TED Talk Data Analysis

Project By: Rohit Kumar (ZR66734), Arti Singh (LN22810)

Abstract

TED is a non-profit American media organization, spreading ideas, usually in the form of short and powerful talks (18-30 min video) covering areas from science to business to education to global issues. Most of these talks are available in more than 100 other languages. In this project, we have taken 3 different datasets for the available TED talks. These datasets include details about ~2500 TED talks, where one of the datasets provides the transcript of each TED Talk topic and second one provides the details like length of talks, views, release date, languages, rating, tags, views, comments etc. The third dataset provides speaker's details which includes speaker's career field, their background detail, etc. We analyzed the available datasets to get useful inferences to conclude how gender, duration, languages, views, comments, speaker's occupation, rating, topic of TED talk and the easiness of following the speech affects the popularity of the talk. We analysed the dataset features to derive meaningful relations between two features and used these derived features to calculate a popularity score. We used this pscore to find the popularity of TED talk. We considered a TED talk to be popular if the calculated popularity score, based on high correlation between the views and comments, is greater than 0.1 (range between 0.0017 – 0.8 value). Further, we did topic modelling on the Transcript dataset using NMF (Non-negative Matrix Factorization) around transcript data to cluster the TED talks and find a trend of how it covers the areas from science to business to global issues. We used these derived topics to draw various meaningful inference, as discussed in the below section, like category in which most of the TED talks falls into, more demanding and viewed topic categories, gender wise representation of the TED talk category etc. Finally, we built a predictive model which could predict the popularity of a TED talk given a talk's duration, transcript and speaker's name and occupation. Our model without much tuning of parameters performs quite well with an accuracy score of 95% and F1 score of 97% using Logistic Regression machine learning model.

Below are the datasets and implementation details in depth.

Dataset, Methodology and Approach

Dataset Details

- Include details about ~2500 TED talks.
- **ted_main.csv**: Contains data on actual TED Talk metadata and TED Talk speaker's information.
- **transcript.csv:** Contains transcript and URL detail of each TED talks.
- **speaker_raw.csv:** Contains speaker's information like which area they come from and their life detail.

Columns description for the ted_main.csv files i.e., the available features

- **comments**: comments made on the talk
- **description**: Talk description
- **duration**: Talk duration in seconds.

- **event**: Contains the TED talk id of the event where the event took place.
- **film_date**: It shows the date of recording of the talk and is present in the UNIX Timestamp
- **languages**: The number of languages in which the talk is available.
- **main_speaker**: Name of the speaker of the talk.
- **name**: name of the TED Talk (contains title and the speaker)
- **num_speaker**: The number of speakers giving the TED talk.
- **published_date**: This date is present in the UNIX timestamp in the original dataset and contains the date of publication of TED talk on TED website
- **ratings**: Various ratings given to the talk, present in the dictionary format in the talk (JSON Dictionary format)
- **related_talks**: list containing recommended talks to watch next.
- **speaker_occupation**: Speacker occupation
- **tags**: Tages associated with the talk.
- **title**: Ted talk title
- **url**: The URL of the talk.
- **views**: The number of view

Columns description for the transcript.csv files

- **transcript** Transcript of the talk
- **url**: The URL of the talk

Columns description for the speaker_raw.csv files

- **name** Name of the TED talk speaker
- **occupation** Occupation of the speaker
- **introduction** Brief introduction about speaker
- **profile** Speaker's profile

Goals and Possible questions that we answered from the available datasets (i.e. accomplished from the dataset)

Our main goal here in this project is find out the parameters, derive meaningful relationship and draw useful inferences about how they impact the popularity of a TED Talk and is it gender based. At the same time, our focus would be to answer the below questions from the dataset available:

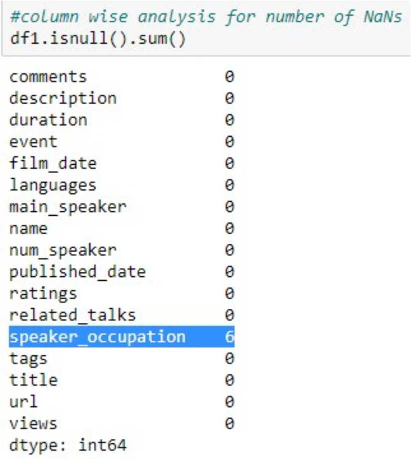
1. The area which has more TED talks being released.
2. TED talk which has more influence on the crowd taking the number of views into consideration.
3. Number of languages TED talk is released in, number of views, number of response received over them. Does Most Popular Ted Talks has been released in more languages?
4. What is making the TED talks popular, considering the duration on talk.
5. Occupation of the speaker and how it is related to TED talks.

6. Tried to identify the answers to some of the basic questions like ‘Is the TED talk on global issues more popular and speakers from which fields are more leaned towards it?’
7. Taking the gender of the speaker into account, to find out which gender has more popular and powerful speaker

Detailed Report on Implementation:

‘ted_main.csv’ Initial Pre-processing and Visualization:
(Jupyter Notebook: ‘TED_Main_Analysis.ipynb’)

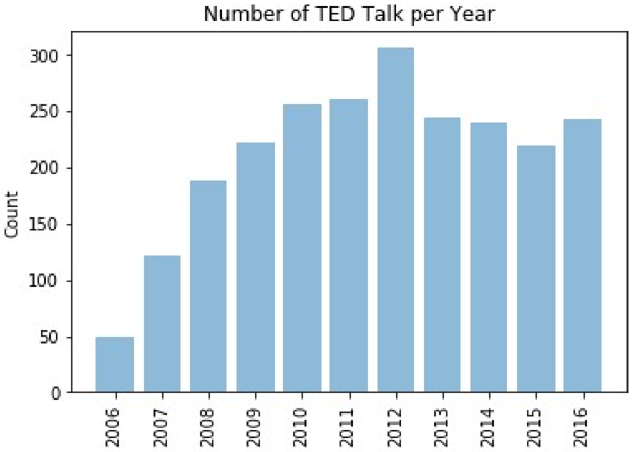
1. We checked for the presence of nan Values, Only 6 nan are present in speaker_occupation column, we dropped those rows
2. Date present in the dataset is in Unix format so we processed Unix Date and changed it MM-DD-YYYY format and created month and year columns for further analysis
3. We plotted 5 different histograms to view data count over five different feature columns like Number of TED Talk per Year, Number of views, Duration of TED Talk in seconds, Number of comments, Number of Language TED-Talks has been released



```
#column wise analysis for number of NaNs
df1.isnull().sum()
```

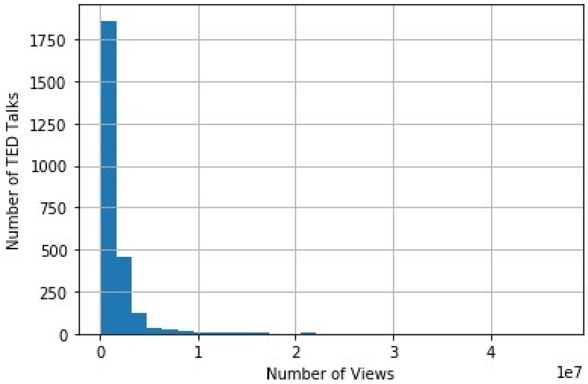
```
comments              0
description           0
duration              0
event                 0
film_date             0
languages             0
main_speaker          0
name                  0
num_speaker           0
published_date        0
ratings               0
related_talks         0
speaker_occupation    6
tags                  0
title                 0
url                   0
views                 0
dtype: int64
```

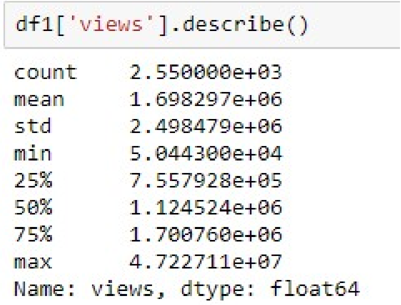
TED Talks per year



- We observe in above figure that number of Ted Talks keep on increasing from 2006 to 2012, showing its growing popularity during the given period and thereafter remains near about same for each following year.

Number of TED Talk Views

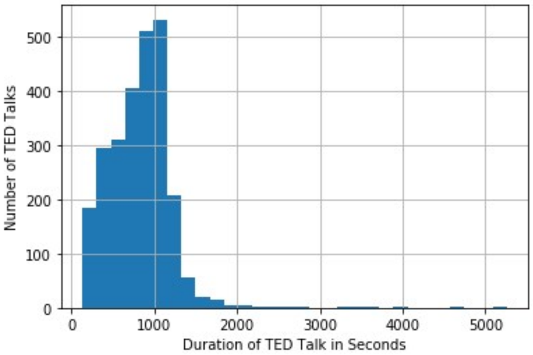


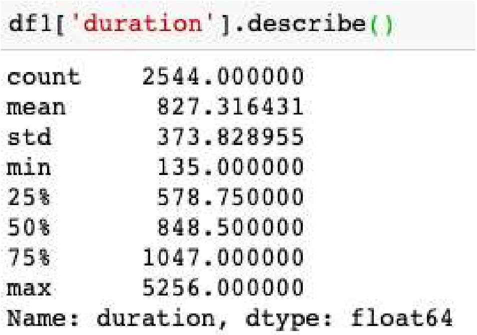


```
df1['views'].describe()

count    2.550000e+03
mean     1.698297e+06
std      2.498479e+06
min      5.044300e+04
25%      7.557928e+05
50%      1.124524e+06
75%      1.700760e+06
max      4.722711e+07
Name: views, dtype: float64
```

- Most of the TED Talks has around 1 million views with some TED Talks between 2 to 10 million views, few are between 10-20 million views range.
- We observe that the maximum number of views for a TED Talk out of all the observed TED Talk is 47.22 million and minimum being 50000 views.
- The above visualization suggests a very high level of popularity of TED Talks. We see the average number of TED Talks views is 1.6 million.

TED Talk Duration

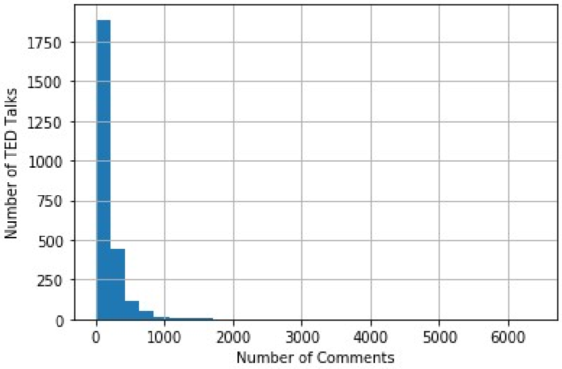


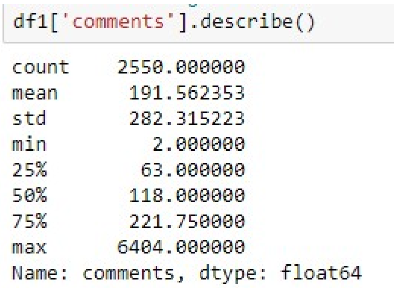


```
df1['duration'].describe()

count    2544.000000
mean      827.316431
std       373.828955
min       135.000000
25%       578.750000
50%       848.500000
75%      1047.000000
max      5256.000000
Name: duration, dtype: float64
```

- The graph looks close to normalized when plotted against the duration of each TED talks. The mean is cantered between 800-1000 seconds as we can see from the graph. However, on right-side the tail of graph is long with few TED Talks with longer durations with the TED talk with the highest duration being 5256 and min duration is 135 second.

Number of TED Talk Comments

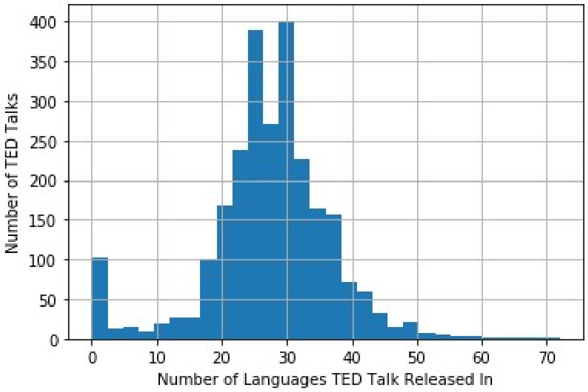


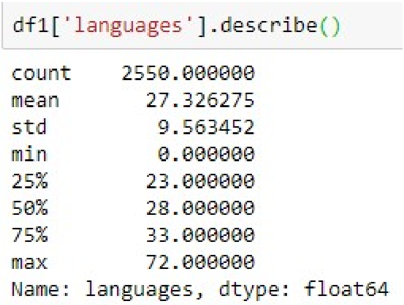


```
df1['comments'].describe()

count    2550.000000
mean      191.562353
std       282.315223
min         2.000000
25%        63.000000
50%       118.000000
75%       221.750000
max      6404.000000
Name: comments, dtype: float64
```

- The graph shows that most of the TED Talks has comments in the range of 0-300, with few TED Talks in range 500-2000.
- So, we find that the average number of comments is 191.56 with a high standard deviation of around 282, a value even higher than mean, thus suggesting it may be sensitive to outliers. The minimum number of comment on any Ted talk is 2 and the maximum is 6404, so there is huge variation between minimum and maximum comments made on Ted talk. Median of comments is 118.

TED Talk Translated in Number of Languages





```
df1['languages'].describe()

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
```

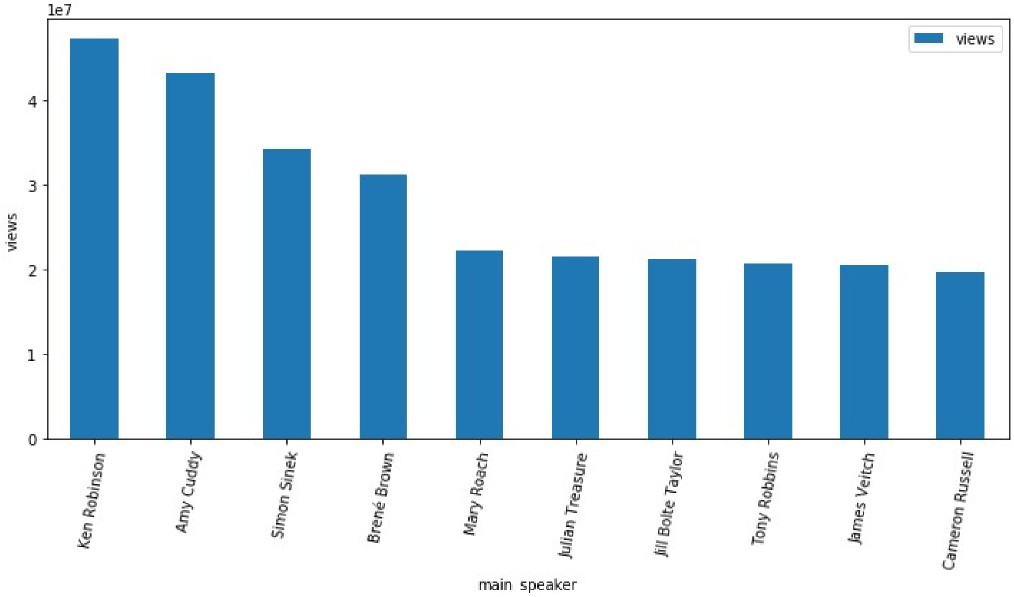
- The graph looks normalized with most of the TED Talks being released in 20-40 languages.
- The average number of languages is 27, in which most of the TED Talk is translated, with highest being in 72 languages and lowest with no translation.

Columns Co-relation Visualization

After exploring the datasets above for some of the columns which we think may be useful to draw the meaningful conclusion in later part of our analysis, we next tried to draw conclusion between some of the pairs of columns like Views-Speaker column, Languages-Speaker, Views-Comment, Views-Language, Views-Duration etc. Why we opted these pairs of column for our analysis?

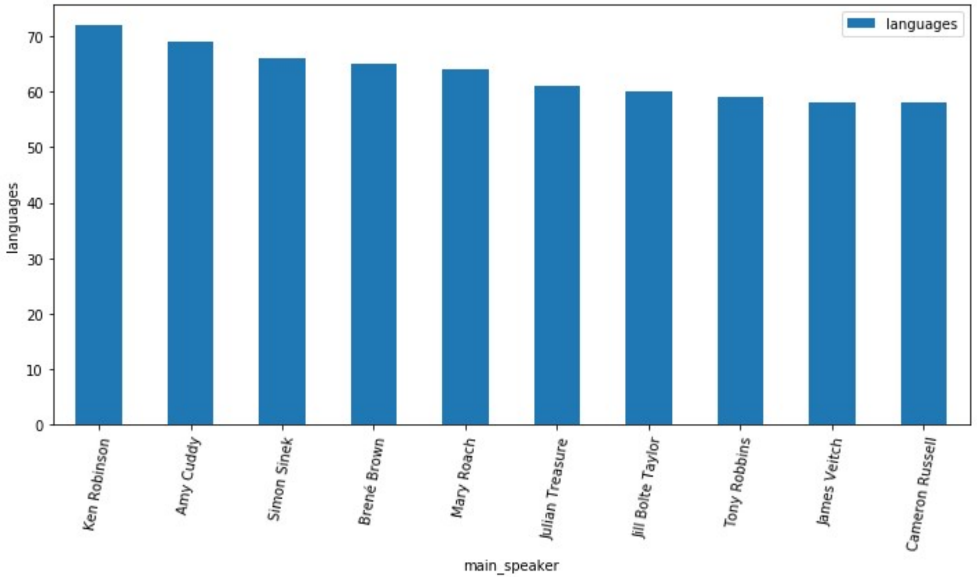
So, our goal here is to find out the parameters and draw meaningful inference using these parameters which is contributing to the popularity of TED talks. As, we have seen above using the histogram plot, the feature wise significance which is pointing to the reason for the popularity of a TED Talks. After some basic visualization of the columns, we further thought that might be TED Talks popularity is affected by person who is giving talks, number of language in which TED-Talk is released, number of comments which is made on the TED Talks released, Languages in which most number of people watch a particular Ted-talk or is the TED-Talk popular among the section of people speaking a particular language and finally we thought is there any significance of the duration of the TED-talks which is affecting the popularity so we tried to find out the relationship between these pairs of features and draw meaningful conclusion using the plot and statistical analysis as below.

1. Views-Speaker



- We plotted the top 10 Speakers which received the higher number of views for their TED Talks, and we observe that 2 Speakers ‘Ken Robinson’ and ‘Amy Cuddy’ crosses the 40M views and other in the top 10 has around 20-30M views having a huge gap among the top two so further thought why this difference is huge between the top 2 and rest so plotted the Language vs Speaker plot as below.

2. Languages-Speaker

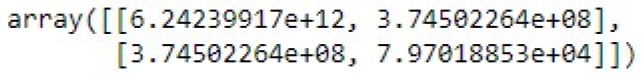


- This plot shows the speaker for which the highest number of translated version is release. And we found an interesting observation that ‘Ken Robinson’ and ‘Amy Cuddy’ who has the TED-talks with largest number of views (we can see from earlier plot) To have their talk translated in more number languages .

3. “Views-Comment” Relation Analysis

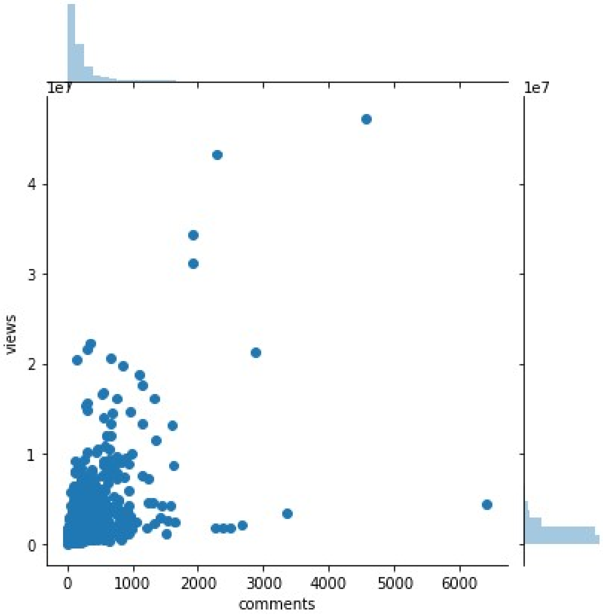
We plotted the number of views vs Number of comments made on that particular TED-Talk

- To see this relation, we calculated the covariance between the features using numpy cov and we got below result:



```
array([[6.24239917e+12, 3.74502264e+08],
       [3.74502264e+08, 7.97018853e+04]])
```

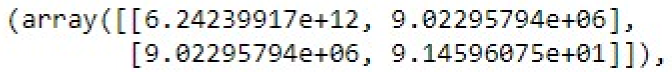
- We can see that the covariance between the two variables is positive, suggesting the variables change in the same direction as we expect. This can be further analyzed using Pearson correlation coefficient which is used to summarize the strength of the linear relationship between two data samples.
- Using pearsonr from ‘scipy.stat’ calculated the **Pearsons correlation** between views & comments and we got this values as **0.531.**
- So, we find that two variables are positively correlated with correlation value as 0.531. This suggests a high level of correlation, e.g. a value above 0.5, which further indicates a medium to strong correlation between the two quantities. Below plot shows how the two variables are correlated using sns joinplot.



4. **“Views-Language” Relation Analysis**

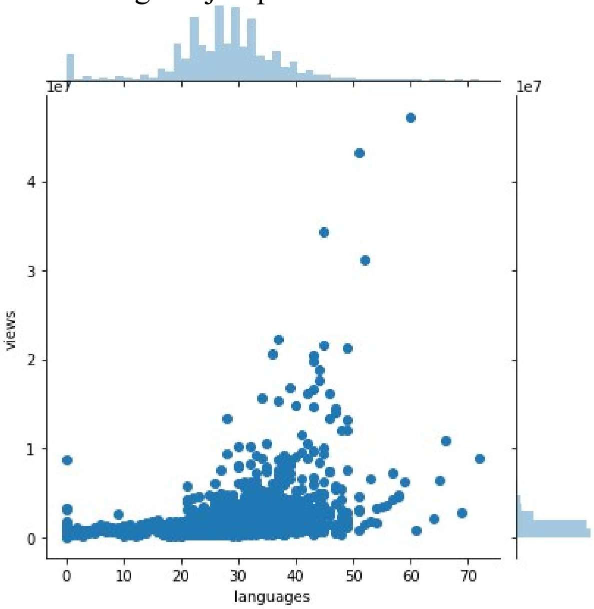
Plot between number of views vs number of language in which a given TED-talks is translated.

We tried to find out the covariance between the 2 features and got below result:



```
(array([[6.24239917e+12, 9.02295794e+06],
        [9.02295794e+06, 9.14596075e+01]]),
```

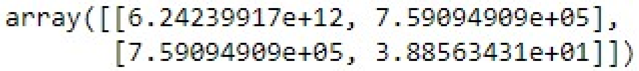
- Using pearsonr from ‘scipy.stat’ calculated the Pearsons correlation between views & language to summarize the strength of the linear relationship between two data samples
- **Pearsons correlation** between views & languages comes out to be **0.378**
- So, we find that two variables are less correlated with correlation value as 0.378. This suggests a low level of correlation between the two features. Below plot shows how the two variables are correlated using sns joinplot



We also observe from the above graph that as the number of language in which a TED-Talk is translated increases, number of views also increases but it is not the general case and this can also be concluded from the pearsons coefficient calculated.

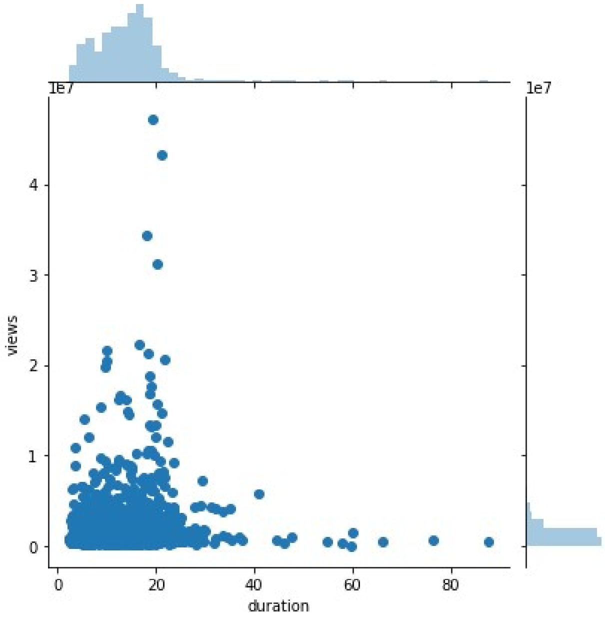
5. “Views-Duration” Relation Analysis

Plot between number of views vs length of duration(in seconds) for a given TED-talk.
We calculated the covariance between the 2 features and got below result:



```
array([[6.24239917e+12, 7.59094909e+05],
       [7.59094909e+05, 3.88563431e+01]])
```

- Using pearsons from ‘scipy.stat’ calculated the Pearsons correlation between views & duration to summarize the strength of the linear relationship between two data samples
- **Pearsons correlation** between views & duration: **0.049**
- So, we find that two variables very less correlated with 0.049 pearsons value. Below plot shows how the two variables are correlated using sns joinplot



Overall analysis for the above co-relation analysis and graphs plotted:

- ***So, we found that ‘view’ and ‘comment’ are strongly correlated and we decided to calculate the popularity score for each TED Talk using these features.***
- ***We calculated a popularity score(also called ‘pscore’) as percent of (number of comments/number of views)***

From observations, we concluded that what we thought was a right way to do because pair-wise feature analysis showed that among all the pairs views and comments are strongly corelated and can be used further to make useful inference as discussed above. Other pairs were also corelated but not as strong as this view-comment pair, so these were some few negative results not considered in further analysis.

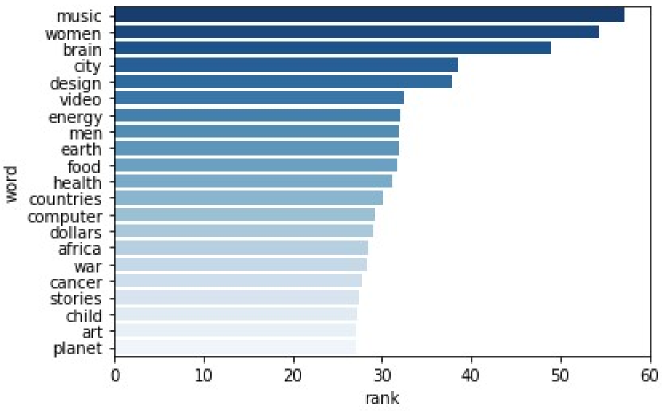
TRANSCRIPT DATA ANALYSIS

(Jupyter Notebook: ‘Source_Code_Analysis.ipynb’)

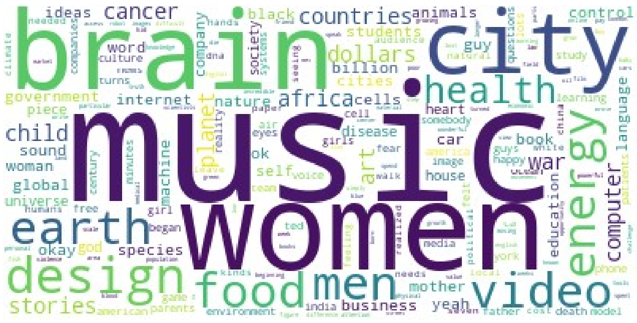
As discussed above Transcript data contains two things, Transcript and URL. So, we decided to perform Topic Modelling on this transcript data and derive some general and useful inference like general category in which most of the TED talks fall and are presented, more demanding and viewed categories of all the TED talks categories, gender wise representation of the TED talk category i.e., out of male and female who gave the most TED talk presentation in a particular category.

Topic Modelling on Transcript

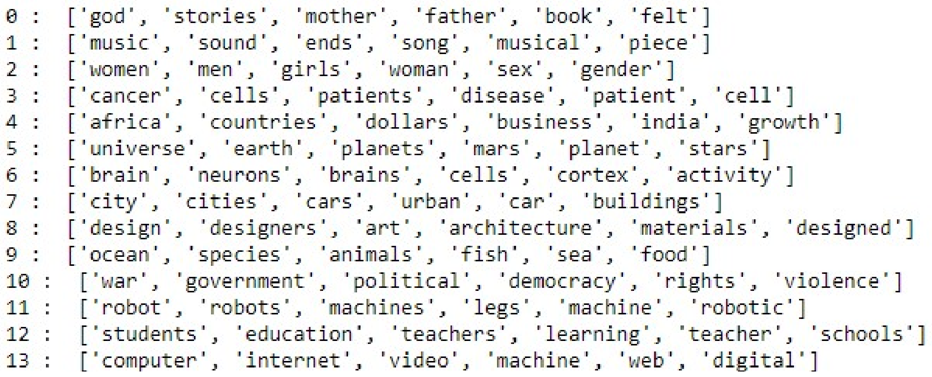
This data contained the ‘transcript’ of each TED Talk and its key as ‘url’ of the TED Talk. To get the vocabulary of the transcript, we used TF-IDF Vectorizer as we can see in the respective notebook and below plot shows the top 20 most occurring words in the vocabulary based on the word importance, by calculating the rank from the TF-IDF vector matrix of word-transcript.



Similar result was observed using the WordCloud over the transcript data.



Further, we did topic modeling on the transcript data using NMF (Non-negative Matrix Factorization). Here we tried and experimented by varying the number of categories and number of similar words to be considered to get different categories, and we found a better convincing result with 14 categories built on 6 similar word group clustered together, as shown below.



```
0 :  ['god', 'stories', 'mother', 'father', 'book', 'felt']
1 :  ['music', 'sound', 'ends', 'song', 'musical', 'piece']
2 :  ['women', 'men', 'girls', 'woman', 'sex', 'gender']
3 :  ['cancer', 'cells', 'patients', 'disease', 'patient', 'cell']
4 :  ['africa', 'countries', 'dollars', 'business', 'india', 'growth']
5 :  ['universe', 'earth', 'planets', 'mars', 'planet', 'stars']
6 :  ['brain', 'neurons', 'brains', 'cells', 'cortex', 'activity']
7 :  ['city', 'cities', 'cars', 'urban', 'car', 'buildings']
8 :  ['design', 'designers', 'art', 'architecture', 'materials', 'designed']
9 :  ['ocean', 'species', 'animals', 'fish', 'sea', 'food']
10 :  ['war', 'government', 'political', 'democracy', 'rights', 'violence']
11 :  ['robot', 'robots', 'machines', 'legs', 'machine', 'robotic']
12 :  ['students', 'education', 'teachers', 'learning', 'teacher', 'schools']
13 :  ['computer', 'internet', 'video', 'machine', 'web', 'digital']
```

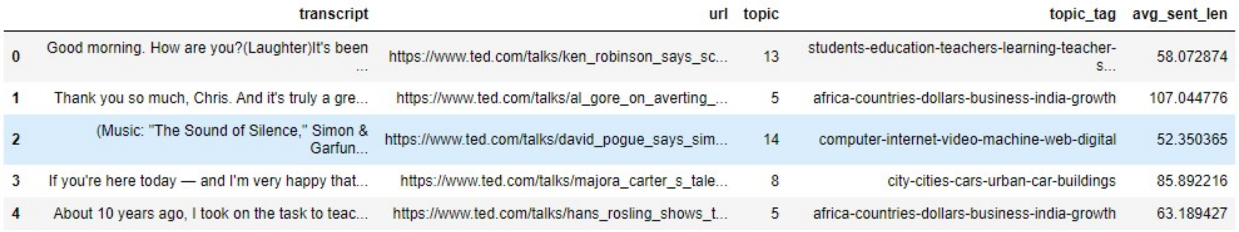
Extracted Topic Tag from Topic Modelling

We have applied NMF model on each TED Talk transcript to get these respective topic tags for above categories. We used Pipeline with NMF model and TFIDF vectorizer built on the data.

1. Average Sentence Length

We calculated the average sentence length for each TED Talk transcript, to get a differentiation over lengthy and small talks, which we planned to use later with main data.

In the end, we have the below dataset (representation) as final dataset of transcript:



	transcript	url	topic	topic_tag	avg_sent_len
0	Good morning. How are you?(Laughter)It's been ...	https://www.ted.com/talks/ken_robinson_says_sc...	13	students-education-teachers-learning-teacher-s...	58.072874
1	Thank you so much, Chris. And it's truly a gre...	https://www.ted.com/talks/al_gore_on_averting_...	5	africa-countries-dollars-business-india-growth	107.044776
2	(Music: "The Sound of Silence," Simon & Garfun...	https://www.ted.com/talks/david_pogue_says_sim...	14	computer-internet-video-machine-web-digital	52.350365
3	If you're here today — and I'm very happy that...	https://www.ted.com/talks/majora_carter_s_tale...	8	city-cities-cars-urban-car-buildings	85.892216
4	About 10 years ago, I took on the task to teac...	https://www.ted.com/talks/hans_rosling_shows_t...	5	africa-countries-dollars-business-india-growth	63.189427

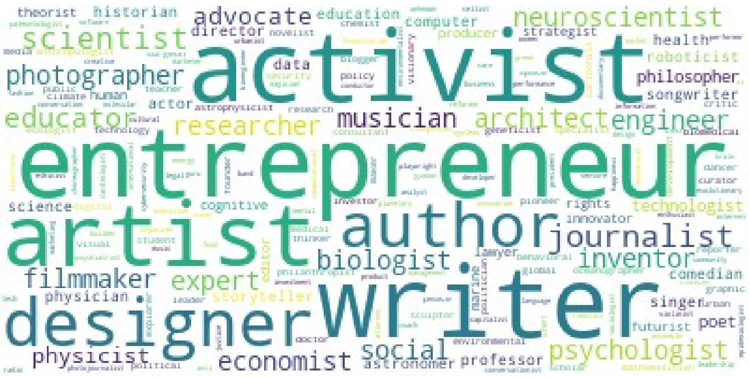
SPEAKER DATA ANALYSIS

(Jupyter Notebook: 'Source_Code_Speaker.ipynb')

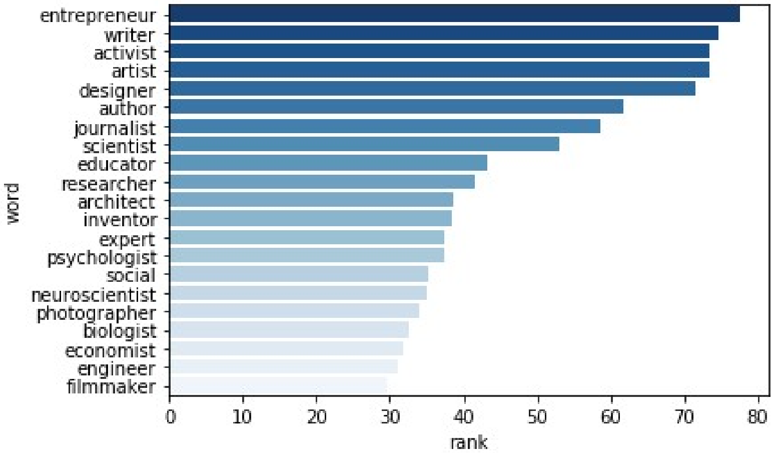
Similar to Transcript dataset, we also applied topic modelling on Speaker dataset on occupation column.

1. Topic Modeling on Occupation

Since the occupation field contained varying value, we applied the similar approach as earlier of Topic modeling on the 'occupation' column, to get a list of 30 speaker occupation category as seen in below WordCloud.

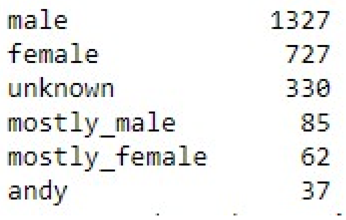


We can clearly observe that most of the TED Talk speakers are writer, entrepreneur, activist, artist, author with moderate ones from journalist, economist, biologist, and musician, social. Similar result is found when plotted over the frequencies of the word as below on the occupation column from Speaker dataset:



2. Finding Gender of the Speaker

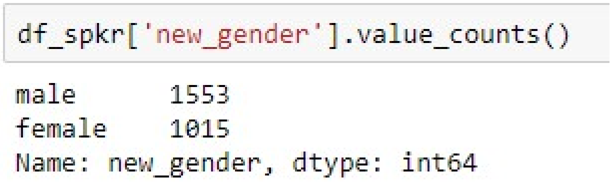
- Initially we used gender_detector to find the gender of speakers from their name, however, we didn't receive good result, as the category varied equally among 'male', 'female', 'mostly_male', 'mostly_female', and 'unknown' as shown below:



```
male              1327
female             727
unknown            330
mostly_male         85
mostly_female       62
andy                37
```

- So, we took the advantage of 'profile' column present in the data, which provides a brief introduction about the speaker.

- We then made use of pronoun to get the gender of the Speaker. So, if the profile have the words like ‘he/his’, are categorized as ‘Male’ and ‘female’ otherwise.



```
df_spkr['new_gender'].value_counts()
```

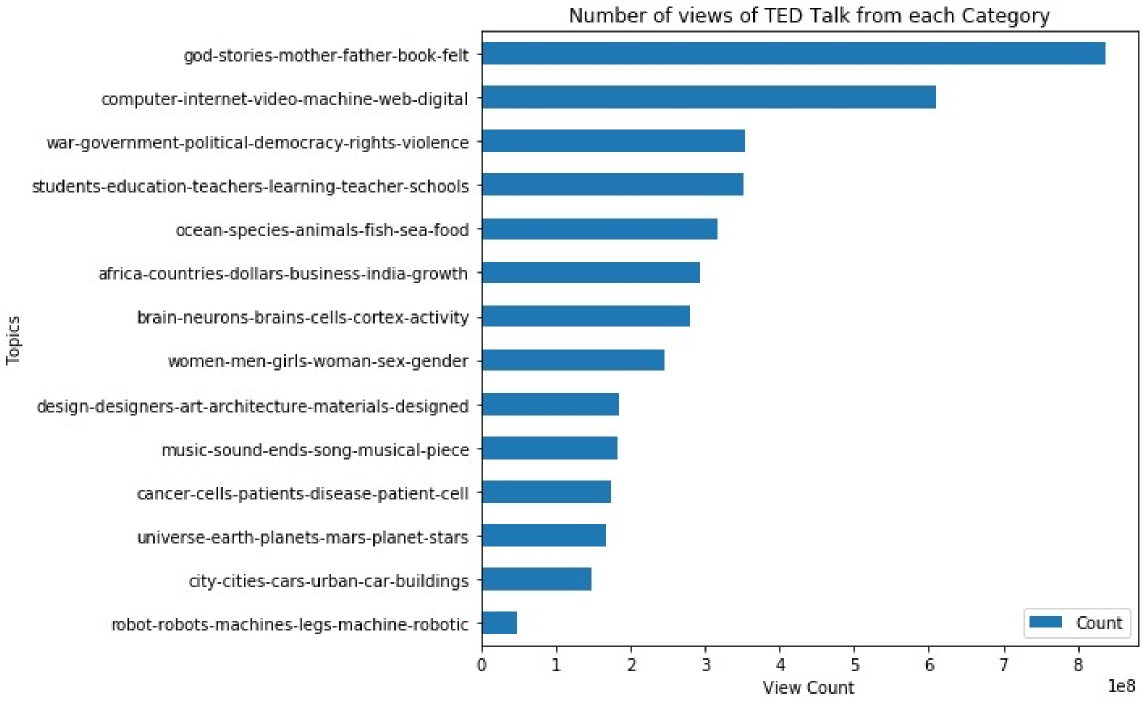
```
male      1553
female    1015
Name: new_gender, dtype: int64
```

Final Analysis by Merging the Dataset from 3 Output Datasets
(Jupyter Notebook: ‘TED_Main_Analysis.ipynb’ After Cell 38)

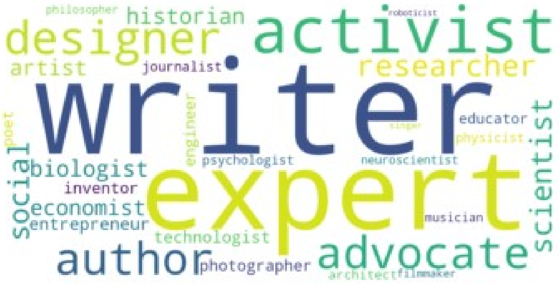
After merging the dataset, we found that we have 508 TED talks for which respective details are not present from the Transcript and Speaker Dataset. So, we decided to drop those rows for now. We are therefor left with 2399 instances of merged data.

Few more plots and visualisation that we tried now based on Topic Models generated and their categories are:

1. Top 20 Topic based on Number of View



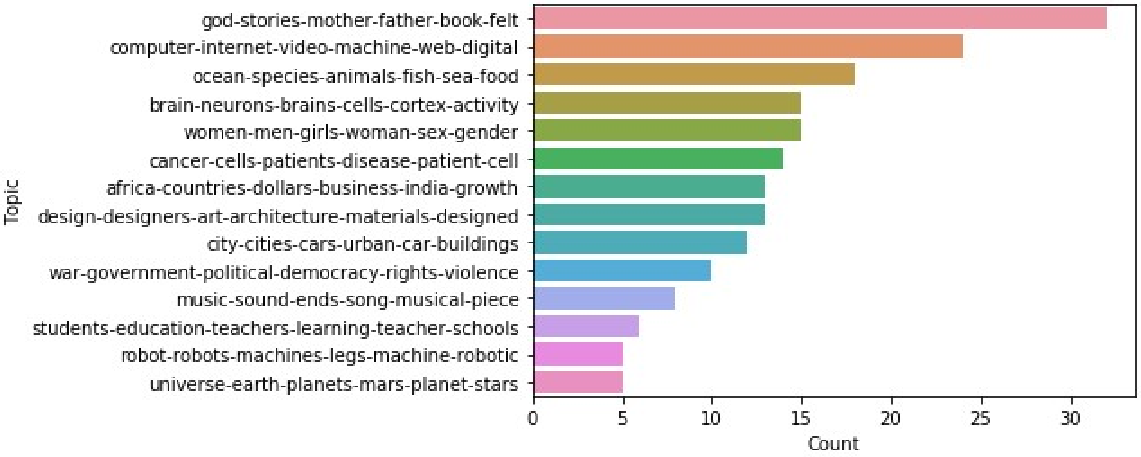
2. **Speaker Occupation based on Number of TED Talk**



We can observe that when combined with number of TED talks per speaker, we find speaker who are writer, expert, author, researcher, designer or an activist more actively involved in TED talks, than other speakers from other field like historian, social, musician.

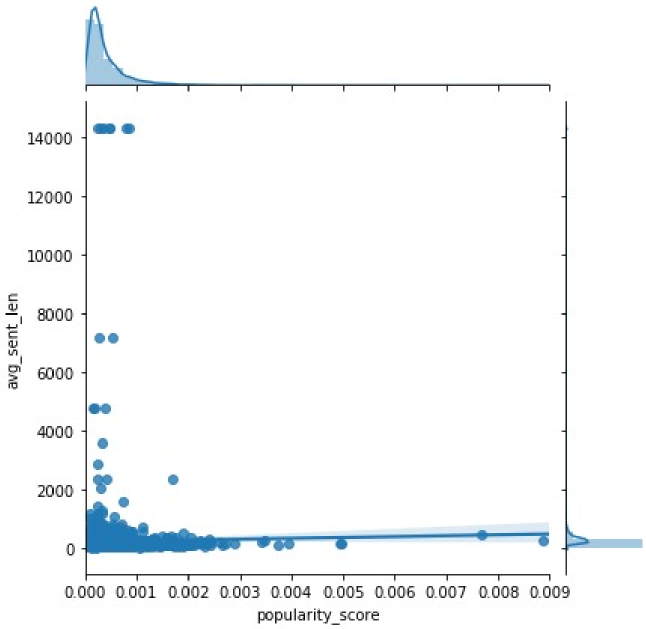
3. **Most Frequently talked about Topic based on Popularity Score >= 0.1 (the Score only range in between 0.0017 – 0.8)**

Priority score we calculated as discussed above based on the strong correlation between Number of Views and Number of Comments and is calculated as percent of number of comments/number of views. We took 0.1 as threashold values for pscore and a TED talk having a score above this value is considered as popular talks.



4. **Does Length of Sentence Impact the Popularity**

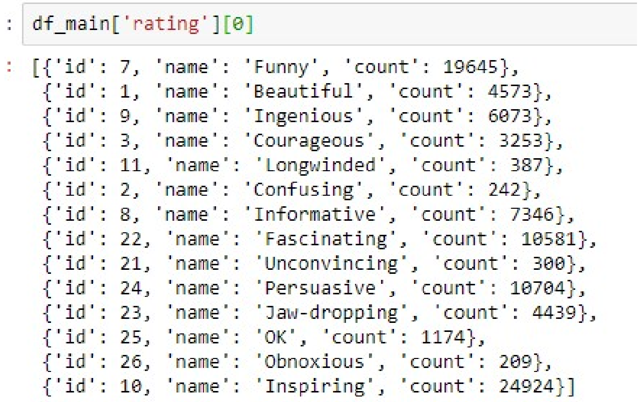
We observed from the below graph that TED talk with lengthy sentences has low popularity score than the TED talks with smaller sentence length. This shows that people may get bored with lengthy talks and prefers to watch lighter and smaller talks that are easy to follow.



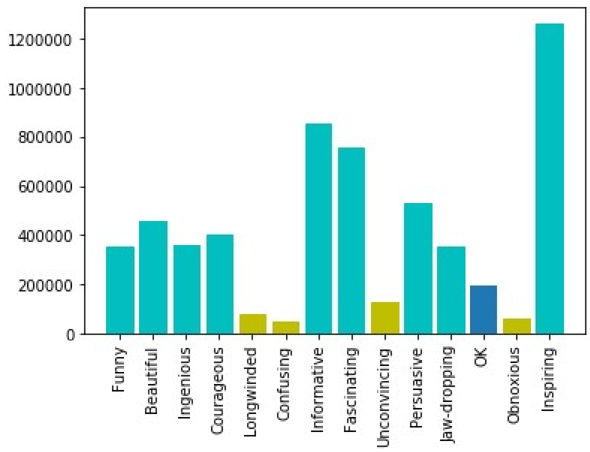
5. **At the end, we tried to visualize and draw inference from the Rating Column**

Rating column includes keyword and its associated hit indicating the emotion of the viewer. For example, the first TED talk has count of 19645 for ‘Funny’ keyword, which may mean that the viewer found this talk a lot funny than ‘Obnoxious’ which has count of 209. This looked like something similar to Facebook’s ‘Like’, ‘Love’, ‘Sad’, ‘Laugh’, ‘Angry’ emoticons.
So, considering these keywords to provides us some useful information about viewer’s sentiment about the TED talk, we first took out all the possible keyword we have in the current TED talk data. We found 14 such keywords, same across the entire TED talks ‘rating’ column, as in the figure below.

The plot on the right shows total numbers of hit received for the particular keyword over the entire TED talks in the dataset where we can clearly observe that the most of the TED talks seemed to be inspiring, informative and fascinating which looked more reasonable and genuine.



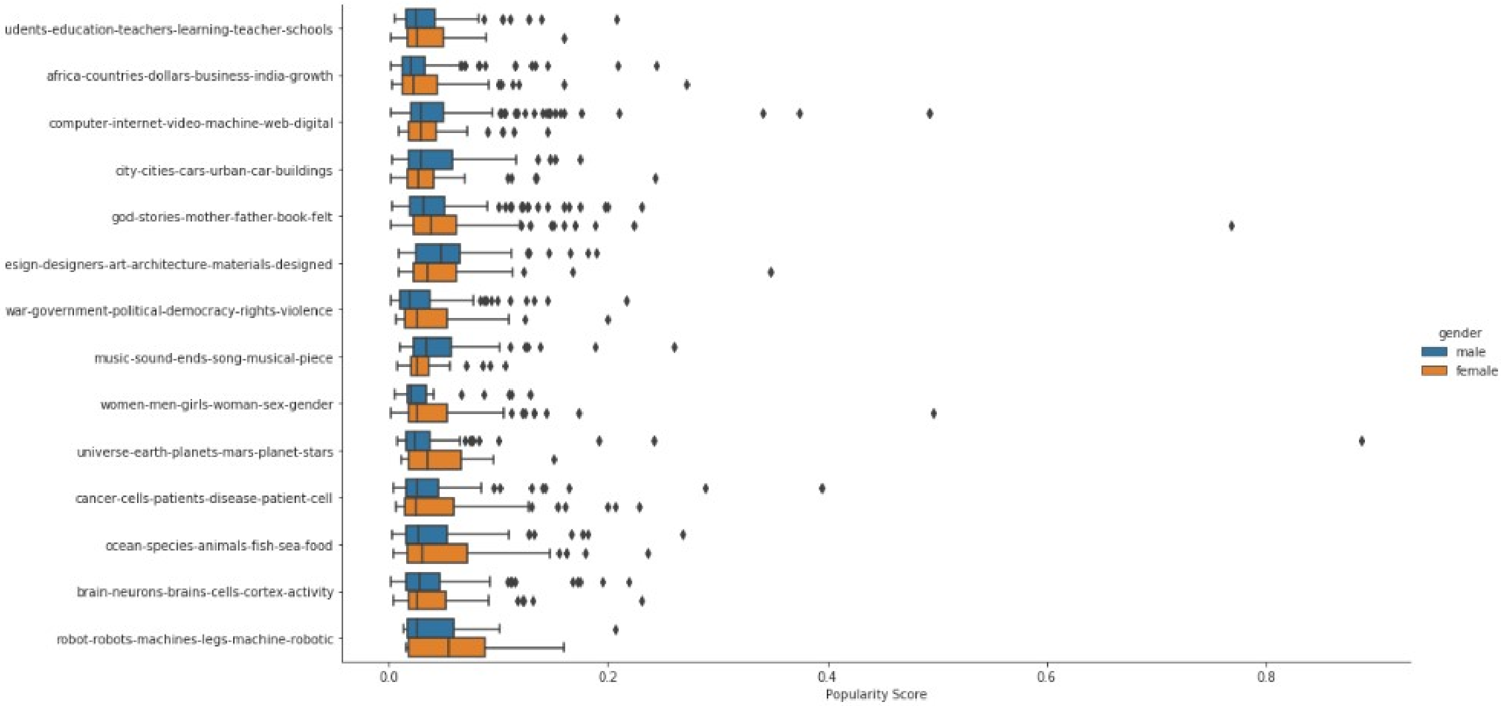
```
: df_main['rating'][0]

: [{'id': 7, 'name': 'Funny', 'count': 19645},
   {'id': 1, 'name': 'Beautiful', 'count': 4573},
   {'id': 9, 'name': 'Ingenious', 'count': 6073},
   {'id': 3, 'name': 'Courageous', 'count': 3253},
   {'id': 11, 'name': 'Longwinded', 'count': 387},
   {'id': 2, 'name': 'Confusing', 'count': 242},
   {'id': 8, 'name': 'Informative', 'count': 7346},
   {'id': 22, 'name': 'Fascinating', 'count': 10581},
   {'id': 21, 'name': 'Unconvincing', 'count': 300},
   {'id': 24, 'name': 'Persuasive', 'count': 10704},
   {'id': 23, 'name': 'Jaw-dropping', 'count': 4439},
   {'id': 25, 'name': 'OK', 'count': 1174},
   {'id': 26, 'name': 'Obnoxious', 'count': 209},
   {'id': 10, 'name': 'Inspiring', 'count': 24924}]
```



We also categorized the TED Talks as having received Positive, Negative and Neutral comments. We created another feature named ‘sentiment’ which indicates whether the TED talk received a positive and negative response from the viewers. These categories are as follows based on the keyword from comments.

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

6. Final Visualization we did on Gender with Topic based on Popularity Score



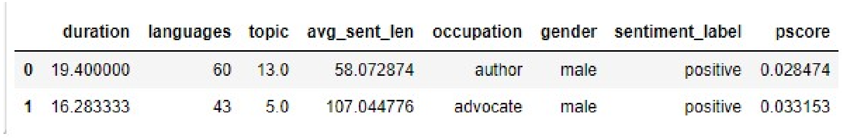
Observation

1. We observe many outliers, on which we can't generalize a particular topic to be popular only if it has a single TED Talk with high popularity.
2. Secondly, one can observe that in 90% of the TED Talk categories, female out-performs male in sense of popularity score. So, we can say that TED Talks given by female are more popular when compared to male.

Predictive Modeling

(Jupyter Notebook: ‘model_code.ipynb’)

Finally, from all the above analysis and inferences, we concluded that the below features are making impact on deciding the popularity of TED talks, and so we opted these features for our predictive modelling tasks:



	duration	languages	topic	avg_sent_len	occupation	gender	sentiment_label	pscore
0	19.400000	60	13.0	58.072874	author	male	positive	0.028474
1	16.283333	43	5.0	107.044776	advocate	male	positive	0.033153

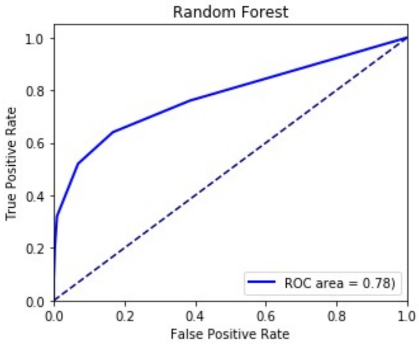
We took 70:15:15 split for train, dev, test of the dataset respectively for training and testing of our model. Below is the training and dev data test result using cross-validation for the respective trained classifiers showing the evaluation along with the performance information.

Random Forest
Cross Validation Scores:

```
    test_accuracy  test_precision  test_recall  test_roc_auc   test_f1
 0       0.919643        0.571429     0.142857      0.713706  0.228571
 1       0.910714        0.250000     0.035714      0.638741  0.062500
 2       0.916667        0.500000     0.071429      0.732549  0.125000
 3       0.931548        1.000000     0.178571      0.731563  0.303030
 4       0.913433        0.000000     0.000000      0.639658  0.000000
```

Dev Test Results:

```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       335
           1       0.83      0.20      0.32        25

    accuracy                           0.94       360
   macro avg       0.89      0.60      0.65       360
weighted avg       0.94      0.94      0.92       360
```

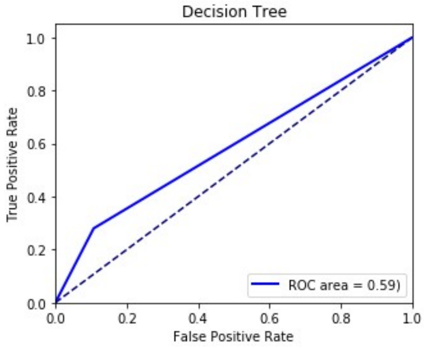


Decision Tree
Cross Validation Scores:

```
    test_accuracy  test_precision  test_recall  test_roc_auc   test_f1
0       0.857143        0.236842     0.321429      0.613636  0.272727
1       0.883929        0.296296     0.285714      0.612013  0.290909
2       0.892857        0.366667     0.392857      0.665584  0.379310
3       0.889881        0.320000     0.285714      0.615260  0.301887
4       0.874627        0.250000     0.250000      0.590798  0.250000
```

Dev Test Results:

```
              precision    recall  f1-score   support

           0       0.94      0.89      0.92       335
           1       0.16      0.28      0.21        25

    accuracy                           0.85       360
```



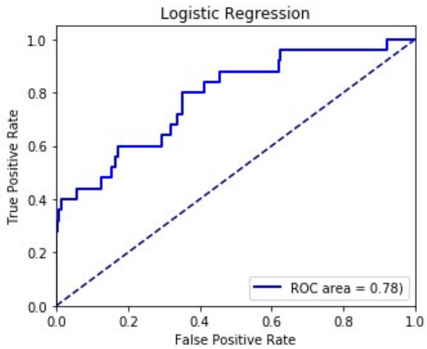
```
      macro avg       0.55      0.59      0.56       360
   weighted avg       0.89      0.85      0.87       360
```

Logistic Regression
Scores:

```
    test_accuracy  test_precision  test_recall  test_roc_auc   test_f1
0        0.928571        0.750000     0.214286      0.786526  0.333333
1        0.919643        0.666667     0.071429      0.742811  0.129032
2        0.916667        0.500000     0.071429      0.663381  0.125000
3        0.928571        0.833333     0.178571      0.744086  0.294118
4        0.913433        0.333333     0.035714      0.743602  0.064516
```

Dev Test Results:

```
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       335
           1       1.00      0.24      0.39        25

    accuracy                           0.95       360
   macro avg       0.97      0.62      0.68       360
weighted avg       0.95      0.95      0.93       360
```

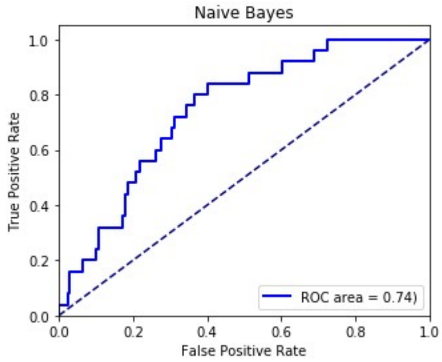


Naive Bayes
Scores:

```
    test_accuracy  test_precision  test_recall  test_roc_auc   test_f1
0        0.342262        0.078603     0.642857      0.528003  0.140078
1        0.377976        0.079070     0.607143      0.476925  0.139918
2        0.422619        0.089109     0.642857      0.518321  0.156522
3        0.407738        0.105991     0.821429      0.644944  0.187755
4        0.337313        0.074561     0.607143      0.514076  0.132812
```

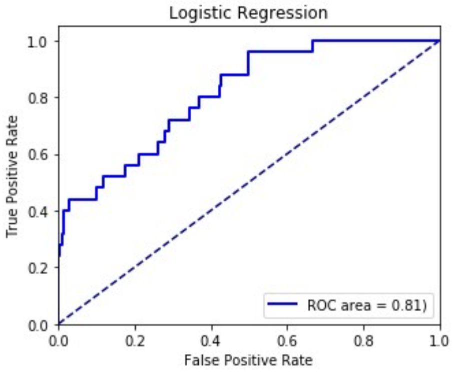
Dev Test Results:

```
              precision    recall  f1-score   support

           0       0.98      0.38      0.55       335
           1       0.10      0.92      0.18        25

    accuracy                           0.42       360
   macro avg       0.54      0.65      0.36       360
weighted avg       0.92      0.42      0.52       360
```



Final Result using Best Model on Test Data:

```
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       335
           1       0.88      0.28      0.42        25

    accuracy                           0.95       360
   macro avg       0.91      0.64      0.70       360
weighted avg       0.94      0.95      0.93       360
```



We performed predictive modelling on the derived important features using several classifiers to determine the accuracy of the model created.
We used classifiers like Random Forest, Decision Tree, Logistic Regression and Naive Bayes. We observe from the accuracy result and F1 score of all of the above trained classifier that Logistic Regression gave the best results with an accuracy score of 95% and F1 score of 97%

Conclusion

On analysing the TED talks dataset to explore the reason for a particular TED talk becoming popular using various parameters like gender, duration, languages, views, comments, topic of TED talk, rating, speaker's occupation etc, we finds that TED talk popularity is highly dependent on these factors. It is observed that the short duration TED talks are more famous and popular, it is more likely that people are interested in shorter duration talks because they are able to grasp the talk's content easily or they don't have time to watch longer duration talks. It is also observed that although language plays a significant factor in determining the popularity of a TED talks but it is not always the case. We observed that Views and comments for a particular TED talk is highly corelated and used this correlation factor to calculate the pscore which is further used to determine the popularity of a TED talk. We also analysed several other pairs for a meaningful correlation but they do not seemed to be strongly corelated to give a useful information so ignored other pairs and used the views- comments pair for our further analysis.

We applied Topic Modelling on the transcript data to derive the topics around which most of the famous TED talks are presented and to determine the gender based TED Talks relation using these generated topics. We calculated the average sentence length from the transcript and it showed that lengthy sentences has low popularity score than the TED talks with smaller sentence length. We identified the occupation of the speaker and found that speakers who are more actively involved in TED talks are writer, expert, author, researcher, designer or an activist etc. We applied topic modelling on the occupation field to derive this result. On analysing the ratings available for various TED talks we observe that the most of the TED talks seemed to be inspiring, informative and fascinating which looked more reasonable and genuine. On doing visualisation on gender with topic based on popularity score, we observe that 90% of the TED talk categories, female out-performs male in sense of popularity score. So, we can say that TED talks given by female are more popular when compared to male. Finally, on applying predictive modelling on the selected features derived from our analysis of datasets, we find that of all classifiers used to train the model, Logistic Regression gives the best results with an accuracy score of 95% and F1 score of 97%. We used various classifiers like Random Forest, Decision Tree, Logistic Regression and Naive Bayes on 70:15: 15 split to test and train the model and Logistic Regression out performed among all.

Our derived results evaluated the TED talks presentation on lots of parameter to explore the reason for the popularity of a particular TED talk. These meaningful derivation, inference and results can be used by several kinds of people with different aims in mind. A TED talk presenter can use it to know the common topic in which people are more interested to listen and thus he can plan the topic of his talk accordingly. Also, from the patterns from rating data for a TED talk can help speakers to know whether their talk has been informative or confusing or ok and improve accordingly. Another audience of our result can be a person looking to gain information from any TED talks from specific area. Our result would help such people to know which talk and speaker has the most popularity in that area as a suggestive guide. It is highly likely that a popular talk will have some informative message and learnings thus the listener can be benefitted most.

Future Work

We tried to discuss several features present in the available dataset in our project which led to positive useful inference towards the reason of TED talk popularity still several other areas are there which are left unexplored and can be further analysed to make the future TED talks more popular and informative. Further analysis can be done over the rating column in the dataset to relate the negative comments with topics of TED talk, and find the area of talk which has received more negative feedback. Also, we can check if it inclined toward any specific gender of the speaker. This can then be used as a data to improve the TED talk popularity by improvising on the details found. Apart from this, we can also make some more analysis over topic and area of TED Talk, by combining some other datasets like news article, social media post etc to find for any pattern between how the hot discussed topics over world found from news article and social media post are included in TED talk topics, around the same time frame as of the hot discussion over the world.