New York Times Data Analysis Using Hive

Bita Faraji, Jay Joshi, Shradha Shinde

Department of Information Systems, California State University

Los Angeles

Tel. 626-438-0899, Fax. 323-343--5209

e-mail: bfaraji@calstatela.edu, jjoshi6@calstatela.edu, sshinde@calstatela.edu

**Abstract:** In this project, our group will analyze “New York Times Comments,”, a data set containing information about the comments made on the articles published in New York Times for the period January-June 2017 and January-May 2018. It illustrates the usage of Hadoop, MapReduce, and Hive on big data for easy summarization by utilizing the knowledge gained in lab sessions by querying big data, hands-on practice, extensive researches and development of HiveQL in order to generate and query about 1.5 GB of data and visualize it on Tableau, Power BI, Excel 3D Maps. The dataset files for comments contain over 2 million comments in total with 34features and those for articles contain 16 features about more than 9,000 articleswhich include comments on articles, number of readers, headline categorization, document type, section name and many more. This data will serve the purpose of understanding and analyzing the public reading interests, analyzing behaviors of the top commentators such as which topics they most likely comment and the sentiment analysis of the comments and the best author for that particular year. Other Elements of this project include a report paper, a tutorial on the queries, and one group presentation.

URL: <https://www.kaggle.com/aashita/nyt-comments>

Dataset size: 1.5GB

Cluster version: Hadoop 2.8.5-amzn-4

No of nodes: 3

HDFS Capacity: 147GB

CPU Speed: 2.20GHz

Storage: 678GB

Hive Version: Hive 2.3.5-amzn-1

**1. Introduction**

Newspapers have been part of people's life for decades and even after the digital revolution, a large amount of masses read and gain awareness about the daily happenings from newspapers. New York Times has wide audience and plays important role in shaping people’s opinion about current affairs, especially in United States of America. The comments sections for articles in the NYT are quite active and give insights to readers’ opinions on the subject matter of the articles. Each comment can receive other readers’ recommendations in the form of upvotes. Our First aim is to classify a given piece of material in NYT as an ‘article’ or ‘blogpost’ and then further categorize in that which is most read material in articles and then to find the article on which the commenters such as people/NYT users/editors most likely comment. Second step is to analyze the public response over these articles and blogposts by seeing the number of reply count and to determine how many of these receive the most recommendations and thirdly to perform the sentiment analysis of these comments and lastly to analyze which is the best author for that particular year which furthermore gives us insight on author likes to write which type of articles more. Our target would be public, editors, authors, users and the commentators here. Performing the steps above would help us to analyze the trending topics of people's interest and the ones which receive a lot of response and recommendations based on the reply count and also most active area from USA which replies to these articles and blogposts.

**2. Manipulating Datasets**

**2.1 Tools and data processing**

Figure 1. Data Processing

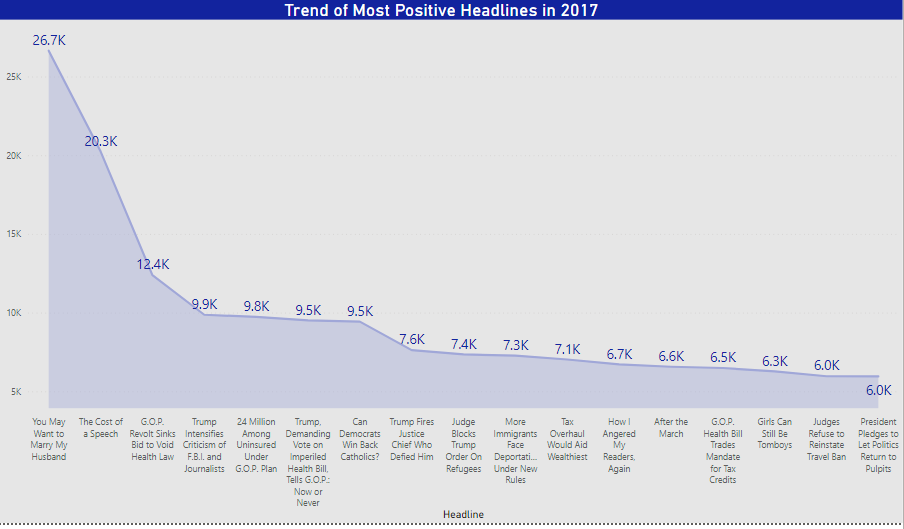
* In order to determine the readers reaction towards the headlines we analyzed and performed sentiment analysis on the dataset April Year 2017, Comment Year 2017 with respect to relative polarity to thoroughly distinguish the positive and negative top headlines.
* Based on the above results which displayed the attitude of readers with respect to specific headlines, we further analyzed the datasets April Year 2017, Comment Year 2017, April Year 2018 and Comment Year 2018 to conclude the best author in these years and also to determine the most active users are from which location with the help of reply count.
* Important commands like wget, mkdir and hive were used to initiate the project. We were able to successfully download and upload the New York Times datasets from Kaggle following which we connected with Hive and built intricate codes to evaluate each dataset.
* Furthermore, we also conducted analyses to derive results based on below queries.

|  |  |  |
| --- | --- | --- |
| Sr No. | Analysis Topic | Categories |
| 1 | Types of Documents that were published more with respect to the year 2017 and 2018 | Articles/Blogs |
| 2 | Types of Materials that were published more with respect to the year 2017 and 2018 | News/Op-Ed/Review/  Editorial/  briefing  /Letter |
| 3 | Document Type: Article that received highest number of replies with respect to the year 2017 and 2018 | Articles |
| 4 | Document Type: Blogs that received highest number of replies with respect to the year 2017 and 2018 | Blogs |
| 5 | Which type of comment received the highest number of counts with respect to the year 2017 and 2018 | Comment/User Reply/  Report Reply |
| 6 | Which month received the highest count of New Desk(Section categories like culture, dining, editorial) for the year 2017 and 2018 for the articles | January, February, March,  April, May, June |
| 7 | Which month received the highest count of New Desk(Section categories like culture, dining, editorial) for the year 2017 and 2018 for the Blogpost | January, February, March,  April, May |
| 8 | Which is a highly recommended New Desk as per the readers with respect to 2017 & 2018 for the Articles | Business, Editorial, Foreign, Learning, National, OpEd |
| 9 | Which is a highly recommended New Desk as per the readers with respect to 2017 & 2018 for the Blogposts | Business, Editorial, Foreign, Learning, National, OpEd |
| 10 | Most Popular Author with respect to reader’s recommendations for the year 2017 and 2018 | Nicholas Kristof - 2017  Susan Chira -2018 |
| 11 | Replies were received greatly from which state in the United States with respect to 2017 & 2018 | New York, California, Chicago |

Table 1

**2.2 Sentiment Analysis (2017)**

We performed the sentiment analysis by evacuating appropriate degree of polarity on the datasets Article Year 2017, Comment Year 2017 and derived the top headlines which received most positive as well as most negative reactions from readers by analyzing their comments.

 Figure 2

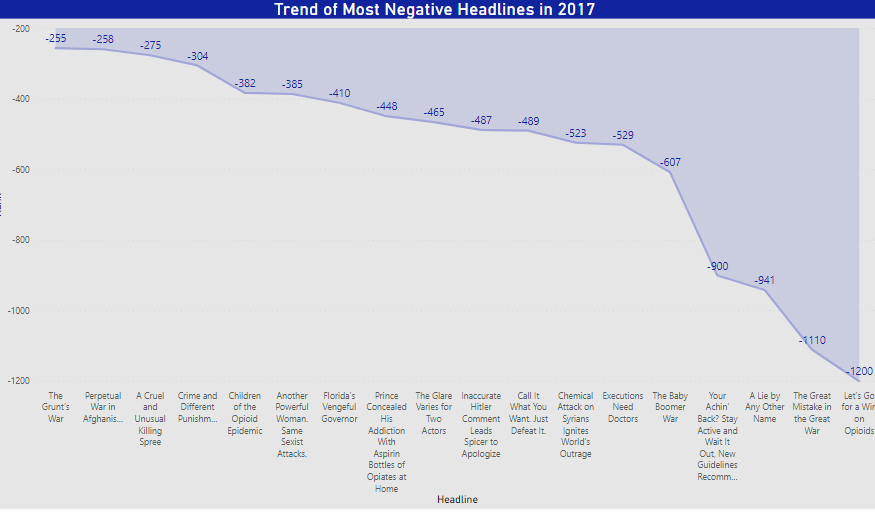


Figure 3

**2.3 Type of Document that were Highly Published and in which month of the year (2017,2018) did the type of Documents Receive Highest Number of Replies.**

New York Times majorly published 2 types of documents which area Articles and Blogs. We wrote queries on hive to analyze which type of documents were published more in 2017 and 2018 and acquired the result that Articles were highly published in both years compared to blogs. Also we categorized articles furthermore into type of material which was News, Op-Ed, Review, Editorial, Briefing, and Letter. We concluded that most read was news and least read was letter in 2017 and Interview in 2018.

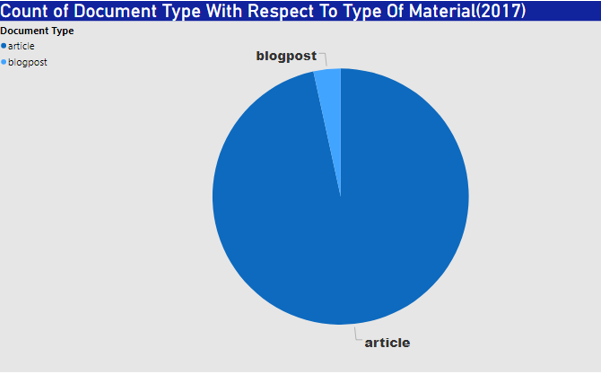


Figure 4

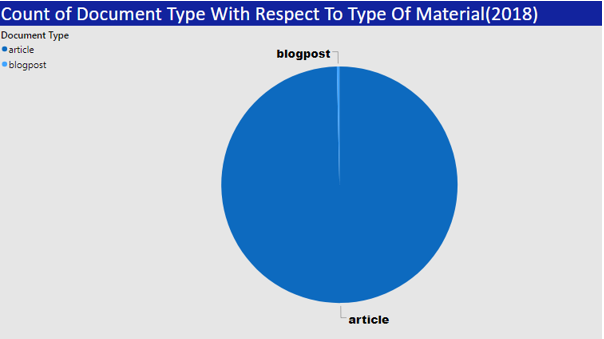


Figure 5

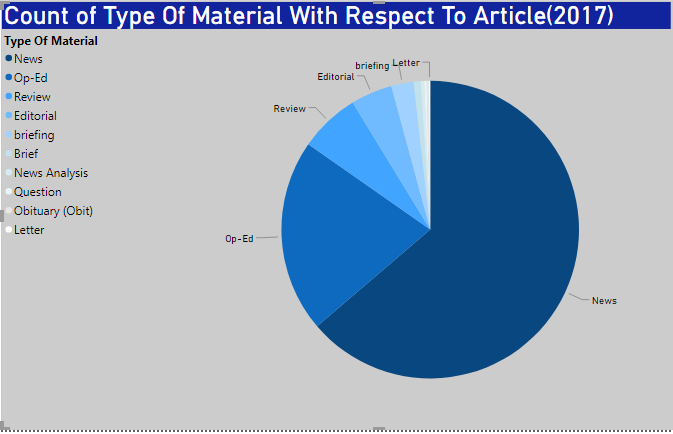
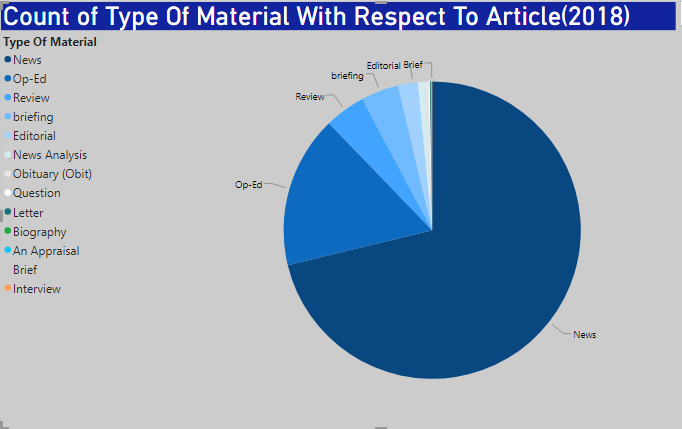


Figure 6

Figure 7

Furthermore, we found out with the help of hive queries, same month of both years i.e. March 2017 and March 2018 received the highest number of count and reader replies. For document type articles and for the blogpost document type March was highest in 2017 and in 2018 it was only January after which there was no response for consecutive 4 months.

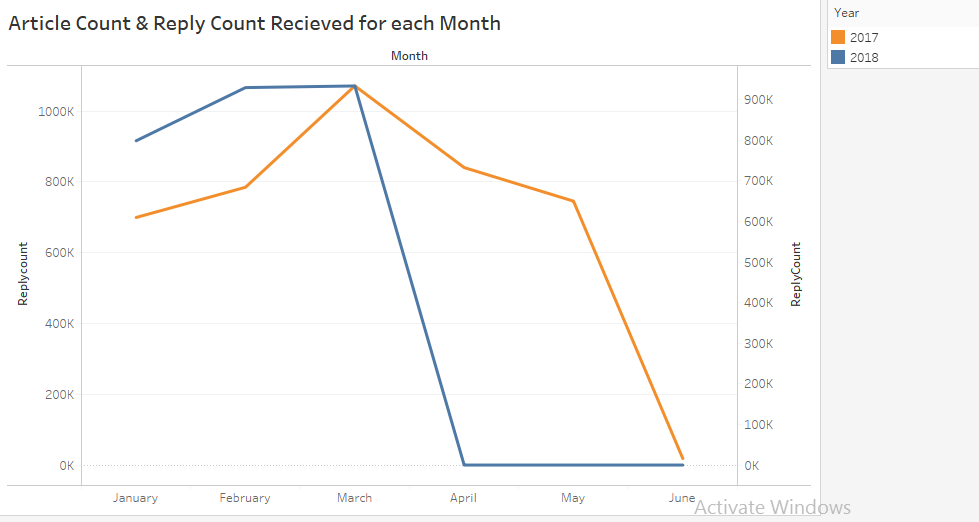


 Figure 8

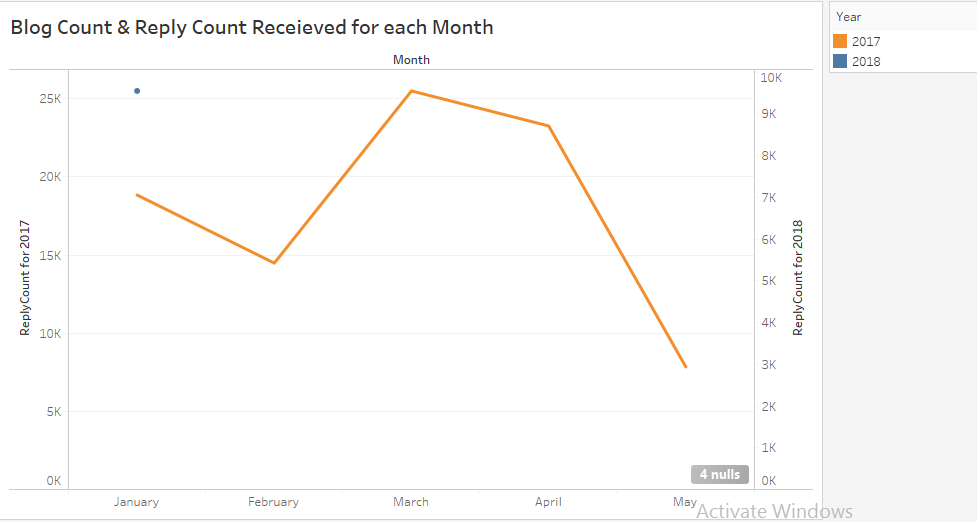


Figure 9

**2.4 Highest Number of Reply Count for Comment Type (Comment, User Reply, Report Reply)**

Comment types of readers are generally classified into 3 broad categories i.e. Comment, User Reply, Report Reply. Here the User Reply are associated with the registered users of New York Times and hence we were keen to know which type of comment raised the highest number of replies.

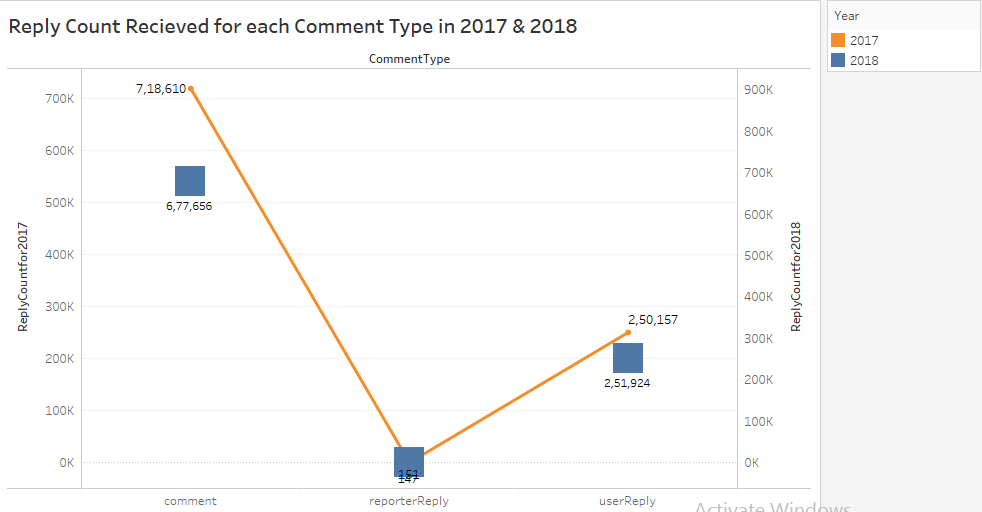


Figure 10

**2.5 Highest number of New Desk for the document type: Articles/Blogposts published for year (2017,2018) and Highly recommended New Desk as per Readers Interests.**

Articles are categorized based on the interests and likings of the readers where the column named New desk involves such categories. Some of the categories under New Desk are Art & Leisure, Business, Dining, Culture, OpEd (Editorial Opinion), Learning, National, Foreign. We wrote queries to know what type of New Desk the readers highly recommended for both the years and in which month of both the years was New Desk highly published for both the document types.

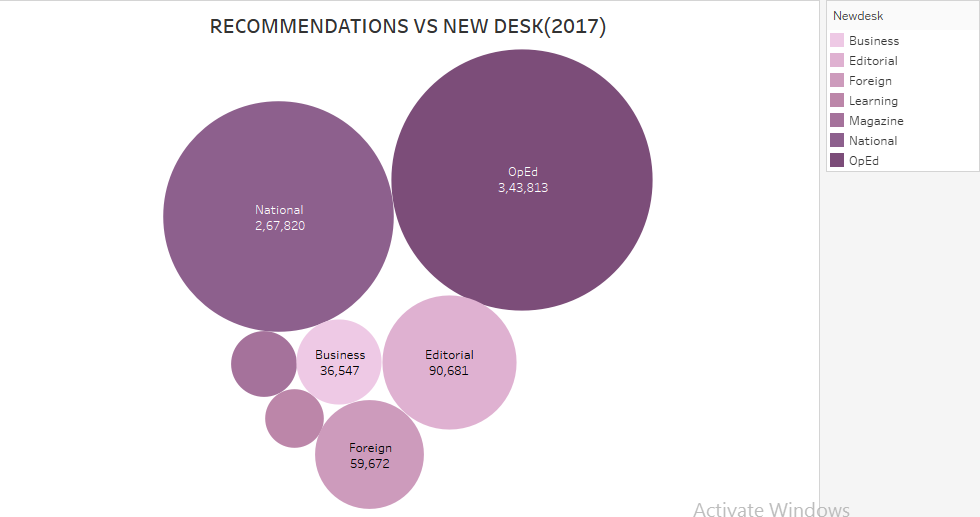


Figure 11

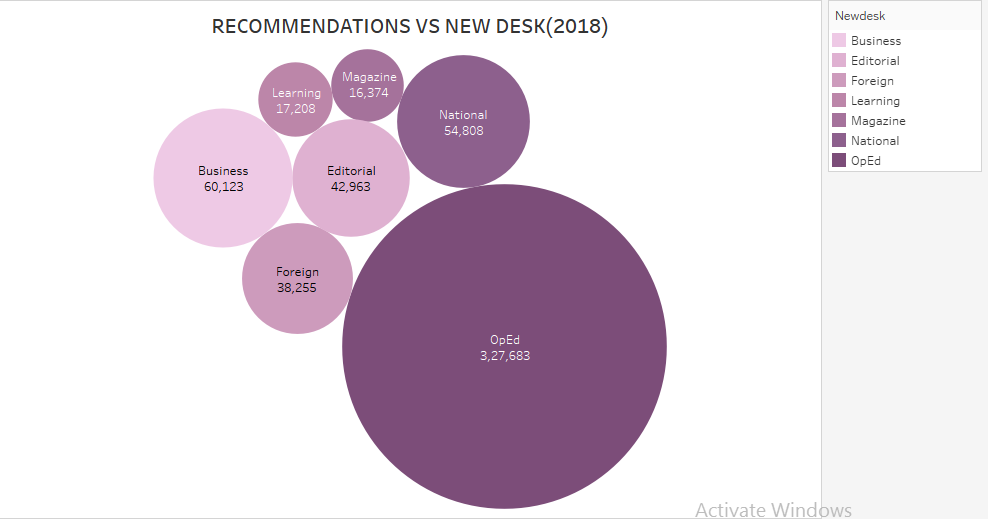


Figure 12

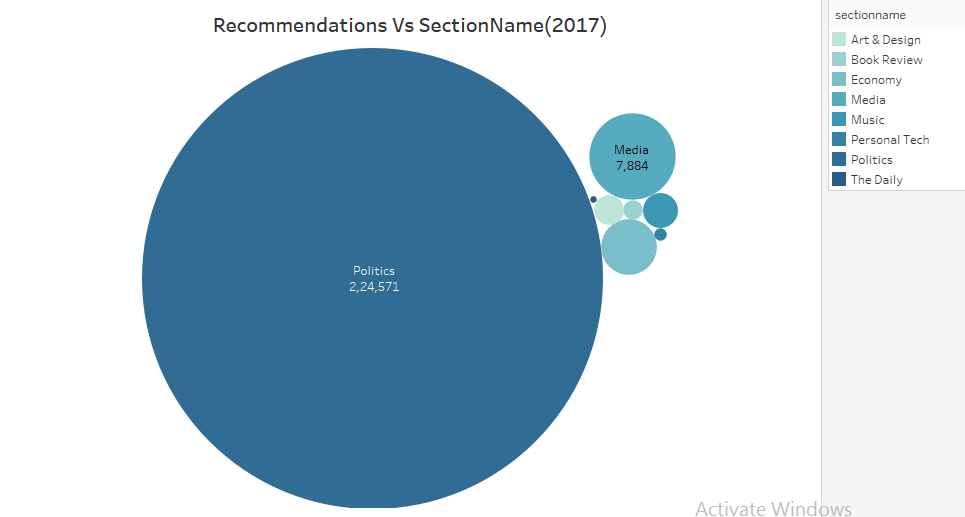


Figure 13

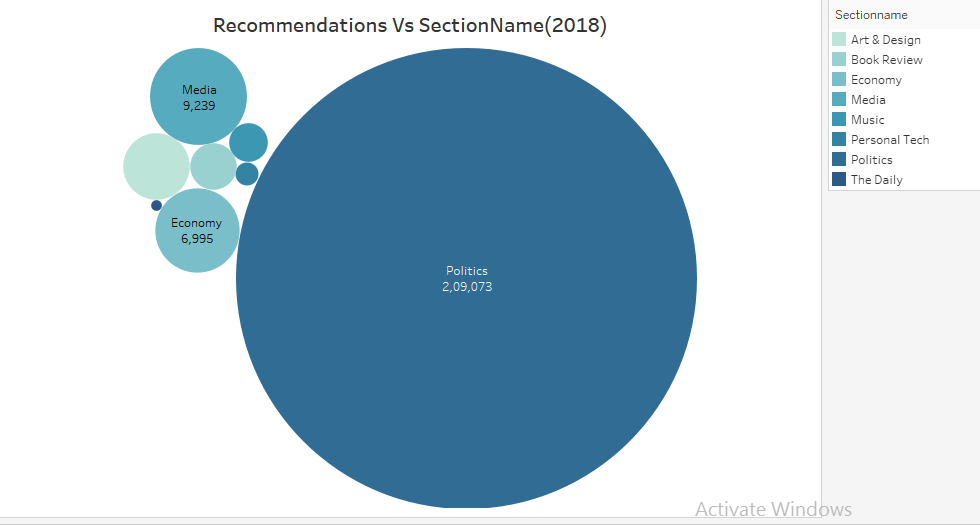


Figure 14

We observed that OpEd which is the opinions provided by the Editorial of the Newspaper were highly recommended by people for articles and Politics for blogposts.

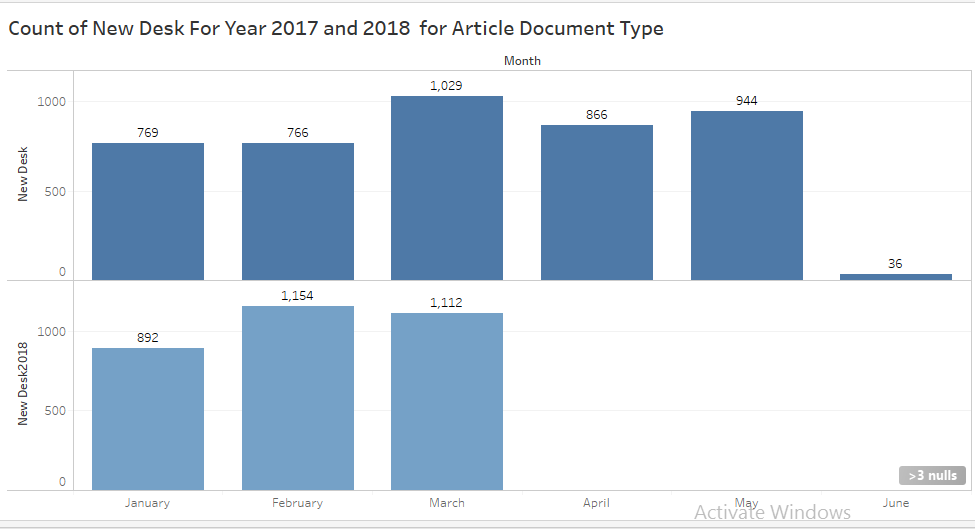


Figure 15

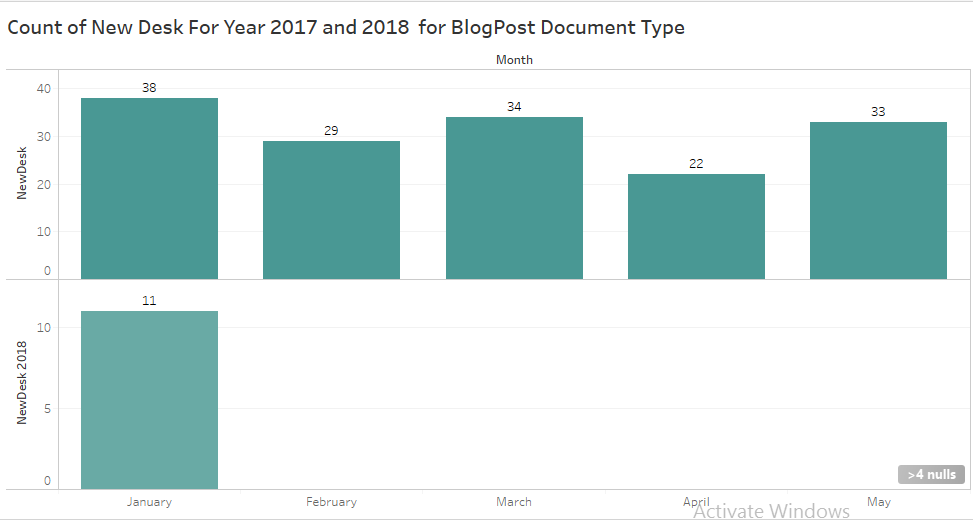


Figure 16

March was the month where New Desk for both the years were highly published for the article document type and January was the month where New Desk for both the years were highly published for the Blogpost document type.

**2.6 Reply Count by User Locations for Year (2017,2018).**

We analyzed to know from which state of United State users were most active and providing the greatest number of replies. For the year 2017 we found the below results.

* NY: 30,638
* Cal: 15,273
* Chicago: 12271

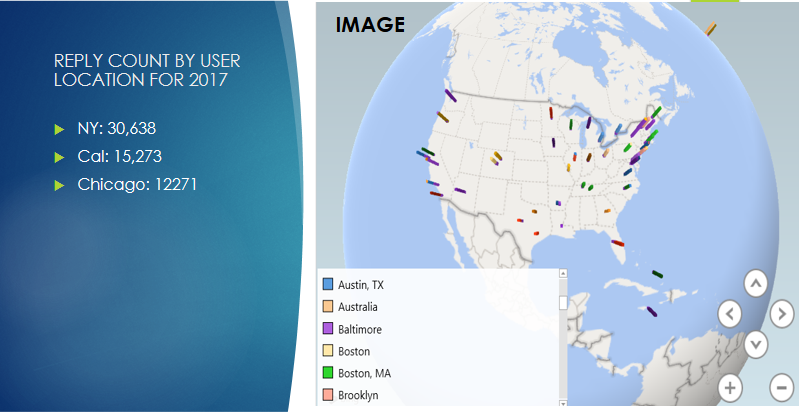


Figure 17

For the year 2018 we found the below results.

* NY: 13990
* Cal: 4251
* Chicago: 3126

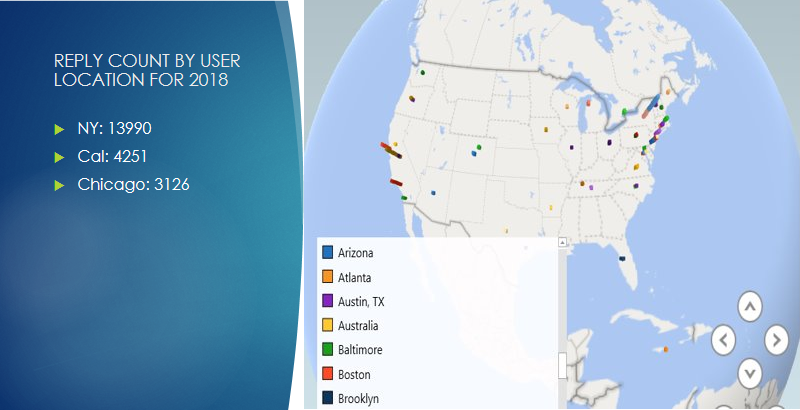


Figure 18

**2.7 Most Popular Author with Respect to Reader’s Recommendations for the Year 2017 and 2018.**

We were intrigued to analyze which authors were highly recommended by the readers in the year 2017 and 2018. We evaluated the below results using the hive queries.

* Highly Recommended author 2017: Nicholas kristof
* Highly Recommended author 2018: Susan Chira

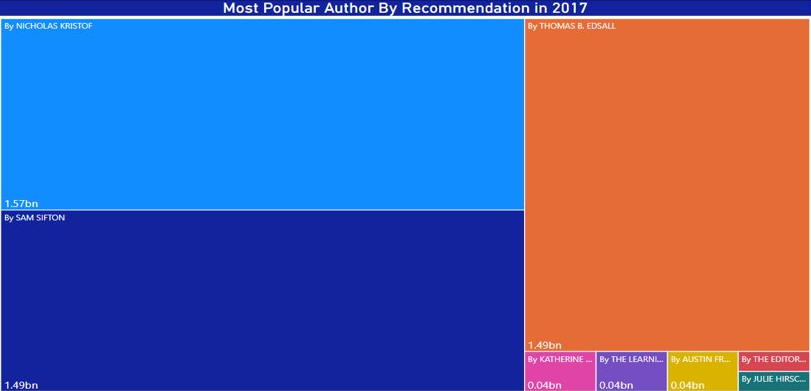


Figure 19



Figure 20

**3.Conclusion**

While exploring the New York Times Dataset, we successfully used Hadoop, HiveQL, 3D -Maps and Tableau to store and manipulate the data in order to gain the maximum insights from it. We analyzed the dataset, being provided with the data for 2 years, that is, 2017 and 2018, right from January to June for 2017 and January to May for 2018. We were able to draw conclusions by analyzing the sentiments of people as positive or negative. Also, we found out that the type of materials being published in NYT were more of the ‘article’ type than the ‘blogpost’ types and in that most used type of material for articles was News. We also investigated the top most areas of people’s interest on which they most likely comment as well as determined those topics for each document type: article/Blogpost that received the highest recommendations from the public as well as the users and editors of NYT. Moderators can focus on these categories when moderating comments added by readers. We even interpreted month wise that the articles received much more replies(responses) in the month of March as compared to other months with a significant decrease of replies for the month of May, for year 2017 and for 2018 and we also observed that the greatest number of replies for Blogpost was in March for year 2017 but for 2018 it was only in January and no one commented after that. While querying the data we also uncovered that ‘news’ was the topic that was most talked(read) about in NYT and by investigating the reply counts for user more we found that users from New York are more active followed by California and Chicago. Also we found out the best author for both the years from the articles and comments data.

**4. GitHub Link**

1. https://github.com/shradhacsula/5200-project

### 5. References

[1]<https://www.kaggle.com/aashita/exploratory-data-analysis-of-comments-on-nyt/notebook>

[2]<https://towardsdatascience.com/predicting-popularity-of-the-new-york-times-comments-part-1-d32f26261f6f>

.