Report on likeliness of person to click on advertisement in a blog.

Precy Mae

9/3/2020

Defining our question

A Kenyan blogger started a cryptocurrency course which she gets to advertise on her blog. Over the years, she has been collecting data on the individuals who visit her blog and whether or not they click on the advertisement.

Moving forward, she would like to know what audience to target in her advertisement. In order to do that she has approached our data science consultancy to give her solutions.

a) Specifying the Question

Performing indepth exploratory data analysis to understand the people who visit her bog and their features in order to best highlight which individuals are most likely to click on her ads.

b) Defining the Metric for Success

- Effectively cleaning our dataset.
- Performing extensive exploratory data analysis.
- Highlighting the individuals most likely to click on the advertisements.

c) Understanding the context

Cryptocurrency is slowly gaining popularity around the world and the need to educate people on this form of money has been on the rise. Many still do not have a clear picture of what cryptocurrency is yet according to the Block-chain Association of Kenya, the total number of bitcoin transactions in Kenya are estimated to be worth over 1.5 million dollars.

Our research topic is therefore, important as it helps us understand which group in the population in general is interested in cryptocurrency.

(Source)

d) Recording the Experimental Design

1) Business Understanding: Understanding the business problem.
2) Reading the data: Getting access to our data and loading it using R-Studio.
3) Checking our data: Understanding our variables and the data types of our data.
4) Data cleaning: Checking for any missing values, duplicates, outliers and solving them.
5) EDA: Visualizing our data using univariate, bivariate and multivariate analysis.

6) Implementing the solution: Using the exploratory data analysis give appropriate solutions to the research problem.

7) Conclusion: Recommend the individuals most likely to click on ad.

PART 1 : Reading our dataset

```
# Load the dataset
df<- read.csv('advertising.csv', header = TRUE)
```

```
# Read the top of our dataset
head(df)
```

```
##   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
```

```
# Read the bottom of our dataset
tail(df)
```

```
##      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
```

```
# To take a look at the numeric and non numeric characters in our dataset.
str(df)
```

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

Our first four columns are numeric while the rest are categorical. Our dataset has 10 columns (variables) 1000 entries (rows)

PART 2: Data Cleaning

a) Missing Values

```
# Checking for missing values in our dataset.
# Columns with missing values
colSums(is.na(df))
```

```
## 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
```

```
# The sum total missing values in the dataset
# Sum of missing
sum(is.na(df))
```

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

There are no missing values in our dataset.

b) Duplicates

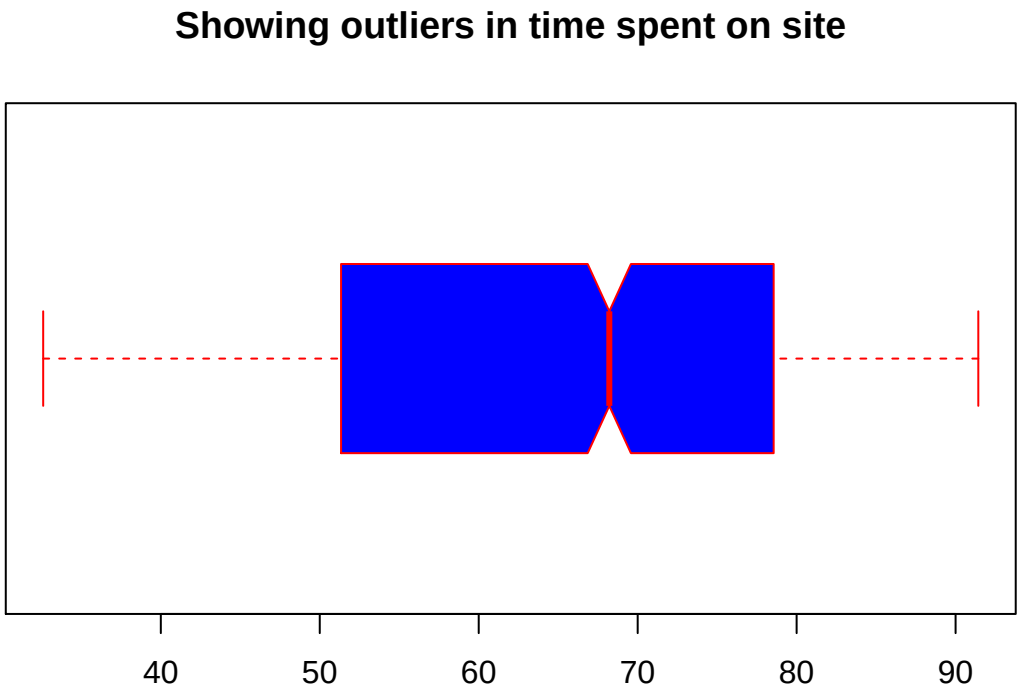
```
# To show our duplicated rows

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

```
##  [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 duplicated rows.

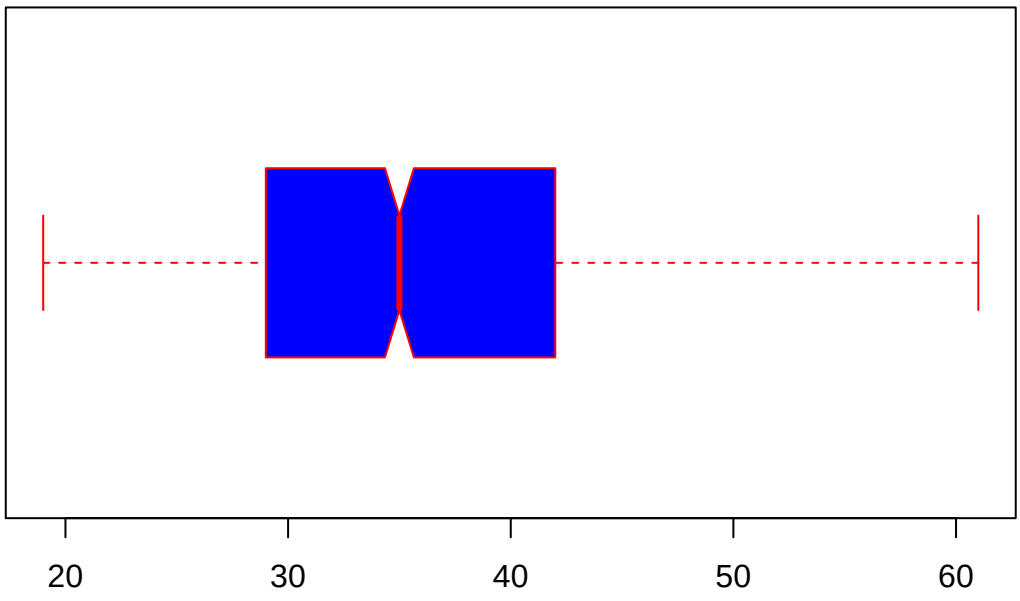
c) Dealing with outliers

```
# Showing outlier using box plots
# A box plot showing outlier in the time spent on site column
boxplot(df$Daily.Time.Spent.on.Site,
main = "Showing outliers in time spent on site",
col = "blue",
border = "red",
horizontal = TRUE,
notch = TRUE)
```



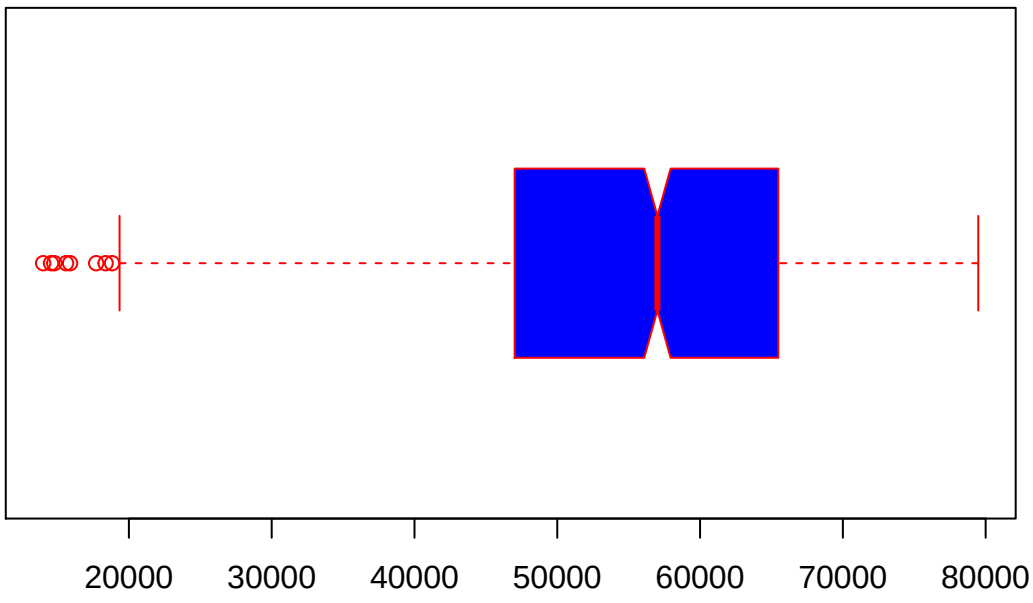
There are no outliers in time spent on site.

```
# A box plot showing outlier in the age column
boxplot(df$Age,
main = "Showing outliers in age",
col = "blue",
border = "red",
horizontal = TRUE,
notch = TRUE)
```



There are no outliers in age

```
# A box plot showing outlier in the area income column
boxplot(df$Area.Income,
main = "Showing outliers in area income column",
col = "blue",
border = "red",
horizontal = TRUE,
notch = TRUE)
```



There are a few outliers below the income 20,000 dollars. We will not remove these outliers

because they are reasonable.

PART 3 : Univariate Analysis

a) Statistical Summary

```
# Viewing the statistical summary of our dataset
summary(df)
```

```
##  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
```

- The average age is 36, the average income is 55,000, the average time spent is 65 and the average internet usage is 180 daily.
- The minimum time spent is 32, youngest person in the site is 19 years old, the smallest income is 13996.

b) Variances and Standard Deviation

- Variance is a way to measure how far a set of numbers is spread out.
- Standard Deviation is a number used to tell how measurements for a group are spread out from the mean. A low standard deviation means that most of the numbers are close to the average.

```
# Variance and standard deviation
age = df$Age     # the age
var(age)         # apply the variance function
```

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

```
sd(age)          # apply the standard deviation function
```

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

Age has a standard deviation of 8(close to mean) and a variance of 77

```r
# Variance and standard deviation
inc = df$Area.Income    # the Area income
var(inc)        # apply the variance function
```

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

```r
sd(inc)         # apply the standard deviation function
```

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

Area income has a standard deviation of 13414 (far from mean) and a variance of 179,952,406

```r
# Variance and standard deviation
time= df$Daily.Time.Spent.on.Site   # the time spent
var(time)        # apply the variance function
```

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

```r
sd(time)         # apply the standard deviation function
```

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

Daily time spent on site has a variance of 251 and a standard deviation of 15 which means most values are closer to the mean.

```r
# Variance and standard deviation
int= df$Daily.Internet.Usage   # the internet usage
var(int)        # apply the variance function
```

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

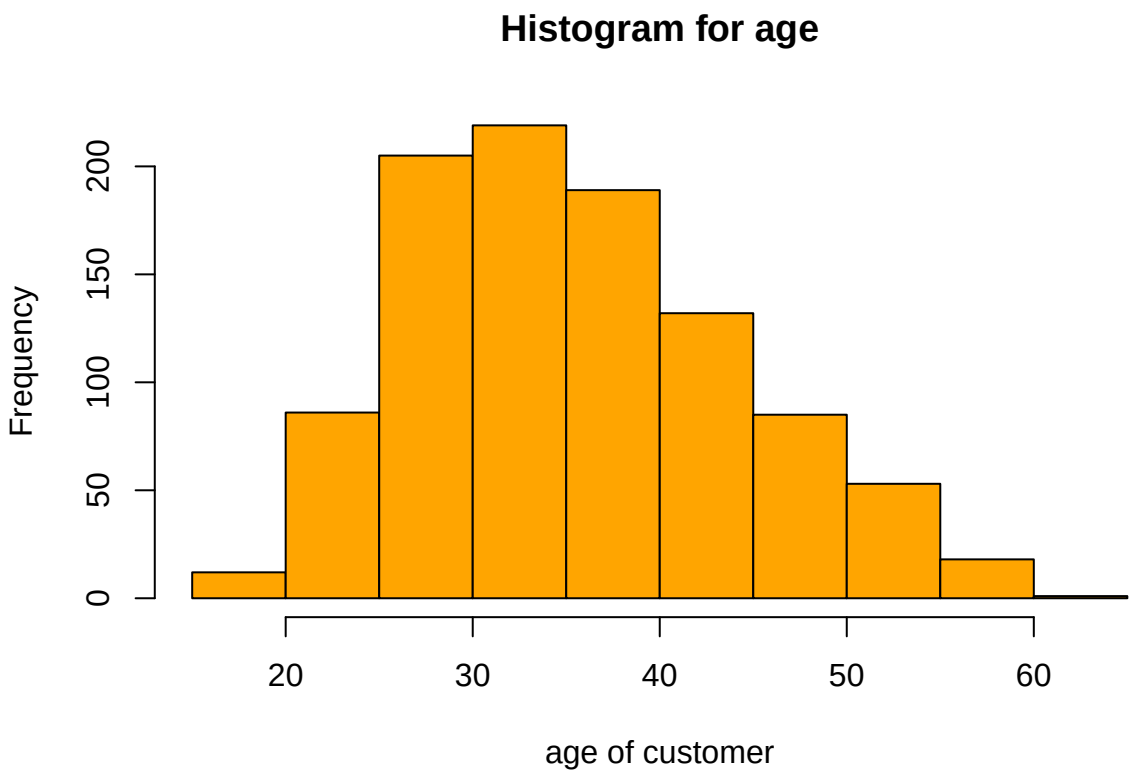
```r
sd(int)         # apply the standard deviation function
```

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

Daily internet usage has a variance of 1927 and a standard deviation of 43 which means that most values are fairly close to the mean.

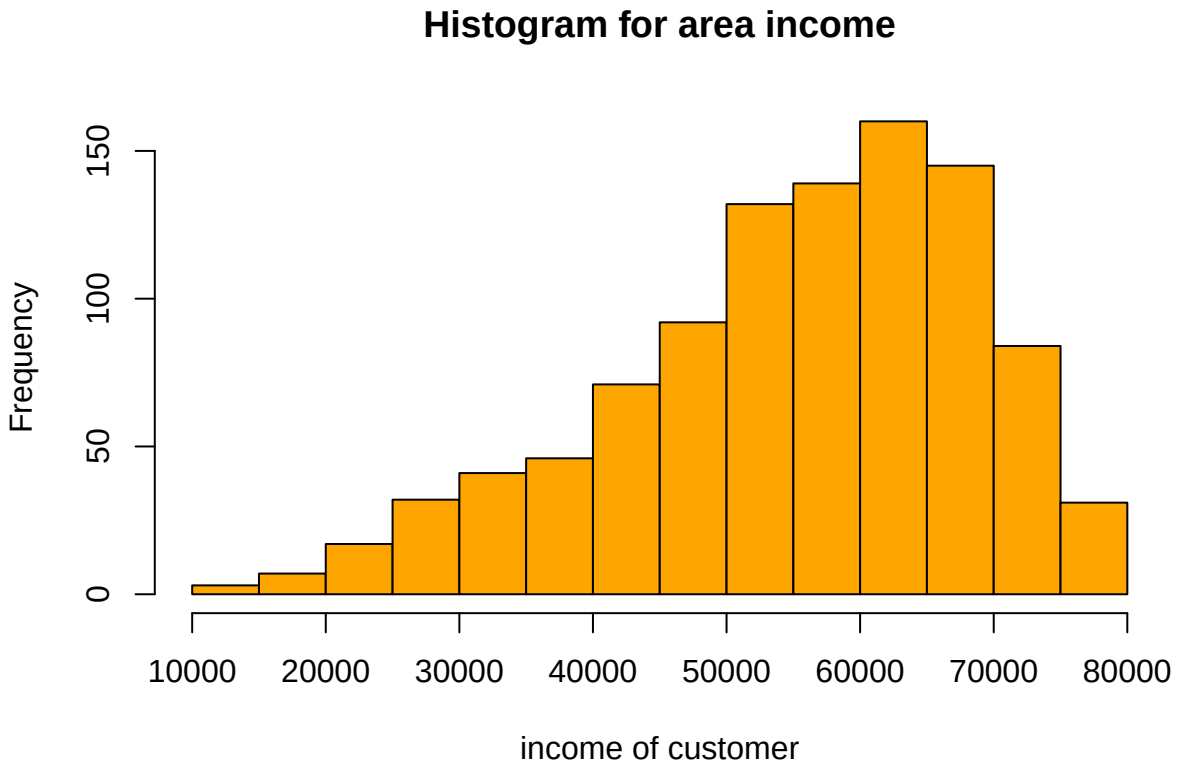
c) histogram

```r
hist(df$Age ,
     col='orange',
     main='Histogram for age',
     xlab= 'age of customer')
```



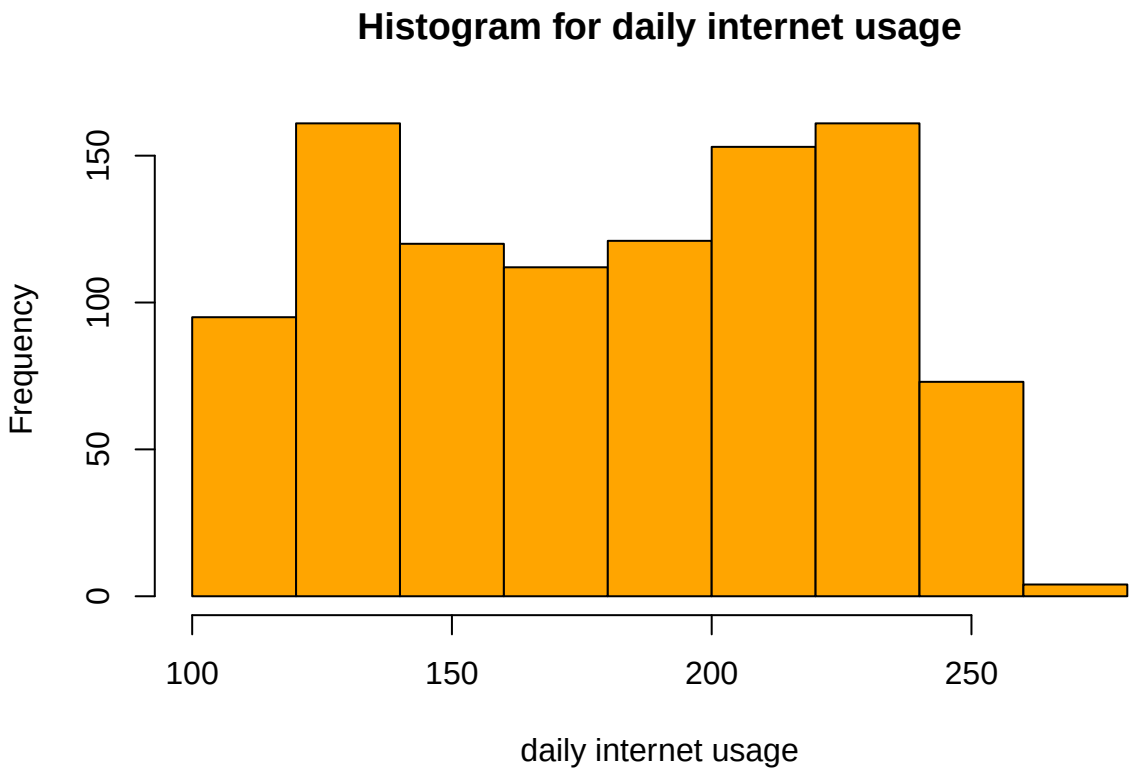
The age variable is almost normally distributed

```
hist(df$Area.Income ,
     col='orange',
     main='Histogram for area income',
     xlab= 'income of customer')
```



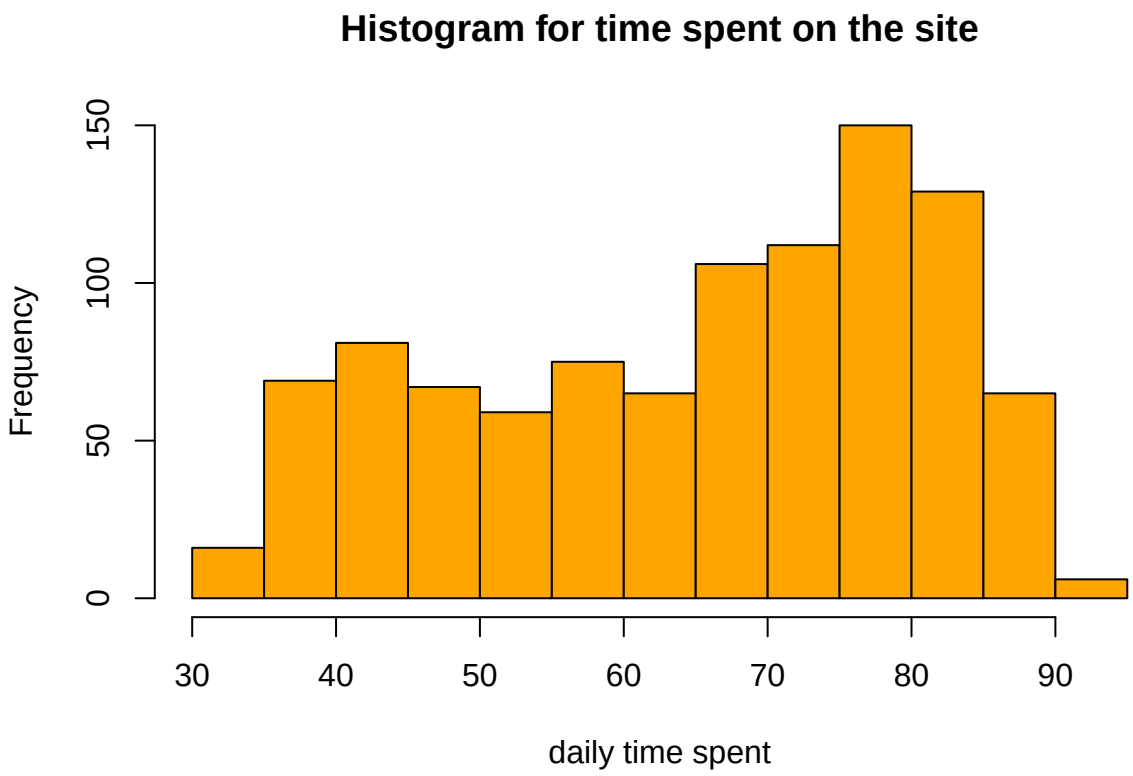
The income of the customers is negatively skewed as the tail is on the left/ left skewed.

```
hist(df$Daily.Internet.Usage ,
     col='orange',
     main='Histogram for daily internet usage',
     xlab= 'daily internet usage')
```



Daily internet usage seems to be bi-modal (with two modes). It also appears to be normally distribution.

```
hist(df$Daily.Time.Spent.on.Site ,
     col='orange',
     main='Histogram for time spent on the site',
     xlab= 'daily time spent')
```



The data on time spent on the site seems to be slightly negatively skewed.

PART 4 : Bivariate Analysis

a) Covariance

```
# Checking the relationship between age and likelihood to click on ad.
# Assigning the age column to the variable age
# ---
#
age <- df$Age

# Assigning the clicked on ad column to the variable clicked on ad
# ---
#
ad<- df$Clicked.on.Ad

# Using the cov() function to determine the covariance
# ---
#
cov(age, ad)
```

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

There is a positive relationship between age and whether or not a person will click on ad.

```
# Checking the relationship between time spent on site and likelihood to click on ad.
time <- df$Daily.Time.Spent.on.Site
cov(time, ad)
```

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

There is a negative relationship between time spent on site and likelihood to click on ad.

```
# Checking the relationship between internet usage and likelihood to click on advert.
int <- df$Daily.Internet.Usage
cov(int, ad)
```

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

There is a negative relationship between internet usage and likelihood of a person to click on ad.

```
# Checking the relationship between area income and whether or not a person will click on ad
inc <- df$Area.Income
cov(inc, ad)
```

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

There is a negative relationship between area income and likelihood to click on ad

b) Correlation

```
# Calculating correlation between age and click on ad
cor(age, ad)
```

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

The relationship is moderately positive between the two.

```
# Showing correlation across variables
install.packages("corrgram")
```

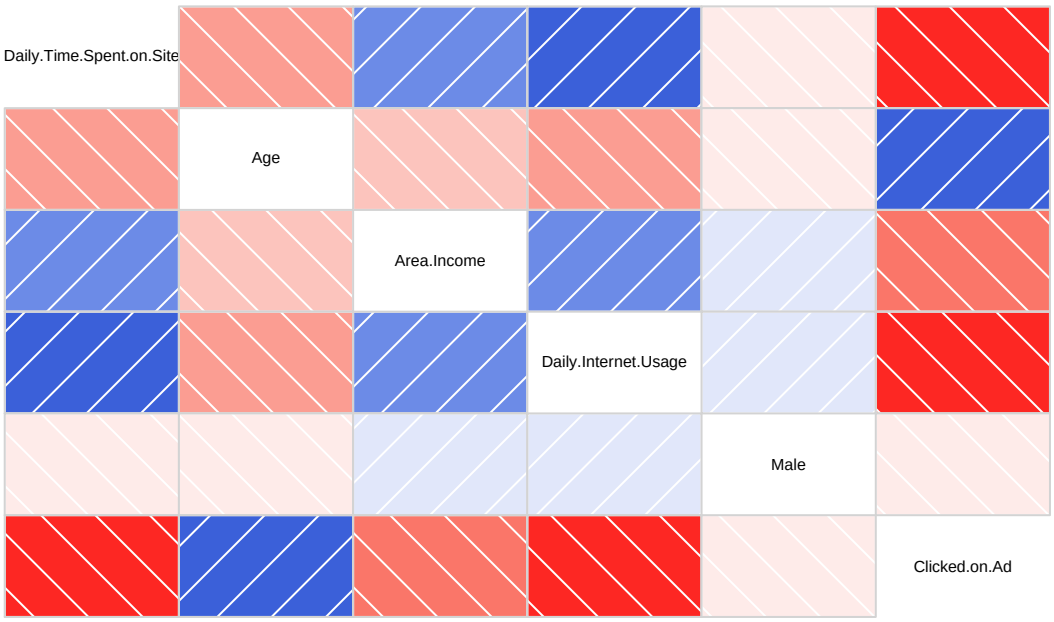
```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(corrgram)
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

```
corrgram(df, order=NULL, panel=panel.shade, text.panel=panel.txt,
           main="Correlogram")
```

Correlogram



- There is a positive correlation between daily internet usage and daily time spent on site.
- There is also a positive relationship between income and daily internet usage.
- There is also a positive relationship between income and daily time spent on site.

c) Count plots

```
# Loading prerequisites
library(ggplot2)
library(ggpubr)
theme_set(theme_pubr())
```

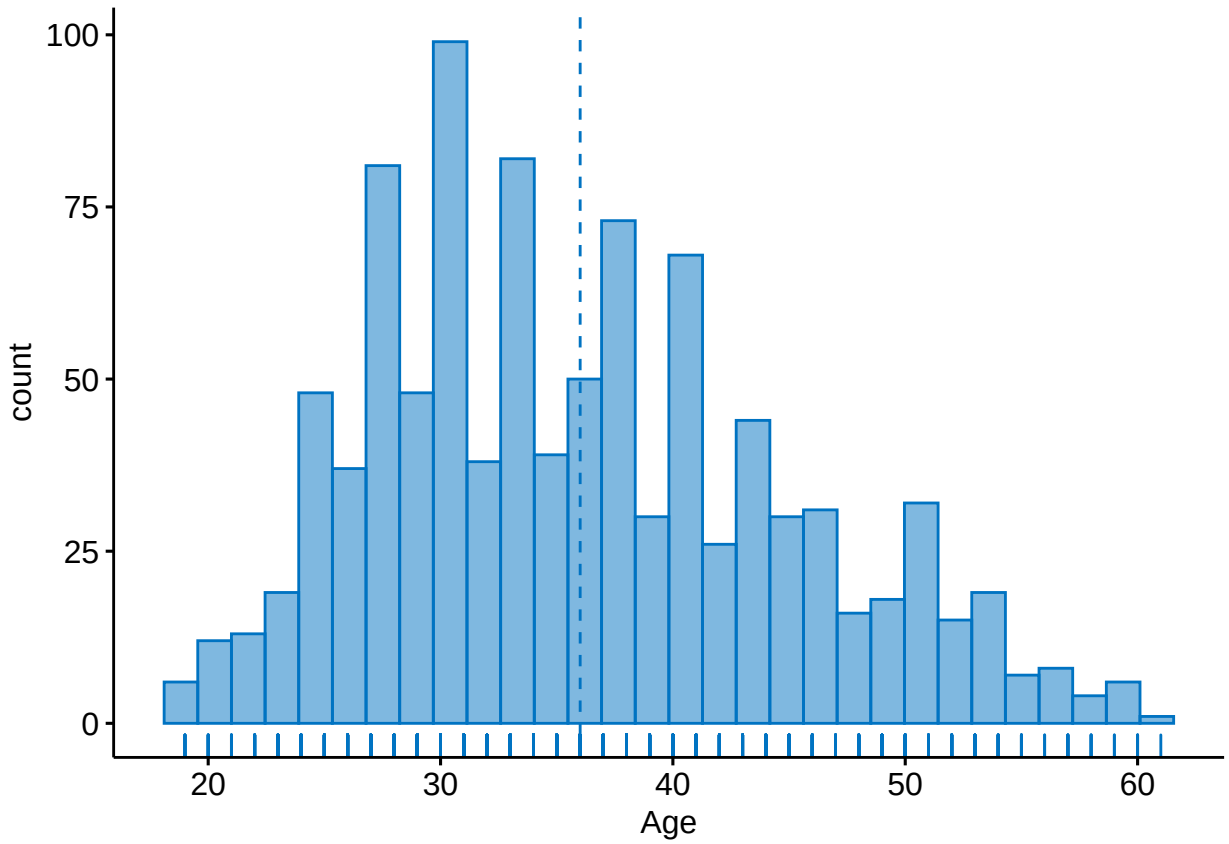
The packages are downloaded

```
# Converting the clicked on ad to categorical
df$Clicked.on.Ad <- as.factor(df$Clicked.on.Ad)
```

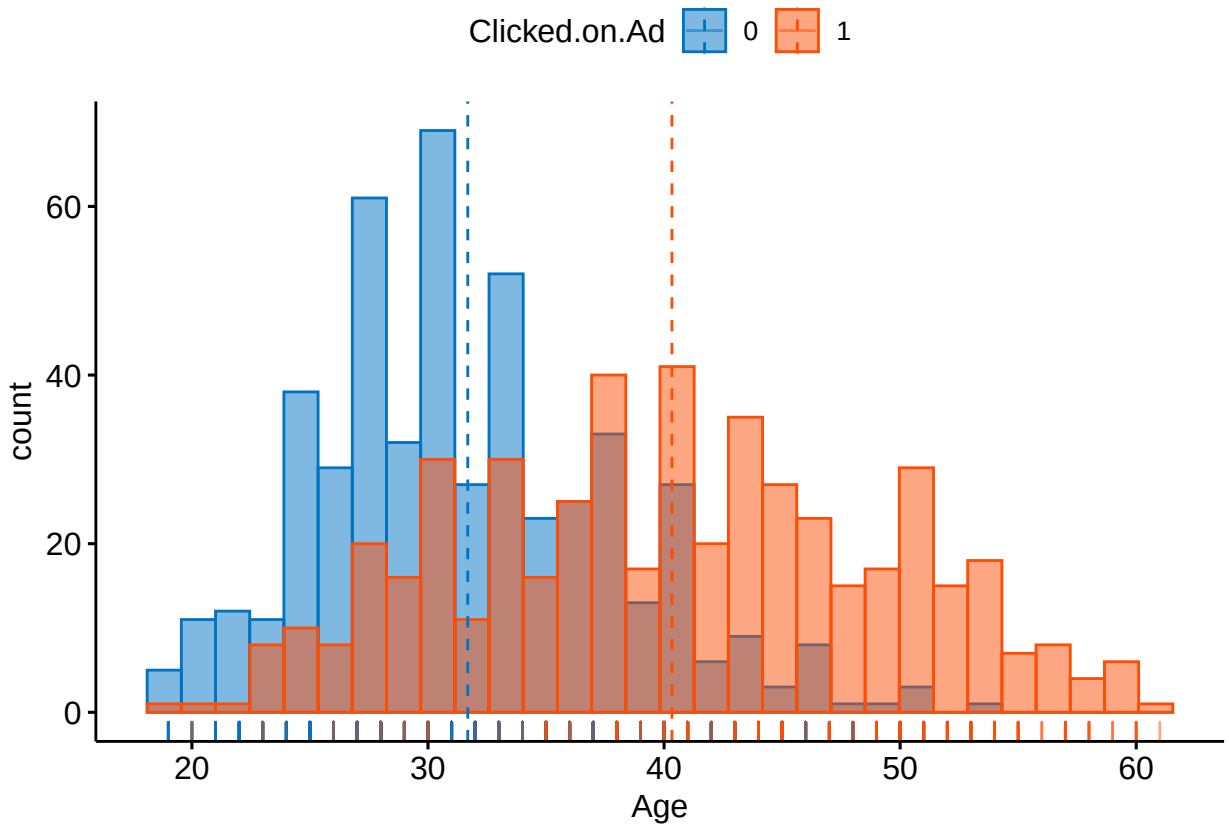
```
# Showing distribution of ages between clicked and not clicked ad
# Use a custom palette
library(ggpubr)
# Basic histogram plot with mean line and marginal rug
gghistogram(df, x = "Age", bins = 30,
            fill = "#0073C2FF", color = "#0073C2FF",
            add = "mean", rug = TRUE)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```

```
## Warning: geom_vline(): Ignoring `data` because `xintercept` was provided.
```




```
# Change outline and fill colors by groups ("sex")
# Use a custom palette
gghistogram(df, x = "Age", bins = 30,
   add = "mean", rug = TRUE,
   color = "Clicked.on.Ad", fill = "Clicked.on.Ad",
   palette = c("#0073C2FF", "#FC4E07"))
```

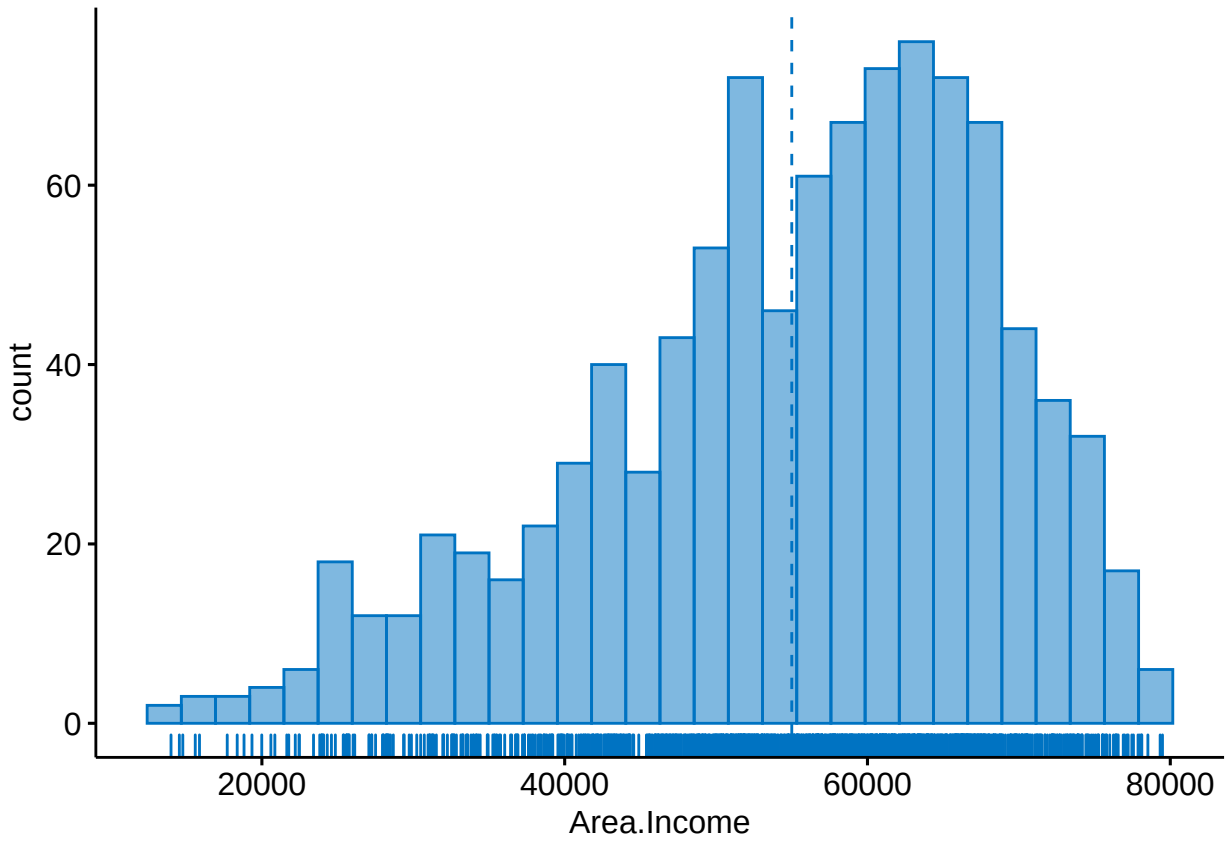


People who clicked on ad are fairly distributed across the ages while those who do not click on ad are mostly under 35 years old.

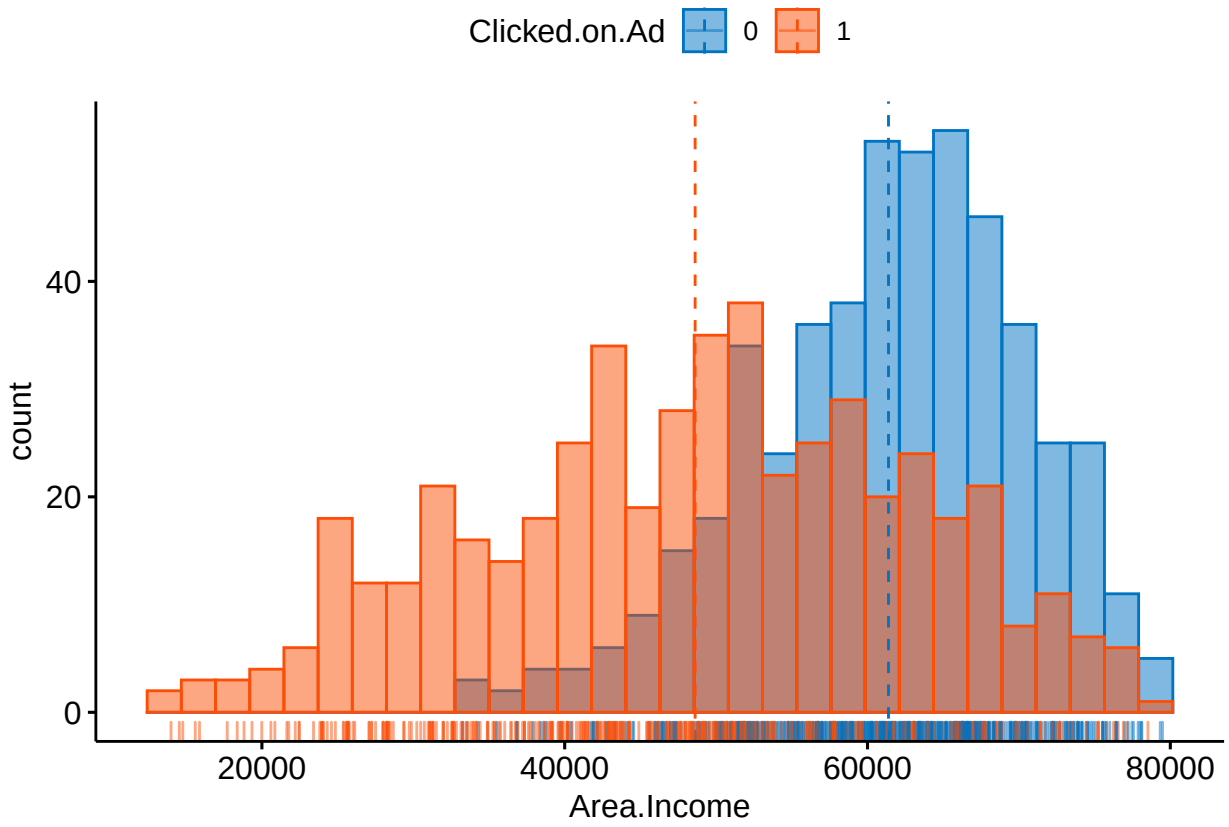
```
# Showing distribution of area income between clicked and not clicked ad
# Use a custom palette
library(ggpubr)
# Basic histogram plot with mean line and marginal rug
gghistogram(df, x = "Area.Income", bins = 30,
            fill = "#0073C2FF", color = "#0073C2FF",
            add = "mean", rug = TRUE)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```

```
## Warning: geom_vline(): Ignoring `data` because `xintercept` was provided.
```




```
# Change outline and fill colors by groups ("sex")
# Use a custom palette
gghistogram(df, x = "Area.Income", bins = 30,
   add = "mean", rug = TRUE,
   color = "Clicked.on.Ad", fill = "Clicked.on.Ad",
   palette = c("#0073C2FF", "#FC4E07"))
```

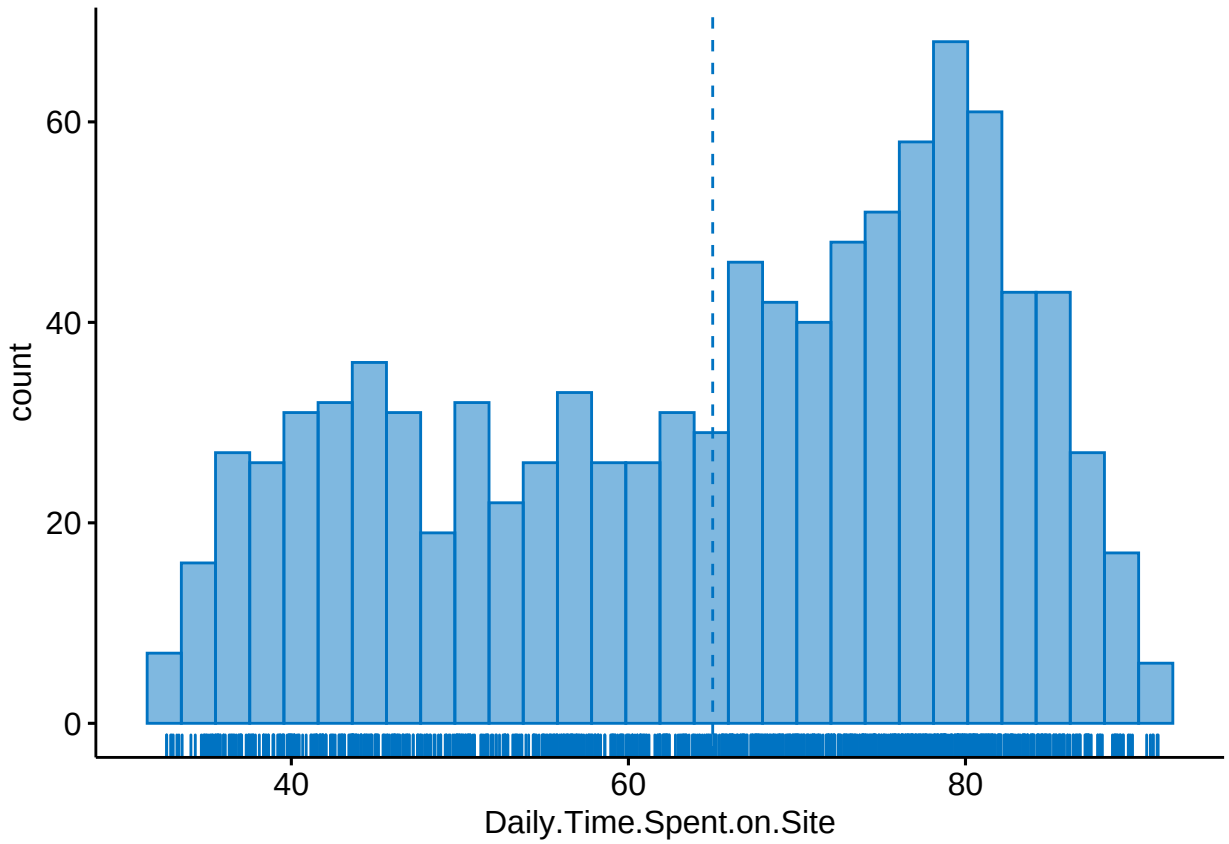


The individuals who clicked on the ad are fairly distributed across area incomes but those with incomes greater than 60,000 were more likely not to click on ad.

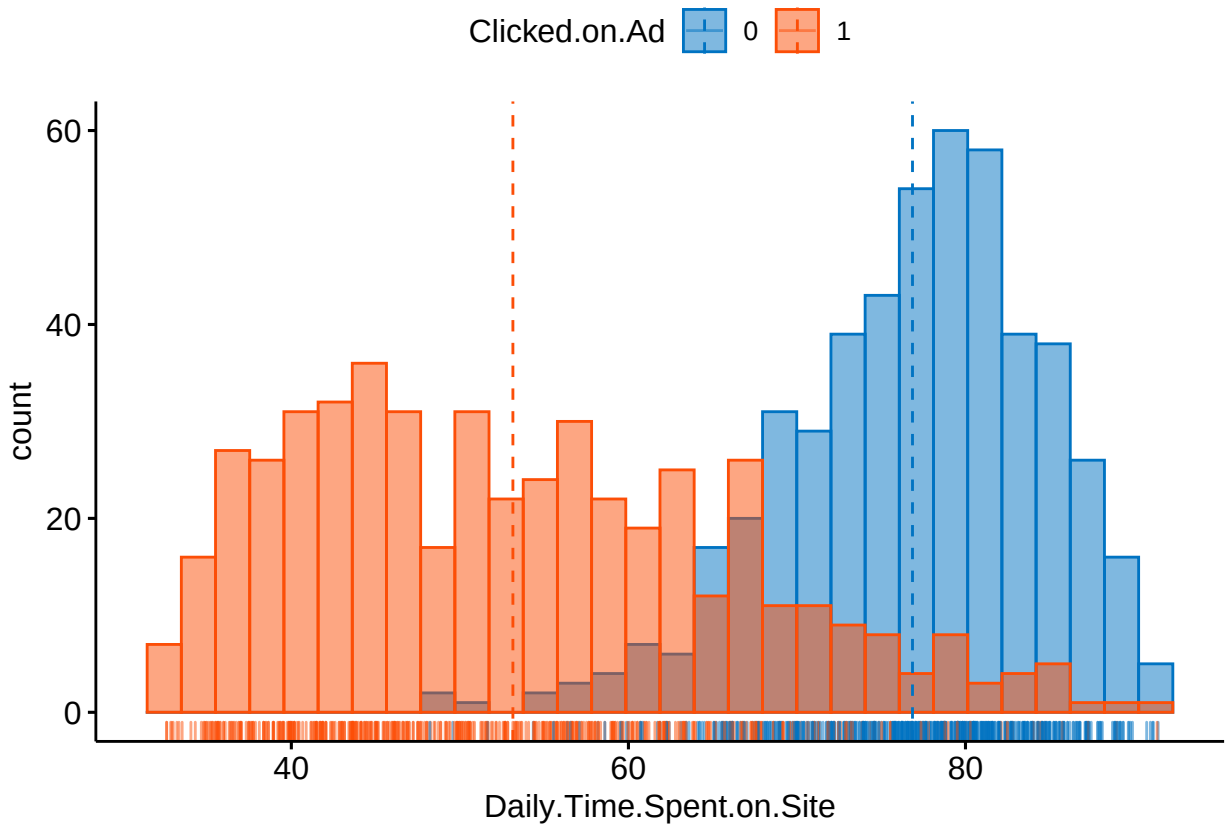
```
# Showing distribution of daily time spent on site between clicked and not clicked ad
# Use a custom palette
library(ggpubr)
# Basic histogram plot with mean line and marginal rug
gghistogram(df, x = "Daily.Time.Spent.on.Site", bins = 30,
            fill = "#0073C2FF", color = "#0073C2FF",
            add = "mean", rug = TRUE)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```

```
## Warning: geom_vline(): Ignoring `data` because `xintercept` was provided.
```




```
# Change outline and fill colors by groups ("sex")
# Use a custom palette
gghistogram(df, x = "Daily.Time.Spent.on.Site", bins = 30,
   add = "mean", rug = TRUE,
   color = "Clicked.on.Ad", fill = "Clicked.on.Ad",
   palette = c("#0073C2FF", "#FC4E07"))
```

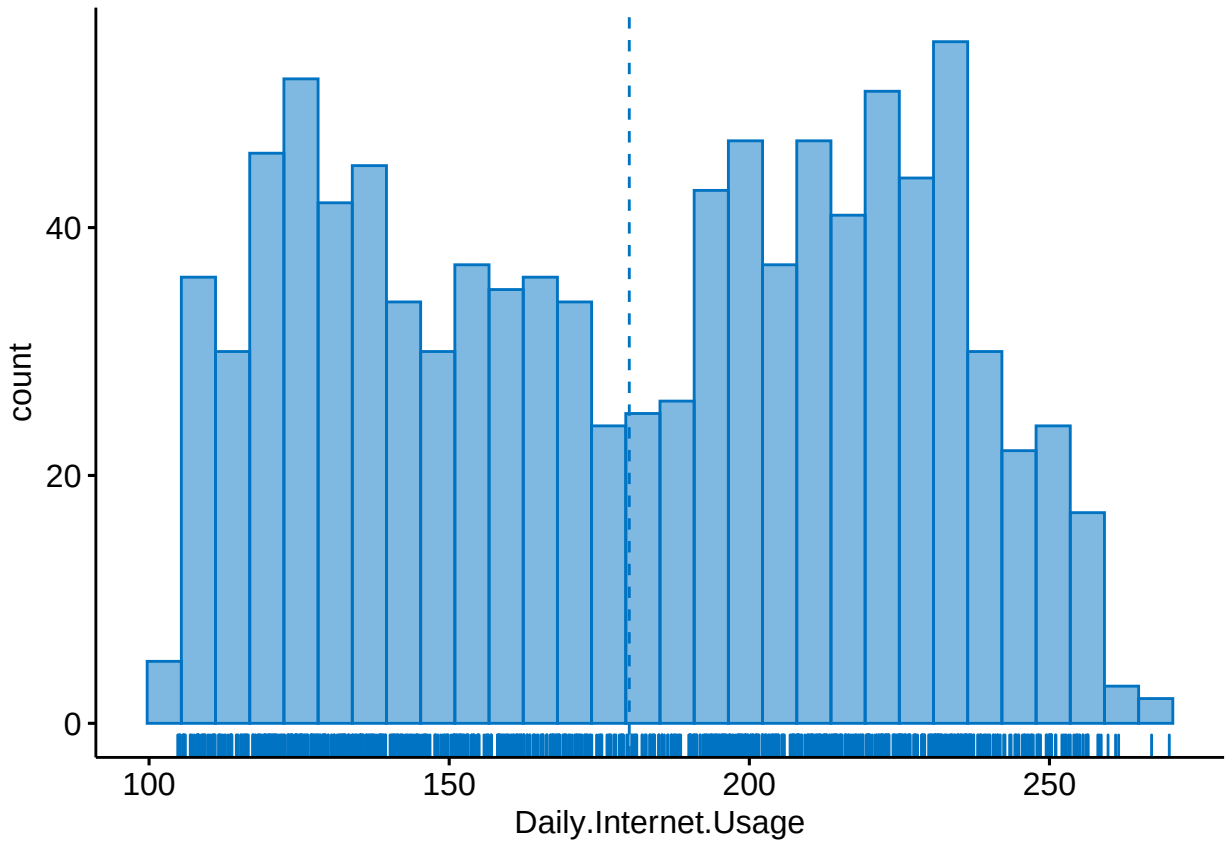


People spending less than 60 minutes on site are more likely to click on ad.

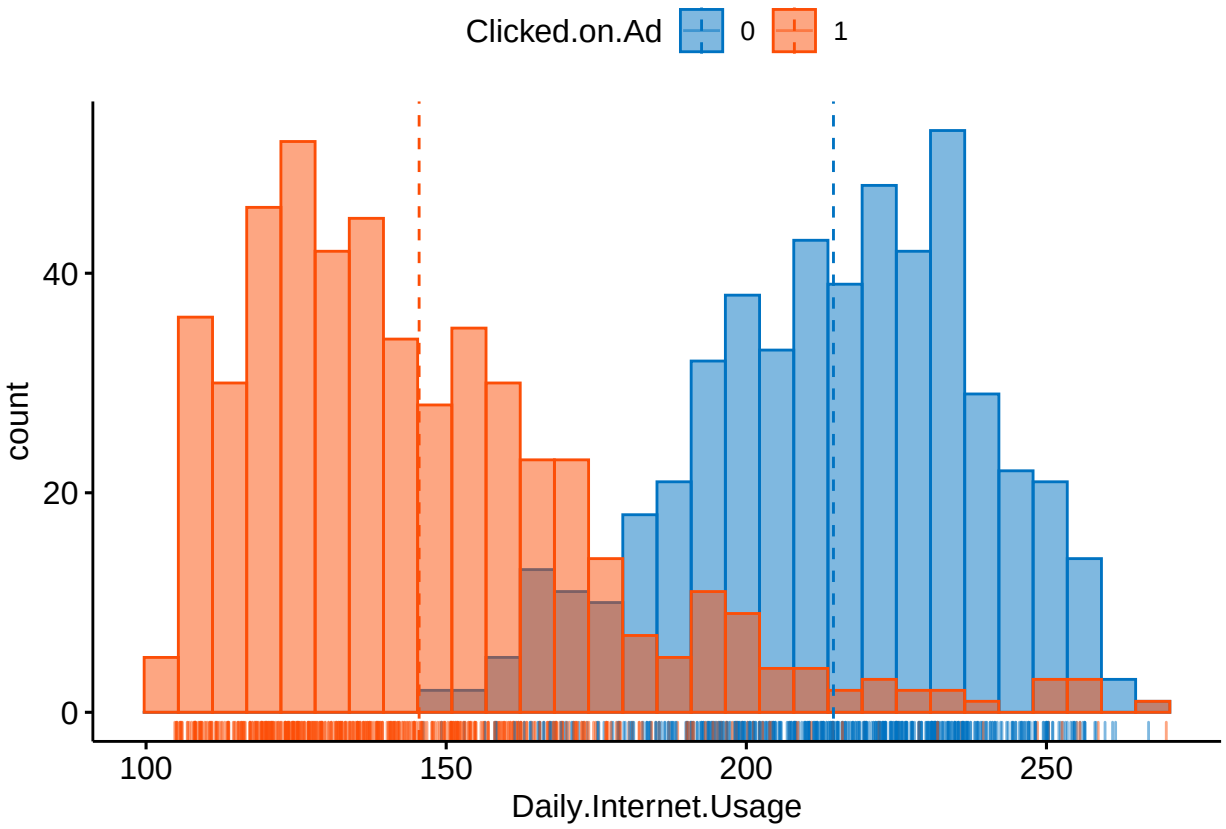
```
#Showing distribution of daily internet usage between clicked and not clicked ad
# Use a custom palette
library(ggpubr)
# Basic histogram plot with mean line and marginal rug
gghistogram(df, x = "Daily.Internet.Usage", bins = 30,
            fill = "#0073C2FF", color = "#0073C2FF",
            add = "mean", rug = TRUE)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```

```
## Warning: geom_vline(): Ignoring `data` because `xintercept` was provided.
```




```
# Change outline and fill colors by groups ("sex")
# Use a custom palette
gghistogram(df, x = "Daily.Internet.Usage", bins = 30,
   add = "mean", rug = TRUE,
   color = "Clicked.on.Ad", fill = "Clicked.on.Ad",
   palette = c("#0073C2FF", "#FC4E07"))
```



Those with high and low internet usage had similar chances to click on advetisement.

PART 4: Multivariate analysis

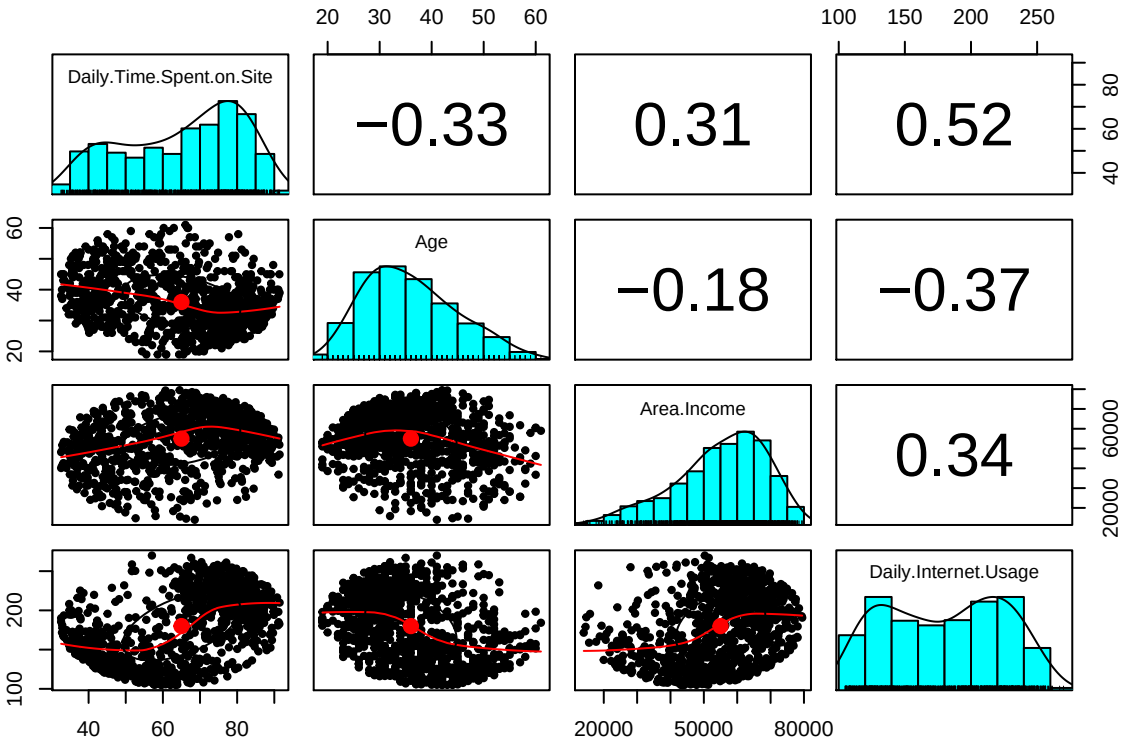
Pairplots

```
# Plotting our pairplots
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(df[,c(1:4)])
```



There is no much relationship between our countinuous variables.

Part 5: Conclusions and Recommendations

- Persons over the age of 40 are more likely to click on the ad compared to the younger counterparts. This could be because they have enough finances to take part in a cryptocurrency transaction.
- Persons who spend less time on the site are most likely to click on the ad therefore we would advise our client to make it interesting to capture someones attention in a short time. The people who spend more time on the site probably have other interests other than cryptocurrency.
- People who click on the ad are fairly distributed across area income levels which means that the course can attract students across income levels.