Week 12 Core IP_Advertising Dataset

Martha Irungu

9/3/2020

Ia) Specifying the question

The objective of this study is to use the advertising dataset provided to support a Kenyan entrepreneur identify which individuals are most likely to click on her online advertisement.

b)Defining the Metrics for success

To meet the objective of the study we will need to do the following:

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

ii) Perform univariate and bivariate analysis.

iii) From the analysis done, share insights and provide a conclusion and recommendation.

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.

Online marketing is the practice of leveraging web-based channels to spread a message about a company's brand, products, or services to its potential customers. The methods and techniques used for online marketing include email, social media, display advertising, search engine optimization, Google AdWords and more. The objective of marketing is to reach potential customers through the channels where they spend their time reading, searching, shopping, and socializing online.

Widespread adoption of the internet for business and personal use in Kenya and the world at large has generated new channels for advertising and marketing engagement, including those mentioned above. There are also many benefits and challenges inherent to online marketing, which uses primarily digital mediums to attract, engage, and convert virtual visitors to customers.

As entreprenuers adopt these options, it is important for one to establish the impact of the advertisements and one way to do this is to establish chances of potential customers clicking on the ads and what are charateristics of such customers in order to maximize return on investment. This is exatly what this entreprenuer would like advise on.

d) Recording the experimental design

The following steps were implemented

1.) Business Understanding.

2.) Reading the data.

3.) Data Exploration and cleaning to prepare the data for analysis

4.) Univariate,Bivariate analysis

5.) Conclusion of the findings and recommendation.

e)Data Relevance

The data provided for this study consists of columns with factors likely to influence an individual to either click on Ad or not.Since the data was collected when the entreprenuer was running another Ad on the same blog, it is relevant to help us establish what kind of audience are likely to click on an Ad or not.

2. Reading and cheking the data

##Reading and previewing that dataset

```
library("data.table")
advertising <- fread("/Users/marthairungu/desktop/R/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
```

##Checking the summary of the dataset

```
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                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
##  Length:1000        Min.   :0.0
##  Class :character   1st Qu.:0.0
##  Mode  :character   Median :0.5
##                     Mean   :0.5
##                     3rd Qu.:1.0
##                     Max.   :1.0
```

#This shows the summary of numeric variables as tabulated.

#Checking the number of rows and columns

```
dim(advertising)
```

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

#We observe that the dataset has 1,000 observations and 10 variables

#Checking the structure of the dataset

```
str(advertising)
```

```
## Classes 'data.table' and 'data.frame':   1000 obs. of  10 variables:
##  $ Daily Time Spent on Site: num  69 80.2 69.5 74.2 68.4 ...
##  $ Age                     : int  35 31 26 29 35 23 33 48 30 20 ...
##  $ Area Income             : num  61834 68442 59786 54806 73890 ...
##  $ Daily Internet Usage    : num  256 194 236 246 226 ...
##  $ Ad Topic Line           : chr  "Cloned 5thgeneration orchestration"
"Monitored national standardization" "Organic bottom-line service-desk" "Triple-
```

```
buffered reciprocal time-frame" ...
##  $ City                    : chr  "Wrightburgh" "West Jodi" "Davidton" "West
Terrifurt" ...
##  $ Male                    : int  0 1 0 1 0 1 0 1 1 1 ...
##  $ Country                 : chr  "Tunisia" "Nauru" "San Marino" "Italy" ...
##  $ Timestamp               : chr  "2016-03-27 00:53:11" "2016-04-04 01:39:02"
"2016-03-13 20:35:42" "2016-01-10 02:31:19" ...
##  $ Clicked on Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

#We observe that out dataset has columns as listed, the datatypes are in numbers, integers and character/string. We will change the columns appropriately.We note that Ad Topic Line, City,Male, country and Clicked on Ad are categorical data whose data type is integer but should be factors.Factors are variables in R which take on a limited number of different values; such variables are often referred to as categorical variables.We will change this to the right datatype format.

##Details of the columns

#Daily Time Spent on Site: the daily time spent in minutes and seconds

#Age: Age of the individuals

#Area Income:Income earned in that area

#Daily Internet Usage:Daily usgae of internet

#Ad Topic Line:Topic of the Ad

#City:City the individual comes from

#Male columns: This represents 0: for female and 1: male

#Country: Name of country

#Time stamp:Time in year,month,date,hour,minutes and seconds

#Clicked on Ad:Chances of clicking on the Ad or not. 0: Not click, 1:click on the Ad

#Checking the class of the dataset

```
class(advertising)
```

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

#We will change the datatypes in integer format to factor. We will leave the ones in character datatype as we will still be able to get the information

```
advertising$Male <- as.factor(advertising$Male)
advertising$Clicked_on_Ad <- as.factor(advertising$Clicked_on_Ad)



#Checking if the changes have been effected
str(advertising$Male)

##  Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 2 2 ...

str(advertising$Clicked_on_Ad)

##  Factor w/ 0 levels: NA NA NA NA NA NA NA NA NA NA ...
```

#Spliting the time stamp column into year, Month, day, hour and minute for ease of determining which year,month,day,hour, minute individuals are likely to click on the Ad or not

```
advertising$year <- format(as.POSIXct(advertising$Timestamp, format="%Y-%m-%d
%H:%M:%S"), "%Y")
advertising$month <- format(as.POSIXct(advertising$Timestamp, format="%Y-%m-%d
%H:%M:%S"), "%m")
advertising$day <- format(as.POSIXct(advertising$Timestamp, format="%Y-%m-%d
%H:%M:%S"), "%d")
advertising$hour <- format(as.POSIXct(advertising$Timestamp, format="%Y-%m-%d
%H:%M:%S"), "%H")
advertising$minute <- format(as.POSIXct(advertising$Timestamp, format="%Y-%m-%d
%H:%M:%S"), "%M")
```

#Printing the head to confirm this has been effected

head(advertising)

#Check the data structure to estblish the data types of date

str(advertising)

#We note that year,month,day, hour and minute are in character datatype. We will change this to factor

```
advertising$year <- as.factor(advertising$year)
advertising$year <- as.factor(advertising$month)
advertising$year <- as.factor(advertising$day)
advertising$year <- as.factor(advertising$hour)
advertising$year <- as.factor(advertising$minute)
```

#Checking for missing values in the dataset

```
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            Clicked_on_Ad                      year
##                        0                      1000                         0
##                    month                       day                      hour
##                        0                         0                         0
##                   minute
##                        0
```

#We note that our dataset does not have missing values. So we will not need to omit or replace them.

#Checking for duplicates in our dataset

```
duplicated_rows <- advertising[duplicated(advertising),]
duplicated_rows
```

```
## Empty data.table (0 rows and 16 cols): Daily Time Spent on Site,Age,Area
Income,Daily Internet Usage,Ad Topic Line,City...
```

#We observe that our dataset does not have duplicates

3. #**Univariate Graphical Exploratory Data Analysis**

#a). Measures of Central Tendency #Checking the mean of numerical Variables

```
Age.mean <- mean(advertising$Age)
Age.mean
```

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

#The mean age of individuals is 36 years

```
Age.median <-median(advertising$Age)
Age.median
```

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

#The median age of individuals is 35 years

getmode <- function(v) { uniqv <- unique(v) uniqv[which.max(tabulate(match(v, uniqv)))] } # Calculating the mode using out getmode() function

```
#Age.mode <- getmode(advertising$Age)
```

#b). Measures of Dispersion

#Checking the minimum age

```
Age.min <-min(advertising$Age)
Age.min

## [1] 19
```

#The minimum age of individual is 19 years

#Checking the maximum age

```
Age.max <-max(advertising$Age)
Age.max

## [1] 61
```

#Maximum age is 61 years

#Checking the range(Difference between highest age and lowest age)

```
Age.range <-range(advertising$Age)
Age.range

## [1] 19 61
```

#Getting the first and third quantile and range using the quantile function

```
Age.quantile <-quantile(advertising$Age)
Age.quantile

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

#The age of 25%centile is 29, 50%centile is 35 and 3rd quantile is 42

#Finding the variance of age. This is a numerical measure of how the data values is dispersed around the mean.

```
Age.variance<-var(advertising$Age)
Age.variance
```

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

#Finding the standard deviation of age.

```
Age.sd<-sd(advertising$Age)
Age.sd
```

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

#Standard deviation of age is 8.78, this is a measure of spread around the mean.

#To get measures of central tendancy and dispersion for the other numerical varibles, we can use the summary function.

```
summary(advertising)
```

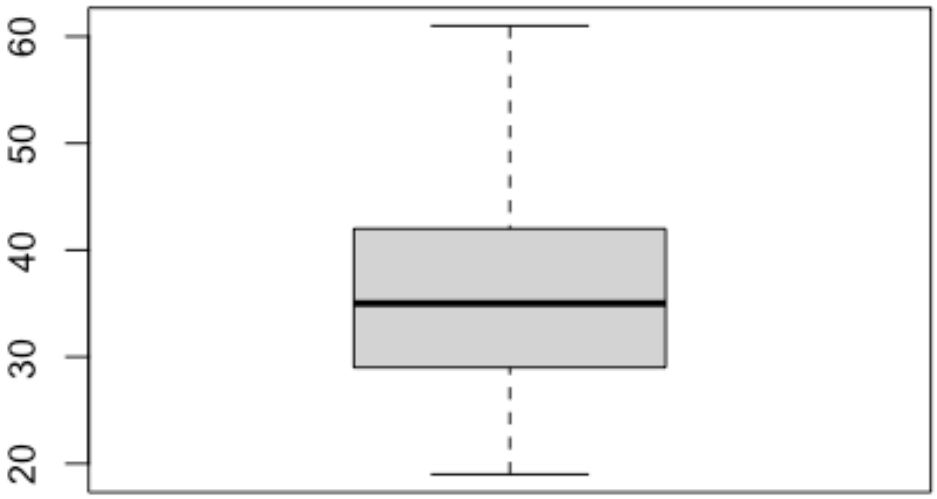
```
##  Daily Time Spent on Site      Age         Area Income    Daily Internet
Usage
##  Min.   :32.60            Min.   :19.00   Min.   :13996   Min.   :104.8
##  1st Qu.:51.36            1st Qu.:29.00   1st Qu.:47032   1st Qu.:138.8
##  Median :68.22            Median :35.00   Median :57012   Median :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            Male      Country
##  Length:1000        Length:1000        0:519   Length:1000
##  Class :character   Class :character   1:481   Class :character
##  Mode  :character   Mode  :character           Mode  :character
##
##
##
##
##   Timestamp         Clicked on Ad Clicked_on_Ad      year
##  Length:1000        Min.   :0.0   NA's:1000     02     : 26
##  Class :character   1st Qu.:0.0                07     : 24
##  Mode  :character   Median :0.5                13     : 24
##                     Mean   :0.5                10     : 22
##                     3rd Qu.:1.0                21     : 21
##                     Max.   :1.0                33     : 21
##                                                (Other):862
##     month              day                hour              minute
##  Length:1000        Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
```

#c). Univariate analysis

#Plotting boxplot for Age variable

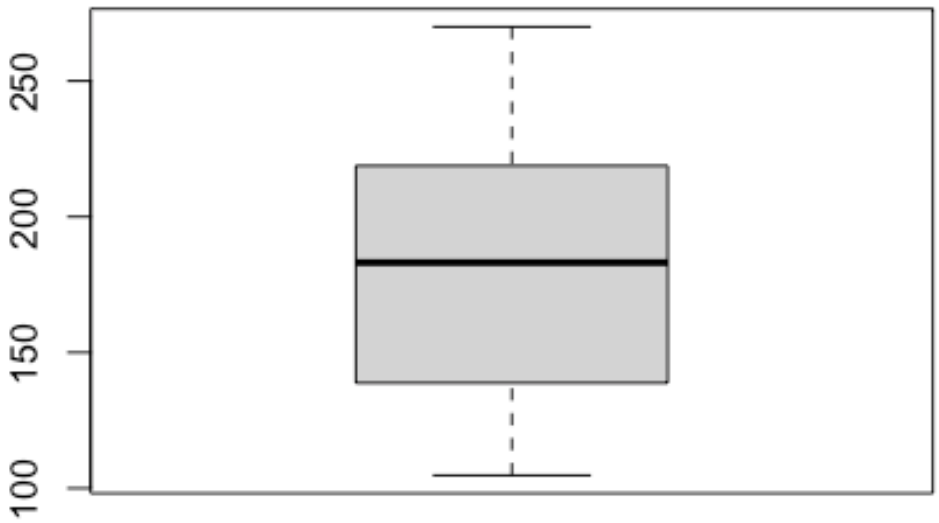
```
boxplot(advertising$Age)
```



#We observe that age has no outliers

#Plotting boxplot for Daily Internet Usage variable

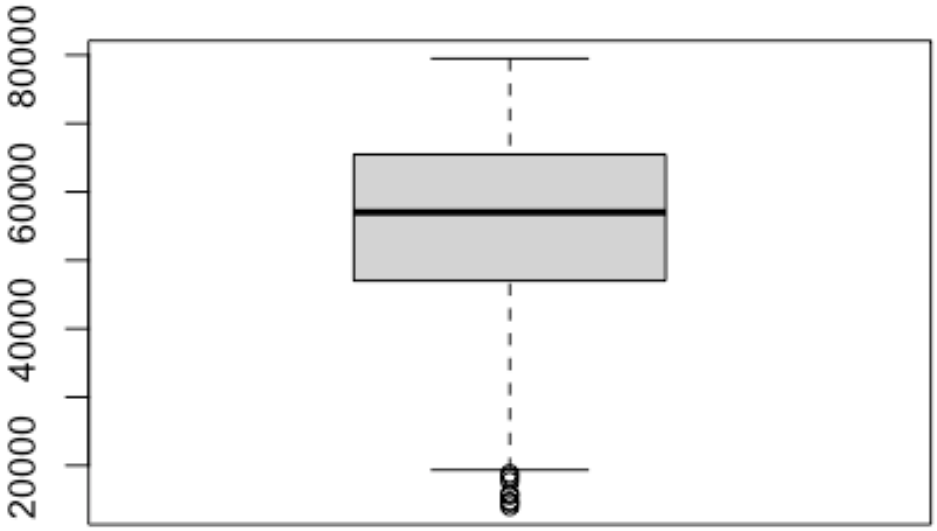
```
boxplot(advertising$'Daily Internet Usage')
```



#Daily Internet Usage has no outliers

#Plotting boxplot for Area Income variable

```
boxplot(advertising$'Area Income')
```



#Area Income has some outliers

#Checking the number of male and female individuals represented

```
Male_table <- table(advertising$Male)
Male_table
```

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

#We observe that we have 519 female and 481 Male in the dataset

#Checking the number of those who clicked on the Ad and those individuals who did not click

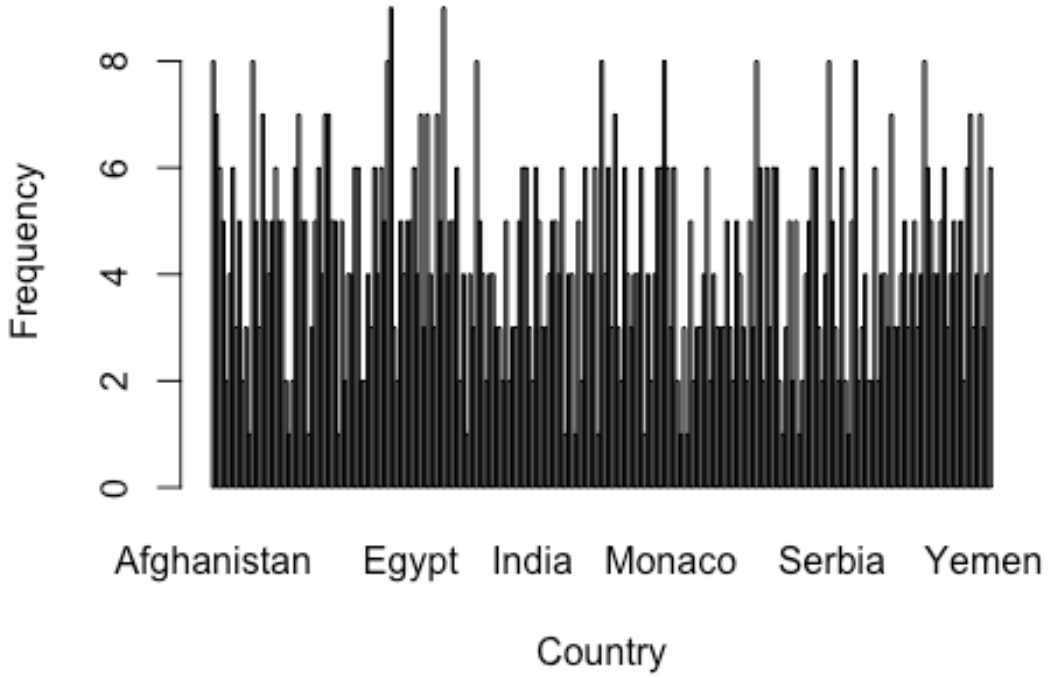
```
Clicked_on_Ad_table <- table(advertising$`Clicked on Ad`)
Clicked_on_Ad_table
```

```
##
##   0   1
## 500 500
```

#We observe that individuals who clicked on Ad 500, same as those who did not click on the Ad

#Plotting barplot frequency by country

```
Country <-advertising$Country
Country_frequency <-table(Country)
barplot(Country_frequency,xlab='Country',ylab='Frequency')
```

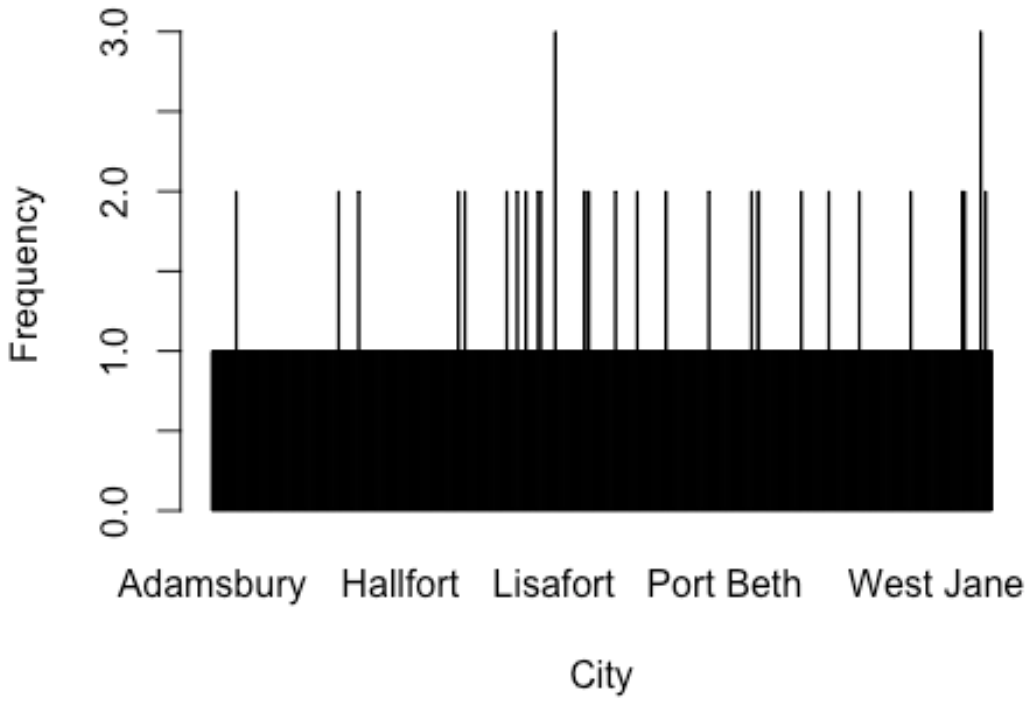


#For better visualization we can check the top 6 Countries that occur freequently

#Czech Republic and France have the highest frequency

#Plotting frequency by city

```
City <-advertising$City
City_frequency <-table(City)
barplot(City_frequency,xlab='City',ylab='Frequency')
```

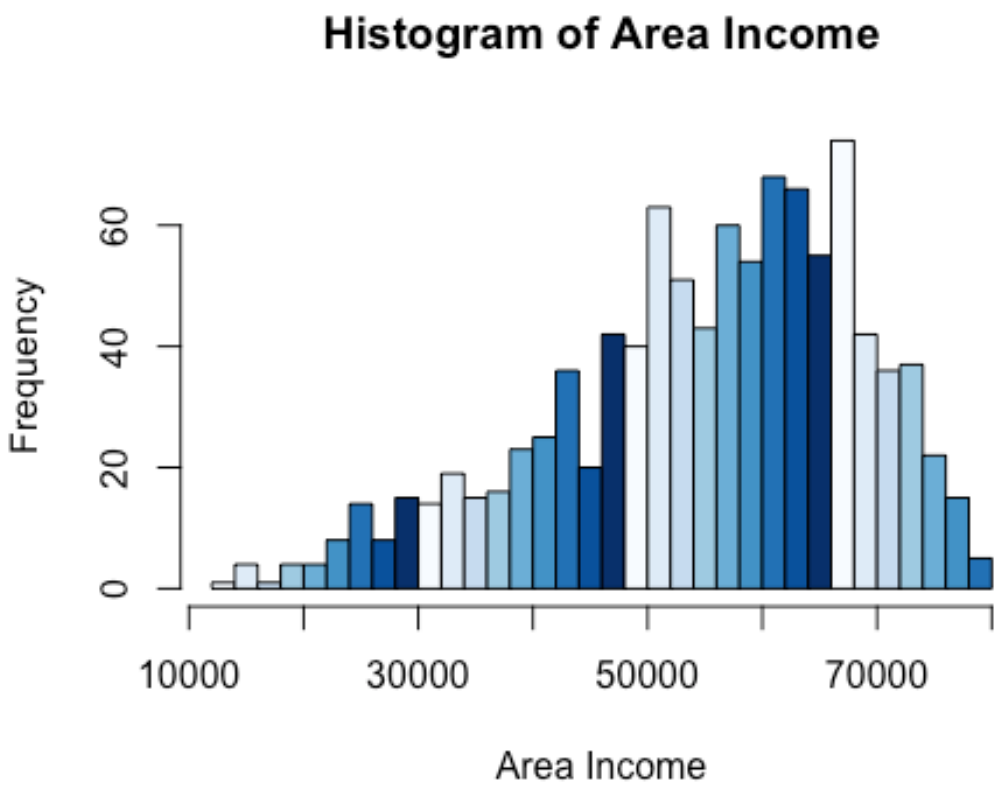


#For better visualization we can check the top 6 Cities that occur frequently

Lisamoth and Williamsport are top 2 cities

#Histogram for Area Income

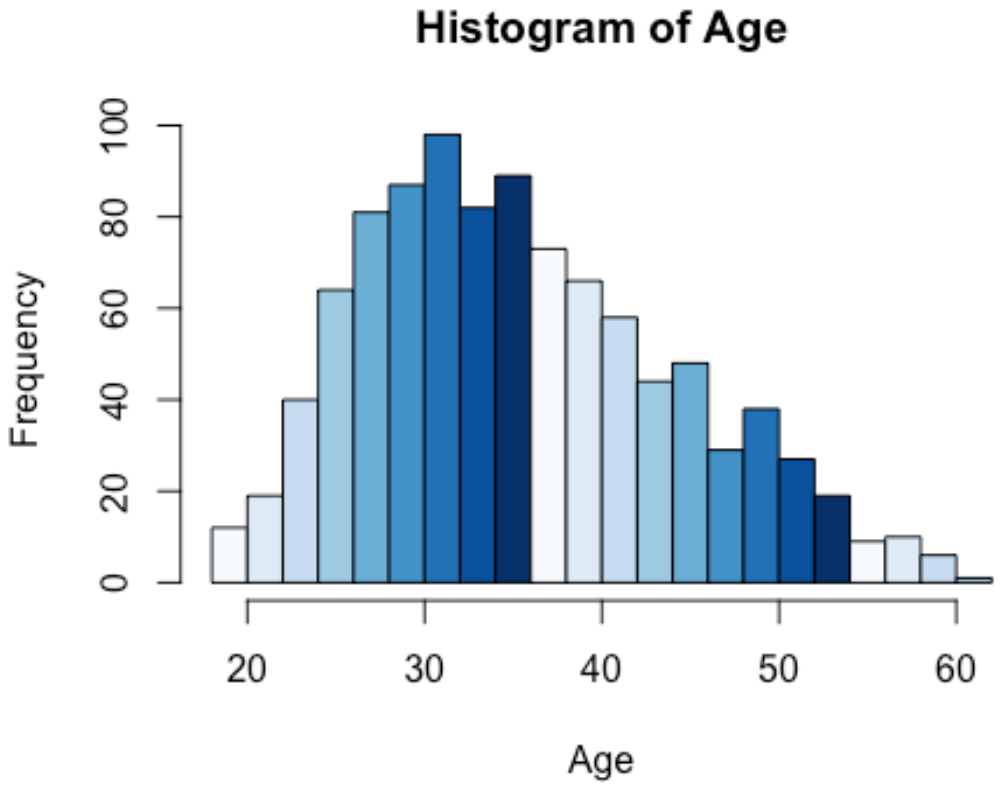
```
hist(advertising$`Area Income`, col=blues9,breaks=25,xlab="Area
Income",main="Histogram of Area Income")
```



#Area income is negatively skewed and most of the area income is about 6000.

#Histogram- Age of individuals

```
hist(advertising$`Age`, col=blues9,breaks=25,xlab="Age",main="Histogram of Age")
```

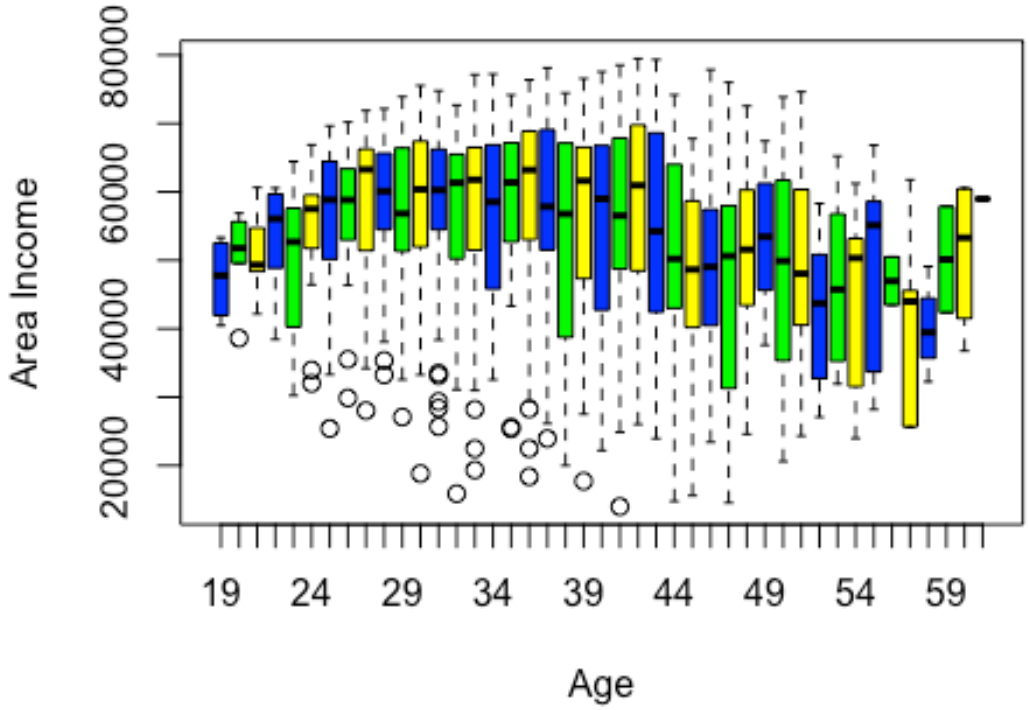


#Age distribution is positively skewed with median age being 35years and mean age 36years.

#c). Bivariate analysis

#Plotting boxplot for Area Income and Age variables

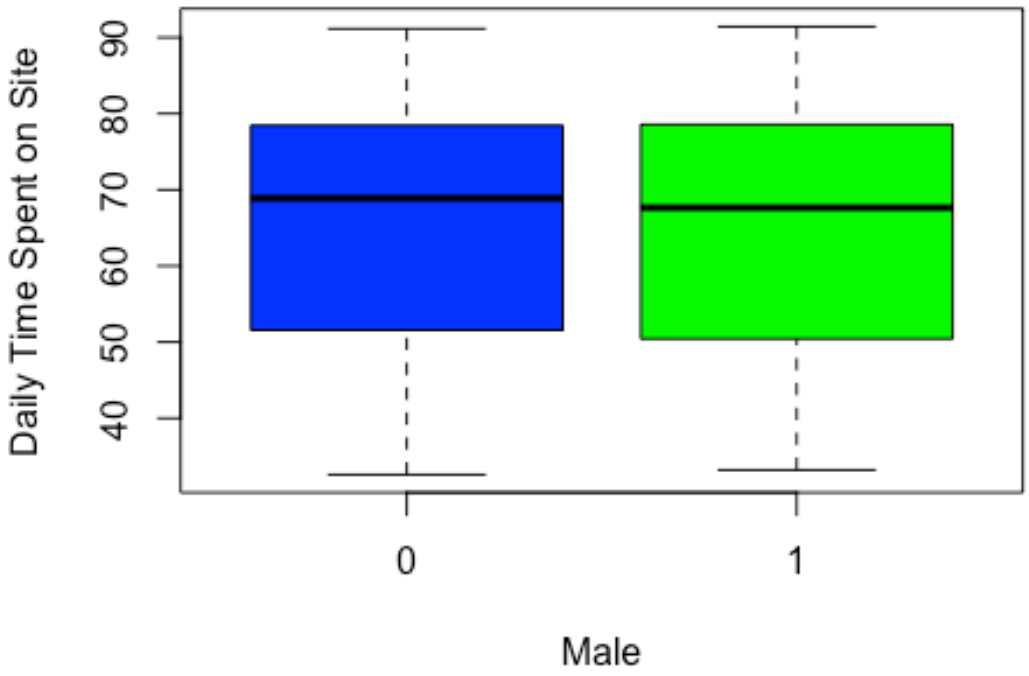
```
boxplot(advertising$'Area Income'~advertising$Age,xlab="Age", ylab="Area
Income", notch=FALSE, col=c("blue","green","yellow"))
```



#We note that Area income increases with increase in age and begins to decline from age 43. We observe outliers in area income below the age of 40.

#Plotting boxplot for Daily Time Spent on Site and Male variables

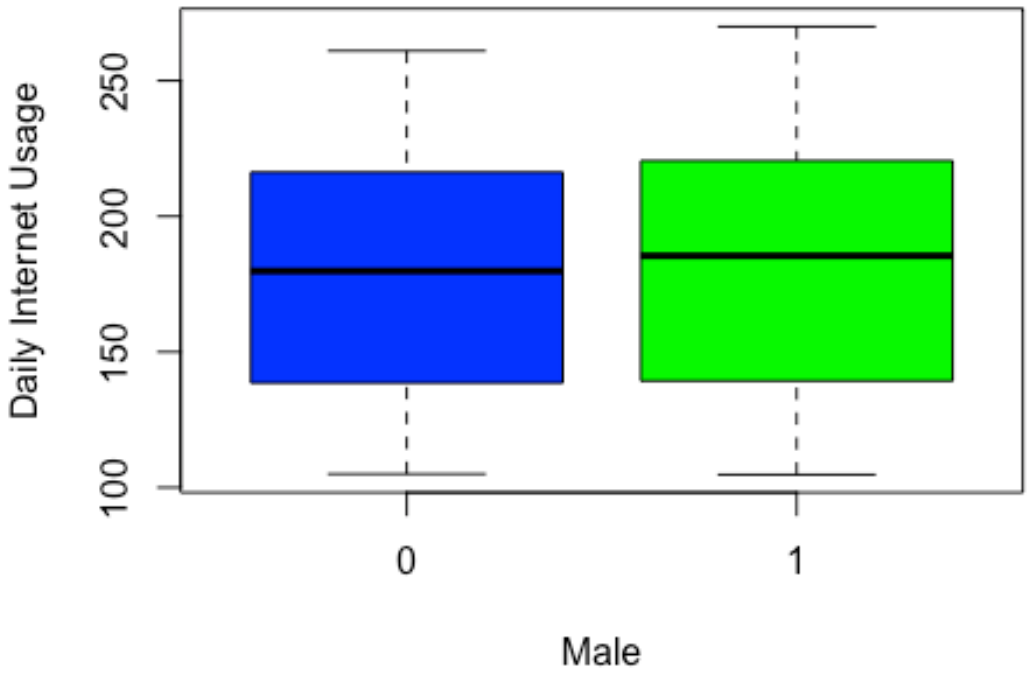
```
boxplot(advertising$'Daily Time Spent on Site'~advertising$Male,xlab="Male",
ylab="Daily Time Spent on Site",notch=FALSE, col=c("blue","green"))
```



#Time spent on the internet for both male and female is more or less the same

#Plotting boxplot for Daily Time Spent on Site and Male variables

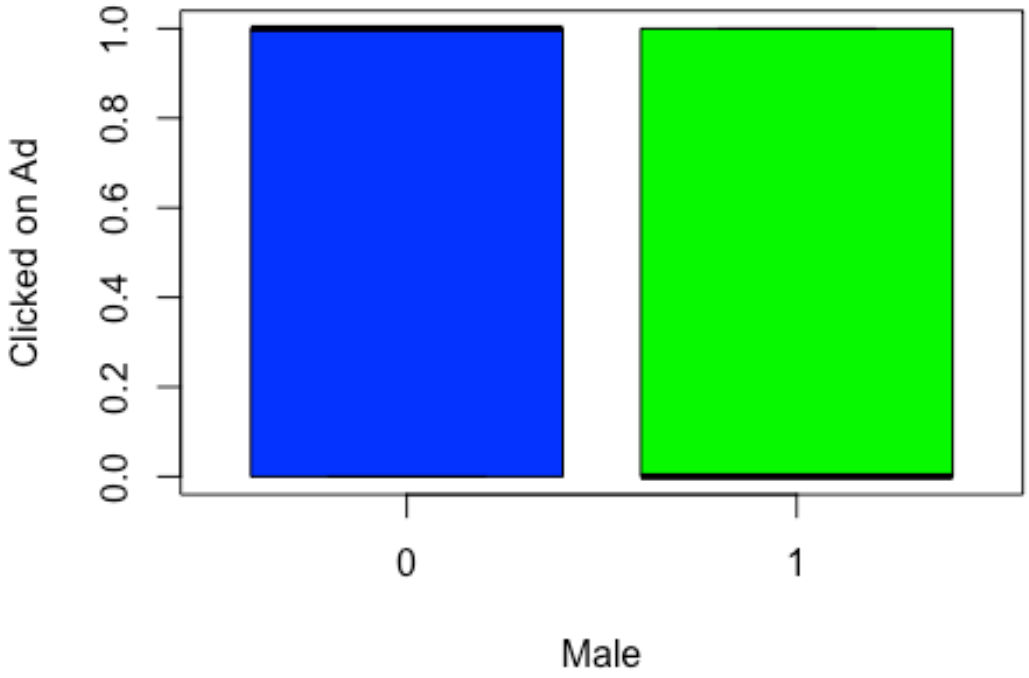
```
boxplot(advertising$'Daily Internet Usage'~advertising$Male,xlab="Male",
ylab="Daily Internet Usage",notch=FALSE, col=c("blue","green"))
```



#Daily internet usage for male is slightly higher than that of female with no outliers.

#Plotting relationship of those who clicked on the Ad by Gender using boxplot

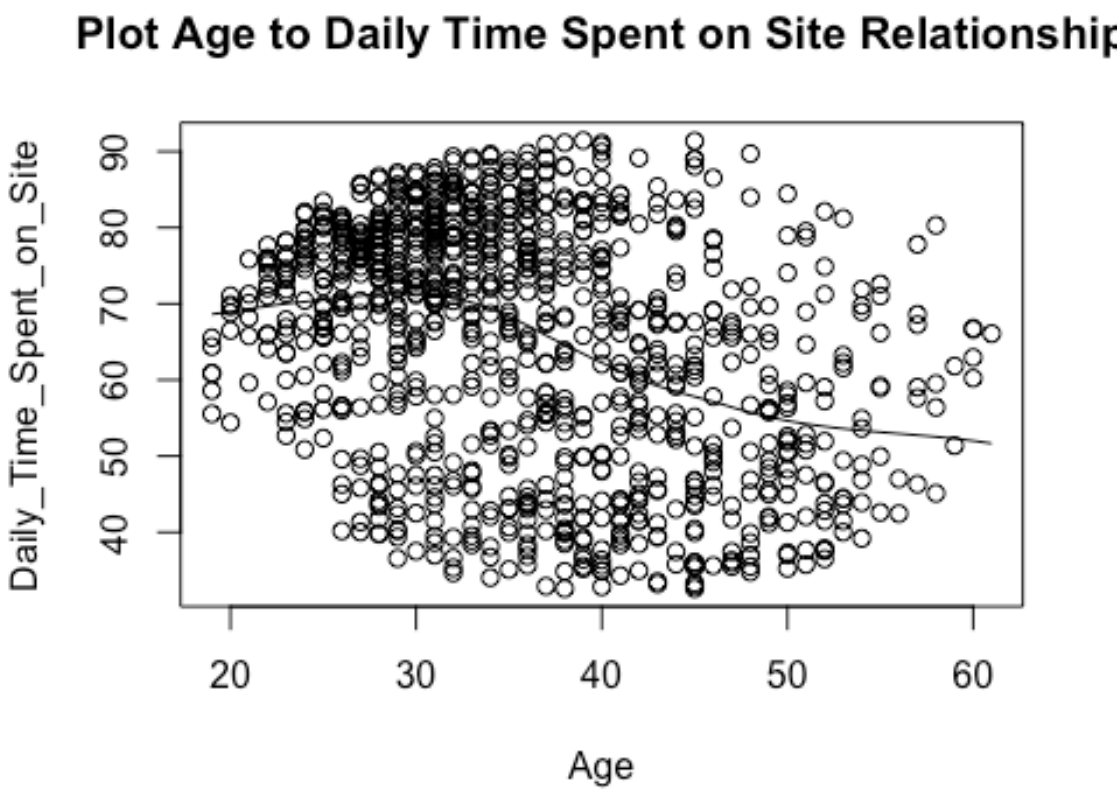
```
boxplot(advertising$'Clicked on Ad'~advertising$Male,xlab="Male", ylab="Clicked
on Ad",notch=FALSE, col=c("blue","green"))
```



#From the above we observe that the number of male and female who clicked on the Ad are more or less the same.Take a 50/50 kind of representation, hence either gender has chances of clicking on the Ad.

#The correlation between age and time spent on the site

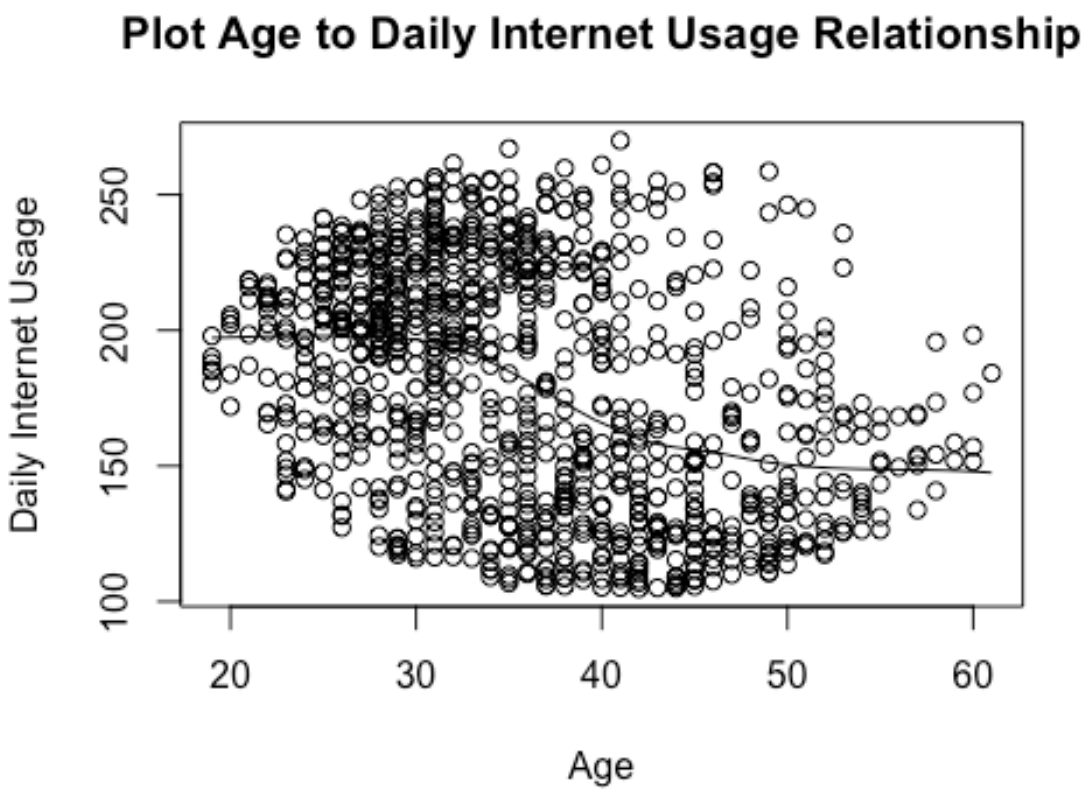
```
scatter.smooth(advertising$Age,advertising$`Daily Time Spent on Site`,main="Plot Age to Daily Time Spent on Site Relationship",xlab = 'Age',ylab = 'Daily_Time_Spent_on_Site')
```



#We observe that daily time spent site is higher with younger individuals, this time declines with age. We observe a peak at age 30years.

#Plotting relationship between Age and Daily Internet Usage

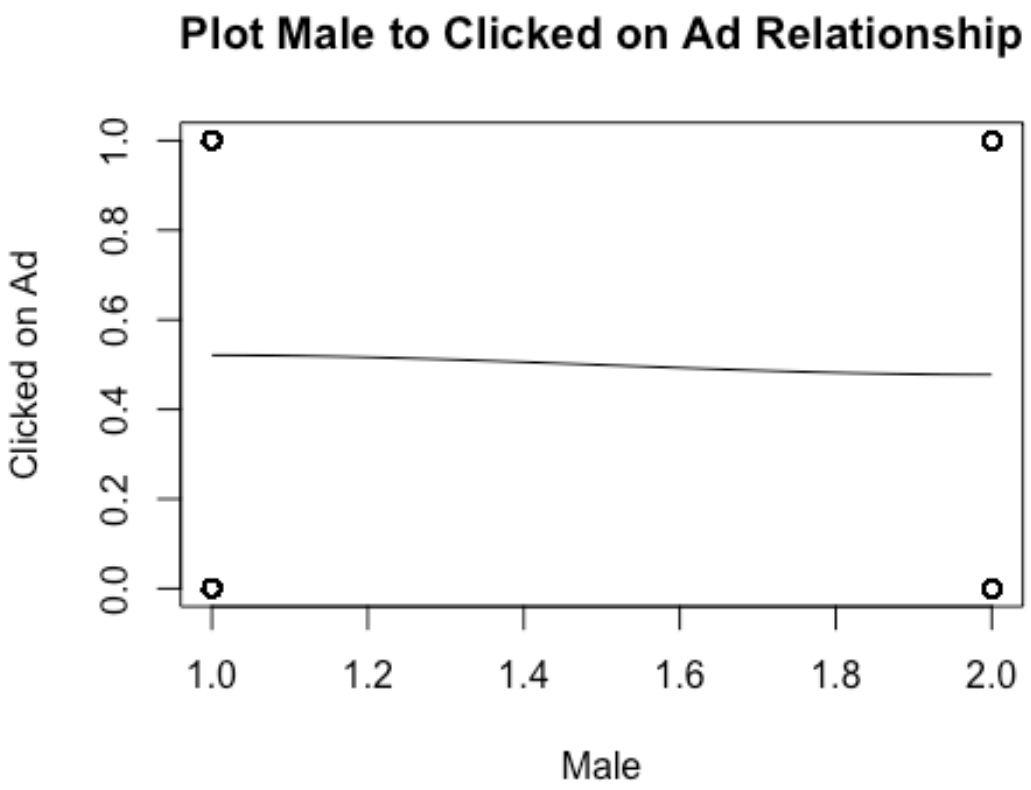
```
scatter.smooth(advertising$Age,advertising$`Daily Internet Usage`,main="Plot Age to Daily Internet Usage Relationship",xlab = 'Age',ylab = 'Daily Internet Usage')
```



#We observe that Daily Internet Usage declines with increase in age.

#Checking the relationship gender and clicking on the ad

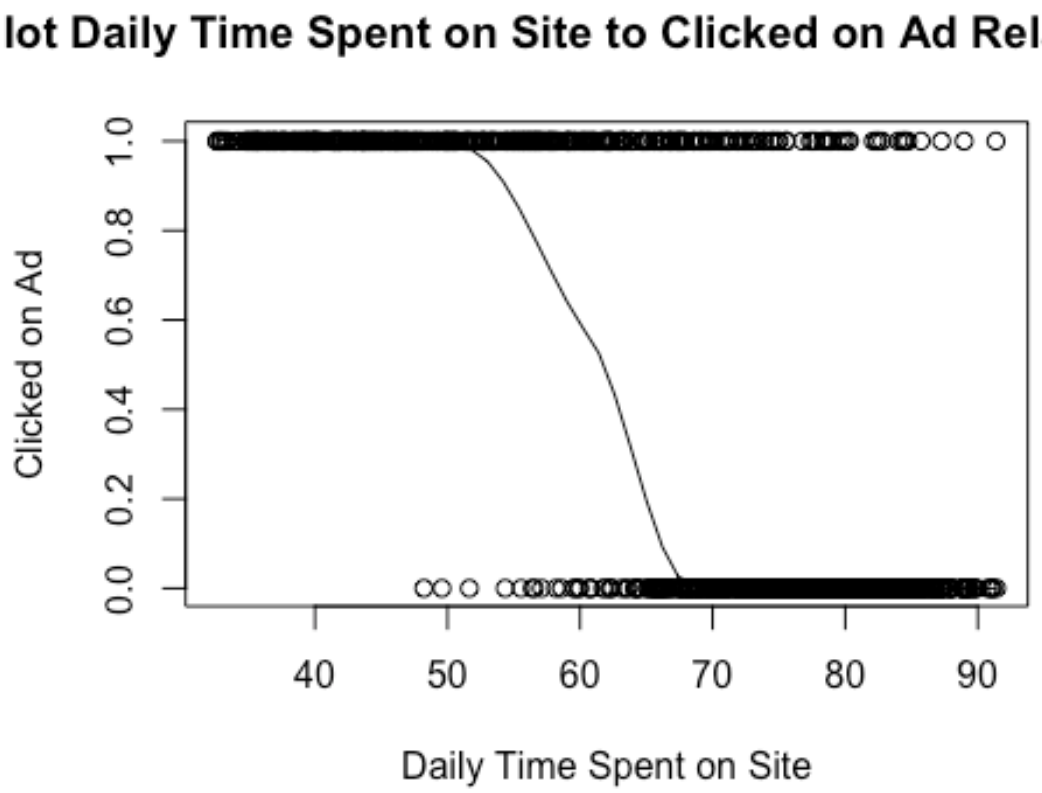
```
scatter.smooth(advertising$Male,advertising$`Clicked on Ad`,main="Plot Male to
Clicked on Ad Relationship",xlab = 'Male',ylab = 'Clicked on Ad')
```



#Gender is not a key determinant when it comes to clicking on the Ad both male and female have equal chances of clicking on the Ad.

#Checking clicking on Ad by Daily Time Spent on Site

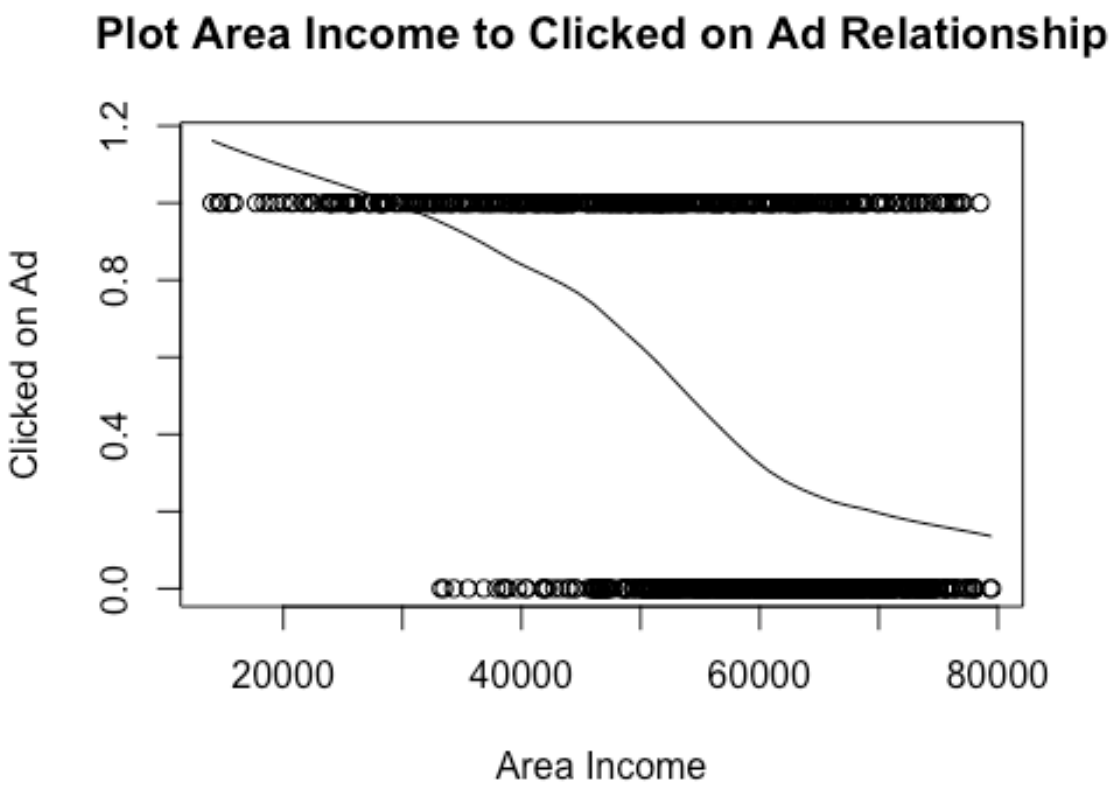
```
scatter.smooth(advertising$`Daily Time Spent on Site`,advertising$`Clicked on Ad`,main="Plot Daily Time Spent on Site to Clicked on Ad Relationship",xlab = 'Daily Time Spent on Site',ylab = 'Clicked on Ad')
```



#The lower the Daily time spent on Site the higher the chances of clicking on Ad.

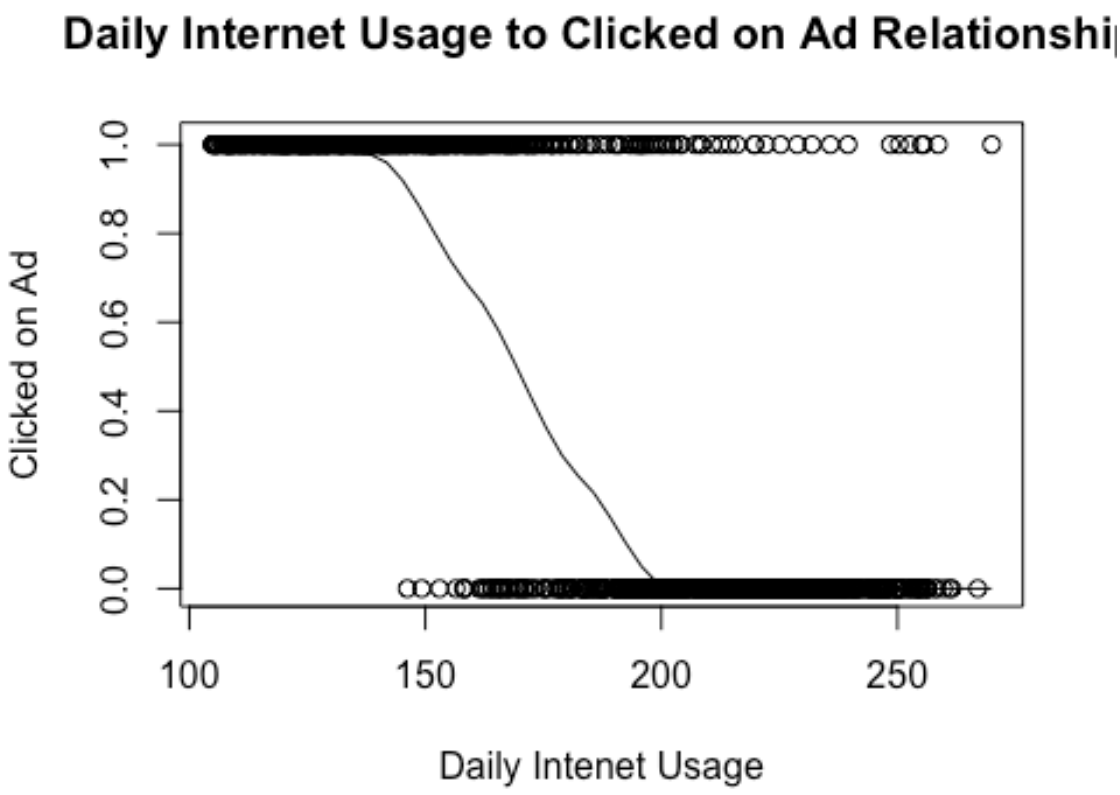
#Checking relationship clicking on Ad to Area Income

```
scatter.smooth(advertising$`Area Income`,advertising$`Clicked on Ad`,main="Plot Area Income to Clicked on Ad Relationship",xlab = 'Area Income',ylab = 'Clicked on Ad')
```



#The lower the Area Income the higher the chances of clicking on Ad.As income increases the probability of clicking on Ad declines.

```
scatter.smooth(advertising$`Daily Internet Usage`,advertising$`Clicked on
Ad`,main="Daily Internet Usage to Clicked on Ad Relationship",xlab = 'Daily
Intenet Usage',ylab = 'Clicked on Ad')
```



#We observe that the lower the daily internet Usage the higher the chances of clicking on the Ad.

#Checking the individuals who clicked the Ad by year

```
year.table <- table(advertising$'Clicked on Ad', advertising$year)
names(dimnames(year.table)) <- c("Clicked on Ad", "year")
year.table
```

```
##               year
## Clicked on Ad 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
21
##             0 11  9 12  1  7  9  8 10  9  3 13  4  5 13  6 11  8  9 12  8  7
13
##             1  9  9 14  5  7  4  8 14 11  7  9 11  9 11 12  4  7 11  8  7  5
8
##               year
## Clicked on Ad 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
43
##             0  7 11  4  7 11 10  7  7 15 10 10 12  9 14  9  9 10  8  6 11  7
```

```
4
##             1  9  8  9  8  5  7  6  2  4  8  8  9 10  5 12 11  7 12  7  9 10
10
##              year
## Clicked on Ad 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
##             0  3  8  7  8 10  7  8  8  6  5  7  9  6  9  9  4
##             1  8  8  7  5  8  8  9 10  9  9 12  5 10  8  8 10
```

#We observe that year 2002 and 2007 achieved the highest click on the Ad with 14 clicks

#Checking the those who clicked the Ad by month

```
month.table <- table(advertising$'Clicked on Ad', advertising$month)
names(dimnames(year.table)) <- c("Clicked on Ad", "month")
month.table

##
##     01 02 03 04 05 06 07
##   0 78 77 82 73 68 71 51
##   1 69 83 74 74 79 71 50
```

#We observe that most the highest number of click on the Ad were achieved in the month of February with 83 clicks followed by May with 79 clicks.

```
hour.table <- table(advertising$'Clicked on Ad', advertising$hour)
names(dimnames(hour.table)) <- c("Clicked on Ad", "hour")
hour.table

##              hour
## Clicked on Ad 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
21
##             0 19 16 19 19 21 23 16 28 22 21 17 16 22 21 22 16 23 18 16 20 26
29
##             1 26 16 17 23 21 21 23 26 21 28 14 24 16 21 21 19 16 23 25 19 24
19
##              hour
## Clicked on Ad 22 23
##             0 24 26
##             1 19 18
```

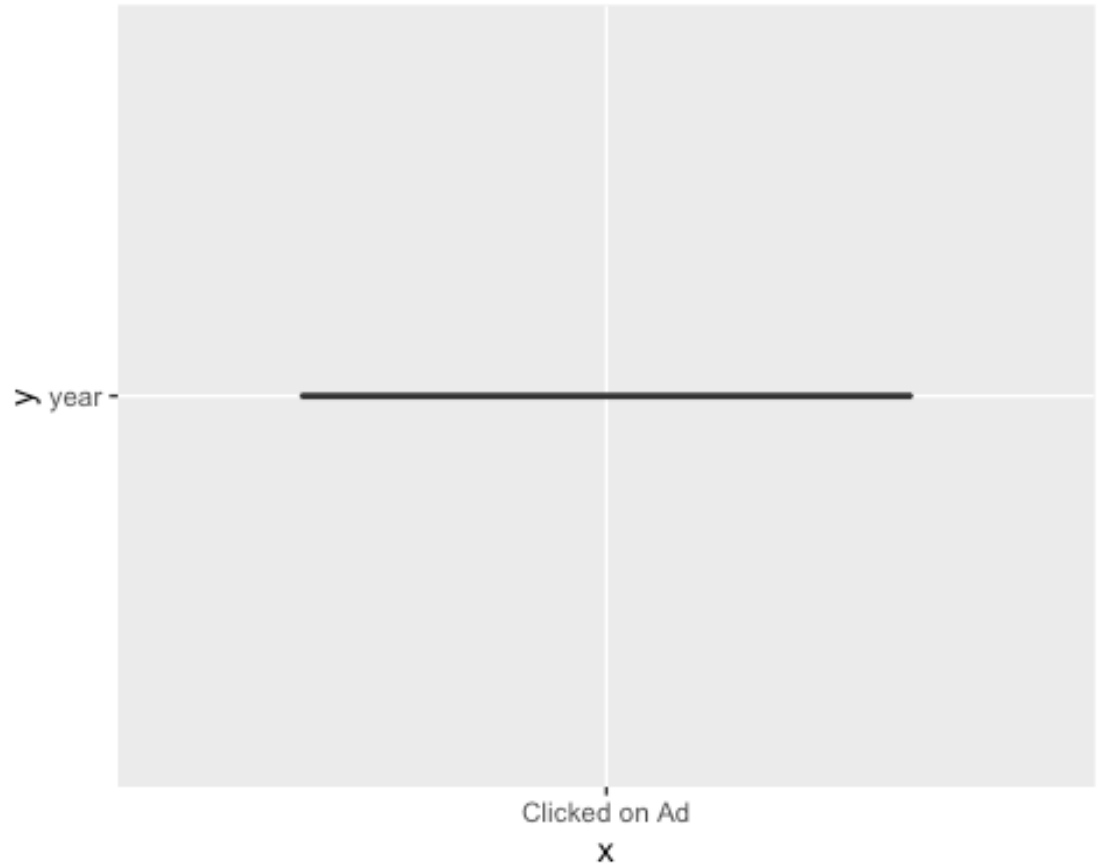
#We observe that the 9th hour achieved most clicks with 29clicks followed but the the 1st hour.

```
minute.table <- table(advertising$'Clicked on Ad', advertising$minute)
names(dimnames(minute.table)) <- c("Clicked on Ad", "minute")
minute.table
```

```
##              minute
## Clicked on Ad 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
21
##             0 11  9 12  1  7  9  8 10  9  3 13  4  5 13  6 11  8  9 12  8  7
13
##             1  9  9 14  5  7  4  8 14 11  7  9 11  9 11 12  4  7 11  8  7  5
8
##              minute
## Clicked on Ad 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
43
##             0  7 11  4  7 11 10  7  7 15 10 10 12  9 14  9  9 10  8  6 11  7
4
##             1  9  8  9  8  5  7  6  2  4  8  8  9 10  5 12 11  7 12  7  9 10
10
##              minute
## Clicked on Ad 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
##             0  3  8  7  8 10  7  8  8  6  5  7  9  6  9  9  4
##             1  8  8  7  5  8  8  9 10  9  9 12  5 10  8  8 10
```

#the 7th minute achieved the highest number of clicks on the Ad.

```
library(ggplot2)
ggplot(advertising, aes(x='Clicked on Ad', y='year')) + geom_boxplot()
```



4.Conclusion and Recommendation

Based on the findings from our analysis we conclude the following:

Gender is not really a key factor to influence clicking on the Ad or not, we established the number of female and those of male that clicked the Ad was more or less the same. So the Ad can target both genders.

We noted that Daily time spent on the site and Daily internet Usage is higher with younger individuals. These 2 variables decrease as age increases. Therefore age is a key factor to consider for success of the Ad. Therefore since we established that those who spend less time are likely to click on the Ad, we conclude that the Ad should target older individuals.

Income is key factor to consider. From our findings, those with lower are income have higher chances of clicking on the Ad than those with higher area income levels. The Ad should therefore target those with lower are income.

Those with lower Daily internet Usage and lower Daily time spent on site have higher chances of clicking on the Ad. The probability of clicking decreases with increase in the two variables. Therefore the individuals who spend less time and use lower internet should be targeted.

Year 2002 and 2007 achieved the highest clicks. This can be investigated further establish what was unique with these 2 years that can be applied into the future.

The month of February achieved moset click followed by the month of May, This could be contributed by May being a school holiday month and February could be as a result of it is not a very busy month, so individuals can afford the time.

The 9th hour and the 1st achieved most clicks, the entrepreneur can target scheduling these hours when placing the Ad.

Zchec republic and France are the top 2 countries that appreared more frequently these can and say 8 more can be targeted with the Ad. If the entrepreneur considers these factors, they will achieve better performance with getting more individuals clicking on the Ad.