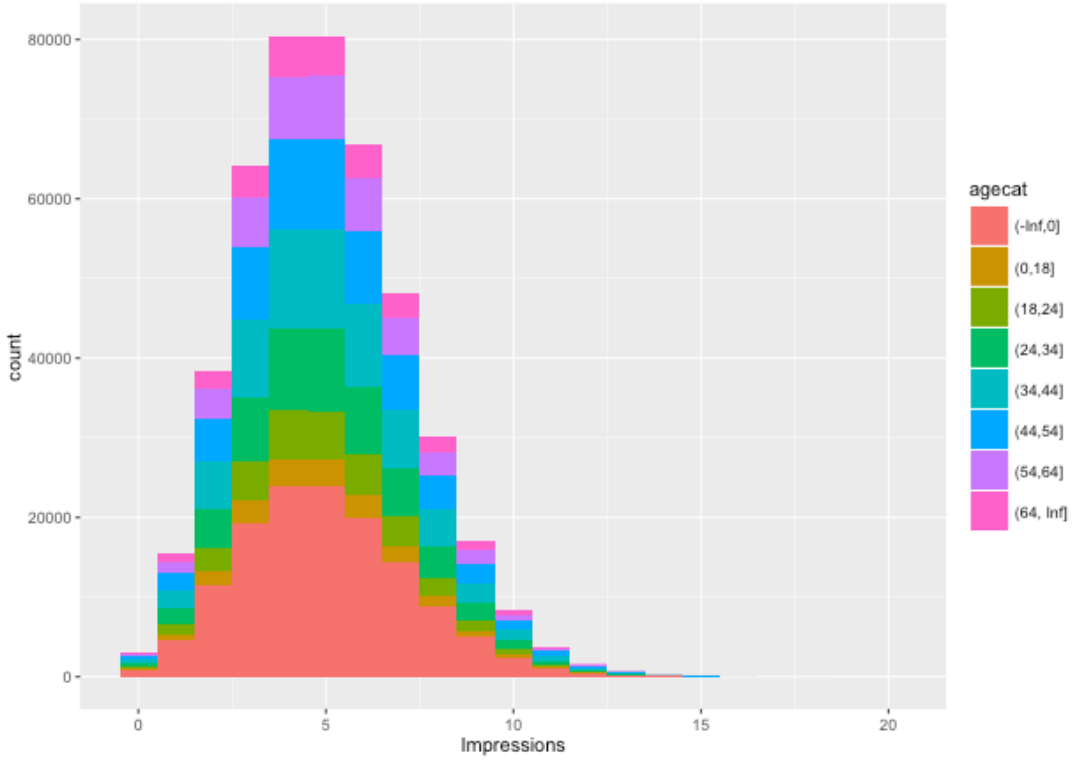
CSE587: DATA INTENSIVE COMPUTING

PROBLEM 2:SIMPLE EDA

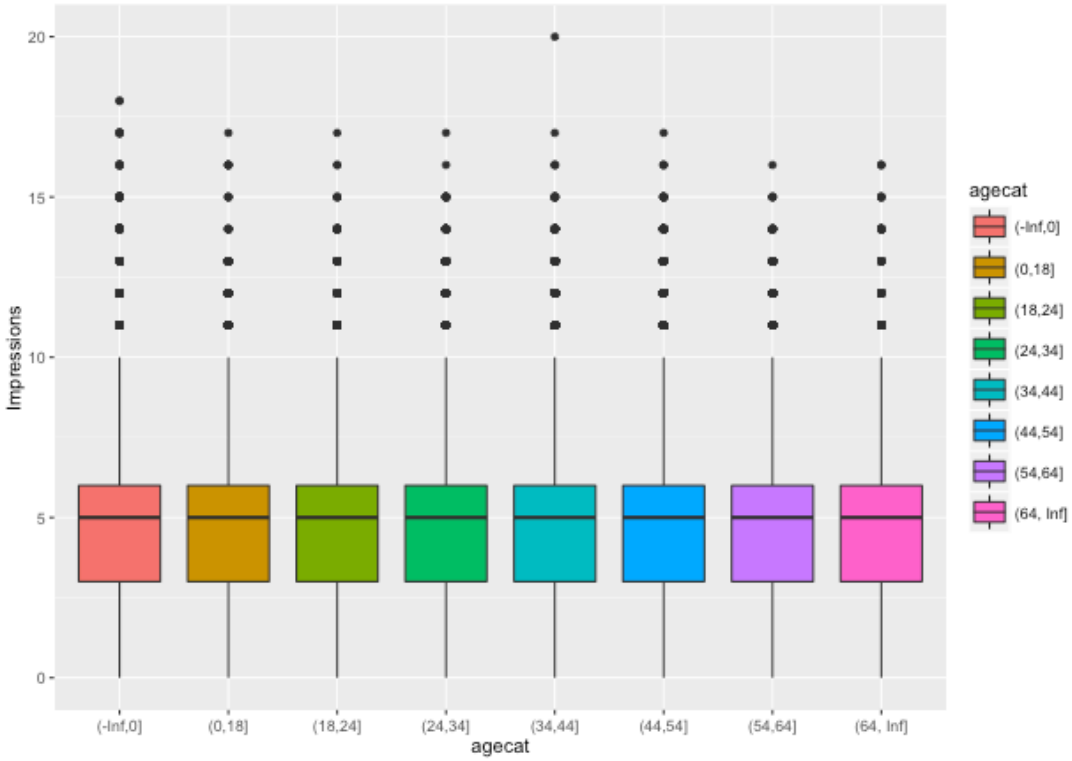
For Problem 2, the dataset used is the “New York Times” dataset that comes as comma separated values or csv files.

The data contains the following information about the New York Times readers: age, gender, number of impressions, number of clicks, and logged in or not.

The first plot is a density plot that gives the density of impressions for every users’ age category as shown below

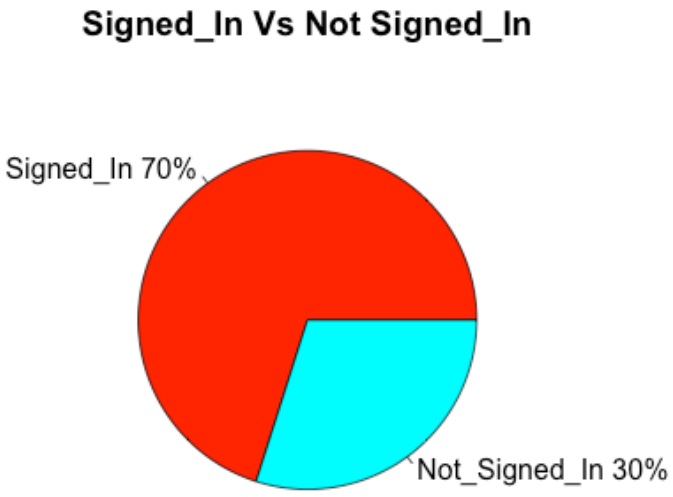


The plot shows that the count of impressions increases with the age category. A similar box plot can be plotted to illustrate the same

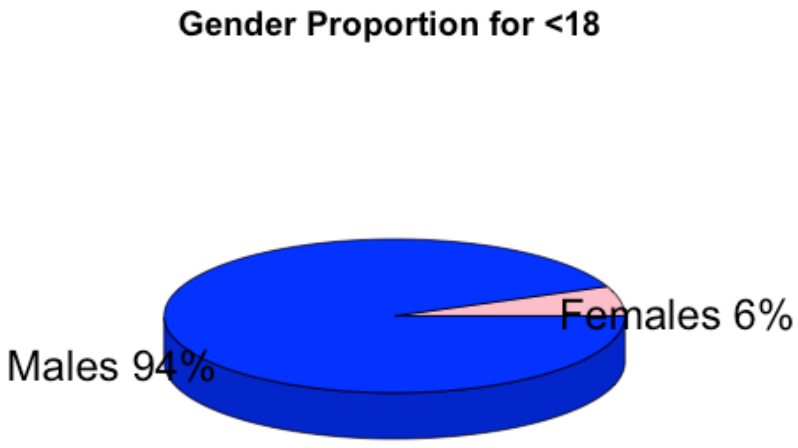


Then we create a new category based on Impressions. We know that the number of clicks is dependent upon the number of impressions. So we use the Impressions to create a new category “scode”.

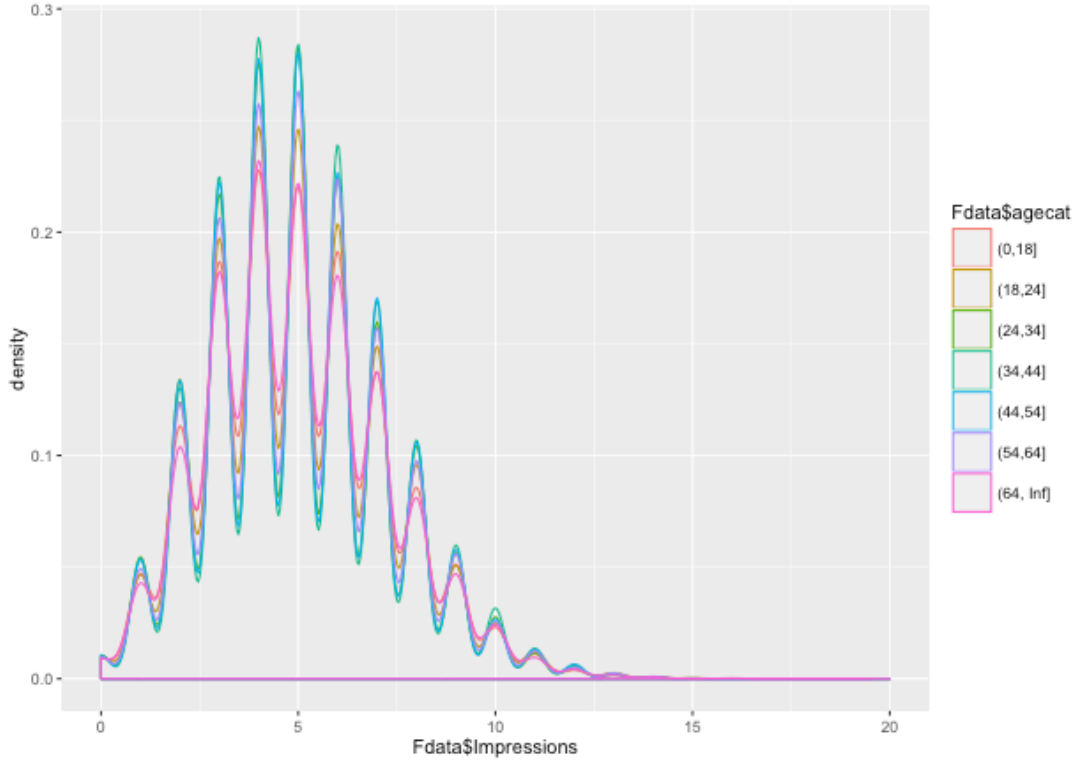
Then we make a quantitative comparison between the number of signed in users versus the number of non-signed in users by taking the count of the logged in and logged out users respectively. The pie chart is as seen below



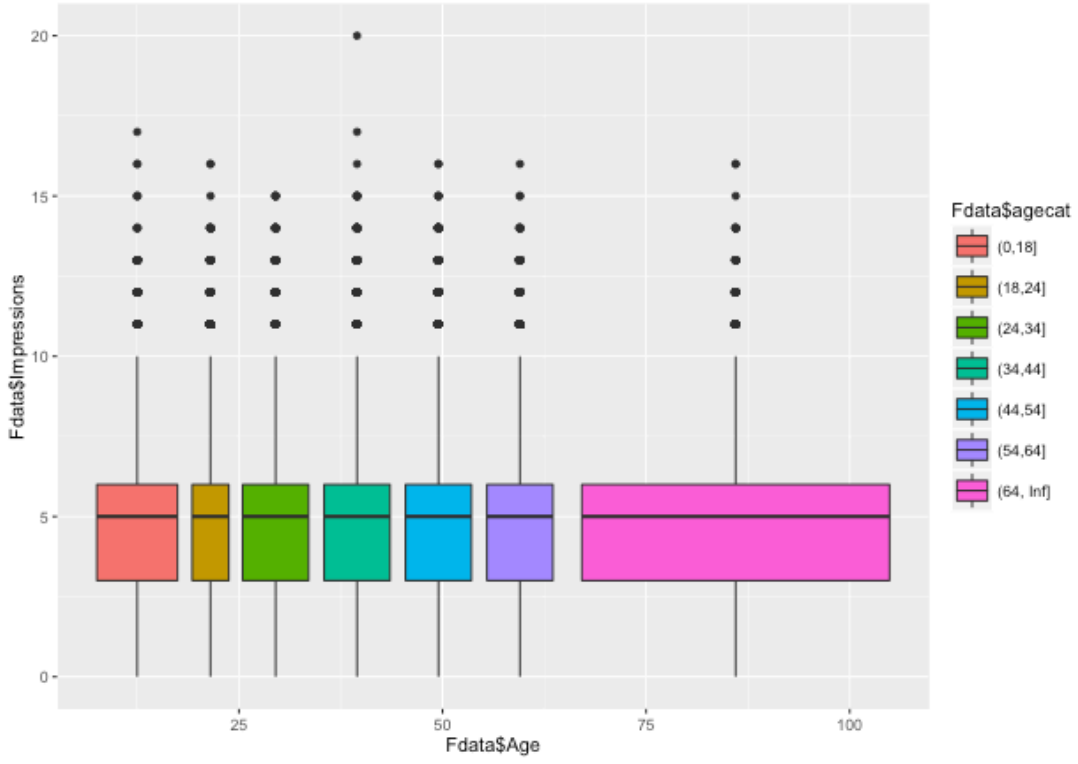
The next pie chart shows the comparison between the number of under 18 females versus the number of under 18 males. The graph shows that the number of males is larger than females.



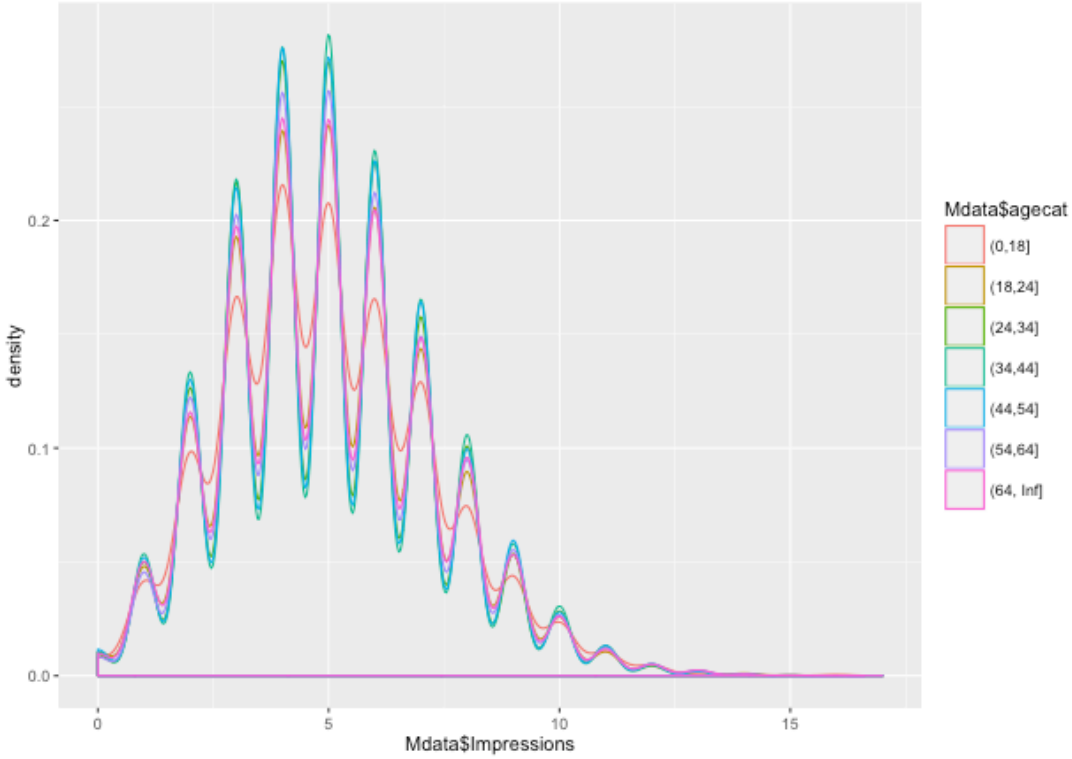
Next we perform separate analysis of the impressions and other factors on both females and males separately.



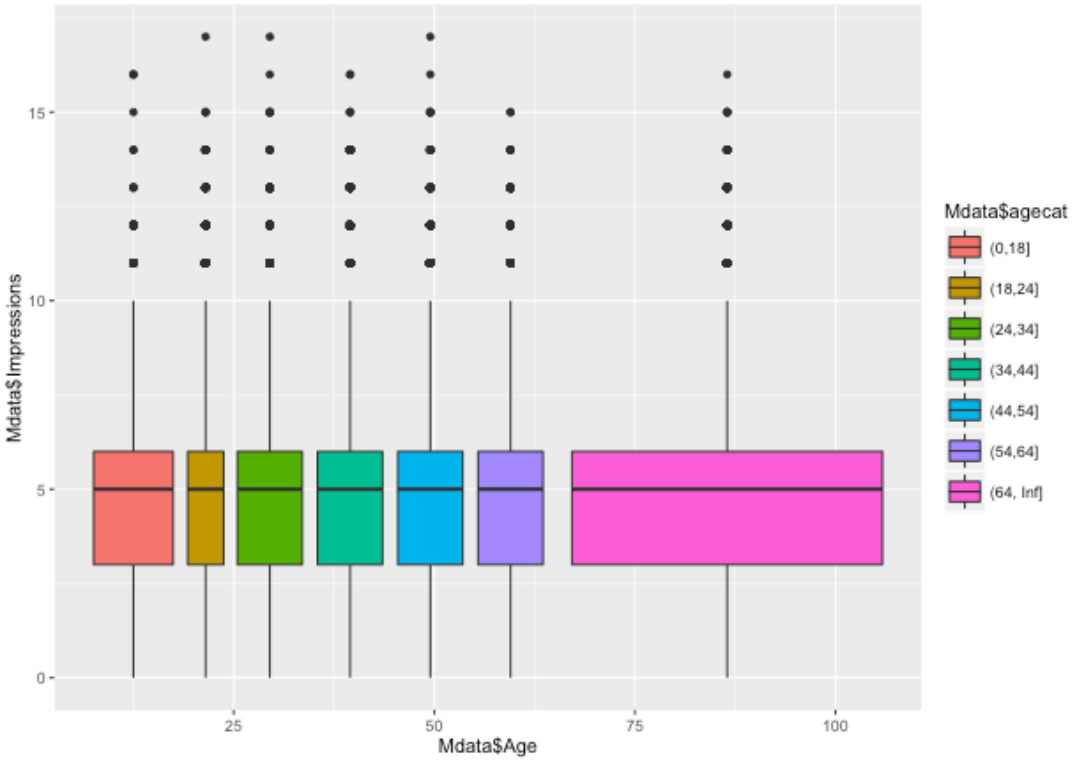
The graph shows that the older females (above 64) have lesser number of impressions on the site than the females around the age of 24 to 34, which is the highest density of all. Similarly the fact can also be illustrated by a box plot.



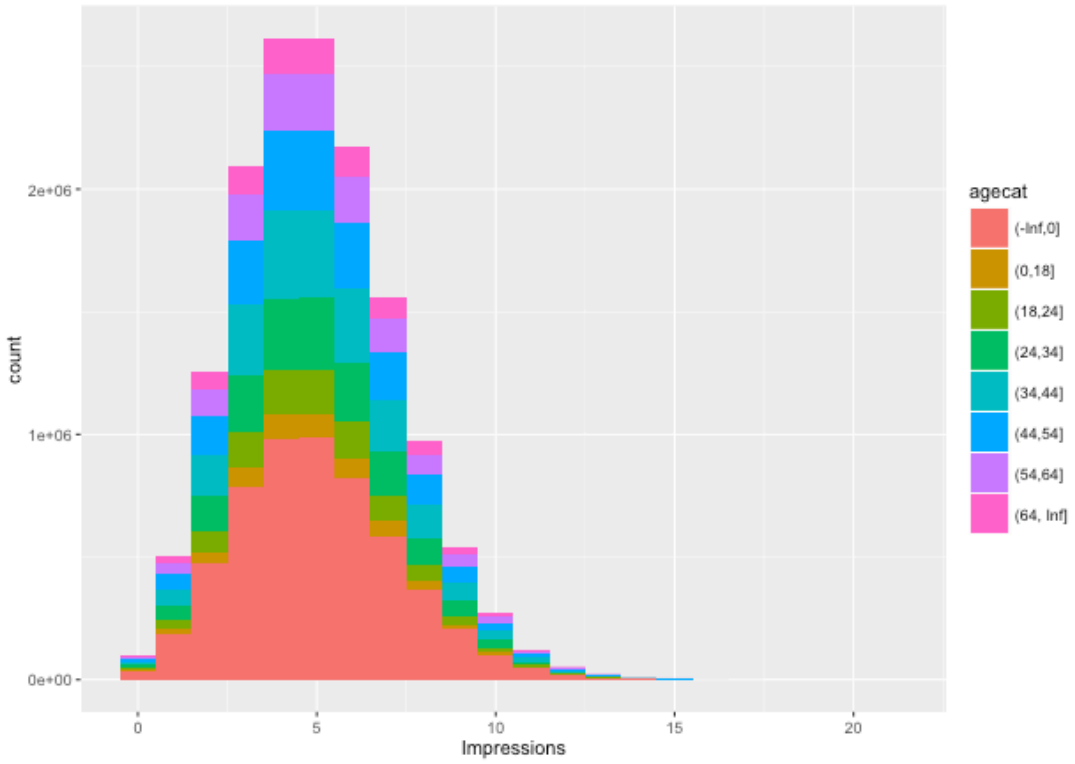
The same analysis is done on the male population as shown below.



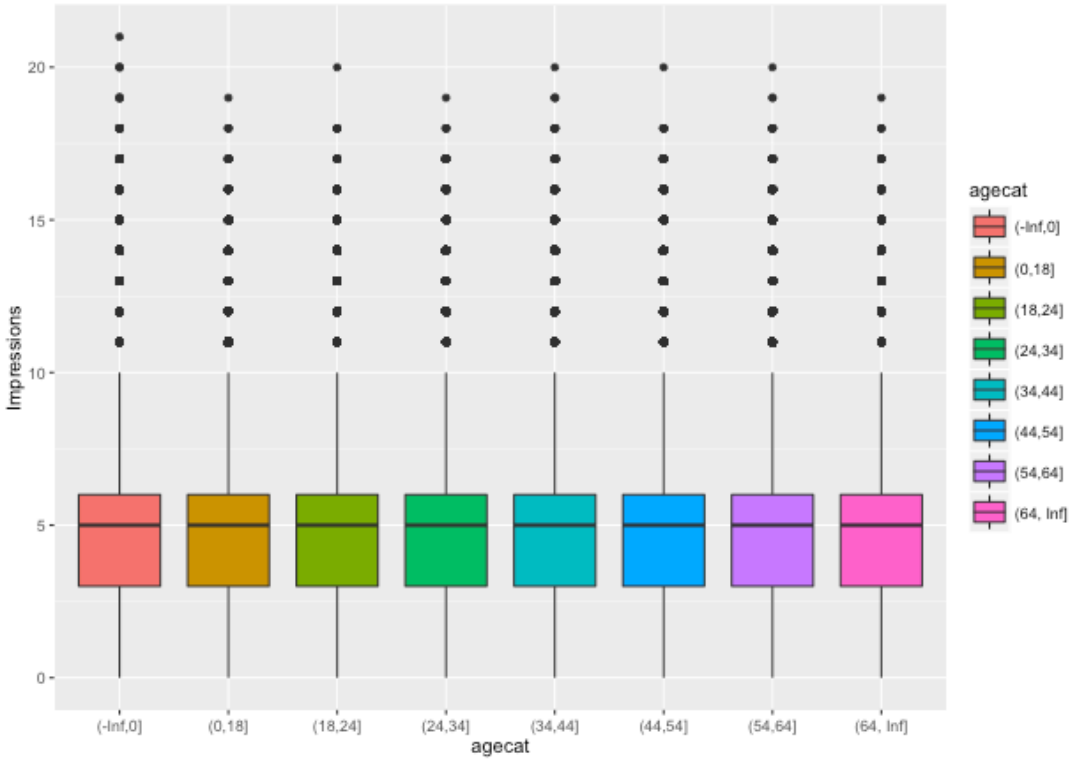
The graph shows that the same pattern is seen in the male population too as the female population with the older males with lesser impressions than the younger males. Although here we can also see the 0 to 18 year old males count a little more prominently than the females . The following is a box plot illustrating the same



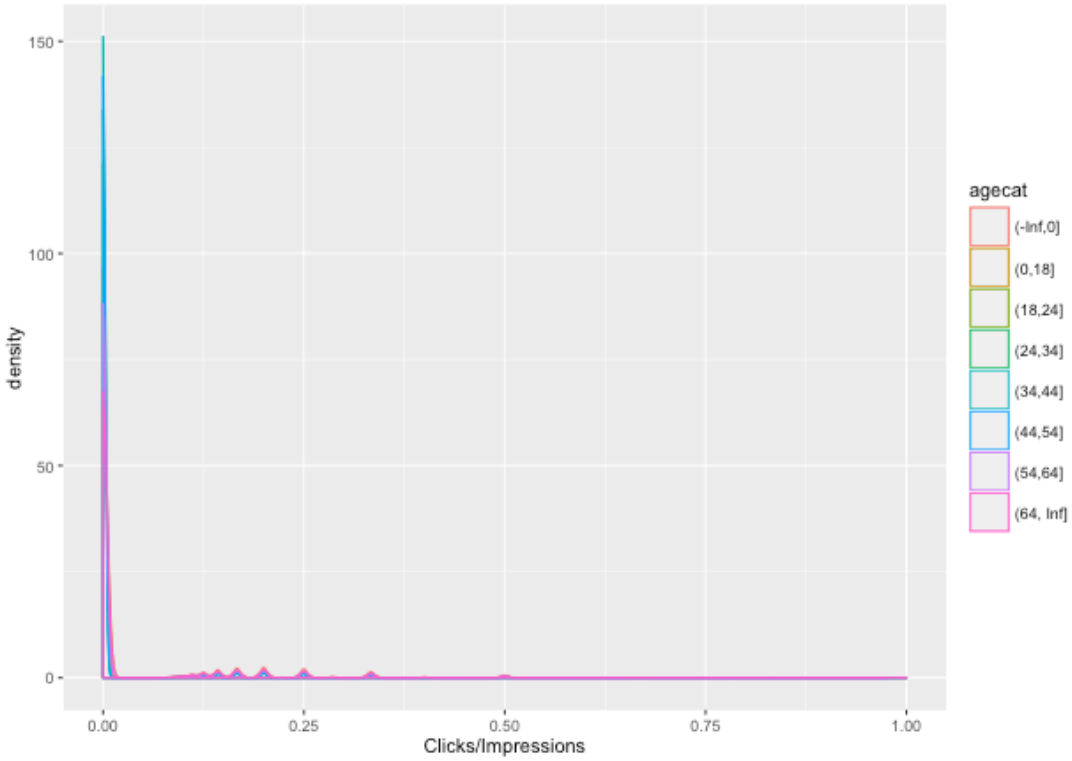
The plots discussed above was for a single day in the New York Times site. When extended over 31 days, the population increases and hence we can see different patterns emerging over a period of time as seen below



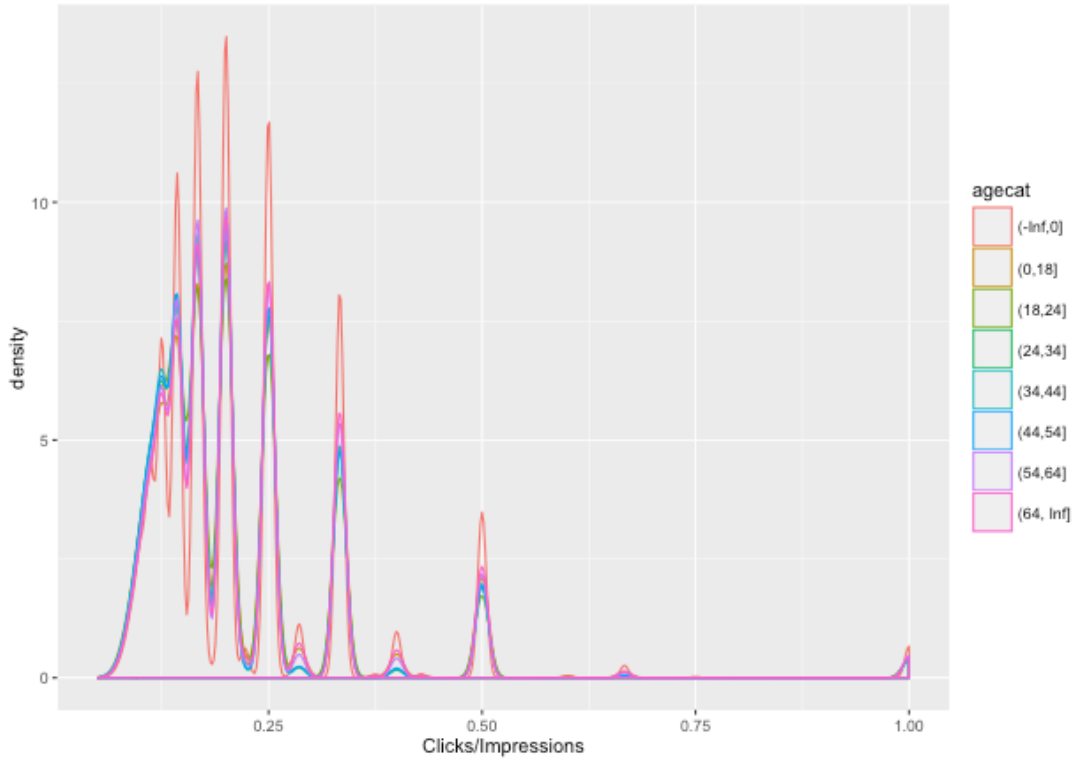
We can see that the density peak is higher than what was observed for a single day data. Similar variation can be seen in the box plot



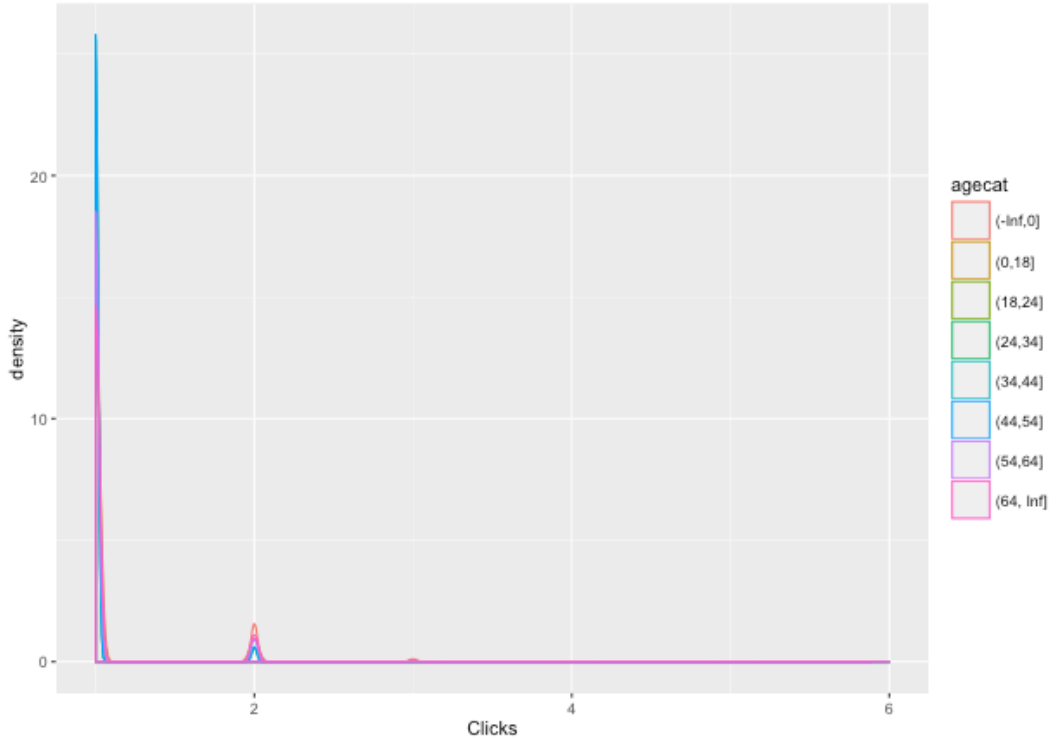
The Click Through Rate is an important feature for analyzing the behavior of users in a website. The following plots show that behavior



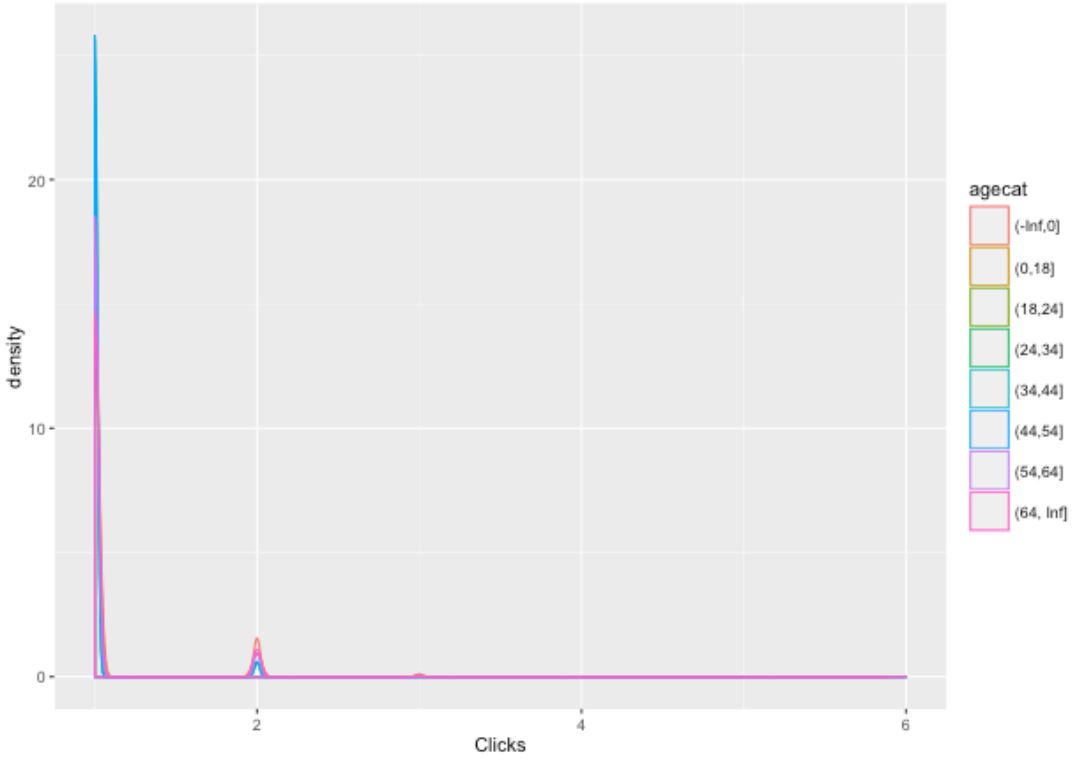
We see that the density of the click through rate is higher for the 44 to 54 age group but it almost becomes flat when the age increases



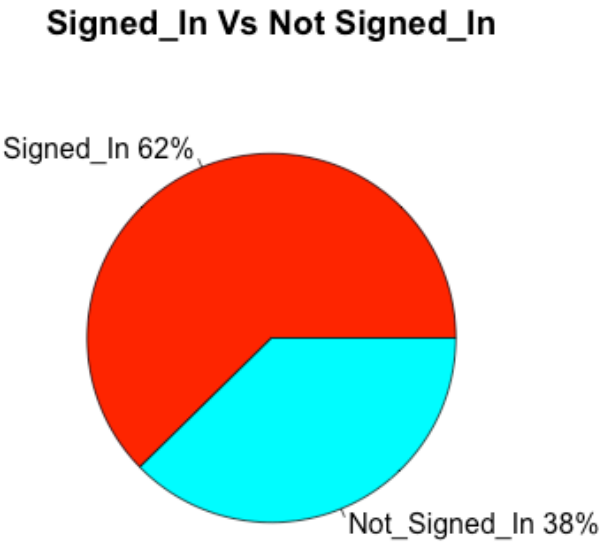
This is the graph for clicks per impression that shows that the younger population tend to have a higher clicks rate than the older population.



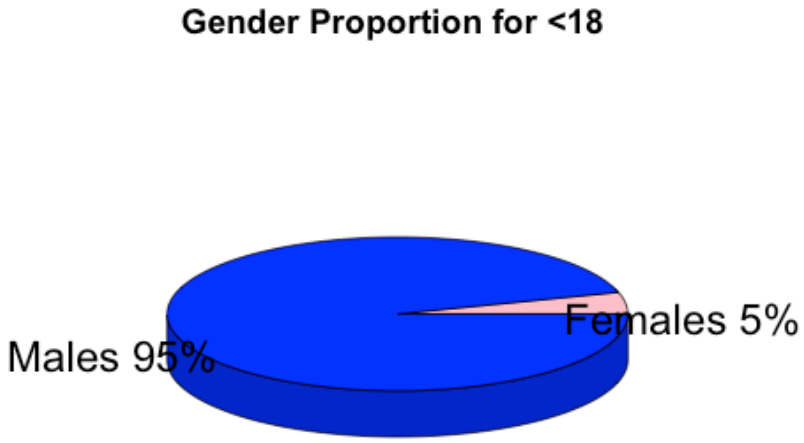
Similarly we draw graphs for clicks density.



Next we analyze the percentage of signed in users to the logged out users as before.

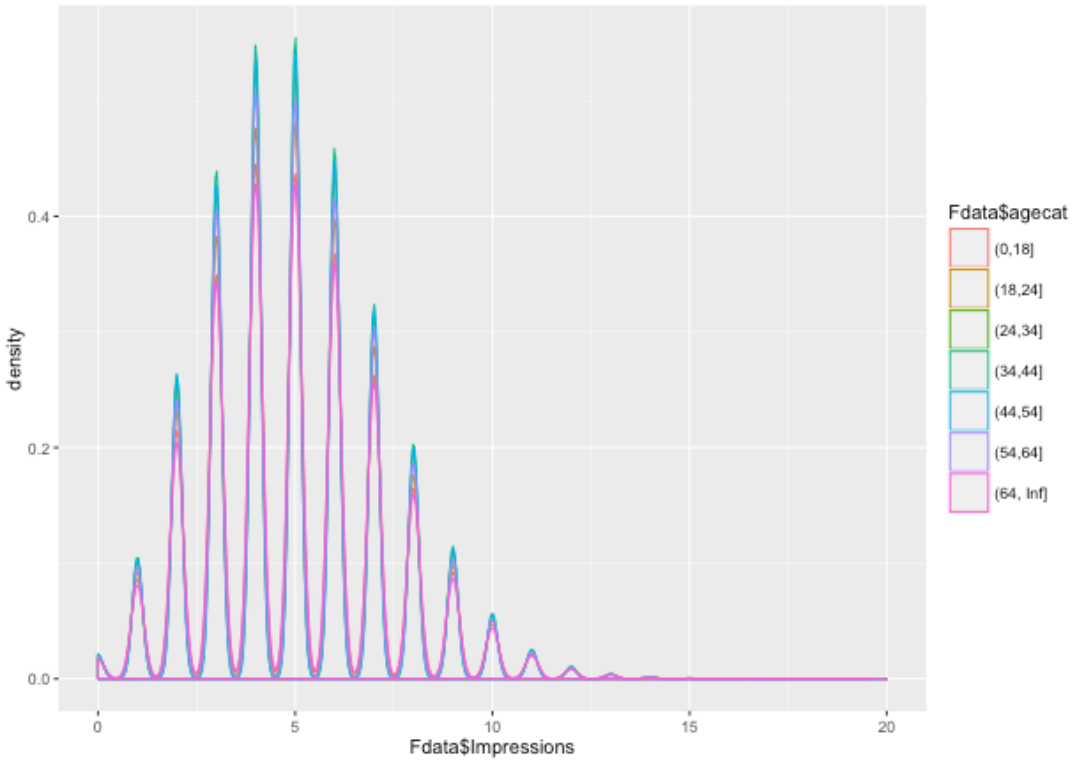


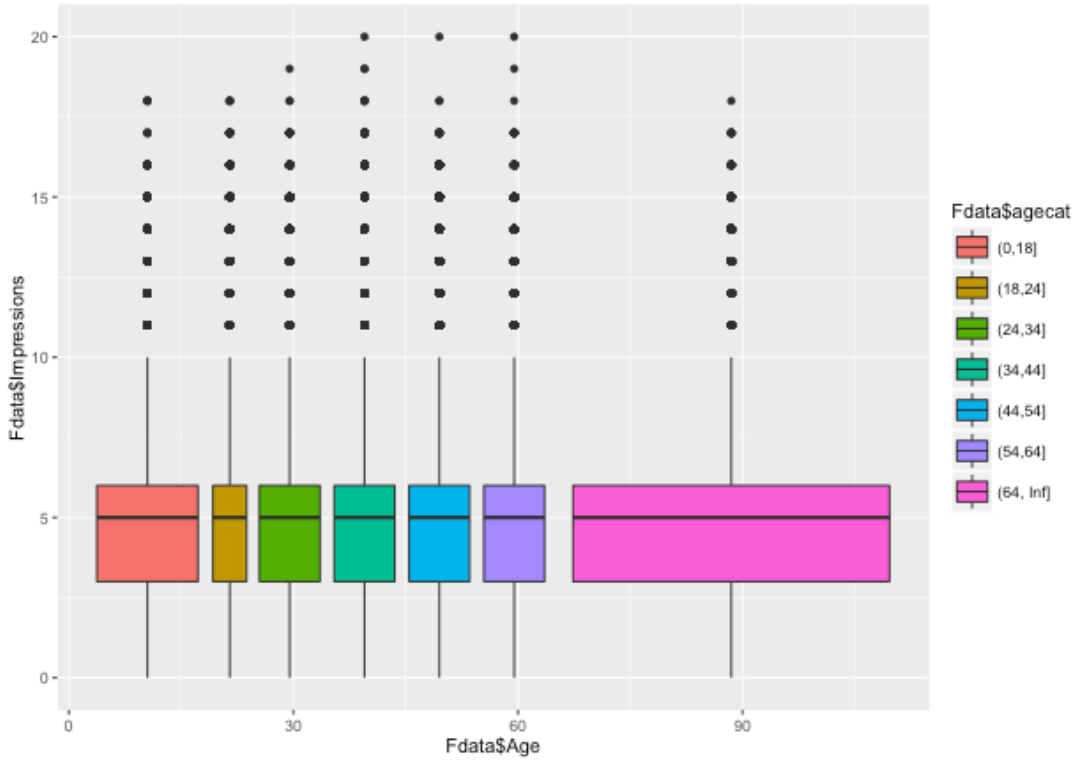
Now we can see that the percentages have slightly changed (which is due to change in amount of population) than for a single day.



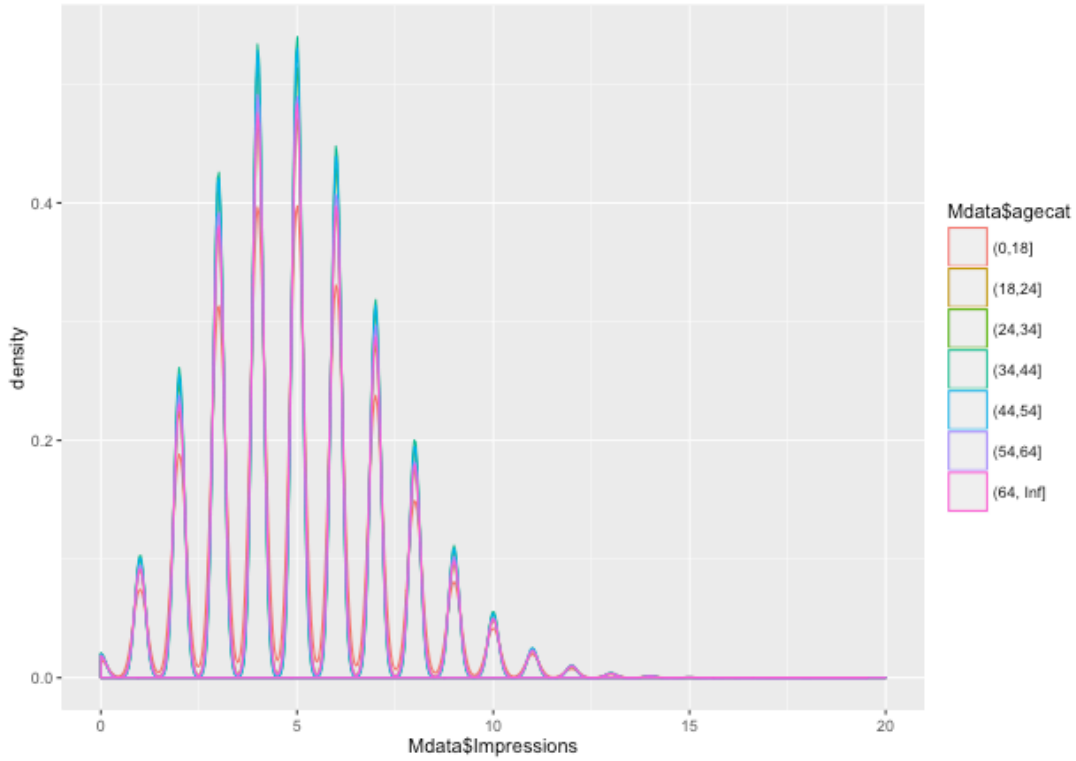
The proportion of under 18 female users vs under 18 male users almost remains the same for the extended dataset

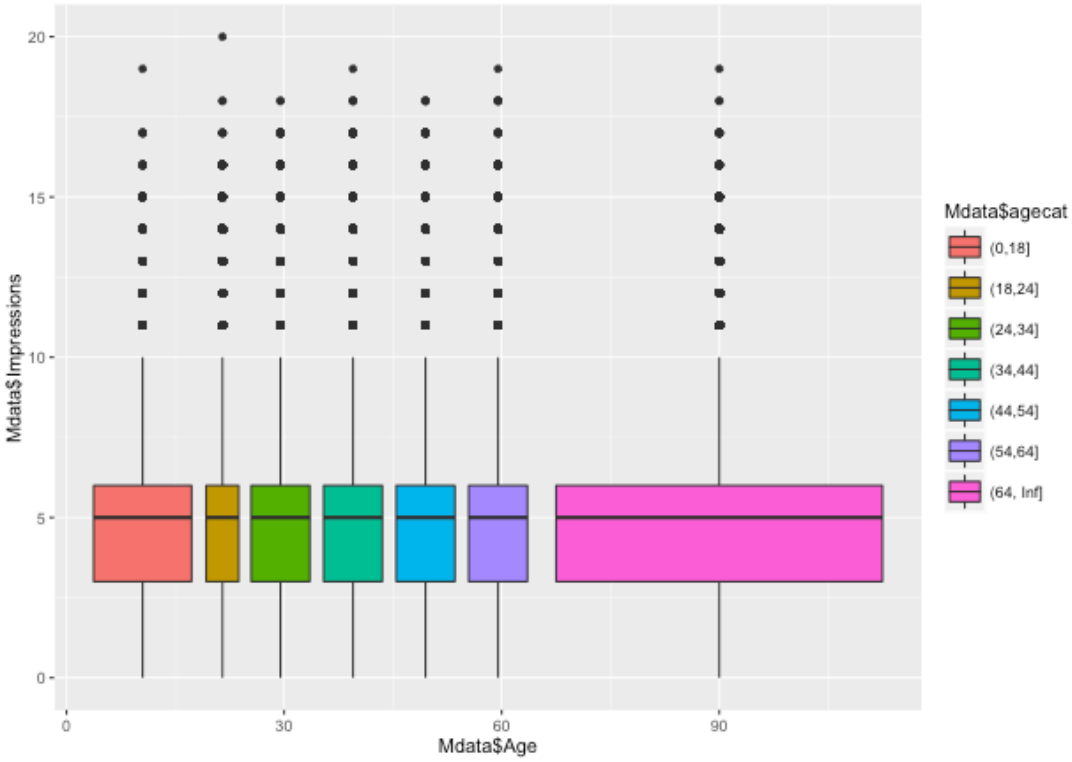
This is the analysis for the impressions left by the female population for the extended dataset. We can see a significant rise in the density contributed by the older females





Next graph shows the density of impressions for the male population . Here too there is a rise in the density of impressions for the older users.




Mdata$ Impressions
20
15
10
5
0
0
30
60
90
Mdata$Age
Mdata$agecat
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64, Inf]