CSE 587 – Data Intensive Computing

Problem 2: Exploratory Data Analysis

Author: Bhavika Jain
bhavikap@buffalo.edu

Objective:

The purpose of this problem was to make us understand the process of Exploratory Data Analysis (EDA) using a data set consisting of details about online readers of New York Times.

New York Times Analysis:

Key variables which were used for analyzing the dataset:

- Age
- Gender
- Signed-In
- Clicks
- Impressions

Age Group:-

In order to do analyses on the age of entire user group who read New York Times online, users are divided into different buckets of age groups based on their age.

1) 0 and below, 0 – 18, 18 – 24, 24 – 34, 34 – 44,
2) 34 – 44, 44 – 54, 54 – 64, 64 and above.

The age group **"-Inf – 0"** indicate those users who have not signed-in and therefore is no age value for them.

Gender:-

As gender information was only available for logged-in users, all the non-logged in users are labelled as **'NOT LOGGED IN'** in the Gender category.

Armed with these details, I proceeded with doing EDA on the New York Times data set for analyzing both daily and monthly trends.

Package Installation:-

Below packages were required to be installed in RStudio :

a) ggplot2: Used for plotting different kinds of plots such as histogram, box plot, scatter plot etc.
b) doBy: Used for doing summary analysis of data such as finding min, max, mean etc.

Author: Bhavika Jain
bhavikap@buffalo.edu

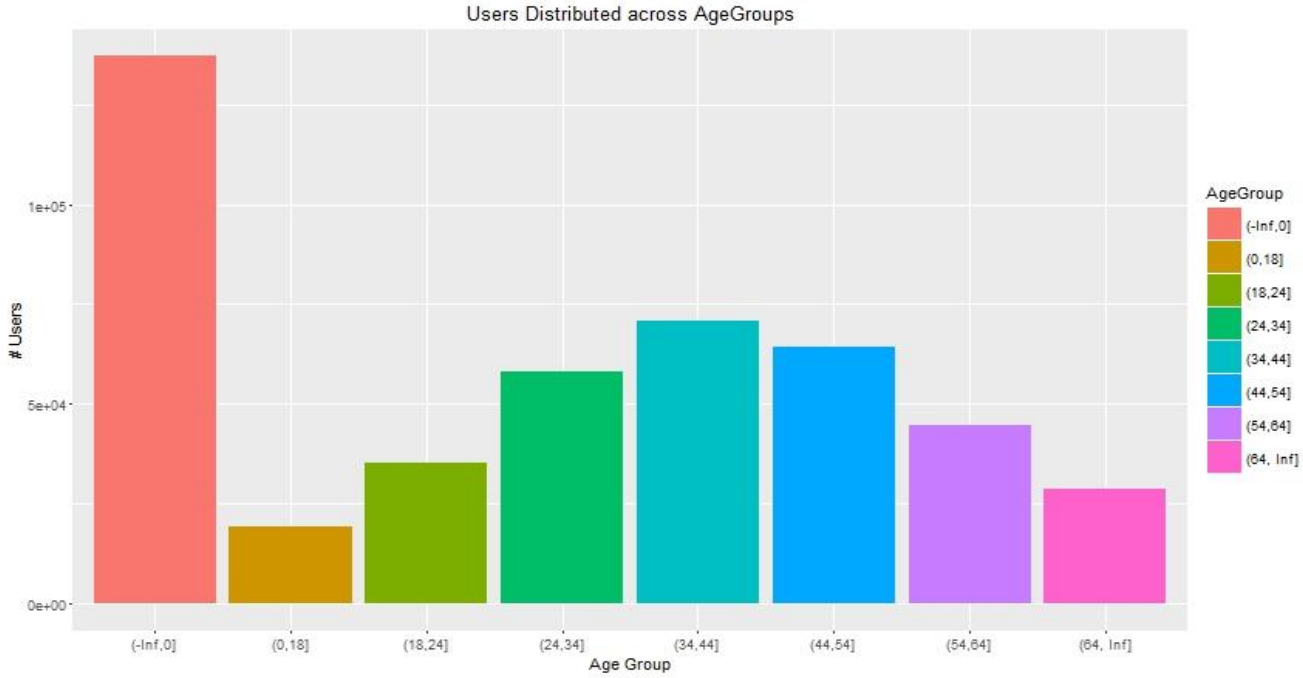
Part A:

In part A, I did EDA on a single data set of New York Times consisting of a single day information about users.

DEMOGRAPHIC/USER DISTRIBUTION:

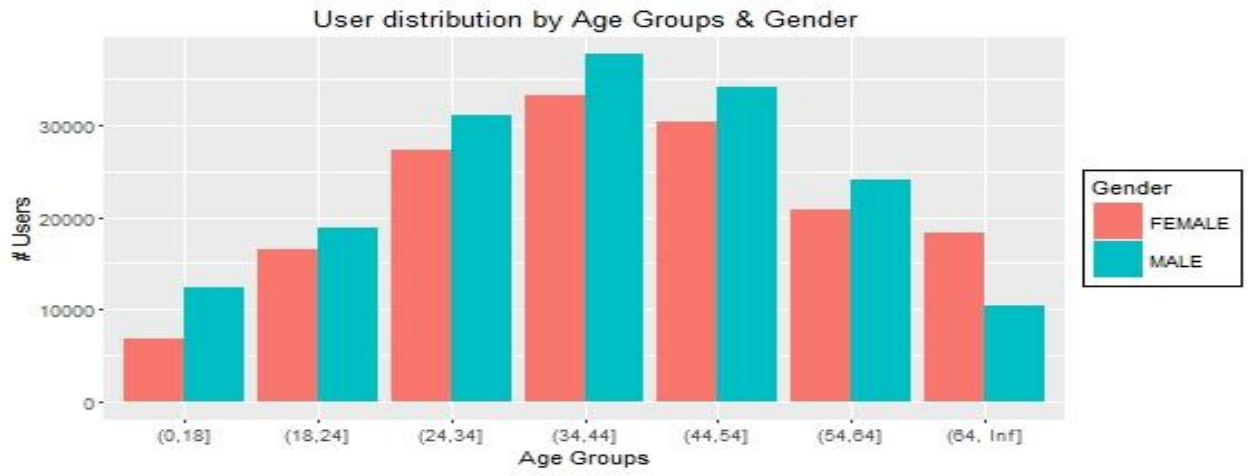
1) ***What is the population of users in each age group visiting the NYT site?.***

As seen from the plot below, NYT site has more visitors with age between 24 – 54 than any other age group.



Author: Bhavika Jain
bhavikap@buffalo.edu

2) *What is the Gender distribution in each of the above age groups?*

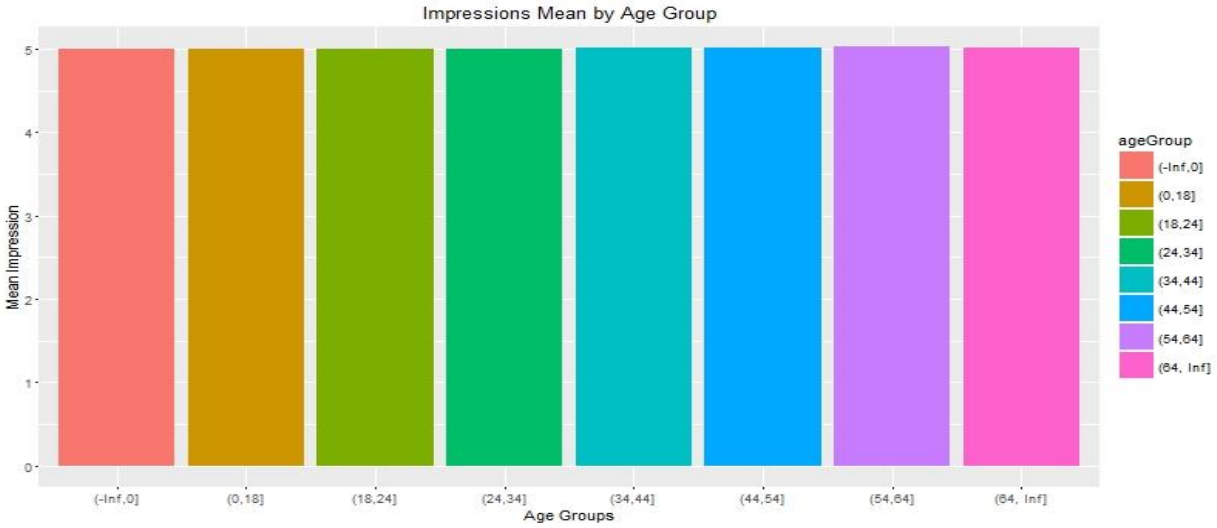


From the plot above, it's evident that NYT site has more male visitors compared to female visitors. Hence NYT site is more popular amongst men of age groups under consideration as shown in the plot.

IMPRESSIONS/CLICK DISTRIBUTION:

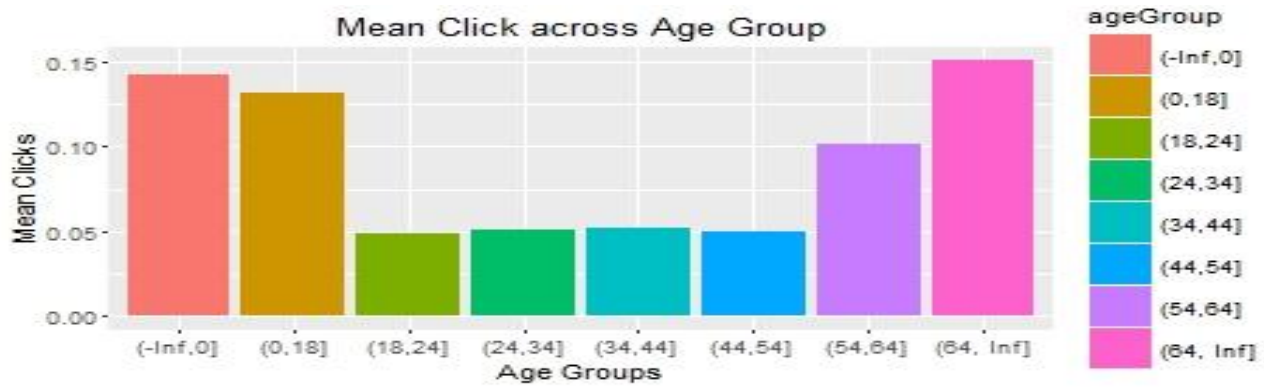
1) *Does NYT's ad engine targets a particular segment of users?*

From the plot below, it's evident that NYT's ad engine does not discriminate amongst its users based on age groups in showing Impressions.



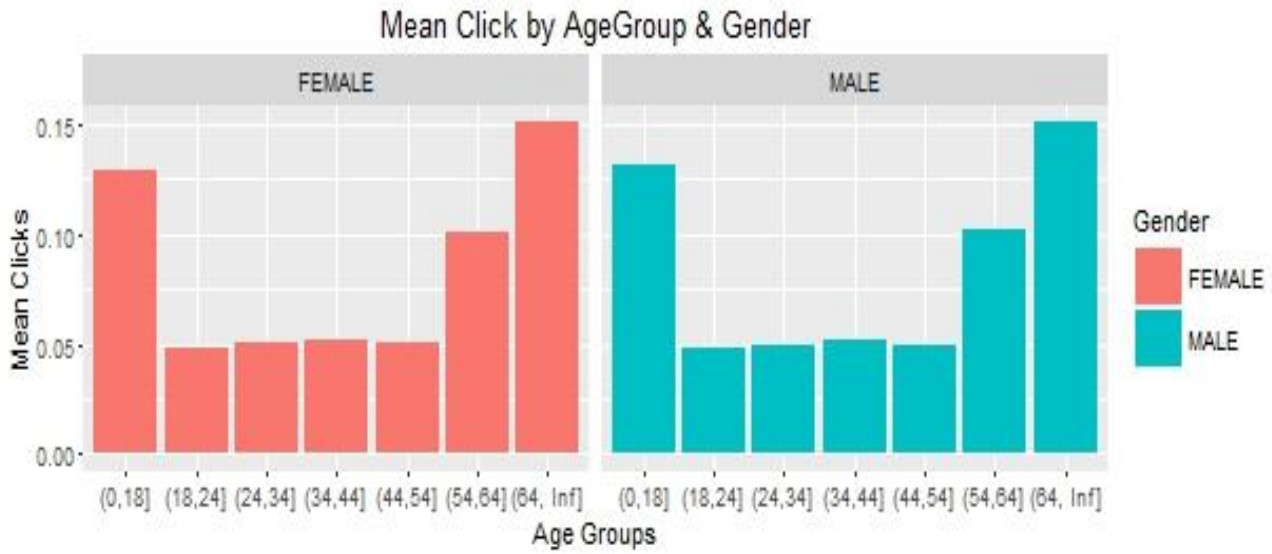
Author: Bhavika Jain
bhavikap@buffalo.edu

2) *How do Users respond to these ads?*



It's evident from that plot above, that there is indeed a pattern amongst users who clicks the most in the age categories under consideration. Users above age of 64 and below age of 18 click more on an average compared to users in other age groups.

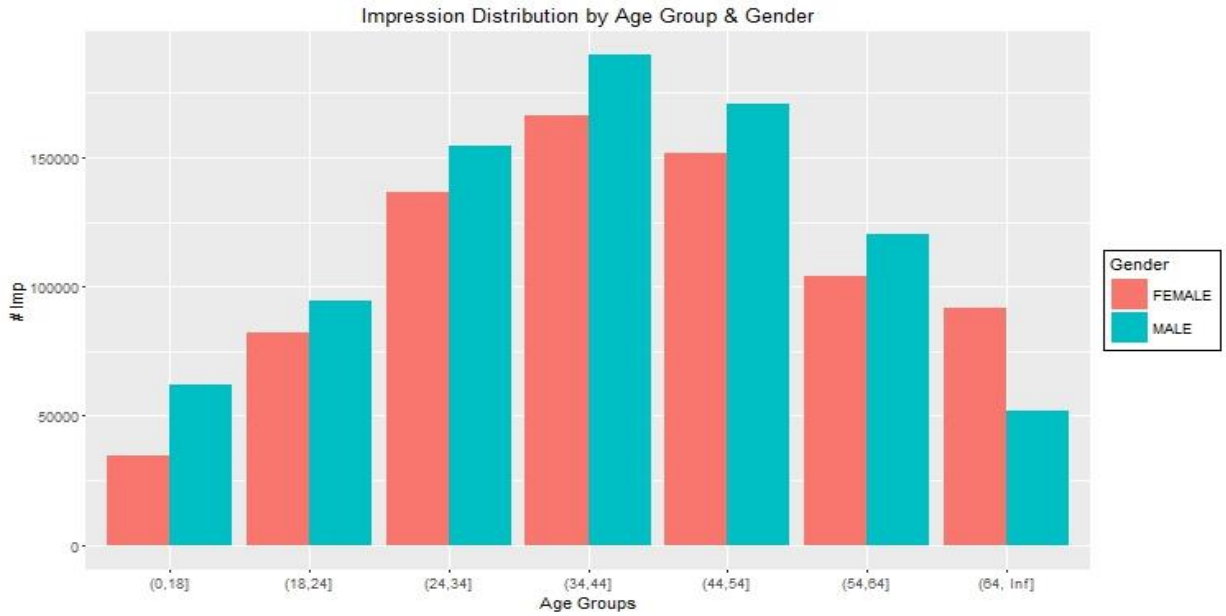
3) *Is there any pattern when Gender is also included for Clicks?*



As evident from the plot above, there is no pattern when gender is included. Males and Females respond equally to the ad-engine.

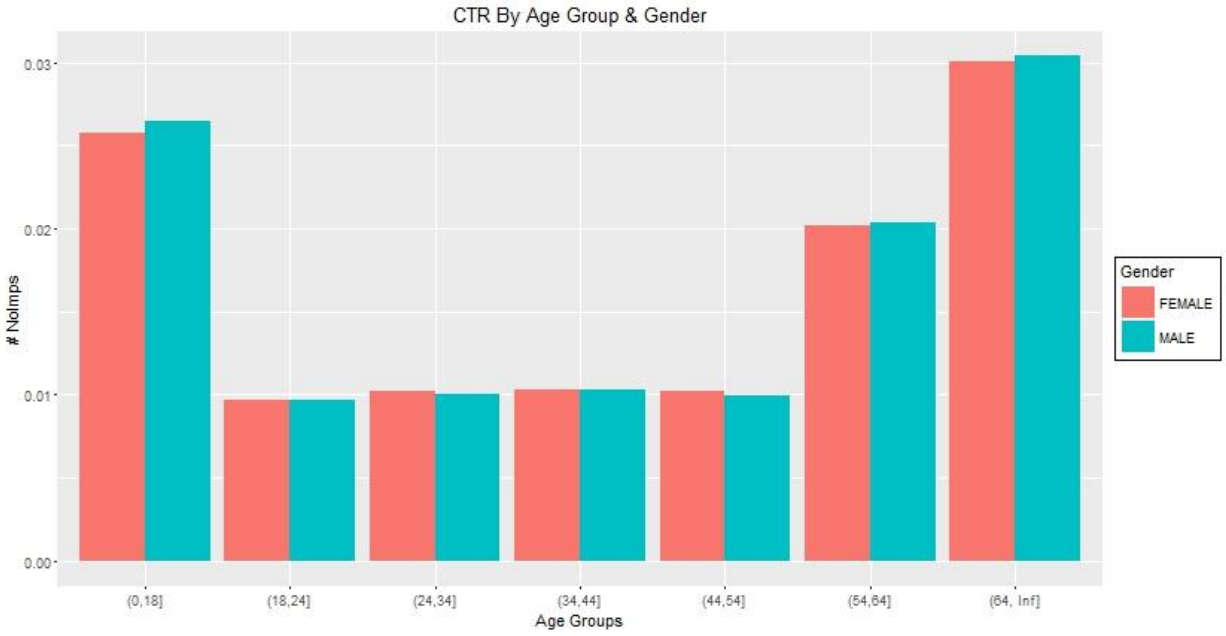
Author: Bhavika Jain
bhavikap@buffalo.edu

4) How are impressions distributed when Gender is also included?.



It is evident from the plot above that as there are more users in the age group of 24-54, the total impression count is higher for them. This behavior is expected.

5) CTR Across Age Groups and Gender



Author: Bhavika Jain
bhavikap@buffalo.edu

As seen from the plot above that the CTR is high among users of age above 64 and below 18. This is expected as the mean clicks is high among users of these age groups.

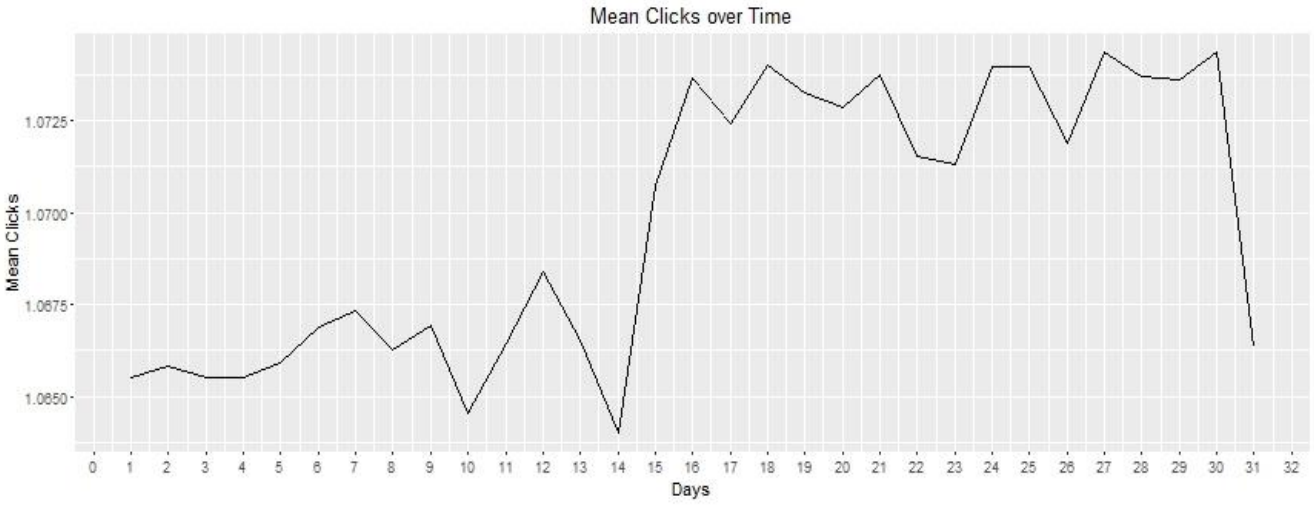
Part B:

In part B, I extended the EDA done for a single data set to entire monthly data set of New York Times.

In order to do find some trends both monthly and daily, I plotted below graphs while doing the EDA.

- ***Average number of clicks per day.***

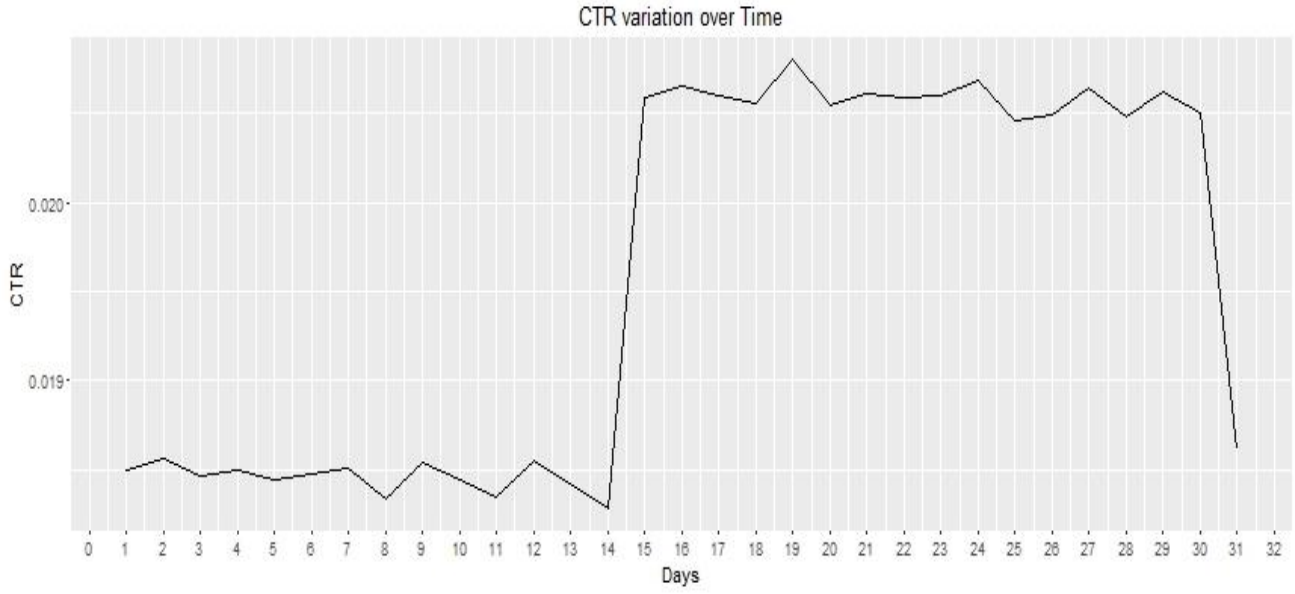
  As seen in the plot below the average number of clicks increased on the 15th day of the month and thereafter there is a sharp decrease after next 15 days.



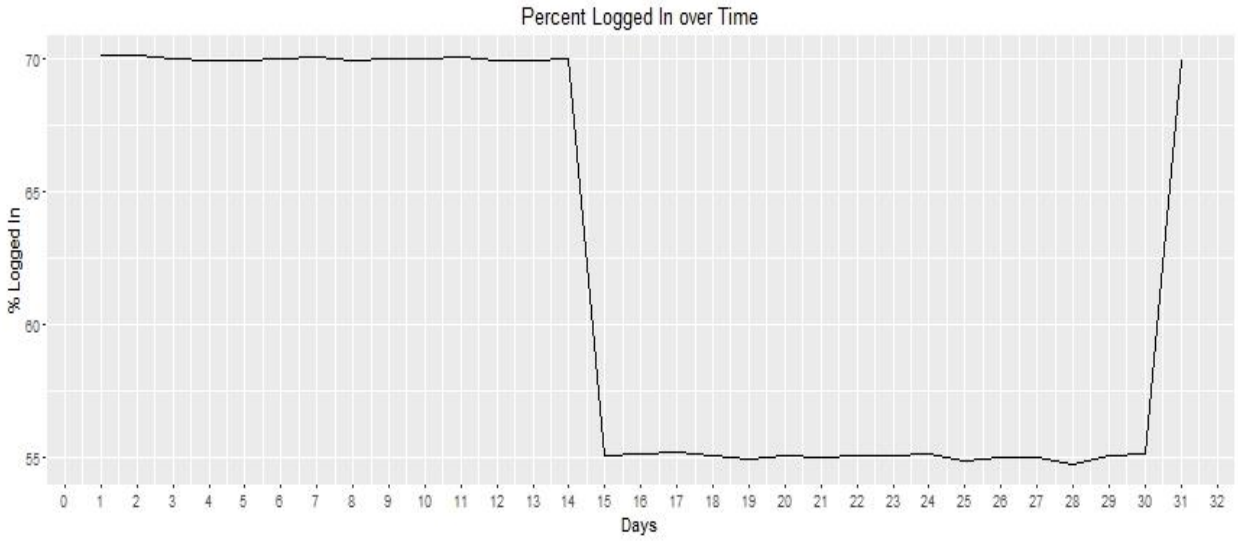
- ***Average CTR per day***

  The sharp increase in CTR on the 15th day of the month follows from the above graph as the number of clicks that day also increased.

Author: Bhavika Jain
bhavikap@buffalo.edu



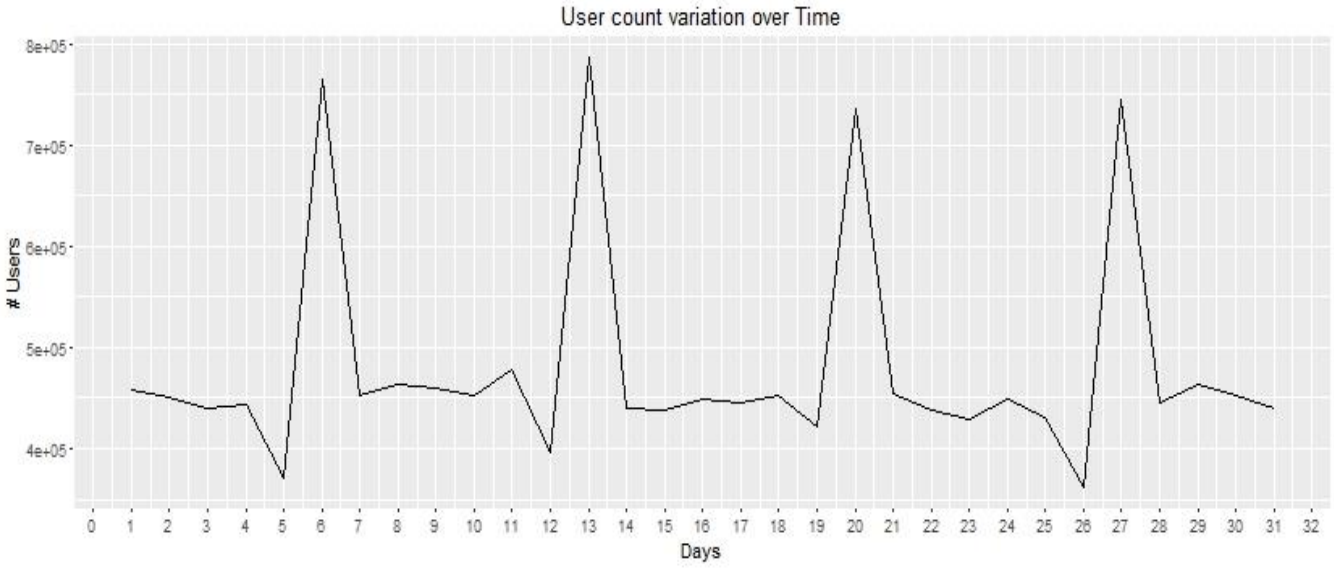
- ***Percentage of users logging per day.***



The reason behind the increase in the average CTR and Clicks on the 15th day of the month is because the number of users logging into the NYT site that day suddenly decreased as shown in the graph below.

Author: Bhavika Jain
bhavikap@buffalo.edu

- ***No of users logging per day***



It is quite evident from the plot above that the user count increases on every 6th day (may be Saturday effect) of the month.

Further Observations:-

- Only in the 65+ age group, there are more females than males.
- In general there more than 300000 users per day.
- On average there are 5 impressions per user.

Conclusion:-

As seen from the plots above, the ad-engine of NYT site does not learn from click behavior of users. The average impression is same across all users whether they are low or high clickers. This is a bad return on investment. The ad-engine algorithm needs to be changed to learn from past history of clicks so that users belonging to low clickers category are not thrown with the same amount of impressions as users in high clickers category. There is lack of algorithmic approach towards targeting users.

Learnings: -

- EDA analysis using R.
- doBy and sub setting data
- ggplot

Author: Bhavika Jain
bhavikap@buffalo.edu