Week 12 Independent Project

Wilkister Mbaka

2022-07-15

```
## R Markdown

##Defining the Question
#a) Specifying the Question
#Perform exploratory data analysis on the provided dataset and identify which individuals are
#most likely to click on her ads.

#b) Defining the metrics of success
#To perform univariate and bivariate data analysis and based on that, provide recommendations
#and on which individuals are most likely to click on her ads.

#c) Understanding the context
#A Kenyan entrepreneur has created an online cryptography course and would want to advertise
#it on her blog. She currently targets audiences originating from various countries. In the past,
#she ran ads to advertise a related course on the same blog and collected data in the process.
#She would now like to employ your services as a Data Science Consultant to help her identify
#which individuals are most likely to click on her ads.

#d) experimental design taken
#1) Load the dataset
#2) Clean the dataset.
#3) perform Exploratory data analysis
```

```
#installing packages
library(data.table)
#
#Loading the dataset
advert <- fread("http://bit.ly/IPAdvertisingData")
#Previewing the first 6 rows
head(advert)
```

```
##    Daily Time Spent on Site Age Area Income Daily Internet Usage
## 1:                    68.95  35    61833.90               256.09
## 2:                    80.23  31    68441.85               193.77
## 3:                    69.47  26    59785.94               236.50
## 4:                    74.15  29    54806.18               245.89
## 5:                    68.37  35    73889.99               225.58
## 6:                    59.99  23    59761.56               226.74
##                            Ad Topic Line           City Male    Country
## 1:    Cloned 5thgeneration orchestration    Wrightburgh    0    Tunisia
## 2:    Monitored national standardization      West Jodi    1      Nauru
```

```
## 3:      Organic bottom-line service-desk       Davidton    0 San Marino
## 4: Triple-buffered reciprocal time-frame West Terrifurt    1      Italy
## 5:         Robust logistical utilization   South Manuel    0    Iceland
## 6:       Sharable client-driven software      Jamieberg    1     Norway
##              Timestamp Clicked on Ad
## 1: 2016-03-27 00:53:11             0
## 2: 2016-04-04 01:39:02             0
## 3: 2016-03-13 20:35:42             0
## 4: 2016-01-10 02:31:19             0
## 5: 2016-06-03 03:36:18             0
## 6: 2016-05-19 14:30:17             0
```

```
#Checking the datatypes
str(advert)
```

```
## Classes 'data.table' and 'data.frame':   1000 obs. of  10 variables:
##  $ Daily Time Spent on Site: num  69 80.2 69.5 74.2 68.4 ...
##  $ Age                     : int  35 31 26 29 35 23 33 48 30 20 ...
##  $ Area Income             : num  61834 68442 59786 54806 73890 ...
##  $ Daily Internet Usage    : num  256 194 236 246 226 ...
##  $ Ad Topic Line           : chr  "Cloned 5thgeneration orchestration" "Monitored national standardi
##  $ City                    : chr  "Wrightburgh" "West Jodi" "Davidton" "West Terrifurt" ...
##  $ Male                    : int  0 1 0 1 0 1 0 1 1 1 ...
##  $ Country                 : chr  "Tunisia" "Nauru" "San Marino" "Italy" ...
##  $ Timestamp               : POSIXct, format: "2016-03-27 00:53:11" "2016-04-04 01:39:02" ...
##  $ Clicked on Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# The dataset has integer, number, character and datetime datatype
```

```
#Finding the total number of missing values in each column
colSums(is.na(advert))
```

```
## Daily Time Spent on Site                      Age              Area Income
##                        0                        0                        0
##     Daily Internet Usage             Ad Topic Line                     City
##                        0                        0                        0
##                     Male                  Country                Timestamp
##                        0                        0                        0
##            Clicked on Ad
##                        0
```

```
#There are no missing values in the dataset
```

```
#Finding duplicated entries within the dataset
duplicated_rows <- advert[duplicated(advert),]
#Printing out the duplicated entries
duplicated_rows
```

```
## Empty data.table (0 rows and 10 cols): Daily Time Spent on Site,Age,Area Income,Daily Internet Usage
```

```
#There are no duplicates in the datset
```

##Checking for outliers

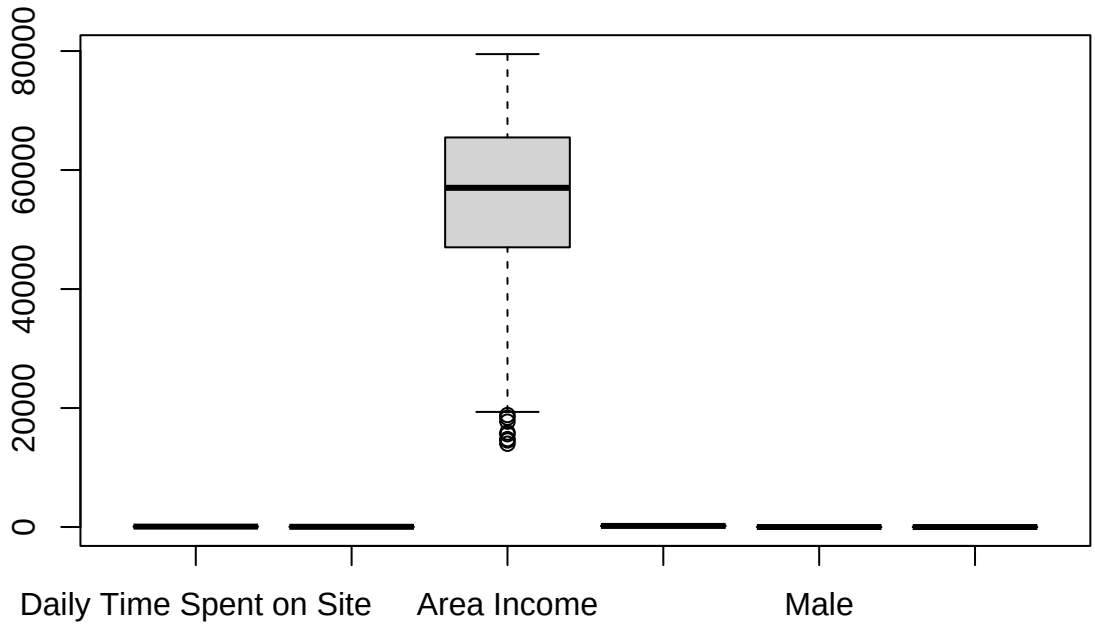
```
#Checking for outliers using boxplot
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

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

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

```
data_num2 <- select_if(advert, is.numeric)
data_num2
```

```
##       Daily Time Spent on Site Age Area Income Daily Internet Usage Male
##    1:                    68.95  35    61833.90                256.09    0
##    2:                    80.23  31    68441.85                193.77    1
##    3:                    69.47  26    59785.94                236.50    0
##    4:                    74.15  29    54806.18                245.89    1
##    5:                    68.37  35    73889.99                225.58    0
##   ---
##  996:                    72.97  30    71384.57                208.58    1
##  997:                    51.30  45    67782.17                134.42    1
##  998:                    51.63  51    42415.72                120.37    1
##  999:                    55.55  19    41920.79                187.95    0
## 1000:                    45.01  26    29875.80                178.35    0
##       Clicked on Ad
##    1:             0
##    2:             0
##    3:             0
##    4:             0
##    5:             0
##   ---
##  996:             1
##  997:             1
##  998:             1
##  999:             0
## 1000:             1
```

```
boxplot(data_num2)
```




```
#We notice that outliers are only available in the Area income column, but since they represent
#income different areas, we fail to drop them
```

##Exploratory Data Analysis ##Univariate Analysis #Measures of Central Tendency

```
#Finding the mean of the numerical columns
advert_mean1 <- mean(advert$`Daily Time Spent on Site`)
advert_mean2 <- mean(advert$Age)
advert_mean3 <- mean(advert$`Area Income`)
advert_mean4 <- mean(advert$`Daily Internet Usage`)
#Printing out the results for daily time spent on site
advert_mean1
```

```
## [1] 65.0002
```

```
#Printing results for Age
advert_mean2
```

```
## [1] 36.009
```

```r
#Printing the results for Area income
advert_mean3
```

```
## [1] 55000
```

```r
#Printing results for daily internet usage
advert_mean4
```

```
## [1] 180.0001
```

```r
#
```

```r
##Median
#Finding the Median of the numerical columns
advert_median1 <- median(advert$`Daily Time Spent on Site`)
advert_median2 <- median(advert$Age)
advert_median3 <- median(advert$`Area Income`)
advert_median4 <- median(advert$`Daily Internet Usage`)
#Printing out the results for daily time spent on site
advert_median1
```

```
## [1] 68.215
```

```r
#Printing results for Age
advert_median2
```

```
## [1] 35
```

```r
#Printing the results for Area income
advert_median3
```

```
## [1] 57012.3
```

```r
#Printing results for daily internet usage
advert_median4
```

```
## [1] 183.13
```

##Mode

```r
#Creating a function for finding mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
#Calculating the mode of each column
advert_mode1 <- getmode(advert$`Daily Time Spent on Site`)
advert_mode2 <- getmode(advert$Age)
advert_mode3 <- getmode(advert$`Area Income`)
advert_mode4 <- getmode(advert$`Daily Internet Usage`)

#Printing out the results for daily time spent on site
advert_mode1
```

```
## [1] 62.26
```

```
#Printing results for Age
advert_mode2
```

```
## [1] 31
```

```
#Printing the results for Area income
advert_mode3
```

```
## [1] 61833.9
```

```
#Printing results for daily internet usage
advert_mode4
```

```
## [1] 167.22
```

##Measures of Dispersion #Maximum values in each numerical column

```
#Finding the maximum values in each cloumn
advert_max1 <- max(advert$`Daily Time Spent on Site`)
advert_max2 <- max(advert$Age)
advert_max3 <- max(advert$`Area Income`)
advert_max4 <- max(advert$`Daily Internet Usage`)

#Printing out the results for daily time spent on site
advert_max1
```

```
## [1] 91.43
```

```
#Printing results for Age
advert_max2
```

```
## [1] 61
```

```
#Printing the results for Area income
advert_max3
```

```
## [1] 79484.8
```

```
#Printing results for daily internet usage
advert_max4
```

```
## [1] 269.96
```

#Minimum values in the numerical columns

```r
#Finding the minimum values in each column
advert_min1 <- min(advert$`Daily Time Spent on Site`)
advert_min2 <- min(advert$Age)
advert_min3 <- min(advert$`Area Income`)
advert_min4 <- min(advert$`Daily Internet Usage`)

#Printing out the results for daily time spent on site
advert_min1
```

```
## [1] 32.6
```

```r
#Printing results for Age
advert_min2
```

```
## [1] 19
```

```r
#Printing the results for Area income
advert_min3
```

```
## [1] 13996.5
```

```r
#Printing results for daily internet usage
advert_min4
```

```
## [1] 104.78
```

##Quantiles

```r
#Finding the quantiles in each cloumn
advert_quan1 <- quantile(advert$`Daily Time Spent on Site`)
advert_quan2 <- quantile(advert$Age)
advert_quan3 <- quantile(advert$`Area Income`)
advert_quan4 <- quantile(advert$`Daily Internet Usage`)
#Printing out the results for daily time spent on site
advert_quan1
```

```
##      0%     25%     50%     75%    100%
## 32.6000 51.3600 68.2150 78.5475 91.4300
```

```r
#Printing results for Age
advert_quan2
```

```
##   0%  25%  50%  75% 100%
##   19   29   35   42   61
```

```r
#Printing the results for Area income
advert_quan3
```

```
##       0%      25%      50%      75%     100%
## 13996.50 47031.80 57012.30 65470.64 79484.80
```

```r
#Printing results for daily internet usage
advert_quan4
```

```
##       0%      25%      50%      75%     100%
## 104.7800 138.8300 183.1300 218.7925 269.9600
```

##Variance

```r
#Finding the variance in each cloumn
advert_var1 <- var(advert$`Daily Time Spent on Site`)
advert_var2 <- var(advert$Age)
advert_var3 <- var(advert$`Area Income`)
advert_var4 <- var(advert$`Daily Internet Usage`)
#Printing out the results for daily time spent on site
advert_var1
```

```
## [1] 251.3371
```

```r
#Printing results for Age
advert_var2
```

```
## [1] 77.18611
```

```r
#Printing the results for Area income
advert_var3
```

```
## [1] 179952406
```

```r
#Printing results for daily internet usage
advert_var4
```

```
## [1] 1927.415
```

##Standard Deviation

```r
#Finding the standard deviation in each cloumn
advert_sd1 <- sd(advert$`Daily Time Spent on Site`)
advert_sd2 <- sd(advert$Age)
advert_sd3 <- sd(advert$`Area Income`)
advert_sd4 <- sd(advert$`Daily Internet Usage`)
#Printing out the results for daily time spent on site
advert_sd1
```

```
## [1] 15.85361
```

```r
#Printing results for Age
advert_sd2
```

```
## [1] 8.785562
```

```
#Printing the results for Area income
advert_sd3
```

```
## [1] 13414.63
```

```
#Printing results for daily internet usage
advert_sd4
```

```
## [1] 43.90234
```

##Skeweness

```
#importing the necessary packages
library(moments)
#Finding the skewness in each cloumn
advert_sk1 <- skewness(advert$`Daily Time Spent on Site`)
advert_sk2 <- skewness(advert$Age)
advert_sk3 <- skewness(advert$`Area Income`)
advert_sk4 <- skewness(advert$`Daily Internet Usage`)

#Printing out the results for daily time spent on site
#negative value which interprets that majority of the data are greater than the mean which
#can also be interpreted that most data are concetrated on the right side of the tail.
advert_sk1
```

```
## [1] -0.3712026
```

```
#Printing results for Age
#From the results below we can note that Age column has a positive skewness meaning majority of
#the data are less than the mean
advert_sk2
```

```
## [1] 0.4784227
```

```
#Printing the results for Area income
#The negative value which interprets that majority of the data are greater than the
#mean which can also be interpreted that most data are concetrated on the right side of the tail.
advert_sk3
```

```
## [1] -0.6493967
```

```
#Printing results for daily internet usage
#Daily internet usage column has a value close to 0 meaning its data is normally distributed.
advert_sk4
```

```
## [1] -0.03348703
```

##Kurtosis

```r
#Finding the skewness in each cloumn
advert_kr1 <- kurtosis(advert$`Daily Time Spent on Site`)
advert_kr2 <- kurtosis(advert$Age)
advert_kr3 <- kurtosis(advert$`Area Income`)
advert_kr4 <- kurtosis(advert$`Daily Internet Usage`)
#Printing out the results for daily time spent on site
advert_kr1
```

```
## [1] 1.903942
```

```r
#Printing results for Age
advert_kr2
```

```
## [1] 2.595482
```

```r
#Printing the results for Area income
advert_kr3
```

```
## [1] 2.894694
```

```r
#Printing results for daily internet usage
advert_kr4
```

```
## [1] 1.727701
```

```r
#All the kurtosis values are less than 3 which is called Platykurtic.
```

Categorical Columns

```r
#ad.topic.line

uniq_topic <- unique(advert$`Ad Topic Line`, )
length(uniq_topic)
```

```
## [1] 1000
```

```r
# There are 1000 unique topic lines meaning it would be impossible to get a good visualization.
```

```r
# city

uniq_city <- unique(advert$City, )
length(uniq_city)
```

city

```
## [1] 969
```

```
# There are 969 unique cities hence it would also be impossible to get a good visualization
```

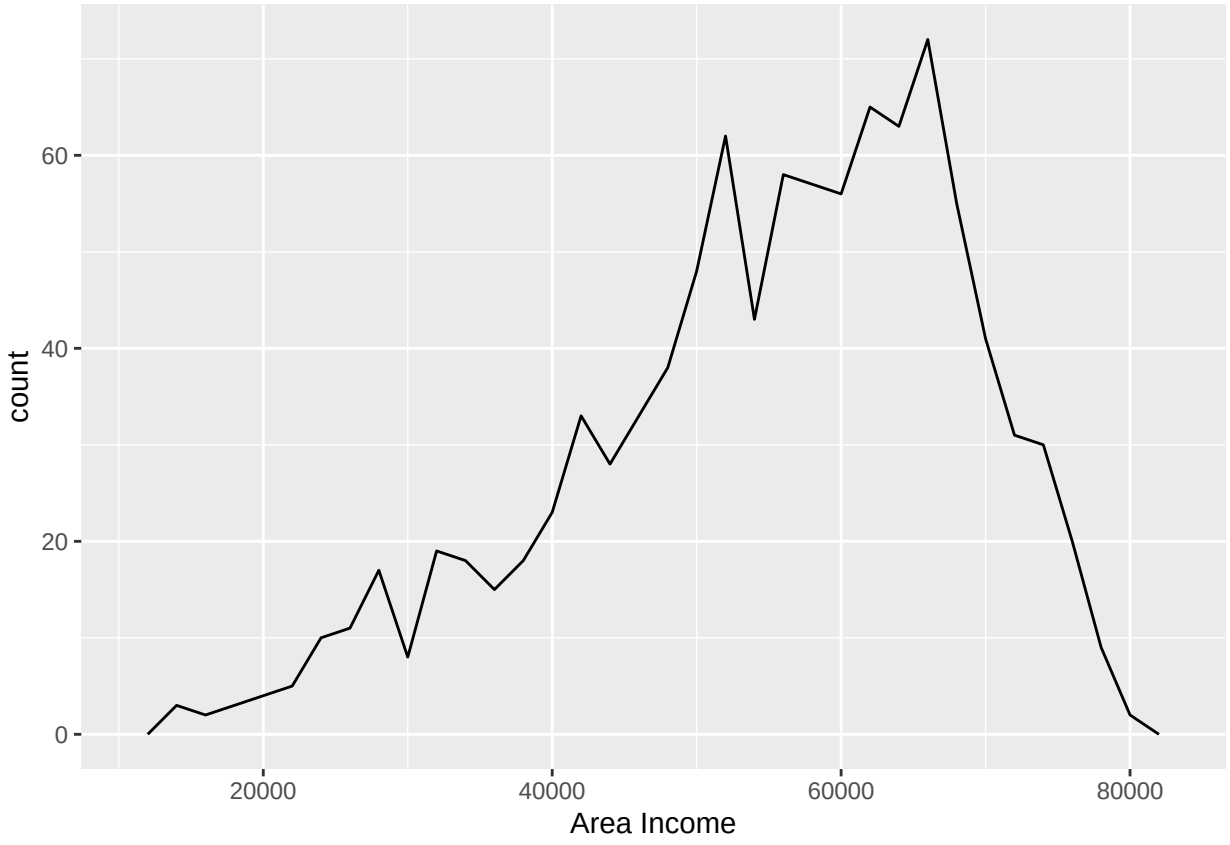
```
#country

uniq_country <- unique(advert$Country)
length(uniq_country)
```

```
## [1] 237
```

```
# There are 237 unique countries.
```

##Bivariate analysis ##Scatter plots

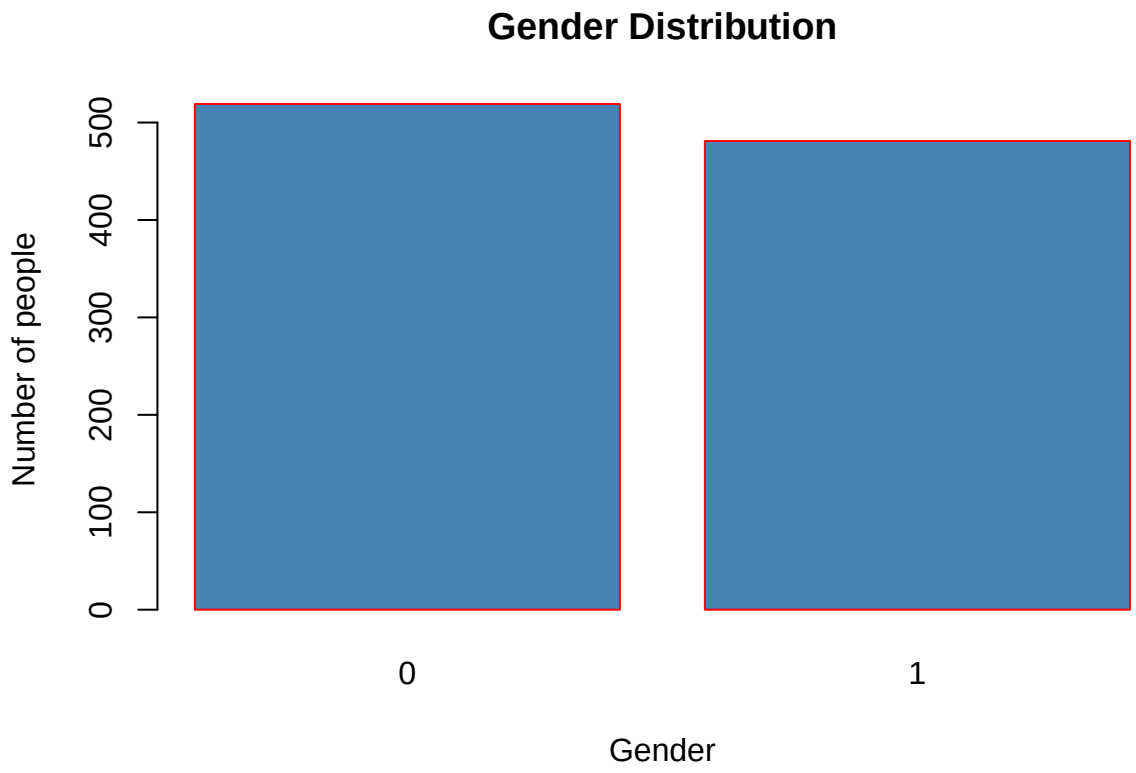
```
#Area income vs clicked on ad
library(ggplot2)
ggplot(data = advert, mapping = aes(x = `Area Income`)) +
  geom_freqpoly(mapping = aes(colour = `Clicked on Ad`), binwidth = 2000)
```




```
# Areas with higher area income clicked on the ad more
```

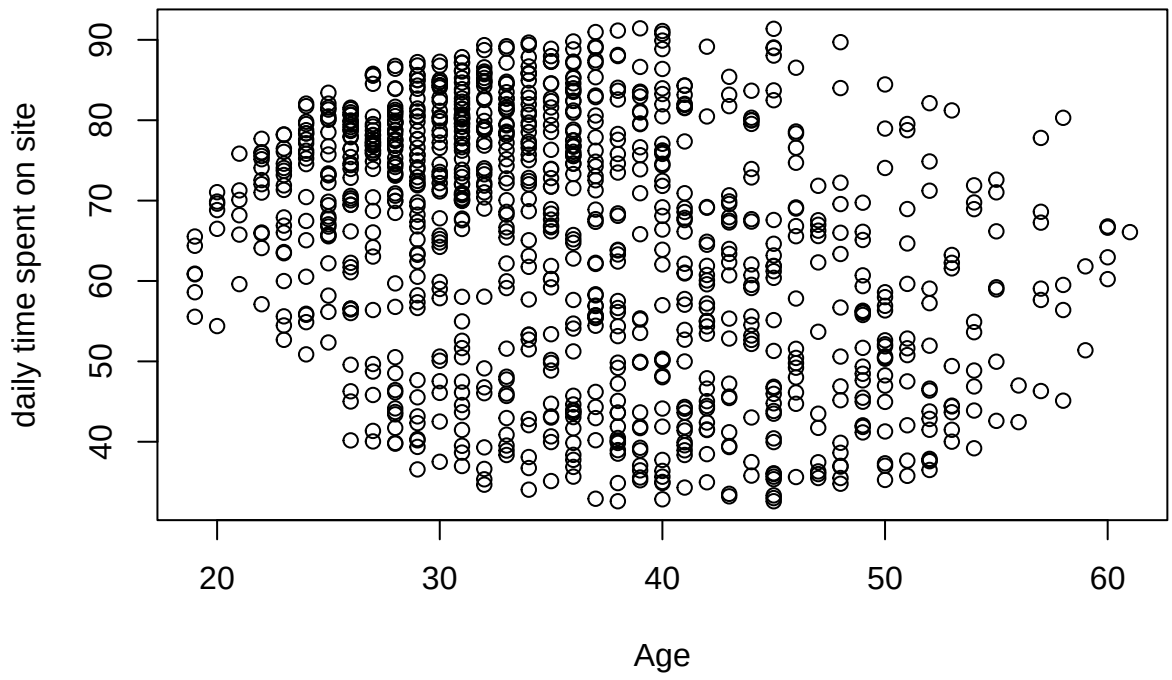
```
### Gender

male <- advert$Male
```

```
male_freq <- table(male)
barplot(male_freq, main= 'Gender Distribution', xlab="Gender",
ylab="Number of people",
border="red",
col="steelblue")
```




```
# Slightly more females than male(0 is female)
```

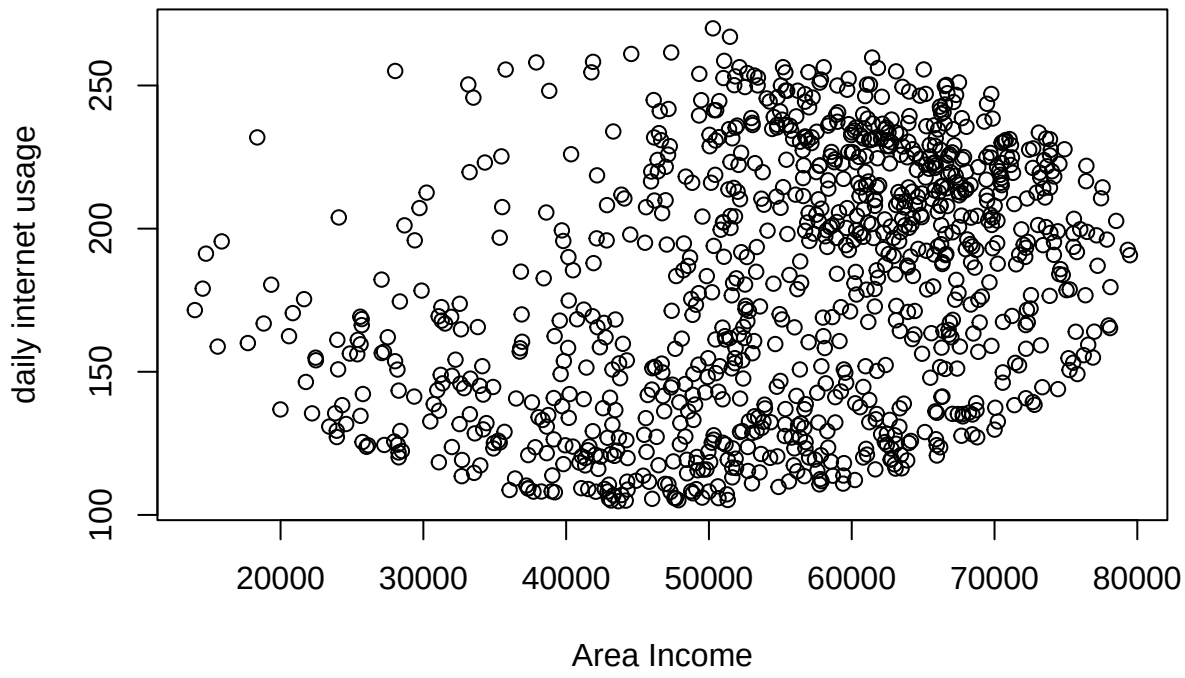
```
#Plotting a scatter plot for age and daily time spent on site
#Assigning age to age column
age <- advert$Age
#Assigning daily time to its column
daily <- advert$`Daily Time Spent on Site`
#Creating a scatter plot
plot(age, daily, xlab = "Age", ylab = "daily time spent on site" )
```




```
#The plot is scattered with no visible relationship bettween age and time spent
```

#Scatter plot between area income and daily internet usage

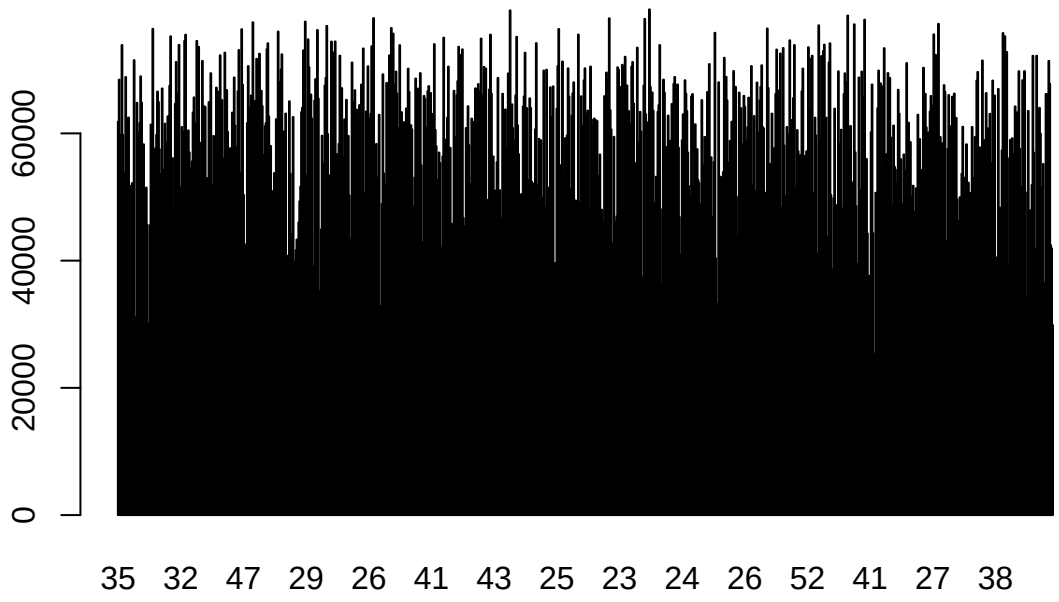
```
#Assigning each column its respective name
area <- advert$`Area Income`
usage <- advert$`Daily Internet Usage`
#Plotting the scatter plot
plot(area, usage, xlab = "Area Income", ylab = "daily internet usage" )
```




```
#As the income increases the number of people using internet increases
```

##Barplots

```
#Assigning values to column names
income <- advert$`Area Income`
age <- advert$Age
#Plotting the bargraph
barplot(income,                                    # Add labels to barplot
        names.arg = age)
```



##Covariance among variables

```
#Printing out covariances
cov(age, income)
```

```
## [1] -21520.93
```

```
cov(area, usage)
```

```
## [1] 198762.5
```

```
cov(age, area)
```

```
## [1] -21520.93
```

```
cov(income, usage )
```

```
## [1] 198762.5
```

##Correlation

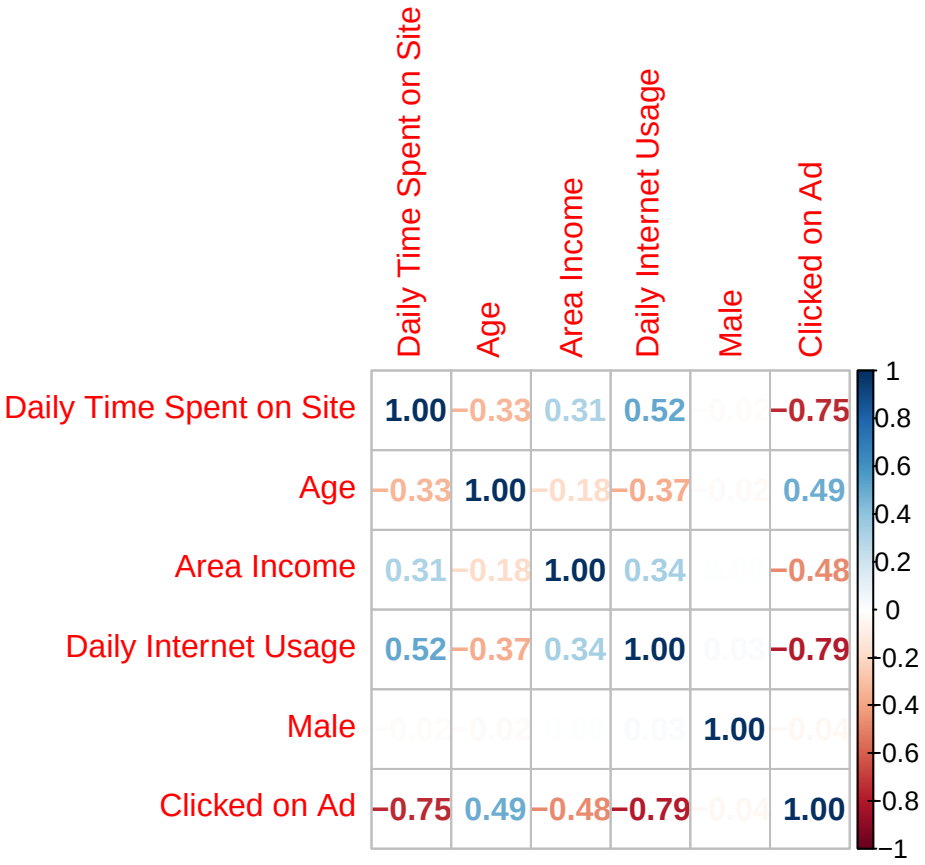
```r
#Finding the correlation of the numerical columns
# Identify numeric columns
library("dplyr")
# Subset numeric columns with dplyr
data_num3 <- select_if(advert, is.numeric)
data_num3
```

```
##       Daily Time Spent on Site Age Area Income Daily Internet Usage Male
##    1:                    68.95  35    61833.90                256.09    0
##    2:                    80.23  31    68441.85                193.77    1
##    3:                    69.47  26    59785.94                236.50    0
##    4:                    74.15  29    54806.18                245.89    1
##    5:                    68.37  35    73889.99                225.58    0
##   ---
##  996:                    72.97  30    71384.57                208.58    1
##  997:                    51.30  45    67782.17                134.42    1
##  998:                    51.63  51    42415.72                120.37    1
##  999:                    55.55  19    41920.79                187.95    0
## 1000:                    45.01  26    29875.80                178.35    0
##       Clicked on Ad
##    1:             0
##    2:             0
##    3:             0
##    4:             0
##    5:             0
##   ---
##  996:             1
##  997:             1
##  998:             1
##  999:             0
## 1000:             1
```

```r
# computing correlation matrix
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
#Assigning m to the correlation
# correlation matrix
M<-cor(data_num2)
corrplot(M, method="number")
```




```
#From the correlation plot
#1.  There is a moderate positive correlation of 0.52 between Daily internet usage and
#Daily time spent on site.
#2.  There is also a weak positive correlation of 0.31 between area income and daily time
#spent on site
#3.  There is a high negative correlation of -0.75 between clicked on ad and daily time
#spent on site
#4.  There is a high negative correlation of -0.79 between clicked on ad and daily internet
#usage
```

##Conclusion

Below are some of the conclusions we have:

1. Most of the individuals in the site are of the average age of 36.
2. Individuals between the age of 30 - 50 spend the most time on the site.
3. Individuals at the age of 31 are the most in the site.
4. Areas with higher area income clicked on the ad more

##Recommendations

Our recommendations are:

1. More advertisement should cater to individuals in their 30s but extend to the age bracket (30-50).