R Notebook

1. DEFINING THE QUESTION

a) Specifying 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.

b) Defining the Metrics of Success

Performing the Exploratory Data Analysis.

c) Understanding the context

Determining the audience the entrepreneur can target.

d) Recording the Experimental Design

1. Defining the question, the metric for success, the context and experimental design.
2. Loading and exploring the dataset.
3. Finding and dealing with outliers, anomalies, and missing data within the dataset.
4. Perform univariate and bivariate analysis.
5. Giving a conclusion and recommendation.

e) Relevance of the data

The data used in this project is for determining which audience should be targeted by the entrepreneur. The dataset link: ('http://bit.ly/IPAdvertisingData')

2. DATA ANALYSIS

a) Checking the Data

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

```
library(ggplot2)
```

```
library(magrittr)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Reading the data
df <- fread('http://bit.ly/IPAdvertisingData')
df
```

```
##       Daily Time Spent on Site   Age Area Income Daily Internet Usage
##                          <num> <int>       <num>                <num>
##    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
##                                      <char>         <char> <int>
##    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
##                        <char>              <POSc>         <int>
##    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
```

```
# Viewing the dataset
View(df)
```

```
# Viewing the column names
colnames(df)
```

```
##  [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"
```

```
# Previewing the dataset
class(df)
```

```
## [1] "data.table" "data.frame"
```

```
# Previewing the top of the dataset
head(df)
```

```
##    Daily Time Spent on Site   Age Area Income Daily Internet Usage
##                       <num> <int>       <num>                <num>
## 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
##                                   <char>         <char> <int>     <char>
## 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
##                 <POSc>         <int>
```

```
## 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
```

```
# Previewing the bottom of the dataset
tail(df)
```

```
##    Daily Time Spent on Site   Age Area Income Daily Internet Usage
##                       <num> <int>       <num>                <num>
## 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
##                                 <char>        <char> <int>
## 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
##                     <char>              <POSc>         <int>
## 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
```

```
# Checking the shape of the dataset
dim(df)
```

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

1000 rows and 10 columns

b) Data Cleaning

Missing Values

```
# Checking for missing values
sum(is.na(df))
```

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

There are no missing values.

```
# Removing all rows with na
na.omit(df)
```

```
##       Daily Time Spent on Site   Age Area Income Daily Internet Usage
##                          <num> <int>       <num>                <num>
##    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
##                                      <char>         <char> <int>
##    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
##                        <char>               <POSc>         <int>
##    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
```

Duplicates

```
# Checking for duplicates
duplicated_rows <- df[duplicated(df),]
duplicated_rows
```

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

There are no duplicates

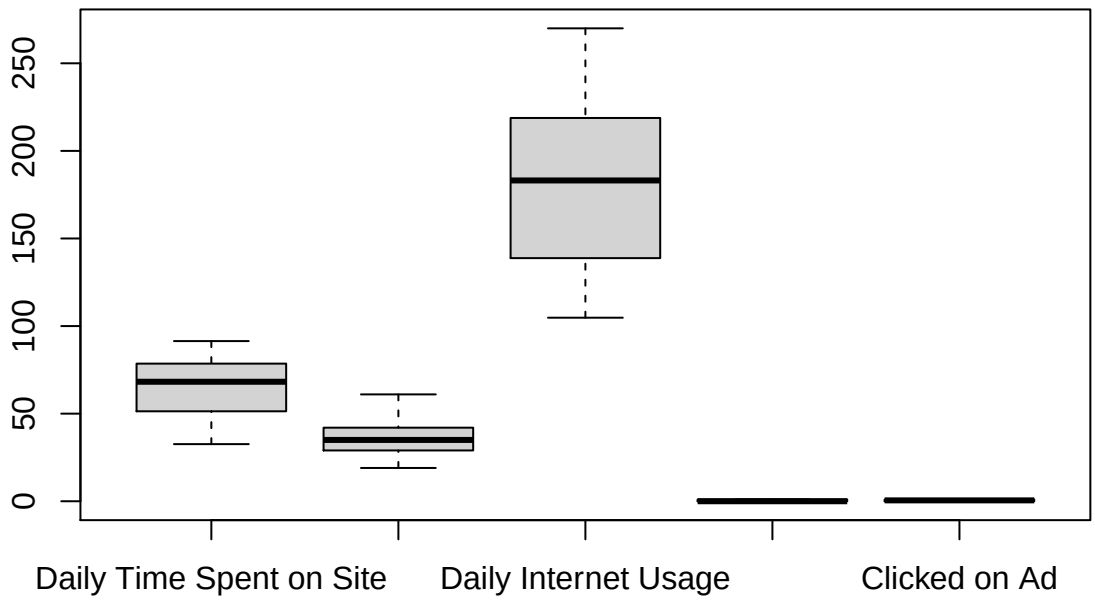
```
# Displaying the unique items and assigning unique_items variable
unique_items <- df[!duplicated(df), ]
unique_items
```

```
##       Daily Time Spent on Site   Age Area Income Daily Internet Usage
##                          <num> <int>       <num>                <num>
##    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
##                                     <char>         <char> <int>
##    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
##                        <char>              <POSc>         <int>
##    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
```

Outliers

```
# Visualizing outliers using boxplot
df1 <- subset(df, select = c("Daily Time Spent on Site", "Age", "Daily Internet Usage", "Male", "Clicke

boxplot(df1)
```




```
# Renaming columns
df1 <- df1 %>% rename(Daily_Time_Spent_on_Site = "Daily Time Spent on Site")
df1 <- df1 %>% rename(Daily_Internet_Usage = "Daily Internet Usage")
df1 <- df1 %>% rename(Clicked_on_Ad = "Clicked on Ad")

df1
```

```
##       Daily_Time_Spent_on_Site   Age Daily_Internet_Usage  Male Clicked_on_Ad
##                          <num> <int>                <num> <int>         <int>
##    1:                    68.95    35               256.09     0             0
##    2:                    80.23    31               193.77     1             0
##    3:                    69.47    26               236.50     0             0
##    4:                    74.15    29               245.89     1             0
##    5:                    68.37    35               225.58     0             0
##   ---
##  996:                    72.97    30               208.58     1             1
##  997:                    51.30    45               134.42     1             1
##  998:                    51.63    51               120.37     1             1
##  999:                    55.55    19               187.95     0             0
## 1000:                    45.01    26               178.35     0             1
```

3. BIVARIATE AND UNIVARIATE ANALYSIS

a) Univariate Analysis

Measures of Central Tendency

```
# Summary statistics of the dataset
summary(df1)
```

```
##  Daily_Time_Spent_on_Site      Age        Daily_Internet_Usage      Male
##  Min.   :32.60            Min.   :19.00   Min.   :104.8        Min.   :0.000
##  1st Qu.:51.36            1st Qu.:29.00   1st Qu.:138.8        1st Qu.:0.000
##  Median :68.22            Median :35.00   Median :183.1        Median :0.000
##  Mean   :65.00            Mean   :36.01   Mean   :180.0        Mean   :0.481
##  3rd Qu.:78.55            3rd Qu.:42.00   3rd Qu.:218.8        3rd Qu.:1.000
##  Max.   :91.43            Max.   :61.00   Max.   :270.0        Max.   :1.000
##  Clicked_on_Ad
##  Min.   :0.0
##  1st Qu.:0.0
##  Median :0.5
##  Mean   :0.5
##  3rd Qu.:1.0
##  Max.   :1.0
```

```
# Median of age
df1.Age.median <- median(df$Age)
df1.Age.median
```

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

```
# Mean of age
df1.Age.mean <- mean(df$Age)
df1.Age.mean
```

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

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

df1.Age.mode <- getmode(df$Age)
df1.Age.mode
```

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

Measures of Dispersion

```r
# Displaying the column names
colnames(df1)
```

```
## [1] "Daily_Time_Spent_on_Site" "Age"
## [3] "Daily_Internet_Usage"     "Male"
## [5] "Clicked_on_Ad"
```

```r
# Minimum code of Daily Time Spent on Site
df1.Daily_Time_Spent_on_Site.min <- min(df1$Daily_Time_Spent_on_Site)
df1.Daily_Time_Spent_on_Site.min
```

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

```r
# Minimum code of Daily Internet Usage
df1.Daily_Internet_Usage.min <- min(df1$Daily_Internet_Usage)
df1.Daily_Internet_Usage.min
```

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

```r
# Minimum code of Age
df1.Age.min <- min(df1$Age)
df1.Age.min
```

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

```r
# Maximum code of age
df1.Age.max <- max(df1$Age)
df1.Age.max
```

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

```r
# Maximum code of Daily Internet Usage
df1.Daily_Internet_Usage.max <- max(df1$Daily_Internet_Usage)
df1.Daily_Internet_Usage.max
```

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

```r
# Maximum code of Daily Time Spent on Site
df1.Daily_Time_Spent_on_Site.max <- max(df1$Daily_Time_Spent_on_Site)
df1.Daily_Time_Spent_on_Site.max
```

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

```r
# Range code of age
df1.Age.range <- range(df1$Age)
df1.Age.range
```

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

```r
# Range code of Daily Time Spent on Site
df1.Daily_Time_Spent_on_Site.range <- range(df1$Daily_Time_Spent_on_Site)
df1.Daily_Time_Spent_on_Site.range
```

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

```r
# Quantile code of Age
df1.Age.quantile <- quantile(df1$Age)
df1.Age.quantile
```

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

```r
# Quantile code of Daily Time Spent on Site
df1.Daily_Time_Spent_on_Site.quantile <- quantile(df1$Daily_Time_Spent_on_Site)
df1.Daily_Time_Spent_on_Site.quantile
```

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

```r
# Variance code of Age
df1.Age.variance <- var(df1$Age)
df1.Age.variance
```

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

```r
# Variance code of Daily Time Spent on Site
df1.Daily_Time_Spent_on_Site.variance <- var(df1$Daily_Time_Spent_on_Site)
df1.Daily_Time_Spent_on_Site.variance
```

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

```r
# Standard deviation code of age
df1.Age.sd <- sd(df1$Age)
df1.Age.sd
```

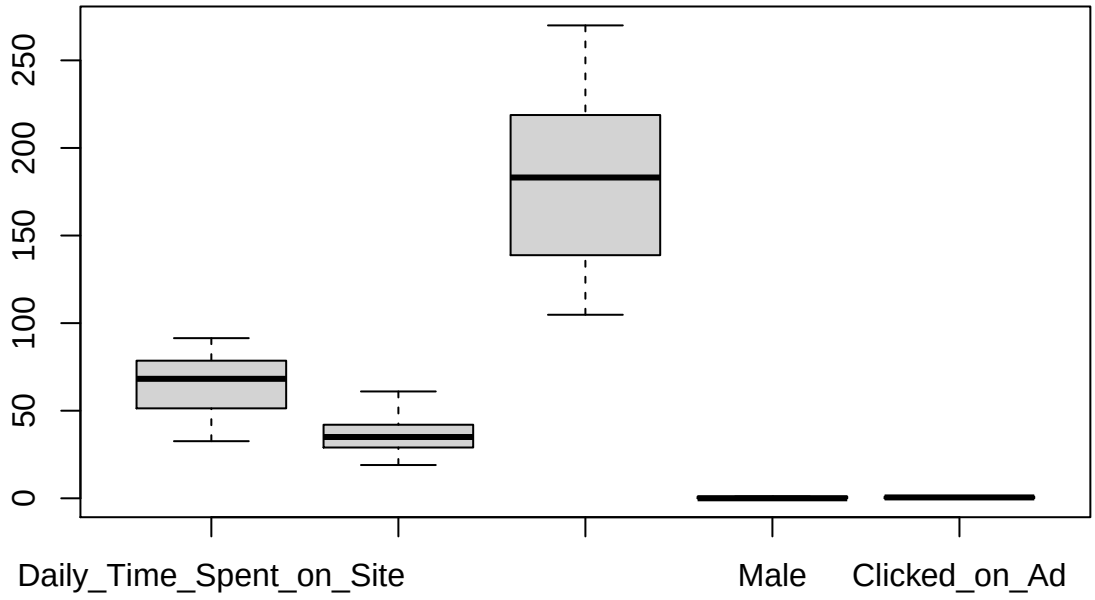
```
## [1] 8.785562
```

```r
# Standard deviation code Daily Time Spent on Site
df1.Daily_Time_Spent_on_Site.sd <- sd(df1$Daily_Time_Spent_on_Site)
df1.Daily_Time_Spent_on_Site.sd
```

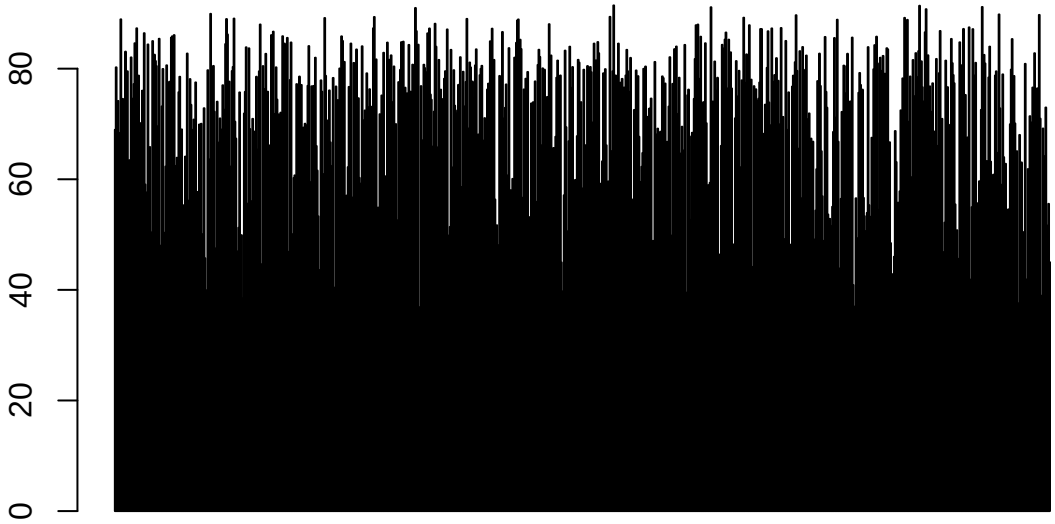
```
## [1] 15.85361
```

Univariate Graphical

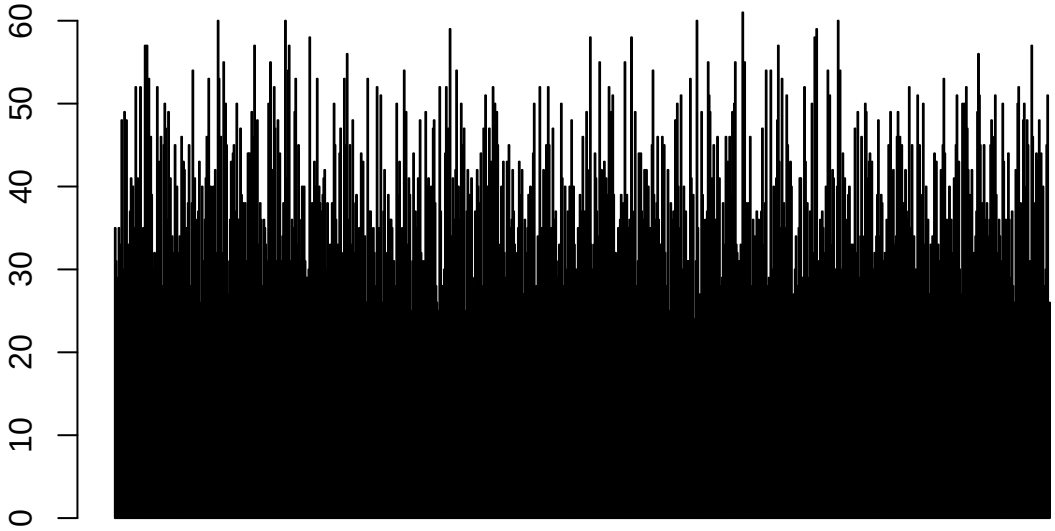
```r
# Visualizing a boxplot of numerical values
boxplot(df1)
```




```r
# Assigning the Daily Time Spent on Site column to the variable Daily Time Spent on Site
Daily_Time_Spent_on_Site <- df1$Daily_Time_Spent_on_Site
# Frequency Distribution
Daily_Time_Spent_on_Site_frequency <- table(Daily_Time_Spent_on_Site)
# Bar plot
barplot(Daily_Time_Spent_on_Site)
```



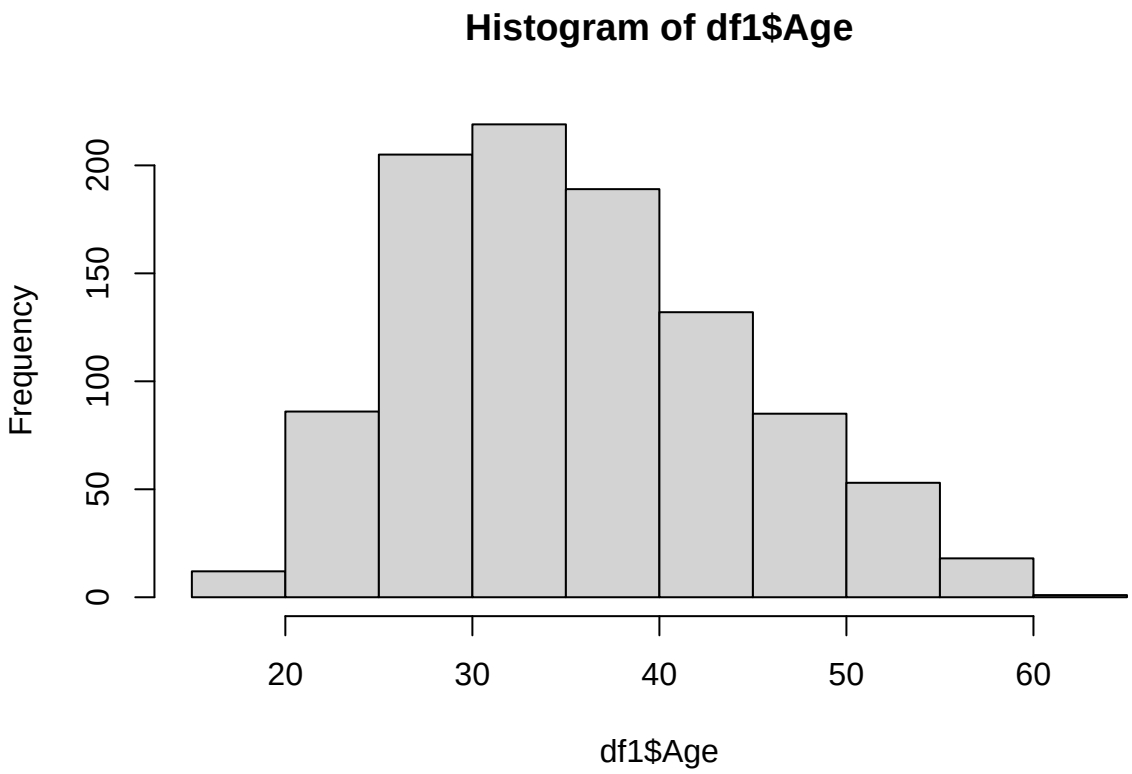

```
# Assigning the age column to the variable age
Age <- df1$Age
# Frequency Distribuution
Age_frequency <- table(Age)
# Bar plot
barplot(Age)
```



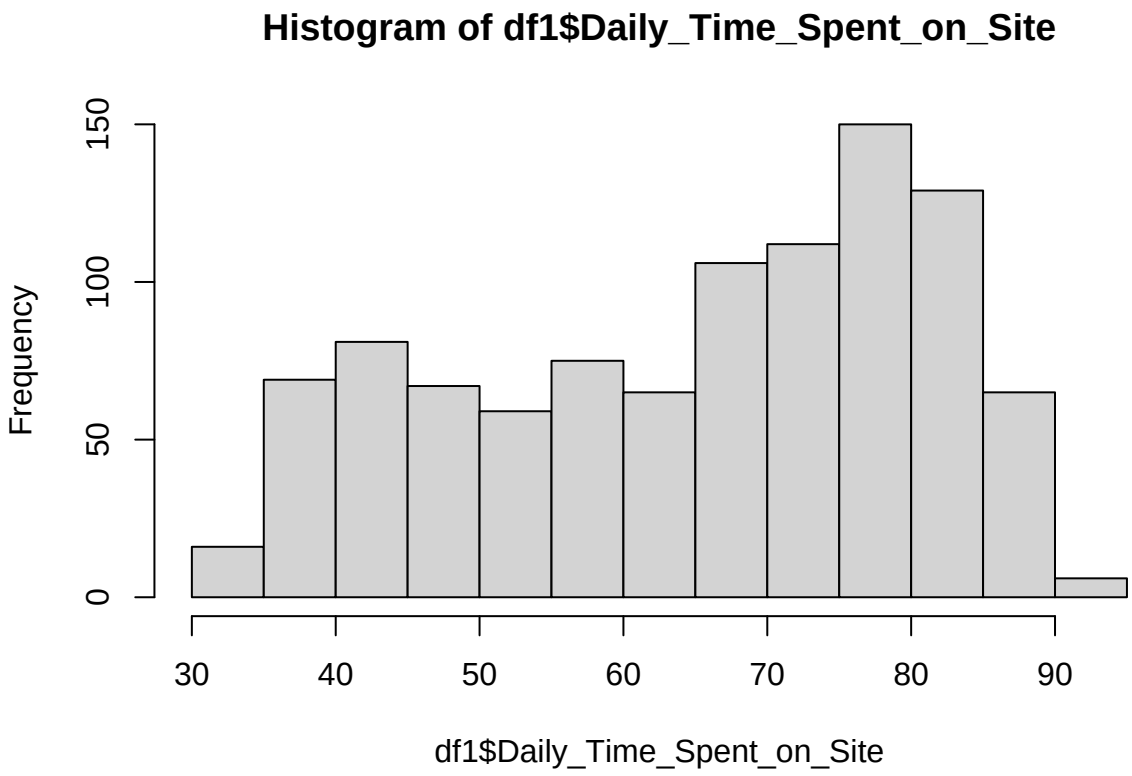

```
# Displaying the column names
colnames(df1)
```

```
## [1] "Daily_Time_Spent_on_Site" "Age"
## [3] "Daily_Internet_Usage"     "Male"
## [5] "Clicked_on_Ad"
```

```
# Histogram of age
hist(df1$Age)
```




```
# Histogram of Daily Time Spent on Site
hist(df1$Daily_Time_Spent_on_Site)
```



Bivariate analysis

```
# Assigning the age column to the variable age
Age<- df1$Age
# Covariance
cov(Daily_Time_Spent_on_Site, Age)
```

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

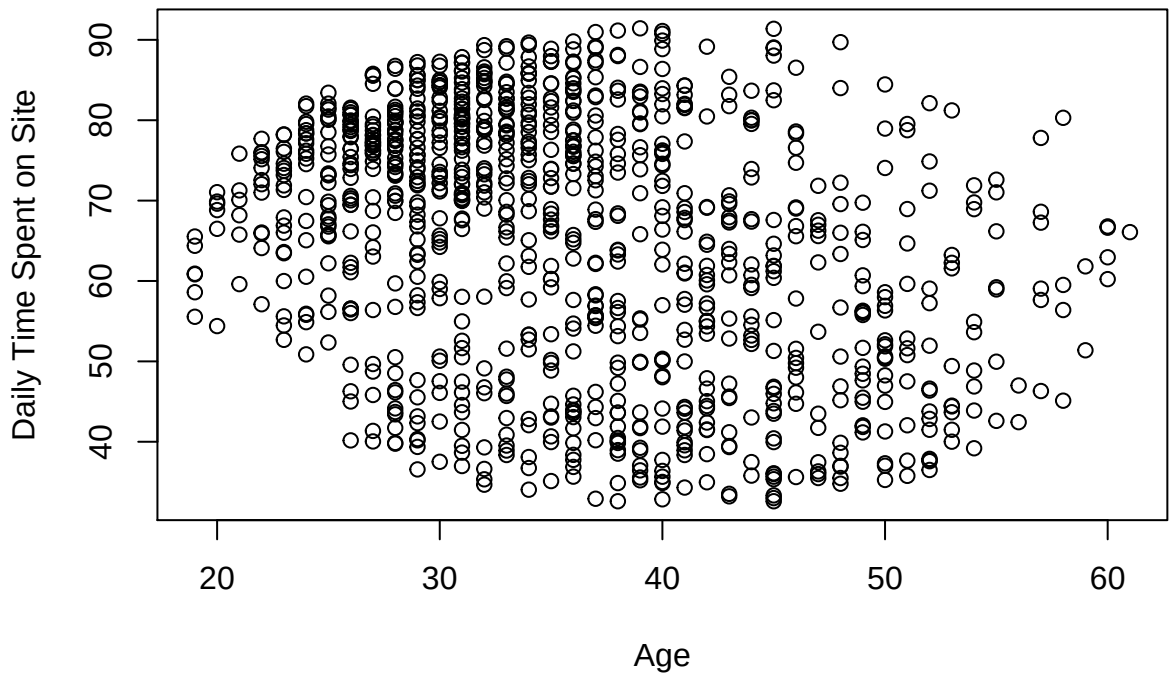
```
# Correlation
cor(Age,Daily_Time_Spent_on_Site)
```

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

There is a negative correlation.

Graphical Techniques

```
# creating a scatterplot
plot(Age, Daily_Time_Spent_on_Site, xlab="Age", ylab="Daily Time Spent on Site")
```




```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
# Rounding the correlation to two decimal places
res <- cor(df1)
round(res, 2)
```

```
##                           Daily_Time_Spent_on_Site   Age Daily_Internet_Usage
## Daily_Time_Spent_on_Site                     1.00 -0.33                 0.52
## Age                                         -0.33  1.00                -0.37
## Daily_Internet_Usage                         0.52 -0.37                 1.00
## Male                                        -0.02 -0.02                 0.03
## Clicked_on_Ad                               -0.75  0.49                -0.79
##                            Male Clicked_on_Ad
## Daily_Time_Spent_on_Site -0.02         -0.75
## Age                      -0.02          0.49
## Daily_Internet_Usage      0.03         -0.79
## Male                      1.00         -0.04
## Clicked_on_Ad            -0.04          1.00
```

4. RECOMMENDATIONS

The people that clicked on the ads on the blog were aged between 19 yrs and 61 years old.

There were no outliers.

The internet usage ranged between 104.8 to 269 units with the time spent on the blog was between 32 to 91 minutes.

There was a negative correlation between age and daily time spent on Site of the individuals.

The ads were mostly viewed by the young and middle aged audience.

5. CONCLUSION

The ads that should be placed on the blog should be relevant to the ages so that the individuals can click on the ads.

For the older people they can minimize the ads and for the younger people they can maximize the ads so that each can relate to ads accordingly.