1.0 Business Understanding

##1.1 Define the question

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.

##1.2 Define the Metric for Success Perform Bivariate and Multivariate analysis giving recommendations

##1.3 Experimental design -Business Understanding -Data Cleaning -Univarite Analysis -Bivariate Analysis -Conclusion -Recommendations

#2.0 Data Preparation

```
#Lets load our dataset

advertising_data = read.csv("advertising.csv")
```

```
#Loading the first 6 rows
head(advertising_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 the dataset dimensions
dim(advertising_data)
```

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

```
#Checking the Column names
colnames(advertising_data)
```

```
##  [1] "Daily.Time.Spent.on.Site" "Age"
##  [3] "Area.Income"              "Daily.Internet.Usage"
##  [5] "Ad.Topic.Line"            "City"
##  [7] "Male"                     "Country"
##  [9] "Timestamp"                "Clicked.on.Ad"
```

```
#Checking the data types
sapply(advertising_data, class)
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                "numeric"                "integer"                "numeric"
##     Daily.Internet.Usage             Ad.Topic.Line                     City
##                "numeric"              "character"              "character"
##                     Male                  Country                Timestamp
##                "integer"              "character"              "character"
##            Clicked.on.Ad
##                "integer"
```

3.0 Data Cleaning

3.1 Missing Values

```
#Checking for missing values per column
colSums(is.na(advertising_data))
```

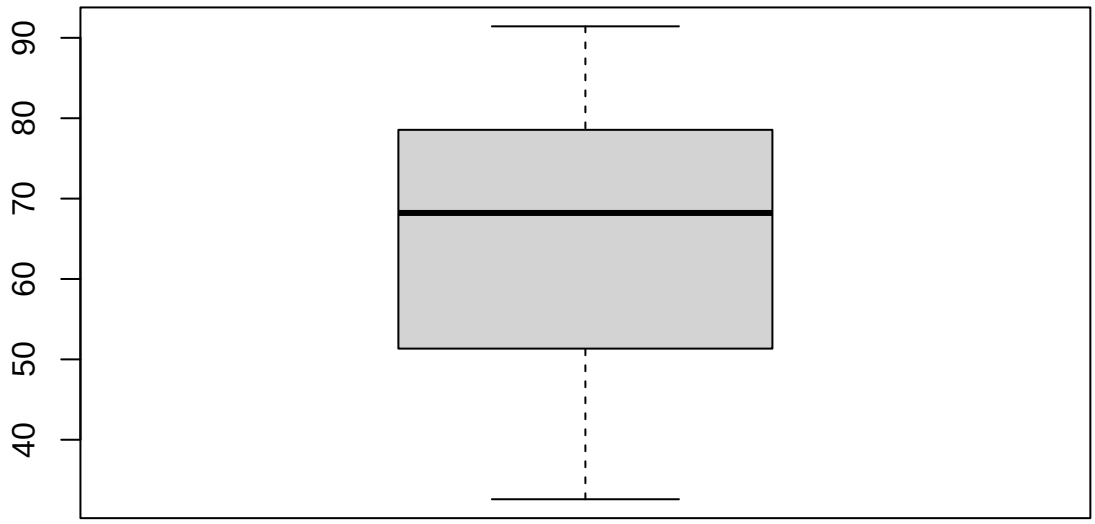
```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                        0                        0                        0
##     Daily.Internet.Usage             Ad.Topic.Line                     City
##                        0                        0                        0
##                     Male                  Country                Timestamp
##                        0                        0                        0
##            Clicked.on.Ad
##                        0
```

#The Dataset has no missing values

Outliers

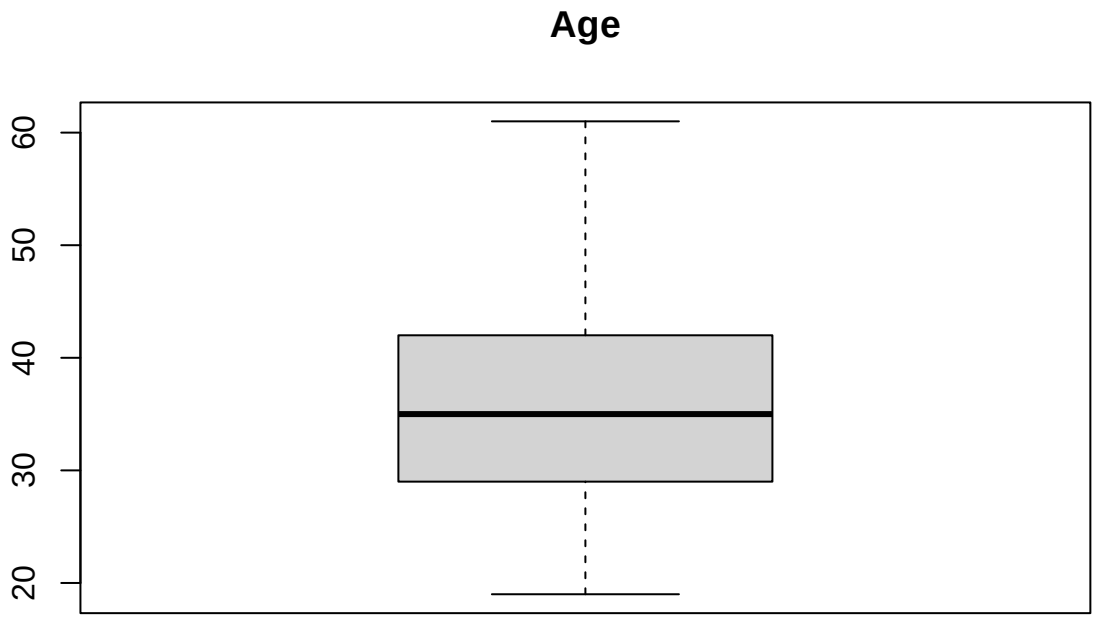
```
#Checking the outliers for Daily time spent, age area income and Daily internet usage

boxplot(advertising_data$Daily.Time.Spent.on.Site, main="Daily Time spent on site")
```

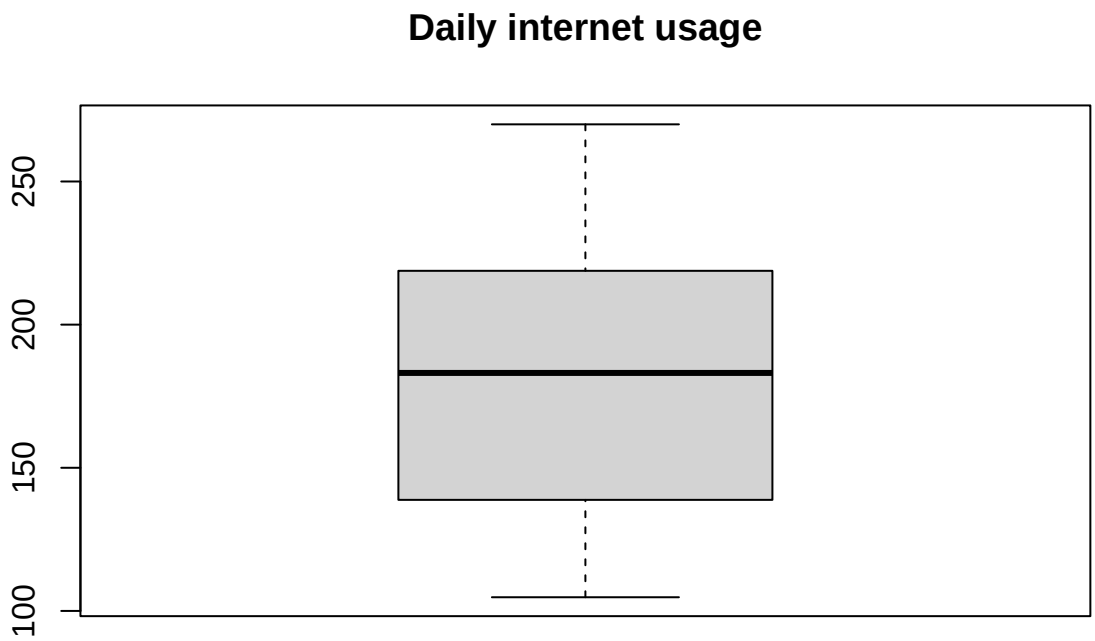



```
boxplot(advertising_data$Area.Income, main = "Area income")
```


```
boxplot(advertising_data$Age, main = "Age")
```




```
boxplot(advertising_data$Daily.Internet.Usage, main = "Daily internet usage")
```



##Age, Daily time spent, and daily internet usage have no outliers while Area of income has a few outliers and this will not have any effect on our data.

Check for Duplicates

```
duplicates <- advertising_data[duplicated(advertising_data),]
```

```
#changing the timestamp datatype from factor to date_time
advertising_data$Timestamp <- as.Date(advertising_data$Timestamp, format = "%Y-%m-%s-%h-%m-
+ %s")
```

```
#checking the new datatype for the Timestamp column
sapply(advertising_data, class)
```

```
## Daily.Time.Spent.on.Site                   Age           Area.Income
##                "numeric"             "integer"             "numeric"
##     Daily.Internet.Usage          Ad.Topic.Line                  City
##                "numeric"           "character"           "character"
##                     Male               Country             Timestamp
##                "integer"           "character"                "Date"
##            Clicked.on.Ad
##                "integer"
```

#4.0 Data Analysis #4.1 Univariate analysis #4.1.1 Measures of central tendancy and dispersion

```
#### Daily time spent on site

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

```
#Getting mean, median and mode for Age
print(paste("The mean is ", mean(advertising_data$Age)))
```

```
## [1] "The mean is  36.009"
```

```
print(paste("The median is ", median(advertising_data$Age)))
```

```
## [1] "The median is  35"
```

```
print(paste("The mode is ",getmode(advertising_data$Age)))
```

```
## [1] "The mode is  31"
```

```
print(paste("The minimum is ",min(advertising_data$Age)))
```

```
## [1] "The minimum is  19"
```

```
print(paste("The maximum is ",max(advertising_data$Age)))
```

```
## [1] "The maximum is  61"
```

```
print(paste("The range is ",range(advertising_data$Age)))
```

```
## [1] "The range is  19" "The range is  61"
```

```
print(paste("The varianceis ",var(advertising_data$Age)))
```

```
## [1] "The varianceis  77.1861051051051"
```

```
print(paste("The standard deviation is ",sd(advertising_data$Age)))
```

```
## [1] "The standard deviation is  8.78556231012592"
```

```
#Getting mean, median,mode, min, max, range, variance and standard deviation for Daily time spent

print(paste("The mean is ", mean(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The mean is  65.0002"
```

```
print(paste("The median is ", median(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The median is  68.215"
```

```
print(paste("The mode is ",getmode(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The mode is  62.26"
```

```
print(paste("The minimum is ",min(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The minimum is  32.6"
```

```
print(paste("The maximum is ",max(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The maximum is  91.43"
```

```
print(paste("The range is ",range(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The range is  32.6"  "The range is  91.43"
```

```
print(paste("The varianceis ",var(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The varianceis  251.337094854855"
```

```
print(paste("The standard deviation is ",sd(advertising_data$Daily.Time.Spent.on.Site)))
```

```
## [1] "The standard deviation is  15.8536145675002"
```

```
#Getting mean, median and mode for Area income
print(paste("The mean is ", mean(advertising_data$Area.Income)))
```

```
## [1] "The mean is  55000.00008"
```

```
print(paste("The median is ", median(advertising_data$Area.Income)))
```

```
## [1] "The median is  57012.3"
```

```
print(paste("The mode is ",getmode(advertising_data$Area.Income)))
```

```
## [1] "The mode is  61833.9"
```

```
print(paste("The minimum is ",min(advertising_data$Area.Income)))
```

```
## [1] "The minimum is  13996.5"
```

```
print(paste("The maximum is ",max(advertising_data$Area.Income)))
```

```
## [1] "The maximum is  79484.8"
```

```
print(paste("The range is ",range(advertising_data$Area.Income)))
```

```
## [1] "The range is  13996.5" "The range is  79484.8"
```

```
print(paste("The varianceis ",var(advertising_data$Area.Income)))
```

```
## [1] "The varianceis  179952405.951775"
```

```
print(paste("The standard deviation is ",sd(advertising_data$Area.Income)))
```

```
## [1] "The standard deviation is  13414.6340222824"
```

```
#Getting mean, median and mode for Daily internet usage

print(paste("The mean is ", mean(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The mean is  180.0001"
```

```
print(paste("The median is ", median(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The median is  183.13"
```

```
print(paste("The mode is ",getmode(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The mode is  167.22"
```

```
print(paste("The minimum is ",min(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The minimum is  104.78"
```

```
print(paste("The maximum is ",max(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The maximum is  269.96"
```

```
print(paste("The range is ",range(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The range is  104.78" "The range is  269.96"
```

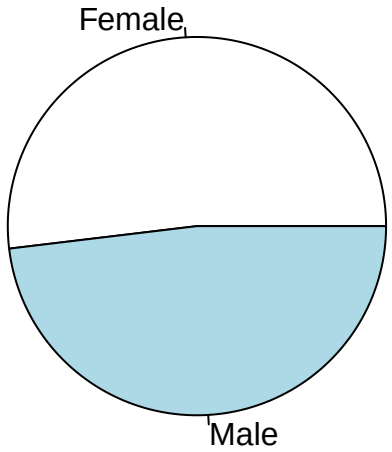
```
print(paste("The varianceis ",var(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The varianceis  1927.41539618619"
```

```
print(paste("The standard deviation is ",sd(advertising_data$Daily.Internet.Usage)))
```

```
## [1] "The standard deviation is  43.9023393019801"
```

```
# Plot the chart.
pie(table(advertising_data$Male), labels <-c("Female", "Male") )
```

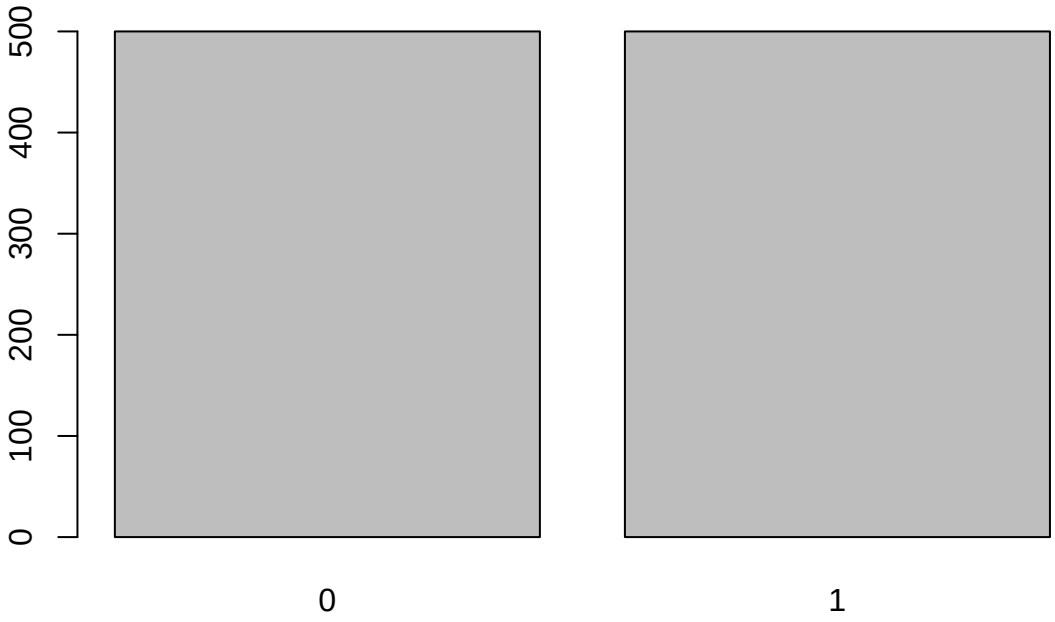


4.1.2 Gender distribution

- We can see that the number of females is slightly higher than that of males.

Clicks rate plot

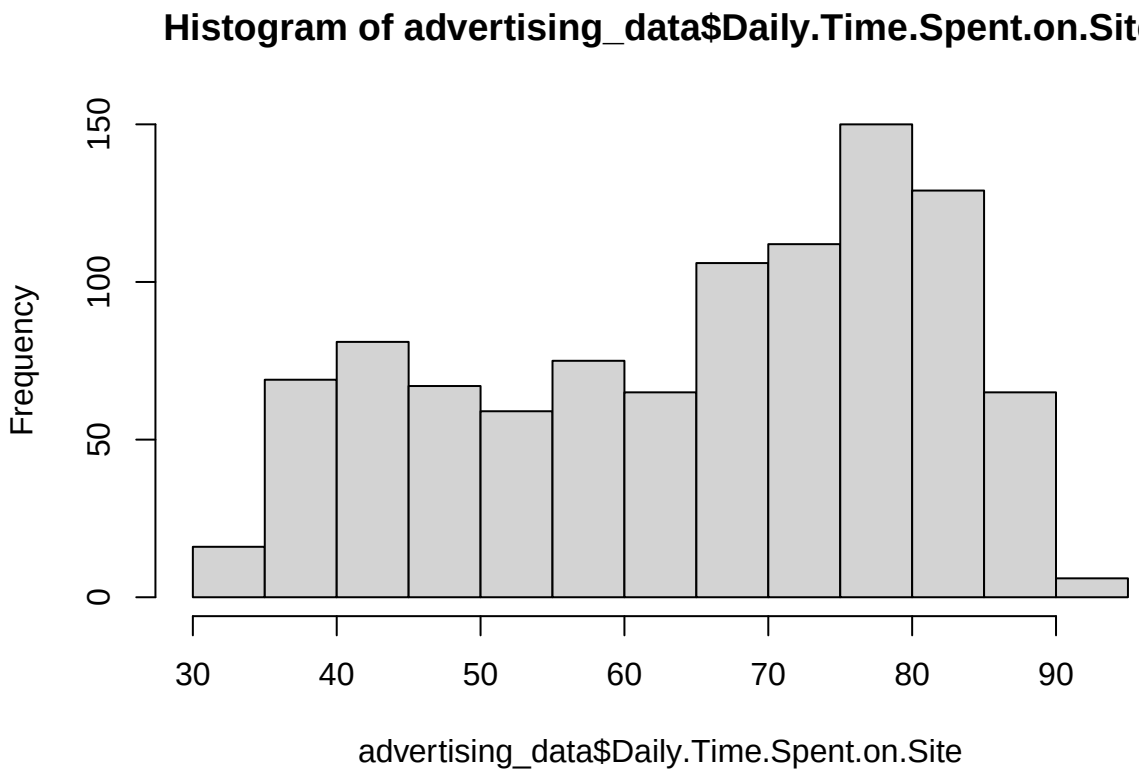
```
barplot(table(advertising_data$Clicked.on.Ad))
```



- The click rate is equal among males and females site visitors

-Time spent online

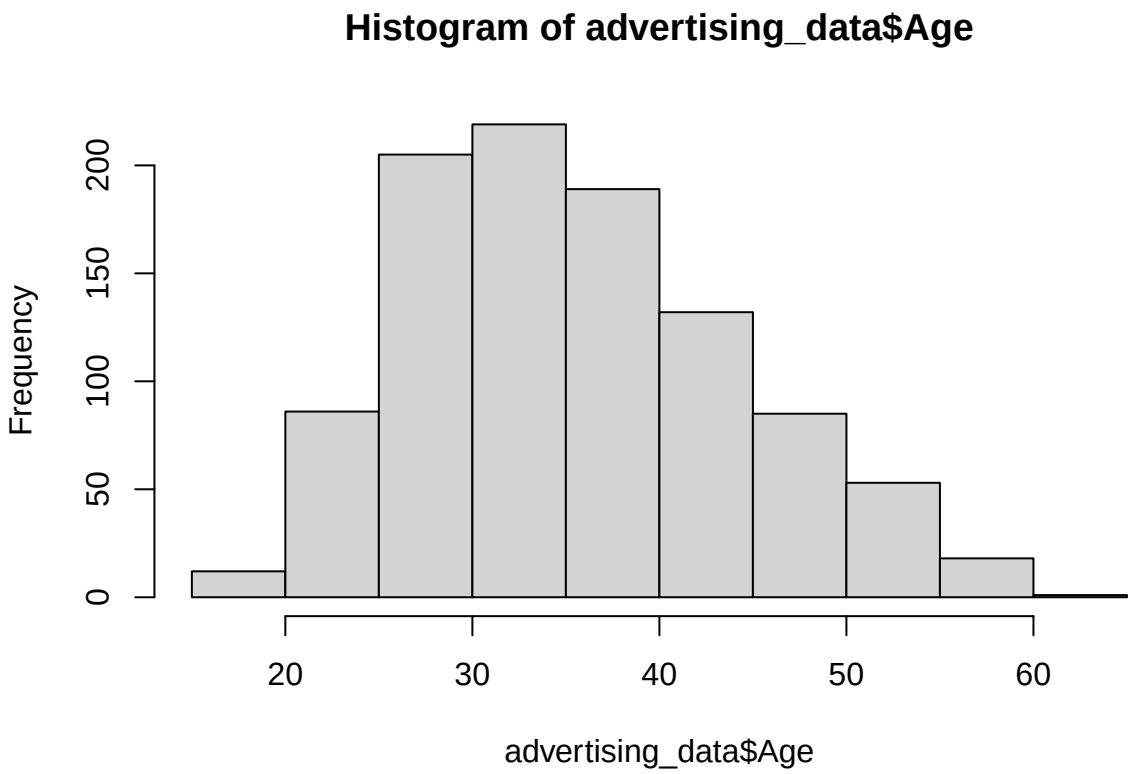
```
hist(advertising_data$Daily.Time.Spent.on.Site)
```



-The time spent online on site was between 65 anad 85 miniutes

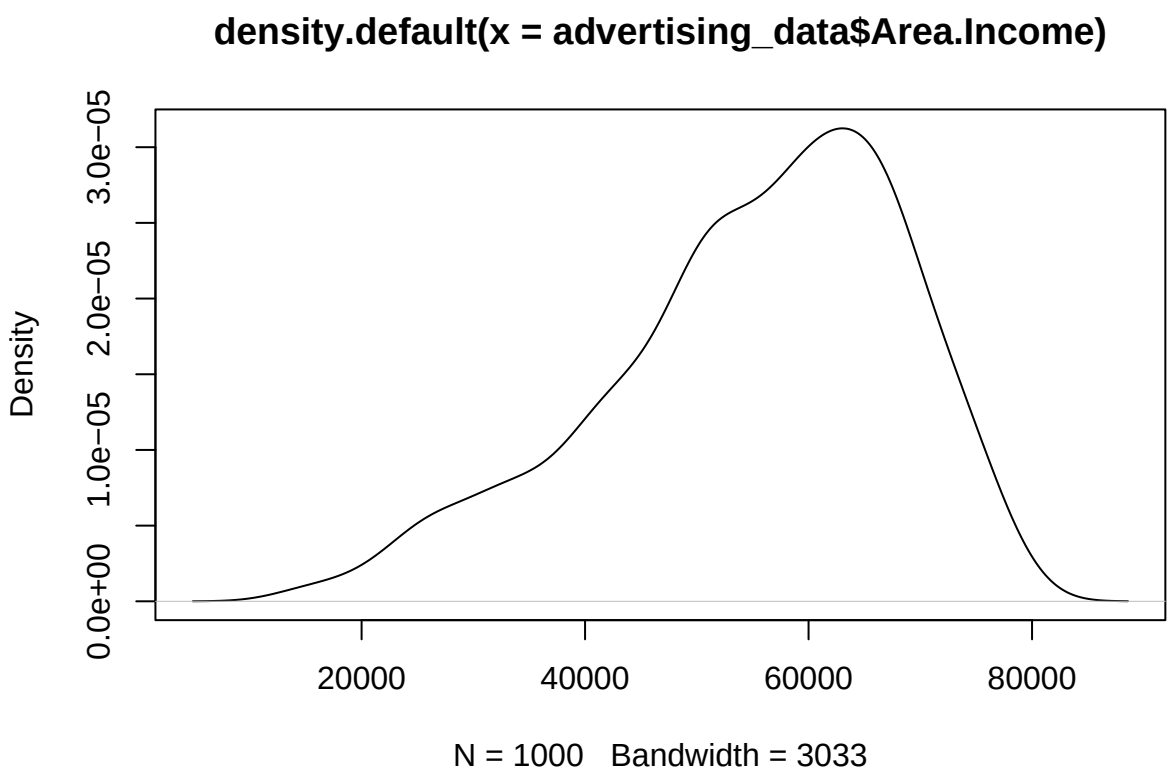
-Age Distribution

```
hist(advertising_data$Age)
```



-The most popular age group is between 25 to 40 years

```
# display density plot
print(plot(density(advertising_data$Area.Income)))
```




```
## NULL
```

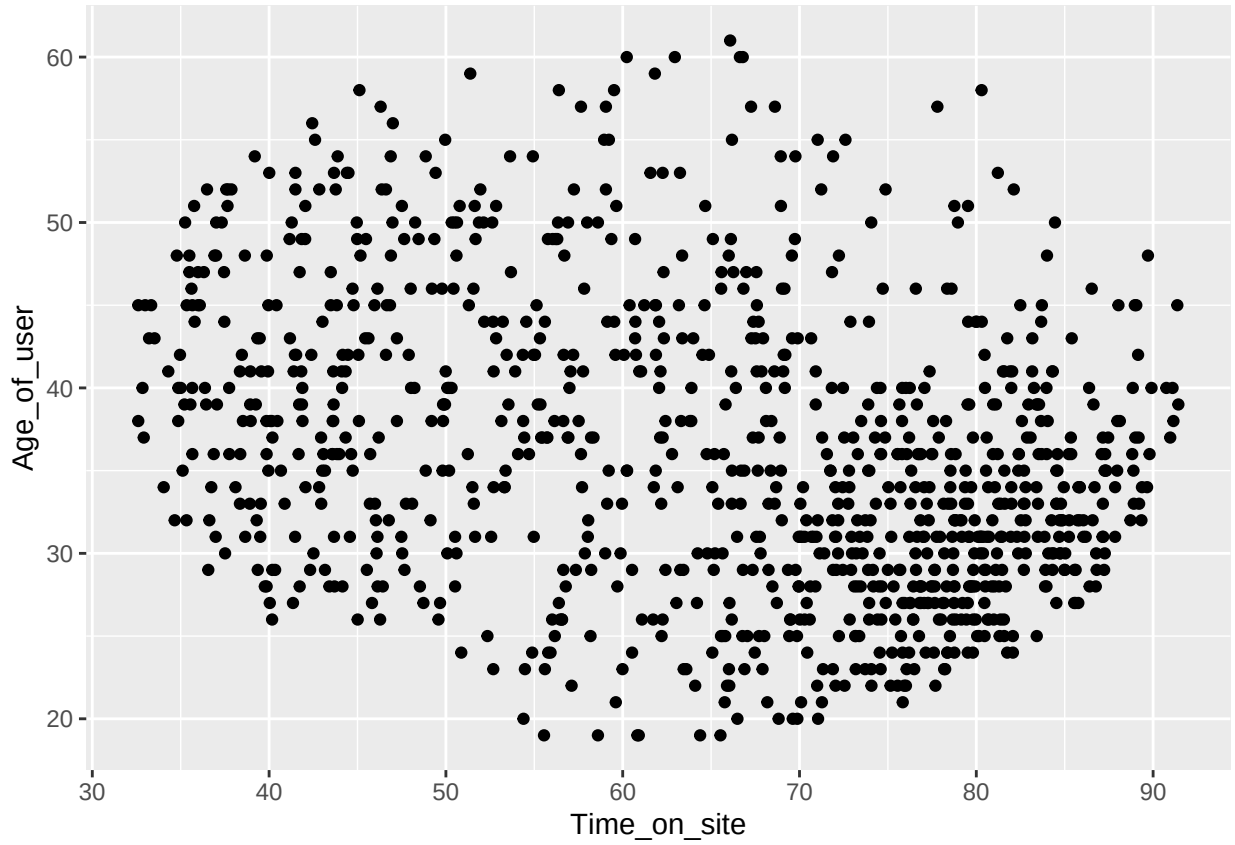
_ The area of income is skewwed to the right showing that internet usage is higher in areas of high income.

#4.2 Bivariate Analysis -We will do away with the data that we don't require in our further analysis

```
advertising_data2 <- subset(advertising_data, select = c(Daily.Time.Spent.on.Site,
Age,Area.Income,Daily.Internet.Usage,Male,Clicked.on.Ad ))
head(advertising_data2)
```

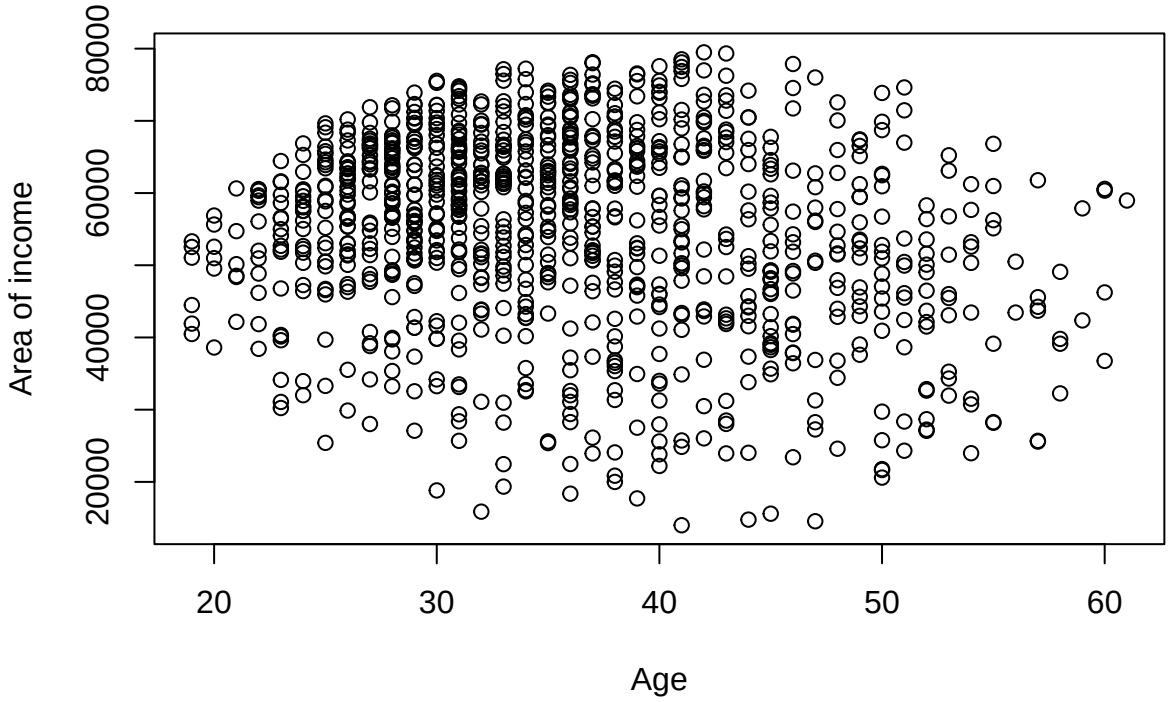
```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage Male
## 1                    68.95  35    61833.90               256.09    0
## 2                    80.23  31    68441.85               193.77    1
## 3                    69.47  26    59785.94               236.50    0
## 4                    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
## 4.2.1 Age vs time on site
#Time spent on the site vs age of the user
# Libraries
library(ggplot2)
# create data
Time_on_site <- advertising_data2$Daily.Time.Spent.on.Site
Age_of_user <- advertising_data2$Age
data <- data.frame(Time_on_site,Age_of_user)
# Plot
ggplot(data, aes(x=Time_on_site, y=Age_of_user)) + geom_point()
```



-People aged between 25 and 40 are the one spending most time on site.

```r
###4.2.2 Ages vs area of income
plot(advertising_data2$Age, advertising_data2$Area.Income, xlab="Age", ylab="Area of income")
```



-60,000 is the most popular area of income

```
###4.3 Covariance
#Covariance between Daily time spent and area of income
cov(advertising_data2$Daily.Time.Spent.on.Site,advertising_data2$Area.Income)
```

```
## [1] 66130.81
```

-Positive covariance. Meaning they show linear similarity

```
#Covariance between Daily time spent and Daily internet usage
cov(advertising_data2$Daily.Time.Spent.on.Site,advertising_data2$Daily.Internet.Usage)
```

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

-Positive covariance. Meaning they show linear similarity

```
#Covariance between Daily time spent and age
cov(advertising_data2$Daily.Time.Spent.on.Site,advertising_data2$Age)
```

```
## [1] -46.17415
```

-Negative covariance meaning Age and daily time spent do not have similarity

#Correlation

```r
#We're going to use Pearson, but we can also compute Spearman or Kendall coefficients.
mydata = cor(advertising_data2, method = c("spearman"))
mydata1= cor(advertising_data2, method = c("kendall"))
mydata2= cor(advertising_data2, method = c("pearson"))
mydata #spearman
```

```
##                           Daily.Time.Spent.on.Site         Age Area.Income
## Daily.Time.Spent.on.Site               1.00000000 -0.31686155  0.28313439
## Age                                   -0.31686155  1.00000000 -0.13595396
## Area.Income                            0.28313439 -0.13595396  1.00000000
## Daily.Internet.Usage                   0.51410805 -0.37086395  0.33916021
## Male                                  -0.01592213 -0.02315468 -0.01436909
## Clicked.on.Ad                         -0.74487253  0.48633733 -0.46722440
##                           Daily.Internet.Usage        Male Clicked.on.Ad
## Daily.Time.Spent.on.Site           0.51410805 -0.01592213   -0.74487253
## Age                               -0.37086395 -0.02315468    0.48633733
## Area.Income                        0.33916021 -0.01436909   -0.46722440
## Daily.Internet.Usage               1.00000000  0.02820432   -0.77660702
## Male                               0.02820432  1.00000000   -0.03802747
## Clicked.on.Ad                     -0.77660702 -0.03802747    1.00000000
```

```r
library("Hmisc")
```

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

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

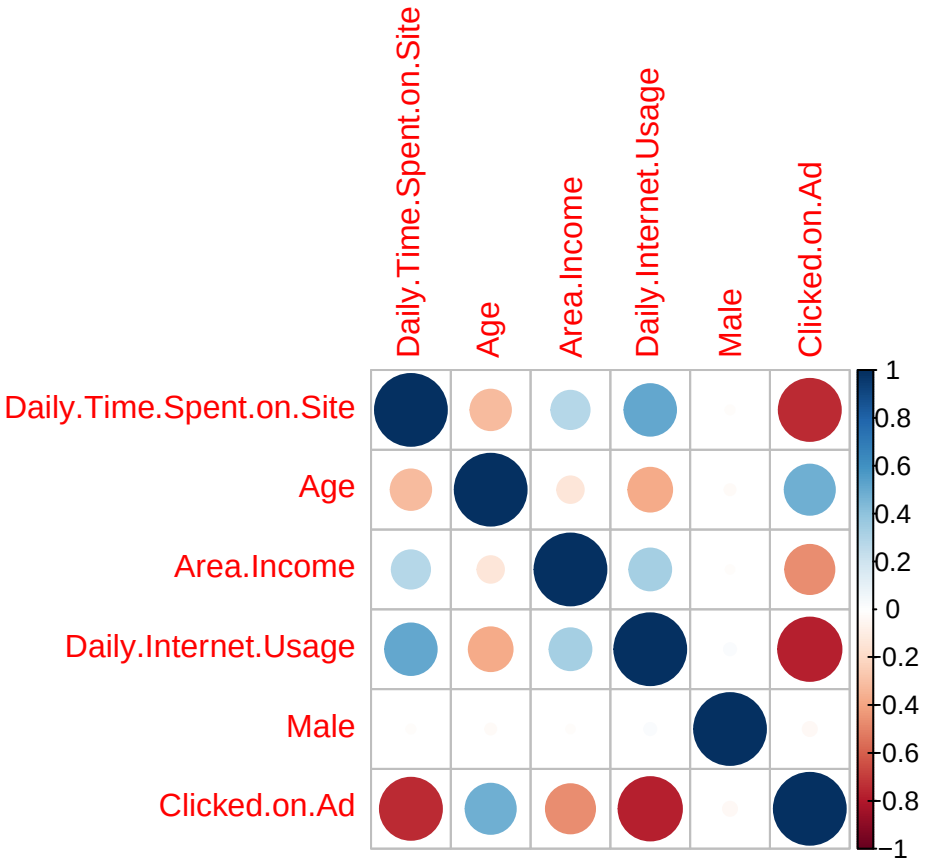
```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(corrplot)
```

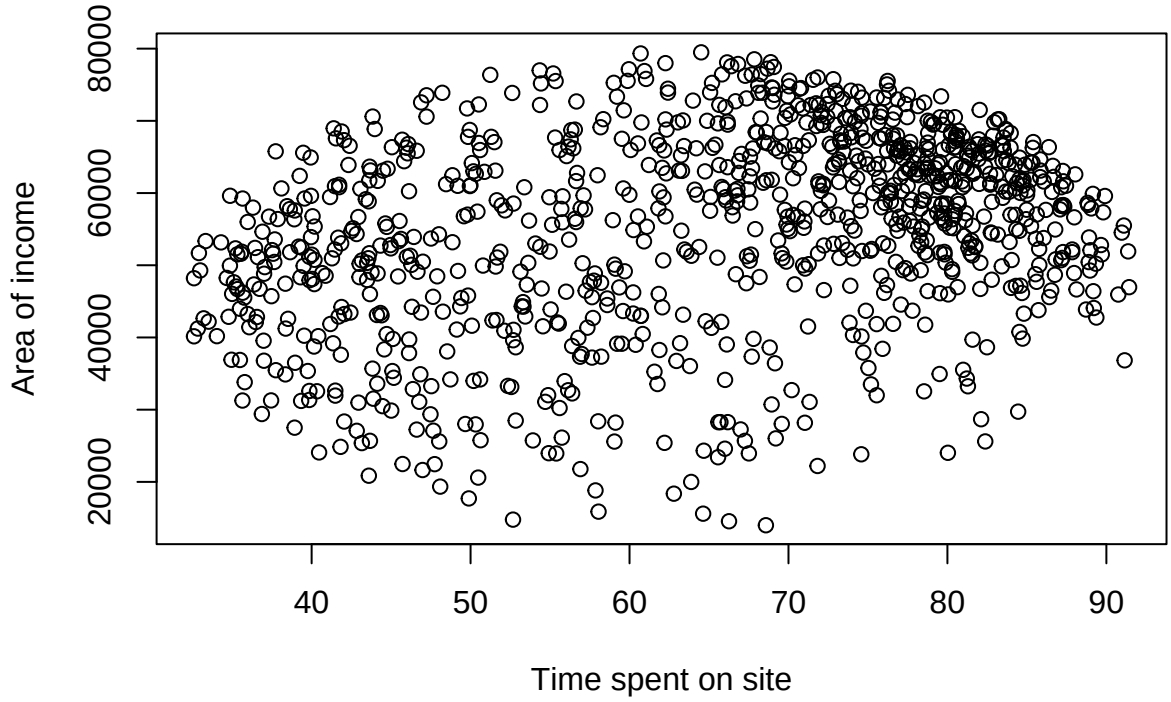
```
## corrplot 0.92 loaded
```

```r
corrplot(mydata)
```



__ There's minimal positive correlation between our dataset variables.

```
plot(advertising_data2$Daily.Time.Spent.on.Site, advertising_data2$Area.Income, xlab="Time spent on sit
```



Age vs clicks
- People who spent more time on site hail from the higher income areas

5.0 Conclusions

```
#-Areas of high income are likely to produce more clicks
#-Female gender is more likely to click on ads
#-Time spent on site will highly influence users to click on ad
```