

Identifying Individuals Likely to Click on Ads

Kibuye Joshua

2022-03-19

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 Metric for Success

Plotting relevant Univariate and Bivariate to identify trends in the dataset. ### c) Understanding the context We are looking at factors contributing to people clicking ads ### d) Recording the Experimental Design 1. Data loading 2. Data Cleaning 3. Exploratory data analysis ### e) Data Relevance ## 2. Reading the Data

```
# Loading the dataset
df <-read.csv("http://bit.ly/IPAdvertisingData")
```

3. Checking the Data

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

```
# Determining the no. of records in our dataset
dim(df)
```

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

There are 1000 records and 10 variables

```
# Previewing 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
```

```
# Previewing 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
```

```
# Checking whether each column has an appropriate datatype
sapply(df, 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"
```

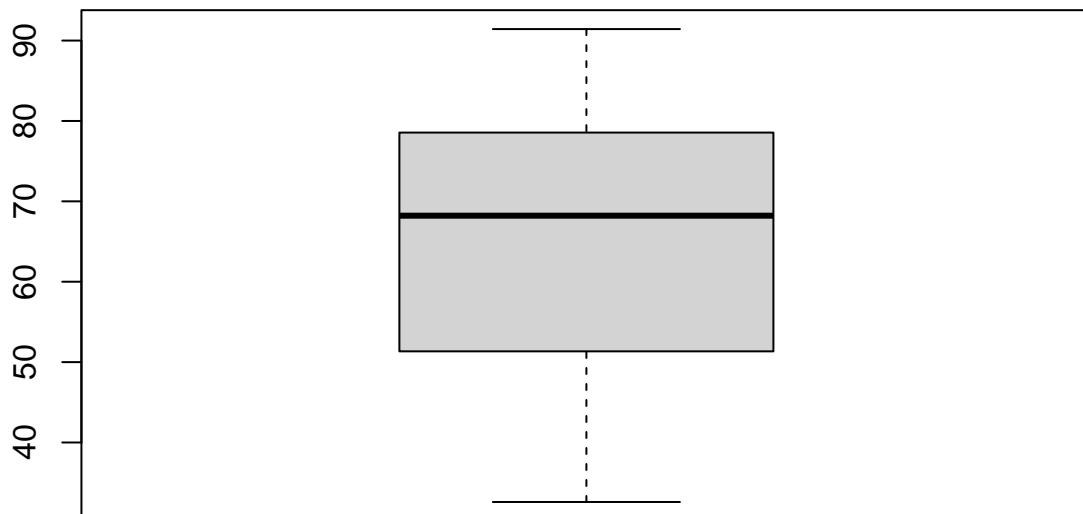
Timestamp should be changed to date time format

4. External Data Source Validation

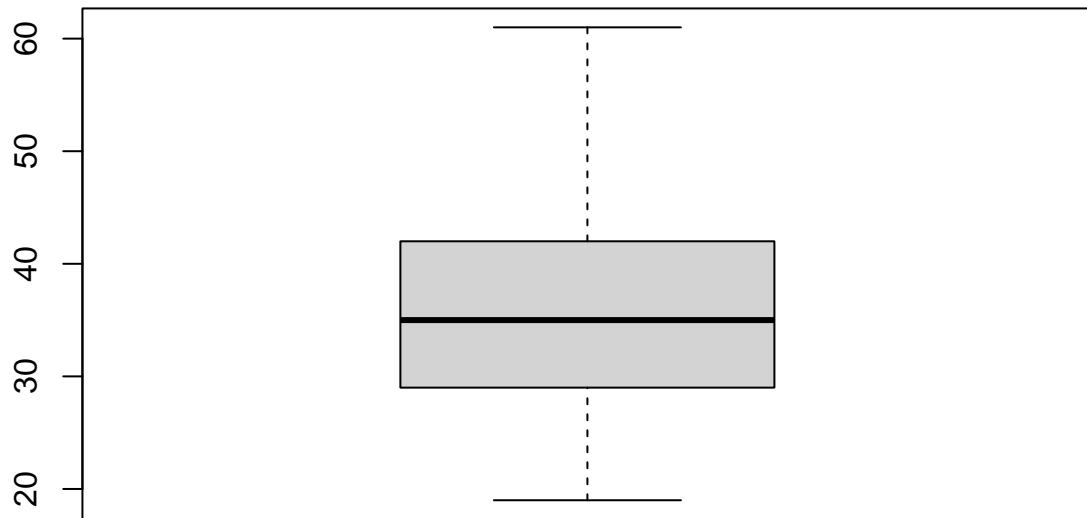
Making sure your data matches something outside of the dataset is very important. It allows you to ensure that the measurements are roughly in line with what they should be and it serves as a check on what other things might be wrong in your dataset. External validation can often be as simple as checking your data against a single number, as we will do here.

5. Tidying the Dataset

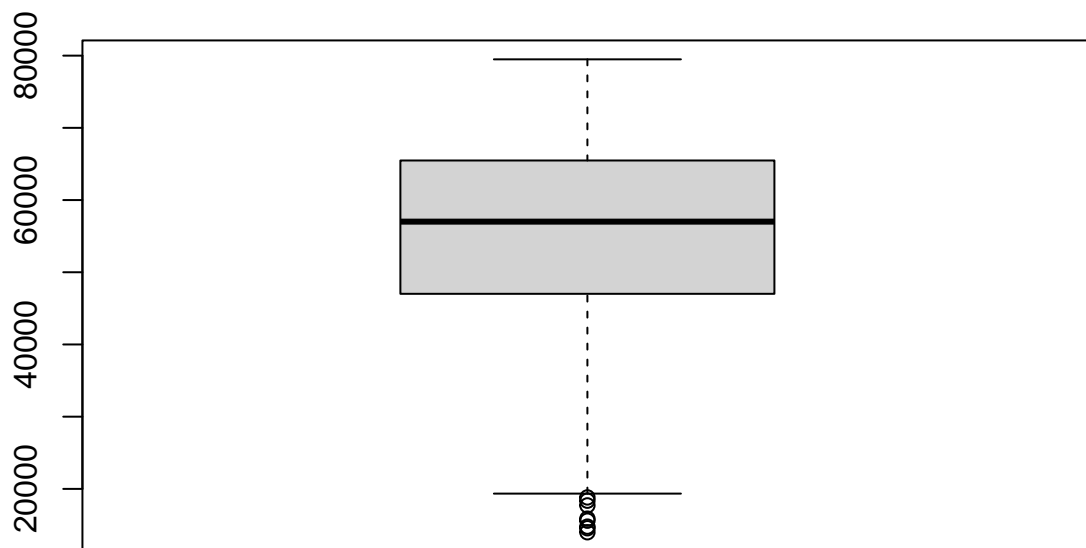
```
# Checking for Outliers
boxplot(df$Daily.Time.Spent.on.Site)
```



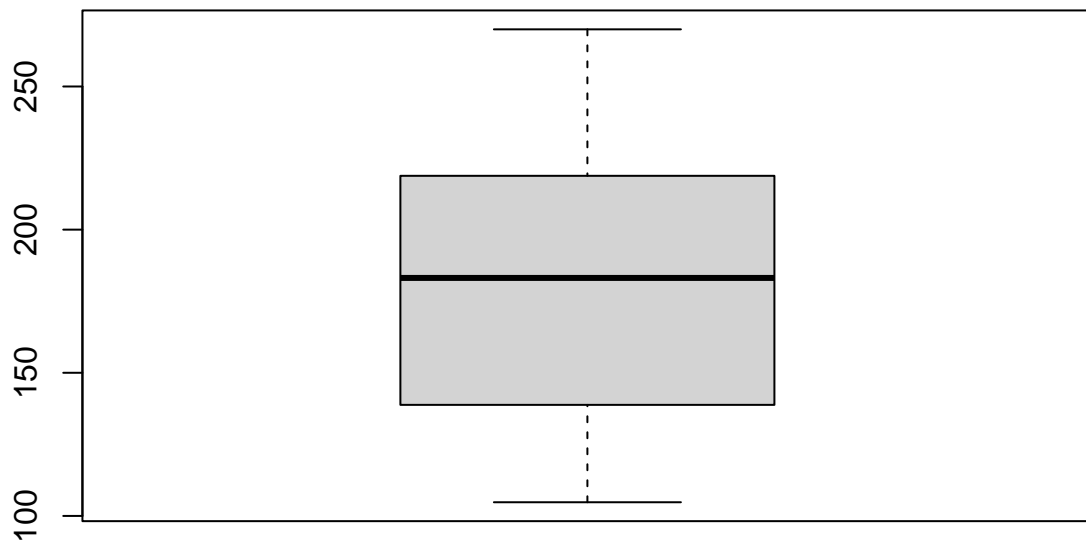
```
boxplot(df$Age)
```



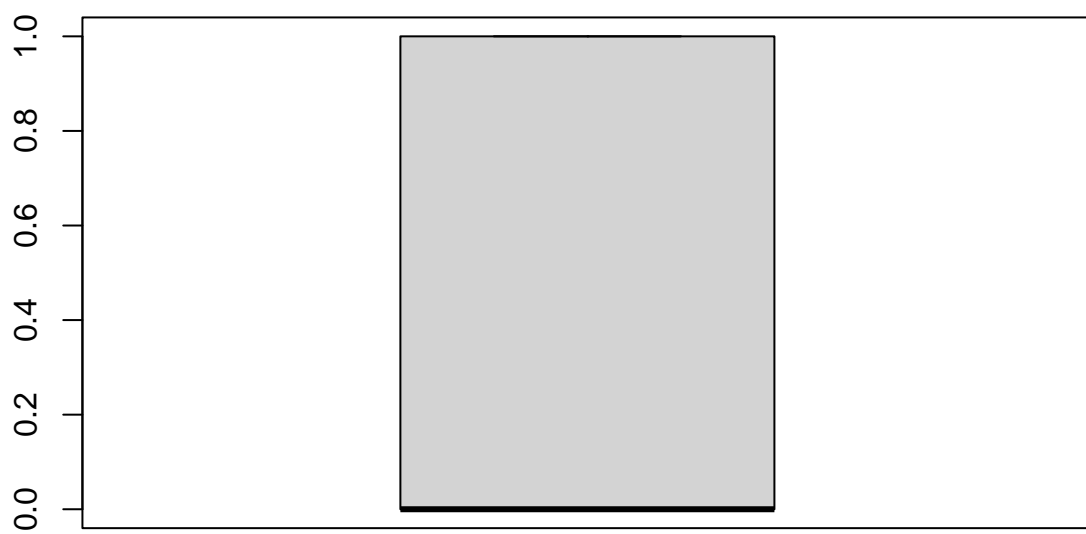
```
boxplot(df$Area.Income)
```



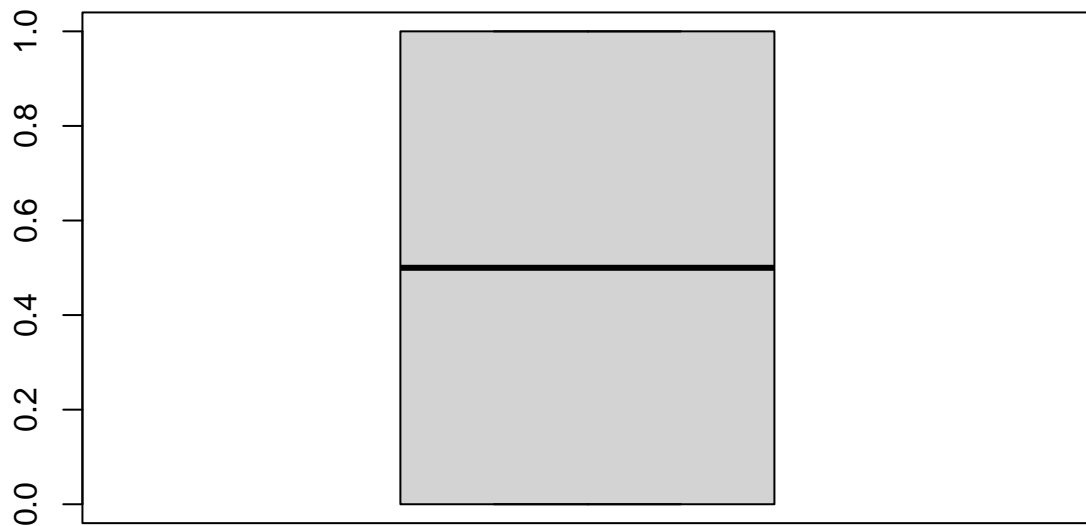
```
boxplot(df$Daily.Internet.Usage)
```



```
boxplot(df$Male)
```



```
boxplot(df$Clicked.on.Ad)
```



The outliers exist in the area income column. However, we do not drop them as they are true values.

```
#Checking for missing data
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
```

There are no missing values

```
#Checking for duplicates
sum(duplicated(df))
```

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

There are no duplicates in the dataset.

```
#Changing column names to lowercase
names(df) <- tolower(names(df))
```



```
#Changing TimeStamp to datetime
df$timestamp <- as.Date(df$timestamp)
```

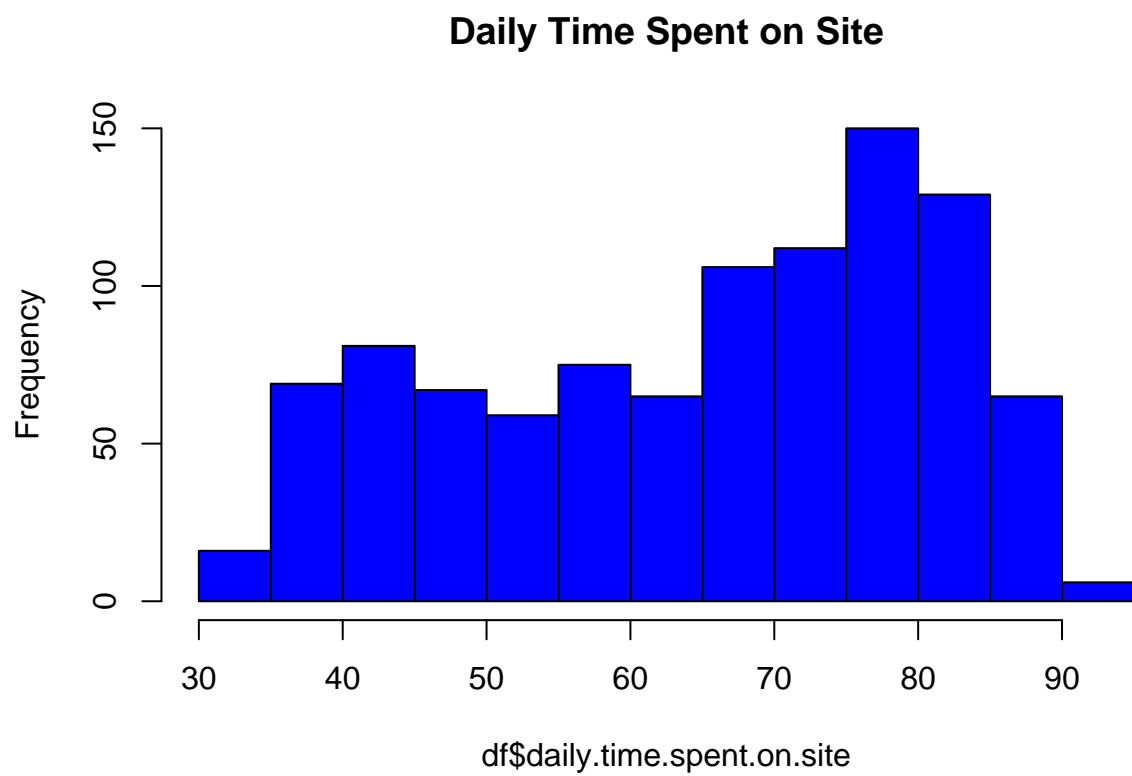
```
head(df)
```

```
##   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            0
## 2 2016-04-04            0
## 3 2016-03-13            0
## 4 2016-01-10            0
## 5 2016-06-03            0
## 6 2016-05-19            0
```

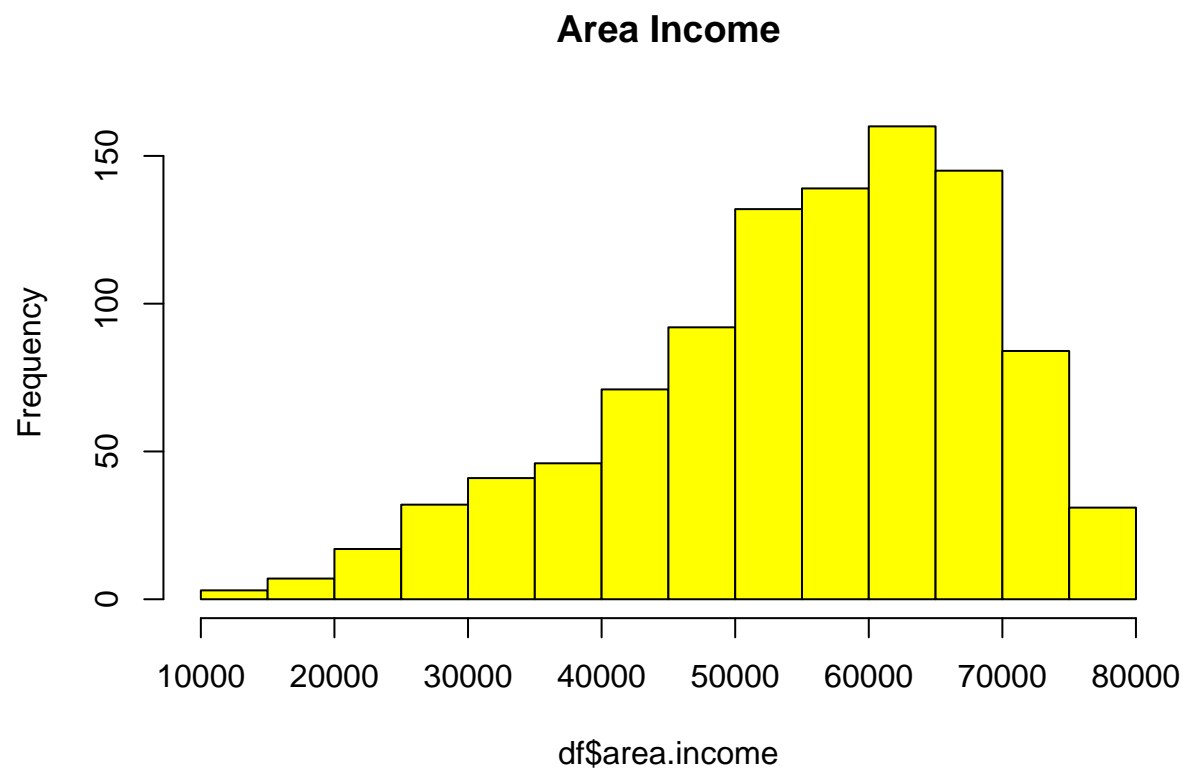
6. Exploratory Analysis

6.1 Univariate analysis

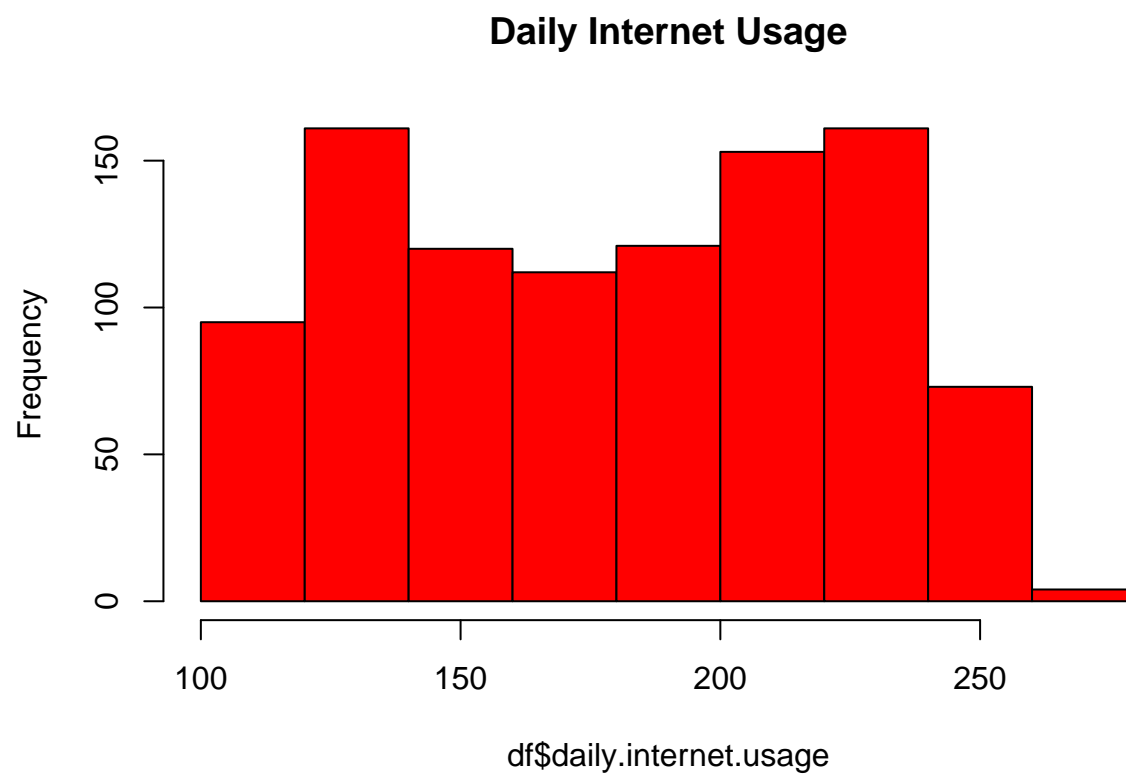
```
nrows = 2
hist(df$daily.time.spent.on.site, main = "Daily Time Spent on Site", col = "blue")
```



```
hist(df$area.income, main = "Area Income", col = "yellow")
```

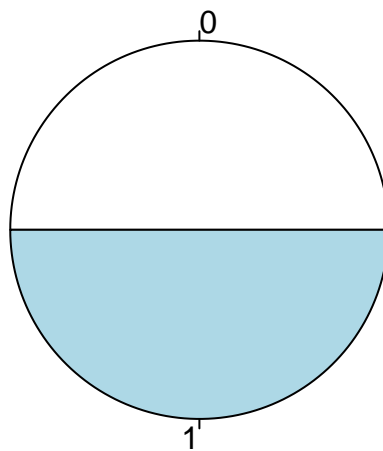


```
hist(df$daily.internet.usage, main = "Daily Internet Usage", col = "red")
```



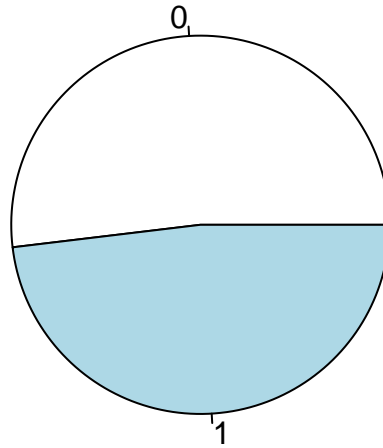
```
pie(table(df$clicked.on.ad), main = "Clicked on Ad")
```

Clicked on Ad



```
pie(table(df$male), main = "Gender of Individual")
```

Gender of Individual



Measures of Dispersion and Central Tendencies

```
#Checking the mean age  
mean(df$age)
```

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

```
mean(df$area.income)
```

```
## [1] 55000
```

6.2 Bivariate Analysis

```
cov(df[, 1:4])
```

```
##               daily.time.spent.on.site      age  area.income  
## daily.time.spent.on.site      251.33709   -46.17415    66130.81  
## age              -46.17415     77.18611   -21520.93  
## area.income      66130.81091 -21520.92580 179952405.95  
## daily.internet.usage      360.99188   -141.63482   198762.53  
##               daily.internet.usage  
## daily.time.spent.on.site      360.9919  
## age              -141.6348  
## area.income      198762.5315  
## daily.internet.usage      1927.4154
```

```
cor(df[,1:4])
```

```
##                daily.time.spent.on.site      age area.income
## daily.time.spent.on.site      1.0000000 -0.3315133  0.3109544
## age                          -0.3315133  1.0000000 -0.1826050
## area.income                  0.3109544 -0.1826050  1.0000000
## daily.internet.usage         0.5186585 -0.3672086  0.3374955
##                daily.internet.usage
## daily.time.spent.on.site      0.5186585
## age                          -0.3672086
## area.income                  0.3374955
## daily.internet.usage         1.0000000
```