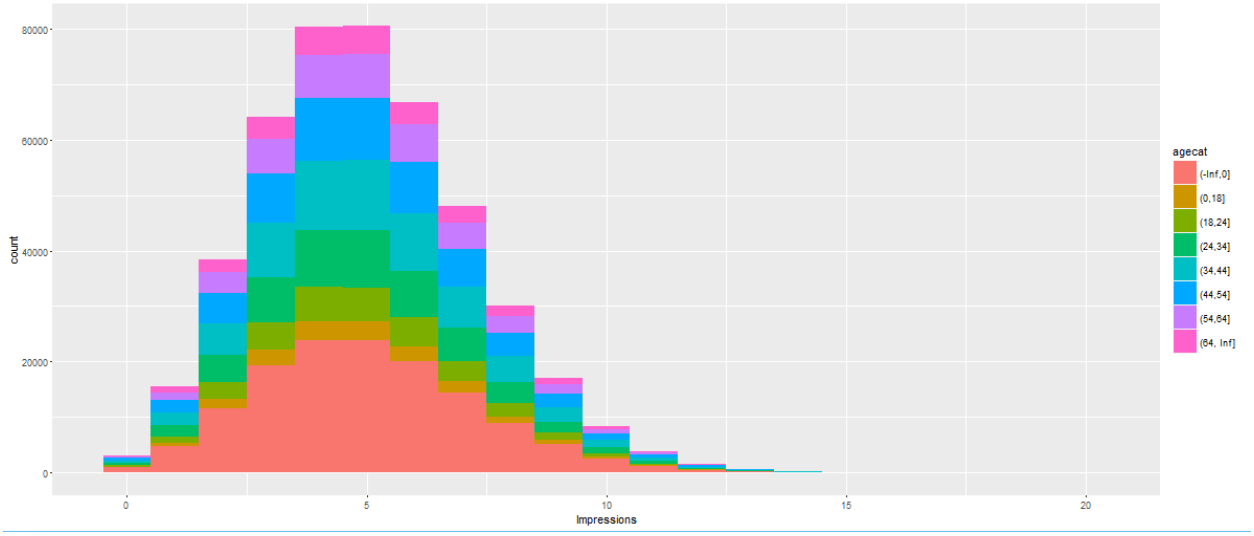
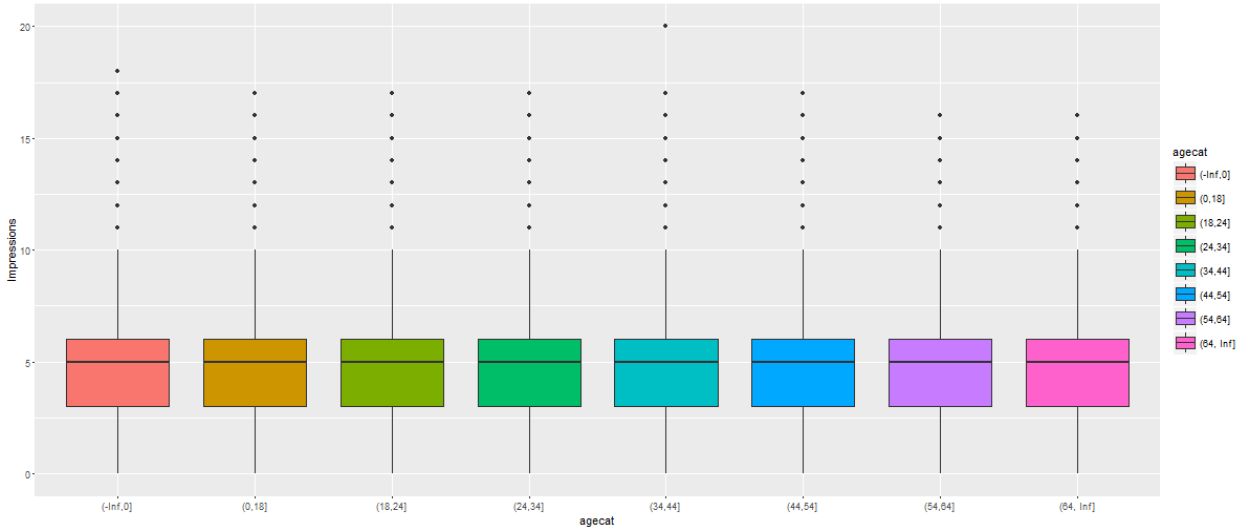
PROBLEM 2

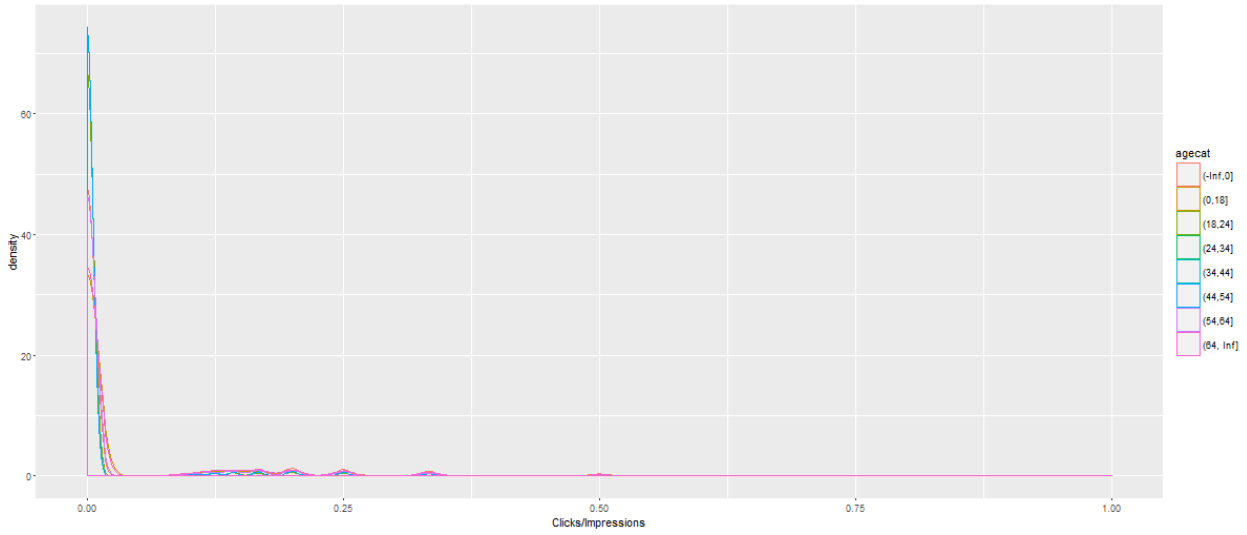
(a) The charts for EDA of data collected by New York Times for a single day of a month are as below:

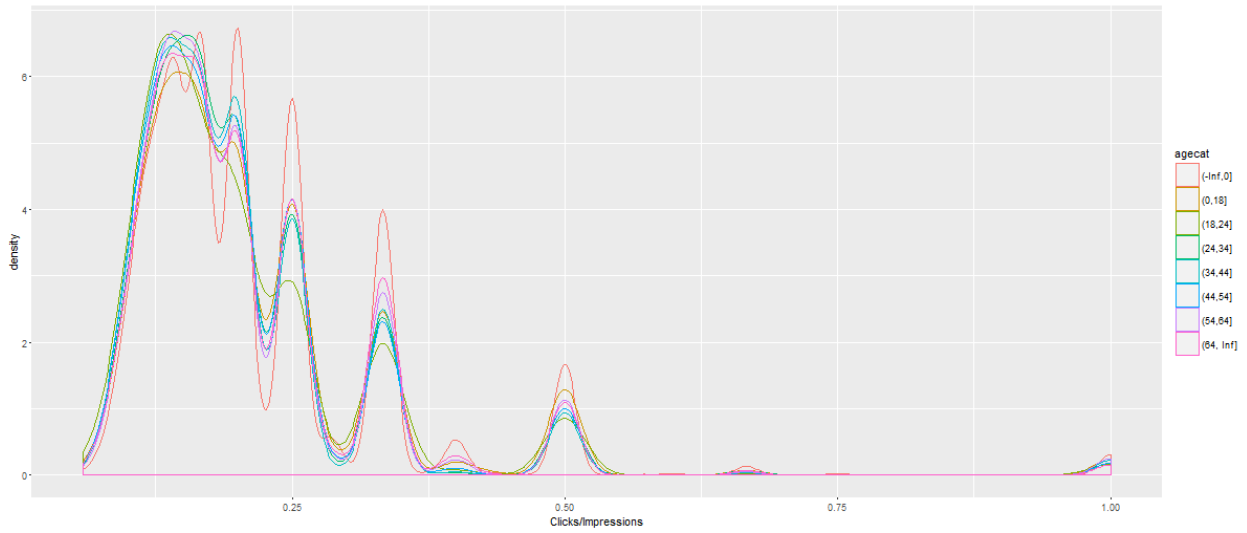


The data here is being categorized based on age of the users. This Histogram the count of Impressions based on these age categories.

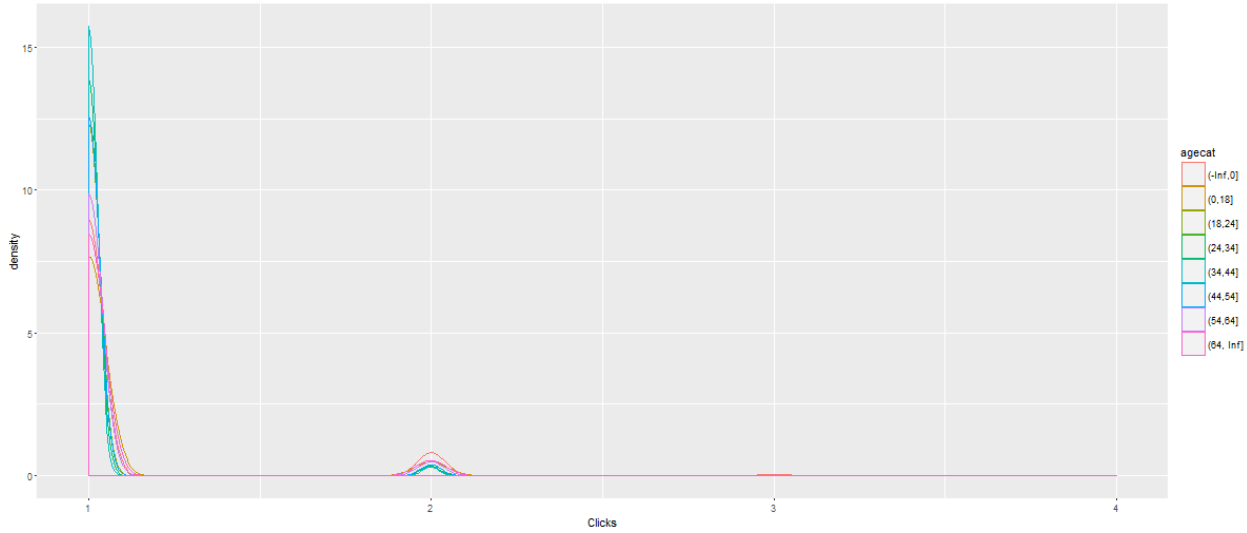


The above is a box plot on the same data.





Density plot on Age Categories vs Click Through Rate (Clicks/Impressions) with Clicks >0



(b) The charts for EDA of data collected by New York Times for the whole month are as below:

The summary of the whole data:

```
      Age              Gender         Impressions      Clicks
 Min.   :  0.00   Min.   :0.0000   Min.   : 0   Min.   :0.00000
 1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.: 3   1st Qu.:0.00000
 Median : 26.00   Median :0.0000   Median : 5   Median :0.00000
 Mean   : 26.24   Mean   :0.3231   Mean   : 5   Mean   :0.09773
 3rd Qu.: 46.00   3rd Qu.:1.0000   3rd Qu.: 6   3rd Qu.:0.00000
 Max.   :115.00   Max.   :1.0000   Max.   :21   Max.   :6.00000

   Signed_In          agecat            gencat            usrcat
 Min.   :0.0000   (-Inf,0]:5613610   (-Inf,0]:10090192   (-Inf,0]:5613610
 1st Qu.:0.0000   (34,44] :2044613   (0, Inf]: 4815673   (0, Inf]:9292255
 Median :1.0000   (44,54] :1859487
 Mean   :0.6234   (24,34] :1673650
 3rd Qu.:1.0000   (54,64] :1299303
 Max.   :1.0000   (18,24] :1022112
                  (Other) :1393090
```

As similar to Problem (a), we categorize the data on Age. The summary is as follows:

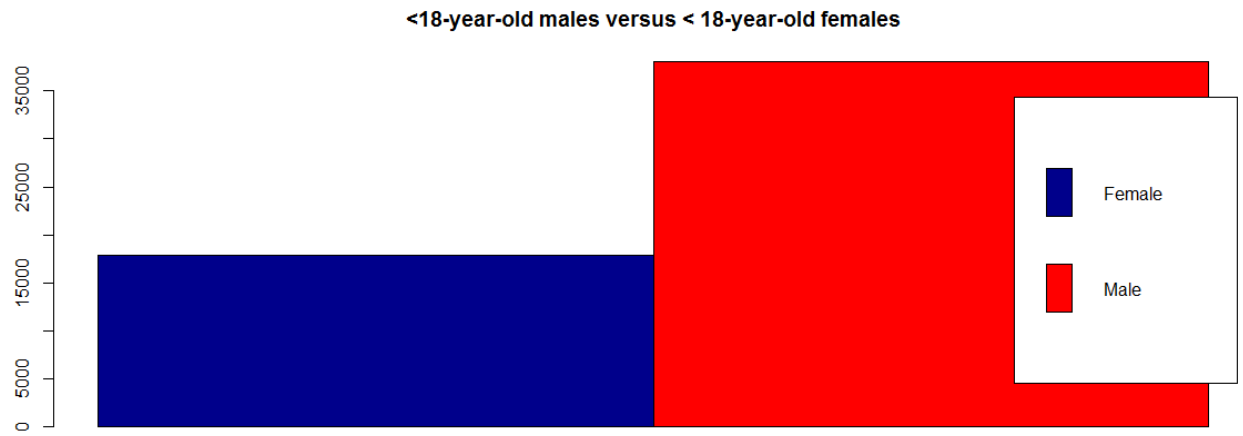
```
> siterange <- function(x){c(length(x), min(x), mean(x), max(x))}
> summaryBy(Age~agecat, data =data2, FUN=siterange)
     agecat Age.FUN1 Age.FUN2 Age.FUN3 Age.FUN4
1  (-Inf,0]  5613610        0  0.00000        0
2    (0,18]   556988        3 16.01827       18
3   (18,24]  1022112       19 21.27320       24
4   (24,34]  1673650       25 29.49960       34
5   (34,44]  2044613       35 39.49901       44
6   (44,54]  1859487       45 49.49890       54
7   (54,64]  1299303       55 59.49965       64
8 (64, Inf]   836102       65 72.98031      115
```

Also, we find the stats of other columns based on agecat.

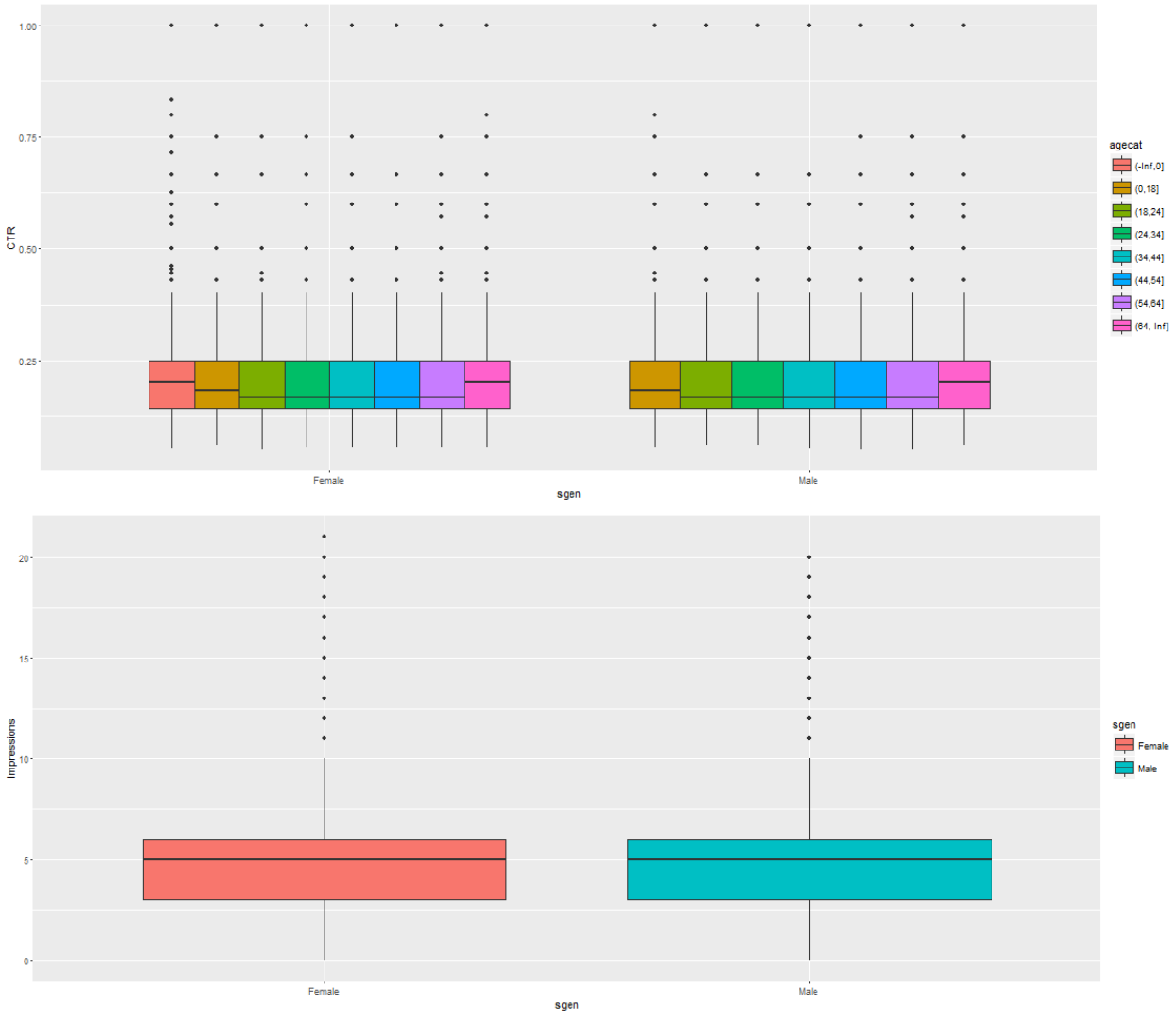
```
> summaryBy(Gender+Signed_In+Impressions+Clicks~agecat,data =data2)
     agecat Gender.mean Signed_In.mean Impressions.mean Clicks.mean
1  (-Inf,0]   0.0000000              0         5.001178  0.14187804
2    (0,18]   0.6386870              1         5.006339  0.13316804
3   (18,24]   0.5263053              1         4.997329  0.04998767
4   (24,34]   0.5265581              1         4.998146  0.05006483
5   (34,44]   0.5272152              1         4.997909  0.05007256
6   (44,54]   0.5259445              1         4.996438  0.05022891
7   (54,64]   0.5264284              1         4.998224  0.10002902
8 (64, Inf]   0.3597492              1         5.001963  0.15012044
```

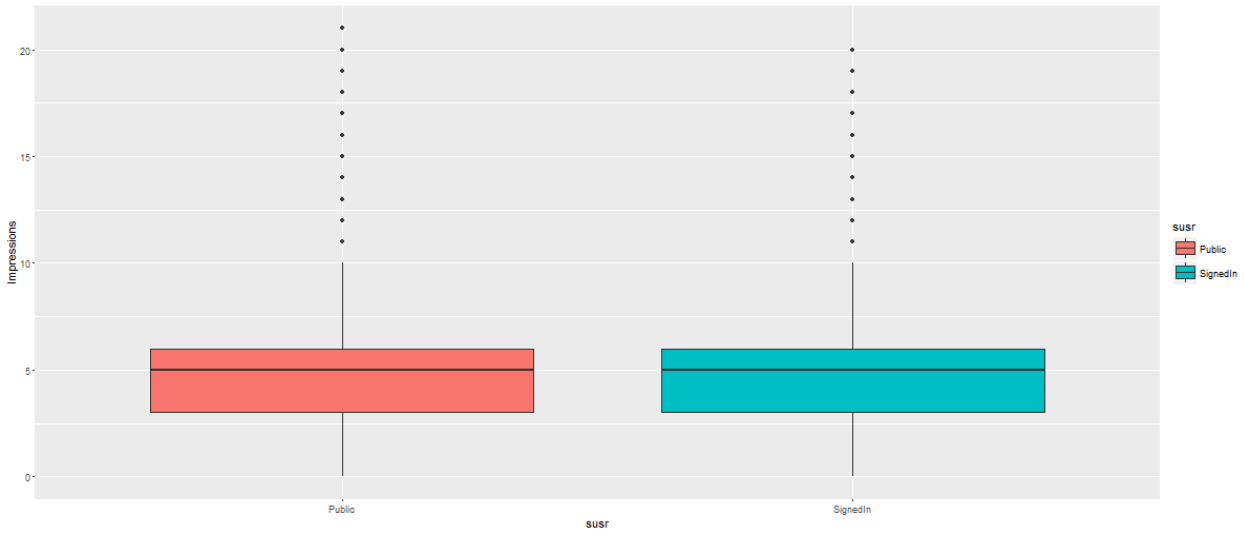
We notice that the Clicks are quite high for teenagers, i.e. age range (0,18] and for the older people (64,Inf], which gives us an idea that people with more free time tend to browse more.

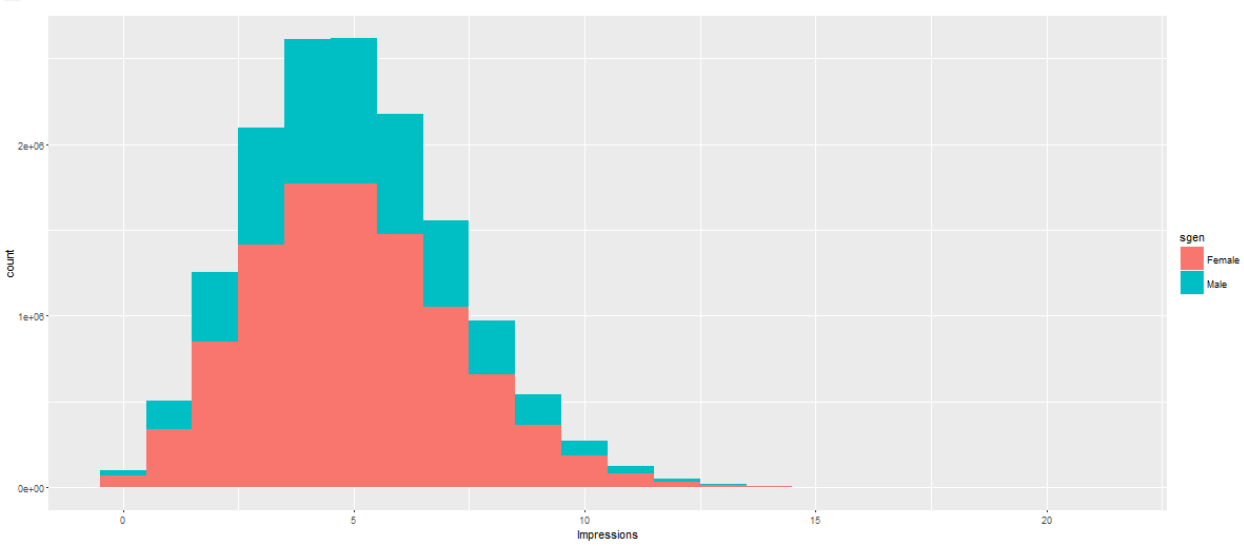
Further we classify the data based on Gender and people who are logged in or logged out (public). The graphs are as follows:

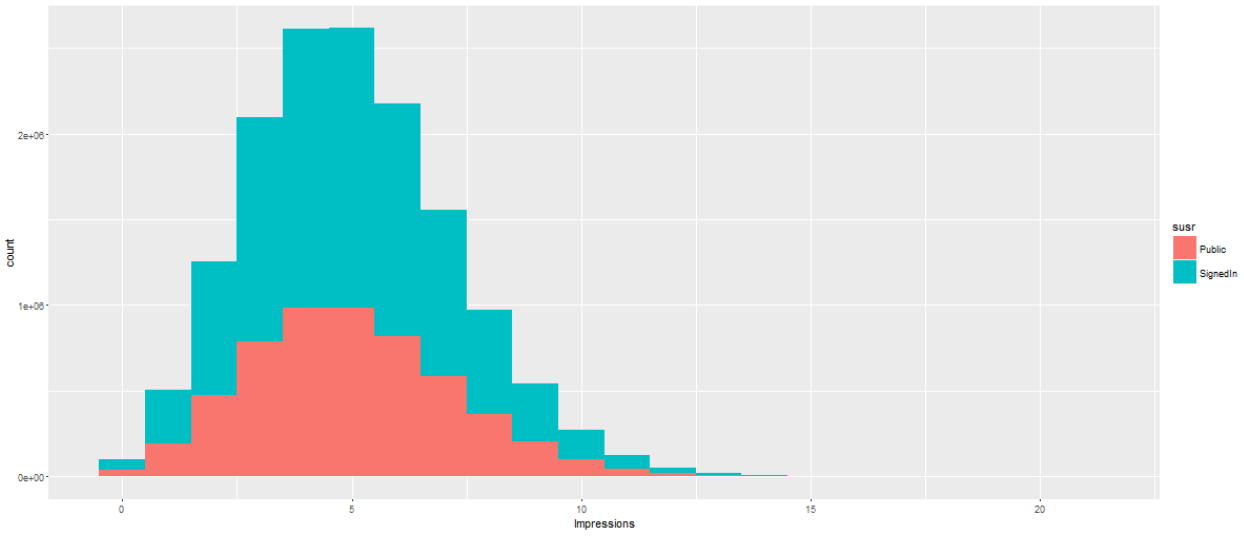


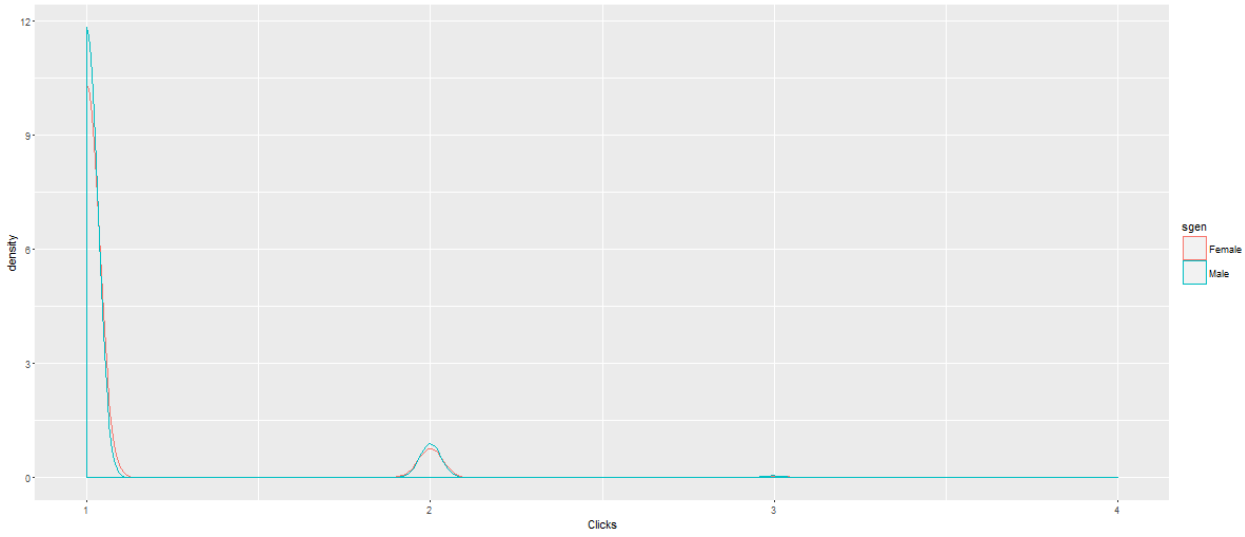
Count of Males and Females under 18.

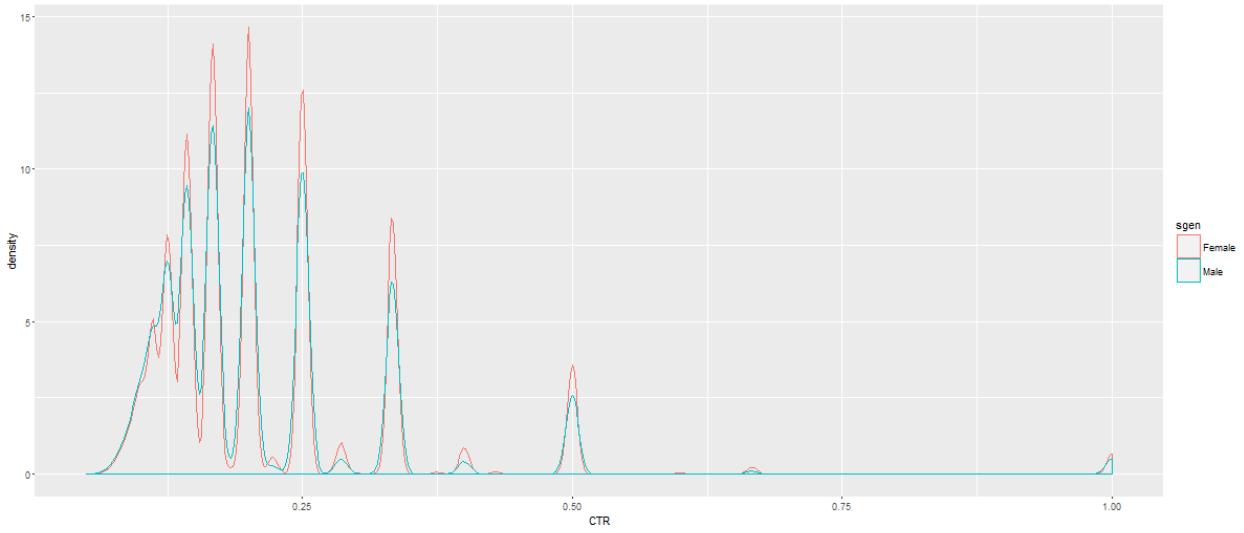


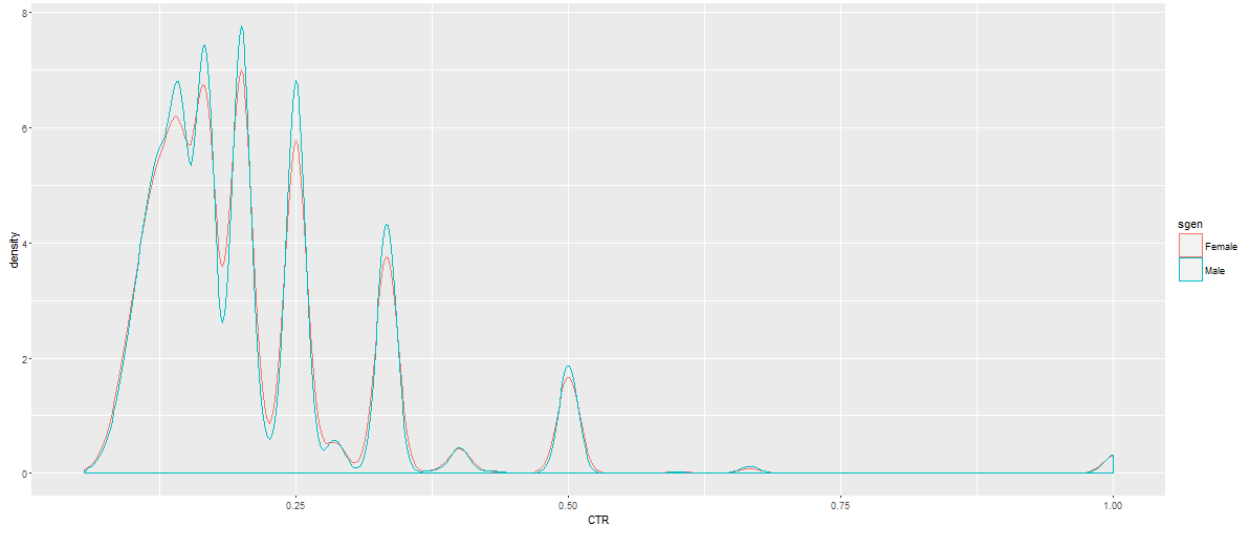

Impressions
20
15
10
5
0
Public
SignedIn
susr
susr
Public
SignedIn


count
2e+06
1e+06
0e+00
0
5
10
15
20
Impressions
sgen
Female
Male


count
2e+06
1e+06
0e+00
0
5
10
15
20
Impressions
susr
Public
SignedIn


density
12
9
6
3
0
1
2
3
4
Clicks
sgen
Female
Male


density
15
10
5
0
0.25
0.50
0.75
1.00
CTR
sgen
Female
Male


density
8
6
4
2
0
0.25
0.50
0.75
1.00
CTR
sgen
Female
Male