Problem 2 Report

Presented By:

Anuj Rastogi

anujrast@buffalo.edu

Person# 50134324

Graphs and Question Answers on a Single Day

INTRODUCTION:

Each one represents one (simulated) days' worth of ads shown and clicks recorded on the *New York Times* home page in May 2012. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged in.

ANSWERS TO BOOK QUESTIONS:

First let's start from answering questions from the book given in page 38-39.

Question 1:

Create a new variable, age_group, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64", and "65+".

Answer Script 1:

```
#Categorise the Data
head(data1)
data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
```

The above script categorizes data based on the age of people. The summary details below how the summary of all the columns and also, the age Category column.

```
> summary(data1)
      Age              Gender        Impressions         Clicks          Signed_In     
 Min.   :  0.00   Min.   :0.000   Min.   : 0.000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.:  0.00   1st Qu.:0.000   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median : 31.00   Median :0.000   Median : 5.000   Median :0.00000   Median :1.0000  
 Mean   : 29.48   Mean   :0.367   Mean   : 5.007   Mean   :0.09259   Mean   :0.7009  
 3rd Qu.: 48.00   3rd Qu.:1.000   3rd Qu.: 6.000   3rd Qu.:0.00000   3rd Qu.:1.0000  
 Max.   :108.00   Max.   :1.000   Max.   :20.000   Max.   :4.00000   Max.   :1.0000  

     agecat      
 (-Inf,0]:137106  
 (34,44] : 70860  
 (44,54] : 64288  
 (24,34] : 58174  
 (54,64] : 44738  
 (18,24] : 35270  
 (Other) : 48005  
```

Figure 1

From the above one can figure out that how many people of different age categories were there online for which that data was collected.

Question 2:

Plot the distributions of number impressions and click through-rate (CTR=# clicks/#impressions) for these six age categories.

Answer Script 2:

```
# create click thru rate
# we don't care about clicks if there are no impressions
# if there are clicks with no imps my assumptions about
# this data are wrong
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()
```

The above script is for the click through rate vs Impressions and click through rate vs clicks based on the age categories. The plots below show both of these distributions based on the age categories.

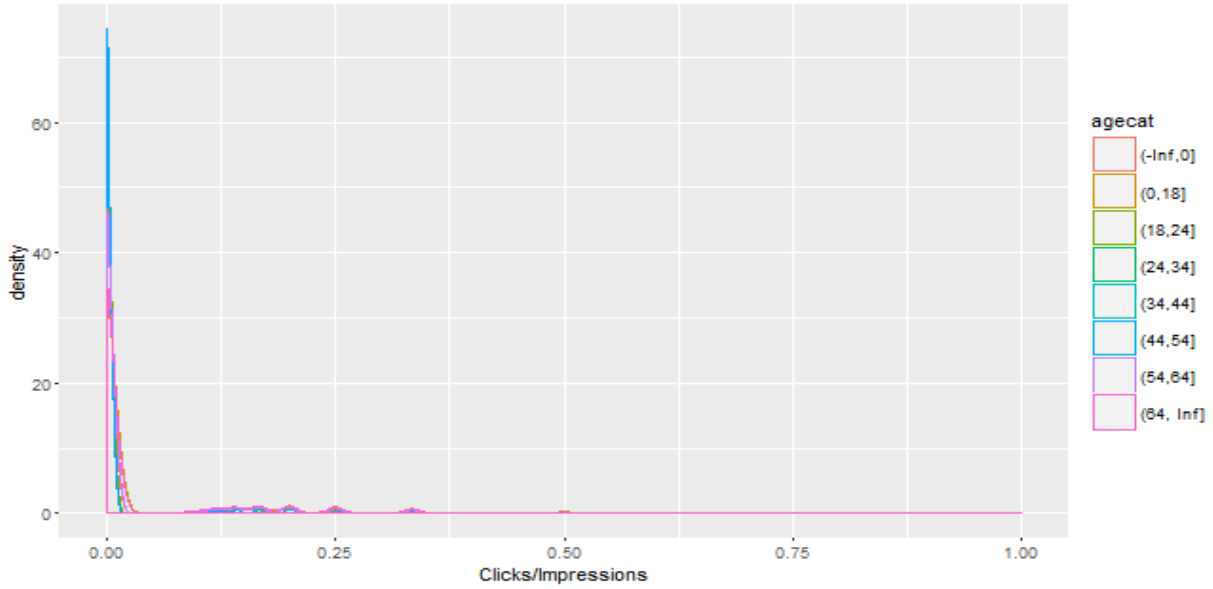

Figure 2

This graph represents the click through rate vs impression density based on different age categories.

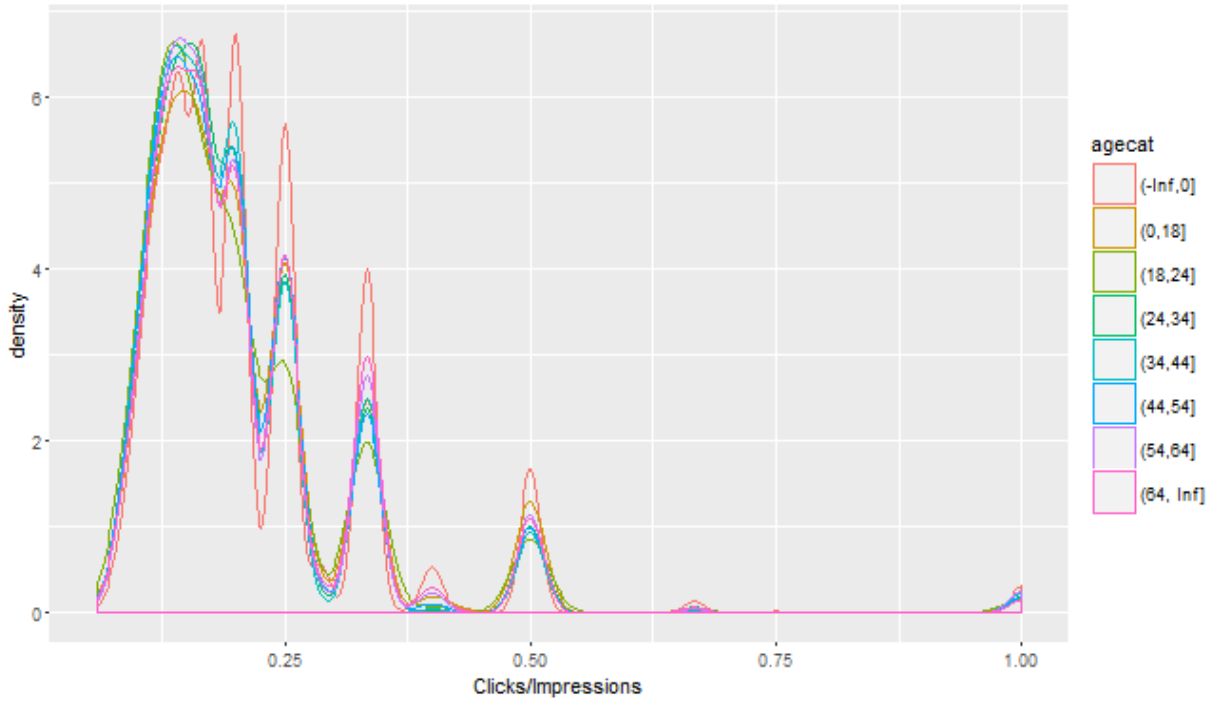

Figure 3

The above graph represents the click through rate vs clicks for different age categories.

Question 3:

Define a new variable to segment or categorize users based on their click behavior.

Answer Script 3:

The below code divides the people based on if they click on a particular advertise or not.

```
# create categories
data1$scode[data1$Impressions==0] <- "NoImps"
data1$scode[data1$Impressions >0] <- "Imps"
data1$scode[data1$Clicks >0] <- "Clicks"

# Convert the column to a factor
data1$scode <- factor(data1$scode)
head(data1)
```

```
> head(data1)
  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
```

Figure 4

Question 4:

Explore the data and make visual and quantitative comparisons across user segments/demographics (<18-year-old males versus < 18-year-old females or logged-in versus not, for example).

Answer Script 4:

The below code categorizes the data based on the clicks of the people and based on different age categories.

```
ggplot(subset(data1, Clicks>0), aes(x=agecat, y=Clicks,fill=agecat)) + geom_boxplot()
ggplot(subset(data1, Clicks>0), aes(x=Clicks, colour=agecat))+ geom_density()
```

The below figure represents the clicks made vs the age category.

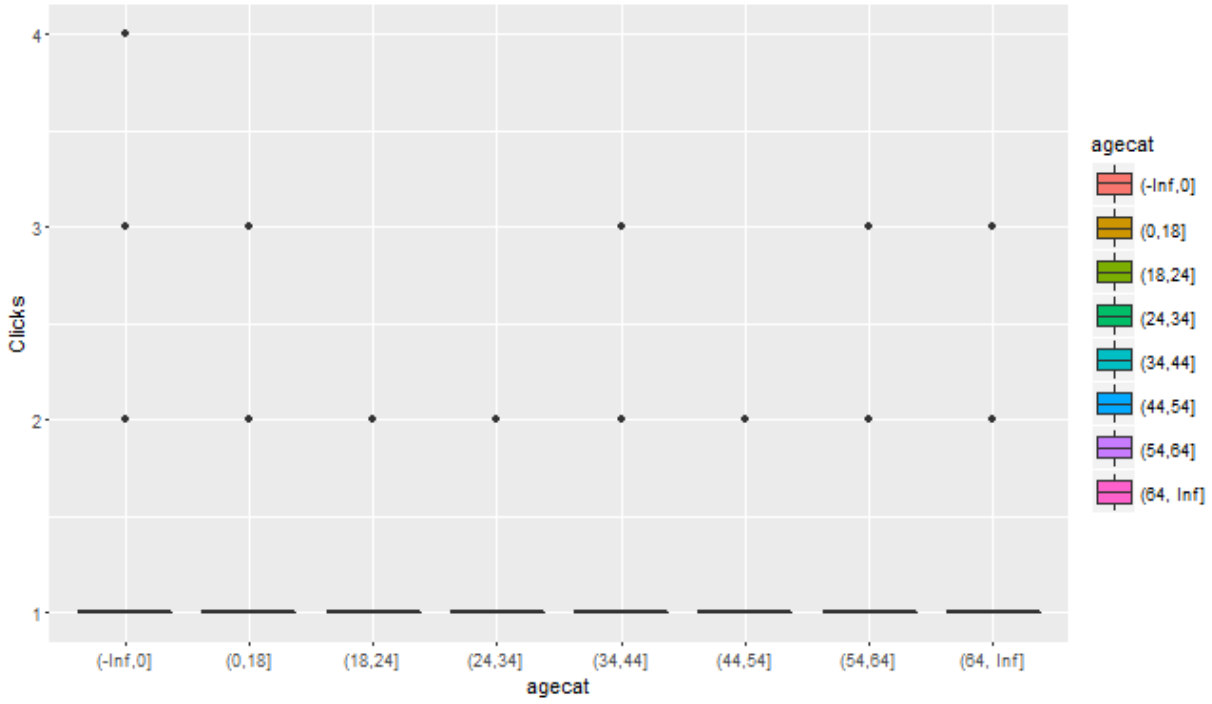

Figure 5

In the figure 6, the density representation of people who click based on their age is given.

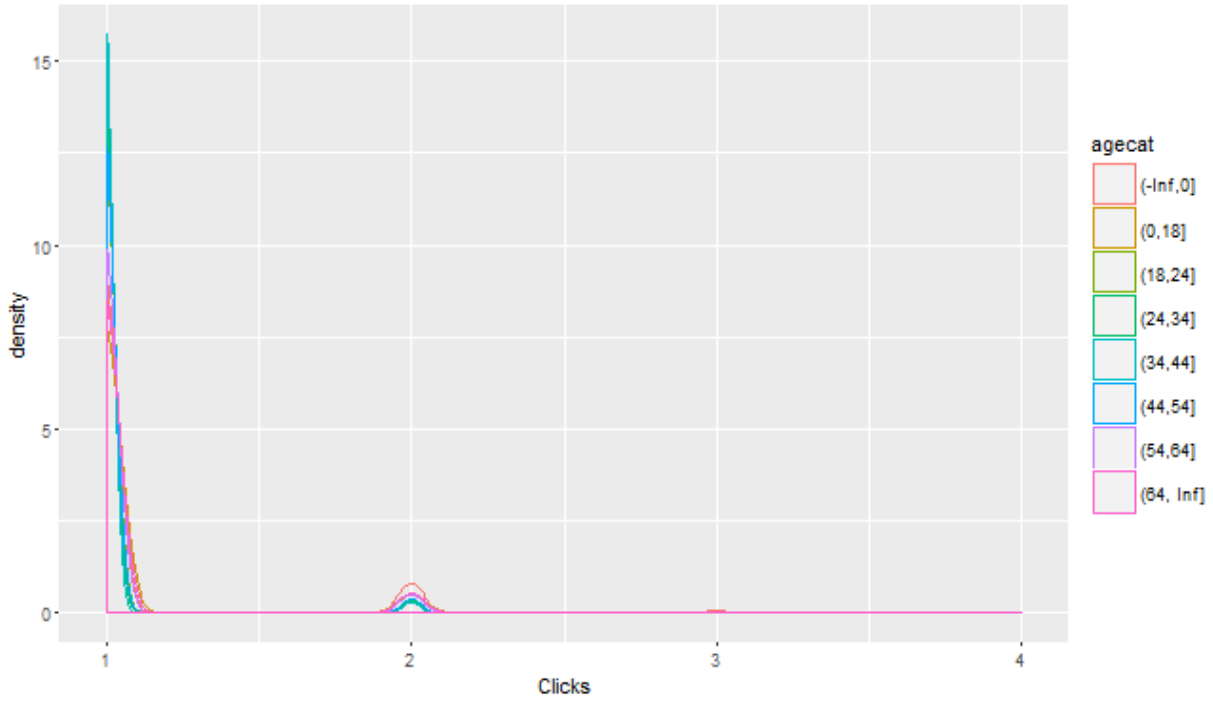

Figure 6

Question 5:

Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median, variance, and max, and these can be calculated across the various user segments. Be selective. Think about what will be important to track over time—what will compress the data, but still capture user behavior.

Answer Script 5:

```
#look at levels
clen <- function(x){c(length(x))}
etable<-summaryBy(Impressions~scode+Gender+agecat,data = data1, FUN=clen)
```

Figure 7

The above script draws the summary based on different levels combined. The figure below represents first few rows of this summary. Here the data has been divided based on the three parameters. These are:

Scode + Gender + agecat

```
> etable
    scode Gender    agecat Impressions.clen
1  clicks      0  (-Inf,0]             17776
2  clicks      0    (0,18]               846
3  clicks      0   (18,24]               779
4  clicks      0   (24,34]              1361
5  clicks      0   (34,44]              1675
6  clicks      0   (44,54]              1494
7  clicks      0   (54,64]              2006
8  clicks      0 (64, Inf]              2598
9  clicks      1    (0,18]              1525
10 clicks      1   (18,24]               890
11 clicks      1   (24,34]              1509
12 clicks      1   (34,44]              1917
13 clicks      1   (44,54]              1645
14 clicks      1   (54,64]              2331
15 clicks      1 (64, Inf]              1486
16   Imps      0  (-Inf,0]            118401
17   Imps      0    (0,18]              6001
18   Imps      0   (18,24]             15538
19   Imps      0   (24,34]             25690
20   Imps      0   (34,44]             31290
21   Imps      0   (44,54]             28563
22   Imps      0   (54,64]             18626
23   Imps      0 (64, Inf]             15585
24   Imps      1    (0,18]             10754
25   Imps      1   (18,24]             17807
26   Imps      1   (24,34]             29241
27   Imps      1   (34,44]             35512
28   Imps      1   (44,54]             32143
29   Imps      1   (54,64]             21499
30   Imps      1 (64, Inf]              8887
31 NoImps      0  (-Inf,0]               929
32 NoImps      0    (0,18]                43
33 NoImps      0   (18,24]               124
34 NoImps      0   (24,34]               165
35 NoImps      0   (34,44]               219
```

Figure 8

ANALYSIS PERFORMED ON THE ABOVE DATA

From Figure1, Figure 3 and Figure 5, we can conclude that out of the real ages the age category with 64-INFINITY has the best click through rate hence, they focus more on advertisements. Though their population is lesser still they make a good ratio by clicking more.

Also, Based on the population distribution, the maximum population is between 34-44 years of age. The clicks recorded for 0-18, 34-44, 54-64, 64-INF are same. But since the maximum population belongs to 34-44 hence according to economics of scale, focusing on this population will yield good probability of clicks. Also, 64-INFINITY is also, a good age category to focus on for better clicks. Therefore, we can focus on 34-44 and from 64-Infinity for advertisements and can target these age groups.

Graphs and Question Answers on a MONTHLY DATA

ANSWERS TO BOOK QUESTIONS:

First let's start from answering questions from the book given in page 38-39.

Question 1:

Create a new variable, age_group, that categorizes users as "<18",
"18-24", "25-34", "35-44", "45-54", "55-64", and "65+".

Answer Script 1:

```
#Distributing the age category
completeTable$AgeCat = cut(completeTable$Age, breaks = c(-Inf,18,24,34,44,54,64,Inf), labels = c('<18','18-24','25-34','35-44','45-54','55-64','65+'))
```

The above script divides the age categories of the people.

Question 2:

Plot the distributions of number impressions and click through-rate (CTR=# clicks/#impressions)
for these six age categories.

Answer Script 2:

The below script calculates the Click Through rate.

```
#Calculating the ClickThoroughRate
completeTable$CTR = (completeTable$Clicks/completeTable$Impressions)
```

```
#Click Through Rate vs Impressions based on AgeCateogry
filepath1 = 'B:/UB_CS/DIC/DICProject/NYT/nyt1.csv'
filepath2 = 'B:/UB_CS/DIC/DICProject/NYT/nyt10.csv'
filepath3 = 'B:/UB_CS/DIC/DICProject/NYT/nyt16.csv'
filepath4 = 'B:/UB_CS/DIC/DICProject/NYT/nyt22.csv'
filepath5 = 'B:/UB_CS/DIC/DICProject/NYT/nyt29.csv'

sampleTable = read.table(filepath1, header = T, sep = ',')
tempTable = read.table(filepath2, header = T, sep = ',')
sampleTable = rbind(sampleTable, tempTable)
tempTable = read.table(filepath3, header = T, sep = ',')
sampleTable = rbind(sampleTable, tempTable)
tempTable = read.table(filepath4, header = T, sep = ',')
sampleTable = rbind(sampleTable, tempTable)
tempTable = read.table(filepath5, header = T, sep = ',')
sampleTable = rbind(sampleTable, tempTable)

head(sampleTable)
sampleTable$agecat1 <-cut(sampleTable$Age,c(-Inf,0,18,24,34,44,54,64,Inf))
#install.packages('doBy')
#install.packages('ggplot2')
library(ggplot2)
library('doBy')
sampleTable$hasimps <-cut(sampleTable$Impressions,c(-Inf,0,Inf))
ggplot(subset(sampleTable, sampleTable$Impressions>0), aes(x=Clicks/Impressions, colour=agecat1)) + geom_density()
ggplot(subset(sampleTable, sampleTable$Clicks>0), aes(x=agecat1, y=Clicks,fill=agecat1)) + geom_boxplot()
ggplot(subset(sampleTable, sampleTable$Clicks>0), aes(x=Clicks/Impressions,colour=agecat1)) + geom_density()
```

The script calculates the CTR and adds a column to the existing table.
The plots plotted are for click through rate based on age category, click through rate based on age category for clicks more than 0 and age category vs clicks made.

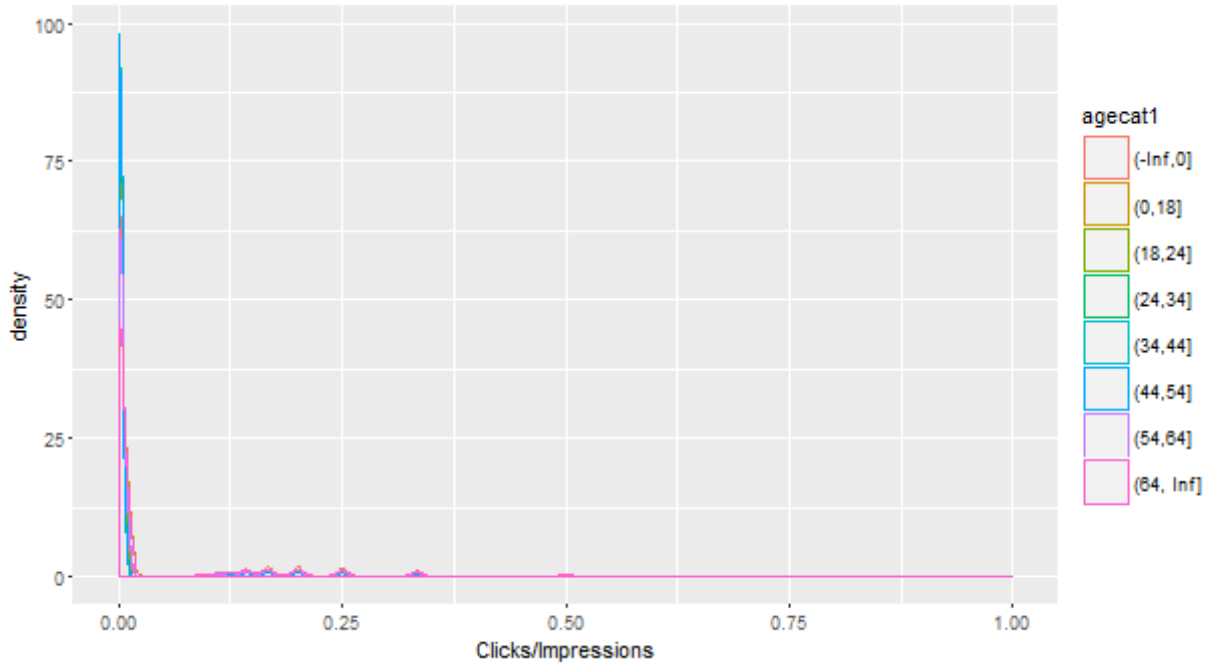

Figure 9- CTR vs impressions based on age category

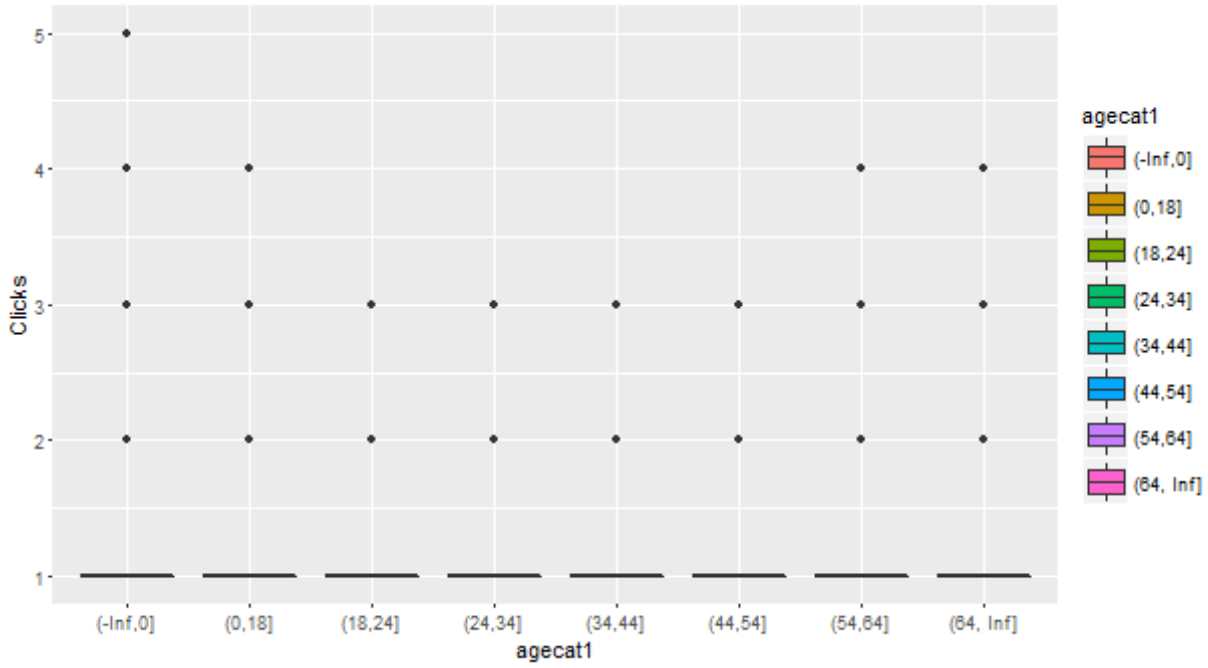

Figure 10 CTR vs Clicks based on age category

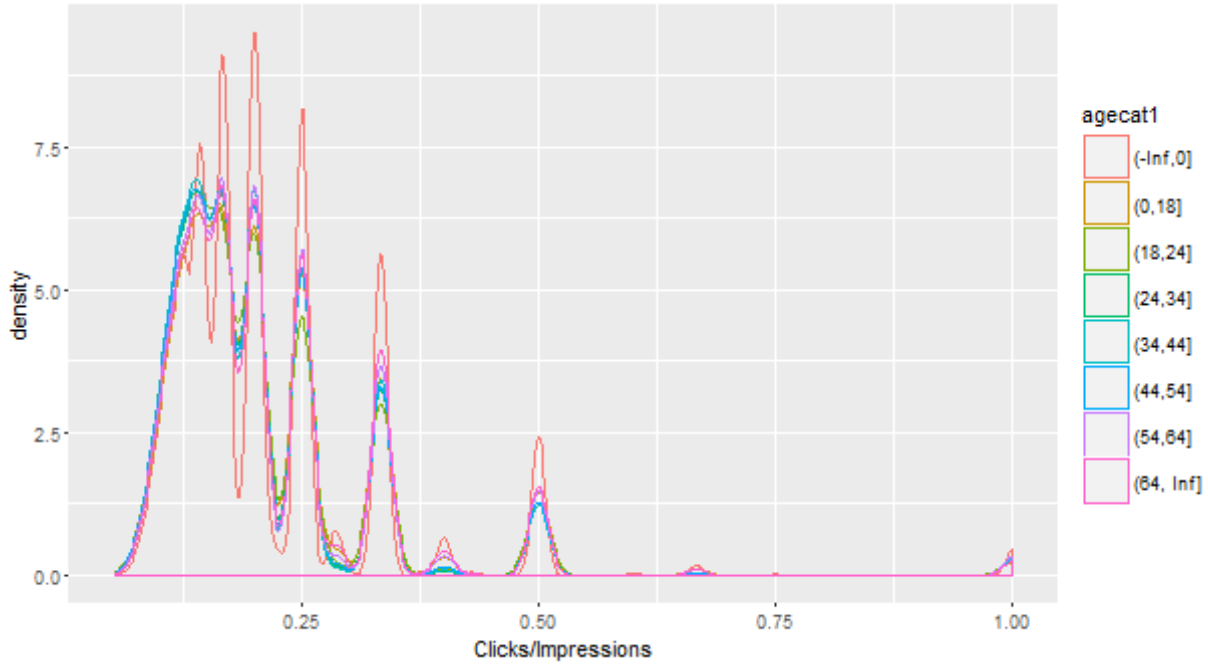

Figure 11- CTR vs age category

The CTR was calculated over a sample data for five days of month

Question 3:

Define a new variable to segment or categorize users based on their click behavior.

Answer Script 3:

```
#Categorise Users Based on Thier Click Behavior
summary(completeTable$Clicks)
completeTable$UserClickBehav = cut(completeTable$Clicks, breaks = c(-Inf,0,Inf), labels = c('NoClicks','Clicks'))
levels(completeTable$UserClickBehav)
table2 <- table(completeTable$UserClickBehav[completeTable$Signed_In == 1],completeTable$Gender[completeTable$Signed_In == 1])
barplot(table2, beside = T, main = 'signed in and Click', las = 1 , xlab = 'Gender', names.arg = c('F','M'),legend.text = c('NoClicks','Clicks'))
box()

#Plotting Impressions based on User Click Behavior and Gender
boxplot(completeTable$Impressions~completeTable$UserClickBehav*completeTable$Gender,las = 2, main = 'Impressions~ [ClickBehavior And Gender]')
```

The above script categories the users based on their click behavior. Below are the two plots, one which plots genders against click behaviors for those who are signed in. The other plots impressions based on click behavior and gender.

Figure 9

The above figure represents users who are signed in and from those who have clicked. It also, categorizes them in two categories based on their gender. The plot below represents females and males who are signed in and have clicked once but have also made impressions.

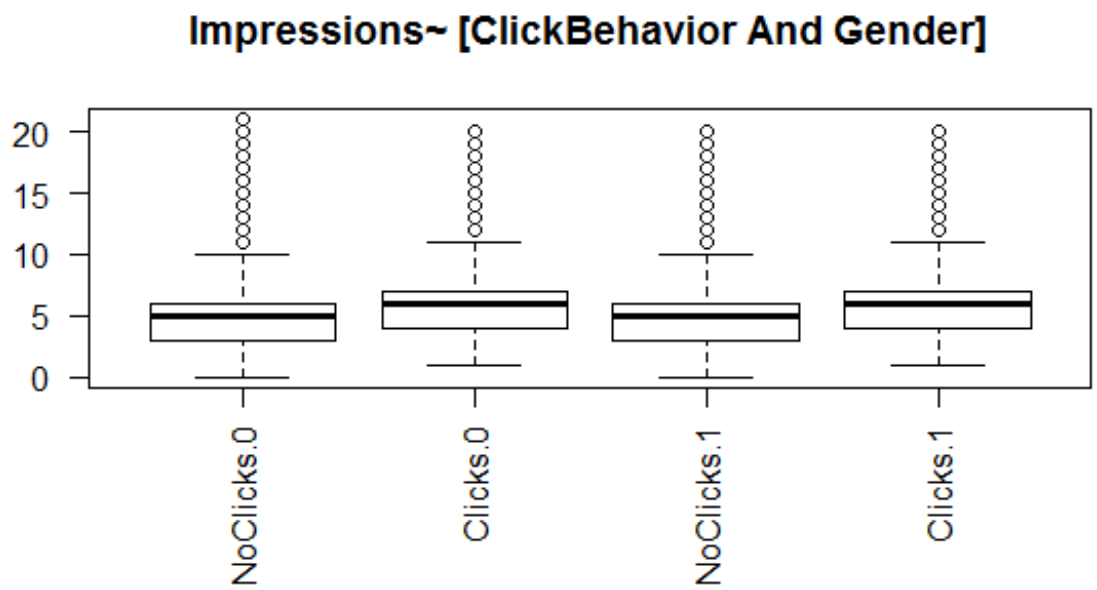

Figure 10

0 represents females
1 represents male

From the above plot it can be identified that the users have been classified based on gender and click behavior.

Question 4:

Explore the data and make visual and quantitative comparisons across user segments/demographics (<18-year-old males versus < 18-year-old females or logged-in versus not, for example).

Answer Script 4:

The below script answers the following questions.

```
#Graph and statistics to show how many females and males <18 are logged IN
completeTable$Signed_In = as.factor(completeTable$Signed_In)
table1 = table(completeTable$Signed_In[completeTable$AgeCat == '<18'],completeTable$Gender[completeTable$AgeCat == '<18'])
table1
barplot(table1, beside = T, legend.text = c('Not-SighnedIn', 'SighnedIn'), xlab = 'Gender' , axes = F, las = 1, names.arg = c('F','M'), main = '<18 Fem
box()
```

The below is the plot for <18 year old males and females who are signed in.

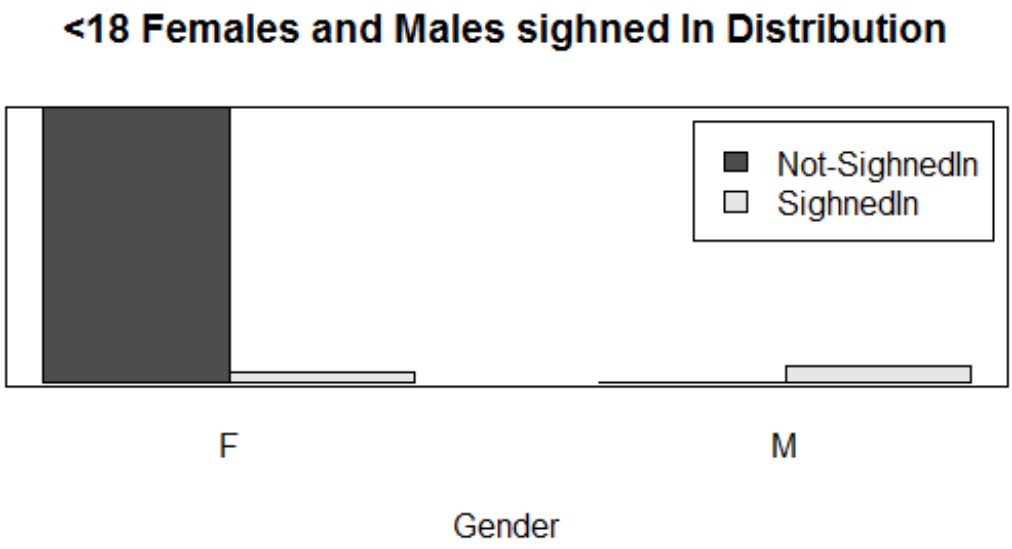

Figure 11

Question 5:

Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median, variance, and max, and these can be calculated across the various user segments. Be selective. Think about what will be important to track over time—what will compress the data, but still capture user behavior.

Answer Script 5:

```
#summary of monthly data
summary(completeTable)
```

The above script gives the summary of the overall data. Along with this summary there are other interesting tables based on gender distribution that can be shown.

```
> summary(completeTable)
      Age          Gender        Impressions      Clicks         Signed_In
 Min.   :  0.00   0:10090192   Min.   : 0   Min.   :0.00000   0:5613610
 1st Qu.:  0.00   1: 4815673   1st Qu.: 3   1st Qu.:0.00000   1:9292255
 Median : 26.00                Median : 5   Median :0.00000
 Mean   : 26.24                Mean   : 5   Mean   :0.09773
 3rd Qu.: 46.00                3rd Qu.: 6   3rd Qu.:0.00000
 Max.   :115.00                Max.   :21   Max.   :6.00000

   AgeCat          UserClickBehav          CTR
 <18  :6170598   NoClicks:13544294   Min.   :0.00
 18-24:1022112   Clicks   : 1361571   1st Qu.:0.00
 25-34:1673650                       Median :0.00
 35-44:2044613                       Mean   :0.02
 45-54:1859487                       3rd Qu.:0.00
 55-64:1299303                       Max.   :1.00
 65+  : 836102                       NA's   :1e+05
```

Figure 12

ANALYSIS PERFORMED ON THE ABOVE DATA

The maximum population belongs to the people with <18 years of age on internet, throughout the month. Also, majority of the people are females. Hence, if we focus on advertisement concerning on females then through quantity we can have a good advertisement click rate. Also, through analysis of the above graph it is evident that in <18 years of age majority of the females are signed in and not males. Hence, on products concerning people less than 18 years age if the advertisement targets female there are better chances of success.

But Figure 9 also shows that out of all those who are signed in male make more clicks than females and hence, advertisements on males should also not be undermined.

COMAPRISION OF THE RESULTS FROM A SINGLE DAY VS MONTH

From the analysis of a single day we assumed that in majority of the people on internet belong to 34-44 years of age. But after monthly analysis, from Figure 12 AgeCat, we see that most of the people are young people with <18 years of age. So redefining our analysis we can say that, by economics of scale we will be able to make a better profit if we focus on this gentry. Also, from the monthly data analysis we see that although the majority of the users are females then too the major number of clicks are recorded by males. Hence, focusing in totality on males can help in longer run. This is evident from figure 9. Also, since <18 years of age people are more, Figure 12 AgeCat, hence, focusing on them tells us that in this age group it's better to focus on females. This can be deciphered from figure 11.

Also, from CTR analysis we see that for <18 years people the clicks increase from 2 to 3 in monthly data compared to single day data. Hence, this is an incremental trend.