**1/12/2022**

**TED Talks  
NLP/Classification**

Sven Meydell

# Objective

To automatically categorize/tag TED Talk videos and **identify key topics/trends based on a series of features provided** in a sample dataset, link below. Once the text data has been cleaned and prepared, a series Classification models will be trained on the sample dataset.

# Problem Statement

**The data provided, in its raw format, does not make for easy classification and statistical analysis of the various TED Talks**. There is no easy way to decipher trends in popular talk topics, categories, or to further explore specific speakers, their occupations, and general talk focus (transcript summary).

# Benefit of Analysis

Being able to peel back and examine various layers of details within the dataset to **isolate key trends, categories, or any other features of interest would be beneficial to the analyst/user** in answering a series of questions, often addressed by some of the best and brightest minds in the world and their respective TED talks.

Being able to quickly classify and assign a given TED talk, based on other similar feature patterns, would allow for **improved viewer experiences from accurate user preference alignment as it relates to Tags, Ratings, and Comment popularity (counts)**.

# Data Review

# Data Collection

There are two datasets, in csv format, pulled from the Kaggle website provided, the main summary for all TED Talks and the respective TED Talk transcripts:

**TED Talks: Summary** <https://www.kaggle.com/rounakbanik/ted-talks?select=ted_main.csv>

**TED Talks: Transcripts** <https://www.kaggle.com/rounakbanik/ted-talks?select=transcripts.csv>

**Data Provided - Summary:**

* **Comments:** The number of first level comments made on the talk
* **Description:** A blurb of what the talk is about
* **Duration:** The duration of the talk in seconds
* **Event:** The TED/TEDx event where the talk took place
* **Film\_date:** The Unix timestamp of the filming
* **Languages:** The number of languages in which the talk is available
* **Main\_speaker:** The first named speaker of the talk
* **Name:** The official name of the TED Talk; includes the title and the speaker
* **Num\_speaker:** The number of speakers in the talk
* **Published\_date:** The Unix timestamp for the publication of the talk on TED.com
* **Ratings:** Groups of ratings assigned to each talk
* **Related\_talks:** Similar talks/links
* **Speaker\_occupation:** Primary speaker occupation/focus
* **Tags:** Tags assigned to the talk
* **Title:** Title of TED talk
* **Url:** The URL of the talk
* **Views:** Total views for talk

**Data Provided - Transcript:**

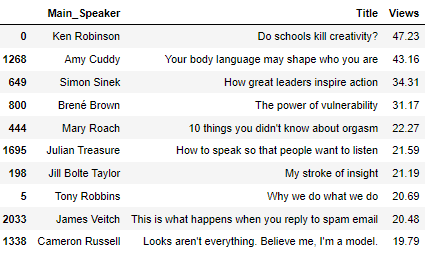
* **Transcript:** The official English transcript of the talk
* **URL:** The URL of the talk

# Data Examination

* The summary file has 2,550 rows of data and 17 columns
* The Transcripts file has slightly less rows (2,467) and only 2 columns
* **Both DataFrames have a URL column**, which can likely serve as the same column for merging both datasets into one overall DataFrame
* There are **no null values** out of the 2,467 entries, however, **Speaker Occupation appears to be missing 6 values** which could be NAs and will require further analysis
* The variables with Object datatypes can be converted to categorical values, saving space and improving EDA presentation

# Data Inspection

### Top TED Talks by Overall Views



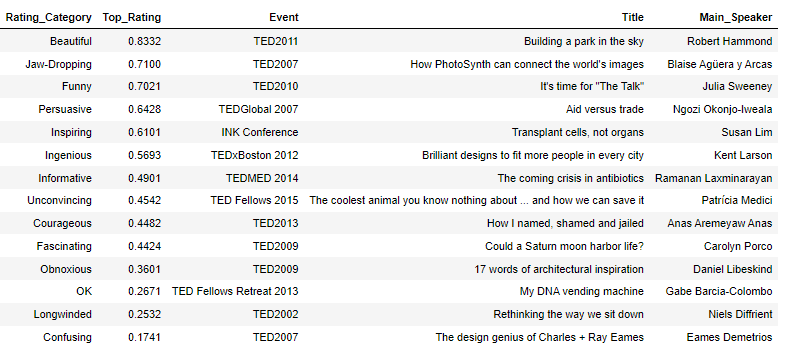
The **top 3 TED Talks of all time**, by total views, are:

* **Do Schools Kill Creativity by Ken Robinson (47.2M)**
* Your Body Language May Shape Who You Are by Amy Cuddy (43.2M)
* How Great Leaders Inspire by Simon Sinek (34.3)

We can see a large drop off in overall views after the top 2 Talks, 43.2M to 34.3M, with **an even larger drop off in total views from 4th to 5th**:

* The Power of Vulnerability by Brené Brown (31.2M)
* 10 Things you Didn't Know about Orgasm by Mary Roach (22.3M)

### Summary Details by Top Rating %



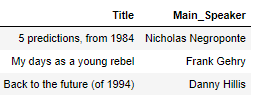
**Top Rated**

* The top rated talk (ranked **Beautiful 83%** of the time) was Building a Park in the Sky by Robert Hammond at the TED2011 event
* The funniest talk (ranked **Funny 70%** of the time) was It's Time for the Talk by Julia Sweeney at the TED2010 event

**Lowest Rated**

* Although The Design Genius of Charles + Ray Eames, by Eames Demetrios at the TED2007 event, was rated as having the highest negative rating for being Confusing, it was **only rated as such 17% of the time which is not a clear indication of overall sentiment**

### First 3 TED Talks Filmed



* The years of: **1984 (first talk)**, 1990, and 1995 each had only 1 TED Talk:
  + **1984: 5 Predictions from 1984 by Nicholas Negroponte**
  + 1990: My days as a young rebel by Frank Gehry
  + 1994: Back to the future (of 1994) by Danny Hillis
  + 1994: Back to the future (of 1994) by Danny Hillis

### Talk Duration Information

* The average TED talk was just under 14 minutes long, with the **maximum length being 1 hour and the minimum being just 2.5 minutes long**
  + Min: The Ancestor of Language by Murray Gell-Mann at the TED2007 event (2.5 minutes)
  + Max: Nationalism vs. Globalism: The New Political Divide by Yuval Noah Harari at the TED Dialogues event (60 minutes)
* The most rated TED talk of all time is:
  + Do Schools Kill Creativity? by Ken Robinson (93,850 total reviews)
* The least rated TED talk of all time is:
  + How your pictures can help reclaim lost history by Chance Coughenour (68 total reviews)
* The majority of talks have an overall primary rating of **Funny (850) or Beautiful (712)**, with **63% of the dataset comprised** of those two categories, 34% and 29% respectively

### Talk Summary Information

* **Hans Rosling** was the most frequent speaker (9 presentations)
* The most common speaker occupation is Writer (45 occurrences)
* The most frequent date for filming, and subsequently publishing, was **4/24/2017 (64 occurrences)**
* The most frequent event (year) is **TED2014 (84 occurrences)**
* 320 unique events listed
* There 65 unique languages and 555 unique comments tracked indicate **unique counts of languages/comments per event (e.g. 4 different languages and 10 comments for a given TED talk)**
* As expected, the majority of speakers per talk is 1 (2,412), however there were 46 cases of 2 speakers presenting

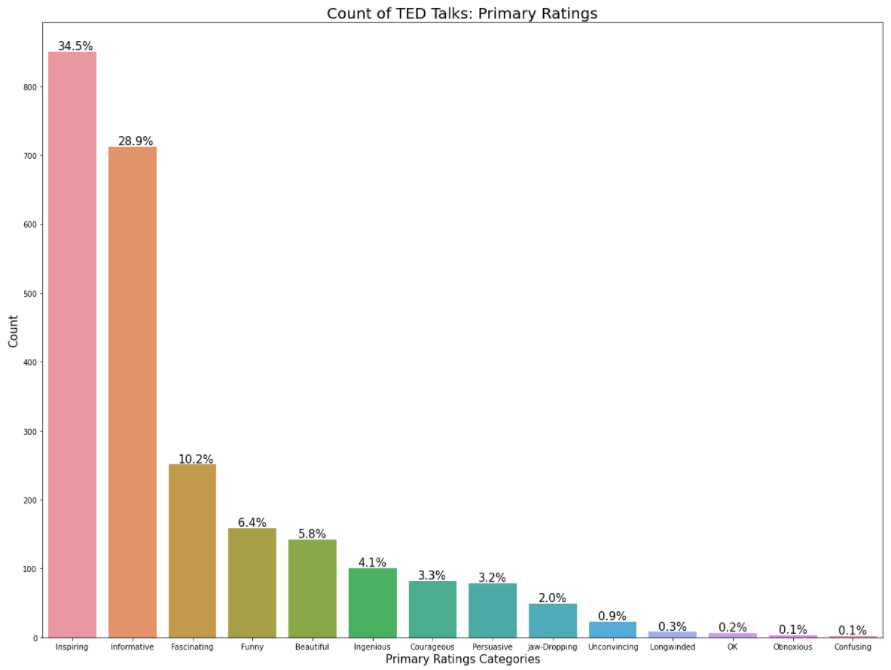
# Data Wrangling

* Merged both datasets on URL column into **TED\_Combined** DataFrame
* Converted all column names to capitalized first letter, each word
* Replaced 6 missing Speaker Occupations via details found online for each
* Converted Date columns to DateTime format and Object datatypes to Categorical
* Divided total Views by 1M and rounded to 2 decimal places
* Divided Duration by 60 seconds to convert to a minutes total
* Removed Unnecessary Columns:
  + Related Talks
  + URL
  + Number of Speakers (Hardly ever more than 1)
  + Name (redundant - concatenation of Speaker and Title)
* **Created Text Statistics summary columns for each row:**
  + **Sentence Count**
  + **Word Count**
  + **Character Count**
* Unpack Ratings & Tags from List and convert to Dictionaries
* Total Ratings column created
  + Sum of each category count within Ratings dictionary
* Total Ratings % column created
  + **Sum of all individual Ratings always equal 100%**
* Individual Ratings Extracted as Columns (% of Total Ratings)
* Unique Ratings column created
* **Extracted the Following from Ratings dictionary:**
  + **Funny, Beautiful, Ingenious, Courageous, Inspiring, Jaw-Dropping**
  + **Longwinded, Confusing, Unconvincing, Obnoxious**
  + **Informative, Fascinating, Persuasive, OK**
* **Primary Rating column created based on max row value for Individual Ratings**
  + e.g. if Funny is highest % of total, Primary Rating will reflect Funny
  + This column will **serve as the Target for predictive modeling**
* Unique Tags column created
* Film and Published Date converted to simple data format
* **Created Transcript and Tag Corpus** for Word Cloud summaries
* **Primary Rating # column created** as a numerical (non-ordinal) target column for predictive modeling

# Exploratory Data Analysis

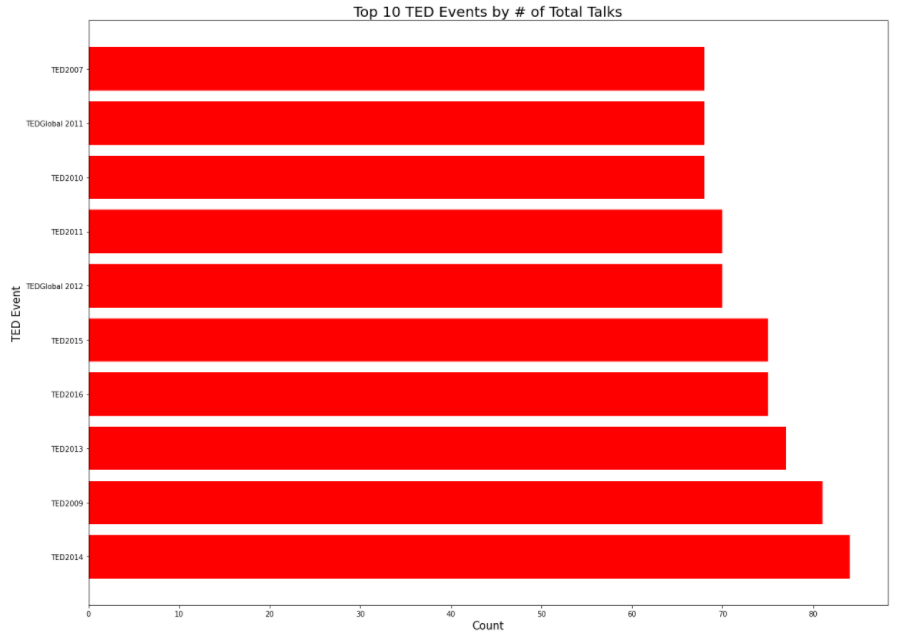
# Univariate

### Count of Ted Talks by Primary Rating (Target)



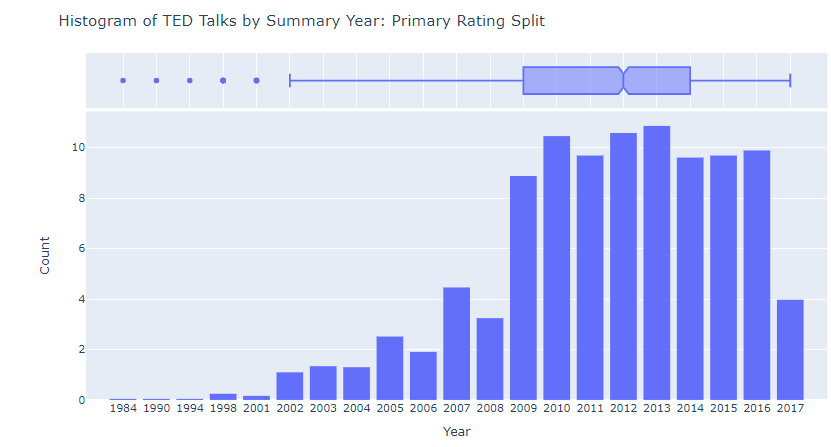
* Over 63% of talks are primarily rated as either **Inspiring (34.5%) or Informative (28.9%)**
* Conversely, **less than 1%** of all talks fall under the following 4 Rating Categories:
  + Longwinded (0.3%)
  + OK (0.2%)
  + Obnoxious (0.1%)
  + Confusing (0.1%)

### Count of Ted Talks by Primary Rating (Target)



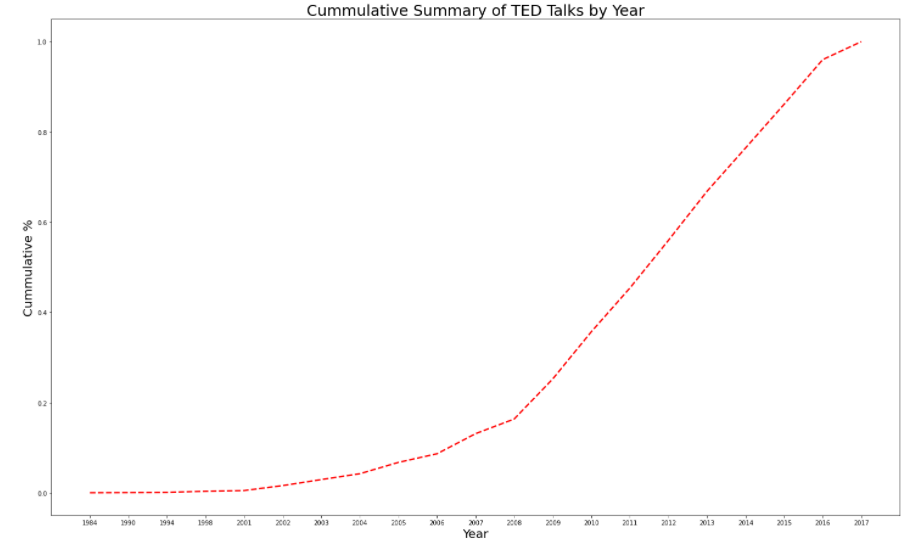
* The most popular event, as it relates to number of talks hosted, was **TED2014, closely followed by TED2009**
* There were numerous TED events with fewer than 10 total talks, some having only 1 unique talk/speaker
  + 1 talk per event - 140
  + 5 or fewer talks per event - 236
  + 10 or fewer talks per event - 276
  + **% of Events with 10 or fewer talks - 11%**

#### Histogram of Talks by Film Year (by Primary Rating)



* Over the 33 years of TED talks, **ranging from 1984 to 2017 (partial through 8/27/17)**, the majority of Talks fall in the years of 2013, 2012, and 2010 comprising of 10.9%, 10.6%, and 10.5% respectively
* Outliers exist on the lower end of the data, before 2002, with 1984, 1990, and 1994 only having 1 TED talk each

### Cumulative Summary of TED Talks by Year



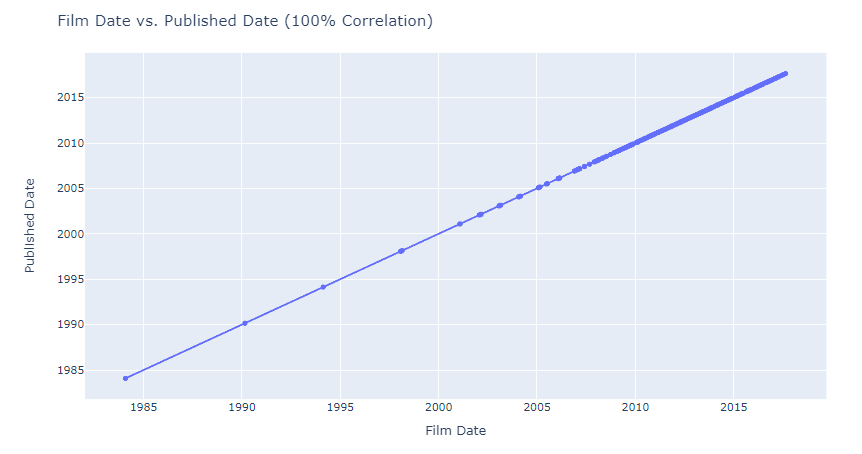
* 50% of the overall TED talks in the sample dataset occurred from 1984 to between 2011/2012 (roughly 28 years), whereas there is a **strong ramp-up from 2012 through partial 2017 (5 years) for the remaining 50% of TED Talks**

### Top 10 Speaker Occupations

Of the top 10 Speaker Occupations sampled, **Writers make up almost 15% of the entire speaker segment**, followed by Journalists (12%) and Artists/Designers (each at 11%)

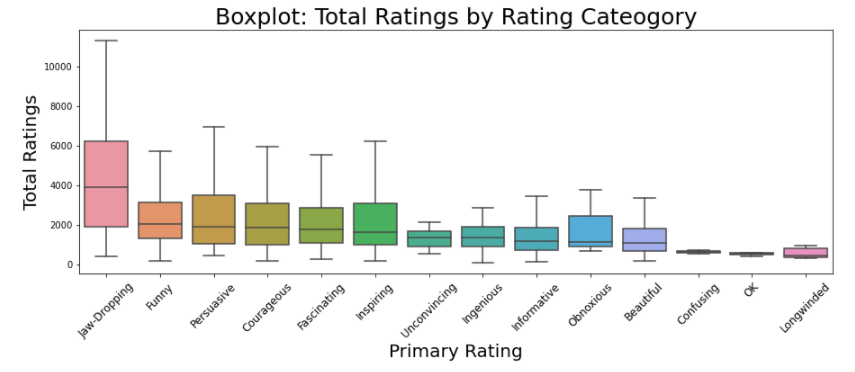
# Bivariate

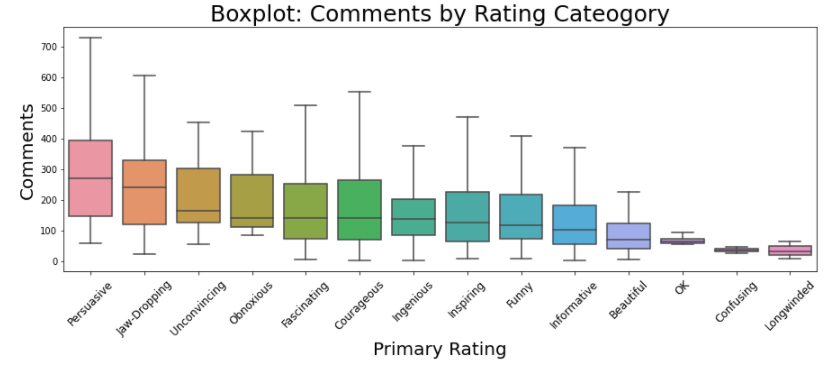
### Count of Ted Talks by Primary Rating (Target)

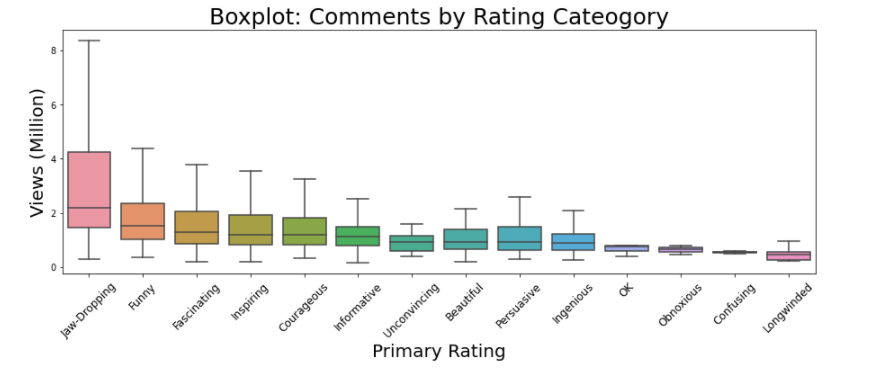


* The plot above confirms that **Film Date and Published Date are equivalent - every TED talk is filmed live and published online that same day** (possibly same time if immediately published after recording)
* There is a **ramp-up in frequency of annual talks hosted per year from around 2008 onwards**

### Ted Talks by Primary Rating (Target): Total Ratings, Comments, & Views

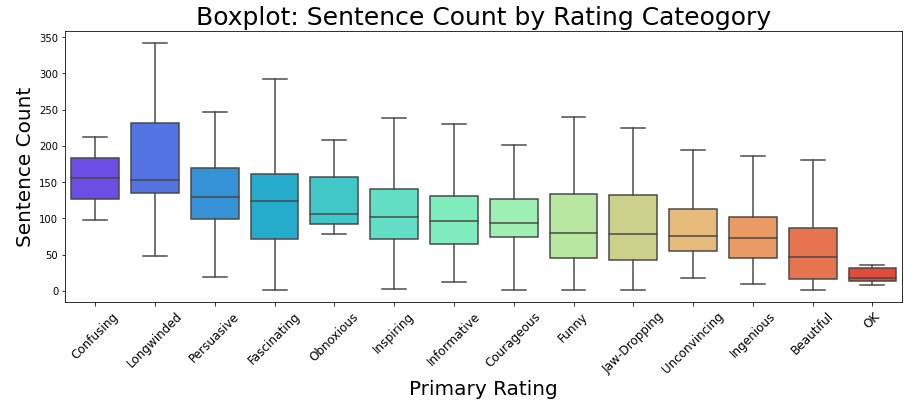


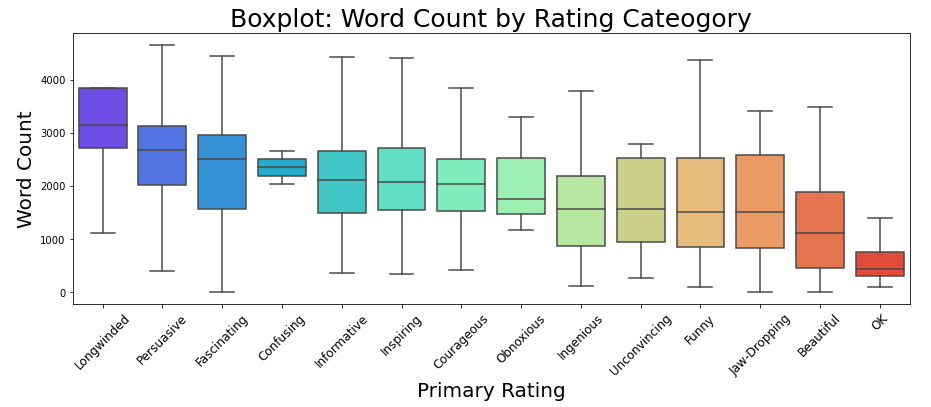


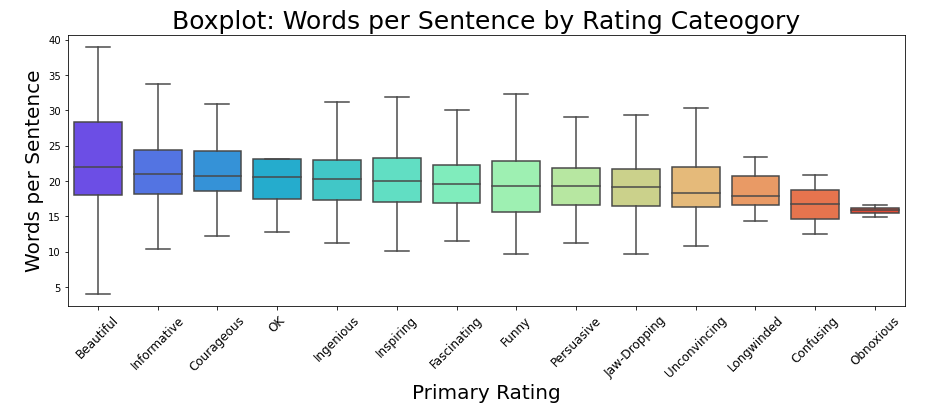


* In general, there are **substantially more positive Rating Categories (9 vs. 5)**
* The majority of Ratings and Comments, particularly outliers are assigned to the **Inspiring** rating category
* Excluding outliers, the **Jaw-Dropping** category has the largest distribution of Ratings, and the **Persuasive** category has the largest distribution of Comments
* The categories for Longwinded, OK, and Confusing, has the lowest distribution of Total Ratings and Comments
  + This, in addition to the large distribution in positive categories, shows that **TED viewers opt to Rate/Comment on a Talk usually do so only usually only when creating a positive reaction for the viewer**
  + **One caveat to this is for talks rated as Obnoxious or Unconvincing**, which do appear to welcome further commentary and ratings form viewers - perhaps invoking stronger feelings and a call to respond
* Talks rated as **Jaw-Dropping have the highest overall views**, substantially higher than all other categories which are relatively similarly distributed
* Talks rated as **Longwinded accounted for the lowest counts for Total Ratings, Comments, and Views**

### Sentence/Word Count and Words per Sentence by Primary Rating



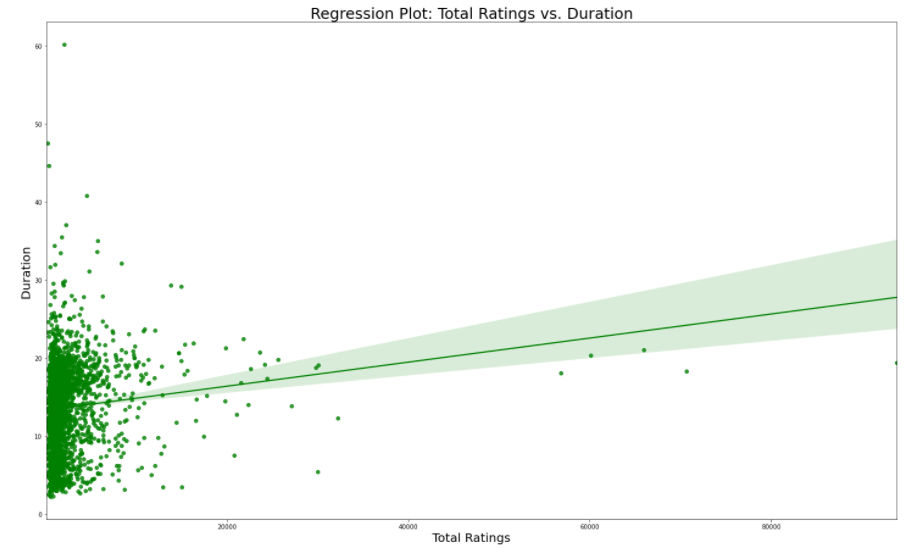


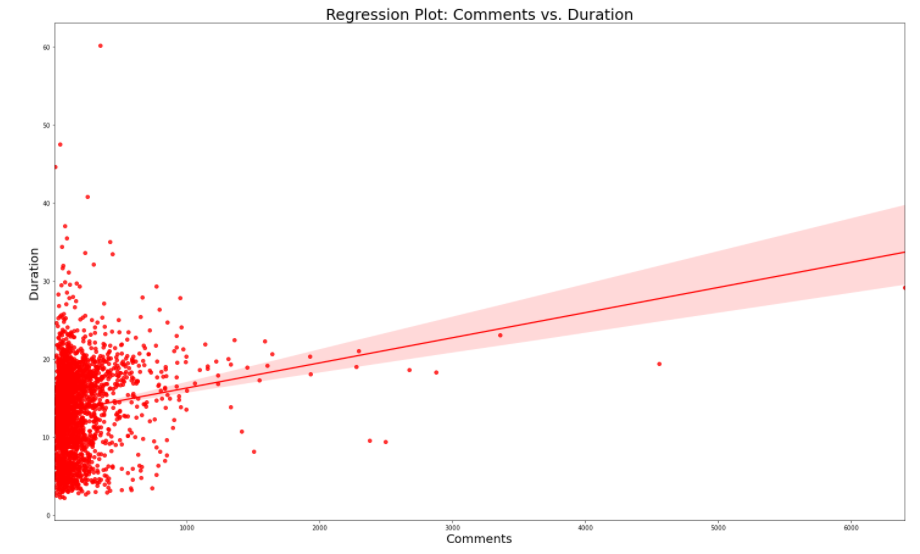


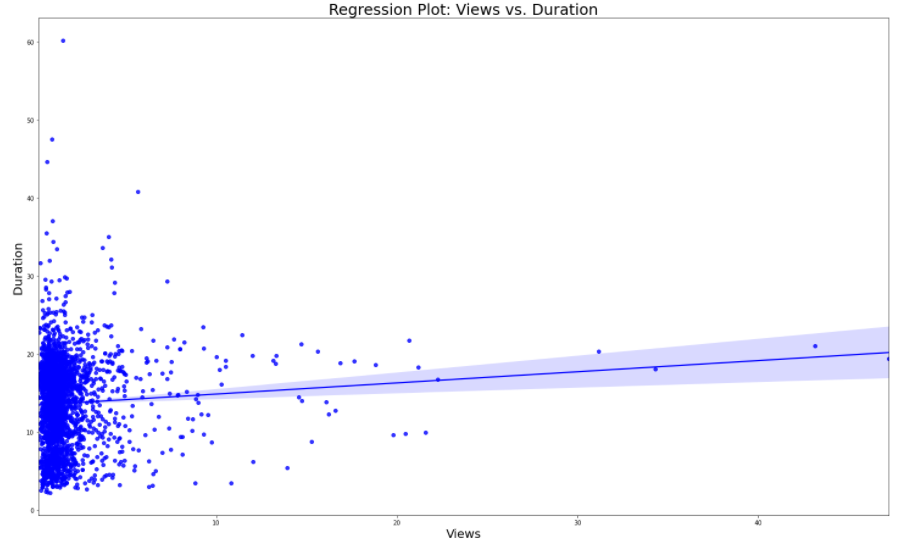
Although **Sentence and Word counts are assumed to be highly correlated for each respective talk**, it is worth further examining the general distribution of each in relation to Primary Rating classifications.

* **Talks classified as OK had the lowest counts for Sentences, Words, and Characters**
  + It seems to show that these talks needed further presentation/delivery in order to impress the audience, but also avoided appearing Longwinded or confusing through being too long
* **Longwinded talks, as the Rating suggests, carried the highest counts for Sentences, Words, and Characters**
  + However, these talks are closely followed by talks rated as Persuasive, Fascinating, and Informative
    - This is perhaps indicative of longer talks, up to a (reasonable) point, showing a greater likelihood of being positively classified
    - There appears to be a **fine line between effective distribution of a message and a Longwinded ramble that could lose audience attention and warrant a negative Primary Rating**

### Correlation of Talk Duration vs. Total Ratings, Comments, & Views



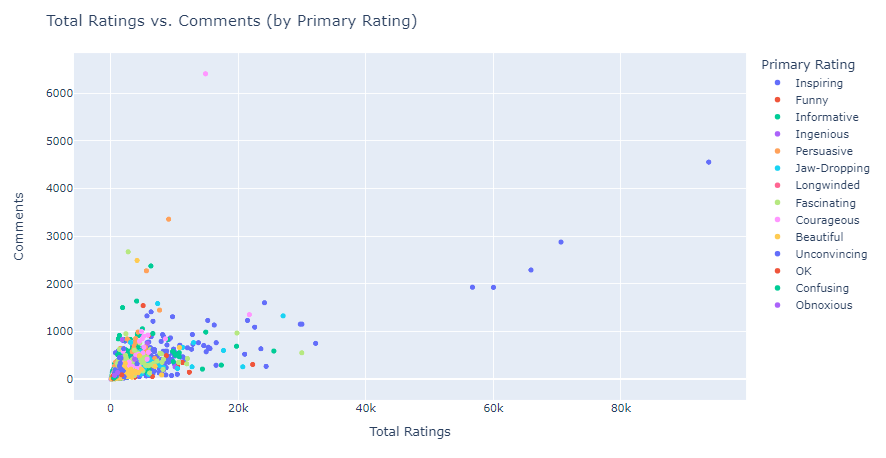


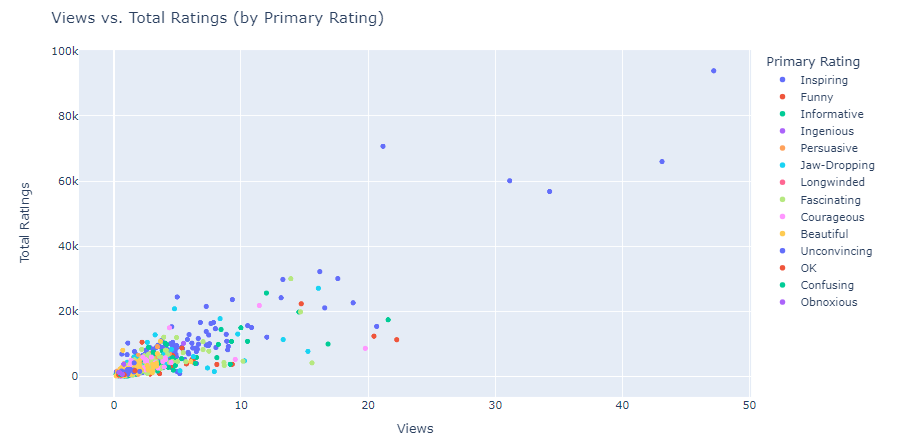


* **Both Comments and Total Ratings have a noticeable positive correlation to Duration**, with Comments have a slightly stronger correlation (higher slope)
  + There is a slight positive correlation between Views and talk Duration with the large majority of talks having a low view count of less than
  + **75% of talks have total views of 1.75M, with the maximum views of 47.2M (Do Schools Kill Creativity by Sir Ken Robinson) largely skewing the dataset**

# Multivariate vs. Target

### Total Ratings vs. Comments & Vies (by Primary Rating)





* There is a slight correlation between increased Rating counts and the amount of Views and Comments created for a given TED talk
* The top rated (quantity) talks are usually those **individually rated as Inspiring, namely the talk by Ken Robinson, Amy Cuddy, and Simon Sinek**
* One notable outlier having substantially higher comments but lower overall views and ratings, is **Militant Atheism by Richard Dawkins** -This talk which was primarily rated as Courageous

### Views/Comments/Total Ratings/Languages vs. Primary Rating

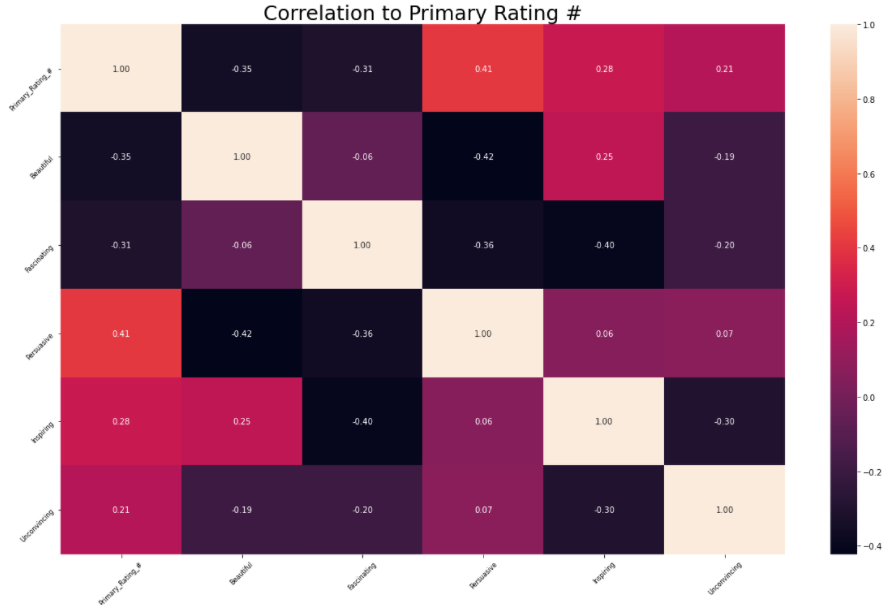
# 

# 

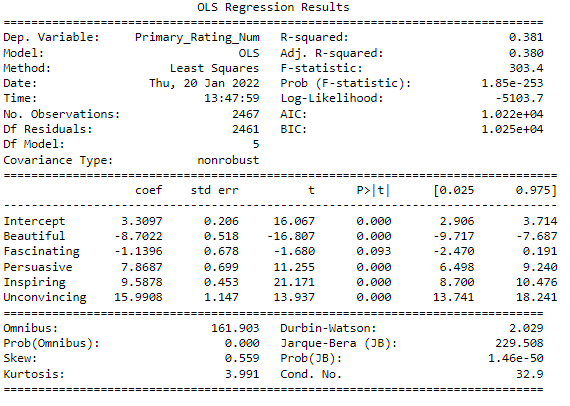
* TED talks classified as **Inspiring show the highest count across almost all metrics, particularly Total Ratings**
  + However, talks classified as Courageous scored the highest Comment counts, followed by talks classified as Inspiring
* Regarding Languages (count), talks classified as: **Confusing, Longwinded, or Obnoxious had substantially lower unique language counts**, with Inspiring, Funny, Persuasive, and OK talks having the highest counts, respectively
* Views and Total Ratings appear to have similar distribution patterns, with Views having more proportionately, as expected since **not everyone who watches a talk will submit a rating**
  + Comments show less of a direct pattern as **certain viewers may be particularly moved by a talk, positively or negatively, and feel compelled to comment, regardless of overall views for that talk**

# Correlation Summary

* **Duration, Sentence Count, and Character Count are all highly correlated**, which makes sense since all are somewhat derived from the initial transcripts detail (character < word < sentence)
* **Views, followed by Comments are strongly correlated (+87% and 64% respectively) to the Total Ratings column**, indicating that the higher the amount of views, and respective comments that follow, the higher the likelihood of increased rating counts
* None of the individual ratings are strongly correlated to the target variable
* Regarding **Rating Categories**, there are a few notable correlations:
  + **Positive**
    - Unconvincing and Obnoxious (+57%)
    - Unconvincing and Confusing (+54%)
    - Confusing and Longwinded (+51%)
  + **Negative**
    - Informative and Beautiful (-52%)
    - Fascinating and Courageous (-50%)
    - Persuasive and Beautiful (-42%)
* **Vs. Primary Rating # (Target)**:
  + **Positive**
    - Persuasive (+41%)
    - Inspiring (+28%)
    - Unconvincing (+21%)
  + **Negative**
    - Beautiful (-35%)
    - Fascinating (-31%)



# Regression Summary - Top 5 Variables Correlated to Target



Although the R-Squared score is only around 40% in the above summary, it shows that **nearly 40% of the variance of the target variable (Primary Rating Number) is explained** by just the 5 Categories/Scores:

**Positive Correlation**

* Unconvincing: Coefficient change of 16 on target for 1 change in variable
* Inspiring: Coefficient change of 9.6 on target for 1 change in variable
* Persuasive: Coefficient change of 7.9 on target for 1 change in variable

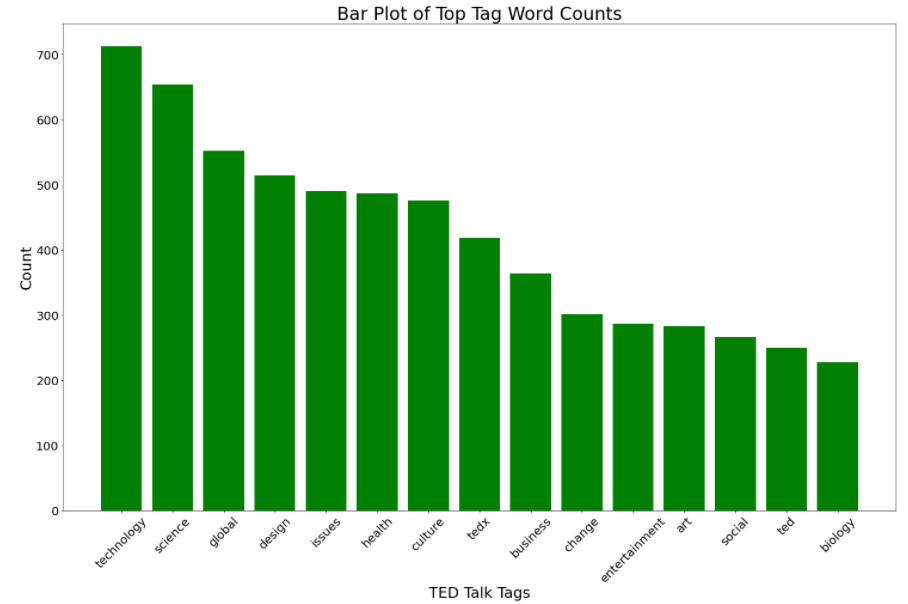
**Negative Correlation**

* Beautiful: Coefficient change of -8.7 on target for 1 change in variable
* Fascinating: Coefficient change of -1.1 on target for 1 change in variable

**Natural Language Processing**

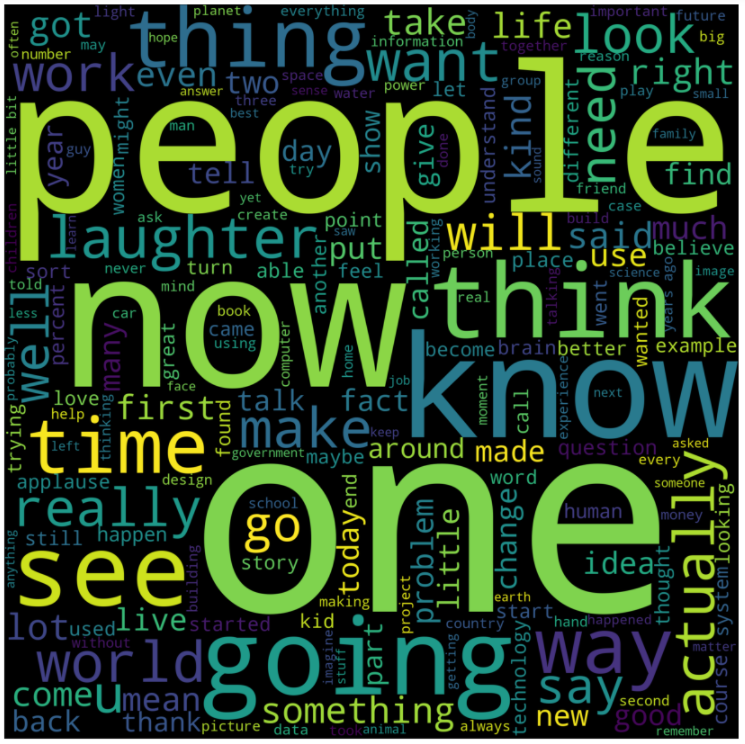
# Word Cloud – Tags

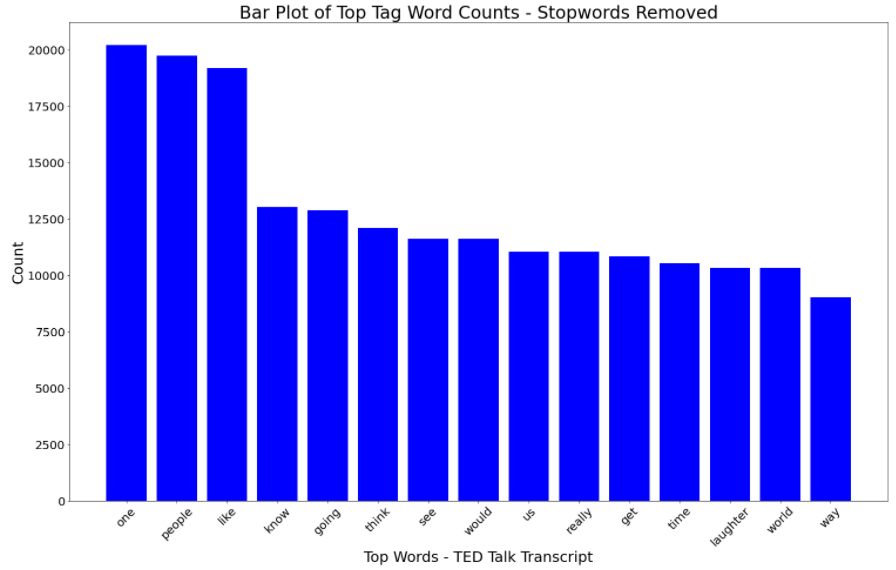




* When looking at the most prevalent individual word tags, we can see that Technology, closely followed by Science are the top tags for all TED Talks
* When looking at top two-worded tags, Global Issues are the top tagged TED Talks
  + When browsing both single/double tags, we can **clearly see that Science, Technology, and Global Issues are key focal points for the large majority of TED Talks**

# Word Cloud – Transcripts





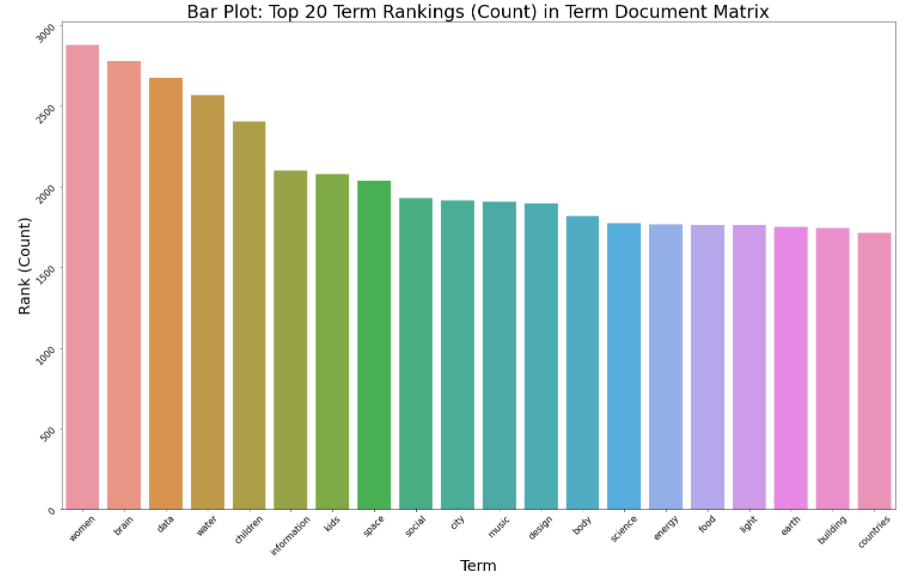
Based on the sizing of key words above, we can see that the following (ordered) are key words/ideas represented within the entire TED dataset:

* Now
* One
* People
* Know
* Going
* Think
* See
* Laughter
* World

These all **encapsulate an overall positive, humanity focused message**, which aligns well with the respective TED Message & Slogan:

* Welcoming People of Every Discipline and Culture who Seek a Deeper Understanding of the World
* Ideas Worth Spreading

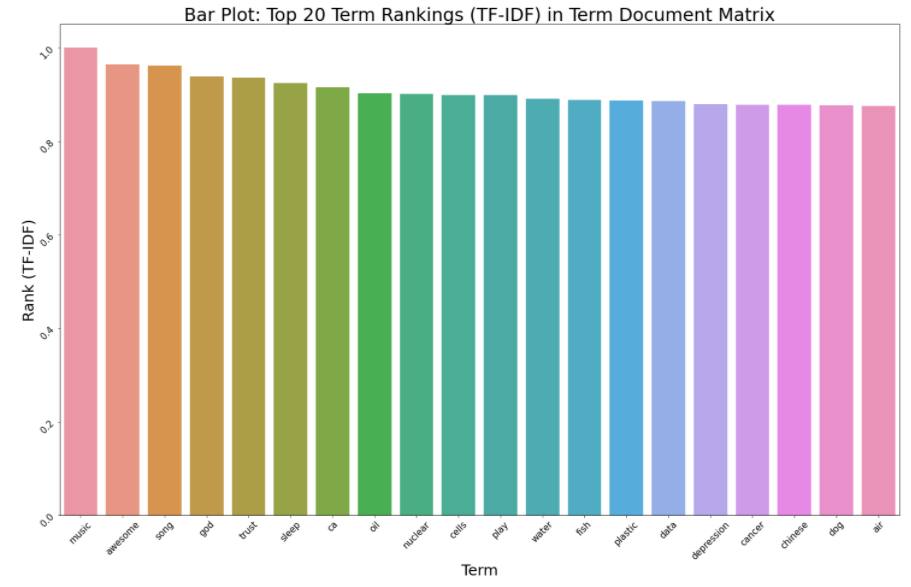
# Term Document Matrix – Total Count:



**From an Overall Count standpoint**, certain keywords/ideas are most prevalent within the total Document Corpus in the Transcript column, including:

* **Humankind/Family**: Women, Children, Kids
* **Science/Learning**: Brain, Data, Information, Science
* **World/Resources**: Water, Space, Energy, Food, Earth, City, Countries
* **Society**: Social, Music, Design, Building

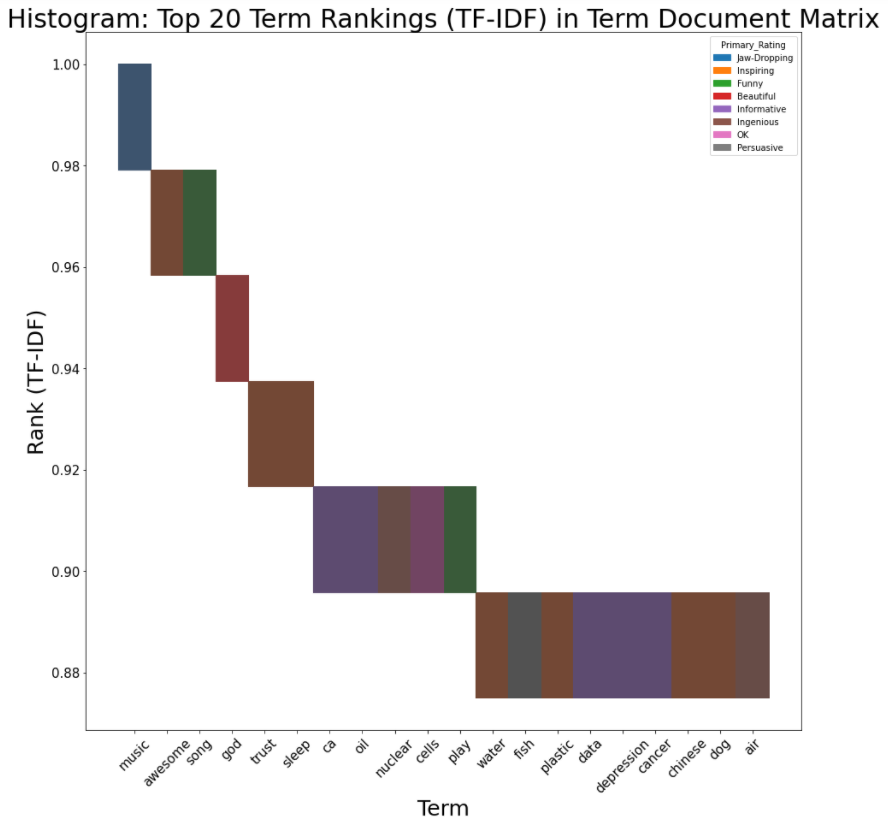
# Term Document Matrix – TFIDF Score:



**From a maximum TF-IDF score standpoint**, certain keywords are most prevalent, however more randomized than when summarizing based on Total Counts, within the total Document Corpus in the Transcript column, including:

* **Musical**: Music, Song
* Awesome (description)
* God
* Trust
* Sleep
* Nuclear
* **Natural Resources**: Water, Fish, Air
* Data
* **Human Issues**: Depression, Cancer
* Chinese
* Dog

# Term Document Matrix – TFIDF Score with Primary Rating:



Adding in the respective Primary Rating categories to the Term Document Matrix, we can see that the top predictive words (based on Maximum TF-IDF Scores) and respective Ratings:

* **Music (Jaw-Dropping)**
* Awesome (Ingenious)
* Song (Funny)
* God (Beautiful)
* Trust & Sleep (Ingenious)