Homework9

Jethin Abraham

March 23, 2017

**MSDS 6306: Introduction to Data Science**

# Live session unit 09 assignment

**Due: No late acceptance; 1 week from the live secession**

**Monday class (Section: 404): 6:30pm, March 27 (Monday)**

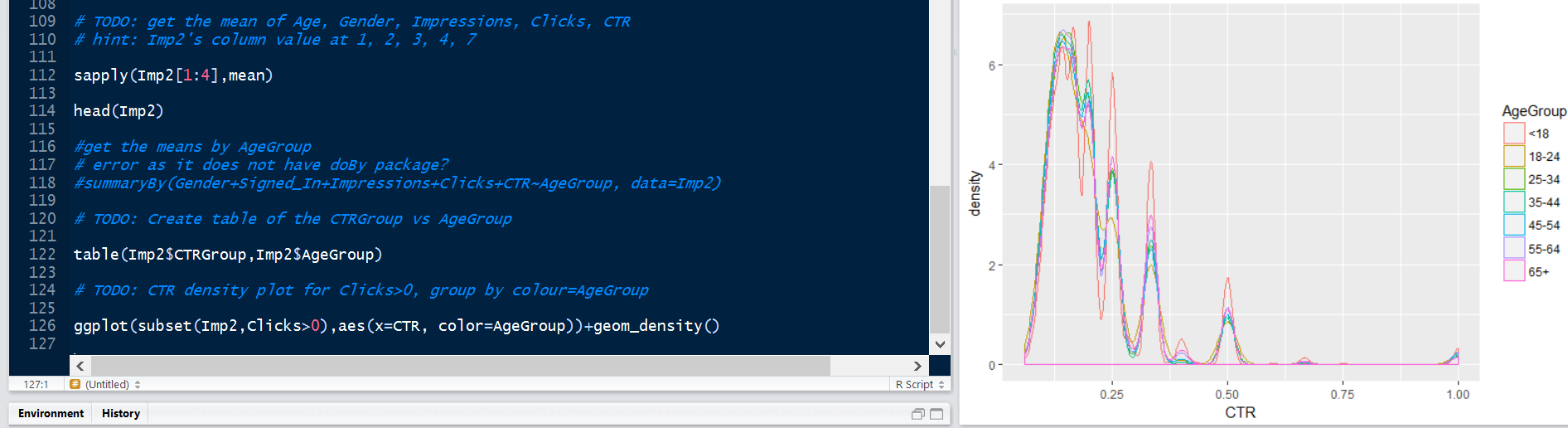
**Wednesday class (Section: 405): 6:30pm, March 29 (Wednesday)**

Calculation Questions (using R) - Use R to calculate the following:

**NOTE:** Submit a word file with the R codes and its screenshot that shows the result of the ggplot chart of the last TODO result that is CTR density plot for Clicks>0, group by colour=AgeGroup.

1. (100%) “hw9\_nyt\_ctr.R” is the R code for keeping track of code used to explore the NYT clicks data from May 1, 2012. You have to complete “TODO” blocks to analyze the click through rate of New York Times web site.

**NOTE:** You can open the R file to complete TODO blocks and when you need hint, you can refer to this document <https://rpubs.com/tmcfl/simulated-click-analysis> .



## R Code



## R Markdown

#=====================================================================================  
# NYTimes Click Through Rate  
# Author: Monnie McGee September 9, 2015  
# Updated: Jongwook Woo, March 17 2017  
# Live Session 10  
# Reference: https://rpubs.com/tmcfl/simulated-click-analysis  
#=====================================================================================  
  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# You don't need to install doBy as it does not work for R3.1.2  
# install.packages ('doBy')  
# library(doBy)  
  
# TODO: set the data pth to fileLocation "http://stat.columbia.edu/~rachel/datasets/nyt1.csv"  
# hint: [fill in] with the file location  
fileLocation <- 'http://stat.columbia.edu/~rachel/datasets/nyt1.csv'  
  
  
nyt\_my <- read.csv(url(fileLocation))  
  
head(nyt\_my)

## 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

# TODO: # What are the variable names? - use names() function  
names(nyt\_my)

## [1] "Age" "Gender" "Impressions" "Clicks" "Signed\_In"

## value of data.frame  
  
str(nyt\_my)

## '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 ...

head(nyt\_my)

## 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

# TODO: Create a new variable ageGroup that categorizes age into following groups:   
# < 18, 18â€“24, 25â€“34, 35â€“44, 45â€“54, 55â€“64 and 65+.  
# hint: [fill in] with c(-Inf, 18, 24, 34, 44, 54, 64, Inf)  
  
nyt\_my$AgeGroup <- cut(nyt\_my$Age, breaks=c(-Inf,18,24,34,44,54,64,Inf))  
  
levels(nyt\_my$AgeGroup) <- c("<18", "18-24", "25-34", "35-44", "45-54", "55-64", "65+")  
  
head(nyt\_my)

## 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

dim(nyt\_my)

## [1] 458441 6

str(nyt\_my)

## 'data.frame': 458441 obs. of 6 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 ...  
## $ AgeGroup : Factor w/ 7 levels "<18","18-24",..: 4 7 3 5 5 5 1 5 1 5 ...

Imp1<-subset(nyt\_my, Impressions == 0)  
  
dim(Imp1)

## [1] 3066 6

# TODO: Use subset of data where Impressions > 0 in your data set.  
# you need to replace [fill in] with the condition  
# hint: similar to the above condition  
  
Imp2<-subset(nyt\_my, Impressions >0)  
  
dim(Imp2)

## [1] 455375 6

# TODO: Create a new variable called click-through-rate (CTR = click/impression).  
# you need to replace [fill in] with impression  
  
Imp2$CTR <- Imp2$Clicks/Imp2$Impressions  
  
  
head(Imp2$CTR)

## [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.09090909

Imp2$CTRGroup <- cut(Imp2$CTR, c(-Inf, 0.2, 0.4, 0.6, 0.8, Inf))  
  
#TODO: Define a new variable to segment users based on click -through-rate (CTR) behavior.  
# 0<=CTR< 0.2, 0.2<=CTR <0.4, 0.4<= CTR<0.6, 0.6<=CTR<0.8, CTR>0.8  
# hint: replace [fill in] with c(variable above)  
  
levels(Imp2$CTRGroup) <- c('0<=CTR< 0.2', '0.2<=CTR <0.4', '0.4<= CTR<0.6', '0.6<=CTR<0.8', 'CTR>0.8')  
  
head(Imp2)

## Age Gender Impressions Clicks Signed\_In AgeGroup CTR CTRGroup  
## 1 36 0 3 0 1 35-44 0.00000000 0<=CTR< 0.2  
## 2 73 1 3 0 1 65+ 0.00000000 0<=CTR< 0.2  
## 3 30 0 3 0 1 25-34 0.00000000 0<=CTR< 0.2  
## 4 49 1 3 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 5 47 1 11 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 6 47 0 11 1 1 45-54 0.09090909 0<=CTR< 0.2

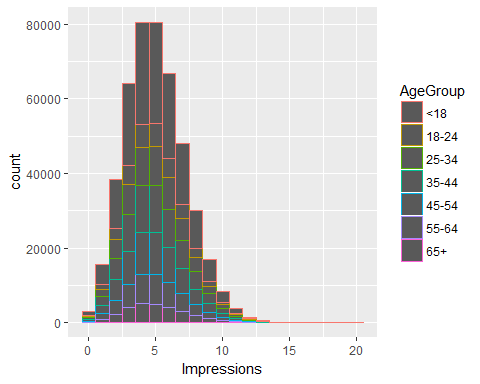
head(nyt\_my)

## 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

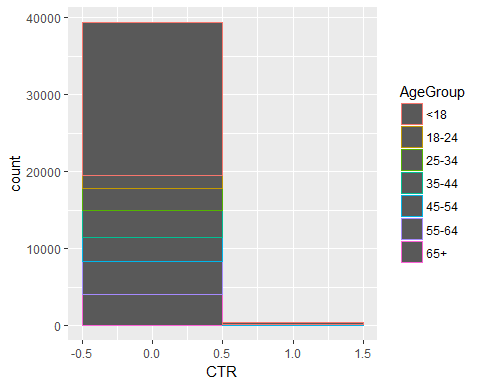
#Impressions density plot for nyt\_my  
#ggplot(nyt\_my,aes(x=Impressions, colour=AgeGroup))+geom\_density()  
  
# TODO: Plot distributions of number impressions and click-through-rate (CTR = click/impression)   
# for the age groups, that is, x: Impression, fill: AgeGroup  
  
nyt\_my$CTR <- nyt\_my$Clicks/nyt\_my$Impressions  
  
head(nyt\_my)

## 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

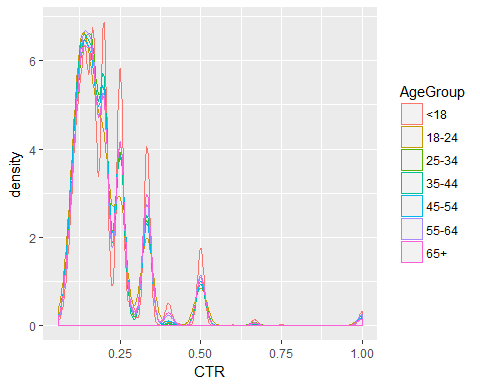
ggplot(nyt\_my, aes(x=Impressions ,color=AgeGroup)) + geom\_histogram(binwidth=1)



ggplot(subset(nyt\_my,CTR>0), aes(x=CTR ,color=AgeGroup)) + geom\_histogram(binwidth=1)



#ggplot(nyt\_my, aes(x=Impressions ,y=CTR,color=AgeGroup)) + geom\_histogram(binwidth=1)  
  
#CTR density plot for Clicks>0, group by day  
  
ggplot(subset(Imp2,Clicks >0),aes(x=CTR, colour=AgeGroup))+geom\_density()



head(Imp2)

## Age Gender Impressions Clicks Signed\_In AgeGroup CTR CTRGroup  
## 1 36 0 3 0 1 35-44 0.00000000 0<=CTR< 0.2  
## 2 73 1 3 0 1 65+ 0.00000000 0<=CTR< 0.2  
## 3 30 0 3 0 1 25-34 0.00000000 0<=CTR< 0.2  
## 4 49 1 3 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 5 47 1 11 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 6 47 0 11 1 1 45-54 0.09090909 0<=CTR< 0.2

# TODO: get the total number of Gender, Impressions, Clicks and Signed\_In  
# hint: Imp2's column value at 2, 3, 4, 5  
sapply(Imp2[2:5],sum)

## Gender Impressions Clicks Signed\_In   
## 167146 2295559 42449 319198

head(Imp2)

## Age Gender Impressions Clicks Signed\_In AgeGroup CTR CTRGroup  
## 1 36 0 3 0 1 35-44 0.00000000 0<=CTR< 0.2  
## 2 73 1 3 0 1 65+ 0.00000000 0<=CTR< 0.2  
## 3 30 0 3 0 1 25-34 0.00000000 0<=CTR< 0.2  
## 4 49 1 3 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 5 47 1 11 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 6 47 0 11 1 1 45-54 0.09090909 0<=CTR< 0.2

# TODO: get the mean of Age, Gender, Impressions, Clicks, CTR  
# hint: Imp2's column value at 1, 2, 3, 4, 7  
  
sapply(Imp2[,c(1,2,3,4,7)],mean)

## Age Gender Impressions Clicks CTR   
## 29.48400988 0.36705133 5.04102992 0.09321768 0.01847053

head(Imp2)

## Age Gender Impressions Clicks Signed\_In AgeGroup CTR CTRGroup  
## 1 36 0 3 0 1 35-44 0.00000000 0<=CTR< 0.2  
## 2 73 1 3 0 1 65+ 0.00000000 0<=CTR< 0.2  
## 3 30 0 3 0 1 25-34 0.00000000 0<=CTR< 0.2  
## 4 49 1 3 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 5 47 1 11 0 1 45-54 0.00000000 0<=CTR< 0.2  
## 6 47 0 11 1 1 45-54 0.09090909 0<=CTR< 0.2

#get the means by AgeGroup  
# error as it does not have doBy package?  
#summaryBy(Gender+Signed\_In+Impressions+Clicks+CTR~AgeGroup, data=Imp2)  
  
# TODO: Create table of the CTRGroup vs AgeGroup  
  
table(Imp2$CTRGroup,Imp2$AgeGroup)

##   
## <18 18-24 25-34 35-44 45-54 55-64 65+  
## 0<=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

# TODO: CTR density plot for Clicks>0, group by colour=AgeGroup  
  
ggplot(subset(Imp2,Clicks>0),aes(x=CTR, color=AgeGroup))+geom\_density()

