Springboard Capstone 1 : TED Talk Text Analysis

Pavan Poosarla; pavanpoosarla01@gmail.com

1 INTRODUCTION

The goal of this capstone project is to automatically assign tags to a TED talk based on its transcript. To do this, we train our model on the existing TED talk dataset. TED talk dataset is taken from the Kaggle repositories.

Kaggle TED talk transcripts : https://www.kaggle.com/rounakbanik/ted-talks/version/3

This data has been scraped from the TED talk official website and is available now by a creative commons license. The database consists of talks scraped till Sep 21st, 2017. It has two files, containing information from about 2550 talks from >350 individual TED and TEDx events across the world. This gives a excellent medium to develop text analysis and NLP techniques to automatically predict sentiments from verbal speech

ted_main.csv : This file contains the metadata associated with each talk. The columns included in this file are as follows

Column	Description
comments	Number of first-level comments on the talk
description	A short note about the talk, usually 1-2 sentences
duration	In seconds
event	Event the talk was presented in
film_date	Unix timestamp of the filming
languages	Number of languages talk is available in
main_speaker	Main Speaker
name	Official name of TED talk, including speaker and title
num_speaker	Number of speakers
published	Unix timestamp of when talk appeared on ted.com
ratings	A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)
related_tags	A list of dictionaries of recommended talks to watch next
speaker_occupation	The occupation of the main speaker
tags	The themes associated with the talk
title	The title of the talk
url	The URL of the talk. This is common column with transcripts file, used for merging the two files on
views	The number of views on the talk

transcripts.csv : It consists of a file consisting of the text of the transcript along with the corresponding url of the talk. The columns on this file are

Column	Description
transcript	This is the transcript of the talk. Text is not divided into paragraphs. It includes comments and audience reactions in parenthesis. For example *(Applause), (Laughter), etc.*
url	URL link of the talk. This is the common column with the *ted_main.csv* used to merge both files on

2 Objective

The overall objective of the project, apart from that of applying machine learning methods to real world dataset and validating the predictions, is to test the efficacy of NLP and text analysis methods for spoken language.

This dataset allows for testing features specific to the spoken word, like words/ minute, audience interactions, etc. In addition, we can also derive metrics to evaluate quality of speech by measuring audience engagement

3 Data Wrangling

The first step in creating a usable dataset to extract statistics from and test machine learning methods is to do data wrangling and output a cleaned dataset. We develop a cleaned dataset, 'df_clean'. To do this, we do the following steps

1. **Join** : The two individual files, 'ted_main.csv' and 'transcripts.csv' are merged on the column url with a inner join. This outputs a dataframe containing information on 2467 talks, each having 18 data columns.
2. **Missing Data** : Looking for missing values in all the columns, we find the dataset to be relatively clean. Only column having missing values is the 'speaker_occupation' column with 6 values missing. We replace missing data with the string 'Unknown'
3. **Remove Outliers :** We also drop the talks with more than 1 speaker. As we have a very few talks with more than 1 speaker, we do not want to contaminate the training data with these talks.
4. **Drop Extra Columns :** Next, we drop the columns which are either redundant or unnecessary for out analysis. The columns dropped and the justification is shown in the table on the next page. Note that some of these columns will provide insights into the data storytelling , but may not be relevant to the machine learning aspects (Categorical with too many categories, etc.)

Format Changes : Another change we do is to convert the date/time fields , i.e, 'film_datestamp' and 'pub_datestamp' into python datetime format. This makes it easier to perform datetime manipulations of these datefields in the future. It also makes these column data human readable.

Dropped Column	**Reasoning/ Justification**
event	Not relevant for learning. Perhaps relevant to data storytelling
languages	Not relevant as we only consider english transcripts. Can be relevant to data storytelling, where it is included.
name	Redundant data. Has title and speaker name
num_speaker	All are 1. Redundant data
related_talks	Dropped. Not relevant
url	Dropped. Not relevant
views	May be too generic a metric to predict or use for training. But included for possible Data Storytelling Assignment and for qualitative validation later on

Once the above operations are completed, we also extract extra features based on analysis of the transcript text. The features extracted from the text are

1. *sentence_count* : Counts the number of sentences in the transcript.
2. *word_count* : Counts the number of words in the transcript
3. *aud_reaction_dict* **:** Creates a dictionary of audience reactions in the transcript (The transcripts contain these in the form of parenthesis)
4. *ratings_dict :* Use a inbuilt function to convert the ratings to a dictionary.

Finally, the two dictionary columns we created, *aud_reaction_dict* and *ratings_dict* are converted to dataframe with multiple columns, with entries in each.

Due to the large number of unique audience reactions (also includes situational descriptions like background music, video playing, etc.), we keep only most popular reactions. Common variants like *laughter* and *laughs* have also been consolidated. The final list of audience reactions are `'laughter', 'applause', 'music', 'cheering', 'sighs', 'video', 'singing', and 'audio'`

Finally, we also separate out the *reaction_dict* into eight columns with the counts corresponding to the following reactions

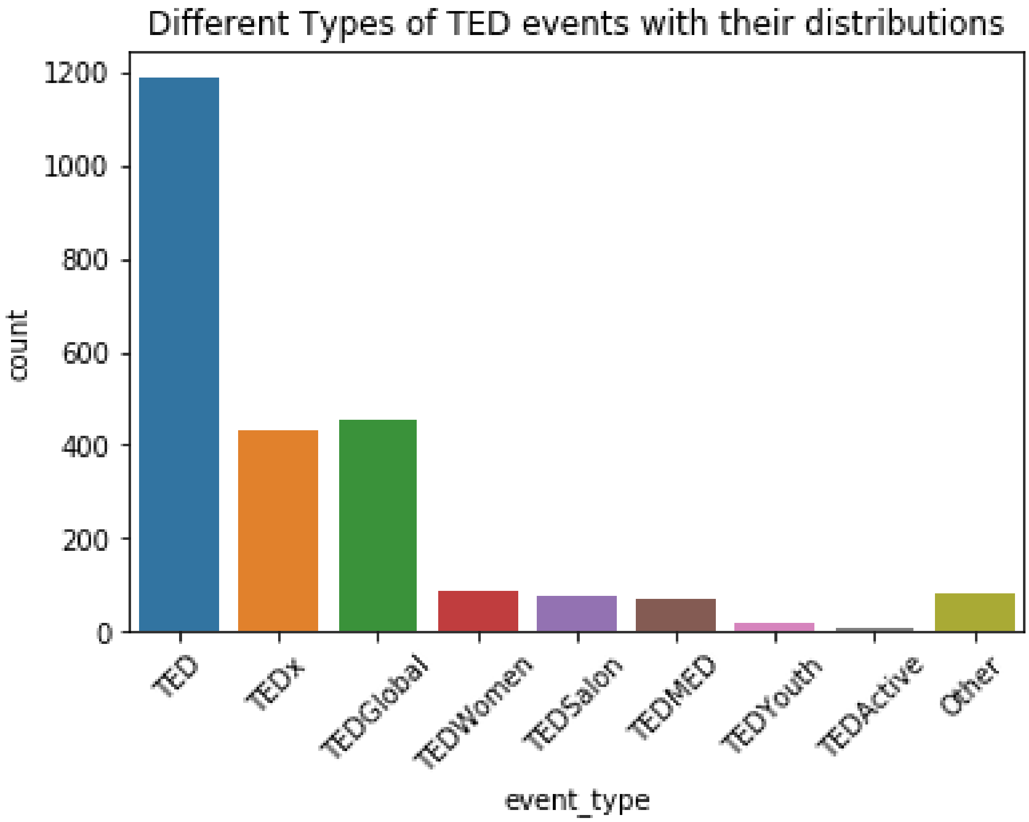
```
'Funny', 'Beautiful', 'Ingenious', 'Courageous', 'Longwinded', 'Confusing',
'Informative', 'Fascinating', 'Unconvincing', 'Persuasive', 'Jaw-dropping',
'OK', 'Obnoxious', 'Inspiring'
```

The final dataframe has 2412 rows and 40 columns. It is written out to a csv file named *'After_DataWrang_Out.csv'.* We will use this file as an input to the data storytelling and the machine learning portions of the project.

4 DATA STORYTELLING

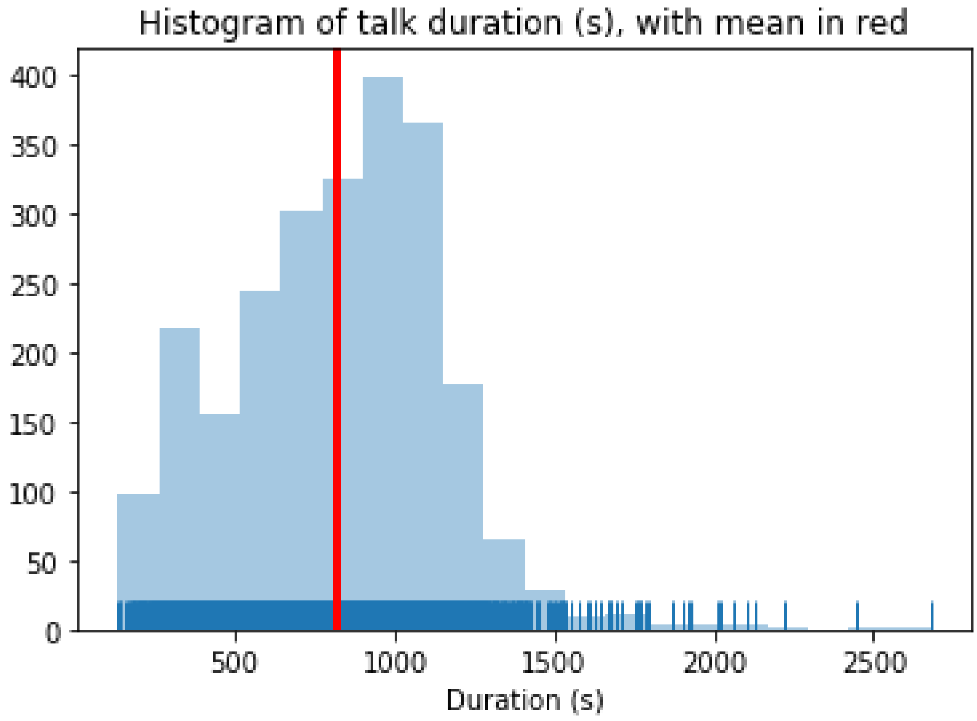
For data storytelling, we start off with the output file from the data wrangling portion of the project. This data is read from the file output after data wrangling (*After_DataWrang_Out.csv*). This file has information on 2412 talks. After dropping redundant columns which have been parsed earlier, each of these talks is left with 35 columns.

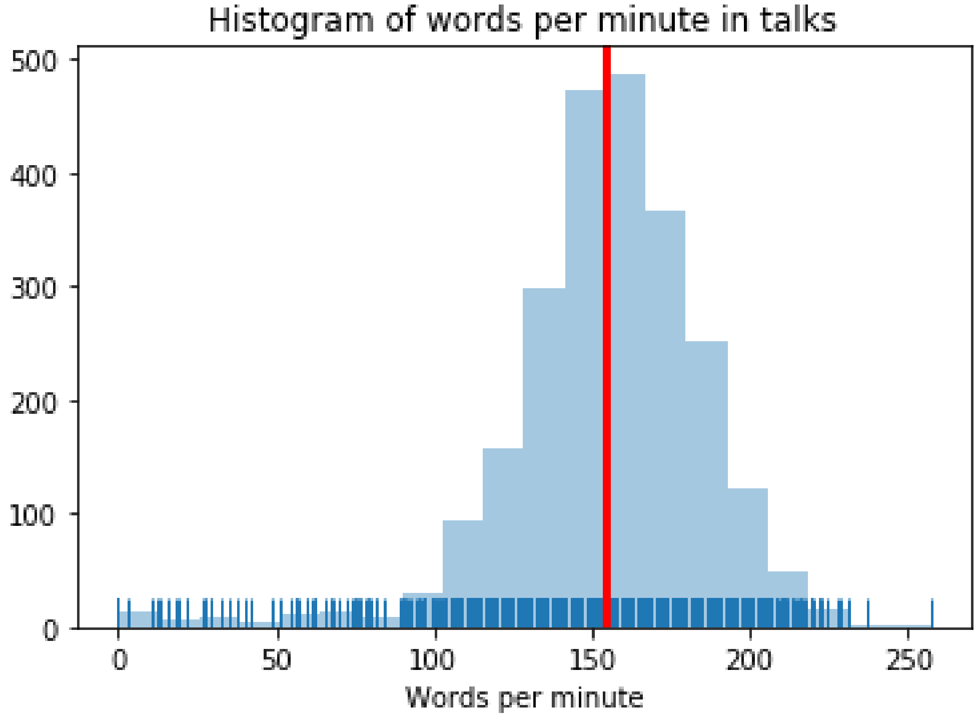
The talks in the dataset are from multiple TED event types. We look at the distribution of talks in each of the major TED categories. The distributions are shown below



We see that based on the number of talks, traditional TED events are the most common. Amongst the variants, TEDx and TEDGlobal are the next most popular versions. Other events, like TED Women, TEDSalon and TEDMED are much smaller in scale. However, TED Youth and TEDActive have the fewest talks of all the events. All other less common variants are grouped under “Other” on the plot.

Looking at the talks themselves, we see the distribution of talk duration and words per minute as shown below. The mean for the distributions are marked in red.

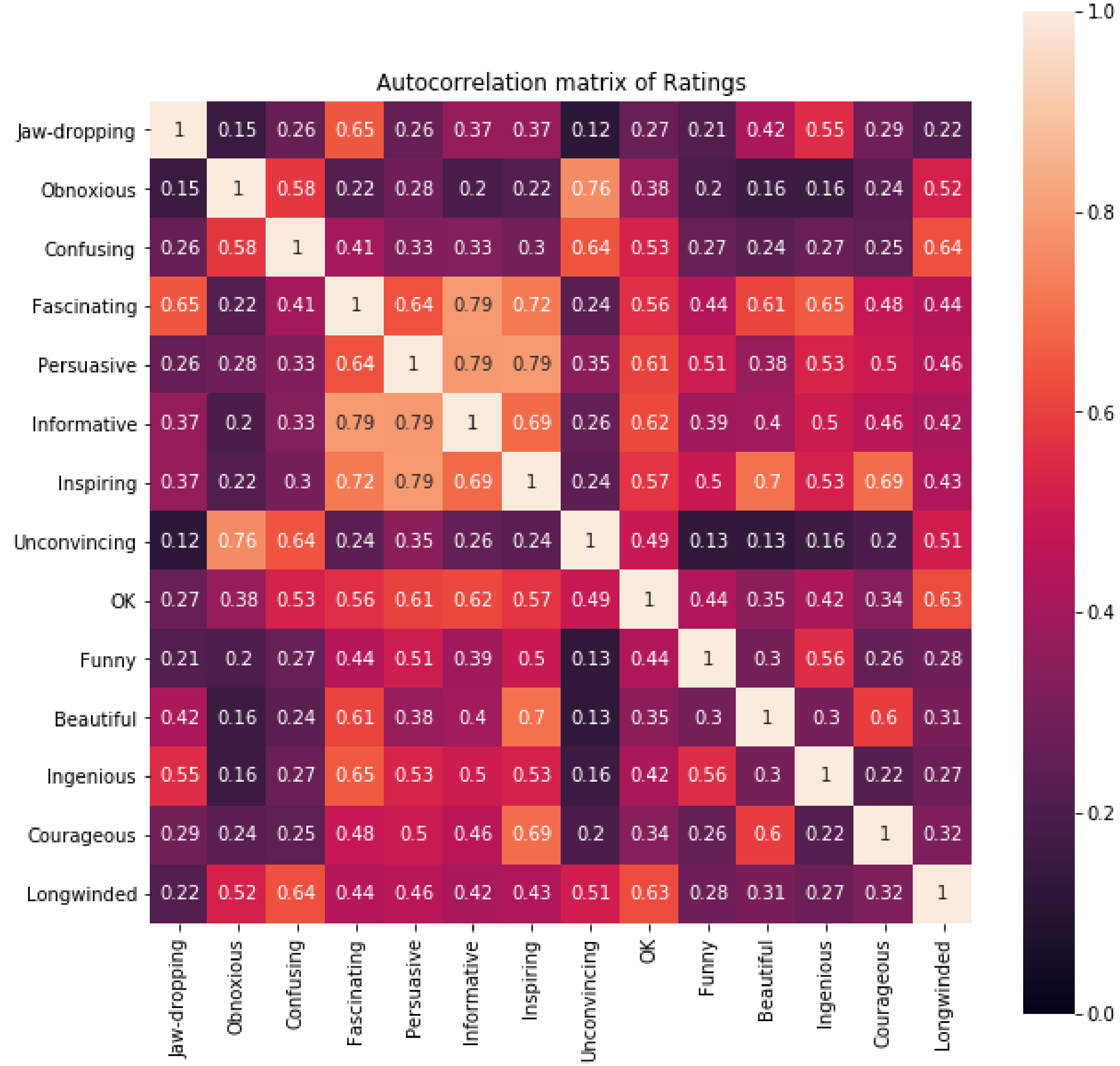




While both have a strong gaussian component, we see that they also have few outliers which are much outside the distribution. Looking at a few examples, we see that these are not traditional talks. These are either musical performances (Those with zero words per minute), product launches, interviews, ad-hoc interviews with notable persons and so on. To avoid these contaminating the dataset, we drop the talks with 'duration' and 'words_per_min' lower than 0.01 percentile and greater than 0.99 percentile.

After filtering, we see that the average duration of a TED talk is about 13 minutes. Most of the talks have a average speed of about 156 words per minute. This is higher than the typical speaking speed for an average person. This is most likely due to the words per minute being contaminated with the audience reactions in the transcript.

To understand the impact of various predictors on our target variable, we will look at the various columns related to ratings. We have 14 columns related to the ratings. We would like to reduce these and also finalize the predictor for applying machine learning to predict the rating. To do this, we look at correlations between different ratings to see which can be dropped and which can be consolidated with others.



	Jaw-dropping	Obnoxious	Confusing	Fascinating	Persuasive	Informative	Inspiring	Unconvincing	OK	Funny	Beautiful	Ingenious	Courageous	Longwinded
Jaw-dropping	1	0.15	0.26	0.65	0.26	0.37	0.37	0.12	0.27	0.21	0.42	0.55	0.29	0.22
Obnoxious	0.15	1	0.58	0.22	0.28	0.2	0.22	0.76	0.38	0.2	0.16	0.16	0.24	0.52
Confusing	0.26	0.58	1	0.41	0.33	0.33	0.3	0.64	0.53	0.27	0.24	0.27	0.25	0.64
Fascinating	0.65	0.22	0.41	1	0.64	0.79	0.72	0.24	0.56	0.44	0.61	0.65	0.48	0.44
Persuasive	0.26	0.28	0.33	0.64	1	0.79	0.79	0.35	0.61	0.51	0.38	0.53	0.5	0.46
Informative	0.37	0.2	0.33	0.79	0.79	1	0.69	0.26	0.62	0.39	0.4	0.5	0.46	0.42
Inspiring	0.37	0.22	0.3	0.72	0.79	0.69	1	0.24	0.57	0.5	0.7	0.53	0.69	0.43
Unconvincing	0.12	0.76	0.64	0.24	0.35	0.26	0.24	1	0.49	0.13	0.13	0.16	0.2	0.51
OK	0.27	0.38	0.53	0.56	0.61	0.62	0.57	0.49	1	0.44	0.35	0.42	0.34	0.63
Funny	0.21	0.2	0.27	0.44	0.51	0.39	0.5	0.13	0.44	1	0.3	0.56	0.26	0.28
Beautiful	0.42	0.16	0.24	0.61	0.38	0.4	0.7	0.13	0.35	0.3	1	0.3	0.6	0.31
Ingenious	0.55	0.16	0.27	0.65	0.53	0.5	0.53	0.16	0.42	0.56	0.3	1	0.22	0.27
Courageous	0.29	0.24	0.25	0.48	0.5	0.46	0.69	0.2	0.34	0.26	0.6	0.22	1	0.32
Longwinded	0.22	0.52	0.64	0.44	0.46	0.42	0.43	0.51	0.63	0.28	0.31	0.27	0.32	1

From the above autocorrelation matrix, we see clear correlations between few of the ratings. For example, most of negative ratings are well correlated. Some of the positive ratings are also correlated (like Inspiring and Persuasive). However, some of the ratings are ambivalent, which might mean that they can be used either with positive or negative connotation (Jaw=dropping, for example). Finally, some positive ratings are not correlated to others. For example, “Funny” and “Inspiring”, although both positive, are not necessarily correlated.

Here is the summary of observations

1. 'Inspring' is strongly correlated to 'Persuasive'. This is understandable based on the similarity in meanings

2. Negative sentiments are highly correlated. Example, "Obnoxious" and "Unconvincing", "Longwinded" and "Confusing", etc.

2. Sone of the ratings seem to have multiple connotations. For example, 'Jaw-dropping' is correlated to 'Fascinating' as well as 'Confusing'. As meaning is unclear, we make a call to drop it, expecting the sentiment to be captured elsewhere

3. 'Funny' is a unique reaction unto itself. So, we do not see any strong correlation to other items. Some talks may be Funny and Persuasive while others may be Funny and Inspiring. However, it is a positive connotation

4. 'Ingenious' is also a independent reaction, not directly correlated to other emotions

5. ‘Informative' is strongly correlated to 'Fascinating' and 'Persuasive'

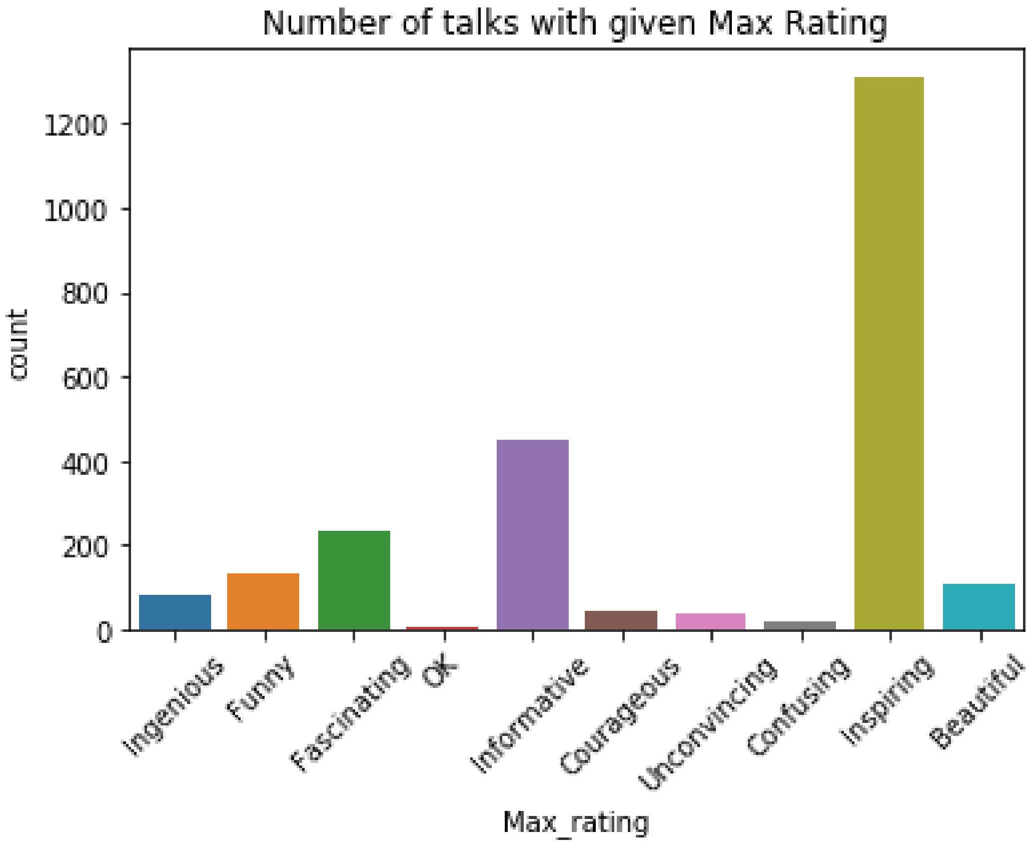
Based on insights from the auto-correlation matrix, we do the following

1. “Inspiring” and 'Persuasive' are very well correlated. We add counts from both columsn and drop “Persuasive”
2. Negative correlations are correlated. We add “Obnoxious” and “Unconvincing”. Also, “Confusing” and “Longwinded”

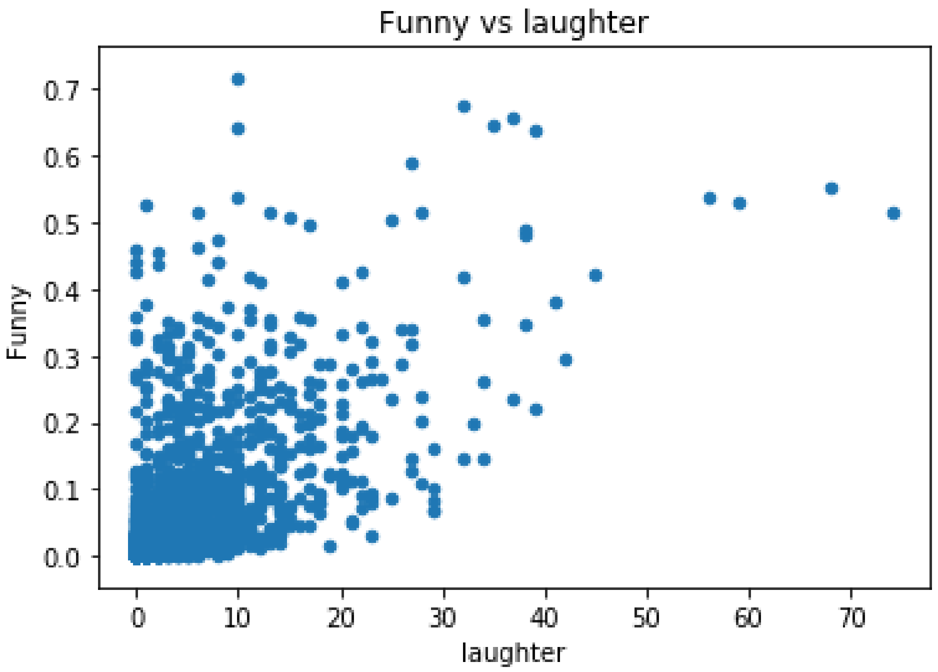
After this, we capture the dominant rating(maximum counts) and look at the distribution. We see that, positive talks are in general more common than negative talks, which is expected as most TED talks have an elaborate screening process. The counts we see are

```
Max_rating
Beautiful        107
Confusing         15
Courageous        46
Fascinating      234
Funny            131
Informative      446
Ingenious         79
Inspiring       1312
OK                 4
Unconvincing      38
```

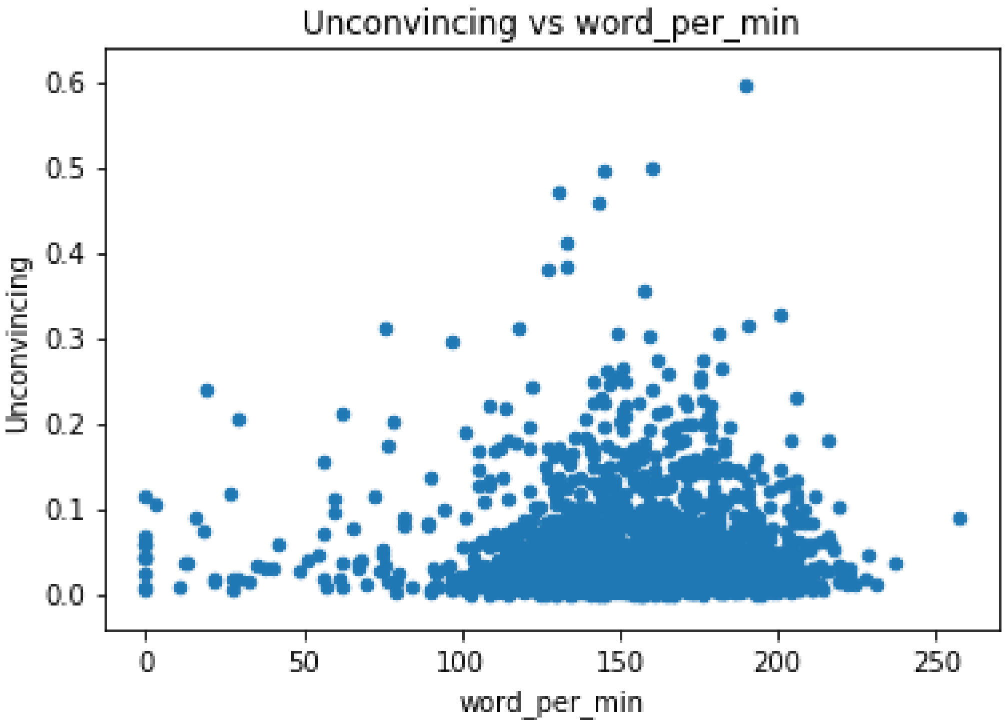
This is graphically shown in the figure below.

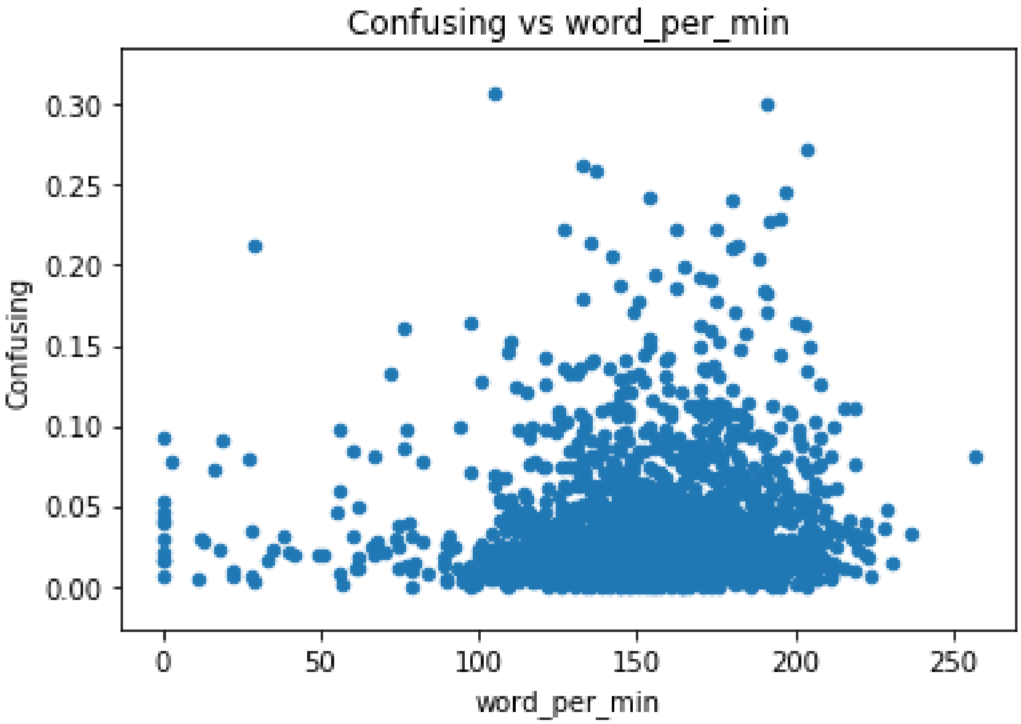


We also see interesting correlations between the ratings and different metadata. As expected, we see that audience are more likely to rate the talk as 'Funny' if there are more audience reactions for 'laughter'.



We also see that negative connotations are correlated to higher words per minute. This may be due to the talks where the speaker is too fast being confusing to the audience.





At this point, the dataset has 2314 talks and 34 columns. Of these, 10 are related to the ratings. They are

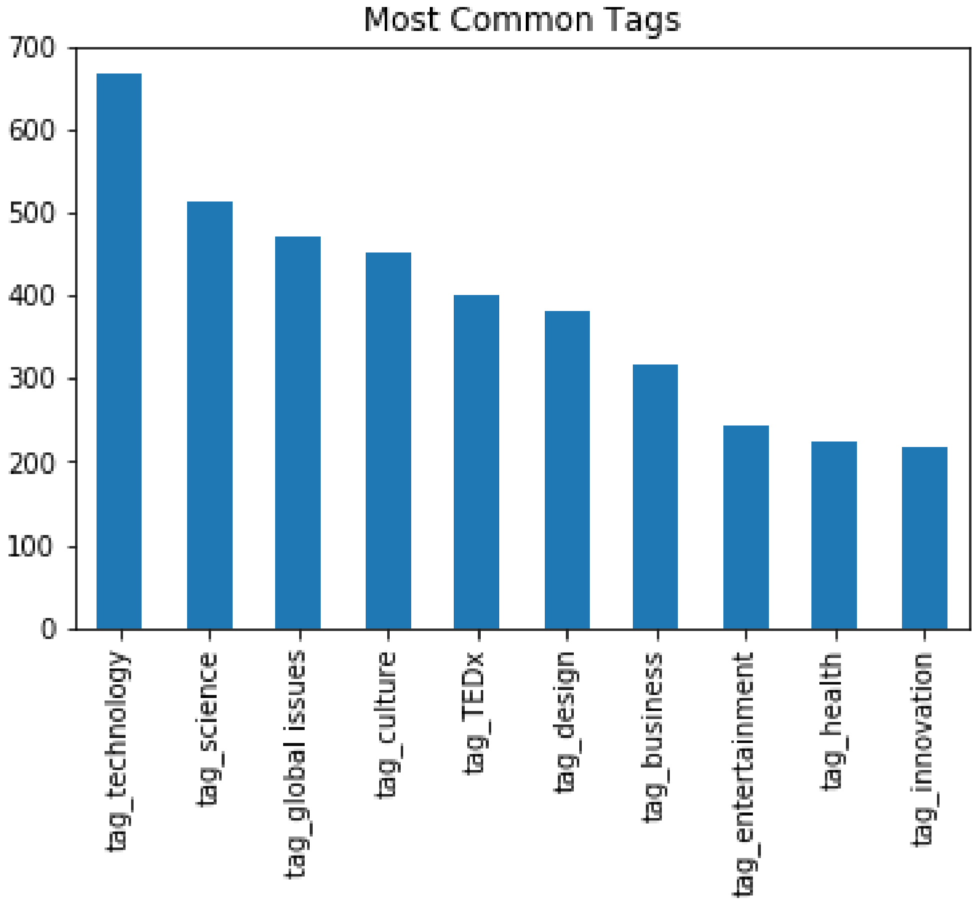
'Courageous', 'Fascinating', 'Ingenious', 'OK', 'Funny', 'Inspiring', 'Confusing', 'Beautiful', 'Unconvincing', 'Informative'

Among these, we will model the main ratings with higher number of examples, like Inspiring, Informative and Funny based on the transcript and metadata.

5 Statistical Data Analysis

For applying machine learning to the rating of the talk, we will try to predict the normalized score of three ratings, 1. Inspiring, 2. Informative and 3. Funny. These are chosen to be distinct characterizations from each other. All three also appear sufficiently often in our dataset.

For statistical data analysis, we would like to evaluate if there is any impact of tags on the ratings. The most common tags in the dataset are shown below



Of these, we create binary columns in our dataset for the following tags : **technology, science, global issues, culture ,design, business and entertainment.** Statistical methods are used to get a quantitative measure of correlation and dependence between various factors and ratings

T-test is used to understand the impact of tags on the ratings. Specifically, t-test is used to answer the following questions,

Q1 : Are talks with tag 'technology' score differently compared to talks without this tag on the rating 'Informative'?

To perform the analysis, we postulate the following null and alternate hypothesis

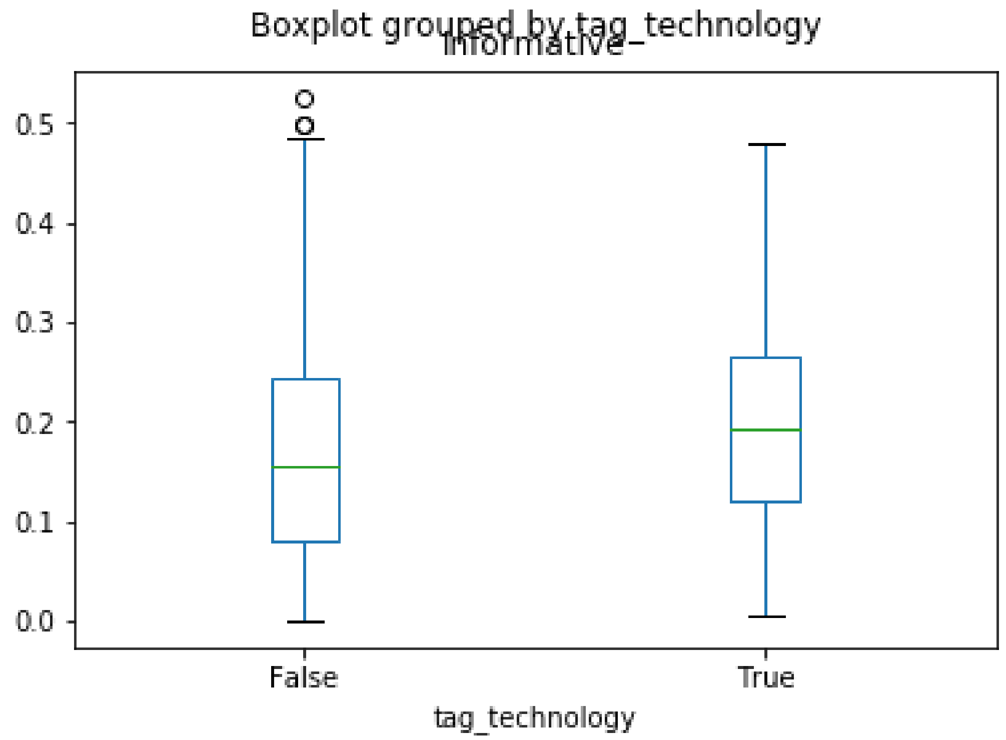
Null Hypotheis : Talks with and without 'technology' tag score similar on 'Informative

Alt Hypotheis : Two groups are different

Performing t-test, the result obtained is

```
Ttest_indResult(statistic=6.227420675983781, pvalue=6.344877075326349e-10)
```

For a alpha of 5%, we can safely conclude that null hypothesis is disproved. Thus, technology tag has an impact on whether a talk is rated as ‘Informative’. As can be seen from the box plot, talks with technology tag rate higher than those without



Similarly, it is also concluded that ‘design’ tag impacts the rating of a talk as ‘Inspiring’. The test result is shown below

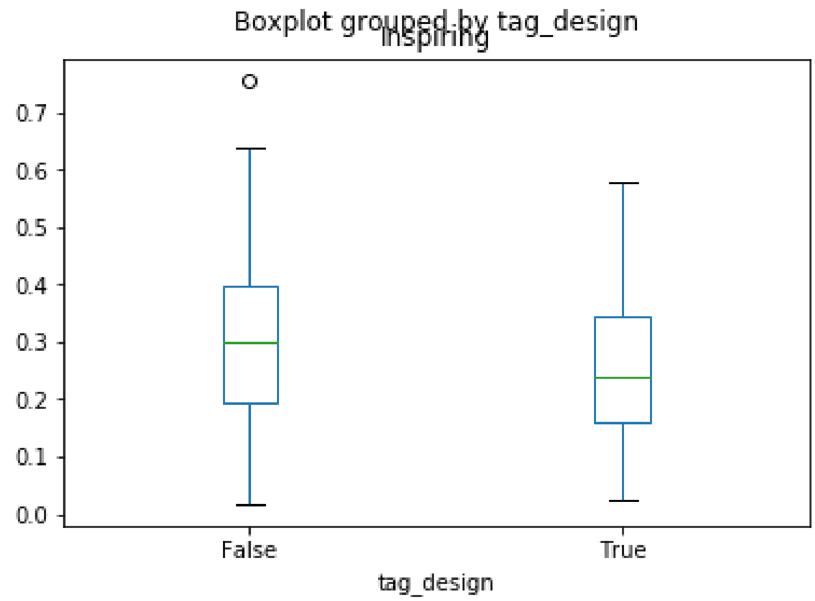
Q2 : Do talks with tag 'design' score differently on 'inspiring' compared to talks without?

Null Hypotheis : Talks with and without tag ‘design’ get rated similarly on ‘Inspiring’

Alt Hypothesis : They rate differently

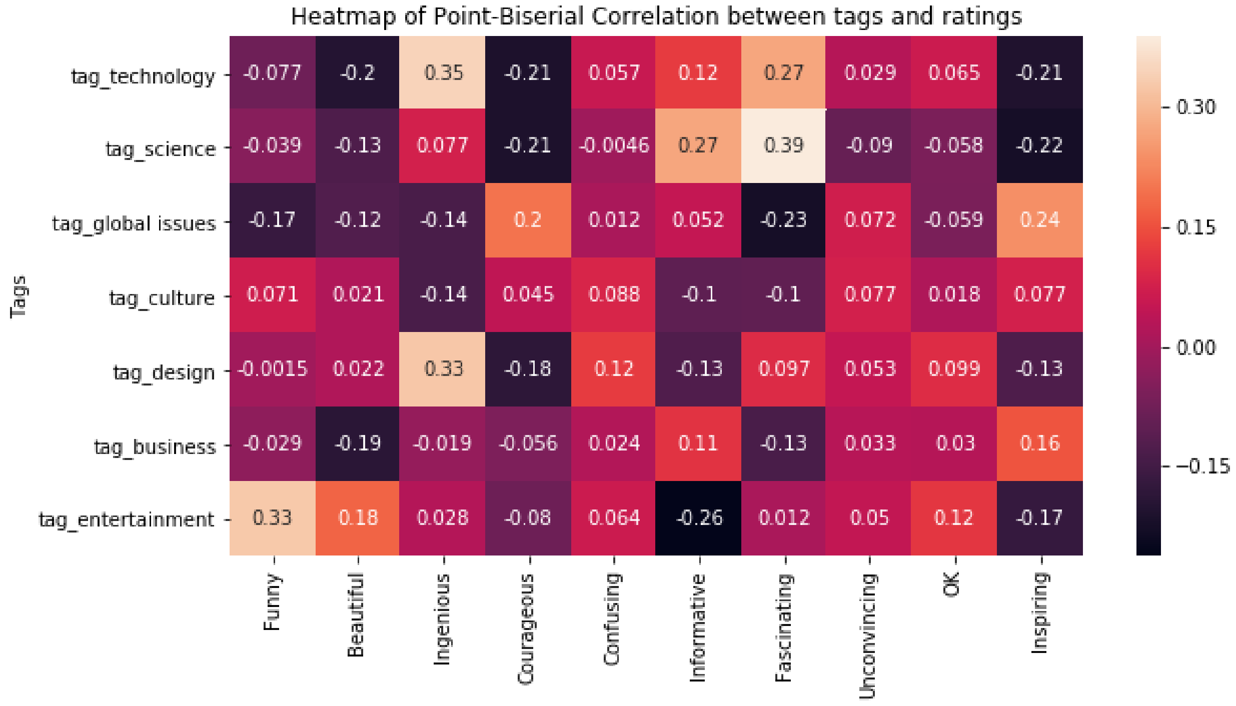
T-test result

```
Ttest_indResult(statistic=-6.615982464579601, pvalue=8.34047782299715e-11)
```



We see that talks with tag ‘design’ get rated differently from those without the tag on ‘Inspiring’. Talks with ‘design’ tag score lower

Finally, we look at correlation between different tags and ratings. As we represent various tags by one-hot encoding, we use point-biserial correlation to calculate the correlation. The result is shown below



From the point-biserial correlation heatmap, we see that the tag ‘science’ is highly correlated to the ratings ‘Fascinating’ and ‘Informative’. Similarly, ‘technology’ and ‘design’ are correlated to ‘Ingenious’. ‘Global Issues’ tend to be correlated to ‘Inspiring’. We also notice that ‘Funny’ is correlated to ‘entertainment’ but not to any other tags considered. Similarly, negative ratings like ‘Confusing’, ‘Unconvincing’ and ‘OK’ are also not correlated to any specific tags