Problem 2: Simple EDA

Problem Statement

This problem is aimed towards understanding steps involved in performing Exploratory Data Analysis.

Data Set

We used New York Times dataset to perform EDA. First step in performing EDA is to understand the data. Each row in the dataset represent one user. Following are the details of each user on a given day with the possible constraints on the values of this field.

Age : Shows the Age of Users

Gender : Shows the gender of user. 0 for female users and 1 for male user.

Impressions: As Wikipedia explains, an impression is when an ad is fetched from its source, and is countable.

Clicks : Number of ad Clicks the user did on the given day.

Signed_In : Shows weather the user is signed in user or a guest user. 0 for Guest Users and 1 for Signed Users

Derived Columns

The given information is generally not sufficient to analyze and interpret the data set. So we derive few more columns to ease our analysis.

Age Category(age_cat): Age being a continuous function is difficult to interpret trends. So we define Age categories to help us interpret trends in age groups. I found following age groups meaningful to categories users:

"<18", "18-24", "25-34", "35-44", "45-54", "55-64", and "65+"

Impression Category(scode): We categorize users based on the fact that an ad was fetched for them or not. For users with 0 Impressions we give it value “NoImps” and for those who have some finite value of Impression we give them a vale of “Imps”.

Click Behavior(clickActivity): We categorized user based on the relative number of clicks in a day. If the number of clicks fall under lowest 25% of clicks the we categorized user’s click behavior as low. If he falls under 25% to 75% we categorize his click behaviou as Mid. And for those with number of clicks falls under top 25% are categorized as high.

Click Through Rate: Click through rate is the fraction of people accessed the ad which was fetched. It is defined as ratio is clicks per impression.

Cleaning the data

On giving a cursory look, we can identify that the data is not consistent. So we need to remove the samples that do not make sense. Following re the inconsistencies I found in data.

1. Zero Impressions, Non Zero Clicks: It is not possible because if the ad is not fetched for the user, the he cannot click it. So we should remove such data points.
2. Zero Age: I found that the users who are not logged in have default age of 0. SO while analyzing ages of users we should only consider logged in users.
3. Gender og Not Logged in users: gender of not logged in users is set as 0, i.e. female. So we should also remove the not logged in users while analyzing related to gender.

Analysis

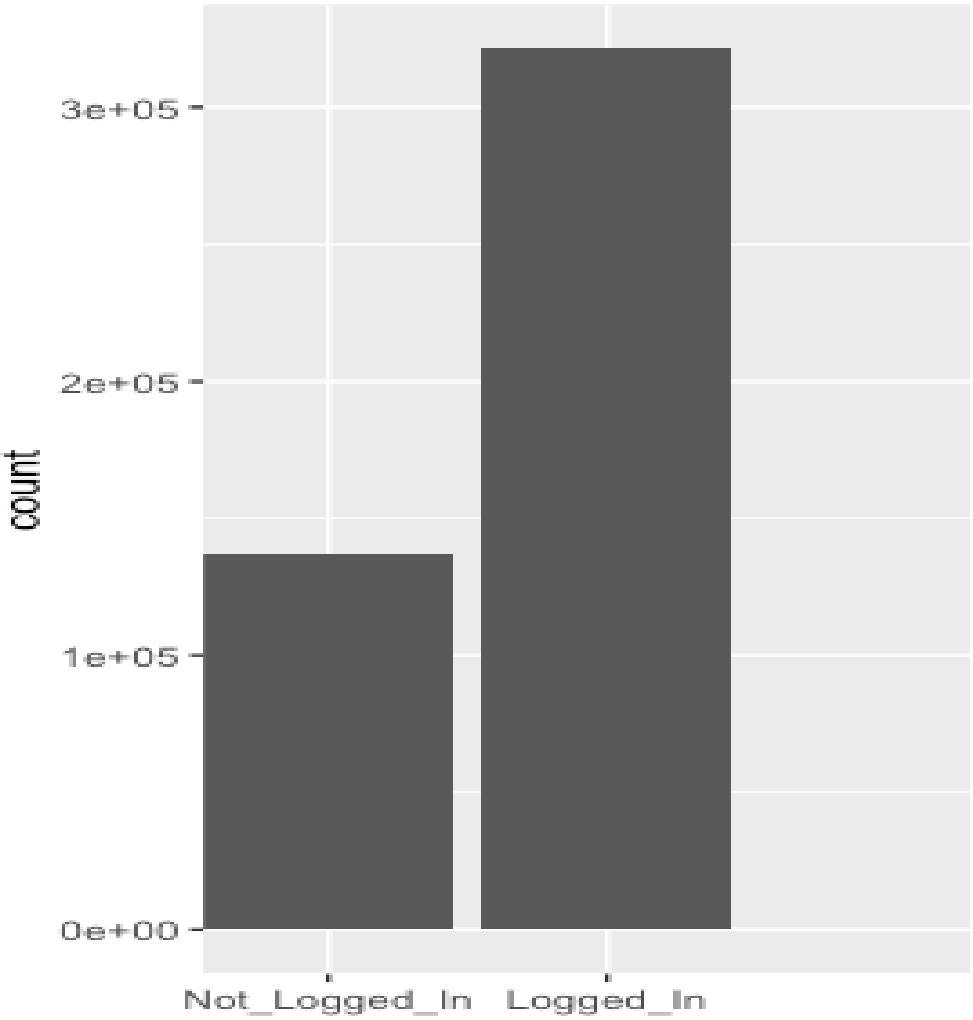
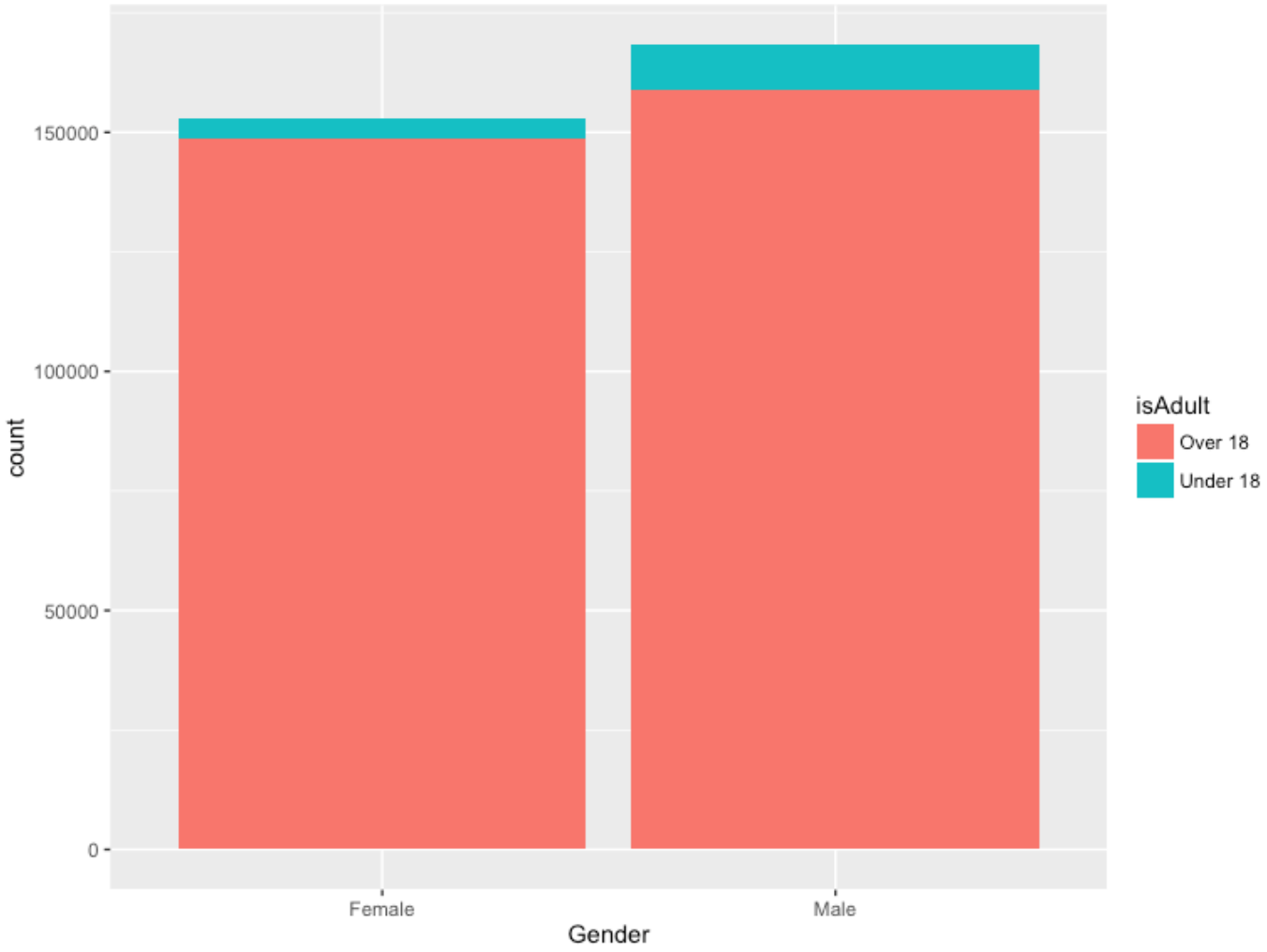


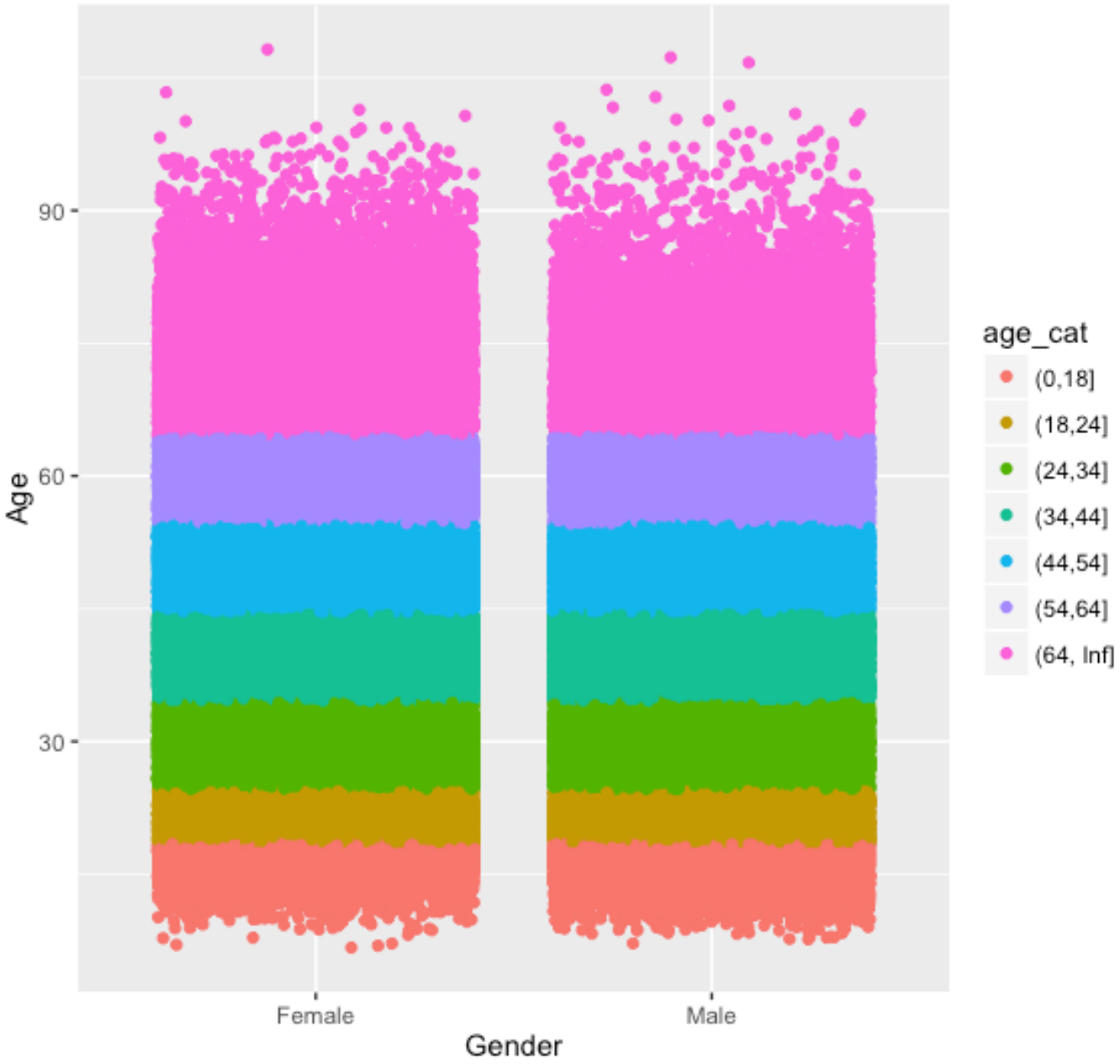
Figure 1: Logged_In vs Guest Users

The above plot shows that almost 2/3rd of users are logged in users. So we can interpret that more than 2/3rd of the readers of New York Times are permanent users.



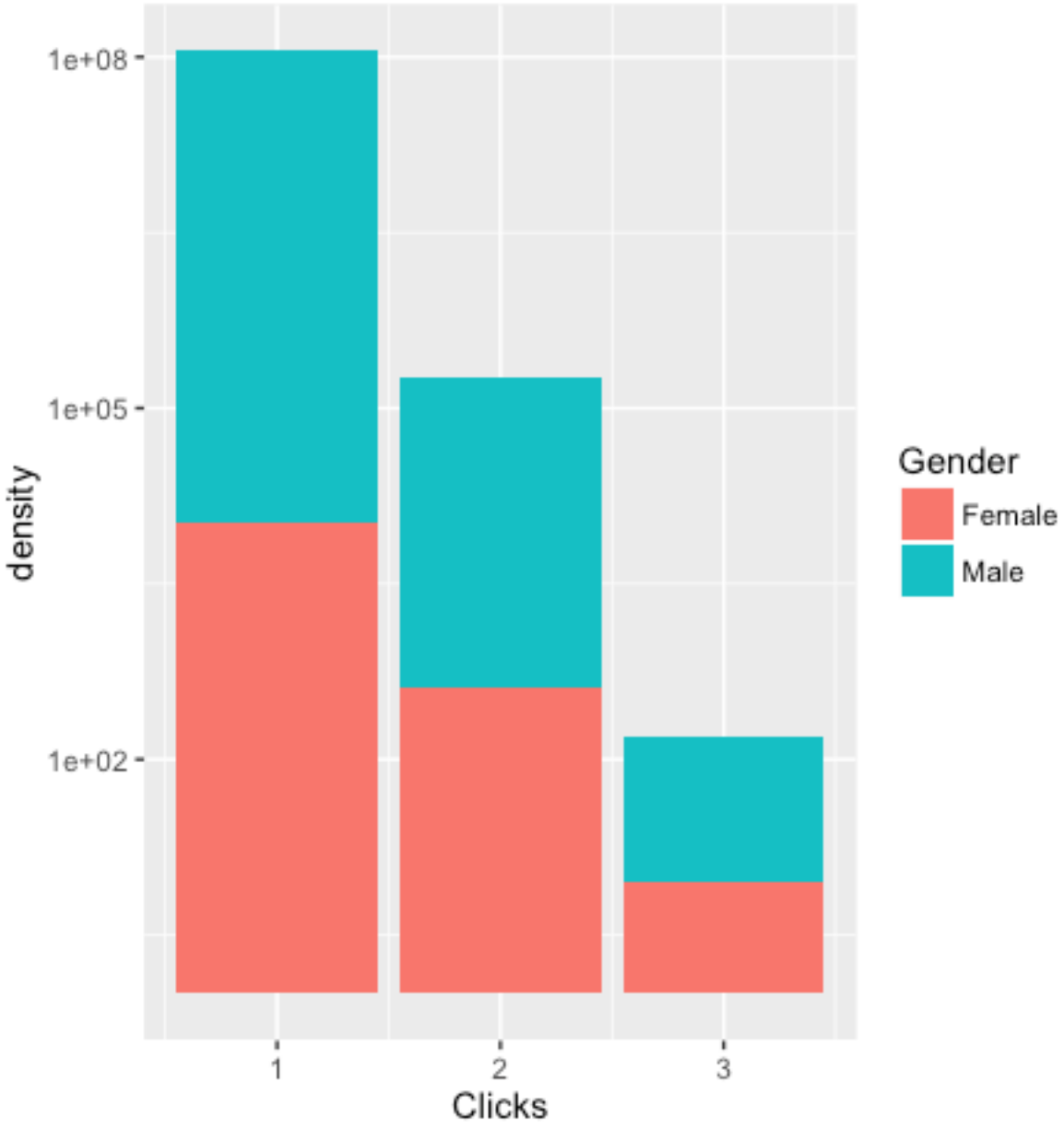
Note that we use subset of data with logged_in user only as the not logged In users have default female gender.

Above graph shows the distribution of logged in users across genders. The graph clearly shows that the number of male and female users are almost same, with male users being on a little higher side. The fill of the graph shows the distribution of ages of users. We ca see that the fraction of under 18 users is less than 10% for both male and female. Another interpretation we can make from the graph is that number of under 18 male users reading NYT is much higher under 18 female users (almost double).



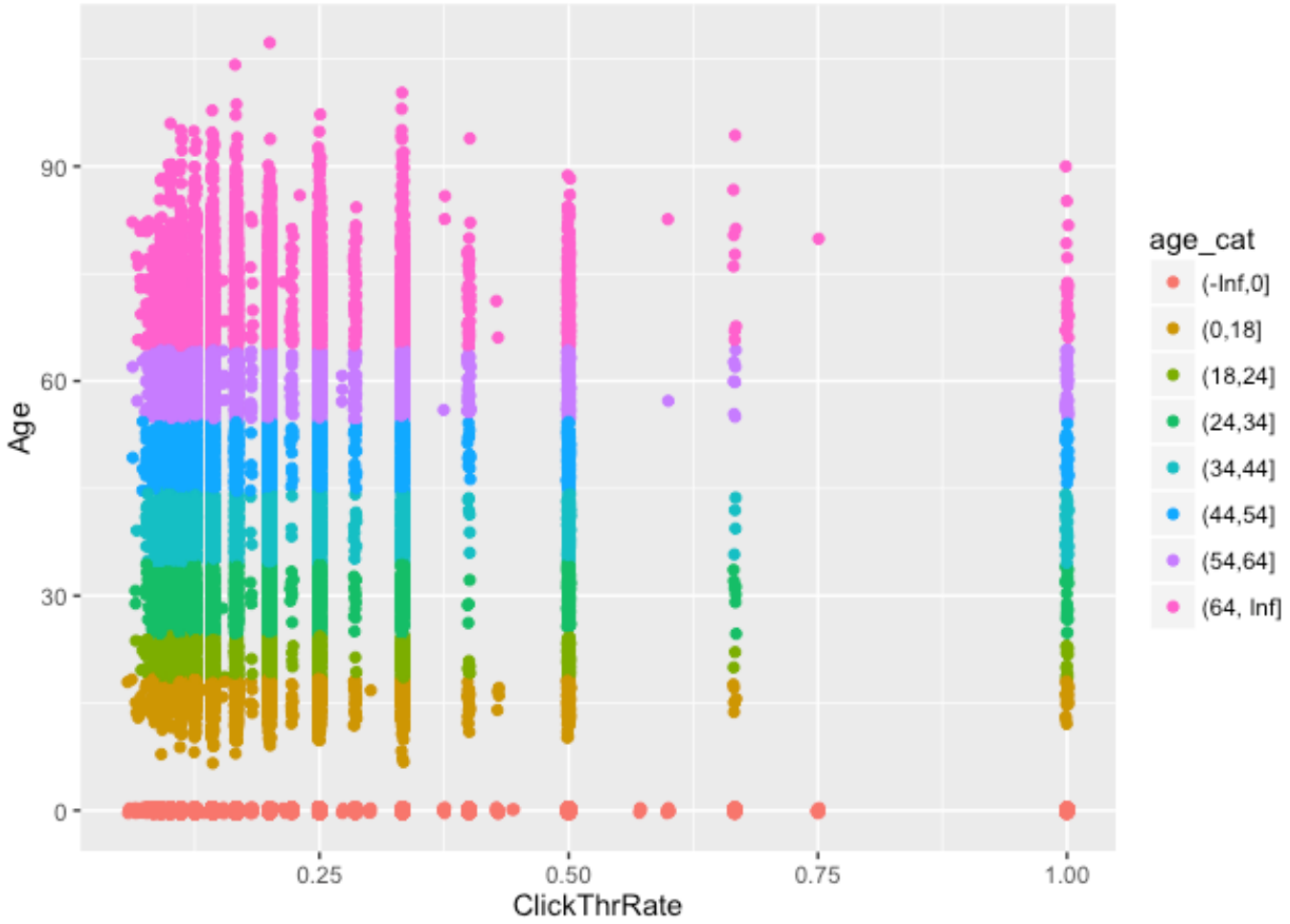
This graph shows that the distribution of Gender of Users over ages. We can see that there are fewer users who are under 18 or 18-24. Almost equal users in age category 24-34, 34-44, 44-54 and 54-64. We have highest number of NYT users over 64 years of age.

For this graph we are taking subset of data with logged in Users only, because not logged in users have zero age and female gender by default



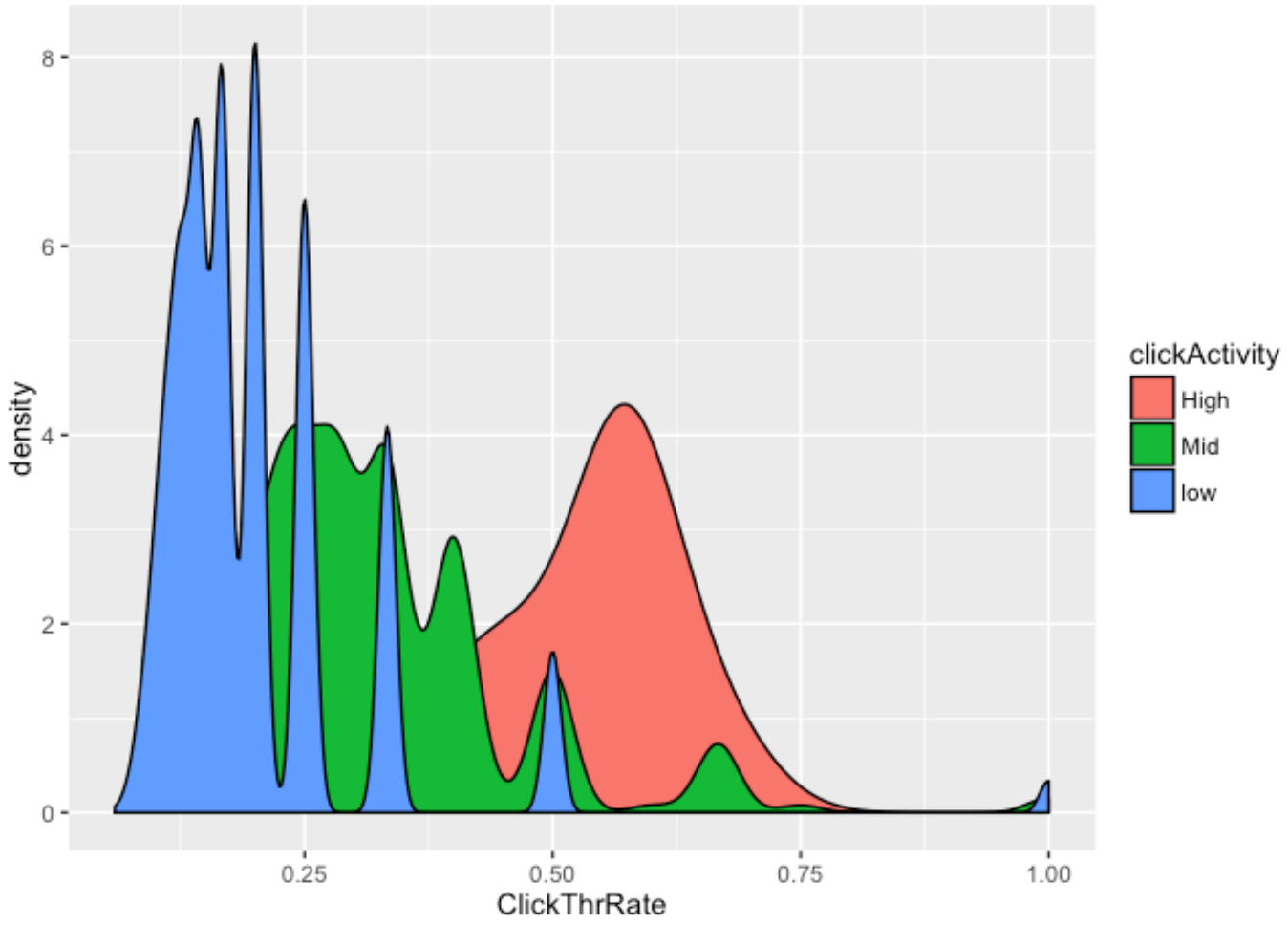
This graph shows, how many users did how many clicks. We cab see that most users did 1 click only. Higher the number of clicks lesser is the number of users. Also we can see that for a given number of clicks the number of males and females are almost same.

We used subset of logged in users, as not logged in user have default fender Female.



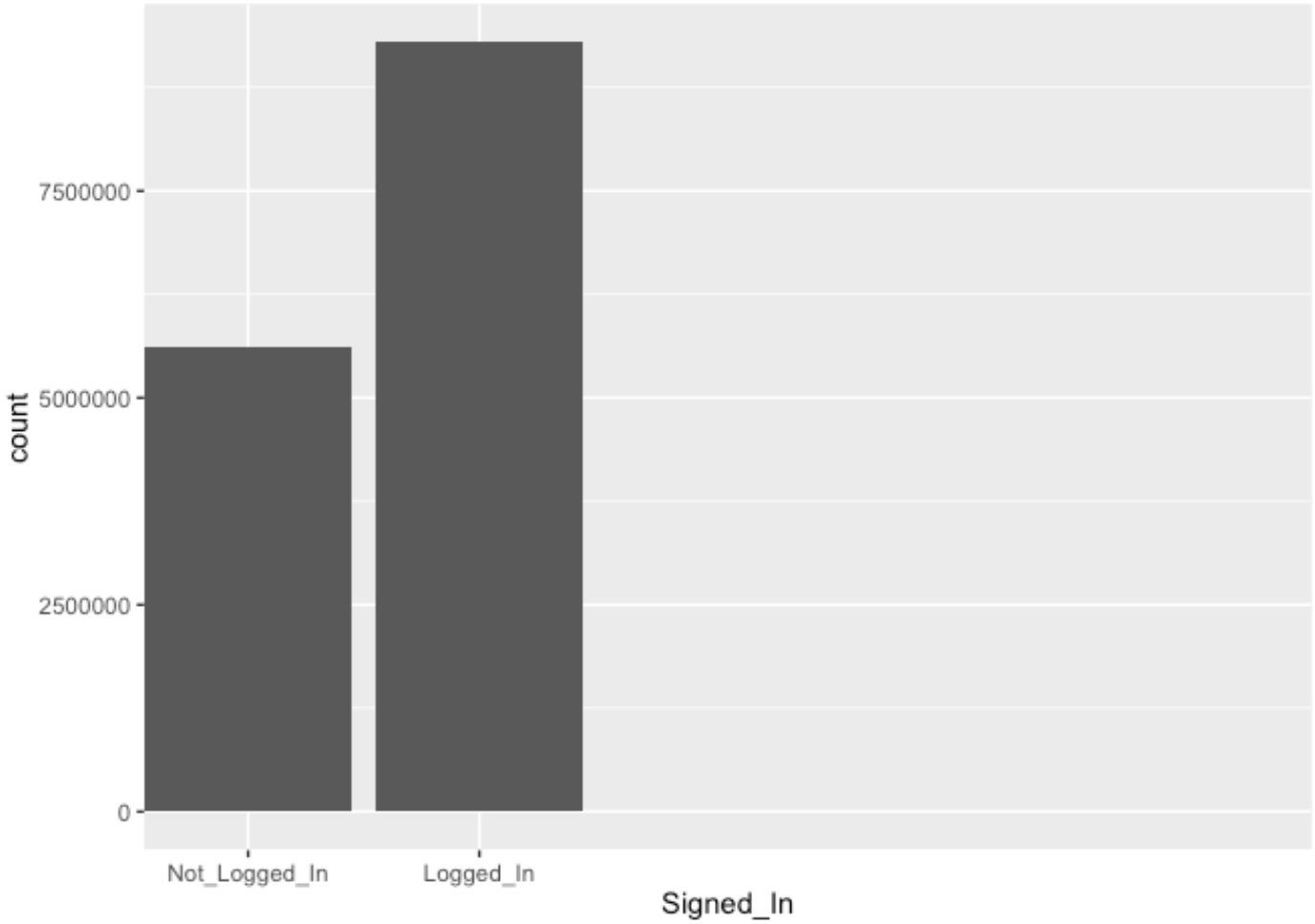
Click through rate is very important measure which shows the activity of a users. It shows the fraction of ad which were clicked by the user.

From the graph we ca see that there is highere density near lower click through rates. That implies that the generally users Click through rate is less than 30%. There are very few people with click through rate of 1, i.e, they click all the plots that is fetched.

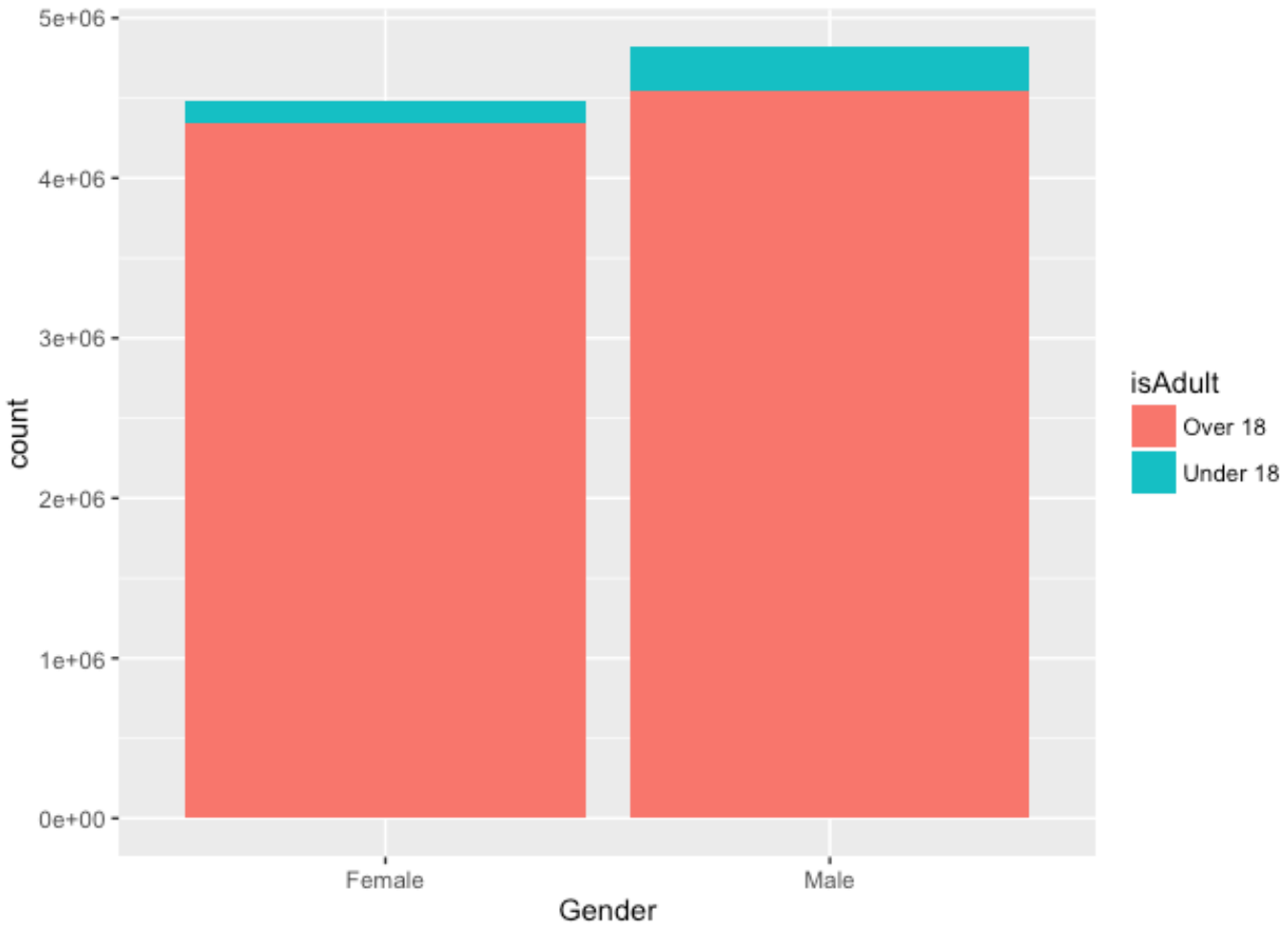


This graph shows that Users with low click activity generally have low % click through rate, less than 25%. The users with medium click activity, generally have click through rate ranging from 20% to 60%. And most of the highly active users have click through rate ranging 40% to 75%.

Analysis Extended to 31 Days

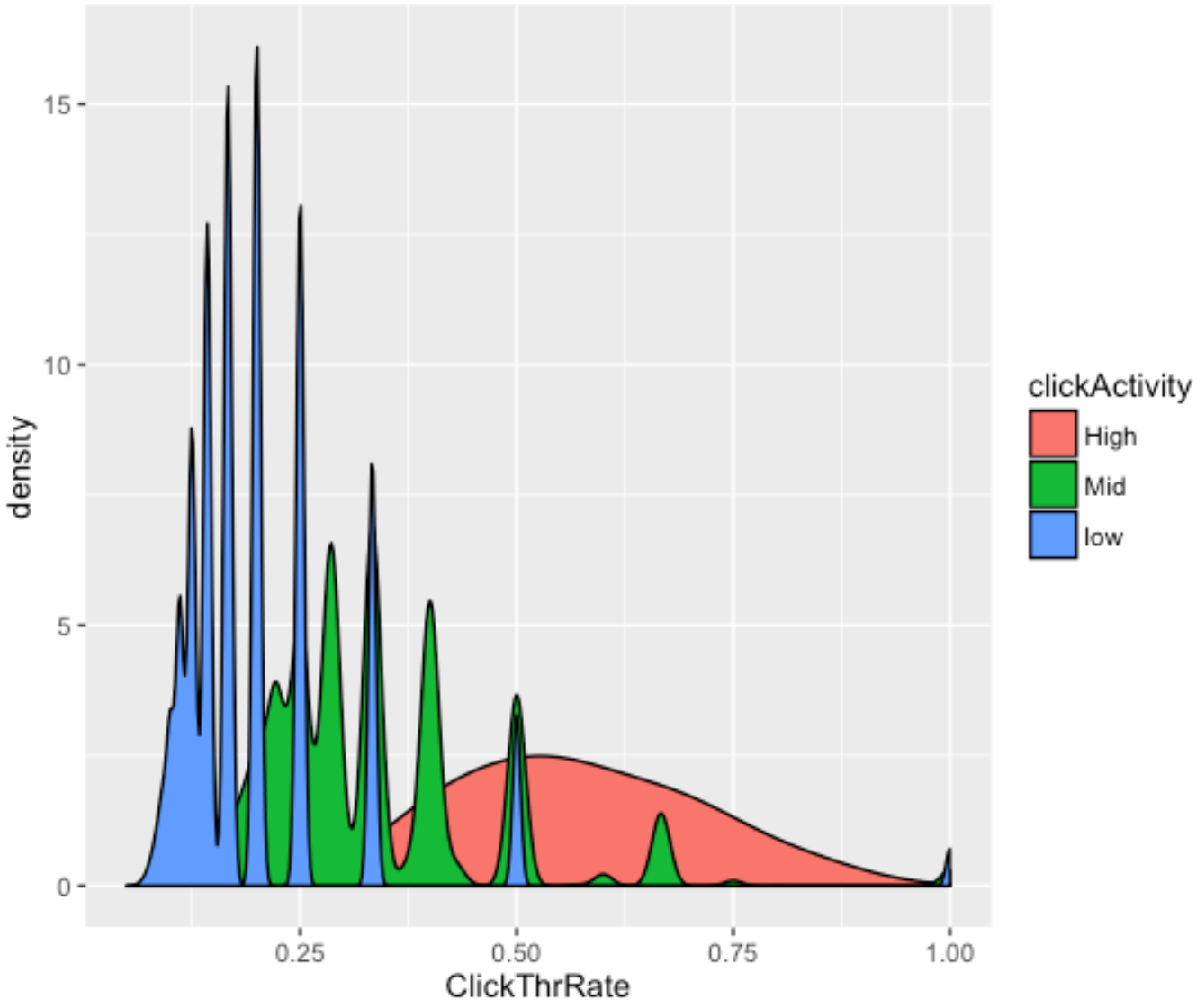


The above plot shows that almost 60% of users are logged in users. So we can interpret that more than 60% of the readers of New York Times are permanent users. As this trend is from 31 days data, we can say that this is more accurate representation of logged in vs Guest Users.

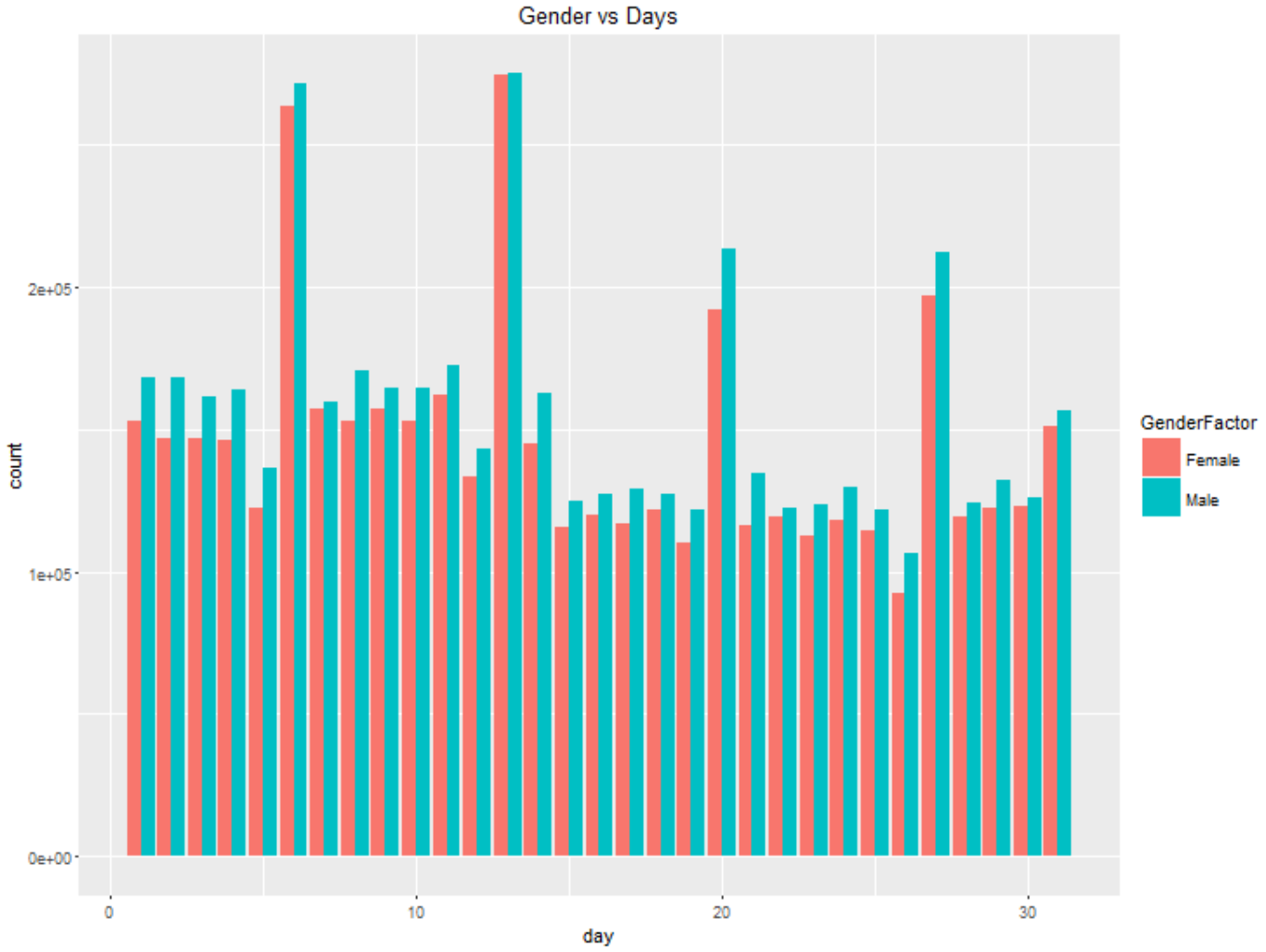


Note that we use subset of data with logged_in user only as the not logged In users have default female gender.

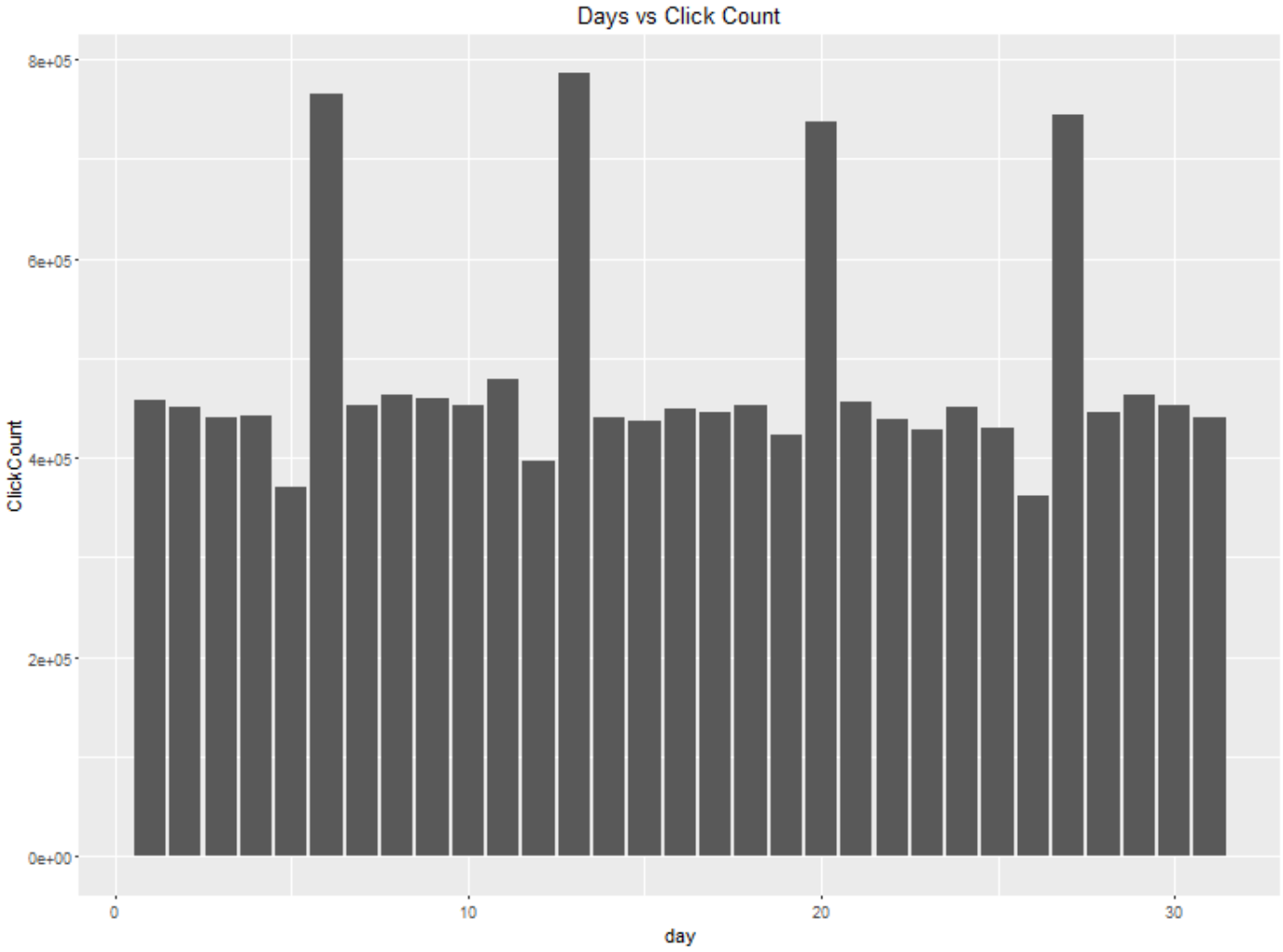
Above graph shows the distribution of logged in users across genders. The graph clearly shows that the number of male and female users are almost same, with male users being on a little higher side. The fill of the graph shows the distribution of ages of users. We ca see that the fraction of under 18 users is less than 10% for both male and female. Another interpretation we can make from the graph is that number of under 18 male users reading NYT is much higher under 18 female users (almost double).



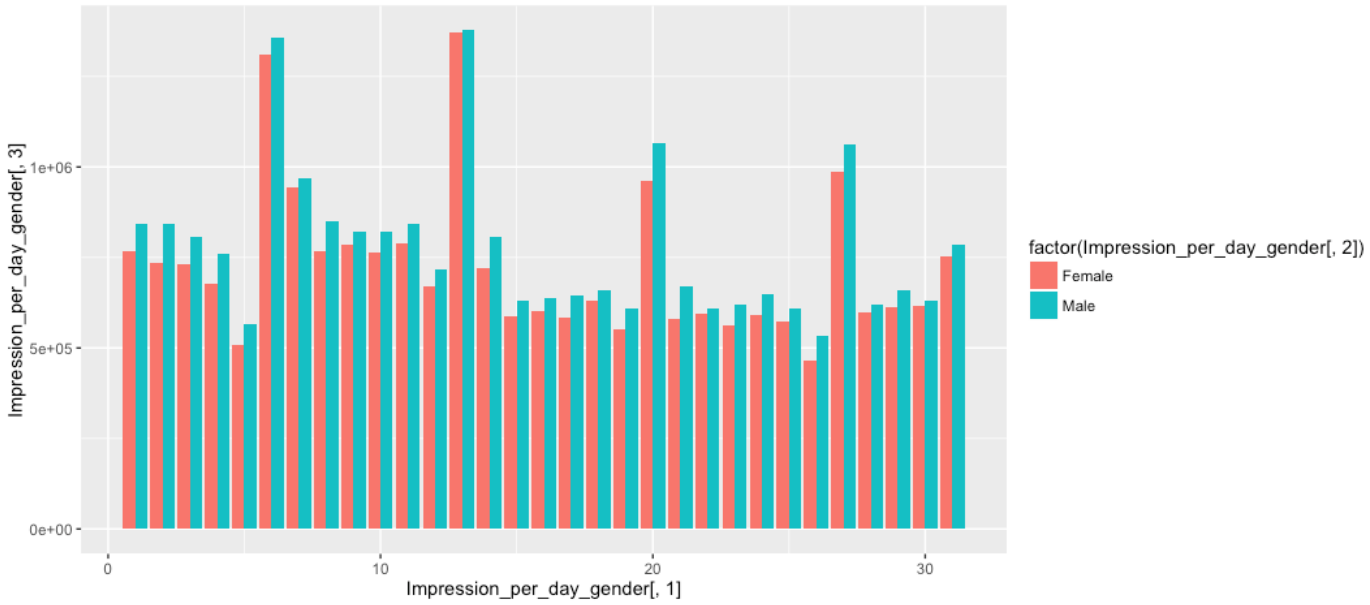
This graph shows that Users with low click activity generally have low % click through rate, less than 25%. The users with medium click activity, generally have click through rate ranging from 20% to 60%. And most of the highly active users have click through rate ranging 40% to 75%. This result is very similar to our result for 1 day.



This graph shows the distribution of Logged In users over dates. We can see that there are sudden peaks in numbers of users. These peaks may weekends when users are more active on news portals.



This graph shows the number of Clicks distribution over days. Again here we can see that during weekdays click are almost same. And The suddenly shoot up during weekend.



This graph shows number of ad fetched per day. Again here we see that the activity of users is more on weekends than on weekdays. So more ad are fetched during weekends. So Using this information the NYT can increase the rate of their ad on weekend.

Conclusion

We can conclude that after cleaning the data we can see some patters which nicely relate to the reality. Like less under 18 users than Adult users, which is true in our social setup. Because children are generally not into reading news and prefer reading comics. Also we saw that users are more active on weekends, which is true given out service class society. Most of the users are busy on weekdays with their offices.

Using this analysis, we NYT can choose their rates, based on the age group the ad targeting and and time of the week.