23, 'name': 'Jaw-dropping', 'count': 31}, {'id': 21, 'name': 'Unconvincing', 'count': 8}, {'id': 11, 'name': 'Longwinded', 'count': 7}, {'id': 2, 'name': 'Confusing', 'count': 8}, {'id': 9, 'name': 'Ingenious', 'count': 6}, {'id': 26, 'name': 'Obnoxious', 'count': 4}, {'id': 3, 'name': 'Courageous', 'count': 22}, {'id': 7, 'name': 'Funny', 'count': 1}, {'id': 8, 'name': 'Informative', 'count': 0}, {'id': 24, 'name': 'Persuasive', 'count': 0}]

Processing →

```
{'Beautiful',
 'Confusing',
 'Courageous',
 'Fascinating',
 'Funny',
 'Informative',
 'Ingenious',
 'Inspiring',
 'Jaw-dropping',
 'Longwinded',
 'OK',
 'Obnoxious',
 'Persuasive',
 'Unconvincing'}
```

These are the different possible rating types for the given data set.

The information regarding these comments can be saved separately for each ted talk and replace the existing ratings column with the total number of comments.

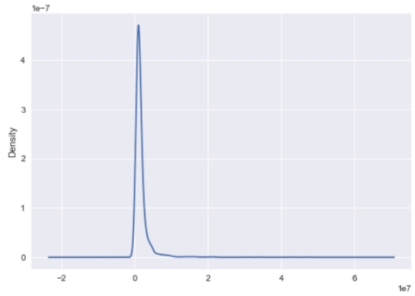
	ratings	Ingenious	OK	Informative	Jaw-dropping	Confusing	Fascinating	Unconvincing	Obnoxious	Funny	Inspiring	Courageous	Longwinded	Persuasive
0	93850	6073	1174	7346	4439	242	10581	300	209	19645	24924	3253	387	10704
1	2936	56	203	443	116	62	132	258	131	544	413	139	113	268
2	2824	183	146	395	54	27	166	104	142	964	230	45	78	230
3	3728	105	85	380	230	32	132	36	35	59	1070	760	53	460
4	25620	3202	248	5433	3736	72	4606	67	61	1390	2893	318	110	2542

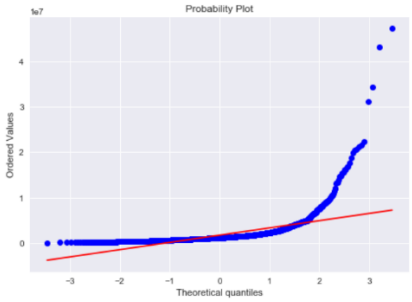
3.ANALYSIS

3.1 Most popular talks

	name	views
0	Ken Robinson: Do schools kill creativity?	47227110
1346	Amy Cuddy: Your body language may shape who yo...	43155405
677	Simon Sinek: How great leaders inspire action	34309432
837	Brené Brown: The power of vulnerability	31168150
452	Mary Roach: 10 things you didn't know about or...	22270883
1776	Julian Treasure: How to speak so that people w...	21594632
201	Jill Bolte Taylor: My stroke of insight	21190883
5	Tony Robbins: Why we do what we do	20685401
2114	James Veitch: This is what happens when you re...	20475972
1416	Cameron Russell: Looks aren't everything. Beli...	19787465
500	Dan Pink: The puzzle of motivation	18830983
1163	Susan Cain: The power of introverts	17629275
1036	Pamela Meyer: How to spot a liar	16861578

- Here popularity is determined by considering the view count. The most popular talks are taken with the view count in top 0.5%.
- The most popular talk show is the first talk show with 47 million views.
- From the list it is observed that there are only 4 talk shows that are slightly nearer to the most popular talk. Rest of the talks have as low as half of the popular talk view count.

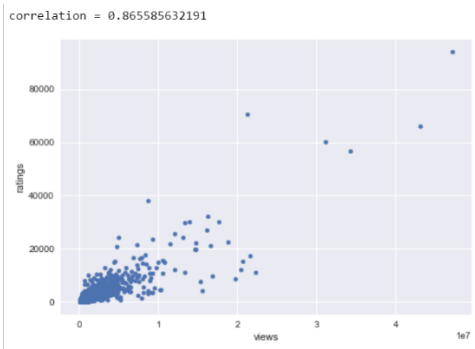


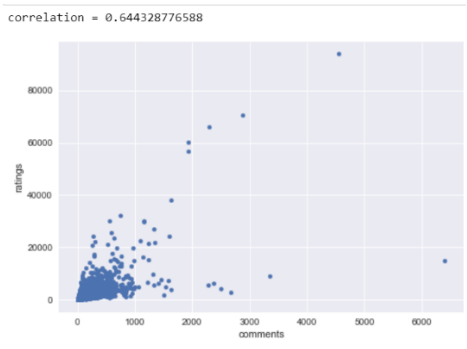


- With the normal distribution and quantile-quantile plots, it can be confirmed that nature of popularity i.e view count is **right skewed**.
- There are very few talks in the upper quartile region which implies, the talks with view count that matches with the max view count are very less.

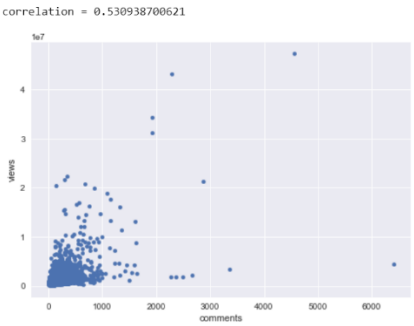
3.2 Relationship between numerical attributes

Correlation plot

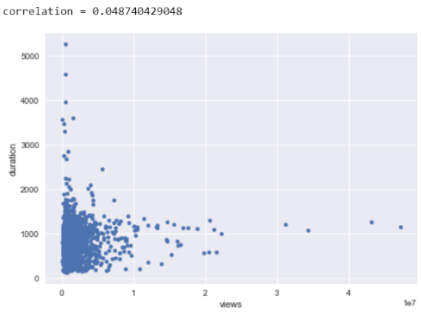




- From the correlation plot it can be observed that **(ratings : popularity)** of the talk have the highest correlation followed by **(comments : ratings)** and **(comments : popularity).**
- All the three correlation values are >0.5 , which can be considered as strong positive correlation



It can be observed that as the correlation value reduces, the relationship between the two attributes is not very strong.



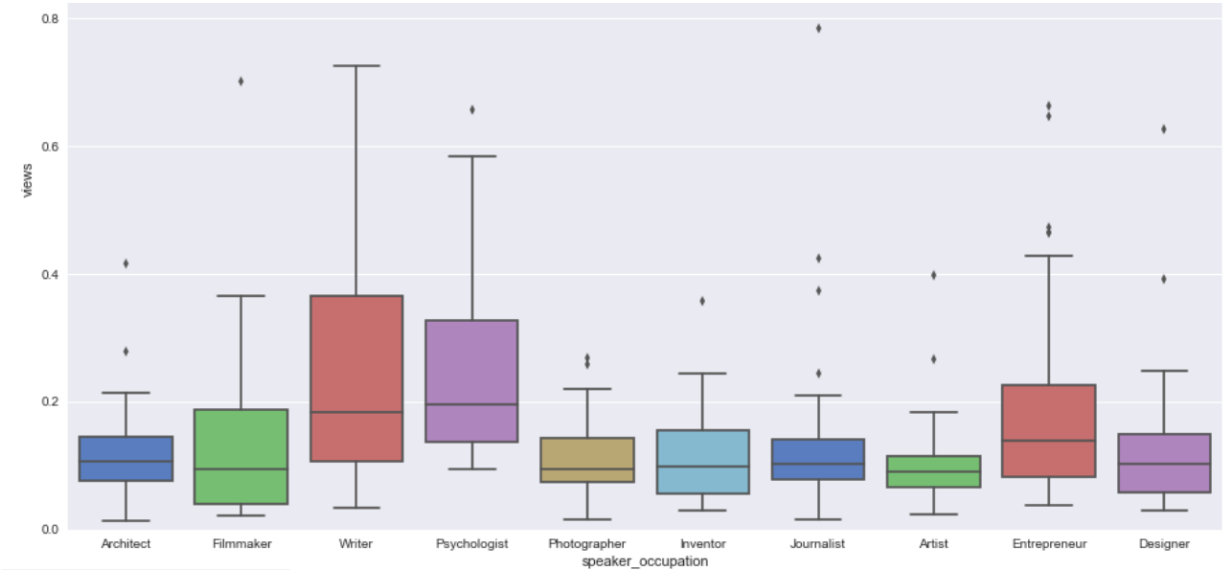
- The rest of the attributes have the least or neutral correlation between them.
- The popularity and duration of the talk have neutral or no correlation

3.3 Speaker occupation

	occupation	appearances
1426	Writer	45
83	Artist	34
413	Designer	34
753	Journalist	33
515	Entrepreneur	31
71	Architect	30
733	Inventor	27
1131	Psychologist	26
1011	Photographer	25
567	Filmmaker	21

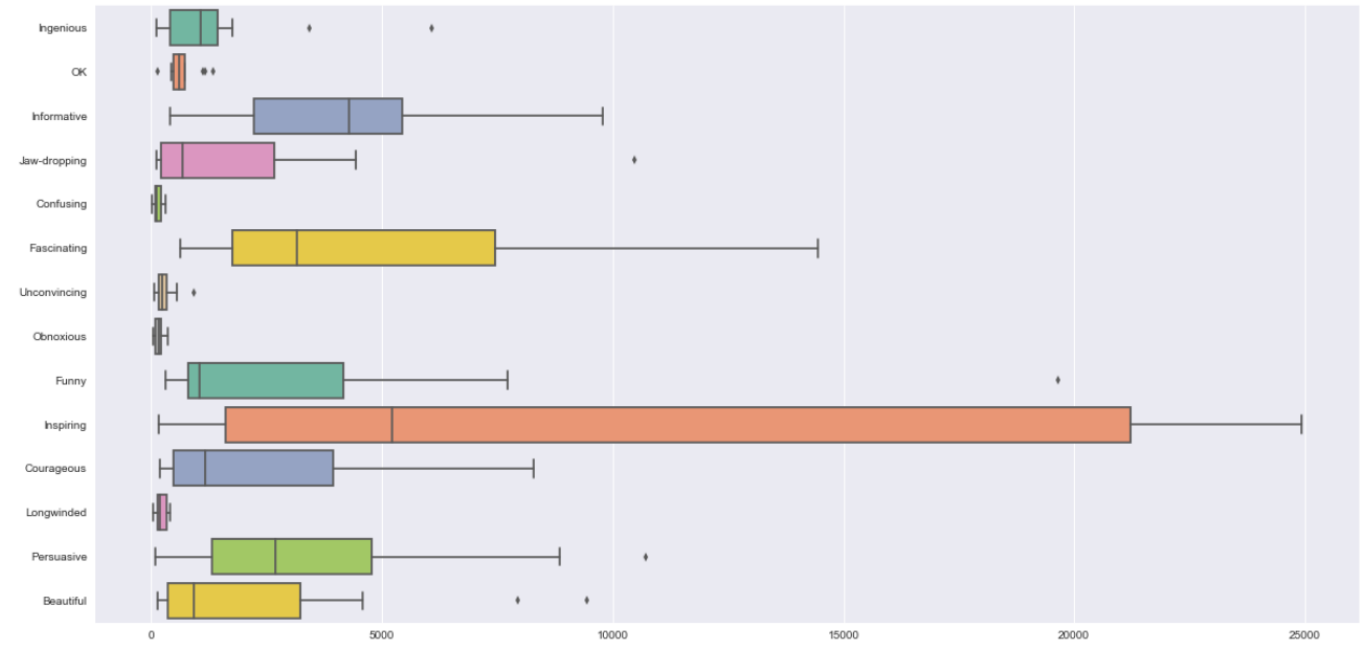
These are the top most occupations of the speakers.

- Writers are the most popular with 45 appearances
- Artists and Designers stand in the second place with 34 appearances
- The below plot of the view count of top ten occupations depicts that, Psychologists have the highest view count followed by Writers.
- Also, the Writers have highest range of values compared to other occupations.

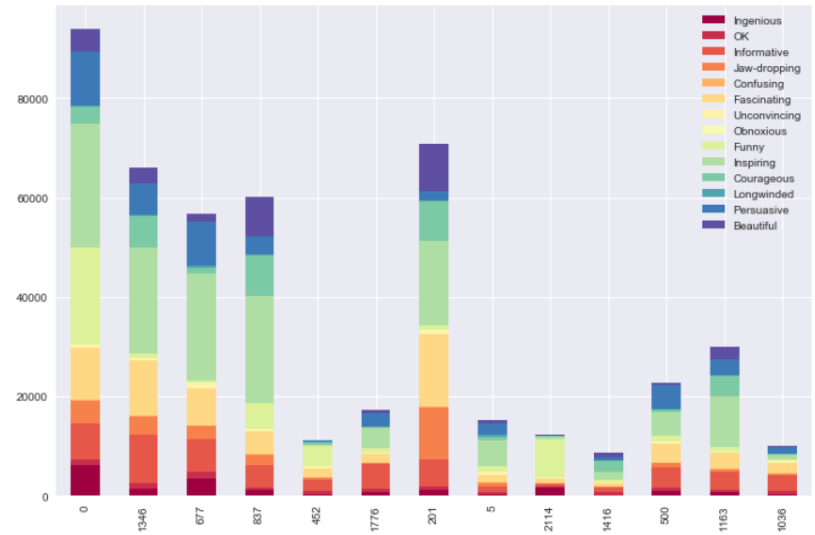


3.4 Ratings

- From the below box plot, it can be observed that 'Inspiring' type of comment is seen in most of the TED Talks followed by 'Informative' and 'Persuasive' rating.
- Very few TED talks have been given rating as 'Longwinded', 'Obnoxious' and 'Confusing'.
- This indicates that percentage of TED talks which are not liked by people is very less compared to the percentage of talks which people liked.

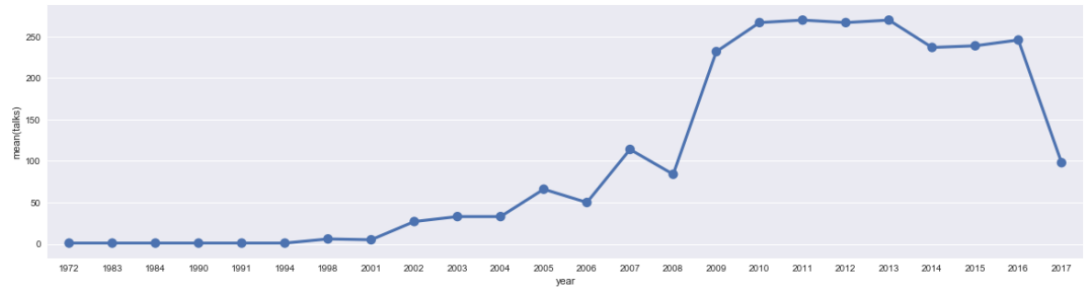


Ratings of Most Popular Talks



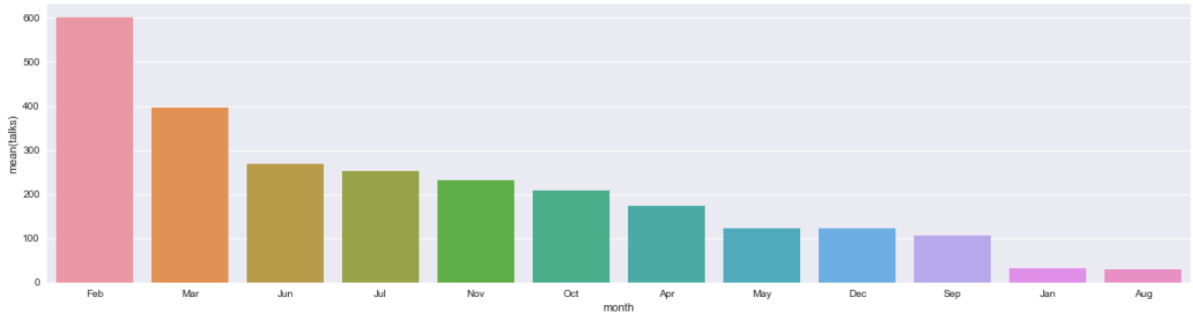
- Most of the popular shows have higher number of ratings of type 'Inspiring' followed by 'Fascinating'
- All these popular shows are rarely given rating as 'Unconvincing' or 'Obnoxious'

3.5 Film Date and Published Date



- The number of talks published are increased consistently over years.
- There is drastic change in number of talks published after the year 2008.
- As the data for the year 2017 is collected only up to September, the number of talks in 2017 could have been reduced for this reason.

Let's check the monthly distribution of data.



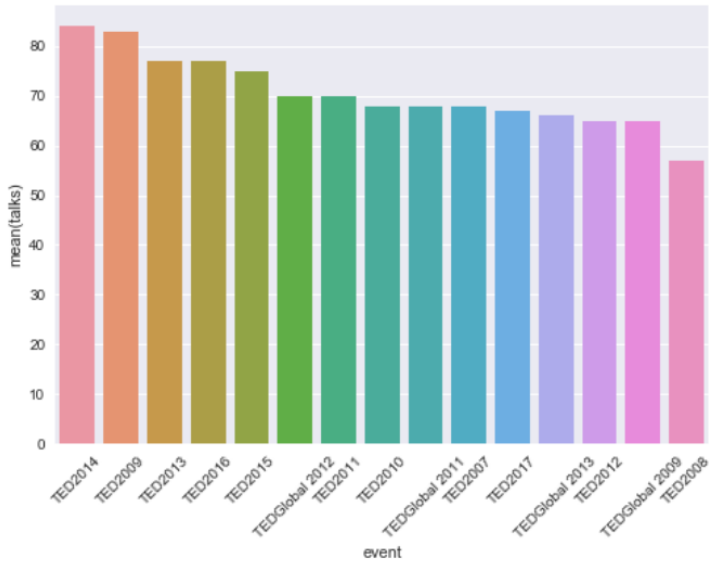
February has the highest talks, where as January and August have the least number of talks.

Difference between Published date and Film date

```
count    2550.000000
mean        8.212852
std        20.258725
min       -11.426723
25%         1.630428
50%         3.292824
75%         6.259888
max       456.012946
Name: diff, dtype: float64
```

- On an average the talks are published 8 months after the filming.
- Negative min indicates that there is some **noise** in the data.
- "Viktor Frankl: Why believe in others" is published after 38 years.

3.6 Events

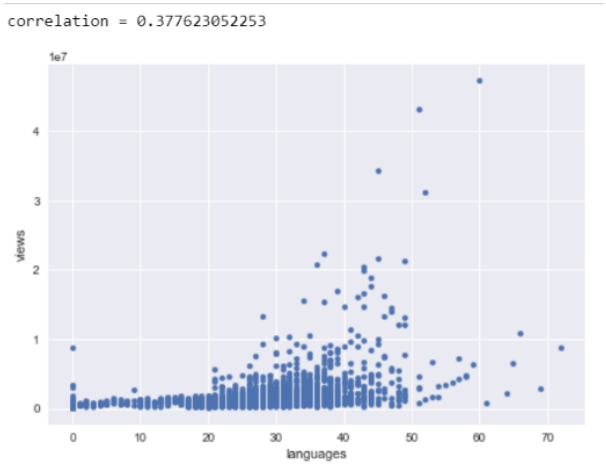



```
count    355.000000
mean       7.183099
std       15.265735
min        1.000000
25%        1.000000
50%        2.000000
75%        5.000000
max       84.000000
Name: talks, dtype: float64
```

- There are a total of 355 events.
- TED2014 has the highest number of talks-84
- On an average, each event has around 7 TED talks.

3.7 Languages

```
count    2550.000000
mean       27.326275
std         9.563452
min         0.000000
25%        23.000000
50%        28.000000
75%        33.000000
max        72.000000
Name: languages, dtype: float64
```



- On an average, there are 27 languages in each talk.
- There is a maximum of 72 languages in one of the show
- There is no strong correlation between popularity and number of languages. But, the number of views of many talks increase with the number of languages.

3.8 Tags

```
technology         727
science            567
global issues      501
culture            486
TEDx               450
design             418
business           348
entertainment      299
health             236
innovation         229
society            224
art                221
social change      218
future             195
```

```
count    416.000000
mean      46.043269
std       76.441760
min        1.000000
25%       10.000000
50%       21.000000
75%       50.000000
max      727.000000
Name: theme, dtype: float64
```

- There are a total of 416 different themes available in the tags.
- Technology appears the highest number of times in the talks followed by science
- On an average each theme appears in 46 talks.

3.8 Duration

```
count    2550.000000
mean      826.510196
std       374.009138
min       135.000000
25%       577.000000
50%       848.000000
75%      1046.750000
max      5256.000000
Name: duration, dtype: float64
```

- On an average, the talks have a duration of 826 minutes.
- The shortest talk is 'The ancestor or language' by Murray Gell-Man.
- The longest talk is 'Parrots, the universe and everything' by Douglas Adams

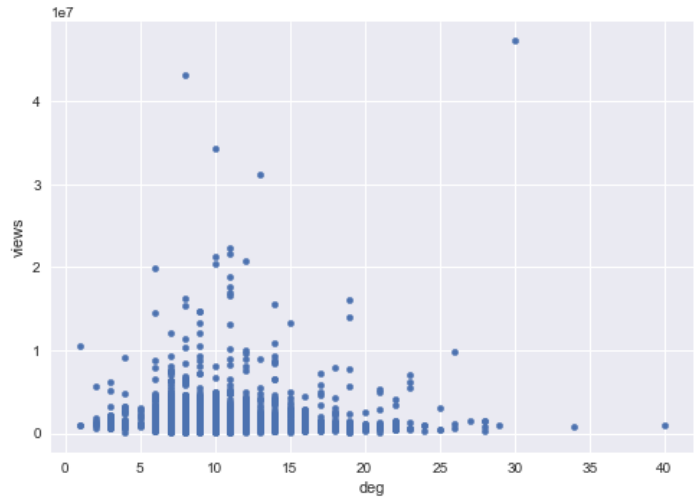
3.10 Related Talks

	name	deg
345	Gregory Petsko: The coming neurological epidemic	40
38	Ben Saunders: Why did I ski to the North Pole?	34
0	Ken Robinson: Do schools kill creativity?	30
107	Tierney Thys: Swim with the giant sunfish	29
812	Conrad Wolfram: Teaching kids real math with c...	28

```
count    13.000000
mean     11.846154
std       5.727800
min       6.000000
25%      10.000000
50%      11.000000
75%      11.000000
max      30.000000
Name: deg, dtype: float64
```

- 'The coming neurological epidemic' by Gregory Petsko has the highest number of references by other talks.
- On an average, every talk is related to 6 of other talks.

```
correlation = 0.0529250836774
```



- The plot and the correlation value indicate that there is no correlation between the popularity and number of related talks associated with particular talk.

References : https://www.kaggle.com/rounakbanik/ted-talks