Advert Analysis

Faith

9/4/2020

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.

From the question derive the problem statement and the solution.

Problem statement

An entrepreneur with an online cryptography would like to advertise her blog. She target audiences from different countries and would like to determine who are more likely to click on her adverts.

Solution

As a data scientist, I will consider all the variables the entrepreneur collected and will be used to identify the target audience.

METRIC FOR SUCCESS

Perform comprehensive data cleaning on the data. Draw important insights from univariate and bivariate analysis by performing measures of central tendency and graphical presentation.

UNDERSTANDING THE CONTEXT

An entrepreneur has collected data to determine her target audience.

EXPERIMENTAL DESIGN

Load the R library and import the data.

Clean the data.

Perform univariate and multivariate analysis.

Conclusions and recommendations.

Load the data

Check file location

```
getwd()
```

```
## [1] "C:/Users/FGakori/Documents/R advert/Advertising-analysis-in-R"
```

```
setwd('C:/Users/FGakori/Documents/R advert/Advertising-analysis-in-R')
```

Import advertisment csv

```
advert <- read.csv('advertising.csv', TRUE, ',')
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
```

```
class(advert)
```

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

Review structure of the dataframe

```r
dim(advert)
```

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

The data frame contains 1000 rows and 10 columns

Review datatypes

```r
str(advert)
```

```
## ’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               : chr  "2016-03-27 00:53:11" "2016-04-04 01:39:02" "2016-03-13 20:35:42"
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

Daily time spent, area income, internet usage are all numerical. Age, male,clicked on ad are integers. Topic line,city,country,time stamp are all factors.

Statistical summary

```r
summary(advert)
```

```
##  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  
```

Displays the statistical summary of the variables. The maximum daily time spent on a site is 91 minutes and 43 seconds. Most people who visit the site are mainly of age 36 having 19 years as minimum and 61 years as the maximum age.

MISSING VALUES AND OUTLIERS

Checking for missing values in the columns

```
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

Checking duplicates

```
anyDuplicated(advert)
```

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

There are no duplicates

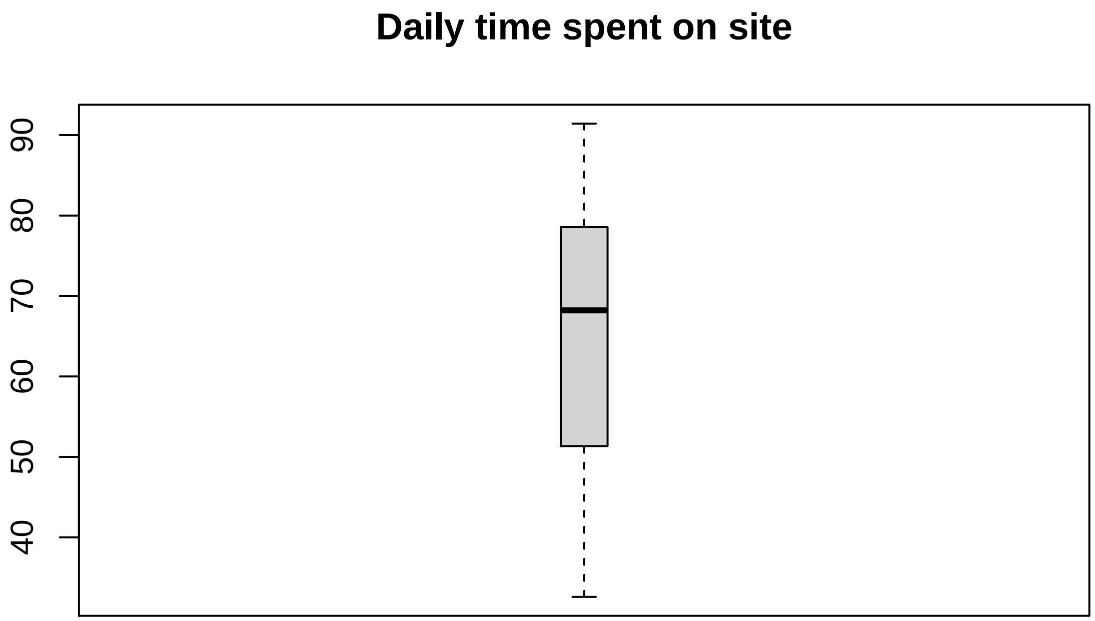
Checking for outliers using box plots

Daily time spent on site

```
time_site <- boxplot.stats(advert$Daily.Time.Spent.on.Site)$out
time_site
```

```
## numeric(0)
```

```
boxplot(advert$Daily.Time.Spent.on.Site, main="Daily time spent on site", boxwex=0.1)
```



There are no outliers in daily time spent on site

Age

```
age <- boxplot.stats(advert$Daily.Time.Spent.on.Site)$out
age
```

```
## numeric(0)
```

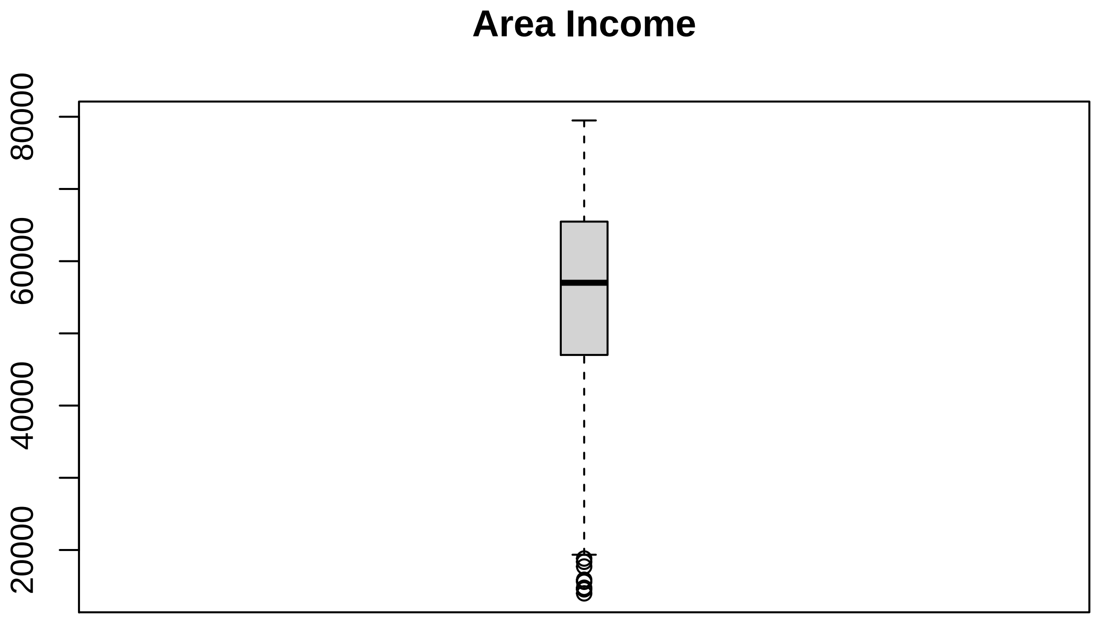
No outliers in age column

Area.income

```
areaincome <- boxplot.stats(advert$Area.Income)$out
areaincome
```

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

```
boxplot(advert$Area.Income, main="Area Income", boxwex=0.1)
```



Area income contains outliers.

Daily internet usage

```
net_usage <- boxplot.stats(advert$Daily.Internet.Usage)$out
net_usage
```

```
## numeric(0)
```

No outliers in internet usage.

UNIVARIATE ANALYSIS

MEAN

clicked on ad

```
mean(advert$Clicked.on.Ad)
```

```
## [1] 0.5
```

```
mean(advert$Daily.Time.Spent.on.Site)
```

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

```
mean(advert$Age)
```

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

```
mean(advert$Area.Income)
```

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

```
mean(advert$Daily.Internet.Usage)
```

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

The average time spent on a site is 65 minutes.

Most of the people who visit the site are of age 35 years.

The average area income is 55K.

The average internet usage is 180.

MEDIAN

It is the middle most value.

```
median(advert$Clicked.on.Ad)
```

```
## [1] 0.5
```

```
median(advert$Daily.Time.Spent.on.Site)
```

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

```
median(advert$Age)
```

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

```
median(advert$Area.Income)
```

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

```r
median(advert$Daily.Internet.Usage)
```

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

MODE

The mode is the value that has highest number of occurrences in a set of data. Created using a function as it does not have a builtin function.

Timestamp

```r
getmode <- function(v) {
   uniqv <- unique(advert$Timestamp)
   uniqv[which.max(tabulate(match(advert$Timestamp, uniqv)))]
}

result <- getmode(v)
print(result)
```

```
## [1] "2016-03-27 00:53:11"
```

```r
getmode <- function(v) {
   uniqv <- unique(advert$Ad.Topic.Line)
   uniqv[which.max(tabulate(match(advert$Ad.Topic.Line, uniqv)))]
}

result <- getmode(v)
print(result)
```

```
## [1] "Cloned 5thgeneration orchestration"
```

The topic 'Cloned 5thgeneration orchestration' occurred multiple times.

```r
getmode <- function(v) {
   uniqv <- unique(advert$City)
   uniqv[which.max(tabulate(match(advert$City.Line, uniqv)))]
}

result <- getmode(v)
print(result)
```

```
## [1] "Wrightburgh"
```

Wrightburgh city occured most times in the data.

Range

Range is a function that produces the smallest and largest values.

```r
range(advert$Age)
```

```
## [1] 19 61
```

```r
range(advert$Clicked.on.Ad)
```

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

```r
range(advert$Daily.Time.Spent.on.Site)
```

```
## [1] 32.60 91.43
```

```r
range(advert$Age)
```

```
## [1] 19 61
```

```r
range(advert$Area.Income)
```

```
## [1] 13996.5 79484.8
```

```r
range(advert$Daily.Internet.Usage)
```

```
## [1] 104.78 269.96
```

MAXIMUM

```r
max(advert$Age)
```

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

```r
max(advert$Daily.Time.Spent.on.Site)
```

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

```r
max(advert$Area.Income)
```

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

```r
max(advert$Daily.Internet.Usage)
```

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

The maximum age of a person who watched the advertisment was 61 years.

The maximum time spent on the advertisment is 91 minutes 43 seconds.

The maximum income is 79484.8

The maximum internet usage on the advert is 269.96

MINIMUM

```
min(advert$Age)
```

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

```
min(advert$Daily.Time.Spent.on.Site)
```

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

```
min(advert$Area.Income)
```

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

```
min(advert$Daily.Internet.Usage)
```

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

The minimum age of a person who visited the site is 19 years.

The minimum time spent on the site is 32min 6secs.

The minimum income earned is 13996.5

The minimum internet used is 104.78.

VARIANCE

Is a numerical measure of how the data values is dispersed around the mean.

```
var(advert$Age)
```

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

```
var(advert$Daily.Time.Spent.on.Site)
```

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

```
var(advert$Area.Income)
```

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

```
var(advert$Daily.Internet.Usage)
```

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

STANDARD DEVIATION

```r
sd(advert$Age)
```

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

```r
sd(advert$Daily.Time.Spent.on.Site)
```

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

```r
sd(advert$Area.Income)
```

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

```r
sd(advert$Daily.Internet.Usage)
```

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

INTERQUARTILE RANGE

```r
IQR(advert$Age)
```

```
## [1] 13
```

```r
IQR(advert$Daily.Time.Spent.on.Site)
```

```
## [1] 27.1875
```

```r
IQR(advert$Area.Income)
```

```
## [1] 18438.83
```

```r
IQR(advert$Daily.Internet.Usage)
```

```
## [1] 79.9625
```

COUNT PLOTS

Determine which gender mainly visited the sites

```r

library(dplyr)
```

Male

```
##
## Attaching package: ’dplyr’

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

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

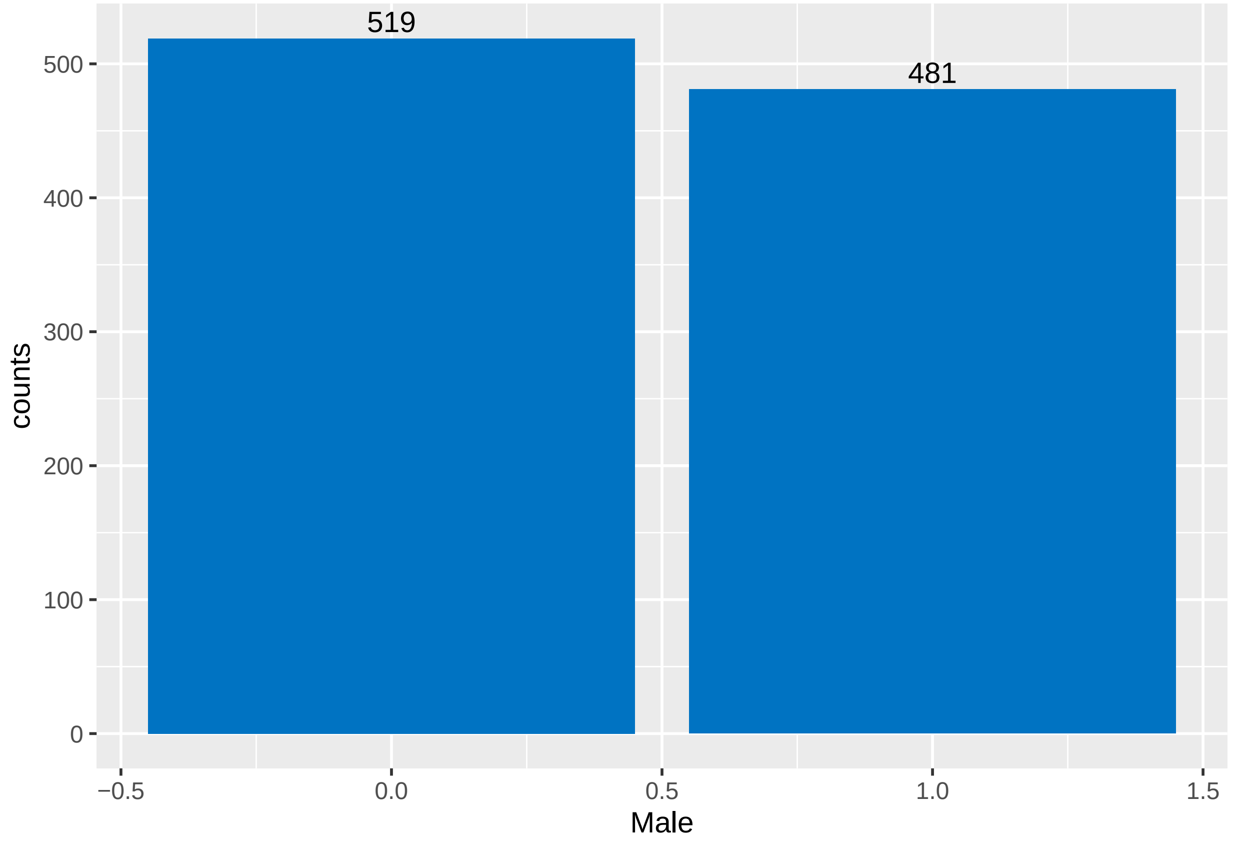
```
df <- advert %>%
  group_by(Male) %>%
  summarise(counts = n())
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
df
```

```
## # A tibble: 2 x 2
##    Male counts
##   <int>  <int>
## 1     0    519
## 2     1    481
```

```
library(’ggplot2’)
ggplot(df, aes(x = Male, y = counts)) +
  geom_bar(fill = "#0073C2FF", stat = "identity") +
  geom_text(aes(label = counts), vjust = -0.3)
```



0 - Female 1 - Male

High numbers of females who visited the site compared to males

Country

```
cntry <- advert %>%
  group_by(Country) %>%
  summarise(counts = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
cntry
```

```
## # A tibble: 237 x 2
##    Country                                      counts
##    <chr>                                         <int>
##  1 Afghanistan                                       8
##  2 Albania                                           7
##  3 Algeria                                           6
##  4 American Samoa                                    5
##  5 Andorra                                           2
##  6 Angola                                            4
##  7 Anguilla                                          6
##  8 Antarctica (the territory South of 60 deg S)       3
##  9 Antigua and Barbuda                                5
## 10 Argentina                                          2
## # ... with 227 more rows
```

Afghanistan had more people who visted the site.

```
city <- advert %>%
  group_by(City) %>%
  summarise(counts = n())
```

city

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
city
```

```
## # A tibble: 969 x 2
##    City          counts
##    <chr>          <int>
##  1 Adamsbury          1
##  2 Adamside           1
##  3 Adamsstad          1
##  4 Alanview           1
##  5 Alexanderfurt      1
##  6 Alexanderview      1
##  7 Alexandrafort      1
##  8 Alexisland         1
##  9 Aliciatown         1
## 10 Alvaradoport       1
## # ... with 959 more rows
```

```
ad <- advert %>%
  group_by(Clicked.on.Ad) %>%
  summarise(counts = n())
```

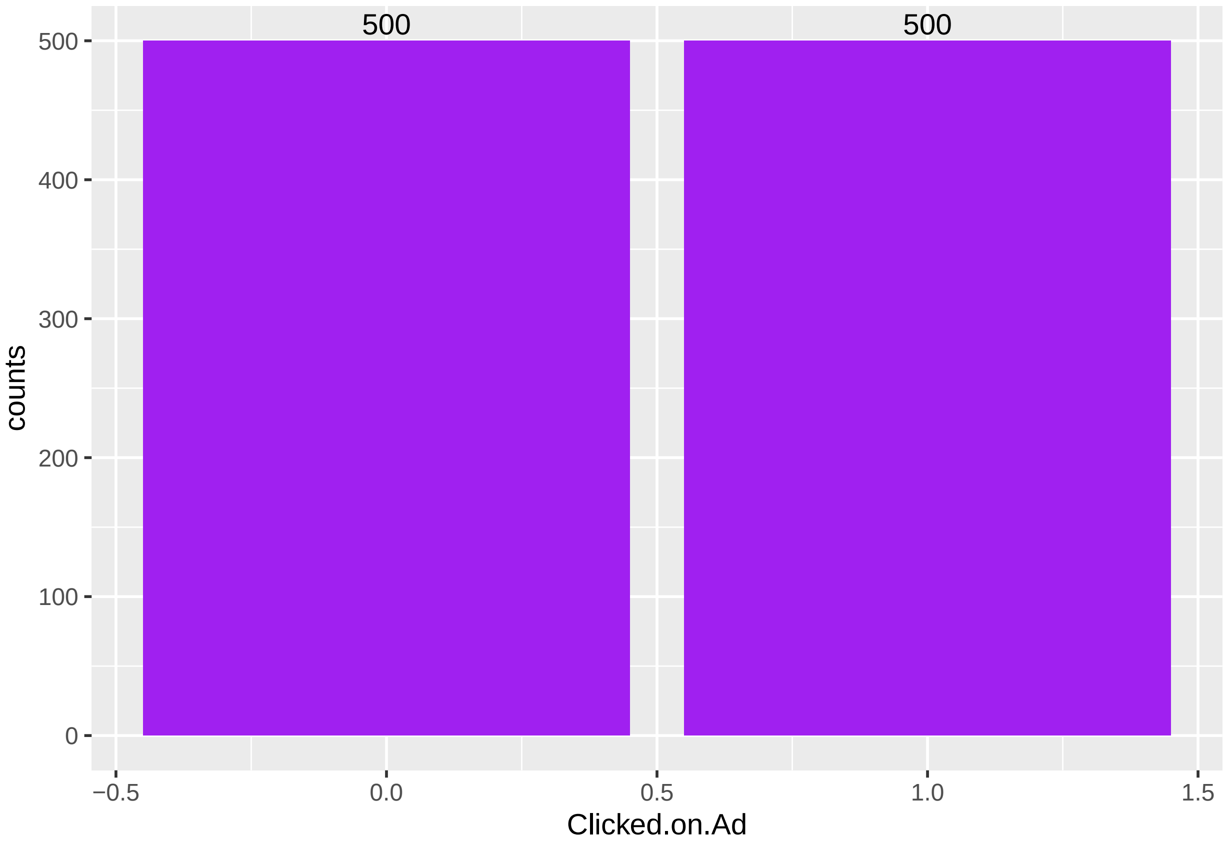
clicked on Ad

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
ad
```

```
## # A tibble: 2 x 2
##   Clicked.on.Ad counts
##           <int>  <int>
## 1             0    500
## 2             1    500
```

```
library('ggplot2')
ggplot(ad, aes(x = Clicked.on.Ad, y = counts)) +
  geom_bar(fill = "Purple", stat = "identity") +
  geom_text(aes(label = counts), vjust = -0.3)
```



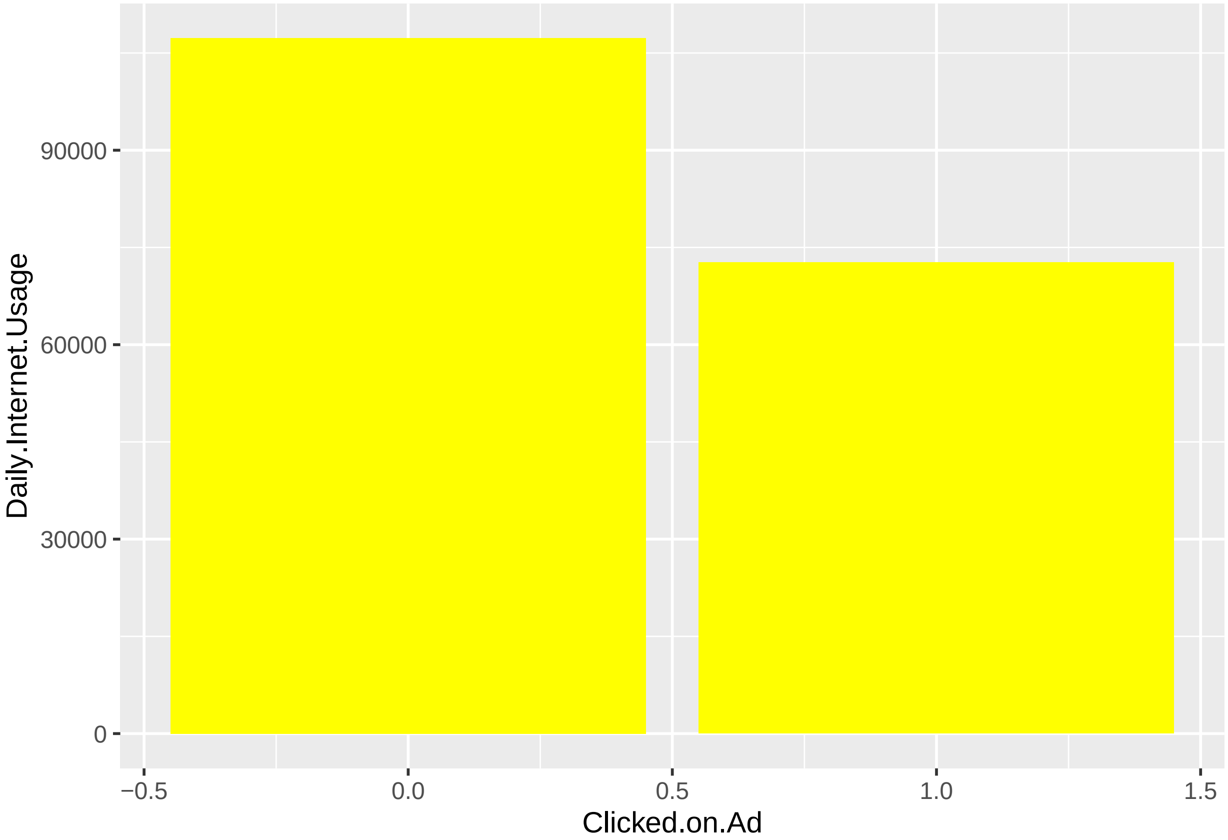
0 - did not click 1 - clicked

There was an equal number of people who clicked and who did not click on the advertisment

BIVARIATE ANALYSIS

clicked on ad vs internet usage

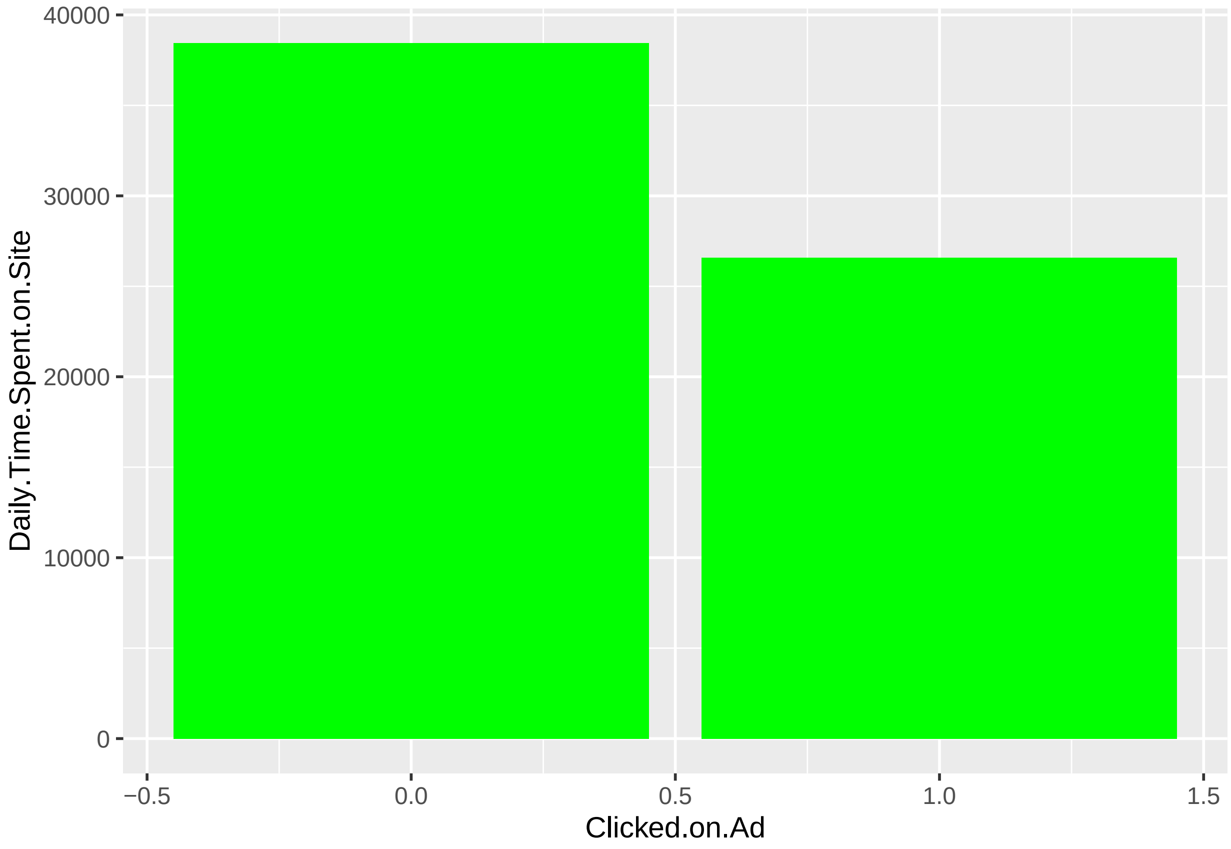
```
ggplot(advert, aes(x=Clicked.on.Ad, y=Daily.Internet.Usage)) +
  geom_bar(fill = 'Yellow' ,stat = "identity")
```



Those who did not click on the advertisment used more internet compared to those who clicked.

clicked on ad vs daily time spent on site

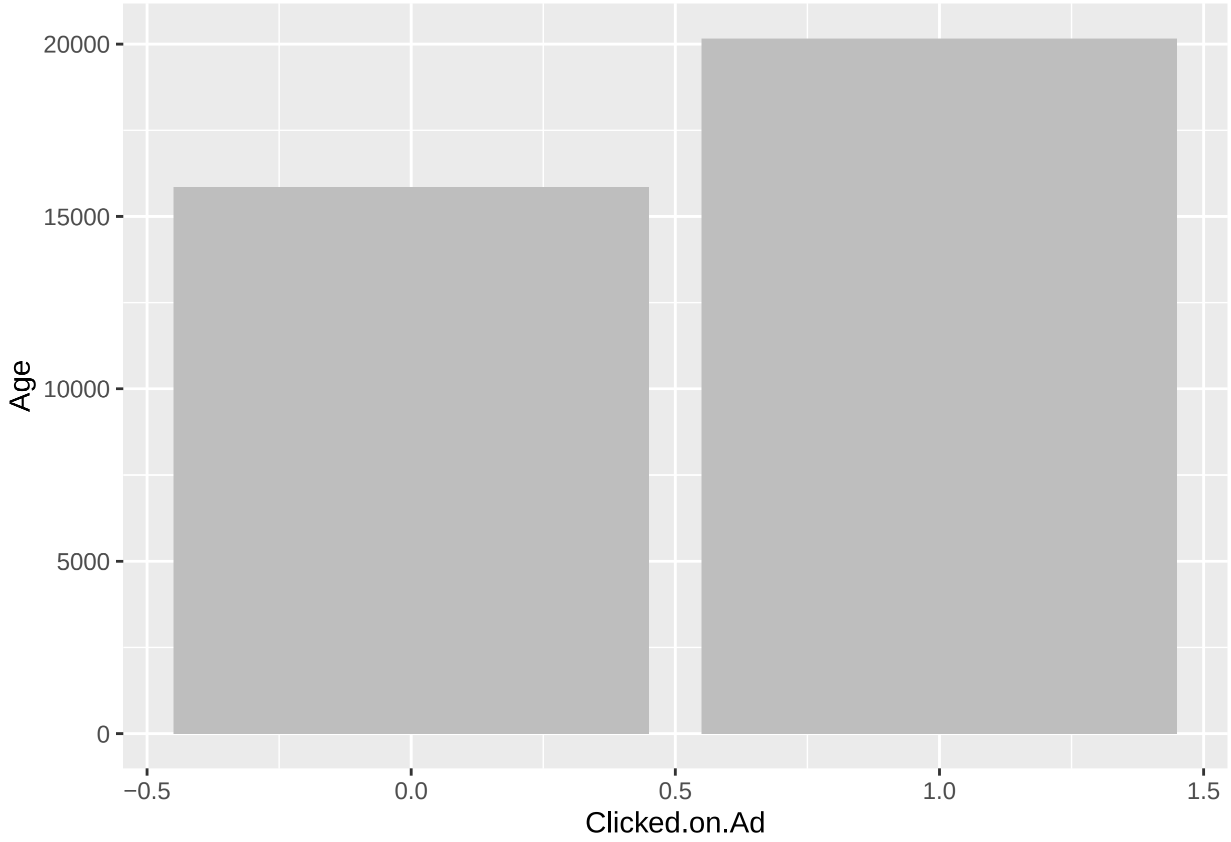
```
ggplot(advert, aes(x=Clicked.on.Ad, y=Daily.Time.Spent.on.Site)) +
  geom_bar(fill = 'Green', stat = "identity")
```



Those who did not click on the Ad spent a lot of time on the site compared to those who clicked.

clicked on ad vs Age

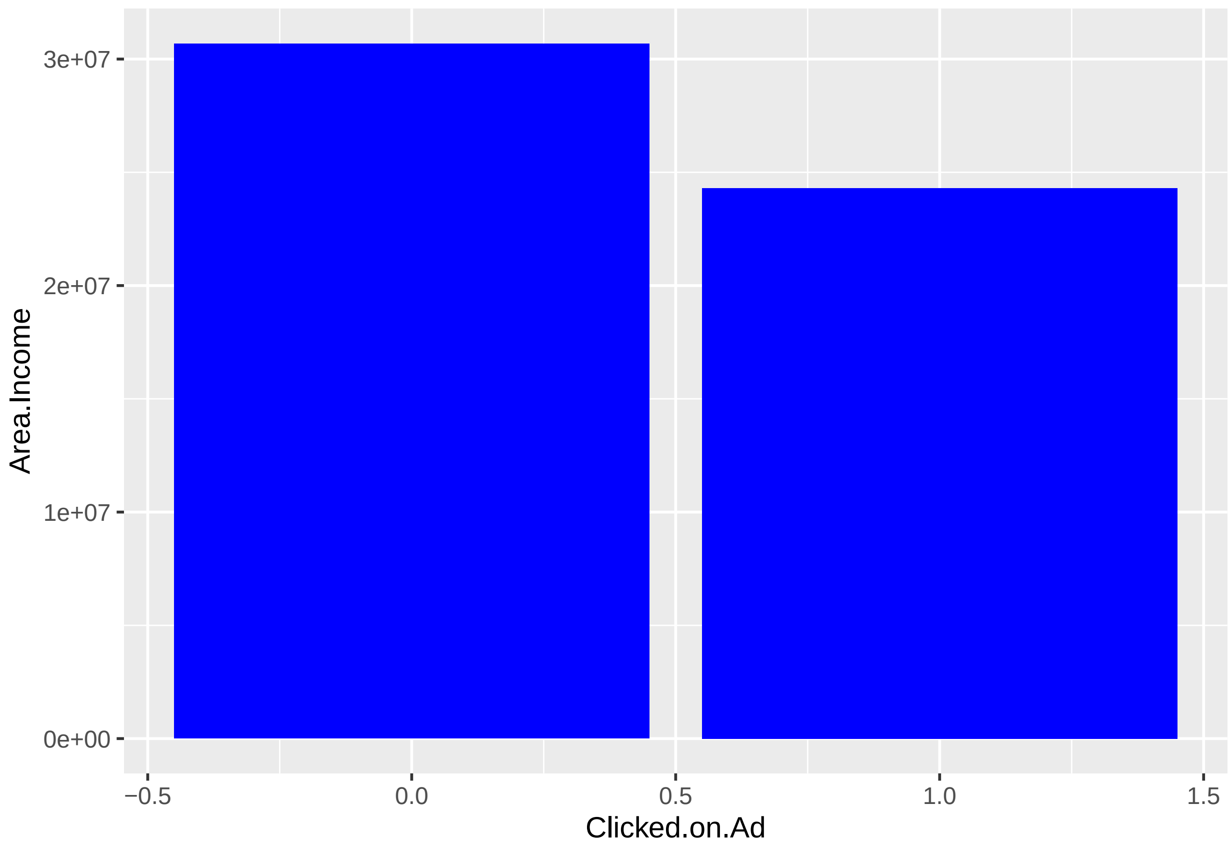
```
ggplot(advert, aes(x=Clicked.on.Ad, y=Age)) +
  geom_bar(fill = 'Grey', stat = "identity")
```



The older people are more likely to click on the Ads

clicked on ad vs Area.Income

```
ggplot(advert, aes(x=Clicked.on.Ad, y=Area.Income)) +
  geom_bar(fill = 'Blue',stat = "identity")
```



Those who did not click on the Ads had more income compared to those who clicked the Ads.

COVARIANCE

Covariance measures how two variables are linearly related.

```
Ad_click <- advert$Clicked.on.Ad
income <- advert$Area.Income
cov(Ad_click, income)
```

```
## [1] -3195.989
```

The clicked on Ad and income have a negative relationship

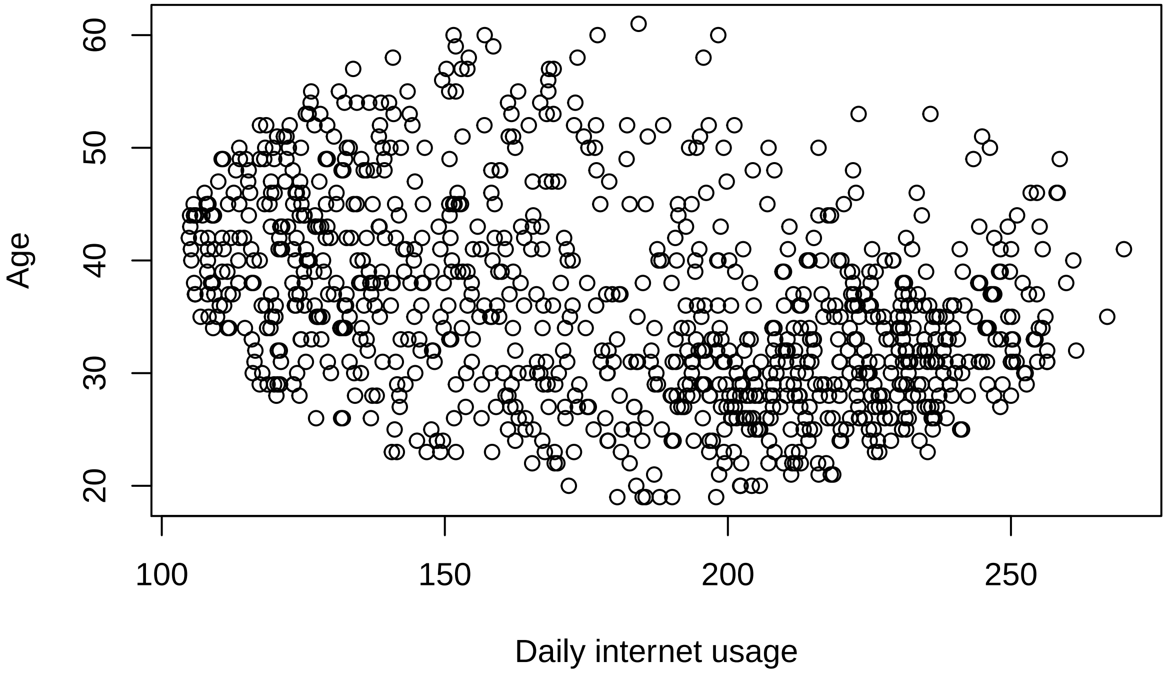
```
Ad_click <- advert$Clicked.on.Ad
net <- advert$Daily.Internet.Usage
cov(Ad_click, net)
```

```
## [1] -17.27409
```

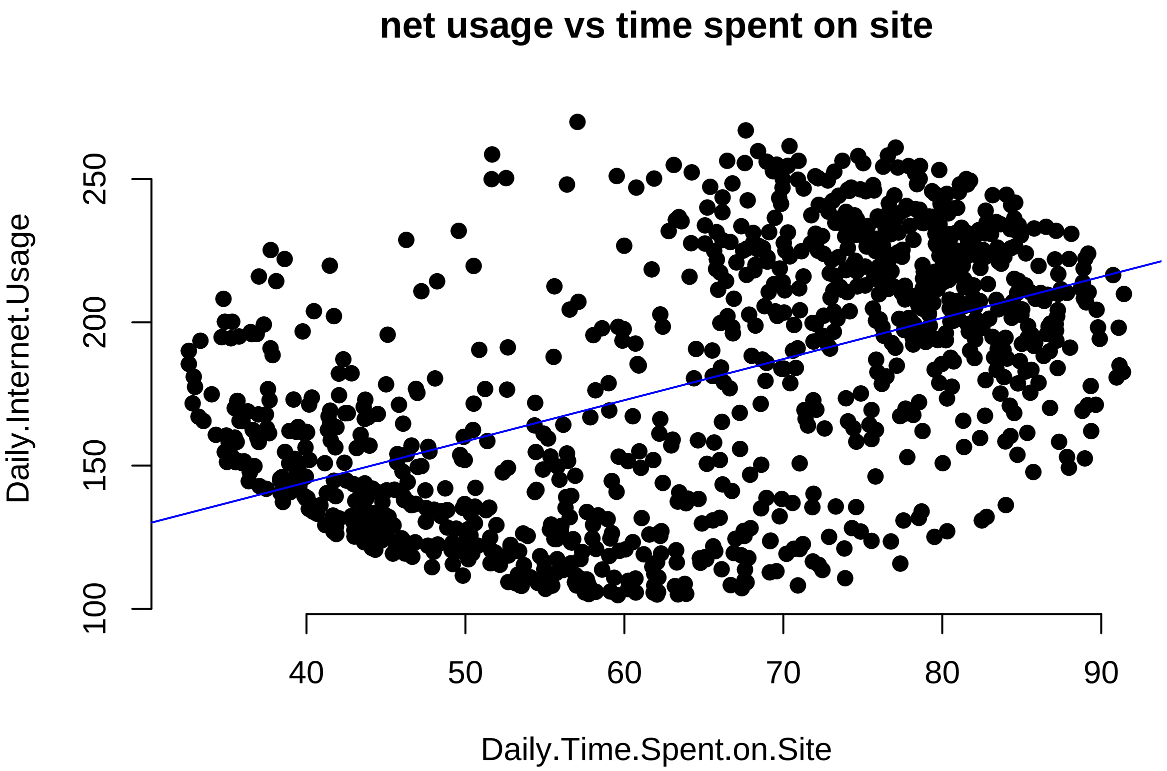
The clicked on Ad and internet usage have a negative relationship

SCATTER PLOTS

```
plot(x = advert$Daily.Internet.Usage,y = advert$Age,
     xlab=’Daily internet usage’,
     ylab = ’Age’)
```

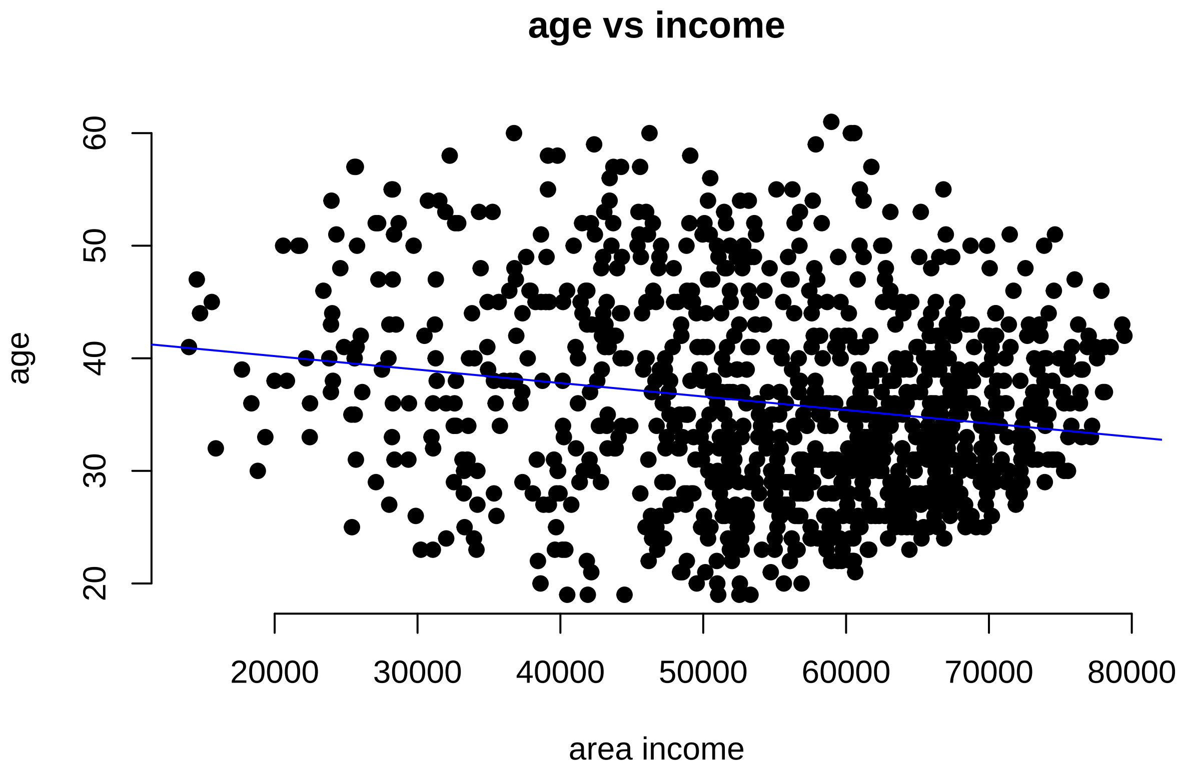


```
plot(advert$Daily.Time.Spent.on.Site, advert$Daily.Internet.Usage, main = "net usage vs time spent on si
     xlab = "Daily.Time.Spent.on.Site", ylab = "Daily.Internet.Usage",
     pch = 19, frame = FALSE)
abline(lm(advert$Daily.Internet.Usage ~ advert$Daily.Time.Spent.on.Site, data = advert), col = "blue")
```



The plot shows a positive relationship between daily net usage and time spent on site

```
plot(advert$Area.Income, advert$Age, main = "age vs income",
     xlab = "area income", ylab = "age",
     pch = 19, frame = FALSE)
abline(lm(advert$Age ~ advert$Area.Income, data = advert), col = "blue")
```



The plot shows a negative correlation between age and income.

CONCLUSION

From the analysis, females are more likely to click the Ads, people living in Afghanistan, those who spent less time on the site are ,ore likely to click on the advertisment.

RECCOMMENDATIONS

The entrepreneur should consider target more females as the are more likely to click on the Ads. She should also target older people, people living in Afghanistan, people with low income and those who internet usage is low.