Ryan Rishi

Professor Mohler

CSCI 183: Data Science

11 April 2015

Homework 1 – *New York Times* May 2012 Demographic Analysis

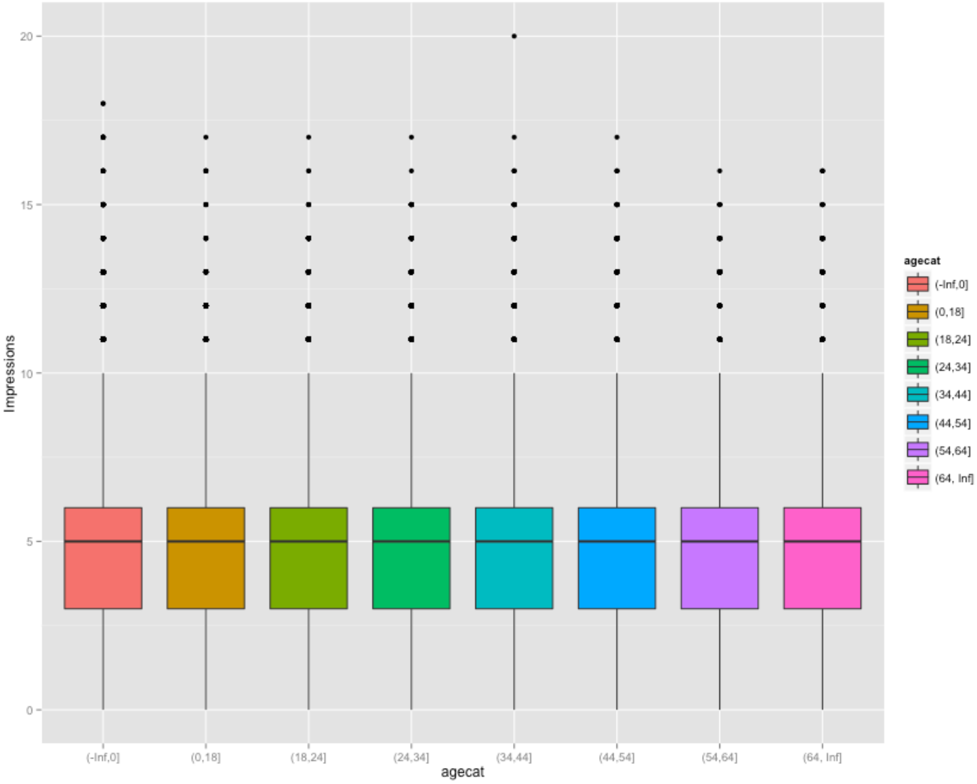
Definition of Terms

Impressions: the number of times a link is seen by a user

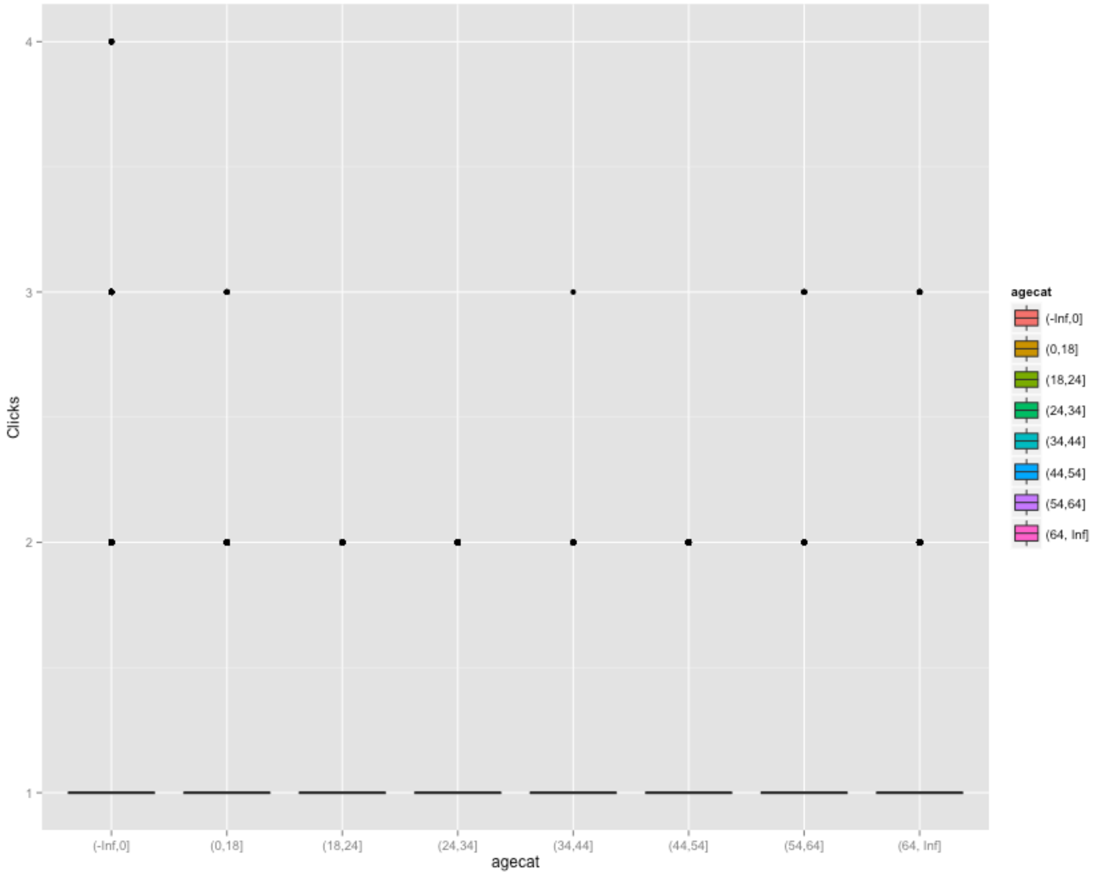
Clicks: The number of times a user clicks on a link

Click-through Rate (CTR): Measurement of the number of users that click on a particular link (impressions / clicks)

Impressions v. Age Category



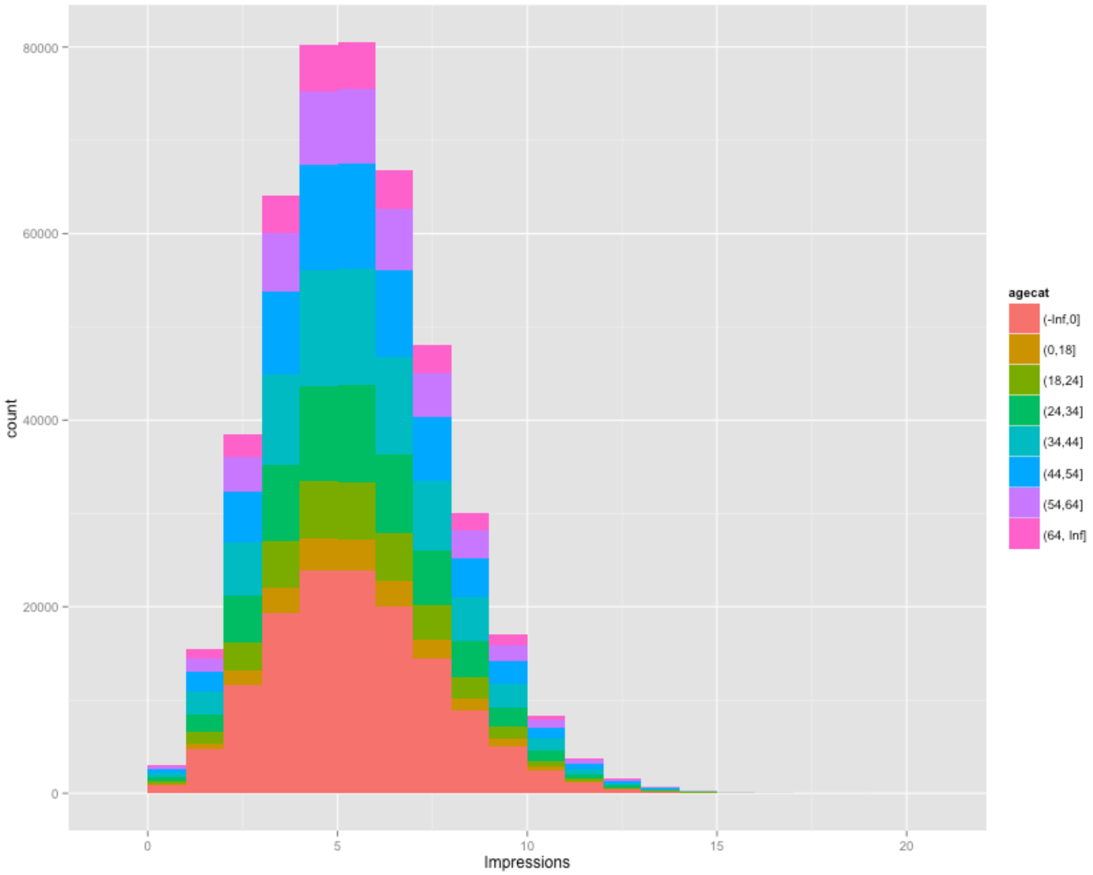
Clicks v. Age Category



It is a little comical that some of the observations in the CSV data file have 0 as the age. I am not sure if this is people actually putting 0 in the age box, or simply a missing data point that defaulted to 0.

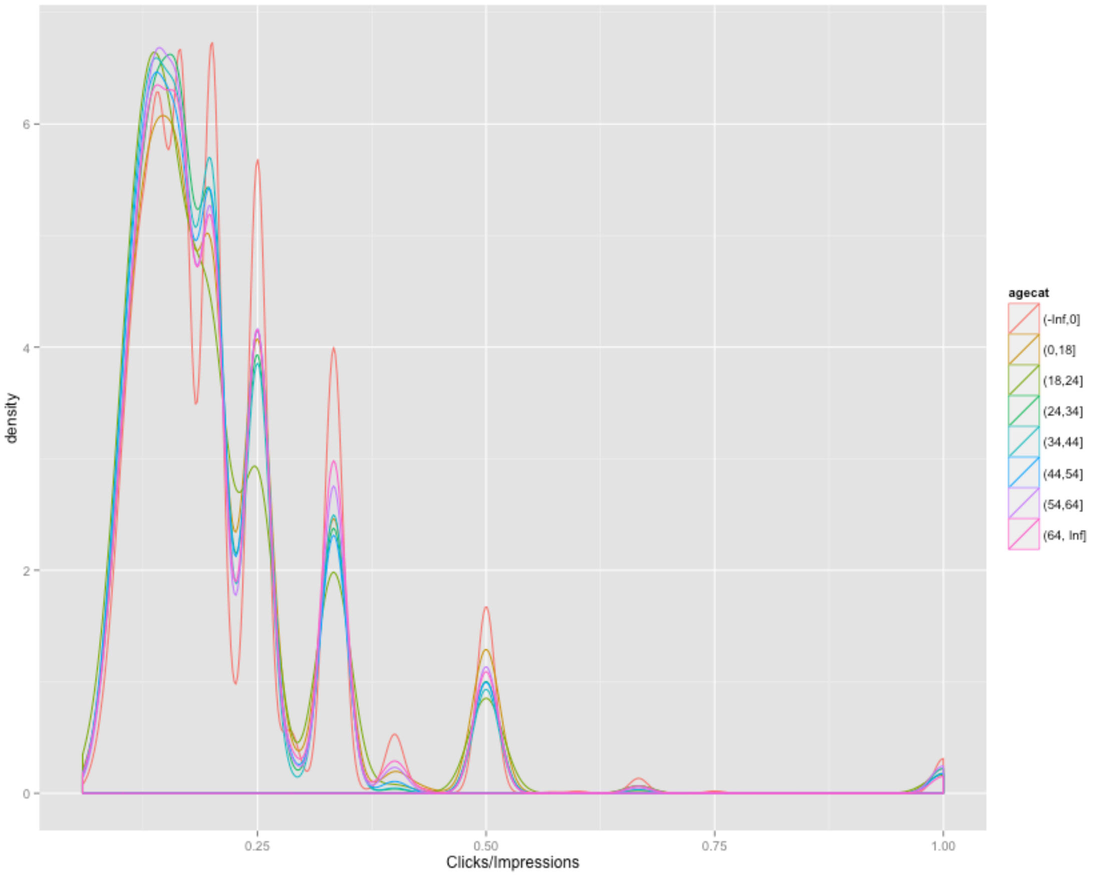
That being said, it is probably best to ignore the first column of these charts (the (-Inf, 0] age category).

Count v. Impressions



It looks like the average number of impressions is around 5. Again, a large portion of this data falls in the (-Inf, 0] age category. It appears that every age category increases by roughly the same ratio for each step in impressions, thus it is safe to say that this is a general trend regardless of age category.

Density v. Click Through Rate



This is the most interesting graph I came across. There is a clear decline in the density as CTR increases, but there is also an error in the way we analyzed this data. Since the number of clicks and the number of impressions are both integers, we limit the CTR to ratios of low integers. This is evident because the peaks are at .125 (1/8), .25 (1/4), .33 (1/3), .375 (3/8), and .50 (1/2). This is not critical in our analysis, but it is interesting to observe.

Other Remarks

This is my first time using R. I had tinkered around with it last spring, but never really got into it. I'm excited to learn more about R's capabilities and explore data in new and exciting ways.