Capstone Project – 2

TED Talk Views Prediction

Presented By:
Aehteshaam Shaikh

Problem Statement :

The TED dataset contains information about all audio-video recordings of TED Talks uploaded to the official TED.com website until the Year 2020.

It contains information about all talks including number of views, number of comments, descriptions, speakers,titles, transcripts, etc.

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:

Sr	Column Name	Description
1	talk_id	Talk identification number provided by TED
2	title	Title of the talk
3	speaker_1	First speaker in TED's speaker list
4	all_speakers	Speakers in the talk
5	occupations	Occupations of the speakers
6	about_speakers	Blurb about each speaker
7	recorded_date	Date the talk was recorded

Data Summary:

Sr	Column Name	Description
8	published_date	Date the talk was published to TED.com
9	event	Event or medium in which the talk was given
10	native_lang	Language the talk was given in
11	available_lang	All available languages (lang_code) for a talk
12	comments	Count of comments
13	duration	Duration in seconds
14	topics	Related tags or topics for the talk

Data Summary:

Sr	Column Name	Description
15	**related_talks**	Related talks (key='talk_id', value='title')
16	**url**	URL of the talk
17	**description**	Description of the talk
18	**transcript**	Full transcript of the talk
19	**views**	Count of views (Target Variable)

Data Cleaning:

The data set consist of around 4005 rows and 19 columns.

Following columns consisted of Null values:

1. all_speakers
2. occupations
3. about_speakers
4. recorded_date
5. Comments

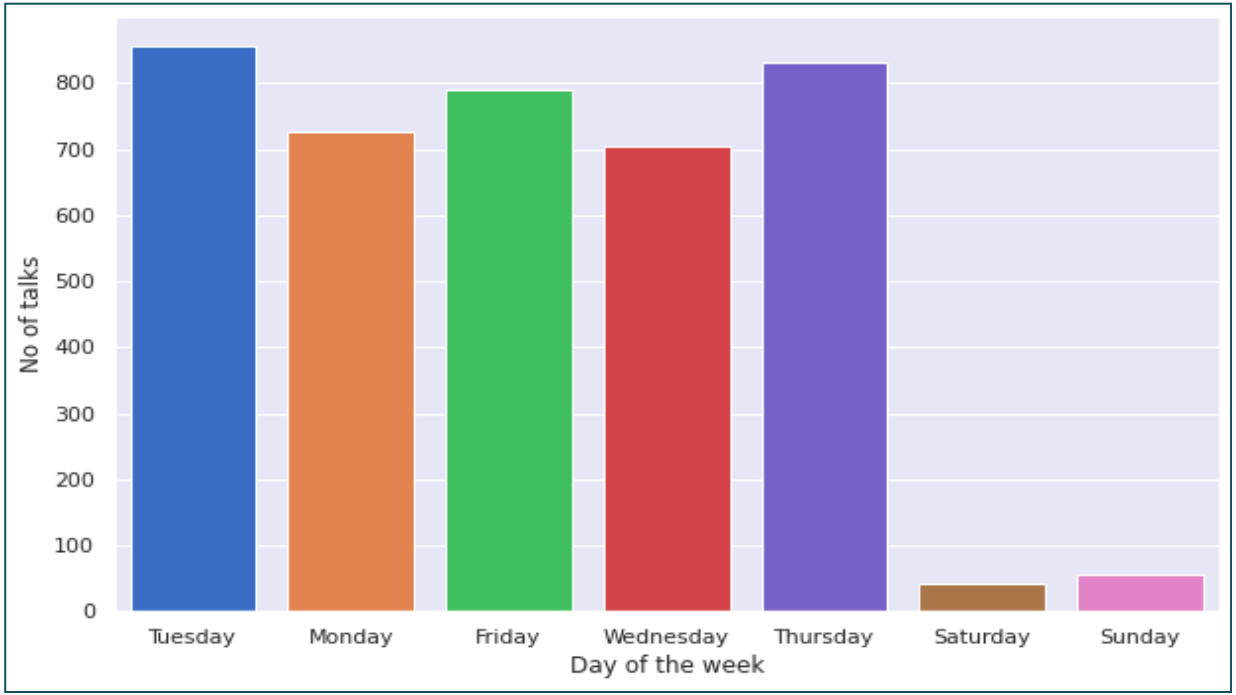
The Null values in the occupations, comments, recorded date, all_speakers were replaced with suitale values.

The column about_speakers has been dropped.

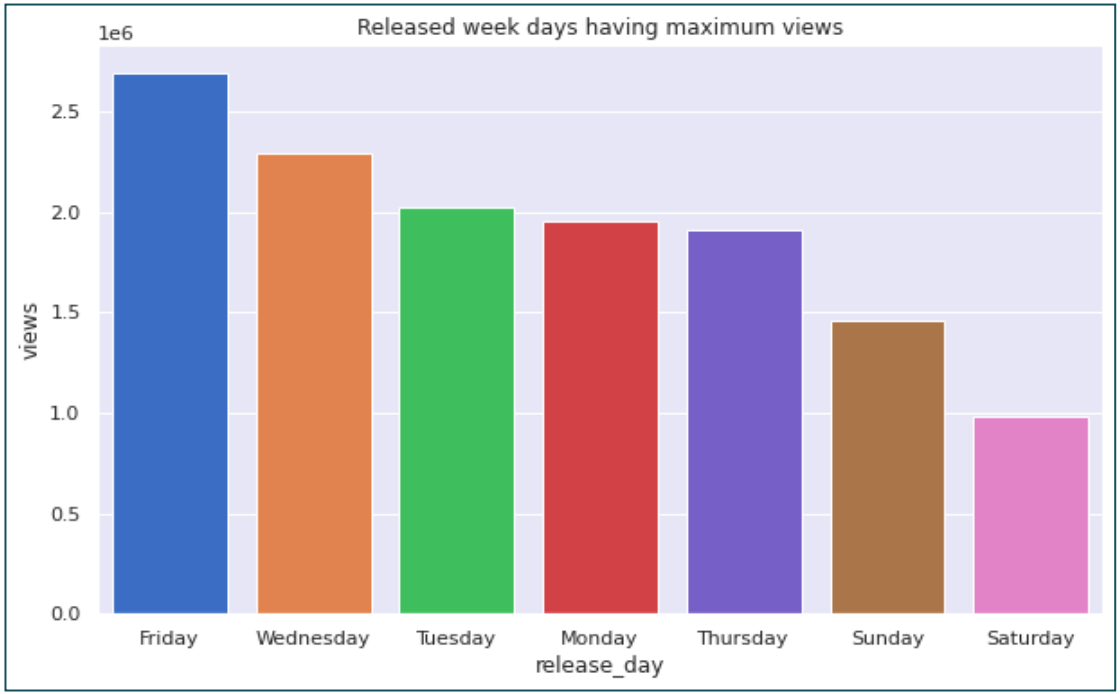
Exploratory Data Analysis

Most no of talks were released on Tuesday followed by Thursday.

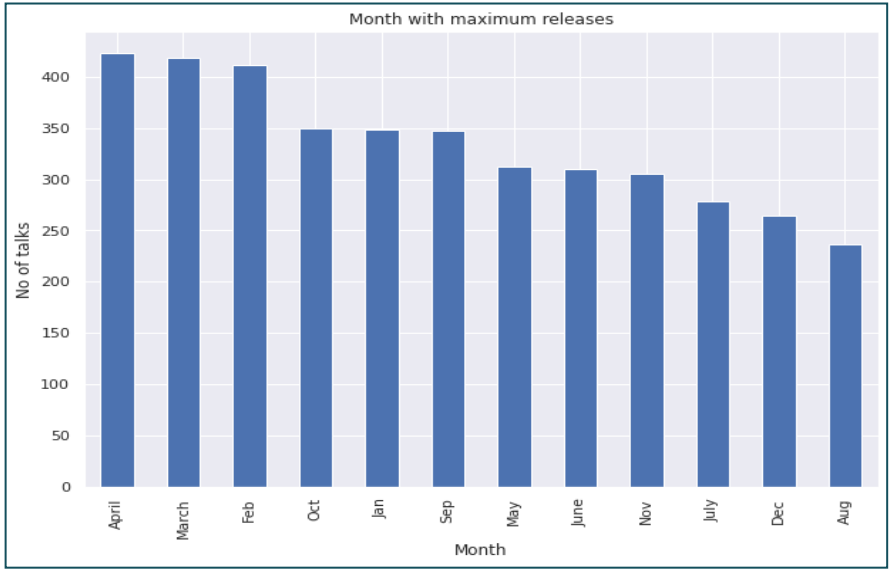
As we can see on saturday and sunday very less ted talk videos were published



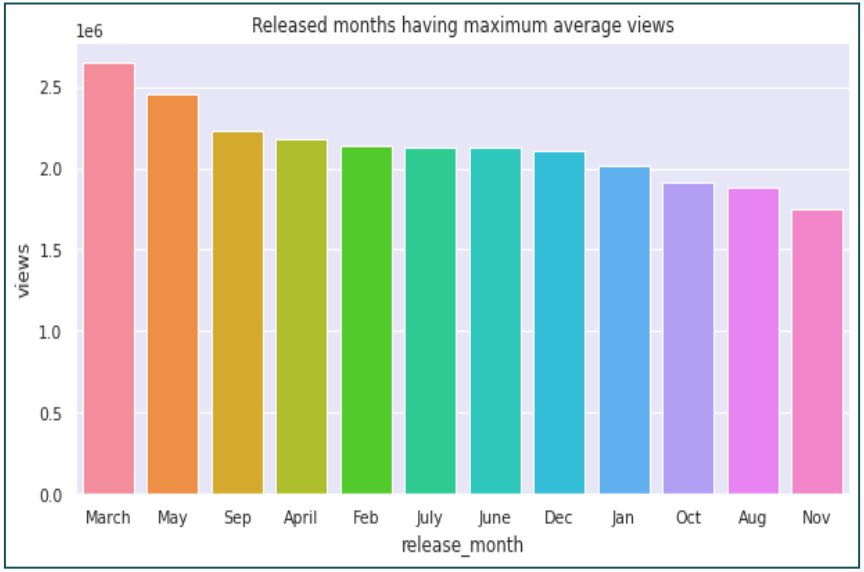
Most no of views are for the videos which are released on Friday.



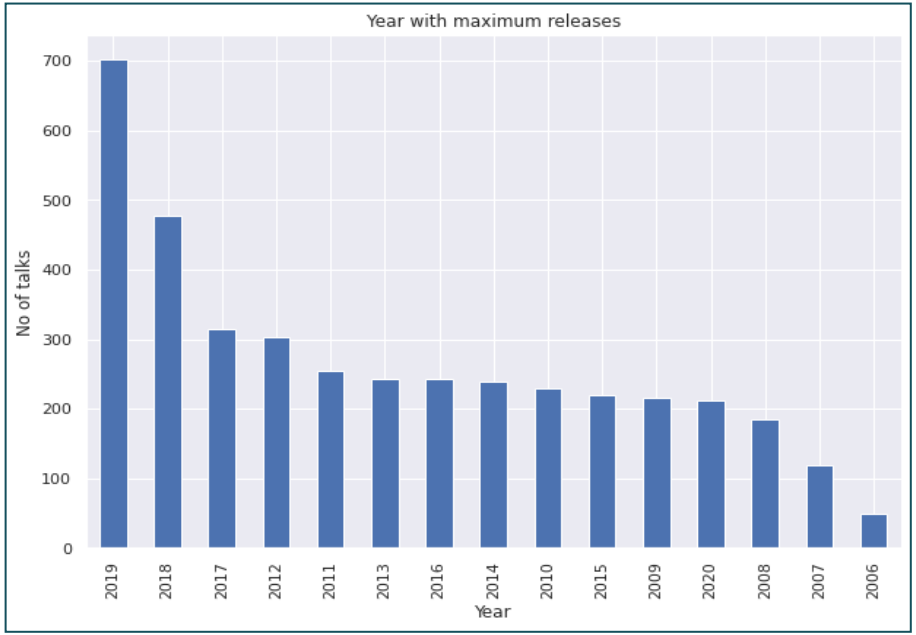
Most no of talks were released in April followed by March and Feb.

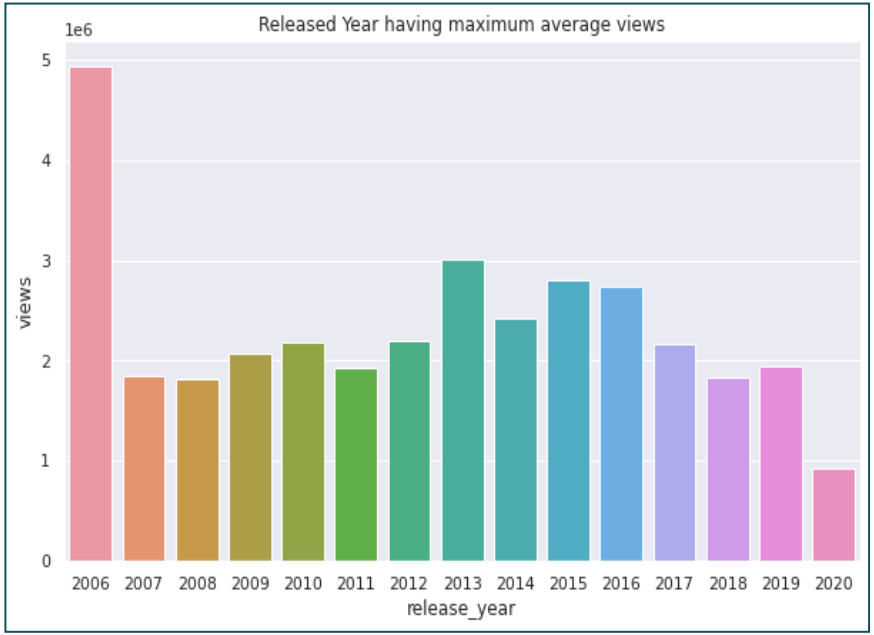


Talks released in March have most views followed by May.



Most no of Talks were released in year 2019 but Talks released in the year 2006 have the maximum average views.





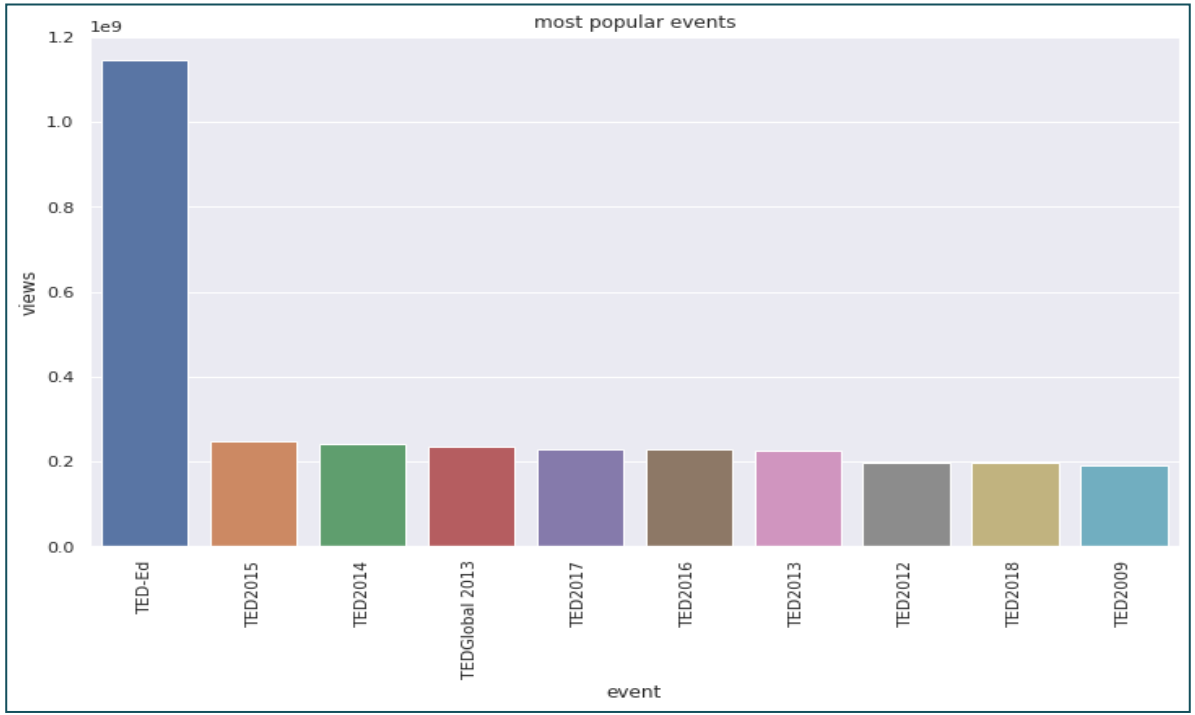
Sir Ken Robinson's 'Do schools kill creativity?' is the most viewed Talk with more than 65 million views.

speaker_1	views	title	release_year
Sir Ken Robinson	65051954	Do schools kill creativity?	2006
Amy Cuddy	57074270	Your body language may shape who you are	2012
James Veitch	56932551	This is what happens when you reply to spam email	2016
Simon Sinek	49730580	How great leaders inspire action	2010
Brené Brown	47544833	The power of vulnerability	2010

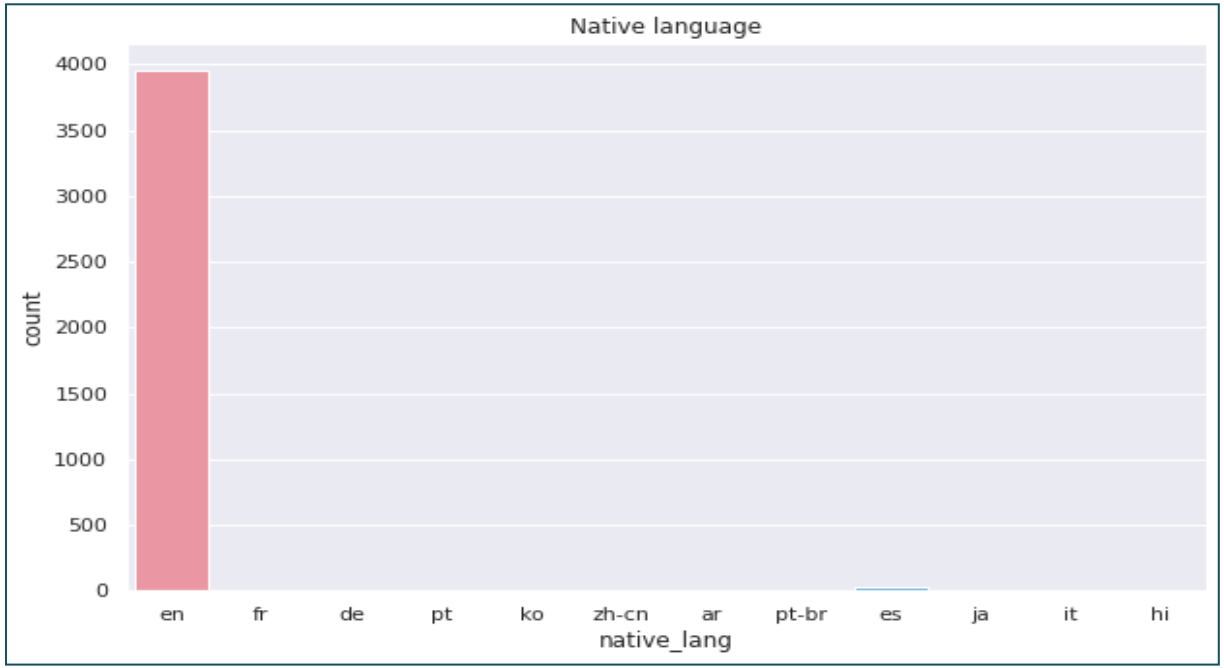
Alex Gendler is the most Popular speaker followed by Sir Ken Robinson

	speaker_1	views
0	Alex Gendler	117619583
1	Sir Ken Robinson	84380518
2	James Veitch	78843641
3	Simon Sinek	62661183
4	Brené Brown	61285977
5	Bill Gates	57107176
6	Amy Cuddy	57074270
7	Julian Treasure	54799681
8	Hans Rosling	39871561
9	Tim Urban	37976820

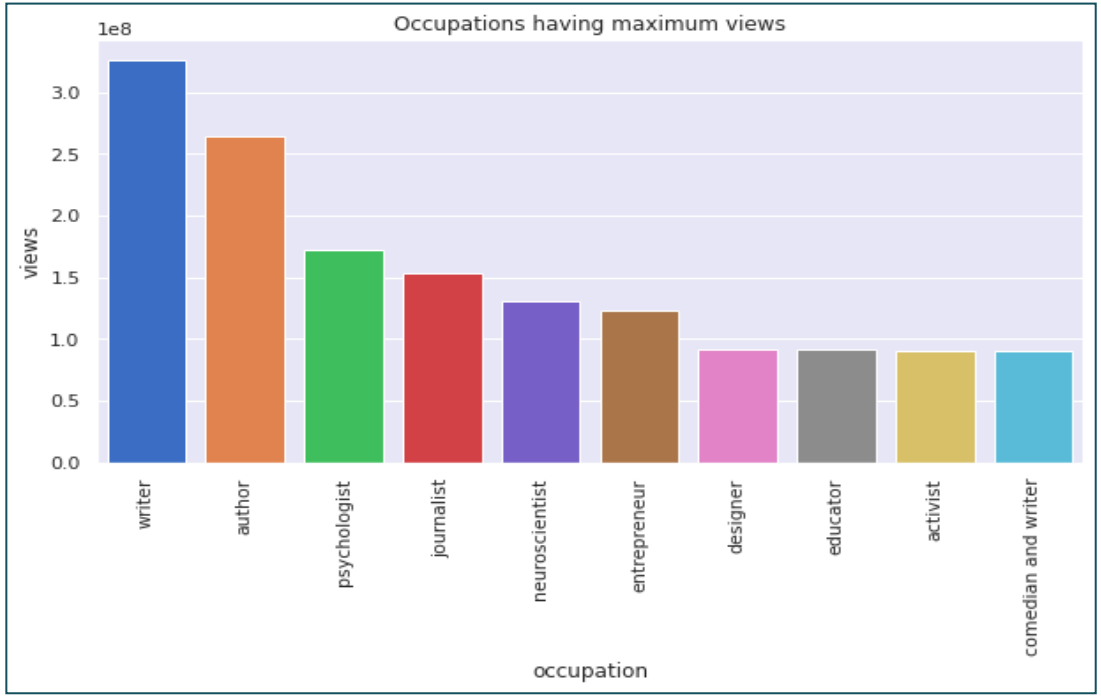
TED-Ed is the most popular event category with most no of views



Almost 99% videos are recorded in English language



Speakers who are Writers are most popular followed by Authors and Psychologists.



It seems that most of the debated topics are related to mainly Science and Religion

	title	comments
0	Militant atheism	6449.0
1	Do schools kill creativity?	4931.0
2	Science can answer moral questions	3424.0
3	How do you explain consciousness?	3006.0
4	My stroke of insight	2984.0
5	Your body language may shape who you are	2633.0
6	Taking imagination seriously	2529.0
7	On reading the Koran	2463.0
8	The danger of science denial	2366.0
9	The power of vulnerability	2209.0

Feature Engineering :

The following Categorical Features are replaced by doing Mean Encoding.

speaker_1 → avg_views_by_speaker

event → avg_views_by_event

The following Categorical Features are replaced by numerical features.

available_lang → total_lang

topics → no_of_topics

added a new column called *video_age* which is the difference of current year and published year.

Feature Engineering :

Following columns consisted Outliers:

1. avg_views_by_event
2. avg_views_by_speaker
3. comments
4. duration
5. no_of_topics
6. total_lang
7. views

The outliers are replaced with the Extreme values.

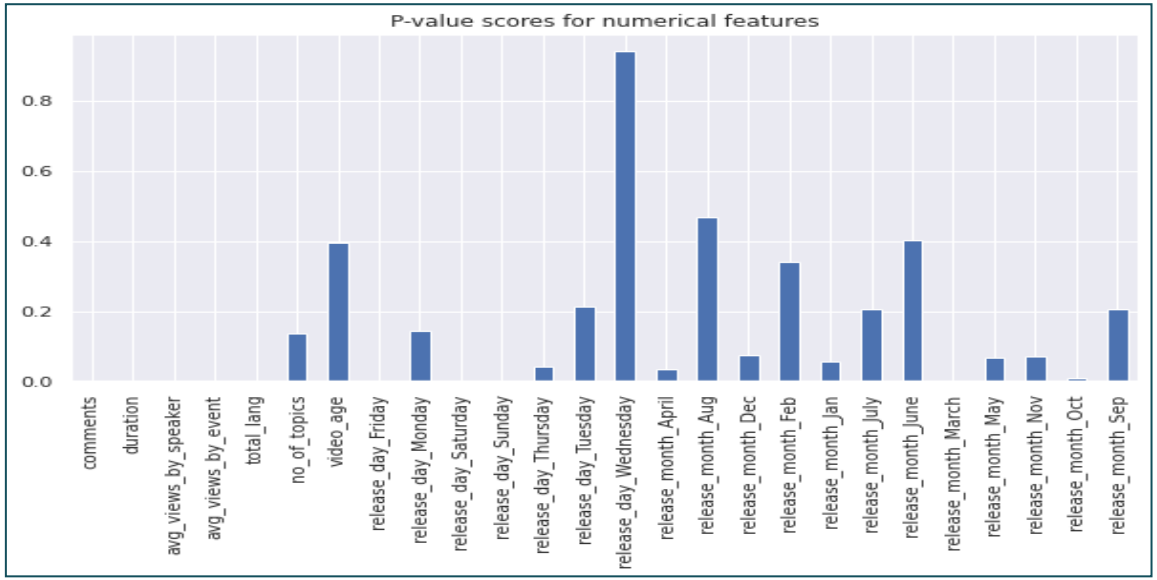
Feature Engineering :

One Hot Encoding is performed on the following categorical Features:

1. release_day
2. release_month

Feature Selection:

Using f-scores, the features with high p-values have been dropped to get the final list of dependent variables.



Machine Learning

ML Models Used :

1. Linear Regression
2. Random Forest Regressor
3. XGBoost Regressor

Model Tuning methods used:

1. Ridge Regression
2. Lasso Regression
3. Elastic Net Regression

Hyper-Parameter Tuning metods used:

1. GridSearch CV
2. RandomSearch CV

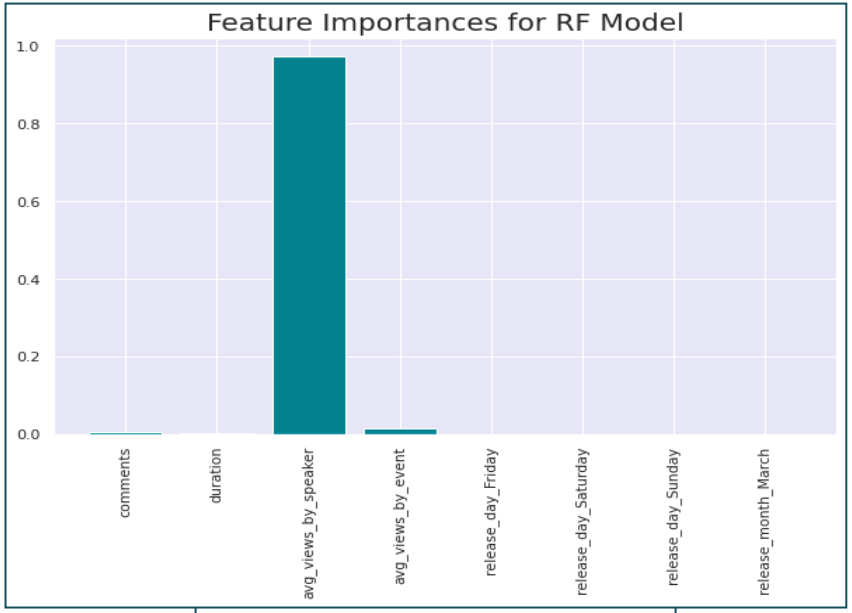
Results obtained after Training the Dataset :

	Model_Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
0	Linear Regressor:	259823.070604	260752.218997	0.816289	0.795937	477776.340431	484604.141408
1	Ridge Regressor:	259767.447921	260672.286088	0.816288	0.795967	477776.923731	484568.983037
2	RandomForest	173867.911902	191102.733149	0.843381	0.798938	441142.150547	481028.240843
3	XGBRegressor:	186342.837263	215350.871952	0.884204	0.818376	379318.551255	457184.533771

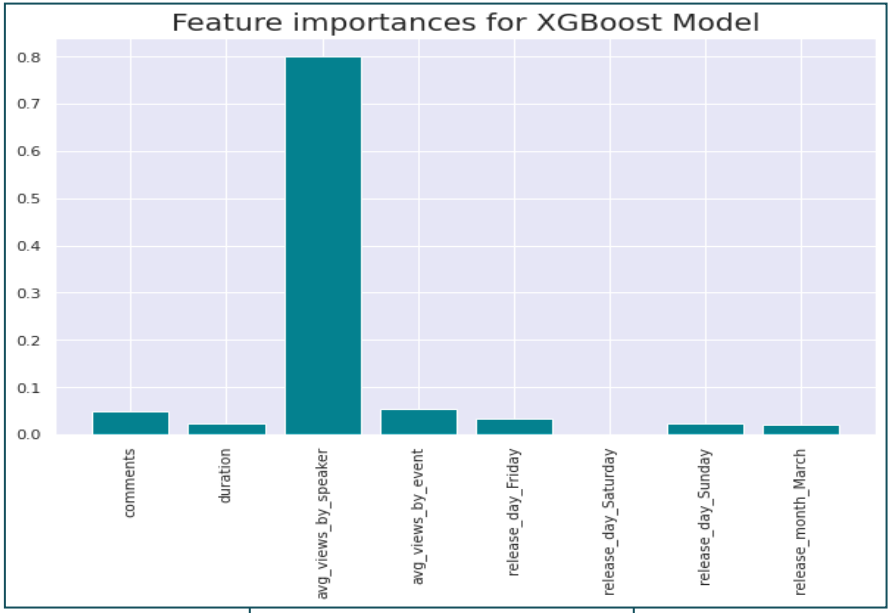
1. In terms of RMSE and R-squared, *XGBoost* is the best performer
2. In terms of Mean Absolute Error, *Random Forest Regressor* is the best performer .

Feature Importance:

In all of the models, it has been observed that the column *avg views by speaker* is the most important feature in the dataset followed by *avg views by event*.



Random Forest Regressor



XGBoost Regressor

Conclusion:

- After loading the dataset, cleaning the data, performing EDA, Feature Engineering and after feature selection, Models are built.
- In all of the models, it has been observed that the column *avg views by speaker* is the most important feature.
- In terms of RMSE score XGBoost Regressor gave the best results. As, to compare the Accuracy among different regression models, RMSE is a better option as it is simple to calculate and differentiable. However, our dataset had outliers, hence MAE is better metric than RMSE as it is robust to outliers.
- So, after comparing MAE values it is evident that *Random Forest Regressor* is the best performer.

that the column avg views by sp

PRESENTATION