Crypto_Course_Advertisement

Kennedy Njoroge

28/02/2020

Business Understanding

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.

Specifying the question

Identify which individuals are most likely to click on her ads based on data collected in the past.

Metric for success

- Outliers, Anomalies and missing data.
- Univariate and bivariate analysis

Understanding the context

Internet adverstising seeks to deliver promotional marketing materials to consumers. Analysis of target audience is necessary so as to reach the right audience who will see conversion of advert to an order.

Recording the experimental design

- Business Understanding
- Data importation and understanding
- Exploratory Data Analysis
- Conclusion

Import Libraries

```
#Impor Latex to facilitate PDF export.
#tinytex::install_tinytex()
#install.packages("tidyverse",dependencies = TRUE)
#library(tidyverse)
```

Exploratory Data Analysis

Import the data

Check Structure of data frame - name, type and preview of data in each column

```
str(df_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           : Factor w/ 1000 levels "Adaptive 24hour Graphic Interface",..: 92 465 567
##  $ City                    : Factor w/ 969 levels "Adamsbury","Adamside",..: 962 904 112 940 806 283
##  $ Male                    : int  0 1 0 1 0 1 0 1 1 1 ...
##  $ Country                 : Factor w/ 237 levels "Afghanistan",..: 216 148 185 104 97 159 146 13 83
##  $ Timestamp               : Factor w/ 1000 levels "2016-01-01 02:52:10",..: 440 475 368 57 768 690
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

Columns are of data type numeric, integers and Factors.

- Numeric Daily.Time.Spent.on.Site, Area.Income, Daily.Internet.Usage : They are numeric as their values are numbers which have decimals.
- Integer Age, Male, Cicked.on.Ad, : Integer as it has whole numbers with no fractions.
- Factors Ad.Topic.Line, City, Country, Timestamp : Are all factors. Ad.Topic.Line and timestamp both have 1000 levels meaning it's distinct values per column. City has 969 levels. Country has 237 meaning data is from 237 countries.

Check the Columns and rows of the dataframe

```
dim(df_advert)
```

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

1000 rows and 10 columns.

Check Null Values

```
#Count the missing values
colSums(is.na(df_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
```

No missing values exist in the datasets as all columns are of value zero.

Check Duplicates

```r
anyDuplicated(df_advert)
```

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

No duplicates exist.

Dataframe Summary Description

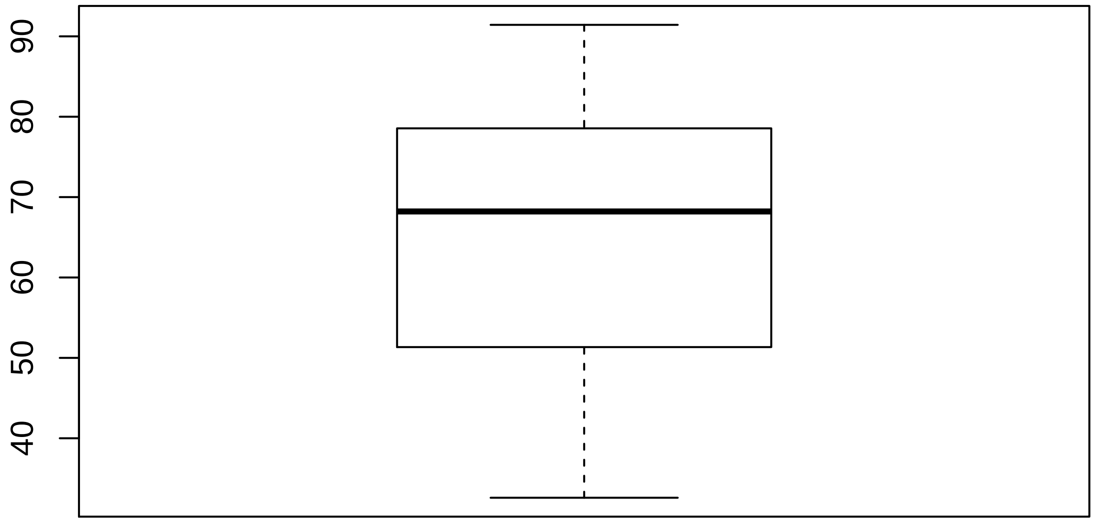
```r
summary(df_advert)
```

```
##  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
##  Adaptive 24hour Graphic Interface      :  1   Lisamouth      :  3
##  Adaptive asynchronous attitude         :  1   Williamsport   :  3
##  Adaptive context-sensitive application :  1   Benjaminchester:  2
##  Adaptive contextually-based methodology:  1   East John      :  2
##  Adaptive demand-driven knowledgebase   :  1   East Timothy   :  2
##  Adaptive uniform capability            :  1   Johnstad       :  2
##  (Other)                                :994   (Other)        :986
##       Male                    Country                         Timestamp    Clicked.on.Ad
##  Min.   :0.000   Czech Republic:  9   2016-01-01 02:52:10:  1   Min.   :0.0
##  1st Qu.:0.000   France        :  9   2016-01-01 03:35:35:  1   1st Qu.:0.0
##  Median :0.000   Afghanistan   :  8   2016-01-01 05:31:22:  1   Median :0.5
##  Mean   :0.481   Australia     :  8   2016-01-01 08:27:06:  1   Mean   :0.5
##  3rd Qu.:1.000   Cyprus        :  8   2016-01-01 15:14:24:  1   3rd Qu.:1.0
##  Max.   :1.000   Greece        :  8   2016-01-01 20:17:49:  1   Max.   :1.0
##                  (Other)       :950   (Other)            :994
```

Outliers

a) Daily Time Spent on Site

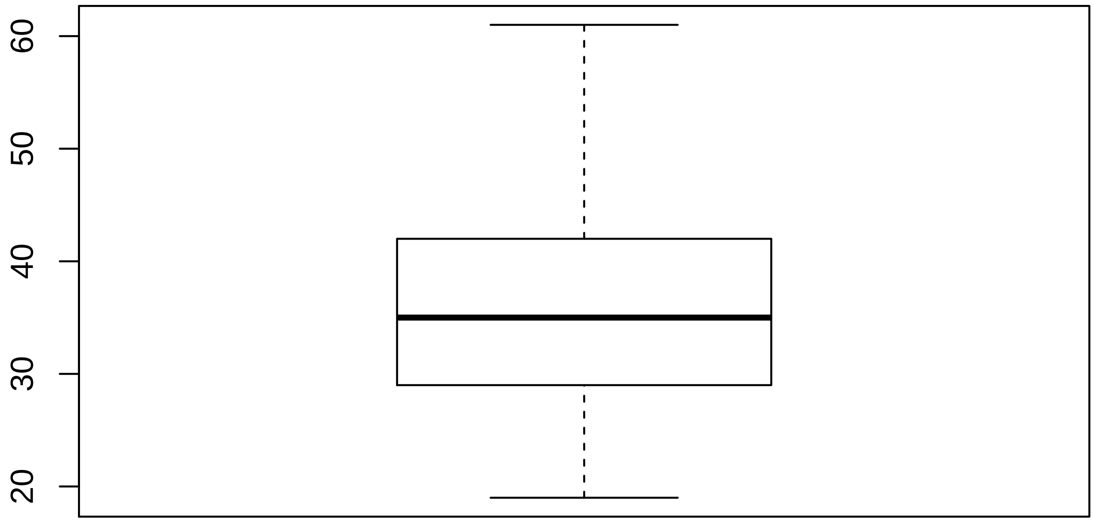
```r
#A = df_advert[c("Daily.Time.Spent.on.Site", "Age", 'Area.Income','Daily.Internet.Usage', 'Male' , 'Cli
boxplot(df_advert["Daily.Time.Spent.on.Site"])
```



No outliers noted in Daily Time spent on site.

b) Age

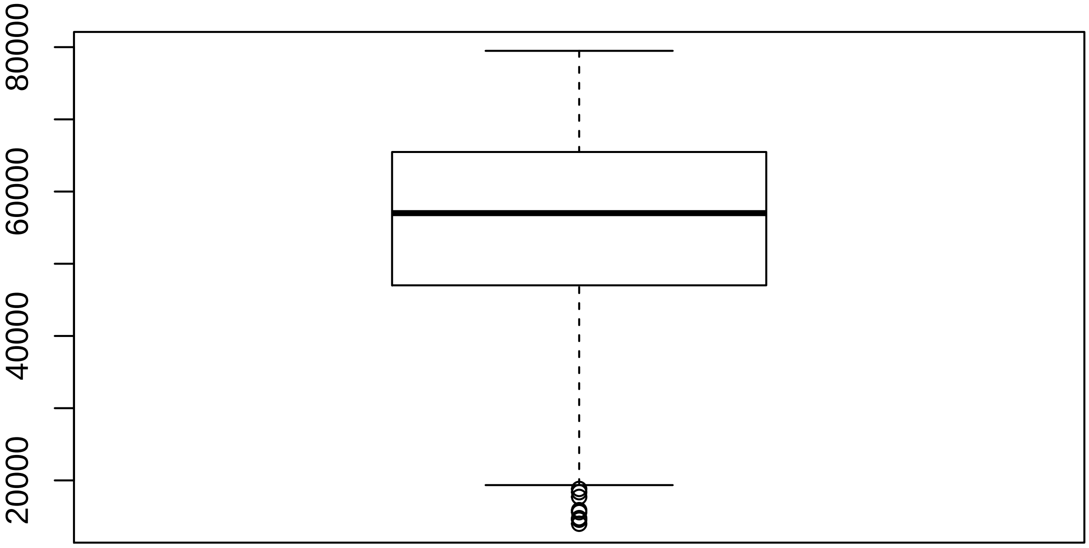
```
boxplot(df_advert["Age"])
```



No outliers noted in age.

c) Area Income

```
boxplot(df_advert['Area.Income'])
```



Outliers noted.

Display of outlier values

```
boxplot.stats(df_advert$Area.Income)$out
```

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

View other dataframe values with outliers

```
df_advert[df_advert$Area.Income %in% c(17709.98,18819.34,15598.29,15879.10,14548.06,13996.50,14775.50,1
```

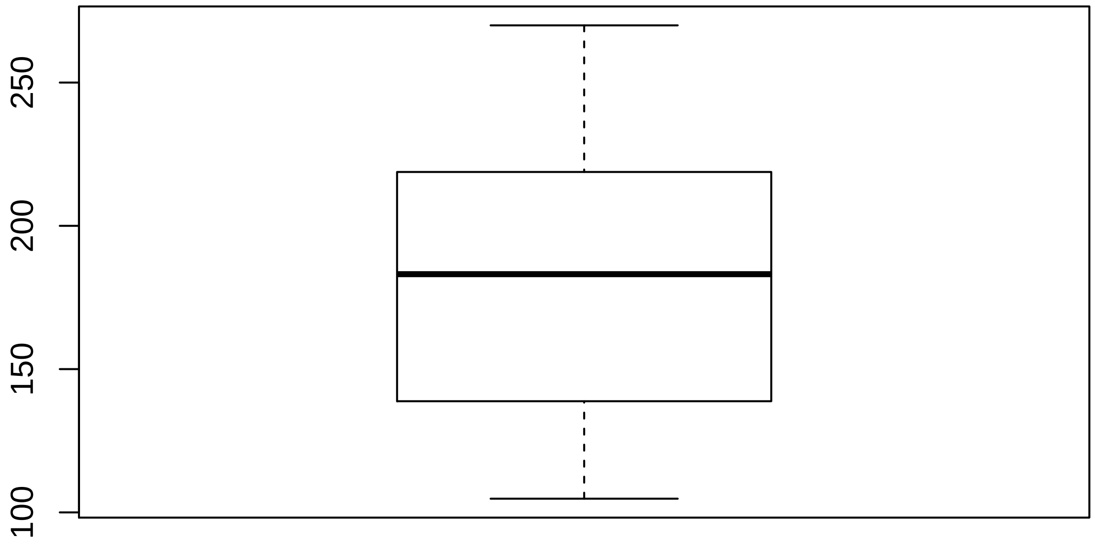
```
##     Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 136                    49.89  39    17709.98               160.03
## 511                    57.86  30    18819.34               166.86
## 641                    64.63  45    15598.29               158.80
## 666                    58.05  32    15879.10               195.54
## 693                    66.26  47    14548.06               179.04
## 769                    68.58  41    13996.50               171.54
## 779                    52.67  44    14775.50               191.26
## 953                    62.79  36    18368.57               231.87
##                                 Ad.Topic.Line          City Male
## 136          Enhanced system-worthy application  East Michele    1
## 511                  Horizontal modular success     Estesfurt    0
## 641 Triple-buffered high-level Internet solution  Isaacborough    1
## 666               Total asynchronous architecture   Sanderstown    1
## 693               Optional full-range projection      Matthewtown    1
## 769                  Exclusive discrete firmware New Williamville    1
## 779     Persevering 5thgeneration knowledge user    New Hollyberg    0
## 953                       Total coherent archive        New James    1
##         Country           Timestamp Clicked.on.Ad
## 136      Belize 2016-04-16 12:09:25             1
## 511     Algeria 2016-07-08 17:14:01             1
## 641  Azerbaijan 2016-06-12 03:11:04             1
## 666  Tajikistan 2016-02-12 10:39:10             1
## 693     Lebanon 2016-04-25 19:31:39             1
## 769 El Salvador 2016-07-06 12:04:29             1
## 779      Jersey 2016-05-19 06:37:38             1
## 953  Luxembourg 2016-05-30 20:08:51             1
```

```r
#c(17709.98,18819.34,15598.29,15879.10,14548.06,13996.50,14775.50,18368.57)
```

The low income areas are noted to be cities in Belize, Algeria, Azerbaijan, Tajikistan, Lebanon, El Salvador, Jersey and Luxembourg. These are not developed countries hence it's understandable why there are low income outliers. Therebeing, the records will be maintained due to their validity.

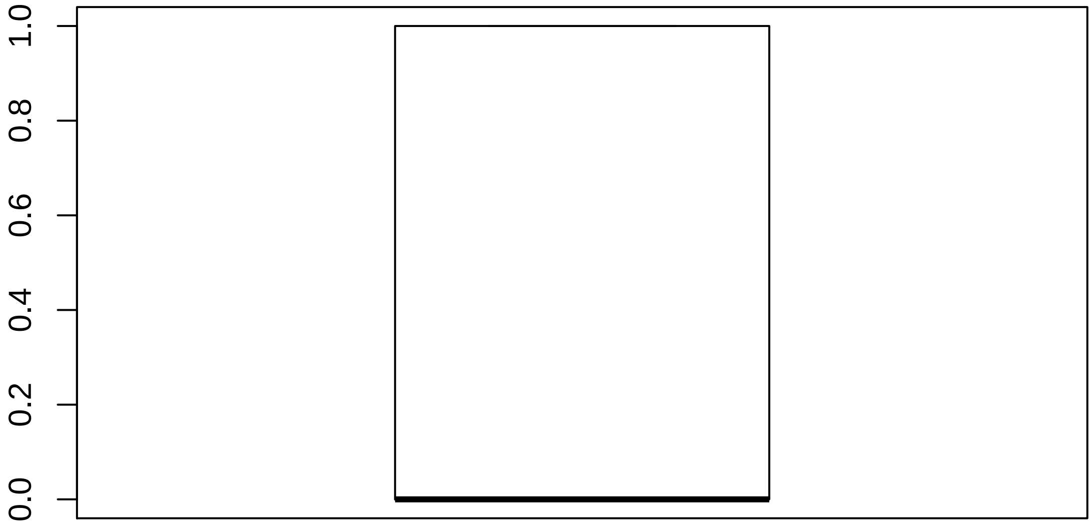
d) Daily Internet Usage

```r
boxplot(df_advert["Daily.Internet.Usage"])
```



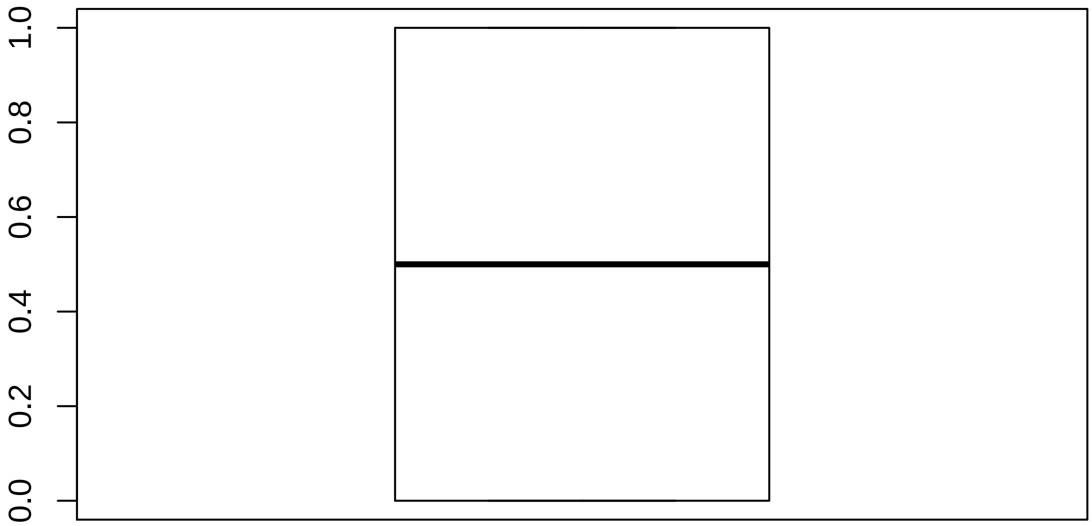
e) Male

```
boxplot(df_advert["Male"])
```



f) Clicked.on.Ad

```
boxplot(df_advert["Clicked.on.Ad"])
```



Feature Engineering

```
df_advert$day <- format(as.POSIXct(strptime(df_advert$Timestamp,"%Y-%m-%d %H:%M:%S",tz="")) ,format = "%
df_advert$month <- format(as.POSIXct(strptime(df_advert$Timestamp,"%Y-%m-%d %H:%M:%S",tz="")) ,format =
```

Bivariate Analysis

Mean

Mean of

```
# Mean of the variables
cat('Age: ',mean(df_advert$Age))
```

```
## Age:  36.009
```

```
cat('\nArea.Income: ',mean(df_advert$Area.Income))
```

```
##
## Area.Income:  55000
```

```
cat('\nDaily Internet Usage: ',mean(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  180.0001
```

```
cat('\nMale: ',mean(df_advert$Male))
```

```
##
## Male:  0.481
```

```
cat('\nClicked.on.Ad: ',mean(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0.5
```

Median

```
# Median of the variables
cat('Age: ',median(df_advert$Age))
```

```
## Age:  35
```

```
cat('\nArea.Income: ',median(df_advert$Area.Income))
```

```
##
## Area.Income:  57012.3
```

```
cat('\nDaily Internet Usage: ',median(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  183.13
```

```
cat('\nMale: ',median(df_advert$Male))
```

```
##
## Male:  0
```

```
cat('\nClicked.on.Ad: ',median(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0.5
```

Mode

```
# Create the function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
cat('Daily.Time.Spent.on.Site: ',getmode(df_advert$Daily.Time.Spent.on.Site))
```

```
## Daily.Time.Spent.on.Site:  62.26
```

```
cat('\nAge: ',getmode(df_advert$Age))
```

```
##
## Age:  31
```

```
cat('\nArea.Income: ',getmode(df_advert$Area.Income))
```

```
##
## Area.Income:  61833.9
```

```
cat('\nDaily Internet Usage: ',getmode(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  167.22
```

```
cat('\nAd Topic Line',getmode(df_advert$Ad.Topic.Line))
```

```
##
## Ad Topic Line 92
```

```
cat('\nCity',getmode(df_advert$City))
```

```
##
## City 427
```

```
cat('\nMale',getmode(df_advert$Male))
```

```
##
## Male 0
```

```
cat('\nCountry',getmode(df_advert$Country))
```

```
##
## Country 55
```

```
cat('\nTimestamp',getmode(df_advert$Timestamp))
```

```
##
## Timestamp 440
```

```r
cat('\nClicked.on.Ad: ',getmode(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0
```

Minimum and Maximum

```r
cat('Min Age: ',min(df_advert$Age))
```

```
## Min Age:  19
```

```r
cat('\nMin Area.Income: ',min(df_advert$Area.Income))
```

```
##
## Min Area.Income:  13996.5
```

```r
cat('\nMin Daily Internet Usage: ',min(df_advert$Daily.Internet.Usage))
```

```
##
## Min Daily Internet Usage:  104.78
```

```r
cat('\nMin Male: ',min(df_advert$Male))
```

```
##
## Min Male:  0
```

```r
cat('\nMin Clicked.on.Ad: ',min(df_advert$Clicked.on.Ad))
```

```
##
## Min Clicked.on.Ad:  0
```

```r
cat('\n')
```

```r
cat('\nMax Age: ',max(df_advert$Age))
```

```
##
## Max Age:  61
```

```r
cat('\nMax Area.Income: ',max(df_advert$Area.Income))
```

```
##
## Max Area.Income:  79484.8
```

```
cat('\nMax Daily Internet Usage: ',max(df_advert$Daily.Internet.Usage))
```

```
##
## Max Daily Internet Usage:  269.96
```

```
cat('\nMax Male: ',max(df_advert$Male))
```

```
##
## Max Male:  1
```

```
cat('\nMax Clicked.on.Ad: ',max(df_advert$Clicked.on.Ad))
```

```
##
## Max Clicked.on.Ad:  1
```

Range, Variance, Quartile, Standard Deviation

```
cat('RANGE: maximum element of the distance \n')
```

```
## RANGE: maximum element of the distance
```

```
cat('Age: ',range(df_advert$Age))
```

```
## Age:  19 61
```

```
cat('\nArea.Income: ',range(df_advert$Area.Income))
```

```
##
## Area.Income:  13996.5 79484.8
```

```
cat('\nDaily Internet Usage: ',range(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  104.78 269.96
```

```
cat('\nMale: ',range(df_advert$Male))
```

```
##
## Male:  0 1
```

```
cat('\nClicked.on.Ad: ',range(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0 1
```

```r
cat('\n\n')
```

```r
cat('VARIANCE: Is a numerical measure of how the data values is dispersed around the mean\n')
```

```
## VARIANCE: Is a numerical measure of how the data values is dispersed around the mean
```

```r
cat('Age: ',var(df_advert$Age))
```

```
## Age:  77.18611
```

```r
cat('\nArea.Income: ',var(df_advert$Area.Income))
```

```
##
## Area.Income:  179952406
```

```r
cat('\nDaily Internet Usage: ',var(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  1927.415
```

```r
cat('\nMale: ',var(df_advert$Male))
```

```
##
## Male:  0.2498889
```

```r
cat('\nClicked.on.Ad: ',var(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0.2502503
```

```r
cat('\n\n')
```

```r
cat('QUARTILE: Lower range, 1st quartile, median, 3rd quartile upper range\n')
```

```
## QUARTILE: Lower range, 1st quartile, median, 3rd quartile upper range
```

```r
cat('Age: ',quantile(df_advert$Age))
```

```
## Age:  19 29 35 42 61
```

```r
cat('\nArea.Income: ',quantile(df_advert$Area.Income))
```

```
##
## Area.Income:  13996.5 47031.8 57012.3 65470.64 79484.8
```

```r
cat('\nDaily Internet Usage: ',quantile(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  104.78 138.83 183.13 218.7925 269.96
```

```r
cat('\nMale: ',quantile(df_advert$Male))
```

```
##
## Male:  0 0 0 1 1
```

```r
cat('\nClicked.on.Ad: ',quantile(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0 0 0.5 1 1
```

```r
cat('\n\n')
```

```r
cat('STANDARD DEVIATION: Deviation from the mean\n')
```

```
## STANDARD DEVIATION: Deviation from the mean
```

```r
cat('Age: ',sd(df_advert$Age))
```

```
## Age:  8.785562
```

```r
cat('\nArea.Income: ',sd(df_advert$Area.Income))
```

```
##
## Area.Income:  13414.63
```

```r
cat('\nDaily Internet Usage: ',sd(df_advert$Daily.Internet.Usage))
```

```
##
## Daily Internet Usage:  43.90234
```

```r
cat('\nMale: ',sd(df_advert$Male))
```

```
##
## Male:  0.4998889
```

```r
cat('\nClicked.on.Ad: ',sd(df_advert$Clicked.on.Ad))
```

```
##
## Clicked.on.Ad:  0.5002502
```

Tabular Frequency Distribution

Selected a few columns which do not have so high distinct distribution

```
cat("Age")
```

```
## Age
```

```
table(df_advert$Age)
```

```
## 
## 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 
##  6  6  6 13 19 21 27 37 33 48 48 39 60 38 43 39 39 50 36 37 30 36 32 26 23 21 
## 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 
## 30 18 13 16 18 20 12 15 10  9  7  2  6  4  2  4  1
```

```
cat("\nMale")
```

```
## 
## Male
```

```
table(df_advert$Male)
```

```
## 
##   0   1 
## 519 481
```

```
cat('\nClicked on Ad')
```

```
## 
## Clicked on Ad
```

```
table(df_advert$Clicked.on.Ad)
```

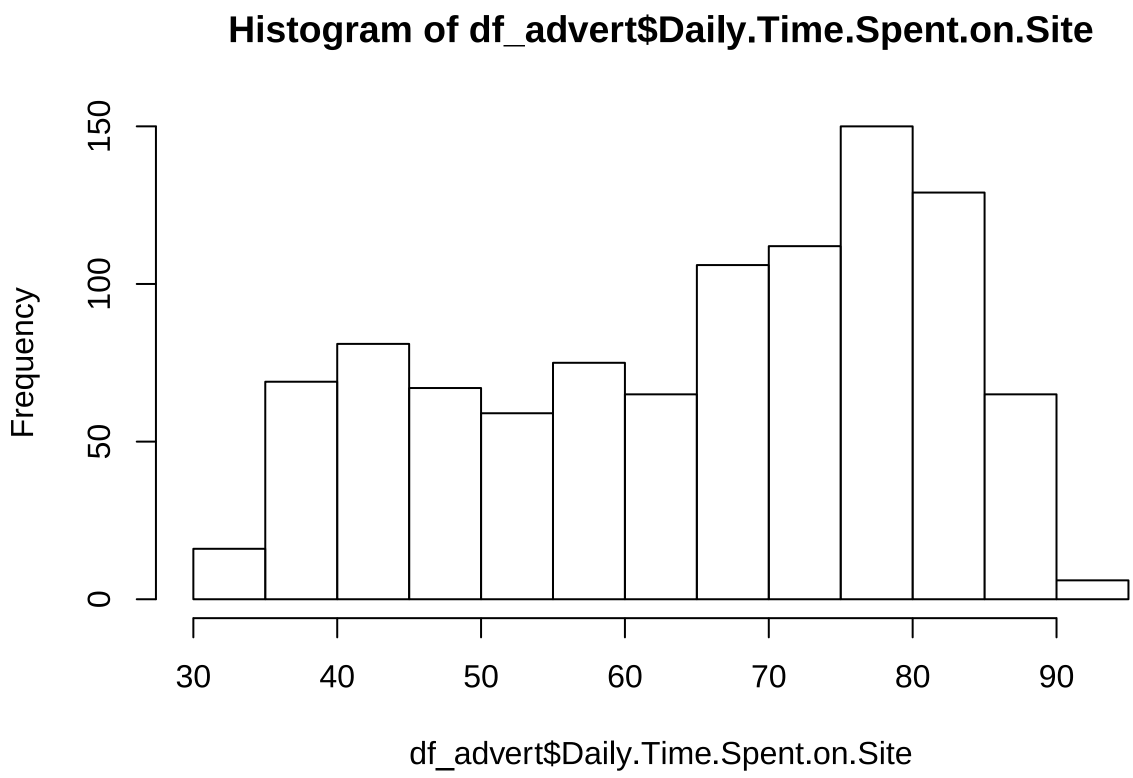
```
## 
##   0   1 
## 500 500
```

Age 31 has the highest distribution of 60 people. The dataset also has more non males than males. There is equal distribution of people who clicked on the add and those that did not click.

Histogram Frequency Distribution

a) Daily Time Spent on Site

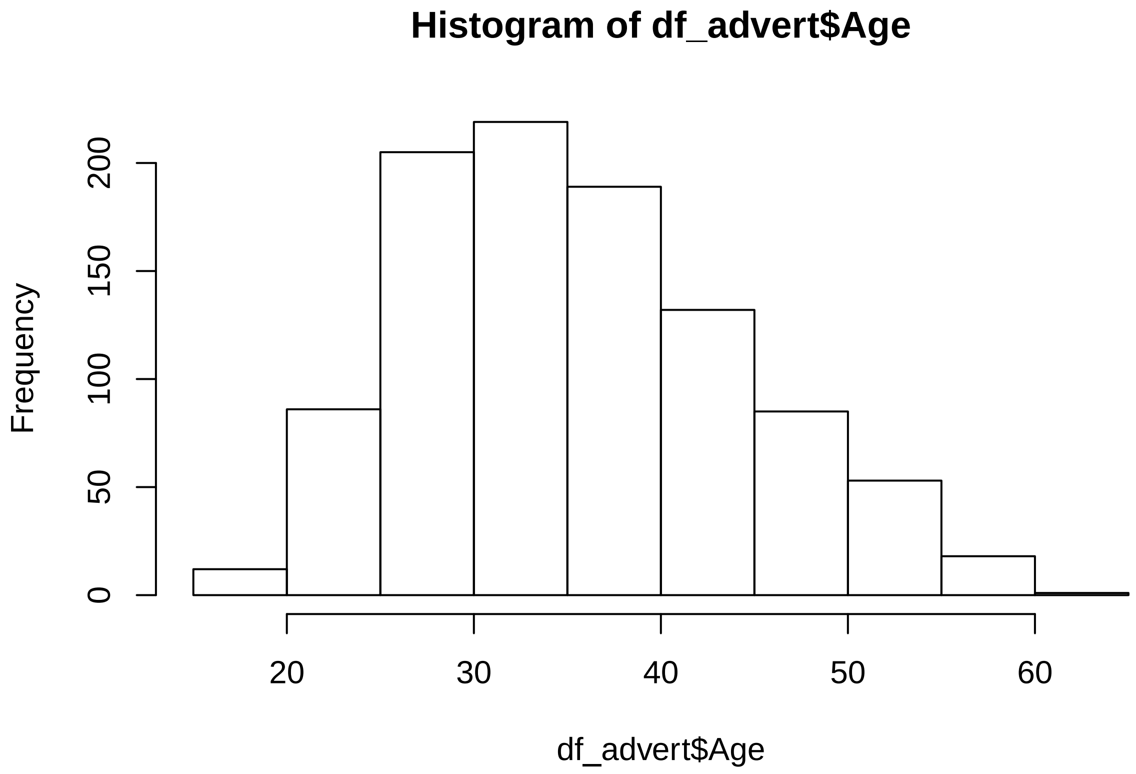
```
hist(df_advert$Daily.Time.Spent.on.Site)
```



The graph is skewed to the left. More people spent time on the website.

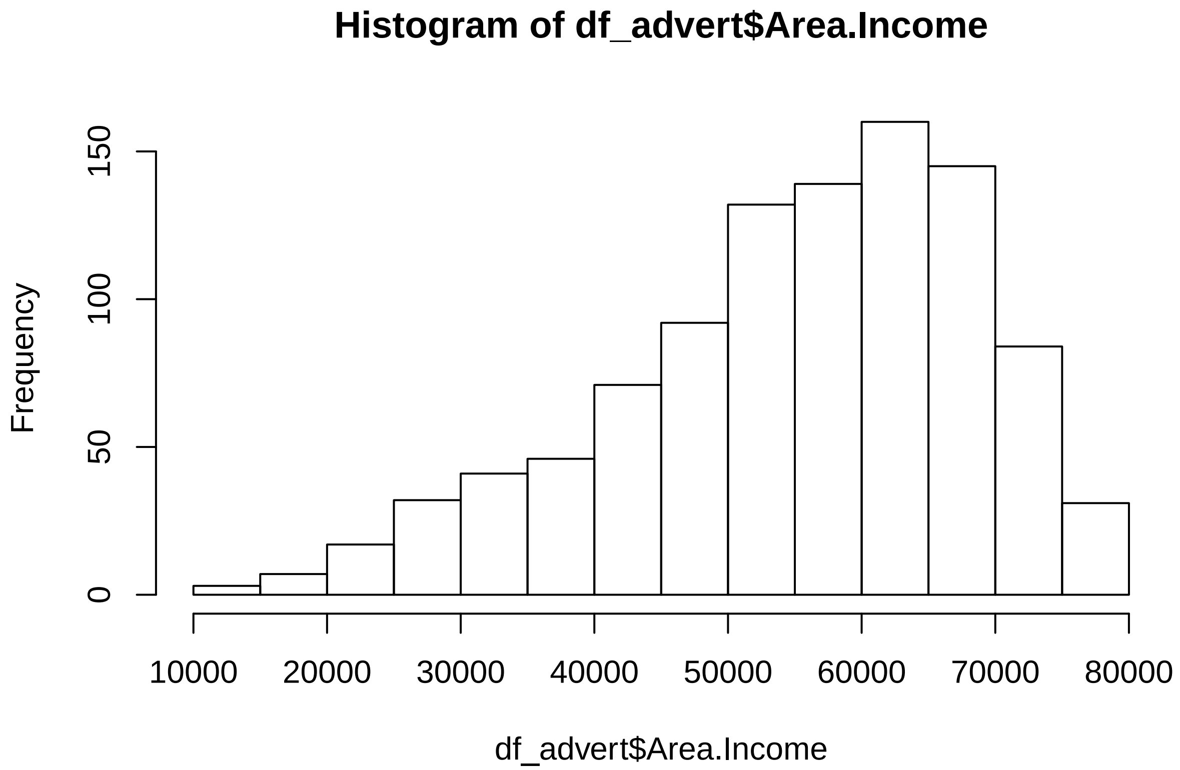
b) Age

```
hist(df_advert$Age)
```



b) Area Income

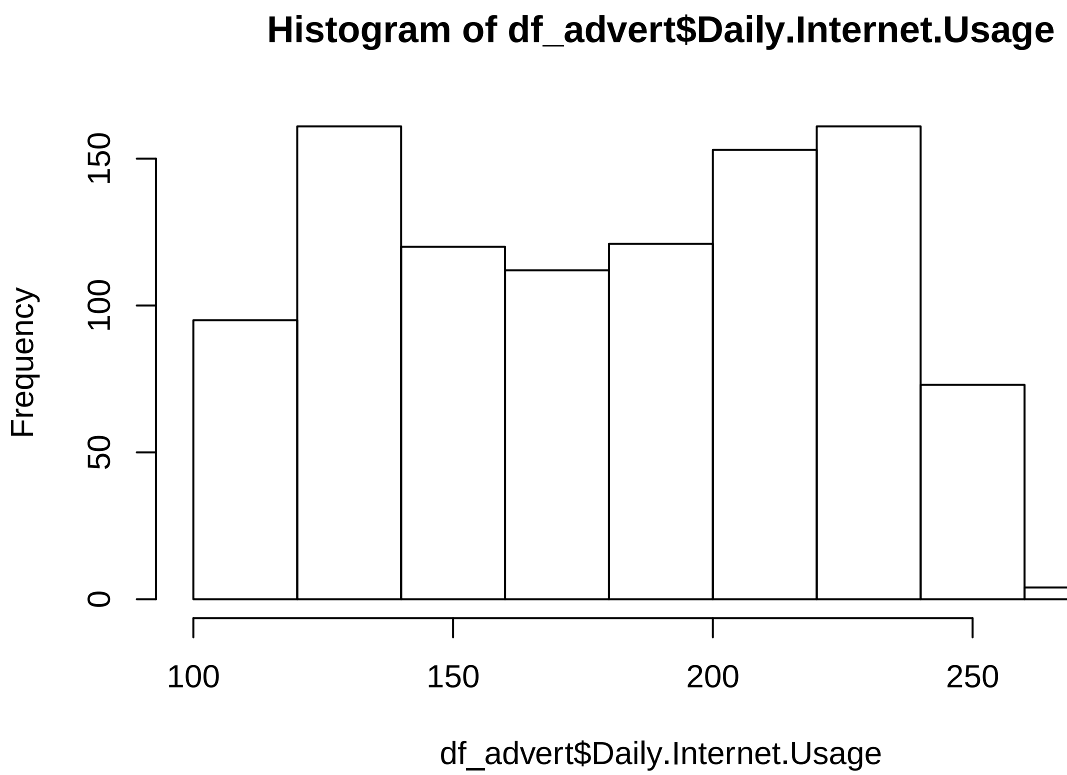
```
hist(df_advert$Area.Income)
```



Income is skewed to the left. Most of the people in the dataset have high income.

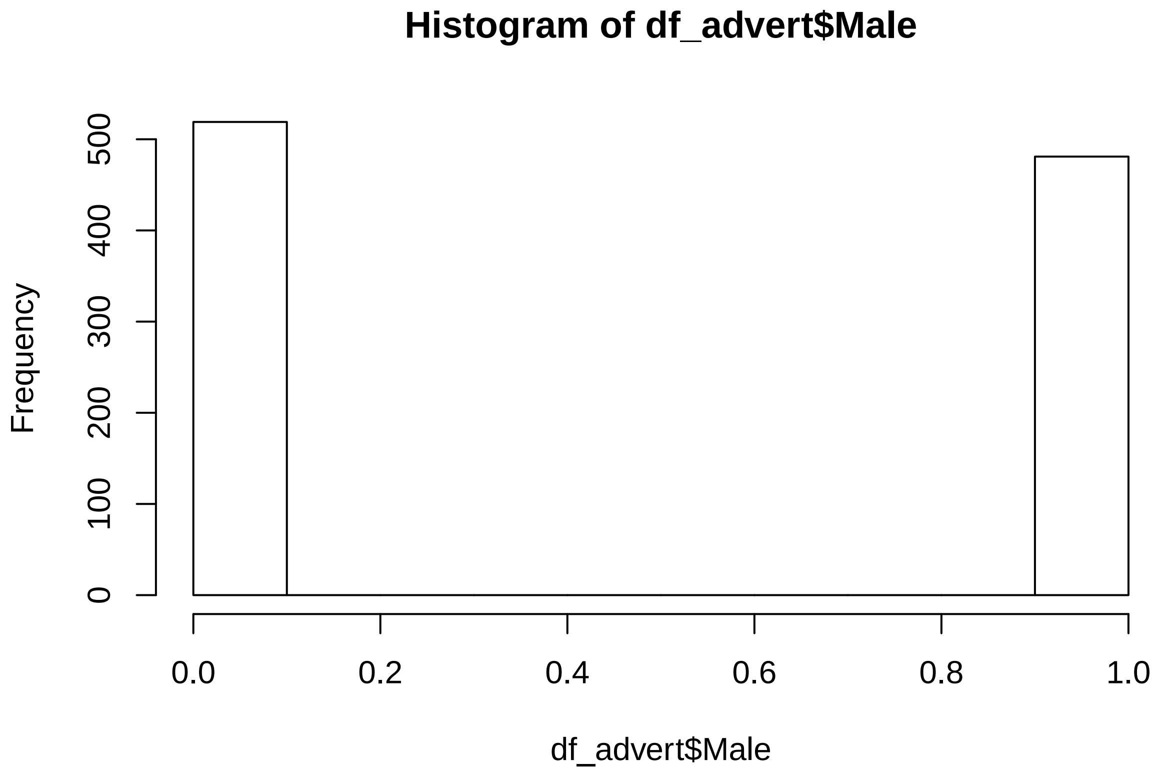
d) Daily.Internet.Usage

```r
hist(df_advert$Daily.Internet.Usage)
```



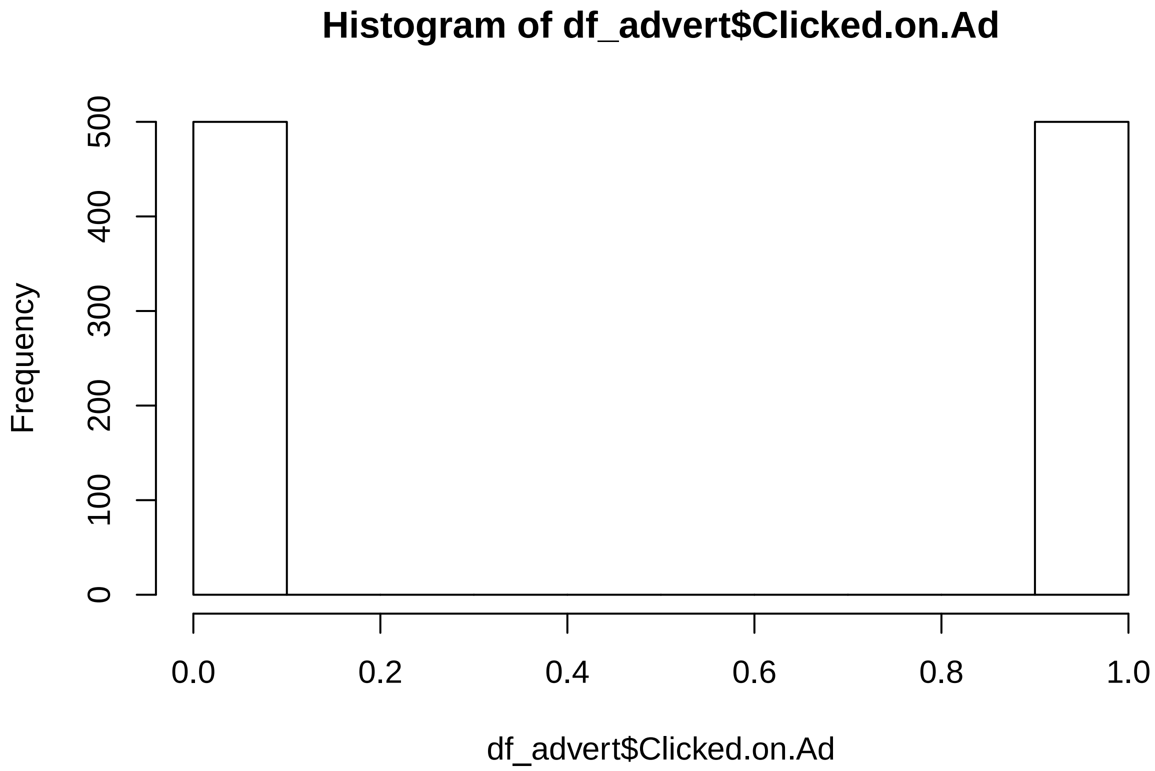
e) Male

```r
hist(df_advert$Male)
```



f) Clicked on Ad

```
hist(df_advert$Clicked.on.Ad)
```



Bivariate Analysis

Covariance

```
num_cols <- Filter(is.numeric, df_advert)
cat('COVARIANCE')
```

```
## COVARIANCE
```

```
cov(num_cols)
```

```
##                          Daily.Time.Spent.on.Site           Age   Area.Income
## Daily.Time.Spent.on.Site              251.3370949 -4.617415e+01  6.613081e+04
## Age                                   -46.1741459  7.718611e+01 -2.152093e+04
## Area.Income                         66130.8109082 -2.152093e+04  1.799524e+08
## Daily.Internet.Usage                  360.9918827 -1.416348e+02  1.987625e+05
## Male                                   -0.1501864 -9.242142e-02  8.867509e+00
## Clicked.on.Ad                          -5.9331431  2.164665e+00 -3.195989e+03
##                          Daily.Internet.Usage        Male Clicked.on.Ad
## Daily.Time.Spent.on.Site         3.609919e+02 -0.15018639 -5.933143e+00
## Age                             -1.416348e+02 -0.09242142  2.164665e+00
## Area.Income                      1.987625e+05  8.86750903 -3.195989e+03
## Daily.Internet.Usage             1.927415e+03  0.61476667 -1.727409e+01
## Male                                  6.147667e-01  0.24988889 -9.509510e-03
## Clicked.on.Ad                        -1.727409e+01 -0.00950951  2.502503e-01
```

```r
cat('\nCORRELATION')
```

```
## 
## CORRELATION
```

```r
cor(num_cols)
```

```
##                          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
```

Age and clicked on Ad have a medium positive correlation as it has a value greater than one.

Daily time spent on site and Daily internet usage have strong negative correlation against clicked on Ad.

a) Age Versus Time Spent On Site The covariance of age versus time spent on site is -46.174. It indicates a medium negative linear relationship between the two variables. The younger the person is, the more the time spent on site

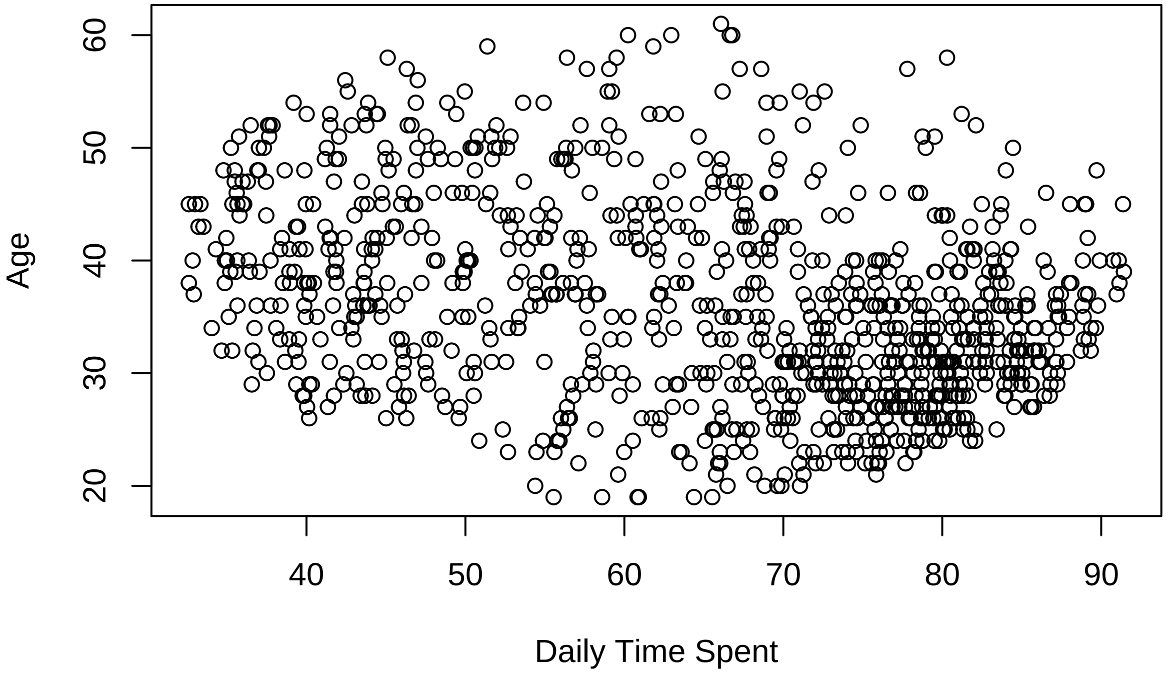
b) Male versus Clicked on A

The covariance of clicked on Ad versus gender is -0.00950951. It indicates no significant difference between Male versus other genders as far as clicking on Ad.

Scatter Plot

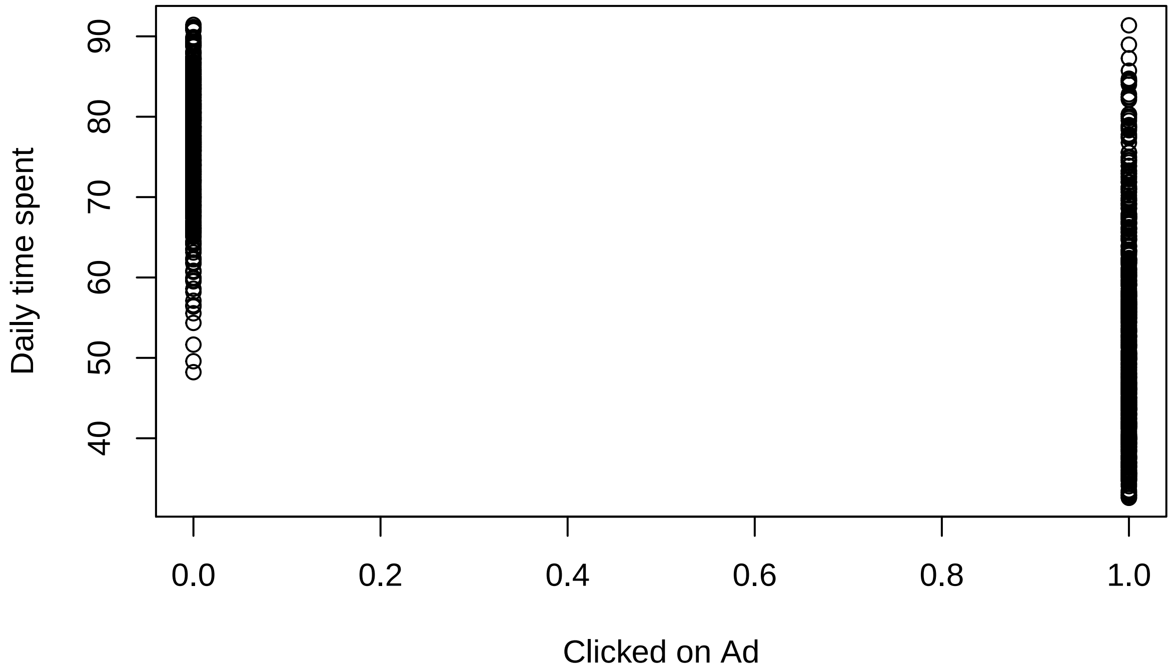
a) Daily time spent versus Age

```r
plot(df_advert$Daily.Time.Spent.on.Site, df_advert$Age, xlab="Daily Time Spent", ylab="Age")
```



b) Clicked on Ad versus Daily time spent

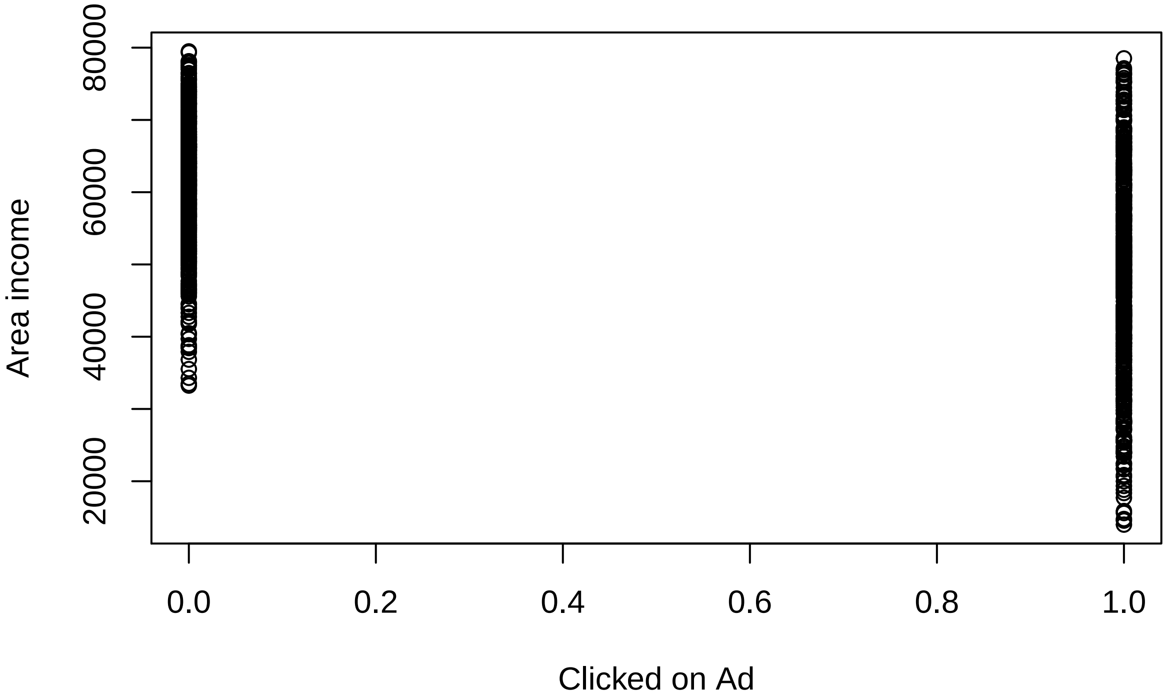
```
plot(df_advert$Clicked.on.Ad, df_advert$Daily.Time.Spent.on.Site , xlab="Clicked on Ad", ylab="Daily ti
```



The users click on add irrespective of time spent. Hence an advert can be placed at the top of the blog and users will still click it.

c) Clicked on Ad versus Daily Time Spent

```
plot(df_advert$Clicked.on.Ad, df_advert$Area.Income , xlab="Clicked on Ad", ylab="Area income")
```



People living in lower income areas clicked on the add.

Summary

- Lower time spent on internet does not limit the clicking of adverts. Hence adverts can be placed on the top of the blog.
- Target audience can include low income areas.
- The higher the age, the more the probability of clicking on the ad. Hence target audience can be users in high age groups who have more disposable income.