output: - default - default —

1.INTRODUCTION

a) Defining the Question

To identify the individuals who are most likely to click on the ad.

b) Defining the metric of success

Finding the audience who are going to be interested in the product advertised.

c) Understanding the Context

By looking at the history of advertisement, we are going to examine the market and get knowledge of the target audience and how to target them.

d) Recording the experimental design

Data preparation and cleaning; • Loading libraries and data table • Check for missing values and duplicates • Check for outliers and anomalies

Performing Exploratory Data Analysis; • Uni variate Analysis • Bivariate Analysis

Conclusions Recommendation

2. DATA PREPARATION AND CLEANING

#loading our dataset

```
data <- read.csv('http://bit.ly/IPAdvertisingData')
head(data)
```

```
##   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 dataset structure

```
str(data)
```

```
## '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 standardi
##  $ 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"
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

rename column names to a uniform case.

```
names(data)[names(data) == "Ad.Topic.Line"] <- "ad_topic_line"
names(data)[names(data) == "City"] <- "city"
names(data)[names(data) == "Male"] <- "male"
names(data)[names(data) == "Country"] <- "country"
names(data)[names(data) == "Timestamp"] <- "timestamp"
names(data)[names(data) == "Clicked.on.Ad"] <- "clicked_on_ad"
names(data)[names(data) == "Daily.Time.Spent.on.Site"] <- "daily_time_spent"
names(data)[names(data) == "Age"] <- "age"
names(data)[names(data) == "Area.Income"] <- "area_income"
names(data)[names(data) == "Daily.Internet.Usage"] <- "daily_internet_usage"
```

#lets review our data to see the changes. #Its established changes have been made.

```
head(data)
```

```
##   daily_time_spent 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 for missing values

```
colSums(is.na(data))
```

```
##     daily_time_spent                  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

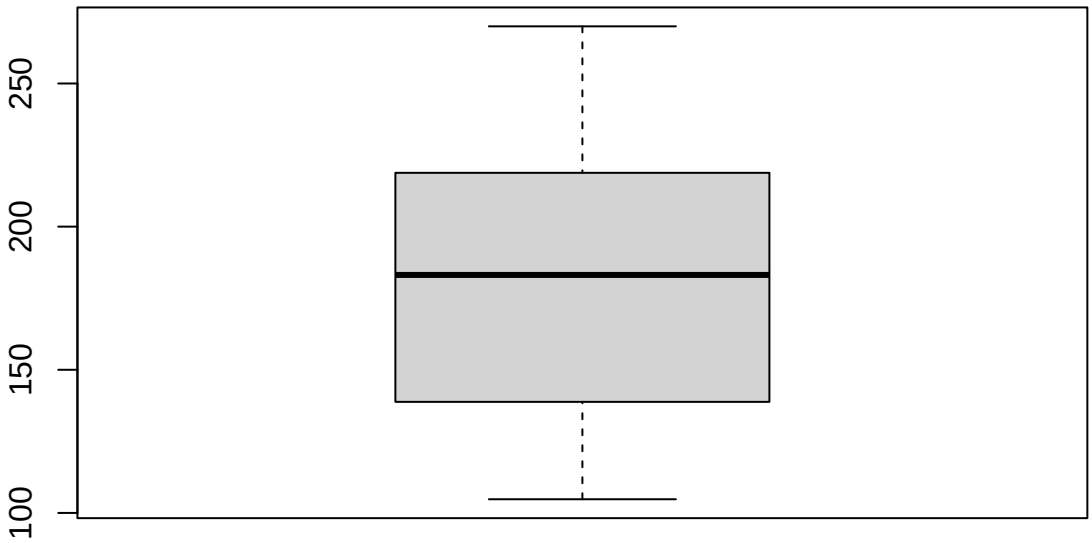
```
anyDuplicated(data)
```

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

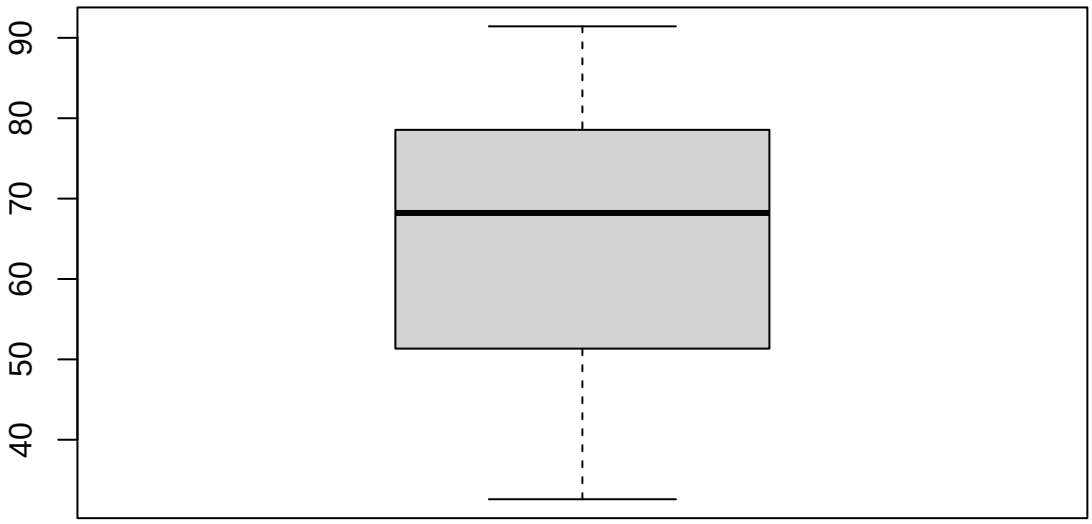
#There are no duplicates

checking for outliers in our numerical values

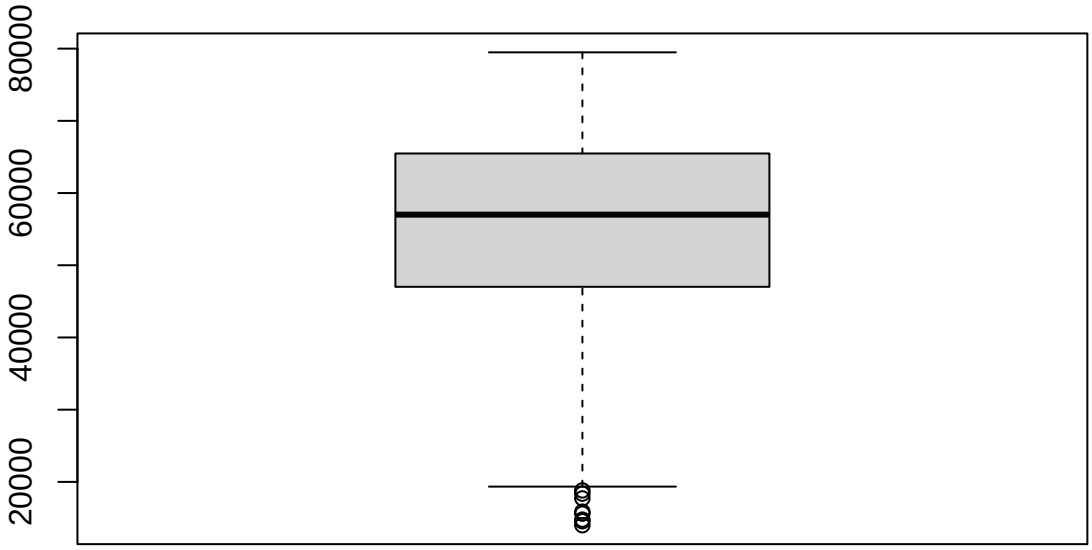
```
boxplot(data$daily_internet_usage)
```



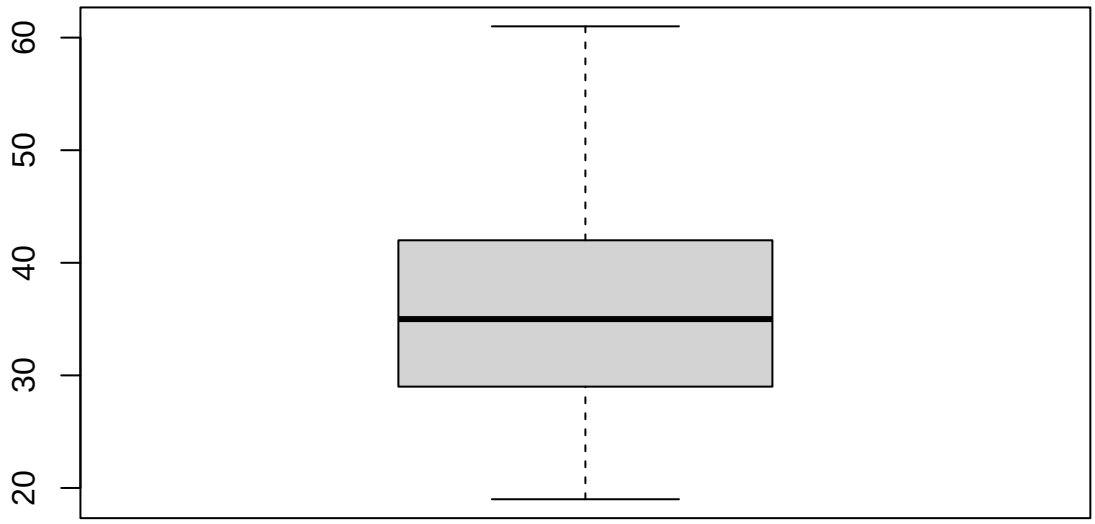

```
boxplot(data$daily_time_spent)
```




```
boxplot(data$area_income)
```




```
boxplot(data$age)
```



#there are no outliers in the variables except in area income but we will keep them because the data is true, income will vary for everyone.

3.EXPLORATORY DATA ANALYSIS

Univariate analysis

We are going to look at variable distribution of our data by analysing the min,max,mean,median and quartile distributions of the variables.

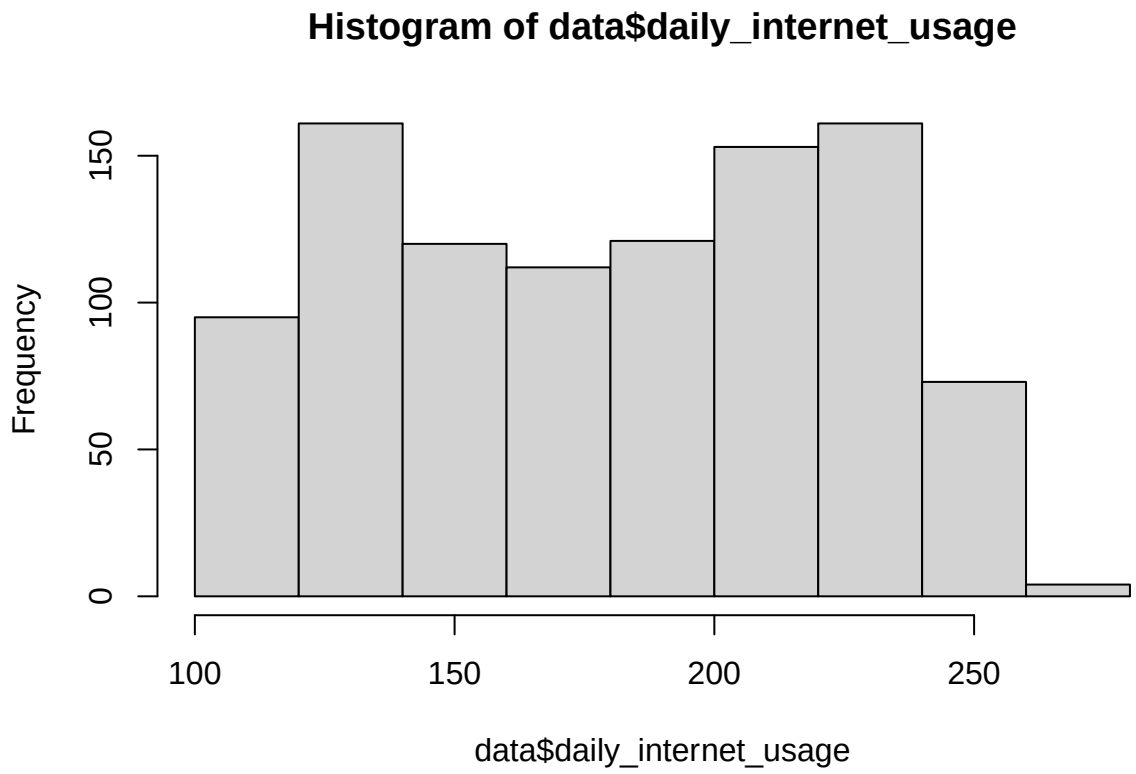
Getting the minimum, maximum, mean,median and quartiles for the variable daily_internet_usage

```
summary(data$daily_internet_usage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.8   138.8   183.1   180.0   218.8   270.0
```

checking the distribution

```
hist(data$daily_internet_usage)
```



#variance

```
var(data$daily_internet_usage)
```

```
## [1] 1927.415
```

#standard deviation

```
sd(data$daily_internet_usage)
```

```
## [1] 43.90234
```

#interquartile range

```
quantile(data$daily_internet_usage, 0.75) - quantile(data$daily_internet_usage, 0.25)
```

```
##     75%
## 79.9625
```

#installing package ‘moments’

```
library(moments)
```

finding the kurtosis

```
kurtosis(data$daily_internet_usage)
```

```
## [1] 1.727701
```

#checking for skewness

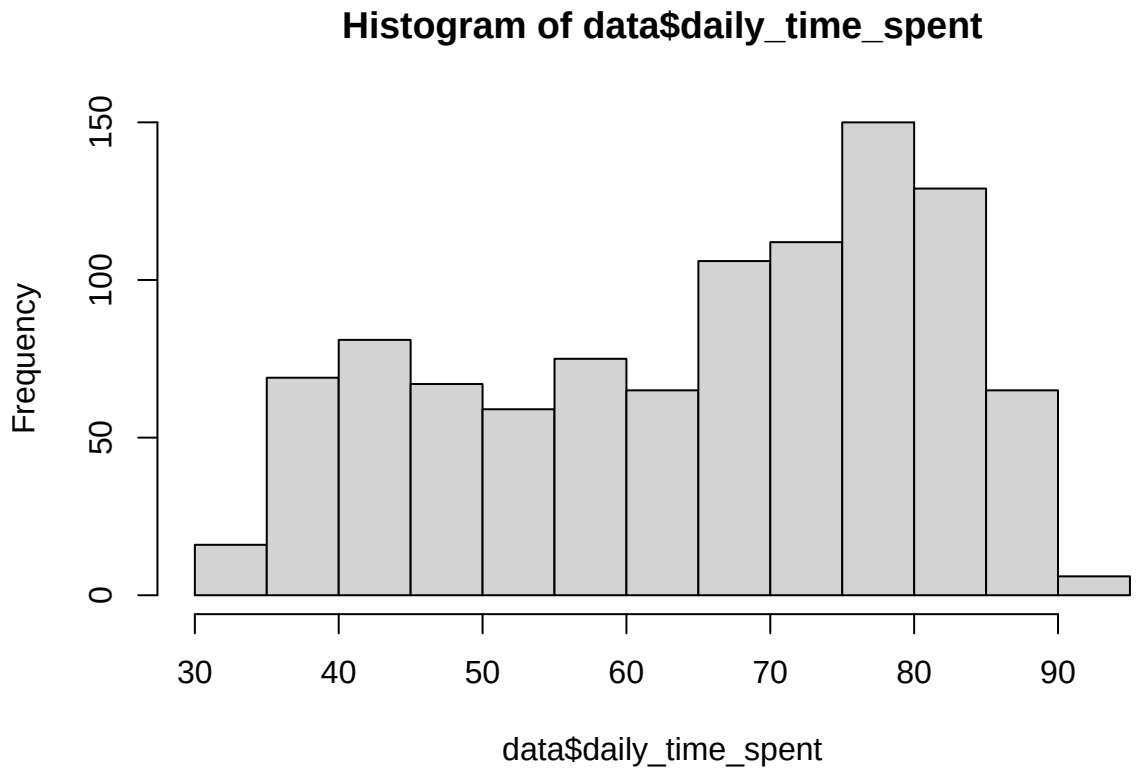
```
skewness(data$daily_internet_usage)
```

```
## [1] -0.03348703
```

#Distribution, Variance, Standard deviation,range,kurtosis,skewness.

checking the distribution

```
hist(data$daily_time_spent)
```



checking for variance

```
var(data$daily_time_spent)
```

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

getting standard deviation

```
sd(data$daily_time_spent)
```

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

checking for skewness

```
skewness(data$daily_time_spent)
```

```
## [1] -0.3712026
```

checking kurtosis

```
kurtosis(data$daily_time_spent)
```

```
## [1] 1.903942
```

Checking for different frequencies on our variables.

checking on the difference in people who clicked the ad and none

)

```
ad_column <- table(data$clicked_on_ad)
ad_column
```

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

most occuring cities

```
library(plyr)
count_city <- count(data$city)
count_city_head <- head(arrange(count_city, desc(freq)))
count_city_head
```

```
##                 x freq
## 1       Lisamouth    3
## 2    Williamsport    3
## 3 Benjaminchester    2
## 4       East John    2
## 5    East Timothy    2
## 6        Johnstad    2
```

most occuring countries

```
count_country <- count(data$country)
count_country_head <- head(arrange(count_country, desc(freq)))
count_country_head
```

```
##                x freq
## 1 Czech Republic    9
## 2         France    9
## 3    Afghanistan    8
## 4      Australia    8
## 5         Cyprus    8
## 6         Greece    8
```

Bivariate analysis

Selecting our numerical variables to check the correlation.

```
numerical <- data[,1:4]
numerical <- cbind(numerical, data[c('male')])
head(numerical)
```

```
##   daily_time_spent age area_income daily_internet_usage male
## 1            68.95  35    61833.90                256.09    0
## 2            80.23  31    68441.85                193.77    1
## 3            69.47  26    59785.94                236.50    0
## 4            74.15  29    54806.18                245.89    1
## 5            68.37  35    73889.99                225.58    0
## 6            59.99  23    59761.56                226.74    1
```

Creating a correlation matrix

```
numerical.cor=cor(numerical,method=c('pearson'))
numerical.cor
```

```
##                       daily_time_spent         age  area_income
## daily_time_spent           1.00000000 -0.33151334  0.310954413
## age                       -0.33151334  1.00000000 -0.182604955
## area_income                0.31095441 -0.18260496  1.000000000
## daily_internet_usage       0.51865848 -0.36720856  0.337495533
## male                      -0.01895085 -0.02104406  0.001322359
##                       daily_internet_usage         male
## daily_time_spent               0.51865848 -0.018950855
## age                           -0.36720856 -0.021044064
## area_income                    0.33749553  0.001322359
## daily_internet_usage           1.00000000  0.028012326
## male                           0.02801233  1.000000000
```

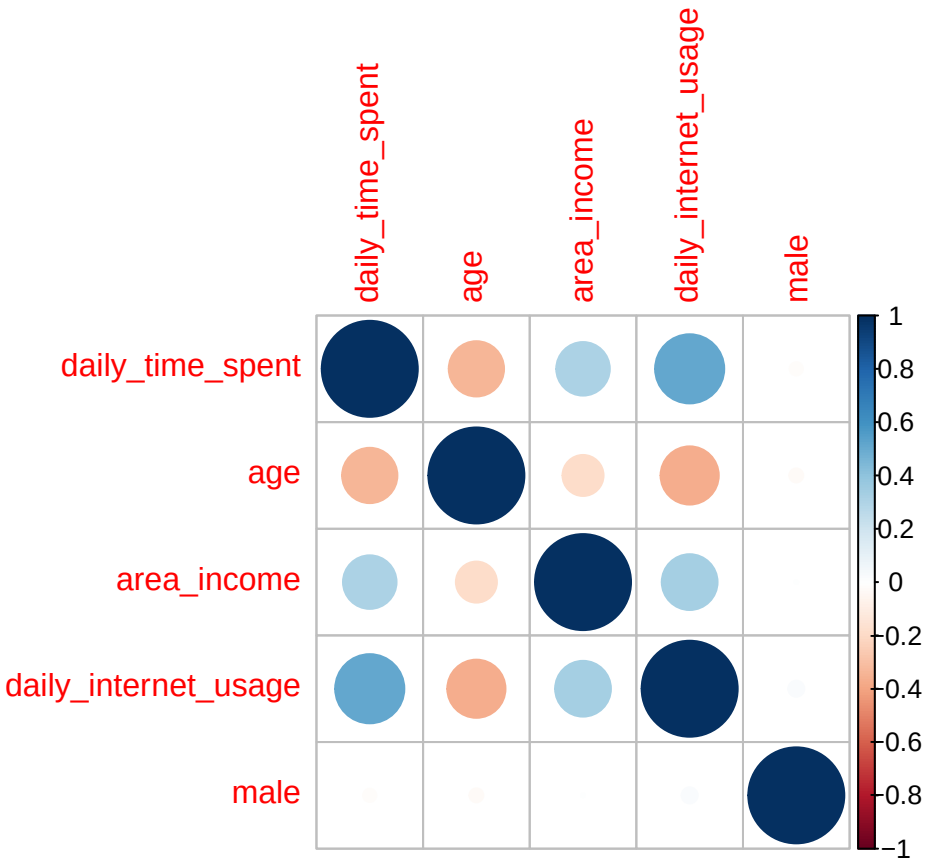
Installing the correlation plot to visualize the correlation coefficients.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

#visualization

```
corrplot(numerical.cor)
```

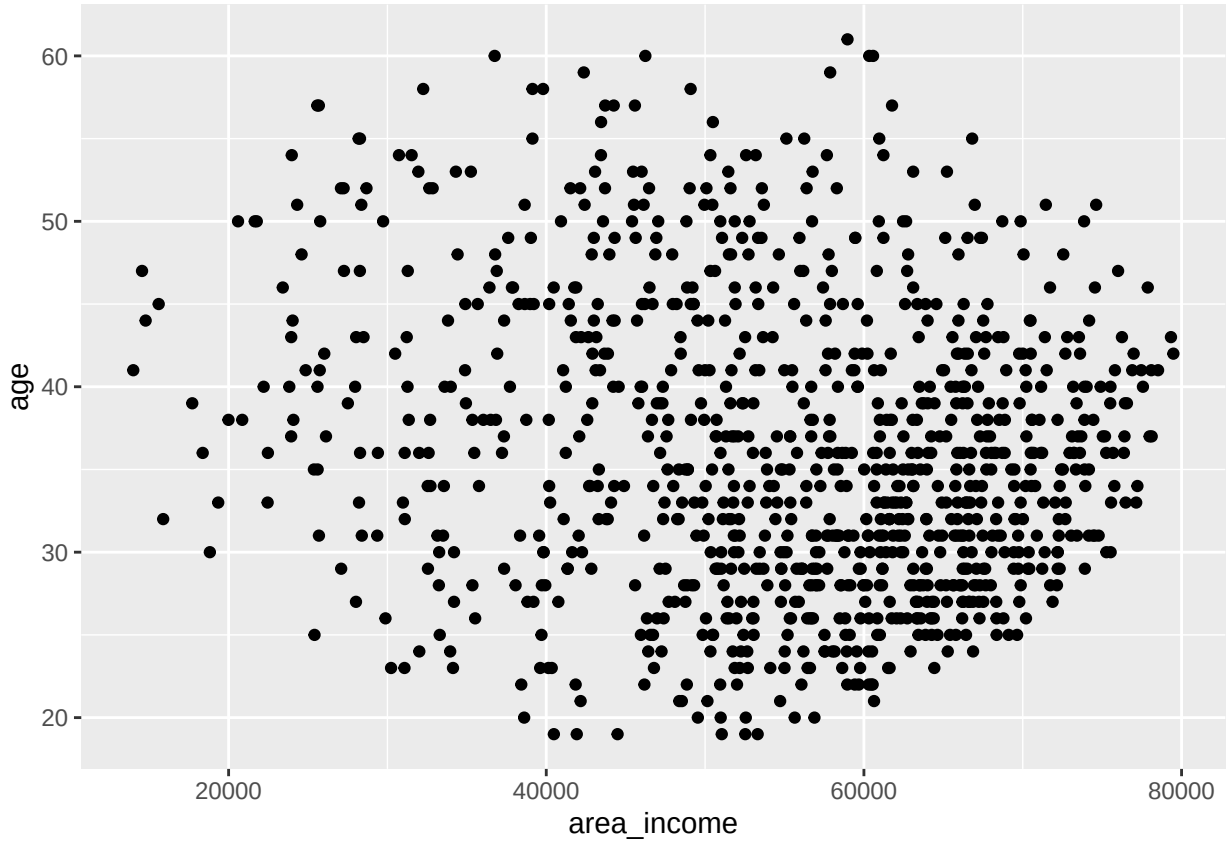


importing library

```
library(ggplot2)
```

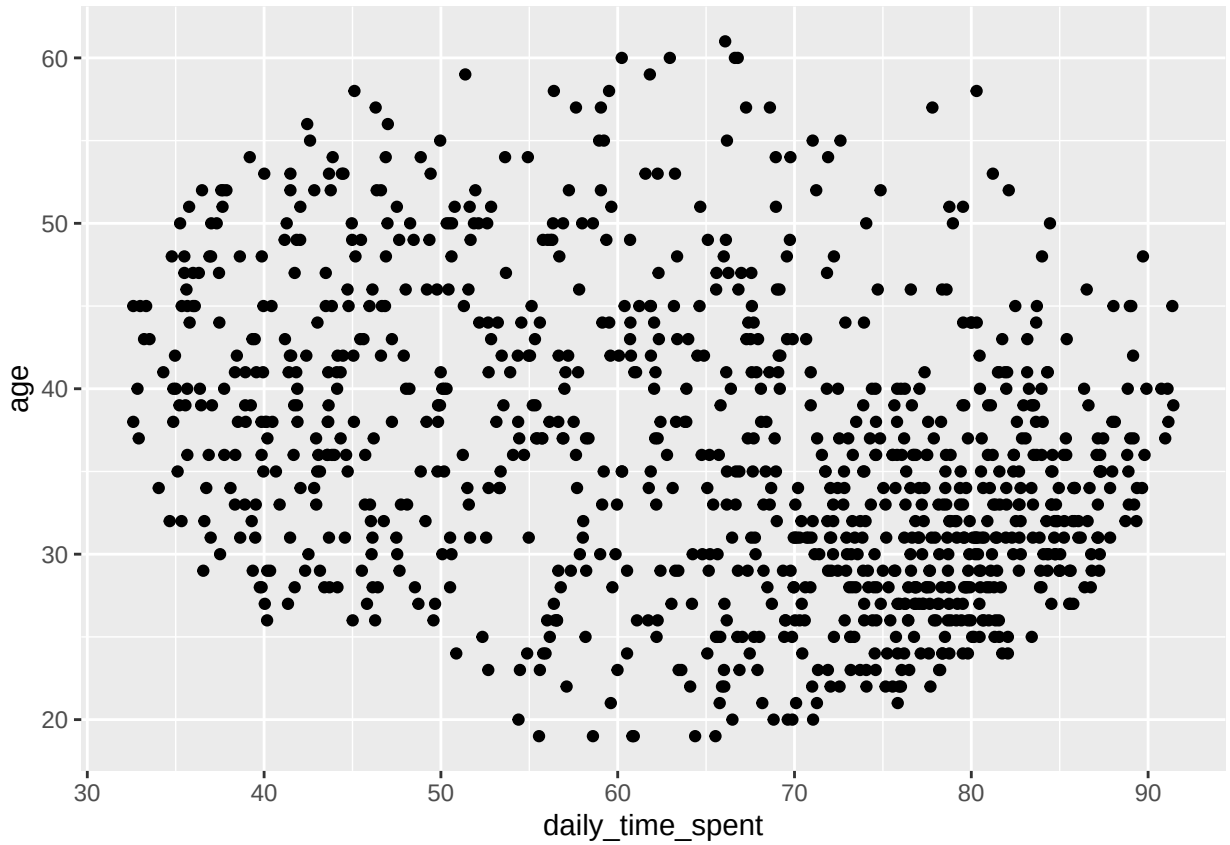
creating a scatter plot of area income and age

```
ggplot(data,
       aes(x = area_income,
           y = age)) +
  geom_point()
```



creating a scatter plot of time spent and age

```
ggplot(data,
       aes(x = daily_time_spent,
           y = age)) +
  geom_point()
```



MODELLING.

```
 # Feature engineering
# Installing libraries
library(caret)
```

```
## Loading required package: lattice
```

```
library(lattice)
```

```
# Randomising the records.
shuffle_index <- sample(1:nrow(data))
data <- data[shuffle_index, ]
dim(data)
```

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

```
# converting to numericals
data$daily_time_spent <- as.numeric(as.character(data$daily_time_spent))
data$age <- as.numeric(as.character(data$age))
data$area_income <- as.numeric(as.character(data$area_income))
```

```
data$ad_topic_line <- as.numeric(as.character(data$ad_topic_line))
```

```
## Warning: NAs introduced by coercion
```

```
data$male <- as.numeric(as.character(data$male))
data$country <- as.numeric(as.character(data$country))
```

```
## Warning: NAs introduced by coercion
```

```
# normalise the dataset.normalize <- function(x){
normalize <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
data$daily_time_spent <- normalize(data$daily_time_spent)
data$age <- normalize(data$age)
data$area_income <- normalize(data$area_income)
data$ad_topic_line <- normalize(data$ad_topic_line)
data$male <- normalize(data$male)
data$country <- normalize(data$country)
head(data)
```

```
##     daily_time_spent       age area_income daily_internet_usage ad_topic_line
## 487        0.8731939 0.5000000   0.7985422               158.42            NA
## 91         0.6340303 0.5714286   0.8764318               138.35            NA
## 952        0.1487336 0.1904762   0.3847550               162.46            NA
## 274        0.6166922 0.4285714   0.9791520               179.58            NA
## 733        0.8594255 0.5238095   0.8579939               194.95            NA
## 920        0.5602584 0.1428571   0.8497678               181.25            NA
##                  city male country           timestamp clicked_on_ad
## 487         West Lisa    1      NA 2016-02-02 11:49:18             0
## 91    Christopherport    0      NA 2016-05-13 11:51:10             1
## 952        Hintonport    1      NA 2016-03-03 03:51:27             1
## 274 Port Whitneyhaven    0      NA 2016-02-09 19:37:52             0
## 733      Robinsonland    0      NA 2016-01-04 04:00:35             0
## 920    West Arielstad    1      NA 2016-01-05 12:59:07             0
```

```
# splitting data into training and testing sets of 30's and 70's
intrain <- createDataPartition(y = data$clicked_on_ad, p = 0.7, list = FALSE)
training <- data[intrain,]
testing <- data[-intrain,]
```

```
# checking the dimensions of our split set
prop.table(table(data$clicked_on_ad))*100
```

```
##
##  0  1
## 50 50
```

```
# converting numeric data into factors
training$daily_time_spent <- as.character(as.numeric(training$daily_time_spent))
training$age <- as.character(as.numeric(training$age))
training$area_income <- as.character(as.numeric(training$area_income))
training$ad_topic_line <- as.character(as.numeric(training$ad_topic_line))
training$male <- as.character(as.numeric(training$male))
training$country <- as.character(as.numeric(training$country))
training$daily_internet_usage <- as.character(as.numeric(training$daily_internet_usage))
str(training)
```

```
## 'data.frame':    700 obs. of  10 variables:
##  $ daily_time_spent    : chr  "0.634030256671766" "0.148733639299677" "0.616692163861975" "0.8026517
##  $ age                 : chr  "0.571428571428571" "0.19047619047619" "0.428571428571429" "0.16666666
##  $ area_income         : chr  "0.876431820645825" "0.384754986768629" "0.97915200119716" "0.57286813
##  $ daily_internet_usage: chr  "138.35" "162.46" "179.58" "223.28" ...
##  $ ad_topic_line       : chr  NA NA NA NA ...
##  $ city                : chr  "Christopherport" "Hintonport" "Port Whitneyhaven" "Jayville" ...
##  $ male                : chr  "0" "1" "0" "1" ...
##  $ country             : chr  NA NA NA NA ...
##  $ timestamp           : chr  "2016-05-13 11:51:10" "2016-03-03 03:51:27" "2016-02-09 19:37:52" "201
##  $ clicked_on_ad       : int  1 1 0 0 1 1 1 1 1 0 ...
```

DECISION TREE.

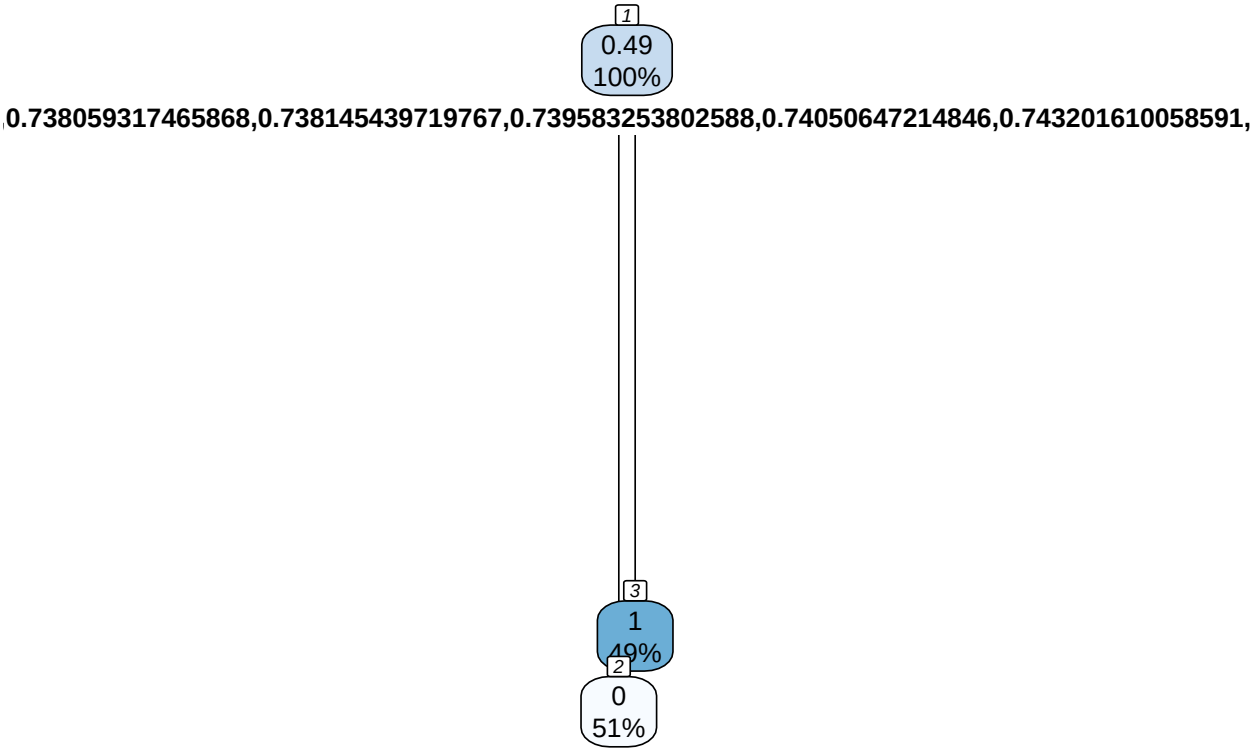
```
# importing libraries
library(rpart)
library(rpart.plot)
```

```
set.seed(12345)
train <- sample(1:nrow(training),size = ceiling(0.80*nrow(training)),replace = FALSE)
#training set
training_train <- training[train,]
#testing set
training.test <- training[-train,]
```

```
penalty.matrix <- matrix(c(0,1,10,0), byrow = TRUE, nrow = 2)
```

```
# building the classification tree with rpart
tree <- rpart(clicked_on_ad~.,
              data = training_train,
              parms = list(loss = penalty.matrix))
```

```
rpart.plot(tree, nn=TRUE)
```



4.CONCLUSION

A) The female gender had the highest numbers who visited the blog.

B) The age group between years 25 and 40 had the highest visits.

C) The income range of the highest visits was in the range of 50k to 70k.

D) The average amount of time spent by those who visited the site was between 75 to 85.

5.RECOMMENDATION

The largest audience for the cryptography course would be the people who would want to seek an extra source of income and it happens that the ages of 25 and 40 are often seeking somewhere to make extra coins,

From the analysis we confirm our target audience spent most time compared to any other age group, I would therefore recommend on creating content that will apeall to the group to reach a large target.