Brian Michira Week 12 IP R Fundamentals

Brian Michira

8/27/2021

Defining the Question

Research 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.

Metrics For Success

This project will be successful when we correctly identify which individuals are most likely to click on the ads.

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.

Experimental Design Taken

1.Specifying the research question. 2. 2.Loading and Previewing the Dataset. 3.Cleaning the Dataset. 4.Explanatory Data Aalysis. 5.Conclusion.

1.Loading and Previewing the dataset

```
library(data.table)

#Loading the dataset
dataset <- read.csv("http://bit.ly/IPAdvertisingData")
#Viewing the top of the dataset
head(dataset)
```

```
##   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
```

```
#Viewing the bottom of the dataset
tail(dataset)
```

```
##      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 number of records
dim(dataset)
```

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

The dataset has 1000 rows and 10 columns.

```
#Checking the Class of our dataset
class(dataset)
```

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

#checking the info
str(dataset)

## '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 ...
```

2.Cleaning the Dataset

```
#checking for missing values
sum(is.na(dataset))

## [1] 0
```

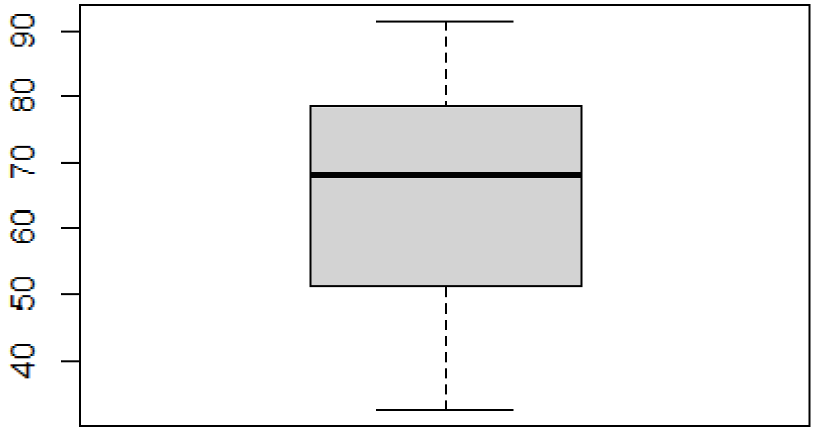
There are no missing values.

```
#checking for duplicates
duplicated <- dataset[duplicated(dataset),]
duplicated

##  [1] Daily.Time.Spent.on.Site Age                      Area.Income
##  [4] Daily.Internet.Usage     Ad.Topic.Line            City
##  [7] Male                     Country                  Timestamp
## [10] Clicked.on.Ad
## <0 rows> (or 0-length row.names)
```

There are no duplicated rows.

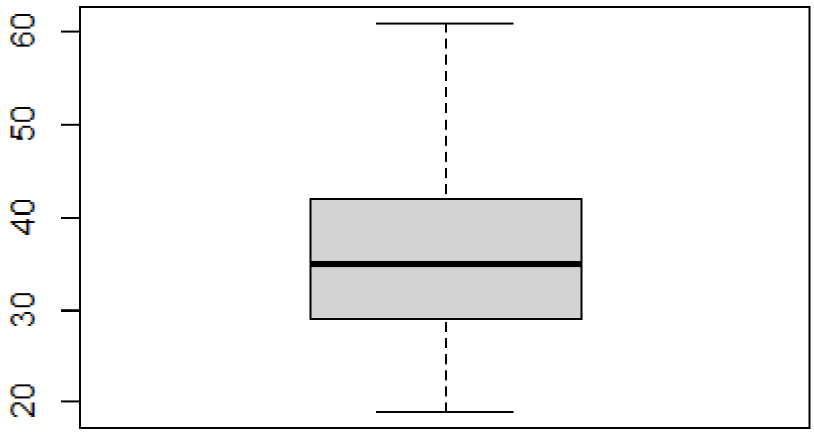
```
#checking the info
str(dataset)

## '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 ...

# overview of the dataset
summary(dataset)

##  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

# checking for outliers on Daily time spent on site
boxplot(dataset$Daily.Time.Spent.on.Site)
```



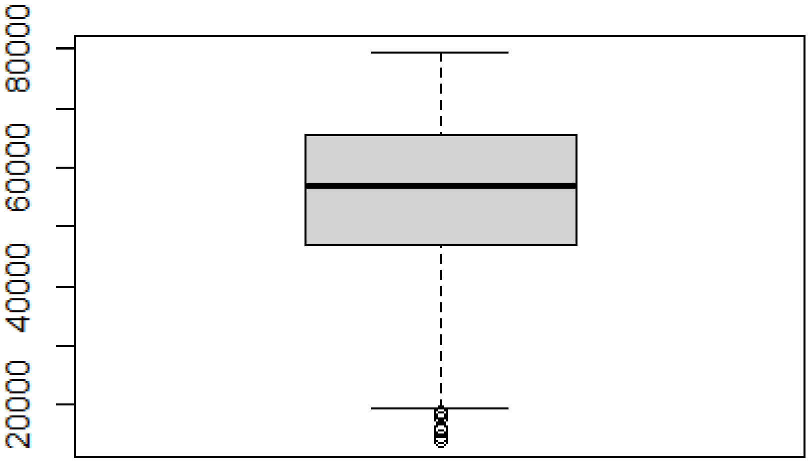
There are no outliers on Daily Time Spent on Site.

```
# checking for outliers on Age
boxplot(dataset$Age)
```



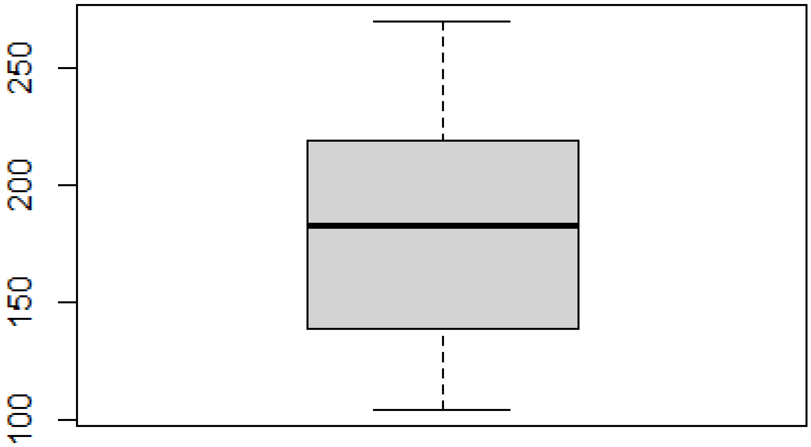
There are no outliers on the Age column.

```
# checking for outliers on Area income
boxplot(dataset$Area.Income)
```



There are outliers on the Area income column.

```
# viewing the exact outliers
boxplot.stats(dataset$Area.Income)$out

## [1] 17709.98 18819.34 15598.29 15879.10 14548.06 13996.50 14775.50
18368.57

# checking for outliers on Daily Internet Usage
boxplot(dataset$Daily.Internet.Usage)
```



There are no outliers on the daily Internet.

3.Exploratory Data Analysis

Univariate Analysis

Daily Time Spent on Site

1. Measures of Central Tendency

```
# mean
mean(dataset$Daily.Time.Spent.on.Site)

## [1] 65.0002
```

The Mean pf the Time Spent on Site Daily is 65.0002.

```
# median
median(dataset$Daily.Time.Spent.on.Site)

## [1] 68.215
```

The Median of the Time Spent on Site Daily is 68.215.

```
# mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

```
getmode(dataset$Daily.Time.Spent.on.Site)

## [1] 62.26
```

The most common time Spent on site is 62.26.

2.Measure of Dispersion

```
#Standard Deviation
sd(dataset$Daily.Time.Spent.on.Site)

## [1] 15.85361
```

The Standard Deviation of Daily Time Spent on Site is 15.85361.

```
#Variance
var(dataset$Daily.Time.Spent.on.Site)

## [1] 251.3371
```

The Variance of Daily Time Spent on Site is 251.3371.

```
#Range
range(dataset$Daily.Time.Spent.on.Site)

## [1] 32.60 91.43
```

The range of Daily Time Spent on Site was 32.60 on the minimum and 91.43 on the maximum.

```
#Quantile
quantile(dataset$Daily.Time.Spent.on.Site)

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

Age

1.Measure of Central Tendancy

```
# mean
mean(dataset$Age)

## [1] 36.009
```

The mean Age is 36 years.

```
# median
median(dataset$Age)

## [1] 35
```

The median Age is 35 years

```
# mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(dataset$Age)

## [1] 31
```

Most people had 31 years.

2.Measure of Dispersion

```
#Variance
var(dataset$Age)

## [1] 77.18611
```

The variance of age is 77.18611.

```
#Standard Deviation
sd(dataset$Age)

## [1] 8.785562
```

The Standard Deviation of Age is 8.785562.

```
#range
range(dataset$Age)

## [1] 19 61
```

The minimum age is 19 years and the maximmum age is 61 years.

```
#Quantile
quantile(dataset$Age)

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

Area Income

1.Measure of Central Tendency

```
#Mean
mean(dataset$Area.Income)

## [1] 55000
```

The mean Area income is 55,000.

```
#median
median(dataset$Area.Income)
```

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

The median Area Income is 57012.3.

```
# mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(dataset$Area.Income)

## [1] 61833.9
```

The common Area Income is 61,833.9.

2.Measure of Dispersion

```
#Variance
var(dataset$Area.Income)

## [1] 179952406
```

The Variance of Area Income is 179952406.

```
#Standard Deviation
sd(dataset$Area.Income)

## [1] 13414.63
```

The Standard Deviation of Area Income is 13414.63.

```
#Range
range(dataset$Area.Income)

## [1] 13996.5 79484.8
```

The minimum Area Income is 13996.5 and the maximum Area Income is 79484.8.

```
#quantile
quantile(dataset$Area.Income)

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

Daily Internet Usage

1.Measure of central Tendecy

```
#mean
mean(dataset$Daily.Internet.Usage)

## [1] 180.0001
```

The Mean Internet Usage is 180.0001.

```
#median
median(dataset$Daily.Internet.Usage)

## [1] 183.13
```

The Median internet usage is 183.13.

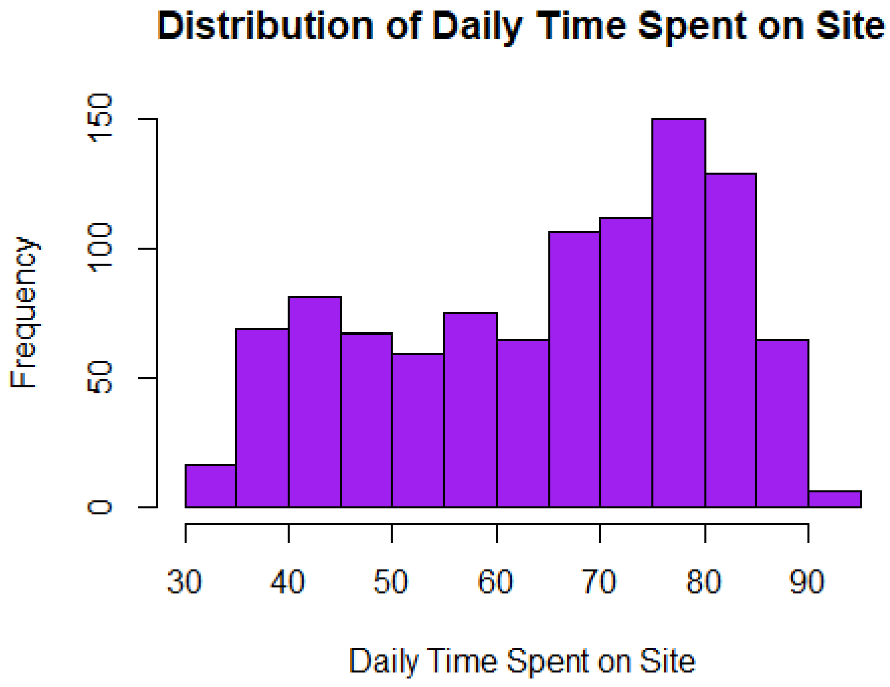
```
# mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(dataset$Daily.Internet.Usage)

## [1] 167.22
```

The most common Internet Usage is 167.22.

Graphical representation of univariate ananlysis

###Daily Time Spent on Site

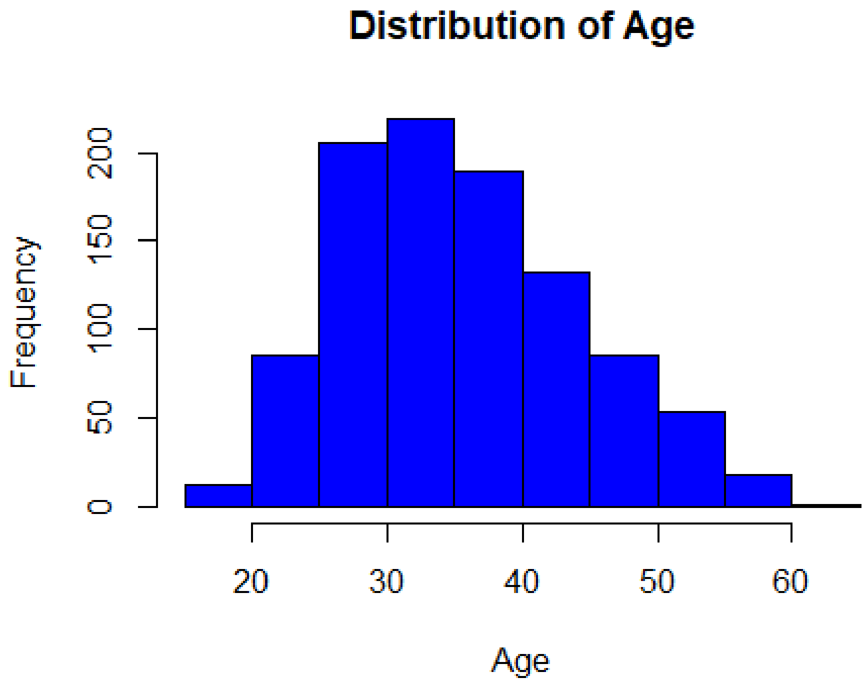
```
hist(dataset$Daily.Time.Spent.on.Site,main = "Distribution of Daily Time
Spent on Site",col="purple"
    ,xlab="Daily Time Spent on Site")
```



The duration between 75 and 85 had the highest frequency.

Age

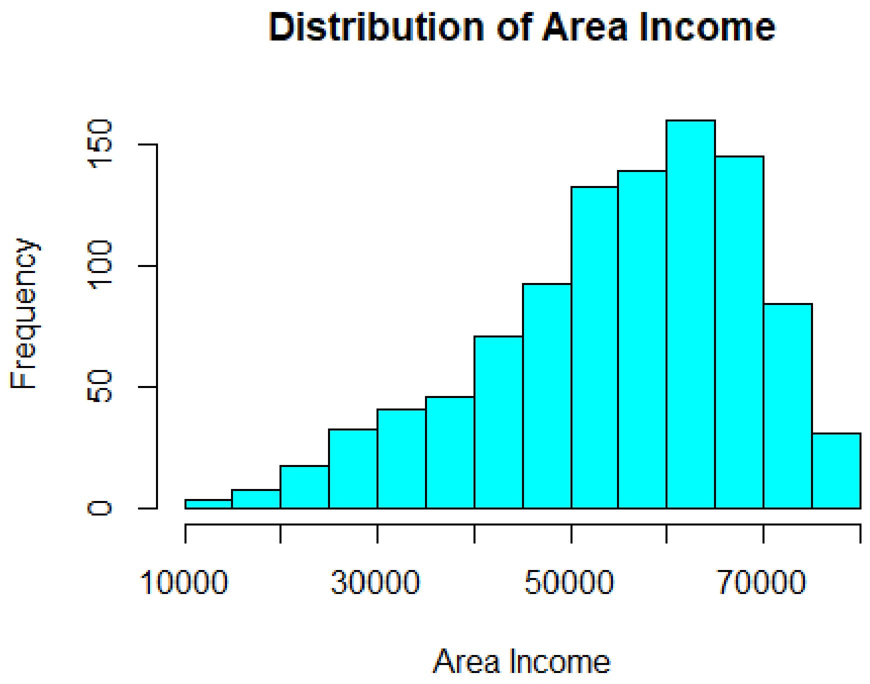
```
hist(dataset$Age,main = "Distribution of Age",col="blue",
     xlab="Age")
```



The age between 25 years and 35 years had the highest frequency. Age is skewed to the right.

Area Income

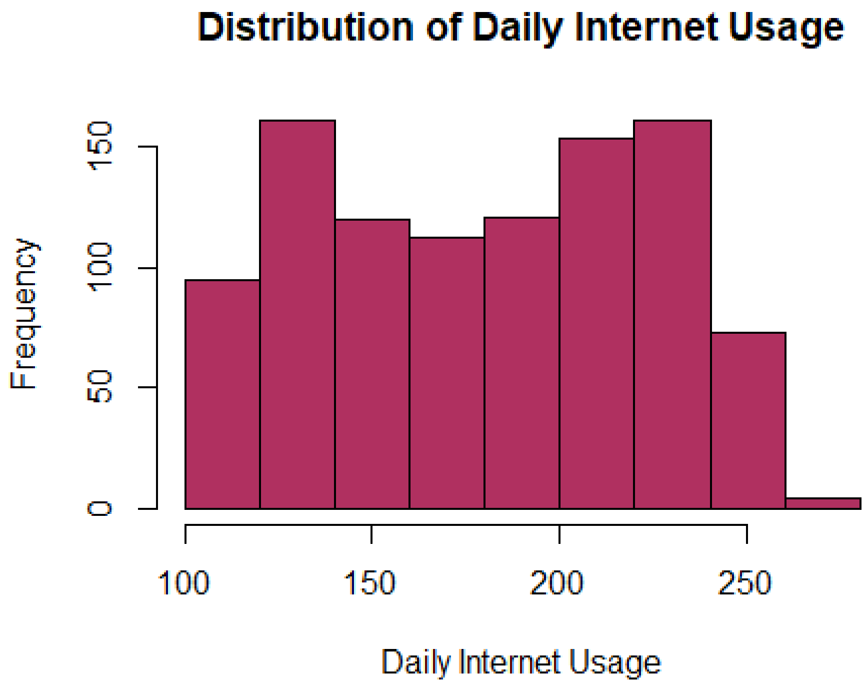
```
hist(dataset$Area.Income,main = "Distribution of Area Income",col="Cyan",
     xlab = "Area Income")
```



The Income between 60,000 and 70,000 had the highest frequency.

Daily Internet Usage

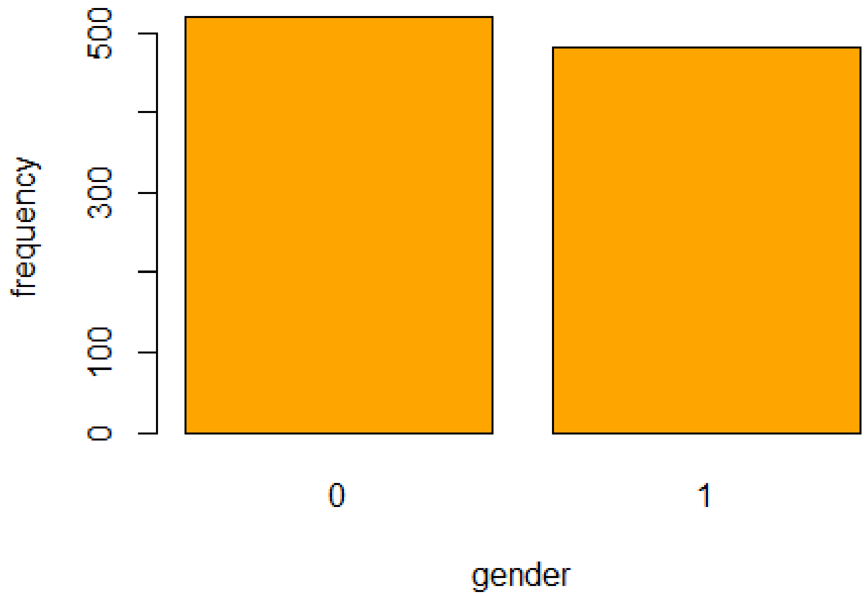
```
hist(dataset$Daily.Internet.Usage,main = "Distribution of Daily Internet
Usage",col="maroon",
     xlab="Daily Internet Usage")
```



The usage between 200 and 250 had the highest frequency. Daily internet usage is Bimodal.

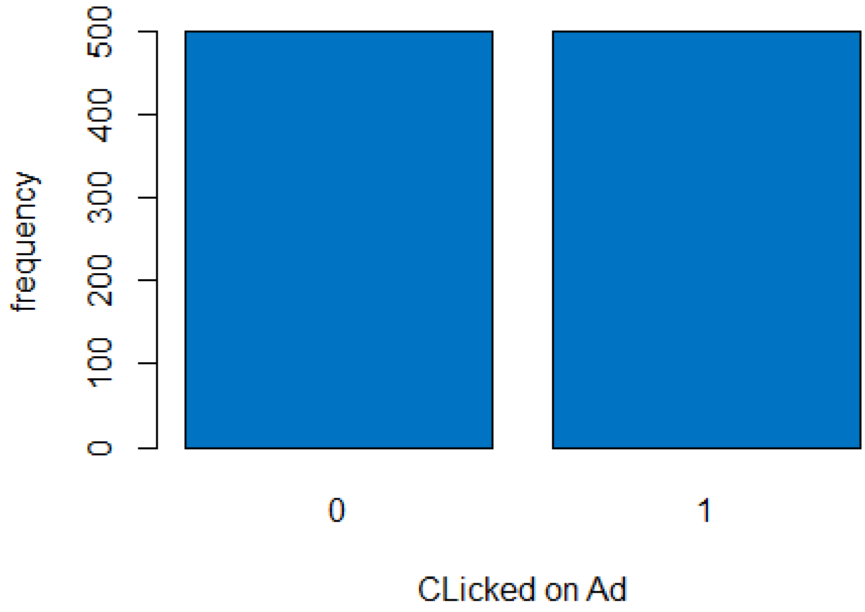
Univariate Analysis of Categorical Data

```
Gender <- dataset$Male
frequency<- table(Gender)
frequency

## Gender
##   0   1 
## 519 481

barplot(frequency,xlab ="gender", ylab = "frequency", col="orange")
```



Majority of the respondents were female.519 females 481 males.

```
Ads <- dataset$Clicked.on.Ad
frequency<- table(Ads)
frequency

## Ads
##   0   1
## 500 500

barplot(frequency,xlab ="CLicked on Ad", ylab = "frequency", col="#0073C2FF")
```



2.Bivariate Analysis

(i)Covariance

```
# covariance between daily time spent on site and age
cov(dataset$Daily.Time.Spent.on.Site,dataset$Age)

## [1] -46.17415
```

The covariance between daily time spent on site and age is -46.17. It indicates a negative linear relationship between the two variables.

```
# covariance between area income and daily internet usage
cov(dataset$Area.Income,dataset$Daily.Internet.Usage)

## [1] 198762.5
```

The covariance between area income and daily internet usage is 198762.5. It indicates a positive linear relationship between the two variables.

```
# covariance between daily internet usage and age
cov(dataset$Daily.Internet.Usage,dataset$Age)

## [1] -141.6348
```

The covariance between daily internet usage and age is -141.6348. It indicates a negative linear relationship between the two variables.

(ii)Correlation

```
# correlation coefficient between area income and daily internet usage
cor(dataset$Area.Income,dataset$Daily.Internet.Usage)

## [1] 0.3374955
```

The correlation coefficient between area income and daily internet usage is 0.3375.

```
# correlation coefficient between daily time spent on site and area income
cor(dataset$Daily.Time.Spent.on.Site,dataset$Area.Income)

## [1] 0.3109544
```

The correlation coefficient between daily time spent on site and area income is 0.311.

```
#correlation matrix
cor(dataset[,unlist(lapply(dataset, is.numeric))])

##                          Daily.Time.Spent.on.Site         Age  Area.Income
## Daily.Time.Spent.on.Site               1.00000000 -0.33151334  0.310954413
## Age                                   -0.33151334  1.00000000 -0.182604955
## Area.Income                            0.31095441 -0.18260496  1.000000000
## Daily.Internet.Usage                   0.51865848 -0.36720856  0.337495533
## Male                                  -0.01895085 -0.02104406  0.001322359
## Clicked.on.Ad                         -0.74811656  0.49253127 -0.476254628
##                          Daily.Internet.Usage         Male Clicked.on.Ad
## Daily.Time.Spent.on.Site           0.51865848 -0.018950855   -0.74811656
## Age                               -0.36720856 -0.021044064    0.49253127
## Area.Income                        0.33749553  0.001322359   -0.47625463
## Daily.Internet.Usage               1.00000000  0.028012326   -0.78653918
## Male                               0.02801233  1.000000000   -0.03802747
## Clicked.on.Ad                     -0.78653918 -0.038027466    1.00000000
```

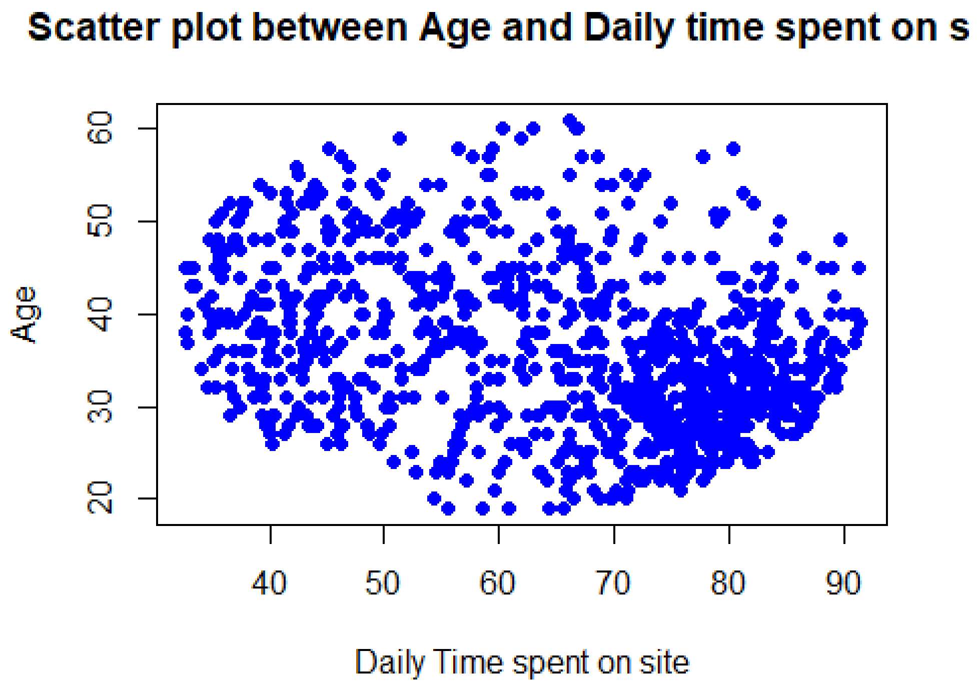
From the correlation matrix looking at the clicked on AD we can see that its only age that has a positive correlation with Clicked on AD. We can confirm this by conducting the point Biserial correlation which is used to test correlation between a continuos and a categorical variable.

```
#Biserial correlation
cor.test(dataset$Age,dataset$Clicked.on.Ad)

## 
##  Pearson's product-moment correlation
## 
## data:  dataset$Age and dataset$Clicked.on.Ad
## t = 17.879, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4440981 0.5380944
## sample estimates:
##       cor 
## 0.4925313
```

We found a correlation coefficient of 0.492 which show there is a positive correlation between Age and Clicked on Ad.

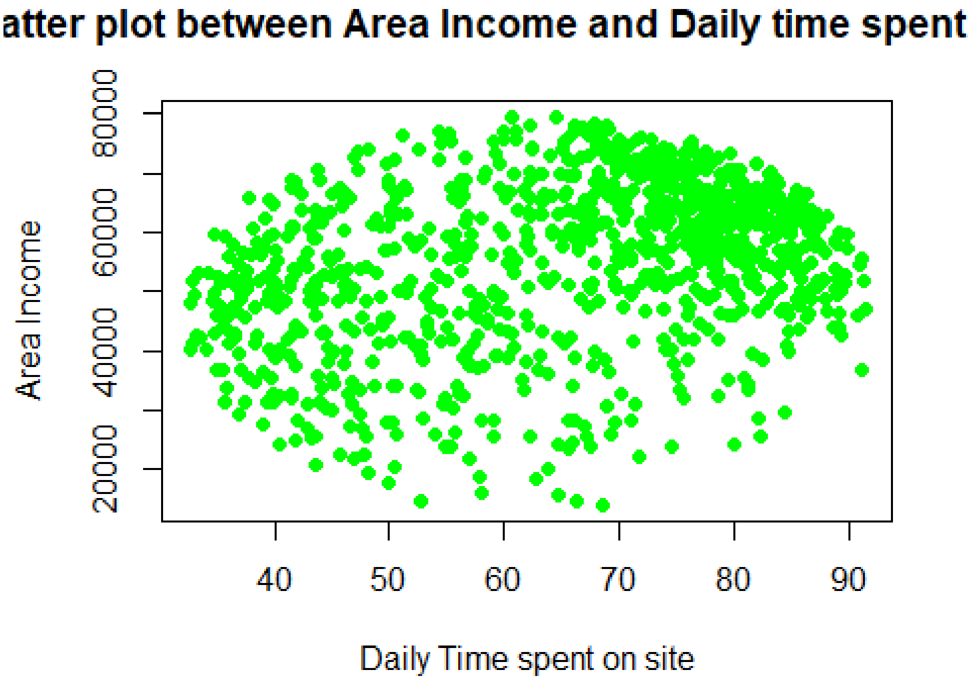
(iii)Scatter plots

```
# scatter plot between age and daily time spent on site
plot(dataset$Daily.Time.Spent.on.Site,dataset$Age,main="Scatter plot between
Age and Daily time spent on site",col='blue',xlab="Daily Time spent on
site",ylab="Age",pch=19)
```



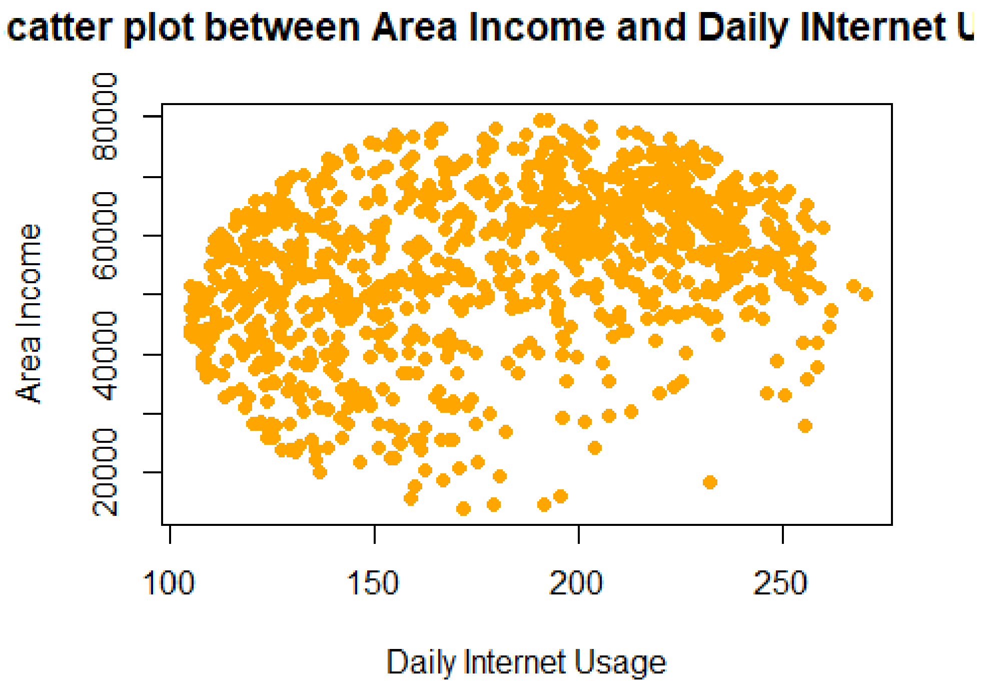
The scatter plot of daily time spent on site and age shows us that between the ages 25 years and 40 years spent more time on site.

```
# scatter plot between Area Income and daily time spent on site
plot(dataset$Daily.Time.Spent.on.Site,dataset$Area.Income,main="Scatter plot
between Area Income and Daily time spent on site",col='green',xlab="Daily
Time spent on site",ylab="Area Income",pch=19)
```



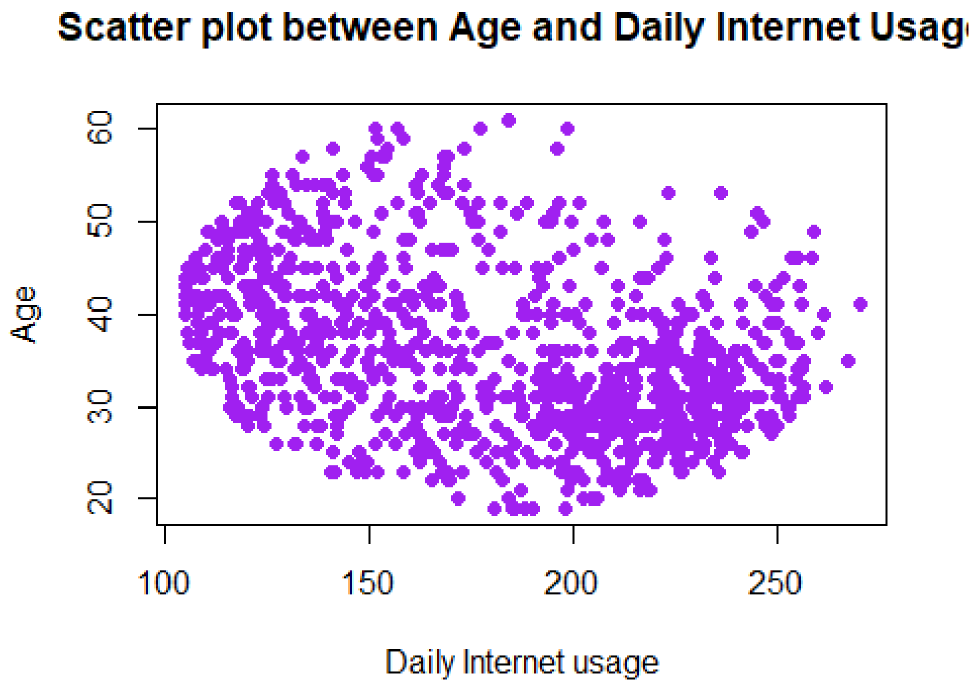
The scatter plot between daily time spent on site and area income shows us that those with an area income between 50,000 and 70,000 are the ones who spend more time on site.

```
# scatter plot between Area Income and daily time spent on site
plot(dataset$Daily.Internet.Usage,dataset$Area.Income,main="Scatter plot
between Area Income and Daily INternet Usage",col='orange',xlab="Daily
Internet Usage",ylab="Area Income",pch=19)
```



The scatter plot of daily internet usage and Area Income shows us that between income 60,000 and 75,000 spent more on internet while those between income 35,000 and 50,000 spent less on internet.

```
# scatter plot between age and daily Internet Usage
plot(dataset$Daily.Internet.Usage,dataset$Age,main="Scatter plot between Age
and Daily Internet Usage",col='purple',xlab="Daily Internet
usage",ylab="Age",pch=19)
```



The scatter plot of daily internet usage and age shows us that between the ages 25 years and 40 years spent more on internet.

Conclusion

1.Majority of the respondents were females.

2.All the variables have a negative correlation with Clicked on Ad apart from Age.

3.There is a positive correlation between Area Income and Daily Time Spent on Site.

4.There is a positive correlation between Area Income and Daily Internet Usage.

5.There is a positive correlation between Age and Clicked on Ad.

6.Respondents aged between 25 years and 40 years spent more time on the internet.

7.Respondents aged between 25 years and 40 years spent a lot on Internet.This is supported by the fact that they spent a lot of time on the internet.

From our Analysis we found out that the elderly are more likely to click on the Ads.

From our analysis we found out that the low income earners are more likely to click on the Ads.