Week12 IP

Brendah

2022-03-19

1. INTRODUCTION

1.1 Defining the Question

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

1.2 Setting the Metric for Success

The project will be considered a success when I am able to identify what makes a user more likely to click on an ad.

1.3 Outlining 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. 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.

1.4 Drafting the Experimental Design

1. Define the question, the metric for success, the context, experimental design taken and the appropriateness of the available data to answer the given question.
2. Load the dataset and previewing it.
3. Check for missing and duplicated values and deal with them where necessary.
4. Check for outliers and other anomalies and deal with them where necessary.
5. Perform univariate and bivariate analysis.
6. Conclude .

1.5 Determining the Appropriateness of the Data

2. Data Preparation and Cleaning

```
#data<- advertising
#head(data)
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
```

```
#to check the number of rows and columns in the dataset
dim(data)
```

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

We can see that there are 1000 rows and 10 columns

```
#to show the structure of the data set specifically the data types of the columns
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 ...
```

The dataset contains 3 num, 3 int and 4 chr variable datatypes

I will change the column names and then change the “Male” and “Clicked on Ad” columns to be categorical variables (Factors) instead of numerical variables . The modification will be to make the dataset easier to work with

```
# get column names
colnames(data)
```

```
##  [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"
```

```
# rename them
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"
#to confirm they've been changed
colnames(data)
```

```
##  [1] "daily_time_spent"     "age"                  "area_income"
##  [4] "daily_internet_usage" "ad_topic_line"        "city"
##  [7] "male"                 "country"              "timestamp"
## [10] "clicked_on_ad"
```

```
# changing the data types of the "male" and "clicked_on_ad" columns from integer to categorical
data$male <- as.factor(data$male)
data$clicked_on_ad<-as.factor(data$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
```

```
str(data)
```

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ daily_time_spent    : 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 standardizatio
##  $ city                : chr  "Wrightburgh" "West Jodi" "Davidton" "West Terrifurt" ...
##  $ male                : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 2 2 ...
##  $ 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
##  $ clicked_on_ad       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
```

We can now see that the data types of the columns 'male' and 'clicked_on_ad' have changed from int to Factor(categorical)

```
# checking for duplicates
anyDuplicated(data)
```

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

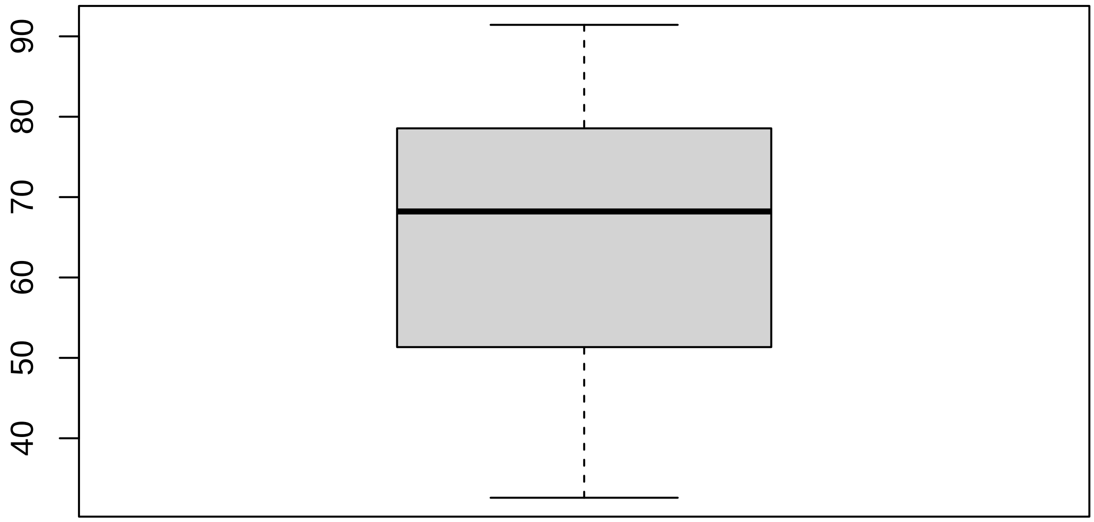
There are no duplicates

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

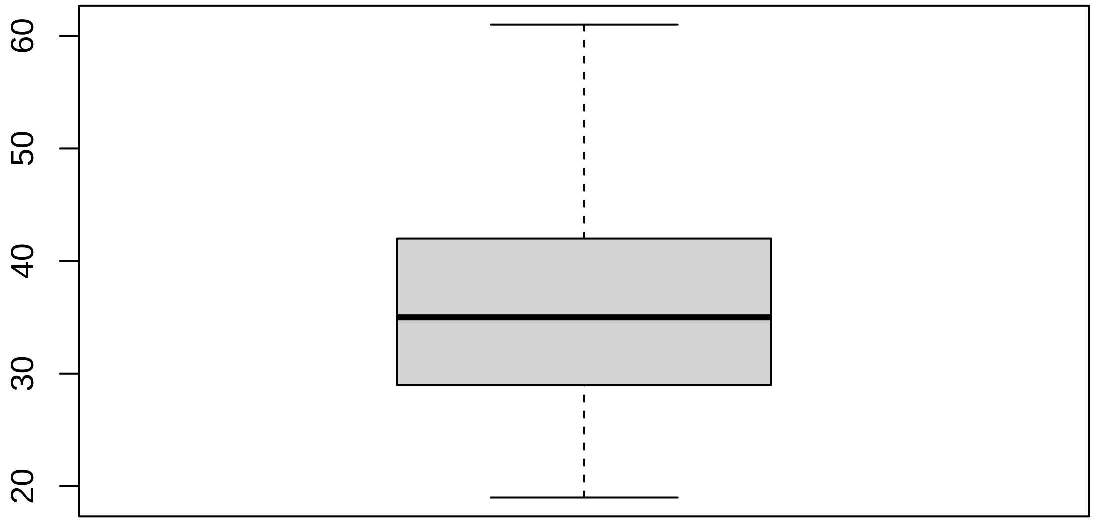
```
##     daily_time_spent                  age          area_income
##                    0                    0                    0
## daily_internet_usage        ad_topic_line                 city
##                    0                    0                    0
##                 male              country            timestamp
##                    0                    0                    0
##        clicked_on_ad
##                    0
```

There are no missing values in each column

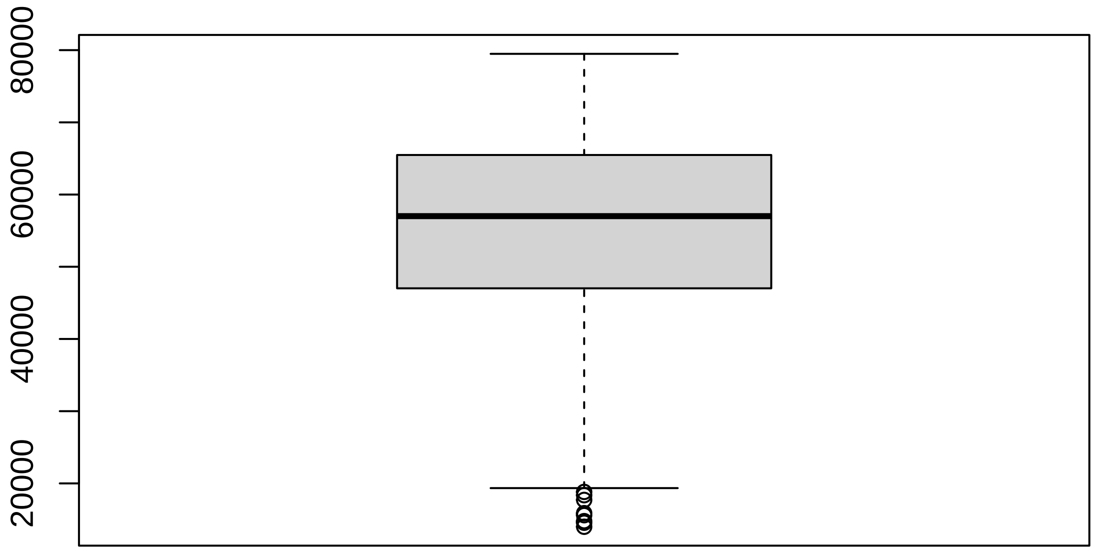
```
# Using boxplot to check for outliers of numerical variables
boxplot(data$daily_time_spent)
```



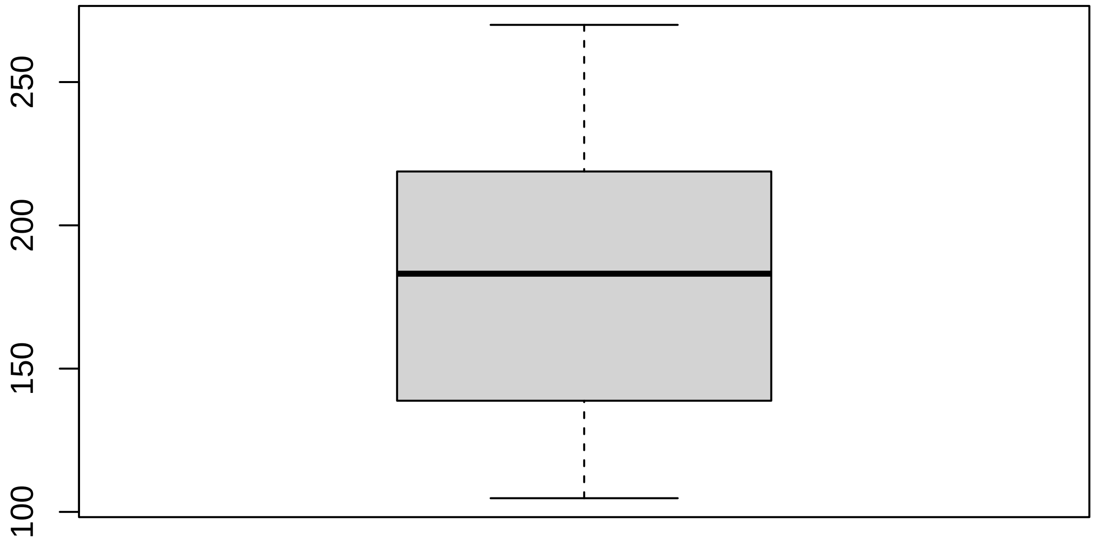
```
boxplot(data$age)
```



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



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



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

```
## [1] 0 1
## Levels: 0 1
```

3. Exploratory Data Analysis

3.1 Univariate Analysis

```
#summary of all columns
summary(data)
```

```
##  daily_time_spent      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
##  Length:1000        0:500
##  Class :character   1:500
##  Mode  :character
##
##
##
```

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

3.1.1 Daily Time Spent

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

```
# create function to calculate mode since R doesn't have an in-built function to do that
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
# now calling the mode function on our column
getmode(data$daily_time_spent)
```

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

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

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

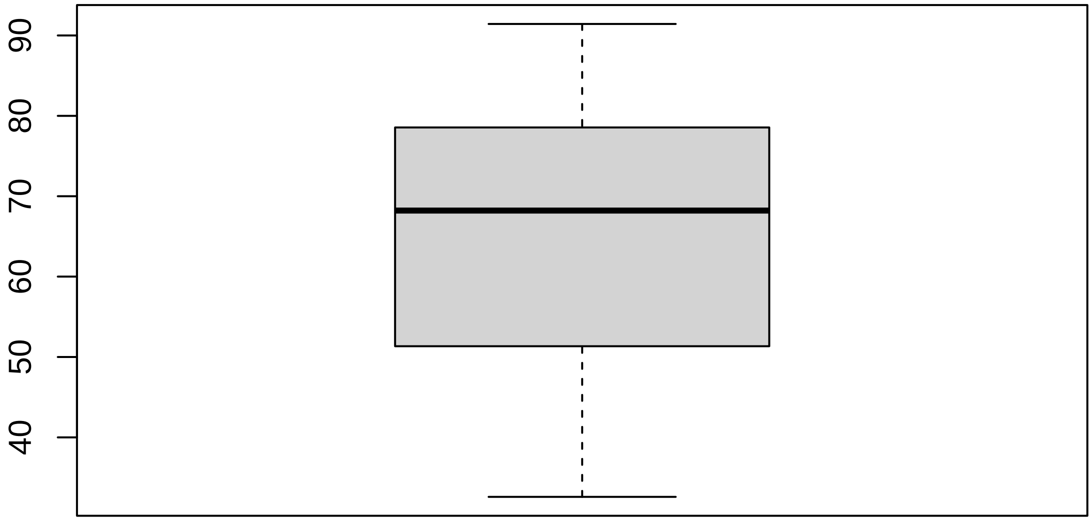
```
# find standard deviation
sd(data$daily_time_spent)
```

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

```
# get interquantile range
quantile(data$daily_time_spent, 0.75) - quantile(data$daily_time_spent, 0.25)
```

```
##     75%
## 27.1875
```

```
# graph boxplot
boxplot(data$daily_time_spent)
```



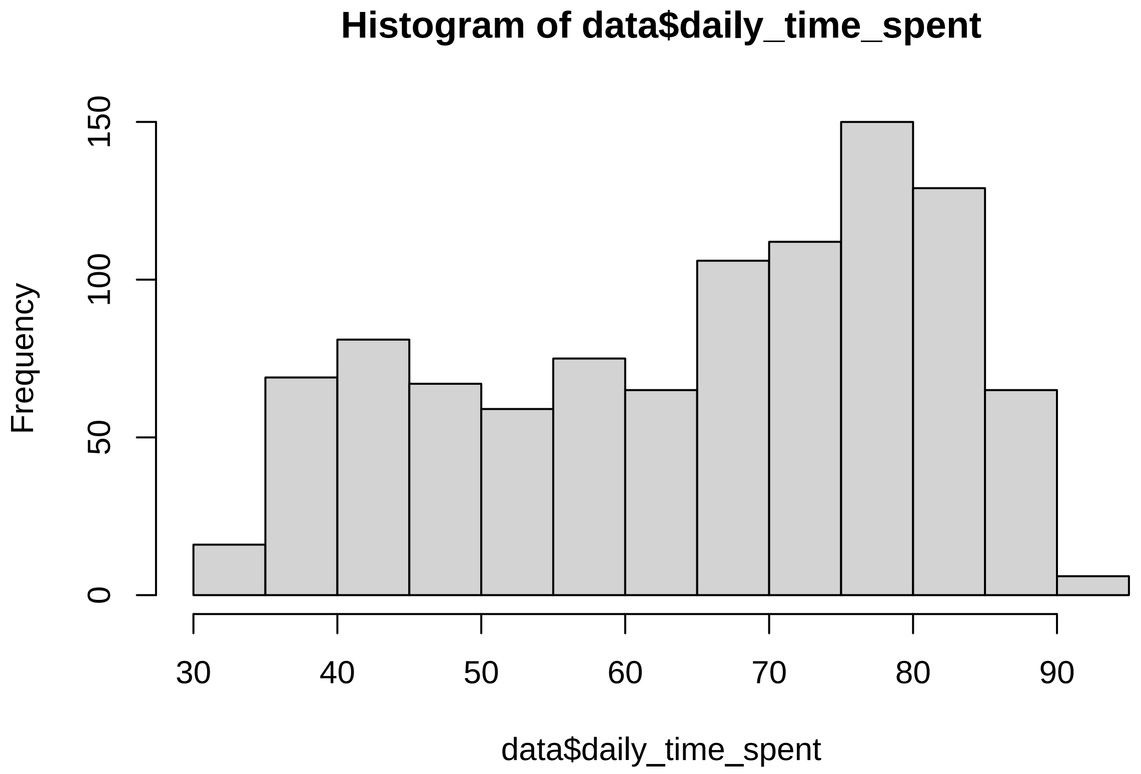
This variable does not have any outliers.

```
# find the kurtosis of this variable

library(moments)
kurtosis(data$daily_time_spent)
```

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

This kurtosis value is less than 3 implying that the distribution of this variable is platykurtic which means that there are few to no outliers.

```
# check distribution
hist(data$daily_time_spent)
```



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

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

This variable is slightly negatively skewed.

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

3.1.2 Age

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

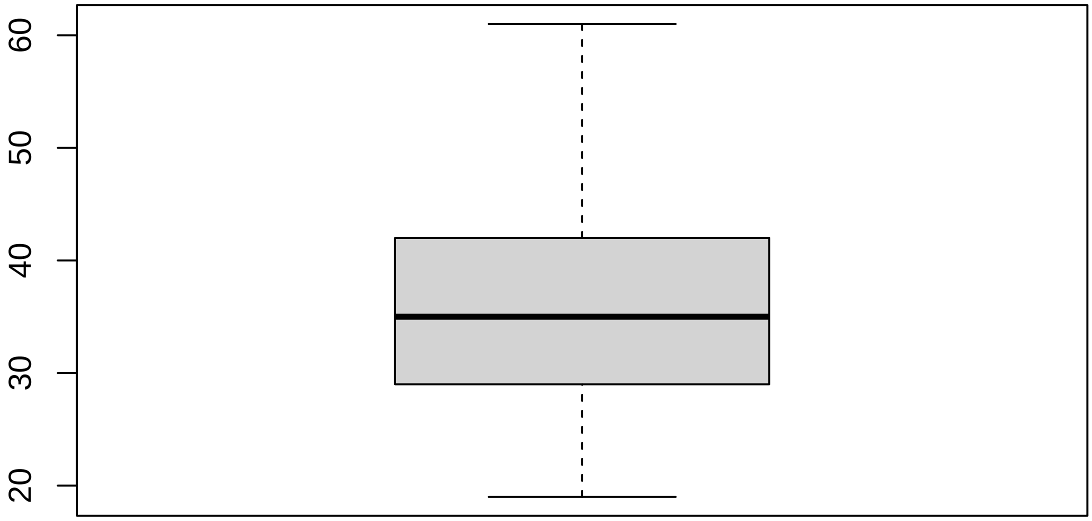
```
# getting mode
getmode(data$age)
```

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

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

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

```
# check for outliers
boxplot(data$age)
```



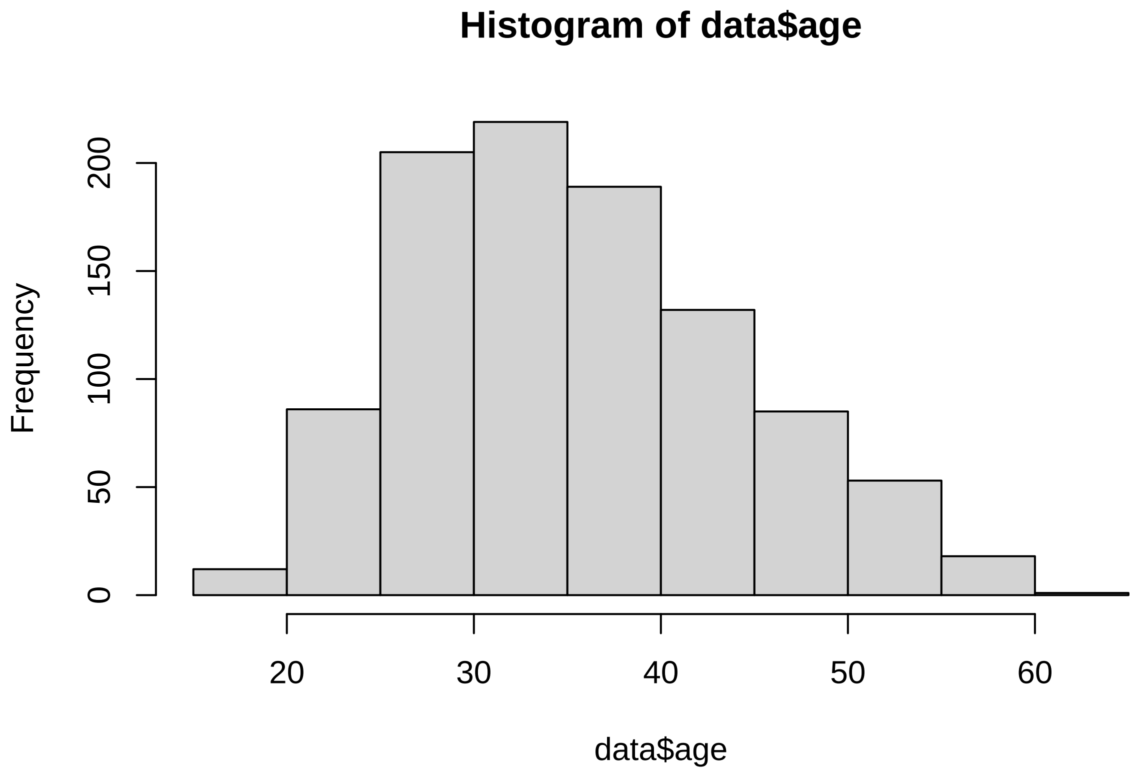
No outliers.

```
# check kurtosis
kurtosis(data$age)
```

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

The distribution is platykurtic implying the existence of few to no outliers.

```
# check distribution
hist(data$age)
```



The distribution looks almost normal

```
skewness(data$age)
```

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

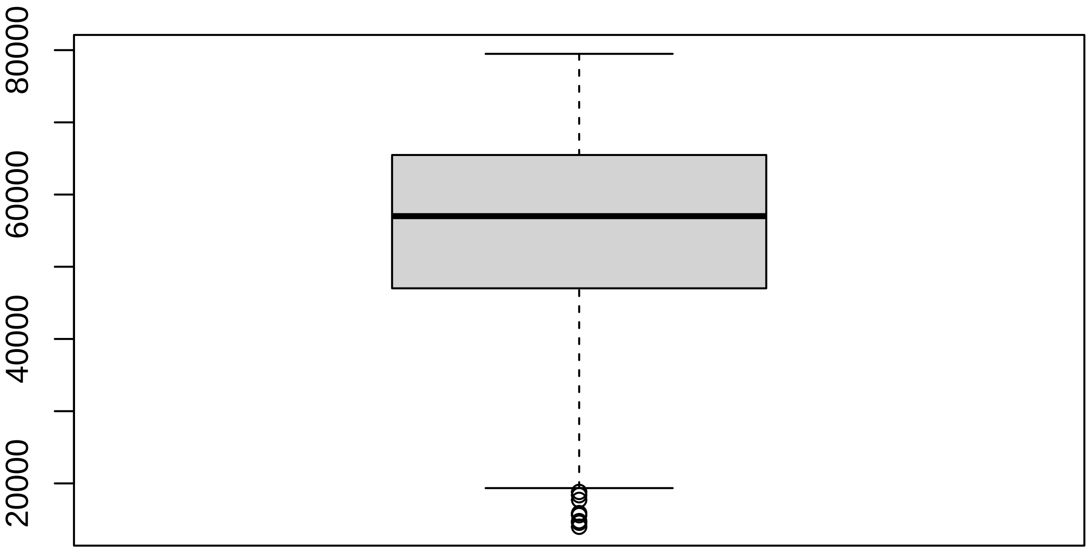
This skewness value implies that the distribution is almost fairly symmetrical unlike our observation from the histogram distribution where it looks almost normal

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

3.1.3 Area Income

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

```
# check for outliers
boxplot(data$area_income)
```



There are outliers below the 20,000 mark.

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

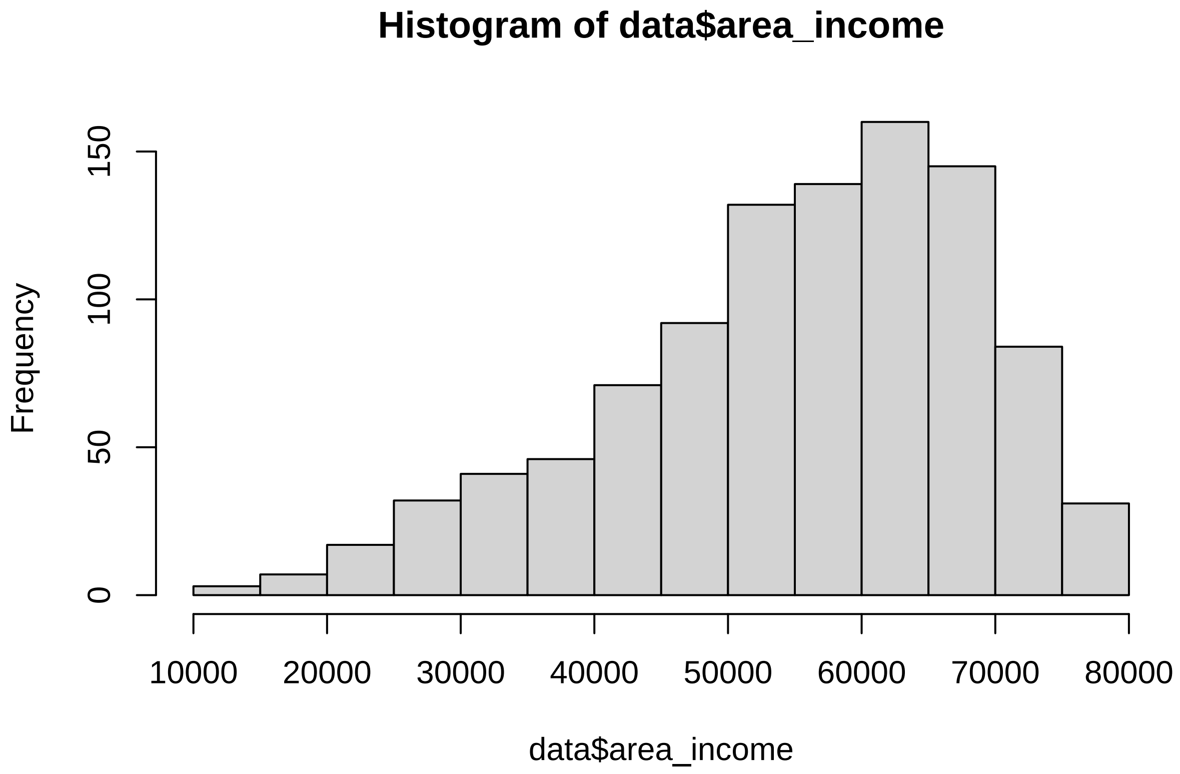
```
## [1] 61833.9
```

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

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

A kurtosis value of 2.89 indicates that the distribution is platykurtic although it is getting very close to being mesokurtic.

```
# check distribution
hist(data$area_income)
```



The distribution is negatively skewed.

```
# check skewness
skewness(data$area_income)
```

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

it's indeed negatively skewed

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

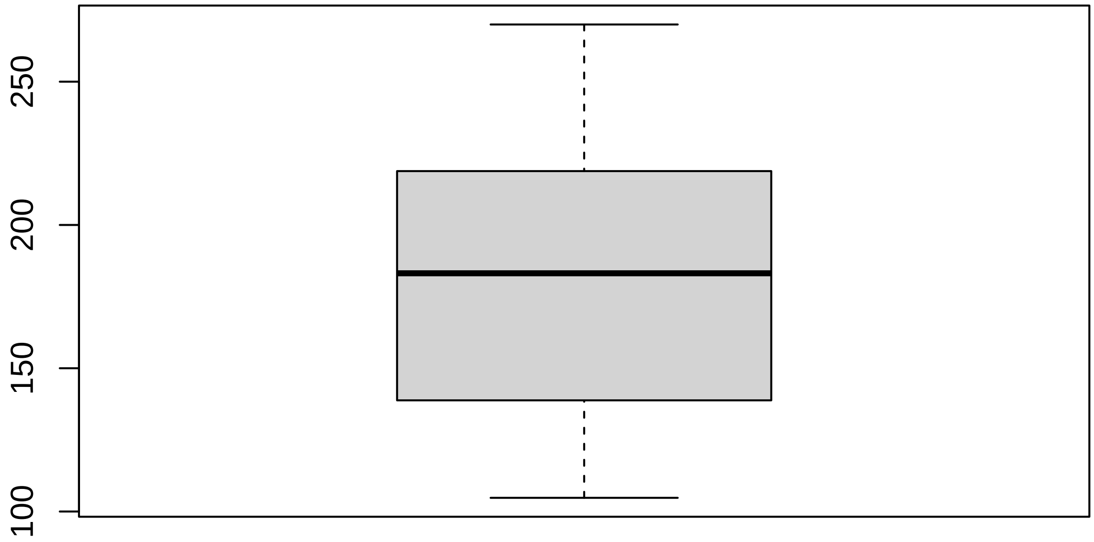
3.1.4 Daily Internet Usage

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

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

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

```
# check for outliers
boxplot(data$daily_internet_usage)
```



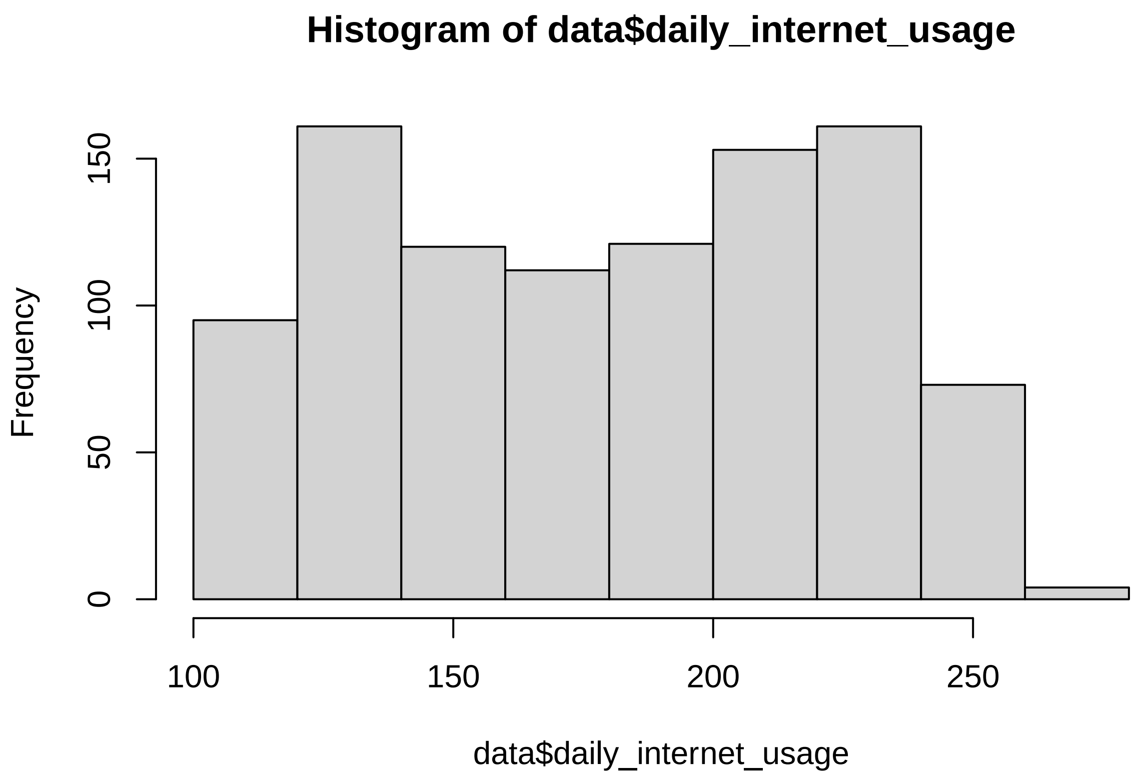
There are no outliers in this column.

```
# check kurtosis
kurtosis(data$daily_internet_usage)
```

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

The distribution is platykurtic.

```
# check distribution
hist(data$daily_internet_usage)
```



The distribution appears to be relatively uniform and bimodal.

```
# check skewness
skewness(data$daily_internet_usage)
```

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

```


# displaying the first 6 frequently occurring cities
 library(plyr)
count_city <- count(data$city)
count_city_head <- head(arrange(count_city, desc(freq)))
count_city_head
```

3.1.5 city

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

Lisamouth,Williamsport,Benjaminchester,East John,East Timothy and Johnstad are 6 frequently occurring cities

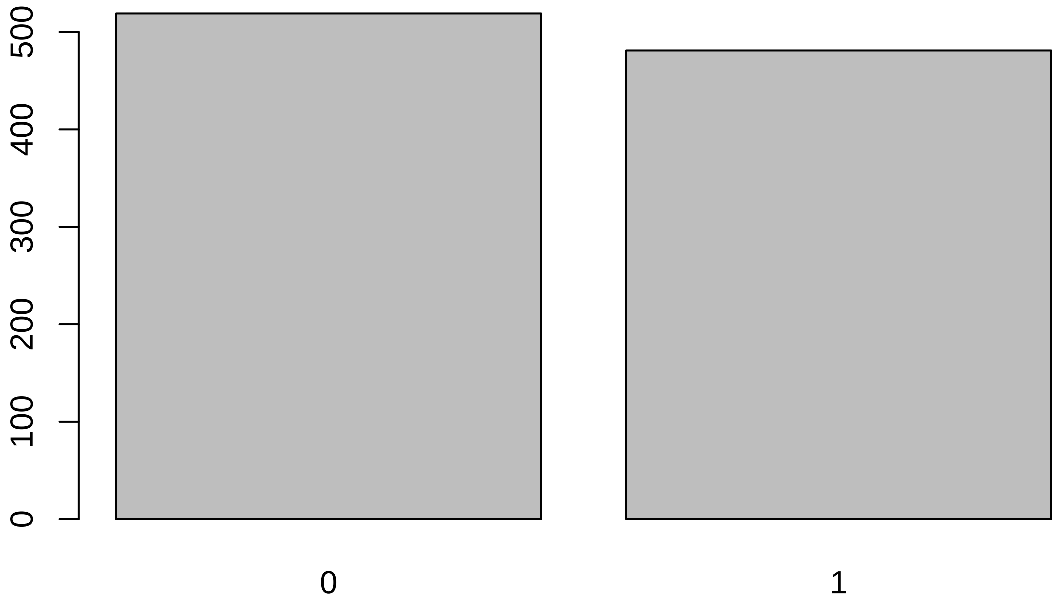
```
male_table <- table(data$male)
male_table
```

3.1.6 male

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

We see here that 519 are not male while 481 are. To easily visualize this:

```
barplot(male_table)
```



```
# displaying the first 10 frequently occuring countries
count_country <- count(data$country)
count_country_head <- head(arrange(count_country, desc(freq)), 10)
count_country_head
```

3.1.7 country

```
##                   x freq
## 1  Czech Republic    9
## 2          France    9
## 3     Afghanistan    8
## 4       Australia    8
## 5          Cyprus    8
## 6          Greece    8
## 7         Liberia    8
## 8      Micronesia    8
## 9            Peru    8
## 10        Senegal    8
```

The table displays the 10 frequently occuring countries with Czech Republic and France leading

```
ad_table <- table(data$clicked_on_ad)
print(ad_table)
```

3.1.8 clicked on ad

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

People who both clicked on the ad and didn't click on the ad is the same (500 each).

3.2 Bivariate Analysis

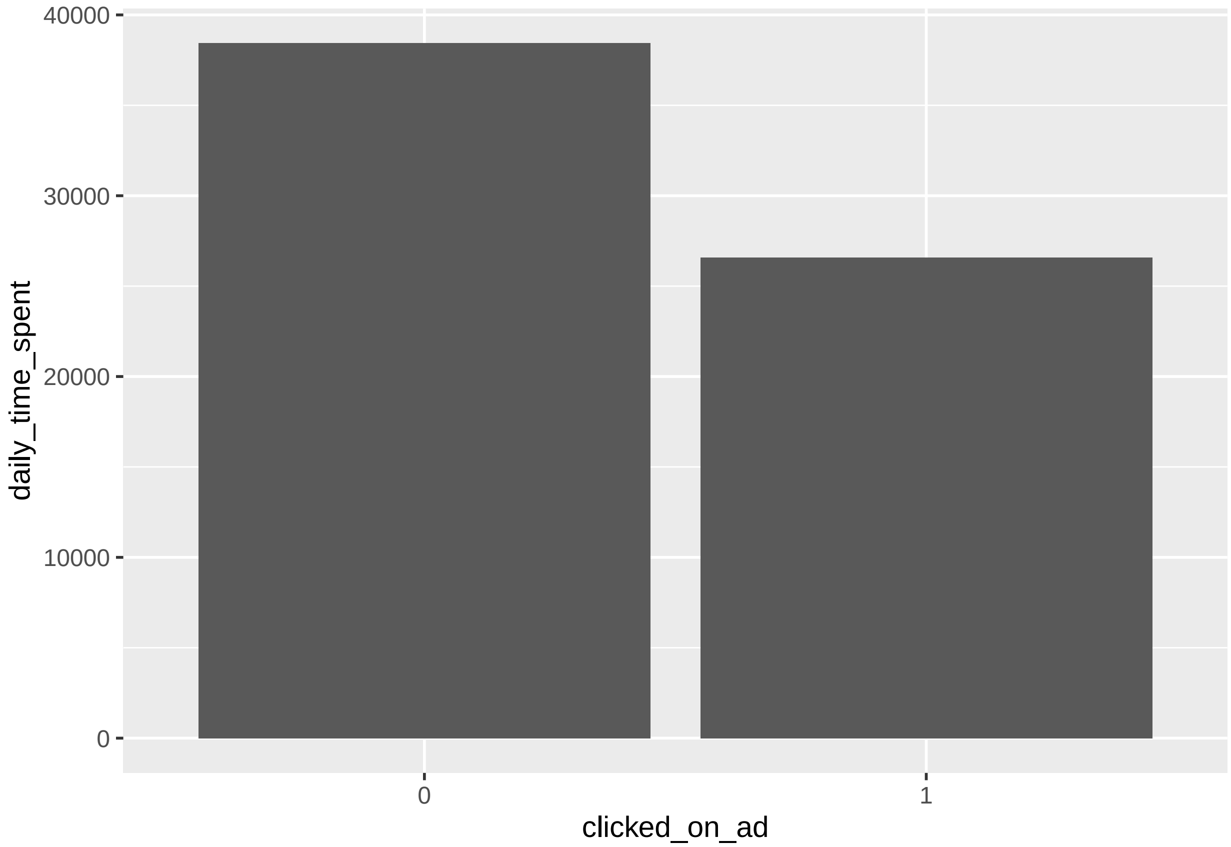
3.2.1 Research-specific Bivariate Analysis relationship between our target variable (clicked_on_ad) and the other variables.

```
# how many males clicked on ads
ad_male.table <- table(data$clicked_on_ad, data$male)
names(dimnames(ad_male.table)) <- c("Clicked on Ad?", "Male?")
ad_male.table
```

```
##                Male?
## Clicked on Ad?   0   1
##              0 250 250
##              1 269 231
```

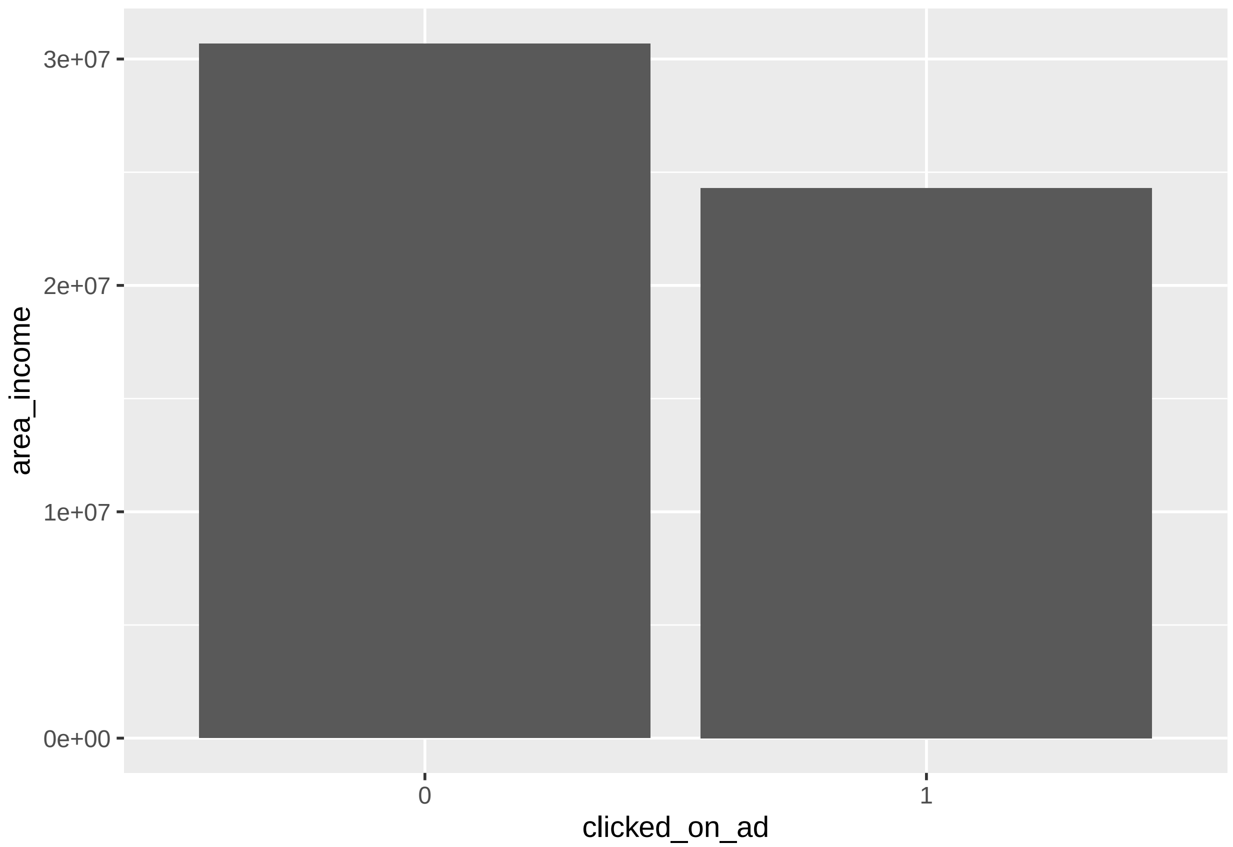
From this we see that of those who clicked on the ad, 269 were female while 231 were male. There was no difference in gender of those who did not click on the ad.

```
library(ggplot2);
ggplot(data, aes(clicked_on_ad,daily_time_spent)) +
    geom_bar(stat ="identity") +
    labs(y ="daily_time_spent",x ="clicked_on_ad")
```



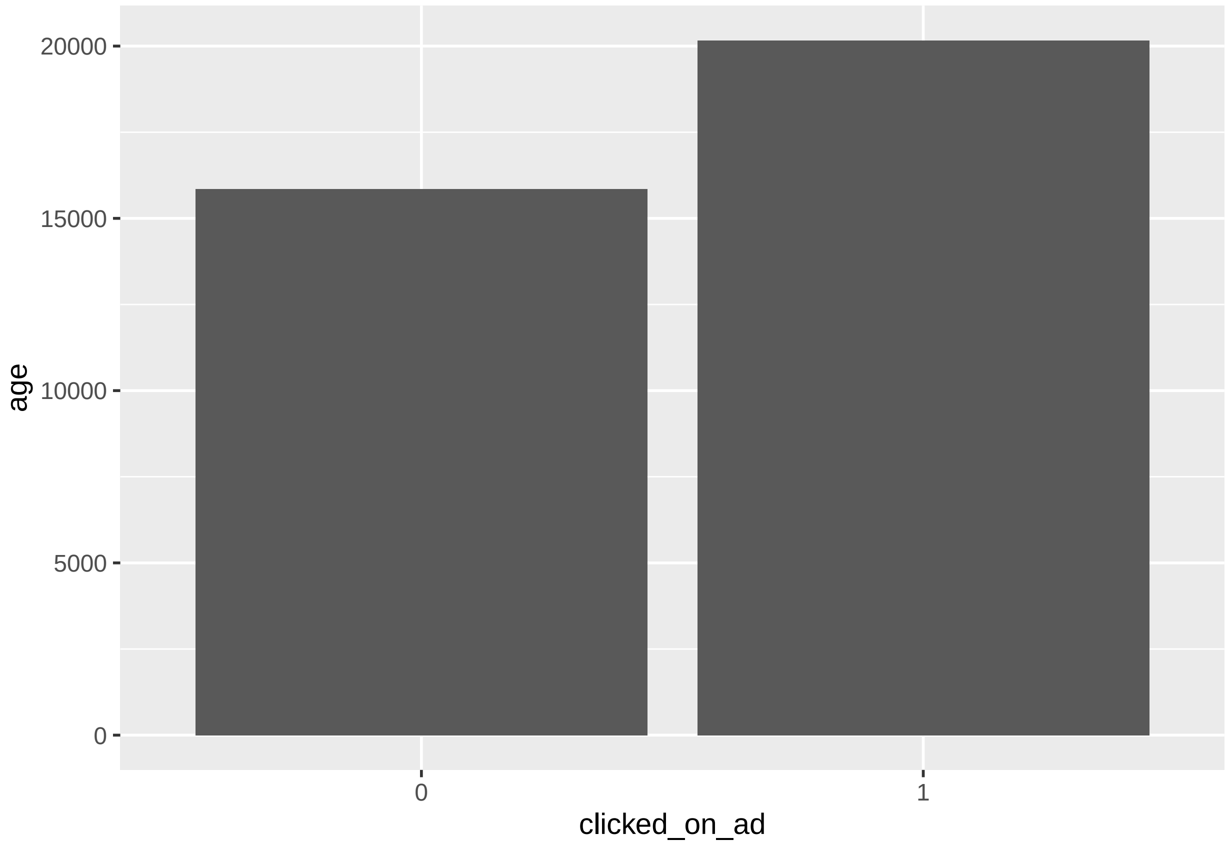
People who spent more time on the site did not click on the ad compared those who did.

```
ggplot(data, aes(clicked_on_ad, area_income)) +
    geom_bar(stat = "identity") +
    labs(y = "area_income", x = "clicked_on_ad")
```



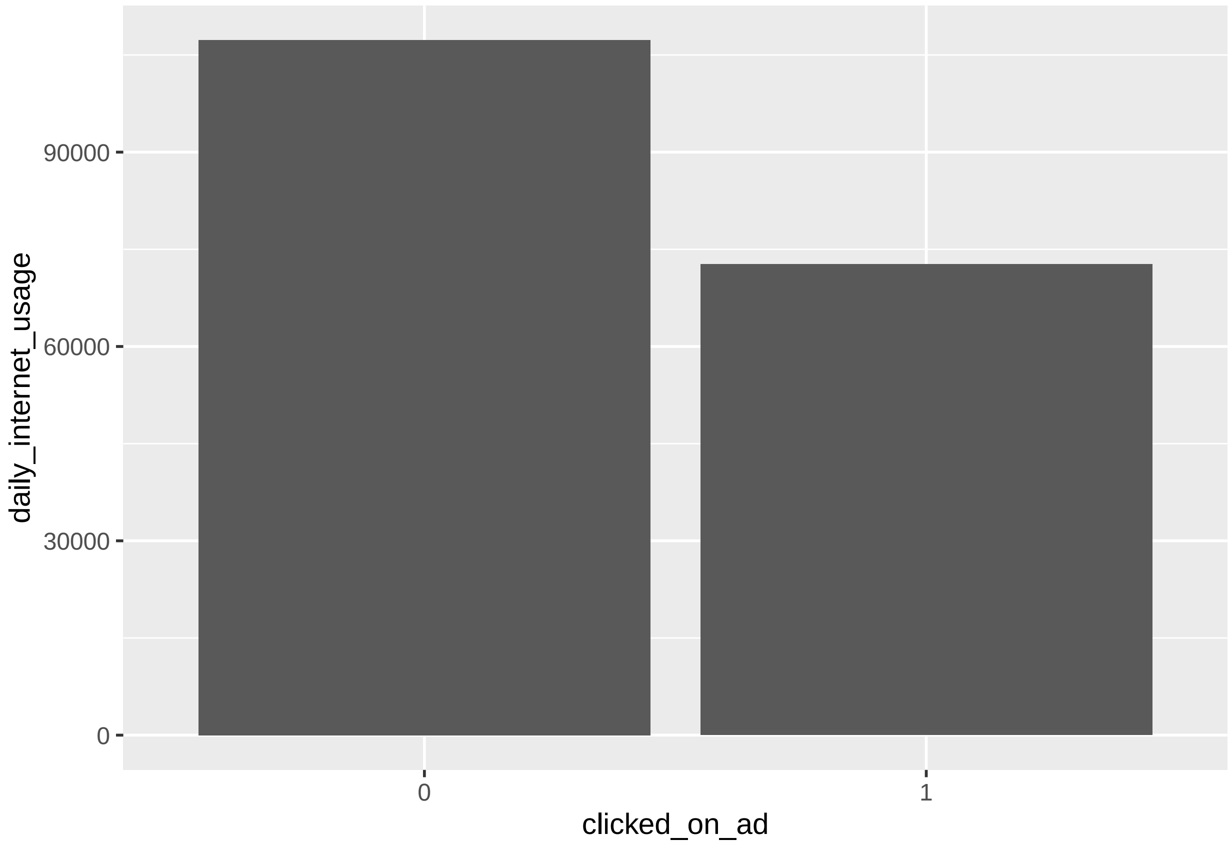
The higher the area income the lower the clicked ads and vice versa.

```
ggplot(data, aes(clicked_on_ad, age)) +
    geom_bar(stat = "identity") +
    labs(y = "age", x = "clicked_on_ad")
```



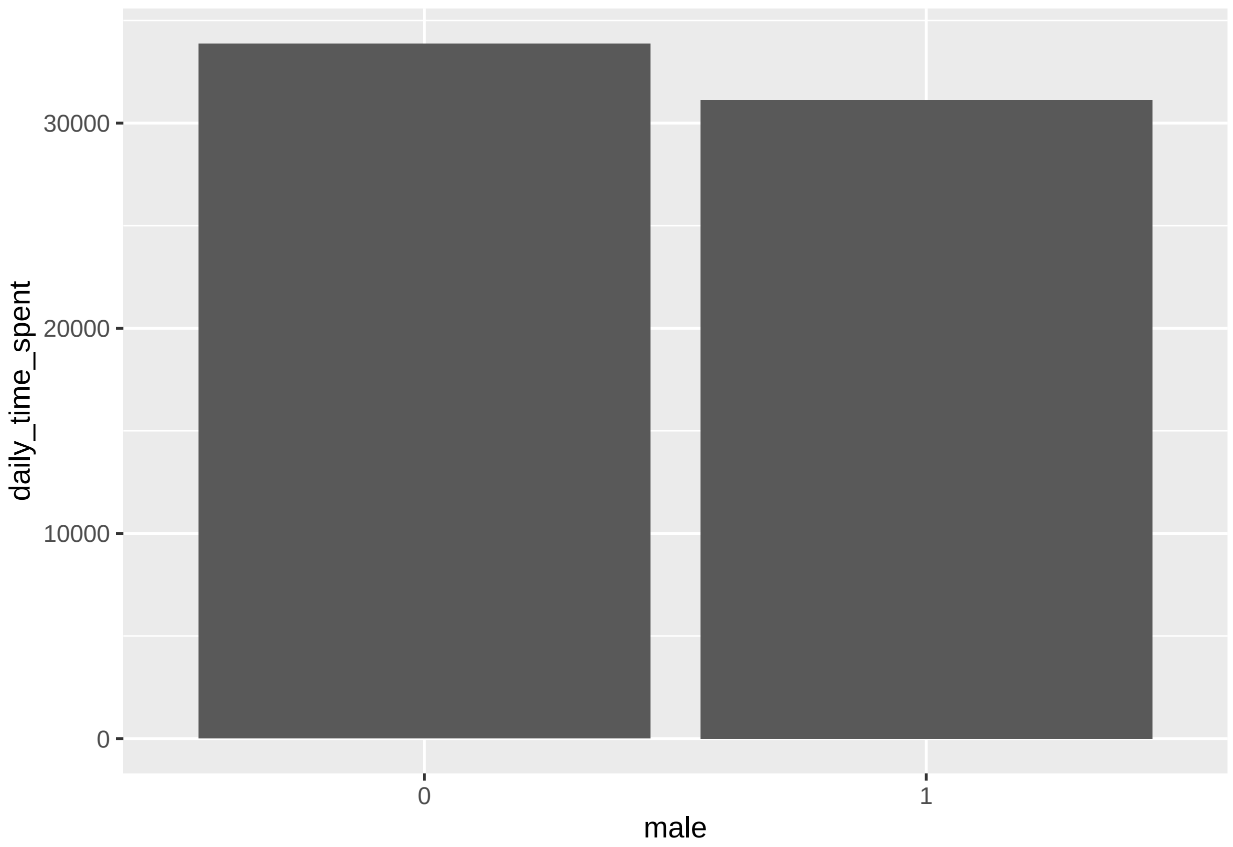
Older people clicked on ads compared to younger people.

```
ggplot(data, aes(clicked_on_ad, daily_internet_usage)) +
    geom_bar(stat = "identity") +
    labs(y = "daily_internet_usage", x = "clicked_on_ad")
```



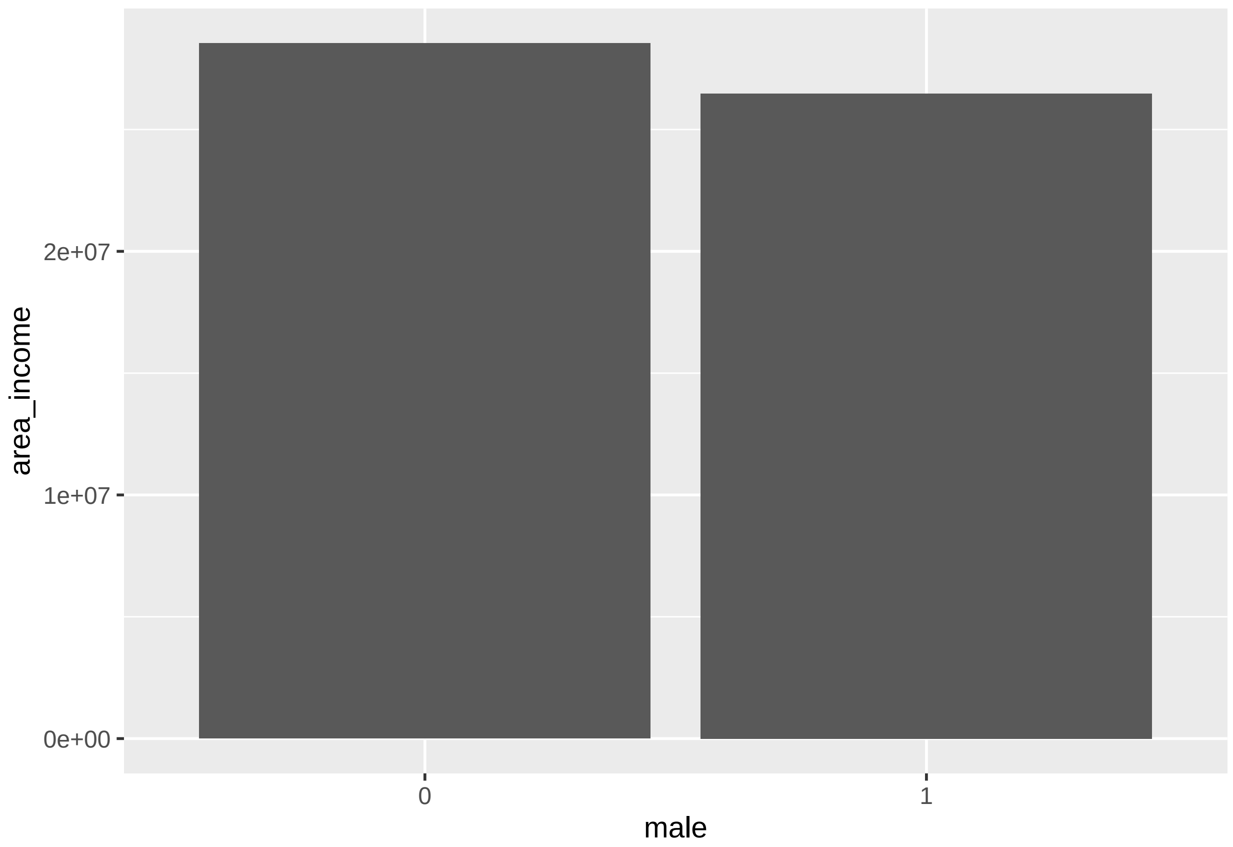
High internet usage did not necessarily mean higher clicks.

```
# a. Using Male Category

# Using bar charts to show the relationship between the Male and other independent variables.

library(ggplot2);
ggplot(data, aes(male,daily_time_spent)) +
    geom_bar(stat = "identity") +
    labs(y ="daily_time_spent",x ="male")
```



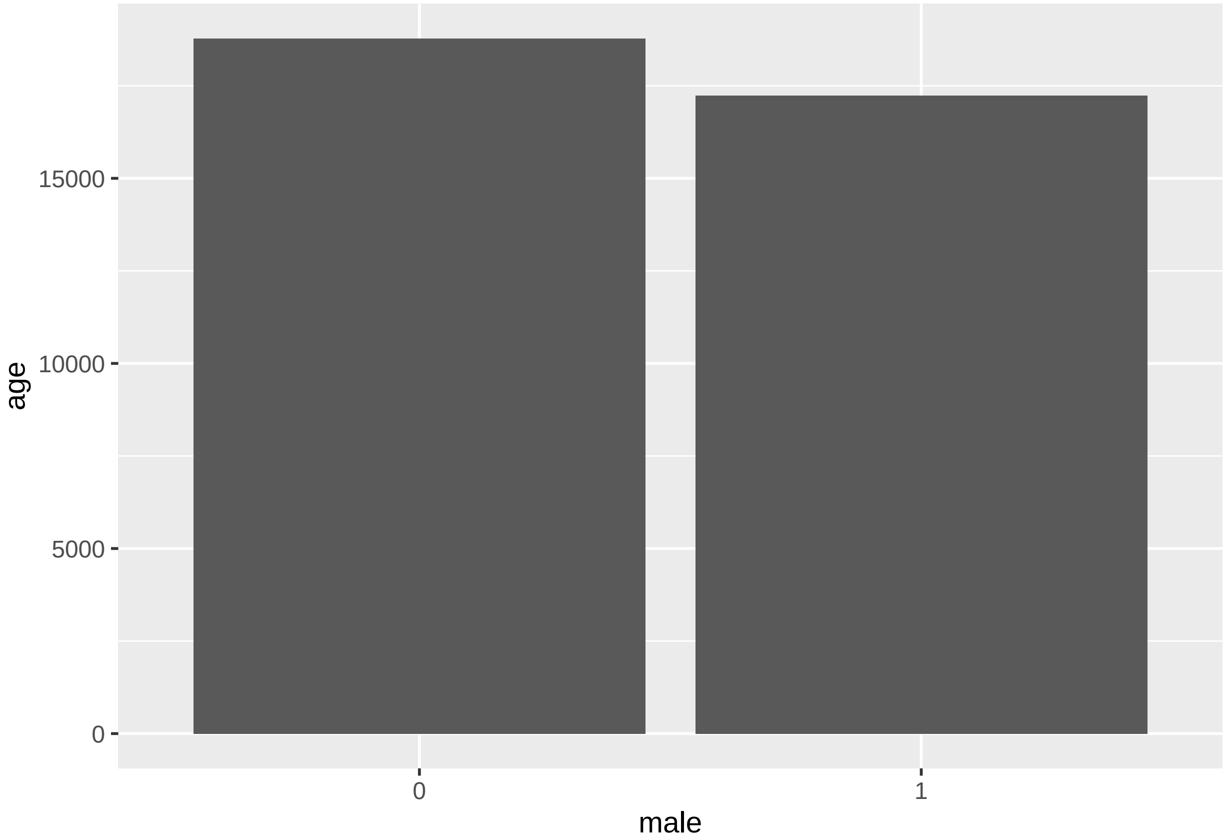
Non-males spend more time on the site compared to the males.

```
ggplot(data, aes(male, area_income)) +
    geom_bar(stat = "identity") +
    labs(y = "area_income", x = "male")
```



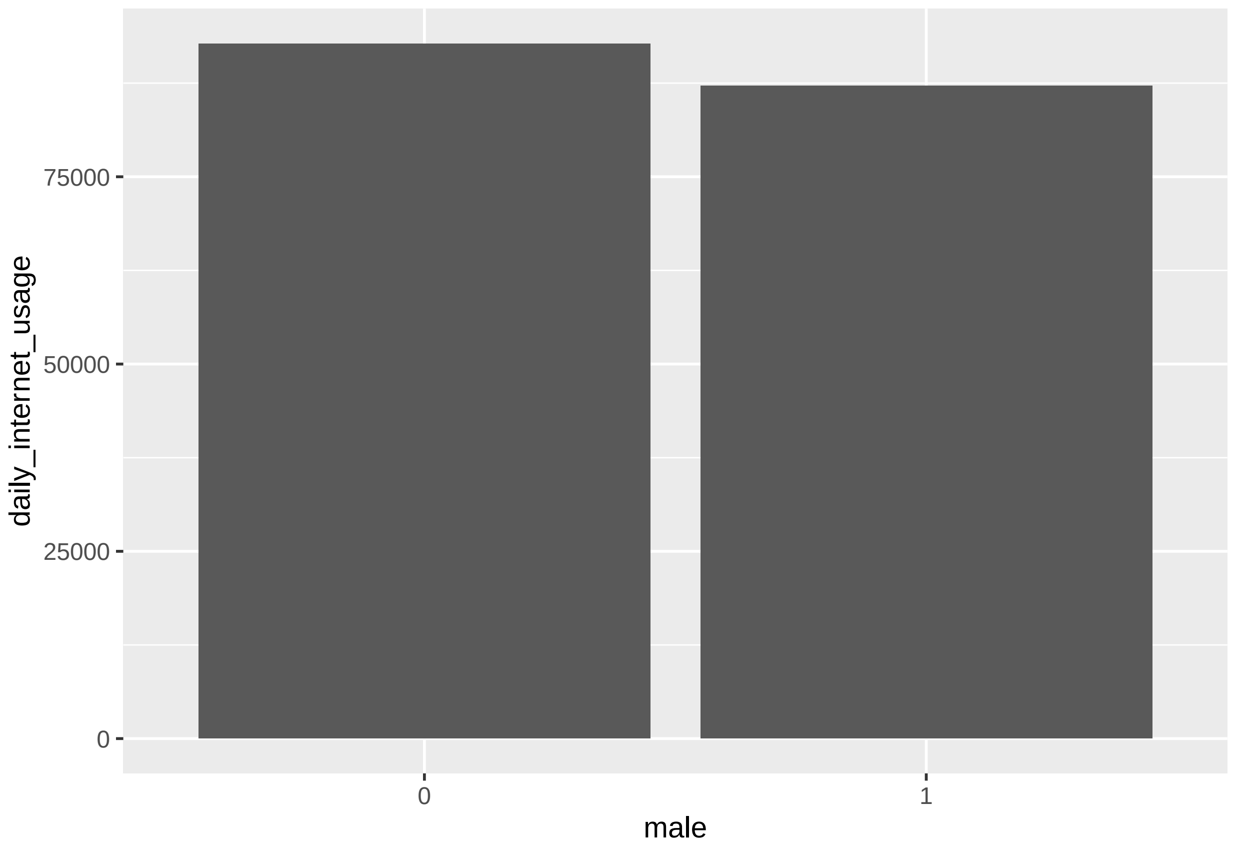
Males have lower area income compared to non-males.

```
ggplot(data, aes(male, age)) +
    geom_bar(stat = "identity") +
    labs(y = "age", x = "male")
```



Older folk are more of non-males than males.

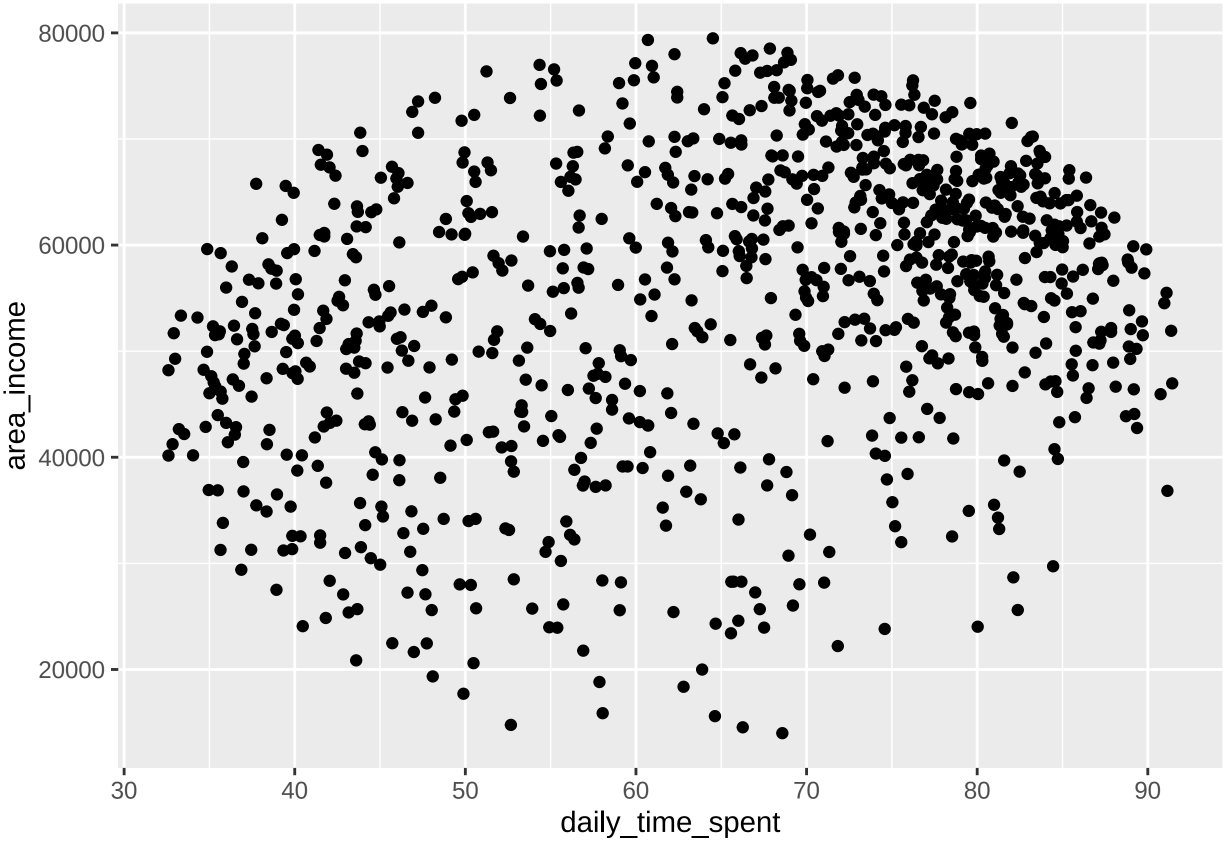
```
ggplot(data, aes(male, daily_internet_usage)) +
    geom_bar(stat = "identity") +
    labs(y = "daily_internet_usage", x = "male")
```



Males use less internet compared to the other gender.

Scatterplot for 2 continous variables.

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

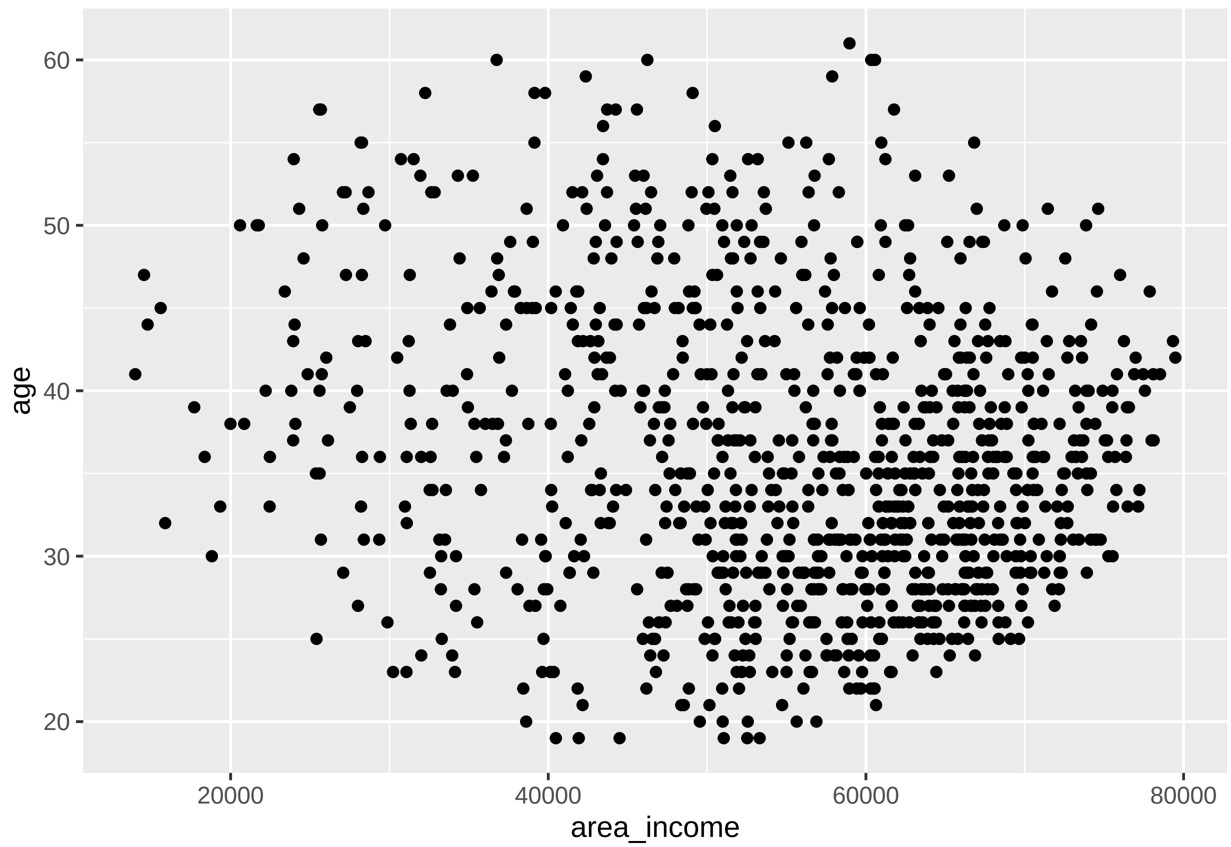


```
cor(data$area_income,data$daily_time_spent)
```

```
## [1] 0.3109544
```

correlation is 0.3109544 showing a positive relationship between the above variables. A change in time spent on the site leads to an increase in area income.

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

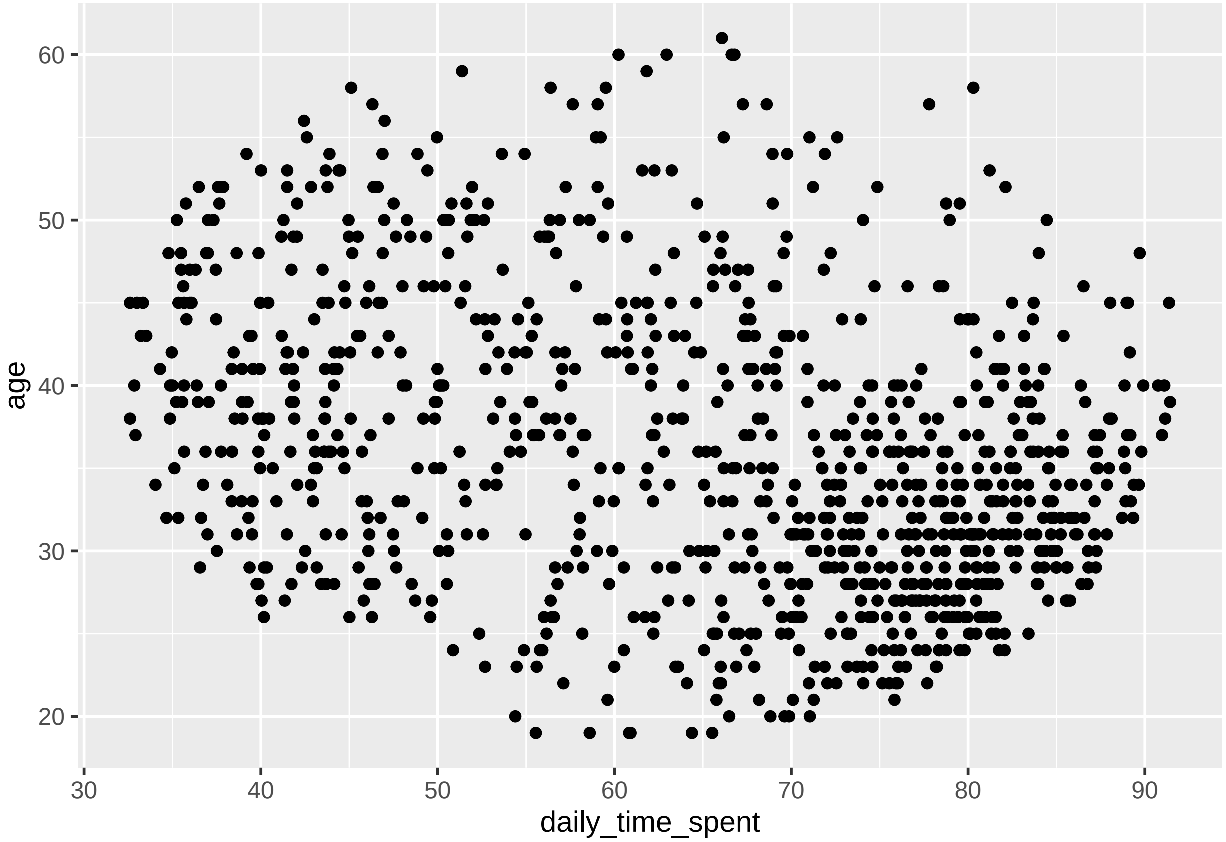


```
cor(data$area_income,data$age)
```

```
## [1] -0.182605
```

Correlation is -0.182605 showing a strong negative relationship between the above variables. Lower ages experience higher area income as they are tech savvy and use online platforms to make money.

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

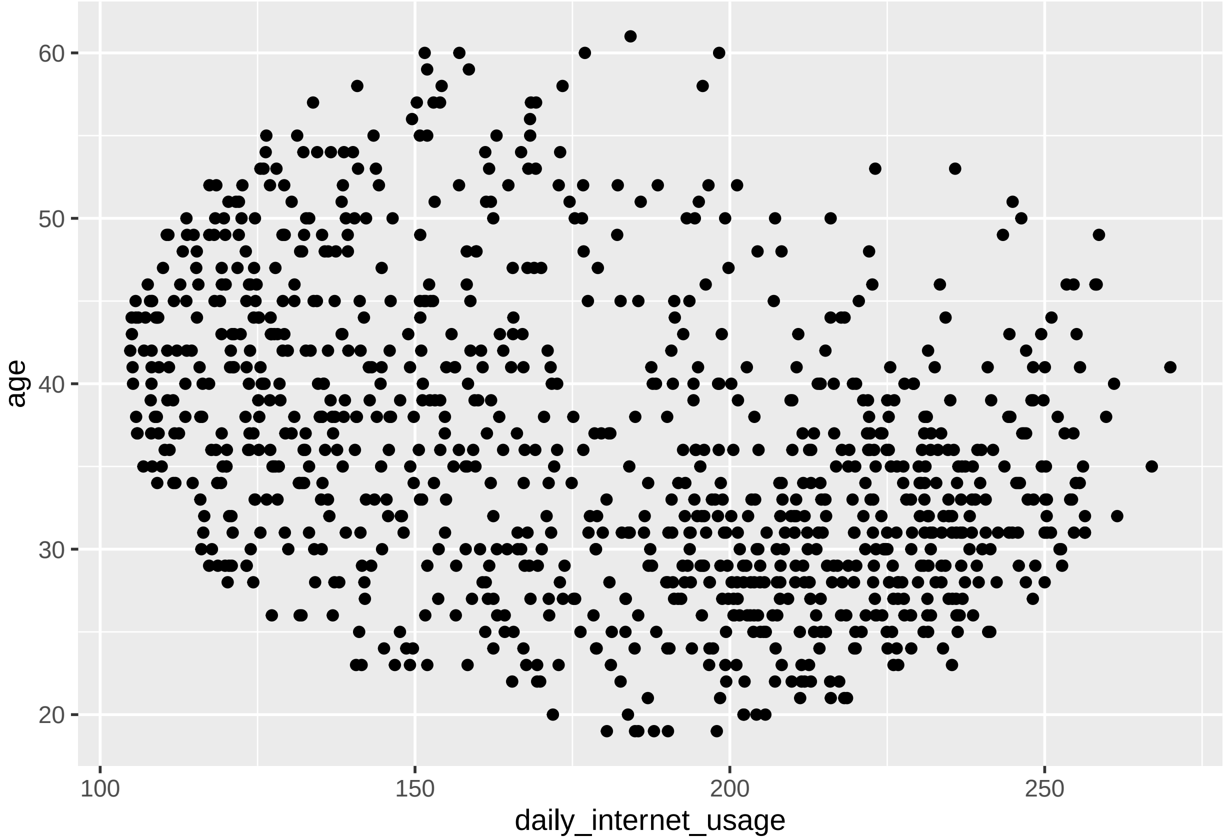


```
cor(data$age,data$daily_time_spent)
```

```
## [1] -0.3315133
```

Correlation is -0.3315133 showing a negative relationship between age and daily time spent on the site because younger people spend more time on the site unlike older people.

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

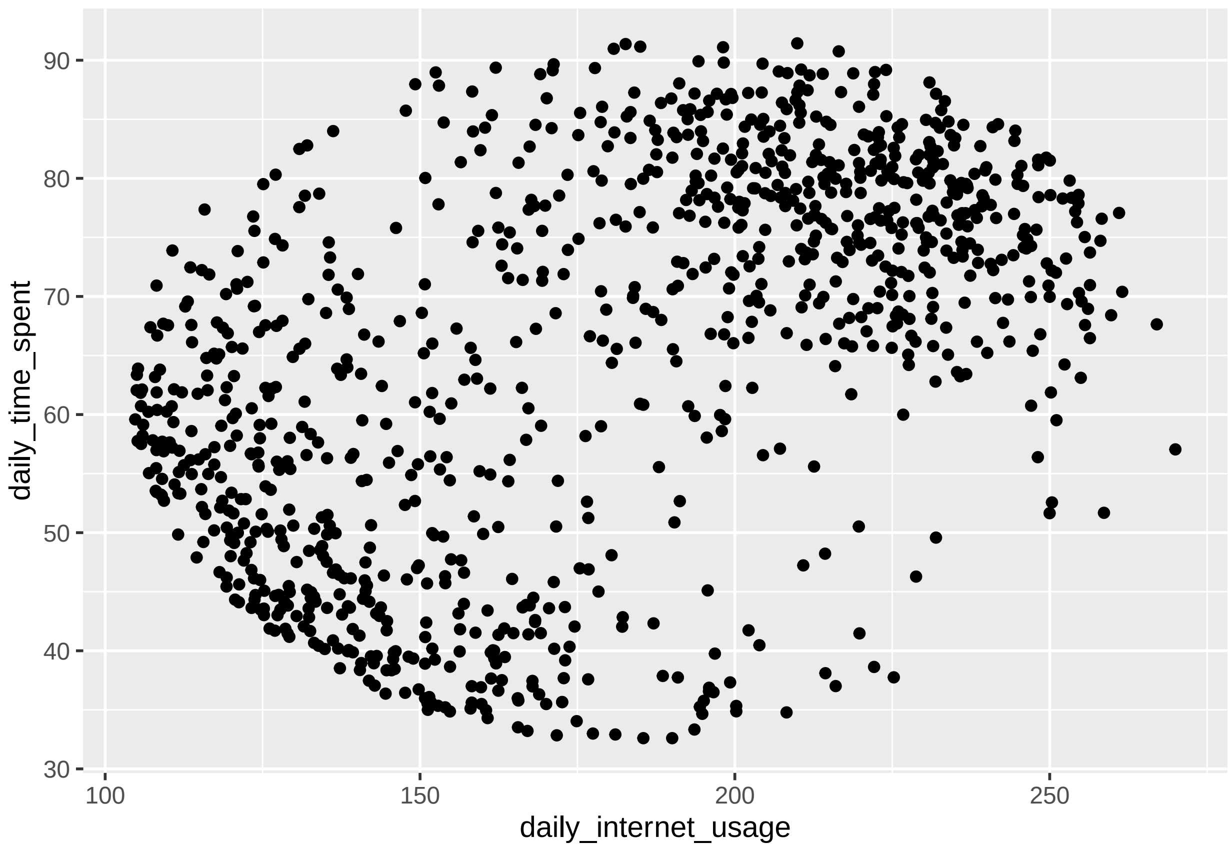


```
cor(data$age,data$daily_internet_usage)
```

```
## [1] -0.3672086
```

Correlation is -0.3672086 showing a negative relationship between age and daily internet usage because younger people spend more time on the site hence using more internet unlike older people.

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

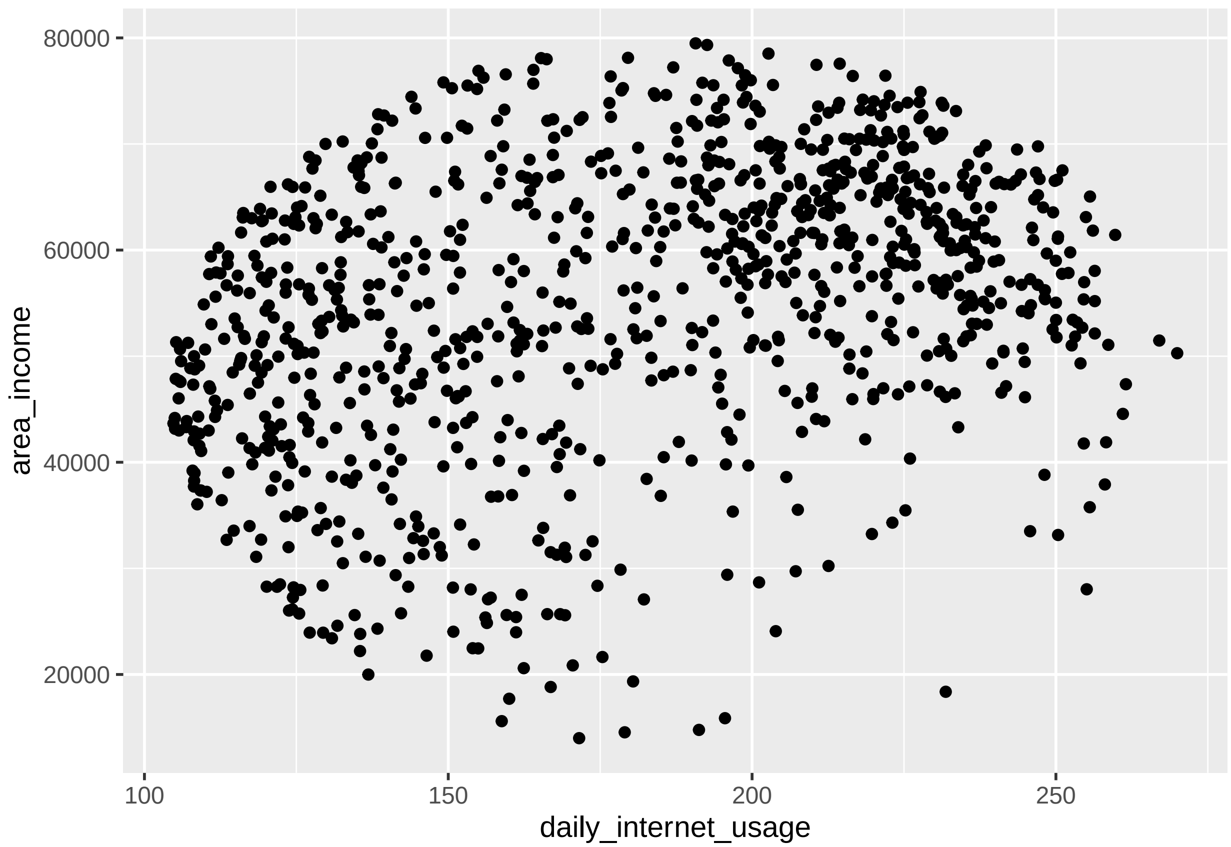


```
cor(data$daily_time_spent,data$daily_internet_usage)
```

```
## [1] 0.5186585
```

Correlation is 0.5186585 showing a strong positive relationship between the above variables obviously because more time on the site results in more internet usage.

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



```
cor(data$area_income,data$daily_internet_usage)
```

```
## [1] 0.3374955
```

Correlation is 0.3374955 showing a positive relationship between the above variables obviously because more area income results in more people being able to afford internet leading to more usage.

```
colnames(data)
```

```
##  [1] "daily_time_spent"     "age"                  "area_income"
##  [4] "daily_internet_usage" "ad_topic_line"        "city"
##  [7] "male"                 "country"              "timestamp"
## [10] "clicked_on_ad"
```

4. Modeling

Feature Engineering

```
#installing libraries to help in computation
# install.packages("caret")
library(lattice)
library(caret)
```

```
str(data)
```

```
## ’data.frame’:    1000 obs. of  10 variables:
##  $ daily_time_spent    : 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 standardizati
##  $ city                : chr  "Wrightburgh" "West Jodi" "Davidton" "West Terrifurt" ...
##  $ male                : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 2 2 ...
##  $ 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" "201
##  $ clicked_on_ad       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
```

```
# shuffling our data set to randomize the records
shuffle_index <- sample(1:nrow(data))
data <- data[shuffle_index, ]
dim(data)
```

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

```
head(data)
```

```
##     daily_time_spent age area_income daily_internet_usage
## 928            91.15  38    36834.04               184.98
## 96             72.84  26    52968.22               238.63
## 655            76.44  26    58820.16               224.20
## 352            73.71  23    56605.12               211.38
## 966            35.25  50    47051.02               194.44
## 541            77.47  36    70510.59               222.91
##                       ad_topic_line          city male     country
## 928         Robust uniform framework   Joshuamouth    0      Cyprus
## 96  Progressive empowering alliance      Wongland    0     Albania
## 655 Distributed intangible database     Troyville    1  Madagascar
## 352               Inverse local hub    New Sheila    1 El Salvador
## 966         Intuitive global website    Waltertown    0        Iran
## 541    Assimilated actuating policy Hatfieldshire    1     Jamaica
##               timestamp clicked_on_ad
## 928 2016-02-22 07:04:05             0
## 96  2016-07-08 21:18:32             0
## 655 2016-01-05 06:34:20             0
## 352 2016-02-15 05:35:54             0
## 966 2016-01-03 04:39:47             1
## 541 2016-03-14 03:29:12             0
```

```
# 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
```

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

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

```r
str(data)
```

```
## 'data.frame':    1000 obs. of  10 variables:
##  $ daily_time_spent    : num  91.2 72.8 76.4 73.7 35.2 ...
##  $ age                 : num  38 26 26 23 50 36 41 37 27 32 ...
##  $ area_income         : num  36834 52968 58820 56605 47051 ...
##  $ daily_internet_usage: num  185 239 224 211 194 ...
##  $ ad_topic_line       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ city                : chr  "Joshuamouth" "Wongland" "Troyville" "New Sheila" ...
##  $ male                : num  0 0 1 1 0 1 1 0 0 1 ...
##  $ country             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ timestamp           : chr  "2016-02-22 07:04:05" "2016-07-08 21:18:32" "2016-01-05 06:34:20" "201
##  $ clicked_on_ad       : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 2 1 1 ...
```

```r
# Normalizing the 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)
```

```r
head(data)
```

```
##     daily_time_spent       age area_income daily_internet_usage ad_topic_line
## 928       0.99524052 0.4523810   0.3487270               184.98            NA
## 96        0.68400476 0.1666667   0.5950944               238.63            NA
## 655       0.74519803 0.1666667   0.6844529               224.20            NA
## 352       0.69879313 0.0952381   0.6506295               211.38            NA
## 966       0.04504505 0.7380952   0.5047393               194.44            NA
## 541       0.76270610 0.4047619   0.8629647               222.91            NA
##              city male country           timestamp clicked_on_ad
## 928   Joshuamouth    0      NA 2016-02-22 07:04:05             0
## 96       Wongland    0      NA 2016-07-08 21:18:32             0
## 655     Troyville    1      NA 2016-01-05 06:34:20             0
## 352    New Sheila    1      NA 2016-02-15 05:35:54             0
## 966    Waltertown    0      NA 2016-01-03 04:39:47             1
## 541 Hatfieldshire    1      NA 2016-03-14 03:29:12             0
```

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

```
# checking the dimensions of our training and testing sets
dim(training)
```

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

```
dim(testing)
```

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

700 of data will be used for training while 300 will be for testing.

```
# checking the dimensions of our split
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
```

```
# convert the numerics 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  "0.99524052354241" "0.684004759476458" "0.762706102328744" "0.129015808
##  $ age                 : chr  "0.452380952380952" "0.166666666666667" "0.404761904761905" "0.42857142
##  $ area_income         : chr  "0.348727024521937" "0.595094390906467" "0.86296468224095" "0.631599537
##  $ daily_internet_usage: chr  "184.98" "238.63" "222.91" "136.99" ...
##  $ ad_topic_line       : chr  NA NA NA NA ...
##  $ city                : chr  "Joshuamouth" "Wongland" "Hatfieldshire" "Adamsbury" ...
##  $ male                : chr  "0" "0" "1" "0" ...
##  $ country             : chr  NA NA NA NA ...
##  $ timestamp           : chr  "2016-02-22 07:04:05" "2016-07-08 21:18:32" "2016-03-14 03:29:12" "2016
##  $ clicked_on_ad       : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 2 2 ...
```

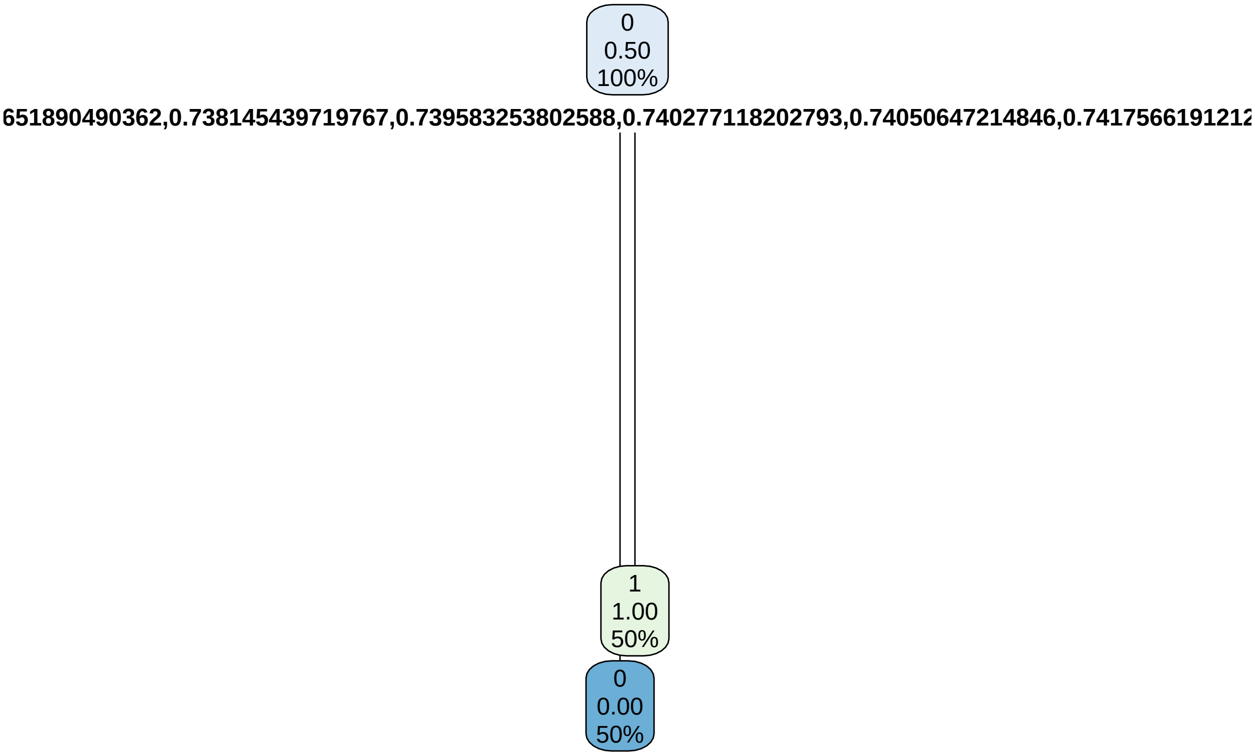
```
str(training)
```

```
## 'data.frame':    700 obs. of  10 variables:
##  $ daily_time_spent    : chr  "0.99524052354241" "0.684004759476458" "0.762706102328744" "0.12901580
##  $ age                 : chr  "0.452380952380952" "0.166666666666667" "0.404761904761905" "0.4285714
##  $ area_income         : chr  "0.348727024521937" "0.595094390906467" "0.86296468224095" "0.6315995
##  $ daily_internet_usage: chr  "184.98" "238.63" "222.91" "136.99" ...
##  $ ad_topic_line       : chr  NA NA NA NA ...
##  $ city                : chr  "Joshuamouth" "Wongland" "Hatfieldshire" "Adamsbury" ...
##  $ male                : chr  "0" "0" "1" "0" ...
##  $ country             : chr  NA NA NA NA ...
##  $ timestamp           : chr  "2016-02-22 07:04:05" "2016-07-08 21:18:32" "2016-03-14 03:29:12" "201
##  $ clicked_on_ad       : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 2 2 ...
```

Decision Tree

```
library(rpart)
library(rpart.plot)
```

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



5. Conclusion

People who spent more time on the site did not click on the ad compared those who did. The higher the area income the lower the clicked ads and vice versa. Older people clicked on ads compared to younger people. High internet usage does not necessarily mean higher clicks. Females spend more time on the site compared to the males. Males have lower area income compared to females. Older people are more of non-males than males. Males use less internet compared to the other gender. A change in time spent on the site leads to an increase in area income. Lower ages experience higher area income as they are tech savvy and use online platforms to make money. Younger people spend more time on the site unlike older people. Younger people spend more time on the site hence using more internet unlike older people. more area income results in more people being able to afford internet leading to more usage.

##6. Recommendation The entrepreneur should focus more on the male clientelle and areas with low income and on people who spend less time on the site and use less internet.

They should also consider older people as they have shown to bring forth more clicks than the younger ones.