Capstone Project-2

Ted Talk Views Prediction
ML Supervised Regression

Individual Project:
Avisikta Majumdar

Contents

- *Problem Statement*
- *Data Summary*
- *Data Analysis*
- *Feature Selection*
- *Data Preparation*
- *Implementing Various Regression Algorithms*
- *Hyperparameter tuning*
- *Conclusions*

Problem Statement

- **Prediction of the views of the videos uploaded on the TEDx website.**

Let's see the features'

Features(19)

Numerical(4)

- *talik_id*
- *views*
- *comments*
- *duration*

Categorical(15)

- About_speakers
- all_speakers
- Available_lang
- Description
- Event
- Native_lang
- Occupations
- Published_date
- Recorded_date
- Related_talks
- Speaker_1
- Title
- Topics
- Transcript
- url

Basic Data Exploration

- This dataset is having 4005 observations & 19 features.
- Most of the features are categorical .
- No duplicate values.

```
Dataset Shape: (4005, 19)
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4005 entries, 1 to 62794
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   title            4005 non-null   object
 1   speaker_1        4005 non-null   object
 2   all_speakers     4001 non-null   object
 3   occupations      3483 non-null   object
 4   about_speakers   3502 non-null   object
 5   views            4005 non-null   int64
 6   recorded_date    4004 non-null   object
 7   published_date   4005 non-null   object
 8   event            4005 non-null   object
 9   native_lang      4005 non-null   object
 10  available_lang   4005 non-null   object
 11  comments         3350 non-null   float64
 12  duration         4005 non-null   int64
 13  topics           4005 non-null   object
 14  related_talks    4005 non-null   object
 15  url              4005 non-null   object
 16  description      4005 non-null   object
 17  transcript       4005 non-null   object
dtypes: float64(1), int64(2), object(15)
memory usage: 116.5 MB
```

Data Exploration(NaN values)

AI

	Feature_Name	Missing	Uniques	%age of missing values
11	comments	655	601	16.35
3	occupations	522	2049	13.03
4	about_speakers	503	2977	12.56
2	all_speakers	4	3306	0.10
6	recorded_date	1	1334	0.02
0	title	0	4005	0.00
16	description	0	4005	0.00
15	url	0	4005	0.00
14	related_talks	0	4005	0.00
13	topics	0	3977	0.00
12	duration	0	1188	0.00
9	native_lang	0	12	0.00
10	available_lang	0	3902	0.00
1	speaker_1	0	3274	0.00
8	event	0	459	0.00
7	published_date	0	2962	0.00
5	views	0	3996	0.00
17	transcript	0	4005	0.00

NaN

- 16% NaN values are present in ***comments***
- 13% NaN values are present in ***occupations***
- 12.5% NaN values are present in *about_speakers*

Unique value

Most of the columns except **native_lang , event** are containing unique values.

Data Processing

- Initially the datatype of ***published_date* , *recorded_date*** *was in string format, i have used pandas* ***to_datetime*** *function to convert the datatype*

```
published_date     object
recorded_date      object
dtype: object
```

```
published_date     datetime64[ns]
recorded_date      datetime64[ns]
dtype: object
```

- *Created* **`month`** *,* **`day`***,* **`year`** *columns based on* **`published_date`** *column*

talk_id	published_date	month	year	day
92	2006-06-27	Jun	2006	27
110	2007-04-14	Apr	2007	14

Data Processing

AI

- Created **`time_since_published`** column based on **`published_date & current_date`**

talk_id	published_date	time_since_published
64	2006-09-06	4983 days
45	2006-08-08	5012 days

- Created **`daily_views`** column based on **`views`** & **`time_since_published_date`**

talk_id	published_date	time_since_published	views	daily_views
820	2010-04-07	3674 days	2248059	611
60	2007-02-09	4827 days	1214012	251
2588	2016-09-26	1310 days	2712894	2069

Feature removing

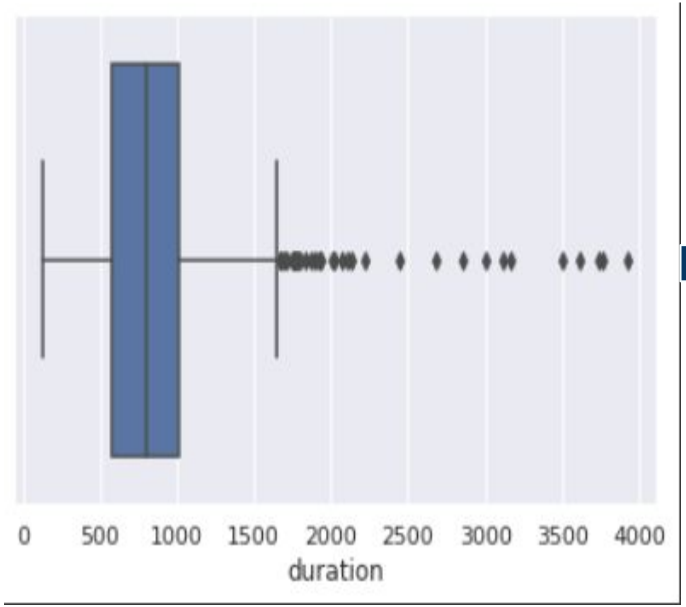
- *Most of the speakers delivered their talk in* ***english***

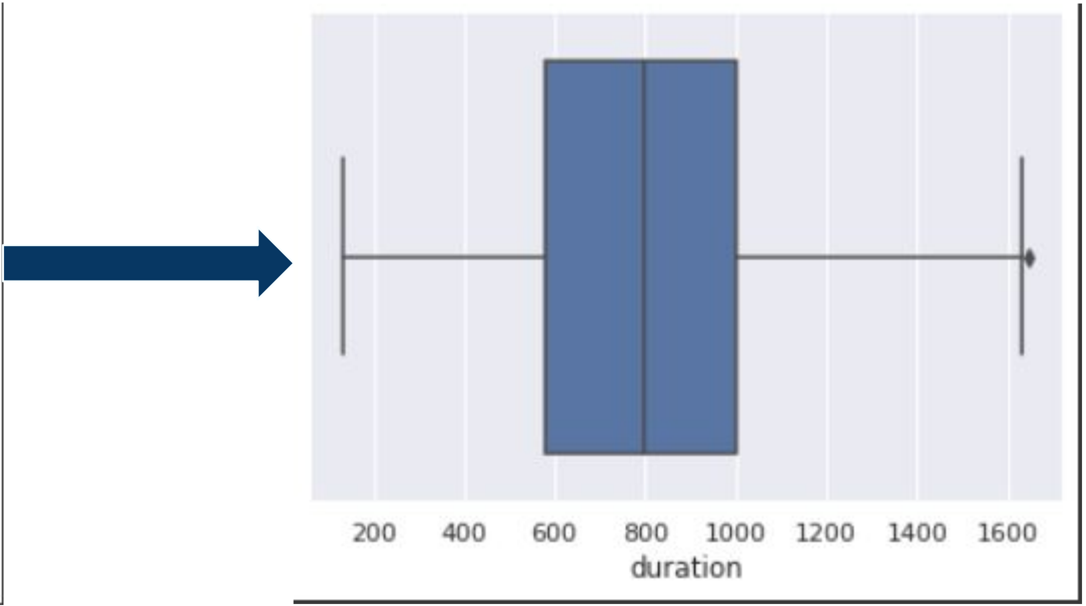
	en	es	fr	hi	pt	it	ko	ja	de	ar	pt-br	zh-cn
native_lang	3306	15	7	2	1	1	1	1	1	1	1	1

- *Removed unnecessary features like*

'talk_id'	'title'	'speaker_1'	'all_speakers'
'occupations'	'about_speakers'	'views'	'recorded_date'
'published_date'	'event'	'native_lang'	'available_lang'
'topics'	'related_talks'	'url'	'description'
'transcript'			

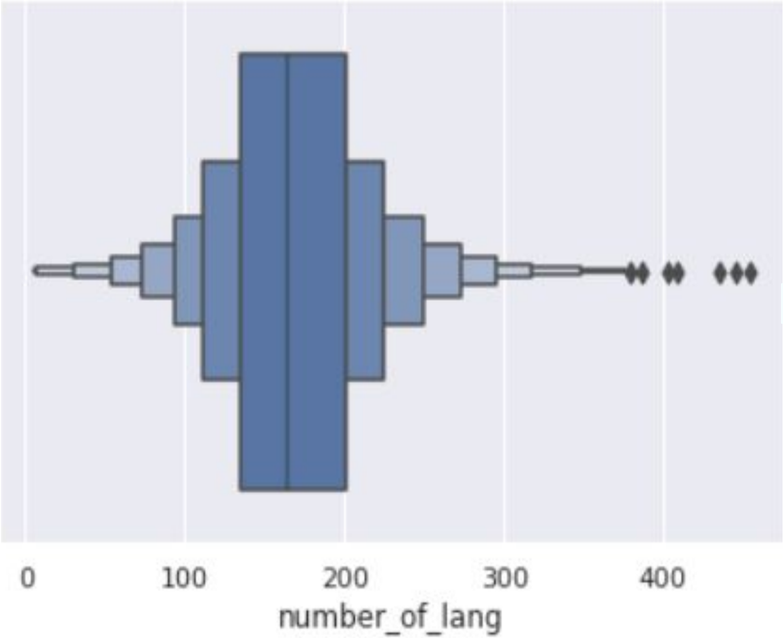
Removing Outliers

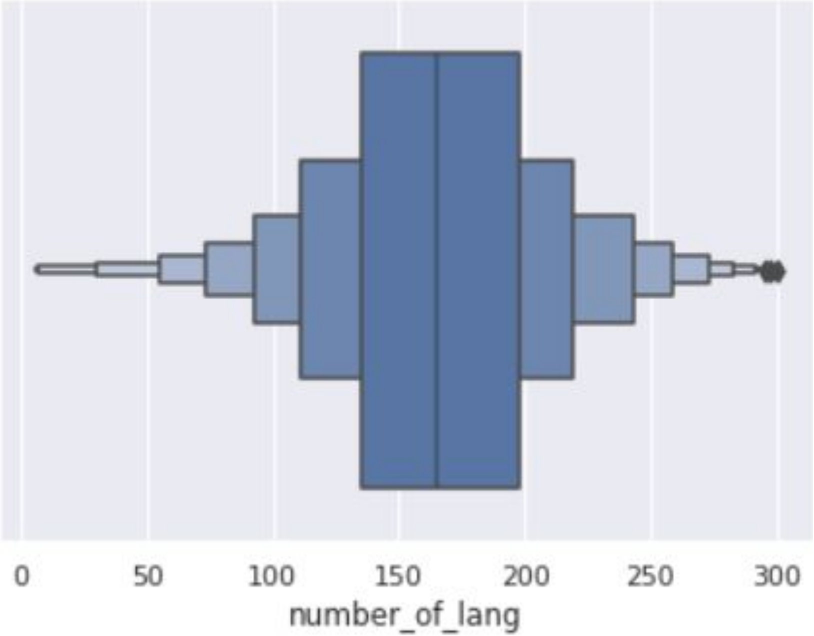




Replaced outliers with mean value of ***duration***

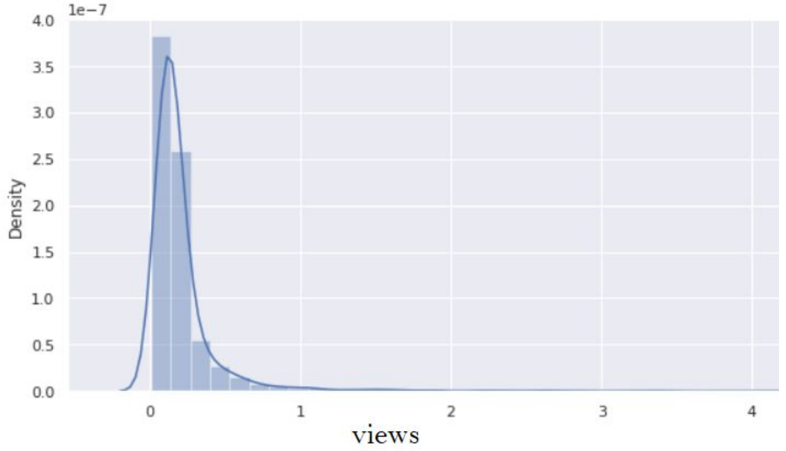
Removing Outliers



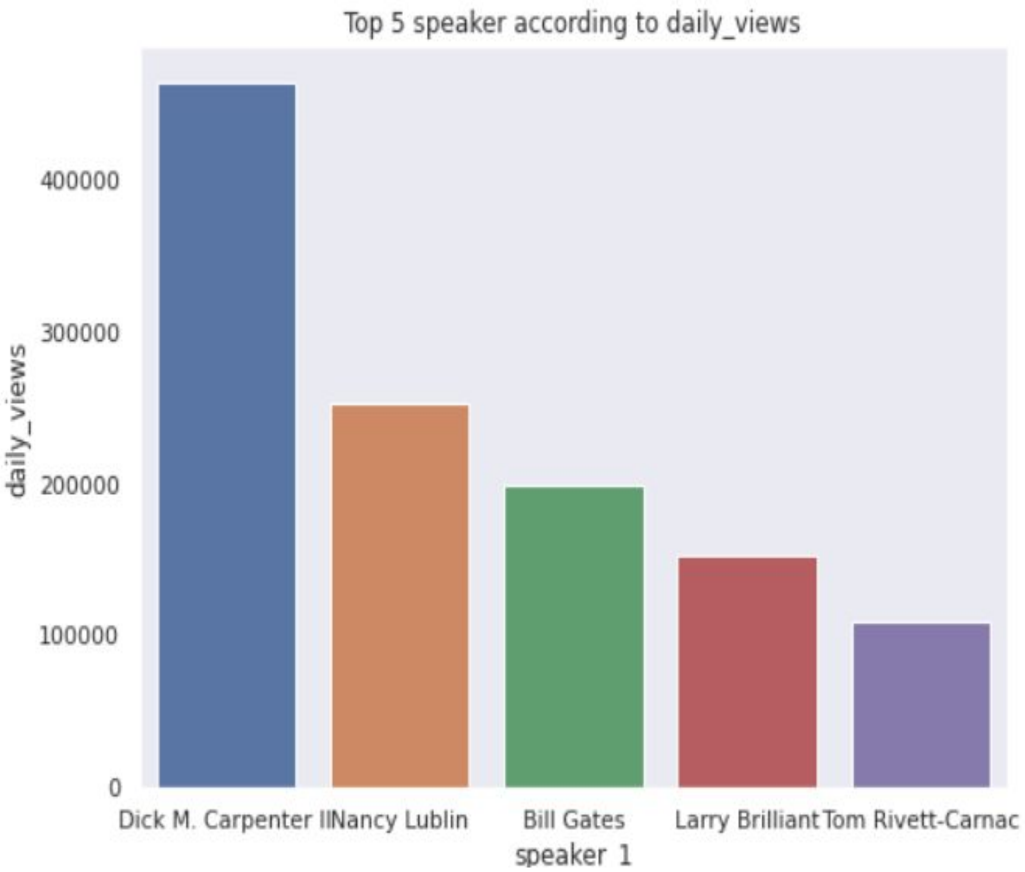


Replaced outliers with mean value of ***number_of_languages***

Visualization

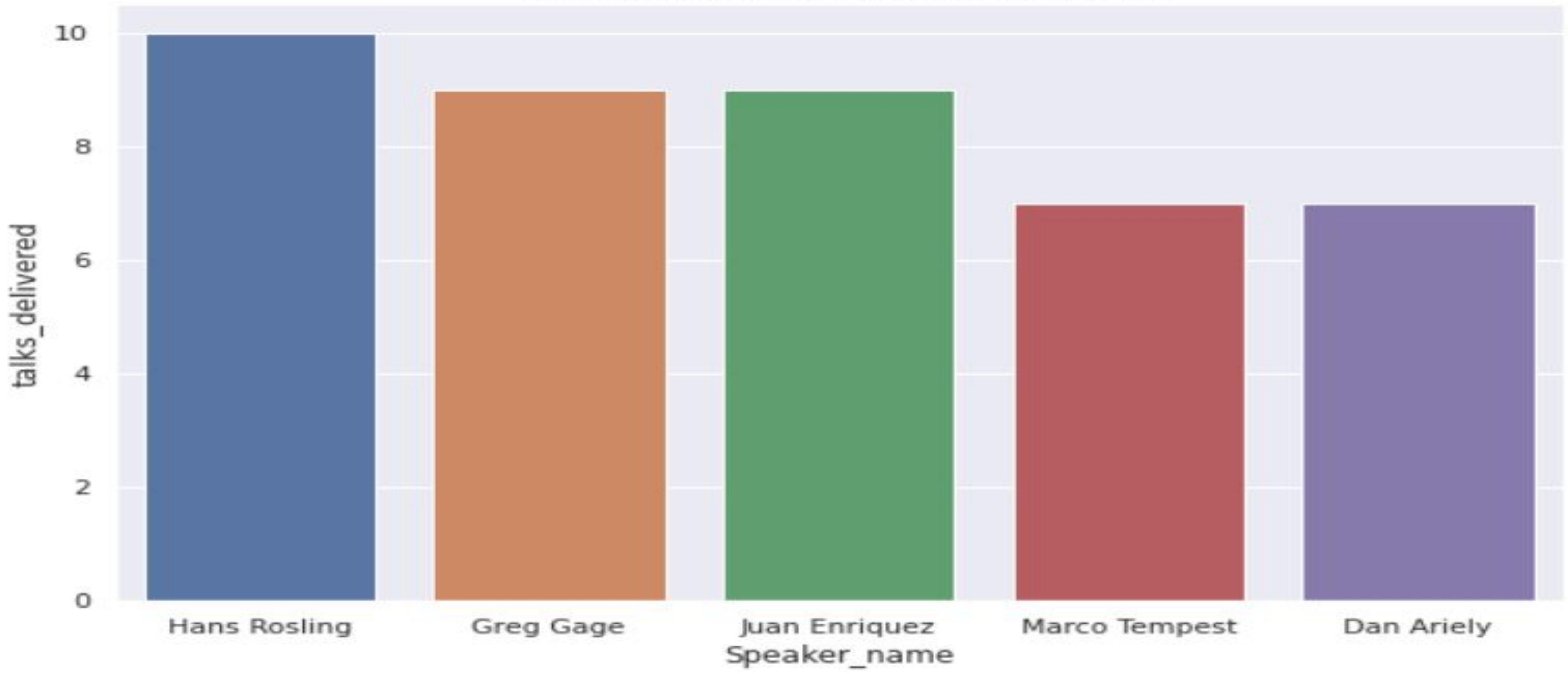


Views is positively skewed



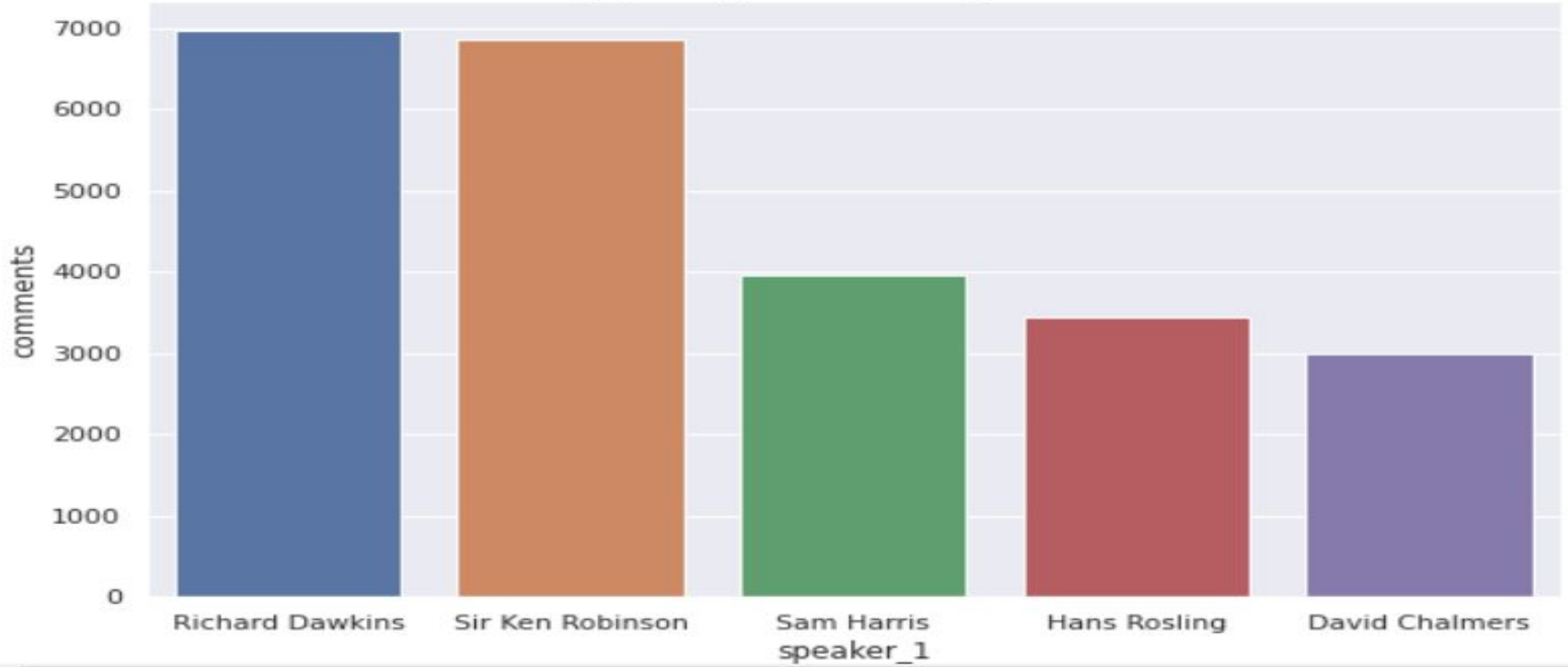
Visualization

(Top 5 speakers who delivered most talks)



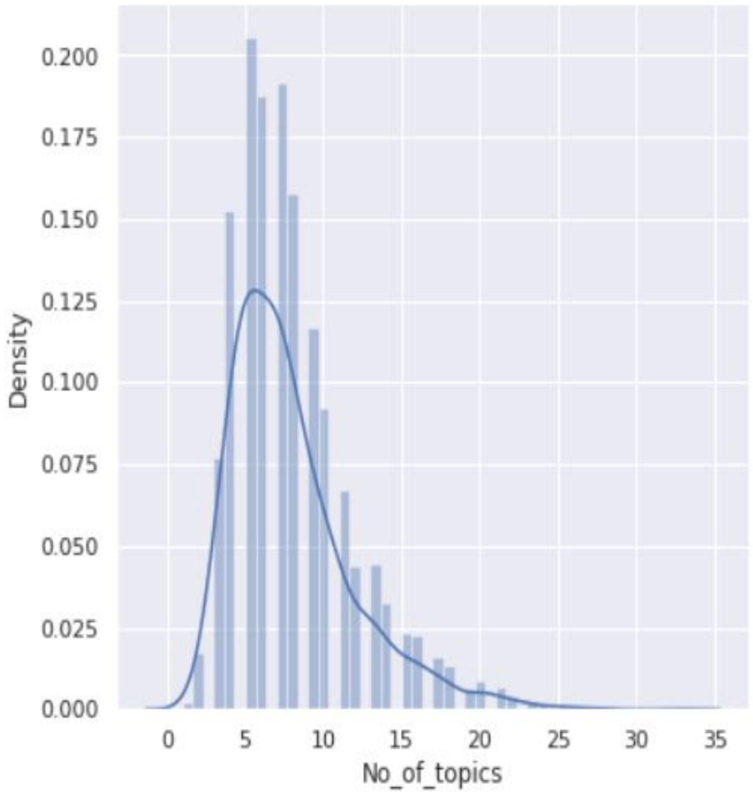
Visualization

(Most popular speakers according to Comments)

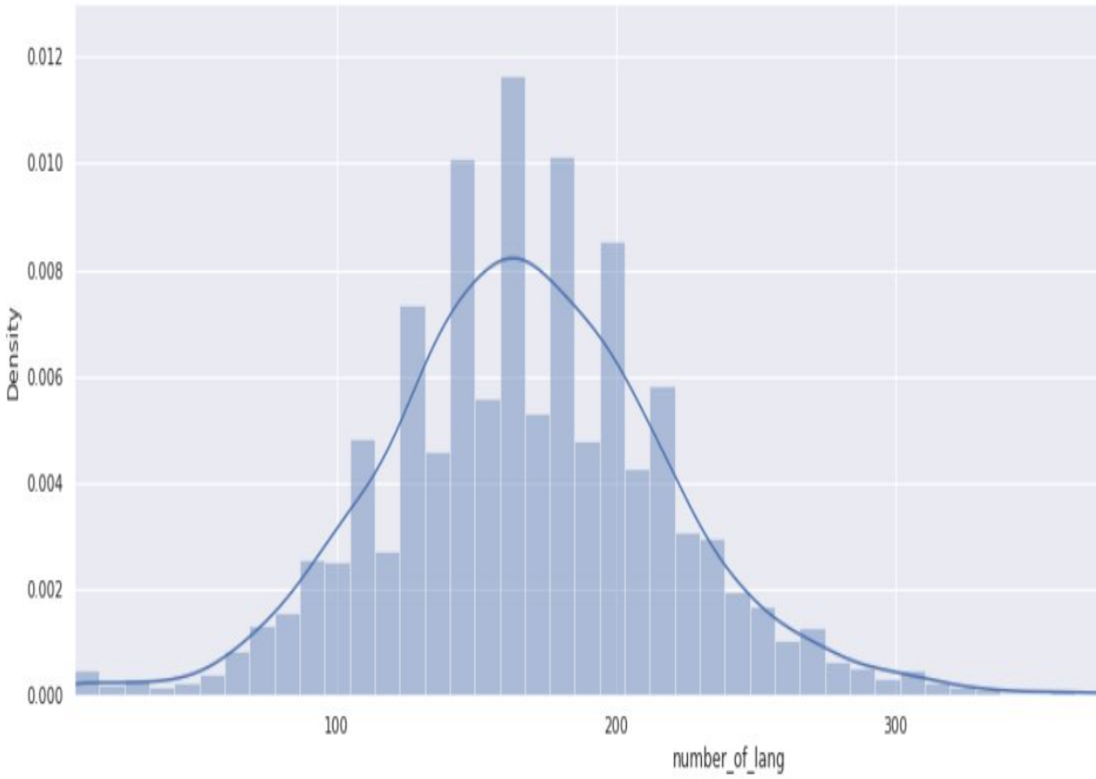


Visualization

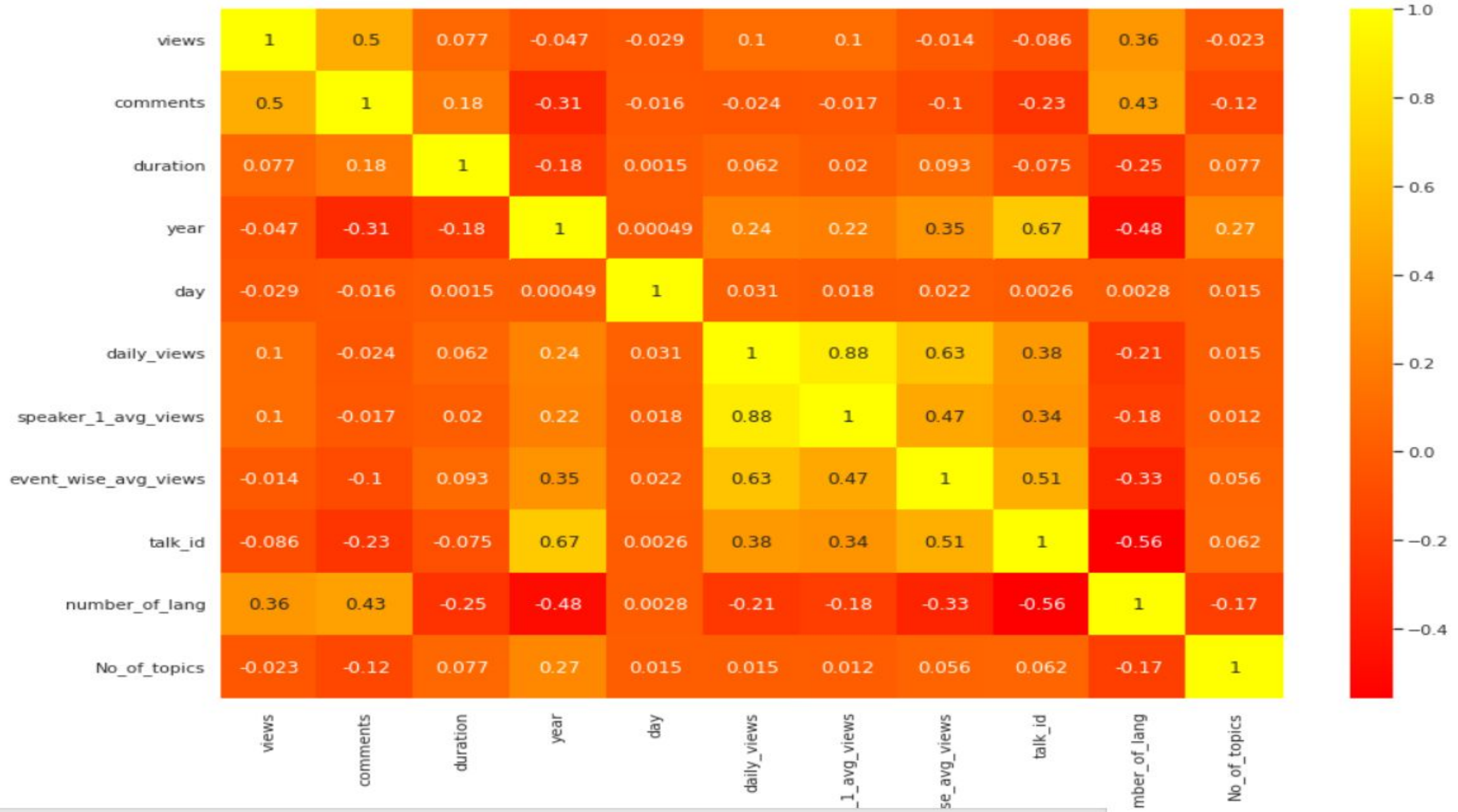
Density plot for no of topics per talk



Density plot for no of languages per talk



Correlation



We can conclude that daily_views column is highly correlated with Speaker_1_avg_views , event_wise_avg_views,

Data Preparation

- ***Independent features*** *:-*

  comments, duration, time_since_published, month, year ,day, daily_views,

  Speaker_1_avg_views , event_wise_avg_views, Number_of_lang , No_of_topics ,

  topics_wise_avg_views
- ***Dependent feature*** *:- daily_views*
- Used **StandScaler**
- Splitted data into 80:20 ratio

Let’s compare those models

	Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
6	GradientBoostingRegressor:	380.283699	759.061577	0.994977	0.399657	857.067785	6248.226254
7	XGBRegressor:	429.726238	680.309046	0.993294	0.766738	990.303692	3894.743516
4	RandomForest	921.695436	839.076255	0.168246	0.335713	11029.234762	6572.562452
3	KNeighborsRegressor:	1031.112739	909.538141	0.541709	0.921037	8186.886733	2266.042923
1	Lasso:	1271.992955	1205.618639	0.859364	0.703730	4535.204569	4389.356547
2	Ridge:	1272.276531	1205.799311	0.859363	0.703867	4535.205410	4388.337808
0	Linear Reg.:	1272.640632	1206.337301	0.859364	0.703543	4535.203672	4390.738157
5	ExtraTreeRegressor :	1528.927152	1371.692837	0.147758	0.305693	11164.243764	6719.433676

We choose MAE and not RMSE as the deciding factor of our model selection because of the following reasons:

- RMSE is heavily influenced by outliers as in the higher the values get the more the RMSE increases.
- MAE doesn’t increase with outliers. MAE is linear and RMSE is quadratically increasing.
- The best performing regressor model for this dataset is XGBRegressor on the basis of MAE.

Hyperparameter Tuning

	Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
0	XGBRegressor_without_hyper	429.888181	680.397815	0.993294	0.766717	990.312172	3894.911052
1	XGBRegressor_with_hyper	102.932939	645.449187	0.999718	0.766717	203.092063	3929.319520

- Used **GridSearchCV** to do hyperparameter tuning
- Hyperparameters I have used :-
  - gamma
  - Learning_rate
  - max_depth
  - n_estimators

Conclusion

Models used

1.Linear Reg.	2. Lasso	3. Ridge	4. ExtraTreeRegressor
5. RandomForest	6. KNeighborsRegressor	7. XGBRegressor	8. GradientBoostingRegressor

Notes: -

- Most of the columns are categorical
- After hyper parameter tuning, we have prevented overfitting out of all these models *XGBRegressor* is the best performer in terms of MAE.
- In all the features *speaker_1_avg_views* is most important this implies that speakers are directly impacting the views.
- R2_score for the final model is 0.99 (train data) & 0.76 (test data)


Thank You!