Capstone Project

News Popularity prediction

By,
Yamini

Problem statement

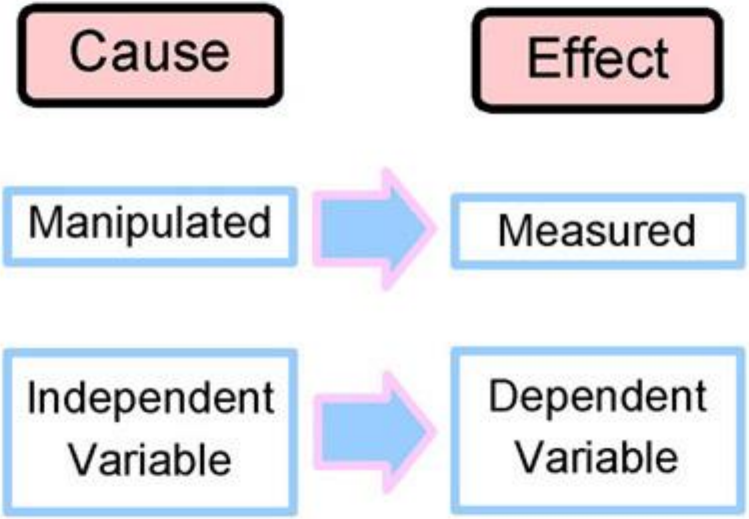
- **This is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn. The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: Economy, Microsoft, Obama and Palestine.**
- There are many variables which are impacting on the dependent variables. The problem here is find the number of likes or reactions generated on different topics discussed in different social media platforms using all other features available in given data with the news obtained from different sources.

Steps involved in social media prediction

a) Understanding the data
b) Missing value analysis
c) Data Cleaning
d) EDA
e) Data Visualization
f) Outlier detection from Statistical method or from boxplots
g) Creating new features from News title and Headline column
h) Feature engineering
i) Dimensionality reduction
j) Feature selection
k) Train and test split
l) Model building
m) Hyperparameter tuning
n) Model Validation
o) Model Selection

Independent variables

- Topic
- Sentiment Title
- Sentiment Headline
- Source
- Title
- Headline
- Publish Date



Cause

Effect

Manipulated

Measured

Independent Variable

Dependent Variable

Dependent variables

- **Facebook**
- **GooglePlus**
- **LinkedIn**

Attributes description

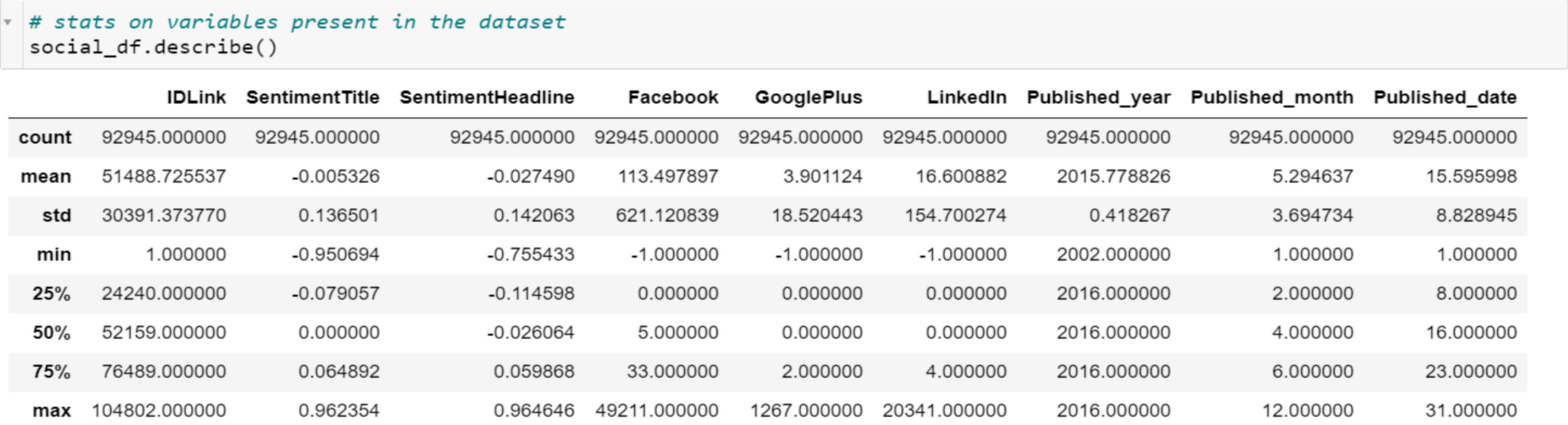
- IDLink (numeric): Unique identifier of news items
- Title (string): Title of the news item according to the official media sources
- Headline (string): Headline of the news item according to the official media sources
- Source (string): Original news outlet that published the news item
- Topic (string): Query topic used to obtain the items in the official media sources
- PublishDate (timestamp): Date and time of the news items' publication
- SentimentTitle (numeric): Sentiment score of the text in the news items' title
- SentimentHeadline (numeric): Sentiment score of the text in the news items' headline
- Facebook (numeric): Final value of the news items' popularity according to the social media source Facebook
- GooglePlus (numeric): Final value of the news items' popularity according to the social media source Google+
- LinkedIn (numeric): Final value of the news items' popularity according to the social media source LinkedIn

Data understanding

- There are a total of 93239 rows and 11 columns in the given data.
- There are 279 missing values in Source and 15 missing values in Headline.
- I have removed those rows as they are very less in percentage compared to the whole and avoids false information and also saves the memory which helps in avoiding memory problem.
- There are 5 object variables(categorical) and 6 numerical variables in the given data. This has been further processed and datetime variable is converted to datetime variable type, created new columns with publish date column.

EDA

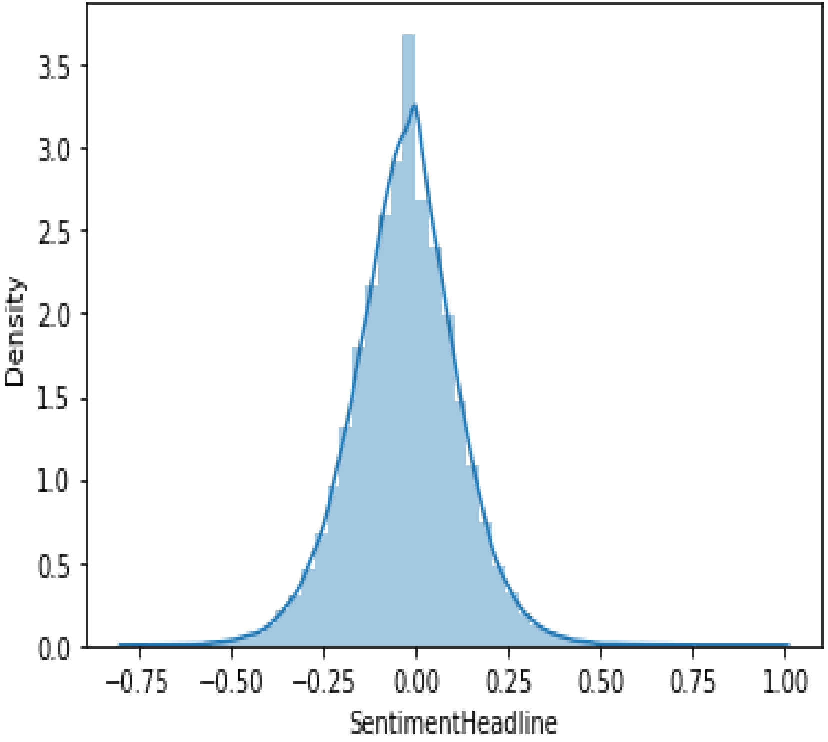
- There are four types of News Topics which are being discussed on the social media platforms. They are Microsoft, Obama, Economy and Palestine.
- Some of the statistics on numerical variables can be found through the describe method and some statistics can be found from pandas_profiling.

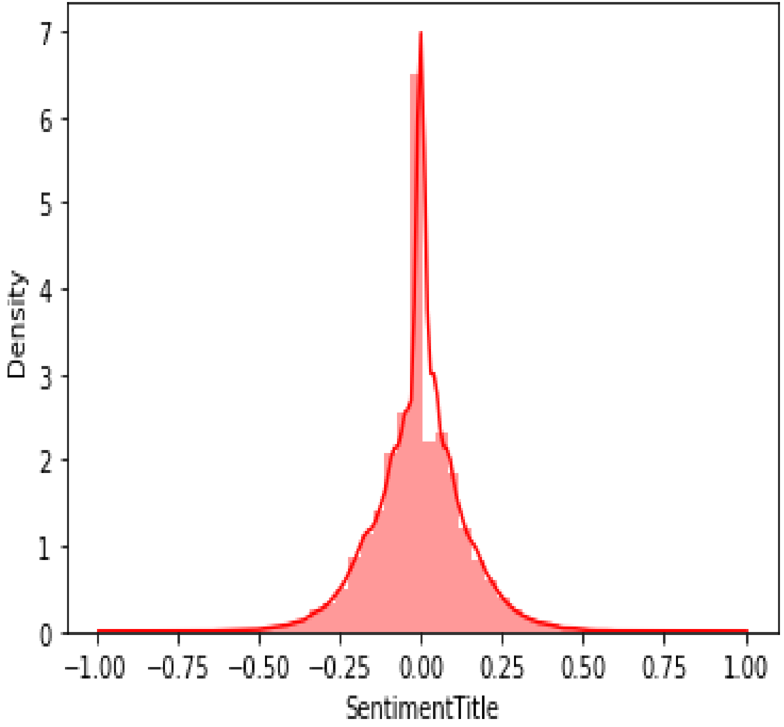


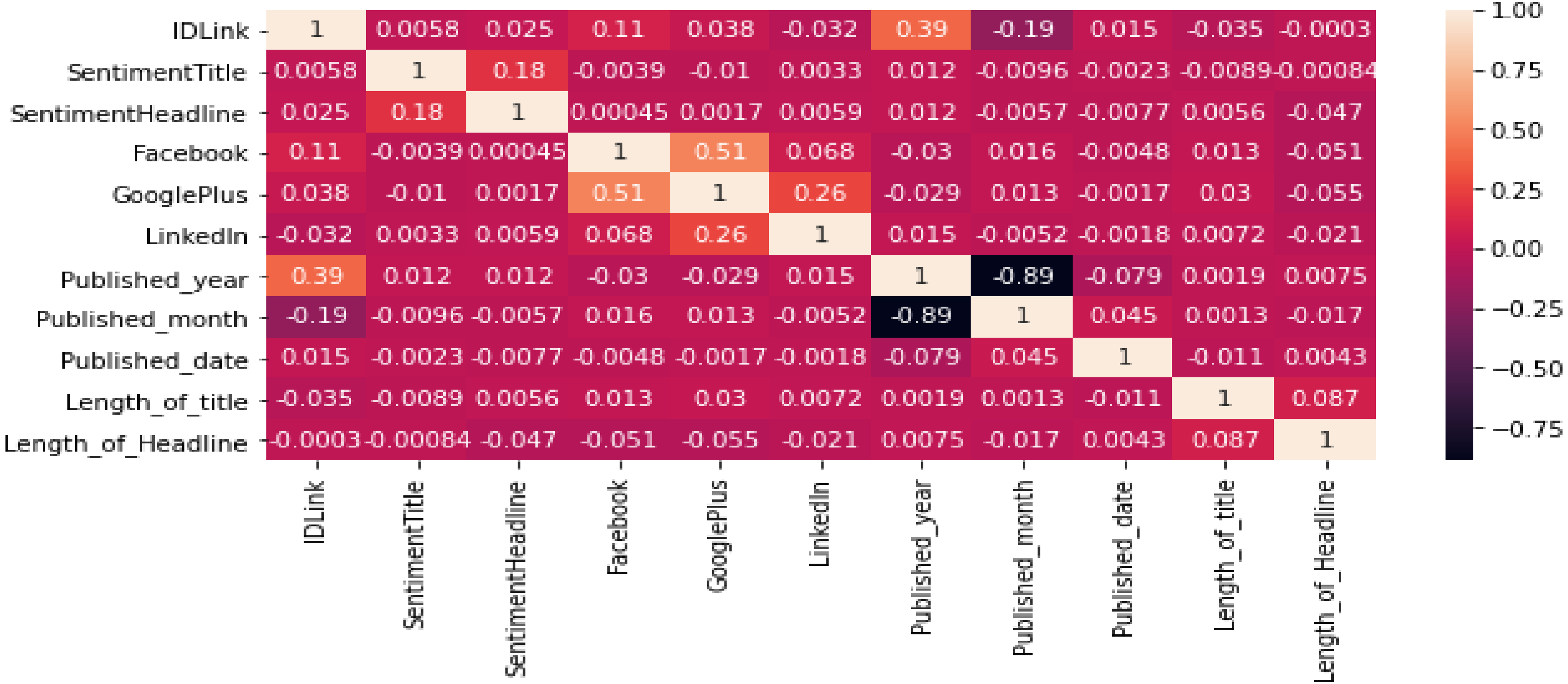
```
# stats on variables present in the dataset
social_df.describe()
```

	IDLink	SentimentTitle	SentimentHeadline	Facebook	GooglePlus	LinkedIn	Published_year	Published_month	Published_date
count	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000	92945.000000
mean	51488.725537	-0.005326	-0.027490	113.497897	3.901124	16.600882	2015.778826	5.294637	15.595998
std	30391.373770	0.136501	0.142063	621.120839	18.520443	154.700274	0.418267	3.694734	8.828945
min	1.000000	-0.950694	-0.755433	-1.000000	-1.000000	-1.000000	2002.000000	1.000000	1.000000
25%	24240.000000	-0.079057	-0.114598	0.000000	0.000000	0.000000	2016.000000	2.000000	8.000000
50%	52159.000000	0.000000	-0.026064	5.000000	0.000000	0.000000	2016.000000	4.000000	16.000000
75%	76489.000000	0.064892	0.059868	33.000000	2.000000	4.000000	2016.000000	6.000000	23.000000
max	104802.000000	0.962354	0.964646	49211.000000	1267.000000	20341.000000	2016.000000	12.000000	31.000000

Data Visualization

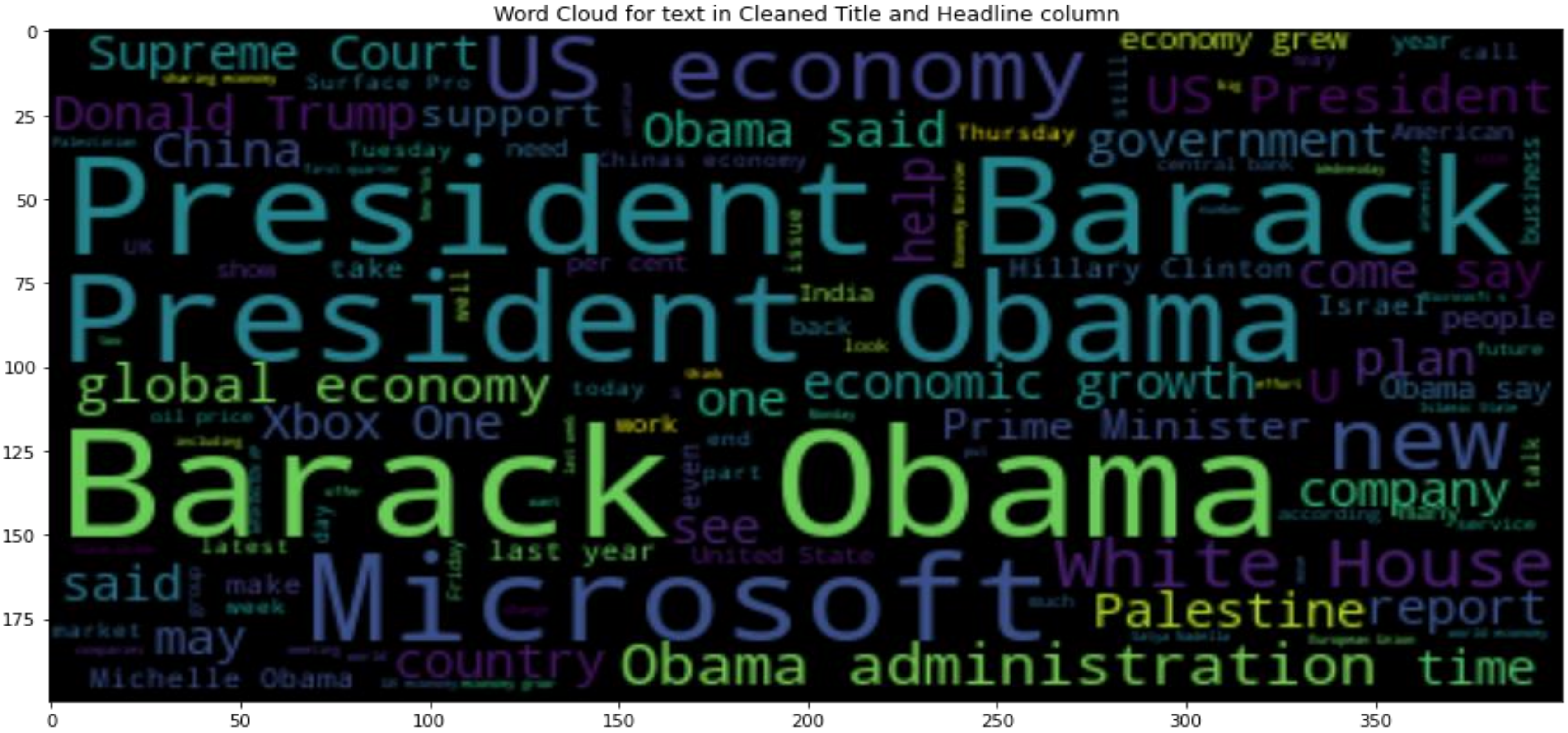




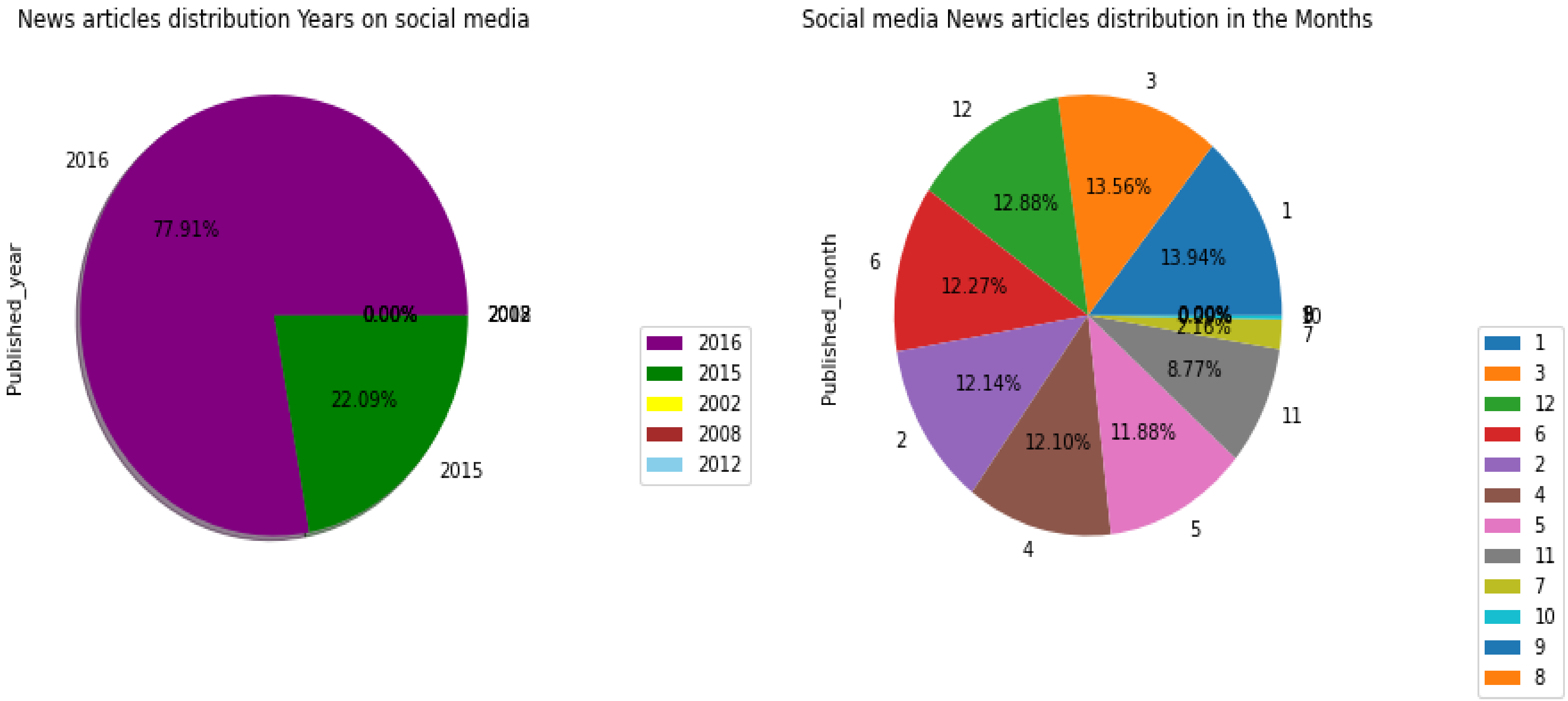


	IDLink	SentimentTitle	SentimentHeadline	Facebook	GooglePlus	LinkedIn	Published_year	Published_month	Published_date	Length_of_title	Length_of_Headline
IDLink	1	0.0058	0.025	0.11	0.038	-0.032	0.39	-0.19	0.015	-0.035	-0.0003
SentimentTitle	0.0058	1	0.18	-0.0039	-0.01	0.0033	0.012	-0.0096	-0.0023	-0.0089	-0.00084
SentimentHeadline	0.025	0.18	1	0.00045	0.0017	0.0059	0.012	-0.0057	-0.0077	0.0056	-0.047
Facebook	0.11	-0.0039	0.00045	1	0.51	0.068	-0.03	0.016	-0.0048	0.013	-0.051
GooglePlus	0.038	-0.01	0.0017	0.51	1	0.26	-0.029	0.013	-0.0017	0.03	-0.055
LinkedIn	-0.032	0.0033	0.0059	0.068	0.26	1	0.015	-0.0052	-0.0018	0.0072	-0.021
Published_year	0.39	0.012	0.012	-0.03	-0.029	0.015	1	-0.89	-0.079	0.0019	0.0075
Published_month	-0.19	-0.0096	-0.0057	0.016	0.013	-0.0052	-0.89	1	0.045	0.0013	-0.017
Published_date	0.015	-0.0023	-0.0077	-0.0048	-0.0017	-0.0018	-0.079	0.045	1	-0.011	0.0043
Length_of_title	-0.035	-0.0089	0.0056	0.013	0.03	0.0072	0.0019	0.0013	-0.011	1	0.087
Length_of_Headline	-0.0003	-0.00084	-0.047	-0.051	-0.055	-0.021	0.0075	-0.017	0.0043	0.087	1

Correlation matrix displaying the correlation between the numerical variables

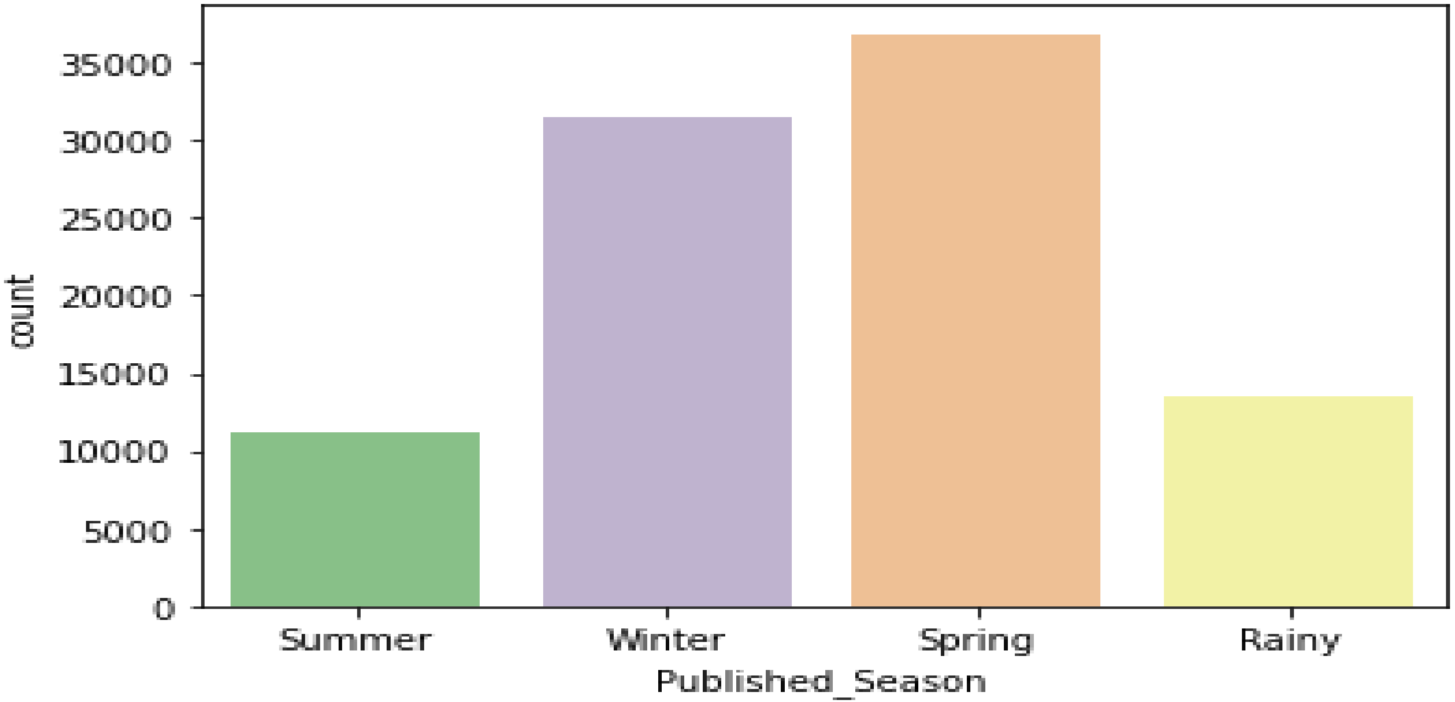


The most important terms from the News title and Headline column

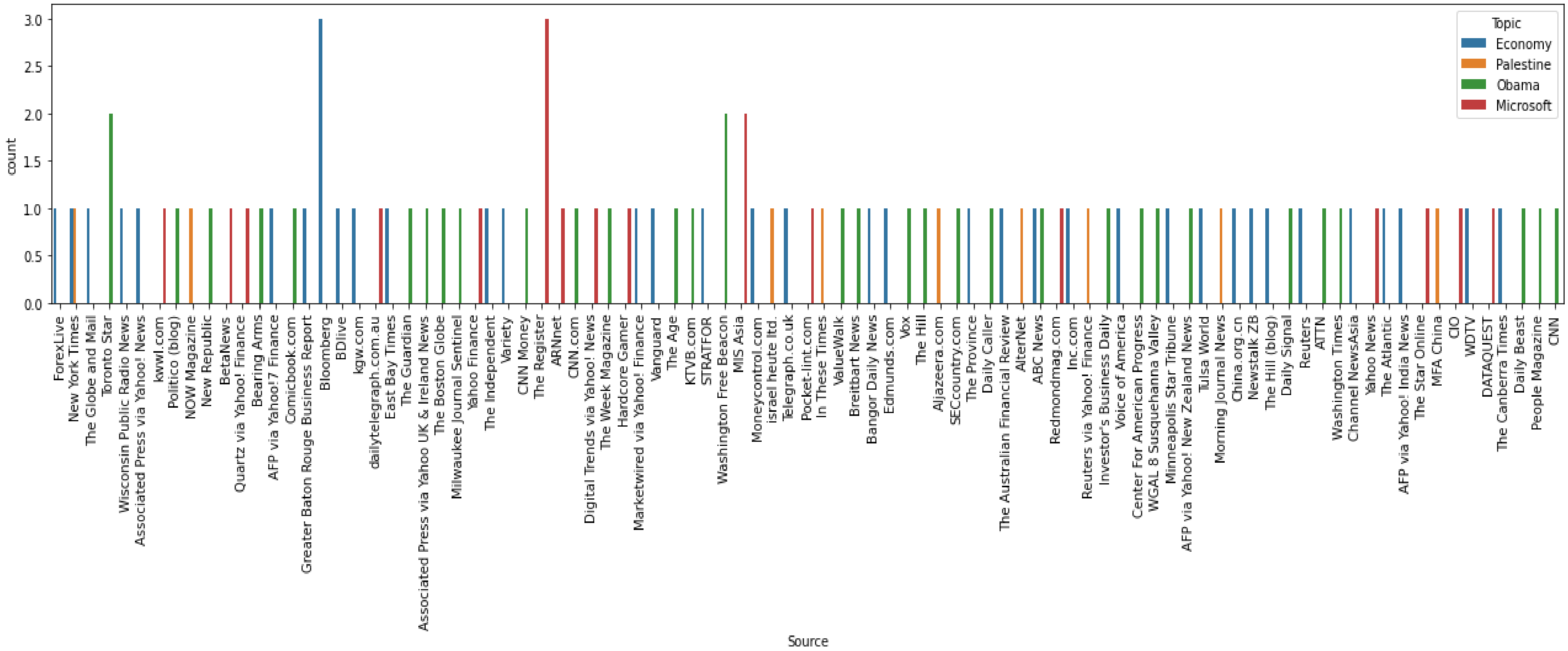

News articles distribution Years on social media

Social media News articles distribution in the Months

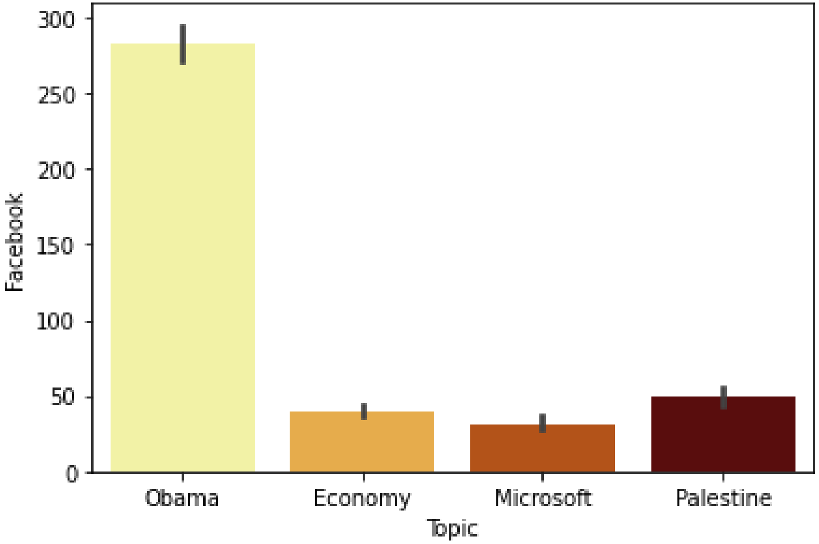
Distribution of News articles in the respective years and months

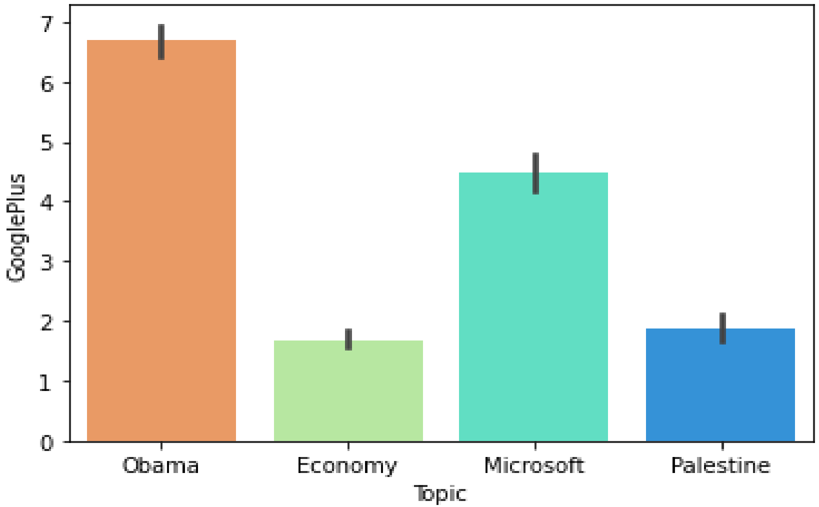


Count of News articles discussed in different Seasons

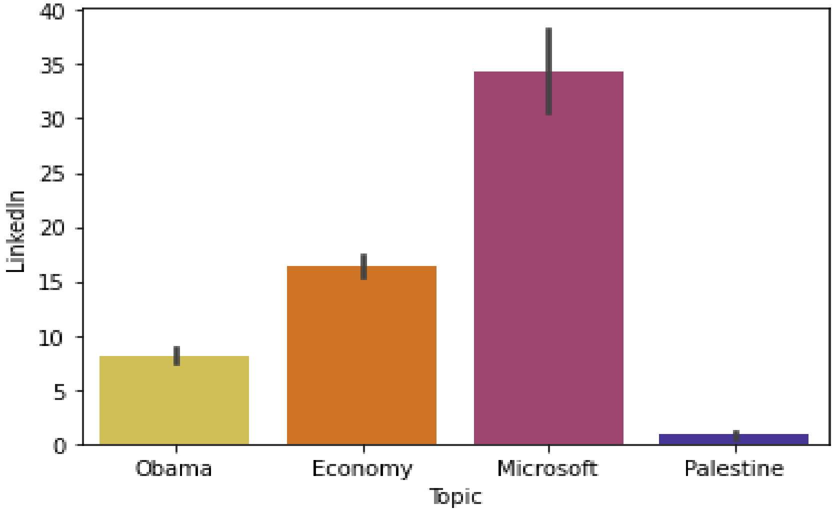


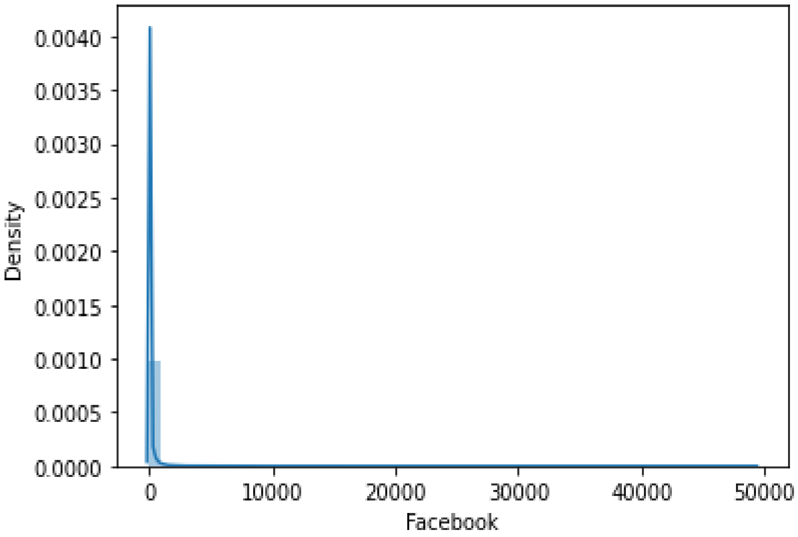
News articles published in social media and their Sources w.r.t the different topics

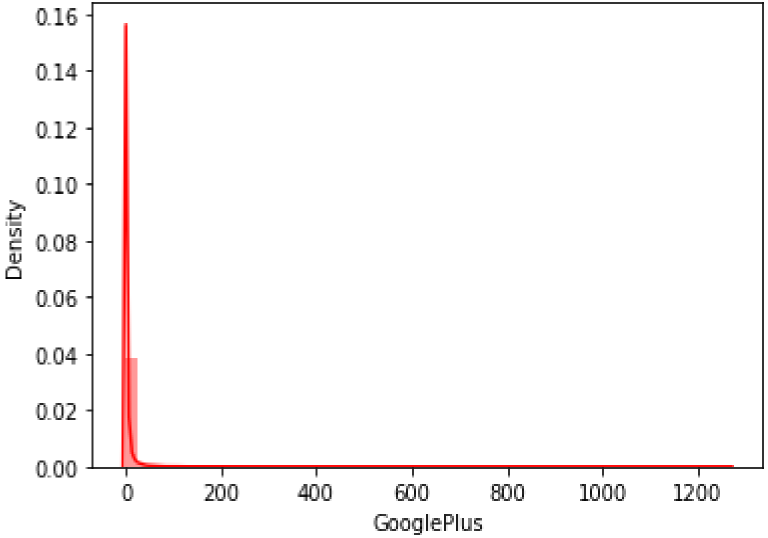


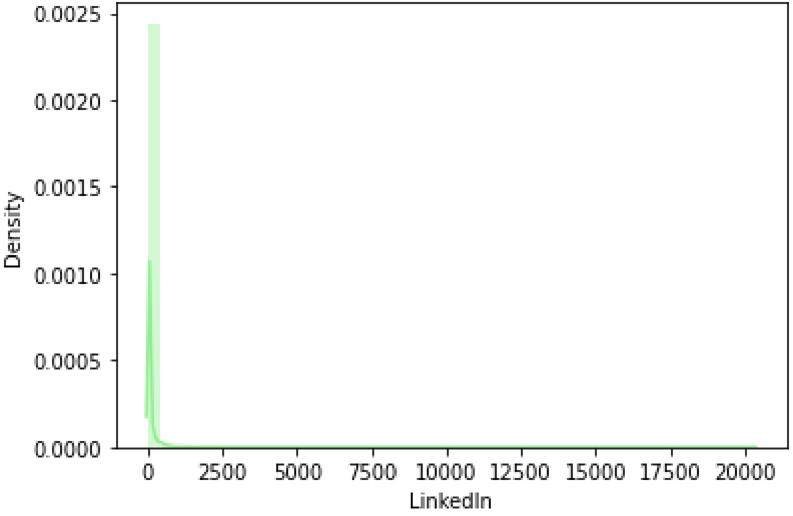


The four different topics news articles being discussed on different Social media platforms Fb, G+ and LinkedIn





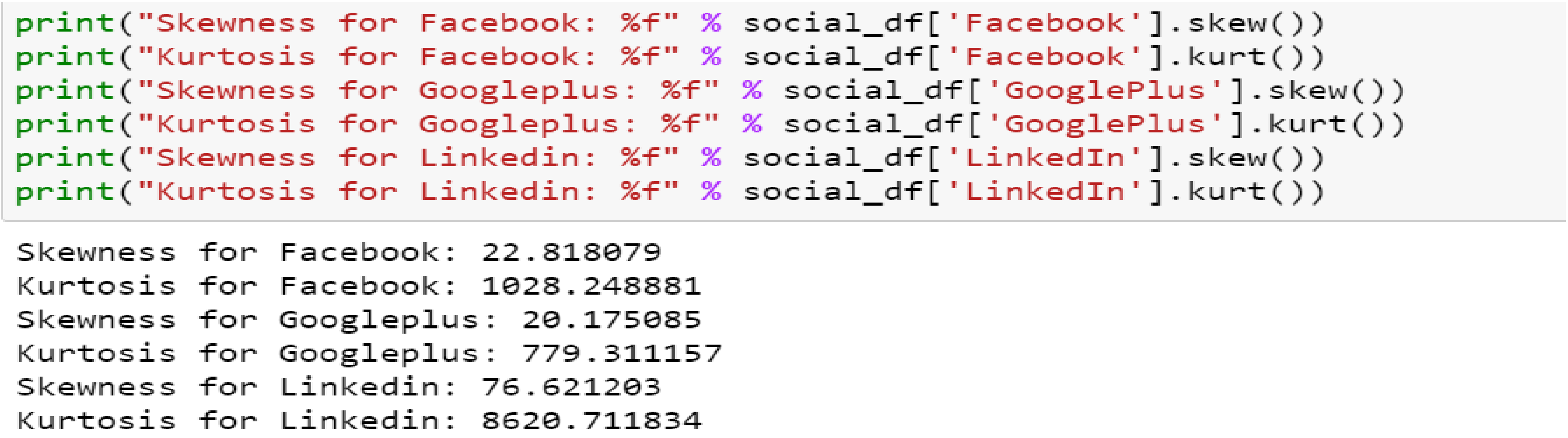




The distribution of values in different social media platforms

Skewness and Kurtosis on target variable

- **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry. ... **Kurtosis** is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high **kurtosis** tend to have heavy tails, or outliers.



```
print("Skewness for Facebook: %f" % social_df['Facebook'].skew())
print("Kurtosis for Facebook: %f" % social_df['Facebook'].kurt())
print("Skewness for Googleplus: %f" % social_df['GooglePlus'].skew())
print("Kurtosis for Googleplus: %f" % social_df['GooglePlus'].kurt())
print("Skewness for Linkedin: %f" % social_df['LinkedIn'].skew())
print("Kurtosis for Linkedin: %f" % social_df['LinkedIn'].kurt())
```

```
Skewness for Facebook: 22.818079
Kurtosis for Facebook: 1028.248881
Skewness for Googleplus: 20.175085
Kurtosis for Googleplus: 779.311157
Skewness for Linkedin: 76.621203
Kurtosis for Linkedin: 8620.711834
```

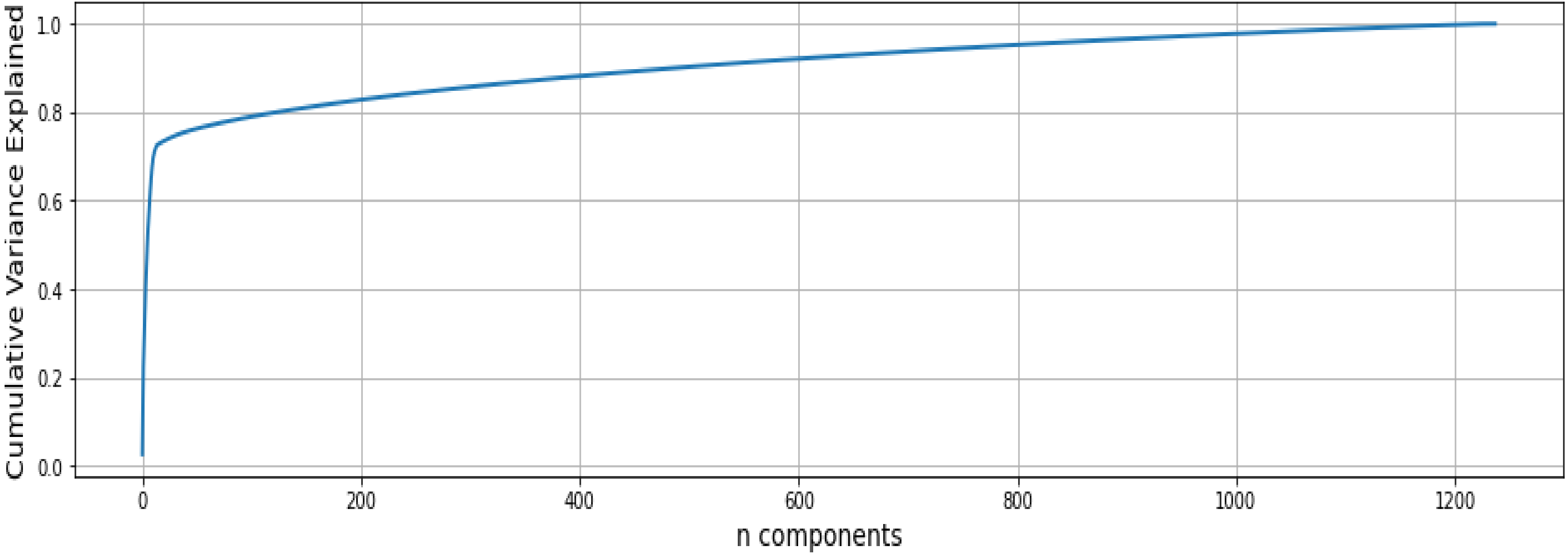
Feature Engineering

- In feature engineering, I have created new features from the existing columns like Published date and also created dummies for categorical variables for using it in model building with without model cannot distinguish between numerical and categorical variables.
- I have standardized the numerical variables with Standard scaler so that all the numerical values stay in one range otherwise there is a chance of misinterpretation.
- Correctly assigning the datatypes to the respective columns
- Combining all the created features into one dataframe to use it for further processing.

Feature Selection

- Drop the redundant variables for further processing and by which it avoids multicollinearity.
- I have used dimensionality reduction technique to reduce the dimension and storing the most important variables which would be useful for predictions with required variance level.
- So that it decreases in size and decreases computation cost and also saves time that is required to process huge dataset.
- Feature Selection plays such an important role because we are required to be cautious while selecting the features and dropping useful variables would bring adverse effect on the final score and in bringing good predictions

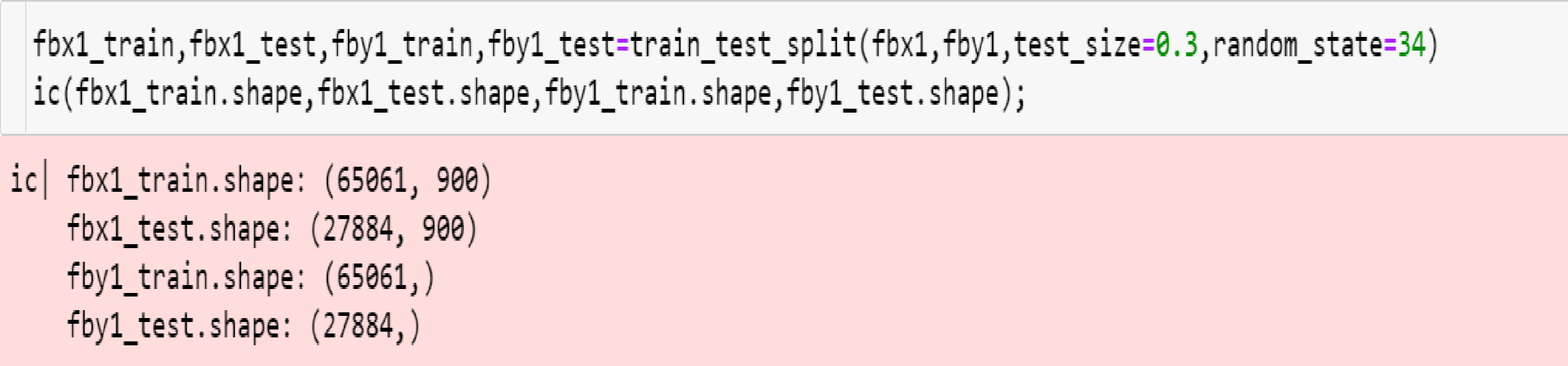
Dimensionality reduction technique result



I used SVD which gives the number of variables that can be used for model building w.r.t Variance required.

Train and test split

- Below is the train and test size for making facebook predictions



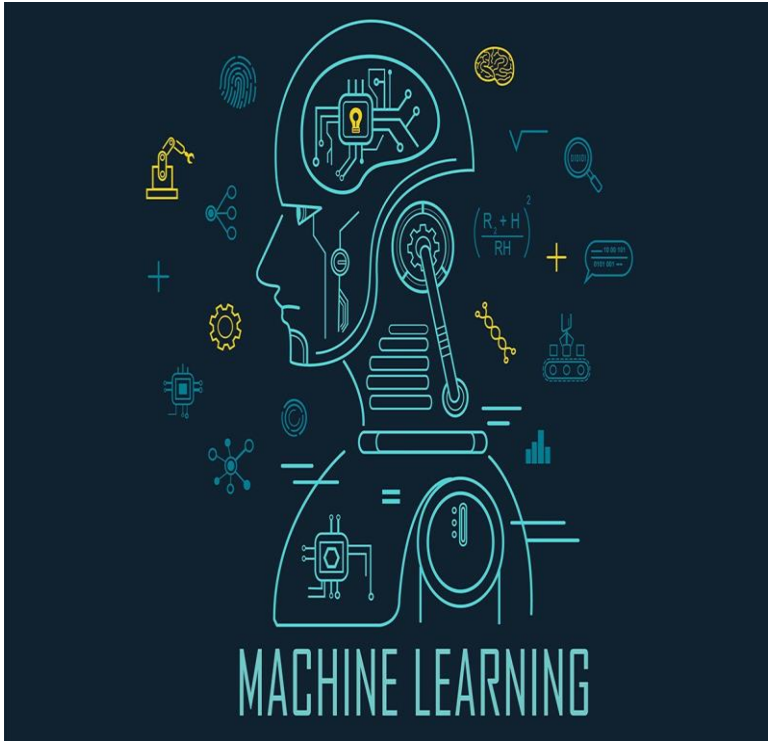
```
fbx1_train,fbx1_test,fby1_train,fby1_test=train_test_split(fbx1,fby1,test_size=0.3,random_state=34)
ic(fbx1_train.shape,fbx1_test.shape,fby1_train.shape,fby1_test.shape);
```

```
ic| fbx1_train.shape: (65061, 900)
    fbx1_test.shape: (27884, 900)
    fby1_train.shape: (65061,)
    fby1_test.shape: (27884,)
```

- The size is the same for other two social media platforms googleplus and linkedin for making their predictions on different topics.

Machine Learning Algorithms/models used for this regression problem solving

1) Linear Regression
2) Decision Tree
3) Gradient Boost Regressor
4) Cat Boost Regressor
5) XGB Regressor
6) KNN Regressor

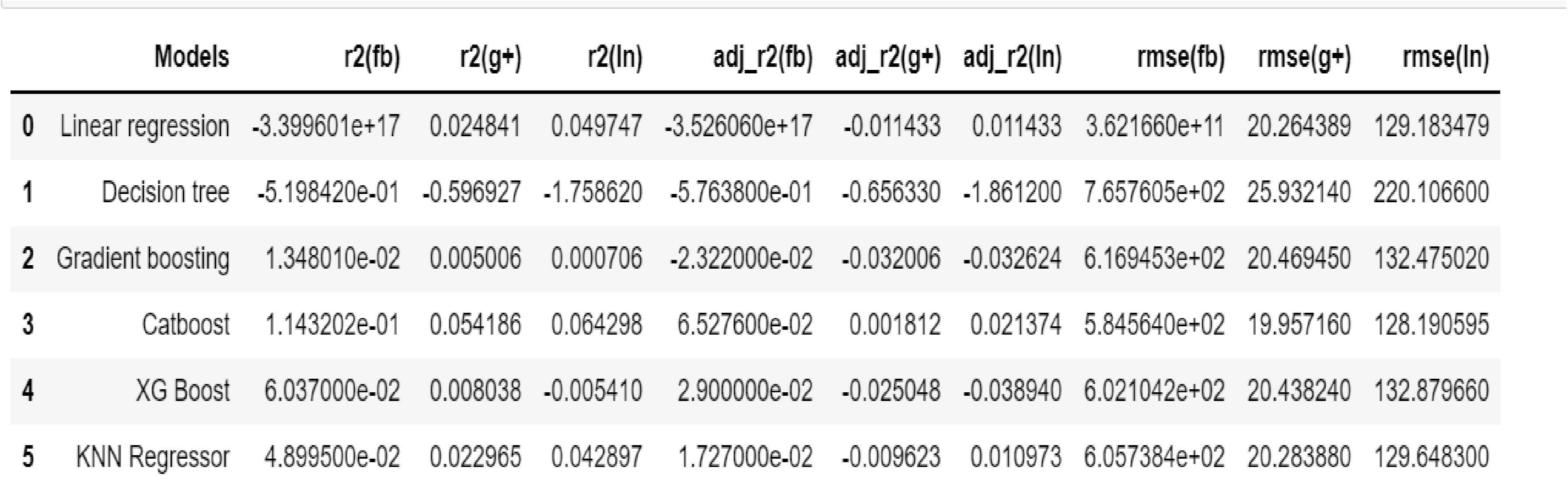


Evaluation metrics

- The evaluation metrics used for model validation are r2_score, adjusted r2_score and rmse.
- R2_score which is also called as the coefficient of determination through which we get to know how well the model is able to capture the variance of the data.
- Adjusted r2 score is same like r2 score then adj r2 score penalizes for adding the useless variables whereas r2 score just ignores them. Using adj r2 score it is easy to say that whether adding additional variables to the model had a made significant impact in making predictions or not.
- **Root Mean Square Error (RMSE)** is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; **RMSE** is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
- As stated above rmse is about finding the error/residuals in the data.

Results

- The r2 scores, adj r2 scores and rmse of different social media platform predictions based on independent features for different news articles can be seen.



	Models	r2(fb)	r2(g+)	r2(ln)	adj_r2(fb)	adj_r2(g+)	adj_r2(ln)	rmse(fb)	rmse(g+)	rmse(ln)
0	Linear regression	-3.399601e+17	0.024841	0.049747	-3.526060e+17	-0.011433	0.011433	3.621660e+11	20.264389	129.183479
1	Decision tree	-5.198420e-01	-0.596927	-1.758620	-5.763800e-01	-0.656330	-1.861200	7.657605e+02	25.932140	220.106600
2	Gradient boosting	1.348010e-02	0.005006	0.000706	-2.322000e-02	-0.032006	-0.032624	6.169453e+02	20.469450	132.475020
3	Catboost	1.143202e-01	0.054186	0.064298	6.527600e-02	0.001812	0.021374	5.845640e+02	19.957160	128.190595
4	XG Boost	6.037000e-02	0.008038	-0.005410	2.900000e-02	-0.025048	-0.038940	6.021042e+02	20.438240	132.879660
5	KNN Regressor	4.899500e-02	0.022965	0.042897	1.727000e-02	-0.009623	0.010973	6.057384e+02	20.283880	129.648300

Model Selection

- From the results it can be seen that CatBoost is performing well on this than any other model with r2 score of 0.1143, 0.05419, 0.0643 for facebook, googleplus and linkedin predictions.
- Even adj r2 scores are high are for this model than any other. Rmse is relatively low for this model making this model best to be chosen for making popularity predictions.
- Even after hyperparameter tuning for these models Catboost is best to choose for making news popularity predictions on the four different topics on three social media platforms Facebook, GooglePlus, LinkedIn.

Conclusion

- Predicting the exact number of likes or interactions based on different topics of news articles on the social media platforms can be hard then approximate or near estimation can be done.
- May be this the reason behind the low r2 scores on the models compared to other regression problems where we can get upto 70-95% scores with tuning and choosing different models.
- There is an imbalance in the news articles topics given in the training data. Having more data on the different topics can increase the performance of the models.
- Based on the news articles people have seen it can be said that there are good number of people who have interacted or discussed or have responded to the news on given social media.

Challenges

- Huge dataset
- Computational cost and time
- Missing values
- Faced with Memory crash problems
- Required more information like promotional activities..etc
- Finding the target feature and in understanding the problem statement
- Data security concerns

Thank you😊