Samuel Kelly
4/6/2014
CSCI 183

I followed the steps and the code given in the book in order to get the outputs that I got. The code I followed the book and inputted into the command line and I just removed my mistakes before posting them below. I was having issue with posting to github so I posted the code below. My interpretation of the results can be found after the outputs.

code:

```
data1<-read.csv(url("http://stat.columbia.edu/~rachel/datasets/nyt1.csv"))
data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
head(data1)
data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
summary(data1)
summaryBy(Age~agecat, data =data1, FUN=siterange)
summaryBy(Gender+Signed_In+Impressions+Clicks~agecat,data =data1)
ggplot(data1, aes(x=Impressions, fill=agecat))+geom_histogram(binwidth=1)
ggplot(data1, aes(x=agecat, y=Impressions, fill=agecat))+geom_boxplot()
data1$hasimps <-cut(data1$Impressions,c(-Inf,0,Inf))
summaryBy(Clicks~hasimps, data =data1, FUN=siterange)
ggplot(subset(data1, Impressions>0), aes(x=Clicks/Impressions,colour=agecat)) +
geom_density()
ggplot(subset(data1, Clicks>0), aes(x=Clicks/Impressions,colour=agecat)) + geom_density()
ggplot(subset(data1, Clicks>0), aes(x=Clicks, colour=agecat))+ geom_density()
data1$scode[data1$Impressions==0] <- "NoImps"
data1$scode[data1$Impressions >0] <- "Imps"
data1$scode[data1$Clicks >0] <- "Clicks"
data1$scode <- factor(data1$scode)
head(data1)
clen <- function(x){c(length(x))}
etable<-summaryBy(Impressions~scode+Gender+agecat,data = data1, FUN=clen)
```

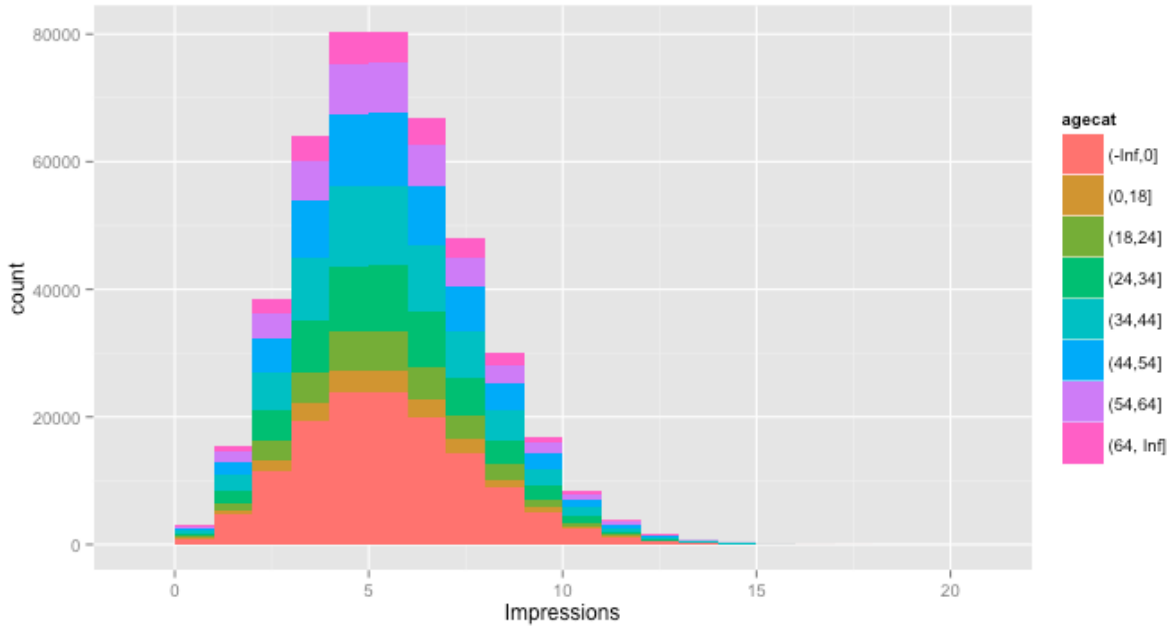
Outputs:

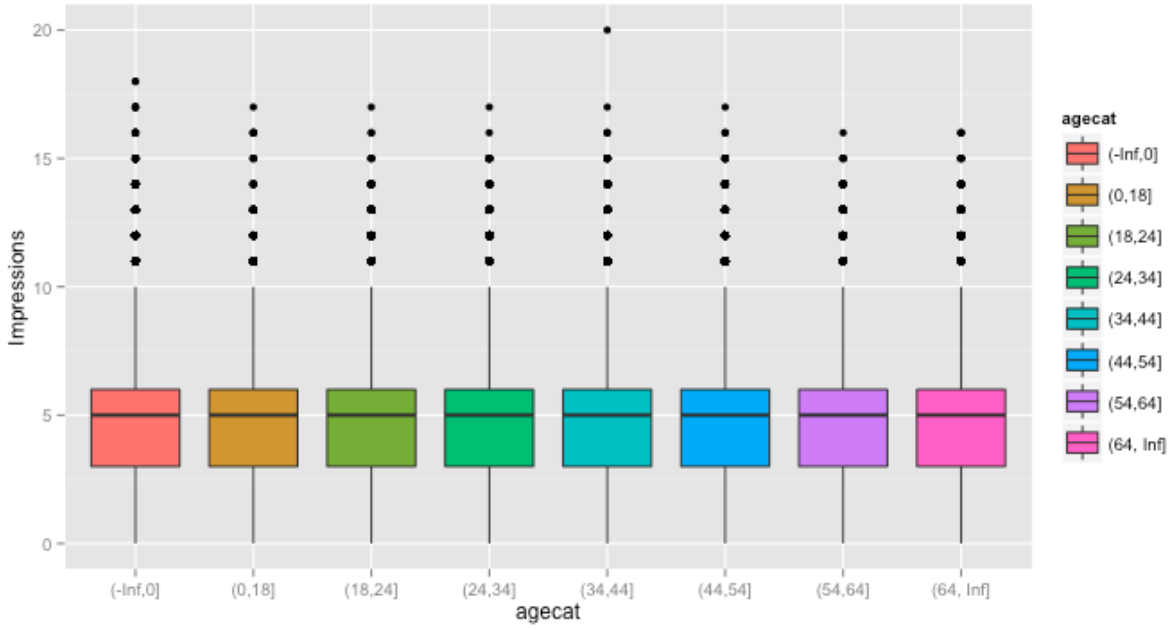
```
Age              Gender       Impressions         Clicks         Signed_In          agecat
 Min.   :  0.00   Min.   :0.000   Min.   : 0.000   Min.   :0.00000   Min.   :0.0000   (-Inf,0]:137106
 1st Qu.:  0.00   1st Qu.:0.000   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.0000   (34,44] : 70860
 Median : 31.00   Median :0.000   Median : 5.000   Median :0.00000   Median :1.0000   (44,54] : 64288
```

```
 Mean   : 29.48   Mean   :0.367   Mean   : 5.007   Mean   :0.09259   Mean   :0.7009   (24,34] :
58174
 3rd Qu.: 48.00   3rd Qu.:1.000   3rd Qu.: 6.000   3rd Qu.:0.00000   3rd Qu.:1.0000   (54,64] :
44738
 Max.   :108.00   Max.   :1.000   Max.   :20.000   Max.   :4.00000   Max.   :1.0000   (18,24] :
35270
                                                                                      (Other) : 48005
```

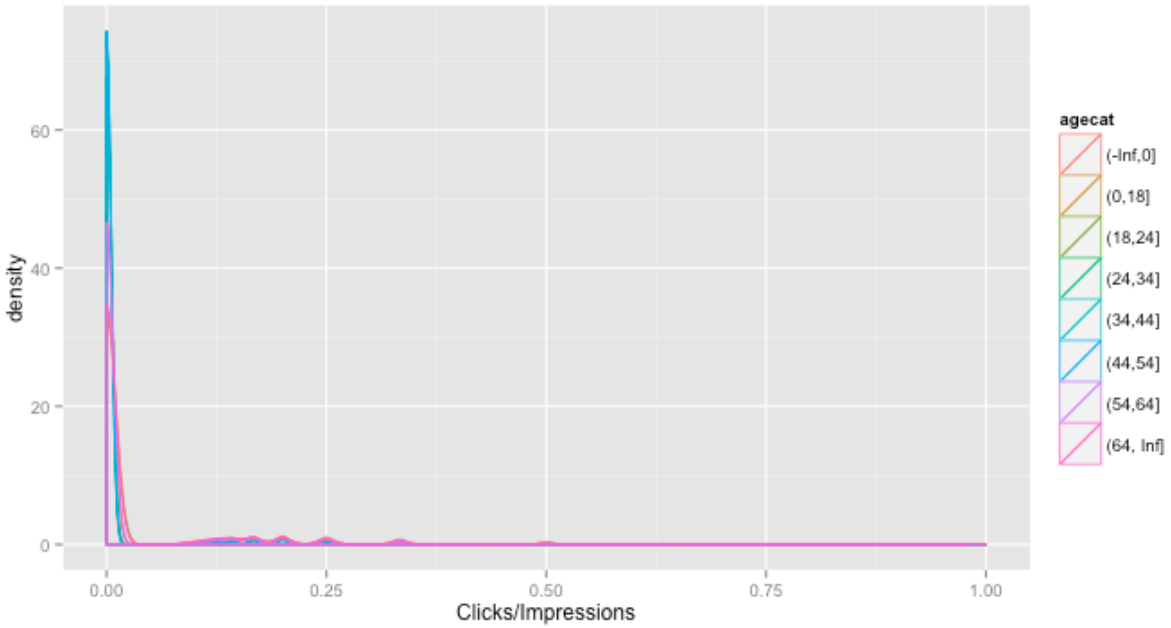
agecat	Age.FUN1	Age.FUN2	Age.FUN3	Age.FUN4
1 (-Inf,0]	137106	0	0.00000	0
2 (0,18]	19252	7	16.03350	18
3 (18,24]	35270	19	21.26904	24
4 (24,34]	58174	25	29.50335	34
5 (34,44]	70860	35	39.49468	44
6 (44,54]	64288	45	49.49258	54
7 (54,64]	44738	55	59.49819	64
8 (64, Inf]	28753	65	72.98870	108

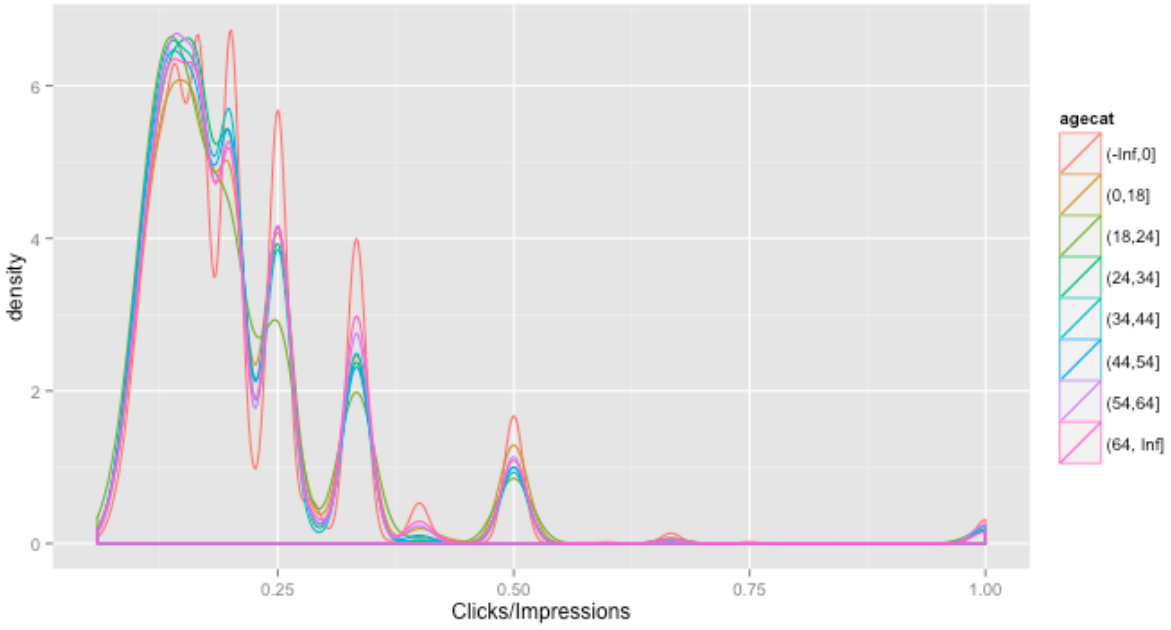
agecat	Gender.mean	Signed_In.mean	Impressions.mean	Clicks.mean
1 (-Inf,0]	0.0000000	0	4.999657	0.14207985
2 (0,18]	0.6421151	1	4.998961	0.13105132
3 (18,24]	0.5338531	1	5.006635	0.04845478
4 (24,34]	0.5321621	1	4.993829	0.05048647
5 (34,44]	0.5316963	1	5.021507	0.05167937
6 (44,54]	0.5289790	1	5.010406	0.05027377
7 (54,64]	0.5361885	1	5.022308	0.10183736
8 (64, Inf]	0.3632664	1	5.012347	0.15128856

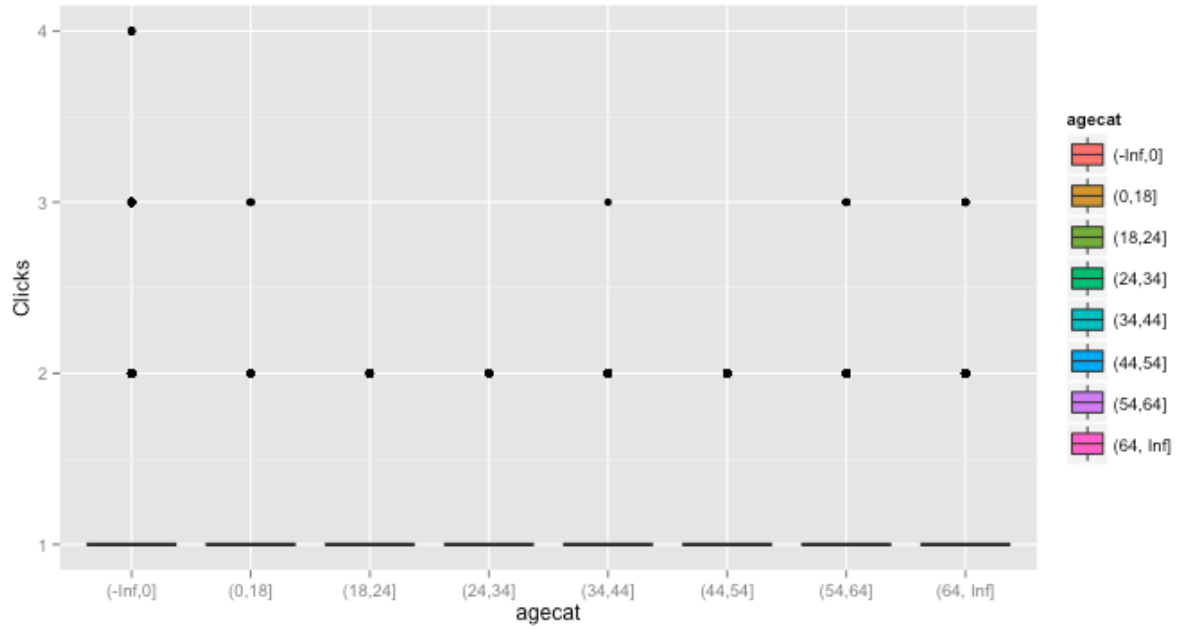


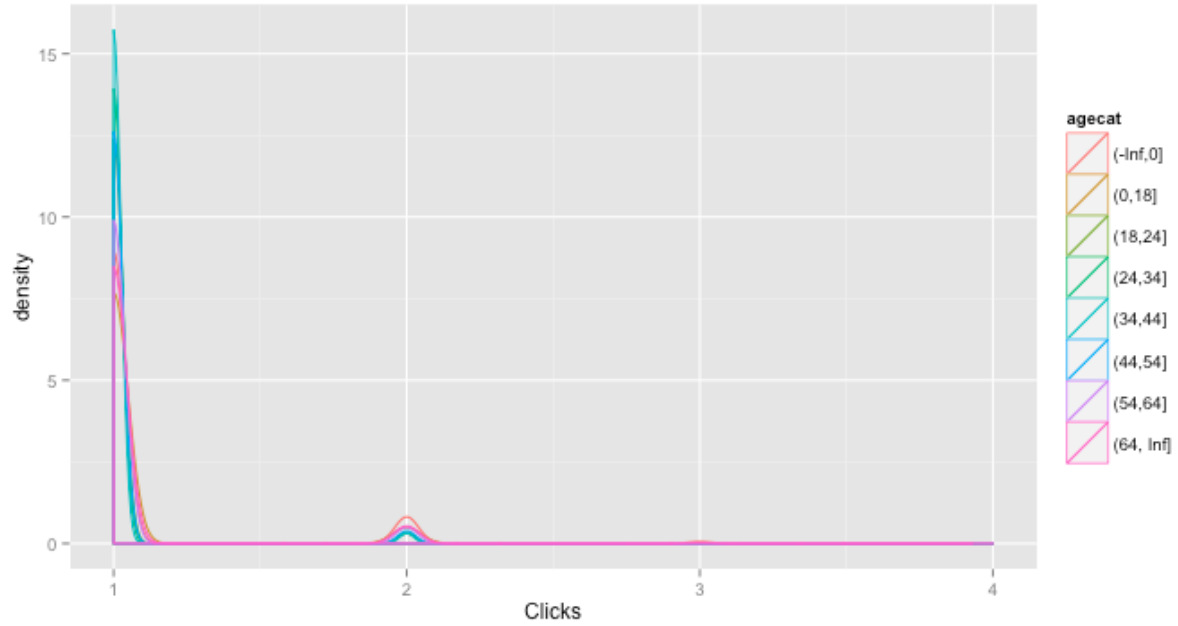


	hasimps	Clicks.FUN1	Clicks.FUN2	Clicks.FUN3	Clicks.FUN4
1	(-Inf,0]	3066	0	0.00000000	0
2	(0, Inf]	455375	0	0.09321768	4


density
60
40
20
0
0.00
0.25
0.50
0.75
1.00
Clicks/Impressions
agecat
(-Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64, Inf]


density
6
4
2
0
0.25
0.50
0.75
1.00
Clicks/Impressions
agecat
(-Inf,0]
(0,18]
(18,24]
(24,34]
(34,44]
(44,54]
(54,64]
(64, Inf]





	Age	Gender	Impressions	Clicks	Signed_In	agecat	hasimps	scode
1	36	0	3	0	1	(34,44]	(0, Inf]	Imps
2	73	1	3	0	1	(64, Inf]	(0, Inf]	Imps
3	30	0	3	0	1	(24,34]	(0, Inf]	Imps
4	49	1	3	0	1	(44,54]	(0, Inf]	Imps
5	47	1	11	0	1	(44,54]	(0, Inf]	Imps
6	47	0	11	1	1	(44,54]	(0, Inf]	Clicks

The first graph that stood out to be was the density vs clicks. If I am reading it correctly it seems as though having a high density of advertisements on a page can really only lead to getting maybe one click. Yet getting two. three or four clicks is a rarity. This graph seems to show that in general for a web page there is a very low click through rate. Looking at the age cat table and impressions also intrigued me. It seems as though the older you are the more likely an ad on a webpage will leave an impression. The first graph also seems to show that ads don't have as much an impression on 18-24 age bracket and the 65+ age bracket compared to the people between 25-64. The fact that 18-24 year olds aren't as susceptible to only ads falls in line with what I would have thought. As for the boxplots I'm not as sure how to interpret the results from these graphs. I also find it interesting that on the clicks/impressions vs density graph it seems as though people in the 0-18 bracket seem to be the most impressionable. While this idea that people under 18 are most likely to click on an ad on the internet makes sense, I'm surprised this holds true for the New York Times website for which this set of data is collected from.