# Capstone Project
## Ted talk view prediction


By-

PUSHKAR SRIVASTAVA
RAHUL PANDEY
HRITIK SHARMA

# Point of discussion

- Introduction
- Problem statement
- Cleaning of data
- Variables for daily views
- Bivariate analysis with dependent variable
- Target Encoding
- Feature Engineering and Data preprocessing
- Outliers Detection
- Removing collinearity
- Variance inflation factor analysis
- Let’s Check Normality in data
- Model Preparation

# Point of discussion

- Error metrics
- Running grid search cross validation for lasso regression
- Running grid search cross validation for ridge regression
- Running grid search cross validation for elastic regression
- Conclusion

# INTRODUCTION

TED talks have been given for many years with the platform of "Ideas Worth Spreading". In the digital world we live in today, TED is a great platform to get your idea out there. But how do you know if your idea will be heard or appreciated? We aim to perform a comprehensive analysis of over 2500 TED talks to determine what it is that makes an idea powerful.

These datasets contain over 4,000 TED talks including transcripts in many languages

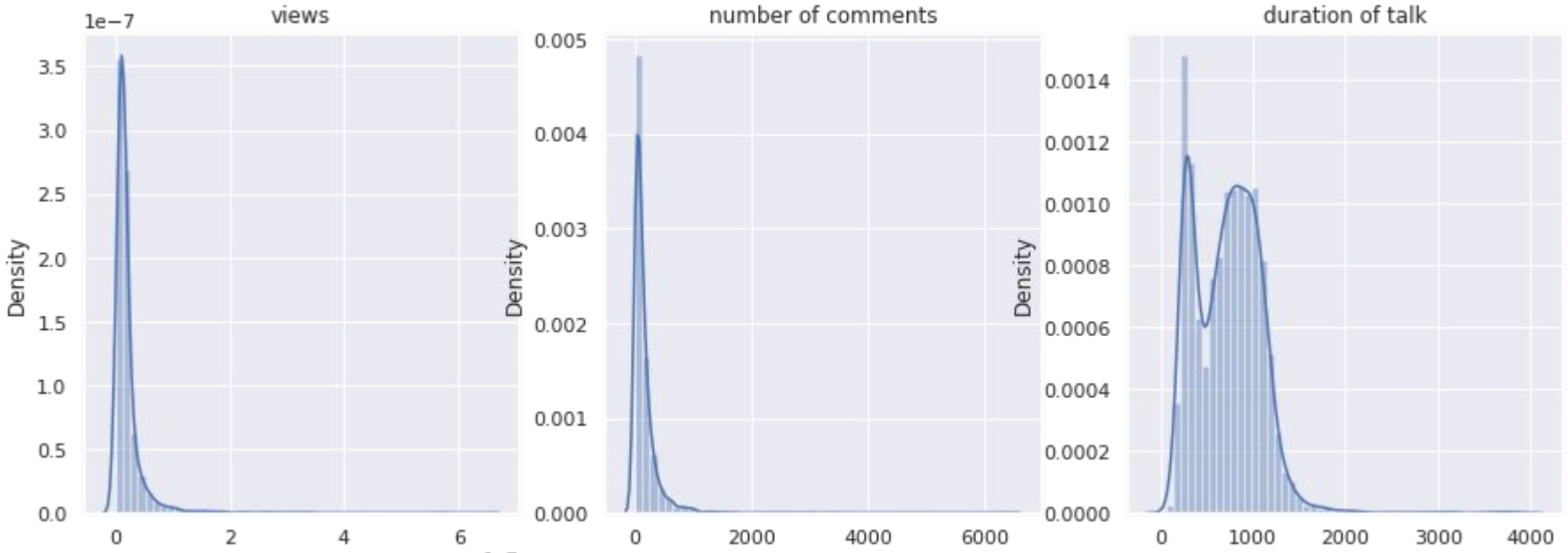
# PROBLEM STATEMENT

The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website."

# Exploratory Data Analysis

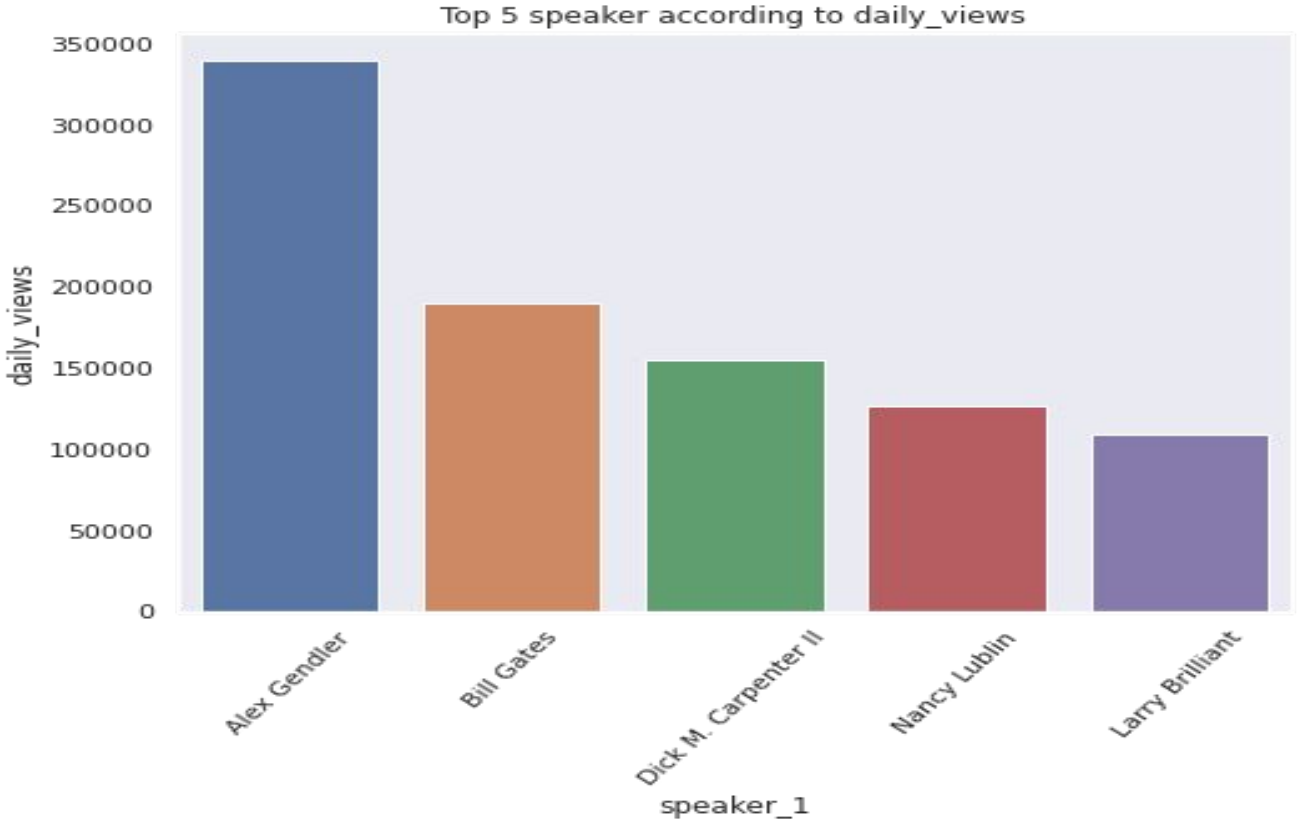
## Univariate analysis

Univariate analysis is the simplest form of analyzing data.



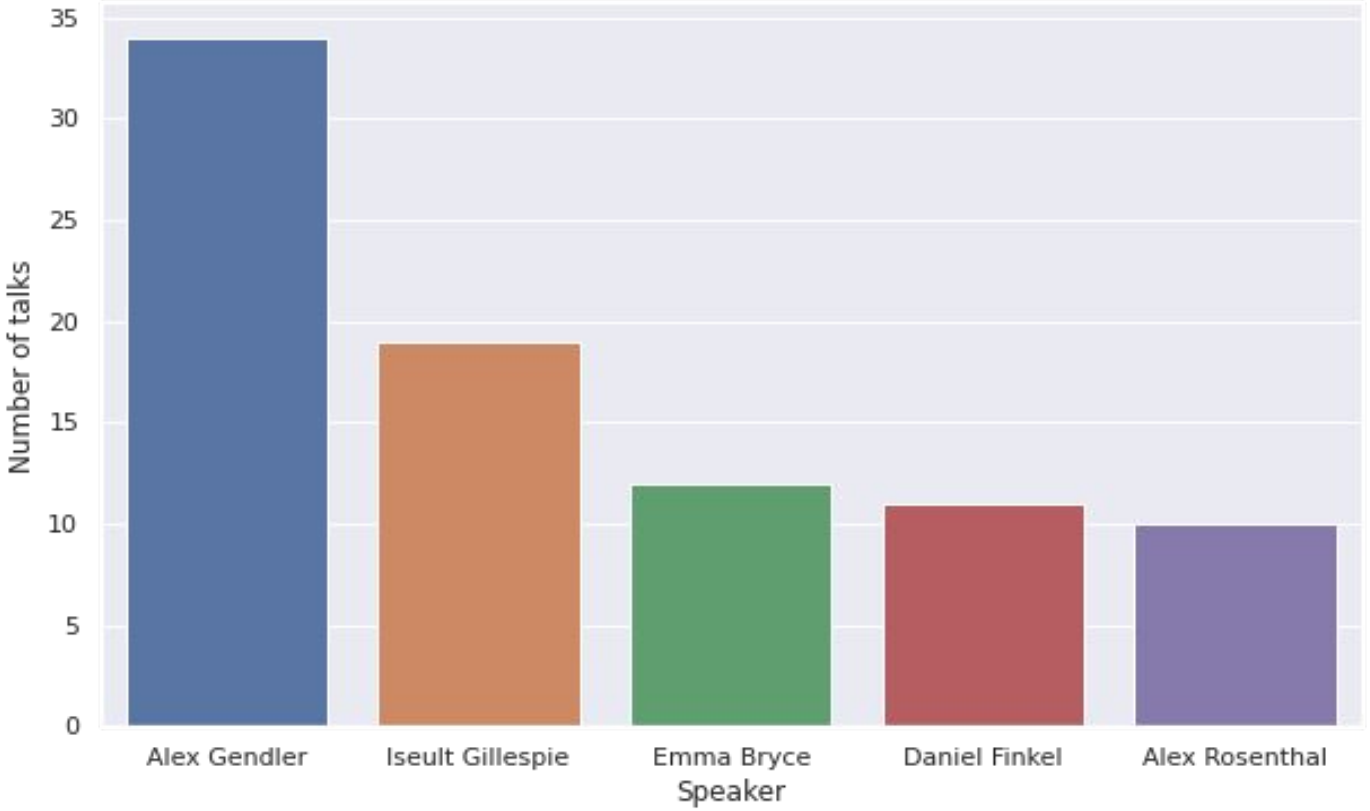
# Bivariate analysis with dependent variable

## speaker_1 vs daily_views



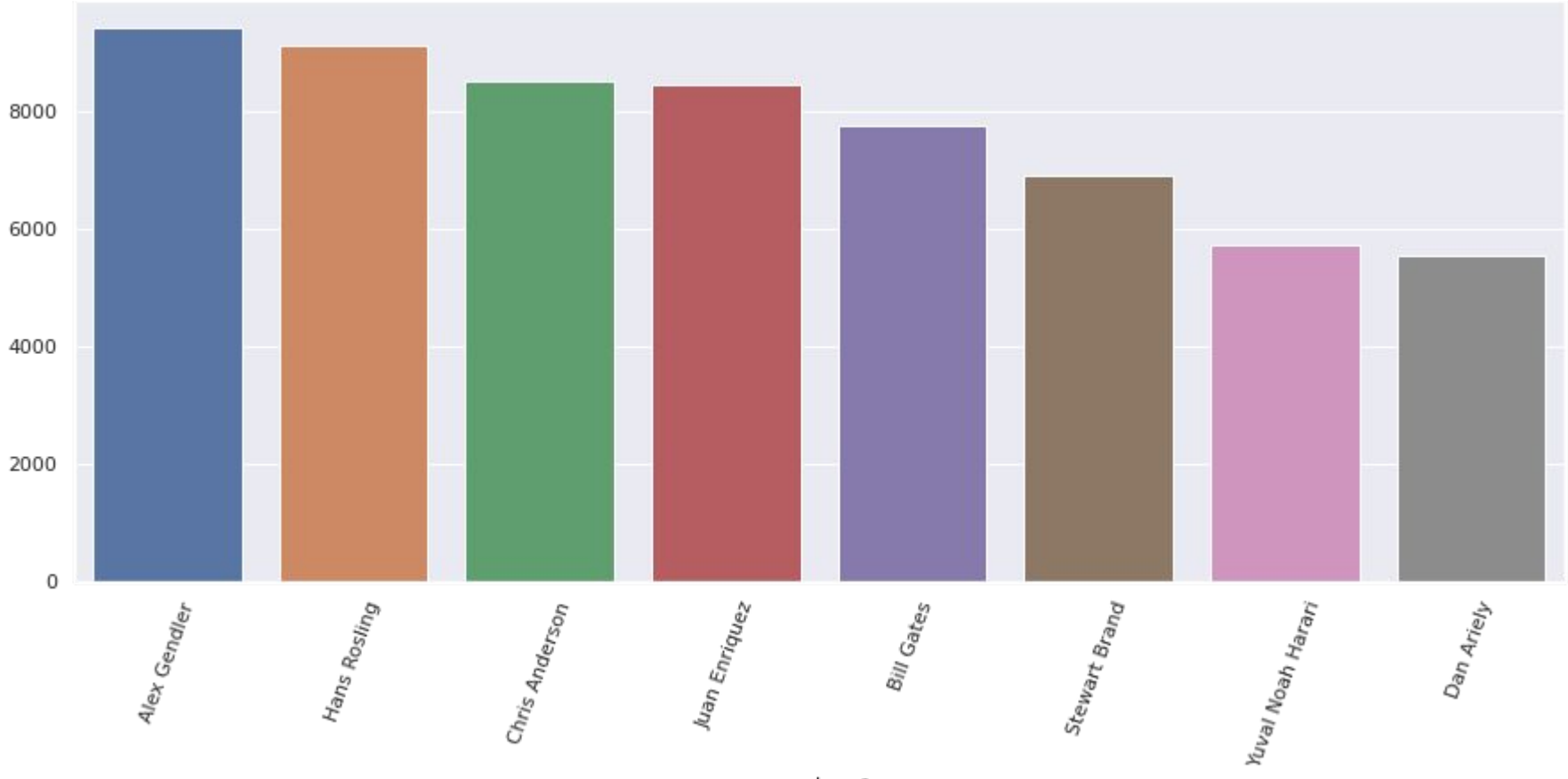

- **Ted Talk by Alex Gendler has the highest daily views followed by Bill Gates.**
- **Here it seems the daily views does depend on the first speaker.**

# Speaker Vs Number of talks delivered

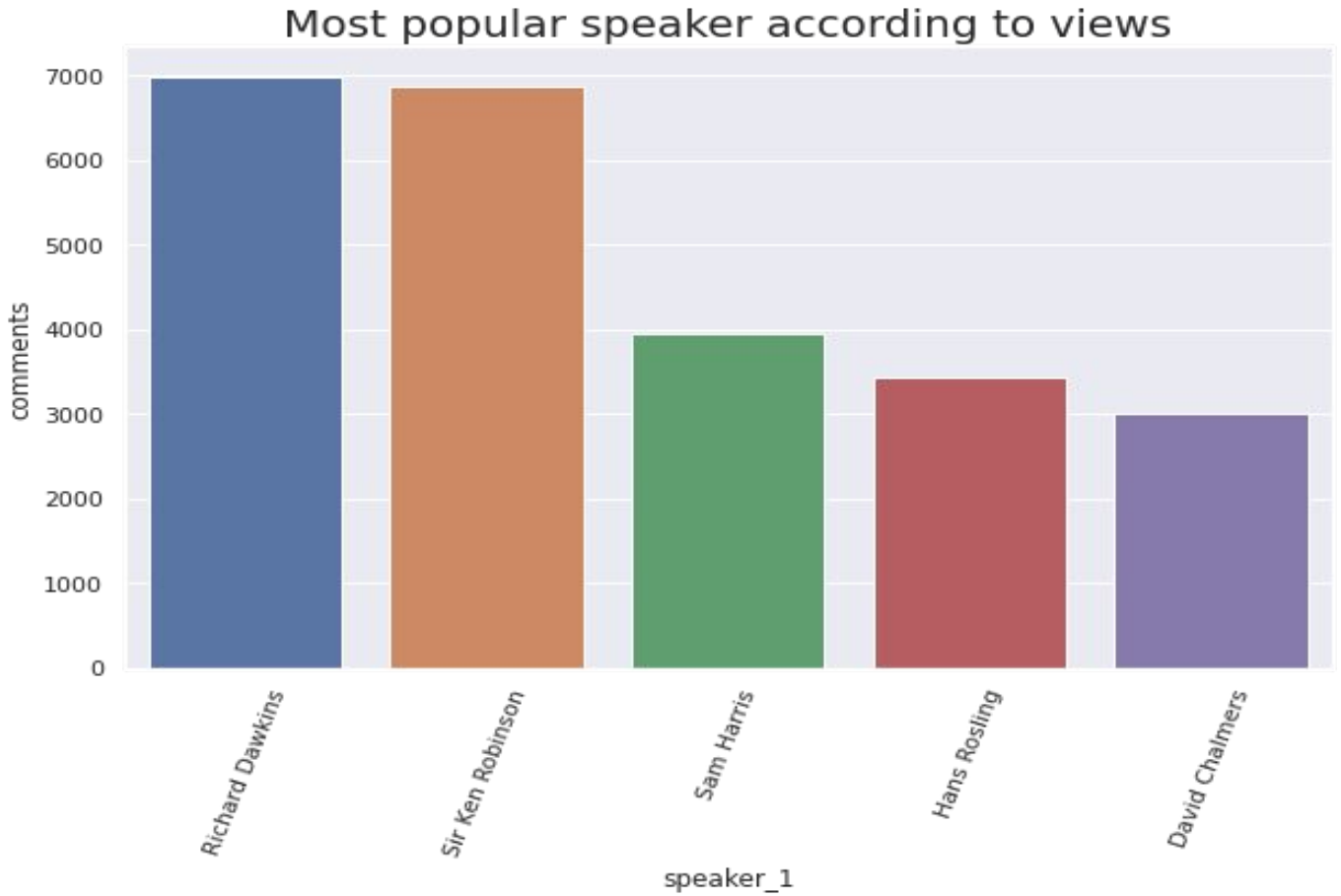



- **Alex Gendler also has highest number of talks that could explain such high overall views.**

# speaker vs duration

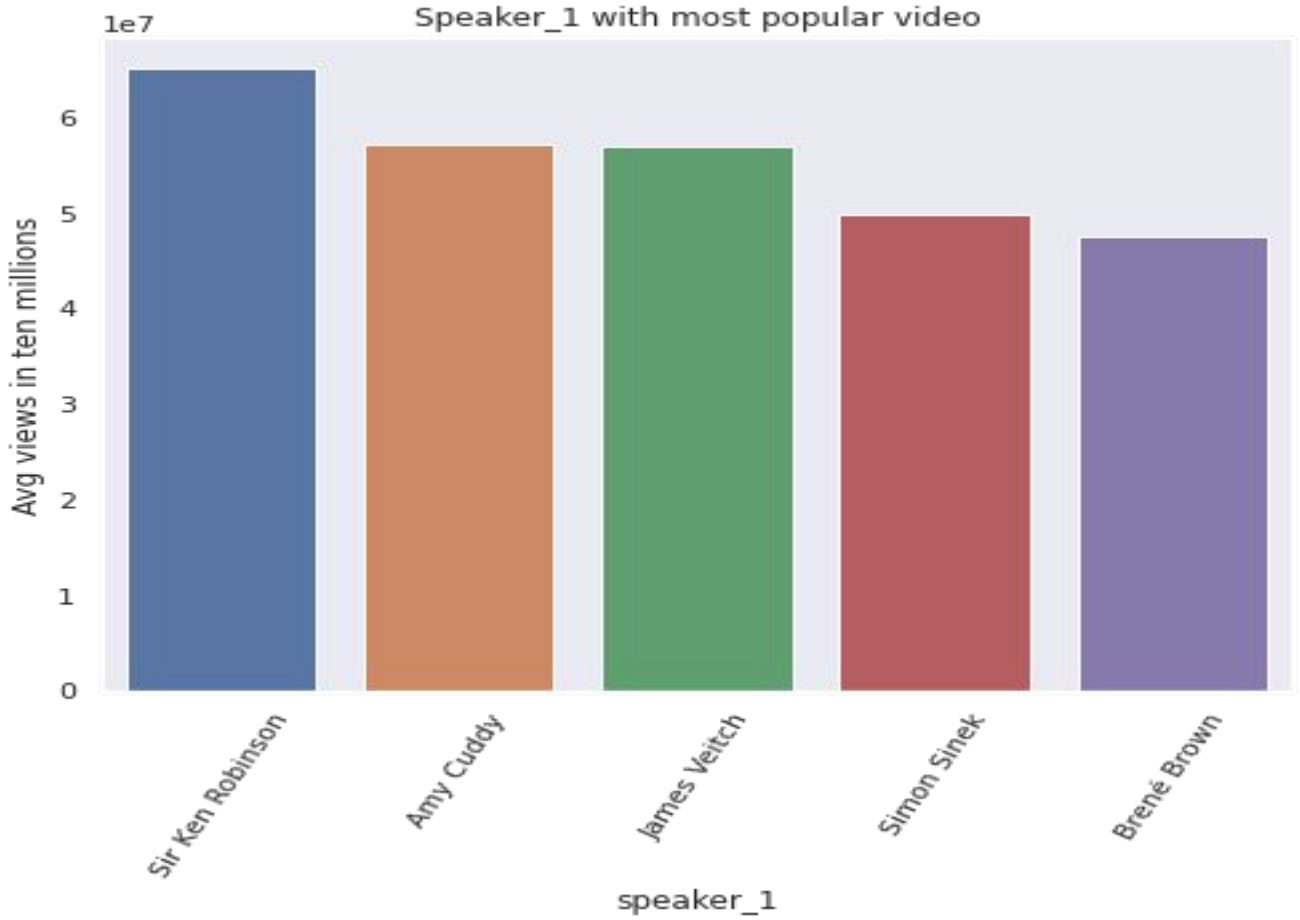


# Speaker vs comments




- **Richard Dawkins has highest number of comments followed by Sir Ken Robinson**

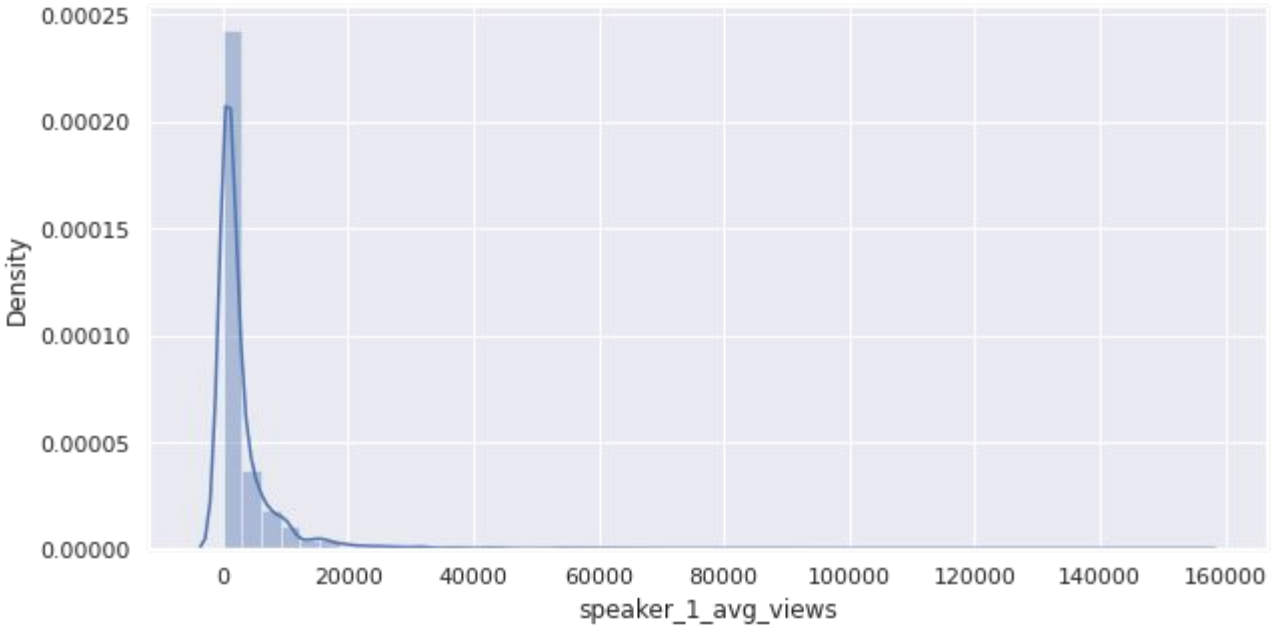
# Speaker vs Average Views



# Target Encoding

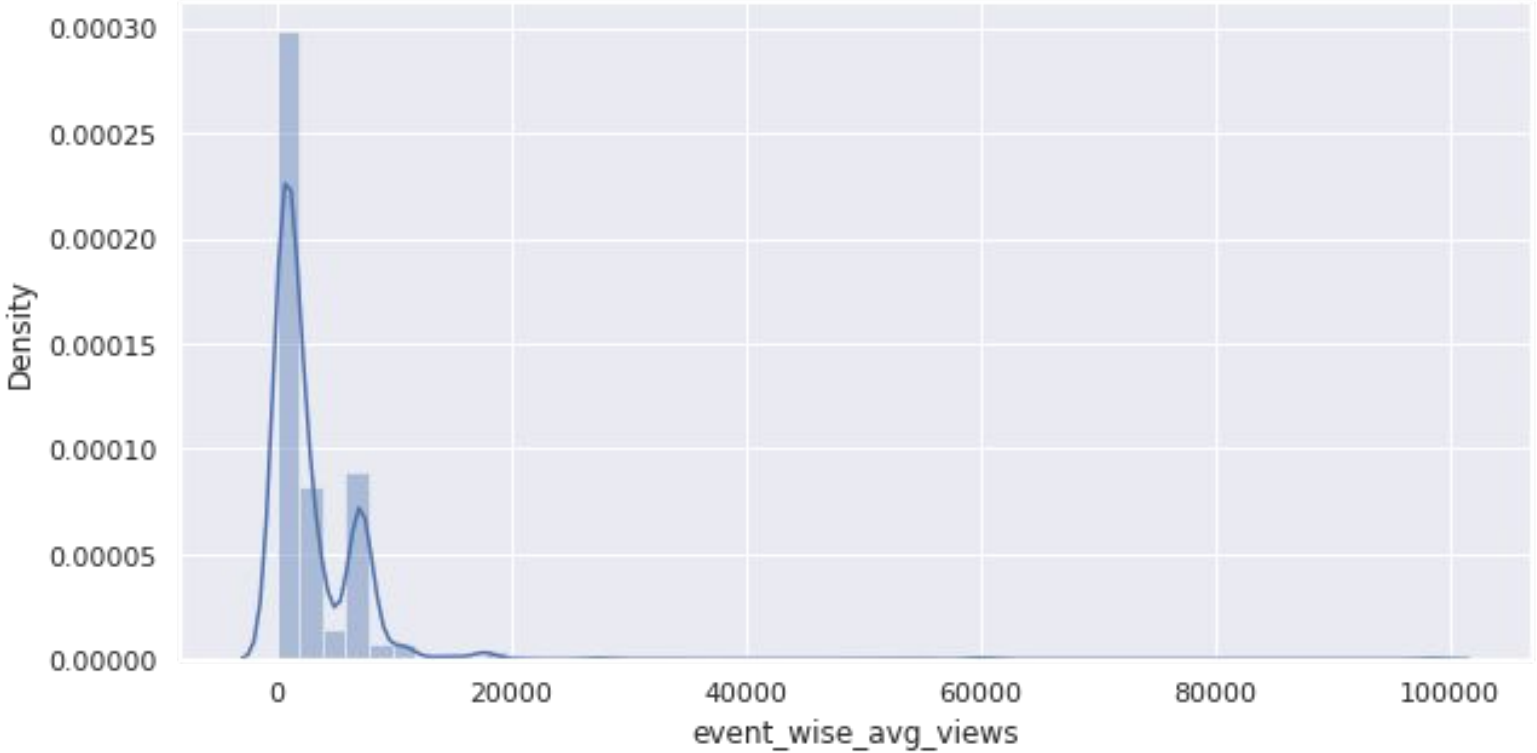
**Target encoding** is the process of replacing a categorical variable values with the mean of the target (dependent variable) variable

## Applying Target encoding on speaker_1

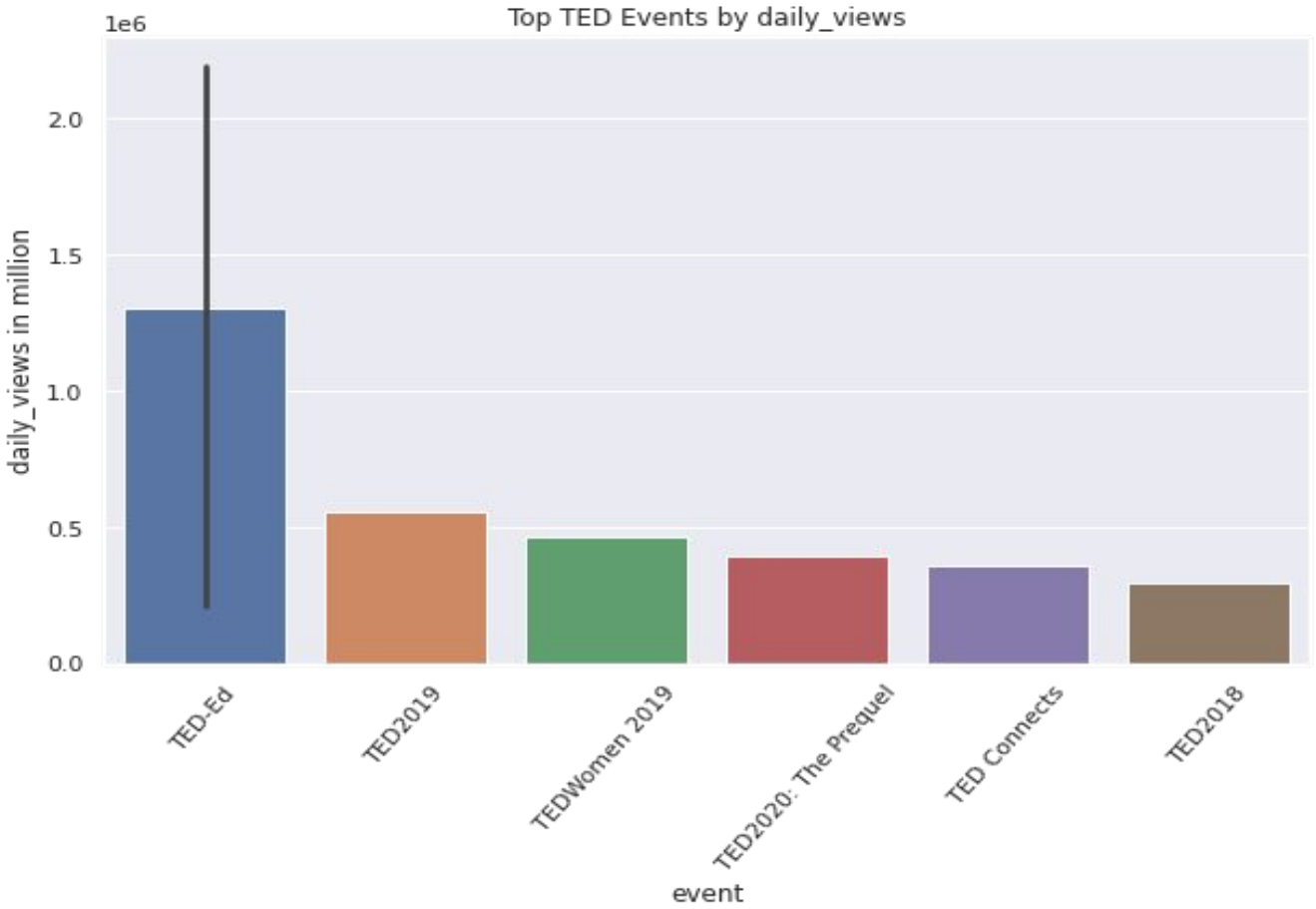


# Event

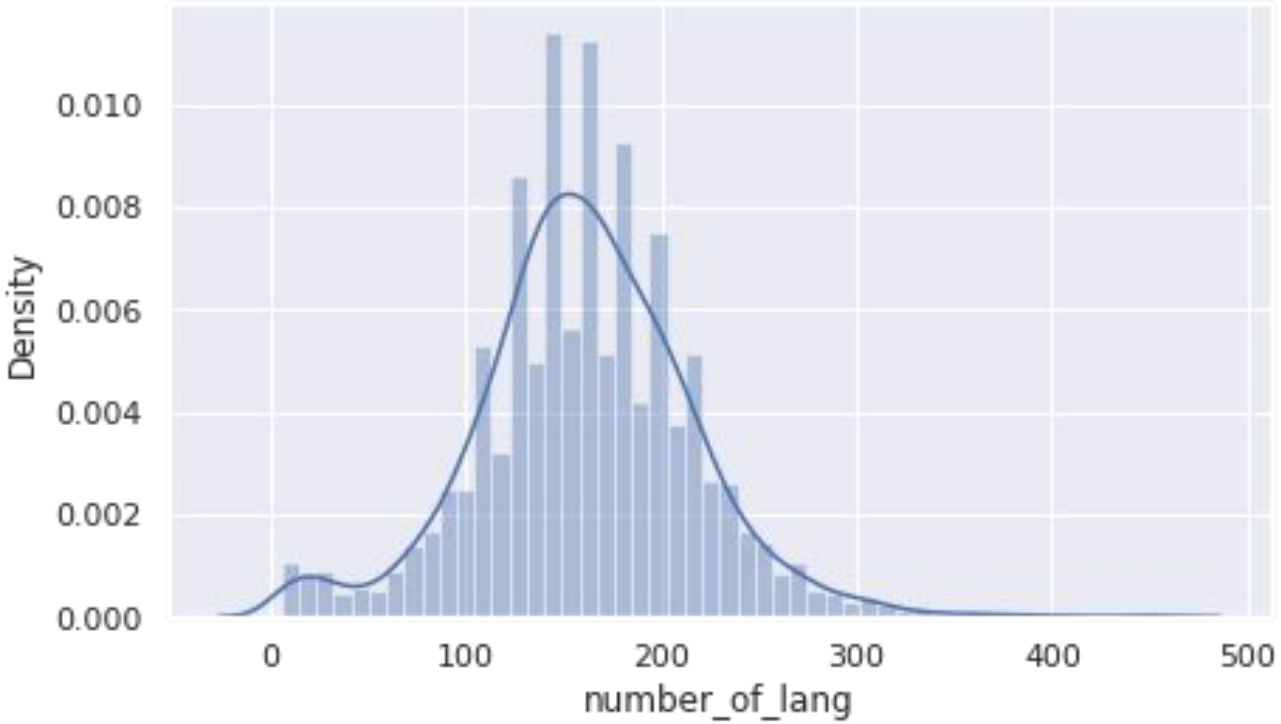
Event is also a categorical variable, therefore we also apply target encoding on it



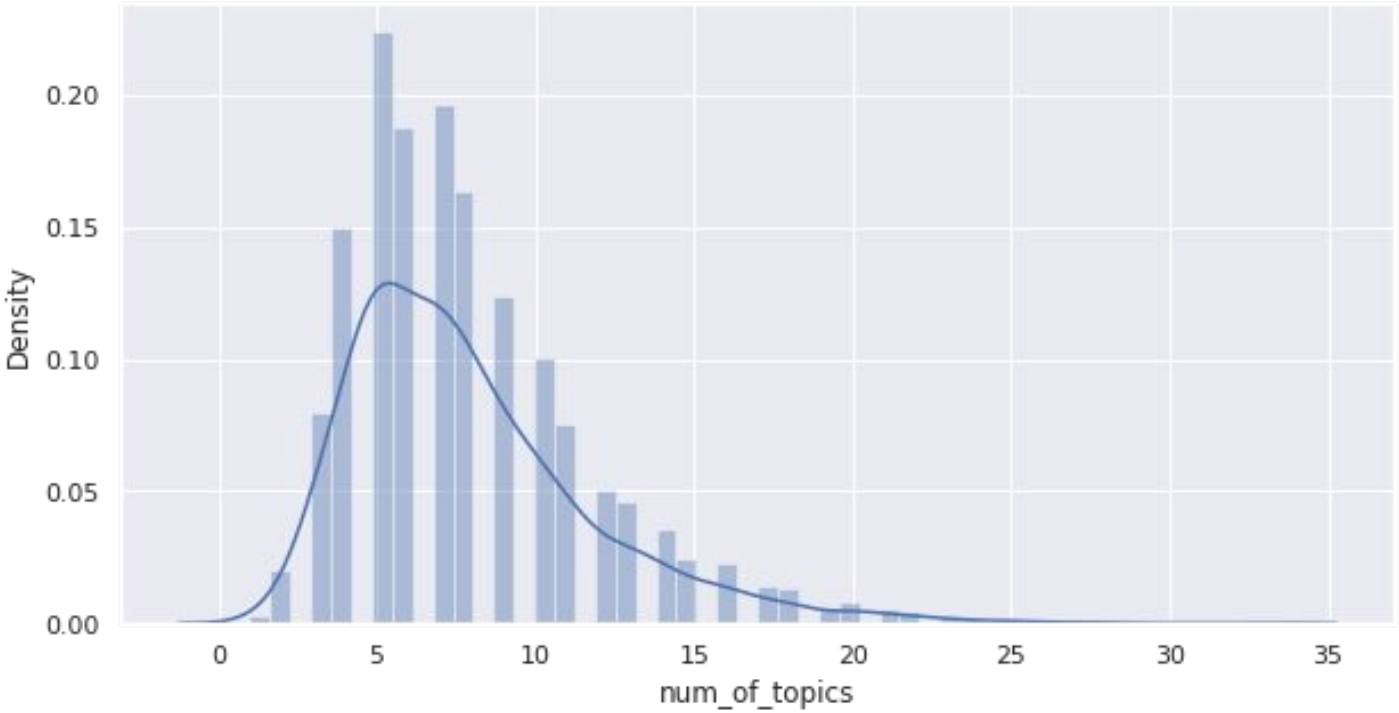
# Top ted talk event



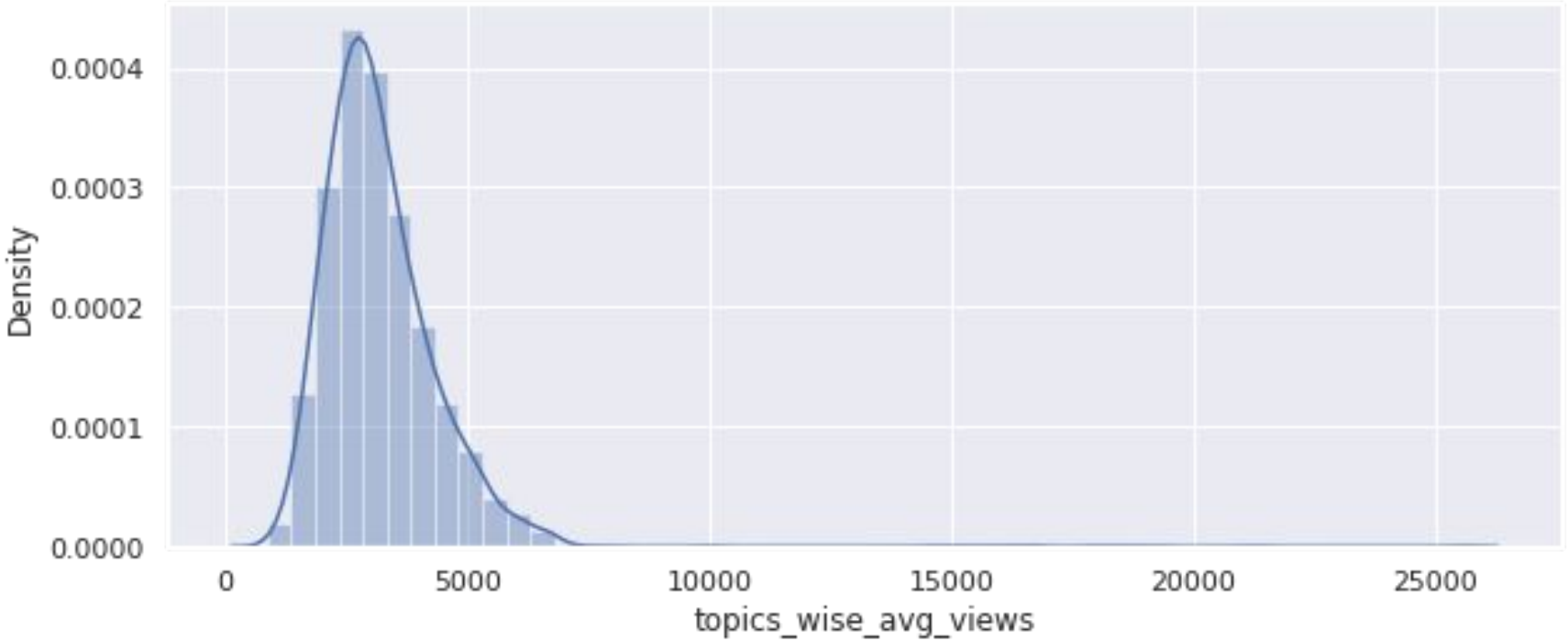
# available_language variable



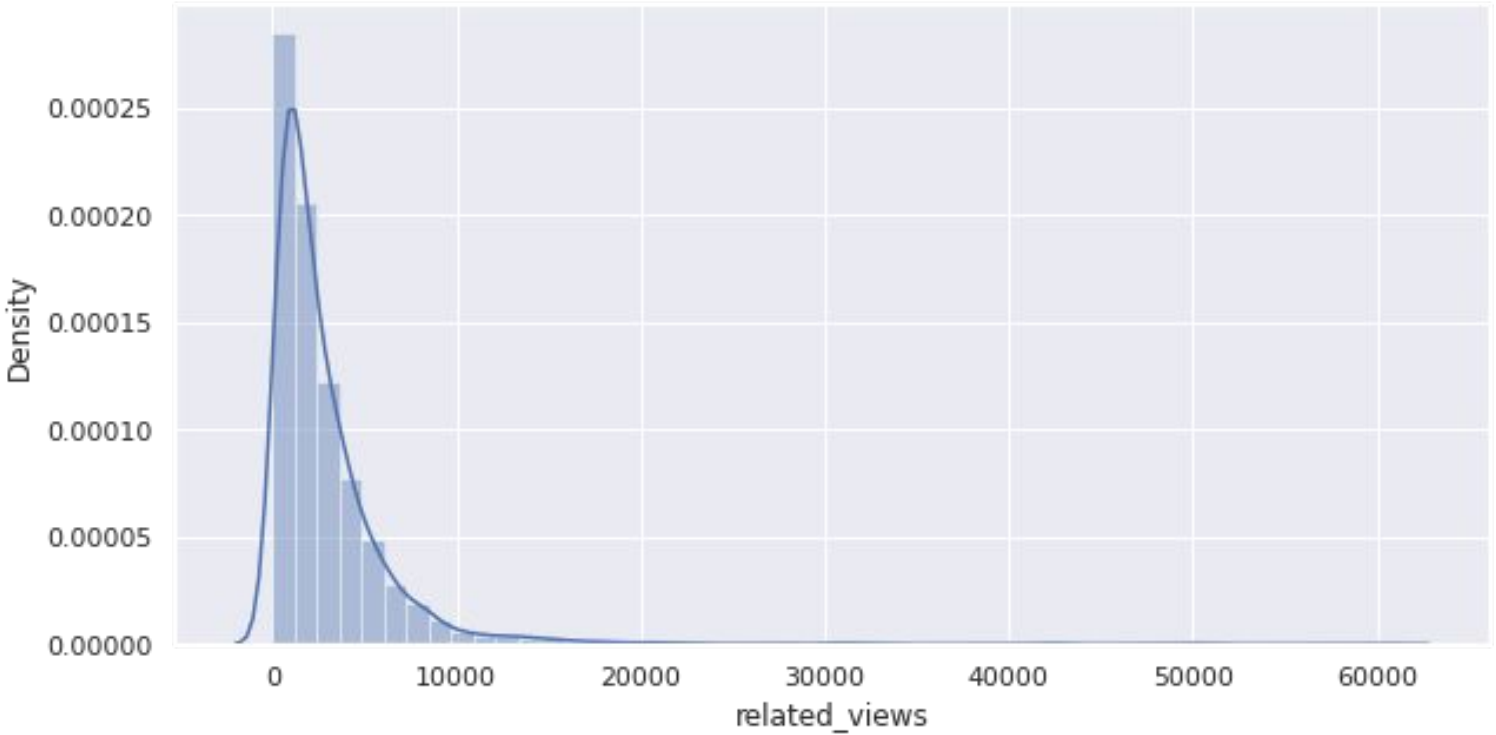
# num_of_topic variable from topic variable



# Target coding on unique topics

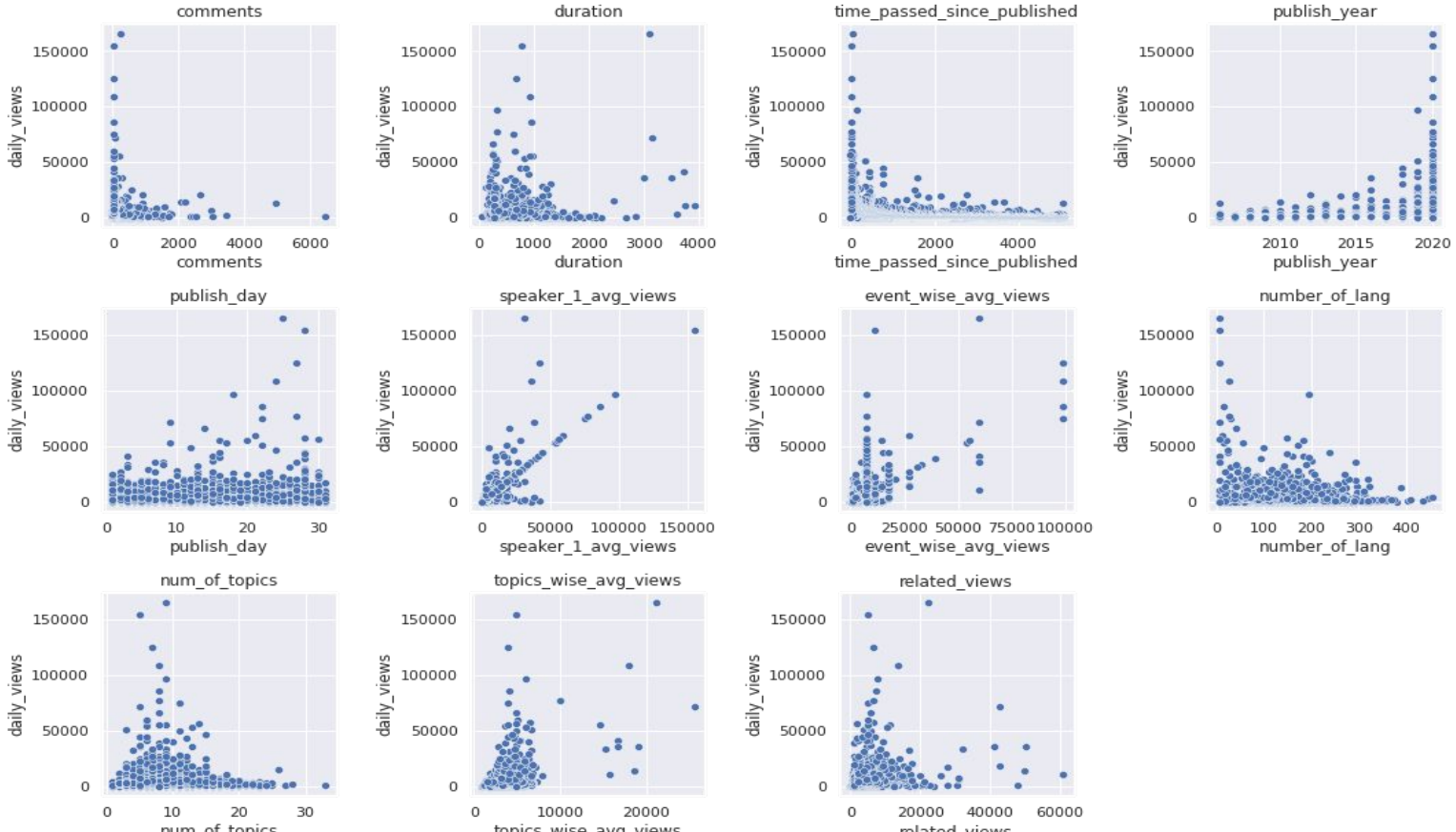


# Related talk variable

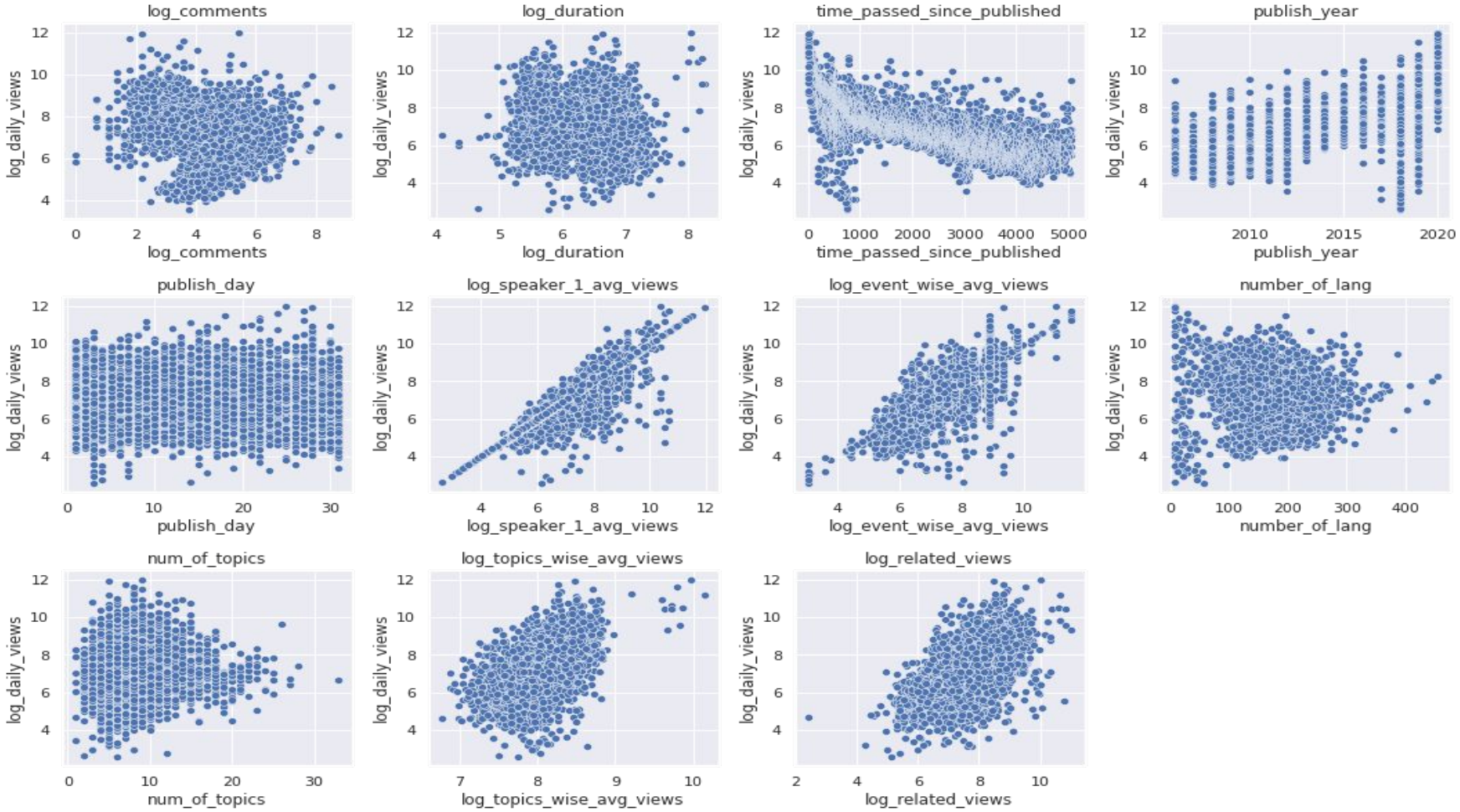


# Feature Engineering and Data Preprocessing

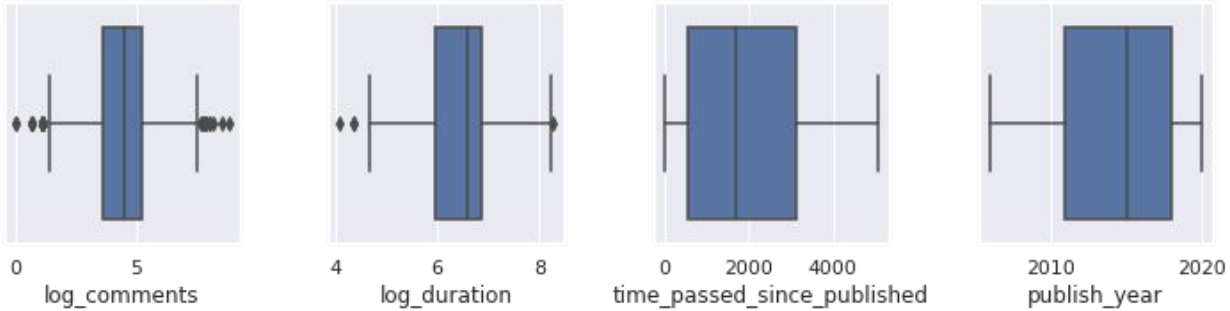
## Verifying OLS assumptions  Linearity



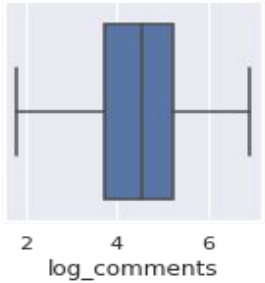
# Transformation for Linearity

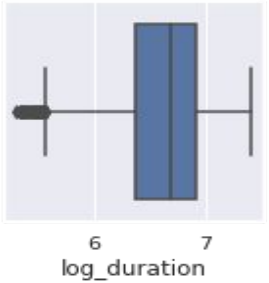


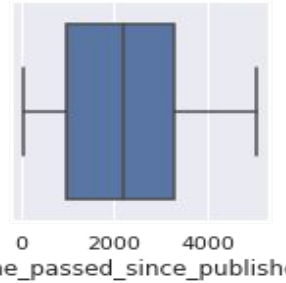
# Outliers Detection (Before)

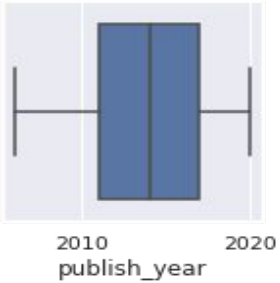


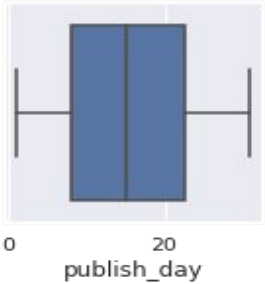
# Outliers Detection (After)

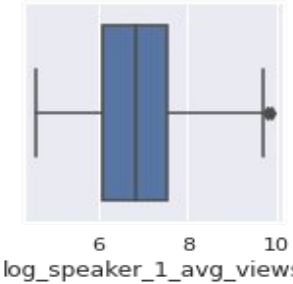


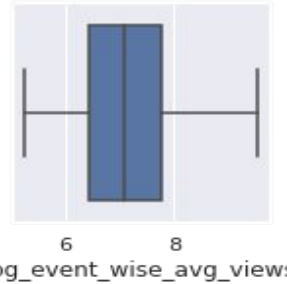



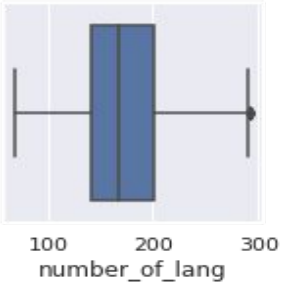



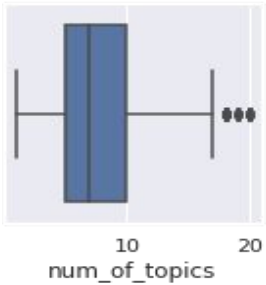



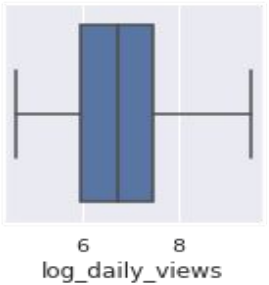



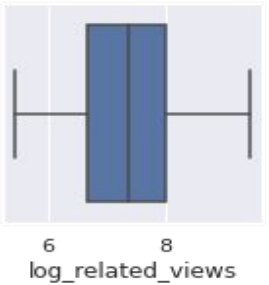



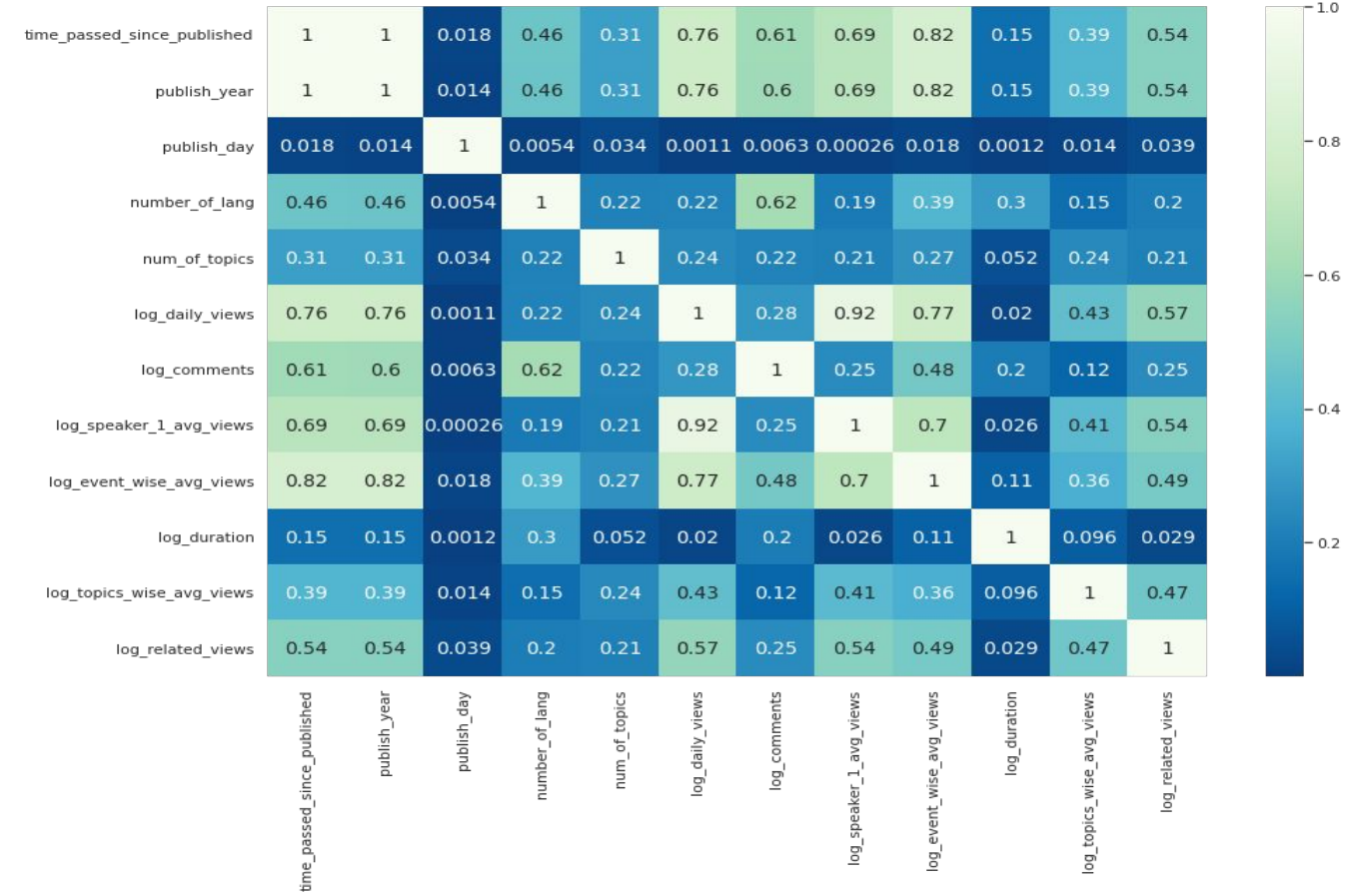







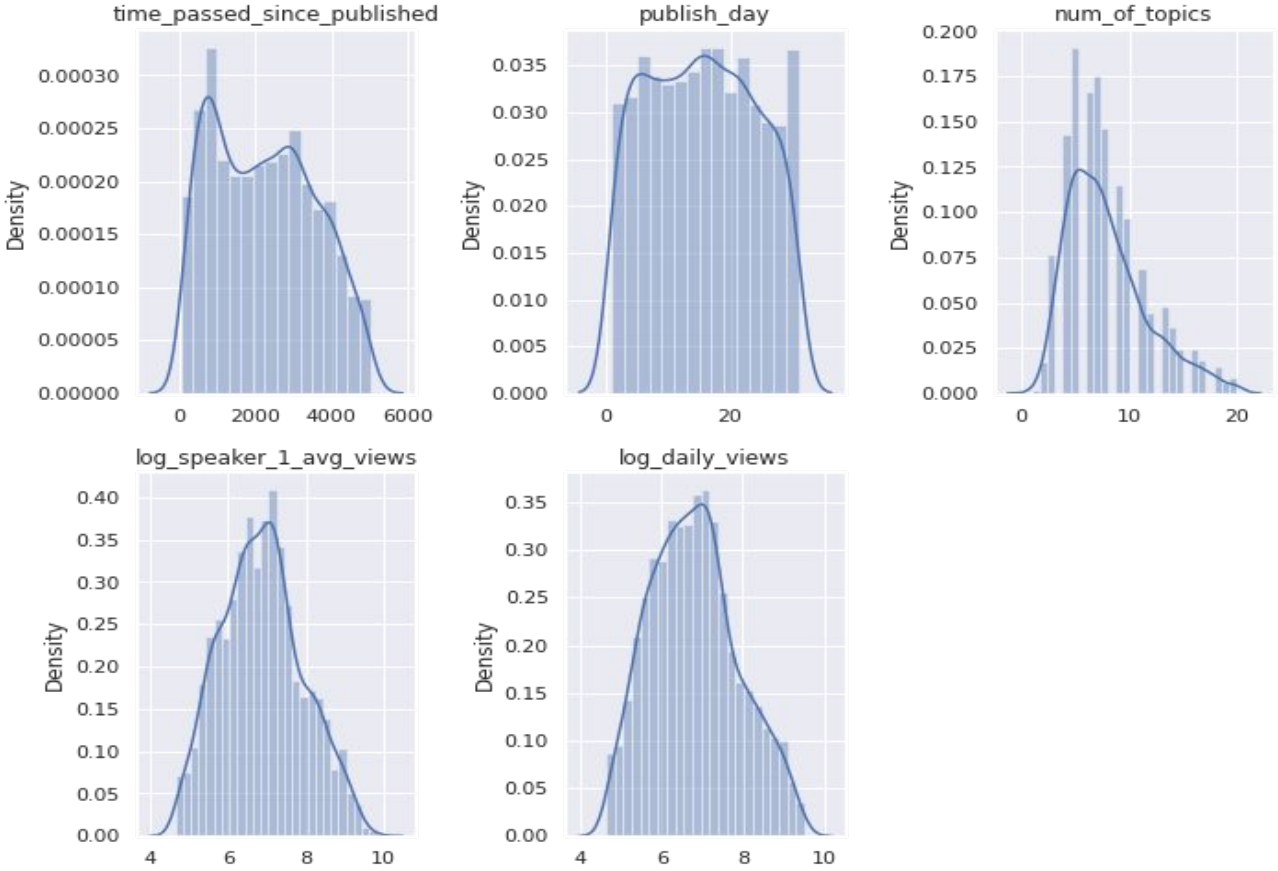
# Removing collinearity



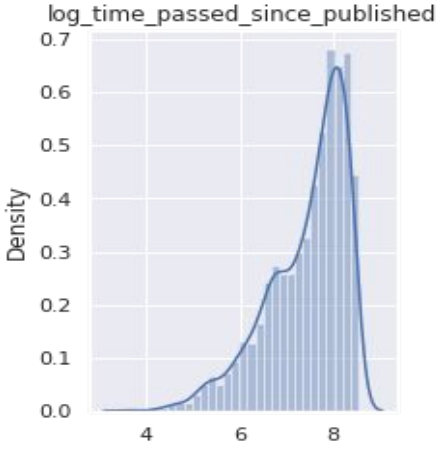

**After removing collinearity from the features we were left with features:**

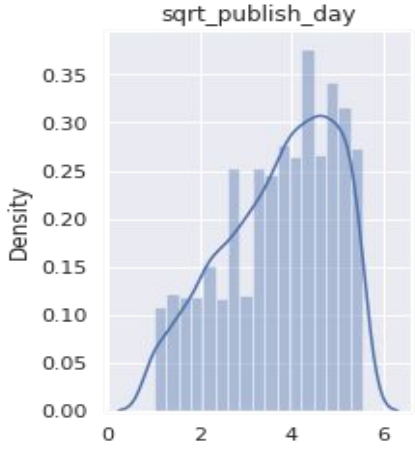
- **time_passed_since_published**
- **publish_day**
- **num_of_topics**
- **log_speaker_1_avg_views**

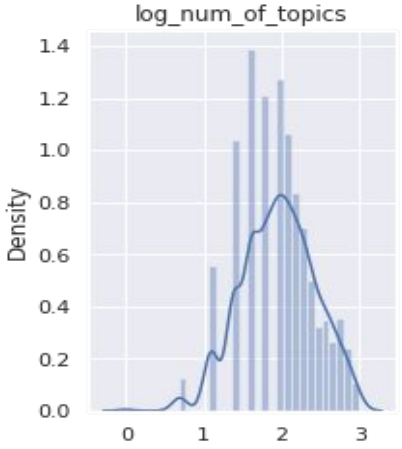
# Normal distribution of features in data

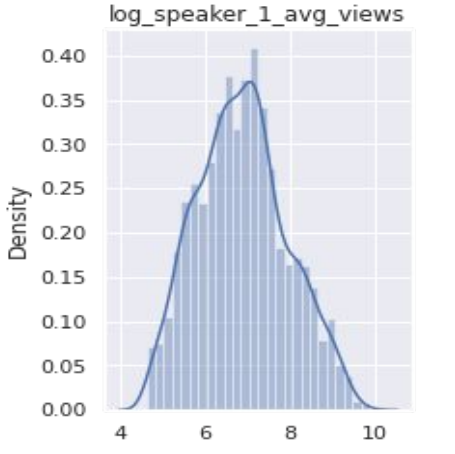


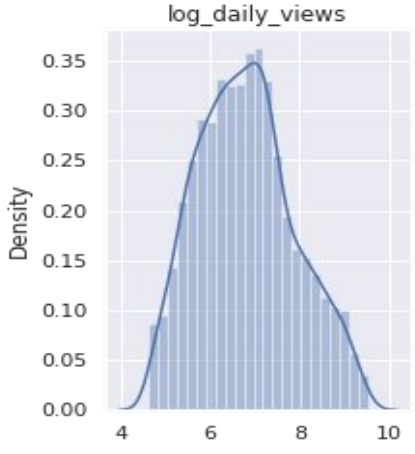
# Transformation



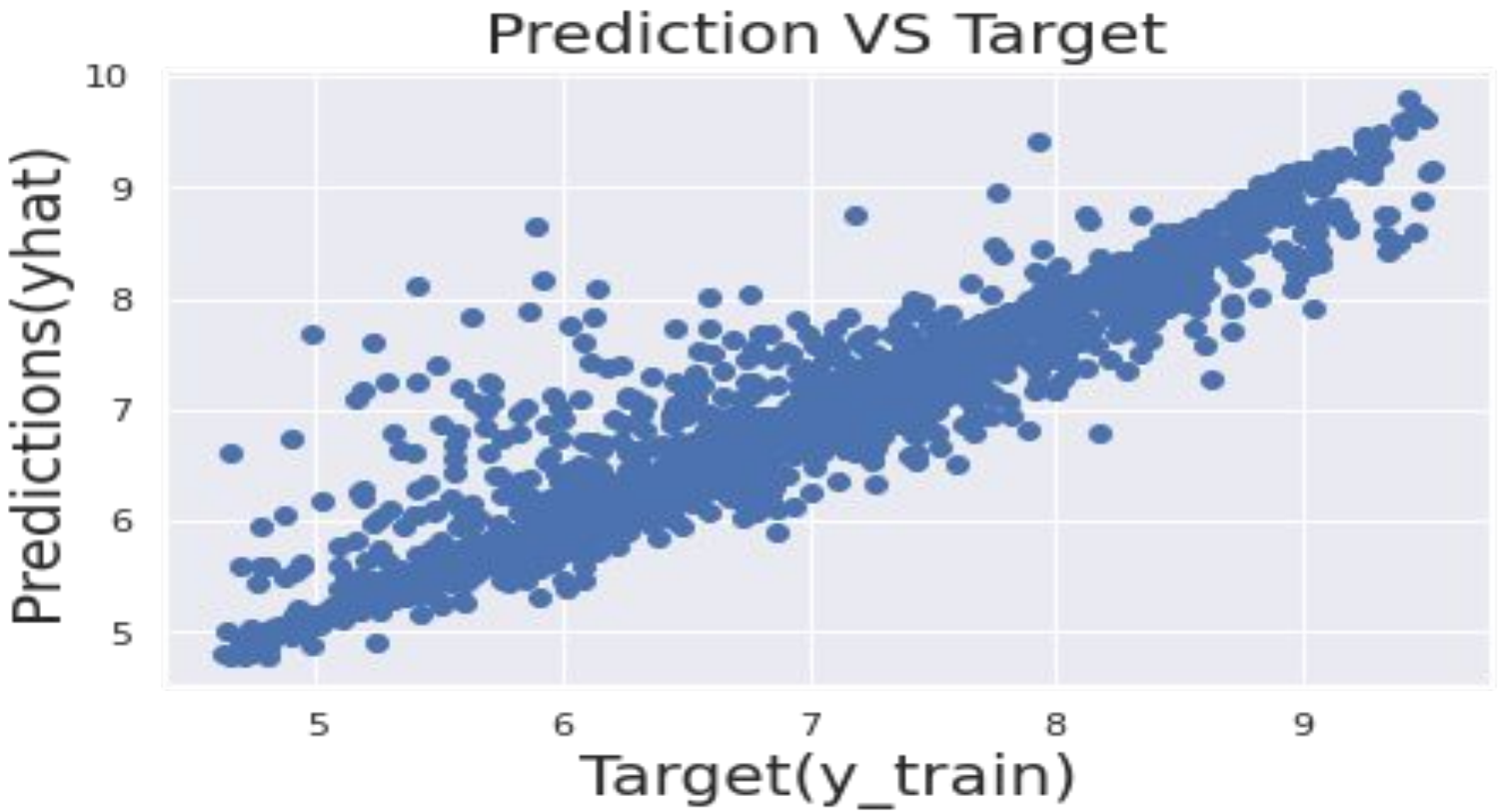



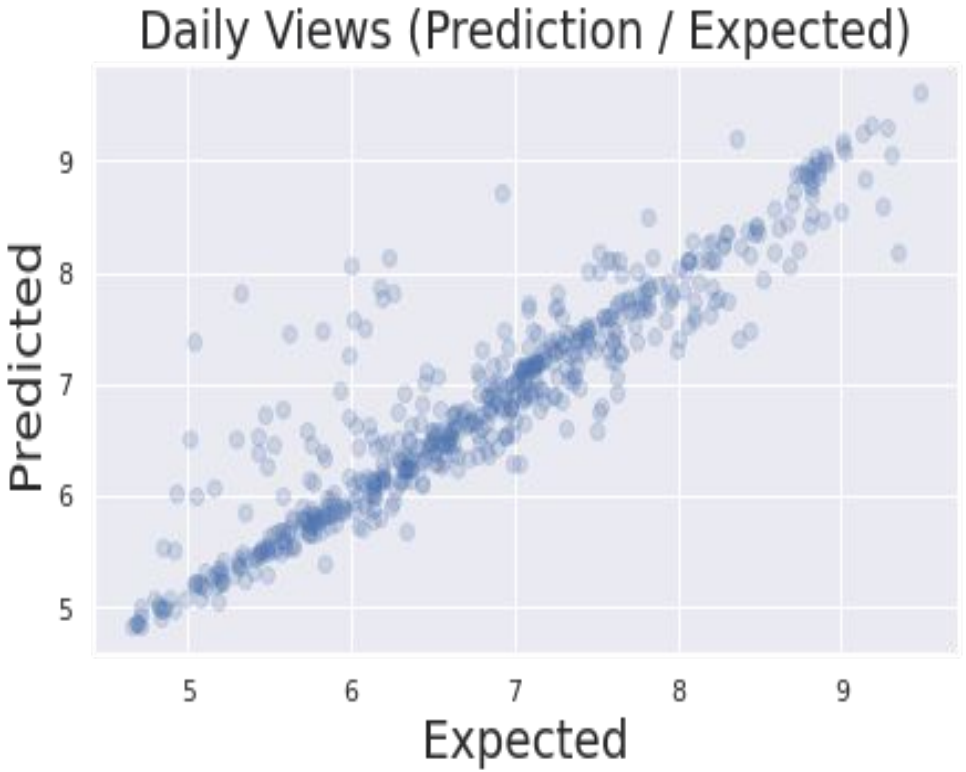
# Model preparation.

- Removing null values from dataset
- Introducing dummy variables for Categorical features
- Defining dependent and independent features
- Next we will standardize the features
- Splitting the data into training and testing
- Implementing Linear Regression Training Models
- Model Accuracy on test data

# Model Accuracy on train data



# Model Accuracy on test data (Base LR Model)

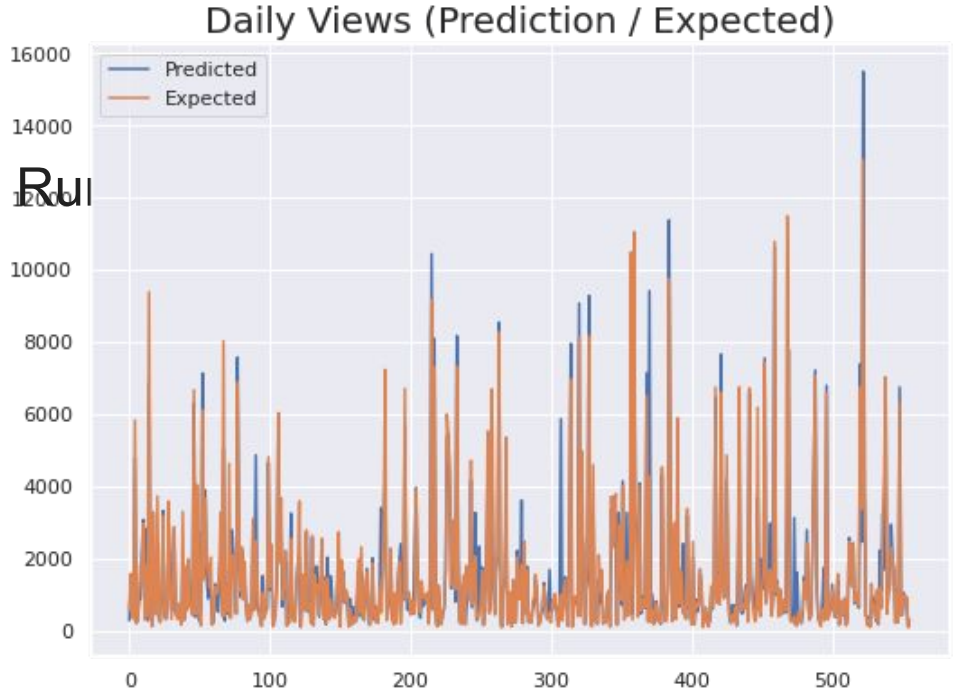


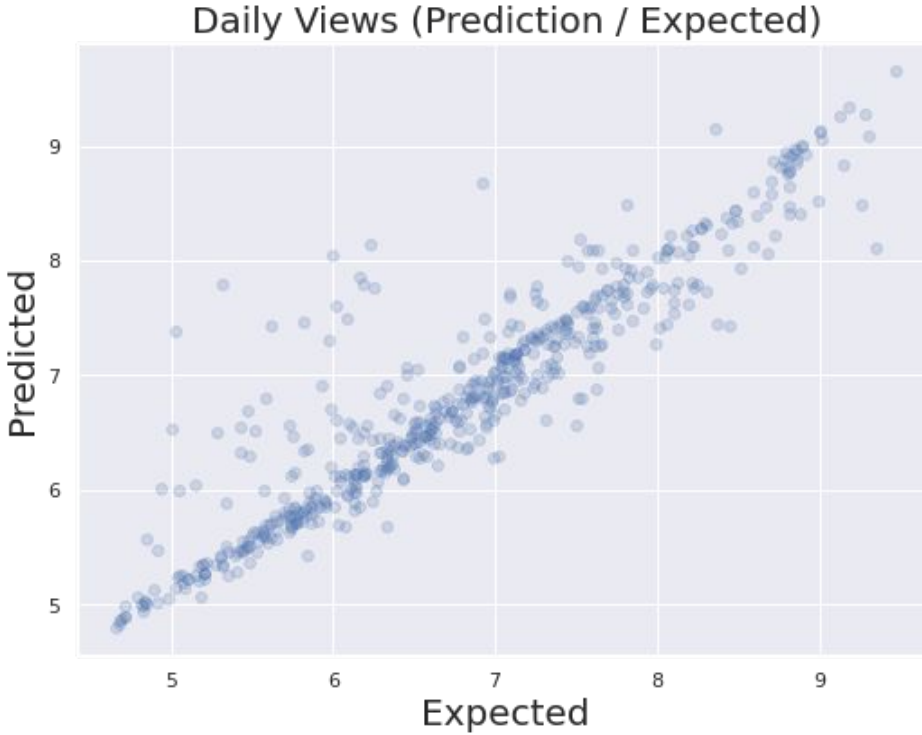

# Error metrics on Base LR Model

Values

| | |
|---|---|
| R-Square- | 0.83691 |
| Adj.R-Square- | 0.830500 |
| MSE- | 651492.850559 |
| RMSE- | 807.151070 |
| MAE- | 368.333887 |
| MAPE- | 0.344262 |

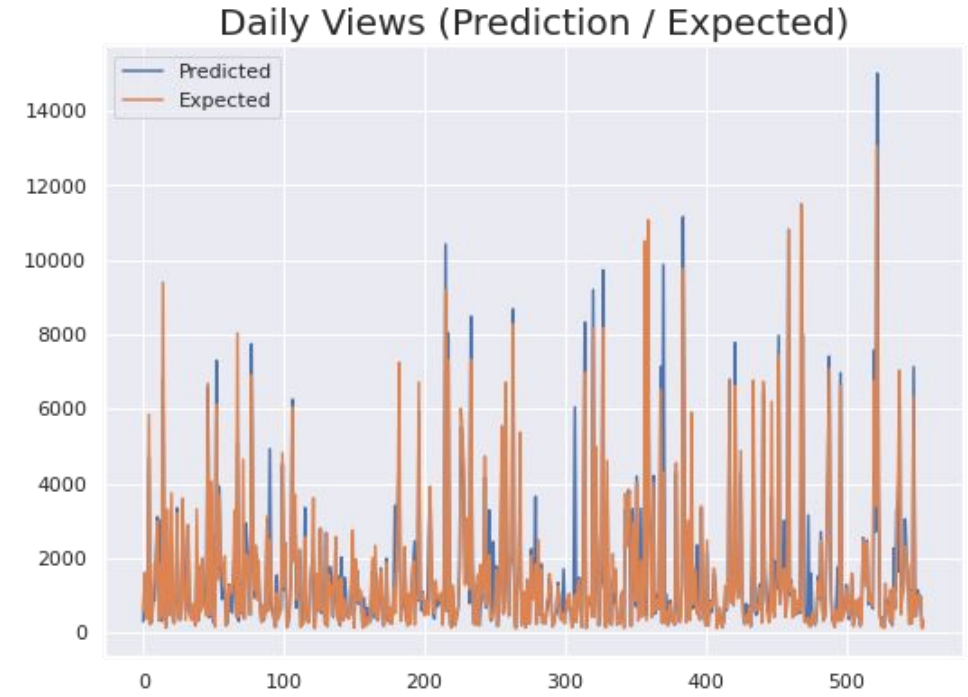
# Model Accuracy on test data (Lasso Regression Model)

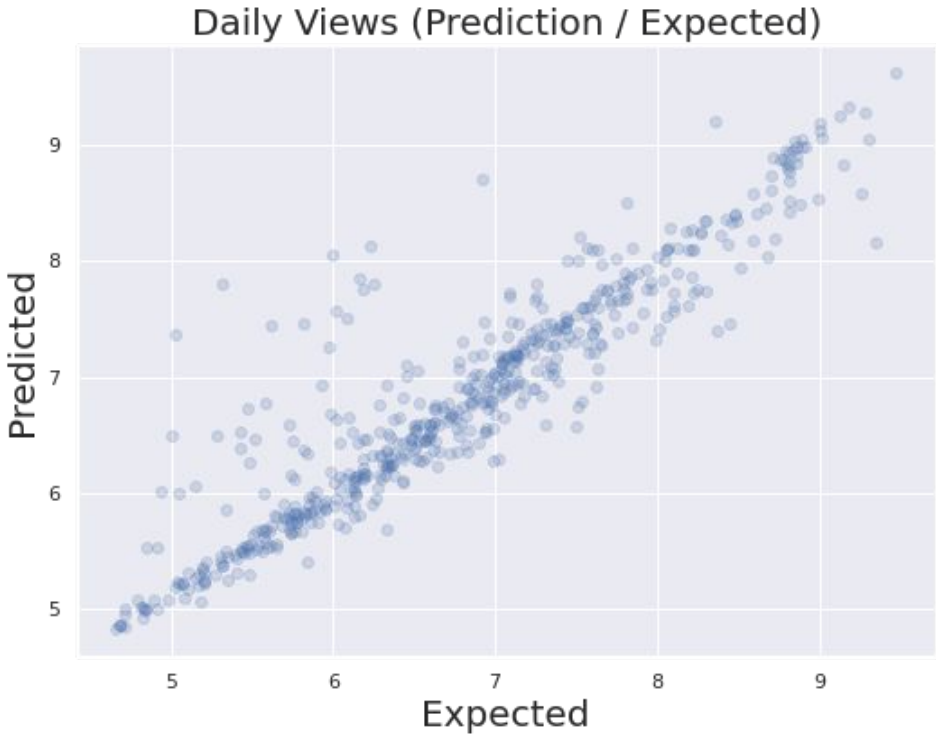


# Ridge regression model

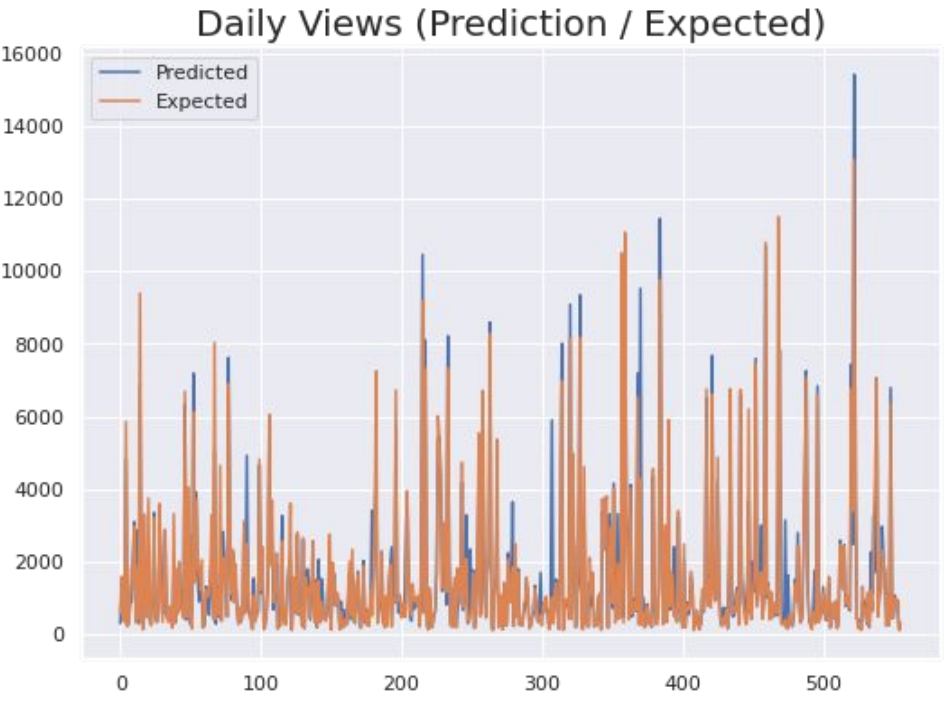
## Running Grid Search Cross Validation

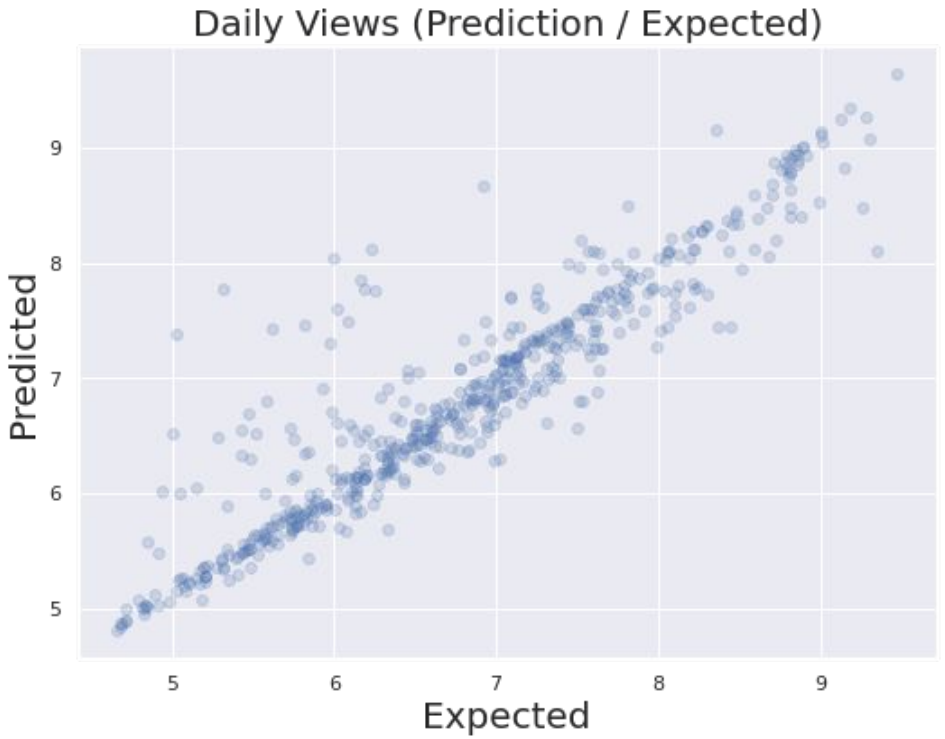


# Elastic regression model

## Running Grid Search Cross Validation



# Observation

On comparing all the models our base linear regression model is still is performing better followed by Lasso, Ridge and ElasticNet Regression model on the basis of RMSE. But our model contains large number of outliers and the value of RMSE is affected by outliers therefore we will use MAE as our evaluation matrix according to which *Lasso Regression* has the best performance

# Conclusion

- We performed EDA, feature engineering, data cleaning, target encoding and one hot encoding of categorical columns, feature selection and then model building.
- Then we checked our model for overfitting by comparing it with Lasso Regression model, Ridge Regression model, ElasticNet Regression model.
- We found that our original base model was overfit and Lasso Regressor has the best accuracy.
- In all of these models our mean errors is 13 %. That implies we have been able to correctly predict views 87 % of the time.
- In all the features speaker_1_avg_views is most important this implies that speakers are directly impacting the views.

# Future Work

- Training our data on other models (XGB, Random Forest, etc)
- More efficient Hyperparameter Tuning through techniques like Random Search

# Q/A