Post Live Session 8 - May 29th

Mike Martos

June 25, 2016

library(downloader)  
library(ggplot2)  
library(vcd)

## Loading required package: grid

## Download data

download("http://stat.columbia.edu/~rachel/datasets/nyt29.csv",  
 destfile="./Data/Raw/clickstream29.csv")  
clickstream29 <- read.csv("./Data/Raw/clickstream29.csv")

## Tidy data

#Get min and max of age  
min <- min(clickstream29$Age)  
max <- max(clickstream29$Age)  
#Create breaks  
breaks <- c(min,18,24,34,44,54,64,max)  
#Create new field for AgeGroup and assign from breaks  
clickstream29$AgeGroup <- cut(clickstream29$Age, breaks,   
 labels = c("18-", "18-24", "25-34", "35-44", "45-54", "55-64", "65+"),   
 include.lowest = TRUE, ordered\_result = TRUE)  
#Make the signed in a factor  
clickstream29$Signed\_In <- factor(clickstream29$Signed\_In)  
#Make the Gender in a factor  
clickstream29$Gender <- factor(clickstream29$Gender)  
#Remove records with no ages  
clickstream29Age <- subset(clickstream29, clickstream29$Age > 0)  
#Subset records with clicks  
clickstream29Click <- subset(clickstream29Age, clickstream29Age$Clicks > 0)

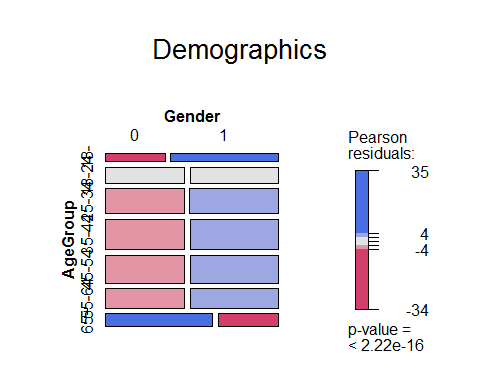
## Store intermediate files

write.csv(clickstream29, file = "./Data/ClickStream29AgeGroup.csv")  
write.csv(clickstream29Age, file = "./Data/ClickStream29AgeGroupWithAge.csv")  
write.csv(clickstream29Click, file = "./Data/ClickStream29AgeGroupWithAgeClicks.csv")

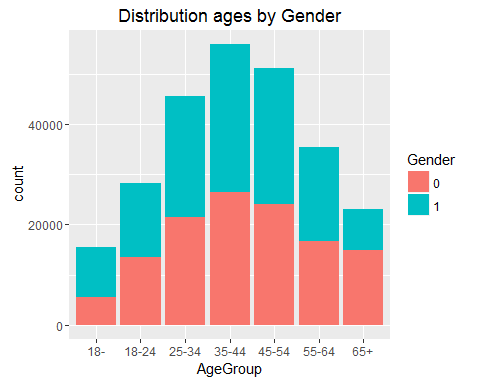
## Including Plots

## Try to learn about the data

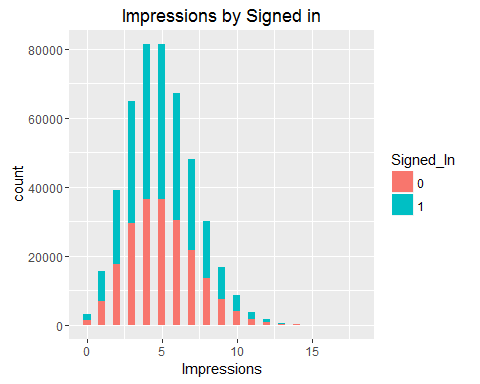
#Show demographics of signed in users  
mosaic(~ AgeGroup + Gender, data = clickstream29Age, shade = TRUE, main="Demographics")



#Distribution ages by Gender  
qplot(AgeGroup, data=clickstream29Age, fill=Gender, main="Distribution ages by Gender")



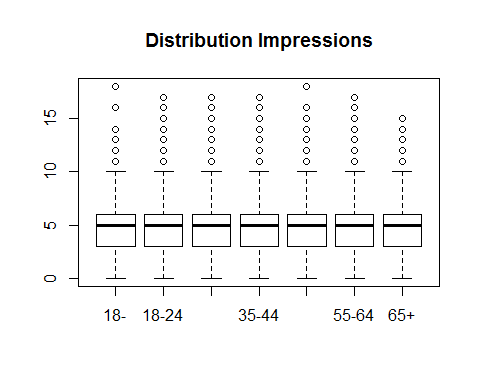
#Distribution Impressions by Signed in  
qplot(Impressions, data=clickstream29, fill=Signed\_In, main="Impressions by Signed in",binwidth = 0.5)



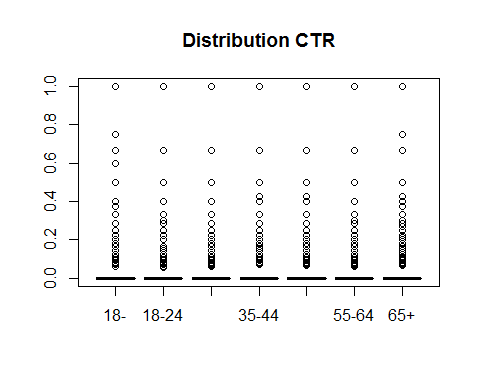
## Answer project questions

### At this point I start using a dataset that has information for Signed In users, as the questions are about Age Group and the only cases where we have that information is of signed in users, and I also focused on observations with one click or more as the data is skewed to the right if not.

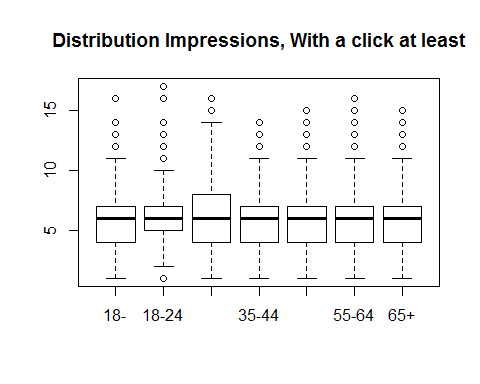
#With 0 Clicks  
plot(clickstream29Age$AgeGroup, clickstream29Age$Impressions, main="Distribution Impressions")



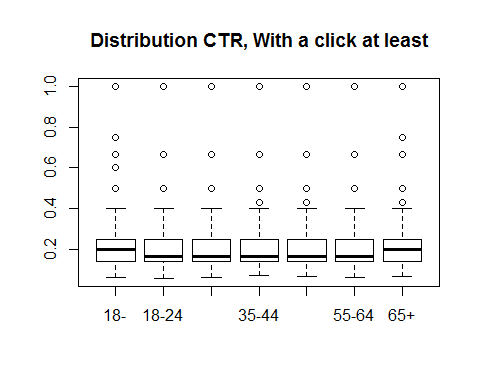
plot(clickstream29Age$AgeGroup, clickstream29Age$Clicks/clickstream29Age$Impressions, main="Distribution CTR")



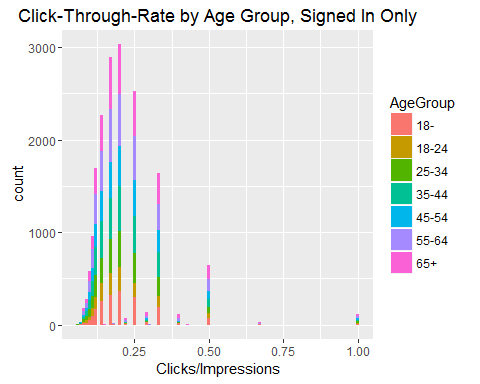
#Removing 0 clicks  
plot(clickstream29Click$AgeGroup, clickstream29Click$Impressions, main="Distribution Impressions, With a click at least")



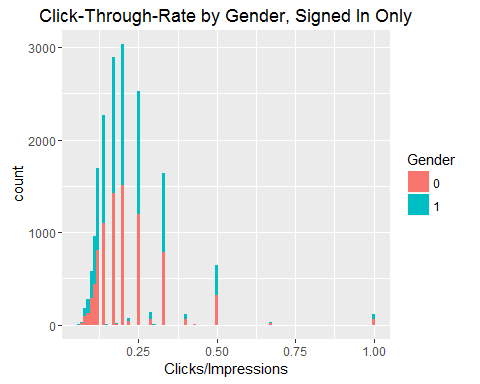
plot(clickstream29Click$AgeGroup, clickstream29Click$Clicks/clickstream29Click$Impressions,   
 main="Distribution CTR, With a click at least")



#CTR Exploratory Data  
qplot(Clicks/Impressions, data=clickstream29Click, fill=AgeGroup,   
 main="Click-Through-Rate by Age Group, Signed In Only",binwidth = 0.01)

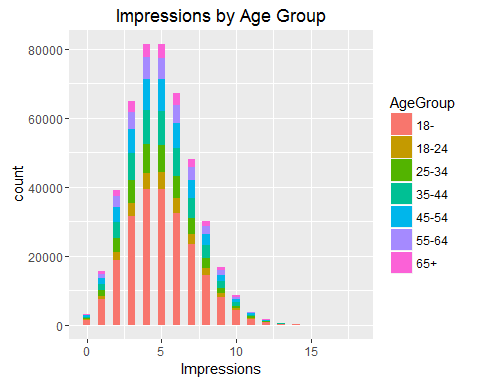


qplot(Clicks/Impressions, data=clickstream29Click, fill=Gender,   
 main="Click-Through-Rate by Gender, Signed In Only",binwidth = 0.01)

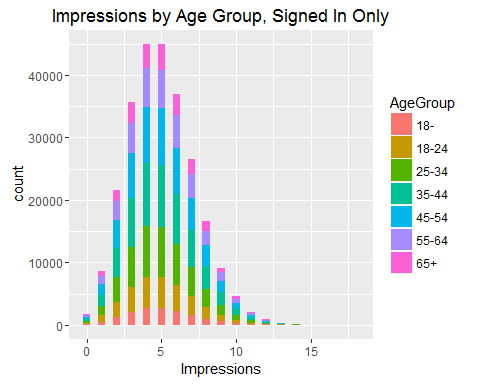


## Look at the data from other angles

#Impressions Exploratory Data  
qplot(Impressions, data=clickstream29, fill=AgeGroup,   
 main="Impressions by Age Group",binwidth = 0.5)



qplot(Impressions, data=clickstream29Age, fill=AgeGroup,   
 main="Impressions by Age Group, Signed In Only",binwidth = 0.5)



qplot(Impressions, data=clickstream29Age, fill=Gender,   
 main="Impressions by Age Group, Signed In Only",binwidth = 0.5)

