Problem 2 Simple EDA

Task: Performing EDA on a simple data set

The first script created i.e. NYTP2agamdeep.R basically contains data for a single day.

Variables used: Data1, data1$Age, data1$has_clicked, data1$Gender, Impressions, Clicks, col, fun, xlab, ylab.

Below are the graphs obtained for this script:

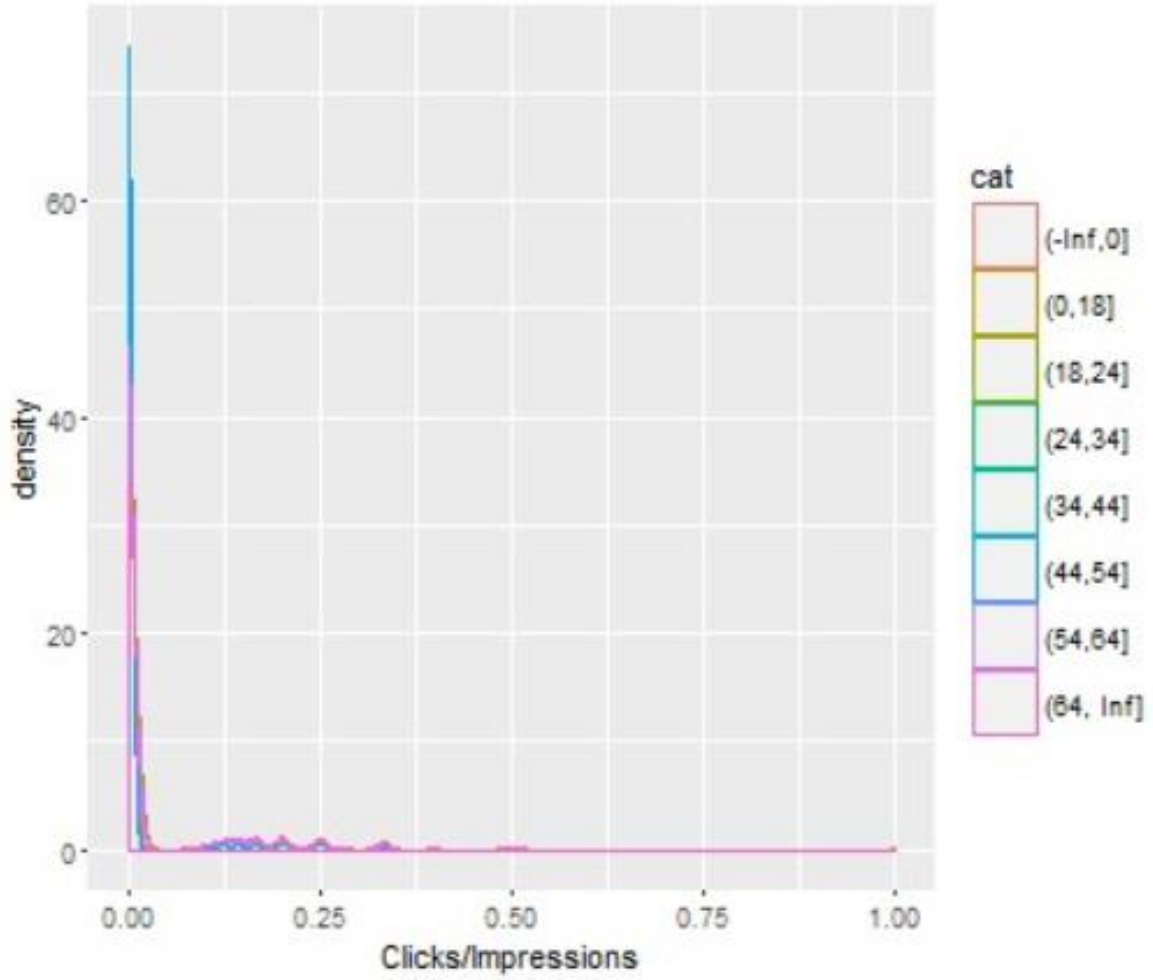


Fig 1: Clicks/Impressions vs density where Impressions > 0

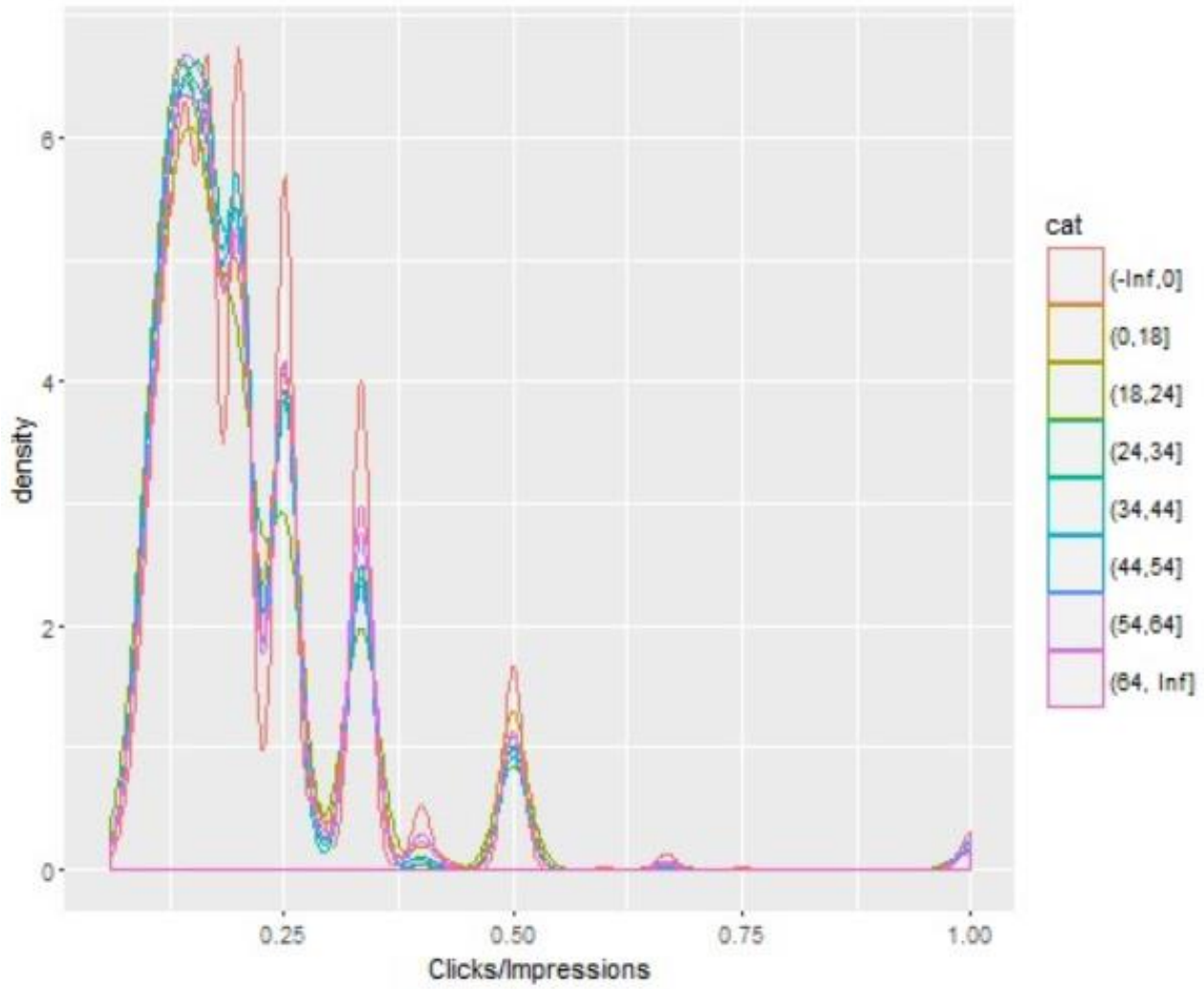


Fig2: Clicks/Impressions vs density where Clicks > 0

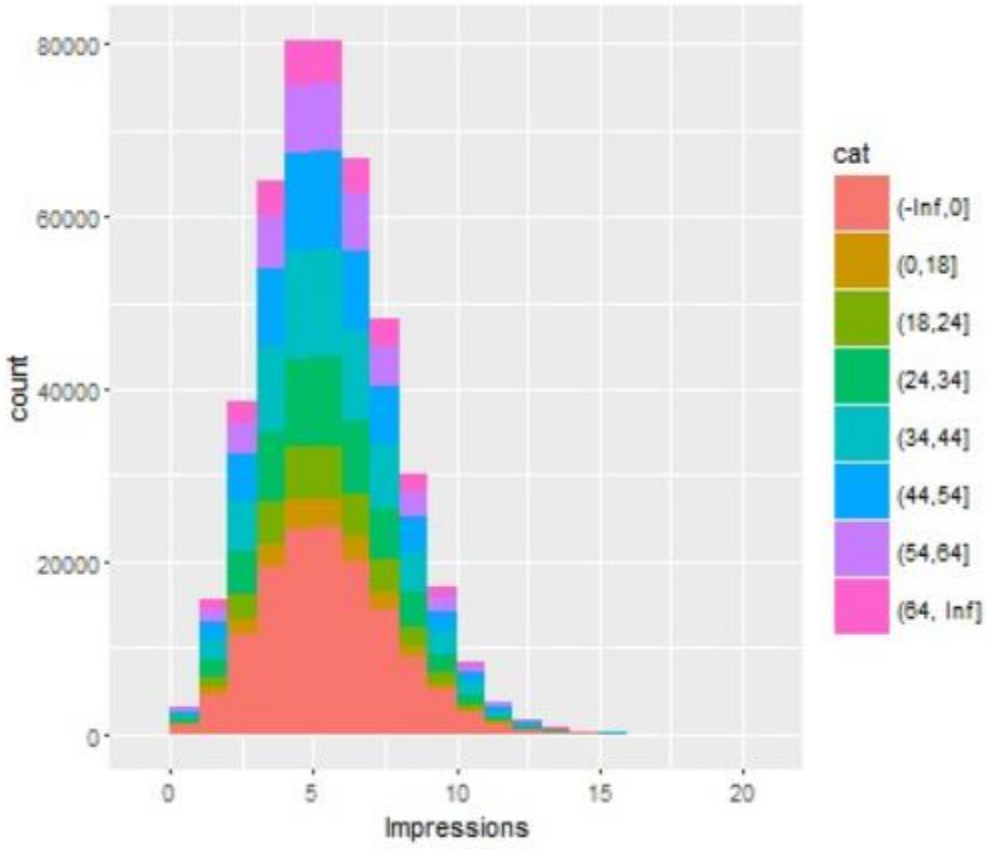


Fig 3: Impressions vs count

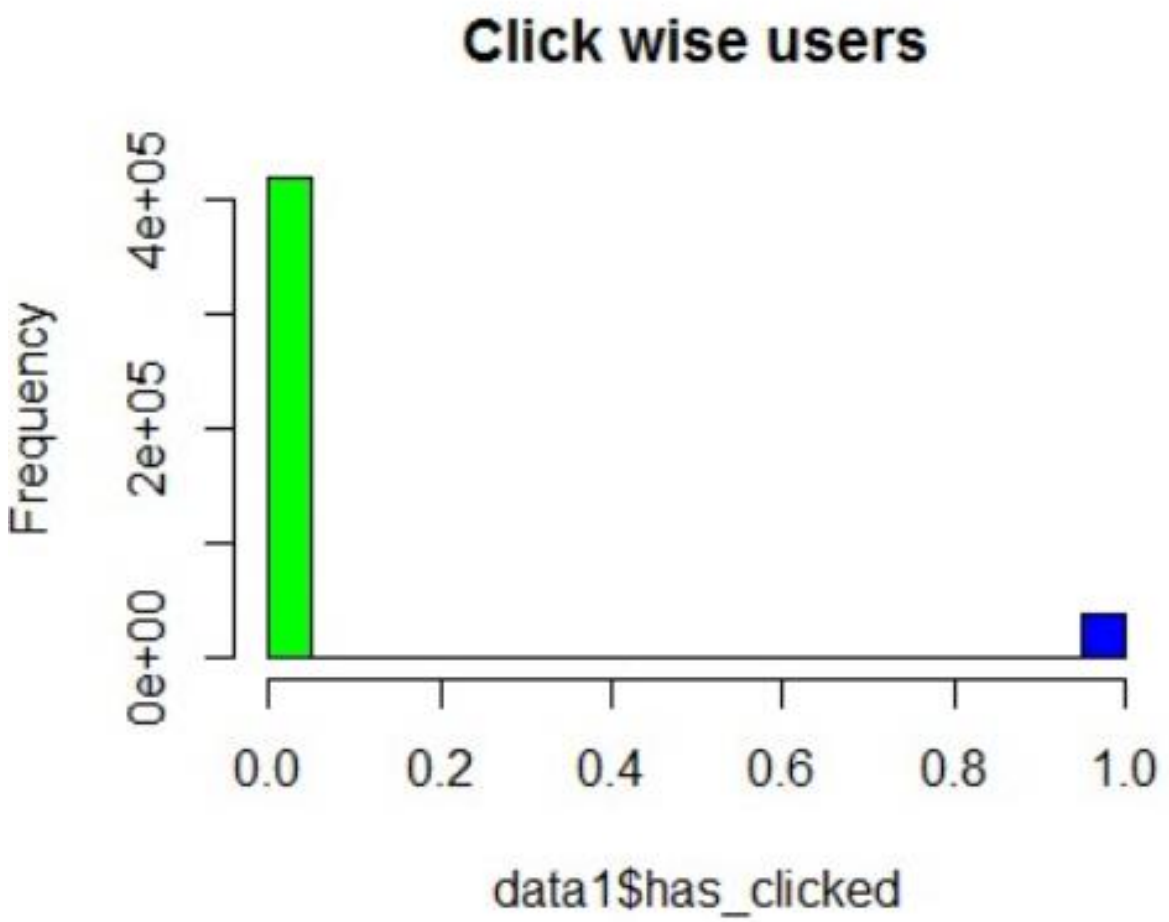


Fig 4: data1$has_clicked vs Frequency

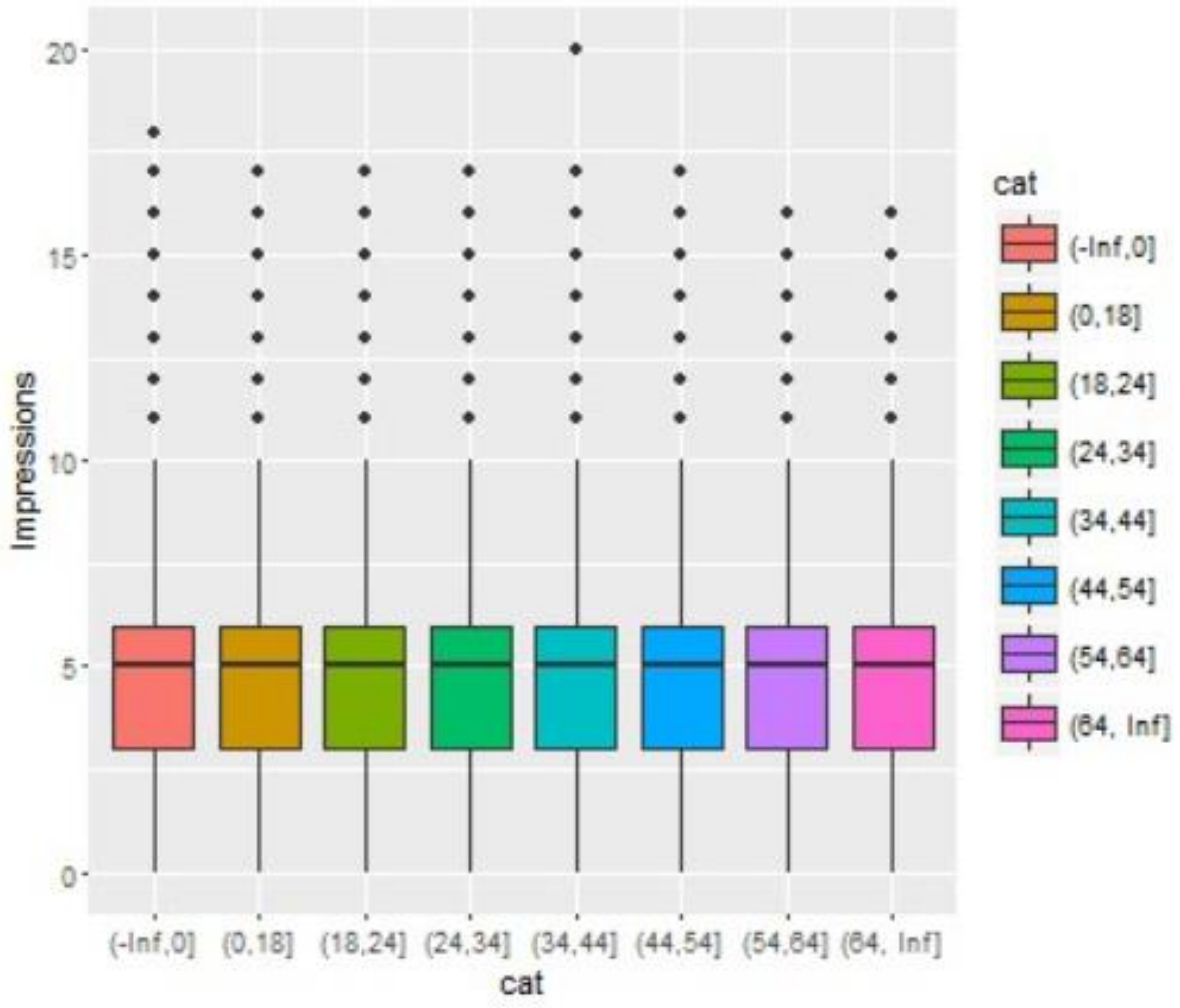


Fig 5: cat vs Impressions

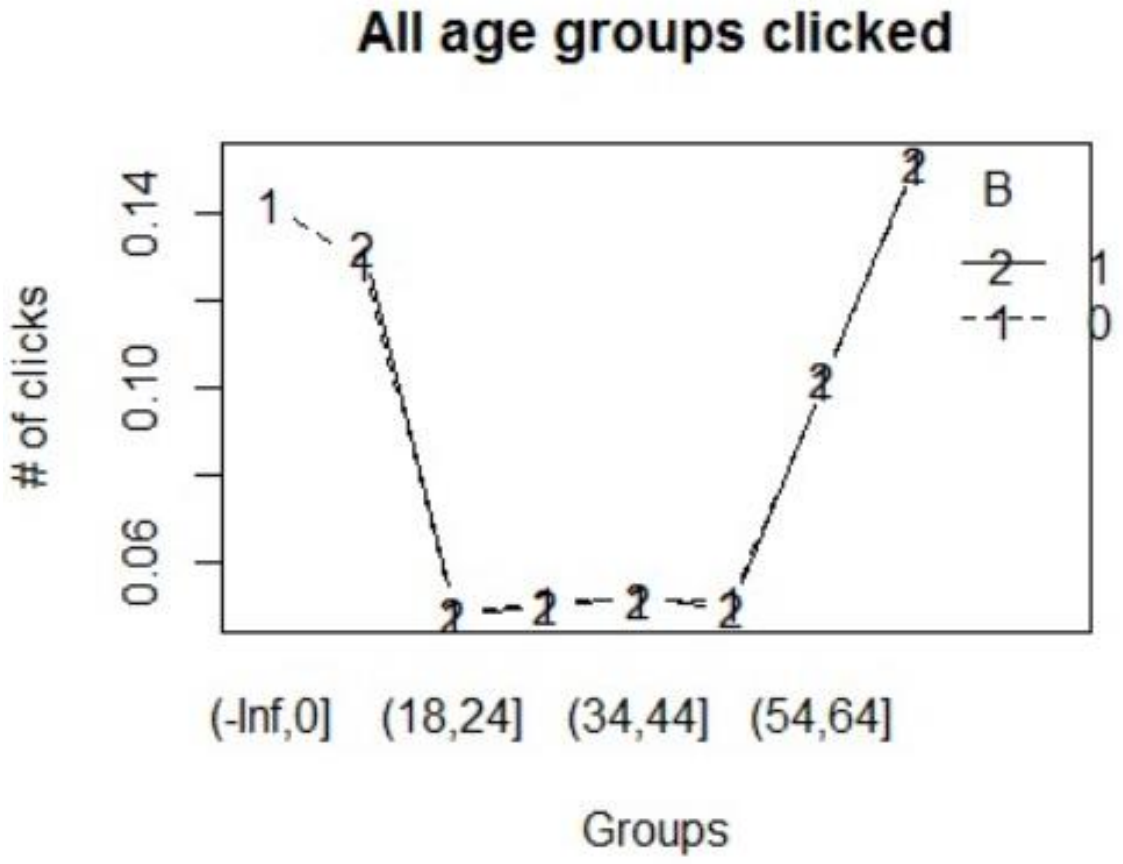


Fig 6: Age Groups vs Number of clicks

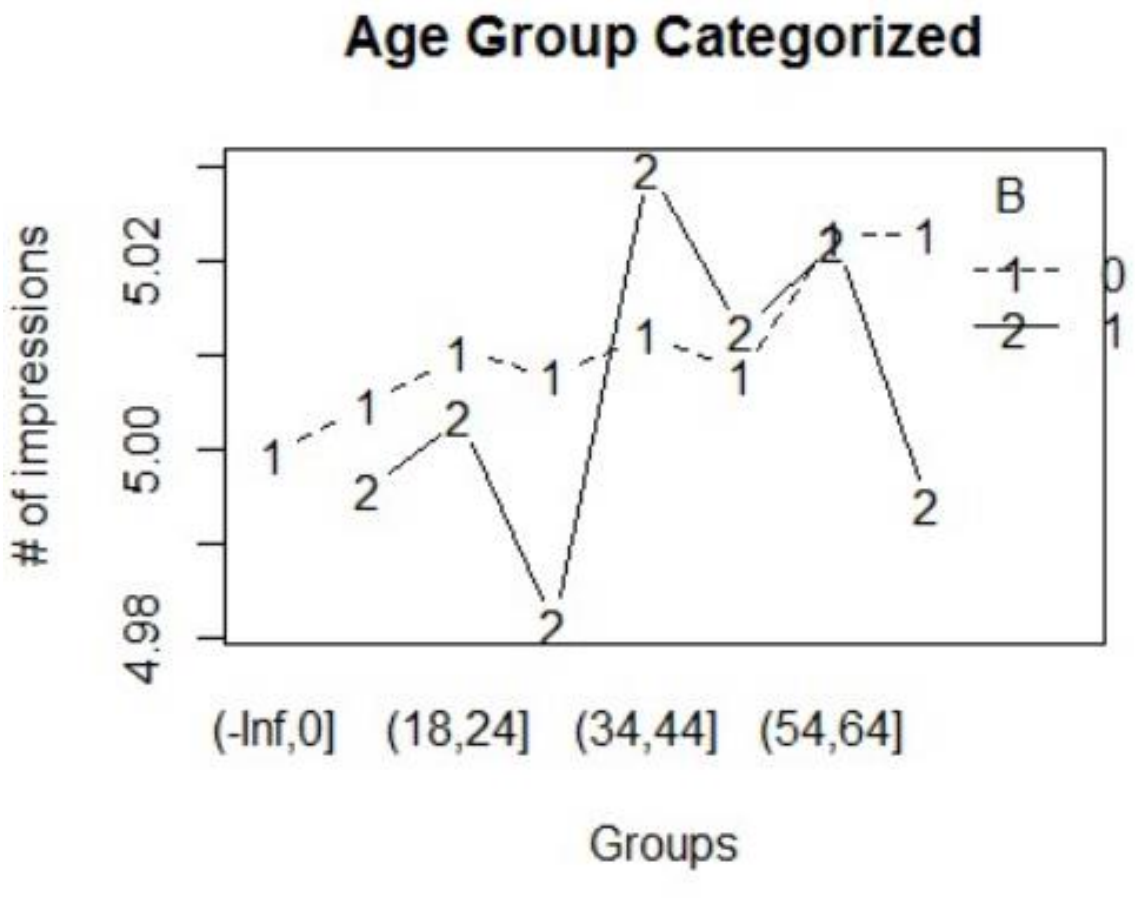


Fig 7: Age groups vs Number of Impressions

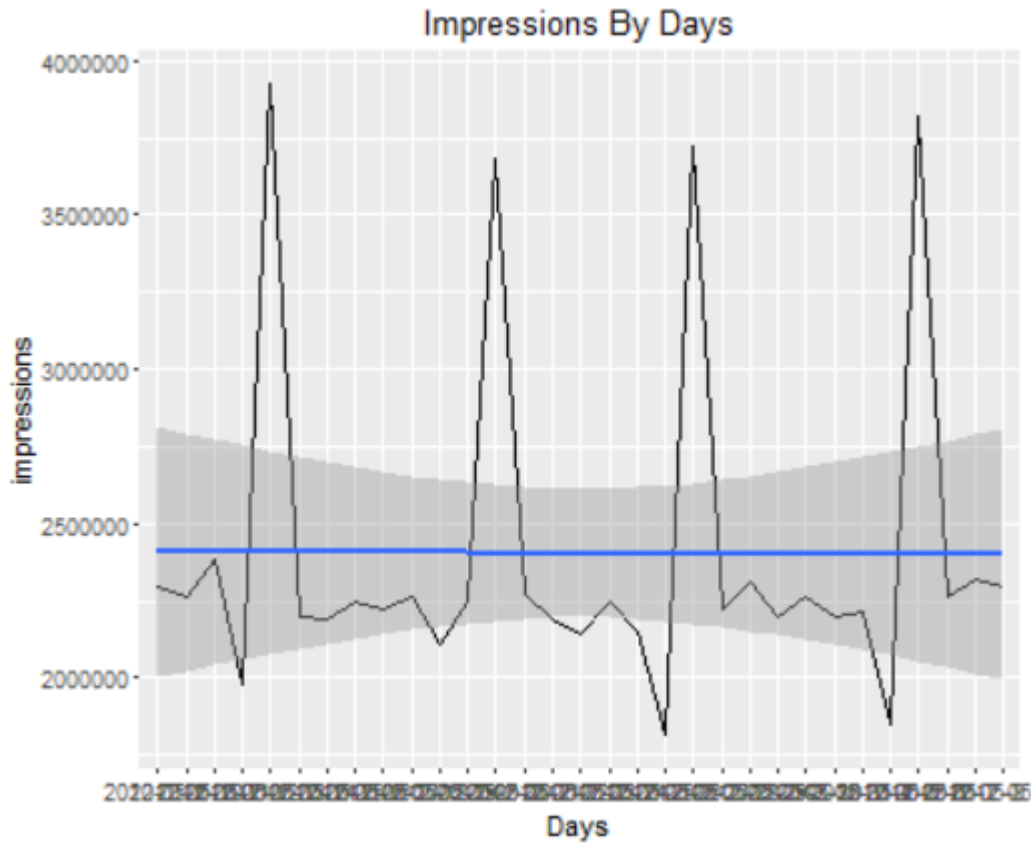
From Fig 6 and Fig 7, we can clearly deduce that there is a significant difference in impressions when we see genders of age groups not between [18,64].

Now in order to reach to a conclusion, it was important to examine to data metric obtained. Below is the snapshot of the metric. It is completely evident that 20% of the distribution is among the first age category

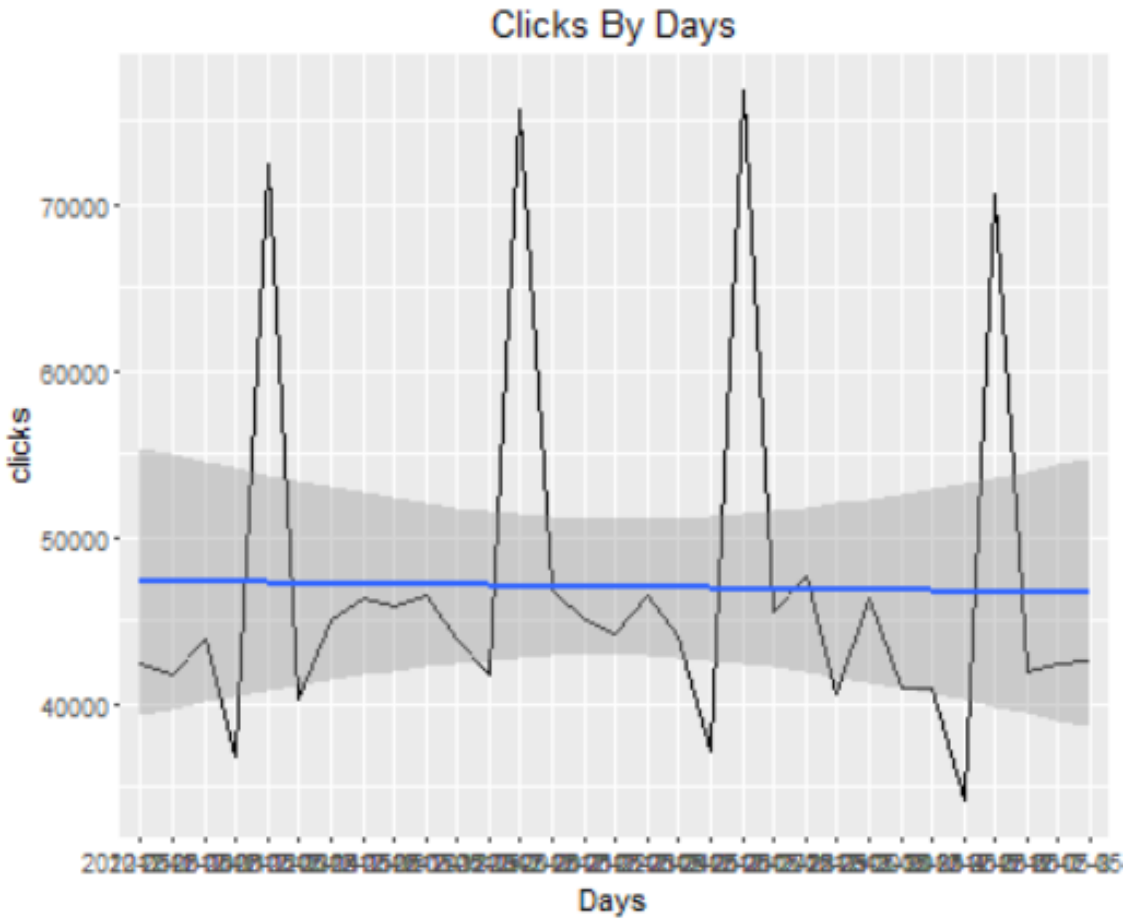
Console ~/

```
> apply(res[, 1:dim(res)[2]], 2, fun)
      Age Gender Impressions Clicks Signed_In cat has_clicked
0%      0      0           0      0         0   1           0
10%     0      0           2      0         0   1           0
20%     0      0           3      0         0   1           0
30%    11      0           4      0         1   2           0
40%    23      0           4      0         1   3           0
50%    31      0           5      0         1   4           0
60%    38      0           5      0         1   5           0
70%    45      1           6      0         1   6           0
80%    51      1           7      0         1   6           0
90%    61      1           8      0         1   7           0
100% 108       1          20      4         1   8           1
>
```

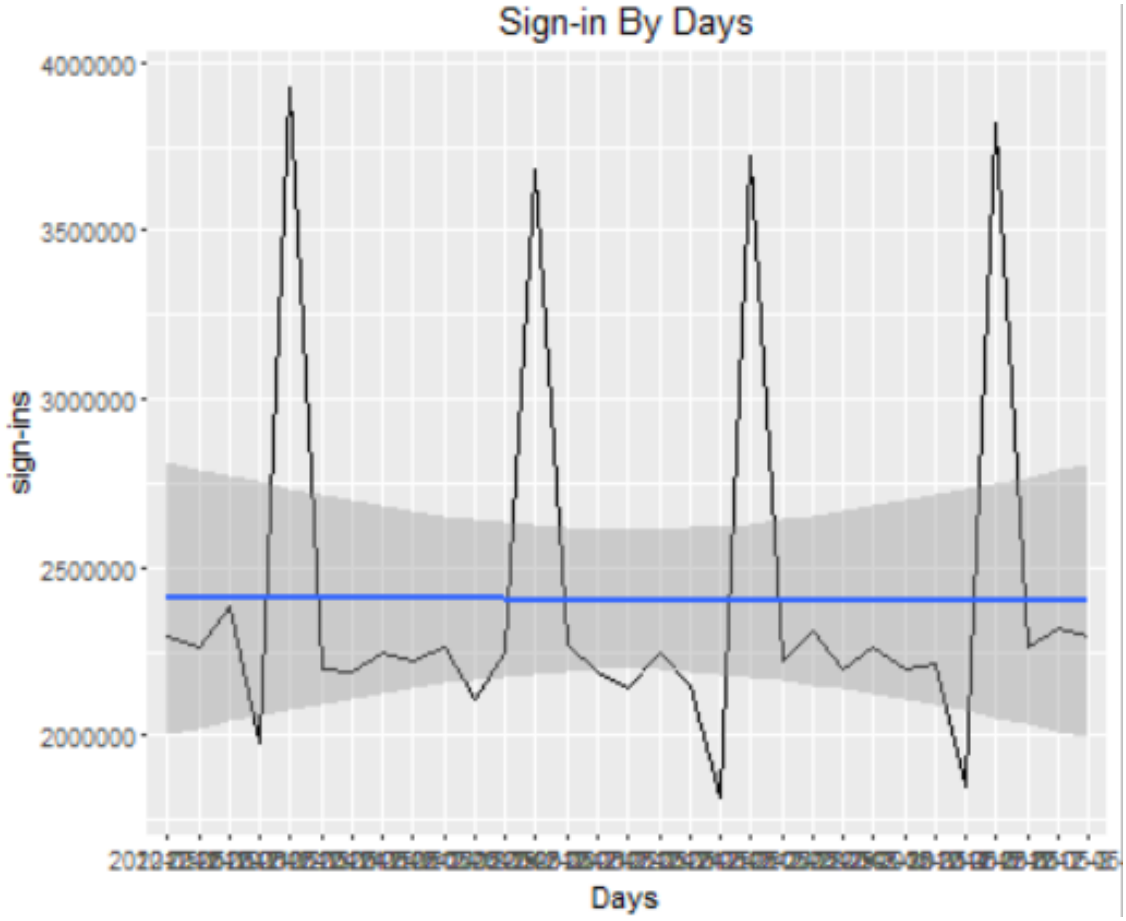
Now let us try to extend the per day analysis to a complete month data



As evident in the graph above, the Impressions are the maximum in the first quarter of the month May. There is a slight dip in the subsequent quarters of May and then there is a slight increase in the final quarter of May.



Now let us examine the graph above. The clicks were maximum in the third quarter of May. The trend also suggests that the second highest number of clicks happened somewhere around the middle of May.



As evident in the graph above, the sign-ins are the maximum in the first quarter of the month May. There is a slight dip in the subsequent quarters of May and then there is a slight increase in the final quarter of May.

Console ~/

```
> summary(model)

Call:
lm(formula = Clicks ~ Impressions + Signed_In + Age)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4546 -0.1230 -0.0832 -0.0395  5.7800

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.412e-02  2.250e-04   196.1   <2e-16 ***
Impressions  1.955e-02  3.636e-05   537.6   <2e-16 ***
Signed_In   -1.131e-01  3.140e-04  -360.2   <2e-16 ***
Age          1.006e-03  6.306e-06   159.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3139 on 14905861 degrees of freedom
Multiple R-squared:  0.03199,   Adjusted R-squared:  0.03199
F-statistic: 1.642e+05 on 3 and 14905861 DF,  p-value: < 2.2e-16

> |
```