TED Talks Data

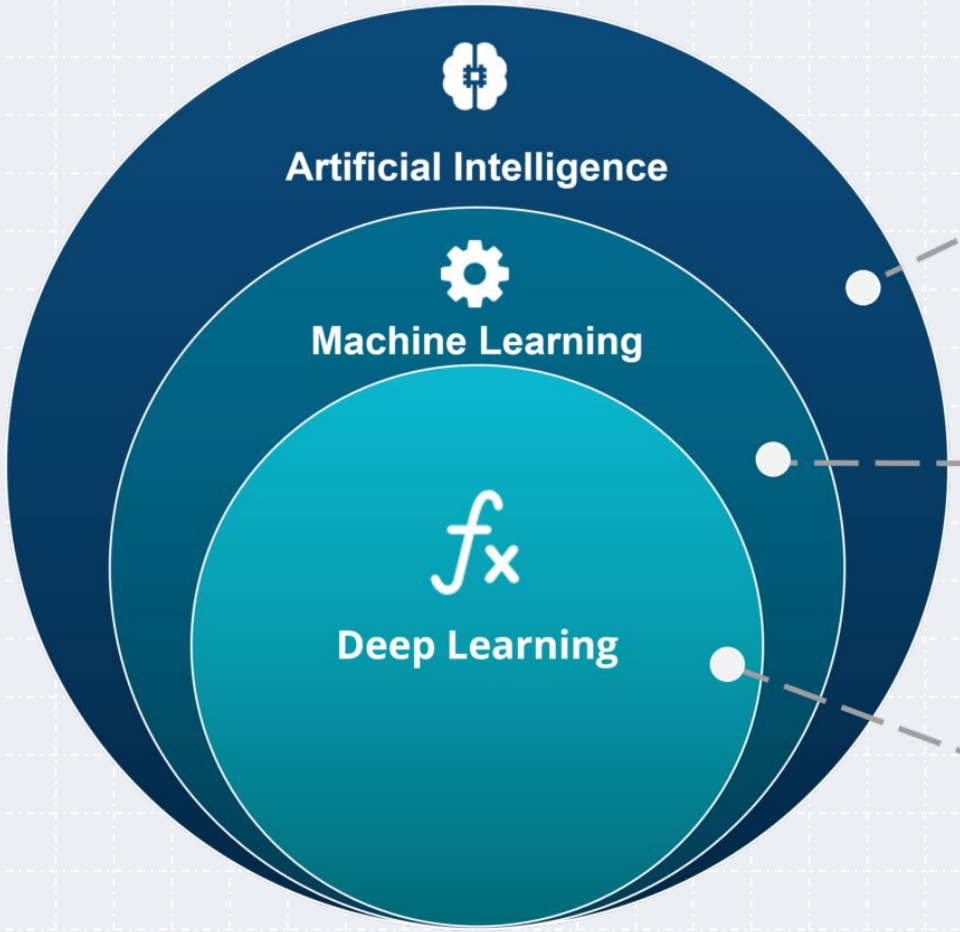
IMT 574 Team Project

Table of contents

01

Intro

A few definitions



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

How Machine Learning Works?

Supervised v/s Unsupervised Learning

How Machine Learning works

Type 1:

Give them labeled pictures of chairs.

After studying those pictures, they should be able to look at new pictures and predict if it's a chair or not.

Supervised Learning

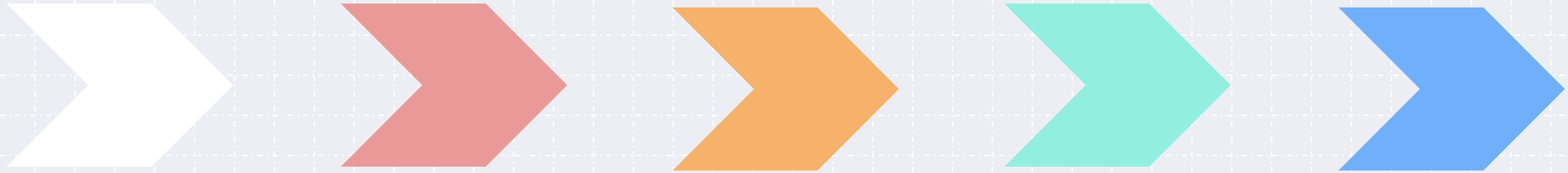
Type 2:

Give them unlabeled pictures of chairs and other items.

They should be able to sort the pictures into groups of chairs or 'not chairs' based on similar characteristics.

Unsupervised Learning

How Data Science Works?



FRAME

Start with a clear, focused question that can be answered with data

PREPARE

Data collection and preparation

ANALYZE

Data acquisition and exploration

INTERPRET

Modelling using various statistical models

COMMUNICATE

Data visualisation and presentation

Business Questions to be answered

1. Analyze and understand what are some of the common characteristics of a TED Talk?
2. What makes a TED Talk successful?
3. What makes a TED Talk popular?

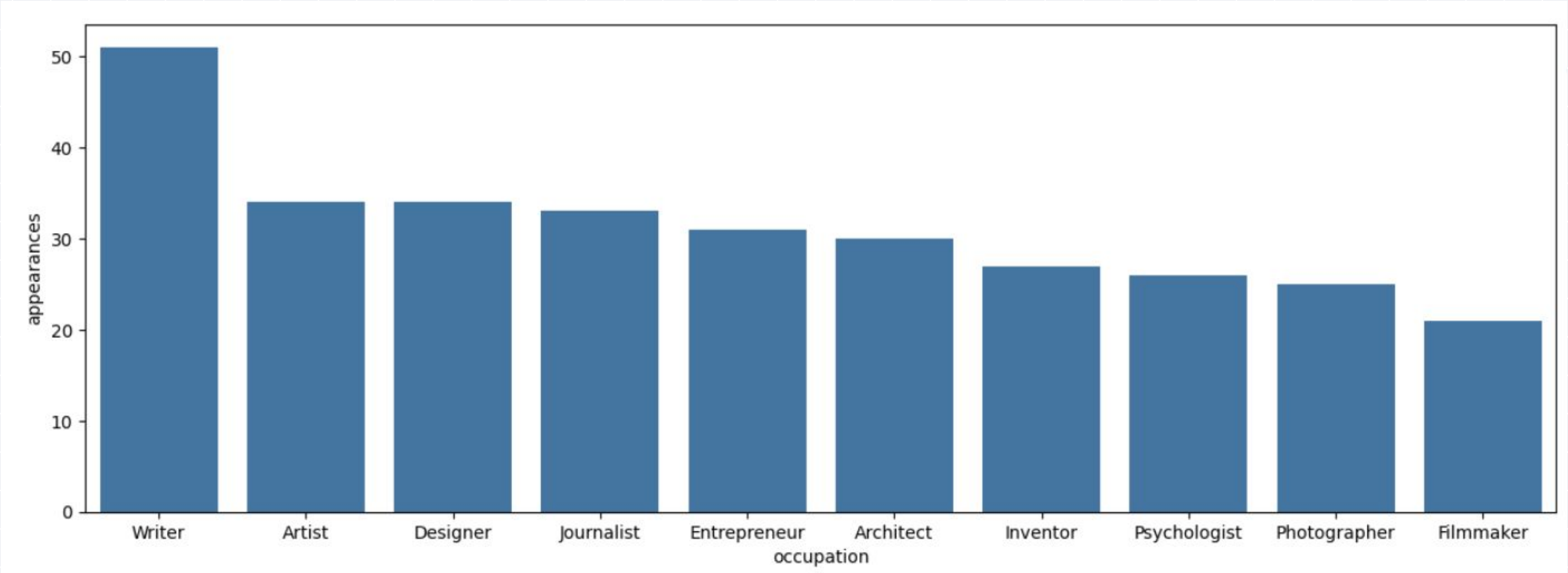
	comments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	ratings	related_talks	speaker_occupation
0	4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	1140825600	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...	[{'id': 865, 'hero': 'https://pe.tedcdn.com/im...	Author/educator
1	265	With the same humor and humanity he exuded in ...	977	TED2006	1140825600	43	Al Gore	Al Gore: Averting the climate crisis	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 544}, {'i...	[{'id': 243, 'hero': 'https://pe.tedcdn.com/im...	Climate advocate
2	124	New York Times columnist David Pogue takes aim...	1286	TED2006	1140739200	26	David Pogue	David Pogue: Simplicity sells	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 964}, {'i...	[{'id': 1725, 'hero': 'https://pe.tedcdn.com/i...	Technology columnist
3	200	In an emotionally charged talk, MacArthur-winn...	1116	TED2006	1140912000	35	Majora Carter	Majora Carter: Greening the ghetto	1	1151367060	[{'id': 3, 'name': 'Courageous', 'count': 760}...	[{'id': 1041, 'hero': 'https://pe.tedcdn.com/i...	Activist for environmental justice
4	593	You've never seen data presented like this. Wi...	1190	TED2006	1140566400	48	Hans Rosling	Hans Rosling: The best stats you've ever seen	1	1151440680	[{'id': 9, 'name': 'Ingenious', 'count': 3202}...	[{'id': 2056, 'hero': 'https://pe.tedcdn.com/i...	Global health expert; data visionary

02

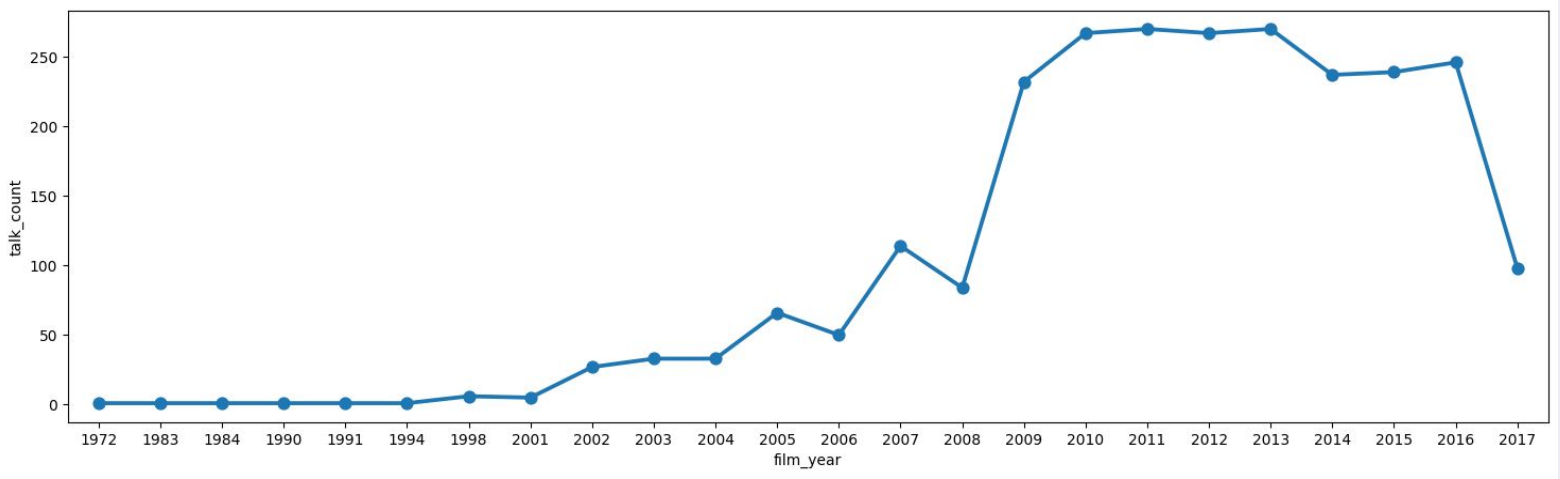
Explorative analysis

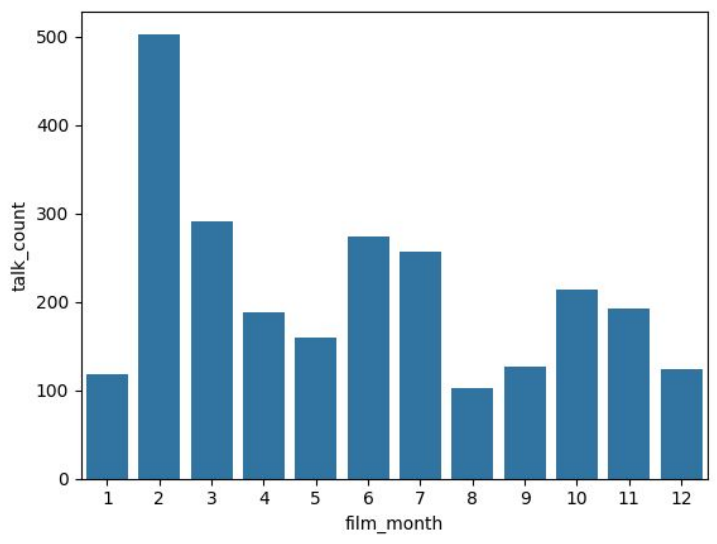
Common characteristics of TED talk data

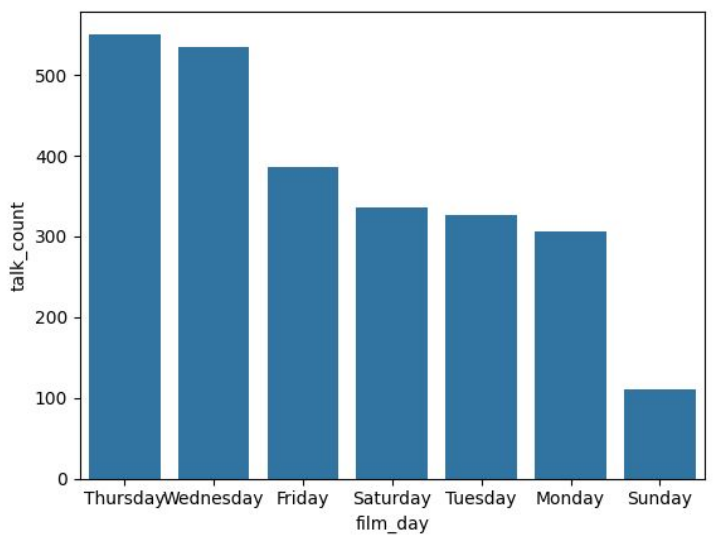
Most popular speaker occupations



Most popular filming time of Talks





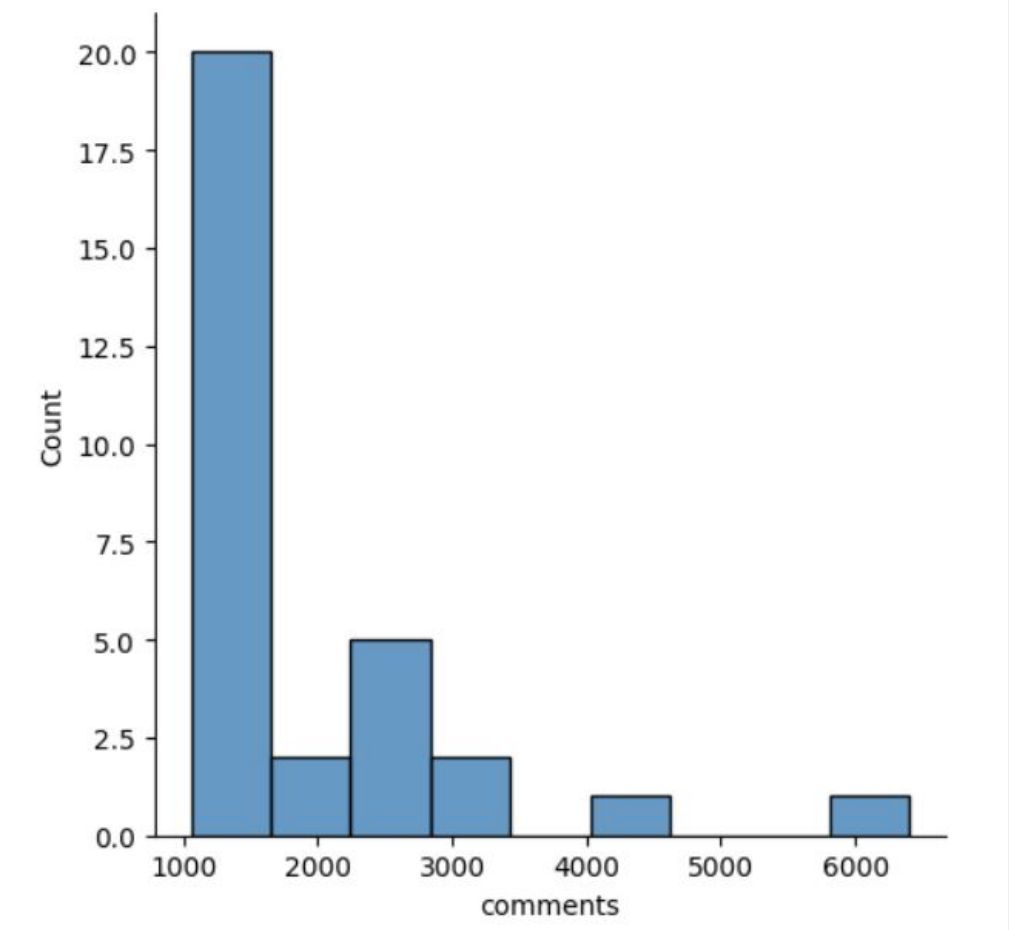


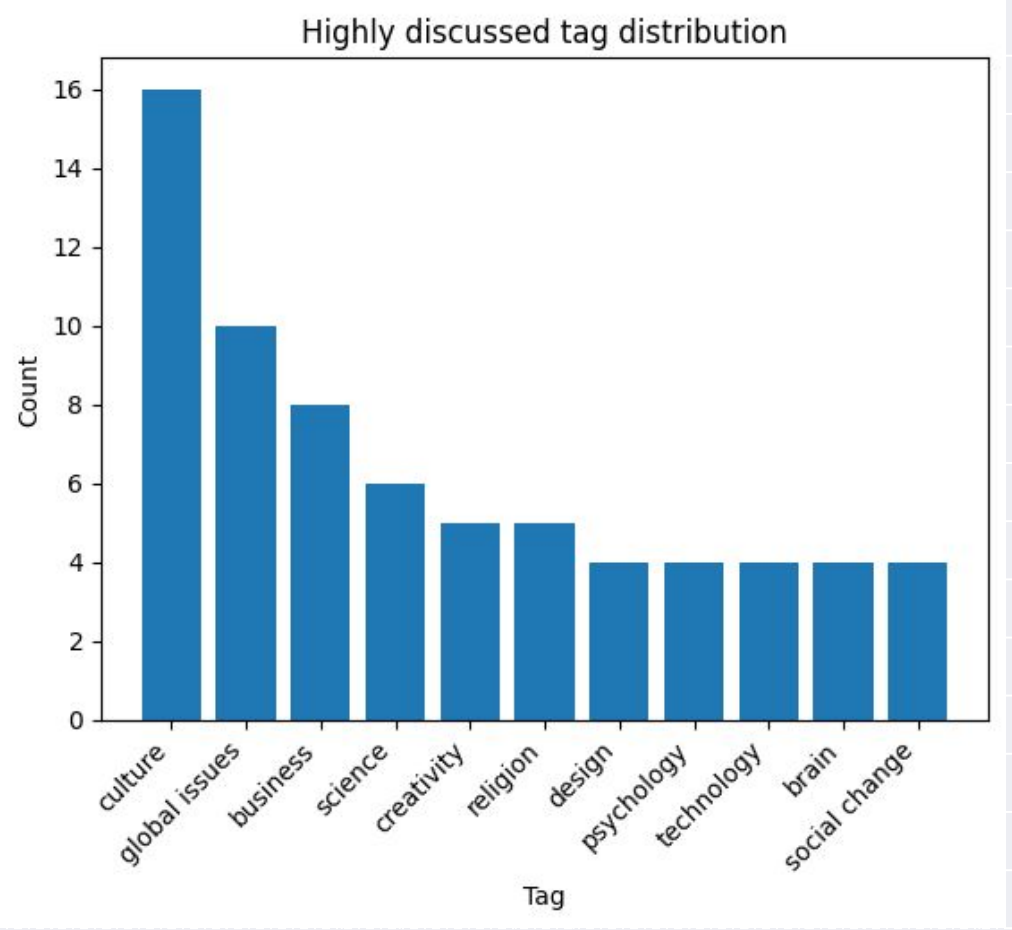
- Year: There was a sharp increase in the number of talks starting in 2008, and the number of talks stayed at a relatively steady level during 2009 to 2016.
- Month: The most popular month is **February**.
- Day: **Wednesday and Thursday** are the most popular days, while Sunday is the least popular day.

Most viewed Talks

- The average number of views per TED talk is **1.6 million**.
- The median number of views per TED talk is **1.2 million**.
- Among the 2550 Talks, only 2 talks win more than 40 million views, and additional 4 talks that win more than 30 million views.

Most discussed Talks

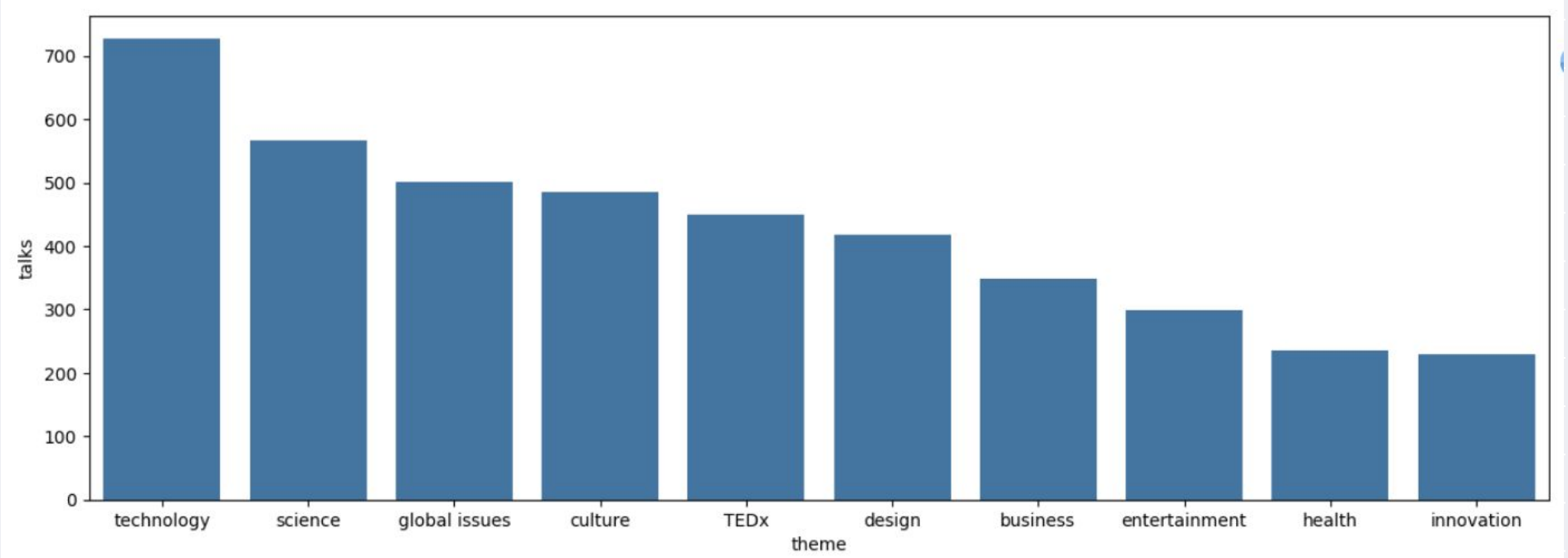




For the highly discussed talks, 32 Ted talks have more than 1,000 comments. Their prevalent themes include **culture**, **global issues**, **business**, **science**, **creativity**, **religion**, **design** and etc.

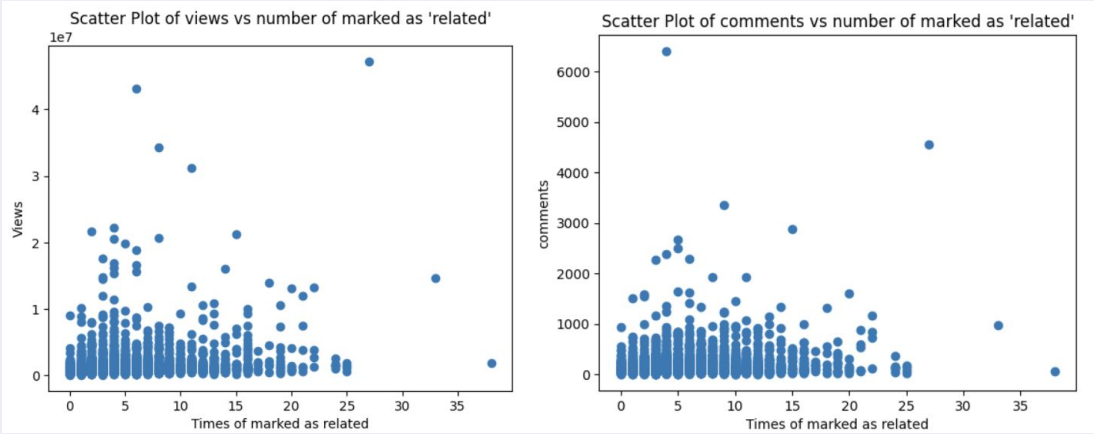
Most popular themes among Talks

	theme	talks
14	technology	727
12	science	567
11	global issues	501
2	culture	486
293	TEDx	450
54	design	418
25	business	348
16	entertainment	299
37	health	236
148	innovation	229



- Based on the tags of the Ted talks, we extracted the popular themes.
- Top 3 most popular themes of TED talks are **technology**, **science**, and **global issues**. They are the only ones that appear more than 500 times, collectively constituting 70% of the total 2,550 talks.

How each TED Talk related to others?

Hypothesis	Method used	Finding
The similarity of talks to be related would be affected by **user engagement**(Views, Comments).	Pearson correlation coefficient	• The correlation scores are both around 15% 
The similarity of talks to be related would be affected by the **specific theme**.	We use BERT (Bidirectional Encoder Representations from Transformers) to vectorize the title texts, identifying the six closet data points for each TED Talk. Subsequently, we calculate the accuracy by comparing the match counts with the actual related talks.	• The matched talks account for 13.96% of all related talks • There is 51.37% talks that is correctly matched at least one related talk.

Reflection on the model results

- **What is the cause of the low accuracy of the model?**
  - Model?
  - Feature?
- **What can we do later?**
  - Transcript(sentiment analysis, text analysis)
  - Combine features to build a single model
    - Views
    - Comments
    - Main title

03

How do you determine if a Ted Talk is Successful?

APPROACH

The question we have in front of us is a very subjective one. How does one evaluate success, especially that of a Ted Talk? Here is how we decided to approach it and our justification behind this:

- We decided to focus on the categorical columns and use Natural Language Processing instead of using the numerical parameters.
- We identified the ratings column as being the most relevant metric of success and used sentiment analysis for classifying.
- Using parameters like views and likes would tell us more about the popularity of the talk rather than success, and this is something which we would be explaining further.

DATA AND RESULTS

The Ratings Data:

```
df_main['sorted_ratings'][0]
```

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

Results Obtained:

```
df_main['sentiment_of_top_three_words'].value_counts()
```

```
positive    2471
neutral       76
negative       3
Name: sentiment_of_top_three_words, dtype: int64
```

```
df_main['sentiment_of_top_five_words'].value_counts()
```

```
positive    2547
negative       2
neutral        1
Name: sentiment_of_top_five_words, dtype: int64
```

```
df_main['sentiment_of_top_ten_words'].value_counts()
```

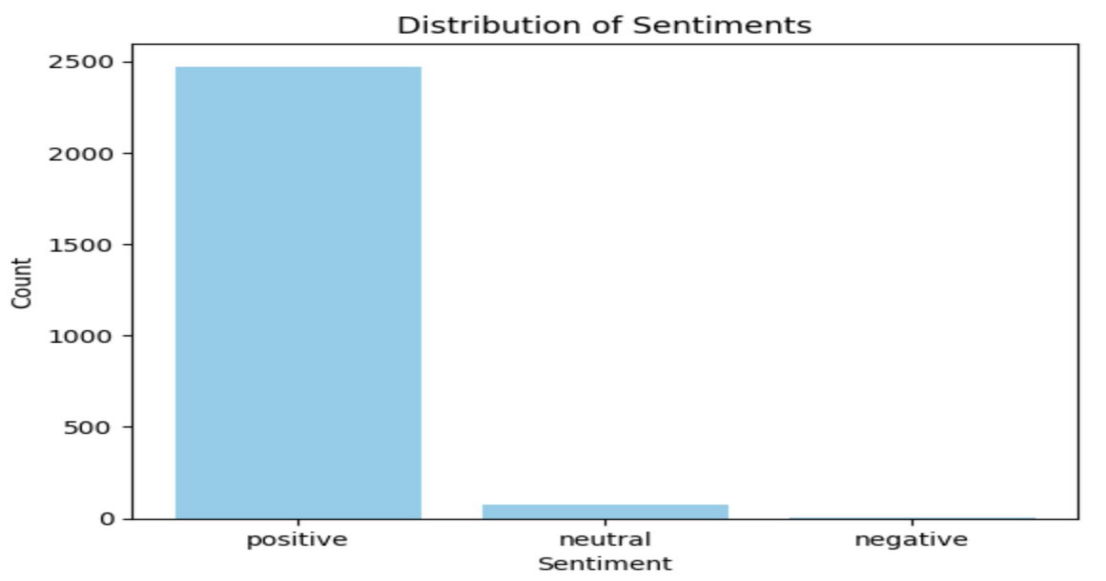
```
positive    2550
Name: sentiment_of_top_ten_words, dtype: int64
```

DISCUSSION OF RESULTS

We obtain different results by changing the number of words that we consider. We can see that when we sort the values and pick the top 3 words associated with the talks, our model is able to distinguish the sentiment in a better way.

For the top 3 words, there are a large number of Ted Talks that have a positive sentiment associated with them. This can be interpreted as follows:

1. The results obtained are biased. This is because of the inherent bias in the data
2. The same 14 words are being used to rate the talks and there is no variability in the sentiment after a certain point.



INSIGHTS

What can be done to avoid the bias in this approach?

1. The main reason for the bias, as we talked about earlier is the fact that the data is inherently biased. So, to avoid this, we can gather more relevant data.
2. We can try to combine the ratings columns with some other columns like comments (in a categorical format) and other user obtained data.

04

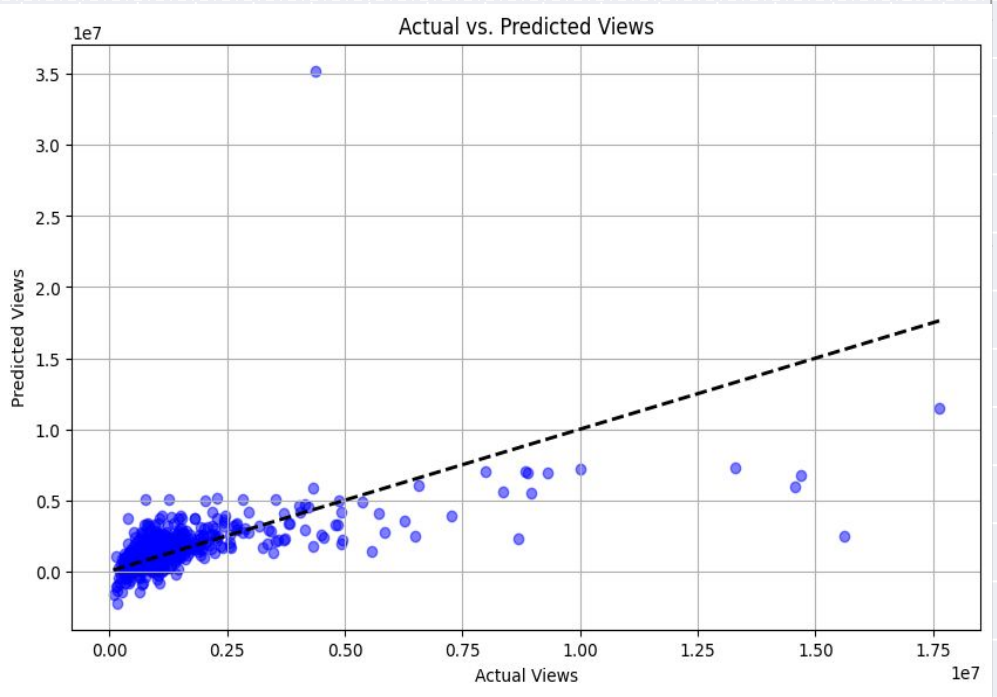
What really makes a Ted Talk Popular?

Linear Regression Models:

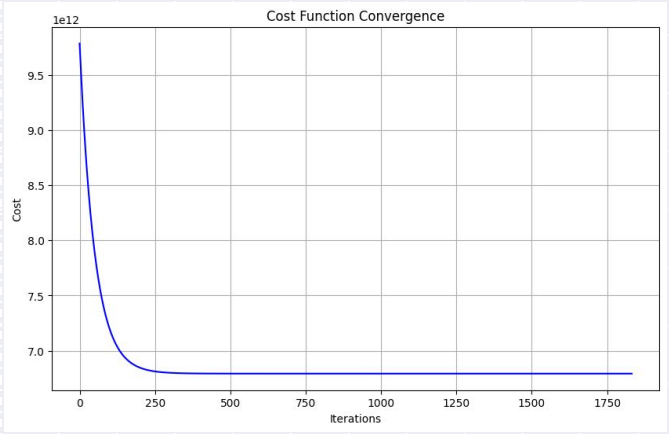
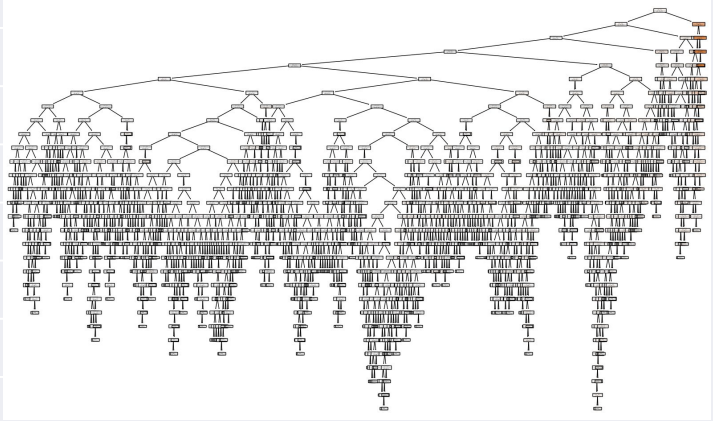
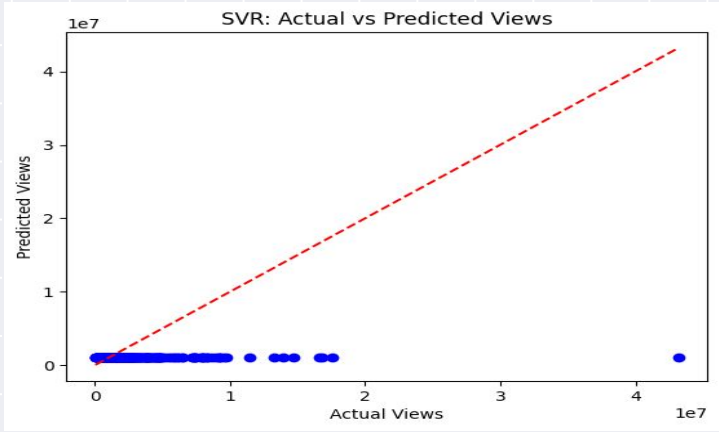
Independent Variable: Views	Model 1	Model 2
Feature Selection	All features	comments ,languages ,duration ,philosopher__spea ,psychology__tags ,power__titl ,expert__spea ,entertainment__tags ,know__titl ,comedian__spea ,might__desc business__tags ,film_date ,secret__titl ,science__titl , politics__tags
R-squared	0.480	0.429
AIC, BIC	6.516e+04, 6.633e+04	6.496e+04, 6.506e+04

INSIGHTS:

Out of the various tags and occupation, it is observed that the keyword "Power" and "Science" in the title gathers more popularity. Along with that, when the ted talk is being delivered by an expert on the matter, then it increases the likeliness of popularity. When the speaker's occupation is that of a comedian, that increases the popularity as well.



Other Failed Models?

Gradient Descent	Random Forest	SVR
R_Squared : -2049.450713972905 MSE : 6869485461510.814 	Mean Squared Error: 2233325324713.815 R-squared: 0.1397046559422412 	Mean Squared Error (Cross-Validation): 7525449822274.568 R-squared: -0.04118489654766044 

Thanks!