Capstone Project On
Ted Talk Views prediction

Team name- **Data Newbies**

Team Members - **Rohit Kumar Raj**
Ranjita Raj
Sandhyarani Patra

Cohort- **Canberra**

Contents-

- Overview
- Problem Statement
- Data Pipeline
- EDA on Features
- Feature Engineering
- Feature Selection
- Models Used
- Which model did I choose and why?
- Challenges
- Conclusion

OVERVIEW:

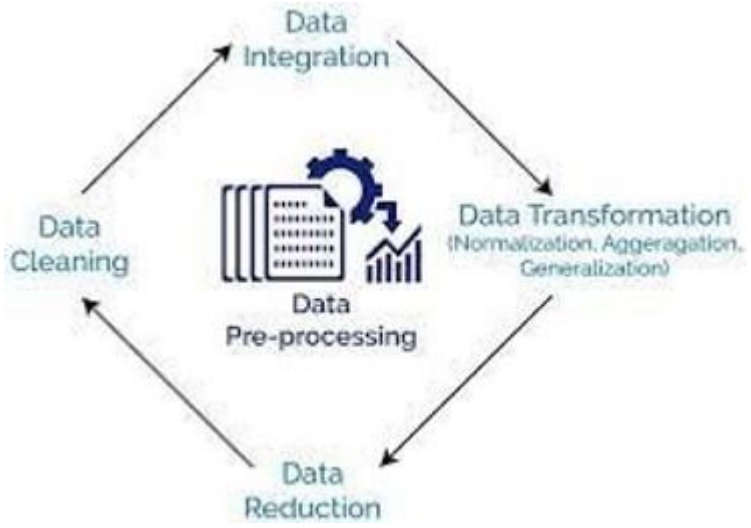
- TED is devoted to spreading powerful ideas on just about any topic. These datasets contain over 4,000 TED talks including transcripts in many languages Founded in 1984 by Richard Salman as a nonprofit organization that aimed at bringing experts from the fields of Technology, Entertainment, and Design 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.

The dataset contains features

- 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
- 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

Data Pipeline

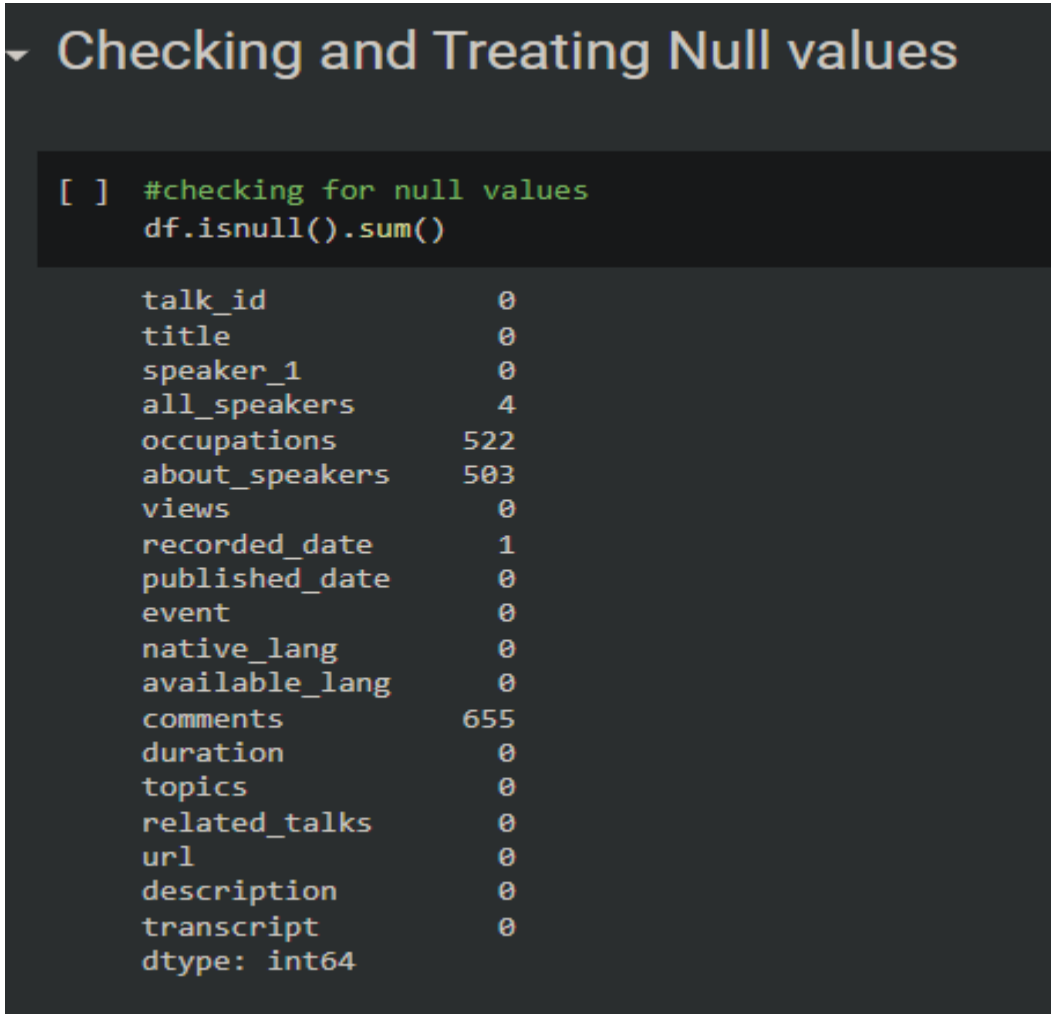
- Understanding the Data
- EDA/Cleaning the Data: The data was checked for duplicate values, null and missing values, and primary inspection was performed. Exploratory data analysis was performed to analyse and visualize the data.
- Feature Engineering: Creating insightful features and transforming data.



Checking for Null values

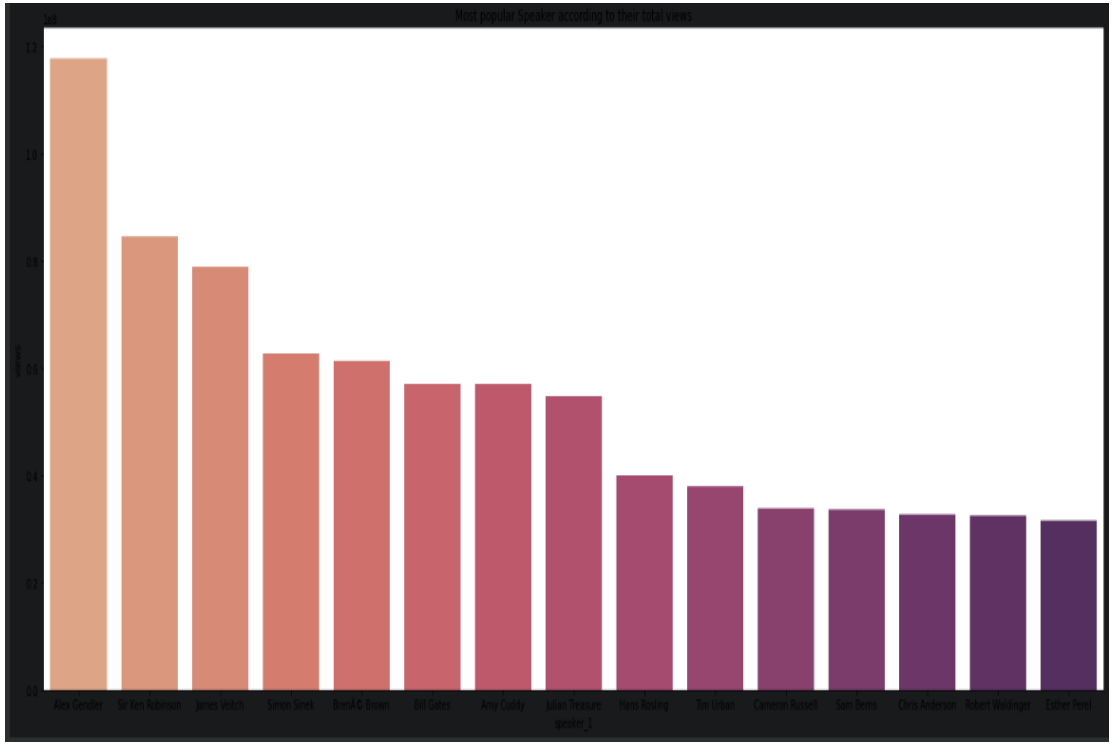
➢ Columns having Null Values

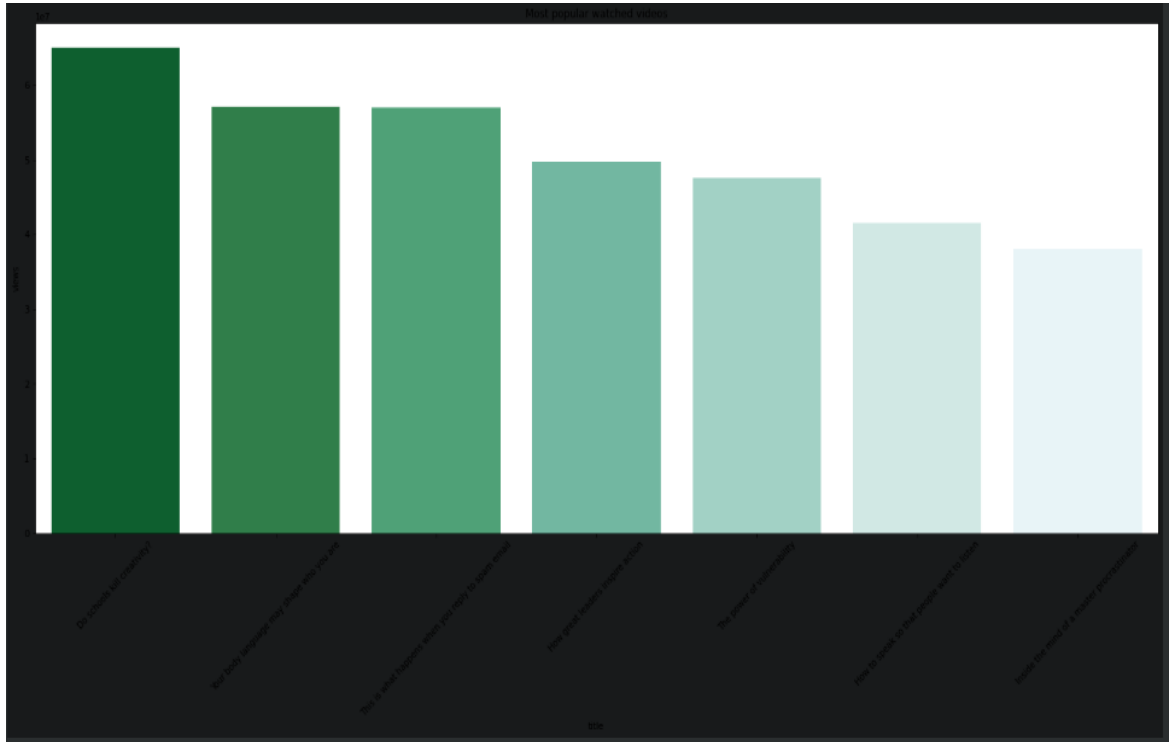
- Occupations
- About_Speakers
- Comments



Exploratory Data Analysis

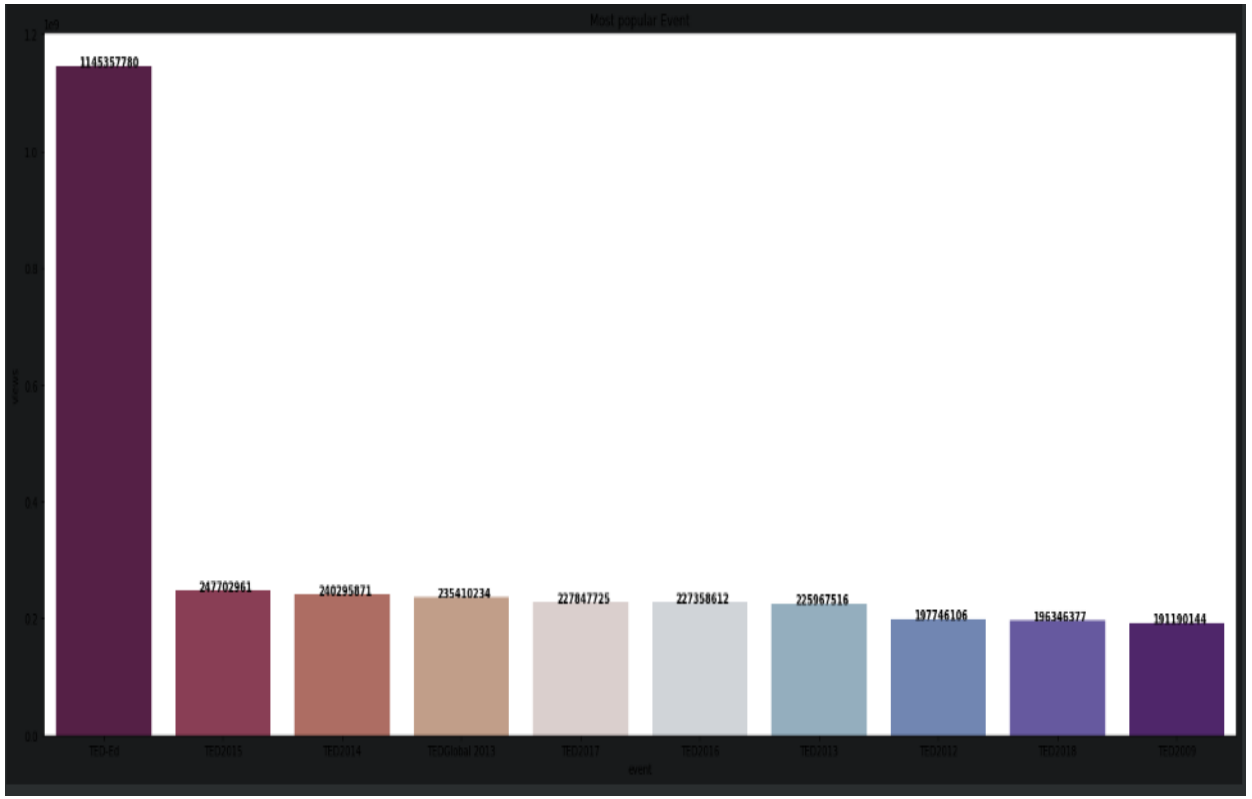
- Alex gendler is the most popular Ted Speaker.
- Do Schools kill creativity is the most watched video on Ted Platform.

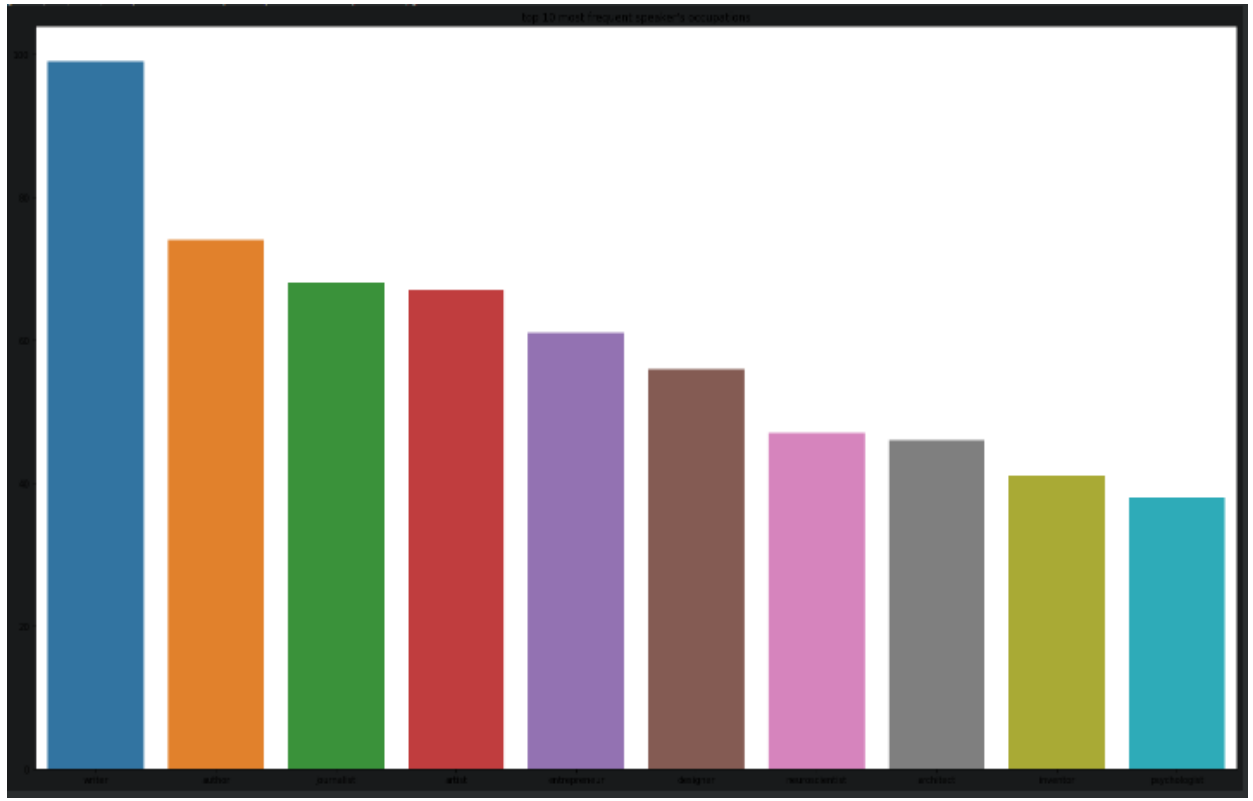




Exploratory Data Analysis

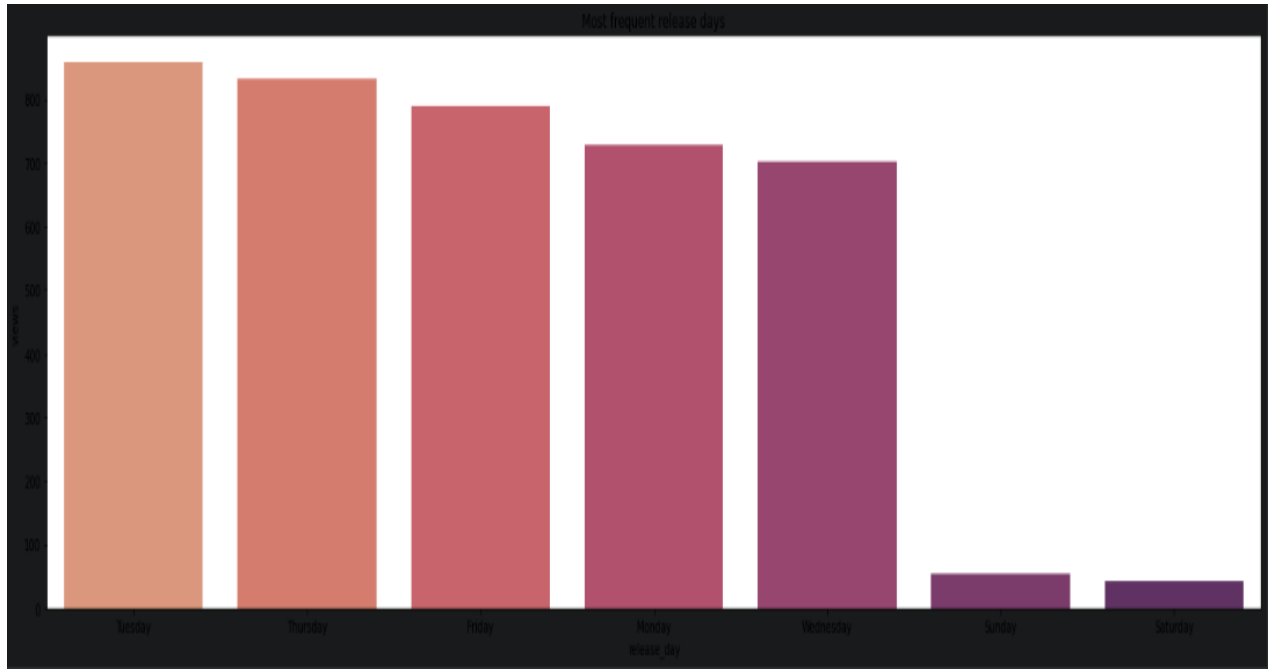
- TED-Ed is Most popular event on TED.
- Occupation of most frequent speaker is Writer followed by Author.

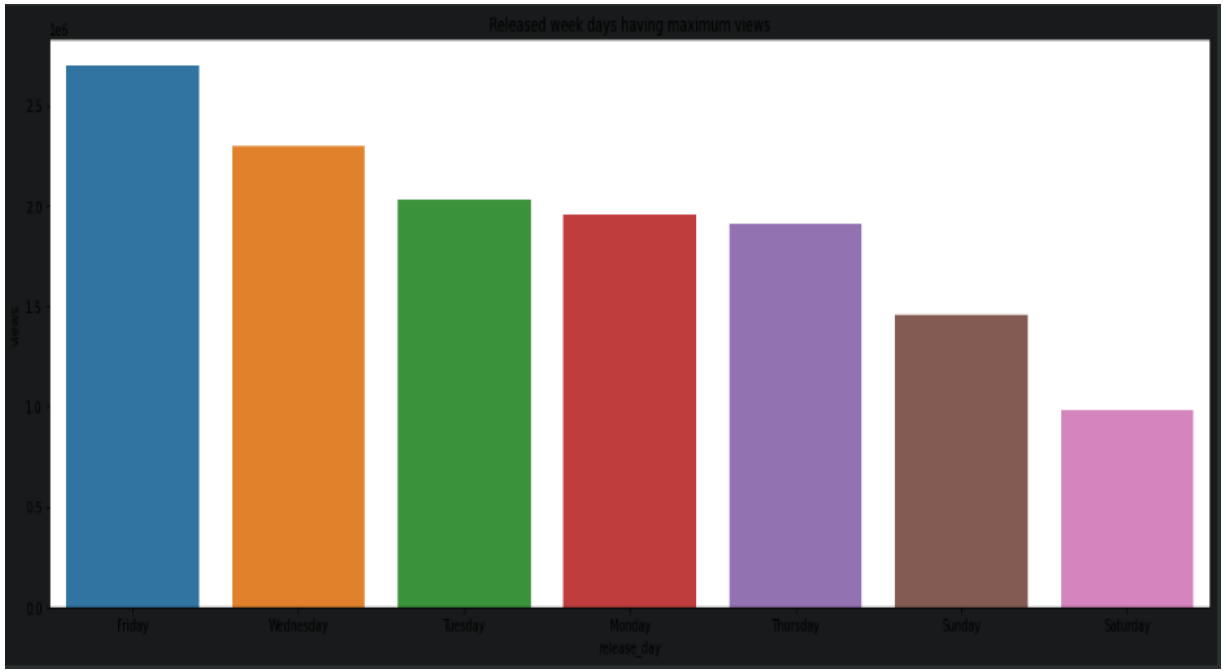




Exploratory Data Analysis

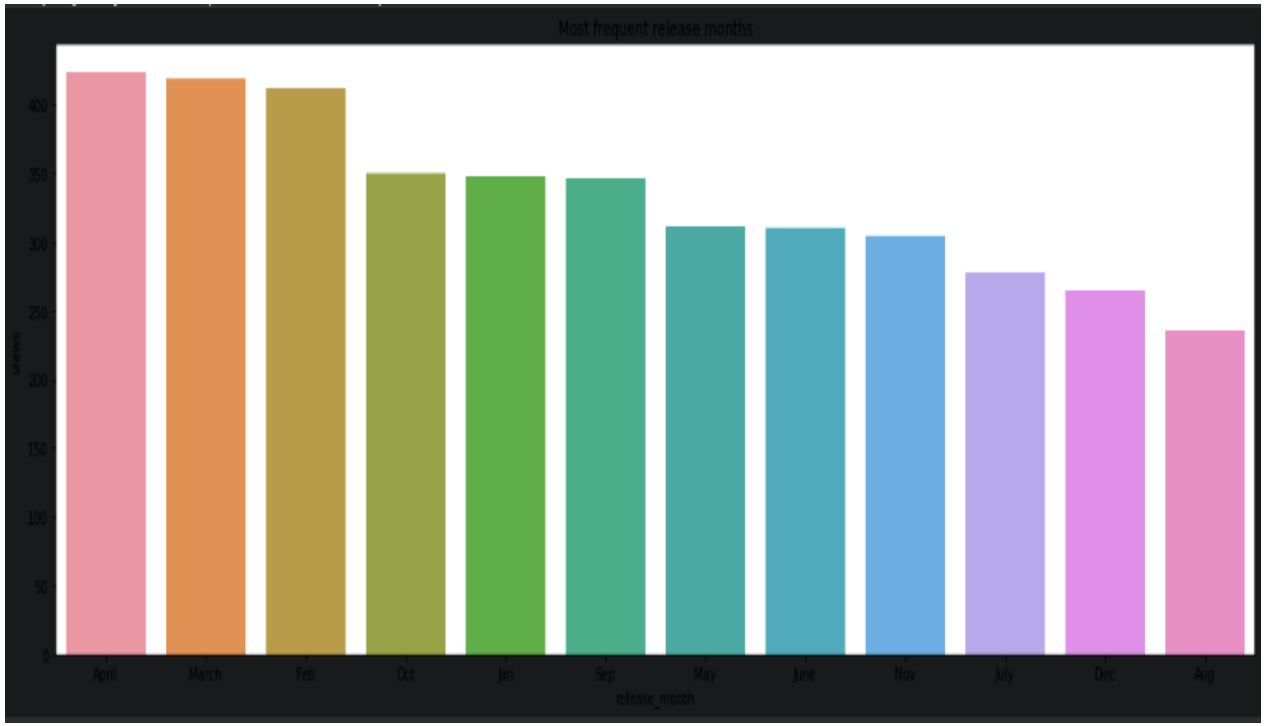
- Tuesday is the most frequent release day of the week followed by Thursday.
- Friday is the best release day of the week with respect to average views.

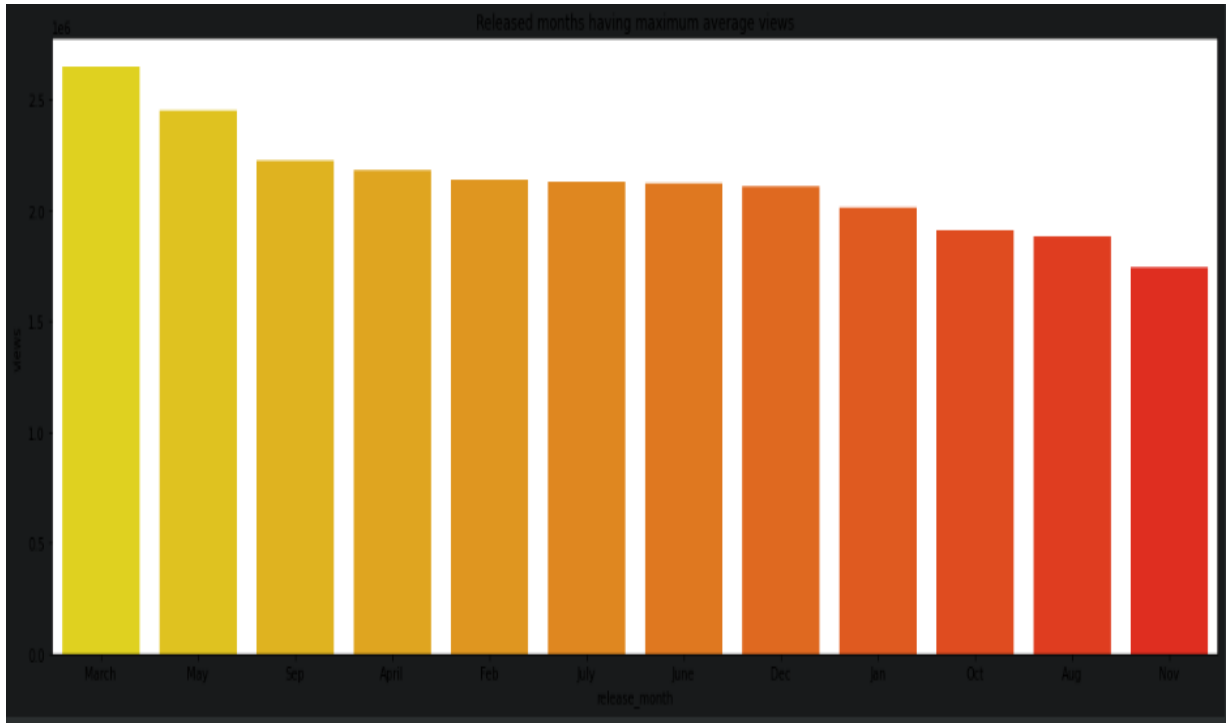




Exploratory Data Analysis

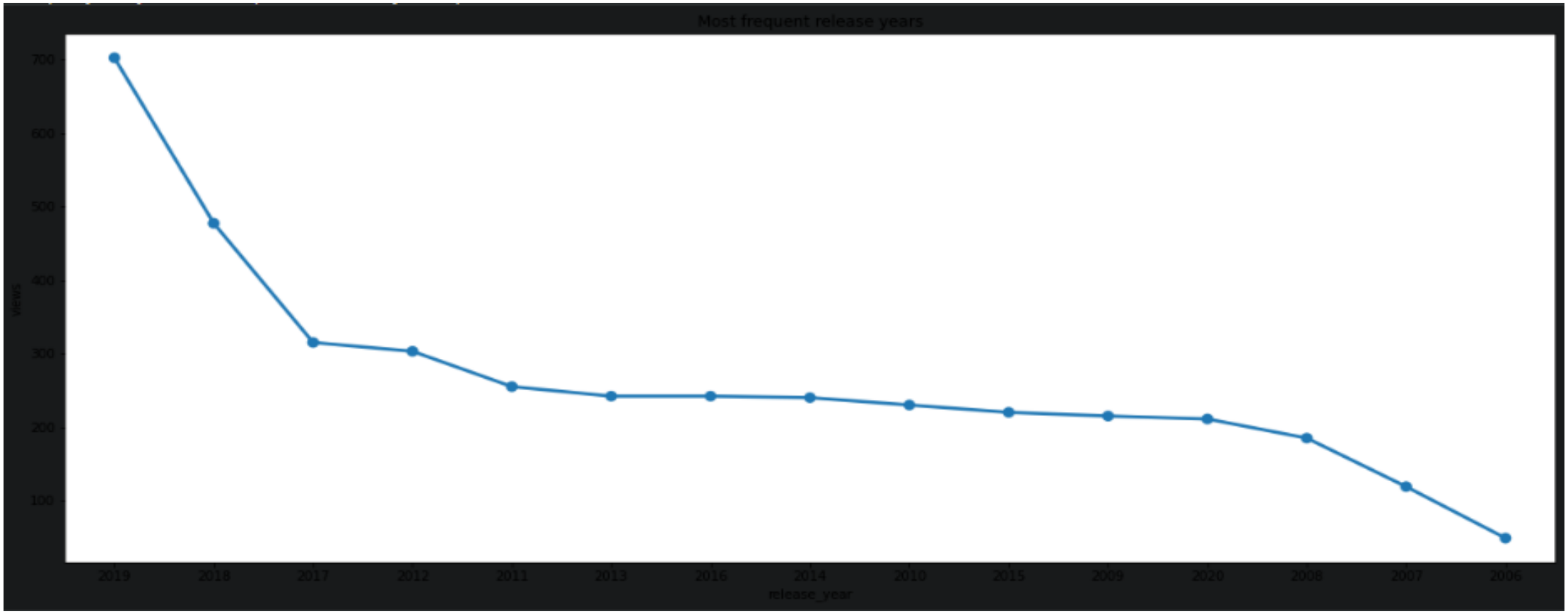
- April is the most frequent release month of the year.
- The month March has maximum average views followed by May.





Exploratory Data Analysis

- Most Videos are published in the year 2019 followed by 2018 and 2017.
- Least number in the year 2006.

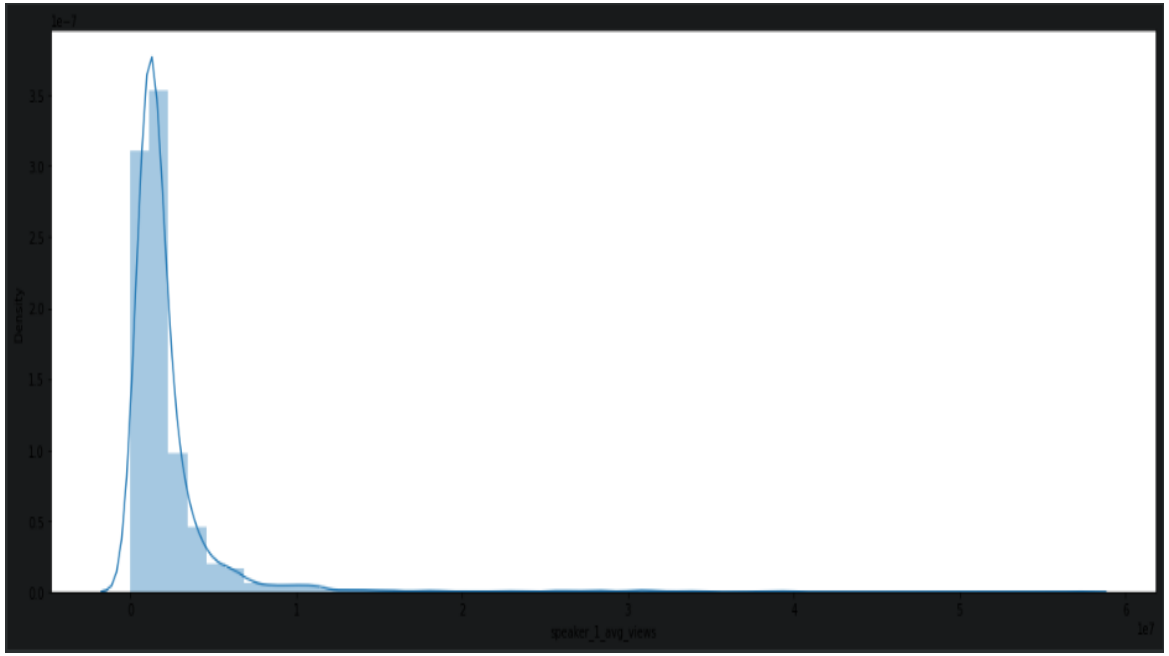


Feature Engineering

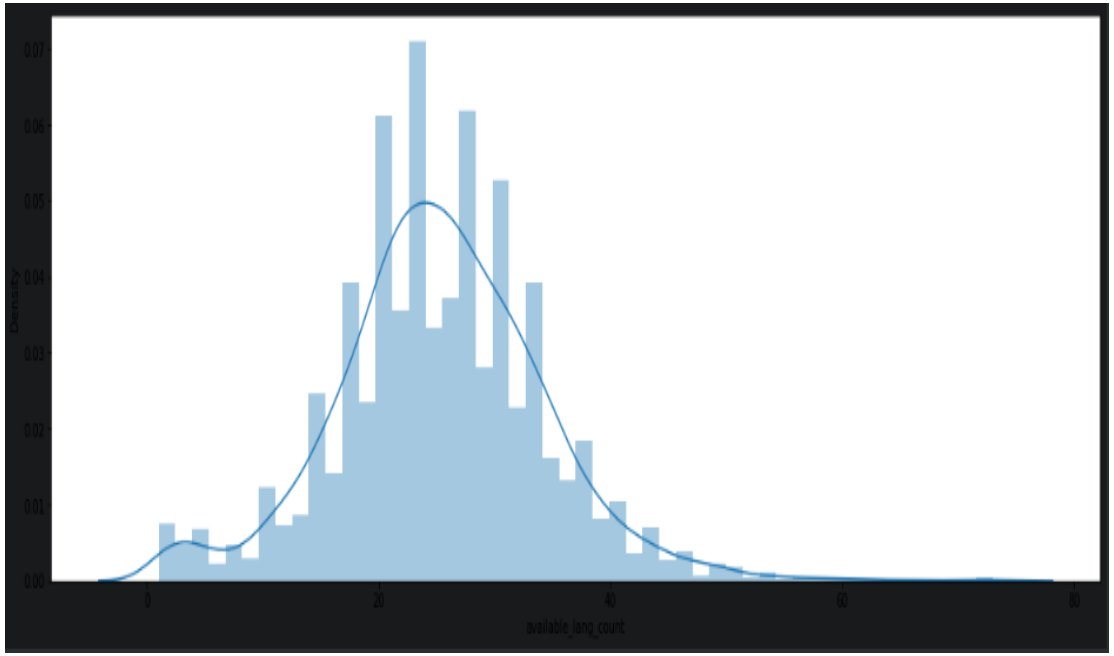
- Feature engineering is the process of using domain knowledge to extract features from raw data. The motivation is to use these extra features to improve the quality of results from a machine learning process, compared with supplying only the raw data to the machine learning process.
- We tried adding and removing some features in this section in order to make a perfect data matrix we can pass to a machine learning model. We will try to interpret categorical features as numeric to be passed to the ML models.

  - Speaker_avg_views
  - Event_wise_avg_views
  - Topic_wise_avg_views
  - Num_of_languages
  - Published_day

Feature Engineering



Feature Engineering on Speaker–1 Column and it is rightly Skewed.



As Skewness is less than 1 , it is normal distribution

Machine Learning Model Used

- Linear Regressor
- Random forest Regressor
- XGB Regressor

Metrics Used

- R2
- RMSE
- MAE

Hyper - Parameter tuning

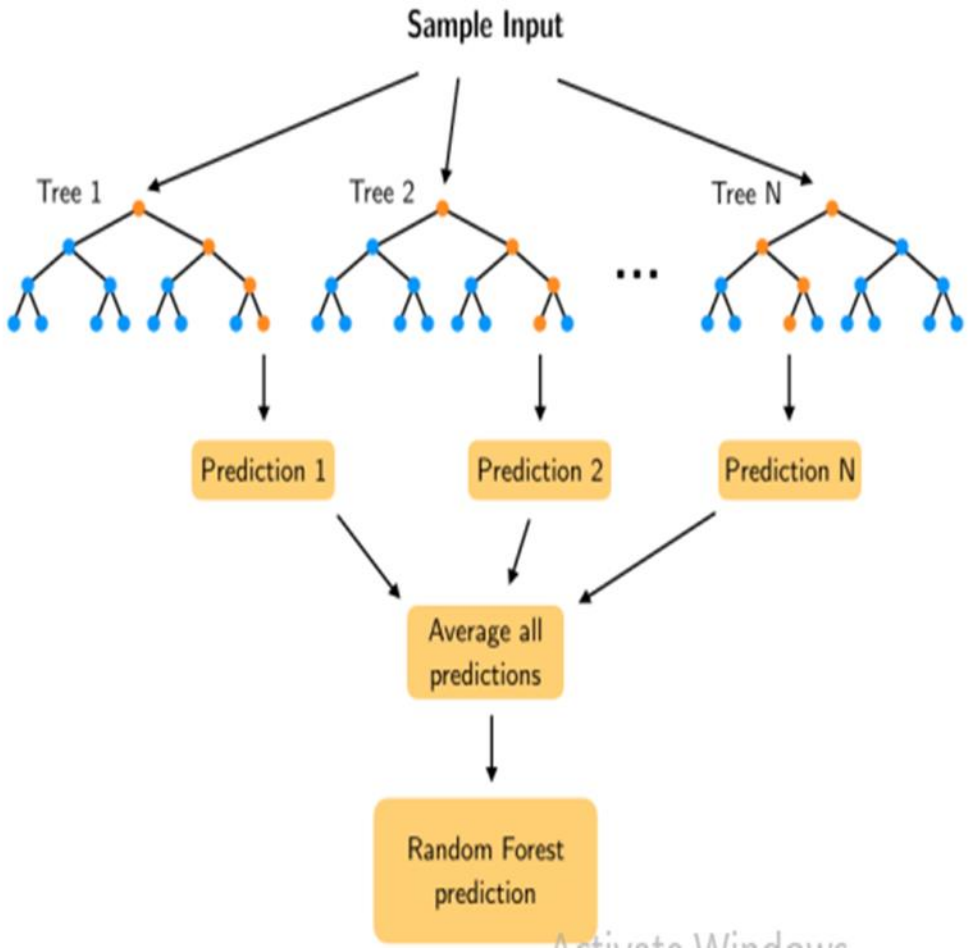
- Random Search CV

Linear Regressor

- R2 Score for Train – 0.8177
- R2 Score for Test – 0.8520
- MAE Score for Train – 271426.40
- MAE Score for Test – 243353.57
- RMSE Score for Train – 470463.24
- RMSE Score for Test – 424644.19

Random Forest Regressor

- R2 Score for Train – 0.8244
- R2 Score for Test – 0.8438
- MAE Score for Train – 185073.04
- MAE Score for Test – 168953.27
- RMSE Score for Train – 461787.90
- RMSE Score for Test – 436254.71



XGB Regressor

- R2 Score for Train – 0.9592
- R2 Score for Test – 0.8649
- MAE Score for Train – 119053.88
- MAE Score for Test – 207756.24
- RMSE Score for Train – 222622.28
- RMSE Score for Test – 405676.23

Which Model Did we Choose And Why?

- We choose **MAE** and not **RMSE** as the deciding factor of our model selection because of the following reasons:
- As we know, RMSE is more influenced by outliers MAE doesn't increase with outliers.
- MAE is linear and RMSE is quadratically increasing.
- So, We choosed MAE as a deciding factor for our model.
- On the basis of MAE, The best performing regression model is Rando m Forest Regressor.

CONCLUSION

- Through our Analysis, we have discovered key insights about what factors influence the Views gained by a video .
- Topics like Technology , Science , Education , Biology attract the attention of viewers more than other topics .
- Most TED Speaker’s Occupation is Writer, followed by Author.
- The Hyperparameter Tuning prevented overfitting and decreased errors by regularzing the models.
- The Random Forest Regressor performed the best in terms of Mean Absolute Error,As we choose MAE as our deciding factor because MAE is not affected by the Outliers.

Challenges and Future Scope

- Dataset have lots of textual and categorical data having high cardinality .
  - So the conversion to meaningful numerical data was a challenge.
- Feature Engineering and Feature Extraction
  - can always be improved in creative ways
  - We can explore more advanced feature encoding

Thank You