REPORT

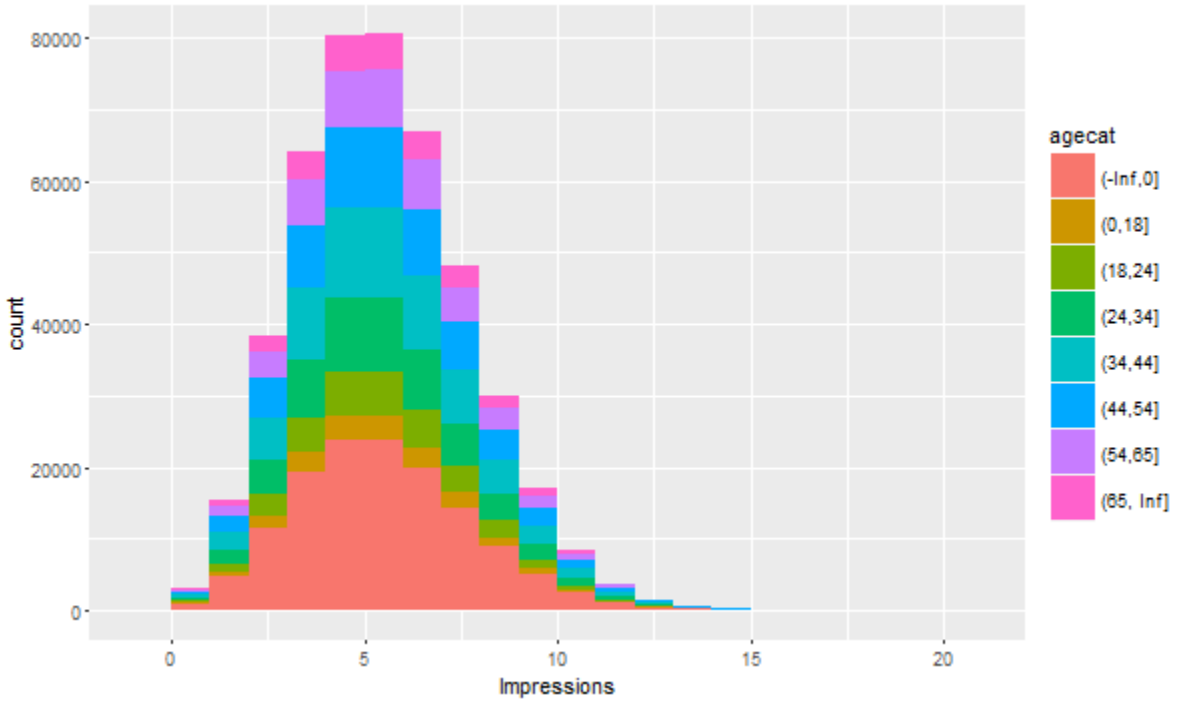
Prob 2: Importing and EDA on nyt data using the description in book and extending the EDA to all the days of the month and making some Inference

Part 1: Importing and EDA as given in the textbook on nyt1 data

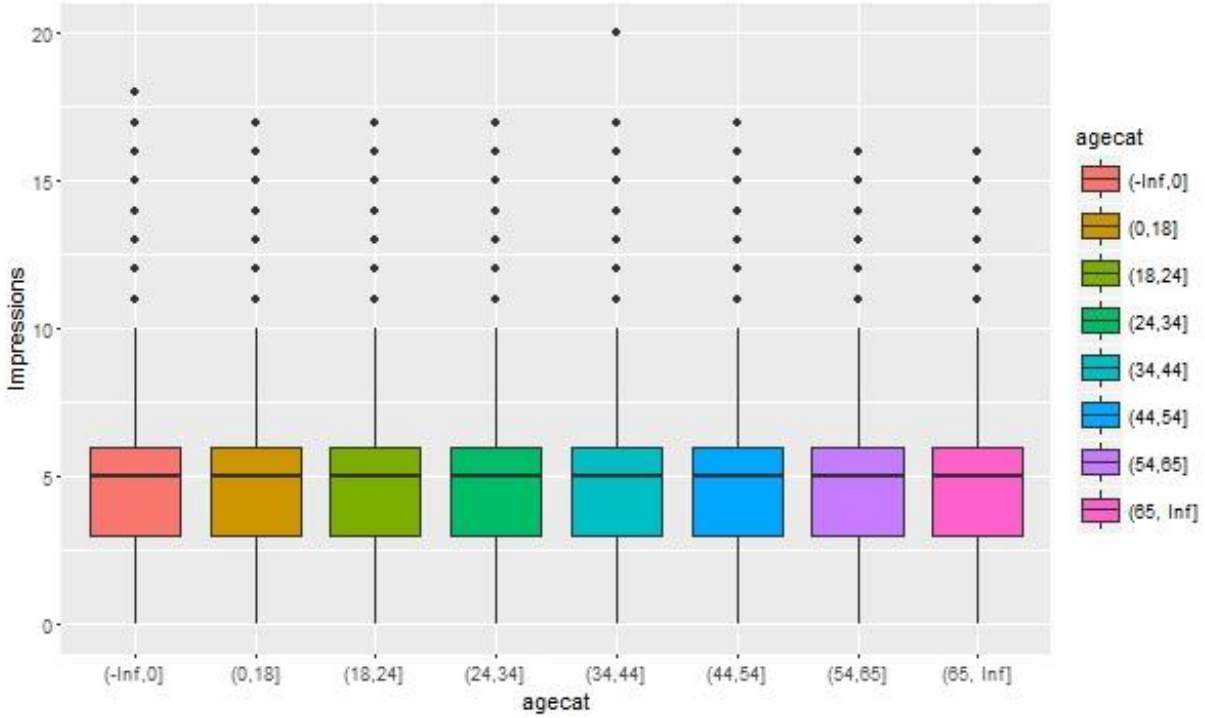
- **Read.csv()** is used to get the csv data directly in R dataframe
- Age categories are added to the dataframe by using the **cut()** method which gives the new column for the dataframe. The head of the dataframe looks something like below:

```
Console ~/
> head(data1)
  Age Gender Impressions Clicks Signed_In    agecat
1  36      0           3      0         1   (34,44]
2  73      1           3      0         1 (65, Inf]
3  30      0           3      0         1   (24,34]
4  49      1           3      0         1   (44,54]
5  47      1          11      0         1   (44,54]
6  47      0          11      1         1   (44,54]
>
```

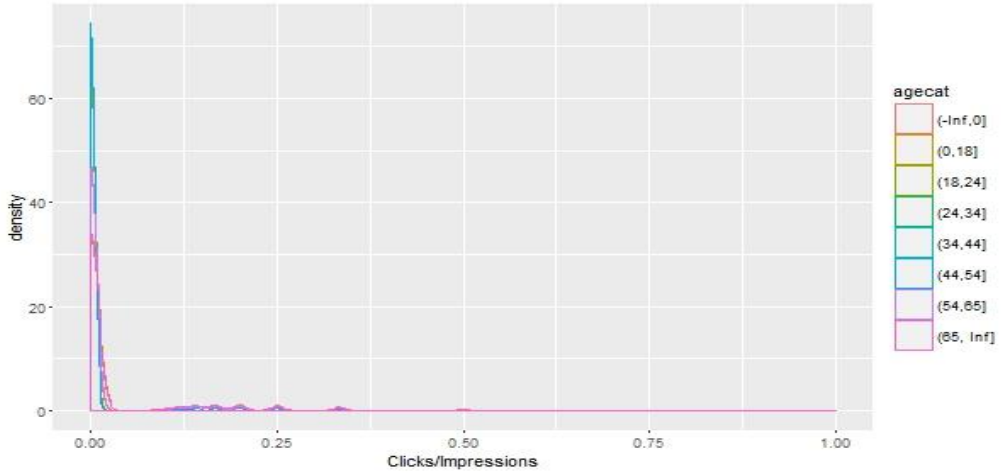
- **ggplot()** is used with **geom_histogram()** to plot the number of impressions of the users against the count with age categories filled in the histogram plot with different colors. The plot looks as below:



- These impressions per users can also be plot against the age categories which has the plot as below:

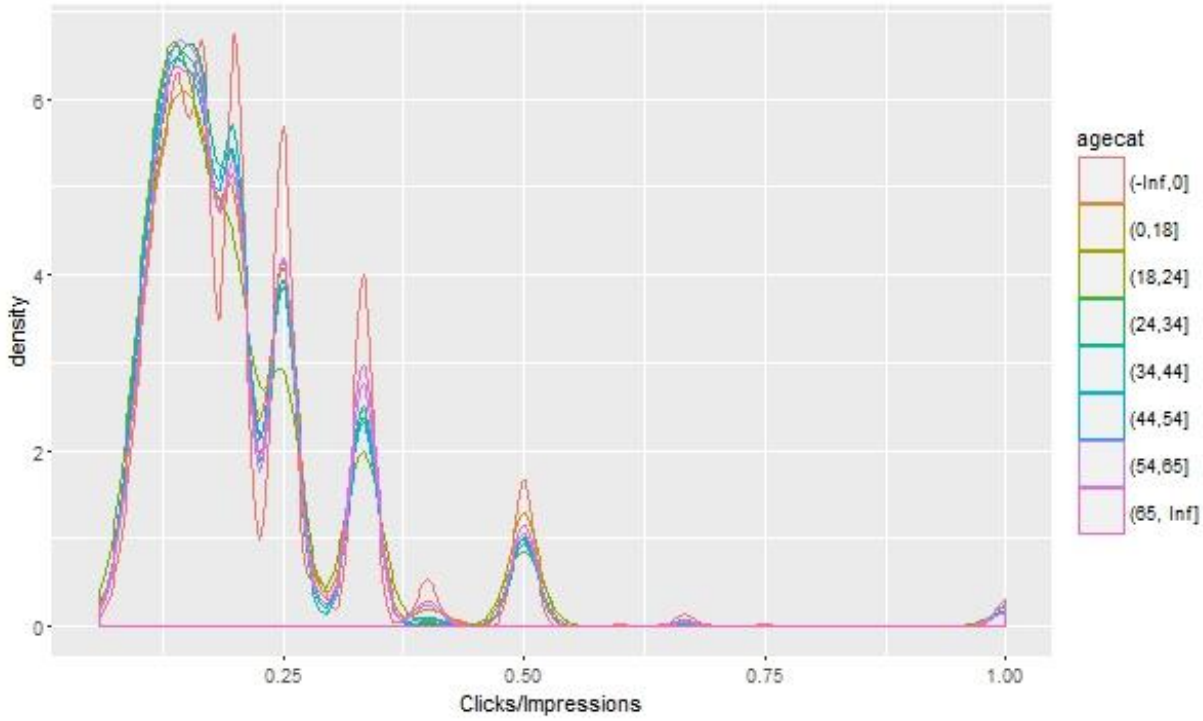


  - Inference: The boxes represent the range to the age categories. The dots represent the number of Impressions per age category. From the plot, we may conclude that most number of Impressions are by the people who has age in the range 34 to 44 while second highest is by the users who did not LogIn.
- As asked the book, next I plot the Click through rate i.e. Clicks per Impression rate using the Impressions and Clicks available in the dataframe.
- Click through rate for Clicks which had Impressions i.e. for which the Impressions were greater than zero are as shown in the below plot



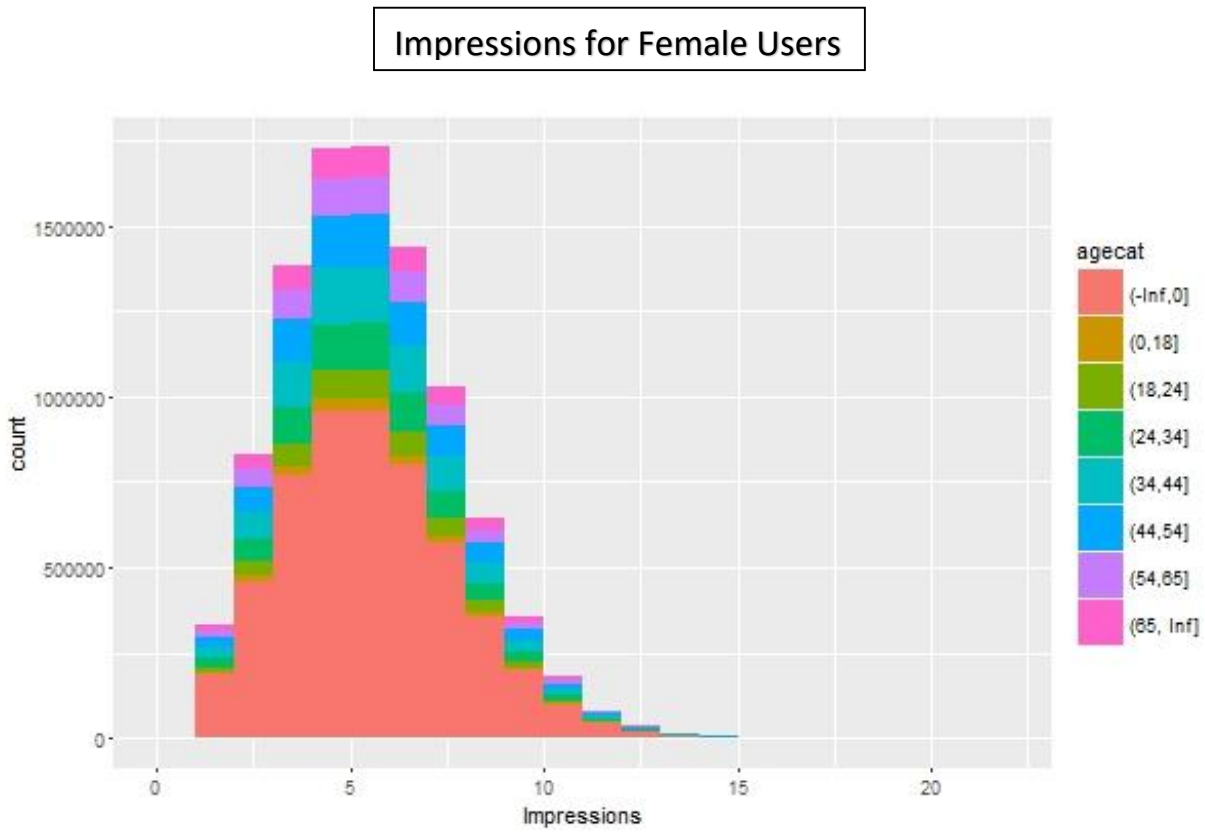
- Inference: The above plot shows the density of the click through rate for Impressions greater than zero and still the max steep is around zero. This means that the click through rate is almost zero because there are rows for which clicks and equal to zero

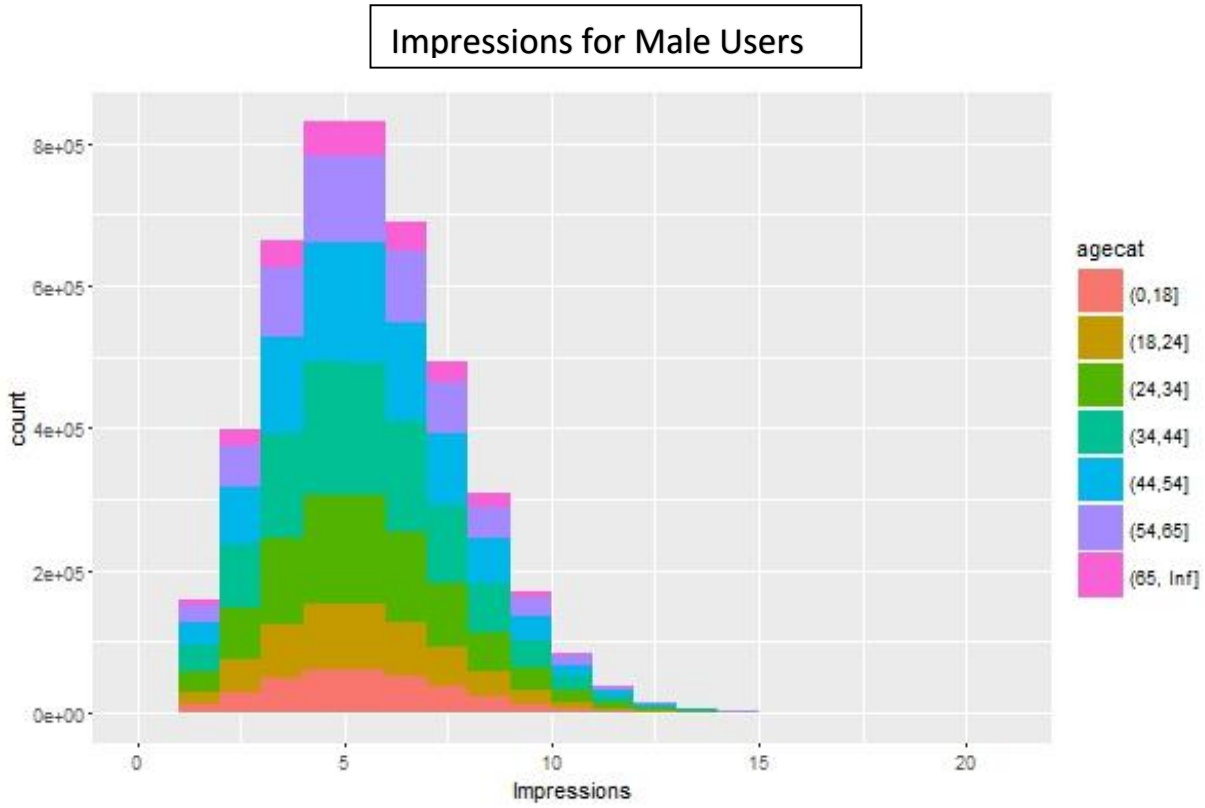
- Next I plot the click through rate for which the clicks are greater than zero. The plot looks as below:



Part 2: Extending the EDA to monthly data

- For taking the data for the whole month to one dataframe I looped over the data files to store and bind the data into a dataframe using the R **rbind()** method.
- Next I added the age categories to the dataframe as I did in the above case.
- Then I planned to plot the Number of Impressions according to the Gender (Female/Male).
- For plotting I cleaned the dataframe like only taking the data for female and male users with Impressions greater than zero. Also I plot the age categories to show the age categories of Female and male users in the respective plots.
- The Male and Female user's plots are as below:

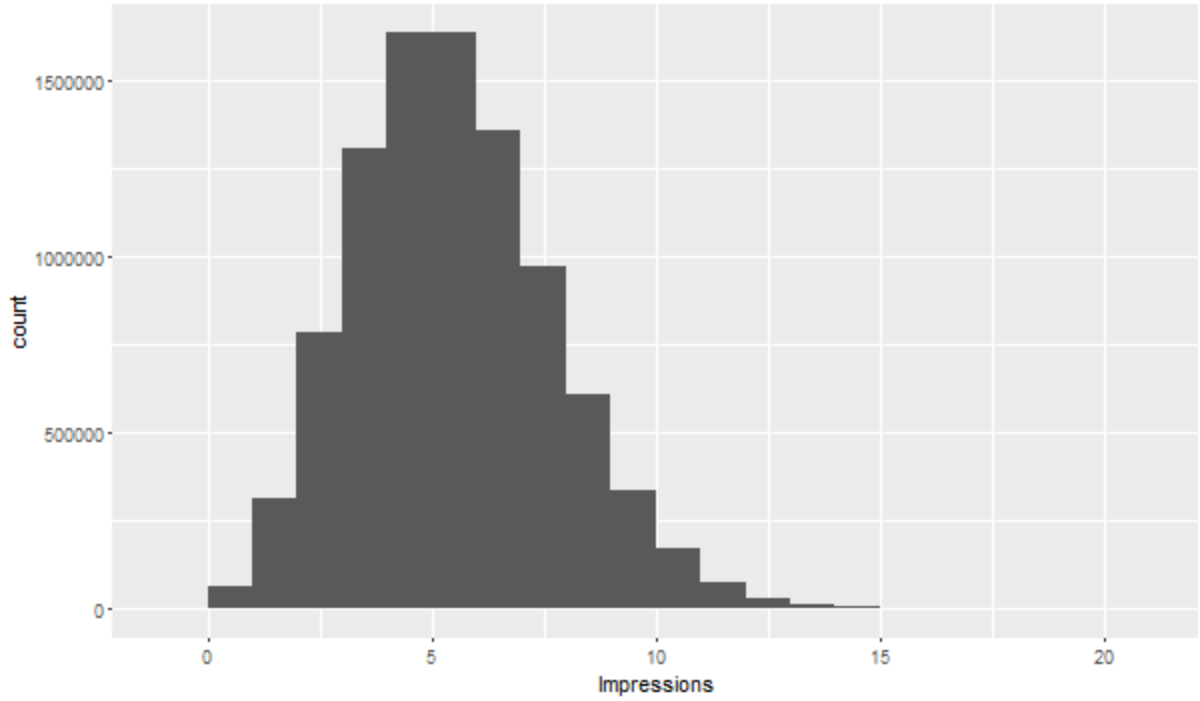




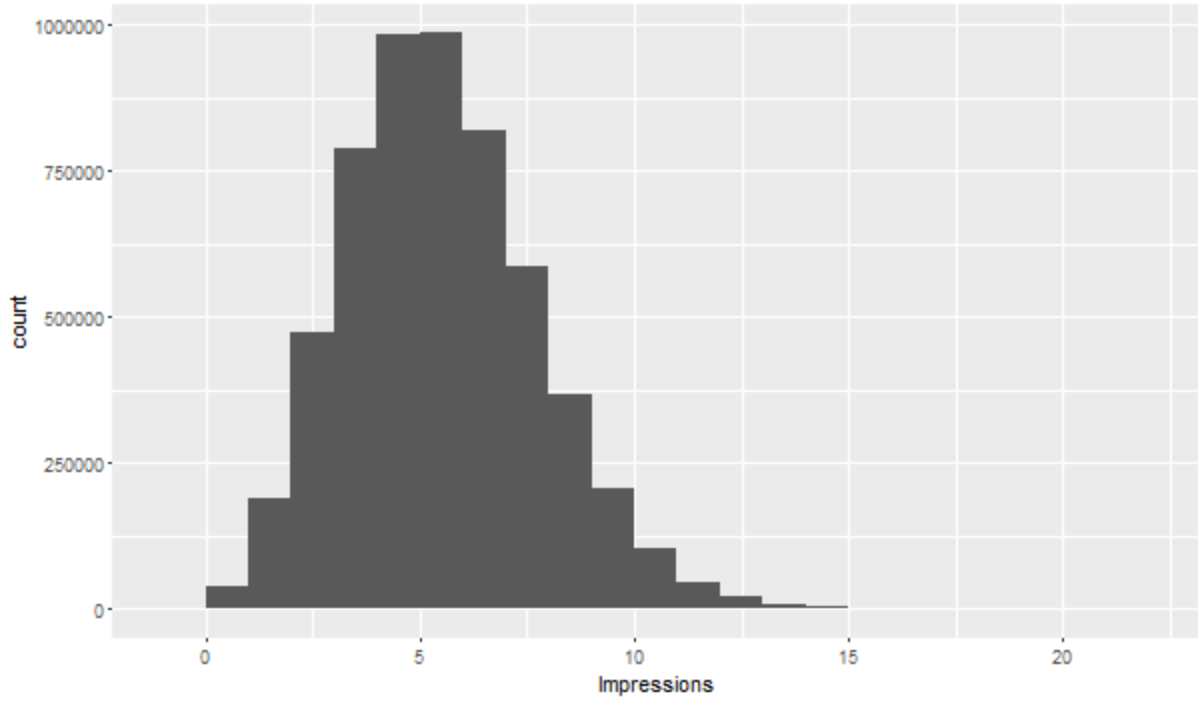
- Inference: The above plot can be used to conclude that are more female impressions that are under 18 or without signing-in. Although, the count scale indicates that there are lot more impressions by Male users in total as compared to female users.

- Next, I plot the Impressions for users on the basis of whether they are signed-in or not. The plots are as below:

Impressions for Logged-In Users



Impressions for Not Logged-In Users



- Inference: The count scale in the plot indicates that there are lot more impressions of the Logged-In users as compared to others. This means that more users prefer to sign-in and use the site as compared to one time non signing-in users.

References:

- TextBook
- Google Search for the methods in R