Advertisement Data Analysis in R

1. Defining the Research Question

(a) Specifying the Question.

An entrprenuer would like to employ your services as a Data Science Consultant to help them identify which individuals are most likely to click on their ads.

(b) Metrics of Success.

The study will perform conclusive Exploratory Data Analysis to enable us identify which individuals who are most likely to click on ads.

(c) Understanding the context.

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. Using the data previously collected, she is looking to do a study to identify which individuals are most likely to click on her ads.

(d) Experimental Design.

1. Define the question, the metric for success, the context, experimental design taken and the appropriateness of the available data to answer the given question

2. Read the dataset into our environment (RStudio)

3. Preview the dataset

4.Find and deal with outliers, anomalies, and missing data within the dataset

5. Perform univariate and bivariate analysis

6.Implement our solution by creating various supervised learning models and choose the best performing one for our research problem

7.From our insights provide conclusions and recommendations

(e) Data Relevance.

Data is provided was collected in the past but from the same blog hence it is very suitable for this study.

#Definition of Variables Daily Time Spent on Site

Age

Area

Income

Daily Internet Usage

Ad Topic Line

City

Male

Country

Timestamp

```r
# Importing Relevant Libraries
library(tidyverse)
```

Clicked on Ad

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

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```r
library(data.table)
```

```
##
## Attaching package: ’data.table’

## The following objects are masked from ’package:dplyr’:
##
##     between, first, last

## The following object is masked from ’package:purrr’:
##
##     transpose
```

2. Loading of the Dataset

```
advert<- fread("http://bit.ly/IPAdvertisingData")
```

3.Previewing the dataset

```
# Previewing the first six entries
head(advert)
```

```
##    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
```

```
# Preview the last five entries
tail(advert)
```

```
##    Daily Time Spent on Site Age Area Income Daily Internet Usage
## 1:                    43.70  28    63126.96               173.01
## 2:                    72.97  30    71384.57               208.58
## 3:                    51.30  45    67782.17               134.42
## 4:                    51.63  51    42415.72               120.37
## 5:                    55.55  19    41920.79               187.95
## 6:                    45.01  26    29875.80               178.35
##                            Ad Topic Line          City Male
## 1:        Front-line bifurcated ability  Nicholasland    0
## 2:        Fundamental modular algorithm     Duffystad    1
## 3:      Grass-roots cohesive monitoring   New Darlene    1
## 4:         Expanded intangible solution South Jessica    1
## 5: Proactive bandwidth-monitored policy   West Steven    0
## 6:      Virtual 5thgeneration emulation   Ronniemouth    0
##                    Country           Timestamp Clicked on Ad
## 1:                Mayotte 2016-04-04 03:57:48             1
## 2:                Lebanon 2016-02-11 21:49:00             1
## 3: Bosnia and Herzegovina 2016-04-22 02:07:01             1
## 4:               Mongolia 2016-02-01 17:24:57             1
## 5:              Guatemala 2016-03-24 02:35:54             0
## 6:                 Brazil 2016-06-03 21:43:21             1
```

```
# Lets check the shape of the data
dim(advert)
```

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

```
# The data contained 10 columns and 1000 rows
```

```
# Lets check for the data types
sapply(advert, class)
```

```
## $‘Daily Time Spent on Site‘
## [1] "numeric"
##
## $Age
## [1] "integer"
##
## $‘Area Income‘
## [1] "numeric"
##
## $‘Daily Internet Usage‘
## [1] "numeric"
##
## $‘Ad Topic Line‘
## [1] "character"
##
## $City
## [1] "character"
##
## $Male
## [1] "integer"
##
## $Country
## [1] "character"
##
## $Timestamp
## [1] "POSIXct" "POSIXt"
##
## $‘Clicked on Ad‘
## [1] "integer"
```

```
# The data contained variables in numeric, characters and integers
```

```
# Lets check for the unique values
sapply(advert, function(x) length(unique(x)))
```

```
## Daily Time Spent on Site                      Age              Area Income
##                        900                       43                     1000
##     Daily Internet Usage            Ad Topic Line                     City
##                        966                     1000                      969
##                       Male                  Country                Timestamp
##                          2                      237                     1000
##             Clicked on Ad
##                          2
```

```
# Lets check for the summary statistics

summary(advert)
```

```
##  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
##  Min.   :2016-01-01 02:52:10.00   Min.   :0.0
##  1st Qu.:2016-02-18 02:55:42.00   1st Qu.:0.0
##  Median :2016-04-07 17:27:29.50   Median :0.5
##  Mean   :2016-04-10 10:34:06.64   Mean   :0.5
##  3rd Qu.:2016-05-31 03:18:14.00   3rd Qu.:1.0
##  Max.   :2016-07-24 00:22:16.00   Max.   :1.0
```

4. Data Cleaning

Checking for missing values

```
# Checking for missing values
colSums(is.na(advert))
```

```
## 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 data did not contain missing values
```

Checking for duplicates

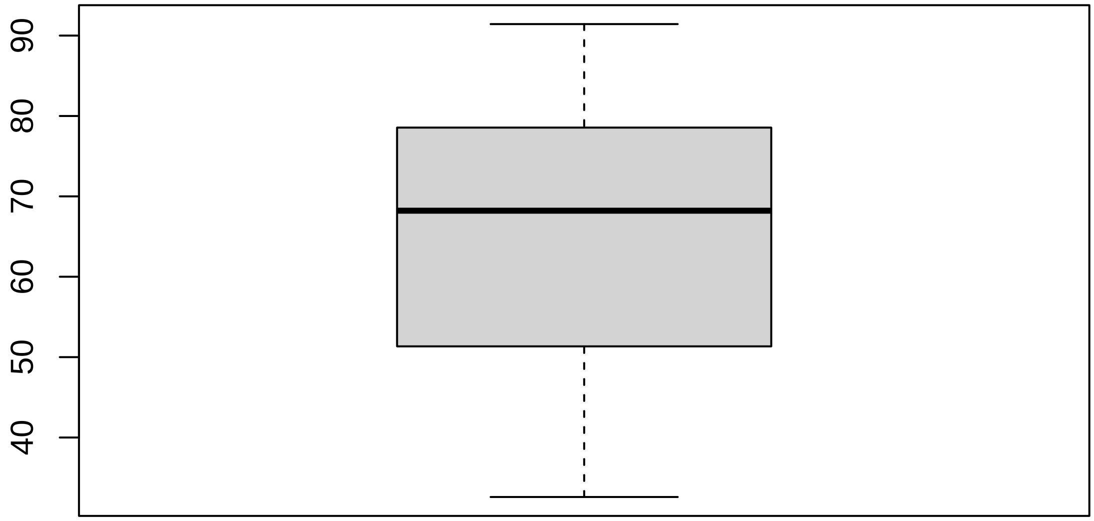
```
# Checking for duplicates
advert[duplicated(advert)]
```

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

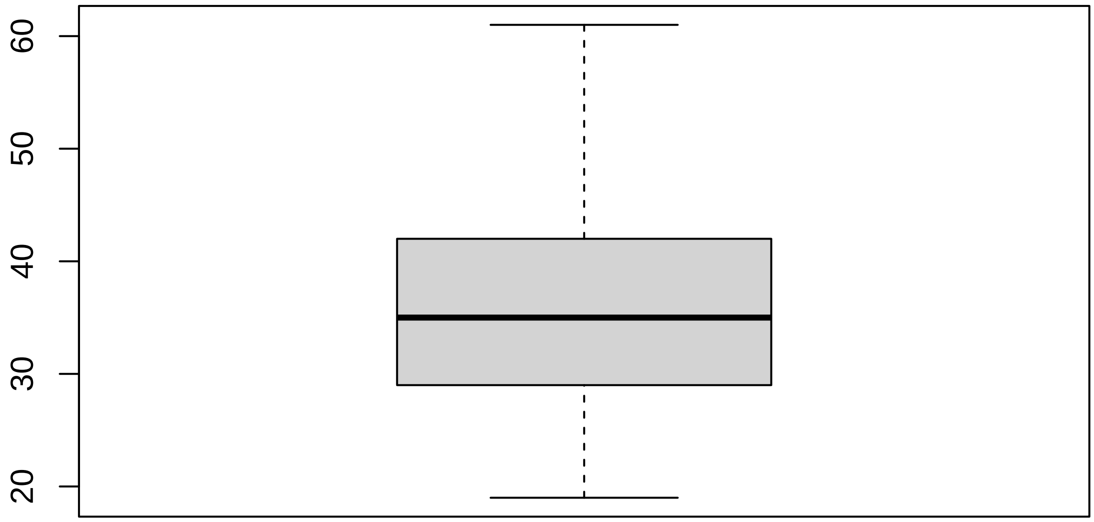
```
# The data did not have duplicates
```

Checking for the outliers

```
# Lets check for outliers using boxplots
# 1. Daily time spent on Site column

boxplot(advert$"Daily Time Spent on Site")
```

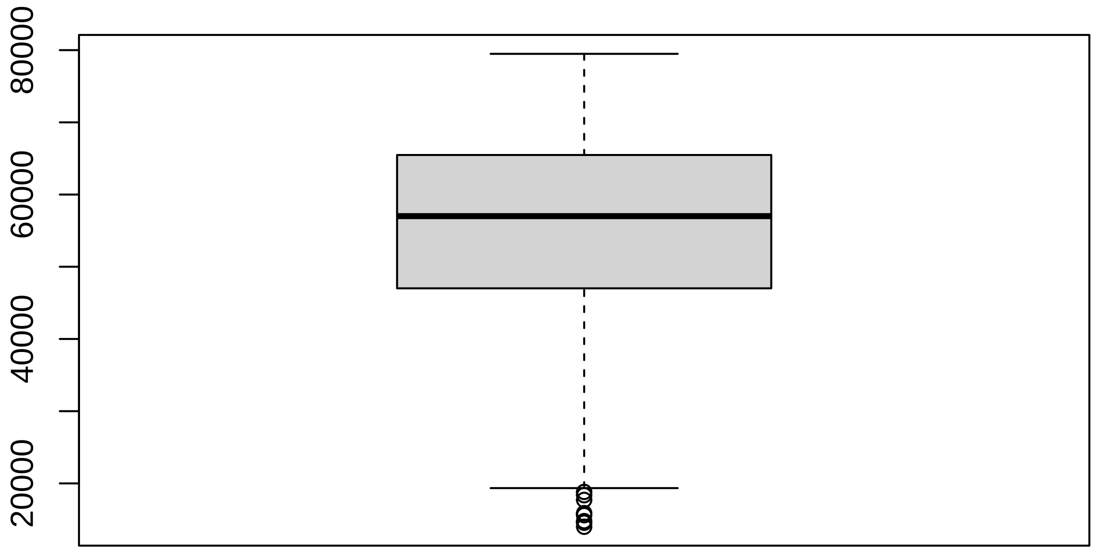


```
# Outliers for the age column
boxplot(advert$"Age")
```



```
# The age column did not contain outliers
```

```
#Outliers for the Area Income
boxplot(advert$"Area Income")
```

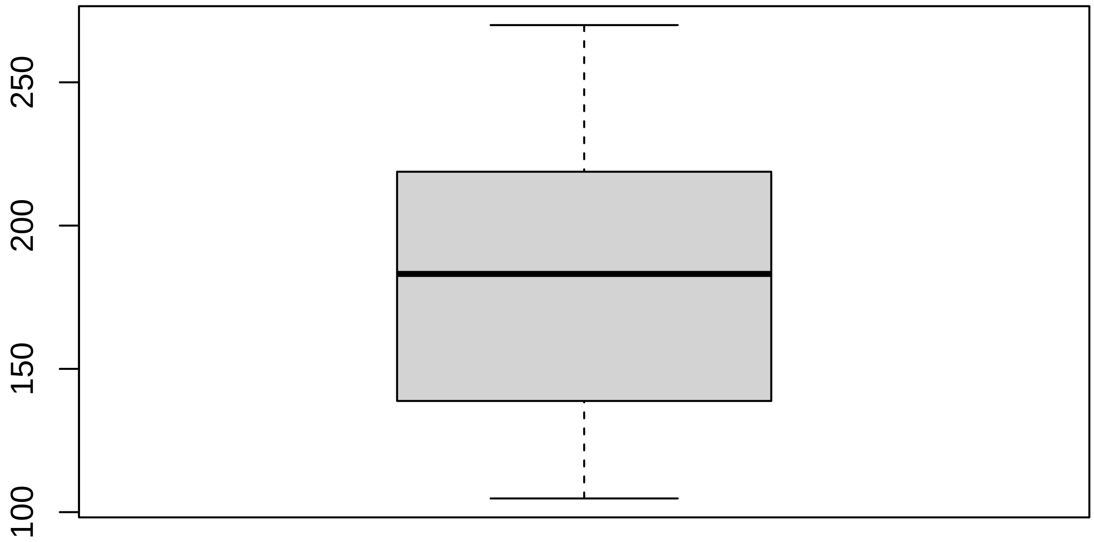


```
# Lets preview the outliers

# income_outliers<- advert %>% dplyr::filter('Area Income') < 20000)
income_outliers<-advert%>% filter('Area.Income' >20000)
income_outliers
```

```
##       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
##   ---
##  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
##    1:    Cloned 5thgeneration orchestration    Wrightburgh    0
##    2:    Monitored national standardization      West Jodi    1
##    3:      Organic bottom-line service-desk       Davidton    0
##    4: Triple-buffered reciprocal time-frame West Terrifurt    1
##    5:         Robust logistical utilization   South Manuel    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
##    1:                 Tunisia 2016-03-27 00:53:11             0
##    2:                   Nauru 2016-04-04 01:39:02             0
##    3:              San Marino 2016-03-13 20:35:42             0
##    4:                   Italy 2016-01-10 02:31:19             0
##    5:                 Iceland 2016-06-03 03:36:18             0
##   ---
##  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
```

```
# We won't remove the outliers since they are true fugures.
```

```
# Outliers in the Daily Internet users
boxplot(advert$`Daily Internet Usage`)
```



### Lets select exclude the clicked.on.ad & male column since they are binary column

5. Exploratory Data Analysis

(a) Univariate Analysis

1. Measures of Central Tendancy

```
# Lets check for the summary statistics of all the numeric values
numeric_ <- advert %>% select('Daily Time Spent on Site','Age','Area Income','Daily Internet Usage')

summary(numeric_)
```

```
##  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
```

1. The minimum amount of time spent on the blog is 32.60 and maximum is 91.43 with a mean at 65 and median at 68.

2. The mean age of people visiting the site is 36, max age is 61 and min age is 19.

3.The maximum income of individuals is 79485 and a min income of 13996

4.The mean daily internet usage on the website is 180 and a median is 183.1

2. Measures of Dispersion

(a) Variance

```
# Lets find the variance of the 'Daily Time Spent on Site' column

var(advert$"Daily Time Spent on Site")
```

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

```


# # Lets find the variance of the 'Age' column

var(advert$Age)
```

The variance for the daily time spent on site from the mean is 251.33.

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

```r
# # Lets find the variance of the 'Daily Internet Usage' column

var(advert$"Daily Internet Usage")
```

The variance for the age column from the mean is 77.18.

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

The variance for the daily internet usage column is 1927.415.

(b) Standard Deviation

```r
# Lets find the S.D for the 'Daily Time Spent on Site' column

sd(advert$"Daily Time Spent on Site")
```

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

```r
# Lets find the S.D for the 'Age' column

sd(advert$Age)
```

The standard deviation for the daily time spent on site is 15.85361.

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

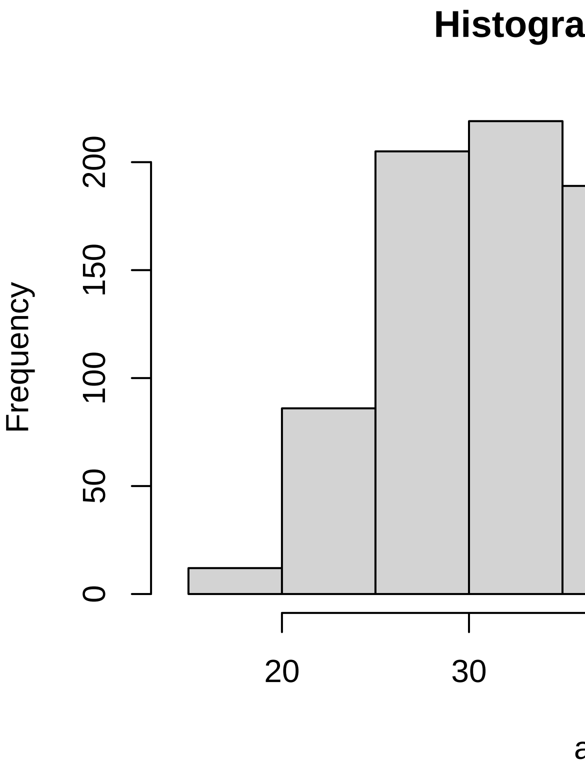
```r
#  Lets find the S.D for the 'Daily Internet Usage' column

sd(advert$"Daily Internet Usage")
```

The standard deviation for the age column is 8.785562.

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

```r
# Importing a library
library(moments)
```

```r
#Plotting a histogram for the column age
hist(advert$Age)
```



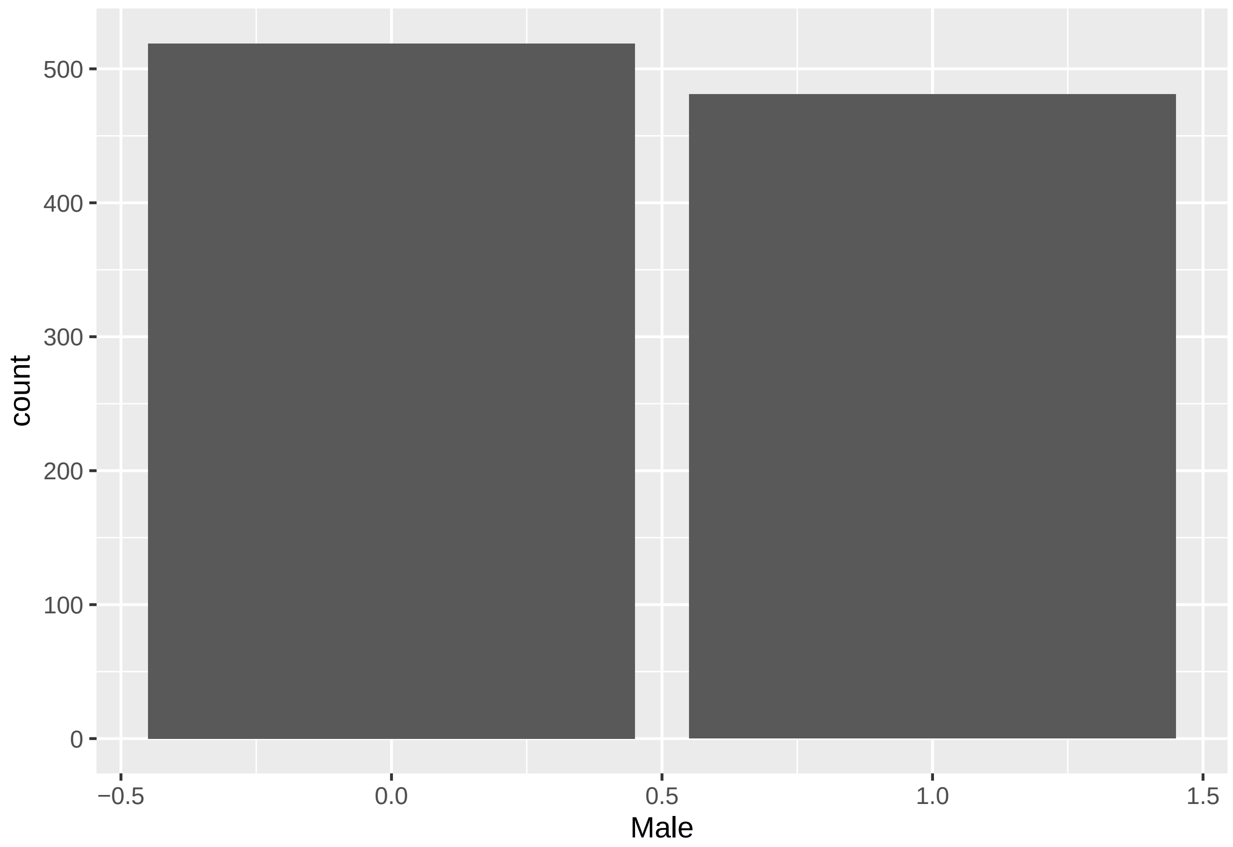
The standard deviation for the daily internet usage is 43.90234.

```r
#Plotting Frequency tables to show the distribution of the data

# 1. Categorical Data
gender <- table(advert$Male)
gender
```

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

There were 519 females and 481 men.

```r
# Plotting a frequency table
ggplot(data = advert) +
  geom_bar(mapping = aes(x = Male))
```



# Bivariate Analysis

```
age <- advert$Age
units <- advert$"Daily Internet Usage"
time <- advert$"Daily Time Spent on Site"
```

```
# Lets find the correlation between age and the Daily Time Spent on Site

cor(age, time)
```

```
## [1] -0.3315133
```

```


# Correlation between age and Daily Internet Usage
cor(age, units)
```

There is a weak negative correlation between age and the daily time spent on the site.

```
## [1] -0.3672086
```

```


# Correlation between daily time spent and the daily internet usage
cor(time, units)
```

There is a weak negative correlation between age and the daily internet usage.

```
## [1] 0.5186585
```

There is a strong positive correlation between the time spent on site and the daily internet usage. This is because the more time you spend on site, the higher the amount of internet usage.

Covariance

```
#Lets find the covariance between age and the daily time spent on the site
cov(age, time)
```

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

```


#Lets find the covariance between age and the daily internet usage
cov(age, units)
```

There is a negative covariance between age and the daily time spent on the site. This means that as a person ages, the less time they spend on the site daily.

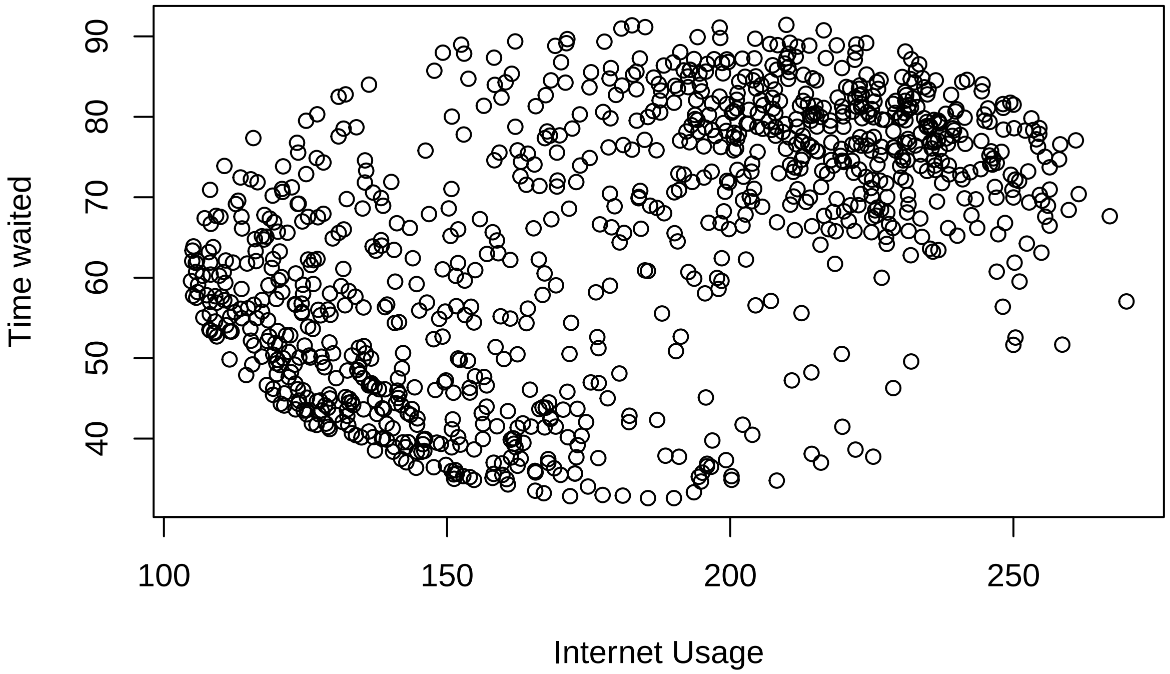
```
## [1] -141.6348
```

```


#Lets find the covariance between daily time spent and the daily internet usage.

cov(time, units)
```

There is a negative covariance between age and the daily internet usage. This means that as a person ages, the less internet usage they spend.

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

```


internet<-advert$`Daily Internet Usage`
time <- advert$`Daily Time Spent on Site`
plot(internet, time, xlab="Internet Usage", ylab="Time waited")
```



# 6. Conclusion

From the univariate data analysis, we can conclude that:

There were more females than males in our dataset.

The dataset was balanced in the sense that 500 individuals clicked on the ads while 500 individuals did not click on the ads.

Individuals who are between 28 and 36 years old were the most in our dataset.

From the bivariate data analysis, we can conclude that:

There is a negative covariance and correlation between age and daily time spent on the site which means that the older an individual is, the less time they spend on the site.

There is also a negative covariance and correlation between age and the daily internet usage which means that the younger an individual is, the higher the internet usage is as compared to an older individual.

On the other hand, there is a positive covariance and correlation between the daily internet usage and the daily time spent on the internet.

7. Recommendation

The study recommends that the entrprenuer creates an ad that targets individuals aged between 25 and 35 years old seeing as they are the most in our dataset.