TedTalk Data Analysis

Casey O'Malley, Junyi Chen, Echo Liu, Nalorm Tay-Agbozo, Agnes Zhang

Introduction & Project Scope

Client

A research university trying to break into the online learning community

Goal

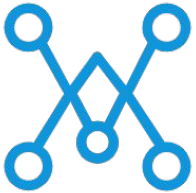
- **Increase** student enrollment rate
- **Enhance** professors recognition
- **Highlight** university's public image
- **Gain status** in higher academic circles

Tool

Ted Talk Dataset

- **Variables** that affect online video views and ratings
- **Transcript data** using NLP analysis
- **Insights** in online educational video industry

Benefit



- **Understand** how different qualities of videos affect views and rating score
- **Predict** video rating and views with exist and new variables from feature engineering and text mining

Database Schema

- **Dataset Description:** all audio-video recordings of TED Talks uploaded to TED.com until Sep 21st, 2017
- **Variables:**

```
 [1] "comments"           "description"        "duration"
 [4] "event"              "film_date"          "languages"
 [7] "main_speaker"       "name"               "num_speaker"
[10] "published_date"     "ratings"            "related_talks"
[13] "speaker_occupation" "tags"               "title"
[16] "url"                "views"
```

- **Additional Details:**
  - **Two datasets:**
    - **Main dataset: 2550** tuples, **17** features; information about all talks
    - **Transcript dataset**: **2464** tuples, **2** attributes; the transcripts for all talks

ETL Process

- Merged data sets
- Split up columns with multiple values
  - Lists
  - Nested jsons
- Created unique identifiers for each table
- Transformed and created new variables
- Ensured all records in each table were unique
- Enforced referential integrities

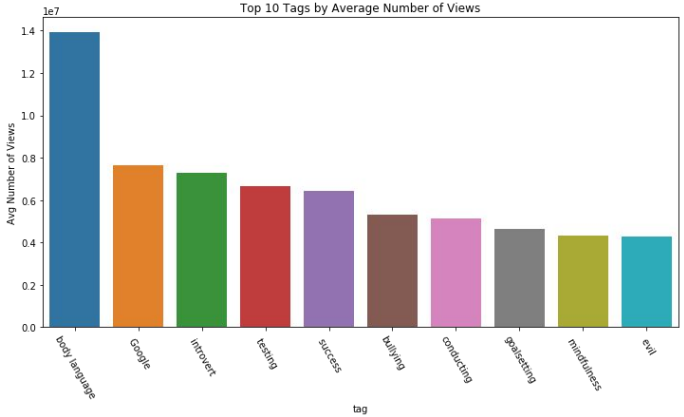
```
main_df.head()
```

	comments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	...	transcript	vide
0	4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	24-02-2006	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	26-06-2006	...	Good morning. How are you? (Laughter)It's been ...	0
0	4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	24-02-2006	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	26-06-2006	...	Good morning. How are you? (Laughter)It's been ...	0
0	4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	24-02-2006	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	26-06-2006	...	Good morning. How are you? (Laughter)It's been ...	0
0	4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	24-02-2006	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	26-06-2006	...	Good morning. How are you? (Laughter)It's been ...	0

Analytical Procedures: Data Exploration

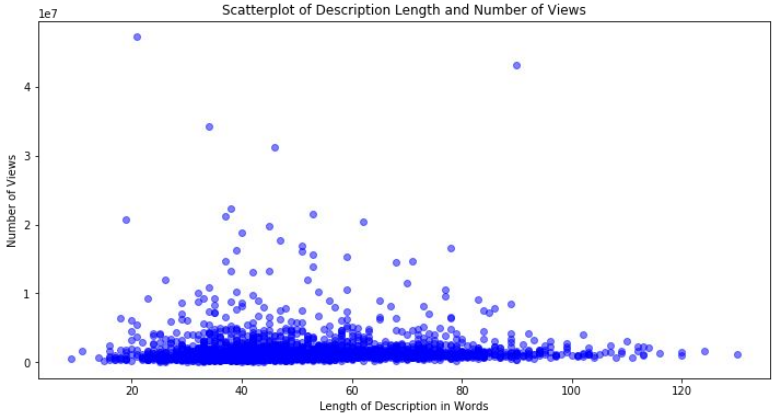
Data Exploration:

- Focused on outcome variable (views)
- Analyzed relationships with predictor variables through Matplotlib & Seaborn



Feature Engineering:

- Title Length and Description Length (both measured in characters and words)
- Data transformation (next slide)

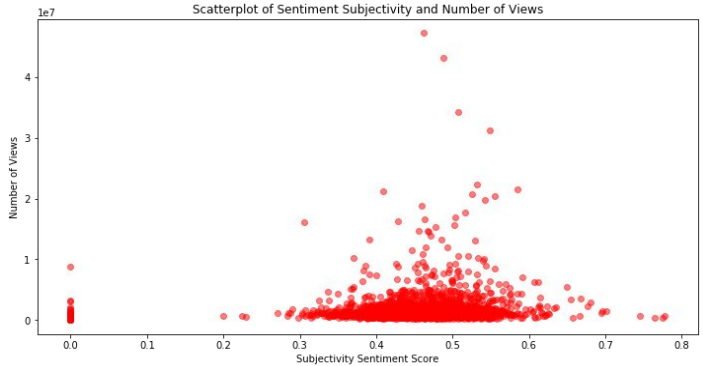


Analytical Procedures: Text Mining

Transcripts for each TED Talk were cleaned using a defined process:

- Remove punctuation
- Remove stop words
- Make characters lower-cased
- Lemmatize remaining words
- Store in list format

Sentiment Analysis:




```
text_data[['title', 'transcript', 'clean_transcript']].head()
```

	title	transcript	clean_transcript
0	Do schools kill creativity?	Good morning. How are you?(Laughter)It's been ...	[good, morning, how, youlaughterits, great, ha...
1	Averting the climate crisis	Thank you so much, Chris. And it's truly a gre...	[thank, much, chris, and, truly, great, honor,...
2	Simplicity sells	(Music: "The Sound of Silence," Simon & Garfun...	[music, the, sound, silence, simon, garfunkelh...
3	Greening the ghetto	If you're here today — and I'm very happy that...	[if, youre, today, im, happy, youve, heard, su...
4	The best stats you've ever seen	About 10 years ago, I took on the task to teac...	[about, 10, year, ago, i, took, task, teach, g...

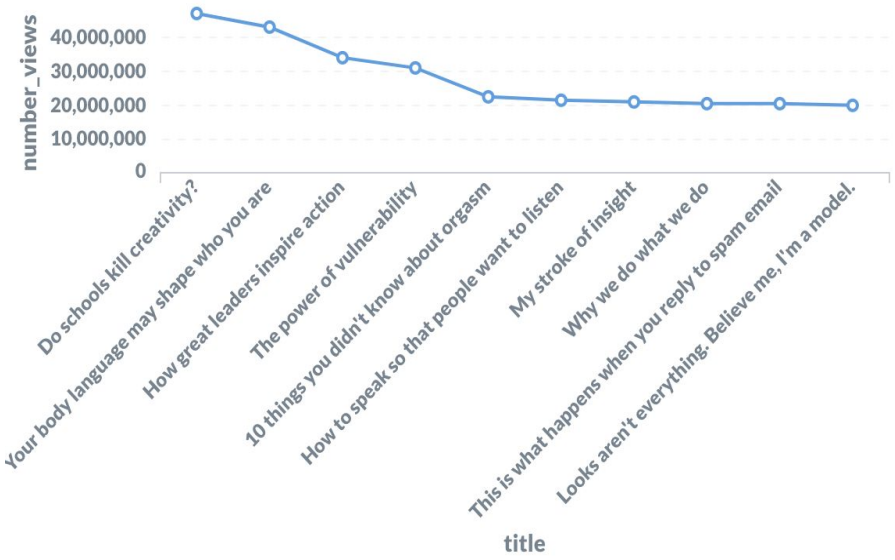
Count Vectorization with Scikit Learn:

	0	01	05	1	10	100	1000	10000	100000	...	zerosum	zimbabwe	zip	zombie	zone	zoning	zoo	zoom	zooming	zurich
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	2	...	0	0	1	0	0	2	0	0	0	0
4	0	0	0	0	0	4	3	1	0	...	0	0	0	0	0	0	0	0	0	0

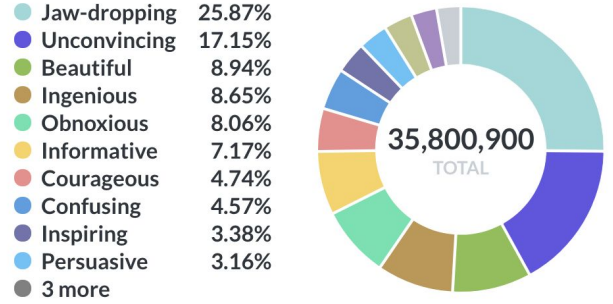
5 rows × 11763 columns

Metabase Dashboard: C Level

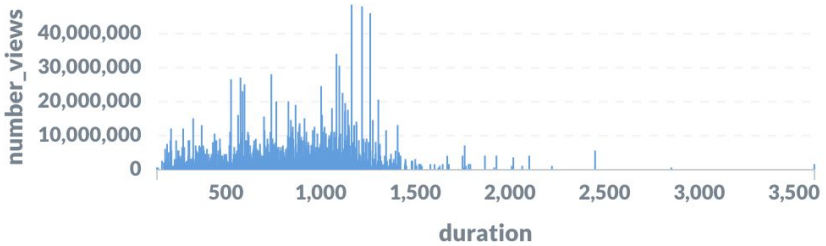
Titles of the top 10 most viewed videos



Top 10 ratings affectimg number of views



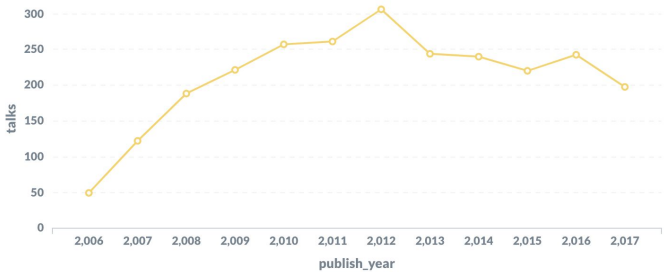
Does duration of a video affect the avg number of views?



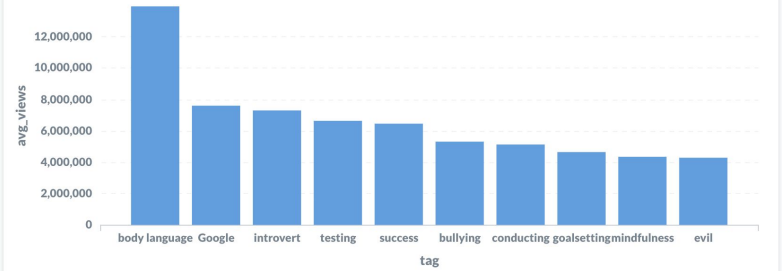
Metabase Dashboard: Analysts

2,012
Year with most published videos

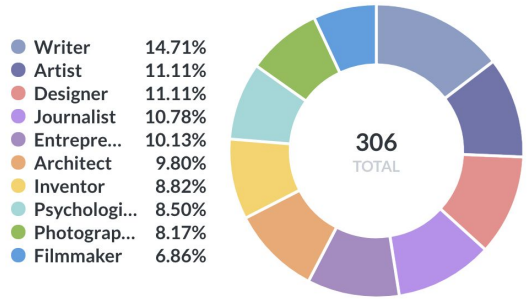
How is the number of published talks going through the years?



what is the top 10 average number of views by tags?



What is the weight of top 10 poplar occupations in Ted-Talk?



Conclusion & Key Takeaways

- PostgreSQL effective way to store this database
    - ETL process important for selected dataset
    - If not in 3NF, data would likely fit best in a NoSQL system such as MongoDB
- Python makes it easy to pull this data from our database and manipulate for new tasks
- Metabase was an effective tool for dealing with our database at all stakeholder levels
    - Analysts can use SQL directly in the system
    - Executives can view dashboards with only the most critical information
- Additional feature engineering and data exploration must be done to better understand the factors that affect video view counts