Capstone Project - 2

Ted Talks Views Prediction

Team Member

Harish Kollana

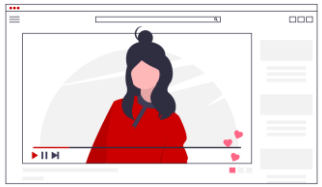
Discussion Points

1. Problem Statement
2. Data Summary
3. Explorative Data Analysis
4. Feature Engineering
5. Data Cleaning
6. Feature Selection
7. Model Fitting
8. Feature Importance
9. Model Comparison
10. Challenges
11. Conclusion

The Dilemma

How Ted Talks Works

Influential Speakers gives speeches on Ted talks



Ted Talks Publish Those videos in their platform

The users Will Watch the videos

Ted Talks: ideas worth spreading

TED Conferences LLC is an American media organization that posts talks online for free distribution under the slogan "ideas worth spreading".

Problem Statement

TED is devoted to spreading powerful ideas on just about any topic. These datasets contain over 4,000 TED talks including transcripts in many languages. Founded in 1984 by Richard Salman as a nonprofit organization that aimed at bringing experts from the fields of Technology, Entertainment, and Design together, TED Conferences have gone on to become the Mecca of ideas from virtually all walks of life. As of 2015, TED and its sister TEDx chapters have published more than 2000 talks for free consumption by the masses and its speaker list boasts of the likes of Al Gore, Jimmy Wales, Shahrukh Khan, and Bill Gates. The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website.

Data Summary

Data Set Name : data_ted_talks

Data Set Information:
Number of instances: 4,005
Number of attributes: 19

Features:
'talk_id', 'title', 'speaker_1', 'all_speakers', 'occupations', 'about_speakers', 'views', 'recorded_date', 'published_date', 'event', 'native_lang', 'available_lang', 'comments', 'duration', 'topics', 'related_talks', 'url', 'description', 'transcript'

Data Summary

talk_id: Talk identification number provided by TED

title: Title of the talk

speaker_1: First speaker in TED's speaker list

all_speakers: Speakers in the talk

occupations: Occupations of the speakers

about_speakers: Blurb about each speaker

recorded_date: Date the talk was recorded

published_date: Date the talk was published to TED.com

event: Event or medium in which the talk was given

Data Summary

native_lang: Language the talk was given in

available_lang: All available languages (lang_code) for a talk

comments: Count of comments

duration: Duration in seconds

topics: Related tags or topics for the talk

related_talks: Related talks (key='talk_id',value='title')

url: URL of the talk

description: Description of the talk

transcript: Full transcript of the talk

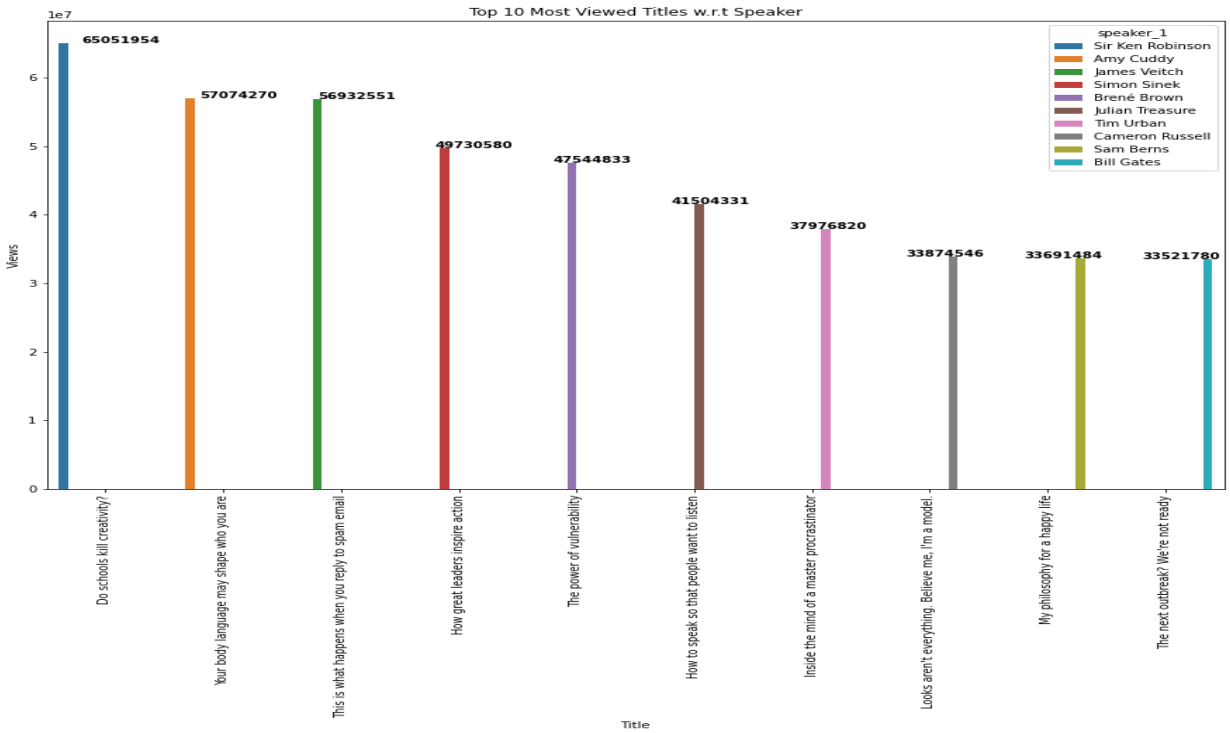
EDA

Most Popular Words In Titles

EDA (continued)

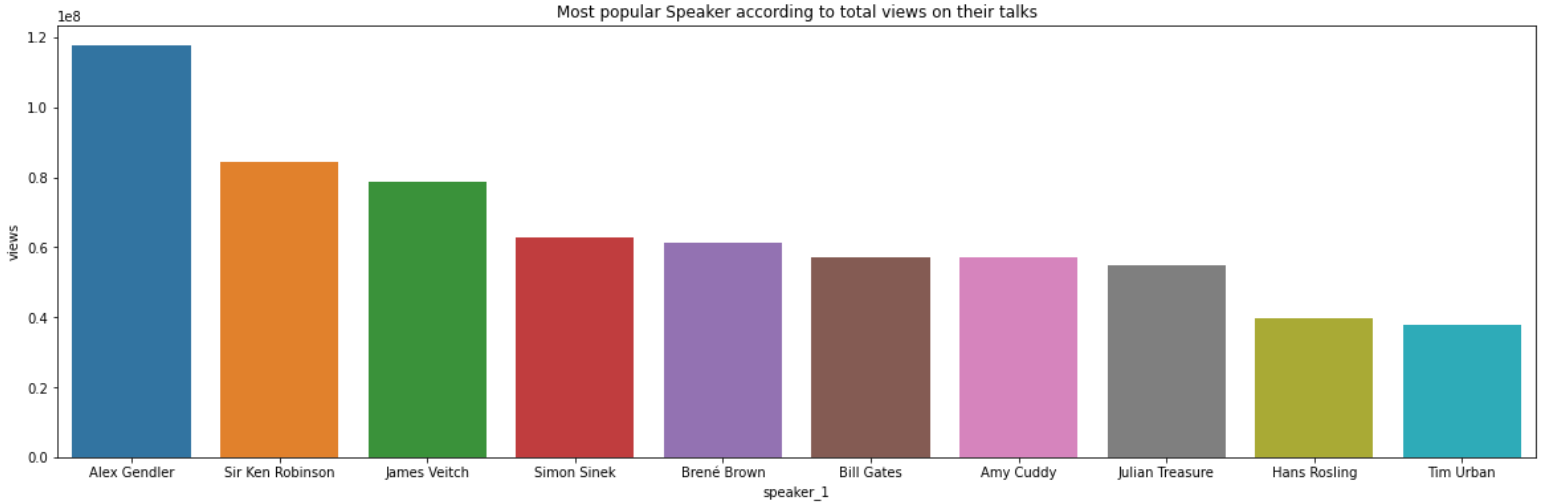
Top 10 Most Viewed Titles w.r.t speakers

Do Schools kill creativity is most popular Title having more than 65 Million views by speaker “Sir Ken Robinson”



EDA (continued)

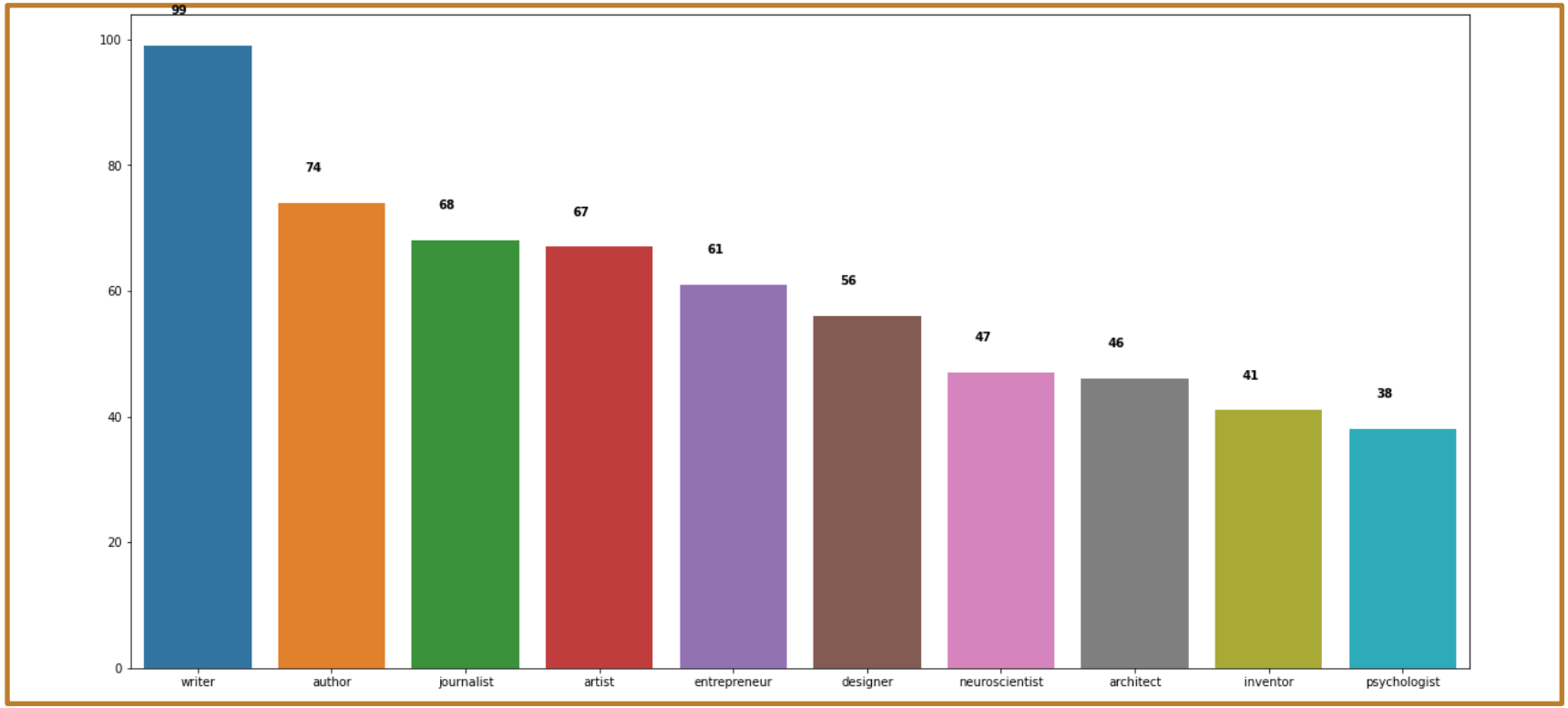
Most popular Speaker according to total views on their talks



Alex Gendler is most popular speaker according to the total views by speakers

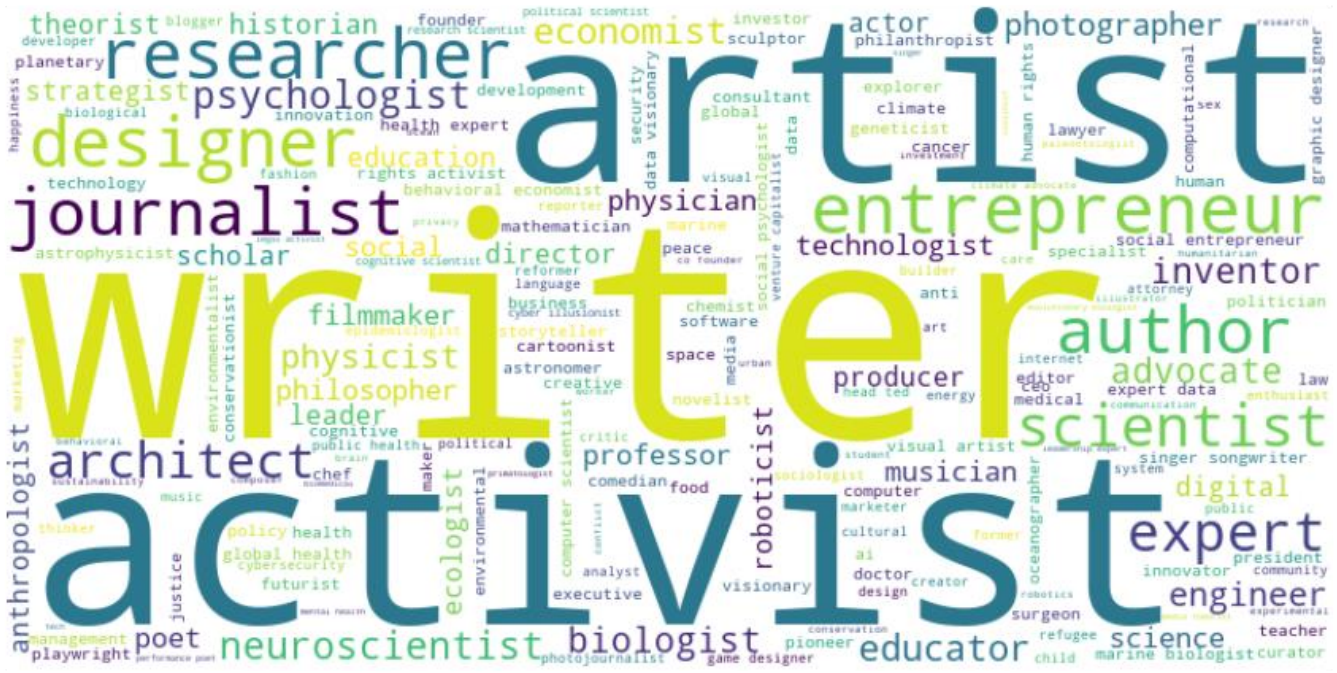
EDA (continued)

Top 10 occupations of speakers



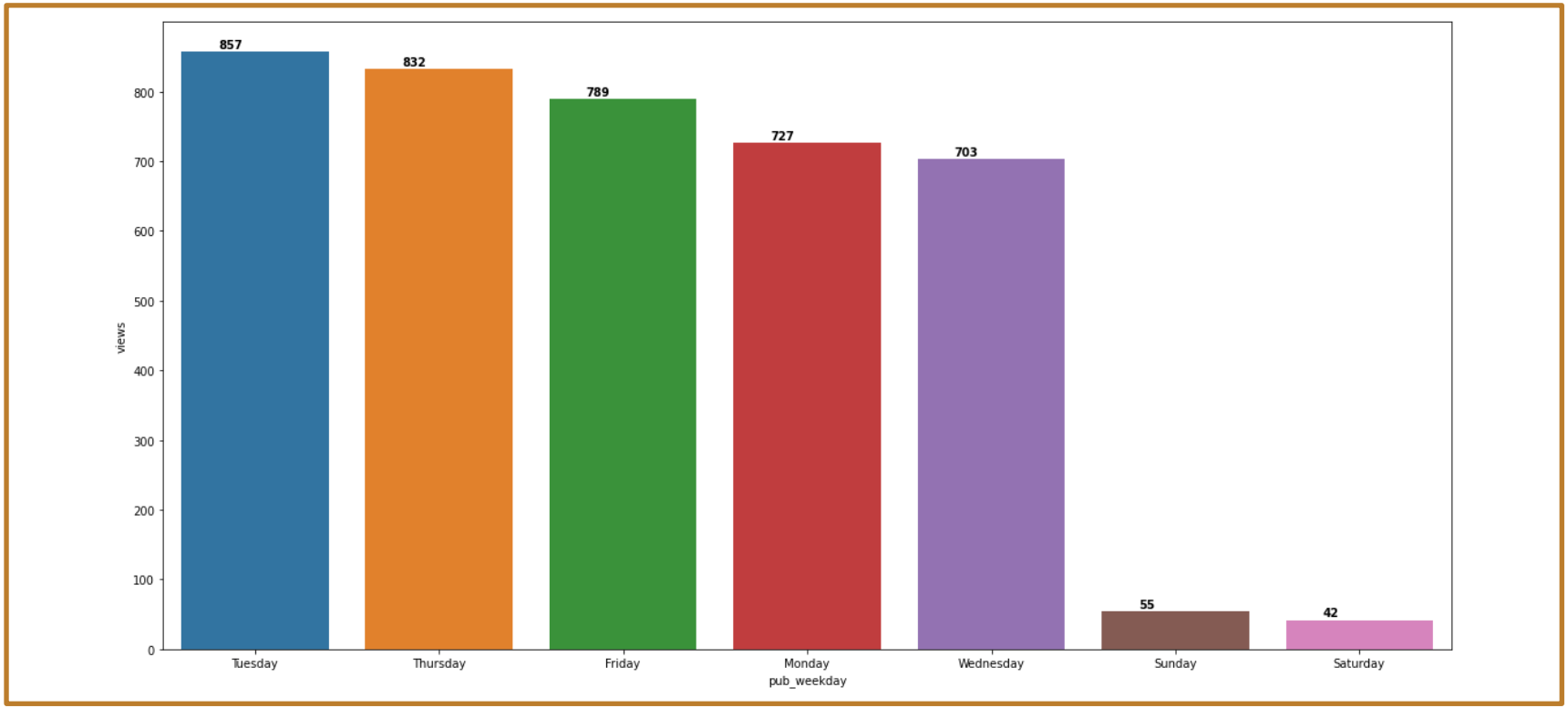
EDA (continued)

Top 10 occupations of speakers



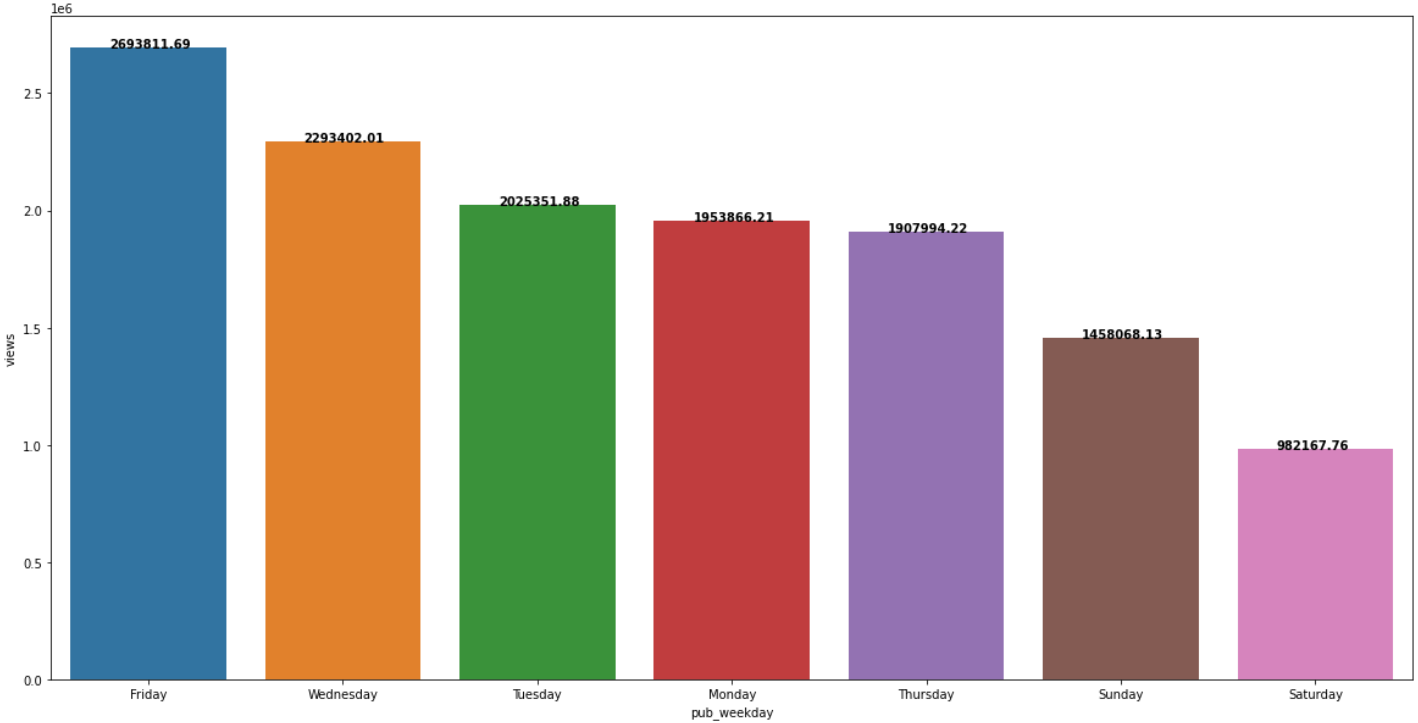
EDA (continued)

weekday having maximum releases



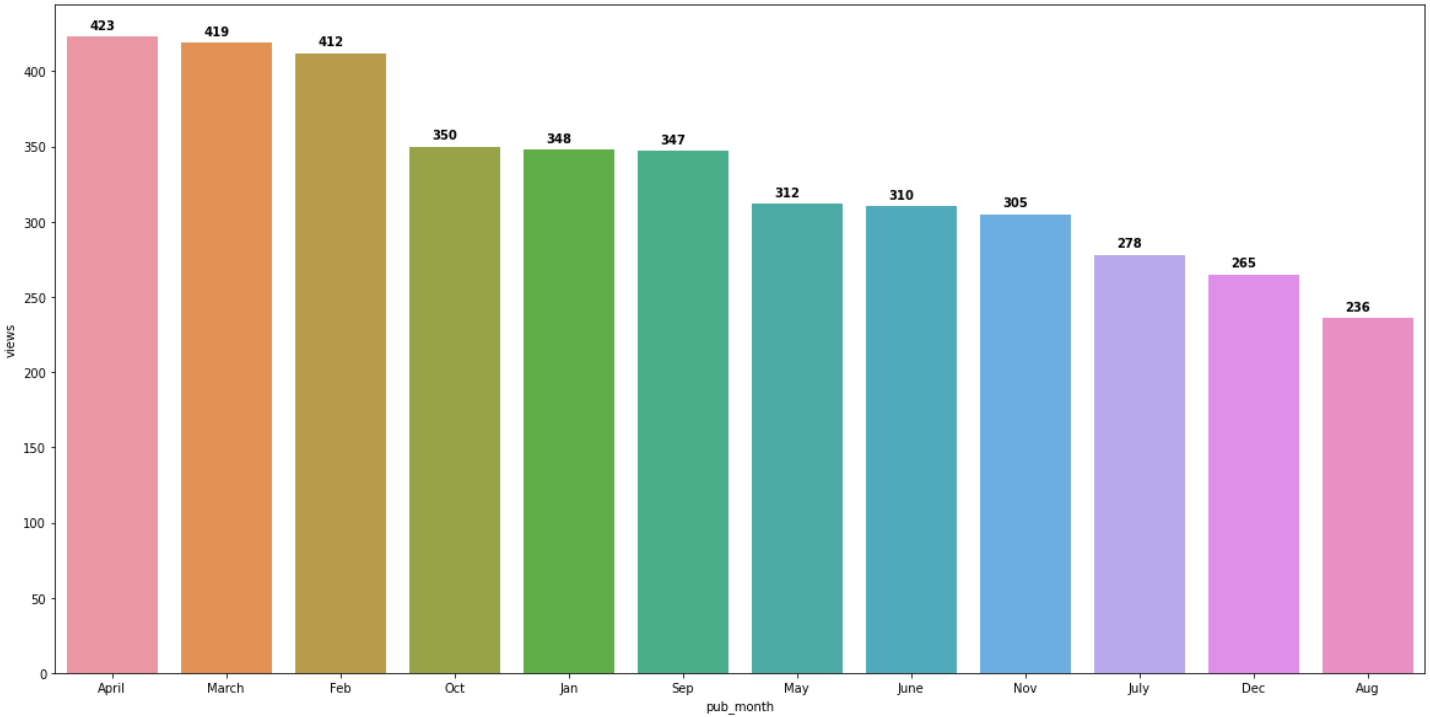
EDA (continued)

Pub_weekday w.r.t Average Views



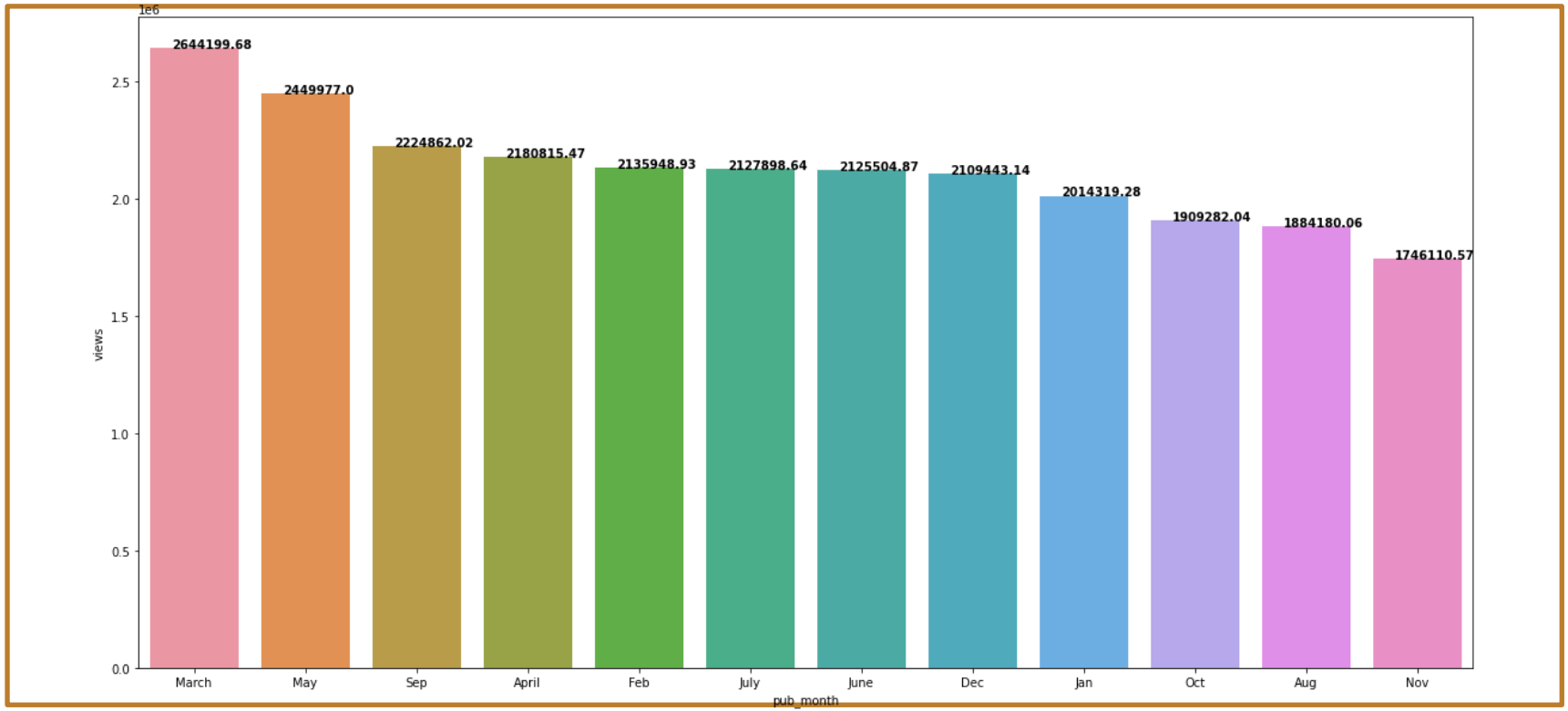
EDA (continued)

Month having maximu releases



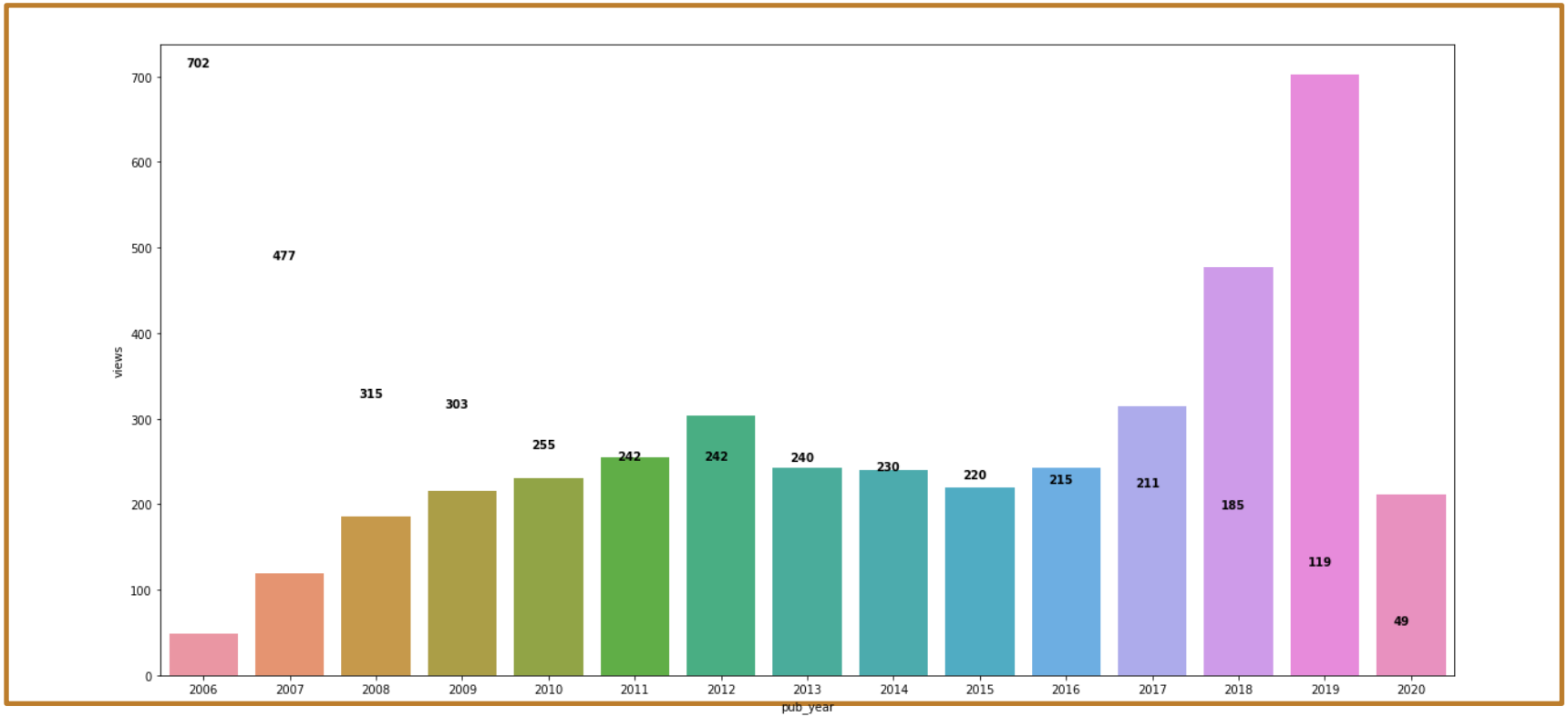
EDA (continued)

pub_month w.r.t to average views



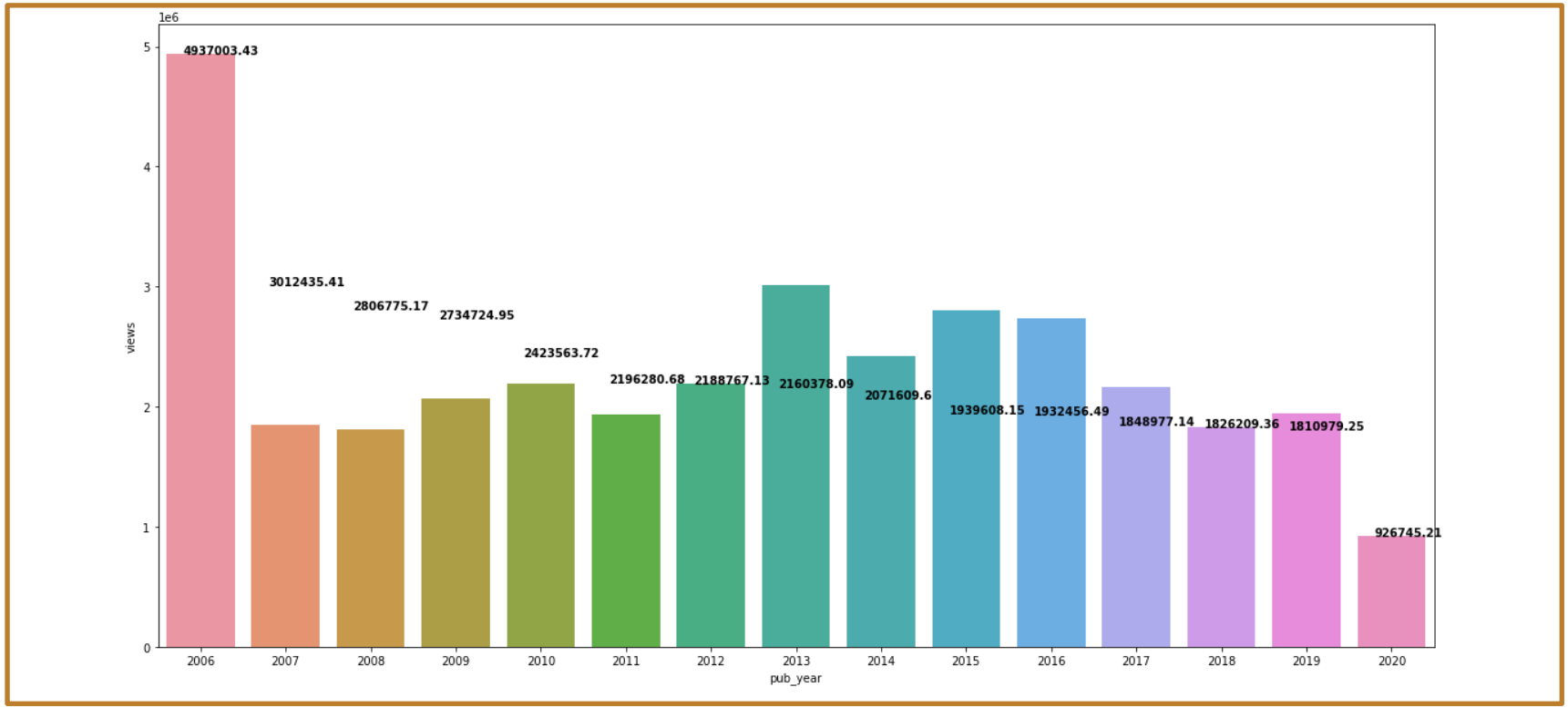
EDA (continued)

year having maximum releases



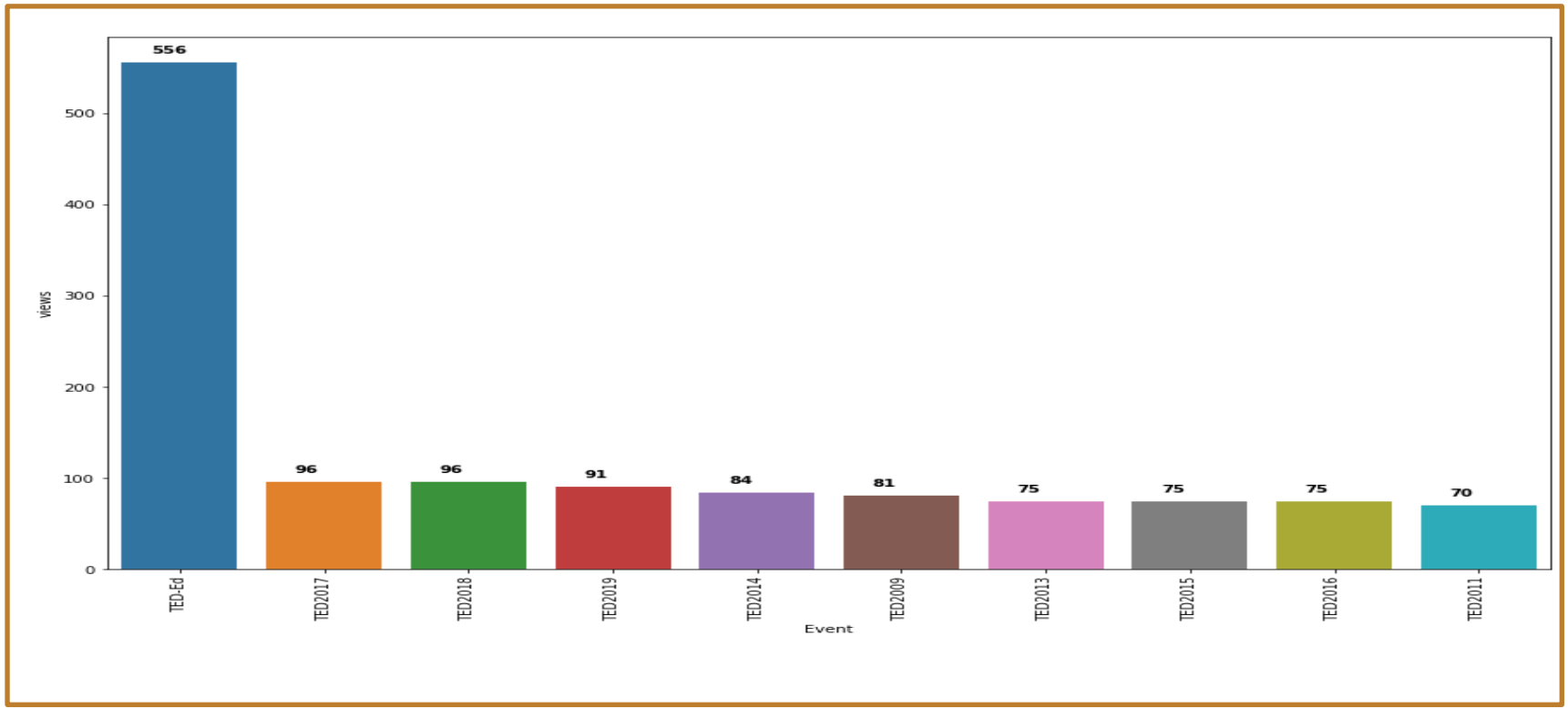
EDA (continued)

year avg w.r.t views



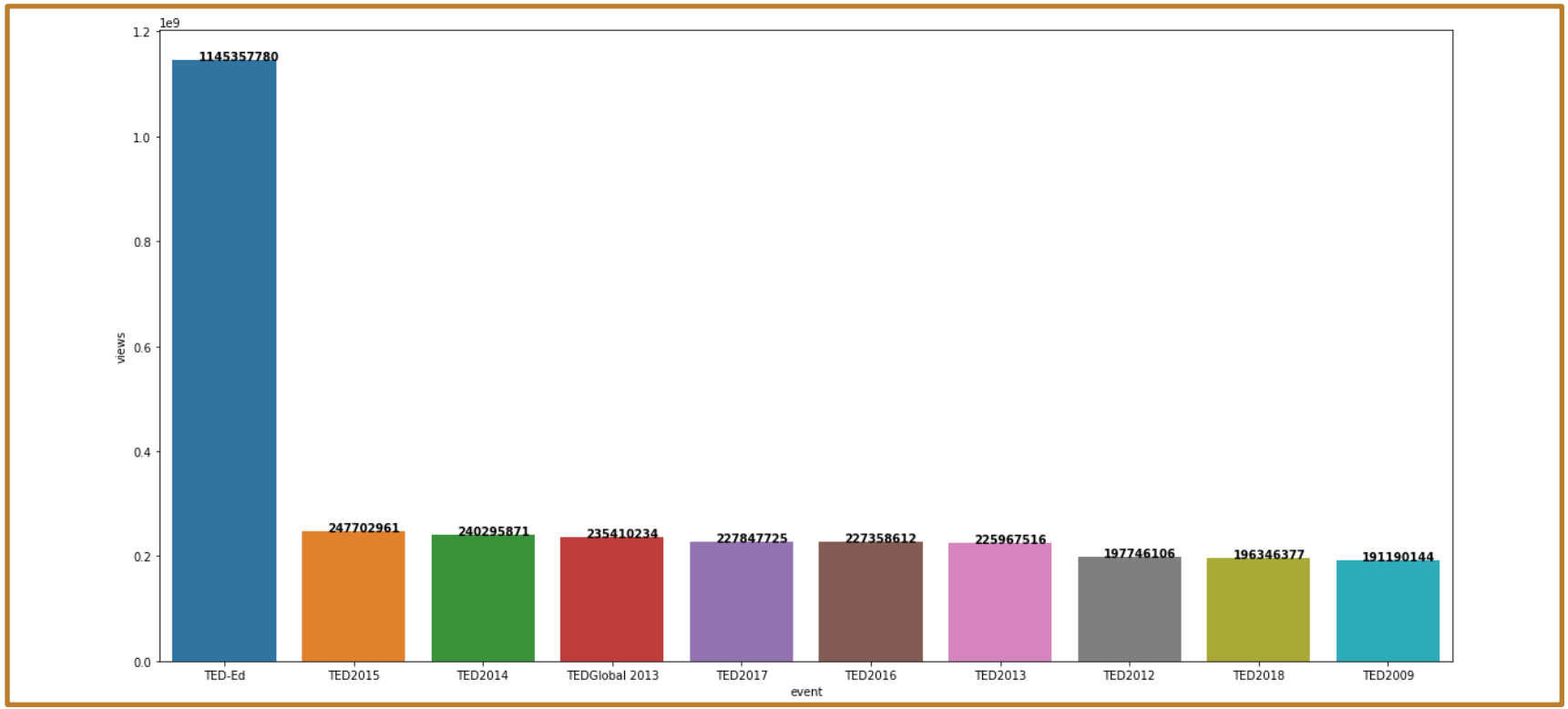
EDA (continued)

Top 10 Events with most no of views



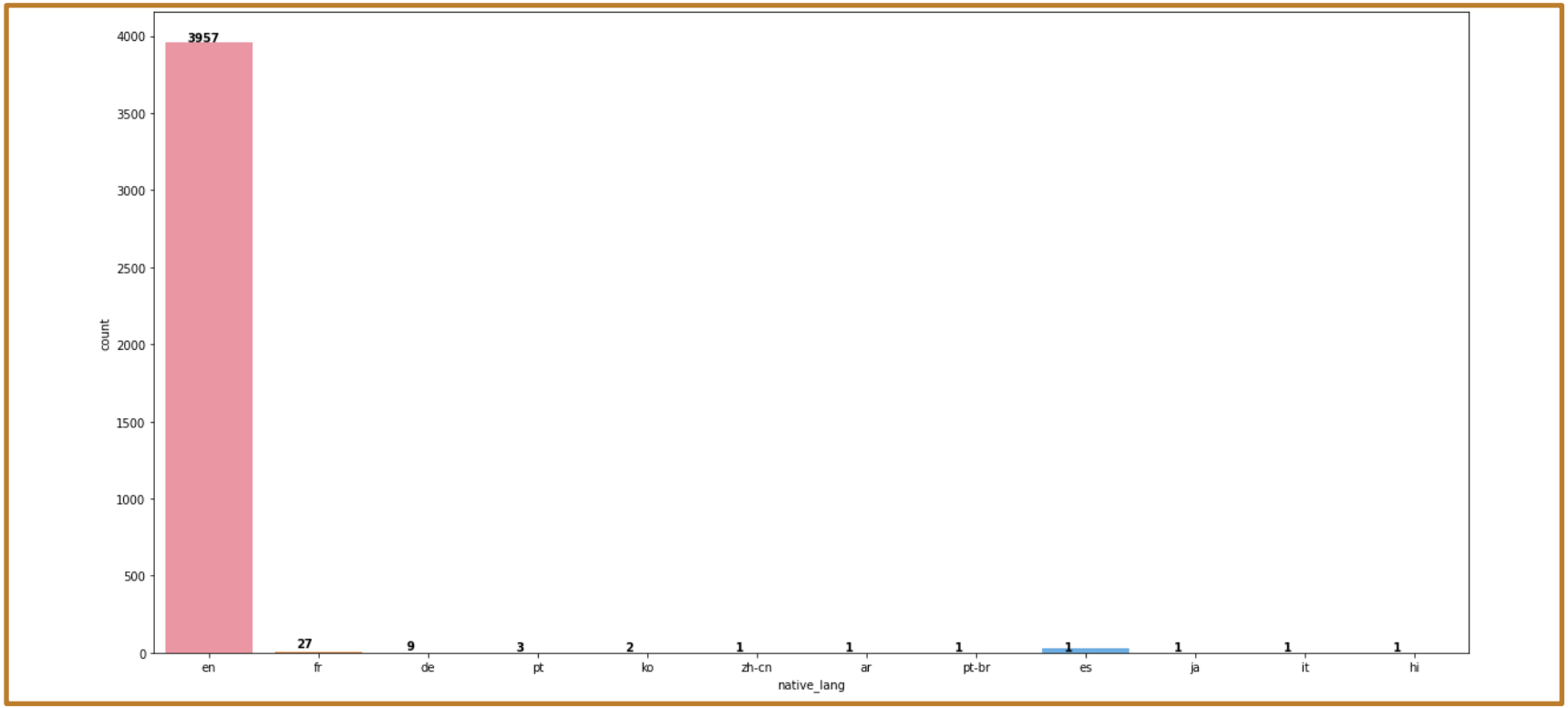
EDA (continued)

Top 10 events with sum of views



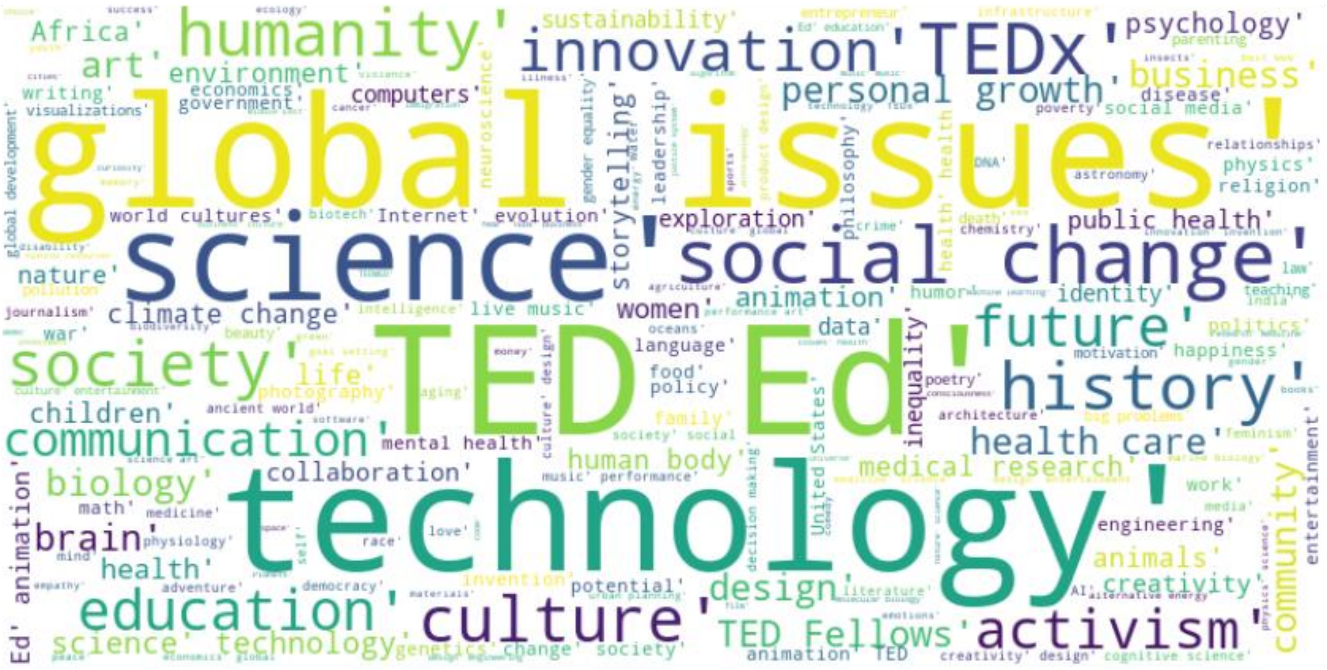
EDA (continued)

Top 10 native languages



EDA (continued)

Most Popular Topic Words



EDA (continued)

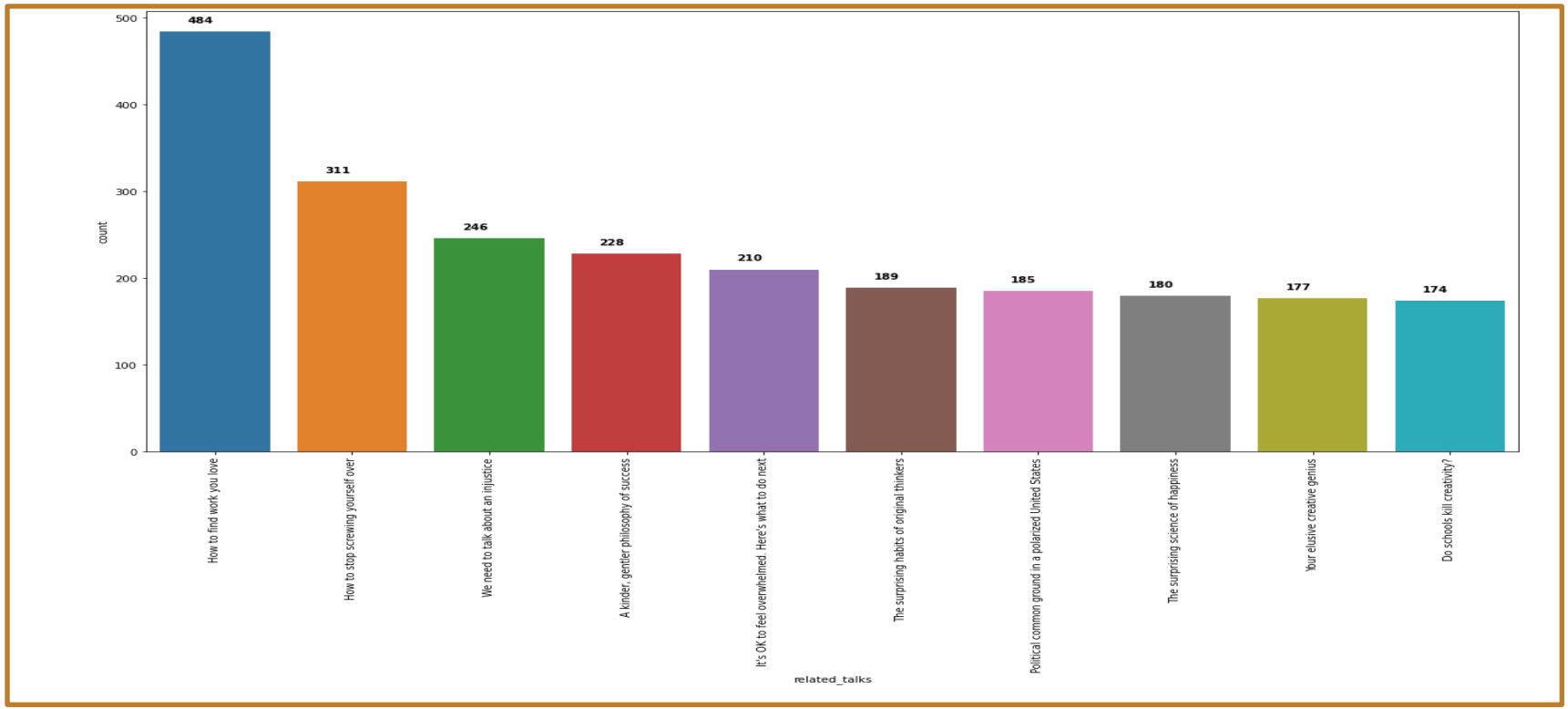
Most Popular Languages in Available Languages

EDA (continued)

Most Popular Words In Related Talks

EDA (continued)

Most frequent related talks categories



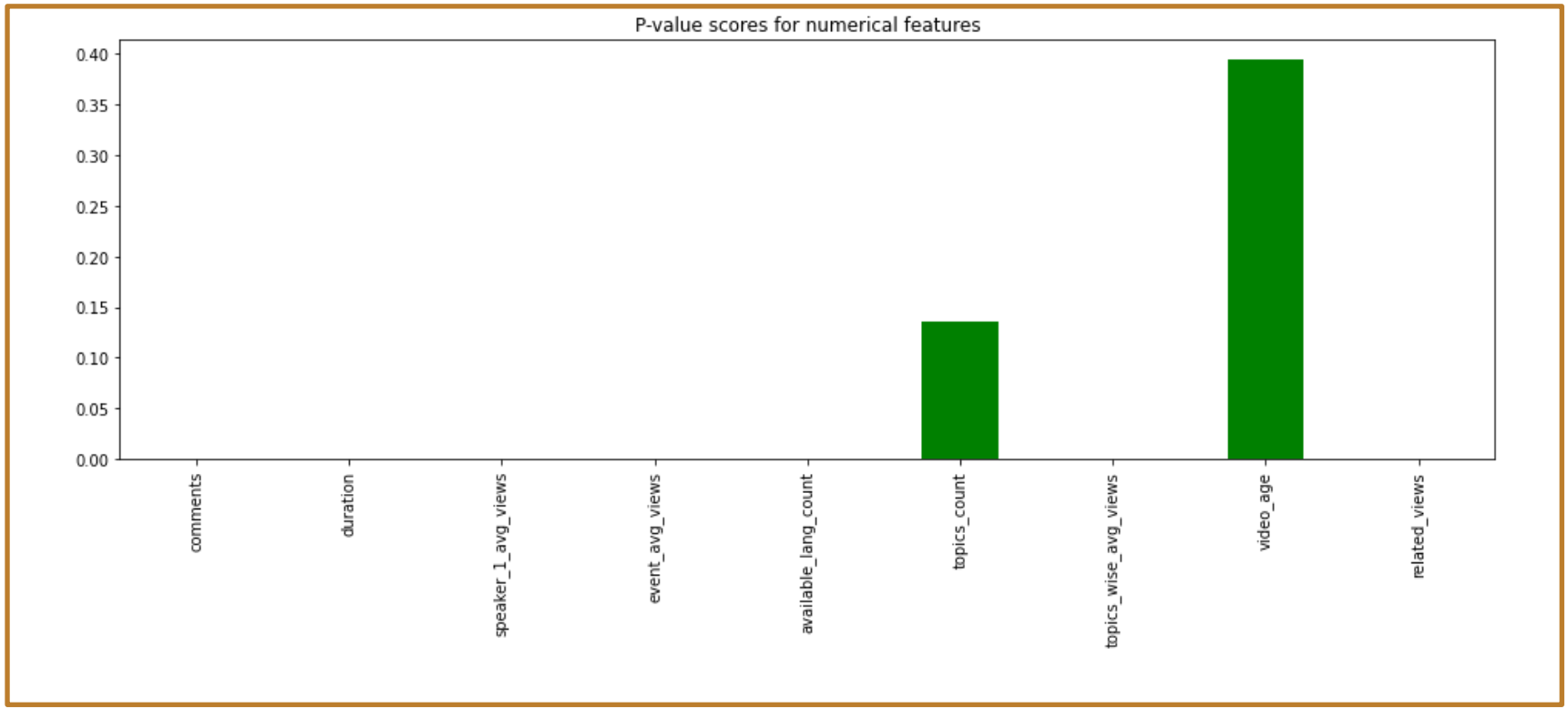
Feature Engineering

1. Speaker_1
2. Event
3. Available_lang
4. Topics
5. Publishes_date
6. Related_talks

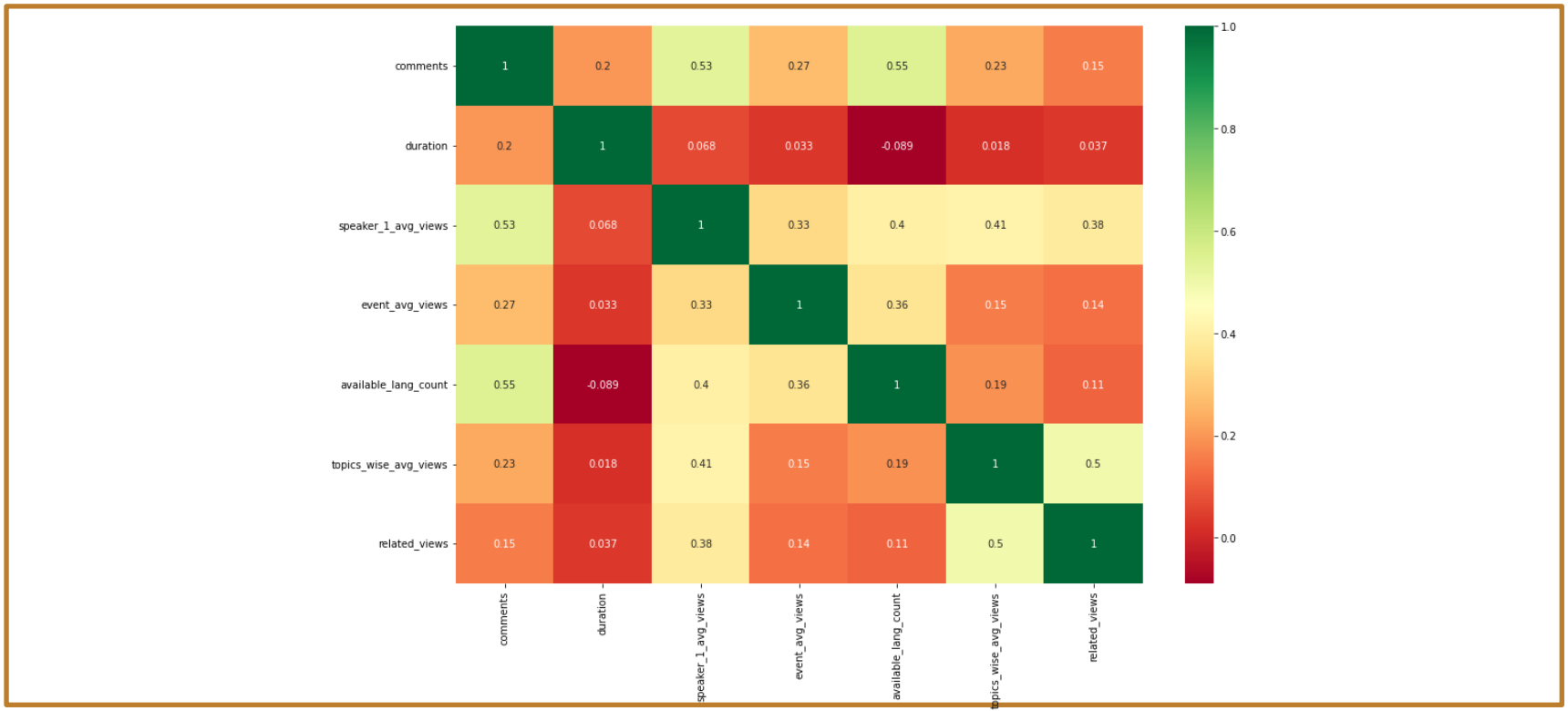
Data Cleaning

1. Outliner Detection & Treatment Using IQR
2. Missing Values Treatment Using Knn Imputer
3. Transformations Using nplog1p

Feature Selection Using F-Regressor



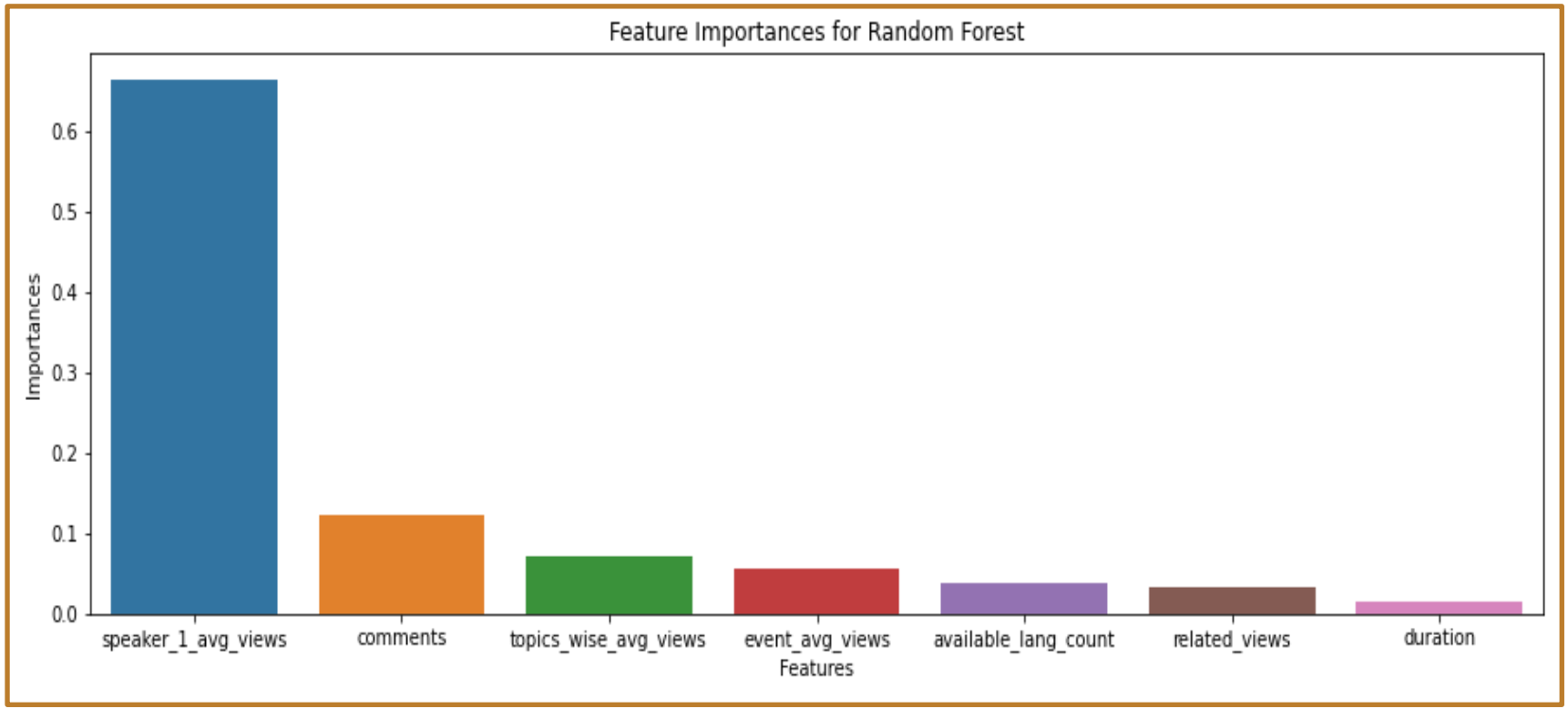
Correlation



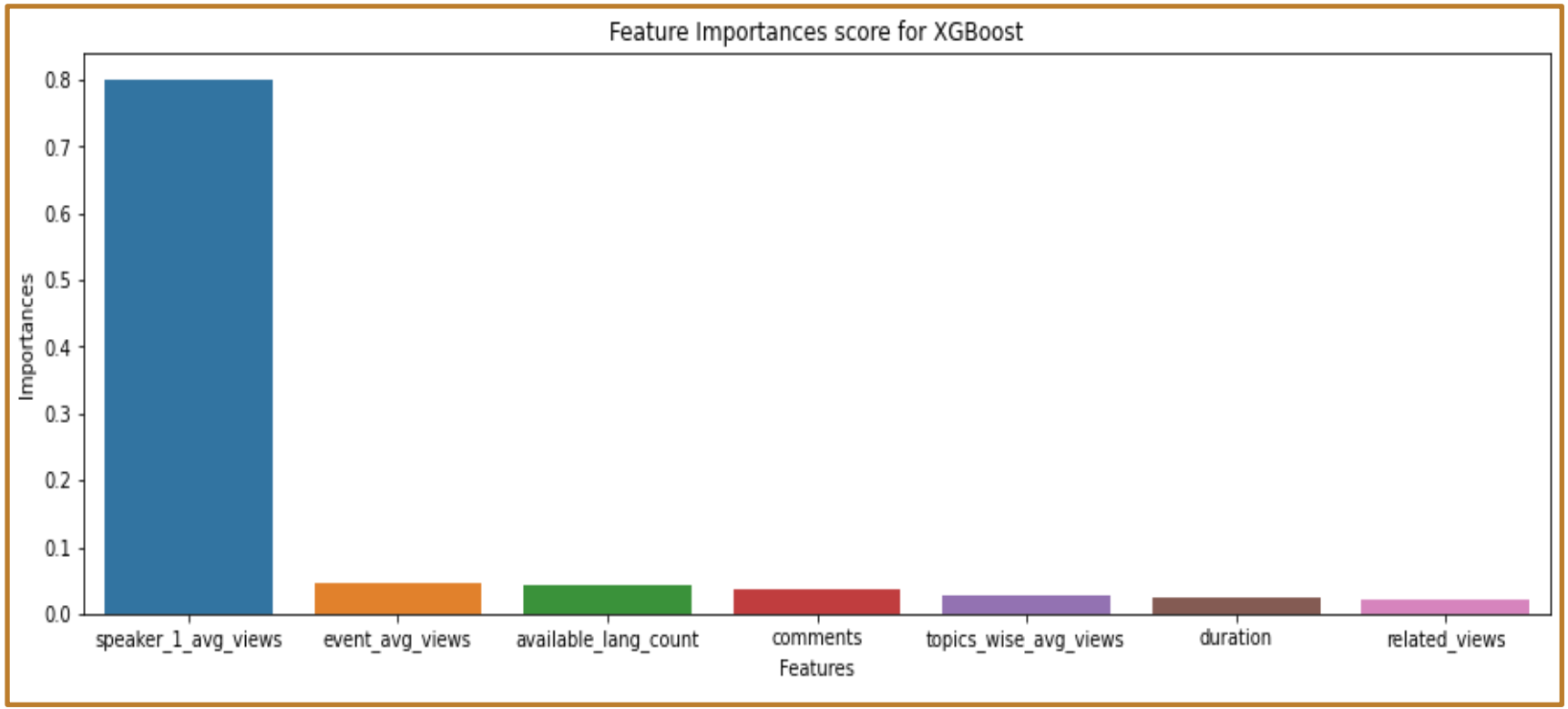
Model Fitting

1. Linear Regression Model
2. Random Forest Model
3. XG Boost Model

Feature Importance by Random Forest Regressor



Feature Importance XG Boost Regressor



Model Comparison

Model	MAE_train	MAE_test	r2_train	r2_test	rmse_train	rmse_test
LinearRegression	268846.003262	261094.062321	0.814269	0.818878	475150.959177	469103.121563
Lasso:	268845.979587	261094.037466	0.814269	0.818878	475150.959177	469103.117093
Ridge:	268845.924378	261093.976795	0.814269	0.818878	475150.959177	469103.103202
KNeighborsRegressor:	231054.636746	281932.195424	0.856594	0.783506	417516.848940	512868.181145
RandomForest	187341.369692	192304.562318	0.805412	0.802992	486348.161894	489242.652464
XGBRegressor:	103147.812151	270187.537440	0.976271	0.792271	169837.507666	502379.329263

Challenges

1. Dataset has lot of features contains a categories in it.
2. Adding new features
3. Outliners In numerical Numbers
4. Selection comparison of models

Conclusion

Linear Regressor, Random Forest Regressor, XGB Regressor are the models used on this project and evalued on MSE, RMSE, MAE ,R2 score and Adjusted R2 scores and finally selected the RandomForest Model as it gining the best score in Mean Square Error i.e. MSE is robust to outliners.

In all these 3 models our erros are ranging2,00,000 which is around 10% of the average views. The model has been able to correctly predict views 90% of the time. After hyper parameter tuning, we have prevented overfitting and decreased errors by regularizing and reducing learning rate. Given that only have 10% errors, our models have performed very well on unseen data due to various factors like feature selection, correct model selection.

Thank You