R-Fundamentals

Vivian Bwana

2022-05-29

Defining the Question

To identify which individuals are most likely to click on her ads.

Metric for success

To be able to identify who is likely to click on the ads

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.

Experimental Design

1. Loading dataset into R
2. Cleaning the dataset
3. Perform EDA
4. Perform Univariate and Bivariate Analysis
5. Provide conclusions and recommendations

Loading the dataset

```
#library(knitr)
#setwd("C:/Users/Desktop/MORINGA_CORE/R PROGRAMMING/R WEEK 1/R WEEK ONE EXERCISES")      # Change worki


# Properly import data
library(data.table)
advert <- fread("http://bit.ly/IPAdvertisingData")
```

```
# Loading the dataset
# If .csv file, use this
# advertising.csv
#library(data.table)
#advert <- read.csv(file.choose())

#advert <- read.csv("advertising.csv")
```

Previewing the dataset

```
#previewing the first 6 rows of the dataset
head(advert)
```

```
##    Daily Time Spent on Site   Age Area Income Daily Internet Usage
##                       <num> <int>       <num>                <num>
## 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
##                                   <char>         <char> <int>     <char>
## 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
##                 <POSc>         <int>
## 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
```

EDA

Exploring the Dataset

```
#Checking the shape of the dataset
dim(advert)
```

```
## [1] 1000   10
```

```
#The dataset has 1000 rows and 10 columns
```

```
#Finding the datatypes of the dataset
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>
```

Data Cleaning

Editting column names

```
# assigning new names to the columns of the data frame
colnames(advert) <- c('Daily.Time.Spent.on.Site','Age','Area.Income','Daily.Internet.Usage','Ad.Topic.L

# printing new data frame
print("New data frame : ")
```

```
## [1] "New data frame : "
```

```
print(advert)
```

```
##       Daily.Time.Spent.on.Site   Age Area.Income Daily.Internet.Usage
##                          <num> <int>       <num>                <num>
##    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
##   ---
##  996:                    72.97    30    71384.57               208.58
##  997:                    51.30    45    67782.17               134.42
##  998:                    51.63    51    42415.72               120.37
##  999:                    55.55    19    41920.79               187.95
## 1000:                    45.01    26    29875.80               178.35
##                            Ad.Topic.Line          City  Male
##                                   <char>        <char> <int>
##    1:    Cloned 5thgeneration orchestration    Wrightburgh     0
##    2:    Monitored national standardization      West Jodi     1
##    3:      Organic bottom-line service-desk       Davidton     0
##    4: Triple-buffered reciprocal time-frame West Terrifurt     1
##    5:         Robust logistical utilization   South Manuel     0
##   ---
##  996:         Fundamental modular algorithm      Duffystad     1
##  997:       Grass-roots cohesive monitoring    New Darlene     1
##  998:          Expanded intangible solution  South Jessica     1
##  999:  Proactive bandwidth-monitored policy    West Steven     0
## 1000:       Virtual 5thgeneration emulation    Ronniemouth     0
##                       Country           Timestamp Clicked.on.Ad
##                        <char>              <POSc>         <int>
##    1:                 Tunisia 2016-03-27 00:53:11             0
##    2:                   Nauru 2016-04-04 01:39:02             0
##    3:              San Marino 2016-03-13 20:35:42             0
##    4:                   Italy 2016-01-10 02:31:19             0
##    5:                 Iceland 2016-06-03 03:36:18             0
##   ---
##  996:                 Lebanon 2016-02-11 21:49:00             1
##  997: Bosnia and Herzegovina 2016-04-22 02:07:01             1
##  998:                Mongolia 2016-02-01 17:24:57             1
##  999:               Guatemala 2016-03-24 02:35:54             0
## 1000:                  Brazil 2016-06-03 21:43:21             1
```

Missing Values

```
#Checking for the sum of Missing values
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 this dataset
```

Duplicates

```
#Checking for duplicates in the dataset
advert.duplicates <- advert[duplicated(advert),]

#printing duplicated rows
advert.duplicates
```

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

```
#There are no duplicated rows in the dataset
```

Outliers

```
#checking for dataframe class
class(advert)
```

```
## [1] "data.table" "data.frame"
```

```
#Exctracting numeric columns to analyse for outliers
num.cols <- unlist(lapply(advert, is.numeric))

#printing numeric columns
num.cols
```

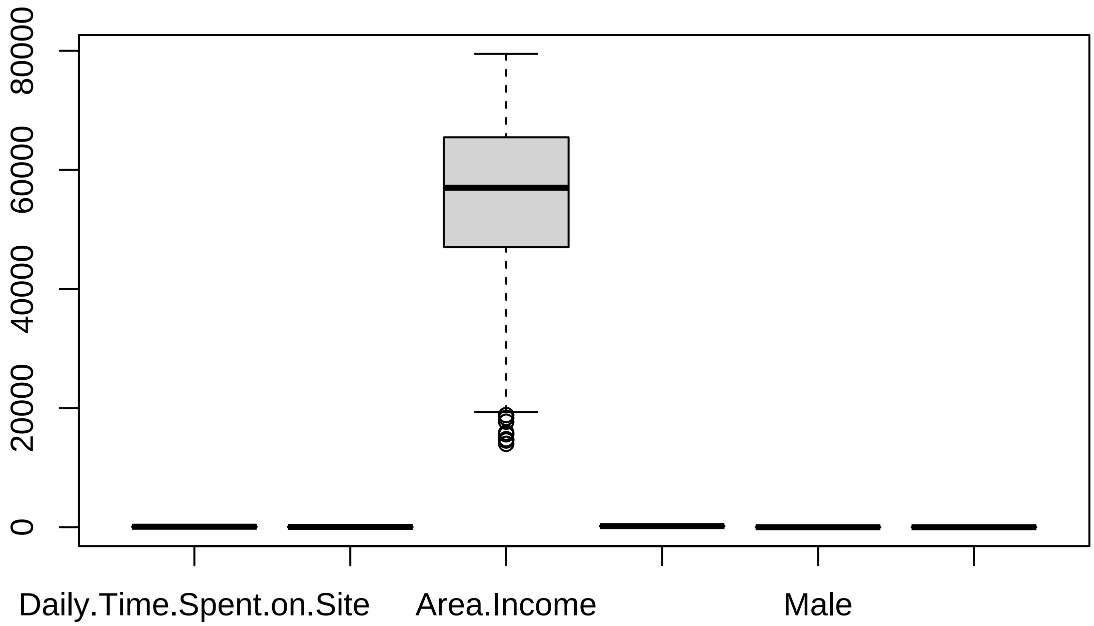
```
## Daily.Time.Spent.on.Site                      Age              Area.Income 
##                     TRUE                     TRUE                     TRUE 
##     Daily.Internet.Usage            Ad.Topic.Line                     City 
##                     TRUE                    FALSE       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

```
#creating a dataframe with numeric columns only so as to plot a boxplot
advert.numeric <-advert[, ..num.cols]

#checking the data types, previewing
str(advert.numeric)
```

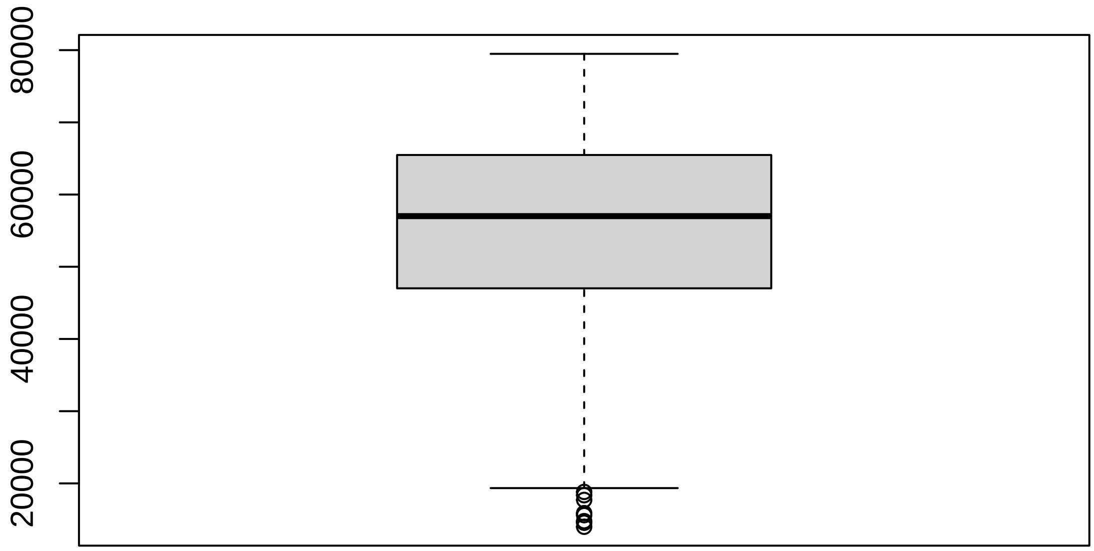
```
## Classes ’data.table’ and ’data.frame’:   1000 obs. of  6 variables:
##  $ Daily.Time.Spent.on.Site: num  69 80.2 69.5 74.2 68.4 ...
##  $ Age                     : int  35 31 26 29 35 23 33 48 30 20 ...
##  $ Area.Income             : num  61834 68442 59786 54806 73890 ...
##  $ Daily.Internet.Usage    : num  256 194 236 246 226 ...
##  $ Male                    : int  0 1 0 1 0 1 0 1 1 1 ...
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
#Plotting a boxplot to check for outliers
#library(data.table)
boxplot(advert.numeric)
```



```
#there are outliers in Area.income column
```

```
#Plotting a boxplot to check for outliers in Area.Income column
#for(y in 1:ncol(advert.numeric)){
 # if (is.na(advert.numeric[1,y])) advert.numeric[1,y] = 0
  #}
#+
boxplot(advert.numeric$Area.Income)#,ylim=c(0,300), main = 'Boxplot of Area Income')$out
```



```
#there are some records appearing as outliers in the lower quartile of the  Area.Income column
#These will be removed before we begin analysis
```

```
#Removing outliers in the lower quartile of the Area.Income
Q1 <- quantile(advert$Area.Income, .25)
Q3 <- quantile(advert$Area.Income, .75)
IQR <- IQR(advert$Area.Income)

#Keeping values above 1.5*IQR of Q1

no.outliers <- subset(advert, advert$Area.Income > (Q1 - 1.5*IQR)) #& advert$Area.Income < (Q3 + 1.5*IQ
dim(no.outliers)
```

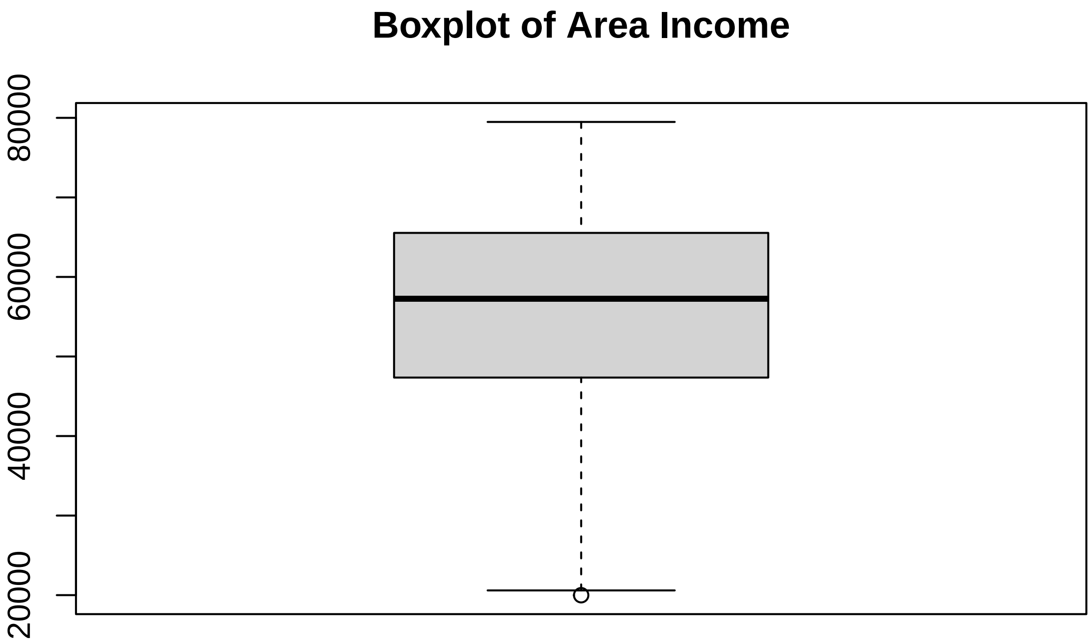
```
## [1] 991  10
```

```
dim(advert)
```

```
## [1] 1000   10
```

```
#9 records were dropped
```

```
#Plotting a boxplot to check if outliers in Area.Income column have been dropped
boxplot(no.outliers$Area.Income, main = 'Boxplot of Area Income')
```



```
#the boxplot indicates that the outliers have been removed
```

Univariate Analysis

```
#previewing the new dataset without outliers
head(no.outliers)
```

```
##    Daily.Time.Spent.on.Site   Age  Area.Income Daily.Internet.Usage
##                       <num> <int>        <num>                <num>
## 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
##                                   <char>         <char> <int>     <char>
## 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
##                 <POSc>         <int>
## 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 summary statistics
summary(no.outliers)
```

```
##  Daily.Time.Spent.on.Site      Age         Area.Income    Daily.Internet.Usage
##  Min.   :32.60            Min.   :19.00   Min.   :19992   Min.   :104.8
##  1st Qu.:51.34            1st Qu.:29.00   1st Qu.:47348   1st Qu.:138.6
##  Median :68.41            Median :35.00   Median :57260   Median :183.4
##  Mean   :65.06            Mean   :35.99   Mean   :55349   Mean   :180.0
##  3rd Qu.:78.59            3rd Qu.:42.00   3rd Qu.:65538   3rd Qu.:218.9
##  Max.   :91.43            Max.   :61.00   Max.   :79485   Max.   :270.0
##  Ad.Topic.Line          City                Male          Country
##  Length:991         Length:991         Min.   :0.0000   Length:991
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.4793
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
##    Timestamp                      Clicked.on.Ad
##  Min.   :2016-01-01 02:52:10.00   Min.   :0.0000
##  1st Qu.:2016-02-17 22:51:14.50   1st Qu.:0.0000
##  Median :2016-04-07 03:56:16.00   Median :0.0000
##  Mean   :2016-04-10 02:20:21.53   Mean   :0.4955
##  3rd Qu.:2016-05-31 01:37:57.50   3rd Qu.:1.0000
##  Max.   :2016-07-24 00:22:16.00   Max.   :1.0000
```

```
#Extracting a numeric subset from the no outliers dataset
no.out.num.cols <-unlist(lapply(no.outliers, is.numeric))
#Exctracting numeric columns to analyse for outliers
#num.cols <- unlist(lapply(advert, is.numeric))

#printing numeric columns
no.out.num.cols
```

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

```
#creating a dataframe with numeric columns only so as to plot a boxplot
no.outliers.numeric <-no.outliers[, ..no.out.num.cols]

#previewing
head(no.outliers.numeric)
```

```
##    Daily.Time.Spent.on.Site   Age Area.Income Daily.Internet.Usage  Male
##                       <num> <int>       <num>                <num> <int>
## 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
## 6:                    59.99    23    59761.56               226.74     1
##    Clicked.on.Ad
##            <int>
## 1:             0
## 2:             0
## 3:             0
## 4:             0
## 5:             0
## 6:             0
```

```
#checking the data types, previewing
#str(no.outliers.numeric)
```

Measures of Central Tendency

i) Mean

```
#means of all numeric columns in the dataset
#this has been exctracted from the datset and named no.outliers.numeric
#the variable for the column means is no.out.col.means

no.out.col.means <- colMeans(data.frame(no.outliers.numeric))

# Printing out
# ---
#
no.out.col.means
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##             6.505689e+01             3.598587e+01             5.534910e+04
##     Daily.Internet.Usage                     Male            Clicked.on.Ad
##             1.799846e+02             4.793138e-01             4.954591e-01
```

The average daily time spent on site was 65.05 units. The average area income was 55,349 units. The average age of respondents was 35.98 years. The average daily internet usage was 179.98 units.

ii) Median

```
#median of all numeric columns in the dataset
#this has been exctracted from the datset and named no.outliers.numeric
#the variable for the column means is no.out.col.median
library(matrixStats)

no.out.col.median <- colMedians(as.matrix.data.frame(no.outliers.numeric))

# Printing out
# ---
#
print(no.out.col.median)
```

```
## [1]    68.41    35.00 57260.41   183.43     0.00     0.00
```

The median of the daily time spent on site was 68.41 units. The median of the area income was 57,260.41 units. The median of the ages of the respondents was 35 years. The median of the daily internet usage was 183.43 units.

iii)Mode

```
# We create the mode function that will perform our mode operation for us
# The mode will give us values that appeared the most number of times
# ---
# library(purrr)
FindMode <- function(no.outliers) {
   uniqv <- unique(no.outliers)
   uniqv[which.max(tabulate(match(no.outliers, uniqv)))]
}

# Calculating the mode using out getmode() function
# ---
#
#no.out.col.mode <- getmode(as.matrix(no.outliers.numeric))
no.out.col.mode <- data.frame(no.outliers)

# Printing out
# ---
#
apply(no.out.col.mode,2, FindMode)
```

```
##             Daily.Time.Spent.on.Site                                  Age
##                              "62.26"                                 "31"
##                          Area.Income                 Daily.Internet.Usage
##                           "61833.90"                             "167.22"
##                        Ad.Topic.Line                                 City
## "Cloned 5thgeneration orchestration"                          "Lisamouth"
##                                 Male                              Country
##                                  "0"                     "Czech Republic"
##                           Timestamp                              Clicked.on.Ad
##                "2016-03-27 00:53:11"                                       "0"
```

```
#The modes of all the variables, both categorical and numerical are as follows:
# For factors male and clicked on ad, 0 = no and 1= yes
# There were lesser male respondents
# Most respondents did not click on the adverts
```

Measures of Dispersion

We will use the numeric data-frame while calculating measures of dispersion

i)Minimum

```
# Minimum
#min <-colMins(as.matrix(no.outliers.numeric[sapply(no.outliers.numeric, is.numeric)]))

#printing
#min

sapply(no.outliers.numeric, min)
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##                    32.60                    19.00                 19991.72
##     Daily.Internet.Usage                     Male            Clicked.on.Ad
##                   104.78                     0.00                     0.00
```

The minimum of the daily time spent on site was 32.60 units. The minimum of the area income was 19,991.72 units. The minimum of the ages of the respondents was 19 years. The minimum of the daily internet usage was 104.78 units. The minimum value of whether male or not is 0. The minimum value of whether clicked on advert or not is 0.

ii)Maximum

```
#minimum, maximum, range, quantile, variance
# and standard deviation
#max <-colMaxs(as.matrix.data.frame(no.outliers.numeric[sapply(no.outliers.numeric, is.numeric)]))

# previewing
#max
#max
sapply(no.outliers.numeric, max)
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##                    91.43                    61.00                 79484.80
##     Daily.Internet.Usage                     Male            Clicked.on.Ad
##                   269.96                     1.00                     1.00
```

The maximum of the daily time spent on site was 91.43 units. The maximum of the area income was 79,484.80 units. The maximum of the ages of the respondents was 61 years. The maximum of the daily internet usage was 269.96 units. The maximum value of whether male or not is 1. The maximum value of whether clicked on advert or not is 1. ### iii) Variance

```
# Finding the variance of all the variables
# area <-sd(no.outliers.numeric$Area.Income)
sapply(no.outliers.numeric, var)
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##             2.528258e+02             7.752303e+01             1.680004e+08
##     Daily.Internet.Usage                     Male             Clicked.on.Ad
##             1.940743e+03             2.498242e-01             2.502319e-01
```

The variance of the daily time spent on site was 252.82. The variance of the area income was 168,000,385. The variance of the ages of the respondents was 77.52. The variance of the daily internet usage was 1940.74. The variance of male column is 0.2498 . The variance of whether ad was clicked or not 0.2502 .

iv) Standard Deviation

```
# Finding the standard deviation for all numeric variables
sapply(no.outliers.numeric, sd)
```

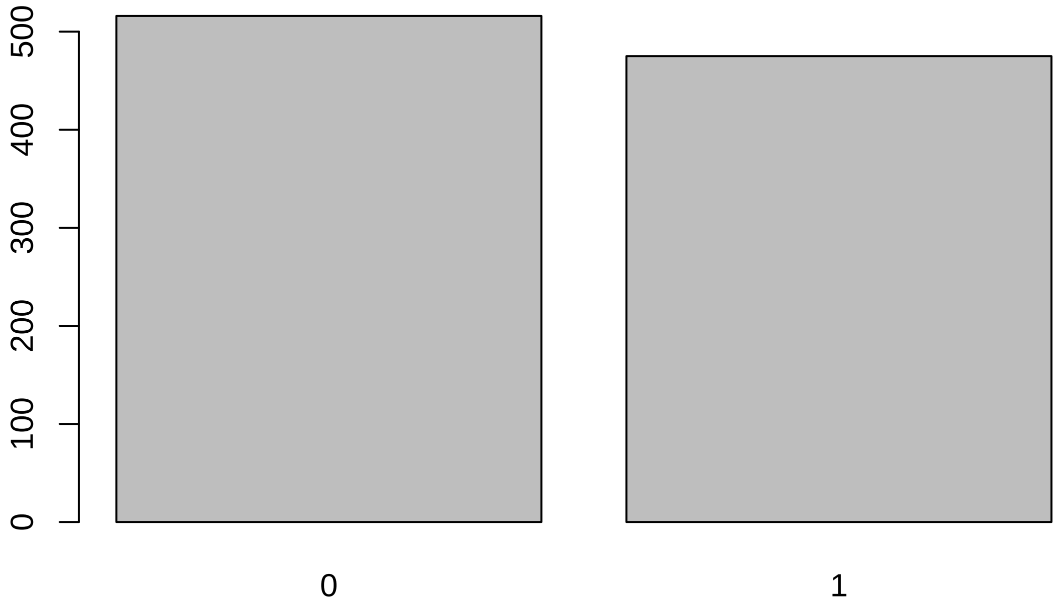
```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##             1.590050e+01             8.804716e+00             1.296150e+04
##     Daily.Internet.Usage                     Male             Clicked.on.Ad
##             4.405386e+01             4.998241e-01             5.002318e-01
```

The standard deviation of the daily time spent on site was 15.90. The standard deviation of the area income was 12,961.5. The standard deviation of the ages of the respondents was 8.80. The standard deviation of the daily internet usage was 44.05.

Univariate Graphicals

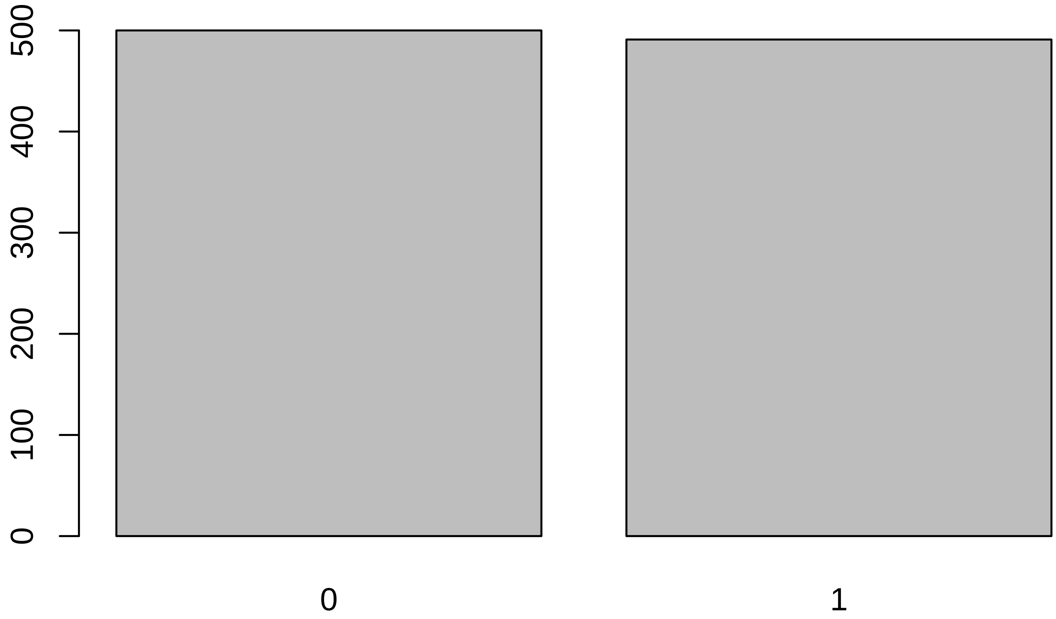
```
# Plotting a bar-graph to see the frequency of the categorical variables
# The table() function computes the frequency distribution of the categorical variables

# for the male column
barplot(table(no.outliers.numeric$Male))
```



```
# The respondents who were not males were fewer than those who were males
```

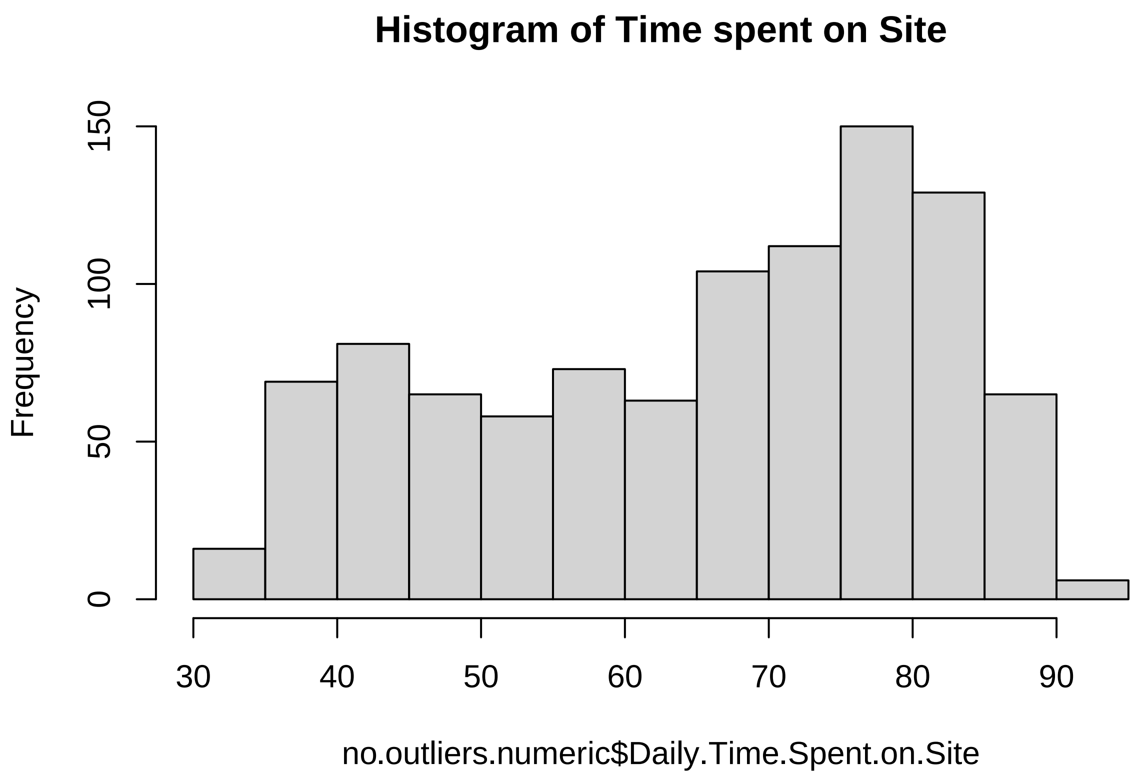
```
# for the male column
barplot(table(no.outliers.numeric$Clicked.on.Ad))
```



```
# The number of respondents who clicked and who did not click on adverts were almost the same
```

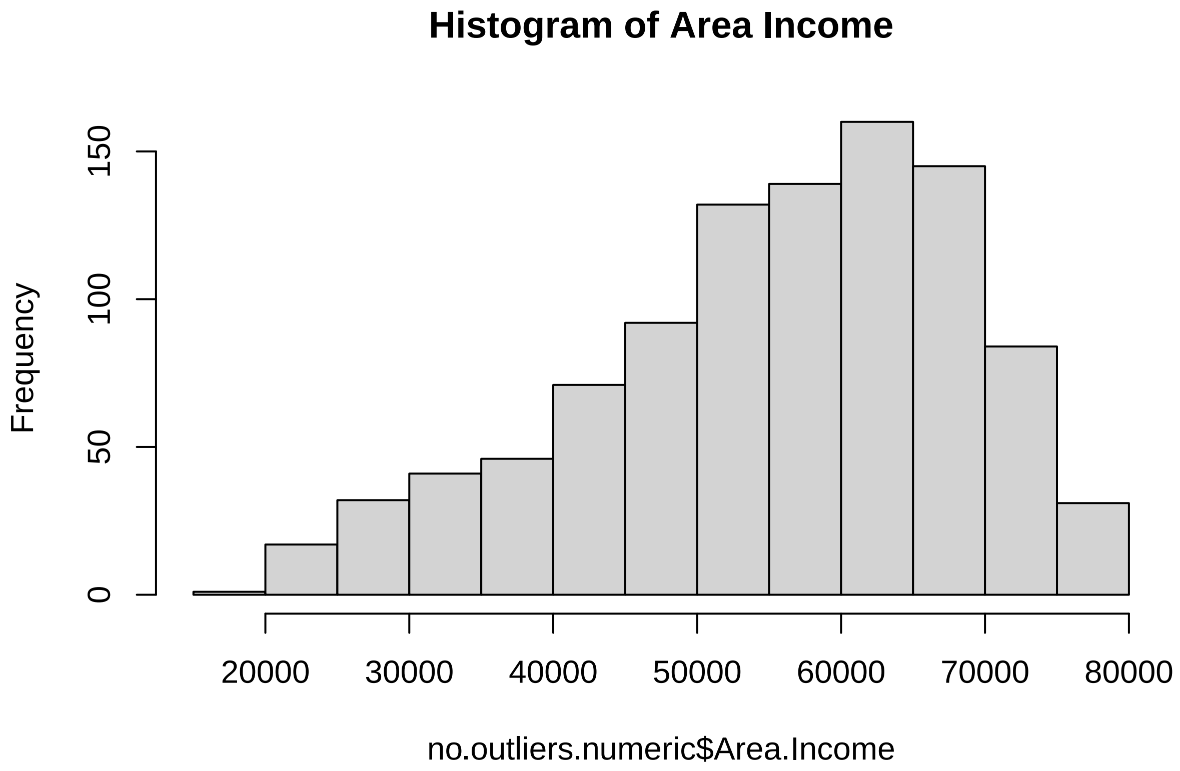
```
# Plotting histograms to show the distribution of the numerical variables

# Histogram of time spent on site
hist(no.outliers.numeric$Daily.Time.Spent.on.Site, main = "Histogram of Time spent on Site")
```



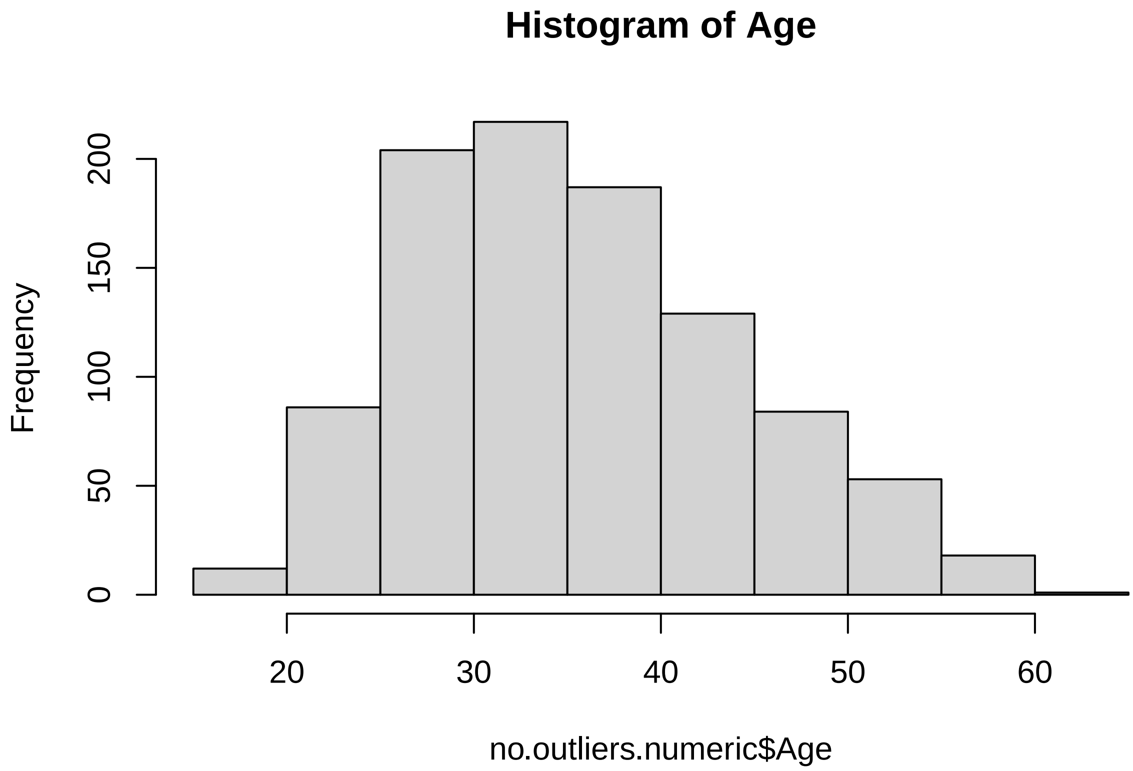
```
# The time spent on sight is not skewed, meaning the data points tend to be evenly distributed
```

```
# Histogram of area income
hist(no.outliers.numeric$ Area.Income, main = "Histogram of Area Income")
```



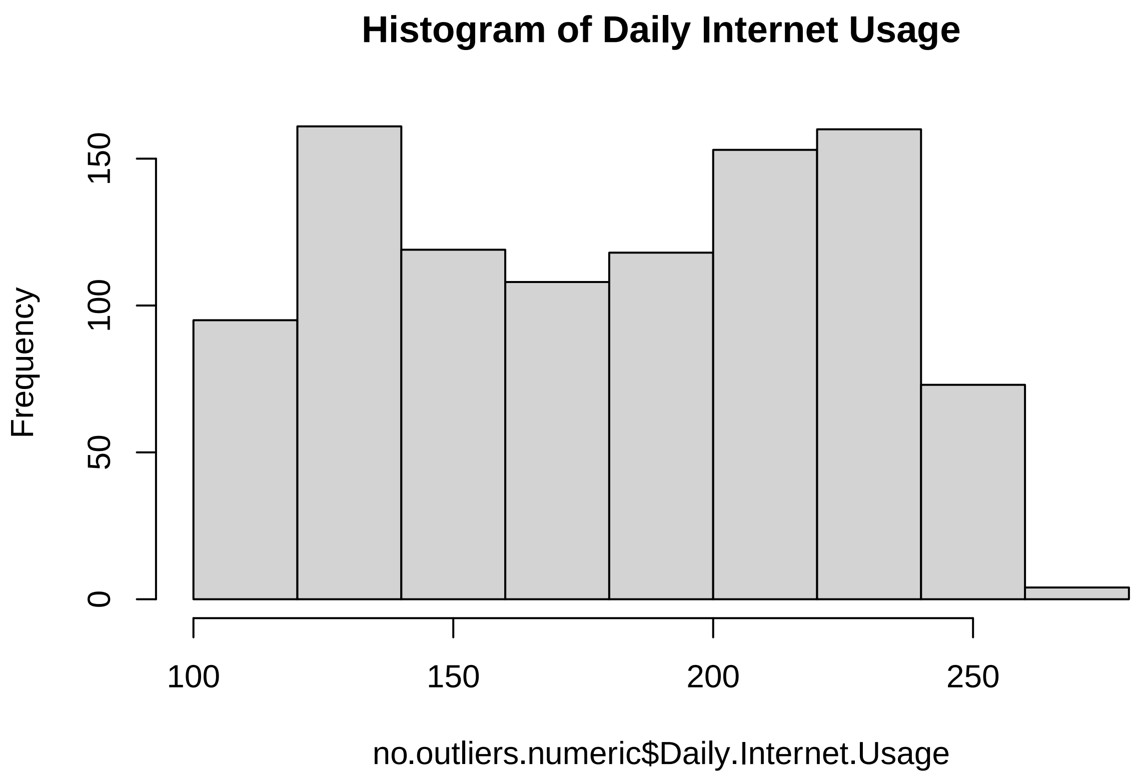
```
# The area income is left skewed, meaning the data points extend to the left of the distribution
```

```
# Histogram of Age
hist(no.outliers.numeric$Age, main = "Histogram of Age")
```



```
# The age variable is right skewed, meaning the data points extend to the right of the data points dist
```

```
# Histogram of daily internet usage
hist(no.outliers.numeric$Daily.Internet.Usage, main = "Histogram of Daily Internet Usage")
```



```
# Daily internet usage is not skewed, meaning the data points tend to be normally distributed
```

Categorical data analysis

```
# Checking the number of countries
# Checking the unique entries
countries <-unique(no.outliers$Country)

# printing the number of unique countries
# we will use the length function to do a unique value count
length(countries)
```

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

```
# there are 237 countires in the dataset
```

```
# Checking the number of cities
# Checking the unique entries
cities <-unique(no.outliers$City)

# printing the number of unique cities
# we will use the length function to do a unique value count
length(cities)
```

```
## [1] 960
```

```
#there are 960 cities in the dataset
```

Bivariate Analysis

i) Covariance

```
# previewing
head(no.outliers)
```

```
##    Daily.Time.Spent.on.Site   Age  Area.Income Daily.Internet.Usage
##                       <num> <int>        <num>                <num>
## 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
##                                     <char>         <char> <int>     <char>
## 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
##                 <POSc>         <int>
## 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
```

```
# finding the covariance of the target variable and other numerical variables
# we assign different variables for the specific columns

# Assigning Daily.Time.Spent.on.Site column to variable time.site
time.site <- no.outliers$Daily.Time.Spent.on.Site

# Assigning Age column to variable age
age <-no.outliers$Age

# Assigning Area.income column to variable area.income
area.income <-no.outliers$Area.Income

# Assigning Daily.Internet.Usage column to variable daily.internet
daily.internet <-no.outliers$Daily.Internet.Usage
```

```
# Assigning Male column to variable male
male <-no.outliers$Male

# Assigning clicked on ads column to variable clicks.target
clicks.target <-no.outliers$Clicked.on.Ad
```

```
# Finding co-variances of the numerical variables

# covariance of age and time spent on site
cov(time.site,age )
```

```
## [1] -46.59899
```

```
 # negative linear relationship between the variables
```

```
# Finding co-variances of the numerical variables

# covariance of age and time spent on site
cov(time.site,area.income )
```

```
## [1] 64600.67
```

```
# strong positive linear relationship between the variables
```

```
# Finding co-variances of the numerical variables

# covariance of age and time spent on site
cov(time.site,daily.internet)
```

```
## [1] 364.2711
```

```
#positive linear relationship between the variables
```

```
# Finding co-variances of the numerical variables

# covariance of age and time spent on site
cov(age,area.income )
```

```
## [1] -20744.22
```

```
# strong negative linear relationship between the variables
```

```
# covariance of age and time spent on site
cov(age,daily.internet )
```

```
## [1] -142.7226
```

```
# negative linear relationship between the variables
```

```
#covariance
cov(area.income,daily.internet )
```

```
## [1] 201115
```

```
# strong positive linear relationship between the variables
```

ii) Correlation

We will use the numeric dataframe

```
# correlation matrix
ad_cor <- cor(no.outliers.numeric, use="pairwise.complete.obs",method = "pearson")
round(ad_cor, 2)
```

```
##                          Daily.Time.Spent.on.Site   Age Area.Income
## Daily.Time.Spent.on.Site                     1.00 -0.33        0.31
## Age                                         -0.33  1.00       -0.18
## Area.Income                                  0.31 -0.18        1.00
## Daily.Internet.Usage                         0.52 -0.37        0.35
## Male                                        -0.02 -0.02        0.01
## Clicked.on.Ad                               -0.75  0.49       -0.47
##                          Daily.Internet.Usage  Male Clicked.on.Ad
## Daily.Time.Spent.on.Site                 0.52 -0.02         -0.75
## Age                                     -0.37 -0.02          0.49
## Area.Income                              0.35  0.01         -0.47
## Daily.Internet.Usage                     1.00  0.03         -0.79
## Male                                     0.03  1.00         -0.04
## Clicked.on.Ad                           -0.79 -0.04          1.00
```

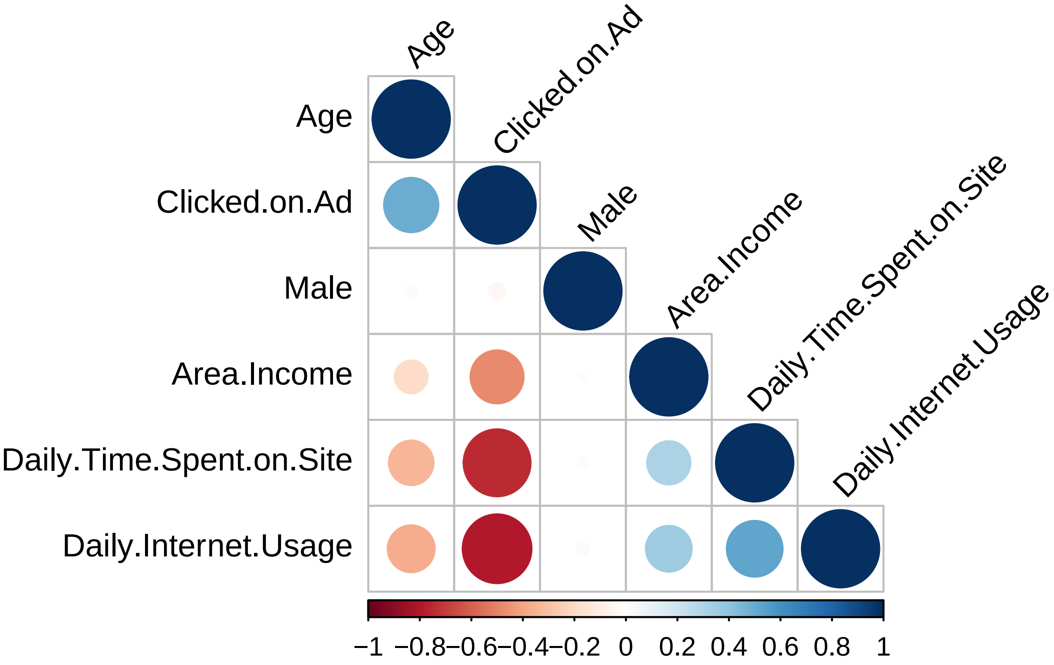
```
# gives correlation co-efficients in pairs and rounding them off to decimal places

# When the correlation the coefficient value is next to 1 it shows a positive linear relationship,
# when next to -1, it indicates that the variables are negatively linearly related
# When close to zero, it would indicate a weak linear relationship between the variables.
```

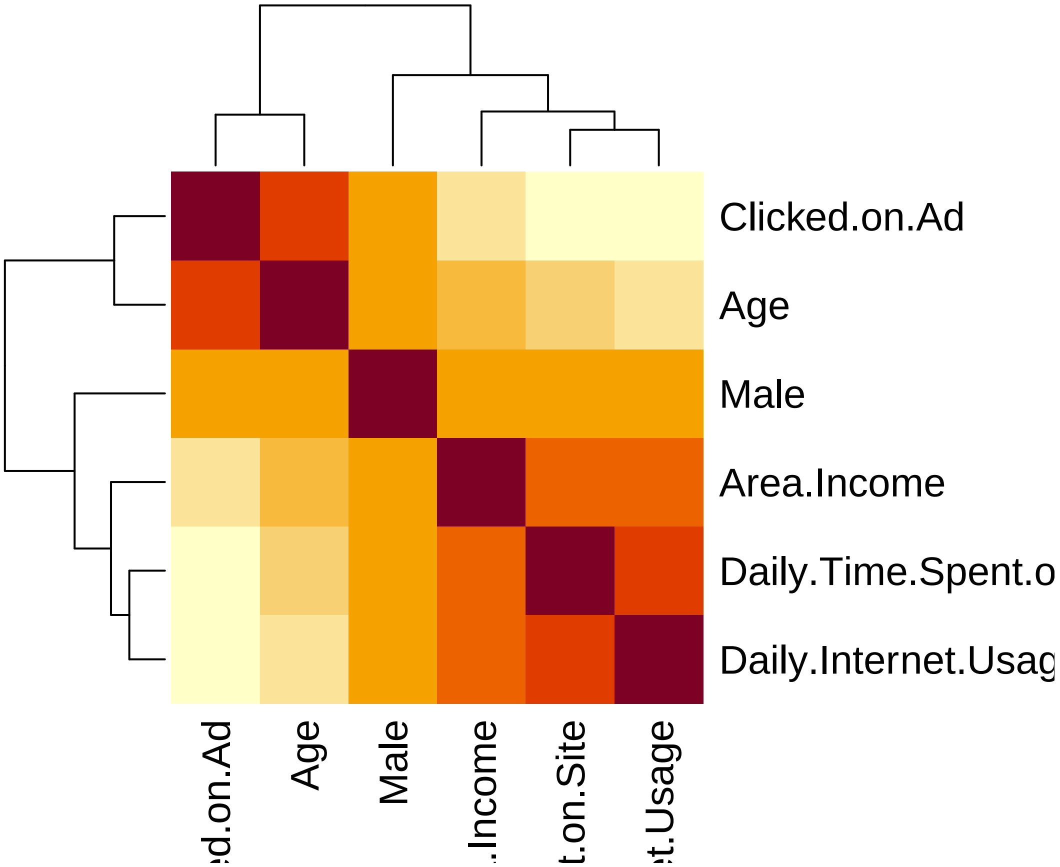
```
# Visualizing the correlation matrix
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(ad_cor, type = "lower", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

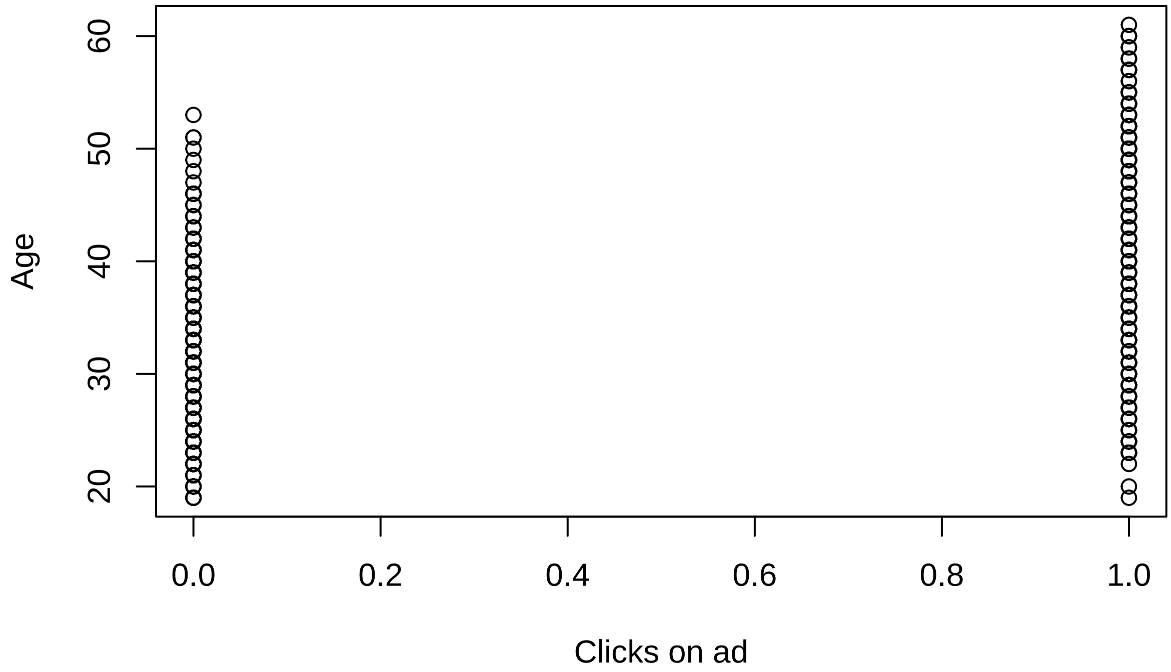


```
# Plotting a correlation Heatmap
# Get some colors
#col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(x = ad_cor, symm = TRUE)
```

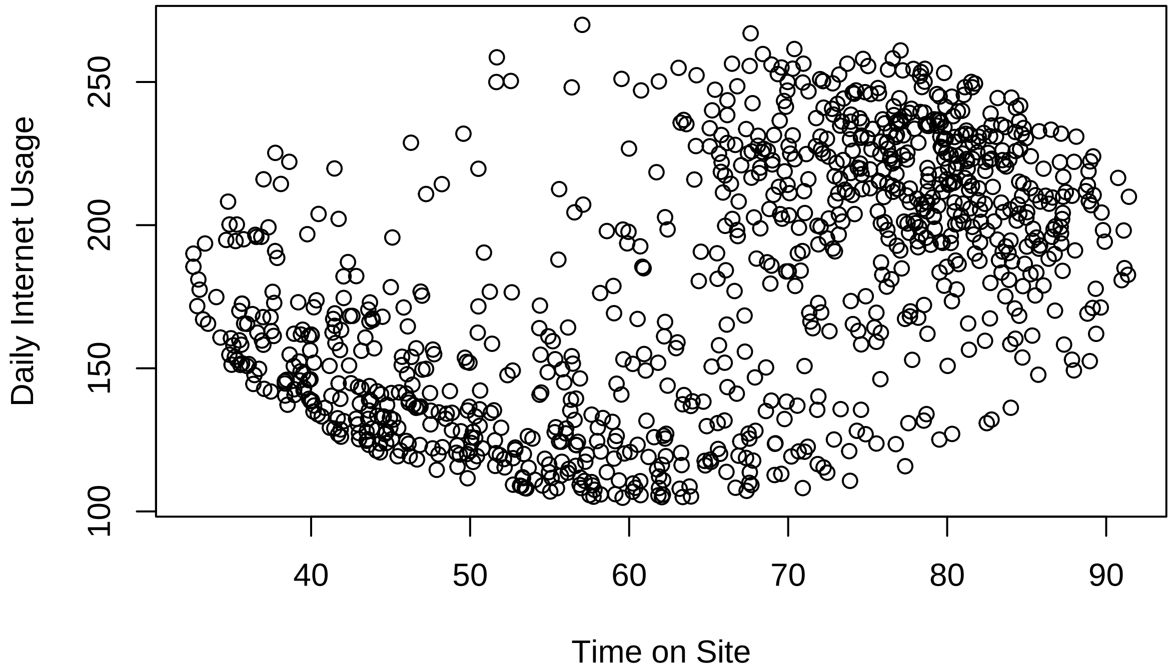


```
# Plotting scatter plots
# we will use the variables we assigned earlier
#time.site
#age
#area.income
#daily.internet
#male
#clicks.target
```

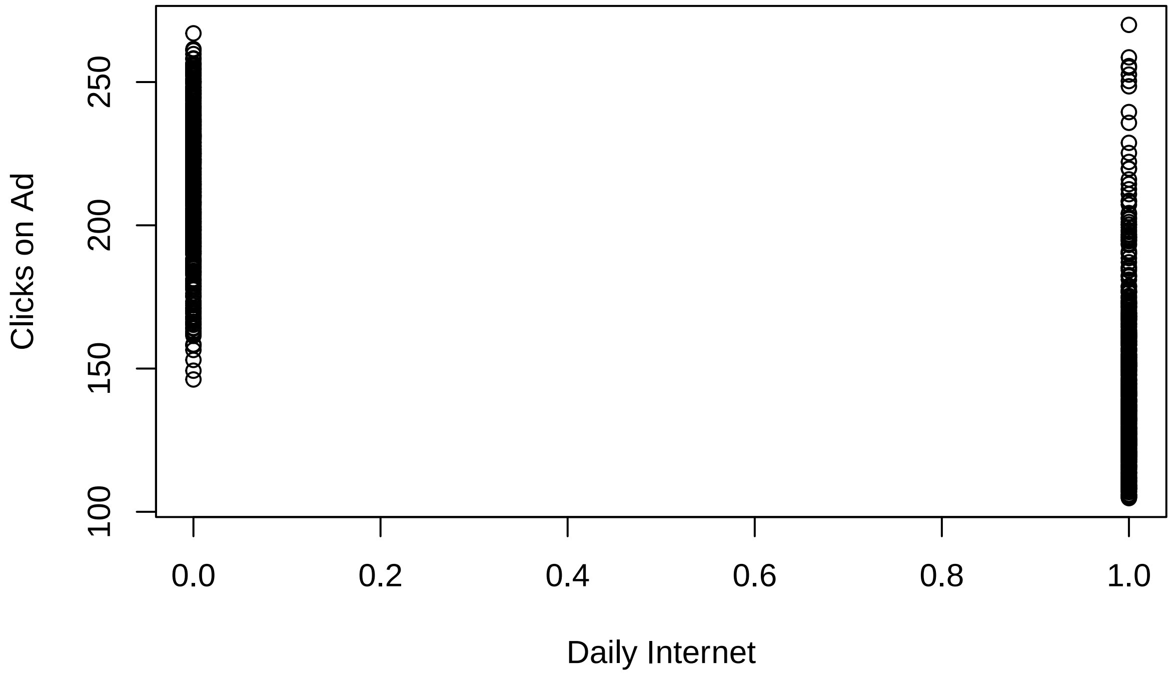
```
# plotting
plot(clicks.target, age, xlab="Clicks on ad", ylab="Age")
```



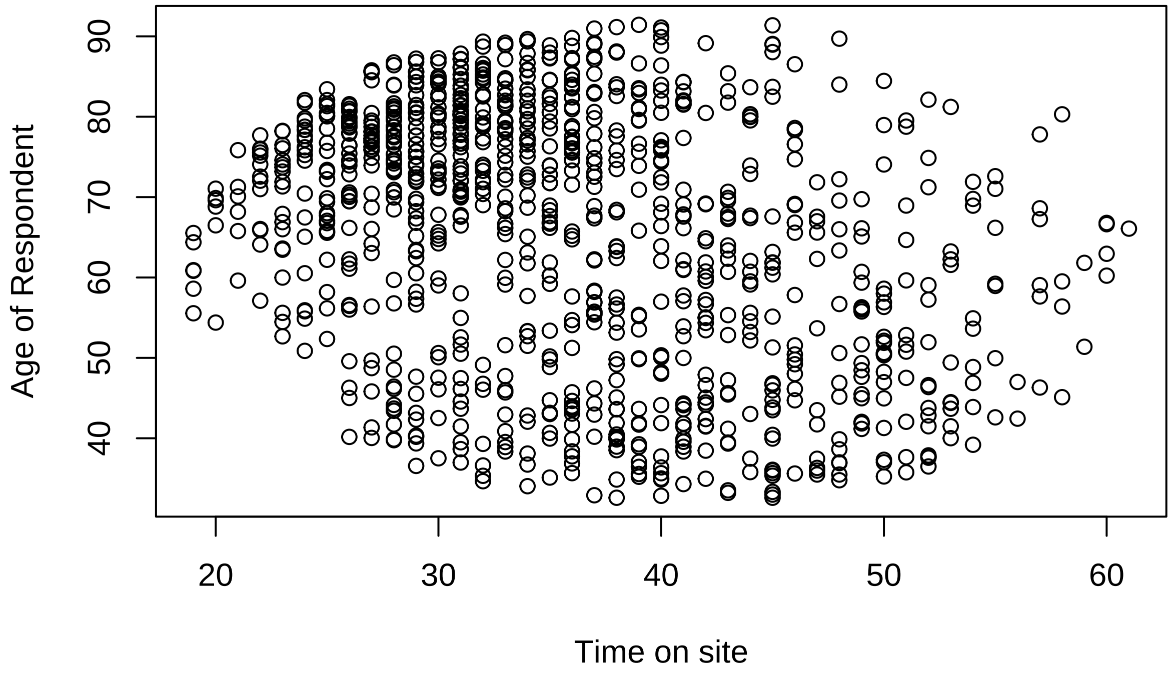
```
# plotting
plot(time.site, daily.internet, xlab="Time on Site", ylab="Daily Internet Usage")
```



```
# plotting
plot(clicks.target, daily.internet, xlab="Daily Internet", ylab="Clicks on Ad")
```

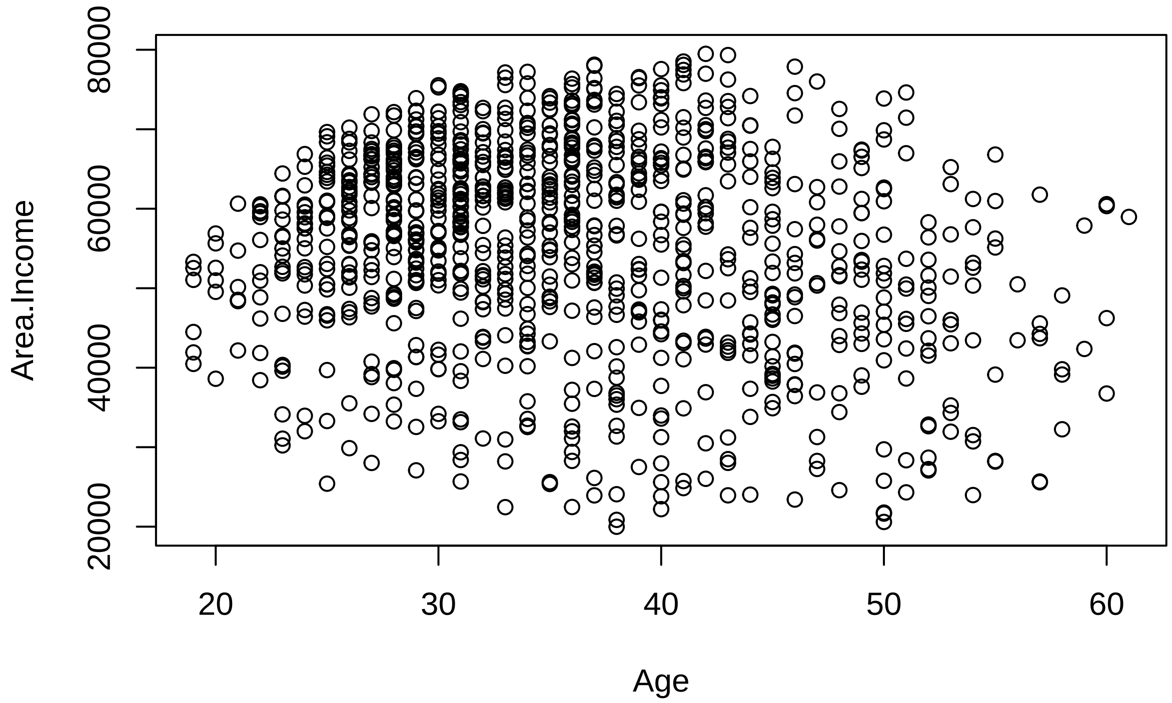


```r
# plotting
plot(age, time.site, xlab="Time on site", ylab="Age of Respondent")
```

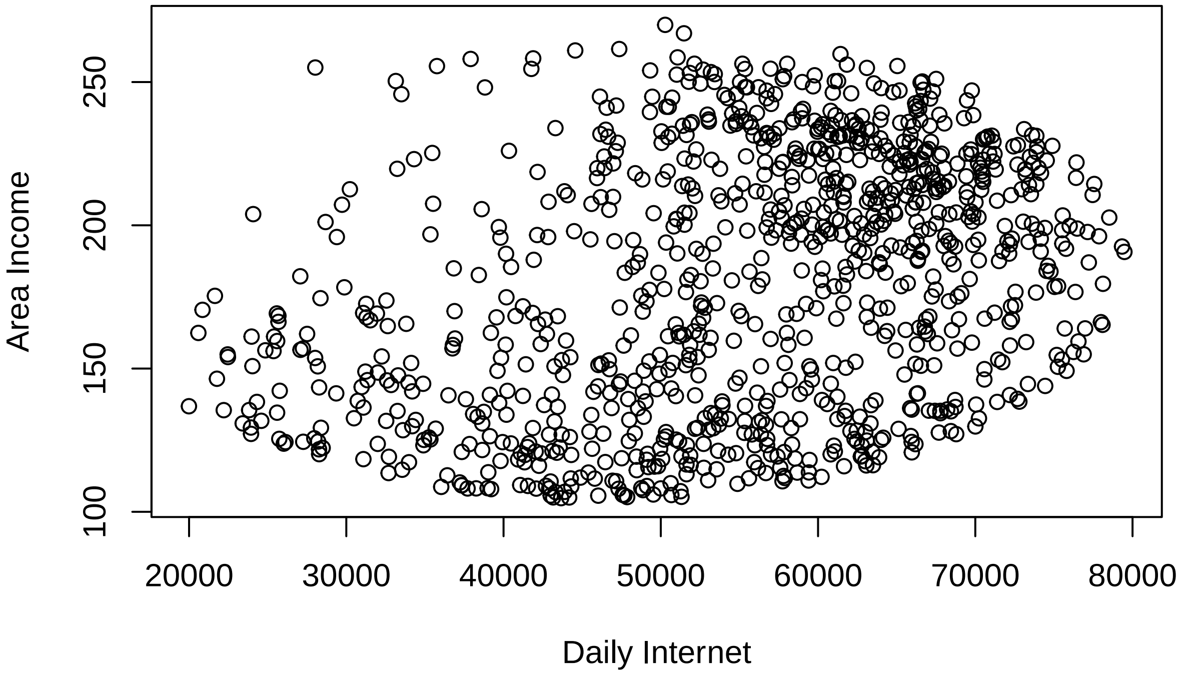


```
# there is no relationship between time spent on site and age
```

```
# plotting
plot(age, area.income, xlab="Age", ylab="Area.Income")
```



```
# plotting
plot(area.income, daily.internet, xlab="Daily Internet", ylab="Area Income")
```



The relationship between the ad being clicked on and other variables are as below Clicked.on.Ad Daily.Time.Spent.on.Site -0.75 Age 0.49 Area.Income -0.47 Daily.Internet.Usage -0.79 Male -0.04 Clicked.on.Ad 1.00

Conclusion and Recommendation

Gender has the least influence on whether the ad is being clicked on or not. Age has a moderately high positive influence on an ad being clicked on, with a mean of about 35 years old, the entrepreneur is advised to custom the advert to target this age group. This data is however skewed and hence could be causing this observation. Area Income has a moderately high negative influence on an ad being clicked on. However since this data is skewed to the right, this could have an influence on this analysis. Daily internet usage and Daily time spent on the site has high negative correlations, this means that when these measurements increase, the chances of an ad being clicked go down. A more balanced data-set could lead to better results.