title: “Blog analytics and visuals”

author: “Christopher Kimani”

date: “2022-07-22”

word_document: default html_document: default pdf_document: default

---

1. Defining the Question**

a) Specifying the Data Analytic Question

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.

b) Defining the Metric for Success

1.Define the question, the metric for success, the context, experimental design taken and the appropriateness of the available data to answer the given question.

2.Perform univariate analysis.

3.Exhaustively perform bivariate analysis.

c) Understanding the context Perform Exploratory Data Analysis for the give data set http://bit.ly/IPAdvertisingData

d) Experimental design taken

1.Reading and checking our data 2.Clean data by finding and dealing with outliers, anomalies, and missing data within the dataset. 3.Perform univariate and bivariate analysis. 4.From the insights provide a conclusion and recommendation.

e) Appropriateness of the data

The data is relevant for this study.

2. Loading the data**

```
data <- read.csv('C:\\Users\\USER\\Downloads\\advertising.csv', header =
TRUE)
```

checking the first 6 rows

```
head(data)
```

```
##   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 last 6 rows

```
tail(data)
```

```
##      Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 995                     43.70  28    63126.96               173.01
## 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
## 995         Front-line bifurcated ability  Nicholasland    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
## 995                  Mayotte 2016-04-04 03:57:48             1
## 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
```

Checking the dimensions of the data There are 1000 rows and 10 columns.

```
dim(data)
```

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

Checking the column names

```
ls(data)
```

```
##  [1] "Ad.Topic.Line"            "Age"
##  [3] "Area.Income"              "City"
##  [5] "Clicked.on.Ad"            "Country"
##  [7] "Daily.Internet.Usage"     "Daily.Time.Spent.on.Site"
##  [9] "Male"                     "Timestamp"
```

Checking the data types and structure The data is made up of numericals, characters and integers.

```
str(data)
```

```
## '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 standardization" "Organic bottom-line service-desk"
"Triple-buffered reciprocal time-frame" ...
##  $ 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               : chr  "2016-03-27 00:53:11" "2016-04-04
01:39:02" "2016-03-13 20:35:42" "2016-01-10 02:31:19" ...
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

Checking the class of the data the data is a data frame

```
class(data)
```

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

3. Data Cleaning

Checking for missing values **There are no missing values in this data**

```
colSums(is.na(data))
```

```
## 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
```

Checking for duplicated values **There are no duplicated values in the data.**

```
sum(duplicated(data))
```

```
## [1] 0
```

Checking for outliers **There are no outliers in all other columns other than in the Area Income column. I will however not remove these outliers as they may be useful in the analysis**

```
length(boxplot.stats(data$'Daily.Time.Spent.on.Site')$out)
```

```
## [1] 0
```

```
length(boxplot.stats(data$'Age')$out)
```

```
## [1] 0
```

```
length(boxplot.stats(data$'Area.Income')$out)
```

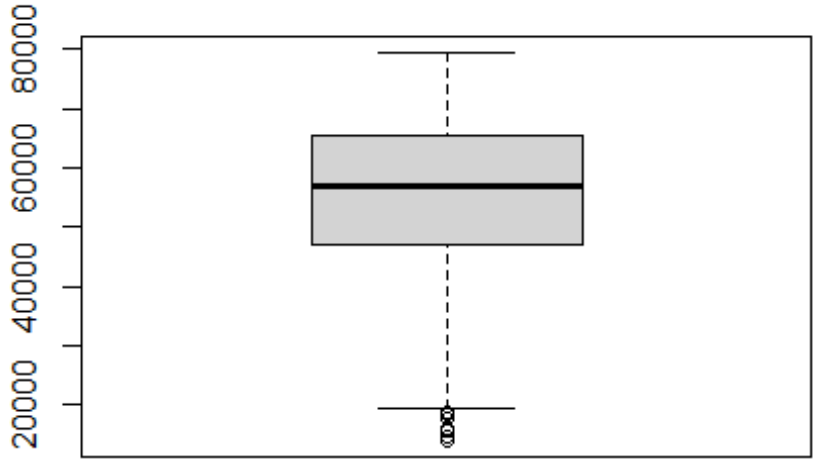
```
## [1] 8
```

```
length(boxplot.stats(data$'Daily.Internet.Usage')$out)
```

```
## [1] 0
```

Plotting outliers in the Area Income column

```
boxplot(data$'Area.Income')
```



Exploratory Data Analysis ## 4. Univariate Analysis A summary of the data **This shows a summary statistic of data. The mean, median, minimum and maximum values, the class of the data and the quartiles**

```
summary(data)

##  Daily.Time.Spent.on.Site      Age         Area.Income
Daily.Internet.Usage
##  Min.   :32.60           Min.   :19.00   Min.   :13996   Min.   :104.8
##  1st Qu.:51.36           1st Qu.:29.00   1st Qu.:47032   1st Qu.:138.8
##  Median :68.22           Median :35.00   Median :57012   Median :183.1
##  Mean   :65.00           Mean   :36.01   Mean   :55000   Mean   :180.0
##  3rd Qu.:78.55           3rd Qu.:42.00   3rd Qu.:65471   3rd Qu.:218.8
##  Max.   :91.43           Max.   :61.00   Max.   :79485   Max.   :270.0
##  Ad.Topic.Line          City               Male         Country
##  Length:1000        Length:1000        Min.   :0.000   Length:1000
##  Class :character   Class :character   1st Qu.:0.000   Class :character
##  Mode  :character   Mode  :character   Median :0.000   Mode  :character
##                                        Mean   :0.481
##                                        3rd Qu.:1.000
##                                        Max.   :1.000
##   Timestamp         Clicked.on.Ad
##  Length:1000        Min.   :0.0
##  Class :character   1st Qu.:0.0
##  Mode  :character   Median :0.5
##                     Mean   :0.5
```

```
##                      3rd Qu.:1.0
##                      Max.   :1.0
```

Obtaining the variances of the data **This shows the variances of the chosen columns**

```
var(data$'Age')
```

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

```
var(data$'Daily.Time.Spent.on.Site')
```

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

```
var(data$'Area.Income')
```

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

```
var(data$'Daily.Internet.Usage')
```

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

Obtaining the standard deviation of the data **This shows the standard deviation of the data**

```
sd(data$'Age')
```

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

```
sd(data$'Daily.Time.Spent.on.Site')
```

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

```
sd(data$'Area.Income')
```

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

```
sd(data$'Daily.Internet.Usage')
```

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

Obtaining the mode of the data **This shows the mode of the data**

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]}
getmode(data$'Age')
```

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

```
getmode(data$'Daily.Time.Spent.on.Site')
```

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

```
getmode(data$'Area.Income')
```

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

```
getmode(data$'Daily.Internet.Usage')
```

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

Getting which countries had the most count

Aruba had the least count France and Czech Republic had the most count

```
country = table(data$Country)
countries <- sort(country, increasing = TRUE)
countries <- sort(country, decreasing = TRUE)
head(countries)
```

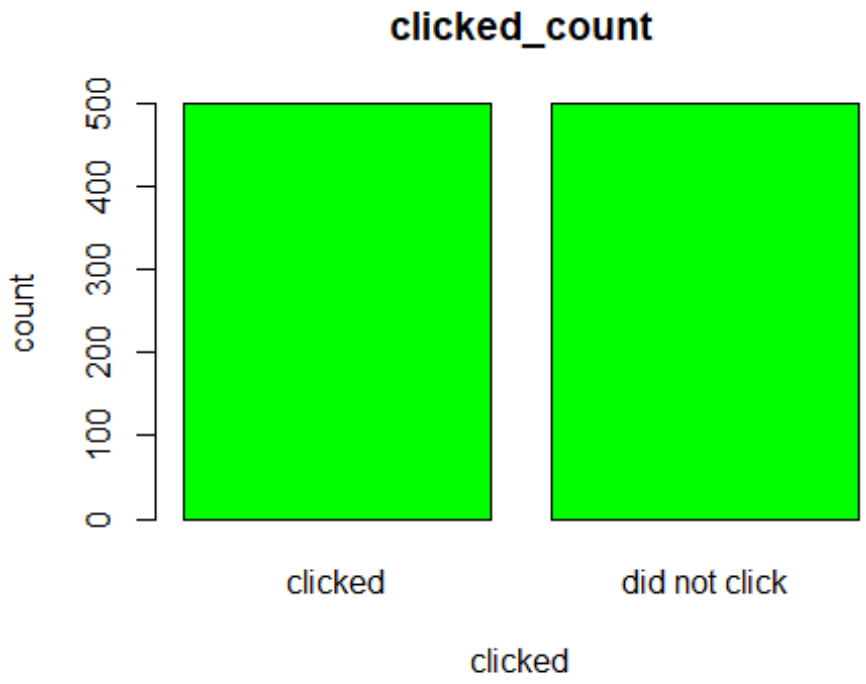
```
##
## Czech Republic         France    Afghanistan      Australia         Cyprus
##              9              9              8              8              8
##         Greece
##              8
```

Obtaining the count and visualisation of Click.on.Ad **There is an equal number of those who clicked and those not clicked of 500 There is no class imbalance**

```
clicked <- table(data$'Clicked.on.Ad')
clicked
```

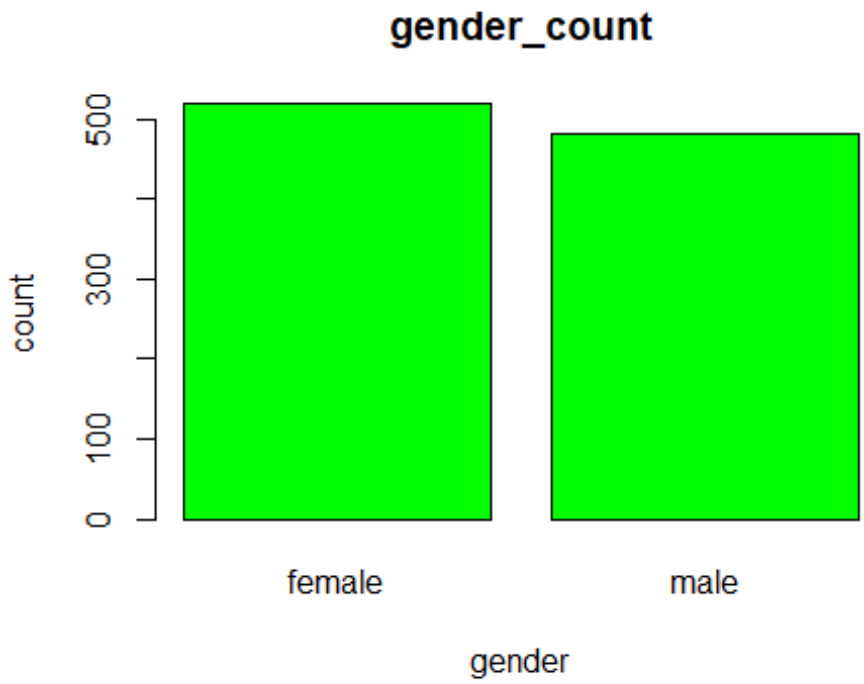
```
##
##   0   1
## 500 500
```

```
labels <- c('clicked','did not click')
barplot(clicked, ylab = 'count', names.arg = labels, xlab = 'clicked', main
='clicked_count', col = 'green')
```



Gender count **The number of females is slightly more than that of men**

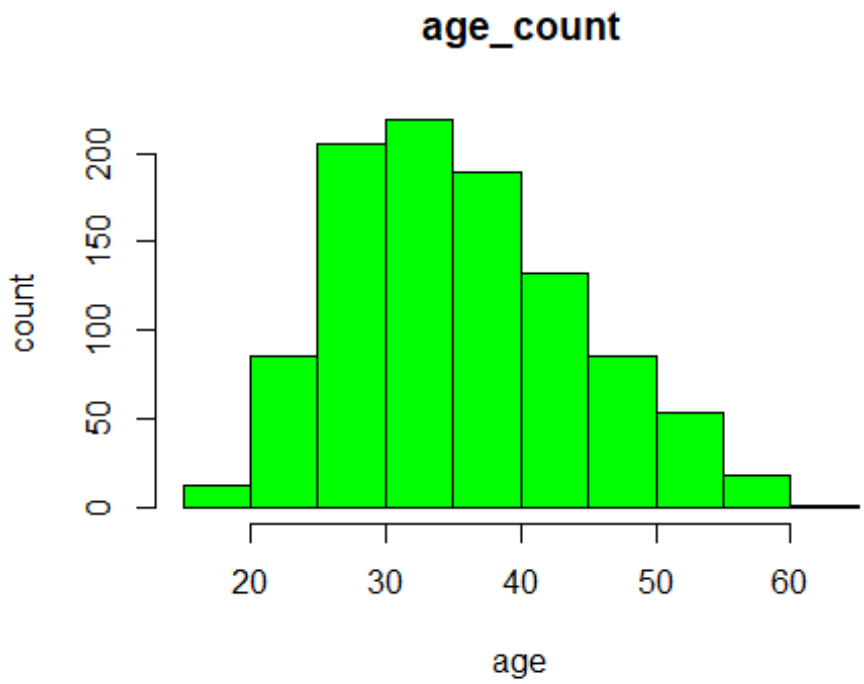
```
gender <- table(data$'Male')
gender

##
##   0   1
## 519 481

label <- c('female','male')
barplot(gender, ylab = 'count',names.arg = label, xlab = 'gender', main =
'gender_count', col = 'green')
```



Histogram for Age

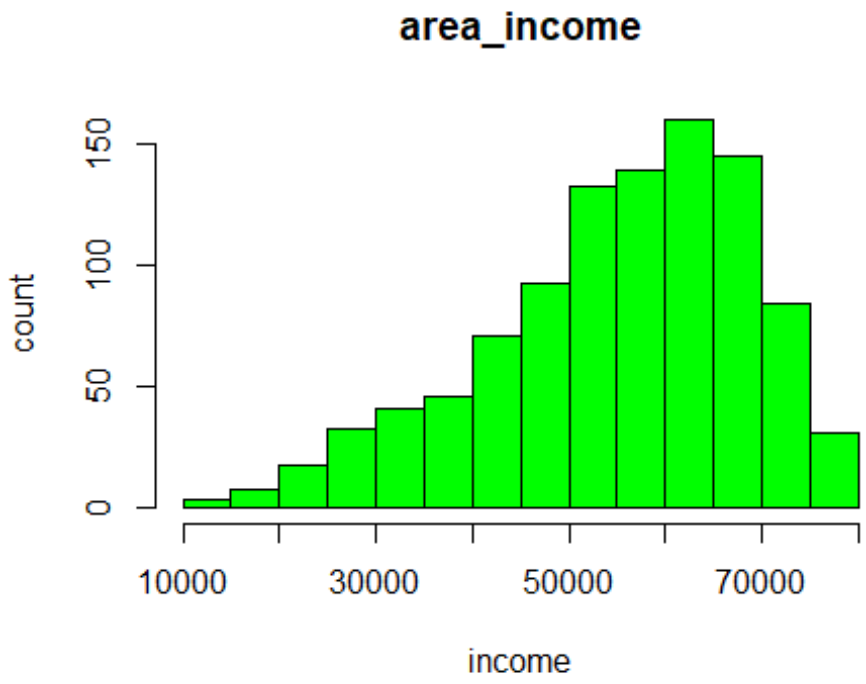
Most people in the data are aged 25-45 years

```
hist(data$'Age', ylab = 'count', xlab = 'age', main = 'age_count', col =
'green')
```



Histogram for Area Income **This chart shows that Area.Income is skewed to the right**

```
hist(data$'Area.Income', ylab = 'count', xlab = 'income', main =
'area_income', col = 'green')
```

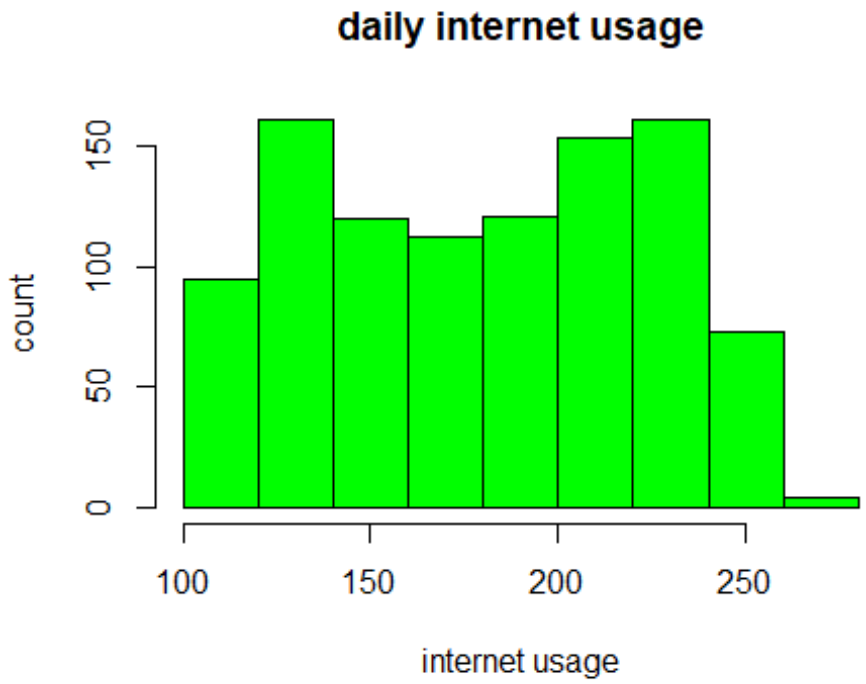


Showing the skewness of the Area Income column

```
library(moments)
skewness(data$Area.Income)

## [1] -0.6493967
```

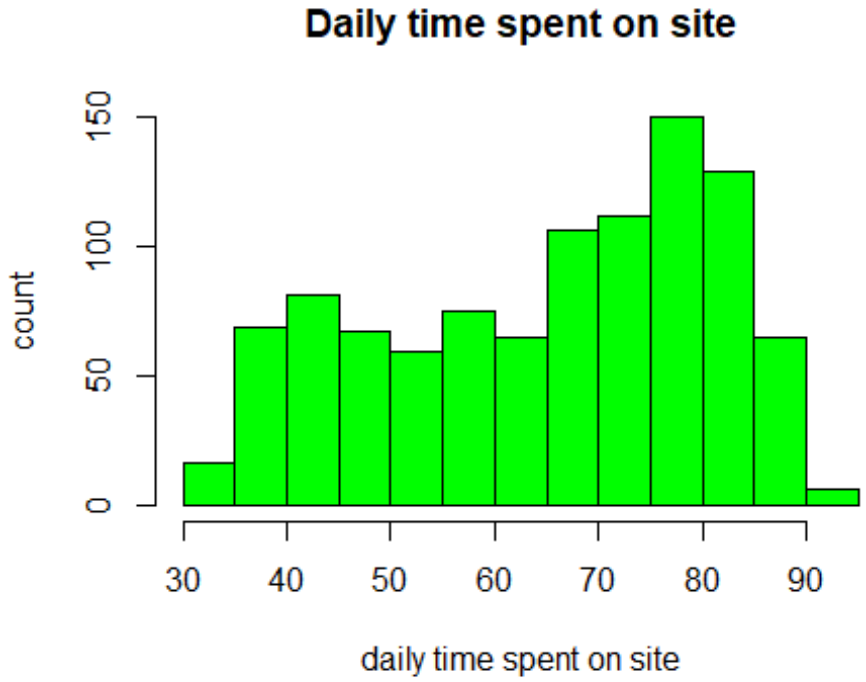
Histogram of Daily Internet Usage **The data is more or less equally distributed**

```
hist(data$'Daily.Internet.Usage', ylab ='count', xlab = 'internet usage',
main = 'daily internet usage', col = 'green')
```



Histogram of Daily Time Spent on Site **The data is fairly skewed to the right**

```
hist(data$Daily.Time.Spent.on.Site, ylab = 'count', xlab = 'daily time spent
on site', main = 'Daily time spent on site', col = 'green')
```



```
skewness(data$'Daily.Time.Spent.on.Site')
```

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

##5. Bivariate Analysis

Average Income per gender

Men had more income than females

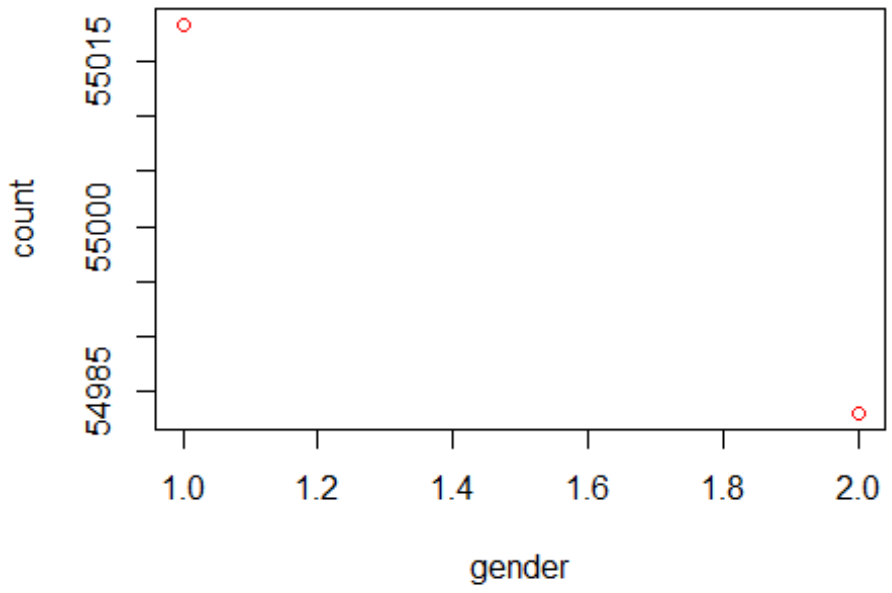
```
mmen <- mean(data[data$Male == '1', 'Area.Income'])
mwomen <- mean(data[data$Male == '0', 'Area.Income'])
mmen
```

```
## [1] 55018.42
```

```
mwomen
```

```
## [1] 54982.93
```

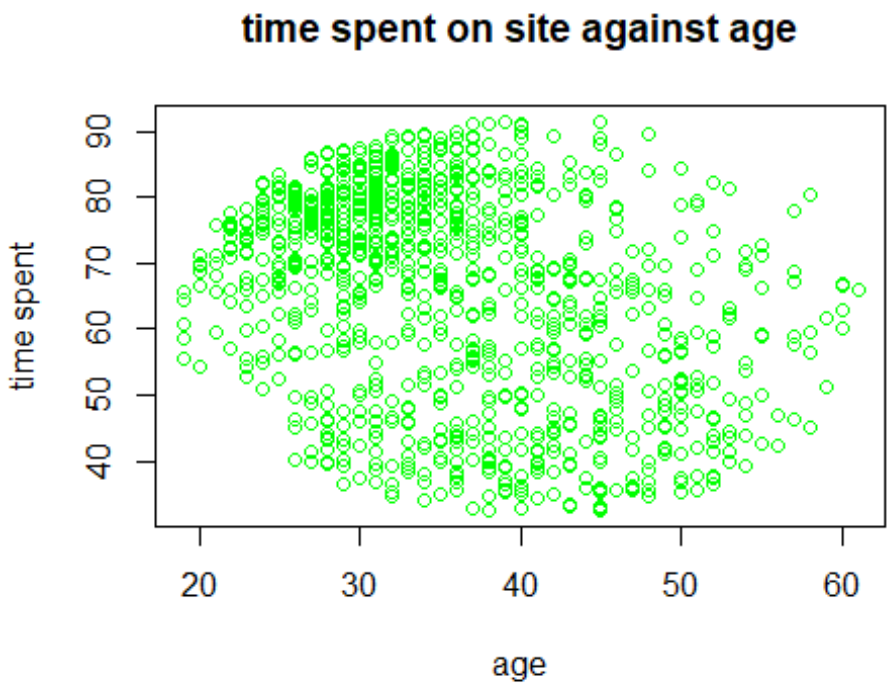
```
mboth <- c(mmen, mwomen)
plot(mboth, ylab = 'count', xlab = 'gender', col = 'red')
```



Relationship between age and daily time spent on site

From this plot it is seen that people around the age of 25-35 years spent the most time on site People above the age of 40 spend lesser time on site. There is a negative correlation between age and time.

```
# Scaling some columns

age <- data$'Age'
nage <- scale(age)
time <- data$'Daily.Time.Spent.on.Site'
ntime <- scale(time)
plot(age, time, ylab = 'time spent', xlab = 'age', main = 'time spent on site
against age', col = 'green')
```

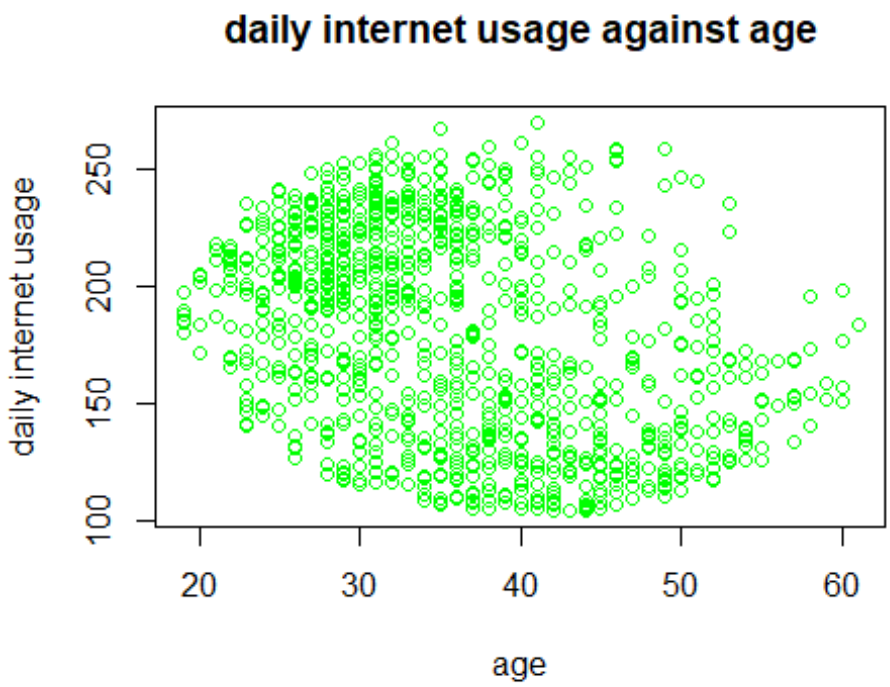



```
cov(nage, ntime)

##             [,1]
## [1,] -0.3315133
```

Relationship between age and daily internet usage

The ages below 40 had more usage than those above 40 years. There is a negative correlation between age and daily internet usage

```
usage <- data$'Daily.Internet.Usage'
nusage <- scale(usage)
cov(usage, age)

## [1] -141.6348

plot(age, usage, ylab ='daily internet usage', xlab ='age', main ='daily
internet usage against age', col = 'green')
```




```
cov(nage, nusage)

##             [,1]
## [1,] -0.3672086
```

8. Conclusion

1. There are more females than males in our data.
2. 500 people clicked on the ads while 500 others did not click on the ads.
3. The average area income is 55000.
4. The average age of most audience is 36 years
5. Lisamouth and Williamsport cities both had the highest number of individuals in the dataset.

9. Recommendations

1. Persons aged between 25 and 35 years old were the most in the data, thus creating ads to target these age group would be very impactful.
2. Creating ads that target men makes more sense since men have more income compared to women.