Homework 1

Liting Hu

October 4, 2016

Problem 1

1. Load and clean data

```
bx <- read.xls("rollingsales_bronx.xls", perl = "/usr/bin/perl", pattern="BOROUGH")
bk <- read.xls("rollingsales_brooklyn.xls", perl = "/usr/bin/perl", pattern="BOROUGH")
mh <- read.xls("rollingsales_manhattan.xls", perl = "/usr/bin/perl", pattern="BOROUGH")
qn <- read.xls("rollingsales_queens.xls", perl = "/usr/bin/perl", pattern="BOROUGH")
si <- read.xls("rollingsales_statenisland.xls", perl = "/usr/bin/perl", pattern="BOROUGH")
#Load in data

cleandata <- function(x) {
    names(x) <- tolower(names(x))
    sale.price.n <- as.numeric(gsub("[^[:digit:]]","",x$sale.price))
    #price should be numeric
    x$gross.square.feet <- as.numeric(gsub("[^[:digit:]]","",x$gross.square.feet))
    x$land.square.feet <- as.numeric(gsub("[^[:digit:]]","",x$land.square.feet))
    #gross square feet and land square feet should be numeric
    x$sale.date <- as.Date(x$sale.date)
    #time series
    x$year.built <- as.numeric(as.character(x$year.built))
    x <- data.frame(x, sale.price.n)
    mvc <- length(x$sale.price.n)-count(is.na(x$sale.price.n))[2]
    if (mvc == 0) message("No missing value")
    else message("There exists at least one missing value")
    #to see if there exists missing value
    x <- subset(x, log(x$sale.price.n) > 5)
    #delete data whose sale price is too low
    return(x)
}
bx <- cleandata(bx)
bk <- cleandata(bk)
mh <- cleandata(mh)
qn <- cleandata(qn)
si <- cleandata(si)
```

2. Add building type
We just want to analyze 1, 2, 3 family homes, coops, and condos. So create a new column called building.type in these data frame as follow

```
addbuildingtype <- function(x) {
    x$building.type <- "6 Others"
    x$building.type[grepl("ONE FAMILY",x$building.class.category)] <- "1 One family"
    x$building.type[grepl("TWO FAMILY",x$building.class.category)] <- "2 Two family"
```

```
    x$building.type[grepl("THREE FAMILY",x$building.class.category)] <- "3 Three family"
    x$building.type[grepl("COOPS",x$building.class.category)] <- "4 Coops"
    x$building.type[grepl("CONDOS",x$building.class.category)] <- "5 Condos"
    x$building.type <- factor(x$building.type)
    return(x)
}

bx <- addbuildingtype(bx)
bk <- addbuildingtype(bk)
mh <- addbuildingtype(mh)
qn <- addbuildingtype(qn)
si <- addbuildingtype(si)
```

3. Combine data In order to use ggplot in R, combine all data

```
aldata <- merge(mh, merge(bk, bx, all = TRUE), all = TRUE)
aldata <- merge(si, merge(qn, aldata, all = TRUE), all = TRUE)
aldata$borough <- as.factor(aldata$borough)
```

4. Data analysis
Firstly, draw the histogram of sale price as figure 1.
Most data are in the first two column. To judge the price more precisely, calculate logs of sale

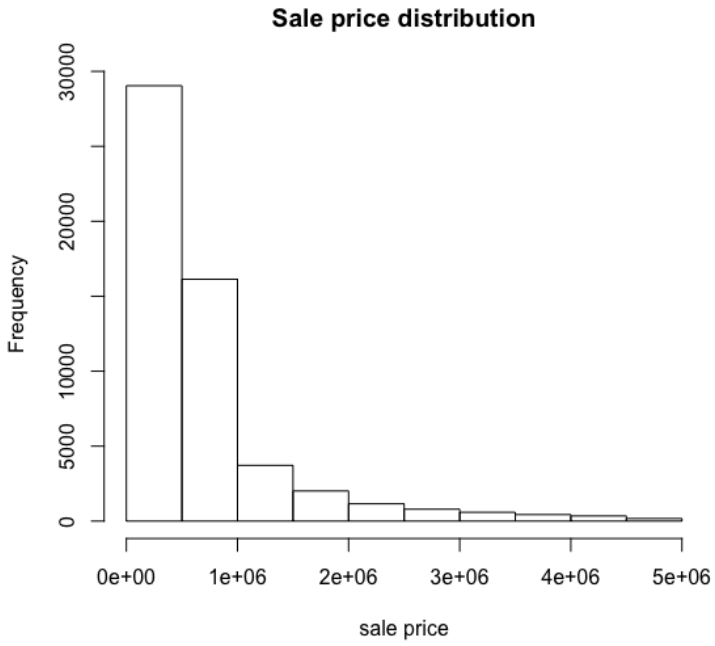


Figure 1: Sale price distribution

prices and draw the histogram as below.

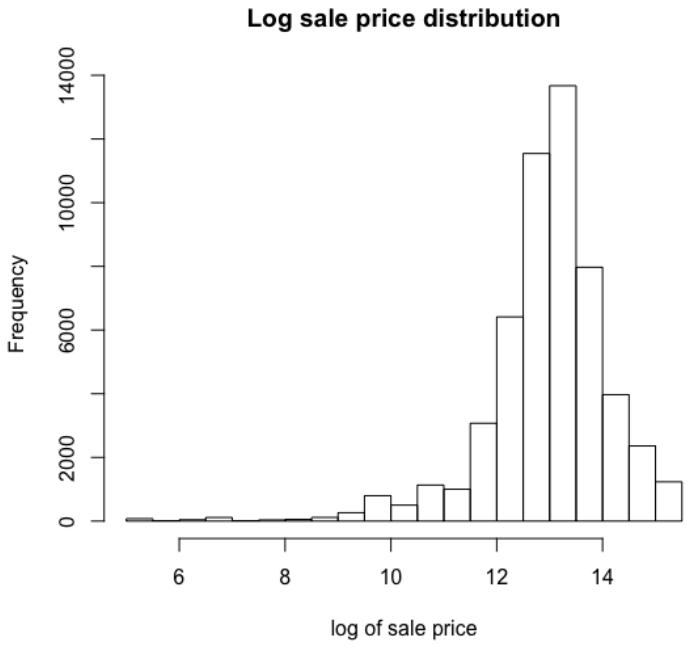


Figure 2: Log sale price distribution

This graph shows that most sale prices of these property is around $\$e^{13}$ ($442413) Than to figure out which patterns have influence on sale prices.

a) Built year

Figure 3: Does built year influence price

As showed above, built year has no much influence in sales prices of family homes while prices of coops and condos do fluctuate as built year changes.

b) Sale date

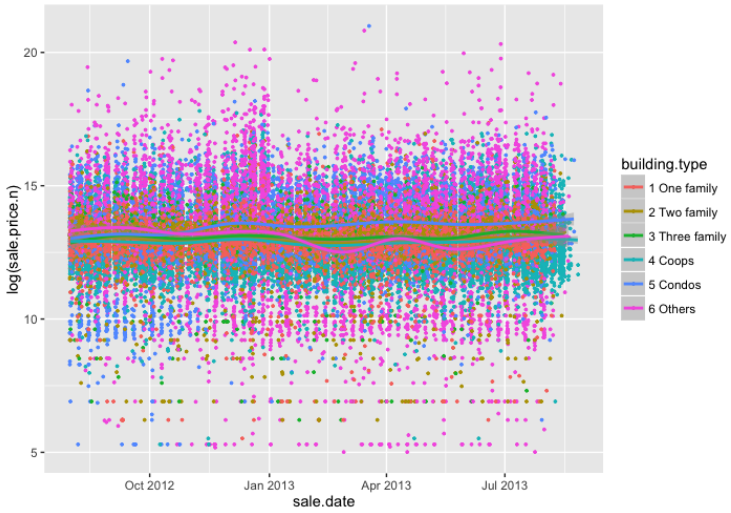


Figure 4: Does sale date influence price

During this year, the average sale price is steady.

c) Borough

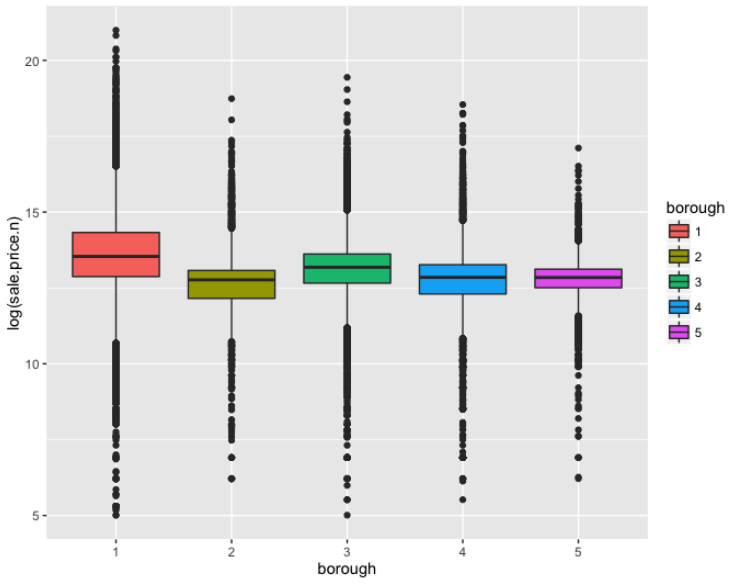


Figure 5: Does borough influence price (1-Manhattan, 2-Bronx, 3-Brooklyn, 4-Queen, 5-Statenisland)

Obviously, price in Manhattan is much higher than other borough while Bronx a little lower than others.

d) Building type

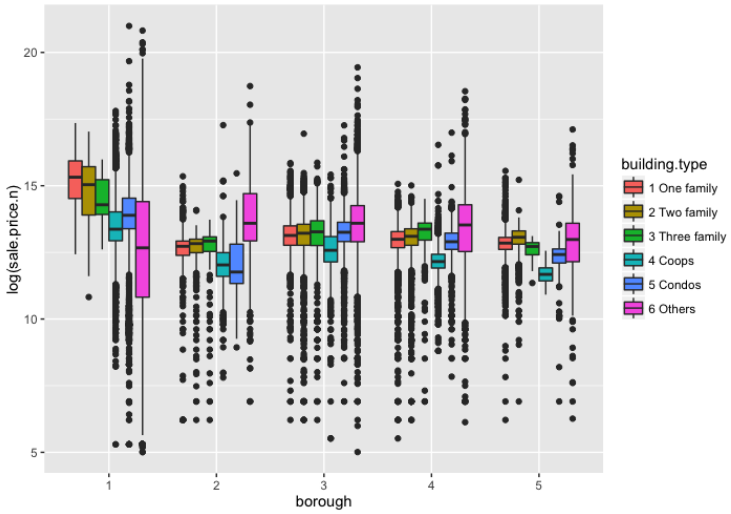


Figure 6: Does building type influence price

We can see that in every borough, price of coops are much lower than other type of building. In Manhattan, family homes are more valuable than coops and condos. But in other borough, price of family homes is close to condos'.

e) Gross square feet

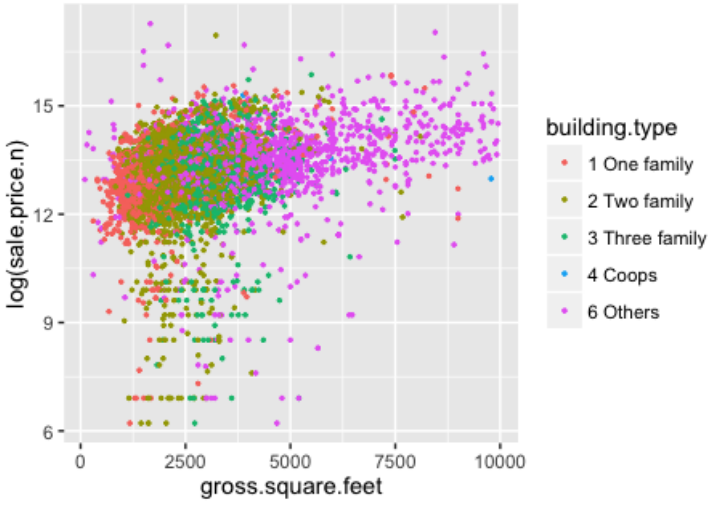


Figure 7: Does gross square feet influence price

Since most data cluster around (2500, 13)(ignore "others"), there is no obvious relationship

between gross square feet and sale price.

By the way, figure 8 shows the composition of estate in each borough.

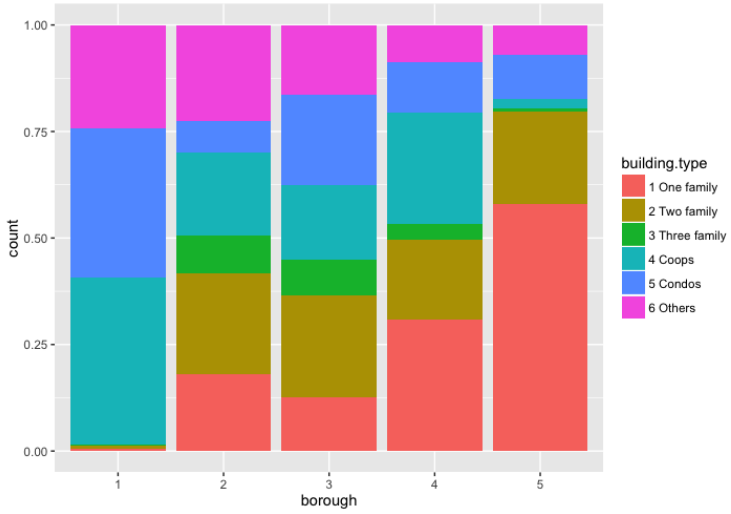


Figure 8: Building type percentage in each borough

We can see that most property on sale in Manhattan is coops and condos while contrarily in Statenisland the most popular property is the house. Maybe this is a factor result in the extra high price homes in Manhattan.

5. Conclusion
Among these elements, borough, building type and build year affect sale price while sale date and gross square feet have little to do with the price.

Problem 2

1. Load data

```
data1 <- read.csv("nyt1.csv")
data2 <- read.csv("nyt2.csv")
data3 <- read.csv("nyt3.csv")
```

2. Create age group and category based on their click behavior.

```
data1$age_group <- cut(data1$Age, c(-Inf,0,19,29,39,49,59,69,Inf))
data2$age_group <- cut(data2$Age, c(-Inf,0,19,29,39,49,59,69,Inf))
data3$age_group <- cut(data3$Age, c(-Inf,0,19,29,39,49,59,69,Inf))
createcate <- function(x) {
    x$scode[x$Impressions==0] <- "NoImps"
    x$scode[x$Impressions >0] <- "Imps"
    x$scode[x$Clicks >0] <- "Clicks"
    x$scode <- factor(x$scode)
    return(x)
}
data1 <- createcate(data1)
```

```
data2 <- createcate(data2)
data3 <- createcate(data3)
```

3. Combine data

```
data1$Day <- "Day 1"
data2$Day <- "Day 2"
data3$Day <- "Day 3"
data123 <- merge(data3, merge(data1, data2, all = T), all = T)
```

4. Analyze data

```
summaryBy(Age~age_group, data=data123, FUN=c(length,min,mean,max))
```

	age_group	Age.length	Age.min	Age.mean	Age.max
1	(-Inf,0]	403761	0	0.00000	0
2	(0,19]	75940	4	16.76189	19
3	(19,29]	169852	20	24.51417	29
4	(29,39]	189026	30	34.74697	39
5	(39,49]	198390	40	44.37435	49
6	(49,59]	160944	50	54.01256	59
7	(59,69]	95611	60	63.45984	69
8	(69, Inf]	55222	70	76.07676	111

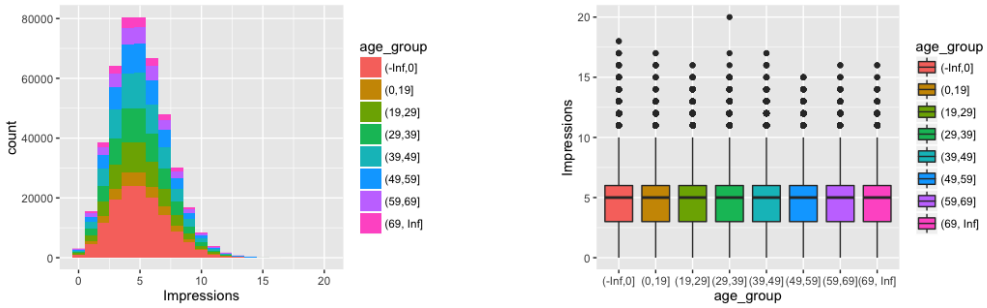
Figure 9: Summary by age group

```
summaryBy(Gender+Signed_In+Impressions+Clicks~age_group, data = data1)
```

	age_group	Gender.mean	Signed_In.mean	Impressions.mean	Clicks.mean
1	(-Inf,0]	0.0000000	0	4.999222	0.14188096
2	(0,19]	0.6208059	1	5.002699	0.11274691
3	(19,29]	0.5352071	1	4.993812	0.04996114
4	(29,39]	0.5359792	1	5.003624	0.05071789
5	(39,49]	0.5353949	1	5.000786	0.05059227
6	(49,59]	0.5352669	1	5.008065	0.07132294
7	(59,69]	0.4889500	1	5.008095	0.11609543
8	(69, Inf]	0.3582992	1	4.996396	0.15019376

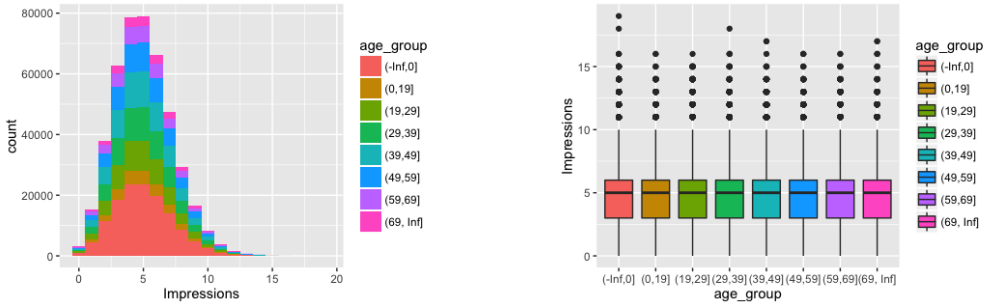
Figure 10: Summary by age group

From these two summaries, we get the information that only signed users have ages and genders. So after that when accessing to analysis about gender and sign status, we should ignore the age group (-Inf, 0].
Then draw the distribution of impression in these three days as follows:



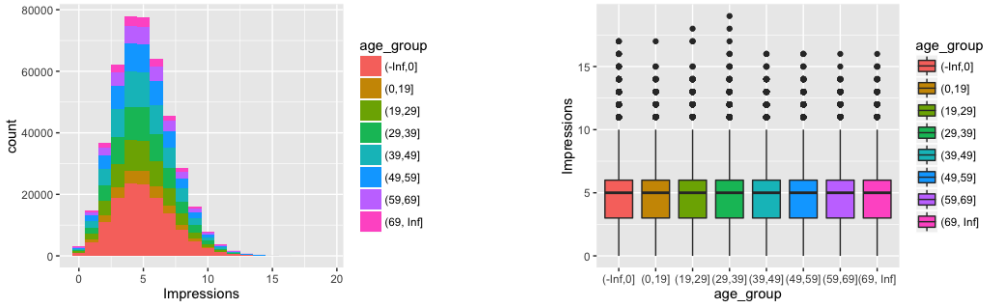
(a) Histogram (b) Boxplot

Figure 11: Impression distribution of day 1



(a) Histogram (b) Boxplot

Figure 12: Impression distribution of day 2



(a) Histogram (b) Boxplot

Figure 13: Impression distribution of day 3

There is no much difference between these three days' data.
To create click-through-rate, we don't consider clicks if there is no impression.

```
data1$hasimp <- cut(data1$Impressions,c(-Inf,0,Inf))
summaryBy(Clicks~hasimp, data=data1, FUN=c(length,min,mean,max))
ggplot(subset(data123, Clicks >0),aes(x=Clicks/Impressions,colour=age_group))+geom_density()
```

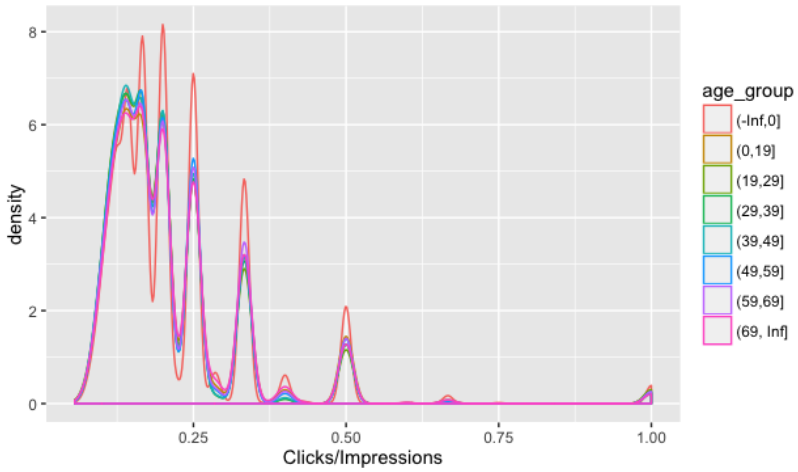


Figure 14: Click-through-rate distribution by age group

All age groups show similar thru rate tendency which reaches its peak at about 0.10 and has several extreme value.
To make comparisons between gender or sign class, create a new column in data frame then draw the plot.

```
data123$gender_group <- cut(data123$Gender, c(-Inf,0,Inf), labels = c("female", "male"))
ggplot(subset(data123, Clicks > 0 & Signed_In > 0),aes(x=Clicks/Impressions,colour=gender_group))+

data123$sign_group <- cut(data123$Signed_In, c(-Inf,0,Inf), labels = c("not signed in", "signed in
ggplot(subset(data123, Clicks >0),aes(x=Clicks/Impressions,colour=sign_group))+geom_density()
```

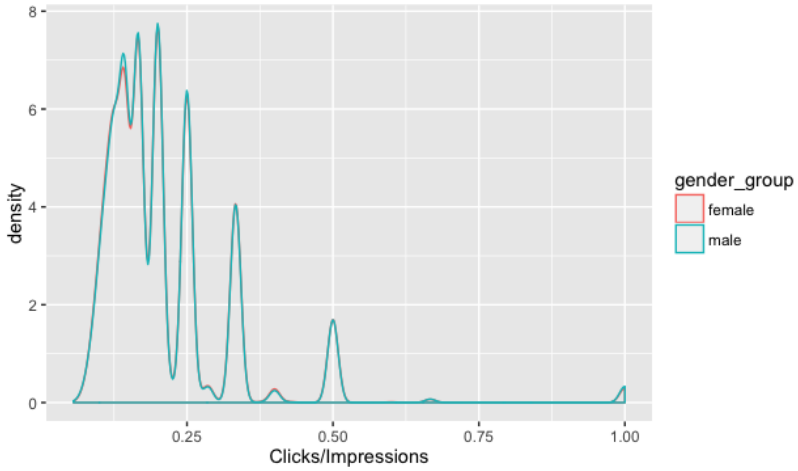


Figure 15: Click-through-rate distribution by gender

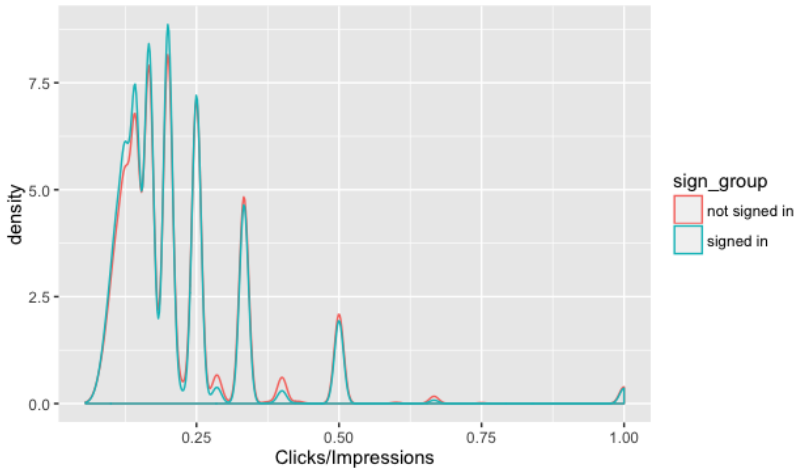


Figure 16: Click-through-rate distribution by sign status

There is no much Click-through-rate difference between gender groups or signed-in groups. As for impression by gender,
a) All age

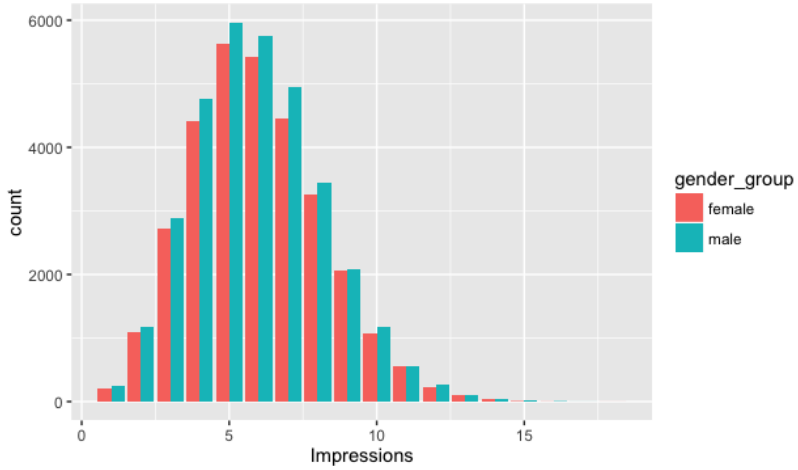


Figure 17: Impression by gender at all ages

b) Age < 20

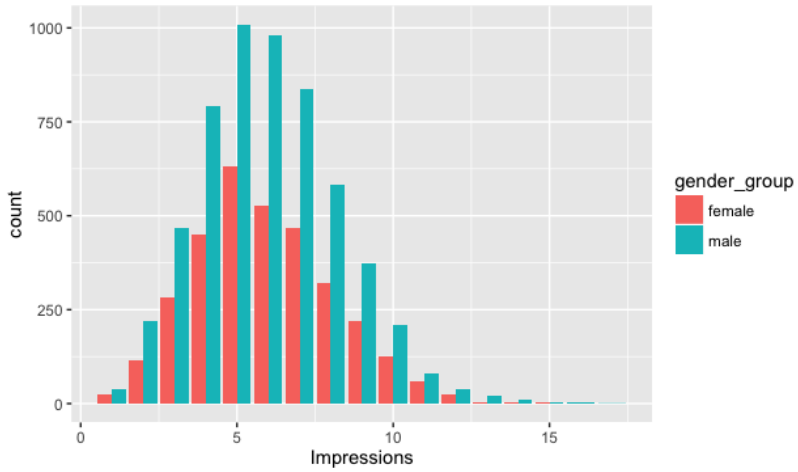


Figure 18: Impression by gender at age < 20

From figure 17 and figure 18, there are much more males than females have impressions on these ads but nor does it at all ages.
5.Conclusion Click-through-rate has similar trend across age group, gender, sign status and days. More young male than female have impression.