Exploratory Data Analysis on User data for New York Times website

Data Set

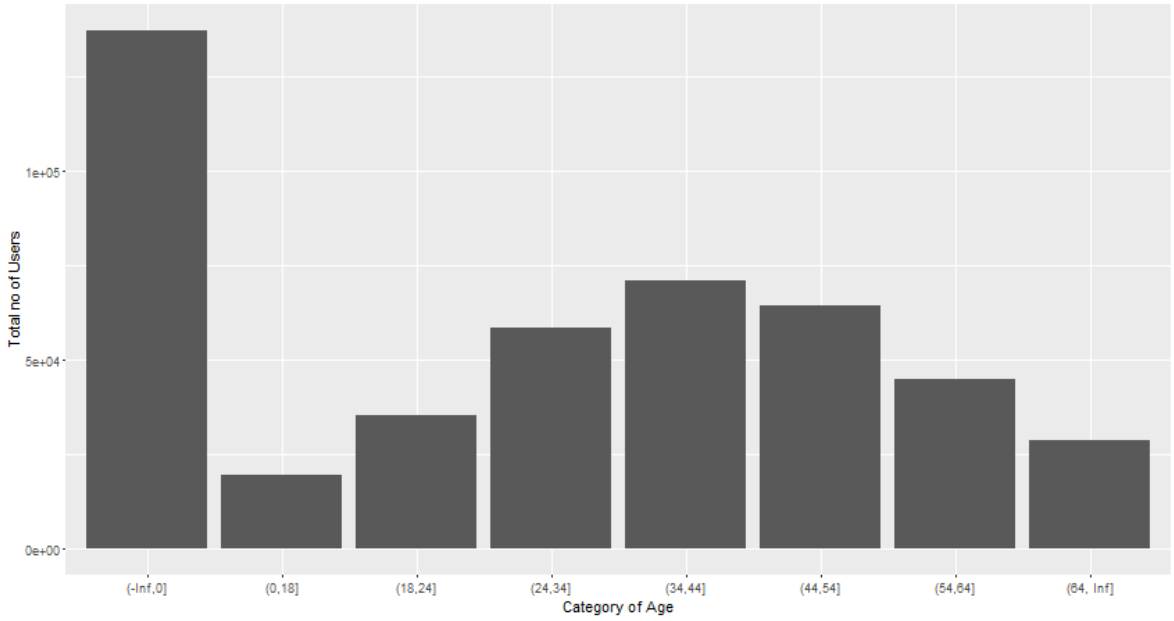
The data used for the analysis was the New York Times user analysis data, with the readings for each day stored as a separate *.csv* file. The data captured user demographics and behavior like, age, clicks, impression, access status (signed in) etc. The data was then analyzed to find relationships between various variables and how the observations derived could help in a better understanding of the data,

1] Analysis on a single day

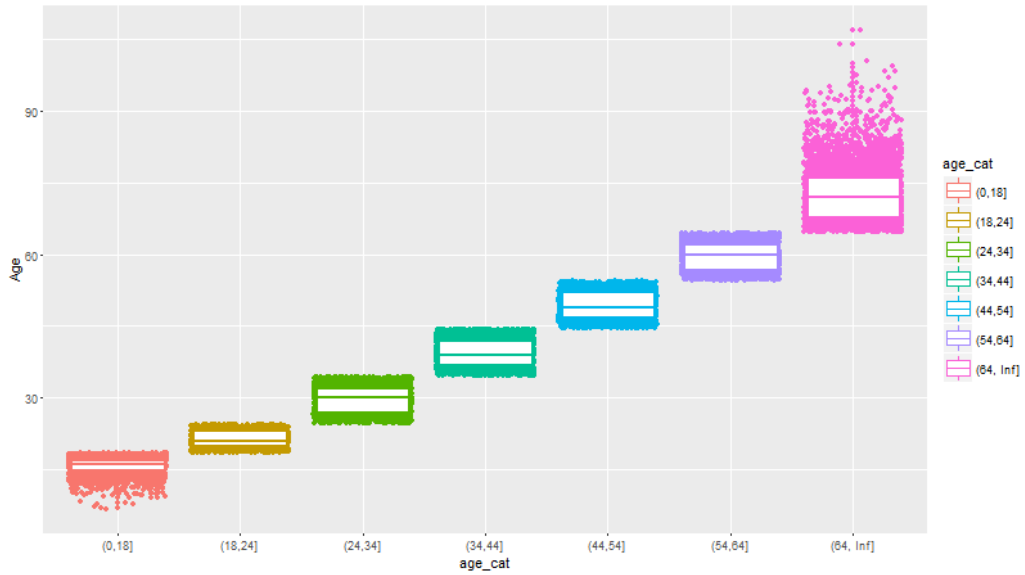
Note: ***impression*** is defined as the number of times a user was exposed to an advertisement and ***clicks*** is the number. Click through ratio is the ratio of ***clicks/impression*** and gives a measure of impact of the advertisement impression on the user behavior. Hence as a measure of user behavior in response to advertisements, clicks are more significant than impressions. The following observations were made keeping this in context.

The data was first categorized based on the age group of the users and a boolean value was set for the users based on whether the users had a an impression or not.

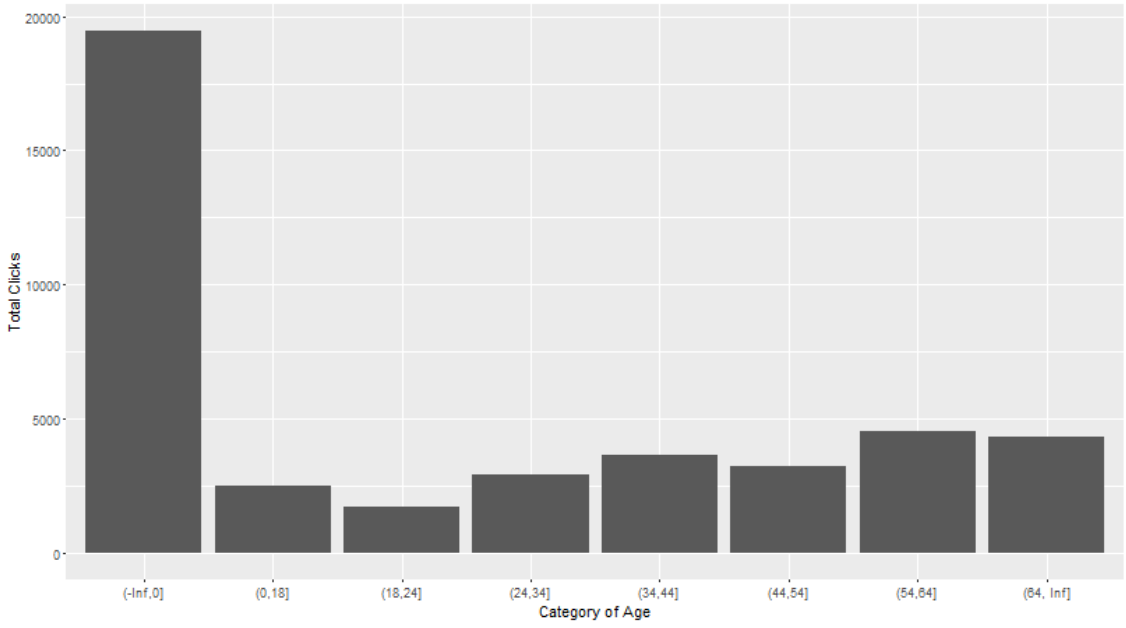
When divided into age groups it was found that the, barring the users that did not specify the age correctly, most of the users came from the 34 to 54 years of age group. This is a useful indication of the age group that is most likely to visit the NYT website and can be used for targeted advertisement tailored for this age group. (*fig 1)*



The overall age distribution is shown, with the horizontal line being the median age within the group.



Using this observation, if the no of clicks for the age category are analyzed, it is observed that the behavior is similar for the number of clicks within the age groups



The mean click through ratio for the different age category was observed as:

	age_cat	nyt_data$ctr.mean
1	(-Inf,0]	2.608588
2	(0,18]	2.566746
3	(18,24]	2.555174
4	(24,34]	2.597655
5	(34,44]	2.620660
6	(44,54]	2.592179
7	(54,64]	2.605101
8	(64, Inf]	2.656349

When comparing male user behavior to female user behavior, their individual density distribution for the different age groups were observed as follows.

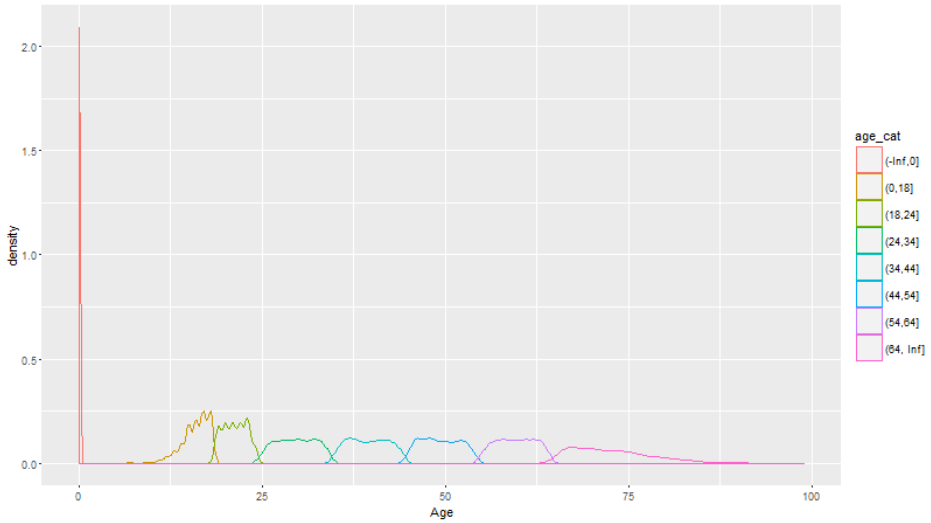


Fig 2. Male age density distribution

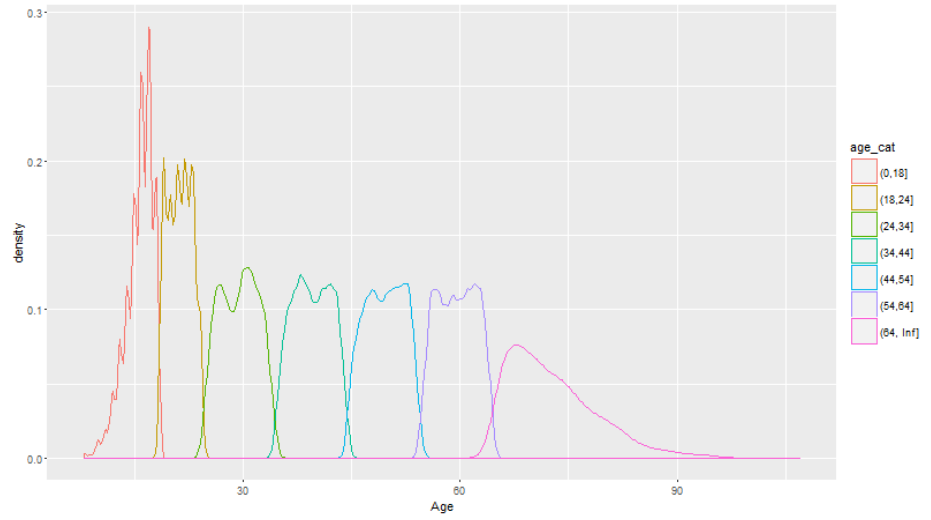
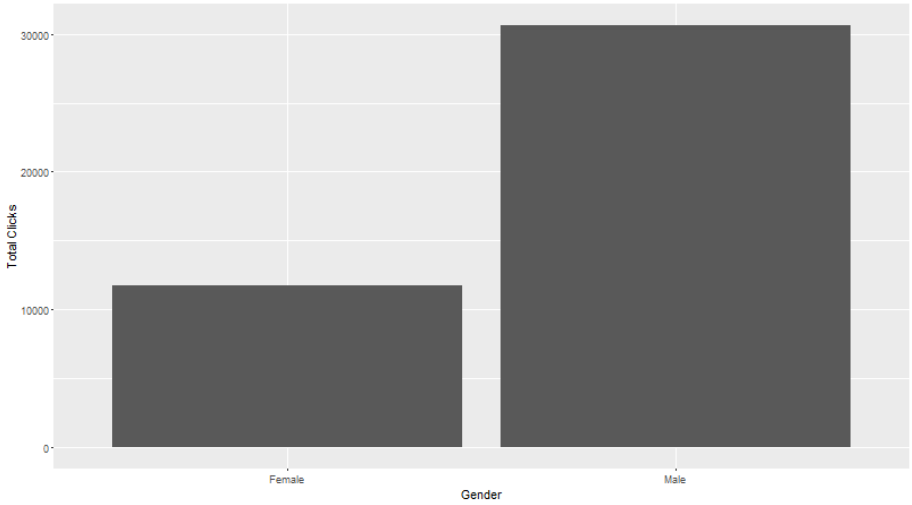


Fig3. Female age density distribution

On comparing the two, it can be seen that for females the age distributions within an age category is more peaked than compared to the male users. This implies that female users are likely to be of similar age within an age category while male users will have more varied ages within the age categories.

When comparing total number of clicks, male users were significantly greater than females.



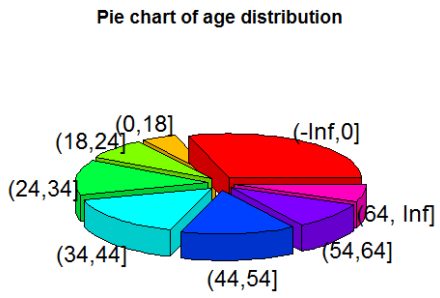
Identifying Bias within the Data

When comparing signed in vs non signed in users, it was found that approx., 41% of the users in the data were signed in (*signed / non signed = .7009*). This implies that the data set **may not** correctly reflect the actual usage of the website as in a more general usage, the proportion of users that are signed in can be lower. Since little is known about the data acquisition process for this set, this bias is inherent within all the observations made.

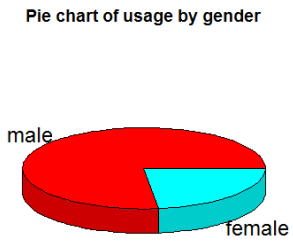
2] Analysis on month data

In order to verify our observations, an analysis of the data for a month was done in order to check whether the observations made in [1] were still true when extrapolated over the whole month.

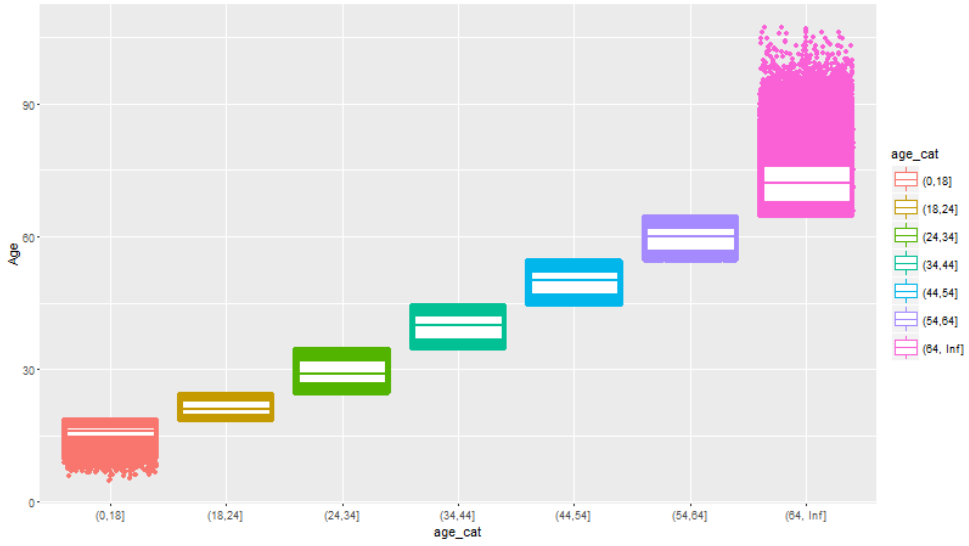
For the monthly analysis, a similar pattern was observed for the total number of clicks noted within each sub group for the user age with the ages greater than 34 showing the highest usage among all the age groups (excluding the outliers.)



While comparing the usage between the 2 genders, similar trend can be seen as observed in the single day analysis.

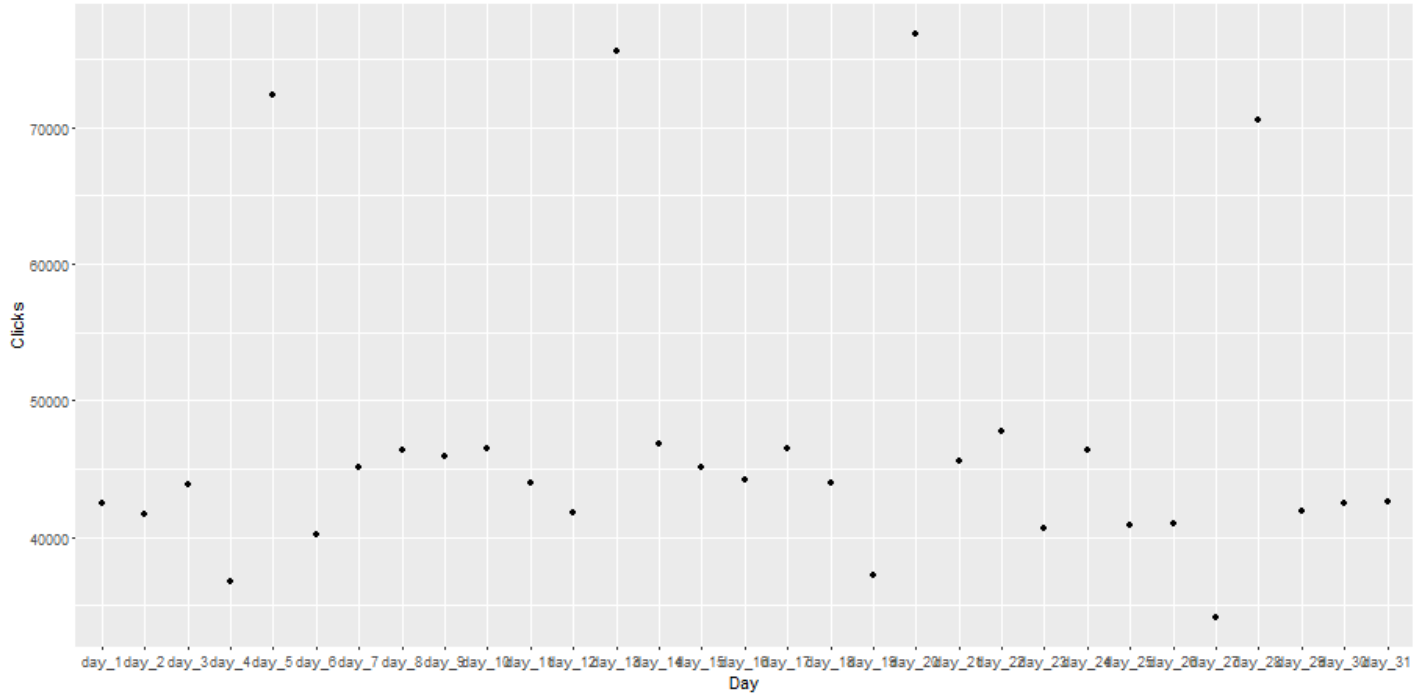


The total age distribution is shown, again the horizontal lines represent the median age in the respective category. Again it can be seen that the [18-24] is the age category with the lowest number of users. This can be used to adjust the content of the website to better suit the demands of this age group as they form the highest number of overall internet users.



Hence the observations made in [1] still hold true when the data is expanded to include a months data.

Observations on the total number of clicks per day for the entire month were:


70000
60000
50000
40000
Clicks
day_1 day_2 day_3 day_4 day_5 day_6 day_7 day_8
Day