Advertisement Analysis — Week 12 IP

1. Define the question

1.1 Main question —which individuals are more likely to click on cryptpgraphy course advertisement?

1.2 Metric for Success

1.3 Understanding the context

A Kenyan entrepreneur, who has previously ran a course advertisement on her blog would like to know: which kinds of individuals are more likely to click on her advertisements posted on the blog. From the initial advertisement, the entrepreneur was able to collect some data and would like a Data Science Consultant to analyze the given data and report recommendations. The findings from the analysis will be crucial in determining which kinds of individuals to target before running another online course ad on the blog. ### 1.4 Experimental Design The approach for this project will include:

1. Importing the necessary libraries
2. Reading the data
3. Checking the data
4. Tidying up the data
5. Implementing the solution using Univariate and Bivariate
6. Conclusion and Recommendations

1.5 Data Relevance

I think that the data was relevant, however, I also think that we could have used more columns to add onto our analysis.

1. Importing libraries

```
library(tidyverse)
```

```
## -- Attaching packages ----------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(ggcorrplot)
```

2. Reading the data

```
# Imported data from the directory where it was saved
advertising <- read.csv("~/Moringa School/R Programming/R datasets/advertising.csv")
view(advertising)
```

3. Checking the data

```
# let's get glimpse of how the data looks like
glimpse(advertising)
```

```
## Rows: 1,000
## Columns: 10
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, ...
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49...
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73...
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 22...
## $ Ad.Topic.Line            <chr> "Cloned 5thgeneration orchestration", "Mon...
## $ City                     <chr> "Wrightburgh", "West Jodi", "Davidton", "W...
## $ Male                     <int> 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, ...
## $ Country                  <chr> "Tunisia", "Nauru", "San Marino", "Italy",...
## $ Timestamp                <chr> "2016-03-27 00:53:11", "2016-04-04 01:39:0...
## $ Clicked.on.Ad            <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, ...
```

Our data has 1,000 records and 10 columns. Some columns have integer, float and character data types. clicked.on.ad is our target variable. The target variable and male columns are listed as integers, but they are both categorical. We will need to convert these columns data types to the appropriate data types during cleaning.

```
#checking the top of the data
head(advertising)
```

```
##   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 the bottom of the data
tail(advertising)
```

```
##      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
```

After looking at the head and tail of the data, there are a few assumptions on the data we need to make before we proceed. These assumptions include:

1. Daily time spend on site column is in minutes.
2. Income is in United States Dollars($).
3. Daily internet usage units is in megabytes. Also assuming that the megabytes refers to data used on the blog site.
4. For male column: if 0 that means not male, but if 1 then yes it's a male.
5. Clicked on ad column: if 1 that means yes somebody clicked on the ad, but if 0 then the add was not clicked.

```
# checking for unique elements in columns if interest
# starting with our target variable
unique(advertising$Clicked.on.Ad)
```

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

There are two elements 0 for not clicked and 1 for clicked(based on the assumption we made above)

```
unique(advertising$Male)
```

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

There is two elements in the male column. 0 for not a male and 1 for male

```
#unique(advertising$Country)
```

In total there is a total of 237 countries. From the countries list, I see funny country names e.g Nauru, Wallis & Futuna, Holy see. Where did these countries come from? Note: country output not added here because the list is too long

4. Tidying the data

4.1 Fixing data types

```
# converting male column to character
advertising$Male <- as.character(advertising$Male)
```

```
# converting clicked on ad column to character
advertising$Clicked.on.Ad <- as.character(advertising$Clicked.on.Ad)
```

```
# confirming that the data types have been converted
glimpse(advertising)
```

```
## Rows: 1,000
## Columns: 10
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, ...
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49...
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73...
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 22...
## $ Ad.Topic.Line            <chr> "Cloned 5thgeneration orchestration", "Mon...
## $ City                     <chr> "Wrightburgh", "West Jodi", "Davidton", "W...
## $ Male                     <chr> "0", "1", "0", "1", "0", "1", "0", "1", "1...
## $ Country                  <chr> "Tunisia", "Nauru", "San Marino", "Italy",...
## $ Timestamp                <chr> "2016-03-27 00:53:11", "2016-04-04 01:39:0...
## $ Clicked.on.Ad            <chr> "0", "0", "0", "0", "0", "0", "0", "1", "0...
```

Male and Clicked on ad have been converted to their appropriate data types

4.2 Missing values

```
# checking for missing values in each column
colSums(is.na(advertising))
```

```
## 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 in our dataset

4.3 Duplicates

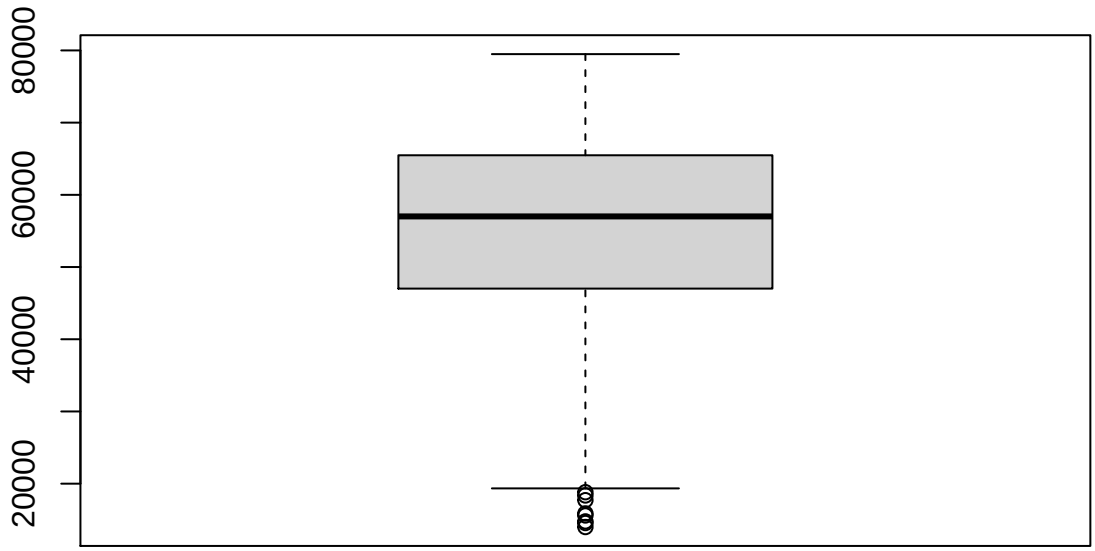
```
# checking for duplicates
advert_duplicates <- advertising[duplicated(advertising),]
advert_duplicates
```

```
##  [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 duplicates in our dataset

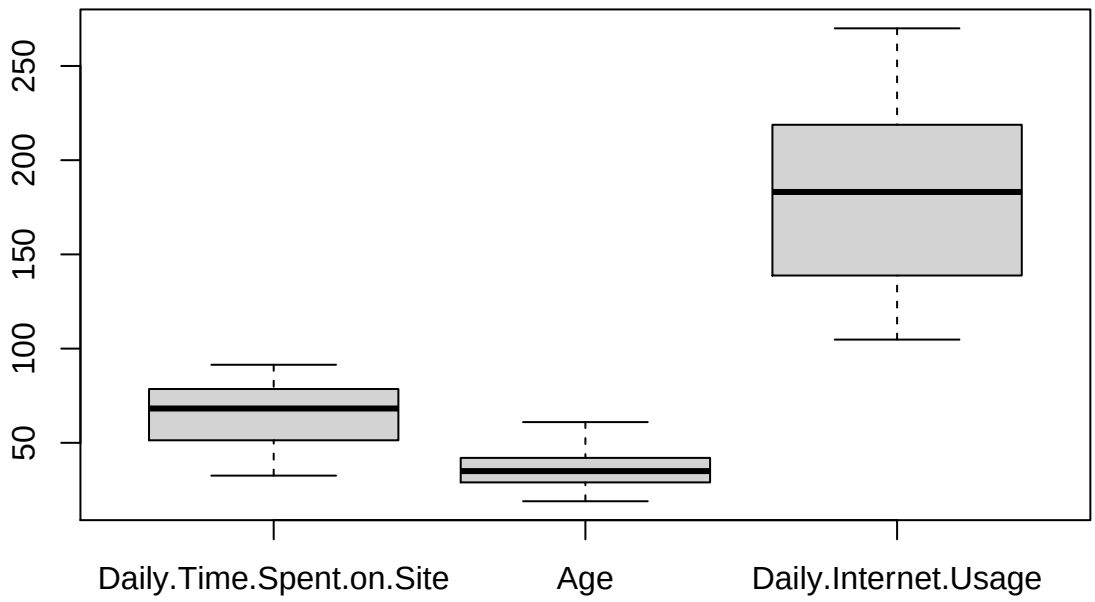
4.4 Outliers

```
# checking for outliers in numerical variables using boxplots
# we will separate area income from other numerical variables because of the difference in scale
boxplot(advertising$Area.Income)
```



We see some outliers from Income column on the lower side of the whisker. We won't be deleting these outliers because we establish that these are true observations. The low income could be as a result of many reasons. An explanation for these observations is that — in every country there are certain areas that have a lot of poor people, which leads to low area income.

```
# checking for outliers in the other numerical columns
num_cols <- select(advertising, Daily.Time.Spent.on.Site, Age, Daily.Internet.Usage)
boxplot(num_cols)
```



There are no observed outliers in these columns. Based on the analysis, so far, our data is pretty clean and we are ready to begin solution implementation using Univariate and Bivariate analysis.

5. Implementing the solution using Univariate and Bivariate

5.1 Univariate analysis

```
#checking for statistical summaries of numerical data. This will allow us to get a summary of: mean
#median, min, max, range, and quantiles
#first select numerical variables
num <- select(advertising, Daily.Time.Spent.on.Site, Age, Area.Income,Daily.Internet.Usage)
glimpse(num)
```

5.1.1 Measures of central tendency and measure of dispersion for numerical variables

```
## Rows: 1,000
## Columns: 4
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, ...
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49...
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73...
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 22...
```

```
summary(num)
```

```
##  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
```

On average daily time spent on the blog is 65 minutes.

Average age that visit the site is about 36 years old, min age = 19 years and max 61 years Average income area is $55,000.

The average daily internet usage on the site is 180 megabytes.

From the statistical summary, we don't see anything strange with the numerical columns.

```
# Variance
# Daily.Time.Spent.on.Site column
sapply(num, var)
```

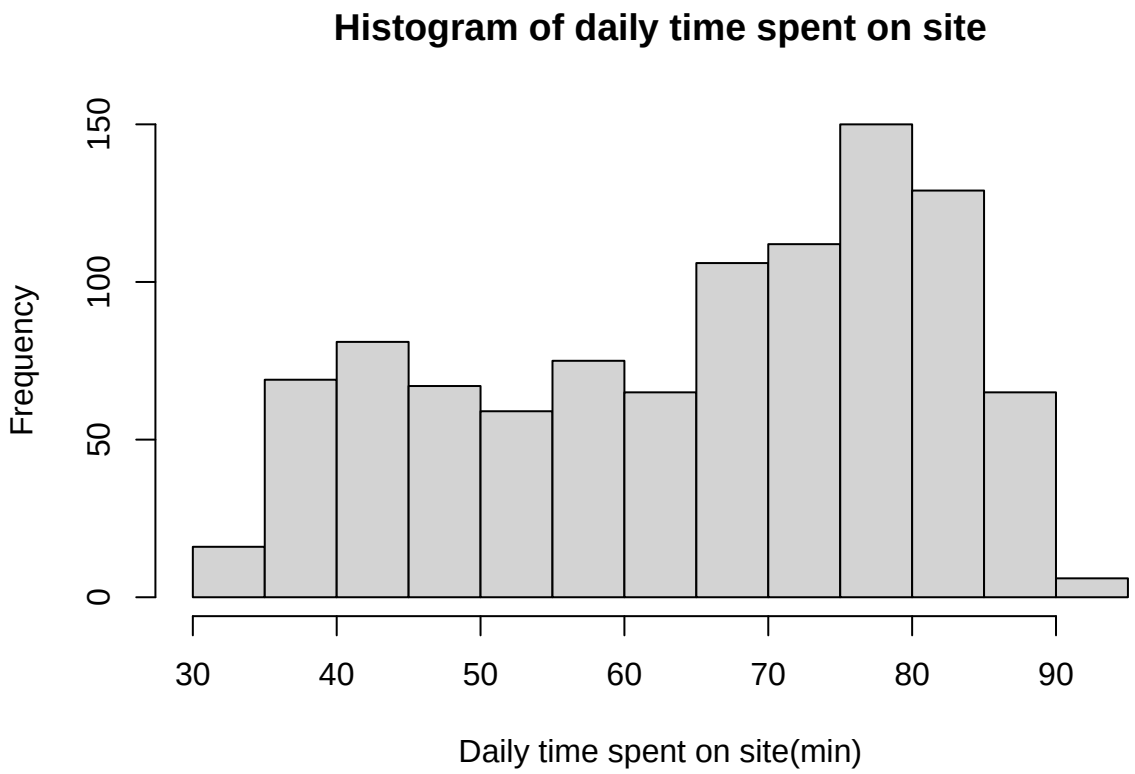
```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##             2.513371e+02            7.718611e+01            1.799524e+08
##     Daily.Internet.Usage
##             1.927415e+03
```

```
# standard deviation
sapply(num, sd)
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                15.853615                8.785562            13414.634022
##     Daily.Internet.Usage
##                43.902339
```

From the standard deviations we see the spreads of the numerical data from their means.

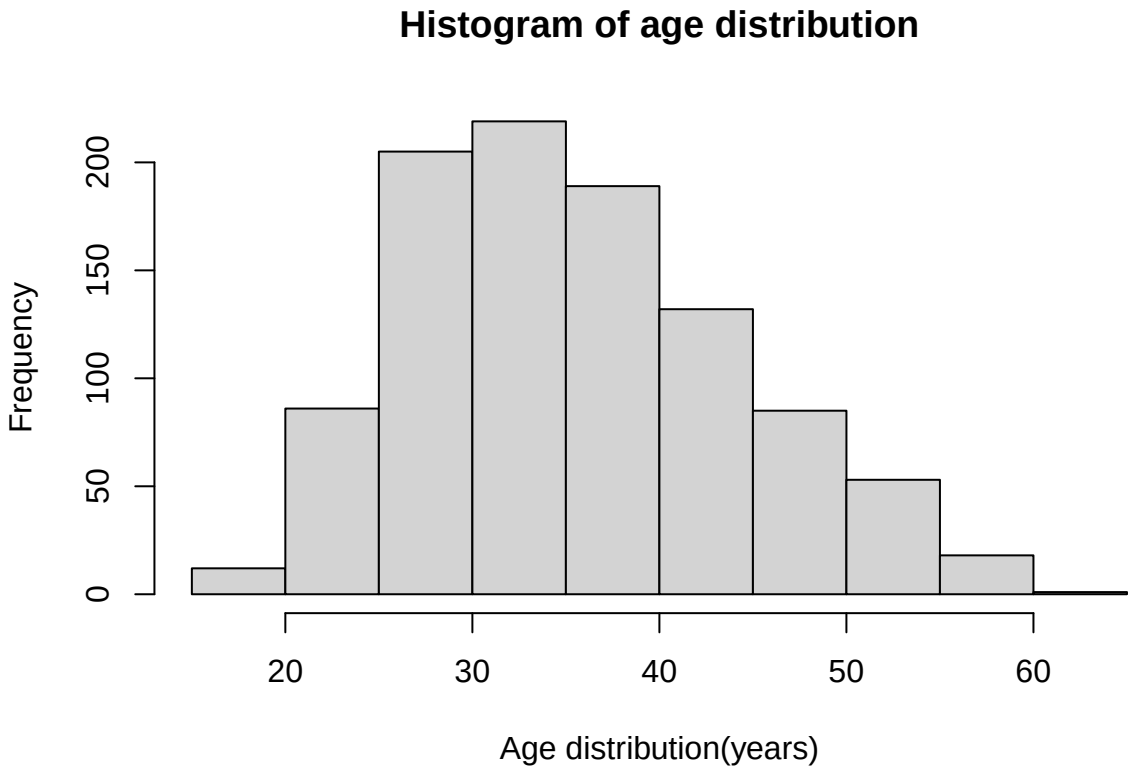
```
hist(num$Daily.Time.Spent.on.Site, xlab= "Daily time spent on site(min)", ylab="Frequency",
     main="Histogram of daily time spent on site")
```



5.1.2 Histograms

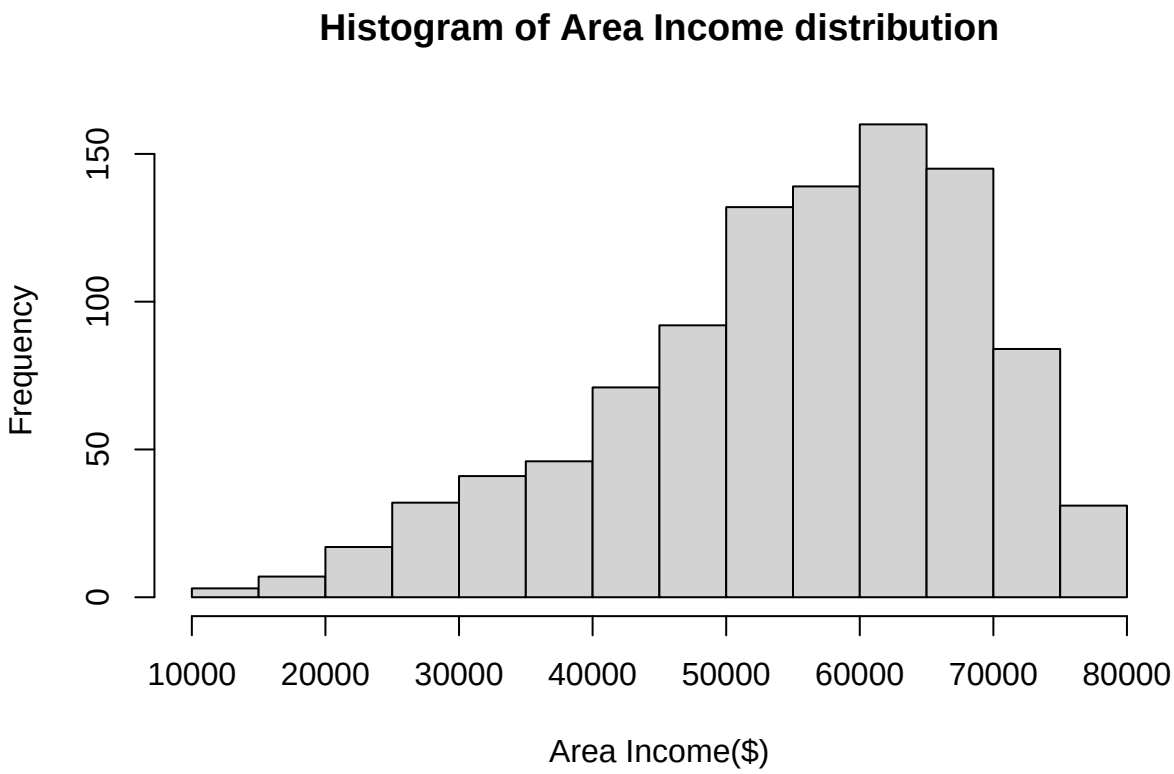
From the graph, we see that the most popular times spent on the site is between 65-85 minutes. The distribution looks like a multi-distribution with left skewness.

```
hist(num$Age, xlab= "Age distribution(years)", ylab="Frequency",
     main="Histogram of age distribution")
```



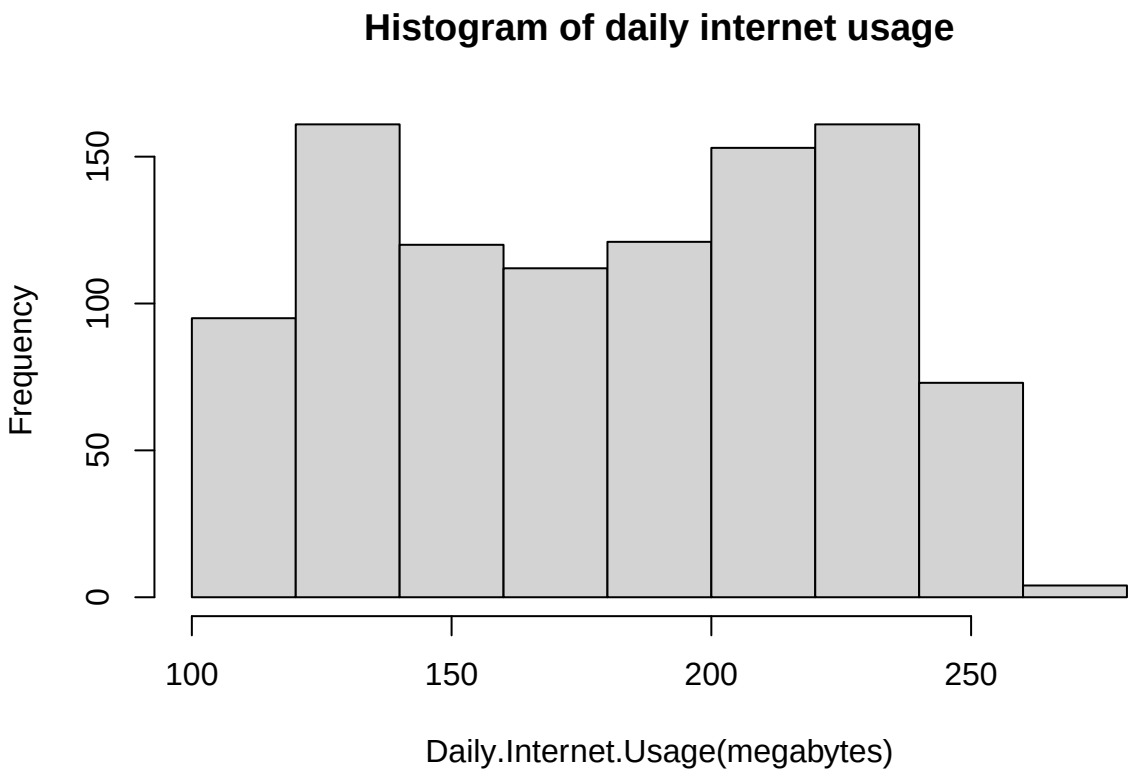
Majority of the people visiting the entrepreneur's blog are between the age of 25-45 years old. We don't know how many of them clicked on the ad, but it's likely that since they are the majority that they will have more clicks. We will check on this during bivariate analysis. The data is almost a normal distribution with a bit of skewness to the right.

```
hist(num$Area.Income, xlab= "Area Income($)", ylab="Frequency",
     main="Histogram of Area Income distribution")
```



The plot shows an obvious left skewness. This is not unusual. Earlier from the boxplot, we saw outliers on the lower side of the whisker. From the plot people who visited the blog are from high income areas($50,000 to 75,000). This makes sense because with high income people are able to easily obtain power and internet needed to visit the blog. Unlike in the low income areas, people can not afford internet as they have to deal with more pressing needs.

```
hist(num$Daily.Internet.Usage, xlab= "Daily.Internet.Usage(megabytes)", ylab="Frequency",
     main="Histogram of daily internet usage")
```



We observe a bimodal distribution here. We have a group of people who majorly use between 125 to about 140 megabytes on the blog, while another group that majorly use between 200 to about 240 megabytes on the blog.

```
advertising %>%
    group_by(Clicked.on.Ad) %>%
    summarize(frequency = n())
```

5.1.3 Univariate analysis of categorical variables using frequency distribution table

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

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

The distribution of the number of people who clicked on adds and the ones that did not click on ad is equal. This is great because it means that our data is balanced.

```
advertising %>%
    group_by(Male) %>%
    summarize(frequency = n())
```

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

```
## # A tibble: 2 x 2
##   Male  frequency
```

```
##   <chr>     <int>
## 1 0           519
## 2 1           481
```

We have majority non males visiting the blog.

```
advertising %>%
    group_by(Country) %>%
    summarize(frequency = n()) %>%
    ungroup() %>%
    arrange(desc(frequency)) %>%
    head(20)
```

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

```
## # A tibble: 20 x 2
##    Country                frequency
##    <chr>                      <int>
##  1 Czech Republic                 9
##  2 France                         9
##  3 Afghanistan                    8
##  4 Australia                      8
##  5 Cyprus                         8
##  6 Greece                         8
##  7 Liberia                        8
##  8 Micronesia                     8
##  9 Peru                           8
## 10 Senegal                        8
## 11 South Africa                   8
## 12 Turkey                         8
## 13 Albania                        7
## 14 Bahamas                        7
## 15 Bosnia and Herzegovina         7
## 16 Burundi                        7
## 17 Cambodia                       7
## 18 Eritrea                        7
## 19 Ethiopia                       7
## 20 Fiji                           7
```

Here we have top 20 most popular countries where the most individuals visit the blog.

5.2 Bivariate Analysis

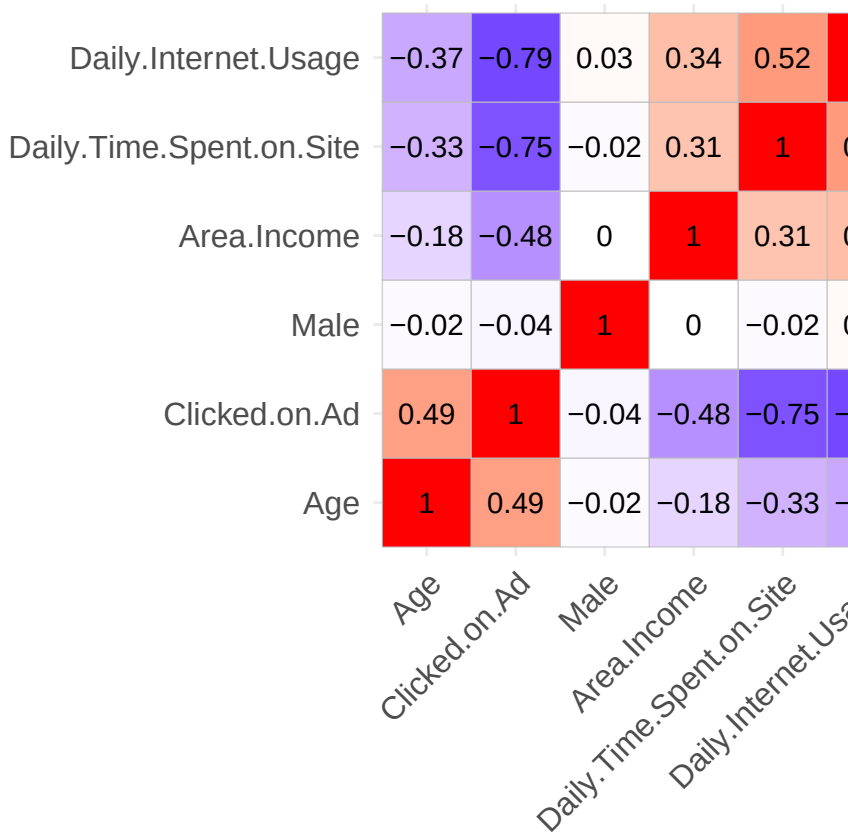
```
advertising$Male <- as.integer(advertising$Male)
advertising$Clicked.on.Ad <- as.integer(advertising$Clicked.on.Ad)
num_var <- select(advertising, Daily.Time.Spent.on.Site, Age, Area.Income, Daily.Internet.Usage, Male, C
#cor(advertising)
cor(num_var, method = "pearson")
```

5.2.1 Correlation calculation using pearson method

```
##                          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
```

```
advertising %>%
    select_if(is.numeric) %>%
    cor %>%
    ggcorrplot(lab = TRUE, hc.order = TRUE)
```



5.2.2 Correlation matrix visualization

Here we get to compare correlations between the target variable (clicked on ad) with other variables. Observations:

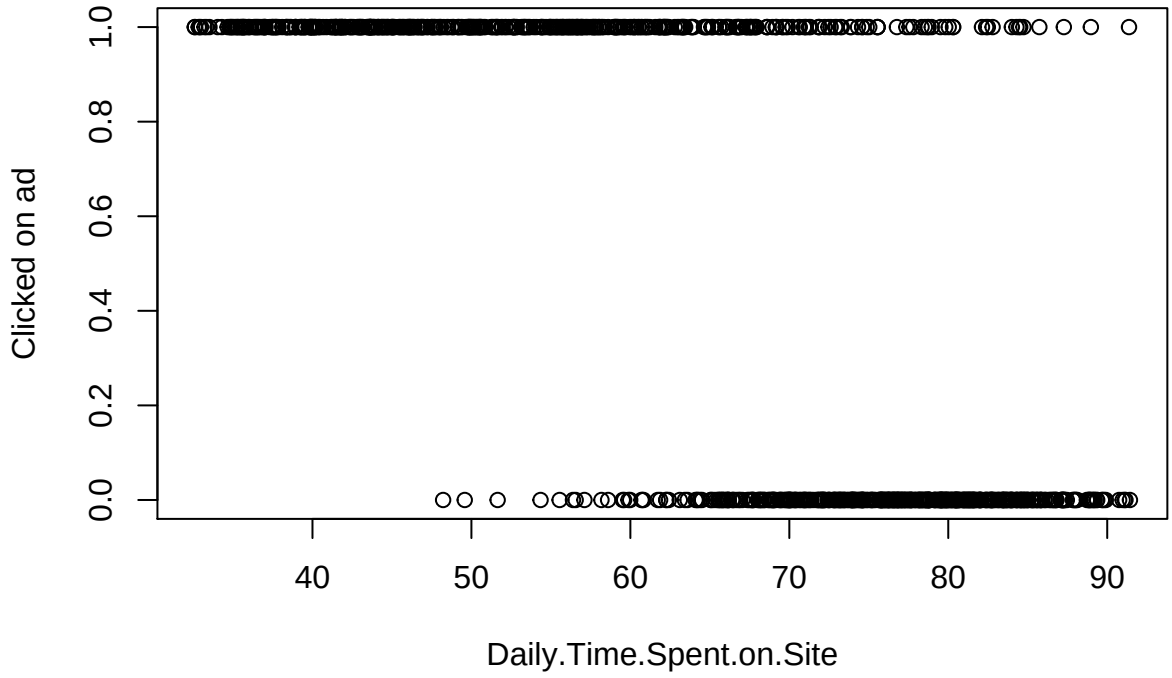
1. The target variable has a positive moderate correlation to age (corr coefficient of 0.49).
2. The target variable is weakly correlated to male column. This means that there is a very weak relationship between a person being male or not and clicking the ad on entrepreneur's blog.
3. The target variable is moderately negatively correlated to Area Income. Meaning that the target variable are usually moving in opposite direction. Which can be interpreted as: the lower income areas have high click rate than low income areas.

4. The target variable is strongly negatively correlated to daily time spent on site. Here we see a strong relationship among the two variable where the two variable are moving in the opposite direction. This could also mean that individuals spending less time on the blog are more likely to click on the ad.
5. We see similar relationship with daily internet usage and target variable as we did with daily time spent on the site. This could also mean that individuals with less internet usage are likely to click on the ads.

5.2.3 Scatter plots

1. Clicked on ad with Daily.Time.Spent.on.Site scatter plot.

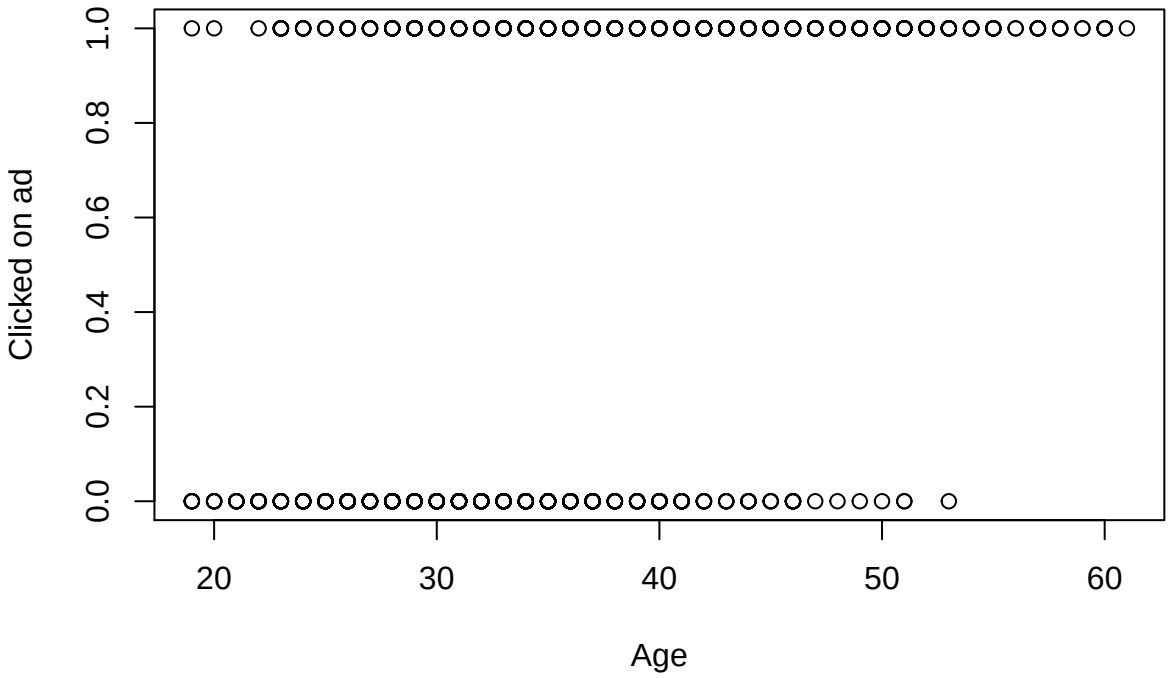
```
plot(advertising$Daily.Time.Spent.on.Site, advertising$Clicked.on.Ad, xlab="Daily.Time.Spent.on.Site", y
```



From the scatter plot, we don't see any relationship between the target variable and daily time spent on site column.

2. Clicked on ad with Age scatter plot

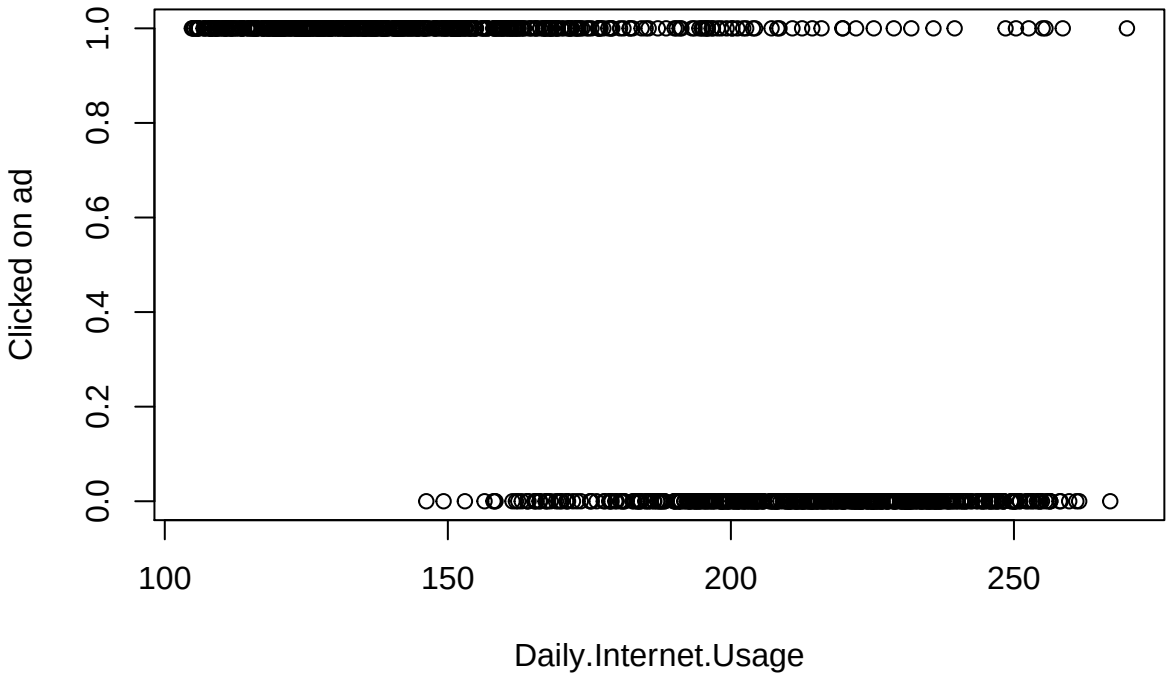
```
plot(advertising$Age, advertising$Clicked.on.Ad, xlab="Age", ylab="Clicked on ad")
```



From the scatter plot, we dont see any relationship between the target variable and age column

3. Clicked on ad with Daily.Internet.Usage scatter plot

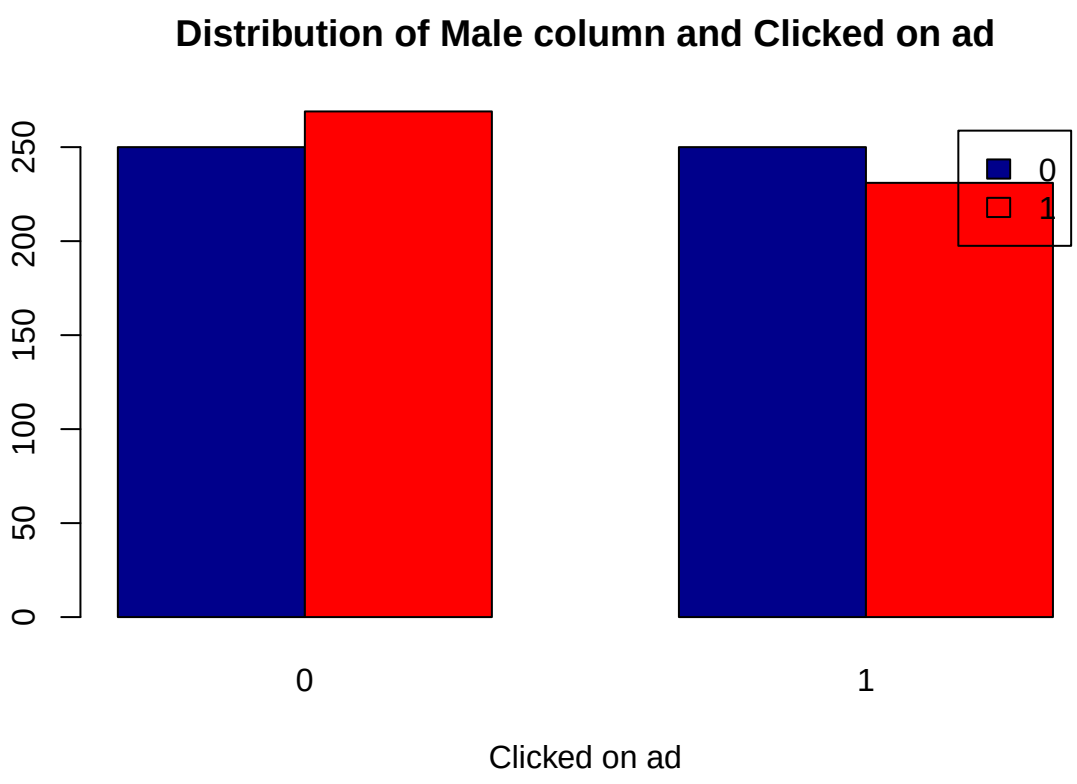
```
plot(advertising$Daily.Internet.Usage, advertising$Clicked.on.Ad, xlab="Daily.Internet.Usage", ylab="Cl
```



There are no insights from the scatterplots as shown above.

5.2.4 More bivariate analysis Visualization of Male column with target variable

```
counts <- table(advertising$Clicked.on.Ad, advertising$Male)
barplot(counts, main="Distribution of Male column and Clicked on ad",
  xlab="Clicked on ad",col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)
```



There was a slightly higher number of non males who clicked on the ad as shown by the frequency table below.

```
advertising %>%
    group_by(Clicked.on.Ad, Male) %>%
    summarize(frequency = n())
```

```
## `summarise()` regrouping output by 'Clicked.on.Ad' (override with `.groups` argument)
```

```
## # A tibble: 4 x 3
## # Groups:   Clicked.on.Ad [2]
##   Clicked.on.Ad  Male frequency
##           <int> <int>     <int>
## 1             0     0       250
## 2             0     1       250
## 3             1     0       269
## 4             1     1       231
```

Comparing countries with click on ad.

```
advertising %>%
    group_by(Clicked.on.Ad, Country) %>%
    summarize(frequency = n())
```

```
## `summarise()` regrouping output by 'Clicked.on.Ad' (override with `.groups` argument)
```

```
## # A tibble: 430 x 3
## # Groups:   Clicked.on.Ad [2]
##    Clicked.on.Ad Country                                    frequency
##            <int> <chr>                                          <int>
```

```
##  1            0 Afghanistan                                        3
##  2            0 Albania                                            3
##  3            0 Algeria                                            3
##  4            0 American Samoa                                     2
##  5            0 Angola                                             3
##  6            0 Anguilla                                           3
##  7            0 Antarctica (the territory South of 60 deg S)       1
##  8            0 Antigua and Barbuda                                1
##  9            0 Argentina                                          1
## 10            0 Armenia                                            2
## # ... with 420 more rows
```

6. Recommendations

1. From the correlation matrix we saw a strong negative correlation between target variable and area income. This means that individuals who live in areas of low income were likely to click on the add compared to individuals from high income areas. We recommend that the entrepreneur focuses on areas of low income because the individuals from this areas will likely click on the ad.

2. We also saw that the target variable was strongly negatively correlated to daily time spent on the blog. It is possible that individuals spending less time on the blog go into the blog to check on any new course ads. We recommend that the entrepreneur find ways of attracting the attention of the individuals who spend longer times in the blog so that they can also click on her ads.

3. Finally, from the correlation matrix we noticed that the target variable was very weakly correlated to the male column. This means that there is no strong relationship between the gender and the target variable so the entrepreneur can continue with the advertisement without worrying about gender bias.