# IPweek12

Alex

02/07/2021

**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.**

IMPORTING OUR DATASET

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

Preview of the data set

```
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 data types

```
typeof(data)
```

```
## [1] "list"
```

Checking the data shape of our data set

```
nrow(data)
```

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

```
ncol(data)
```

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

```
#From our data set, we can see that we have a total of 1000 rows and a total of 10 columns.
```

Checking for missing values

```
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
```

```
#When we check for null values in each column, we can see that there are no missing values in our data
```

Checking for duplicates

```
duplicated_rows <- data[duplicated(data),]

duplicated_rows
```

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

```
#From the above code, we can clearly see the output which shows that there are no duplicates in our dat
```

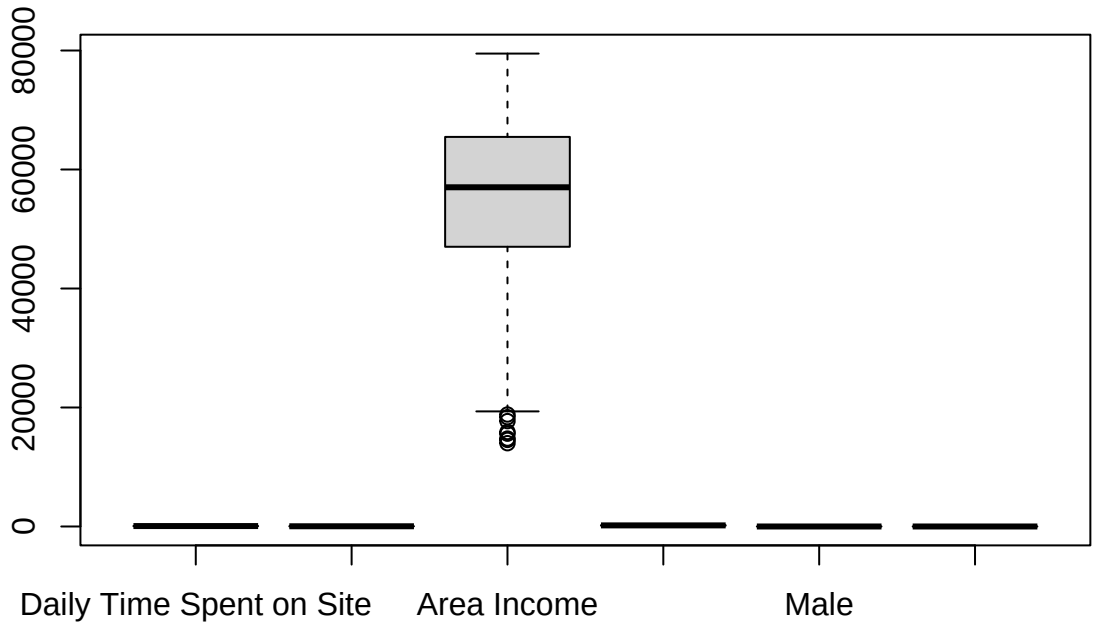
Checking for Outliers

```
#creating a variable with only numeric columns
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```
data2 <-data %>% select(1,2,3,4,7,10)

#Previewing outliers for numeric columns using box plots
boxplot(data2)
```




From the box plot above, we can conclude that there aren't any outliers in our given data set

# Exploratory Data Analysis

## Univariate Analysis

Finding the mean

```
lapply(data2,FUN=mean)
```

```
## $`Daily Time Spent on Site`
## [1] 65.0002
##
## $Age
## [1] 36.009
##
## $`Area Income`
## [1] 55000
##
## $`Daily Internet Usage`
## [1] 180.0001
##
## $Male
## [1] 0.481
##
## $`Clicked on Ad`
## [1] 0.5
```

Finding the median

```
lapply(data2,FUN=median)
```

```
## $`Daily Time Spent on Site`
## [1] 68.215
##
## $Age
## [1] 35
##
## $`Area Income`
## [1] 57012.3
##
## $`Daily Internet Usage`
## [1] 183.13
##
## $Male
## [1] 0
##
## $`Clicked on Ad`
## [1] 0.5
```

Finding the mode

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

lapply(data,FUN=getmode)
```

```
## $`Daily Time Spent on Site`
## [1] 62.26
##
## $Age
## [1] 31
##
## $`Area Income`
## [1] 61833.9
##
## $`Daily Internet Usage`
## [1] 167.22
##
## $`Ad Topic Line`
## [1] "Cloned 5thgeneration orchestration"
##
## $City
## [1] "Lisamouth"
##
## $Male
## [1] 0
##
## $Country
## [1] "Czech Republic"
##
## $Timestamp
## [1] "2016-03-27 00:53:11 UTC"
##
## $`Clicked on Ad`
## [1] 0
```

Measures of Dispersion

Minumum

```
lapply(data2,FUN=min)
```

```
## $`Daily Time Spent on Site`
## [1] 32.6
##
## $Age
## [1] 19
##
## $`Area Income`
## [1] 13996.5
##
## $`Daily Internet Usage`
## [1] 104.78
##
## $Male
## [1] 0
##
## $`Clicked on Ad`
## [1] 0
```

Maximum

```
lapply(data2,FUN=max)
```

```
## $`Daily Time Spent on Site`
## [1] 91.43
##
## $Age
## [1] 61
##
## $`Area Income`
## [1] 79484.8
##
## $`Daily Internet Usage`
## [1] 269.96
##
## $Male
## [1] 1
##
## $`Clicked on Ad`
## [1] 1
```

Range

```
lapply(data2,FUN=range)
```

```
## $`Daily Time Spent on Site`
## [1] 32.60 91.43
##
## $Age
## [1] 19 61
##
## $`Area Income`
## [1] 13996.5 79484.8
##
## $`Daily Internet Usage`
## [1] 104.78 269.96
##
## $Male
## [1] 0 1
##
## $`Clicked on Ad`
## [1] 0 1
```

Quantile

```
lapply(data2,FUN=quantile)
```

```
## $`Daily Time Spent on Site`
##      0%     25%     50%     75%    100%
## 32.6000 51.3600 68.2150 78.5475 91.4300
##
```

```
## $Age
##   0%  25%  50%  75% 100%
##   19   29   35   42   61
##
## $`Area Income`
##       0%      25%      50%      75%     100%
## 13996.50 47031.80 57012.30 65470.64 79484.80
##
## $`Daily Internet Usage`
##       0%      25%      50%      75%     100%
## 104.7800 138.8300 183.1300 218.7925 269.9600
##
## $Male
##   0%  25%  50%  75% 100%
##    0    0    0    1    1
##
## $`Clicked on Ad`
##   0%  25%  50%  75% 100%
##  0.0  0.0  0.5  1.0  1.0
```

Variance

```
lapply(data2,FUN=var)
```

```
## $`Daily Time Spent on Site`
## [1] 251.3371
##
## $Age
## [1] 77.18611
##
## $`Area Income`
## [1] 179952406
##
## $`Daily Internet Usage`
## [1] 1927.415
##
## $Male
## [1] 0.2498889
##
## $`Clicked on Ad`
## [1] 0.2502503
```

Standard Deviation

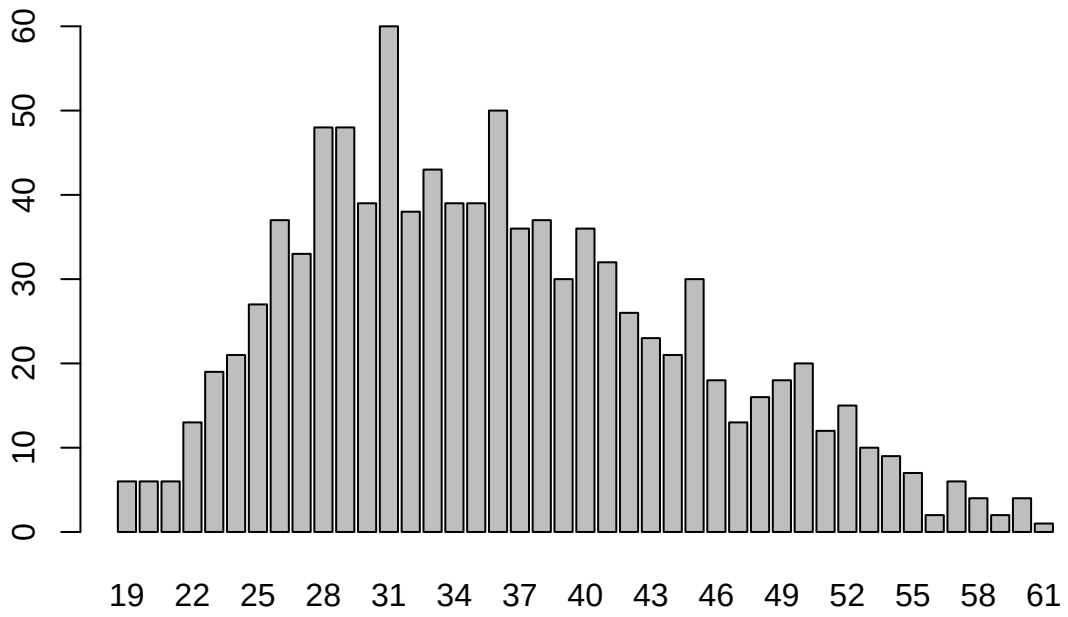
```
lapply(data2,FUN=sd)
```

```
## $`Daily Time Spent on Site`
## [1] 15.85361
##
## $Age
## [1] 8.785562
##
```

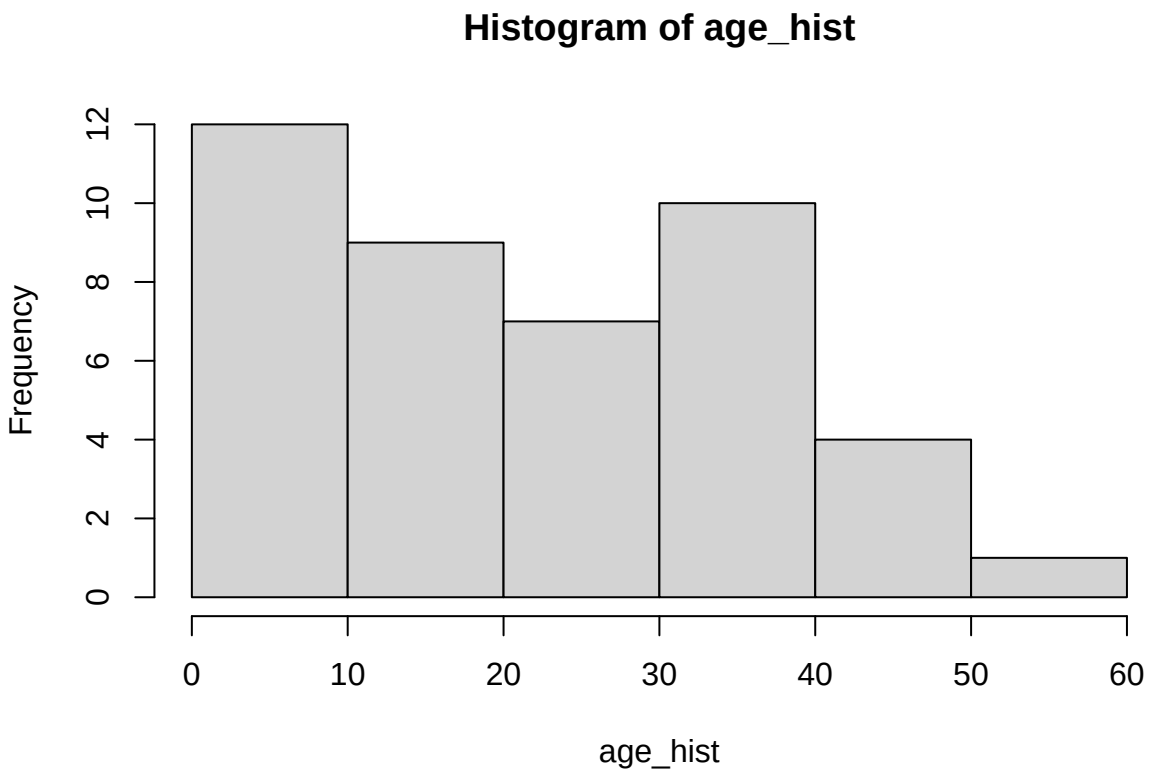
```
## $`Area Income`
## [1] 13414.63
##
## $`Daily Internet Usage`
## [1] 43.90234
##
## $Male
## [1] 0.4998889
##
## $`Clicked on Ad`
## [1] 0.5002502
```

### Univariate Graphical

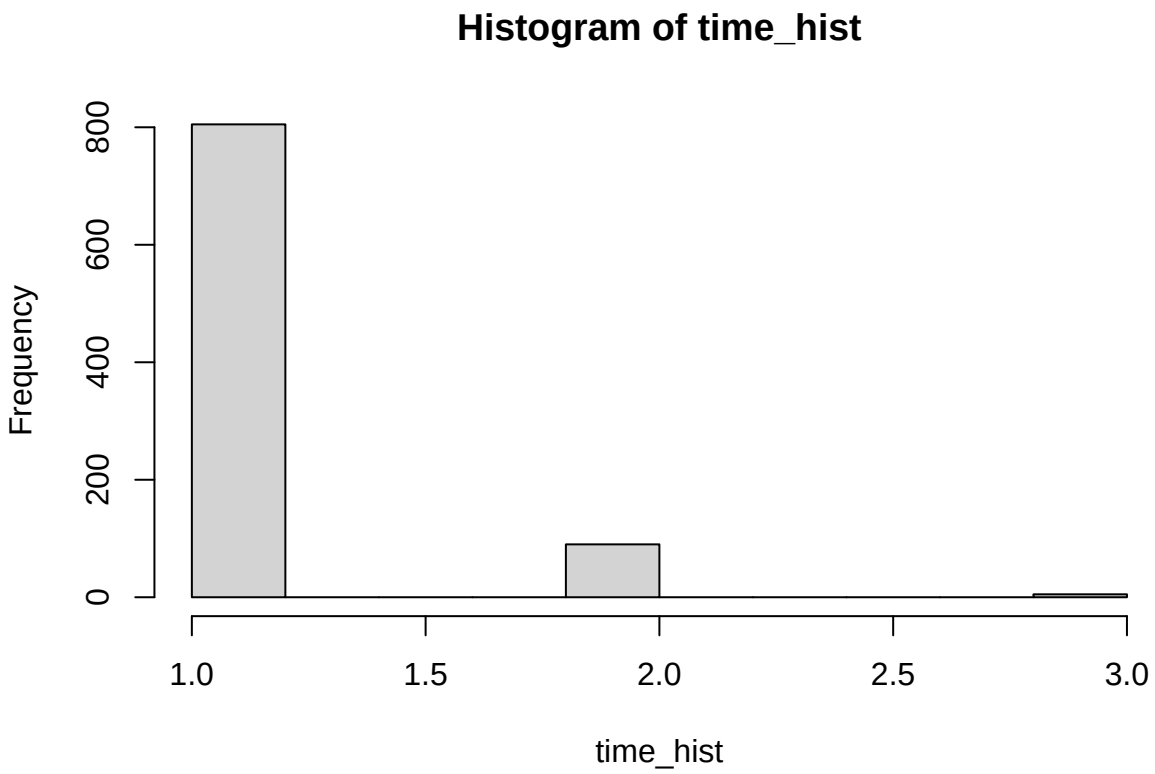
```
age_frequency <- table(data$`Age`)

barplot(age_frequency)
```




```
age_hist <- table(data$`Age`)

hist(age_hist)
```




```
time_hist <- table(data$`Daily Time Spent on Site`)

hist(time_hist)
```




## Bivariate Analysis

Finding the covariance between internet Usage and internet time

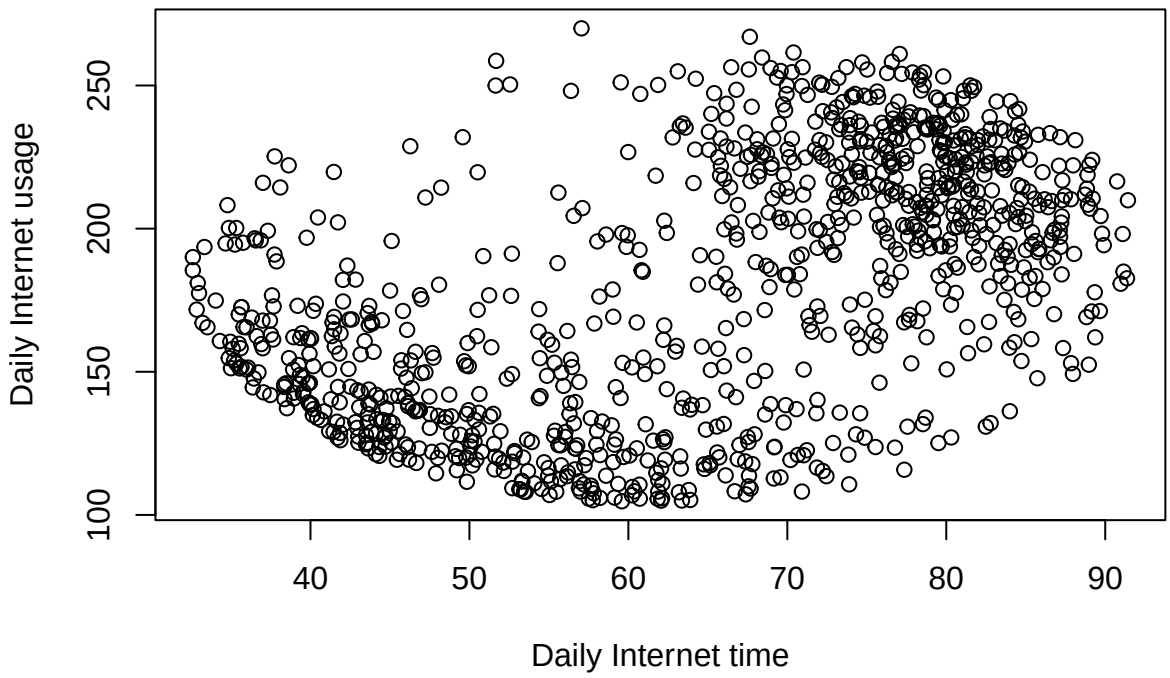
```
time <- (data$`Daily Time Spent on Site`)


usage  <- (data$`Daily Internet Usage`)

cov(time, usage)
```

```
## [1] 360.9919
```

### Graphical Techniques

Scatter plot of Daily Internet usage and Daily internet time spent

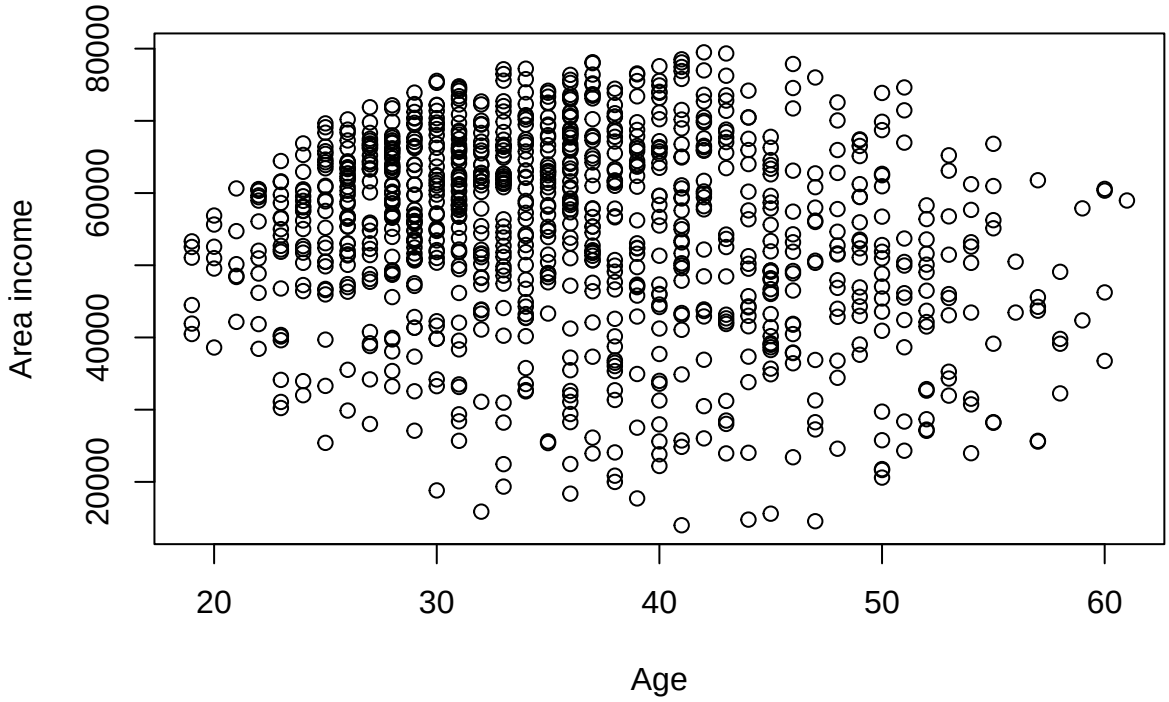
```
plot(time, usage, xlab="Daily Internet time", ylab="Daily Internet usage")
```




From the above scatter plot, we can see that there is a positive correlation between daily time spent and daily internet usage

Scatter plot of age and Area income

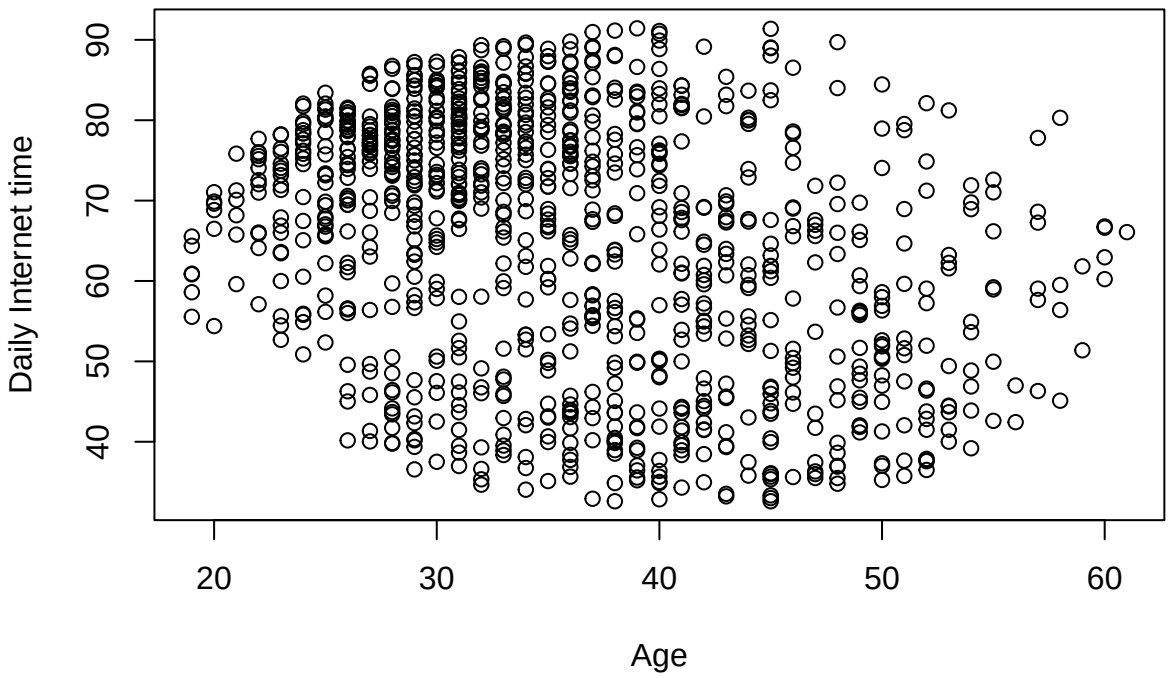
```
age <- (data$Age)


income  <- (data$`Area Income`)

plot(age, income, xlab="Age", ylab="Area income")
```




From the graph above, we can see that there is no correlation between age and area income

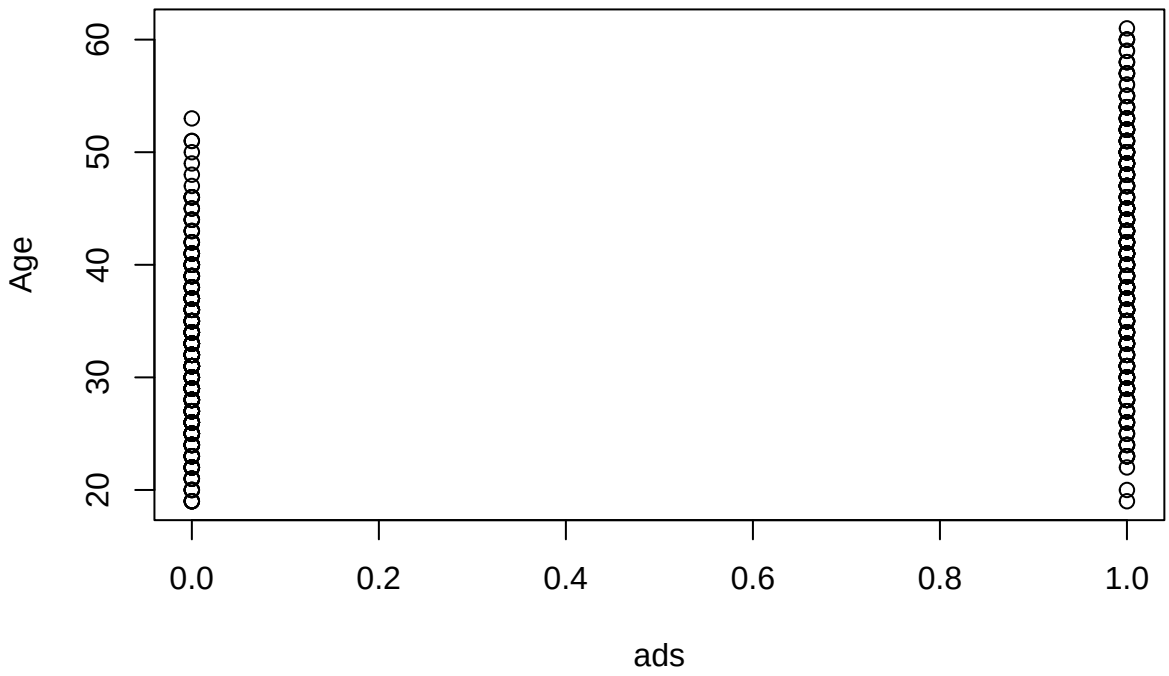
Scatter plot of age and Daily internet time spent

```
plot(age, time, xlab="Age", ylab="Daily Internet time")
```



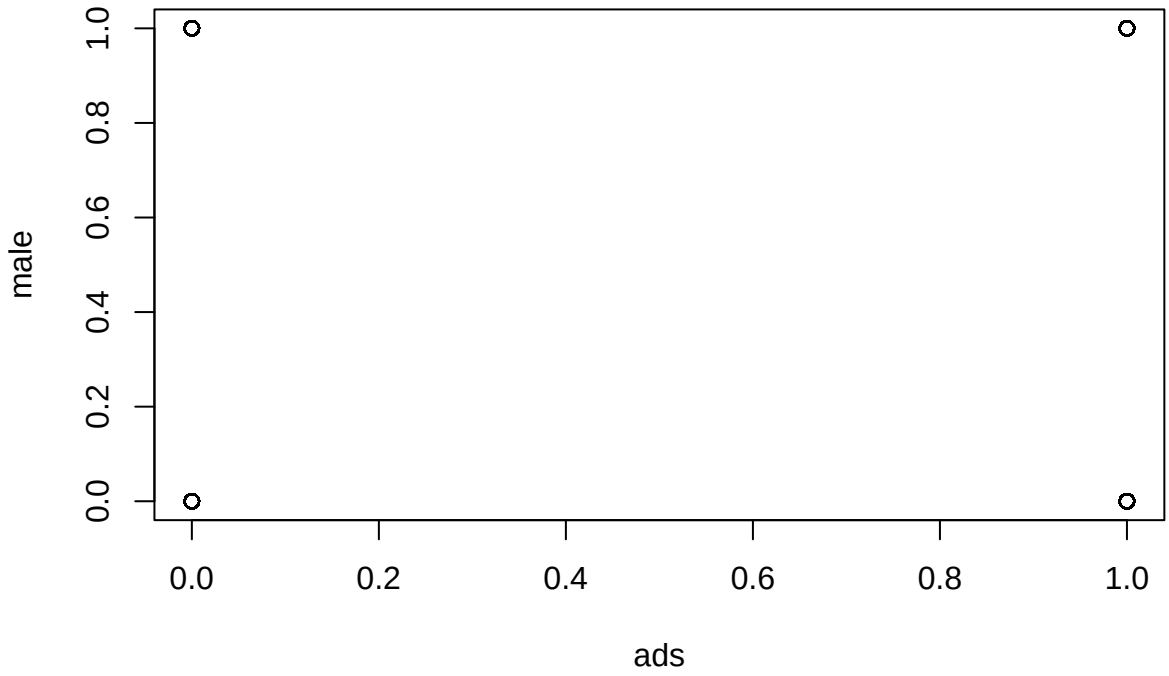

From the above plot, we can see no correlation between Age and duration of time spent daily on the internet.

```
ad <- (data$`Clicked on Ad`)


country  <- (data$Country)

plot(ad, age, xlab="ads", ylab="Age")
```



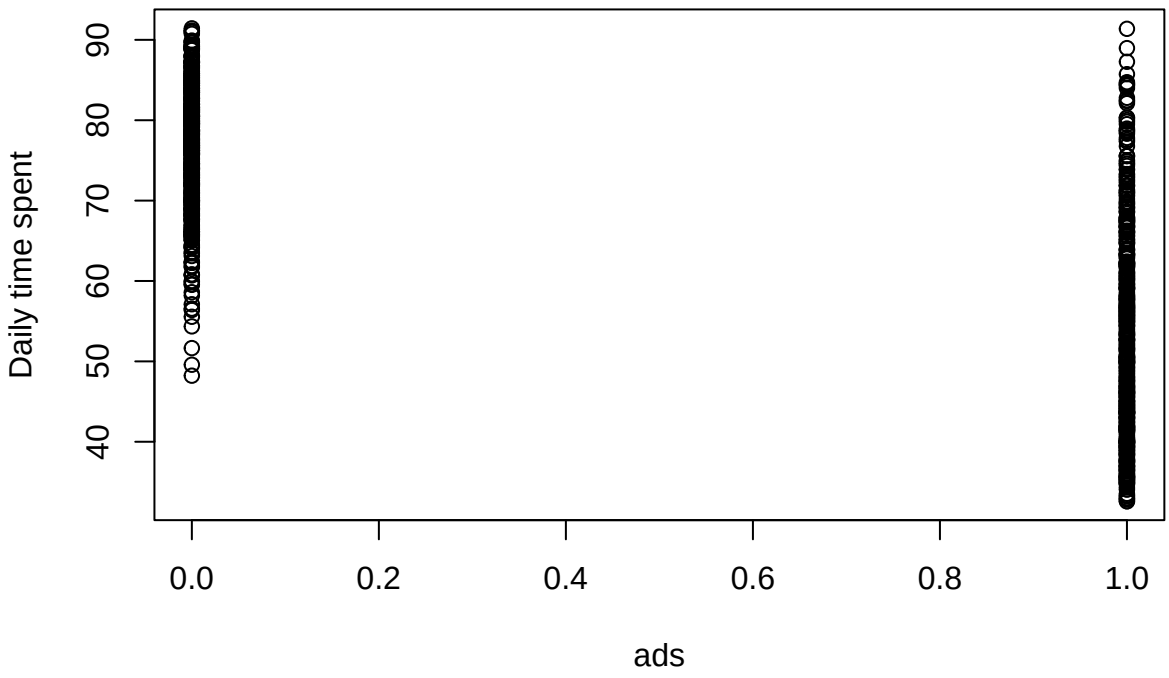

From the plot above, we can see that people from all age end up clicking on the ads. We can also see that people who click on ads are also from the age of 50-60.

```
ad <- (data$`Clicked on Ad`)


male  <- (data$Male)

plot(ad, male, xlab="ads", ylab="male")
```




From the plot, we can see that gender isn't a factor that affects whether people click on ads or not.

```
plot(ad, time, xlab="ads", ylab="Daily time spent")
```




From the plot we can see that majority of the people who click on ads are the ones who spend most time daily on the internet.

## Decision trees

```
library(rpart)
library(rpart.plot)
library(mlbench)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Defining features and target variables

library(tidyverse)
ad <-data %>% select(1,2,3,4,7,10)
```

```r
add <-data %>% select(10)

head(ad)
```

```
##    Daily Time Spent on Site Age Area Income Daily Internet Usage Male
## 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
## 1:             0
## 2:             0
## 3:             0
## 4:             0
## 5:             0
## 6:             0
```

```r
ad$Country <- as.integer(as.factor(ad$Country))
```

```r
data_train <- ad[1:800, ]
data_test <- ad[801:1000,]
```

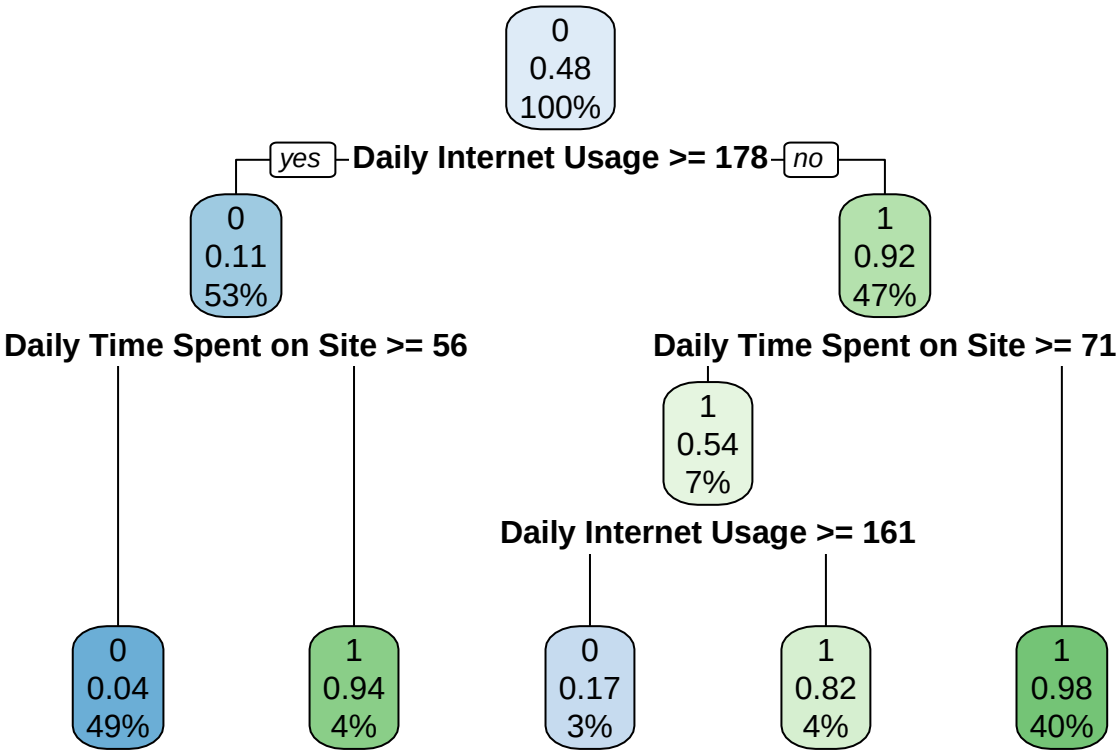
```r
dim(ad)
```

```
## [1] 1000    7
```

```r
head(ad)
```

```
##    Daily Time Spent on Site Age Area Income Daily Internet Usage Male
## 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 Country
## 1:             0      NA
## 2:             0      NA
## 3:             0      NA
## 4:             0      NA
## 5:             0      NA
## 6:             0      NA
```

```r
model <- rpart(`Clicked on Ad`~., data = data_train, method = 'class')
rpart.plot(model)

# Visualizing the model

rpart.plot(model)
```




## Conclusion

From the graphs above we can conclude that majority of the people who will click on the ads are the ones who spend more time on the internet, are from all ages including the age from 50-60 and use more internet usage since it has a positive correlation with daily internet time.

## Recommendations

The company should target people from all age mostly from the age of 50-60 maybe by adverting something close to what that age group may like. The company should focus less on gender and more on people who spend more daily internet time maybe by giving free internet usage on clicking the ads.