CSE 587: Project 1
Problem 2 -Simple EDA

Lalith Vikram Natarajan, UBIT ID: lalithvi, Person No: 50169243

March 5, 2016

1 Objective

Performing EDA on a simple data set

2 Data Source

The data got from the link - **http://stat.columbia.edu/ rachel/datasets/nyt1.csv**

3 Plots, Analysis and Results

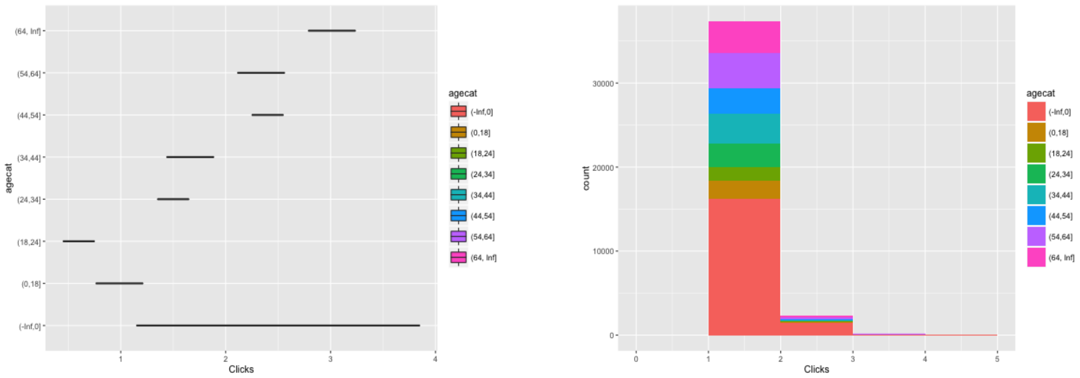
From the charts in figure 1(a) and 1 (b), seniors take a while to go to the place they want to in a webpage, hence the increased number of clicks, majority of the users click once or twice.

A click-through is the actual number of times someone has taken their cursor, placed it on the advertising image and used their mouse to click on that image. It evens out in the naive graph (figure 2(a)) and generally has a density close to zero for different age groups (figure 2(b)). This indicates that more efforts have to be taken to engage the users in the site.

Males of age 34-44 access the site than any other demographic. After cleaning (making sure we record the gender only of people who have logged in), we see that the no. of males accessing the site is more than half of the females. (figure 3(a) and figure 3(b))

figure 4 shows a graph between the impressions and age category. Mean is 5 impressions. figure 5(a) We now compare male and female impressions density. As expected, males have greater impression density.

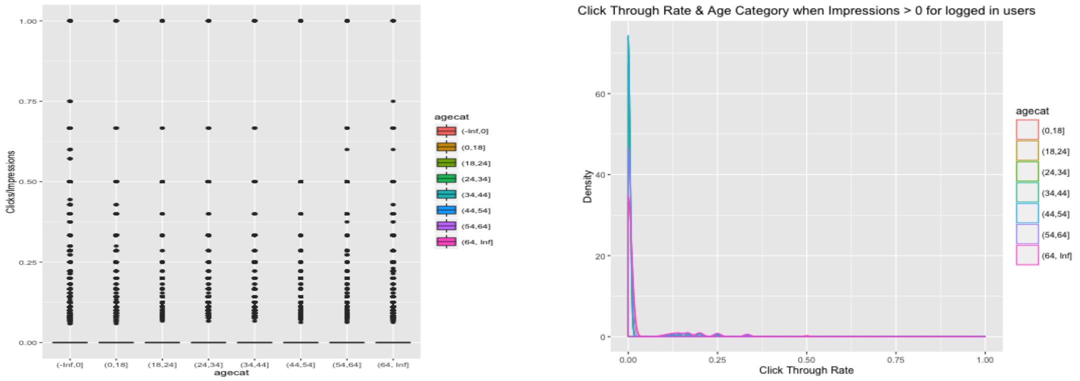
Male and female impression density are nearly equal among different age groups. The application makes a flaw of recording everyone as Gender 0 when they are not logged in.



(a) BoxPlot of Clicks vs Age categories

(b) Clicks vs Age categories

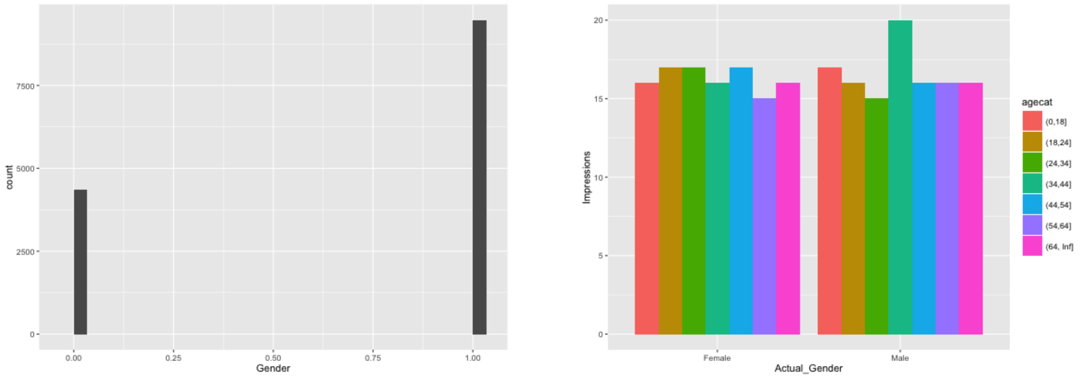
Figure 1: Click plots



(a) CTR vs Age Categories Naive Plot

(b) CTR vs Age Category Bar Chart

Figure 2: CTR plots



(a) Bar Plot of Gender Count

(b) Bar Chart of Gender vs Age Impressions

Figure 3: Gender Demographic

This explains the fat pink bar. Plotting this also enables us to figure out the flaws in the existing system. (figure 6)

Combining multiple days tweets gives us an impression that on a Friday, (assuming day 1 starts on Sunday), more people log in to the site. This is shown in both the charts (figure 7). Also they are completely different data sets, one is from days 1 -7 (figure 7(a))and the other is from days 14 - 21 (figure 7(b)) (this chart also plots only persons of the female gender). This proves that this is a prevalent pattern.

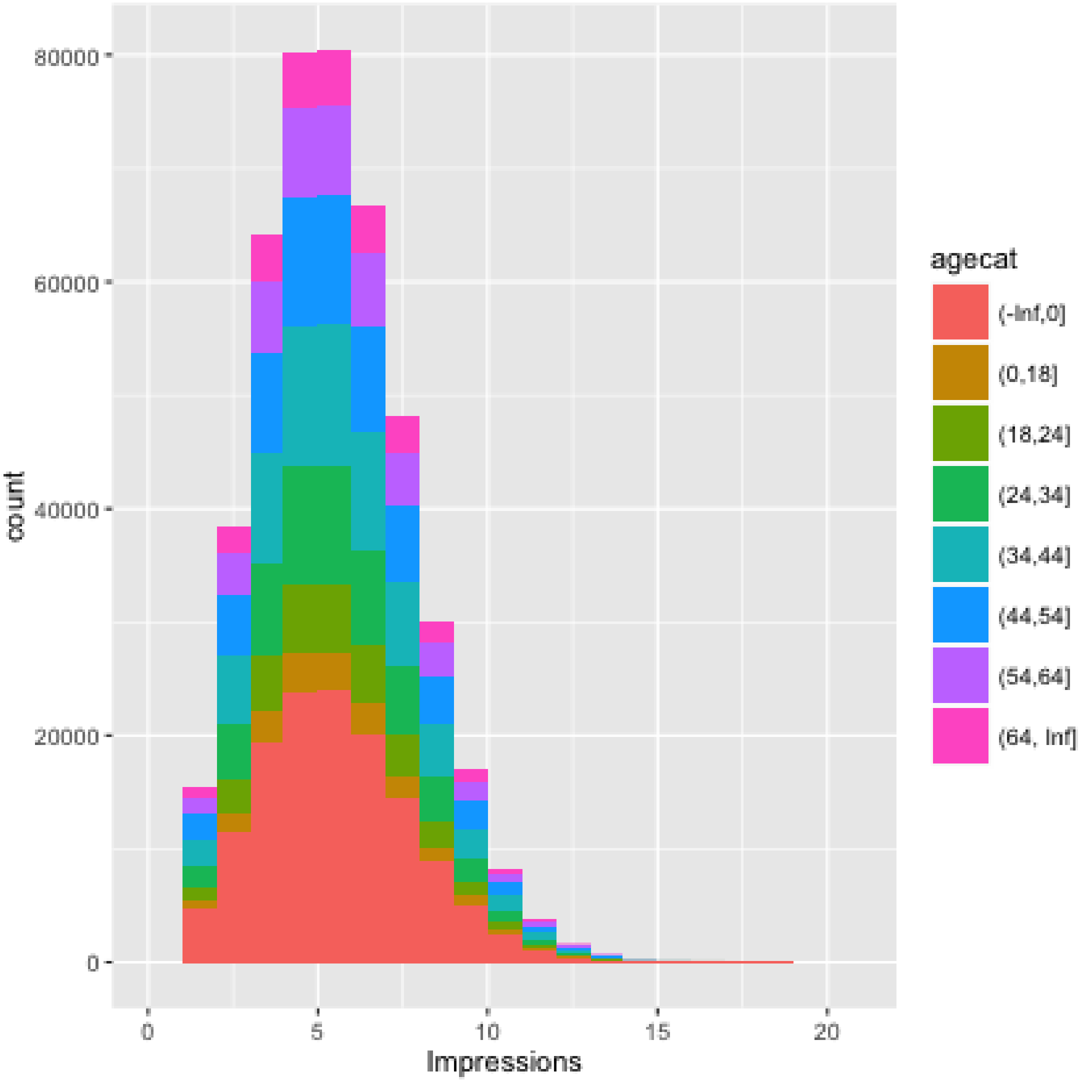
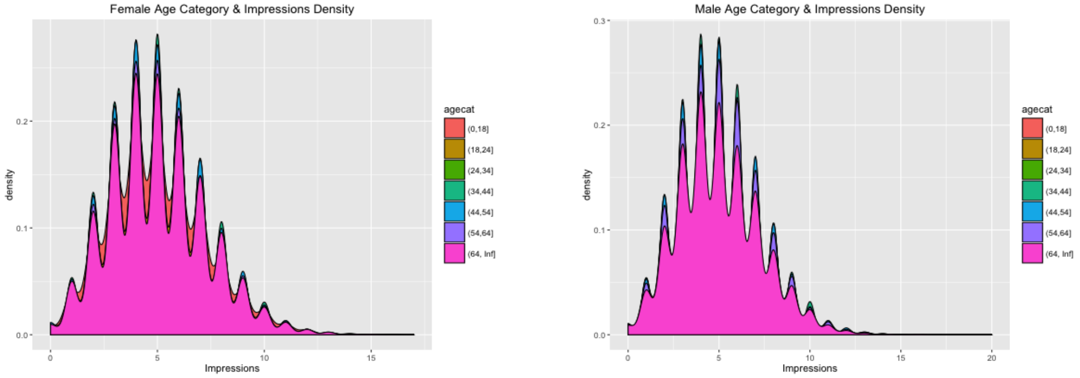


Figure 4: Impressions vs Age Category



(a) Density chart of Female Age and Impressions (b) Density chart of Female Age and Impressions

Figure 5: Genderwise Impression Density

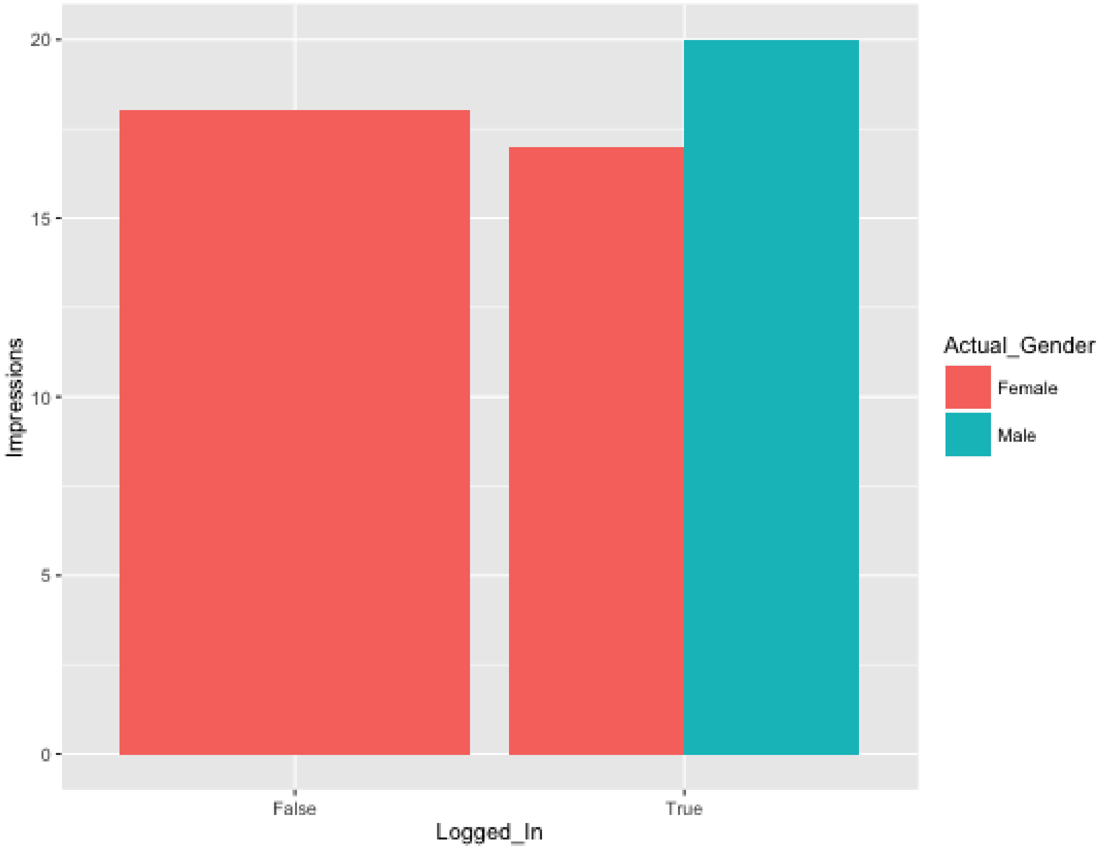
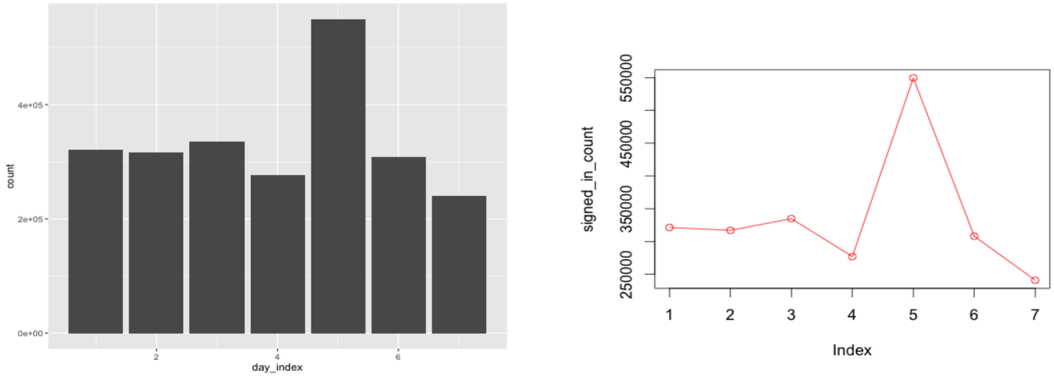


Figure 6: Logged in Status vs Gender



(a) Logged in users over a week (days 1-7)

(b) Signed in users over a week (days 14-21)

Figure 7: Weekly Plots