Nile University
School of Computer Science and Information Technology
CSCI-322

Data Analysis

Fall 2020

How to Give the Perfect TED Talk?

Project report

Submitted by:

1. Hossam El-Din Sayed Hassan ID: 1710700
2. Menna-Allah Sayed Mostafa ID: 1710223

Under Supervision of

Dr. Mustafa El-Attar

Engr. Hossam El-Ghamry

Engr. Passant Amin

Table of Contents

Abstract

In this project, we aim to find if there are any factors, other than content, that would improve a TED talk and increase the number of views it has. First, we give a brief introduction and overview about TED conferences. Second, we discuss the method of data acquisition and a brief description of it. After that, we go through a deep and detailed analysis of the features we believe to be playing a significant role in determining the number of views for a talk and lastly, we discuss several regression techniques to predict how different features truly affect the number of views on a TED talk.

Introduction and Background Review

"Ideas worth spreading." This is the how TED describes and defines all the talks presented on its website. TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less). TED began in 1984 as a conference where Technology, Entertainment and Design converged, and today covers almost all topics — from science to business to global issues — in more than 100 languages. Meanwhile, independently run TEDx events help share ideas in communities around the world. As of 2015, TED and its sister TEDx chapters have published more than 4000 talks for free consumption by the masses and its speaker list boasts of the likes of Al Gore, Jimmy Wales, Shahrukh Khan and Bill Gates.

The main TED conference is held annually in Vancouver, British Columbia, Canada at the Vancouver Convention Centre. Prior to 2014, the conference was held in Long Beach, California, United States. TED events are also held throughout North America and in Europe, Asia, and Africa, offering live streaming of the talks. They address a wide range of topics within the research and practice of science and culture, often through storytelling. The speakers are given a maximum of 18 minutes to present their ideas in the most innovative and engaging ways they can.

Since June 2006, TED Talks have been offered for free viewing online, under an Attribution-Noncommercial-No Derivatives Creative Commons license, through TED.com. As of January 2018, over 2,600 TED Talks are freely available on the website. In June 2011, TED Talks' combined viewing figure stood at more than 500 million, and by November 2012, TED Talks had been watched over one billion times worldwide. TED Talks given by academics tend to be watched more online while art and design videos tend to be watched less than average.

The question we are asking is how one can give an outstanding talk at TED? TED can be thought of as the ultimatum of public speaking and is an excellent platform for delivering your ideas and thoughts. We believe that there are many factors that collaborate in making a perfect TED talk and public speaking such as duration of the topic, occupation of the speaker, the timing of the talk and many others. We aim to utilize different data analysis techniques really find out.

Data Acquisition and Description

We have used a python-based scrapper to scrap TED.com for all the information it is offering on the talks from different conferences made available on the website. We managed to collect data on different 4928 TED talks and videos (from TED-Ed which are not technically a talk) with the first talk being filmed on 1972 and the most recent one being on the 10th of December 2020 and published on 11th of January 2021. For each talk, we have acquired 49 attributes (features) shown below:

1.	talk__id	int64
2.	talk__name	object
3.	talk__description	object
4.	view_count	int64
5.	comment_count	float64
6.	duration	int64
7.	transcript	object
8.	video_type_name	object
9.	event	object
10.	number_of__speakers	int64
11.	speaker__id	float64
12.	speaker__name	object
13.	speaker__description	object
14.	speaker__who_he_is	object
15.	speaker__why_listen	object
16.	speaker__what_others_say	object
17.	speaker__is_published	object
18.	all_speakers_details	object
19.	is_talk_featured	bool
20.	has_talk_citation	object
21.	recording_date	object
22.	published_timestamp	object
23.	talks__tags	object
24.	number_of__tags	int64
25.	language	object
26.	native_language	object
27.	language_swap	bool
28.	is_subtitle_required	bool
29.	url__webpage	object
30.	url__audio	object
31.	url__video	object
32.	url__photo__talk	object
33.	url__photo__speaker	object
34.	url__subtitled_videos	object
35.	number_of__subtitled_videos	int64
36.	talk__download_languages	object
37.	number_of__talk__download_languages	int64
38.	talk__more_resources	object
39.	number_of__talk__more_resources	int64
40.	talk__recommendations__blurb	object
41.	talk__recommendations	object
42.	number_of__talk__recommendations	int64
43.	related_talks	object
44.	number_of__related_talks	int64
45.	intro_duration	float64
46.	ad_duration	float64
47.	post_ad_duration	float64

```
48.  external__duration                    float64
49.  external__start_time                  float64
```

We haver only chosen a subset of these columns for our analysis and renamed the columns to be more self-explanatory.

Data Cleaning

Luckily, the scraped data did not have a lot of null values, only one or two entries in the chosen columns that we have either decided to drop and/or substitute. Below is the cleaning that we have done for each column. A column not mentioned is a column that did not require any cleaning.

- We dropped the entries with null values as the speaker name and film date.
- The null values in the speaker's occupation column are substituted with "unknown."
- We have also removed all talks falling under TED-Ed event since they are short, animated videos and not technically "The Talk" we are aiming to analyze.
- All videos with zero views were replaced with the next minimum value of views since we have assumed that these videos had very low views that TED preferred not to provide them.

In some analysis, we did not consider all the data for the purpose of wanting to analyze the common case or did not want to include these outliers. But we did not remove them completely from the data frame as they represent valuable and valid sample to be considered.

Analysis

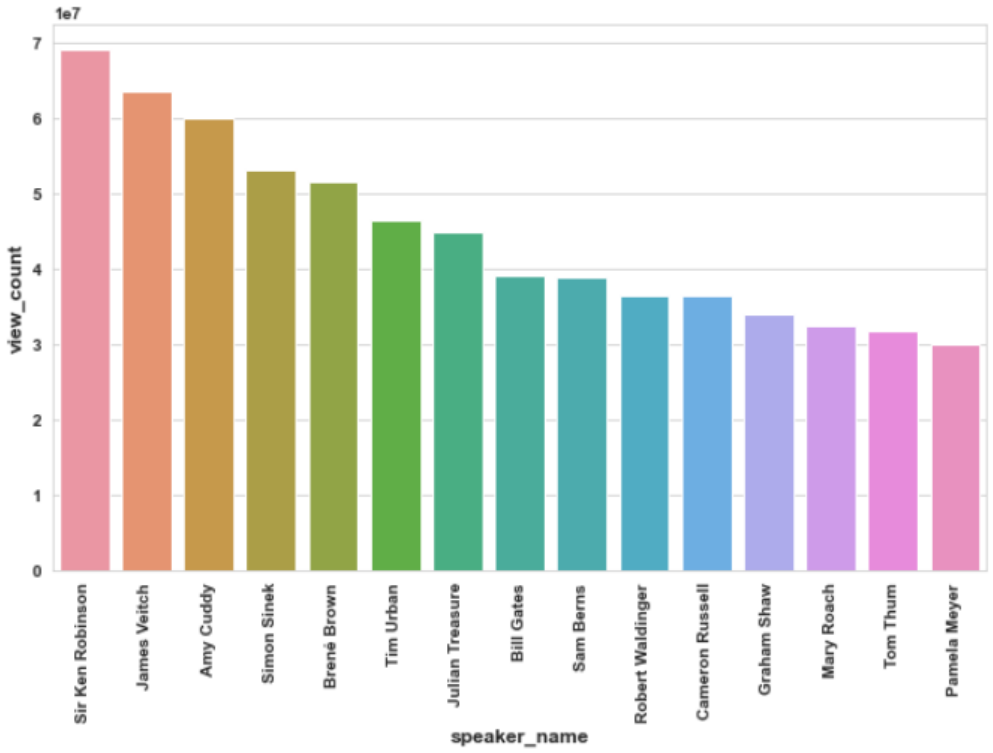
Most viewed talks

First step towards conducting the best TED-talk ever, is getting to know the competition. We will start with finding the most viewed talks so far on TED.

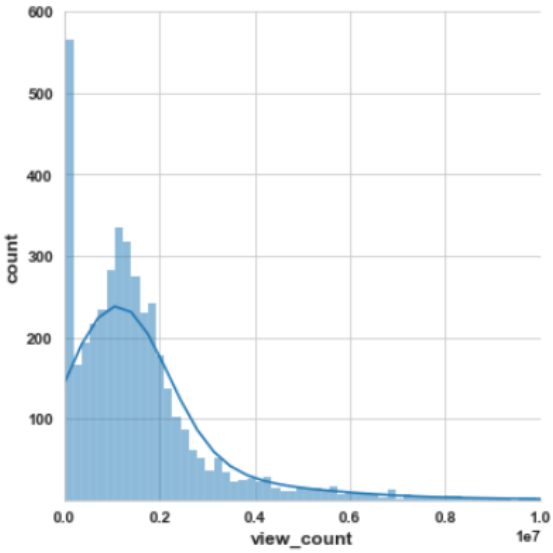
We have extracted the top 15 talks according to their views:

	name	speaker_name	view_count	published_date
0	Do schools kill creativity?	Sir Ken Robinson	69154230	2006-06-27
1	This is what happens when you reply to spam email	James Veitch	63556835	2020-10-23
2	Your body language may shape who you are	Amy Cuddy	59958589	2012-10-01
3	How great leaders inspire action	Simon Sinek	53105906	2010-05-04
4	The power of vulnerability	Brené Brown	51501858	2010-12-23
5	Inside the mind of a master procrastinator	Tim Urban	46492849	2016-03-15
6	How to speak so that people want to listen	Julian Treasure	44971966	2014-06-27
7	The next outbreak? We're not ready	Bill Gates	39167037	2015-04-03
8	My philosophy for a happy life	Sam Berns	38875621	2018-03-28
9	What makes a good life? Lessons from the longe...	Robert Waldinger	36500033	2015-12-23
10	Looks aren't everything. Believe me, I'm a model.	Cameron Russell	36376952	2013-01-16
11	Why people believe they can't draw	Graham Shaw	33919773	2018-03-28
12	10 things you didn't know about orgasm	Mary Roach	32371337	2009-05-20
13	The orchestra in my mouth	Tom Thum	31675280	2013-07-19
14	How to spot a liar	Pamela Meyer	30078569	2011-10-13

- Do schools kill creativity? by Sir Ken Robinson is the most viewed talk with almost 70 million views and nearly 20 million of them is in the years between 2018 till 2021
- Robinson's talk is closely followed by Amy Cuddy's talk on Your Body Language May Shape Who You Are.
- Bill Gates' talk about the pandemic is in 7th place with nearly 40 million views . This talk was not even in the top 10 talks in 2017 and only had 2.2 millions , despite the fact that there have been other talks on 2015 that made it to the top 10 in 2017



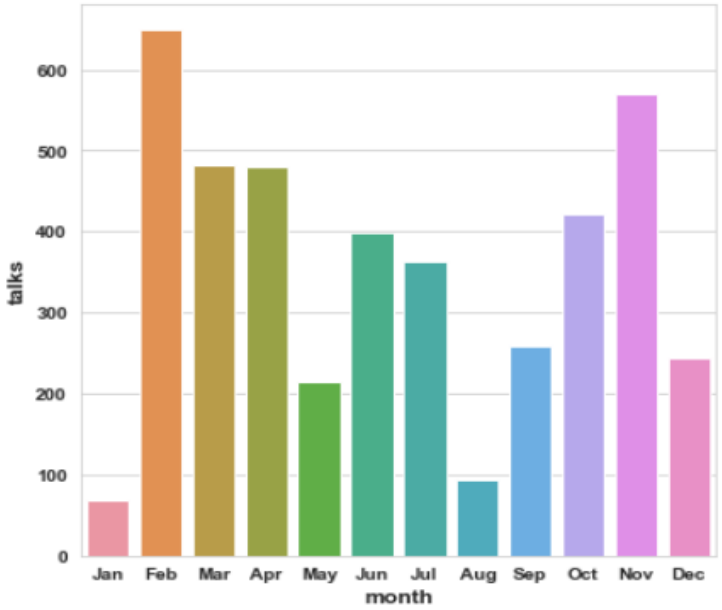
Then we investigated the summary statistics and the distribution of the views on various TED Talks.



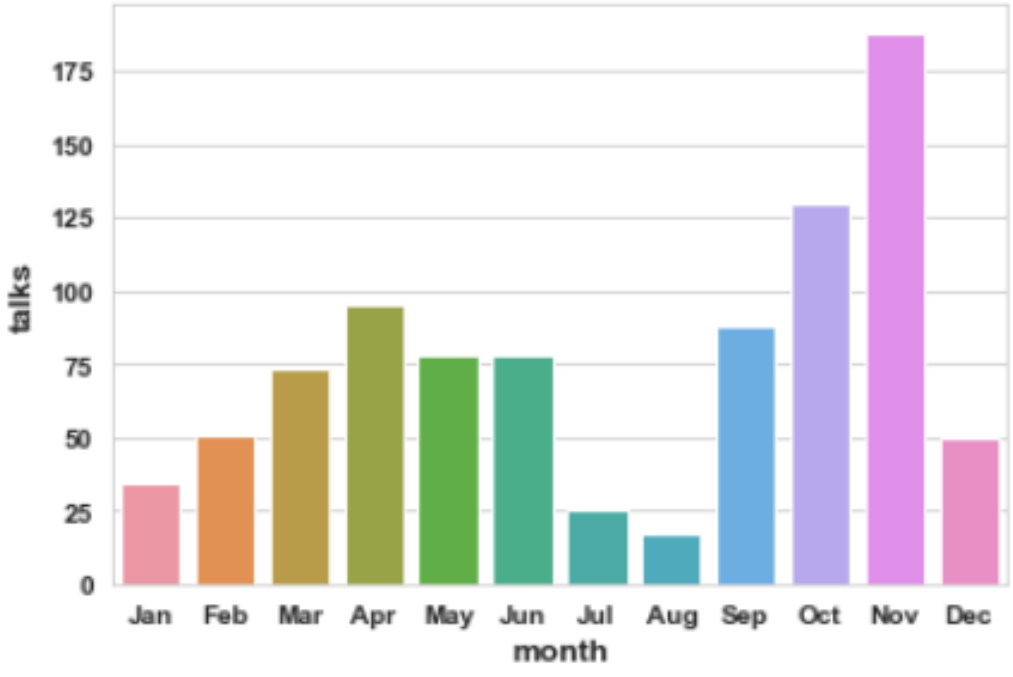
It is found that the distribution of the views is skewed to the left with most of the videos being less than 4 million views, In fact, it was found that The average number of views on TED Talk videos is nearly 2 million and the median is 1.3 million.

Analysis by month and year

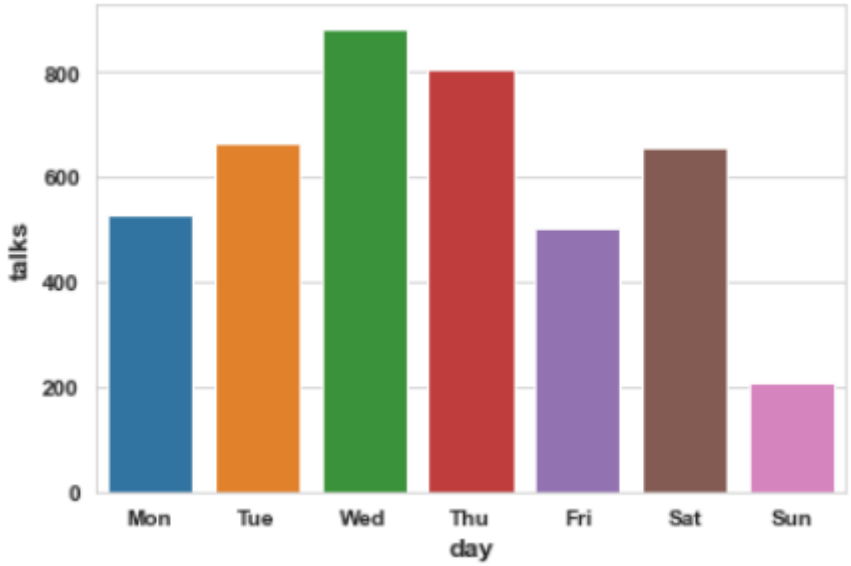
We wanted to see which months have the most frequent talks. So we plotted the number of talks in each month.



February is clearly the most popular month for TED Conferences followed by November whereas August and January are the least popular. February's popularity is largely due to the fact that the official TED Conferences are held in February. Since TEDx talks are more frequent we decided to analyze them separately.



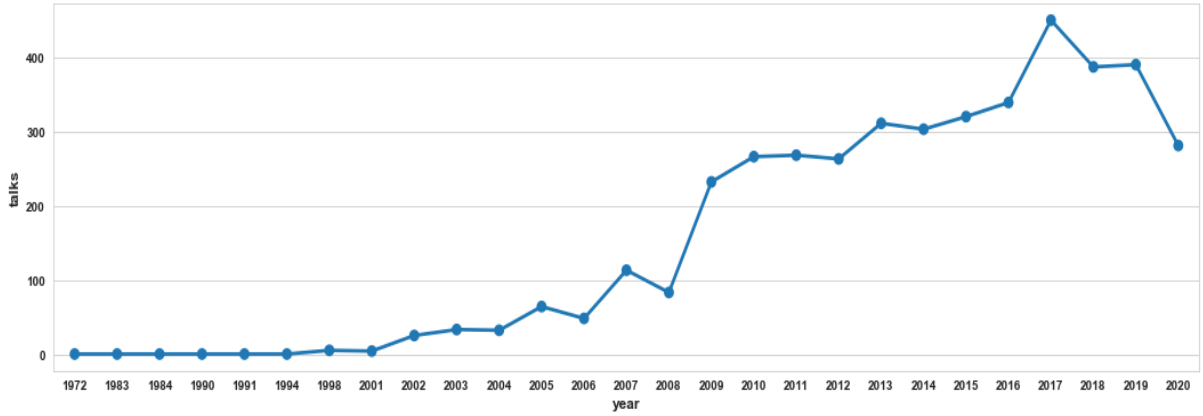
Afterwards, we seek to see which day has the greatest number of ted talks



The distribution is almost a bell curve shape with most days being Wednesday and Thursday.

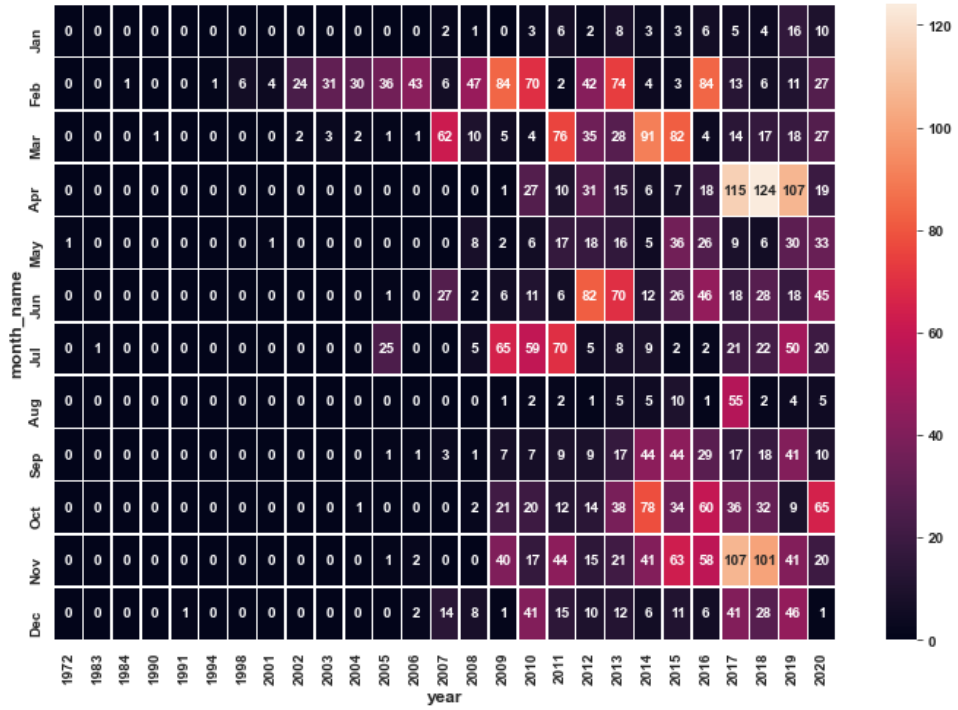
In conclusion, Most TED conferences are in February and November and take place on Wednesday and Thursday. It seems reasonable for someone who wants to give a TED talk be prepared at these times.

Finally, we observed the evolution and increase in TED talks over the years to really prove that having a successful TED talk will help greatly in making your ideas reach a large and growing audience,



It appears that the number of talks have been increasing over the years and truly spiked in 2009. It would be interesting if we could find the real purpose behind this spike. There are many events and breakthroughs such as Barack Obama becoming president and the introduction of blockchain. There is also a sharp decrease in 2020 because of the pandemic.

Finally, we constructed a heat map to see the number of talks by month and year. This will give us a good summary of the distribution of talks.

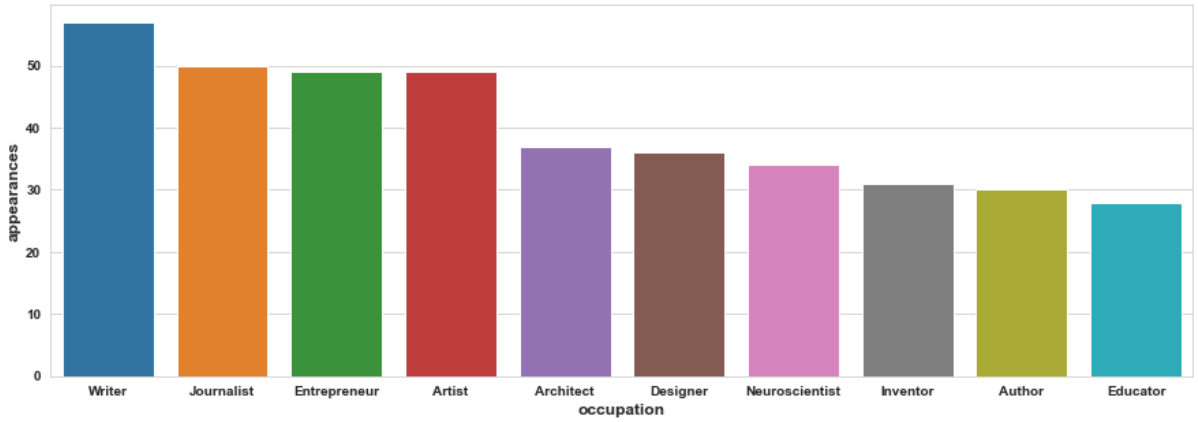


The most Popular Ted Talker

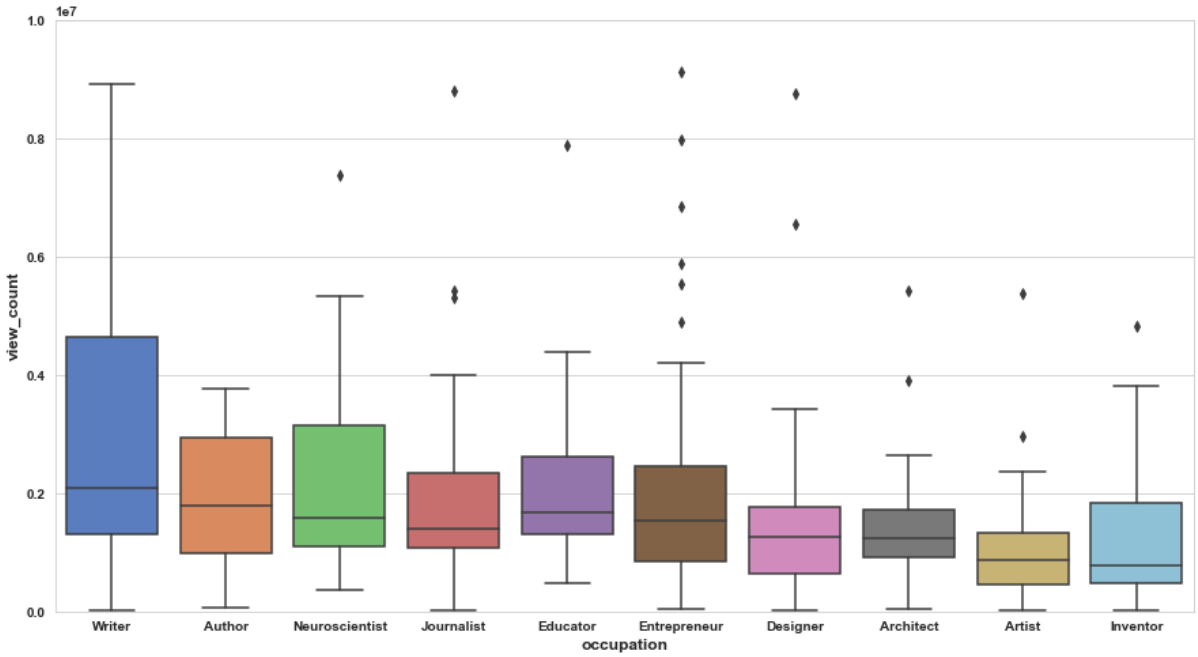
To really ace a TED Talk, it certainly would be beneficial to see what the traits of those who are frequently invited to give out TED talks. We have extracted the top speakers according to their number of appearances.

	main_speaker	appearances
1802	Juan Enriquez	10
2319	Matt Walker	10
1340	Hans Rosling	9
1298	Greg Gage	8
2220	Marco Tempest	7
440	Bill Gates	7
1480	Jacqueline Novogratz	6
2906	Rives	6
728	Dan Ariely	6
2383	Michael Green	5

Afterwards, we wanted to see which occupation has the most appearances and views. This can help us figure which occupation is more likely to be invited to have a TED talk in addition to enabling one to deliver his ideas best.



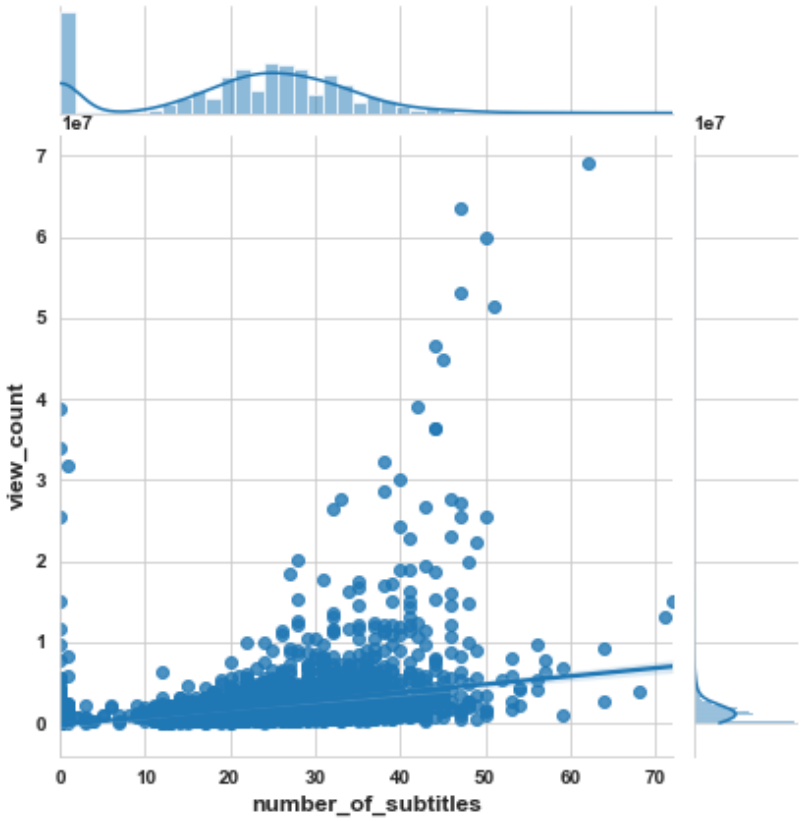
It was found that Writers are the most popular with more than 50 speakers identifying themselves as the aforementioned. In addition, Journalists and Entrepreneurs come in second and third place with very near results.



We can observe that writers tend to have the most views and have the greatest range. It is also interesting to see that most entrepreneurs do not give talks that gain a large number of views except for a few outliers. Also being an inventor seems not to guarantee a large number of views.

Number of languages a talk is available in

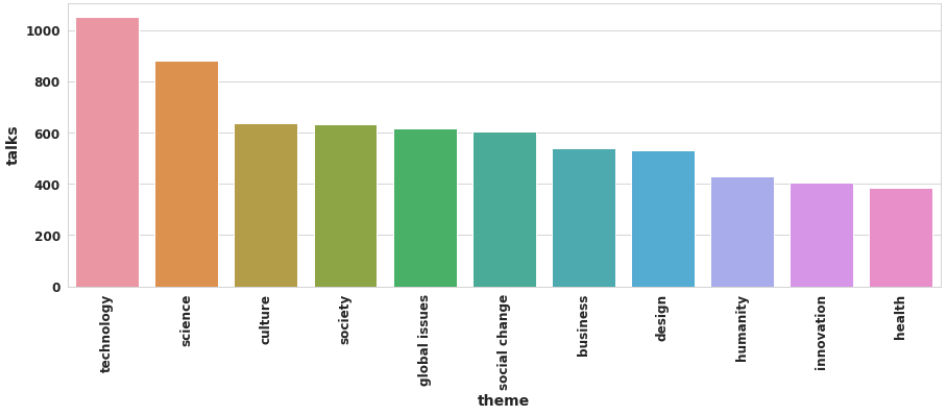
One remarkable aspect of TED Talks is the sheer number of languages in which it is accessible. We used to believe that the more languages a certain talk is available in, the more views it will has. After measuring the Pearson coefficient between these two features, This belief turned out to be true but the correlation is not that strong.



Which topic attracts most views?

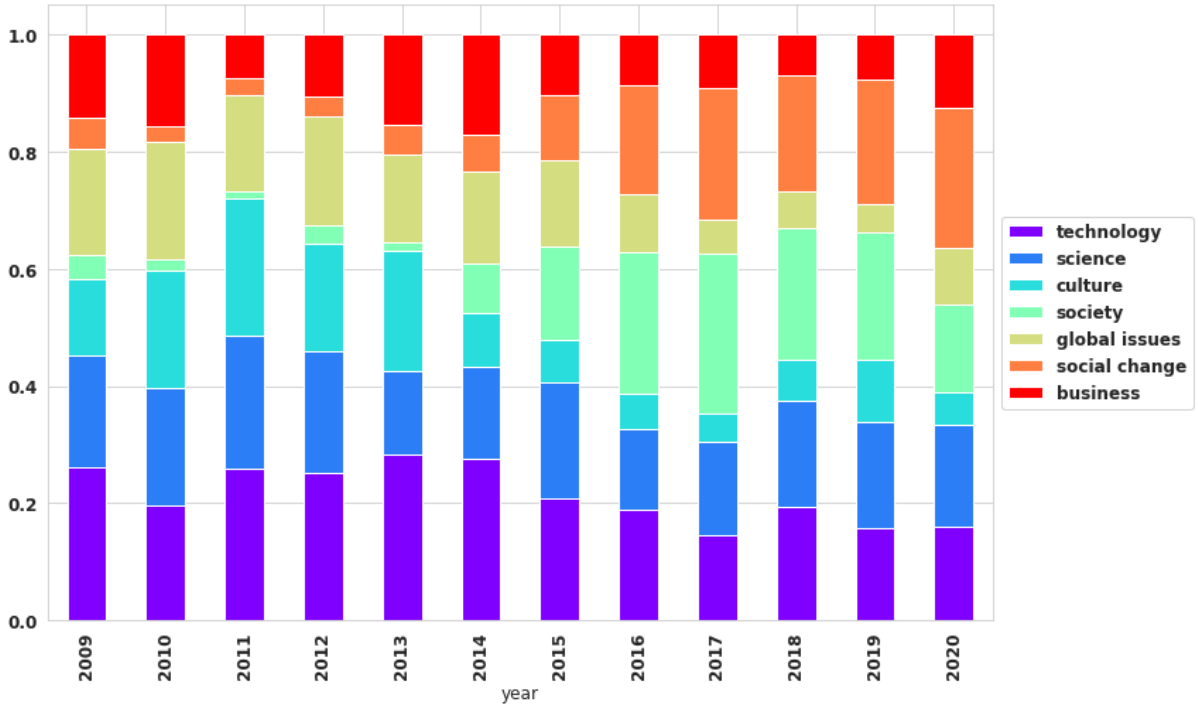
In this section, we will try to find out the most popular themes in the TED conferences. Although TED started out as a conference about technology, entertainment, and design, it has since diversified into virtually every field of study and walk of life. It will be interesting to see if this conference with Silicon Valley origins has a bias towards certain topics.

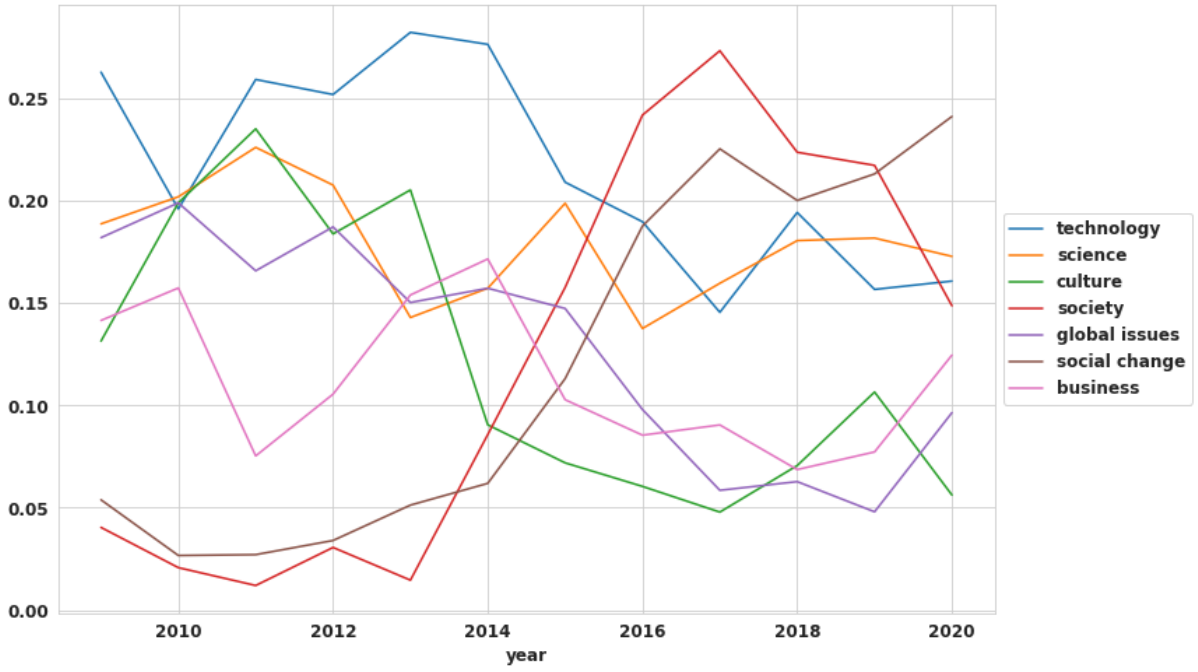
It was found that there are more than 455 themes in different TED talks and the following plot shows the most 11 themes that TED talks were about.



As may have been expected, Technology is the most popular topic for talks, followed by science and culture. By comparing these results with the results of 2017, it semes that the interest of viewers in global issues have decreased as it ranked third in 2017. We believe that this because of the uprising of conspiracy theories and people may have had enough of people telling them about the end of the world.

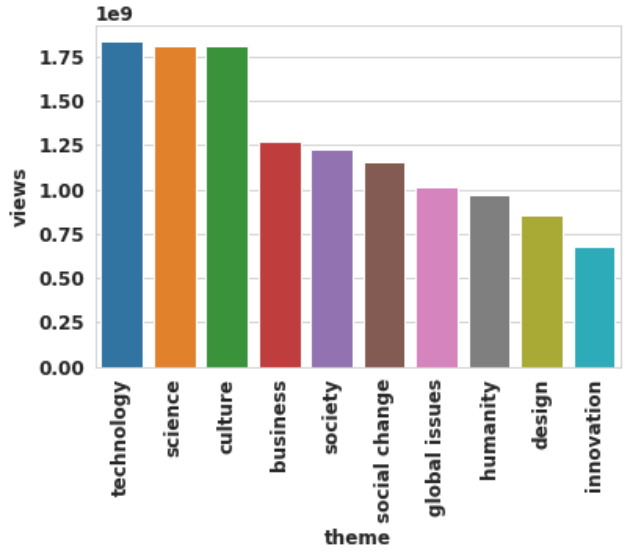
We currently seek to find how the number of talks about a certain topic has changed over the years as we believe that a lot has happened in the past years that has shifted the interest of the public.





It appears that the number of talks about science is most stable throughout the years. Talks about technology have been decreasing over the years but it appears to be re-gaining momentum. Finally, Society has spiked from 2013 till 2017 and then began to decline again

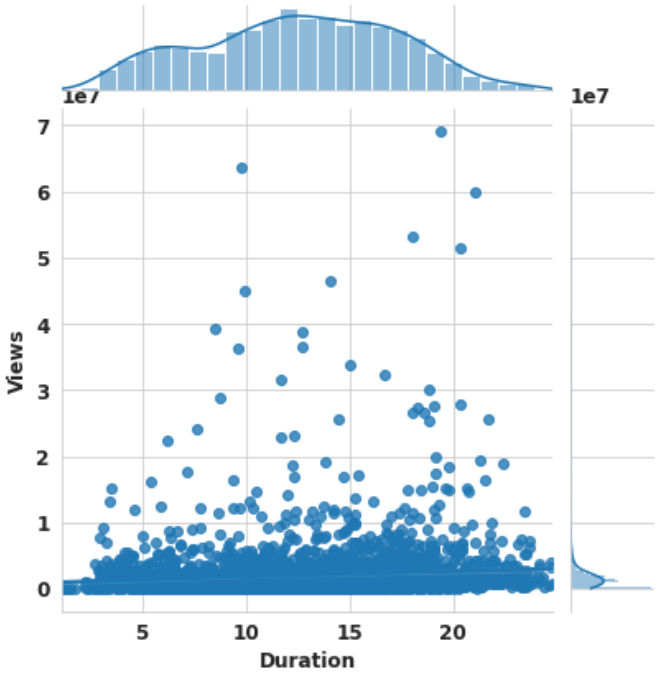
It should be noted that this analysis contradicts with the analysis made on the dataset collected till 2017. We believe that this is because many talks are uploaded later than their film date.



In conclusion, it is apparent that having a talk about technology, science and culture is the safest option to have your talk about.

Duration of the talk and number of views

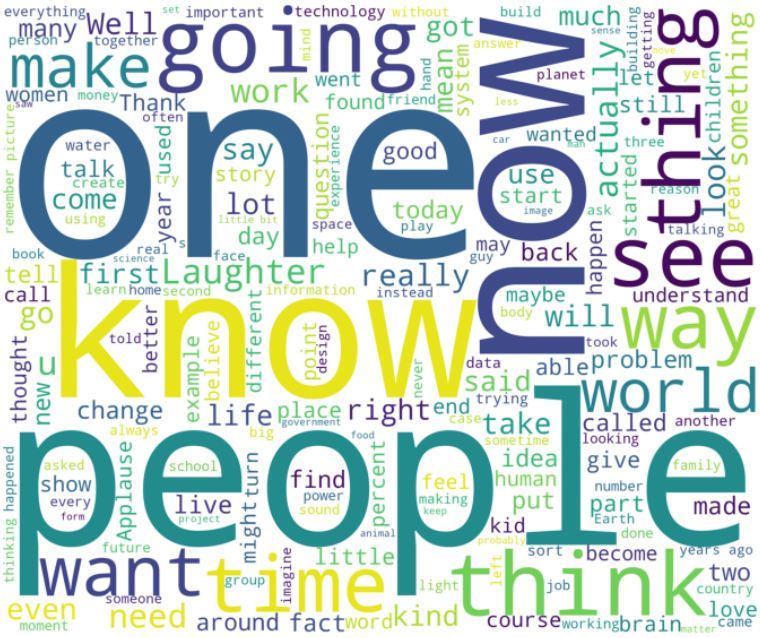
In this section we aim to find the relation, if exists, between the length of a talk and the number of views it has. It should be reminded that TED has a strict 18-minute rule, but the time of TED talks vary greatly.



It was found that the Pearson coefficient between the duration and the number of views is extremely low (less than 0.1) which proves that there is no tangible correlation between the length and the popularity of a TED Talk.

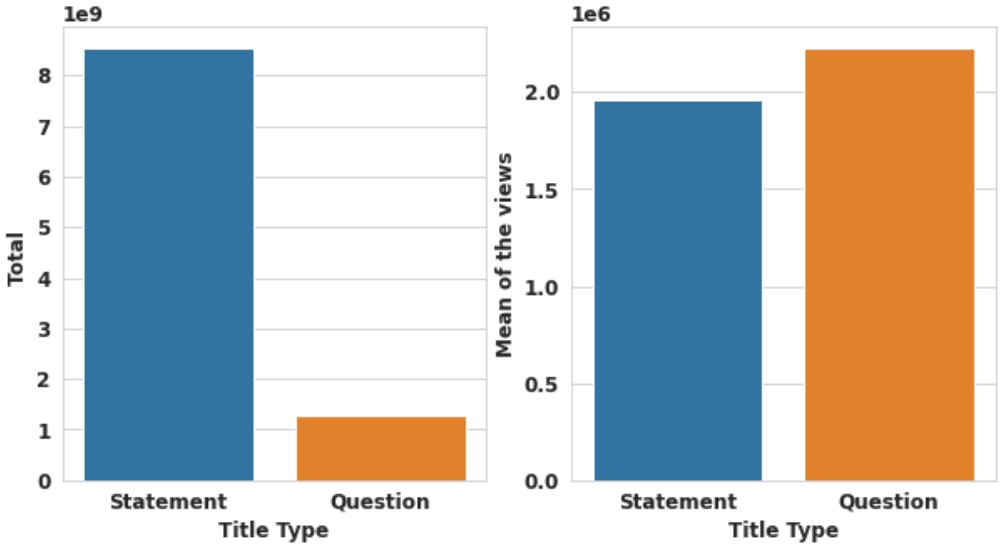
Words that appear most frequently

We decided to dive in even further and search if there are certain words that appear most frequently in all TED talks. We have constructed a word cloud from the available transcripts, which are 4359 transcripts. The words one, now and people are the most popular words across the corpus of all TED Transcripts which we think encapsulates the idea of TED well. TED speeches seem to place a lot of emphasis on knowledge, insight, the present and of course, the people.



The title: Make a Statement or Ask a Question

For our last analysis, we seek to know which form your title should take. More formally, should the title of your talk be a question or a statement?



We have found that Number of talks with Questions as its title is 577 while the number of talks with Statements as its title is 4351. This makes the fact, which is shown in the above graph, that the total views of videos that have their titles as statements is much more than that of videos that has extremely reasonable. However, the mean of the views of the videos with questions as a title is higher. This makes us recommend having the title of your talk is a question rather than being a statement.

Categorizing the talks according to its tags

Since each talk has its own tags, it will be viewed if people are interested in categories related to those tags. In fact, different words could have the same meaning; so considering the words that have the same meaning of the mentioned tags may help in categorization efficiently. We used "word-to-vector" method to obtain a vector of words that are equivalent in meaning for each tag. We used **KeyedVectors** from **gensim** library to extract those vectors from a collected data set from Google named "GoogleNews-vectors-negative300.bin.gz".

Now, for each tag we have a vector that contains the equivalent words. By using PCA, the parameters of that used to relate the vectors with each other is reduced to only 2 components. By K-Means algorithms, the vectors were clustered into 15 categories.

Giving each cluster a name was a challenge as in each run the results of the clustering differ. But in most cases, those are the categories that are notices in the words:

(Economics/Bussiness Category, Global issues Category, Exploration Category, Humanity/Progress Category, Music Category, Scientific Fields Category, Technology/Computer Category, Ecology Category, Epidemics Category, Philanthropy/Religion Category, Arts Category, Animals/Organisms Category, Social Category, Media/Entertainment Category)

Prediction

In this section, and after that thorough analysis of different features of the data, we now aim to design a model to predict the number of views a TED talk with certain features would have and perhaps, in the process, determine what the most important feature of them all is.

Selection of features

Although some features could have a large factor in the analysis like published time year, it is not reliable to consider them our prediction because we may not go to some year to present the talk. Therefore, the selection of features was done by selecting the following features:

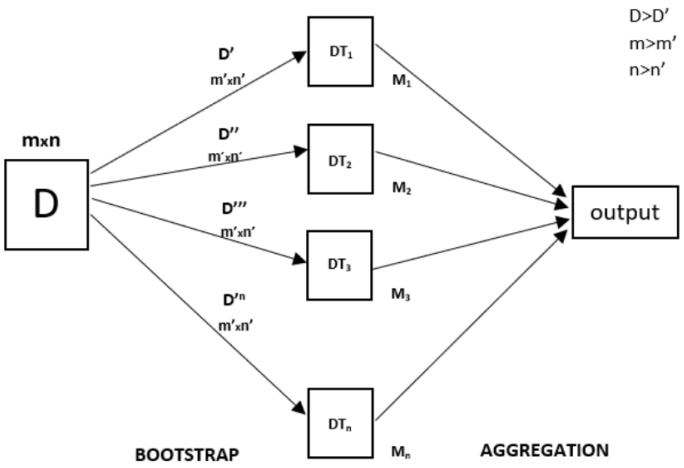
1. Duration
2. Type name
3. Number of related views
4. Published time day
5. Published time month
6. The Categories of the talk
7. Number of talk's categories

In prediction step, four regression methods were used to predict the number of views according to the prepared features. Let's have a quick overview in each in order to understand their results

Regression Techniques

Random Forest Regression:

The random forest regression is a classification method that is composed of parallel independent decision trees. This technique overcomes the drawback of the high variance of the decision trees, because it considers all the results produced from each decision trees. The following figure illustrates the division of the whole data into random rows (records) with some columns (features) to go through a decision tree; the result is the average of those decision trees' results.



Extra Trees Regression:

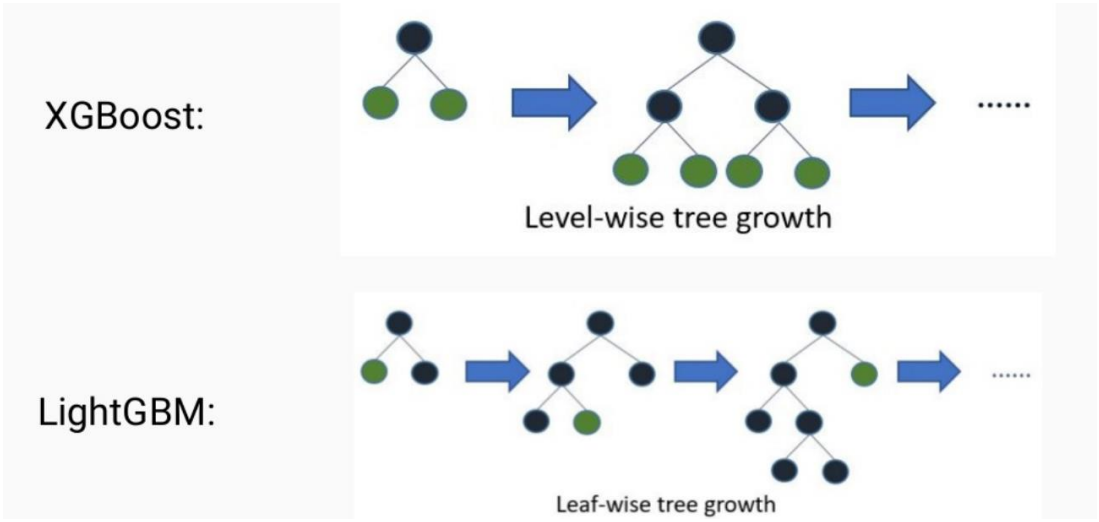
The random forest regression is a classification method that is similar to the random forest regression method in having parallel independent decision trees. They also have the same method to retrieve their result from the decision trees result by taking the average. The main difference is in the selection of cut points as the extra trees choose its cut points in the decision trees randomly, while the random forest has constant cut points.

XGBoost Regression:

XGBoost Regression is a gradient classification method that is composed of sequential dependent decision trees. Each decision considers the results of the previous one to enhance its model in regression. Although this technique could produce more accuracy in some cases, they don't process in parallel; therefore they had a large processing time.

LightGBM Regression:

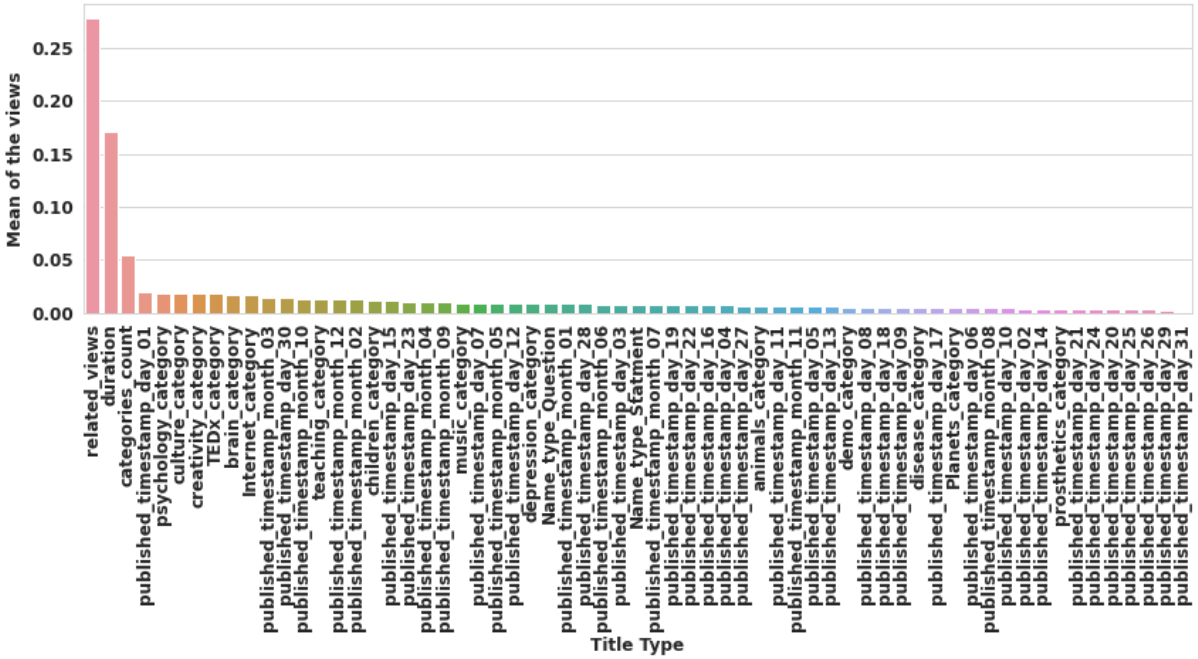
XGBoost Regression is also a gradient classification method that is composed of sequential dependent decision trees like the XGBoost regression. It works in reducing the time of processing by ignoring some leaf of the decision tree as showed in the following figure. In fact, this reduced the prediction accuracy as it missed some factors which may be important in prediction.



It becomes clear now that regression methods could be compared with two parameters which are Accuracy and Time. In order to compare the accuracy, we used MAE (Mean Absolute Error) estimator. Moreover, we calculated the time taken in each regression method

Regression Method	Training MAE	Test MAE	Total Fitting Time
Random Forest Regression	936782.56	1428783.78	204.3782018
Extra Trees Regression	649228.46	1413310.64	112.3990384
XGBoost Regression	957188.17	1704590.00	0.1934174
LightGBM Regression	1313007.53	1526545.78	0.0604686

As shown in the result table, the **Extra Trees Regression** was the most accurate regression method as it has the lowest MAE either on the training records or in the test records. On the other hand, the LightGBM Regression was the fastest regression method. Actually, reducing the n_estimators in the random forest or the Extra trees could reduce their fitting time, but unfortunately their accuracies will also decrease. The below figure shows the weight of each feature selected after applying regression. It appears that two factors that truly play a significant role is the number of views on related TED talks and the duration of the talk itself.



Conclusion

In conclusion, it was found that there are many factors that may have a direct effect on having the perfect TED talk. These are our concluding remarks:

1- Make your talk about topics such as technology and science.
2- Make sure to be prepared to give your talk on February as it is the month with most talks.
3- The more subtitle languages you can provide for you talk the better.
4- Make sure to not pass the 18-minute time limit of TED.
5- Make your talk title a question.
6- And most importantly, and this is what the different analysis techniques have shown, make sure that your talk is related to the most previously viewed talks on TED.

Finally, public speaking in general and acing a TED talk is difficult and challenging but it is a great quality to have. In this study we have tried to find other factors that need to be looked out for, after content of course, to give the perfect TED talk. We do not claim that this work is perfect nor complete, but we are truly proud with all the insights (whether they were predicted or not) that we have found.