596: Machine Learning
Term Project

Nrithya, Liz, Xiqiao, Akash
and Rohith

TED Data Analysis

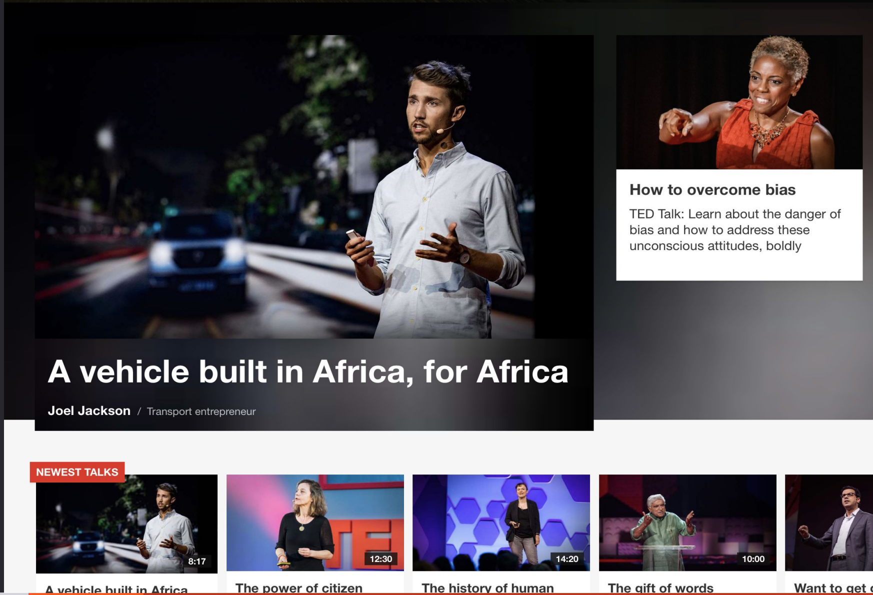
Dec 20, 2017

What is TED?

“A nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less).”

Example TED Talks:

- “Do schools kill creativity?”
- “Extreme swimming with the world’s most dangerous jellyfish”
- “The enchanting music of sign language”



How to overcome bias

TED Talk: Learn about the danger of bias and how to address these unconscious attitudes, boldly

A vehicle built in Africa, for Africa

Joel Jackson / Transport entrepreneur

NEWEST TALKS

8:17 A vehicle built in Africa

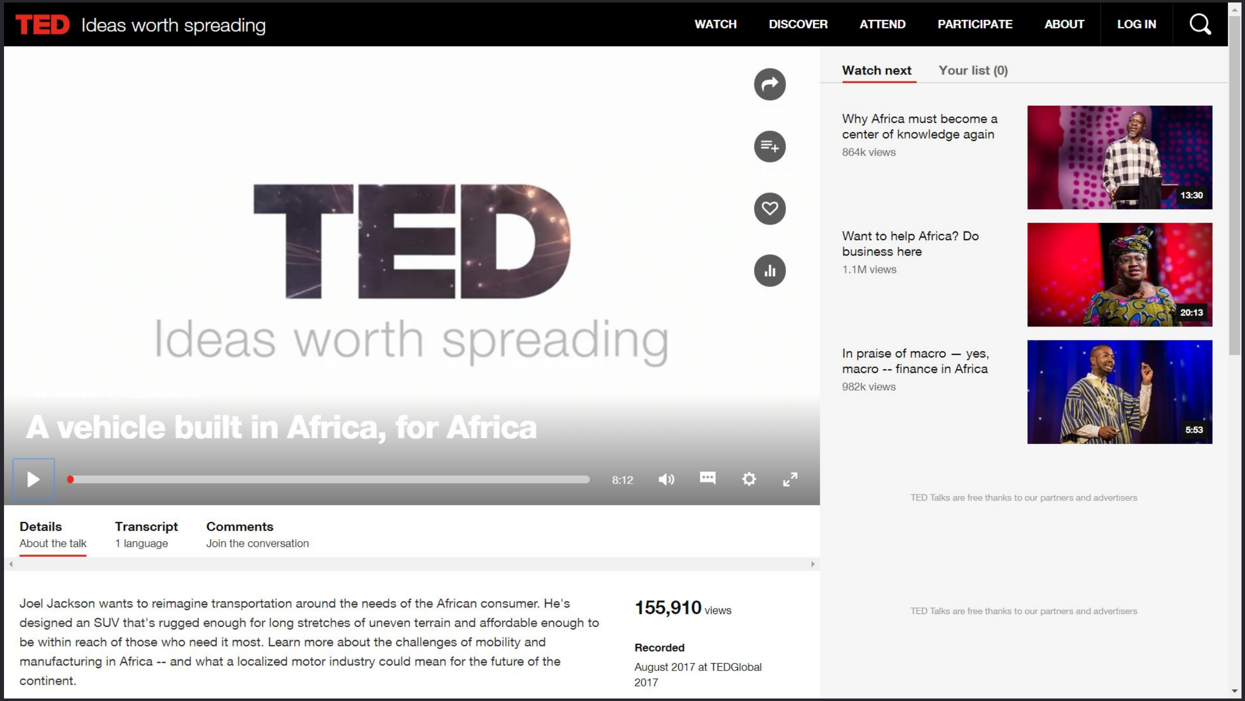
12:30 The power of citizen

14:20 The history of human

10:00 The gift of words

Want to get

Descriptive Findings



TED Ideas worth spreading
WATCH
DISCOVER
ATTEND
PARTICIPATE
ABOUT
LOG IN
TED
Ideas worth spreading
8:12
Watch next
Your list (0)
13:30
20:13
5:53
TED Talks are free thanks to our partners and advertisers
TED Talks are free thanks to our partners and advertisers

A vehicle built in Africa, for Africa

Why Africa must become a center of knowledge again
864k views

Want to help Africa? Do business here
1.1M views

In praise of macro — yes, macro -- finance in Africa
982k views

Details
About the talk

Transcript
1 language

Comments
Join the conversation

Joel Jackson wants to reimagine transportation around the needs of the African consumer. He's designed an SUV that's rugged enough for long stretches of uneven terrain and affordable enough to be within reach of those who need it most. Learn more about the challenges of mobility and manufacturing in Africa -- and what a localized motor industry could mean for the future of the continent.

155,910 views

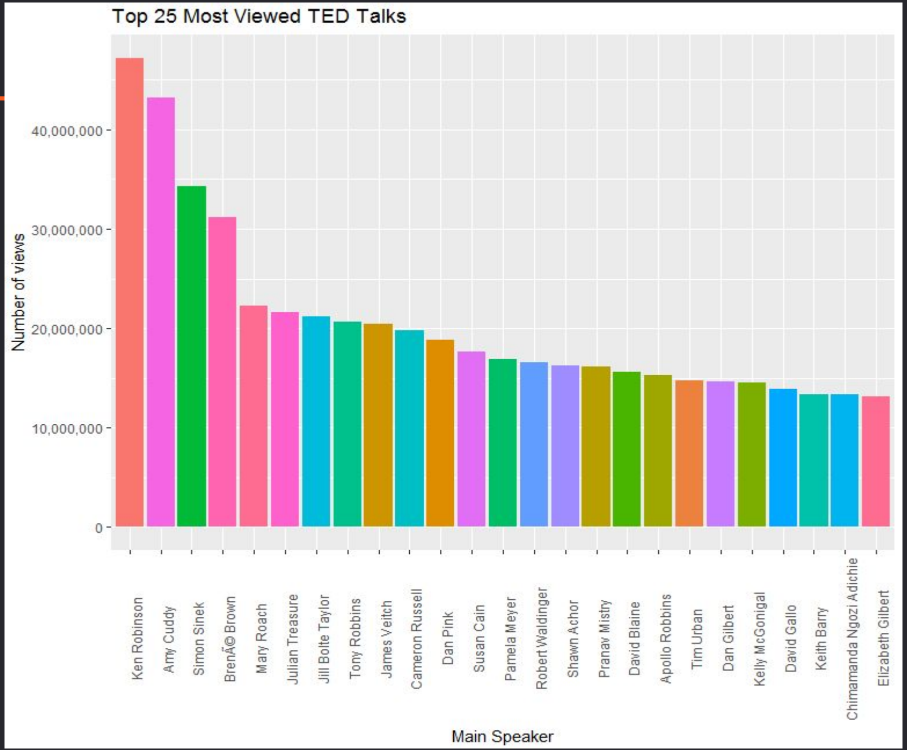
Recorded
August 2017 at TEDGlobal 2017

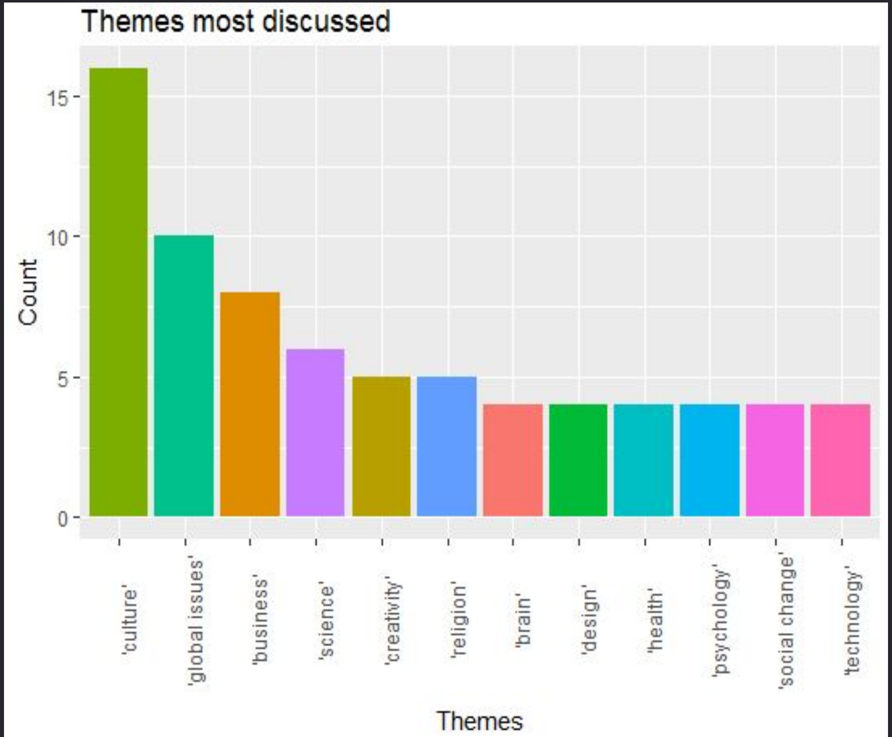
- Looking at TED.com the videos are ordered by newest released first
  - 2017 - 2006

- Transcripts, Details, Favoriting, and Rating Features
  - Controlled List

- Commenting and Discussions are encouraged and are Monitored for Spam

Initial Descriptive Findings





- **“Do schools kill creativity?”**
- **“How to escape education's death valley”**

- Culture, Global Issues, Business, Science

Business questions

1. Can we predict views and comments for new videos?
2. Can we categorize videos by popularity?
3. What kind of impact did a video have on viewers?
4. Can we cluster videos based on tags?

Data Processing and Curation

Aggregate Data

- Unix and Excel to separate columns
  - REGEX, grep, and pivot tables
- R to clean/curate remaining data
  - Incomplete data, unix timestamps,
    string manipulation

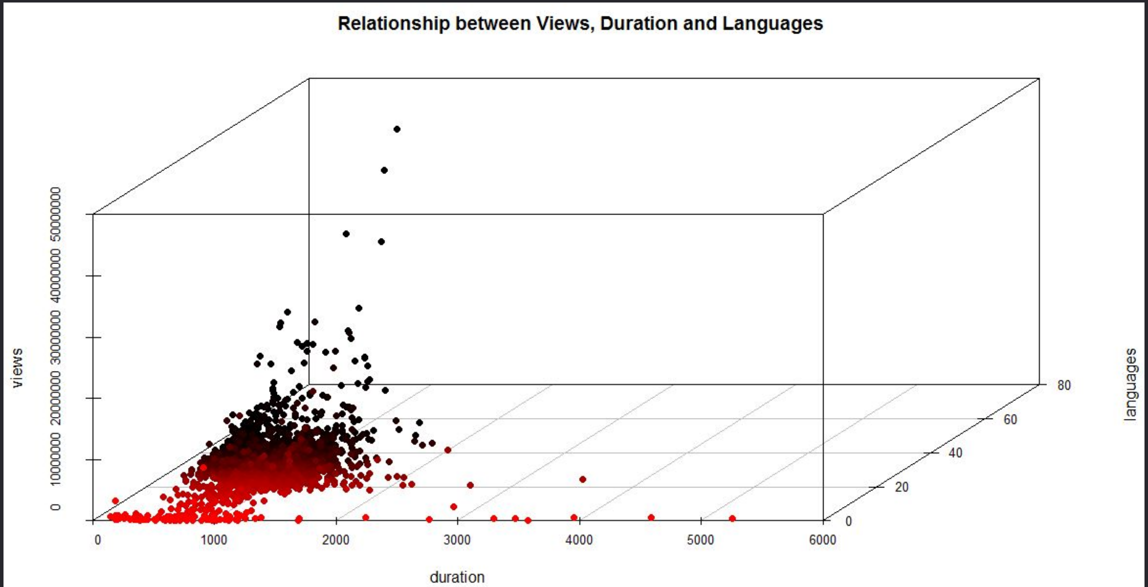
Prioritize Metrics

- Ratings
- Tags/Themes
- Number of views
- Number of comments
- Duration
- Number of Languages

QUESTION 1
PREDICTING
VIEWS & COMMENTS

Using linear regression and artificial neural networks to predict the number of views and the number of comments a video will receive.

Linear Regression - Relationship



Videos clustered between

0 - 10,000,000 views

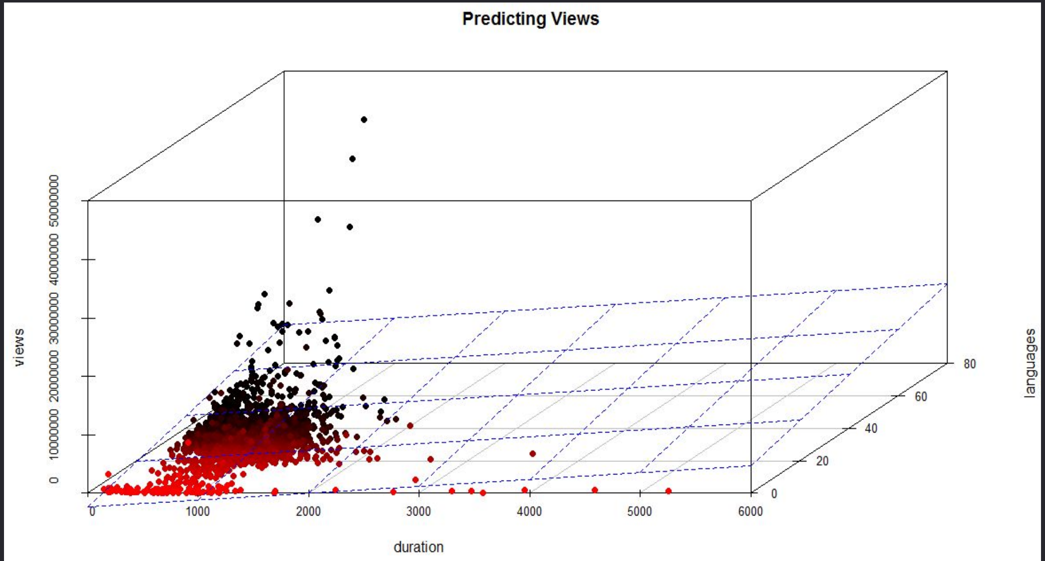
0 - 1,000 seconds

0 - 40 languages

Linear Regression - Predicting Views

Views = 11741.1(duration) + 112232.3(languages) - 2339034.6

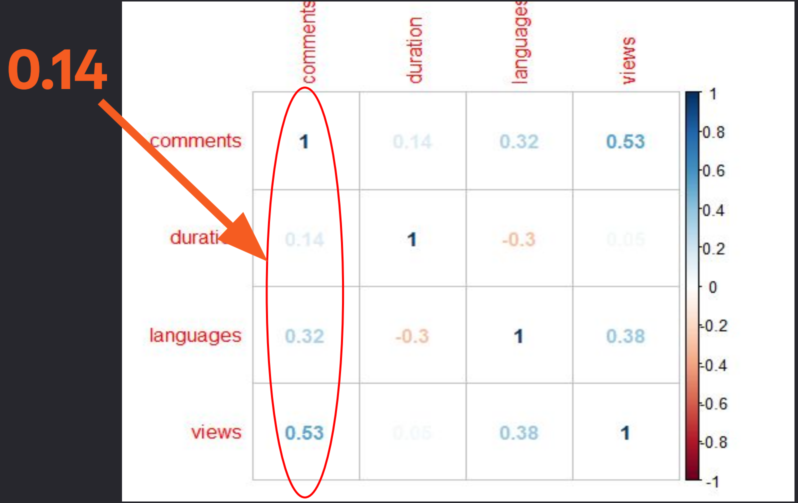
Correlation	Duration	Language
Views	0.38	.005



Linear Regression - Predicting comments

Comments = -.01693 + 0.1351(duration) - 0.0005035(views) + 5.991(languages)

Correlation	Views	Duration	Language
Comments	0.53	0.32	.014

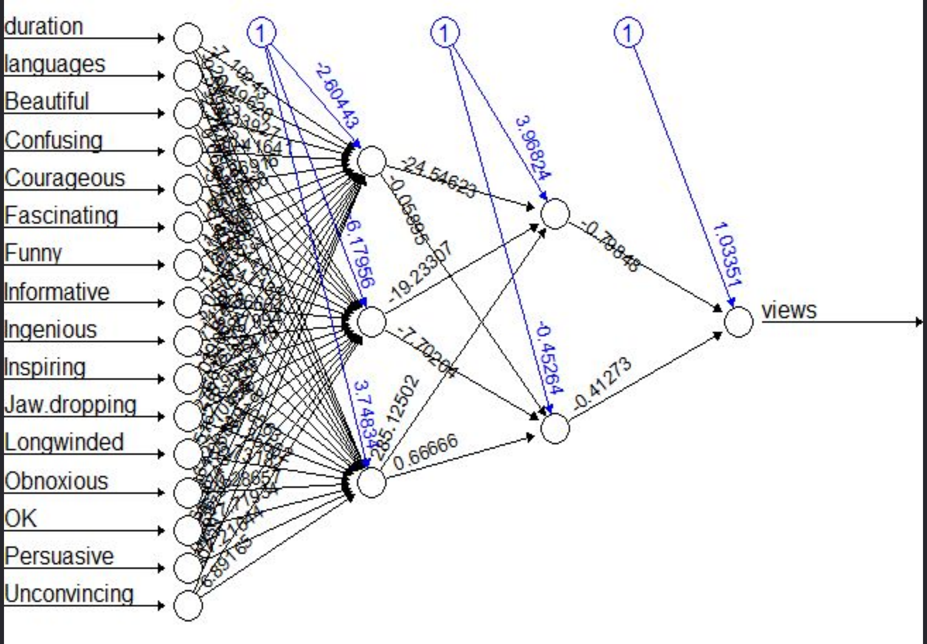


0.14

	comments	duration	languages	views
comments	1	0.14	0.32	0.53
duration	0.14	1	-0.3	0.05
languages	0.32	-0.3	1	0.38
views	0.53	0.05	0.38	1

Artificial Neural Networks - Views

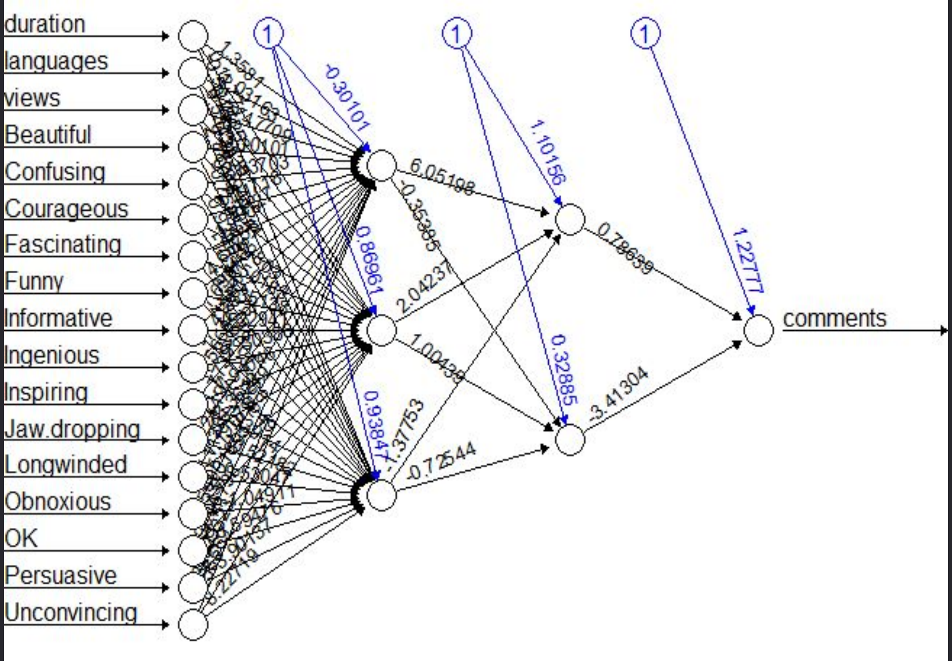
Model : Views ~ Durations + Languages + Ratings



MSE: 6283596155909

Artificial Neural Networks - Comments

Model : Comments ~ Duration + Languages + Ratings



MSE: 4068261144209

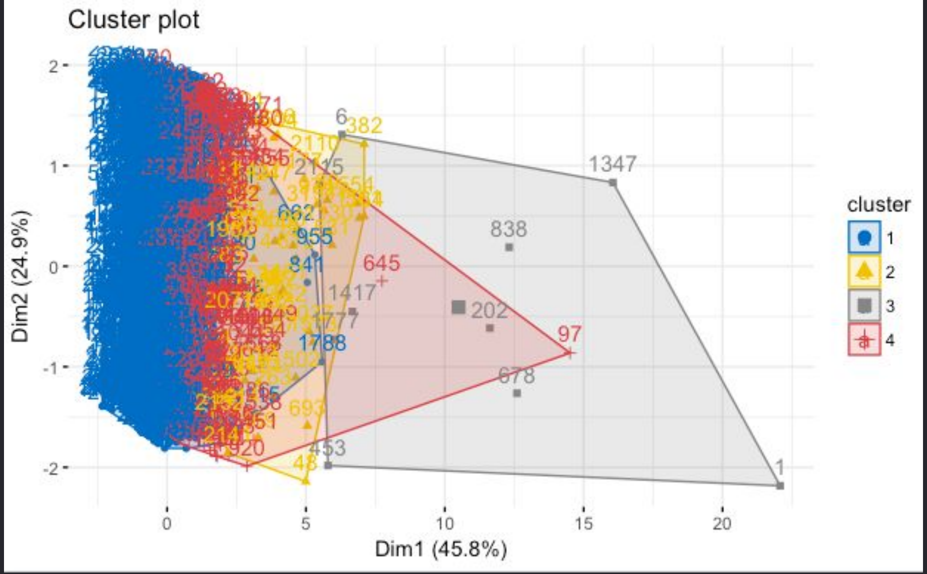
QUESTION 2
VIDEO POPULARITY GROUPS

Unsupervised clustering using K-means to group videos according to the popularity, determining with the number of its views, comments, and languages available.

K-means clustering : Clusters

Model : Popularity ~ Views + comments + languages

- 75 : 25 data split
- Four clusters of sizes:
  - 318, 10, 2167, 55
- Between_SS / Total_SS = 84.7 %



K-means: Findings

- The majority of the spots are highly concentrated on the left side of the plot, which indicates that most of the videos share similar popularity according to the number of its views, comments, and languages available.

- On the other hand, a few spots are spread apart in the center and on the right of the plot, which demonstrates there are small number of videos that have different popularity compared to the majority.

QUESTION 3
UNDERSTANDING IMPACT

Using KNN & Random Forest to classify videos based on views, comments, and ratings

Understanding Ratings

RATINGS	METRIC	IMPACT
Beautiful, Fascinating, Jaw-dropping, Inspiring, Ingenious	4	High Impact
Informative, Courageous, Persuasive, Funny	3	Good Impact
OK, Long-winded	2	Neutral
Confusing, Obnoxious, Unconvincing	1	Bad Impact

- 14 Rating options
- Viewer can choose one, more than one, or all
- 4-level metric to understand ratings.

KNN

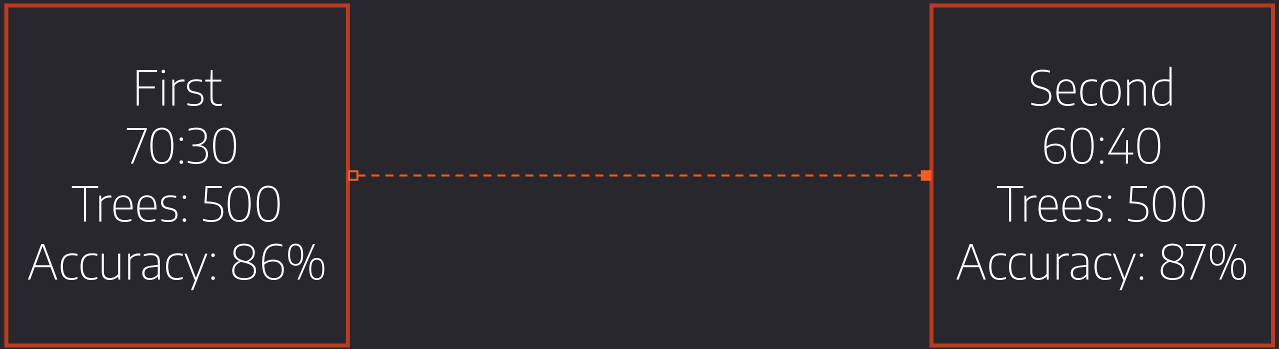
Model : Metric ~ (Ratings + View + Comments + Duration + Language)

First
75:25
Neighbors- 3
Accuracy- 50%

Second
75:25
Neighbors- 5
Accuracy- 55%

Random Forest

Model : Metric ~ (Ratings + View + Comments + Duration + Language)



First
70:30
Trees: 500
Accuracy: 86%

Second
60:40
Trees: 500
Accuracy: 87%

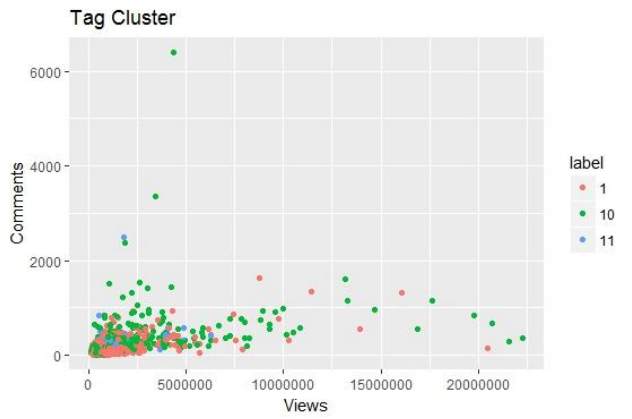
QUESTION 4
CLUSTERING FOR RECOMMENDATION

Using Agglomerative & Divisive Clustering to classify videos based on tags for better recommendations

Understanding Tags

- 147 unique tags
- Approximately 10 to 20 tags on each video
- Recommend videos with viewer’s choice of tags.
- 2 of the top most tags-“technology” and “culture”

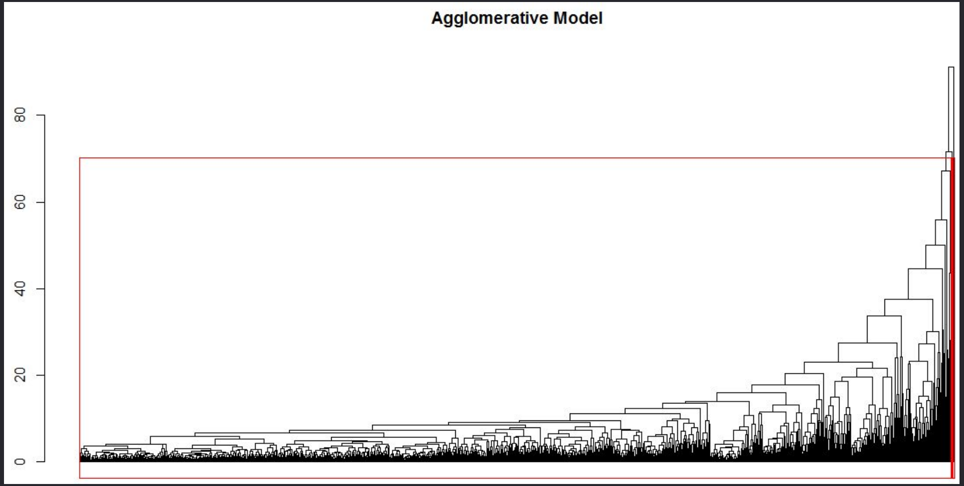
TAGS	LABEL
Technology	01
Culture	10
Technology, Culture	11



Agglomerative

Model : Tag Labels ~ Views+Comments+Duration+Languages+Ratings

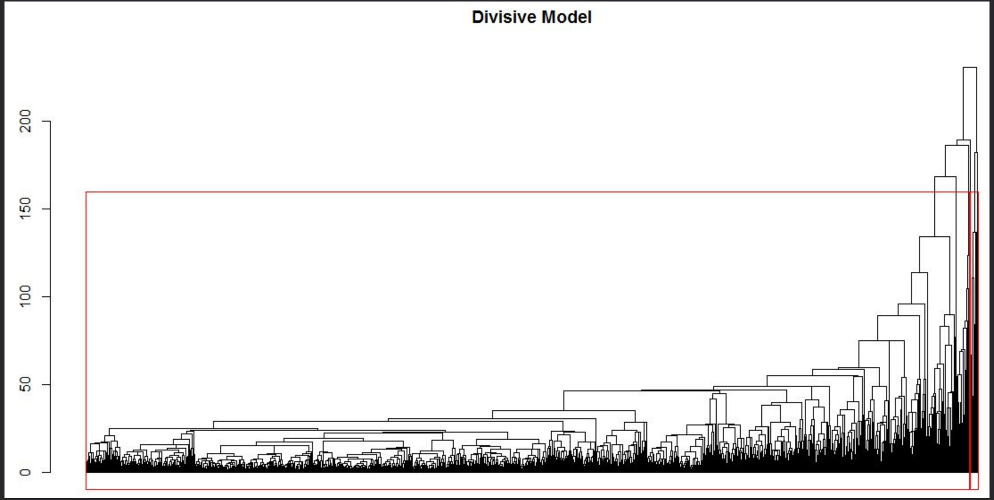
No of clusters : 3 Height : 70 Accuracy : 56%



Divisive

Model : Tag Labels ~ Views+Comments+Duration+Languages+Ratings

No of clusters : 3 Height : 160 Accuracy : 57%



Limitations

- Limited analysis scope due to finite numerical variables
  - Lot of categorical variables

- Time span of videos - Comparing views of 2006 to 2017
  - Exposure on videos
  - Number of Tags have changed after 2013

- Data restricted to 2,550 unique entries with skewed clusters

Summary

- We found that the data was skewed and that was difficult to manage due to different exposure on videos and topics
- We used the ‘tags’, ‘views’, ’topic’ to forecast the popularity of the talks and increase viewership
- Having more dimensional aspects like Rating out of 5, Likes, etc. can help analysts create better models to predict the viewership

THANKS!

ANY QUESTIONS?