1. Problem Definition

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.

Research Question

A model that assist the enterpreneur best decide which factors to use in rolling out their advertisements to get highest engagement on them(calculated by number of users that click on the ads)

2. Data Sourcing

The data used for the analysis was sourced from here.

The data contains infromation about interactions of customers with advertisements placed in a crypography course website.

3.Loading and Checking our Dataset

```
ad = read.csv("http://bit.ly/IPAdvertisingData")
```

Previewing the Dataset

```
head(ad)
```

```
##   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
```

```
tail(ad)
```

```
##      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
```

##4. Data Cleaning

```
#Checking for null values in our dataset

colSums(is.na(ad))
```

```
## 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
```

our dataset does not contain any missing values.

```
# Checking for duplicates

anyDuplicated(ad)
```

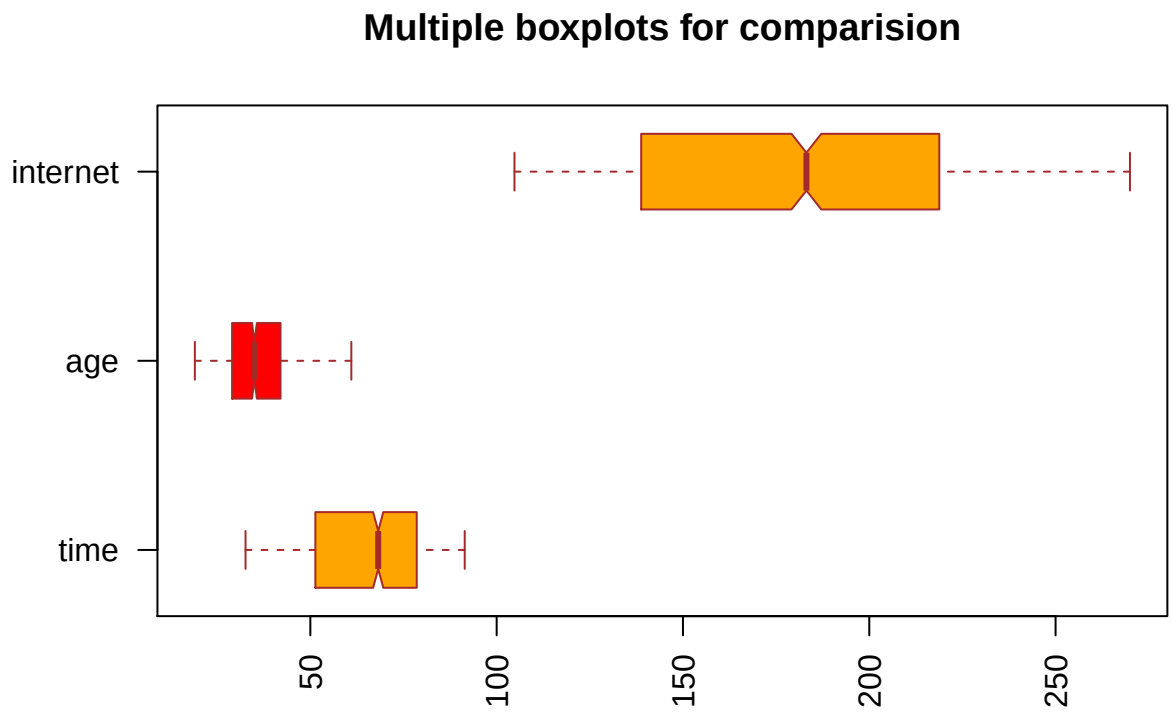
```
## [1] 0
```

Our dataset does not contain any duplicates.

```r
str(ad)
```

```
## '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 ...
```

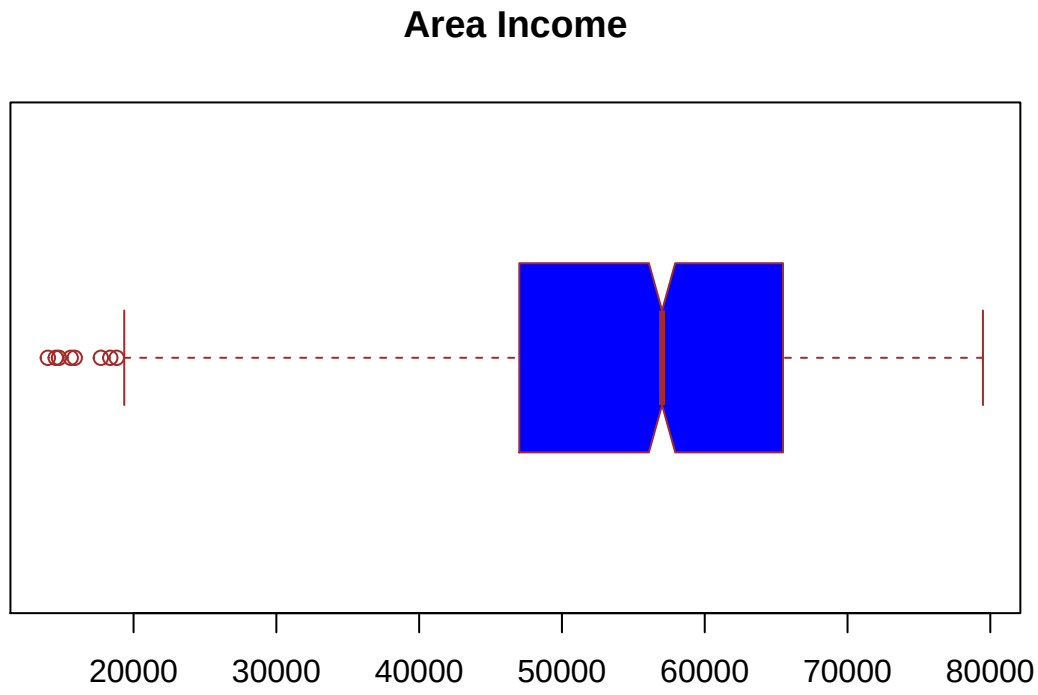
```r
# Checking for outliers

# prepare the data
time <- ad$Daily.Time.Spent.on.Site
age <- ad$Age
internet <- ad$Daily.Internet.Usage

boxplot(time, age, internet,
main = "Multiple boxplots for comparision",
at = c(1,3,5),
names = c("time", "age", "internet"),
las = 2,
col = c("orange","red"),
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```



The “Daily time spent on Site”, “Age” and “Daily Internet Usage” columns do not contain any outliers.

The values on the income column were far too different to be plotted with the other columns while still making the box plots make sense. A box plot for income values is done below.

```
boxplot(ad$Area.Income,
main = "Area Income",
col = "blue",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```



The "Area Income" has several outliers. However, since these represent income earned by different people in different geographical areas, the outliers are assumed to be factual and will not be removed or replaced.

5.Exploratory Data Analysis

Univariate Exploratory Data Analysis

Calculating/identifying several measures of central tendency.

To start of we will create a function for calculating mode.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
```

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

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

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

Finding the mode

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

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

```
getmode(ad$Age)
```

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

```
getmode(ad$Area.Income)
```

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

```
getmode(ad$Daily.Internet.Usage)
```

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

The mode for daily time spent on the site is 62.26 The mode for age is 31 The mode for area income is 6.18339×10^4 The mode for daily internet usage is 167.22

Finding the mean

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

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

```
mean(ad$Age)
```

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

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

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

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

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

The mean daily time spent on the site is 65.0002 The mean for the age is 36.009 The mean area income is 5.5×10^4 The mean daily internet usage is 180.0001

Finding the Median

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

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

```r
median(ad$Age)
```

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

```r
median(ad$Area.Income)
```

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

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

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

The median daily time spent on the site is 68.215 The median for the age is 35 The median area income is 5.70123×10^4 The median daily internet usage is 183.13

Calculating Measures of Dispersion

Finding the maximum values

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

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

```r
max(ad$Age)
```

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

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

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

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

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

The maximum daily time spent on the site is 91.43 The maximum for the age is 61 The maximum area income is 7.94848×10^4 The maximum daily internet usage is 269.96

Finding the minimum values

```r
min(ad$Daily.Time.Spent.on.Site)
```

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

```r
min(ad$Age)
```

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

```r
min(ad$Area.Income)
```

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

```r
min(ad$Daily.Internet.Usage)
```

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

The minmum daily time spent on the site is 32.6 The minmum for the age is 19 The minmum area income is 1.39965×10^4 The minmum daily internet usage is 104.78

Finding the Range

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

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

```r
range(ad$Age)
```

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

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

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

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

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

The range daily time spent on the site is 32.6, 91.43 The range for the age is 19, 61 The range area income is 1.39965×10^4, 7.94848×10^4 The range daily internet usage is 104.78, 269.96

Finding the Variance

```r
var(ad$Daily.Time.Spent.on.Site)
```

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

```r
var(ad$Age)
```

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

```r
var(ad$Area.Income)
```

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

```r
var(ad$Daily.Internet.Usage)
```

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

The variance for daily time spent on the site is 251.3370949 The variance for the age is 77.1861051 The variance for area income is 1.7995241×10^8 The variance for daily internet usage is 1927.4153962

Finding the Standard Deviation

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

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

```r
sd(ad$Age)
```

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

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

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

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

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

The Standard deviation for daily time spent on the site is 15.8536146 The Standard deviation for the age is 8.7855623 The Standard deviation for area income is 1.3414634×10^4 The Standard deviation for daily internet usage is 43.9023393

```r
quantile(ad$Daily.Time.Spent.on.Site)
```

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

```r
quantile(ad$Age)
```

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

```r
quantile(ad$Area.Income)
```

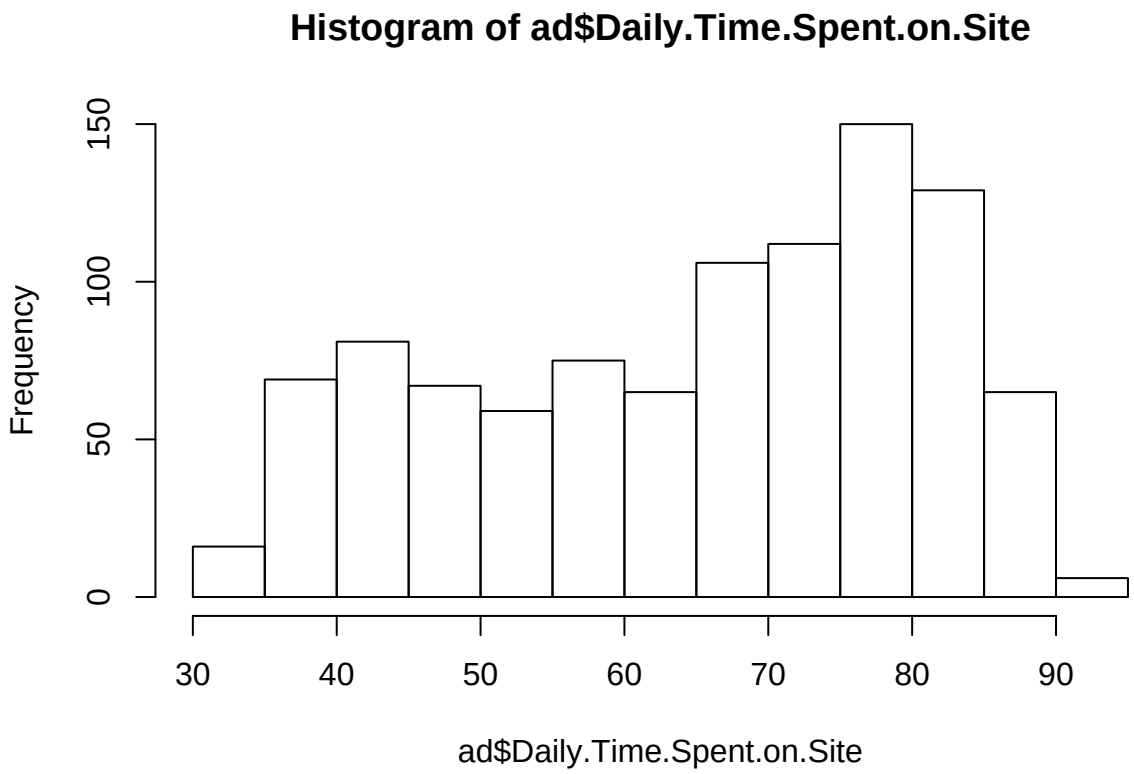
```
##       0%      25%      50%      75%     100%
## 13996.50 47031.80 57012.30 65470.64 79484.80
```

```r
quantile(ad$Daily.Internet.Usage)
```

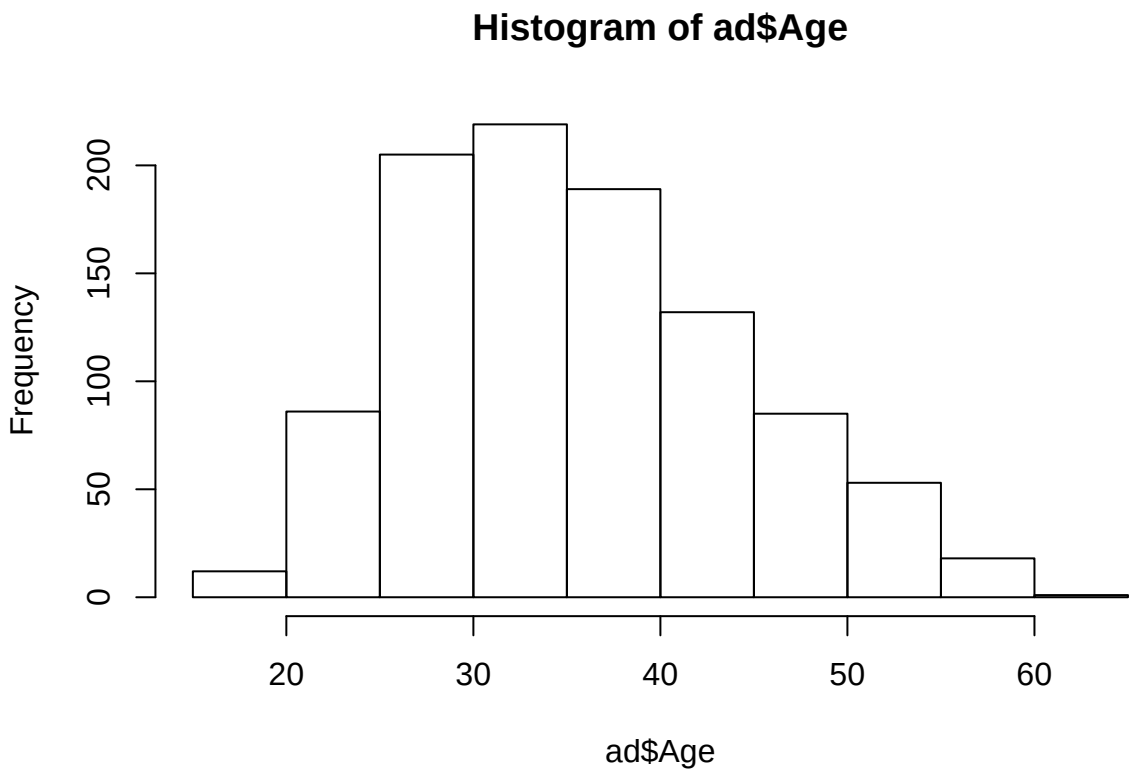
```
##       0%      25%      50%      75%     100%
## 104.7800 138.8300 183.1300 218.7925 269.9600
```

Visualisations

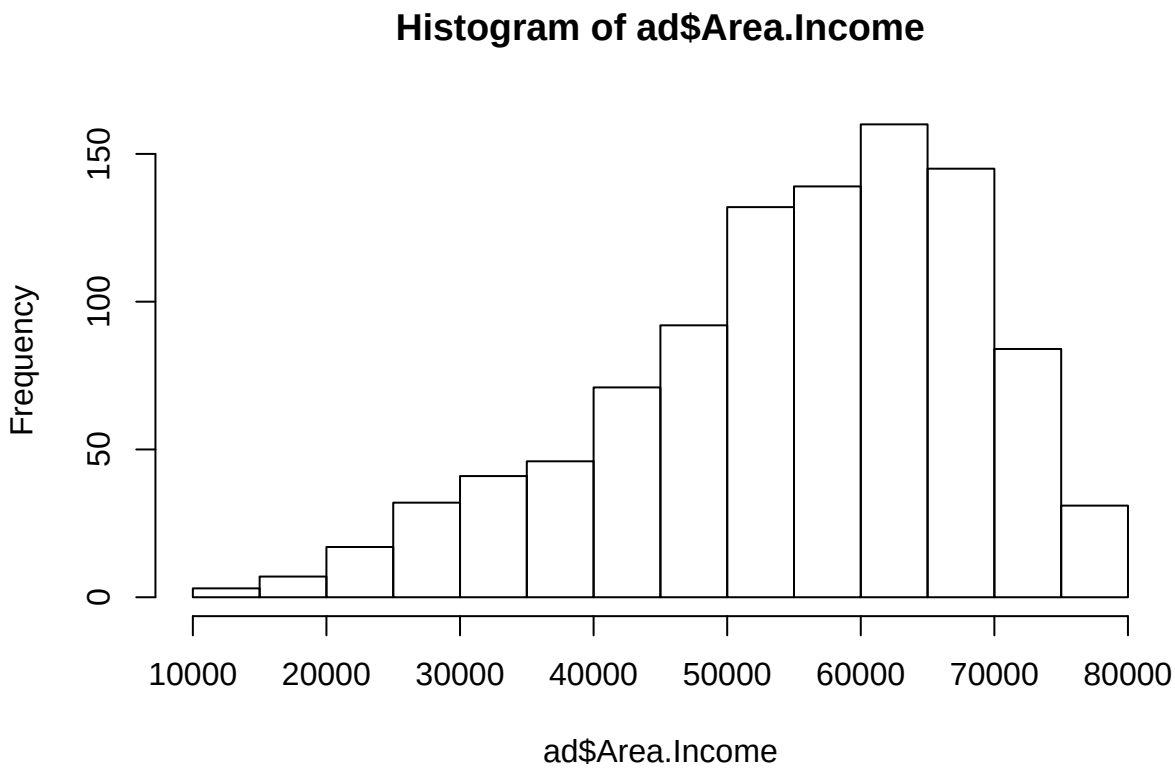
```
hist(ad$Daily.Time.Spent.on.Site)
```



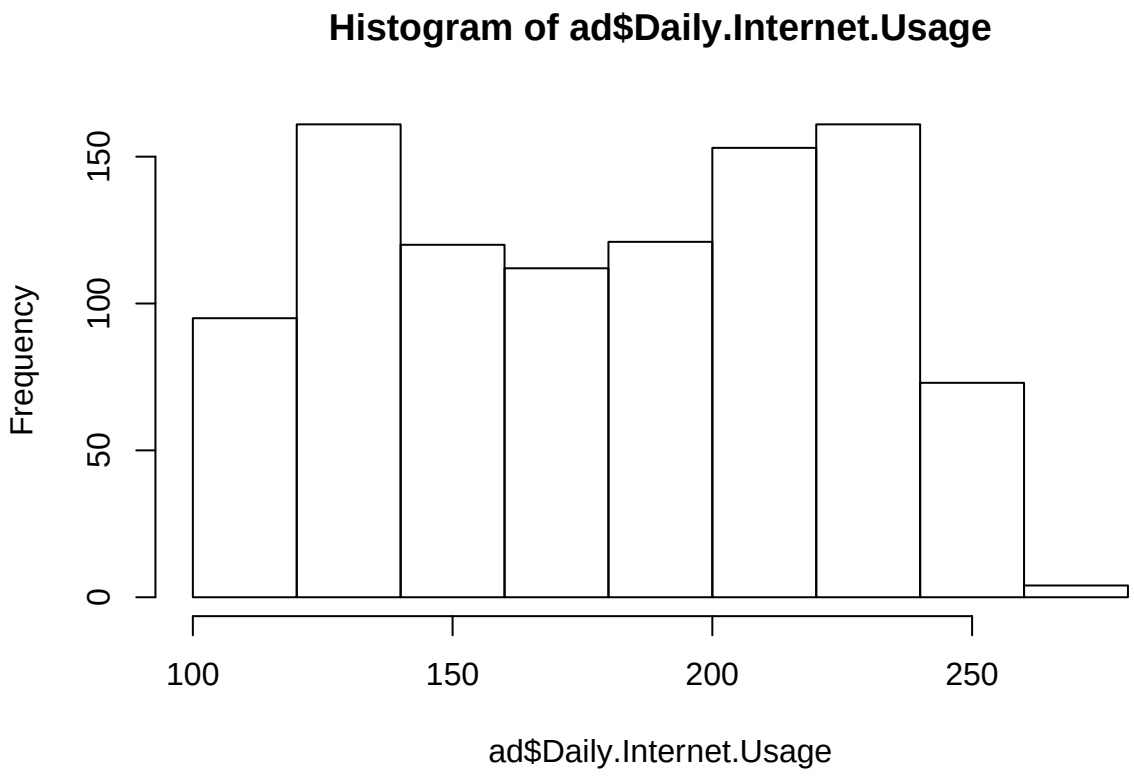

```
hist(ad$Age)
```




```r
hist(ad$Area.Income)
```



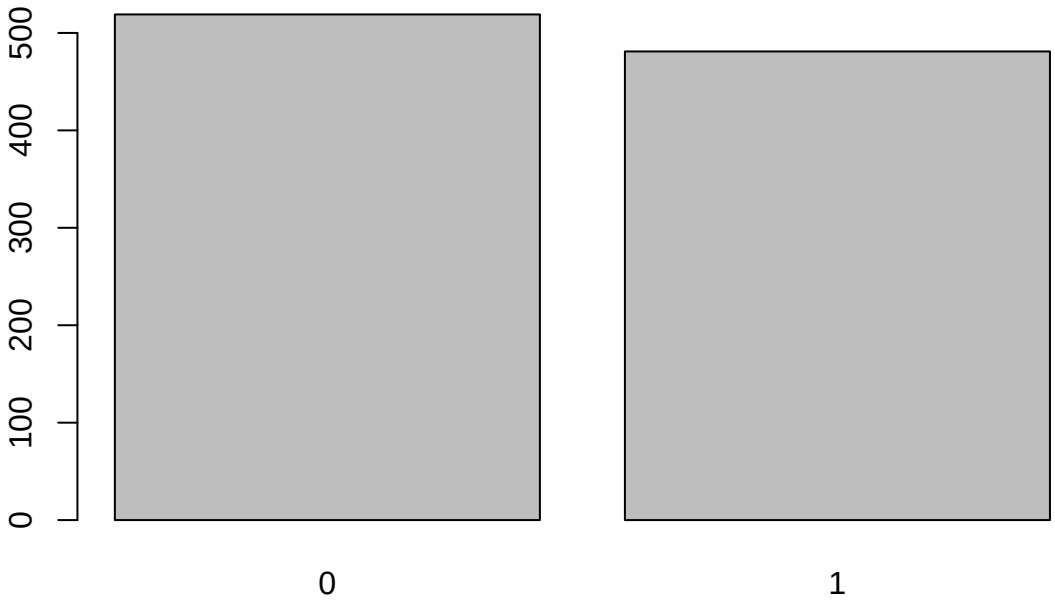

```
hist(ad$Daily.Internet.Usage)
```




```
table(ad$Male)
```

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

```
barplot(table(ad$Male))
```

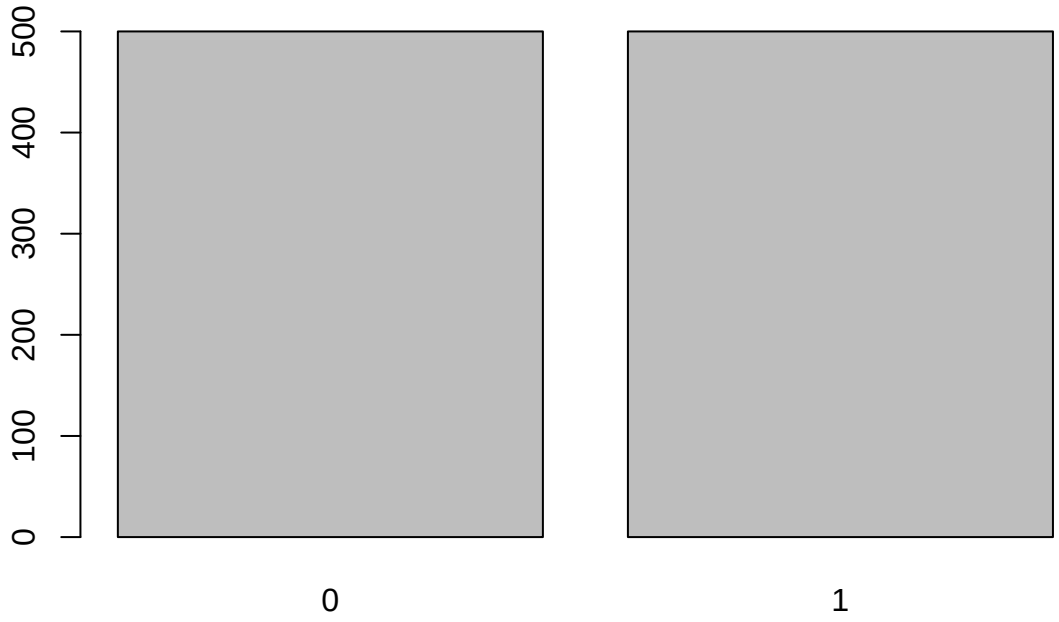


The dataset has 519 females and 481 males.

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

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

```
barplot(table(ad$Clicked.on.Ad))
```



500 people clicked on the ad, while 500 others did not click on the ad.

Bivariate Analysis

```
# Creating variables to used in checking for covariance
time <- ad$Daily.Time.Spent.on.Site
age <- ad$Age
income <- ad$Area.Income
usage <- ad$Daily.Internet.Usage
gender <- ad$Male
click <- ad$Clicked.on.Ad
```

Covariance between likelihood of clicking the ad and other variables

```
cov(click,time)
```

```
## [1] -5.933143
```

```
cov(click,age)
```

```
## [1] 2.164665
```

```r
cov(click,income)
```

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

```r
cov(click,usage)
```

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

```r
cov(click,gender)
```

```
## [1] -0.00950951
```

The chance of clicking an ad has a positive linear relationship with the age of the site user. **Correlation between likelihood of clicking the ad and other variables**

```r
cor(click,time)
```

```
## [1] -0.7481166
```

```r
cor(click,age)
```

```
## [1] 0.4925313
```

```r
cor(click,income)
```

```
## [1] -0.4762546
```

```r
cor(click,usage)
```

```
## [1] -0.7865392
```

```r
cor(click,gender)
```

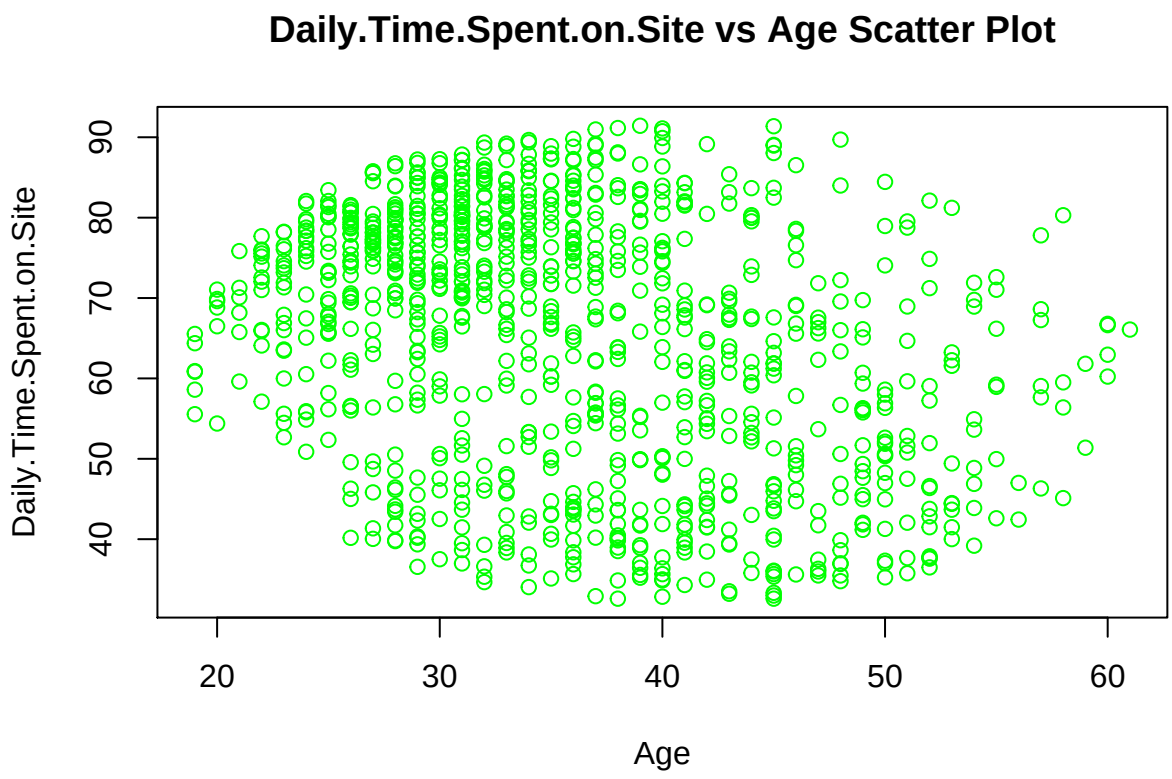
```
## [1] -0.03802747
```

There is a high negative correlation between chance of clicking the ad with time spent on the site and daily internet usage There is a week correlation between chance of clicking on an ad and the age of the user 0.4925313

```r
my_data <- ad[, c(1,2,3,4,7,10)]

head(my_data)
```

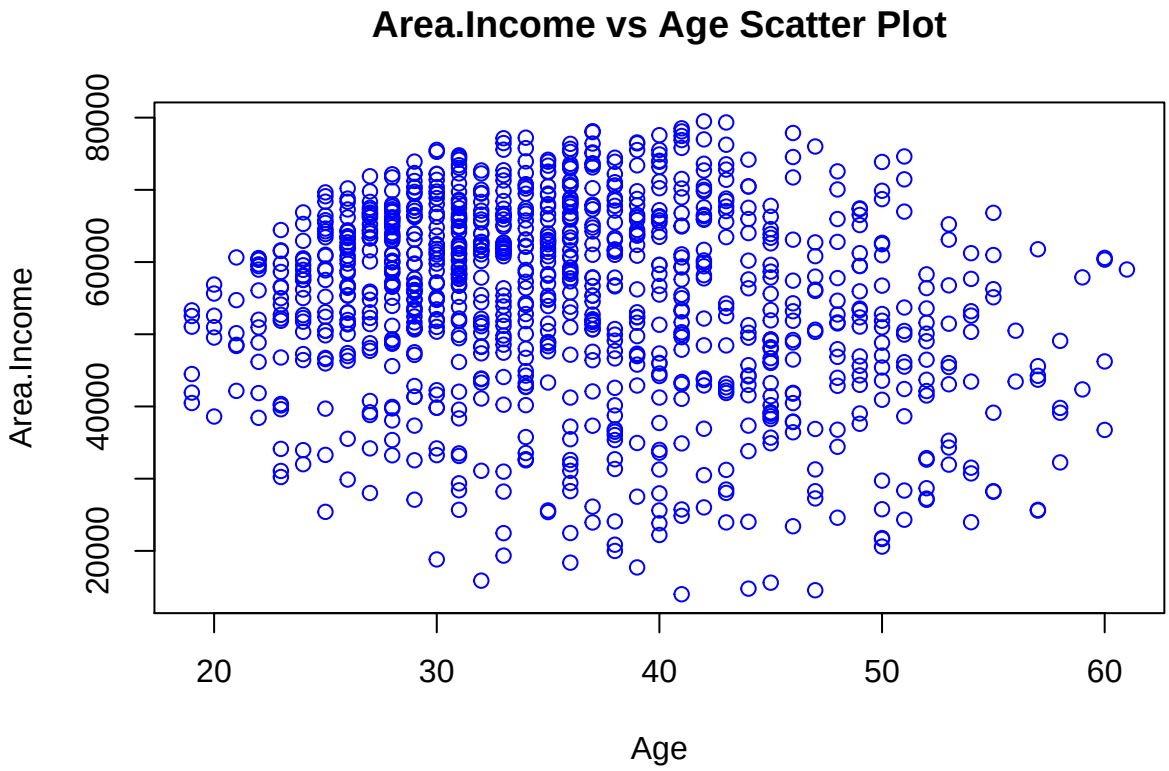
```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage Male
## 1                    68.95  35    61833.90                256.09    0
## 2                    80.23  31    68441.85                193.77    1
## 3                    69.47  26    59785.94                236.50    0
## 4                    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
```

```
plot( Daily.Time.Spent.on.Site~Age , dat = my_data,
      col = "green",
      main = "Daily.Time.Spent.on.Site vs Age Scatter Plot")
```



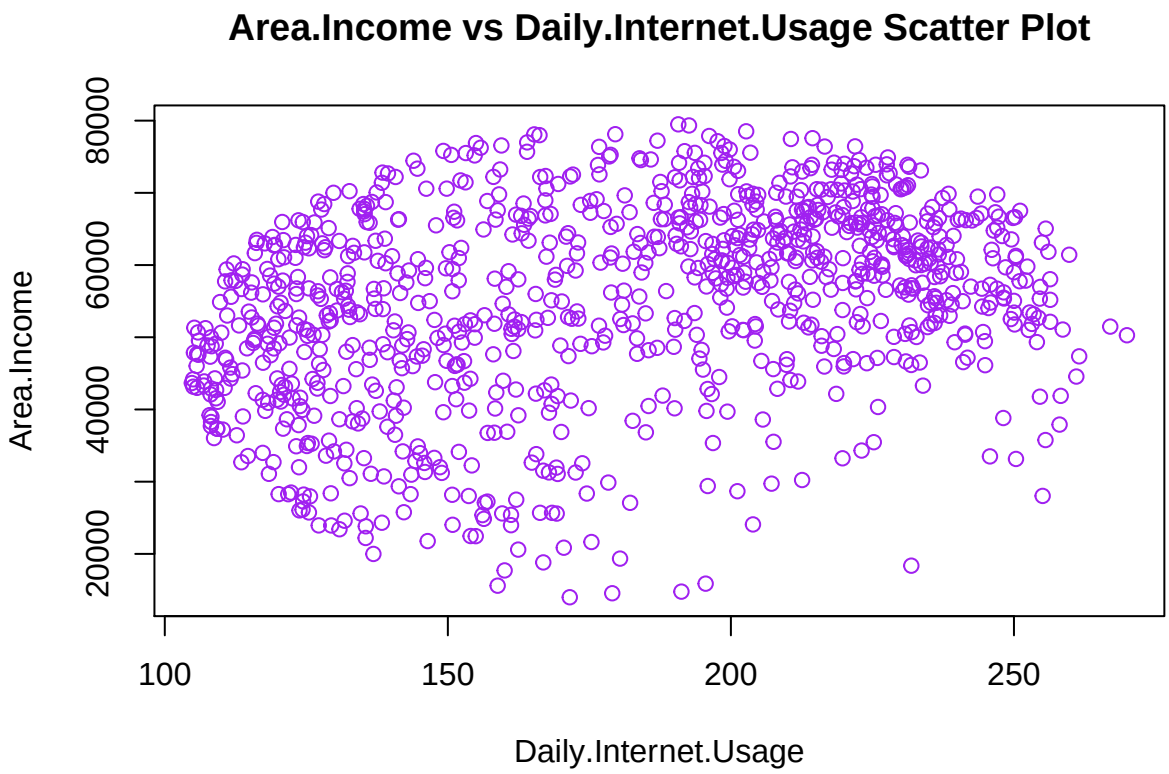
The age of the user and the time spent on the site show no corelation.

```
plot( Area.Income~Age , dat = my_data,
      col = "blue",
      main = "Area.Income vs Age Scatter Plot")
```



There is also no linear corelation between the area income and the age.

```
plot( Area.Income~Daily.Internet.Usage , dat = my_data,
      col = "purple",
      main = "Area.Income vs Daily.Internet.Usage Scatter Plot")
```



There is also no linear corelation between the area income and the daily internet usage

```
res <- cor(my_data)
round(res, 2)
```

```
##                           Daily.Time.Spent.on.Site   Age Area.Income
## Daily.Time.Spent.on.Site                      1.00 -0.33        0.31
## Age                                          -0.33  1.00       -0.18
## Area.Income                                   0.31 -0.18        1.00
## Daily.Internet.Usage                          0.52 -0.37        0.34
## Male                                         -0.02 -0.02        0.00
## Clicked.on.Ad                                -0.75  0.49       -0.48
##                           Daily.Internet.Usage  Male Clicked.on.Ad
## Daily.Time.Spent.on.Site                 0.52 -0.02         -0.75
## Age                                     -0.37 -0.02          0.49
## Area.Income                              0.34  0.00         -0.48
## Daily.Internet.Usage                     1.00  0.03         -0.79
## Male                                     0.03  1.00         -0.04
## Clicked.on.Ad                           -0.79 -0.04          1.00
```

```
cor(my_data, use = "complete.obs")
```

```
##                           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
```

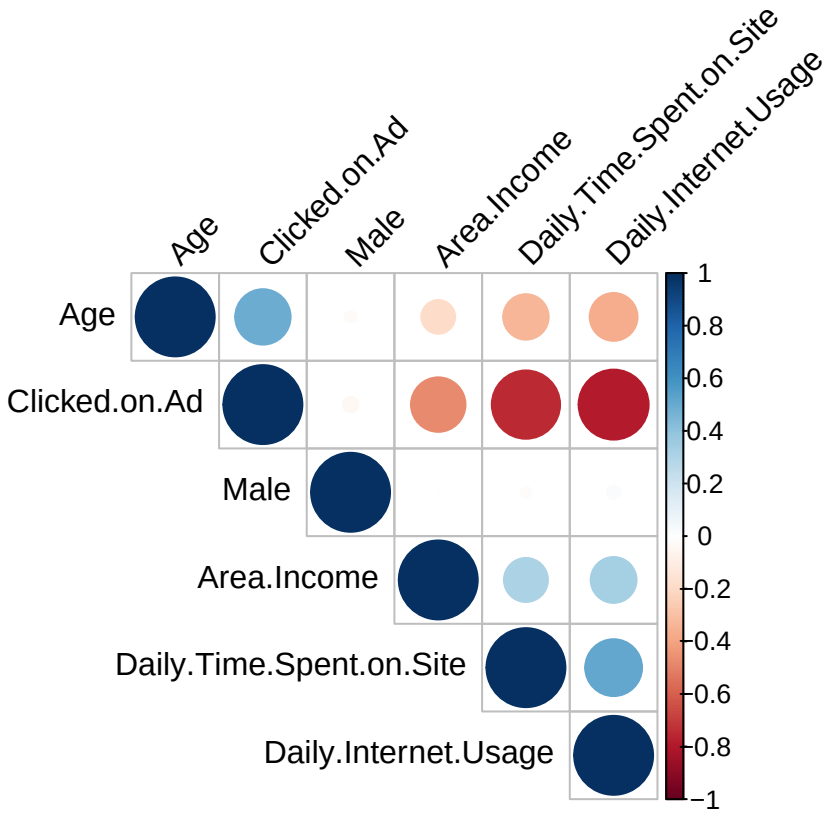
```
## Daily.Internet.Usage                  0.51865848 -0.36720856  0.337495533
## Male                                 -0.01895085 -0.02104406  0.001322359
## Clicked.on.Ad                        -0.74811656  0.49253127 -0.476254628
##                           Daily.Internet.Usage         Male Clicked.on.Ad
## 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
```

```
library(corrplot)
```

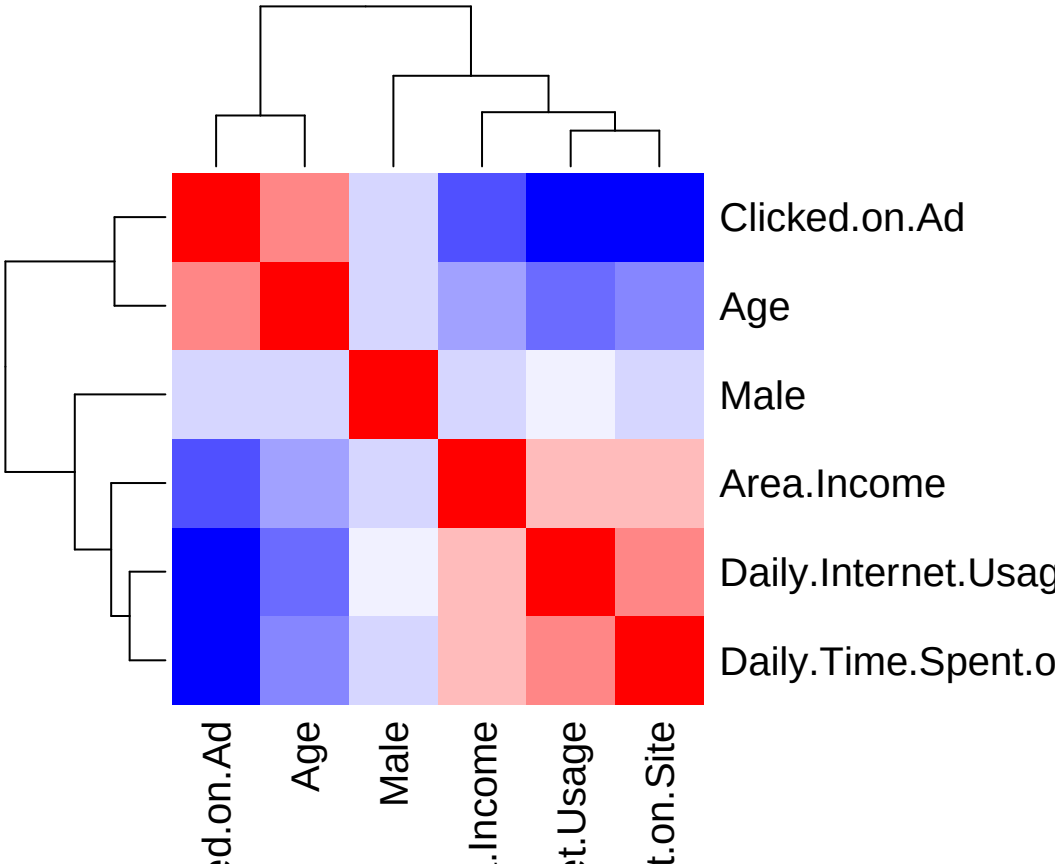
```
## Warning: package 'corrplot' was built under R version 3.6.3
```

```
## corrplot 0.84 loaded
```

```
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```




```
# Get some colors
col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(x = res, col = col, symm = TRUE)
```



6. Implementing the Solution

Baseline Model For a our baseline model we will create a linear regression model, we will use this to compare with our more advanced models to see whether we are improving the accuracy of the new models.

```
# Applying the lm() function.
multiple_lm <- lm(Clicked.on.Ad~ ., my_data)
```

```
# Generating the anova table
anova(multiple_lm)
```

```
## Analysis of Variance Table
##
## Response: Clicked.on.Ad
##                            Df  Sum Sq Mean Sq   F value  Pr(>F)
## Daily.Time.Spent.on.Site    1 139.920 139.920 3162.2238 < 2e-16 ***
## Age                         1  16.793  16.793  379.5306 < 2e-16 ***
## Area.Income                 1  13.721  13.721  310.0920 < 2e-16 ***
## Daily.Internet.Usage        1  35.372  35.372  799.4083 < 2e-16 ***
## Male                        1   0.213   0.213    4.8183 0.02839 *
## Residuals                 994  43.982   0.044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Then performing our prediction
pred <- predict(multiple_lm, my_data)
```

```r
# Then performing our prediction
summary(multiple_lm)
```

```
##
## Call:
## lm(formula = Clicked.on.Ad ~ ., data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65251 -0.11577 -0.03069  0.05081  1.03147
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.309e+00  5.755e-02  40.113   <2e-16 ***
## Daily.Time.Spent.on.Site -1.279e-02  5.058e-04 -25.294   <2e-16 ***
## Age                          8.983e-03  8.283e-04  10.845   <2e-16 ***
## Area.Income                -6.173e-06  5.351e-07 -11.536   <2e-16 ***
## Daily.Internet.Usage      -5.260e-03  1.867e-04 -28.169   <2e-16 ***
## Male                        -2.926e-02  1.333e-02  -2.195   0.0284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2104 on 994 degrees of freedom
## Multiple R-squared:  0.8241, Adjusted R-squared:  0.8232
## F-statistic: 931.2 on 5 and 994 DF,  p-value: < 2.2e-16
```

SVM

We shall use SVM to create our first model

```r
intrain <- createDataPartition(y = my_data$Clicked.on.Ad, p= 0.7, list = FALSE)
training <- my_data[intrain,]
testing <- my_data[-intrain,]
```

```r
# We check the dimensions of out training dataframe and testing dataframe
# ---
#
dim(training);
```

```
## [1] 700   6
```

```r
dim(testing);
```

```
## [1] 300   6
```

```r
#
training[["Clicked.on.Ad"]] = factor(training[["Clicked.on.Ad"]])
```

```r
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

svm_Linear <- train(Clicked.on.Ad ~., data = training, method = "svmLinear",
trControl=trctrl,
preProcess = c("center", "scale"),
tuneLength = 10)
```

```r
# We can then check the reult of our train() model as shown below
# ---
#
svm_Linear
```

```
## Support Vector Machines with Linear Kernel
##
## 700 samples
##   5 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (5), scaled (5)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 630, 630, 630, 630, 630, 630, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9719048  0.9438095
##
## Tuning parameter 'C' was held constant at a value of 1
```

```r
# We can use the predict() method for predicting results as shown below.
# We pass 2 arguements, our trained model and our testing data frame.
# ---
#
test_pred <- predict(svm_Linear, newdata = testing)
```

```r
# Now checking for our accuracy of our model by using a confusion matrix
# ---
#
confusionMatrix(table(test_pred, testing$Clicked.on.Ad))
```

```
## Confusion Matrix and Statistics
##
##
## test_pred   0   1
##         0 146   7
##         1   4 143
##
##                Accuracy : 0.9633
##                  95% CI : (0.9353, 0.9816)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9267
```

```
##
##  Mcnemar's Test P-Value : 0.5465
##
##             Sensitivity : 0.9733
##             Specificity : 0.9533
##          Pos Pred Value : 0.9542
##          Neg Pred Value : 0.9728
##              Prevalence : 0.5000
##          Detection Rate : 0.4867
##    Detection Prevalence : 0.5100
##       Balanced Accuracy : 0.9633
##
##        'Positive' Class : 0
##
```

using SVM we get a model that is 97 accurate. The model is very accurate.

The model was able to correctly predict 146 and 145 clicks or no clicks, while wrongly predicting 5 and 4 clicks or no clicks.

7. Challenge the Solution.

The SVM Model produced an accuracy of 97%, for such a dataset, this level of accuracy is good enough. However, more models could be created to try and come up with models that are more accurate.

8. Follow up Question.

To better understand our data we could create a clustering algorithm to test how well it perfoms in predicting in which cluster the users are placed.