IFasere_LiveSession10a Assignment

Ireti Fasere

April 12, 2017

Introduction

Download click through rate data set from the given site and perform necessary analysis to understand and visualize the clicks recorded on the New York Times.

Setup and Importing the Raw Data.

Set my work directory to my dataset file location.

```
setwd("D:/DODATAHW/")
getwd()
```

```
## [1] "D:/DODATAHW"
```

```
##Download the Url to my Directory
site <- "http://stat.columbia.edu/~rachel/datasets/nyt1.csv"
download.file(site,destfile="./nyt1.csv")
##Importing the nyt1 data file.
nyt1=df<-read.csv("nyt1.csv",header = TRUE, sep=",",na.strings=c("", "NA"))
head(nyt1) #Checking the data.
```

```
##   Age Gender Impressions Clicks Signed_In
## 1  36      0           3      0         1
## 2  73      1           3      0         1
## 3  30      0           3      0         1
## 4  49      1           3      0         1
## 5  47      1          11      0         1
## 6  47      0          11      1         1
```

```
str(nyt1)#Checking the data.
```

```
## 'data.frame':    458441 obs. of  5 variables:
##  $ Age        : int  36 73 30 49 47 47 0 46 16 52 ...
##  $ Gender     : int  0 1 0 1 1 0 0 0 0 0 ...
##  $ Impressions: int  3 3 3 3 11 11 7 5 3 4 ...
```

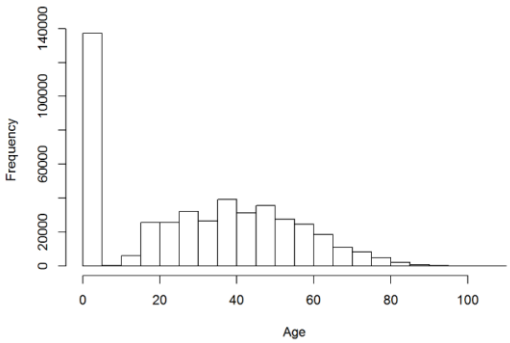
```
##  $ Clicks     : int  0 0 0 0 0 1 1 0 0 0 ...
##  $ Signed_In  : int  1 1 1 1 1 1 0 1 1 1 ...
```

```
summary(nyt1)#Checking the data.
```

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

Checking the Data

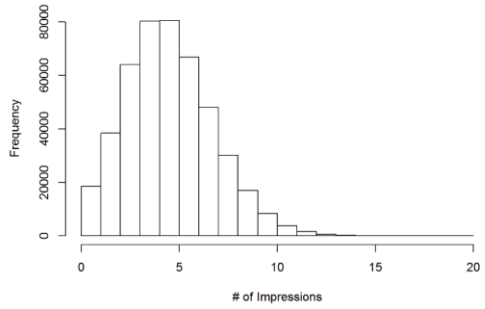
```
# Visualization of the Age column
hist(nyt1$Age, main="", xlab="Age")
```




```
range(nyt1$Age)
```

```
## [1]   0 108
```

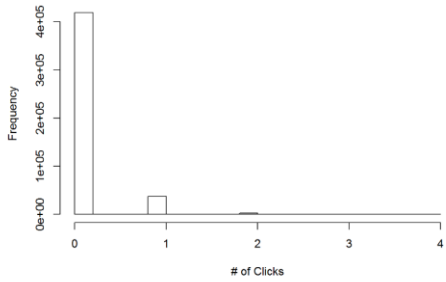
```
# Visualization of the Impressions column
hist(nyt1$Impressions, main="", xlab="# of Impressions")
```




```
range(nyt1$Impressions)
```

```
## [1]  0 20
```

```
# Visualization of the Clicks column
hist(nyt1$Clicks, main="", xlab="# of Clicks")
```




```
range(nyt1$Clicks)
```

```
## [1] 0 4
```

Performing the Analysis

```
1.#Create a new variable called AgeGroup
```

```
## [1] 1
```

```
nyt1$AgeGroup <- cut(nyt1$Age, c(-Inf, 18, 24, 34, 44, 54, 64, Inf))
levels(nyt1$AgeGroup) <- c("<18", "18-24", "25-34", "35-44", "45-54", "55-64"
, "65+")


# Take a look at the Data set after changes.
head(nyt1)
```

```
##   Age Gender Impressions Clicks Signed_In AgeGroup
## 1  36      0           3      0         1    35-44
```

```
## 2  73      1           3      0         1      65+
## 3  30      0           3      0         1    25-34
## 4  49      1           3      0         1    45-54
## 5  47      1          11      0         1    45-54
## 6  47      0          11      1         1    45-54
```

```
2.# Create subset  "ImpSub" where Impressions > 0 ) in your data set.
```

```
## [1] 2
```

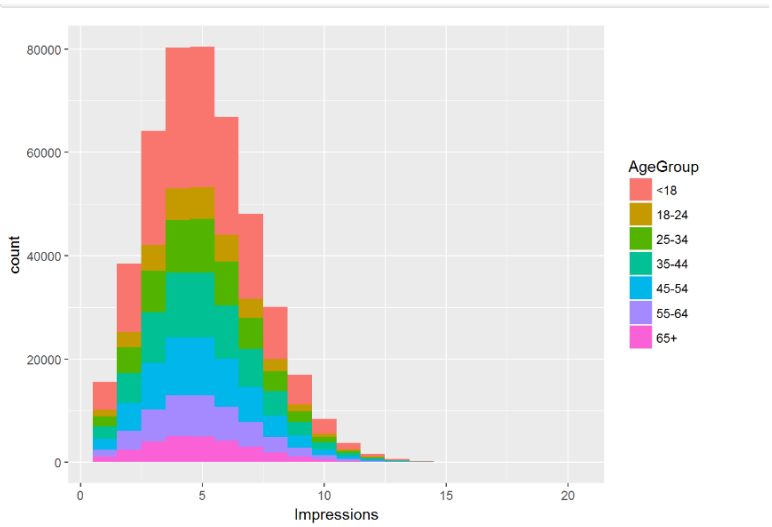
```
ImpSub <- subset(nyt1, Impressions>0)
# Create variable cick-through-rate (CTR)and add to ImpSub
ImpSub$CTR <- ImpSub$Clicks/ImpSub$Impressions


head(ImpSub)
```

```
##   Age Gender Impressions Clicks Signed_In AgeGroup        CTR
## 1  36      0           3      0         1    35-44 0.00000000
## 2  73      1           3      0         1      65+ 0.00000000
## 3  30      0           3      0         1    25-34 0.00000000
## 4  49      1           3      0         1    45-54 0.00000000
## 5  47      1          11      0         1    45-54 0.00000000
## 6  47      0          11      1         1    45-54 0.09090909
```

```
3.# Plot the distribution of number of Impressions and CTR by AgeGroup.
```

```
## [1] 3
```

```
library(ggplot2) # used for visualizations
ggplot(subset(ImpSub, Impressions>0), aes(x=Impressions, fill=AgeGroup))+
geom_histogram(binwidth=1)
```




```
4.# Define new variable to segment users based on CTR
```

```
## [1] 4
```

```
ImpSub$CTRGroup <-cut(ImpSub$CTR,c(-Inf,0.2,0.4,0.6,0.8,Inf))
```

```
head(ImpSub)
```

```
##   Age Gender Impressions Clicks Signed_In AgeGroup        CTR   CTRGroup
## 1  36      0           3      0         1    35-44 0.00000000 (-Inf,0.2]
## 2  73      1           3      0         1      65+ 0.00000000 (-Inf,0.2]
## 3  30      0           3      0         1    25-34 0.00000000 (-Inf,0.2]
## 4  49      1           3      0         1    45-54 0.00000000 (-Inf,0.2]
## 5  47      1          11      0         1    45-54 0.00000000 (-Inf,0.2]
## 6  47      0          11      1         1    45-54 0.09090909 (-Inf,0.2]
```

```
levels(ImpSub$CTRGroup) <- c("CTR<0.2","0.2<=CTR<0.4","0.4<=CTR<0.6","0.6<=CTR<0.8","CTR>0.8")
```

```
levels(ImpSub$CTRGroup)
```

```
## [1] "CTR<0.2"       "0.2<=CTR<0.4" "0.4<=CTR<0.6" "0.6<=CTR<0.8"
## [5] "CTR>0.8"
```

```
5.# Get the total number of Male, Impressions, Clicks and Signed_In
```

```
## [1] 5
```

```
ImpSub_male <- dim(ImpSub[ImpSub$Gender==1,])[1]
```

```
ImpSub_male
```

```
## [1] 167146
```

```
ImpSub_imp <- dim(ImpSub[ImpSub$Impressions>0,])[1]
```

```
ImpSub_imp
```

```
## [1] 455375
```

```
ImpSub_clicks <- dim(ImpSub[ImpSub$Clicks>0,])[1]
ImpSub_clicks
```

```
## [1] 39838
```

```
ImpSub_signedIn <- dim(ImpSub[ImpSub$Signed_In>0,])[1]
ImpSub_signedIn
```

```
## [1] 319198
```

```
cat("the total number of Male = ",ImpSub_male,",Impressions =", ImpSub_imp,",
Clicks=",ImpSub_clicks,",Signed_In = ",ImpSub_signedIn)
```

```
## the total number of Male =  167146 ,Impressions = 455375 ,Clicks= 39838 ,S
igned_In =  319198
```

```
6.#Get the mean of Age, Impressions, Clicks, CTR and percentage of males and
signed_In
```

```
## [1] 6
```

```
sub_means <- sapply(list(nyt1$Age,nyt1$Impressions,nyt1$Clicks,nyt1$CTR,nyt1$
Signed_In), FUN = mean)
```

```
## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA
```

```
names(sub_means) <- c("Avg.Age", "Avg.Imperssions", "Avg.Clicks","Avg.CTR","A
vg.Signed_in")
print(sub_means)
```

```
##         Avg.Age Avg.Imperssions      Avg.Clicks         Avg.CTR
##     29.48255064      5.00731610      0.09259425              NA
##   Avg.Signed_in
##      0.70092989
```

```
## A subdata created as ImpSub_means
ImpSub_means <- sapply(list(ImpSub$Age,ImpSub$Impressions,ImpSub$Clicks,ImpSu
b[(ImpSub$CTR > 0),],ImpSub$Signed_In), FUN = mean)
```

```
## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA
```

```
names(ImpSub_means) <- c("Avg.Age", "Avg.Imperssions", "Avg.Clicks","Avg.CTR"
,"Avg.Signed_in")
print(ImpSub_means)
```

```
##         Avg.Age Avg.Imperssions      Avg.Clicks         Avg.CTR
##     29.48400988      5.04102992      0.09321768              NA
##   Avg.Signed_in
##      0.70095635
```

```
7.#Get the means of Impressions, Clicks, CTR and percentage of males and sign
ed_In by AgeGroup.
```

```
## [1] 7
```

```
aggregate(nyt1[c(1:5)],
by = list(nyt1$AgeGroup),
FUN = mean)
```

```
##   Group.1       Age     Gender Impressions     Clicks Signed_In
## 1     <18  1.974168 0.07906215    4.999571 0.14072193 0.1231277
## 2   18-24 21.269039 0.53385313    5.006635 0.04845478 1.0000000
## 3   25-34 29.503352 0.53216213    4.993829 0.05048647 1.0000000
## 4   35-44 39.494680 0.53169630    5.021507 0.05167937 1.0000000
## 5   45-54 49.492580 0.52897897    5.010406 0.05027377 1.0000000
## 6   55-64 59.498189 0.53618848    5.022308 0.10183736 1.0000000
## 7     65+ 72.988697 0.36326644    5.012347 0.15128856 1.0000000
```

```
8.#CTRGroup vs. AgeGroup counts
```

```
## [1] 8
```

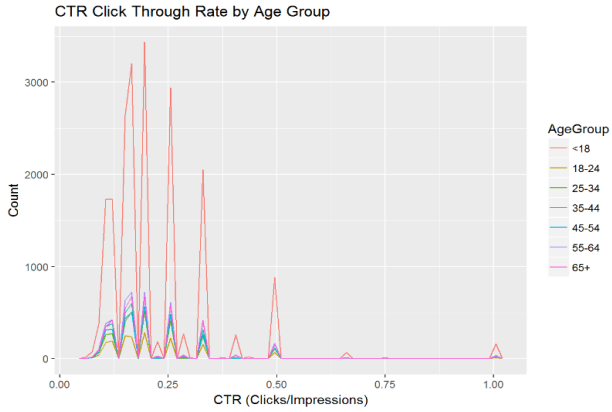
```
CTRAgeGroup <- table(ImpSub$CTRGroup,ImpSub$AgeGroup)
head(CTRAgeGroup)
```

```
##
##                    <18  18-24  25-34  35-44  45-54  55-64    65+
##   CTR<0.2       148412  34540  56980  69424  62936  43147  27261
##   0.2<=CTR<0.4    5735    391    689    820    776   1104   1108
##   0.4<=CTR<0.6     918     68    106    118    113    168    156
##   0.6<=CTR<0.8      76      2      7      4      0      7     10
##   CTR>0.8          162     13     19     28     20     36     21
```

```
9.
```

```
## [1] 9
```

```
#Plot distributions of number impressions and click-through-rate for the age
groups


ggplot(subset(ImpSub,CTR>0), aes(x=CTR, colour = AgeGroup)) +
  geom_freqpoly(binwidth = 0.015) +
  ggtitle("CTR Click Through Rate by Age Group") + xlab("CTR (Clicks/Impressi
ons)") + ylab("Count")
```



10.

```
## [1] 10
```

```
#One more plot you think which is important to look at.
hist(CTRAgeGroup, main="", xlab=" CTRGroup")
```