Week 12-IP

Definition of Question

Context

An entrepreneur from Kenya has created an online cause which she would like to advertise on her blog. Based on the data she collected from running the ads on her blog, ehe would like a data scientist to help her determine which individuals click on the ads.

Objective

To do an analysis to establish individuals that click on the ads most.

Data Appropriateness

This dataset is suitable for the analysis.

Load Dataset

```
advertising <- read.csv("~/Downloads/advertising.csv")
head(advertising)
```

```
##   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
```

Data Understanding

Dataset summary

```
summary(advertising)
```

```
##  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
##  Adaptive 24hour Graphic Interface      :  1   Lisamouth      :  3
##  Adaptive asynchronous attitude         :  1   Williamsport   :  3
##  Adaptive context-sensitive application :  1   Benjaminchester:  2
##  Adaptive contextually-based methodology:  1   East John      :  2
##  Adaptive demand-driven knowledgebase   :  1   East Timothy   :  2
##  Adaptive uniform capability            :  1   Johnstad       :  2
##  (Other)                                :994   (Other)        :986
##       Male                 Country                  Timestamp   Clicked.on.Ad
##  Min.   :0.000   Czech Republic:  9   2016-01-01 02:52:10:  1   Min.   :0.0
##  1st Qu.:0.000   France        :  9   2016-01-01 03:35:35:  1   1st Qu.:0.0
##  Median :0.000   Afghanistan   :  8   2016-01-01 05:31:22:  1   Median :0.5
##  Mean   :0.481   Australia     :  8   2016-01-01 08:27:06:  1   Mean   :0.5
##  3rd Qu.:1.000   Cyprus        :  8   2016-01-01 15:14:24:  1   3rd Qu.:1.0
##  Max.   :1.000   Greece        :  8   2016-01-01 20:17:49:  1   Max.   :1.0
##                  (Other)       :950   (Other)            :994
```

Checking column names

```
names(advertising)
```

```
##  [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 datatypes

```
#obtaining the datatypes
sapply(data, class)
```

```
##       ...      list   package   lib.loc   verbose     envir overwrite
##    "name"    "call"    "NULL"    "NULL"    "call"    "name" "logical"       "{"
```

Checking for null values and duplicates

```r
colSums(is.na(advertising))
```

```
## 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
```

```r
is.double(advertising)
```

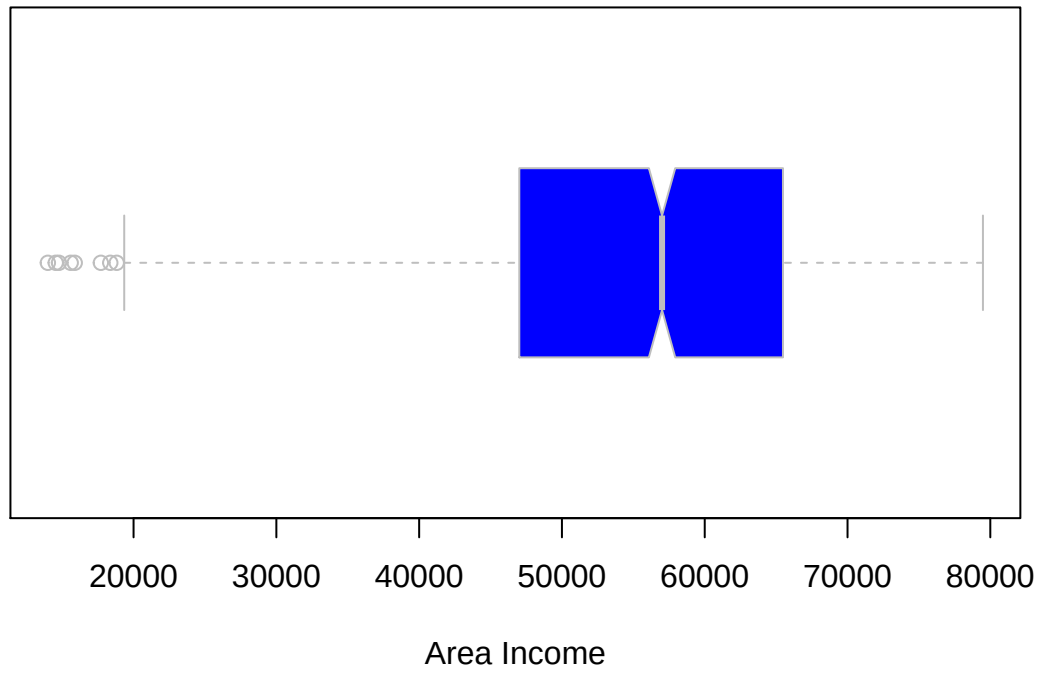
```
## [1] FALSE
```

Conclusion

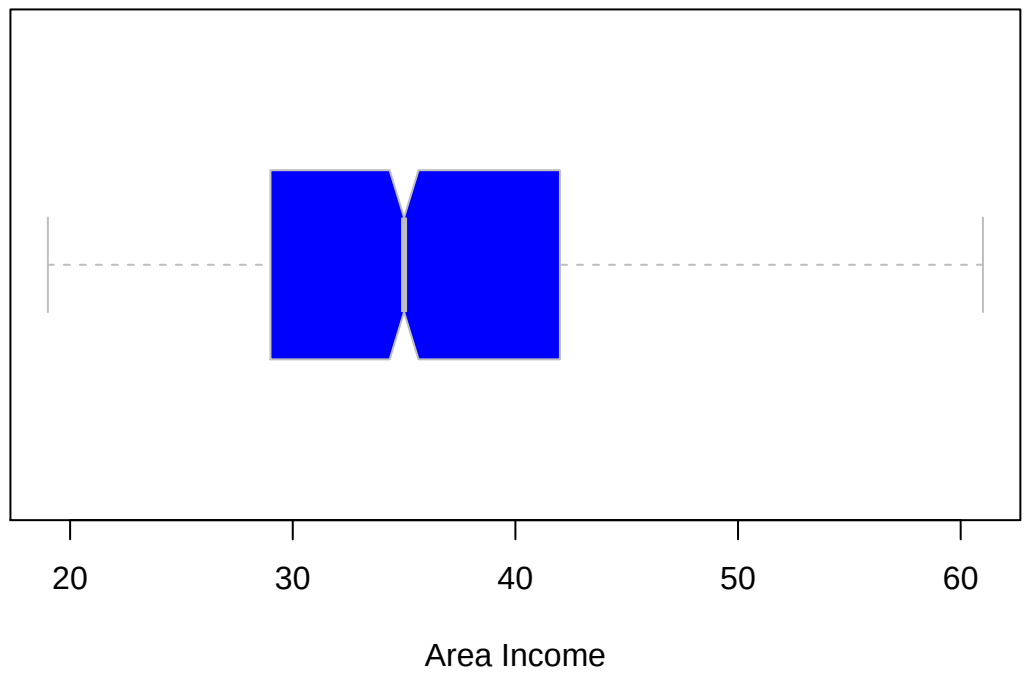
- There are no null values and duplicates in the dataset.

Univariate Analysis

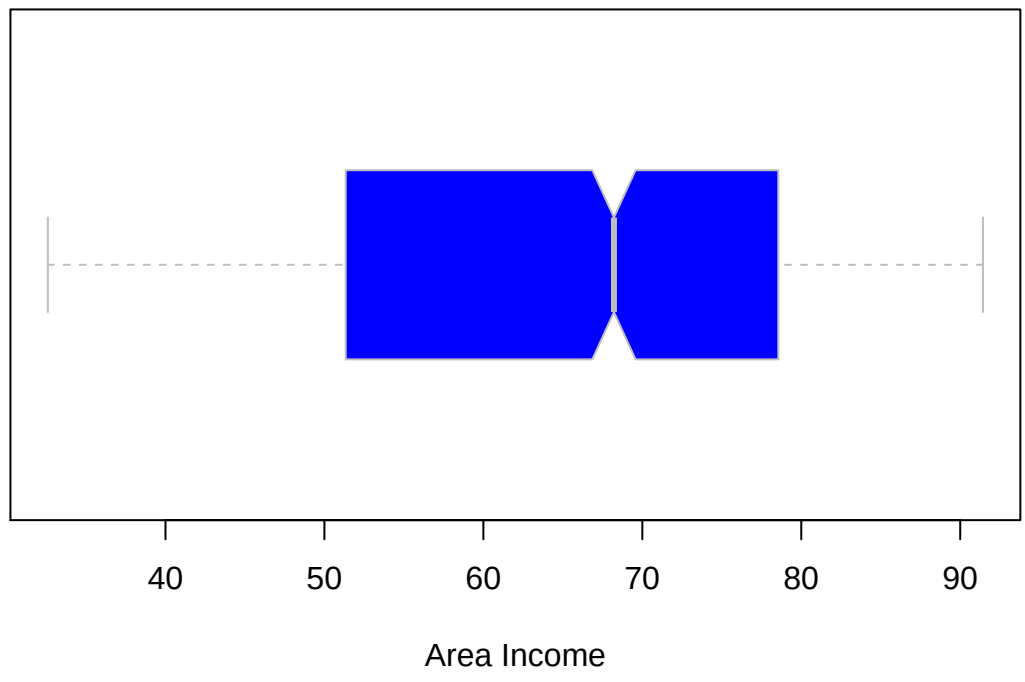
```r
bxplt_Area.Income = boxplot(advertising$Area.Income,
                    main = "Boxplot for Area.Income variable",
                    xlab = "Area Income",
                    col = "blue",
                    border = "grey",
                    horizontal = TRUE,
                    notch = TRUE)
```



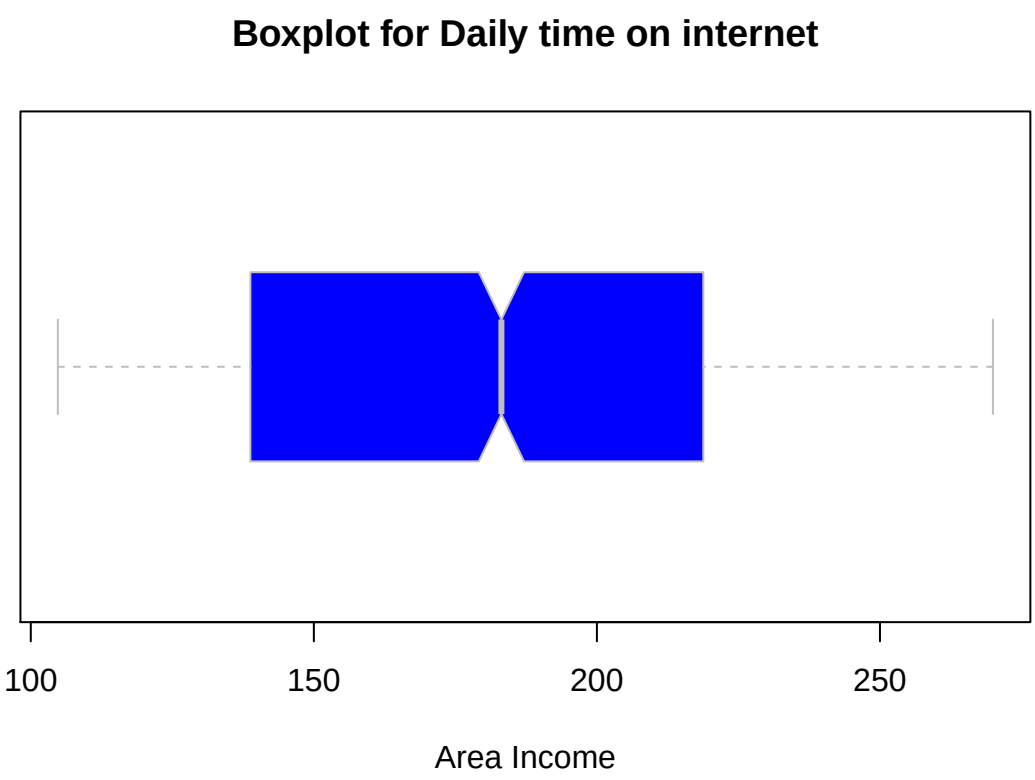

```
bxplt_Area.Income = boxplot(advertising$Age,
                      main = "Boxplot for Age",
                      xlab = "Area Income",
                      col = "blue",
                      border = "grey",
                      horizontal = TRUE,
                      notch = TRUE)
```




```
bxplt_Area.Income = boxplot(advertising$Daily.Time.Spent.on.Site,
                      main = "Boxplot for Time spent on the site",
                      xlab = "Area Income",
                      col = "blue",
                      border = "grey",
                      horizontal = TRUE,
                      notch = TRUE)
```




```
bxplt_Area.Income = boxplot(advertising$Daily.Internet.Usage,
                      main = "Boxplot for Daily time on internet",
                      xlab = "Area Income",
                      col = "blue",
                      border = "grey",
                      horizontal = TRUE,
                      notch = TRUE)
```



Bivariate Analysis

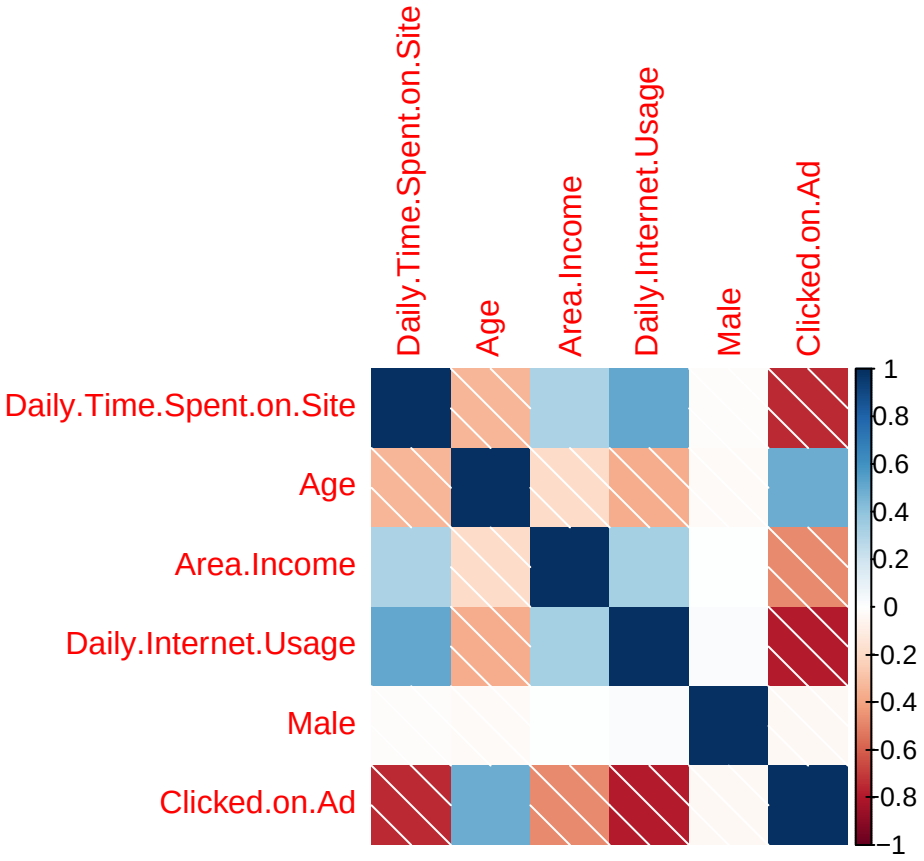
Correlation between variables

```
#Accesing corrplot library
library(corrplot)
```

```
## corrplot 0.84 loaded
```

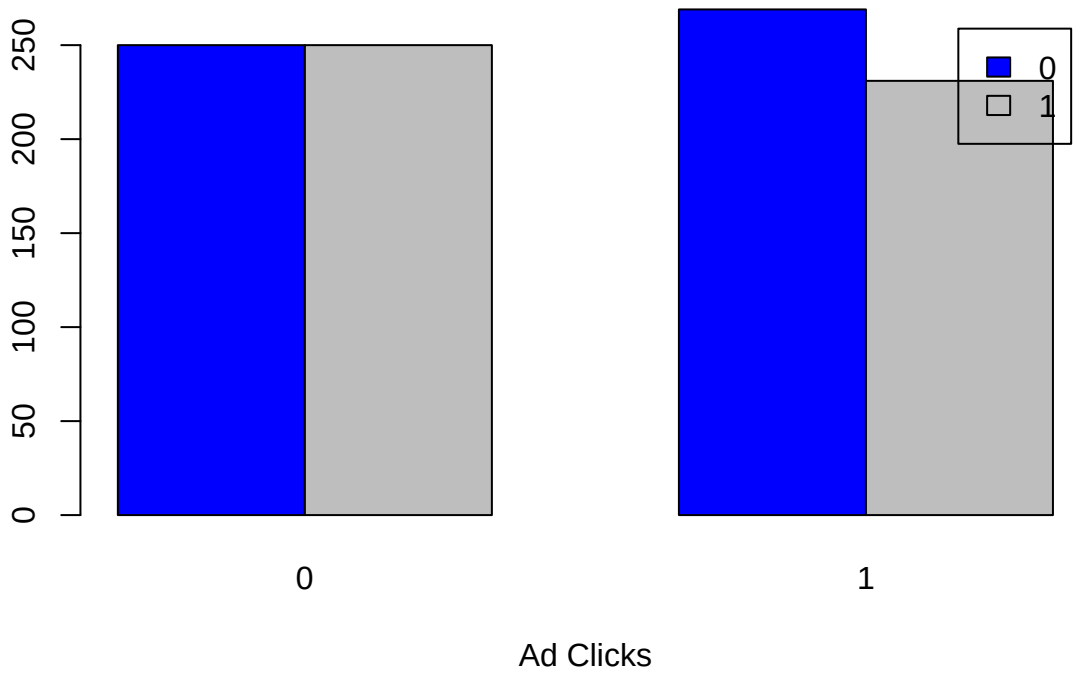
```
#`getting the numeric values of our dataaset
data = advertising[, sapply(advertising, is.numeric)]

#plotting the numeric values.
corrplot(cor(data), method = 'shade')
```



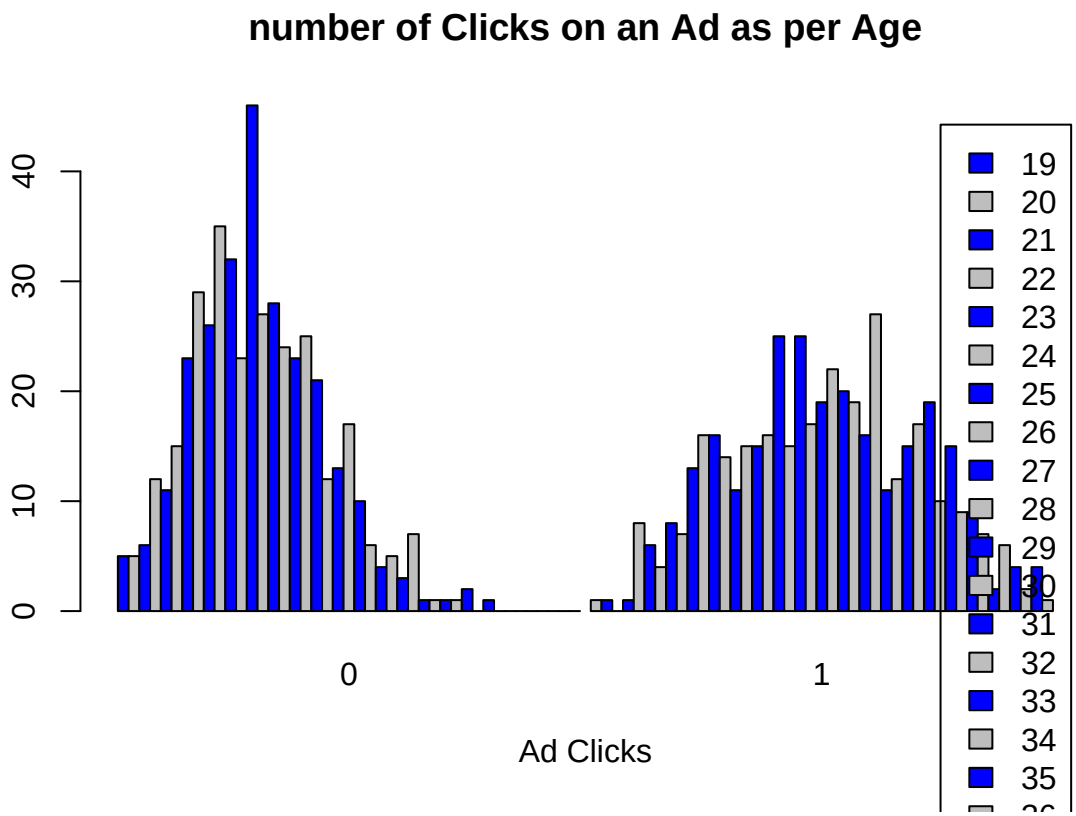
Sex versus Clicks

```
counts = table(advertising$Male, advertising$Clicked.on.Ad)
barplot(counts, main="number of Clicks on an Ad as per each sex, 0=Female, 1=male",
xlab="Ad Clicks", col=c("blue","grey"),
legend = rownames(counts), beside=TRUE)
```



Age versus clicked

```
counts = table(advertising$Age, advertising$Clicked.on.Ad)
barplot(counts, main="number of Clicks on an Ad as per Age",
xlab="Ad Clicks", col=c("blue","grey"),
legend = rownames(counts), beside=TRUE)
```



Conclusion

- People between the age of 37 and 45 year old are more likely to click on the ads
- Females are more likely to clock on the ads