Week12_IP

Stacy Mwende

#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. 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.

#Defining the question Identify the target group in this case, which individuals who are likely to click on the ad

#Metric for success Get the best variables that should be considered while posting the add to be enable maximum views which will translate to more people enrolling for the course.

#Experimental Design I undertook three steps in my project: Data Understanding Business Understanding Analysis Conclusion and Recommendation

#Loading the dataset ##Get the directory for the dataset

```
getwd()

## [1] "C:/Users/comp5/Downloads/R_Program"
```

###Reading the dataset

```
advertise <- read.csv("advertising.csv")
head(advertise)

##   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
```

#Getting more information about the dataset

```
nrow(advertise)

## [1] 1000
```

The dataset has 1000 entries

```
NCOL(advertise)

## [1] 10
```

The dataset has a total of 10 columns

#Getting the Datatypes of the variables

```
sapply(advertise, class)

## Daily.Time.Spent.on.Site                      Age              Area.Income 
##                "numeric"                "integer"                "numeric" 
##     Daily.Internet.Usage            Ad.Topic.Line                     City 
##                "numeric"                 "factor"                 "factor" 
##                     Male                  Country                Timestamp 
##                "integer"                 "factor"                 "factor" 
##            Clicked.on.Ad 
##                "integer"
```

#Data Cleaning ##Checking for null values

```
sum(is.na(advertise))

## [1] 0
```

There is no missing data in the dataset provided for analysis

##Checking for duplicates

```
duplicated_adv <- advertise[duplicated(advertise),]
duplicated_adv

##  [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           
## <0 rows> (or 0-length row.names)
```

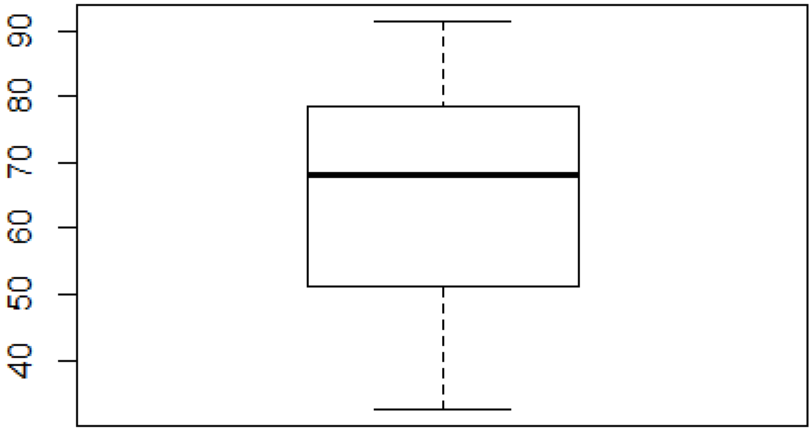
There are no duplicates in the dataset

##Converting datatypes

```
advertise$Timestamp <- as.Date(advertise$Timestamp, format= "%Y-%m-%s-%h-%m-
%s")
```

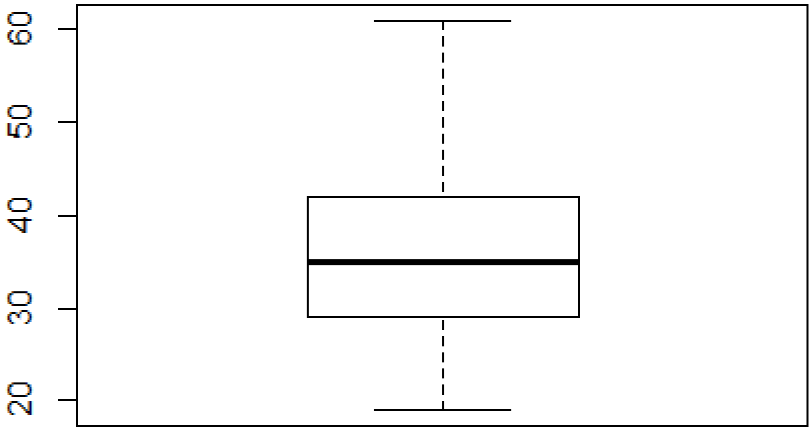
##Checking for outliers

```
OutVals = boxplot(advertise$
Daily.Time.Spent.on.Site)
```



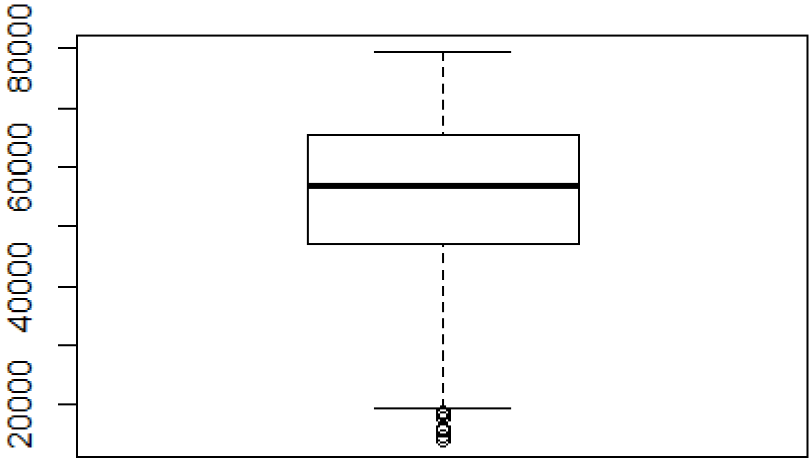
There are no outliers in the Daily.Time.Spent.on.Site column

```
OutVals = boxplot(advertise$Age)
```



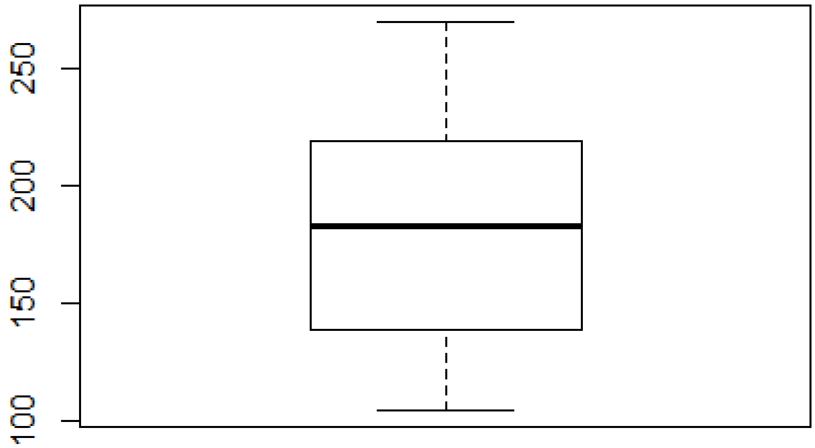
No outliers in the Age column

```
OutVals = boxplot(advertise$Area.Income)
```



There are outliers in the Area.Income column.

```r
OutVals = boxplot(advertise$
Daily.Internet.Usage)
```



There are no outliers in the Daily internet usage column

#Correcting wrongly labeled columns Get the column names

```r
colnames(advertise)

##  [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"
```

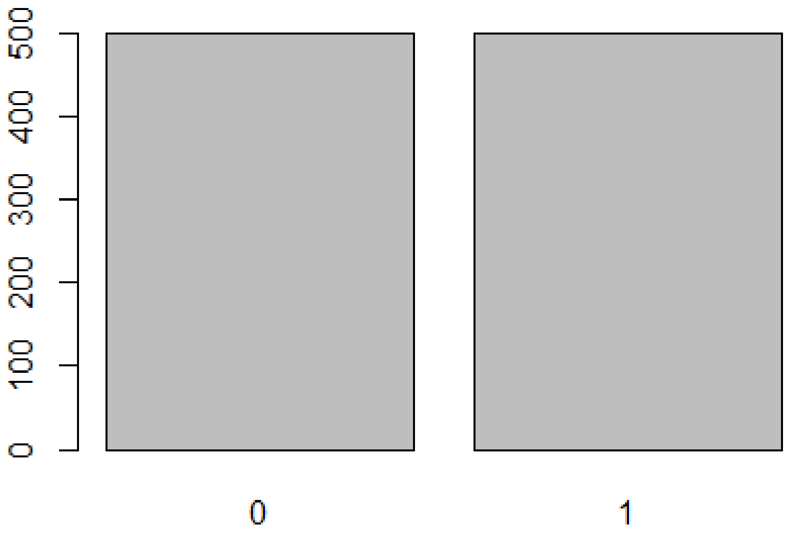
The column labelled Male should be Gender which will consist of Male and Female

```r
names(advertise)[names(advertise) == "Male"] <- "Gender"

colnames(advertise)

##  [1] "Daily.Time.Spent.on.Site" "Age"
##  [3] "Area.Income"              "Daily.Internet.Usage"
##  [5] "Ad.Topic.Line"            "City"
##  [7] "Gender"                   "Country"
##  [9] "Timestamp"                "Clicked.on.Ad"
```

The column has been renamed to Gender as observed above

Checking if the data is balanced We consider our target variable i.e Clicked.on.Ad

```
Clicked.on.Ad_freq <- table(advertise$Clicked.on.Ad)
barplot(Clicked.on.Ad_freq)
```



For the visualization above, the data is balanced in that equal number of people viewed and also an equal number did not view the Ad.

#Getting a Summary of the dataset

```
summary(advertise)

##  Daily.Time.Spent.on.Site      Age         Area.Income
##  Min.   :32.60            Min.   :19.00   Min.   :13996
##  1st Qu.:51.36            1st Qu.:29.00   1st Qu.:47032
##  Median :68.22            Median :35.00   Median :57012
##  Mean   :65.00            Mean   :36.01   Mean   :55000
##  3rd Qu.:78.55            3rd Qu.:42.00   3rd Qu.:65471
##  Max.   :91.43            Max.   :61.00   Max.   :79485
##
##  Daily.Internet.Usage                                 Ad.Topic.Line
##  Min.   :104.8        Adaptive 24hour Graphic Interface      :  1
##  1st Qu.:138.8        Adaptive asynchronous attitude         :  1
##  Median :183.1        Adaptive context-sensitive application :  1
##  Mean   :180.0        Adaptive contextually-based methodology:  1
##  3rd Qu.:218.8        Adaptive demand-driven knowledgebase   :  1
##  Max.   :270.0        Adaptive uniform capability            :  1
##                       (Other)                                :994
##                City          Gender                Country       Timestamp
##  Lisamouth      :  3   Min.   :0.000   Czech Republic:  9   Min.   :NA
##  Williamsport   :  3   1st Qu.:0.000   France        :  9   1st Qu.:NA
##  Benjaminchester:  2   Median :0.000   Afghanistan   :  8   Median :NA
##  East John      :  2   Mean   :0.481   Australia     :  8   Mean   :NA
##  East Timothy   :  2   3rd Qu.:1.000   Cyprus        :  8   3rd Qu.:NA
##  Johnstad       :  2   Max.   :1.000   Greece        :  8   Max.   :NA
##  (Other)        :986                   (Other)       :950   NA's   :1000
##  Clicked.on.Ad
##  Min.   :0.0
##  1st Qu.:0.0
##  Median :0.5
##  Mean   :0.5
##  3rd Qu.:1.0
##  Max.   :1.0
##
```

The summary code gives us a couple of outputs including Minimum,1st quartile,Median,Mean,3rd Quartile and the Maximum value. The different variables are distributed with the above stated output.

Checking the country that had more people visit the site

```
country_tbl <- advertise$Country
names(table(country_tbl))[table (country_tbl)==max(table(country_tbl))]

## [1] "Czech Republic" "France"
```

Czech Republic, France topped the list with most people who visited the site

#Performing Exploratory Data Analysis ##Univariate Analysis Covert the dataframe to tibble for easier analysis

```
library("tibble")
my_data <- as_tibble(advertise)
class(my_data)

## [1] "tbl_df"     "tbl"        "data.frame"
```

###Measures of Central Tendency 1.*Mean*

```
mean(my_data$Daily.Time.Spent.on.Site)

## [1] 65.0002
```

People avearagely spent 65 minutes per day on the the site

```
mean(my_data$Area.Income)

## [1] 55000
```

Average area income is 55,000 from the considered dataset

```
mean(my_data$Daily.Internet.Usage)

## [1] 180.0001
```

Most people used an average of 180MBs per day on the very website in consideration

```
library(dplyr)

##
## Attaching package: 'dplyr'

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

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

library(pander)
my_data %>%
    summarize(variable = "Age", mean_age = mean(Age), st_dev_age = sd(Age))
%>%
    pander
```

variable	mean_age	st_dev_age
Age	36.01	8.786

Mean age of the people who visited the site was 36 years of age with a spread of 8 from the mean

Getting a quartile distribution of the Age column

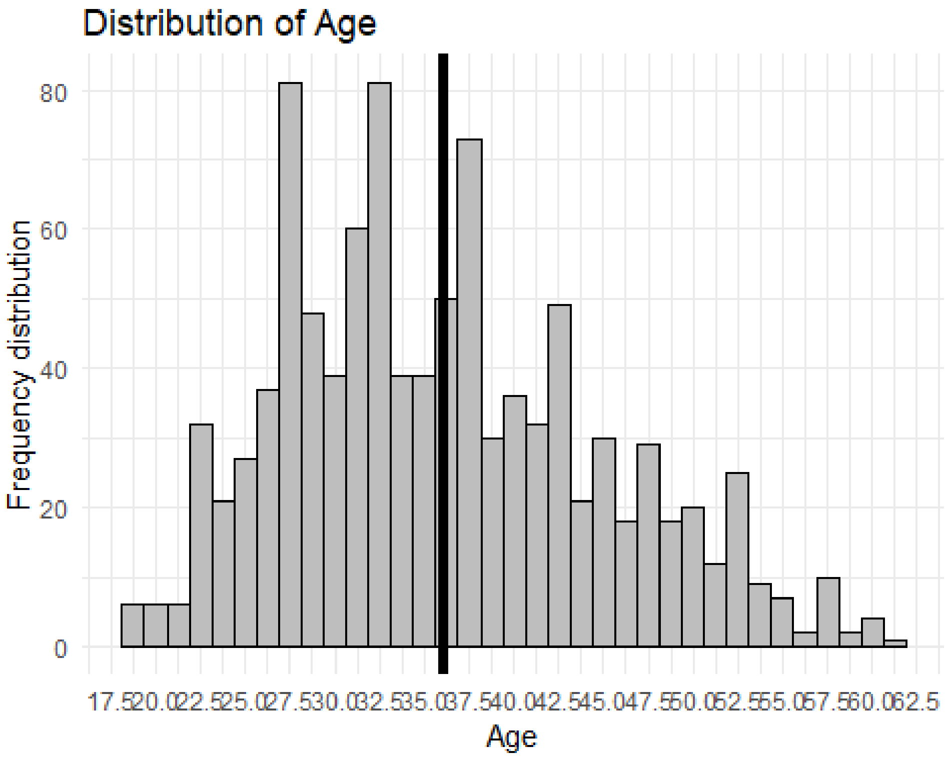
```
my_data %>%
    summarize(variable = "Age",
              q0.25 = quantile(Age, 0.25),
              q0.50 = quantile(Age, 0.50),
              q0.75 = quantile(Age, 0.75),
              q1 = quantile(Age, 1)) %>%
    pander
```

variable	q0.25	q0.50	q0.75	q1
Age	29	35	42	61

The above distribution of quartiles shows the range of the age column showing that the least age of the people who visited the site was 29 years and the max age in that case was 61 years.

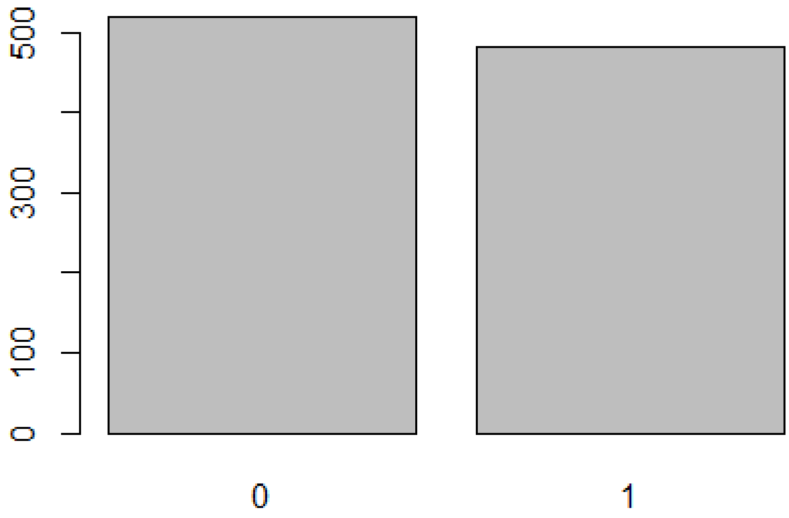
Visualizing the above age column in a ggplot as shown below

```
library("ggplot2")
library(dplyr)

my_data %>%
    ggplot(aes(Age)) +
    geom_histogram(binwidth = 1.25, color = "black",fill = "grey") +
    geom_vline(xintercept = mean(my_data$Age), lwd = 2) +
    labs(title = "Distribution of Age",
         x = "Age",
         y = "Frequency distribution") +
    theme_minimal() +
    scale_x_continuous(breaks = seq(7.5,100,2.5))
```



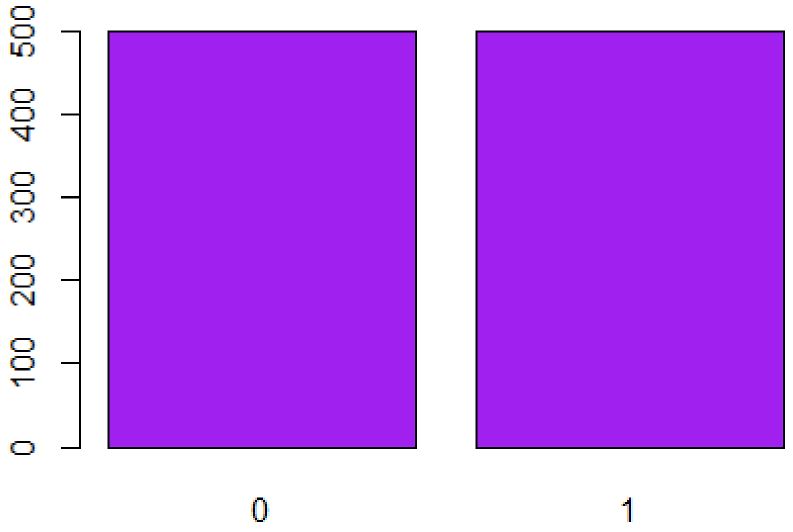
###Bar Graph *Gender Column*

```
Gender <- my_data$Gender
Gender_freq <- table(Gender)
barplot(Gender_freq)
```



From the visualization above, more females(0) visited the site compare to male counterparts(1)

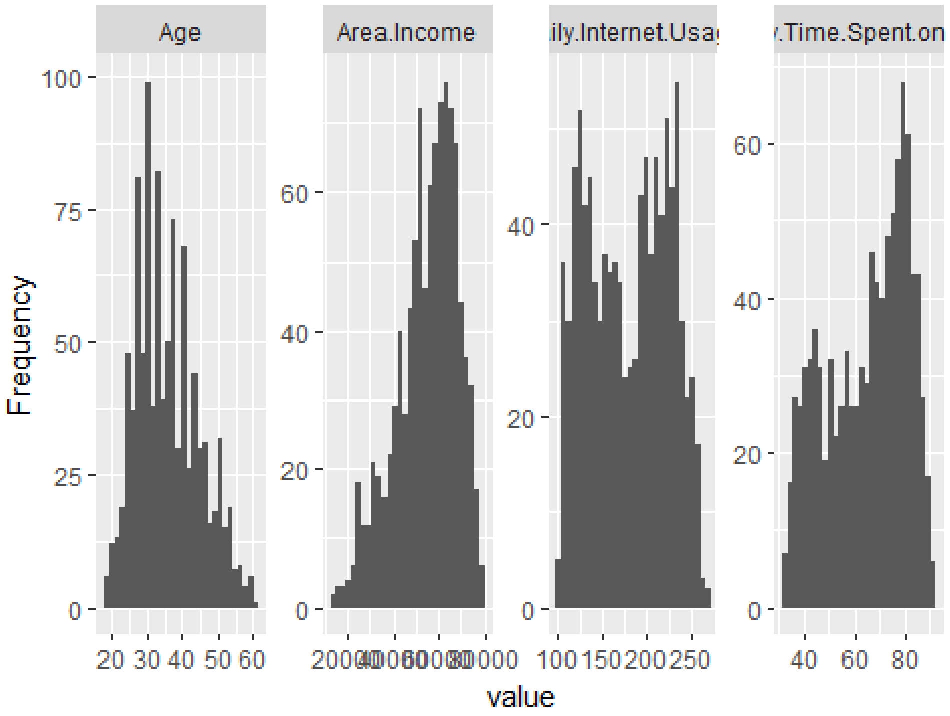
```
Clicked.on.Ad <- my_data$Clicked.on.Ad
Clicked.on.Ad_freq <- table(Clicked.on.Ad)
barplot(Clicked.on.Ad_freq, col = "purple")
```



The visualization above shows that there was equal distribution of the people who either viewed or did not view on the add.

###Histogram

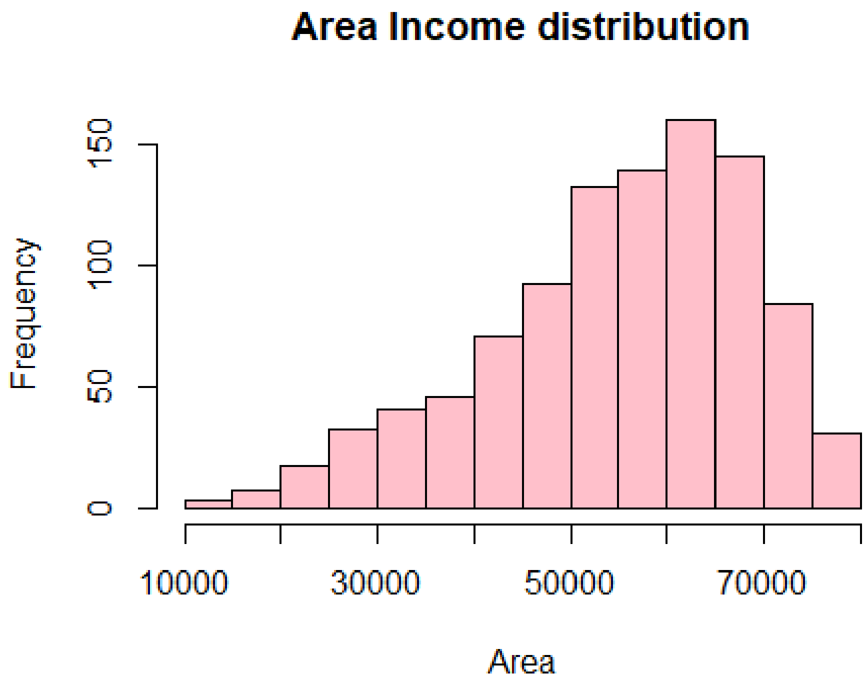
```
library(DataExplorer)
plot_histogram(my_data)
```



Age column is skewed to the left with most people who have visited the site being of age 25-45 years Area income is skwed to the right. Daily internet usage is related to the daily time spent on the site even from the visualization above

###AreaIncome Column

```
z = hist(my_data$Area.Income,
         main = "Area Income distribution",
         xlab = "Area",
         col = "pink"
         )
```



##Bivariate Analysis *Covariance* Covariance is a number that reflects the degree to which two variable vary together **Daily.Time.Spent.on.Site Vs Age**

```
Daily.Time.Spent.on.Site <- my_data$Daily.Time.Spent.on.Site
Age <- my_data$Age
cov(Daily.Time.Spent.on.Site, Age)

## [1] -46.17415
```

The above value indicates that there is a negative correlation between the 2 variables. This means that a change in either variable leads to a decrease in the other.

Age Vs Gender

```
Age <- my_data$Age
Gender <- my_data$Gender
cov(Age, Gender)

## [1] -0.09242142
```

There is a negative relation between the 2 variables though not so big.

Clicked.on.Ad Vs Age

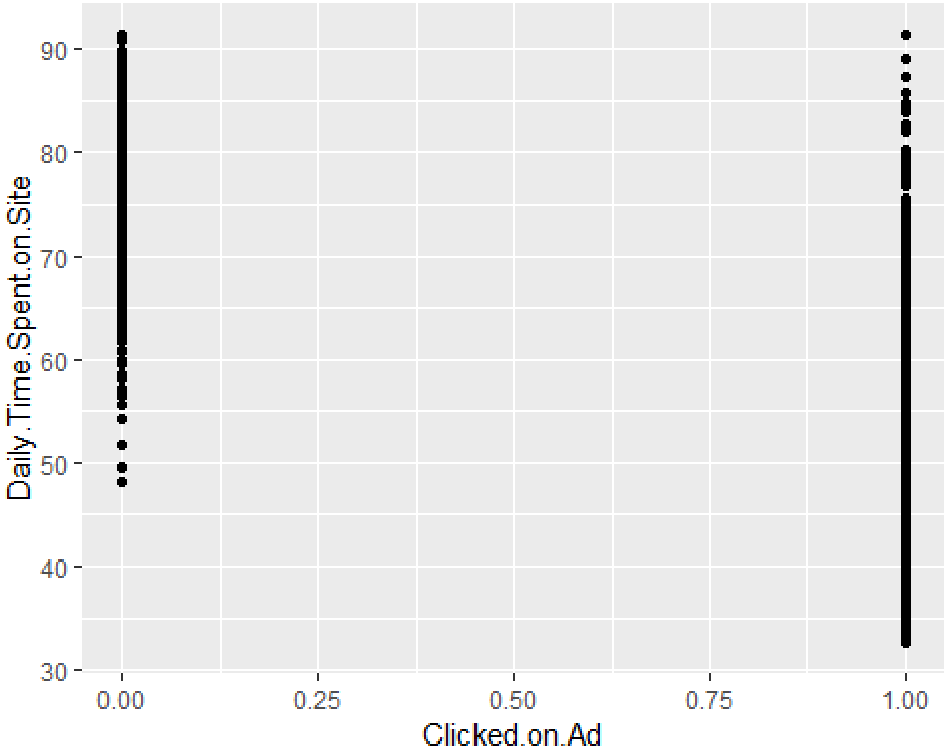
```
Clicked.on.Ad <- my_data$Clicked.on.Ad
Age <- my_data$Age
cov(Clicked.on.Ad, Age)
```

```
## [1] 2.164665
```

There is a positive relation between the 2 variables, in that an increase in one would increase the value of the

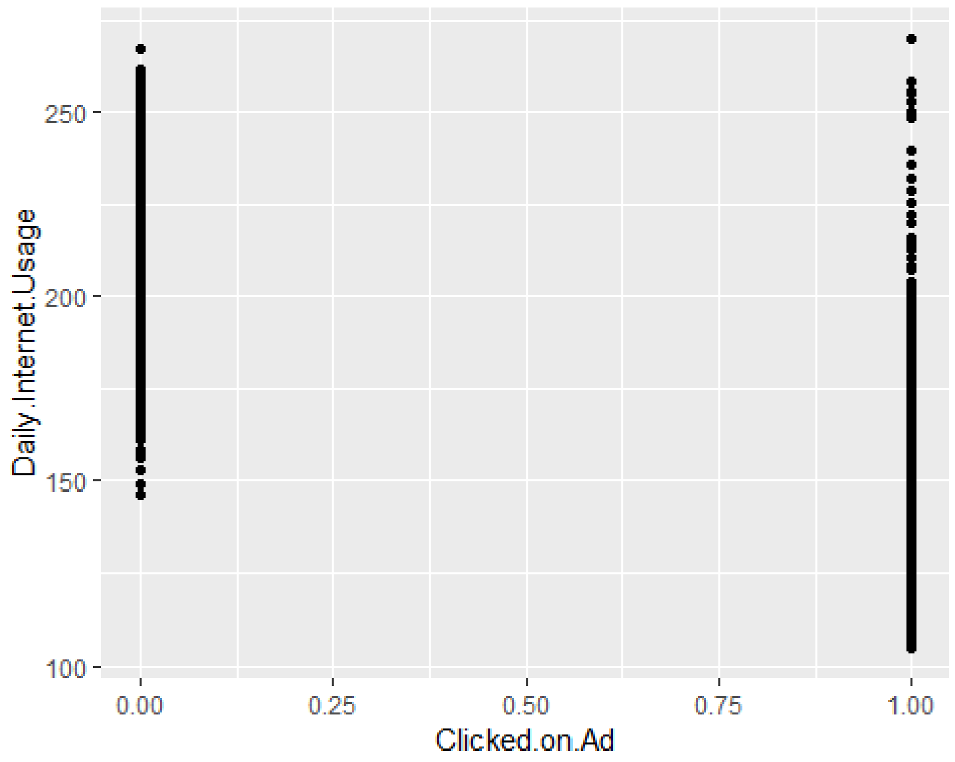
DailyTimeSpentOnSite Vs ClickedOnAd

```
library(ggplot2)
ggplot(my_data, aes(x = Clicked.on.Ad, y = Daily.Time.Spent.on.Site)) +
    geom_point()
```

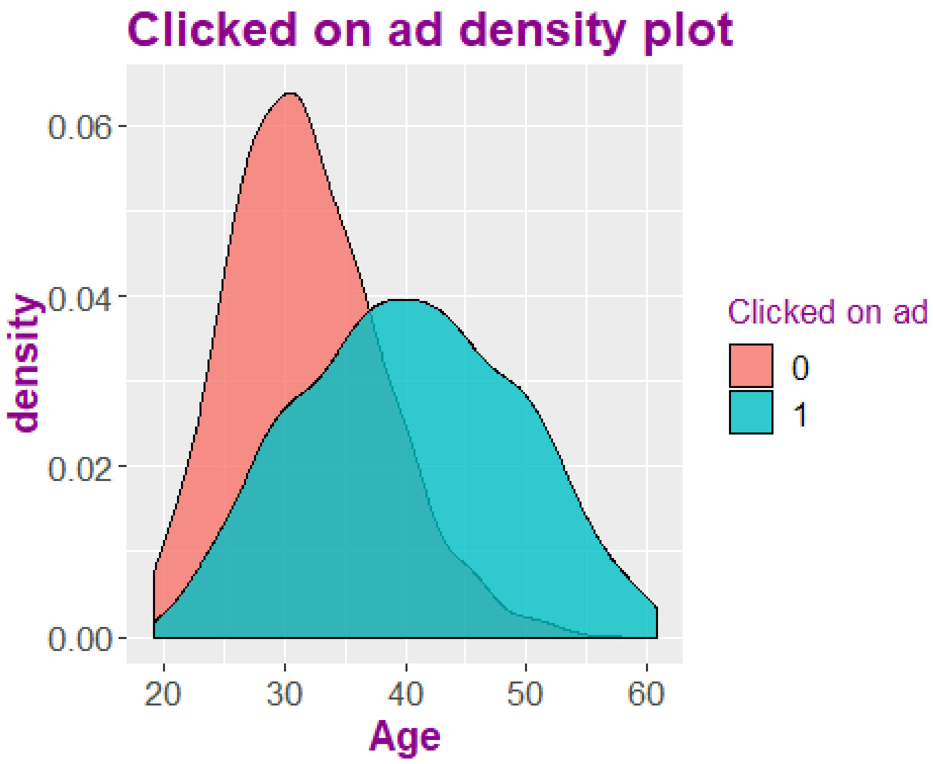


The above graph shows that ClickedOnAd is more distributed than DailyTimeSpent OnSite which means that spending more time on the site does not neccesarily mean that they will click on the Ad

```
library(ggplot2)
ggplot(my_data, aes(x = Clicked.on.Ad, y = Daily.Internet.Usage)) +
    geom_point()
```



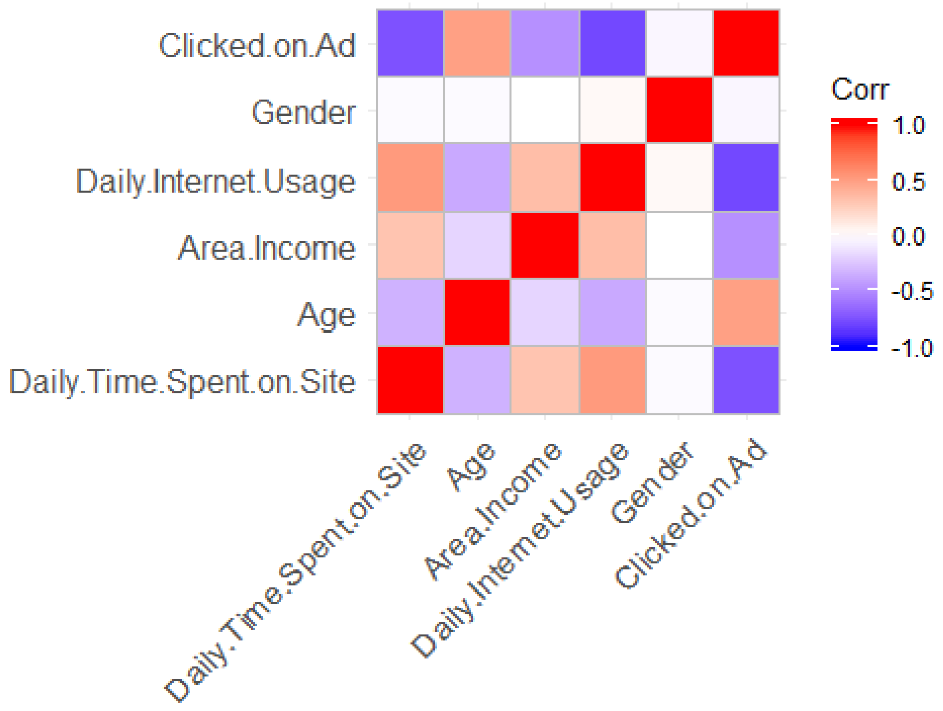
```
library(ggplot2)
options(repr.plot.width = 13, repr.plot.height = 7)
R2 = ggplot(data = my_data, aes(Age)) +
        geom_density(aes(fill=factor(Clicked.on.Ad)), alpha = 0.8) +
        labs(title = 'Clicked on ad density plot', x = 'Age', fill = 'Clicked on ad') +
        scale_color_brewer(palette = 1) +
        theme(plot.title = element_text(size = 18, face = 'bold', color = 'darkmagenta'),
             axis.title.x = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.title.y = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.text.x = element_text(size = 13, angle = 0),
             axis.text.y = element_text(size = 13),
             legend.title = element_text(size = 13, color = 'darkmagenta'),
             legend.text = element_text(size = 12))

plot(R2)
```



0 means that they did not click on the Ad with 1 indicating they clicked on the Ad The distribution of the people who clicked on the Ad is normally distributed compared to the people who did not clicl on the Ad which is skwed to the left.

#Multivariate Analysis

```
library(ggcorrplot)
my_data %>%
    select_if(is.numeric) %>%
    cor %>%
    ggcorrplot()
```



There is a high correlation between a variable and it's self which is normal There is also a highly relative correlation between DailyTimeSpentOnSite and the DailyInternetUsage We also observe a correlatio between Age and ClickedOnAd column There is no correlation between Gender and Age

#Conclusion I can conclude the variables provided for analysis are fit to answer the question inplace. I also observed that the type of Ad topic determined on who clicked on it.

#Recommendation From the visualization above, the entrepreneur should consider being diverse in the different topics posted on her channel since all ages are involved. From the Area income distribution, the enterpreneur should consider enrolling people from the high end countries with more income. The enterpreneur should have interesting Ad topics to attract someone to click on the Ad.