Giovanni Briggs
4/6/14
CSCI 183

HW1: Exploratory Data Analysis

Overview

The dataset we are examining is user data from the *New York Times'* website during April 2012. It contains data such as the user's gender, age, clicks and impressions (number of times an ad was displayed to the user) for a given day. A user's age and gender are displayed only if the user is signed in. Since having gender and age available makes the analysis more interesting, this analysis looks only at the subset of data where users are signed in. This analysis will refer to "users" as a collective whole, but it really only applies to signed-in users.

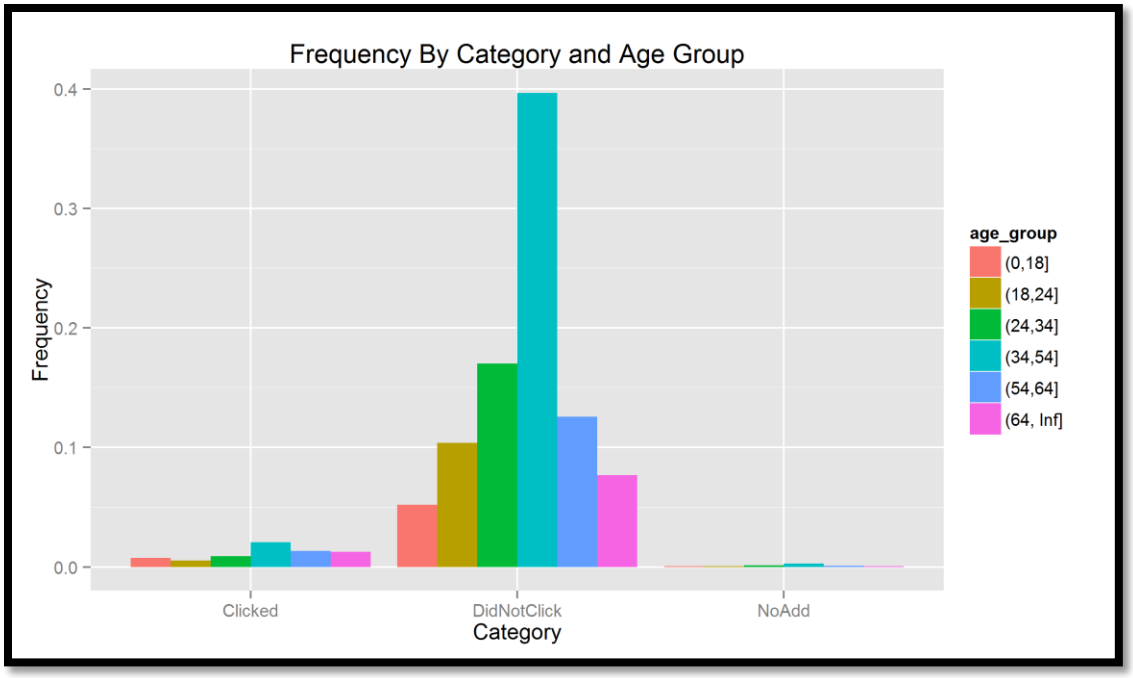
In particular, this analysis looks at how many users click on advertisements, and the average click-through-rate of users. This analysis will also examine if age and/or gender make a difference in a user's behavior.

User Behavior

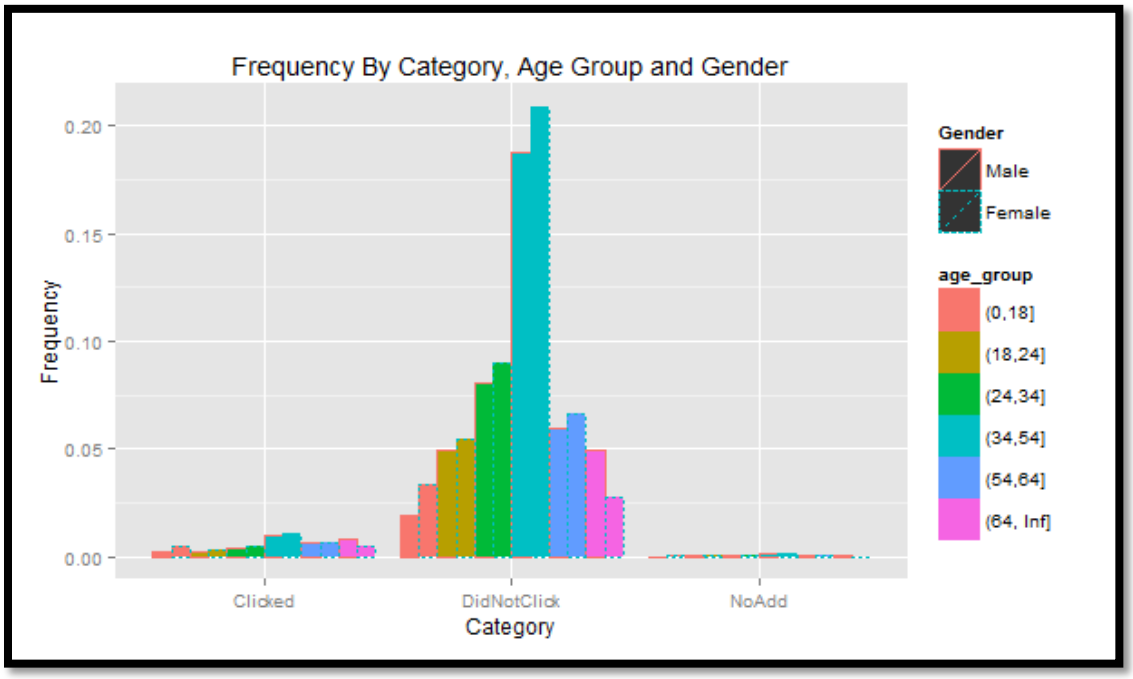
There are three categories that this analysis uses when examining user behavior:

1. Clicked
2. DidNotClick
3. NoImpressions

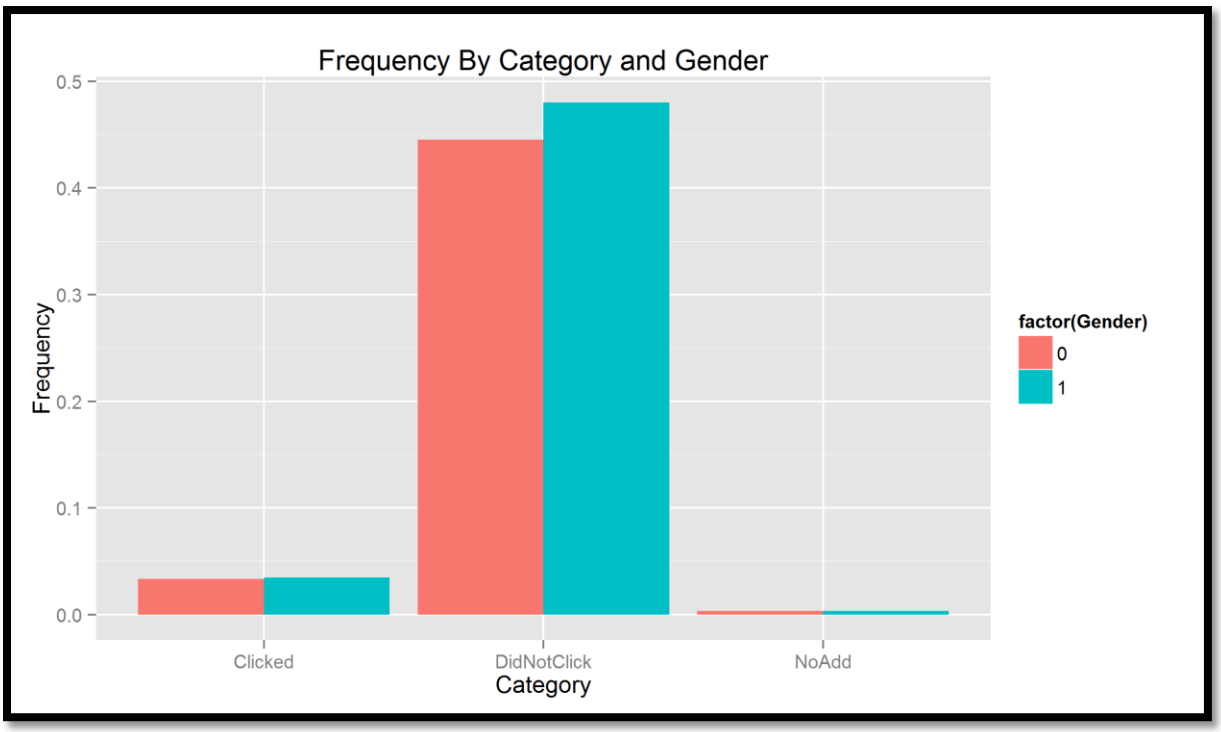
Clicked is for users who saw an advertisement and clicked on it, *DidNotClick* is for users who saw at least one advertisement but did not click on any, and *NoImpressions* is for users who were never shown an impression.



The graph above gives a breakdown of the behavior of users. An overwhelming majority of users never click on an advertisement even though nearly every user comes across at least on advertisement. We can also see from this graph that the age distribution across users is not equal. The 34-54 age group makes up the largest portion of users. We can then further subdivide each category by gender.



Male users actually make up slightly less of the population than female users and that is holds true across almost every group. The 54-64 and 64+ age groups have the opposite trend, but in general, female users are a slight majority. We can see this even more clearly if we remove age groups from the graph.

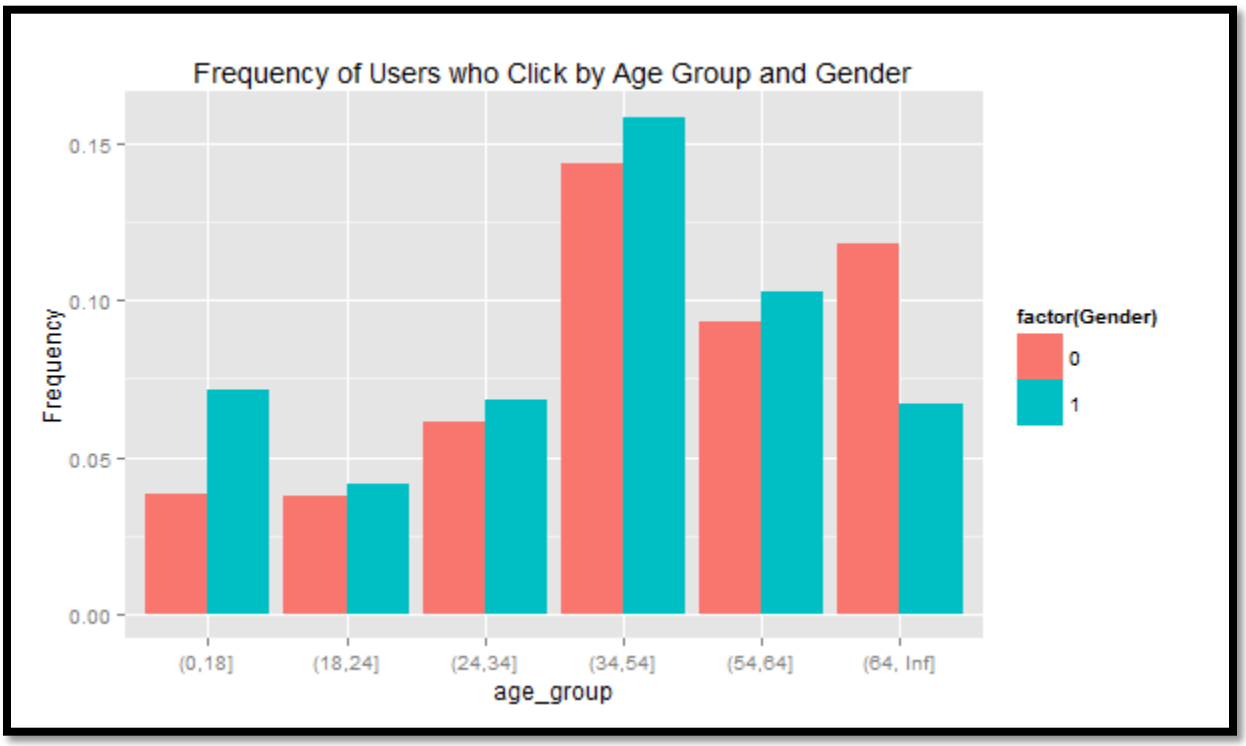


What these graphs tell us then is that very few users will click on advertisements, and there is no apparent difference in behavior between males and females. While age is not equally distributed throughout our dataset, gender is.

Click-Through-Rate

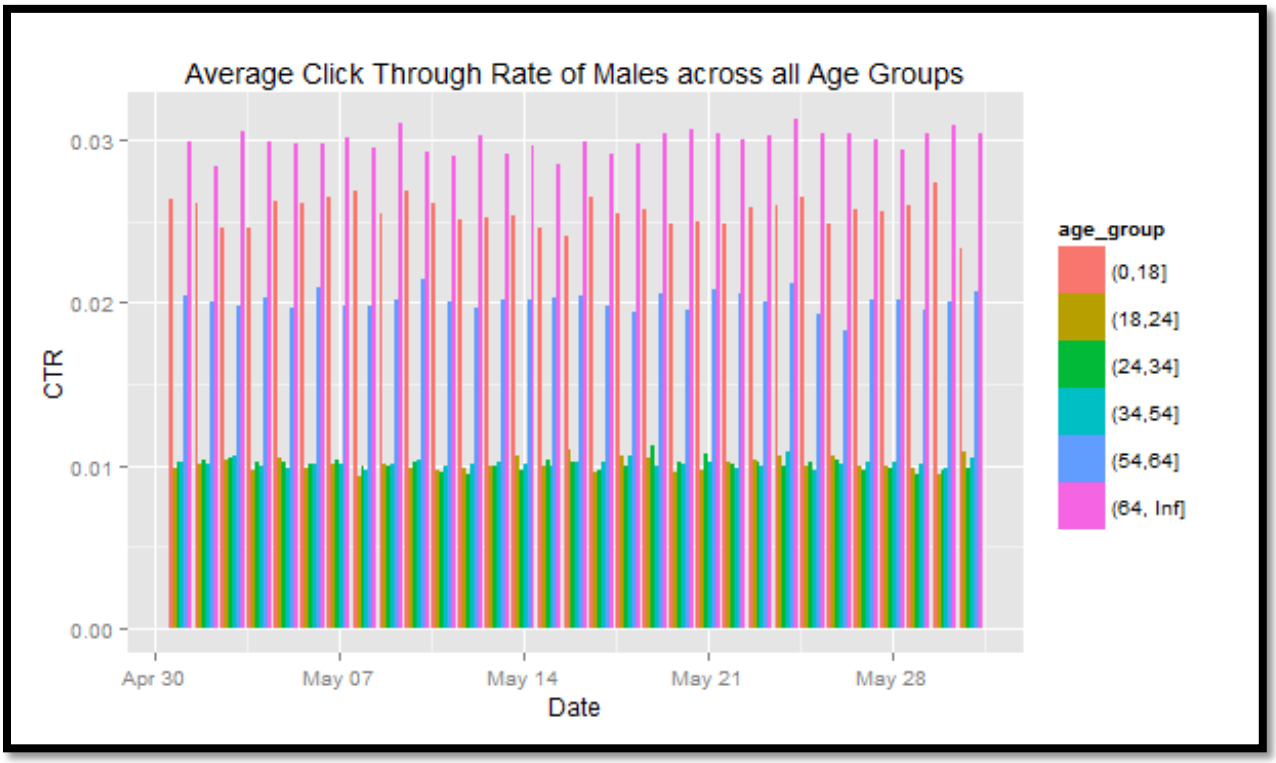
A user's click-through-rate is defined as: $CTR = \frac{\#clicks}{\#impressions}$. In order for a user to have a CTR, they need to have at least 1 impression. In other words, they need to have seen at least one advertisement. A CTR of 0 means that the user decided the advertisements were not worth their time and therefore did not click on them. A CTR of 100 means the user was very excited and clicked on everything.

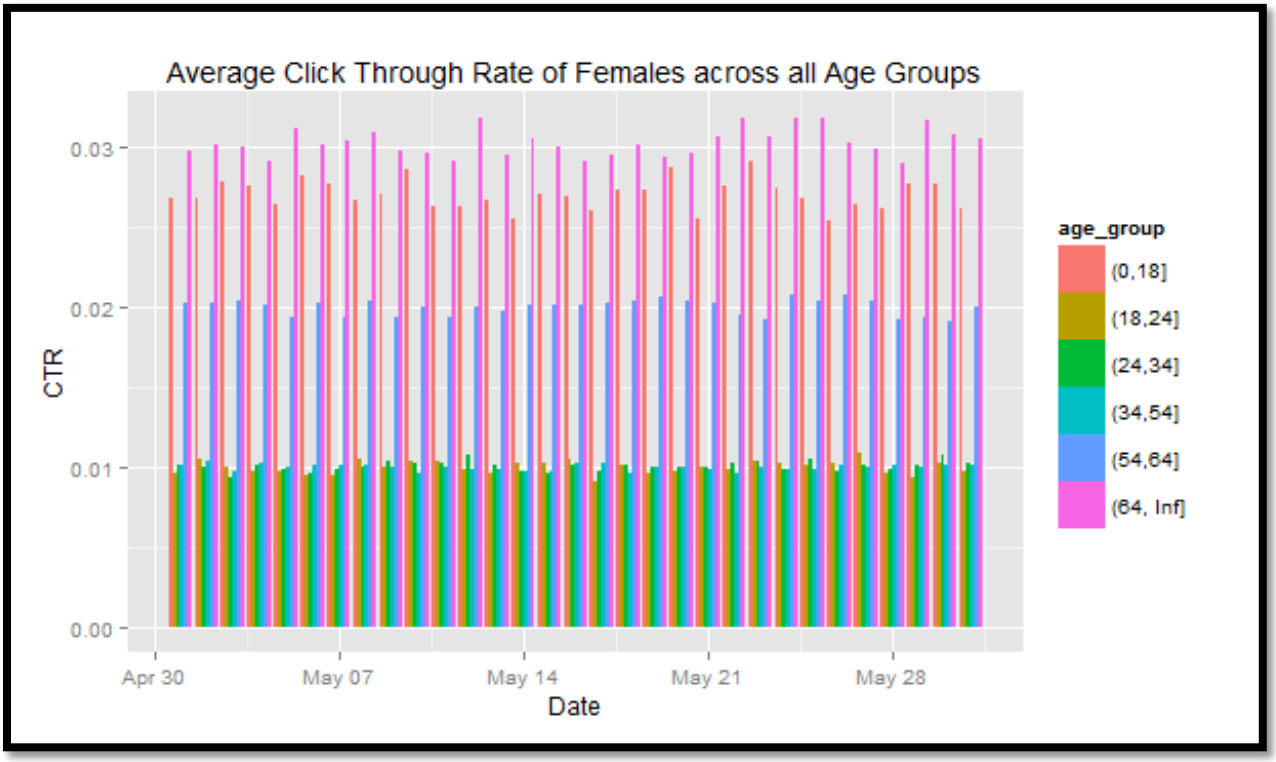
While very few users tend to click on advertisements, it is still useful to see if certain users click on advertisements more often than others. If a certain demographic has a higher CTR than others, then the advertisements used are more effective at targeting those groups. Those users are the most likely to click on an advertisement if one is displayed. We can start by looking at the frequency of each age group who clicked on an advertisement at least once.



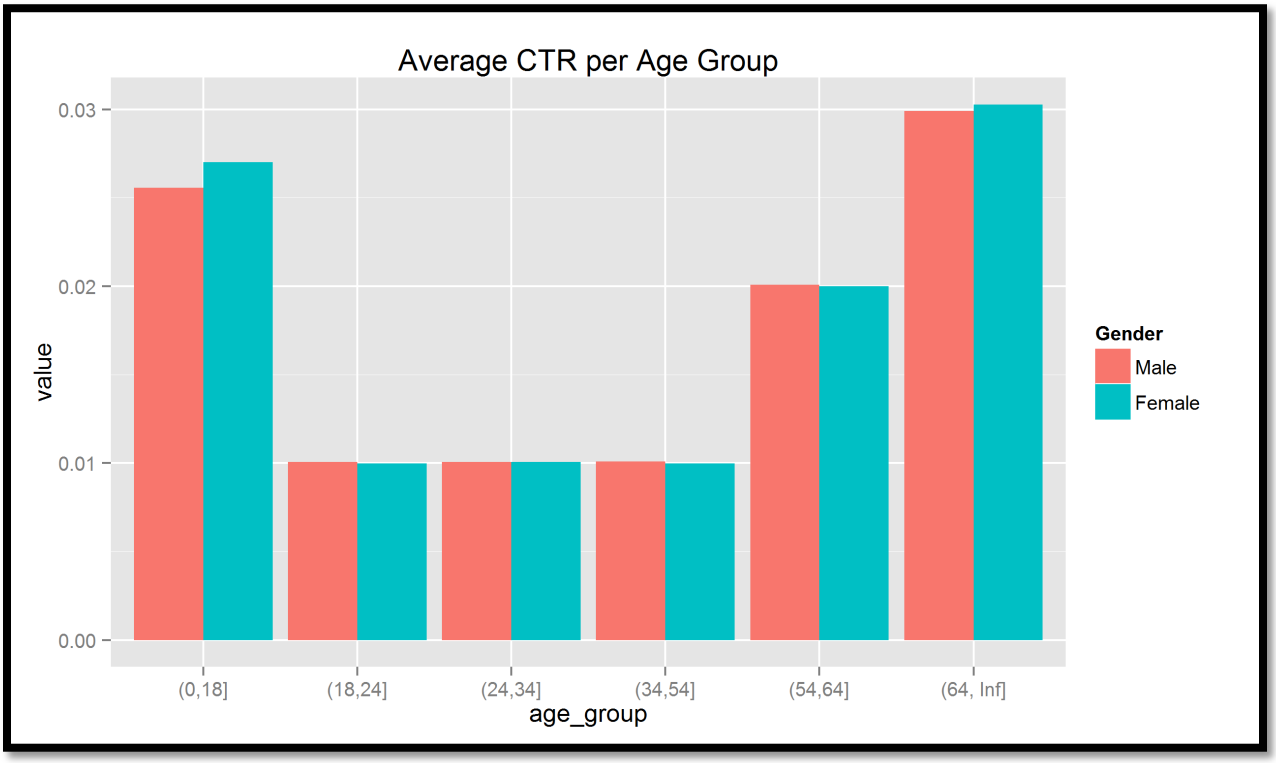
Again, we can see that gender is very close to evenly distributed but age is not. The 34-54 age group still makes up the largest percentage of users. However, this only shows the users

who clicked on one or more advertisements. By taking the average CTR of each age group and subdividing that by gender for each day, we can look at which group has the highest CTR.





What we see from these two plots is that the 64+ age group for both genders has the highest CTR. Furthermore, this appears to be consistent across time. The 0-18 age group comes in second and rarely comes close to surpassing the 64+ age group. The 35-54 group, which is the largest, has one of the lowest number of CTRs across both genders. We also see that there is no serious difference in CTR across genders. We can reduce the plots above into single points by taking the average of the average CTRs for the month of April.



The 64+ and 0-18 age groups nearly triple the CTR of the 34-54 age group. This graph also makes it apparent that gender is not a factor in a user's click thru rate. Interestingly, the 18-24, 24-34, and 34-54 age groups all have almost identical click through rates.

Summary

The results from this analysis show that gender does not make an impact in a user's behavior, but age does. All results were consistent across gender, but varied with age. However, across all age groups, users do not frequently click on advertisements. Most users will not click on an advertisement when they see one. The 64+ and 0-18 age groups click the most frequently. Their click through rates are nearly triple the rates of other users. This

indicates that advertisements on the *New York Times'* website are more effective at targeting these groups. These ads are especially bad at targeting users between the ages of 19 and 54, even though this group makes up over 75% of all users.