Week 12 IP

Ed Sang

2022-05-30

1.INTRODUCTION

a) Defining the Question

To identify which factors determining whether a user clicks on an ad or not.

b) Defining the metric of success

Finding and recommending the different feature characteristics that will increase the number of clicks in an ad.

c) Understanding the Context

Monitoring of ads helps entrepreneurs understand their effectiveness and being able to make adjustment that will be of gain to the firm and also its target audience.

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;

- Univariate Analysis
- Bivariate Analysis

Conclusions

Recommendation

2. DATA PREPARATION AND CLEANING

```r
#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
```

```r
#checking our dataset
dim(data)
```

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

```r
#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 ...
```

```r
# renaming column names
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"
```

```
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"
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 duplicates
anyDuplicated(data)
```

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

Our dataset has no duplicates

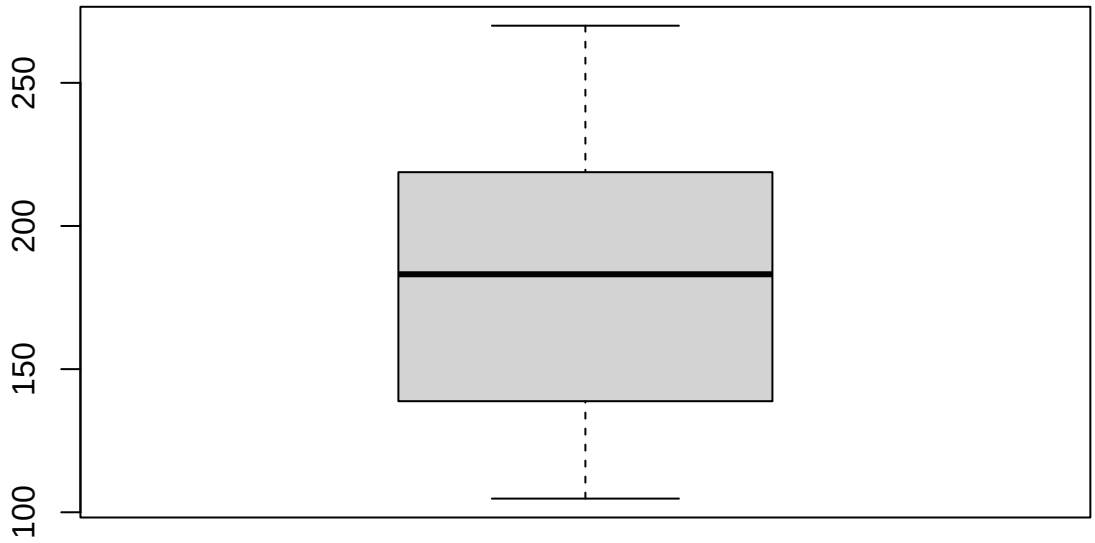
```
#checking for missing values
colSums(is.na(data))
```

```
##     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

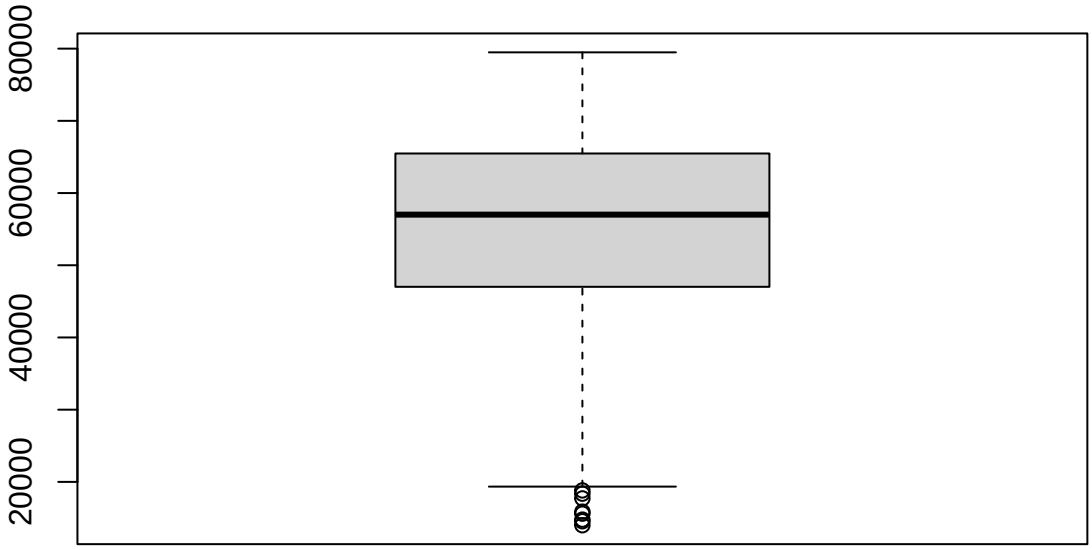
checking for outliers

```
# checking for outliers in our numerical values
boxplot(data$daily_internet_usage)
```



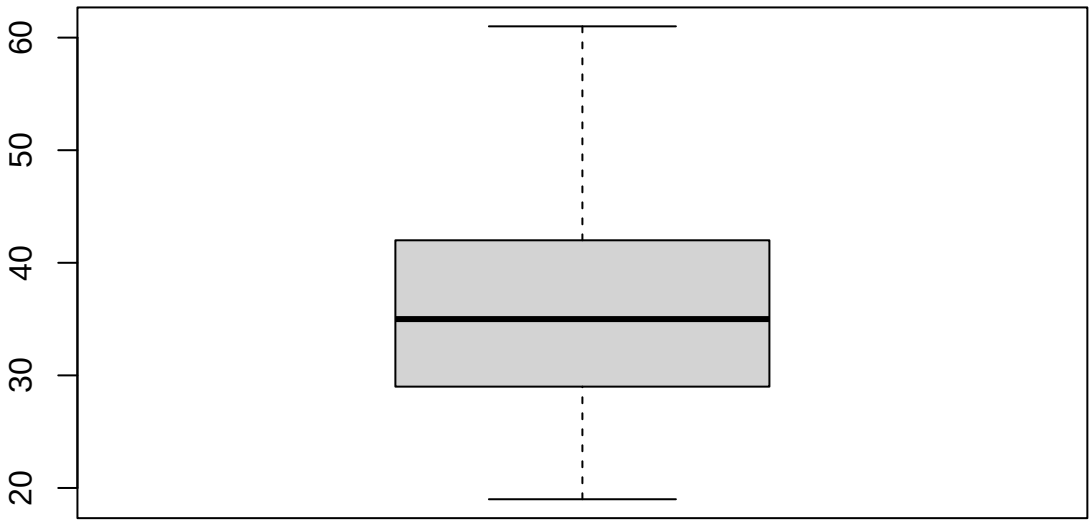
There are no outliers on daily_internet_usage column

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



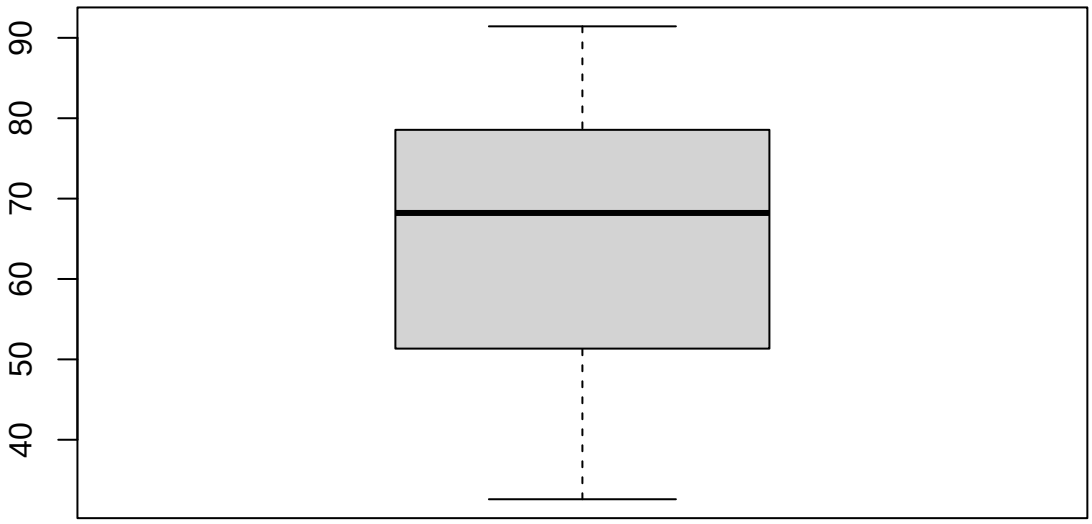
Presence of outliers in area_income column but we are going to keep them

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



There are no outliers on the age column

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



There are no utliers on the daily_time_spent

```
#checking for anomalies
unique_male<-unique(data$male)
unique_male
```

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

3.EXPLORATORY DATA ANALYSIS

Univariate analysis

```
# getting the minimum, maximum, mean,median and quartiles
summary(data$daily_internet_usage)
```

daily internet usage column

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

```r
# create function to calculate mode
getmode <- function(v) {
uniqv <- unique(v)
uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

```r
# mode in our column
getmode(data$daily_internet_usage)
```

```
## [1] 167.22
```

```r
#variance
var(data$daily_internet_usage)
```

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

```r
#standard deviation
sd(data$daily_internet_usage)
```

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

```r
#interquartile range
quantile(data$daily_internet_usage, 0.75) - quantile(data$daily_internet_usage, 0.25)
```

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

```r
#installing package 'moments'
library(moments)
```

```r
#checking for skewness
skewness(data$daily_internet_usage)
```

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

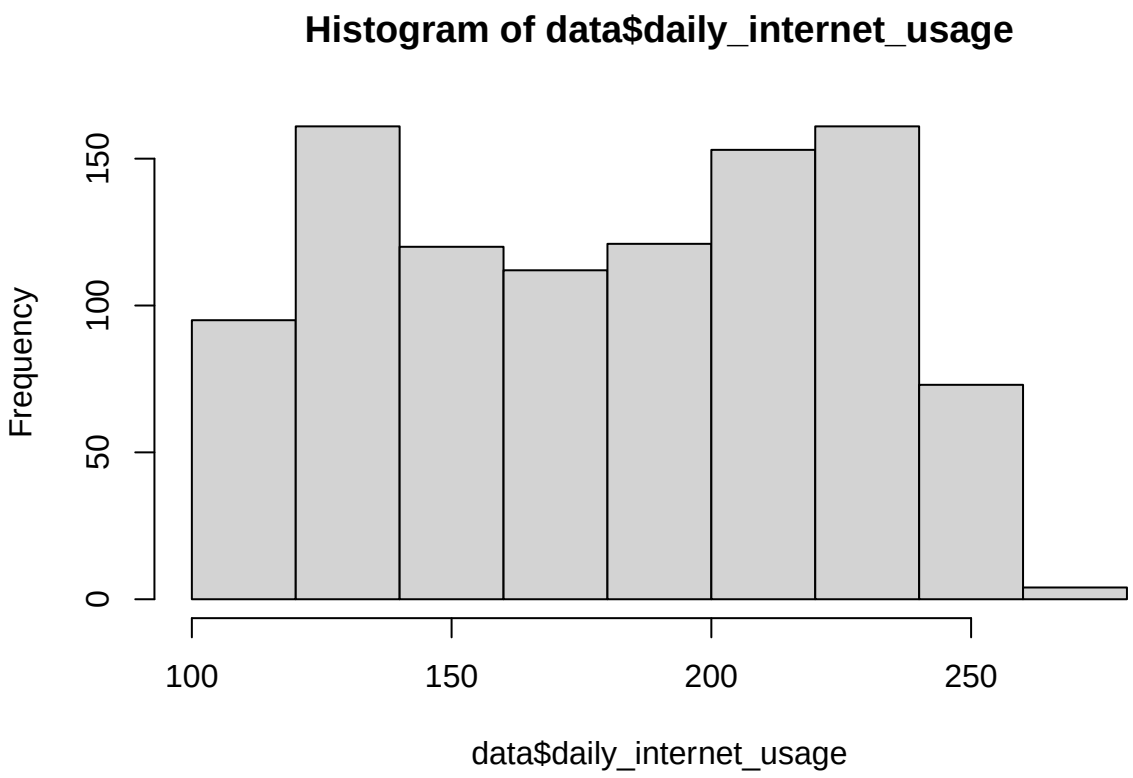
The variable is negatively skewed.

```r
# finding the kurtosis
kurtosis(data$daily_internet_usage)
```

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

The distributio is leptokurtic

```r
# checking the distribution
hist(data$daily_internet_usage)
```




```
# getting the minimum, maximum, mean,median and quartiles
summary(data$area_income)
```

area_income column

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13996   47032   57012   55000   65471   79485
```

```
# getting mode
getmode(data$area_income)
```

```
## [1] 61833.9
```

```
# getting variance
var(data$area_income)
```

```
## [1] 179952406
```

```
# getting standard deviation
sd(data$area_income)
```

```
## [1] 13414.63
```

```
# checking  kurtosis
kurtosis(data$area_income)
```

```
## [1] 2.894694
```

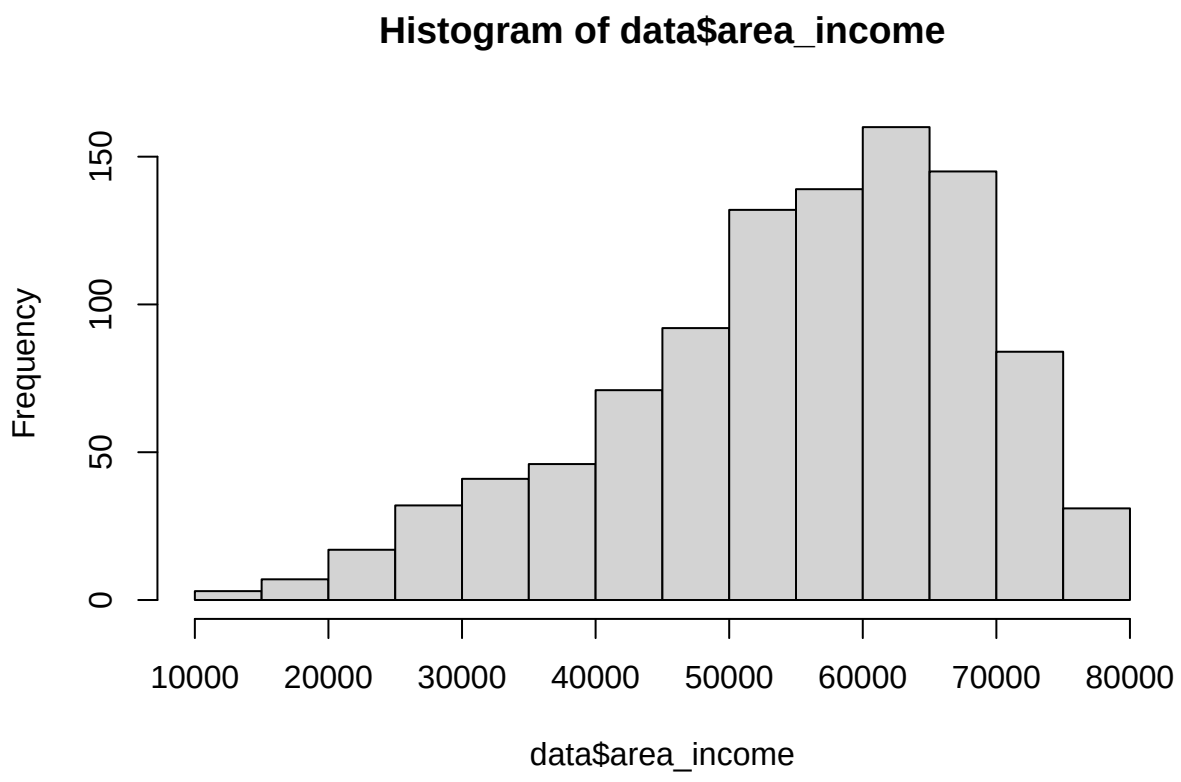
The distribution is leptokurtic

```
#checking for skewness
skewness(data$area_income)
```

```
## [1] -0.6493967
```

The distribution is negatively skewed

```
# checking for distribution
hist(data$area_income)
```




```
# getting the minimum, maximum, mean,median and quartiles
summary(data$age)
```

age column

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   29.00   35.00   36.01   42.00   61.00
```

```
#getting mode
getmode(data$age)
```

```
## [1] 31
```

```
#getting variance
var(data$age)
```

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

```
#getting standard deviation
sd(data$age)
```

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

```
#checking for kurtosis
kurtosis(data$age)
```

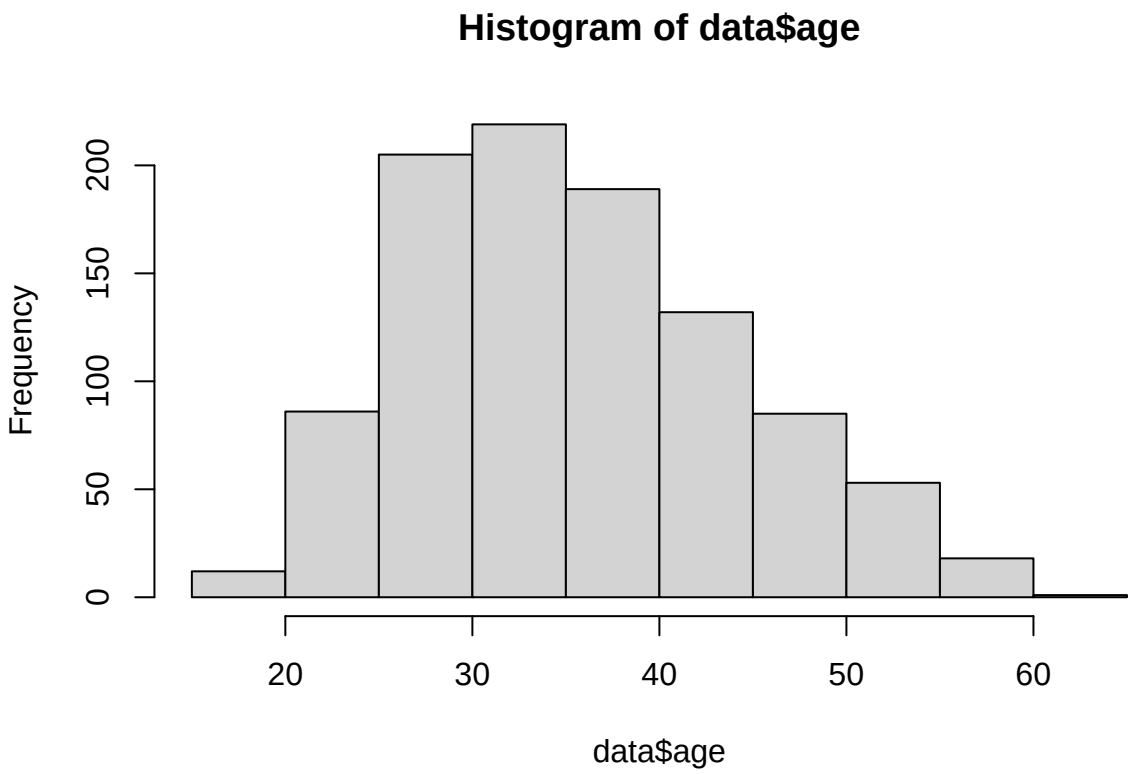
```
## [1] 2.595482
```

The distribution is leptokurtic

```
#checking for skewness
skewness(data$age)
```

```
## [1] 0.4784227
```

```
#checking the distribution
hist(data$age)
```




```
# getting the minimum, maximum, mean,median and quartiles
summary(data$daily_time_spent)
```

daily_time_spent column

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.60   51.36   68.22   65.00   78.55   91.43
```

```
# getting mode
getmode(data$daily_time_spent)
```

```
## [1] 62.26
```

```
# getting variance
var(data$daily_time_spent)
```

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

```
# getting standard deviation
sd(data$daily_time_spent)
```

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

```
# checking kurtosis
kurtosis(data$daily_time_spent)
```

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

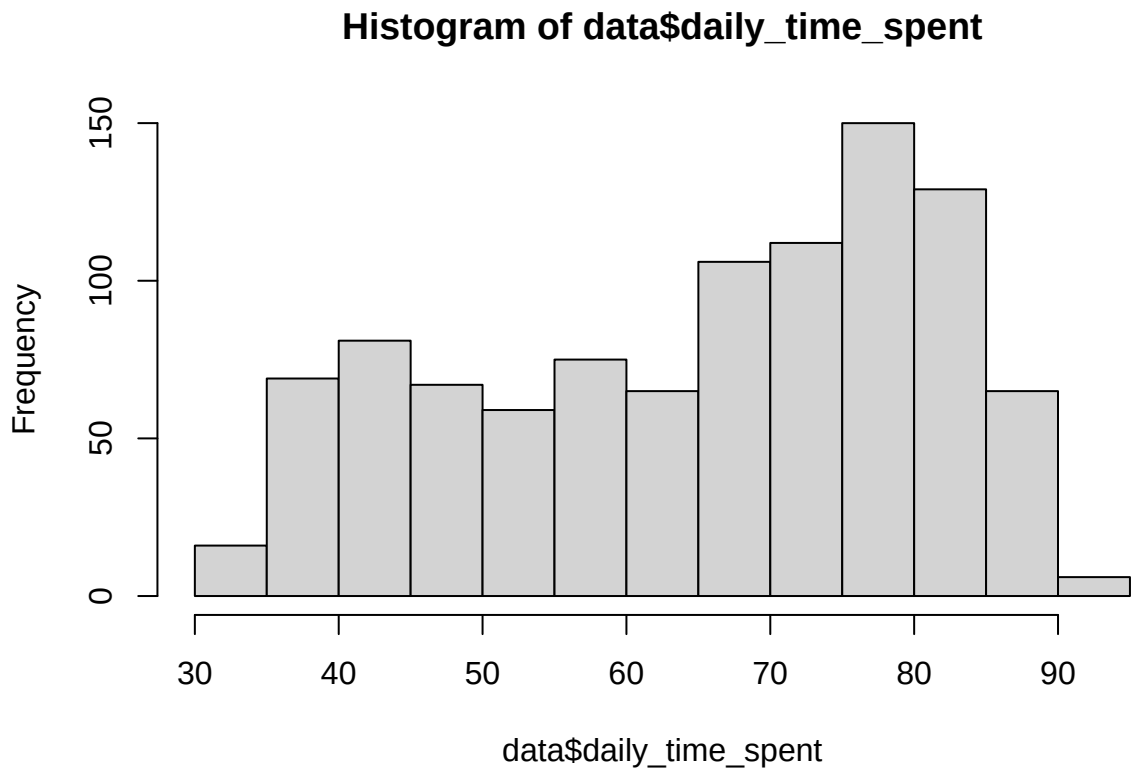
The distribution is leptokurtic

```
# checking for skewness
skewness(data$daily_time_spent)
```

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

The distribution is negatively skewed

```
# checking the distribution
hist(data$daily_time_spent)
```

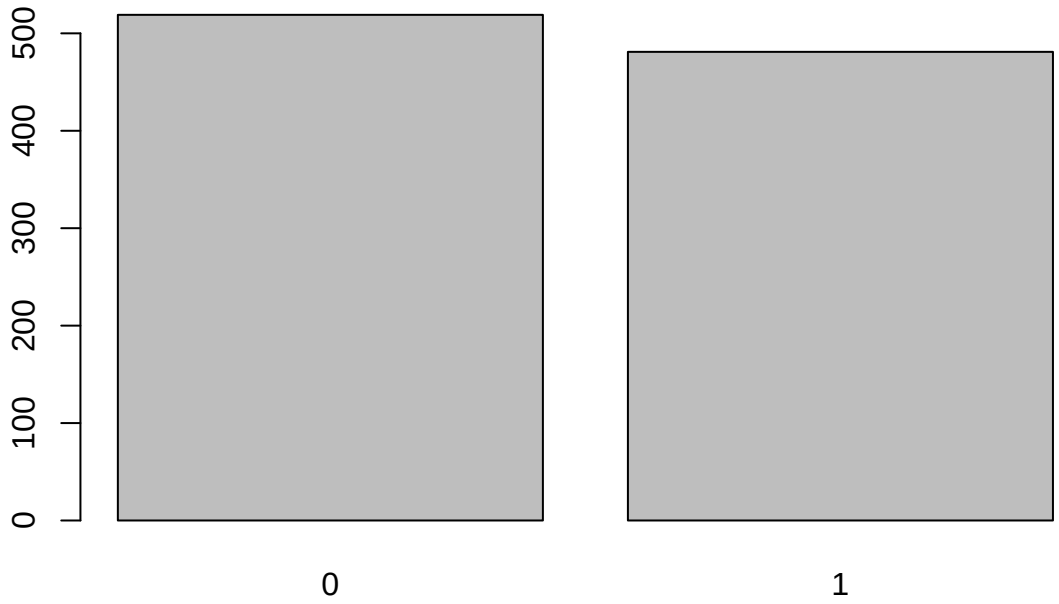



```
# checking the distribution of males and females
male_column <- table(data$male)
male_column
```

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

There are 519 females and 481 males.

```
# visual representation of the above information
barplot(male_column)
```




```
#displaying the 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
```

Lisamouth and Williamsport are the most occuring citites in the distribution.

```r
# showing 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
```

Czech Republic and France are leading as the most frequently occuring countries.

```r
# showing the number who clicked on add and those who did not
ad_column <- table(data$clicked_on_ad)
print(ad_column)
```

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

This shows that the number of people who clicked on ad is the same as the number who did not click on the ad

Bivariate analysis

```r
# 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
```

```r
# 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
```

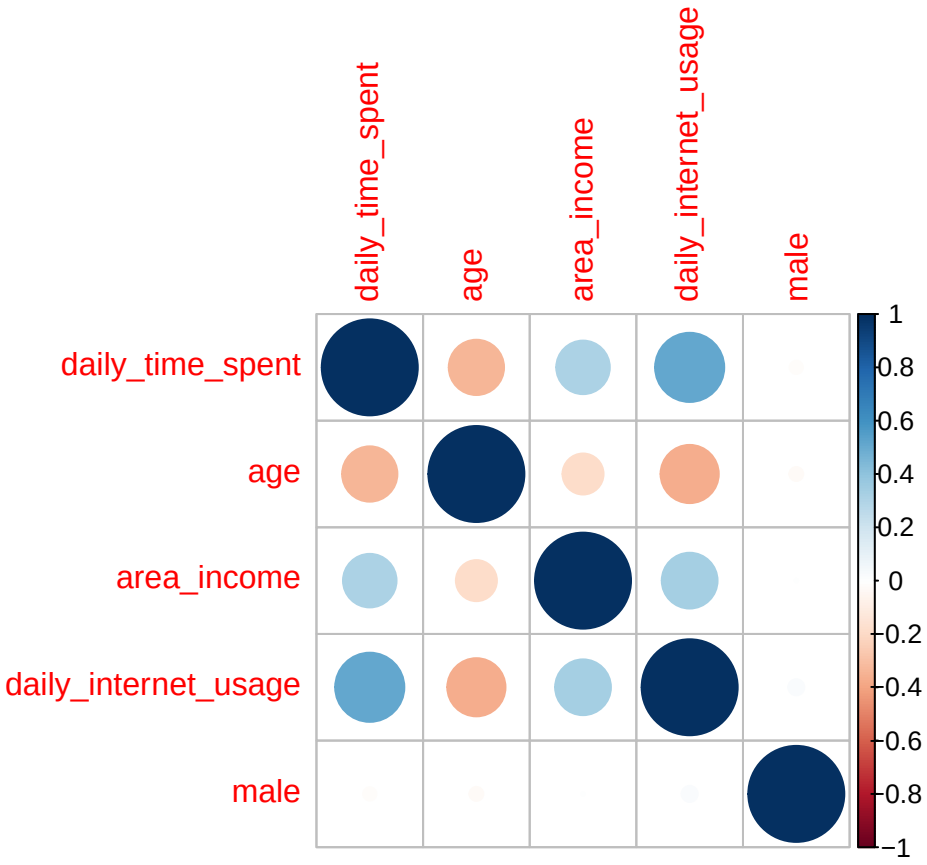
```
## daily_internet_usage       0.51865848 -0.36720856  0.337495533
## male                      -0.01895085 -0.02104406  0.001322359
##                      daily_internet_usage         male
## daily_time_spent              0.51865848 -0.018950855
## age                          -0.36720856 -0.021044064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

From our matrix we can see that there is a positive correlation between 'daily_time_spent' and 'daily_internet_usage' columns of 0.5186.

```
# Installing the correlation plot to visualize the correlation coefficients.
library(corrplot)
```

```
## corrplot 0.92 loaded
```

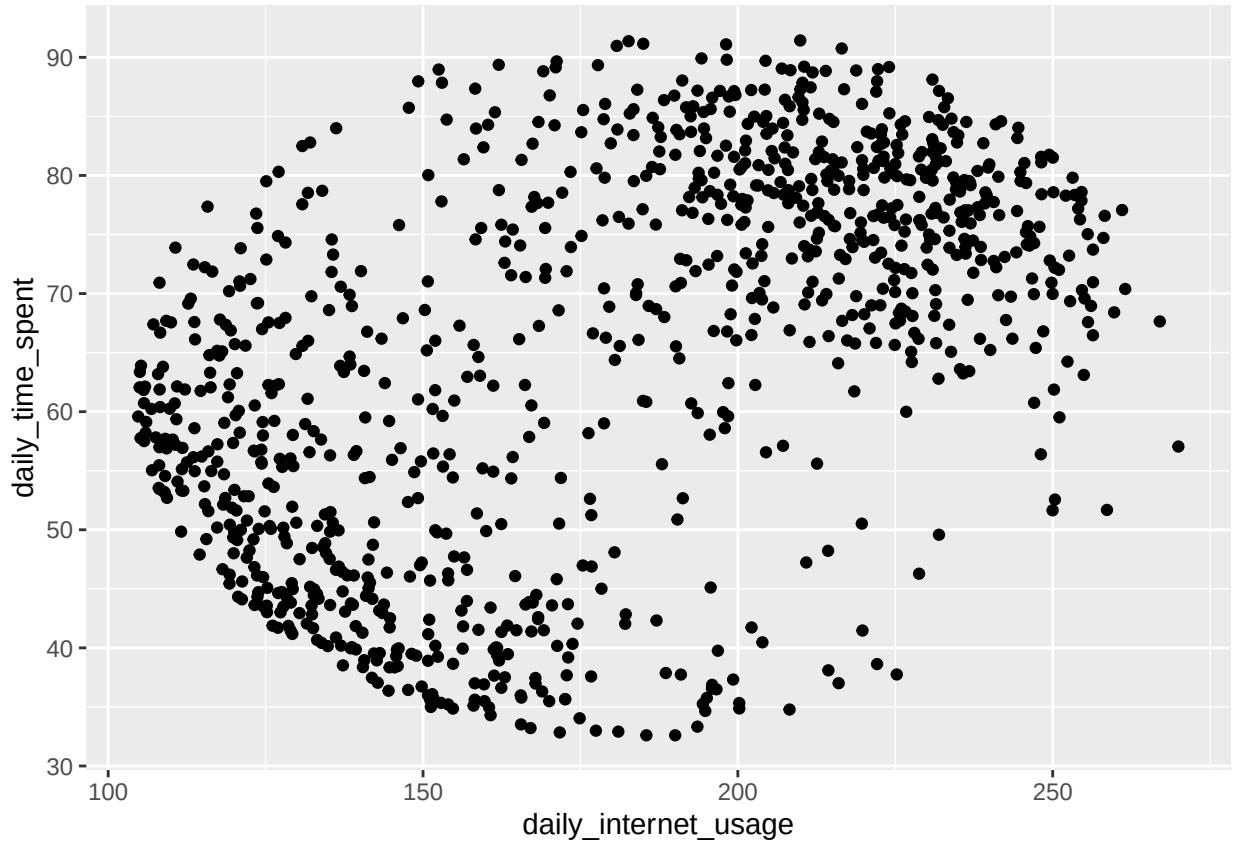
```
#visualization
corrplot(numerical.cor)
```



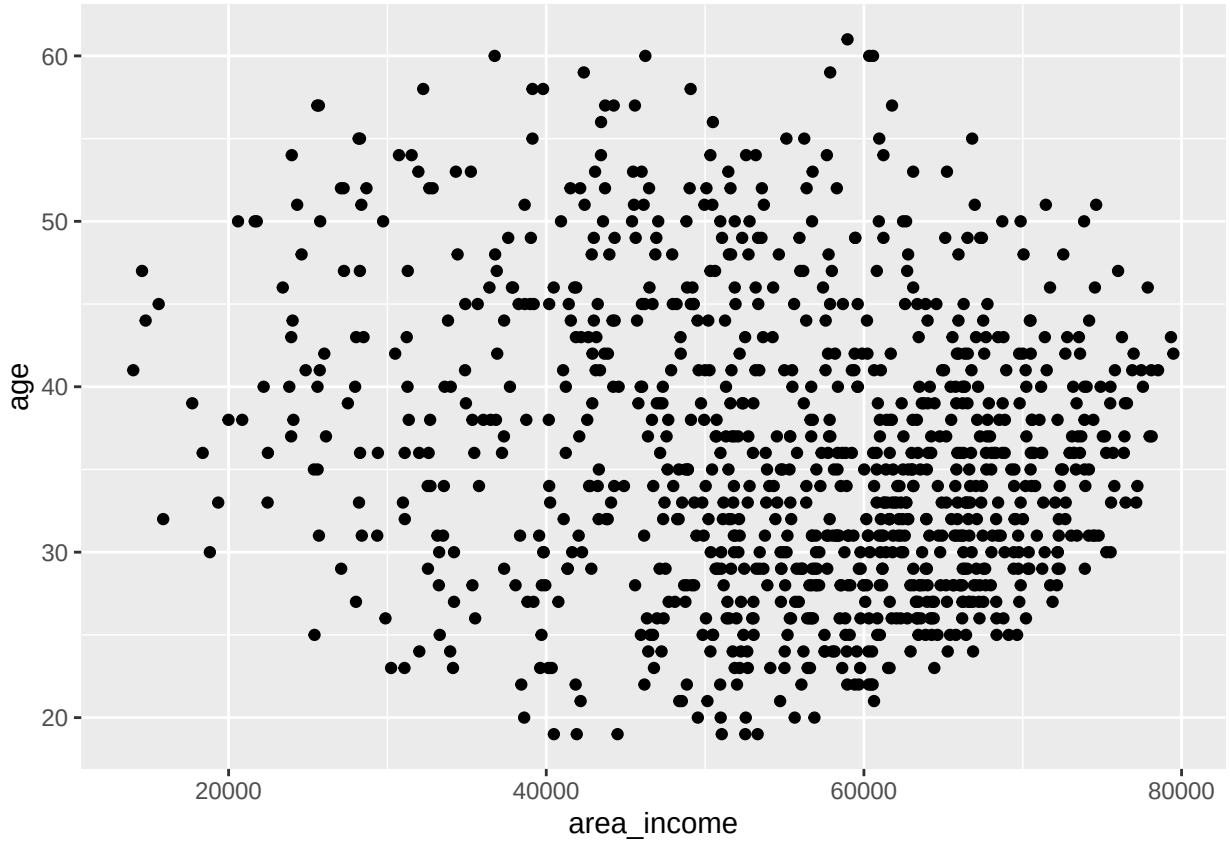
We can see that majority of the columns have no correlation with each other.

```
# importing library
library(ggplot2)
```

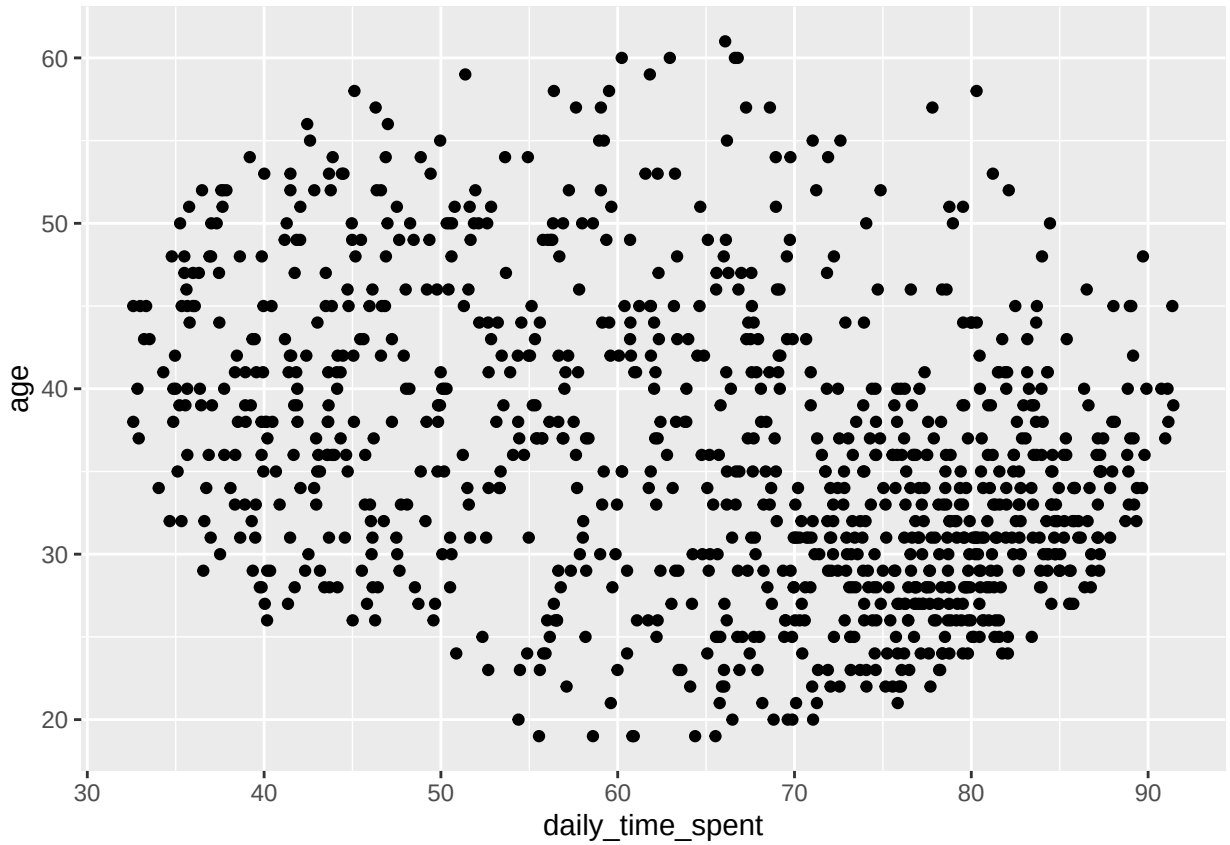
```
#creating a scatter plot to show our most positively correlated columns
ggplot(data,aes(x = daily_internet_usage,y = daily_time_spent)) + geom_point()
```




```
# 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()
```



4.MODELLING

Feature Engineering

```
#installing libraries to help in computation
library(lattice)
library(caret)
```

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

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

```
# convert the factors into numerics
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
```

```
str(data)
```

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ daily_time_spent    : num  41.5 74.4 66.9 41.9 79.2 ...
##  $ age                 : num  42 26 23 38 26 26 42 52 58 28 ...
##  $ area_income         : num  52177 64396 64434 68520 62312 ...
##  $ daily_internet_usage: num  129 163 208 163 203 ...
##  $ ad_topic_line       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ city                : chr  "Robertstown" "Dianashire" "Sandersland" "Hartmanchester" ...
##  $ male                : num  1 0 1 0 0 0 1 0 0 0 ...
##  $ country             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ timestamp           : chr  "2016-05-23 08:06:24" "2016-07-23 04:37:05" "2016-01-04 07:28:43" "201
##  $ clicked_on_ad       : int  1 0 0 1 0 0 1 1 1 1 ...
```

```
# Normalizing our dataset
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
## 833        0.1506034 0.5476190   0.5830186                128.98            NA
## 174        0.7106918 0.1666667   0.7695932                163.05            NA
## 579        0.5828659 0.0952381   0.7701756                208.24            NA
## 212        0.1579126 0.4523810   0.8325680                163.38            NA
## 270        0.7912630 0.1666667   0.7377765                203.23            NA
## 373        0.7961924 0.1666667   0.7153805                206.79            NA
##                city male country           timestamp clicked_on_ad
## 833    Robertstown    1      NA 2016-05-23 08:06:24             1
## 174     Dianashire    0      NA 2016-07-23 04:37:05             0
## 579    Sandersland    1      NA 2016-01-04 07:28:43             0
## 212 Hartmanchester    0      NA 2016-02-03 04:21:14             1
## 270     Port Robin    0      NA 2016-01-20 19:09:37             0
## 373    West Thomas    0      NA 2016-05-15 13:18:34             0
```

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

```
#checking our train set
dim(training)
```

```
## [1] 700  10
```

```
#checking our test set
dim(testing)
```

```
## [1] 300  10
```

Our training data will be 700 while our testing data will be 300

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

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

```
prop.table(table(training$clicked_on_ad))*100
```

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

```
prop.table(table(testing$clicked_on_ad))*100
```

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

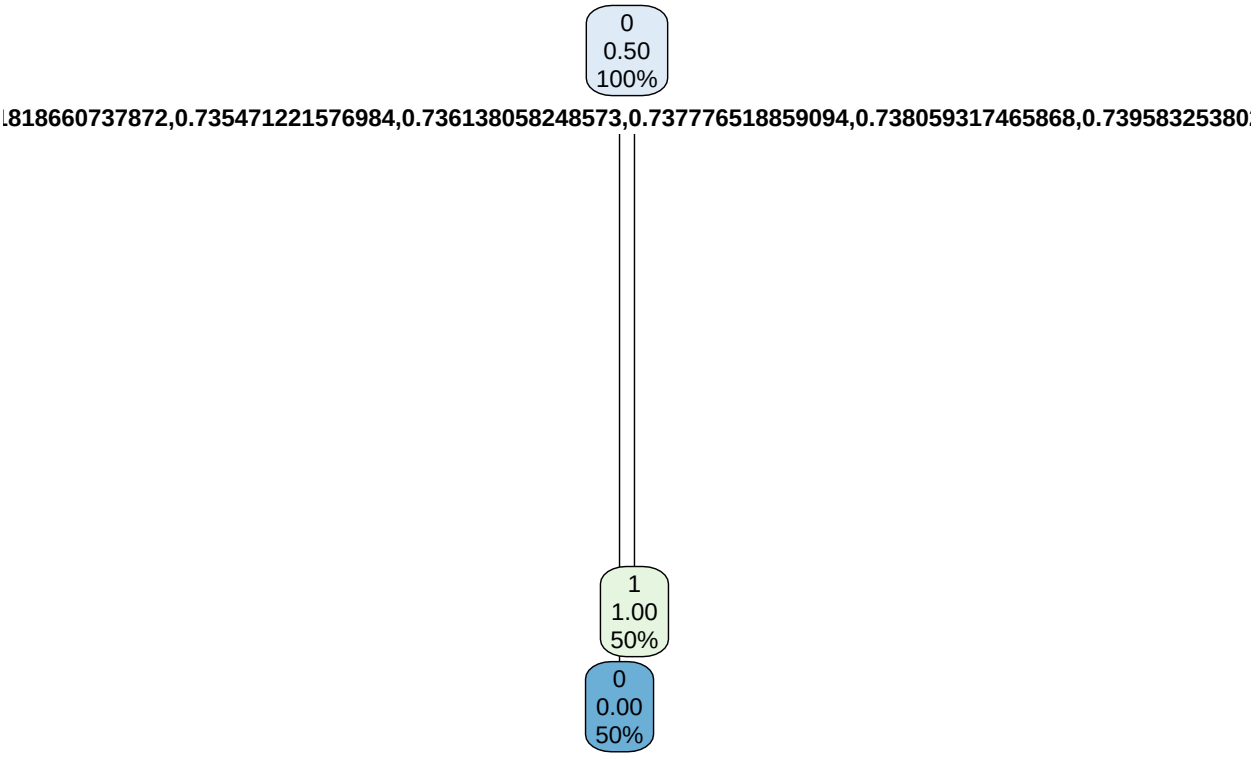
```
# converting numeric data into factors
training$daily_time_spent <- as.character(as.numeric(training$daily_time_spent))
training$age <- as.character(as.numeric(training$age))
training$area_income <- as.character(as.numeric(training$area_income))
training$ad_topic_line <- as.character(as.numeric(training$ad_topic_line))
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  "0.150603433622302" "0.710691823899371" "0.582865884752677" "0.1579126
##  $ age                 : chr  "0.547619047619048" "0.166666666666667" "0.0952380952380952" "0.4523809
##  $ area_income         : chr  "0.583018646078765" "0.769593194509554" "0.770175588616593" "0.8325679
##  $ daily_internet_usage: chr  "128.98" "163.05" "208.24" "163.38" ...
##  $ ad_topic_line       : chr  NA NA NA NA ...
##  $ city                : chr  "Robertstown" "Dianashire" "Sandersland" "Hartmanchester" ...
##  $ male                : chr  "1" "0" "1" "0" ...
##  $ country             : chr  NA NA NA NA ...
##  $ timestamp           : chr  "2016-05-23 08:06:24" "2016-07-23 04:37:05" "2016-01-04 07:28:43" "2016
##  $ clicked_on_ad       : int  1 0 0 1 0 1 1 1 1 0 ...
```

Decision Tree

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

```
# fitting and training the model using decision tree
fit <- rpart(clicked_on_ad ~ ., data = training, method = 'class')
rpart.plot(fit, extra = 106)
```



5.CONCLUSION

A) The largest number of people who visited the blog were females.

B) Majority of those who visited the blog are in the 50,000 to 70,000 area income.

C) Most of the people who visited the blog are between the ages of 25 to 40.

D) Most of the people who visited the blog spent 75 to 85 minutes on the site.

E) A change in time spent on site causes an increase in area income

6.RECOMMENDATION

The entrepreneur should focus on creating more content that is focused on the ages of 25 to 40 years since this is her clientele base. She should focus on more on the male clients to increase her market since they are fewer in numbers. Since most of her people come from high income areas, she should focus more on that to expose her business. She should offer discount to her most frequent customers so that they would continue clicking on the ad.