STEVENS INSTITUTE OF TECHNOLOGY
FE582 Homework 1
Instructor: Dragos Bozdog

Assignment 1

Student: Aidana Bekboeva

Contents

1 Problem 1

Explore realdirect.com thinking about how buyers and sellers would navigate, and how the website is organized. Use the datasets provided for Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Do the following:

- Load in and clean the data.

- Conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.

- Conduct exploratory data analysis to visualize and make comparisons for residential building category classes across boroughs and across time (select the following: 1-, 2-, and 3-family homes, coops, and condos). Use histograms, boxplots, scatterplots or other visual graphs. Provide summary statistics along with your conclusions.

1.1 Solution: 1. Loading the data:

Before starting the first problem, I loaded the given Excel files as spreadsheets into the new document main.xlsx. The filee can be accessed via link: `https://drive.google.com/drive/folders/1MbZCGEfvq4EORdI6Od9Uht1CVfrK4Z9s?usp=sharing` Then, I access them independently, and merge into the dataframe:

```
bronx <- read.xlsx("C:\\Users\\Aidana Bekboeva\\Documents\\HW1_F22\\main.xlsx", 1, startRow = 5)
brooklyn <- read.xlsx("C:\\Users\\Aidana Bekboeva\\Documents\\HW1_F22\\main.xlsx", 2, startRow = 5)
manhattan <- read.xlsx("C:\\Users\\Aidana Bekboeva\\Documents\\HW1_F22\\main.xlsx", 3, startRow = 5)
queens <- read.xlsx("C:\\Users\\Aidana Bekboeva\\Documents\\HW1_F22\\main.xlsx", 4, startRow = 5)
statenisland <- read.xlsx("C:\\Users\\Aidana Bekboeva\\Documents\\HW1_F22\\main.xlsx", 5, startRow = 5)

#Merging the spreadsheets
df <- rbind(bronx, brooklyn, manhattan, queens, statenisland)
```

Listing 1: Loading the data

1.2 Cleaning the data, exploratory data analysis:

Cleaning the data involved modification of invalid data types, such as dates in column SALE.DATE (from numeric to datetime), adjusting variable names and investigating empty columns, resulting in the column ‘EASE-MENT‘ being removed, and data from APART.MENT.NUMBER being transferred to ADDRESS, and then removed.

```
#Modifying data types
glimpse(df)
df$SALE.DATE <- openxlsx::convertToDateTime(df$SALE.DATE)
```

```
#Investigating empty columns and dropping them
df$`EASE-MENT`
df = subset(df, select = -c(`EASE-MENT`) )
glimpse(df)

#Transferring important, yet rare information in APART.MENT.NUMBER
df$APART.MENT.NUMBER
df$ADDRESS <- paste(df$ADDRESS, df$APART.MENT.NUMBER)
df$ADDRESS
df = subset(df, select = -c(APART.MENT.NUMBER) )

#Adjusting not descriptive variable names
df$BOROUGH <- gsub("1", "1. MANHATTAN", df$BOROUGH)
df$BOROUGH <- gsub("2", "2. BRONX", df$BOROUGH)
df$BOROUGH <- gsub("3", "3. BROOKLYN", df$BOROUGH)
df$BOROUGH <- gsub("4", "4. QUEENS", df$BOROUGH)
df$BOROUGH <- gsub("5", "5. STATEN ISLAND", df$BOROUGH)
df$BOROUGH
```

Listing 2: Cleaning the data

In columns, where 0 provides misleading information and is supposed to indicate a missing value, we replace 0 with NA. Such columns include data on the year the building was built, land sq feet, gross sq feet, sale price and zip code.

Additionally, the data also has invalid entries, such as: 1. price being unrealistically low, 2. land sq feet and gross sq feet less than the legal minimum (70 sq feet), 3. buildings with 0 units (one whole building is 1 unit minimum).

Taking a closer look on five-number summaries for each variable, we are able to identify obvious outliers that damage the consistency of data. Outliers include rows with excessive number of residential units (Residential units - Max.: 8270.00). Restricting the max and min for those variables terminates outliers without hurting the data.

```

#Formatting missing values where 0 is inexplicable
df$YEAR.BUILT[df$YEAR.BUILT == 0] <- NA
df$LAND.SQUARE.FEET[df$LAND.SQUARE.FEET == 0] <- NA
df$GROSS.SQUARE.FEET[df$GROSS.SQUARE.FEET == 0] <- NA
df$SALE.PRICE[df$SALE.PRICE == 0] <- NA
df$ZIP.CODE[df$ZIP.CODE == 0] <- NA

#Removing the invalid entries
df <- df[-c(which(df$SALE.PRICE < 10000)), ]
df <- df[-c(which(df$LAND.SQUARE.FEET < 70)), ]
df <- df[-c(which(df$GROSS.SQUARE.FEET < 70)), ]
df <- df[-c(which(df$TOTAL.UNITS == 0)), ]

#Detecting outliers
summary(df)
df <- df[-c(which(df$RESIDENTIAL.UNITS > 1500)), ]
```

Listing 3: Missing values and outliers

1.3 Visualizing and making comparisons

The most descriptive histogram of all the numerical data in our dataframe illustrates the number of buildings built each year:

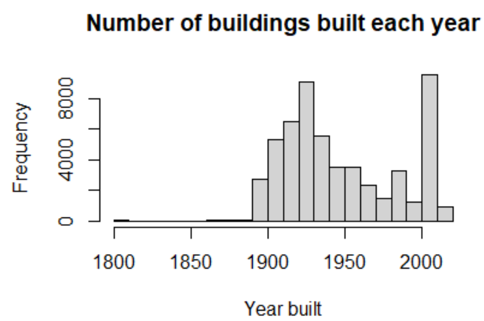
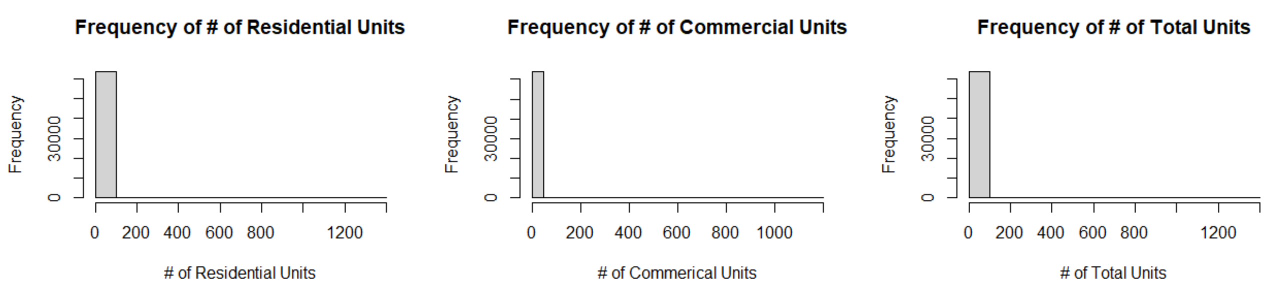


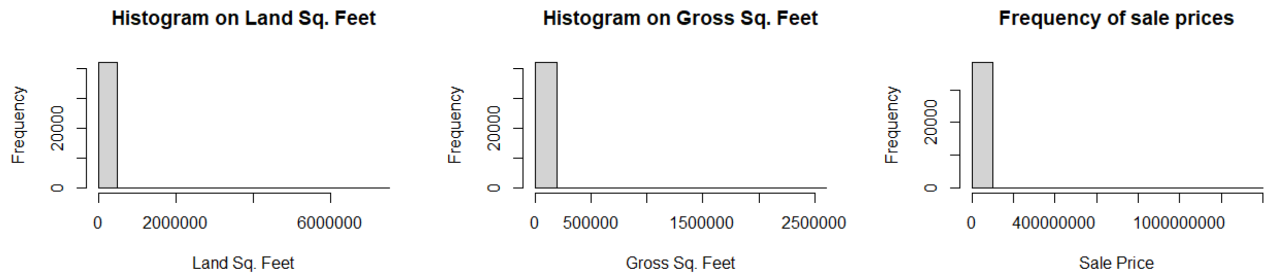
Figure 1: Task 6

Based on the histogram, since the 1900s the number of buildings had been following a bell-curved distribution, skewed to the right closer to the end of the XX century. Then, in the early 2000s we observe a dramatic rise in the number of buildings built, however, this tendency does not last long, and in the following years, the number of newly built buildings falls back down.

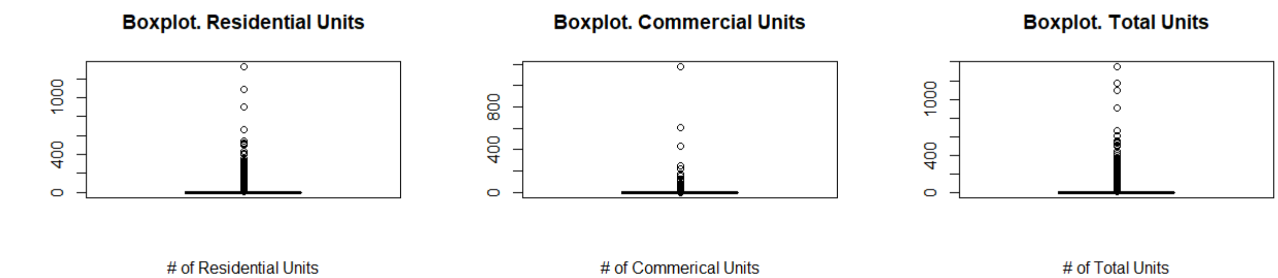
The histogram clearly depicts that the majority of the buildings across five boroughs (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) were established in the 1920s and 2000s. It is safe to assume that a large number of new establishments in the 1920s, the so-called, "Golden Twenties" is correlated with the economic boom following World War I; the drastic decline (end of 1920s), on contrary, could correlate with the crash of the stock market in 1929, and the Great Depression.

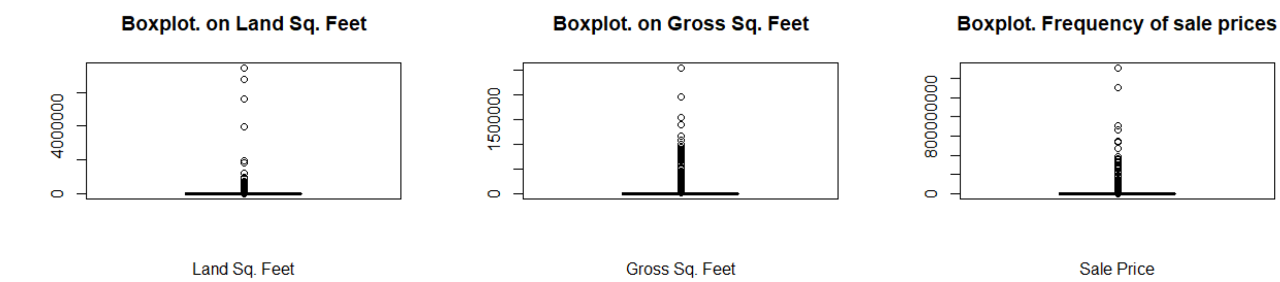
Although the following histograms accurately represent the data, they are not descriptive. All of them indicate that a large population of buildings acquire the least amount of each determining factor: the number of residential or commercial units, land square feet and price.





On the other hand, the histograms illustrated that the data is versatile and numbers, that drastically differ from the majority, need further investigation. Next tools to visualize the data are the five-number summaries, and their graphic representations in boxplots.





1.4 Exploratory analysis for 1-, 2-, and 3-family homes, coops, and condos

```
unique(df$BUILDING.CLASS.CATEGORY)

BUILDING.CLASS.CATEGORY.PLOT <- df%>%filter(str_detect(df$BUILDING.CLASS.
    CATEGORY,
                      "01  ONE FAMILY HOMES                         |
                      02  TWO FAMILY HOMES                         |
                      03  THREE FAMILY HOMES                       |
                      13  CONDOS - ELEVATOR APARTMENTS             |
                      12  CONDOS - WALKUP APARTMENTS               |
```

```
                  11A  CONDO-RENTALS                              |
                  09   COOPS - WALKUP APARTMENTS                  |
                  10   COOPS - ELEVATOR APARTMENTS                |
                  15   CONDOS - 2-10 UNIT RESIDENTIAL             |
                  16   CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT |
                  17   CONDOPS                                    "))

class(BUILDING.CLASS.CATEGORY.PLOT)
summary(BUILDING.CLASS.CATEGORY.PLOT)

plot(BUILDING.CLASS.CATEGORY.PLOT$YEAR.BUILT, BUILDING.CLASS.CATEGORY.PLOT
   $SALE.PRICE, main="Price and year",
     xlab="Year built", ylab="Sale Price", pch=20)

plot(BUILDING.CLASS.CATEGORY.PLOT$YEAR.BUILT, BUILDING.CLASS.CATEGORY.PLOT
   $GROSS.SQUARE.FEET, main="Gross sq.feet and year",
     xlab="Year built", ylab="Gross Sq Feet", pch=20)
```

Listing 4: Building class categories

Plot 1. Price and Year

The following plot is used to describe the trend of selling prices throughout the decades.

It is noticeable that in the first half of the XX century, the selling prices were higher overall, and, additionally, had generally more understandingly high prices. Throughout the years, the general trend for the majority of the prices has not been drastically moving. However, the closer look at the last decade suggests that the prices have been rising.

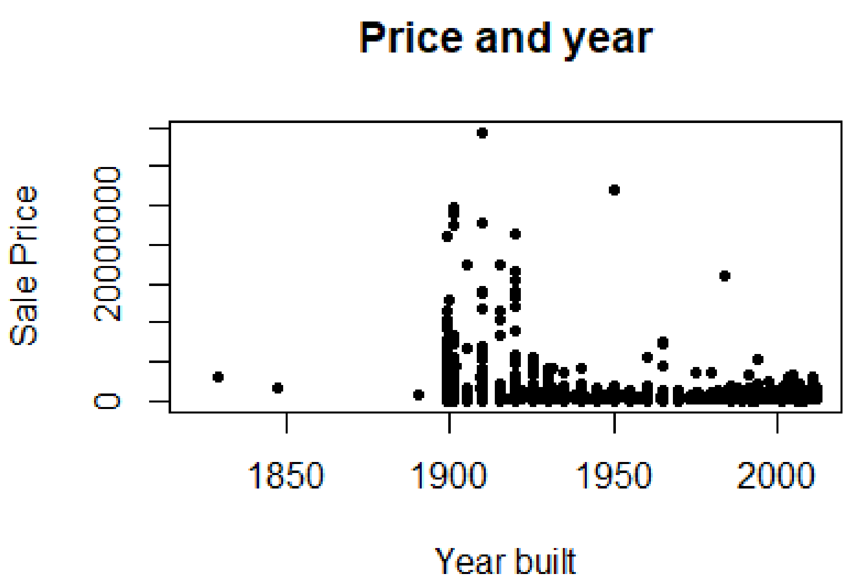


Figure 2: Task 6

Plot 2. Gross sq. feet and year

This plot illustrates the correlation between the year the building had been built, and the gross square feet indicator.

Overall, the peaks of the general trend form a U-shape with a significant condensing towards the right-hand side, indicating a larger number of units of data that lie between years 1970 and 2000s.

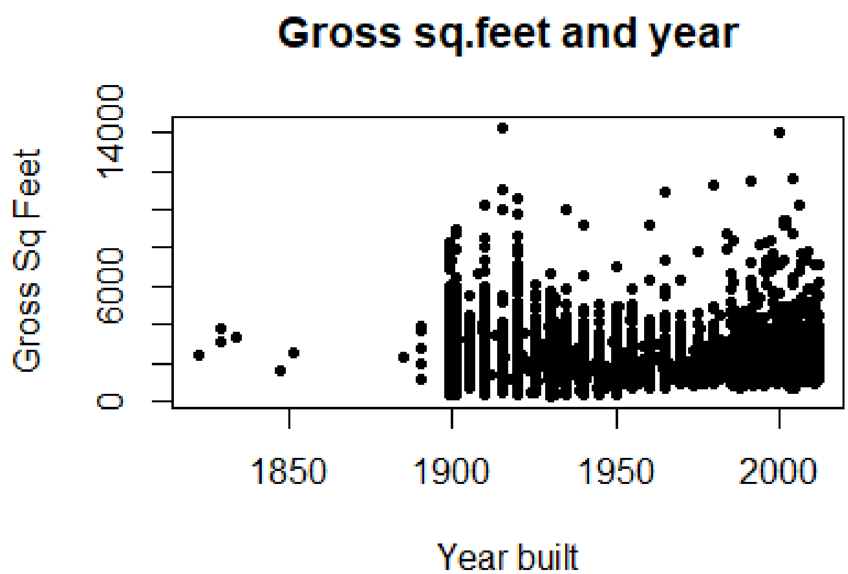


Figure 3: Task 6

2 Problem 2

The datasets provided nyt1.csv, nyt2.csv, and nyt3.csv represents three (simulated) days of ads shown and clicks recorded on the New York Times homepage. Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged-in. Use R to handle this data. Perform some exploratory data analysis:

1. Create a new variable, age_group, that categorizes users as "< 20", "20-29", "30-39", "40-49", "50-59", "60-69", and "70+".

2. For each day:
- Plot the distribution of number of impressions and click-through-rate (CTR = clicks / impressions) for these age categories
- Define a new variable to segment or categorize users based on their click behavior.
- Explore the data and make visual and quantitative comparisons across user segments/demographics (¡20-year-old males versus ¡20-year-old females or logged-in versus not, for example).
- Extend your analysis across days. Visualize some metrics and distributions over time.

2.1 Categorizing users into age groups

```

#Assigning NA to user whose age is 0 to avoid confusion
nyt1$Age[nyt1$Age == 0] <- NA
nyt2$Age[nyt2$Age == 0] <- NA
nyt3$Age[nyt3$Age == 0] <- NA

#Create a new variable, age_group, that categorizes users

nyt1$Age_Group <- cut(nyt1$Age,
                 breaks = c(-Inf, 20, 30, 40, 50, 60, 70, 120),
                 labels = c("<20", "20-29", "30-39",
                            "40-49", "50-59", "60-69", "70+"),
                 right=FALSE)

nyt2$Age_Group <- cut(nyt2$Age,
                      breaks = c(-Inf, 20, 30, 40, 50, 60, 70, 120),
                      labels = c("<20", "20-29", "30-39",
                                 "40-49", "50-59", "60-69", "70+"),
                      right=FALSE)

nyt3$Age_Group <- cut(nyt3$Age,
                      breaks = c(-Inf, 20, 30, 40, 50, 60, 70, 120),
                      labels = c("<20", "20-29", "30-39",
                                 "40-49", "50-59", "60-69", "70+"),
                      right=FALSE)
```

Listing 5: Categorizing users

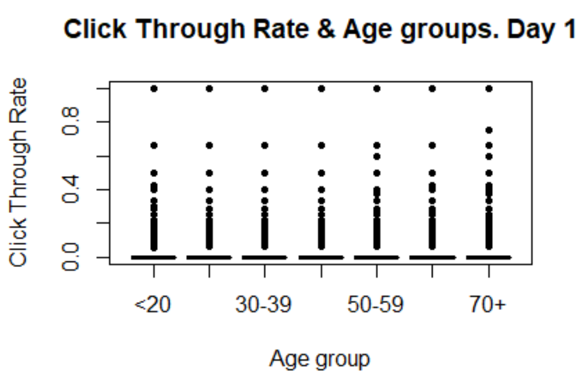
2.2 Number of impressions and click-through-rate

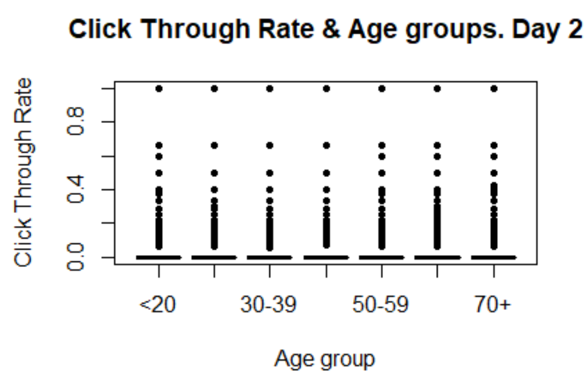
1. Plotting distribution of click-through-rate

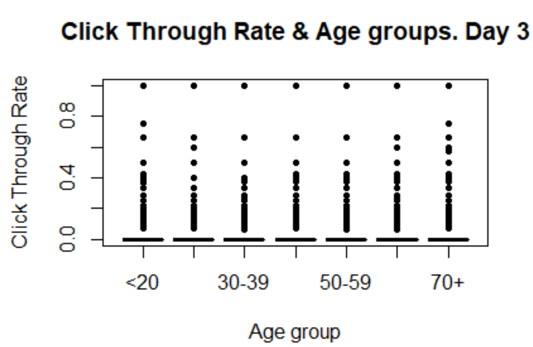
Given the formula, we create a new variable in each of the data sets (nyt1, nyt2, and nyt3) that represents the click-through-rate value. Next, we plot the distribution of click-through-rate along different age groups. We expand our observations across 3 days, which results in three scatter-plots shown below.

```
#Plotting distribution of click-through-rate

plot(nyt1$Age_Group, nyt1$Click_Through_Rate, main="Click Through Rate & Age groups. Day 1",
     xlab="Age group", ylab="Click Through Rate", pch=20)

plot(nyt2$Age_Group, nyt2$Click_Through_Rate, main="Click Through Rate & Age groups. Day 2",
     xlab="Age group", ylab="Click Through Rate", pch=20)

plot(nyt3$Age_Group, nyt3$Click_Through_Rate, main="Click Through Rate & Age groups. Day 3",
     xlab="Age group", ylab="Click Through Rate", pch=20)
```

Listing 6: CTR Distribution





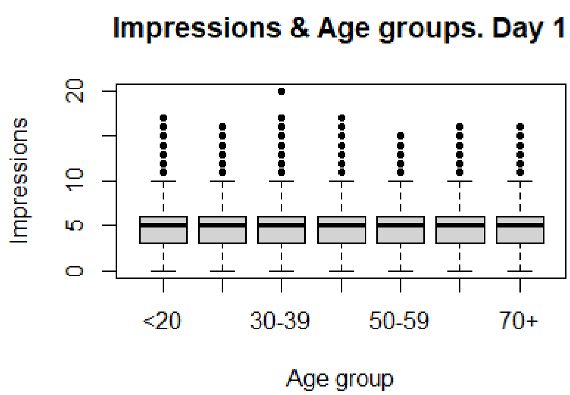


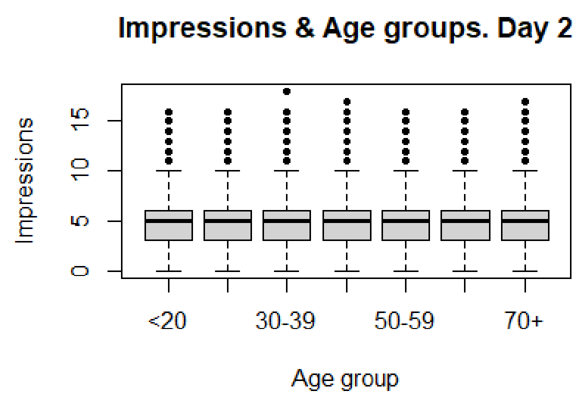
2. Plotting distribution of impressions

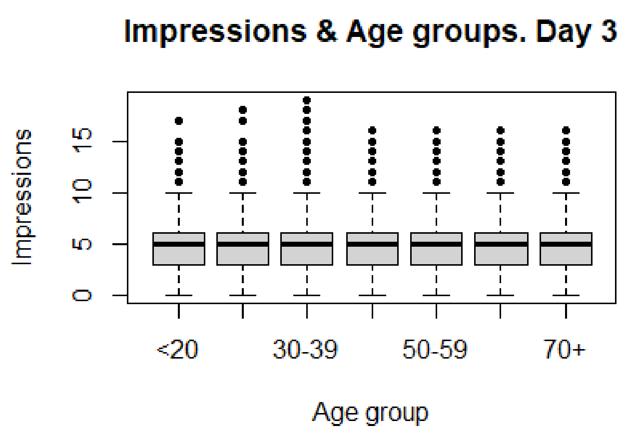
Next, we plot the distribution of impressions along different age groups. Again, we plot observations across 3 days.

```
#Plotting distribution of impressions

plot(nyt1$Age_Group, nyt1$Impressions, main="Impressions & Age groups. Day
    1", xlab="Age group", ylab="Impressions", pch=20)

plot(nyt2$Age_Group, nyt2$Impressions, main="Impressions & Age groups. Day
    2", xlab="Age group", ylab="Impressions", pch=20)

plot(nyt3$Age_Group, nyt3$Impressions, main="Impressions & Age groups. Day
    3", xlab="Age group", ylab="Impressions", pch=20)
```

Listing 7: CTR







2.3 Categorizing users based on their click behavior

```
#Categorizing users based on their click behavior

summary(nyt1$Clicks)
unique(nyt1$Clicks)

nyt1$User_Type <- cut(nyt1$Clicks,
                      breaks = c(0, 1, 2, 3, 4, 5),
                      labels = c("Passive", "Somewhat passive",
                                 "Moderately active", "Intensly active", "
   Extremely active"),
                      right=FALSE)

nyt2$User_Type <- cut(nyt2$Clicks,
                      breaks = c(0, 1, 2, 3, 4, 5),
                      labels = c("Passive", "Somewhat passive",
                                 "Moderately active", "Intensly active", "
   Extremely active"),
                      right=FALSE)

nyt3$User_Type <- cut(nyt3$Clicks,
                      breaks = c(0, 1, 2, 3, 4, 5),
                      labels = c("Passive", "Somewhat passive",
                                 "Moderately active", "Intensly active", "
   Extremely active"),
                      right=FALSE)
```

Listing 8: Segmentation

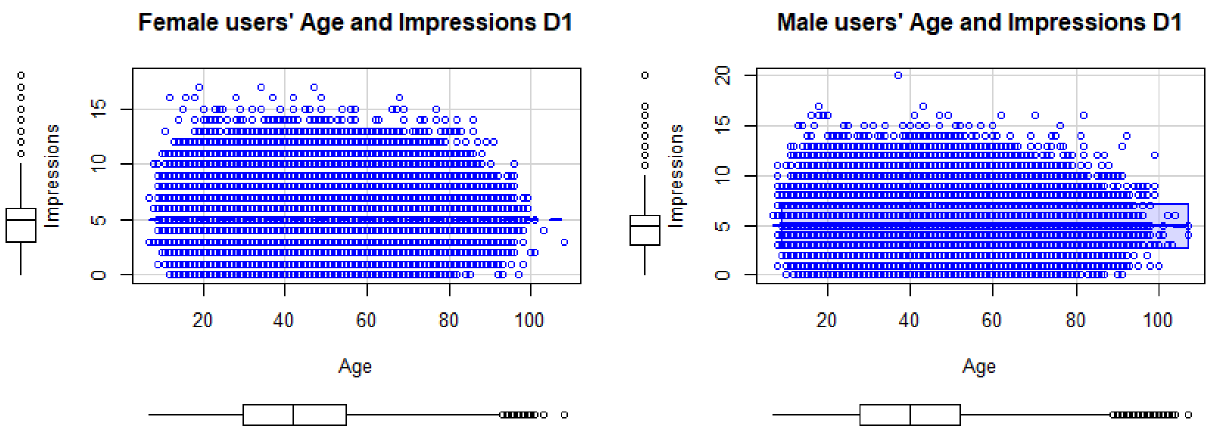
2.4 Visual and quantitative comparisons

Day 1.

We begin with creating 2 new dataframes by subsetting the female (0) and male (1) categories in the Gender column of our main data frame for day 1 - nyt1. Next, we plot scatterplots for female and male users, with impressions plotted for both of these groups.

```
femaleusers <- subset(nyt1, Gender==0, select=Age:Age_Group)
maleusers <- subset(nyt1, Gender==1, select = Age:Age_Group)

scatterplot(Impressions ~ Age, data = femaleusers, main="Female users' Age
    and Impressions D1")
scatterplot(Impressions ~ Age, data = maleusers, main="Male users' Age and
    Impressions D1")
```

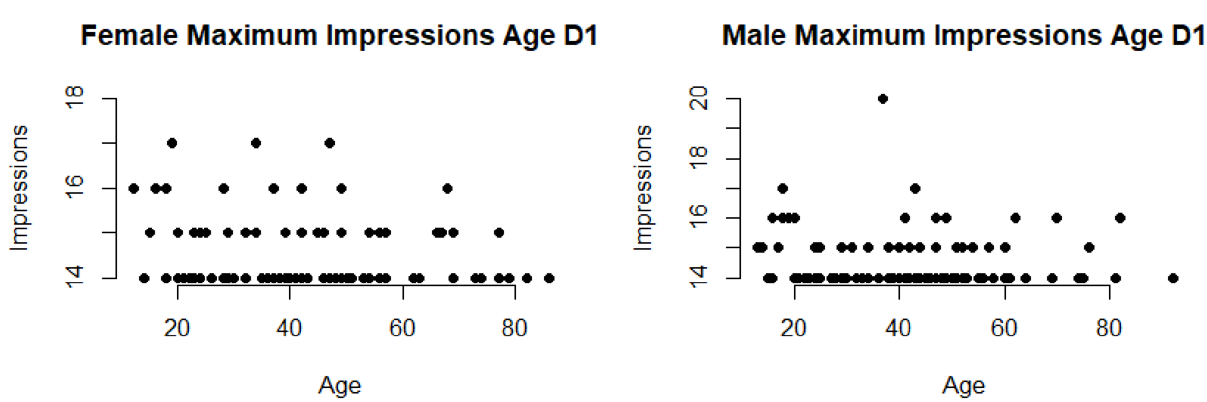
Listing 9: Subsetting the daily data by gender. Day 1 (nyt1)



The scatterplotts presented above illustrate allocation of all the female and male users based off of their impressions index and their age. Now, we are interested to plot a detailed representation of what ages are predominantly having higher number of impressions classified in two groups of women and men.

```
summary(femaleusers$Impressions)

MaxImpressions_F_D1 <- subset(femaleusers, Impressions > 13)
plot(MaxImpressions_F_D1$Age, MaxImpressions_F_D1$Impressions, main = "
    Female Maximum Impressions Age D1",
      xlab = "Age", ylab = "Impressions",
      pch = 19, frame = FALSE)

summary(maleusers$Impressions)

MaxImpressions_M_D1 <- subset(maleusers, Impressions > 13)
plot(MaxImpressions_M_D1$Age, MaxImpressions_M_D1$Impressions, main = "
    Male Maximum Impressions Age D1",
      xlab = "Age", ylab = "Impressions",
      pch = 19, frame = FALSE)
```

Listing 10: Highest impressions. Day 1

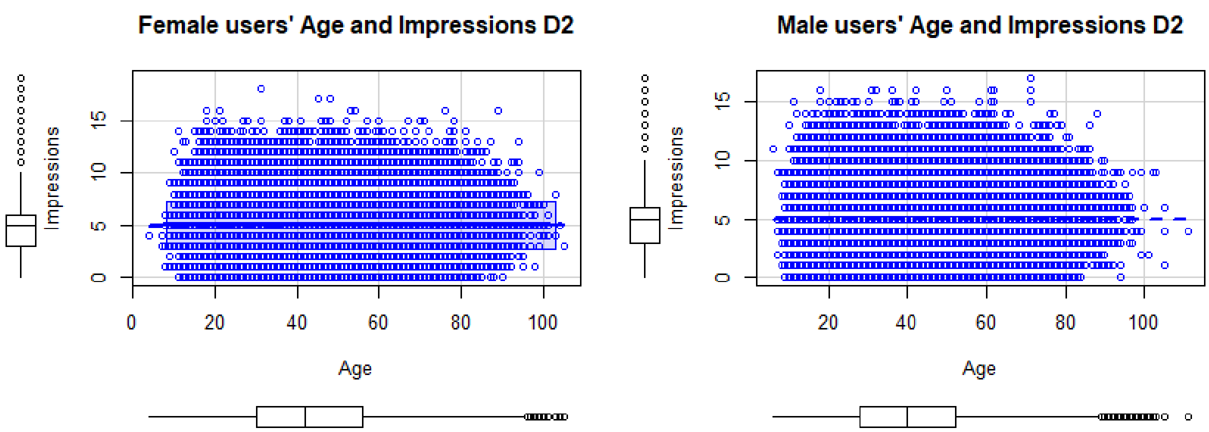


Day 2.

Let's plot similar graphics for the second and the third day.

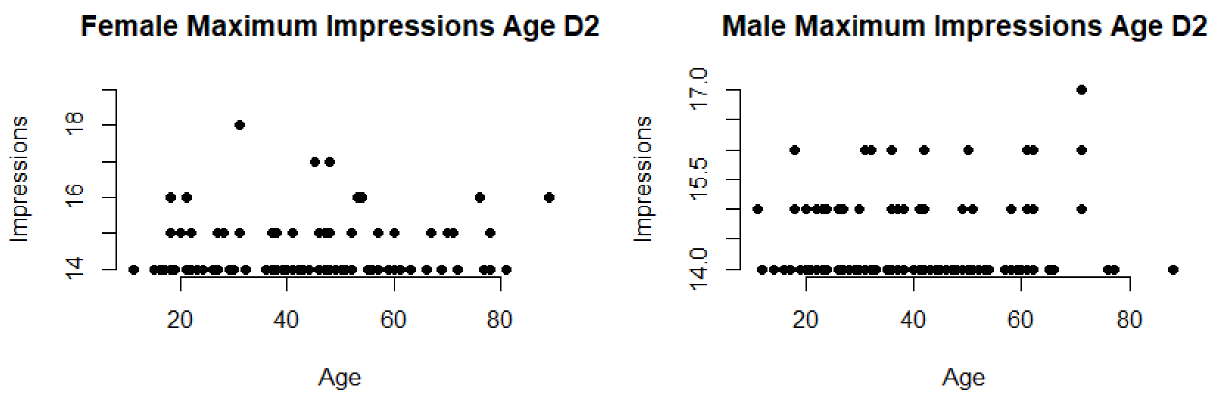
```r
#Day 2

femaleusers2 <- subset(nyt2, Gender==0, select=Age:Age_Group)
maleusers2 <- subset(nyt2, Gender==1, select = Age:Age_Group)

scatterplot(Impressions ~ Age, data = femaleusers2, main="Female users' Age and Impressions D2")
scatterplot(Impressions ~ Age, data = maleusers2, main="Male users' Age and Impressions D2")
```

Listing 11: Subsetting the daily data by gender. Day 2 (nyt1)




```r

MaxImpressions_F_D2 <- subset(femaleusers2, Impressions > 13)
plot(MaxImpressions_F_D2$Age, MaxImpressions_F_D2$Impressions, main = "Female Maximum Impressions Age D2",
     xlab = "Age", ylab = "Impressions",
     pch = 19, frame = FALSE)

MaxImpressions_M_D2 <- subset(maleusers2, Impressions > 13)
plot(MaxImpressions_M_D2$Age, MaxImpressions_M_D2$Impressions, main = "Male Maximum Impressions Age D2",
     xlab = "Age", ylab = "Impressions",
     pch = 19, frame = FALSE)
```

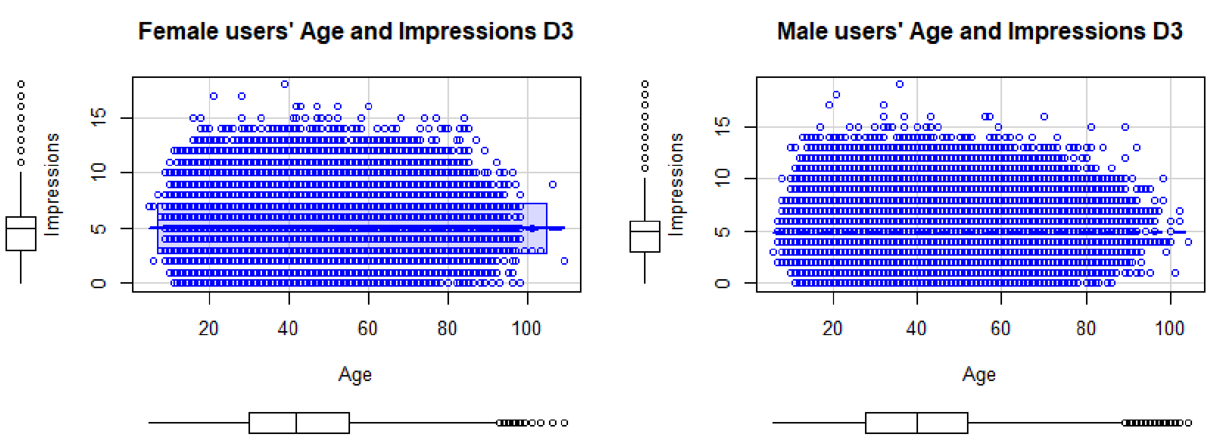
Listing 12: Highest impressions. Day 2



Day 3.

```
femaleusers3 <- subset(nyt3, Gender==0, select=Age:Age_Group)
maleusers3 <- subset(nyt3, Gender==1, select = Age:Age_Group)

scatterplot(Impressions ~ Age, data = femaleusers3, main="Female users' Age and Impressions D3")
scatterplot(Impressions ~ Age, data = maleusers3, main="Male users' Age and Impressions D3")
```

Listing 13: Subsetting the daily data by gender. Day 3 (nyt1)




```
MaxImpressions_F_D3 <- subset(femaleusers3, Impressions > 13)
plot(MaxImpressions_F_D3$Age, MaxImpressions_F_D3$Impressions, main = "Female Maximum Impressions Age D3",
     xlab = "Age", ylab = "Impressions",
     pch = 19, frame = FALSE)

MaxImpressions_M_D3 <- subset(maleusers3, Impressions > 13)
plot(MaxImpressions_M_D3$Age, MaxImpressions_M_D3$Impressions, main = "Male Maximum Impressions Age D3",
```

```
    xlab = "Age", ylab = "Impressions",
    pch = 19, frame = FALSE)
```

Listing 14: Highest impressions. Day 3